主管单位：中华人民共和国工业和信息化部
主办单位：北京航空航天大学

ISSN 1001-5965
CN 11-2625/V

北京航空航天大學学报

JOURNAL OF BEIJING UNIVERSITY OF AERONAUTICS AND ASTRONAUTICS



基于多尺度空间注意力引导的图像超分辨率重建网络

程德强　王培杰　董彦强　寇旗旗　江　鹤

7 | 2025
Vol.51

北京航空航天大学学报

第 51 卷　第 7 期　（总第 389 期）　2025 年 7 月

目　　次

基于多尺度空间注意力引导的图像超分辨率重建网络 …… 程德强，王培杰，董彦强，寇旗旗，江鹤（2185）

基于模糊逻辑与自适应策略的红外可见光图像融合 …… 杨勇，刘家祥，黄淑英，王晓争，夏钰锟（2196）

空间信息增强的室内多任务 RGB-D 场景理解 …… 孙国栋，熊晨韵，刘俊杰，张杨（2209）

基于坐标感知注意的多帧自监督单目深度估计 …… 程德强，范舒铭，钱建生，江鹤，寇旗旗（2218）

递归门控增强与金字塔预测的铁路全景分割 …… 陈永，周方春，张娇娇（2229）

用于长时视觉跟踪的级联目标漂移判定网络 …… 侯志强，赵佳鑫，陈语，马素刚，余旺盛，范九伦（2240）

基于动态嵌入特征的鲁棒半监督视频目标分割 …… 陈亚当，赵翊冰，吴恩华（2253）

基于联合交互注意力的图文情感分析方法 …… 胡慧君，丁子毅，张耀峰，刘茂福（2262）

基于语义信息引导的多标签图像分类 …… 黄俊，范浩东，洪旭东，李雪（2271）

基于双分支特征增强和多级轨迹关联的多目标跟踪算法 …… 马素刚，段帅鹏，侯志强，余旺盛，蒲磊，杨小宝（2282）

基于变换学习的快速多切片 MRI 重建算法 …… 段继忠，刘欢（2290）

结合上下文关联的图像情感分析 …… 罗改芳，张浩，徐丹（2304）

基于 IcD-FDRL 的应急监控视频边缘智能传输优化 …… 李彦，万征，邓承志，汪胜前（2314）

基于自适应多态蚁群优化的智能体路径规划 …… 邢娜，邸昊天，尹文杰，韩亚君，周洋（2330）

语言引导视觉的小样本航拍图像目标检测 …… 张智，易华挥，郑锦（2338）

基于特征增强与自适应阈值非极大值抑制的目标检测算法 …… 孟伟君，安雯，马素刚，杨小宝（2349）

基于信噪比的学习型哈里斯鹰优化算法 …… 张林，沈佳颖，胡传陆，朱东林（2360）

面向低延迟视频压缩感知的搜索窗自适应重构 …… 孙仁慧，刘浩，邓开连，燕帅（2374）

无模态融合的高效弱监督视频时刻检索算法 …… 蒋寻，徐行，沈复民，王国庆，杨阳（2384）

面向红外弱小舰船检测的轻量化神经网络设计 …… 唐文婷，李波，季梦奇（2394）

基于多视口自适应融合的全景图像质量评价 …… 冯晨曦，张地，林敢，叶龙（2404）

基于小波变换和平行注意力的多源遥感图像分类 …… 王嘉毅，高峰，张天戈，甘言海（2415）

显著性感知三重正则化相关滤波无人机目标跟踪算法 …… 贺冰，王法胜，王星，孙福明（2423）

基于无监督深度学习的航拍图像拼接算法 …… 梁镇锋，夏海英，谭玉枚，宋树祥（2437）

基于语义规则的自适应情感词典自动构建算法 …… 卫青蓝，何雨，宋金宝（2450）

基于差异性特征蒸馏的多模态连续学习方法 …… 贺驰原，程少旭，许林峰，孟凡满，吴庆波（2460）
基于多元优化的批量图像隐写载体选择方法 …… 王仰光，姚远志，俞能海（2468）
基于多标签对抗领域自适应的行人属性识别算法 …… 胡强梁，陈琳，尚明生（2478）
基于双路视觉 Transformer 的图像风格迁移 …… 纪宗杏，贝佳，刘润泽，任桐炜（2488）
基于旋转目标感知网络的 SAR 船舶检测方法 …… 王梓懿，尹嘉豪，黄博斌，高峰（2498）
针对航拍小目标检测的 YOLOv7 改进方法 …… 刘一诺，张琪，王蓉，李冲（2506）
多目视觉下基于融合特征的密集行人跟踪方法 …… 黄煜杰，陈凯，王子源，王紫腾（2513）
基于三元组哈希损失的半监督图像检索 …… 邵伟志，熊思宇，潘丽丽（2526）
面向基于知识图谱个性化推荐的诱导信息识别 …… 倪文锴，彭舒凡，杜彦辉（2538）
基于人眼视觉机制的伪装目标检测网络 …… 张冬冬，王春平，付强（2553）
基于自适应阈值和速度优化的轻量化语义 VSLAM 方法 ……
…… 齐浩，付悦欣，胡祝华，吴佳琪，赵瑶池（2562）

期刊基本参数：CN 11-2625/V * 1956 * m * A4 * 388 * zh * P * ￥50.00 * 400 * 36 * 2025-07

（**编　辑**　张　嵘　孙　芳　万玉敏　卞欢欢　王　茜　李亚泰）

JOURNAL OF BEIJING UNIVERSITY OF
AERONAUTICS AND ASTRONAUTICS

Vol. 51 No. 7 （Sum 389） July 2025

CONTENTS

Image super-resolution reconstruction network based on multi-scale spatial attention guidance
…… CHENG Deqiang，WANG Peijie，DONG Yanqiang，KOU Qiqi，JIANG He （2185）

Fuzzy logic and adaptive strategy for infrared and visible light image fusion
…… YANG Yong，LIU Jiaxiang，HUANG Shuying，WANG Xiaozheng，XIA Yukun （2196）

Spatial information-enhanced indoor multi-task RGB-D scene understanding
…… SUN Guodong，XIONG Chenyun，LIU Junjie，ZHANG Yang （2209）

Coordinate-aware attention-based multi-frame self-supervised monocular depth estimation
…… CHENG Deqiang，FAN Shuming，QIAN Jiansheng，JIANG He，KOU Qiqi （2218）

Railway panoramic segmentation based on recursive gating enhancement and pyramid prediction
…… CHEN Yong，ZHOU Fangchun，ZHANG Jiaojiao （2229）

Cascaded object drift determination network for long-term visual tracking
…… HOU Zhiqiang，ZHAO Jiaxin，CHEN Yu，MA Sugang，YU Wangsheng，FAN Jiulun （2240）

Robust semi-supervised video object segmentation based on dynamic embedding feature
…… CHEN Yadang，ZHAO Yibing，WU Enhua （2253）

Analysis of image and text sentiment method based on joint and interactive attention
…… HU Huijun，DING Ziyi，ZHANG Yaofeng，LIU Maofu （2262）

Semantic information-guided multi-label image classification
…… HUANG Jun，FAN Haodong，HONG Xudong，LI Xue （2271）

Multi-object tracking algorithm based on dual-branch feature enhancement and multi-level trajectory association
…… MA Sugang，DUAN Shuaipeng，HOU Zhiqiang，YU Wangsheng，PU Lei，YANG Xiaobao （2282）

Fast multi-slice MRI reconstruction algorithm based on transform learning
…… DUAN Jizhong，LIU Huan （2290）

Image sentiment analysis by combining contextual correlation
…… LUO Gaifang，ZHANG Hao，XU Dan （2304）

Edge-intelligent transmission optimization of emergency surveillance video based on IcD-FDRL
…… LI Yan，WAN Zheng，DENG Chengzhi，WANG Shengqian （2314）

Path planning for agents based on adaptive polymorphic ant colony optimization
…… XING Na，DI Haotian，YIN Wenjie，HAN Yajun，ZHOU Yang （2330）

Few-shot object detection of aerial image based on language guidance vision
…… ZHANG Zhi，YI Huahui，ZHENG Jin （2338）

An object detection algorithm based on feature enhancement and adaptive threshold non-maximum suppression
…… MENG Weijun，AN Wen，MA Sugang，YANG Xiaobao （2349）

Learning Harris Hawks optimization algorithm with signal-to-noise ratio
…… ZHANG Lin，SHEN Jiaying，HU Chuanlu，ZHU Donglin （2360）

Adaptive search window reconstruction for low-delay video compressive sensing
…… SUN Renhui，LIU Hao，DENG Kailian，YAN Shuai （2374）

Efficient weakly-supervised video moment retrieval algorithm without multimodal fusion
…… JIANG Xun，XU Xing，SHEN Fumin，WANG Guoqing，YANG Yang （2384）

Lightweight neural network design for infrared small ship detection
…… TANG Wenting，LI Bo，JI Mengqi （2394）

Omnidirectional image quality assessment based on adaptive multi-viewport fusion
…… FENG Chenxi，ZHANG Di，LIN Gan，YE Long （2404）

Multi-source remote sensing image classification based on wavelet transform and parallel attention
…… WANG Jiayi，GAO Feng，ZHANG Tiange，GAN Yanhai （2415）

Saliency-aware triple-regularized correlation filter algorithm for UAV object tracking
…… HE Bing，WANG Fasheng，WANG Xing，SUN Fuming （2423）

Aerial image stitching algorithm based on unsupervised deep learning
…… LIANG Zhenfeng，XIA Haiying，TAN Yumei，SONG Shuxiang （2437）
An adaptive automatic construction algorithm for sentiment dictionaries based on semantic rules
…… WEI Qinglan，HE Yu，SONG Jinbao （2450）
Continual learning method based on differential feature distillation for multimodal network
…… HE Chiyuan，CHENG Shaoxu，XU Linfeng，MENG Fanman，WU Qingbo （2460）
Cover selection method for batch image steganography based on multivariate optimization
…… WANG Yangguang，YAO Yuanzhi，YU Nenghai （2468）
Pedestrian attribute recognition algorithm based on multi-label adversarial domain adaptation
…… HU Qiangliang，CHEN Lin，SHANG Mingsheng （2478）
Dual-channel vision Transformer-based image style transfer
…… JI Zongxing，BEI Jia，LIU Runze，REN Tongwei （2488）
A rotated content-aware retina network for SAR ship detection
…… WANG Ziyi，YIN Jiahao，HUANG Bobin，GAO Feng （2498）
Improved YOLOv7 method for aerial small target detection in aerial photography
…… LIU Yinuo，ZHANG Qi，WANG Rong，LI Chong （2506）
A dense pedestrian tracking method based on fusion features under multi-vision
…… HUANG Yujie，CHEN Kai，WANG Ziyuan，WANG Ziteng （2513）
Semi-supervised image retrieval based on triplet hash loss
…… SHAO Weizhi，XIONG Siyu，PAN Lili （2526）
Identification of induced information for personalized recommendations based on knowledge graph
…… NI Wenkai，PENG Shufan，DU Yanhui （2538）
Camouflaged object detection network based on human visual mechanisms
…… ZHANG Dongdong，WANG Chunping，FU Qiang （2553）
A lightweight semantic VSLAM approach based on adaptive thresholding and speed optimization
…… QI Hao，FU Yuexin，HU Zhuhua，WU Jiaqi，ZHAO Yaochi （2562）

http://bhxb.buaa.edu.cn jbuaa@buaa.edu.cn
DOI: 10.13700/j.bh.1001-5965.2023.0547

基于多尺度空间注意力引导的图像超分辨率重建网络

程德强[1]，王培杰[1]，董彦强[2]，寇旗旗[3]，江鹤[1,*]

(1. 中国矿业大学 信息与控制工程学院，徐州 221116；2. 国家电力投资集团 内蒙古白音华煤电有限公司露天矿，锡林郭勒盟 026000；3. 中国矿业大学 计算机科学与技术学院，徐州 221116)

摘 要：针对基于注意力机制的图像超分辨率重建网络忽视了注意力特征的差异性，仅将注意力机制直接引入到网络模型中，对不同层次特征进行相同处理的问题，设计了一种多尺度空间注意力引导的图像超分辨率重建网络SAGN。提出了增强特征提取残差块(ERB)，完善了局部信息的表征能力；集成了多尺度空间注意力(MSA)模块，获取了MSA特征信息；引入了注意力引导模块(AGM)，对不同的特征分配个性化的权重，以实现有效的上下文全局特征融合和冗余信息抑制。实验结果表明：量化测试和主观效果上，相比于传统的注意力结构，SAGN在4个基准数据集上都展现出了优越性，其4倍重建结果的峰值信噪比(PSNR)较次优模型平均提高了0.05 dB，进一步证实了SAGN在恢复图像的几何结构和细节方面的优势。

关 键 词：超分辨率重建；卷积神经网络；注意力机制；多尺度空间注意力；注意力引导
中图分类号：TP391
文献标志码：A **文章编号**：1001-5965(2025)07-2185-11

随着信息技术的发展，图像作为传递信息的载体在各行各业都不可或缺，但由于成像设备、拍摄环境等因素影响，分辨率低影响图像的信息量及其视觉效果。图像超分辨率(super resolution, SR)重建是一种在不依赖硬件设备改进的情况下提高图像分辨率的技术，可以将低分辨率(low resolution, LR)图像恢复至相应的高分辨率(high resolution, HR)图像。因此，图像超分辨率重建是图像恢复任务中的研究热点，并应用于诸多领域，如视频监控[1]、遥感影像[2]、医学诊断[3]等。

基于卷积神经网络(convolutional neural network, CNN)的超分辨率重建技术取得了较大的发展，性能极大地优于传统方法。Dong等[4]在超分辨率重建中应用CNN，提出了由3层构成的超分辨率卷积神经网络(super-resolution convolution neural network, SRCNN)，尽管取得了优于传统方法的性能，但由于LR图像上采样在前，限制了网络重建效率。因此，Dong等[5]提出了快速超分辨率卷积神经网络(fast super-resolution convolutional neural network, FSRCNN)，在SRCNN[4]的基础上引入了反卷积操作，降低了输入图像的分辨率，使网络训练更快更有效。Kim等[6]引入了残差学习，设计了更深的CNN结构来提高重建性能，提出了更深层CNN的单幅图像超分辨率(very deep convolutional neural network for single image super-resolution, VDSR)方法，有效缓解了训练深度网络时带来的梯度消失问题。Lim等[7]提出了单幅图像超分辨率增强深度残差网络(enhanced deep residual network for single image super-resolution, EDSR)方法，通过移除ResNet[8]残差块中的批归一化(batch normalization, BN)层，提升了

收稿日期：2023-08-28；**录用日期**：2023-10-29；**网络出版时间**：2023-11-15 16:02
网络出版地址：link.cnki.net/urlid/11.2625.V.20231115.1126.001
基金项目：国家自然科学基金(52204177,52304182)；中央高校基本科研业务费专项资金(2020QN49)
*通信作者. E-mail：jianghe@cumt.edu.cn

引用格式：程德强，王培杰，董彦强，等. 基于多尺度空间注意力引导的图像超分辨率重建网络[J]. 北京航空航天大学学报，2025，51(7)：2185-2195. CHENG D Q，WANG P J，DONG Y Q，et al. Image super-resolution reconstruction network based on multi-scale spatial attention guidance[J]. Journal of Beijing University of Aeronautics and Astronautics，2025，51(7)：2185-2195 (in Chinese).

图像超分辨率重建效果。程德强等[9]提出了利用递归残差网络的多通道超分辨率重建网络MCSR，通过多通道网络交叉融合特征信息，减少参数量的同时提高模型重建能力。Hui等[10]将信息蒸馏技术引入超分辨率重建，并提出信息多蒸馏网络(information multi-distillation network, IMDN)，模型能够更好地选择和利用层次特征。Li等[11]提出了多尺度残差网络(multi-scale residual network, MSRN)，引入不同大小的卷积核来获取多尺度特征，并利用层次特征融合的方式提高了重建性能。Liu等[12]提出了交叉卷积超分辨率网络(cross convolution super resolution network, Cross-SRN)，利用交叉卷积来学习图像边缘特征，并将多尺度特征融合嵌入交叉卷积块中，能有效重建出图像边缘信息。

传统的基于CNN的模型通过加深网络来获取更好的性能，无差别地对待各层特征，但随着注意力机制的提出，使CNN可以对各层特征进行差异化处理，进而提升模型的性能。Liu等[13]提出了残差特征聚合网络(residual feature aggregation network, RFANet)，并进一步引入了增强空间注意力(enhanced spatial attention, ESA)，使残差特征有效关注空间内容。Wang等[14]提出了具有注意力辅助特征学习的超分辨率网络AAF，利用注意力辅助特征促进信息传递，有效提升了网络的性能。Chen等[15]提出了图像超分辨率注意力网络中的注意力(attention in attention network, AAN)及动态注意力模块，动态分配2个分支的权重，抑制不重要的注意力特征，权重可以根据输入特征自适应变化。Liang等[16]提出了基于近期流行的Transformer的超分模型SwinIR，计算量较大，对硬件要求高。基于注意力机制的超分辨率网络都取得了满意的效果，但仍存在一些局限性。像素注意力网络(pixel attention network, PAN)[17]等方法的注意力模块关注的感受野相对较小，大多通过全连接层局部提取注意力特征，无法捕捉到整幅图像像素之间的全局信息和远程依赖关系，且提取注意力特征尺度单一，对高频和低频信息的辨别能力有限。RFANet[13]等方法将注意力机制直接引入到图像超分辨率方法中，对不同层次的特征做相同的处理，忽略了注意力是否都能选择性地捕获有效的特征信息，即并非所有注意力分支都有同样的贡献。虽然AAN[15]考虑到这一问题，但其只是在同一残差块中对注意力和非注意力进行动态平衡，并没有考虑到不同残差信息之间长期远程依赖的关系。

本文提出了一种基于多尺度空间注意力引导的图像超分辨率重建网络SAGN架构，该方法有效捕获长期远距离依赖关系的通道和空间上下文信息，扩大感受野并关注不同尺度信息，以实现更有效的特征表示，对恢复纹理细节多的图像具有一定的优势。①提出了增强特征提取残差块(enhanced feature extraction residual block, ERB)，构造出直接传播信息的局部路径。②集成了多尺度空间注意力(multi-scale spatial attention, MSA)模块，根据空间缩放信息学习到不同尺度的注意力特征信息。③引入了注意力引导模块(attention guidance module, AGM)，将MSA特征有效融合至残差块中，抑制了冗余信息，提高了残差特征的表征能力。

1 本文方法

1.1 模型整体结构

本文提出的SAGN网络包括3个部分：浅层特征提取(shallow feature extraction, SFE)、空间注意力引导残差组(spatial attention guidance residual group, SAGG)和RM。RM采用高效亚像素卷积神经网络(efficient sub-pixel convolutional neural network, ESPCN)[18]中像素洗牌的方式，如图1所示。

首先，给定$\boldsymbol{I}_{\mathrm{LR}}$和$\boldsymbol{I}_{\mathrm{SR}}$分别作为模型输入的LR图像块和输出的超分辨率图像块。SAGN先经过3×3单层卷积从$\boldsymbol{I}_{\mathrm{LR}}$中提取浅层特征$\boldsymbol{F}_0$，该层操作可表示为

$$\boldsymbol{F}_0 = H_{\mathrm{SFE}}(\boldsymbol{I}_{\mathrm{LR}}) = \mathrm{Conv3\times3}(\boldsymbol{I}_{\mathrm{LR}}) \tag{1}$$

式中：H_{SFE}为浅层特征提取函数；$\mathrm{Conv}_{3\times3}(\cdot)$表示3×3卷积层。

然后，将浅层特征$\boldsymbol{F}_0$输入到SAGG中进行深层特征学习，进一步可表示为

$$\boldsymbol{F}_{\mathrm{DF}} = \boldsymbol{F}_0 + \boldsymbol{F}_G = \boldsymbol{F}_0 + H_G(H_{G-1}(\cdots(H_1(\boldsymbol{F}_0))\cdots)) \tag{2}$$

式中：$\boldsymbol{F}_{\mathrm{DF}}$为提取的深层特征，即$\boldsymbol{F}_G$和$\boldsymbol{F}_0$逐元素相加；$H_G$为第$G$个SAGG函数；$\boldsymbol{F}_G$为最后一个SAGG的输出。

最后，通过图像RM对深层特征$\boldsymbol{F}_{\mathrm{DF}}$进行上采样重建，即

$$\boldsymbol{I}_{\mathrm{SR}} = H_{\mathrm{RM}}(\boldsymbol{F}_{\mathrm{DF}}) = H_{\mathrm{SAGN}}(\boldsymbol{I}_{\mathrm{LR}}) \tag{3}$$

式中：H_{RM}为重建函数；H_{SAGN}为SAGN网络函数。

为对SAGN模型进行优化，本文利用L_1损失函数使其最小。给定N个LR图像及其对应HR图像的训练集，参数集为θ的网络L_1损失函数为

$$L_1(\theta) = \frac{1}{N}\sum_{i=1}^{N}\left\|H_{\mathrm{SAGN}}(\boldsymbol{I}_{\mathrm{LR}}^i) - \boldsymbol{I}_{\mathrm{HR}}^i\right\|_1 \tag{4}$$

1.2 空间注意力引导残差组

如图1所示，SAGG能够有效利用残差块中每层提取的特征，这些特征被分层地聚合到空间注意力分支中，以便于下一个残差块使用，其目的是减

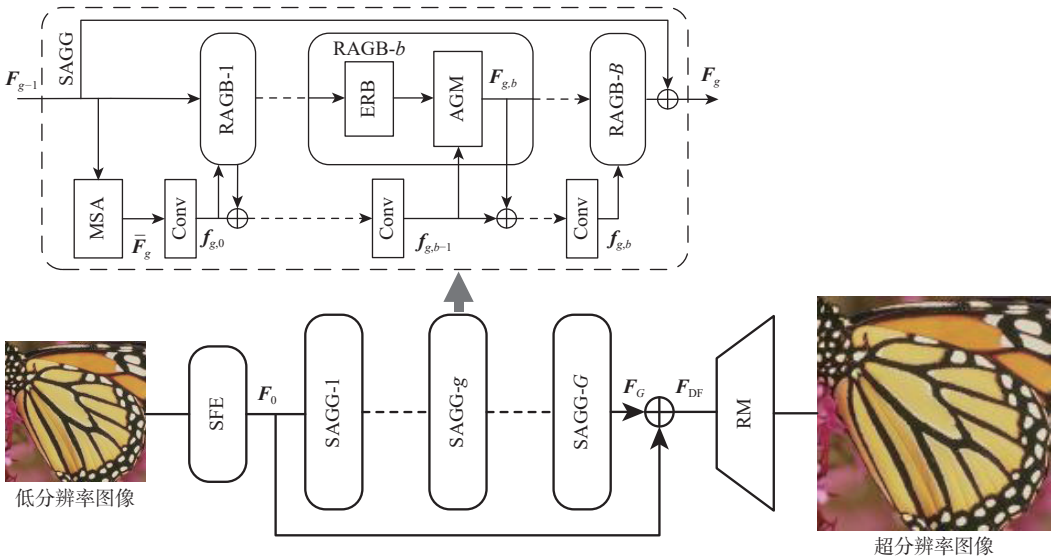


图 1　基于多尺度空间注意力引导的图像超分辨率重建网络

Fig. 1　Image super-resolution reconstruction network based on multi-scale spatial attention guidance

少由连续的残差聚集引起的特征退化, 提高特征信息的流动性。SAGG 由 1 个 MSA 和 8 个注意力引导残差块 (residual attention guidance block, RAGB) 组成。输入特征 $\boldsymbol{F}_{g-1}$ 分别进入 SAGG 中的网络主干和注意力支路, 其中, 注意力支路通过 MSA 得到注意力映射权重 $\boldsymbol{W}_{\mathrm{MSA}}$ 与输入特征 $\boldsymbol{F}_{g-1}$ 逐元素相乘, 即

$$\overline{\boldsymbol{F}}_g = \boldsymbol{W}_{\mathrm{MSA}} \otimes \boldsymbol{F}_{g-1} \tag{5}$$

式中: $\overline{\boldsymbol{F}}_g$ 为第 g 个 SAGG 带有空间注意力的加权特征; ⊗表示逐元素相乘。

第 g 个 SAGG 特征提取过程 $\boldsymbol{F}_g$ 可以表示为

$$\begin{aligned}\boldsymbol{F}_g &= \boldsymbol{F}_{g-1} + H_g(\boldsymbol{F}_{g-1}) = \\ &\boldsymbol{F}_{g-1} + H_{g,b}(H_{g,b-1}(\cdots H_{g,1}(\boldsymbol{F}_{g-1}, \boldsymbol{f}_{g,0})\cdots), \boldsymbol{f}_{g,b-1})\end{aligned} \tag{6}$$

式中: $\boldsymbol{F}_g$ 为第 g 个 SAGG 的输出特征; $H_{g,b}$ 为第 g 个 SAGG 中第 b 个 RAGB 函数。如图 1 所示, 需要将网络主干特征 $\boldsymbol{F}_{g,b}$ 与注意力支路特征图 $\boldsymbol{f}_{g,b}$ 并行输入到 RAGB 中, 且 $\boldsymbol{f}_{g,b} = \mathrm{Conv1\times1}(\boldsymbol{F}_{g,b} + \boldsymbol{f}_{g,b-1})$。

1.3　多尺度空间注意力

为最大限度地发挥 AGM 的有效性, 受到 RFA-Net[13] 中 ESA 的启示, 提出了一种 MSA 模块, 如图 2 所示, 其中, k 为卷积核大小, s 为步长, p 为填充大小。该模块根据空间信息缩放特征, 使模块学习到更多有区别的不同尺度的注意力特征, 与 AGM 配合使用, 增强特征选择的能力。表 1 为多尺度空间注意模块的张量输入输出尺寸, 其中, C 为通道数, H 为高, W 为宽。先使特征图进入 1×1 卷积层, 压缩其通道数, 连续使用 2 个不同卷积核的跨步卷积 (stride=2) 进行下采样来增大其感受野, 中间通过 1 个 3×3 卷积层后, 连续使用 2 个不同卷积核的跨步反卷积 (stride=2), 再经过 1 个 1×1 卷积层恢复通道数, 用于上采样恢复图像尺寸的同时, 还能提高对空间信息的敏感性。过程中, 该模块还使用了 2 个跳跃连接, 保留更多的高频信息。最终, 通过 Sigmoid 层生成带有权重注意力的特征图。

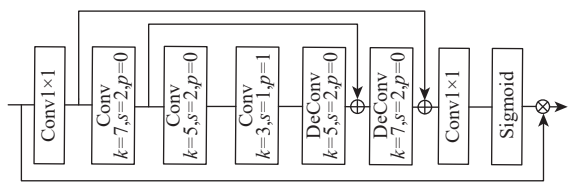


图 2　多尺度空间注意力模块

Fig. 2　Multi-scale spatial attention module

表 1　多尺度空间注意力模块张量尺寸

Table 1　Tensor dimensions of multi-scale spatial attention module

卷积层名称	张量输入尺寸	张量输出尺寸
Conv1×1	(C,H,W)	$(C/2,H,W)$
Conv k=7,s=2,p=0	$(C/2,H,W)$	$(C/4,H_1=(H-7)/2+1,$ $W_1=(W-7)/2+1)$
Conv k=5,s=2,p=0	$(C/4,H_1,W_1)$	$(C/8,H_2=(H_1-5)/2+1,$ $W_2=(W_1-5)/2+1)$
Conv k=3,s=1,p=1	$(C/8,H_2,W_2)$	$(C/8,H_2,W_2)$
DeConv k=5,s=2,p=0	$(C/8,H_2,W_2)$	$(C/4,H_1,W_1)$
DeConv k=7,s=2,p=0	$(C/4,H_1,W_1)$	$(C/2,H,W)$
Conv1×1	$(C/2,H,W)$	(C,H,W)

1.4　注意力引导残差块

本节构建 RAGB, 由 ERB 和 AGM 构成, 从优化基础残差块 (residual block, RB) 和引入 AGM 这 2 个方面克服深层次导致信息丢失的缺陷, 并保持了其处理空间信息的能力, 提高了网络重建效率。其思想是: 权衡主干网络特征和注意力分支特征的比例, 以便在残差块输出后使用。由此, 第 g 个 SAGG

中的第 b 个 RAGB 可表示为

$$\boldsymbol{F}_{g,b}=\boldsymbol{F}_{g,b-1}+\mathrm{AGM}_{g,b}(\mathrm{ERB}(\boldsymbol{F}_{g,b-1}),\boldsymbol{f}_{g,b-1}) \tag{7}$$

式中：$\boldsymbol{F}_{g,b-1}$ 和 $\boldsymbol{F}_{g,b}$ 分别为 RAGB 的输入和输出；$\mathrm{AGM}_{g,b}(\cdot)$为注意力引导函数；ERB(·)为增强特征提取残差块函数；$\boldsymbol{f}_{g,b-1}$ 为注意力分支的特征输入。

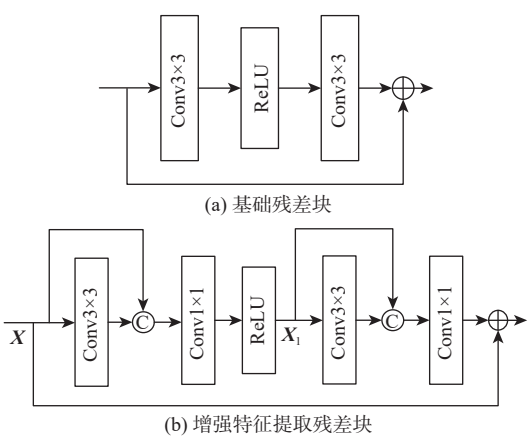
1.4.1 增强特征提取残差块

在基于 CNN 的图像超分辨率重建模型中，大多数模型都采用了 EDSR[7] 中的 RB，如图 3(a) 所示，其由 2 个 3×3 标准卷积和 1 个 ReLU 激活函数构成。由于图像特征信息传播的直接路径限制在具有相同特征映射大小的局部区域，限制了特征信息在上下文之间的流动性，为构造局部路径直接传播信息，对 RB 进行了改进，并提出了 ERB，该残差块通过局部路径传播特征信息，如图 3(b) 所示。在 RB 的基础上增加 2 个 1×1 逐点卷积，并将经过 3×3 标准卷积前后的图像特征分别在通道维度上拼接融合，通过逐点卷积还原特征映射的维数，实现全局交互和信息融合，有效减少图像信息丢失。整个残差学习过程可表示为

$$\boldsymbol{X}_1=\delta(\mathrm{Conv1\times1}(\mathrm{Cat}(\boldsymbol{X},\mathrm{Conv3\times3}(\boldsymbol{X})))) \tag{8}$$

$$\mathrm{ERB}(\boldsymbol{X})=\boldsymbol{X}+\mathrm{Conv1\times1}(\mathrm{Cat}(\boldsymbol{X}_1,\mathrm{Conv3\times3}(\boldsymbol{X}_1))) \tag{9}$$

式中：$\mathrm{Conv}_{i\times i}(\cdot)$表示卷积核尺寸为 $i\times i$ 的卷积操作；Cat(·)表示特征拼接融合操作；$\delta(\cdot)$表示 ReLU 激活函数；$\boldsymbol{X}$ 为输入特征；$\boldsymbol{X}_1$ 为中间特征。



(a) 基础残差块

(b) 增强特征提取残差块

图 3 卷积残差块

Fig. 3 Convolution residual block

1.4.2 注意力引导模块

由于网络需要经过连续重复的若干个残差块，导致提取图像特征信息较为单一且剩余特征利用率低，这也是网络深度不断增加所带来的缺陷。注意力机制能够缓解这一缺陷，但不是所有的注意力特征都能准确捕获有效的信息，有时网络中的冗余参数也会成为阻碍网络性能提升的重要原因之一。因此，受 AAN[15] 在同一残差块中对注意力和非注意力进行动态平衡的启示，提出了一种 AGM，其能加强不同残差块之间远程信息的依赖关系，是一种个性化分配权重的双输入单输出的结构设计。如图 4 所示，该模块同时处理主干特征 $\boldsymbol{X}_{\mathrm{att}}$ 和注意力特征 $\boldsymbol{Y}_{\mathrm{att}}$，学习和平衡 2 个分支的权重比例，能够在解决特征退化问题的同时，更有效地帮助 SAGN 融合上下文全局特征。

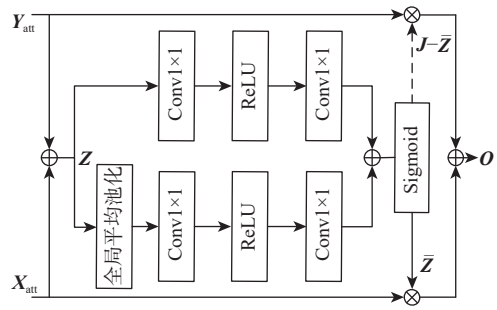


图 4 注意力引导模块

Fig. 4 Attention guidance module

给定 2 个输入 $\boldsymbol{X}_{\mathrm{att}}$、$\boldsymbol{Y}_{\mathrm{att}}$。进入模块后，先进行加融合的操作，即：$\boldsymbol{Z}=\boldsymbol{X}_{\mathrm{att}}+\boldsymbol{Y}_{\mathrm{att}}$。对 $\boldsymbol{Z}$ 进行注意力特征提取，即

$$\mathrm{Global}(\boldsymbol{Z})=\mathrm{Conv1\times1}(\delta(\mathrm{Conv1\times1}(\mathrm{global}(\boldsymbol{Z})))) \tag{10}$$

$$\mathrm{Local}(\boldsymbol{Z})=\mathrm{Conv1\times1}(\delta(\mathrm{Conv1\times1}(\boldsymbol{Z}))) \tag{11}$$

式中：Global($\boldsymbol{Z}$) 为 $\boldsymbol{Z}$ 的全局通道注意力特征；Local($\boldsymbol{Z}$) 为 $\boldsymbol{Z}$ 的局部通道注意力特征；global($\boldsymbol{Z}$) 为全局平均池化，即

$$\mathrm{global}(\boldsymbol{Z})=\frac{1}{H\times W}\sum_{i=1}^{H}\sum_{j=1}^{W}Z(i,j) \tag{12}$$

对全局注意力特征和局部注意力特征相加融合，再通过 Sigmoid 激活函数生成注意力权值。

$$\overline{\boldsymbol{Z}}=\sigma(\mathrm{Local}(\boldsymbol{Z})+\mathrm{Global}(\boldsymbol{Z})) \tag{13}$$

式中：$\overline{\boldsymbol{Z}}$为归一化注意力特征；$\sigma$表示 Sigmoid 激活函数。

$\overline{\boldsymbol{Z}}$分别对 2 个分支的特征逐点相乘，分配注意力权重比例。

$$\boldsymbol{O}=\boldsymbol{X}_{\mathrm{att}}\otimes\overline{\boldsymbol{Z}}+\boldsymbol{Y}_{\mathrm{att}}\otimes(\boldsymbol{J}-\overline{\boldsymbol{Z}}) \tag{14}$$

式中：$\boldsymbol{J}$ 为全 1 矩阵。$\boldsymbol{J}-\overline{\boldsymbol{Z}}$如图 4 中虚线所示。

2 实 验

2.1 实验设置

实验训练使用 ubuntu18.04，编程语言 Python 3.6，深度学习框架为 Pytorch 1.7，Cuda11.0 版本，处理器为 Intel(R) Core (TM) i9-10980XE CPU@ 3.00 GHz，18 核 36 线程，系统内存为 32 GB，显卡为 NVIDIA RTX 3090，显存为 24 GB。训练数据集采用 DIV2K

数据集中的 800 张训练图像, 测试集使用基准数据集 Set5[19]、Set14[20]、B100[21] 和 Urban100[22]。每批次从 LR 图像中随机裁剪 16 个大小为 48×48 的图像块输入模型, 训练时对训练图像进行数据增强, 包括随机旋转 90°、180°、270°和水平翻转, 共训练 500 轮, 初始学习率为 10^{-4}, 学习率每 200 个周期减半。使用 ADAM 优化器训练, $\beta_1=0.9$, $\beta_2=0.999$, $\varepsilon=10^{-8}$。模型性能评价指标采用峰值信噪比 (peak signal to noise ratio, PSNR) 和结构相似性 (structural similarity, SSIM)[23]。PSNR 是最常用衡量图像失真的评价指标, 单位为 dB, PSNR 值越高, 表示图像超分辨率重建质量越好。SSIM 从亮度、对比度和结构 3 个方面衡量图像质量, 当 2 幅图像结构越相似时, SSIM 值越接近 1。在图像 y 通道上计算 PSNR 和 SSIM。

2.2　模型性能分析

2.2.1　不同模块对模型性能的影响

设置 5 组实验, 验证 SAGN 中各模块的有效性。在尺度因子为 4 下训练 500 轮, 使用 PSNR 评价指标进行消融研究。表 2 为 5 组实验分别加入各模块后在基准数据集上的 PSNR 结果。其中, 模型 EDSR_RB 使用通道数为 96、残差块数为 16 的 EDSR 作为此消融实验的基准, 与加入其他模块后的结果相比较。由表 2 可以看出, 模型 SAGN_ERB 是将 RB 替换为 ERB, 在各数据集上均有提升, 尤其在 Set5 数据集上, 相较于模型 EDSR_RB, PSNR 提升了 0.07 dB, ERB 增加了特征信息在残差块之间的流动性。模型 SAGN_ESA 与模型 SAGN_MSA 对比可见, 使用 MSA 模块在各数据集的指标大幅提升, SAGN 并没有大量堆积空间注意力, 因此, MSA 更适应 SAGN, 对网络性能的提升有很大帮助。当单独加入 AGM 时, 模型 SAGN_AGM 相较于 SAGN_ERB 的指标有所提升, 尤其是在 Urban100 数据集上, PSNR 提升了 0.1 dB。将本文提出的所有模块加入使用, SAGN 的性能指标最高, 由此可见, 当 AGM 和 SAGN 配合使用时, 模型能够主动权衡主干特征和注意力特征配比, 融合了上下文全局特征, 有效提升了模型性能。

表 2　尺度因子为 4 时加入不同模块后对模型性能的影响

Table 2　Effect of model performance after adding different modules when scale factor is 4

模型	RB	ERB	ESA	MSA	AGM	参数量	浮点计算量	PSNR/dB			
								Set5	Set14	B100	Urban100
EDSR_RB	√					3.58×10^6	223.85 GFLOPs	32.21	28.64	27.59	26.12
SAGN_ERB		√				4.17×10^6	252.52 GFLOPs	32.28	28.68	27.61	26.16
SAGN_ESA		√	√			4.18×10^6	266.83 GFLOPs	32.30	28.67	27.61	26.18
SAGN_MSA		√		√		4.32×10^6	284.58 GFLOPs	32.37	28.71	27.63	26.37
SAGN_AGM		√			√	4.33×10^6	293.99 GFLOPs	32.35	28.70	27.61	26.26
SAGN		√		√	√	4.48×10^6	326.05 GFLOPs	**32.45**	**28.73**	**27.68**	**26.40**

注: 性能最优作加粗处理。

2.2.2　不同模块的特征图可视化分析

为验证加入不同模块对深层重建特征信息的影响, 将经过第 1 个 SAGG 后输出的“Set5: bird”特征图进行可视化分析, 如图 5 所示。可以看出, “SAGN_ERB”特征图略显平滑, “SAGN_MSA”特征图能预测更多的边缘细节, 同时“SAGN_AGM”特征图更关注高频的纹理信息, 在 MSA 和 AGM 的作用下, SAGN 充分将有效的特征信息融合到残差块中, “SAGN”特征图显示更加清晰的边缘纹理。因此, 这些可视化特征图证明了本文方法能够更好地恢复图像清晰的边缘和纹理。

2.2.3　不同引导方式对模型性能的影响

为研究 AGM 引导方式对性能的影响, 在尺度因子为 4 下训练 500 轮, 对 4 种注意力引导方式做对比实验。图 6 为 SAGG 的 4 种注意力引导方式。表 3 为 4 种引导方式对模型性能的影响对比。其中, 浮点计算量是在输入图像块 96×96 后测量的。据表 3 观察, 引导方式 1 在注意力分支上去掉 1×1 卷积层后, 虽然参数量减少, 但性能也随之降低; 引导方式 2 仅在 Set5 上有较高的性能, 而在 Urban100 细节纹理较多的数据集上却表现最差; 引导方式 3 与本文方式仅差了相加操作, 性能上表现良好; 引导方式 4 在

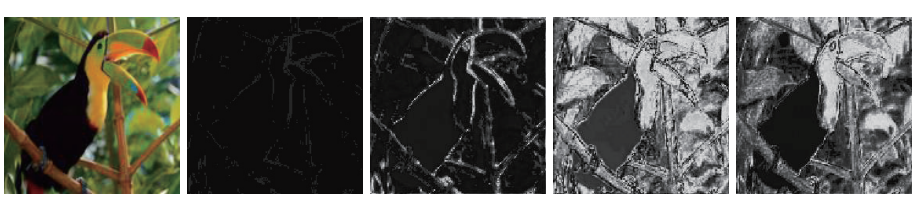


图 5　加入不同模块的特征图可视化

Fig. 5　Feature visualization after adding different modules

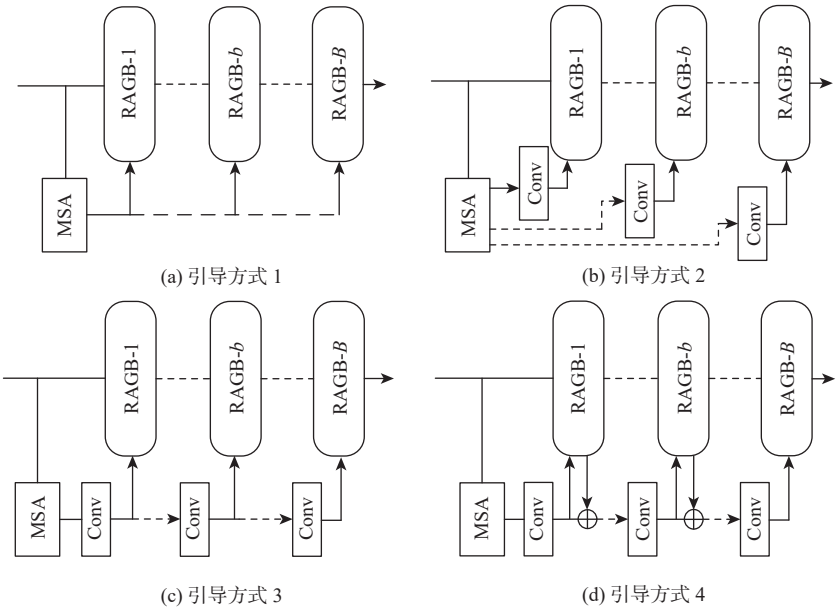


图 6　4 种注意力引导方式

Fig. 6　Four ways of attention guidance methods

表 3　4 种注意力引导方式对模型性能的影响

Table 3　Effect of four ways of attention guidance methods on model performance

引导方式	参数量	浮点计算量	PSNR/dB			
			Set5	Set14	B100	Urban100
引导方式1	4.23×10^6	278.98 GFLOPs	32.38	28.55	27.62	26.25
引导方式2	4.48×10^6	325.04 GFLOPs	**32.45**	28.60	27.63	26.23
引导方式3	4.48×10^6	326.04 GFLOPs	32.41	28.72	27.67	26.38
引导方式4	4.48×10^6	326.04 GFLOPs	**32.45**	**28.73**	**27.68**	**26.40**

注: 指标最优作加粗处理。

注意力分支特征输入给 AGM 后, 残差块输出特征又反馈给分支做相加操作, 对性能提升有一定帮助, 且在训练时前期收敛速度较其他方式快, 在不同数据集上也都能有较好的重建性能。综合计算成本与性能, 本文主要使用引导方式 4。

2.3　模型超参数分析

设置 3 组实验, 如图 7 所示, 进一步分析 SAGN 中残差组数量、每组 SAGG 中 RAGB 数量及 AGM 数量、卷积层通道数对网络性能的影响。从图 7(a) 可以看出, 在“SAGN_G2B8”与“SAGN_G4B4”网络深度相同情况下, 当 SAGG 数量减少、RAGB 的数量增加时, 模型的 PSNR 提高; 当“SAGN_G2B4”网络深度减小时, 参数量减少, 网络性能随之降低。其中, SAGN_G2B8 表示残差组数量为 2, RAGB 数量为 8。由图 7(b) 可知, 固定残差组数量为 2, RAGB 数量为 8, AGM 数量取 8 时, 性能达到最高, 相反, AGM 数量减少会影响网络的性能。由图 7(c) 可知, 通道数越大, 网络性能越好, 但通道数为 128 时, PSNR 提升不多反而会增加大量参数。由以上实验得到, 综合考虑计算成本和模型性能, 最终采用 2 个 SAGG, 每个 SAGG 中包含 8 个 RAGB 和 8 个 AGM, 且通道数为 96。

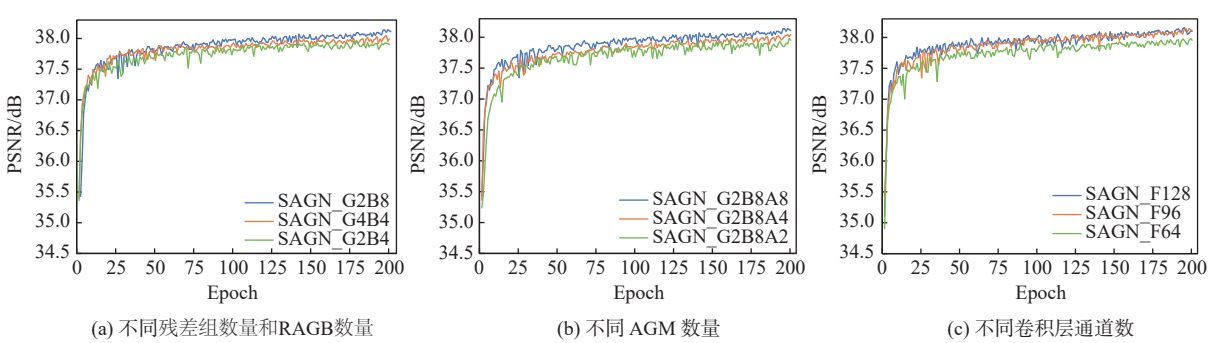


图 7　在 200 个 Epoch、Set5 数据集尺度因子为 2 下, 不同超参数对网络性能的影响

Fig. 7　Influence of different hyperparameters on network performance under 200 Epoch and scale factor of Set5 datasets is 2

2.4　模型性能比较

2.4.1　客观指标对比

为验证本文 SAGN 方法的有效性(见表 4[4-6,10-12,14-16,24-30]), 选择与 VDSR[6] 等经典方法、IMDN[10] 等轻量型方法、MSRN[11] 等同量级方法、AAN[15] 等注意力机制模型、SwinIR[16] 等基于 Trans-

表 4　尺度因子为 2、3、4 时基准数据集的客观评价指标对比

Table 4　Comparison of objective evaluation indexes with scale factors of 2, 3 and 4 for benchmark datasets

尺度因子	模型	PSNR/dB				SSIM			
		Set5	Set14	B100	Urban100	Set5	Set14	B100	Urban100
2	SRCNN[4]	36.66	32.45	31.36	29.50	0.954 2	0.906 7	0.887 9	0.894 6
	FSRCNN[5]	37.00	32.63	31.53	29.88	0.955 8	0.908 8	0.892 0	0.902 0
	VDSR[6]	37.53	33.03	31.90	30.76	0.958 7	0.912 4	0.896 0	0.914 0
	IMDN[10]	38.00	33.63	32.19	32.17	0.960 5	0.917 7	0.899 6	0.928 3
	MSRN[11]	38.08	33.74	32.23	32.22	0.960 5	0.917 0	0.901 3	0.932 6
	Cross-SRN[12]	38.03	33.62	32.19	32.28	0.960 6	0.918 0	0.899 7	0.929 0
	AAF-L[14]	38.09	33.78	32.23	32.46	0.960 7	0.919 2	0.900 2	0.931 3
	AAN[15]	38.06	33.75	32.22	32.43	0.960 8	0.919 4	0.900 2	0.931 1
	SwinIR-light[16]	38.14	33.86	32.31	**32.76**	0.961 1	0.920 6	0.901 2	0.934 0
	CARN[24]	37.76	33.52	32.09	31.92	0.959 0	0.916 6	0.897 8	0.925 6
	MRFN[25]	37.98	33.41	32.14	31.45	0.961 1	0.915 9	0.899 7	0.922 1
	LatticeNet[26]	38.15	33.78	32.25	32.43	0.961 0	0.919 3	0.900 5	0.930 2
	OISR-RK2[27]	38.12	33.80	32.26	32.48	0.960 9	0.919 3	0.900 6	0.931 7
	BSRN[28]	38.10	33.74	32.24	32.34	0.961 0	0.919 3	0.900 6	0.930 3
	NGSwin[29]	38.05	33.79	32.27	32.53	0.961 0	0.919 9	0.900 8	0.932 4
	VapSR[30]	38.08	33.77	32.27	32.45	0.961 2	0.919 5	0.901 1	0.931 6
	SAGN	**38.18**	**33.87**	**32.31**	**32.76**	**0.961 2**	**0.921 0**	**0.901 2**	**0.934 1**
3	SRCNN[4]	32.75	29.28	28.41	26.24	0.909 0	0.820 9	0.786 3	0.798 9
	FSRCNN[5]	33.16	29.43	28.53	26.43	0.914 0	0.824 2	0.791 0	0.808 0
	VDSR[6]	33.66	29.77	28.82	27.14	0.921 3	0.831 4	0.797 6	0.827 9
	IMDN[10]	34.36	30.32	29.09	28.17	0.927 0	0.841 7	0.804 6	0.851 9
	MSRN[11]	34.38	30.34	29.08	28.08	0.926 2	0.839 5	0.804 1	0.855 4
	Cross-SRN[12]	34.43	30.33	29.09	28.23	0.927 5	0.841	0.805 0	0.853 5
	AAF-L[14]	35.54	30.41	29.14	28.40	0.928 3	0.843 6	0.806 2	0.857 4
	AAN[15]	34.47	30.44	29.14	28.41	0.927 9	0.843 7	0.805 9	0.857 0
	SwinIR-light[16]	34.62	30.54	29.20	28.66	0.928 9	0.846 3	0.808 2	0.862 4
	CARN[24]	34.29	30.29	29.06	28.06	0.925 5	0.840 7	0.803 4	0.849 3
	MRFN[25]	34.21	30.03	28.99	27.53	0.926 7	0.836 3	0.802 9	0.858 9
	LatticeNet[26]	34.53	30.39	29.15	28.33	0.928 1	0.842 4	0.805 9	0.853 8
	OISR-RK2[27]	34.55	30.46	29.18	28.50	0.928 2	0.844 3	0.807 5	0.859 7
	BSRN[28]	34.46	30.48	29.18	28.39	0.927 7	0.844 9	0.806 8	0.856 7
	NGSwin[29]	34.52	30.53	29.19	28.52	0.928 2	0.845 6	0.807 8	0.860 3
	VapSR[30]	34.52	30.53	29.19	28.43	0.928 4	0.845 2	0.807 7	0.858 3
	SAGN	**34.63**	**30.55**	**29.23**	**28.67**	**0.929 0**	**0.846 5**	**0.808 2**	**0.862 5**
4	SRCNN[4]	30.48	27.49	26.90	24.52	0.862 8	0.750 3	0.710 1	0.722 1
	FSRCNN[5]	30.71	27.59	26.98	24.62	0.865 7	0.753 5	0.715 0	0.728 0
	VDSR[6]	31.35	28.01	27.29	25.18	0.883 8	0.767 4	0.725 1	0.752 4
	IMDN[10]	32.21	28.58	27.56	26.04	0.894 8	0.781 1	0.735 3	0.783 8
	MSRN[11]	32.07	28.60	27.52	26.04	0.890 3	0.775 1	0.727 3	0.789 6
	Cross-SRN[12]	32.24	28.59	27.58	26.16	0.895 4	0.781 7	0.736 4	0.788 1
	AAF-L[14]	32.32	28.67	27.62	26.32	0.896 4	0.783 9	0.737 9	0.793 1
	AAN[15]	32.30	28.71	27.61	26.27	0.896 6	0.784 2	0.737 4	0.792 0
	SwinIR-light[16]	32.44	28.77	27.69	26.47	0.897 6	0.785 8	0.740 6	0.798 0
	CARN[24]	32.13	28.60	27.58	26.07	0.893 7	0.780 6	0.734 9	0.783 7
	MRFN[25]	31.90	28.31	27.43	25.46	0.891 6	0.774 6	0.730 9	0.765 4
	LatticeNet[26]	32.30	28.68	27.62	26.25	0.896 2	0.783 0	0.736 7	0.787 3
	OISR-RK2[27]	32.32	28.72	27.66	26.37	0.896 5	0.784 3	0.739 0	0.795 3
	BSRN[28]	32.35	28.73	27.65	26.27	0.896 6	0.784 8	0.738 7	0.790 8
	NGSwin[29]	32.33	28.78	27.66	26.45	0.896 3	0.785 9	0.739 6	0.796 3
	VapSR[30]	32.38	28.77	27.68	26.35	0.897 8	0.785 2	0.739 8	0.794 1
	SAGN	**32.52**	**28.82**	**27.73**	**26.51**	**0.898 0**	**0.786 0**	**0.739 9**	**0.797 1**

注: 最优方法作加粗处理。

former 的模型及 VapSR[30] 等最先进的方法在 4 个基准数据集上进行 PSNR 和 SSIM[23] 性能比较。据表 4 数据显示，本文 SAGN 方法与目前的先进方法比较，在 4 个基准数据集上都表现出最优性能，其 4 倍重建结果的 PSNR 较 SwinIR-light[16] 平均提高了 0.05 dB。实验结果充分证明了本文 SAGN 方法在处理空间条纹信息能力上的突出表现，重建结果显示最优的性能。

2.4.2 主观视觉对比

图 8 为本文方法与 10 个有代表性的模型在尺度因子为 4 下基准数据集图像超分辨率重建结果的主观视觉图对比。如图 8 中的 Set14: Barbara 所示，MSRN 等方法恢复“书本”的纹理均是倾斜的，与 HR 图像不符，存在严重失真。虽然 NGSwin 等方法恢复的纹理方向是正确的，但存在纹理模糊和边缘不清晰的问题，相比较之下，本文 SAGN 方

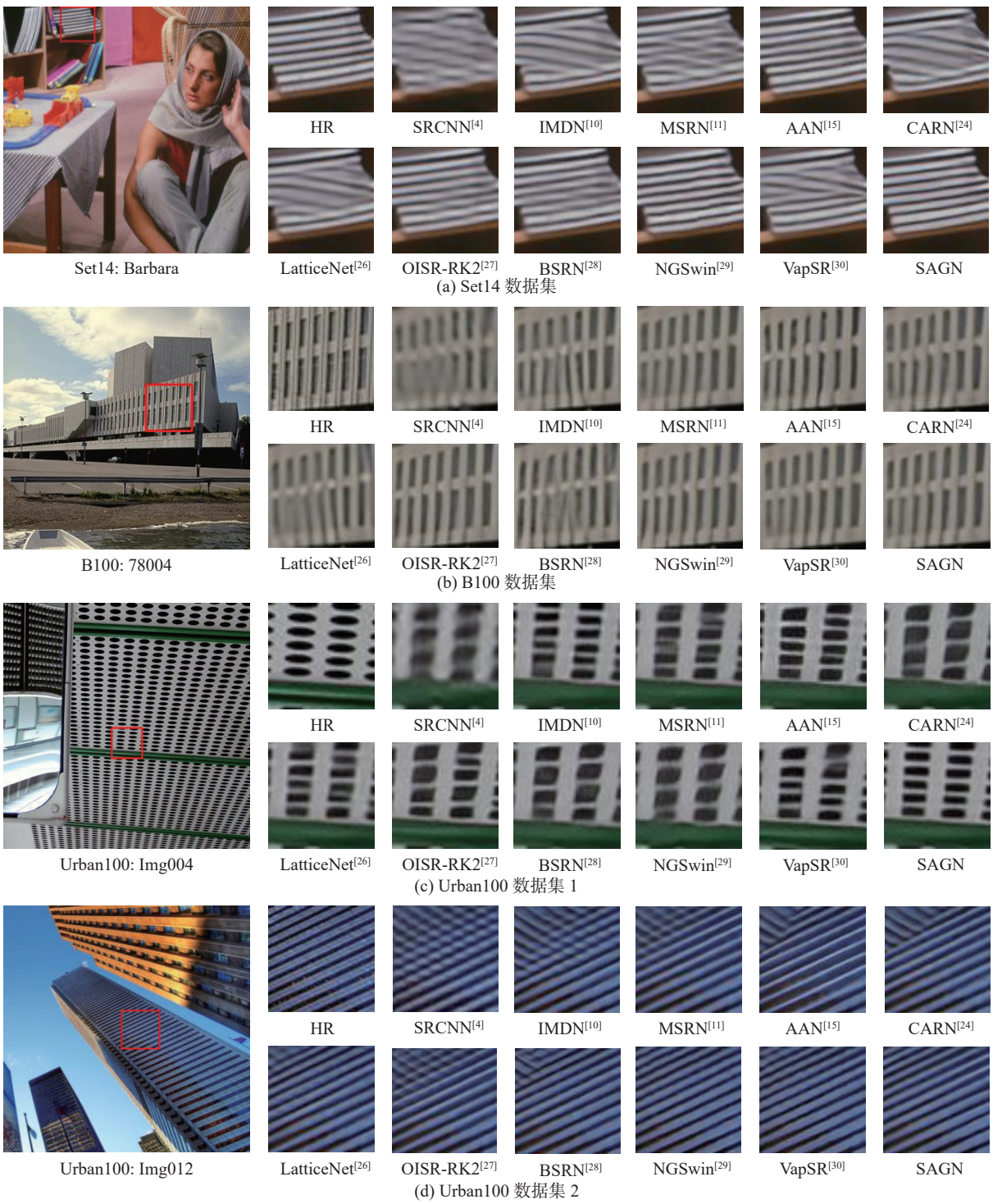


图 8 尺度因子为 4 时基准数据集图像的超分辨率重建结果视觉对比

Fig. 8 Visual comparison of SR reconstruction results with scale factor of 4 for benchmark datasets

法恢复“书本”的层次纹理更曲直; 在具有重复图案和排列规律的图像上, SAGN 的优势更为明显, 如图 8 中的 B100: 78004 所示, BSRN 等方法恢复的“窗户”发生了变形, 而 SAGN 方法恢复“窗户”的结构更接近矩形; 如图 8 中的 Urban100: Img004 所示, 大多数方法都无法重建正确的“椭圆形”结构, 甚至会产生伪影, 而本文 SAGN 方法虽然不能完全重建出原有形状, 但还可以有规律地排列; 如图 8 中的 Urban: Img012 所示, 清楚观察到“高楼大厦”的楼层越高, 视觉效果越密集, BSRN 等方法会生成一些方向错误的线条, 甚至发生了严重的扭曲, 而本文 SAGN 方法能够较好地恢复楼层的纹理信息。重建结果视觉对比图表明, 本文 SAGN 方法在恢复图像几何结构和细节方面具有一定优势, 证明了该方法的有效性。

2.4.3　模型复杂度对比

为分析 SAGN 的复杂度, 本节将基于 Transformer 的方法及轻量型方法等进行复杂度分析对比, 如表 5 所示[7,11,16,27-31]。在 OISR-RK2、MSRN、DBPN、SAGN 这 4 个同量级方法中, SAGN 在参数量最少的情况下, 取得较优的性能和较快的运行时间; 在 EDSR 参数量远超于本文方法的情况下, PSNR 仍高 0.07 dB; 虽然 NGSwin、BSRN 等轻量型方法的参数量较少且运行时间短, 但其性能指标远低于本文 SAGN 方法; 另外, 与最近流行的基于 Transformer 的超分辨率重建方法 SwinIR 相比, SAGN 在参数量减少 63.91% 情况下, 性能仅降低了 0.44%。综合参数量、性能和运行时间 3 项指标, 实验结果表明, 本文方法的复杂度适中。

表 5　尺度因子为 2 时 Set5 数据集上不同方法复杂度对比

Table 5　Complexity comparison of different methods on Set5 dataset under scale factor is 2

模型	参数量	PSNR/dB	SSIM	运行时间/ms
EDSR[7]	40.73×10^6	38.11	0.960 2	885
MSRN[11]	5.89×10^6	38.08	0.960 5	685
SwinIR[16]	11.5×10^6	**38.35**	**0.962 0**	889
SwinIR-light[16]	0.87×10^6	38.14	0.961 1	355
OISR-RK2[27]	4.97×10^6	38.12	0.960 9	558
BSRN[28]	0.32×10^6	38.10	0.961 0	**205**
DBPN[31]	5.95×10^6	38.09	0.960 0	775
NGSwin[29]	0.99×10^6	38.05	0.961 0	298
VapSR[30]	0.32×10^6	38.08	0.961 2	223
SAGN	4.15×10^6	38.18	0.961 2	425

注: 指标最优和次优分别作加粗和下划线处理。

3　结　论

1) 增强特征提取残差块可通过局部路径增加特征信息的流动性, 相比基础残差块, 在增加 0.59×10^6 参数量后, Set5 数据集的 PSNR 提升了 0.07 dB。

2) 多尺度空间注意力和注意力引导模块配合使用后, 本文 SAGN 有效权衡主干特征和注意力特征比例, 更关注图像的边缘纹理。

3) 本文 SAGN 可用于重建几何结构复杂和细节纹理丰富的图像, 虽模型重建时间有所增加, 但恢复图像效果明显提升。

为使本文 SAGN 能够适配可移动设备端, 仍需要优化模型参数量及重建时长, 使模型更加轻量化。

参考文献 (References)

[1] 陈嘉琪, 刘祥梅, 李宁, 等. 一种超分辨 SAR 图像水域分割算法及其应用[J]. 电子与信息学报, 2021, 43(3): 700-707.
CHEN J Q, LIU X M, LI N, et al. A high-precision water segmentation algorithm for SAR image and its application[J]. Journal of Electronics & Information Technology, 2021, 43(3): 700-707(in Chinese).

[2] XIAO Y, YUAN Q Q, JIANG K, et al. From degrade to upgrade: learning a self-supervised degradation guided adaptive network for blind remote sensing image super-resolution[J]. Information Fusion, 2023, 96: 297-311.

[3] GEORGESCU M I, IONESCU R T, MIRON A I, et al. Multimodal multi-head convolutional attention with various kernel sizes for medical image super-resolution[C]//Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. Piscataway: IEEE Press, 2023.

[4] DONG C, LOY C C, HE K M, et al. Image super-resolution using deep convolutional networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2016, 38(2): 295-307.

[5] DONG C, LOY C C, TANG X O. Accelerating the super-resolution convolutional neural network[C]//Proceedings of the European Conference on Computer Vision. Berlin: Springer, 2016: 391-407.

[6] KIM J, LEE J K, LEE K M. Accurate image super-resolution using very deep convolutional networks[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2016: 1646-1654.

[7] LIM B, SON S, KIM H, et al. Enhanced deep residual networks for single image super-resolution[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. Piscataway: IEEE Press, 2017: 1132-1140.

[8] HE K M, ZHANG X Y, REN S Q, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2016: 770-778.

[9] 程德强, 郭昕, 陈亮亮, 等. 多通道递归残差网络的图像超分辨率重建[J]. 中国图象图形学报, 2021, 26(3): 605-618.
CHENG D Q, GUO X, CHEN L L, et al. Image super-resolution reconstruction from multi-channel recursive residual network[J]. Journal of Image and Graphics, 2021, 26(3): 605-618(in Chinese).

[10] HUI Z, GAO X B, YANG Y C, et al. Lightweight image super-resolution with information multi-distillation network[C]//Proceedings of the 27th ACM International Conference on Multimedia.

New York: ACM, 2019: 2024-2032.

[11] LI J C, FANG F M, MEI K F, et al. Multi-scale residual network for image super-resolution[C]//Proceedings of the European Conference on Computer Vision. Berlin: Springer, 2018: 527-542.

[12] LIU Y Q, JIA Q, FAN X, et al. Cross-SRN: structure-preserving super-resolution network with cross convolution[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2022, 32(8): 4927-4939.

[13] LIU J, ZHANG W J, TANG Y T, et al. Residual feature aggregation network for image super-resolution[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2020: 2356-2365.

[14] WANG X H, WANG Q, ZHAO Y Z, et al. Lightweight single image super-resolution network with attentive auxiliary feature learning[C]//Proceedings of the Asian Conference on Computer Vision. Berlin: Springer, 2021: 268-285.

[15] CHEN H, GU J, ZHANG Z, et al. Attention in attention network for image super-resolution[C]//Proceedings of the Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2021.

[16] LIANG J Y, CAO J Z, SUN G L, et al. SwinIR: image restoration using swin Transformer[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops. Piscataway: IEEE Press, 2021: 1833-1844.

[17] ZHAO H Y, KONG X T, HE J W, et al. Efficient image super-resolution using pixel attention[C]//Proceedings of the European Conference on Computer Vision. Berlin: Springer, 2020: 56-72.

[18] SHI W Z, CABALLERO J, HUSZÁR F, et al. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2016: 1874-1883.

[19] BEVILACQUA M, ROUMY A, GUILLEMOT C, et al. Low-complexity single-image super- resolution based on nonnegative neighbor embedding[C]//Proceedings of the British Machine Vision Conference. Surrey: BMVA Press, 2012: 135.1-135.10.

[20] ZEYDE R, ELAD M, PROTTER M. On single image scale-up using sparse-representations[C]//Proceedings of the International Conference on Curves and Surfaces. Berlin: Springer, 2012: 711-730.

[21] MARTIN D, FOWLKES C, TAL D, et al. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics[C]//Proceedings of the 8th IEEE International Conference on Computer Vision. Piscataway: IEEE Press, 2001: 416-423.

[22] HUANG J B, SINGH A, AHUJA N. Single image super-resolution from transformed self-exemplars[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2015: 5197-5206.

[23] WANG Z, BOVIK A C, SHEIKH H R, et al. Image quality assessment: from error visibility to structural similarity[J]. IEEE Transactions on Image Processing, 2004, 13(4): 600-612.

[24] AHN N, KANG B, SOHN K A. Fast, accurate, and lightweight super-resolution with cascading residual network[C]//Proceedings of the European Conference on Computer Vision. Berlin: Springer, 2018: 256-272.

[25] HE Z W, CAO Y P, DU L, et al. MRFN: multi-receptive-field network for fast and accurate single image super-resolution[J]. IEEE Transactions on Multimedia, 2020, 22(4): 1042-1054.

[26] LUO X T, XIE Y, ZHANG Y L, et al. LatticeNet: towards lightweight image super-resolution with lattice block[C]//Proceedings of the European Conference on Computer Vision. Berlin: Springer, 2020: 272-289.

[27] HE X Y, MO Z T, WANG P S, et al. ODE-inspired network design for single image super-resolution[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2019: 1732-1741.

[28] LI Z Y, LIU Y Q, CHEN X Y, et al. Blueprint separable residual network for efficient image super-resolution[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. Piscataway: IEEE Press, 2022: 832-842.

[29] CHOI H, LEE J, YANG J. N-gram in swin Transformers for efficient lightweight image super-resolution[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2023: 2071-2081.

[30] ZHOU L, CAI H M, GU J J, et al. Efficient image super-resolution using vast-receptive-field attention[C]//Proceedings of the European Conference on Computer Vision Workshops. Berlin: Springer, 2023: 256-272.

[31] HARIS M, SHAKHNAROVICH G, UKITA N. Deep back-projection networks for super-resolution[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2018: 1664-1673.

Image super-resolution reconstruction network based on multi-scale spatial attention guidance

CHENG Deqiang[1]，WANG Peijie[1]，DONG Yanqiang[2]，KOU Qiqi[3]，JIANG He[1,*]

(1. School of Information and Control Engineering，China University of Mining and Technology，Xuzhou 221116，China；
2. Open-pit Mine of Inner Mongolia Baiyinhua Coal and Power Company Limited，State Power Investment Corporation，Xilingol League 026000，China；
3. School of Computer Science and Technology，China University of Mining and Technology，Xuzhou 221116，China)

Abstract: Aiming at the problem that the attention-mechanism-based image super-resolution reconstruction network ignores the heterogeneity of attentional features and treats features at different levels uniformly by directly incorporating the attention mechanism into the network model. This study designs a novel multi-scale spatial attention guidance network, namely SAGN, which makes the following key contributions. Firstly, an enhanced feature extraction residual block (ERB) is proposed to enhance the representation capacity of local information. Secondly, to record spatial attention features at various scales, a multi-scale spatial attention (MSA) module is incorporated. Lastly, an attention-guided module (AGM) is introduced to assign individualized weights to different features, facilitating effective fusion of contextual global features and suppression of redundant information. On four benchmark datasets, extensive experimental results show that SAGN outperforms standard attention structures in terms of both subjective visual perception and objective evaluation criteria. Notably, SAGN achieves an average 0.05 dB higher than that of the suboptimal model in peak signal-to-noise ratio (PSNR) for 4 times reconstruction results, further underscoring its efficacy in recovering image geometric structures and fine details.

Keywords: super-resolution reconstruction; convolutional neural network; attention mechanism; multi-scale spatial attention; attention guidance

Received：2023-08-28；**Accepted**：2023-10-29；**Published Online**：2023-11-15 16：02

URL：link.cnki.net/urlid/11.2625.V.20231115.1126.001

Foundation items：National Natural Science Foundation of China (52204177,52304182); The Fundamental Research Funds for the Central Universities (2020QN49)

* **Corresponding author**. E-mail：jianghe@cumt.edu.cn

http://bhxb.buaa.edu.cn jbuaa@buaa.edu.cn
DOI: 10.13700/j.bh.1001-5965.2023.0383

基于模糊逻辑与自适应策略的红外可见光图像融合

杨勇[1,2]，刘家祥[3]，黄淑英[4,*]，王晓争[5]，夏钰锟[3]

(1. 天津工业大学 计算机科学与技术学院，天津 300387；2. 沧州市天津工业大学研究院，沧州 061000；
3. 江西财经大学 信息管理与数学学院，南昌 330032；4. 天津工业大学 软件学院，天津 300387；
5. 天津工业大学 控制科学与工程学院，天津 300387)

摘 要：由于成像机制不同，红外图像能捕捉目标信息，可见光图像提供纹理细节，需融合两者以提升视觉感知与机器识别效果。基于模糊逻辑理论，提出一种多级模糊逻辑判别与自适应参数融合策略 (MFD-APFS) 的红外与可见光图像融合方法。将红外图像与可见光图像分别进行结构块分解，得到由信号强度分量重构的对比度细节图像组；将源图像组与对比度细节图像组分别输入设计的模糊逻辑判别系统，对图像组进行模糊逻辑判别得到各自的显著性图像，并对得到的显著性图像组进行二次模糊逻辑判别，得到联合的显著性图像；利用引导滤波技术，将显著性图像引导源图像，得到多幅决策图，通过自适应参数的融合策略，得到最终的融合图像。将 MFD-APFS 方法在红外和可见光图像公开数据集上进行实验测试，结果表明，相比 7 种主流的融合方法，对于客观度量指标 SSIM-F 和 $Q^{AB/F}$，在 TNO 数据集上分别提升了 0.169 和 0.140 3，在 RoadScenes 数据集上分别提升了 0.175 3 和 0.053 7；主观视觉效果表明，所提方法可以生成目标清晰、细节丰富的融合图像，较好地保留了红外图像目标信息及可见光图像纹理信息。

关 键 词：红外与可见光图像融合；自适应参数融合策略；多级模糊逻辑；引导滤波；决策图
中图分类号：TP391.41
文献标志码：A **文章编号**：1001-5965(2025)07-2196-13

单传感器获取的图像信息通常是不完整、不精确的，其他传感器可以提供补充的图像信息，图像融合技术能够融合多传感器源的图像信息，得到更完整、精确的图像。图像融合技术[1]是一种可保留源图像组中互补信息并能提取源图像中的有利信息，进而生成利于视觉感知和机器处理的图像的处理技术。红外图像与可见光图像融合是图像融合技术领域的一个重要分支。红外图像反映场景的温度或辐射差异信息，不易受到复杂的自然条件影响。然而，红外图像细节信息不明显，成像性能差，可见性差。与红外图像不同，可见光图像仅与目标场景的反射有关，可捕获丰富的细节信息。因此，红外与可见光图像融合能够使同一场景的信息更加丰富，从而提高复杂场景下图像的可识别性。

近年来，红外与可见光图像融合技术发展迅速。按照融合的基础理论，图像融合技术可以分为基于多尺度变换的方法[2-7]、基于深度学习的方法[8-16]、基于稀疏表示的方法[17-19]和基于显著性的方法[20]等。基于多尺度变换的方法通常利用多尺度变换工具将源图像分解成不同尺度的分量系数，根据融合规则对多尺度分量系数进行融合，利用多尺度逆变换得到融合后的图像。常见的多尺度变换工具有离散小波变换(discrete wavelet transform,

收稿日期：2023-06-16；**录用日期**：2023-11-24；**网络出版时间**：2025-04-14 17:58
网络出版地址：link.cnki.net/urlid/11.2625.V.20250414.1647.003
基金项目：国家自然科学基金 (62072218,61862030)；天津市自然科学基金 (24JCZDJC00130)；沧州市天津工业大学研究院项目 (TGCYY-Z-0303)
* **通信作者**. E-mail：shuyinghuang2010@126.com

引用格式：杨勇，刘家祥，黄淑英，等. 基于模糊逻辑与自适应策略的红外可见光图像融合 [J]. 北京航空航天大学学报，2025，51（7）：2196-2208. YANG Y，LIU J X，HUANG S Y，et al. Fuzzy logic and adaptive strategy for infrared and visible light image fusion [J]. Journal of Beijing University of Aeronautics and Astronautics，2025，51（7）：2196-2208（in Chinese）.

DWT)[5]、轮廓波变换(contourlet transform, CT)[6]、非下采样剪切波变换(nonsubsampled shearlet transform, NSST)[7] 等, 但是这些变换方法的计算复杂度较高。

随着深度学习的不断发展, 研究人员提出基于深度学习的红外与可见光图像的融合方法。Xu 等[8]提出了基于卷积神经网络(convolutional neural network, CNN)的并行场景和纹理学习的图像融合方法, 通过关注粗细粒度的图像信息提升融合效果, 但该网络对处理极端过曝条件的图像能力较差, 容易丢失过曝区域的场景信息。Li 等[9]提出了一种基于 CNN 的网络 DenseFuse, 引入稠密块来提取特征, DenseFuse 通过编码和解码网络生成融合图像, 但该方法训练时间长, 由于训练过程没有针对红外和可见光图像, 导致提取的特征不够全面。Tang 等[10]提出了一种基于照明感知的渐进式图像融合网络(progressive illumination-aware fusion, PIAFusion), 该网络自适应地保持显著目标的强度分布, 并保留背景中的纹理信息, 但在某些极端环境下融合质量降低。为在极端环境下生成具有突出目标和丰富纹理的高质量融合图像, Tang 等[11]提出了一种暗光下的红外与可见光图像融合方法(darkness-free infrared and visible image fusion, DIVFusion), 促进亮度信息互补聚合。Ma 等[12]提出了一种用于图像融合的生成对抗网络(generative adversarial network-based model for image fusion, FusionGAN), 网络中, 生成器生成融合图像, 鉴别器鉴别融合图像是否包含可见光图像的信息, 但由于 FusionGAN 过于依赖对抗学习来学习图像特征, 导致其相对不稳定, 可见光图像的纹理信息容易丢失。Yang 等[13]提出了基于纹理条件生成对抗网络(texture conditional generative adversarial network, TC-GAN), 用于捕获图像融合中的梯度变化, 并设计了一种基于多决策图的融合策略, 从源图像中获取重要的纹理信息。然而, 大多数基于生成对抗网络(generative adversarial network, GAN)的方法将 2 个源图像串联作为输入, 因此, 提取的特征不足以保留背景和细节信息[14]。Yang 等[14]提出了一种新的基于 GAN 的融合框架 DSG-Fusion, 鉴于红外和可见光图像的模态不同, DSG 网络的生成器通过 2 个独立的数据流提取 2 个图像的特征。Xu 等[15]提出了一种统一的无监督图像融合网络 (unified unsupervised image fusion network, U2Fusion), 通过特征提取和信息度量来保持信息, 但 U2Fusion 的融合策略未充分考虑红外和可见光图像的本质特征。上述基于深度学习的图像融合方法大多需要较长的训练时间和大量参考的训练集, 对于红外和可见光图像往往比较困难。

综合上述分析, 基于显著性特征提取的思想, 本文提出一种多级模糊逻辑判别与自适应参数融合策略(multistage fuzzy discrimination and adaptive parameter fusion strategy, MFD-APFS)的红外与可见光图像融合方法。首先, 将红外图像与可见光图像分别进行结构块分解(structural patch decomposition, SPD), 得到由信号强度分量重构组成的对比度细节图像组; 然后, 将红外和可见光图像组与对比度细节图像组分别进行模糊逻辑判别, 得到显著性图像, 并对显著性图像进行第 2 次模糊逻辑判别, 得到联合的显著性图像; 最后, 利用自适应参数引导滤波(adaptive guided filter, AGF)技术[13], 采用显著性图像引导源图像的方式, 得到多幅决策图[21], 再通过自适应参数的融合规则(adaptive image fusion rule, AIFR)得到最终的融合图像。本文提出的红外与可见光图像融合主要贡献如下:

1) 采用结构块分解方法分解红外与可见光图像, 通过提取其信号强度可有效提取源图像中的对比度信息。

2) 设计多级模糊逻辑判别系统, 可有效提取图像中的显著性特征信息。

3) 提出自适应参数融合策略, 融合结果可得到红外图像的完整目标, 同时避免重构过程中可能带来的额外噪声。

1　相关工作

1.1　结构块分解

Ma 等[22]提出一种简单有效的多曝光图像融合技术, 其对伪影具有很强的鲁棒性。该方法将图像进行结构块分解, 得到 3 个分量, 分别为信号强度、信号结构、平均强度, 如图 1 所示。结构块分解是一种简单有效的基于显著性特征的融合方法, 其基于著名的图像质量评估指标, 即结构相似性(structural similarity, SSIM)[23]而提出。在结构块分

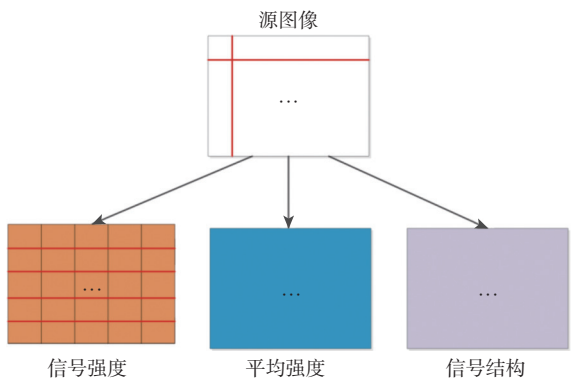


图 1　结构块分解图示

Fig. 1　Structural patch decomposition

解中，信号强度对应SSIM中的对比度测度，因此，可以使用信号强度分量提取对比度特征，将这些不同的特征用于确定每个输入图像的融合区域。通过结构块分解和重建，可以从源图像中获得信号强度。结构块分解可表示为

$$\boldsymbol{x}_k=\|\boldsymbol{x}_k-\boldsymbol{\mu}_{\boldsymbol{x}_k}\|\frac{\boldsymbol{x}_k-\boldsymbol{\mu}_{\boldsymbol{x}_k}}{\|\boldsymbol{x}_k-\boldsymbol{\mu}_{\boldsymbol{x}_k}\|}+\boldsymbol{\mu}_{\boldsymbol{x}_k}=\|\tilde{\boldsymbol{x}}_k\|\frac{\tilde{\boldsymbol{x}}_k}{\|\tilde{\boldsymbol{x}}_k\|}+\boldsymbol{\mu}_{\boldsymbol{x}_k}=\boldsymbol{c}_k\boldsymbol{s}_k+\boldsymbol{l}_k \tag{1}$$

式中：$\boldsymbol{x}_k$为图像中第k个图像块；$\|\cdot\|$表示L_2范数；$\boldsymbol{\mu}_{\boldsymbol{x}_k}$为图像块的像素均值；$\tilde{\boldsymbol{x}}_k$为去均值后的图像块；$\boldsymbol{c}_k$、$\boldsymbol{s}_k$、$\boldsymbol{l}_k$分别为图像的信号强度、平均强度、信号结构。

1.2 模糊逻辑判别图像融合

模糊逻辑[24]建立在布尔逻辑之上，运用模糊集来解决建模不确定导致的问题。因此，当存在不确定性且没有明显的数学关系可用时，通常使用模糊方法[25]。使用模糊逻辑判别系统的优点是：可以满足有关图像融合的各种输入和输出数据条件，且允许用户通过使用语言标签作为规则直接定义编码[26]。基于模糊逻辑判别的图像融合方法通常比传统的图像融合方法表现出更好的性能。文献[27]提出一种使用直觉模糊集的简单通用融合方法，可以获得具有高对比度的融合图像；文献[28]提出一种基于模糊变换的多模态医学图像融合方法。Yang等[24]通过对非下采样轮廓波变换(non-subsampled contourlet transform, NSCT)和type-2模糊逻辑技术进行多尺度几何分析，设计了一种新颖的多模态医学图像融合方法。与流行的融合方法相比，这些方法丰富了融合图像的细节，使其与源图像的差异缩小，并且提升了视觉效果。

2 本文方法

本文提出的MFD-APFS红外与可见光图像融合方法的过程如图2所示，步骤如下：

步骤1 将红外图像与可见光图像分别进行结构块分解，得到由信号强度分量重构而成的对比度细节图像组。

步骤2 将红外、可见光图像组与对比度细节图像组分别进行模糊逻辑判别1，得到各自的显著性图像$\boldsymbol{S}_1$、$\boldsymbol{S}_2$，将显著性图像$\boldsymbol{S}_1$、$\boldsymbol{S}_2$再次进行模糊逻辑判别2，得到联合的显著性图像组。

步骤3 采用AGF技术，利用显著性图像引导源图像，得到多幅决策图，通过设计的AIFR模块，得到最终的融合图像。

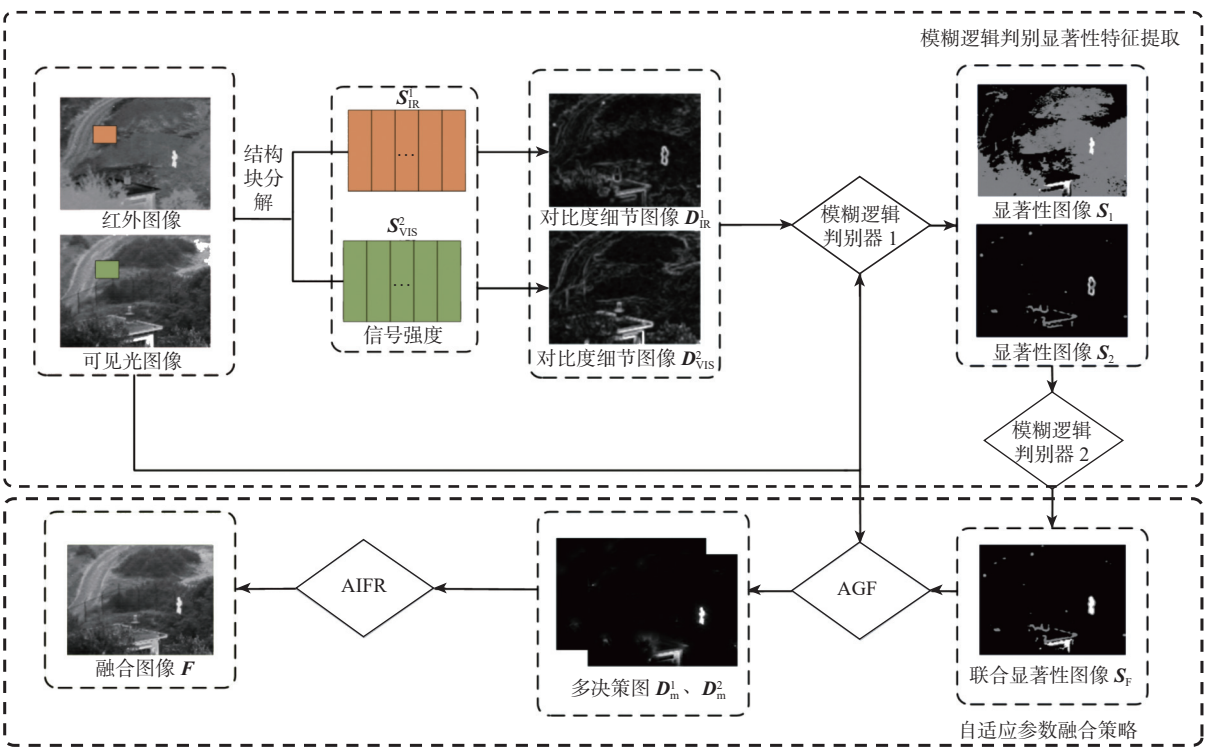


图2 多级模糊逻辑判别与自适应参数融合策略框架

Fig. 2 Framework based on multistage fuzzy logic discrimination and adaptive parameter fusion strategy

2.1 基于结构块分解的对比度细节提取

为更利于计算机视觉中人眼视觉感知系统(human visual system, HVS)，使用基于SSIM提出的结构块分解方法对红外与可见光图像进行细节提取。在红外与可见光图像融合中，希望能够在源图像中提取准确的红外目标信息，故需要对比度细节

信息来区分图像中的目标与背景。本文使用结构块分解技术分解红外与可见光图像,得到图像的信号强度,对其重构可获得源图像的对比度细节图像。根据红外信息的特性,对比度细节信息中包含红外图像目标的细节信息,故信号强度越大,信息越详细,同时,这些信息又能代表红外图像的显著目标信息。

2.2　基于多级模糊逻辑判别的显著性图像重构

最常用的模糊逻辑判别系统称为 Mamdani 和 Sugeno[25], 两者之间的主要区别在于输出类型。Mamdani 输出为常数,而 Sugeno 模型提供多项式输出。由于本文期望所提出融合方法的输出权重为常数,使用 Mamdani 型模糊逻辑判别系统获取显著性图像。

在信号强度的对比度细节图像中,信号强度 $\boldsymbol{D}_{\mathrm{IR}}^{1}$ 和 $\boldsymbol{D}_{\mathrm{VIS}}^{2}$ 中的每个像素都可作为模糊逻辑判别器的输入。模糊逻辑判别器将确定的值转换为由隶属度函数(membership function, MF)定义的模糊集。在模糊推理引擎块[25]中,使用预定义规则和模糊运算符将输入模糊集映射到输出模糊集权重。进行去模糊化,从聚合的模糊集中对输入图像的特定像素计算单个权重值。模糊逻辑判别步骤如图 3 所示。对于隶属度函数,将信号强度图及源图像的隶属度函数集合分别表示为 MF-ss 与 MF-so。由于源图像和信号强度图的灰度阈为 [0,255], 选择灰色映射[29]来评估隶属度参数的近似范围,并将模糊集的输入分为 3 个区域:低显著区域、中显著区域和高显著区域。选择三角模糊度[30]作为隶属度函数。输入隶属度函数 $\mu_{\mathrm{in_1}}$ 可以表示为

$$\mu_{\mathrm{in_1}}=\mu_{\mathrm{in_1}}^{\mathrm{L}}(x)*\mu_{\mathrm{in_1}}^{\mathrm{M}}(x)*\mu_{\mathrm{in_1}}^{\mathrm{H}}(x) \tag{2}$$

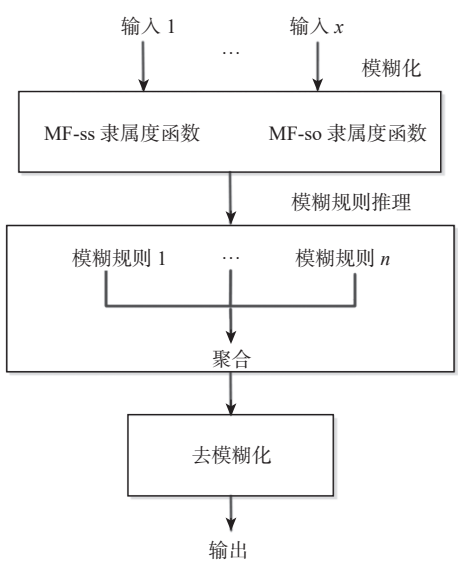


图 3　模糊逻辑判别结构

Fig. 3　Fuzzy logic discrimination structure

式中:“$*$”表示模糊交集运算;$\mu_{\mathrm{in_1}}^{\mathrm{L}}(x)$、$\mu_{\mathrm{in_1}}^{\mathrm{M}}(x)$、$\mu_{\mathrm{in_1}}^{\mathrm{H}}(x)$ 分别为低显著区域、中显著区域、高显著区域的三角隶属度函数,表示为

$$\mu_{\mathrm{in_1}}^{k}(x)=\begin{cases}\dfrac{1}{m-x}x-\dfrac{l}{m-l} & x\in[l,m)\\ \dfrac{1}{m-u}x-\dfrac{u}{m-u} & x\in[m,u]\\ 0 & \text{其他}\end{cases} \tag{3}$$

其中:参数 m、l、u 用于控制 $\mu_{\mathrm{in_1}}^{\mathrm{L}}(x)$、$\mu_{\mathrm{in_1}}^{\mathrm{H}}(x)$、$\mu_{\mathrm{in_1}}^{\mathrm{M}}(x)$ 的形状。

输出隶属度函数需要获得显著区域的最大值和过滤区域的最小值。由于高斯隶属度函数的平滑度具有简洁表示形式的非零值及间隔之间的连续过渡等优点,使用高斯隶属度函数计算输出权重,并使用梯形隶属度函数作为辅助。输出隶属度函数定义为

$$\mu_{\mathrm{out_1}}=\mu_{\mathrm{out_1}}^{\mathrm{L}}(x)*\mu_{\mathrm{out_1}}^{\mathrm{M}}(x)*\mu_{\mathrm{out_1}}^{\mathrm{H}}(x) \tag{4}$$

式中:$\mu_{\mathrm{out_1}}^{\mathrm{L}}(x)$ 和 $\mu_{\mathrm{out_1}}^{\mathrm{H}}(x)$ 为梯形隶属度函数(见图 4),定义为

$$\mu_{\mathrm{out_1}}^{\mathrm{L,H}}(x)=\begin{cases}1-(b-x)/a & x\in[b-a,b)\\ 1 & x\in[b,c)\\ 1-(x-c)/d & x\in[c,c+d]\\ 0 & \text{其他}\end{cases} \tag{5}$$

其中:$b-a=1$, $d-c=1$。

$\mu_{\mathrm{out_1}}^{\mathrm{L}}(x)$ 和 $\mu_{\mathrm{out_1}}^{\mathrm{H}}(x)$ 关于中心轴对称,$\mu_{\mathrm{out_1}}^{\mathrm{M}}(x)$ 为高斯子集,定义为

$$\mu_{\mathrm{out_1}}^{\mathrm{M}}=\mathrm{e}^{\frac{-(x-\bar{x})^2}{2\sigma^2}} \tag{6}$$

式中:σ 和 $\bar{x}$ 分别为方差和均值,取值为 0.15 和 0.5。

对于隶属度函数的输入和输出设计,本文参考文献 [31] 使用梯形和高斯函数作为隶属度函数的输入和输出,并根据红外与可见光图像和多聚焦图像之间的特性不同进行调整。因此,模糊逻辑判别器将数值(先验值)限制在输入变量的讨论范围之内,并为模糊语言集提供相应的隶属度。

输入变量通过输入隶属度计算得到相应的隶属度后,需使用模糊规则进行判别。本文使用 Mamdani 模糊蕴涵最小运算规则。当 2 个输入满足某一条件,输出一个值。例如,如果 MF-ss 为 $\mu_{\mathrm{in_1}}^{\mathrm{H}}$(见表 1),且 MF-so 为 $\mu_{\mathrm{in_1}}^{\mathrm{H}}$(见表 2),则输出权重为 $\mu_{\mathrm{out_1}}^{\mathrm{H}}$,本文模糊规则表示为

$$I=\mu_{\mathrm{in_1}}^{\mathrm{H}}(\text{MF-ss})\mu_{\mathrm{in_1}}^{\mathrm{H}}(\text{MF-so})\rightarrow\mu_{\mathrm{out_1}}^{\mathrm{H}} \tag{7}$$

式(7)表示为:如果 MF-ss 和 MF-so 都大于某一阈值,权重取最大值。MF-ss 和 MF-so 为 2 个输入,权重为输出。定义的模糊规则如表 1 和表 2 所

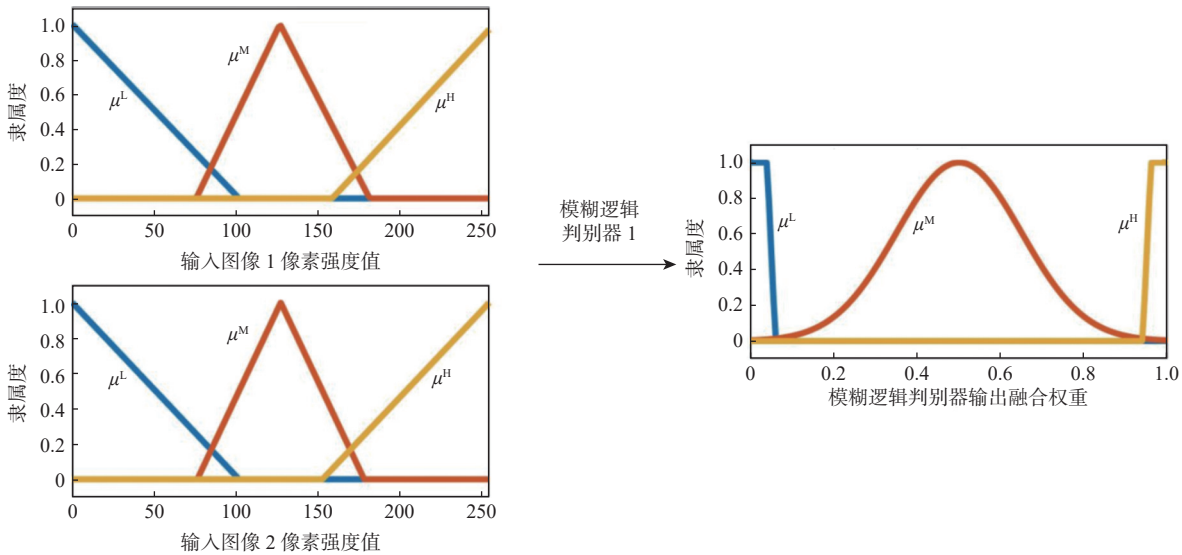


图 4 模糊逻辑判别器 1 隶属度函数

Fig. 4 Membership function of fuzzy logic discriminator 1

表 1 模糊逻辑判别器 1 的 MF-ss 模糊规则

Table 1 MF-ss fuzzy rules for fuzzy logic discriminator 1

MF-sf	MF-ss		
	$\mu^{\mathrm{L}}_{\mathrm{in_1}}$	$\mu^{\mathrm{M}}_{\mathrm{in_1}}$	$\mu^{\mathrm{H}}_{\mathrm{in_1}}$
$\mu^{\mathrm{L}}_{\mathrm{in_1}}$	μ_{w1}	μ_{w1}	μ_{w1}
$\mu^{\mathrm{M}}_{\mathrm{in_1}}$	μ_{w2}	μ_{w2}	μ_{w1}
$\mu^{\mathrm{H}}_{\mathrm{in_1}}$	μ_{w3}	μ_{w3}	μ_{w3}

表 2 模糊逻辑判别器 1 的 MF-so 模糊规则

Table 2 MF-so fuzzy rules for fuzzy logic discriminator 1

MF-sf	MF-so		
	$\mu^{\mathrm{L}}_{\mathrm{in_1}}$	$\mu^{\mathrm{M}}_{\mathrm{in_1}}$	$\mu^{\mathrm{H}}_{\mathrm{in_1}}$
$\mu^{\mathrm{L}}_{\mathrm{in_1}}$	μ_{w1}	μ_{w1}	μ_{w1}
$\mu^{\mathrm{M}}_{\mathrm{in_1}}$	μ_{w2}	μ_{w2}	μ_{w1}
$\mu^{\mathrm{H}}_{\mathrm{in_1}}$	μ_{w3}	μ_{w3}	μ_{w3}

示。本文中,MF-ss 和 MF-so 使用相同的隶属度函数。

使用模糊规则确定每个输入的结果,各个输出聚合得到整体的判别结果,本文使用最大值函数进行输出:

$$\boldsymbol{\mu}_{\mathrm{A}}(x)=\max\{\boldsymbol{\mu}_1(x),\boldsymbol{\mu}_2(x),\cdots,\boldsymbol{\mu}_9(x)\} \tag{8}$$

式中:$\boldsymbol{\mu}_{\mathrm{A}}(x)$为聚合曲线;$\boldsymbol{\mu}_j(x)$为第 j 个模糊规则的输出。

通过模糊规则的建立,从模糊逻辑判别器中输出模糊集,去模糊化,得到权重图。去模糊化的任务是指定最佳模糊集 $\boldsymbol{\mu}(z)$ 输出域的一个点,本文使用重心去模糊指定一个精确点,显著性图像 $\boldsymbol{M}_{\mathrm{SR}}$ 的定义如下:

$$\boldsymbol{M}_{\mathrm{SR}}(x,y)=\frac{\int z\boldsymbol{\mu}_i(z)\mathrm{d}z}{\int\boldsymbol{\mu}_i(z)\mathrm{d}z} \tag{9}$$

式中:i=1, 2, ⋯,n, n 为输入图像数量;$\boldsymbol{M}_{\mathrm{SR}}$为通过模糊逻辑判别得到的显著性图像,$(x,y)$为像素位置。

将信号强度重构的对比度细节图像组及源图像组分别经过模糊逻辑判别器 1 得到显著性图像 $\boldsymbol{S}_1$ 和显著性图像 $\boldsymbol{S}_2$,将两者经过模糊逻辑判别器 2,能够得到最终显著性图像。与模糊逻辑判别器 1 类似,模糊逻辑判别器 2 也是利用 Mamdani 型的模糊推理系统构建。

在模糊逻辑判别器 2 中,输入隶属度由梯形隶属度函数构成,为捕获隶属度函数曲线下最大面积,梯形的顶部被设计为在有效范围内的最小长度。同时,为使梯形函数与高斯隶属度函数具有较小重叠,故选择一个较小方差的高斯函数作为输入的隶属度函数。输入隶属度函数 MF-sp 可以表示为

$$\mu_{\mathrm{in_2}}=\mu^{\mathrm{L}}_{\mathrm{in_2}}(x)*\mu^{\mathrm{M}}_{\mathrm{in_2}}(x)*\mu^{\mathrm{H}}_{\mathrm{in_2}}(x) \tag{10}$$

式中:$\mu^{\mathrm{L}}_{\mathrm{in_2}}(x)$和$\mu^{\mathrm{H}}_{\mathrm{in_2}}(x)$为梯形隶属度函数,$\mu^{\mathrm{M}}_{\mathrm{in_2}}(x)$为式(6)计算得到的高斯函数。

输出隶属度函数定义为

$$\mu_{\mathrm{out_2}}=\mu^{\mathrm{L}}_{\mathrm{out_2}}(x)*\mu^{\mathrm{M}}_{\mathrm{out_2}}(x)*\mu^{\mathrm{H}}_{\mathrm{out_2}}(x) \tag{11}$$

式中:$\mu^{\mathrm{L}}_{\mathrm{out_2}}(x)$和$\mu^{\mathrm{H}}_{\mathrm{out_2}}(x)$为梯形隶属度函数;$\mu^{\mathrm{M}}_{\mathrm{out_2}}(x)$为高斯子集(见图 5)。

模糊逻辑判别器 2 的聚合规则与模糊逻辑判别器 1 一致,定义的模糊规则如表 3 所示。

去模糊化的任务是指定最佳模糊集 $\boldsymbol{\mu}(z)$ 输出域的一个点[32],在本文中,联合显著性图像$\boldsymbol{S}_{\mathrm{F}}$定义如下:

$$\boldsymbol{S}_{\mathrm{F}}(x,y)=\frac{\int z\boldsymbol{\mu}_i(z)\mathrm{d}z}{\int\boldsymbol{\mu}_i(z)\mathrm{d}z} \tag{12}$$

式中:$\boldsymbol{S}_{\mathrm{F}}$为通过模糊逻辑判别 2 得到的联合显著性图像。

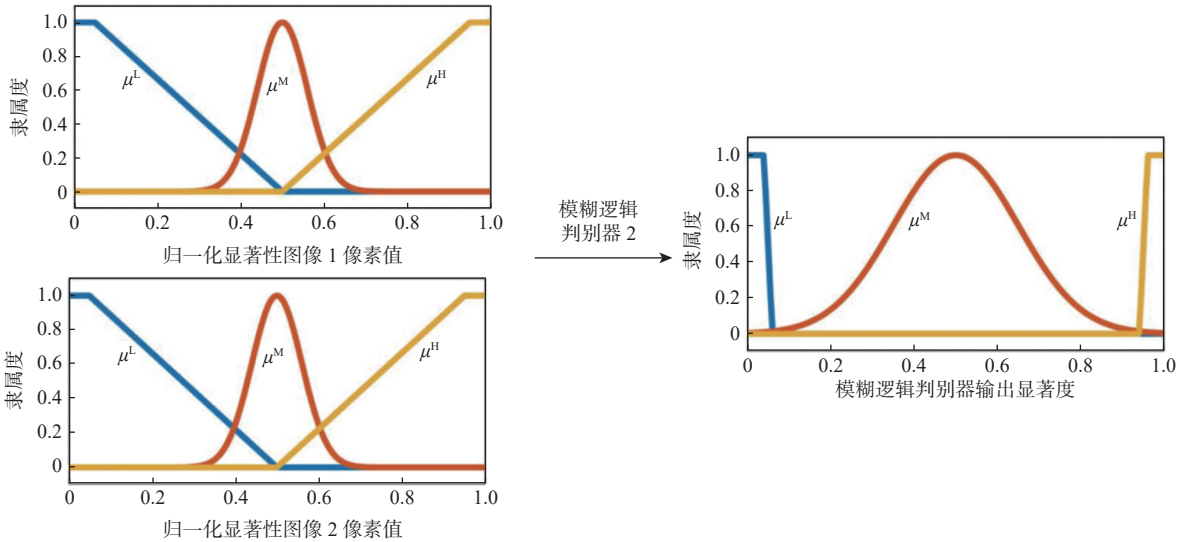


图 5　模糊逻辑判别器 2 隶属度函数

Fig. 5　Membership function of fuzzy logic discriminator 2

表 3　模糊逻辑判别器 2 的模糊规则

Table 3　Fuzzy rules for fuzzy logic discriminator 2

MF-sf	MF-sp		
	$\mu^{\mathrm{L}}_{\mathrm{in_2}}$	$\mu^{\mathrm{M}}_{\mathrm{in_2}}$	$\mu^{\mathrm{H}}_{\mathrm{in_2}}$
$\mu^{\mathrm{L}}_{\mathrm{in_2}}$	μ_{w1}	μ_{w2}	μ_{w3}
$\mu^{\mathrm{M}}_{\mathrm{in_2}}$	μ_{w1}	μ_{w2}	μ_{w3}
$\mu^{\mathrm{H}}_{\mathrm{in_2}}$	μ_{w3}	μ_{w3}	μ_{w3}

2.3　自适应参数融合策略

本文的 APFS 包括 AIFR 及 AGF[13]。图像融合规则是突出图像中感兴趣的特征并使其不重要的特征最小化。有效的融合规则在融合图像的客观质量评估中起着重要作用。本文提出一种基于 SSIM 的 AIFR, 实现红外和可见光图像的融合。本文提出的融合策略可以分为以下 2 个步骤:

步骤 1　得到联合显著性图像 $\boldsymbol{S}_{\mathrm{F}}$ 后, 利用 $\boldsymbol{S}_{\mathrm{F}}$ 作为滤波图像的引导图像, 再通过 AGF 得到多个决策图。

步骤 2　使用步骤 1 中得到的决策图, 采用设计的 AIFR 获得融合图像。

引导滤波是一种基于梯度信息保留图像边缘信息的滤波算法。因此, 引导滤波能够在多个决策图上保留融合显著性图像的显著信息。AGF 根据源图像的特征调整滤波参数, 从而达到优化实验结果的目的[13]。本文引入 AGF, 红外图像中纹理细节越多, 滤波半径越小, 反之亦然。实验中, 利用图像的平均梯度 I_{AG}[33] 评估纹理信息。I_{AG} 值越高, 纹理信息越丰富。

$$r = \mathrm{AGF}(I_{\mathrm{AG}}) \tag{13}$$

式中: r 为滤波半径。

利用联合显著性图像引导源图像得到多融合决策图, 定义为

$$\boldsymbol{D}^1 = G_{r_1,\varepsilon_1}(\boldsymbol{I}_{\mathrm{IR}}, \boldsymbol{S}_{\mathrm{F}}) \tag{14}$$

$$\boldsymbol{D}^2 = G_{r_2,\varepsilon_2}(\boldsymbol{I}_{\mathrm{VIS}}, \boldsymbol{S}_{\mathrm{F}}) \tag{15}$$

式中: $G_{r,\varepsilon}(\boldsymbol{I},\boldsymbol{S})$表示引导滤波操作; r_1、r_2为引导滤波半径[13]; ε_1、ε_2为引导滤波规整化因子, 取值为 0.01、0.01; $\boldsymbol{I}_{\mathrm{IR}}$ 和 $\boldsymbol{I}_{\mathrm{VIS}}$ 分别为红外和可见光图像; $\boldsymbol{D}^1$ 和 $\boldsymbol{D}^2$ 为决策图。

将决策图归一化至 0～1 区间:

$$\boldsymbol{D}^i_{\mathrm{N}} = \frac{\boldsymbol{D}^i - \min(\boldsymbol{D}^i)}{\max(\boldsymbol{D}^i) - \min(\boldsymbol{D}^i)} \qquad i = 1,2 \tag{16}$$

基于多融合决策图, 设计基于 SSIM 的 AIFR, SSIM[34] 是图像融合中有效的质量评价指标, 主要评估图像的结构信息、对比度信息、亮度信息。通过多融合决策图后, 得到 2 张初始融合图像, 对初始融合图像 $\boldsymbol{F}_1$ 和 $\boldsymbol{F}_2$ 分别与源图像比较基于 SSIM 的测度值。符合人眼视觉效果的初始融合图像权重增加(测度值更高), 越能够得到更加符合人眼视觉效果的融合结果, 详细参数设置将在 3.2 节实验部分进行相关讨论。具体融合规则如下:

$$\boldsymbol{F}_1 = \boldsymbol{I}_{\mathrm{IR}}(x,y)\boldsymbol{D}^1_{\mathrm{N}}(x,y) + \boldsymbol{I}_{\mathrm{VIS}}(x,y)(\boldsymbol{J} - \boldsymbol{D}^1_{\mathrm{N}}(x,y)) \tag{17}$$

$$\boldsymbol{F}_2 = \boldsymbol{I}_{\mathrm{IR}}(x,y)\boldsymbol{D}^2_{\mathrm{N}}(x,y) + \boldsymbol{I}_{\mathrm{VIS}}(x,y)(\boldsymbol{J} - \boldsymbol{D}^2_{\mathrm{N}}(x,y)) \tag{18}$$

$$\lambda_1 = \frac{\mathrm{SSIM\text{-}F}(\boldsymbol{F}_1, \boldsymbol{I}_{\mathrm{IR}}, \boldsymbol{I}_{\mathrm{VIS}})}{\mathrm{SSIM\text{-}F}(\boldsymbol{F}_1, \boldsymbol{I}_{\mathrm{IR}}, \boldsymbol{I}_{\mathrm{VIS}}) + \mathrm{SSIM\text{-}F}(\boldsymbol{F}_2, \boldsymbol{I}_{\mathrm{IR}}, \boldsymbol{I}_{\mathrm{VIS}})} \tag{19}$$

$$\lambda_2 = \frac{\mathrm{SSIM\text{-}F}(\boldsymbol{F}_2, \boldsymbol{I}_{\mathrm{IR}}, \boldsymbol{I}_{\mathrm{VIS}})}{\mathrm{SSIM\text{-}F}(\boldsymbol{F}_1, \boldsymbol{I}_{\mathrm{IR}}, \boldsymbol{I}_{\mathrm{VIS}}) + \mathrm{SSIM\text{-}F}(\boldsymbol{F}_2, \boldsymbol{I}_{\mathrm{IR}}, \boldsymbol{I}_{\mathrm{VIS}})} \tag{20}$$

$$\boldsymbol{F}(x,y) = \sum_{i=1,2} \lambda_i [\boldsymbol{I}_{\mathrm{IR}}(x,y)\boldsymbol{D}^i_{\mathrm{N}}(x,y) + \boldsymbol{I}_{\mathrm{VIS}}(x,y)(1 - \boldsymbol{D}^i_{\mathrm{N}}(x,y))] \tag{21}$$

式中: $\boldsymbol{J}$ 为全 1 矩阵; SSIM-F($\boldsymbol{F}_1, \boldsymbol{I}_{\mathrm{IR}}, \boldsymbol{I}_{\mathrm{VIS}}$)和SSIM-F($\boldsymbol{F}_2$,

$\boldsymbol{I}_{\mathrm{IR}}, \boldsymbol{I}_{\mathrm{VIS}}$)分别为相应初始融合图像与源图像的SSIM测度值; λ_1和λ_2为基于SSIM的自适应参数。

3 实验结果分析

3.1 实验数据集

为测试本文方法的性能,选取TNO[35]数据集中10组不同场景的图像组及RoadScenes[36]数据集中10组不同场景的图像组进行实验,TNO和RoadScenes是红外与可见光图像领域公开的数据集,包含常用的红外与可见光测试图像。

将本文方法与7个主流的红外与可见光图像融合方法进行比较,分别为梯度转移和总变异最小化法GTF[37]、深度融合网络DeepFuse[38]、低秩融合方法LatLRR[39]、FusionGAN[12]、DenseFuse[9]、多级低秩分解(multi-level image decomposition method based on latent low-rank representation, MDLatLRR)[40]、U2Fusion[15]。所有源代码来自于论文,比较方法中的参数与其论文一致。

3.2 实验参数

在结构块分解中,5×5的滑块大小能够提高算法效率,且能够提高实验结果[32],故选择5×5的滑块进行结构块分解。

模糊逻辑判别中有多个参数,这些参数对于决策判别十分重要。隶属度函数沿着中心轴对称,本文主要讨论右半轴的参数。在模糊逻辑判别器1中,隶属度函数输入的参数l和m控制$\mu_{\mathrm{in_1}}^{\mathrm{L}}$、$\mu_{\mathrm{in_1}}^{\mathrm{M}}$、$\mu_{\mathrm{in_1}}^{\mathrm{H}}$,如式(3)所示,根据大量实验,取得最好效果的输入隶属度函数参数值如表4所示。隶属度函数输出的参数c和d控制$\mu_{\mathrm{out_1}}^{\mathrm{L}}$的形状,$\sigma$控制$\mu_{\mathrm{out_1}}^{\mathrm{M}}$的形状,如式(4)和式(5)所示,根据大量实验,取得最好效果的输出隶属度函数参数值如表5所示。在模糊逻辑判别器2中,取得最好效果的输入隶属度函数参数值如表6所示。根据文献[32],对于$\mu_{\mathrm{in_2}}^{\mathrm{M}}$,参数$\sigma$取0.06,$\bar{x}$取0.5。对于$\mu_{\mathrm{out_2}}^{\mathrm{M}}$,参数$\sigma$取0.2,$\bar{x}$取0.5。取得最好效果的输出隶属度函数参数值如表7所示。

在APFS中,设计一种基于SSIM的自适应融合规则,将其与平均加权规则进行对比。图6(a)为红外图像,图6(b)为可见光图像,图6(c)为APFS得到的图像,图6(d)为固定参数引导滤波(式(14)、式(15)中r_1、r_2均取10)与等权重融合规则(式(21)中λ_1和λ_2取0.5)得到的图像,从选框放大处可看出,APFS能够得到对比度更好的图像,而固定参数方法会降低对比度,不利于人眼观测目标。

3.3 实验结果

3.3.1 主观视觉实验结果

如图7所示,选取包含本文方法的8种融合方式对“nato_camp”图像组融合进行主观视觉对比。可以看出,GTF方法的融合结果对比度丢失严重,丢失了可见光图像纹理信息。LatLRR、DeepFuse、DenseFuse与U2Fusion方法的融合结果中,可见光图像的纹理保持较好,但图像对比度信息丢失,FusionGAN方法在纹理信息保留方面存在不足,造成图像边缘处模糊,相对来说,MDLatLRR方法结果较好,突出了红外目标图像,但引入少量噪声导致融合结果产生锐化效果。本文方法很好地保留了红外图像的整体目标信息及可见光图像的纹理细节信息,相对于LatLRR、GTF、DeepFuse、DenseFuse、FusionGAN和U2Fusion方法,本文方法的细节信息明显增多,更好地保留了可见光图像的梯度纹理信息及红外图像的目标信息。相对于MDLatLRR方法,从选框放大处可以看出,本文方法没有引入额外噪声,图像更加符合人眼视觉感知。

表4 模糊逻辑判别器1输入隶属度函数的参数

Table 4 Fuzzy logic discriminator 1 parameters of input membership function

隶属度函数	l	m	u
$\mu_{\mathrm{in_1}}^{\mathrm{L}}$	−102	76.5	153
$\mu_{\mathrm{in_1}}^{\mathrm{M}}$	0	127.5	255
$\mu_{\mathrm{in_1}}^{\mathrm{H}}$	102	178.5	357

表5 模糊逻辑判别器1输出隶属度函数的参数

Table 5 Fuzzy logic discriminator 1 parameters of output membership function

隶属度函数	a	b	c	d
$\mu_{\mathrm{out_1}}^{\mathrm{L}}$	−0.36	−0.04	0.04	0.06
$\mu_{\mathrm{out_1}}^{\mathrm{M}}$				
$\mu_{\mathrm{out_1}}^{\mathrm{H}}$	0.94	0.96	1.04	1.36

表6 模糊逻辑判别器2输入隶属度函数的参数

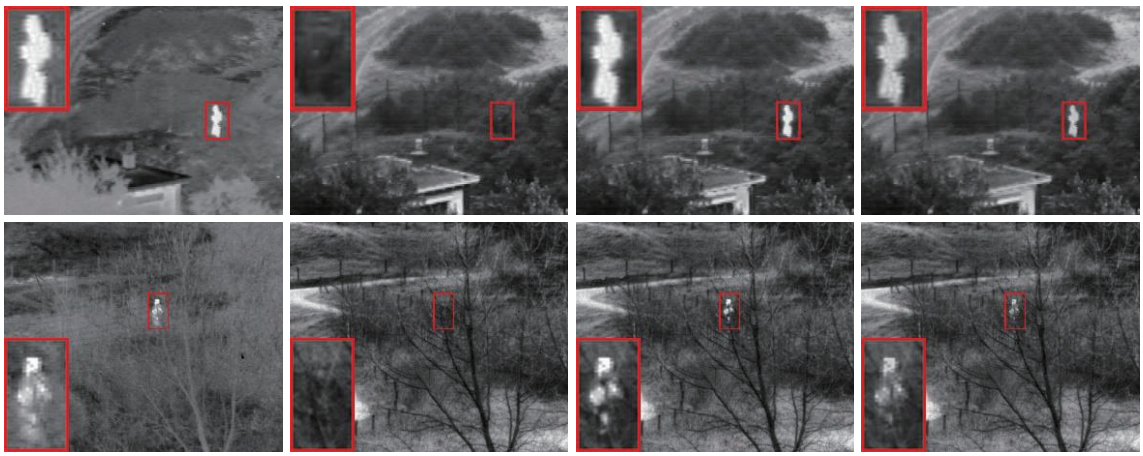
Table 6 Fuzzy logic discriminator 2 parameters of input membership function

隶属度函数	a	b	c	d
$\mu_{\mathrm{in_2}}^{\mathrm{L}}$	−0.271 1	−0.128 9	0.05	0.5
$\mu_{\mathrm{in_2}}^{\mathrm{M}}$				
$\mu_{\mathrm{in_2}}^{\mathrm{H}}$	0.5	0.95	1.129	1.271

表7 模糊逻辑判别器2输出隶属度函数的参数

Table 7 Fuzzy logic discriminator 2 parameters of output membership function

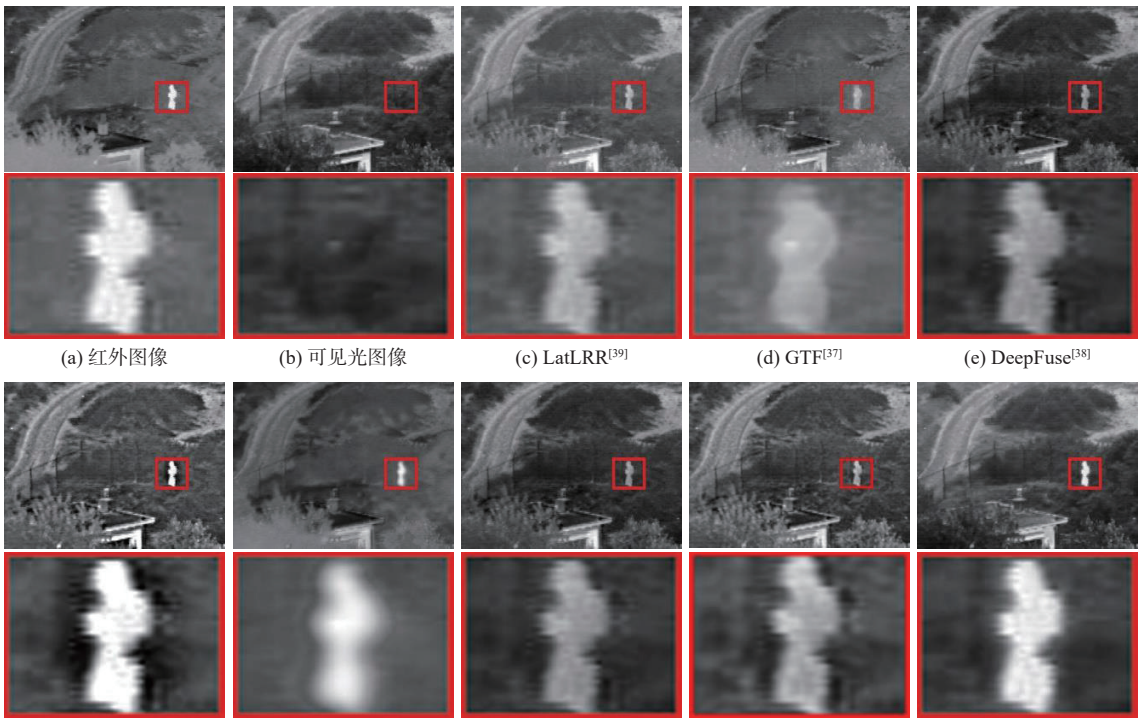
隶属度函数	a	b	c	d
$\mu_{\mathrm{out_2}}^{\mathrm{L}}$	−0.36	−0.04	0.025	0.05
$\mu_{\mathrm{out_2}}^{\mathrm{M}}$				
$\mu_{\mathrm{out_2}}^{\mathrm{H}}$	0.95	0.975	1.04	1.36



(a) 红外图像　(b) 可见光图像　(c) MFD-APFS　(d) 固定参数

图 6　APFS 与固定参数融合结果比较

Fig. 6　Comparison of fusion results between APFS and fixed parameters



(a) 红外图像　(b) 可见光图像　(c) LatLRR[39]　(d) GTF[37]　(e) DeepFuse[38]

(f) MDLatLRR[40]　(g) FusionGAN[12]　(h) DenseFuse[9]　(i) U2Fusion[15]　(j) MFD-APFS

图 7　“nato_camp”图像组融合结果

Fig. 7　Fusion results of “nato_camp” image set

为增加实验的可信度, 另外选取“Sandpath”图像组进行实验比较, 如图 8 所示。从融合结果来看, GTF 方法的融合图像整体模糊, 红外图像信息丢失严重。DeepFuse 和 DenseFuse 方法的融合结果中, 选框区域红外图像信息保持较好, 但树枝信息丢失严重。LatLRR 方法的融合图像整体上丢失了对比度信息; FusionGAN 方法融合图像整体模糊; MDLatLRR 和 U2Fusion 方法图像引入噪声, 造成图像锐化。相较于以上方法, 本文方法没有引入额外的噪声, 较好地保留了红外图像目标信息及可见光图像纹理信息。对“Kaptein_1654”图像组进行实验比较, 如图 9 所示, 本文方法的放大选框放大部分最接近可见光图像的纹理, 同时局部选框(未放大)中的红外人物部分在所有对比方法中也最接近红外图像。综上所述, 本文方法对于红外与可见光图像融合, 能够有效地保留可见光图像的细节和红外图像的热信息, 强化了图像的真实感, 并提升了视觉质量。

为证明模型的泛化性, 在 RoadScenes[36] 数据集上同样进行 8 种融合方法的实验验证, 选取“FLIR”图像组进行实验比较。从融合结果来看, LatLRR 和 GTF 方法的融合图像整体模糊, 红外图像信息丢失严重。DeepFuse 和 DenseFuse 方法的结果中车牌信息丢失严重, 几乎不可见; FusionGAN 方法融合



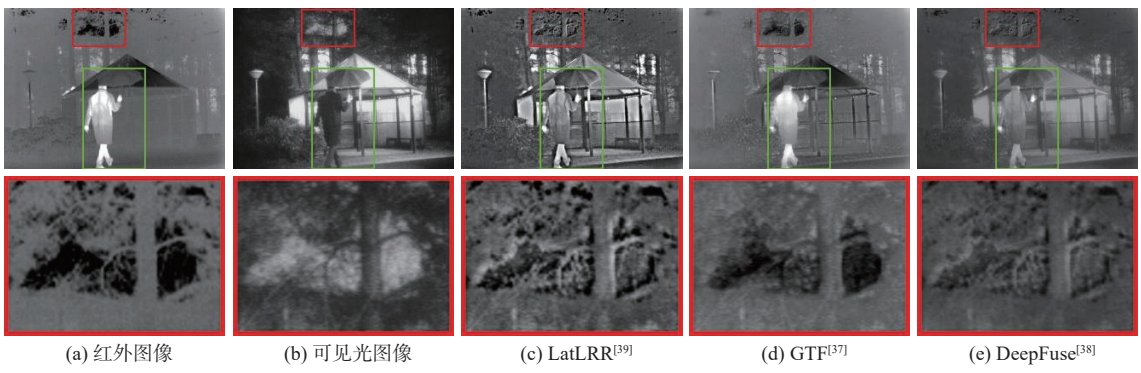
(a) 红外图像 (b) 可见光图像 (c) LatLRR[39] (d) GTF[37] (e) DeepFuse[38]



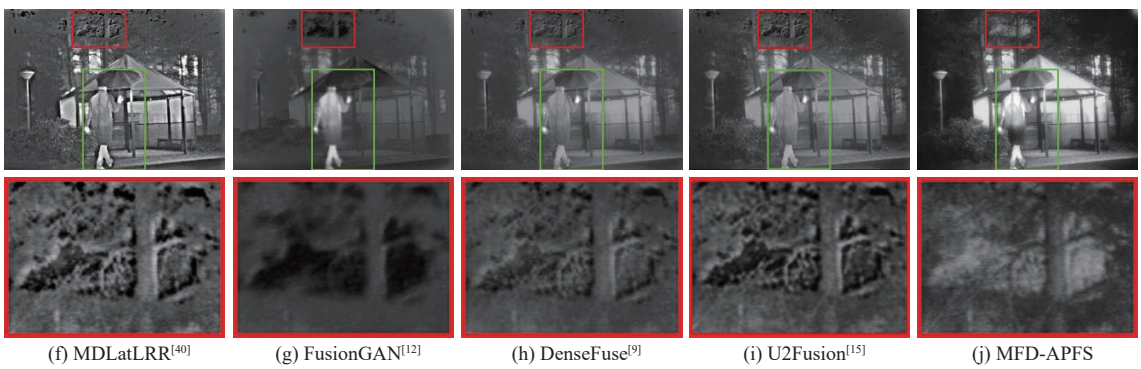
(f) MDLatLRR[40] (g) FusionGAN[12] (h) DenseFuse[9] (i) U2Fusion[15] (j) MFD-APFS

图 8 “Sandpath”图像组融合结果

Fig. 8 Fusion results of “Sandpath” image set



(a) 红外图像 (b) 可见光图像 (c) LatLRR[39] (d) GTF[37] (e) DeepFuse[38]



(f) MDLatLRR[40] (g) FusionGAN[12] (h) DenseFuse[9] (i) U2Fusion[15] (j) MFD-APFS

图 9 “Kaptein_1654”图像组融合结果

Fig. 9 Fusion results of “Kaptein_1654” image set

图像整体模糊；MDLatLRR 方法的融合图像引入较大噪声，造成图像锐化。U2Fusion 方法取得了较好的效果，但相较于本文方法，其对比度显著降低。综上所述，本文方法整体效果优于对比的 7 种方法，能够较好地保留红外图像目标显著信息及可见光图像的纹理信息。

3.3.2 客观融合指标实验结果

为进一步验证本文方法的效果，进行客观质量评

价。利用熵 E[41]、小波变换域互信息 FMIw[42]、离散余弦变换域互信息 FMIdct[42]、SSIM-F[43]、基于边缘的相似性度量 $Q^{AB/F}$[44] 和标准差融合指标 S_D[45] 6 种评价指标对融合结果进行定量分析, SSIM-F 越大, 表示融合结果在人眼视觉上越接近于源图像, FMIw、FMIdct 为测量互信息转换度, 表示融合图像从源图像中转换而来的信息数量, E、$Q^{AB/F}$ 和 S_D 为无参考测度指标, 指标越高, 融合图像质量越好。在 TNO[35] 数据集上, 各方法的融合结果客观指标的测定值如表 8 所示。从表 8 可以看出, 本文方法与所有对比方法相比, 6 种指标的量化评估结果均最好。本文方法的客观指标明显优于传统融合方法 GTF 和 LatLRR, 也优于 DeepFuse、FusionGAN、DenseFuse 和 U2Fusion 等深度学习融合方法。从平均定量指标来看, 本文方法的结果也为最优。在 RoadScenes[36] 数据集上, 各方法的融合结果客观指标的测定值如表 9 所示。可以看出, 本文方法与所有对比方法相比, 虽然在 E、FMIdct、S_D 指标上, GTF 优于其他方法, 但根据可视化的结果, GTF 方法引入了过多噪声信息, 导致融合图像整体模糊, 在指标 FMIw、SSIM-F 及 $Q^{AB/F}$ 上, 本文方法优于传统与深度学习融合方法。综合以上分析, 本文方法在定量与定性的分析结果上达到最优。

表 8　不同融合方法在 TNO 数据集上的客观指标对比

Table 8　Comparison of objective indicators of different fusion methods on TNO dataset

方法	E	FMIw	FMIdct	SSIM-F	$Q^{AB/F}$	S_D
LatLRR[39]	6.507 5	0.379 5	0.326 4	0.741 9	0.410 2	28.010 6
GTF[37]	6.579 9	0.439 3	**0.428 0**	0.758 6	0.444 5	28.234 4
DeepFuse[38]	6.907 3	0.421 6	0.411 8	0.753 2	0.459 0	36.828 1
MDLatLRR[40]	6.871 1	0.445 9	0.400 1	0.797 9	0.500 4	34.443 1
FusionGAN[12]	6.370 8	0.368 1	0.357 3	0.522 3	0.215 5	25.089 9
DenseFuse[9]	6.838 1	0.427 2	0.416 6	0.755 0	0.461 0	36.443 1
U2Fusion[15]	6.854 6	0.362 4	0.340 0	0.752 2	0.445 1	32.838 6
MFD-APFS	**7.030 1**	**0.524 2**	**0.428 0**	**0.966 9**	**0.640 7**	**48.743 1**

注: 加粗数值表示最优结果。

表 9　不同融合方法在 RoadScenes 数据集上的客观指标对比

Table 9　Comparison of objective indicators of different fusion methods on RoadScenes dataset

方法	E	FMIw	FMIdct	SSIM-F	$Q^{AB/F}$	S_D
LatLRR[39]	6.816 2	0.340 8	0.288 0	0.685 5	0.420 0	36.776 8
GTF[37]	**7.702 0**	0.381 1	**0.418 4**	0.703 2	0.445 4	**62.417 2**
DeepFuse[38]	7.319 3	0.404 7	0.364 6	0.727 3	0.477 1	50.720 7
MDLatLRR[40]	6.586 9	0.390 0	0.345 7	0.689 8	0.482 6	41.376 8
FusionGAN[12]	7.312 2	0.369 1	0.371 7	0.569 5	0.329 6	48.168 3
DenseFuse[9]	7.318 5	0.407 4	0.365 7	0.729 3	0.478 6	50.669 2
U2Fusion[15]	7.296 6	0.403 4	0.352 4	0.743 2	0.477 0	47.151 2
MFD-APFS	7.002 4	**0.499 0**	0.316 7	**0.918 5**	**0.532 3**	47.450 2

注: 加粗数值表示最优结果。

4　结　论

1) 本文提出的多级模糊逻辑判别机制可有效提取红外与可见光图像中的显著信息, 显著提升融合图像的对比度与细节保持能力。

2) 设计的自适应参数融合策略具备动态调节能力, 能根据不同源图像的特性自适应增强对比度, 改善红外图像信息不足的问题。

3) 本文提出的融合方法在多组公开数据集上均取得优异的主客观评价指标, 性能优于 7 种主流对比方法, 验证了其融合效果的先进性与稳定性。

4) 多级模糊逻辑判别与结构块分解结合的方案展示出良好的可扩展性, 为后续多模态图像融合任务提供了可行路径。

下一步研究将聚焦于模糊推理规则的自学习机制与融合模块的轻量化设计, 以提升模型的实用性与实时处理能力。

参考文献 (References)

[1] MA W H, WANG K, LI J W, et al. Infrared and visible image fusion technology and application: a review[J]. Sensors, 2023,

23(2): 599.

[2] LI H, MANJUNATH B S, MITRA S K. Multisensor image fusion using the wavelet transform[J]. Graphical Models and Image Processing, 1995, 57(3): 235-245.

[3] SINGH S, SINGH H, GEHLOT A, et al. IR and visible image fusion using DWT and bilateral filter[J]. Microsystem Technologies, 2023, 29(4): 457-467.

[4] REN Z G, REN G Q, WU D H. Fusion of infrared and visible images based on discrete cosine wavelet transform and high pass filter[J]. Soft Computing, 2023, 27(18): 13583-13594.

[5] AGHAMALEKI J A, GHORBANI A. Image fusion using dual tree discrete wavelet transform and weights optimization[J]. The Visual Computer, 2023, 39(3): 1181-1191.

[6] HUANG Z H, LI X, WANG L, et al. Spatially adaptive multi-scale image enhancement based on nonsubsampled contourlet transform [J]. Infrared Physics & Technology, 2022, 121: 104014.

[7] YANG C X, HE Y N, SUN C, et al. Infrared and visible image fusion based on QNSCT and guided filter[J]. Optik, 2022, 253: 168592.

[8] XU M L, TANG L F, ZHANG H, et al. Infrared and visible image fusion via parallel scene and texture learning[J]. Pattern Recognition, 2022, 132: 108929.

[9] LI H, WU X J. DenseFuse: a fusion approach to infrared and visible images[J]. IEEE Transactions on Image Processing, 2019, 28(5): 2614-2623.

[10] TANG L F, YUAN J T, ZHANG H, et al. PIAFusion: a progressive infrared and visible image fusion network based on illumination aware[J]. Information Fusion, 2022, 83: 79-92.

[11] TANG L F, XIANG X Y, ZHANG H, et al. DIVFusion: darkness-free infrared and visible image fusion[J]. Information Fusion, 2023, 91: 477-493.

[12] MA J Y, YU W, LIANG P W, et al. FusionGAN: a generative adversarial network for infrared and visible image fusion[J]. Information Fusion, 2019, 48: 11-26.

[13] YANG Y, LIU J X, HUANG S Y, et al. Infrared and visible image fusion via texture conditional generative adversarial network[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2021, 31(12): 4771-4783.

[14] YANG X, HUO H T, LI J, et al. DSG-Fusion: infrared and visible image fusion via generative adversarial networks and guided filter[J]. Expert Systems with Applications, 2022, 200: 116905.

[15] XU H, MA J Y, JIANG J J, et al. U2Fusion: a unified unsupervised image fusion network[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 44(1): 502-518.

[16] YANG Y, LIU J X, HUANG S Y, et al. VMDM-fusion: a saliency feature representation method for infrared and visible image fusion [J]. Signal, Image and Video Processing, 2021, 15(6): 1221-1229.

[17] ZHANG C F, ZHANG Z Y, FENG Z L. Image fusion using online convolutional sparse coding[J]. Journal of Ambient Intelligence and Humanized Computing, 2023, 14(10): 13559-13570.

[18] LI H Y, ZHANG C F, HE S D, et al. A novel fusion method based on online convolutional sparse coding with sample-dependent dictionary for visible-infrared images[J]. Arabian Journal for Science and Engineering, 2023, 48(8): 10605-10615.

[19] TAO T W, LIU M X, HOU Y K, et al. Latent low-rank representation with sparse consistency constraint for infrared and visible image fusion[J]. Optik, 2022, 261: 169102.

[20] GUO Z Y, YU X T, DU Q L. Infrared and visible image fusion based on saliency and fast guided filtering[J]. Infrared Physics & Technology, 2022, 123: 104178.

[21] YANG Y, LIU J X, HUANG S Y, et al. Infrared and visible image fusion based on modal feature fusion network and dual visual decision[C]//Proceedings of the IEEE International Conference on Multimedia and Expo. Piscataway: IEEE Press, 2021: 1-6.

[22] MA K D, LI H, YONG H W, et al. Robust multi-exposure image fusion: a structural patch decomposition approach[J]. IEEE Transactions on Image Processing, 2017, 26(5): 2519-2532.

[23] WANG Z, BOVIK A C, SHEIKH H R, et al. Image quality assessment: from error visibility to structural similarity[J]. IEEE Transactions on Image Processing, 2004, 13(4): 600-612.

[24] YANG Y, QUE Y, HUANG S Y, et al. Multimodal sensor medical image fusion based on type-2 fuzzy logic in NSCT domain[J]. IEEE Sensors Journal, 2016, 16(10): 3735-3745.

[25] ZAND M D, ANSARI A H, LUCAS C, et al. Risk assessment of coronary arteries heart disease based on neuro-fuzzy classifiers [C]//Proceedings of the 17th Iranian Conference of Biomedical Engineering. Piscataway: IEEE Press, 2010: 1-4.

[26] RAHMAN M A, LIU S, WONG C Y, et al. Multi-focal image fusion using degree of focus and fuzzy logic[J]. Digital Signal Processing, 2017, 60: 1-19.

[27] BALASUBRAMANIAM P, ANANTHI V P. Image fusion using intuitionistic fuzzy sets[J]. Information Fusion, 2014, 20: 21-30.

[28] MANCHANDA M, SHARMA R. A novel method of multimodal medical image fusion using fuzzy transform[J]. Journal of Visual Communication and Image Representation, 2016, 40: 197-217.

[29] RIZZI A, GATTA C, PIACENTINI B, et al. Human-visual-system-inspired tone mapping algorithm for HDR images[C]//Proceedings of the Human Vision and Electronic Imaging IX. Bellingham: SPIE, 2004.

[30] WANG Z J, TONG X Y. Consistency analysis and group decision making based on triangular fuzzy additive reciprocal preference relations[J]. Information Sciences, 2016, 361: 29-47.

[31] CHAIRA T. A rank ordered filter for medical image edge enhancement and detection using intuitionistic fuzzy set[J]. Applied Soft Computing, 2012, 12(4): 1259-1266.

[32] YANG Y, WU J H, HUANG S Y, et al. Multimodal medical image fusion based on fuzzy discrimination with structural patch decomposition[J]. IEEE Journal of Biomedical and Health Informatics, 2019, 23(4): 1647-1660.

[33] CUI G M, FENG H J, XU Z H, et al. Detail preserved fusion of visible and infrared images using regional saliency extraction and multi-scale image decomposition[J]. Optics Communications, 2015, 341: 199-209.

[34] WANG Z, SIMONCELLI E P, BOVIK A C. Multiscale structural similarity for image quality assessment[C]//Proceedings of the Asilomar Conference on Signals, Systems & Computers. Piscataway: IEEE Press, 2003: 1398-1402.

[35] TOET. TNO image fusion dataset[EB/OL]. (2022-10-15)[2023-06-01]. https://figshare.com/articles/TN_Image_Fusion_Dataset/1008029.

[36] XU H, MA J Y, LE Z L, et al. FusionDN: a unified densely connected network for image fusion[J]. Proceedings of the AAAI Confer-

ence on Artificial Intelligence, 2020, 34(7): 12484-12491.

[37] MA J Y, CHEN C, LI C, et al. Infrared and visible image fusion via gradient transfer and total variation minimization[J]. Information Fusion, 2016, 31: 100-109.

[38] PRABHAKAR K R, SRIKAR V S, BABU R V. DeepFuse: a deep unsupervised approach for exposure fusion with extreme exposure image pairs[C]//Proceedings of the IEEE International Conference on Computer Vision. Piscataway: IEEE Press, 2017: 4724-4732.

[39] LI H, WU X J. Multi-focus image fusion using dictionary learning and low-rank representation[C]//Proceedings of the International Conference on Image and Graphics. Berlin: Springer, 2017: 675-686.

[40] LI H, WU X J, KITTLER J. MDLatLRR: a novel decomposition method for infrared and visible image fusion[J]. IEEE Transactions on Image Processing, 2020, 29: 4733-4746.

[41] VAN AARDT J. Assessment of image fusion procedures using entropy, image quality, and multispectral classification[J]. Journal of Applied Remote Sensing, 2008, 2(1): 023522.

[42] HAGHIGHAT M, RAZIAN M A. Fast-FMI: non-reference image fusion metric[C]//Proceedings of the IEEE 8th International Conference on Application of Information and Communication Technologies. Piscataway: IEEE Press, 2014: 1-3.

[43] YANG C, ZHANG J Q, WANG X R, et al. A novel similarity based quality metric for image fusion[J]. Information Fusion, 2008, 9(2): 156-160.

[44] QU G H, ZHANG D L, YAN P F. Information measure for performance of image fusion[J]. Electronics Letters, 2002, 38(7): 313.

[45] MA J Y, MA Y, LI C. Infrared and visible image fusion methods and applications: a survey[J]. Information Fusion, 2019, 45: 153-178.

Fuzzy logic and adaptive strategy for infrared and visible light image fusion

YANG Yong[1,2], LIU Jiaxiang[3], HUANG Shuying[4,*], WANG Xiaozheng[5], XIA Yukun[3]

(1. School of Computer Science and Technology, Tiangong University, Tianjin 300387, China;
2. Cangzhou Institute of Tiangong University, Cangzhou 061000, China;
3. School of Information Management and Mathematics, Jiangxi University of Finance and Economics, Nanchang 330032, China;
4. School of Software, Tiangong University, Tianjin 300387, China;
5. School of Control Science and Engineering, Tiangong University, Tianjin 300387, China)

Abstract: Due to different imaging mechanisms, infrared imaging can capture target information under special conditions where the targetis obstructed, while visible light imaging can capture the texture details of the scenarios. Therefore, to obtain a fusion image containing both target information and texture details, infrared imaging and visible light imaging are generally combined to facilitate visual perception and machine recognition. Based on fuzzy logic theory, an infrared and visible light image fusion method was proposed,combining multistage fuzzy discrimination and adaptive parameter fusion strategy (MFD-APFS). First, the infrared and visible light images were decomposed into structural patches, obtaining the contrast-detail image set reconstructed by the signal intensity component. Second, the source image stand contrast-detail image set were processed through a designed fuzzy discrimination system, generating saliency maps for each set. A second-stage fuzzy discrimination was then applied to produce a unified saliency map. Finally, the guided filtering technique was used, with the saliency map guiding the source image to obtain multiple decision graphs. The final fusion image was obtained through the adaptive parameter fusion strategy. The proposed MFD-APFS method was experimentally evaluated on publicly available infrared and visible light datasets. Compared to the seven mainstream fusion methods, the proposed method shows improvements in objective metrics. On the TNO dataset, SSIM-F and $Q^{AB/F}$ were improved by 0.169 and 0.140 3, respectively, and on the RoadScenes dataset, they were improved by 0.175 3 and 0.053 7, respectively. Furthermore, the subjective visual analysis indicates that the proposed method is capable of generating fusion images with clear targets and enriched details while retaining infrared target information and visible texture information.

Keywords: infrared and visible light image fusion; adaptive parameter fusion strategy; multistage fuzzy logic; guided filtering; decision map

Received: 2023-06-16; **Accepted**: 2023-11-24; **Published Online**: 2025-04-14 17:58

URL: link.cnki.net/urlid/11.2625.V.20250414.1647.003

Foundation items: National Natural Science Foundation of China (62072218,61862030); Natural Science Foundation of Tianjin (24JCZDJC00130); Project of Cangzhou Institute of Tiangong University (TGCYY-Z-0303)

* **Corresponding author**. E-mail: shuyinghuang2010@126.com

http://bhxb.buaa.edu.cn jbuaa@buaa.edu.cn
DOI: 10.13700/j.bh.1001-5965.2023.0391

空间信息增强的室内多任务 RGB-D 场景理解

孙国栋，熊晨韵，刘俊杰，张杨*
(湖北工业大学 机械工程学院，武汉 430068)

摘 要：移动机器人在探索三维空间时需要获取大量场景信息，这些信息包含语义、实例对象、位置关系等多个方面。理解场景信息的准确性和计算复杂性是移动端关注的2个焦点。基于此，提出了一种适用于室内场景理解的空间信息增强的多任务学习方法。该方法由包含通道-空间注意力融合模块的编码器及多任务头的解码器组成，可同时实现语义分割、全景分割(实例分割)和方向估计多个任务。其中，通道-空间注意力融合模块旨在增强RGB和深度各自的模态特征，由简单卷积构成的空间注意力机制可降低收敛速度，与通道注意力机制信息融合后，进一步强化全局信息的位置特征。语义分支的上下文模块位于解码器后，为像素级语义信息提供有力支持，有助于减小模型大小。同时，设计了一种基于硬参数共享且能均衡训练任务的损失函数，探讨合适的轻量级骨干网络和任务数量对提升场景理解算法性能的影响。在新增标签注释的室内数据集NYUv2和SUN RGB-D上，评估了多任务学习方法的有效性，综合性全景分割精度分别提高了2.93%和4.87%。

关 键 词：场景理解；多任务学习；RGB-D；空间信息；室内场景
中图分类号：V221+.3; TB553
文献标志码：A **文章编号**：1001-5965(2025)07-2209-09

场景理解是一项多模态信息学习任务，需要对三维动态场景进行感知、分析和解释[1]。使用多任务学习能够方便且有效地提高模型的泛化能力，并实现多个重要任务，如语义分割、实例分割、全景分割和方向估计，这些任务提供了关于场景的丰富信息，如不同物体的位置、大小、形状和运动方向等。本文的目标是通过移动机器人来探索室内场景信息，因此，更集中研究以上任务。这些任务相互关联，通过多任务学习可以更好地协同工作，提高场景理解的准确性和效率。

在早期的场景理解任务中，通常只依赖RGB图像作为输入，因此会引入新的问题，如忽略图像中目标对象的纹理和光照条件对图像输入产生的影响。随着基于结构光法和飞行时间法(time of flight, TOF)的RGB-D相机不断问世，RGB与深度图像像素可一一对应，双输入逐渐成为研究热点[2]。然而，由于摄像机和外部环境的干扰，深度图像中往往存在噪声。因此，亟需一个高效的机制过滤无关因素。Hu等[3]提出一种有效捕捉通道特征的注意力模块，但仅适用于有明显通道特征的任务(如目标检测等)。深度图中的空间信息与RGB图像中的语义信息有所不同，且空间注意力与通道注意力的匹配具有实现最优解的能力，因此，需要一种捕捉空间特征的注意力机制[4]。

目前，基于编码-解码器的多任务场景理解方法得到广泛关注[5]。这种运用于移动端的三维场景理解方法在不断改进和发展，越来越符合工业实际需求。然而，分割任务中，目标与场景常常具有很大的尺度差异。基于求和的跳过连接模块连接网络前后的信息，可以在获得不同尺度信息的同时，有

收稿日期：2023-06-19；**录用日期**：2024-03-08；**网络出版时间**：2025-05-29 14:40
网络出版地址：link.cnki.net/urlid/11.2625.V.20250529.0856.001
基金项目：国家自然科学基金(51775177)；湖北省揭榜制科技项目(2024BEB018)
* **通信作者**. E-mail：yzhangcst@hbut.edu.cn

引用格式：孙国栋，熊晨韵，刘俊杰，等. 空间信息增强的室内多任务RGB-D场景理解[J]. 北京航空航天大学学报，2025，51(7)：2209-2217.
SUN G D，XIONG C Y，LIU J J，et al. Spatial information-enhanced indoor multi-task RGB-D scene understanding[J]. Journal of Beijing University of Aeronautics and Astronautics，2025，51(7)：2209-2217 (in Chinese).

效减少梯度消失并解决网络退化问题[6]。基于注意力的网络也可以捕获全局依赖关系和随机上下文信息[7]。在此基础上，一些方法构建自适应的上下文模块[8-10]来增强编码器的特性，扩大感受野，减少降采样造成的分辨率损失，构建编码器和解码器之间的信息桥梁。然而，以上方法均需要复杂的主干来获得丰富的语义信息，消耗大量的计算资源。从优化计算效率角度，图像级联网络（image cascade network, ICNet）[11]通过减小输入图像的大小，以较小的计算量实现快速推理，但随之而来的是信息丢失和目标边缘模糊等问题。一些特殊的网络结构通过减少通道数和卷积层[12]来减少计算。深度可分离卷积[13-16]等轻量化卷积神经网络结构能代替传统卷积，减少计算参数量。

针对以上问题，本文提出了一种面向室内环境的多任务场景理解方法，采用RGB和深度图作为信息输入；为持续关注特定的输入信息，在编码过程中添加捕捉空间和通道信息的注意力模块；同时，在语义解码分支增加了能充分捕获全局信息的上下文模块，帮助移动机器人获取丰富的语义信息，为后续任务提供有力支持；选择硬参数共享的多任务学习方式，在每个任务的末端进行选择性的信息交互，减少计算参数，合适的骨干网络也有助于高效实现多项任务的场景理解。综上所述，本文的主要创新和贡献如下：

1）构建了一种强化空间信息的室内多任务RGB-D场景理解网络，该网络由进一步提取空间信息的注意力融合模块和实现多项任务的解码分支组成，保证分割精度的同时减少参数计算。

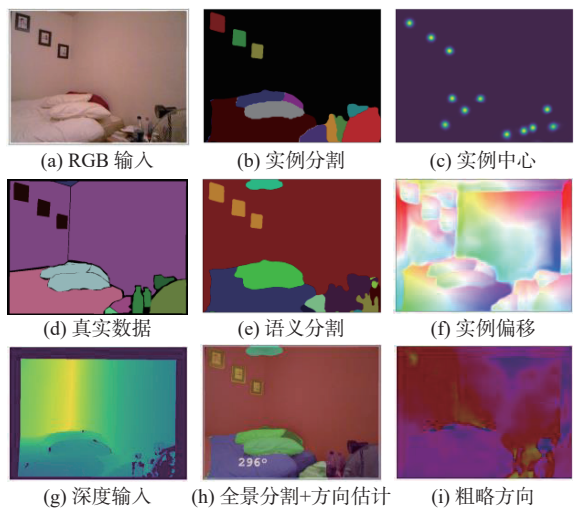
2）设计了一种关注信道和空间特征的注意力融合模块，可以强化RGB和深度图的特征表达，且结构简单，有效提高了网络分割精度。

3）在语义解码分支增加了能充分捕获全局信息的上下文模块，扩大不同尺度的感受野，保证网络分割精度的同时减少冗余计算。

1 相关工作

1.1 全景分割

全景分割通常是语义分割和实例分割的结合[17]。现有的全景分割算法存在机遇和挑战。其中，统一的全景分割网络（unified panoptic segmentation network, UPSNet）[18]是一种端到端的神经网络模型，能够通过映射变换将目标实例分割子分支与背景分割子分支的输出结合起来，从而得到全景头部输出的特征张量，然而，该算法的计算复杂度较高，需要大量的计算资源和时间。高效的全景分割（efficient panoptic segmentation, EfficientPS）[19]具备自适应调整实例和语义分割掩码融合的能力，但其对数据集的要求较高，很可能出现过拟合和欠拟合的情况。自下而上的全景分割基准Panoptic-DeepLab[20]能够通过语义分割生成前景掩码，并将前景像素分组到最近的实例中心，从而获得不可知实例分割，在语义与实例分割的特征融合后，生成最终的全景分割。图1展示了本文室内RGB-D场景理解方法的多任务输出，其中，图1(a)、(g)为网络输入，图1(d)为真实数据标注，图1(b)、(c)、(e)、(f)、(h)、(i)为任务输出结果。为了能够实现移动边缘部署，选择自下而上的全景分割方法。



(a) RGB输入 (b) 实例分割 (c) 实例中心
(d) 真实数据 (e) 语义分割 (f) 实例偏移
(g) 深度输入 (h) 全景分割+方向估计 (i) 粗略方向

图1 本文室内RGB-D场景理解方法的多任务输出

Fig. 1 Multi-task output of indoor RGB-D scene understanding method

1.2 方向估计

方向估计是在实例分割的基础上进一步估计每个已分割对象的三维方向，在深度学习中，被广泛采用以识别物体或场景的朝向，更准确地预测目标的运动趋势，优化追踪性能。方向估计还能应用在机器人、人体运动跟踪、航空航海等跟踪领域[21]。Mahfoudi等[22]提出了回归方法MultiBin，能够在不需要精确三维场景信息的情况下进行三维目标检测，但该方法需要进行图像预处理，将二维图像投影到三维空间中，增加了计算成本。Christie等[23]采用经典的编码-解码网络结构，将均方误差（mean squared error, MSE）损失运用到建筑物的角度估计上，无需多个相机就能估计建筑物的方向角，但该方法对图像质量要求很高，需要清晰、无噪声的图像才能获得较好的结果。为进一步改进该问题，Panoptic-DeepLab[20]通过离散的训练标签，引入von Mises损失进行连续、周期性的方向估计。

1.3 多任务学习

多任务学习比单任务学习更贴近人类学习方

式, 一个网络模型就能实现多项相关任务。好的多任务学习应该在满足以上要求的同时, 实现多个任务的均衡发展。软参数共享方法[24] 对不同的任务采用独立的主体结构, 在底层阶段相互传递信息, 在特定层之间没有信息交互。这样做的好处是模型具有更好的泛化性能, 减少对特定任务的依赖, 使模型更容易适应新任务。但同时软参数共享需要复杂的权重分配方式来确定损失函数的权重。硬参数共享的方式[25] 共享模型的主体参数, 而在任务阶段彼此分开。这种方法可以减少训练过程中的参数, 降低模型过拟合的风险。由于本文主要探讨语义分割、实例分割、全景分割、方向估计 4 个任务, 训练过程较为复杂, 为提高模型的稳定性和实用性, 缓解权重分配不均的情况, 选择硬参数共享的多任务学习方法。

2　室内多任务 RGB-D 场景理解

2.1　多任务场景理解网络

如图 2 所示, 本文提出了一种基于强化空间信息的室内多任务 RGB-D 场景理解网络。该结构采用典型的编码器-解码器结构, 在 0～4 层编码器中, 加入注意力融合模块, 层层融合后, 加强空间信息对场景对象的几何理解, 获得更加丰富的尺度特征。

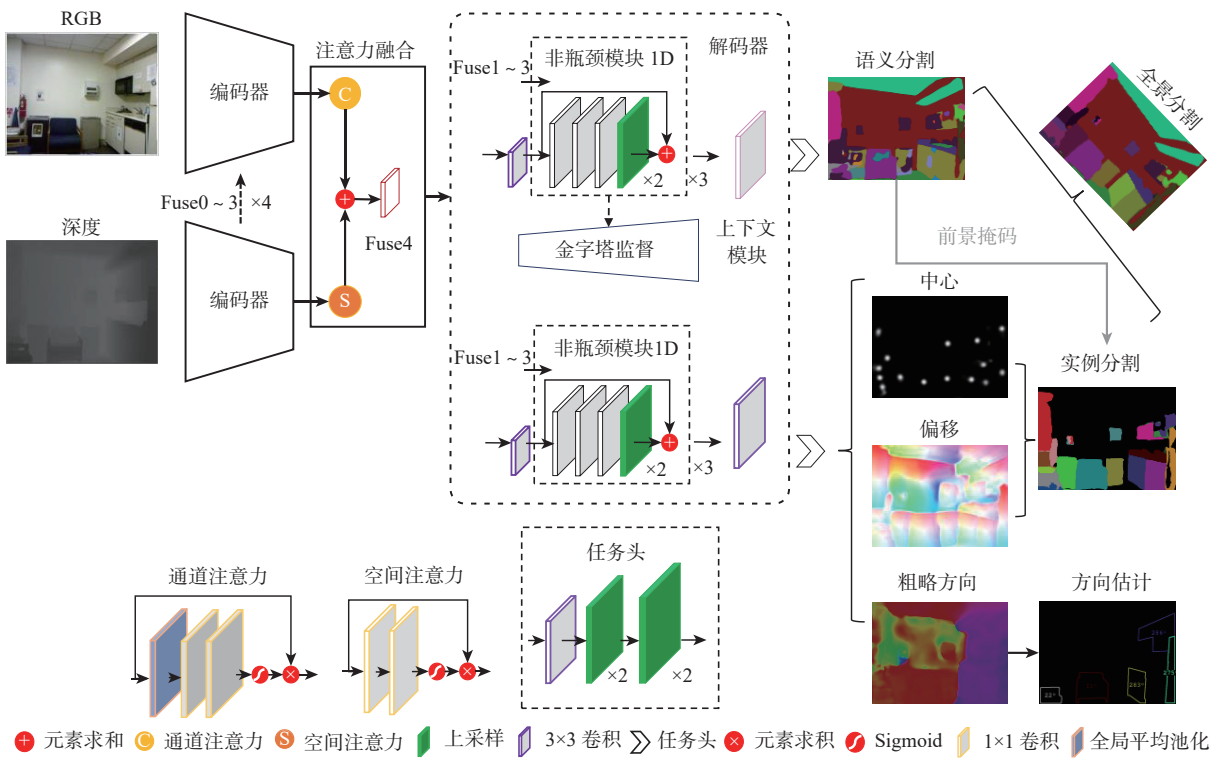


图 2　强化空间信息的室内多任务 RGB-D 场景理解网络结构

Fig. 2　Structure of spatial information-enhanced indoor multi-task RGB-D scene understanding network

值得注意的是, 实例分割通过一般卷积层就可以完成目标检测和语义分割的任务, 直接对实例进行像素级建模。在实例分割任务上增加上下文模块反而会导致模型过拟合, 增加计算量[26]。因此, 本文提出将金字塔池化及上下文模块仅放在语义分割的解码器末端, 充分利用不同尺度的全局信息。跳跃连接(Fuse1～3)将来自骨干网络中的低分辨率特征经过上采样与高分辨率特征进行融合, 获得更精确的特征。

在解码部分, 用到了 2 个解码器, 分别作用于语义分割和实例分割, 两者架构相似。其中, 语义解码器通过 1 个 3×3 卷积、3 个非瓶颈模块 1D 和 2 倍上采样来完成, 上采样使用的是可学习的 3×3 上采样[27], 其可以自适应地调整上采样方式, 从而提高模型的泛化能力。同时, 增加解码器的金字塔监督输出也是为了学习更多不同分辨率的特征, 以捕捉不同尺度的物体和语义信息。

通过不同的任务头, 分别引出实例中心预测、每个像素关于实例中心点坐标偏移的预测、粗略方向预测。每个任务头的结构也是相似的, 由 1 个 3×3 卷积和 2 个上采样块组成, 每个上采样块为 2 层最近邻上采样叠加在一起。语义分割为实例分割提供前景掩码, 与预测的实例中心和坐标偏移结合生成预测的实例分割结果。全景分割是语义分割和实例分割结合的产物, 其中, 每个实例的语义类由语义分割投票最多的确定。精确的方向估计

建立在实例分割的基础上，预测实例中所有像素的方向。

为简化网络表达，图2中的网络框架省略了批量归一化BN和激活函数ReLU，但在实际应用中，它们是必要的，并常常组合使用以提高模型性能，使训练更加稳定和高效[28]。

2.2 通道-空间注意力融合模块

RGB分支和深度分支的信息输入不同，RGB图像可以获取关于物体的形状、外观和表面属性的信息，深度图中的每个像素点则表示该点与相机的距离或深度值。但以前的方法往往使用同一注意力机制来处理[8-10]，压抑某一分支的信息特征。因此，为有效提取RGB与深度图的相对特征，使深度信息更好地辅助通道信息的理解，本文设计一种适应不同信息特征的注意力融合模块，空间注意力模块专注于提取深度通道中的重要特征，而通道注意力模块则对RGB通道进行全局上下文感知，从而更好地捕捉RGB通道的关键特征。这种分别处理通道的方式可以更好地利用2个通道的特点，提高语义分割的性能。2种模态特征经过通道注意力模块和空间注意力模块的重新元素求和后形成新的特征(Fuse0～4)，如图2中注意力融合模块所示。

各注意力模块具体结构如图2所示。输入特征$x_i \in X=\{x_1,x_2,\cdots,x_i\}$，$X \in \mathbf{R}^{C\times H\times W}$，其中，$C$为通道数，$H$、$W$分别为输入特征的长度、宽度，$i$表示第$i$个特征输入。

通道注意力V_{rgb}表示为

$$V_{\mathrm{rgb}}(x_i)=x_i+A[\sigma W_{1\times1}(W_{1\times1}(x_i))] \tag{1}$$

式中：A为全局平均池化；$W_{1\times1}$表示1×1的卷积计算；σ表示激活函数Sigmoid。

全局平均池可以强化类别与特征图的联系，更重视整体信息，但会提高收敛速度。而空间注意力更关注特征在空间中的位置，用卷积层就已经能满足传递相应信息的要求。因此，空间注意力V_{d}及最终融合的注意力V为

$$\begin{cases} V_{\mathrm{d}}(x_i)=x_i+\sigma W_{1\times1}(W_{1\times1}(x_i)) \\ V=V_{\mathrm{rgb}}(x_i)+V_{\mathrm{d}}(x_i) \end{cases} \tag{2}$$

还有其他能强化深度信息的注意力方法，但这些方法无法在边缘设备上部署，或计算量太大，综合以上因素，确定了适合的注意力融合模块。

2.3 上下文模块

在解码器中，本文网络实际感受野范围比空间操作所产生的理论感受野小得多。因此，在上下文模块中，采用了类似于金字塔池模块[8]。特征信息经过自适应平均池化后分解成多个分支，得到不同分辨率的上下文信息，再聚合成带有不同尺度的全局特征表示(如图3中双线性上采样模块所示)。本文的上下文模块有2层，每层的尺寸分别为1×1和5×5。图3中，n为模块的层数，C_{in}为输入通道数。在编码-解码器结构中，上下文模块放在编码器后，其主要作用是通过上下文信息来提高编码器对输入图像的特征提取能力。但本文在编码器中已有跳跃连接层(Fuse1～3)与解码器信息相融合，能补充因输入图像降维而丢失的重要信息。因此，相比于更换上下文模块，探讨其位置对于图像特征提取和语义分割能力更有意义。

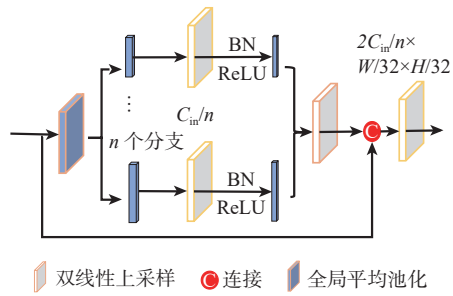


图3 多尺度上下文模块

Fig. 3 Multi-scale context module

2.4 多任务损失函数

1）语义分割。语义分割的本质是对像素分类，交叉熵函数L_{s}表示为

$$L_{\mathrm{s}}(u_i)=-\sum_{i=1}^{n}p(u_i)\lg q(u_i) \tag{3}$$

式中：$p(u_i)$为真实分布；$\lg q(u_i)$为模型输出的概率分布；u_i为类别，i表示第i个特征输入；n为样本总数。该过程等价于Softmax函数、log函数和NLLLoss函数的综合计算。

2）实例分割。实例中心和距离实例中心的偏移是需要计算的。因此，用平均绝对误差函数L_1来预测目标变量(实例中心)与预测变量之间的绝对差值，即偏移计算。实例分割的中心损失函数L_1为

$$L_1(y_i)=\frac{1}{n}\sum_{i=1}^{n}|y_i-\hat{y}_i| \tag{4}$$

式中：y_i为第i个样本的真实值；$\hat{y}_i$为预测值。

采用均方误差函数L_2来计算中心预测，且实例分割任务采用来自语义分割的真实实例掩码，距离实例中心的偏移损失函数L_2表示为

$$L_2(y_i)=\frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2 \tag{5}$$

3）全景分割。参考Panoptic-DeepLab[20]，在后处理实例中心时，应用0.1的阈值和池化大小为17的最大池化来执行关键点非最大抑制，过滤前64个实例，将实例与语义分割合并，生成全景分割结果。在全景分割中，还要考虑这2种损失函数的

平衡搭配, 即需要仔细调整初始化和损失权重。通过对实例中心和偏移采用相对应的激活函数 Sigmoid、tanh, 来稳定实例分割方面的性能。

4) 方向估计。本文考虑更多的是物体在场景中的角度。角度是连续且带有周期性的, 因此, 引入 von Mises 分布。von Mises 分布是一种圆上连续概率分布的模型[29], 符合正态分布的循环模拟。结合二元数编码[29], 周期性的角度表达为 $f=(\cos\varphi,\sin\varphi)$, φ 为角度。方向预测损失函数 L_{vm} 为

$$L_{\mathrm{vm}}(f|t)=1-\mathrm{e}^{\kappa(ft-1)} \tag{6}$$

式中: κ 与近似高斯的方差成反比, 是一个简单的超参数, 用于控制损失函数的尾部; t 为目标值。采用以 e 为底的指数函数是为了减小微小错误惩罚的影响力。

综上, 其他任务的总损失函数 L_{all} 定义为

$$L_{\mathrm{all}}=a(bL_1+L_2)+L_{\mathrm{s}}+cL_{\mathrm{vm}} \tag{7}$$

式中: a、c 分别为实例分割、方向估计任务的损失权重; b 为实例中心与实例偏移的初始权重比。

在神经网络框架下学习到的多个任务间存在冲突, 因此, 如何平衡多个任务的损失权重是值得考虑的问题。

3　实　验

3.1　数据集及实验细节

1) 室内数据集。在 NYUv2[30] 和 SUN RGB-D[31] 数据集上验证本文方法在语义分割、实例分割和方向估计 3 个任务上的有效性。这 2 个数据集在标签数量和用途上存在差异, 具体如表 1 所示, 包含了 NYUv2 和 SUN RGB-D 的训练集和测试集样本数量。为使 RGB 图像更生动, 在 HSV 空间上进行了轻微的颜色抖动处理。

① NYUv2[30]。原本就有针对 RGB 和深度图的语义分割标签; 带有实例分割、方向估计标签的数据集来自 EMSANet[31]。为规避非常小的实例, 实例分割是在语义分割标签的基础上限制实例面积不小于图像区域的 0.25%。手动注释实例方向的标签在点云[22] 中进行, 由于图像数据集是拍摄三维场景的二维平面图像, 需要考虑透视变形。全景分割将墙、地板和天花板作为背景 stuff, 其他类作为物品类 thing。

② SUN RGB-D[32]。没有实例分割和方向估计的注释, 但有三维物体边框和三维物体方向的标签, 因此, 可以通过创建语义与三维框类的映射, 再将其与三维语义点云匹配, 为边框获取实例掩码和方向。全景分割情况与 NYUv2 一致。

表 1　多任务数据标签

Table 1　Multi-task data label

主要任务	NYUv2		SUN RGB-D	
	训练集样本数量	测试集样本数量	训练集样本数量	测试集样本数量
语义分割	795	654	5 285	5 050
实例分割	12 092	9 874	18 171	16 961
方向估计	2 696	2 096	13 076	12 440

2) 评价指标。为验证网络的有效性, 依据不同任务提出对应的评价指标。①语义分割: 均交并集 mIoU, 计算所有类别交集和并集之比的平均值。②全景分割: 均交并集 mIoU; 全景质量 PQ, 关注每个实例的分割质量; 分割质量 SQ, 表示匹配后预测与标注分割的平均分割精度; 识别质量 RQ, 即 F_1 分数, 用来计算全景分割中每个实例物体识别的准确性; 平均绝对角度误差 MAAE, 用于实例方向评估。③实例分割: 全景质量 PQ=SQ·RQ; 平均绝对角度误差 MAAE。④方向估计: 为评估实例方向, 采用类似位姿估计[23] 的平均绝对角度误差 MAAE, 与其他指标相比, 越低越好, 最大误差为 180。参数量是衡量模型复杂度的重要指标。

3) 参数设置。实验在 Intel Xeon E5-2640v4 CPU、64 GB RAM 和单个 NVIDIA RTX3090Ti GPU 上进行。实验环境为 PyTorch1.10.1, CUDA V10.1, Python 3.8。采用 SGD 优化器, 动量设置为 0.9, 权重衰减 0.001; 迭代 500 个 epoch, 训练批次为 8, 学习率通过网格搜索确定, 最终 NYUv2 的学习率设置为 0.04, SUN RGB-D 学习率设置为 0.005。不同数据集输入图像的分辨率不同, NYUv2 的分辨率为 640×480, SUN RGB-D 的分辨率为 730×530。

3.2　消融实验

消融实验主要涉及 2 个问题: ①验证本文提出的注意力融合模块的有效性, 将 SENet[8]、CBAM[9]、NonLocal[10] 方法及本文设计的注意力融合模块进行对比实验; ②探讨上下文位置对本文多任务学习框架的影响。

表 2 中, 相较于基准方法, 方法 3 在多方面性能上都实现了提升, 实例分割中, PQ 增加 0.70%, MAAE 增加 2.05%, 全景分割中, mIoU 增加 0.75%, PQ 增加 2.62%, RQ 增加 1.30%, SQ 减少 0.13%, MAAE 减少 5.29%, 语义分割中, mIoU 减少 0.38%。这说明注意力融合模块能够提升实例分割和全景分割的性能, 且在综合性全景分割任务上, 方法 3 的表现仅次于方法 2。

因此, 方法 5 和本文方法旨在进一步研究方法 2 和方法 3 之间的差异, 并探索上下文位置对整体网

表2 注意力融合方式和上下文位置的消融实验结果

Table 2 Ablation experiment results on attention fusion mode and context position %

方法	融合方式	上下文位置	实例分割		全景分割					语义分割
			PQ	MAAE	mIoU	PQ	RQ	SQ	MAAE	mIoU
基准	2×SENet[8]	解码器前	60.99	17.59*	49.31	42.30	50.73	82.32*	16.83	50.16
方法1	2×CBAM[9]	解码器前	60.71	17.71	50.27	42.70	51.30	82.17	15.79*	50.35
方法2	2×NonLocal[10]	解码器前	61.75*	18.08	50.52*	42.98	51.75*	82.13	16.07	50.61*
方法3	SENet/Excite	解码器前	61.42	17.95	49.68	43.41*	51.39	82.21	15.94	49.97
方法4	2×SENet[8]	解码器后	61.61	18.33	49.19	42.79	51.48	82.68	16.10	49.76
方法5	2×NonLocal[10]	解码器后	61.02	19.11	48.42	41.85	50.19	82.31	16.52	48.45
本文方法	SENet/Excite	解码器后	61.88	18.45	49.36	43.54	52.08	82.60	15.76	49.75

注：“*”标志的数据表示注意力消融实验的最佳结果；SENet表示通道注意力模块，Excite表示空间注意力模块。

络性能的影响。调整上下文模块位置后，本文方法和方法4的精度都显著提高，并在与其他对比方法的比较中，本文方法表现最优，充分说明上下文位置对本文多任务学习框架的积极作用。此外，本文推理模型的参数量也由64.24×10^6减少至63.50×10^6，具有一定减少计算负担的效果。方法5则充分验证了模型架构的调整对于任务和数据的适应性不同，导致在不同的设置下表现出差异。

3.3 结果分析

1）任务组合。将ResNet34-NBt1D作为骨干网络，先进行单任务实验，得到语义分割的mIoU为49.10%，学习率0.04；实例分割的PQ为61.39%，学习率0.08；方向估计的MAAE为18.06%，学习率0.02。

根据语义分割、实例分割、方向估计的重要性进行手动权重分配，按照先两两任务组合，再增加一个任务的训练顺序，探索各任务之间的关系，并通过不断试验确定最终权重。

由表3可知，在NYUv2数据集上，语义分割+实例分割+方向估计3个任务的学习方法在其他组合任务中取得了最优越的性能，相比单个任务，语义分割的mIoU从49.10%提升至49.75%，实例分割的PQ提高了0.80%。图4展示了本文方法与基准方法在NYUv2数据集的比较结果，可以很明显观察到本文方法对于场景的理解更细致，框表示分割改进的地方。

从表3、表4的结果可知，单个任务到4个任务的过程中，并不是任务越多，网络的性能就越好。原因可能是：虽然语义分割本身包含分类任务，但没有专门用于分类任务的分类器，这样的分类任务与像素级信息相关性不大，反而拖累了综合网络的性能。

表4比较了基于SUN RGB-D数据集的3种任务和4种任务网络的性能，实验分析表明，本文方法在全景分割这样的综合任务上表现更佳。

2）任务权重。实例中心与实例偏移的初始权重比$b=2$。一方面，任务程度越重要，获得的权重系数越大，如在NYUv2数据集中，语义分割、实例分割、方向估计3个任务的权重分别为1、3、1，即实例分割损失权重$a=3$，方向估计损失权重$c=3$。另一方面，不同的数据集导致任务权重发生变化，如在SUN RGB-D数据集中，语义分割、实例分割、方向估计3个任务的权重变为1、2、0.5，即实例分割损失权重$a=2$，方向估计损失权重$c=0.5$，这可能使以上提及的2个数据集样本类别数量存在差异。

3）与先进方法比较。除了对任务数量和其中的相关性展开讨论，还就语义分割和全景分割任务与最先进的方法作比较。如表5所示，本文方法在全景分割的PQ上取得最优，为43.54%。在同等骨干网络和模态下，语义分割的mIoU仅次于EMSA-Net，明显优于其他深层骨干网络，这可能是因为本

表3 在NYUv2数据集中验证不同任务的组合效果

Table 3 Combination effect of different tasks verified on NYUv2 dataset

任务	任务权重	学习率	实例分割		全景分割					语义分割
			PQ/%	MAAE/%	mIoU/%	PQ/%	RQ/%	SQ/%	MAAE/%	mIoU/%
语义分割+实例分割	1∶3	0.03	61.81		48.56	42.98	51.33	82.43		48.77
实例分割+方向估计	3∶1	0.04	59.72	17.66						
语义分割+实例分割+方向估计(本文)	1∶3∶1	0.04	61.88	18.45	49.36	43.54	52.08	82.60	15.76	49.75
语义分割+实例分割+方向估计+分类	1∶0.25∶3∶1	0.04	60.14	18.64	49.63	42.02	50.34	82.45	16.70	49.33

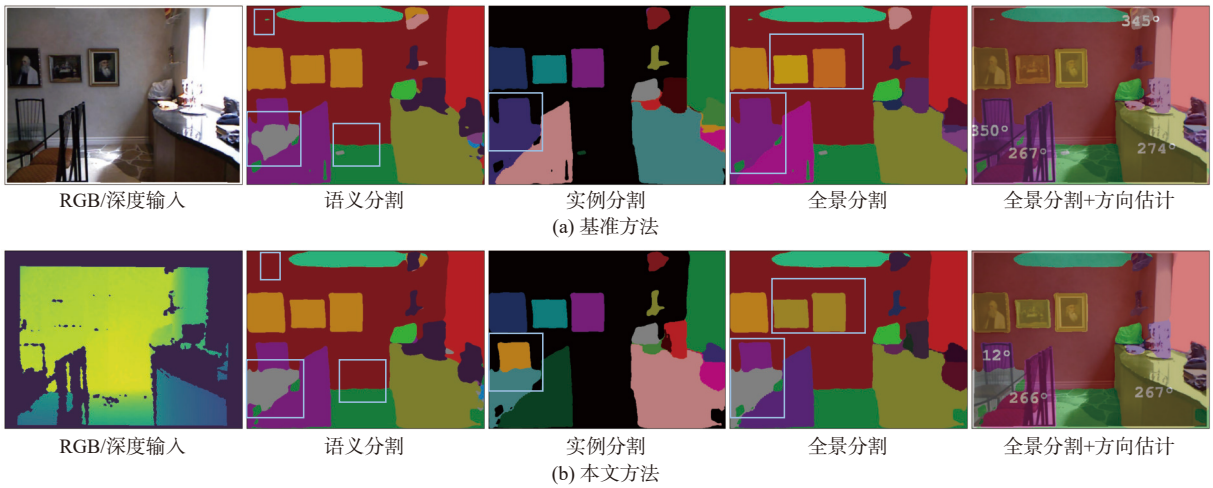


图 4　基准方法与本文方法在 NYUv2 数据集上的分割结果比较

Fig. 4　Comparison of segmentation results between benchmark method and the proposed method on NYUv2 dataset

表 4　在 SUN RGB-D 数据集中验证不同任务的组合效果

Table 4　Combination effect of different tasks verified on SUN RGB-D dataset

任务	任务权重	学习率	实例分割		全景分割					语义分割
			PQ/%	MAAE/%	mIoU/%	PQ/%	RQ/%	SQ/%	MAAE/%	mIoU/%
语义分割+实例分割+方向估计(基准)	1:2:0.5	0.005	60.64	16.29	45.51	48.46	55.90	83.08	14.10	47.79
语义分割+实例分割+方向估计(本文)	1:2:0.5	0.005	61.17	16.01	45.00	50.82	58.96	85.05	14.66	47.75
语义分割+实例分割+方向估计+分类	1:0.25:2:0.5	0.005	61.00	16.00	44.91	50.81	58.88	84.87	13.76	47.96

表 5　在 NYUv2 数据集上与最先进方法的比较

Table 5　Comparison with state-of-the-art methods on NYUv2 dataset　%

方法	骨干网络	模态	mIoU	PQ
MTI-Net[13]	HRNet48	RGB	49.00	
RefineNet[14]	ResNet152	RGB	47.60	
3DGNN[15]	ResNet101	RGB-D	48.60	
SGNet[16]	ResNet101	RGB-D	49.00	
Link-RGBD[26]	ResNet50	RGB-D	49.50	
SGACNet[27]	ResNet34-NBt1D	RGB-D	49.40	
Panoptic-DeepLab[20]	ResNet50	RGB	39.42	30.99
	ResNet101	RGB	42.55	35.32
EMSANet[31]	ASMLP-S	RGB-D	36.02	28.52
	ResNet18-NBt1D	RGB-D	47.35	40.04
	ResNet34-NBt1D	RGB	44.16	36.38
	ResNet34-NBt1D	RGB-D	50.16	42.30
本文方法	ResNet34-NBt1D	RGB-D	49.75	43.54

表 6　在 SUN RGB-D 数据集上与最先进方法的比较

Table 6　Comparison with state-of-the-art methods on SUN RGB-D dataset

方法	骨干网络	模态	mIoU/%	参数量
RefineNet[14]	ResNet152	RGB	45.90	129.50×10^6
Link-RGBD[26]	ResNet50	RGB-D	48.40	116.00×10^6
3DGNN[15]	ResNet101	RGB-D	45.90	
SGNet[16]	ResNet101	RGB-D	47.10	64.70×10^6
ESANet[32]	ResNet34-NBt1D	RGB-D	47.46	46.95×10^6
SGACNet[27]	ResNet34-NBt1D	RGB-D	46.73	35.65×10^6
EMSANet[31]	ResNet34-NBt1D	RGB	45.41	49.37×10^6
	ResNet34-NBt1D	RGB-D	47.96	64.24×10^6
本文方法	ResNet34-NBt1D	RGB-D	47.75	63.50×10^6

文方法更偏重于多尺度全局信息的获取, 对获取方向估计、语义分割任务所需局部特征的能力有限。有关室内方向估计的任务工作较少, 未做比较。

从表 5、表 6 的输入模态来看, RGB-D 模态比单 RGB 模态的性能更好, 说明深度信息能够提供更多三维场景信息。此外, 对在 SUN RGB-D 数据集上的分割精度和计算复杂度进行分析, 具体结果如表 6 所示。相较于分割性能优越但计算复杂度大的 Link RGBD[26] 和计算复杂度低但分割效果较差的 SGACNet[27], 本文方法在性能和参数量之间取得了一种平衡, 具备了更好的实用性。

4　结　论

本文提出了一种能实现语义分割、实例分割、方向估计的多任务室内场景理解方法, 并通过对注意力融合方式、上下文位置等改进与其他方法进行对比, 根据任务组合表现, 得出以下结论:

1) 网络通过通道-空间注意力融合模块能有效增强深度信息, 放在语义分支的上下文模块专注提供更多像素的语义信息, 减少冗余计算。

2）多任务室内场景理解方法在综合性全景分割任务上表现优异，证实了并不是任务越多，网络的性能就越好，本文提出的3项任务结合能最优把握全局场景信息，并在较浅的网络深度下，比其他改进方法的全景分割精度更高，在性能和参数量之间取得了一种平衡。

3）在扩展标签注释的室内数据集NYUv2和SUN RGB-D上评估本文方法的有效性，综合性全景分割精度分别提高了2.93%和4.87%，证明多任务方法只有相关任务才能互相受益。

参考文献（References）

[1] 马素刚, 张子贤, 蒲磊, 等. 结合空间注意力机制的实时鲁棒视觉跟踪[J]. 北京航空航天大学学报, 2024, 50(2): 419-432.
MA S G, ZHANG Z X, PU L, et al. Real-time robust visual tracking based on spatial attention mechanism[J]. Journal of Beijing University of Aeronautics and Astronautics, 2024, 50(2): 419-432(in Chinese).

[2] WENGEFELD T, SEICHTER D, LEWANDOWSKI B, et al. Enhancing person perception for mobile robotics by real-time RGB-D person attribute estimation[C]//Proceedings of the IEEE/SICE International Symposium on System Integration. Piscataway: IEEE Press, 2024: 914-921.

[3] HU J, SHEN L, SUN G. Squeeze-and-excitation networks[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2018: 7132-7141.

[4] HE Y, XIAO L, SUN Z G, et al. Bimodal feature propagation and fusion for real-time semantic segmentation on RGB-D images[C]//Proceedings of the 7th International Conference on Intelligent Computing and Signal Processing. Piscataway: IEEE Press, 2022: 1897-1902.

[5] ISLAM M A, ROCHAN M, BRUCE N D B, et al. Gated feedback refinement network for dense image labeling[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2017: 4877-4885.

[6] QUAN T M, HILDEBRAND D G C, JEONG W K. FusionNet: a deep fully residual convolutional neural network for image segmentation in connectomics[J]. Frontiers in Computer Science, 2021, 3: 613981.

[7] FU J, LIU J, TIAN H J, et al. Dual attention network for scene segmentation[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2019: 3141-3149.

[8] ZHANG R F, LI G B, LI Z, et al. Adaptive context selection for polyp segmentation[C]//Proceedings of the Medical Image Computing and Computer Assisted Intervention. Berlin: Springer, 2020: 253-262.

[9] YANG C L, ZHANG C C, YANG X Q, et al. Performance study of CBAM attention mechanism in convolutional neural networks at different depths[C]//Proceedings of the IEEE 18th Conference on Industrial Electronics and Applications. Piscataway: IEEE Press, 2023: 1373-1377.

[10] ZOU W B, PENG Y Q, ZHANG Z Y, et al. RGB-D gate-guided edge distillation for indoor semantic segmentation[J]. Multimedia Tools and Applications, 2022, 81(25): 35815-35830.

[11] ZHAO H S, QI X J, SHEN X Y, et al. ICNet for real-time semantic segmentation on high-resolution images[C]//Proceedings of the European Conference on Computer Vision. Berlin: Springer, 2018: 418-434.

[12] GE W X, YANG X B, JIANG R, et al. CD-CTFM: a lightweight CNN-transformer network for remote sensing cloud detection fusing multiscale features[J]. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2024, 17: 4538-4551.

[13] VANDENHENDE S, GEORGOULIS S, VAN GOOL L. MTI-Net: multi-scale task interaction networks for multi-task learning[C]//Proceedings of the European Conference on Computer Vision. Berlin: Springer, 2020: 527-543.

[14] LIN G S, LIU F Y, MILAN A, et al. RefineNet: multi-path refinement networks for dense prediction[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020, 42(5): 1228-1242.

[15] CAO J M, LENG H C, COHEN-OR D, et al. RGB × D: learning depth-weighted RGB patches for RGB-D indoor semantic segmentation[J]. Neurocomputing, 2021, 462: 568-580.

[16] CHEN L Z, LIN Z, WANG Z Q, et al. Spatial information guided convolution for real-time RGBD semantic segmentation[J]. IEEE Transactions on Image Processing, 2021, 30: 2313-2324.

[17] BORSE S, PARK H, CAI H, et al. Panoptic, instance and semantic relations: a relational context encoder to enhance panoptic segmentation[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2022: 1259-1269.

[18] XIONG Y W, LIAO R J, ZHAO H S, et al. A unified panoptic segmentation network[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2019: 8810-8818.

[19] MOHAN R, VALADA A. EfficientPS: efficient panoptic segmentation[J]. International Journal of Computer Vision, 2021, 129(5): 1551-1579.

[20] CHENG B W, COLLINS M D, ZHU Y K, et al. Panoptic-DeepLab: a simple, strong, and fast baseline for bottom-up panoptic segmentation[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2020: 12472-12482.

[21] ZOU Q, DU X Z, LIU Y Z, et al. Dynamic path planning and motion control of microrobotic swarms for mobile target tracking[J]. IEEE Transactions on Automation Science and Engineering, 2023, 20(4): 2454-2468.

[22] MAHFOUDI M N, TURLETTI T, PARMENTELAT T, et al. ORION: orientation estimation using commodity Wi-Fi[C]//Proceedings of the IEEE International Conference on Communications Workshops. Piscataway: IEEE Press, 2017: 1233-1238.

[23] CHRISTIE G, ABUJDER R R R M, FOSTER K, et al. Learning geocentric object pose in oblique monocular images[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2020: 14500-14508.

[24] TANG H Y, LIU J N, ZHAO M, et al. Progressive layered extraction (PLE): a novel multi-task learning (MTL) model for personalized recommendations[C]//Proceedings of the 14th ACM Confer-

ence on Recommender Systems. New York: ACM, 2020: 269-278.

[25] LIU X D, HE P C, CHEN W Z, et al. Multi-task deep neural networks for natural language understanding[C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: ACL, 2019: 4487-4496.

[26] WU P, GUO R Z, TONG X Z, et al. Link-RGBD: cross-guided feature fusion network for RGBD semantic segmentation[J]. IEEE Sensors Journal, 2022, 22(24): 24161-24175.

[27] ZHANG Y, XIONG C Y, LIU J J, et al. Spatial information-guided adaptive context-aware network for efficient RGB-D semantic segmentation[J]. IEEE Sensors Journal, 2023, 23(19): 23512-23521.

[28] HUANG L, GUO H. Research on multi-task learning method based on causal features[C]//Proceedings of the International Conference on Image Processing, Computer Vision and Machine Learning. Piscataway: IEEE Press, 2023: 924-927.

[29] BEYER L, HERMANS A, LEIBE B. Biternion Nets: continuous head pose regression from discrete training labels[J]. Pattern Recognition, 2015, 9358: 157-168.

[30] SILBERMAN N, HOIEM D, KOHLI P, et al. Indoor segmentation and support inference from RGBD images[C]//Proceedings of the European Conference on Computer Vision. Berlin: Springer, 2012: 746-760.

[31] SEICHTER D, FISCHEDICK S B, KÖHLER M, et al. Efficient multi-task RGB-D scene analysis for indoor environments[C]//Proceedings of the International Joint Conference on Neural Networks. Piscataway: IEEE Press, 2022: 1-10.

[32] SONG S R, LICHTENBERG S P, XIAO J X. SUN RGB-D: a RGB-D scene understanding benchmark suite[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2015: 567-576.

Spatial information-enhanced indoor multi-task RGB-D scene understanding

SUN Guodong，XIONG Chenyun，LIU Junjie，ZHANG Yang*

(School of Mechanical Engineering，Hubei University of Technology，Wuhan 430068，China)

Abstract: To explore 3D space, mobile robots need to obtain a large amount of scene information, which includes semantic, instance objects, and positional relationships. The accuracy and computational complexity of scene analysis are the two focuses of mobile terminals. Therefore, a spatial information-enhanced multi-task learning method for indoor scene understanding was proposed. This method consists of an encoder with a channel-spatial attention fusion module and a decoder with multi-task heads for semantic segmentation, panoptic segmentation (instance), and orientation estimation. The channel-spatial attention fusion module aims to enhance the modal characteristics of RGB and depth, and the spatial attention mechanism, composed of simple convolutions, can reduce the convergence speed. After fusing with the channel attention mechanism, it further strengthens the position features of global information. The context module of the semantic branch is located after the decoder, providing strong support for pixel-level semantic classification and helping to reduce the model size. A loss function based on hard parameter sharing was designed, enabling balanced training tasks. The influence of an appropriate lightweight backbone network and the number of tasks on improving the performance of scene understanding was discussed. Finally, on the NYUv2 and SUN RGB-D indoor datasets with newly added label annotations, the effectiveness of the proposed multi-task learning method was evaluated. Results show that the comprehensive panoramic segmentation accuracy is improved by 2.93% and 4.87%, respectively.

Keywords: scene understanding; multi-task learning; RGB-D; spatial information; indoor scene

Received：2023-06-19；**Accepted**：2024-03-08；**Published Online**：2025-05-29 14：40

URL：link.cnki.net/urlid/11.2625.V.20250529.0856.001

Foundation items：National Natural Science Foundation of China (51775177); Hubei Province Science and Technology Project of Open Bidding for Selecting the Best Candidates (2024BEB018)

* **Corresponding author**. E-mail：yzhangcst@hbut.edu.cn

http://bhxb.buaa.edu.cn jbuaa@buaa.edu.cn
DOI: 10.13700/j.bh.1001-5965.2023.0417

基于坐标感知注意的多帧自监督单目深度估计

程德强[1]，范舒铭[1]，钱建生[1]，江鹤[1]，寇旗旗[2,*]

(1. 中国矿业大学 信息与控制工程学院，徐州 221000；2. 中国矿业大学 计算机科学与技术学院，徐州 221000)

摘 要：为解决单目深度估计方法中物体细节边缘深度预测模糊不清的问题，提出了一种基于坐标感知注意的多帧自监督单目深度估计方法。提出了一种坐标感知注意模块，以增强编码器最下层输出特征，并加强成本体的特征利用；提出了一种基于像素洗牌的深度预测解码器，可有效分离低分辨编码器特征中的多物体融合特征，以细化深度估计结果中的物体边缘。在KITTI和Cityscapes数据集上的实验测试结果表明：所提方法优于目前主流方法，显著提升了主观视觉效果和客观评价指标，尤其在物体边缘细节上具有更好的深度估计性能。

关 键 词：图像处理；深度学习；深度估计；自监督学习；注意力机制

中图分类号：TP391

文献标志码：A **文章编号**：1001-5965(2025)07-2218-11

单目深度估计是计算机视觉领域的一个基础课题和热门方向，已经在三维重建[1]、虚拟现实、环境感知、无人机探测、自动驾驶等任务上有广泛应用。在有监督的单目深度估计中，可以利用传感器采集到的图像信息中像素的现实真实距离作为训练的真实值，但采集设备成本高昂。自监督的单目深度估计可以在不使用真实深度作为参考的情况下，通过训练使网络估计出图像中像素的深度，成为当前深度估计研究领域中的热点。

Zhou等[2]提出自监督单目深度估计方法，以视频序列中连续帧作为训练数据，联合训练深度和位姿网络，来预测目标图像的对应深度图，该方法中的深度预测网络的输入为单帧图像，但其使用相机位姿运动和图像深度联合构造监督信息，受到物体遮挡、运动及弱纹理区域等因素的影响。Godard等[3]使用最小重投影损失减少物体遮挡影响，并构造了一种自动掩模来处理场景中与相机运动相对静止的物体。

Watson等[4]提出了多帧自监督单目深度估计方法(其中，深度预测网络的输入为多帧图像)，并利用多帧图像和相机运动构造成本体来优化目标图像的深度估计，但成本体作为多帧图像构造出的融合特征也受到多帧图像间运动物体、弱纹理区域等因素的影响，为解决该问题，使用额外的自监督单帧深度预测网络来规范成本体中特征的利用。Guizilini等[5]提出了使用交叉注意力及纹理自适应层来引导成本体特征训练的方法。

上述方法极大促进了自监督单目深度估计的发展，但生成的深度图仍存在很多问题。这些方法都采用基于卷积神经网络的编码器-解码器架构，虽然在编码器中使用多层下采样提取到的特征可以有效建立一张图像中不同位置像素点间的联系，但对于深度估计，如何通过解码器有效地解码这些特征又成为新的挑战。下层编码器分辨率小，不同物体间的特征融合度高，尤其是前景物体和背景间的实际深度差值较大的情况，在解码器中很难有效区分，解码困难，导致最终预测出的深度图中物体边缘模糊。

针对上述问题，本文提出一种坐标感知注意模

收稿日期：2023-06-29；**录用日期**：2023-11-10；**网络出版时间**：2023-11-16 11:36
网络出版地址：link.cnki.net/urlid/11.2625.V.20231115.1718.002
基金项目：国家自然科学基金(52204177,52304182)；中央高校基本科研业务费专项资金(2020QN49)
*通信作者. E-mail：kouqiqi@cumt.edu.cn

引用格式：程德强，范舒铭，钱建生，等. 基于坐标感知注意的多帧自监督单目深度估计[J]. 北京航空航天大学学报，2025，51(7)：2218-2228.
CHENG D Q，FAN S M，QIAN J S，et al. Coordinate-aware attention-based multi-frame self-supervised monocular depth estimation[J]. Journal of Beijing University of Aeronautics and Astronautics，2025，51(7)：2218-2228 (in Chinese).

块, 专注于低分辨率编码器特征增强及成本体的特征提取增强; 并提出一种基于像素洗牌[6]的解码器, 旨在以通道维度分担低分辨率下同一场景中不同物体的融合特征, 以达到更好的解码效果。本文提出基于坐标感知注意的多帧自监督单目深度估计 (coodinate-aware attention-based multi-frame self-supervised monocular depth estimation, CAMD) 网络, 通过实验比较与结果分析, 在 KITTI 数据集[7]和 Cityscapes 数据集上验证了 CAMD 网络结构及注意模块的有效性。

1　多帧自监督单目深度估计方法

多帧自监督单目深度估计模型训练过程中, 利用源图像和预测深度目标图得到目标图像的重建图像, 通过最小化目标图像与重建图像之间的差异来优化模型。多帧自监督单目深度估计框架采用联合深度预测网络和位姿预测网络的方式进行训练。其中, 位姿预测网络仅用于预测相邻帧间的位姿变化。深度预测网络采用传统编解码器架构, 该架构在编解码的过程中会丢失大量的细节信息, 无法充分利用高层次特征和低层次特征, 导致在一些复杂纹理区域及远景区域上产生错误的深度估计。本文在文献 [4] 的基础上改进了深度预测网络的整体架构, 提出了 CAMD 网络框架, 其需要联合深度预测网络、位姿预测网络进行训练。整体网络框架如图 1 所示。

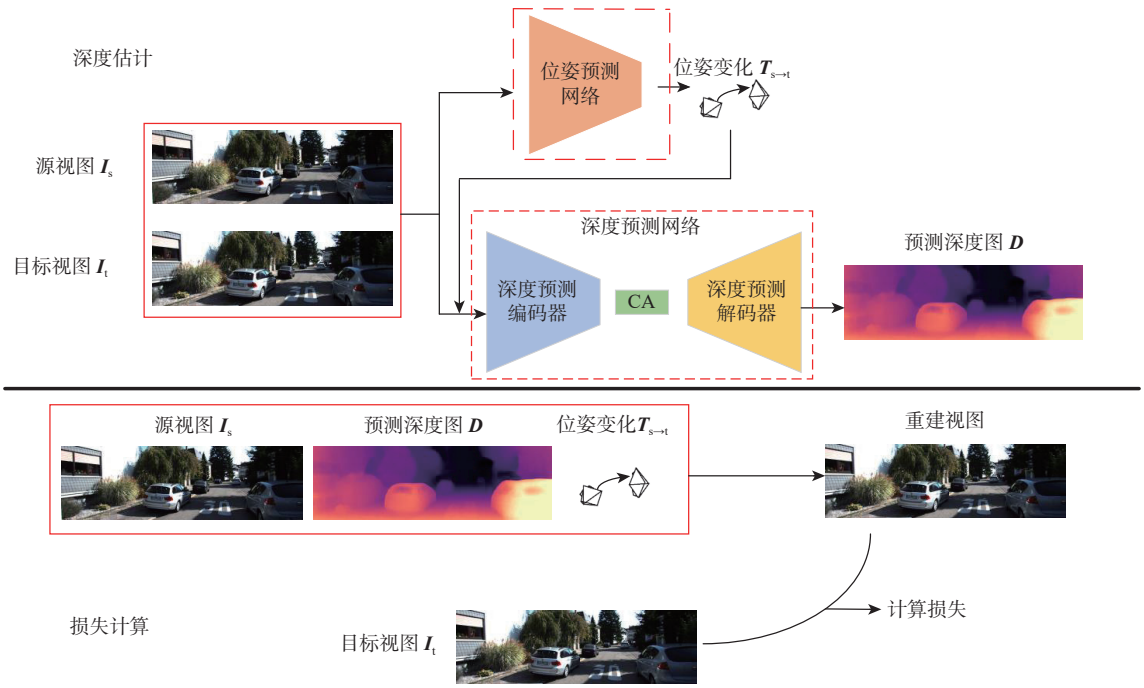


图 1　多帧自监督单目深度估计方法框架

Fig. 1　Framework of multi-frame self-supervised monocular depth estimation method

不同于单帧自监督单目深度估计, 多帧自监督单目深度估计框架中的深度预测网络的输入为连续的 2 帧图像。如图 2 所示, 该深度预测网络对于连续帧的处理主要为生成一个成本体, 以此作为 2 帧的融合特征参与网络训练。

关于成本体 $\boldsymbol{C}_{\mathrm{V}}\in\mathbf{R}^{H/4\times W/4\times N_d}$ 的计算, 假定有 N_d 个深度图 $\boldsymbol{D}_i\in\mathbf{R}^{H/4\times W/4\times 1}$, $\boldsymbol{D}_i$ 中所有点的深度值均等于 d_i, 其中, H、W 分别为高、宽。对于给定的最小深度值 $d_{\min}$ 和最大深度值 $d_{\max}$, N_d 个离散深度 d_i 的计算公式为

$$d_i=d_{\min}+(i-1)\frac{d_{\max}-d_{\min}}{N_d} \tag{1}$$

对于目标视图 $\boldsymbol{I}_{\mathrm{t}}\in\mathbf{R}^{H\times W\times 3}$ 与源视图 $\boldsymbol{I}_{\mathrm{s}}\in\mathbf{R}^{H\times W\times 3}$, 输入位姿预测网络中得到相机移动位姿 $\boldsymbol{T}_{\mathrm{t}\to\mathrm{s}}$; 再输入深度预测网络, 各自经过同一个特征层提取得到目标视图特征 $\boldsymbol{F}_{\mathrm{t}}\in\mathbf{R}^{H/4\times W/4\times C_{\mathrm{f}}}$ 与源视图特征 $\boldsymbol{F}_{\mathrm{s}}\in\mathbf{R}^{H/4\times W/4\times C_{\mathrm{f}}}$ (C_{f} 为特征的通道维度)。在每个离散深度图 $\boldsymbol{D}_i$ 条件下, 源视图特征可向目标视图投影, 得到对应的重建特征 $\boldsymbol{F}^i_{\mathrm{s}\to\mathrm{t}}\in\mathbf{R}^{H/4\times W/4\times C_{\mathrm{f}}}$。计算公式为

$$\boldsymbol{F}^i_{\mathrm{s}\to\mathrm{t}}=\boldsymbol{F}_{\mathrm{t}}\langle\mathrm{proj}(\boldsymbol{D}_i,\boldsymbol{T}_{\mathrm{t}\to\mathrm{s}},\boldsymbol{K})\rangle \tag{2}$$

式中: $\boldsymbol{K}\in\mathbf{R}^{3\times 3}$ 为相机的内参矩阵; proj( )为源视图向目标视图投影的二维坐标图, 可以由深度图、相机移动位姿、相机内参计算得到; $\langle\rangle$ 表示采样操作, 本文使用可微的双线性插值采样。

因此, 成本体 $\boldsymbol{C}_{\mathrm{V}}\in\mathbf{R}^{H/4\times W/4\times N_d}$ 中每个特征层

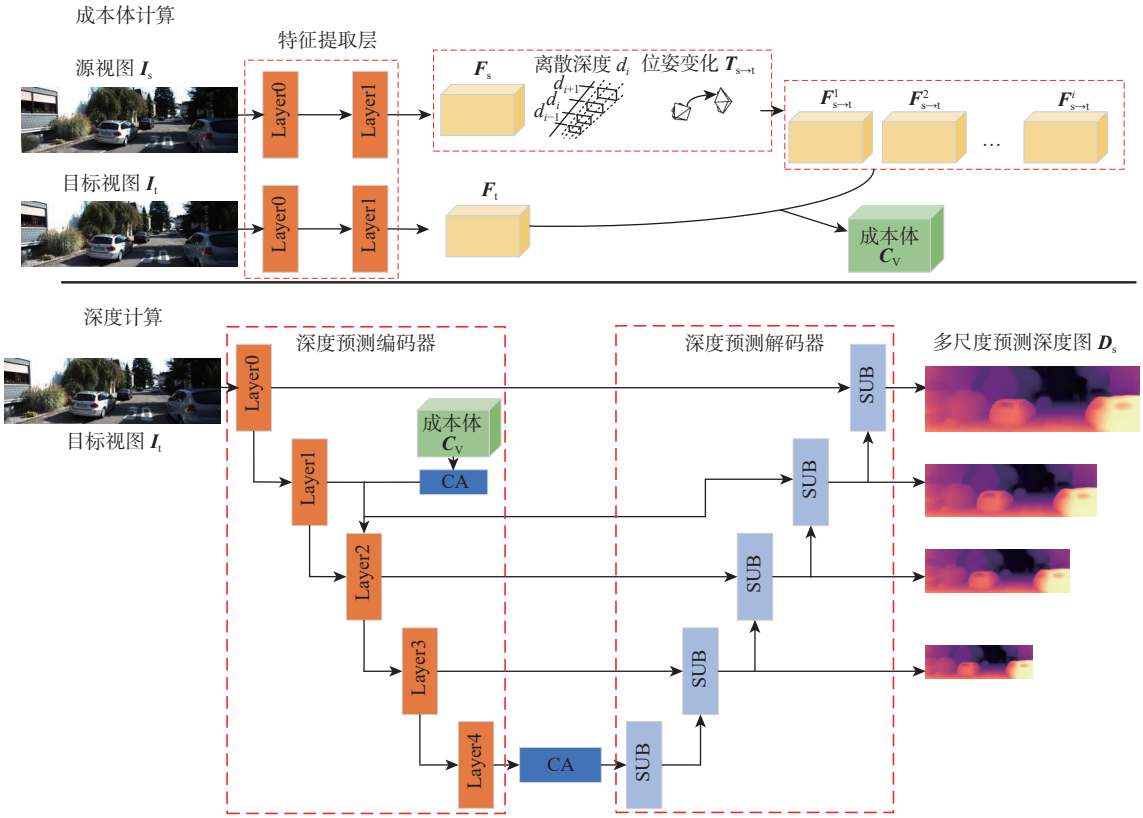


图2 多尺度空间注意力模块

Fig. 2 Multi-scale spatial attention module

$C_{Vi}\in\mathbf{R}^{H/4\times W/4\times 1}$可以表示为某一离散深度下，重建特征$\boldsymbol{F}^i_{s\to t}$与目标视图特征$\boldsymbol{F}_t$间的$L_1$范数后，在通道维度上的平均。

对于通道数为C_f的特征$\boldsymbol{F}_t(c)$与$\boldsymbol{F}^i_{s\to t}(c)$，公式为

$$C_{Vi}=\frac{1}{C_f}\sum_{c=1}^{C_f}(|\boldsymbol{F}_t(c)-\boldsymbol{F}^i_{s\to t}(c)|) \tag{3}$$

式中：c为通道维度，数值上等于N_d。

集合$\{C_{V1},C_{V2},\cdots,C_{VN_d}\}$可依次在通道维度上经拼接操作得到$\boldsymbol{C}_V\in\mathbf{R}^{N_d\times H/4\times W/4}$。

在深度预测网络得到目标视图$\boldsymbol{I}_t$的对应深度图$\boldsymbol{D}_t$后，与式(2)原理相同，可以得到由源视图向目标视图的重建图$\boldsymbol{I}_{s\to t}$：

$$\boldsymbol{I}_{s\to t}=\boldsymbol{I}_t\langle\mathrm{proj}(\boldsymbol{D}_t,\boldsymbol{T}_{t\to s},\boldsymbol{K})\rangle \tag{4}$$

理想状态下，多帧自监督单目深度估计框架中的重建视图$\boldsymbol{I}_{s\to t}$应与原目标视图$\boldsymbol{I}_t$一致，联合结构相似性指标(structural similarity, SSIM)结合L_1范数作为光度损失L_p：

$$L_p=\alpha\frac{1-\mathrm{SSIM}(\boldsymbol{I}_t,\boldsymbol{I}_{s\to t})}{2}+(1-\alpha)|\boldsymbol{I}_t-\boldsymbol{I}_{s\to t}| \tag{5}$$

式中：α为权重。

为处理相对静止物体等问题，应用自动掩码方法过滤掉在视频帧间保持运动的物体，使用最小重投影处理遮挡问题。同时，采用边缘平滑损失L_{sm}对低梯度区域的深度进行正则化，以平滑预测深度。

$$L_{sm}=|\partial_x d_t^*|\mathrm{e}^{-|\partial_x \boldsymbol{I}_t|}+|\partial_y d_t^*|\mathrm{e}^{-|\partial_y \boldsymbol{I}_t|} \tag{6}$$

式中：∂_x和∂_y分别为水平方向和竖直方向的梯度；$d_t^*=d_t/\bar{d}$为平均归一化深度，d_t和$\bar{d}$分别为深度和平均深度。

由于相邻帧之间可能存在运动物体和远景中的无纹理区域，而导致相应区域的深度表示在成本体中出现异常。为过滤成本体中深度表示的异常值，加强其特征利用，使用额外的教师深度预测网络[4]来约束其训练。最小化多帧深度预测网络预测的深度图$\boldsymbol{D}\in\mathbf{R}^{H\times W\times 1}$应与教师深度预测网络预测的深度图$\boldsymbol{D}^*\in\mathbf{R}^{H\times W\times 1}$的$L_1$损失一致。其一致性损失$L_t$表达为

$$L_t=|\boldsymbol{D}-\boldsymbol{D}^*| \tag{7}$$

教师深度预测网络为一个单独的自监督单目深度预测网络，输入为单帧图像，与多帧深度预测网络共享同一个位姿预测网络，在训练过程中同步参与训练，因此，具有对应的光度重建损失L_p^*与边缘平滑损失L_s^*。

总损失函数为

$$L=ML_p+\beta L_{sm}+(1-M)L_t+L_p^*+\beta L_s^* \tag{8}$$

式中：β为边缘平滑损失的权重系数；$M\in\mathbf{R}^{H\times W\times 1}$为

异常深度像素掩模, 使用教师深度预测网络生成的深度图$\boldsymbol{D}^*$与成本体所计算的粗糙深度图$\boldsymbol{D}_{\rm cv}\in\mathbf{R}^{H\times W\times 1}$计算得到, 取值为 0 或 1。

粗糙深度图$\boldsymbol{D}_{\rm cv}$的计算公式为

$$\boldsymbol{D}_{\rm cv}=\text{upsample}\left(\arg\min_c(\boldsymbol{C}_{\rm V})\frac{d_{\max}-d_{\min}}{N_d}\right) \tag{9}$$

式中: upsample( )表示双线性插值上采样操作。

异常深度像素掩模M的计算公式为

$$M=\max\left(\frac{\boldsymbol{D}^*-\boldsymbol{D}_{\rm cv}}{\boldsymbol{D}_{\rm cv}},\frac{\boldsymbol{D}_{\rm cv}-\boldsymbol{D}^*}{\boldsymbol{D}}\right)<1 \tag{10}$$

在M中, 异常深度像素点位置的值为 0, 其余为 1。

因此, 对于多帧深度预测网络, 在正常深度点处, 计算光度损失, 总损失为$ML_{\rm p}$; 在异常深度点处, 计算一致性损失, 总损失为$(1-M)L_{\rm t}$。

2　CAMD 网络结构

现有许多深度预测网络都建立在 ResNet18[8] 作为主干网络的基础上。为保证编码器网络提取到的特征有足够多的通道数适配本文提出的解码器, 编码器使用 Densenet121[9], 并提取高通道数的多尺度特征。相对于 ResNet18, Densenet121 具有训练参数少、提取到的多尺度特征细节更多、表现力更强等特点。其中, 越高的层次特征具有更强的语义信息, 但分辨率较低, 原分辨率下各种物体的空间布局, 尤其是背景与前景, 在低分辨率中混合在一起, 对于深度估计来说, 不利于各物体的预测深度分离。因此, 自适应地感知低分辨率下物体的空间布局显得尤为重要。本文针对该问题, 提出了 CAMD 框架。CAMD 采用多帧图像作为输入, 输出其中定义为目标图像的多尺度深度图。

在传统的深度估计中, 解码器的上采样方法使用双线性插值。这种上采样方法不仅计算量大, 而且丢失了许多原本的特征信息。针对该问题, 本文基于像素洗牌提出了一种新的适用于深度估计的解码器。解码器采用像素洗牌作为上采样方法, 加强了特征的空间与通道间的联系, 且减少了计算量。

2.1　坐标感知注意模块

在深度估计中, 预测网络对图像中物体布局的感知对于预测精度的提升有很大帮助。坐标注意力模块[10] 可以利用所提取的特征在x轴方向与y轴方向的相关分布, 以增强特征在空间布局上的表示。

本文基于 Hou 等[10] 的研究提出了一种坐标感知注意模块, 如图 3 所示。

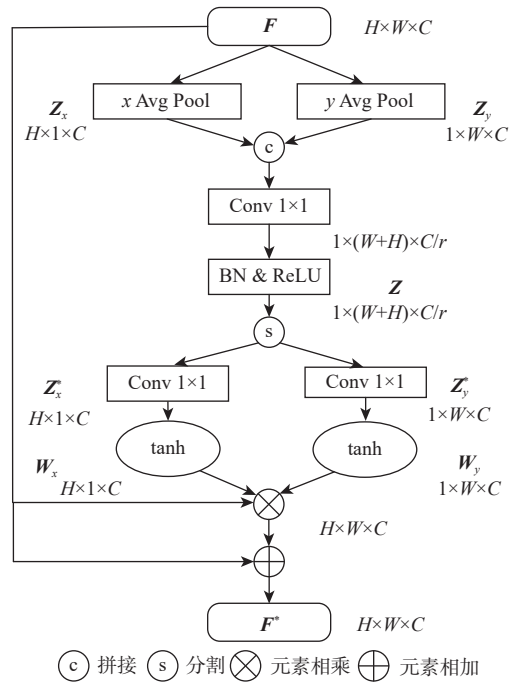


图 3　坐标感知注意模块

Fig. 3　Coordinate-aware attention module

对于给定的编码器特征$\boldsymbol{F}\in\mathbf{R}^{H\times W\times C}$, 以$W$方向为$x$轴, 以$H$方向为$y$轴, 分别通过在$x$轴上的平均池化与$y$轴上的平均池化得到$\boldsymbol{Z}_x\in\mathbf{R}^{H\times 1\times C}$与$\boldsymbol{Z}_y\in\mathbf{R}^{1\times W\times C}$:

$$z_x(h)=\frac{1}{W}\sum_{i=0}^{W}f_i(h) \tag{11}$$

$$z_y(w)=\frac{1}{H}\sum_{i=0}^{H}f_i(w) \tag{12}$$

对于式 (11), h为y轴上的任意坐标, 其取值范围为$[0,H]$。式 (11) 可以解释为: 对于任意y轴上的坐标h, 都在x轴方向进行求和再平均, 得到$z_x(h)$。对于式 (12), w为x轴上的任意坐标, 其取值范围为$[0,W]$。式 (12) 可以解释为: 对于任意x轴上的坐标w, 都在y轴方向上进行求和再平均, 得到$z_y(w)$。

$\boldsymbol{Z}_x$经过维度交换后与$\boldsymbol{Z}_y$拼接, 再依次通过一个卷积核大小为1×1的卷积、BN 层、ReLU 层, 输出中间特征为$\boldsymbol{Z}\in\mathbf{R}^{1\times(W+H)\times C/r}$, 其中, r为中间特征的维度缩放因子, 本文设$r=16$。

$$\boldsymbol{Z}=\text{ReLU}(\text{BN}(\text{Conv}(\text{Cat}(\boldsymbol{Z}_x,\boldsymbol{Z}_y)))) \tag{13}$$

式中: Cat( )表示拼接操作; Conv( )表示卷积层。

中间特征$\boldsymbol{Z}\in\mathbf{R}^{1\times(W+H)\times C/r}$在特定维度上分割为 2 个特征, 并分别经过一个卷积核大小为1×1的卷积, 得到$\boldsymbol{Z}_x^*\in\mathbf{R}^{H\times 1\times C}$与$\boldsymbol{Z}_y^*\in\mathbf{R}^{1\times W\times C}$。

$\boldsymbol{Z}_x^*$与$\boldsymbol{Z}_y^*$经过非线性激活函数得到x轴上的注意力

权重$\boldsymbol{W}_x \in \mathbf{R}^{H\times 1\times C}$与$y$轴上的注意力权重$\boldsymbol{W}_y \in \mathbf{R}^{1\times W\times C}$。nolinear function()为某一非线性激活函数。

$$\boldsymbol{W}_x = \text{nolinear function}(\boldsymbol{Z}_x^*) \tag{14}$$

$$\boldsymbol{W}_y = \text{nolinear function}(\boldsymbol{Z}_y^*) \tag{15}$$

模块输出为$\boldsymbol{F}^* \in \mathbf{R}^{H\times W\times C}$，对于$\boldsymbol{F}^*$上每个$f^*(i,j)$，都可以由$\boldsymbol{F}$上对应的$f(i,j)$、$\boldsymbol{W}_x$中对应的$w_x(i)$、$\boldsymbol{W}_y$中对应的$w_y(j)$计算得出。其中，$i$为$x$轴坐标，$j$为$y$轴坐标。

$$f^*(i,j) = f(i,j)w_x(i)w_y(j) + f(i,j) \tag{16}$$

不同于Hou等[10]的研究，本文选用tanh()而不是Sigmoid()作为式(14)和式(15)的非线性激活函数，降低了该模块的计算复杂度；且引入原特征进行相加，重点在于学习原特征的残差项，使模块可以更快地收敛。经过实验验证，在多帧单目深度估计任务上，引入坐标感知注意模块比直接引入Hou等[10]提出的坐标注意力模块性能更优，在各评价指标上均有提升。如图2所示，本文深度预测网络有2处使用了改进坐标注意力模块，对于网络性能有相应的提升。

2.2 解码器SUB模块

如图2所示，解码器由5个连续的SUB模块组成，除最上层的SUB模块仅有1个下层特征作为输入，其他SUB模块均有2个输入：$\boldsymbol{F}_{\mathrm{d}}^{i}$和$\boldsymbol{F}_{\mathrm{e}}^{i-1}$。$\boldsymbol{D}_i$为输入到SUB模块需要上采样的特征，$\boldsymbol{F}_{\mathrm{e}}^{i-1} \in \{\boldsymbol{F}_{\mathrm{e}}^{1},\boldsymbol{F}_{\mathrm{e}}^{2},\boldsymbol{F}_{\mathrm{e}}^{3},\boldsymbol{F}_{\mathrm{e}}^{4},\boldsymbol{F}_{\mathrm{e}}^{5}\}$为编码器每层的输出特征。其中，$\boldsymbol{F}_{\mathrm{d}}^{5}$由$\boldsymbol{F}_{\mathrm{e}}^{5}$经过坐标注意力模块得到。第$i$个SUB模块的输出为$\boldsymbol{F}_{\mathrm{d}}^{i-1}$，其作为第$i-1$个SUB模块的输入。除最下层的SUB模块，每个模块还有一个不同尺度的深度图$\boldsymbol{D}_i$输出，用于多尺度的重建损失计算。

如图2和图4所示，对于第i个SUB模块，输入特征$\boldsymbol{F}_{\mathrm{d}}^{i} \in \mathbf{R}^{H/2^i\times W/2^i\times C_d^i}$依次经过一个卷积核大小为3×3的卷积层、一个ELU层及一个2倍的像素洗牌上采样，得到$\boldsymbol{F}_{\mathrm{d}^*}^{i} \in \mathbf{R}^{H/2^{i-1}\times W/2^{i-1}\times C}$，其公式表达为

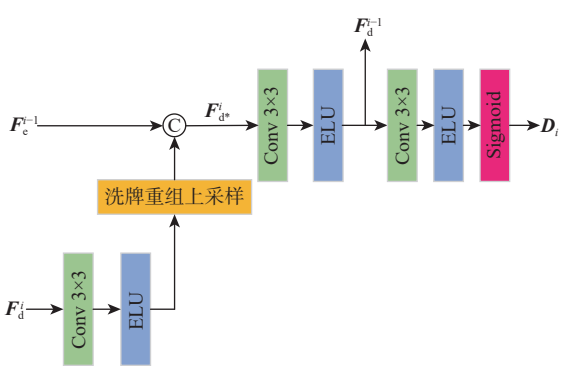


图4 SUB模块

Fig. 4 SUB module

$$\boldsymbol{F}_{\mathrm{d}^*}^{i} = \text{PixelShuffle}(\text{ELU}(\text{Conv}(\boldsymbol{F}_{\mathrm{d}}^{i}))) \tag{17}$$

式中：PixelShuffle()为2倍的像素洗牌上采样操作。

$\boldsymbol{F}_{\mathrm{d}^*}^{i}$与$\boldsymbol{F}_{\mathrm{e}}^{i-1}$在通道维度上拼接后，通过一个卷积核大小为3×3的卷积层和一个ELU层得到输出$\boldsymbol{F}_{\mathrm{d}}^{i-1} \in \mathbf{R}^{H/2^{i-1}\times W/2^{i-1}\times C_d^{i-1}}$。缺少$\boldsymbol{F}_{\mathrm{e}}^{i-1}$为另一特征输入的最下层SUB模块，在此过程中无拼接操作。

$$\boldsymbol{F}_{\mathrm{d}}^{i-1} = \text{ELU}(\text{Conv}(\text{Cat}(\boldsymbol{F}_{\mathrm{d}^*}^{i},\boldsymbol{F}_{\mathrm{e}}^{i-1}))) \tag{18}$$

$\boldsymbol{F}_{\mathrm{d}}^{i-1}$依次通过一个卷积核大小为3×3的卷积层、一个ELU层和Sigmoid层得到深度图$\boldsymbol{D}_i \in \mathbf{R}^{H/2^{i-1}\times W/2^{i-1}\times 1}$。

$$\boldsymbol{D}_i = \text{Sigmoid}(\text{ELU}(\text{Conv}(\boldsymbol{F}_{\mathrm{d}}^{i-1}))) \tag{19}$$

3 实 验

3.1 数据集

KITTI数据集是评测多种自动驾驶任务常用的数据集。本文使用Eigen等[11]的数据分割。对于使用单目序列的训练，遵循Zhou等[2]的预处理来去除静态帧。39 810个单目样本组(每个样本组包括3张连续帧)用于训练，4 424个用于验证。每张图像的分辨率为192×640。对所有图像使用相同的相机内参矩阵，将相机的主点设置为图像中心，将焦距设置为KITTI数据集中所有焦距的平均值。

Cityscapes数据集面向自动驾驶场景，包含50个不同城市街道场景的立体图像序列，图像分辨率为1 028×2 048，共19个类别。选用该数据集的图像用于测试可视化结果，以评估方法的泛化性。

3.2 实施细节

实验训练测试采用的操作系统为Windows10，CPU为Interl(R) Core(TM) i7-12700KF@3.6 GHz，显卡为NVIDIA GeForce RTX 3090Ti，显存为24 GB。软件环境采用Python3.7，训练时使用Pytorch框架搭建所使用的整体框架和网络。使用Adam优化器，学习率为0.000 1，训练批次大小为12，训练轮次为20轮。根据文献[2]，设定损失参数如下：式(5)中α=0.85，式(8)中β=0.001。

3.3 评价指标

为评价不同单目深度估计方法的效果，通常采用4种误差指标及3项准确率指标作为评测指标，包括绝对值相对误差L_{Abs}、平方相对误差L_{Sq}、均方误差L_{RMS}、对数均方误差L_{RMSlog}及3项不同阈值的准确率。各指标具体定义如下：

$$L_{\mathrm{Abs}}=\frac{1}{|N|}\sum_{d\in N}\frac{|d-d^*|}{d^*} \tag{20}$$

$$L_{\mathrm{Sq}}=\frac{1}{|N|}\sum_{d\in N}\frac{|d-d^*|^2}{d^*} \tag{21}$$

$$L_{\mathrm{RMS}}=\sqrt{\frac{1}{|N|}\sum_{d\in N}(d-d^*)^2} \tag{22}$$

$$L_{\mathrm{RMSlog}}=\sqrt{\frac{1}{|N|}\sum_{d\in N}|\ln d-\ln d^*|^2} \tag{23}$$

$$\delta=\max\left(\frac{d_{\mathrm{r}}}{d^*},\frac{d^*}{d_{\mathrm{r}}}\right)<r_{\mathrm{th}}\qquad r_{\mathrm{th}}=1.25,1.25^2,1.25^3 \tag{24}$$

式中: d_{r}和d^*分别为地面真值和预测的深度; N为单幅图像中所有预测深度值的集合。L_{Abs}、L_{Sq}、L_{RMS}、L_{RMSlog}的值越小, 表示深度估计效果越好。δ为衡量模型准确率的指标, 值越大, 表示深度估计效果越好。

3.4　实验结果

3.4.1　定量分析

为评估本文深度预测网络的性能, 与目前最先进的单目深度估计方法进行对比。所有模型都在 192×640 分辨率图像下进行测试, 结果如表 1 所示[3-4,12-27]。本文方法比其他方法有更准确的深度估计结果, 评估结果的数值对比上有了全面提升。对于单目深度估计中最重要的精度指标, 本文方法的$\delta<1.25$达到了 0.907, $\delta<1.25^2$达到了 0.966。对比基线模型[4], L_{Abs}指标优化了 4.2%(0.098 → 0.094), L_{Sq}指标优化了8.7%(0.770 → 0.703), L_{RMS}指标优化了2.5%(4.459 → 4.349), L_{RMSlog}指标优化了 1.1%(0.176 → 0.174)。

3.4.2　消融实验

为评估本文所提模块的有效性, 在 KITTI 数据集上进行了消融实验。本文方法的基线模型[4]使用 ResNet18 作为深度预测网络编码器的主干网络; 如表 2 中第 2 行所示, 为公平起见, 额外添加了基

表 1　KITTI 数据集上的定量结果[3-4,12-27]

Table 1　Quantitative results on KITTI dataset[3-4,12-27]

方法	训练方法	L_{Abs}	L_{Sq}	L_{RMS}	L_{RMSlog}	$\delta<1.25$	$\delta<1.25^2$	$\delta<1.25^3$
文献[3]	M	0.114	0.864	4.817	0.192	0.877	0.959	0.981
文献[12]	M	0.115	0.871	4.778	0.191	0.874	0.963	**0.984**
文献[13]	M	0.120	0.896	4.869	0.198	0.868	0.957	0.981
文献[14]	M+Se	0.113	0.835	4.693	0.191	0.879	0.961	0.981
文献[15]	M+Se	0.112	0.777	4.772	0.191	0.872	0.959	0.982
文献[16]	M	0.111	0.863	4.756	0.188	0.881	0.961	0.982
文献[17]	M	0.111	0.821	4.650	0.187	0.883	0.961	0.982
文献[18]	M+Se	0.112	0.788	4.582	0.187	0.878	0.963	0.983
文献[19]	M	0.109	0.792	4.632	0.185	0.884	0.962	0.983
文献[20]	M	0.110	0.812	4.686	0.187	0.882	0.962	0.983
文献[21]	M+Se	0.105	0.722	4.547	0.182	0.886	0.964	**0.984**
文献[22]	M	0.107	0.765	4.561	0.183	0.886	0.961	0.982
文献[23]	M	0.113	0.851	4.809	0.191	0.877	0.959	0.981
文献[24]	M	0.100	0.805	4.678	0.187	0.879	0.961	0.983
文献[25]	M	0.111	0.829	4.393	0.185	0.883	0.961	0.983
文献[26]	M	0.100	0.747	4.455	0.177	0.895	**0.966**	**0.984**
文献[27]	M	0.102	0.757	4.493	0.178	0.896	0.965	**0.984**
文献[4]	M	0.098	0.770	4.459	0.176	0.900	0.965	0.983
本文	M	**0.094**	**0.703**	**4.349**	**0.174**	**0.907**	**0.966**	**0.984**

注: 粗体表示最优结果, 下划线表示次优结果。M表示以单目视频为训练集训练, Se表示使用语义标签作为监督训练。

表2 消融实验结果

Table 2 Ablation study results

方法	CA(E)	CA(C)	SUB	L_{Abs}	L_{Sq}	L_{RMS}	L_{RMSlog}	$\delta<1.25$	$\delta<1.25^2$	$\delta<1.25^3$
基线模型				0.098	0.770	4.459	0.176	0.900	0.965	0.984
基线模型+Densenet 101				0.096	0.760	4.441	0.175	0.902	0.965	0.984
	√			0.100	0.781	4.466	0.176	0.905	0.965	0.983
		√		0.094	0.747	4.405	0.175	0.903	0.965	0.983
			√	0.094	0.723	4.429	0.175	0.904	0.965	0.983
CAMD	√	√		0.095	0.762	4.501	0.173	0.904	0.965	0.983
		√	√	0.094	0.738	4.496	0.177	0.904	0.965	0.983
	√		√	0.105	0.780	4.486	0.176	0.906	0.966	0.984
	√	√	√	0.094	0.703	4.349	0.174	0.907	0.966	0.984

注: CA(E)表示本文坐标注意力模块用来增强编码器最下层的输出特征, CA(C)表示本文坐标注意力模块用来增强成本体的特征利用。

线模型以 Densenet101 作为深度预测网络编码器的主干网络的实验结果。

从表2可以看出, 本文方法对于深度估计在指标上都有不同的提升效果。其中, 本文坐标注意力机制用来增强编码器最下层输出特征, 将$\delta<1.25$提高至0.905, 明显改善准确率指标, 但对于一些误差指标有负面影响; 用于增强成本体特征利用, 可以明显改善误差指标, 且不损害准确率指标。将两者结合可以很好地优势互补。本文提出的基于像素洗牌的解码器, 无论是对于误差指标, 还是准确率指标, 都有明显的优化。

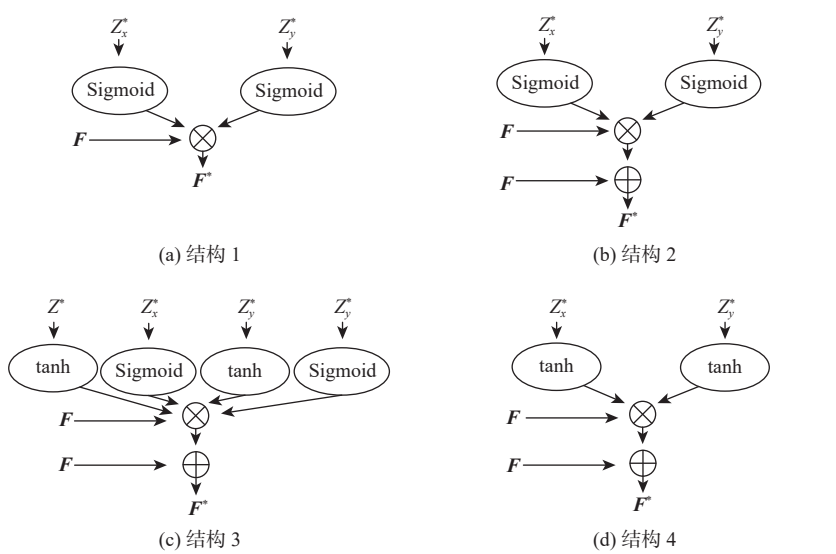
3.4.3 不同激活函数及结构的坐标注意力

如图5所示, 探究了4种不同结构的非线性激活函数对深度估计性能的影响。如表3所示, 结构1为原坐标注意力模块[10]。该实验中, 解码器均为本文提出的解码器, 且在网络中的2处都使用坐标注意力模块。

从评价指标结果上来看, 使用 Sigmoid 作为计算权重量的非线性激活函数, 对于深度估计的性能几乎没有提升。而单使用 tanh, 并引入原特征作为元素相加的残差项, 可以明显改善深度估计评价的各项指标。

3.4.4 定性分析

图6为本文方法在 KITTI 数据集中的定性分析结果。可以看出, 本文方法即使没有用到任何分割信息作为辅助, 仍然可以很好地分离不同深度的物体轮



(a) 结构1 (b) 结构2

(c) 结构3 (d) 结构4

图5 坐标感知注意模块的不同激活函数及结构

Fig. 5 Different activation functions and structures of coordinate-aware attention module

表 3　不同结构的坐标感知注意模块结果

Table 3　Results of different structure of coordinate-aware attention modules

结构	L_{Abs}	L_{Sq}	L_{RMS}	L_{RMSlog}	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
结构1	0.098	0.761	4.401	0.181	0.904	0.965	0.984
结构2	0.097	0.748	4.436	0.181	0.904	0.965	0.984
结构3	0.095	0.750	4.415	0.181	0.905	0.965	0.984
结构4	0.094	0.703	4.349	0.174	0.907	0.966	0.984

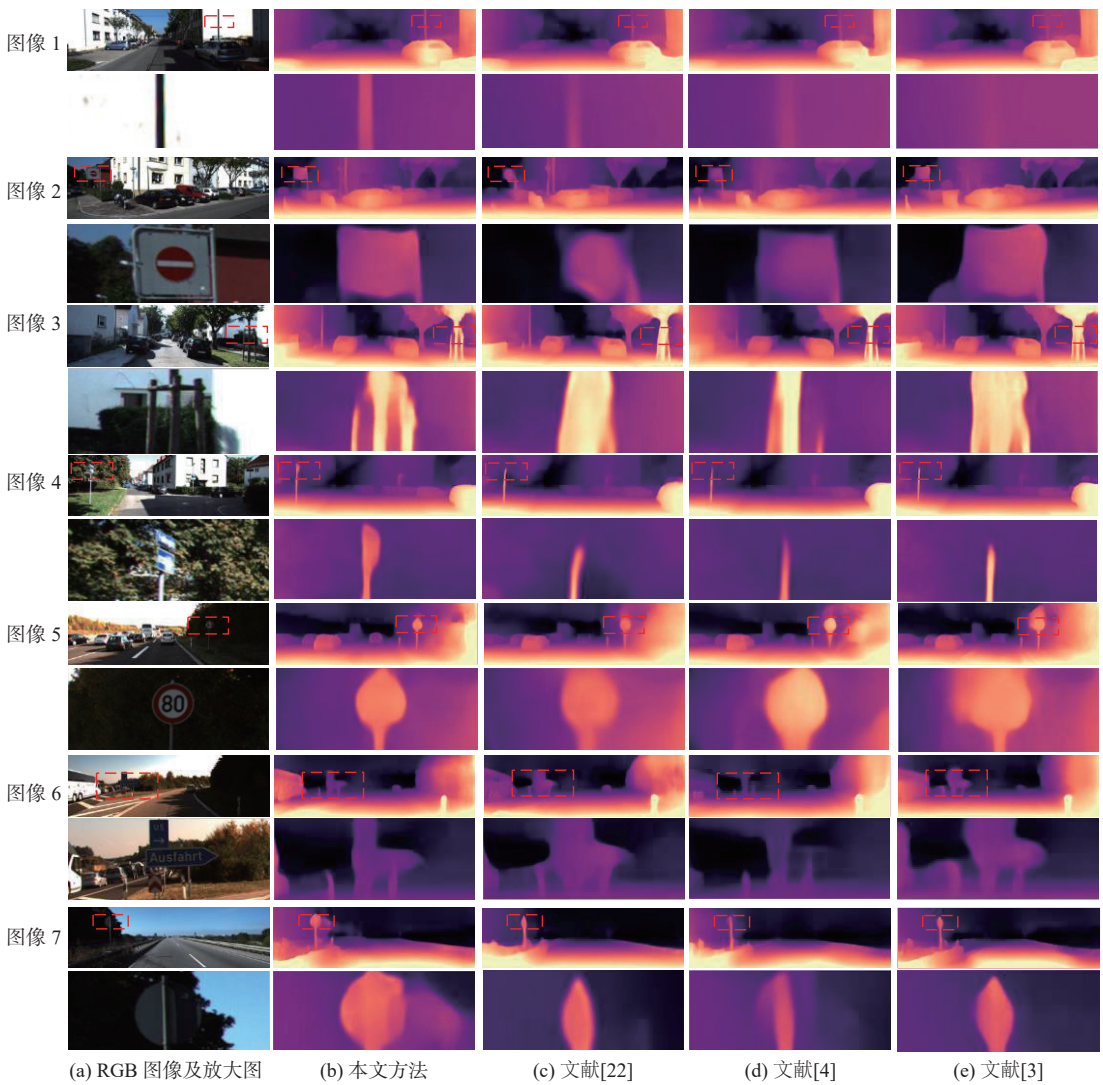


图 6　KITTI 数据集上的可视化结果

Fig. 6　Visualization results on KITTI dataset

廓, 且预测深度相对更加符合对应物体轮廓, 如图 6 中第 6 幅图中远景的路标; 可以很好地分离前景物体与背景, 如图 6 中第 4 幅图中几乎隐于背景树木中的路标; 对于光线变化, 也有很强的鲁棒性, 即使处于阴影处的物体, 也可以很好地识别其轮廓并分离出来, 如图 6 中第 5、7 幅图中处于阴影中的路标。

3.4.5　泛化性能分析

为评估本文模型的通用性, 直接使用 Cityscapes 数据集进行测试。图 7 为 Cityscapes 数据集中的定性分析结果, 表 4 为迁移数据结果。可以看出, 本文方法在直接迁移的情况下依然可以很好地预测出场景的深度, 并且相较于其他方法保留了更多的物体轮廓细节。在面对复杂背景中的物体时, 如图 7 中第 4 幅图中的树干, 仍然可以在深度估计中得以分辨, 保持了物体的完整性和深度一致性。

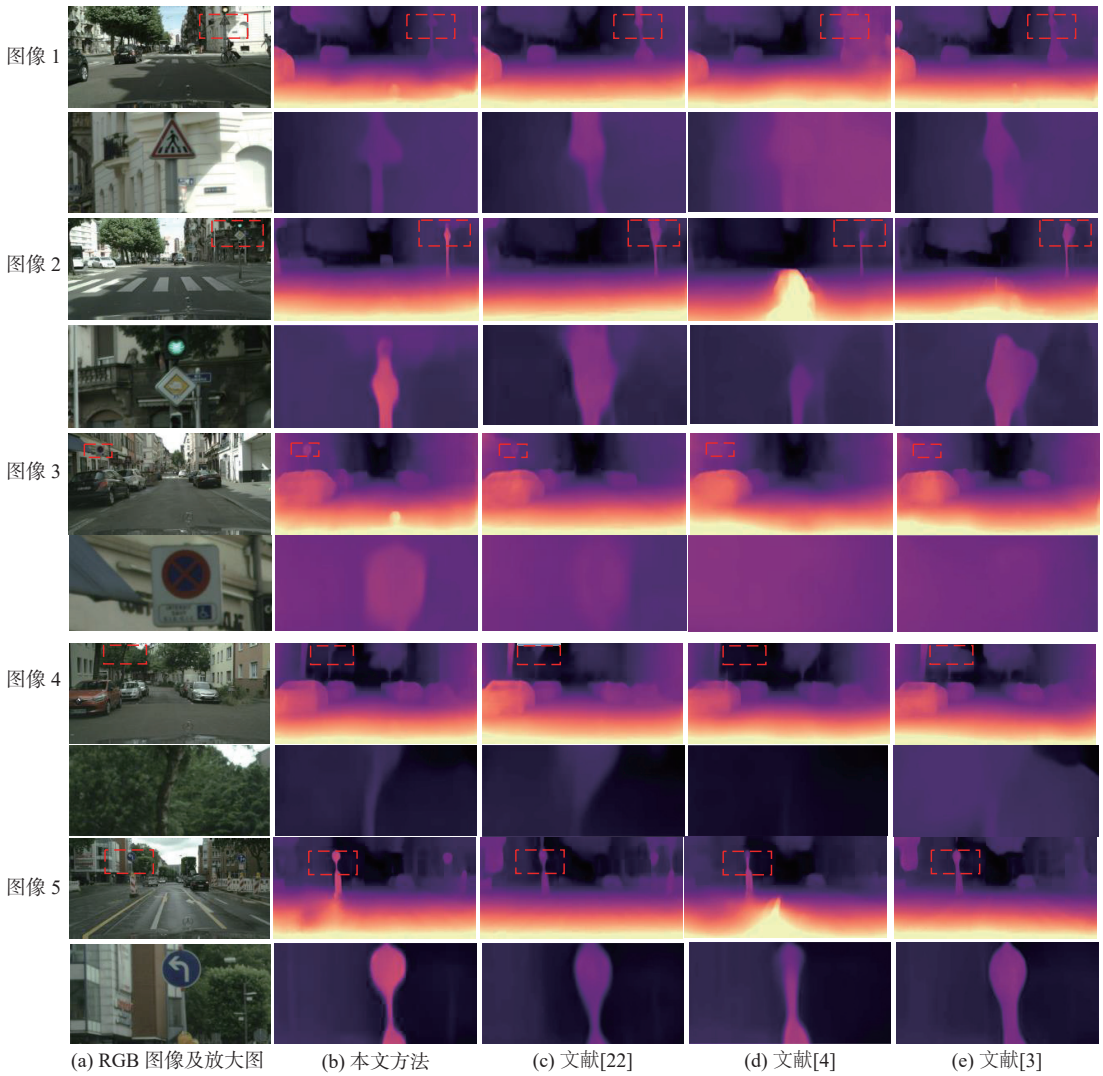


图 7 Cityscapes 数据集上的可视化结果

Fig. 7 Visualization results on Cityscapes dataset

表 4 Cityspaces 数据集上迁移结果

Table 4 Transfer results on Cityspaces dataset

方法	L_{Abs}	L_{Sq}	L_{RMS}	L_{RMSlog}	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
文献[3]	0.178	1.802	8.520	0.249	0.731	0.929	0.977
文献[22]	0.178	1.797	8.537	0.244	0.731	0.928	0.977
文献[4]	0.170	1.789	8.357	0.236	0.748	0.932	0.978
本文	0.165	1.781	8.282	0.221	0.761	0.939	0.982

4 结 论

本文提出了一种基于坐标注意的多帧自监督单目深度估计方法。

1）通过坐标感知注意机制增强低分辨率编码器特征。

2）加强成本体中的特征提取。

3）提出了一种基于像素洗牌的新型深度预测解码器，能够在解码过程中有效分离原图各物体在低分辨率特征中的融合信息，解决深度估计中物体边缘模糊不清的问题。

4）在 KITTI 数据集上进行了实验，取得了良好的预测效果与评测指标，为进一步验证方法的鲁棒性，在 Cityscapes 数据集上进行了测试和比较。实验结果表明，本文方法在深度估计方向取得了较好性能，可以在深度估计结果中细化物体边缘，在未知环境中具有良好的泛化性能。

未来工作需进一步提高深度预测网络的泛化能力，降低网络参数量，提高计算效率。

参考文献（References）

[1] 李杰, 李一轩, 吴天生, 等. 基于 FPGA 无人机影像快速低功耗高

精度三维重建[J]. 北京航空航天大学学报, 2021, 47(3): 486-499.
LI J, LI Y X, WU T S, et al. Fast, low-power and high-precision 3D reconstruction of UAV images based on FPGA[J]. Journal of Beijing University of Aeronautics and Astronautics, 2021, 47(3): 486-499(in Chinese).

[2] ZHOU T H, BROWN M, SNAVELY N, et al. Unsupervised learning of depth and ego-motion from video[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2017: 6612-6619.

[3] GODARD C, MAC AODHA O, FIRMAN M, et al. Digging into self-supervised monocular depth estimation[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE Press, 2019: 3827-3837.

[4] WATSON J, MAC AODHA O, PRISACARIU V, et al. The temporal opportunist: self-supervised multi-frame monocular depth[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2021: 1164-1174.

[5] GUIZILINI V, AMBRUŞ R, CHEN D, et al. Multi-frame self-supervised depth with Transformers[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2022: 160-170.

[6] SHI W Z, CABALLERO J, HUSZÁR F, et al. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2016: 1874-1883.

[7] GEIGER A, LENZ P, STILLER C, et al. Vision meets robotics: the KITTI dataset[J]. The International Journal of Robotics Research, 2013, 32(11): 1231-1237.

[8] HE K M, ZHANG X Y, REN S Q, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2016: 770-778.

[9] HUANG G, LIU Z, VAN DER MAATEN L, et al. Densely connected convolutional networks[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2017: 2261-2269.

[10] HOU Q B, ZHOU D Q, FENG J S. Coordinate attention for efficient mobile network design[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2021: 13708-13717.

[11] EIGEN D, PUHRSCH C, FERGUS R. Depth map prediction from a single image using a multi-scale deep network[EB/OL]. (2014-06-09)[2023-06-01]. https://arxiv.org/abs/1406.2283v1.

[12] ZOU Y L, JI P, TRAN Q H, et al. Learning monocular visual odometry via self-supervised long-term modeling[C]//Proceedings of the 16th European Conference on Computer Vision. Berlin: Springer, 2020: 710-727.

[13] GUIZILINI V, HOU R, LI J, et al. Semantically-guided representation learning for self-supervised monocular depth[EB/OL]. (2020-02-27)[2023-06-01]. https://arxiv.org/abs/2002.12319v1.

[14] KLINGNER M, TERMÖHLEN J A, MIKOLAJCZYK J, et al. Self-supervised monocular depth estimation: solving the dynamic object problem by semantic guidance[C]//Proceedings of the 16th European Conference on Computer Vision. Berlin: Springer, 2020: 582-600.

[15] LEE S, IM S, LIN S, et al. Learning monocular depth in dynamic scenes via instance-aware projection consistency[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2021, 35(3): 1863-1872.

[16] POGGI M, ALEOTTI F, TOSI F, et al. On the uncertainty of self-supervised monocular depth estimation[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2020: 3224-3234.

[17] PATIL V, VAN GANSBEKE W, DAI D X, et al. Don't forget the past: recurrent depth estimation from monocular video[J]. IEEE Robotics and Automation Letters, 2020, 5(4): 6813-6820.

[18] CHOI J, JUNG D, LEE D, et al. SAFENet: self-supervised monocular depth estimation with semantic-aware feature extraction[EB/OL]. (2020-11-29)[2023-06-01]. https://arxiv.org/abs/2010.02893v3.

[19] WANG J D, SUN K, CHENG T H, et al. Deep high-resolution representation learning for visual recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021, 43(10): 3349-3364.

[20] YAN J X, ZHAO H, BU P H, et al. Channel-wise attention-based network for self-supervised monocular depth estimation[C]//Proceedings of the International Conference on 3D Vision. Piscataway: IEEE Press, 2021: 464-473.

[21] JUNG H, PARK E, YOO S. Fine-grained semantics-aware representation enhancement for self-supervised monocular depth estimation[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE Press, 2021: 12622-12632.

[22] ZHANG N, NEX F, VOSSELMAN G, et al. Lite-Mono: a lightweight CNN and transformer architecture for self-supervised monocular depth estimation[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2023: 18537-18546.

[23] SURI Z K. Pose constraints for consistent self-supervised monocular depth and ego-motion[C]//Proceedings of the Scandinavian Conference on Image Analysis. Berlin: Springer, 2023: 340-353.

[24] BOULAHBAL H, VOICILA A, COMPORT A. STDepthFormer: predicting spatio-temporal depth from video with a self-supervised transformer model[EB/OL]. (2023-05-02)[2023-06-01]. https://arxiv.org/abs/2303.01196v1.

[25] LIU Y X, XU Z H, HUANG H Y, et al. FSNet: redesign self-supervised MonoDepth for full-scale depth prediction for autonomous driving[EB/OL]. (2023-04-21)[2023-06-01]. https://arxiv.org/abs/2304.10719v1.

[26] SAUNDERS K, VOGIATZIS G, MANSO L J. Self-supervised monocular depth estimation: let's talk about the weather[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE Press, 2023: 8873-8883.

[27] WANG B Y, WANG S, YE D, et al. Deep neighbor layer aggregation for lightweight self-supervised monocular depth estimation [EB/OL]. [2023-06-01]. https://arxiv.org/abs/2309.09272v2.

Coordinate-aware attention-based multi-frame self-supervised monocular depth estimation

CHENG Deqiang[1], FAN Shuming[1], QIAN Jiansheng[1], JIANG He[1], KOU Qiqi[2,*]

(1. School of Information and Control Engineering, China University of Mining and Technology, Xuzhou 221000, China;
2. School of Computer Science and Technology, China University of Mining and Technology, Xuzhou 221000, China)

Abstract: A novel multi-frame self-supervised single-image depth estimation technique based on coordinate-aware attention has been presented to tackle the issue of hazy depth prediction near object edges in single-image depth estimation methods. Firstly, a coordinate-aware attention module is proposed to enhance the output features of the bottom layer of the encoder and improve the feature utilization of the cost volume. To improve the object edges in depth prediction results, a new pixel-shuffle-based depth prediction decoder is also suggested. This decoder can efficiently separate the multi-object fusion features in low-resolution encoder features. Experimental results on the KITTI and Cityscapes datasets demonstrate that the proposed method is superior to current mainstream methods, significantly improving subjective visual effects and objective evaluation indicators, especially with better depth prediction performance in object edge details.

Keywords: image processing; deep learning; depth estimation; self-supervised learning; attention mechanism

Received: 2023-06-29; **Accepted**: 2023-11-10; **Published Online**: 2023-11-16 11:36

URL: link.cnki.net/urlid/11.2625.V.20231115.1718.002

Foundation items: National Natural Science Foundation of China (52204177,52304182); The Fundamental Research Fund for the Central Universities (2020QN49)

* **Corresponding author**. E-mail: kouqiqi@cumt.edu.cn

http://bhxb.buaa.edu.cn jbuaa@buaa.edu.cn
DOI: 10.13700/j.bh.1001-5965.2023.0492

递归门控增强与金字塔预测的铁路全景分割

陈永[1,2,*]，周方春[1]，张娇娇[1]

(1. 兰州交通大学 电子与信息工程学院，兰州 730070；2. 甘肃省人工智能与图形图像处理工程研究中心，兰州 730070)

摘 要：针对高速铁路场景全景分割时存在目标特征提取不充分、边缘轮廓分割模糊等问题，提出了一种递归门控增强与金字塔预测的铁路全景分割网络。在DETR模型的基础上，构建改进多尺度级联CSP-DarkNet53特征提取网络，提升对不同尺度的铁路场景目标特征提取能力；提出递归门控与类特征增强模块，获取更丰富的边缘特征信息，增强对边缘轮廓信息的提取和分割的能力；将多尺度可变形注意力引入编码骨干网络中，进一步捕获多尺度上下文信息，减少分割细节特征丢失；通过改进金字塔预测与像素类别分割模块，实现铁路全景的分割输出。实验结果表明：相比于原始DETR模型，所提方法的全景分割质量指标PQ提升了7.4%，前景实例目标评价指标PQ^{Th}提升了9.7%，背景填充区域质量评价指标PQ^{St}提升了6.6%。所提方法在铁路场景下图像全景分割具有较好的性能，主观评价均优于对比方法。

关 键 词：全景分割；DETR；递归门控增强；金字塔预测；高速铁路
中图分类号：TP391.4
文献标志码：A **文章编号**：1001-5965(2025)07-2229-11

随着中国高速铁路的快速发展，列车运行安全备受关注[1]。行人、车辆等或山体滑坡等异物侵入铁路限界，极易造成交通事故，严重危及乘客人身安全。鉴于此，对高速铁路安全提出了更高的要求，实现列车运行时对铁路场景中物体的分割和检测对于中国智能高铁发展至关重要[2]。列车运行过程中，获得铁路场景中的全景信息能够增强对铁路异物的判断能力，进一步提升列车的运行安全防护。但在铁路的全景分割中，由于物体类别繁多且数量较大[3]，目标边缘易与背景信息发生混淆，造成分割准确率低、边缘轮廓分割模糊等问题，严重影响图像分割的效果。

全景分割[4]的任务是对图像中每个像素点分配语义类别标签和实例编号。随着计算机视觉技术的不断发展，全景分割逐渐成为热点研究方向。Kirillov等[5]基于掩膜区域卷积神经网络(mask region-convolutional neural nerwork, Mask R-CNN)，通过引用特征金字塔和多任务损失函数方法进行全景分割，但该方法网络结构单一，存在模型特征提取能力弱的问题，整体分割效果较差。Zhang等[6]提出了一种基于动态内核的深度学习分割模型，采用动态内核的方法进行全景分割，但在特征提取时存在边缘细节丢失问题，导致分割效果较差。Carion等[7]以端到端的Transformer深度学习全景分割网络为基础，利用ResNet网络实现对图像的特征提取，但存在对图像中特征提取不充分、分割结果不精细的问题。Cheng等[8]基于Transformer结构的深度学习分割模型，结合自注意力机制学习图像的特征，但目标分割后边缘较为粗糙。Wang等[9]提出了一种双路径Mask Transformer的深度学习模型，通过引入像素卷积路径对图像进行全景分割，但在特征融合时会造成语义鸿沟，导致图像分割时边缘轮廓

收稿日期：2023-07-27；**录用日期**：2023-09-26；**网络出版时间**：2023-10-07 17:42
网络出版地址：link.cnki.net/urlid/11.2625.V.20231007.1038.001
基金项目：国家自然科学基金(62462043,61963023)；兰州交通大学重点研发项目(ZDYF2304)
*通信作者. E-mail：edukeylab@126.com

引用格式：陈永，周方春，张娇娇. 递归门控增强与金字塔预测的铁路全景分割[J]. 北京航空航天大学学报，2025，51(7)：2229-2239.
CHEN Y，ZHOU F C，ZHANG J J. Railway panoramic segmentation based on recursive gating enhancement and pyramid prediction[J]. Journal of Beijing University of Aeronautics and Astronautics，2025，51(7)：2229-2239 (in Chinese).

易出现模糊的问题。Cheng等[10]采用语义和实力分割的双空洞空间卷积池化金字塔(atrous spatial pyramid pooling, ASPP)和双解码结构,通过神经网络和Atrous卷积实现图像的全景分割,但存在误分割及分割不完整的问题。Cheng等[11]采用基于Transformer的双解码器网络结构,结合自注意模块和交叉注意模块实现图像的全景分割,但在连续的解码器层之间存在不一致的掩码预测,导致全景分割预测效果较差。Yang等[12]提出了一种自底向上的图像解析网络,利用边界框的4个角及中心点进行实例掩码分割,但对边缘细节处理效果较差,存在分割边缘不平滑的问题。

综上所述,现有大多数铁路场景下,图像全景分割方法存在目标特征提取不充分、边缘轮廓分割模糊的问题[13]。针对上述问题,本文提出了递归门控增强与金字塔预测的铁路全景分割网络。首先,提出改进多尺度级联CSP-DarkNet53特征提取网络,提升对不同尺度的铁路场景目标特征提取能力,改善模型中对目标特征提取不充分的问题。其次,提出递归门控与类特征增强(recursive gating and class feature enhancement, RG-CFE)模块,增强对多种类别铁路场景下物体边缘轮廓提取和分割的能力。然后,将多尺度可变形注意力引入编码骨干网络中,实现对多尺度上下文信息的捕捉,减少细节特征的丢失。最后,通过改进金字塔预测与像素类别分割模块相结合实现对铁路场景下的图像全景分割。实验结果表明,本文方法和比较方法相比,对铁路场景下图像全景分割具有较好的性能。

1 本文方法

1.1 网络整体架构

DETR模型是一种目标检测结构,使用基于Transformer的网络结构进行目标检测及分割,被广泛应用于目标分割中[7]。本文基于DETR模型,提出了一种递归门控增强与金字塔预测的铁路全景分割网络。网络整体结构如图1所示。所提全景分割模型由以下模块组成:①改进多尺度级联CSP-DarkNet53特征提取网络;②递归门控与类特征增强模块;③多尺度可变形注意力编码器;④改进金字塔预测与像素类别分割模块。模型工作时,首先,采用提出的改进多尺度级联CSP-DarkNet53特征提取网络,通过密集连接替换残差网络实现多尺度特征提取,改善

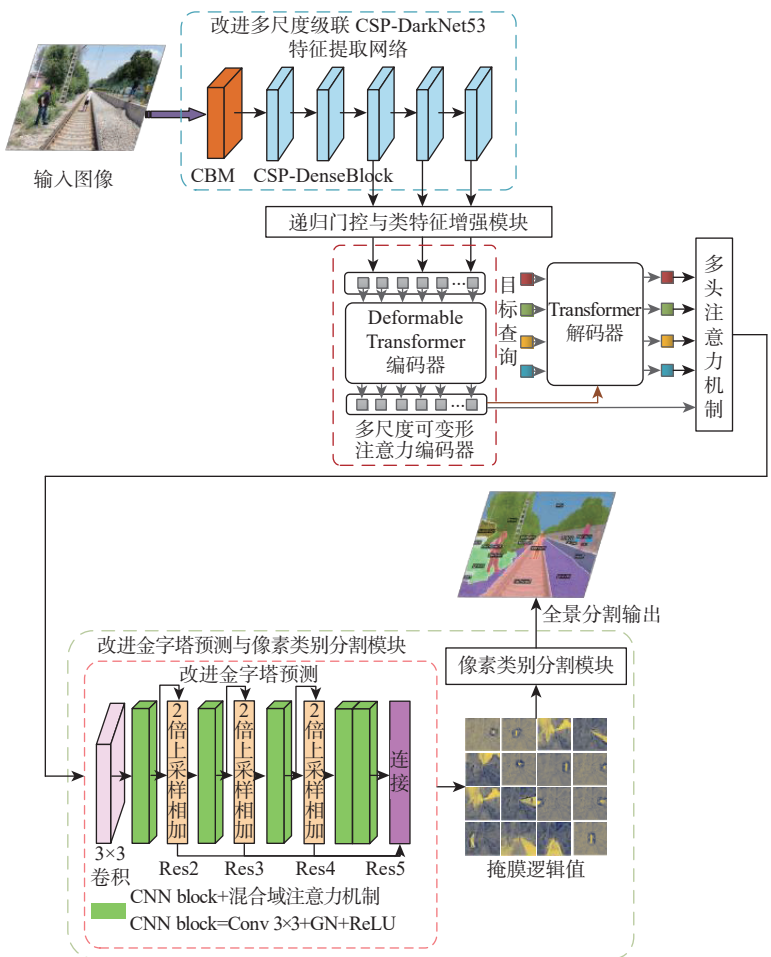


图1 本文方法整体模型框架

Fig. 1 Overall modelling framework of the proposed method

全景分割模型中对目标特征提取不足的问题。其次，利用类特征注意力机制与递归门控卷积，设计递归门控与类特征增强模块，增强对多种类别铁路场景下物体边缘轮廓检测和分割的精度。然后，改进编码器，利用多尺度可变形注意力减少细节特征丢失，提升分割性能。最后，通过改进金字塔预测并与像素类别分割模块结合对图像信息进行预测和分割，实现对铁路场景的全景分割。

1.2　改进多尺度级联 CSP-DarkNet53 特征提取网络

在原始 DETR 全景分割过程中，通过 ResNet 特征提取网络对图像进行单尺度特征提取，但由于单尺度特征提取只考虑一个尺度的特征信息，无法充分捕获输入图像的多层次特征，导致全景分割中目标提取不充分[14]。针对上述问题，本文提出多尺度级联的 CSP-DarkNet53 网络进行特征提取，与 ResNet 单尺度特征提取网络相比，所提方法采用密集连接结构替换 CSP-DarkNet53 模块中的残差网络作为特征提取骨干网络，实现对铁路场景下图像的目标特征提取。通过多尺度级联结构结合不同层次下的特征和语义信息，减少特征信息丢失。本文所采用的多尺度级联 CSP-DarkNet53 特征提取网络结构如图 2 所示。

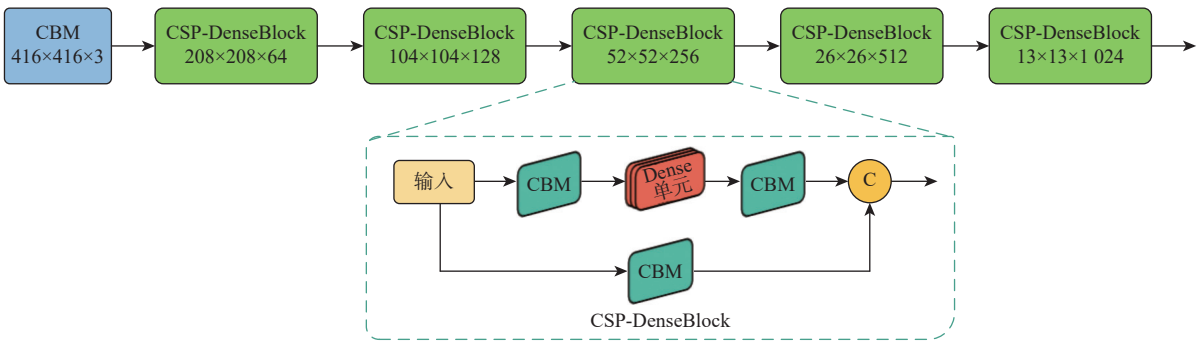


图 2　多尺度级联 CSP-DarkNet53 特征提取网络结构

Fig. 2　Structure of multi-scale cascaded CSP-DarkNet53 feature extract network

改进前 CSP-DarkNet53 作为骨干特征提取网络，主要由 CBM 卷积和 CSP-ResBlock 卷积残差组合模块构成。其中，CSP-ResBlock 由 3 个 CBM 卷积块和多个残差模块张量拼接而成[15]，残差模块由主干卷积路径和残差边组成[16]。在骨干提取网络中，使用残差模块引用残差结构，通过增加网络层次来提高特征提取能力，但由于残差量较大，堆叠层数较深，难以对信息量较小的特征进行有效提取。DenseNet[17] 模型的网络整体结构与 ResNet 结构相似，属于密集型的端到端的网络模型，DenseNet 每层直接建立密集连接，因此，在本文特征提取网络中使用 DenseNet 模型替换 ResNet 模型能获得更加丰富的特征信息。DenseNet 模型中的特征计算表达式为

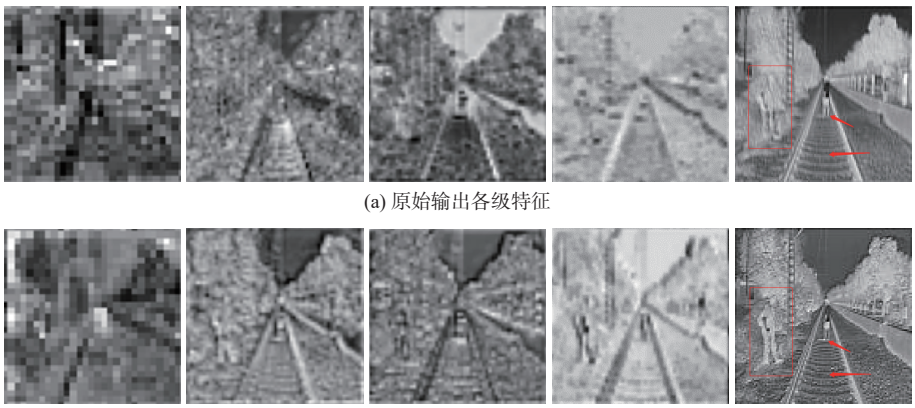
$$\boldsymbol{x}_l = H_l([\boldsymbol{x}_0, \boldsymbol{x}_1, \cdots, \boldsymbol{x}_{l-1}]) \tag{1}$$

式中：$\boldsymbol{x}_l$为输出；$H_l(\cdot)$为非线性函数；l为当前层；$[\boldsymbol{x}_0, \boldsymbol{x}_1, \cdots, \boldsymbol{x}_{l-1}]$表示特征图通道维度合并。

在原始 CSP-DarkNet53 网络的基础上，通过 DenseNet 网络替换 ResNet 构成 CSP-DenseBlock 模块。随着下采样的不断加深，网络不断地从浅层特征学习到深层语义特征学习，有效改善了提取中目标特征信息丢失的问题，此外，通过密集连接可以充分利用特征层间的语义信息，加强不同层间特征图的关联性，提升获取更多关键信息的能力。为直观展示所提级联特征网络的有效性，采用特征可视化方法进行对比实验，比较结果如图 3 所示。其中，图 3(a) 为原始 DETR 骨干提取网络 ResNet 输出的各级特征图，图 3(b) 为本文方法改进后输出的各级特征图。从图 3(a) 和图 3(b) 整体可以看出，改进前后特征图随着浅层特征到深层特征逐级进行，其语义特征都呈现出逐渐清晰化的趋势。其中，浅层特征离输入端较近，该级特征主要包含更多的像素点信息，由于其经过卷积操作较少，语义性最低，噪声更多；随着特征提取级数的增加，提取的特征逐渐包含更多的纹理、边缘等信息；最终在离输出端较近的深层特征中，其包含更多的抽象信息，即更加丰富的语义信息。此外，可以看出，图 3(a) 最后一幅为原始方法最后一级特征提取结果，仍存在提取图像不清晰、边缘模糊的问题，而图 3(b) 最后一幅为本文方法最后一级提取特征图，可以看出，通过本文多尺度级联特征提取后，其特征边缘和细节更加清晰，提升了模型对目标特征的提取能力。

1.3　递归门控与类特征增强模块

铁路场景中，目标种类多且复杂，对其特征进行准确提取存在较大难度，全景分割中，对相邻类别边缘区域的区分能力较弱，边缘检测较难。为增强类别之间的相关性，提取和处理不同类别的语义信息，本文通过引用类特征注意力机制[18] 捕捉类之间的依



(a) 原始输出各级特征

(b) 本文方法输出各级特征

图3 各级特征图可视化比较

Fig. 3 Visual comparison of feature maps at different levels

赖关系，增强对各类信息的理解能力，并通过递归门控卷积(gnConv)[19]从空间交互维度实现对空间特征权重的学习和更新。为解决图像中边缘轮廓模糊的问题，本文提出递归门控与类特征增强模块，以便增强网络对边缘特征信息的获取能力，提升对不同物体边缘轮廓信息的提取和分割能力，如图4所示。

图4中，在类特征注意力模块中，输入特征 $\boldsymbol{A}\in\mathbf{R}^{C\times H\times W}$ 通过1×1卷积生成特征图 $\boldsymbol{B}\in\mathbf{R}^{C'\times H\times W}$ 和类别关注度特征图 $\boldsymbol{D}\in\mathbf{R}^{N\times H\times W}$，其中，$C$ 为通道数，H 为高度，W 为宽度，C' 为 $\boldsymbol{B}$ 缩减后的通道数，N 为类别数。将 $\boldsymbol{B}$ 转换为 $\boldsymbol{B}\in\mathbf{R}^{C'\times HW}$，同时将 $\boldsymbol{D}$ 经过 Softmax 函数后变换得到 $\boldsymbol{D}\in\mathbf{R}^{N\times HW}$，将 $\boldsymbol{B}\in\mathbf{R}^{C'\times HW}$ 和 $\boldsymbol{D}\in\mathbf{R}^{N\times HW}$ 的转置相乘，并通过 Softmax 函数生成类相似性映射图 $\boldsymbol{X}\in\mathbf{R}^{C'\times N}$。将 $\boldsymbol{X}\in\mathbf{R}^{C'\times N}$ 和 $\boldsymbol{D}\in\mathbf{R}^{N\times HW}$ 相乘得到 $\boldsymbol{Y}\in\mathbf{R}^{C'\times H\times W}$，将其通过1×1卷积操作后与 $\boldsymbol{A}$ 通道数相加，最终输出类增强特征图，如下：

$$\boldsymbol{F}_u=f\left(f\left(\sum_{k=1}^{N}\left(x_{u,k}\mathrm{e}^{D_k}\Big/\sum_{j=1}^{N}\mathrm{e}^{D_j}\right)+A_u\right)\right) \tag{2}$$

式中：$\boldsymbol{F}_u$ 为输出特征 $\boldsymbol{F}\in\mathbf{R}^{C\times H\times W}$ 第 u 个通道；$f(\cdot)$ 表示1×1卷积、BN层、ReLU激活函数3层叠加运算；$x_{u,k}$ 为类相似性映射 $\boldsymbol{X}\in\mathbf{R}^{C'\times N}$ 中位置 u 与类别 k 的关联权重，$u\in[1,2,\cdots,C'],k\in[1,2,\cdots,N]$。

式(2)显示每个通道的最终输出是类特征注意图中所有通道基于类别的加权和，表示特征图之间基于类别的语义依赖，以获得更为丰富的图像中各类别的细节信息。

类增强特征图经过递归门控卷积对图像中局部信息建立依赖关系，递归门控卷积由门控卷积和递归设计来实现，先通过2个卷积层调整通道数，再将深度可分离卷积的输出特征划分为多块，并逐元素相乘得到最终的输出特征图，操作公式如下：

$$\boldsymbol{F}_{\text{RG-CFE}}=G_n(\boldsymbol{F}_u) \tag{3}$$

式中：G_n 表示门控卷积。

为直观验证本文引入递归门控与类特征增强模块的有效性，进行改进前后可视化比较实验。将特征图与输入原图进行空间映射，得到相应的热力图，比较结果如图5所示。在热力图中，越偏红色说明该区域对于最终图像分割结果的贡献越大，紫色说明越小。从图5可以看出，改进后，本文方法对于钢轨、行人异物等可以聚焦到更多的类别。此外，继续对类特征增强改进前后进行可视化比较，结果如图6所示。可以看出，未加入类特征增强模块前，其细节信息缺失严重，如铁轨、轨旁电杆孔洞、侵入铁轨行人等细节信息缺失，改进后可以发

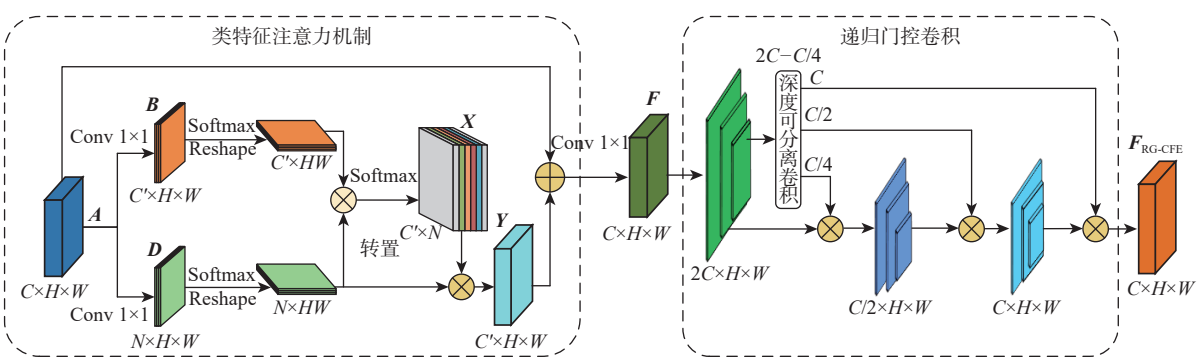
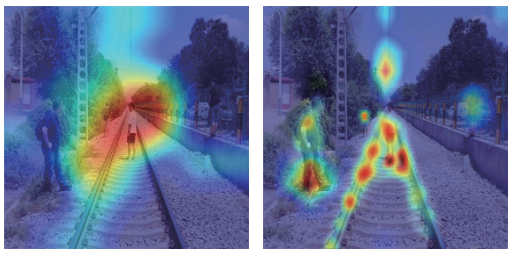


图4 递归门控与类特征增强模块

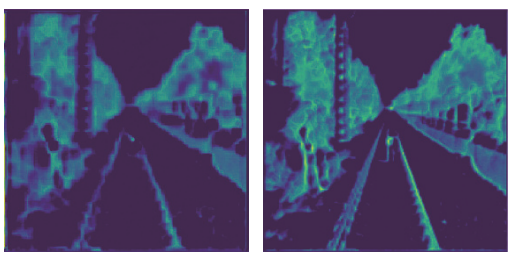
Fig. 4 Recursive gating and class feature enhancement module



(a) 改进前　　(b) 改进后

图 5　类别关注度特征热力图可视化

Fig. 5　Visualization of heat maps based on class attention feature



(a) 未添加类特征增强模块　　(b) 添加类特征增强模块

图 6　类特征增强特征可视化

Fig. 6　Visualization of class feature enhancement

现优于未加入类特征增强模块的结果，说明加入本文所提类特征增强模块后，对目标能够有效进行精确定位和聚焦，进一步提升了模型对目标边缘等特征信息的针对性。

1.4　多尺度可变形注意力编码器

在编码器中，采用注意力检测模块用于图像特征的提取，但注意力检测模块难以有效利用含有较多目标信息的多尺度特征图，会造成部分目标语义信息在下采样过程中信息丢失[20]。为更好地提取多尺度重要上下文特征信息，在编码器中引入多尺度可变形注意力机制。多尺度可变形注意力编码器[21]能够自然地聚合不同尺度的特征图，减少目标语义信息下采样过程中的损失，提升对重要上下文信息的捕获能力。多尺度可变形注意力机制示意图如图 7 所示。

图 7 中，通过骨干提取网络获得多尺度特征图 $\{\boldsymbol{F}^l_{\text{RG-CFE}}\}_{l=1}^{L}$，其中，$\boldsymbol{F}^l_{\text{RG-CFE}} \in \mathbf{R}^{C\times H_l\times W_l}$（$H_l$为第 l层特征图的高度，W_l为第l层特征图的宽度），按位置添加可学习的位置编码后，送入可变形注意力机制中进行注意力权重更新，送入的多尺度特征图通过全连接层转化为像素点注意力权重W'_m，查询特征z_q通过全连接层分别转化为采样偏移量Δp_{mlqk}及注意力权重系数A_{mlqk}。在每个查询中，所有层都采样k个点，融合了不同层的特征。多尺度可变形注意力机制如下：

$$\text{MSDeformAttn}\left(z_q,\hat{p}_q,\{\boldsymbol{F}^l_{\text{RG-CFE}}\}_{l=1}^{L}\right)=\sum_{m=1}^{M}W_m\left[\sum_{l=1}^{L}\sum_{k=1}^{K}A_{mlqk}W'_m\boldsymbol{F}^l_{\text{RG-CFE}}\left(\phi_l\left(\hat{p}_q\right)+\Delta p_{mlqk}\right)\right] \tag{4}$$

式中：q为查询元素的索引；z_q为查询元素的特征表示；$\hat{p}_q\in[0,1]^2$为查询元素参考点的归一化坐标；L为特征层的总数；$\boldsymbol{F}_{\text{RG-CFE}}$为特征图；$W_m$为多头注意力；$\Delta p_{mlqk}$和$A_{mlqk}$分别为第$l$个特征层中的第$m$个注意力头中第$k$个采样点的查询$q$的采样偏移和注意力权重，$\Delta p_{mlqk}$和$A_{mlqk}$均为可学习参数，采样偏移量$\Delta p_{mlqk}$将输入特征图查询特征$z_q$通过一个全连接 Linear 线性层学习得到，通过该操作后可以得到每个 l 层上相应的 k 个采样点，注意力权重系数A_{mlqk}将输入查询特征z_q通过一个全连接 Linear 层和

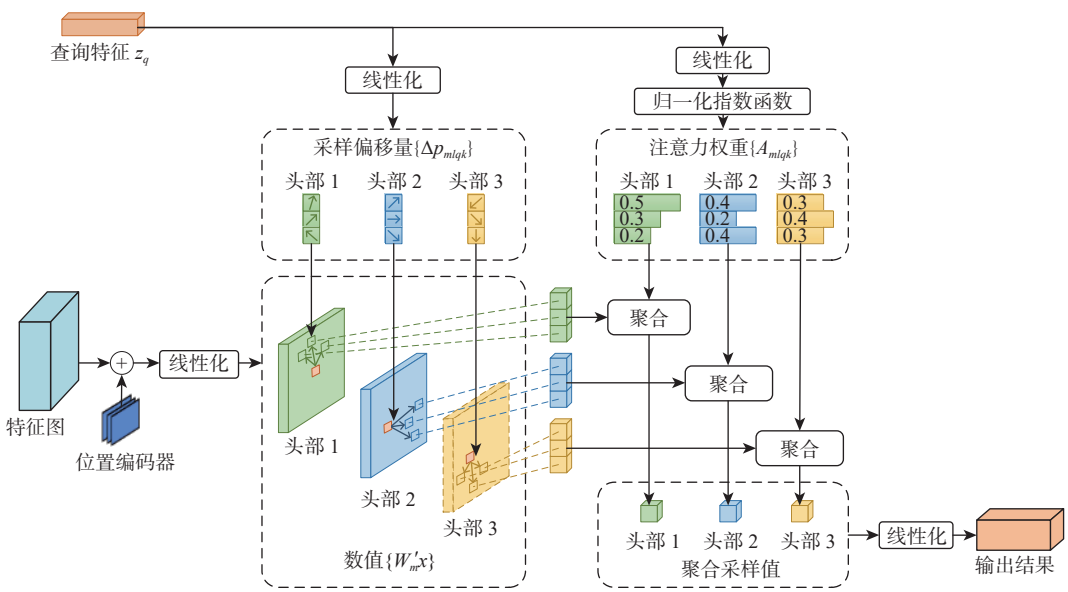


图 7　多尺度可变形注意力机制示意图

Fig. 7　Schematic of multi-scale deformable attention mechanism

Softmax激活函数学习得到，其取值范围为[0, 1]，且注意力权重之和约束为1[7]；$\phi_l(\hat{p}_q)$对$\hat{p}_q$坐标进行缩放以对应相应的特征层。

1.5 改进金字塔预测与像素类别分割模块

通过主干网络对图像特征进行提取，将提取的特征传递到多尺度可变形注意力编码器中，同时将一定数量的目标查询嵌入解码器中作为输入，把编码器和解码器的输出传入多头注意力机制中生成每个查询的目标图像，但随着特征图尺寸不断增加，目标间的特征信息逐渐降低。为准确捕捉图像之间的依赖关系，增强图像掩码预测的准确性，通过改进金字塔预测网络，在金字塔卷积神经网络中增加混合域注意力机制(mixed domain attention mechanism, MDAM)[22]改善全景分割模型分割精度低和鲁棒性差的问题。混合域注意力机制结构如图8所示。

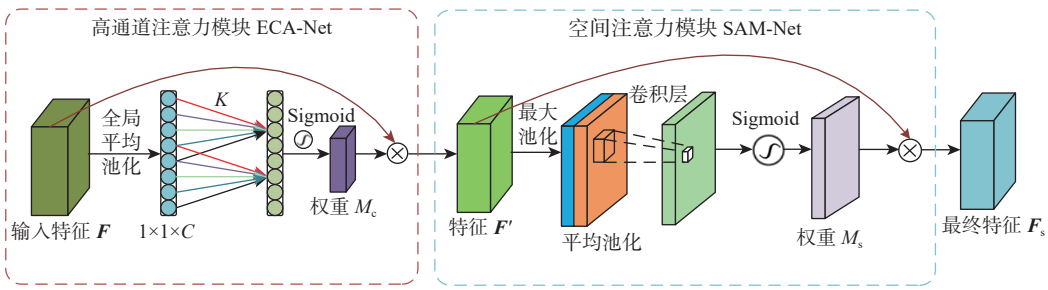


图8 混合域注意力机制结构

Fig. 8 Architecture of mixed domain attention mechanism

混合域注意力机制由高通道注意力模块ECA-Net[23]和空间注意力模块SAM-Net[24]组成，分别捕获特征矩阵各通道和各空间之间的特征依赖关系。首先，将加强特征提取中得到的输入维度为$W\times H\times C$的特征图$\boldsymbol{F}$送入到ECA-Net模块中进行池化，将其压缩为$1\times1\times C$的特征图；然后，采用Sigmoid激活函数处理，恢复为$1\times1\times C$特征图尺寸，自适应确定卷积核大小。

$$k=\left|\frac{\log_2 C+b}{\gamma}\right|_{\text{odd}} \tag{5}$$

式中：k为卷积核大小；$|\,|_{\text{odd}}$表示k只能取奇数；γ和b用于改变通道数C和卷积核大小k之间的比例。

经过上述操作，最终得到更新后的权重$\boldsymbol{M}_{\text{c}}$，将$\boldsymbol{M}_{\text{c}}$与输入特征映射相乘，生成新的特征映射$\boldsymbol{F}'$，再将其输入到SAM-Net模块中，通过卷积层和Sigmoid函数得到权重$\boldsymbol{M}_{\text{s}}$，计算式如下：

$$\begin{aligned}\boldsymbol{M}_{\text{s}}&=\sigma\left(f^{7\times7}\left(\left[\text{AvgPool}(\boldsymbol{F}');\text{MaxPool}(\boldsymbol{F}')\right]\right)\right)=\\&\sigma\left(f^{7\times7}\left(\left[\boldsymbol{F}^{\text{s}}_{\text{Avg}};\boldsymbol{F}^{\text{s}}_{\text{Max}}\right]\right)\right)\end{aligned} \tag{6}$$

式中：σ为Sigmoid函数；$f^{7\times7}$表示7×7的卷积；$\boldsymbol{F}^{\text{s}}_{\text{Avg}}$和$\boldsymbol{F}^{\text{s}}_{\text{Max}}$分别表示全局平均和最大池化。

将生成的权重$\boldsymbol{M}_{\text{s}}$与ECA-Net模块输出特征$\boldsymbol{F}'$相乘，得到最终特征$\boldsymbol{F}_{\text{s}}$，计算公式如下：

$$\boldsymbol{F}_{\text{s}}=\boldsymbol{M}_{\text{s}}\otimes\boldsymbol{F}' \tag{7}$$

通过所提模块增强对掩码的预测精度，预测的掩码通过像素类别分割模块[7]将每个像素点分配到预测的不同类别中，经过像素点对应的类别权重对每个掩码区域预测对应的标签，从而得到整个图像的全景分割。

1.6 模型损失函数

本文铁路场景下全景分割网络模型中，解码器输出后需要用前馈网络对类别和检测框进行预测。计算损失值是将预测结果与GT进行匹配，在网络模型中推测出固定大小的N个预测集。网络模型中，计算损失函数分为2步：第1步，定义检测框和真实框的匹配代价，公式如下：

$$\hat{\nu}=\arg\min\sum_{i=1}^{N}L_{\text{match}}\left(y_i,\hat{y}_{\nu(i)}\right) \tag{8}$$

$$L_{\text{match}}=-1_{\{c_i\neq\varnothing\}}\hat{p}_{\nu(i)}(c_i)+1_{\{c_i\neq\varnothing\}}L_{\text{box}}\left(b_i,\hat{b}_{\nu(i)}\right) \tag{9}$$

式中：y_i为第i个真实标签；$\hat{y}_{\nu(i)}$为第$\nu(i)$个预测框的输出；$1_{\{c_i\neq\varnothing\}}$为布尔函数，当$c_i\neq\varnothing$时为1，否则为0；$\hat{p}_{\nu(i)}(c_i)$表示算法预测的$\nu(i)$个检测框为类别$c_i$的概率；$b_i$和$\hat{b}_{\nu(i)}$分别为第$i$个目标的真实框坐标和算法预测的第$\nu(i)$个检测框坐标；$L_{\text{box}}$由GIOU损失[25]和$L_1$损失组成。

计算式(8)和式(9)可得到$\hat{\nu}$，即预测值与GT的对应关系。

$$L_{\text{box}}\left(b_i,\hat{b}_{\nu(i)}\right)=\lambda_{\text{iou}}L_{\text{GIOU}}\left(b_i,\hat{b}_{\nu(i)}\right)+\lambda_{L_1}L_{L_1}(b_i,\hat{b}_{\nu(i)}) \tag{10}$$

式中：λ_{iou}和λ_{L_1}分别为GIOU损失和L_1损失的权重系数；L_{GIOU}为GIOU损失函数；L_{L_1}为L_1损失函数。

第2步，通过定义的匹配代价可以得到真实框和检测框的最优二部图匹配，根据匹配的结果计算损失函数，如下：

$$L_{\text{Huangarian}}(y,\hat{y})=\sum_{i=1}^{N}\left(-\ln\hat{p}_{\hat{\nu}(i)}(c_i)+1_{\{c_i\neq\varnothing\}}L_{\text{box}}\left(b_i,\hat{b}_{\hat{\nu}(i)}\right)\right) \tag{11}$$

2　实验结果与分析

2.1　实验数据与环境配置

由于目前缺少公开的铁路全景分割数据集，为获取足够的铁路全景分割数据集，本文模型采用微软公司发布的 COCO 数据集进行训练，并结合铁路场景下实地拍摄图像及相关铁路图像，通过 LabelImg 工具制作了铁路场景下的数据集，实现数据集扩充，共计 15 000 张图像，其中，70% 作为训练集，30% 作为测试集。实验使用 PyTorch 深度学习框架，处理器为 Intel(R) Core(TM) i7-9700K CPU@3.6 GHz，显卡采用 NVIDIA GeForce GTX 1660。为验证本文方法的有效性，与文献 [6] 及原始 DETR 方法进行比较，原始 DETR 方法主要用于目标检测，在原始 DETR 模型解码器的输出端增加 mask 检测头即可完成全景分割的任务[7]。本文方法及对比方法均在相同环境进行实验。

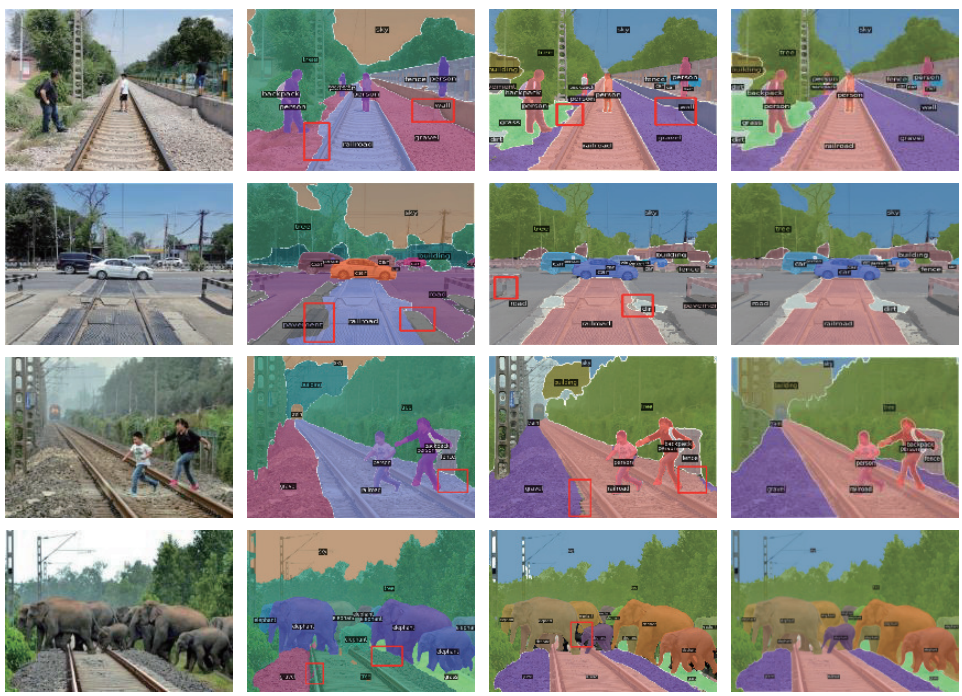
2.2　不同异物侵入铁路场景下全景分割实验

为验证本文方法在不同铁路场景下对铁路异物的全景分割效果，进行真实铁路场景下异物侵限图像的全景分割。选取不同场景下的铁路图像，对于图像中出现的行人、机动车及动物入侵铁路的图像进行全景分割，实验结果如图 9 所示。可以看出，文献 [6] 采用动态内核的深度学习方法对图像中不同类别的物体分割有一定效果，但分割效果差，存在边缘分割模糊及误分割的问题，如图中红色框标注的地方。图 9(c) 为原始 DETR 全景分割方法，通过分割结果可以看出，原方法对图像中特征提取效果不佳，分割后不同类别边缘之间信息判断不准确，分割不完整。图 9(d) 为本文方法分割结果，可以看出，采用本文提出的递归门控增强与金字塔预测方法，较原始 DETR 方法图像分割精度有了大幅度提升，对于图像中不同类别的边缘细节更为清晰。

为进一步检测结果对细节比较，对图 9 检测结果进行局部放大，得到图 10。其中，图 10(a) 为文献 [6] 方法分割结果局部放大图，可以看出，该方法出现了分割边缘不准确、分割模糊的问题，这是因为文献 [6] 方法采用动态内核的方法进行目标匹配与关联，但该方法存在误匹配的问题，导致全景分割存在误分割及边缘模糊的问题。图 10(b) 为原始 DETR 方法检测结果局部放大图，虽然该方法采用 ResNet 特征提取网络和多头注意力机制，在一定程度上提升了图像中对不同类别物体的分割精度，但仍存在信息丢失、边缘不准确的问题。

2.3　山体滑坡铁路全景分割实验

在高速铁路运行过程中，山体滑坡灾害给铁路运行安全带来巨大隐患，在完成真实的不同铁路异物侵限场景下的全景分割实验后，进一步进行铁路山体滑坡中的全景分割实验，实验结果如图 11 所示。可以看出，文献 [6] 方法在全景分割过程中，对不同类型物体之间的边缘轮廓存在分



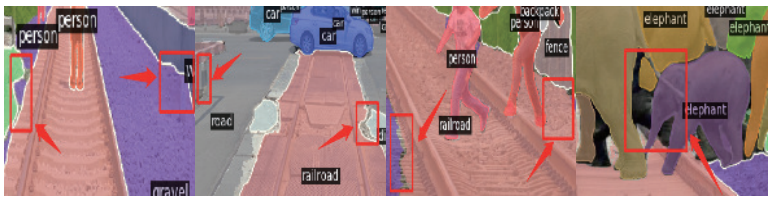
(a) 铁路场景图像　(b) 文献[6]方法　(c) DETR 方法　(d) 本文方法

图 9　不同铁路场景下全景分割实验结果

Fig. 9　Panoramic segmentation experiment results in different railway scenes



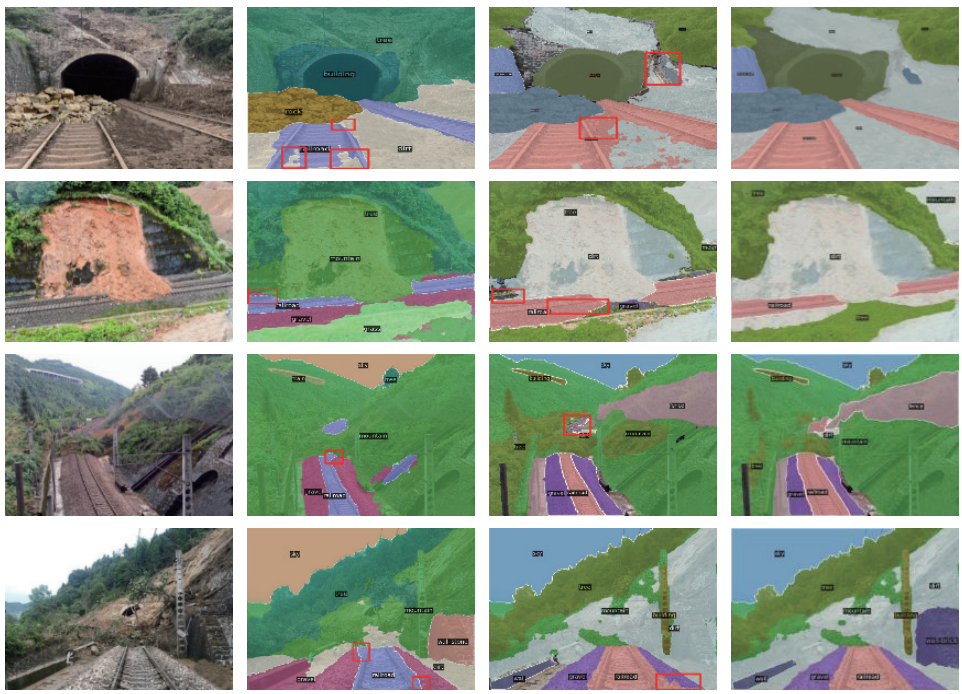
(a) 文献[6] 方法局部放大图



(b) DETR 方法局部放大图

图 10 铁路场景下全景分割局部放大结果

Fig. 10 Partial enlarged diagram for panoramic segmentation in railway scenes



(a) 铁路山体滑坡图像 (b) 文献[6]方法 (c) DETR 方法 (d) 本文方法

图 11 铁路山体滑坡全景分割实验结果

Fig. 11 Experiment results on panoramic segmentation of railway landslides

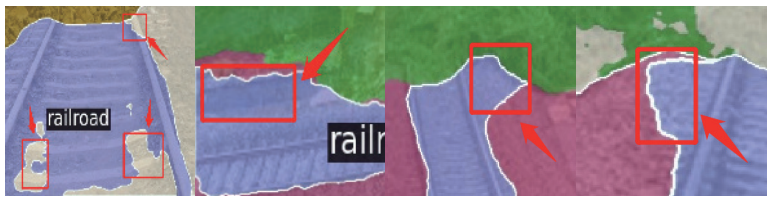
割效果不佳的问题。图 11(c) 为原始 DETR 方法的分割结果，可以看出，该方法存在对不同的目标检测边缘不清晰、分割精度较差的问题。图 11(d) 为本文方法的分割结果，可以看出，本文方法不仅能获得更多特征信息显示出的不同物体，而且对物体之间分割时边缘轮廓的检测更清晰，全景分割质量高于其他方法。

为进一步进行全景分割结果细节比较，对图 11 全景分割结果进行局部放大，得到图 12。其中，图 12(a) 为文献 [6] 方法检测结果局部放大图，可以看出，该方法无法有效地对山体滑坡中不同类别中的边缘进行很好的检测，如图 12(a) 第 1 幅图存在漏分割的问题，其余 3 幅图均出现了误分割的问题。图 12(b) 为原始 DETR 方法的分割结果局部放大图，该方法针对铁路山体滑坡场景中的物体，同样存在特征提取不充分的问题，导致山体滑坡场景下边缘检测遗漏，全景分割质量偏低。

2.4 定量比较分析

为对全景分割性能进行客观评价，采用全景分割质量指标 PQ、前景实例目标评价指标 PQ^{Th} 和背景填充区域质量评价指标 PQ^{St} 进行衡量。上述 3 个指标作为定量比较的标准，值越大，表示方法分割性能越好。比较结果如表 1 所示。

对表 1 数据进行分析，文献 [6] 方法采用动态内核的深度学习方法进行图像的全景分割，其全景分割质量优于原始 DETR 方法。原始 DETR 方法



(a) 文献[6]方法局部放大图



(b) DETR 方法局部放大图

图 12　铁路山体滑坡全景分割局部放大结果

Fig. 12　Partial enlarged map for panoramic segmentation of railway landslides

表 1　不同全景分割方法性能对比

Table 1　Performance comparison of different panoramic segmentation methods　%

方法	PQ	PQ^{Th}	PQ^{St}
文献[6]	48.4	52.6	41.0
DETR	47.2	51.7	40.3
本文	54.6	61.4	46.9

采用端到端的 Transformer 深度学习全景分割方法，该网络模型采用 ResNet 单尺度特征提取网络，对图像特征提取的能力有限，图像中目标特征提取不充分，边缘轮廓不清晰，导致 PQ 仅为 47.2%，前景实例目标评价为 51.7%，背景填充区域质量评价为 40.3%。本文方法定量评价优于对比方法，这是因为：本文方法通过改进多尺度级联 CSP-DarkNet53 特征提取网络和递归门控与类特征增强模块，增强了模型对图像中目标特征的提取能力，并通过多尺度可变形注意力编码器和改进金字塔预测与像素类别分割模块提升了分割质量，本文方法在铁路场景中保持了较高的全景质量，综合性能指标优于对比方法。

2.5　消融实验

为进一步反映本文方法中多尺度级联 CSP-DarkNet53 特征提取网络、递归门控与类特征增强模块、多尺度可变形注意力编码器和改进金字塔预测网络等改进点在铁路场景下全景分割过程中的有效性，采用消融实验进行验证。通过在原 DETR 网络模型中依次嵌入对应的改进模型进行训练及测试，将 PQ 作为指标，验证各改进模块对模型性能提升的作用，消融实验结果如表 2 所示。

从表 2 可以看出，所提铁路场景下全景分割模型进行级联特征提取网络增强提取后，通过改进后的 CSP-DenseBlock 模块增强了对图像中目标特征的提取能力，将网络模型的 PQ 由 47.2% 提升到 48.5%，提升了 1.3%。在此基础上，加入递归门控与类特征增强模块，在训练的过程中增强模型对目标边缘轮廓的关注，将模型的 PQ 提升了 2.6%；在编码器阶段采用多尺度可变形注意力编码器进行优化，PQ 由 51.1% 提升到 52.5%。在预测过程中采用改进的金字塔预测网络进行优化，本文方法的 PQ 达到 54.6%。通过消融实验，直观体现各部分改进模块提升的图像的全景质量，验证本文方法全景分割的有效性。

表 2　模型消融实验比较

Table 2　Comparison of ablation experimental of model

基准	多尺度级联 CSP-DarkNet53 特征提取网络	递归门控与类特征增强模块	多尺度可变形注意力编码器	改进金字塔预测网络	PQ/%
√					47.2
√	√				48.5
√	√	√			51.1
√	√	√	√		52.5
√	√	√	√	√	54.6

2.6　计算复杂度分析

在深度学习全景分割网络模型中，一般通过模型的运算次数来衡量模型计算复杂度。该评价值越小，说明模型计算复杂度越低。不同方法计算复杂度比较如表 3 所示。

从表 3 可以看出，原始 DETR 方法因为采用 ResNet 单尺度特征提取网络对图像进行提取，并且编码器和解码器均采用传统 Transformer 模型，其计算量较低，但结合表 1 可知，DETR 方法易出现边缘检测遗漏的问题，其全景分割质量偏低。文献 [6] 方法采用基于动态内核实现全景分割，但该方法采用互相匹配的策略为每个卷积核分配学习目标，计算复杂度高于 DETR 方法。本文方法的计算量高于对比

表3 模型计算量对比实验结果

Table 3 Comparative experiment results of model calculation

方法	计算量
文献[6]	157 GFLOPs
DETR	93 GFLOPs
本文	223 GFLOPs

注: GFLOPs为模型计算量的标准单位, 表示10亿次浮点运算。

方法, 这是为了克服全景分割方法目标特征提取不充分、边缘轮廓分割模糊等问题, 在骨干提取网络中采用改进的多尺度级联CSP-DarkNet53特征提取网络进行特征提取, 并在编码器中引入多尺度可变形注意力机制, 导致本文方法相比其他方法计算复杂度较高, 但本文方法全景分割保持了较高的全景质量, 综合分割性能指标优于对比方法。

3 结 论

1) 本文提出一种递归门控增强与金字塔预测的铁路全景分割, 设计改进多尺度级联CSP-DarkNet53特征提取网络, 提升对不同尺度的铁路场景目标特征提取能力, 克服了目标特征提取不充分的缺点。

2) 利用类特征注意力机制与递归门控卷积, 提出了递归门控与类特征增强模块, 增强了边缘轮廓的提取和分割的能力, 提升了铁路场景下全景分割的精度。

3) 实验结果表明, 本文方法较原始DETR模型, PQ提升了7.4%, PQ^{Th}提升了9.7%, PQ^{St}提升了6.6%。综上所述, 本文方法对于铁路场景下的图像分割具有更好的性能。

参考文献 (References)

[1] 陈永, 王镇, 周方春. 空间定位与特征泛化增强的铁路异物跟踪检测[J]. 北京航空航天大学学报, 2025, 51(1): 9-18.
CHEN Y, WANG Z, ZHOU F C. Infrared railway foreign objects tracking based on spatial location and feature generalization enhancement[J]. Journal of Beijing University of Aeronautics and Astronautics, 2025, 51(1): 9-18(in Chinese).

[2] LI C, XIE Z Y, QIN Y, et al. A multi-scale image and dynamic candidate region-based automatic detection of foreign targets intruding the railway perimeter[J]. Measurement, 2021, 185: 109853.

[3] 路通, 余祖俊, 郭保青, 等. 网状多尺度与双向通道注意力的铁路场景语义分割[J]. 交通运输系统工程与信息, 2023, 23(2): 233-241.
LU T, YU Z J, GUO B Q, et al. Semantic segmentation of railway scene based on reticulated multi-scale and bidirectional channel attention[J]. Journal of Transportation Systems Engineering and Information Technology, 2023, 23(2): 233-241(in Chinese).

[4] KIRILLOV A, HE K, GIRSHICK R, et al. Panoptic segmentation[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2019: 9396-9405.

[5] KIRILLOV A, GIRSHICK R, HE K M, et al. Panoptic feature pyramid networks[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2019: 6392-6401.

[6] ZHANG W, PANG J, CHEN K, et al. K-Net: towards unified image segmentation[EB/OL]. (2021-10-01)[2023-07-01]. http://arxiv.org/abs/2106.14855?context=cs.AI.

[7] CARION N, MASSA F, SYNNAEVE G, et al. End-to-end object detection with Transformers[C]//Proceedings of the 16th European Conference on Computer Vision. Berlin: Springer, 2020: 213-229.

[8] CHENG B, SCHWING A G, KIRILLOV A. Per-pixel classification is not all you need for semantic segmentation[EB/OL]. (2021-10-31)[2023-07-01]. https://arxiv.org/abs/2107.06278?context=cs.

[9] WANG H Y, ZHU Y K, ADAM H, et al. MaX-DeepLab: end-to-end panoptic segmentation with mask Transformers[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2021: 5459-5470.

[10] CHENG B W, COLLINS M D, ZHU Y K, et al. Panoptic-DeepLab: a simple, strong, and fast baseline for bottom-up panoptic segmentation[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2020: 12472-12482.

[11] CHENG B W, MISRA I, SCHWING A G, et al. Masked-attention mask Transformer for universal image segmentation[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2022: 1280-1289.

[12] YANG T J, COLLINS M D, ZHU Y, et al. DeeperLab: single-shot image parser[C]//Proceedings of theIEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2019: 1902-5093.

[13] 毛琳, 任凤至, 杨大伟, 等. 实例特征深度链式学习全景分割网络[J]. 光学精密工程, 2020, 28(12): 2665.
MAO L, REN F Z, YANG D W, et al. INFNet: deep instance feature chain learning network for panoptic segmentation[J]. Optics and Precision Engineering, 2020, 28(12): 2665(in Chinese).

[14] 宁欣, 田伟娟, 于丽娜, 等. 面向小目标和遮挡目标检测的脑启发CIRA-DETR全推理方法[J]. 计算机学报, 2022, 45(10): 2080-2092.
NING X, TIAN W J, YU L N, et al. Brain-inspired CIRA-DETR full inference model for small and occluded object detection[J]. Chinese Journal of Computers, 2022, 45(10): 2080-2092(in Chinese).

[15] CHEN Y T, CHEN S F. Localizing plucking points of tea leaves using deep convolutional neural networks[J]. Computers and Electronics in Agriculture, 2020, 171: 105298.

[16] ZHAO Z Q, ZHENG P, XU S T, et al. Object detection with deep learning: a review[J]. IEEE Transactions on Neural Networks and Learning Systems, 2019, 30(11): 3212-3232.

[17] HUANG G, LIU Z, VAN DER MAATEN L, et al. Densely connected convolutional networks[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2017: 2261-2269.

[18] WANG Z M, WANG J S, YANG K, et al. Semantic segmentation of high-resolution remote sensing images based on a class feature attention mechanism fused with Deeplabv3+[J]. Computers and Geosciences, 2022, 158: 104969.

[19] RAO Y M, ZHAO W L, TANG Y S, et al. HorNet: efficient high-order spatial interactions with recursive gated convolutions[EB/OL]. (2022-10-11)[2023-07-01]. https://arxiv.org/abs/2207.14284v3.

[20] 冷冰, 冷敏, 常智敏, 等. 基于 Transformer 结构的深度学习模型用于外周血白细胞检测[J]. 仪器仪表学报, 2023, 44(5): 113-120.
LENG B, LENG M, CHANG Z M, et al. Deep learning model based on Transformer architecture for peripheral blood leukocyte detection[J]. Chinese Journal of Scientific Instrument, 2023, 44(5): 113-120(in Chinese).

[21] ZHU X, SU W, LU L, et al. Deformable DETR: deformable Transformers for end-to-end object detection[EB/OL]. (2021-05-18)[2023-07-01]. https://arxiv.org/abs/2010.04159?context=cs.

[22] 李飞, 胡坤, 张勇, 等. 基于混合域注意力 YOLOv4 的输送带纵向撕裂多维度检测[J]. 浙江大学学报 (工学版), 2022, 56(11): 2156-2167.
LI F, HU K, ZHANG Y, et al. Multi-dimensional detection of longitudinal tearing of conveyor belt based on YOLOv4 of hybrid domain attention[J]. Journal of Zhejiang University (Engineering Science), 2022, 56(11): 2156-2167(in Chinese).

[23] WANG Q L, WU B G, ZHU P F, et al. ECA-Net: efficient channel attention for deep convolutional neural networks[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2020: 11531-11539.

[24] YING X Y, WANG Y Q, WANG L G, et al. A stereo attention module for stereo image super-resolution[J]. IEEE Signal Processing Letters, 2020, 27: 496-500.

[25] REZATOFIGHI H, TSOI N, GWAK J, et al. Generalized intersection over union: a metric and a loss for bounding box regression[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2019: 658-666.

Railway panoramic segmentation based on recursive gating enhancement and pyramid prediction

CHEN Yong[1, 2, *], ZHOU Fangchun[1], ZHANG Jiaojiao[1]

(1. School of Electronic and Information Engineering, Lanzhou Jiaotong University, Lanzhou 730070, China;
2. Gansu Provincial Engineering Research Center for Artificial Intelligence and Graphics & Image Processing, Lanzhou 730070, China)

Abstract: A recursive gated enhancement and pyramid prediction railway panoramic segmentation network is proposed to address the issues of insufficient target feature extraction and blurred edge contour segmentation in high-speed railway scene panoramic segmentation. On the basis of the DETR panoramic segmentation model, firstly, an improved multi-scale cascaded CSP-DarkNet53 feature network is constructed to enhance the ability to extract target features from railway scenes of different scales. Then, to improve the ability to extract and segment edge contour information and acquire richer edge feature information, a recursive gating and class feature augmentation module is proposed. Then, deformable attention is introduced into the coding backbone network to further capture context information and reduce the loss of segmentation details. Finally, by improving the pyramid prediction and pixel category segmentation module, the segmentation output of the railway panoramic view is achieved. The suggested approach enhances the original DETR model's panorama segmentation quality index PQ by 7.4%, the foreground instance target's panoramic quality index PQ^{Th} by 9.7%, and the background-filled area's panoramic quality index PQ^{St} by 6.6%, according to experimental data. The proposed method has good performance in panoramic image segmentation in railway scenes, and its subjective evaluation is superior to the comparison method.

Keywords: panoramic segmentation; DETR; recursive gating enhancement; pyramid prediction; high-speed railway

Received: 2023-07-27; **Accepted**: 2023-09-26; **Published Online**: 2023-10-07 17:42

URL: link.cnki.net/urlid/11.2625.V.20231007.1038.001

Foundation items: National Natural Science Foundation of China (62462043,61963023); Key Research and Development Project of Lanzhou Jiaotong University (ZDYF2304)

* **Corresponding author**. E-mail: edukeylab@126.com

http://bhxb.buaa.edu.cn jbuaa@buaa.edu.cn
DOI: 10.13700/j.bh.1001-5965.2023.0504

用于长时视觉跟踪的级联目标漂移判定网络

侯志强[1,2,*]，赵佳鑫[1,2]，陈语[1,2]，马素刚[1,2]，余旺盛[3]，范九伦[1]

(1. 西安邮电大学 计算机学院，西安 710121；2. 西安邮电大学 陕西省网络数据分析与智能处理重点实验室，西安 710121；
3. 空军工程大学 信息与导航学院，西安 710077)

摘 要：针对现有目标漂移判定准则中需要人为选定阈值和判定性能不佳的问题，提出一种自适应选取阈值的级联目标漂移判定网络。通过2个子判定网络的级联设计，判定跟踪结果是否漂移；在所提网络中使用静态模板、长时模板和短时模板联合判定跟踪结果，提高判定的准确性，为使模板适应判定过程中目标的外观变化，设计长短时模板更新策略以保证模板质量；将所提级联目标漂移判定网络联合短时跟踪器TransT与全局重检测方法GlobalTrack，搭建长时视觉跟踪算法TransT_LT。在UAV20L、LaSOT、VOT2018-LT和VOT2020-LT等4个长时视觉跟踪数据集上对所提算法进行性能测试，实验结果表明：所提长时视觉跟踪算法具有优越的长时视觉跟踪性能，特别是在UAV20L数据集上，相较于基准算法，跟踪成功率和精度分别提升了7.7%和10.3%。所提目标漂移判定网络的判定速度为100帧/s，对长时视觉跟踪算法的速度影响不大。

关 键 词：长时视觉跟踪；深度学习；级联目标漂移判定网络；模板更新；多尺度特征融合
中图分类号：TP391.4
文献标志码：A **文章编号**：1001-5965(2025)07-2240-13

视觉目标跟踪作为计算机视觉领域中的重要研究方向，在视频监控、智能交通、人机交互和医学诊断等领域具有广泛应用，是当今的热点研究问题[1-3]。

根据视频帧数可将视觉目标跟踪分为短时视觉跟踪和长时视觉跟踪。短时视觉跟踪中，视频帧数一般在2 000帧内，面临的挑战大多为快速运动、部分遮挡等；而在长时视觉跟踪中，视频帧数往往在2 000帧以上，有时甚至达到数万帧，面临目标全遮挡、尺度变化和长时间出视野等更加困难的挑战。由于长时视觉跟踪更符合实际应用场景，近年来对其研究日益增多。

主流长时视觉跟踪框架通常由基础跟踪器、目标漂移判定准则及全局重检测方法组成[4-7]。基础跟踪器用来对目标进行跟踪，其通常为短时视觉跟踪器，跟踪速度较快，能在目标面临简单挑战时准确跟踪目标。目标漂移判定准则用来对基础跟踪器的跟踪结果进行判定，若判定跟踪结果准确，则在下一帧中继续使用基础跟踪器进行跟踪；若判定跟踪丢失，则调用全局重检测方法在全局范围内搜索目标，以便在目标重新回到视野中时跟踪到目标。目标漂移判定准则作为长时视觉跟踪中的关键一环，对长时视觉跟踪性能具有决定性的影响。如果目标漂移判定准则的判定结果足够准确，则长时视觉跟踪算法就能在跟踪结果正确时一直使用基础跟踪器进行跟踪，保证长时视觉跟踪算法的速度不会受额外全局重检测的影响；当判定跟踪结果丢失时，长时视觉跟踪算法能够立刻调用全局重检测方法，保证了长时视觉跟踪算法的成功率和准确率。

收稿日期：2023-08-02；**录用日期**：2023-11-10；**网络出版时间**：2023-11-23 20:24
网络出版地址：link.cnki.net/urlid/11.2625.V.20231123.0918.001
基金项目：国家自然科学基金(62072370)；陕西省自然科学基金(2023-JC-YB-598)
*通信作者. E-mail：hzq@xupt.edu.cn

引用格式：侯志强，赵佳鑫，陈语，等. 用于长时视觉跟踪的级联目标漂移判定网络[J]. 北京航空航天大学学报，2025，51(7)：2240-2252.
HOU Z Q，ZHAO J X，CHEN Y，et al. Cascaded object drift determination network for long-term visual tracking[J]. Journal of Beijing University of Aeronautics and Astronautics，2025，51(7)：2240-2252 (in Chinese).

近年来出现的长时视觉跟踪算法中, Wang 等[4]在目标感知网络 (target-aware network, TANet) 中通过基础跟踪器的最大响应值是否大于某个阈值来判定目标是否丢失; Bolme 等[8]在最小输出平方和误差 (minimum output sum square error, MOSSE) 方法中提出峰值旁瓣比 (peak to sidelobe ratio, PSR) 来判定跟踪是否失败; Wang 等[9]在基于循环特征映射的大边界目标跟踪 (large margin object tracking with circulant feature map, LMCF) 算法中提出用平均峰值相关能量 (average peak-to-correlation energy, APCE) 来判定跟踪结果是否可靠。上述目标漂移判定准则存在相同的问题: 需要在不同的数据集上选择不同的阈值来达到最优的判定性能。为此, 许多专家和学者提出了无需进行阈值选择的深度判定网络, 对跟踪结果进行判定。Dai 等[5]在基于元更新器的高效长时跟踪 (high-performance long-term tracking with meta-updater, LTMU) 算法中使用实时多域卷积网络 (real-time MDNet, RT-MDNet)[10]作为判定网络, 当判定跟踪失败时, 调用全局重检测方法, 并在下一帧中继续使用全局搜索; Zhang 等[11]在回归和验证网络 MBMD 中使用多域卷积网络 MDNet[12]作为判定网络, 从相似度分数最高的候选框开始, 判定其是否为目标; Hou 等[13]提出了一种轻量级的目标漂移判定网络, 当判定跟踪失败时, 在新的搜索区域内重新搜索目标。实验结果表明, 这些深度判定网络虽然不需要在不同数据集上进行阈值选取, 但其判定性能却不如需要阈值选取的判定准则。

为解决该问题, 本文对目标漂移判定准则进行深入研究, 提出一种级联目标漂移判定网络, 并在此基础上构建一种基于级联目标漂移判定网络的长时视觉跟踪算法。

级联目标漂移判定网络由 2 个子判定网络构成。其中, 第 1 级目标漂移判定网络在大部分情况下能够准确判定跟踪结果是否丢失, 但在目标发生尺度变化时判定性能不佳。因此, 设计了第 2 级目标漂移判定网络, 在其中引入金字塔池化模块 (pyramid pool module, PPM) 和全局上下文模块 (global context network, GC-Net), 当目标发生尺度变化时能够准确判定跟踪结果是否丢失。为提高判定的准确性, 采用静态模板、长时模板和短时模板联合判定跟踪结果是否漂移, 并设计了长短时模板更新策略, 以保证判定过程中级联目标漂移判定网络所使用的模板质量。本文网络能够在不同的数据集上自适应地选取阈值, 避免了人为选定阈值的困难, 且判定性能超过了大部分现有的目标漂移判定准则。本文以 TransT[14]作为基础跟踪器, 联合全局重检测方法 GlobalTrack[15], 搭建了长时视觉跟踪算法 TransT_LT, 在长时视觉跟踪数据集 UAV-20L[16]、LaSOT[17]、VOT2018-LT[18]和 VOT2020-LT[19]上进行测试和分析, 实验结果表明, 相较于基础跟踪器 TransT, 长时视觉跟踪性能有了显著提升。

1　本文算法

本文使用的长时视觉跟踪框架主要包括短时目标跟踪、级联目标漂移判定网络和全局重检测 3 个部分。图 1 为长时视觉跟踪算法的整体框架。

本文提出了一个级联目标漂移判定网络, 基于所提判定网络, 以 TransT[14]为短时目标跟踪算法, 以 GlobalTrack[15]为全局重检测方法, 根据图 1 所示的长时视觉跟踪算法框架, 搭建了一个新的长时视觉跟踪算法 TransT_LT。

1.1　级联目标漂移判定网络

本文提出的级联目标漂移判定网络由 2 个子判定网络构成。在第 1 级目标漂移判定网络中, 设计了一个基于 SENet[20]的局部特征融合模块 (local feature fusion module, LFFM), 以增强模板和跟踪结果局部特征的表达能力。通过实验发现, 第 1 级目标漂移判定网络在大部分情况下能准确判定跟踪结果是否丢失, 但在目标发生尺度变化时判定性能不佳。因此, 设计了第 2 级目标漂移判定网络。在该网络中, 提出了一个基于 PPM[21]和 GC-Net[22]的多尺度全局特征融合模块 (multi-scale global feature fusion module, MGFFM), 使第 2 级目标漂移判定网络更专注于提取模板和跟踪结果的多尺度全局特征, 从而提高目标发生尺度变化时判定的准确性。此外, 所提级联目标漂移判定网络使用 3 个模板联合判定跟踪结果是否准确, 有效提升了判定性能。为保证判定过程中的模板质量, 提出了一个有效的长短时模板更新策略。具体的级联目标漂移判定网络如图 2 所示。

1.1.1　第 1 级目标漂移判定网络

如图 2 所示, 将静态模板、由基础跟踪器得到的初始结果、长时模板和短时模板共同输入权重共享的卷积神经网络 VGG16[23]中, 提取深层特征 $\boldsymbol{X}_1$、$\boldsymbol{Y}_1$、$\boldsymbol{Z}_1$和 $\boldsymbol{K}_1$。其中, 静态模板、长时模板和短时模板使用第 1 帧的 GroundTruth 进行初始化。将得到的特征 $\boldsymbol{X}_1$、$\boldsymbol{Y}_1$、$\boldsymbol{Z}_1$和 $\boldsymbol{K}_1$分别经过 LFFM 模块, 得到 $\boldsymbol{X}_2$、$\boldsymbol{Y}_2$、$\boldsymbol{Z}_2$和 $\boldsymbol{K}_2$(由于静态模板、初始结果、长时模板和短时模板 4 路的处理步骤相同, 下面以静态模板为例说明 LFFM 模块的处理过程)。

如图 3 所示, 首先, 通过 SENet 提取图像块的关键特征, 具体如下:

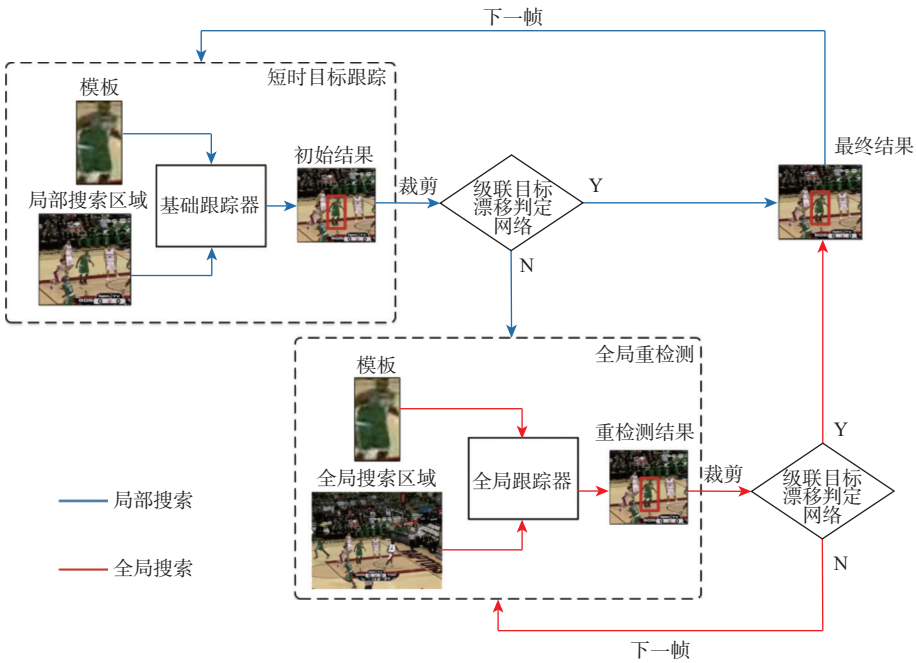


图 1 长时视觉跟踪算法框架

Fig. 1 Framework of long-term visual tracking algorithm

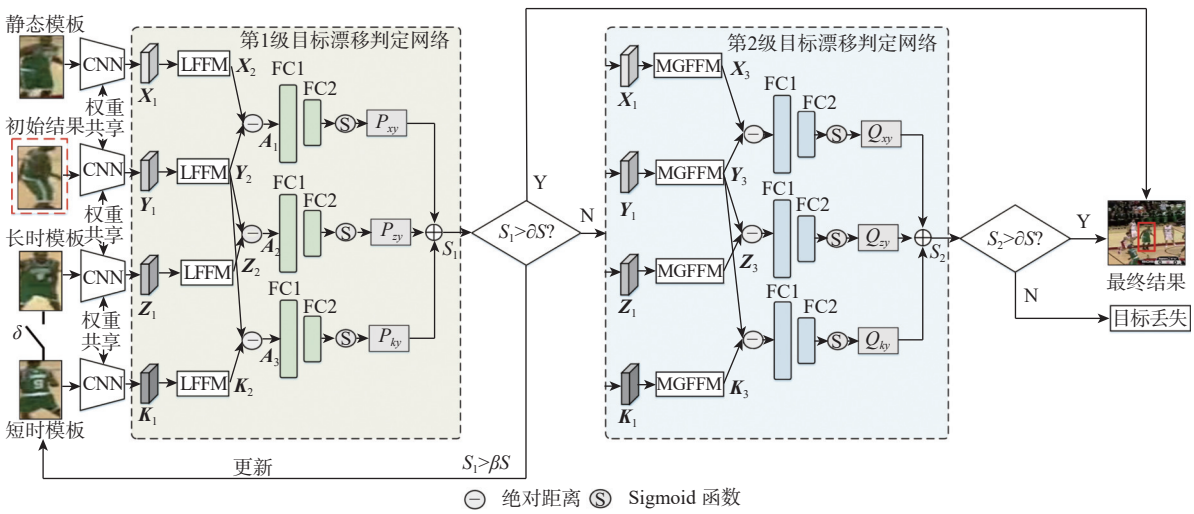


图 2 级联目标漂移判定网络结构

Fig. 2 Structure of cascaded object drift determination network

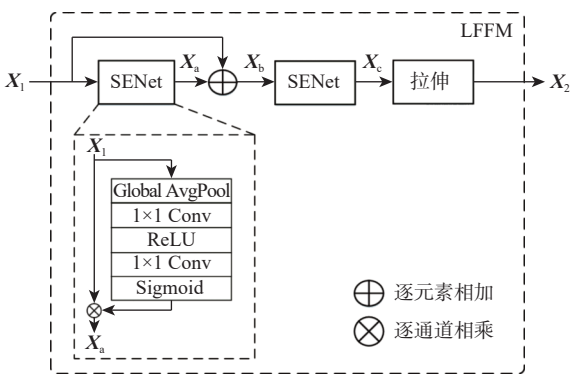


图 3 LFFM 模块示意

Fig. 3 Schematic diagram of LFFM module

$$\boldsymbol{X}_{\mathrm{a}}=\mathrm{SE}(\boldsymbol{X}_1) \tag{1}$$

式中：$\boldsymbol{X}_1$表示经过卷积神经网络提取的深层特征；SE表示图 3 所示的通道注意力 SENet[20]。

由于特征复用能使得到的特征更加丰富，设置残差连接，将$\boldsymbol{X}_{\mathrm{a}}$与$\boldsymbol{X}_1$相加，得到$\boldsymbol{X}_{\mathrm{b}}$。将融合后的特征$\boldsymbol{X}_{\mathrm{b}}$再经过 SENet，得到突出关键特征的$\boldsymbol{X}_{\mathrm{c}}$。

为便于计算特征间的距离，将$\boldsymbol{X}_{\mathrm{c}}$经过拉伸操作，得到 LFFM 模块的输出$\boldsymbol{X}_2$，具体如下：

$$\boldsymbol{X}_2=\mathrm{Flatten}(\boldsymbol{X}_{\mathrm{c}}) \tag{2}$$

式中：$\boldsymbol{X}_2$为静态模板得到的一维向量，向量内共有 4 608 个元素；Flatten 表示拉伸操作。

同时，如图 2 所示，其他 3 路分别得到了 LFFM 模块的输出$\boldsymbol{Y}_2$、$\boldsymbol{Z}_2$和$\boldsymbol{K}_2$，表示初始结果、长时模板和短时模板得到的一维向量。

其次，分别计算$\boldsymbol{Y}_2$与$\boldsymbol{X}_2$、$\boldsymbol{Y}_2$与$\boldsymbol{Z}_2$及$\boldsymbol{Y}_2$与$\boldsymbol{K}_2$之间的

绝对距离$\boldsymbol{A}_1$、$\boldsymbol{A}_2$和$\boldsymbol{A}_3$。其中, $\boldsymbol{A}_1=[\varepsilon_1,\cdots,\varepsilon_i,\cdots,\varepsilon_C]$, $\boldsymbol{A}_2=[\gamma_1,\cdots,\gamma_i,\cdots,\gamma_C]$, $\boldsymbol{A}_3=[\lambda_1,\cdots,\lambda_i,\cdots,\lambda_C]$, $\varepsilon_i=\left|X_2^i-Y_2^i\right|_1$表示$\boldsymbol{A}_1$中的第$i$个元素, $\gamma_i=\left|Z_2^i-Y_2^i\right|_1$表示$\boldsymbol{A}_2$中的第$i$个元素, $\lambda_i=\left|K_2^i-Y_2^i\right|_1$表示$\boldsymbol{A}_3$中的第$i$个元素, X_2^i、Y_2^i、Z_2^i和K_2^i分别表示一维向量$\boldsymbol{X}_2$、$\boldsymbol{Y}_2$、$\boldsymbol{Z}_2$和$\boldsymbol{K}_2$中的第i个元素。

然后, 将得到的特征$\boldsymbol{A}_1$、$\boldsymbol{A}_2$和$\boldsymbol{A}_3$分别经过 2 个全连接层 (FC1、FC2), 在第 2 个全连接层中将结果映射到一个节点上, 经过 Sigmoid 函数, 将网络输出映射到 [0,1] 之间, 分别得到当前帧与 3 个模板之间的相似度分数P_{xy}、P_{zy}和P_{ky}。将得到的相似度分数相加, 得到当前帧的相似度分数总和S_1。具体如下:

$$S_1=P_{xy}+P_{zy}+P_{ky} \tag{3}$$

式中: P_{xy}、P_{zy}和P_{ky}分别为当前帧初始结果与静态模板、长时模板和短时模板之间的相似度分数。

由于在目标跟踪领域中, 第 1 帧目标是由人工标注的, 跟踪过程从第 2 帧开始, 因此, 第 2 帧的跟踪结果往往是非常可靠的[24]。为在进行目标漂移判定时有对比参考值, 将第 2 帧的初始跟踪结果y_0作为参考来计算基准相似度。计算y_0与 3 个模板之间的相似度分数总和记为基准相似度S, 具体如下:

$$S=P_{xy_0}+P_{zy_0}+P_{ky_0} \tag{4}$$

式中: P_{xy_0}、P_{zy_0}和P_{ky_0}分别为第 2 帧的初始跟踪结果与静态模板、长时模板和短时模板之间的相似度分数。

最后, 将当前帧的相似度分数总和S_1与基准相似度S进行比较, 判定跟踪结果是否准确, 具体如下:

$$S_1>\partial S \tag{5}$$

式中: ∂为比例系数, 根据表 1, 最终确定∂为 0.6。即若当前帧的相似度分数总和S_1满足式(5), 则认为跟踪结果准确, 否则, 输入第 2 级子判定网络中再次判定。

1.1.2　第 2 级目标漂移判定网络

实验发现, 第 1 级目标漂移判定网络能在大部分情况下保持良好的判定性能, 但当目标尺度发生变化时, 判定性能不佳。为使本文所提级联目标漂移判定网络能在目标发生尺度变化时仍保持较好的判定性能, 在第 2 级目标漂移判定网络中设计了一个基于 PPM 模块[21] 和 GC-Net 模块[22] 的 MGFFM 模块。

如图 2 所示, 当第 1 级目标漂移判定网络判定跟踪结果不准确时, 将 VGG16[23] 提取的深层特征$\boldsymbol{X}_1$、$\boldsymbol{Y}_1$、$\boldsymbol{Z}_1$和$\boldsymbol{K}_1$输入第 2 级目标漂移判定网络中, 分别经过 MGFFM 模块得到$\boldsymbol{X}_3$、$\boldsymbol{Y}_3$、$\boldsymbol{Z}_3$和$\boldsymbol{K}_3$(由于静态模板、初始结果、长时模板和短时模板 4 路的处理步骤相同, 下面以静态模板为例说明 MGFFM 模块的处理过程)。

表 1　不同∂与β参数值对长时视觉跟踪性能的影响

Table 1　Impact of different parameter values on long-term visual tracking performance for ∂ and β

(∂,β)	UAV20L		VOT2020-LT		
	成功率	精确度	F分数	精确度	召回率
(0.5,0.6)	0.635	0.826	0.648	0.693	0.608
(0.5,0.7)	0.649	0.839	0.660	0.703	0.623
(0.5,0.8)	0.654	0.862	0.661	0.711	0.618
(0.5,0.9)	0.666	0.866	0.664	0.653	0.675
(0.6,0.7)	0.671	0.870	0.664	0.709	0.625
(0.6,0.8)	0.672	0.872	0.665	0.653	**0.677**
(0.6,0.9)	**0.693**	**0.898**	**0.671**	**0.715**	0.632
(0.7,0.8)	0.673	0.870	0.665	0.713	0.624
(0.7,0.9)	0.672	0.875	0.665	0.711	0.622

注: 加粗字体表示最优值。

如图 4 所示, 首先, 将$\boldsymbol{X}_1$通过 PPM 模块和 GC-Net 模块, 分别得到图像块的多尺度特征$\boldsymbol{M}_{\mathrm{a}}$和全局上下文特征$\boldsymbol{G}_{\mathrm{a}}$, 具体如下:

$$\boldsymbol{M}_{\mathrm{a}}=\mathrm{PPM}(\boldsymbol{X}_1) \tag{6}$$

$$\boldsymbol{G}_{\mathrm{a}}=\mathrm{GC}(\boldsymbol{X}_1) \tag{7}$$

式中: $\boldsymbol{X}_1$为静态模板经过 VGG16 提取的深度特征; PPM表示图 4 所示改进的 PPM 模块; GC表示图 4 所示的 GC-Net 模块。

为使得到的特征更加丰富, 将$\boldsymbol{X}_1$、$\boldsymbol{M}_{\mathrm{a}}$和$\boldsymbol{G}_{\mathrm{a}}$相加融合, 得到$\boldsymbol{C}_{\mathrm{a}}$, 具体如下:

$$\boldsymbol{C}_{\mathrm{a}}=\boldsymbol{X}_1+\boldsymbol{M}_{\mathrm{a}}+\boldsymbol{G}_{\mathrm{a}} \tag{8}$$

将融合后的特征$\boldsymbol{C}_{\mathrm{a}}$经过 GC-Net[22] 进行全局上下文特征的提取, 使第 2 级目标漂移判定网络更聚焦于图像块的多尺度特征和全局特征, 得到$\boldsymbol{C}_{\mathrm{b}}$。将$\boldsymbol{C}_{\mathrm{b}}$经过拉伸操作, 得到 MGFFM 模块的输出$\boldsymbol{X}_3$, 具体如下:

$$\boldsymbol{X}_3=\mathrm{Flatten}(\boldsymbol{C}_{\mathrm{b}}) \tag{9}$$

式中: $\boldsymbol{X}_3$为静态模板在第 2 级目标漂移判定网络中得到的一维向量, 向量内共有 4 608 个元素。

然后, 如图 2 所示, 经过与第 1 级目标漂移判定网络相同的步骤, 得到第 2 级目标漂移判定网络中初始结果与静态模板、长时模板和短时模板之间的相似度分数Q_{xy}、Q_{zy}和Q_{ky}。将得到的相似度分数相加, 得到第 2 级目标漂移判定网络输出的相似度总和S_2, 具体如下:

$$S_2=Q_{xy}+Q_{zy}+Q_{ky} \tag{10}$$

最后, 将当前帧在第 2 级目标漂移判定网络中

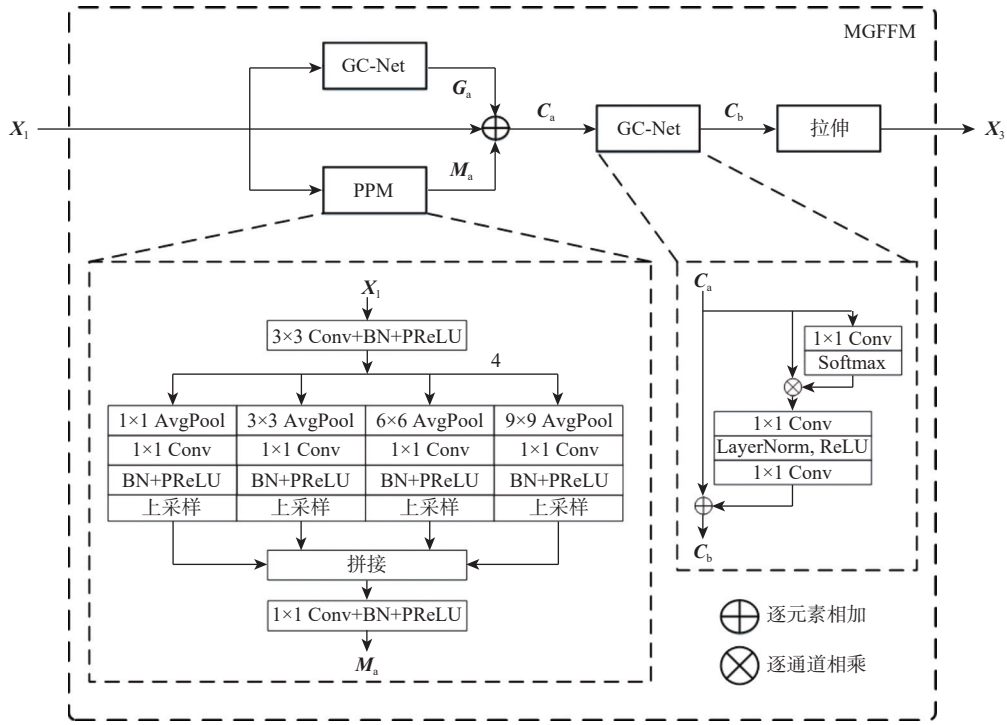


图4 MGFFM模块示意图

Fig. 4 Schematic diagram of MGFFM module

的相似度分数总和S_2与基准相似度S进行比较，判定跟踪结果是否准确，具体如下：

$$S_2 > \partial S \tag{11}$$

式中：∂=0.6。若当前帧的相似度分数总和S_2满足式(11)，则认为跟踪结果准确，否则，认为当前帧跟踪结果错误，使用全局重检测方法GlobalTrack[15]对当前帧重新进行跟踪。

1.1.3 长短时模板更新策略

在本文提出的级联目标漂移判定网络中，使用静态模板、长时模板和短时模板联合判定跟踪结果是否准确，有效提升了网络的判定性能。其中，静态模板始终保持不变，而长短时模板不断更新，选取短时模板的更新条件与长时模板的更新时间间隔十分重要。如果短时模板的更新条件不准确，就有可能将错误的跟踪结果引入短时模板中；如果长时模板的更新时间间隔不合适，网络将不能适应目标的外观变化。上述2种情况都会影响后续多帧的判定结果，导致长时视觉跟踪算法的性能有所下降。

为选择跟踪过程中质量较高的跟踪结果更新短时模板，提出短时模板的更新条件，具体如下：

$$S_1 > \beta S \tag{12}$$

如果当前帧跟踪结果的相似度分数总和S_1满足式(12)，则使用当前帧的跟踪结果更新短时模板；否则，不更新短时模板。其中，β为比例系数。根据实际应用的要求进行仔细分析，认为模板更新的条件应该比判定跟踪结果准确的条件更加严谨，才能选出高质量的模板，从而提高所提级联目标漂移判定网络的判定性能。因此，式(12)中的比例系数β要大于式(5)中判定跟踪结果是否准确的比例系数∂。此外，根据表1，最终确定β为0.9。

为确定图2中长时模板更新的时间间隔参数δ，选定不同的时间间隔在UAV20L[16]和VOT2020-LT[19]数据集上进行实验，结果如表2所示。当δ设为100帧时，长时视觉跟踪算法的性能达到最优。

综上，本文所提长短时模板更新策略如下：若当前帧跟踪结果的相似度分数总和满足$S_1 > \beta S$，则用当前帧跟踪结果更新短时模板；若跟踪时间间隔达到100帧，则用当前的短时模板更新长时模板。实验结果表明，上述模板更新条件能准确选择长短时模板，保证长短时模板的质量，有效提升所提级联目标漂移判定网络的判定性能。

1.2 训练数据集

为提高本文所提级联目标漂移判定网络的准确性，需要大量的训练数据。本文将OTB2015[25]数据集中与GroundTruth的IoU值大于0.7的图像块作为正样本，与GroundTruth的IoU值小于0.3的图像块作为负样本。为保证训练的平衡，获得同样数量的正负样本各67 549个。部分正负样本如图5所示。

为对网络进行训练，将训练数据按照9∶1的比例划分为训练集和验证集，其中，训练集中正负样本各60 794个，验证集中正负样本各6 755个。

表 2　长时模板更新时间间隔参数δ对长时视觉跟踪性能的影响

Table 2　Performance impact of long-term template update interval parameter on long-term visual tracking

δ/帧	UAV20L		VOT2020-LT		
	成功率	精确度	F分数	精确度	召回率
50	0.666	0.859	0.665	0.713	0.623
100	**0.693**	**0.898**	**0.671**	**0.715**	**0.632**
150	0.689	0.893	0.669	0.711	0.631
200	0.686	0.888	0.664	0.713	0.622
250	0.685	0.886	0.661	0.712	0.618
300	0.672	0.866	0.660	**0.715**	0.613

注: 加粗字体表示最优值。



图 5　部分正负样本

Fig. 5　Partial positive and negative sample plots

1.3　长时视觉跟踪算法的构建

将 TransT[14] 作为基础跟踪器, 联合本文所提级联目标漂移判定网络与全局重检测方法 GlobalTrack[15], 搭建了一个新的长时视觉跟踪算法 TransT_LT。长时视觉跟踪算法的流程如图 1 所示。

算法具体步骤如下:

步骤 1　使用基础跟踪器 TransT 对目标进行跟踪, 得到跟踪的初始结果。

步骤 2　将初始结果输入本文所提级联目标漂移判定网络中, 利用网络中的静态模板、长时模板和短时模板联合判定初始结果是否准确。

步骤 3　当判定初始结果准确时, 将初始结果视为最终结果, 并在下一帧中继续使用 TransT 对目标进行局部搜索; 否则, 使用 GlobalTrack 在全局搜索区域内对当前帧重新进行跟踪, 将输出的重检测结果输入所提判定网络中进行判断。当判定重检测结果准确时, 将重检测结果视为最终结果, 并在下一帧中使用 TransT 进行局部搜索; 当判定重检测结果不准确时, 在下一帧中继续使用 GlobalTrack 在全局搜索区域内搜索目标, 直到目标被成功跟踪时, 在下一帧使用 TransT 进行局部搜索。

此外, 在判定过程中, 若当前帧跟踪结果满足短时模板更新条件, 则及时更新短时模板, 当达到一定时间间隔后更新长时模板。

2　实验结果与分析

2.1　实验平台及参数设置

实验平台为 Ubuntu16.04, 处理器为 Intel Corei5-8400 2.80 GHz, GPU 为 GeForce RTX 1080Ti。训练图像大小裁剪成 105×105, 优化算法使用随机梯度下降法, 损失函数使用交叉熵损失函数, 批量大小为 8, 初始学习率为 0.01, 训练 100 批次。

2.2　实验数据集

为验证本文提出的级联目标漂移判定网络的有效性, 在 4 个长时视觉跟踪数据集 UAV20L[16]、LaSOT[17]、VOT2018-LT[18] 和 VOT2020-LT[19] 上进行各项实验。

1) UAV20L 数据集是 UAV123[16] 中 20 个时间跨度大、目标视野变化剧烈、目标形变更大的视频序列, 有利于评估跟踪算法在长时视觉跟踪环境下的有效性和鲁棒性。UAV20L 数据集将视觉跟踪中存在的各种挑战归纳为 12 种, 分别对应视频的不同属性, 可有效评估跟踪算法的综合性能。该数据集采用跟踪成功率和跟踪精确度比较各跟踪算法的性能。

2) LaSOT[17] 数据集有 1 400 个视频序列, 每个视频平均有 2 512 帧, 最短的视频有 1 000 帧, 最长的包含 11 397 帧。LaSOT 按照精确度、归一化精确度和成功率来比较不同跟踪算法的性能。

3) VOT2018-LT[18] 数据集包括 35 个不同对象(如人、汽车、自行车和动物)的序列, 总帧长为 146 847 帧, 分辨率范围在 290×217 和 1 280×720 之间。该数据集充分考虑了目标消失的情况, 其中, 每个序列平均包含 12 次长期目标消失, 每个消失平均持续 40 帧。该数据集的准确度评估使用 F 分数来衡量给定长期跟踪算法的性能。

4) VOT2020-LT[19] 在 VOT2018-LT 的基础上引入了 15 个更困难的视频和不常见的目标(如船、牛和降落伞)。VOT2020-LT 数据集的准确度评估指标与 VOT2018-LT 数据集一致, 根据 F 分数对不同的跟踪算法进行排名。

2.3　与其他目标漂移判定准则的对比

为验证本文所提级联目标漂移判定网络的优越性, 在相同的基础跟踪器、全局重检测方法及长时视觉跟踪框架下, 将所提级联目标漂移判定网络替换为不同的目标漂移判定准则, 如 Fmax[4]、PSR[8]、APCE[9]、RT-MDNet[10] 及 MDNet[12], 在 UAV20L[16]

和 VOT2020-LT[19] 数据集上进行实验与分析，实验结果如表 3 和表 4 所示。

实验结果表明，Fmax、PSR 和 APCE 这类基于最大响应值的判定准则，在不同的数据集上需要选定不同的阈值来达到最优的判定性能；RT-MDNet 和 MDNet 这类基于深度学习的判定网络，虽不需要选定不同的阈值，但其判定性能却不及 APCE 和 PSR。本文所提级联目标漂移判定网络，相较于这 2 类目标漂移判定准则，不仅像 RT-MDNet 和 MDNet 一样不需要在不同的数据集上选定不同的阈值，并且判定性能也超过了 APCE 和 PSR。

表 3 不同判定准则在 2 个数据集上的最优阈值

Table 3 Optimal thresholds for different determination criteria on two datasets

判定准则	UAV20L	VOT2020-LT
Fmax[4]	>0.9	>0.95
PSR[8]	>3	>5
APCE[9]	<17	<19
RT-MDNet[10]	/	/
MDNet[12]	/	/
本文	/	/

注："/" 表示不需要选取阈值。

表 4 不同判定准则的最优阈值在 2 个数据集上的性能

Table 4 Performance of the optimal thresholds of different determination criteria on two datasets

判定准则	UAV20L		VOT2020-LT		
	成功率	精确度	*F*分数	精确度	召回率
Fmax[4]	0.632	0.817	0.644	0.696	0.599
PSR[8]	0.639	0.824	0.648	0.704	0.600
APCE[9]	0.641	0.828	0.651	0.709	0.602
RT-MDNet[10]	0.624	0.806	0.641	0.712	0.590
MDNet[12]	0.630	0.813	0.646	0.698	0.601
本文	**0.693**	**0.898**	**0.671**	**0.715**	**0.632**

注：加粗字体表示最优值。

2.4 消融实验

在 UAV20L[16] 和 VOT2020-LT[19] 数据集上进行消融实验，验证本文所提级联目标漂移判定网络中各个模块的有效性，实验结果如表 5 所示。其中，3 模板判定 (three template determination, TTD) 表示在级联目标漂移判定网络中使用静态模板、长时模板和短时模板联合判定跟踪结果（此时，长时模板的更新时间间隔设为 50 帧，还未确定为最优值 100 帧）；长时模板更新策略 (long-term template update strategy, LTUS) 表示将长时模板更新时间间隔设为最优值 100 帧；级联漂移判定网络 (cascaded drift determination network, CDDN) 表示构建的第 2 级目标漂移判定网络。当级联的 2 个目标漂移判定网络都判定初始结果不准确时，才认为跟踪失败。实验结果表明，本文所提各个模块都提升了长时视觉跟踪算法 TransT_LT 的跟踪性能。当使用静态模板、长时模板和短时模板共同判定，长时模板更新时间间隔设为 100 帧，级联第 2 级目标漂移判定网络时，TransT_LT 的跟踪性能达到最优。

表 5 消融实验结果

Table 5 Ablation experiment results

TransT	TTD	LTUS	CDDN	UAV20L		VOT2020-LT		
				成功率	精确度	*F*分数	精确度	召回率
√				0.616	0.795	0.638	0.682	0.599
√	√			0.658	0.857	0.657	0.669	0.645
√	√	√		0.688	0.892	0.668	0.660	0.676
√	√	√	√	0.693	0.898	0.671	0.715	0.632

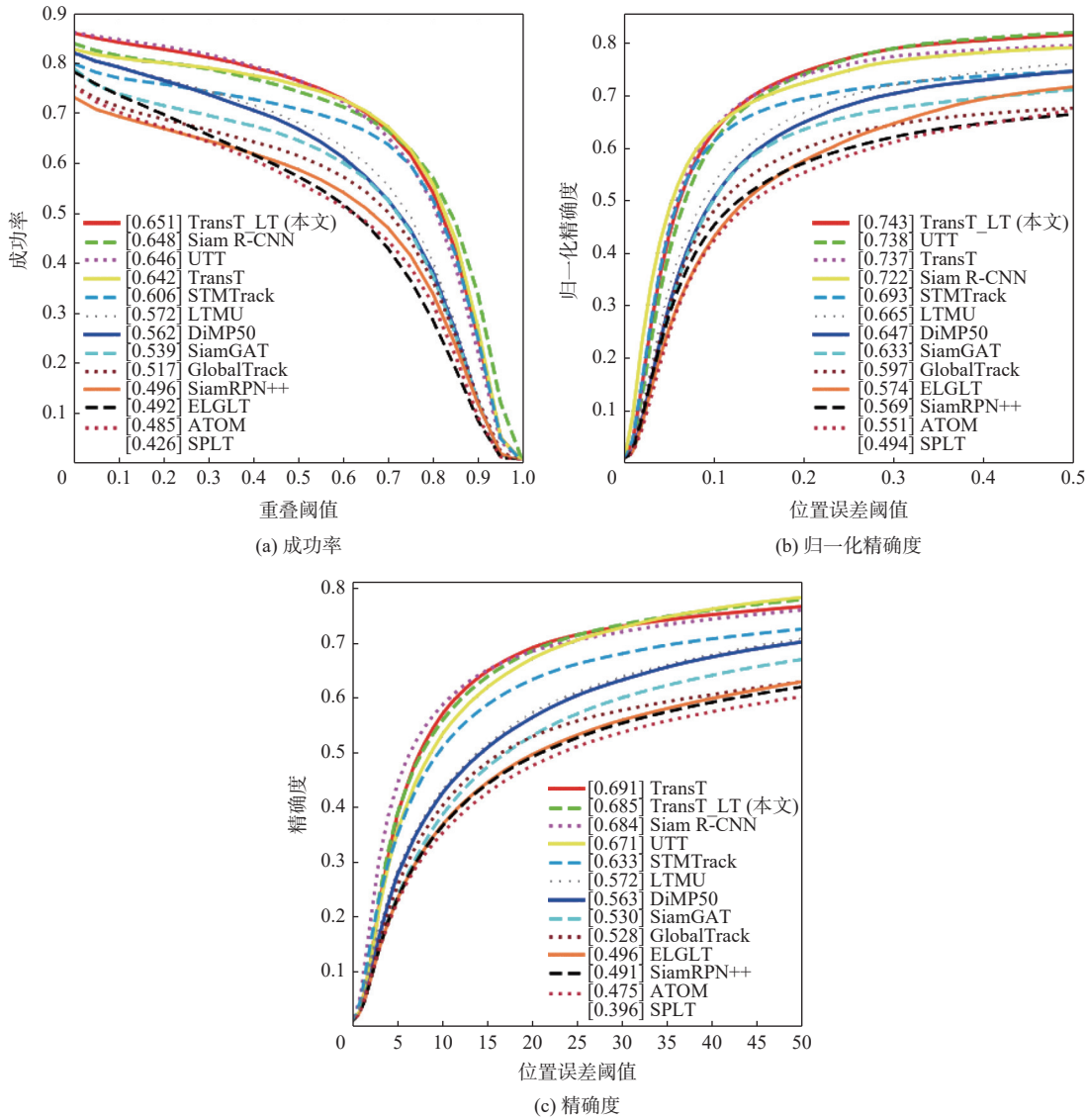
2.5 定量分析

2.5.1 LaSOT 数据集

将本文算法与 TransT 及其他 11 个跟踪算法进行比较，比较算法包括 LTMU[5]、用于快速长时跟踪的高效局部全局搜索 (effective local and global search for fast long-term tracking, ELGLT) 方法[6]、GlobalTrack[15]、用于实时和鲁棒长时跟踪的略读和精读框架 SPLT[26]、基于重检测的视觉跟踪 Siam R-CNN 算法[27]、用于目标跟踪的统一 Transformer 跟踪器 (unified Transformer tracker for object tracking, UTT)[28]、基于时空记忆网络的无模板视觉跟踪 (template-free visual tracking with space-time memory network, STMTrack) 方法[29]、DiMP50[30]、SiamGAT[31]、SiamRPN++[32] 和基于重叠最大化的精确跟踪 (accurate tracking by overlap maximization, ATOM) 方法[33]。图 6 展示了本文所提长时视觉跟踪算法 TransT_LT 及各视觉跟踪算法在 LaSOT[17] 数据集上的跟踪成功率、跟踪归一化精确度和跟踪精确度曲线对比。可以看出，本文算法相较于基准算法 TransT 在成功率上提升了 0.009，提升并不明显。出现类似情况的还有 LTMU 算法[5] 和 ELGLT 算法[6]，其在长时视觉跟踪数据集上的性能都非常突出，但在 LaSOT 数据集上的成功率也都较低，分别为 0.572 和 0.492。这是由于 LaSOT 数据集中的目标面临的挑战更多是部分遮挡、运动模糊、亮度变化等，并没有频繁出现长时视觉跟踪中所面临的出视野、全遮挡等挑战。

2.5.2 UAV20L 数据集

将本文算法与 TransT 及其他 11 个跟踪算法进行比较，比较算法包括 GlobalTrack[15]、STMTrack[29]、SuperDiMP[30]、DiMP50[30]、SiamRPN++[32]、ATOM[33]、TrDiMP[34]、用于视觉跟踪的空间正则化相关滤波器 SRDCF[35]、基于核的结构化输出跟踪器 (structured



(a) 成功率　(b) 归一化精确度

(c) 精确度

图 6　LaSOT 跟踪结果评估曲线

Fig. 6　LaSOT tracking result evaluation curves

output tracking with kernels, Struck)[36]、判别尺度空间跟踪器 (discriminative scale space tracking, DSST)[37] 和跟踪-学习-检测 (tracking-learning-detection, TLD) 方法[38]。图 7 展示了本文算法与对比的各视觉跟踪算法在 UAV20L[16] 数据集上的跟踪成功率和跟踪精确度曲线对比。可以看出, 本文构建的长时视觉跟踪算法 TransT_LT 的所有评估指标均有提升, 相比于基准算法 TransT, 跟踪成功率和跟踪精确度分别提升了 0.077 和 0.103。

针对 UAV20L 数据集中包含的 12 个不同的视频属性, 图 7(c) 定量展示了本文所提长时视觉跟踪算法 TransT_LT 与基准算法 TransT 及各对比算法在应对复杂跟踪因素下的平均成功率。

2.5.3　VOT2018-LT 数据集

表 6 为本文所提长时视觉跟踪算法 TransT_LT 及各视觉跟踪算法在 VOT2018-LT[18] 数据集上性能比较结果。

可以看出, 本文构建的长时视觉跟踪算法相较于基准算法 TransT, 在 F 分数上提升了 0.023。相较于使用 RT-MDNet 作为判定网络的 LGST, 本文算法在 F 分数上高出 0.063; 相较于使用 MDNet 作为判定网络的 MBMD, 本文算法在 F 分数上高出 0.083; 相较于使用最大响应值作为判定准则的 TANet, 本文算法在 F 分数上高出 0.107。

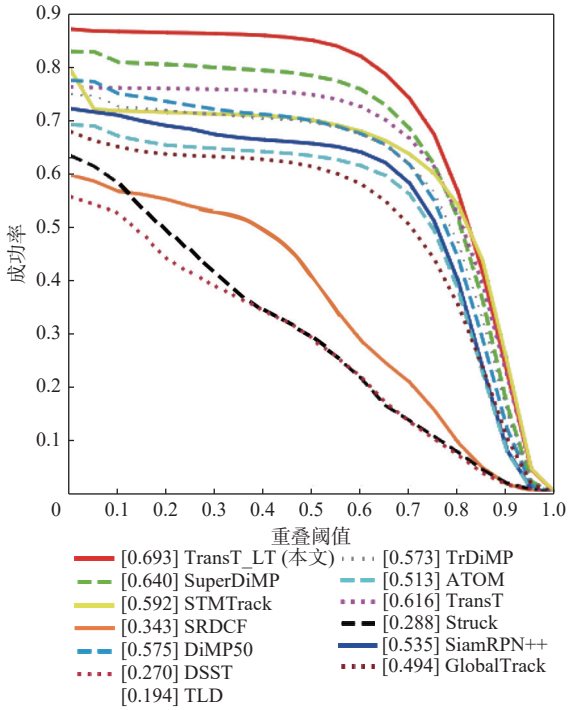
2.5.4　VOT2020-LT 数据集

表 7 为本文所提长时视觉跟踪算法 TransT_LT 及各视觉跟踪算法在 VOT2020-LT[19] 数据集上的性能比较结果。

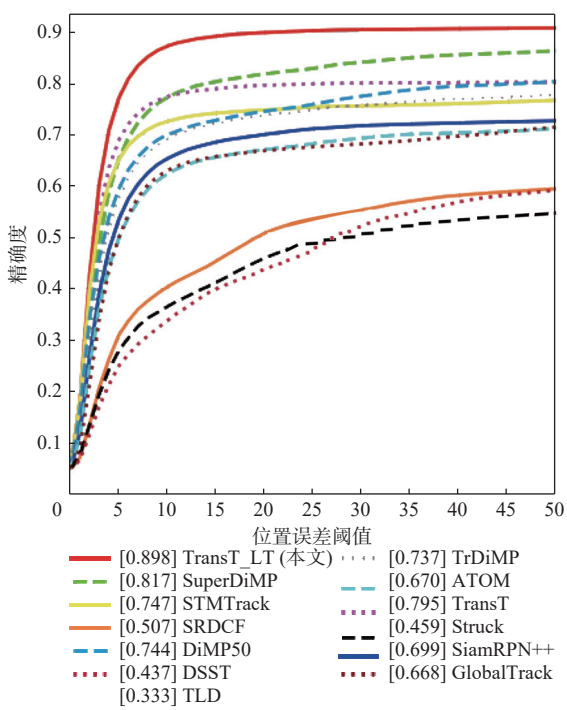
可以看出, 本文构建的长时视觉跟踪算法相较于基准算法 TransT, 在 F 分数上提升了 0.033。相较于使用相同长时视觉跟踪框架的 LGST, 本文算法在 F 分数上高出 0.093。

2.6　定性分析

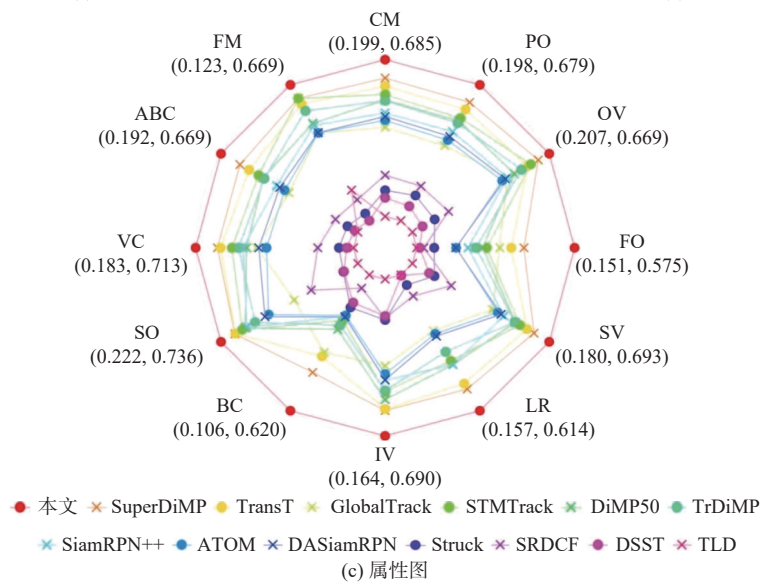
为直观体现本文所提长时视觉跟踪算法的优



(a) 成功率



(b) 精确度



(c) 属性图

图 7 UAV20L 跟踪结果评估曲线及属性的平均成功率对比结果

Fig. 7 UAV20L tracking result evaluation curves and average success rate comparison results for attributes

势，将所构建的长时视觉跟踪算法 TransT_LT 与基础跟踪器 TransT[14] 和 2 个经典的长时视觉跟踪算法 MBMD[11]、SPLT[26] 在 VOT2020-LT 数据集上进行定性分析。

图 8 选择的视频序列包括目标长时间出视野、全遮挡、尺度变化、背景杂乱和部分遮挡等多种挑战。可以看出，相比于 SPLT、MBMD 和基础跟踪器 TransT，改进后的算法 TransT_LT 能在目标丢失后及时使用全局重检测方法在全局范围内搜索目标，当目标重新出现时能再次找到目标，有效提升了长时视觉跟踪算法的性能。实验结果表明，本文提出的长时视觉跟踪算法能在各种复杂的情况下准确跟踪目标，有一定的鲁棒性。

2.7 速度分析

对所提长时视觉跟踪算法 TransT_LT 进行速度分析，图 9 展示了在 VOT2020-LT[19] 数据集上基准算法 TransT 及改进后的算法 TransT_LT 的精度与平均速度。此外，本文还与其他 11 种长时视觉跟踪算法进行了精度和平均速度的比较，对比算法包括 TANet[4]、ELGLT[6]、基于动态记忆网络的局部全局联合搜索机制 LGST[7]、MBMD[11]、GlobalTrack[15]、SPLT[26]、Siam R-CNN[27]、基于记忆注意网络的长期分段跟踪器 (long-term segmentation tracker with memory attention network, LTST)[39]、基于改进判别模型

表 6　VOT2018-LT 数据集上不同视觉跟踪算法的跟踪结果

Table 6　Tracking results of different visual tracking algorithms on VOT2018-LT dataset

算法	*F*分数	精确度	召回率
TANet[4]	0.586	0.649	0.535
LTMU[5]	0.690	0.710	0.672
ELGLT[6]	0.638	0.669	0.610
LGST[7]	0.630	0.637	0.622
MBMD[11]	0.610	0.634	0.588
TransT[14]	0.670	0.714	0.631
GlobalTrack[15]	0.555	0.503	0.528
SPLT[26]	0.616	0.633	0.600
Siam R-CNN[27]	0.668	0.667	0.675
LTST[39]	0.636	0.653	0.620
MTTNet[40]	0.621	0.634	0.609
TransT_LT(本文)	0.693	0.736	0.654

表 7　VOT2020-LT 数据集上不同视觉跟踪算法的跟踪结果

Table 7　Tracking results of different visual tracking algorithms on VOT2020-LT dataset

算法	*F*分数	精确度	召回率
TANet[4]	0.515	0.568	0.513
ELGLT[6]	0.607	0.611	0.560
LGST[7]	0.578	0.607	0.552
MBMD[11]	0.575	0.623	0.534
TransT[14]	0.638	0.682	0.599
GlobalTrack[15]	0.555	0.503	0.528
SPLT[26]	0.565	0.587	0.544
Siam R-CNN[27]	0.670	0.658	0.676
LTST[39]	0.587	0.631	0.548
RLT-DiMP[41]	0.670	0.657	0.684
TransT_LT(本文)	0.671	0.715	0.632

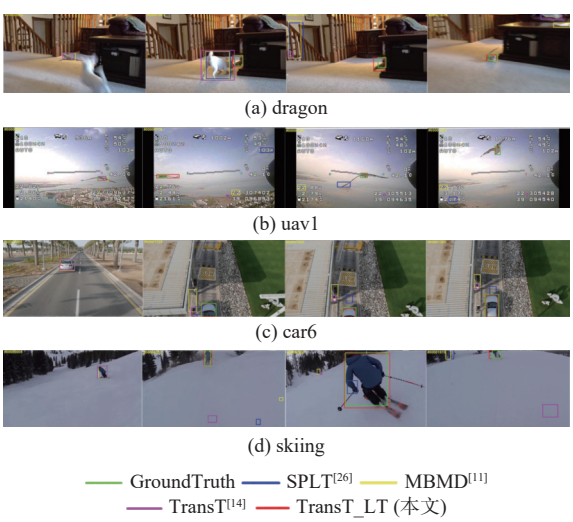


图 8　定性分析图

Fig. 8　Qualitative analysis diagram

预测的鲁棒长时目标跟踪 (robust long-term object tracking via improved discriminative model prediction, RLT-DiMP)[41]、mbdet 和 SiamDW_LT(mbdet 和 SiamDW_LT 来源于 https://www.votchallenge.net/vot2019/results.html)。

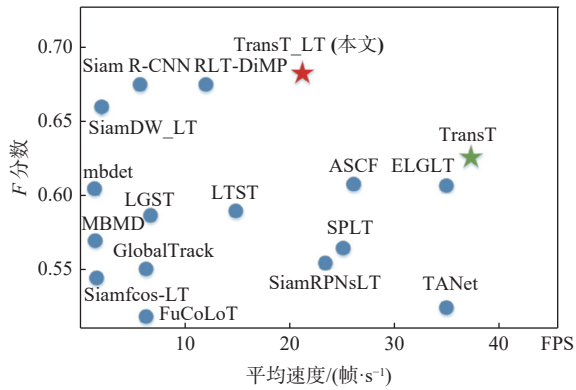


图 9　视觉跟踪算法平均速度对比结果

Fig. 9　Average speed comparison results of vision tracking algorithm

本文使用的基础跟踪器 TransT 的速度为 38 帧/s, 使用的全局重检测方法 GlobalTrack 的速度为 6 帧/s, 搭建的长时视觉跟踪算法 TransT_LT 的速度为 21 帧/s。本文所提级联目标漂移判定网络的判定速度为 100 帧/s, 对所提长时视觉跟踪算法的速度影响不大, 但由于使用的全局重检测方法 GlobalTrack 的速度仅为 6 帧/s, 会对所提长时视觉跟踪算法的速度产生较大影响。但本文所提长时视觉跟踪算法 TransT_LT 仍比一些经典长时视觉跟踪算法的速度快。

2.8　第 2 级目标漂移判定网络在尺度变换属性上的效果

为更清楚地展示加入第 2 级目标漂移判定网络之后在尺度变换属性上的效果, 在 UAV20L 数据集[16] 上进行了定量分析, 在 VOT2020-LT 数据集[19] 上进行了定性分析。表 8 展示了本文所提长时视觉跟踪算法 TransT_LT 在加入第 2 级目标漂移判定网络之后相比于没有加入第 2 级目标漂移判定网络在尺度变换属性上的成功率与精度对比。

可以看到, 加入第 2 级目标漂移判定网络后, 在 UAV20L 数据集上针对尺度变换的属性上成功率增加了 0.005, 精确度增加了 0.006, 这与表 5 中

表 8　UAV20L 数据集上第 2 级目标漂移判定网络在尺度变换属性上的实验结果

Table 8　Experimental results of the second-level target drift determination network on scale transformation attributes on UAV20L dataset

有无级联	成功率	精确度
TransT_LT(无级联)	0.688	0.886
TransT_LT(有级联)	0.693	0.892

加入第2级目标漂移判定网络后在UAV20L数据集上整体的成功率和精确度的增加相同。说明加入第2级目标漂移判定网络后，不仅在尺度变换属性上有所提升，而且还不会对其他属性产生影响。

图10从定性分析的角度更直观地体现了加入第2级目标漂移判定网络后在尺度变换属性上判定准确性的提高。其中，绿色框代表GroundTruth，红色框代表所提目标漂移判定网络判定当前帧跟踪结果错误，黄色框代表所提目标漂移判定网络判定当前帧跟踪结果正确。第1行与第3行为没有加入第2级目标漂移判定网络的实验结果，第2行与第4行为加入了第2级目标漂移判定网络的实验结果。可以看到，当目标尺度正常时，2个网络都能准确判定跟踪结果是否正确，但当目标尺度发生变化时，加入第2级目标漂移判定网络后，相比于没有加入第2级目标漂移判定网络，能够更准确地判定跟踪结果是否正确。

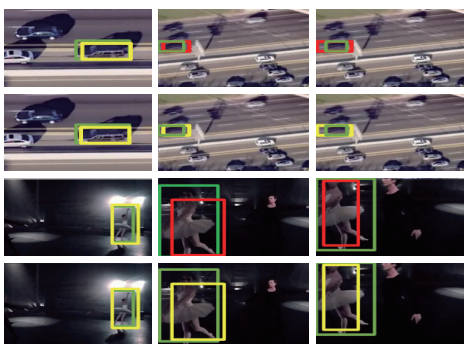


图10 加入第2级目标漂移判定网络后在尺度变换属性上的定性分析

Fig. 10 Qualitative analysis on scale transformation attributes after incorporating the second-level target drift determination network

2.9 结果分析

尽管算法TransT本身已涵盖了基于Transformer结构的全局特征关联检测方法，但其搜索区域仍是一个局部的搜索区域(256×256)。因此，当目标丢失时，若继续使用算法TransT进行跟踪，可能在一个没有目标的局部搜索区域内进行无意义的搜索。相比之下，GlobalTrack专为全局重检测而设计，具备更强大的全局特征关联和检测性能，其搜索区域涵盖整张图像，因此，能够在目标丢失后，从全局范围内重新寻找目标。即使当前帧中目标不在视野内，但当目标重新进入视野时，GlobalTrack也能够在第一时间重新跟踪到，这是算法TransT无法实现的。因此，在本文搭建的长时视觉跟踪算法中使用GlobalTrack作为全局重检测方法是合理且必要的。

3 结 论

1）提出了一个级联目标漂移判定网络，该网络在不同数据集上能自适应地选取阈值对跟踪结果进行准确判定。

2）在级联目标漂移判定网络中提出了一个长短时模板更新策略，保证判定过程中长短时模板的质量，从而提高网络的判定性能。

3）基于所提级联目标漂移判定网络，以TransT为基础跟踪器搭建了一个长时视觉跟踪算法TransT_LT，能有效应对长时视觉跟踪中目标消失再出现、尺度变化等挑战。对所提长时视觉跟踪算法TransT_LT在4个长时视觉数据集上进行了大量实验，验证了其出色的性能。

但是，由于全局重检测方法的速度一般较慢，导致构建的长时视觉跟踪算法的跟踪速度出现明显下降，下一步工作将在继续优化目标漂移判定网络的基础上，对全局重检测方法进行改进，进一步提升长时视觉跟踪算法的性能和跟踪速度。

参考文献（References）

[1] 李玺, 查宇飞, 张天柱, 等. 深度学习的目标跟踪算法综述[J]. 中国图象图形学报, 2019, 24(12): 2057-2080.
LI X, ZHA Y F, ZHANG T Z, et al. Survey of visual object tracking algorithms based on deep learning[J]. Journal of Image and Graphics, 2019, 24(12): 2057-2080(in Chinese).

[2] 刘芳, 孙亚楠, 王洪娟, 等. 基于残差学习的自适应无人机目标跟踪算法[J]. 北京航空航天大学学报, 2020, 46(10): 1874-1882.
LIU F, SUN Y N, WANG H J, et al. Adaptive UAV target tracking algorithm based on residual learning[J]. Journal of Beijing University of Aeronautics and Astronautics, 2020, 46(10): 1874-1882 (in Chinese).

[3] 蒲磊, 冯新喜, 侯志强, 等. 基于级联注意力机制的孪生网络视觉跟踪算法[J]. 北京航空航天大学学报, 2020, 46(12): 2302-2310.
PU L, FENG X X, HOU Z Q, et al. Siamese network visual tracking algorithm based on cascaded attention mechanism[J]. Journal of Beijing University of Aeronautics and Astronautics, 2020, 46(12): 2302-2310(in Chinese).

[4] WANG X, TANG J, LUO B, et al. Tracking by joint local and global search: a target-aware attention-based approach[J]. IEEE Transactions on Neural Networks and Learning Systems, 2022, 33(11): 6931-6945.

[5] DAI K N, ZHANG Y H, WANG D, et al. High-performance long-term tracking with meta-updater[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2020: 6297-6306.

[6] ZHAO H J, YAN B, WANG D, et al. Effective local and global search for fast long-term tracking[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45(1): 460-474.

[7] GAO Z, ZHUANG Y, GU J J, et al. A joint local-global search

mechanism for long-term tracking with dynamic memory network [J]. Expert Systems with Applications, 2023, 223: 119890.

[8] BOLME D S, BEVERIDGE J R, DRAPER B A, et al. Visual object tracking using adaptive correlation filters[C]//Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2010: 2544-2550.

[9] WANG M M, LIU Y, HUANG Z Y. Large margin object tracking with circulant feature maps[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2017: 4800-4808.

[10] JUNG I, SON J, BAEK M, et al. Real-time MDNet[C]//Proceedings of the European Conference on Computer Vision. Berlin: Springer, 2018: 83-98.

[11] ZHANG Y, WANG D, WANG L, et al. Learning regression and verification networks for long-term visual tracking[EB/OL]. (2018-11-19)[2023-08-01]. https://arxiv.org/abs/1809.04320v1.

[12] NAM H, HAN B. Learning multi-domain convolutional neural networks for visual tracking[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2016: 4293-4302.

[13] HOU Z Q, WANG Z, PU L, et al. Target drift discriminative network based on deep learning in visual tracking[J]. Journal of Electronic Imaging, 2022, 31: 043052.

[14] CHEN X, YAN B, ZHU J, et al. Transformer tracking[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2021: 8126-8135.

[15] HUANG L H, ZHAO X, HUANG K Q. GlobalTrack: a simple and strong baseline for long-term tracking[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2020, 34(7): 11037-11044.

[16] MUELLER M, SMITH N, GHANEM B. A benchmark and simulator for UAV tracking[C]//Proceedings of the European Conference on Computer Vision. Berlin: Springer, 2016: 445-461.

[17] FAN H, LIN L T, YANG F, et al. LaSOT: a high-quality benchmark for large-scale single object tracking[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2019: 5369-5378.

[18] LUKEŽIČ A, ZAJC L Č, VOJÍŘ T, et al. Now you see me: evaluating performance in long-term visual tracking[C]//Proceedings of the European Conference on Computer Vision. Berlin: Springer, 2018: 4991391.

[19] KRISTAN M, LEONARDIS A, MATAS J, et al. The eighth visual object tracking VOT2020 challenge results[C]//Proceedings of the European Conference on Computer Vision. Berlin: Springer, 2020: 547-601.

[20] HU J, SHEN L, SUN G. Squeeze-and-excitation networks[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2018: 7132-7141.

[21] ZHAO H S, SHI J P, QI X J, et al. Pyramid scene parsing network[C]//Proceedings of theIEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2017: 6230-6239.

[22] NI J J, WU J H, TONG J, et al. GC-Net: global context network for medical image segmentation[J]. Computer Methods and Programs in Biomedicine, 2020, 190: 105121.

[23] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition[EB/OL]. (2015-04-10)[2023-08-01]. https://arxiv.org/abs/1409.1556.

[24] DONG Q J, HE X D, GE H Y, et al. Improving model drift for robust object tracking[J]. Multimedia Tools and Applications, 2020, 79(35): 25801-25815.

[25] WU Y, LIM J, YANG M H. Object tracking benchmark[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015, 37(9): 1834-1848.

[26] YAN B, ZHAO H J, WANG D, et al. 'skimming-perusal' tracking: a framework for real-time and robust long-term tracking[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE Press, 2019: 2385-2393.

[27] VOIGTLAENDER P, LUITEN J, TORR P H S, et al. Siam R-CNN: visual tracking by re-detection[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2020: 6577-6587.

[28] MA F, SHOU M Z, ZHU L C, et al. Unified Transformer tracker for object tracking[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2022: 8771-8780.

[29] FU Z H, LIU Q J, FU Z H, et al. STMTrack: template-free visual tracking with space-time memory networks[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2021: 13769-13778.

[30] BHAT G, DANELLJAN M, VAN GOOL L, et al. Learning discriminative model prediction for tracking[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE Press, 2019: 6181-6190.

[31] GUO D Y, SHAO Y Y, CUI Y, et al. Graph attention tracking [C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2021: 9538-9547.

[32] LI B, WU W, WANG Q, et al. Evolution of Siamese visual tracking with very deep networks[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2019: 16-20.

[33] DANELLJAN M, BHAT G, KHAN F S, et al. ATOM: accurate tracking by overlap maximization[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2019: 4655-4664.

[34] WANG N, ZHOU W G, WANG J, et al. Transformer meets tracker: exploiting temporal context for robust visual tracking[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2021: 1571-1580.

[35] DANELLJAN M, HÄGER G, KHAN F S, et al. Learning spatially regularized correlation filters for visual tracking[C]//Proceedings of the IEEE International Conference on Computer Vision. Piscataway: IEEE Press, 2015: 4310-4318.

[36] HARE S, GOLODETZ S, SAFFARI A, et al. Struck: structured output tracking with kernels[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2016, 38(10): 2096-2109.

[37] DANELLJAN M, HÄGER G, KHAN F S, et al. Discriminative scale space tracking[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(8): 1561-1575.

[38] KALAL Z, MIKOLAJCZYK K, MATAS J. Tracking-learning-

detection[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2012, 34(7): 1409-1422.

[39] YU L, QIAO B J, ZHANG H L, et al. LTST: long-term segmentation tracker with memory attention network[J]. Image and Vision Computing, 2022, 119: 104374.

[40] SANG H, LI G, ZHAO Z. Multi-scale global retrieval and temporal-spatial consistency matching based long-term tracking network[J]. Chinese Journal of Electronics, 2022, 32: 1-11.

[41] CHOI S, LEE J, LEE Y, et al. Robust long-term object tracking via improved discriminative model prediction[C]//Proceedings of the European Conference on Computer Vision. Berlin: Springer, , 2020: 602-617.

Cascaded object drift determination network for long-term visual tracking

HOU Zhiqiang[1, 2, *], ZHAO Jiaxin[1, 2], CHEN Yu[1, 2], MA Sugang[1, 2], YU Wangsheng[3], FAN Jiulun[1]

(1. School of Computer Science and Technology, Xi'an University of Posts and Telecommunications, Xi'an 710121, China;
2. Shaanxi Key Laboratory of Network Data Analysis and Intelligent Processing, Xi'an University of Posts and Telecommunications, Xi'an 710121, China;
3. College of Information and Navigation, Air Force Engineering University, Xi'an 710077, China)

Abstract: Aiming at the problems of artificially selecting the threshold and poor determination performance in the existing object drift determination criteria, this paper proposes a cascaded object drift determination network with adaptive threshold selection. Firstly, through the cascade design of the two sub-networks, determine whether the tracking results drift. The results are then jointly determined by the proposed network using the static template, long-term template, and short-term template. A long-term and short-term template update strategy is then designed to guarantee the quality of the template and adapt it to the object's changing appearance during the determination process. Finally, the proposed network is combined with the short-term tracker TransT and the global re-detection method GlobalTrack to build a long-term tracking algorithm TransT_LT. The proposed algorithm's performance test on four datasets (UAV20L, LaSOT, VOT2018-LT, and VOT2020-LT) demonstrates that it performs better over the long term in tracking, particularly on the UAV20L dataset, where it outperforms the benchmark algorithm by 7.7% and 10.3%, respectively, in tracking success rate and accuracy. The determination speed of the proposed network is 100 frames per second, which has little effect on the speed of the long-term tracking algorithm.

Keywords: long-term visual tracking; deep learning; cascaded object drift determination network; template updates; multi-scale feature fusion

Received: 2023-08-02; **Accepted**: 2023-11-10; **Published Online**: 2023-11-23 20:24

URL: link.cnki.net/urlid/11.2625.V.20231123.0918.001

Foundation items: National Natural Science Foundation of China (62072370); Natural Science Foundation of Shaanxi Province (2023-JC-YB-598)

* **Corresponding author**. E-mail: hzq@xupt.edu.cn

http://bhxb.buaa.edu.cn jbuaa@buaa.edu.cn
DOI: 10.13700/j.bh.1001-5965.2023.0354

基于动态嵌入特征的鲁棒半监督视频目标分割

陈亚当[1,2,*]，赵翊冰[1,2]，吴恩华[3]

(1. 南京信息工程大学 计算机学院，南京 210044；2. 南京信息工程大学无锡研究院，无锡 214100；
3. 中国科学院软件研究所，北京 100190)

摘 要：针对半监督视频目标分割 (VOS) 方法存在推理时内存占用不断增加及仅依赖低级像素特征训练困难的问题，提出一种基于动态嵌入特征和辅助损失函数的半监督视频目标分割方法。使用动态嵌入特征建立恒定大小的记忆库；通过时空聚合方法，利用历史信息生成和更新动态嵌入特征；使用内存更新感应器来自适应控制记忆库的更新间隔，适应不同视频的运动模式；使用辅助损失函数，在高级语义特征层面上给网络提供辅助指导，并通过在多重特征层面多方面指导，提高模型精度和训练效率；针对视频前背景中相似目标误匹配的问题，设计一种时空约束模块，以利用视频的时间连续性特性更好地捕获前一帧掩码信息与当前帧之间的关联。实验结果表明：所提方法在 DAVIS 2017 验证集上达到 84.5% *J&F* 的精度，在 YouTube-VOS 2019 验证集达到 82.4% *J&F* 的精度。

关 键 词：视频目标分割；时空记忆网络；时空约束；内存更新感应；动态嵌入特征
中图分类号：TP391
文献标志码：A **文章编号**：1001-5965(2025)07-2253-09

视频目标分割 (video object segmentation, VOS) 是视频理解的基本任务，包括很多现实应用，如自动驾驶和视频编辑。本文专注于半监督学习下的视频目标分割，即给定第 1 帧的目标掩码，推理出剩余视频帧中的目标掩码。

传统的半监督视频目标分割方法可分为基于掩码传播[1]的方法、基于在线微调[2-3]的方法和混合方法[4-5]。基于掩码传播[1]的方法使用了先前帧的信息，将图像和前 1 帧的分割掩码作为输入进行当前帧的推理，虽然对物体形变的鲁棒性较强，但误差会随着推理过程而累加。基于在线微调[2-3]的方法学习分割任务所需的通用特征，在推理阶段利用视频第 1 帧进行微调，该方法对目标遮挡的鲁棒性强，但目标在后续帧中的形变会影响分割精度。混合方法[4-5]同时使用第 1 帧和前 1 帧的信息进行分割，对目标遮挡和形变都有较强的鲁棒性，但未充分利用视频中的历史帧，导致性能提升有限。

最近，基于时空记忆的方法[6-26]成为主流方法。该方法充分利用已分割的历史帧信息，统一编码为嵌入特征，存储在记忆库中以协助后续帧的分割。虽然该方法取得了较好的效果，但存在以下问题：①随着时间的推移，基于时空记忆的方法的存储需求不断增长，而硬件无法满足这种增长需求；②大部分方法仅依靠低级像素特征 (掩码) 对网络进行训练，缺乏高级语义特征层面的辅助指导，导致模型训练困难。

针对第 1 个问题，受 Li 等[9]的启发，本文设计了一种时空聚合方法，用于生成与视频长度无关、固定大小的动态嵌入特征。时空聚合方法包括提取、增强和压缩 3 个部分，可以自适应地组织历史信息和最新帧的嵌入，维护时空聚合模式记忆库。此外，本文还设计了一种内存更新感应器，其逐帧

收稿日期：2023-06-13；**录用日期**：2023-12-01；**网络出版时间**：2024-01-10 15:11
网络出版地址：link.cnki.net/urlid/11.2625.V.20240110.0919.001
基金项目：国家自然科学基金 (62473201,62477026)；无锡市产业创新研究院先导技术预研项目
*通信作者. E-mail：adamchen@nuist.edu.cn

引用格式：陈亚当，赵翊冰，吴恩华. 基于动态嵌入特征的鲁棒半监督视频目标分割 [J]. 北京航空航天大学学报，2025，51（7）：2253-2261. CHEN Y D，ZHAO Y B，WU E H. Robust semi-supervised video object segmentation based on dynamic embedding feature [J]. Journal of Beijing University of Aeronautics and Astronautics，2025，51（7）：2253-2261 （in Chinese）.

评估帧与帧之间的变化程度，避免无差别更新内存导致精度下降。

针对第2个问题，本文使用了一种辅助损失，可以为网络提供高级语义特征层面的辅助指导，帮助时空聚合模块高效选择用于维护新型记忆库的有用特征。通过在低级和高级特征层面上共同学习，为网络提供更丰富的特征空间学习目标，从而使网络的训练更加稳定和高效。

同时，针对视频前背景中出现相似目标，容易出现错误匹配的问题，受Zhang等[6]的启发，本文设计了一种时空约束模块，其充分利用了视频的时空约束信息，使用来自前1帧视频的掩码作为粗略位置约束来引导模型对前景目标进行分割，更好地捕获前1帧掩码信息与当前帧之间的关联。

本文工作的主要创新点如下：

1）设计了一种时空聚合方法，通过自适应地组织历史信息和最新帧的嵌入，生成与视频长度无关、固定大小的记忆库。

2）使用了一种辅助损失，为网络提供高级语义特征层面的辅助指导，使网络训练更加稳定和高效。

3）在基准数据集上与其他方法相比表现良好。在DAVIS 2017验证集上获得了84.5% *J&F*的精度，在YouTube-VOS 2019验证集上获得了82.4% *J&F*的精度。

1 相关工作

在半监督视频目标分割工作中，时空记忆[10]网络模型一直是该领域的主流方法，该方法通过对已分割的视频帧进行特征映射，以记忆编码方式存储在内存中，为视频后续帧提供丰富的分割线索，其取得了令人印象深刻和优异的分割结果。但该方法最大的问题在于所存储的记忆编码会随视频时长而不断累加，造成大量数据冗余，使得记忆匹配产生混淆，造成分割结果错误。

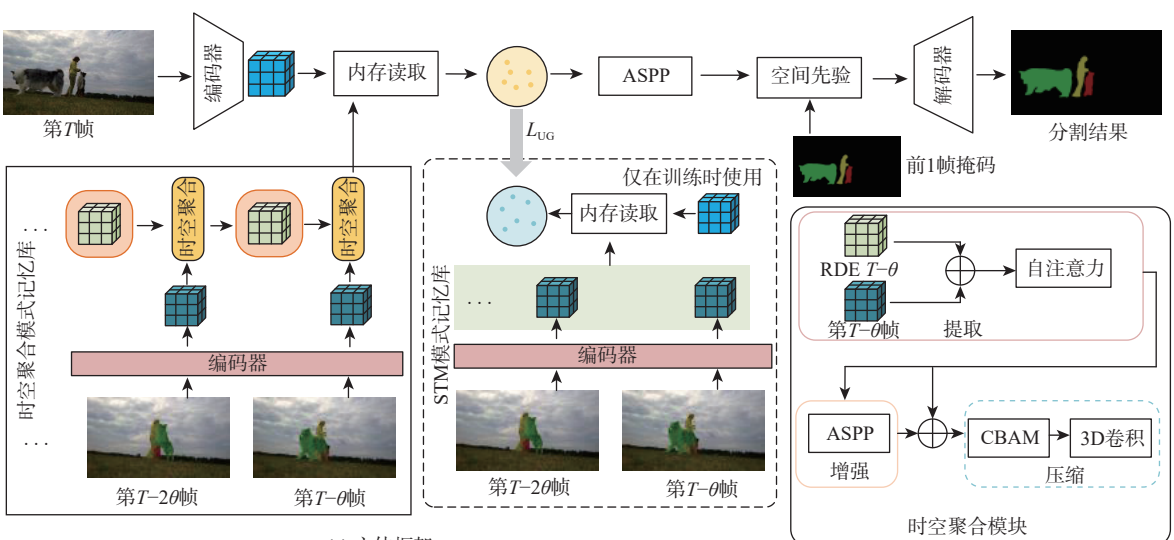
为此，Zhang等[6]通过局部读取内存信息，减少冗余内存信息对分割目标的干扰。类似的，Xie等[13]通过仅提取目标物体所在区域的特征，并进行局部特征匹配，从而减少了全局特征匹配所带来的混淆干扰。虽然上述方法成功降低了数据冗余，提高了计算效率，缓解了视频前背景中目标误匹配的问题，但这种局部记忆的方法难免会遗漏重要分割线索，对任意变化的视频目标而言，该方案是一种次优解，并非全局最优解。

Li等[9]为此提出一种全局最优方法，基于时空记忆网络试图通过维持一个记忆库来利用过去帧的所有信息，生成的记忆编码长度与视频时长无关，但由于设计简单，造成特征信息在记忆库中会被丢失，无法提供准确的分割线索。针对这一问题，本文设计了一种时空聚合方法，通过固定大小的动态嵌入特征，并使用辅助损失，为网络提供高级语义特征层面的辅助指导，大幅提高了分割精度与效果。

2 本文方法

2.1 时空聚合模式记忆库

STM[10]模式存储器组可以无损地存储历史信息，但其记忆库会随着时间推移逐步变大。而维持记忆库大小不变的重点是如何高效选择记忆特征。为此，本文设计了一种时空聚合模式记忆库。本文方法主体框架如图1所示。可以看出，在训练



(a) 主体框架

图1 本文方法主体框架

Fig. 1 Main framework of the proposed method

阶段, 同时维护 STM[10] 和时空聚合模式记忆库。在推理阶段, 只使用时空聚合模式记忆库, 其占用恒定内存, 不会随着视频时间推移变大。将 STM[10] 模式记忆库记为 M, 其包含 $\{\boldsymbol{k}_t^M, \boldsymbol{v}_{t,i}^M\}$, $\boldsymbol{k}$ 和 $\boldsymbol{v}$ 分别为关键词矩阵和原值矩阵。记时空聚合模式记忆库为 m, 其包含 $\{\boldsymbol{k}_t^m, \boldsymbol{v}_{t,i}^m\}$。其中, t为时间, i为第 i个目标。

2.1.1　动态嵌入特征

为充分利用历史帧和最新帧的信息, 使用一种动态嵌入特征, 将历史帧和最新帧的嵌入融合, 提供了一种固定大小的记忆库表达形式。将时间t处的动态嵌入特征表示为$\{\boldsymbol{k}_t^{\mathrm{RDE}}, \boldsymbol{v}_{t,i}^{\mathrm{RDE}}\} \in \{\boldsymbol{k}_t^m, \boldsymbol{v}_{t,i}^m\}$。

2.1.2　时空聚合方法

为生成和更新动态嵌入特征, 设计了一种时空聚合模块, 负责提取历史帧和最新帧嵌入中的重要信息。

时空聚合方法包括 3 个部分: 提取、增强和压缩, 如图 1 所示。提取部分负责组织前 1 帧嵌入 $\{\boldsymbol{k}_{t-\theta}^{\mathrm{RDE}}, \boldsymbol{v}_{t-\theta,i}^{\mathrm{RDE}}\}$($\theta$为采样间隔)和最新帧嵌入 $\{\boldsymbol{k}_t^{\mathrm{L}}, \boldsymbol{v}_{t,i}^{\mathrm{L}}\}$之间的时空关系。以 $\boldsymbol{k}_t^{\mathrm{RDE}}$更新为例, 拼接前 1 帧嵌入 $\{\boldsymbol{k}_{t-\theta}^{\mathrm{RDE}}, \boldsymbol{v}_{t-\theta,i}^{\mathrm{RDE}}\}$和最新帧嵌入$\{\boldsymbol{k}_t^{\mathrm{L}}, \boldsymbol{v}_{t,i}^{\mathrm{L}}\}$以获得特征$\boldsymbol{x}$, 具体公式可表示为

$$\boldsymbol{x} = \mathrm{Cat}(\boldsymbol{k}_{t-\theta}^{\mathrm{RDE}}, \boldsymbol{k}_t^{\mathrm{L}}) \qquad \boldsymbol{x} \in \mathbf{R}^{C_k \times 2 \times \frac{H}{16} \times \frac{W}{16}} \tag{1}$$

式中: Cat表示在时间维度进行拼接操作; C_k为特征图通道数; H 和 W 分别为特征图的高和宽。

受自注意力机制的启发, 组织先前的动态嵌入特征 $\boldsymbol{k}_{t-\theta}^{\mathrm{RDE}}$和最新帧 $\boldsymbol{k}_{t-\theta}^{\mathrm{L}}$嵌入之间的时空关系, 以获得聚合特征$\boldsymbol{x}_{\mathrm{agg}}$, 具体公式可表示为

$$\boldsymbol{x}_{\mathrm{agg}} = \frac{1}{C(\boldsymbol{x})}(\omega(\boldsymbol{x}))^{\mathrm{T}} \phi(\boldsymbol{x}\downarrow) g(\boldsymbol{x}\downarrow) \tag{2}$$

式中: $C(\boldsymbol{x})$ 为归一化因子, 其值为$\boldsymbol{x}$空间位置数量的总和; 函数ω、ϕ和g为 1×1×1 卷积, $\boldsymbol{x}\downarrow$表示经过最大池化操作后的$\boldsymbol{x}$。

在增强部分, 通过空洞空间卷积池化金字塔[27-28] (atrous spatial pyramid pooling, ASPP) 以残差的形式增强$\boldsymbol{x}_{\mathrm{agg}}$。在压缩部分, 使用 CBAM [29] 模块捕捉重要信息, 通过 2×3×3 卷积压缩增强的特征, 记为 Squeeze 函数, 公式可表示为

$$\boldsymbol{k}_t^{\mathrm{RDE}} = \mathrm{Squeeze}(\boldsymbol{x}_{\mathrm{agg}} + \mathrm{ASPP}(\boldsymbol{x}_{\mathrm{agg}})) \tag{3}$$

通过该方法, 先前动态嵌入特征和最新帧嵌入自适应地融合, 并保持记忆库大小不变。

2.1.3　内存更新感应器

先前的方法根据不同的数据集采用不同的固定存储间隔, 这种不加区分的减少时间冗余方式会导致精度损失。

基于以上考虑, 并受 Wang 等[17] 的启发, 引入一种内存更新感应器, 其会评估帧与帧之间的变化程度, 当变化程度超出设定的阈值P_{th}时, 对内存进行更新。给定x_t、y_t和x_{t-1}、y_{t-1}, 分别计算图像差异D_{f}和掩码差异D_{m}, 公式为

$$D_{\mathrm{f}}^j = \sum_{c\in\{R,G,B\}} |x_t^{j,c} - x_{t-1}^{j,c}|/255 \tag{4}$$

$$D_{\mathrm{m}}^j = |y_t^j - y_{t-1}^j| \tag{5}$$

在处理每个像素j时, 将全局的像素变化程度指标P更新为

$$P = \begin{cases} P+1 & D_{\mathrm{f}}^j > 1 \text{或} D_{\mathrm{m}}^j > \tau_{\mathrm{m}} \\ P & \text{其他} \end{cases} \tag{6}$$

一旦P超过阈值P_{th}, 将触发新一轮的内存更新。实验表明, P_{th}和τ_{m}分别等于 160 和 0 时取得了较好的性能。

2.2　高层语义辅助损失

目前, 大多数方法只使用低级像素特征 (掩码) 对网络进行训练, 缺乏高级语义特征层面的辅助指导, 导致训练变得困难。此外, 动态嵌入特征中的关键词矩阵和原值矩阵是由 2 个参数不同的时空聚合方法生成, 如果仅依靠低级像素特征的指导, 很难直接定义这种新型记忆库。

基于以上考虑, 受 Li 等[9] 的启发, 本文认为成熟的 STM[10] 模式记忆库中的高级语义特征是有价值的学习目标。为帮助时空聚合模式学习这些特征, 在训练阶段额外维护了一个 STM[10] 模式记忆库, 该记忆库存储并提取出这些特征, 以供时空聚合模式学习使用。具体来说, 使用了一种引导损失L_{UG}, 使得从时空聚合模式记忆库中读出的高级语义特征趋近于从 STM[5] 模式记忆库读出的高级语义特征。辅助损失的公式如下:

$$L_{\mathrm{UG}} = \sum_i \mathrm{KL}\left(\boldsymbol{v}_{t,i}^{M\to Q} \| \boldsymbol{v}_{t,i}^{m\to Q}\right) \tag{7}$$

式中: KL 表示 KL 散度; $\boldsymbol{v}_{t,i}^{M\to Q}$为 STM[10] 模式记忆库读出的特征; $\boldsymbol{v}_{t,i}^{m\to Q}$为时空聚合模式记忆库读出的特征。

2.3　时空约束模块

为应对目标形变和遮挡等问题, 时空记忆网络利用全局匹配的方式在特征空间中检索有用的信息。然而, 这种方式并未考虑空间一致性。因此, 当前背景中目标相似时, 模型可能会出现错误的预测。

本文设计了一种时空约束模块(见图 2)以确保相邻帧之间的空间一致性。定义前 1 帧的预测掩码为形状 H×W 的 0-1 掩码, 其与当前帧嵌入特征

$H\times W\times C$连接以获得形状为$H\times W\times(C+1)$的特征图。本文引入通道注意力机制，首先，使用全局平均池化，降低输入特征的分辨率，将其映射到1，并使用批量归一化Sigmoid和1×1卷积以调节特征图的值，将其与原始特征图相乘以生成通道注意力分支的输出；然后，经过大小为3×3的卷积层和Sigmoid函数生成空间先验，即形状为$H\times W\times 1$的特征图，将空间先验与当前帧嵌入特征相乘。

图3展现了一些帧相应的时空约束模块可视化结果，模块粗略地捕获了目标的位置。

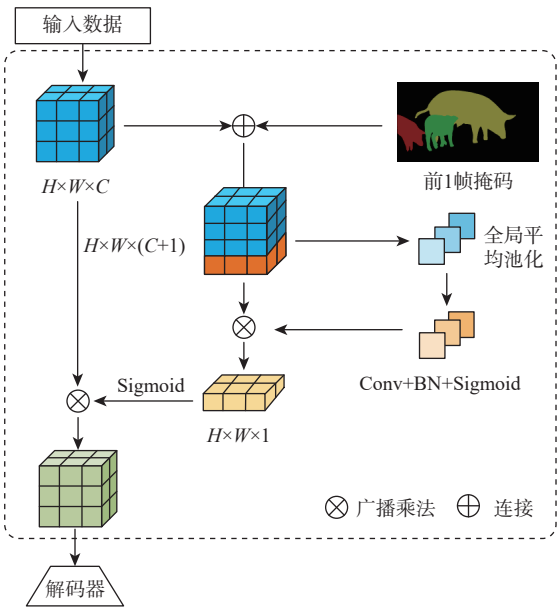


图2 时空模块的内部构造

Fig. 2 Internal structure of spatiotemporal constraint module

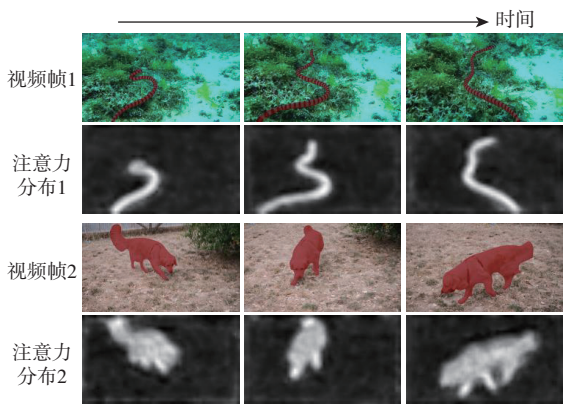


图3 时空约束模块注意力可视化结果

Fig. 3 Results of attention visualization of spatiotemporal constraint module

2.4 总体损失函数

在训练阶段，与先前方法[9]相同，本文利用时空聚合模式记忆库对第3帧和第5帧进行分割，以处理不同速率的视频，利用STM[10]模式记忆库对第2帧和第4帧进行分割以提高训练稳定性。公式如下：

$$L_{\mathrm{Seg}}=\frac{1}{2}\left(\sum_i\sum_{t=2,4}\underbrace{\mathrm{BCE}\left(\tilde{\boldsymbol{y}}^M_{t,i},\boldsymbol{y}_{t,i}\right)}_{\mathrm{STM}}+\sum_i\sum_{t=3,5}\underbrace{\mathrm{BCE}\left(\tilde{\boldsymbol{y}}^m_{t,i},\boldsymbol{y}_{t,i}\right)}_{\mathrm{SAM}}\right) \tag{8}$$

式中：$\tilde{\boldsymbol{y}}^M_{t,i}$和$\tilde{\boldsymbol{y}}^m_{t,i}$分别为从STM[10]模式记忆库和时空聚合模式记忆库读取的分割结果；$\boldsymbol{y}_{t,i}$为第i个目标在第t帧的真实掩码。

同时，本文使用与之前工作[4]相同的自矫正策略，具体公式如下：

$$L_{\mathrm{MC}}=\mathrm{KL}(\boldsymbol{k}_1\|\ddot{\boldsymbol{k}}_1)+\sum_i\mathrm{KL}(\boldsymbol{v}_{1,i}\|\ddot{\boldsymbol{v}}_{1,i}) \tag{9}$$

式中：$\ddot{\boldsymbol{k}}_1$和$\ddot{\boldsymbol{v}}_{1,i}$为扰动后的特征；$\boldsymbol{k}_1$和$\boldsymbol{v}_{1,i}$为正常特征。

整体损失函数如下：

$$L=L_{\mathrm{Seg}}+L_{\mathrm{UG}}+L_{\mathrm{MC}} \tag{10}$$

3 实 验

3.1 数据集和指标

3.1.1 DAVIS验证集

DAVIS 2016是基于单目标的视频目标分割数据集，共有50个视频，其中，训练集30个，验证集20个。DAVIS 2017是基于多目标的视频目标分割数据集，共包含120个视频，其验证集和测试集分别为30个逐像素标注的视频。

3.1.2 YouTube-VOS 2019验证集

YouTube-VOS 2019是多目标视频分割最新的大规模基准，提供了3 471个视频用于训练(65个类别)和507个视频用于验证。验证集存在训练集中没有的类别，用来进一步评估模型的泛化性。

3.1.3 评估指标

对于DAVIS数据集，使用区域相似度J、轮廓准确度F及其平均值$J\&F$来评估分割结果。对于YouTube-VOS 2019数据集，展示了见过和未见过类别的J和F及其平均值$J_{\mathrm{seen/unseen}}\&F_{\mathrm{seen/unseen}}$。

3.2 实施细节

3.2.1 训练阶段

模型使用Adam优化器和PyTorch进行训练。与之前的工作[20]相同，在由静态图像执行变形操作形成的假视频[1]上进行预训练，并使用相同的数据处理方法，在YouTube-VOS 2019和DAVIS数据集上进行主要训练。与STM[10]相同，在训练过程中将冻结批归一化层。在训练阶段，将同时维护STM[10]模式记忆库和时空聚合模式记忆库，利用辅助损失对网络进行训练。

3.2.2 推理阶段

在推理过程中，只使用时空聚合模式记忆库。具体来说，除了通过时空聚合方法维护本文的动态

嵌入特征外, 为在准确的模板信息(第 1 帧)和动态信息(最新帧或动态嵌入特征)之间保持采样平衡, 还采样了最新帧的嵌入和第 1 帧的 2 次重复嵌入, 将最小更新间隔设置为 2。与先前方法[11]相同, 使用前 k 个过滤器[7]读取内存。

3.3　与其他方法对比

本文将在推理过程中记忆库大小恒定不变的方法标记为常数消耗, 使用 CC 表示。随着视频时间的推移, 该类方法的占用内存相对稳定。

3.3.1　定量比较

表 1 显示了本文方法在 DAVIS 2017 验证集上与其他方法的比较。在验证集上, 本文方法达到了 84.5% *J*&*F* 的精度, 相较于基线方法[9]提升了 0.3%。表 2 显示了本文方法在 DAVIS 2017 测试集上与其他方法的比较。在测试集上, 本文方法达到了 78.0% *J*&*F* 的精度, 相较于基线方法[9]提升了 0.6%, 获得了有竞争力的结果。表 3 展示了本文方法在 YouTube-VOS 2019 验证集上与其他方法的比较。在验证集上, 本文方法达到了 82.4% *J*&*F* 的精度, 相较于基线方法[9]提升了 0.5%, 获得了有竞争力的结果。表 4 显示了本文方法在 DAVIS 2016 验证集上与其他方法的比较。可以看到, 在 DAVIS 2016 验证集上, 本文方法达到了 91.2% *J*&*F* 的精度, 与基线方法相当。尽管本文方法在某些数据集上的精度低于 STCN[11], 但在推理过程中, 本文方法占用的内存恒定, 而 STCN[11]可能因为内存耗尽而导致推理失败。

表 1　DAVIS 2017 验证集上的实验结果

Table 1　Experimental results on DAVIS 2017 validation set　%

方法	OL	CC	*J*&*F*	*J*	*F*
MiVOS[12]		×	84.5	81.7	87.4
QDMN[24]		×	84.6	81.8	87.3
HMMN[8]		×	84.7	81.9	87.5
STCN[11]		×	85.3	82.0	88.6
XMem[25]		×	86.2	82.9	89.5
OnAVOS[2]	√	√	65.3	61.6	69.1
OSVOS[3]	√	√	60.2	56.6	63.9
GCNet[16]		√	71.4	69.3	73.5
文献[18]		√	74.6	73.0	76.1
PReMVOS[21]	√	√	77.8	73.9	81.7
SwiftNet[17]		√	81.1	78.3	83.9
SSTVOS[22]		√	82.5	79.9	85.1
文献[23]		√	82.7	80.2	85.3
SITVOS[26]		√	83.5	80.4	86.5
RDE[9]		√	84.2	80.8	87.5
本文		√	84.5	81.2	87.7

注: OL表示基于在线微调的方法, “√” 表示使用在线微调的方法。CC表示该方法占用恒定内存, “√” 表示内存占用保持不变, “×” 表示内存占用会逐渐增加。

表 2　DAVIS 2017 测试集上的实验结果

Table 2　Experimental results on DAVIS 2017 test set　%

方法	600p	CC	*J*&*F*	*J*	*F*
KMN[7]	√	×	77.2	74.1	80.3
STCN[11]	×	×	76.1	72.7	79.6
MiVOS[12]	×	×	76.0	72.6	79.3
QDMN[24]	×	×	77.7	74.2	81.2
XMem[25]	×	×	81.0	77.4	84.5
CFBI[4]	×	√	74.8	71.1	78.5
文献[23]	×	√	75.2	72,0	78.3
CFBI+[5]	×	√	75.6	71.6	79.6
RDE[9]	×	√	77.4	73.6	81.2
本文	×	√	78.0	74.3	81.6

注: 600p表示使用600分辨率视频进行推理, “√” 表示使用, “×” 表示不使用。CC表示该方法占用恒定内存, 其中 “√” 表示内存占用保持不变, “×” 表示内存占用会逐渐增加。

表3 YouTube-VOS 2019验证集上的实验结果

Table 3 Experimental results on YouTube-VOS 2019 validation set %

方法	CC	$J_{seen/unseen}\&F_{seen/unseen}$	F_{unseen}	F_{seen}	J_{unseen}	J_{seen}
STM[10]	×	79.2	79.6	83.6	73.0	80.6
MiVOS[12]	×	80.3	79.3	83.7	75.3	82.8
STCN[11]	×	82.7	81.1	85.4	78.2	85.9
XMem[25]	×	85.5	84.3	88.6	80.3	88.6
CFBI+[5]	√	81.0	80.6	85.1	75.2	83.0
SSTVOS[22]	√	81.8	80.9		76.6	
RDE[9]	√	81.9	81.1	85.5	76.2	84.8
本文	√	82.4	81.3	85.4	77.2	85.7

注：CC表示该方法占用恒定内存，“√”表示内存占用保持不变，“×”表示内存占用会逐渐增加。

表4 DAVIS 2016验证集上的实验结果

Table 4 Experimental results on DAVIS 2016 validation set %

方法	CC	J&F	J	F
STM[10]	×	89.3	88.7	89.9
KMN[7]	×	90.5	89.5	91.5
HMMN[8]	×	90.8	89.6	92.0
MiVOS[12]	×	91.0	89.7	92.4
QDMN[24]	×	91.0	90.2	91.7
STCN[11]	×	91.7	90.4	93.0
XMem[25]	×	91.5	90.4	92.7
GCNet[16]	√	86.6	87.6	85.7
CFBI+[5]	√	89.9	88.7	91.1
SwiftNet[17]	√	90.4	90.7	90.3
SITVOS[26]	√	90.5	89.5	91.4
RDE[9]	√	91.1	89.7	92.5
本文	√	91.2	89.8	92.5

注：CC表示该方法占用恒定内存，“√”表示内存占用保持不变，“×”表示内存占用会逐渐增加。

3.3.2 定性比较

图4展示了分割失败的示例，并与其他主流方法[11,25]进行了对比。图5显示了其他方法和本文模型在存在相似目标场景下的定性比较。图6展现了本文方法在数据集上更多的实验结果。可以看出，当前主流方法在应对光线变化时，分割精度普遍下降。这主要是因为光线使得目标的外观产生变化，数据增强是缓解该问题的一种方法。

3.3.3 长视频上的表现

数据集中视频长度平均只有67帧，然而实际

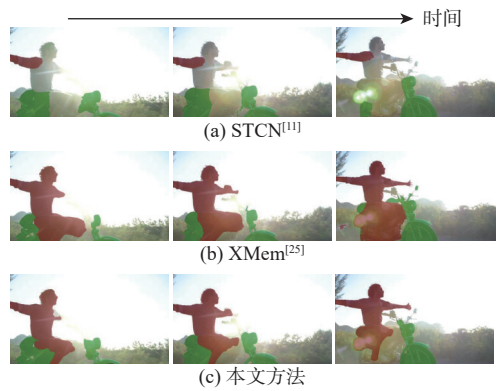


图4 本文方法的局限性

Fig. 4 Limitations of the proposed method



图5 DAVIS验证集和YouTube-VOS 2019验证集上与其他方法的对比

Fig. 5 Comparison with other methods on DAVIS and YouTube-VOS validation datasets



图 6　DAVIS 验证集和 YouTube-VOS 2019 验证集上更多的定性结果

Fig. 6　More qualitative results on DAVIS and YouTube-VOS validation datasets

应用中视频帧数更多。为证明本文方法在实际应用中的有效性, 在长视频[9]上进行实验。如图 7 所示, 随着视频长度的增加, 本文方法的速度几乎没有受到影响, 而 STCN[11] 的速度明显下降。

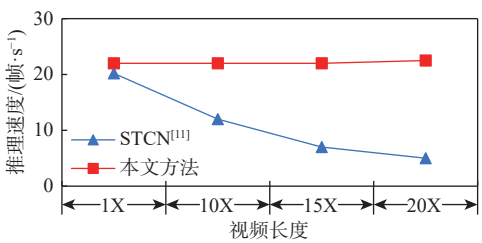


图 7　STCN[11] 和本文方法在速度上的对比

Fig. 7　Speed comparison of STCN[11] and the proposed method

3.4　消融实验

表 5 和表 6 对各模块进行了消融研究。本文提出的高层语义辅助损失、内存更新感应器和时空约束模块都可以提升模型的性能。如表 6 所示, 在

表 5　时空约束模块和高层语义辅助损失消融实验结果

Table 5　Ablation experiment results on spatial constraint modules and high-level semantic auxiliary losses　%

消融项	*J*&*F*	*J*	*F*
不使用时空约束	84.0	80.6	87.3
不使用辅助损失	83.3	80.2	86.4
同时使用时空约束和辅助损失	84.5	81.2	87.7

表 6　内存更新感应器消融实验结果

Table 6　Ablation experiment results on memory update sensor　%

内存更新方式	*J*&*F*	$J_{\text{seen/unseen}}$&$F_{\text{seen/unseen}}$
固定间隔(3帧)	84.5	81.9
固定间隔(4帧)	84.2	82.4
使用内存更新感应	84.5	82.4

注: $J_{\text{seen/unseen}}$&$F_{\text{seen/unseen}}$表示在YouTube-VOS 2019验证集上的结果。

DAVIS验证集上使用固定间隔3帧,可以获得最佳性能。而在YouTube-VOS 2019验证集上,需要将固定间隔增加到4才能获得最佳性能。使用本文提出的内存更新感应器可以根据视频的运动模式自适应地选择合适的更新间隔,适应不同数据集,获得更好的性能。

4 结 论

本文提出一种基于动态嵌入特征的鲁棒半监督视频目标分割方法,通过实验分析得出以下结论:

1)可以提升物体表征能力,对复杂背景中相似干扰物体有更好的鉴别能力。

2)可以有效提高记忆内存使用效率,并且内存损耗不会随视频时长推移而增加。

3)在DAVIS验证集和YouTube-VOS 2019验证集上的对比实验结果显示,本文方法分割效果优于多种先进网络模型方法,大幅提高了长时视频中的目标分割精度。

后续将继续探索长时视频中时空一致的记忆关联特性,减少内存记忆冗余,进一步提高分割精度;同时优化网络构建方法,减少网络参数量,提高推理速度。

参考文献(References)

[1] PERAZZI F, KHOREVA A, BENENSON R, et al. Learning video object segmentation from static images[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2017: 3491-3500.

[2] VOIGTLAENDER P, LEIBE B. Online adaptation of convolutional neural networks for video object segmentation[EB/OL]. (2017-08-01)[2023-06-01]. https://arxiv.org/abs/1706.09364v2.

[3] CAELLES S, MANINIS K K, PONT-TUSET J, et al. One-shot video object segmentation[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2017: 5320-5329.

[4] YANG Z X, WEI Y C, YANG Y. Collaborative video object segmentation by foreground-background integration[C]//Proceedings of the European Conference on Computer Vision. Berlin: Springer, 2020: 332-348.

[5] YANG Z X, WEI Y C, YANG Y. Collaborative video object segmentation by multi-scale foreground-background integration[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 44(9): 4701-4712.

[6] ZHANG P, HU L, ZHANG B, et al. Spatial consistent memory network for semi-supervised video object segmentation[C]//Proceedings of the DAVIS Challenge on Video Object Segmentation. Piscataway: IEEE Press, 2020: 1-4.

[7] SEONG H, HYUN J, KIM E. Kernelized memory network for video object segmentation[C]//Proceedings of the European Conference on Computer Vision. Berlin: Springer, 2020: 629-645.

[8] SEONG H, OH S W, LEE J Y, et al. Hierarchical memory matching network for video object segmentation[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE Press, 2021: 12869-12878.

[9] LI M X, HU L, XIONG Z W, et al. Recurrent dynamic embedding for video object segmentation[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2022: 1322-1331.

[10] OH S W, LEE J Y, XU N, et al. Video object segmentation using space-time memory networks[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE Press, 2019: 9226-9235.

[11] CHENG H K, TAI Y W, TANG C K. Rethinking space-time networks with improved memory coverage for efficient video object segmentation[EB/OL]. (2021-10-08)[2023-06-01]. https://arxiv.org/abs/2106.05210?context=cs.CV.

[12] CHENG H K, TAI Y W, TANG C K. Modular interactive video object segmentation: interaction-to-mask, propagation and difference-aware fusion[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2021: 5555-5564.

[13] XIE H Z, YAO H X, ZHOU S C, et al. Efficient regional memory network for video object segmentation[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2021: 1286-1295.

[14] 季传俊, 陈亚当, 车洵. 融合视觉词与自注意力机制的视频目标分割[J]. 中国图象图形学报, 2022, 27(8): 2444-2457.
JI C J, CHEN Y D, CHE X. Visual words and self-attention mechanism fusion based video object segmentation method[J]. Journal of Image and Graphics, 2022, 27(8): 2444-2457(in Chinese).

[15] 征煜, 陈亚当, 郝川艳. 特征一致性约束的视频目标分割[J]. 中国图象图形学报, 2020, 25(8): 1558-1566.
ZHENG Y, CHEN Y D, HAO C Y. Video object segmentation algorithm based on consistent features[J]. Journal of Image and Graphics, 2020, 25(8): 1558-1566(in Chinese).

[16] LI Y, SHEN Z R, SHAN Y. Fast video object segmentation using the global context module[C]//Proceedings of the European Conferenceon Computer Vision. Berlin: Springer, 2020: 735-750.

[17] WANG H C, JIANG X L, REN H B, et al. SwiftNet: real-time video object segmentation[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2021: 1296-1305.

[18] LIANG Y Q, LI X, JAFARI N, et al. Video object segmentation with adaptive feature bank and uncertain-region refinement[EB/OL]. (2020-10-15)[2023-06-01]. https://arxiv.org/abs/2010.07958v1.

[19] CHEN Y D, HAO C Y, YANG Z X, et al. Fast target-aware learning for few-shot video object segmentation[J]. Science China Information Sciences, 2022, 65(8): 182104.

[20] CHO S, LEE H, LEE M, et al. Tackling background distraction in video object segmentation[C]//Proceedings of the European Conference on Computer Vision. Berlin: Springer, 2022: 446-462.

[21] LUITEN J, VOIGTLAENDER P, LEIBE B. PReMVOS: proposal-generation, refinement and merging for video object segmentation [C]//Proceedings of the European Conference on Computer Vision. Berlin: Springer, 2019: 565-580.

[22] DUKE B, AHMED A, WOLF C, et al. SSTVOS: sparse spatio-temporal Transformers for video object segmentation[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2021: 5908-5917.

[23] GE W B, LU X K, SHEN J B. Video object segmentation using global and instance embedding learning[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2021: 16831-16840.

[24] LIU Y, YU R, YIN F, et al. Learning quality-aware dynamic memory for video object segmentation[C]//Proceedings of the European Conference on Computer Vision. Berlin: Springer, 2022: 468-486.

[25] CHENG H K, SCHWING A G. XMem: long-term video object segmentation with an Atkinson-Shiffrin memory model[C]//Proceedings of the European Conference on Computer Vision. Berlin: Springer, 2022: 640-658.

[26] LAN M, ZHANG J, HE F X, et al. Siamese network with interactive Transformer for video object segmentation[J]. Proceedings of theAAAIConferenceonArtificialIntelligence,2022,36(2):1228-1236.

[27] CHEN L C, PAPANDREOU G, KOKKINOS I, et al. DeepLab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018, 40(4): 834-848.

[28] 陈亚当, 陈柳任, 余文斌, 等. 多尺度特征融合的知识蒸馏异常检测方法[J]. 计算机辅助设计与图形学学报, 2022, 34(10): 1542-1549.
CHEN Y D, CHEN L R, YU W B, et al. Knowledge distillation anomaly detection with multi-scale feature fusion[J]. Journal of Computer-Aided Design & Computer Graphics, 2022, 34(10): 1542-1549(in Chinese).

[29] WOO S, PARK J, LEE J Y, et al. CBAM: convolutional block attention module[C]//Proceedings of the European Conference on Computer Vision. Berlin: Springer, 2018: 3-19.

Robust semi-supervised video object segmentation based on dynamic embedding feature

CHEN Yadang[1, 2, *], ZHAO Yibing[1, 2], WU Enhua[3]

(1. School of Computer Science, Nanjing University of Information Science and Technology, Nanjing 210044, China;
2. Wuxi Research Institute, Nanjing University of Information Science and Technology, Wuxi 214100, China;
3. Institute of Software, Chinese Academy of Sciences, Beijing 100190, China)

Abstract: A semi-supervised video object segmentation (VOS) method was proposed to address the issues of increasing memory consumption during inference and the difficulty of training relying solely on low-level pixel features. The method is based on dynamic embedding features and an auxiliary loss function. First, a dynamic embedding feature was employed to establish a constant-sized memory bank. Through spatiotemporal aggregation, historical information was utilized to generate and update dynamic embedding features. Simultaneously, a memory update sensor was employed to adaptively control the update interval of the memory bank, accommodating different motion patterns in various videos. Second, an auxiliary loss function was utilized to provide the network with guidance at the high semantic feature level, enhancing model accuracy and training efficiency by offering diverse guidance across multiple feature levels. Finally, to address the issue of misalignment between similar objects in the foreground and background of videos, a spatial constraint module was designed, which leveraged the temporal continuity of videos to better capture the correlation between the mask from the previous frame and the current frame. Experimental results demonstrate that the proposed method achieves an accuracy of 84.5% *J&F* on the DAVIS 2017 validation set and 82.4% *J&F* on the YouTube-VOS 2019 validation set.

Keywords: video object segmentation; spatiotemporal memory network; spatiotemporal constraint; memory update sensing; dynamic embedding feature

Received: 2023-06-13; **Accepted**: 2023-12-01; **Published Online**: 2024-01-10 15:11

URL: link.cnki.net/urlid/11.2625.V.20240110.0919.001

Foundation items: National Natural Science Foundation of China (62473201,62477026); Preliminary Research Project on Leading Technologies by Wuxi Industrial Innovation Research Institute

* **Corresponding author**. E-mail: adamchen@nuist.edu.cn

http://bhxb.buaa.edu.cn jbuaa@buaa.edu.cn
DOI: 10.13700/j.bh.1001-5965.2023.0365

基于联合交互注意力的图文情感分析方法

胡慧君[1,2]，丁子毅[1,2]，张耀峰[3,*]，刘茂福[1,2]

(1. 武汉科技大学 计算机科学与技术学院，武汉 430065；2. 武汉科技大学 智能信息处理与实时工业系统湖北省重点实验室，武汉 430065；3. 湖北经济学院 湖北数据与分析中心，武汉 430205)

摘　　要：社交媒体中的图文情感对于引导舆论走向具有重要意义，越来越受到自然语言处理（NLP）领域的广泛关注。当前，社交媒体图文情感分析的研究对象主要为单幅图像文本对，针对无时序性及多样性的图集文本对的研究相对较少，为有效挖掘图集中图像与文本之间情感一致性信息，提出基于联合交互注意力的图文情感分析（SA-JIA）方法。该方法使用 RoBERTa 和双向门控循环单元 (Bi-GRU) 来提取文本表达特征，使用 ResNet50 获取图像视觉特征，利用联合注意力来找到图文情感信息表达一致的显著区域，获得新的文本和图像视觉特征，采用交互注意力关注模态间的特征交互，并进行多模态特征融合，进而完成情感分类任务。在 IsTS-CN 数据集和 CCIR20-YQ 数据集上进行了实验验证，结果表明：所提方法能够提升社交媒体图文情感分析的性能。

关 键 词：社交媒体；图文情感分析；联合注意力；交互注意力；多模态融合

中图分类号：TP391

文献标志码：A　　**文章编号**：1001-5965(2025)07-2262-09

随着网络技术的日益发展，越来越多的用户选择通过社交媒体表达观点，社交媒体已成为人们分享和日常交流中必不可少的工具，而用户在表达观点时，不仅会发布文本内容，还会配以图像内容，这些图像信息通常蕴含着用户的情感信息。因此，社交媒体的图文情感分析与理解也逐渐受到广泛关注。

事实上，用户在编辑社交媒体文案时，通常配以多幅图像。本文将包含多幅图像的集合统称为图集。然而，现有社交媒体情感分析多以文本[1]或单幅图文对[2]为主，少见针对图集文本对的情感分析研究。Truong 等[3]开展了面向图集文本对的社交媒体情感分析研究，指出图集中每幅图像对社交媒体整体情感判断的重要性有所不同，因此，以图像视觉特征来对齐文本，针对数据来源单一的餐厅评论图集进行了图文情感分析。而本文聚焦于社交媒体图集，数据来源多样，蕴含情感信息更加丰富，更具有挑战性。随着信息量的不断增加和网络新词的频繁出现，影响文本情感的情感因子信息粒度加大，而情感词典则只是固定长度的情感词汇或短语，虽然准确但难以及时更新，在颗粒度上也很难适应社交媒体上迅速更迭的情感数据。现如今人们常以文本结合图像的方式表达情感，选取能够相互映射的文本和图像来突显自身的情感。但由于图集中图像所表达情感的多样性，从而导致文本与图集之间的情感会表达不同。可以看出，现有方法未能充分考虑图集中部分图像与文本情感不一致的问题。例如：文本“#当代大学生暑假日常#你鲨了我吧，生活压得我喘不过气”，配图为一张生活压垮自己的表情包和一张带有微笑的自拍照。从“鲨了我吧”和配图表情包可以看出，图文的整体情感与文本的情感均表达为负面情感。从配图表

收稿日期：2023-06-15；**录用日期**：2023-10-12；**网络出版时间**：2023-11-23 15:37

网络出版地址：link.cnki.net/urlid/11.2625.V.20231122.1324.001

基金项目：国家自然科学基金 (62271359)；国家社科基金重点项目 (23ATJ005)；“十四五”湖北省优势特色学科（群）项目 (2023D0302)；湖北省教育厅科研重点项目 (20192202)

* **通信作者**. E-mail：yfzhang@hbue.edu.cn

引用格式：胡慧君，丁子毅，张耀峰，等. 基于联合交互注意力的图文情感分析方法 [J]. 北京航空航天大学学报，2025，51（7）：2262-2270.
HU H J，DING Z Y，ZHANG Y F，et al. Analysis of image and text sentiment method based on joint and interactive attention [J]. Journal of Beijing University of Aeronautics and Astronautics，2025，51（7）：2262-2270 （in Chinese）.

情包可以看出, 图像表达为负面情感, 与图文整体情感保持一致, 但是带微笑的自拍照却表达正面情感, 与图文整体情感不一致。这种图集中存在图像与文本表达情感不一致的现象, 会影响图集文本情感分析的效果。

为解决上述问题, 有效挖掘图集中图像和文本情感一致性信息, 本文提出基于联合交互注意力的社交媒体图文情感分析(images-text sentiment analysis in social media based on joint and interactive attention, SA-JIA)方法, 该方法采用联合注意力关注文本与图集之间的情感一致性表达; 将经过联合注意力得到的图文增强特征输入到交互注意力中进行图文的特征融合, 进而完成图文情感分类任务。

本文的主要创新点包括以下 3 个方面:

1) 本文采用联合注意力来关注图文特征的情感显著区域, 对提取的图文特征进行模态间情感信息一致性匹配, 从而增强图文特征表示。

2) 本文提出交互注意力, 对增强的图像特征和文本特征经交互注意力来实行特征信息互补并进行图文特征融合, 进而完成图文情感分析。

3) 为证明 SA-JIA 方法的有效性, 在中文社交媒体图文情感数据集 IsTS-CN 和疫情情感数据集 CCIR20-YQ 上进行了实验验证。与现有主流方法相比, 本文方法在中文社交媒体图集文本对情感分析效果与性能中取得了更好的结果。

1　相关工作

1.1　文本情感分析

早期的文本情感分析主要为基于情感词典的方法。Stone[4] 通过构建情感词典来计算文本的整体情感极性。随着机器学习和深度学习的发展, 对于文本情感分析的方法也逐渐变得复杂多样, Wiebe 等[5] 构建大型语料库, 利用机器学习分类器建立基于情感的预测模型。Basiri 等[6] 利用 2 个独立的双向长短期记忆(bidirectional long short term memory, Bi-LSTM)网络和门控循环单元(gated recurrent unit, GRU)来提取时间流中过去和未来的上下文信息, 极大提高了文本情感分析的准确性。

近年来, 以基于双向变换器的编码器表示(bidirectional encoder representation from transformers, BERT)[7] 为主的预训练语言模型被广泛应用于情感分析领域。Dai 等[8] 基于鲁棒优化的 BERT 预训练(robustly optimized BERT pretraining approach, RoBERTa)[9] 方法生成以情感词为导向的依存句法树, 并将其用于文本方面级情感分析任务。Wu 等[10] 提出一种相对位置编码层, 整合给定方面词的位置信息, 并使用方面注意力机制考虑方面词和上下文词之间的语义关系。

虽然文本能独立地表示一定的情感, 但是人们进行交流总是通过信息的综合表现来进行。因此, 多模态的情感分析更符合人们对情感的感知, 更符合人们表达情感的模式。研究结论也表明, 相比单一文本情感分析, 多模态情感分析效果更好。

1.2　图文情感分析

随着社交媒体迅猛发展, 用户开始以图文方式分享内容, 多模态情感分析吸引了研究者的注意。不同于文本情感分析, 多模态情感分析不仅要学习单模态特征, 并且需要进行模态融合, 捕捉不同模态间的交互信息。Wu 等[11] 提出基于多头自注意力机制的融合网络, 通过合理地分配声学-视觉、声学-文本等特征权重, 获得重要的情感特征。但是该模型无法获得模态间的层次关联信息, 并且会引入大量的噪声, 造成特征冗余。Han 等[12] 提出双向双模态融合网络, 对 2 种模态特征表示进行差异增量操作, 进而提取模态间的相关信息。王靖豪等[13] 提出一种基于多层次特征融合注意力网络, 通过计算“图文”与“文图”特征, 实现多模态情感特征互补。Wen 等[14] 提出跨模态上下文门控卷积网络, 并引入了跨模态上下文门的概念, 使其能有效地捕获模态间的交互信息, 同时减少不相关信息的影响, 从而生成具有情感语义信息的图文特征。并且, 研究者在单模态预训练模型基础上提出包括多模式双向 Transformer (multimodal bitransformers MMBT)[15]、EF-CapTrBERT[16] 在内的多模态预训练模型, 刻画语言和视觉之间的关联信息, 虽能在图像文本检索、视觉问答等诸多下游任务中取得不错效果, 但这些多模态预训练模型多针对单幅图文对, 未能考虑图集与文本之间的内在联系。

1.3　Transformer 网络

Transformer 网络[17] 较早被引入用于神经机器翻译 (neural machine translation, NMT) 任务, 其中, 编码器和解码器端各自利用包含自注意力的 Transformer。在每层自注意力之后, 编码器和解码器通过一个额外的解码器子层连接, 其中, 解码器针对目标文本的每个元素关注源文本的每个元素。Tsai 等[18] 在 Transformer 的基础上设计了一种未对齐多模态序列的多模态 Transformer 来关注跨不同时间步长的多模态序列之间的交互, 并潜在地使一种模态适应另一种模态, 以端到端的形式解决了跨模态之间数据未对齐的问题。本文将 Transformer 应用到多模态, 使用交互注意力替代原有结构, 用以关注图文模态间的特征信息互补, 进而对图文

特征进行融合。

分析上述研究工作发现，已有社交媒体图文情感分析研究存在如下问题：①多模态社交媒体的图像大多不止一幅，而现有研究主要针对单幅图文对进行情感分析，针对图像多样性、无时序性的图集文本对的情感分析研究相对较少；②现有的方法未能充分考虑到图集与文本之间的关系。因此，为有效挖掘图集与文本之间情感一致性的信息，本文提出SA-JIA方法。

2 本文方法

本文SA-JIA方法可以划分为特征提取模块、联合注意力模块、交互注意力融合模块，本文方法框架如图1所示。

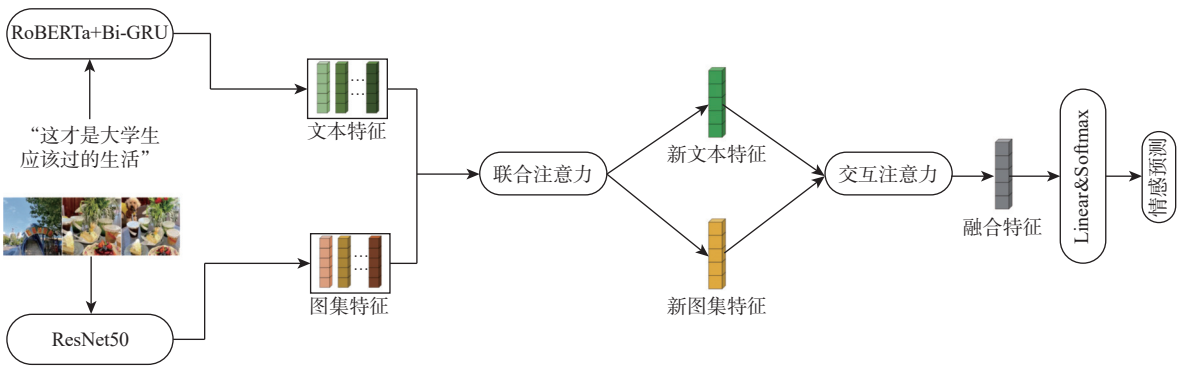


图1 SA-JIA方法框架

Fig. 1 SA-JIA method framework

特征提取模块使用现有的图像和文本特征提取的方法，分别提取图像和文本中表达情感的特征；联合注意力模块使用注意力挖掘图集文本一致性情感信息，以增强图集和文本的特征表达；交互注意力以Transformer为基础，用以关注图文特征之间的互补信息并融合输出，经Linear和Softmax后，对情感进行分类预测。

2.1 特征提取模块

首先，需要单独提取文本和图集的特征。对于文本数据，由于RoBERTa可以学习到更高层次的语义信息，而双向GRU（Bidirectional GRU，Bi-GRU）[19]可以学习到更低层次的语义信息，两者结合可以在编码文本时实现多层次、多粒度的语义信息提取，从而更好地学习文本的语义表示。因此，本文采用RoBERTa和Bi-GRU相结合的方式对文本进行特征提取。使用预训练好的RoBERTa对文本进行字编码，得到文本特征$\boldsymbol{F}_{\boldsymbol{T}-\mathrm{RoBERTa}}$：

$$\boldsymbol{F}_{\boldsymbol{T}-\mathrm{RoBERTa}}=\left[\boldsymbol{f}_1,\boldsymbol{f}_2,\cdots,\boldsymbol{f}_m,\cdots,\boldsymbol{f}_N\right]\in\mathbf{R}^{d\times N} \tag{1}$$

式中：下标$\boldsymbol{T}$为文本输入；$\boldsymbol{f}_m$为第m个字的上下文语义特征；N为RoBERTa字编码最大长度；d为RoBERTa输出维度，取值768。

其次，采用Bi-GRU进一步提取文本的上下文信息：

$$\boldsymbol{h}_m=\left[\overrightarrow{\mathrm{GRU}}(\boldsymbol{f}_m)\oplus\overleftarrow{\mathrm{GRU}}(\boldsymbol{f}_m)\right]\in\mathbf{R}^{H} \tag{2}$$

式中：$\overrightarrow{\mathrm{GRU}}$、$\overleftarrow{\mathrm{GRU}}$表示双向门控循环单元；字$\boldsymbol{f}_m$的$\boldsymbol{h}_m$经过Bi-GRU得到；$H=d_h\times2$为Bi-GRU隐藏层维度2倍；⊕表示进行异或运算，d_h为Bi-GRU的隐藏层维度。

最后，可由此得到最终的文本语义特征$\boldsymbol{F}_{\boldsymbol{T}}$为

$$\boldsymbol{F}_{\boldsymbol{T}}=[\boldsymbol{h}_1,\boldsymbol{h}_2,\cdots,\boldsymbol{h}_{\mathrm{N}}]\in\mathbf{R}^{H\times N} \tag{3}$$

对于图像数据，由于ResNet50引入了残差块，减轻了梯度消失的问题。ResNet50可以处理不同大小、角度和光照条件下的图像，而且ResNet50具有更深的网络结构，可以更好地提取图像特征。因此，本文采用ResNet50对图像进行特征提取。首先，将图像设定为K幅图像；然后，采取迁移学习方式，使用在ImageNet数据集[20]上预训练的ResNet50模型[21]提取每幅图像的视觉特征。将每幅图像裁剪成224×224大小作为ResNet50输入$\boldsymbol{I}_i$，对所有图像特征进行拼接得到图集视觉特征$\boldsymbol{F}_{\boldsymbol{I}_r}$：

$$\boldsymbol{F}_{\boldsymbol{I}_r}=\left[\boldsymbol{F}_{\boldsymbol{I}_1},\boldsymbol{F}_{\boldsymbol{I}_2},\cdots,\boldsymbol{F}_{\boldsymbol{I}_K}\right]\in\mathbf{R}^{1\,000\times K} \tag{4}$$

使用单层感知机将视觉特征$\boldsymbol{F}_{\boldsymbol{I}_r}$换成与文本特征具有相同维度的向量，得到最终的图集视觉特征$\boldsymbol{F}_{\boldsymbol{I}}$：

$$\boldsymbol{F}_{\boldsymbol{I}}=\tanh(\boldsymbol{W}_{\boldsymbol{I}}\boldsymbol{F}_{\boldsymbol{I}_r})\in\mathbf{R}^{H\times K} \tag{5}$$

式中：$\boldsymbol{W}_{\boldsymbol{I}}\in\mathbf{R}^{H\times1\,000}$为可训练权重矩阵。

2.2 联合注意力模块

为捕捉图集和文本之间的情感一致性信息，SA-JIA方法采用联合注意力捕捉文本和图集之间的一致性信息，对图文情感表达一致的特征权重进行增强，而图文情感表达不一致的特征权重进行减弱。

联合注意力的结构如图2所示，通过计算得到

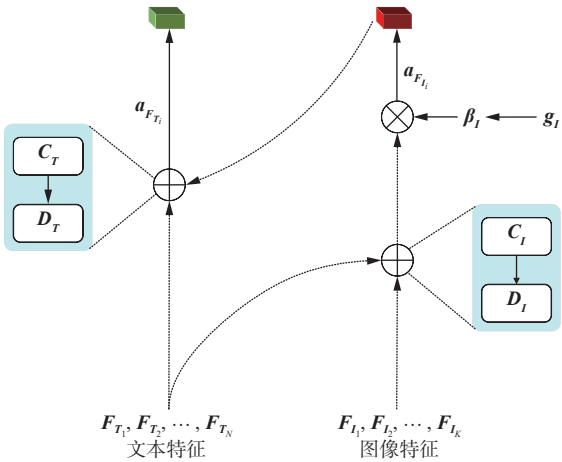


图 2　联合注意力

Fig. 2　Joint attention

图像和文本的交叉矩阵$\boldsymbol{C}_I$:

$$\boldsymbol{C}_I = \tanh(\boldsymbol{F}_I^{\mathrm{T}}\boldsymbol{W}_C\boldsymbol{F}_T) \in \mathbf{R}^{K\times N} \tag{6}$$

式中: 图像特征$\boldsymbol{F}_I \in \mathbf{R}^{H\times K}$, 文本特征$\boldsymbol{F}_T \in \mathbf{R}^{H\times N}$已由 2.1 节得出; $\boldsymbol{W}_C \in \mathbf{R}^{H\times H}$为可训练权重矩阵。

文本引导的图文相似度矩阵为

$$\boldsymbol{D}_I = \tanh\left(\boldsymbol{W}_{F_I}\boldsymbol{F}_I + (\boldsymbol{W}_{F_T}\boldsymbol{F}_T)\boldsymbol{C}_I^{\mathrm{T}}\right) \in \mathbf{R}^{K\times K} \tag{7}$$

式中: $\boldsymbol{W}_{F_I} \in \mathbf{R}^{K\times H}$、$\boldsymbol{W}_{F_T} \in \mathbf{R}^{K\times H}$为可训练权重矩阵。其次, 引入门控机制生成图像情感权重矩阵$\boldsymbol{\beta}_I$和门控机制中对情感一致性信息的表达$\boldsymbol{g}_I$:

$$\boldsymbol{g}_I = \sigma(\boldsymbol{W}_{F_I}\boldsymbol{D}_I) \tag{8}$$

$$\boldsymbol{\beta}_I = \boldsymbol{W}_s^{\mathrm{T}}\tanh(\boldsymbol{g}_I) \tag{9}$$

式中: $\boldsymbol{W}_{F_I} \in \mathbf{R}^{H\times H}$和$\boldsymbol{W}_s \in \mathbf{R}^{K\times H}$为可训练权重矩阵。计算得到视觉注意力权重$\boldsymbol{a}_{F_I}$:

$$\boldsymbol{a}_{F_I} = \mathrm{softmax}(\boldsymbol{\beta}_I) \in \mathbf{R}^{K} \tag{10}$$

对特征进行加权求和获得包含文本一致性信息的图集视觉特征$\tilde{\boldsymbol{F}}_I$为

$$\tilde{\boldsymbol{F}}_I = \sum_{i=1}^{K} \boldsymbol{a}_{F_{I_i}}\boldsymbol{F}_{I_i} \in \mathbf{R}^{H\times K} \tag{11}$$

式中: $\boldsymbol{a}_{F_I} = [\boldsymbol{a}_{F_{I_1}}, \boldsymbol{a}_{F_{I_2}}, \cdots, \boldsymbol{a}_{F_{I_K}}]$; $\boldsymbol{F}_{I_i}$为第 i 幅图像的视觉特征表示。由于获得包含文本一致性情感信息的增强的图集特征, 因此, 可以直接由图集特征引导的文本注意力特征权重$\boldsymbol{a}_{F_T}$, 获得带有文本注意力的文本语义特征$\tilde{\boldsymbol{F}}_T$为

$$\boldsymbol{C}_T = \tanh(\boldsymbol{F}_T^{\mathrm{T}}\boldsymbol{W}_C\tilde{\boldsymbol{F}}_I) \in \mathbf{R}^{K\times N} \tag{12}$$

$$\boldsymbol{D}_T = \tanh\left(\boldsymbol{W}_{F_T}\boldsymbol{F}_T + (\boldsymbol{W}_{\tilde{F}_I}\tilde{\boldsymbol{F}}_I)\boldsymbol{C}_T^{\mathrm{T}}\right) \in \mathbf{R}^{N\times N} \tag{13}$$

$$\boldsymbol{a}_{F_T} = \mathrm{softmax}(\boldsymbol{D}_T) \in \mathbf{R}^{N} \tag{14}$$

$$\tilde{\boldsymbol{F}}_T = \sum_{i=1}^{N} \boldsymbol{a}_{F_{T_i}}\boldsymbol{F}_T \in \mathbf{R}^{H\times N} \tag{15}$$

式中: $\boldsymbol{a}_{F_T} = [\boldsymbol{a}_{F_{T_1}}, \boldsymbol{a}_{F_{T_2}}, \cdots, \boldsymbol{a}_{F_{T_N}}]$; $\boldsymbol{D}_T$为图像引导的图文相似度; $\boldsymbol{W}_{F_T} \in \mathbf{R}^{N\times H}$、$\boldsymbol{W}_{\tilde{F}_I} \in \mathbf{R}^{N\times H}$为可训练权重矩阵。得到的新的文本特征$\tilde{\boldsymbol{F}}_T$和新的图像特征$\tilde{\boldsymbol{F}}_I$即为增强的图文特征。

2.3　交互注意力融合模块

为关注图集和文本之间的情感互补信息, 更好地进行模态的交互融合, SA-JIA 方法在 Transformer[17] 的基础上, 设计了一种融合不同模态信息的交互注意力方法来提供不同模态之间的交互融合。

交互注意力的结构如图 3 所示, 在注意力机制中, 缩放点积注意力问题的输入由 Queries、Keys、Values 组成。对于文本向着图集进行交互融合, 图集特征作为 Queries, 将文本特征作为 Keys 和 Values, 表达 Queries 为$\boldsymbol{Q}_I = \boldsymbol{W}_{Q_I}$, Keys 为$\boldsymbol{K}_T = \boldsymbol{W}_{K_T}$, 并将 Values 表达为$\boldsymbol{V}_I = \boldsymbol{W}_{V_T}$。从文本向着图集进行交互的注意力表示$\boldsymbol{f}_I$为

$$\boldsymbol{f}_I = \boldsymbol{I}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}) = \mathrm{softmax}\left(\boldsymbol{Q}_I(\boldsymbol{K}_T)^{\mathrm{T}}/\sqrt{d_K}\right)\boldsymbol{V}_T \tag{16}$$

式中: $\boldsymbol{W}_{Q_I} \in \mathbf{R}^{d_I}$, $\boldsymbol{W}_{K_T} \in \mathbf{R}^{d_T}$, $\boldsymbol{W}_{V_T} \in \mathbf{R}^{d_T}$为可训练权重矩阵; d_T为文本特征维度; d_I为图集特征维度; d_K为键向量的维度。

同理, 从图集向着文本进行交互融合, 将文本特征作为 $\boldsymbol{Q}$, 将图集特征作为 $\boldsymbol{K}$ 和 $\boldsymbol{V}$。从图集向文本进行适应融合的注意力表示$\boldsymbol{f}_T$为

$$\boldsymbol{f}_T = \boldsymbol{T}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}) = \mathrm{softmax}\left(\boldsymbol{Q}_T(\boldsymbol{K}_I)^{\mathrm{T}}/\sqrt{d_K}\right)\boldsymbol{V}_I \tag{17}$$

式中: $\boldsymbol{W}_{Q_T} \in \mathbf{R}^{d_T}$, $\boldsymbol{W}_{K_I} \in \mathbf{R}^{d_I}$, $\boldsymbol{W}_{V_I} \in \mathbf{R}^{d_I}$为可训练权重矩阵。

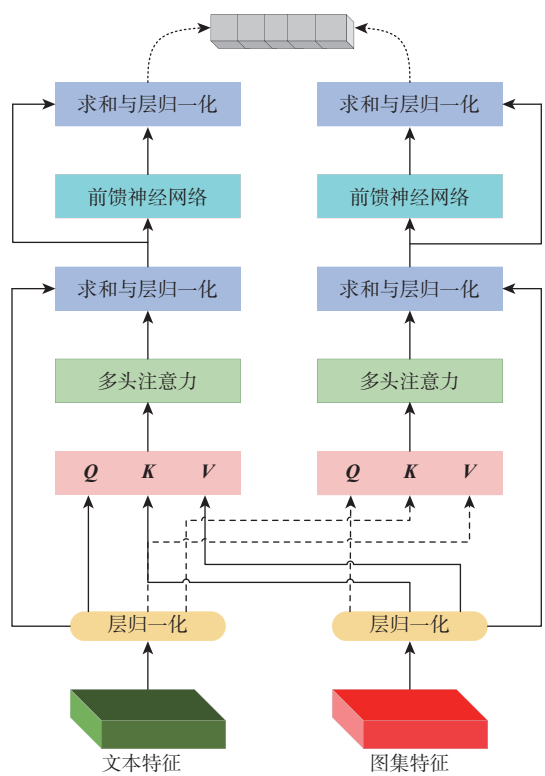


图 3　交互注意力

Fig. 3　Interactive attention

经过前馈神经网络(feedforward neural network, FNN)[22]和层归一化(layer normalization, LN)[23]后得到文本交互特征$\boldsymbol{V}_T$:

$$\boldsymbol{V}_T = \mathrm{LN}(\mathrm{FNN}(\boldsymbol{f}_T(\tilde{\boldsymbol{F}}_T, \tilde{\boldsymbol{F}}_I, \tilde{\boldsymbol{F}}_I))) \tag{18}$$

式中: LN 表示层归一化; FNN 表示前馈神经网络。

同理, 经过交互注意力模块的图集交互特征$\boldsymbol{V}_I$为

$$\boldsymbol{V}_I = \mathrm{LN}(\mathrm{FNN}(\boldsymbol{f}_I(\tilde{\boldsymbol{F}}_I, \tilde{\boldsymbol{F}}_T, \tilde{\boldsymbol{F}}_T))) \tag{19}$$

将文本交互特征$\boldsymbol{V}_T$与图集交互特征$\boldsymbol{V}_I$进行拼接得到所需的图文情感融合特征$\boldsymbol{V}$:

$$\boldsymbol{V} = \boldsymbol{V}_T \oplus \boldsymbol{V}_I \tag{20}$$

将图文特征向量$\boldsymbol{V}$输入全连接层得到结果向量$\boldsymbol{P}_{\mathrm{r}}$, 再经过 softmax 函数得到最终的情感分类结果p:

$$\boldsymbol{P}_{\mathrm{r}} = \mathrm{Linear}(\boldsymbol{W}_{P_{\mathrm{r}}} \boldsymbol{V}) \tag{21}$$

$$p = \mathrm{softmax}(\boldsymbol{W}_p \boldsymbol{P}_{\mathrm{r}}) \tag{22}$$

式中: $\boldsymbol{W}_{P_{\mathrm{r}}}$, $\boldsymbol{W}_p$为可训练权重矩阵。

3 实 验

3.1 数据集

为验证 SA-JIA 方法的有效性, 本文在 IsTS-CN 数据集和 CCIR20-YQ 数据集上进行实验。

IsTS-CN 数据集是由曹梦丽[24]构建的中文社交媒体图文情感数据集。通过设置主题词与“考研”“毕业季”“校招”等与大学生相关的词汇, 进行人工标注后得到的 10 347 条图集文本对数据, 其中, 数据情感表达为正面、中性和负面的分别有 5 449 条、2 191 条和 2 707 条。数据集包含的图像共有 34 807 幅, 平均每条数据包含约有 3 幅图像。

CCIR20-YQ 数据集来源于全国信息检索学术会议 CCIR2020 的疫情期间网民情绪识别评测任务(https://www.datafountain.cn/competitions/423?CCIR2020), 通过设置与“新冠肺炎”相关的 230 个主题词, 进行数据采集并进行人工标注 10 万条数据, 情感标签为正面、中性和负面。本文选取 10 万条中包含图像的 67 556 条图文数据构成 CCIR20-YQ 数据集, 其中, 正面 19 639 条、负面 8 864 条、中性 39 053 条。CCIR20-YQ 数据集共有 138 320 幅图像, 平均每条数据约有 2 幅图像。

在实验过程中, 将 IsTS-CN 和 CCIR20-YQ 数据集按照 8∶1∶1 的比例划分为训练集、验证集、测试集, 数据集详细信息如表 1 所示。

表 1 数据集划分

Table 1 Dataset partitioning

数据集	训练集/条	验证集/条	测试集/条
IsTS-CN	8 277	1 035	1 035
CCIR20-YQ	54 044	6 756	6 756

3.2 实验设置

本文所有模块优化器选择 Adam, Bi-GRU 隐藏层的维度设置为 200, 最优的超参数组合根据在验证集上的实验表现确定, 其中, Dropout 下降率设置为 0.5, 损失函数 Loss 使用交叉熵损失函数 CrossEntropyLoss, 并在测试集上验证实验效果。经过统计, 图集中图像至多为 9 幅, 因此, 图文情感分析图集填充时输入的图片数量统一取值为 9。如表 2 所示。

表 2 SA-JIA 方法参数设置

Table 2 SA-JIA method parameters setting

学习率	句子最大长度	下降率	注意力机制头数	Loss	隐藏层维度
2×10^{-5}	200	0.5	5	CrossEntropyLoss	200

3.3 实验结果与分析

图文情感分析为分类任务, 所以主要使用精确率P、召回率R和F_1值作为评价指标。

3.3.1 对比实验分析

为评估 SA-JIA 方法的性能, 本文从文本和图文两方面进行对比实验。对比方法如下:

Att-Bi-LSTM[25]。基于 Bi-LSTM 与注意力机制的文本情感分析方法。

BERT[7]。基于 BERT 的文本情感分析方法, 通过 BERT 模型上的微调, 有效提取文本语义特征用于微博文本情感分析。

RoBERTa[9]。基于 RoBERTa 的文本情感分析方法。

VistaNet[3]。基于图像与文本对齐的图(集)文情感分析方法。

mBERT[26]。基于 BERT 与图文交互特征的图文情感分析方法。

MMBT[15]。多模态双向 Transformer 方法, 利用卷积神经网络架构增强纯文本表征。

EF-CapTrBERT[16]。利用 BERT 实现基于输入空间转换的多模态目标感知分类方法。

TIBERT[27]。利用特征引导方法获得多模态特征, 然后, 拼接单峰特征并通过注意力模型进行多模态情感分类。

表3 为本文方法与对比方法在IsTS-CN 和CCIR20-YQ 这 2 个数据集上的图文情感分析实验结果。

相较于多模态图文情感分析, 基于文本的情感分析方法性能相对较差, 这说明引入图像视觉特征

表 3　IsTS-CN 和 CCIR20-YQ 数据集上的图文情感分析对比实验结果

Table 3　Comparative experimental results of image and text sentiment analysis on IsTS-CN and CCIR20-YQ datasets

模态	方法	P		R		F_1	
		IsTS-CN	CCIR20-YQ	IsTS-CN	CCIR20-YQ	IsTS-CN	CCIR20-YQ
文本	Att-Bi-LSTM[25]	0.574	0.691	0.601	0.583	0.566	0.607
	BERT[7]	0.617	0.601	0.608	0.634	0.612	0.617
	RoBERTa[9]	0.637	0.682	0.638	0.764	0.637	0.708
图文	VistaNet[3]	0.634	0.627	0.626	0.619	0.628	0.623
	mBERT[26]	0.733	0.709	0.723	0.702	0.727	0.705
	MMBT[15]	0.711	0.732	0.747	0.670	0.722	0.691
	EF-CapTrBERT[16]	0.723	0.717	0.720	0.692	0.721	0.703
	TIBERT[27]	0.716	0.695	0.726	0.687	0.721	0.692
	SA-JIA	0.722	0.741	0.778	0.708	0.734	0.718

能够有效改善单模态文本情感分析方法在社交媒体情感分析任务上的性能表现。而基于 RoBERTa 微调的情感分析方法优于其他方法, 原因在于, 该方法采用 Transformer 捕捉文本语义特征, 能够更加高效地获取文本语义信息, 同时, RoBERTa 在 BERT 上进行优化, 更能充分挖掘文本情感语义信息。

由于现有情感分析方法大多针对单幅图文对进行情感分析, 对于图集中情感信息的捕捉和抽象能力相对较弱, 因此, 图集文本情感分析效果欠佳。VistaNet 虽然针对图集进行情感分析, 但在所有图文情感分析方法中表现最差, 在 IsTS-CN 数据集与 CCIR20-YQ 数据集上较其他方法效果均较低, 可能原因是社交媒体数据表现更复杂, 社交媒体中蕴含着丰富的情感信息, 图集中图像情感表达复杂, 而上述方法未充分考虑到文本与图集之间的一致性表达而导致图文情感分析的效果变差。

由结果看出, SA-JIA 方法的实验结果优于其他主流方法, 表明 SA-JIA 方法能够有效处理社交媒体图文情感分析任务。当前性能表现较好的深度学习方法, 如 EF-CapTrBERT、TIBERT 方法虽然均采用预训练语言模型进行文本特征提取, 但具体的方法不同。EF-CapTrBERT 方法利用图像描述生成方法将图像翻译为辅助句子, 为语言模型提供多模态信息; TIBERT 方法则使用预训练的 ResNet 模型处理输入图像, 将图像与文本特征一起输入到多头注意力模型中, 对文本与图像特征进行融合。虽然以上做法在性能上接近或优于其他图文情感分析方法, 但性能改善并不显著, 这说明在图集文本情感分析任务中有效捕捉图集与文本之间的关联特征和情感信息对结果有重要影响。因此, SA-JIA 方法首先采用联合注意力机制来聚焦图集和文本之间的一致性情感信息, 以此来增强图集和文本的特征表达, 然后, 采用交互注意力来融合图集和文本特征进行情感分类, 最终在社交媒体图集文本情感分析效果上有提升, 在 IsTS-CN 数据集上和在 CCIR20-YQ 数据集上的实验结果可以看出, SA-JIA 方法相较于其他的方法都有较好的效果。

3.3.2　消融实验分析

为评估 SA-JIA 方法中不同模块、不同结构对于情感分析性能的影响程度, 在 IsTS-CN 和 CCIR20-YQ 数据集上对 SA-JIA 方法进行消融实验, 实验结果如表 4 所示, 消融实验所涉方法如下:

JIA-JA。使用联合注意力增强的图文特征融合的情感分析方法。

JIA-IA。使用交互注意力的图文特征融合情感分析方法。

SA-JIA。基于联合交互注意力的图文情感分析方法。

如表 4 所示, JIA-IA 方法在评测指标 F_1 值上要低于 JIA-JA 方法, 主要的原因在于, JIA-IA 方法只是简单地对文本和图集特征直接进行了拼接, 并未进行图文模态的情感一致性匹配增强, 导致图文之

表 4　IsTS-CN 和 CCIR20-YQ 数据集上的消融实验结果

Table 4　Ablation experimental results on IsTS-CN and CCIR20-YQ datasets

方法	P		R		F_1	
	IsTS-CN	CCIR20-YQ	IsTS-CN	CCIR20-YQ	IsTS-CN	CCIR20-YQ
JIA-JA	0.670	0.738	0.724	0.671	0.611	0.661
JIA-IA	0.649	0.718	0.614	0.611	0.602	0.608
SA-JIA	0.722	0.741	0.778	0.708	0.734	0.718

间的特征表示缺少一致性的关联，因此，情感分析效果要低于JIA-JA方法。而SA-JIA方法在评测指标上均高于JIA-IA，主要原因在于，SA-JIA方法首先引入联合注意力，从文本和图像2个方面引导下得到新的图文特征，减少噪声的同时增强了图文模态的情感一致性表达，然后，经过交互注意力处理，使文本和图像进行交互融合，从而提升了情感分析的效果。由实验结果可知，联合注意力和交互注意力的结合优于其他方法，也进一步说明了SA-JIA方法的有效性。

3.3.3 实验实例分析

本文对SA-JIA方法在数据集IsTS-CN上进行实例分析。

对于图4中的正确预测样例，文本标题为“#yqy的大学生活##晚安打工人#”，数据集标注的情感类别为中性，SA-JIA方法的预测情感类别也为中性。从文本信息看，标题“yqy的大学生活”“晚安打工人”是比较普通的一句话，是不带任何情感，从图集来看，每张图像展示的都只是普通城市风景照，也可以不带任何情感，因此，数据集标注情感为中性。SA-JIA方法很好地分析文本语义信息并捕捉到图集特征，所给出的情感预测类别与数据集标注一致，也为中性。

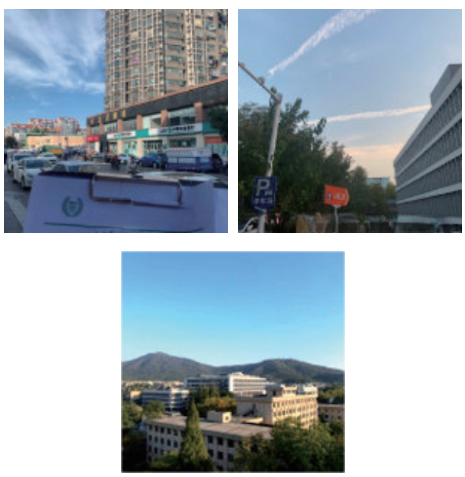


图4 正确预测样例

Fig. 4 Correct prediction example

对于图5中的错误预测样例，文本标题为“#毕业季#认识了814个日夜在第813天拍了毕业照”，数据集标注的整体图文情感类别为正面，但是SA-JIA方法给出的情感类别为中性。单独从文本信息看，“#毕业季#认识了814个日夜在第813天拍了毕业照”表达的情感可以被判断为中性；从图像的角度看，图像内容表现的是毕业的情景，情感表达可以判断为中性，并且，用户在发表图像时，有时候会将大小不一的图像合成一张长图，在处理图像时



图5 错误预测样例

Fig. 5 Incorrect prediction example

由于受神经网络的输入限制，会对图像进行分割，导致丢失部分视觉信息，因此，SA-JIA方法错误地将图文情感判断为中性情感，从而导致预测的情感类别错误。

总结分析结果发现，导致图文情感分析不准确的原因不仅在于文本的情感表达，而且图像处理的方式也会影响图文情感的判断，在一定程度上影响最终图文情感分析效果，因此，对图集的处理方式及图集与文本之间的信息关联极为重要。

4 结 论

1）本文SA-JIA方法解决了主流方法中未充分考虑图集文本对之间情感一致性信息匹配的问题。

2）本文方法通过联合注意力网络，从文本和图集2个方面去关注用户的情感表达，根据相关性来增强文本和图集的情感一致性表达；通过交互注意力对图文特征进行交互融合并进行社交媒体的图文情感分析，从而提升图文情感分析性能。

3）实验结果表明，本文方法与传统单独处理图文信息的方法不同，本文方法以图文信息互补，相互挖掘图文之间的语义关联，提供更全面的情感表达理解，为图集文本对情感分析提供了新的思考方向。

此外，本文方法仍存在局限性，图集中图像视觉信息存在不连续性，无法很好地对其整体处理获取图集视觉特征，在未来的工作中将尝试使用更多

的方法提取图集视觉信息。

参考文献（References）

[1] XIA E, YUE H, LIU H F. Tweet sentiment analysis of the 2020 U. S. presidential election[C]//Proceedings of the 30th Web Conference . New York: ACM, 2021.

[2] BONHEME L, GRZES M. SESAM at SemEval-2020 task 8: investigating the relationship between image and text in sentiment analysis of memes[C]//Proceedings of the Fourteenth Workshop on Semantic Evaluation. Barcelona : International Committee for Computational Linguistics, 2020.

[3] TRUONG Q T, LAUW H W. VistaNet: visual aspect attention network for multimodal sentiment analysis[C]//Proceedings of the AAAI Conference on Artificial Intelligence. Palo Alto: AAAI Press, 2019, 33(1): 305-312.

[4] STONE P J. Thematic text analysis: new agendas for analyzing text content[M]. London: Routledge, 2020: 35-54.

[5] WIEBE J M, BRUCE R F, O'HARA T P. Development and use of a gold-standard data set for subjectivity classifications[C]//Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics on Computational Linguistics. Stroudsburg: Association for Computational Linguistics, 1999.

[6] BASIRI M E, NEMATI S, ABDAR M, et al. ABCDM: an attention-based bidirectional CNN-RNN deep model for sentiment analysis[J]. Future Generation Computer Systems, 2021, 115: 279-294.

[7] DEVLIN J, CHANG M W, LEE K, et al. BERT: pre-training of deep bidirectional transformers for language understanding[C]//Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: human Language Technologies. Stroudsburg: Association for Computational Linguistics, 2019: 4171-4186.

[8] DAI J Q, YAN H, SUN T X, et al. Does syntax matter? A strong baseline for aspect-based sentiment analysis with RoBERTa[C]//Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: human Language Technologies. Stroudsburg: Association for Computational Linguistics, 2021: 1816-1829.

[9] LIU Y H, OTT M, GOYAL N, et al. RoBERTa: a robustly optimized BERT pretraining approach[EB/OL]. (2019-07-26)[2023-05-20]. https://arxiv.org/abs/1907.11692v1.

[10] WU C, XIONG Q Y, GAO M, et al. A relative position attention network for aspect-based sentiment analysis[J]. Knowledge and Information Systems, 2021, 63(2): 333-347.

[11] WU T, PENG J J, ZHANG W Q, et al. Video sentiment analysis with bimodal information-augmented multi-head attention[J]. Knowledge-Based Systems, 2022, 235: 107676.

[12] HAN W, CHEN H, GELBUKH A, et al. Bi-bimodal modality fusion for correlation-controlled multimodal sentiment analysis[C]//Proceedings of the International Conference on Multimodal Interaction. New York: ACM, 2021.

[13] 王靖豪, 刘箴, 刘婷婷, 等. 基于多层次特征融合注意力网络的多模态情感分析[J]. 中文信息学报, 2022, 36(10): 145-154.
WANG J H, LIU Z, LIU T T, et al. Multimodal sentiment analysis based on multilevel feature fusion attention network[J]. Journal of Chinese Information Processing, 2022, 36(10): 145-154(in Chinese).

[14] WEN H L, YOU S D, FU Y. Cross-modal context-gated convolution for multi-modal sentiment analysis[J]. Pattern Recognition Letters, 2021, 146: 252-259.

[15] KIELA D, BHOOSHAN S, FIROOZ H, et al. Supervised multimodal bitransformers for classifying images and text[EB/OL]. (2020-11-12)[2023-05-20]. https://arxiv.org/abs/1909.02950.

[16] KHAN Z, FU Y. Exploiting BERT for multimodal target sentiment classification through input space translation[C]//Proceedings of the 29th ACM International Conference on Multimedia. New York: ACM, 2021.

[17] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[EB/OL]. (2017-06-12)[2023-05-20]. https://arxiv.org/abs/1706.03762.

[18] TSAI Y H, BAI S J, LIANG P P, et al. Multimodal transformer for unaligned multimodal language sequences[EB/OL]. (2019-06-01)[2023-05-20]. https://arxiv.org/abs/1906.00295.

[19] CHO K, VAN MERRIENBOER B, GULCEHRE C, et al. Learning phrase representations using RNN encoder-decoder for statistical machinetranslation[EB/OL].(2014-06-03)[2023-05-20].https://arxiv.org/abs/1406.1078v3.

[20] DENG J, DONG W, SOCHER R, et al. ImageNet: a large-scale hierarchical image database[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2009: 248-255.

[21] HE K M, ZHANG X Y, REN S Q, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2016: 770-778.

[22] BENGIO Y, SCHWENK H, SENÉCAL J S, et al. Neural probabilistic language models[M]. Berlin: Springer, 2006: 137-186.

[23] BA J L, KIROS J R, HINTON G E. Layer normalization[EB/OL]. (2016-07-21)[2023-05-22]. https://arxiv.org/abs/1607.06450v1.

[24] 曹梦丽. 基于辅助信息抽取与融合的社交媒体图文情感分析方法研究[D]. 武汉: 武汉科技大学, 2022.
CAO M L. Research on sentiment analysis method of social media graphics and text based on auxiliary information extraction and fusion[D]. Wuhan: Wuhan University of Science and Technology, 2022(in Chinese).

[25] ZHOU P, SHI W, TIAN J, et al. Attention-based bidirectional long short-term memory networks for relation classification[C]//Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: Association for Computational Linguistics, 2016.

[26] YU J F, JIANG J. Adapting BERT for target-oriented multimodal sentiment classification[C]//Proceedings of the 28h International Joint Conference on Artificial Intelligence. Vienna : IJCAI, 2019: 5408-5414.

[27] YU B H, WEI J X, YU B, et al. Feature-guided multimodal sentiment analysis towards industry 4.0[J]. Computers and Electrical Engineering, 2022, 100: 107961.

Analysis of image and text sentiment method based on joint and interactive attention

HU Huijun[1, 2], DING Ziyi[1, 2], ZHANG Yaofeng[3, *], LIU Maofu[1, 2]

(1. School of Computer Science and Technology, Wuhan University of Science and Technology, Wuhan 430065, China;
2. Hubei Key Laboratory of Intelligent Information Processing and Real-time Industrial System, Wuhan University of Science and Technology, Wuhan 430065, China;
3. Hubei Center for Data and Analysis, Hubei University of Economics, Wuhan 430205, China)

Abstract: The image and text sentiment in social media is an important factor affecting public opinion and is receiving increasing attention in the field of natural language processing (NLP). Currently, the analysis of image and text sentiment in social media has mainly focused on single image and text pairs, while little attention has been given to image and text pairs of atlas that are non-chronological and diverse. To explore the sentiment consistency between images and texts in the atlas, a method for analyzing image and text sentiment in social media based on joint and interactive attention (SA-JIA) was proposed. The method used RoBERTa and bidirectional gated recurrent unit (Bi-GRU) to extract textual expression features and ResNet50 to obtain image visual features. Joint attention was employed to identify salient regions where image and text sentiment align, obtaining new textual and image visual features. Interactive attention was utilized to focus on inter-modal feature interactions and multimodal feature fusion, finally obtaining the sentiment categories. Experimental validation was conducted on the IsTS-CN dataset and the CCIR20-YQ dataset, showing that the proposed method can enhance the performance of analyzing image and text sentiment in social media.

Keywords: social media; image and text sentiment analysis; joint attention; interactive attention; multimodal fusion

Received: 2023-06-15; **Accepted**: 2023-10-12; **Published Online**: 2023-11-23 15:37

URL: link.cnki.net/urlid/11.2625.V.20231122.1324.001

Foundation items: National Natural Science Foundation of China (62271359); Key Projects of the National Social Science Foundation of China (23ATJ005); The "14th Five Year Plan" Hubei Province Advantage Characteristic Discipline (Group) Project (2023D0302); Key Scientific Research Project of Hubei Provincial Department of Education (20192202)

* **Corresponding author**. E-mail: yfzhang@hbue.edu.cn

http://bhxb.buaa.edu.cn jbuaa@buaa.edu.cn
DOI: 10.13700/j.bh.1001-5965.2023.0382

基于语义信息引导的多标签图像分类

黄俊[1, 2, *]，范浩东[1]，洪旭东[1]，李雪[1]

(1. 安徽工业大学 计算机科学与技术学院，马鞍山 243032；2. 合肥综合性国家科学中心人工智能研究院，合肥 230088)

摘　　要：多标签图像分类旨在为给定的输入图像预测一组标签，基于语义信息的研究主要利用语义和视觉空间的相关性指导特征提取过程生成有效的特征表示，或利用语义和标签空间的相关性学习能够捕获标签相关性的加权分类器，未能同时建模语义、视觉和标签空间相关性。针对该问题，提出一种基于语义信息引导的多标签图像分类(SIG-MLIC)方法，SIG-MLIC方法可以同时利用语义、视觉和标签空间，通过语义引导的注意力（SGA）机制增强标签与图像区域的关联性而生成语义特定的特征表示，同时利用标签的语义信息生成一个具有标签相关性约束的语义字典对视觉特征进行重建，获得归一化的表示系数作为标签出现的概率。在3个标准的多标签图像分类数据集上的实验结果表明：SIG-MLIC方法中的注意力机制和字典学习可以有效提高分类性能，验证了所提方法的有效性。

关 键 词：多标签图像分类；语义空间；视觉空间；标签空间；注意力机制；字典学习
中图分类号：TP37; TP183; TP181
文献标志码：A　　**文章编号**：1001-5965(2025)07-2271-11

多标签图像分类任务旨在识别一张图像中的多个对象或属性，与单标签图像分类任务相比更具有现实意义和实用价值。目前，多标签图像分类已被广泛应用于医疗诊断识别[1]、生物信息学[2]和自动图像标注[3]等领域。然而随着类别标签数量的增多，输出空间的组合方式也将以指数级增长，因此，多标签图像分类的研究也更具挑战性。

随着深度卷积神经网络(convolutional neural networks, CNN)的快速发展，骨干模型也逐渐完善，因此，在模型学习过程中引入附加信息成为研究热点[4]。而标签的语义信息是一种非常有价值的附加信息，基于标签语义信息的研究在多标签图像分类领域中受到广泛关注。一方面，利用语义空间和视觉空间之间的相关性，标签的语义信息可以指导特征提取过程生成有效的特征表示。例如，Wang等提出的CNN-循环神经网络(recurrent neural networks, RNN)[5]框架学习联合图像和标签的嵌入，以表征语义标签依赖性和图像与标签的相关性；Zhou等提出一种类别原型合成学习(category prototype compositional learning, CPCL)方法[6]，通过组合和分解与类别相关的特征及其对应的类别原型，从而提高了分类的准确性；Chen等提出的语义特定图表示学习(semantic-specific graph representation learning, SSGRL)[7]利用语义引导的注意力(semantic guided attention, SGA)机制，结合标签语义引导语义特定的视觉特征的学习。这些工作大都只将语义信息作为视觉特征的补充，且没有考虑到对标签空间的开发利用。

另一方面，利用语义空间和标签空间之间的相关性，通过图卷积神经网络结合标签的语义信息学

收稿日期：2023-06-16；**录用日期**：2023-09-15；**网络出版时间**：2023-11-08 11:50
网络出版地址：link.cnki.net/urlid/11.2625.V.20231107.1105.005
基金项目：国家自然科学基金(61806005)；安徽高校协同创新项目(GXXT-2022-052)；安徽省高校优秀青年人才支持计划项目(gxyqZD2022032)；安徽省高校科学研究重点项目(KJ2021A0373)；多模态认知计算安徽省重点实验室（安徽大学）开放基金(MMC202008)；安徽省自然科学基金(2008085QF305)

* **通信作者**. E-mail：huangjun.cs@ahut.edu.cn

引用格式：黄俊，范浩东，洪旭东，等. 基于语义信息引导的多标签图像分类[J]. 北京航空航天大学学报，2025，51（7）：2271-2281.
HUANG J，FAN H D，HONG X D，et al. Semantic information-guided multi-label image classification[J]. Journal of Beijing University of Aeronautics and Astronautics，2025，51（7）：2271-2281（in Chinese）.

习标签相关性也受到研究者们的广泛关注。Chen等提出的基于图卷积网络的多标签分类模型(multi-label classification model based on graph convolutional networks, ML-GCN)[8]方法在学习标签分类器的同时建模了标签相关性;Li等提出的自适应图卷积网络(adaptive graph convolutional networks, A-GCN)[9]是一个即插即用的标签图模块,即在标签嵌入上自动构建标签相关矩阵,从而使得标签嵌入具有自适应性;Wang等提出的快速图卷积网络(fast graph convolution networks, F-GCN)[10]方法将标签嵌入和视觉特征作为整体进行融合,有效提高了模型收敛速度;Chen等提出的胸透X光图卷积网络(chest X-ray graph convolution networks, CheXGCN)[11]方法将标签的语义信息应用于医疗图像诊断领域,实验验证了该方法的有效性。这些工作大都只将语义信息作为指导判别分类器推导的辅助信息,但没有足够关注对视觉空间的开发利用。

此外,字典学习作为一种利用视觉和标签空间相关性的有效工具已被应用在多标签图像分类领域。例如,Niu等提出的多标签耦合字典学习(mutli-label coupled dictionary learning, ML-CDL)[12]将视觉空间和标签空间视为耦合域,提出一种多标签耦合字典学习算法来解决标签空间降维问题;Zheng等提出的四元组字典学习(quadruplet dictionary learning, QDL)模型[13]用于多类别的多标签分类,其视觉空间和标签空间由2个字典对互连,充分开发2个空间之间的相关性;Jing等[14]提出一种具有标签一致性正则化和部分相同标签嵌入的多标签字典学习,以充分利用视觉和标签空间之间的关系。然而,这些方法并没有引入标签的语义信息,语义和视觉空间之间的相关性没有得到开发利用。

综上,在多标签图像分类领域,利用到标签语义信息的方法大都将其作为开发视觉空间或标签空间的辅助信息,且很少有同时开发语义、视觉和标签空间三者之间相关性的研究。标签的语义信息仍然具有有待开发利用的潜在价值。

为探究上述问题,本文提出基于语义信息引导的多标签图像分类(semantic information guided multi-label image classification, SIG-MLIC)方法。SIG-MLIC方法将多标签图像分类视为一项字典学习任务,利用具有标签相关性约束的语义字典对语义特定的视觉特征进行重建,同时协调了语义、视觉和标签空间三者之间的相关性,提高了标签预测的可靠性。

具体地说,首先,SIG-MLIC方法探索了标签空间和视觉空间之间的相关性,其利用由标签的语义信息直接生成的语义字典作为分类器对视觉特征进行重建,获得归一化的表示系数作为标签预测的概率;其次,SIG-MLIC方法利用了语义空间和视觉空间之间的相关性,其利用标签的语义信息引导注意力机制指导特征提取过程,使得视觉特征更加关注与语义相关的区域,减少了无关信息的干扰,这更有利于语义字典对视觉特征的重建过程。此外,SIG-MLIC方法对语义空间进行了标签相关性约束,使得语义字典能够更好地反映标签之间的关系,增强语义信息的表达能力。本文的创新点可以总结如下:

1)提出SIG-MLIC方法。SIG-MLIC方法同时将标签的语义信息用于图像的特征提取过程和分类器学习过程,充分利用和协调语义、视觉和标签空间。

2)SIG-MLIC方法采用将注意力机制和字典学习相结合的方式,利用具有标签相关性约束的语义字典重建语义特定的视觉特征,标签相关性约束使得字典的表达能力更强,语义特定的视觉特征减少了字典对视觉特征重建过程中无关信息的干扰,两方面共同作用提升模型性能。

3)通过在Pascal VOC 2007[15]、Pascal VOC 2012[15]和Microsoft COCO[16]这3个多标签数据集上对SIG-MLIC方法进行评估,实验结果均证实本文方法的有效性。

1 相关工作

1.1 多标签图像分类

随着多标签图像分类任务的发展,解决相关问题的方法和模型框架相继被提出,其主要思想可以分为以下3类。

1)建模标签相关性。由于在同一张图像中通常出现多个对象,因此,需要对标签相关性进行建模以提高分类性能。Zhu等提出的全局与局部标签关联(global and local label correlation, GLOCAL)的多标签学习[17]方法通过流行正则化学习全局和局部标签相关性;Weng等提出的基于全局与局部标签关联的标签特定特征学习(label-specific features learning with global and local label correlation, LFGLC)[18]方法集成了全局和局部标签相关性学习标签特定特征。Qu等提出的多层语义表征网络(multi-layered semantic representation networks, MSRN)[19]模型通过建模标签相关性来发现标签的局部语义和全局语义。

2)定位感兴趣区域。多标签图像通常在不同位置具有多个对象,因此,定位到所关注的区域可

以提高分类性能。Chen 等提出的循环注意力强化学习(recurrent attentional reinforcement learning, RARL)[20]模型通过引入强化学习定位注意区域, 并基于生成的区域顺序预测标签分布; Gao 等提出的多类别注意力区域(multi-class attentional region, MCAR)[21]模型具备简单有效的双流框架, 可以识别从全局图像到局部区域的多个类别对象; Wang 等提出的跨模态融合的多标签图像分类(cross-modal fusion for multi-label image classification, CFMIC)[22]模型通过具有注意力机制的特征提取模块, 帮助生成每个输入图像的准确特征。

3) 改进损失函数。多标签的 one-vs-all 策略造成了多标签图像分类任务中正负样本不平衡的问题。为解决此问题, Wu 等[23]针对呈现长尾分布的多标签识别问题提出一种分布平衡损失; Ridnic 等提出的非对称损失函数(asymmetric loss function, ASL)[24]用于解决多标签分类的不对称损失问题, 其包含 2 种互补的不对称机制, 对正样本和负样本的作用不同; Dong[25]提出基于焦点损失的方法来解决类别不平衡问题, 从而提高算法的性能。

1.2　字典学习

字典学习是稀疏表示理论中最具代表性的研究, 其可以通过稀疏性表示对图像数据进行分类。先前的一些研究中, 字典学习已经被应用在图像注释、图像分类和人脸识别等领域。具体地说, Cao 等提出的语义标签嵌入字典(semantic label embedding dictionary, SLED)[26]模型将标签信息明确地融合到字典表示中, 并探索共现标签的语义相关性; Zhao 等[27]提出一种基于多层字典学习的图像分类新方法, 通过增强编码系数特征和筛选有益的分类特征来提高分类精度; Ou 等[28]通过字典学习的帮助, 系统地研究了一种有效的人脸识别差分隐私方法以解决人脸识别的隐私问题。

最近, Zhou 等提出的深度语义字典学习(deep semantic dictionary learning, DSDL)[29]模型较早将字典学习引入多标签图像分类领域, 利用由自动编码器从标签语义中生成的语义字典作为分类器对视觉特征进行重建以获得标签预测概率。SIG-MLIC 方法与其主要区别在于不仅将标签语义用于生成分类器, 同时在特征提取过程中利用语义引导的注意力机制使得学习到的视觉特征更加关注与标签相关的区域, 进而在对视觉特征进行重建时减少了无关信息的干扰。此外, 在语义空间中引入标签相关性约束, 使得语义字典能够更好地反映标签之间的关系, 增强了语义信息的表达能力。

1.3　语义信息

随着自然语言处理技术的发展, 语义信息在多个领域已被证明具有引入模型学习过程的价值。例如, Huang 等提出的类原型判别网络(class-prototype discriminative networks, CPDN)模型[30]用于零样本学习, 该模型通过每类语义表示生成视觉原型, 在度量网络中测量样本和视觉原型的相似性以完成分类; Xing 等[31]针对少镜头学习提出一种机制, 根据要学习的图像类别自适应地组合来自视觉信息和语义表示 2 种模态的信息; He 等提出视觉和语言相结合(combining vision and language, CVL)[32]的双流模型用于细粒度图像分类, 其中, 语言流基于自然语言描述指出每个图像中的判别性信息。

2　SIG-MLIC 框架

2.1　本文方法概述

本文提出 SIG-MLIC 方法, 其由 3 个关键模块组成, 即字典学习模块、特征学习模块和视觉特征表示模块。字典学习模块利用标签词嵌入和标签相关性损失学习具有标签相关性约束的语义字典。首先, 特征学习模块通过 CNN 提取输入图像的特征图; 然后, 将标签词嵌入和特征图输入基于语义引导的注意力机制模块中, 以获得语义特定的特征表示。视觉特征表示模块用语义字典对视觉特征进行重建, 获得具有归一化属性的表示系数, 并将其作为标签出现的概率。SIG-MLIC 方法总体框架如图 1 所示, 其中, FC 表示全连接层, $\boldsymbol{x}_i$ 为词嵌入, $\boldsymbol{D}$ 为语义字典, F_{linear} 为全连接层函数, $\boldsymbol{a}_i$ 为字典原子, $\boldsymbol{f}^k$ 为特征图, $\boldsymbol{F}$ 为特征向量, $\boldsymbol{F}=[\boldsymbol{f}^1, \boldsymbol{f}^2, \cdots, \boldsymbol{f}^c]$, $\boldsymbol{f}$ 为全局特征, $\boldsymbol{\gamma}$ 为字典原子的系数向量; m 为语义嵌入的维数; d 为语义字典的维度。

2.2　字典学习模块

字典学习在各个领域已经被证明是一种利用标签空间和视觉空间之间相关性的有效工具[29]。本文利用标签的语义信息生成与视觉空间对齐的语义字典, 使得 CNN 提取的视觉特征可以用语义字典来重建, 从而将多标签图像分类问题转换为字典学习任务来解决。

字典学习模块先采用在维基百科上训练的 Glove 模型[33]$\boldsymbol{f}_{\text{g}}$ 获得每个类别标签的 m 维语义嵌入向量:

$$\boldsymbol{x}_i = \boldsymbol{f}_{\text{g}}(w_i) \tag{1}$$

式中: w_i 为第 i 个标签的语义词; $\boldsymbol{x}_i$ 为与之对应的词嵌入。由此获得类别标签的词嵌入 $\boldsymbol{X}=[\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_c] \in \mathbf{R}^{m\times c}$, 其中, m=300, c 为类别数。其次, 将 $\boldsymbol{X}$ 投影到视觉空间中以生成语义字典 $\boldsymbol{D}=[\boldsymbol{a}_1, \boldsymbol{a}_2, \cdots, \boldsymbol{a}_c] \in \mathbf{R}^{d\times c}$, 其中, $\boldsymbol{a}_i \in \mathbf{R}^d$ 为第 i 个字典原子, d=2 048, $\boldsymbol{D}$ 的表达式为

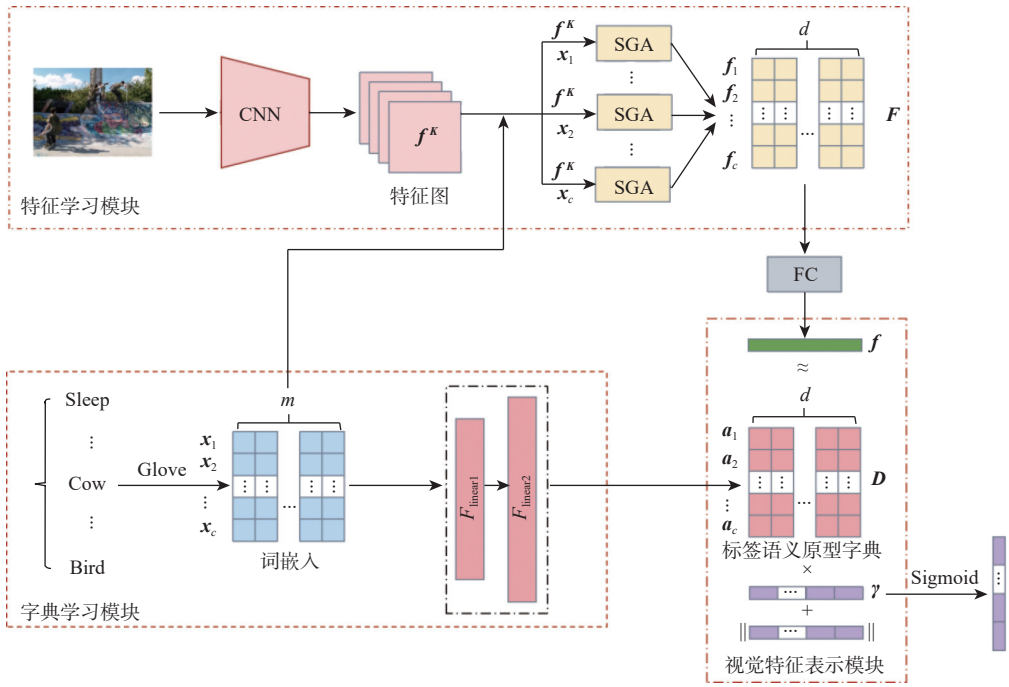


图1 基于语义信息引导的多标签图像分类框架

Fig. 1 SIG-MLIC framework

$$\boldsymbol{D}=F_{\text{linear2}}(F_{\text{linear1}}(\boldsymbol{X};\boldsymbol{\varphi}_1);\boldsymbol{\varphi}_2) \tag{2}$$

式中：$F_{\text{linear1}}(\cdot)$和$F_{\text{linear2}}(\cdot)$由2个全连接层实现，$\boldsymbol{\varphi}_1$和$\boldsymbol{\varphi}_2$为可学习参数矩阵。

利用标签相关性可以提高多标签分类器的性能。本文通过训练数据集直接得到标签矩阵$\boldsymbol{Y}\in\{0,1\}^{n\times c}$，其中，$n$为训练样本总数。通过$\boldsymbol{Y}$计算任意标签对之间的余弦相似度获得相关矩阵$\boldsymbol{P}\in\mathbf{R}^{c\times c}$，$P_{ij}$表示第$i$个标签和第$j$个标签的相关性。进一步通过计算具有标签相关性约束的任意一对字典原子之间的欧几里得距离，即$P_{ij}\|\boldsymbol{a}_i-\boldsymbol{a}_j\|_2^2$，来合并每对字典原子之间的成对标签相关性。如果第$i$个标签和第$j$个标签之间具有很强的相关性，那么其应当具有相似的字典原子。相应的，经过字典学习获得的标签预测概率应该更接近。由此，将标签相关性损失表示为

$$L_{\text{cor}}=\text{tr}(\boldsymbol{D}^{\text{T}}\boldsymbol{S}\boldsymbol{D}) \tag{3}$$

式中：函数tr(·)表示求矩阵的迹，$\boldsymbol{S}\in\mathbf{R}^{c\times c}$为相关矩阵$\boldsymbol{P}$的图拉普拉斯矩阵。$L_{\text{cor}}$在语义字典$\boldsymbol{D}$中的字典原子之间充分考虑了标签相关性。

2.3 特征学习模块

在特征学习模块，受SSGRL[7]工作启发，本文采用语义引导的注意力机制，利用标签的语义信息引导学习语义特定的特征表示。

给定输入图像$\boldsymbol{K}$，先采用在ImageNet上预训练的ResNet-101作为特征学习的主干网络f_{cnn}提取$\boldsymbol{K}$的特征图$\boldsymbol{f}^{\boldsymbol{K}}$：

$$\boldsymbol{f}^{\boldsymbol{K}}=f_{\text{cnn}}(\boldsymbol{K};\boldsymbol{\theta})\qquad \boldsymbol{f}^{\boldsymbol{K}}\in\mathbf{R}^{W\times H\times N} \tag{4}$$

式中：$\boldsymbol{\theta}$为CNN的可学习的参数张量；W、H和N分别为特征图的宽度、高度和通道数。

语义引导的注意力结合了在字典学习模块中生成的每个类别标签的词嵌入向量$\boldsymbol{x}_i$，以引导网络更多地关注语义感知区域，进而学习与该类别相对应的特征向量。SGA的具体实现过程如图2所示。

首先，对于特征图$\boldsymbol{f}^{\boldsymbol{K}}$上的每个位置$(w,h)$，使用低秩双线性池方法[34]，融合图像特征$\boldsymbol{f}^{\boldsymbol{K}}_{(w,h)}$和词嵌入向量$\boldsymbol{x}_i$这2种模态：

$$\boldsymbol{f}^{\boldsymbol{K}}_{i,(w,h)}=\boldsymbol{Q}^{\text{T}}(\tanh((\boldsymbol{U}^{\text{T}}\boldsymbol{f}^{\boldsymbol{K}}_{(w,h)})\odot(\boldsymbol{V}^{\text{T}}\boldsymbol{x}_i)))+\boldsymbol{b} \tag{5}$$

式中：tanh(·)为双曲正切函数；$\boldsymbol{Q}\in\mathbf{R}^{d_1\times d_2}$、$\boldsymbol{U}\in\mathbf{R}^{N\times d_1}$、$\boldsymbol{V}\in\mathbf{R}^{m\times d_1}$和$\boldsymbol{b}\in\mathbf{R}^{d_2}$都是可学习参数矩阵，$d_1$为联合嵌入的维度，$d_2$为输出特征的维度；⊙表示Hadamard乘积。

其次，在$\boldsymbol{x}_i$的指导下计算表示位置(w,h)重要性的注意力系数，并使用SoftMax函数进行归一化：

$$\boldsymbol{a}_{i,(w,h)}=\text{SoftMax}(f_{\text{a}}(\boldsymbol{f}^{\boldsymbol{K}}_{i,(w,h)})) \tag{6}$$

式中：$f_{\text{a}}(\cdot)$为由全连接网络实现的一个注意力函数。对特征图上的所有位置执行上述过程。并采用加权平均池化获得特征向量：

$$\boldsymbol{f}_i=\sum_{(w,h)}\boldsymbol{a}_{i,(w,h)}\boldsymbol{f}^{\boldsymbol{K}}_{i,(w,h)} \tag{7}$$

对所有类别重复式(6)和式(7)，以获得与c个类别分别相关的特征向量$\boldsymbol{F}=(\boldsymbol{f}_1,\boldsymbol{f}_2,\cdots,\boldsymbol{f}_c)$。最后，SIG-MLIC使用一个全连接层(fully connected layer,

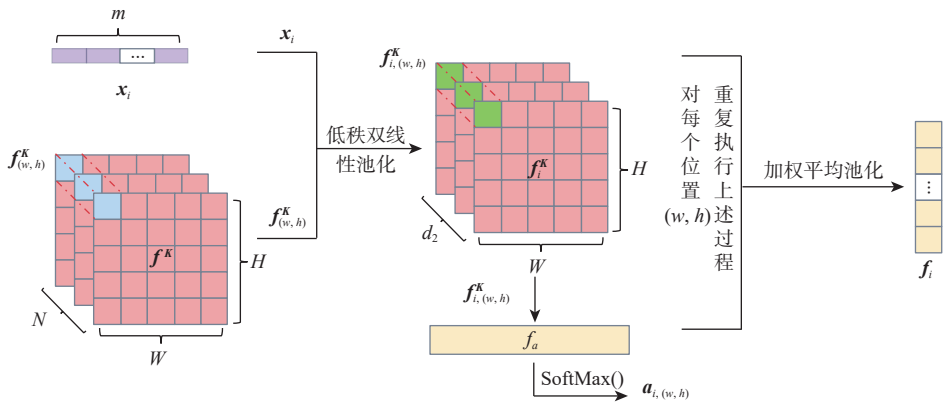


图 2　语义引导的注意力模块示意图

Fig. 2　Schematic diagram of SGA module

FC) 对语义特定的特征向量进行融合, 得到全局特征 $\boldsymbol{f}\in\mathbf{R}^d$。

2.4　视觉特征表示模块

在视觉特征表示模块中, 通过生成的语义字典 $\boldsymbol{D}$ 将视觉特征 $\boldsymbol{f}$ 的重建作为字典查询任务。字典学习损失可以表示为

$$L_{\text{dic}}=\left\|\boldsymbol{f}-\boldsymbol{D}\boldsymbol{\gamma}\right\|_2^2+\beta\left\|\boldsymbol{\gamma}\right\|_2^2 \tag{8}$$

式中: $\boldsymbol{\gamma}=[\gamma_1,\gamma_2,\cdots,\gamma_c]\in\mathbf{R}^c$, γ_i 为对应于 $\boldsymbol{D}$ 中字典原子 $\boldsymbol{a}_i$ 的表示系数; $\|\boldsymbol{f}-\boldsymbol{D}\boldsymbol{\gamma}\|_2^2$ 为视觉特征的重建损失; $\|\boldsymbol{\gamma}\|_2^2$ 为 $\boldsymbol{\gamma}$ 的 L2 范数正则化, 用以避免模型出现过拟合; β 为用来平衡 2 种损失的超参数。利用 sigmoid 函数对 $\boldsymbol{\gamma}$ 进行归一化处理:

$$\hat{\boldsymbol{y}}=\text{sigmoid}(\boldsymbol{\gamma})=\left[\hat{\boldsymbol{y}}_1,\hat{\boldsymbol{y}}_2,\cdots,\hat{\boldsymbol{y}}_c\right]\in\mathbf{R}^c \tag{9}$$

将 $\hat{\boldsymbol{y}}$ 视为视觉特征 $\boldsymbol{f}$ 的标签出现概率, 关于 $\boldsymbol{\gamma}$ 的求解详见 2.5 节。真实标签分布 $\boldsymbol{y}=[\boldsymbol{y}_1,\boldsymbol{y}_2,\cdots,\boldsymbol{y}_c]\in\mathbf{R}^c$ 和标签的预测概率 $\hat{\boldsymbol{y}}$ 之间的差异用交叉熵损失进行衡量:

$$L_{\text{ce}}=-\sum_{j=0}^{n-1}\sum_{i=0}^{c-1}y_i^j\ln(\hat{y}_i^j)+(1-y_i^j)\ln(1-\hat{y}_i^j) \tag{10}$$

2.5　模型优化

综合式 (3)、式 (8) 和式 (10), SIG-MLIC 总体优化目标为

$$\arg\min_{\boldsymbol{\theta},\boldsymbol{\varphi}}L_{\text{total}}=L_{\text{ce}}+\delta L_{\text{cor}}+\alpha L_{\text{dic}} \tag{11}$$

式中: δ 和 α 为用来平衡损失的超参数。由于语义字典 $\boldsymbol{D}$ 和表示系数向量 $\boldsymbol{\gamma}$ 在训练阶段是可学习的, 本文借鉴 DSDL[29] 提出的交替参数更新策略进行优化求解。首先, 在正向传播过程中, 固定语义字典 $\boldsymbol{D}$, 将式 (8) 的优化问题转换为正则化最小二乘问题。令字典学习损失 L_{dic} 关于 $\boldsymbol{\gamma}$ 的导数为 $\mathbf{0}$, 则:

$$\boldsymbol{\gamma}=(\boldsymbol{D}^{\mathrm{T}}\boldsymbol{D}+\beta\boldsymbol{I})^{-1}\boldsymbol{D}^{\mathrm{T}}\boldsymbol{f} \tag{12}$$

由此得到 $\boldsymbol{\gamma}$ 的闭式解; 其次, 在反向传播过程中, 固定表示系数 $\boldsymbol{\gamma}$, 通过反向传播优化 $\boldsymbol{\theta}$ 和 $\boldsymbol{\varphi}$, 从而更新语义字典 $\boldsymbol{D}$ 和特征学习网络; 最后, 重复执行上述更新步骤直至模型收敛。

3　实　验

3.1　实验参数与评价指标

在特征学习模块, 输入图像被随机裁剪并调整像素大小为 448×448, 采用在 ImageNet 上预训练的 ResNet-101 作为主干网络提取特征图 $\boldsymbol{f}^K$, $\boldsymbol{f}^K$ 的宽度 W 为 14, 高度 H 为 14, 通道数 N 为 2 048。对于 SGA 模块中的低秩双线性池操作, d_1 和 d_2 都设置为 1 024, $f_{\text{a}}(\cdot)$ 由一个 1 024 对 1 的全连接层实现。用于组合类别相关特征的全连接网络 FC 由类别数 c 对 1 的全连接层实现。在字典学习模块, $F_{\text{linear1}}(\cdot)$ 和 $F_{\text{linear2}}(\cdot)$ 分别由 300 对 1 024 和 1 024 对 2 048 的全连接层实现, 后跟负斜率为 0.2 的 LeakyReLu 函数。模型采用 SGD 优化器进行优化, 动量设置为 0.9, 权重衰减为 10^{-4}。初始学习率设置为 0.01, 且学习率在 100 回合中每 40 回合衰减 0.1。本实验在一台 NVIDIA_GeForce_RTX_3 090 的 GPU 上训练, 并将批量大小设为 8。

模型评价指标包括每个类别的平均精度 (average precision, AP) 和所有类别的平均精度 (mean average precision, mAP)。对于每个图像, 如果标签置信度大于 0.5, 则将其预测为正。对于 MS-COCO 数据集, 使用每个类别的精度 (class-wise precision, CP)、每个类别召回率 (class-wise recall, CR)、每个类别 F_1 分数 (class-wise F_1, CF1) 和所有类别平均精度 (overall precision, OP)、所有类别平均召回率 (overall recall, OR)、所有类别平均 F_1 分数 (overall F_1, OF1) 作为详细的性能指标, 为公平比较, 实验给每个图像分配得分前 3 的标签, 并将其与真实标签进行比较。一般地, OF1、CF1 和 mAP 对于性能评估更重要。

3.2 对比方法

本文选取的对比方法根据是否引入标签的语义信息可以总结如下：

1）利用标签的语义信息引导注意力机制的方法。CPCL[6]、SSGRL[7]和MSRN[19]采用了类似的注意力机制，学习与标签语义相关的区域。CPCL强调通过合成和分解操作在特征空间对合成性质进行建模；SSGRL强调探索特征向量之间的语义交互；MSRN强调在CNN的多层学习图像的语义表示。

2）利用标签的语义信息训练分类器的方法。ML-GCN[8]在学习标签分类器的同时，强调对标签相关性的建模；DSDL[29]是较早将字典学习引入多标签图像分类领域的方法。

3）未利用语义信息的方法。Yang等[35]利用边界框注释进行多标签对象识别；Wang等[36]提出一种随机裁剪池方法，让网络在训练过程中发现图像的判别区域；He等[37]将一种强化学习算法用于多标签图像分类；MCAR[21]旨在仅使用图像自身的语义信息而不依赖标签信息提高模型性能。

3.3 实验结果

3.3.1 在Pascal VOC 2007数据集上的结果

Pascal VOC 2007[15]数据集包含4个主要类别：车辆、家庭、动物和人，其又被细分为20个类别。该数据集分为训练集(2 501张)、验证集(2 510张)和测试集(4 952张)，遵循先前的工作，本文实验在训练集和验证集上训练模型，在测试集上评估模型。

基于VOC2007数据集进行实验，将SIG-MLIC方法与其他的标签图像分类方法进行对比，实验结果如表1所示。SIG-MLIC方法获得了95.5%的mAP评分，相比于利用标签的语义信息引导注意力机制的CPCL、SSGRL和MSRN方法，mAP分别提高了1.1%、2.1%和0.6%；相比于利用标签的语义信息指导生成判别分类器的ML-GCN的mAP值，其性能增益为1.5%；较早将字典学习的思想引入到多标签图像分类中的DSDL方法获得了94.4%的mAP值，SIG-MLIC方法的mAP比其高出1.1%。这种性能增益得益于SIG-MLIC方法通过对语义信息的充分开发和利用，同时协调了语义、视觉和标签空间的相关性。由于MS-COCO数据集包含了VOC 2007数据集中的所有种类，本文通过先将模型在MS-COCO数据集上进行预训练，训练后的模型简称为SIG-MLIC(pre)，再迁移到VOC 2007数据集上进行训练的方式进一步提升模型性能。具体地，SIG-MLIC(pre)在VOC 2007数据集上获得了最高为96.5%的mAP评分，并在该数据集中的15个类别中获得了第1的AP评分，4个类别中获得了第2的AP评分。

表1 在Pascal VOC 2007数据集上不同方法的比较

Table 1 Comparison with different methods on Pascal VOC 2007 dataset %

方法	AP																				mAP
	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	motor	person	plant	sheep	sofa	train	TV	
CPCL[6]	99.6	98.6	98.5	98.8	81.9	95.1	97.8	98.2	83.0	95.5	85.5	98.4	98.5	97.0	99.0	86.6	97.0	84.9	99.1	94.3	94.4
SSGRL[7]*	99.5	97.1	97.6	97.8	82.6	94.8	96.7	98.1	78.0	97.0	85.6	97.8	98.3	96.4	98.8	84.9	96.5	79.8	98.4	92.8	93.4
SSGRL(pre)[7]*	99.7	98.4	98.0	97.6	85.7	96.2	98.2	98.8	82.0	98.1	89.7	98.8	98.7	97.0	99.0	86.9	98.1	85.8	99.0	93.7	95.0
MSRN[19]	**100.0**	98.8	98.9	**99.1**	81.6	95.5	98.0	98.2	84.4	96.6	87.5	98.6	98.6	97.2	99.1	87.0	97.6	86.5	99.4	94.4	94.9
MSRN(pre)[19]	99.7	98.9	98.7	**99.1**	86.6	97.9	98.5	98.9	**86.0**	98.7	89.1	99.0	99.1	97.3	99.2	**90.2**	99.2	**89.7**	**99.8**	95.3	96.0
ML-GCN[8]	99.5	98.5	98.6	98.1	80.8	94.6	97.2	98.2	82.3	95.7	86.4	98.2	98.4	96.7	99.0	84.7	96.7	84.3	98.9	93.7	94.0
DSDL[29]	99.8	98.7	98.4	97.9	81.9	95.4	97.6	98.3	83.3	95.0	88.6	98.0	97.9	95.8	99.0	86.6	95.9	86.4	98.6	94.4	94.4
MCAR[21]	99.7	99.0	98.5	98.2	85.4	96.9	97.4	98.9	83.7	95.0	88.8	99.1	98.2	95.1	99.1	84.8	97.1	87.8	98.3	94.8	94.8
SIG-MLIC	99.9	98.6	98.9	98.3	83.6	**98.2**	97.9	99.1	82.7	97.5	89.6	99.3	99.1	98.8	99.1	86.9	99.5	86.6	99.4	**97.2**	95.5
SIG-MLIC(pre)	**100.0**	**99.1**	**99.4**	98.7	**88.2**	**98.2**	**98.8**	**99.6**	85.0	**98.8**	**92.0**	**99.7**	**99.5**	**99.3**	**99.3**	88.8	**99.7**	89.4	99.6	**97.2**	**96.5**

注：粗体数字为最优值，下划线数字为次优值，“pre”表示模型在MS-COCO数据集上进行了预训练，*表示输入图片像素大小为576×576。

3.3.2 在VOC 2012上的结果

Pascal VOC 2012[15]包含与VOC 2007相同的20个类别，但其图像与VOC 2007没有交集。该数据集分为训练集(5 717张)、验证集(5 823张)和测试集(10 991张)。由于该数据集没有提供测试集的真实标签，实验在训练集上进行训练，在验证集上进行微调，最后，通过向Pascal VOC评估服务器提交测试结果来评估。

本文SIG-MLIC方法与其他7种多标签图像分类方法在VOC 2012数据集上进行比较，结果如表2所示。SIG-MLIC方法在该数据集上的表现更为突出，具体地，SIG-MLIC方法比对比方法中排名第1的MSRN方法在mAP上提升了0.8%，同时经过预训练后的模型SIG-MLIC(pre)仍比MSRN(pre)

表 2　在 Pascal VOC 2012 数据集上不同方法的比较

Table 2　Comparison with different methods on Pascal VOC 2012 dataset　　%

方法	AP																				mAP
	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	motor	person	plant	sheep	sofa	train	TV	
FeV+LV[35]	98.4	92.8	93.4	90.7	74.9	93.2	90.2	96.1	78.2	89.8	80.6	95.7	96.1	95.3	97.5	73.1	91.2	75.4	97.0	88.2	89.4
RCP[36]	99.3	92.2	97.5	94.9	82.3	94.1	92.4	98.5	83.8	93.5	83.1	98.1	97.3	96.0	98.8	77.7	95.1	79.4	97.7	92.4	92.2
RMIC[37]	98.0	85.5	92.6	88.7	64.0	86.8	82.0	94.9	72.7	83.1	73.4	95.2	91.7	90.8	95.5	58.3	87.6	70.6	93.8	83.0	84.4
SSGRL[7]*	99.5	95.1	97.4	96.4	85.8	94.5	93.7	98.9	86.7	96.3	84.6	98.9	98.6	96.2	98.7	82.2	98.2	84.2	98.1	93.5	93.9
SSGRL(pre)[7]*	99.7	96.1	97.7	96.5	86.9	95.8	95.0	98.9	88.3	97.6	87.4	99.1	99.2	97.3	**99.0**	84.8	98.3	85.8	99.2	94.1	94.8
MSRN[19]	99.7	95.2	98.3	96.3	84.8	96.5	93.3	**99.6**	87.4	96.0	86.3	98.9	98.3	96.9	98.8	80.6	97.7	80.5	99.4	94.7	94.0
MSRN(pre)[19]	99.8	96.3	98.4	96.8	85.2	**97.5**	95.2	**99.6**	88.0	96.6	**89.8**	99.0	98.8	96.8	**99.0**	84.5	97.1	85.8	99.3	95.9	95.0
DSDL[29]	99.4	95.3	97.6	95.7	83.5	94.8	93.9	98.5	85.7	94.5	83.8	98.4	97.7	95.9	98.5	80.6	95.7	82.3	98.2	93.2	93.2
MCAR[21]	99.6	97.1	98.3	96.6	87.0	95.5	94.4	98.8	87.0	96.9	85.0	98.7	98.3	97.3	**99.0**	83.8	96.8	83.7	98.3	93.5	94.3
SIG-MLIC	99.9	96.4	98.4	97.3	86.5	96.5	95.4	99.3	87.9	98.0	85.0	98.9	99.2	97.9	98.5	85.5	98.5	83.7	99.3	95.0	94.8
SIG-MLIC(pre)	**100.0**	**97.8**	**98.8**	**98.1**	**88.9**	**97.5**	**96.0**	99.5	**89.7**	**98.8**	88.1	**99.3**	**99.7**	**98.4**	98.9	**88.4**	**98.7**	**87.0**	**99.6**	**96.7**	**96.0**

注: 粗体数字为每列最优值, 下划线数字为每列次优值, “pre” 表示模型在MS-COCO数据集上进行了预训练, *表示输入图片大小为576×576。

高出 1% 的 mAP 值。SIG-MLIC 方法比同样采用字典学习的 DSDL 方法高出 1.6% 的 mAP 值, 与 DSDL 方法中语义字典不同的是, SIG-MLIC 方法中的语义字典经过标签相关性约束, 语义信息的表达能力得到了增强。SIG-MLIC(pre) 在 17 个类别上取得了最优 AP 评分, 仅在 3 个类别上的 AP 评分以微小的差距低于其他方法。

3.3.3　在 MS-COCO 上的结果

近年来, Microsoft COCO[16] 被广泛应用到多标签图像分类任务当中, 其包含 122 218 张图像, 其中, 82 081 张作为训练集, 40 137 张作为验证集。图像中由 80 个常见类别组成, 平均每张图像有 2.9 个标签。由于测试集的真实标签不可用, 本实验在验证集上评估所有方法的性能。

在 MS-COCO 数据集上的比较结果如表 3 所示。SIG-MLIC 方法在所有比较的方法中获得了最高的 mAP, 与 CPCL、SSGRL 和 MSRN 方法相比分别取得了 2.8%、1.8% 和 2.2% 的性能增益。与采用字典学习的 DSDL 方法相比在 mAP 上取得了 3.9% 的性能增益。这种较为明显的性能提升的原因是, 面对更加庞大和复杂的数据集, SIG-MLIC 方法中对语义字典的标签相关性约束发挥了更为重要的作用, 捕获了复杂标签之间的共现关系。此外, 语义引导的注意力机制使得模型学习到的图像特征包含相对较少的干扰信息。对于其他的评估标准, SIG-MLIC 方法在所有类别上的 CF1、OF1、CR 和 OR 上同样取得了最高的评估得分, 为每个图像分配得分前 3 的标签后, SIG-MLIC 在所有指标上的评估得分均排在第 2 位。

3.4　消融实验

为验证语义引导的注意力机制和字典学习在各模块中发挥的作用, 本文以 4 种方案在 Pascal VOC 2007 数据集上进行消融实验:

1) ResNet-101 表示只使用 ResNet-101 作为主干网络。将特征图 $\boldsymbol{f}^K$ 经过全局最大池化 (global max pooling, GMP) 获得 2 048 维的特征向量, 再用

表 3　在 MS-COCO 数据集上不同方法的比较

Table 3　Comparison with different methods on MS-COCO dataset　　%

方法	mAP	CP		CR		CF1		OP		OR		OF1	
		得分前3标签	所有标签	得分前3标签	所有标签	得分前3标签	所有标签	得分前3标签	所有标签	得分前3标签	所有标签	得分前3标签	所有标签
CPCL[6]	82.8	89.0	85.6	63.5	71.1	74.1	77.6	90.5	86.1	65.9	74.5	76.3	79.9
SSGRL[7]*	83.8	91.9	89.9	62.5	68.5	72.7	76.8	93.8	91.3	64.1	70.8	76.2	79.7
MSRN[19]	83.4	84.5	86.5	72.9	71.5	78.3	78.3	84.3	86.1	76.8	75.5	80.4	80.4
ML-GCN[8]	83.0	89.2	85.1	64.1	72.0	74.6	78.0	90.5	85.8	66.5	75.4	76.7	80.3
DSDL[29]	81.7	88.1	84.1	62.9	70.4	73.4	76.7	89.6	85.1	65.3	73.9	75.6	79.1
MCAR[21]	83.8	88.1	85.0	65.5	72.1	75.1	78.0	91.0	88.0	66.3	73.9	76.7	80.3
SIG-MLIC	85.6	90.2	86.4	66.6	75.1	76.6	80.3	91.9	87.7	67.9	77.5	78.1	82.3

注: 下划线数字为每列次优值。

一个2 048对20的全连接层作为分类器预测每个类别的概率，损失函数采用交叉熵损失。

2）SIG-MLIC w/o $\boldsymbol{D}$ 表示未使用语义字典。将由特征学习模块获得的全局特征$\boldsymbol{f}$输入一个由2 048对20的全连接层构成的分类器中预测每个类别的概率，损失函数采用交叉熵损失。

3）SIG-MLIC w/o SGA_1 表示使用GMP代替SGA模块。将特征图$\boldsymbol{f}^K$经过GMP获得2 048维的特征向量，并输入视觉特征表示模块，损失函数和SIG-MLIC保持一致。

4）SIG-MLIC w/o SGA_2 表示使用全局平均池化(global average pooling, GAP)代替SGA模块。将特征图$\boldsymbol{f}^K$经过GAP获得2 048维的特征向量，并输入视觉特征表示模块，损失函数和SIG-MLIC保持一致。

在表4中比较了4种消融方案与SIG-MLIC在mAP上的评分变化。

表4 在Pascal VOC 2007数据集上的消融实验

Table 4 Ablation experiment on Pascal VOC 2007 dataset

方案	名称	mAP/%
1	ResNet-101	92.6
2	SIG-MLIC w/o $\boldsymbol{D}$	93.4
3	SIG-MLIC w/o SGA_1	95.1
4	SIG-MLIC w/o SGA_2	89.5
本文方法	SIG-MLIC	**95.5**

注：粗体数字为最优值。

3.4.1 语义引导的注意力机制分析

为分析SGA模块的实际贡献，首先，对比方案1和方案2。根据表4中的结果，方案2在mAP评分上比方案1高出0.8%，这表明视觉特征通过SGA模块关注到与标签相关的区域，提升了模型预测的可靠性。其次，对比方案3、方案4和本文方法SIG-MLIC发现，三者分别通过GMP、GAP和SGA生成全局特征输入由字典学习模块训练得到的分类器中对标签进行预测，依次得到了95.1%、89.5%和95.5%的mAP值。方案4表现最差的可能原因是GAP强调对图像的整体特征信息进行下采样，获取的特征表示并不具备判别性，这不利于语义字典对视觉特征的重建。而GMP关注的正是图像中具有判别性的区域，所以方案3也获得了较为理想的效果，但仍低于本文方法。因此，语义引导的注意力机制有效探索了语义和视觉空间的相关性，减少了视觉特征中无关信息的干扰，对语义字典对视觉特征的重建发挥了正向的促进作用。

3.4.2 字典学习分析

为分析字典学习在模型中的实际贡献，在消融实验中对方案1和方案3进行比较。可以观察到相比于方案1，方案3的mAP评分提升了2.5%，这表明标签相关性约束下的语义字典具有反映标签之间关系的语义表达能力，有效利用了标签和视觉空间之间的相关性提升模型预测能力。此外，DSDL方法较早将深度字典学习引入多标签图像分类领域的方法，作为所提出框架的基线，本文强调与DSDL方法的比较，以证明SIG-MLIC方法的有效性。在表1～表3中，分别对比了SIG-MLIC方法与DSDL方法在VOC 2007、VOC 2012和MS-COCO这3个数据集上的表现。相比于DSDL方法，SIG-MLIC方法在mAP评分上分别提升了1.1%、1.6%和3.9%。

字典学习带来的性能增益得益于SIG-MLIC方法利用注意力机制提取到语义特定的特征表示，这正向促进了语义字典对视觉特征进行重建。此外，在语义空间中引入的标签相关性约束使得语义字典能够更好地反映标签之间的关系，增强了语义信息的表达能力。本节通过可视化分析直观展示了字典学习在本文方法中的贡献。

3.5 可视化分析

为更直观地展现加入语义字典以后，模型提取得到的特征有进一步的改进，图3为一些示例的可视化效果，其中，C_{f}表示置信度。每行中的4个图像依次是输入图像和最高的3个置信度类别相对应的语义特征注意图，输入图像的下方是真实标签，类激活图下方是对应的预测标签名和置信度分数。其中，第1行和第3行是由SIG-MLIC方法生成的，第2行和第4行是由3.3节中未使用语义字典的方案2生成的。

对比前2行可以发现，SIG-MLIC方法很好地突出了真实标签“chair”、“TV”和“person”的语义相关区域，并给出了较高的置信度分数。而没有使用语义字典的网络没能准确判断出标签“TV”的存在，且误判出了“sofa”。对比后2行也可以发现类似的情况。值得一提的是，后2行的输入图片中虽然标签“chair”对应的目标出现在边缘的位置，且不完整，但SIG-MLIC方法仍准确地关注到了其对应的语义区域，并给出了较高的置信度分数。

3.6 参数敏感性分析

在总体优化目标中存在3个可学习的超参数α、β和δ，其分别用来调节字典学习损失、L2正则化和标签相关性损失。由于篇幅原因，本文仅在VOC 2007数据集上对参数敏感性进行分析。实验结果如图4所示。

不同α值的结果如图4(a)所示，当α从0取到

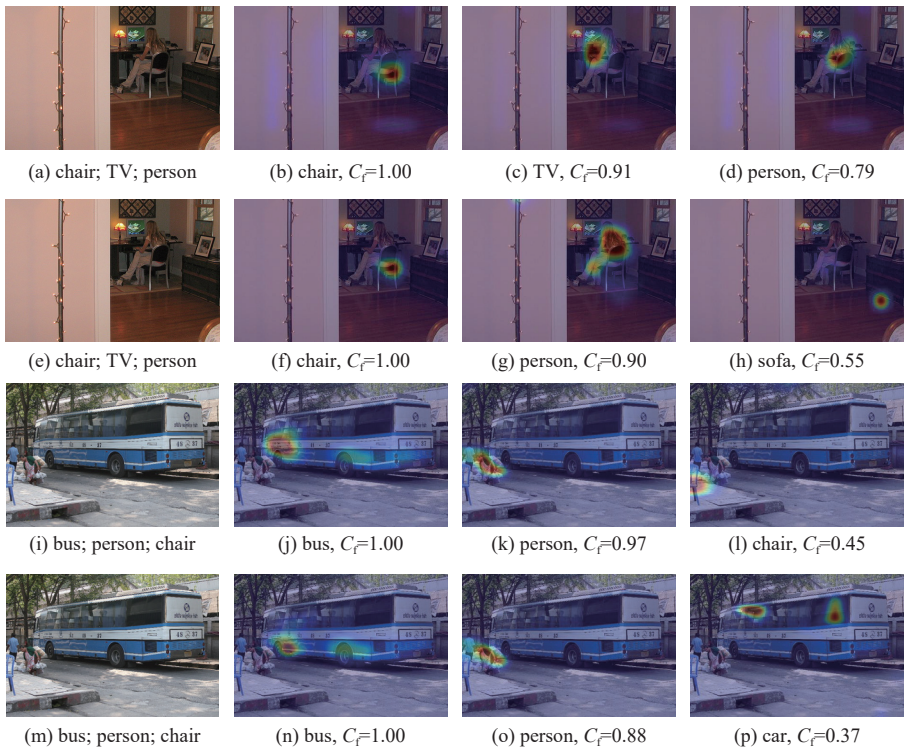


图 3　语义特征注意图可视化示例

Fig. 3　Visualization examples of semantic feature attention maps

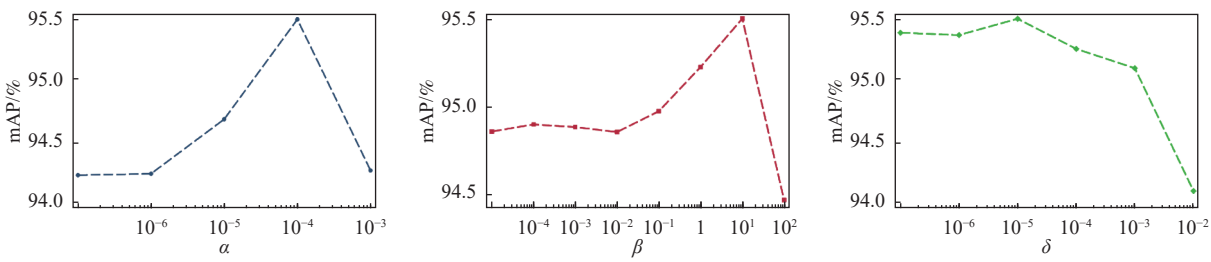


图 4　在 Pascal VOC 2007 数据集上不同α、β和δ的比较

Fig. 4　Comparison of different α, β and δ values on Pascal VOC 2007 dataset

10^{-4}时, 模型性能稳定提升。这表明字典学习损失在模型优化中发挥了重要作用。不同β值的结果如图 4(b) 所示, β越大通常会得到更好的结果, 直到β太大而导致模型崩溃。不同δ值的结果如图 4(c) 所示, 模型对δ不太敏感, 当δ取10^{-5}时, 标签相关性约束带来的效果最好。

4　结　论

1) SIG-MLIC 方法在特征提取过程中利用语义引导的注意机制生成语义特定的特征表示; 在分类器训练过程中设计了一种标签相关性损失将标签的语义信息转换成具有标签相关性约束的语义字典。消融实验和可视化分析证明, 这两方面的特性能够使语义字典对视觉特征更好地进行重建, 以获得较为可靠的归一化表示系数作为预测标签出现的概率。

2) SIG-MLIC 方法实际上是利用标签的语义信息同时开发和协调语义、视觉和标签空间。实验结果证明, 这种解决多标签图像分类任务的路径有效提升了模型的性能。

在未来的研究中, 或考虑将模型迁移到多标签医疗图像诊断领域, 探究其更多的实际应用。

参考文献 (References)

[1] GE Z Y, JIANG X H, TONG Z, et al. Multi-label correlation guided feature fusion network for abnormal ECG diagnosis[J]. Knowledge-Based Systems, 2021, 233: 107508.

[2] CHEN L, LI Z D, ZENG T, et al. Predicting gene phenotype by multi-label multi-class model based on essential functional features[J]. Molecular Genetics and Genomics, 2021, 296(4): 905-918.

[3] DING X M, LI B, XIONG W H, et al. Multi-instance multi-label learning combining hierarchical context and its application to image annotation[J]. IEEE Transactions on Multimedia, 2016, 18(8): 1616-1627.

[4] LIU S L, ZHANG L, YANG X, et al. Query2Label: a simple transformer way to multi-label classification[EB/OL]. (2021-07-22)

[2023-06-02]. https://arxiv.org/abs/2107.10834.

[5] WANG J, YANG Y, MAO J H, et al. CNN-RNN: a unified framework for multi-label image classification[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2016: 2285-2294.

[6] ZHOU F T, HUANG S, LIU B, et al. Multi-label image classification via category prototype compositional learning[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2022, 32(7): 4513-4525.

[7] CHEN T S, XU M X, HUI X L, et al. Learning semantic-specific graph representation for multi-label image recognition[C]//Proceedings of the IEEE International Conference on Computer Vision. Piscataway: IEEE Press, 2019: 522-531.

[8] CHEN Z M, WEI X S, WANG P, et al. Multi-label image recognition with graph convolutional networks[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2019: 5177-5186.

[9] LI Q, PENG X J, QIAO Y, et al. Learning label correlations for multi-label image recognition with graph networks[J]. Pattern Recognition Letters, 2020, 138: 378-384.

[10] WANG Y T, XIE Y Z, LIU Y, et al. Fast graph convolution network based multi-label image recognition via cross-modal fusion[C]//Proceedings of the 29th ACM International Conference on Information & Knowledge Management. New York: ACM, 2020: 1575-1584.

[11] CHEN B Z, LI J X, LU G M, et al. Label co-occurrence learning with graph convolutional networks for multi-label chest X-ray image classification[J]. IEEE Journal of Biomedical and Health Informatics, 2020, 24(8): 2292-2302.

[12] NIU S J, XU Q, ZHU P F, et al. Coupled dictionary learning for multi-label embedding[C]//Proceedings of the International Joint Conference on Neural Networks. Piscataway: IEEE Press, 2019: 1-8.

[13] ZHENG J Y, ZHU W C, ZHU P F. Multi-label quadruplet dictionary learning[C]//Proceedings of the Artificial Neural Networks and Machine Learning. Berlin: Springer, 2020: 119-131.

[14] JING X Y, WU F, LI Z Q, et al. Multi-label dictionary learning for image annotation[J]. IEEE Transactions on Image Processing, 2016, 25(6): 2712-2725.

[15] EVERINGHAM M, VAN GOOL L, WILLIAMS C K I, et al. The pascal visual object classes (VOC) challenge[J]. International Journal of Computer Vision, 2010, 88(2): 303-338.

[16] LIN T-Y, MAIRE M, BELONGIE S, et al. Microsoft COCO: common objects in context[C]//Proceedings of the European conference on computer vision. Berlin: Springer, 2014: 740-755.

[17] ZHU Y, KWOK J T, ZHOU Z H. Multi-label learning with global and local label correlation[J]. IEEE Transactions on Knowledge and Data Engineering, 2018, 30(6): 1081-1094.

[18] WENG W, WEI B W, KE W, et al. Learning label-specific features with global and local label correlation for multi-label classification[J]. Applied Intelligence, 2023, 53(3): 3017-3033.

[19] QU X W, CHE H, HUANG J, et al. Multi-layered semantic representation network for multi-label image classification[J]. International Journal of Machine Learning and Cybernetics, 2023, 14(10): 3427-3435.

[20] CHEN T S, WANG Z X, LI G B, et al. Recurrent attentional reinforcement learning for multi-label image recognition[C]//Proceeding of the AAAI Conference on Artificial Intelligence. Palo Alto: AAAI Press, 2017, 6730-6737.

[21] GAO B B, ZHOU H Y. Learning to discover multi-class attentional regions for multi-label image recognition[J]. IEEE Transactions on Image Processing, 2021, 30: 5920-5932.

[22] WANG Y T, XIE Y Z, ZENG J F, et al. Cross-modal fusion for multi-label image classification with attention mechanism[J]. Computers and Electrical Engineering, 2022, 101: 108002.

[23] WU T, HUANG Q Q, LIU Z W, et al. Distribution-balanced loss for multi-label classification in long-tailed datasets[C]//Proceeding of the European Conference on Computer Vision. Berlin: Springer, 2020: 162-178.

[24] RIDNIK T, BEN-BARUCH E, ZAMIR N, et al. Asymmetric loss for multi-label classification[C]//Proceedings of the IEEE/CVF Inter-national Conference on Computer Vision. Piscataway: IEEE Press, 2021: 82-91.

[25] DONG J X. Focal loss improves the model performance on multi-label image classifications with imbalanced data[C]//Proceedings of the 2nd International Conference on Industrial Control Network and System Engineering Research. New York: ACM, 2020: 18-21.

[26] CAO X C, ZHANG H, GUO X J, et al. SLED: semantic label embedding dictionary representation for multilabel image annotation [J]. IEEE Transactions on Image Processing, 2015, 24(9): 2746-2759.

[27] ZHAO D D, YI M H, GUO J X, et al. A novel image classification method based on multi-layer dictionary learning[C]//Proceedings of the CAA Symposium on Fault Detection, Supervision, and Safety for Technical Processes. Piscataway: IEEE Press, 2021: 1-6.

[28] OU L, HE Y, LIAO S L, et al. FaceIDP: face identification differential privacy via dictionary learning neural networks[J]. IEEE Access, 2023, 11: 31829-31841.

[29] ZHOU F T, HUANG S, XING Y. Deep semantic dictionary learning for multi-label image classification[C]//Proceedings of the AAAI Conference on Artificial Intelligence. Palo Alto: AAAI Press, 2021: 3572-3580.

[30] HUANG S, LIN J K, HUANGFU L W. Class-prototype discriminative network for generalized zero-shot learning[J]. IEEE Signal Processing Letters, 2020, 27: 301-305.

[31] XING C, ROSTAMZADEH N, ORESHKIN B N, et al. Adaptive cross-modal few-shot learning[C]//Proceedings of the Annual Conference on Neural Information Processing Systems. La Jolla: NIPS, 2019: 4848-4858.

[32] HE X T, PENG Y X. Fine-grained image classification via combining vision and language[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2017: 7332-7340.

[33] PENNINGTON J, SOCHER R, MANNING C. Glove: global vectors for word representation[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing. Stroudsburg: Association for Computational Linguistics, 2014: 1532-1543.

[34] WANG Z, FANG Z L, LI D D, et al. Semantic supplementary network with prior information for multi-label image classification[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2022, 32(4): 1848-1859.

[35] YANG H, ZHOU J T, ZHANG Y, et al. Exploit bounding box an-

notations for multi-label object recognition[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2016: 280-288.

[36] WANG M, LUO C Z, HONG R C, et al. Beyond object proposals: random crop pooling for multi-label image recognition[J]. IEEE Transactions on Image Processing, 2016, 25(12): 5678-5688.

[37] HE S, XU C, GUO T, et al. Reinforced multi-label image classification by exploring curriculum[C]//Proceedings of the AAAI Conference on Artificial Intelligence. Palo Alto: AAAI Press, 2018: 3183-3190.

Semantic information-guided multi-label image classification

HUANG Jun[1, 2, *], FAN Haodong[1], HONG Xudong[1], LI Xue[1]

(1. School of Computer Science and Technology, Anhui University of Technology, Ma'anshan 243032, China;
2. Institute of Artificial Intelligence, Hefei Comprehensive National Science Center, Hefei 230088, China)

Abstract: Multi-label image classification aims to predict a set of labels for a given input image. Existing studies based on semantic information either use the correlation between semantic and visual space to guide the feature extraction process to generate effective feature representations or use the correlation between semantic and label spaces to learn weighted classifiers that capture label correlation. Most of these works use semantic information as auxiliary information for exploiting the visual space or label space, and few studies have exploited semantic, visual, and label space correlations simultaneously. However, these approaches fail to model the correlations across semantic, visual, and label spaces simultaneously. To solve this problem, a semantic information-guided multi-label image classification (SIG-MLIC) method was proposed. SIG-MLIC could simultaneously utilize semantic, visual, and label spaces, generating semantically specific feature representations via the association of image regions with labels reinforced by a semantic-guided attention (SGA) mechanism. Besides, the semantic information of labels was used to generate a semantic dictionary with label relevance constraints to reconstruct visual features, obtaining normalized representation coefficients as the probability of label occurrence. Experimental results on three standard multi-label image classification datasets show that both the attention mechanism and dictionary learning in SIG-MLIC can effectively improve classification performance, and the effectiveness of the proposed method has been verified.

Keywords: multi-label image classification; semantic space; visual space; label space; attention mechanism; dictionary learning

Received: 2023-06-16; **Accepted**: 2023-09-15; **Published Online**: 2023-11-08 11:50

URL: link.cnki.net/urlid/11.2625.V.20231107.1105.005

Foundation items: National Natural Science Foundation of China (61806005); The University Synergy Innovation Program of Anhui Province (GXXT-2022-052); Outstanding Young Talents Support Program of Anhui Province (gxyqZD2022032); Natural Science Foundation of the Educational Commission of Anhui Province (KJ2021A0373); Open Project of Anhui Provincial Key Laboratory of Multimodal Cognitive Computation, Anhui University (MMC202008); Natural Science Foundation of Anhui Province (2008085QF305)

* **Corresponding author**. E-mail: huangjun.cs@ahut.edu.cn

http://bhxb.buaa.edu.cn jbuaa@buaa.edu.cn
DOI: 10.13700/j.bh.1001-5965.2023.0472

基于双分支特征增强和多级轨迹关联的多目标跟踪算法

马素刚[1,2]，段帅鹏[1,*]，侯志强[1,2]，余旺盛[3]，蒲磊[4]，杨小宝[1,5]

(1. 西安邮电大学 计算机学院，西安 710121；2. 西安邮电大学 陕西省网络数据分析与智能处理重点实验室，西安 710121；
3. 空军工程大学 信息与导航学院，西安 710077；4. 火箭军工程大学 作战保障学院，西安 710025；
5. 西安邮电大学 西安市大数据与智能计算重点实验室，西安 710121)

摘 要：在多目标跟踪 (MOT) 算法中，经常出现目标特征提取不足、身份切换及轨迹缺失问题，降低跟踪性能。为解决以上问题，提出一种基于双分支特征增强和多级轨迹关联 (MTA) 的 MOT 算法。采用双分支特征学习网络对检测和跟踪 2 种任务的特殊性和相关性进行学习，缓解了两任务之间的过度竞争，提取到充足的目标特征信息；引入关联矩阵 (AM)，利用更多的时序信息预测偏移向量，减少身份切换次数；采用多级轨迹关联策略，保留一部分低分检测框，并将检测框重新划分为高分框和低分框，采用不同的匹配方式与轨迹进行关联，减少轨迹缺失次数。在典型多目标跟踪数据集 MOT17 和 MOT20 上，对 JDE、CenterTrack 等 6 种相关算法进行对比实验。实验结果表明：所提算法在 MOT17 数据集上的多目标跟踪准确度 (MOTA) 和身份 F_1 分数 (IDF1) 值分别达到 68.2% 和 68.5%，与基准算法 CenterTrack 相比，分别提升了 2.1%、4.3%；在 MOT20 数据集上，MOTA 和 IDF1 值分别达到 52.7% 和 48.2%，分别提升了 1.4%、7.9%。所提算法在复杂场景下取得了优异的跟踪性能。

关 键 词：多目标跟踪；双分支特征增强；关联矩阵；偏移向量；多级轨迹关联
中图分类号：TP391
文献标志码：A **文章编号**：1001-5965(2025)07-2282-08

多目标跟踪 (multiple object tracking, MOT) 是计算机视觉领域的重要研究方向，旨在检测连续视频帧中多个目标的位置，识别哪些属于同一目标并生成目标的运动轨迹[1]。在视频分析[2]、自动驾驶[3]、机器人[4] 等实际场景中有着非常广泛的应用。

近年来，随着目标检测器准确性和速度的提升，MOT 算法得到了快速发展[5]，当前，MOT 算法模型分为 2 种，即两阶段和单阶段模型。两阶段模型[6-7] 遵循先检测后跟踪的范式，其将 MOT 分为 2 个独立的任务，先通过检测器对视频帧进行检测得到目标的边界框，再进行目标间的关联跟踪。两阶段模型的跟踪准确度虽然处于领先水平，但是模型比较复杂，计算代价高。单阶段模型[8-9] 遵循联合检测和跟踪范式，其在同一个网络中同时进行检测和关联。与两阶段模型相比，单阶段模型简化了模型复杂性，较好地提升了跟踪速度，能够兼顾算法准确度和跟踪速度。

然而，面临复杂的场景时，单阶段模型会出现目标特征提取不足、身份切换及轨迹缺失等问题，降低 MOT 的性能。针对多数 MOT 算法的检测器，

收稿日期：2023-07-17；**录用日期**：2023-10-25；**网络出版时间**：2023-10-30 15:50
网络出版地址：link.cnki.net/urlid/11.2625.V.20231030.1414.005
基金项目：国家自然科学基金 (62072370)；陕西省自然科学基金 (2023-JC-YB-598)；西安市科技计划项目 (22GXFW0125)
* **通信作者**. E-mail：984187621@stu.xupt.edu.cn

引用格式：马素刚，段帅鹏，侯志强，等. 基于双分支特征增强和多级轨迹关联的多目标跟踪算法 [J]. 北京航空航天大学学报，2025，51（7）：2282-2289. MA S G，DUAN S P，HOU Z Q，et al. Multi-object tracking algorithm based on dual-branch feature enhancement and multi-level trajectory association [J]. Journal of Beijing University of Aeronautics and Astronautics，2025，51（7）：2282-2289（in Chinese）.

如果对目标特征提取不足, 将无法输出高质量的检测结果, 从而限制了跟踪器的性能。CSTrack 算法[10]发现检测与跟踪 2 个任务之间存在过度竞争, 导致不能充分提取目标特征。CenterTrack [9] 算法仅关联相邻帧的目标, 不会重新连接已丢失的远程轨迹, 容易造成身份切换 (IDentity switch, IDs)。此外, 当遇到目标遮挡或模糊等场景时, 一部分检测框置信度可能会降低, 进而被过滤掉, 从而导致一部分轨迹被迫产生中断。

本文基于 CenterTrack 算法, 提出一种基于双分支特征增强和多级轨迹关联(multi-level trajectory association, MTA)的 MOT 算法, 可以有效解决目标特征提取不足、身份切换及轨迹缺失问题, 提高复杂场景下算法的跟踪性能。主要创新点如下:

1) 本文提出一种双分支特征学习网络, 能够对目标特征进行增强。通过对检测与跟踪的特殊性和相关性进行学习, 缓解了检测与跟踪的过度竞争, 能够得到充足的目标特征信息。

2) 本文引入关联矩阵(association matrix, AM), 减少身份切换的次数。替换跟踪分支中的回归预测为重识别(re-identification, ReID), 进而构建关联矩阵, 通过学习相邻帧之间的相似关系, 得到可靠的偏移向量, 从而在轨迹关联部分能够实现更准确的关联。

3) 本文提出一种多级轨迹关联策略, 能够减少轨迹缺失。保留一部分低分检测框, 并将检测框重新划分为高分框与低分框, 对 2 种得分框采用不同的匹配方式与轨迹进行关联。

1　本文算法

本文基于 CenterTrack 算法框架, 从双分支特征学习(dual-branch feature learning, DFL)网络、AM 和 MTA 策略 3 个方面进行改进, 总体框架如图 1 所示。主干网络提取图像共享特征送入 DFL 网络, DFL 网络将共享特征解耦为分别用于检测和跟踪的特征, 输入不同分支。检测分支结构与 CenterNet 算法类似。跟踪分支通过 ReID 提取特征, 进一步构建相邻帧之间的 AM。在轨迹关联阶段, 使用 MTA 策略, 减少轨迹缺失。

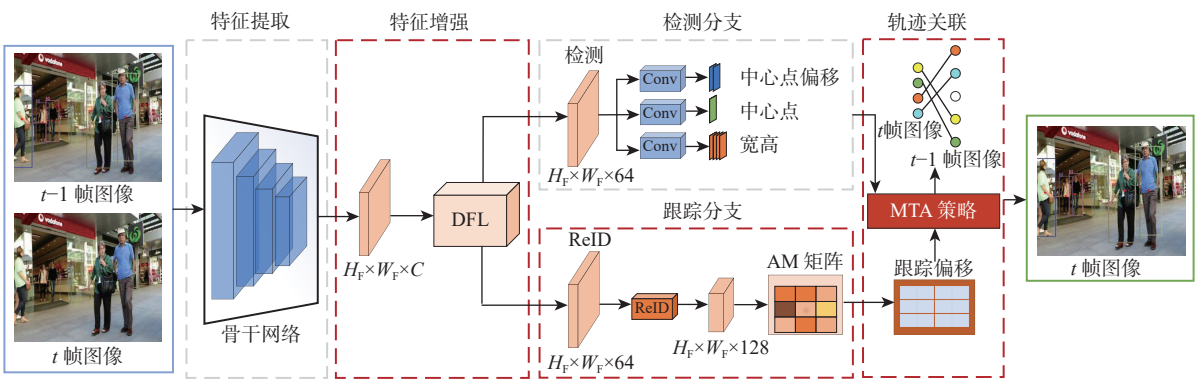


图 1　本文算法总体框架

Fig. 1　Overall framework of the proposed algorithm

1.1　CenterTrack 算法

CenterTrack 算法建立在 CenterNet[11] 检测器基础上, 添加一个额外的两通道输出分支, 其输出结果为偏移向量。然后, CenterTrack 算法基于偏移向量使用简单的贪心算法进行轨迹关联。

1.2　DFL 网络

受 CSTrack[10] 算法提出的互惠网络启发, 本文采用 DFL 网络对用于检测和跟踪 2 种任务的特征进行学习。DFL 网络结构如图 2 所示, 从骨干网络中得到的共享特征 $\boldsymbol{f}\in\mathbf{R}^{H_F\times W_F\times C}$, 其中, H 和 W 表示图像分辨率大小, C 为通道数, F 表示经过骨干网络处理后的图像。首先, 为减少计算量, 对特征图进行池化操作, 其中, 对检测任务进行平均池化, 对跟踪任务进行最大池化。特征 $\boldsymbol{f}$ 经池化后, 得到包含局部信息的 2 种特征, 即检测特征 $\boldsymbol{f}_1\in\mathbf{R}^{H'_F\times W'_F\times C}$ 和跟踪特征 $\boldsymbol{f}_2\in\mathbf{R}^{H'_F\times W'_F\times C}$。其次, 将 $\boldsymbol{f}_1$、$\boldsymbol{f}_2$ 分别通过卷积层进行编码, 生成用于检测和跟踪的特征图 $\boldsymbol{A}_1$ 和 $\boldsymbol{A}_2$, 并且重塑成尺寸为 $C\times H'_F W'_F$ 的二维张量 $\boldsymbol{M}_1$ 和 $\boldsymbol{M}_2$。然后, 对 $\boldsymbol{M}_k$ 及其对应的转置张量进行矩阵乘法并进行归一化, 计算出每个任务的特殊性响应图 $\boldsymbol{S}_k\in\mathbf{R}^{C\times C}$, 计算式为

$$s_k^{ij}=\exp(\boldsymbol{m}_k^i\cdot\boldsymbol{m}_k^j)\Big/\sum_{l=1}^{C}\exp(\boldsymbol{m}_k^i\cdot\boldsymbol{m}_k^l)\qquad k\in\{1,2\}\quad(1)$$

式中: $\boldsymbol{m}_k^i$、$\boldsymbol{m}_k^j$ 和 $\boldsymbol{m}_k^l$ 表示 $\boldsymbol{M}_k$ 的第 i、j 和 l 行; s_k^{ij} 为 $\boldsymbol{S}_k$ 上 (i,j) 位置的值, 表示特征图中第 4 个通道和第 j 个通道之间的相关关系。接下来, $\boldsymbol{M}_1$ 和 $\boldsymbol{M}_2^{\mathrm{T}}$, 以及 $\boldsymbol{M}_2$ 和 $\boldsymbol{M}_1^{\mathrm{T}}$ 进行矩阵乘法, 以学习不同任务之间的相关性,

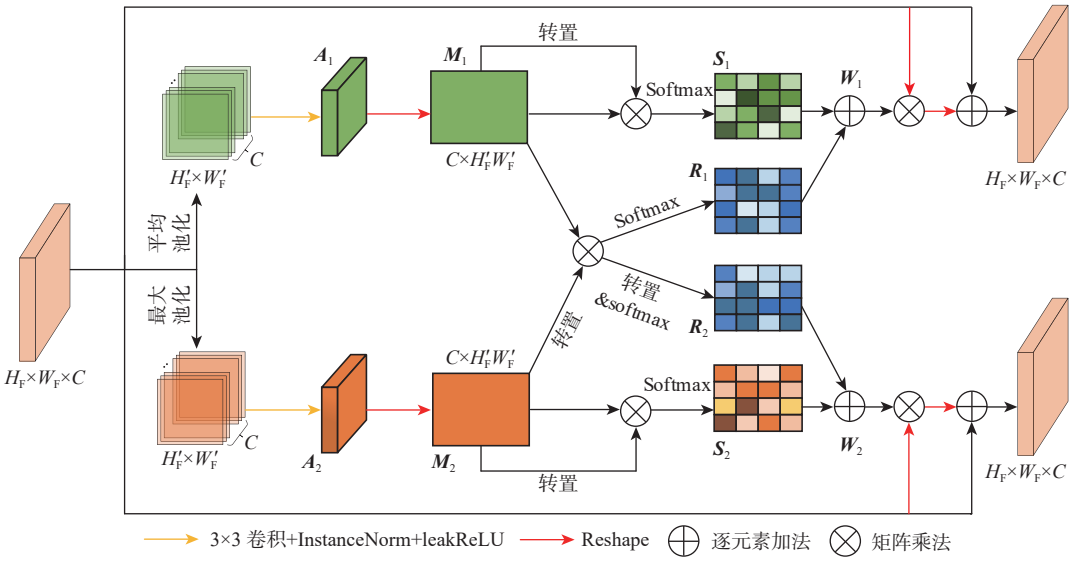


图 2 DFL 网络结构

Fig. 2 DFL networks structure

并经过归一化得到任务之间的相关性响应图 $\boldsymbol{R}_k \in \mathbf{R}^{C\times C}$，计算式为

$$r_k^{ij} = \exp(\boldsymbol{m}_k^i \cdot \boldsymbol{m}_h^j) \Big/ \sum_{l=1}^{C} \exp(\boldsymbol{m}_k^i \cdot \boldsymbol{m}_h^l) \quad (k,h) \in \{(1,2),(2,1)\} \tag{2}$$

式中：r_1^{ij}表示任务 1 的第4个通道对任务 2 的第 j个通道的相关关系；r_2^{ij}表示任务 2 的第4个通道对任务 1 的第 j个通道的相关关系。r_k^{ij}值越大，表示该通道的特征信息被 2 个任务共同关注程度越大。最后，通过一个可训练的参数λ来融合特殊性响应图和相关性响应图，得到特征增强响应图 $\boldsymbol{W}_k \in \mathbf{R}^{C\times C}$，计算式为

$$\boldsymbol{W}_k = \lambda_k \boldsymbol{S}_k + (1-\lambda_k)\boldsymbol{R}_k \qquad k \in \{1,2\} \tag{3}$$

将不同任务对应的特征增强响应图 $\boldsymbol{W}_1$和 $\boldsymbol{W}_2$与重塑后的输入特征进行矩阵相乘，获得每个任务的增强特征，将其重塑为与输入特征 $\boldsymbol{f}$相同形状，并与 $\boldsymbol{f}$进行融合，以防止信息的丢失。最终得到用于检测分支和跟踪分支输入的 2 种特征。

1.3 关联矩阵

将 CenterTrack 算法中的回归预测分支替换为 ReID 分支后，为利用更多的时序信息，本文引入了 AM[12]。首先，给定上一模块输出的增强特征 $\boldsymbol{f}_2^t \in \mathbf{R}^{H_F\times W_F\times 64}$，通过 ReID 提取网络 $\mu(\cdot)$提取其特征，即 $\boldsymbol{e}^t = \mu\left(\boldsymbol{f}_2^t\right) \in \mathbf{R}^{H_F\times W_F\times 128}$。其中，$\mu(\cdot)$由 4 个卷积层组成，前 2 层为 1×1 的深度可分离卷积中的深度卷积和逐点卷积，后 2 层为 3×3 卷积和 1×1 卷积。

然后，构建相邻帧之间的 AM。构建的 AM 表示为 $\boldsymbol{A} \in \mathbf{R}^{H_F\times W_F\times H_F\times W_F}$，可以通过 $\boldsymbol{e}^t$和 $\boldsymbol{e}^{t-1}$对应的转置张量进行矩阵乘法得到，代表图像 $\boldsymbol{I}^t$和 $\boldsymbol{I}^{t-1}$之间的相似度，计算式为

$$\boldsymbol{A}_{i,j,m,n} = \boldsymbol{e}_{i,j}^t (\boldsymbol{e}_{m,n}^{t-1})^{\mathrm{T}} \tag{4}$$

式中：$\boldsymbol{A}_{i,j,m,n}$为第 t帧图像目标点 (i,j)与第 $t-1$帧图像目标点 (m,n)之间特征相似度。对于第 t帧目标 x的中心点 (i,j)，可以从矩阵 $\boldsymbol{A}$中得到其对应的二维关联矩阵 $\boldsymbol{A}_{i,j} \in \mathbf{R}^{H_F\times W_F}$，其表示目标 x和第 $t-1$帧图像上所有点之间的特征相似度。

接下来，通过 $\boldsymbol{A}_{i,j}$求偏移向量。分别以水平和垂直 2 个方向的最大池化作用于 $\boldsymbol{A}_{i,j}$，池化核分别为 $H_F \times 1$和 $1 \times W_F$，得到 2 个方向的矩阵 $\boldsymbol{A}_{i,j}^W \in \mathbf{R}^{1\times W_F}$和 $\boldsymbol{A}_{i,j}^H \in \mathbf{R}^{H_F\times 1}$，将 $\boldsymbol{A}_{i,j}^W$和 $\boldsymbol{A}_{i,j}^H$经过 softmax 函数进行归一化，得到 2 个向量 $\boldsymbol{G}_{i,j}^W \in [0,1]^{1\times W_F}$和 $\boldsymbol{G}_{i,j}^H \in [0,1]^{H_F\times 1}$，$\boldsymbol{G}_{i,j}^W$和 $\boldsymbol{G}_{i,j}^H$分别表示目标出现在第 $t-1$帧上的水平方向位置和垂直方向位置的概率向量。根据输出图像分辨率的大小定义 2 个方向的偏移模板，分别为 $\boldsymbol{X}_{i,j} \in \mathbf{R}^{1\times W_F}$和 $\boldsymbol{Y}_{i,j} \in \mathbf{R}^{H_F\times 1}$，其表示目标实际出现在其他位置的偏移值，计算式为

$$X_{i,j,n} = (n-j)s \qquad 1 \leqslant n \leqslant W_F \tag{5}$$

$$Y_{i,j,m} = (m-i)s \qquad 1 \leqslant m \leqslant H_F \tag{6}$$

式中：s表示下采样倍率，设置为 4；$X_{i,j,n}$和 $Y_{i,j,\mathrm{m}}$分别指目标在第 $t-1$帧上出现在位置 $(*,n)$和位置 $(m,*)$的偏移量。最终的跟踪偏移就是由实际目标位置的偏移值与其对应位置上的偏移概率的点积得到：

$$\boldsymbol{A}_{i,j}^{\mathrm{O}} = \left[\boldsymbol{A}_{i,j}^{HT} \cdot \boldsymbol{Y}_{i,j}, \boldsymbol{A}_{i,j}^{W} \cdot \boldsymbol{X}_{i,j}^{\mathrm{T}}\right]^{\mathrm{T}} \tag{7}$$

水平位置与垂直位置的偏移分别通过 2 个通道学习，最终得到跟踪的偏移为 $\boldsymbol{A}^{\mathrm{O}} \in \mathbf{R}^{H_F\times W_F\times 2}$，用于后续的轨迹关联。

1.4　MTA 策略

本文提出一种 MTA 策略, 以减少轨迹缺失, 如图 3 所示, 其中, η 表示检测框置信度得分。多数跟踪器在检测框得分低于阈值时, 会过滤掉这些低分框。然而, 这些低分框可能是因为遮挡等原因, 导致得分的下降, 因此, 简单地过滤掉这些低分框容易造成轨迹的缺失。本文根据设定的阈值将检测框重新划分为高分框与低分框。高分框可以提取到准确的目标特征信息, 以此实现目标的长期关联, 低分框用来恢复缺失的轨迹, 对这 2 种得分框分别采用不同的匹配方式实现与轨迹的关联。

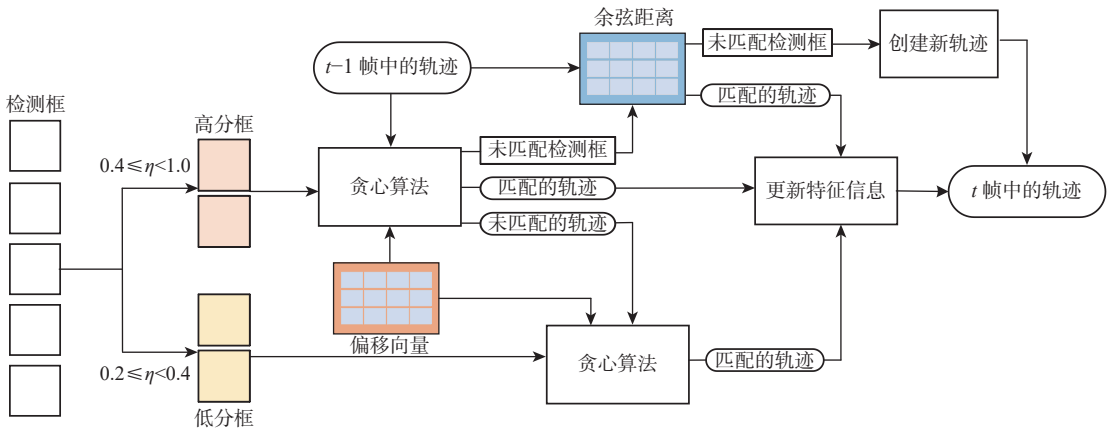


图 3　MTA 策略

Fig. 3　MTA strategy

在轨迹关联时, 首先, 高分检测框基于偏移向量与上一帧的轨迹进行简单的贪心算法匹配, 产生未匹配检测框、匹配的轨迹和未匹配轨迹; 然后, 未匹配检测框与上一帧的跟踪轨迹计算特征之间的余弦相似性, 进行二次匹配, 若低于阈值则创建新的轨迹, 以此可以实现目标遮挡后的重新关联。若高于阈值则匹配成功, 将之添加进轨迹, 并更新第 i 个轨迹在第 t 帧的特征信息 $\boldsymbol{f}_i^t$ 为

$$\boldsymbol{f}_i^t = \varepsilon \boldsymbol{f}_i^{t-1} + (1-\varepsilon)\boldsymbol{e}_i^t \tag{8}$$

式中: $\boldsymbol{e}_i^t$ 为 ReID 网络提取的当前帧图像特征; ε 为权重。而未匹配的轨迹, 可能是低分检测框因遮挡被过滤掉引起的, 所以让低分检测框与未匹配的轨迹进行二次匹配。

2　实　验

2.1　数据集和评估指标

在 MOT17 和 MOT20 数据集上对本文算法进行评估, 利用多目标跟踪准确度(multiple object tracking accuracy, MOTA)、身份 F_1 分数(Identity F_1 Score, IDF1)指标验证本文算法的有效性。

MOT17。由 7 个用于训练的视频序列和 7 个用于测试的序列组成。MOT17 带有 3 个不同对象检测器产生的边界框, 即可变形部件模型(deformable parts model, DPM)、Faster 基于区域的卷积神经网络(region-based convolutional neural networks, R-CNN)和堆叠可变形部件模型(stacked deformable part model, SDP)。本文使用 MOT17 训练集中每个视频的前半部分进行训练, 后半部分用于测试。

MOT20。与 MOT17 相比, MOT20 更具挑战性, 其由在 3 个非常拥挤的场景中捕获的 8 个视频序列组成, 其数据呈现高度多样性。实验部分仍然将 MOT20 训练序列分为两部分, 一部分用于训练, 一部分用于测试。

评估指标。主要的指标[13]是 MOTA 和 IDF1 分数。MOTA 表示跟踪器的整体性能, 其通过评估假阴性 (false negative, FN)、假阳性 (false positive, FP) 和 IDs 次数 3 个来源衡量。IDF1 分数表示跟踪器的关联性能。本文还使用了其他常见的评估指标, 包括最多跟踪目标比率 (mostly tracked, MT) 和最多丢失目标比率 (mostly lost, ML)。

实验使用的硬件配置为: NVIDIA DGX-STATION 工作站, Intel(R) Xeon(R) CPU E5-2698 v4 CPU, 4 块 Tesla v100 DGXS 32 GB GPU; 软件环境为: Ubuntu 18.04 操作系统、PyTorch 1.3.1 框架。本文实验中主干网络、图像尺寸、预训练、分数阈值等参数采用与 CenterTrack[11] 算法相同的设置。

2.2　消融实验

表 1 为本文算法在 MOT17 上的消融实验结果, 验证了本文算法各个模块的有效性, 表中, 最优结果加粗显示。表中, 第 1 行数据为基准算法 CenterTrack 的结果。

AM 的有效性。表 1 中第 2 行数据表示引入 AM 矩阵的结果。由表可知, 使用 AM 矩阵进行偏

表1 本文算法在MOT17测试集上的消融实验

Table 1 Ablation experiments of the proposed algorithm on MOT17 test set

基准算法	AM	DFL	MTA	MOTA/%	IDF1/%	IDs/次	MT/%	ML/%
CenterTrack				66.1	64.2	528	41.3	21.2
	√			67.1	68.6	369	41.0	19.2
	√	√		67.4	**69.1**	368	40.4	18.6
	√		√	67.9	67.8	344	44.2	18.3
	√	√	√	**68.2**	68.5	**333**	**44.5**	**16.8**

移预测，MOTA值提升了1.0%，IDF1提升了4.4%，IDs从528降到了369。这是因为AM由相邻帧之间的相似关系构成，包含更多的时序信息，因此，预测的偏移向量更加准确。以AM预测的偏移向量进行轨迹关联，可以减少目标IDs的次数，提升关联能力IDF1。

DFL网络的有效性。表1中第3行数据表示采用了DFL网络的结果，其中，MOTA提升了0.3%，IDF1提升了0.5%。这是因为DFL网络的2个分支将输入的共享特征分别进行增强，得到检测特征和跟踪特征，缓解了检测和跟踪之间的竞争问题。

MTA策略的有效性。由表1中第4行数据可知，在引入AM矩阵的基础上采用MTA策略，MOTA提升了0.8%，但是IDF1分数降低了0.8%，跟踪器总体性能提升了但是关联能力降低了。因为MTA策略利用了一部分低分框，保留其会导致假阳性FP的升高，IDF1分数因此会下降。其中MT增加了3.2%，这表明缺失轨迹大量减少。MT表示视频跟踪过程中目标至少80%的轨迹都被正确跟踪的比率。

如表1最后一行数据所示，同时加入AM、DFL和MTA这3个模块后，MOTA和IDF1分别提升了2.1%、4.3%，IDs从528次降低到333次，本文算法在复杂场景下有效提升了MOT的性能。

2.3 定性分析

本文选取MOT17测试集对本文算法进行定性分析，并与CenterTrack算法对比跟踪效果。

图4为对比算法和本文算法在MOT17-04视频序列中可视化结果。在远景场景下，目标容易受到



(a1) 第271帧 (a2) 第317帧 (a3) 第350帧



(a4) 第316帧 (a5) 第341帧 (a6) 第247帧

(a) CenterTrack算法在MOT17-04下的跟踪结果



(b1) 第271帧 (b2) 第317帧 (b3) 第350帧



(b4) 第316帧 (b5) 第341帧 (b6) 第247帧

(b) 本文算法在MOT17-04下的跟踪结果

图4 远景下局部遮挡的对比结果

Fig. 4 Comparison results of localized occlusion in distant view

局部遮挡影响, 从图 4(a) 可以看到, 红色箭头处的目标 103 和 128 在被遮挡后产生了 IDs 现象。而如图 4(b) 所示, 本文算法能够在遮挡场景中保持原有的目标 ID, 提高 MOT 的跟踪性能。

图 5 为对比算法与本文算法在 MOT17-09 的可视化结果。在近景场景中, 目标容易受到严重遮挡的影响。对比算法中, 图 5(a) 红色箭头处目标 21 在经历完全遮挡后, 就产生了新的 ID 值为 29, 而本文算法中, 图 5(b) 可以实现遮挡后的重新关联, 即在完全遮挡后, 目标再出现时可以通过 ReID 提取外观特征与轨迹进行余弦距离匹配, 以应对强遮挡场景, 提高跟踪器跟踪性能。图 6 为 MTA 策略处理轨迹缺失场景, 可以看出本文算法可以有效减少轨迹的缺失。



(a1) 第 198 帧　(a2) 第 205 帧　(a3) 第 215 帧

(a) CenterTrack 算法在 MOT17-09 下的跟踪结果



(b1) 第 198 帧　(b2) 第 205 帧　(b3) 第 215 帧

(b) 本文算法在 MOT17-09 下的跟踪结果

图 5　近景下严重遮挡的对比结果

Fig. 5　Comparison results of severe occlusion in close-up view



(a1) 第 88 帧　(a2) 第 108 帧

(a) CenterTrack 算法在 MOT17-11 下的跟踪结果



(b1) 第 88 帧　(b2) 第 108 帧

(b) 本文算法在 MOT17-11 下的跟踪结果

图 6　MTA 处理轨迹缺失场景

Fig. 6　MTA processing trajectory missing scenarios

2.4　定量分析

为进一步验证本文算法的有效性, 选择了近年来 6 种先进的 MOT 算法: CTracker[14]、JDE(joint detection and embedding)[8]、CenterTrack[9]、QuasiDense[15]、TransTrack[16] 和 MOTR[17], 在 MOT17 和 MOT20 数据集上进行对比分析, 如表 2 所示, 其中, MOT17 中对最优和次优数据加粗, MOT20 对最优数据加粗, speed 表示每秒处理图像的帧数。

本文算法在 MOT17 和 MOT20 数据集上的 MOTA 和 IDF1 指标分别达到了 68.2%、68.5% 和 52.7%、48.2%, 均取得了最优的跟踪结果。

在 MOT17 上, 尽管本文算法的 MOTA 和 IDF1 指标达到了最优, 但 IDs 与 FP 处于次优。与 TransTrack 算法相比, 其 IDs 与 FP 值优于本文算法, 这

表 2　不同跟踪算法实验对比

Table 2　Experimental comparison of different tracking algorithms

数据集	算法	MOTA/%	IDF1/%	IDs/次	FP	FN	speed/(帧·s^{-1})
MOT17	CTracker[14]	63.1	60.9	755	2 955	16 174	6.8
	JDE[8]	60.0	63.6	473	2 923	18 158	**22.2**
	CenterTrack[9]	66.1	64.2	528	2 442	15 286	17.5
	QuasiDense[15]	**67.3**	67.8	377	2 637	**14 605**	**20.3**
	TransTrack[16]	67.1	**68.3**	**254**	**1 652**	15 817	10.0
	MOTR[17]	64.7	67.2	346	5 278	**13 452**	7.5
	本文算法	**68.2**	**68.5**	**333**	**2 062**	14 764	16.8
MOT20	CenterTrack[9]	51.3	40.3	7 731	**10 080**	281 757	**10.2**
	本文算法	**52.7**	**48.2**	**3 043**	13 403	**274 419**	7.8

是因为部分低分框中存在背景干扰，保留其会导致本文算法错将一些背景与目标进行关联。

在MOT20上，与基准算法CenterTrack相比，本文算法MOTA指标提升了1.4%，IDF1指标提升了7.9%，IDs从7 731降低到3 043，FP从10 080增加到134 03，FN从281 757降低到274 419。结果表明，本文算法能够在复杂的环境下取得优异的跟踪性能。

在跟踪速度上，本文算法与基准算法CenterTrack差别不大，但是低于JDE和QuasiDense算法。JDE算法仅使用ReID。而QuasiDense算法通过密集学习对ReID进行增强，但其在关联部分仅使用最近邻搜索。所以这2种算法有着更快的跟踪速度。而本文算法通过学习ReID构建AM矩阵来得到位移预测，且在关联阶段使用位移和外观特征两类信息，虽然取得了更高的跟踪性能，但是跟踪速度受到了影响。

3 结 论

本文提出一种基于双分支特征增强和多级轨迹关联的MOT算法。

1）本文算法利用双分支特征学习网络DFL缓解了检测与跟踪2种任务的过度竞争，提取充足的目标特征信息。

2）本文算法引入包含更多时序信息的关联矩阵，预测得到更准确的偏移向量，减少身份切换次数。

3）本文算法设计了一种新颖的轨迹关联策略，有效减少了轨迹缺失。

4）本文算法在MOT17和MOT20数据集上均取得了优异的跟踪性能，较好地解决了目标遮挡、模糊等复杂场景下的目标跟踪失败问题。

对于长时间遮挡的目标，本文算法仍然会产生IDs问题，下一步将利用更多的时序信息进行跟踪，更好地解决长时遮挡场景下的MOT问题。

参考文献（References）

[1] 朱姝姝, 王欢, 严慧. 基于帧内关系建模和自注意力融合的多目标跟踪方法[J]. 控制与决策, 2023, 38(2): 335-344.
ZHU S S, WANG H, YAN H. Multi-object tracking based on intra-frame relationship modeling and self-attention fusion mechanism[J]. Control and Decision, 2023, 38(2): 335-344(in Chinese).

[2] CAO M W, ZHENG L P, JIA W, et al. Joint 3D reconstruction and object tracking for traffic video analysis under IoV environment[J]. IEEE Transactions on Intelligent Transportation Systems, 2021, 22(6): 3577-3591.

[3] SUN P, KRETZSCHMAR H, DOTIWALLA X, et al. Scalability in perception for autonomous driving: waymo open dataset[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2020: 2443-2451.

[4] PEREIRA R, CARVALHO G, GARROTE L, et al. Sort and deep-SORT based multi-object tracking for mobile robotics: evaluation with new data association metrics[J]. Applied Sciences, 2022, 12(3): 1319.

[5] 吴孙勇, 周于松, 谢芸, 等. 基于MM-GGIW-PMBM滤波的扩展目标跟踪算法[J]. 北京航空航天大学学报, 2022, 48(12): 2356-2364.
WU S Y, ZHOU Y S, XIE Y, et al. Extended target tracking algorithm based on MM-GGIW-PMBM filter[J]. Journal of Beijing University of Aeronautics and Astronautics, 2022, 48(12): 2356-2364(in Chinese).

[6] BEWLEY A, GE Z Y, OTT L, et al. Simple online and realtime tracking[C]//Proceedings of the IEEE International Conference on Image Processing. Piscataway: IEEE Press, 2016: 3464-3468.

[7] WOJKE N, BEWLEY A, PAULUS D. Simple online and realtime tracking with a deep association metric[C]//Proceedings of the IEEE International Conference on Image Processing. Piscataway: IEEE Press, 2017: 3645-3649.

[8] WANG Z D, ZHENG L, LIU Y X, et al. Towards real-time multi-object tracking[C]//Proceedings of the 16th European Conference on Computer Vision. Berlin: Springer, 2020: 107-122.

[9] ZHOU X Y, KOLTUN V, KRÄHENBÜHL P. Tracking objects as points[C]//Proceedings of the 16th European Conference on Computer Vision. Berlin: Springer, 2020: 474-490.

[10] LIANG C, ZHANG Z P, ZHOU X, et al. Rethinking the competition between detection and ReID in multiobject tracking[J]. IEEE Transactions on Image Processing, 2022, 31: 3182-3196.

[11] ZHOU X Y, WANG D Q, KRÄHENBÜHL P. Objects as points [EB/OL]. (2019-04-16)[2023-07-01]. https://arxiv.org/abs/1904.07850.

[12] WU J L, CAO J L, SONG L C, et al. Track to detect and segment: an online multi-object tracker[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2021: 12347-12356.

[13] STIEFELHAGEN R, BERNARDIN K, BOWERS R, et al. The CLEAR 2006 evaluation[C]//Proceedings of the International Evaluation Workshop on Classification of Events, Activities and Relationships. Berlin: Springer, 2006: 1-44.

[14] PENG J L, WANG C G, WAN F B, et al. Chained-tracker: chaining paired attentive regression results for end-to-end joint multiple-object detection and tracking[C]//Proceedings of the 16th European Conference on Computer Vision. Berlin: Springer, 2020: 145-161.

[15] PANG J M, QIU L L, LI X, et al. Quasi-dense similarity learning for multiple object tracking[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2021: 164-173.

[16] SUN P Z, CAO J K, JIANG Y, et al. Transtrack: multiple object tracking with transformer [EB/OL]. (2020-12-31)[2023-07-01]. https://arxiv.org/abs/2012.15460.

[17] ZENG F G, DONG B, ZHANG Y A, et al. MOTR: end-to-end multiple-object tracking with transformer[C]//Proceedings of the 17th European Conference on Computer Vision. Berlin: Springer, 2022: 659-675.

Multi-object tracking algorithm based on dual-branch feature enhancement and multi-level trajectory association

MA Sugang[1, 2], DUAN Shuaipeng[1, *], HOU Zhiqiang[1, 2], YU Wangsheng[3], PU Lei[4], YANG Xiaobao[1, 5]

(1. School of Computer Science and Technology, Xi'an University of Posts and Telecommunications, Xi'an 710121, China;
2. Shaanxi Key Laboratory of Network Data Analysis and Intelligent Processing, Xi'an University of Posts and Telecommunications, Xi'an 710121, China;
3. School of Information and Navigation, Air Force Engineering University, Xi'an 710077, China;
4. School of Operational Support, Rocket Force Engineering University, Xi'an 710025, China;
5. Xi'an Key Laboratory of Big Data and Intelligent Computing, Xi'an University of Posts and Telecommunications, Xi'an 710121, China)

Abstract: Insufficient target feature extraction and target occlusion situations frequently occur in single-stage multipal object tracking(MOT) algorithms, resulting in a large number of identity switches and degraded tracking performance. A multi-target tracking algorithm based on dual-branch feature enhancement and multi-level trajectory association(MTA) is proposed to solve this problem. In order to alleviate the excessive competition between detection and tracking branches and fully extract the target improvement features for detection and tracking, respectively, a dual-branch feature learning network is utilized to learn the specificity and relevance of both detection and tracking tasks. An association matrix(AM) is introduced to predict a more accurate offset vector for data association by learning the similarity relationship between two frames. In order to recover lost trajectories and achieve long-term target association, a hierarchical trajectory association strategy is used to divide detection frames into high-score frames and low-score frames, and different matching methods are used to associate with the trajectories. On the typical multi-target tracking datasets MOT17 and MOT20, six related algorithms, including CenterTrack and QuasiDense, are tested. The algorithms' multipal object tracking accuracy(MOTA)and identity F_1 score (IDF1) values on MOT17 and MOT20 are 68.2% and 68.5%, respectively, 2.1% and 4.3% higher than those of the benchmark algorithm CenterTrack. On MOT20, the MOTA and IDF1 values are 52.7% and 48.2%, respectively, 1.4% and 7.9% higher. The algorithm solves the identity switching problem better and achieves excellent tracking performance in complex scenarios.

Keywords: multi-object tracking; dual-branch feature enhancement; association matrix; offset vector; multi-level trajectory association

Received: 2023-07-17; **Accepted**: 2023-10-25; **Published Online**: 2023-10-30 15:50

URL: link.cnki.net/urlid/11.2625.V.20231030.1414.005

Foundation items: National Natural Science Foundation of China(62072370); Natural Science Foundation of Shaanxi Province(2023-JC-YB-598); Science and Technology Project of Xi'an City(22GXFW0125)

* **Corresponding author**. E-mail: 984187621@stu.xupt.edu.cn

http://bhxb.buaa.edu.cn jbuaa@buaa.edu.cn
DOI: 10.13700/j.bh.1001-5965.2023.0561

基于变换学习的快速多切片 MRI 重建算法

段继忠*，刘欢
(昆明理工大学 信息工程与自动化学院，昆明 650504)

摘　　要：二维 (2D) 多切片磁共振数据在相邻切片之间具有高度的相关性，通过利用切片间的冗余性能够重建出更高质量的切片图像，但由于硬件条件的限制，2D 多切片磁共振成像（MRI）需要耗费大量时间。为提高 2D 多切片磁共振图像的重建质量和重建速度，将联合稀疏变换学习正则项引入到多切片 Hankel 张量完成（MS-HTC）模型中，提出一种快速 2D 多切片磁共振成像重建（FMS-JTLHTC）算法。该算法使用交替方向乘子法对目标问题进行求解；引入快速迭代收缩阈值法加快收敛，并使用图形处理器对算法进行加速。使用 4 组脑部数据集在 2 种不同采样模式下进行实验，结果表明：FMS-JTLHTC 算法的峰值信噪比（PSNR）相较于同时自动校准和 K 空间估计（SAKE）算法、并行成像数据的局部 K 空间领域的低秩建模（PLORAKS）算法和 MS-HTC 算法分别平均提高了 4.04 dB、3.67 dB 和 2.07 dB，而且重建速度相比 MS-HTC 算法提高了 14 倍。

关　键　词：多切片磁共振成像；Hankel 张量完成；联合稀疏变换学习；交替方向乘子法；快速迭代收缩阈值法
中图分类号：TP391
文献标志码：A　　**文章编号**：1001-5965(2025)07-2290-14

磁共振成像 (magnetic resonance imaging, MRI) 是现代医学影像诊断工具之一，其没有电离辐射，并可提供脑部、心脏、肝脏、血管等软组织的高分辨率图像。然而，由于 MRI 数据是沿着 K 空间中的相位编码方向逐行采集的，获取全采样数据将需要很长的时间。尤其是儿童、老年人这样的患者在对心脏及腹部这些需要屏气的器官进行扫描时，很难在医生进行扫描检测时长时间保持不动，患者在扫描时轻微的动作均可能导致重建出的图像存在伪影，而达不到医生诊断所需的图像质量。因此，减少 MRI 的成像时间具有很重要的实际意义，研究者们已经提出很多加快 MRI 成像的方法。

并行磁共振成像(parallel magnetic resonance imaging, PMRI)中，使用磁共振扫描仪上的多个接收线圈同时采集数据，因为每个独立的接收线圈都具有独特的线圈灵敏度分布，这使得采集到的多线圈数据具有不同的灵敏度信息，从而能够利用这些线圈灵敏度信息重建出质量良好的图像。现有的并行成像技术有灵敏度编码 (sensitivity encoding, SENSE)[1] 算法、空间谐波同时采集 (simultaneous acquisition of spatial harmonics, SMASH)[2] 算法、基于特征向量的迭代自一致性并行成像 (eigenvector maps-based iterative self-consistent parallel imaging reconstruction using, ESPIRiT)[3] 算法，这类算法需要应用到数据中明确的灵敏度信息，但在实际运用中很难精确的获取到线圈灵敏度信息；而另一类则不需要明确的灵敏度信息，这类算法有：广义自动校准部分并行采集 (generalized autocalibarating partially parallel acquisition, GRAPPA)[4] 算法、迭代自一致性并行成像 (iterative self-consistent parallel imaging reconstruction, SPIRiT)[5] 算法、同时自动校准和 K 空间估计 (simultaneous auto-calibrating and K-space estima-

收稿日期：2023-08-31；**录用日期**：2023-12-05；**网络出版时间**：2024-03-28 18:26
网络出版地址：link.cnki.net/urlid/11.2625.V.20240327.1738.002
基金项目：国家自然科学基金 (61861023)；云南省基础研究计划项目 (202301AT070452)
* **通信作者**. E-mail：duanjz@kust.edu.cn

引用格式：段继忠，刘欢. 基于变换学习的快速多切片 MRI 重建算法 [J]. 北京航空航天大学学报，2025，51（7）：2290-2303.
DUAN J Z，LIU H. Fast multi-slice MRI reconstruction algorithm based on transform learning[J]. Journal of Beijing University of Aeronautics and Astronautics，2025，51（7）：2290-2303（in Chinese）.

tion, SAKE)[6] 算法、具有并行成像数据的局部 K 空间领域的低秩建模 (low-rank modeling of local K-space neighborhoods with parallel imaging data, PLORAKS)[7] 算法。在 MRI 中使用并行成像技术能够有效的减少扫描时间。

Donoho[8] 提出压缩感知 (compressed sensing, CS) 理论, 该理论表明: 若信号是稀疏的或在某种变换域下具有稀疏性, 就能通过最优化非线性重建算法利用远低于奈奎斯特采样率的采样数据重建出质量良好的信号。Lustig 等[9] 将 CS 理论运用到 MRI 中, 通过降低 MRI 数据的采集量来减少磁共振成像时间。CS-MRI 重建算法主要利用特殊变换域中磁共振图像的稀疏性, 常见的变换域有: 离散余弦变换域、离散小波变换域[10]、全变分 (total variation, TV)[11], 近年来, Ravishankar 等[12] 提出一种基于字典学习 (dictionary learning, DL) 的 MRI 重建算法——DLMRI, 该算法通过使用 K-奇异值分解 (K-singular value decomposition, K-SVD)[13] 方法从欠采样的 K 空间数据中迭代学习稀疏字典, 并同时重建 MRI 图像。DLMRI 算法在重建过程中需要不断更新学习基于补丁的字典, 这会使得该算法的计算复杂度非常高, 从而导致重建时间长。为解决 DLMRI 算法计算复杂度高的问题, Ravishankar 等[14-15] 提出基于稀疏变换学习 (transform learning, TL) 的 MRI 重建算法, 因为算法中变换学习和稀疏编码 2 个步骤均具有解析解, 所以算法的计算复杂度大大降低, 减少了图像重建时间。CS 的运用, 使得重建 MRI 图像时所需的 K 空间数据大大减少, 因此, 将并行成像与 CS 理论结合起来进一步减少了 MRI 的成像时间, 而且图像的重建质量会比单独使用其中一种方法好[16-20]。CS 与并行成像的结合, 也使得磁共振成像在临床获得更广泛的运用[21-22], 例如: 心脏成像、腹部成像、对比增强成像等。

在 2D 多切片 MRI 中, 当切片间的间隙非常小时 (通常为 0 或几毫米), 相邻的 2D 切片通常是高度相关的, 通过利用切片之间的数据冗余能够有效提升切片的重建质量。而并行成像技术和 CS 方法都只在单独切片中运用, 并没有利用切片之间的相关信息进行图像重建。一些研究者提出了联合多切片重建图像的算法[23-28], 这些算法称为插值压缩感知 (interpolated compressed sensing, iCS) 算法。Murad 等[25] 基于二维变密度欠采样 (2D variable density under-sampling, 2D-VRDU) 模式提出一种新的快速插值压缩感知 (fast interpolated compressed sensing, FiCS) 算法, 相较于 Datta 等[24] 在一维变密度欠采样 (1D variable density under-sampling, 1D-VRDU) 模式下的 iCS 算法有更好的重建性能并且实现了更低的采样率。Deka 等[29] 提出一种基于小波域森林稀疏性和联合全变分 (joint total variation, JTV) 的 CS pMRI 重建算法, 同时利用了多切片信息和多线圈通道信息, 并在多核中央处理器 (central processing unit, CPU) 和通用图形处理器 (general-purpose computing on graphics processing units, GP-GPU) 上执行并行计算来实现算法。

Liu 等[30] 提出多切片 Hankel 张量完成 (multi-slice Hankel tensor completion, MS-HTC) 模型, 通过将 K 空间数据转换成逐块 Hankel 张量来利用多切片多线圈 MRI 数据中相似的线圈灵敏度和相邻切片间相似的图像信息, 以无校准的方式重建出多切片 MRI 图像, 但该方法在高度欠采样情况下, 重建出的多切片图像存在明显的混叠伪影。除此之外, MS-HTC 算法中的张量分解及 K 空间数据和分块 Hankel 张量之间的相互转换的过程计算复杂度高, 导致图像重建时间很长。

本文在 MS-HTC 模型上进一步改进, 将联合稀疏变换学习 (joint sparsifying transform learning, JTL) 正则项与 MS-HTC 算法相结合, 提出一种基于联合稀疏变换学习的快速 2D 多切片 MRI 重建 (fast 2D multi-slice MRI reconstruction, FMS-JTLHTC) 算法。针对目标问题, 首先, 使用交替方向乘子法 (alternating direction method of multipliers, ADMM) 算法[11] 将问题分解成若干子问题来分别进行求解; 其次, 使用快速迭代收缩阈值法 (fast iterative shrinkage-thresholding algorithm, FISTA)[31] 加快本文算法的收敛速度; 最后, 使用图形处理器 (graphics processing unit, GPU) 并行计算加快重建速度。将 FMS-JTLHTC 算法在 1D-VRDU、2D-VRDU 和 1D-Possion 模式下对 4 个数据集进行图像重建, 实验结果表明: 本文算法不但在重建质量上有很大提升, 而且在重建速度上有很大的提高, 对于 4 个切片数据同时重建平均只需要 18 s。

本文 FMS-JTLHTC 算法相较于 MS-HTC 算法进行了几点改进: ①在 MS-HTC 模型的基础上引入 JTL 正则项; ②使用 ADMM 算法对目标问题进行求解; ③由于算法需要大量的迭代, 本文使用 FISTA 技术减少迭代次数来加快收敛速度, 并且使用 GPU 对整个计算过程进行加速。

1　MS-HTC 模型介绍

张量是高维 MRI 数据更自然的表示, 因此, 低秩张量先验被广泛运用到高维或动态 MRI 中[32-36]。低秩张量完成需要对由三阶张量展开而成的 2D 矩阵执行低秩约束, 而将三阶张量展开成为 2D 矩阵会丢失部分其本身所具有的信息, 由此研究者们提

出通过使用张量分解模型的方法来解决这个问题。例如：高阶奇异值分解 (higher-order singular value decomposition, HOSVD)[37]，也被称为 Tucker 分解[38]、多线性奇异值分解 (singular value decomposition, SVD)[39])、张量奇异值分解 (tensor singular value decomposition, t-SVD)[40] 等。HOSVD 通过计算核心张量$\boldsymbol{Q}$及 3 个标准正交基$\boldsymbol{U}^{(n)}$ $(n=1,2,3)$来分解张量，可表示为

$$\boldsymbol{T}=\boldsymbol{Q}\times_1\boldsymbol{U}^{(1)}\times_2\boldsymbol{U}^{(2)}\times_3\boldsymbol{U}^{(3)} \tag{1}$$

式中：$\times_n$表示张量和矩阵的n模乘积；$\boldsymbol{U}^{(n)}$为对n模式展开矩阵$\boldsymbol{T}(n)$进行奇异值分解得到的酉矩阵。

Liu 等[30] 提出 MS-HTC 模型，首先，将所获取的多切片多线圈 K 空间数据构造成为一个三阶张量，在此过程中，将构造的逐块 Hankel 矩阵沿着线圈通道串联起来；其次，应用顺序截断 HOSVD[41] 将构造出来的三阶张量分解并执行低秩近似；最后，执行数据一致性和结构一致性将三阶张量还原为多切片多线圈 K 空间数据。其中，顺序截断 HOSVD 通过顺序秩截断酉矩阵$\boldsymbol{U}^{(n)}$进行来近似低秩张量。

MS-HTC 模型可表述为

$$\begin{cases}\min\limits_{\boldsymbol{f}}\|\boldsymbol{D}\boldsymbol{f}-\boldsymbol{y}\|_2^2\\ \quad\text{s.t. }\boldsymbol{\Gamma}=\boldsymbol{R}\boldsymbol{f},\ \boldsymbol{\Gamma}=\boldsymbol{Q}\times_1\boldsymbol{U}^{(1)}\times_2\boldsymbol{U}^{(2)}\times_3\boldsymbol{U}^{(3)}\end{cases} \tag{2}$$

式中：$\boldsymbol{f}\in\mathbf{C}^{NN_CN_S}$为待重构的多切片多线圈 K 空间数据，$N=N_x\times N_y$为单切片单线圈的总像素点，$N_x$和$N_y$分别为磁共振图像的列数和行数，$N_S$为切片数量，$N_C$为单切片的线圈数量；$\boldsymbol{y}\in\mathbf{C}^{MN_CN_S}$为欠采样的多切片多线圈 K 空间数据，$M$为单切片中单线圈欠采样 K 空间数据的数量；$\boldsymbol{R}(\cdot):\mathbf{C}^{NN_CN_S}\to\mathbf{C}^{bN_C\times N_b\times N_S}$表示将多切片多线圈 K 空间数据构造成三阶张量的算子；b为从单切片单线圈 K 空间中提取数据块的总点数，N_b为从单切片单线圈 K 空间提取的数据块的个数；$\boldsymbol{D}\in\mathbf{C}^{MN_CN_S\times NN_CN_S}$为多切片多线圈欠采样算子；$\boldsymbol{\Gamma}\in\mathbf{C}^{bN_C\times N_b\times N_S}$为低秩近似张量。

2 问题描述

为进一步提高 2D 多切片 MRI 的成像质量，在 MS-HTC 模型的基础上引入 JTL 正则项，得到如下优化问题：

$$\begin{cases}\min\limits_{\boldsymbol{f},\boldsymbol{\Gamma},\boldsymbol{W}_s}\|\boldsymbol{D}\boldsymbol{f}-\boldsymbol{y}\|_2^2+\alpha\|\boldsymbol{\Gamma}-\boldsymbol{R}(\boldsymbol{f})\|_F^2+\\ \quad\beta\sum\limits_{s=1}^{N_S}\sum\limits_{j=1}^{N_p}\left\|\boldsymbol{W}_s\boldsymbol{P}_j\boldsymbol{E}_s(\boldsymbol{F}^{-1}\boldsymbol{f})\right\|_{0,2}\\ \quad\text{s.t. }\boldsymbol{\Gamma}=\boldsymbol{Q}\times_1\boldsymbol{U}^{(1)}\times_2\boldsymbol{U}^{(2)}\times_3\boldsymbol{U}^{(3)},\ \boldsymbol{W}_s^{\mathrm{H}}\boldsymbol{W}_s=\boldsymbol{I}\end{cases} \tag{3}$$

式中：$\boldsymbol{P}_j(\cdot):\mathbf{C}^{NN_C}\to\mathbf{C}^{m\times N_C}$为第$j$个图像块提取算子，表示从单切片图像的单个线圈中提取大小为$\sqrt{m}\times\sqrt{m}$的图像块并组合成大小为$m\times N_C$的矩阵；N_p为从单线圈图像中提取出来的图像块数量；$\boldsymbol{W}_s\in\mathbf{C}^{m\times m}$为从图像块学习到的自适应稀疏变换；$\boldsymbol{E}_s:\mathbf{C}^{NN_CN_S}\to\mathbf{C}^{NN_C}$为第$s$个切片图像的提取算子，表示从多切片多线圈图像中提取单切片多线圈图像的矩阵；$\boldsymbol{F}=\boldsymbol{I}_S\otimes\boldsymbol{I}_C\otimes\boldsymbol{F}_x\otimes\boldsymbol{F}_y$为多切片多线圈傅里叶变换算子，$\boldsymbol{F}_x$、$\boldsymbol{F}_y$分别为$N_x$、$N_y$点傅里叶变换矩阵，$\boldsymbol{I}_C$和$\boldsymbol{I}_S$分别为$N_C\times N_C$和$N_S\times N_S$的单位矩阵，$\otimes$表示克罗内克积；$\|\cdot\|_F$表示 Frobenius 范数；$\|\boldsymbol{w}\|_{0,2}=\left\|\sum\limits_{c=1}^{N_C}|\boldsymbol{w}_c|^2\right\|_0$表示联合 L0 范数。

3 问题求解

为求解目标问题 (3)，使用 ADMM 算法将问题分解为如下几个子问题分别求解：

$$\begin{cases}\boldsymbol{W}_s^{k+1}=\arg\min\limits_{\boldsymbol{W}_s}\sum\limits_{j=1}^{N_p}\left\|\boldsymbol{B}_{sj}^k-\boldsymbol{W}_s\boldsymbol{P}_j\boldsymbol{E}_s(\boldsymbol{F}^{-1}\boldsymbol{f}^k)-\boldsymbol{d}_{\boldsymbol{B}_{sj}}^k\right\|_F^2\\ \quad\text{s.t. }\boldsymbol{W}_s^{\mathrm{H}}\boldsymbol{W}_s=\boldsymbol{I}\end{cases} \tag{4}$$

$$\boldsymbol{B}_{sj}^{k+1}=\arg\min_{\boldsymbol{B}_{sj}}\left\{\beta\left\|\boldsymbol{B}_{sj}\right\|_{0,2}+\frac{\mu_1}{2}\left\|\boldsymbol{B}_{sj}-\boldsymbol{W}_s^{k+1}\boldsymbol{P}_j\boldsymbol{E}_s(\boldsymbol{F}^{-1}\boldsymbol{f}^k)-\boldsymbol{d}_{\boldsymbol{B}_{sj}}^k\right\|_F^2\right\} \tag{5}$$

$$\begin{cases}\boldsymbol{\Gamma}^{k+1}=\arg\min\limits_{\boldsymbol{\Gamma}}\left\|\boldsymbol{\Gamma}-\boldsymbol{R}(\boldsymbol{f}^k)\right\|_F^2\\ \quad\text{s.t. }\boldsymbol{\Gamma}=\boldsymbol{Q}\times_1\boldsymbol{U}^{(1)}\times_2\boldsymbol{U}^{(2)}\times_3\boldsymbol{U}^{(3)}\end{cases} \tag{6}$$

$$\boldsymbol{f}^{k+1}=\arg\min_{\boldsymbol{f}}\left\{\|\boldsymbol{D}\boldsymbol{f}-\boldsymbol{y}\|_2^2+\alpha\left\|\boldsymbol{\Gamma}^{k+1}-\boldsymbol{R}(\boldsymbol{f})\right\|_F^2+\frac{\mu_1}{2}\sum_{s=1}^{N_S}\sum_{j=1}^{N_p}\left\|\boldsymbol{B}_{sj}^{k+1}-\boldsymbol{W}_s^{k+1}\boldsymbol{P}_j\boldsymbol{E}_s(\boldsymbol{F}^{-1}\boldsymbol{f})-\boldsymbol{d}_{\boldsymbol{B}_{sj}}^k\right\|_F^2\right\} \tag{7}$$

问题 (4) 中变换学习矩阵$\boldsymbol{W}_s^{k+1}$可通过 SVD 进行求解[15]，其解表示为

$$\boldsymbol{W}_s^{k+1}=\boldsymbol{V}_s\boldsymbol{U}_s^{\mathrm{H}} \tag{8}$$

问题 (5) 中稀疏编码$\boldsymbol{B}_{sj}^{k+1}$通过联合硬阈值法[42] 进行求解，其解为

$$\boldsymbol{B}_{sj}^{k+1}=H_J\left(\boldsymbol{W}_s^{k+1}\boldsymbol{P}_j\boldsymbol{E}_s(\boldsymbol{F}^{-1}\boldsymbol{f}^k)+\boldsymbol{d}_{\boldsymbol{B}_{sj}}^k,\sqrt{2\beta/\mu_1}\right) \tag{9}$$

其中：$H_J(\boldsymbol{x},\theta)$表示联合硬阈值算子，定义如下：

$$H_J(\boldsymbol{x},\theta)=\begin{cases}\boldsymbol{0} & \sqrt{\sum\limits_{c=1}^{N_C}|\boldsymbol{x}_c|^2}<\theta\\ \boldsymbol{x} & \sqrt{\sum\limits_{c=1}^{N_C}|\boldsymbol{x}_c|^2}\geqslant\theta\end{cases} \tag{10}$$

问题 (6) 中低秩张量$\boldsymbol{\Gamma}^{k+1}$使用 MS-HTC 算法中求解近似低秩张量的方法来进行求解，即使用顺序截断 HOSVD 对构造出来的三阶张量分解并执行低

秩近似[30], 其解为

$$\boldsymbol{\mathcal{I}}^{k+1}=\boldsymbol{\mathcal{Q}}\times_1\hat{\boldsymbol{U}}^{(1)}\times_2\hat{\boldsymbol{U}}^{(2)}\times_3\hat{\boldsymbol{U}}^{(3)} \tag{11}$$

问题 (7) 中多切片 K 空间数据$\boldsymbol{f}^{k+1}$可通过令目标函数的导数为 0 求出, 其解为

$$\boldsymbol{f}^{k+1}=\left(2\boldsymbol{D}^{\mathrm{T}}\boldsymbol{D}+2\alpha\boldsymbol{R}^{*}\boldsymbol{R}+\mu_1 m\boldsymbol{I}\right)^{-1}\left[2\boldsymbol{D}^{\mathrm{T}}\boldsymbol{y}+2\alpha\boldsymbol{R}^{*}(\boldsymbol{\mathcal{I}}^{k+1})+\mu_1\boldsymbol{F}\sum_{s=1}^{N_{\mathrm{S}}}\sum_{j=1}^{N_{\mathrm{p}}}\boldsymbol{E}_s^{*}\left(\boldsymbol{P}_j^{*}\left((\boldsymbol{W}_s^{k+1})^{\mathrm{H}}(\boldsymbol{B}_{sj}^{k+1}-\boldsymbol{d}_{\boldsymbol{B}_{sj}}^{k})\right)\right)\right] \tag{12}$$

式中: $\boldsymbol{R}^{*}(\cdot):\mathbf{C}^{bN_{\mathrm{C}}\times N_b\times N_{\mathrm{S}}}\to\mathbf{C}^{NN_{\mathrm{C}}N_{\mathrm{S}}}$表示将三阶张量还原为多切片多线圈 K 空间数据的算子; 上标 "*" 表示伴随算子。

因为 ADMM 算法求解问题需要大量的迭代, 所以进一步引入 FISTA 技术来减少算法的迭代次数, 以此减少图像的重建时间。

在附录 A 中给出了求解目标问题 (3) 的详细推导过程。

将$\boldsymbol{f}^{k+1}$进行快速傅里叶变换, 然后执行平方和的平方根 (square root of the sum of squares, SOS) 操作得到各切片相对应的图像, 该步骤可将其表示为: $\hat{\boldsymbol{x}}=\sqrt{\sum_{c=1}^{N_{\mathrm{C}}}\mathrm{modulus}\left((\boldsymbol{F}^{-1}\boldsymbol{f})_c\right)^2}$, $\hat{\boldsymbol{x}}\in\mathbf{R}^{NN_{\mathrm{S}}}$为本文算法重建出的多切片图像, modulus(·) 表示复数的模运算。

对于本文算法, 当迭代次数k大于主循环的最大迭代次数K或当相对误差 (relative error, RE) 小于预定容差时停止算法, FMS-JTLHTC 算法实现的具体步骤如下所示。相对误差定义如下:

$$V_{\mathrm{RE}}=\left\|\hat{\boldsymbol{x}}^{k+1}-\hat{\boldsymbol{x}}^{k}\right\|_2/\left\|\hat{\boldsymbol{x}}^{k}\right\|_2 \tag{13}$$

MS-HTC 算法使用 Cadzow 方法[43] 对目标问题进行求解, 而本文 FMS-JTLHTC 算法先使用 ADMM 算法对目标问题进行求解; 然后, 使用 FISTA 技术减少迭代次数来加快收敛速度, 并且使用 GPU 对整个计算过程进行加速。

虽然 2 个算法对目标问题的求解方法不同, 但在求解低秩张量时, MS-HTC 算法和本文 FMS-JTLHTC 算法均使用顺序截断 HOSVD 对三阶张量分解并执行低秩近似。

算法　FMS-JTLHTC。

输入: undersampled K-space data $\boldsymbol{y}$ and sampling mask $\boldsymbol{D}$。

Initial: $\boldsymbol{f}^0=\boldsymbol{D}^{\mathrm{T}}\boldsymbol{y},\boldsymbol{W}_s^0=\mathrm{DCT}_{\sqrt{m}}\otimes\mathrm{DCT}_{\sqrt{m}}$,

$\boldsymbol{B}_s^0=H_{\mathrm{J}}\left(\boldsymbol{W}_s^0\boldsymbol{A}_s^0+\boldsymbol{d}_{\boldsymbol{B}_s}^0,\sqrt{2\beta/\mu_1}\right)$,

$\tilde{\boldsymbol{f}}^0=0,t^0=0,k=0,\boldsymbol{d}_{\boldsymbol{B}_s}^0=\boldsymbol{0}$,

$\boldsymbol{A}_s^0=\left[\boldsymbol{P}_1\boldsymbol{E}_s(\boldsymbol{F}^{-1}\boldsymbol{f}^0),\cdots,\boldsymbol{P}_{N_{\mathrm{p}}}\boldsymbol{E}_s(\boldsymbol{F}^{-1}\boldsymbol{f}^0)\right]$.

Repeat

for $s=1:N_{\mathrm{S}}$ do

Compute the SVD of $\boldsymbol{A}_s^k(\boldsymbol{B}_s^k-\boldsymbol{d}_{\boldsymbol{B}_s}^k)^{\mathrm{H}}$ to get:

$\boldsymbol{A}_s^k(\boldsymbol{B}_s^k-\boldsymbol{d}_{\boldsymbol{B}_s}^k)^{\mathrm{H}}=\boldsymbol{U}_s\boldsymbol{\Sigma}_s\boldsymbol{V}_s^{\mathrm{H}}$

Transform Update: $\boldsymbol{W}_s^{k+1}=\boldsymbol{V}_s\boldsymbol{U}_s^{\mathrm{H}}$

Sparse Coding: $\boldsymbol{B}_s^{k+1}=H_J\left(\boldsymbol{W}_s^{k+1}\boldsymbol{A}_s^k+\boldsymbol{d}_{\boldsymbol{B}_s}^k,\sqrt{2\beta/\mu_1}\right)$

end

Compute: $\boldsymbol{\mathcal{I}}^{k+1}=\boldsymbol{\mathcal{Q}}\times_1\hat{\boldsymbol{U}}^{(1)}\times_2\hat{\boldsymbol{U}}^{(2)}\times_3\hat{\boldsymbol{U}}^{(3)}$

Compute: $\tilde{\boldsymbol{f}}^{k+1}$ by 式 (12)

Compute: $t^{k+1}=\left[1+\sqrt{1+4(t^k)^2}\right]/2$

Compute: $\boldsymbol{f}^{k+1}=\tilde{\boldsymbol{f}}^{k+1}+\frac{t^k-1}{t^{k+1}}(\tilde{\boldsymbol{f}}^{k+1}-\tilde{\boldsymbol{f}}^{k})$

for $s=1:N_{\mathrm{S}}$ do

Compute:

$\boldsymbol{A}_s^{k+1}=\left[\boldsymbol{P}_1\boldsymbol{E}_s(\boldsymbol{F}^{-1}\boldsymbol{f}^{k+1}),\cdots,\boldsymbol{P}_{N_{\mathrm{p}}}\boldsymbol{E}_s(\boldsymbol{F}^{-1}\boldsymbol{f}^{k+1})\right]$

Compute: $\boldsymbol{d}_{\boldsymbol{B}_s}^{k+1}=\boldsymbol{d}_{\boldsymbol{B}_s}^k+(\boldsymbol{W}_s^{k+1}\boldsymbol{A}_s^{k+1}-\boldsymbol{B}_s^{k+1})$

end

Compute: $\hat{\boldsymbol{x}}^{k+1}$ via $\hat{\boldsymbol{x}}^{k+1}=\sqrt{\sum_{c=1}^{N_{\mathrm{C}}}\left|(\boldsymbol{F}^{-1}\boldsymbol{f}^{k+1})_c\right|^2}$

$k=k+1$

Compute: $V_{\mathrm{RE}}=\left\|\hat{\boldsymbol{x}}^{k+1}-\hat{\boldsymbol{x}}^{k}\right\|_2/\left\|\hat{\boldsymbol{x}}^{k}\right\|_2$

Until $V_{\mathrm{RE}}<\mathrm{tol}$ or $k>K$

输出: $\hat{\boldsymbol{x}}=\hat{\boldsymbol{x}}^{k+1}$。

4　实验结果与分析

4.1　实验设置和数据

为验证 FMS-JTLHTC 算法的重建性能, 本文选择 3 种算法作为对比: SAKE 算法、PLORAKS 算法和 MS-HTC 算法, 所有算法均用 MATLAB 来实现, 所有的实验均在一台配置为 Intel (R) Core (TM) i9-13900K @ 3.00 GHz 处理器, 64 GB 内存, NVIDIA GeForce RTX 4090 GPU 的台式机上进行。

实验中选择了 brain_chall_tr4、brain_chall_tr5、brain_chall_tr7 及 brain_chall_tr8 这 4 个数据集[44] 进行实验, 依次将其记为数据集 dataset 1、dataset 2、dataset 3、dataset 4。所用的 4 个数据集均在具有 12 通道线圈的临床扫描仪 (Discovery MR750) 上使用 3D T1 加权梯度回波序列进行采集。扫描参数为: 重复时间为 6.3 ms, 回波时间为 2.6 ms, 反转时间为 650 ms, 视场为 256 mm×218 mm, 切片厚度为 1 mm。分别选择 4 个数据集中的 4 个连续切片进行实验。

相邻切片间使用不同的欠采样模式能够增强不相干性, 有效提升 MS-HTC 模型的性能[30], 因此, 为实验所用到的 4 个切片生成了不同的欠采样掩模。为评估本文 FMS-JTLHTC 算法性能, 使用如下 4 个评价指标来评估 FMS-JTLHTC 算法重建出

的图像质量：峰值信噪比 (peak signal-to-noise ratio, PSNR)[45]、平均结构相似性指数 (mean structural similarity index, MSSIM)[46]、特征相似性指数 (feature similarity index, FSIM)[47] 和高频误差范数 (high-frequency error norm, HFEN)[12]。4个评价指标中PSNR、SSIM和FSIM的数值越大，HFEN数值越小，则表明重建出的图像质量越高。实验中所有的指标都是在图像感兴趣区域下计算得出。

PSNR定义如下：

$$V_{\mathrm{PSNR}}=10\lg\left(\frac{{V_{\max}}^2}{V_{\mathrm{MSE}}}\right) \tag{14}$$

式中：V_{MSE}为重建图像$\hat{\boldsymbol{x}}$和参考图像$\boldsymbol{x}$之间的均方误差；$V_{\max}$为$\boldsymbol{x}$的最大像素值。

MSSIM定义为

$$V_{\mathrm{MSSIM}}=\frac{1}{N_{\mathrm{S}}}\sum_{s=1}^{N_{\mathrm{S}}}\frac{\left(2u_{x_s}u_{\hat{x}_s}+c_1\right)\left(2\sigma_{x_s,\hat{x}_s}+c_2\right)}{\left(u_{x_s}^2+u_{\hat{x}_s}^2+c_1\right)\left(\sigma_{x_s}^2+\sigma_{\hat{x}_s}^2+c_2\right)} \tag{15}$$

式中：$\boldsymbol{x}_s$和$\hat{\boldsymbol{x}}_s$分别表示参考图像和重建图像的第s个切片图像；u_{x_s}和$u_{\hat{x}_s}$为$\boldsymbol{x}_s$和$\hat{\boldsymbol{x}}_s$的均值；$\sigma_{x_s}^2$和$\sigma_{\hat{x}_s}^2$为$\boldsymbol{x}_s$和$\hat{\boldsymbol{x}}_s$的方差；$\sigma_{x_s,\hat{x}_s}$为$\boldsymbol{x}$与$\hat{\boldsymbol{x}}$的协方差；c_1和c_2均为常数，$c_1=0.01$，$c_2=0.03$。

FSIM定义如下：

$$V_{\mathrm{FSIM}}=\frac{\sum_{i,j}\left(S_{\mathrm{PC}}S_{\mathrm{G}}\right)\left(\max(\mathrm{PC}(\hat{\boldsymbol{x}}),\mathrm{PC}(\boldsymbol{x}))\right)}{\sum_{i,j}\max(\mathrm{PC}(\hat{\boldsymbol{x}}),\mathrm{PC}(\boldsymbol{x}))} \tag{16}$$

式中：$\mathrm{PC}(\boldsymbol{x})$和$\mathrm{PC}(\hat{\boldsymbol{x}})$分别为参考图像$\boldsymbol{x}$和重建图像$\hat{\boldsymbol{x}}$的相位一致性；$S_{\mathrm{PC}}$为$\mathrm{PC}(\boldsymbol{x})$和$\mathrm{PC}(\hat{\boldsymbol{x}})$相似性度量；$S_{\mathrm{G}}$为参考图像和重建图像的梯度幅度值的相似性度量；max(·)表示最大值函数。

HFEN定义如下：

$$V_{\mathrm{HFEN}}=\|\mathrm{filter}(\boldsymbol{x})-\mathrm{filter}(\hat{\boldsymbol{x}})\|_2/\|\mathrm{filter}(\hat{\boldsymbol{x}})\|_2 \tag{17}$$

式中：filter(·)为拉普拉斯高斯滤波器，其作用是获取图像的边缘信息。

传统无校准模型的性能会随着内核大小的增加而有一定的提升，但图像重建的时间会随着内核大小的增加而显著增加，所以需要选择合适的内核大小。在实验中，将SAKE算法和MS-HTC算法的内核大小设置为：6×6，将PLORAKS算法的K空间邻域半径设置为3。对于FMS-JTLHTC算法将其参数设置为：$\alpha=10^{-3}$，$\mu_1=10^{-5}$，内核 (算子$\boldsymbol{R}(\cdot)$和$\boldsymbol{P}_j(\cdot)$) 大小均设置为：6×6。

4.2 视觉比较

本文使用不同加速因子的1D-VRDU采样模式对4个数据集分别进行实验。图1为4种算法在加速因子 (acceleration factor, AF) 为5的1D-VRDU采样模式下，在数据集dataset 1的重建图像及误差图。由图1(a)可知，SAKE算法和PLORAKS算法的重建图像存在明显的混叠伪影，MS-HTC算法较大程度的改善了混叠伪影，但重建的图像仍然存在模糊伪影，FMS-JTLHTC算法很好的去除了残留的模糊伪影。由图1(b)可知，MS-HTC算法相对于SAKE算法和PLORAKS算法白点明显减少，FMS-JTLHTC算法中的白点进一步减少，误差更小，说明重建图像质量更高。

由于图像的低频信息主要集中在K空间的中心区域，使用1D-VRDU采样模式在高度欠采样的情况下会导致重建出的MRI图像产生大量的混叠伪影，而使用2D-VRDU采样模式能够采集到包含更多低频信息和较少高频信息的K空间数据，这样能够在高度欠采样的情况下重建出质量良好的图像[25]。所以实验中使用不同加速因子的2D-VRDU采样模式在4个数据集上分别进行实验。图2为4种算法在AF=10的2D-VRDU采样模式下，在数据集dataset 2的重建图像及误差图。由图可知，FMS-JTLHTC算法相对于SAKE算法、PLORAKS算法及MS-HTC算法有效的去除了模糊伪影，从误差图中也能看到FMS-JTLHTC算法白点整体减少，误差更小。

为进一步验证本文算法的适用性，使用4种算法对AF=3的一维泊松 (1D-Poisson) 欠采样模式下的数据集dataset 1进行重建实验。重建图像和误差图如图3所示，从重建图像中能够清晰看出SAKE算法和PLORAKS算法重建出图像中心存在明显的混叠伪影，MS-HTC算法重建出的图像很好的消除了混叠伪影，但中心部分还存在模糊伪影。本文FMS-JTLHTC算法很好的抑制了残留的模糊伪影，图像更清晰，从误差图中也能看出FMS-JTLHTC算法相对于MS-HTC算法中心部分的白点明显减少，误差更小。

4.3 评价指标分析

为客观评价本文算法的重建性能，使用PSNR、MSSIM、FSIM及HFEN这4个指标对4种算法进行比较。表1和表2分别为4种算法在不同加速因子的1D-VRDU和2D-VRDU采样模式下4个数据集的PSNR、MSSIM、FSIM及HFEN的平均值，最优结果加粗显示。由表1可知，在1D-VRDU采样模式下，FMS-JTLHTC算法相对于SAKE算法、PLORAKS算法及MS-HTC算法在PSNR值上平均提高了5.84 dB、5.24 dB、2.60 dB。由表2可知，在2D-VRDU采样模式下，FMS-JTLHTC算法相对于SAKE算法、PLORAKS算法及MS-HTC算法在PSNR值上平均提高了2.22 dB、2.10 dB、1.54 dB。表1和

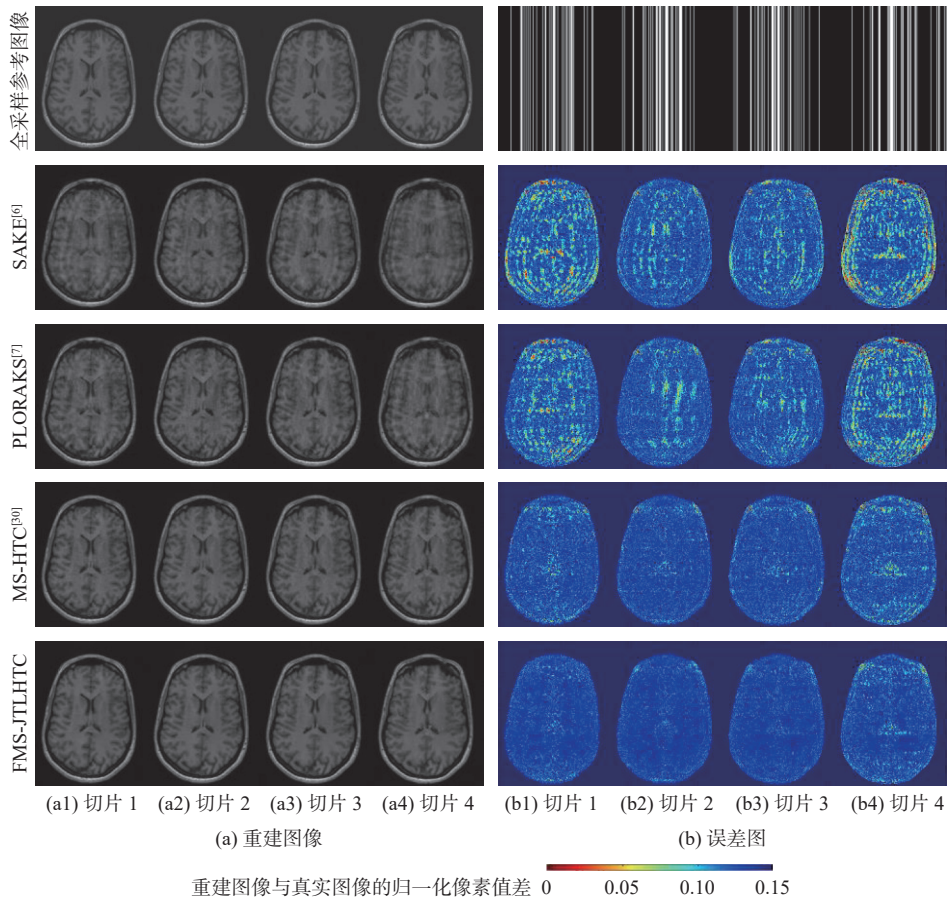


图 1　4 种算法对 AF=5 的 1D-VRDU 采样模式下的 dataset 1 重建的图像和误差图

Fig. 1　Images and error maps of dataset 1 reconstruction by four algorithms for 1D-VRDU sampling pattern with AF=5

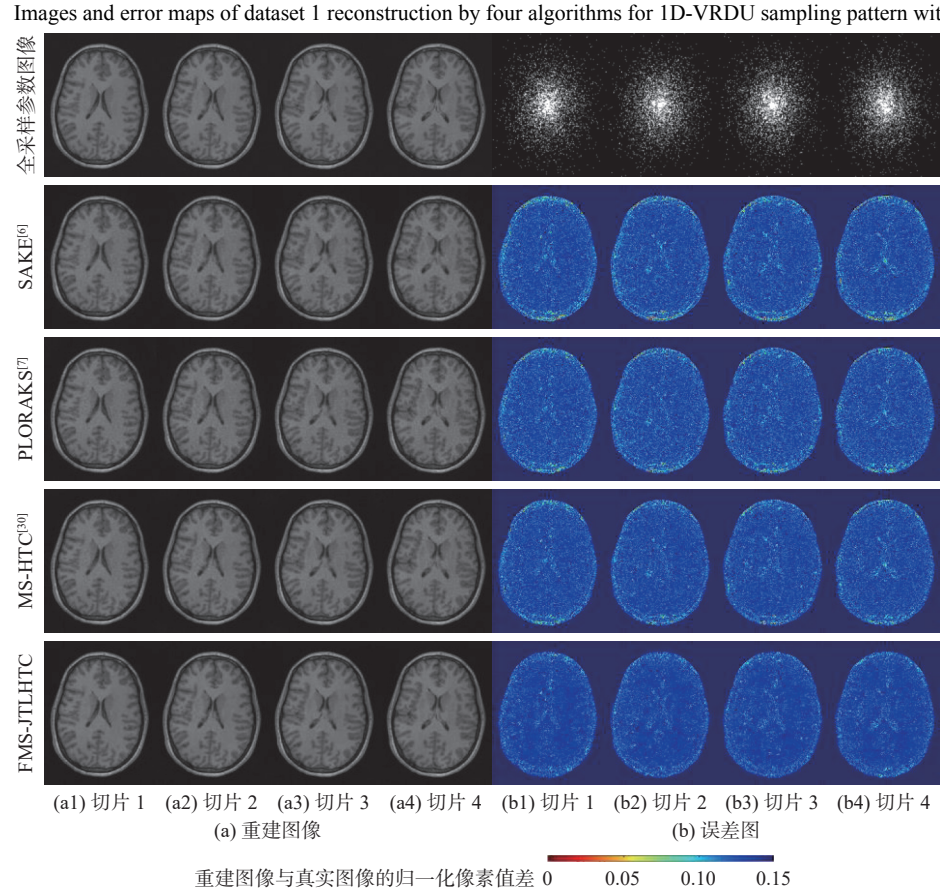


图 2　4 种算法对 AF=10 的 2D-VRDU 采样模式下 dataset 2 重建的图像和误差图

Fig. 2　Images and error maps of dataset 2 reconstruction by four algorithms for 2D-VRDU sampling pattern with AF=10

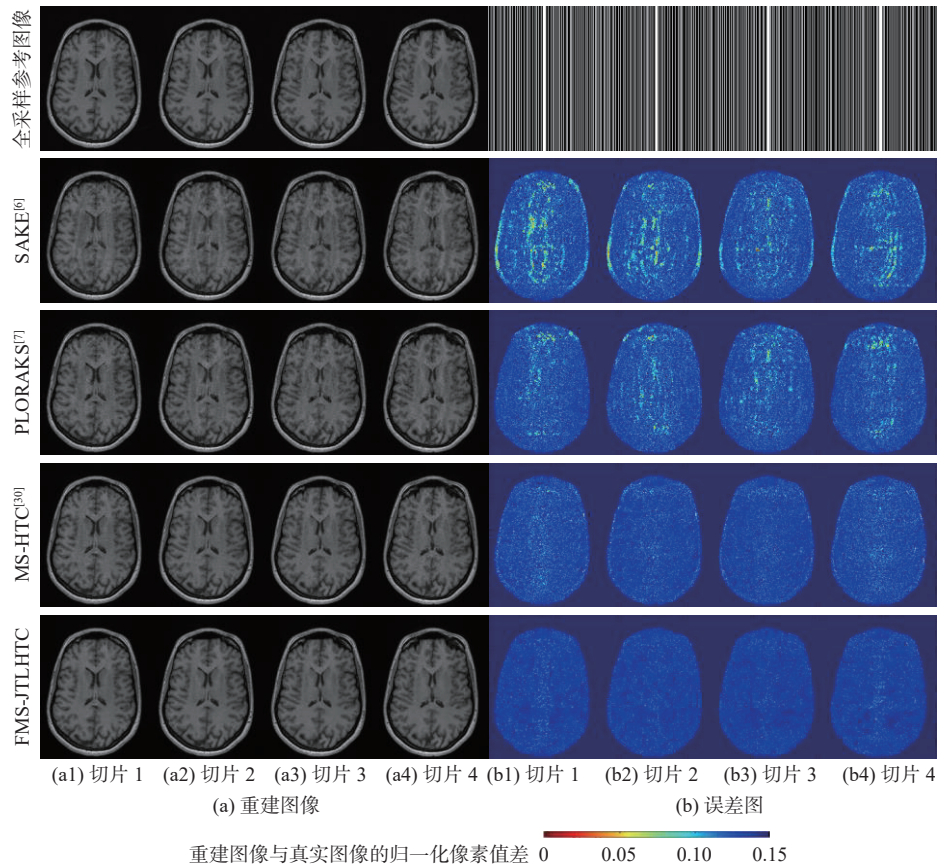


图 3 4 种算法对 AF=3 的 1D-Poisson 欠采样模式下 datase 1 重建的图像和误差图

Fig. 3 Images and error maps of dataset 1 reconstruction by four algorithms for 1D-Poisson undersampling mode with AF=3

表 1 4 种算法对不同加速因子的 1D-VRDU 采样模式下 4 个数据集重建的评价指标平均值

Table 1 Mean values of evalution indicators reconstructed by four algorithms for four datasets in 1D-VRDU sampling pattern with different acceleration factors

算法	AF	PSNR/dB	MSSIM	FSIM	HFEN
SAKE[6]	3	32.86	0.907 0	0.996 1	0.207 0
	5	28.55	0.837 0	0.991 2	0.312 6
	6	27.66	0.823 8	0.988 9	0.343 7
	7	27.29	0.806 2	0.987 5	0.368 0
PLORAKS[7]	3	33.85	0.915 1	0.997 3	0.181 4
	5	29.27	0.848 9	0.992 9	0.289 2
	6	28.11	0.830 4	0.990 6	0.328 2
	7	27.53	0.815 4	0.988 3	0.365 9
MS-HTC[30]	3	35.19	0.931 7	0.998 1	0.136 5
	5	31.82	0.884 4	0.995 4	0.208 3
	6	31.32	0.880 5	0.994 9	0.226 5
	7	30.98	0.871 1	0.993 7	0.241 6
FMS-JTLHTC	3	**38.01**	**0.964 6**	**0.999 0**	**0.099 7**
	5	**34.59**	**0.940 1**	**0.998 1**	**0.150 4**
	6	**33.71**	**0.931 8**	**0.997 6**	**0.169 8**
	7	**33.39**	**0.926 7**	**0.997 3**	**0.183 0**

表 2 4 种算法对不同加速因子的 2D-VRDU 采样模式下 4 个数据集重建的评价指标平均值

Table 2 Mean values of evalution indicators reconstructed by four algorithms for four datasets in 2D-VRDU sampling pattern with different acceleration factors

算法	AF	PSNR /dB	MSSIM	FSIM	HFEN
SAKE[6]	3	37.89	0.957 1	0.999 6	0.066 4
	5	35.36	0.932 6	0.999 0	0.098 0
	8	33.48	0.914 8	0.998 4	0.136 1
	10	32.67	0.906 8	0.997 9	0.160 9
	12	32.09	0.897 7	0.997 4	0.179 7
PLORAKS[7]	3	37.86	0.953 7	0.999 6	0.065 7
	5	35.55	0.935 4	0.999 1	0.096 6
	8	33.65	0.916 8	0.998 4	0.135 4
	10	32.83	0.905 9	0.997 9	0.154 9
	12	32.24	0.898 7	0.997 5	0.173 0
MS-HTC[30]	3	38.55	0.962 6	0.999 7	0.059 1
	5	35.98	0.943 5	0.999 2	0.085 0
	8	34.21	0.926 3	0.998 7	0.114 4
	10	33.26	0.913 6	0.998 1	0.134 3
	12	32.92	0.911 5	0.997 9	0.145 8
FMS-JTLHTC	3	**39.99**	**0.973 8**	**0.999 7**	**0.052 9**
	5	**37.58**	**0.961 9**	**0.999 5**	**0.072 2**
	8	**35.76**	**0.949 7**	**0.999 2**	**0.095 1**
	10	**34.94**	**0.943 5**	**0.998 9**	**0.109 4**
	12	**34.34**	**0.938 3**	**0.998 8**	**0.120 3**

表 2 评价指标分析所得到的结论与图 1 和图 2 视觉分析结论一致。综合来说，FMS-JTLHTC 算法的

PSNR 相较于 SAKE 算法、PLORAKS 算法和 MS-HTC 算法分别平均提高了 4.04 dB、3.67 dB 和 2.07 dB。

此外, 表 3 为在 AF=3 的 1D-Possion 欠采样模式下, 4 种算法对数据集 dataset 1 进行重建的 PSNR、MSSIM、FISM、HFEN 的值, 最优结果加粗显示。由表可知, FMS-JTLHTC 算法在其他采样模式下仍然具有良好的重建质量。

图 4 为 4 个数据集在不同加速因子的 1D-VRDU 和 2D-VRDU 采样模式下, FMS-JTLHTC 算法与 SAKE 算法、PLORAKS 算法及 MS-HTC 算法重建图像的 PSNR 差值。从箱线图中可以明显的看出 FMS-JTLHTC 算法的性能相对 SAKE 算法、PLORAKS 算法及 MS-HTC 算法有较大的提升。

表 3　4 种算法对 AF=3 的 1D-Possion 欠采样模式下数据集 dataset 1 重建的评价指标值

Table 3　Values of evalution indicators for four algorithms for reconstruction of dataset dataset 1 in 1D-Possion undersampling pattern with AF=3

算法	PSNR/dB	MSSIM	FSIM	HFEN
SAKE[6]	31.73	0.890 4	0.992 2	0.211 9
PLORAKS[7]	33.05	0.907 6	0.996 1	0.179 5
MS-HTC[30]	35.19	0.932 2	0.997 8	0.141 1
FMS-JTLHTC	**38.54**	**0.967 8**	**0.999 1**	**0.105 9**

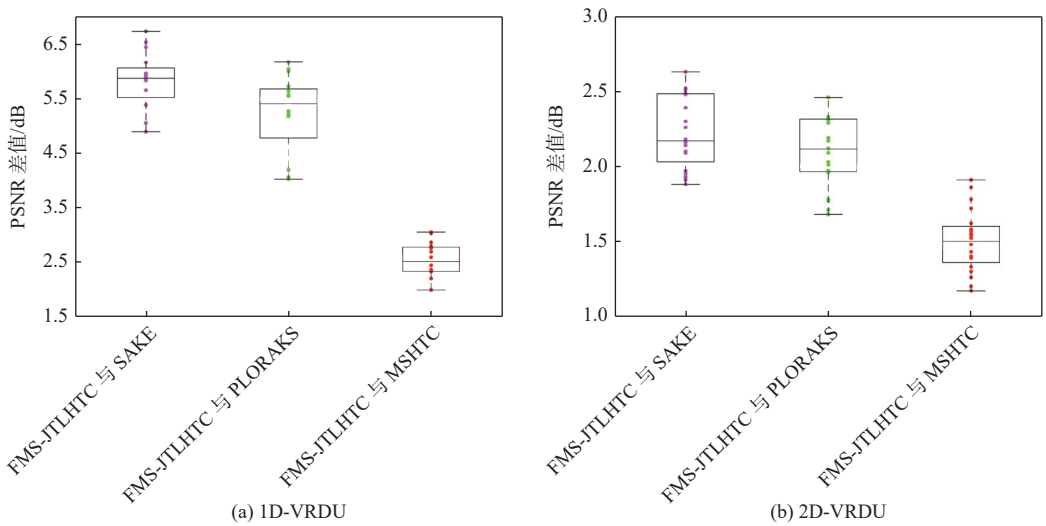


图 4　FMS-JTLHTC 算法与其他算法对不同加速因子的 2 种采样模式下 4 个数据集重建图像的 PSNR 的差值箱线图

Fig. 4　Boxplot of difference in PSNR between FMS-JTLHTC algorithm and other algorithms for reconstructed images of four datasets with different acceleration factors for two sampling patterns

4.4　收敛性分析

通过 PSNR、HFEN 及 RE 随迭代次数 K 的变化来分析 FMS-JTLHTC 算法的收敛性。图 5 分别为数据集 dataset 3 在 AF=5 的 1D-VRDU 和 AF=10 的 2D-VRDU 欠采样模式下的收敛性。由图 5 可知, 当 $V_{RE} < 10^{-3}$时, PSNR 及 HFEN 的数值基本不再变化, 由此可以判定 FMS-JTLHTC 算法此时已经达到收敛, 可以停止算法迭代, 减少重建时间, 图中, “*” 表示算法停止时的迭代次数。

4.5　重建时间分析

为解决算法计算复杂度高的问题, FMS-JTLHTC 算法利用 GPU 进行加速, rawFMS-JTLHTC 表示未

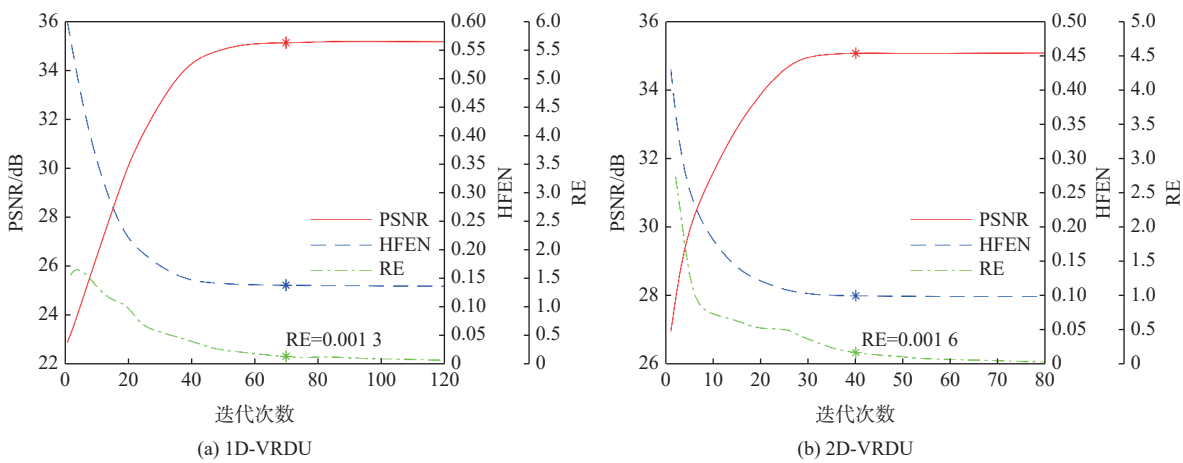


图 5　数据集 dataset 3 在 2 种采样模式下进行重建的收敛性分析

Fig. 5　Convergence analysis of dataset dataset 3 reconstructed in two sampling patterns

使用GPU进行加速的算法。图6分别为FMS-JTLHTC、rawFMS-JTLHTC和MS-HTC算法在AF=10的2D-VRDU采样模式下，4个数据集进行重建所需的时间和重建质量，图6中用“*”标记的位置为算法的停止时间。由图6可知，FMS-JTLHTC算法相较于MS-HTC算法不但在重建质量有较为明显的提升，而且在重建速度上提升了10倍以上。FMS-JTLHTC算法相较于rawFMS-JTLHTC算法在重建速度上提高了26倍。

表4和表5中分别为5种算法对AF=5的1D-VRDU采样模式和AF=10的2D-VRDU采样模式下，4个数据集的重建时间及FMS-JTLHTC算法相较于其他几种算法在重建时间上的提升倍数，表中，最快的计算时间加粗显示。由表可知，FMS-JTLHTC算法能够在20 s左右的时间内重建出4个切片图像，在实际应用中更具有优势。

4.6 问题分析

ADMM算法将原始的优化问题分解成易于求解的较小子问题，然后，交替求解这些子问题，并不断迭代以逼近最优解。这使得ADMM算法的性能在很大程度上取决于算法中的参数选择，如罚参数的选择。选择合适的罚参数能够有效协调ADMM

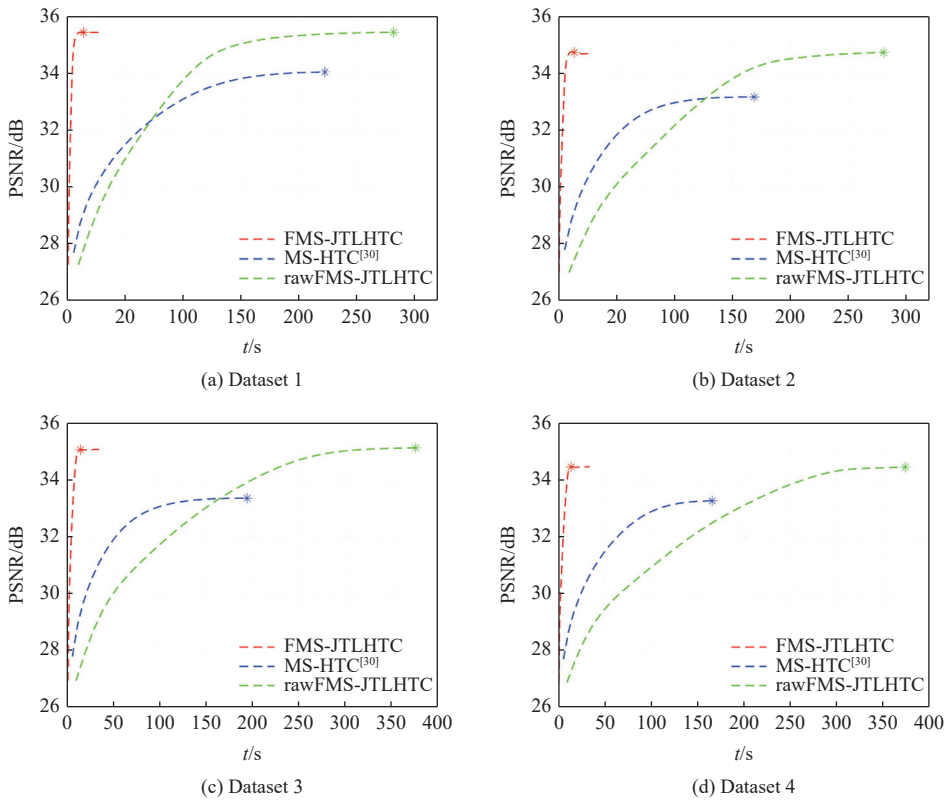


图6 FMS-JTLHTC、rawFMS-JTLHTC和MS-HTC算法对4个数据集进行重建的时间比较

Fig. 6 Comparison of reconstruction times of FMS-JTLHTC, rawFMS-JTLHTC and MS-HTC algorithms for four datasets

表4 5种算法对AF=5的1D-VRDU采样模式下4个数据集的重建时间与速度

Table 4 Reconstruction times and velocities of five algorithms for four datasets in 1D-VRDU sampling mode with AF=5

算法	t/s				速度提升倍数			
	Dataset 1	Dataset 2	Dataset 3	Dataset 4	Dataset 1	Dataset 2	Dataset 3	Dataset 4
SAKE[6]	779.35	779.50	779.39	779.17	44.82	36.79	32.14	23.62
PLORAKS[7]	570.30	517.17	578.35	777.84	32.79	24.41	23.85	23.58
MS-HTC[30]	272.46	271.92	271.43	275.41	15.67	12.83	11.19	8.35
rawFMS-JTLHTC	456.81	550.35	624.76	860.09	26.27	25.97	25.76	26.07
FMS-JTLHTC	**17.39**	**21.19**	**24.25**	**32.99**	1.00	1.00	1.00	1.00

表5 5种算法对AF=10的2D-VRDU采样模式下四个数据集的重建时间与速度

Table 5 Reconstruction times and velocities of the five algorithms for four datasets in 2D-VRDU sampling mode with AF=10

算法	t/s				速度提升倍数			
	Dataset 1	Dataset 2	Dataset 3	Dataset 4	Dataset 1	Dataset 2	Dataset 3	Dataset 4
SAKE[6]	311.95	207.80	259.67	207.94	29.10	19.37	18.15	14.21
PLORAKS[7]	99.62	96.39	98.85	398.66	9.29	8.98	6.91	27.25
MS-HTC[30]	222.17	168.68	193.70	165.82	20.72	15.72	13.54	11.33
rawFMS-JTLHTC	282.55	281.26	376.39	374.50	26.32	26.21	26.30	25.60
FMS-JTLHTC	**10.72**	**10.73**	**14.31**	**14.63**	1.00	1.00	1.00	1.00

算法分解的子问题的解, 确保子问题的解保持一致, 从而得到最优解。

文中使用 FMS-JTLHTC 算法对 5 倍加速的 1D-VRDU 采样模式下 dataset 1 重建时参数α和μ_1的变化对重构图像 PSNR 值的影响进行分析, 如图 7 所示。由图可知, 当$\alpha \in [3\times10^{-4}, 2\times10^{-3}]$和$\mu_1 \in [5\times10^{-6}, 3\times10^{-5}]$时, 重构图像的 PSNR 值均能保证在 34.5 dB 以上, 而当参数α和μ_1的值超出这个范围, 重构图像的 PSNR 值会大幅下降。

另外, ADMM 算法求解优化问题可能会导致算法出现收敛较慢的问题。针对这个问题, 在求解过程中引入 FISTA 技术来减少迭代次数从而加快收敛速度。图 8 显示了有无 FISTA 技术对本文算法迭代次数的影响, MS-JTLHTC 算法表示未加 FISTA 技术, FMS-JTLHTC 表示已加 FISTA 技术。其中, *表示算法停止时的迭代次数。由图 8 可知: 加了 FISTA 技术的算法比没有加 FISTA 技术的算法收敛速度提高了 5 倍左右。

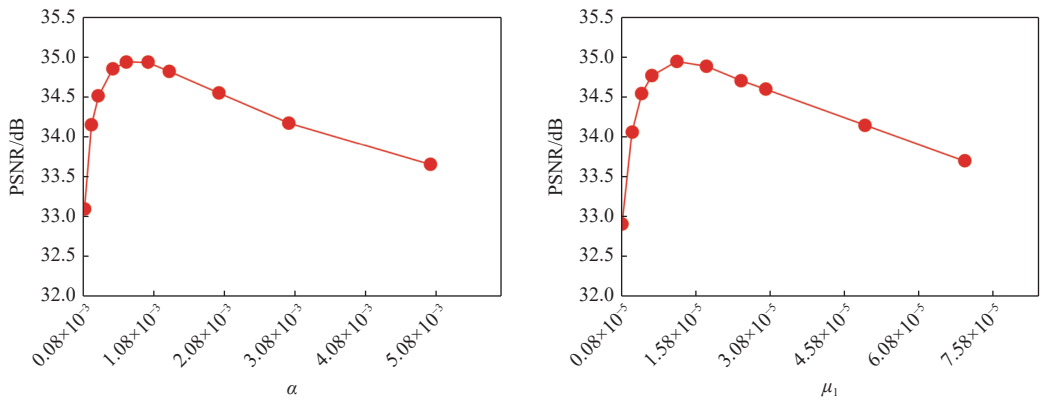


图 7　FMS-JTLHTC 算法对 5 倍加速的 1D-VRDU 采样模式下 dataset 1 重建时参数α和μ_1的变化对重构图像 PSNR 值的影响

Fig. 7　Effect of variation of parameters α and μ_1 on PSNR values of reconstructed images during dataset 1 reconstruction by FMS-JTLHTC algorithm for 1D-VRDU sampling mode with 5-fold acceleration

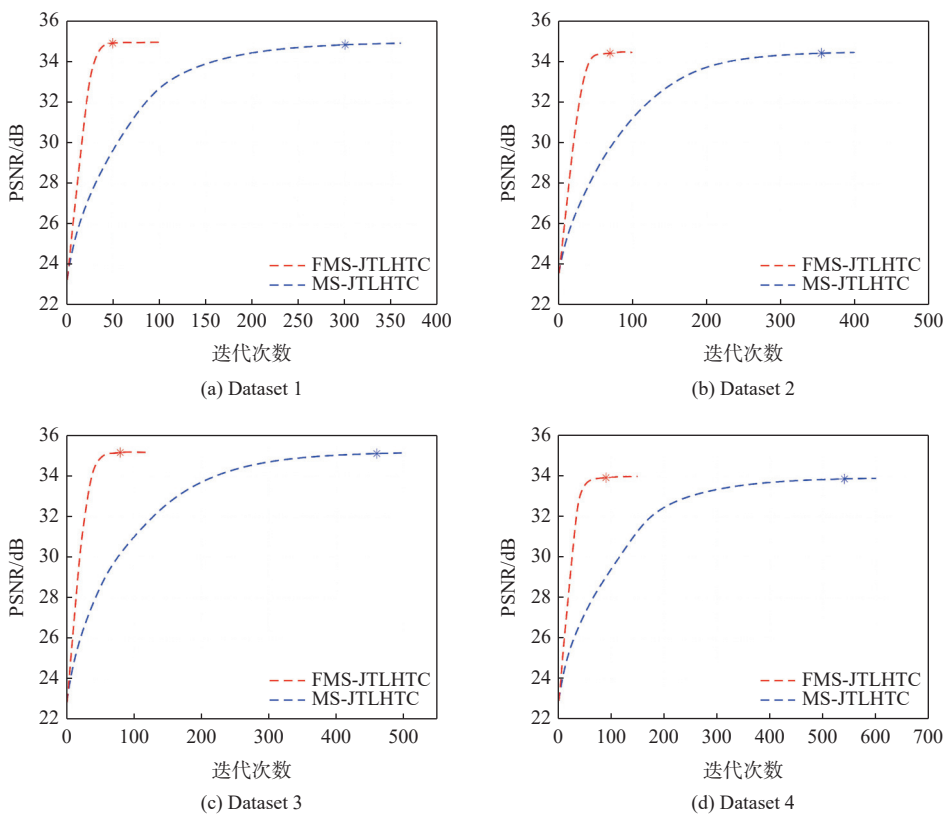


图 8　FMS-JTLHTC 和 MS-JTLHTC 算法对 5 倍加速的 1D-VRDU 采样模式下 4 个数据集重建时 PSNR 值随迭代次数的变化

Fig. 8　Variation of PSNR values with number of iterations for four datasets reconstructed by FMS-JTLHTC and MS-JTLHTC algorithm with 5-fold acceleration of 1D-VRDU sampling mode

5 结 论

1）FMS-JTLHTC算法相较于SAKE算法、PLORAKS算法及MSHTC算法在重建质量上有较大的提升。

2）FMS-JTLHTC算法中FISTA技术有效解决了由ADMM算法导致的收敛速度慢的问题，相对于没有使用FISTA技术的MS-JTLHTC算法在收敛速度上提高了5倍左右。

3）FMS-JTLHTC算法中GPU的利用弥补了张量分解、K空间数据和分块Hankel张量之间的相互转换及稀疏变换学习所带来的计算复杂度高的问题，在重建速度上比MS-HTC算法提高了10倍以上，比未利用GPU的rawFMS-JTLHTC算法提高了26倍以上。

相较于笛卡儿采样，非笛卡儿采样能有效使用MRI梯度硬件快速覆盖K空间，并具有较高的K空间覆盖率、较少的欠采样相干伪影、受运动影响较小、具有自导航等特点。除此之外，FMS-JTLHTC算法的性能取决于模式展开的秩大小、内核大小、正则项参数，目前这些参数根据经验确定，而自动参数选择在MRI应用中更具有实际意义。未来将对非笛卡儿数据应用到本文的算法中和算法自动参数选择的可行性进行研究。

参考文献（References）

[1] PRUESSMANN K P, WEIGER M, SCHEIDEGGER M B, et al. SENSE: sensitivity encoding for fast MRI[J]. Magnetic Resonance in Medicine, 1999, 42(5): 952-962.

[2] SODICKSON D K, MANNING W J. Simultaneous acquisition of spatial harmonics (SMASH): fast imaging with radiofrequency coil arrays[J]. Magnetic Resonance in Medicine, 1997, 38(4): 591-603.

[3] UECKER M, LAI P, MURPHY M J, et al. ESPIRiT: an eigenvalue approach to autocalibrating parallel MRI: where SENSE meets GRAPPA[J]. Magnetic Resonance in Medicine, 2014, 71(3): 990-1001.

[4] GRISWOLD M A, JAKOB P M, HEIDEMANN R M, et al. Generalized autocalibrating partially parallel acquisitions (GRAPPA)[J]. Magnetic Resonance in Medicine, 2002, 47(6): 1202-1210.

[5] LUSTIG M, PAULY J M. SPIRiT: iterative self-consistent parallel imaging reconstruction from arbitrary k-space[J]. Magnetic Resonance in Medicine, 2010, 64(2): 457-471.

[6] SHIN P J, LARSON P E Z, OHLIGER M A, et al. Calibrationless parallel imaging reconstruction based on structured low-rank matrix completion[J]. Magnetic Resonance in Medicine, 2014, 72(4): 959-970.

[7] HALDAR J P, ZHUO J W. P-LORAKS: low-rank modeling of local K-space neighborhoods with parallel imaging data[J]. Magnetic Resonance in Medicine, 2016, 75(4): 1499-1514.

[8] DONOHO D L. Compressed sensing[J]. IEEE Transactions on Information Theory, 2006, 52(4): 1289-1306.

[9] LUSTIG M, DONOHO D, PAULY J M. Sparse MRI: the application of compressed sensing for rapid MR imaging[J]. Magnetic Resonance in Medicine, 2007, 58(6): 1182-1195.

[10] LAI Z Y, QU X B, LIU Y S, et al. Image reconstruction of compressed sensing MRI using graph-based redundant wavelet transform[J]. Medical Image Analysis, 2016, 27: 93-104.

[11] YANG J F, ZHANG Y, YIN W T. A fast alternating direction method for TVL1-L2 signal reconstruction from partial Fourier data[J]. IEEE Journal of Selected Topics in Signal Processing, 2010, 4(2): 288-297.

[12] RAVISHANKAR S, BRESLER Y. MR image reconstruction from highly undersampled K-space data by dictionary learning[J]. IEEE Transactions on Medical Imaging, 2011, 30(5): 1028-1041.

[13] AHARON M, ELAD M, BRUCKSTEIN A. K-SVD: an algorithm for designing overcomplete dictionaries for sparse representation[J]. IEEE Transactions on Signal Processing, 2006, 54(11): 4311-4322.

[14] RAVISHANKAR S, BRESLER Y. Learning sparsifying transforms[J]. IEEE Transactions on Signal Processing, 2013, 61(5): 1072-1086.

[15] RAVISHANKAR S, BRESLER Y. Efficient blind compressed sensing using sparsifying transforms with convergence guarantees and application to magnetic resonance imaging[J]. SIAM Journal on Imaging Sciences, 2015, 8(4): 2519-2557.

[16] 李玺兰, 段继忠. 基于稀疏变换学习的改进灵敏度编码重建算法[J]. 北京邮电大学学报, 2022, 45(5): 97-102.
LI X L, DUAN J Z. An improved sensitivity encoding reconstruction algorithm based on sparse transform learning[J]. Journal of Beijing University of Posts and Telecommunications, 2022, 45(5): 97-102(in Chinese).

[17] 段继忠, 钱青青. 基于SIDWT和迭代自一致性的快速并行成像重建方法[J]. 上海交通大学学报, 2023, 57(5): 582-592.
DUAN J Z, QIAN Q Q. Fast parallel imaging reconstruction method based on SIDWT and iterative self-consistency[J]. Journal of Shanghai Jiao Tong University, 2023, 57(5): 582-592(in Chinese).

[18] DUAN J Z, PAN T, LIU Y, et al. Non-local low-rank constraint-based self-consistent PMRI reconstruction using eigenvector maps[J]. IET Signal Processing, 2023, 17(3): e12180.

[19] PAN T, DUAN J Z, WANG J F, et al. Iterative self-consistent parallel magnetic resonance imaging reconstruction based on nonlocal low-rank regularization[J]. Magnetic Resonance Imaging, 2022, 88: 62-75.

[20] DUAN J Z, LIU C, LIU Y, et al. Adaptive transform learning and joint sparsity based PLORAKS parallel magnetic resonance image reconstruction[J]. IEEE Access, 2020, 8: 212315-212326.

[21] LIN C Y, FESSLER J A. Efficient dynamic parallel MRI reconstruction for the low-rank plus sparse model[J]. IEEE Transactions on Computational Imaging, 2019, 5(1): 17-26.

[22] OTAZO R, CANDÈS E, SODICKSON D K. Low-rank plus sparse matrix decomposition for accelerated dynamic MRI with separation of background and dynamic components[J]. Magnetic Resonance in Medicine, 2015, 73(3): 1125-1136.

[23] PANG Y, ZHANG X L. Interpolated compressed sensing for 2D multiple slice fast MR imaging[J]. PLoS One, 2013, 8(2): e56098.

[24] DATTA S, DEKA B. Magnetic resonance image reconstruction using fast interpolated compressed sensing[J]. Journal of Optics, 2018, 47(2): 154-165.

[25] MURAD M, BILAL M, JALIL A, et al. Efficient reconstruction technique for multi-slice CS-MRI using novel interpolation and 2D sampling scheme[J]. IEEE Access, 2020, 8: 117452-117466.

[26] DATTA S, DEKA B. Efficient interpolated compressed sensing reconstruction scheme for 3D MRI[J]. IET Image Processing, 2018, 12(11): 2119-2127.

[27] DATTA S, DEKA B, MULLAH H U, et al. An efficient interpolated compressed sensing method for highly correlated 2D multi-slice MRI[C]//Proceedings of the International Conference on Accessibility to Digital World. Piscataway: IEEE Press, 2016: 187-192.

[28] MURAD M, JALIL A, BILAL M, et al. Radial undersampling-based interpolation scheme for multislice CSMRI reconstruction techniques[J]. BioMed Research International, 2021, 2021: 6638588.

[29] DEKA B, DATTA S. Calibrationless joint compressed sensing reconstruction for rapid parallel MRI[J]. Biomedical Signal Processing and Control, 2020, 58: 101871.

[30] LIU Y L, YI Z Y, ZHAO Y J, et al. Calibrationless parallel imaging reconstruction for multislice MR data using low-rank tensor completion[J]. Magnetic Resonance in Medicine, 2021, 85(2): 897-911.

[31] BECK A, TEBOULLE M. A fast iterative shrinkage-thresholding algorithm for linear inverse problems[J]. SIAM Journal on Imaging Sciences, 2009, 2(1): 183-202.

[32] LIU Y P, LIU T T, LIU J N, et al. Smooth robust tensor principal component analysis for compressed sensing of dynamic MRI[J]. Pattern Recognition, 2020, 102: 107252.

[33] LIU D, ZHOU J J, MENG M M, et al. Highly undersampling dynamic cardiac MRI based on low-rank tensor coding[J]. Magnetic Resonance Imaging, 2022, 89: 12-23.

[34] HE J F, LIU X T, LUAN N N. Bi-smooth constraints for accelerated dynamic MRI with low-rank plus sparse tensor decomposition[J]. Biomedical Signal Processing and Control, 2023, 82: 104530.

[35] ZHANG Y H, HU Y. Dynamic cardiac MRI reconstruction using combined tensor nuclear norm and casorati matrix nuclear norm regularizations[C]//Proceedings of the IEEE 19th International Symposium on Biomedical Imaging. Piscataway: IEEE Press, 2022: 1-4.

[36] WU T L, GAO B, FAN J C, et al. Low-rank tensor completion based on self-adaptive learnable transforms[J]. IEEE Transactions on Neural Networks and Learning Systems, 2024, 35(7): 8826-8838.

[37] BERGQVIST G, LARSSON E G. The higher-order singular value decomposition: theory and an application [lecture notes][J]. IEEE Signal Processing Magazine, 2010, 27(3): 151-154.

[38] TUCKER L R. Some mathematical notes on three-mode factor analysis[J]. Psychometrika, 1966, 31(3): 279-311.

[39] DE LATHAUWER L, DE MOOR B, VANDEWALLE J. A multilinear singular value decomposition[J]. SIAM Journal on Matrix Analysis and Applications, 2000, 21(4): 1253-1278.

[40] LIU Y P, CHEN L X, ZHU C. Improved robust tensor principal component analysis via low-rank core matrix[J]. IEEE Journal of Selected Topics in Signal Processing, 2018, 12(6): 1378-1389.

[41] VANNIEUWENHOVEN N, VANDEBRIL R, MEERBERGEN K. A new truncation strategy for the higher-order singular value decomposition[J]. SIAM Journal on Scientific Computing, 2012, 34(2): A1027-A1052.

[42] DUAN J Z, LIU Y, WANG J F. Accelerated SPIRiT parallel MR image reconstruction based on joint sparsity and sparsifying transform learning[J]. IEEE Transactions on Computational Imaging, 2023, 9: 276-288.

[43] GILLARD J. Cadzow's basic algorithm, alternating projections and singular spectrum analysis[J]. Statistics and Its Interface, 2010, 3(3): 335-343.

[44] SOUZA R, BENTO M, NOGOVITSYN N, et al. Dual-domain cascade of U-nets for multi-channel magnetic resonance image reconstruction[J]. Magnetic Resonance Imaging, 2020, 71: 140-153.

[45] CORRIVEAU P, WEBSTER A. VQEG evaluation of objective methods of video quality assessment[J]. SMPTE Journal, 1999, 108(9): 645-648.

[46] WANG Z, BOVIK A C, SHEIKH H R, et al. Image quality assessment: from error visibility to structural similarity[J]. IEEE Transactions on Image Processing, 2004, 13(4): 600-612.

[47] ZHANG L, ZHANG L, MOU X Q, et al. FSIM: a feature similarity index for image quality assessment[J]. IEEE Transactions on Image Processing, 2011, 20(8): 2378-2386.

附录 A:

对于问题 (3), 使用 ADMM 算法对其进行求解, 引入辅助变量 $\boldsymbol{B}_{sj}=\boldsymbol{W}_s\boldsymbol{P}_j\boldsymbol{E}_s(\boldsymbol{F}^{-1}\boldsymbol{f})\in\mathbf{C}^{m\times N_c}$ 及对应的拉格朗日乘子 $\boldsymbol{d}_{\boldsymbol{B}_{sj}}\in\mathbf{C}^{m\times N_c}$, 则其增广拉格朗日函数定义如下:

$$\begin{cases} L(\boldsymbol{f},\boldsymbol{W}_s,\boldsymbol{\Gamma},\boldsymbol{B}_{sj};\boldsymbol{d}_{\boldsymbol{B}_{sj}})=\|\boldsymbol{D}\boldsymbol{f}-\boldsymbol{y}\|_2^2+\alpha\|\boldsymbol{\Gamma}-\boldsymbol{R}(\boldsymbol{f})\|_{\mathrm{F}}^2+ \\ \quad \sum_{s=1}^{N_{\mathrm{S}}}\sum_{j=1}^{N_{\mathrm{P}}}\left\{\beta\left\|\boldsymbol{B}_{sj}\right\|_{0,2}+\frac{\mu_1}{2}\left\|\boldsymbol{B}_{sj}-\boldsymbol{W}_s\boldsymbol{P}_j\boldsymbol{E}_s(\boldsymbol{F}^{-1}\boldsymbol{f})\right\|_{\mathrm{F}}^2-\right. \\ \quad \left.\mu_1\left\langle\boldsymbol{d}_{\boldsymbol{B}_{sj}},\boldsymbol{B}_{sj}-\boldsymbol{W}_s\boldsymbol{P}_j\boldsymbol{E}_s(\boldsymbol{F}^{-1}\boldsymbol{f})\right\rangle\right\} \\ \quad \text{s.t. } \boldsymbol{\Gamma}=\boldsymbol{Q}\times_1\boldsymbol{U}^{(1)}\times_2\boldsymbol{U}^{(2)}\times_3\boldsymbol{U}^{(3)},\ \boldsymbol{W}_s^{\mathrm{H}}\boldsymbol{W}_s=\boldsymbol{I} \end{cases} \tag{A1}$$

则可以得到以下等式:

$$\begin{cases} L(\boldsymbol{f},\boldsymbol{W}_s,\boldsymbol{\Gamma},\boldsymbol{B}_{sj};\boldsymbol{d}_{\boldsymbol{B}_{sj}})=\|\boldsymbol{D}\boldsymbol{f}-\boldsymbol{y}\|_2^2+ \\ \quad \alpha\|\boldsymbol{\Gamma}-\boldsymbol{R}(\boldsymbol{f})\|_F^2+\sum_{s=1}^{N_{\mathrm{S}}}\sum_{j=1}^{N_{\mathrm{P}}}\left\{\beta\left\|\boldsymbol{B}_{sj}\right\|_{0,2}+\right. \\ \quad \left.\frac{\mu_1}{2}\left\|\boldsymbol{B}_{sj}-\boldsymbol{W}_s\boldsymbol{P}_j\boldsymbol{E}_s\left(\boldsymbol{F}^{-1}\boldsymbol{f}\right)-\boldsymbol{d}_{\boldsymbol{B}_{sj}}\right\|_{\mathrm{F}}^2\right\} \\ \quad \text{s.t. } \boldsymbol{\Gamma}=\boldsymbol{Q}\times_1\boldsymbol{U}^{(1)}\times_2\boldsymbol{U}^{(2)}\times_3\boldsymbol{U}^{(3)},\ \boldsymbol{W}_s^{\mathrm{H}}\boldsymbol{W}_s=\boldsymbol{I} \end{cases} \tag{A2}$$

由此可以得到问题 (A2) 中各变量的子问题:

$$\boldsymbol{d}_{\boldsymbol{B}_{sj}}^{k+1}=\boldsymbol{d}_{\boldsymbol{B}_{sj}}^{k}+\left[\boldsymbol{W}_s^{k+1}\boldsymbol{P}_j\boldsymbol{E}_s(\boldsymbol{F}^{-1}\boldsymbol{f}^{k+1})-\boldsymbol{B}_{sj}^{k+1}\right] \tag{A3}$$

求解变换学习 $\boldsymbol{W}_s$: 为方便求解计算, 令

$$\boldsymbol{B}_s^k=\left[\boldsymbol{B}_{s1}^k,\cdots,\boldsymbol{B}_{sj}^k,\cdots,\boldsymbol{B}_{sN_\mathrm{p}}^k\right]\in\mathbf{C}^{m\times N_\mathrm{C}N_\mathrm{p}}$$

$$\boldsymbol{d}_{\boldsymbol{B}_s}^k=\left[\boldsymbol{d}_{\boldsymbol{B}_{s1}}^k,\cdots,\boldsymbol{d}_{\boldsymbol{B}_{sj}}^k,\cdots,\boldsymbol{d}_{\boldsymbol{B}_{sN_\mathrm{p}}}^k\right]\in\mathbf{C}^{m\times N_\mathrm{C}N_\mathrm{p}}$$

$$\boldsymbol{A}_s^k=\left[\boldsymbol{P}_1\boldsymbol{E}_s(\boldsymbol{F}^{-1}\boldsymbol{f}^k),\cdots,\boldsymbol{P}_{N_\mathrm{p}}\boldsymbol{E}_s(\boldsymbol{F}^{-1}\boldsymbol{f}^k)\right]\in\mathbf{C}^{m\times N_\mathrm{C}N_\mathrm{p}}$$

$\boldsymbol{A}_s$为以向量块$\boldsymbol{P}_j\boldsymbol{E}_s(\boldsymbol{F}^{-1}\boldsymbol{f}^k)$为列的矩阵, 则问题式 (4) 可以写为

$$\begin{cases}\boldsymbol{W}_s^{k+1}=\arg\min\limits_{\boldsymbol{W}_s}\left\|\boldsymbol{W}_s\boldsymbol{A}_s^k-(\boldsymbol{B}_s^k-\boldsymbol{d}_{\boldsymbol{B}_s}^k)\right\|_\mathrm{F}^2\\ \quad\text{s.t. }\boldsymbol{W}_s^\mathrm{H}\boldsymbol{W}_s=\boldsymbol{I}\end{cases}\tag{A4}$$

问题 (A4) 通过奇异值分解进行求解[15], 即对$\boldsymbol{A}_s^k(\boldsymbol{B}_s^k-\boldsymbol{d}_{\boldsymbol{B}_s}^k)^\mathrm{H}$进行奇异值分解得到$\boldsymbol{A}_s^k(\boldsymbol{B}_s^k-\boldsymbol{d}_{\boldsymbol{B}_s}^k)^\mathrm{H}=\boldsymbol{U}_s\boldsymbol{\Sigma}_s\boldsymbol{V}_s^\mathrm{H}$, 则问题 (A4) 的解为式 (8)。

求解低秩张量$\boldsymbol{\Gamma}$: 使用顺序截断 HOSVD 对其进行求解[30]。其操作是: 顺序对$\boldsymbol{R}(\boldsymbol{f}^k)$进行 n 模式展开得到$\boldsymbol{\Gamma}_{(n)}=\mathrm{refold}(\boldsymbol{R}(\boldsymbol{f}^k),n)$, 然后对 n 模式展开矩阵$\boldsymbol{\Gamma}_{(n)}$进行奇异值分解(SVD)并对酉矩阵$\boldsymbol{U}^{(n)}$进行秩截断得到$\boldsymbol{\Gamma}_{(n)}=\hat{\boldsymbol{U}}^{(n)}\hat{\boldsymbol{\Sigma}}^{(n)}\hat{\boldsymbol{V}}^{(n)\mathrm{H}}$, 式中, $\hat{\boldsymbol{U}}^{(n)}$、$\hat{\boldsymbol{\Sigma}}^{(n)}$及$\hat{\boldsymbol{V}}^{(n)}$是对$\boldsymbol{R}\boldsymbol{f}^k$顺序进行$n$模式展开执行奇异值分解并对$\boldsymbol{U}^{(n)}$, $\boldsymbol{\Sigma}^{(n)}$和$\boldsymbol{V}^{(n)}$进行低秩估计和截断得到的。将计算得到的核心张量$\boldsymbol{Q}$和酉矩阵$\hat{\boldsymbol{U}}^{(1)},\hat{\boldsymbol{U}}^{(2)},\hat{\boldsymbol{U}}^{(3)}$执行多线性乘法[39], 生成最终的三阶近似低秩张量$\boldsymbol{\Gamma}^{k+1}$。则问题 (6) 中的低秩张量$\boldsymbol{\Gamma}^{k+1}$的解可将其表示为式 (11)。

更新图像$\boldsymbol{f}$, 为求解$\boldsymbol{f}$, 令问题 (7) 的目标函数的导数为 0, 可得

$$\begin{aligned}&\left(\mu_1\sum_{s=1}^{N_\mathrm{S}}\sum_{j=1}^{N_\mathrm{p}}\boldsymbol{F}\boldsymbol{E}_s^*\boldsymbol{P}_j{}^*(\boldsymbol{W}_s^{k+1})^\mathrm{H}\boldsymbol{W}_s^{k+1}\boldsymbol{P}_j\boldsymbol{E}_s\boldsymbol{F}^{-1}\right)\boldsymbol{f}^{k+1}+\\&\left(2\boldsymbol{D}^\mathrm{T}\boldsymbol{D}+2\alpha\boldsymbol{R}^*\boldsymbol{R}\right)\boldsymbol{f}^{k+1}=2\boldsymbol{D}^\mathrm{T}\boldsymbol{y}+2\alpha\boldsymbol{R}^*(\boldsymbol{\Gamma}^{k+1})+\\&\mu_1\sum_{s=1}^{N_\mathrm{S}}\sum_{j=1}^{N_\mathrm{p}}\boldsymbol{F}\boldsymbol{E}_s^*\left\{\boldsymbol{P}_j{}^*\left[(\boldsymbol{W}_s^{k+1})^\mathrm{H}(\boldsymbol{B}_{sj}^{k+1}-\boldsymbol{d}_{\boldsymbol{B}_{sj}}^k)\right]\right\}\end{aligned}\tag{A5}$$

式中: $\boldsymbol{R}^*$为$\boldsymbol{R}$的伴随算子, $\boldsymbol{R}^*\boldsymbol{R}$为对角矩阵; $\boldsymbol{P}_j{}^*$为$\boldsymbol{P}_j$的伴随算子, $\sum\limits_{j=1}^{N_\mathrm{p}}\boldsymbol{P}_j{}^*\boldsymbol{P}_j$为对角矩阵; $\boldsymbol{E}_s^*$为$\boldsymbol{E}_s$的伴随算子, $\sum\limits_{s=1}^{N_\mathrm{S}}\boldsymbol{E}_s{}^*\boldsymbol{E}_s$为对角矩阵, 且$\boldsymbol{W}_s^\mathrm{H}\boldsymbol{W}_s=\boldsymbol{I}$, 则有

$$\begin{aligned}&\sum_{s=1}^{N_\mathrm{S}}\sum_{j=1}^{N_\mathrm{p}}\boldsymbol{F}\boldsymbol{E}_s{}^*\boldsymbol{P}_j{}^*(\boldsymbol{W}_s^{k+1})^\mathrm{H}\boldsymbol{W}_s^{k+1}\boldsymbol{P}_j\boldsymbol{E}_s\boldsymbol{F}^{-1}=\\&\sum_{s=1}^{N_\mathrm{S}}\boldsymbol{F}\boldsymbol{E}_s{}^*\left(\sum_{j=1}^{N_\mathrm{p}}\boldsymbol{P}_j{}^*\boldsymbol{P}_j\right)\boldsymbol{E}_s\boldsymbol{F}^{-1}=\\&\boldsymbol{F}\left(\sum_{s=1}^{N_\mathrm{S}}\boldsymbol{E}_s{}^*\boldsymbol{E}_s\right)\boldsymbol{F}^{-1}=m\boldsymbol{I}\end{aligned}\tag{A6}$$

Fast multi-slice MRI reconstruction algorithm based on transform learning

DUAN Jizhong*, LIU Huan

(Faculty of Information Engineering and Automation, Kunming University of Science and Technology, Kunming 650504, China)

Abstract: Due to the significant correlation between neighboring slices in two-dimensional (2D) multi-slice magnetic resonance data, higher quality slice pictures can be reconstructed by taking use of the redundancy between slices. However, 2D multi-slice magnetic resonance imaging requires an amount of time. To improve the reconstruction quality and speed of 2D multi-slice (MRI) images, proposes a fast 2D multi-slice MRI reconstruction (FMS-JTLHTC) algorithm, which introduces the joint transform learning regular term into the multi-slice hankel tensor completion (MS-HTC) model. Prior to introducing the fast iterative shrinkage-thresholding procedure to accelerate convergence and utilize the graphics processing unit to speed up the procedure, the alternating direction method of multipliers is used to solve the objective issue. Experiments using four brain datasets in two different sampling modes show that the peak signal-to-noise ratio (PSNR) of the FMS-JTLHTC algorithm is improved by an average of 4.04 dB, 3.67 dB, and 2.07 dB compared to the simultaneous atuo-calibrating and K-space estimation (SAKE), low-rank modeling of local K-space neighborhoods with parallel imaging data (PLORAKS) and MS-HTC algorithms, respectively, the reconstruction speed is improved by a factor of 14 compared to the MS-HTC algorithm.

Keywords: multi-slice magnetic resonance imaging; Hankel tensor completion; joint sparse transform learning; alternating direction multiplier method; fast iterative shrinkage-thresholding algorithm

Received: 2023-08-31; **Accepted**: 2023-12-05; **Published Online**: 2024-03-28 18:26

URL: link.cnki.net/urlid/11.2625.V.20240327.1738.002

Foundation items: National Natural Science Foundation of China (61861023); Supported by Yunnan Fundamental Research Projects (202301AT070452)

* **Corresponding author**. E-mail: duanjz@kust.edu.cn

http://bhxb.buaa.edu.cn jbuaa@buaa.edu.cn
DOI: 10.13700/j.bh.1001-5965.2023.0345

结合上下文关联的图像情感分析

罗改芳[1,2]，张浩[1]，徐丹[1,*]

(1. 云南大学 信息学院，昆明 650500；2. 山西农业大学 软件学院，晋中 030801)

摘 要：图像情感分析旨在分析和理解视觉内容所传达的情感，其挑战在于弥合潜在视觉特征与抽象情感间的情感鸿沟。现有的深度模型试图一次性通过直接在全局范围内学习有辨别力的高级情感表征来弥合鸿沟，但忽略了深度模型各层特征之间的层次关系，导致上下文特征间的关联缺失。为此，提出一种上下文层次交互网络（CHINet）来建立层次结构中的上下文信息和情感之间的相关性模型。该模型包含2个分支：自下而上的主分支直接在高级语义层次上学习全局情感表征，针对该分支的不同层次特征，通过构建浅层风格编码器和情感激活注意力机制来分别提取风格表示并定位潜在情感激活区域；所提取的特征被级联到金字塔结构作为自上而下分支，从而建模上下文层次相关性并为情感表示提供浅层视觉特征。通过全局和局部学习将低级风格属性和高级图像语义整合到一起。实验结果表明：所提模型在FI数据集上较同类方法（包括多层次特征融合方法和结合了局部情感区域的方法）提升了情感识别准确率。

关 键 词：图像情感分析；情感鸿沟；上下文关联；层次结构；情感分类
中图分类号：TP391
文献标志码：A **文章编号**：1001-5965(2025)07-2304-10

“一图胜千言”，与具体的图像语义感知任务相比，在抽象层面研究视觉情感更具挑战性，这一开创性的课题被称为图像情感分析[1]。根据心理学的描述，图像情感大致可分为2种情感极性，即积极极性和消极极性，其又可以进一步细分为子情感类别[2]。图像情感与复杂的视觉特征密切相关，对于观者而言，不论是图像中的高级语义概念还是低级视觉特征，如颜色、纹理等视觉元素都可以诱发人类情感[3-4]。抽象情感与具体视觉特征紧密相关却又无法被直观地通过数学语言表达，因此，设计一个综合表征来弥合情感鸿沟是情感分析的关键[5-6]。

现有的基于深度学习的方法试图通过在图像全局范围内直接学习得出最终的高级情感表征进行情感表达，然而深度神经网络的特征在信息传递上存在上下文关联，即信息传递越深，特征表达的层次就越高[7-8]。这些方法忽略了视觉特征之间的层次关系，导致深度神经网络层次关系中的上下文信息难以关联，因此，仅使用高层次特征来表示情感是不够的。同时，已有研究也表明颜色、纹理等底层信息也是与视觉情感相关的重要组成部分[9-10]。

为解决上述问题，本文提出一种上下文层次交互网络（context hierarchical interaction network, CHINet）模型用于图像情感推理。本文模型由2个分支组成：自下而上的分支通过主干网络从全局角度对图像语义中的情感进行分类；自上而下的分支通过浅层风格编码器从主干网络中提取细粒度的图像风格。图像风格用于将颜色、纹理和线条等低级视觉特征整合到一起[11-12]，而语义则从高级图像内容形成情感表征。此外，为融合层次上下文特征，设计了一种情感注意力机制（emotional attention mechan-

收稿日期：2023-06-12；**录用日期**：2023-09-08；**网络出版时间**：2023-11-07 14:39
网络出版地址：link.cnki.net/urlid/11.2625.V.20231106.1919.001
基金项目：国家自然科学基金(62162068,62061049)；云南省万人计划“云岭学者”专项(YNWR-YLXZ-2018-022)；云南省教育厅科学研究基金(2025J0008)

*通信作者. E-mail：danxu@ynu.edu.cn

引用格式：罗改芳，张浩，徐丹. 结合上下文关联的图像情感分析[J]. 北京航空航天大学学报，2025，51(7)：2304-2313.
LUO G F，ZHANG H，XU D. Image sentiment analysis by combining contextual correlation[J]. Journal of Beijing University of Aeronautics and Astronautics，2025，51(7)：2304-2313 (in Chinese).

ism, EAM)作用于 2 个分支间的金字塔结构中, 抽取神经网络中的各层级属性, 并构建相邻上下文特征图的相关性。整个模型通过 2 个分支将图像分离为高级语义和低级风格, 并通过浅层风格编码器及情感激活区域提取模块建立网络层级中的上下文关联。引入全局和局部联合学习来优化模型, 用于从全局到局部视角将 CHINet 集成到统一的端到端框架中。本文的主要创新点如下:

1) 提出一种浅层风格编码器将卷积神经网络(convolutional neural networks, CNN)层次结构中不同级别的特征进行分离, 充分考虑浅层风格特征和高级语义的融合来进行情感推理。

2) 设计了一种情感注意力机制用于 CNN 主干与浅层风格编码器之间, 并通过图像金字塔建模层次特征中的上下文关联。

3) 使用全局学习和局部学习的统一框架, 完成上下文特征之间的协调优化, 从而提高情感分类的准确度。

1　相关工作

1.1　情感分析模型

计算机视觉和多媒体领域通常以心理学常用的 2 种方式描述情感: 离散情感状态(categorical emotion states, CES)和维度情绪空间(dimensional emotion space, DES)[13]。本文更关注 CES, 与 DES 相比, CES 可解释性更强。由于情绪的复杂性, 研究者对离散情感状态存在多种界定方式, 分类数量也各不相同。本文选择具有代表性的 2 类情感极性理论和八类情感状态来综合评估本文模型。2 类情感极性将情绪分为积极和消极, 而 8 类情感理论包括愉悦、敬畏、满足、激动、恐惧、悲伤、恶心和愤怒, 不同类型的情感图像示例如图 1 所示。早

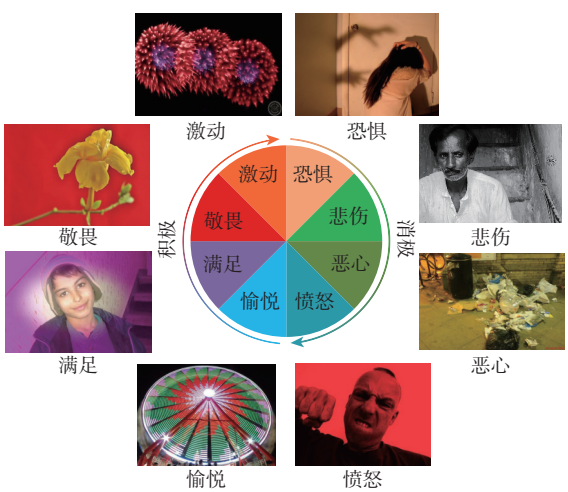


图 1　来自 FI 数据集的情感图像示例

Fig. 1　Affective image examples from FI dataset

期, 研究者从心理学的角度出发, 人为设计并提取视觉特征来表达情感, 与从 CNN 中提取的特征相比, 这些手工制作的特征主要集中在低级视觉线索上, 而高级语义没有得到充分利用[13]。为此, 本文更注重采用深度学习的思路, 构造一种能够提取图像上下文关系的神经网络模型用于情感分析。

1.2　基于深度学习的全局图像情感特征提取

已有大量工作将深度神经网络用于情感分析。Peng 等[14]利用迁移学习将预训练的 CNN 微调, 并在 EmotionROI 数据集上进行情感识别; Campos 等[15]通过消融实验对 AlexNet[16]进行逐层分析。结果均证明 CNN 可以捕获比手动方法更好的情感表征。一些工作将高级语义与低级视觉特征相结合来指导情感分类, 这类方法与本文工作直接相关, 都采用了来自于视觉领域被广泛使用的 CNN, 并根据不同层次设计出针对性的特征提取模块。Zhu 等[17]设计了一个联合 CNN 和循环神经网络(recurrent neural networks, RNN)的模型, 通过卷积层提取视觉和语义特征, 然后, 使用双向循环神经网络进行融合。文献 [10] 和文献 [18] 提出多层次深度网络, 结果表明图像情感可以通过图像语义、美学和低级特征来识别。She 等[19]则从弱监督的角度切入, 提出一种用于视觉情感识别的多核网络。You 等[20]提出一种渐进式 CNN(Progressive CNN, PCNN)架构, 有效利用弱标记数据提高模型泛化能力。DeepSentiBank[21]是一种基于 CNN 的视觉情感概念提取块, 用于发现形容词-名词对, 可提取 2 089 维向量作为中层表示。基于贝叶斯网络, Zhang 等[22]提出一种对象语义情感相关模型来指导图像情感分类。Yamamoto 等[23]提出一种同时使用视觉特征和语义的图像情感识别算法, 通过现有对象检测模型的高级嵌入获得语义, 而 CNN 提取基本视觉特征, 最终实现情感分类。

然而, 目前大多数深度模型均是从输入图像的全局视角中直接学习到网络最后的深层次表征, 并没有仔细考虑情感分析中涉及的情感特征演化过程, 从而忽略了深度模型中的层次关联。基于以上因素, 本文充分考虑网络中不同层次特征的上下文相关性, 针对不同层次的特征设计不同的情感特征提取模块进行最终的图像情感推理。

1.3　融合情感激活区域的局部情感特征提取

结合情感激活区域的方法也与本文模型密切相关, 这类方法结合了图像的局部和全局特征[24-26]。Peng 等[27]提出的监督情感区域检测方法使用情感激活图来表示每个像素对诱发情感的贡献, 该方法

综合考虑了不同区域对情感的影响。

She等[3]提出一种带有侧分支的弱监督耦合网络，通过侧分支检测情感诱发区域并将其反馈到模型主分支用于最终的图像情感分类，取得了良好的效果。Sun等[28-29]提供了一个新的思路，其使用双分支策略结合局部信息和图像整体表征：先检测出若干情感候选区域并对每个区域进行评分；然后，根据评分确定出首选情感激活区域；最后，将图像整体和局部情感激活区域信息进行融合。此外，Zhang等[30]提出一种多层次的情感区域相关性分析（multi-level sentiment region correlation analysis, MSRCA）模型，该模型首先从原始情感图像多个层次中发现与情感潜在相关的区域，然后，使用Transformer编码器从多个角度探索区域之间的相互作用和依赖关系。Xiong等[31]提出的基于群稀疏正则化的区域卷积神经网络（region-based convolutional neural networks using group sparse regularization, R-CNNGSR）也通过自动检测情感区域来提升情感识别准确率。

基于情感区域构建的情感分类器可有效地从情感区域中提取情感表征，增加模型对于局部信息的关注，同时避免接触到非情感区域的冗余噪声。然而，以往的方法大多采用现有的目标检测工具来定位情感激活区域，这类检测方案通常是为目标检测等视觉任务设计的，其丧失了模型自动探索用于情感识别的区域局部特征的可能性。因此，如何定位潜在情感诱发区域也是值得考虑的问题。基于以上挑战，本文提出情感激活注意力机制，探索用于图像局部情感表示的解决方案。

2 情感推理模型

2.1 上下文层次交互网络

相关研究表明，虽然图像语义更能唤起情绪，但纹理、颜色等潜在的底层特征也有利于情感表达[32-33]。现有的用于情感分析的深度模型大多从图像级别学习高级图像语义以弥合情感鸿沟，导致浅层特征的缺失。为此，本文通过分离浅层风格特征和高级语义来进行情感推理。

受图像风格迁移的启发，采用图像风格表示浅层视觉特征，因为其包含了线条、纹理、颜色等。本文提出的上下文层次网络由2个分支组成，模型的整体结构如图2所示，其中，$\boldsymbol{F}_l$为特征图，MLP表示多层感知器。自下而上的分支采用了预训练的ResNet[34]，预训练的模型为提取视觉语义提供了先验知识。ResNet由若干串联的残差块组成，通过极大池化进行下采样。ResNet模型共包含4个阶段，根据模型中特征图的大小是否相同将模型中的层划分为是否在同一阶段。最后的高级语义表征经过一个全局平均池化层所得出。为充分利用中间网络层特征的上下文关系，浅层风格编码器提取每个阶段后的特征进行图像风格表示。不同级别特征的风格表示被合并到自上而下的分支中，并通过层次金字塔结构连接在一起，用于浅层特征表示，以消除情感差距。图3为浅层风格编码器的计算过程，特征具体维度被标识在组件下方以便于查看。

形式化地，记模型第l层的特征图$\boldsymbol{F}_l \in \mathbf{R}^{H\times W\times C}$，其中，$H$和$W$为特征图的高和宽，$C$为通道数，$\boldsymbol{F}_l$中

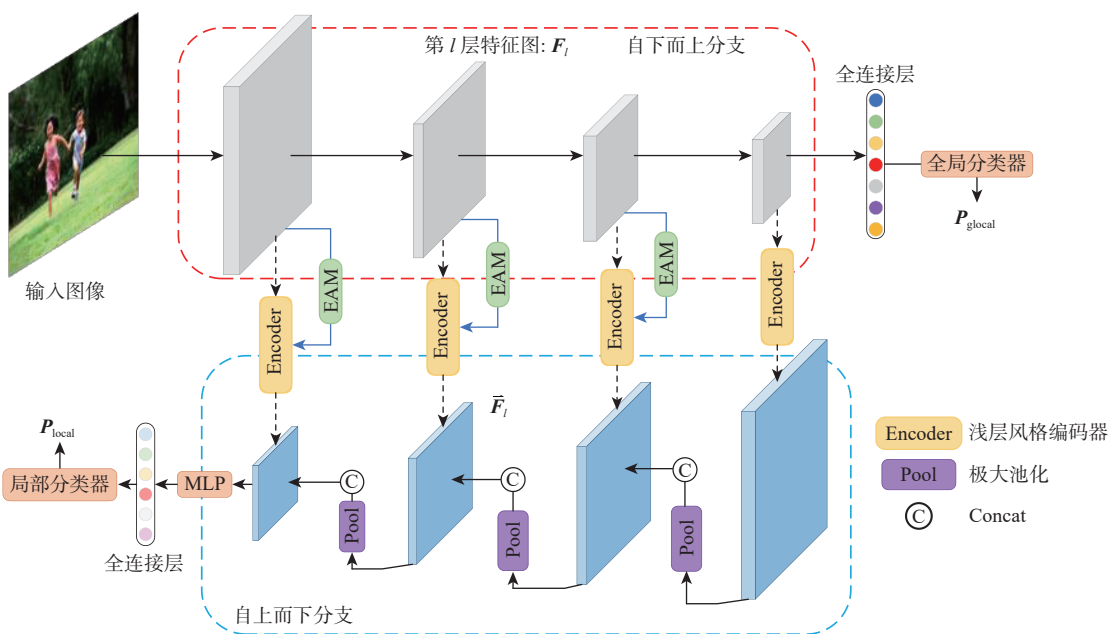


图2 CHINet整体结构

Fig. 2 Overall structure of CHINet

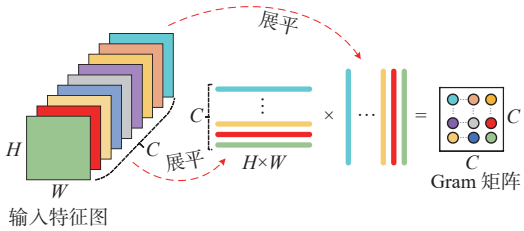


图 3　浅层风格编码器

Fig. 3　Shallow style encoder

的第 i 个通道的特征图为 $\boldsymbol{x}_i \in \mathbf{R}^{H\times W}$。对于第 l 层特征图 $\boldsymbol{F}_l$ 中的每个特征图 $\boldsymbol{x}_i$, 将 $\boldsymbol{x}_i$ 展平为大小为 $H\times W$ 的一维向量。然后, 将变换后的特征图 $\boldsymbol{F}'_l \in \mathbf{R}^{H\times W\times C}$ 乘以其转置形成 Gram 矩阵 $\boldsymbol{G}\in \mathbf{R}^{C\times C}$, 其表达式为

$$\boldsymbol{G} = (\boldsymbol{F}_{l:C} \to \boldsymbol{F}'_l)\otimes(\boldsymbol{F}_{l:C} \to \boldsymbol{F}'_l)^{\mathrm{T}} \tag{1}$$

式中: ⊗表示矩阵乘法。Gram 矩阵中的每个元素 $G_{i,j}$ 是扁平化后的特征 $\boldsymbol{x}_i$ 和 $\boldsymbol{x}_j$ 的内积:

$$G_{i,j} = \frac{1}{C\times H\times W}\sum_{k=1}^{n} x_{ik}x_{jk} \tag{2}$$

其中: $1/(C\times H\times W)$ 为归一化因子, 用于模型优化。

2.2　情感激活区域提取模块

盲目地考虑整幅图像的所有信息可能会削弱特征表达情绪的能力, 从而降低情感分类的准确性[35]。本节提出一种情感注意力机制来获取情感激活图, 模块结构如图 4 所示。与以往的注意力机制不同, 本文 EAM 用于获取情感诱发区域的注意力图, 同时过滤掉大部分噪声, 模块强大的自适应学习能力可以将原始输入特征图映射到若干弱相关的注意力图中。

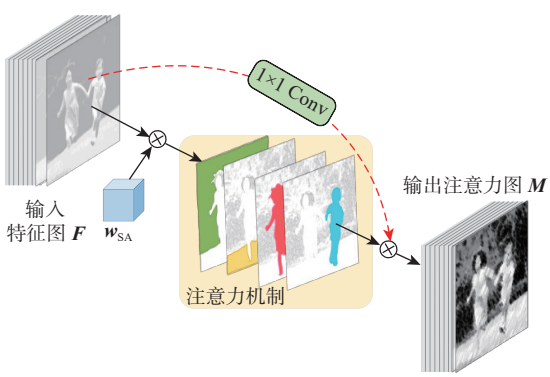


图 4　情感激活注意力机制

Fig. 4　Emotion activation attention mechanism

对于第 l 层特征图 $\boldsymbol{F}_l \in \mathbf{R}^{H\times W\times C}$, 使用逐点卷积将每个位置像素 $\boldsymbol{x}_i \in \mathbf{R}^{H\times W}$ 映射到 n 个情感激活图之一。在每个映射中, 在空间维度上池化像素以获得情感激活图 $\boldsymbol{M} \in \mathbf{R}^{n\times C}(n << HW)$:

$$\boldsymbol{M} = \underset{HW}{\mathrm{softmax}}(\boldsymbol{F}_l\boldsymbol{w}_{\mathrm{SA}})^{\mathrm{T}}\boldsymbol{F}_l \tag{3}$$

式中: 自适应滤波器 $\boldsymbol{w}_{\mathrm{SA}} \in \mathbf{R}^{C\times n}$ 用于从输入特征 $\boldsymbol{F}_l$ 转换成情感激活图, 经 softmax(·) 激活将其转化为维度为 $H\times W$ 的注意力权重。将情感激活图 $\boldsymbol{M}$ 与图像 Gram 矩阵相乘, 得到输入特征图 $\boldsymbol{F}_l$ 对应的 $\tilde{\boldsymbol{F}}_l$, 最后, 由 1×1 卷积进行不同通道之间的信息交互:

$$\tilde{\boldsymbol{F}}_l = \max(\boldsymbol{0},\ \mathrm{Conv1\times 1}(\boldsymbol{M}\otimes\boldsymbol{G})) \tag{4}$$

2.3　全局和局部联合学习

本文 CHINet 模型包含 2 个分类器, 即全局分类器和局部分类器, 预测概率分别为 $\boldsymbol{P}_{\mathrm{global}}$ 和 $\boldsymbol{P}_{\mathrm{local}}$, 这使得模型不仅能够专注于特定级别的情感表征, 如低级视觉细节或高级语义特征, 同时也能更好地提取层级特征, 这对于不同类型的情感图像至关重要。例如, 社交媒体图像更倾向高级语义, 而抽象绘画更倾向于颜色、线条等低级视觉风格。同时优化 2 个分类器可以帮助模型发挥优势并避免劣势。基于以上因素, 本节采用全局和局部学习来优化模型。

对于局部分类器, 给定训练样本 $\{(x, y)\}$, 其中, x 为图像, y 为对应情感标签, CNN 使用卷积层和全连接层提取输入图像的逐层表示, 在最后的 softmax 层之后, 最后一个全连接层的输出被转换为 c 类图像情感的概率分布 $\boldsymbol{P}\in \mathbf{R}^{1\times c}$。在本文中, c=2 或 8 表示 2 个或 8 个情感类别。情感图像样本属于特定情感的概率被定义如下:

$$p_i = \frac{1}{N}\sum_{j=0}^{N} \mathrm{e}^{h_{i,j}} \Big/ \sum_{i=0}^{c} \mathrm{e}^{h_{i,j}} \qquad i\in[1,c] \tag{5}$$

式中: $h_{i,j}$ 为全连接层的输出; N 为样本数。预测概率的损失值通过交叉熵衡量:

$$L_{\mathrm{cls}} = -\sum_{i=0}^{c} y_i \lg p_i \tag{6}$$

其中: $y_i = \{y_i | y_i \in \{0,1\}, i = 1,2,\cdots,c\}$表示图像的情感标签; $\sum_{i=1}^{c} p_i = 1$。

在本文实验过程中, 首先, 通过全局分类器对主分支(自下而上分支)进行训练, 即固定住自上而下分支的权重, 调整主分支权重; 然后, 通过局部分类器对自上而下的分支进行训练, 即类似地固定住主分支的权重, 调整自上而下分支权重。然而, 单纯的使用以上训练方式将导致模型的预测结果更加偏向最后完成训练的分类器的结果, 即偏向局部分类器的预测概率 $\boldsymbol{P}_{\mathrm{local}}$, 导致情感识别准确率不高(实验结果见 3.2.2 节)。为使得模型作为一个端到端的整体架构, 在优化用于全局学习的分支局部分类器之后, 将 2 个情感表征分类器的预测结果进行加权并同时优化:

$$\boldsymbol{P} = \left[(1-\alpha)\boldsymbol{P}_{\mathrm{global}} + \alpha\boldsymbol{P}_{\mathrm{local}}\right]/N \qquad \alpha\in[0,1] \tag{7}$$

式中：α为超参数，用于平衡全局分类器和局部分类器对于模型最终预测结果的影响。由于 2 个分类器是相互独立事件，类似的，通过最小化交叉熵损失函数来学习最终的情感概率表达：

$$L_{\mathrm{cls}}=-\frac{1}{N}\sum_{j=0}^{N}\sum_{i=0}^{c}y_i\lg p_i \tag{8}$$

其中：c 为情感类别数。

3 实验结果和分析

3.1 实验细节

本文所有实验均在开源深度学习开发平台 MXNet 上实现。模型参数通过小批量梯度下降法进行优化，对于二元情感分类和多元情感识别任务，批量分别设置为 16 和 32。采用动量法加速模型收敛，动量系数设为 0.9。学习率初始化为 0.001，每 30 轮乘 0.1 直至模型收敛。同时，采用图像增强方案用于随机改变训练样本以减少模型对特定属性的依赖，从而提高泛化能力，即将输入图像尺寸调整为 256×256，然后，随机裁剪回 224×224，同时按 0.5 概率镜像翻转。最后，所有输入图像被规范化后输入模型进行训练。所有实验均在 2 块 Nvidia RTX2080ti 显卡上进行。

本文选择 5 个图像情感数据集验证本文模型的情感识别准确率，包括 Abstract[9]、ArtPhoto[9]、Twitter I[20]、EmotionROI[14] 和 FI[1]，表 1 为数据集对应详细信息。为充分验证本文模型的有效性，本文不仅讨论了二元情感分类，还探索了多元情感识别，二元情感分类即将该任务视为积极、消极的情感极性识别问题，多元情感识别则按原始数据集标签进行 8 类情感识别任务。对于二元情感极性分类任务，除 Twitter I 数据集最初用二元情感标记外，其他数据集则根据维度情绪空间中的愉悦度（Valance）将多情感标签转换为积极和消极极性。

表 1 图像情感数据集详细信息

Table 1 Details of image emotion datasets

数据集	积极极性数量					消极极性数量					合计
	愉悦	敬畏	满足	激动	小记	愤怒	恶心	恐惧	悲伤	小记	
Abstract[9]	25	15	63	36	139	3	18	36	32	89	228
ArtPhoto[9]	101	102	70	105	378	77	70	115	166	428	806
EmotionROI[14]			330	330	660	330	330	330	330	1 320	1 980
FI[1]	4 942	3 151	5 374	2 963	16 430	1 266	1 658	1 032	2 922	6 878	23 308
Twitter I[20]					769					500	1 269

3.2 消融实验和参数分析

3.2.1 主干选择

为确定主干网络，先分析了几种广泛使用的 CNN 模型，包括 AlexNet[16]、VGG[36]、DenseNet[37] 和 ResNet[34]，因为其在视觉领域稳定且优秀的性能表现和相对统一且一致的框架结构，所有模型参数均来自开源开发平台 MXNet 模型库并在 FI 数据集上进行了微调。表 2 为各模型在 FI 上的情感分类准确率。由于无论是局部学习还是全局学习，CHINet 的模型最后都采用了全连接层 fc，所以表中包含了对最后一个全连接层在情感分析任务中作用的分析，这与本文工作密切相关。由表可知，ResNet 系列取得了最好的效果。特别是 152 层的 ResNet 表现最好，准确率为 85.19%。

相比之下 AlexNet 准确率最差，因为其只有 5 个卷积层，但对于同一系列的网络，具有更多层的模型则效果也更好。为此，本文选择 ResNet-152 作为主干模型。还发现 DenseNet 虽然有 100 多个卷积层，但其效果仍然不够理想，因为每个卷积层中的通道数有限。不仅如此，VGG 优于具有更多卷积层的 ResNet-18，因为最后的全连接层提供了更多可学习的参数，可以将学习到的“分布式表示”映射到特定的情感标签。表 2 中的实验结果表明，网络末端的全连接层有助于情感图像识别，因为其

表 2 FI 上不同层的 CNN 情感分类准确率

Table 2 Emotion classification accuracy of CNN models with different layers on FI

模型	层	分类准确率/%	
		c=2	c=8
AlexNet[16]	5 Conv + 2 fc	65.43	42.78
VGG-16[36]	13 Conv + 2 fc	81.35	63.84
VGG-19[36]	16 Conv + 2 fc	82.62	64.23
DenseNet-100 w/o fc[37]	99 Conv	83.77	64.76
DenseNet-100[37]	99 Conv + 2 fc	84.21	65.08
ResNet-18 w/o fc[34]	17 Conv	80.16	60.76
ResNet-18[34]	17 Conv + 2 fc	81.27	61.55
ResNet-50 w/o fc[34]	49 Conv	83.52	62.81
ResNet-50[34]	49 Conv + 2 fc	84.06	63.74
ResNet-152 w/o fc[34]	151 Conv	84.73	66.27
ResNet-152[34]	151 Conv + 2 fc	85.19	66.44

提供了冗余参数来构建情感推理的高级概念。基于以上原因, 本文同样引入类似的全连接层。

3.2.2　全局和局部学习

在 2.3 节中提到, 应用全局和局部学习来优化本文提出的深度模型。本节讨论优化策略对模型性能的贡献, 结果如表 3 所示, 表中, CHINet w/o G 和 CHINet w/o L 分别表示在没有全局学习和没有局部学习训练情况下模型的情感识别准确率结果, CHINet G&L 则表示同时使用全局和局部学习训练下的模型结果。基线模型是经过微调的 ResNet-152。由表 3 可知: 与局部学习相比, 同时优化局部和全局分类器可提高分类精度 2.38%; 其次, 使用全局学习比局部学习更好, 因为全局学习可以综合 2 个分支不同层次的情感表征。结果证明局部学习和全局学习的优化方法之间能够有效互补。

此外, 分类器权重参数 α 也是关键因素之一, 其决定了模型优化时 2 个分类器对模型的贡献。图 5 为不同 α 下性能的变化趋势, 当 α=0 时, 模型退化为只使用局部学习, 反之则代表只使用了全局学习。可以发现当 α=0.6 时, 模型的情感分类准确率最好, 说明同步优化 2 个分类器有利于模型的情感推理, 在后续实验中遵循了 α=0.6 这一设置。

表 3　基于局部和全局学习的情感分类准确率

Table 3　Emotion classification accuracy based on local and global learning

模型	准确率/%
基线模型	85.19
CHINet w/o G	87.36
CHINet w/o L	88.41
CHINet G&L	89.74

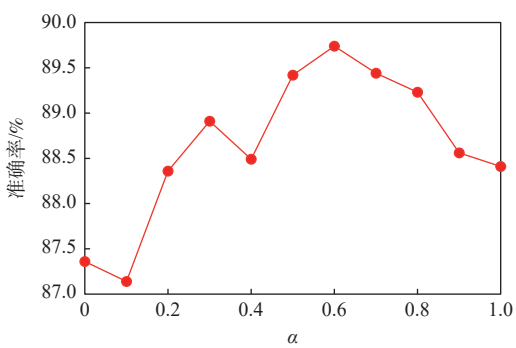


图 5　不同 α 对模型性能的影响

Fig. 5　Effect of different α on model performance

3.3　结果对比

为证明本文模型的有效性, 本节将其与其他方法进行比较, 包括几种经典的手工提取特征算法和最新的深度模型。

首先, 进行二元情感分类。表 4 为本文模型在 FI 上的性能。可以观察到, 本文模型优于为对象分类任务设计的深度 CNN, 如 AlexNet、VGG、DenseNet 和 ResNet。同时, 本文 CHINet 比其他为情感分类设计的模型(包括 PCNN[20] 和 DeepSentiBank[21])拥有更好的性能。与同样构建侧分支, 如情感区域(affective regions, AR)[29]、视觉和情感特征(visual and semantic features, VSF)[23] 相比, CHINet 仍然以 89.74% 的情感识别准确率实现最佳性能。同时, 基于多层次表示的 CHINet 比类似的基于多级特征融合(multi-level representations, MLR)[18] 和 MSRCA[30] 获得更好的性能, 这表明本文模型受益于上下文层次表示和更合适的优化方法。

表 4　FI 数据集上二元情感极性分类准确率比较

Table 4　Classification accuracy comparison of binary emotion polarity on FI dataset

模型	网络	准确率/%
基准模型	AlexNet[16]	60.54
	VGGNet[36]	70.64
	ResNet[34]	72.22
	Fine-tuned AlexNet[6]	72.43
	Fine-tuned VGGNet[36]	83.05
	Fine-tuned ResNet-50[34]	85.19
对比方法	PCNN (VGGNet)[20]	75.34
	DeepSentiBank[21]	61.54
	AR[29]	86.35
	VSF[23]	88.11
	MSRCA[30]	87.40
	MLR[18]	87.87
本文模型	CHINet w/o G	87.36
	CHINet w/o L	88.41
	CHINet G&L	89.74

此外, 表 5 为当前最先进的方法在几个小规模数据集上的二元情感表现, 表中, 无数据表示本文未能复现出的结果。由表可知, CHINet 在所有数据集上始终优于其他模型。对于 ArtPhoto 和 Abstract 数据集, 本文模型有相对明显的性能提升, 分类准确率分别为 83.27% 和 82.33%。值得一提的是, ArtPhoto 和 Abstract 数据集由缺乏对象或语义但充满低级特征(如纹理、颜色等)的抽象绘画组成。这证明了浅层视觉细节的重要性。除此之外, 全局和局部学习也是必不可少的组成部分, 可以使模型的性能全面超越其他方法。

为进一步评估本文模型的有效性, 本文还在 8 类别情感识别任务上进行了实验, 表 6 为具体实验结果。表中手工特征算法被删除对比, 因为其效果较弱不具备直接对比的价值。实验中采用预训

表5 小规模数据集上的二元情感极性分类准确率比较

Table 5 Classification accuracy comparison of binary emotion polarity on small-scale datasets %

模型	网络	ArtPhoto[9]	Abstract[9]	Twitter I[20]			EmotionROI[14]
				Twitter I 5	Twitter I 4	Twitter I 3	
手工特征	PAEF[13]	67.85	70.05	72.90	69.61	67.92	75.24
	SentiBank	67.74	64.95	71.32	68.28	66.63	66.18
深度学习	DeepSentiBank[21]	68.73	71.19	76.35	70.15	71.25	70.11
	PCNN (VGGNet)[20]	70.96	70.84	82.54	76.52	76.36	73.58
	Fine-tuned VGG[36]	70.09	72.48	84.35	82.26	76.75	77.02
	AR[29]	74.80	76.03	88.65	85.10	81.06	81.26
	R-CNNGSR[31]	75.02	75.89				81.36
	VSF[23]	81.62	81.82			83.10	83.17
	MLR[18]	75.63	77.85	89.77	85.72	81.49	83.08
本文模型	CHINet w/o G	78.61	79.07	86.54	83.04	81.36	82.26
	CHINet w/o L	81.12	81.15	89.51	82.95	82.17	82.47
	CHINet G&L	83.27	82.33	90.52	86.43	84.37	83.53

练的CNN，然后，在FI上进行微调。与结合了情感激活区域的模型，包括对象语义情感关联模型（object semantics sentiment correlation model, OSSCM）[22]、弱监督耦合网络（weakly supervised coupled networks, WSCNet）[3]和VSF[23]相比，CHINet取得了最高的识别精度，因为本文模型可以自适应地选择浅层视觉特征或深层语义，以更好的建模层次结构的上下文关联。此外，与基于多级特征融合的MSRCA[30]和MLR[18]相比，CHINet达到超越2.51%的性能提升，同时，即使在不考虑全局与局部学习的统一学习框架下，其表现也能取得近0.9%的准确率提升，证明了本文模型的有效性。总的来说，本文模型可以有效地学习更具辨别力的情感表示，提高分类性能。

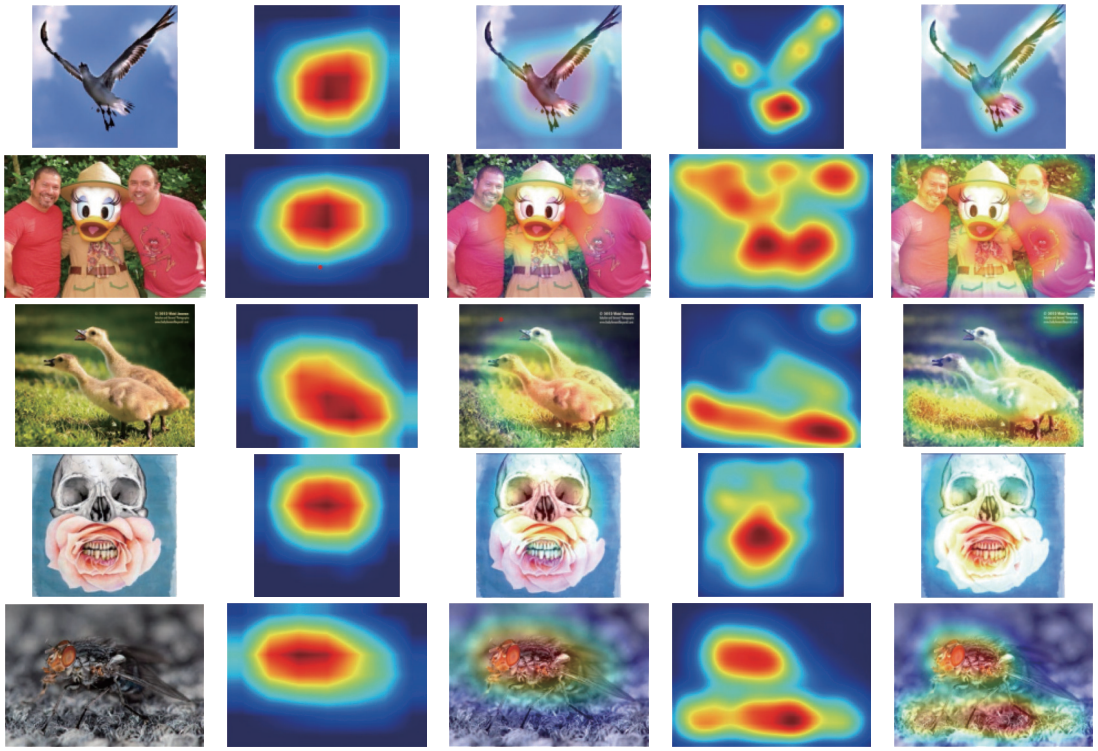
表6 FI数据集上多类别情感分类性能比较

Table 6 Comparison of multi-category emotion classification performance on FI dataset

模型	网络	准确率/%
基线模型	AlexNet[16]	49.54
	VGGNet-19[36]	61.74
	ResNet[34]	53.08
	Inception[38]	60.12
	Inception-Resnet[39]	62.77
对比模型	DeepSentiBank[21]	51.29
	PDANet[40]	69.42
	MKN[19]	63.92
	CycleEmotionGAN[41]	66.79
	Deep metric learning[42]	68.37
	MldrNet[16]	67.24
	OSSCM[22]	69.32
	WSCNet[3]	70.07
	VSF[23]	70.46
	MSRCA[30]	69.05
	MLR[18]	67.49
本文模型	CHINet w/o G	69.94
	CHINet w/o L	70.32
	CHINet G&L	71.56

3.4 可视化分析和讨论

为观察模型的情感推理过程并通过可视化分析模型的潜在缺陷，采用Grad-类激活图（class activation map, CAM）[43]展示微调的ResNet和CHINet对情感图像样本预测的CAM可视化结果。预测结果如图6所示，图6(a)为情感图像样本，图6(b)和图6(d)是ResNet和CHINet在Grad-CAM方法下的预测结果，图6(c)和图6(e)是预测结果与原图的叠加。可以观察到ResNet的激活图更为聚焦，表明其更注重高度集成的语义信息和显著的对象或目标，这类对象具有足够的语义代表性，可以进行初步的情绪预测，然而其无法获得进一步的背景信息。相比之下，本文模型不仅可以捕获具有强烈语义意义的对象，还可以捕获对预测情绪很重要的空间细节，如背景、颜色、纹理细节等。如第2行的结果，ResNet只聚焦于黛西，而CHINet还注意到黛西旁的2个人物的笑脸及人物衣服上的人物动作；对于第4行，本文模型不仅注意到图片上半部分的骷髅，还考虑了下半部分莲花中龇牙咧嘴的怪相。以上可视化的结果表明，CHINet的情感激活图不仅关注特定的对象，而且充分考虑到所有可能潜在的对象，以上可视化分析结果证实了关注层次结构中的上下文关联是有效的。



(a) 情感图像　(b) ResNet CAM　(c) ResNet CAM+情感图像　(d) CHINet CAM　(e) CHINet CAM+情感图像

图 6　用于情感推理的类激活图可视化

Fig. 6　Visualization of class activation map for emotion reasoning

4　结　论

1) 本文提出一种上下文层次交互网络来构建网络中间特征中的上下文关系, 并结合不同的视觉属性进行视觉情感分析。自下而上的分支学习全局的高级语义表示, 而自上而下的分支分离出图像风格以获取浅层情感特征, 最后, 通过全局和局部学习同时优化 2 个局部分类器。

2) 实验细节分析和消融实验结果验证本文模型的有效性。可视化的相关结果进一步表明, 本文模型可以同时捕获显著性对象和潜在情感诱发区域。

未来, 为使本文模型能够适应各种类型的情感图像, 仍需考虑如何构造更加有效且扩展性更好的模块对于不同层次特征的挖掘。同时, 如何更加有效地建模上下文特征关联也是值得考虑的方向。

参考文献（References）

[1] YOU Q Z, LUO J B, JIN H L, et al. Building a large scale dataset for image emotion recognition: the fine print and the benchmark[C]// Proceedings of the AAAI Conference on Artificial Intelligence. Palo Alto: AAAI Press, 2016, 308-314.

[2] XU L W, WANG Z T, WU B, et al. MDAN: multi-level dependent attention network for visual emotion analysis[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2022: 9469-9478.

[3] SHE D Y, YANG J F, CHENG M M, et al. WSCNet: weakly supervised coupled networks for visual sentiment classification and detection[J]. IEEE Transactions on Multimedia, 2020, 22(5): 1358-1371.

[4] YANG H S, FAN Y Y, LV G Y, et al. Exploiting emotional concepts for image emotion recognition[J]. The Visual Computer, 2023, 39(5): 2177-2190.

[5] YANG J Y, LI J, WANG X M, et al. Stimuli-aware visual emotion analysis[J]. IEEE Transactions on Image Processing, 2021, 30: 7432-7445.

[6] LIANG Y, MAEDA K, OGAWA T, et al. Chain centre loss: a psychology inspired loss function for image sentiment analysis[J]. Neurocomputing, 2022, 495: 118-128.

[7] 张浩, 李海鹏, 彭国琴, 等. 多层次特征融合表征的图像情感识别[J]. 计算机辅助设计与图形学学报, 2023, 35(10): 1566-1576.
ZHANG H, LI H P, PENG G Q, et al. Image emotion recognition via fusion multi-level representations[J]. Journal of Computer-Aided Design & Computer Graphics, 2023, 35(10): 1566-1576(in Chinese).

[8] ZHANG H M, XU M. Multiscale emotion representation learning for affective image recognition[J]. IEEE Transactions on Multimedia, 2022, 25: 2203-2212.

[9] MACHAJDIK J, HANBURY A. Affective image classification using features inspired by psychology and art theory[C]//Proceedings of the 18th ACM International Conference on Multimedia. New York: ACM, 2010.

[10] RAO T R, LI X X, XU M. Learning multi-level deep representations for image emotion classification[J]. Neural Processing Letters, 2020, 51(3): 2043-2061.

[11] 詹明. 融合风格特征的抽象图像情感识别[D]. 吉林: 吉林大学, 2023.

ZHAN M. Affective analysis of abstract images using style representation[D]. Jilin : Jilin University, 2023(in Chinese).

[12] 尹朝. 基于内容生成与特征提取的图像情感识别模型研究[J]. 系统仿真技术, 2023, 19(2): 141-147.
YIN C. Research on image emotion recognition model based on feature extraction and content generation[J]. System Simulation Technology, 2023, 19(2): 141-147(in Chinese).

[13] ZHAO S C, YAO X X, YANG J F, et al. Affective image content analysis: two decades review and new perspectives[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 44(10): 6729-6751.

[14] PENG K C, CHEN T, SADOVNIK A, et al. A mixed bag of emotions: model, predict, and transfer emotion distributions[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2015: 860-868.

[15] CAMPOS V, SALVADOR A, GIRO-I-NIETO X, et al. Diving deep into sentiment: understanding fine-tuned CNNs for visual sentiment prediction[C]//Proceedings of the 1st International Workshop on Affect & Sentiment in Multimedia. New York: ACM, 2015.

[16] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. ImageNet classification with deep convolutional neural networks[J]. Communications of the ACM, 2017, 60(6): 84-90.

[17] ZHU X G, LI L, ZHANG W G, et al. Dependency exploitation: a unified CNN-RNN approach for visual emotion recognition[C]//Proceedings of the 26th International Joint Conference on Artificial Intelligence. New York: ACM, 2017: 3595-3601.

[18] ZHANG H, XU D, LUO G F, et al. Learning multi-level representations for affective image recognition[J]. Neural Computing and Applications, 2022, 34(16): 14107-14120.

[19] SHE D Y, SUN M, YANG J F. Learning discriminative sentiment representation from strongly- and weakly supervised CNNs[J]. ACM Transactions on Multimedia Computing, Communications, and Applications, 2019, 15(3s): 1-19.

[20] YOU Q Z, LUO J B, JIN H L, et al. Robust image sentiment analysis using progressively trained and domain transferred deep networks[C]//Proceedings of the AAAI Conference on Artificial Intelligence. Palo Alto: AAAI Press, 2015, 29(1): 1-8.

[21] CHEN T, BORTH D, DARRELL T, et al. DeepSentiBank: visual sentiment concept classification with deep convolutional neural networks[EB/OL]. (2014-10-30)[2023-06-01]. https://arxiv.org/abs/1410.8586.

[22] ZHANG J, CHEN M, SUN H, et al. Object semantics sentiment correlation analysis enhanced image sentiment classification[J]. Knowledge-Based Systems, 2020, 191: 105245.

[23] YAMAMOTO T, TAKEUCHI S, NAKAZAWA A. Image emotion recognition using visual and semantic features reflecting emotional and similar objects[J]. IEICE Transactions on Information and Systems, 2021, 104(10): 1691-1701.

[24] YANG J Y, GAO X B, LI L D, et al. SOLVER: scene-object interrelated visual emotion reasoning network[J]. IEEE Transactions on Image Processing, 2021, 30: 8686-8701.

[25] DENG Z L, ZHU Q R, HE P, et al. A saliency detection and gram matrix transform-based convolutional neural network for image emotion classification[J]. Security and Communication Networks, 2021, 2021: 6854586.

[26] 彭国琴. 基于深度学习的图像情感语义分析关键问题研究[D]. 昆明: 云南大学, 2021.
PENG G Q. Research on key issues of image emotional semantic analysis based on deep learning[D]. Kunming: Yunnan University, 2021(in Chinese).

[27] PENG K C, SADOVNIK A, GALLAGHER A, et al. Where do emotions come from? predicting the emotion stimuli map[C]//Proceedings of the IEEE International Conference on Image Processing. Piscataway: IEEE Press, 2016: 614-618.

[28] SUN M, YANG J F, WANG K, et al. Discovering affective regions in deep convolutional neural networks for visual sentiment prediction[C]//Proceedings of the IEEE International Conference on Multimedia and Expo. Piscataway: IEEE Press, 2016: 1-6.

[29] YANG J F, SHE D Y, SUN M, et al. Visual sentiment prediction based on automatic discovery of affective regions[J]. IEEE Transactions on Multimedia, 2018, 20(9): 2513-2525.

[30] ZHANG J, LIU X Y, CHEN M, et al. Image sentiment classification via multi-level sentiment region correlation analysis[J]. Neurocomputing, 2022, 469: 221-233.

[31] XIONG H T, LIU Q, SONG S Y, et al. Region-based convolutional neural network using group sparse regularization for image sentiment classification[J]. EURASIP Journal on Image and Video Processing, 2019, 2019(1): 30.

[32] 申朕, 崔超然, 董桂鑫, 等. 基于深度多任务学习的图像美感与情感联合预测研究[J]. 软件学报, 2023, 34(5): 2494-2506.
SHEN Z, CUI C R, DONG G X, et al. Unified image aesthetic and emotional prediction based on deep multi-task learning[J]. Journal of Software, 2023, 34(5): 2494-2506(in Chinese).

[33] ZHANG H, LUO G F, YUE Y Y, et al. Affective image recognition with multi-attribute knowledge in deep neural networks[J]. Multimedia Tools and Applications, 2024, 83(6): 18353-18379.

[34] HE K M, ZHANG X Y, REN S Q, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition . Piscataway: IEEE Press, 2016: 770-778.

[35] 张浩, 徐丹. 基于深度学习的少数民族绘画情感分析方法[J]. 中国科学: 信息科学, 2019, 49(2): 204-215.
ZHANG H, XU D. Ethnic painting analysis based on deep learning[J]. Scientia Sinica (Informationis), 2019, 49(2): 204-215 (in Chinese).

[36] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition[EB/OL]. (2015-04-10)[2023-06-01]. https://arxiv.org/abs/1409.1556.

[37] HUANG G, LIU Z, VAN DER MAATEN L, et al. Densely connected convolutional networks[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2017: 2261-2269.

[38] SZEGEDY C, VANHOUCKE V, IOFFE S, et al. Rethinking the inception architecture for computer vision[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2016: 2818-2826.

[39] SZEGEDY C, IOFFE S, VANHOUCKE V, et al. Inception-v4, inception-ResNet and the impact of residual connections on

learning[C]//Proceedings of the AAAI Conference on Artificial Intelligence. Palo Alto: AAAI Press, 2017, 31(1): 1-7.

[40] ZHAO S C, JIA Z Z, CHEN H, et al. PDANet: polarity-consistent deep attention network for fine-grained visual emotion regression[C]//Proceedings of the 27th ACM International Conference on Multimedia. New York: ACM, 2019.

[41] SUN Z K, SARMA P, SETHARES W, et al. Learning relationships between text, audio, and video via deep canonical correlation for multimodal language analysis[C]//Proceedings of the AAAI Conference on Artificial Intelligence. Palo Alto: AAAI Press, 2020, 34(5): 8992-8999.

[42] YAO X X, SHE D Y, ZHANG H W, et al. Adaptive deep metric learning for affective image retrieval and classification[J]. IEEE Transactions on Multimedia, 2020, 23: 1640-1653.

[43] SELVARAJU R R, COGSWELL M, DAS A, et al. Grad-CAM: visual explanations from deep networks *via* gradient-based localization[J]. International Journal of Computer Vision, 2020, 128(2): 336-359.

Image sentiment analysis by combining contextual correlation

LUO Gaifang[1, 2], ZHANG Hao[1], XU Dan[1, *]

(1. School of Information Science & Engineering, Yunnan University, Kunming 650500, China;
2. School of Software, Shanxi Agricultural University, Jinzhong 030801, China)

Abstract: Image sentiment analysis aims to analyze the emotions conveyed by visual content. A key challenge in this field is to bridge the affective gap between latent visual features and abstract emotions. Existing deep learning models attempt to address this issue by directly learning discriminative high-level emotional representations globally at once but overlook the hierarchical relationship between features at each layer of the deep model, resulting in a lack of correlation between contextual features. Therefore, this paper proposed a context-hierarchical interaction network (CHINet) to model the correlation between contextual information and sentiment within the hierarchy. The model consists of two branches: a bottom-up branch, which first directly learns the global emotional representation at the high-level semantic level; then, for different feature level within the branch, it extracts the style representation and localizes potential emotion activation regions by shallow style encoder and emotion activation attention mechanism. The extracted features are then cascaded into a pyramid structure as top-down branches, modeling contextual hierarchical dependencies and providing shallow visual features for emotion representation. Finally, global and local learning integrate shallow image styles with high-level semantics. Experiments show that the proposed model improves emotion recognition accuracy on the FI dataset compared with related methods, including multi-level feature fusion methods and approaches incorporating local emotional regions.

Keywords: image sentiment analysis; affective gap; contextual correlation; hierarchy structure; emotion classification

Received: 2023-06-12; **Accepted**: 2023-09-08; **Published Online**: 2023-11-07 14:39

URL: link.cnki.net/urlid/11.2625.V.20231106.1919.001

Foundation items: National Natural Science Foundation of China (62162068,62061049); Yunnan Province Ten Thousand Talents Program and Yunling Scholars Special Project (YNWR-YLXZ-2018-022); Yunnan Provincial Department of Education Scientific Research Fund (2025J0008)

*** Corresponding author**. E-mail: danxu@ynu.edu.cn

http://bhxb.buaa.edu.cn jbuaa@buaa.edu.cn
DOI: 10.13700/j.bh.1001-5965.2023.0378

基于 IcD-FDRL 的应急监控视频边缘智能传输优化

李彦[1,2]，万征[1,*]，邓承志[2]，汪胜前[2]

(1. 江西财经大学 信息管理与数学学院，南昌 330032；2. 江西水利电力大学 信息工程学院，南昌 330099)

摘 要：应急监控视频传输作为提升突发事件监测、公共安全事件处理、灾后重建等情况下应急工作处理能力的关键技术手段，逐渐成为国家智慧应急体系建设重点支持的专业领域和研究方向。随着5G技术、决策型人工智能技术的不断发展，为实现自适应的高质量应急监控视频传输，针对局部区域内公共安全和应急救援监控，建立一种应急监控视频边缘智能传输架构，设计了应急监控视频重要性度量方法，提出簇内动态联邦深度强化学习(IcD-FDRL)算法，并实现了基于簇内动态联邦深度强化学习的应急监控视频边缘智能传输优化，以打破监控数据孤岛，提升算法学习效率，实现重要应急监控视频的低时延、低成本、高质量和优先传输。通过仿真实验进行了对比分析，验证了所提模型和算法的有效性。

关 键 词：应急监控视频；边缘集群；动态联邦深度强化学习；边缘智能；无线视频传输；移动边缘计算

中图分类号：TN929.5

文献标志码：A **文章编号**：1001-5965(2025)07-2314-16

当今社会，如公共安全、紧急事件、自然灾害、恐怖袭击等众多突发情况不断地考验着政府及其相应的职能机构的应急工作处理能力，促使应急通信技术不断发展。中国是世界上受自然灾害、事故灾难等影响较为严重的国家之一，亟须大力提升安全风险防范和应急处置能力，构建科学的应急管理体系迫在眉睫。《“十四五”国家应急体系规划》中指出，到2035年，建立与基本实现现代化相适应的中国特色大国应急体系，全面实现依法应急、科学应急、智慧应急，形成共建共治共享的应急管理新格局[1]。

应急通信依据通信方式、内容和覆盖面可分为各种不同的种类[2]。在突发事件监测、公共安全监管、灾后重建等事件的处理过程中，国家或地方需要实时的掌握现场的状况，此时语音、图像、视频等应急通信的必要性显得尤其突出，在区域内建立应急监控视频移动通信系统是至关重要的。

应急监控视频数据传输方式可以分为有线传输和无线传输。无线传输包括无线网络传输、微波传输、WiFi传输和5G传输等。随着移动通信技术的发展，无线传输成为主流的视频业务服务方式。而在应急场景下，应急监控视频传输根据应急场景的不同需求，对无线视频传输技术提出了更个性化的要求。一方面，应急监控视频传输仍然需要解决监控视频的高质量、低时延传输问题，以满足应急事件及时处理和准确决策的需要。另一方面，作为应急监控视频传输，需要解决重要视频的优先传输问题，以避免应急事件的发生，提升应急监管智能化水平。

针对应急监控视频传输中存在的上述问题，解决思路聚焦在如何有效利用无线视频传输技术为重要性高的监控视频优先提供自适应的低时延、低

收稿日期：2023-06-16；**录用日期**：2024-03-29；**网络出版时间**：2024-04-20 16:18

网络出版地址：link.cnki.net/urlid/11.2625.V.20240419.0927.001

基金项目：国家自然科学基金(61961021)；江西省教育厅科技计划重点项目(GJJ180251)；江西水利电力大学博士科研启动项目(2024kyqd062)

* **通信作者**. E-mail：wanzheng97@163.com

引用格式：李彦，万征，邓承志，等. 基于IcD-FDRL的应急监控视频边缘智能传输优化[J]. 北京航空航天大学学报，2025，51(7)：2314-2329.
LI Y，WAN Z，DENG C Z，et al. Edge-intelligent transmission optimization of emergency surveillance video based on IcD-FDRL[J]. Journal of Beijing University of Aeronautics and Astronautics，2025，51(7)：2314-2329 (in Chinese).

成本和高质量传输。无线视频传输中提高视频传输质量的研究集中在视频编解码算法和视频传输优化算法, 以及包括对两者相结合的联合优化。无论是视频编码算法还是传输优化算法, 都必须根据网络环境的动态变化进行相应的调整, 才能适应环境的频繁变化, 提高视频传输质量。

当前基于多种无线网络共用和流行的时代, 无线视频传输进入了多变异构无线网络阶段, 对无线视频传输业务的要求更多, 对视频传输质量标准的要求更高, 对视频传输智能优化算法的需求更甚。5G 通信技术的应用, 使得诸多低时延任务处理成为现实, 也促使移动通信网络向随时连接、随地连接的泛在化时代迈进[3]。而以深度学习为核心的人工智能(artificial intelligence, AI)技术的发展, 使得感知型人工智能技术崛起和盛行, 包括图像识别、语音识别、自然语言理解等智能感知技术[4]。随着感知型人工智能技术的产业化发展, 智能感知模块已经日趋成熟, 并可使用这些模块数据进行决策优化, 促使决策型人工智能技术不断发展。同时, 强化学习作为决策型人工智能的主流方法, 其理论和技术体系不断完善, 预计将推动决策型人工智能技术迎来快速发展的黄金十年。近年来, 深度强化学习(deep reinforcement learning, DRL)[5-6] 被广泛应用在无线网络传输优化中, 以解决计算任务卸载、网络资源分配和计算、缓存资源调度等相关问题。通信技术、人工智能技术的不断发展, 也使得应急监控视频系统逐渐向智能化、泛在化发展。

然而, DRL 在如何平衡探索与利用、容易陷入局部最优、训练数据采集有效性差及泛化能力弱、不稳定等方面仍影响着算法的性能和应用。其中, 一种有效的解决方法即通过将联邦学习(federal learning, FL)和 DRL 进行融合, 形成联邦 DRL(federated deep reinforcement learning, FDRL)[7-9], 以打破训练数据孤岛, 改善训练数据有效性, 协同提升算法学习效率。然而, 联邦学习促进 DRL 算法性能提升, 其关键在于联邦学习中参与客户端的选择、模型训练和更新规则的设计等核心技术。

因此, 本文针对应急场景下视频传输的特点和需求, 进行区域应急监控视频移动通信系统设计, 考虑到单独边缘节点缓存、计算和通信能力的局限性, 首先, 提出移动边缘计算(mobile edge computing, MEC)网络中边缘节点分簇算法, 并构建应急监控视频边缘智能传输网络架构。然后, 结合应急监控视频传输优化的需要, 建立应急监控视频传输数学模型, 设计应急监控视频的重要性度量方法。最后, 在边缘集群分簇网络结构基础上, 将 DRL、联邦学习[10] 融合引入应急监控视频传输优化中, 提出基于簇内动态 FDRL(Intra-clustered dynamic FDRL, IcD-FDRL)算法的应急监控视频边缘智能传输优化方法, 以实现为重要性高的监控视频优先提供自适应的低时延、低成本和高质量传输。

本文的主要创新点如下:

1) 本文提出边缘节点分簇算法。针对区域公共安全和应急监控对边缘网络能力的需要, 根据边缘节点分簇算法, 建立边缘集群簇, 并结合边缘计算和人工智能技术, 构建了应急监控视频边缘智能传输网络架构, 为应急监控视频低时延、低成本的高质量传输提供基础。

2) 本文设计了应急监控视频重要性度量方法。基于移动监控终端和视频监控中心的监控视频数据调阅观看的特点, 结合监控视频内容, 对应急监控视频的重要性进行度量, 以便于更重要的监控数据能优先显示在监控终端。

3) 本文提出基于 IcD-FDRL 算法的应急监控视频边缘智能传输优化方法。基于上述应急监控视频边缘智能传输架构, 根据监控智能终端网络条件和需求, 并结合边缘簇内各边缘节点计算和负载情况, 通过 DRL 算法感知决策, 为重要性高的监控视频优先提供自适应的高质量视频传输, 实现基于 DRL 的应急监控视频边缘智能传输方法。然后, 为提升 FDRL 效率, 打破监控数据孤岛, 构建了基于 IcD-FDRL 的应急监控视频边缘智能传输架构, 以进一步有效实现应急监控视频的低时延、低成本和高质量传输。

1　相关工作

1.1　无线监控视频传输

随着 5G 技术的普及应用, 5G 赋能的无线视频监控更具优势: ①监控点的可移动性, 借助于无线网络传输, 智能摄像头可以在移动环境下进行视频监控; ②监控点的广覆盖性, 借助移动网络的广覆盖特点, 在一些不具备有线宽带接入的区域或采用有线宽带接入监控成本过高的地区, 可方便地部署监控点; ③监控终端的可移动性, 借助移动客户终端, 可满足客户远程视频查看、应急指挥的要求。具备传输峰值超过 10 Gbit/s 的 5G 技术将会有效改善现有视频监控中存在的反应迟钝、监控效果差等问题, 能够以更快的速度提供更加高清的监控视频。因此, 当前更多采用的是无线视频传输方式。

无线视频监控系统的研究主要集中在 3 个方面: ①智能视频识别处理, 即无线视频监控中的视频场景解读和目标识别; ②无线视频监控系统中的

用户隐私保护和视频数据安全，确保监控网络系统的安全和敏感数据的安全；③无线监控视频传输与通信系统的优化设计，即在无线视频监控中如何进行有效的视频采集，并通过无线视频传输与通信系统设计的优化，实现低成本、低延迟、高质量的视频传输处理和决策反馈。这部分研究需要基于无线视频传输、云计算和边缘计算等理论。本文主要聚焦在无线监控视频的传输优化与通信系统设计，而其与监控视频“端-边-云”协同处理又密不可分[11]。

为获得较高的平均识别准确率和较低的平均处理延迟，Yan等提出一种用于人脸识别应用的多摄像头场景下移动边缘视频监控系统[12]。Ke等设计了一种基于边缘计算的智能停车监控任务[13]，在真实场景中实现了95%以上的准确率。为有效获取移动边缘网络的特征，Ming等提出一种基于边缘计算的图神经网络辅助智能构建视频监控系统[14]，可以实现可靠的传输准确性和较低的延迟。为有效分析监控视频并提取视频关键帧，Hou等[15]提出一种云-边协作架构，并使用DRL算法执行任务调度和自适应任务卸载。在无线视频监控系统下，Kim等提出一种基于深度Q学习的摄像机智能控制方法[16]，以实现监控视频的高效智能采集。

无线监控视频传输通信系统设计的目标是在保证监控视频识别准确性的同时实现更快的实时处理，这就需要高质量和高准确的无线监控视频传输保障技术。因此，监控视频可靠传输优化问题仍是无线视频监控亟需解决的关键问题，在提高视频传输质量的同时追求低延迟、低成本。

1.2 移动边缘计算视频传输

在大数据和万物互联的时代背景下，各类设备产生的数据量与日俱增，将所有数据传送到云中心几乎是不可能完成的任务[17-18]。事实上，大多数时候云中心只是需要处理后的结果而不是规模庞大的原始数据。于是，移动云计算、雾计算、海云计算等新型计算模型相继被提出，而边缘计算[19]则被认为是这些模型最终走向融合的产物，其利用云中心到终端之间各种现有或新增设备的处理能力，就地执行计算和存储任务，以实现减少骨干网带宽消耗、提高系统响应能力、降低云中心能耗和保护数据安全与用户隐私的目标。

随着移动通信技术的发展和智能终端的普及，人们越来越习惯于使用手机等移动设备处理日常事务和休闲娱乐，这使得大多数用户终端总是处在移动网络当中。为改善移动网络的用户体验，欧洲电信标准化协会提出MEC模型[20-21]，通过在移动网络边缘部署计算节点，就近处理移动终端提交的数据和计算任务。由于具有边缘计算的天然优势，MEC被视为向5G过渡的关键技术和架构。

虽然无线带宽一直在稳步增长，且应用了内容分发网络（content delivery networks, CDN）、对等通信（peer to peer, P2P）等视频传输的优化技术，但视频清晰度的提升、无线网络和用户终端的异构性、无线信道的不稳定和终端的移动等因素仍然导致无线视频的接收质量无法满足用户的需求[22]。MEC的提出，则为无线视频传输的优化带来了新的机遇。利用边缘设备的计算、存储和通信能力，可以在移动网络边缘提供视频转码和分发服务，从而减轻骨干网的负担并提升用户的观看体验。不过，MEC环境下的视频传输同样面临许多挑战：

1）兼容现有网络和视频传输技术。视频传输的优化必须与异构的无线网络环境（3G/4G/5G、WiFi）和主流的视频传输技术（CDN、P2P、直播等）融合。

2）突破单个边缘节点的能力限制。只有与邻近节点有效开展“边-边”协作，并分别与云中心和用户终端开展“云-边”和“边-端”协作，才能充分发挥MEC的作用。

3）平衡资源消耗与用户体验质量（quality of experience, QoE）提升的矛盾。绝大多数情况下用户QoE的提升往往意味着更多的资源消耗。如何平衡两者之间的矛盾，或者在特定情况下明确两者间的优先级，是影响视频传输性能的关键。

4）符合视频业务的特点。现有大多数关于计算任务卸载、缓存放置和内容分发的策略并非专为视频业务设计，只有充分考虑视频业务的特点、视频数据的重要性差异、不同视频版本间的关联及直播、点播、视频分析等业务的特殊需求，才能设计出高效的视频优化方案。

MEC通过将计算和存储任务下沉，有效减轻了云中心的负担，降低了带宽消耗并缩短了服务响应时间，同时也为无线视频的传输优化提供了有效手段。

1.3 DRL应用于网络传输优化

DRL是将深度学习和强化学习结合起来，从而实现端对端学习的一种全新的算法[23]，是通用的人工智能框架，目前已经成为网络优化的重要方法和工具。其在MEC环境下视频传输优化方面的研究逐年增加，目前该领域的研究主要包括以下几个方面：

1）基于DRL的缓存策略[24-27]。将DRL应用于缓存技术，主要解决缓存内容位置决策和更新，缓存策略及缓存内容交付等问题，利用DRL算法对网络信息进行分析和学习，提前进行资源分配和缓

存调度, 对相应视频内容和比特率版本进行缓存, 从而提高缓存命中率和缓存资源的利用率。

2) 基于 DRL 的实时转码调度决策[28-30]。DRL 在转码调度决策中的应用, 主要结合客户 QoE 及系统的服务质量(quality of service, QoS)等要求, 基于来自网络和客户的大量实时信息在边缘做出实时转码智能决策, 以求实现以最小化系统成本来适应个性化 QoE 和网络系统 QoS 要求, 从而满足不同客户对高质量视频服务体验的要求。

3) 基于 DRL 的网络通信资源分配[31-33]。DRL 在网络通信资源分配中的应用, 主要结合网络系统的 QoS 要求, 采用 DRL 算法, 做出动态分配无线电资源的决策。

4) 基于 DRL 的联合资源分配和调度[34-36]。在使用 DRL 解决 MEC 中联合资源分配和调度问题中, 深度 Q 学习网络(deep Q networks, DQN)等典型的 DRL 算法都得到了充分应用。

5) 基于 DRL 的其他实际应用[37]。文献 [36] 提出一种基于 DQN 的摄像机自动控制方法, 以提高视频监控系统中异常动作的识别精度。

边缘计算环境下 DRL 应用研究取得了一些有价值的研究成果, 特别是在计算、缓存和通信资源的协同优化和资源分配领域, 为后续研究打下了坚实的基础。

综上所述, 视频传输优化方面已有一定的研究成果, 可以拓展到应急视频监控领域, 但仍然存在一些能够改进的地方。因此, 研究一套适用于 MEC 环境下的无线监控视频传输方案, 极大地改善用户应急监控效率和管理水平具有重要意义。

1) 在应急视频监控系统设计上, 对于新型智能视频监控来说, 已经陆续引入了深度学习、强化学习等人工智能技术, 亟需建立一个更加满足智能视频监控需求的应急视频监控传输网络系统, 从而提升应急监控系统的运行效率。

2) 对传统的应急视频传输, 由于通信开销和时延较高, 并不适用于动态网络和大规模的监控网络。而设计实现基于 MEC 的应急监控视频传输, 更能满足应急视频监控需求, 实现低时延的应急视频传输。

3) 在复杂多变异构网络中, 如何应用 DRL、联邦学习等人工智能算法实现应急监控视频的高质量智能传输仍有待进一步研究。

2　应急监控视频边缘智能传输架构

本节将从应急监控视频的网络拓扑结构和传输流程 2 个方面阐述本文所构建的应急监控视频边缘智能传输架构。

2.1　网络拓扑结构

本文提出如图 1 所示的应急监控视频边缘智能传输架构, 与当前主流视频传输技术 CDN 无缝对接。在该网络结构中, 基于区域位置, CDN 末梢服务器与局部区域内的众多边缘节点共同组成边缘区域, 利用边缘节点的计算、存储和通信能力辅助部署稀疏的 CDN 末梢服务器以优化应急监控视频的全网传输。在此模型中, CDN 末梢服务器即为边缘区域的边缘云中心, 其上与云中心相连, 下连接到区域内各边缘簇。边缘簇内各边缘节点可部署于微基站、宏基站及高于宏基站的位置, 应急通信中可部署于紧急通讯车、无人机等设备。在本文当前构建的应急监控视频边缘智能传输网络架构下, 边缘节点选取为微基站、宏基站等设备, 其之间是进行有线传输的。

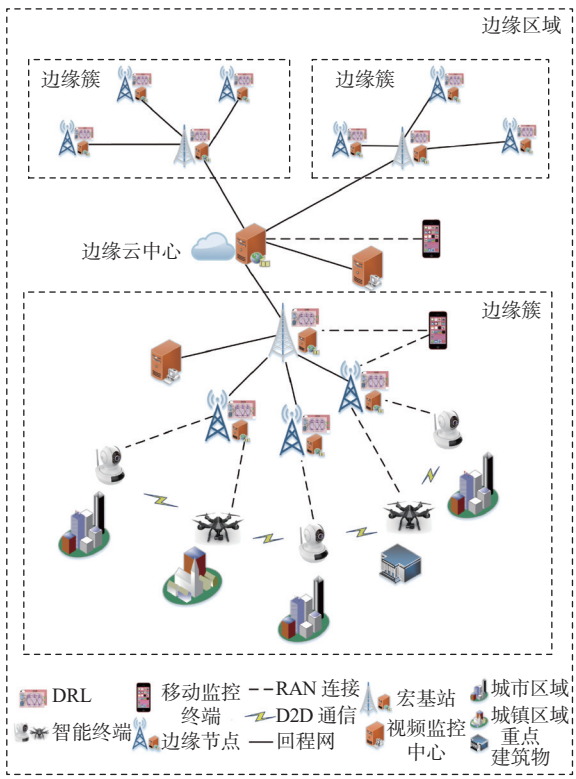


图 1　应急监控视频边缘智能传输网络拓扑结构

Fig. 1　Topology of edge-intelligent transmission networks for emergency surveillance video

由于边缘节点数量较多且彼此间的能力差异较大, 本文基于区域位置, 提出将边缘节点分簇的层次化管理模式。边缘节点分簇主要依据地理位置分布、边缘节点能力、用户数量与活跃度等影响因素。边缘节点分簇算法原则具体如图 1 所示。

1) 位置分布原则。边缘簇划分首先基于地理位置为第一原则, 同一区域的边缘节点划分在同一个边缘簇内。

2）能力平衡原则。簇内边缘节点的总体服务能力应与其区域内应急监控视频采集、处理和传输需求相匹配，不同簇之间的边缘智能服务能力在一定程度上保持均衡。

3）满足需求原则。如果边缘节点位于2个簇的交叉区域内，则基于边缘簇内视频监控中心和移动监控终端的数量来为其选择数量更多的边缘簇，以尽可能满足应急视频监控的需要。

4）分层管理原则。分簇完成后，根据边缘节点的计算、存储、通信能力，以及边缘节点之间的通信延时来选举簇头，其余则为簇元。边缘簇之间由簇头进行通信，建立了云中心-边缘区域（边缘云）-边缘簇（簇头）-边缘节点的分层网络结构。

根据以上原则，提出以下边缘节点分簇算法：

输入：边缘节点数据集 $E=\{E_1, E_2, \cdots, E_X\}$；边缘簇 $C=\{C_1, C_2, \cdots, C_Y\}$。

输出：Y 个簇的集合；从 E 中的宏基站边缘节点随机选取 Y 个对象作为边缘簇中心 E_{C_Y}。

1. Repeat:

2. FOR $j = 1, 2, \cdots, X$ DO

3. FOR $i = 1, 2, \cdots, Y$ DO

4. 计算边缘节点 E_j 到各簇中心 E_{C_i} 的距离

5. IF $|E_j - E_{C_i}| < \psi$: $E_j \in C_i$

6. ELIF $|E_j - E_{C_i}| = \psi$:

计算各边缘簇服务终端数量，将边缘节点 E_j 分配给终端数量最多的簇

7. ELSE

计算此时各边缘簇计算、存储和通信等服务能力 A_i，将剩余的边缘节点 E_j 分配给服务能力最小的边缘簇

8. END FOR

9. END FOR

10. UNTIL 各簇区域分布合理，服务能力 A_i 均衡

在应急监控视频边缘智能传输网络架构中，一方面其与主流的CDN技术兼容，可以借鉴其缓存替换、内容预取和负载均衡等方面的研究成果；另一方面能利用边缘集群协作能力进一步提升应急监控视频传输的质量。上述网络结构可支持以下协作模式：“云-边”协作，“边-边”协作，“边-端”协作，“端-端”协作。

2.2 传输流程

在本文构建的应急监控视频边缘智能传输网络场景下，应急监控视频的传输可以分为监控视频采集、缓存、智能传输决策、传输和监控终端播放5个阶段。

1）采集。在监控区域通过多个智能摄像头获取监控视频，并设置监控视频编码格式和比特率版本。

2）缓存。根据智能摄像头的缓存能力大小，可选择将监控视频缓存在智能摄像头终端，或者缓存到最邻近的边缘节点服务器，为后续监控视频传输做好准备。

3）智能传输决策。根据监控终端网络条件和需求，结合边缘簇内各边缘节点计算能力和服务负载情况，通过DRL算法感知决策，确定优先传输视频队列、传输视频比特率版本和实施转码任务边缘节点等信息，为重要性高的监控视频优先提供自适应的高质量视频传输。

4）传输。根据监控视频智能传输决策结果，由实施转码任务边缘节点提供应急监控视频转码和传输服务，并将视频传输到相应的监控终端。

5）监控终端视频播放。监控中心终端或移动监控终端依据传输视频的顺序，实现多摄像头监控视频的自动轮播。

3 基于IcD-FDRL的应急监控视频边缘智能传输优化方法

在第2节应急监控视频边缘智能传输网络中，应急监控视频传输的智能决策优化方法可以采取边缘节点独立决策方式和边缘节点联邦决策方式。本节从应急监控视频重要性度量和分组方法、系统模型、问题建模、DRL算法和IcD-FDRL算法5个方面阐述本文方法主要内容。

3.1 应急监控视频重要性度量和分组方法

基于移动监控终端和视频监控中心的监控视频调阅观看数据特点，结合监控视频内容，对监控视频的重要性进行度量，便于更重要的监控数据能优先显示在监控终端。

应急监控视频重要性主要与监控终端点击查看情况、规定重点关注区域和监控对象异常信息等因素有关。假设边缘区域某一边缘簇内有 K 个智能监控摄像头和 N 个边缘节点服务器。则定义应急监控视频的重要度为

$$\begin{cases} V_{k,t}^{\mathrm{imp}} = \alpha P_{k,t}^{\mathrm{popul}} + \beta P_{k,t}^{\mathrm{atten}} + \delta P_{k,t}^{\mathrm{objec}} \\ \quad \text{s.t.} \quad \alpha + \beta + \delta = 1 \\ \quad 0 \leqslant P_{k,t}^{\mathrm{popul}}, P_{k,t}^{\mathrm{atten}}, P_{k,t}^{\mathrm{objec}} \leqslant 1 \end{cases} \tag{1}$$

式中：$P_{k,t}^{\mathrm{popul}}$、$P_{k,t}^{\mathrm{atten}}$、$P_{k,t}^{\mathrm{objec}}$ 分别为应急监控视频监控终端点击流行率、重点区域关注率、目标信息监测率；α、β、δ 为相应参数系数；k 为智能监控摄像头序号；t 为具体时刻。

监控终端点击流行率是指视频监控中心和移动监控终端在观看应急监控视频时，点击查看 K 个

智能监控摄像头应急监控视频的概率。重点区域关注率是指应急视频监控过程中会根据需要设定重要监控区域, 其对应摄像头采集的监控视频重点区域关注率将会更高。目标信息监测率是指通过自动识别算法监测到应急监控视频中含有目标异常信息的概率, 此值越大说明存在的安全风险越大, 视频更为重要, 需要尽快进行高质量传输, 以及实时进行预警和处理。

为更好地实现应急监控视频高效传输, 基于上述应急监控视频重要度, 本文将应急监控视频根据重要度进行视频分组, 分为 I 级、II 级和 III 级, 级别越高, 表示视频紧要程度更高。边缘节点服务器依据传输决策决定转发传输哪一个分组虚拟队列。具体视频分组类型判断标准如下:

Ⅲ级:

$$V_{k,t}^{\mathrm{imp}} \geqslant 0.6 \tag{2}$$

Ⅱ级:

$$0.3 \leqslant V_{k,t}^{\mathrm{imp}} < 0.6 \tag{3}$$

Ⅰ级:

$$0 \leqslant V_{k,t}^{\mathrm{imp}} < 0.3 \tag{4}$$

同时, 结合应急监控视频码率大小创建了视频传输虚拟队列, 定义如下:

$$Q_k^T(t) = h(V_{k,t}^{\mathrm{imp}}, R_k(t), C_{\mathrm{s}}(t)) \tag{5}$$

式中: $R_k(t)$为应急视频流的带宽消耗(即视频编码码率); T 为视频分组类型; $h(x)$ 为映射函数; $C_{\mathrm{s}}(t)$为视频传输流畅度, 其与监控视频播放端缓冲区的重新缓冲情况密切相关, 当视频播放速率大于视频传输速率时, 缓冲区将减小, 如果在下一个区块到达之前播放了多个区块, 那么缓冲区将耗尽, 并进行重新缓冲, 将致使视频不能流畅播放。

视频传输流畅度(视频播放缓冲区占用率)表达式为

$$C_{\mathrm{s}}(t) = \frac{B_{\mathrm{target}}(t)}{W(t)} \tag{6}$$

式中: $B_{\mathrm{target}}(t)$为边缘网络提供给监控用户的视频比特率; $W(t)$为监控用户自身网络条件下的最大网络速率。

3.2　系统模型

通常情况下, MEC 网络下视频传输优化方法均需建立视频传输优化模型, 给出优化目标函数。根据视频传输优化目标的不同, MEC 网络下智能视频传输优化方法分为面向网络 QoS 的边缘智能视频传输优化和面向用户 QoE 的边缘智能视频传输优化两大类。而在应急监控视频传输中, 更为重要的是监控终端用户的视频播放质量和流畅度(即用户体验质量 QoE), 以满足应急事件及时处理和准确决策的需要。在本文应急监控视频传输架构下, 是基于边缘簇为监控终端用户提供多摄像头重要监控视频的自动轮播, 以实现为重要性高的监控视频优先提供自适应的低时延、低成本和高质量传输。因此, 本文聚焦于面向监控终端用户 QoE 的边缘智能视频传输优化方法研究, 需要全面考虑监控用户播放端网络状态。本节主要介绍视频质量、视频缓存、视频转码、视频传输等模型, 包括播放端视频传输速率、重新缓冲时间、播放时延等。

1) 视频质量模型。在视频处理中, 峰值信噪比度量是在原始帧和压缩帧之间提供客观质量评估的事实标准[38]。视频的峰值信噪比曲线可以作为视频编码速率的函数, 即如果源对等端以这些速率编码和发送其视频, 则可以通过视频编码码率来表示视频质量。一般来说, 视频的比特率版本决定了视频的质量, 而较高比特率的视频具有较高的视频质量。设定时刻 t 时边缘网络提供给监控用户的应急监控视频质量为

$$Q(t) = \rho \ln(B_{\mathrm{target}}(t)) \tag{7}$$

式中: ρ为可以在监控视频编码采集时获得的常数。

2) 视频缓存模型。应急监控视频的缓存功能主要考虑 2 个方面: 视频采集端的视频缓存和视频播放端的视频缓存。在采集端, 应急监控视频优先缓存在智能摄像头, 智能摄像头缓存资源不足时, 应急监控视频被缓存到邻近边缘节点。在播放端, 需要考虑视频播放的重新缓冲情况, 高质量的监控视频传输应该没有重新缓冲情况的出现。重新缓冲通常发生在缓冲区占用率较低的情况下, 即下载新视频块的时间 $D(t)$ 大于缓冲视频块的播放时间 $T_{\mathrm{play}}(t)$。则视频重新缓冲时间 $T_{\mathrm{rebuf}}(t)$ 可以定义为

$$T_{\mathrm{rebuf}}(t) = \max(D(t) - T_{\mathrm{play}}(t), 0) \tag{8}$$

下载新视频块的时间 $D(t)$ 和缓冲视频块的播放时间 $T_{\mathrm{play}}(t)$ 均可根据视频块数据量大小除以当前传输速率计算得到。

3) 视频转码模型。应急监控视频在提供给视频监控终端播放时, 需要根据播放端用户需求、网络条件等采取自适应视频转码。视频转码主要考虑视频比特率版本转码和视频编码格式转码, 其转码成本定义为

$$C_{\mathrm{transc}}(t) = \mu_1(B_{\mathrm{input}} - B_{\mathrm{target}})DN_{\mathrm{cpu}}^{\mathrm{b}} + \mu_2 N_{\mathrm{cpu}}^{\mathrm{f}} \tag{9}$$

式中: B_{input}和B_{target}分别为转码源视频比特率版本和转码目标视频比特率版本; $N_{\mathrm{cpu}}^{\mathrm{b}}$和$N_{\mathrm{cpu}}^{\mathrm{f}}$分别为完成比特率版本转码和格式转码所需的 CPU 核数; D 为转码视频块长度; μ_1、μ_2为可调比重参数。

4）视频传输模型。在应急监控视频传输过程中，主要考虑监控视频播放时延、传输带宽成本和监控视频传输准确率。视频播放时延主要考虑视频重新缓冲时间和排队时延，其计算式为

$$T_{\text{delay}}(t)=T_{\text{rebuf}}(t)+T_{\text{queue}}(t) \tag{10}$$

其中，排队时延$T_{\text{queue}}(t)$仅在边缘节点负载相对比较紧张时才予以考虑。

传输成本可以从用户播放端累计计算：

$$C_{\text{transm}}(t)=D\sum_{i=1}^{N_{\text{user}}}B_{\text{target}}^{i}(t) \tag{11}$$

应急监控视频优先传输准确率根据应急监控视频重要度和传输排队顺序来具体计算。假设在一个学习回合内需要提供应急监控视频传输决策次数为N_{episode}，则定义视频优先传输准确率$P_{\text{acc}}(t)$为

$$P_{\text{acc}}(t)=\left(\sum_{j=1}^{N_{\text{episode}}}l\left(G_{Q_k^T}^{j}(t)\right)\right)/N_{\text{episode}} \tag{12}$$

式中：$G_{Q_k^T}^{j}(t)$为时刻t转发传输的视频分组序列；$l(\cdot)$为指示函数，即

$$l\left(G_{Q_k^T}^{j}(t)\right)=\begin{cases}1 & G_{Q_k^T}^{j}(t)\in Q_k^T(t)\ \text{且}\ T\in\{\text{III}\}\\ 0 & G_{Q_k^T}^{j}(t)\in Q_k^T(t)\ \text{且}\ T\in\{\text{I,II}\}\end{cases} \tag{13}$$

式(13)表示此次应急监控视频传输中重要度最高的监控视频序列获得了优先传输则取值为1，否则为0。

3.3 问题建模

基于上述模型，结合监控终端的需求和边缘网络环境，本文采用DRL算法来解决复杂网络环境下应急监控视频传输问题。首先，建立DRL算法中的状态、动作和奖励等基本元素。然后，给出QoE计算方法和优化目标函数。

1）状态变量。根据建立的系统模型，应急监控视频传输环境与视频播放缓冲区占用率、视频播放时延、视频质量和传输带宽成本等因素密切相关，因此，状态变量可定义为

$$S(t)=\left\{C_{\text{s}}(t),T_{\text{delay}}(t),Q(t),C_{\text{transm}}(t)\right\} \tag{14}$$

2）动作变量。通过智能体与环境的不断交互学习，最终需要进行感知决策，决定转发传输哪一个视频分组序列，并给定实施转码和服务的边缘节点，以及应急监控视频传输的比特率版本和编码格式。因此，动作变量可以定义为

$$A(t)=\left\{G_{Q_k^T}(t),M(t),B_{\text{target}}(t),F_{\text{target}}(t)\right\} \tag{15}$$

3）奖励函数。在应急监控视频传输中，主要是为实现重要应急视频的高质量、低延时和低成本的传输。因此，奖励函数主要包含视频质量、视频质量变化、视频优先传输准确率、播放延时和传输成本等因素。其可定义为

$$R(t)=\omega_1Q(t)+\omega_2P_{\text{acc}}(t)-\omega_3|Q(t)-Q(t-1)|-\omega_4T_{\text{delay}}(t)-\omega_5C_{\text{transc}}(t)-\omega_6C_{\text{transm}}(t) \tag{16}$$

式中：ω_i为可变权值参数。

4）QoE。基于上述奖励函数，结合用户对应急监控视频的播放需求，QoE值定义为

$$Q_{\text{oE}}(t)=\omega_1Q(t)-\omega_3|Q(t)-Q(t-1)|+\omega_2P_{\text{acc}}(t)-\omega_4T_{\text{delay}}(t) \tag{17}$$

5）优化目标函数。根据DRL算法理论，为获得模型的最优解，要使长期收益函数最大化，综合考虑应急监控视频传输的高质量、低延时和低成本。即

$$\begin{cases}\max\limits_{G_{Q_k^T}(t),M(t),B_{\text{target}}(t),F_{\text{target}}(t)}J(t)=E\left(\sum\limits_{t=0}^{T-1}\gamma^tR(t)\right)\\ \text{s.t.}\ \ \text{C1}:M(t)\in\{0,1,\cdots,N_{\text{MEC}}\},\forall t\\ \qquad\text{C2}:G_{Q_k^T}(t)\in Q_k^T(t),\forall t\\ \qquad\text{C3}:B_{\text{target}}(t)\in\{B_1,B_2,\cdots,B_{\max}\},\forall t\\ \qquad\text{C4}:\ \gamma\in(0,1]\end{cases} \tag{18}$$

3.4 DRL算法

强化学习的核心机制是用试错来学会在给定的情境下选择最恰当的行为动作，而从环境状态到动作的映射学习即为策略。智能体自动寻找在时间序列里的最优策略，而最优策略通常是指最大化长期累积奖励期望，如目标函数式(18)所示。

自DQN[39-40]问世以来，在DRL领域出现了大量用于解决实际应用问题的算法和论文。强化学习的数学基础理论是基于马尔可夫性质的马尔可夫决策过程（Markov decision processes, MDP）。在Q学习算法中，其基本思想是通过使用Bellman方程作为迭代更新来估计Q值函数，这种值迭代算法收敛于最优Q值函数：

$$\begin{cases}Q_{i+1}(s,a)=E\left(r+\gamma\max_{a'}Q_i(s',a')|s,a\right)\\ \qquad\text{s.t.}\ \ i\to\infty\text{时},\ \ Q_i\to Q^*\end{cases} \tag{19}$$

但在实际应用中，当状态空间是高维空间时，这种基本方法会面临维度灾难，是完全不切实际的，几乎无法估计Q值函数。针对这个问题，通常可以使用非线性函数逼近器来估计Q值函数，如神经网络。因此，将Q学习与深度神经网络相结合，实现一系列动作选择决策，便是DQN的基本思想出发点。在DQN中，Mnih等使用权值作为Q网络的神经网络函数逼近器[40]，可通过最小化每次迭代时变化的损失函数序列来训练Q网络。即

$$\begin{cases}L_i(\theta_i)=E_{s,a\sim\rho(\cdot)}\left[(y_i-Q(s,a;\theta_i))^2\right]\\ y_i=r+\gamma\max\limits_{a'}Q(s',a';\theta_{i-1})|s,a)\end{cases} \tag{20}$$

式中: y_i为 Q 现实网络和 Q 估计网络联合迭代的 Q 现实值; $Q(s,a;\theta_i)$为 Q 估计网络每次状态时迭代的 Q 估计值; $\rho(\cdot)$为状态和动作的概率分布, 也称之为行为分布。在 DQN 算法中, 通过经验回放、固定 Q 现实参数和贪心算法, 消除学习样本之间的相关性, 减少策略的振荡和分歧, 控制算法的开发与探索, 使得 DQN 算法的学习效率越来越高。

基于 DQN 算法在处理离散状态空间和动作空间时的优异性能, 本文采用 DQN 算法学习集群 MEC 网络内应急监控视频传输和转码策略, 其实现过程如图 2 所示。首先, Q 估计网络根据当前状态 $S(t)$学习得到 Q 估计值 $Q(s,a;\theta_i)$, 选取 Q 估计中最大值的动作 $A(t)$来换取环境中的奖励 $R(t)$; 然后, 基于下一状态 $S(t+1)$和当前奖励 $R(t)$通过 Q 现实网络计算得到正确的 Q 现实值; 最后, 利用上述 Q 估计值和 Q 现实值来计算得到损失函数 $L_i(\theta_i)$, 并通过最小化每次迭代时变化的损失函数序列来训练 Q 估计网络, 以此不断更新和迭代, 直到满足迭代终止条件。

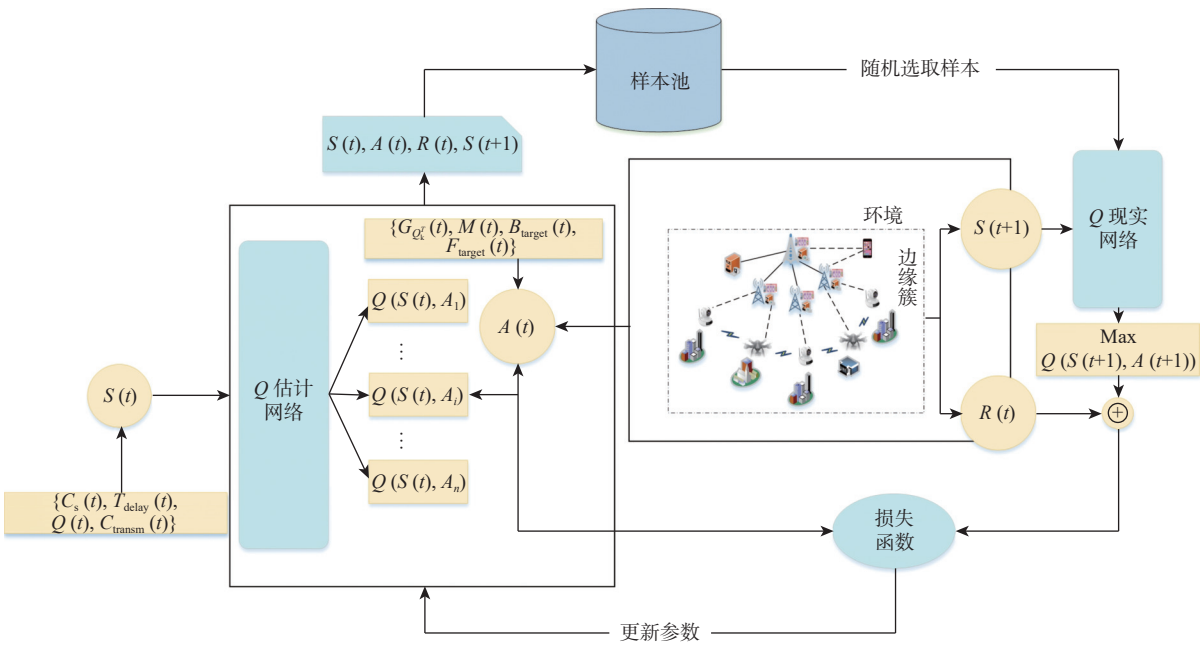


图 2　基于 DQN 算法的应急监控视频智能传输和转码优化

Fig. 2　Intelligent transmission and transcoding optimization of emergency surveillance video based on DQN algorithm

根据实现过程框图, 设计基于 DRL 的应急监控视频边缘智能传输优化算法 (简称 DRL-EIT) 具体流程如下:

1. 初始化:

2. 设置记忆存储器容量为 N_D

3. 用随机权重初始化 Q 估计网络和 Q 现实网络

4. 初始化用户应急监控的服务记录矩阵 $\boldsymbol{S}_{\text{es}}$

5. 初始化最大迭代步数 M 和一个回合内次数 T

6. FOR episode = 1, 2, ⋯, M DO

7. 生成用户应急监控视频传输请求数据

8. 根据式 (14), 获得初始状态 S_1

9. FOR t = 1, 2, ⋯, T DO

10. 随意给定一个随机概率 $\varsigma \in [0,1]$

11. 按照下式选取式 (15) 的动作

$$A(t)=\begin{cases} a^*(t)=\arg\max\limits_a Q(s,a;\theta) & \varsigma>\varepsilon \\ a(t)\neq a^*(t), \text{randomly select } a(t) & \text{其他} \end{cases}$$

12. 根据智能体选取的动作 $A(t)$, 执行应急监控视频传输决策 $\left\{G_{Q_k^T}(t), M(t), B_{\text{target}}(t), F_{\text{target}}(t)\right\}$

13. 获得奖励 $R(t)$, 进入下一个状态 $S(t+1)$

14. 存储记忆经验 $(S(t), A(t), R(t), S(t+1))$ 到重放记忆存储器

15. 更新用户应急监控服务记录矩阵 $\boldsymbol{S}_{\text{es}}$

16. 每隔固定步数在重放记忆存储器中随机选取小批量样本 $(S(t), A(t), R(t), S(t+1))$

17. 根据式 (20) 执行梯度下降算法, 其中令

$$y_j=\begin{cases} r_j & \text{for terminal } s' \\ r_j+\gamma\max\limits_{a'}Q(s',a';\theta_{i-1})|s,a) & \text{for non-terminal } s' \end{cases}$$

18. 更新 Q 估计网络中的参数

19. 每 G 步将 Q 估计网络参数赋给 Q 现实网络

20. 更新学习率

21. END FOR

22. END FOR

3.5　IcD-FDRL 算法

DRL 被广泛应用到各类应用场景下的复杂优

化问题中，而且取得了不错的效果。但是DRL仍存在弱点和局限，包括不稳定性、易陷入局部最优、奖励函数敏感且设计困难、泛化性弱等问题。针对不稳定性和易陷入局部最优，其中，一种解决思路是通过更多不相干的训练数据来提升学习效率，促进算法尽快获取全局最优。但是，更多的训练数据往往就会需要更多的网络传输成本消耗和存在数据隐私安全威胁，联邦学习的提出为这一思路提供了解决办法。

联邦学习本质是一种分布式机器学习框架，在保障数据隐私安全的基础上，可以实现训练数据共享，共同训练优化模型。将联邦学习融入DRL的实施中，形成FDRL，以解决训练数据孤岛的问题。然而，联邦学习中恶意的边缘节点很容易导致联邦学习产生梯度爆炸，影响学习效率。

基于2.1节所构建的应急监控视频边缘智能传输网络架构，本文IcD-FDRL的应急监控视频边缘智能传输优化架构，如图3所示。MEC网络基于边缘簇提供应急监控视频缓存和转码服务，针对区域内的用户需求，动态选择簇内可靠边缘节点共同参与实现联邦学习，每个边缘节点通过DRL-EIT算法进行学习训练，并上传簇头进行模型聚合，再将新的模型参数下载到各边缘节点，以此不断更新和迭代，直到满足迭代终止条件。各边缘节点依据当时最新神经网络模型来确定优先传输视频队列、传输视频比特率版本和实施转码任务边缘节点等信息，并为监控中心终端或移动监控终端提供应急监控视频的传输服务。

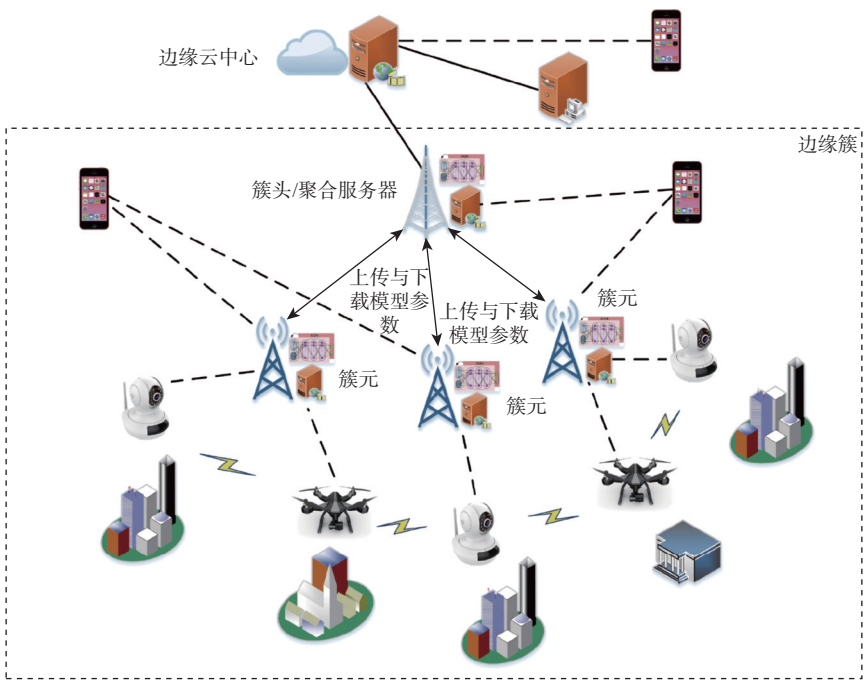


图3 边缘簇内基于FDRL的应急监控视频边缘智能传输优化架构

Fig. 3 Architecture for edge-intelligent transmission optimization of emergency surveillance video based on FDRL in edge cluster

在FDRL过程中，为杜绝边缘簇内低质量的恶意边缘节点的参与，在每轮联邦学习神经网络模型聚合之前，先根据各边缘节点的本轮训练结果选取允许参与联邦学习训练的可靠边缘节点，再根据聚合公式进行模型聚合计算，以实现簇内动态的FDRL算法。基于动态FDRL的应急监控视频边缘智能传输优化算法(IcD-FDRL-EIT)流程如下：

1. 初始化：

2. 设置边缘簇 C 中能够参与联邦学习边缘节点 E_i 的数量为 N_C

3. 初始化联邦学习最大学习轮数 R

4. 初始化用户应急监控的服务记录矩阵 $\boldsymbol{S}_{\text{es}}$

5. 用随机权重初始化聚合服务器 Q 估计网络和 Q 现实网络

6. FOR round = 1, 2, ⋯, R DO

7. 下载聚合服务器神经网络模型参数 θ'_{aggre} 到各边缘节点

8. 各边缘节点使用本地数据执DRL-EIT算法

9. 得到本地神经网络模型参数 θ^i_{local} 和损失函数 L^i_{local}

10. 将本轮本地训练模型参数 θ^i_{local} 上传到簇头

11. 根据以下要求选取参与联邦学习训练的边缘节点 E_i

$$L^i_{\text{local}} \leqslant \lambda \frac{1}{N_C} \sum_{i=0}^{N_C} L^i_{\text{local}}$$

12. 执行下面聚合公式进行模型聚合计算

$$\theta_{\text{aggre}}=\left(\sum_{i=1}^{N_{\text{E}}}\theta_{\text{local}}^{i}\right)\Big/N_C$$

13. 用θ_{aggre}初始化聚合服务器神经网络模型

14. 使用聚合服务器本地数据执行 DRL-EIT 算法

15. 更新聚合服务器神经网络模型参数θ'_{aggre}

16. END FOR

4　实验分析

本节对本文算法通过 Python 集成开发环境进行模拟仿真实验, 并从收敛性、训练效率、时延性、传输准确率、传输质量和成本等方面进行实验分析。

本文 IcD-FDRL-EIT 算法中涉及到 MEC 网络环境相关变量, 选择的对比算法需要转化应用到当前网络模型下。为体现对比算法的有效性和针对性, 本文选取下面 2 种基于集群 MEC 网络环境下的相关算法作为对比算法。针对文献 [30] 中的算法, 其构建了集群移动边缘网络下基于 DRL 的协同视频缓存和转码算法, 将其转化应用到当前应急监控视频传输实验环境下, 记作 DRL-CCT 算法。文献 [9] 在集群 MEC 网络环境下构建了基于 FDRL 的缓存和码率自适应算法, 将其相应适用到当前应急监控视频传输网络架构下, 记作 FDRL-CBA 算法。然后, 将本文提出的 IcD-FDRL-EIT 算法与上述 2 种算法进行算法性能对比分析。

4.1　仿真设置

1) 传输网络设置

在本文构建的应急监控视频边缘智能传输网络架构下, 边缘节点选择部署于微基站、宏基站等位置。由于边缘节点服务器为监控终端用户提供视频监控服务是无线接入网络传输, 而参与联邦学习边缘节点之间是进行有线传输的, 时延很低且不占用边缘节点为监控终端用户提供的无线传输带宽, 所以没有考虑联邦学习带宽消耗和时延。

本文对一个边缘簇内边缘节点进行应急监控视频智能传输模拟仿真实验。簇内含有 7 个边缘节点, 每个边缘节点配置连接 10 个智能摄像头, 为 30 个(N_{user})移动监控终端提供应急监控视频传输服务。采集端和播放端视频编码格式均使用 H.265, 比特率版本分为低清、标清、准高清、高清、4K 这 5 种, 其对应的低码率编码码率分别为 0.25, 0.5, 1, 2, 16 Mbit/s。

2) DQN 网络设置

本文通过 TensorFlow 构建了一个 4 层神经网络结构(4, 256, 512, 105), 其中, 2 个全连接隐藏层的大小分别为 256 和 512, 采用的激活函数均为 Relu 函数, 输入层神经元个数为 4, 输出层神经元个数是应急监控视频分组、簇内边缘节点和视频比特率版本的全部组合个数为 105。实验中设置使用的损失函数为均方误差, 经验池大小为 2 000, 随机梯度下降的小批量大小为 32。在训练过程中, 随着学习的进行, 不断降低贪心算法的探索概率, 以促进传输优化算法尽快收敛。

3) 参数设置

仿真实验中各参数的取值情况如表 1 所示。

表 1　实验中主要参数设置

Table 1　Setting of main parameters in experiments

参数	取值	参数	取值
α	0.4	ω_5	0.1
β	0.3	ω_6	0.1
δ	0.3	N_c	7
ρ	7	K	10
学习率	0.01	N_{user}	30
μ_1	1	$\{N_{\text{cpu}}^b\}$	{2,4,6,8}
μ_2	0	D/s	10
ω_1	1.2	γ	0.9
ω_2	1.2	ε	0.9
ω_3	1.0	ψ	600
ω_4	1.0	λ	0.65

4) 训练数据生成和训练设置

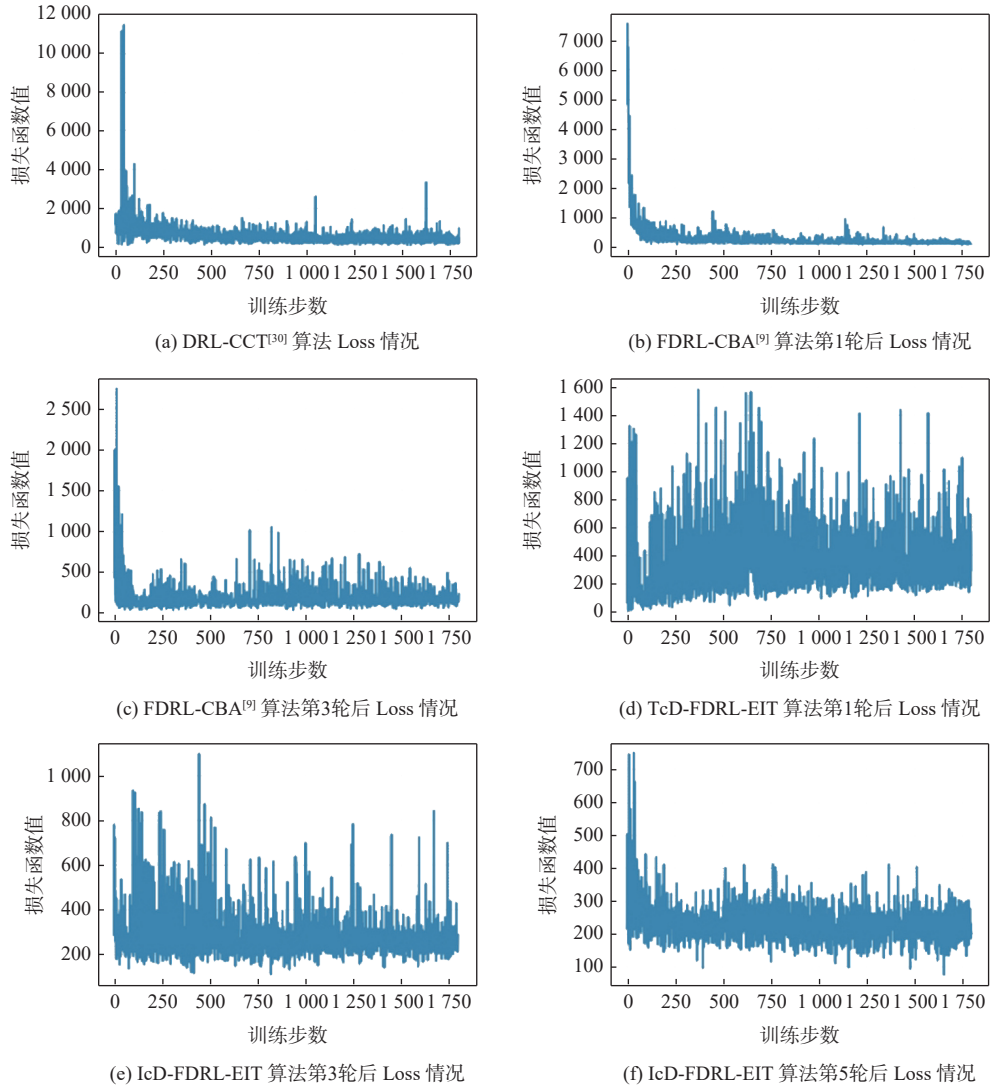
模拟仿真实验中, 监控终端网络情况和训练数据通过程序生成。监控终端用户应急监控视频点击流行率服从 Zipf 分布, 应急监控视频的重点区域关注率、目标信息监测率由服从随机分布。在仿真过程中, DRL-CCT 算法中的 DQN 运行 40 个回合, 其中, 一个回合包含 50 次监控视频服务。在 IcD-FDRL-EIT 和 FDRL-CBA 算法中, 联邦学习迭代 5 轮, 簇头边缘节点上 DQN 运行 40 个回合, 其余边缘节点上 DQN 运行 12 个回合, 每个回合也包含 50 次监控视频服务。

4.2　实验结果分析

由于 DRL 的不稳定性, 实验结果均取 30 次运行结果的平均值。本文从收敛性、训练效率、时延性、传输准确率、传输质量和成本 5 个方面进行实验仿真结果分析。由于是模拟仿真实验, 针对时延、传输质量和成本没有设置量纲, 主要比较各评价指标的大小变化和趋势情况。

1) 收敛性

图4 为IcD-FDRL-EIT 算法与DRL-CCT 和FDRL-CBA 算法的 Loss 函数收敛情况对比。由图可知, 相比 DRL-CCT 和 FDRL-CBA 算法, 本文算法在第 1 轮联邦学习迭代之后 Loss 函数就已经快速收敛, 并且在之后联邦学习的不断迭代过程中 Loss 函数



(a) DRL-CCT[30] 算法 Loss 情况　(b) FDRL-CBA[9] 算法第1轮后 Loss 情况

(c) FDRL-CBA[9] 算法第3轮后 Loss 情况　(d) TcD-FDRL-EIT 算法第1轮后 Loss 情况

(e) IcD-FDRL-EIT 算法第3轮后 Loss 情况　(f) IcD-FDRL-EIT 算法第5轮后 Loss 情况

图 4　Loss 函数收敛情况对比

Fig. 4　Comparison of Loss function convergence

取值范围迅速缩小，说明动态选择参与联邦学习的边缘节点，促使联邦学习效率得到有效提升。这是由于动态选择簇内边缘节点，促进了训练样本数据的有效性，加强了 DRL 的学习效率，有力地促进了算法的快速收敛。

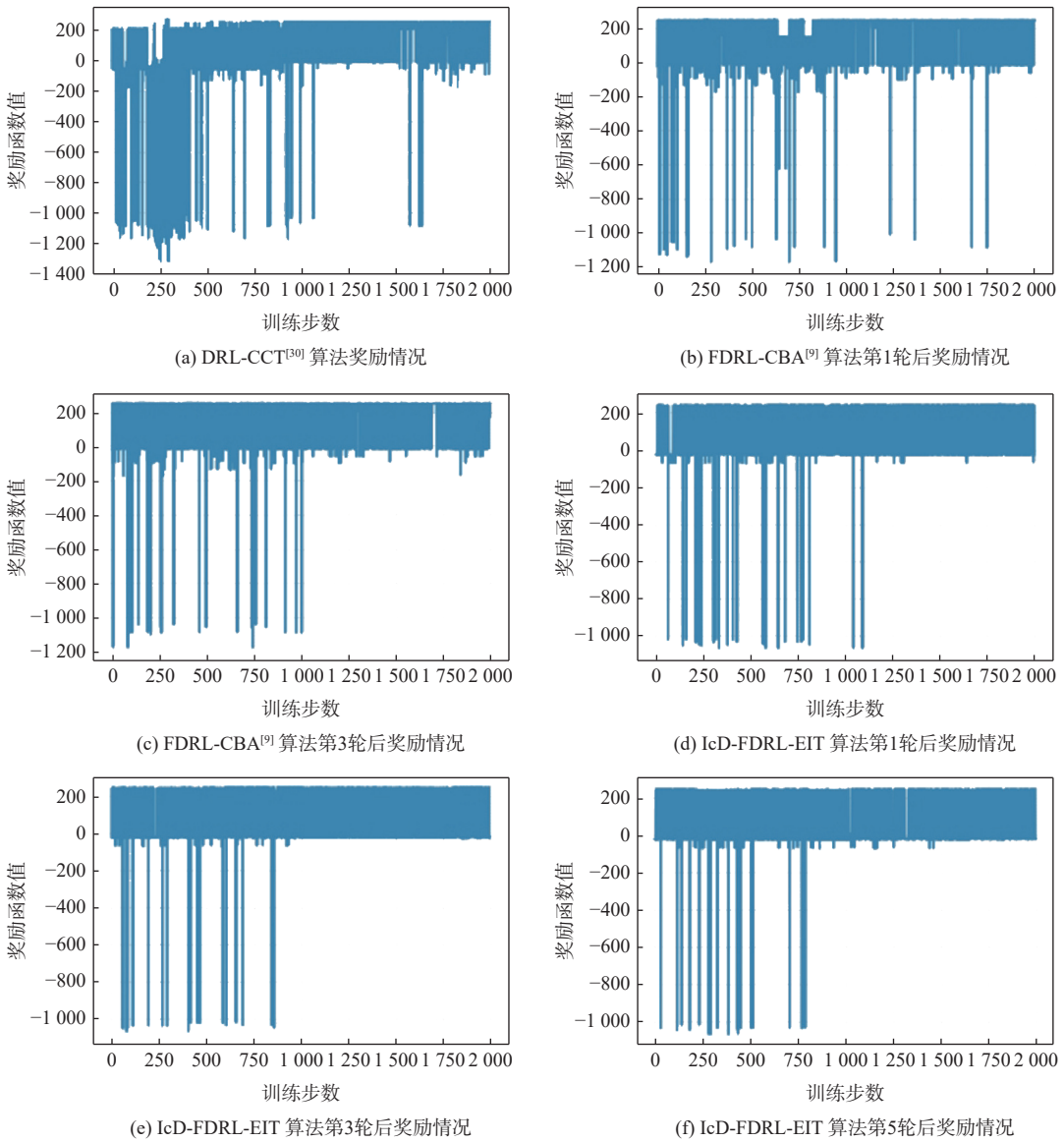
2）训练效率

本文从算法奖励函数的稳定性来分析本文算法的有效性。一般来说，奖励稳定性越好说明强化学习算法训练效率越高，越能尽快获得全局最优，抖动性越大说明强化学习算法训练效率较低，容易陷入局部最优。图 5 为 3 种算法的奖励稳定性情况对比，与其他 2 种算法相比，IcD-FDRL-EIT 算法的稳定性更好。随着簇内动态联邦学习的进行，IcT-FDRL-EIT 算法奖励函数中的波动条变得越来越稀疏、越来越少，并且达到稳定状态的速度更快。上述表明，本文 IcD-FDRL-EIT 算法的奖励函数稳定性更好，可以更快速获得全局最优，并趋于保持稳定。

3）时延性

本文主要研究应急监控视频的传输优化，重点关注的是监控终端用户的视频体验质量。因此，时延性主要考虑用户接收端播放时延，即视频重新缓冲时间。

图 6 为应急监控视频服务过程中时延对比情况。根据仿真结果可以看出，DRL-CCT 算法的时延随着学习的进行不断慢慢的变小，而 IcD-FDRL-EIT 和 FDRL-CBA 算法的时延可以实现迅速变小，并且本文算法时延相对更低。为保障监控终端用户的视频体验质量，本文算法可以自适应减小播放端重新缓冲时间，获得高质量的应急监控视频，为应急事件处理提供坚实的支撑。本文基于监控终端用户的视频体验质量，主要考虑的播放端视频重新缓冲时间，但当终端用户数量快速增加时，集群边缘簇的缓存、计算和通信能力就可能出现不足的



(a) DRL-CCT[30] 算法奖励情况　(b) FDRL-CBA[9] 算法第1轮后奖励情况

(c) FDRL-CBA[9] 算法第3轮后奖励情况　(d) IcD-FDRL-EIT 算法第1轮后奖励情况

(e) IcD-FDRL-EIT 算法第3轮后奖励情况　(f) IcD-FDRL-EIT 算法第5轮后奖励情况

图 5　奖励函数稳定性对比

Fig. 5　Comparison of reward function stability

现象, 此时便要充分考虑排队时延对视频传输和播放质量的影响, 也是后续进一步深入研究的方向。

4) 传输准确率

本部分主要分析在本文网络架构下 3 种算法重要应急监控视频优先传输准确率情况。如表 2 所示, 在应急监控视频源(智能摄像头、无人机等)数量不同的情况下, 分别给出了 3 种算法应急监控视频优先传输准确率。总体上看, IcD-FDRL-EIT 算法的优先传输准确率均优于其他算法, 且都在 93% 以上。在摄像头终端数量为 140 个时, 应急监控视频优先传输准确率提高达 6% 以上。同时对每个边缘节点配置连接不同智能摄像头数量时的传输准确率情况进行了对比, 可以看出, 随着智能摄像头数量的增加, 应急监控视频数量也就增加, 导致传输准确率略有下降, 但 IcD-FDRL-EIT 算法的优先传输准确率仍是最高, 且相对更加稳定。这表明, 本文算法在环境参数变化时, 模型和算法仍能保持较稳定的性能, 具有一定的鲁棒性。并且, 随着联邦学习迭代轮数和 DQN 运行回合数的增加, 应急监控视频优先传输准确率还将进一步提升。

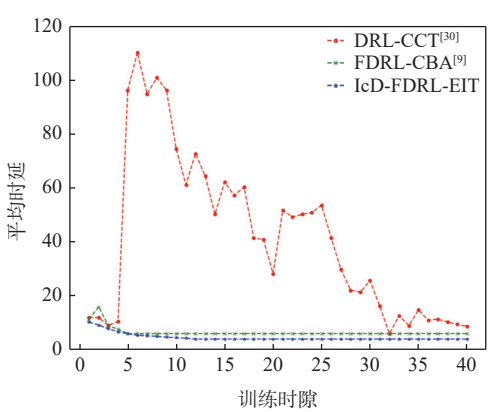


图 6　平均时延对比

Fig. 6　Comparison of average delay

表2 应急监控视频优先传输准确率对比

Table 2 Comparison of priority transmission accuracy of emergency surveillance video %

算法	摄像头数量为70	摄像头数量为140
DRL-CCT[30]	92.5	85.1
FDRL-CBA[9]	97.2	92.3
IcD-FDRL-EIT	98.3	93.5

5）传输质量和成本

根据文中系统模型，传输质量主要考虑移动监控终端的应急监控视频观看体验质量，即QoE值。而传输成本关注的是应急监控视频传输带宽成本和转码成本。图7和图8分别为3种算法的平均QoE和平均传输成本对比情况。

由图7和图8可知，随着学习的进行，3种算法的平均QoE均不断提升，平均传输成本均不断降低，足以说明DRL可以实现在对传输成本控制的基础上，追求监控终端QoE的不断提升，以期获得高质量、低成本的应急监控视频传输。同时，根据图中数据可以明确看出基于FDRL的2种算法的平均QoE提升和平均传输成本降低都比DRL-CCT算法迅速，说明联邦学习架构促进了DRL的快速收敛和高效训练。同时，相比其余2种算法，本文IcD-FDRL-EIT算法在降低平均传输成本和提升平均QoE值时更加平稳，并且性能更好，这与文中算法收敛性和训练效率分析也是一致的。

为更好地验证本文模型和算法的鲁棒性和泛化性，对边缘簇内不同边缘节点数量和不同监控终端数量时传输质量和传输成本进行了对比实验。

图9和图10分别为当监控终端数量为30，边缘簇内不同边缘节点数量时的平均QoE和平均传输成本对比。可以看出，边缘簇内边缘节点数量的变化对边缘簇内边缘节点的总体平均传输成本和平均QoE值影响不大。但是簇内边缘节点数量增加必然使得每个MEC服务器所承载的传输和计算

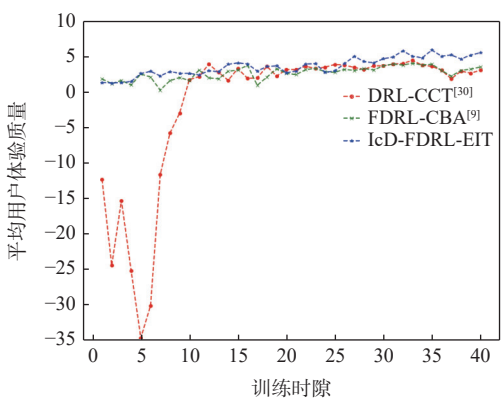


图7 平均QoE对比

Fig. 7 Comparison of average QoE

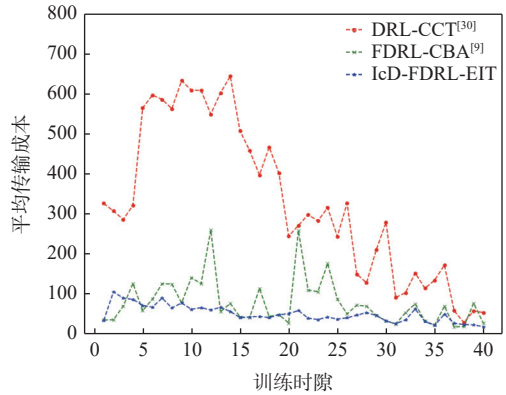


图8 平均传输成本对比

Fig. 8 Comparison of average transmission cost

压力减小，而簇内每个MEC服务器的负载均衡和资源利用率情况是需要下一步深入探究的内容。

边缘簇内MEC服务器数量为7时，不同监控终端数量下的平均QoE和平均传输成本对比如图11和图12所示。由图可知，终端数量为20时平均QoE值略高一点，但不同监控终端数量时，平均QoE值基本稳定。而随着监控终端数量的增加，平均传输成本是增加的，随着算法学习的进行，平均

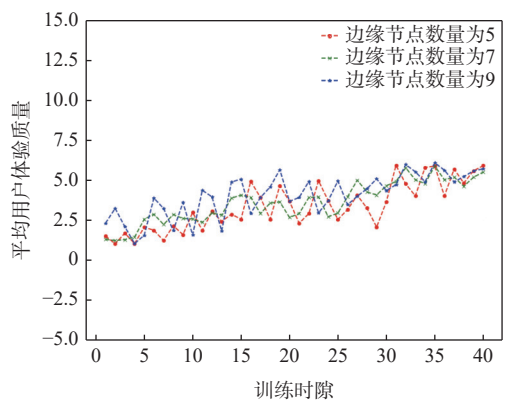


图9 边缘簇内不同边缘节点数量时平均QoE对比

Fig. 9 Comparison of average QoE for different edge nodes in edge cluster

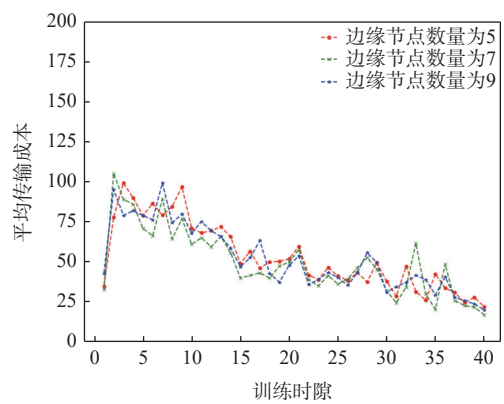


图10 边缘簇内不同边缘节点数量时平均传输成本对比

Fig. 10 Comparison of average transmission cost for different edge nodes in edge cluster

传输成本在逐步降低。综上所述, 本文算法能够在追求视频传输质量的同时, 尽量降低视频传输成本, 并且在不同网络环境下具有较好的鲁棒性。

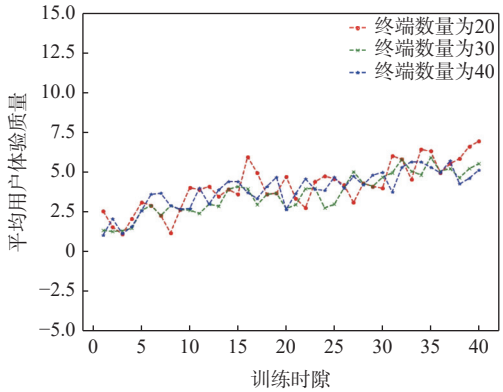


图 11　不同监控终端数量时平均 QoE 对比

Fig. 11　Comparison of average QoE for different monitored terminals

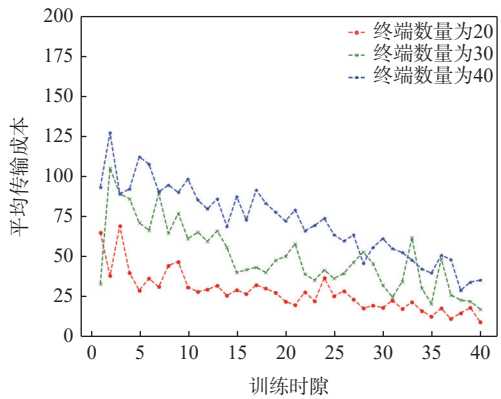


图 12　不同监控终端数量时平均传输成本对比

Fig. 12　Comparison of average transmission cost for different monitored terminals

5　结　论

1) 基于移动边缘计算网络构建了一种应急监控视频边缘智能传输架构, 用于搭建区域内的应急监控视频无线通信系统, 以满足突发事件监测、公共安全事件处理、灾后重建等应急通信需要。

2) 针对应急监控视频的特性, 提出应急监控视频重要性度量方法, 实现了对应急监控视频的重要性甄别, 促进应急监控视频优先传输准确率提高 6% 以上。

3) 针对应急监控视频传输, 基于集群分簇视频传输模型, 提出基于簇内动态联邦深度强化学习的应急监控视频边缘智能传输优化方法, 实现重要应急监控视频的优先和高质量传输。

4) 仿真实验结果表明, 本文算法收敛性、训练效率得到提升, 在应急监控视频的优先传输准确率、平均用户体验质量、平均延迟、平均传输成本等方面取得了较好的性能。

然而, 本文 IcD-FDRL-EIT 算法中采用的模型聚合更新方法仍是比较简单的联邦平均算法, 在模型聚合精度上, 还存在着提升的空间。同时, 应急监控视频的重要性度量目前主要考虑的监控终端点击查看情况、规定重点关注区域和监控对象异常信息等因素, 其度量因素可以根据实际情况需要进行补充和优化。未来将继续密切结合应急监控视频传输实际, 深入探索研究更有效的联邦学习模型聚合更新算法和重要性度量方法, 以充分满足生活中智慧应急通信的需要。

参考文献（References）

[1] 张铮, 李政华. 中国特色应急管理制度体系构建: 现实基础、存在问题与发展策略[J]. 管理世界, 2022, 38(1): 138-144.
ZHANG Z, LI Z H. Emergency management system construction with Chinese characteristics: realistic foundation, existing problems and development strategy[J]. Journal of Management World, 2022, 38(1): 138-144(in Chinese).

[2] 刘鲲. 智能视频监控中事件理解的关键技术研究[D]. 北京: 北京邮电大学, 2021.
LIU K. Research on key technologies of event understanding in intelligent video surveillance[D]. Beijing: Beijing University of Posts and Telecommunications, 2021(in Chinese).

[3] 黄韬, 刘江, 汪硕, 等. 未来网络技术与发展趋势综述[J]. 通信学报, 2021, 42(1): 130-150.
HUANG T, LIU J, WANG S, et al. Survey of the future network technology and trend[J]. Journal on Communications, 2021, 42(1): 130-150(in Chinese).

[4] 伏玉笋, 杨根科. 人工智能在移动通信中的应用: 挑战与实践[J]. 通信学报, 2020, 41(9): 190-201.
FU Y S, YANG G K. Application of artificial intelligence in mobile communication: challenge and practice[J]. Journal on Communications, 2020, 41(9): 190-201(in Chinese).

[5] 梁应敞, 谭俊杰, NIYATO D. 智能无线通信技术研究概况[J]. 通信学报, 2020, 41(7): 1-17.
LIANG Y C, TAN J J, NIYATO D. Overview on intelligent wireless communication technology[J]. Journal on Communications, 2020, 41(7): 1-17(in Chinese).

[6] LI Y. A suvey on edge intelligent video surveillance with deep reinforcement learning[J]. Journal of Network Intelligence, 2022, 7(1): 70-83.

[7] WANG X F, LI R B, WANG C Y, et al. Attention-weighted federated deep reinforcement learning for device-to-device assisted heterogeneous collaborative edge caching[J]. IEEE Journal on Selected Areas in Communications, 2021, 39(1): 154-169.

[8] LEI M, LI Q, GE X H, et al. Partially collaborative edge caching based on federated deep reinforcement learning[J]. IEEE Transactions on Vehicular Technology, 2023, 72(1): 1389-1394.

[9] LI Y. Federated deep reinforcement learning-based caching and bitrate adaptation for VR panoramic video in clustered MEC networks[J]. Electronics, 2022, 11(23): 3968.

[10] WEI Y K, ZHOU S P, LENG S P, et al. Federated learning empowered end-edge-cloud cooperation for 5G HetNet security[J]. IEEE Network, 2021, 35(2): 88-94.

[11] QUINTANA-RAMIREZ I, SEQUEIRA L, RUIZ-MAS J. An edge-cloud approach for video surveillance in public transport vehicles[J]. IEEE Latin America Transactions, 2021, 19(10): 1763-1771.

[12] YAN K P, SHAN H G, SUN T X, et al. Reinforcement learning-based mobile edge computing and transmission scheduling for video surveillance[J]. IEEE Transactions on Emerging Topics in Computing, 2022, 10(2): 1142-1156.

[13] KE R M, ZHUANG Y F, PU Z Y, et al. A smart, efficient, and reliable parking surveillance system with edge artificial intelligence on IoT devices[J]. IEEE Transactions on Intelligent Transportation Systems, 2021, 22(8): 4962-4974.

[14] MING Z X, CHEN J S, CUI L Z, et al. Edge-based video surveillance with graph-assisted reinforcement learning in smart construction[J]. IEEE Internet of Things Journal, 2022, 9(12): 9249-9265.

[15] HOU B, ZHANG J X. Real-time surveillance video salient object detection using collaborative cloud-edge deep reinforcement learning[C]//Proceedings of the International Joint Conference on Neural Networks. Piscataway: IEEE Press, 2021: 1-8.

[16] KIM D, KIM K, PARK S. Automatic PTZ camera control based on deep-Q network in video surveillance system[C]//Proceedings of the International Conference on Electronics, Information, and Communication. Piscataway: IEEE Press, 2019: 1-3.

[17] 施巍松, 张星洲, 王一帆, 等. 边缘计算: 现状与展望[J]. 计算机研究与发展, 2019, 56(1): 69-89.
SHI W S, ZHANG X Z, WANG Y F, et al. Edge computing: state-of-the-art and future directions[J]. Journal of Computer Research and Development, 2019, 56(1): 69-89(in Chinese).

[18] 赵梓铭, 刘芳, 蔡志平, 等. 边缘计算: 平台、应用与挑战[J]. 计算机研究与发展, 2018, 55(2): 327-337.
ZHAO Z M, LIU F, CAI Z P, et al. Edge computing: platforms, applications and challenges[J]. Journal of Computer Research and Development, 2018, 55(2): 327-337(in Chinese).

[19] 施巍松, 孙辉, 曹杰, 等. 边缘计算: 万物互联时代新型计算模型[J]. 计算机研究与发展, 2017, 54(5): 907-924.
SHI W S, SUN H, CAO J, et al. Edge computing: an emerging computing model for the Internet of everything era[J]. Journal of Computer Research and Development, 2017, 54(5): 907-924(in Chinese).

[20] HU Y C, PATEL M, SABELLA D, et al. Mobile edge computing-a key technology towards 5G[J]. European Telecommunications Standards Institute, 2015: 1-16.

[21] TALEB T, SAMDANIS K, MADA B, et al. On multi-access edge computing: a survey of the emerging 5G network edge cloud architecture and orchestration[J]. IEEE Communications Surveys & Tutorials, 2017, 19(3): 1657-1681.

[22] 吴冀衍, 程渤, 南国顺, 等. 面向异构无线网移动视频传输的联合信源信道编码方式[J]. 计算机学报, 2015, 38(2): 439-454.
WU J Y, CHENG B, NAN G S, et al. A novel joint source-channel coding scheme for mobile video delivery in heterogeneous wireless networks[J]. Chinese Journal of Computers, 2015, 38(2): 439-454 (in Chinese).

[23] SAHA N, ZANGOOEI M, GOLKARIFARD M, et al. Deep reinforcement learning approaches to network slice scaling and placement: a survey[J]. IEEE Communications Magazine, 2023, 61(2): 82-87.

[24] CHIEN W C, WENG H Y, LAI C F. Q-learning based collaborative cache allocation in mobile edge computing[J]. Future Generation Computer Systems, 2020, 102: 603-610.

[25] WANG S, ZHANG X, ZHANG Y, et al. A survey on mobile edge networks: convergence of computing, caching and communications [J]. IEEE Access, 2017, 5: 6757-6779.

[26] 李佳, 谢人超, 贾庆民, 等. 面向视频流的 MEC 缓存转码联合优化研究[J]. 电信科学, 2018, 34(8): 76-86.
LI J, XIE R C, JIA Q M, et al. A survey on joint optimization of MEC caching and transcoding for video streaming[J]. Telecommunications Science, 2018, 34(8): 76-86(in Chinese).

[27] MANIOTIS P, THOMOS N. Viewport-aware deep reinforcement learning approach for 360° video caching[J]. IEEE Transactions on Multimedia, 2021, 24: 386-399.

[28] ZHONG C, GURSOY M C, VELIPASALAR S. Deep reinforcement learning-based edge caching in wireless networks[J]. IEEE Transactions on Cognitive Communications and Networking, 2020, 6(1): 48-61.

[29] PANG Z Y, SUN L F, HUANG T C, et al. Towards QoS-aware cloud live transcoding: a deep reinforcement learning approach[C]//Proceedings of the IEEE International Conference on Multimedia and Expo. Piscataway: IEEE Press, 2019: 670-675.

[30] WAN Z, LI Y. Deep reinforcement learning-based collaborative video caching and transcoding in clustered and intelligent edge B5G networks[J]. Wireless Communications and Mobile Computing, 2020, 2020(1): 6684293.

[31] AN B, SUN S, WANG R D. Deep reinforcement learning for quantitative trading: challenges and opportunities[J]. IEEE Intelligent Systems, 2022, 37(2): 23-26.

[32] HE X M, WANG K, HUANG H W, et al. Green resource allocation based on deep reinforcement learning in content-centric IoT[J]. IEEE Transactions on Emerging Topics in Computing, 2020, 8(3): 781-796.

[33] MENG F, CHEN P, WU L N, et al. Power allocation in multi-user cellular networks: deep reinforcement learning approaches[J]. IEEE Transactions on Wireless Communications, 2020, 19(10): 6255-6267.

[34] PARK S, KIM J, KWON D, et al. Joint offloading and streaming in mobile edges: a deep reinforcement learning approach[C]//Proceedings of the IEEE VTS Asia Pacific Wireless Communications Symposium. Piscataway: IEEE Press, 2019: 1-4.

[35] ZHANG H, WU W J, WANG C Y, et al. Deep reinforcement learning-based offloading decision optimization in mobile edge computing[C]//Proceedings of the IEEE Wireless Communications and Networking Conference. Piscataway: IEEE Press, 2019: 1-7.

[36] HE Y, ZHAO N, YIN H X. Integrated networking, caching, and computing for connected vehicles: a deep reinforcement learning approach[J]. IEEE Transactions on Vehicular Technology, 2018, 67(1): 44-55.

[37] WANG F X, ZHANG C, WANG F, et al. Intelligent edge-assisted crowdcast with deep reinforcement learning for personalized QoE[C]//Proceedings of the IEEE Conference on Computer Com-

munications. Piscataway: IEEE Press, 2019: 910-918.

[38] CHEN M H, PONEC M, SENGUPTA S, et al. Utility maximization in peer-to-peer systems with applications to video conferencing[J]. IEEE/ACM Transactions on Networking, 2012, 20(6): 1681-1694.

[39] SUTTON R S, BARTO A G. Reinforcement learning: an introduction[M]. 2nd ed. Cambridge: MIT Press, 2018.

[40] MNIH V, KAVUKCUOGLU K, SILVER D, et al. Human-level control through deep reinforcement learning[J]. Nature, 2015, 518(7540): 529-533.

Edge-intelligent transmission optimization of emergency surveillance video based on IcD-FDRL

LI Yan[1, 2], WAN Zheng[1, *], DENG Chengzhi[2], WANG Shengqian[2]

(1. School of Information Management and Mathematics, Jiangxi University of Finance and Economics, Nanchang 330032, China;
2. School of Information Engineering, Jiangxi University of Water Resources and Electric Power, Nanchang 330099, China)

Abstract: Emergency surveillance video transmission is a key technical means to improve emergency handling capability under circumstances such as emergency monitoring, public security incident handling, and post-disaster reconstruction. It has gradually become a key focus of research and development in the construction of the national smart emergency system. With the continuous development of 5G technology and decision-making artificial intelligence technology in recent years, an edge-intelligent transmission architecture for emergency surveillance video was established, aimed at public safety and emergency rescue monitoring in local areas. This model seeks to achieve adaptive and high-quality transmission of emergency surveillance video. Furthermore, the importance measurement method of emergency surveillance video was designed, and an intra-clustered dynamic federated deep reinforcement learning algorithm was proposed. The proposed optimization method based on intra-clustered dynamic federated deeps reinforcement learning (IcD-FDRL) enhances the edge-intelligent transmission of emergency surveillance video, breaks monitoring data silos, improves algorithm learning efficiency, and realizes low-delay, low-cost, high-quality, and priority transmission of important emergency surveillance video. Finally, a simulation experiment was performed and its results were compared, verifying the effectiveness of the proposed model and algorithms.

Keywords: emergency surveillance video; edge cluster; dynamic federated deep reinforcement learning; edge intelligence; wireless video transmission; mobile edge computing

Received: 2023-06-16; **Accepted**: 2024-03-29; **Published Online**: 2024-04-20 16:18

URL: link.cnki.net/urlid/11.2625.V.20240419.0927.001

Foundation items: National Natural Science Foundation of China (61961021); Key Project of Science and Technology Research of Jiangxi Provincial Education Department (GJJ180251); Start-up Project of Doctoral Research in Jiangxi University of Water Resources and Electric Power (2024kyqd062)

* **Corresponding author**. E-mail: wanzheng97@163.com

http://bhxb.buaa.edu.cn jbuaa@buaa.edu.cn
DOI: 10.13700/j.bh.1001-5965.2023.0432

基于自适应多态蚁群优化的智能体路径规划

邢娜，邸昊天，尹文杰*，韩亚君，周洋
(北方工业大学 信息学院，北京 100144)

摘　　要：在智能体路径规划中，蚁群算法是较为流行的路径求解策略，且得到了广泛的应用。然而，传统蚁群算法存在局部最优和多余拐点问题。基于此，提出自适应多态蚁群优化算法，通过多群体划分和协作机制，极大的提高了搜索和收敛速度，有助于增强全局搜索能力，避免陷入局部最优解。改进的信息素更新策略和路径选择记录表构造进一步提高路径规划的准确性。通过3次B样条平滑曲线对路径进行处理，有效减少拐点，实现路径的平滑化。经过MATLAB和机器人操作系统(ROS)-Gazebo仿真验证，结果表明：所提算法在复杂环境下具有良好的可行性。综上所述，所提算法为智能体全局搜索带来了显著的优化和改进。

关 键 词：路径规划；自适应多态蚁群算法；B样条；机器人操作系统；Gazebo平台
中图分类号：TP301.6; TP391.9
文献标志码：A　　**文章编号**：1001-5965(2025)07-2330-08

近年来，智能体相关技术得到了迅猛发展，智能体在生产制造、医疗卫生、教育娱乐等领域的应用越来越广泛。这些应用场景要求智能体能够避免与障碍物碰撞，快速到达目标点[1-2]。因此，路径规划成为智能体领域中重要研究方向之一[3]。

目前，已有大量算法应用于智能体路径规划，如A*算法[4]、蚁群算法[5-7]、人工势场法[8]、Dijkstra与蚁群的结合算法[9]等。其中，蚁群算法是一种经典的路径规划方法。Liu等[10]采用蚁群算法改善路径规划的智能优化机制。Zhang等[11]采用回退策略解决了刮擦障碍和死锁问题。Yuan等[12]通过增强蚁群算法的启发式功能、信息素分配机制和路径选择策略，提高路径规划效率。Brand等[13]研究了蚁群算法在动态环境下智能体路径规划问题。然而，传统蚁群算法易产生局部最优解[14]。为此，文献[15-17]引入自适应多态蚁群算法，提出一种基于改进蚁群优化的智能体路径规划算法，但生成路径仍存在多余拐点。

本文为优化局部最优解及解决多余拐点的问题，提出自适应多态蚁群优化算法，将伪随机规则进行状态转移，并与自适应并行策略相结合，避免在复杂环境中算法陷入局部最优解。同时构造路径选择记录表，改进信息素更新策略以确定下一个路径点方向。从而提高全局搜索能力和智能体路径规划的准确性。然后，进一步采用3次B-Spline平滑曲线对所得路径进行平滑，以减少轨迹拐点，获得更平滑路径。最后，为验证本文算法的可行性，本文运用MATLAB和机器人操作系统(robot operating system, ROS)-Gazebo分别对本文算法进行仿真验证。

1 问题描述

考虑智能体在含有障碍物的环境中移动，需要找到从起点到目标点的最佳路径。其中，约束条件分为硬约束和软约束2类：硬约束是必须满足的，即智能体需要避开障碍物；软约束是尽量缩短路径

收稿日期：2023-07-03；**录用日期**：2023-08-15；**网络出版时间**：2023-09-05 10:03
网络出版地址：link.cnki.net/urlid/11.2625.V.20230904.2053.002
基金项目：国家重点研发计划(2020YFC0811004)；北京市教育委员会科研计划项目(KM202310009001)；北方工业大学科研启动基金(110051360023XN224-9)
* **通信作者**. E-mail：yin__wenjie@163.com

引用格式：*邢娜，邸昊天，尹文杰，等. 基于自适应多态蚁群优化的智能体路径规划[J]. 北京航空航天大学学报，2025，51(7)：2330-2337.*
XING N，DI H T，YIN W J，et al. Path planning for agents based on adaptive polymorphic ant colony optimization[J]. Journal of Beijing University of Aeronautics and Astronautics，2025，51(7)：2330-2337 (in Chinese).

长度, 降低智能体能耗。

令 x_i 表示路径上的第 i 个点, 考虑智能体的起点和终点约束:

$$\begin{cases} x_0 = S \\ x_n = G \end{cases} \tag{1}$$

式中: S 为起点; G 为终点; x_0 为路径上的第 1 个点; x_n 为路径上的最后一个点。如果路径规划生成的路径不在可行空间内, 那么智能体就无法沿着这条路径移动。因此:

$$x_i \in X \qquad i = 0,1,\cdots,n \tag{2}$$

其中: X 为智能体的可行空间。

智能体在移动过程中需要避开障碍物, 假设环境中存在 M 个障碍物, 令 $O_{m'}$ 表示第 m' 个障碍物所在区域, 障碍物满足:

$$\bigcap_{m'=1}^{M} O_{m'} = \varnothing \tag{3}$$

式中: $\varnothing$ 为空集。因此, 智能体的避障约束可以表示为

$$x_i \notin O_{m'} \qquad i = 0,1,\cdots,n; m' = 1,2,\cdots,M \tag{4}$$

上述约束属于硬约束。此外, 本文考虑的软约束主要包含以下 2 种:

1) 路径长度。路径长度通常与时间或成本有关。在寻找最短路径的同时, 考虑到时间或成本的限制, 在可接受的时间或成本范围内得到最优路径。

2) 优先级。在有多个优化目标或需求的情况下, 需要权衡不同目标或需求的优先级。例如, 在寻找最短路径的同时, 需要优先避开某些区域。

假设 $L(x_i)$ 表示从起点到 x_i 的路径长度, $w(x_i)$ 表示在 x_i 处的优先级权重, 本文考虑如下多目标优化: $\min f_1(x_i), \min f_2(x_i), \cdots, \min f_l(x_i)$,

$f_l(x_i)$ 为关于目标或需求的函数:

$$\begin{cases} f_1(x_i) = L(x_i) \\ f_2(x_i) = w(x_i) \end{cases} \tag{5}$$

本文旨在通过解决上述多目标优化问题, 得到一个合理的、在不同目标或需求之间做出权衡的路径规划方案。

2　基于改进蚁群算法的路径规划

2.1　路径生成

为防止陷入局部最优, 本文引入自适应多态并行规则和路径选择记录表[18], 对多态蚁群算法进行改进以生成智能体最佳移动路径。设置侦查蚂蚁和搜索蚂蚁执行以下任务。

侦查蚂蚁。将数量为 K 的侦察蚂蚁分别放置在智能体 n 个路径节点上。每个侦察蚂蚁的目标是调查智能体的另一个路径节点(数量为 $n-1$)。它们以智能体的当前路径节点 i 为中心, 将新节点 j 的调查结果与已有信息相结合, 构成调查元素 s_{ij}, 并将其标记在路径节点 i 到路径节点 j 的路径上, 当路径节点 j 在路径节点 i 的范围内时 $s_{ij}(i,j = 1,2,\cdots,n-1, i \neq j)$ 表达式如式(6)所示, 其余情况为 0:

$$s_{ij} = d_{ij}^{\min} / d_{ij} \tag{6}$$

式中: d_{ij} 为路径 (i,j) 之间的距离; $d_{ij}^{\min}$ 为以智能体路径节点 i 为中心, 到其他 $n-1$ 个路径节点的最小距离。

根据此结果, 设置初始时刻每条路径上的信息素浓度:

$$\tau_{ij}(0) = \begin{cases} H s_{ij} & s_{ij} \neq 0 \\ H d_{ij}^{\min} / d_{ij}^{\max} & \text{其他} \end{cases} \tag{7}$$

式中: $d_{ij}^{\max}$ 为以路径节点 i 为中心到其他 $n-1$ 个路径节点的最大距离; H 为常数。

搜索蚂蚁。搜索蚂蚁 $k(k = 1,2,\cdots,K)$ 在搜索阶段选择要移动到哪个路径节点时, 根据自适应多态并行规则来确定转移方向, 当 $j \notin T_k$ 且 $s_{ij} \neq 0$ 时, 自适应转移概率 $p_{ij}^k(t)$ 的计算式如式(8)所示, 其余情况均为 0:

$$P_{ij}^k(t) = \frac{\tau_{ij}(t)^{\frac{4\alpha t}{N}+1} \eta_{ij}(t)^{\frac{2\beta t}{N}+1}}{\sum_{h \notin T_k} \tau_{ih}(t)^{\frac{4\alpha t}{N}+1} \eta_{ih}(t)^{\frac{2\beta t}{N}+1}} \tag{8}$$

式中: $\tau_{ij}(t)$ 为第 t 次迭代路径 (i,j) 上的信息素浓度; α 为信息素因子, 用于调整路径的相对重要性; β 为启发函数因子, 用于调整可见性的相对重要性; T_k 为非可达集合; N 为最大迭代次数; $\eta_{ij}(t)$ 为启发式函数, 其计算式为 $\eta_{ij}(t) = 1/d_{ij}$。

当所有蚂蚁完成一次迭代后, 蚂蚁 k 经过 (i,j) 且 $s_{ij} \neq 0$ 时每条路径上的信息素更新为如式(9)所示, 其余情况均为 0:

$$\Delta\tau_{ij}^k(t) = Q(d_{ij}^{\min} / d_{ij}) / L_k \tag{9}$$

式中: Q 为信息素增加强度系数; L_k 为蚂蚁 k 走过的路程长度。

然后, 根据伪随机规则进行状态转移, 采用自适应并行策略确定状态转移函数中的最佳组合参数。状态转移规则表示为

$$S = \begin{cases} \underset{j \in A_k}{\operatorname{argmax}} \left\{ [\tau_{ij}(t)]^{\alpha} [\eta_{ij}(t)]^{\beta} \right\} & q < q_0 \\ p_{ij}^k(t) & \text{其他} \end{cases} \tag{10}$$

式中: A_k 为蚂蚁 k 可选路径集; q 为一个均匀分布在 [0,1] 区间的随机数; q_0 为随着迭代次数增加逐渐趋

近于1，本文$q_0=1-\mathrm{e}^{-1/\bar{K}}$，其中，$\bar{K}$表示迭代次数，范围为1～$N$。

在多态蚁群算法中，搜索蚂蚁会在可能的最优解路径上留下信息素，而忽略其他路径，这可能导致最终的解决方案不是最佳的。为避免上述情况，本文引入自适应信息更新策略，其中，自适应信息素浓度表示为

$$\Delta\tau_{ij}^{k}(t)=\begin{cases}\left(\dfrac{Q\left(d_{ij}^{\min}/d_{ij}\right)}{L_{\text{best}}\displaystyle\sum_{k=1}^{l}d_k^L}\right)^{\frac{2t}{N}+1} & k\text{经过}(i,j)\text{且}s_{ij}\neq 0\\ 0 & \text{其他}\end{cases} \tag{11}$$

式中：L_{best}为当前最优路径长度；L为前一次迭代中的最优路径；d_k^L为未排序的路径长度。

在搜索过程中，为让智能体在移动时尽可能避开障碍物，通过自适应地更新信息素浓度，根据改进信息素更新策略以确定下一个路径点方向。可以避免对安全路径的忽视，即使较少的信息素被留在路径上，也能有效地维持搜索，并防止信息素随着迭代次数的增加而减少。当蚂蚁产生多拐点时，本文算法会更新路径选择记录表，设置奖惩函数，对处于多余拐点路径上的信息素进行惩罚。因此，蚂蚁能够调整其在原位置的前进方向，快速找到没有障碍物的方向，并沿该方向继续前进。信息素折扣函数设置为

$$\tau_{\text{rs}}=(1-\lambda)\tau_{\text{rs}} \tag{12}$$

式中：$(1-\lambda)$为折扣系数，为防止信息素无限积累，需将其取值范围设定为[0,1]。

为验证本文算法的有效性，如图1所示，智能体从(3.5,17.5)移动到(15.5,7.5)，红色线为传统蚁群算法生成的路径，轨迹“*”为基于改进的自适应多态蚁群算法生成的路径。由图1可知，本文算法能够有效处理传统蚁群算法局部最优问题，使得搜索能够继续进行，并找到可行的路径解决方案。本文算法流程如下。

首先，对环境进行网格建模。

初始化：参数Q、信息素浓度H、最大迭代次数N及智能体的起始位置；

开始循环：

1. 将侦察蚂蚁放置在路径节点上，对其他$n-1$个路径进行侦察，并存储调查结果；

2. 计算初始时刻每条路径上的信息素，设置迭代次数初始值为0；

3. 随机选择每个搜索蚂蚁的初始位置，并将其放置在路径选择记录表中；

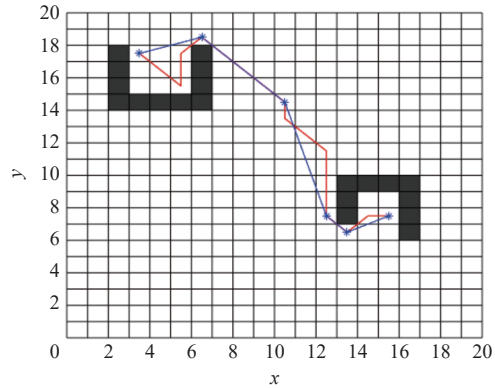


图1 生成路径对比

Fig. 1 Comparison of path generation

4. 根据式(11)计算搜索蚂蚁k移动到的位置j，及上一位置i，并将j放入搜索蚂蚁k的路径选择记录表，直至每只搜索蚂蚁完成迭代求解，并处理可能生成局部最优的情况；

5. 计算每个搜索蚂蚁的目标函数值，记录当前最优解；

6. 根据预设条件及式(13)判断是否继续迭代，如果满足条件则执行第8步，否则根据式(12)更新每条路径上的信息素浓度；

7. 清空路径选择记录表和信息素变化量，更新迭代次数，返回第3步，直到到达目标点；

8. 输出最优解。

结束循环。

综上所述，本文算法通过自适应性地调整信息素显著提高了路径搜索能力和解的多样性。

2.2 轨迹平滑处理

生成路径后，还需要对智能体移动轨迹进行平滑处理。本文使用3次B-Spline来实现轨迹平滑，3次B-Spline由一系列控制点和节点构成，具有高阶连续性和平滑性，因此，能够更好地适应复杂的轨迹形状，同时具有较好的计算效率。

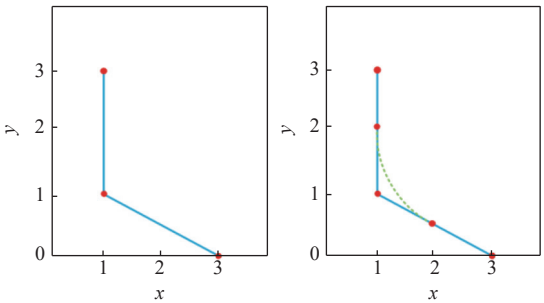
为生成平滑的3次B-Spline曲线，需要在控制点之间插值，同时保持曲线的平滑性和形状。通常将轨迹平滑问题转化为寻找最优控制点的问题，通过不断更新信息素浓度和搜索空间，找到最优的控制点组合，从而生成更加平滑和合理的轨迹。3次B-Spline曲线计算式为

$$B_i^0(t')=\begin{cases}1 & t_i'\leqslant t'<t_{i+1}'\\ 0 & \text{其他}\end{cases} \tag{13}$$

$$B_i^u(t')=\frac{t'-t_i'}{t_{i+u}'-t_i'}B_i^{u-1}(t')+\frac{t_{i+u+1}'-t'}{t_{i+u+1}'-t_{i+1}'}B_{i+1}^{u-1}(t') \tag{14}$$

式中：u为B-Spline的阶数；b为控制点的数量。对于3次B-Spline，$u=3$。如果路径由3个路径点组成，拐点为路径分割点，那么3个点可分割成2段

路径, 如图 2 所示, 3 次 B-Spline 方法在每段路径起点和终止点路径中插入一个中点, 以有效平滑路径。



(a) B-Spline 平滑前　　(b) B-Spline 平滑后

图 2　轨迹拐点与 B-Spline 平滑对比

Fig. 2　Comparison of trajectory inflection points and B-Spline smoothing

给定b个控制点的 B-Spline 曲线表示为

$$S(x)=\sum_{i=0}^{b}\left(P_i\cdot B_{i,\mathrm{b}}(z)\right) \tag{15}$$

式中: P_i为第i个控制点; $B_{i,n}(z)$为第c次曲线的 B-Spline 基函数; z为归一化曲线长度参数。

令a_i表示由m个非递减实数组成的节点, 其中, $m=b+c+1$。如果路径由 3 个路径点组成, 即$b=3$, $c=5$, 则$m=9$。$B_{i,0}$表示 B-Spline 一阶基函数, 其定义为

$$B_{i,0}=\begin{cases}1 & z\in[z_i,z_{i+1}]\\ 0 & \text{其他}\end{cases} \tag{16}$$

式中: z_i、z_{i+1}为各自的节点。高阶基函数(从 2～c)定义为

$$B_{i,\mathrm{b}}(z)=\frac{z-z_i}{z_{i+1}-z_i}B_{i,\mathrm{b}-1}(z)+\frac{z_{i+k}-z}{z_{i+k}-z_{i+1}}B_{i+1,u-1}(z) \tag{17}$$

智能体路径规划完成后, 将会生成一条连续平滑的 B-spline 曲线。

3　基于 MATLAB 的仿真实验

3.1　环境建模

为验证本文算法的性能指标, 本实验在 MATLAB 2018b 上, 通过仿真实验验证本文算法的收敛速度和优化能力。实验参数如表 1 所示。

构造栅格地图, 将整体环境划分为合适的网格单元格, 然后, 设置障碍物和可移动区域。通过遍历环境中的每个位置, 可以确定各位置是否为障碍物或可移动区域, 以“1”表示障碍物网格, “0”表示可移动区域网格。由此获得具有标记的网格。

图 3 为一个20×20的网格地图示例, 黑色网格表示障碍物, 白色网格表示智能体可移动空间。

3.2　基于 MATLAB 的仿真结果分析

传统蚁群算法在搜索过程中会产生朝向障碍的多余轨迹, 如图 4 所示, 进行多次传统蚁群算法的实验。但是引入自适应多态蚁群算法和路径选择记录表, 通过改进信息素更新策略可以提高规划方向的准确性, 有效提升算法的全局路径搜索能

表 1　实验参数设置表

Table 1　Experimental Parameter Setting Table

参数	数值
起点	381
终点	18
蚂蚁个数m	50
最大迭代次数N	100
信息素因子α	1
启发函数因子β	5
信息素挥发因子ρ	0.6
信息素增加强度系数Q	1
折扣系数$(1-\lambda)$	0.7

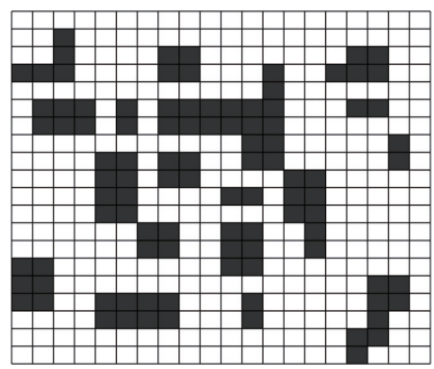
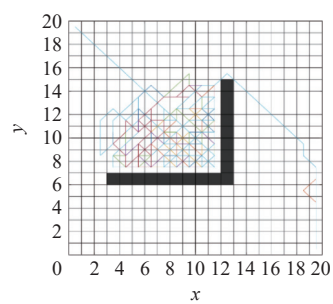
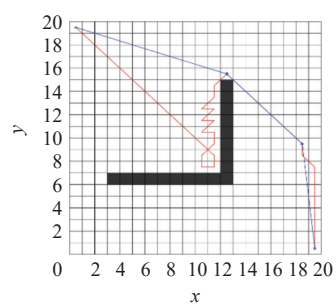


图 3　栅格地图下环境建模

Fig. 3　Environment modeling under grid map



(a) 传统蚁群算法



(b) 本文算法

图 4　传统蚁群算法蚂蚁路线与改进算法对比

Fig. 4　Comparison between traditional ant colony algorithm ant route and improved algorithm

力，避免了传统蚁群算法的盲目搜索。由此可见，本文算法最终会得到一条更优路径，如图 4 轨迹“*”所示，其中，红色轨迹是传统蚁群的实验结果。

当场景中存在很多障碍物时，仿真实验表明，蚁群算法仍会产生多余轨迹，而引入自适应多态蚁群算法和路径选择记录表后的轨迹（路径“*”）可以绕过障碍，有效缩短了路径。通过采用 3 次 B 样条曲线对轨迹进行平滑，使拐点更少，如图 5 所示。

为进一步证明本文算法的有效性。将本文算法与基于改进 A*的蚁群算法[19]进行对比。图 6 和图 7 分别为基于改进 A*的蚁群算法和本文算法生成的路径，地图中左上角为智能体起点，右下角为智能体终点。图 8 为基于改进 A*的蚁群算法和本文算法最优路径的路径长度与迭代次数的对比结果。由此可见，与基于改进 A*的蚁群算法相比，本文算法可以生成更短的路径长度，收敛速度更快，平滑度更好。

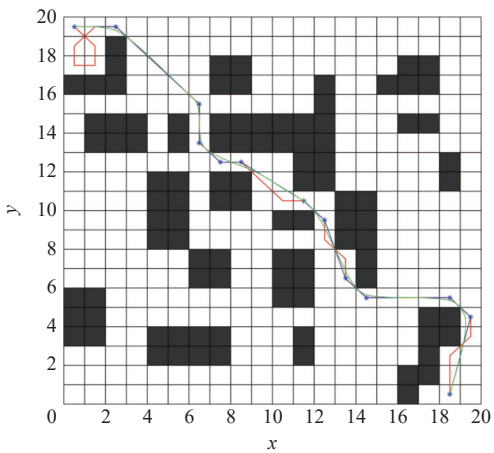


图 5 复杂特殊场景下本文算法轨迹

Fig. 5 Proposed algorithm trajectories in complex special scenarios

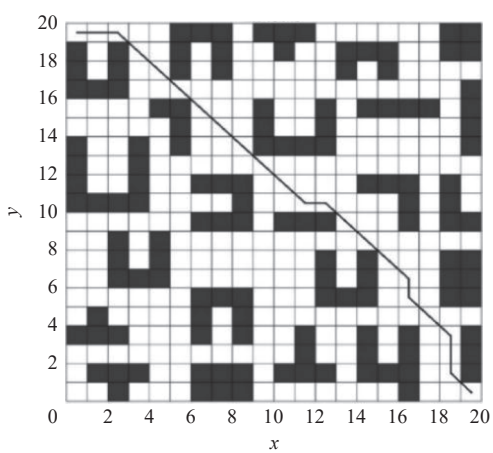


图 6 基于改进 A*的蚁群算法生成的轨迹

Fig. 6 Trajectory generated by an ant colony algorithm based on improved A*

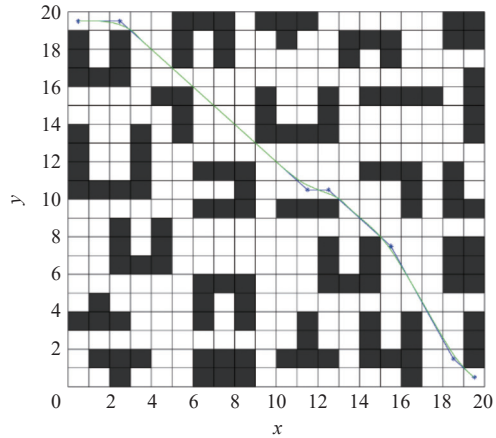


图 7 本文算法生成的轨迹

Fig. 7 Trajectory generated by the proposed algorithm

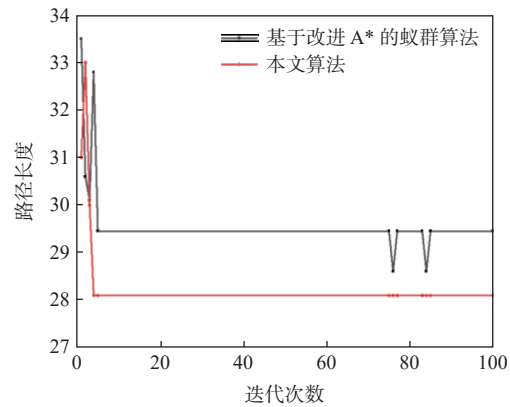


图 8 路径长度与迭代次数对比

Fig. 8 Path length versus number of iterations

4 基于 ROS-Gazebo 的仿真实验

与 MATLAB 仿真相比，ROS-Gazebo 仿真更贴近真实物理情景，仿真的实时性和真实性更强，同时能够克服 MATLAB 仿真曲线表示缺乏直观性的缺点，实现仿真结果的三维立体视景演示。为进一步验证本文算法的有效性，在 ROS-Gazebo 仿真环境下进行实验，从而实现完整的“建模-仿真-可视化”的仿真流程，基本框架如图 9 示。

4.1 算法初始化

为在 ROS 中验证本文算法，对信息素和蚂蚁位置进行初始化。

1）初始化信息素。首先，从 ROS 的参数服务器中获取地图尺寸和信息素初始值；然后，使用 ROS 中的 Occupancy Grid 消息类型创建一个信息素地图，将地图尺寸和信息素初始值设置为消息的属性；最后，初始化所有像素点的信息素值，并将信息素地图发布到 ROS 中的话题中。

2）初始化蚂蚁位置。在 ROS 中，可以通过随机数或启发式算法初始化蚂蚁的位置，如将所有蚂

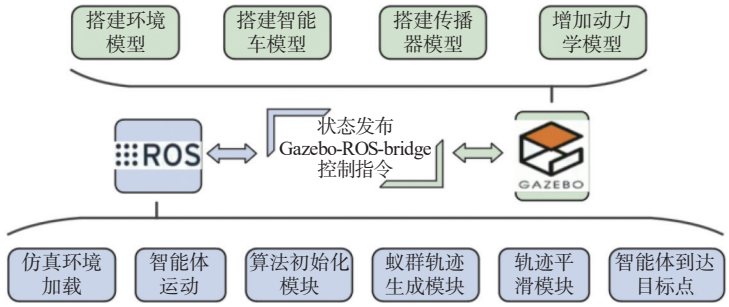


图 9　ROS-Gazebo 仿真平台架构

Fig. 9　ROS-Gazebo simulation platform architecture

蚁放置在地图的中心或将其随机分布在地图上。首先, 本文从 ROS 的参数服务器中获取蚂蚁数量和地图尺寸; 然后, 使用 Ant 类创建一个蚂蚁, 并将蚂蚁的位置随机分布在地图上; 最后, 将所有蚂蚁添加到一个列表中。其整体流程如图 10 所示。



图 10　蚂蚁位置初始化流程

Fig. 10　Ant position initialization flow

4.2　路径生成模块设计

完成初始化后, 需要设计路径生成模块。

1) 更新信息素。使用 Navigation Stack 中 costmap_2d 包的 Costmap2D 类来表示地图, 并在其上设置信息素含量。根据当前位置到下一步位置的信息素含量和启发值计算出每个可选方向的期望值, 然后, 根据期望值大小选择下一步方向。

2) 路径规划器设计。利用导航包的路径规划器实现路径规划功能, 定义路径生成函数, 在每次蚂蚁到达终点时调用。整体路径生成过程如图 11 所示。

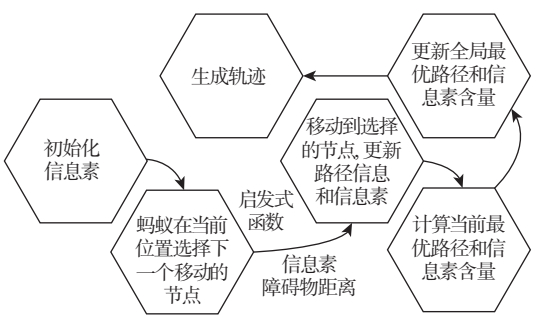


图 11　路径生成流程

Fig. 11　Path generation flow

路径规划器使用本文算法来搜索最优路径, 并根据信息素含量和起点、终点及障碍物的位置来生成路径。

3) 路径平滑器设计。使用 3 次 B-Spline 曲线对智能体路径进行平滑。本文用 Spline 库计算 B-Spline 曲线, 流程如图 12 所示。

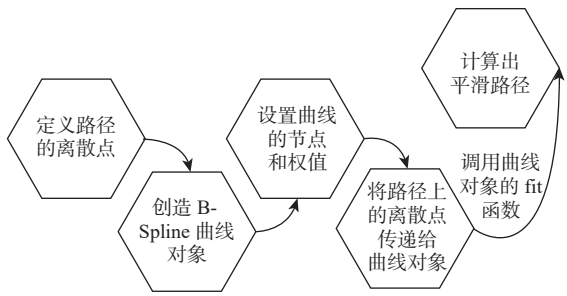


图 12　路径平滑流程

Fig. 12　Path smoothing flow

4.3　基于 ROS-Gazebo 的仿真结果分析

给定智能体的目标点, 采用本文算法计算最优路径。图 13～图 15 分别为智能体初始位置、智能体避障过程、智能体目标点位置。由图可知, 本文算法能够有效对智能体进行路径规划。

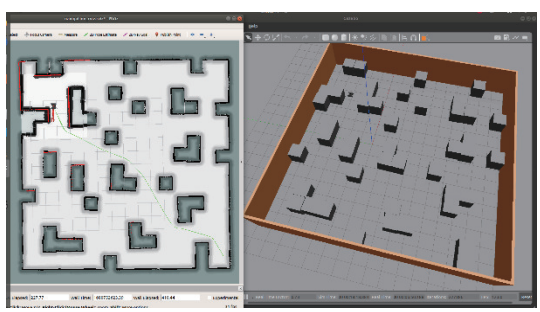


图 13　智能体初始位置

Fig. 13　Initial position of the agent

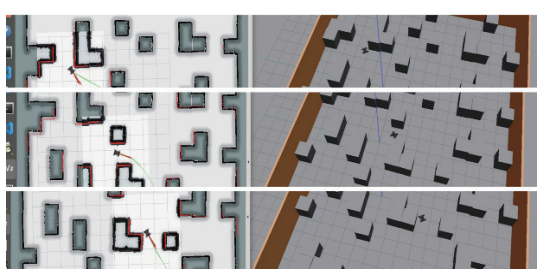


图 14　智能体避障过程

Fig. 14　Obstacle avoidance process of the agent

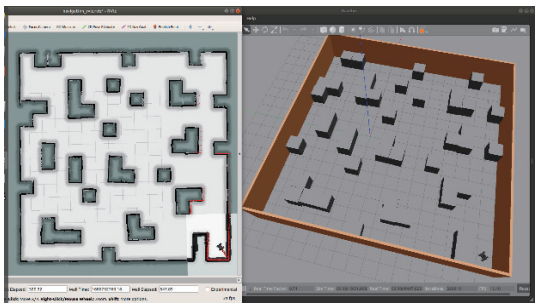


图15 智能体目标点位置

Fig. 15 Position of the target point of the agent

5 结 论

1)为解决传统蚁群算法易陷入局部最优,本文提出一种自适应多态蚁群优化路径规划算法,并采用3次B-spline对得到的路径进行平滑处理。

2)为验证本文算法的有效性,本文基于MATLAB和ROS-Gazebo平台进行了仿真实验,仿真结果表明本文算法能够为智能体自主规划出一条避障路径。

为使本文算法适用更多场景,在未来的研究中,将考虑引入一些更加实用的智能体路径规划的模型。

参考文献(References)

[1] 王梓强, 胡晓光, 李晓筱, 等. 移动机器人全局路径规划算法综述[J]. 计算机科学, 2021, 48(10): 19-29.
WANG Z Q, HU X G, LI X X, et al. Overview of global path planning algorithms for mobile robots[J]. Computer Science, 2021, 48(10): 19-29(in Chinese).

[2] ZHANG L, ZHANG Y J, LI Y F. Mobile robot path planning based on improved localized particle swarm optimization[J]. IEEE Sensors Journal, 2021, 21(5): 6962-6972.

[3] 霍凤财, 迟金, 黄梓健, 等. 移动机器人路径规划算法综述[J]. 吉林大学学报 (信息科学版), 2018, 36(6): 639-647.
HUO F C, CHI J, HUANG Z J, et al. Review of path planning for mobile robots[J]. Journal of Jilin University (Information Science Edition), 2018, 36(6): 639-647(in Chinese).

[4] SANG H Q, YOU Y S, SUN X J, et al. The hybrid path planning algorithm based on improved A* and artificial potential field for unmanned surface vehicle formations[J]. Ocean Engineering, 2021, 223: 108709.

[5] ZANG X N, JIANG L, DING B, et al. A hybrid ant colony system algorithm for solving the ring star problem[J]. Applied Intelligence, 2021, 51(6): 3789-3800.

[6] YI G H, FENG Z L, MEI T C, et al. Multi-AGVs path planning based on improved ant colony algorithm[J]. The Journal of Supercomputing, 2019, 75(9): 5898-5913.

[7] 张松灿, 普杰信, 司彦娜, 等. 蚁群算法在移动机器人路径规划中的应用综述[J]. 计算机工程与应用, 2020, 56(8): 10-19.
ZHANG S C, PU J X, SI Y N, et al. Survey on application of ant colony algorithm in path planning of mobile robot[J]. Computer Engineering and Applications, 2020, 56(8): 10-19(in Chinese).

[8] SONG J, HAO C, SU J C. Path planning for unmanned surface vehicle based on predictive artificial potential field[J]. International Journal of Advanced Robotic Systems, 2020, 17(2): 172988142091846.

[9] ZHOU Y L, HUANG N N. Airport AGV path optimization model based on ant colony algorithm to optimize Dijkstra algorithm in urban systems[J]. Sustainable Computing: Informatics and Systems, 2022, 35: 100716.

[10] LIU G Q, LI T J, PENG Y Q, et al. The ant algorithm for solving robot path planning problem[C]//Proceedings of the 3rd International Conference on Information Technology and Applications. Piscataway: IEEE Press, 2005: 25-27.

[11] ZHANG Y, CAO Y, HAN Z. Path planning of vehicle based on improved ant colony algorithm[C]//Proceedings of the International Conference on Modelling, Identification and Control. Piscataway: IEEE Press, 2012: 797-801.

[12] YUAN Z R, YU H Y, HUANG M H. Improved ant colony optimization algorithm for intelligent vehicle path planning[C]//Proceedings of the International Conference on Industrial Informatics-Computing Technology, Intelligent Technology, Industrial Information Integration. Piscataway: IEEE Press, 2017: 1-4.

[13] BRAND M, MASUDA M, WEHNER N, et al. Ant colony optimization algorithm for robot path planning[C]//Proceedings of the International Conference on Computer Design and Applications. Piscataway: IEEE Press, 2010: V3-436-V3-440.

[14] AKKA K, KHABER F. Mobile robot path planning using an improved ant colony optimization[J]. International Journal of Advanced Robotic Systems, 2018, 15(3): 1729881418774673.

[15] WU L, HUANG X D, CUI J G, et al. Modified adaptive ant colony optimization algorithm and its application for solving path planning of mobile robot[J]. Expert Systems with Applications, 2023, 215: 119410.

[16] 甄然, 张春悦, 矫阳, 等. 基于自适应多态融合蚁群算法的无人机航迹规划[J]. 河北科技大学学报, 2019, 40(6): 526-532.
ZHEN R, ZHANG C Y, JIAO Y, et al. Research on UAV route planning based on adaptive polymorphic ant colony algorithm[J]. Journal of Hebei University of Science and Technology, 2019, 40(6): 526-532(in Chinese).

[17] MIAO C W, CHEN G Z, YAN C L, et al. Path planning optimization of indoor mobile robot based on adaptive ant colony algorithm[J]. Computers & Industrial Engineering, 2021, 156: 107230.

[18] JIAO Z Q, MA K, RONG Y L, et al. A path planning method using adaptive polymorphic ant colony algorithm for smart wheelchairs[J]. Journal of Computational Science, 2018, 25: 50-57.

[19] LAN X, LV X F, LIU W, et al. Research on robot global path planning based on improved A-star ant colony algorithm[C]//Proceedings of the 5th Advanced Information Technology, Electronic and Automation Control Conference. Piscataway: IEEE Press, 2021: 613-617.

Path planning for agents based on adaptive polymorphic ant colony optimization

XING Na，DI Haotian，YIN Wenjie*，HAN Yajun，ZHOU Yang

(School of Information Science and Technology，North China University of Technology，Beijing 100144，China)

Abstract：In the realm of intelligent agent path planning, the ant colony algorithm stands as a prominent path-solving strategy that has garnered extensive adoption. However, conventional ant colony algorithms exhibit issues pertaining to local optima and excessive inflection points. The adaptive polymorphic ant colony optimization algorithm, through the mechanism of multi-colony partitioning and collaboration, remarkably enhances the search and convergence velocities, thereby bolstering global search capabilities and circumventing entrapment within local optima. To significantly increase the accuracy of path planning, this research proposes an enhanced pheromone updating technique and path selection record table building. Lastly, the adoption of trice B-Spline smoothing curves facilitates an effective reduction in inflection points, achieving path smoothness. Substantiated by MATLAB and robot operating system (ROS)-Gazebo simulations, the results demonstrate the algorithm's sound feasibility in complex environments. In conclusion, the global search of the intelligent agent is greatly optimized and enhanced by the suggested enhanced adaptive polymorphic ant colony method.

Keywords：path planning；adaptive polymorphic ant colony algorithm；B-Spline；robot operating system；Gazebo

Received：2023-07-03；**Accepted**：2023-08-15；**Published Online**：2023-09-05 10：03

URL：link.cnki.net/urlid/11.2625.V.20230904.2053.002

Foundation items：National Key Research and Development Program of China (2020YFC0811004); Scientific Research Project of Beijing Educational Committee (KM202310009001); North China University of Technology Research Initiation Fund (110051360023XN224-9)

* **Corresponding author**. E-mail：yin__wenjie@163.com

http://bhxb.buaa.edu.cn jbuaa@buaa.edu.cn
DOI: 10.13700/j.bh.1001-5965.2023.0491

语言引导视觉的小样本航拍图像目标检测

张智[1]，易华挥[1]，郑锦[2,*]

(1. 中国民航大学 计算机科学与技术学院，天津 300300；2. 北京航空航天大学 计算机学院，北京 100191)

摘 要：针对现有航拍图像目标检测方法在航拍数据集变化时，即拍摄视角、图像质量、照明条件、背景环境等发生大幅变化，以及目标外观变化明显、目标类别新增时，不经过对新数据集全样本训练，而采用原有数据集直接推理，检测精度大幅下降的问题，提出语言引导视觉的小样本航拍图像目标检测方法。采用词语-区域对齐分支取代传统目标检测网络中的分类分支，得到同时具有语言和视觉信息的词语-区域对齐分类分数作为预测分类结果，进而将目标检测和词语定位统一为一个任务，并利用语言引导提升视觉目标检测精度。针对输入文本语言变化引起小样本目标检测精度波动的情况，设计语言视觉偏置网络，挖掘语言特征和视觉特征的关联关系，提升语言视觉的匹配度，缩小精度波动，并进一步提升小样本目标检测精度。在UAVDT、Visdrone、AeriaDrone、VEDAI、CARPK_PUCPR数据集上的大量实验结果证明了所提方法的优越性能，在UAVDT航拍数据集上所提方法在30样本时平均精度均值（mAP）可达14.6%，相比航拍图像检测方法簇检测器（ClusDet）、密度图引导的目标检测网络（DMNet）、全局-局部自适应网络（GLSAN）和粗粒度密度图网络（CDMNet）在全样本训练的精度，分别提高了0.9%、−0.1%、−2.4%和−2.2%；在CARPK_PUCPR数据集上所提方法在30样本时mAP可达58.0%，相比通用目标检测方法全卷积单阶段目标检测器（FCOS）、自适应训练样本选择（ATSS）、广义焦点损失V2（GFLV2）和交并比感知密集目标检测器（VFNET）在全样本训练的精度，分别提高了1.0%、0.8%、0.1%和0.3%，体现了所提方法强大的小样本泛化和迁移能力。

关 键 词：航拍图像；小样本；词语-区域对齐分支；语言视觉偏置网络；目标检测
中图分类号：V279; TP391.4
文献标志码：A **文章编号**：1001-5965(2025)07-2338-11

无人机航拍图像具有视角灵活、视域广阔、场景易切换等特点，近年来受到越来越多的关注。作为无人机地面感知的基础，航拍图像目标检测被更多的研究者关注，大量的航拍数据集被提出，如VisDrone-DET2020数据集[1]、UAVDT数据集[2]、CARPK_PUCPR+数据集[3]、VEDAI数据集[4]和Aeria-Drone数据集[5]等。

上述航拍数据集具有的共同特点是存在大量小目标、目标新类多且背景复杂。研究者们针对这些特点，在现有通用目标检测器的基础上提出了一系列优秀的航拍图像目标检测方法，如航拍图像中的簇检测器(clustered detection, ClusDet[6])、密度图引导的目标检测网络(density map guided object detection networks, DMNet[7])、全局-局部自适应网络(global-local self-adaptive networks, GLSAN[8])、粗粒度密度图网络(coarse-grained density map networks, CDMNet[9])等。但这些方法都存在一个问题，即在某一个航拍数据集下进行训练、得到优秀的检测结果时，如果在另一个数据集下不经过全样本训练而直接推理，其检测结果往往出现精度下降的问题。

收稿日期：2023-07-27；**录用日期**：2023-09-22；**网络出版时间**：2023-10-30 20:23
网络出版地址：link.cnki.net/urlid/11.2625.V.20231030.1001.001
基金项目：国家自然科学基金(62372029);中央高校基本科研业务费专项资金(3122023007)
*通信作者. E-mail：JinZheng@buaa.edu.cn

引用格式：张智，易华挥，郑锦. 语言引导视觉的小样本航拍图像目标检测[J]. 北京航空航天大学学报，2025，51(7)：2338-2348.
ZHANG Z，YI H H，ZHENG J. Few-shot object detection of aerial image based on language guidance vision[J]. Journal of Beijing University of Aeronautics and Astronautics，2025，51(7)：2338-2348 (in Chinese).

无人机航拍图像视角灵活、视域广阔、场景易切换的优点反而会导致新数据集与原数据集有较大差异, 尤其当出现大量新类时, 检测精度下降尤为明显。但此时若收集新数据集并标注新数据集、重新训练模型, 带来的额外花费太大, 成本太高。

近年来, 为使图像分类、图像分割和目标检测等任务更加智能, 小样本学习得到了长足的进步[10-16]。小样本学习目的在于通过对大量公开可用数据进行预训练检测模型后, 仅对新类目标做少量实例标注后就能快速实现目标检测。小样本分类被定义为 C-way K-shot 问题, 即在训练阶段, 会从训练集中随机抽取 C 个类别, 其中, 每个类别包含 K 个样本, 共 $C\times K$ 个数据, 作为模型的输入; 再从这 C 个类别中剩余的数据中抽取一批样本作为模型的预测对象。即要求模型从 $C\times K$ 个数据中学会如何区分这 C 个类别。以有 10 个类别的 Visdrone 数据集进行 5-shot 训练推理为例, 首先, 在公开数据集 Object365 和 GoldG 对模型进行预训练, 其次, 在 Visdrone 数据集的训练集中随机筛选得到一组小样本数据对模型进行微调, 其中, 每组数据都包含 10 个类别, 且每个类别都至少包含 5 个样本, 然后, 将训练好的模型在测试集上进行预测, 得到模型的推理精度。现有的小样本目标检测方法通常是为常规视角下拍摄的各种目标而设计的, 而在无人机航拍图像中, 各类目标的拍摄视角、所处环境、目标大小比例变化更大, 小样本的特点更为突出, 这使得无人机航拍图像目标检测更加具有挑战性。为此研究者们根据航拍遥感图像的特点, 提出许多航拍遥感图像小样本目标检测方法[17-21]。但这些方法主要还是通过挖掘视觉图像目标的特征提升小样本检测的精度, 在一定程度上对推理数据集依赖程度高, 在跨数据集进行小样本目标检测时泛化性能依然较低。

针对这一特点, 一种基于文本-图像匹配对进行预训练, 通过引入新的语言模态数据来提升视觉模型迁移能力的方法被研究者们提出[22]。相较于重新构建大规模视觉数据集, 语言文本-图像对的获取难度和成本大大降低, 可获取的样本数量也是非常巨大的。在此基础上, 本文提出语言引导视觉的小样本航拍图像目标检测方法。在现有传统的目标检测网络框架下, 采用多模态学习, 引入词语-区域对齐分支替代原有的分类分支, 利用语言信息进一步提升视觉检测精度, 弥补航拍图像整体标注数据相对常规视角自然图像标注数据少的不足。与此同时, 针对输入文本语言变化引起小样本目标检测精度波动的情况, 设计语言视觉偏置网络, 对输入语言进行语言编码后的语言特征, 依据每张输入图像视觉特征进行自动学习调整, 进而为后续两者的融合提供更有效的支撑, 使得小样本目标检测精度进一步提升。本文主要创新点如下:

1) 采用词语-区域对齐分支取代传统检测网络中的分类分支, 得到同时具有语言和视觉信息的词语-区域对齐分类分数作为预测分类结果, 利用语言信息的引导提升视觉检测精度。

2) 设计语言视觉偏置网络, 挖掘语言特征和视觉特征的关联关系, 提升语言视觉匹配度, 缩小精度波动, 并进一步提升小样本目标检测精度。

3) 在多个航拍数据集上的实验验证了本文方法的有效性, 在新的数据集上只需 30-shot 训练即可达到接近经典航拍图像目标检测方法全样本训练的精度。

1　相关工作

1.1　经典航拍图像目标检测方法

针对航拍图像背景多变、目标尺寸小等特点, 在现有通用目标检测器的基础上, 研究者们提出一系列优秀的航拍图像目标检测方法。例如, ClusDet[6] 通过设计群集网络获得航拍图像中的密集困难区域, 对其进行增强检测, 有效提升密集小目标检测的精度和效率; DMNet[7] 通过引入人群计数领域的密度掩膜图, 用其反映图像中的目标分布情况, 得到密集目标的区域, 从而进一步提升检测精度; GLSAN[8] 提出一个渐进性式的尺度变化网络, 并加入局部超分辨率网络, 有效提升密集小目标和分布不均匀目标的检测精度; CDMNet[9] 设计粗粒度密度图, 相较于 DMNet 利用密度图描述目标分布, CDMNet 提取密度图中的目标尺度信息来实现目标聚类并量化目标尺度, 进而提升目标检测精度。

上述方法提升航拍图像目标检测精度的核心思路是对局部困难区域进行增强检测, 其中, 增强处理流程大致可分为局部困难区域生成和局部区域特征增强 2 个阶段: 局部困难图像区域通常是一些目标密集且可能存在大量遮挡的区域, 可采用聚类方法、群集网络、密度图等方法得到局部困难区域; 得到局部困难图像区域后, 可以进一步利用超分辨率、图像自适应缩放等方法进行局部区域特征增强。最后, 将得到的局部增强检测结果和原始输入图像整图全局检测结果进行融合, 得到最终的更优异的检测结果如图 1 所示。

1.2　小样本航拍图像目标检测方法

尽管上述基于深度学习的航拍图像目标检测方法取得了突破, 但这些方法存在一个共同前提是

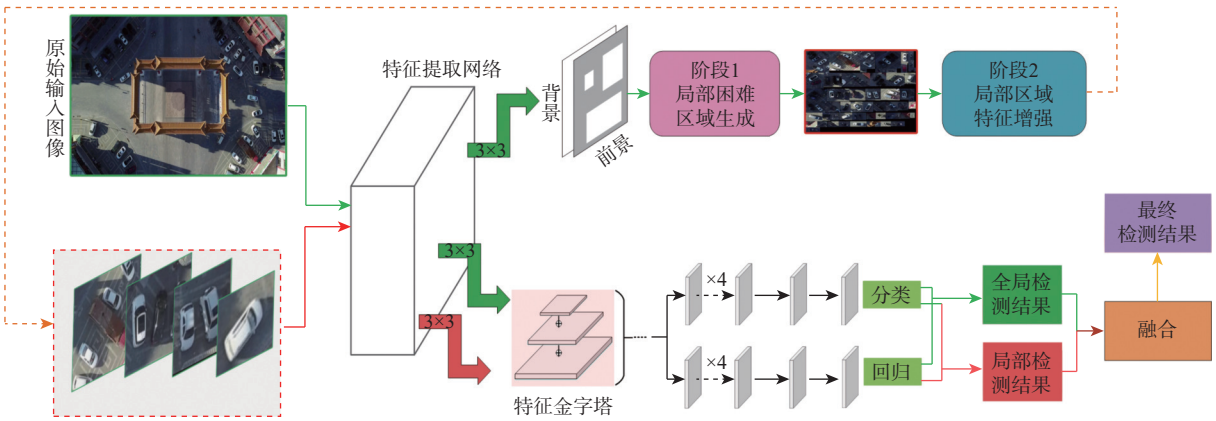


图1 现有的航拍图像目标检测方法处理流程

Fig. 1 Implementation process of existing aerial image object detection method

需要用一个相对较大的航拍数据集来训练该方法模型。这些方法在不同数据集之间的迁移泛化能力相对较弱，尤其是当新数据集与原数据集的背景和目标类别差异大时，仅使用原航拍图像数据集训练后的模型，对新数据集中的新类别样本检测能力大大下降。甚至在新数据集全训练后，可能由于航拍图像目标检测方法针对原数据集进行的特定设计不能在新数据集中起到更好的辅助作用，使得检测结果不好。想要取得更好的检测精度，需要对数据集或检测方法进行调整。但是对数据集已有类别进行调整需要花费大量金钱进行补充标注，而进行检测方法的重新设计和调整，同样也要花费大量时间去重新训练方法模型。

针对这一情况，近年来，小样本目标检测[10-11]越来越受到研究者们的重视。例如，基于通用目标检测器，Yan 等[12]在快速区域卷积神经网络(faster region-based convolutional neural networks, Faster R-CNN[13])基础上设计了检测头重构网络(predictor-head remodeling networks, PRN)模块，用元学习的方法对感兴趣区域上产生的特征图进行处理并完成预测，进而实现小样本目标检测；Fan 等[14]基于 Faster R-CNN 提出包含带注意力机制的区域生成网络、多关联检测器和对比训练策略的小样本目标检测方法，通过度量小样本支持集和查询集之间的相似性来解决小样本的问题，同时很好地抑制了背景干扰；Sun 等[15]提出基于对比建议框编码的小样本目标检测方法，在文献 [13] 的基础上引入对比学习的方法，减小类内差异并增大类间差异，进而提高小样本目标检测的准确度；Han 等[16]认为通过改进现有检测头中查询和支持分支之间的交互、提高相似性学习的方法忽视了特征提取，进而提出基于全交叉 Transformer 方法，该方法将交叉注意力合并到特征主干和检测头，通过非对称批次的交叉注意力聚合来自不同批次大小的 2 个分支的关键信息，引入多层交互来改进 2 个分支之间的小样本相似性学习。

在这些方法的基础上，研究者们根据遥感航拍图像的特点，设计了一系列小样本目标检测方法。Deng 等[17]设计了元特征提取模块，用于从输入的查询图像中学习多种不同尺度的元特征，结合重新加权模块和边界框预测模块完成航拍图像小样本目标检测，提升了小样本检测中新类样本的泛化能力；Cheng 等[18]提出基于原型 CNN 的小样本目标检测，设计了原型指导区域生成网络，在一定程度上解决了小样本中新样本数量较少的问题；Huang 等[19]针对新类样本和基类样本数量严重失衡的问题，提出平衡微调策略和带有注意力的微调网络；李红光等提出基于元学习小样本遥感图像目标检测[20]，引入元学习技术并通过设计多尺度重加权与场景修正方法，来克服小样本条件下目标与背景易混淆的问题。Zhou 等[21]采用上下文感知的像素聚集和特征融合来解决遥感图像小样本目标检测中的尺度变化问题。

1.3 语言结合视觉的小样本目标检测方法

上述小样本航拍图像目标检测方法，在通用小样本目标检测的基础上，针对遥感航拍图像的特点提出各种改进模块，来进一步提升遥感图像中新类的检测精度。然而这些小样本学习的方式，主要在于设计更优的策略对遥感数据集中目标图像的特征进行挖掘，相对更依赖于数据集的分布特点，在跨数据集进行目标检测时泛化性能依然较弱。

如何使得小样本目标检测方法更有效快速地扩展到不同类别的数据集，即方法的迁移能力得到提升，成为研究者们关注的热点和难点问题。一种传统的思路，是构建更加鲁棒的、通用的、包含数万类别及数亿标注目标的大规模数据集。在这样的数据集上进行训练后，目标检测方法在一定程度上可以相对有效地扩展到其他数据集上。但这种

方法不论是数据集的制作成本, 还是后期对模型训练的硬件成本均是非常可观的。另一方面, 一种对比语言-图像预训练(contrastive language-image pre-training, CLIP[22])被提出, 文本被作为模型的监督信号来训练可迁移的视觉模型, 通过引入具有高级语义特征的语言模态数据来提升视觉模型的迁移能力。相较于构建大规模视觉数据集, 语言-图像对的获取难度和成本大大降低, 可获取的样本数量也是非常巨大的。但 CLIP 的强大性能主要体现在图像分类任务上, 如果直接应用到目标检测任务上性能不佳。这是因为在 CLIP 的训练过程中, 网络学习的是一整张图像和相对应的文字描述匹配, 而非图像中的某一区域, 所以 CLIP 并未学习到更细粒度的文本图像区域表示。

在此基础上, 基于区域的对比语言-图像预训练(region-based contrastive language-image pretraining, Region-CLIP[23])被提出, 研究者通过构建一种不需要人工标注的区域-文本对齐方法来克服 CLIP 中只有图像-文本匹配对的困难, 进而实现语言结合视觉的目标检测。Region-CLIP 通过使用 CLIP 的视觉编码器来提取对应区域的视觉特征, 计算其和文本池中的文本语言向量的距离, 得分最高者被分配到对应区域构成文本-区域匹配对, 用于后期的训练。与此同时, 针对缺少文本-区域标记匹配对的问题, 谷歌提出一种采用 CLIP 进行零样本目标检测的方法[24], 通过使用视觉和语言模型共同挖掘未标注图像, 来生成语义丰富的边界框-文本伪标签, 进而能够更好地利用未标注图像数据训练目标检测网络, 克服 CLIP 在进行目标检测任务时由于定位质量不高导致的检测性能不佳的问题。上述这些语言结合视觉的小样本或零样本目标检测方法在 COCO 等公开数据集上展现了其优越的检测性能。受到这些优秀方法的启发, 本文在传统视觉目标检测模型的基础上, 引入了词语-区域对齐分支替代经典检测网络中的分类分支, 同时设计语言视觉偏置网络, 依据图像视觉特征对语言特征进行自动学习调整, 最终构成了一个性能优越的语言引导视觉的小样本航拍图像目标检测方法。

2　本文方法

2.1　整体框架

本文方法的整体网络框架如图 2 所示, 可分为词语-区域对齐分支和定位分支 2 部分。首先, 词语-区域对齐分支对给定的目标类别文本进行语言编码, 得到词语特征 $\boldsymbol{P}$, $\boldsymbol{P}$通过语言视觉偏置网络调整后, 得到与输入图像匹配度更高的词语特征 $\boldsymbol{P}^0$, 基于转换器的双向编码器表示(bidirectional encoder representations from transformers, BERT[25])中的 BERT 层, 期间不断与区域特征进行深度融合得到最终的词语特征 $\boldsymbol{P}'$; 定位分支对给定的图像进行视觉编码得到区域特征 $\boldsymbol{O}^0$, 然后, 通过动态头(dynamic head, DyHead[26]): 基于注意力机制的目标检测头统一框架中的多个 DyHead 模块, 期间不断和词语特征进行深度融合, 得到最终图像区域特征 $\boldsymbol{O}'$, 用于进行目标的位置定位损失计算; $\boldsymbol{O}'$和 $\boldsymbol{P}'$共同构成了词语-区域对齐分数, 该分数取代了原有检测模型中分类分支的分类分数, 用于计算词语-区域对齐损失, 并指导目标类别的识别。

2.2　词语-区域对齐分支

传统的航拍图像目标检测方法, 大多是在经典的目标检测方法上根据航拍图像目标密集、目标小且存在遮挡等特点进行改进的。这些方法针对特定的航拍数据集得到优秀检测结果的同时, 如果直接迁移到另一个航拍数据集上时则表现不佳, 究其原因在于不同航拍数据集差异大, 体现为拍摄视角、图像质量、照明条件、背景环境变化大, 目标外观变化明显、目标类别容易新增。针对这一特点, 如果仅从视觉图像角度入手, 难以有效提升模型

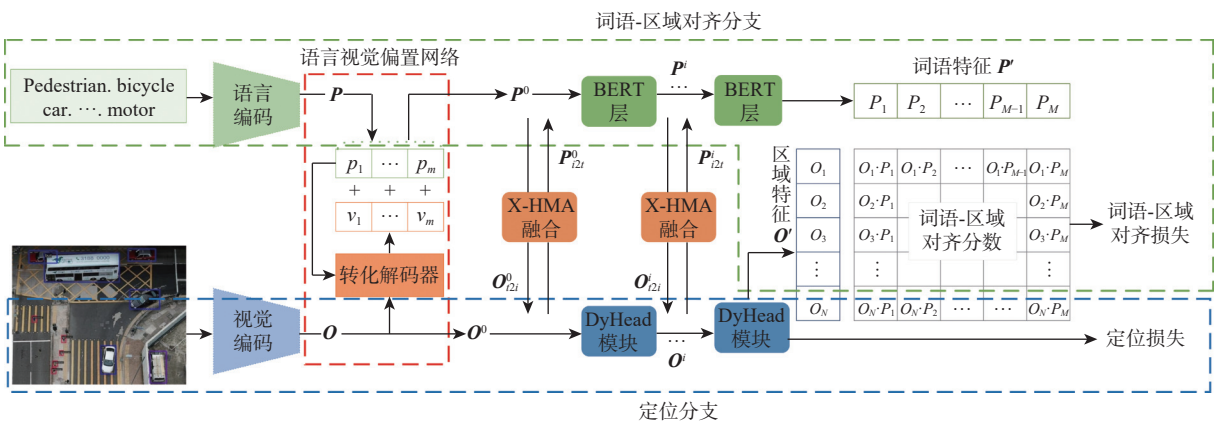


图 2　本文方法的整体框架

Fig. 2　Overall framework of the proposed method

迁移后的目标检测精度。本文引入跨模态对齐的语言-图像预训练(grounded language-image pre-training, GLIP[5])中的词语-区域对齐分支替代经典检测网络中的分类分支,通过将图像目标检测和短语定位任务统一为一个任务,利用大量的语言信息引导视觉分类定位,不仅提升了已有类别目标的检测精度,还大幅度提升模型迁移后对新类别目标的检测精度。

词语-区域对齐分支如图2绿色虚线框内所示。以VisDrone数据集为例,首先,将数据集中存在的目标类别组成一个句子,这里对应为Prompt = “pedestrian. people. bicycle. ⋯. motor”,接下来采用预训练BERT[25]模型进行语言编码提取特征$\boldsymbol{P}$,通过语言视觉偏置网络得到特征$\boldsymbol{P}^0$。然后,与定位分支的区域特征$\boldsymbol{O}^0$通过跨模态多头注意力(cross-modality multi-head attention, X-MHA)[5]模块融合得到$\boldsymbol{P}^0_{i2t}$,并与$\boldsymbol{P}^0$一同输入到BERT层中得到$\boldsymbol{P}^1$,重复多次,最终得到词语特征$\boldsymbol{P}'$。在定位分支中,VisDrone中的图像数据,首先,通过DyHead[26]模型进行视觉编码提取特征$\boldsymbol{O}^1$,一方面作为语言视觉偏置网络的输入,另一方面作为$\boldsymbol{O}^0$($\boldsymbol{O}^0$==$\boldsymbol{O}$)与词语特征$\boldsymbol{P}^0$通过X-MHA融合合得到$\boldsymbol{O}^0_{i2i}$,同时输入DyHead模块得到$\boldsymbol{O}^1$,重复多次,得到区域特征$\boldsymbol{O}'$。最后,通过将词语和区域特征进行点积得到词语-区域对齐分数$\boldsymbol{S}_{\text{ground}}$,完成语言引导图像分类。大致流程为

$$\begin{cases}\boldsymbol{O}=\text{Enc}_{\text{I}}(\boldsymbol{I}_{\text{img}})\\ \boldsymbol{P}=\text{Enc}_{\text{L}}(\boldsymbol{P}_{\text{prompt}})\\ \boldsymbol{S}_{\text{ground}}=\boldsymbol{O}'\boldsymbol{P}'^{\text{T}}\end{cases}\tag{1}$$

式中:Enc_{I}为视觉编码;$\boldsymbol{I}_{\text{img}}$为输入图像;$\boldsymbol{O}$为提取的区域特征;$\text{Enc}_{\text{L}}$为语言编码;$\boldsymbol{P}_{\text{prompt}}$为输入的数据类别组成的句子;$\boldsymbol{P}$为提取的词语特征;$\boldsymbol{S}_{\text{ground}}$为词语-区域对齐分数;$\boldsymbol{O}'$和$\boldsymbol{P}'$为经过一系列融合后得到的最终区域和视觉特征;$\boldsymbol{P}'^{\text{T}}$为词语特征$\boldsymbol{P}'$的转置。

在目标检测过程中,为提取更有利于检测的目标图像特征,会进行高层语义和低层细节特征各种形式的特征融合。词语-区域对齐分支为更好地利用语言和视觉信息,使得视觉特征拥有语言感知能力,进行了语言视觉深度融合。融合操作为:采用X-MHA,对区域特征$\boldsymbol{O}^i$和词语特征$\boldsymbol{P}^i$这2种模态特征进行融合,随后再利用DyHead模块和BERT层对视觉区域特征和语言词语特征进行单一模态的融合和更新,具体计算式为

$$\boldsymbol{O}^i_{t2i},\boldsymbol{P}^i_{i2t}=\text{X-MHA}(\boldsymbol{O}^i,\boldsymbol{P}^i)\qquad i\in\{0,1,\cdots,L-1\}\tag{2}$$

$$\boldsymbol{O}^{i+1}=\text{DyHead}(\boldsymbol{O}^i+\boldsymbol{O}^i_{t2i})\qquad \boldsymbol{O}'=\boldsymbol{O}^L\tag{3}$$

$$\boldsymbol{P}^{i+1}=\text{BERT}(\boldsymbol{P}^i+\boldsymbol{P}^i_{i2t})\qquad \boldsymbol{P}'=\boldsymbol{P}^L\tag{4}$$

式中:X-MHA为采用跨模态多头注意力模块的融合方式;L为BERT层和DyHead模块的数量;$\boldsymbol{P}^0$为词语特征$\boldsymbol{P}$通过语言偏置网络后获得的词语特征;$\boldsymbol{O}^0$为原始图像通过视觉编码器后得到的区域特征;$\boldsymbol{O}^i_{t2i}$和$\boldsymbol{P}^i_{i2t}$分别为X-MHA深度融合后的区域特征和词语特征。

航拍图像目标检测模型通常被训练成针对特定航拍数据集、目标类别固定的模型,如CDMNet针对Visdrone数据集训练后,只对Visdrone的10类目标检测效果较好,这极大地限制了模型的泛化能力和实际推广效果,因为每次更换场景和数据集,便面临着额外的标注数据工作和模型重新训练工作。本文通过引入词语-区域对齐分支,在原有只使用视觉特征的基础上增加了对应的语言特征。由于各种无人机场景的文本描述数据远远多于无人机图像数据,故而本文方法可以在少量样本、甚至零样本条件下实现在不同航拍数据集上的迁移,且接近经典航拍图像目标检测方法在全样本训练时的检测精度。

2.3 语言视觉偏置网络

词语-区域对齐分支有效引入文本语言数据,从而在不需要大量图像标注数据的情况下,使得本文方法的小样本目标检测精度接近全样本训练精度,同时迁移能力得到极大的提升。然而,本文发现在语言输入端输入不同的语言文本,会造成检测精度的波动。例如,当在VisDrone数据集上进行零样本实验时,数据集类别组成Prompt =“pedestrian. people. bicycle. ⋯. motor”中的motor被修改为motorcycle,检测精度则出现明显变化。出现这种情况的原因在于语言视觉信息特征的匹配程度不够高。针对这一情况,一方面可以考虑人工直接改变语言的输入,另一方面可以考虑对语言特征依据图像视觉特征进行自动学习调整,进而为后期两者的融合提供更有效的支撑。

考虑到人工标注的随意性,以及现有各种不同无人机图像数据集标注类别的差异,为方便各数据集之间的迁移,本文选择了后者,即对语言特征依据图像视觉特征进行自动学习调整。设计语言视觉偏置网络,整体结构如图2红色虚线框所示。该网络通过使用转化解码器,将输入语言词语特征$\boldsymbol{P}$和其对应的视觉区域特征$\boldsymbol{O}$建模为一组连续特征表示,对其进行学习并生成归纳偏置权重$\boldsymbol{V}$,然后,与输入语言词语特征$\boldsymbol{P}$融合得到被视觉信息微调后的语言词语特征$\boldsymbol{P}^0$,这样使得语言词语特征$\boldsymbol{P}^0$具有更好的伸缩性[27],达到与每张输入图像区域更适配

的效果。

转化解码器结构如图 3 所示, 将语言编码器得到的语言特征 $\boldsymbol{P}=[p_1,p_2,\cdots,p_m]$, 首先, 输入到多头自注意力中, 依次进行特征相加和层标准化后, 与视觉特征 $\boldsymbol{O}$ 一起输入到交叉注意力, 依次进行特征相加和层标准化后, 将得到的特征再次输入到多头自注意力中, 依次进行特征相加和层标准化后, 再与视觉特征 $\boldsymbol{O}$ 一起输入到交叉注意力, 依次进行特征相加和层标准化后, 得到最终的归纳偏置权重 $\boldsymbol{V}=[v_1,v_2,\cdots,v_m]$, 其中, m 为数据集类别。语言视觉偏置网络的整个过程可以归纳为

$$\begin{cases}\boldsymbol{V}=\text{TransDecoder}(\boldsymbol{P},\boldsymbol{O})\\ \boldsymbol{P}^0=\boldsymbol{P}+\boldsymbol{V}\end{cases}\tag{5}$$

式中: TransDecoder 表示转化解码器。

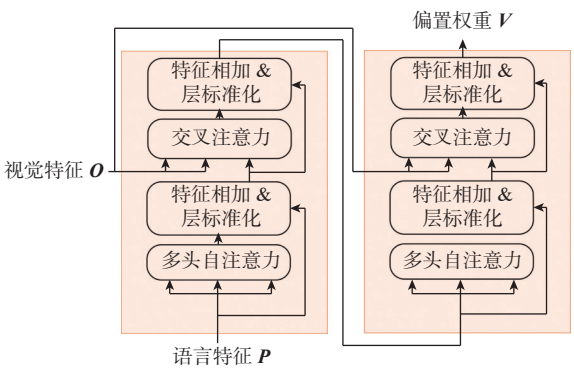


图 3　转化解码器

Fig. 3　Transformation decoder

输入的语言词语特征 $\boldsymbol{P}$ 通过多头自注意力、特征相加和层标准化操作, 学习到语言特征中更具有上下文信息的稳定特征, 进而和对应的视觉区域特征 $\boldsymbol{O}$ 交叉学习生成与视觉目标更贴近的语言特征。上述操作执行后, 生成的归纳偏置权重 $\boldsymbol{V}$ 更好地刻画了视觉特征与语言特征之间稳定的关联关系。通过采用语言视觉偏置网络, 将原本输入的语言特征 $\boldsymbol{P}=[p_1,p_2,\cdots,p_m]$ 根据视觉信息进行微调, 生成与视觉目标更贴近的语言特征, 即原本固定的语言特征有了学习能力。如原本固定语言输入分别为 motor 和 motorcycle 时, 会导致不同的检测精度, 而且输入 motorcycle, 精度相对更高。语言视觉偏置网络使得输入语言特征有了学习能力后, 让原本输入为 motor 的语言文本在结合视觉特征后生成更贴近 motorcycle, 甚至是更好的语言特征, 进而提升小样本的检测结果, 同时提升了模型的泛化能力。

3　实验分析与比较

3.1　实验设置

本文实验基于 Facebook 的 MaskRCNN-Benchmark 工具箱完成。实验的预训练模型选择 GLIP-T[6], 其中, 预训练使用大规模自然图像数据集 Object365 和自然场景图文匹配数据集 GoldG 训练完成, 实验新类数据集为 UAVDT、Visdrone、AeriaDrone、VEDAI、CARPK_PUCPR 无人机数据集。实验环境为 TeslaV-100, 显存 32 GB。采用单尺度训练, 输入图像大小设置短边最低为 800 像素, 长边最长为 1 333 像素, 批量大小(batch_size)均为 4。优化器选择 AdamW, 权重衰减为 0.05, 学习率预热为 2 000 个迭代, 学习率为 0.000 1。评测指标采用 COCO 检测评估指标。

3.2　实验结果

为验证本文方法的有效性, 本文对提出的词语-区域对齐分支和语言视觉偏置网络分别在 5 个具有不同风格环境的航拍数据集上进行验证, 这 5 个数据集的部分图像和类别展示分别如图 4 和表 1 所示。由表 1 可知, 这些数据集的类别存在较多不同, 由图 4 可知, 5 种无人机所拍摄的环境风格大不相同, 不仅观察距离存在差异, 而且拍摄视角、图像质量、照明条件、背景环境也存在较大不同。通过这些数据集可以很好地验证本文方法的迁移能力。与此同时, 本文将多种最新经典航拍图像方法、多种最新通用目标检测方法在全样本训练的情况下和本文方法的小样本训练进行了精度比较。

3.2.1　消融实验

为验证词语-区域对齐分支和语言视觉偏置网络的有效性, 本文进行了大量的实验, 实验结果如表 2 和图 5 所示, 表中, 数据加黑是指该数值是该列中的最高值。其中, Baseline 表示只使用传统视觉目标检测 DyHead 模型, Baseline+词语区域对齐分支表示在 DyHead 模型基础上增加了词语-区域对齐分支。Baseline+all 表示同时使用词语-区域对齐分支和语言视觉偏置网络。实验结果均为平均精度均值(mean average precision, mAP), 由每类的平均精度(average precision, AP)(交并比为 0.50∶0.05∶0.95)值, 从交并比为 0.5 到交并比为 0.95, 每隔



(a) Visdrone[1]　(b) AeriaDrone[5]　(c) UAVDT[2]　(d) VEDAI[4]　(e) CARPK_PUCPR[3]

图 4　无人机数据集展示

Fig. 4　Display of drone datasets

表 1 无人机数据集类别分析

Table 1 Classification analysis of drone datasets

数据集	类别	样本数量
Visdrone[1]	pedestrian, people, bicycle, car, van, truck, tricycle, awning-tricycle, bus, motor	54 200
AerialDrone[5]	docks, boats, lifts, jetskis, cars	1 600
UAVDT[2]	car, truck, bus	841 500
VEDAI[4]	boat, camping car, pickup, plane, tractor, truck, vans	733
CARPK_PUCPR[3]	car	89 800

表 2 消融实验

Table 2 Ablation experiment

数据集	方法	mAP/%				
		1-shot	3-shot	5-shot	10-shot	30-shot
AerialDrone[5]	Baseline	0	17.5	23.8	30.5	33.1
	Baseline + 词语-区域对齐分支	19.4	28.9	33.0	37.4	42.9
	Baseline + all	**19.8**	**32.7**	**33.7**	**39.0**	**43.1**
VisDrone-DET[1]	Baseline	0.4	1.1	2.1	1.3	1.6
	Baseline + 词语-区域对齐分支	15.9	16.6	18.3	18.7	21.8
	Baseline + all	**16.1**	**17.2**	**18.4**	**18.8**	**21.8**
CARPK_PUCPR[3]	Baseline	0	1.4	11.1	15.3	30.3
	Baseline + 词语-区域对齐分支	46.2	52.0	55.5	56.0	57.8
	Baseline + all	**46.9**	**55.0**	**56.4**	54.2	**58.0**
VEDAI[4]	Baseline	0	3.6	6.8	8.1	17.4
	Baseline + 词语-区域对齐分支	8.7	11.6	12.7	13.8	26.8
	Baseline + all	8.2	**13.8**	**15.1**	**15.6**	**27.1**

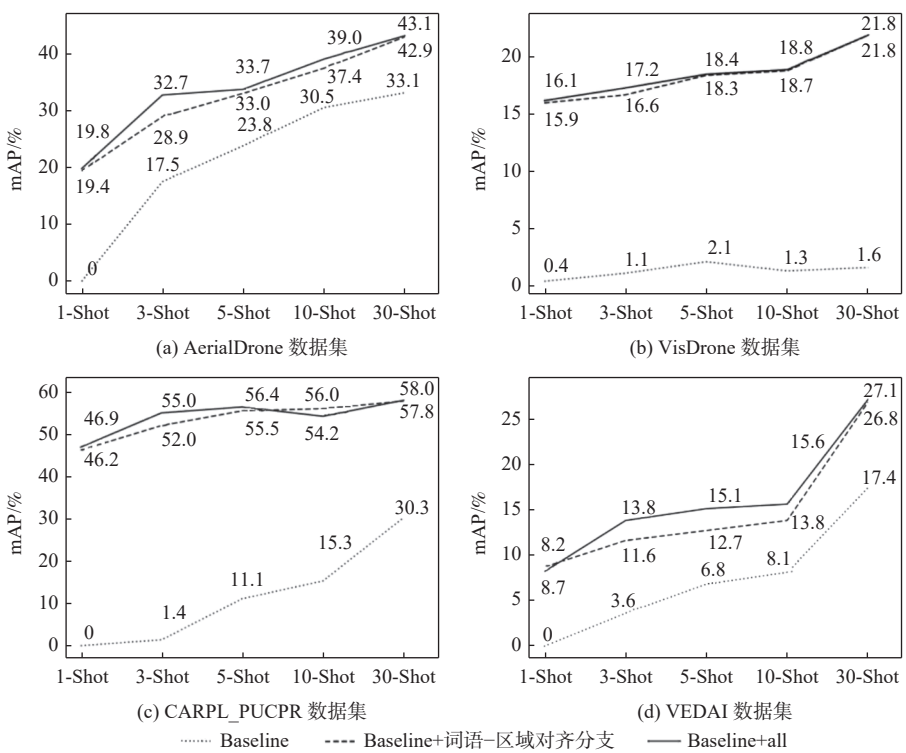


图 5 消融实验的可视化结果

Fig. 5 Visualization results of ablation experiment

0.05 个交并比计算一次 AP, 再求各类的平均值所得。

通过实验可以发现传统视觉目标检测 DyHead 模型仅使用常规视角自然图像进行预训练的情况

下, 在多个不同航拍数据集上的小样本检测性能都比较弱。尤其在 5-shot 以下, 多个数据集上 mAP 都比较差。在 1-shot 训练的情况下, 不采用词语-区域对齐分支, 在 VisDrone-DET、AerialDrone、CARPK_PUCPR 和 VEDAI(CO) 数据集上的 mAP 都只有 0% 左右; 而有词语-区域对齐分支的 mAP 则表现良好, 在针对只有 car 这一常见类别的 CARPK_PUCPR 数据集, 其 1-shot mAP 甚至从 0% 提升到了 46.2%。在 VisDrone-DET 数据集上, 当数据集类别较多、样本目标小且遮挡严重、数据量较大时, 不采用词语-区域对齐分支的情况下, 在 30-shot 训练时 mAP 只有 1.6%, 比 5-shot 的 2.1% 还低 0.5%, 相对 1-shot 训练无法有效提升 mAP。同时也从侧面反映单纯使用视觉模型, 在 30-shot 以内的小样本训练的情况下, 网络模型甚至不能完全有效收敛。而采用词语-区域对齐分支, 增加了语言模态后, 1-shot 训练便有 15.9% 的 mAP, 在增加到 30-shot 训练后, mAP 从 15.9% 提升到了 21.8%。由表 2 和图 5 可知, 在采用了词语-区域对齐分支之后, 在 UAVDT、Visdrone、AerialDrone、VEDAI、CARPK_PUCPR 数据集上, 在 30-shot 样本时 mAP 分别可达到 14.6%、21.8%、43.1%、27.1%、58.0%, mAP 有了性能的飞跃。这些实验结果整体验证了词语-区域对齐分支可以在大量非标注无人机图像上, 通过其对应文本语言的引导, 得到更高的 mAP。

进一步, 由表 2 和图 5 可知, 通过在词语-区域对齐分支基础上增加语言视觉偏置网络, 从 1-shot 到 30-shot, 在 VisDrone-DET、AerialDrone、CARPK_PUCPR 和 VEDAI(CO) 这 4 个数据集上, mAP 基本都得到了一定程度的提升。其中, 在 AerialDrone 数据集上, 增加语言视觉偏置网络后, 3-shot mAP 提升了 3.8%, 在 CARPK_PUCPR 数据集上, 5-shot mAP 提高了 0.9%, 在 VEDAI(CO) 数据集上, 10-shot 提升了 1.8% 的 mAP。这些实验共同验证了本文提出的语言视觉偏置网络, 可以根据输入图像视觉特征对输入语言的语言特征进行自适应地学习调整, 能够更好地匹配视觉和语言的关系, 使得区域特征与词语特征得到更好融合, 进而缩小了小样本的波动, 并提升了小样本的整体 mAP。

3.2.2　与其他方法的比较

为进一步验证本文方法的性能, 本文在 VisDrone-DET 和 UAVDT 数据集上与近年来提出的经典航拍图像目标检测方法进行对比。此时, 本文方法仅采用 5/10/30-shot 进行训练, 而对比方法采用训练集全样本进行训练, 其中, VisDrone-DET 和 UAVDT 训练集样本均超过 10 000 个目标样本。如表 3 所示, 在 UAVDT 数据集上, 本文方法在 30-shot 的 mAP 可达 14.6%, 相比 ClusDet、DMNet、GLSAN 和 CDMNet 在全样本的 mAP 分别提高了 0.9%、−0.1%、−2.4% 和−2.2%, 本文方法非常具有竞争力。此外, 现有这些经典航拍图像目标检测方法在全样本训练的情况下, 在 VisDrone-DET 数据集上相对本文方法 30-shot 有着 5%～10% 点的 mAP 优势。其主要原因在于 VisDrone-DET 类别相对较多, 部分少见样本在语言训练样本中也相对少见, 同时, 在训练和推理数据集中分布不均匀, 导致小样本训练时, 对该类样本学习相对不足。

表 3　与经典航拍图像目标检测方法对比

Table 3　Comparison with aerial image object detection

数据集/方法	mAP/%				
	ClusDet[6]	DMNet[7]	GLSAN[8]	CDMNet[9]	本文方法 (5/10/30-shot)
VisDrone-DET	26.7	28.2	30.7	29.2	18.4/18.8/21.8
UAVDT	13.7	14.7	17.0	16.8	10.1/13.6/14.6

另一方面, 在 VisDrone-DET 和 CARPK_PUCPR 数据集上与近年来的通用目标检测方法进行对比。本文对全卷积单阶段目标检测器(fully convolutional one-stage object detector, FCOS)、自适应训练样本选择(adaptive training sample selection, ATSS)、广义焦点损失 V2(generalized focal loss V2, GFLV2)和交并比感知密集目标检测器(VarifocalNet, VFNET)方法在 2 个数据集上进行了全样本训练, 并进行推理。如表 4 所示, 这些方法在 VisDrone-DET 数据集上进行全样本训练, 相对本文方法 30-shot 同样有着 5%～10% 的 mAP 优势。但在类别相对较少的 CARPK_PUCPR 数据集上, 本文方法在 30-shot 时 mAP 可达 58%, 相比 FCOS、ATSS、GFLV2 和 VFNET 分别高了 1.0%、0.8%、0.1% 和 0.3%, 而本文方法 5-shot 的结果也和这些方法全样本训练的检测结果相当。这表明: 本文方法在新数据集下即使只通过极少量样本的训练, 也能达到与通用目标检测 SOTA 方法相当的 mAP。由表 3 和表 4 可知, 相较于传统视觉模型通常需要成千上万的数据才能实现充分训练和迁移, 本文方法则只需要不足 100 张甚至 30 张图像, 在不到 20 min 的训练时间, 即可达到相近的 mAP。故而无论是在新数据集的

表 4　与经典通用目标检测方法对比

Table 4　Comparison with classical general object detection methods

数据集/方法	mAP/%				
	FCOS[28]	ATSS[29]	GFLV2[30]	VFNET[31]	本文方法 (5/10/30-shot)
VisDrone-DET	27.2	28.8	29.3	30.2	18.4/18.8/21.8
CARPK_PUCPR	57.0	57.2	57.9	57.7	56.4/54.2/58.0

样本标注代价上，还是训练所需时间，本文方法的实现效率高效且性能优越。

3.2.3 可视化检测结果

图6为本文方法在Visdrone、AeriaDrone、UAVDT、VEDAI、CARPK_PUCPR数据集上30-shot的检测效果。由图可知，在30-shot的检测推理中，对图像中尺寸相对较小且目标排列密集的目标基本实现正确检测。这些实验共同证明了本文方法拥有良好的小样本学习能力和迁移学习，能够适配无人机航拍图像的小样本目标检测任务。



(a) Visdrone[1] (b) AeriaDrone[5] (c) UAVDT[2] (d) VEDAI[4] (e) CARPK_PUCPR[3]

图6 本文方法30-shot的可视化结果

Fig. 6 Visualization results of the proposed method at 30-shot

4 结 论

本文方法在UAVDT、Visdrone、AeriaDrone、VEDAI、CARPK_PUCPR数据集上，在30-shot时mAP分别可以达到14.6%、21.8%、43.1%、27.1%、58.0%。

1）本文方法采用词语-区域对齐分支，将图像目标检测和短语定位任务统一为一个任务，利用语言信息特征引导视觉分类定位，在提升已有类别的检测精度的基础上，还大幅度提升模型迁移后对新类别目标的检测精度。

2）本文方法所设计的语言视觉偏置网络，可以进一步挖掘语言特征和视觉特征的关联关系，提升语言视觉的匹配度，缩小精度波动，同时提升小样本检测精度。

在本文基础上，可进一步研究如何利用视觉-语言预训练大模型增强对未知类别的分类检测理解等任务的性能。

参考文献（References）

[1] DU D W, WEN L Y, ZHU P F, et al. VisDrone-DET2020: the vision meets drone object detection in image challenge results[C]//Proceedings of the Computer Vision – ECCV 2020 Workshops. Berlin: Springer, 2020: 692-712.

[2] DU D W, QI Y K, YU H Y, et al. The unmanned aerial vehicle benchmark: object detection and tracking[M]//Computer Vision – ECCV 2018. Berlin: Springer, 2018: 375-391.

[3] HSIEH M R, LIN Y L, HSU W H. Drone-based object counting by spatially regularized regional proposal network[C]//Proceedings of the IEEE International Conference on Computer Vision. Piscataway: IEEE Press, 2017: 4165-4173.

[4] RAZAKARIVONY S, JURIE F. Vehicle detection in aerial imagery: a small target detection benchmark[J]. Journal of Visual Communication and Image Representation, 2016, 34: 187-203.

[5] LI L H, ZHANG P C, ZHANG H T, et al. Grounded language-image pre-training[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2022: 10955-10965.

[6] YANG F, FAN H, CHU P, et al. Clustered object detection in aerial images[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE Press, 2019: 8310-8319.

[7] LI C L, YANG T, ZHU S J, et al. Density map guided object detection in aerial images[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. Piscataway: IEEE Press, 2020: 737-746.

[8] DENG S T, LI S, XIE K, et al. A global-local self-adaptive network for drone-view object detection[J]. IEEE Transactions on Image Processing, 2020, 30: 1556-1569.

[9] DUAN C Z, WEI Z W, ZHANG C, et al. Coarse-grained density map guided object detection in aerial images[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops. Piscataway: IEEE Press, 2021: 2789-2798.

[10] 宋闯, 赵佳佳, 王康, 等. 面向智能感知的小样本学习研究综述[J]. 航空学报, 2020, 41(S1): 723756.
SONG C, ZHAO J J, WANG K, et al. A review of small sample learning for intelligent perception[J]. Acta Aeronautica et Astronautica Sinica, 2020, 41(S1): 723756(in Chinese).

[11] 段立娟, 袁莹, 王文健, 等. 基于多模态联合语义感知的零样本目标检测[J]. 北京航空航天大学学报, 2024, 50(2): 368-375.
DUAN L J, YUAN Y, WANG W J, et al. Zero-shot object detection based on multi-modal joint semantic perception[J]. Journal of Beijing University of Aeronautics and Astronautics, 2024, 50(2): 368-375(in Chinese).

[12] YAN X P, CHEN Z L, XU A N, et al. Meta R-CNN: towards general solver for instance-level low-shot learning[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE Press, 2019: 9576-9585.

[13] REN S Q, HE K M, GIRSHICK R, et al. Faster R-CNN: towards real-time object detection with region proposal networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(6): 1137-1149.

[14] FAN Q, ZHUO W, TANG C K, et al. Few-shot object detection with attention-RPN and multi-relation detector[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2020: 4012-4021.

[15] SUN B, LI B H, CAI S C, et al. FSCE: few-shot object detection via contrastive proposal encoding[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscat-

away: IEEE Press, 2021: 7348-7358.

[16] HAN G X, MA J W, HUANG S Y, et al. Few-shot object detection with fully cross-transformer[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2022: 5311-5320.

[17] DENG J Y, LI X, FANG Y. Few-shot object detection on remote sensing images[EB/OL]. (2020-06-14)[2023-02-01]. https://arxiv.org/abs/2006.07826v2.

[18] CHENG G, YAN B W, SHI P Z, et al. Prototype-CNN for few-shot object detection in remote sensing images[J]. IEEE Transactions on Geoscience and Remote Sensing, 2021, 60: 5604610.

[19] HUANG X, HE B K, TONG M, et al. Few-shot object detection on remote sensing images *via* shared attention module and balanced fine-tuning strategy[J]. Remote Sensing, 2021, 13(19): 3816.

[20] 李红光, 王玉峰, 杨丽春. 基于元学习的小样本遥感图像目标检测[J]. 北京航空航天大学学报, 2024, 50(8): 2503-2513.
LI H G, WANG Y F, YANG L C. Meta-learning-based few-shot object detection for remote sensing images[J]. Journal of Beijing University of Aeronautics and Astronautics, 2024, 50(8): 2503-2513(in Chinese).

[21] ZHOU Y, HU H, ZHAO J Q, et al. Few-shot object detection via context-aware aggregation for remote sensing images[J]. IEEE Geoscience and Remote Sensing Letters, 2022, 19: 6509605.

[22] RADFORD A, KIM J W, HALLACY C, et al. Learning transferable visual models from natural language supervision[EB/OL]. (2021-02-26)[2023-02-01]. https://arxiv.org/abs/2103.00020.

[23] ZHONG Y W, YANG J W, ZHANG P C, et al. RegionCLIP: region-based language-image pretraining[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2022: 16772-16782.

[24] ZHAO S Y, ZHANG Z X, SCHULTER S, et al. Exploiting unlabeled data with vision and language models for object detection [M]//Computer Vision – ECCV 2022. Berlin: Springer, 2022: 159-175.

[25] DEVLIN J, CHANG M W, LEE K, et al. BERT: pre-training of deep bidirectional transformers for language understanding[EB/OL]. (2019-05-24)[2023-02-01]. https://arxiv.org/abs/1810.04805v2.

[26] DAI X Y, CHEN Y P, XIAO B, et al. Dynamic head: unifying object detection heads with attentions[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2021: 7369-7378.

[27] ZHOU K Y, YANG J K, LOY C C, et al. Conditional prompt learning for vision-language models[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2022: 16795-16804.

[28] TIAN Z, SHEN C H, CHEN H, et al. FCOS: a simple and strong anchor-free object detector[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 44(4): 1922-1933.

[29] ZHANG S F, CHI C, YAO Y Q, et al. Bridging the gap between anchor-based and anchor-free detection *via* adaptive training sample selection[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2020: 9756-9765.

[30] LI X, WANG W H, HU X L, et al. Generalized focal loss V2: learning reliable localization quality estimation for dense object detection[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2021: 11627-11636.

[31] ZHANG H Y, WANG Y, DAYOUB F, et al. VarifocalNet: an IoU-aware dense object detector[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2021: 8510-8519.

Few-shot object detection of aerial image based on language guidance vision

ZHANG Zhi[1], YI Huahui[1], ZHENG Jin[2,*]

(1. School of Computer Science and Technology, Civil Aviation University of China, Tianjin 300300, China;
2. School of Computer Science and Engineering, Beihang University, Beijing 100191, China)

Abstract: This paper presents a few-shot object detection of aerial images based on language guidance vision in response to the problem of decreased detection accuracy in existing aerial image object detection brought on by a lack of training data and changes in the aerial image dataset, such as changes in shooting angle, image quality, lighting conditions, background environment, and significant changes in object appearance and the addition of new object categories. First, a word-region alignment branch replaces the classification branch in conventional object detection networks, and a word-region alignment classification score with both language and visual information is obtained as a prediction classification result. Then, object detection and phrase grounding are unified into a single task, leveraging language information to enhance visual detection accuracy. Furthermore, to address the challenge of accuracy fluctuations caused by changes in textual prompts during few-shot object detection, a language visual bias network is designed to mine the association between language features and visual features. This network aims to improve the alignment between language and vision, mitigate accuracy fluctuations, and further enhance the accuracy of few-shot detection. Extensive experimental results on UAVDT, Visdrone, AeriaDrone, VEDAI, and CARPK_PUCPR demonstrate the superior performance of the proposed method. Particularly, on the UAVDT dataset, the method achieves an impressive mAP of 14.6% at 30-shot. Compared with aerial image detection algorithms clustered detection (ClusDet), density map guided object detection network (DMNet), global-local self-adaptive network (GLSAN) and coarse-grained density map network (CDMNet), the detection accuracy is 0.9%, −0.1%, −2.4% and −2.2% higher during full data training. The mAP of the approach can achieve 58.0% on the PUCPR dataset at 30-shot, which is 1.0%, 0.8%, 0.1% and 0.3% higher than the accuracy of full data trained generic object detection methods fully convolutional one-stage object detector (FCOS), adaptive training sample selection (ATSS), generalized focal loss V2 (GFLV2) and VarifocalNet (VFNET), respectively. These results highlight the robust few-shot generalization and transfer capabilities of the proposed method.

Keywords: aerial images; few-shot; word-region aligned branch; language visual bias network; object detection

Received: 2023-07-27; **Accepted**: 2023-09-22; **Published Online**: 2023-10-30 20:23

URL: link.cnki.net/urlid/11.2625.V.20231030.1001.001

Foundation items: National Natural Science Foundation of China (62372029); The Fundamental Research Funds for the Central Universities (3122023007)

* **Corresponding author**. E-mail: JinZheng@buaa.edu.cn

http://bhxb.buaa.edu.cn jbuaa@buaa.edu.cn
DOI: 10.13700/j.bh.1001-5965.2023.0534

基于特征增强与自适应阈值非极大值抑制的目标检测算法

孟伟君[1,*]，安雯[1]，马素刚[1,2]，杨小宝[1,3]

(1. 西安邮电大学 计算机学院，西安 710121；2. 陕西省网络数据分析与智能处理重点实验室，西安 710121；
3. 西安市大数据与智能计算重点实验室，西安 710121)

摘 要：为进一步解决目标漏检和重复检测等问题,提升目标检测的性能,提出一种基于特征增强与自适应阈值的非极大值抑制（NMS）目标检测算法。将注意力引导的多尺度上下文模块(AMCM)用于检测器颈部，在利用空洞卷积提升特征语义信息的基础上，通过注意力捕获跨通道位置信息，增强网络的特征表达能力；通过基于目标密度的自适应阈值NMS（ADT-NMS），针对不同场景的实例应用动态抑制阈值，降低目标的误检率。所提算法在PASCAL VOC数据集上误检率为13.7%，相比基准算法YOLOv4降低了1%，检测精度、召回率分别达到83.7%、96.6%，分别提高了1.7%、0.9%；在KITTI数据集上误检率为22.1%，相比基准算法降低了1.3%，检测精度、召回率分别达到83.6%、91.8%，分别提高了1.8%、2.3%。实验结果表明：所提算法较好地解决了目标漏检和重复检测问题。

关 键 词：自适应阈值；非极大值抑制；目标检测；空洞卷积；注意力机制
中图分类号：TP391
文献标志码：A **文章编号**：1001-5965(2025)07-2349-11

目标检测作为计算机视觉领域一个重要的研究方向,广泛应用于自动驾驶、智能监控、缺陷检测等实际场景中。基于深度学习的目标检测方法主要分为2类：①基于卷积神经网络(convolutional neural networks, CNN)和锚框(Anchor)的检测算法,又可细分为两阶段目标检测算法和一阶段目标检测算法,前者先生成感兴趣区域,再执行分类和回归预测任务,如区域卷积神经网络(region-based convolutional neural networks, R-CNN)[1],后者直接对整幅图像进行特征提取,一次性完成分类和坐标回归,如YOLO算法[2];②基于Transformer[3]架构的目标检测算法,不同于基于CNN和Anchor架构的检测器,其利用Cross-attention和Object query提取全局信息生成对所有目标的检测结果,如检测用变换器(detection transformer, DETR)[4]。基于Transformer[3]架构的目标检测算法虽然舍弃了Anchor和非极大值抑制(non-maximum suppression, NMS),但对于硬件要求很高。

随着卷积神经网络的发展,出现了许多具有较强特征提取能力的网络结构,随着网络深度的不断增加,目标信息由浅层向深层传递过程中其边缘轮廓信息逐渐减弱,导致检测器对目标正确预测和定位的能力不足,造成目标漏检问题。

当前基于Anchor的目标检测算法通过在每个位置设置不同大小和宽高比的锚框,用来匹配图像中不同尺度的目标,这种方式可以使网络模型更好

收稿日期：2023-08-24；**录用日期**：2023-10-13；**网络出版时间**：2023-11-14 13:34
网络出版地址：link.cnki.net/urlid/11.2625.V.20231114.0933.001
基金项目：国家自然科学基金（62072370）；陕西省自然科学基金（2023-JC-YB-598）
* **通信作者**. E-mail：mengweijun@xupt.edu.cn

引用格式：孟伟君，安雯，马素刚，等. 基于特征增强与自适应阈值非极大值抑制的目标检测算法 [J]. 北京航空航天大学学报，2025，51（7）：2349-2359. MENG W J，AN W，MA S G，et al. An object detection algorithm based on feature enhancement and adaptive threshold non-maximum suppression [J]. Journal of Beijing University of Aeronautics and Astronautics，2025，51（7）：2349-2359（in Chinese）.

地适应多尺度目标，但是这种匹配机制可能生成数倍于基准真值 (ground truth, GT) 数量的预测结果，因此，需要通过高效的后处理方法进行冗余过滤。传统的 Greedy-NMS[5] 算法采用一种硬机制，将与当前候选目标的交并比（intersection over union, IoU）值高于预定义阈值的目标过滤掉，只将分类得分最高的预测结果保留并输出，得分较低的预测结果即使正确也会被删除，从而导致目标重复检测或漏检现象的发生。分析当前研究文献可以看出，这些问题的处理主要从以下 2 个方面考虑。

在特征增强方面，通过特征金字塔网络[6]（feature pyramid networks, FPN）实现自深向浅的多尺度特征融合，丰富了浅层特征的语义信息。路径聚合网络（path aggregation networks, PANet）[7] 在 FPN[8] 的基础上增加了浅层向深层的特征融合路径，在丰富浅层语义特征的同时增加深层的空间位置特征。文献 [6] 通过分布于每层特征金字塔对象的统计特性估计融合因子，用来控制不同层特征的融合权重。这些方法均从网络纵向考虑，通过不同层特征融合增强网络的特征提取能力。文献 [9] 的 Inception 结构通过并联多个不同大小的卷积获取目标的多尺度信息，增强空间定位准确性。文献 [10] 在特征提取网络中引入空洞卷积在增大感受野的同时，又将增强后的特征与深层特征进行融合，提升目标的检测精度。文献 [11] 通过将可分离卷积和空洞卷积结合使用，在有效获取多尺度上下文的基础上减少模型参数量。这些方法从网络横向考虑，通过扩大感受野获得丰富的目标上下文信息。文献 [12] 通过增加自下而上的路径，减少浅层特征传递到顶层时的信息损失，提出双向多尺度特征融合网络。文献 [13] 采用 SE-ResNeXt 作为特征提取网络，通过多分路卷积和残差连接提高特征提取的效率。

在推理阶段，NMS 也对模型性能有重大影响，如 Soft-NMS[14] 提出一类“软机制”方法，在处理重叠度超过 NMS 阈值的预测框时，采用加权方式降低预测框得分而不是直接删除，在一定程度上可以降低目标被误删的概率。动态阈值非极大值抑制（dynamic threshold non-maximum suppression, DT-NMS）[15] 通过双阈值平衡了单阈值算法对目标漏检和误检问题的关系。IoU-Guided NMS[16] 通过使用 IoU 得分取代分类得分进行 NMS 处理。文献 [17] 将 NMS 过程加入训练阶段，使用平均精度作为训练损失，真正实现端到端训练 CNN 检测器。这些针对传统 NMS 改进的算法，都取得了一定的检测性能提升，但是它们依然采用先验固定阈值。如果该阈值设置过高，会引入更多假正例；设置过低，会删除真正例，造成目标漏检，因此，选取合适的 NMS 阈值对提升检测性能起着至关重要的作用。

总之，上述算法虽然在一定程度解决了目标漏检和重复检测问题，提升了目标检测精度，但是仍不能满足精确检测和鲁棒检测的要求。基于此，本文主要创新点如下：

1）提出一种注意力引导的多尺度上下文模块 (attention-guided multi-scale context module, AMCM)，增强网络对检测目标的特征表达能力。首先，通过并联多组不同空洞率的空洞卷积，获取目标不同尺度的上下文信息，更好地利用目标的空间位置信息，然后，通过通道注意力调整多尺度上下文特征融合比例，同时获取目标的跨通道信息，使特征更具鲁棒性，达到增强目标特征的目的。

2）针对推理过程，NMS 中固定阈值问题，提出一种基于目标密度的自适应阈值 NMS 方法 (adaptive density threshold of NMS, ADT-NMS)，依据每幅图像中存在的目标个数与推理过程中阈值的关系，使网络自适应地学习 NMS 阈值。合理的 NMS 应该给重叠度高的目标设置较高的过滤阈值，对重叠度低的目标分配较低阈值。本文设计了一个目标密度预测网络，将其嵌入到检测器头部形成新的分支，将该分支的输出结果作为 NMS 的阈值，使其能够随着目标密度自适应变化，消除了当前 NMS 方法针对所有目标手工设定单一阈值的局限性。

3）在 PASCAL VOC 和 KITTI 数据集上进行大量实验，本文算法的平均检测精度（mean average precision, mAP）和召回率（Recall）都有较明显的提升，误检率有明显下降，证明了本文算法的有效性。

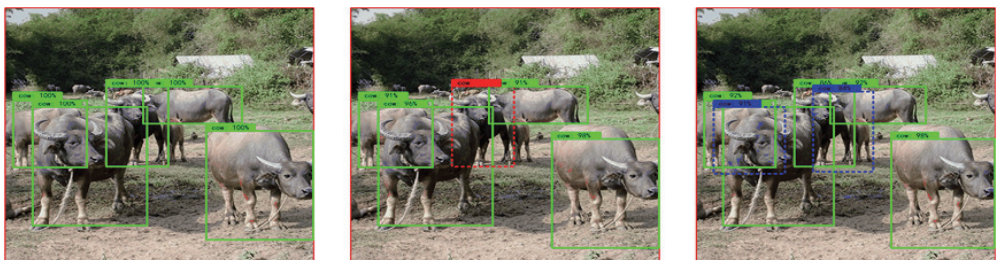
1 基准算法 YOLOv4

本文算法将 YOLOv4[18] 作为基准。YOLOv4[18] 是当前具有代表性的一阶段目标检测算法，采用 CSPDarknet53 作为骨干网络。利用该骨干网络提取目标特征，在性能上优于 ResNet-101[19] 和 ResNet-152[19]，同时因为参数量少运算速度快，可以在一块 GPU 上运行，增强了算法的通用性和可移植性。YOLOv4[18] 通过空间金字塔池化（spatial pyramid pooling, SPP[20]）模块增强骨干网络提取的最深层的特征，并通过 PANet[7] 双向特征融合模块将增强后的特征传递到各个预测层。在后处理阶段，YOLOv4[18] 采用 DIoU-NMS[21]，在考虑预测框 IoU 的同时考虑目标中心点的距离影响，处理方式如下：

$$S_{\mathrm{f}}=\begin{cases}S_{\mathrm{j}} & D_{\mathrm{DIoU}}(M,b_{\mathrm{j}})<N_{\mathrm{t}}\\ 0 & D_{\mathrm{DIoU}}(M,b_{\mathrm{j}})\geqslant N_{\mathrm{t}}\end{cases} \tag{1}$$

式中: N_t为抑制阈值; b_j为候选预测框; M 为当前分类得分最高的预测框; $D_{DIoU}(M,b_j)$为 M 和b_j的距离交并比 (Distance-IoU, DIoU) 值; S_f为经 NMS 处理过的最终预测框得分。

对于 DIoU-NMS[21] 策略, 如图 1 所示的目标密集又重叠场景下, 如果设置较低的阈值, 可能会导致目标漏检。在分组处理中, 能够抑制掉所有与 M 的 IoU 值大于阈值的预测框, 即使它们是对附近不同 GT 的正确预测, 如图 1(b) 所示, 设置的阈值为 0.3, 红色虚线框代表漏检的目标。设置较高的阈值又会引入更多的错误预测, 如图 1 (c) 所示, 设置的阈值为 0.7, 蓝色虚线框代表引入的假正例。因此, 如何设定 NMS 阈值对于提升目标检测算法性能至关重要。



(a) 真值　　(b) N_t=0.3　　(c) N_t=0.7

图 1　DIoU-NMS 在不同阈值下的检测结果

Fig. 1　Detection results of DIoU-NMS under different thresholds

2　本文算法

2.1　本文算法整体框架

本文算法整体框架由特征提取网络、特征增强模块、检测头和后处理模块组成, 如图 2 所示。首先, 利用 CSPDarknet53 网络提取特征, 得到最后 3 层的特征图 F_4、F_5、F_6; 然后, 将其输入 AMCM, 丰富骨干网络输出的特征多样性, 再将增强后的特征 C_1、C_2、C_3 输入路径聚合网络 PANet。AMCM 和 PANet 组成特征增强模块; 接着, 通过嵌入目标密度预测网络的检测头结构, 同时得到预测目标的坐标、分类得分、置信度及后处理 NMS 的处理阈值; 最后, 通过 ADT-NMS, 根据每幅图像的目标密度自适应地调整抑制阈值, 有效增强 NMS 的普适性。

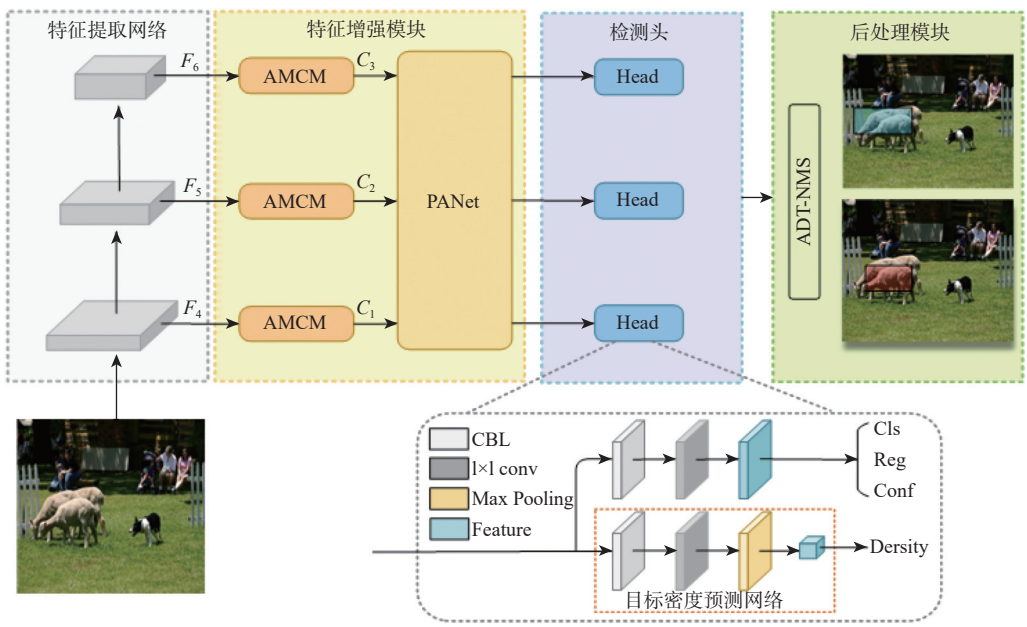


图 2　本文算法整体框架

Fig. 2　Overall framework of the proposed algorithm

本文将目标密度预测作为一个回归任务, 通过添加一个平行的检测头, 将其与目标的位置、分类一样当成同等任务处理。目标密度预测网络如图 2 所示, 其由 3 部分组成, 第 1 部分为标准卷积模块, 由一个 3×3 的卷积层、一个批量归一化层和一个 Leaky ReLU 激活函数层组成; 第 2 部分是一个 1×1 的卷积层; 第 3 部分是一个最大池化层。该网络通过第 1 部分的升维卷积丰富目标的信息, 使

预测的目标密度更准确，第2部分的降维卷积达到减少参数量的效果，第3部分用来获得当前特征图上的目标密度。

2.2 注意力引导的多尺度上下文特征增强模块

在检测场景中，目标往往不是孤立存在的，其周围总存在着各类对象，同时这些对象能够对目标本身产生影响。这些和目标相关或无关的信息就是目标的上下文信息，文献[22]指出这类信息往往对正确预测目标起着至关重要的作用。例如，如果目标背景信息是斑马线，则预测目标分类是行人的可能性比较大；如果目标背景信息是天空，则预测目标是飞机的可能性比较大。通过上下文信息，可以利用目标周围的背景信息增强目标的置信度。

受此启发，本文设计了AMCM结构，增强网络对目标多尺度特征的提取能力，包含多尺度上下文模块和通道注意力2部分，如图3所示。首先，通过空洞空间金字塔池化(atrous spatial pyramid pooling, ASPP)[23]的并行结构将不同空洞率的卷积层获取的特征图在通道维度进行拼接；接着，将它和经过通道注意力增强的特征进行拼接，进一步增强特征的有效信息，抑制冗余信息；最后，通过1×1的卷积进行通道降维。

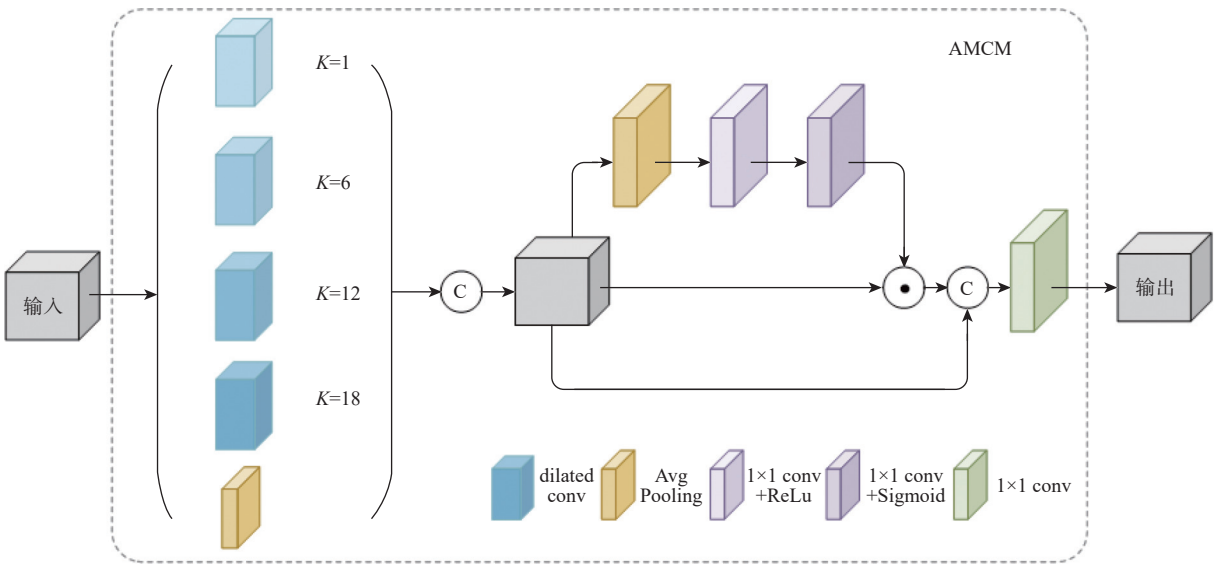


图3 注意力引导的多尺度上下文模块

Fig. 3 Attention-guided multi-scale context module

多尺度上下文模块由4个空洞卷积和一个全局平均池化5部分并联组成，其中，4个空洞卷积的卷积核r大小为3，空洞率K取值分别为1、6、12和18。4个空洞卷积的感受野大小通过式$R_f=(r-1)\cdot(K-1)+r$计算得到，分别为3、13、25、37。采用并联方式，相比级联处理速度更快，在扩大感受野的同时，能够保留更多的多尺度目标特征信息。多尺度上下文模块从空间域出发，学习目标在空间维度上的尺度关联性，但是感受野增大会引入较多的图像背景噪声，可能降低检测准确率。为消除冗余信息对检测网络的负面作用，引入注意力机制的压缩激励网络(squeeze-and-excitation networks, SENet)[24]，通过对各个通道特征之间的重要性建模，赋予不同通道相应的权值，衰减作用较小甚至无用的通道特征，增强强相关的通道特征。首先，通道注意力通过全局平均池化(Avg Pooling)操作，得到特征图上每个通道的整体均值；然后，通过1×1卷积学习通道之间的相关性，并利用激活函数将关联关系映射到非线性空间，完成对所有特征通道的影响力评分；最后，将通道权值和原始特征向量点乘完成通道加权。

2.3 基于目标密度的自适应阈值NMS方法

通过在PASCAL VOC2007测试集上，统计每幅图像中经NMS方法处理前的预测目标数量，找到每幅图像中目标密度到NMS阈值的一种映射关系。图4为每幅图像中经NMS方法处理前的预测目标数量，可以看出，预测目标数N主要集中于区间[500,200]内。将N通过$N^{\beta}-1$归一化处理，映射到区间[0,1]中，使它和NMS阈值在相同的分布区间。基于此，本文设计了一种映射方式，根据目标密度自适应地生成NMS阈值：

$$N_t=0.5\left(\alpha^2+\left(N^{\beta}-1\right)^2\right)^{\frac{1}{2}} \tag{2}$$

式中：N为每幅图像中经NMS方法处理前的预测目标数量；β为N的衰减指数；α为阈值的平衡因子。

为获得超参数α、β的值，采用YOLOv4[18]原始

网络结构, 对预设的 $\alpha \in \{0.4, 0.45, 0.5, 0.55, 0.6\}$和 $\beta \in \{1/11, 1/12, 1/13, 1/14\}$进行联合测试。在超参选取过程中, 以算法的 mAP 作为判断条件。如果实验中存在多组数据平均精度相同, 再以 Recall 作为确定超参设置的准则。不同参数设置对比如图 5 和图 6 所示, 其中, α取 0.5, β取 1/12 时, 在 PASCAL VOC 数据集上 mAP 最高且 Recall 最高, mAP 达到 82.15%。

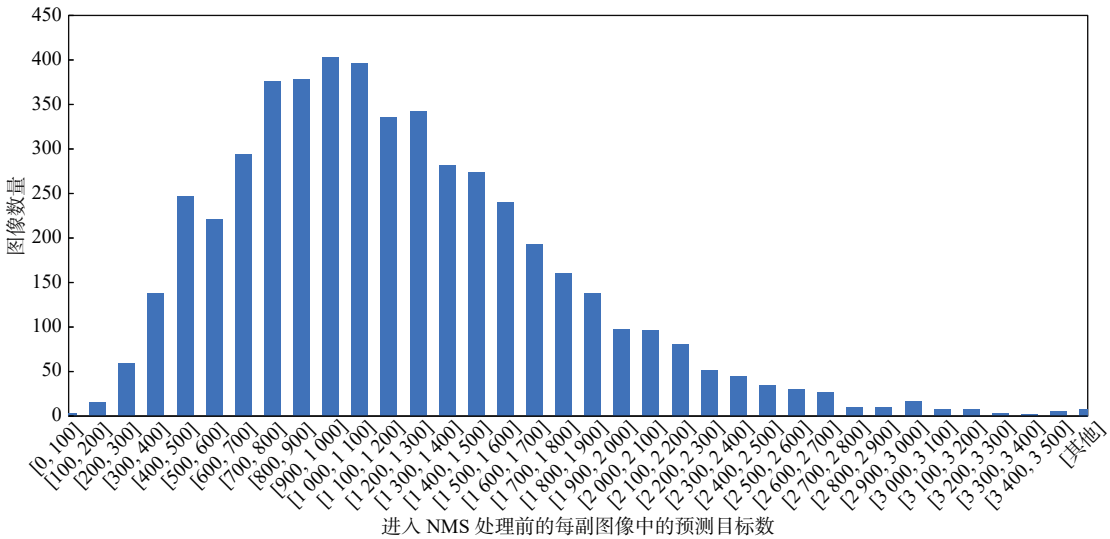


图 4　预测目标数量统计

Fig. 4　Statistical of predicted object quantity

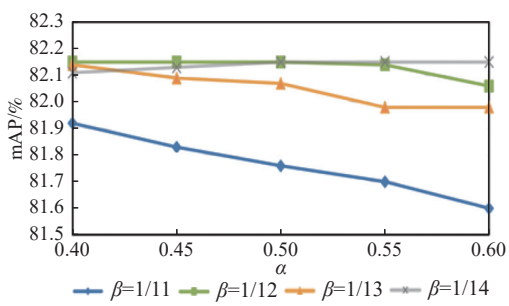


图 5　不同超参数设置下的检测精度

Fig. 5　Detection accuracy under different hyperparameter settings

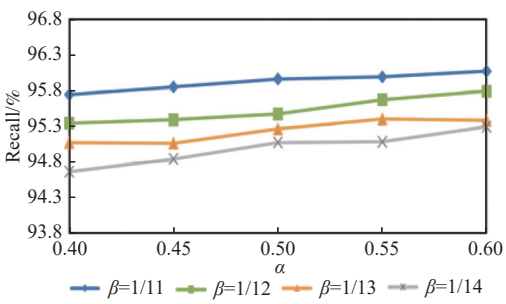


图 6　不同超参数设置下的召回率

Fig. 6　Recall under different hyperparameter settings

进一步将式(2)的自适应生成 NMS 阈值方法用到网络训练中, 增强 ADT-NMS 的鲁棒性。在训练过程中将 ADT-NMS 的抑制阈值 N_M定义为式(3), 并使用式(4)进行 NMS 抑制:

$$N_M = \max(d_S, d_M, d_L) \tag{3}$$

$$S_f = \begin{cases} S_i & D_{DIoU}(M, b_i) < N_M \\ 0 & D_{DIoU}(M, b_i) \geqslant N_M \end{cases} \tag{4}$$

式中: d_S、d_M、d_L分别表示图 2 的 3 个检测头中的目标密度预测分支得到的目标密度, 阈值 N_M取三者中的最大值。ADT-NMS 算法具有 3 个优势: ①摆脱了单一阈值的限制, 通过网络生成每幅图像对应的 NMS 阈值; ②将整体密度融合到个体, 如果整体的密度高, 意味着目标之间的聚集度高, 目标发生重叠的几率高, 自适应地提高 NMS 阈值, 可以将相邻高重叠的目标保留下来; ③仍然采用 Greedy-NMS 的流程, 通过动态阈值抑制密集目标的高重叠预测, 虽然增加少量的开销, 但更好地解决目标检测中的漏检和重复检测问题。

本文算法整体网络损失函数表达式为

$$l_{oss} = L_{CIoU} + L_{cls} + L_{conf} + \lambda |\max(d_S, d_M, d_L) - N_t| \tag{5}$$

式中: L_{CIoU}为目标的回归损失; L_{cls}为分类损失; L_{conf}为置信度损失; $\lambda|\max(d_S, d_M, d_L) - N_t|$为本文所提出的新型损失函数, 定义为目标密度回归损失; 将N_t作为监督信号用来监督式(3)通过目标密度预测网络生成的 NMS 阈值, 其中, λ设置为 50。

ADT-NMS 抑制方式伪代码如下:

输入: 检测框的初始化列表 $B = \{b_1, b_2, \cdots, b_N\}$, 对应的检测值 $S = \{s_1, s_2, \cdots, s_N\}$, 检测密度 $D = \{d_S, d_M, d_L\}$。

输出: $F = \{f_1, f_2, \cdots, f_N\}$, $S = \{s_1, s_2, \cdots, s_N\}$。

```
BEGIN
    F ← { }
    WHILE B ≠ empty DO
        m ← arg max S
        M ← b_m
        N_M ← max(D)
        F ← F ∪ M; B ← B − M
        FOR b_i in B DO
            IF iou(M, b_i) ⩾ N_M THEN
                B ← B − b_i; S ← S − s_i
            END
        END
    END
    RETURN F, S
END
```

3 实验方法与分析

3.1 数据集和评价指标

本文算法将PASCAL VOC2007和PASCALVOC 2012的训练集和验证集作为本实验的训练集，共包含16 551张训练图像，测试集采用PASCAL VOC 2007的测试集，共包含4 952张测试图像。为进一步证明算法的鲁棒性，使用KITTI数据集训练和测试，共包含5 984张训练图像和1 497张测试图像，该数据集是当前自动驾驶场景下最大的评测数据集。

在检测数据集上将mAP作为评价标准。同时，为说明本文算法在重叠目标检测上的优势，将Recall和对数平均漏检率(log-average miss rate, MR^{-2})也作为验证指标。MR^{-2}是行人检测中常用的指标，由于人群的聚集性特点，重叠现象在行人检测中常见，因此，通过该指标可以证明算法在重叠目标上的有效性。

mAP的定义如下：

$$m_{AP}=\sum_{k=0}^{N}P(k)\Delta r(k) \tag{6}$$

式中：N为图片的数量；$P(k)$为截止第k个图像的精度；$\Delta r(k)$为截止第k个图像和$k-1$个图像之间的Recall变化区间。

Recall的定义如下：

$$R_{ecall}=\frac{T_P}{T_P+F_N} \tag{7}$$

式中：T_P表示所有测试样本中，正确预测的正样本数量；F_N表示错误预测的负样本数量；T_P和F_N之和是整个测试集正样本目标总数量，该值越高说明漏检的目标越少。

MR和FPPI的计算式分别为

$$M_R=1-R_{ecall} \tag{8}$$

$$F_{PPI}=F_P/N \tag{9}$$

式中：F_P表示错误预测的正样本数量。通过计算均匀分布在区间$[10^{-2},10^{0}]$内，每张图像误报数-漏检率曲线(false positives per image-miss rate, FPPI-MR)的平均值得到MR^{-2}。

3.2 实验参数设置

以一阶段目标检测器YOLOv4[18]为基准，特征提取网络仍然使用CSPDarkNet53。将AMCM和ADT-NMS加入训练中，两者采用相同的初始化方式，其中，卷积层采用均值为0，方差为0.01的高斯分布初始化，偏差设置为0，BN层采用权重为1，偏置为0进行初始化。在训练中采用随机梯度下降(stochastic gradient descent, SGD)优化器，其中，动量和权重衰减分别设置为0.9和0.000 5，Batch size设置为4，初始学习率设置为10^{-5}，采用余弦退火的学习衰减策略，通过在一个GTX 1080Ti GPU上训练并验证，总训练轮数为200。

3.3 消融实验

3.3.1 AMCM模块有效性

1) PASCAL VOC数据集

如表1所示，第1行表示基准算法YOLOv4的结果，第2行表示在基准算法中引入ASPP模块，mAP提升了0.6%，Recall提升了0.5%，MR^{-2}下降了0.3%，表明通过空洞卷积扩大感受野，可以获取丰富的目标上下文信息提升了对目标的预测精度，第3行表示引入AMCM模块，mAP提升了1.2%，Recall提升了0.7%，MR^{-2}下降了0.6%，表明该模块能更好的重组目标的空间信息和通道信息，达到更好的检测效果。

表1 AMCM模块在PASCAL VOC上的有效性分析

Table 1 Effectiveness analysis of AMCM module on PASCAL VOC

基准算法	阈值	ASPP	AMCM	mAP/%	Recall/%	MR^{-2}/%
YOLOv4	$N_t=0.5$			82.0	95.7	14.7
		√		82.6	96.2	14.4
			√	83.2	96.4	14.1

2) KITTI 数据集

如表 2 所示, 第 2 行表示在基准上引入 ASPP 模块, mAP 提升了 1.1%, Recall 提升了 0.9%, MR^{-2} 降低了 0.6%, 第 3 行引入 AMCM 模块, mAP 提升了 1.4%, Recall 提升了 1.5%, MR^{-2} 下降了 1.1%。

在 2 个数据集上的实验结果表明: 通过 AMCM 模块能够提升对目标的上下文信息捕获能力, 来增强检测算法对目标的准确定位和分类预测能力, 有效的降低目标的漏检概率。AMCM 模块能稳定的提升目标检测效果, 具有较强的鲁棒性。

3.3.2　自适应阈值 NMS 的性能

1) PASCAL VOC 数据集

在 PASCAL VOC 数据集上, 如表 3 所示, 由表第 2 行和第 3 行可知, ADT-NMS 在 mAP 上相较 YOLOv4 算法有提升, 第 2 行表示引入未加入网络训练的 ADT-NMS, 采用式 (2) 自适应生成 NMS 阈值, 其中, 阈值平衡因子α和预测目标数的衰减指数β分别设置为 0.5 和 1/12, mAP 提升了 0.15%, Recall 不变, MR^{-2} 降低了 0.4%, 提高了检测结果。第 3 行表示将 ADT-NMS 加入网络训练, 采用式(3)自适应的生成 NMS 阈值, mAP 提升了 0.7%, Recall 虽然下降了 0.2%, 但是同时 MR^{-2} 下降了 0.3%, 说明在对预测样本处理时, 本文的 ADT-NMS 算法正确删除的假阳性样本比错误删除的正样本多, 总体性能依然是提升的。第 4 行表示引入 AMCM 模块和 ADT-NMS, mAP 提升了 1.7%, Recall 提升了 0.9%, MR^{-2} 降低了 1%。

表 2　AMCM 模块在 KITTI 上的有效性分析

Table 2　Effectiveness analysis of AMCM module on KITTI

基准算法	阈值	ASPP	AMCM	mAP/%	Recall/%	MR^{-2}/%
YOLOv4	$N_t = 0.5$			81.8	89.5	23.4
		√		82.9	90.4	22.8
			√	83.2	91.0	22.3

表 3　ADT-NMS 在 PASCAL VOC 上的有效性分析

Table 3　Effectiveness analysis of ADT-NMS on PASCAL VOC

基准算法	AMCM	ADT-NMS	阈值	mAP/%	Recall/%	MR^{-2}/%
YOLOv4			$N_t = 0.5$	82.00	95.7	14.7
		√	$N_t = 0.5[\alpha^2 + (N^\beta - 1)^2]^{1/2}$	82.15	95.7	14.3
		√	$N_t = \max(d_S, d_M, d_L)$	82.70	95.5	14.4
	√	√	$N_t = \max(d_S, d_M, d_L)$	83.70	96.6	13.7

2) KIITI 数据集

通过对 PASCAL VOC 的实验数据总结分析, 应用到 KITTI 数据集中的 ADT-NMS 动态阈值映射式(2)的阈值平衡因子α和衰减指数β的数值最终设置为 0.6 和 1/11。如表 4 所示, 由表第 2 行和第 3 行可知, 本文的 ADT-NMS 在检测精度上相较 YOLOv4 算法有提升, 第 2 行表示引入未参与神经网络训练的 ADT-NMS, mAP 提升了 0.5%, Recall 提升了 1.4%, MR^{-2} 降低了 0.3%, 第 3 行将 ADT-NMS 引入训练中, mAP 提升了 0.7%, Recall 提升了 1.7%, MR^{-2} 降低了 0.4%, 说明本文的 ADT-NMS 是一种有效的处理拥挤场景中重叠目标的方法。第 4 行表示同时引入 AMCM 模块和 ADT-NMS, mAP 提升了 1.8%, Recall 提升了 2.3%, MR^{-2} 降低了 1.3%, 结果表明, AMCM 模块和 ADT-NMS 两者相辅相成, 组成一种提升目标 mAP 更加行之有效的方法, 更能有效解决漏检问题。

表 4　ADT-NMS 在 KITTI 上的有效性分析

Table 4　Effectiveness analysis of ADT-NMS on KITTI

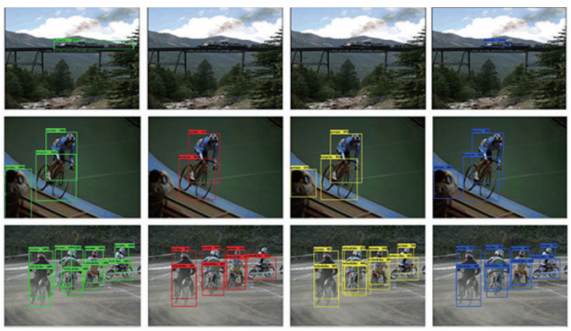
Baseline algorithm	AMCM	ADT-NMS	阈值	mAP/%	Recall/%	MR^{-2}/%
YOLOv4			$N_t = 0.5$	81.8	89.5	23.4
		√	$N_t = 0.5(\alpha^2 + (N^\beta - 1)^2)^{1/2}$	82.3	90.9	23.1
		√	$N_t = \max(d_S, d_M, d_L)$	82.5	91.2	23.0
	√	√	$N_t = \max(d_S, d_M, d_L)$	83.6	91.8	22.1

3.4 定性分析

1）PASCAL VOC数据集

通过将本文算法的检测结果与YOLOv4、Soft-NMS进行比较，选取具有代表性的检测结果进行可视化展示，说明算法的检测性能。

如图7所示，第1行，在背景和目标本身差异较小的情形下，仅本文算法正确检测到了火车，说明本文算法增强了网络对目标特征的提取能力，从而提高目标的定位准确度和分类置信度。第2行，Soft-NMS漏检了左下角的人，本文算法不仅正确检测到人，而且相比YOLOv4算法检测的结果置信度更高。第3行，Soft-NMS漏检了图中从左往右的第2个人，原算法又重复检测了这个人，只有本文算法正确检测了图像中的所有目标。



(a) GT (b) YOLOv4[18] (c) Soft-NMS[14] (d) 本文算法

图7 PASCAL VOC上的实验效果对比

Fig. 7 Comparison of experimental results on PASCAL VOC

实验结果表明，本文算法优于YOLOv4和Soft-NMS算法，主要表现在检测相同目标时，能给出更高置信度得分，而对重叠目标检测更精确。

2）KITTI数据集

在KITTI数据集的检测结果如图8所示。在左边的一组对比图中，YOLOv4算法存在误检现象，将草丛错误识别为车辆，而本文算法没有产生误识别现象，且对场景中出现的其他车辆识别准确。在右边一组对比图中，本文在有重叠的场景没有产生重复识别和漏检现象。

3.5 定量分析

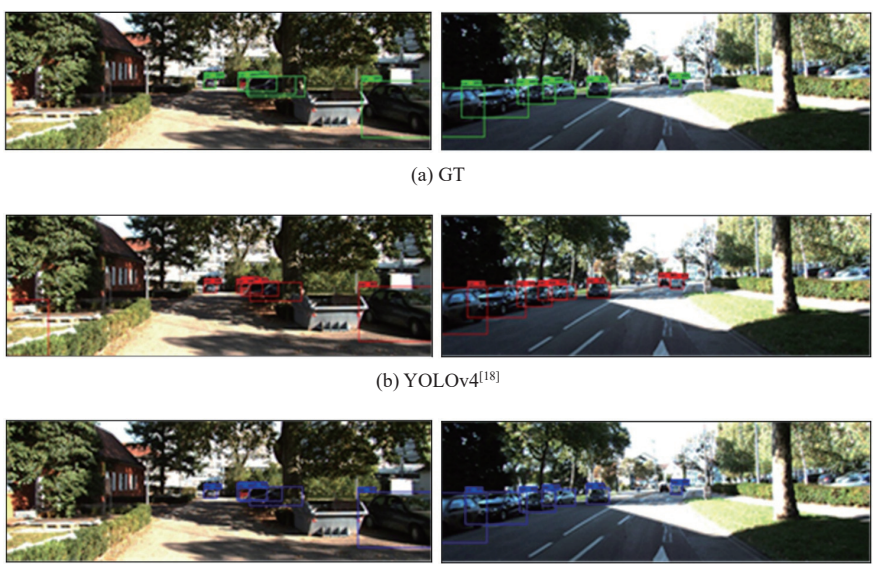
为进一步说明本文算法的检测性能，分别在PASCAL VOC2007数据集和KITTI数据集上对本文算法进行定量分析。

1）PASCAL VOC数据集

本文算法在当前主流算法中具有明显的优势，整体性能上，本文算法在检测精度和检测速度方面取得了很好的平衡，取得了83.7%的mAP。相比二阶段检测算法如Faster-RCNN[25]、R-FCN[26]、CoupleNet[27]，本文算法在速度和精度方面都有明显的提升，分别提升了10.5%、3.1%、1%。对比一阶段的目标检测算法DSSD513[28]、YOLOv3[29]、RefineDet-512[30]、FCOS[31]、CenterNet[32]、CenterNet-DHRNet[33]、YOLOv4[18]等，本文算法在检测精度方面表现最优[34]，如表5所示，最优结果表中粗体显示。

2）KITTI数据集

对比当前主流算法，本文算法在KITTI数据集上检测效果显著，最终取得了83.6%的mAP，相对于SSD[28]、CenterNet[32]、YOLOv4[18]的mAP分别提高了22.4%、13.1%、1.8%。在分类检测关于小车、行人、骑行的人的mAP相对于YOLOv4分别提升了2.0%、5.8%、6.2%，如表6和表7所示。



(a) GT

(b) YOLOv4[18]

(c) 本文算法

图8 KITTI上的实验效果对比

Fig. 8 Comparison of experimental results on KITTI

表 5　本文算法和其他算法在 PASAL VOC 上整体性能对比

Table 5　Comparison of the overall performance the proposed algorithm and other algorithms on PASAL VOC

算法	骨干网络	图像大小/像素	mAP/%	帧率/(帧·s^{-1})
Faster-RCNN [25]	VGG-16	1 000×600	73.2	7
R-FCN [26]	ResNet-101	1 000×600	80.5	9
CoupleNet [27]	ResNet-101	1 000×600	82.7	8.2
SSD [28]	VGG-16	512×512	76.8	19
DSSD513 [28]	ResNet-101	513×513	81.5	5.5
YOLOv3 [29]	DarkNet-53	544×544	79.3	26
RefineDet512 [30]	VGG-16	512×512	81.8	24.1
ExtremeNet [32]	Hourglass-104	512×512	79.3	3
FCOS [31]	ResNet-50	800×1 333	81.1	10
CenterNet [32]	ResNet-101	512×512	78.7	30
CenterNet-DHRNet [33]	DHRNet	512×512	82.3	27.6
YOLOv4 [18]	CSPDarkNet53	416×416	82.0	19.1
本文算法	CSPDarkNet53	**416×416**	**83.7**	**16**

表 6　本文算法和其他算法在 KITTI 上整体性能的对比

Table 6　Comparison of the overall performance the proposed algorithm and other algorithms on KITTI

算法	图像大小/像素	mAP/%	帧率/(帧·s^{-1})
SSD[28]	512×512	61.2	28.9
CenterNet[32]	512×512	70.5	36.0
YOLOv4[18]	416×416	81.8	18.3
本文算法	416×416	83.6	17.0

表 7　本文算法和其他算法在 KITTI 上分类性能对比

Table 7　Comparison of the classification performance the proposed algorithm and other algorithms on KITTI

算法	mAP/%		
	小车	行人	骑行的人
SSD[28]	85.1	48.0	50.6
CenterNet[32]	81.3	60.5	69.7
YOLOv4[18]	92.8	74.0	78.6
本文算法	94.8	79.8	84.8

4　结　论

1）针对目标检测中的目标漏检和重复检测问题，本文提出了一种基于特征增强和自适应阈值 NMS 的目标检测算法。

2）相对于基准算法 YOLOv4，本文算法主要对特征增强和后处理方法进行改进，在特征增强方面主要采用多尺度上下文特征，并结合多通道注意力机制，进行特征提取处理。在后处理方法上提出了根据目标密度的监测值动态生成 NMS 阈值，从而改善了算法的性能。

3）在 PASCAL VOC 数据集和 KITTI 数据集上进行了大量实验，证明了本文算法的有效性。

在实验过程中发现，本文算法对于小目标检测性能不高，在下一步工作中将尝试使用更好的特征增强策略或者正负样本分配进一步提高对小目标检测的性能。

参考文献（References）

[1] GIRSHICK R, DONAHUE J, DARRELL T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2014: 580-587.

[2] REDMON J, DIVVALA S, GIRSHICK R, et al. You only look once: unified, real-time object detection[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2016: 779-788.

[3] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[EB/OL]. (2017-06-12)[2023-02-01]. https://arxiv.org/abs/1706.03762.

[4] ZHU X Z, SU W J, LU L W, et al. Deformable DETR: deformable transformers for end-to-end object detection[EB/OL]. (2021-03-18)[2023-02-01]. https://arxiv.org/abs/2010.04159.

[5] NEUBECK A, VAN GOOL L. Efficient non-maximum suppression[C]//Proceedings of the 18th International Conference on Pattern Recognition. Piscataway: IEEE Press, 2006: 850-855.

[6] GONG Y Q, YU X H, DING Y, et al. Effective fusion factor in FPN for tiny object detection[C]//Proceedings of the IEEE Winter Conference on Applications of Computer Vision. Piscataway: IEEE Press, 2021: 1159-1167.

[7] LIU S, QI L, QIN H F, et al. Path aggregation network for instance segmentation[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2018: 8759-8768.

[8] LIN T Y, DOLLÁR P, GIRSHICK R, et al. Feature pyramid networks for object detection[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE

Press, 2017: 936-944.

[9] SZEGEDY C, VANHOUCKE V, IOFFE S, et al. Rethinking the inception architecture for computer vision[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2016: 2818-2826.

[10] 许腾, 唐贵进, 刘清萍, 等. 基于空洞卷积和 Focal Loss 的改进 YOLOv3 算法[J]. 南京邮电大学学报 (自然科学版), 2020, 40(6): 100-108.
XU T, TANG G J, LIU Q P, et al. Improved YOLOv3 based on dilated convolution and focal loss[J]. Journal of Nanjing University of Posts and Telecommunications (Natural Science Edition), 2020, 40(6): 100-108(in Chinese).

[11] 王囡, 侯志强, 蒲磊, 等. 空洞可分离卷积和注意力机制的实时语义分割[J]. 中国图象图形学报, 2022, 27(4): 1216-1225.
WANG N, HOU Z Q, PU L, et al. Real-time semantic segmentation analysis based on cavity separable convolution and attention mechanism[J]. Journal of Image and Graphics, 2022, 27(4): 1216-1225(in Chinese).

[12] 肖进胜, 张舒豪, 陈云华, 等. 双向特征融合与特征选择的遥感影像目标检测[J]. 电子学报, 2022, 50(2): 267-272.
XIAO J S, ZHANG S H, CHEN Y H, et al. Remote sensing image object detection based on bidirectional feature fusion and feature selection[J]. Acta Electronica Sinica, 2022, 50(2): 267-272 (in Chinese).

[13] 谢学立, 李传祥, 杨小冈, 等. 基于动态感受野的航拍图像目标检测算法[J]. 光学学报, 2020, 40(4): 0415001.
XIE X L, LI C X, YANG X G, et al. Dynamic receptive field-based object detection in aerial imaging[J]. Acta Optica Sinica, 2020, 40(4): 0415001(in Chinese).

[14] BODLA N, SINGH B, CHELLAPPA R, et al. Soft-NMS: improving object detection with one line of code[C]//Proceedings of the IEEE International Conference on Computer Vision. Piscataway: IEEE Press, 2017: 5562-5570.

[15] 侯志强, 刘晓义, 余旺盛, 等. 基于双阈值-非极大值抑制的 Faster R-CNN 改进算法[J]. 光电工程, 2019, 46(12): 190159.
HOU Z Q, LIU X Y, YU W S, et al. Improved algorithm of Faster R-CNN based on double threshold-non-maximum suppression[J]. Opto-Electronic Engineering, 2019, 46(12): 190159(in Chinese).

[16] JIANG B R, LUO R X, MAO J Y, et al. Acquisition of localization confidence for accurate object detection[M]// Computer Vision——ECCV 2018. Berlin: Springer, 2018: 816-832.

[17] HENDERSON P, FERRARI V. End-to-end training of object class detectors for mean average precision[M]//Computer Vision——ACCV 2016. Berlin: Springer, 2017: 198-213.

[18] BOCHKOVSKIY A, WANG C Y, LIAO H Y M. Yolov4: optimal speed and accuracy of object detection[EB/OL]. (2020-04-23)[2023-02-01]. https://arxiv.org/abs/2004.10934.

[19] HE K M, ZHANG X Y, REN S Q, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2016: 770-778.

[20] HE K M, ZHANG X Y, REN S Q, et al. Spatial pyramid pooling in deep convolutional networks for visual recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015, 37(9): 1904-1916.

[21] ZHENG Z H, WANG P, LIU W, et al. Distance-IoU loss: faster and better learning for bounding box regression[C]//Proceedings of the AAAI Conference on Artificial Intelligence. Palo Alto: AAAI Press, 2020, 34(7): 12993-13000.

[22] LIU W, LIAO S C, REN W Q, et al. High-level semantic feature detection: a new perspective for pedestrian detection[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2019: 5182-5191.

[23] CHEN L C, PAPANDREOU G, KOKKINOS I, et al. DeepLab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018, 40(4): 834-848.

[24] HU J, SHEN L, SUN G. Squeeze-and-excitation networks[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2018: 7132-7141.

[25] REN S Q, HE K M, GIRSHICK R, et al. Faster R-CNN: towards real-time object detection with region proposal networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(6): 1137-1149.

[26] DAI J F, LI Y, HE K M, et al. R-FCN: object detection via region-based fully convolutional networks[EB/OL]. (2016-05-20)[2023-02-01]. https://arxiv.org/abs/1605.06409.

[27] ZHU Y S, ZHAO C Y, WANG J Q, et al. CoupleNet: coupling global structure with local parts for object detection[C]//Proceedings of the IEEE International Conference on Computer Vision. Piscataway: IEEE Press, 2017: 4146-4154.

[28] LIU W, ANGUELOV D, ERHAN D, et al. SSD: single shot MultiBox detector[M]//Computer Vision——ECCV 2016. Berlin: Springer, 2016: 21-37.

[29] REDMON J, FARHADI A. YOLOv3: an incremental improvement[EB/OL].(2018-04-08)[2023-02-01].https://arxiv.org/abs/1804.02767.

[30] ZHANG S F, WEN L Y, BIAN X, et al. Single-shot refinement neural network for object detection[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2018: 4203-4212.

[31] TIAN Z, SHEN C H, CHEN H, et al. FCOS: fully convolutional one-stage object detection[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE Press, 2019: 9626-9635.

[32] DUAN K W, BAI S, XIE L X, et al. CenterNet: keypoint triplets for object detection[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE Press, 2019: 6568-6577.

[33] ZHOU X Y, WANG D Q, KRÄHENBÜHL P. Objects as points[EB/OL]. (2019-04-25)[2023-02-01]. https://arxiv.org/abs/1904.07850.

[34] 高扬, 安雯. 基于可变空间感知的目标检测算法[J]. 现代电子技术, 2023, 46(12): 91-95.
GAO Y, AN W. Object detection algorithm based on variable spatial perception[J]. Modern Electronics Technique, 2023, 46(12): 91-95 (in Chinese).

An object detection algorithm based on feature enhancement and adaptive threshold non-maximum suppression

MENG Weijun[1, *], AN Wen[1], MA Sugang[1, 2], YANG Xiaobao[1, 3]

(1. School of Computer Science and Technology, Xi'an University of Posts and Telecommunications, Xi'an 710121, China;
2. Shaanxi Key Laboratory of Network Data Analysis and Intelligent Processing, Xi'an 710121, China;
3. Xi'an Key Laboratory of Big Data and Intelligent Computing, Xi'an 710121, China)

Abstract: To further solve the problems of object omission and repeated detection and improve the accuracy of object detection, this paper proposes an object detection algorithm based on feature enhancement and adaptive threshold non-maximum suppression(NMS). The attention-guided multi-scale context module(AMCM) is applied to the neck of the detector. Based on improving the semantic information of features by dilated convolution, the cross-channel location information is captured by the attention mechanism, so as to enhance the feature expression ability of the network. The dynamic suppression threshold is adaptively applied to instances of the scenes through the adaptive density threshold of NMS(ADT-NMS), which lowers the false detection rate for objects. In comparison to the baseline algorithm YOLOv4, the suggested approach's false detection rate on the PASCAL VOC dataset is 13.7%, a 1% decrease. The recall rate and detection accuracy increase by 0.9% and 1.7%, respectively, to 96.6% and 83.7%. The false detection rate of the proposed algorithm on the KITTI dataset achieves 22.1%, reduced by 1.3%. The detection accuracy and recall rate achieved 83.6% and 91.8%, improved by 1.8% and 2.3%, respectively. The experimental results show that the algorithm can better solve the problems of object omission and repeated detection.

Keywords: adaptive threshold; non-maximum suppression; object detection; dilated convolution; attention mechanism

Received: 2023-08-24; **Accepted**: 2023-10-13; **Published Online**: 2023-11-14 13:34

URL: link.cnki.net/urlid/11.2625.V.20231114.0933.001

Foundation items: National Natural Science Foundation of China (62072370); Natural Science Foundation of Shaanxi Province (2023-JC-YB-598)

* **Corresponding author**. E-mail: mengweijun@xupt.edu.cn

http://bhxb.buaa.edu.cn jbuaa@buaa.edu.cn
DOI: 10.13700/j.bh.1001-5965.2023.0433

基于信噪比的学习型哈里斯鹰优化算法

张林[1]，沈佳颖[2]，胡传陆[1]，朱东林[2,*]
(1. 江西理工大学 软件工程学院，南昌 330000；2. 浙江师范大学 计算机科学与技术学院，金华 321004)

摘 要：针对哈里斯鹰优化（HHO）算法存在种群学习性与适应性不足的问题，提出一种基于信噪比的学习型哈里斯鹰优化（SLHHO）算法。该算法通过引入信噪比的概念来判断个体的位置信息，设计了一种协调学习策略，可以更合理地更新种群内个体的位置，进而对逃逸距离重新设计，提升了算法的适应与寻优能力。以12个基准函数为标准，将所提算法与哈里斯鹰算法的变体及其他算法进行性能测试，并在时间复杂度、多样性、探索与开发等评价指标中进行对比分析，结果显示，SLHHO算法具有较强的竞争力与可行性，在压力容器设计问题中，验证了SLHHO算法的实用性。

关 键 词：哈里斯鹰优化；信噪比；协调学习；逃逸距离；基准函数
中图分类号：TP301.6
文献标志码：A **文章编号**：1001-5965(2025)07-2360-14

随着军事、工业、医疗的快速发展，越来越多的优化问题随之产生，由于这类问题是黑盒问题，因此，对这些问题的探索还在深入当中。随着这些优化问题的复杂性增加，传统的数学方法如牛顿法和梯度下降法求解效率较低，因为它们耗时严重，仅取决于问题类型，并在局部状态中挣扎。近十年来，受自然界启发的元启发式算法被提出来解决这些问题，元启发式算法的特点是求解效率高和计算成本低，其模拟了自然界的生物或物理行为现象，较为经典与热门的算法有：粒子群优化 (particle swarm optimization, PSO) 算法[1]、蚁群优化 (ant colony optimization, ACO) 算法[2]、灰狼优化 (grey wolf optimizer, GWO) 算法[3]、鲸鱼优化算法 (whale optimization algorithm, WOA)[4] 等，这些算法均在工程设计、运筹学、能源等领域得到了广泛的应用。

近几年，元启发式算法的研究工作掀起了学者们的研究热潮，并提出了一系列元启发式算法，如麻雀搜索算法[5]、蝠鲼优化算法[6]、金枪鱼算法[7]、天牛须算法[8] 等，基于世界上没有免费午餐的定理，可知没有一个算法能最佳地解决所有的优化问题，那么如何使一个算法在某个或多个优化问题中具有一定的优势是学者们共同追求的目标。哈里斯鹰优化 (Harris hawks optimization, HHO) 算法[9] 是2019年提出的一种新型启发式算法，模拟哈里斯鹰捕食的手段，在一定空间内通过群体独特的协作方式来进行捕食。该算法具有优化能力强、可控参数少、易理解等优势，但在合作觅食中，存在内部交流不足的问题，在寻优精度与速度上还存在可改善的空间。

为进一步提升HHO算法优化问题的能力，学者们也在探索其改进的方法，Jiao等结合正交学习和反向学习，以有效和准确地估计太阳能电池和光伏组件的参数[10]；闫晓斌等提出自适应高斯变异来合理地控制种群的多样性[11]；Elgamal等在HHO算法的初始化阶段引入混沌映射，以提高搜索空间中的群体多样性，并用模拟退火算法来处理当前的最佳解决方案，以提高HHO算法的利用率[12]；胡春安等在HHO算法中引入混沌局部搜索策略，提高算

收稿日期：2023-07-03；录用日期：2023-11-27；网络出版时间：2024-02-04 10:56
网络出版地址：link.cnki.net/urlid/11.2625.V.20240202.1346.003
基金项目：国家自然科学基金 (62272418, 62002046)
*通信作者. E-mail：2078570674@qq.com

引用格式：张林，沈佳颖，胡传陆，等. 基于信噪比的学习型哈里斯鹰优化算法[J]. 北京航空航天大学学报，2025，51(7)：2360-2373.
ZHANG L，SHEN J Y，HU C L，et al. Learning Harris Hawks optimization algorithm with signal-to-noise ratio[J]. Journal of Beijing University of Aeronautics and Astronautics，2025，51(7)：2360-2373 (in Chinese).

法的开发能力, 精英备选池与分布估计来动态调整种群的寻优方向[13]; Li 等提出一种基于对数螺旋和基于反向学习的探索策略来提高 HHO 算法的探索能力[14]; 李岩等提出多种群的进化方向, 同时引入了协同量子化策略避免算法陷入局部极值[15]; 柴岩等引入双向经验与极端个体来引导个体寻优方向[16]; 刘小龙等设计了多子群邻域来引导子群觅食, 同时通过固定的置换概率来加强各子群的内部交流, 并通过算子选择更好地利用现有问题的信息[17]; 展广涵等采用了精英混沌反向学习策略初始化种群, 并引入动态自适应权重机制; 采用拉普拉斯交叉算子策略生成高质量个体[18]; 赵峰等引入指数能量方程和正弦跳跃距离方程, 对个体进行柯西变异或 Circle 混沌扰动[19]; 贾鹤鸣等在初始化阶段引入了动态反向学习, 寻优过程中融合了阿奎拉鹰优化 (Aquila optimizer, AO) 算法[20]。上述研究虽在一定程度上考虑了种群多样性及完善的寻优手段, 但因种群内部缺乏信息交流及对当前环境的判断能力不足, 导致算法不能在多类场景下均有较好的寻优效果。为此, 本文提出一种基于信噪比 (signal-noise ratio,SNR) 的学习型哈里斯鹰优化 (learning Harris hawk optimization with SNR, SLHHO) 算法, 引入了 SNR 的概念来判断种群的位置信息, 再设计一种协调学习机制来更新个体的位置, 并对逃逸距离重新进行调整, 提升了种群寻优的适应能力。主要创新点如下:

1) 本文提出一种新型的逃逸距离设计公式, 摆脱了 HHO 算法中逃逸距离 J 的搜索范围不全面的劣势, 使得设计后的值能够分布地更加均匀, 从而使寻优方式多样, 适应不同环境的寻优。

2) 通过引入 SNR 的概念来判断个体的位置信息, 设计了一种协调学习策略, 动态调整了个体的位置更新方式, 使得算法具有充分的学习能力。

3) 对 SLHHO 算法的优化性能进行测试, 通过 12 个变维的基准函数来验证 SLHHO 算法的有效性与可行性, 并对比其他算法与 HHO 算法的变体, 结合寻优过程中的多样性, 探索与开发指标进行分析, 结果显示 SLHHO 算法具有较强的寻优能力与优势。

4) 在压力容器设计问题中, SLHHO 算法优化后的解更加稳定, 验证了 SLHHO 算法的实用性。

1　相关理论

1.1　SNR 距离

在启发式优化算法的寻优过程中, 准确地判断当前个体的位置是否趋向最优解, 并做出合理的寻优方式是极其重要的, 这样才能减少个体走向不利解, 趋向合理的最优解, 提升算法的效率。以往的算法采用的方式是利用欧式距离来判断个体之间的位置信息, 但这种方式存在一定的局限性。

在机器学习领域, 距离的概念从诞生之日起就一直有着广泛的应用, 其通常用来度量数据间的相似性, 即确保相似的数据彼此接近, 而使不同的数据之间的映射距离相离。欧氏距离作为一种有效的度量工具被广泛的应用于诸多的度量学习算法中。然而, 随着维数的增大, 欧氏空间中任意 2 点距离度量的差异性越来越弱, 从而导致数据趋向于均匀分布, 这种现象称为“度量集中”, 出现在 Yu 等描述高维概率分布[21], 而且欧氏距离的计算过程在高维中需要耗费很长的时间, 因此, 判断数据相似性需要引入更合理的方法来替代欧氏距离, 例如, Weinberger 等使用马氏距离度量对 K-最近邻分类[22]; Yuan 等提出一种基于 SNR 的新型深度度量学习方法, 通过构建具有几何可解释性的 SNR 距离度量, 有效量化图像对之间的语义相似性[23]。由于基于 SNR 的距离度量可以产生比欧氏距离更多的鉴别特征。本文引入了 SNR 距离来衡量不同数据间的差异情况, 一对数据 p_i 和 p_j 的 SNR 距离定义为[24]

$$d_{\mathrm{S}}\left(p_i, p_j\right)=\frac{v\left(P_j-P_i\right)}{v\left(P_i\right)}=\frac{v\left(h_{ij}\right)}{v\left(P_i\right)} \tag{1}$$

式中: $v(z)=\left[\sum_{i=1}^{n}\left(z_i-\mu\right)^2\right]/n$ 为 z 的方差, μ 为 x 的均值, n 为 x 的维度。较长的 SNR 距离说明锚定数据和比较数据间的差异较大, 因此, 本文算法使用 SNR 距离度量作为相似性度量来代替欧氏距离度量, 通过个体间的数据差异来判断位置信息, 从而做出合理的寻优手段

1.2　HHO 算法简介

HHO 算法是 2019 年提出的一种新型群体智能优化算法, 该算法模拟哈里斯鹰的捕食行为, 主要分为探索阶段、探索与开发转换阶段和开发阶段。

HHO 算法可以从探索转移到开发, 然后, 根据猎物的逃逸能量在不同的开发行为之间进行转换。在猎物的逃避过程中, 能量将大大降低。猎物的逃逸能量 E 为

$$E=2E_0(1-t/T) \tag{2}$$

式中: T 为最大迭代次数; E_0 为能量的初始状态。在 HHO 算法的迭代过程中, E_0 在 (−1,1) 内均匀随机变化。HHO 算法通过 E 的值采用不同的行为策略。

1）探索阶段（$|E|\geqslant 1$）

HHO算法在初始化后再进行探索行为，具体模型为

$$x(t+1)=\begin{cases}x_{\text{rand}}(t)-r_1|x_{\text{rand}}(t)-2r_2x(t)| & q\geqslant 0.5\\ x_{\text{rabbit}}(t)-x_{\text{m}}(t)-r_3[L_{\text{B}}+r_4(U_{\text{B}}-L_{\text{B}})] & q<0.5\end{cases}\tag{3}$$

式中：$x(t+1)$为t+1次迭代中鹰的位置；$x_{\text{rabbit}}(t)$为最优猎物的位置，表示个体捕捉到最优食物的位置；$r_1\sim r_4$与q均为[0,1]的均匀随机数；L_{B}和U_{B}分别为变量的上限和下限；$x_{\text{rand}}(t)$为当前种群内部随机选择的鹰的位置；$x_{\text{m}}(t)$为当前种群的平均位置，具体形式为

$$x_{\text{m}}(t)=\frac{1}{N}\sum_{i=1}^{N}x_i(t)\tag{4}$$

2）开发阶段（$|E|<1$）

当鹰围攻猎物时，猎物会试图逃脱追捕，在这个阶段结合E与随机参数r提出4种不同的行为方式来模拟攻击阶段。

①软包围（$|E|\geqslant 0.5$，$r\geqslant 0.5$）

猎物此时具有较高的逃逸能量，需要对当前区域进行彻底地搜索，具体数学模型为

$$\begin{cases}x(t+1)=\Delta x(t)-E|Jx_{\text{rabbit}}(t)-x(t)|\\ \Delta x(t)=x_{\text{rabbit}}(t)-x(t)\\ J=2(1-r_5)\end{cases}\tag{5}$$

式中：r_5为[0,1]的均匀随机数，表示兔子在整个逃逸过程中的随机跳跃强度；J值每次迭代中在[0,2]随机变化，以模拟兔子运动的距离。

②强包围（$|E|<0.5$，$r\geqslant 0.5$）

兔子能量不足，鹰对兔子进行压迫式围攻，具体如下：

$$x(t+1)=x_{\text{rabbit}}(t)-E|\Delta x(t)|\tag{6}$$

③渐进式快速俯冲的软包围（$|E|\geqslant 0.5, r<0.5$）

猎物有机会从包围圈中逃脱，且逃逸能量足够，因此，哈里斯鹰需要在进攻前形成一个更加快速智能的软包围圈，通过以下2个策略实施：

$$\begin{cases}Y=x_{\text{rabbit}}(t)-E|Jx_{\text{rabbit}}(t)-x(t)|\\ Z=Y+SL_{\text{F}}(D)\\ x(t+1)=\begin{cases}Y & F(Y)<F(x(t))\\ Z & F(Z)<F(x(t))\end{cases}\end{cases}\tag{7}$$

式中：L_{F}表示levy飞行；D为维度。

④渐进式快速俯冲的硬包围（$|E|<0.5$，$r<0.5$）

猎物虽有一定的能量进行逃逸，但逃逸能量不足，鹰要趁机采用强硬且灵活的方式围攻，采用以下策略进行狩猎：

$$\begin{cases}x(t+1)=\begin{cases}Y & F(Y)<F(x(t))\\ Z & F(Z)<F(x(t))\end{cases}\\ Y=x_{\text{rabbit}}(t)-E|Jx_{\text{rabbit}}(t)-x_{\text{m}}(t)|\\ Z=Y+SL_{\text{F}}(D)\end{cases}\tag{8}$$

从式(3)～式(8)可以看出，HHO算法通过4个阶段进行寻优，其中，探索与开发能力的分配是通过随机数值进行处理，但会造成种群适应性较低[25]。种群适应性是面对不同的寻优环境时都能够较好地找到优质解，由式(2)～式(8)可知，HHO算法的策略仅通过随机数值来选择，缺乏进化性与学习性，同时逃逸能量是个体寻优范围的关键所在，其变化更是缺乏多样性，仅是通过自适应函数来控制，在面对简单与复杂函数寻优时，并不能使个体有较好的寻优视野，因此，需要对种群的寻优方式进行合理的设计，是研究的关键所在。

2 本文SLHHO算法

2.1 参数控制

智能优化算法中，合理的参数控制能够使得算法更具全局性，提升解的精度，在HHO算法中，J为猎物逃跑的距离，控制鹰搜索食物范围的能力，用来控制算法的搜索能力，为展现HHO算法中J数值的变化情况，图1为迭代1 000次时J的变化。

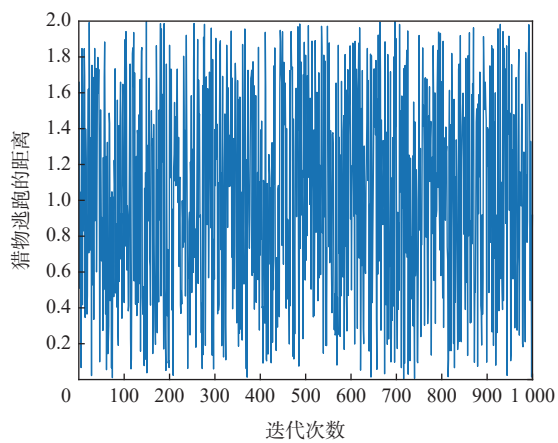


图1 HHO算法中迭代变化的J

Fig. 1 Iterative change of J in HHO algorithm

由图1可知，在1 000次的迭代里，J的变化区间基本一致，但数值没有更好地遍布整个范围，在0.2以下的数值非常少，同时1.8以上的数值也不多，这样的变化方式不能具有较好的适应性，尤其是在简单函数中，并不能开发有利的精度问题，而带来的无效性搜索会增加。因此，J需要一种更加均匀分布的值，才能应对不同的寻优环境。本文设

计了一种新型的 J 值变化方式, 用来适应不同的寻优环境, 具体如下:

$$J = 2(1-\cos(\rho\pi/2)) \tag{9}$$

式中: ρ 为 [0,1] 的均匀随机数, 具体参数变化如图 2 所示。由图 2 可知, 新设计的 J 在变化过程中每个区间都能较好地覆盖, 特别是较小的数值上也能覆盖, 同时在 J 值的变化路线中并不是一致的, 这样有利于开发解的精度, 在算法能够找到理想的解时, 是需要一定小的 J 来提升局部搜索, 找到更优质的解; 在复杂优化问题中, 较大的数值 J 能够避开不利的局部最优解, 探索出合理的高质量解。

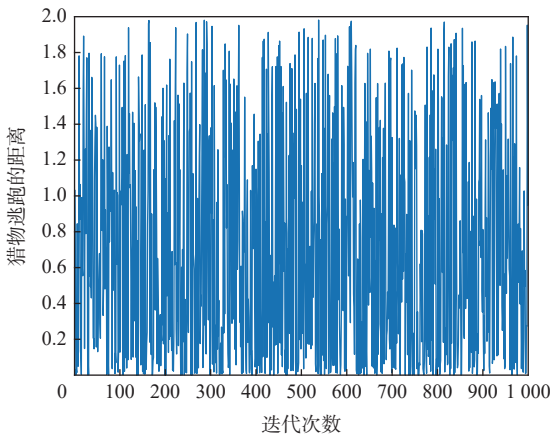


图 2　SLHHO 算法中迭代变化的 J

Fig. 2　Iterative change of J in SLHHO algorithm

2.2　协调学习

个体需要对种群外的解进行勘探, 同时还需要结合自己种群内的信息, 通过这些信息来更加可靠地找到优质解。在 HHO 算法中鹰的搜索及捕食方式仅是根据猎物来展开, 但自身种群内的个体缺乏信息交流, 仅跟最优个体、平均位置与随机位置进行一定的关联, 缺乏内部维度的交流, 同时缺乏智能化地更新, 不能对不同的寻优环境进行决策。为此, 本文设计了一种协调学习来提升 HHO 算法中个体之间的信息交流, 通过 SNR 距离来判断个体的状态, 来提升种群的学习性与适应性。具体如下:

$$\begin{cases} X_{i,j}(t) = \begin{cases} x_{\text{rabbit},l}(t) + b_1\left(x_{\text{worst},k}(t) - x_{\text{i},h}(t)\right) & d_1 \geqslant d_2 \\ x_{\text{rabbit},l}(t) + b_1\left(x_{\text{worst},k}(t) - x_{\text{M},h}(t)\right) & \text{其他} \end{cases} \\ d_1 = v(x_i(t) - x_{\text{worst}}(t))/v(x_{\text{worst}}(t)) \\ d_2 = v(x_i(t) - x_{\text{rabbit}}(t))/v(x_{\text{rabbit}}(t)) \\ w = \max P(f) \\ x_{\text{worst}}(t) = X_w(t) \end{cases} \tag{10}$$

式中: b_1 为 [0,1] 的均匀随机数; $x_{\text{M}}(t)$ 为减去最优位置与最差位置后得到的平均位置, 排除极端位置的影响, 这样就能更加明确地代表种群的大致方向; 以最小化问题为例, f 为种群个体计算后对应的适应度值数组; $\max P(f)$ 为提取最大值的对应下标位置; w 为找到的最大值的下标; $x_{\text{worst}}(t)$ 为 t 次迭代时的最差位置。对得到的 $x_{\text{M}}(t)$、$x_{\text{worst}}(t)$、$x_{\text{rabbit}}(t)$ 位置信息数据进行随机不重复的选择, l、k、h 为随机选取的维度, $x_{\text{worst},k}(t)$ 为在最差位置 $x_{\text{worst}}(t)$ 中随机选取的第 k 维的值, $x_{\text{rabbit},l}(t)$、$x_{\text{i},h}(t)$、$x_{\text{M},h}(t)$ 的含义以此类推。通过这些个体各维度上的信息随机进行交叉学习, 来提升种群的搜索效率。$d_1 \geqslant d_2$ 说明当前个体位置信息与最优位置相似, 需要将最优位置进行较远的搜索, 进行个体间信息交流, 最后, 更新出新的位置。反之则说明当前个体与最差位置相似, 则需要在最优位置进行较小的搜索, 需要将当前个体与平均位置保持一致, 放弃原来的位置且进行位置替换。2 种方式的紧密结合, 共同推进种群内部纵横交错的寻优路线, 并找到合理的最优位置。

为更好地理解个体之间的交流与变化, 假定维度为 3, 协调学习原理如图 3 所示。图 3 中黑色代表所随机选择的维度, 由 3 个不同的位置信息来共同引导产生新的个体。值得注意的是, 这类方案需要保证各维度的上下界均是一致, 否则变换后的无效性更新会增多。

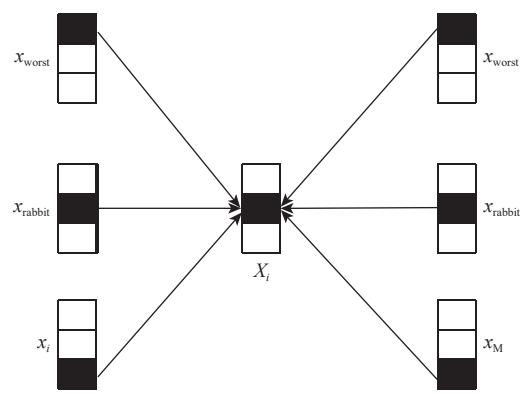


图 3　协调学习原理

Fig. 3　Coordinated learning schematic

2.3　本文算法的流程

为提升 HHO 算法的种群内部交流且提升算法的学习性与适应性, 本文算法设计了一种协调学习策略, 用来判断种群内部个体的位置特征, 便于做出合理的位置更新方式, 再设计了一种新的逃逸步长, 适应不同环境的寻优, 从而找到更优质的解。本文算法具体流程如下:

步骤 1　设定初始参数: 种群数量, 问题的上下界及维度, 最大迭代次数 (评估)。

步骤 2 随机初始化种群。

步骤 3 迭代开始，计算个体对应的适应度值。

步骤 4 采用协调学习进行更新位置信息，并找到当前最优猎物。

步骤 5 更新能量值 E 与逃逸距离 J。

步骤 6 若$|E|>1$，进行探索阶段则按式(3)～式(4)执行探索阶段的更新；若$|E|<1$，会随机生成判断值 r，哈里斯鹰需要根据能量值 E 和判断值 r 做出以下更新：①若$|E|\leqslant 0.5$ 且 $r\leqslant 0.5$，则按式(8)执行渐进式快速俯冲的硬包围；若$|E|\leqslant 0.5$ 且 $r>0.5$，则按式(6)执行强包围；②若$|E|>0.5$ 且 $r\leqslant 0.5$，则按式(7)执行进式快速俯冲的软包围；若$|E|>0.5$ 且 $r>0.5$，则按式(5)执行软包围。

步骤 7 判断是否到达结束条件，若是则输出最终的最优解及位置，否则返回步骤3。

2.4 时间复杂度分析

算法的时间复杂度分析是评价算法好坏的一个重要因素，在合理的时间内求取最好的效果是每个算法的目标。在原HHO算法中，寻优过程分为3个阶段，计算2次种群的适应度值，其时间复杂度为

$$O_{\mathrm{HHO}}=O(T(O(T_1+T_2+T_3)))=O(T(N_D+N_D))=O(TN_D) \tag{11}$$

SLHHO算法中 J 与HHO算法中一样，都需要计算2次数种群的适应度值，仅增加了协调学习这一过程，其时间复杂度为

$$O_{\mathrm{SLHHO}}=O(T(O(T_1+T_2+T_3+T_4)))=O(T(N_D+N_D+N_D))=O(TN_D) \tag{12}$$

由此可见，SLHHO算法并没有过分地增加时间复杂度，在统一评价次数的标准下HHO算法与SLHHO算法具有相同的时间复杂度，保持了一定的公平性。

3 性能测试

为检验SLHHO算法的优化能力，本文采用12个基准测试函数来验证，具体函数信息如表1所示。所有函数均为变维函数，f_1~f_7 为单峰函数，f_8~f_{12} 为多峰函数，2种类型的函数共同验证算法的全局优化能力。

表1 各函数信息

Table 1 Information for each function

函数	区间	最小值
$f_1(x)=\sum_{i=1}^{n}x_i^2$	[−100,100]	0
$f_2(x)=\sum_{i=1}^{n}\lvert x_i\rvert+\prod_{i=1}^{n}\lvert x_i\rvert$	[−10,10]	0
$f_3(x)=\sum_{i=1}^{n}\left(\sum_{j=1}^{i}x_j\right)^2$	[−100,100]	0
$f_4(x)=\max_i\{\lvert x_i\rvert,1\leqslant i\leqslant n\}$	[−100,100]	0
$f_5(x)=\sum_{i=1}^{n-1}[100(x_{i+1}-x_i^2)^2+(x_i-1)^2]$	[−30,30]	0
$f_6(x)=\sum_{i=1}^{n}(x_i+0.5)^2$	[−100,100]	0
$f_7(x)=\sum_{i=1}^{n}ix_i^4+\mathrm{random}[0,1)$	[−1.28,1.28]	0
$f_8(x)=\sum_{i=1}^{n}-x_i\sin\left(\sqrt{\lvert x_i\rvert}\right)$	[−500,500]	$-418.982\,9n$
$f_9(x)=\sum_{i=1}^{n}(x_i^2-10\cos(2\pi x_i)+10)$	[−5.12,5.12]	0
$f_{10}(x)=-20\exp\left(-0.2\sqrt{\frac{1}{n}\sum_{i=1}^{n}x_i^2}\right)-\exp\left(\frac{1}{n}\sum_{i=1}^{n}\cos(2\pi x_i)\right)+20+\mathrm{e}$	[−32,32]	0
$f_{11}(x)=\frac{1}{4\,000}\sum_{i=1}^{n}x^2-\prod_{i=1}^{n}\cos\left(\frac{x_i}{\sqrt{i}}\right)$	[−600,600]	0
$\begin{cases}f_{12}(x)=\frac{\pi}{n}\left[10\sin(\pi y_1)+\sum_{i=1}^{n-1}(y_i-1)^2\left(1+10\sin^2(\pi y_{i+1})\right)+(y_n-1)^2\right]+\sum_{i=1}^{n}u(x_i,10,100,4)\\ y_i=1+\frac{x_i+1}{4}\\ u(x_i,a,k,m)=\begin{cases}k(x_i-a)^m & x_i>a\\ 0 & -a<x_i<a\\ k(-x_i-a)^m & x_i<-a\end{cases}\end{cases}$	[−50,50]	0

3.1　与其他算法的变体进行比较

为证明 SLHHO 算法的新颖性与优势性, 将 SLHHO 算法与近几年提出的算法(见表 1)的变体进行比较, 这些算法有 ATOA[26]、DAOA[27]、EJAYA[28]、IGWO[29]、MFPA[30]、PSOGSA[31]、TACPSO[32], 这些算法都是近几年提出的算法, 并在基准函数中进行了验证, 具有一定的竞争性及代表性。各算法的内部参数如表 2 所示。每个算法的内部结构不一样, 为保证各算法实验的公平性, 每个算法的种群数量为 100, 最大评估次数为 50 000, 维度为 30, 其中, 最大评估次数是计算目标函数的次数。每个算法独立运行 30 次, 并计算 30 次数据的平均值 f_{age}, 最小值 f_{min}, 标准差 f_{std}, 用来衡量算法的寻优能力及稳定性。实验环境为 MATLAB 2019a, Windows10 操作系统, 运行内存为 8 GB; 另一方面, 为清晰地看出各算法之间地差异, 采用 Wilcoxon 检验对 30 次优化的结果统计测试, 各算法在显著水平为 0.05 下进行, 如表 3 所示, "+" 表示 SLHHO 算法优于其他算法, "=" 表示 SLHHO 算法与其他算法性能相当, "−" 则表示性能不如其他算法, f_{cm} 表示两算法的对比结果, 最后, 结合与实验指标来给出排名 f_{rank}。为直观展现 SLHHO 算法的优势, 本文实验数据均将 SLHHO 算法较优(最小)的指标进行加粗; 最后, 为清晰的看出各算法的综合寻优能力及普适性, 将表 3 中的排名进行汇总。

表 2　各算法需设置的内部参数

Table 2　Internal parameters to be set for each algorithm

算法	参数
ATOA[26]	f_{min}=0.2, f_{max}=1.0, α=5, μ=0.499, ε=2.220 4×10^{-16}
MFPA[30]	n=50, P=0.8, γ_1=1, γ_2=3, λ=1.5, cloning array = [9,8,7,6,5,4,3,2,1,1,1,1,1,1]
PSOGSA[31]	w=rand(1), c_1=0.5, c_2=1.5, G_0=1, α=20
TACPSO[32]	W_{max}=0.9, W_{min}=0.4, c_1=c_2=2

由表 3 可知, SLHHO 算法在所有函数上的指标均是较优的, 综合排名是较为靠前的, 验证了 SLHHO 算法具有较强的普适性。与其他算法的变体相比, SLHHO 算法具有较好的寻优能力, 同时在 f_1、f_8、f_9、f_{11} 均能直接找到理论最优值, 较其他算法具有明显的优势, 另一方面, HHO 算法在 f_8、f_9、f_{11} 也能直接找到理论最优值, 由此可见, SLHHO 算法并没有削弱 HHO 算法原本的寻优能力, 在其基础之上开发了更好的寻优手段, 彰显了 2 种策略的有效性, 使得种群寻优机制变得更加智能化。

收敛速度是元启发式算法追求的一大目标, 也是需要考量的一大因素。为彰显各算法在寻优速度上的差异, 本文将各算法在 30 次运行结果下的平均收敛轨迹进行展示, 具体如图 4 所示, 图中纵坐标为平均目标函数值, 横坐标为评估次数。

由图 4 可知, SLHHO 算法在各函数中的收敛速度均较快, 除 f_8 外, SLHHO 算法在其他函数上的收敛效果具有一定的优势。进一步来看, 在 f_1~f_7 单峰函数中, SLHHO 算法的收敛速度明显较快且精度得到有效的提高; 在 f_8~f_{12} 这些多峰函数中, SLHHO 算法的收敛速度存在一定的优势且精度较高, 具有一定的抗局部吸引能力, HHO 算法的收敛效果仅排在 SLHHO 算法之后, 其他算法在前期出现早熟现象。由此可知, 种群内的交流加快使得收敛速度得到一定的提升, 同时也证明了改进 HHO 算法具有一定的价值。

3.2　与 HHO 算法的变体对比

为展现 SLHHO 算法的竞争力与策略的有效性, 本节将 SLHHO 算法与 HHO 算法的变体进行比较, 同时考虑 2 种策略组合的合理性, 将其进行拆分进行试验, 设定调整 J 的 HHO 算法为 HHO-1, 设定加入协调学习的 HHO 算法为 HHO-2, 另外也将近几年提出的 HHO 算法变体加入对比实验中, 算法有 NCHHO[33]、LMHHO[34]、LHHO[35]、JOS-HHO[36]、THHO[19]、DAHHO[20]。测试维度分别为 30 与 100。其他实验参数设置与细则如上一致, 具体结果如表 4 和表 5 所示。考虑各算法在全部函数上的综合优化能力, 将各算法的平均值进行 Friedman 检验[37], 在显著水平 0.05 下进行, 综合排名 f_{rank} 结果如表 6 中所示。由表 4 和表 5 可知, SLHHO 算法在各函数中的较优指标数目依然是较多的, 仅在 30 维中的 f_3、f_4、f_{12} 与 100 维的 f_1、f_4、f_6、f_7 中, SLHHO 算法不是最优的数值; 在 30 维的 f_1、f_6、f_7、f_9、f_{10}、f_{11} 与 100 维的 f_5、f_9、f_{10}、f_{11}、f_{12} 中, 可看出 SLHHO 算法的寻优结果的优势较大, 从 Friedman 的综合排名来看, SLHHO 算法在 30 维与 100 维中分别为 2.916 7、3.166 7, 在与 HHO 算法的其他变体进行对比后, 排名依旧是最高的。另一方面, HHO-2 的综合排名相对靠前, 可以看出协调学习的有效性。由此可见, SLHHO 算法的综合寻优

表3 SLHHO算法与其他算法的变体

Table 3 SLHHO algorithm and other algorithm variants

算法	f_{min}											
	f_1	f_2	f_3	f_4	f_5	f_6	f_7	f_8	f_9	f_{10}	f_{11}	f_{12}
SLHHO	**0**	**8.60×10^{-231}**	**4.89×10^{-100}**	**1.25×10^{-59}**	**0**	**0**	**2.82×10^{-6}**	**-1.26×10^{4}**	**0**	**8.88×10^{-16}**	**0**	**3.58×10^{-19}**
ATOA[26]	5.59×10^{-12}	1.18×10^{-39}	4.12×10^{-3}	2.79×10^{-3}	$2.88\times10^{+1}$	5.56×10^{-1}	6.13×10^{-4}	-6.67×10^{3}	0	3.97×10^{-10}	0	2.53×10^{-1}
DAOA[27]	4.21×10^{4}	0	0	7.57×10^{1}	7.36×10^{3}	4.61×10^{4}	3.49×10^{-2}	-4.03×10^{3}	3.57×10^{1}	8.88×10^{-16}	9.19×10^{1}	1.85×10^{8}
EJAYA[28]	6.17×10^{-6}	1.52×10^{-3}	1.57	1.27×10^{-1}	2.31×10^{1}	1.37×10^{-5}	6.34×10^{-3}	-7.30×10^{3}	8.05	3.06×10^{-3}	1.48×10^{-5}	3.39×10^{-7}
IGWO[29]	6.14×10^{-34}	1.85×10^{-21}	3.52×10^{-5}	8.41×10^{-9}	2.51×10^{1}	3.93×10^{-2}	6.78×10^{-4}	-7.55×10^{3}	5.68	1.15×10^{-14}	0	1.12×10^{-3}
MFPA[30]	1.27	9.02×10^{-4}	3.84×10^{2}	8.28	2.20×10^{1}	2.69	7.71×10^{-2}	-1.08×10^{4}	2.15×10^{1}	2.57	8.57×10^{-1}	7.38×10^{-1}
PSOGSA[31]	4.88×10^{-19}	3.05×10^{-9}	1.08×10^{3}	1.85×10^{1}	1.87×10^{1}	4.84×10^{-19}	1.89×10^{-2}	-9.76×10^{3}	7.86×10^{1}	5.80×10^{-10}	7.94×10^{-3}	3.10×10^{-2}
TACPSO[32]	1.59×10^{-8}	8.43×10^{-5}	2.02	6.52×10^{-1}	2.78	2.33×10^{-9}	5.95×10^{-3}	-1.04×10^{4}	2.89×10^{1}	1.04×10^{-4}	4.90×10^{-9}	4.05×10^{-8}
HHO[9]	3.64×10^{-44}	3.02×10^{-19}	3.15×10^{-32}	2.08×10^{-20}	1.14×10^{-5}	4.97×10^{-9}	2.58×10^{-6}	-1.26×10^{4}	0	8.88×10^{-16}	0	1.65×10^{-8}

算法	f_{ave}											
	f_1	f_2	f_3	f_4	f_5	f_6	f_7	f_8	f_9	f_{10}	f_{11}	f_{12}
SLHHO	**0**	**8.85×10^{-58}**	**5.45×10^{-82}**	**1.95×10^{-50}**	**1.66×10^{-7}**	**6.05×10^{-12}**	**4.34×10^{-5}**	**-1.26×10^{4}**	**0**	**8.88×10^{-16}**	**0**	**1.48×10^{-8}**
ATOA[26]	7.56×10^{-4}	7.02×10^{-4}	4.00×10^{2}	3.47×10^{-2}	2.92×10^{1}	8.05×10^{-1}	5.62×10^{-3}	-5.34×10^{3}	0	9.30×10^{-3}	5.17×10^{-7}	9.47×10^{-1}
DAOA[27]	6.12×10^{4}	4.23×10^{1}	7.82×10^{4}	8.40×10^{1}	1.06×10^{4}	5.95×10^{4}	3.11×10^{1}	-2.75×10^{3}	1.13×10^{2}	1.08×10^{1}	5.13×10^{2}	4.69×10^{8}
EJAYA[28]	7.52×10^{-5}	6.34×10^{-3}	9.76	4.80×10^{-1}	$2.41\times10^{+1}$	9.62×10^{-5}	1.48×10^{-2}	-6.37×10^{3}	5.00×10^{1}	1.64	1.70×10^{-2}	1.03×10^{-1}
IGWO[29]	1.04×10^{-31}	8.36×10^{-21}	4.15×10^{-3}	4.58×10^{-7}	2.55×10^{1}	8.72×10^{-2}	2.14×10^{-3}	-5.88×10^{3}	6.45×10^{1}	1.76×10^{-14}	6.55×10^{-3}	3.80×10^{-3}
MFPA[30]	1.12×10^{1}	3.29×10^{-2}	1.63×10^{3}	1.56×10^{1}	2.99×10^{1}	6.86	1.54×10^{-1}	-8.20×10^{3}	5.66×10^{1}	1.50×10^{1}	1.07	4.24
PSOGSA[31]	2.02×10^{-1}	1.87×10^{-2}	6.94×10^{3}	3.56×10^{1}	2.44×10^{1}	3.37×10^{2}	5.89×10^{-2}	-8.32×10^{3}	1.23×10^{2}	1.02×10^{1}	9.21	4.50
TACPSO[32]	6.55×10^{-5}	4.87×10^{-3}	4.29×10^{1}	2.02	3.33×10^{1}	8.81×10^{-5}	1.88×10^{-2}	-9.27×10^{3}	5.82×10^{1}	7.01×10^{-1}	1.32×10^{-2}	2.07×10^{-1}
HHO[9]	1.44×10^{-27}	2.18×10^{-15}	3.00×10^{-17}	1.83×10^{-15}	8.10×10^{-4}	1.08×10^{-5}	1.12×10^{-4}	-1.26×10^{4}	0	2.79×10^{-14}	0	7.97×10^{-7}

算法	f_{std}											
	f_1	f_2	f_3	f_4	f_5	f_6	f_7	f_8	f_9	f_{10}	f_{11}	f_{12}
SLHHO	**0**	**4.85×10^{-57}**	**2.22×10^{-81}**	**9.12×10^{-50}**	**9.07×10^{-7}**	**2.63×10^{-11}**	**4.05×10^{-5}**	**2.48×10^{-2}**	**0**	**1.00×10^{-31}**	**0**	**7.42×10^{-8}**
ATOA[26]	9.91×10^{-4}	3.83×10^{-3}	$2.18\times10^{+3}$	1.83×10^{-2}	1.16	1.83×10^{-1}	4.12×10^{-3}	5.50×10^{2}	0	1.23×10^{-2}	2.83×10^{-6}	5.89×10^{-1}
DAOA[27]	6.62×10^{3}	1.35×10^{1}	3.41×10^{4}	3.79	1.37×10^{3}	5.42×10^{3}	2.55×10^{1}	4.73×10^{2}	3.60×10^{1}	9.62	1.03×10^{2}	1.08×10^{8}
EJAYA[28]	9.26×10^{-5}	4.82×10^{-3}	9.07	2.16×10^{-1}	5.34×10^{-1}	7.18×10^{-5}	4.96×10^{-3}	6.31×10^{2}	3.22×10^{1}	6.13×10^{-1}	3.75×10^{-2}	1.41×10^{-1}
IGWO[29]	1.84×10^{-31}	7.11×10^{-21}	9.53×10^{-3}	3.44×10^{-7}	1.97×10^{-1}	7.33×10^{-2}	1.22×10^{-3}	6.19×10^{2}	4.94×10^{1}	3.51×10^{-15}	1.30×10^{-2}	2.34×10^{-3}
MFPA[30]	1.93×10^{1}	8.68×10^{-2}	7.72×10^{2}	4.10	7.96	4.40	4.36×10^{-2}	1.12×10^{3}	1.81×10^{1}	6.80	7.77×10^{-2}	2.31
PSOGSA[31]	5.74×10^{-1}	5.64×10^{-2}	3.89×10^{3}	1.96×10^{1}	1.92	1.84×10^{3}	1.62×10^{-2}	7.08×10^{2}	2.83×10^{1}	6.78	2.75×10^{1}	3.25
TACPSO[32]	3.28×10^{-4}	4.94×10^{-3}	4.02×10^{1}	1.32	2.33×10^{1}	4.33×10^{-4}	7.82×10^{-3}	6.19×10^{2}	1.90×10^{1}	7.05×10^{-1}	1.44×10^{-2}	2.79×10^{-1}
HHO[9]	6.83×10^{-27}	4.05×10^{-15}	1.64×10^{-16}	4.48×10^{-15}	1.13×10^{-3}	1.82×10^{-5}	1.45×10^{-4}	1.75×10^{-1}	0	1.31×10^{-13}	0	8.28×10^{-7}

算法	f_{cm}											
	f_1	f_2	f_3	f_4	f_5	f_6	f_7	f_8	f_9	f_{10}	f_{11}	f_{12}
SLHHO												
ATOA[26]	+	+	+	+	+	+	+	+	=	+	+	+
DAOA[27]	+	+	+	+	+	+	+	+	+	+	+	+
EJAYA[28]	+	+	+	+	+	+	+	+	+	+	+	+
IGWO[29]	+	+	+	+	+	+	+	+	+	+	+	+
MFPA[30]	+	+	+	+	+	+	+	+	+	+	+	+
PSOGSA[31]	+	+	+	+	+	+	+	+	+	+	+	+
TACPSO[32]	+	+	+	+	+	+	+	+	+	+	+	+
HHO[9]	+	+	+	+	+	+	+	+	=	+	=	+

算法	f_{rank}											
	f_1	f_2	f_3	f_4	f_5	f_6	f_7	f_8	f_9	f_{10}	f_{11}	f_{12}
SLHHO	1	1	1	1	1	1	1	1	1	1	1	1
ATOA[26]	6	4	6	4	6	6	4	8	1	4	2	6
DAOA[27]	9	9	9	9	9	9	9	9	6	8	8	9
EJAYA[28]	5	6	4	5	3	4	5	6	2	6	5	4
IGWO[29]	2	2	3	3	5	5	3	7	5	2	3	3
MFPA[30]	8	8	7	7	7	7	8	5	3	9	6	7
PSOGSA[31]	7	7	8	8	4	8	7	4	7	7	7	8
TACPSO[32]	4	5	5	6	8	3	6	3	4	5	4	5
HHO[9]	3	3	2	2	2	2	2	2	1	3	1	2

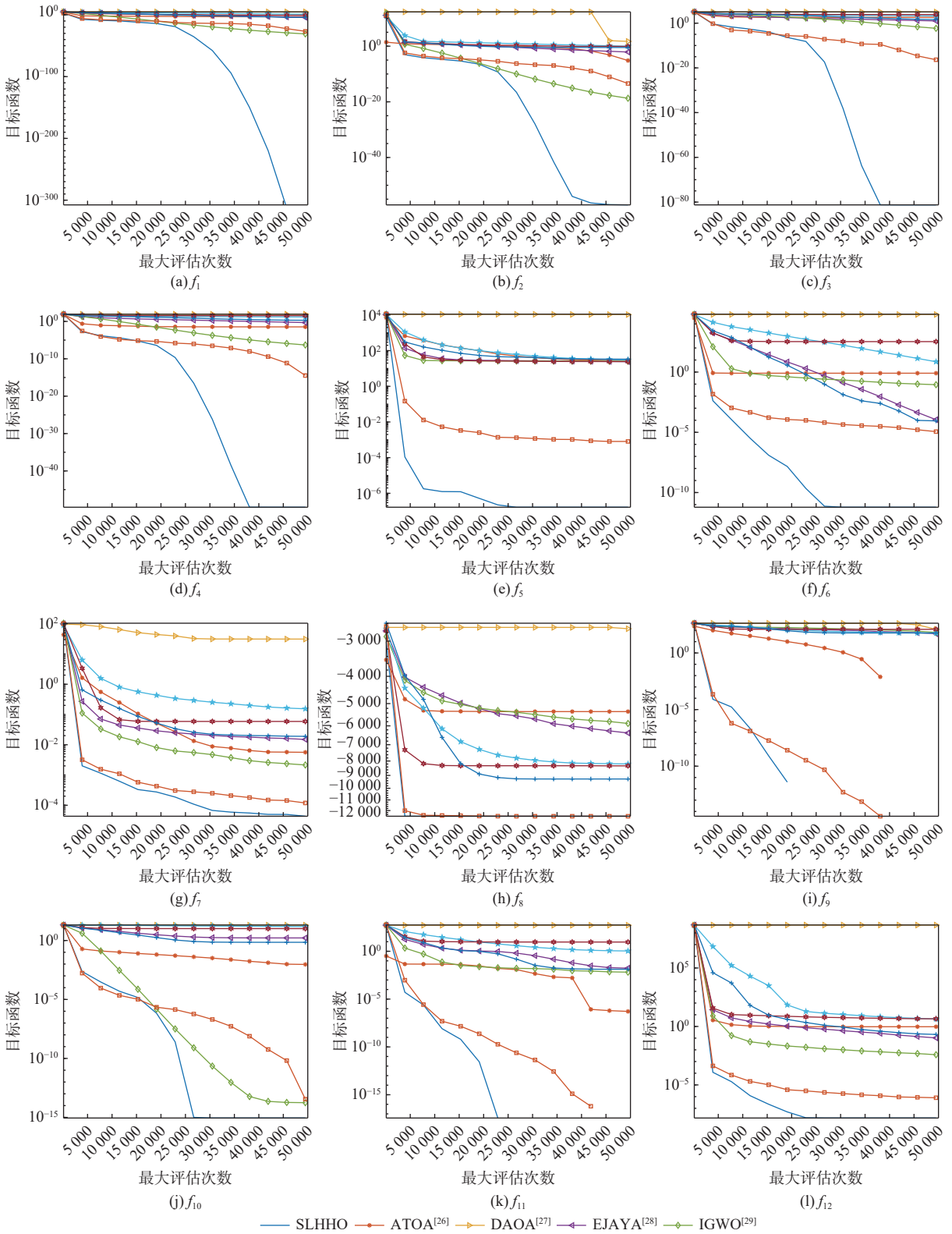


图 4　各算法平均收敛图

Fig. 4　Average convergence plot of each algorithm

能力是较强的, 进一步说明了 SLHHO 算法有力的竞争性与可行性。

3.3　勘探开发与多样性检验

算法的多样性是指在寻优过程中个体分布空间的离散程度, 离散程度较大时, 说明算法在一定空间内, 开拓解的能力也较强。探索与开发是指算法在全局与局部空间内挖掘有效解的能力, 是显示算法合理性的要素之一, 探索与开发表现地越出色

表4 SLHHO算法与HHO算法的变体（维度为30）

Table 4 SLHHO algorithm and HHO algorithm variants (dim=30)

算法	$f_{\min}$											
	f_1	f_2	f_3	f_4	f_5	f_6	f_7	f_8	f_9	f_{10}	f_{11}	f_{12}
SLHHO	0	0	4.89×10^{-100}	1.25×10^{-59}	0	0	2.82×10^{-6}	−8.60	0	8.88×10^{-16}	0	3.58×10^{-19}
NCHHO[33]	3.93×10^{-246}	3.93×10^{-246}	1.75×10^{-220}	1.73×10^{-115}	2.88×10^{1}	1.62	4.24×10^{-6}	−6.80	0	8.88×10^{-16}	0	4.56×10^{-2}
LMHHO[34]	3.07×10^{-44}	3.07×10^{-44}	3.01×10^{-44}	1.31×10^{-25}	7.35×10^{-9}	1.33×10^{-10}	5.00×10^{-7}	−8.60	0	8.88×10^{-16}	0	5.54×10^{-10}
LHHO[35]	3.14×10^{-75}	3.14×10^{-75}	3.09×10^{-70}	4.16×10^{-39}	2.27×10^{-6}	1.10×10^{-10}	1.23×10^{-5}	−8.60	0	8.88×10^{-16}	0	3.61×10^{-10}
THHO[10]	1.28×10^{-246}	1.28×10^{-246}	4.75×10^{-246}	2.62×10^{-127}	7.20×10^{-8}	2.95×10^{-13}	1.48×10^{-6}	−8.60	0	8.88×10^{-16}	0	9.20×10^{-12}
DAHHO[20]	2.03×10^{-60}	2.03×10^{-60}	1.30×10^{-50}	1.03×10^{-30}	2.54×10^{-6}	4.11×10^{-8}	2.37×10^{-6}	−8.60	0	8.88×10^{-16}	0	6.93×10^{-8}
JOS-HHO[36]	1.50×10^{-16}	1.50×10^{-16}	1.39×10^{-15}	1.11×10^{-9}	8.06×10^{-5}	5.24×10^{-5}	6.60×10^{-6}	−8.60	0	8.77×10^{-11}	0	9.23×10^{-7}
HHO-1	8.19×10^{-42}	8.19×10^{-42}	2.68×10^{-30}	3.25×10^{-18}	0	1.51×10^{-8}	4.06×10^{-5}	−8.60	0	8.88×10^{-16}	0	3.74×10^{-10}
HHO-2	7.82×10^{-123}	7.82×10^{-123}	4.38×10^{-104}	1.48×10^{-59}	5.19×10^{-8}	5.48×10^{-7}	2.88×10^{-6}	−8.60	0	8.88×10^{-16}	0	5.90×10^{-24}

算法	f_{ave}											
	f_1	f_2	f_3	f_4	f_5	f_6	f_7	f_8	f_9	f_{10}	f_{11}	f_{12}
SLHHO	0	8.85×10^{-58}	5.45×10^{-82}	1.95×10^{-50}	1.66×10^{-7}	6.05×10^{-12}	4.34×10^{-5}	−8.60	0	8.88×10^{-16}	0	1.48×10^{-8}
NCHHO[33]	3.61×10^{-222}	9.70×10^{-115}	3.26×10^{-202}	2.25×10^{-106}	2.88×10^{1}	2.36	6.59×10^{-5}	-6.16×10^{1}	0	8.88×10^{-16}	0	1.02×10^{-1}
LMHHO[34]	1.33×10^{-29}	6.45×10^{-16}	7.63×10^{-28}	3.78×10^{-16}	1.70×10^{-4}	5.98×10^{-7}	6.22×10^{-5}	−8.60	0	1.36×10^{-15}	0	2.11×10^{-7}
LHHO[35]	5.71×10^{-64}	4.29×10^{-32}	2.86×10^{-54}	8.66×10^{-33}	3.85×10^{-3}	2.21×10^{-6}	9.27×10^{-5}	−8.60	0	8.88×10^{-16}	0	1.03×10^{-6}
THHO[10]	3.67×10^{-234}	9.42×10^{-117}	3.74×10^{-226}	9.85×10^{-117}	1.46×10^{-3}	6.16×10^{-7}	7.59×10^{-5}	−8.60	0	8.88×10^{-16}	0	2.46×10^{-6}
DAHHO[20]	1.36×10^{-47}	2.04×10^{-26}	2.55×10^{-35}	4.94×10^{-26}	5.28×10^{-3}	4.96×10^{-5}	8.77×10^{-5}	−8.60	0	8.88×10^{-16}	0	3.76×10^{-6}
JOS-HHO[36]	6.37×10^{-10}	6.25×10^{-6}	2.59×10^{-7}	6.87×10^{-6}	1.51×10^{-2}	1.47×10^{-3}	3.38×10^{-4}	−8.60	4.15×10^{-10}	1.86×10^{-6}	3.34×10^{-9}	7.92×10^{-5}
HHO-1	5.49×10^{-24}	1.68×10^{-12}	3.18×10^{-12}	6.57×10^{-13}	1.53×10^{-21}	4.75×10^{-5}	2.45×10^{-4}	−8.60	0	1.30×10^{-12}	0	2.63×10^{-6}
HHO-2	9.46×10^{-102}	2.31×10^{-55}	3.35×10^{-89}	7.67×10^{-53}	1.84×10^{-4}	3.42×10^{-5}	5.91×10^{-5}	−8.60	0	8.88×10^{-16}	0	8.22×10^{-14}

算法	f_{std}											
	f_1	f_2	f_3	f_4	f_5	f_6	f_7	f_8	f_9	f_{10}	f_{11}	f_{12}
SLHHO	0	4.85×10^{-57}	2.22×10^{-81}	9.12×10^{-50}	9.07×10^{-7}	2.63×10^{-11}	4.05×10^{-5}	2.48×10^{-2}	0	1.00×10^{-31}	0	7.42×10^{-8}
NCHHO[33]	0	5.21×10^{-114}	0	1.18×10^{-105}	2.11×10^{-2}	4.51×10^{-1}	7.15×10^{-5}	$1.74\times10^{+3}$	0	1.00×10^{-31}	0	3.90×10^{-2}
LMHHO[34]	5.41×10^{-29}	1.59×10^{-15}	3.94×10^{-27}	1.78×10^{-15}	4.21×10^{-4}	9.84×10^{-7}	5.70×10^{-5}	3.49×10^{-2}	0	2.59×10^{-15}	0	4.59×10^{-7}
LHHO[35]	1.68×10^{-63}	1.71×10^{-31}	8.21×10^{-54}	2.98×10^{-32}	7.80×10^{-3}	3.19×10^{-6}	5.07×10^{-5}	1.82×10^{-2}	0	1.00×10^{-31}	0	1.77×10^{-6}
THHO[10]	0	5.10×10^{-116}	0	5.10×10^{-116}	5.77×10^{-3}	1.92×10^{-6}	8.68×10^{-5}	1.16	0	1.00×10^{-31}	0	7.00×10^{-6}
DAHHO[20]	7.25×10^{-47}	8.76×10^{-26}	1.09×10^{-34}	1.25×10^{-25}	5.54×10^{-3}	7.85×10^{-5}	1.04×10^{-4}	5.42×10^{-1}	0	1.00×10^{-31}	0	5.35×10^{-6}
JOS-HHO[36]	1.97×10^{-9}	8.74×10^{-6}	7.26×10^{-7}	1.02×10^{-5}	2.11×10^{-2}	1.77×10^{-3}	2.93×10^{-4}	1.35	8.53×10^{-10}	2.66×10^{-6}	1.45×10^{-8}	7.29×10^{-5}
HHO-1	2.82×10^{-23}	4.49×10^{-12}	1.22×10^{-11}	1.90×10^{-12}	8.40×10^{-21}	6.01×10^{-5}	2.08×10^{-4}	5.75×10^{-2}	0	5.98×10^{-12}	0	3.55×10^{-6}
HHO-2	5.18×10^{-101}	8.26×10^{-55}	1.83×10^{-88}	2.63×10^{-52}	1.96×10^{-4}	4.72×10^{-5}	5.16×10^{-5}	3.45×10^{-2}	0	1.00×10^{-31}	0	4.06×10^{-13}

表 5　SLHHO 算法与 HHO 算法的变体 (维度为 100)

Table 5　SLHHO algorithm and HHO algorithm variants (dim=100)

算法	f_{min}											
	f_1	f_2	f_3	f_4	f_5	f_6	f_7	f_8	f_9	f_{10}	f_{11}	f_{12}
SLHHO	1.10×10^{-112}	2.48×10^{-68}	2.05×10^{-105}	2.70×10^{-60}	0	5.56×10^{-8}	1.85×10^{-6}	−37.9	0	8.88×10^{-16}	0	1.54×10^{-20}
NCHHO[33]	1.25×10^{-239}	6.64×10^{-122}	3.12×10^{-214}	1.28×10^{-116}	9.82×10^{1}	5.87	2.51×10^{-6}	−36.6	0	8.88×10^{-16}	0	1.15×10^{-1}
LMHHO[34]	2.28×10^{-42}	5.70×10^{-23}	9.74×10^{-38}	2.61×10^{-23}	2.51×10^{-7}	5.25×10^{-11}	4.95×10^{-6}	−37.9	0	8.88×10^{-16}	0	2.49×10^{-12}
LHHO[35]	6.78×10^{-80}	9.25×10^{-38}	8.52×10^{-71}	6.23×10^{-39}	1.22×10^{-5}	3.99×10^{-8}	9.72×10^{-6}	−37.9	0	8.88×10^{-16}	0	7.51×10^{-11}
THHO[10]	7.71×10^{-254}	9.11×10^{-125}	2.86×10^{-248}	1.53×10^{-126}	1.27×10^{-7}	1.66×10^{-11}	3.42×10^{-7}	−37.9	0	8.88×10^{-16}	0	8.16×10^{-13}
DAHHO[20]	6.55×10^{-58}	2.58×10^{-32}	3.92×10^{-50}	3.62×10^{-32}	0	1.10×10^{-6}	1.41×10^{-5}	−37.9	0	8.88×10^{-16}	0	5.71×10^{-9}
JOS-HHO[36]	2.34×10^{-17}	2.57×10^{-8}	1.81×10^{-11}	2.17×10^{-12}	7.43×10^{-3}	2.72×10^{-4}	1.72×10^{-5}	−37.9	0	1.46×10^{-10}	1.89×10^{-15}	7.14×10^{-7}
HHO-1	9.94×10^{-21}	2.09×10^{-11}	2.41×10^{-20}	2.00×10^{-10}	2.11×10^{-5}	4.57×10^{-5}	3.40×10^{-6}	−37.9	0	2.30×10^{-11}	0	6.70×10^{-7}
HHO-2	9.26×10^{-77}	4.64×10^{-41}	5.63×10^{-63}	6.69×10^{-39}	7.85×10^{-5}	1.53×10^{-8}	1.24×10^{-7}	−37.9	0	8.88×10^{-16}	0	1.32×10^{-8}

算法	f_{ave}											
	f_1	f_2	f_3	f_4	f_5	f_6	f_7	f_8	f_9	f_{10}	f_{11}	f_{12}
SLHHO	5.33×10^{-99}	3.46×10^{-54}	1.73×10^{-77}	2.41×10^{-50}	0	2.01×10^{-4}	4.90×10^{-5}	−37.9	0	8.88×10^{-16}	0	2.89×10^{-9}
NCHHO[33]	5.23×10^{-220}	2.09×10^{-115}	2.42×10^{-198}	3.53×10^{-107}	9.83×10^{1}	7.95	8.19×10^{-5}	−21.2	0	8.88×10^{-16}	0	1.87×10^{-1}
LMHHO[34]	1.16×10^{-30}	2.17×10^{-14}	1.01×10^{-25}	3.41×10^{-16}	5.20×10^{-4}	1.74×10^{-6}	7.11×10^{-5}	−37.9	0	1.72×10^{-15}	0	1.86×10^{-8}
LHHO[35]	1.56×10^{-61}	3.36×10^{-31}	6.89×10^{-48}	5.71×10^{-32}	1.21×10^{-2}	1.97×10^{-5}	1.05×10^{-4}	−37.9	0	8.88×10^{-16}	0	8.80×10^{-7}
THHO[10]	5.65×10^{-229}	2.20×10^{-117}	2.12×10^{-228}	2.10×10^{-117}	2.26×10^{-3}	8.72×10^{-6}	4.88×10^{-5}	−37.9	0	8.88×10^{-16}	0	1.94×10^{-8}
DAHHO[20]	1.07×10^{-46}	1.42×10^{-26}	1.20×10^{-37}	7.80×10^{-26}	3.32×10^{-6}	2.26×10^{-3}	3.04×10^{-4}	−37.9	0	8.88×10^{-16}	0	9.55×10^{-6}
JOS-HHO[36]	4.38×10^{-9}	3.01×10^{-5}	6.64×10^{-5}	5.43×10^{-6}	1.76×10^{-1}	8.32×10^{-3}	2.71×10^{-4}	−37.9	7.18×10^{-9}	4.93×10^{-6}	5.14×10^{-9}	5.67×10^{-5}
HHO-1	2.98×10^{-10}	3.02×10^{-7}	3.06×10^{-8}	5.27×10^{-7}	5.05×10^{-2}	5.54×10^{-3}	1.22×10^{-4}	−37.9	3.16×10^{-12}	3.93×10^{-8}	2.18×10^{-12}	1.50×10^{-5}
HHO-2	5.80×10^{-62}	2.03×10^{-34}	5.49×10^{-47}	4.17×10^{-32}	1.74×10^{-3}	4.57×10^{-5}	3.27×10^{-5}	−37.9	0	8.88×10^{-16}	0	4.15×10^{-7}

算法	f_{std}											
	f_1	f_2	f_3	f_4	f_5	f_6	f_7	f_8	f_9	f_{10}	f_{11}	f_{12}
SLHHO	2.88×10^{-98}	1.84×10^{-53}	8.16×10^{-77}	7.89×10^{-50}	0	3.14×10^{-4}	3.69×10^{-5}	2.41×10^{-1}	0	1.00×10^{-31}	0	6.78×10^{-9}
NCHHO[33]	0	8.89×10^{-115}	0	1.77×10^{-106}	6.68×10^{-2}	1.95	7.59×10^{-5}	8.50×10^{3}	0	1.00×10^{-31}	0	6.79×10^{-2}
LMHHO[34]	4.72×10^{-30}	8.97×10^{-14}	5.52×10^{-25}	1.27×10^{-15}	1.25×10^{-3}	3.20×10^{-6}	6.49×10^{-5}	1.14	0	3.32×10^{-15}	0	3.66×10^{-8}
LHHO[35]	8.32×10^{-61}	1.75×10^{-30}	3.13×10^{-47}	2.76×10^{-31}	2.22×10^{-2}	3.56×10^{-5}	6.56×10^{-5}	3.48×10^{-2}	0	1.00×10^{-31}	0	1.28×10^{-6}
THHO[10]	0	1.08×10^{-116}	0	7.70×10^{-117}	8.19×10^{-3}	2.88×10^{-5}	4.07×10^{-5}	3.57	0	1.00×10^{-31}	0	5.66×10^{-8}
DAHHO[20]	5.66×10^{-46}	4.52×10^{-26}	4.68×10^{-37}	1.99×10^{-25}	1.82×10^{-5}	3.11×10^{-3}	3.33×10^{-4}	3.05	0	1.00×10^{-31}	0	2.15×10^{-5}
JOS-HHO[36]	1.07×10^{-8}	4.91×10^{-5}	2.04×10^{-4}	1.09×10^{-5}	2.07×10^{-1}	8.01×10^{-3}	2.48×10^{-4}	2.68×10^{1}	3.64×10^{-8}	1.20×10^{-5}	1.19×10^{-8}	6.68×10^{-5}
HHO-1	1.62×10^{-9}	6.43×10^{-7}	6.42×10^{-8}	1.47×10^{-6}	6.57×10^{-2}	4.61×10^{-3}	1.76×10^{-4}	7.88	1.14×10^{-11}	7.55×10^{-8}	6.06×10^{-12}	2.02×10^{-5}
HHO-2	3.15×10^{-61}	9.48×10^{-34}	2.23×10^{-46}	1.51×10^{-31}	2.60×10^{-3}	6.19×10^{-5}	2.90×10^{-5}	3.59×10^{-1}	0	0	0	5.59×10^{-7}

表 6　不同维度下 SLHHO 算法与 HHO 算法的变体排序值

Table 6　Variant sorting values of SLHHO algorithm and HHO algorithm in different dimensions

算法	f_{rank}(维度为30)	f_{rank}(维度为100)
SLHHO	2.916 7	3.166 7
NCHHO[33]	5.125 0	5.041 7
LMHHO[34]	4.958 3	4.791 7
LHHO[35]	4.833 3	4.583 3
THHO[10]	3.333 3	2.666 7
DAHHO[20]	5.666 7	5.250 0
JOS-HHO[36]	8.375 0	8.291 7
HHO-1	6.208 3	7.375 0
HHO-2	3.583 3	3.833 3

说明算法寻优能力越好。为更好地看出复杂函数下的寻优情况，本文选取了f_7与f_{12}的多样性与探索开发的表现进行展示，2种为不同类型的函数。将最大迭代次数设为500，种群维度为100，SLHHO算法与HHO算法的展示具体如图5和图6所示。

由图6可知，SLHHO算法在f_7与f_{12}探索与开发的过程中，前期搜索范围较广，勘探能力强，提升算法的全局搜索能力，中期具有一定的勘探及开发能力，平衡算法的全局与局部搜索能力，后期搜索范围逐渐变小，开发能力较强，勘探能力较差，提高解的精度；而HHO算法前期勘探能力较差，同时在寻优过程中，勘探与开发的平衡较

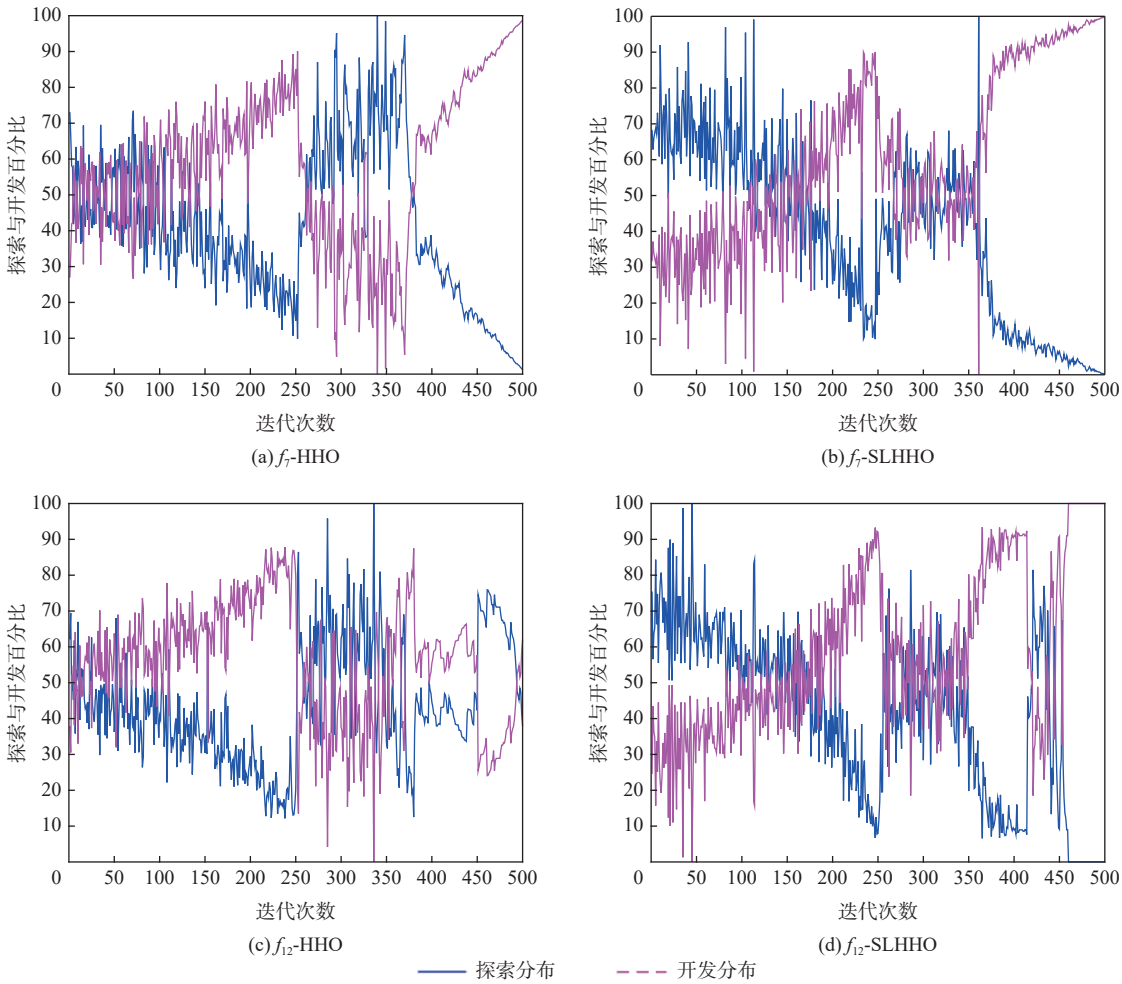


图5 HHO算法与SLHHO算法的探索与开发分布

Fig. 5 Distribution of exploration and development of HHO algorithm and SLHHO algorithm

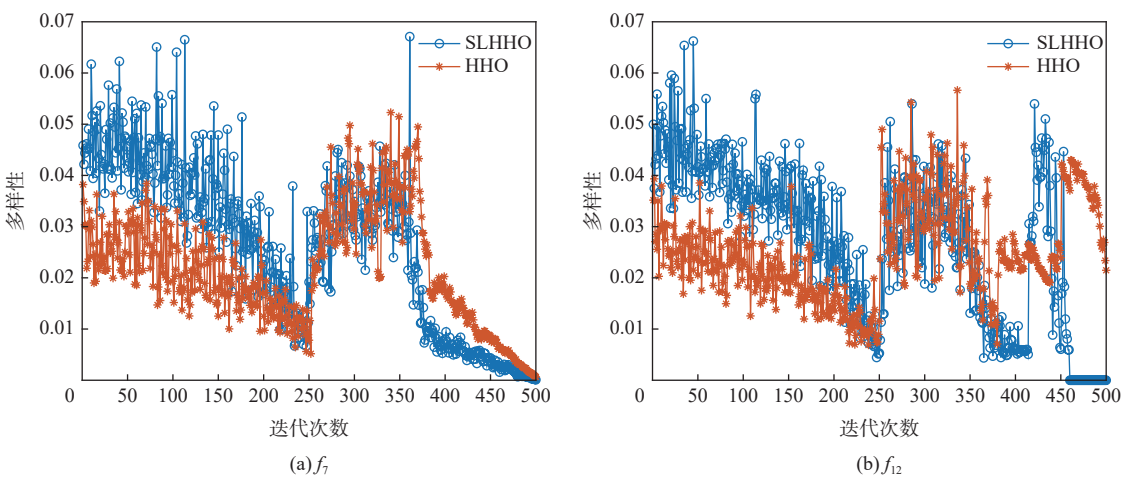


图6 HHO算法与SLHHO算法的多样性分布

Fig. 6 Diversity distribution of HHO algorithm and SLHHO algorithm

为紊乱; 由此可知, SLHHO 算法勘探与开发的相互作用, 能够适应不同环境的寻优。在种群多样性分布的图 6 中, 在单峰 f_7 函数中, SLHHO 算法前期具有较好的多样性, 后期能够较快地降低种群多样性, 提升收敛精度; 在多峰 f_{12} 函数中, 与 f_7 函数不同的是, SLHHO 算法能够在后期具有较快的变化, 这能够减小算法陷入局部最优的概率, 具有较好的抗局部吸引能力[38]。综合来看, SLHHO 算法在面对不同类型的函数时能够做出不同的响应, 提升解的精度, 体现了 SLHHO 算法具有种群学习型及适应性的特点。

4　压力容器设计问题

为验证 SLHHO 算法的实用性, 将其与 HHO 算法在压力容器设计问题中进行实验。压力容器设计的最大目标是减小圆柱形压力容器的制造成本, 同时能够满足实际的约束要求。其变量包括壳体的厚度 T_s、头部的厚度 T_h、内半径 R 和不带头部的圆柱形部分的长度 L。通过算法找到最小经济成本的最优变量[39]。将 PSO、GWO、WOA、HHO 与 SLHHO 进行比较, 各算法独立运行 30 次, 且种群数量为 50, 最大评估次数为 40 000, 同时记录 30 次数据的最优值 f_{best}, 平均值 f_{mean}, 最差值 f_{worst}, 最优参数, 具体结果如表 7 所示。

由表 7 可知, SLHHO 算法能够在 30 次运行结果中, 找到合理的参数且 f_{best} 指标最小, 其他算法在 f_{best} 指标都较差; 另一方面, SLHHO 算法在 f_{mean} 指标也最小, 意味着 SLHHO 算法是比较稳定的。综合来看, SLHHO 算法在压力容器设计问题中具有一定的可行性, 同时验证了 SLHHO 算法的实用性。

表 7　压力容器设计问题优化结果

Table 7　Pressure vessel design problem optimization results

函数	f_{best}	f_{mean}	f_{worst}	T_s/mm	T_h/mm	R/mm	L/mm
SLHHO	5 887.536 4	5 974.215 5	7 225.623 1	0.778 6	0.385 3	40.340 3	199.715 3
HHO[9]	5 981.944 0	6 909.828 1	7 654.063 5	1.175 2	0.578 6	59.893 4	37.962 5
PSO[1]	5 951.927 0	6 200.667 7	7 217.045 7	0.815 3	0.403 0	42.244 8	174.830 9
GWO[3]	5 981.944 0	6 909.828 1	7 654.063 5	1.175 2	0.578 6	59.893 4	37.962 5
WOA[4]	6 103.979 4	7 578.926 9	10 436.704 4	0.844 2	0.408 9	42.748 7	168.742 4

5　结　论

为提升 HHO 算法在寻优过程中的学习性及适应性, 本文提出一种基于 SNR 的学习型算法 SLHHO 算法。主要工作及未来方向如下:

1) 本文提出 SNR 驱动的个体差异评估机制, 通过量化解的质量差异; 设计协调学习策略实现精英引导与邻域探索的自适应平衡; 并改进逃逸距离公式。

2) 在 CEC2017 基准测试中, SLHHO 算法在 30 维问题上优于 7 种对比算法; 压力容器设计案例中, 成本控制较好, 验证了工程实用性。

3) 从基准函数的实验测试结果来看, 在 100 维的情况下, SLHHO 算法的优化能力有待提高, 特别是在 f_1 与 f_6 中, 寻优结果存在明显的下降, 其主要原因是随着维度的升高, 协调学习参考维度信息的错误率也在提高。

下一步的工作重点是在种群具有一定的学习性与适应性的基础上, 如何提升维度间的可靠性交流, 并将算法用于更多的实际场景中, 展现更好的实用性。

参考文献 (References)

[1] KENNEDY J, EBERHART R. Particle swarm optimization[C]// Proceedings of the ICNN'95 - International Conference on Neural Networks. Piscataway: IEEE Press, 1995: 1942-1948.

[2] DORIGO M, BIRATTARI M, STUTZLE T. Ant colony optimization[J]. IEEE Computational Intelligence Magazine, 2006, 1(4): 28-39.

[3] MIRJALILI S, MIRJALILI S M, LEWIS A. Grey wolf optimizer[J]. Advances in Engineering Software, 2014, 69: 46-61.

[4] MIRJALILI S, LEWIS A. The whale optimization algorithm[J]. Advances in Engineering Software, 2016, 95: 51-67.

[5] XUE J K, SHEN B. A novel swarm intelligence optimization approach: sparrow search algorithm[J]. Systems Science & Control Engineering, 2020, 8(1): 22-34.

[6] ZHAO W G, ZHANG Z X, WANG L Y. Manta ray foraging optimization: an effective bio-inspired optimizer for engineering applications[J]. Engineering Applications of Artificial Intelligence, 2020, 87: 103300.

[7] XIE L, HAN T, ZHOU H, et al. Tuna swarm optimization: a novel swarm-based metaheuristic algorithm for global optimization[J]. Computational Intelligence and Neuroscience, 2021, 2021(1): 9210050.

[8] JIANG X Y, LI S. BAS: beetle antennae search algorithm for optimization problems[EB/OL]. (2017-10-30)[2023-02-01]. https://arxiv.org/abs/1710.10724.

[9] HEIDARI A A, MIRJALILI S, FARIS H, et al. Harris hawks optimization: algorithm and applications[J]. Future Generation Computer Systems, 2019, 97: 849-872.

[10] JIAO S, CHONG G S, HUANG C C, et al. Orthogonally adapted

Harris hawks optimization for parameter estimation of photovoltaic models[J]. Energy, 2020, 203: 117804.

[11] 闫晓斌, 方洋旺, 彭维仕. 基于自适应高斯变异的多目标哈里斯鹰优化算法[J]. 北京航空航天大学学报, 2024, 50(8): 2636-2645.
YAN X B, FANG Y W, PENG W S. Multi-objective Harris hawk optimization algorithm based on adaptive Gaussian mutation[J]. Journal of Beijing University of Aeronautics and Astronautics, 2024, 50(8): 2636-2645(in Chinese).

[12] ELGAMAL Z M, YASIN N B M, TUBISHAT M, et al. An improved Harris hawks optimization algorithm with simulated annealing for feature selection in the medical field[J]. IEEE Access, 2020, 8: 186638-186652.

[13] 胡春安, 熊昱然. 多策略改进的混沌哈里斯鹰优化算法[J]. 计算机工程与科学, 2023, 45(9): 1648-1660.
HU C A, XIONG Y R. A multi-strategy improved chaotic Harris hawk optimization algorithm[J]. Computer Engineering & Science, 2023, 45(9): 1648-1660(in Chinese).

[14] LI C Y, LI J, CHEN H L, et al. Enhanced Harris hawks optimization with multi-strategy for global optimization tasks[J]. Expert Systems with Applications, 2021, 185: 115499.

[15] 李岩, 钱谦. 基于多种群与协同量子化的哈里斯鹰优化算法[J]. 控制与决策, 2024, 39(7): 2169-2176.
LI Y, QIAN Q. Harris hawks optimization algorithm based on multigroup and collaborative quantization[J]. Control and Decision, 2024, 39(7): 2169-2176(in Chinese).

[16] 柴岩, 任生. 双向经验引导与极端个体调控的 HHO 算法[J]. 计算机科学与探索, 2023, 17(9): 2118-2136.
CHAI Y, REN S. HHO algorithm based on bidirectional experience guidance and extreme individual regulation[J]. Journal of Frontiers of Computer Science and Technology, 2023, 17(9): 2118-2136(in Chinese).

[17] 刘小龙, 梁彤缨. 基于方形邻域和随机数组的哈里斯鹰优化算法[J]. 控制与决策, 2022, 37(10): 2467-2476.
LIU X L, LIANG T Y. Harris hawk optimization algorithm based on square neighborhood and random array[J]. Control and Decision, 2022, 37(10): 2467-2476(in Chinese).

[18] 展广涵, 王雨虹, 刘昊. 混合策略改进的哈里斯鹰优化算法及其应用[J]. 传感技术学报, 2022, 35(10): 1394-1403.
ZHAN G H, WANG Y H, LIU H. Improved Harris hawks optimization algorithm with hybrid strategy and its application[J]. Chinese Journal of Sensors and Actuators, 2022, 35(10): 1394-1403(in Chinese).

[19] 赵峰, 徐丹华. 基于混沌扰动与柯西变异的 HHO 算法[J]. 安徽大学学报 (自然科学版), 2023, 47(4): 25-34.
ZHAO F, XU D H. HHO algorithm based on chaotic perturbation and Cauchy mutation[J]. Journal of Anhui University (Natural Science Edition), 2023, 47(4): 25-34(in Chinese).

[20] 贾鹤鸣, 刘庆鑫, 刘宇翔, 等. 融合动态反向学习的阿奎拉鹰与哈里斯鹰混合优化算法[J]. 智能系统学报, 2023, 18(1): 104-116.
JIA H M, LIU Q X, LIU Y X, et al. Hybrid Aquila and Harris hawks optimization algorithm with dynamic opposition-based learning[J]. CAAI Transactions on Intelligent Systems, 2023, 18(1): 104-116(in Chinese).

[21] YU J H, YANG J H. Adaptive multi-strategy learning particle swarm optimization with evolutionary state estimation[M]//Bio-Inspired Computing: theories and Applications. Berlin: Springer, 2022: 174-186.

[22] WEINBERGER K Q, SAUL L K. Distance metric learning for large margin nearest neighbor classification[J]. Journal of Machine Learning Research, 2009, 10: 207-244.

[23] YUAN T T, DENG W H, TANG J, et al. Signal-to-noise ratio: a robust distance metric for deep metric learning[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2019: 4810-4819.

[24] YANG J H, YU J H, HUANG C. Adaptive multistrategy ensemble particle swarm optimization with signal-to-noise ratio distance metric[J]. Information Sciences, 2022, 612: 1066-1094.

[25] 高建瓴, 喻明毫. 混合差分进化与二次插值改进哈里斯鹰算法[J]. 小型微型计算机系统, 2023, 44(7): 1352-1359.
GAO J L, YU M H. Hybrid differential evolution and quadratic interpolation improved Harris hawks optimization algorithm[J]. Journal of Chinese Computer Systems, 2023, 44(7): 1352-1359(in Chinese).

[26] DEVAN P A M, HUSSIN F A, IBRAHIM R B, et al. An arithmetic-trigonometric optimization algorithm with application for control of real-time pressure process plant[J]. Sensors, 2022, 22(2): 617.

[27] KHODADADI N, SNASEL V, MIRJALILI S. Dynamic arithmetic optimization algorithm for truss optimization under natural frequency constraints[J]. IEEE Access, 2022, 10: 16188-16208.

[28] ZHANG Y Y, CHI A N, MIRJALILI S. Enhanced Jaya algorithm: a simple but efficient optimization method for constrained engineering design problems[J]. Knowledge-Based Systems, 2021, 233: 107555.

[29] NADIMI-SHAHRAKI M H, TAGHIAN S, MIRJALILI S. An improved grey wolf optimizer for solving engineering problems[J]. Expert Systems with Applications, 2021, 166: 113917.

[30] NABIL E. A modified flower pollination algorithm for global optimization[J]. Expert Systems with Applications, 2016, 57: 192-203.

[31] MIRJALILI S, HASHIM S Z M. A new hybrid PSOGSA algorithm for function optimization[C]//Proceedings of the International Conference on Computer and Information Application. Piscataway: IEEE Press, 2010: 374-377.

[32] TANG Z Y, ZHANG D X. A modified particle swarm optimization with an adaptive acceleration coefficients[C]//Proceedings of the Asia-Pacific Conference on Information Processing. Piscataway: IEEE Press, 2009: 330-332.

[33] DEHKORDI A A, SADIQ A S, MIRJALILI S, et al. Nonlinear-based chaotic Harris hawks optimizer: algorithm and Internet of vehicles application[J]. Applied Soft Computing, 2021, 109: 107574.

[34] HUSSAIN K, ZHU W, MOHD SALLEH M N. Long-term memory Harris' hawk optimization for high dimensional and optimal power flow problems[J]. IEEE Access, 2019, 7: 147596-147616.

[35] NAIK M K, PANDA R, WUNNAVA A, et al. A leader Harris hawks optimization for 2-D Masi entropy-based multilevel image thresholding[J]. Multimedia Tools and Applications, 2021, 80(28): 35543-35583.

[36] ARINI F Y, CHIEWCHANWATTANA S, SOOMLEK C, et al. Joint opposite selection (JOS): a premiere joint of selective leading

opposition and dynamic opposite enhanced Harris' hawks optimization for solving single-objective problems[J]. Expert Systems with Applications, 2022, 188: 116001.

[37] DEMSAR J. Statistical comparisons of classifiers over multiple data sets[J]. Journal of Machine Learning Research, 2006, 7: 1-30.

[38] POLÁKOVÁ R, TVRDÍK J, BUJOK P. Differential evolution with adaptive mechanism of population size according to current population diversity[J]. Swarm and Evolutionary Computation, 2019, 50: 100519.

[39] 张吉祥, 张孟健, 王德光. 增强型白骨顶鸡优化算法及其应用[J]. 小型微型计算机系统, 2024, 45(10): 2401-2410.
ZHANG J X, ZHANG M J, WANG D G. Enhanced coot optimization algorithm and its application[J]. Journal of Chinese Computer Systems, 2024, 45(10): 2401-2410.

Learning Harris Hawks optimization algorithm with signal-to-noise ratio

ZHANG Lin[1], SHEN Jiaying[2], HU Chuanlu[1], ZHU Donglin[2,*]

(1. School of Software Engineering, Jiangxi University of Science and Technology, Nanchang 330000, China;
2. School of Computer Science and Technology, Zhejiang Normal University, Jinhua 321004, China)

Abstract: Aiming at the problem of insufficient population learning and adaptability of the Harris hawks optimization (HHO) algorithm, this paper proposes a learning Harris hawks optimization based on the signal-to-noise ratio(SLHHO)algorithm. By using the signal-to-noise ratio as a metric to assess individual position information, the algorithm creates a coordinated learning strategy that can more realistically update the positions of individuals within the population. It then reworks the escape distance to enhance the algorithm's capacity for adaptation and optimization seeking. With 12 benchmark functions as the standard, the proposed algorithm was tested for its performance with the variants of the Harris hawk algorithm and other algorithms, and compared and analyzed in the evaluation indexes such as time complexity, diversity, exploration and development, etc. The results show that SLHHO algorithm is highly competitive and feasible, and finally, its practicability is verified in the pressure vessel design problem.

Keywords: Harris hawks optimization; signal to noise ratio; coordinated learning; escape distance; benchmark function

Received: 2023-07-03; **Accepted**: 2023-11-27; **Published Online**: 2024-02-04 10:56
URL: link.cnki.net/urlid/11.2625.V.20240202.1346.003
Foundation items: National Natural Science Foundation of China (62272418, 62002046)
* **Corresponding author**. E-mail: 2078570674@qq.com

http://bhxb.buaa.edu.cn jbuaa@buaa.edu.cn
DOI: 10.13700/j.bh.1001-5965.2023.0333

面向低延迟视频压缩感知的搜索窗自适应重构

孙仁慧，刘浩*，邓开连，燕帅
(东华大学 信息科学与技术学院，上海 201620)

摘 要：面向分布式视频压缩感知，帧间多假设预测能够降低编码端的运算量、提高解码端非关键帧的恢复质量，因此，近年来出现了很多与其相关的优化算法。然而在现有算法中，假设集的搜索窗口是大小经验固定的正方形区域。为进一步提高假设集质量、降低解码端时延，提出一种搜索窗口位置和大小自适应变化的重构算法。所提算法根据光流法快速确定相邻非关键帧之间的运动向量；联合该运动向量和前向相邻非关键帧与关键帧之间的运动信息，在关键帧中确定搜索窗口的中心块位置；由当前重构块与搜索窗口中心块的相对位置关系自适应地确定一个符合运动变化的矩形搜索窗口。在低延迟框架下对多个视频序列进行实验分析。实验结果表明：所提算法能够有效提高非关键帧的恢复质量，并减少运行时间。

关 键 词：自适应；光流法；运动向量；低延迟框架；矩形搜索窗口
中图分类号：TN919.8
文献标志码：A **文章编号**：1001-5965(2025)07-2374-10

随着多媒体通信技术的发展，人们可以随时随地从手机、电脑上获得各类信息，对于信息的表达从原本的文字和声音转向了更为生动形象的图像和视频。各类多媒体应用已成为人们日常生活中必不可少的部分。目前网络数据流量正在呈爆发性增长，提升信号传输能力非常必要。传统的奈奎斯特采样定理认为想要让采样之后的数字信号完整的保留原始信号中的信息，采样频率必须大于信号中最高频率的2倍。压缩感知(compressed sensing, CS)理论证明，如果信号是稀疏的或在其变换域中稀疏，那么其可以由远低于奈奎斯特采样定理要求的采样点数量重构恢复[1-4]。CS技术在采样的过程中就完成了压缩，减少了编码端的计算资源浪费。对于视频编解码系统，由于视频的信号量庞大，CS技术能降低编码端计算复杂度的特性对其更具有重要意义[5]。在视频CS的实际应用中，特别是高分辨率图像，视频序列的每帧都含有大量的像素点，需要有足够的硬件资源支撑。为降低视频CS的计算复杂度，文献[6]提出基于块的CS(block compressed sensing, BCS)采样方案，同时将视频序列划分为多个帧数相同的图像组(group of pictures, GOP)。目前视频和图像CS技术中的采样方案大多基于GOP-BCS系统。

为将BCS理论更好的运用于视频传输系统，文献[7]提出一种分布式编码和流媒体系统，一定程度上降低了编码端的计算复杂度，但该系统对于关键帧采用的是传统的H.26X帧内编解码技术。文献[8]提出一种完全基于GOP-BCS采样方案的分布式视频CS(distributed compressive video sensing, DCVS)系统，是目前视频CS研究的主流架构。提高DCVS系统解码端非关键帧的恢复质量是视频CS技术的难点。文献[9]提出运动补偿(motion compensation, MC)重构算法。该算法先通过运动估计在参考帧中确定最优匹配块，再使用运动补偿的残

收稿日期：2023-06-09；**录用日期**：2023-11-24；**网络出版时间**：2023-12-21 15:25
网络出版地址：link.cnki.net/urlid/11.2625.V.20231220.1726.001
基金项目：国家自然科学基金(62001099)
*通信作者. E-mail：liuhao@dhu.edu.cn

引用格式：孙仁慧，刘浩，邓开连，等. 面向低延迟视频压缩感知的搜索窗自适应重构[J]. 北京航空航天大学学报，2025，51(7)：2374-2383. SUN R H，LIU H，DENG K L，et al. Adaptive search window reconstruction for low-delay video compressive sensing[J]. Journal of Beijing University of Aeronautics and Astronautics，2025，51(7)：2374-2383 (in Chinese).

差重构非关键帧。文献 [10-11] 均采用多假设(multihypothesis, MH)预测的线性组合来估计运动补偿帧。MC 和 MH 算法都融合了基于 BCS 采样的平滑投影 Landweber(block compressed sensing with smoothed projected Landweber, BCS-SPL)重构技术[12]。MH 算法的解码端有 2 个主要组成部分, 即获取假设集和假设权重估计。优化 MH 算法是目前的研究热点。文献 [13] 使用 Tikhonov 正则化项来优化假设权重估计, 但是该方法不能精准的控制每个预测块的权重并去除相关度低的假设候选块。文献 [14] 为使权重合理, 在 L2 正则化项中加入了权重调整函数。为求解最优权值, 文献 [15] 将 Lasso 模型中的 L1 正则化和回归模型中的 L2 正则化相结合来解决加权问题。对于帧间的块匹配, 文献 [16] 提供了一种假设预测的最优选择算法, 能够提高匹配的精度。

在 DCVS 系统中, 每个 GOP 包含多个非关键帧, 随着关键帧与非关键帧的时域距离增加, 帧间相关性会显著降低。针对这一问题, 文献 [17] 提出一种基于多参考系的局部二次重构算法。为获取高质量的假设集, 文献 [18] 提出一种新的多假设(new multihypothesis, NMH)预测方案, 该方案联合前向的关键帧和相邻非关键帧获取最优匹配块, 充分的利用视频序列的时空相关性。文献 [19] 提出一种基于迭代渐进假设(iterative progressive hypothesis, IPH)预测的视频 CS 帧间重构算法。非关键帧在迭代重构的过程中, 其细节信息逐渐明显, IPH 算法通过自适应地剔除假设集中相似度低的假设候选块以提高恢复质量。

在上述基于 DCVS 系统的算法中, 每个非重叠块在重构时, 一般以参考帧中对应位置块为中心向外扩展固定像素数构建搜索窗口。即搜索窗口的形状是大小固定的正方形。然而, 视频序列中不同空间位置、不同时间段的运动变化程度存在差异, 对于搜索窗口的大小要求不同。为此, 本文提出一种面向低延迟视频 CS 的搜索窗自适应运动补偿重构(window-adaptive motion-compensated reconstruction, WMR)算法。本文算法的主要创新点如下: ①使用光流法快速计算相邻非关键帧之间的运动向量; ②联合帧间运动向量和前向相邻非关键帧到关键帧的运动信息来确定关键帧中搜索窗口的中心块位置; ③根据当前块和搜索窗口中心块的相对位置关系判断帧间的运动变化趋势, 自适应地确定一个符合运动变化的矩形搜索窗口。本文在帧间多假设预测下使用不同大小的搜索窗口进行对比实验, 并将本文算法与 MC、NMH 和 IPH 算法进行对比分析。结果表明: 本文 WMR 算法能够获取质量更高的假设集, 且在客观指标和主观观察方面都优于其他算法。

1　相关工作

1.1　DCVS 系统

图 1 为 DCVS 的系统架构。DCVS 系统的编码端完全基于 GOP-BCS 采样, 其中, 关键帧使用高采样率, 非关键帧使用低采样率。在 DCVS 系统的解码端, 关键帧采用帧内重构, 非关键帧利用关键帧提供的边信息辅助重构。DCVS 系统极大地简化了编码端的计算, 这对于上行流媒体应用的发展具有重要意义[20]。提升 DCVS 系统解码端非关键帧的恢复质量是目前的研究热点和难点, 经典的算法如 MC 算法、MH 算法等。

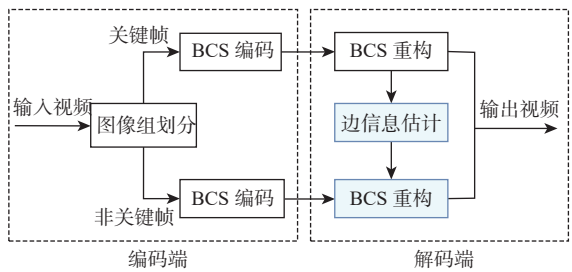


图 1　DCVS 系统架构

Fig. 1　DCVS system architecture

首先, MC 算法通过运动估计在参考帧中找到当前帧上每个非重叠块的最优匹配块, 然后, 由最优匹配块构建的运动补偿帧与当前帧的残差进行 BCS-SPL 重构。运动估计常用块匹配算法(block-matching algorithm, BMA)。该算法假设块中所有像素位移量相同。图 2 为使用 BMA 进行运动估计的过程, 在参考帧中以当前重构块的对应位置为中心向外扩展 w 像素数构建搜索窗口, 按照一定的匹配准则在窗口中寻找最优匹配块。运动向量 $\boldsymbol{V}$ 由当前块和最优匹配块的相对位移来表示。最优匹配块的搜索方式有全局匹配和快速匹配。全局匹配在搜索窗口中逐像素寻找最优匹配块, 耗时长, 匹配次数为 $(2w+1)^2$。快速匹配的搜索方法有三步法、菱形搜索法、六边形搜索算法等, 这些算法减少了

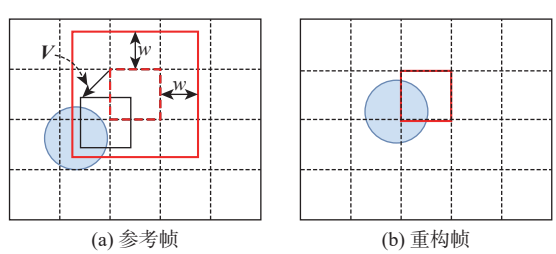


图 2　重构中的运动估计过程

Fig. 2　Motion estimation process in reconstruction

匹配次数，但容易陷入局部最优[21]。匹配准则用来衡量匹配块和当前重构块的相似性，常用平均绝对差异 (mean absolute difference, MAD) 来表示：

$$\mathrm{MAD}(a,b)=\frac{1}{B^2}\sum_{m=0}^{B-1}\sum_{n=0}^{B-1}|C(x+m,y+n)-R(x+a+m,y+b+n)| \tag{1}$$

式中：B为块的边尺寸；m和n为块内像素索引值；a和b分别为水平和垂直位移；(x,y)为当前块的坐标；$C(x+m,y+n)$为当前重构块的像素值；$R(x+a+m,y+b+n)$为参考帧上匹配块的像素值。

则运动向量$\boldsymbol{V}$的(u,v)表示如下，其中，u和v分别为运动向量$\boldsymbol{V}$水平和垂直分量：

$$(u,v)=\arg\min_{(a,b)}\mathrm{MAD}(a,b)\qquad a,b\in[-w,w] \tag{2}$$

在 MH 算法中，首先，在参考帧的搜索窗口中寻找多个相似度高的匹配块，也称假设候选块；然后，通过多个假设候选的线性组合可以得到运动补偿帧。在选取假设候选块时，先通过运动估计寻找最优匹配块，再以最优匹配块为中心向外扩展固定像素构建搜索窗口。现有算法为降低计算复杂度，大多直接使用参考帧中的对应位置块作为搜索窗口中心。NMH 和 IPH 算法是最新的针对 MH 的优化算法。NMH 算法充分利用了视频序列的时空相关性，联合使用前向的非关键帧和关键帧寻找最优匹配块。IPH 算法考虑到了假设候选块数量对于非关键帧恢复的影响，在迭代重构过程中不断剔除低质量的假设候选块以提高恢复质量。

上述基于 DCVS 系统的非关键帧重构算法从关键帧中获得了有效的边信息，提高了非关键帧恢复质量。但是上述算法的假设集搜索范围都是固定大小的正方形窗口。对于运动变化微弱的位置，帧间的运动向量模值小，小尺寸的搜索窗口中就能找到最优匹配块和高质量的假设集，过大的搜索窗口会增加非必要的块匹配运算。对于运动变化剧烈的位置，搜索窗口的尺寸过小，最优匹配块可能会在搜索窗口之外。且运动变化剧烈的位置，图像发生形变的可能性大，适当扩大假设集的搜索范围，选取窗口中其他结构相似的块作为假设候选块有助于提高恢复质量。同时，正方形的搜索窗口没有考虑到水平方向和竖直方向运动幅度的差异。本文算法可有效解决上述问题。

1.2 低延迟框架

低延迟框架是指解码端接收一帧就立即重构输出一帧，即非关键帧重构只能使用已重构完成的前向帧作为参考。图 3 为基于 DCVS 系统的低延迟框架。其中，key1表示当前 GOP 中的关键帧，

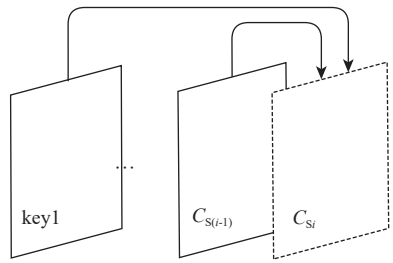


图 3 低延迟框架

Fig. 3 Low-delay framework

C_{Si}表示第i个非关键帧。因为关键帧的采样率高，恢复质量好，而相邻前向非关键帧$C_{S(i-1)}$的时域距离近，故C_{Si}重建时，key1和$C_{S(i-1)}$常被用作参考帧。低延迟框架常被应用于智能传感系统中的即时通信，所以面向低延迟框架的非关键帧重构算法需要着重考虑运行时间。

2 本文算法

2.1 本文算法主要流程

针对现有算法的问题，本文提出 WMR 算法。图 4 为本文算法的流程，展现了一个 GOP 中第i个非关键帧在关键中确定假设集搜索窗口位置和范围过程。$i=1$表示 GOP 中的第一个非关键帧，与关键帧相邻。本文 WMR 算法主要目标是在低延迟框架下从关键帧中确定一个位置和范围更加合理的假设集搜索窗口，从而获得高质量的假设集和准确的最优匹配块。

首先，使用 BCS-SPL 算法初步重构当前帧。如

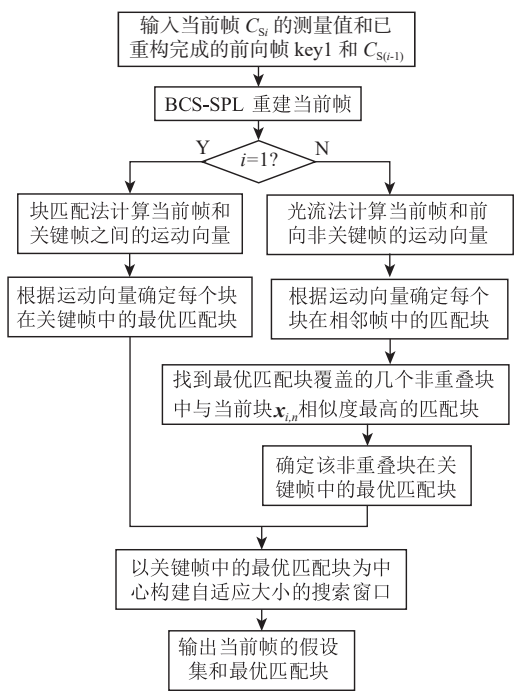


图 4 本文算法流程

Fig. 4 Flow of the proposed algorithm

果前向相邻帧是关键帧, 使用块匹配法计算帧间运动向量; 否则, 使用光流法计算相邻非关键帧之间的运动向量。由运动向量可以在相邻帧中确定一匹配块, 当前重构帧与关键帧相邻时, 则以该匹配块为假设集搜索窗口的中心; 如果当前帧和关键帧不相邻, 则计算该匹配块覆盖的几个非重叠块与当前块的相似度。找到相似度最高的块, 以该非重叠块在关键帧中所对应的最优匹配块为假设集搜索窗口的中心块。在已经重构完成的前向帧中, 所有非重叠块在关键帧中的最优匹配块已在重构时确定。最后, 考虑水平方向和垂直方向的运动程度不同, 根据搜索窗口的中心块和当前块的相对位置关系自适应地确定一个符合运动变化的矩形搜索窗口。该方案所确定的假设集搜索范围更加精准, 有效提高了假设集质量。同时, 该方案减少了非必要的块匹配运算, 降低了解码端的运行时间。

2.2　本文算法细节

2.2.1　光流法确定相邻帧间运动向量

传统的 MH 算法, 先使用 BMA 进行运动估计, 寻找最优匹配块。再以最优匹配块为中心构造搜索窗口。大量的块匹配算法会增加解码端的运行时间。在低延迟框架下, 随着非关键帧与关键帧的时域距离增加, BMA 难以获得准确的运动向量。

光流法通过图像序列中像素在时间域上的变化和相邻帧之间的相关性来确定帧间的运动关系。光流约束方程假设像素点在时间间隔很小的情况下, 其灰度值维持不变。若光流矢量在 x 轴和 y 轴上的分量分别用 u 和 v 来表示。在 x 轴方向、y 轴方向、时间 t 上灰度值的偏导数分别用 G_x、G_y、G_t 来表示, 则光流约束方程为

$$G_x u + G_y v + G_t = 0 \tag{3}$$

求解光流约束方程, 还需联立其他约束条件。LK(Lucas-Kanade)光流法[22] 是目前应用最广泛、最具代表性的方法, 其是在局部空间邻域上建立约束方程实现的。LK 光流法假设在一个较小的空间邻域 Ω(边长为 n, $n\geqslant 2$)内所有像素点的光流值一致, 并满足光流约束基本方程, 通过对邻域中的不同像素点赋予不同的权值来实现光流估计。光流估计误差为

$$\varepsilon = \sum_{\Omega} W^2(x,y)(G_x u + G_y v + G_t)^2 \tag{4}$$

式中: ε 为光流估计误差; $W^2(x,y)$ 为像素点 (x,y) 的权重函数, 则光流矢量为

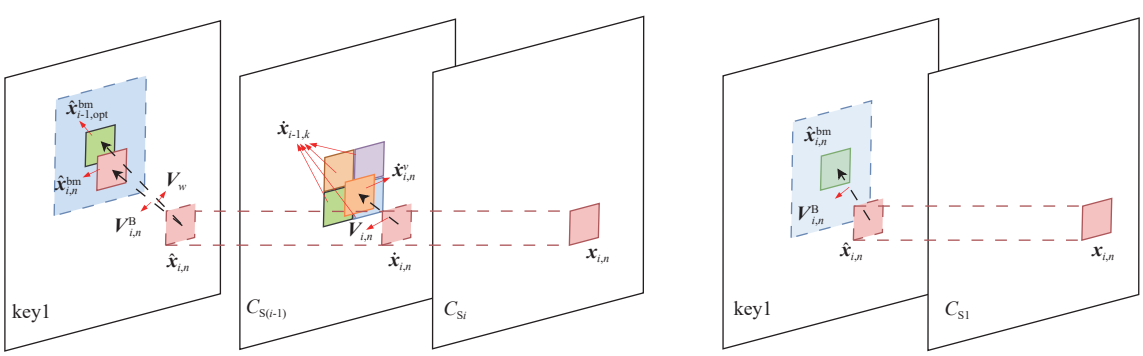
$$\begin{bmatrix} u \\ v \end{bmatrix} = \begin{bmatrix} \sum W^2(x,y)G_x^2(x) & \sum W^2(x,y)G_x(x)G_y(x) \\ \sum W^2(x,y)G_y(x)G_x(x) & \sum W^2(x,y)G_y^2(x) \end{bmatrix}^{-1} \cdot \begin{bmatrix} -\sum W^2(x,y)G_x(x)G_t(x) \\ -\sum W^2(x,y)G_y(x)G_t(x) \end{bmatrix} \tag{5}$$

LK 光流法的实时性好, 能够快速计算帧间的运动向量。LK 光流法的前提条件为: 亮度恒定; 2 帧之间的运动幅度较小; 空间一致, 临近点的运动变化相似。多数情况下, 相邻帧之间变化幅度不大, LK 光流法能够准确计算大多数位置的运动向量。但是对于运动幅度较大的位置光流法求得的运动向量可能存在误差。

光流法的目的是从相邻前向非关键帧中找到相似度最高的非重叠块作为参考块。由运动向量可在相邻帧中确定一匹配块, 计算该匹配块所覆盖的非重叠块与当前块的相似度。只要该匹配块与最优的非重叠块有重合, 就能够找到准确的参考块。所以光流法的误差在一定范围内不会影响最终结果。且后续确定搜索窗口中心位置的过程中起主要作用的是前向相邻非关键帧到关键帧的运动向量。

2.2.2　确定搜索窗口的位置

图 5(a) 为 GOP 中第 i 个非关键帧的重构过程。$\boldsymbol{x}_{i,n}$ 为当前重构块, 表示第 i 个非关键帧上的第 n 个非重叠块信息。首先, 通过光流法可以确定当前帧



(a) 重构帧与关键帧不相邻　　(b) 重构帧关键帧相邻

图 5　非关键帧 C_{Si} 的重构过程

Fig. 5　Reconstruction process of non-keyframe C_{Si}

C_{Si}和前向相邻帧$C_{S(i-1)}$之间的运动向量$\boldsymbol{V}_{i,n}$。$\dot{\boldsymbol{x}}_{i,n}$为$\boldsymbol{x}_{i,n}$在$C_{S(i-1)}$上的对应位置块，由其坐标和向量$\boldsymbol{V}_{i,n}$可以得到参考块位置：

$$\dot{\boldsymbol{x}}^{v}_{i,n}(j)=\dot{\boldsymbol{x}}_{i,n}(j)+\boldsymbol{V}_{i,n} \tag{6}$$

式中：j为像素索引，$\dot{\boldsymbol{x}}^{v}_{i,n}(j)$和$\dot{\boldsymbol{x}}_{i,n}(j)$为对应块的像素坐标。由于非关键帧在重构时都是按照在编码端划分的原始非重叠块逐块重构，故已重构完成的$C_{S(i-1)}$中只有非重叠块中含有到关键帧之间的运动信息。

进一步计算$\dot{\boldsymbol{x}}^{v}_{i,n}$覆盖的几个非重叠块$\dot{\boldsymbol{x}}_{i-1,k}(k\in[1,4])$和$\boldsymbol{x}_{i,n}$之间的欧氏距离，表达式如下：

$$d(\boldsymbol{x}_{i,n},\dot{\boldsymbol{x}}_{i-1,k})=\|\boldsymbol{y}_{i,n}-\boldsymbol{\Phi}\dot{\boldsymbol{x}}_{i-1,k}\|_2 \tag{7}$$

式中：$\boldsymbol{y}_{i,n}$为$\boldsymbol{x}_{i,n}$的测量值；$\boldsymbol{\Phi}$为测量矩阵。

选择与$\boldsymbol{x}_{i,n}$相似度最高的块$\dot{\boldsymbol{x}}^{\mathrm{opt}}_{i,n}$作为参考块：

$$\dot{\boldsymbol{x}}^{\mathrm{opt}}_{i,n}=\arg\min_{\dot{\boldsymbol{x}}_{i-1,k}} d(\boldsymbol{x}_{i,n},\dot{\boldsymbol{x}}_{i-1,k}) \tag{8}$$

$\dot{\boldsymbol{x}}^{v}_{i,n}$和理想的$\dot{\boldsymbol{x}}^{\mathrm{opt}}_{i,n}$只要有一点重叠的区域，都能够准确找相似度最高的非重叠参考块$\dot{\boldsymbol{x}}^{\mathrm{opt}}_{i,n}$，即光流法求相邻非关键帧之间的运动向量允许存在一定的误差。在低延迟框架下，此时前向帧均已重构完成，即已知$C_{S(i-1)}$与关键帧之间的运动信息，以$\dot{\boldsymbol{x}}^{\mathrm{opt}}_{i,n}$在关键帧中最优匹配块$\hat{\boldsymbol{x}}^{\mathrm{bm}}_{i-1,\mathrm{opt}}$为中心构建搜索窗口，通过逐像素滑动从搜索窗口中收集假设候选块$\boldsymbol{h}_l$。选取与当前块相似度最高的L个块作为假设集。得到假设集$H=\{\boldsymbol{h}_1,\boldsymbol{h}_2,\cdots,\boldsymbol{h}_L\}$，通过计算$H$中假设候选块的加权和，可以得到当前块的边信息。其中与当前块$\boldsymbol{x}_{i,n}$之间的欧氏距离最小，相似度最高的块为最优匹配块$\hat{\boldsymbol{x}}^{\mathrm{bm}}_{i,n}$。由$\hat{\boldsymbol{x}}^{\mathrm{bm}}_{i,n}$与当前块在关键帧中对应位置块$\hat{\boldsymbol{x}}_{i,n}$之间的相对位置可以得到运动向量$\boldsymbol{V}^{\mathrm{B}}_{i,n}$，该运动向量可以为后一非关键帧重构提供参考。

图5(b)为GOP第一个非关键帧C_{S1}的重构过程。该非关键帧和关键帧相邻，没有相邻的前向非关键帧作参考。这时使用BMA求当前帧C_{S1}和关键帧key1之间的运动向量比光流法更准确。由$\boldsymbol{V}^{\mathrm{B}}_{i,n}$求得$\boldsymbol{x}_{i,n}$在关键帧上的匹配块$\hat{\boldsymbol{x}}^{\mathrm{bm}}_{i,n}$。以该匹配块为中心构建搜索窗口。$\boldsymbol{V}^{\mathrm{B}}_{i,n}$为后一非关键帧重构提供参考。上述确定搜索窗口位置的方案，参考了前向相邻非关键帧到关键帧中的运动信息，减少了计算量。同时，相比直接使用关键帧的对应位置块为搜索窗口中心选取假设集更加准确。

2.2.3 自适应确定搜索窗口的大小

关键帧中的对应位置块$\hat{\boldsymbol{x}}_{i,n}$和$\hat{\boldsymbol{x}}^{\mathrm{bm}}_{i-1,\mathrm{opt}}$的坐标关系可以求得向量$\boldsymbol{V}_w$，从而判断当前块的运动幅度和方向。因此，本文根据向量$\boldsymbol{V}_w$来确定搜索窗口的范围。令向量$\boldsymbol{V}_w$的水平分量为u，垂直分量为v。则u和v可以分别表示当前块在水平方向和竖直方向的运动变化程度。如图6所示，根据u和v的数值，以$\hat{\boldsymbol{x}}^{\mathrm{bm}}_{i-1,\mathrm{opt}}$为中心水平方向外扩展$w_x$像素，竖直方向向外扩展$w_y$像素构建一个矩形搜索窗口。

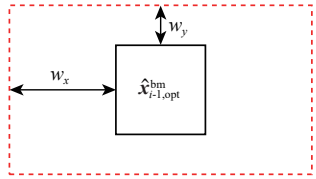


图6 搜索窗口构建方案

Fig. 6 Construction scheme of search window

假设图像被划分成了N个非重叠块，每个块都与关键帧存在一个运动向量$\boldsymbol{V}_w$。每个块的运动向量的水平分量绝对值为$|u_1|,|u_2|,\cdots,|u_N|$。令w_x的范围为$[w_{x,\min},w_{x,\max}]$。其中：

$$w_{x,\min}=\max[5,\min|u_n|] \qquad 1\leqslant n\leqslant N \tag{9}$$

$$w_{x,\max}=\frac{2}{N}\sum_{n=1}^{N}|u_n| \tag{10}$$

即w_x的下限为$\boldsymbol{V}_w$的最小水平分量和5中的较大值，上限为$\boldsymbol{V}_w$水平分量均值的2倍。确定上下限是为了将搜索窗口的大小控制在合理的范围内。文献[9]提出的MC算法w的大小设置为20，所测试的视频序列均可获得不错的恢复质量，文献[10-11]提出的帧内MH算法w的大小设置为6，说明在多数情况下，该尺寸的搜索窗口中有具有参考价值的假设候选块，因此，下限需保证不小于5同时不小于$\boldsymbol{V}_w$的最小水平分量。而w的上限应当与视频序列的运动变化程度相关，水平分量的均值能够表征当前帧相对关键帧的运动变化程度，因此，本文将上限设置成了均值的2倍，将$|u_1|,|u_2|,\cdots,|u_N|$线性映射到w_x的区间中得到对应的$w_{x,n}$：

$$w_{x,n}=\alpha(|u_n|-\min|u_n|)+w_{x,\min} \tag{11}$$

系数α为

$$\alpha=\frac{w_{\max}-w_{\min}}{\max|u_n|-\min|u_n|} \tag{12}$$

则第n个块在关键帧中的搜索窗口中心位置确定后，水平方向向外扩展$w_{x,n}$像素。$w_{y,n}$求解方法与$w_{x,n}$相同。

3 实验与分析

本节通过实验检验本文WMR算法在低延迟框架下的性能。对于测试源视频，使用News、Foreman和Football这3个不同运动程度的视频序列作为标准测试应用于本实验。3.1节和3.2节是2个消融

实验, 分别验证了根据运动向量自适应地确定搜索窗口的大小和位置对于非关键帧的恢复质量和重构时间的影响。3.3 节验证本文算法相较现有算法的性能优势。

对于参数设置, GOP 大小设为 8。在编码端, 视频帧被划分为固定大小 16×16 的非重叠块。关键帧的采样率为 0.7, 非关键帧采样率从 0.1～0.3 不等。峰值信噪比(peak signal-to-noise ratio, PSNR)指标用于评估不同算法的恢复质量, 通过 MATLAB 计算不同算法的速度性能。所有实验均在配备 Intel Core i7-7700HQ CPU、8GB 内存和 Windows 10 操作系统的计算机上使用 MATLAB R2018b 完成。

3.1　搜索窗口大小对于恢复质量的影响

为验证搜索窗口大小对于非关键帧重构的影响, 本文使用不同尺寸的固定大小的搜索窗口及本文提出的根据当前重构帧与关键帧之间运动向量确定一个自适应大小的矩形范围搜索窗口进行实验比较。所有搜索窗口的构建都是以当前重构块在关键帧中的对应位置为中心。为保证公平性, 本次实验除了搜索窗口范围设置方案不同外, 帧间多假设预测方案保持不变。

图 7 为不同采样率下, 使用不同大小的搜索窗口重建非关键帧的恢复质量对比。w 表征搜索窗口的大小。构建搜索窗口时, 以对应位置块为中心向外扩展固定像素数 w。3 个视频序列的恢复质量随着 w 的增加而增加。对于视频序列 News 和 Foreman, w 为 15 像素相比 w 为 10 像素时, 恢复质量提升明显, w 从 15 像素提升至 30 像素时, 恢复质量提高微弱。综合考虑运行时间和恢复质量, 对于视频序列 News, w 取 15 像素左右较为合适; 对于视频序列 Foreman, w 取 15～20 像素之间较为合适。对于运动变化程度较大的视频序列 Football, w 对于恢复质量的提升有限, w 设为 30 像素时能够获得相对较高的恢复质量。可见运动程度不同的视频序列所适合的搜索窗口尺寸不同。在 3 个视频序列中, 使用本文提出的根据运动向量将 w 设为自适应尺寸方案时, 任意采样率下均可获得相对较高的恢复质量, 且采样率越高, 自适应尺寸的搜索窗口对于恢复质量的提升越明显。

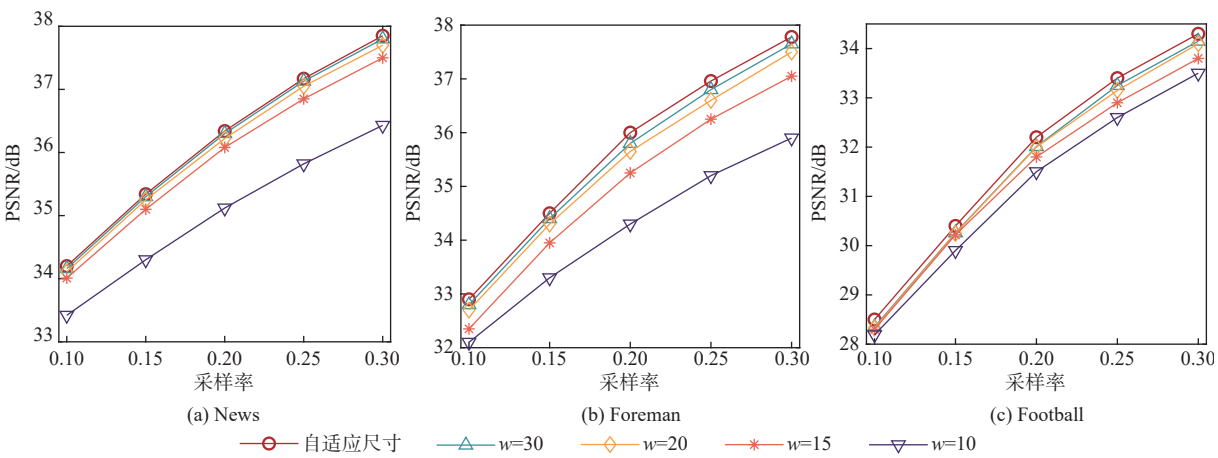


图 7　不同尺寸搜索窗口的恢复质量

Fig. 7　Restoration quality for search windows with different sizes

表 1 为不同尺寸的搜索窗口下, 块匹配过程 CPU 平均运行时间。如表 1 所示, CPU 的运行时间会随着搜索窗口尺寸的增加大幅增加。在不同采样率下, 使用自适应大小的搜索窗口比 w 设为 20 像素时运行时间减少了 5～9 s。但是恢复质量相对搜索窗口大小为 20 像素时, 3 个视频序列分别提升了 0.27 dB、0.41 dB、0.48 dB。综合考虑运行时间和恢复质量, 本文的自适应大小的搜索窗口方案不仅能够提高恢复质量, 实时性也更好, 更符合低延迟框架的要求。

表 1　不同窗口尺寸下块匹配过程的 CPU 运行时间

Table 1　CPU runtime of block matching process under different window sizes

窗口尺寸/像素	CPU运行时间/s				
	采样率为0.10	采样率为0.15	采样率为0.20	采样率为0.25	采样率为0.30
10	33.5	37.5	42.3	45.2	47.3
20	49.3	52.3	57.3	64.3	74.6
30	70.2	76.4	82.4	86.6	95.3
自适应尺寸	42.2	46.9	51.6	59.1	63.1

3.2　搜索窗口位置对于恢复质量的影响

为验证搜索窗口位置对于非关键帧重构的影响, 重构非关键帧时, 假设集搜索窗口分别使用本文提出的根据运动向量自适应确定搜索窗口位置的方案和固定位置的方案进行实验对比。所谓固定位置, 就是

搜索窗口中心位置始终与当前重构块位置相对应。

图 8 为 w 取不同固定值的情况下，使用 2 种方案重构非关键帧的恢复质量对比。在 3 个视频序列中，使用自适应位置方案时，w 取任意固定值均可取得相对较高的恢复质量。随着 w 的增加，恢复质量的提高逐渐变缓。由表 1 可知，w 越小，重构时间越短。为兼顾时效性和恢复质量，w 应当取图中恢复质量增加开始变缓位置的对应数值。使用自适应位置方案时，令搜索窗口尺寸为 w_1 像素；使用固定位置方案时，令搜索窗口尺寸为 w_2 像素。对于运动变化较慢的视频序列 News，w_1 取 10 像素与 w_2 取 15 像素恢复质量相当；对于运动变化适中的视频序列 Foreman，w_1 取 15 像素与 w_2 取 20 像素恢复质量相当；对于运动变化较快的视频序列 Football，w_1 取 15 像素与 w_2 取 30 像素恢复质量相当。即本文根据运动向量确定搜索窗口位置的方案能够降低对搜索窗口大小的要求，从而减少块匹配运算，提高非关键帧重建速度。

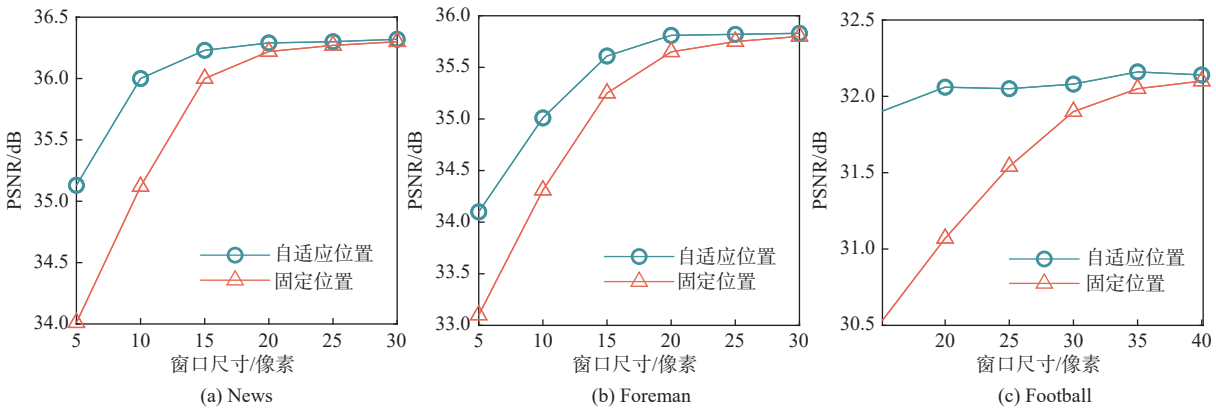


图 8 自适应位置与固定位置的搜索窗口恢复质量

Fig. 8 Restoration quality for adaptive-position and fixed-position search windows

3.3 综合性能比较

本文算法的目标是通过运动向量自适应地确定假设集的搜索窗口，从而得到高质量的假设集，以实现对帧间多假设预测的改进。为检验本文算法整体性能，将本文 WMR 算法和当前研究中几个主流算法进行比较。实验中所有算法方案都在低延迟框架下运行，即只有前向帧做参考。实验对比算法包括 MC、NMH 和 IPH。MC 算法通过运动估计在参考帧中寻找最优匹配块，直接由最优匹配块构建运动补偿帧。NMH 算法为得到高质量的假设集，联合相邻的非关键帧和前向关键帧来寻找最优匹配块、构建搜索窗口。充分利用视频序列的时空相关性。IPH 算法为充分的利用假设集中的有效信息，在迭代重构过程自适应地剔除低质量的假设候选块。NMH 算法使用弹性网络优化假设权重以提高恢复质量。为保证公平性，本文算法与其他算法的区别仅在于获取假设集的方案不同。假设权重设置方案及编码端的采样方案等保持一致。

表 2 为低延迟框架下，对于运动变化程度不同的视频序列，不同算法重构非关键帧的平均 PSNR

表 2 不同运动程度视频序列上不同算法的 PSNR

Table 2 PSNR of different algorithms on video sequences with different motion degrees

视频	算法	PSNR/dB					
		采样率为0.10	采样率为0.15	采样率为0.20	采样率为0.25	采样率为0.30	均值
News	MC[9]	31.19	32.26	33.38	34.24	34.97	33.21
	NMH[18]	33.92	34.86	36.04	36.88	37.24	35.79
	IPH[19]	34.12	35.02	36.46	37.32	37.63	36.11
	WMR	**34.35**	**35.29**	**36.72**	**37.56**	**37.94**	**36.37**
Foreman	MC[9]	31.49	32.57	34.12	34.78	35.79	33.75
	NMH[18]	33.52	34.70	36.35	37.04	38.14	35.95
	IPH[19]	33.77	34.97	36.65	37.36	38.48	36.25
	WMR	**34.11**	**35.42**	**36.94**	**37.66**	**38.76**	**36.58**
Football	MC[9]	25.34	27.11	28.63	29.93	31.10	28.42
	NMH[18]	27.35	29.16	30.83	32.26	33.45	30.61
	IPH[19]	27.61	29.43	31.10	32.57	33.77	30.90
	WMR	**27.83**	**29.67**	**31.35**	**32.84**	**34.06**	**31.15**

结果, 其中, 加粗数值为最优结果。可以看出, 在同一视频序列和采样率下, 使用 WMR 算法重构的非关键帧平均恢复质量优于其他 3 个算法。对于视频序列 News, WMR 算法相较于 MC、NMH 和 IPH 分别提高了 3.16 dB、0.58 dB、0.26 dB。对于视频序列 Foreman, 分别提高了 2.83 dB、0.63 dB、0.33 dB。对于视频序列 Football, 分别提高了 2.73 dB、0.54 dB、0.25 dB。可以看出对于不同运动程度的视频序列, WMR 算法重构的图像质量相对较好。

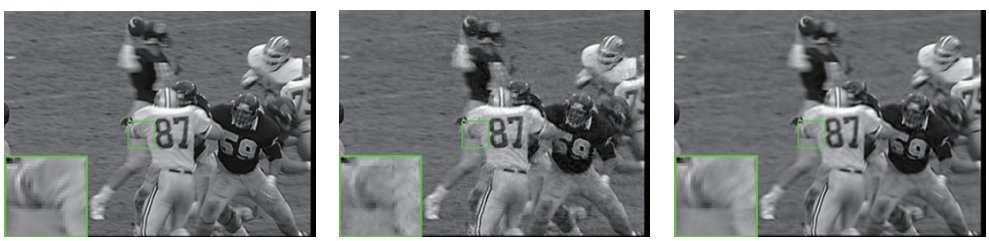
表 3 为不同算法的运行时间, 其中, 加粗数值为最优结果。可以看出, WMR 算法在不同的采样率下, CPU 运行的时间比最新的 NMH 和 IPH 算法减少了 4～12 s。

图 9 为非关键帧的采样率为 0.3 的情况下, 不同的算法对于 Football 视频序列中第 3 个非关键帧的重构结果。使用 MC、NMH、IPH 和 WMR 算法的恢复质量分别为 31.13 dB、33.48 dB、33.81 dB 和 34.11 dB。由图可知, 使用 MC 重构的图像有比较明显的块效应, 球员手臂处的衣服褶皱纹理丢失, 球员的白色上衣有明显的噪声。使用 NMH 和 IPH 算法相比较 MC 算法图像恢复质量明显提高, 但是相较原图, 手臂处的衣服褶皱依旧不够清晰, 同时 87 号所遮挡球员的腿部和草坪的分界线相较原图也更模糊。WMR 算法重构的图像与其他算法相比, 球员手臂处衣服褶皱的纹理更加明显, 球员头盔边缘的细节信息也更加清楚。综上所述, 从实验结果的客观指标数据及主观观察可以说明, 本文 WMR 算法相比最新算法 NMH 和 IPH 重构的图像质量更高, 且运行时间更短。虽然运行时间比 MC 算法长, 但是恢复质量相比 MC 算法有很大提升。

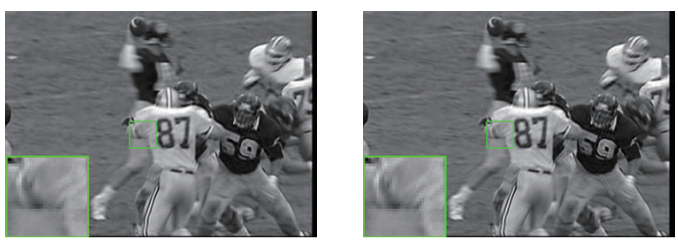
表 3　不同算法的 CPU 运行时间

Table 3　CPU runtime of different algorithms

算法	CPU运行时间/s				
	采样率为0.1	采样率为0.15	采样率为0.2	采样率为0.25	采样率为0.3
MC[9]	35.3	37.8	40.4	43.6	45.3
NMH[18]	60.3	34.6	67.7	70.2	75.6
IPH[19]	56.2	59.7	63.3	65.3	69.9
WMR	**52.2**	**55.4**	**57.3**	**59.5**	**63.8**



(a) 原图　(b) MC[9]　(c) NMH[18]



(d) IPH[19]　(e) WMR

图 9　不同算法恢复质量的主观比较

Fig. 9　Subjective comparison of restoration quality for different algorithms

4　结　论

1) 本文算法使用光流法快速计算相邻非关键帧之间的运动向量, 联合帧间运动向量和前向重构帧的已知运动信息自适应地在关键帧中确定搜索窗口的中心位置。

2) 本文算法根据当前块和关键帧中搜索窗口中心块的相对位置关系自适应地确定一个符合运动变化的矩形搜索窗口。

3) 本文搜索窗口大小自适应机制可以减少块匹配运算、提升解码端图像恢复质量; 本文搜索窗口中心位置自适应机制可以降低解码端对搜索窗口大小的要求、减少重构的复杂度。

4) 在低延迟框架下, 本文算法相较于基于帧间多假设预测的现有算法不仅提升了恢复质量, 也降低了解码端的运行时间。

参考文献 (References)

[1] CANDES E J, ROMBERG J, TAO T. Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information[J]. IEEE Transactions on Information Theory, 2006, 52(2): 489-509.

[2] 刘浩, 郑浩然, 黄荣. 面向量化分块 CS 的区域层次化预测编码

[J]. 北京航空航天大学学报, 2022, 48(8): 1376-1382.
LIU H, ZHENG H R, HUANG R. Region-hierarchical predictive coding for quantized block compressive sensing[J]. Journal of Beijing University of Aeronautics and Astronautics, 2022, 48(8): 1376-1382(in Chinese).

[3] JI Y, KANG Z, ZHANG X, et al. Model recovery for multi-input signal-output nonlinear systems based on the compressed sensing recovery theory[J]. Journal of the Franklin Institute, 2022, 359(5): 2317-2339.

[4] CANDES E J, WAKIN M B. An introduction to compressive sampling[J]. IEEE Signal Processing Magazine, 2008, 25(2): 21-30.

[5] WANG S W, YU L, XIANG S. A low complexity compressed sensing-based codec for consumer depth video sensors[J]. IEEE Transactions on Consumer Electronics, 2019, 65(4): 434-443.

[6] GAN L. Block compressed sensing of natural images[C]//Proceedings of the 15th International Conference on Digital Signal Processing. Piscataway: IEEE Press, 2007: 403-406.

[7] DO T T, CHEN Y, NGUYEN D T, et al. Distributed compressed video sensing[C]//Proceedings of the 16th IEEE International Conference on Image Processing. Piscataway: IEEE Press, 2009: 1393-1396.

[8] KANG L W, LU C S. Distributed compressive video sensing[C]//Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing. Piscataway: IEEE Press, 2009: 1169-1172.

[9] MUN S, FOWLER J E. Residual reconstruction for block-based compressed sensing of video[C]//Proceedings of the Data Compression Conference. Piscataway: IEEE Press, 2011: 183-192.

[10] CHEN C, TRAMEL E W, FOWLER J E. Compressed-sensing recovery of images and video using multihypothesis predictions[C]//Proceedings of the Conference Record of the 45th Asilomar Conference on Signals, Systems and Computers. Piscataway: IEEE Press, 2011: 1193-1198.

[11] TRAMEL E W, FOWLER J E. Video compressed sensing with multihypothesis[C]//Proceedings of the Data Compression Conference. Piscataway: IEEE Press, 2011: 193-202.

[12] FOWLER J E. Block-based compressed sensing of images and video[J]. Foundations and Trends® in Signal Processing, 2010, 4(4): 297-416.

[13] TROCAN M, TRAMEL E W, FOWLER J E, et al. Compressed-sensing recovery of multiview image and video sequences using signal prediction[J]. Multimedia Tools and Applications, 2014, 72(1): 95-121.

[14] KUO Y H, WU K, CHEN J. A scheme for distributed compressed video sensing based on hypothesis set optimization techniques[J]. Multidimensional Systems and Signal Processing, 2017, 28(1): 129-148.

[15] CHEN J, CHEN Y Z, QIN D, et al. An elastic net-based hybrid hypothesis method for compressed video sensing[J]. Multimedia Tools and Applications, 2015, 74(6): 2085-2108.

[16] ZHAO C, MA S W, GAO W. Image compressive-sensing recovery using structured Laplacian sparsity in DCT domain and multi-hypothesis prediction[C]//Proceedings of the IEEE International Conference on Multimedia and Expo. Piscataway: IEEE Press, 2014: 1-6.

[17] ZHENG S, ZHANG X P, CHEN J, et al. A high-efficiency compressed sensing-based terminal-to-cloud video transmission system[J]. IEEE Transactions on Multimedia, 2019, 21(8): 1905-1920.

[18] ZHENG S, CHEN J, ZHANG X P, et al. A new multihypothesis-based compressed video sensing reconstruction system[J]. IEEE Transactions on Multimedia, 2020, 23: 3577-3589.

[19] LIU H, SUN R H. Iterative progressive-hypothesis prediction for forward interframe reconstruction of video compressive sensing[C]//Proceedings of the IEEE 24th International Workshop on Multimedia Signal Processing. Piscataway: IEEE Press, 2022: 1-6.

[20] 刘浩, 黄荣, 袁浩东. 面向上行流媒体的CS视频流技术前沿[J]. 中国图象图形学报, 2021, 26(7): 1545-1557.
LIU H, HUANG R, YUAN H D. Survey on compressive sensing video stream for uplink streaming media[J]. Journal of Image and Graphics, 2021, 26(7): 1545-1557(in Chinese).

[21] 刘泉洋, 刘云清, 史俊, 等. 视频图像运动估计中的一维块匹配算法[J]. 计算机辅助设计与图形学学报, 2021, 33(3): 424-430.
LIU Q Y, LIU Y Q, SHI J, et al. One-dimensional block matching algorithm in video image motion estimation[J]. Journal of Computer-Aided Design & Computer Graphics, 2021, 33(3): 424-430(in Chinese).

[22] LIU Q P, XI U, ZHANG W C, et al. Improved image matching algorithm based on LK optical flow and grid motion statistics[J]. International Journal of Computer Applications in Technology, 2022, 68(1): 49-57.

Adaptive search window reconstruction for low-delay video compressive sensing

SUN Renhui，LIU Hao*，DENG Kailian，YAN Shuai

(College of Information Science and Technology，Donghua University，Shanghai 201620，China)

Abstract: For distributed video compressive sensing, the inter-frame multi-hypothesis prediction offers low computational complexity at the encoding end and good restoration quality for non-key frames at the decoding end. In recent years, many optimization algorithms related to it have been proposed. However, in existing algorithms, the search window of the hypothesis set is a square area whose size is empirically fixed. To further improve the quality of the hypothesis set and reduce the delay at the decoding end, this paper proposed a reconstruction algorithm that adaptively adjusted the position and size of the search window. The proposed algorithm first quickly determined the motion vector between adjacent non-key frames using the optical flow method. Then, by combining the motion vector and the motion information between the forward adjacent non-key frame and the key frame, the central block position of the search window in the key frame was determined. Finally, based on the relative position relationship between the current reconstruction block and the central block of the search window, a rectangular search window that aligned with motion changes could be determined adaptively. In the paper, several video sequences were experimentally analyzed within a low-delay framework. The results show that the proposed algorithm can effectively improve the restoration quality of non-key frames and reduce the runtime.

Keywords: adaptively; optical flow method; motion vector; low-delay framework; rectangular search window

Received：2023-06-09；**Accepted**：2023-11-24；**Published Online**：2023-12-21 15:25

URL：link.cnki.net/urlid/11.2625.V.20231220.1726.001

Foundation item：National Natural Science Foundation of China (62001099)

* **Corresponding author**. E-mail：liuhao@dhu.edu.cn

http://bhxb.buaa.edu.cn jbuaa@buaa.edu.cn
DOI: 10.13700/j.bh.1001-5965.2023.0379

无模态融合的高效弱监督视频时刻检索算法

蒋寻，徐行*，沈复民，王国庆，杨阳
(电子科技大学 计算机科学与工程学院，成都 611731)

摘 要：弱监督视频时刻检索（WSVMR）旨在基于视频与自然语言文本的匹配关系训练深度学习算法模型，以实现根据自然语言查询文本从未经修剪的视频中检索特定事件内容的起始与结束时间。大多数现有的WSVMR算法采用多模态融合机制来理解视频内容以完成时刻检索，限制了现有算法的运行效率，降低了该项技术在多媒体应用中的实用性。基于此，提出一种可实现快速WSVMR的无融合多模态对齐网络（FMAN）算法。该算法可以将复杂的跨模态交互计算全部限制在训练阶段，从而允许模型对视频数据和文本数据都进行离线编码，显著提高了视频时刻检索的推理速度。在Charades-STA数据集和ActivityNet-Captions数据集上的实验结果表明：FMAN算法所取得的检索性能与效率都优于现有算法：对于衡量检索性能的指标R_1召回率和R_5召回率，在Charades-STA数据集上，所提算法分别平均取得了2.66%和1.57%的性能提升；在ActivityNet-Captions数据集上，所提算法分别平均取得了0.19%和3.35%的性能提升；在检索效率上，所提算法将在线每秒浮点运算次数降低至原有算法的1%以下。

关 键 词：视频时刻检索；弱监督学习；媒体内容检索；模型效率；跨模态对齐
中图分类号：TP37
文献标志码：A **文章编号**：1001-5965(2025)07-2384-10

随着近年来移动互联网和网络技术的快速发展，视频数据出现爆炸式增长，大规模多媒体检索应用也得到了快速的发展。视频时刻检索(video moment retrieval, VMR)[1-4]旨在通过给定自然语言查询，高效地理解并检索未修剪长视频中的特定事件，从而提高多媒体应用用户对长视频内容的信息获取效率。传统的VMR算法采用完全监督学习范式，由专门的标注者标记的训练数据中每个事件在视频中的起始和结束时间，用作监督信号。然而，为每个事件标记时间边界是繁琐且主观的，这无形中增加了部署VMR模型的成本。为降低数据的标注成本，文献[5-7]通过将精确的时序标注松弛为点级标注或引入主动学习技术[8]，达到减少数据标注成本的目的。但是该类算法仍然无法避免额外的人工数据标注成本。为此，本文专注于一项更实用但也更具挑战性的任务，即弱监督视频时刻检索(weakly-supervised video moment retrieval, WSVMR)，其训练数据仅标记有视频-文本匹配对，不依赖于具体的时刻标注信息，因此，显著降低了注释成本。

现有WSVMR算法[9-12]普遍采用多模态融合机制，联合2个模态学习得到视频文本的统一表征来理解视频内容。通过细粒度的跨模态交互，基于融合的算法取得了显著的性能改进，但也带来了新的问题：在多媒体检索应用中，这些基于融合的算法必须先等待用户的文本查询，再进行细粒度多模态融合计算。因此，这类算法中绝大部分计算动作都必须在应用中在线推进，导致了大量的冗余时间和计算资源消耗，致使模型无法真正应用于多媒体检

收稿日期：2023-06-16；**录用日期**：2023-09-15；**网络出版时间**：2023-11-28 14:11
网络出版地址：link.cnki.net/urlid/11.2625.V.20231128.0939.001
基金项目：国家自然科学基金(61976049,62072080)
*通信作者. E-mail：xing.xu@uestc.edu.cn

引用格式：蒋寻，徐行，沈复民，等. 无模态融合的高效弱监督视频时刻检索算法[J]. 北京航空航天大学学报，2025，51(7)：2384-2393. JIANG X，XU X，SHEN F M，et al. Efficient weakly-supervised video moment retrieval algorithm without multimodal fusion[J]. Journal of Beijing University of Aeronautics and Astronautics，2025，51(7)：2384-2393 (in Chinese).

索系统中。此外, 由于上述限制, 这些算法也无法应用于视频级别的检索, 因为每次针对用户的检索查询输入, 都必须对视频数据集中的每个视频进行联合表征计算, 这在多数跨媒体检索应用场景下是无法接受的。

为解决上述问题, 本文提出一种无融合多模态对齐网络 (fusion-free multimodal alignment network, FMAN) 算法, 以绕开现有模型在推理阶段大量的细粒度模态融合计算, 在保证检索性能的同时实现高效时刻检索。FMAN 算法在测试阶段仅需要进行模态内建模, 无需模态交互, 从而使得所有视频建模的计算都可以采用离线的预处理方式实现, 而在实际多媒体检索应用中仅需要对用户输入的查询文本进行模态内编码即可。具体来说, 本文 FMAN 算法主要包括 3 个关键组件: ①事件候选生成器 (event candidate generator, ECG); ②模态内编码器 (intra-modality encoders, IME); ③多模态对齐学习模块 (multimodal alignment learning module, MALM)。ECG 用于学习未修剪长视频中的潜在事件, 并生成多个部分视频作为事件候选。随后, 所有潜在事件候选都将被输入到 IME, 其通过重新加权的注意力机制强化每个候选事件对应的视频时序特征。MALM 旨在进行视频文本对齐学习, 且仅在训练阶段启用, 达到将所有的跨模态交互动作限制在训练阶段的目的。本文 FMAN 采用 2 阶段训练策略: 首先, 使用 MALM 来训练 ECG; 然后, 冻结 ECG 的参数, 并使用重置参数的 MALM 来训练 IME。在测试阶段, FMAN 将直接计算文本编码和每个事件候选的视频编码之间的余弦相似度并排序, 以检索得到最相关的时刻。本文的主要创新点如下:

1) 本文提出一种快速 WSVMR 算法——FMAN。该算法绕过了测试阶段复杂的细粒度多模态融合, 实现了更高的性能并显著降低计算资源的消耗。

2) 本文设计了一种 MALM, 其可以灵活地应用于学习视频和文本之间的跨模态对齐, 实现视频内容理解, 并有效地将全部跨模态交互限制在训练阶段内。

3) 在 2 个广泛使用的 WSVMR 数据集上进行大量实验, 实验结果表明, 本文算法在检索性能和效率两方面都取得了显著的提高。

1　弱监督视频时刻检索相关技术

VMR 旨在从未经修剪的视频中使用给定的文本查询检索事件。作为多模态视频理解中的基本问题, 其在学术界和工业界都得到了高度关注, 并且大多数现有算法可以分为 2 类: 基于候选的算法和无候选的算法。基于候选的算法[1-2] 遵循将 VMR 转换为视频文件检索的原则, 首先, 生成大量的时刻提案作为候选, 然后, 测量视频片段和文本查询之间的相似度以选择最佳结果。无候选的算法[3-4] 则放弃了生成时序候选框, 而是侧重于通过回归或估计起始和结束时序分布, 从而直接预测得到目标时刻, 因此, 实现了更高的模型推理效率。

然而, 以上算法都需要大量的时间标注, 在实际多媒体检索应用中存在较大局限性。因此, 弱监督学习被广泛应用于减少时间标注成本。弱监督学习[13-16] 是机器学习中的学习范式概念, 用于训练具有部分或没有监督信号的模型, 可有效降低监督学习中存在的高额标注成本。通过直接采用不带有时刻标注的视频-文本数据训练模型, 可以极大的减轻数据标注的负担。同时, 弱监督学习也要求模型更多的侧重于视频内容与文本的对齐学习, 因此, 现有的 WSVMR 算法模型都采用多模态融合的方式来实现模态协同表征, 促进视频与文本在语义层级上的对齐。

融合多个模态以提取重要信息是多媒体智能研究的一个重要组成部分, 典型地应用包括视觉语言学习[17]、视听学习[18]、底层视觉[19-20] 及多媒体内容检索[21] 等领域。但是由于检索系统中用户查询文本的不可预测性, 多模态融合计算不得不在接收到用户查询文本后才能推进, 从而导致更多地计算动作需要在检索系统中在线进行, 带来冗余计算开销, 极大地限制多媒体检索算法在实际应用场景中的部署能力。如图 1(a) 所示, 现有的基于模态融合的 VMR 算法中大量的计算都必须在检索系统中在线进行, 对于海量的视频数据不得不按照串行的方式进行计算, 进而到导致实际检索效率低下。如图 1(b) 所示, 本文算法通过将视频内容和文本映射到一个公共的多模态空间实现模态对齐, 在模型推理阶段

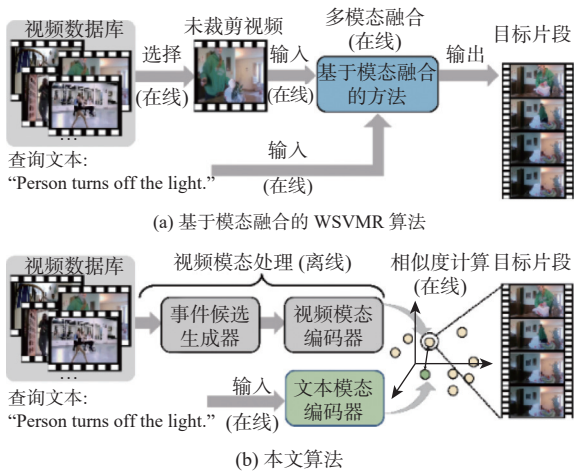


图 1　现有基于模态融合的算法与本文算法的示意图

Fig. 1　Diagrams of existing modal fusion-based algorithms and the proposed algorithm

绕开了繁琐的多模态融合动作，显著提升了模型部署后的检索效率。

2 本文算法

2.1 特征提取

本文算法的总体架构如图2所示，与现有算法类似，本文算法采用了离线特征提取算法。具体来说，给定未修剪的视频V，本文算法采用预训练的视觉骨干网络[22]来提取视觉特征，同时使用GloVe[23]来提取每个查询句子Q的单词级别嵌入。利用视频和文本模态的预训练特征提取器，可以得到提取的视频和文本特征序列 $\boldsymbol{F}_{\mathrm{v}}=\left[\boldsymbol{f}_{\mathrm{v}}^{i}\right]_{i=1}^{l_{\mathrm{v}}}$ 和 $\boldsymbol{F}_{\mathrm{q}}=\left[\boldsymbol{f}_{\mathrm{q}}^{j}\right]_{j=1}^{l_{\mathrm{Q}}}$，其中，$\boldsymbol{F}_{\mathrm{v}}\in\mathbf{R}^{l_{\mathrm{v}}\times d_{\mathrm{v}}}$ 和 $\boldsymbol{F}_{\mathrm{q}}\in\mathbf{R}^{l_{\mathrm{Q}}\times d_{\mathrm{q}}}$ 分别为提取的视频和文本特征。其中，d_{v} 和 d_{q} 分别为视频和文本模态的特征维度，l_{v} 和 l_{Q} 分别为视频特征的总长度和查询中单词的数量。在实现过程中，本文算法在2种模态的特征序列末尾填充了一个〈CLS〉标记，可分别表示为 $\boldsymbol{F}'_{\mathrm{v}}=\left[\boldsymbol{f}'_{\mathrm{v}},\boldsymbol{f}_{\mathrm{v}}^{\mathrm{cls}}\right]_{i=1}^{l_{\mathrm{v}}}$ 和 $\boldsymbol{F}'_{\mathrm{q}}=\left[\boldsymbol{f}_{\mathrm{q}}^{j},\boldsymbol{f}_{\mathrm{q}}^{\mathrm{cls}}\right]_{j=1}^{l_{\mathrm{Q}}}$。

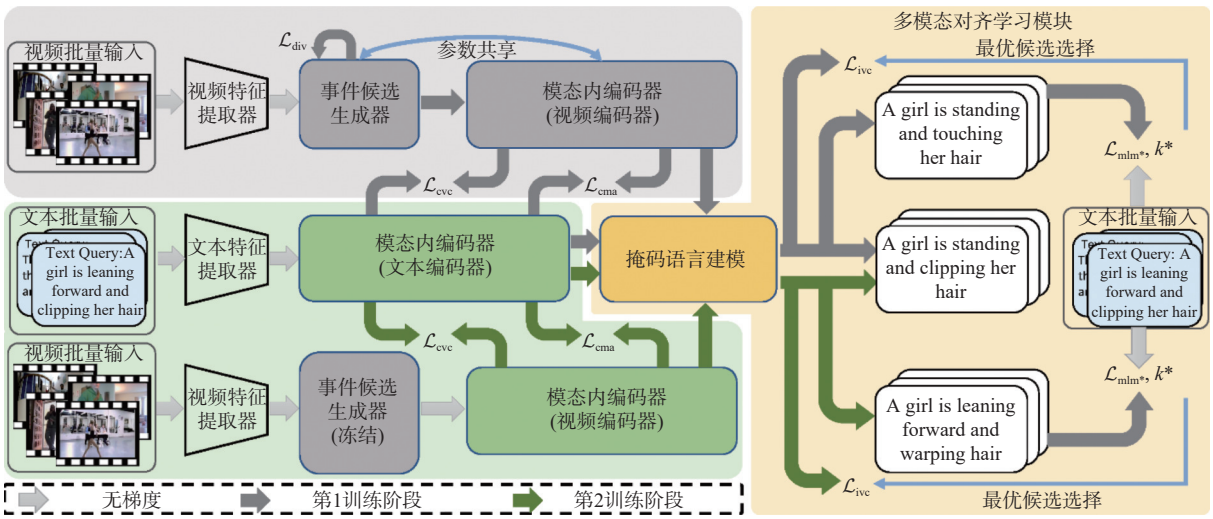


图2 本文算法的总体架构

Fig. 2 Overall architecture of the proposed algorithm

2.2 事件候选生成器

2.2.1 视频上下文编码

本文基于Transformer编码器[24]和多层感知机设计了一个ECG，学习与查询文本相关的潜在视频时序区域并生成多个候选片段：

$$\boldsymbol{H}_{\mathrm{e}}^{l+1}=\boldsymbol{\Phi}_{\mathrm{ECG}}^{l}\left(f_{\mathrm{e}}\left(\boldsymbol{H}_{\mathrm{e}}^{l}\right)\right) \tag{1}$$

式中：$\boldsymbol{\Phi}_{\mathrm{ECG}}^{l}$ 为第l个Transformer编码器层；其中，$\boldsymbol{H}_{\mathrm{e}}^{1}=\boldsymbol{F}'_{\mathrm{v}}$。$\boldsymbol{H}_{\mathrm{e}}^{l}=\left[\boldsymbol{h}_{\mathrm{e}}^{i}\right]_{i=1}^{l_{\mathrm{v}}+1},\boldsymbol{h}_{\mathrm{e}}^{j}\in\mathbf{R}^{d_{\mathrm{h}}}$ 为第l个编码器层的编码视频特征序列；$\boldsymbol{h}_{0}^{l_{\mathrm{v}}+1}$ 为视频的全局上下文编码；$f_{\mathrm{e}}(\cdot)$ 为ECG模块中的线性投影层；d_{h} 为隐藏特征维度。

2.2.2 候选时刻生成

基于得到的全局上下文编码，本文算法使用2个多层感知机预测目标事件的中心和宽度。设 $\mathrm{MLP}_{\mathrm{e}}^{\mathrm{c}}(\cdot)$ 和 $\mathrm{MLP}_{\mathrm{e}}^{\mathrm{w}}(\cdot)$ 为2个不同的多层感知机，则目标事件的中心和宽度可表示为

$$\begin{cases}\boldsymbol{c}=\delta\left(\mathrm{MLP}_{\mathrm{e}}^{\mathrm{c}}\left(\boldsymbol{h}_{\mathrm{e}}^{j+1}\right)\right)\\ \boldsymbol{w}=\delta\left(\mathrm{MLP}_{\mathrm{e}}^{\mathrm{w}}\left(\boldsymbol{h}_{\mathrm{e}}^{j+1}\right)\right)\end{cases} \tag{2}$$

式中：$\boldsymbol{c}=\left[c^{k}\right]_{k=1}^{K}$；$\boldsymbol{w}=\left[w^{k}\right]_{k=1}^{K}$；$\delta(\cdot)$ 为Sigmoid函数；K 为预测的候选数。

同时，为引导模型生成不同的候选时间段，本文算法引入了多样性损失和Zheng等[9]提出的基于高斯分布的掩码生成策略。给定第 k 个候选时间段的中心 c^{k} 和宽度 w^{k}，本文算法沿时间维度生成相应的高斯分布 $\boldsymbol{G}^{k}=\left[p_{i}^{k}\right]_{i=1}^{k}$，第 i 个视频特征的密度 p_{i}^{k} 可表示为

$$p_{i}^{k}=\frac{1}{\sqrt{2\pi}(w^{k}/\gamma)}\exp\left(-\frac{\left(i/l_{\mathrm{V}}-c^{k}\right)^{2}}{2(w^{k}/\gamma)^{2}}\right) \tag{3}$$

式中：γ 为一个缩放超参数。

将一个高斯分布掩码记作列向量，则多个高斯分布掩码可构成矩阵 $\boldsymbol{S}=\left[\boldsymbol{G}^{k}\right]_{k=1}^{K},\boldsymbol{S}\in\mathbf{R}^{Kl_{\mathrm{v}}}$，本文采用多样性损失以避免模型在训练ECG初期陷入退化解：

$$\mathcal{L}_{\mathrm{div}}=\left\|\boldsymbol{S}\boldsymbol{S}^{\mathrm{T}}-\epsilon\boldsymbol{I}\right\|_{\mathrm{F}}^{2} \tag{4}$$

式中：$\|\cdot\|_{\mathrm{F}}$ 和 $\boldsymbol{I}$ 分别为矩阵Frobenius范数运算和单位矩阵；ϵ 为控制不同候选分布之间重叠程度的超参数。

2.3 模态内编码器

2.3.1 文本编码器

文本编码器旨在学习自然语言查询文本的全局上下文编码。该模块采用多个Transformer编码

器堆叠实现。具体来说, 设Φ_{TE}^{l}为第l层文本 Transformer 编码器, 则文本模态的编码过程记为

$$\boldsymbol{H}_{\mathrm{q}}^{l+1}=\Phi_{\mathrm{TE}}^{l}\left(f_{\mathrm{q}}\left(\boldsymbol{H}_{\mathrm{q}}^{l}\right)\right) \tag{5}$$

式中: $\boldsymbol{H}_{\mathrm{q}}^{l}=\left[\boldsymbol{h}_{\mathrm{q}}^{j}\right]_{j=1}^{l_{\mathrm{Q}}+1}$, $\boldsymbol{h}_{\mathrm{q}}^{j}\in\mathbf{R}^{d_{\mathrm{h}}}$为第$l$个编码器层中的文本特征序列; $\boldsymbol{h}_{\mathrm{q}}^{l_{\mathrm{Q}}+1}$为文本全局上下文编码; $f_{\mathrm{q}}(\cdot)$为 IME 模块中文本编码器中的线性投影层。

2.3.2　视频编码器

不同于文本编码器, 视频编码器主要用于学习 ECG 模块生成的候选片段对应的局部上下文编码。具体来说, 该模块也是一个基于 Transformer 的架构。针对每个生成的候选片段, 视频编码器需要增强其对应时序区域中的上下文信息。具体来说, 给定 ECG 生成的高斯分布掩码矩阵 $\boldsymbol{S}=\left[\boldsymbol{G}^{k}\right]_{k=1}^{K}$。该模块将采用高斯分布掩码沿时间维度重新加权自注意力矩阵。假设Φ_{VE}^{l}是第l层视频 Transformer 编码器, 则上述过程可表示为

$$\boldsymbol{H}_{\mathrm{vk}}^{l+1}=\Phi_{\mathrm{VE}}^{L}\left(f_{\mathrm{v}}\left(\boldsymbol{H}_{\mathrm{vk}}^{l}\right),\boldsymbol{G}^{k}\right) \tag{6}$$

式中: $\boldsymbol{H}_{\mathrm{vk}}^{l}=\left[\boldsymbol{h}_{\mathrm{vk}}^{i}\right]_{i=1}^{l_{\mathrm{v}}+1}$, $\boldsymbol{h}_{\mathrm{vk}}^{i}\in\mathbf{R}^{d_{\mathrm{h}}}$为第$l$编码器层上第$k$个候选片段的编码视频特征序列; $f_{\mathrm{v}}(\cdot)$为线性投影层。对于自注意力矩阵重新加权, 每个编码器层中多头自注意力权重矩阵的第i行将和$\boldsymbol{G}^{k}$中每个元素计算点积, 实现重新加权。因此, $\boldsymbol{h}_{\mathrm{vk}}^{l_{\mathrm{v}}+1}$表示第$k$个候选片段的视频局部上下文编码。此外, 通过反转高斯分布的掩码或使用单位向量替换高斯分布的掩码, 本文算法还同步计算了每个候选片段对应的负采样局部上下文编码和全局上下文编码。

2.4　多模态对齐学习

2.4.1　最优候选选择

本文算法采用了掩码语言建模来从查询文本中学习深层次语义。具体来说, 其通过随机掩盖文本查询中的一部分关键词, 如特定的动词或名词, 并使用 Transformer 解码器根据输入句子的给定前缀和 2 种模态的上下文编码重构掩码词。设Ψ表示多个 Transformer 解码器层, 上述过程可表示为

$$\begin{cases}\boldsymbol{P}_{\mathrm{v}}^{k}=\sigma\left(f_{\mathrm{m}}\left(\Psi\left(\boldsymbol{h}_{\mathrm{vk}}^{l_{\mathrm{v}}+1},\hat{\boldsymbol{F}}_{\mathrm{q}}\right)\right)\right)\\ \boldsymbol{P}_{\mathrm{q}}=\sigma\left(f_{\mathrm{m}}\left(\Psi\left(\boldsymbol{h}_{\mathrm{q}}^{l_{\mathrm{Q}}+1},\hat{\boldsymbol{F}}_{\mathrm{q}}\right)\right)\right)\end{cases} \tag{7}$$

式中: $\sigma(\cdot)$和$f_{\mathrm{m}}(\cdot)$分别为 softmax 函数和线性投影层; $\hat{\boldsymbol{F}}_{\mathrm{q}}$为掩码后查询文本的特征, 其中掩码词被替换为特定的填充标记; $\boldsymbol{P}_{\mathrm{v}}^{k}$和$\boldsymbol{P}_{\mathrm{q}}$分别为用第$k$个视频局部上下文编码和文本全局上下文编码重构得到的单词预测概率, $\boldsymbol{P}_{\mathrm{v}},\boldsymbol{P}_{\mathrm{q}}\in\mathbf{R}^{l_{\mathrm{Q}}\times l_{\mathrm{W}}}$, l_{W}为词典的长度。

本文采用对数似然损失函数来评估第k个候选时刻:

$$\mathcal{L}_{\mathrm{mlm}}=-\sum_{j=1}^{l_{\mathrm{Q}}}\sum_{i=1}^{l_{\mathrm{W}}}\hat{\boldsymbol{P}}\left(\lg\boldsymbol{P}_{\mathrm{v}}+\lg\boldsymbol{P}_{\mathrm{q}}\right) \tag{8}$$

式中: $\hat{\boldsymbol{P}}$为真实标签的独热化向量。本文算法选择具有最小$\mathcal{L}_{\mathrm{mlm}}$的时刻候选作为最佳候选, 其编号和对应的掩码语言建模损失分别表示为k^{*}和$\mathcal{L}_{\mathrm{mlm}}^{*}$。类似地, 本文还计算出每个候选框对应的负采样视频局部上下文编码的掩码语言建模损失, 记作$\mathcal{L}_{\mathrm{mlm}}^{\prime}$。

基于上述 2 部分损失函数, 最优候选选择阶段的损失函数可表示为

$$\mathcal{L}_{\mathrm{opt}}=\mathcal{L}_{\mathrm{mlm}}^{*}+\mathrm{Max}\left(\mathcal{L}_{\mathrm{mlm}}^{*}-\mathcal{L}_{\mathrm{mlm}}^{\prime}+\alpha_{2},0\right) \tag{9}$$

式中: α_{2}为超参数。

2.4.2　跨模态对齐学习

本文算法将视频局部上下文编码和文本全局上下文编码投影到一个公共空间, 并通过度量学习来实现模态对齐。具体来说, 本文将$\boldsymbol{h}_{\mathrm{vk}^{*}}^{l_{\mathrm{v}}+1}$和$\boldsymbol{h}_{\mathrm{q}}^{l_{\mathrm{Q}}+1}$视为正样本对, 记作$T_{\mathrm{p}}$, 而$\boldsymbol{h}_{\mathrm{vk}^{*}}^{l_{\mathrm{v}}+1}$和$\boldsymbol{h}_{\mathrm{q}}^{l_{\mathrm{Q}}+1}$视为负样本对, 记作$T_{\mathrm{n}}$, 则跨模态对齐学习的训练目标可表示为

$$\mathcal{L}_{\mathrm{cma}}=\mathrm{Max}\left(\mathrm{Sim}\left(T_{\mathrm{p}}\right)-\mathrm{Sim}\left(T_{\mathrm{n}}\right)+\alpha_{1},0\right)+\varrho\left(T_{\mathrm{p}}\right) \tag{10}$$

式中: Sim(·) 和 $\varrho(\cdot)$分别为余弦相似度计算操作和均方误差损失; α_{1}为一个超参数。

2.4.3　跨视频对比学习

本文算法进一步在不同视频之间考虑对比学习, 以引导 IME 模块区分不同视频部分和文本查询之间的匹配关系。具体来说, 给定最优视频局部上下文编码$\boldsymbol{h}_{\mathrm{vk}^{*}}^{l_{\mathrm{v}}+1}$、视频全局上下文嵌入$\boldsymbol{h}_{\mathrm{g}}^{l_{\mathrm{v}}+1}$和文本上下文嵌入$\boldsymbol{h}_{\mathrm{q}}^{l_{\mathrm{Q}}+1}$, 本文在它们之间互相计算组合对比学习损失。该部分损失函数可以表示为

$$\begin{aligned}\mathcal{L}_{\mathrm{cvc}}=&\mathcal{L}_{\mathrm{ccl}}\left(\boldsymbol{h}_{\mathrm{vk}^{*}}^{l_{\mathrm{v}}+1},\boldsymbol{h}_{\mathrm{g}}^{l_{\mathrm{v}}+1}\right)+\mathcal{L}_{\mathrm{ccl}}\left(\boldsymbol{h}_{\mathrm{vk}^{*}}^{l_{\mathrm{v}}+1},\boldsymbol{h}_{\mathrm{q}}^{l_{\mathrm{Q}}+1}\right)+\\&\mathcal{L}_{\mathrm{ccl}}\left(\boldsymbol{h}_{\mathrm{q}}^{l_{\mathrm{Q}}+1},\boldsymbol{h}_{\mathrm{g}}^{l_{\mathrm{v}}+1}\right)\end{aligned} \tag{11}$$

式中: $\mathcal{L}_{\mathrm{ccl}}(\cdot)$为组合对比学习损失函数, 计算式为

$$\begin{aligned}\mathcal{L}_{\mathrm{ccl}}(x,y)=&\frac{1}{N}\sum_{x\in\mathcal{B}}\mathrm{Max}\left(S\left(x,y^{+}\right)-S\left(x,y^{-}\right)+\alpha_{3},0\right)+\\&\frac{1}{N}\sum_{x\in\mathcal{B}}\lg\left(\frac{S\left(x,y^{+}\right)}{S\left(x,y^{+}\right)+\sum\limits_{y^{-}\in\mathcal{B}}S\left(x,y^{-}\right)}\right)\end{aligned} \tag{12}$$

2.5　模型训练与推理

本文 FMAN 算法采用两阶段训练策略。具体来说, 即在第 1 阶段使用提出的 MALM 来训练 ECG 模块, 该阶段的损失函数可表示为

$$\mathcal{L}_{\mathrm{ecg}}=\mathcal{L}_{\mathrm{div}}+\mathcal{L}_{\mathrm{cma}}+\mathcal{L}_{\mathrm{opt}}+\mathcal{L}_{\mathrm{cve}} \tag{13}$$

在该阶段, ECG 模块的参数将同步用于 IME 模块中的视频编码器。因此, 反向传播将只需更新 ECG 模块。在第 2 阶段, ECG 模块的参数将被冻结且 MALM 模块的参数将被重置, 取消多样性损失

函数$\mathcal{L}_{\text{div}}$并训练IME模块。

在测试阶段，MALM模块将被禁用，只部署ECG和IME模块，直接计算每个视频局部上下文编码和文本上下文编码之间的余弦相似度，并对每个候选时刻进行排序。此外，由于在推理阶段无需任何的多模态融合，故所有视频建模都可以在离线环境中进行，对多媒体检索应用程序仅需要在IME中的文本编码器中进行实时推理。

3 实验与评估

3.1 实验设置

3.1.1 实验数据集

Charades-STA[1]数据集包含6 672个日常生活视频和16 128个视频-文本对，其中，训练集和测试集分别包含12 408个和3 720个视频-文本对。每个视频数据的平均持续时间为29.76 s。ActivityNet-Captions[25]数据集包含约20 000个来自互联网的开放域视频，平均每个视频包含3.65个查询，每个查询平均有13.48个单词。该数据集被划分为训练集、验证集和测试集，分别为37 421、17 505和17 031个视频-文本对。

3.1.2 模型实现细节

本文算法采用预训练的I3D骨干网络来提取视频特征，且每个视频的特征序列长度l_{v}被均匀采样为200。自然语言查询文本的最大长度l_0被设置为20。对于生成的基于高斯的掩码，在时序长度上的缩放超参数γ被设置为9。在ECG模块中，用于Charades-STA和ActivityNet-Captions数据集，生成的候选数分别为8和5。在训练期间，最优候选选择、跨模态对齐学习、跨视频对比学习3个损失函数中的度量学习超参数α_1、α_2和α_3分别设置为0.30、0.15和0.10。FMAN的两阶段训练策略依赖于单块Nvidia RTX 3090 GPU完成，批处理大小被设置为32，用Adam优化器[26]更新参数，学习率设置为4×10^{-4}。此外，本文算法对于Charades-STA和ActivityNet-Captions数据集分别对ECG模块进行了10次迭代和5次迭代的一阶段训练。在验证和测试阶段，本文算法采用召回率作为性能度量指标，记为R_n,IoU$=m$，其中，n为候选生成数，m为阈值。对于Charades-STA和ActivityNet-Captions数据集，本文算法分别将n设置为5，m设置分别为0.5、0.7和0.3、0.5，同时采用了该条件下的平均召回率(记作Avg.)评估模型综合性能。此外，本文算法采用了每秒浮点运算次数和模型参数量作为模型效率评估指标。

3.2 模型评估与对比

3.2.1 检索性能评估

本文FMAN算法和现有算法在时刻检索性能方面的比较如表1所示，表中，加黑数值表示性能最优。由表可知：①本文算法在大多数指标上优于

表1 在不同数据集上的检索性能对比

Table 1 Comparison of retrieval performance on different datasets

类型	算法	召回率/%											
		Charades-STA数据集						ActivityNet-Captions数据集					
		n=1, m=0.5	n=1, m=0.7	n=1, Avg.	n=5, m=0.5	n=5, m=0.7	n=5, Avg.	n=1, m=0.3	n=1, m=0.5	n=1, Avg.	n=5, m=0.3	n=5, m=0.5	n=5, Avg.
全监督	2DTAN[2]	39.70	23.31	31.51	80.32	51.26	65.79	59.45	44.51	51.98	85.53	77.13	81.33
	LGI[3]	59.46	35.48	47.47				58.52	41.51	50.02			
	SDN[4]	60.27	41.48	50.88				63.00	42.41	52.71			
点监督	ViGA[5]	45.05	20.27	32.66				59.61	35.79	47.70			
	D3G[27]	41.64	19.60	30.62	79.25	49.30	64.28	58.25	36.68	47.47	87.84	74.21	81.03
	CFMR[7]	48.14	22.58	35.36	80.06	56.09	68.08	59.97	36.97	48.47	82.68	69.28	75.98
弱监督	SCN[15]	23.58	9.97	16.78	71.80	38.87	55.34	47.23	29.22	38.23	71.56	55.69	63.63
	LCNet[10]	39.19	18.87	29.03	80.56	45.24	62.90	48.49	26.33	37.41	82.51	62.66	72.59
	DCCP[28]	29.80	11.90	20.85	77.20	32.20	54.70	41.60	23.20	32.40	61.40	41.70	51.55
	EVA[11]	40.21	18.22	29.22				49.89	29.43	39.66			
	ProTeGe[6]	31.84	17.51	24.68				45.02	27.85	36.44			
	CWG[12]	31.02	16.53	23.78	77.53	41.91	59.72	46.62	29.52	38.07	80.92	66.61	73.77
	CPL[9]	49.24	22.39	35.82	84.71	52.37	68.54	50.07	30.14	40.11	81.32	65.79	73.56
	FMAN	**51.40**	**25.05**	**38.23**	**86.29**	**53.93**	**70.11**	50.01	**30.58**	**40.30**	**83.75**	**70.48**	**77.12**

目前的最新 WSVMR 算法。例如, 与最新的 WSVMR 算法 CPL 相比, 本文 FMAN 算法在本文 FMAN 算法在 Charades-STA 数据集的 R_1, IoU=0.7 指标上取得了 2.66% 的绝对性能提升。同时, 其也在 Charades-STA 数据集和 ActivityNet-Captions 数据集的 R_5 指标方面取得了 1.57% 和 3.35% 的平均绝对性能提升, 但差距并不明显, 仅取得了 0.19% 的绝对性能提升。这是因为算法采用了可以将模态交互计算限制在训练阶段 MALM 模块, 其充分考虑了跨视频对比学习和视频内对比学习。②与全监督学习条件下训练得到的 2D-TAN 算法相比, 本文算法仍然取得了一定的优势。这是因为本文 FMAN 算法通过可学习的方式生成事件候选而非滑动窗口, 因此, 可以更加精细地对齐视频和文本 2 个模态。③本文算法在 ActivityNet-Captions 上 R_1 指标相较于现有算法取得了略高的平均性能, 但差距并不明显。一个潜在的原因是该数据集具有较多事件为完整视频, 弱化了事件候选生成模块的作用。因此, 在该数据集上, FMAN 算法在 R_5 上取得了显著提升, 而在 R_1 上优势并不明显。

3.2.2　检索效率评估

表 2 和表 3 分别为本文算法在 2 个数据集上的模型效率测试结果, 具体包括在线每秒浮点运算次数、模型参数量和 R_1～R_5 的平均召回率 (IoU 对于 Charades-STA 和 ActivityNet-Captions 数据集分别设置为 0.5 和 0.3), 表中, 加黑数值表示性能最优。为保证实验对比的公平性, 该部分实验省略了对比的现有算法中可以离线处理的组件, 仅比较在线运算开销。实验结果表明: ① 本文算法在效率上显著优于其他 WSVMR 算法。受多模态融合的限制, 大多数现有算法不得不将大量的模态融合处理计算部署于在线应用中, 等待用户输入文本。而本文算法只需要在离线视觉特征提取期间进行视频建模并生成多个候选时刻, 并存储于数据库中, 用户进行在线检索时仅需对文本编码器进行前向推理。因此, 本文算法节省了超过 99% 的在线每秒浮点运算次数, 并在具有较短的查询文本测试样例的 Charades-STA 数据集上, 使在线部署的模型参数减少了约 50%。②在 2 个数据集上的 R_1 比最新的最先进算法 CPL 分别提高了 1.56% 和 4.69%。这说明本文算法不仅显著减少了计算资源的消耗, 也有效地保持了检索性能。

表 2　在 Charades-STA 数据集上的检索效率对比

Table 2　Comparison of retrieval efficiency on Charades-STA dataset

模型名称	每秒浮点运算次数	模型参数量	R_1/%	R_5/%
2DTAN[2]	$4\,973.81\times10^8$	84.94×10^6	44.51	77.13
LGI[3]	56.94×10^8	47.21×10^6	41.64	
SCN[15]	11.01×10^8	7.01×10^6	29.22	55.69
CPL[9]	51.65×10^8	7.01×10^6	30.14	65.79
FMAN	$\mathbf{0.20\times10^8}$	$\mathbf{6.67\times10^6}$	**30.58**	**70.48**

表 3　Charades-STA 数据集上的消融实验

Table 3　Ablation study on Charades-STA dataset　%

$\mathcal{L}_{\text{cma}}$	$\mathcal{L}_{\text{ivc}}$	$\mathcal{L}_{\text{cvc}}$	R_1		R_5	
			m=0.5	m=0.7	m=0.5	m=0.7
×	×	×	46.82	20.66	82.06	50.48
√	×	×	48.41	21.52	82.92	50.70
√	√	×	48.87	22.17	84.40	51.67
√	×	√	50.86	23.09	86.29	53.04
√	√	√	**51.40**	**25.05**	**86.29**	**53.93**

3.3　定量分析与深入探究

3.3.1　多模态对齐学习分析

为进一步探究 MALM 模块的有效性, 本文对采用的 3 个重要损失函数 $\mathcal{L}_{\text{cma}}$、$\mathcal{L}_{\text{ivc}}$和 $\mathcal{L}_{\text{cvc}}$进行消融研究。根据表 3 中的实验结果, 可知: ①如果没有 $\mathcal{L}_{\text{cvc}}$, 模型在 R_1 和 R_5 指标上的性能降分别下降约 3% 和 2%。该结果证明了跨视频对比学习对于在弱监督学习环境下改进 VMR 起着至关重要的作用。② 在结合 $\mathcal{L}_{\text{ivc}}$和 $\mathcal{L}_{\text{cvc}}$的情况下, 模型的整体性能将得到进一步提升。这再次证明了视频间对比学习和跨视频对比学习可以有效地提升 WSVMR 模型的性能。这是由于对比学习中的大多数困难的负样本局部视频上下文样例来自相同的视频, 而视频间对比学习在训练我们的模型中恰恰起到了困难负样例挖掘的作用。

3.3.2　候选框生成分析

本文通过研究基于高斯掩码的注意力重加权机制和候选生成的数量, 进一步探索了 ECG 模块的有效性。具体来说, 本文在实验中设计了一种消融模型, 其采用二值化掩码替换原有模型的基于高斯分布的掩码。此外, 本文也采用了不同的候选者数量 K重新训练模型以观察模型性能变化。如图 3(a) 所示, 与完整的 FMAN 模型相比, 采用二值化掩码的消融模型性能下降明显。这种现象是由于基于高斯的掩码提供了一种软性的重加权机制, 使得深度学习模型在训练时梯度的反向传播更有效。其对于更好地学习局部视频上下文编码是必不可少的。此外, 如图 3(b) 所示, 候选数 K对模糊

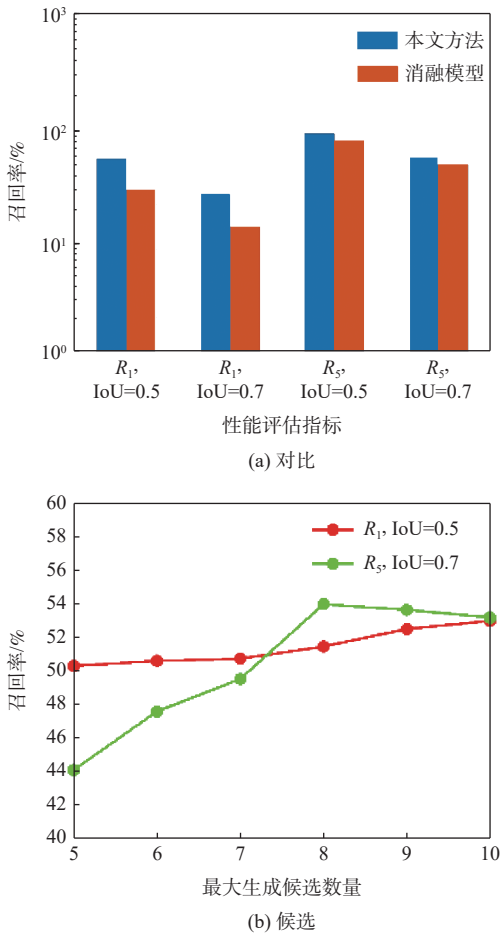


图3 候选生成模块的定量分析

Fig. 3 Quantitative analysis of candidate generation module

检索的 R_5 指标影响较大。这是由于生成更多候选将倾向于覆盖整个视频，但增加了模型筛选最佳候选的难度。

3.3.3 超参数敏感性分析

为分析模型对部分超参数的敏感性，此部分实验对高斯掩码时序缩放参数和视频采样长度进行了消融实验，通过调整上述2个参数的取值，在Charades-STA数据集上重新训练模型并观察模型在 R_1, IoU=0.5 和 R_5, IoU=0.5 这2项指标上的性能变化。实验中，时序缩放超参数和视频采样长度分别取值为 [11, 9, 7, 5] 和 [400, 300, 200, 100]。如图4所示，本文算法在上述2项参数同步增大或减小时可以保持较稳定的性能，反之仅调整其中一项超参数则会导致模型性能的显著下降。这是由于时序缩放超参数直接关系到高斯掩码在时间序列上的加权宽度，随着视频帧数量的增多，不匹配的缩放参数将使得高斯掩码无法准确的对应区域的局部视频内容，造成性能下降。

3.4 定性分析与结果可视化

3.4.1 候选框生成可视化

本文通过可视化ECG模块中基于高斯分布的掩码并按照对应的检索得分排序，对ECG模块的候选生成过程进行定量分析。由图4可知，生成的候选框与查询文本的语义高度相关，重叠的时间区域从 R_1 到 R_5 逐渐减少，与相似度得分呈正相关。这些结果再次证明本文算法有效学习到了视频和文本模态之间的语义一致性。此外，可以观察到ActivityNet-Captions中的一个典型案例 R_2 和 R_3 尽管与人工标注有较大出入，但仍然具有很高的相似度分数，且存在较明显的语义联系。这是由VMR中标记时间边界的主观性和不可靠性造成的，也进一步揭示了弱监督学习在该任务上的重要性。

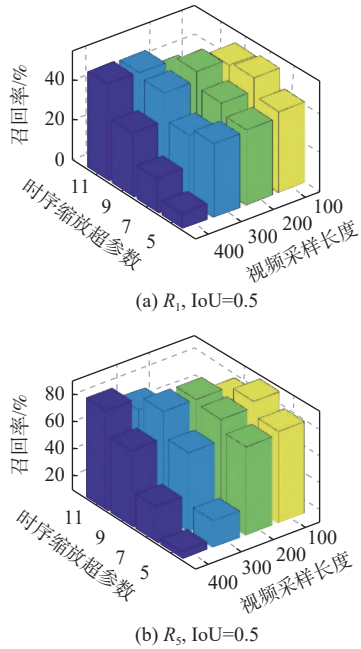


图4 超参数敏感性分析

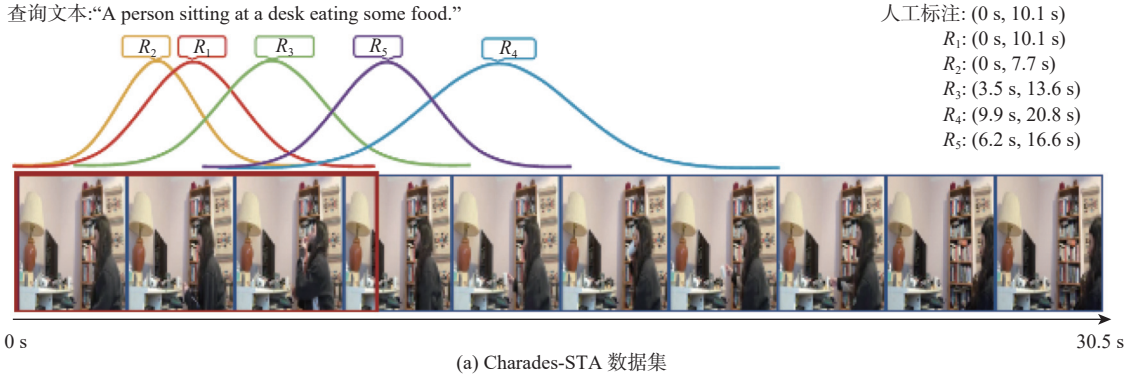
Fig. 4 Sensitivity analysis of hyperparameters

3.4.2 视频时刻检索可视化

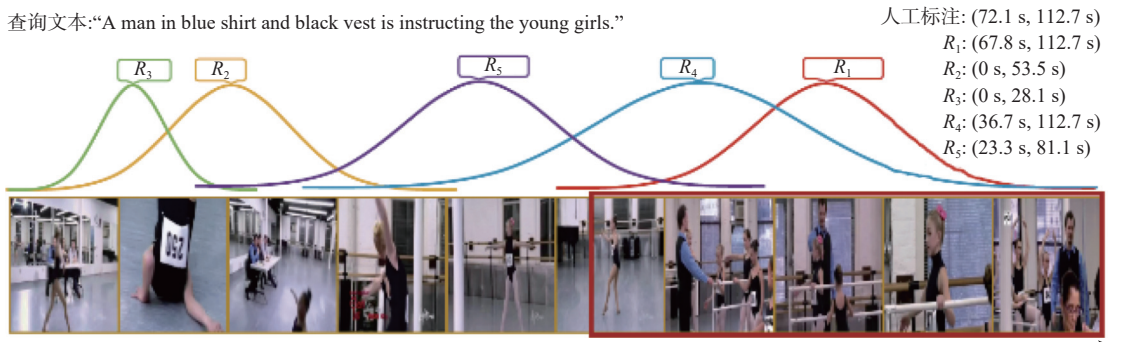
图5与图6分别为候选框生成结果的可视化情况和VMR结果示例，并与最新的算法CPL进行比较。由图可知，相较于现有算法，本文算法实现了更高的时刻检索精度。图6中本文FMAN算法精确定位了目标事件，而现有算法CPL则给出了错误的检索结果。这再次说明本文算法显着降低了多媒体检索应用计算成本，并有效地提高了检索事件的性能。

3.4.3 算法局限性分析

为进一步探究本文算法存在的局限性，图7展示了测试集中一个典型的检索失败案例。由图可知，该案例中本文算法尽管检索到了与查询文本相关的片段，但同时也包含了大量非目标片段内容。这是由于本文算法在对齐视频内容与查询文本时，仅在视频层面考虑了全局与局部的关系，并未对查询文本的语义层次进行深入探究。



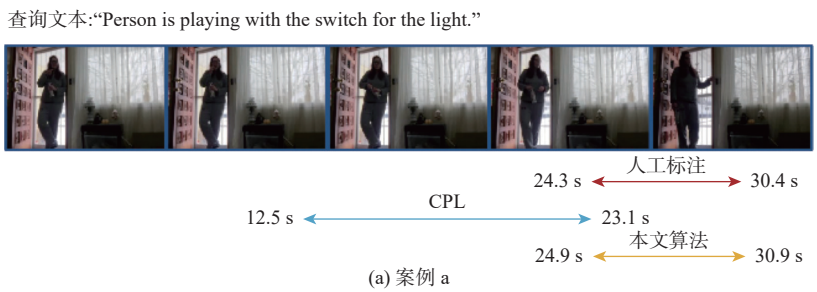
(a) Charades-STA 数据集



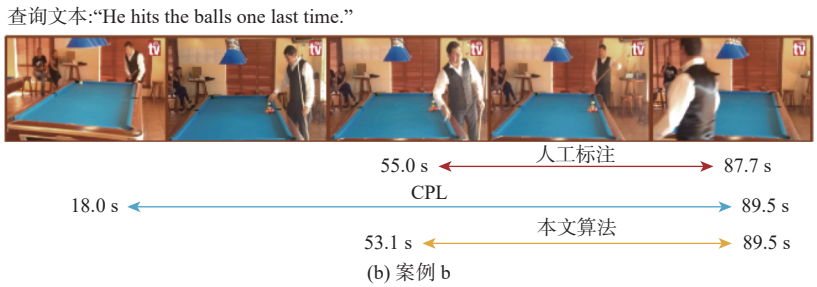
(b) Activitynet-captions 数据集

图 5　候选框生成结果可视化

Fig. 5　Visualization of candidate box generation



(a) 案例 a



(b) 案例 b

图 6　VMR 结果可视化

Fig. 6　Visualization of video moment retrieval results



图 7　检索失败案例可视化

Fig. 7　Visualization of unsuccessful retrieval case

在本案例中，查询文本由2个并列的事件组成，而本文算法最终给出的检索结果显然受到其中一部分对人物姿态描述的事件影响，从而给出了符合姿态描述语句的检索结果，导致结果的匹配度下降。

4 结 论

1）本文提出FMAN算法，以更好地解决WSVMR任务，提升其在应用场景下的实用性。

2）通过将视频内容和查询文本映射到公共空间并进行多模态对齐学习，本文算法取得了较好的VMR性能。

3）相较于基于多模态融合的现有算法，本文算法使得模型减少了超过99%的在线计算成本，实现了VMR效率的显著提升。

4）本文在2个广泛采用的WSVMR数据集上验证了本文算法的有效性。

本文FMAN算法主要针对单个视频中具体内容的检索，而该算法对于在视频数据库中进行视频文件检索是否仍然有效尚缺乏实验验证。在未来的工作中，将结合视频文件检索算法，实现针对大规模数据集的细粒度跨模态视频内容检索，以促进智能多媒体技术的发展。

参考文献（References）

[1] GAO J Y, SUN C, YANG Z H, et al. TALL: temporal activity localization via language query[C]//Proceedings of the IEEE International Conference on Computer Vision. Piscataway: IEEE Press, 2017: 5277-5285.

[2] ZHANG S Y, PENG H W, FU J L, et al. Learning 2D temporal adjacent networks for moment localization with natural language[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2020, 34(7): 12870-12877.

[3] MUN J, CHO M, HAN B. Local-global video-text interactions for temporal grounding[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2020: 10807-10816.

[4] JIANG X, XU X, ZHANG J R, et al. SDN: semantic decoupling network for temporal language grounding[J]. IEEE Transactions on Neural Networks and Learning Systems, 2024, 35(5): 6598-6612.

[5] CUI R, QIAN T W, PENG P, et al. Video moment retrieval from text queries *via* single frame annotation[C]//Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM, 2022: 1033-1043.

[6] WANG L, MITTAL G, SAJEEV S, et al. ProTéGé: untrimmed pretraining for video temporal grounding by video temporal grounding[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2023: 6575-6585.

[7] JIANG X, ZHOU Z L, XU X, et al. Faster video moment retrieval with point-level supervision[EB/OL]. (2023-01-23)[2023-02-01]. https://arxiv.org/abs/2305.14017v1.

[8] JI W, LIANG R J, ZHENG Z D, et al. Are binary annotations sufficient? video moment retrieval via hierarchical uncertainty-based active learning[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2023: 23013-23022.

[9] ZHENG M H, HUANG Y J, CHEN Q C, et al. Weakly supervised temporal sentence grounding with Gaussian-based contrastive proposal learning[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2022: 15534-15543.

[10] YANG W F, ZHANG T Z, ZHANG Y D, et al. Local correspondence network for weakly supervised temporal sentence grounding [J]. IEEE Transactions on Image Processing, 2021, 30: 3252-3262.

[11] CAI W T, HUANG J B, GONG S G. Hybrid-learning video moment retrieval across multi-domain labels[EB/OL]. (2022-12-23)[2023-02-01]. https://arxiv.org/abs/2406.01791.

[12] CHEN J M, LUO W X, ZHANG W, et al. Explore inter-contrast between videos *via* composition for weakly supervised temporal sentence grounding[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2022, 36(1): 267-275.

[13] WU Q, CHEN Y F, HUANG N, et al. Weakly-supervised cerebrovascular segmentation network with shape prior and model indicator[C]//Proceedings of the International Conference on Multimedia Retrieval. New York: ACM, 2022: 668-676.

[14] 丁家满, 刘楠, 周蜀杰, 等. 基于正则化的半监督弱标签分类算法[J]. 计算机学报, 2022, 45(1): 69-81.
DING J M, LIU N, ZHOU S J, et al. Semi-supervised weak-label classification method by regularization[J]. Chinese Journal of Computers, 2022, 45(1): 69-81(in Chinese).

[15] LIN Z J, ZHAO Z, ZHANG Z, et al. Weakly-supervised video moment retrieval *via* semantic completion network[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2020, 34(7): 11539-11546.

[16] JIANG X, XU X, ZHANG J R, et al. Semi-supervised video paragraph grounding with contrastive encoder[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2022: 2456-2465.

[17] XU X, WANG T, YANG Y, et al. Cross-modal attention with semantic consistence for image–text matching[J]. IEEE Transactions on Neural Networks and Learning Systems, 2020, 31(12): 5412-5425.

[18] JIANG X, XU X, CHEN Z G, et al. DHHN: dual hierarchical hybrid network for weakly-supervised audio-visual video parsing[C]//Proceedings of the 30th ACM International Conference on Multimedia. New York: ACM, 2022: 719-727.

[19] 刘耀胜, 廖育荣, 林存宝, 等. 基于特征融合与抗遮挡的卫星视频目标跟踪算法[J]. 北京航空航天大学学报, 2022, 48(12): 2537-2547.
LIU Y S, LIAO Y R, LIN C B, et al. Feature-fusion and anti-occlusion based target tracking method for satellite videos[J]. Journal of Beijing University of Aeronautics and Astronautics, 2022, 48(12): 2537-2547(in Chinese).

[20] 张燕咏, 张莎, 张昱, 等. 基于多模态融合的自动驾驶感知及计算

[J]. 计算机研究与发展, 2020, 57(9): 1781-1799.
ZHANG Y Y, ZHANG S, ZHANG Y, et al. Multi-modality fusion perception and computing in autonomous driving[J]. Journal of Computer Research and Development, 2020, 57(9): 1781-1799(in Chinese).

[21] JIANG X, XU X, ZHANG J R, et al. GTLR: graph-based transformer with language reconstruction for video paragraph grounding[C]//Proceedings of the IEEE International Conference on Multimedia and Expo. Piscataway: IEEE Press, 2022: 1-6.

[22] CARREIRA J, ZISSERMAN A. Quo vadis, action recognition? a new model and the kinetics dataset[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2017: 4724-4733.

[23] PENNINGTON J, SOCHER R, MANNING C. Glove: global vectors for word representation[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing. Stroudsburg: USAACL, 2014: 1532-1543.

[24] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[EB/OL]. (2017-06-12)[2023-02-01]. https://arxiv.org/abs/1706.03762.

[25] KRISHNA R, KENJI H T, REN F, et al. Dense-captioning events in videos[C]//Proceedings of the IEEE International Conference on Computer Vision. Piscataway: IEEE Press, 2017: 706-715.

[26] KINGMA D P, BA J, HAMMAD M M. Adam: a method for stochastic optimization[EB/OL]. (2017-01-30)[2023-02-01]. https://arxiv.org/abs/1412.6980v9.

[27] LI H J, SHU X J, HE S N, et al. D3G: exploring Gaussian prior for temporal sentence grounding with glance annotation[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE Press, 2023: 13688-13700.

[28] MA F, ZHU L C, YANG Y. Weakly supervised moment localization with decoupled consistent concept prediction[J]. International Journal of Computer Vision, 2022, 130(5): 1244-1258.

Efficient weakly-supervised video moment retrieval algorithm without multimodal fusion

JIANG Xun，XU Xing*，SHEN Fumin，WANG Guoqing，YANG Yang

(School of Computer Science and Engineering，University of Electronic Science and Technology of China，Chengdu 611731，China)

Abstract：The weakly-supervised video moment retrieval (WSVMR) task retrieves in and out points of specific events from untrimmed videos based on natural language query text, using a deep learning algorithm model trained through video and natural language text matching relationships. Most existing WSVMR algorithms use multimodal fusion mechanisms to understand video content for moment retrieval. However, the cross-modal interactions required for effective fusion are quite complex, and this process can only be initiated once a clear user query is received. These hinder the operational efficiency of these algorithms, limiting their deployment in multimedia applications. To address this issue, a novel fusion-free multimodal alignment network (FMAN) algorithm for achieving rapid WSVMR was proposed. By restricting complex cross-modal interaction computation in the training phase, this algorithm enables offline encoding of both video and text data by the model, thereby significantly improving the inference speed of video moment retrieval. Experimental results on Charades-STA and ActivityNet-Captions datasets show FMAN's superior retrieval performance and efficiency compared to existing methods. The experiment on retrieval performance metrics, specifically recall rates (R_1 and R_5), shows an average improvement of 2.66% and 1.57% on Charades-STA, as well as 0.19% and 3.35% on ActivityNet-Captions, respectively. Additionally, the proposed algorithm reduces online floating-point operations per second to no more than 1% of the original algorithm.

Keywords：video moment retrieval; weakly-supervised learning; media content retrieval; model efficiency; cross-modal alignment

Received：2023-06-16；**Accepted**：2023-09-15；**Published Online**：2023-11-28 14：11

URL：link.cnki.net/urlid/11.2625.V.20231128.0939.001

Foundation items：National Natural Science Foundation of China (61976049,62072080)

* **Corresponding author**. E-mail：xing.xu@uestc.edu.cn

http://bhxb.buaa.edu.cn jbuaa@buaa.edu.cn
DOI: 10.13700/j.bh.1001-5965.2024.0747

面向红外弱小舰船检测的轻量化神经网络设计

唐文婷[1]，李波[2]，季梦奇[2, 3, 4, *]

(1. 北京航空航天大学 计算机学院，北京 100191；2. 北京航空航天大学 人工智能学院，北京 100191；
3. 杭州市北京航空航天大学国际创新研究院（北京航空航天大学国际创新学院），杭州 311115；
4. 信息系统工程全国重点实验室，长沙 410003)

摘 要：为高效提取红外遥感图像中弱小舰船的深度特征，提出一种轻量化骨干网络设计方法。受视觉注意力驱动的感受野调节机制启发，提出包含多尺寸感受野感知与选择过程的视觉感受野调节机制模拟方法，提高红外弱小舰船目标的表征效果；结合特征复用与卷积核分解的设计思想优化了多尺寸感受野模拟过程，实现轻量特征选择算子模拟多尺寸感受野选择过程，进一步降低网络的运算开销。在红外弱小舰船检测数据集上的实验结果表明：该网络检测精度提高了2%，且相较通用轻量化网络参数量减少 2.3×10^6，计算量降低 9.1 GFLOPs 次；在存在相似地物干扰的港口及离岸复杂场景下，所提方法有效降低了虚警，并抑制了漏检。

关 键 词：小目标检测；卫星遥感图像；感受野；神经网络设计；轻量化神经网络
中图分类号：V474.2+7; V474.2+92; TP753
文献标志码：A **文章编号**：1001-5965(2025)07-2394-10

利用红外遥感成像技术对海事活动进行监测，已成为侦察卫星与多用途卫星应用的一个重要领域。该技术能够通过捕捉目标的热辐射特征，实现对海上船舶及其他活动的精确探测与实时监控，对于维护海上交通安全、打击海上非法行为以及提升海洋资源管理效率等方面发挥着关键作用[1]。然而，红外成像技术的空间分辨率较低，且星载平台可用运算资源有限，即便部署通用轻量深度神经网络也难以高效检测红外弱小舰船目标[2]。一方面，红外成像技术的空间分辨率低于可见光成像技术的空间分辨率，从而导致舰船目标成像尺寸小且细节模糊，面向可见光设计的通用深度神经网络难以精准表征弱小红外舰船；另一方面，星载平台载荷有限，难以支持面向 GPU 部署所设计的通用轻量神经网络推理与训练所需的能耗开销[3]。因此，研究面向红外弱小舰船检测的轻量级深度神经网络对智能化海事监测具有重要的实际意义。

优化骨干网络或目标检测算法可提升表征质量与运算性能，但兼顾目标检测算法的实时性与泛化性是遥感图像处理领域的研究热点与应用难点[3-9]。骨干网络是指深度学习算法中进行特征提取的网络模型。以 YOLOv5 检测算法为例[10]，在推理过程中，骨干网络将占据近 70% 的运算开销，因此，骨干网络的优化方式决定目标检测算法的泛化性与实时性。为实现无人机航拍图像小目标精准检测，文献 [4] 采用 Transformer 与可形变卷积进行骨干网络优化，在提升小目标表征效果的同时带来了近 10×10^6 的参数量上升；为精准检测光学遥感舰船关键部位，文献 [5] 在主干网络引入全局上下文模块以联合舰船全局信息进行关键部位推理，然而该检测算法的参数量约 32×10^6，不适用于运算资源极其紧张的星载平台。目标检测算法包括骨干网

收稿日期：2024-10-16；**录用日期**：2024-11-22；**网络出版时间**：2025-02-26 16:39
网络出版地址：link.cnki.net/urlid/11.2625.V.20250226.1305.002
基金项目：国家自然科学基金 (62171253)； 青年人才托举工程 (2022QNRC001)；信息系统工程全国重点实验室稳定支持资助项目 (WDZC20255290411)
* **通信作者**. E-mail：jimengqi@buaa.edu.cn

引用格式：唐文婷，李波，季梦奇. 面向红外弱小舰船检测的轻量化神经网络设计 [J]. 北京航空航天大学学报，2025，51（7）：2394-2403.
TANG W T，LI B，JI M Q. Lightweight neural network design for infrared small ship detection [J]. Journal of Beijing University of Aeronautics and Astronautics，2025，51（7）：2394-2403 （in Chinese）.

络、特征融合网络与末端检测网络(检测头)3 部分, 通过优化目标函数或特征融合策略同样能提升表征质量。文献 [3] 重新设计了特征金字塔网络与末端检测网络以提升机场区域遥感图像的检测精确率; 文献 [9] 提出注意力引导网络与基于多样本决策的最大似然损失函数以提升目标检测算法的 SAR 舰船定位能力。然而, 这类算法参数量大, 不适宜星载红外舰船检测应用场景。

借鉴轻量化深度神经网络模块设计思想与视觉注意力驱动的感受野调节机制, 可优化红外弱小舰船目标的表征效果。首先, 特征复用的网络模块设计思想可以更低的参数量实现更优的表征效果。特征复用是通过利用特征提取过程中的中间计算结果, 以多分支的特征提取形式实现以更低的参数量逼近高参数量网络模块的表征效果[11-12]。以包含两次卷积的幽灵卷积模块为例, 该模块中第 2 个卷积的输入为第 1 次卷积的输出, 从而实现以更低的参数量实现特征处理过程中的感受野扩张。已有红外图像处理技术表明, 大尺寸的感受野意味着大尺寸的滤波核, 大尺寸滤波核对于表征红外弱小目标至关重要[12]。其次, 利用通道级注意力机制可进一步实现特征选择。通道级注意力机制是指通过重新分配通道权重以实现通道级特征增强。“挤压-激励”(squeeze and excitation, SE)模块是常用的通道级注意力机制[13], 该模块根据空间维度降采样后特征进行通道注意力预测以完成通道权重估计。特征选择是剔除冗余、低效特征的过程。由于预测通道权重在 (0,1) 之间, 因此其可作为通道压缩的线索以完成特征降维, 从而实现重要特征的筛选。最后, 考虑到注意力驱动的感受野调节机制有助于生物快速识别复杂场景下弱小目标[14], 提出在轻量化网络设计时也应遵循感受野缩放这一设计思路。复杂场景会导致视觉拥挤效应(由目标周围物体存在导致辨别目标困难的现象), 而生物视觉皮层区通过降低群感受野大小降低视觉拥挤效应对识别的影响[14]。在复杂场景下, 红外目标的信噪比较低, 导致在目标周围存在复杂地物时其识别难度进一步增加。因此, 理论感受野尺寸较大的深度神经网络推理复杂场景下的弱小红外目标时, 视觉拥挤效应同样存在。此时, 缩放感受野尺寸至关重要。

基于以上分析, 本文以 YOLOv5 检测算法为检测框架, 其轻量化骨干网络模型 YOLOv5-N 与 YOLOv5-S 作为网络架构的深度与宽度基准, 通过重构骨干网络的特征提取模块实现骨干网络建模与表征优化。将常见轻量化网络模块(CSP 瓶颈模块[10]、嵌入空间特征金字塔的 CSP 瓶颈模块[15] 与幽灵卷积模块[11])作为对比方法, 通过对红外舰船检测数据集 (infrared ship detection dataset, ISDD)[16] 进行弱小舰船目标检测验证并分析本文方法的有效性与实用性。主要工作及创新点如下:

1) 为抑制视觉拥挤效应, 提出了一种视觉感受野调节机制模拟方法以提升红外舰船目标的表征效果。

2) 为实现网络感受野动态缩放, 提出了一种轻量化骨干网络模块重构方法, 应用了特征复用与卷积核分解技术降低各尺度特征提取模块的运算开销。与主流轻量化网络模块相比, 本文方法可以更低的参数量实现相当甚至更优的检测精度。

3) 为实现注意力聚焦, 本文方法设计了特征选择算子主动抑制低效特征。通过特征质量分析实验, 直观展示了本文方法提出的网络注意力选择与抑制过程。

1　本文方法

为有效降低深度神经网络复杂度并满足星载红外舰船检测应用场景的部署需求, 本文提出了一种面向红外弱小舰船检测的轻量化神经网络设计方法。受视觉注意力驱动的感受野调节机制启发, 本文重构了 YOLOv5 检测算法中骨干网络的茎层与特征提取模块, 使骨干网络在特征提取过程中具备感受野缩放能力。视觉拥挤效应下的生物感受野调节机制与本文方法框架如图 1 所示。其中, 卷积算子 CONV 括号后内容为滤波器的长、宽和步长, 基于 YOLOv5 检测算法的骨干网络中的 3, 6, 9, 3 为各个尺度采用的特征提取模块数量。MRF-STEM (multi-receptive field stem), 即通过多尺度(多感受野)卷积操作来增强模型对不同尺度特征的感知能力, DMRF(depthwise multi-receptive field), 即基于深度可分离卷积的多感受野结构, 结合了深度卷积和多感受野设计。

首先, 设计一种视觉感受野调节机制模拟方法, 提升骨干网络对弱小红外目标的表征效果; 其次, 基于该模拟方法, 提出了一种轻量化骨干网络模块重构方法, 针对性地优化红外图像与深度特征表征过程; 然后, 设计了轻量特征选择算子主动抑制低效特征, 进一步降低表征开销; 最后, 通过嵌入重构后模块至目标检测骨干网络实现红外弱小目标的精准检测。

1.1　视觉感受野调节机制模拟方法

当发生视觉拥挤效应时, 人脑视觉皮层(secondary visual cortex, V2)通过降低群感受野尺寸识别目标。群感受野(也称感受野)是指视觉系统中神经元对视觉刺激空间区域的敏感区域。如图 1 所示, 红外遥感场景下, 当停泊船只周围的地物建筑复杂时, V2 主动调节处理该区域的感受野尺寸, 从

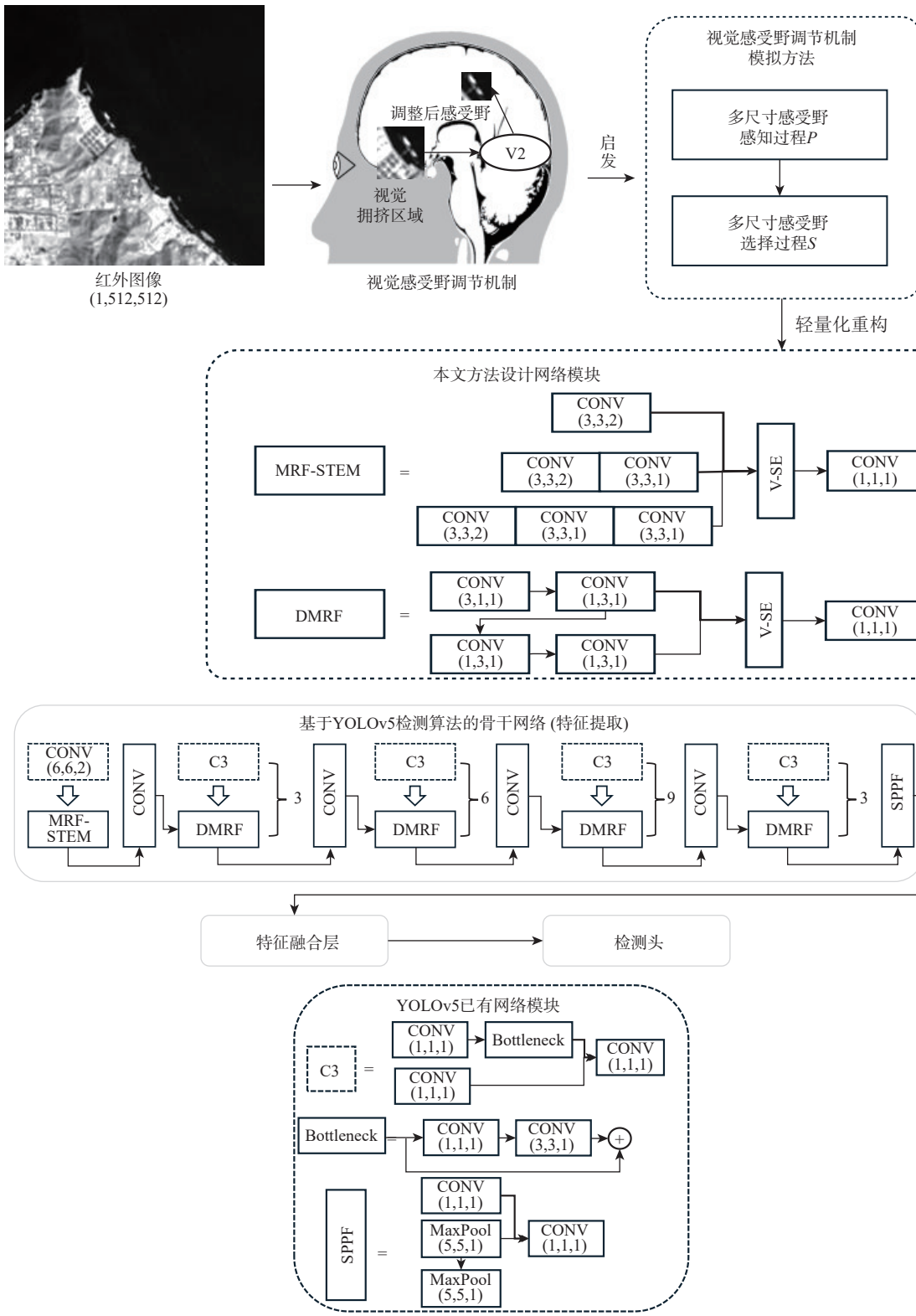


图1 感受野调节机制与本文方法框架

Fig. 1 Receptive field adjustment mechanism and the proposed algorithm framework

而达到注意力聚焦于目标的效果。然而,骨干网络中感受野尺寸固定,无法实现不同尺寸感受野之间的选择。因此,人工神经网络无法像视觉神经系统一样,通过主动调节感受野尺寸实现注意力聚焦。为此,本文提出视觉感受野机制模拟方法以优化复杂场景下红外弱小目标深度特征表征方式。

具体而言,对于高为H、宽为W、维度为C且像素值为实数的输入特征或图像$\boldsymbol{F}\in\mathbf{R}^{H\times W\times C}$,视觉感受野调节机制可简化为多尺寸感受野感知过程$P$与多尺寸感受野选择过程$S$:

$$\begin{cases}\boldsymbol{F}_P=P_\theta(\boldsymbol{F})\\ \boldsymbol{F}_S=S_\rho(\boldsymbol{F}_P)\end{cases}\tag{1}$$

式中: θ为模拟多尺寸感受野感知过程的参数; $\boldsymbol{F}_P$为该感知过程提取的n种感受野特征, 即$\boldsymbol{F}_P=\{\hat{\boldsymbol{F}}_{p_i\times q_i}|i=1,\cdots,n\}$, $\hat{\boldsymbol{F}}_{p_i\times q_i}$为第$i$个感受野区域特征, q_i与p_i分别为其区域宽与高; ρ为建模多尺寸感受野选择规则所需参数, 聚焦特征$\boldsymbol{F}_S\in\mathbf{R}^{H\times W\times n'}$, $n'<n$为聚焦特征维度, 且第i维聚焦特征$\boldsymbol{F}_S^i$的计算方式为

$$\boldsymbol{F}_S^i=\begin{cases}w_i\hat{\boldsymbol{F}}_{p_i\times q_i} & w_i\neq 0\\ \varnothing & w_i=0\end{cases} \tag{2}$$

其中: $w_i\in[0,1]$为$\boldsymbol{F}_S$中第i个感受野的调节系数。在容易引起视觉拥挤效应的复杂场景下, 通过抑制调节系数为0的特征实现感受野选择。

为实现人工神经网络的动态聚焦, 根据视觉感受野调节机制模拟方法进行网络模块设计。如图 2(a) 所示, 该网络模块包含多尺寸感受野特征提取与选择 2 个子模块。多尺寸感受野特征提取模块建模了 P 过程并输出$\boldsymbol{F}_P$。其中, 第i个感受野特征的获取方式为

$$\hat{\boldsymbol{F}}_{p_i\times q_i}=\mathrm{CONV}(p_i,q_i)(\boldsymbol{F}) \tag{3}$$

式中: $\mathrm{CONV}(p_i,q_i)$表示以高、宽分别为p_i、q_i的滤波核初始化的二维卷积算子。

多尺寸感受野特征选择模块建模了 S 并输出$\boldsymbol{F}_S$。其中, 采用多层感知机 MLP 预测各感受野调节系数$\omega=\{w_i|i=1,\cdots,n\}$的方式为

$$\begin{cases}\omega=\mathrm{MLP}(\boldsymbol{F}',2)\\ \boldsymbol{F}'=\mathrm{reshape}(\boldsymbol{F}_P,<H\times W,n>)\\ \mathrm{MLP}(\boldsymbol{F}',i)=\begin{cases}\mathrm{LP}(H\times W,n)(\boldsymbol{F}') & i=1\\ \mathrm{LP}(n,n)(\mathrm{MLP}(\boldsymbol{F}',i-1)) & i>1\end{cases}\\ \mathrm{LP}(a,b)=\mathrm{act}[\mathrm{Linear}(a,b)]\end{cases} \tag{4}$$

式中: reshape()表示张量维度变换操作; act()表示进行非线性变换的激励函数; $\mathrm{Linear}(a,b)$表示输入通道为a、输出通道为b的线性变换。

1.2　轻量化骨干网络重构方法

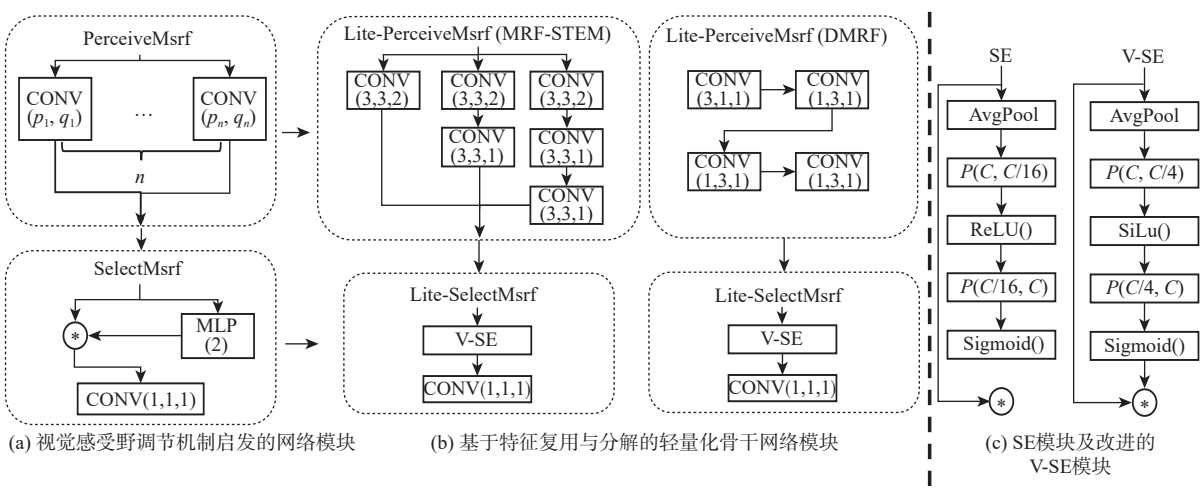
星载平台存在运算资源瓶颈, 直接模拟视觉感受野调节机制中多尺寸感受野感知过程开销较高。处理高为H、宽为W且维度为C的输入特征或图像时, 提取各尺度感受野的二维卷积算子$\mathrm{CONV}(p_i,q_i)$的存储开销$S(\theta)$与计算开销$T(\theta)$为

$$\begin{cases}S(\theta)=C\times 1\times p_i\times q_i\\ T(\theta)=H\times W\times C\times p_i\times q_i\end{cases} \tag{5}$$

因此, 模拟多尺度感受野感知过程 P 时空开销满足: $S(\theta)\propto C,p_i,q_i$, $T(\theta)\propto H,W,C,p_i,q_i$。当感受野尺寸、处理图像或特征的长、宽与高增加时, 模拟 P 过程的时空开销呈指数级增长。

本文提出通过轻量化重构骨干网络中的茎层与特征提取模块, 实现多尺寸感知过程的高效模拟。轻量化重构骨干网络是指通过重新设计或调整网络模块结构, 降低骨干网络的参数量与计算量。骨干网络中, 茎层可直接提取未经下采样的图像特征, 是一种特殊的特征提取模块。由于PerceiveMsrf过程的存储和时间开销与图像尺寸成正比, 且图像的宽高远大于图像波段数(红外图像波段数为 1), 故茎层在模拟 P 过程时应限制感受野尺寸的种类。如图 2(b) 所示, 将茎层感受野尺寸的种类限制为 3 种, 而各深度特征的感受野尺寸设计为 2 种, 且各感受野的长与宽相同。

在此基础上, 基于特征复用与卷积核分解的网络设计思想进一步降低茎层与深度特征提取过程模拟 P 过程的开销。特征复用的设计思想是指在神经网络的不同层之间共享和重用特征以提升网络性能。已知输入特征$\boldsymbol{F}\in\mathbf{R}^{H\times W\times C}$, 输出特征$\boldsymbol{F}_P\in\mathbf{R}^{H\times W\times\left(\sum_{i=1}^{n}\phi_i\right)}$, ϕ_i为第i种感受野尺寸的特征数量, 假设要求$\boldsymbol{F}_P$输出特征的感受野尺寸$R_{\mathrm{F}}=\{r_{fi}|i=$



(a) 视觉感受野调节机制启发的网络模块　(b) 基于特征复用与分解的轻量化骨干网络模块　(c) SE模块及改进的V-SE模块

图 2　本文方法网络模块结构

Fig. 2　Architecture of the modules of the proposed method

$1,\cdots,n\}$，则特征复用的设计方法对各输出感受野尺寸进行如下约束：

$$r_{f_i}=\begin{cases}(p_i-1)\prod_{i=1}^{n}s_i+\xi & i=1\\ r_{i-1}+(p_i-1)\prod_{i=1}^{n}s_i & i>1\end{cases} \tag{6}$$

式中：s_i为第i种感受野的尺度；p_i为第i种感受野的宽或高；ξ为该模块的已有感受野尺寸，若该模块为第1个DMRF模块，则$\xi=12$，若该模块为MRF-STEM，则$\xi=0$。

如图2(b)中DMRF网络模块所示，感受野尺寸为r_2的特征通过重用感受野尺寸为r_1的特征实现感受野高效扩张。卷积核分解是指通过将大尺寸的卷积核分解为一系列小的卷积核的组合设计思想。已知分解前卷积核的感受野尺寸r_i，分解后小卷积核的卷积核宽k，则分解次数D_i为

$$D_i=r_i-\xi/(k-1)\prod_{i=1}^{n}s_i \tag{7}$$

如MRF-STEM模块中通过最多进行3次分解获得感受野尺寸为12的特征，而DMRF模块最多通过4次分解获得感受野尺寸为$4\prod_{i=1}^{n}s_i+\xi$。此时，提取各感受野的卷积运算期望开销均下降$k/2$倍。

1.3 轻量特征选择模块设计

采用多层感知机直接模拟多尺寸感受野特征选择开销较高。多层感知机MLP($\boldsymbol{F}'$,i)的存储与时间开销$S(\rho)$、$T(\rho)$为

$$\begin{cases}S(\rho)=H\times W\times C'\\ T(\rho)=H\times W\times C'\\ C'=\sum_{i=1}^{n}\phi_i\end{cases} \tag{8}$$

因此，当处理图像或特征的长、宽与高增加时，预测各感受野调节系数的开销飙升。

使用SE模块可高效预测感受野调节系数，但其具有一定的决策局限性。如图2(c)所示，SE模块通过平均池化算子对输入特征降维AvgPool：$\mathbf{R}^{H\times W\times C'}\rightarrow\mathbf{R}^{C'}$。经过降维，使用多层感知机的存储与时间开销变为

$$\begin{cases}S(\rho')=C'\times C'\\ T(\rho')=C'\times C'\end{cases} \tag{9}$$

式中：ρ'为轻量化后建模多尺寸感受野选择规则所需参数。此时，预测各感受野调节系数的时空开销仅与特征的高C'相关，存储与时间开销大幅缩减。然而，SE模块使用ReLU算子作为激活函数，会过度抑制负值特征。卷积神经网络中，茎层与低尺度特征提取模块的卷积算子与Gabor变换功能相似，均对图像中纹理信息进行深入分析[17]。然而，经过Gabor变换后获得的负特征包含相位信息，表征了图像中的特定纹理或结构。因此，以ReLU算子作为激活函数的SE模块在预测感受野调节系数时具有一定的局限性。

为解决这一问题，本文设计了改进的V-SE模块。如图2(c)所示，V-SE模块将ReLU算子替换为了SiLu算子，缓解了SE模块对于负特征的抑制效果。此外，为了提升预测感受野调节系数的准确性，将SE模块的挤压因子从16调整至4，以轻微时空开销为代价提升多层感知机的拟合能力。具体地，挤压因子调整前存储和计算开销均为$C'^2/8$，调整后开销为$C'^2/2$，仍远低于采用多层感知机直接模拟所需开销。

2 实验结果及分析

2.1 数据集和评价指标

为验证本文方法在包含弱小目标的红外遥感场景的处理效果，选择使用ISDD[16]数据集进行训练和验证。ISDD[16]数据集包含1 284张涵盖拖尾、近岸、靠岸、海浪和多云等真实复杂场景，共包含3 061个弱小舰船实例。

为度量检测正确性，采用交并比(intersection of union, IOU)等于0.3为基准计算准确率与召回率，并进一步使用AP-0.5(取IOU≥0.5时计算的平均精度)、AP-0.5:0.95(将真实框的IOU∈[0.5,0.95]的检测框依次看作正确预测时其正确预测占比期望)评估检测算法的预测准确性；为评估模型复杂度，采用参数量(模型权重中非零元素个数)与浮点运算操作数评估模型的存储复杂度与计算复杂度。

2.2 实验平台及参数设置

采用戴尔XPS 8930台式机实现各方法的训练与推理。该台式机搭载英特尔第8代酷睿i7-8700K处理器与GTX1080显卡(8GB显存)。各方法训练时，初始学习率为1×10^{-4}，优化算法为Adam，批大小为32，训练轮数为300轮。所有网络的输入图像尺寸为512×512。在图像预处理过程中，将输入图像沿波段维度复制3次以避免对比方法对茎层的修改，而本文方法直接采用单波段红外图像作为输入。为验证本文方法在不同模型复杂度时的表现，参考YOLOv5-N与YOLOv5-S进行网络结构设置，其轻量特征提取模块为CSP[10]，本文采用“[轻量化模块名]-[网络结构设置]”的形式对各先进方法进行简称，如CSP-S是指采用YOLOv5-S的网络结构设置并以CSP作为骨干网络构建模块的目标检测

模型。若对比方法的网络结构设置相同, 如均为“-S”, 则代表其骨干网络的宽度与深度相同。不同网络结构设置之间的大小关系为: “-N”的宽度与深度最低, “-S”次之。

2.3　对比实验

选取多种应用特征复用与卷积核分解设计思想的先进方法进行对比。具体而言, CSP[10] 与 SPP[15] 为采用了特征复用设计思想的轻量化网络模块, Ghost[11] 与 C3X[18] 为采用了特征复用与卷积核分解设计思想的轻量化网络模块。其中, CSP[10] 为应用广泛的轻量化目标检测算法 YOLOv5 的骨干网络模块; SPP[15] 自 YOLOv3 起便集成至 YOLO 系列目标检测骨干网络中以高效提升特征表征质量; Ghost[11] 模块是由华为诺亚方舟实验室提出, 旨在提升端侧骨干网络表征效率的轻量化模块; C3X[18] 模块为应用交叉卷积进行特征提取的 CSP[10] 模块, 旨在提升弱小目标结构与细节的表征能力。其中, SPP[15] 通过最大池化操作引入多个尺寸感受野特征, 而 C3X[18] 应用交叉卷积实现了卷积核分解, 但上述方法均不具备感受野筛选机制; Ghost[11] 根据特征重要性设计其特征提取方式, 但不显式地预测感受野调节系数。

如表 1 所示, 与改进前 YOLOv5-N 网络使用的轻量化网络模块 CSP[10] 相比, 本文方法在参数量下降 0.4×10^6 的同时将准确率与 AP-0.5:0.95 分别提升了 0.04 与 0.01; 与其他先进方法相比, 本文方法与 C3X[18] 为仅有的实现以低参数量(采用 C3X 改进后参数量下降 0.1×10^6)完成 AP-0.5:0.95 提升 (采用 C3X 改进后精度提升 0.02) 的优化方法。与改进前 YOLOv5-S 网络使用的轻量化网络模块 CSP[10] 相比, 本文方法在参数量下降 1.9×10^6 的同时将准确率与 AP-0.5:0.95 分别提升了 0.02 与 0.03; 与其他先进方法相比, 本文方法实现了与其他先进方法相近的泛化性。与最优的改进方法 C3X[18] 相比, 本文方法的准确率高 0.02, AP-0.5 与 AP-0.5:0.95 均低 0.02。由于本文方法的参数量低了 1.4×10^6, 建模复杂度更低, 因此, 本文方法并未逊于最先进的改进方法[18]。在实现本文方法的过程中, 并未采用混合精度或 BN 层融合等加速技术, 因此其计算量略逊于所比较的各先进方法。

表 1　ISDD[16] 数据集上本文方法与先进方法的比较

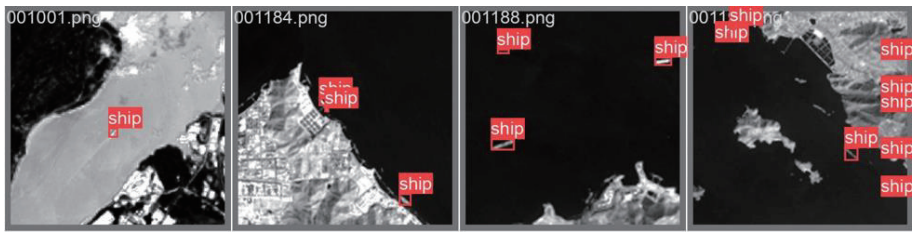
Table 1　Comparison between the proposed algorithm and advanced algorithms on the ISDD[16] dataset

方法	准确率	召回率	AP-0.5	AP-0.5:0.95	参数量	浮点运算操作数
CSP-N[10]	0.84	0.82	0.86	0.40	1.7×10^6	4.2 GFLOPs
C3X-N[18]	0.88	0.79	0.87	0.42	1.6×10^6	3.9 GFLOPs
Ghost-N[11]	0.87	0.77	0.85	0.40	1.5×10^6	2.4 GFLOPs
SPP-N[15]	0.88	0.77	0.85	0.40	1.6×10^6	3.1 GFLOPs
DMRF-N (本文方法)	0.88	0.78	0.86	0.41	1.3×10^6	3.6 GFLOPs
CSP-S[10]	0.87	0.83	0.87	0.40	7.0×10^6	16.0 GFLOPs
C3X-S[18]	0.88	0.83	0.89	0.45	6.5×10^6	5.4 GFLOPs
Ghost-S[11]	0.91	0.79	0.87	0.43	5.9×10^6	4.4 GFLOPs
SPP-S[15]	0.88	0.84	0.88	0.43	6.4×10^6	5.3 GFLOPs
DMRF-S (本文方法)	0.90	0.79	0.87	0.43	5.1×10^6	6.9 GFLOPs

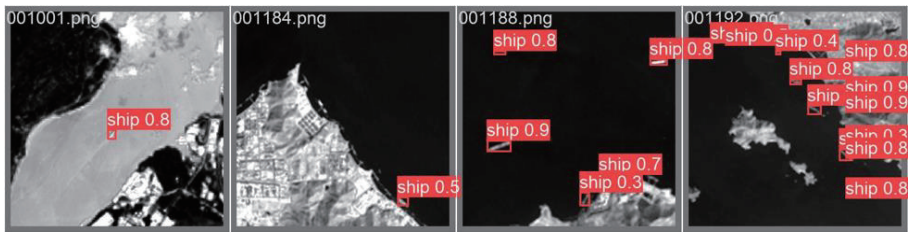
为了分析本文方法提升模型的泛化性, 对比了改进前 (CSP[10]) 后 (本文方法) 在复杂场景下的检测结果。如图 3(a) 所示, 依次选取了云雾单目标场景、港口多目标场景、离岸多目标场景与近岸多目标场景这 4 个复杂场景。在云雾单目标场景下, 改进前后模型均能正确检测红外弱小舰船, 本文方法的检测置信度提升了 0.1; 在港口多目标场景下, 模型需检测在港口停靠的 2 艘呈相弱小的舰船目标, 改进前方法漏检了目标结构信息较弱的舰船目标, 而本文方法仍以高置信度准确检测出 2 艘舰船; 在离岸多目标场景下, 复杂地物容易导致检测模型虚警, 改进前方法将港口口岸检测为目标, 且其中一个虚警置信度极高 (0.7), 而本文方法仍未出现虚警且以高置信度正确检出目标; 近岸多目标场景下包含 9 个目标, 其中部分目标为密集目标。对于采用锚框回归的 YOLOv5 检测算法而言, 准确检测密集目标具备极大挑战。此时, 简单进行轻量化的 CSP 模块难以准确表征红外弱小密集目标, 以较高的置信度指示该场景下存在 11 个目标。本文方法检测框个数和位置与真实目标数量和位置一致, 不存在虚警。综上, 认为本文方法通过抑制虚警与提升检测框的质量(置信度、位置)从而实现了检测准确率与 AP-0.5:0.95 的提升。该实验结果表明, 通过轻量化地模拟感受野调节机制, 可以可控地提升模型泛化性。

为分析本文方法表征优势的由来, 记录了改进前后模型于训练过程中的验证集损失的变化趋势。验证集损失可以反映检测算法在训练过程中的学习策略, 定位损失反映了模型指示前景大致位置的能力, 而回归框损失反映了模型指示前景位置的精度。如图 4(a) 所示, 对于应用本文方法改进的检测模型而言, 定位损失从 0.25 收敛至 0.01, 变化极其剧烈; 而回归框损失变化平缓, 但当定位损失上升时也随之上升。由于回归框损失随定位损失而变化, 因此认为改进后模型倾向于先习得定位能力, 并同时强化与其相关的检测框质量。该学习策略符合检测算法的训练预期, 即目标检测任务需具备目标定位及其识别能力。

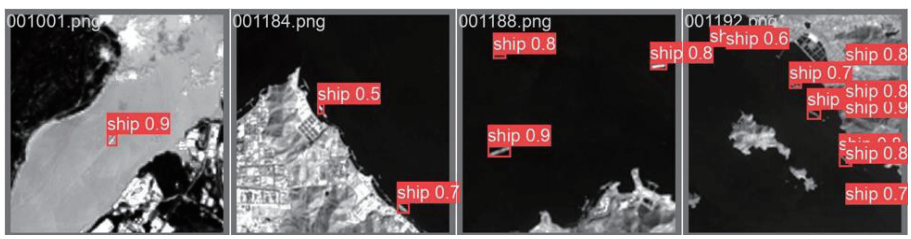
同时, 改进前模型 (CSP-S) 的学习过程不符合检测算法的训练预期。如图 4(b) 所示, 训练过程中



(a) 真实目标位置



(b) 改造前检测结果



(c) 改造后检测结果

图 3 复杂场景下真实目标位置及改进前后检测结果

Fig. 3 Object localization and the detection results before and after improvement in challenging scenarios

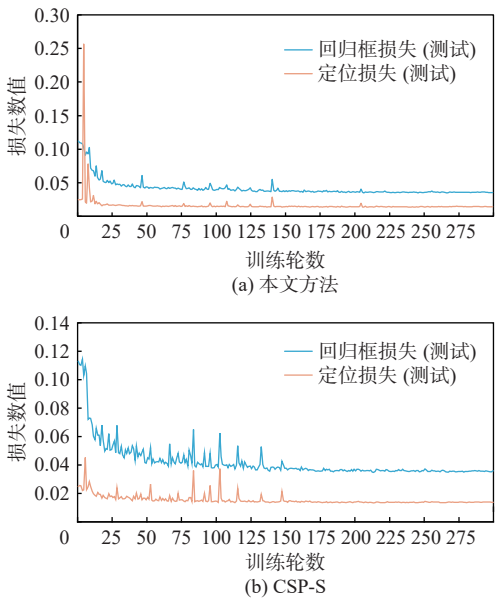


图 4 本文方法与 CSP-S 的验证集损失

Fig. 4 Validation set loss of the proposed method and CSP-S

前 150 训练轮数，改进前模型的定位与回归框损失波动频率与强度均高于本文方法改进模型，均呈振荡收敛趋势。该趋势意味着对于改进前模型而言，复杂场景下的红外弱小目标检测任务学习困难。此外，改进前模型的定位损失变化平缓而回归框损失变化剧烈，这意味着该模型的定位损失随回归框损失而变化。对于基于锚框的 YOLOv5 检测算法，该收敛趋势意味着改进前模型过度依赖先验的锚框，不符合检测算法的训练预期。

为了分析改进前后训练策略与泛化性之间的关系，对应用本文方法改进前后模型的检测精度变化趋势进行比较。AP-0.5 与 AP-0.5:0.95 可准确反映检测精度，前者代表将真实框的交并比大于 0.5 的检测框看作正确预测时正确预测占比，后者代表将真实框的交并比范围为 [0.5,0.95] 的检测框依次看作正确预测时其正确预测占比期望。该指标可以直接反映模型的泛化性，因此可验证不同模型的收敛能力。如图 5(a) 所示，改进后本文方法的验证集检测精度于 25 轮后收敛，并于第 140 轮、210 轮附近出现较高的精度波动；如图 5(b) 所示，改进前模型同样于 25 轮后收敛，但在 100 轮、150 轮时精度骤降至训练 10 轮的拟合效果。结合图 4(a) 所示改进前模型的训练策略，该精度骤降意味着即便改进前模型采用更多的参数进行建模，其表征红外弱小目标的能力有限。而本文方法通过轻量化模拟视觉感受野调节机制习得更优的红外弱小目标表征。

2.4 消融实验

为了验证 1.2 节轻量化骨干网络重构方法与其结合 1.3 节轻量化特征选择模块模拟视觉感受野调节机制的有效性，分别开展各重构方法的消融实验

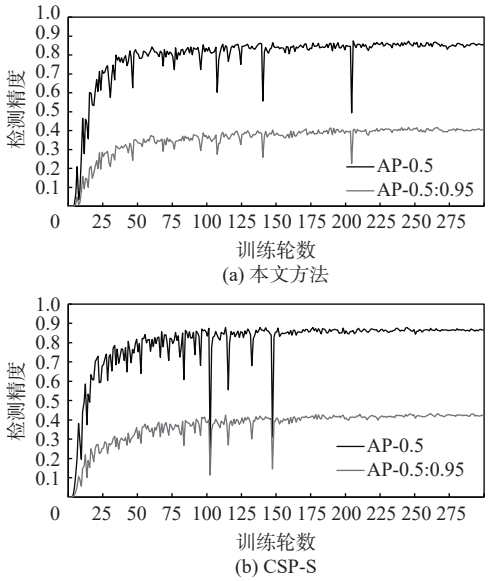


图 5　本文方法与 CSP-S 的验证集检测精度

Fig. 5　Validation set detection accuracy of the proposed method and CSP-S

与特征质量分析实验。具体地, 各重构方法的消融实验通过禁用不同轻量化设计思想后进行网络模块重构, 通过分析其检测精度与模型参数量验证各重构方法在弱小目标表征上的优劣; 特征质量分析实验采用 Eigen-CAM 算法实现, 该算法以深度特征为输入, 输出空间注意力[19]。虽然基于类激活映射 (class activation mapping, CAM) 的方法繁多, 但 Eigen-CAM[19] 仅关注于网络输出的深度特征, 且选取其对权重进行奇异值分解后最大的特征值对应的深度特征来进行空间注意力分析。对于不同结构的骨干网络, 该方法选取特征的标准统一且客观, 可有效分析网络的空间注意力。在分析时, 以完成 300 轮训练的不同骨干网络输出的深度特征作为输入单独进行分析。

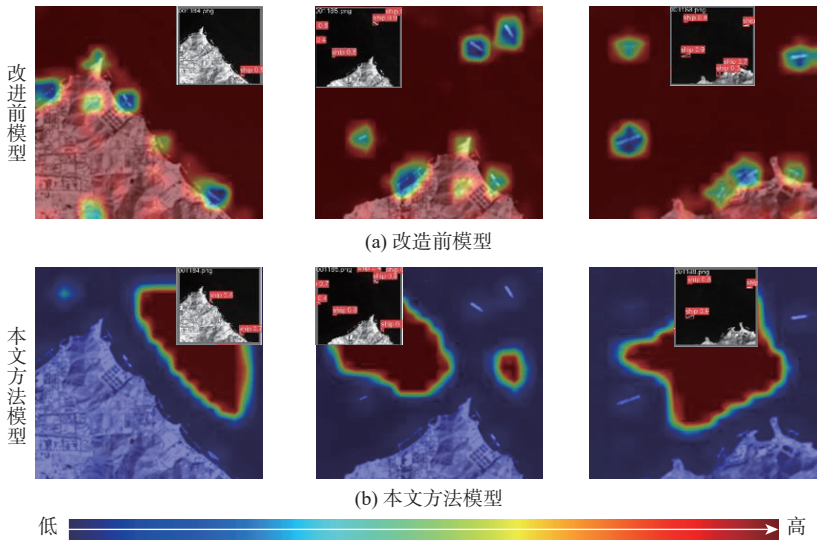
通过与仅进行特征复用的 CSP[10] 模块进行对比, 验证了 1.2 节所述轻量化骨干网络重构方法有效性。分别在“-S”与“-N”2 种宽度与深度不同的网络架构上进行消融实验, 旨在证明本文骨干网络重构方法适用于不同复杂度的轻量化骨干网络。如表 2 所示, 当网络宽度与深度极低时, 应用不同轻量化设计方法不会大幅降低模型参数量。CSP-N[10] 的参数量为 1.7×10^6, 仅应用本文方法提出的多尺寸感受野特征复用方法后参数量为 1.3×10^6, 而进一步应用卷积核分解后参数量为 1.2×10^6。然而, 对于该类复杂度较低的轻量化模型而言, 参数量下降伴随着模型精度上升的风险。复用改进-N 与 DMRF-N 的召回率分别下降了 0.05 与 0.04, 换来了准确率与 AP-0.5:0.95 分别提升 0.04 与 0.01。由于复用改进-N 召回率下降幅度大于准确率上升幅度, 且其他检测精度与 CSP-N[10] 持平, 因此认为复用改进-N 模型实现了性能优化, 但未实现检测精度优化; 而 DMRF-N 实现了 2 项检测精度提升, 且准确率上升幅度与准确率下降幅度持平, 故在参数量极低时本文方法提出的骨干网络重构方法优于 CSP-N[10]。同理, 当对 YOLOv5-S 进行骨干网络重构时, 复用改进-S 与 DMRF-S 均将参数量降低了 1.9×10^6, 降低了骨干网络的复杂度。然而, 复用改进-S 虽然提升了 AP-0.5:0.95 至 0.43, 但其召回率下降了 0.04 且 AP-0.5 下降了 0.01。结果表明, 当模型复杂度上升时, 应用多尺度感受野可以提升弱小目标检测的精准定位能力, 但仅使用特征复用不是最优的多尺度感受野感知模拟方式。通过对比 DMRF-S 与 CSP-S, 联合应用特征复用与卷积核分解设计思想虽然导致召回率下降了 0.04, 但准确率与 AP-0.5:0.95 均上升 0.03, 故 1.2 节所述轻量化骨干网络重构方法可实现骨干网络表征方式的优化。

通过采用 Eigen-CAM[19] 分析改进前模型 (CSP-S[10]) 与本文方法模型 (DMRF-S) 的深度特征, 发现 DMRF-S 通过调节空间注意力实现红外弱小目标的高效表征。图 6 展示了复杂场景下改进前后模型的特征分析结果, 其中, 模型的检测结果以缩略图形式展示, 而深度特征的空间注意力形式以热力图形式展示。热力图中, 蓝色代表空间注意力数值低, 红色代表空间注意力数值高。除图 3 中港口多目标场景与离岸多目标场景, 分析时额外选取了一副离岸多目标场景以进一步说明相似地物对弱小目标检测造成的注意力干扰。基于能量学习理论, 深度学习的优化目标为使机器毫不费力地完成学习任务[20]。以损失函数建模能力, 实现优化目标的深度学习模型在完成训练时损失数值收敛且数值极低。以图 4 与图 5 为例, 完成训练的 CSP-S[10] 与 DMRF-S 均满足该优化目标, 即此时其识别目标的能量开销低于背景, 进行特征空间可视化后模型认为的目标区域应呈蓝色, 与目标差异大的区域为红色。如图 6(a)所示, 改进前模型注意力分配不均, 与上述预期不符, 并出现漏检或虚警。在港口多目

表 2　相关轻量化设计方法在 ISDD[16] 数据集上的检测性能

Table 2　The detection performance of light-weight design approaches on the ISDD[16] dataset

设计方法	特征复用	卷积核分解	准确率	召回率	AP-0.5	AP-0.5:0.95	参数量
CSP-N[10]	√	×	0.84	0.82	0.86	0.40	1.7×10^6
复用改进-N	√	×	0.88	0.77	0.86	0.40	1.3×10^6
DMRF-N	√	√	0.88	0.78	0.86	0.41	1.2×10^6
CSP-S[10]	√	×	0.87	0.83	0.87	0.40	7.0×10^6
复用改进-S	√	×	0.88	0.79	0.86	0.43	5.1×10^6
DMRF-S	√	√	0.90	0.79	0.87	0.43	5.1×10^6



(a) 改造前模型

(b) 本文方法模型

图 6 通过 Eigen-CAM[19] 可视化, CSP-S 和 DMRF-S(本文方法) 的空间注意力

Fig. 6 The spatial attention of the CSP-S and DMRF-S (the proposed method) visualized via Eigen-CAM[19]

标与离岸多目标场景下,均出现注意力为黄色的区域且全图为大面积的红色区域,表明该模型在复杂场景下难以准确区分目标与相似地物之间的差距,从而漏检左图与中图内的停靠船只,并将右图中的堤岸错误检出。与之相对,本文方法模型的红色注意力区域有限且位于海洋区域,而在于目标相似的地物区域为大面积蓝色区域。这表明通过模拟感受野调节机制,完成训练的改进后模型可轻易区分前景和与其相似的地物。对于改进后模型而言,与目标相似的地物区域信息量更低,因此该区域的注意力抑制而与其表观差异大的背景区域注意力集中,符合实验预期且与感受野调节机制中的集中与抑制特性高度相似。

3 结 论

1) 本文针对红外遥感图像中弱小舰船目标的高效表征问题,提出了一种模拟视觉感受野调节机制的轻量化神经网络设计方法,显著提升了检测精度和,并降低了建模复杂度。

2) 通过在红外遥感弱小目标场景下进行对比,本文方法可显著提升检测精度 AP-0.5:0.95 高达 0.03 并同步降低了 1.9×10^6 的参数量,其轻量化程度优于应用广泛的先进方法。

3) 通过消融实验对深度特征分析,证明了未模拟感受野调节机制的方法存在视觉拥挤现象,而本文方法通过模拟感受野调节机制实现了注意力的集中与抑制机制。

4) 鉴于视觉拥挤现象可能普遍影响卷积神经网络在遥感图像处理时对弱小目标的检测性能,未来的研究将探索通过嵌入模拟感受野调节机制的网络模块来提升基于卷积神经网络的目标检测算法在可见光遥感弱小目标检测场景下的性能。具体而言,将在全色、合成孔径雷达和可见光遥感图像中验证该模块的有效性,并量化其对目标检测算法性能的具体提升。

参考文献 (References)

[1] 田祥瑞, 贾茆钧, 罗欣, 等. 基于嵌入式平台的航拍目标智能识别[J]. 计算机测量与控制, 2022, 30(11): 153-160.
TIAN X R, JIA Y J, LUO X, et al. Target intelligent recognition of aerial photography images based on embedded platform[J]. Computer Measurement & Control, 2022, 30(11): 153-160(in Chinese).

[2] 陈立群, 邹旭, 张磊, 等. 基于国产商用器件的星载智能目标检测技术[J]. 航空学报, 2023, 44(S2): 728860.
CHEN L Q, ZOU X, ZHANG L, et al. Spaceborne intelligent target detection technology based on domestic commercial devices[J]. Acta Aeronautica et Astronautica Sinica, 2023, 44(S2): 728860(in Chinese).

[3] 韩永赛, 马时平, 何林远, 等. 改进的深度神经网络下遥感机场区域目标检测[J]. 北京航空航天大学学报, 2021, 47(7): 1470-1480.
HAN Y S, MA S P, HE L Y, et al. Regional object detection of remote sensing airport based on improved deep neural network[J]. Journal of Beijing University of Aeronautics and Astronautics, 2021, 47(7): 1470-1480(in Chinese).

[4] 王玺坤, 姜宏旭, 林珂玉. 基于改进型 YOLO 算法的遥感图像舰船检测[J]. 北京航空航天大学学报, 2020, 46(6): 1184-1191.
WANG X K, JIANG H X, LIN K Y. Remote sensing image ship detection based on modified YOLO algorithm[J]. Journal of Beijing University of Aeronautics and Astronautics, 2020, 46(6): 1184-1191 (in Chinese).

[5] 张冬冬, 王春平, 付强. 基于 anchor-free 的光学遥感舰船关重部位检测算法[J]. 北京航空航天大学学报, 2024, 50(4): 1365-1374.
ZHANG D D, WANG C P, FU Q. Ship's critical part detection algorithm based on anchor-free in optical remote sensing[J]. Journal of Beijing University of Aeronautics and Astronautics, 2024, 50(4): 1365-1374(in Chinese).

[6] 王殿伟, 胡里晨, 房杰, 等. 基于改进 Double-Head RCNN 的无人机航拍图像小目标检测算法[J]. 北京航空航天大学学报, 2024, 50(7): 2141-2149.
WANG D W, HU L C, FANG J, et al. Small target detection algorithm based on improved Double-Head RCNN for UAV aerial

images[J]. Journal of Beijing University of Aeronautics and Astronautics, 2024, 50(7): 2141-2149(in Chinese).

[7] 陈金林, 吴一全, 苑玉彬. 无人机视角下目标检测的 YOLO 系列算法研究进展[J/OL]. 北京航空航天大学学报, 2024: 1-33. [2024-05-29]. http://kns.cnki.net/KCMS/detail/detail.aspx?filename=BJHK20240822002&dbname=CJFD&dbcode=CJFQ.
CHEN J L, WU Y Q, YUAN Y B. Research progress of YOLO series algorithms for target detection from UAV perspective[J/OL]. China Industrial Economics, 2024: 1-33. [2024-05-29]. http://kns.cnki.net/KCMS/detail/detail.aspx?filename=BJHK20240822002&dbname=CJFD&dbcode=CJFQ(in Chinese).

[8] 李红光, 王玉峰, 杨丽春. 基于元学习的小样本遥感图像目标检测[J]. 北京航空航天大学学报, 2024, 50(8): 2503-2513.
LI H G, WANG Y F, YANG L C. Meta-learning-based few-shot object detection for remote sensing images[J]. Journal of Beijing University of Aeronautics and Astronautics, 2024, 50(8): 2503-2513 (in Chinese).

[9] 吕奕龙, 苟瑶, 李敏, 等. 基于注意力引导和多样本决策的舰船检测方法[J]. 北京航空航天大学学报, 2025, 51(1): 202-213.
LYU Y L, GOU Y, LI M, et al. Ship detection method based on attentional guidance and multi-sample decision[J]. Journal of Beijing University of Aeronautics and Astronautics, 2025, 51(1): 202-213(in Chinese).

[10] JOCHER G. YOLOv5 by ultralytics (version 7.0) [EB/OL]. (2022-11-22)[2024-05-29]. https://doi.org/10.5281/zenodo.3908559.

[11] HAN K, WANG Y H, TIAN Q, et al. GhostNet: more features from cheap operations[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2020: 1577-1586.

[12] SZEGEDY C, LIU W, JIA Y Q, et al. Going deeper with convolutions[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2015: 1-9.

[13] HU J, SHEN L, SUN G. Squeeze-and-excitation networks[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2018: 7132-7141.

[14] HE D J, WANG Y Y, FANG F. The critical role of V2 population receptive fields in visual orientation crowding[J]. Current Biology, 2019, 29(13): 2229-2236.

[15] HE K, ZHANG X, REN S, et al. Spatial pyramid pooling in deep convolutional networks for visual recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015, 37(9): 1904-1916.

[16] HAN Y Q, LIAO J W, LU T S, et al. KCPNet: knowledge-driven context perception networks for ship detection in infrared imagery[J]. IEEE Transactions on Geoscience and Remote Sensing, 2022, 61: 5000219.

[17] LUAN S Z, CHEN C, ZHANG B C, et al. Gabor convolutional networks[J]. IEEE Transactions on Image Processing, 2018, 27(9): 4357-4366.

[18] LIU Y Q, JIA Q, FAN X, et al. Cross-SRN: structure-preserving super-resolution network with cross convolution[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2022, 32(8): 4927-4939.

[19] BANY M M, YEASIN M. Eigen-CAM: visual explanations for deep convolutional neural networks[J]. SN Computer Science, 2021, 2(1): 47.

[20] LECUN Y, CHOPRA S, HADSELL R, et al. A tutorial on energy-based learning[M]//BAKIR G, HOFMANN T, SCHLKOPF B, et al. Predicting structured data. Cambridge: MIT Press, 2006: 45-49.

Lightweight neural network design for infrared small ship detection

TANG Wenting[1], LI Bo[2], JI Mengqi[2, 3, 4, *]

(1. School of Computer Science and Engineering, Beihang University, Beijing 100191, China;
2. School of Artificial Intelligence, Beihang University, Beijing 100191, China;
3. Hangzhou International Innovation Institute of Beihang University, Hangzhou 311115, China;
4. National Key Laboratory of Information Systems Engineering, Changsha 410003, China)

Abstract: A lightweight neural network design method is proposed to efficiently represent small ships in infrared remote sensing images. To improve the representation effect of infrared dim and small targets, a method for simulating the visual receptive field adjustment mechanism that incorporates multi-scale receptive field perception and selection processes is proposed. This method is inspired by the visual attention-driven receptive field adjustment mechanism. A lightweight feature selection operator is devised to enhance the receptive field selection, and feature reuse and convolution kernel decomposition are used to optimize the multi-scale receptive field perception process in order to further increase efficiency. Experimental results on an infrared dim and small ship detection dataset show that the network detection accuracy increased by 2%, with a reduction of 2.3×10^6 parameters and 9.1×10^9 computations compared to general lightweight networks. In complex scenarios with similar ground interference, this method effectively reduces false alarms and suppresses missed detections.

Keywords: small object detection; satellite remote sensing images; receptive field; neural network design; lightweight neural networks

Received: 2024-10-16; **Accepted**: 2024-11-22; **Published Online**: 2025-02-26 16:39

URL: link.cnki.net/urlid/11.2625.V.20250226.1305.002

Foundation items: National Natural Science Foundation of China (62171253); Young Elite Scientists Sponsorship Program by CAST (2022QNRC001); The Open Fund (WDZC20255290411)

* **Corresponding author**. E-mail: jimengqi@buaa.edu.cn

http://bhxb.buaa.edu.cn jbuaa@buaa.edu.cn
DOI: 10.13700/j.bh.1001-5965.2023.0381

基于多视口自适应融合的全景图像质量评价

冯晨曦[1]，张地[1,2]，林敢[3]，叶龙[2,4,*]

(1. 中国传媒大学 媒体融合与传播国家重点实验室，北京 100024；2. 中国传媒大学 数据科学与智能媒体学院，北京 100024；
3. 中国传媒大学 海南国际学院，陵水 572400；4. 中国传媒大学媒介音视频教育部重点实验室，北京 100024)

摘 要：现有的全景图像质量评价 (OIQA) 模型在提取各视口局部特征时相对独立，导致计算复杂度较高，且难以利用端到端的融合模型对各视口间的关联性进行刻画。针对该问题，提出了一种基于特征共享和多视口自适应融合的质量评价方法。利用共享的骨干网络，将现有方法中互相独立的视口分割与计算任务转换到特征域，使得仅经1次前馈计算即可提取整个图像的局部特征。在此基础上，引入基于球面均匀采样的特征域视口分割方法以保证观察空间和表示空间的像素密度一致，并用语义信息指导各视口局部质量特征的自适应融合。所提方法在压缩的虚拟现实图像质量 (CVIQ) 数据集和 OIQA 数据集上的 Pearson 线性相关系数 (PLCC) 和 Spearman 秩相关系数 (SRCC) 均在 0.96 以上，与现有主流评价方法相比达到最优。相较于传统评价方法结构相似性指数 (SSIM)，所提方法在2个数据集上的平均 PLCC 和平均 SRCC 分别提高了 9.52% 和 8.7%；相较于最新评价方法多感知特征图像质量评价 (MPFIQA)，所提方法在2个数据集上的平均 PLCC 和平均 SRCC 分别提高了 1.71% 和 1.44%。

关 键 词：全景图像；无参考图像质量评价；球面均匀采样；自适应特征融合；等距柱状投影
中图分类号：TN911.73
文献标志码：A **文章编号**：1001-5965(2025)07-2404-11

得益于网络信息技术的飞速发展和移动智能终端软硬件的迭代升级，视觉信息正从传统的平面显示走向沉浸式呈现。同 5G、超高清、高动态范围等高新视频技术相结合，虚拟现实 (virtual reality, VR) 给用户带来了身临其境的观看体验[1]。然而，全景图像在采集、编码、传输等过程中会难以避免地引入模糊、块效应、过曝/欠曝等各种降质问题[2]，影响用户的观看体验。因此，全景图像感知质量的有效评价是保障用户视听体验、提升并优化视频平台服务质量的关键因素之一[3]。

现有的全景图像质量评价 (omnidirectional image quality assessment, OIQA) 算法根据处理流程，主要可分为面向投影、面向图像块和面向视口3类。面向投影的算法直接对等距柱状投影 (equirectangular projection, ERP) 等方式映射后的内容进行特征提取和融合[4-5]，其优点是可直接将现有面向普通二维图像质量评价的模型应用到 OIQA 领域，但并未深入考虑普通图像与全景图像在观看行为模式之间的区别。同时，该类算法直接将具有几何畸变的投影图像送入深度学习模型，没有分析这种畸变对于现有深度学习模型的潜在影响。面向图像块的算法[6]一般将全景图像划分为多个局部区域分别进行训练，其目的是增加训练样本数量以促进模型收敛，但各局部区域共享整个图像的平均意见分数 (mean opinion scores, MOS)，对于一些空域感知质量分布不均匀的失真类型，会引入明显的标注误差。面向视口的算法首先将 ERP 内容映射到球体，随后将局部范围的视口内容投影到二维平面以消

收稿日期：2023-06-16；**录用日期**：2023-12-15；**网络出版时间**：2024-02-01 15:38
网络出版地址：link.cnki.net/urlid/11.2625.V.20240131.1721.002
基金项目：国家重点研发计划 (2021YFF0900503)；国家自然科学基金 (61971383,62001432)；北京自然科学基金 (4244109)
* **通信作者**. E-mail：yelong@cuc.edu.cn

引用格式：冯晨曦，张地，林敢，等. 基于多视口自适应融合的全景图像质量评价 [J]. 北京航空航天大学学报，2025，51（7）：2404-2414.
FENG C X，ZHANG D，LIN G，et al. Omnidirectional image quality assessment based on adaptive multi-viewport fusion [J]. Journal of Beijing University of Aeronautics and Astronautics，2025，51（7）：2404-2414（in Chinese）.

除几何畸变, 最后通过模拟用户观看 VR 内容时的行为模式设计评价模型[7-8]。然而, 该类算法需要将各视口内容分别送入骨干网络提取特征, 使得特征提取过程消耗大量的计算资源。

由以上分析可知, 面向视口的评价模型充分考虑了用户观察全景图像时的行为特性, 因此, 可以给出更符合人类视觉系统感知特性的质量预测分值, 但其视口生成及特征提取过程互相独立, 即各视口内容需分别送入骨干网络提取局部特征, 复杂度较高。如何在有效降低模型复杂度的同时, 对模型进行更为充分的端到端训练, 是全景内容质量评价领域的重要问题。

对此, 本文设计了一种特征共享的多视口自适应融合的质量评价框架——仅前馈一次网络 (you only feedforward once network, YOFONet)。其核心思路在于将面向像素域的视口分割、多视口融合转化为特征域的局部特征提取与融合, 从而在实现特征共享, 且仅需前馈计算 1 次的同时, 也能实现更为灵活的局部特征划分与聚合。

在局部特征的提取方面, 由于 ERP 投影后不同纬度上的像素密度存在差异, 若直接在表示空间对特征均匀分割, 会导致高纬度、低像素密度区域与低纬度、高像素密度区域被等同对待, 不但使各局部视口和观察者的视场角不一致, 还将影响算法精度。针对该问题, 本文通过球面均匀采样的视口分割方法, 在观察空间对图像进行球面均匀采样, 将各采样点的三维坐标映射至表示空间, 在表示空间根据各采样点纬度计算其视口范围, 以此保持观察空间和表示空间的像素密度一致。

在局部特征的聚合方面, 本文利用 Transformer 变换器[9] 统筹各视口特征的自适应融合。此外, 考虑到感知质量和语义识别具有紧密关联, 有效刻画两者间的相互作用关系对于指导网络模型的设计具有重要作用, 本文以骨干网络的高层抽象特征作为语义信息, 并综合多尺度特征得到质量信息, 进而在多头注意力的计算过程中对语义-质量信息联合分析, 最终实现语义信息引导下多视口质量特征的自适应融合。

综上所述, 本文主要创新点如下: ①针对现有方法在视口特征提取过程中互相独立, 导致计算复杂度较高且难以进行端到端训练的问题, 设计了各视口在特征域的统一描述方法, 有效避免了模型计算过程中的多次前馈操作; ②针对 ERP 投影后不同纬度上像素密度不同的特性, 设计了基于球面均匀采样的非均匀视口分割方法, 从而维持观察空间和表示空间像素密度的一致性, 并确保所提取各局部特征对应相同的视场角; ③综合考虑人类视觉系统中语义识别和质量感知任务的高度相关性, 以对应位置的语义特征信息指导各视口局部质量特征的自适应融合, 进一步提高 OIQA 模型的精度。

1　相关工作

由于全景图像具有超高的分辨率和全方位的观看视角, 投影处理会带来不同程度的畸变, 并导致不同区域像素密度具有显著差别, 因此, 面向二维平面图像的质量评价模型难以有效评估全景图像的感知质量。为解决该问题, 研究者们分别在全参考和无参考范畴开展了一系列工作。

在全参考 OIQA 方面, 研究主要集中在通过消除与投影相关的几何失真影响来改进现有的普通图像全参考质量评价指标。Yu 等[10] 提出了基于球面的峰值信噪比 (spherical peak signal-to-noise ratio, S-PSNR) 算法, 从球体中均匀地采样像素位置, 根据球坐标与二维坐标的关系, 从参考图像和失真图像中获取像素值, 计算参考图像与失真图像的像素值差异以预测其感知质量。Yule 等[11] 提出了加权球面峰值信噪比 (weighted spherical peak signal-to-noise ratio, WS-PSNR) 算法, 通过在平面图像上引入权重值来计算图像质量, 其权重由采样区域在投影平面中被拉伸的程度决定。Zakharchenko 等[12] 在计算峰值信噪比 (peak signal-to-noise ratio, PSNR) 之前, 将参考图像和失真图像变换到克拉斯特抛物线投影空间, 此投影方式在相同空间分辨率下的形状失真优于 ERP 投影。

与全参考 OIQA 算法相比, 无参考算法不需要参考图像的任何信息, 在实际系统中有更广泛的应用。Liu 等[4,13] 提出了基于结构和自然特征的 OIQA 算法, 对 ERP 形式的全景图像, 提取其纹理和梯度信息作为结构特征来解决全景图像的结构失真问题, 提取其亮度、颜色和图像熵信息作为自然特征, 应用支持向量回归建立上述 2 类特征与主观分值之间的关系, 得到最终的质量分数。

由于全景图像具有超高分辨率, 基于图像块的计算方法被普遍使用。例如, Kim 等[14] 提出了基于人类感知引导的虚拟现实全景图像质量评价 (deep virtual reality image quality assessment, DeepVR-IQA) 模型, 将 ERP 形式的图像划分为不重叠等大的图像块, 各图像块通过位置编码器和空间编码器分别提取位置特征和空间特征, 权重估计器同时考虑这 2 种特征来估计各图像块的权重, 质量分数估计器聚合空间特征得到各图像块的质量分数。DeepVR-IQA 采用简单的平均分块方法, 但根据 ERP 投影原理, 从两极区域到赤道区域投影像素分布极不均匀, 平均分块并不能反映这种不均匀性。

Truong 等[6]只考虑了靠近赤道周围的区域，将数据增强方法[15-16]用在了全景图像中，在训练阶段将全景图像的中间区域分为等大的图像块，并为每个图像块分配和原图一样的 MOS 分值，以确保模型可以从有限的数据集中学习到最多的信息。

上述基于图像块的方法并未考虑 ERP 投影两极区域会产生严重形变。因此，Jiang 等[5]提出了基于立方体投影的 VR 图像质量评价模型，并设计了 3 种方案将各立方体面的质量分数同头动、眼动等信息综合，得到最终质量预测分值。与 ERP 相比，立方体投影的畸变要小很多，但这并不符合用户的观看特性，即用户佩戴头显观看全景图像时，一次只能看到其中一小部分区域。

为了与在 VR 设备中看到的全景图像视觉内容保持一致，研究者提出了各种基于视口的算法。Sun 等[8]将 ERP 形式的全景图像投影到前、后、左、右、上、下 6 个视口中，并使用 6 个并行的骨干网络提取对应视口图像的特征以回归质量分值。类似地，近年来许多工作都聚焦于如何提取视口和建模视口间的关系上，如 Xu 等[7]提出的面向视口的图卷积网络根据人类对结构信息的视觉敏感度提取视口，采用图卷积网络隐式建模不同视口之间的交互，预测全景图像质量分数。但此方式无法反映视口的高阶交互，且在图构造时只考虑视口的空间位置，忽略了其内容特征。为解决上述问题，Fu 等[17]采用超图卷积网络，通过超图建模视口间的交互，为每个视口构建 1 个基于位置的超边和 1 个基于内容的超边。

与上述工作只考虑 ERP 图像或只考虑视口图像不同，Zhou 等[18]将 ERP 投影图像和多个视口图像相结合，从 ERP 图像提取自然场景统计特征作为全景自然性的度量，同时依据观看模式，以各视口自然场景统计特征刻画局部自然性，以低频、高频小波子带的熵强度来衡量全景图像的多频信息，并使用支持向量回归进行聚合。

由以上分析可知，面向视口的 OIQA 算法可以更好地模拟用户观看 VR 内容时的真实行为特性，但在像素域进行多视口分割后，各视口内容需分别利用深度学习模型进行特征提取，且由于各视口内容在送入骨干网络前已确定，该框架较难适配端到端的训练过程。而本文所设计的 YOFONet 通过特征共享的多视口自适应融合研究思路，为解决上述问题提供了有效途径。

2 YOFONet 的结构与实现

本文提出的 YOFONet 的整体框架如图 1 所示，主要包括基于卷积神经网络 (convolutional neural networks, CNN) 的多尺度特征融合模块、基于球面均匀采样的非均匀视口分割模块和基于 Transformer 的语义-质量聚合模块。

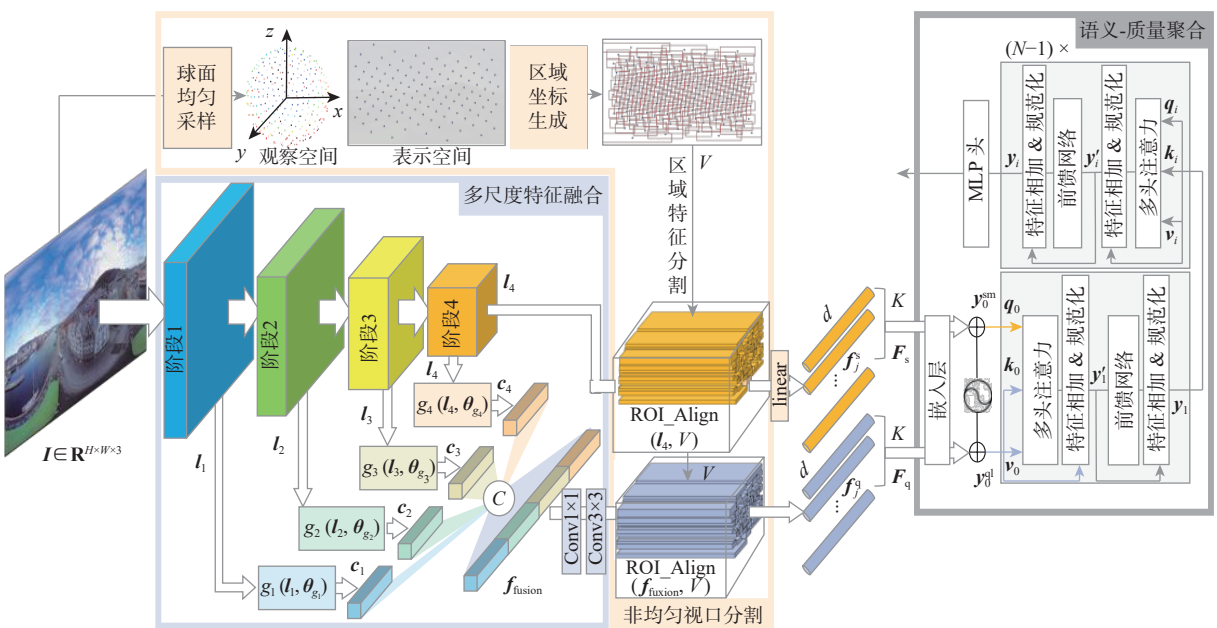


图 1 YOFONet 模型整体结构

Fig. 1 Overall structure of YOFONet model

其中，多尺度特征融合模块通过不同尺度特征的联合描述来提高网络的特征表达能力；非均匀视口分割模块通过分析 ERP 投影特性，在观察空间进行球面均匀采样以保证各视口区域的像素密度具有一致性；语义-质量聚合模块充分利用注意力机制，以语义特征指导局部质量特征的自适应融合，最终形成仅需 1 次前馈过程即可实现多视口信息提取与自适应融合的 OIQA 模型。

2.1　基于 CNN 的多尺度特征融合模块

机器视觉领域的诸多研究表明[19-21], 通过将不同尺度的特征相融合, 可有效提高网络的特征表达能力。因此, 将输入图像 $\boldsymbol{I}\in\mathbf{R}^{H\times W\times 3}$ 送入在 ImageNet 数据集[22] 上预训练的骨干网络 f_{b}, H 和 W 分别为图像的高度和宽度分辨率。如图 2 所示, 提取其多尺度特征 $\boldsymbol{L}$:

$$\boldsymbol{L}=\{\boldsymbol{l}_i|i\in[1,m]\}=f_{\mathrm{b}}(\boldsymbol{I};\boldsymbol{\theta}_{\mathrm{b}}) \tag{1}$$

式中: $\boldsymbol{l}_i$ 为 m 个尺度中第 i 个尺度的特征; $\boldsymbol{\theta}_{\mathrm{b}}$ 为骨干网络的预训练参数。将每个尺度的特征通过不同步长的卷积核映射至相同维度:

$$\boldsymbol{c}_i=g_i\left(\boldsymbol{l}_i,\boldsymbol{\theta}_{g_i}\right)\qquad \forall i\in[1,m] \tag{2}$$

式中: g_i 为第 i 个尺度上的映射函数, 其参数为 $\boldsymbol{\theta}_{g_i}$; $\boldsymbol{c}_i\in\mathbf{R}^{h\times w\times D}$ 为映射后第 i 个尺度上的特征, h、w 和 D 分别为特征 $\boldsymbol{c}_i$ 的高度、宽度分辨率和通道数。记 $C=\{\boldsymbol{c}_i|i\in[1,m]\}$, 将 m 个映射后的特征在通道上拼接形成多尺度质量特征 $\boldsymbol{f}_{\mathrm{fusion}}\in\mathbf{R}^{h\times w\times m\times D}=\mathrm{cat}(C)=\boldsymbol{c}_1\cup\boldsymbol{c}_2\cup\cdots\cup\boldsymbol{c}_m$。

经上述操作所提取的多尺度特征可充分反映在全景图像的感知质量变化时, 各尺度特征的相应变化情况。因此, 将其作为局部质量特征描述送入后续自适应融合模块。需要注意的是, 为了避免几何畸变对所提取特征表示能力的影响, 与其他算法以较小尺寸图像作为输入不同, 本文输入图像的分辨率 $H\times W=1\,024\times 512$(见 3.3.5 节)。

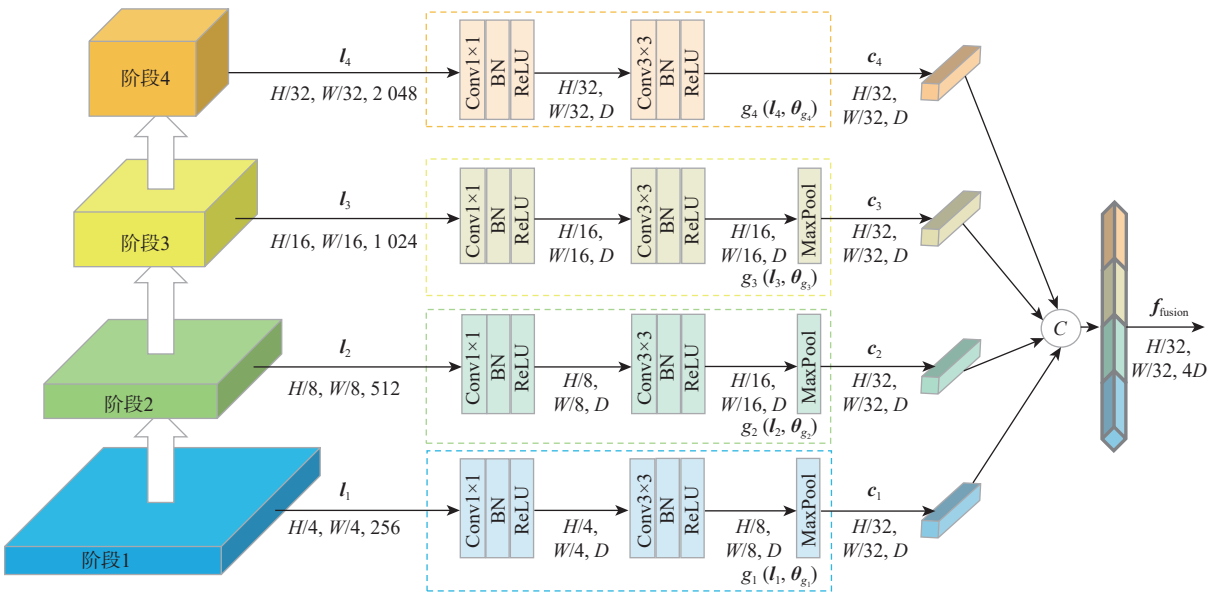


图 2　基于 CNN 的多尺度特征融合模块

Fig. 2　CNN-based multi-scale feature fusion module

2.2　基于球面均匀采样的非均匀视口分割模块

以融合的多尺度特征为输入, 在特征域设计一种灵活的视口分割方法, 模拟用户的视口观看方式。该方法共包含 3 步: ①通过球面均匀采样在观察空间选取在球面上分布均匀的采样点, 获得各采样点的空间直角坐标、球坐标及像素坐标; ②根据每个采样点所在的纬度确定以该采样点为中心的视口宽度, 计算出视口左上顶点和右下顶点的像素坐标, 从而保证不同视口的像素密度一致性; ③基于感兴趣区域对齐 (region of interest align, ROI_Align)[23] 分别对多尺度质量特征和高级语义特征按照各个视口的坐标分割出对应局部特征。

2.2.1　球面均匀采样

ERP 投影会对球面图像的极性区域过采样, 导致该区域的图像包含大量重复像素。通过对观察空间的球面图像而非表示空间的投影图像做均匀采样, 可以避免对极性区域进行不必要的计算。

如图 3 所示, 利用斐波那契球面均匀采样算法[24]

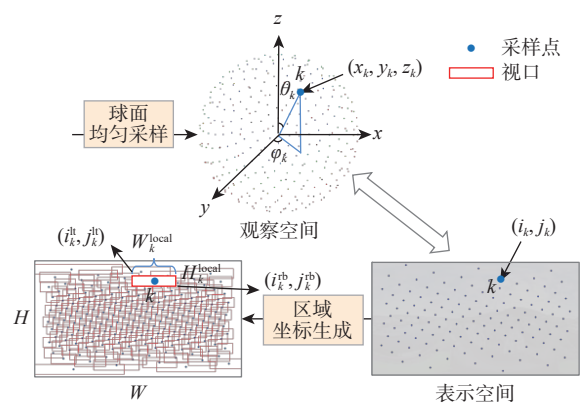


图 3　基于球面均匀采样的非均匀视口分割

Fig. 3　Non-uniform viewport segmentation based on spherical uniform sampling

计算采样点坐标。设球面半径$r=1$，采样点数为K，则第k个采样点的空间直角坐标(x_k,y_k,z_k)为

$$\begin{cases} z_k=(2k-1)/(K-1) \\ x_k=\sqrt{1-z_k^2}\cos(2\pi k\phi) \\ y_k=\sqrt{1-z_k^2}\sin(2\pi k\phi) \end{cases} \tag{3}$$

式中：ϕ为斐波那契采样间隔。在获取到采样点k的空间直角坐标后，先将其转换为球坐标(r,θ_k,φ_k)，再转换为平面像素坐标(i_k,j_k)：

$$\begin{cases} \varphi_k=\arctan(x_k,y_k) \\ \theta_k=\arcsin(z_k/r) \\ i_k=\dfrac{(\pi/2-\theta_k)H}{\pi}-0.5 \\ j_k=\dfrac{(\varphi_k+\pi)W}{2\pi}-0.5 \end{cases} \tag{4}$$

2.2.2 视口坐标生成

由于ERP投影后不同纬度上的像素密度不同，若直接对表示空间中所有采样点选取同样尺寸的视口，会导致各局部区域和观看者的视场角及像素密度不一致，影响算法精度。因此，根据采样点所在的纬度设置不同的视口宽度W^{local}，对于视口的高度H^{local}选取同样的值（如图3所示）。采样点k的视口宽度$W_k^{\text{local}}=W/(a\cos\theta_k)$，其中，$a$为控制缩放量的常数。根据视口的宽度和高度可计算出视口左上角和右下角顶点的坐标，视口左上角顶点的像素坐标为$\left(i_k^{\text{lt}},j_k^{\text{lt}}\right)=\left(i_k-H_k^{\text{local}}/2,j_k-W_k^{\text{local}}/2\right)$，右下角顶点坐标为$\left(i_k^{\text{rb}},j_k^{\text{rb}}\right)=\left(i_k+H_k^{\text{local}}/2,j_k+W_k^{\text{local}}/2\right)$，记所有视口的左上角和右下角顶点坐标集合为$V=\left\{\boldsymbol{v}_k=\left(i_k^{\text{lt}},j_k^{\text{lt}},i_k^{\text{rb}},j_k^{\text{rb}}\right)|k\in1,2,\cdots,K\right\}$。

2.2.3 视口特征分割

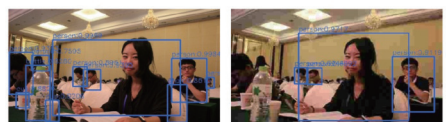
结合2.2.2节中各视口位置信息V及多尺度特征融合所提取的整个图像层面的质量特征，可利用ROI_Align方法[23]提取各视口对应的特征信息：$\boldsymbol{F}_{\text{q}}=\text{ROI_Align}(\boldsymbol{f}_{\text{fusion}},V)=\left\{\boldsymbol{f}_k^{\text{q}}|k\in[1,K]\right\}$，其中，$\boldsymbol{f}_k^{\text{q}}\in\mathbf{R}^{1\times d}$为多尺度质量特征$\boldsymbol{f}_{\text{fusion}}$在第$k$个采样点形成的视口特征，$d$为特征维度。为了使语义和质量视口特征维度一致，特征$\boldsymbol{l}_4$在经过ROI_Align后需经线性函数linear改变其维度，即$\boldsymbol{F}_{\text{s}}=\text{linear}\left(\text{ROI_Align}(\boldsymbol{l}_4,V)\right)=\left\{\boldsymbol{f}_k^{\text{s}}|k\in[1,K]\right\}$，其中，$\boldsymbol{f}_k^{\text{s}}\in\mathbf{R}^{1\times d}$为高级语义特征$\boldsymbol{l}_4$在第$k$个采样点形成的视口特征。

2.3 基于Transformer的语义-质量聚合模块

感知质量和语义识别间具有紧密关联[25-26]：一方面，语义信息是进行质量评价的前提，如图4(a)所示，虽然2幅图像间PSNR值很低，但由于其并不包含任何语义信息，人类视觉系统感知不到两者明显的质量差异，而图4(b)能感知到右图在花瓣和天空部分相较于左图具有明显的块效应损伤；另一方面，感知质量的降低也会影响语义识别任务的精度，如图4(c)所示，压缩后图像在识别物体数量、置信度方面均小于原始图像。由此可见，感知质量评估和语义识别任务间具有明显的相互依赖关系，因此，利用语义特征指导感知质量特征的回归，对于提升模型性能具有重要作用。



(a) 噪声图像，PSNR≈4.8 (b) 原图和压缩后的图像，PSNR≈22.3



(c) 对原图和压缩后的图像做目标检测

图4 质量感知与语义识别任务间的相互依赖关系示例

Fig. 4 Illustration of interdependent relationships between quality perception and semantic recognition

考虑到Transformer[9]的多头注意力机制可对输入的上下文信息进行编码，使模型可以学习到不同输入符号之间的依赖关系，因此，本文基于Transformer构建语义-质量聚合模块进行特征的融合和质量回归，以得到最终的质量预测分数。具体地，该模块将视口分割输出的语义视口特征$\boldsymbol{F}_{\text{s}}$和质量视口特征$\boldsymbol{F}_{\text{q}}$作为输入。为$\boldsymbol{F}_{\text{s}}$和$\boldsymbol{F}_{\text{q}}$分别附加额外的嵌入向量$\boldsymbol{f}_0\in\mathbf{R}^{1\times d}$以汇聚质量特征，并添加可训练的位置嵌入$\boldsymbol{P}\in\mathbf{R}^{(K+1)d}$以刻画位置信息，得到$\boldsymbol{y}_0^{\text{sm}}$和$\boldsymbol{y}_0^{\text{ql}}$作为Transformer编码器第1层的输入，编码器的质量汇聚过程可描述为

$$\begin{cases} \boldsymbol{y}_0^{\text{sm}}=\left[\boldsymbol{f}_0+\boldsymbol{P}_0,\boldsymbol{f}_1^{\text{s}}+\boldsymbol{P}_1,\cdots,\boldsymbol{f}_K^{\text{s}}+\boldsymbol{P}_K\right] \\ \boldsymbol{y}_0^{\text{ql}}=\left[\boldsymbol{f}_0+\boldsymbol{P}_0,\boldsymbol{f}_1^{\text{q}}+\boldsymbol{P}_1,\cdots,\boldsymbol{f}_K^{\text{q}}+\boldsymbol{P}_K\right] \\ \boldsymbol{q}_0=\boldsymbol{y}_0^{\text{sm}},\boldsymbol{k}_0=\boldsymbol{v}_0=\boldsymbol{y}_0^{\text{ql}} \\ \boldsymbol{y}_1'=\text{LN}\left(\text{MHA}(\boldsymbol{q}_0,\boldsymbol{k}_0,\boldsymbol{v}_0)+\boldsymbol{y}_0^{\text{ql}}\right) \\ \boldsymbol{y}_1=\text{LN}\left(\text{MHA}\left(\boldsymbol{y}_1'\right)+\boldsymbol{y}_1'\right) \\ \boldsymbol{q}_i=\boldsymbol{k}_i=\boldsymbol{v}_i=\boldsymbol{y}_{i-1}\qquad i=2,3,\cdots,N \\ \boldsymbol{y}_i'=\text{LN}(\text{MHA}(\boldsymbol{q}_i,\boldsymbol{k}_i,\boldsymbol{v}_i)+\boldsymbol{y}_{i-1}) \\ \boldsymbol{y}_i=\text{LN}\left(\text{MHA}\left(\boldsymbol{y}_i'\right)+\boldsymbol{y}_i'\right) \\ \boldsymbol{y}_N=\left[\boldsymbol{F}_0',\boldsymbol{F}_1',\cdots,\boldsymbol{F}_K'\right] \end{cases} \tag{5}$$

式中：N为Transformer编码器的层数；$\boldsymbol{q}_i$、$\boldsymbol{k}_i$、$\boldsymbol{v}_i$用于计算质量特征在空域上的注意力；LN为层归一化模块；MHA为多头注意力模块。编码器的输出$\boldsymbol{y}_N\in\mathbf{R}^{(K+1)d}$与$\boldsymbol{y}_0^{\text{ql}}$具有相同的维度。经上述操作，$\boldsymbol{y}_N$的第1个向量$\boldsymbol{F}_0'\in\mathbf{R}^{1\times d}$可在语义信息的引导下有效汇聚感知质量特征，因此，将其作为多层感知机（multilayer perceptron, MLP）的输入以学习图像的整体质量预测分值。MLP头由2个全连接层组成，第1个全连接后接ReLU[27]激活函数，第2个全连

接的输出即为质量预测分数Q。

由于本节模块任务为汇聚质量特征, 式(5)中需以质量特征作为$\boldsymbol{v}_0$, 进而利用质量特征和语义特征的位置相关信息对$\boldsymbol{v}_0$加权, 即$\boldsymbol{q}_0=\boldsymbol{y}_0^{\mathrm{sm}}$, $\boldsymbol{k}_0=\boldsymbol{y}_0^{\mathrm{ql}}$。需要注意的是, 置换$\boldsymbol{q}_0$和$\boldsymbol{k}_0$并不会对模型性能产生显著影响(见 3.3.2 节)。

2.4　训练设置

2.4.1　具体参数配置

对于基于 CNN 的多尺度特征融合模块, 选取在 ImageNet 上预训练的 ResNet-50[28] 作为骨干网络, 分别从 conv2-10、conv3-22、conv4-40 和 conv5-49 提取特征, 经过映射函数g后, 特征通道数$D=128$; 对于基于球面均匀采样的非均匀视口分割模块, 选择$K=128$个采样点, 计算视口宽度W^{local}时的缩放量$a=8$, 经过ROI_Align后, 提取的特征维度$d=256$; 对于语义-质量聚合模块, 实验使用$N=2$个堆叠的 Transformer 编码器, 每个编码器具有 6 个多头注意力。

2.4.2　损失函数设计

采用均方误差 (mean square error, MSE) 作为目标函数T, 如下:

$$T=\frac{1}{B}\sum_{i=1}^{B}(Q_i-G_i)^2 \tag{6}$$

式中: B为批次大小; Q_i和G_i分别为第i个失真图像的质量预测分值和 MOS 分值。

2.4.3　实验平台与训练流程

本文所有实验在 NVIDIA Quadro RTX 8000 GPU 上进行, 所提出的 YOFONet 框架由 Pytorch1.7 实现, 采用 ResNet-50 作为骨干网络, 为了获取语义信息以指导局部质量特征的聚合, 利用在 ImageNet 上的预训练参数初始化骨干网络, 且在训练过程中保持参数固定。后续模块的可学习参数采用 Adam[29] 优化器进行优化, 学习率为1×10^{-4}, 权重衰减为1×10^{-5}, 迭代次数为 50, 批次大小B为 12。此外, 为了防止模型过拟合, 在最后的 MLP 头中使用随机失活[30], 失活率设为 0.5。同其他工作[8,13,17] 类似, 在训练集和测试集的配置方面, 根据参考图像划分, 选取 80% 的参考图像所对应的失真图像用于模型训练, 剩余图像用于测试, 从而保证测试集的图像内容不在训练过程中出现。

3　实验与分析

3.1　实验数据集

分别在压缩的虚拟现实图像质量 (compressed virtual reality image quality, CVIQ)[31] 数据集和 OIQA[32] 全景图像数据集上验证所提模型性能。表 1 总结了各数据集的具体信息, 包括参考图像数量、失真图像数量、失真类型、分辨率、主观实验人数、量表等级及所使用的头显设备。

表 1　CVIQ 和 OIQA 数据集信息统计

Table 1　Information on CVIQ and OIQA datasets

数据集	参考图像数量	失真图像数量	失真类型	分辨率	主观实验人数	量表等级	头显设备
CVIQ	16	528	JPEG/AVC/HEVC	4 096×2 048	20	10	HTC Vive
OIQA	16	320	JPEG/JPEG2000/GB/WGN	11 332×5 666/13 322×6 661	20	10	HTC Vive

CVIQ 数据集[31] 包含 16 幅参考图像和 528 幅对应的失真图像。失真类型均为压缩失真, 包括 JPEG 压缩、H.264(AVC) 压缩和 H.265(HEVC) 压缩。

OIQA 数据集[32] 包含 16 幅参考图像和 320 幅失真图像。失真类型包括 JPEG 压缩失真、JPEG2000 压缩失真、高斯模糊(Gaussian blur, GB)和白高斯噪声(white Gaussian noise, WGN), 每种类型包含 5 种失真等级。

CVIQ 和 OIQA 数据集采用 ITU-R BT.500-11[33] 定义的单刺激方法采集并分析主观评价人员给出的质量评分, 共 20 名被试参加了主观实验, 采用 10 分离散制打分。

3.2　评价指标和对比算法

3.2.1　评价指标

采用图像质量评价领域常用的 2 个评价指标比较不同算法的性能, Pearson 线性相关系数(Pearson linear correlation coefficient, PLCC)用于评估算法预测准确性, Spearman 秩相关系数(Spearman rank order correlation coefficient, SRCC)用于评估算法预测单调性, 2 个指标的取值范围均为 [−1,1], 其值越接近 1, 说明预测值与真实值的一致程度越高, 具体计算方法如下:

$$M_{\mathrm{PLCC}}=\frac{\sum_{i=1}^{N_{\mathrm{S}}}(Q_i-\bar{Q})(G_i-\bar{G})}{\sqrt{\sum_{i=1}^{N_{\mathrm{S}}}(Q_i-\bar{Q})^2\sum_{i=1}^{N_{\mathrm{S}}}(G_i-\bar{G})^2}} \tag{7}$$

$$M_{\mathrm{SRCC}}=1-\frac{6\sum_{i=1}^{N_{\mathrm{S}}}\delta_i^2}{N_{\mathrm{S}}(N_{\mathrm{S}}^2-1)} \tag{8}$$

式中: $\bar{G}$和$\bar{Q}$分别为主观质量分数平均值和客观质

量分数平均值；N_S为样本数量；δ_i为第i幅图像主观质量分数排名与客观质量分数排名的差值。

3.2.2 对比算法

选取9个有代表性的图像质量评价算法进行性能对比，包括普通图像全参考质量评价算法PSNR、结构相似度 (structural similarity index measure, SSIM)[34]、全景图像全参考质量评价算法WS-PSNR[11]、多尺度结构相似度 (multiscale structural similarity index measure, MS-SSIM)[35]、普通图像无参考质量评价算法盲/无参考图像空间质量评估器 (blind/referenceless image spatial quality evaluator, BRISQUE)[36]、自然图像质量评估器 (natural image quality evaluator, NIQE)[37]、全景图像无参考质量评价算法用于盲360°图像质量评估的多通道卷积神经网络 (multi-channel convolution neural network for blind 360° image quality assessment, MC360IQA)[8]、MPFIQA[13]和自适应超图卷积网络 (adaptive hypergraph convolutional network, AHGCN)[17]。

3.3 实验结果

3.3.1 整体性能比较

表2为各模型在CVIQ和OIQA数据集上的性能比较。可知，本文设计的YOFONet在2个数据集上均取得了最高的PLCC和SRCC值。在CVIQ数据集上，本文模型相较于传统方法SSIM，PLCC和SRCC分别提高了9.69%和7.43%；相较于最新评价方法多感知特征图像质量评价 (multi-perceptual features image quality assessment, MPFIQA)[13]，PLCC和SRCC分别提高了1.94%和1.83%；相较于当前最优方法AHGCN，PLCC和SRCC分别提高了0.59%和0.23%。在OIQA数据集上，本文模型相较于传统方法SSIM，PLCC和SRCC分别提高了9.36%和9.97%；相较于最新评价方法MPFIQA，PLCC和SRCC分别提高了1.47%和1.05%；相较于当前最优方法AHGCN，PLCC和SRCC分别提高了0.43%和0.29%。上述实验结果表明，相比于现有代表性算法，YOFONet的性能均有提升，验证了本文所设计特征共享和多视口自适应融合方法的有效性。

表2 CVIQ和OIQA数据集上模型性能比较

Table 2 Performance comparison of models on CVIQ and OIQA datasets

类型	算法	PLCC		SRCC	
		CVIQ	OIQA	CVIQ	OIQA
全参考	PSNR	0.732 0	0.691 8	0.766 2	0.680 2
	WS-PSNR[11]	0.749 8	0.698 5	0.775 5	0.693 2
	SSIM[34]	0.885 7	0.889 2	0.897 2	0.879 8
	MS-SSIM[35]	0.876 2	0.842 7	0.887 5	0.833 2
无参考	NIQE[37]	−0.512 6	0.470 1	0.532 9	−0.330 6
	BRISQUE[36]	−0.744 8	0.777 8	0.764 1	−0.779 4
	MC360IQA[8]	0.950 6	0.924 7	0.913 9	0.918 7
	MPFIQA[13]	0.953 0	0.958 3	0.946 6	0.957 4
	AHGCN[17]	0.965 8	0.968 2	0.961 7	0.964 7
	本文	**0.971 5**	**0.972 4**	**0.963 9**	**0.967 5**

注：粗体表示性能最好的模型。

3.3.2 消融实验

为了系统地验证本文所提模型各模块对模型预测性能的影响，在CVIQ数据集上进行消融实验，训练集-测试集采用50%-50%的划分方式，结果如表3所示。其中，基准模型“Baseline”直接将ResNet-50的输出回归得到质量分数，没有使用Transformer聚合质量特征；“Model1”使用ResNet-50的输出，不进行视口分割，直接作为Transformer的输入；“Model2”在“Model1”的基础上加入多尺度特征融合模块，将多尺度特征作为Transformer的输入；“Model3”在“Model2”的基础上加入本文提出的语义-质量聚合模块，分别将高级语义特征和多尺度质量特征作为第1层编码器的$\boldsymbol{q}$、$\boldsymbol{k}$和$\boldsymbol{v}$；YOFONet在“Model3”的基础上加入非均匀视口分割模块，在特征域分割局部视口特征，为本文提出的完整模型；“Model4”同YOFONet的整体结构一致，但Transformer结构中以质量特征作为$\boldsymbol{q}$，以语义特征作为$\boldsymbol{k}$。

表3 消融实验结果（CVIQ数据集）

Table 3 Ablation experiment results (CVIQ dataset)

名称	模块			PLCC	SRCC
	多尺度特征融合	语义-质量聚合	非均匀视口分割		
Baseline	×	×	×	0.920 1	0.884 9
Model1	×	×	×	0.940 0	0.908 0
Model2	√	×	×	0.944 4	0.917 9
Model3	√	√	×	0.959 0	0.935 4
Model4	√	√	√	0.962 7	0.945 5
YOFONet	√	√	√	**0.967 4**	**0.950 6**

注：粗体表示性能最好。

从表3可以看出：①“Model2”的PLCC和SRCC均高于“Model1”，表明融合不同尺度的特征可提升模型性能；②“Model3”的2个指标均高于“Model2”，说明以语义特征与质量特征的相关性来融合局部质量特征，可有效建模人类视觉系统中语义识别和感知质量2个任务的内在联系，从而提升模型性能；③YOFONet的2个指标均高于“Model3”，表明在特征域分割视口的方式可以很好地模拟用户观看方式，此外，本文所设计的分割方法可有效应对ERP投影特性；④由于注意力机制计算过程中$\boldsymbol{q}$和$\boldsymbol{k}$是等价的，置换Transformer结构中$\boldsymbol{q}$和$\boldsymbol{k}$的赋值并不会显著影响模型性能，表3的实验结果同分析相

符, 这也进一步验证了本文方法的稳定性。

为进一步验证所提出 YOFONet 的有效性, 利用群体最大差异化(group maximum differentiation, gMAD)竞争方法[38-39]比较基准模型“Baseline”和 YOFONet 的性能差异, 如图 5 所示。在 gMAD 竞争中, 可以自动选择出不同质量等级下防守者认为质量相同而攻击者预测质量差异最大的图像对。由图 5(a) 和图 5(b) 可以看出, 当 YOFONet 作为防守者, “Baseline”作为攻击者时, “Baseline”未能找到具有不同质量的图像对, 而 YOFONet 预测图像对具有相同或相似的质量, 与人类的质量感知高度一致。在图 5(c) 和图 5(d) 中, YOFONet 预测顶部图像的质量优于底部图像, 而“Baseline”认为它们具有相同的质量, 可以看出, 作为攻击者的 YOFONet 搜寻到反例成功攻击了防守者。由此可见, 在不同质量等级下, YOFONet 不管作为防守者还是攻击者, 其预测结果相较于“Baseline”同人类视觉系统更一致。



图 5　YOFONet 与基准模型的 gMAD 竞争结果

Fig. 5　gMAD competition between YOFONet and baseline models

3.3.3　算法复杂度分析

在算法复杂度方面, 若各视口采用独立的特征提取方法, 则当视口分辨率为 448×448 时, 每个视口在前馈过程中的浮点运算次数(floating point operations, FLOPs)为 1.82×10^{10}, 即当选取 R 个视口时, 由于各视口需分别进行前馈操作以提取特征, 则其浮点运算次数为 $R\times1.82\times10^{10}$。与之相比, YOFONet 在 $K=128$ 时, 其 FLOPs 为 4.75×10^{10}, 因此本文所设计的特征共享的质量评价方法可有效降低算法复杂度, 实现更为轻量级的局部特征提取模型。

此外, 由于本文采用 ROI_Align 方法提取质量特征, 其视口位置作为单独的配置信息可根据实际应用场景的变化灵活调整, 使本文模型具有更广泛的应用前景。

3.3.4　Transformer 注意力权重可视化

为了更直观地探索 Transformer 中注意力的工作机制, 对编码器各层注意力权重进行可视化, 结果如图 6 所示, 其中, 第 1 列图像为高斯模糊, 第 2 列图像为 JPEG2000 压缩, 其分析了图中各区域的注意力权重变化及局部质量。

以图 6 中 $\boldsymbol{A}$ 和 $\boldsymbol{A}'$ 这 2 个对应区域为例说明注意力占比变化和局部质量的计算方法。记原图像和失真图像分别为 $\boldsymbol{I}$ 和 $\boldsymbol{I}'$, 则注意力权重变化由 $(S_{\text{att}}(\boldsymbol{A})\cdot S_{\text{att}}(\boldsymbol{I}'))/(S_{\text{att}}(\boldsymbol{A}')S_{\text{att}}(\boldsymbol{I}))$ 得出, 其中, $S_{\text{att}}(\cdot)$ 表示计算某区域的注意力权重矩阵并求和, 该值可以反应局

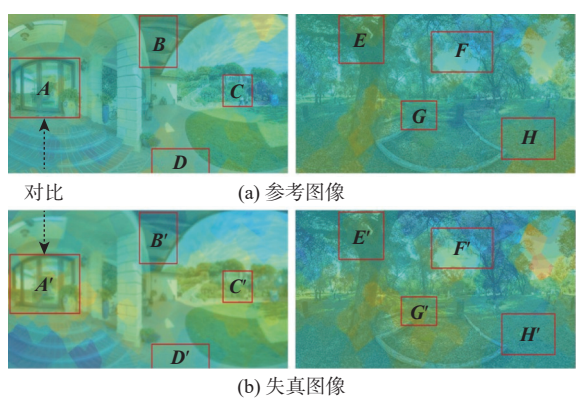


图 6　Transformer 注意力权重可视化

Fig. 6　Visualization of Transformer attention weights

部区域注意力占比在原始图像和失真图像间的变化，其大于1时说明原始图像对应区域注意力占比更高，反之说明失真图像对应区域占比更高。而局部质量分值Q_{local}通过计算$\boldsymbol{A}$和$\boldsymbol{A}'$间的梯度幅度相似性偏差 (gradient magnitude similarity deviation, GMSD) 分值[40]$Q_{\text{local}}^{\text{GMSD}}$后归一化得出，即$Q_{\text{local}}=\left(Q_{\max}-Q_{\text{local}}^{\text{GMSD}}\right)/(Q_{\max}-Q_{\min})$，其中，$Q_{\max}$和$Q_{\min}$分别为统计多个局部区域$Q_{\text{local}}^{\text{GMSD}}$值所得的最大和最小值，$Q_{\text{local}}$值越高，说明局部质量越好。

由图6可以看出，区域$\boldsymbol{A}$与$\boldsymbol{A}'$、$\boldsymbol{C}$与$\boldsymbol{C}'$、$\boldsymbol{E}$与$\boldsymbol{E}'$、$\boldsymbol{H}$与$\boldsymbol{H}'$的注意力权重占比大于1，局部质量分值在0.7～0.9之间，而其他区域的注意力权重占比均小于1，局部质量分值在0.1～0.4之间。可知，当局部质量较好时，原始图像对应区域具有更多的注意力权重占比，而随着质量劣化，注意力向失真明显的区域转移，表明局部质量特征聚合模块可有效识别图像中失真明显的位置，并在汇聚过程中赋予其更高的权重。

3.3.5 投影畸变对不同分辨率的影响

分析面向普通图像的深度学习模型对于全景图像投影畸变的敏感程度，如图7所示，物体A相较于物体B具有更明显的投影畸变，将该图像以不同的分辨率送入物体识别网络YOLOv3[41]，可以看到，分辨率较高时，投影畸变对识别置信度的影响不大，但随着分辨率降低，具有几何畸变物体的识别置信度显著下降。

由上述分析可知，当图像分辨率较小时，全景图像的几何畸变显著影响了深度学习模型的特征表示，与之相反，当图像分辨率较大时，这种影响较为轻微。由于深度学习模型的感受野是固定的，图像分辨率越大，相对而言，其感受野的观察尺度越小，此时几何畸变也更加不明显，因此可有效解释上述现象。综上分析，本文以较大分辨率图像送入骨干网络提取相应特征，在尽可能避免几何畸变对模型表示能力影响的同时，得到各视口在特征域的统一描述。

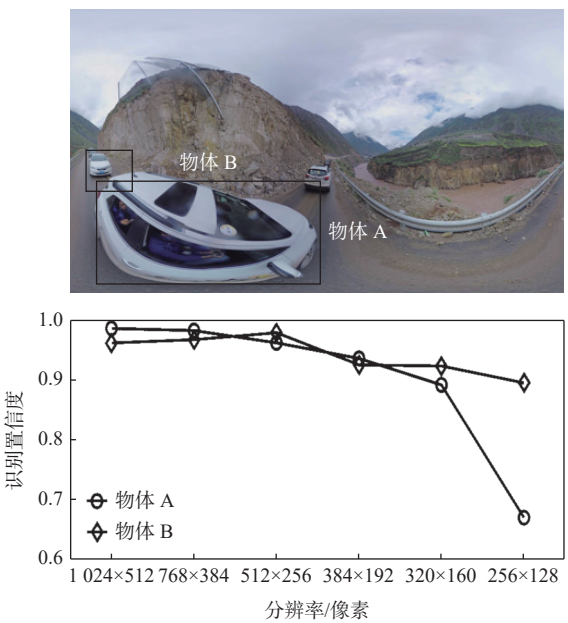


图7 投影畸变对识别模型性能的影响

Fig. 7 Influence of projection distortion on recognition model performance

为了进一步验证上述观察结果，在CVIQ数据集上进行实验，分别观察YOFONet面向不同输入分辨率的预测性能，训练集和测试集采用50%-50%的划分方式，结果如图8所示。可以看出，当分辨率较高时（≥1 024×512像素），由于几何畸变不会对特征提取造成显著影响，其预测准确性均较高，但随着分辨率下降，骨干网络的感受野范围相较于整个图像越来越大，几何畸变会对特征提取过程造成显著影响，导致预测性能降低。

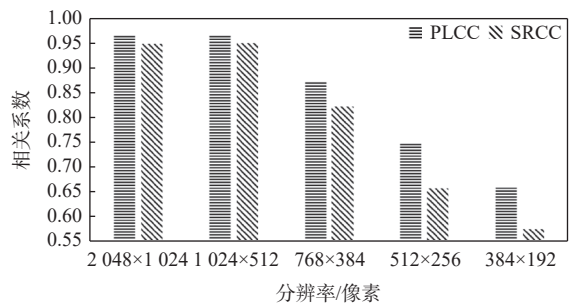


图8 不同分辨率下YOFONet的性能比较

Fig. 8 Performance of YOFONet at different resolutions

4 结 论

1）本文提出的YOFONet输出的全景图像质量预测结果同人类视觉系统具有较高一致性，其在CVIQ和OIQA数据集上的PLCC和SRCC均达0.96以上。

2）现有方法特征提取过程的计算复杂度随着视口数量线性增加，而本文方法由于采用共享特征提取方法，其局部特征提取模块计算复杂度固定，在视口数量较多时可实现更为轻量级的计算。

3）消融实验表明，本文提出的多尺度特征融合、基于球面均匀采样的视口分割及语义引导的局部质量特征聚合方法均能有效提高质量评价模型的预测准确性。

4）本文将在后续工作中通过自行采集数据集、组织主观评价实验等方法，进一步测试所设计方法的跨数据库预测性能，并结合光流、帧差、时空切片等时域信息，探索YOFONet面向视频质量评价的适配方法。

参考文献（References）

[1] 叶成英, 李建微, 陈思喜. VR 全景视频传输研究进展[J]. 计算机应用研究, 2022, 39(6): 1601-1607.
YE C Y, LI J W, CHEN S X. Research progress of VR panoramic video transmission[J]. Application Research of Computers, 2022, 39(6): 1601-1607(in Chinese).

[2] ZHAI G T, MIN X K. Perceptual image quality assessment: a survey[J]. Science China Information Sciences, 2020, 63(11): 211301.

[3] 曾焕强, 丁瑞, 黄海靓, 等. 沉浸式视频质量评价综述[J]. 信号处理, 2022, 38(9): 1853-1862.
ZENG H Q, DING R, HUANG H L, et al. Immersive video quality assessment: an overview[J]. Journal of Signal Processing, 2022, 38(9): 1853-1862(in Chinese).

[4] LIU Y, YU H W, HUANG B Q, et al. Blind omnidirectional image quality assessment based on structure and natural features[J]. IEEE Transactions on Instrumentation and Measurement, 2021, 70: 5014011.

[5] JIANG H, JIANG G Y, YU M, et al. Cubemap-based perception-driven blind quality assessment for 360-degree images[J]. IEEE Transactions on Image Processing, 2021, 30: 2364-2377.

[6] TRUONG T Q, TRAN H T T, THANG T C. Non-reference quality assessment model using deep learning for omnidirectional images[C]//Proceedings of the IEEE 10th International Conference on Awareness Science and Technology. Piscataway: IEEE Press, 2019: 1-5.

[7] XU J H, ZHOU W, CHEN Z B. Blind omnidirectional image quality assessment with viewport oriented graph convolutional networks[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2021, 31(5): 1724-1737.

[8] SUN W, MIN X K, ZHAI G T, et al. MC360IQA: a multi-channel CNN for blind 360-degree image quality assessment[J]. IEEE Journal of Selected Topics in Signal Processing, 2020, 14(1): 64-77.

[9] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, et al. An image is worth 16×16 words: transformers for image recognition at scale[C]//The 9th International Conference on Learning Representations. [S.l.:s.n.], 2021: 1-21.

[10] YU M, LAKSHMAN H, GIROD B. A framework to evaluate omnidirectional video coding schemes[C]//Proceedings of the IEEE International Symposium on Mixed and Augmented Reality. Piscataway: IEEE Press, 2015: 31-36.

[11] YULE S, LU A, LU Y. Weighted-to-spherically-uniform quality evaluation for omnidirectional video[J]. IEEE Signal Processing Letters, 2017, 24(9): 1408-1412.

[12] ZAKHARCHENKO V, CHOI K P, PARK J H. Quality metric for spherical panoramic video[C]//Proceedings of the Optics and Photonics for Information Processing X. Bellingham: SPIE Press, 2016, 9970: 57-65.

[13] LIU Y, YIN X H, WAN Z L, et al. Toward a no-reference omnidirectional image quality evaluation by using multi-perceptual features[J]. ACM Transactions on Multimedia Computing, Communications, and Applications, 2023, 19(2): 1-19.

[14] KIM H G, LIM H T, RO Y M. Deep virtual reality image quality assessment with human perception guider for omnidirectional image[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2020, 30(4): 917-928.

[15] BOSSE S, MANIRY D, MÜLLER K R, et al. Deep neural networks for no-reference and full-reference image quality assessment [J]. IEEE Transactions on Image Processing, 2018, 27(1): 206-219.

[16] MA K D, LIU W T, ZHANG K, et al. End-to-end blind image quality assessment using deep neural networks[J]. IEEE Transactions on Image Processing, 2018, 27(3): 1202-1213.

[17] FU J, HOU C, ZHOU W, et al. Adaptive hypergraph convolutional network for no-reference 360-degree image quality assessment[C]//Proceedings of the 30th ACM International Conference on Multimedia. New York: ACM, 2022: 961-969.

[18] ZHOU W, XU J H, JIANG Q P, et al. No-reference quality assessment for 360-degree images by analysis of multifrequency information and local-global naturalness[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2022, 32(4): 1778-1791.

[19] HE K M, ZHANG X Y, REN S Q, et al. Spatial pyramid pooling in deep convolutional networks for visual recognition[C]//European Conference on Computer Vision. Berlin: Springer, 2014: 346-361.

[20] KE J J, WANG Q F, WANG Y L, et al. MUSIQ: multi-scale image quality transformer[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE Press, 2021: 5128-5137.

[21] SU S L, YAN Q S, ZHU Y, et al. Blindly assess image quality in the wild guided by a self-adaptive hyper network[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2020: 3664-3673.

[22] DENG J, DONG W, SOCHER R, et al. ImageNet: a large-scale hierarchical image database[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2009: 248-255.

[23] HE K, GKIOXARI G, DOLLAR P, et al. Mask R-CNN[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE Press, 2017: 2961-2969.

[24] SAFF E B, KUIJLAARS A B J. Distributing many points on a sphere[J]. The Mathematical Intelligencer, 1997, 19(1): 5-11.

[25] LI D Q, JIANG T T, LIN W S, et al. Which has better visual quality: the clear blue sky or a blurry animal?[J]. IEEE Transactions on Multimedia, 2019, 21(5): 1221-1234.

[26] MA X Y, WANG Y Q, LIU C, et al. ADGNet: attention discrepancy guided deep neural network for blind image quality assessment[C]//Proceedings of the 30th ACM International Conference on Multimedia. New York: ACM, 2022: 1309-1318.

[27] GLOROT X, BORDES A, BENGIO Y. Deep sparse rectifier neural networks[C]//Proceedings of the 14th International Conference on Artificial Intelligence and Statistics. Ft. Lauderdale: Journal of Machine Learning Research Workshop and Conference Proceedings, 2011: 315-323.

[28] HE K M, ZHANG X Y, REN S Q, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2016: 770-778.

[29] KINGMA D P, BA J. Adam: a method for stochastic optimization [EB/OL].(2017-01-30)[2023-04-10].https://arxiv.org/abs/1412.6980.

[30] SRIVASTAVA N, HINTON G, KRIZHEVSKY A, et al. Dropout: a simple way to prevent neural networks from overfitting[J]. Journal of Machine Learning Research, 2014, 15: 1929-1958.

[31] SUN W, GU K, MA S W, et al. A large-scale compressed 360-degree spherical image database: from subjective quality evaluation to objective model comparison[C]//Proceedings of the IEEE 20th International Workshop on Multimedia Signal Processing. Piscataway: IEEE Press, 2018: 1-6.

[32] DUAN H Y, ZHAI G T, MIN X K, et al. Perceptual quality assessment of omnidirectional images[C]//Proceedings of the IEEE International Symposium on Circuits and Systems. Piscataway: IEEE Press, 2018: 1-5.

[33] ITU-R. Methodology for the subjective assessment of the quality of television pictures: ITU-R BT. 500-11[S]. Geneva: International Telecommunication Union, 2002: 2-30.

[34] WANG Z, BOVIK A C, SHEIKH H R, et al. Image quality assessment: from error visibility to structural similarity[J]. IEEE Transactions on Image Processing, 2004, 13(4): 600-612.

[35] WANG Z, SIMONCELLI E P, BOVIK A C. Multiscale structural similarity for image quality assessment[C]//The 37th Asilomar Conference on Signals, Systems & Computers. Piscataway: IEEE Press, 2003: 1398-1402.

[36] MITTAL A, MOORTHY A K, BOVIK A C. No-reference image quality assessment in the spatial domain[J]. IEEE Transactions on Image Processing, 2012, 21(12): 4695-4708.

[37] MITTAL A, SOUNDARARAJAN R, BOVIK A C. Making a "completely blind" image quality analyzer[J]. IEEE Signal Processing Letters, 2013, 20(3): 209-212.

[38] MA K D, DUANMU Z F, WANG Z, et al. Group maximum differentiation competition: model comparison with few samples[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020, 42(4): 851-864.

[39] MA K D, WU Q B, WANG Z, et al. Group MAD competition? a new methodology to compare objective image quality models[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2016: 1664-1673.

[40] XUE W F, ZHANG L, MOU X Q, et al. Gradient magnitude similarity deviation: a highly efficient perceptual image quality index[J]. IEEE Transactions on Image Processing, 2014, 23(2): 684-695.

[41] REDMON J, FARHADI A. YOLOv3: an incremental improvement[EB/OL]. (2018-04-08)[2023-04-10]. https://arxiv.org/abs/1804.02767.

Omnidirectional image quality assessment based on adaptive multi-viewport fusion

FENG Chenxi[1], ZHANG Di[1, 2], LIN Gan[3], YE Long[2, 4, *]

(1. State Key Laboratory of Media Convergence and Communication, Communication University of China, Beijing 100024, China;
2. School of Data Science and Intelligent Media, Communication University of China, Beijing 100024, China;
3. Hainan International College, Communication University of China, Lingshui 572400, China;
4. Key Laboratory of Media Audio & Video (Communication University of China), Ministry of Education, Beijing 100024, China)

Abstract: Existing omnidirectional image quality assessment (OIQA) models extract local features from each viewport independently, increasing computational complexity and making it difficult to describe the correlations between viewports using an end-to-end fusion model. To solve these issues, a quality assessment method was proposed based on feature sharing and adaptive fusion of multiple viewports. By utilizing shared backbone networks, the method transformed the viewport segmentation and computation that were independent of each other to the feature domain, enabling local feature extraction of the image through one-shot feed-forward computation. In addition, a viewport segmentation method in the feature domain using spherical uniform sampling was employed to guarantee consistent pixel density between view space and observation space, with semantic information guiding the adaptive fusion of local quality features of viewpoints. The Pearson linear correlation coefficient (PLCC) and Spearman rank order correlation coefficient (SRCC) on the compressed virtual reality image quality (CVIQ) and OIQA datasets were both above 0.96, showing superior performance compared with mainstream evaluation methods. Compared with the traditional evaluation method structural similarity index measure (SSIM), its average PLCC and average SRCC on the above two datasets were improved by 9.52% and 8.7%, respectively; compared with the latest evaluation method multi-perceptual features image quality assessment (MPFIQA), its average PLCC and average SRCC were improved by 1.71% and 1.44%, respectively.

Keywords: omnidirectional image; no-reference image quality assessment; spherical uniform sampling; adaptive feature fusion; equirectangular projection

Received: 2023-06-16; **Accepted**: 2023-12-15; **Published Online**: 2024-02-01 15:38

URL: link.cnki.net/urlid/11.2625.V.20240131.1721.002

Foundation items: National Key Research and Development Program of China (2021YFF0900503); National Natural Science Foundation of China (61971383,62001432); Beijing Natural Science Foundation (4244109)

* **Corresponding author**. E-mail: yelong@cuc.edu.cn

http://bhxb.buaa.edu.cn jbuaa@buaa.edu.cn
DOI: 10.13700/j.bh.1001-5965.2023.0329

基于小波变换和平行注意力的多源遥感图像分类

王嘉毅，高峰*，张天戈，甘言海
(中国海洋大学 计算机科学与技术学院，青岛 266100)

摘　　要：充分挖掘多源遥感图像数据特征的依赖关系，实现不同模态图像数据间的优势互补，已成为遥感领域的研究热点方向之一。现有的高光谱和合成孔径雷达(SAR)数据联合分类任务存在图像特征提取和特征表达不充分的问题，高频信息容易损失，不利于后续的分类任务，以及多源图像特征交互有限，多模态特征关联不紧密的关键难题。针对上述问题，围绕图像特征的鲁棒表达和多源特征的高效关联开展研究，提出了基于小波变换和平行注意力机制的多源遥感图像分类网络(WPANet)。基于小波变换的特征提取器可以充分利用频域分析技术，在可逆下采样的过程中充分捕捉粗/细粒度级别特征；基于平行注意力机制的特征融合器充分综合多模态遥感数据的一致性和差异性，完成强相关性特征的融合和生成，以提升分类准确度。在Augsburg和Berlin这2个真实多源遥感数据集上的实验表明：所提分类方法具有显著优势，总体准确率分别达到90.40%和76.23%，相比于深度特征交互网络(DFINet)等主流方法，在2个数据集上的总体准确率分别至少提升2.66%和12.22%。

关 键 词：高光谱图像；合成孔径雷达；小波变换；多源特征融合；遥感图像
中图分类号：TP753
文献标志码：A　　**文章编号**：1001-5965(2025)07-2415-08

近年来，中国的卫星遥感技术发展日新月异，多源遥感卫星数据，如高光谱、合成孔径雷达(synthetic aperture radar, SAR)、红外、可见光等，已经应用于国民经济、社会生活与国家安全的各个方面，初步形成了遥感卫星应用体系。遥感图像分类技术通过对遥感图像中的地物特征进行捕捉、加工和分类，达到对图像中多种地物信息进行识别的目的[1]。其中，高光谱传感器能够获取地物丰富的光谱信息，捕获反映物体材质的差异信息[2]。SAR传感器能够全天时、全天候工作，有效弥补恶劣天气(云雾遮挡、雾霾等)情况下高光谱传感器观测的局限性[3]。然而，SAR传输图像往往受到相干斑噪声的影响，目标细节模糊，视觉解译难度大。高光谱图像(hyperspectral image, HSI)相比于SAR图像具有光谱信息丰富、目标结构明显等特点，但容易受光谱、云雾、阴影等因素的影响。融合HSI和SAR图像的特征用于地物分类，能够将多源遥感信息优势互补，实现地物信息更全面、更丰富的解译。

HSI和SAR图像成像机理的差异导致其数据存在较大的差异，多源图像特征的关联非常困难。针对多源遥感图像分类任务，国内外学者提出了很多方法。Man等[4]采用像素级特征融合分类器，应用极大似然分类和支持向量机(support vector machine, SVM)等方法，实现了HSI和激光雷达(light detection and ranging, LiDAR)数据的融合分类。Hu等[5]提出了一种基于对象的融合方法，用于联合使用极化SAR和HSI数据的地物分类任务。Chen等[6]

收稿日期：2023-06-09；**录用日期**：2023-09-15；**网络出版时间**：2023-11-07 16:05
网络出版地址：link.cnki.net/urlid/11.2625.V.20231107.0859.001
基金项目：国家自然科学基金(42106191)；科技创新2030-新一代人工智能重大项目(2022ZD0117202)
* **通信作者**. E-mail：gaofeng@ouc.edu.cn

引用格式：王嘉毅，高峰，张天戈，等. 基于小波变换和平行注意力的多源遥感图像分类[J]. 北京航空航天大学学报，2025，51(7)：2415-2422.
WANG J Y，GAO F，ZHANG T G，et al. Multi-source remote sensing image classification based on wavelet transform and parallel attention[J]. Journal of Beijing University of Aeronautics and Astronautics，2025，51(7)：2415-2422 (in Chinese).

利用卷积神经网络(convolutional neural network, CNN)处理多源遥感数据,分别提取其光谱空间特征和高度特征,以获取深层次的特征信息并完成融合分类。Hong等[7]提出了深层编码器-解码器网络,利用强制融合单元对不同模态的特征进行重构,提升了多模态信息的利用率,更好地完成了多源遥感图像分类任务。Zhao等[8]提出了分层随机游走网络模型以更好地解决遥感图像融合分类问题,通过双隧道卷积神经网络对空间特征和光谱特征进行全局先验,采用像素级亲和分支以增强样本特征空间的一致性,从而实现多源遥感数据的高效融合分类。Feng等[9]提出了一种线性自注意力融合网络模型,以实现HSI和LiDAR图像的特征融合,该模型由特征提取单元、自注意力单元和特征融合单元构成,显著提高了多源遥感图像融合分类的精度和准确率。Li等[10]提出了一种基于非对称特征融合的分类方法,利用权重共享的剩余块进行特征提取,设计了特征校准模块,利用多源特征的空间依赖性提升识别能力,提高了多源遥感图像融合分类的准确率。

尽管国内外学者近年来围绕HSI和SAR融合分类技术开展了大量的研究工作,在多个国内外公开数据集上取得了较好的效果,但仍难以满足遥感应用的实际需求,主要因为以下2个技术难点亟待攻克:①图像中的高频信息容易损失。现有特征提取方法通常会使用池化、空洞卷积等操作来降低计算成本,但容易损失图像中的纹理细节。②多模态特征的交互有限。当前的多源图像特征融合方法大多简单拼接多源特征,或仅采用加权融合的方式,对不同模态特征交互信息的利用较为有限,亟需针对遥感场景的复杂性进行重构和优化。

为了解决以上2个问题,本文提出了基于小波变换和平行注意力机制的多源遥感图像分类网络(wavelet transform and parallel attention-based network for multi-source remote sensing image classification, WPANet)。具体而言,为了进一步充分挖掘HSI数据中的高频特征和低频特征,设计了基于小波变换的特征提取器,利用频域空间分析原理,充分捕捉多源图像特征之间的依赖关系。此外,设计了基于平行注意力机制的特征融合器,分别提取多源数据间的一致性和差异性特征。其中,一致性分析更加关注多源特征的一致性,抑制非显著信息的干扰;差异性分析更加关注多模态特征在显著区域的差异,从而实现多模态特征的对齐。在Augsburg和Berlin多源遥感数据集上的实验结果表明,WPANet在多源遥感图像地物分类任务中具有较好的效果表现。

本文的主要贡献如下:

1)在图像特征鲁棒表达方面,提出了基于小波变换的特征提取器,利用频域分析原理完成对粗/细粒度级别特征的提取,通过中间特征的可逆下采样,完成特征的充分挖掘。

2)在多源特征高效关联方面,构建了基于平行注意力的特征融合器,分别关注多源特征的一致性和差异性,充分融合多源数据的特征以得到更高质量的联合特征,从而实现地物分类准确度的有效提升。

3)在Augsburg和Berlin多源遥感数据集上进行了大量实验,从定量分析和定性分析2个角度证明了本文提出的WPANet模型的有效性及可扩展性。

1 本文方法

本文提出的WPANet模型整体架构如图1所

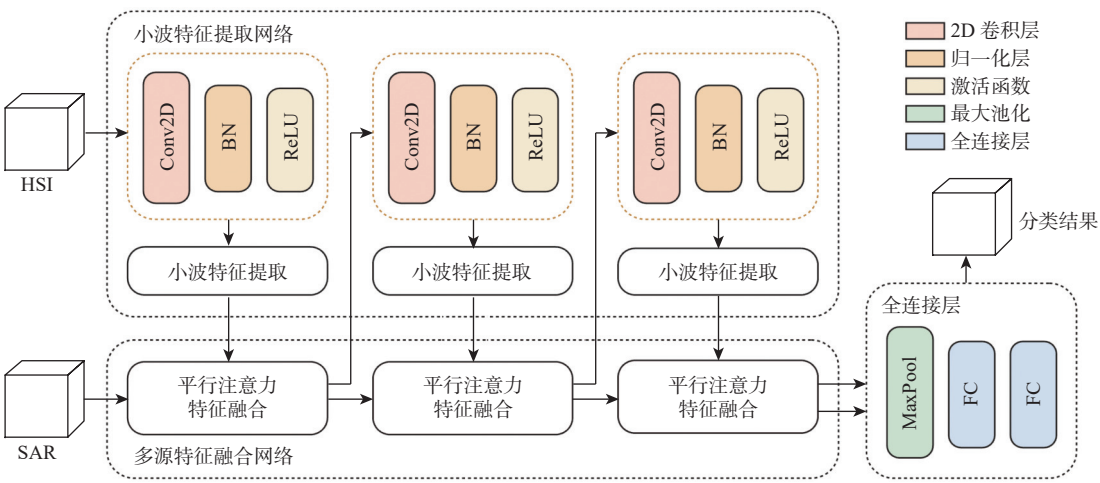


图1 基于小波变换和平行注意力机制的多源遥感图像分类网络

Fig. 1 Multi-source remote sensing image classification network based on wavelet transform and parallel attention mechanism

示，由小波特征提取网络、多源特征融合网络及全连接层组成。WPANet 主要包括 2 个关键模块：基于小波变换的特征提取模块和基于平行注意力机制的特征融合模块。

1.1　网络整体架构

选择常见的双流网络架构以完成跨模态信息特征的交互和融合。首先，给定 2 组输入信息：HSI 图像 $\boldsymbol{I}^{\mathrm{HSI}} \in \mathbf{R}^{H\times W\times C}$ 和 SAR 图像 $\boldsymbol{I}^{\mathrm{SAR}} \in \mathbf{R}^{H\times W\times C}$，其中，$H$、$W$、$C$ 分别为图像高度、宽度和通道数，将这 2 种不同模态的图像作为 WPANet 的输入信息。采用主成分分析（principal component analysis, PCA）方法对 $\boldsymbol{I}^{\mathrm{HSI}}$ 在通道维度进行降维处理以保留其关键信息，从而减少计算量，提升网络训练效率。

在小波特征提取网络中，$\boldsymbol{I}^{\mathrm{HSI}}$ 经过基于小波变换的特征提取模块，能够更好地实现图像特征在频率域的信息整合，进而更好地实现高频和低频特征的提取。在多源特征融合网络中，使用基于平行注意力的特征融合器，对经过特征提取的 $\boldsymbol{I}^{\mathrm{HSI}}$ 和上一级的 $\boldsymbol{I}^{\mathrm{SAR}}$ 完成跨尺度和跨模态的特征融合。最后，融合的特征输入全连接层处理，将已充分融合的高维度特征映射为不同的地物特征种类，使用 Softmax 函数进行归一化，得到最终的地物分类结果。

1.2　基于小波变换的特征提取器

小波变换特征提取器将小波变换引入传统的 Transformer 特征编码器中，利用频域空间分析原理，充分挖掘粗/细粒度级别的特征信息，从而更好地实现中间特征的可逆下采样。如图 2 所示，给定输入特征图 $\boldsymbol{I} \in \mathbf{R}^{H\times W\times C}$，通过可学习的嵌入矩阵 $\boldsymbol{W}_{\mathrm{d}} \in \mathbf{R}^{C\times C/4}$ 将其线性变换为 $\bar{\boldsymbol{I}}$：

$$\bar{\boldsymbol{I}} = \boldsymbol{I}\boldsymbol{W}_{\mathrm{d}} \in \mathbf{R}^{H\times W\times \frac{C}{4}} \tag{1}$$

此时，通道数缩小为原来的 1/4。通过离散小波变换（discrete wavelet transform, DWT）将 $\bar{\boldsymbol{I}}$ 分解为 4 个小波子带实现下采样。

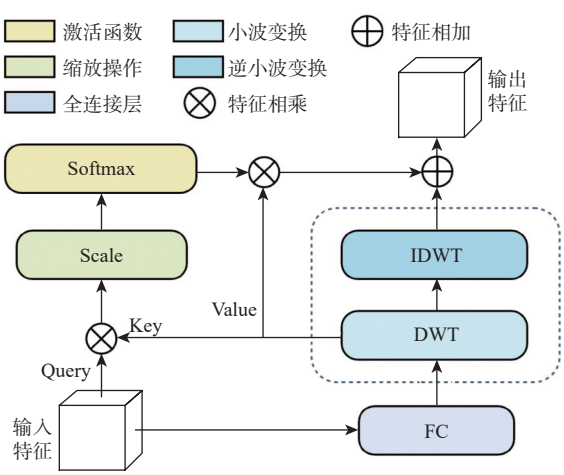


图 2　小波变换特征提取器

Fig. 2　Wavelet transform-based feature extractor

选择 Haar 小波函数，2×2 Haar 变换矩阵可表示为

$$\boldsymbol{H}_2 = \frac{1}{\sqrt{2}}\begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} \tag{2}$$

式中：$\boldsymbol{H}_2$ 的行可用于分别定义分析滤波器 $\boldsymbol{F}_{\mathrm{L}} = \left(1/\sqrt{2}, 1/\sqrt{2}\right)$ 和 $\boldsymbol{F}_{\mathrm{H}} = \left(1/\sqrt{2}, -1/\sqrt{2}\right)$。沿行应用低通滤波器 $\boldsymbol{F}_{\mathrm{L}}$ 和高通滤波器 $\boldsymbol{F}_{\mathrm{H}}$ 将 $\bar{\boldsymbol{I}}$ 分解为 2 个小波子带 $\boldsymbol{I}_{\mathrm{L}}$ 和 $\boldsymbol{I}_{\mathrm{H}}$，进一步沿列方向使用相同的低通滤波器 $\boldsymbol{F}_{\mathrm{L}}$ 和高通滤波器 $\boldsymbol{F}_{\mathrm{H}}$ 将 $\boldsymbol{I}_{\mathrm{L}}$ 和 $\boldsymbol{I}_{\mathrm{H}}$ 分解为 4 个小波子带：$\boldsymbol{I}_{\mathrm{LL}}$（原始的低频信息）、$\boldsymbol{I}_{\mathrm{LH}}$（水平方向的高频信息）、$\boldsymbol{I}_{\mathrm{HL}}$（垂直方向的高频信息）和 $\boldsymbol{I}_{\mathrm{HH}}$（对角线方向的高频信息）：

$$(\boldsymbol{I}_{\mathrm{LL}}, \boldsymbol{I}_{\mathrm{LH}}, \boldsymbol{I}_{\mathrm{HL}}, \boldsymbol{I}_{\mathrm{HH}}) = \mathrm{DWT}\left(\bar{\boldsymbol{I}}\right) \in \mathbf{R}^{\frac{H}{2}\times\frac{W}{2}\times\frac{C}{4}} \tag{3}$$

此时，每个小波子带可以看作 $\bar{\boldsymbol{I}}$ 经过无损下采样得到的特征信息，此过程中无特征信息丢失。如图 3 所示，$\bar{\boldsymbol{I}}$ 为原始的输入信息，经过 Haar 小波变换，被分解成 4 部分：$\boldsymbol{I}_{\mathrm{LL}}$、$\boldsymbol{I}_{\mathrm{LH}}$、$\boldsymbol{I}_{\mathrm{HL}}$ 和 $\boldsymbol{I}_{\mathrm{HH}}$。$\boldsymbol{I}_{\mathrm{LL}}$ 在粗粒度级别反映对象的基本结构特征，而 $\boldsymbol{I}_{\mathrm{LH}}$、$\boldsymbol{I}_{\mathrm{HL}}$ 和 $\boldsymbol{I}_{\mathrm{HH}}$ 在细粒度级别保留对象的纹理细节信息。

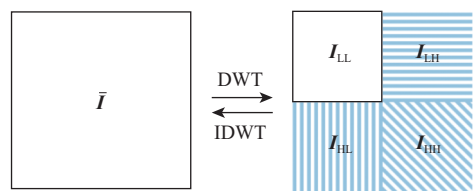


图 3　DWT 和 IDWT 过程

Fig. 3　DWT and IDWT process

沿通道维度拼接 4 个小波子带，得到频域综合特征 $\hat{\boldsymbol{I}}$：

$$\hat{\boldsymbol{I}} = \mathrm{Concat}(\boldsymbol{I}_{\mathrm{LL}}, \boldsymbol{I}_{\mathrm{LH}}, \boldsymbol{I}_{\mathrm{HL}}, \boldsymbol{I}_{\mathrm{HH}}) \in \mathbf{R}^{\frac{H}{2}\times\frac{W}{2}\times C} \tag{4}$$

将 $\hat{\boldsymbol{I}}$ 变换为经过下采样后的 Key（$\boldsymbol{K}^{\mathrm{w}}$）和 Value（$\boldsymbol{V}^{\mathrm{w}}$），而 Query（$\boldsymbol{Q}$）为原始输入特征图 $\boldsymbol{I}$。此时基于小波变换的多头自注意力计算在 $\boldsymbol{Q}$、$\boldsymbol{K}^{\mathrm{W}}$ 和 $\boldsymbol{V}^{\mathrm{w}}$ 上进行：

$$\boldsymbol{h}_i = \mathrm{Attention}\left(\boldsymbol{Q}_i, \boldsymbol{K}_i^{\mathrm{w}}, \boldsymbol{V}_i^{\mathrm{w}}\right) = \mathrm{Softmax}\left(\frac{\boldsymbol{Q}_i(\boldsymbol{K}_i^{\mathrm{w}})^{\mathrm{T}}}{\sqrt{D_{\mathrm{h}}}}\right)\boldsymbol{V}_i^{\mathrm{w}} \tag{5}$$

式中：下标 i 表示第 i 个头；D_{h} 为每个头的维度。由于多头自注意力的聚合性，$\boldsymbol{h}_i$ 为输入的长距离全局依赖特征信息。对 $\hat{\boldsymbol{I}}$ 应用逆小波变换（inverse discrete wavelet transform, IDWT）得到重构的特征图 $\boldsymbol{I}^{\mathrm{r}}$：

$$\boldsymbol{I}^{\mathrm{r}} = \mathrm{IDWT}\left(\hat{\boldsymbol{I}}\right) \in \mathbf{R}^{H\times W\times\frac{C}{4}} \tag{6}$$

根据小波变换理论，重构得到的特征图 $\boldsymbol{I}^{\mathrm{r}}$ 包含

原始输入细节，从而实现对局部特征信息的可逆下采样，更好地在多尺度信息上完成自注意力学习。小波特征提取器的整体操作过程表示如下：

$$\begin{aligned}&\text{Wavelet}(\boldsymbol{I})=\text{MultiHead}^{\text{w}}(\boldsymbol{Q},\boldsymbol{K}^{\text{w}},\boldsymbol{V}^{\text{w}},\boldsymbol{I}^{\text{r}})=\\&\quad\text{MultiHead}^{\text{w}}\left(\boldsymbol{I}\boldsymbol{W}^{\boldsymbol{Q}},\hat{\boldsymbol{I}}\boldsymbol{W}^{\boldsymbol{K}},\hat{\boldsymbol{I}}\boldsymbol{W}^{\boldsymbol{V}},\boldsymbol{I}^{\text{r}}\right)=\\&\quad\text{Concat}\left(\boldsymbol{h}_0,\cdots,\boldsymbol{h}_{N_{\text{h}}},\boldsymbol{I}^{\text{r}}\right)\boldsymbol{W}^{\text{O}}\end{aligned}\tag{7}$$

式中：N_{h}为头的数量；$\boldsymbol{W}^{\boldsymbol{K}}$、$\boldsymbol{W}^{\boldsymbol{V}}$和$\boldsymbol{W}^{\boldsymbol{Q}}$分别为对应$\boldsymbol{K}^{\text{w}}$、$\boldsymbol{V}^{\text{w}}$和$\boldsymbol{Q}$的中间矩阵；$\boldsymbol{W}^{\text{O}}$为变换矩阵。与常见的注意力模块相比，在加入小波变换之后，小波特征提取器中的DWT-IDWT过程可以在不计计算成本的前提下，将特征信息进行强相关的局部上下文化处理，其注意力模块具有更大的感受野，能够更好地实现局部特征信息和上下文之间的交互。

1.3 基于平行注意力机制的特征融合器

基于平行注意力机制的特征融合器可以更好地将不同层级的2种模态图像特征在深层次完成特征信息的交互和融合。经过平行注意力特征融合器的计算，类间特征的差异性得以增强，对应局部差异的细节纹理得以保留；类内特征的一致性可以更充分地保证2种模态图像特征完成高效的综合。

如图4所示，不同层级的2种模态图像特征$\boldsymbol{F}_{\text{HSI}}$和$\boldsymbol{F}_{\text{SAR}}$首先采用逐元素乘法，得到初步的一致性特征信息$\boldsymbol{F}_{\text{HSI}}\otimes\boldsymbol{F}_{\text{SAR}}$。然后将所得到的初步一致性特征经过通道注意力(channel attention, CA)的计算，在通道维度上完成注意力的聚合，以关注图像特征信息量较大的通道，抑制相关性较低的通道特征信息[11]，得到通道注意力权重$\boldsymbol{W}$：

$$\boldsymbol{W}=\text{CA}(\boldsymbol{F}_{\text{HSI}}\otimes\boldsymbol{F}_{\text{SAR}})\tag{8}$$

利用$\boldsymbol{W}$来增强原始的2个模态图像的显著性特征信息，通过逐元素加法完成一致性联合特征$\boldsymbol{F}_{\text{C}}$的计算：

$$\boldsymbol{F}_{\text{C}}=(\boldsymbol{W}\otimes\boldsymbol{F}_{\text{HSI}})\oplus(\boldsymbol{W}\otimes\boldsymbol{F}_{\text{SAR}})\tag{9}$$

$\boldsymbol{F}_{\text{C}}$在跨层级和跨模态的图像特征信息中实现了特征一致性的显著提升。差异性联合特征$\boldsymbol{F}_{\text{D}}$描述了图像特征的差异信息，跨模态图像的类间差异得以保留和强化，对于跨模态图像特征融合有高效的提升作用。$\boldsymbol{F}_{\text{D}}$关注类间差异，其定义为

$$\boldsymbol{F}_{\text{D}}=|\boldsymbol{W}\otimes\boldsymbol{F}_{\text{HSI}}\ominus\boldsymbol{W}\otimes\boldsymbol{F}_{\text{SAR}}|\tag{10}$$

式中：$\ominus$表示逐元素减法；$|\cdot|$表示逐元素计算绝对值。通过结合一致性联合特征$\boldsymbol{F}_{\text{C}}$和差异性联合特征$\boldsymbol{F}_{\text{D}}$得到平行注意力特征融合器输出的融合信息$\boldsymbol{F}_{\text{CD}}$：

$$\boldsymbol{F}_{\text{CD}}=\boldsymbol{F}_{\text{C}}\oplus\boldsymbol{F}_{\text{D}}\tag{11}$$

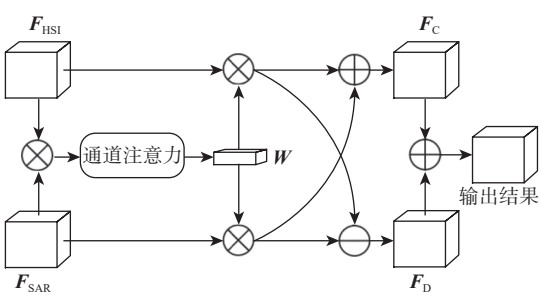


图4 基于平行注意力机制的特征融合器

Fig. 4 Feature fuser based on parallel attention mechanism

2 实验结果与分析

2.1 数据集

Augsburg数据集采集于德国奥格斯堡市上空，包括星载HSI图像和双pol偏振合成孔径雷达(PolSAR)图像[12-13]。Augsburg数据集图像场景包含332像素×485像素和180个光谱波段，其中，HSI图像的光谱范围为0.4～2.5 μm；SAR数据包含了垂直发射-垂直接收强度、垂直发射-水平接收强度及PolSAR协方差矩阵非对角元素的实部和虚部4个特征。

Berlin数据集采集于柏林城区及其周围农村地区上空，包括HSI图像和SAR图像[14]。Berlin数据集中，HSI图像场景的空间分辨率为30 m，包含了797像素×220像素和244个光谱带，光谱带范围为0.4～2.5 μm；相应的，SAR图像空间分辨率为13.89 m，图像大小为1 723像素×476像素。由于这2种模态图像大小不同，使用最近邻插值算法来扩展HSI的大小与SAR图像保持一致。

2.2 实验参数分析

Augsburg和Berlin数据集中，HSI的光谱波段数分别为180和244，选择PCA方法对HSI中的光谱特征维度进行降维处理，仅保留关键的特征信息，减少神经网络的输入数据量以提升计算效率。为探究高光谱主成分数n和图像分类效果的关系，在Augsburg和Berlin数据集上完成实验，选择总体准确率(overall accuracy, OA)作为评价指标，设置n的大小从集合{10, 20, 30, 40, 50}中取值，得到n与总体准确率的映射关系，如图5所示。可以看出，

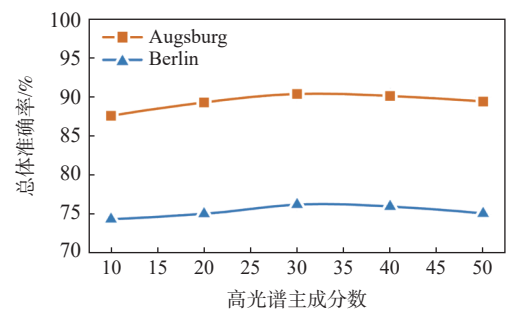


图5 高光谱主成分数与总体准确率关系

Fig. 5 Relationship between hyperspectral principal component score and overall accuracy

当n取值为 30 时, WPANet 达到最理想的实验结果。当n小于 30 时, n与总体准确率呈正相关。当n大于 30 后, 总体准确率无明显提升效果, 甚至出现退化的情况, 因此, 最终将高光谱主成分数设置为 30。

2.3　对比实验分析

为了进一步验证本文提出的 WPANet 地物分类模型的有效性, 选择 SVM[15]、局部二进制模式-极限学习机(local binary pattern-extreme learning machine, LBP-ELM)[16]、双分支卷积神经网络(two-branch CNN, TBCNN)[17]、上下文深度卷积网络 ContextCNN[18]和深度特征交互网络(depthwise feature interaction network, DFINet)[19]5 种不同的遥感图像地物分类方法与 WPANet 进行对比实验分析。

SVM[15] 通过将数据样本特征映射到高维空间, 寻找分类超平面以解决分类的目标任务。LBP-ELM[16] 首先使用局部二进制模式(local binary pattern, LBP)充分挖掘图像局部特征信息, 如边缘特征、纹理特征、斑点特征等, 然后和全局 Gabor 特征与原始光谱特征相融合, 输入特征分类器以得到最终的分类结果。TBCNN[17] 可以高效地完成多源遥感图像特征提取任务。首先, 在特征融合阶段融合 HSI 图像的光谱和空间特征, 然后, 将 HSI 特征与其他数据特征相结合, 完成多源遥感图像的高效融合。ContextCNN[18] 通过设计多尺度卷积滤波组模块, 利用局部的语境交互信息, 充分挖掘上下文信息的潜在交互作用, 以实现对空间和光谱特征的高效分类。DFINet[19] 提出的深度特征交互网络可充分利用多源遥感数据的互补性, 深入挖掘多源遥感数据中的自相关和交叉相关特征信息, 增强影响分类任务的关键信息, 以提高分类准确度。

使用相同的实验环境、相同的训练样本集和测试样本集, 以保证实验过程的科学性和实验结果的可靠性。在 Augsburg 和 Berlin 数据集上完成的对比实验结果分别如表 1 和表 2 所示, 内容包括对各地物样本类别的分类准确度评价指标总体准确率、平均准确度(average accuracy, AA)和 Kappa 系数。对于 Augsburg 数据集, WPANet 取得了最高的 3 类评价指标结果, 并且对于低矮植物和商业区 2 种地物取得了最高的类别样本分类准确度。对于 Berlin 数据集, WPANet 的综合指标和其余分类方法相比, 同样取得了最优的地物分类效果。尤其对于住宅区、工业区、低矮植物和土壤 4 类地物, WPANet 的分类准确度已明显优于其他分类方法, 这是因为基于小波变换的特征提取器可以充分利用频域分析原理挖掘图像的深层特征信息, 同时基于平行注意力机制的特征融合器可以完成不同模态之间的高效融合。综合 WPANet 在 2 种数据集的对比实

表 1　Augsburg 数据集对比实验结果

Table 1　Comparative experimental results for Augsburg dataset　%

方法	分类准确度							总体准确率	平均准确度	Kappa 系数
	森林 (146/13 345)	住宅区 (264/30 065)	工业区 (21/3 830)	低矮植物 (248/26 543)	配额地 (52/523)	商业区 (7/1 632)	水域 (23/1 502)			
SVM[15]	90.55	89.81	23.03	83.73	34.23	9.71	**45.92**	81.60	53.82	73.17
LBP-ELM[16]	93.65	86.81	35.12	83.21	49.33	7.94	44.99	81.47	57.29	73.41
TBCNN[17]	94.77	95.01	**71.17**	85.33	56.41	15.14	22.30	87.11	62.87	81.69
ContextCNN[18]	94.57	**97.25**	51.46	86.25	56.02	13.68	21.57	87.24	60.11	81.82
DFINet[19]	**95.38**	95.84	69.79	86.65	**64.05**	13.86	28.47	88.06	64.86	82.98
WPANet	94.81	93.66	67.52	**95.49**	50.10	**19.91**	44.81	**90.40**	**66.61**	**86.28**

注: 加粗数值表示最优结果。括号中数值为训练/测试样本数。

表 2　Berlin 数据集对比实验结果

Table 2　Comparative experimental results for Berlin dataset　%

方法	分类准确度								总体准确率	平均准确度	Kappa 系数
	森林 (443/54 484)	住宅区 (423/268 219)	工业区 (499/19 067)	低矮植物 (376/58 906)	土壤 (331/17 095)	配额地 (280/13 025)	商业区 (298/24 526)	水域 (170/6 502)			
SVM[15]	50.08	61.07	30.68	84.29	87.30	54.00	26.61	65.40	60.48	57.43	45.36
LBP-ELM[16]	**86.17**	36.95	45.46	84.09	89.72	0.00	0.35	50.17	48.32	49.25	34.65
TBCNN[17]	76.47	62.42	43.22	78.82	76.33	73.44	**49.76**	**82.28**	65.81	67.84	41.79
ContextCNN[18]	77.22	63.69	61.44	73.77	87.22	**82.88**	31.13	74.24	66.31	68.95	54.03
DFINet[19]	68.95	67.52	43.42	81.77	75.58	80.05	40.94	79.87	67.93	67.26	55.22
WPANet	69.35	**81.06**	**62.22**	**85.17**	**90.47**	61.21	25.08	80.04	**76.23**	**69.32**	**64.36**

注: 加粗数值表示最优结果。括号中数值为训练/测试样本数。

验结果可以看出，WPANet 在多源遥感图像地物分类任务中的优势。在 Augsburg 和 Berlin 数据集上完成的对比实验地物分类可视化结果分别如图 6 和图 7 所示。可以看出，WPANet 和其余 5 种地物分类方法相比，对不同的地物类别均表现出了较好的地物分类效果，表明了本文提出的遥感图像地物分类模型 WPANet 在地物分类任务上的显著优势。

2.4 消融实验

为了验证本文提出的小波变换特征提取器和平行注意力特征融合器的有效性，分别在 Augsburg 和 Berlin 数据集上完成消融实验，测试 2 个模块对整体网络架构的影响。消融实验结果如表 3 所示。可以看出，基础的卷积特征提取网络在 Augsburg 和 Berlin 这 2 个数据集上的总体准确率分别为 87.27% 和 73.86%，加入小波变换特征提取器后分别提升 2.79% 和 2.70%；而加入平行注意力特征融合器后分别提升 1.94% 和 1.77%；同时加入两者后，总体准确率获得了进一步提升，相较于原始网络分别提升 3.59% 和 3.21%，这是因为小波变换特征提取器充分利用了频域空间分析原理，完成了对粗/细粒度级别特征的充分挖掘；而平行注意力特征融合器高度

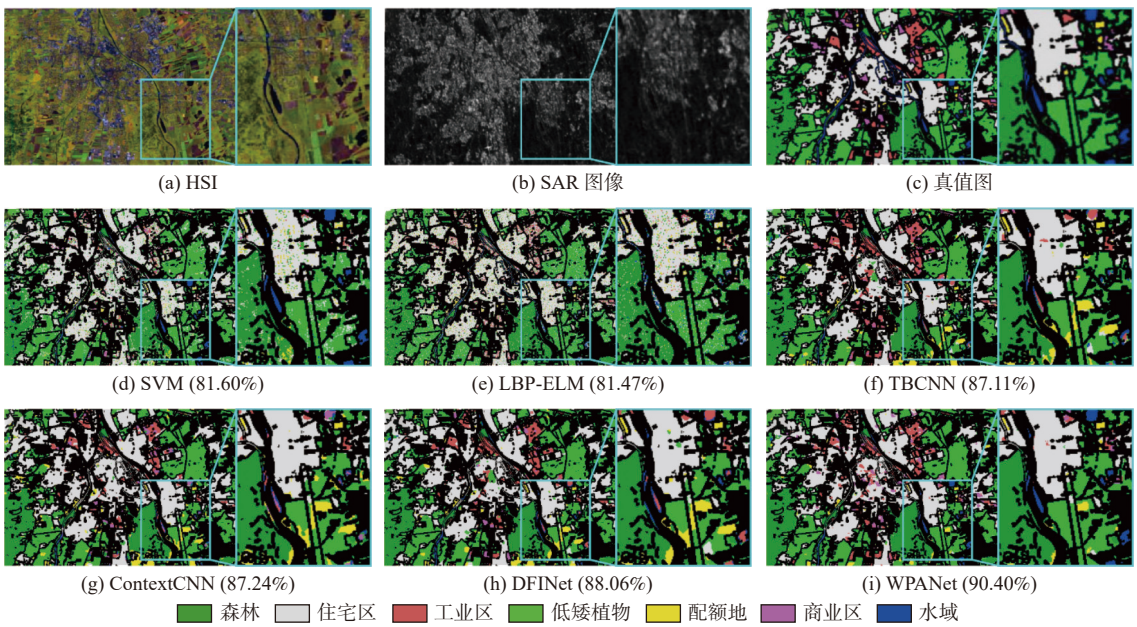


图 6 Augsburg 数据集上不同方法的分类结果

Fig. 6 Classification results of different methods on Augsburg dataset

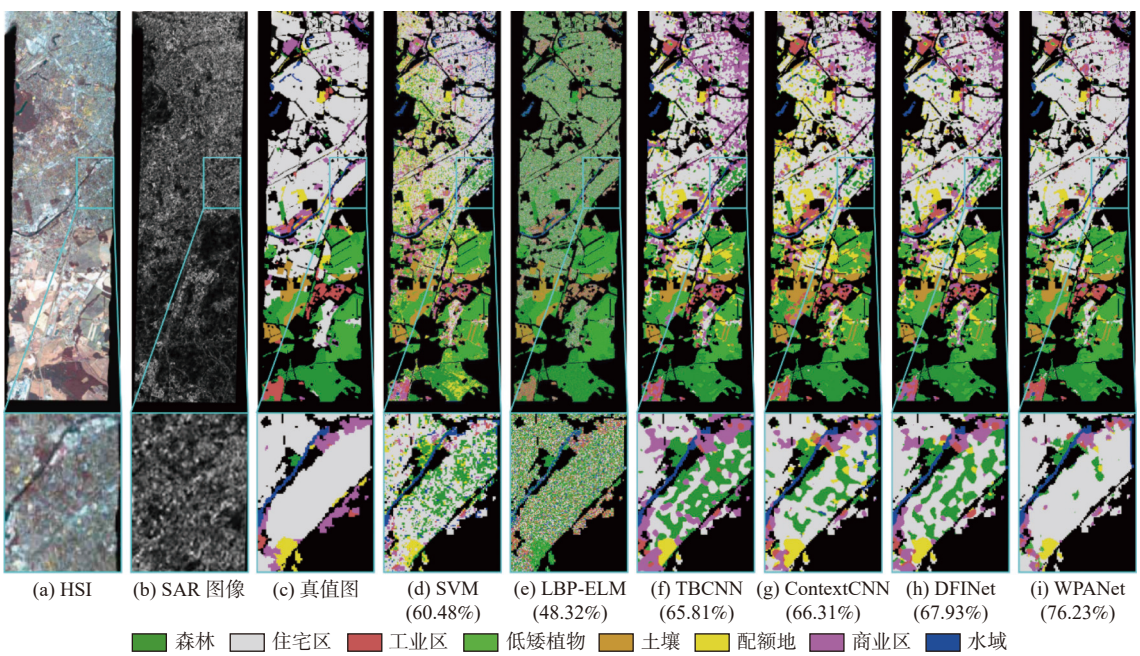


图 7 Berlin 数据集上不同方法的分类结果

Fig. 7 Classification results of different methods on Berlin dataset

表 3　小波变换特征提取器和平行注意力特征融合器消融实验结果

Table 3　Results of ablation experiments with wavelet transform feature extractor and parallel attention-based feature fuser

网络结构	总体准确率/%	
	Augsburg	Berlin
卷积特征提取网络	87.27	73.86
小波变换特征提取器	89.71	75.86
平行注意力特征融合器	88.96	75.17
小波变换特征提取器+平行注意力特征融合器	**90.40**	**76.23**

注：加粗数值表示最优结果。

关注中间特征信息的一致性和差异性，可实现对跨层级和跨模态 2 个层面的特征信息融合补偿，显著提升分类网络的特征融合能力。

3　结　论

1）基于小波变换分析理论，设计并通过编程实现了 WPANet 多源遥感图像分类网络模型。提出了基于小波变换的特征提取器，基于其强大的频域分析能力，通过中间特征的可逆下采样，完成对粗/细粒度级别特征的充分提取，改善了 HSI 和 SAR 图像的特征捕捉效果，对多源遥感图像地物分类准确度有所提升。

2）基于平行注意力机制的特征融合器，通过结合跨层级和跨模态 2 个维度的融合能力，充分挖掘 HSI 特征和 SAR 图像特征的一致性和差异性，改善不同模态图像特征的融合效果和分类效果。

3）在 Augsburg 和 Berlin 数据集上，通过总体准确率、平均准确度和 Kappa 系数 3 个模型评价指标对实验结果完成评价和分析，结合地物分类可视化结果，从定量和定性分析 2 个角度综合表明了本文提出的 WPANet 的有效性和可扩展能力。

4）本文仅实现了 HSI 和 SAR 这 2 种模态图像的融合分类，可继续尝试更多模态（如多光谱图像、热红外图像等）遥感图像的融合分类，更进一步地研究其他模态图像和 HSI、SAR 图像之间的融合分类方法，以扩展 WPANet 的应用场景，在遥感图像处理领域为社会带来更多的实际应用价值。

参考文献（References）

[1] WANG C, ZHANG L, WEI W, et al. Dynamic super-pixel normalization for robust hyperspectral image classification[J]. IEEE Transactions on Geoscience and Remote Sensing, 2023, 61: 5505713.

[2] BIOUCAS-DIAS J M, PLAZA A, CAMPS-VALLS G, et al. Hyperspectral remote sensing data analysis and future challenges[J]. IEEE Geoscience and Remote Sensing Magazine, 2013, 1(2): 6-36.

[3] MOREIRA A, PRATS-IRAOLA P, YOUNIS M, et al. A tutorial on synthetic aperture radar[J]. IEEE Geoscience and Remote Sensing Magazine, 2013, 1(1): 6-43.

[4] MAN Q X, DONG P L, GUO H D. Pixel-and feature-level fusion of hyperspectral and lidar data for urban land-use classification[J]. International Journal of Remote Sensing, 2015, 36(6): 1618-1644.

[5] HU J L, GHAMISI P, SCHMITT A, et al. Object based fusion of polarimetric SAR and hyperspectral imaging for land use classification[C]//2016 8th Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing. Piscataway: IEEE Press, 2016: 1-5.

[6] CHEN Y S, LI C Y, GHAMISI P, et al. Deep fusion of remote sensing data for accurate classification[J]. IEEE Geoscience and Remote Sensing Letters, 2017, 14(8): 1253-1257.

[7] HONG D F, GAO L R, HANG R L, et al. Deep encoder-decoder networks for classification of hyperspectral and LiDAR data[J]. IEEE Geoscience and Remote Sensing Letters, 2020, 19: 5500205.

[8] ZHAO X D, TAO R, LI W, et al. Joint classification of hyperspectral and LiDAR data using hierarchical random walk and deep CNN architecture[J]. IEEE Transactions on Geoscience and Remote Sensing, 2020, 58(10): 7355-7370.

[9] FENG M, GAO F, FANG J, et al. Hyperspectral and lidar data classification based on linear self-attention[C]//2021 IEEE International Geoscience and Remote Sensing Symposium. Piscataway: IEEE Press, 2021: 2401-2404.

[10] LI W, GAO Y H, ZHANG M M, et al. Asymmetric feature fusion network for hyperspectral and SAR image classification[J]. IEEE Transactions on Neural Networks and Learning Systems, 2023, 34(10): 8057-8070.

[11] HU J, SHEN L, SUN G. Squeeze-and-excitation networks[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2018: 7132-7141.

[12] KURZ F, ROSENBAUM D, LEITLOFF J, et al. Real time camera system for disaster and traffic monitoring[C]//Proceedings of International Conference on SMPR 2011. Tehran: University of Tehran, 2011: 1-6.

[13] BAUMGARTNER A, GEGE P, KÖHLER C, et al. Characterisation methods for the hyperspectral sensor HySpex at DLR's calibration home base[C]//Sensors, Systems, and Next-Generation Satellites XVI. Bellingham: SPIE, 201285331H.

[14] HONG D F, HU J L, YAO J, et al. Multimodal remote sensing benchmark datasets for land cover classification with a shared and specific feature learning model[J]. ISPRS Journal of Photogrammetry and Remote Sensing, 2021, 178: 68-80.

[15] MELGANI F, BRUZZONE L. Classification of hyperspectral remote sensing images with support vector machines[J]. IEEE Transactions on Geoscience and Remote Sensing, 2004, 42(8): 1778-1790.

[16] LI W, CHEN C, SU H J, et al. Local binary patterns and extreme learning machine for hyperspectral imagery classification[J]. IEEE Transactions on Geoscience and Remote Sensing, 2015, 53(7): 3681-3693.

[17] XU X D, LI W, RAN Q, et al. Multisource remote sensing data classification based on convolutional neural network[J]. IEEE Transactions on Geoscience and Remote Sensing, 2018, 56(2): 937-949.

[18] LEE H, KWON H. Going deeper with contextual CNN for hyper-

spectral image classification[J]. IEEE Transactions on Image Processing, 2017, 26(10): 4843-4855.

[19] GAO Y H, LI W, ZHANG M M, et al. Hyperspectral and multispectral classification for coastal wetland using depthwise feature interaction network[J]. IEEE Transactions on Geoscience and Remote Sensing, 2021, 60: 5512615.

Multi-source remote sensing image classification based on wavelet transform and parallel attention

WANG Jiayi，GAO Feng*，ZHANG Tiange，GAN Yanhai

(School of Computer Science and Technology，Ocean University of China，Qingdao 266100，China)

Abstract：Exploring the dependency relationships of multi-source remote sensing image data features to leverage the complementary advantages between different modalities has become a prominent research direction in the field of remote sensing. Existing joint classification tasks of hyperspectral and synthetic aperture radar (SAR) data face two key challenges: insufficient feature extraction and representation in images, resulting in the loss of high-frequency information, which hinders subsequent classification tasks, and limited interaction among multi-source image features and weak correlation between multimodal features. To address these challenges, research was conducted on robust representation of image features and efficient correlation of multi-source features, and a multi-source remote sensing image classification method based on wavelet transform and parallel attention mechanism (WPANet) was proposed. The feature extractor based on wavelet transform could effectively utilize frequency domain analysis techniques, capturing coarse- and fine-grained features during the process of reversible downsampling. The feature fuser based on the parallel attention mechanism comprehensively integrated the consistency and differences of multimodal remote sensing data, accomplishing the fusion and generation of highly correlated features to enhance classification accuracy. Experimental results on two real multi-source remote sensing datasets demonstrate the significant advantages of the proposed classification method. The overall accuracy on the Augsburg and Berlin datasets reaches 90.40% and 76.23%, respectively, with at least a 2.66% and 12.22% improvement in overall accuracy compared to mainstream methods like depthwise feature interaction network (DFINet) on the two datasets.

Keywords：hyperspectral image；synthetic aperture radar；wavelet transform；multi-source feature fusion；remote sensing image

Received：2023-06-09；**Accepted**：2023-09-15；**Published Online**：2023-11-07 16：05

URL：link.cnki.net/urlid/11.2625.V.20231107.0859.001

Foundation items：National Natural Science Foundation of China (42106191); Major Innovation 2030-Major Project of New Generation Artificial Intelligence (2022ZD0117202)

* **Corresponding author**. E-mail：gaofeng@ouc.edu.cn

http://bhxb.buaa.edu.cn jbuaa@buaa.edu.cn
DOI: 10.13700/j.bh.1001-5965.2023.0362

显著性感知三重正则化相关滤波无人机目标跟踪算法

贺冰，王法胜*，王星，孙福明
(大连民族大学 信息与通信工程学院，大连 116600)

摘　　要：无人机(UAV)场景中的目标跟踪在很多现实任务中得到广泛应用。与一般场景中的目标跟踪任务不同，UAV目标跟踪更易受到复杂环境干扰和算力的限制。基于此，提出了一种显著性感知三重正则化相关滤波(TRCF)UAV目标跟踪算法。采用高效的显著性目标检测算法动态生成对偶空间正则化器来抑制边界效应，惩罚不相关的背景噪声系数。引入时间正则化应对目标因外观变化而导致的滤波器退化问题，提供更鲁棒的外观模型。此外，引入轻量型的深度网络CF-VGG来提取目标的深度特征，并与手工特征线性融合描述目标的语义信息，提高跟踪精度。在5个公开的UAV基准数据集上进行了充分实验，结果表明：所提算法在5个数据集上的整体性能均有不同程度提升，证明了算法的有效性和鲁棒性，且算法的实时跟踪速度约为21帧/s，能够胜任UAV的目标跟踪任务。

关 键 词：无人机目标跟踪；相关滤波；显著性特征图；时间正则化；轻量型深度网络
中图分类号：TP391
文献标志码：A　　**文章编号**：1001-5965(2025)07-2423-14

近年来，由于在救援救灾、空中监视、侦察、交通巡逻等方面的广泛应用，无人机(unmanned aerial vehicle, UAV)目标跟踪受到了广泛的关注[1-3]。在这些应用中，UAV能够通过视觉传感器感知周围环境，实现对指定目标的自主检测和跟踪。虽然近年来已经出现了许多先进的UAV目标跟踪算法，大大提升了跟踪的准确性，但由于目标的快速运动、尺度变化、完全遮挡等因素的影响，UAV目标跟踪仍然是一个极具困难和挑战性的任务[4]。

现有的主流目标跟踪算法主要有2类，即基于相关滤波(correlation filter, CF)的目标跟踪算法和基于深度学习(deep learning, DL)的目标跟踪算法。为了提高跟踪的准确性和鲁棒性，基于DL的目标跟踪算法使用1个或者多个深度网络对目标进行跟踪，但需要消耗较多的算力。由于考虑到单位成本和UAV本身的质量，大多数UAV目前没有配备图形处理单元(graphics processing unit, GPU)，导致DL目标跟踪算法在单中央处理单元(central processing unit, CPU)的UAV平台上难以满足跟踪精度的要求。基于CF的目标跟踪算法可以在跟踪精度和运行速度之间实现良好平衡，能够快速有效地部署到UAV平台，成为资源有限的UAV平台的首选[5]。

为了满足实时跟踪要求，传统的基于CF的目标跟踪算法普遍使用灰度特征[6]、方向梯度直方图及颜色名称[7]等手工特征描述目标，但手工特征并不能完整地表述目标的语义信息。由于残差网络(residual network, ResNet)、视觉几何组网络(visual geometry group network, VGG)等卷积神经网络(convolutional neural network, CNN)在图像分类和图像识别等具有挑战性的视觉任务上取得巨大成功，研究人员为提高跟踪性能，将深度特征应用到CF的

收稿日期：2023-06-15；**录用日期**：2023-09-15；**网络出版时间**：2023-12-07 07:00
网络出版地址：link.cnki.net/urlid/11.2625.V.20231206.1417.005
基金项目：国家自然科学基金(61972068,61976042)；辽宁省“兴辽英才计划”(XLYC2007023)
* **通信作者**. E-mail：wangfasheng@dlnu.edu.cn

引用格式：贺冰，王法胜，王星，等. 显著性感知三重正则化相关滤波无人机目标跟踪算法[J]. 北京航空航天大学学报，2025，51(7)：2423-2436. HE B，WANG F S，WANG X，et al. Saliency-aware triple-regularized correlation filter algorithm for UAV object tracking[J]. Journal of Beijing University of Aeronautics and Astronautics，2025，51(7)：2423-2436 (in Chinese).

目标跟踪框架中[8-11]。然而，由于这些深度模型的复杂性，使用深度特征的目标跟踪算法在单CPU平台上运行速度往往只有2～3帧/s，无法满足实时跟踪的要求。受到UAV平台本身计算性能的限制，现有的面向UAV的目标跟踪算法大多使用手工特征描述目标。

CF目标跟踪算法所依赖的循环相关操作会对训练图像样本进行循环移位，导致样本在边界处出现不连续而引起边界效应问题，降低了滤波器的鉴别能力，影响跟踪性能。为了缓解边界效应，研究人员提出了2种方法：①如有限边界CF[12]和背景感知CF(background aware CF, BACF)[13]，直接从背景区域提取负训练样本用于训练相关滤波器。文献[14]在BACF算法的基础上通过在训练阶段引入2个相邻帧之间的特征残差来增强滤波器学习，提高目标跟踪算法在复杂环境下的跟踪性能。由于这类算法用于训练的所有正样本和负样本都被赋予相等的权重，使其难以适应复杂目标变化，且不能完全抑制边界效应。②包括空间正则化CF(spatially regularized discriminative CF, SRDCF)[15]和通道及空间可靠性CF[16]等，引入了空间正则化组件，对滤波器系数施加预定义的空间约束，以便滤波器可以在目标区域中学习。文献[17]在SRDCF跟踪框架中嵌入高斯形状掩模，通过增加中心样本的响应从而降低样本污染的影响。响应推理CF算法[18]在SRDCF算法基础之上引入响应推断正则化和历史响应正则化以更好地应对复杂的UAV场景。内容感知SRDCF算法[19]通过空间偏差隐式对不同的训练样本进行加权，重新构建一个能够抑制边界效应的能量函数。文献[20]通过引入上下文补丁在多个方向将响应抑制为零，以此来抑制边界效应。然而，上述算法添加的约束信息通常是固定的，并且不会根据跟踪过程中的目标变化而改变，使得模型无法适应跟踪目标的变化。同时，UAV场景目标跟踪中被跟踪目标的大小和视角易发生巨大变化，基于CF的目标跟踪算法采用基于固定速率的模型更新机制，忽略了不同帧之间的变化，当被跟踪对象外观发生显著变化时，从上一帧中学习到的滤波器就无法根据外观变化进行调整，从而导致滤波器退化。

为了解决以上问题，本文提出了一种显著性感知三重正则化CF(triple-regularized CF, TRCF)目标跟踪算法，引入一种能够有效提取目标信息的显著性目标检测算法构建对偶空间正则化因子，同时加入时间正则化与空间正则化构成三重正则化共同优化滤波器。为有效提高跟踪鲁棒性，在不额外增加算力的前提下，引入轻量型深度网络模块提取目标的深度特征。本文主要贡献如下：

1）为抑制边界效应，提高跟踪器辨别目标的能力，引入一种能够准确从背景中辨别目标的显著性目标检测算法构造动态空间正则化因子，使得相关滤波器可以对目标执行更有针对性的滤波器系数惩罚，并抑制不相关区域的背景噪声。

2）为增强目标跟踪算法在遭遇遮挡、形变等挑战因素时的抗干扰能力，在跟踪框架中引入时间正则化与空间正则化同时优化滤波器，使用交替方向乘子法(alternating direction method of multipliers, ADMM)进行优化求解，在目标发生较大外观变化时提供鲁棒的外观模型，防止滤波器退化。

3）为提高目标跟踪算法的精确度，引入一个轻量型的深度网络提取目标的深度特征，与手工特征按比例结合表示目标的语义信息。该轻量型深度网络计算量小，在单CPU支持下可以达到实时跟踪速度。

4）在5个公开的UAV基准数据集UAV123[21]、VisDrone-SOT2018[22]、UAV20L[21]、UAVTrack112[23]和UAVDT[24]上进行了充分实验，结果表明，TRCF目标跟踪算法具有较高的精确度和鲁棒性。

1 本文算法

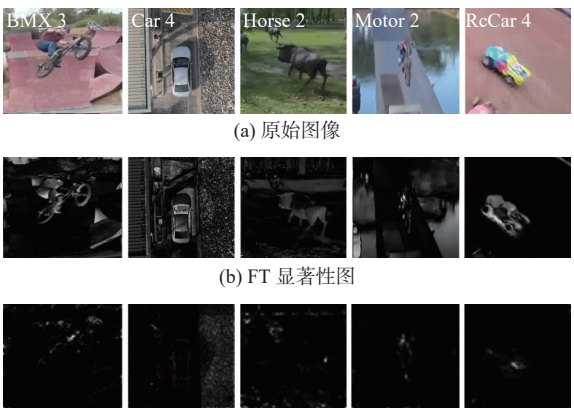
1.1 显著性目标检测模块

近年来，不断有研究者将显著性目标检测引入到CF目标跟踪算法框架中[25]，以提供更加全面的信息附加特征，帮助目标跟踪算法从背景中区分目标。此外，显著性目标检测还可以用来指示目标可能出现的位置，以缓解模型漂移问题[26]。文献[27]使用光谱残差(spectral residual, SR)[28]显著性目标检测算法构建双正则化因子，用来动态关注更多来自对象区域的信息，在相关运算中抑制边界效应。但由于SR算法使用低通滤波和下采样会生成64像素×64像素统一尺寸的显著性图，如果输入图像尺寸产生变化，在尺度空间对于图像频率产生的损失会导致模糊的显著性图。不清晰的显著性图会直接影响目标跟踪算法对目标对象的判断，进而导致滤波器性能下降。为了更加清晰地描述对象目标轮廓，产生能精准突出目标对象轮廓的显著性特征图，本文采用一种基于Lab颜色空间像素向量与平均像素向量欧氏距离的显著性检测算法，即频率协调(frequency-tuned, FT)显著性目标检测算法[29]来构建对偶正则化因子s_1和s_2，动态强调目标区域形状。首先，裁剪比目标边界框大λ倍的区域执行显著性目标检测，获得原始显著性图。然后，乘以余弦窗后，将原始映射调整为滤波器训练使用的外观模型的相应大小，通过阈值ε将其系数重新映射

用来调节权重。最后, 将规则映射$\boldsymbol{s}_{\mathrm{f}}$与另外预定义的正则化因子$\boldsymbol{s}_{\mathrm{r}}$相交, 得到最终的正则化因子, 如下:

$$\boldsymbol{s}_i = \boldsymbol{s}_{\mathrm{f}} \cap \boldsymbol{s}_{\mathrm{r}} \qquad i = 1,2 \tag{1}$$

为了展示 FT 显著性目标检测算法的效果, 选取 UAV 基准数据集中部分图像对 SR 与 FT 显著性目标检测算法进行测试, 实验对比结果如图 1 所示。可以看出, FT 显著性目标检测算法相较于 SR 显著性目标检测算法能够产生对于目标描述更加清晰的显著性图, 帮助跟踪器更加关注目区域, 使滤波器受到更少背景噪声的影响, 在检测目标时获得更强的响应。



(a) 原始图像

(b) FT 显著性图

(c) SR 显著性图

图 1　FT 与 SR 显著性目标检测结果图对比

Fig. 1　Comparison of saliency object detection results of FT and SR methods

1.2　轻量型深度网络

近年来, 同时使用深度特征和手工特征结合的 CF 目标跟踪算法表现出了较强的竞争力, 但多数跟踪器采用结构复杂的大型深度网络提取深度特征, 大大增加计算成本, 只有在搭载 GPU 的跟踪平台上才能够达到实时的跟踪效率。由于 UAV 的主要优势就在于体积小、灵活度高, 导致其仅能搭载 CPU。为了满足实时速率, 目前面向 UAV 的跟踪器大都使用手工特征进行特征提取, 这对跟踪器的精度造成一定影响。文献 [30] 采用知识蒸馏技术设计了一个轻量型的深度网络 CF-VGG 提取目标的深度特征, 并且该网络在单个 CPU 平台上就能达到实时的跟踪速率, 适合 UAV 目标跟踪的应用场景。

CF-VGG 将 VGG-M[31] 作为教师网络, 使用知识蒸馏[32-33] 生成轻量型的学生网络。CF-VGG 的网络架构如图 2 所示。

本文将 CF-VGG 引入到算法的跟踪框架中来提取目标的深度特征, 与手工提取的特征一起描述目标特征。具体而言, 对于第$t+1$帧的搜索区域, CF-VGG 使用 stride = 2的步长进行下采样, 并用

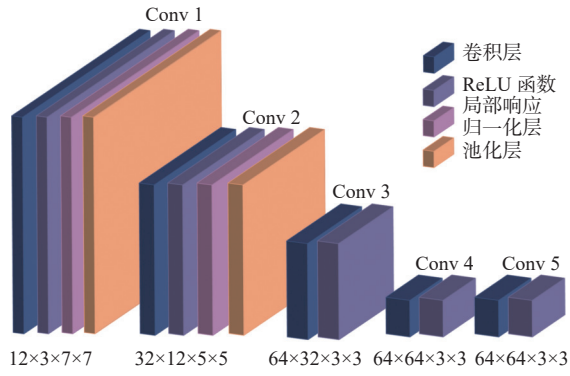


图 2　CF-VGG 网络架构

Fig. 2　Architecture of CF-VGG

ReLU函数对更有助于精确定位的第 3 卷积层上的 CNN 特征进行处理, 以描述目标的语义信息。

1.3　时间正则化模块

CF 目标跟踪算法中使用循环移位操作增加样本, 从而引起边界效应, 降低了跟踪性能。Danelljan 等[15] 提出空间正则化项, 根据相关滤波系数的空间位置对其进行惩罚, 可以减轻边界效应并提高相关滤波器的性能。但面对复杂的 UAV 跟踪场景, 目标更容易因为遮挡、形变、尺度变化等挑战因素导致目标外观发生显著变化, 如果上一帧中的滤波器无法根据外观变化进行调整, 就会导致过滤器退化从而影响跟踪器性能。针对该问题, 本文将时间正则化[34] 引入到基线跟踪框架中构建三重正则化器。由于时间正则化是对多个样本的合理近似, 几乎不额外增加计算复杂度, 能够使算法在不损失效率的同时更好处理边界效应; 并且时间正则化项可以自适应平衡主动和被动模型学习, 对于遭遇遮挡、形变和尺度变化等会产生较大外观变化的序列提供更加稳健的模型, 防止滤波器退化。

本文算法的整体跟踪流程如图 3 所示。

1.4　算法的建模和优化

1.4.1　TRCF 目标跟踪算法的建模和优化

本文算法的目标函数为

$$\arg\min_{\boldsymbol{w}} \frac{1}{2}\left\|\sum_{d=1}^{D}\boldsymbol{x}_d \star (\boldsymbol{s}_1 \odot \boldsymbol{w}_d) - \boldsymbol{y}\right\|_2^2 + \frac{1}{2}\sum_{d=1}^{D}\|\boldsymbol{s}_2 \odot \boldsymbol{w}_d\|_2^2 + \frac{\mu}{2}\|\boldsymbol{w} - \boldsymbol{w}_{t-1}\|_2^2 \tag{2}$$

式中: $\boldsymbol{x} = \{\boldsymbol{x}_d\}_{d=1:D}$ 和 $\boldsymbol{w} = \{\boldsymbol{w}_d\}_{d=1:D}$分别为第$d$个特征通道中的训练样本和滤波器; D为特征通道数; $\boldsymbol{y} \in \mathbf{R}^{M\times N}$为预定义的高斯形状期望响应, 其空间大小尺寸为$M \times N$; $\star$表示循环相关运算; $\odot$表示逐元素相乘; $\boldsymbol{s}_1$和$\boldsymbol{s}_2 \in \mathbf{R}^{M\times N}$为由显著性特征图生成的动态双正则化器; t为第t帧; $\boldsymbol{w}_{t-1}$为$t-1$帧滤波器; μ为正则化参数; $\|\boldsymbol{w} - \boldsymbol{w}_{t-1}\|_2^2$定义为时间正则化因子。

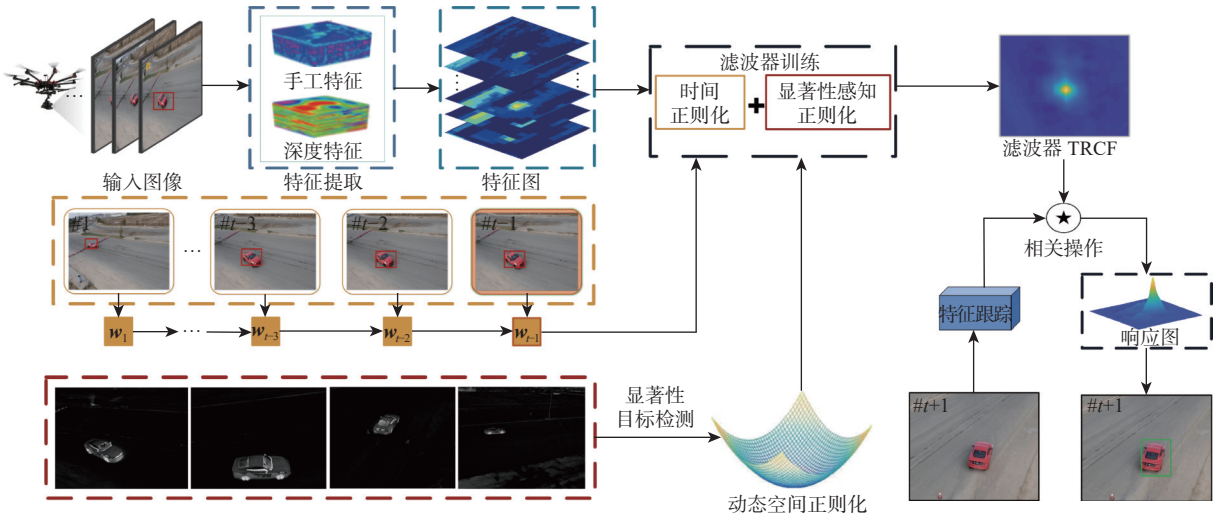


图3 TRCF目标跟踪算法框架

Fig. 3 Framework of TRCF object tracking algorithm

通过引入约束$\boldsymbol{g}=\boldsymbol{s}_1\odot\boldsymbol{w}_d$、拉格朗日乘数$\boldsymbol{r}$及相应的步长参数$\rho\in\mathbf{R}$，对式(2)使用增广拉格朗日方法求解。式(2)利用增广拉格朗日法转化为

$$\begin{aligned}L(\boldsymbol{w},\boldsymbol{g},\boldsymbol{r})=&\frac{1}{2}\left\|\sum_{d=1}^{D}\boldsymbol{x}_d\star\boldsymbol{g}_d-\boldsymbol{y}\right\|_2^2+\frac{1}{2}\sum_{d=1}^{D}\|\boldsymbol{s}_2\odot\boldsymbol{w}_d\|_2^2+\\&\sum_{d=1}^{D}\operatorname{trace}\left[(\boldsymbol{g}_d-\boldsymbol{s}_1\odot\boldsymbol{w}_d)^{\mathrm{T}}\boldsymbol{r}_d\right]+\\&\frac{\rho}{2}\sum_{d=1}^{D}\|\boldsymbol{g}_d-\boldsymbol{s}_1\odot\boldsymbol{w}_d\|_2^2+\frac{\mu}{2}\|\boldsymbol{w}-\boldsymbol{w}_{t-1}\|_2^2\end{aligned}\tag{3}$$

通过引入$\boldsymbol{h}=\dfrac{1}{\rho}\boldsymbol{r}$，式(3)可以重新表述为

$$\begin{aligned}L(\boldsymbol{w},\boldsymbol{g},\boldsymbol{r})=&\frac{1}{2}\left\|\sum_{d=1}^{D}\boldsymbol{x}_d\star\boldsymbol{g}_d-\boldsymbol{y}\right\|_2^2+\frac{1}{2}\sum_{d=1}^{D}\|\boldsymbol{s}_2\odot\boldsymbol{w}_d\|_2^2+\\&\frac{\rho}{2}\sum_{d=1}^{D}\|\boldsymbol{g}_d-\boldsymbol{s}_1\odot\boldsymbol{w}_d+\boldsymbol{h}_d\|_2^2-\frac{\rho}{2}\sum_{d=1}^{D}\|\boldsymbol{h}_d\|_2^2+\\&\frac{\mu}{2}\|\boldsymbol{w}-\boldsymbol{w}_{t-1}\|_2^2\end{aligned}\tag{4}$$

采用ADMM方法对算法进行优化，将原问题转化为关于滤波器和辅助变量求解的子问题。式(4)可以被分解为3个子问题进行迭代求解，且所有子问题都有封闭解：

$$\begin{cases}\boldsymbol{g}^{k+1}=\arg\min\limits_{\boldsymbol{g}}\left\{\dfrac{1}{2}\left\|\sum\limits_{d=1}^{D}\boldsymbol{x}_d\star\boldsymbol{g}_d-\boldsymbol{y}\right\|_2^2+\dfrac{\rho}{2}\left\|\boldsymbol{g}-\boldsymbol{s}_1\odot\boldsymbol{w}^k+\boldsymbol{h}^k\right\|_2^2\right\}\\\boldsymbol{w}^{k+1}=\arg\min\limits_{\boldsymbol{w}}\left\{\dfrac{1}{2}\|\boldsymbol{s}_2\odot\boldsymbol{w}\|_2^2+\dfrac{\rho}{2}\left\|\boldsymbol{g}-\boldsymbol{s}_1\odot\boldsymbol{w}+\boldsymbol{h}^k\right\|_2^2+\dfrac{\mu}{2}\|\boldsymbol{w}-\boldsymbol{w}_{t-1}\|_2^2\right\}\\\boldsymbol{h}^{k+1}=\boldsymbol{h}^k+\boldsymbol{g}^{k+1}-\boldsymbol{s}_1\odot\boldsymbol{w}^{k+1}\end{cases}\tag{5}$$

式中：k为迭代次数。

1)子问题$\boldsymbol{g}^{k+1}$

利用帕塞瓦尔定理和卷积定理，式(5)中的第1个子问题可以转换到傅里叶域中，如下：

$$\begin{aligned}\hat{\boldsymbol{g}}^{k+1}=&\arg\min_{\hat{\boldsymbol{g}}}\left\{\frac{1}{2}\left\|\sum_{d=1}^{D}\hat{\boldsymbol{x}}_d^*\odot\hat{\boldsymbol{g}}_d-\hat{\boldsymbol{y}}\right\|_2^2+\right.\\&\left.\frac{\rho}{2}\left\|\hat{\boldsymbol{g}}-\left(\widehat{\boldsymbol{s}_1\odot\boldsymbol{w}}\right)+\hat{\boldsymbol{h}}\right\|_2^2\right\}\end{aligned}\tag{6}$$

式中：$*$表示复共扼；“^”表示离散傅里叶变换，$\hat{\boldsymbol{a}}=\boldsymbol{F}\cdot\boldsymbol{a}$，$\boldsymbol{F}$为离散傅里叶变换矩阵。由于$\hat{\boldsymbol{y}}$仅依赖于所有通道上的$\hat{\boldsymbol{g}}_d(i,j)$和$\hat{\boldsymbol{x}}_d^*(i,j)$，$\boldsymbol{e}_{ij}\in\mathbf{C}^{D\times1}$表示覆盖所有$D$个通道上的$\hat{\boldsymbol{x}}_d^*(i,j)$元素。式(6)可以分解为$M\times N$个子问题，每个子问题的形式如下：

$$\begin{aligned}&\arg\min_{\boldsymbol{e}_{ij}(\hat{\boldsymbol{g}})}\left\{\frac{1}{2}\left\|\sum_{d=1}^{D}\boldsymbol{e}_{ij}(\hat{\boldsymbol{x}})^{\mathrm{H}}\boldsymbol{e}_{ij}(\hat{\boldsymbol{g}})-\hat{\boldsymbol{y}}_{ij}\right\|_2^2+\right.\\&\left.\frac{\rho}{2}\left\|\boldsymbol{e}_{ij}(\hat{\boldsymbol{g}})-\boldsymbol{e}_{ij}\left(\widehat{\boldsymbol{s}_1\odot\boldsymbol{w}}\right)+\boldsymbol{e}_{ij}\left(\hat{\boldsymbol{h}}\right)\right\|_2^2\right\}\end{aligned}\tag{7}$$

其中，$(\hat{\boldsymbol{x}})^{\mathrm{H}}$表示Hermitian转置，$(\hat{\boldsymbol{x}})^{\mathrm{H}}=(\hat{\boldsymbol{x}}^*)^{\mathrm{T}}$。将式(7)关于$\boldsymbol{e}_{ij}(\hat{\boldsymbol{g}})$的导数设为零，可得

$$\boldsymbol{q}\cdot\boldsymbol{e}_{ij}(\hat{\boldsymbol{g}})=\rho\left[\boldsymbol{e}_{ij}(\boldsymbol{s}_1\odot\boldsymbol{w})-\boldsymbol{e}_{ij}\left(\hat{\boldsymbol{h}}\right)\right]+\boldsymbol{e}_{ij}(\hat{\boldsymbol{x}})^*\hat{\boldsymbol{y}}_{ij}\tag{8}$$

式中：$\boldsymbol{q}=\boldsymbol{e}_{ij}(\hat{\boldsymbol{x}})^*\boldsymbol{e}_{ij}(\hat{\boldsymbol{x}})^{\mathrm{H}}+\rho\boldsymbol{I}$，$\boldsymbol{I}$为单位矩阵。式(8)中含有矩阵的求逆过程会影响计算速度，为了减少计算量，采用Sherman-Morrison公式，可以获得$\boldsymbol{e}_{ij}(\hat{\boldsymbol{g}})$的封闭解：

$$\begin{cases}\boldsymbol{e}_{ij}(\hat{\boldsymbol{g}})=\dfrac{1}{\rho}\left[\boldsymbol{I}-\dfrac{\boldsymbol{e}_{ij}(\hat{\boldsymbol{x}})^*\boldsymbol{e}_{ij}(\hat{\boldsymbol{x}})^{\mathrm{H}}}{\rho+\boldsymbol{e}_{ij}(\hat{\boldsymbol{x}})^{\mathrm{H}}\boldsymbol{e}_{ij}(\hat{\boldsymbol{x}})^*}\right]\boldsymbol{\psi}\\\boldsymbol{\psi}=\rho\left[\boldsymbol{e}_{ij}\left(\widehat{\boldsymbol{s}_1\odot\boldsymbol{w}}\right)-\boldsymbol{e}_{ij}\left(\hat{\boldsymbol{h}}\right)\right]+\boldsymbol{e}_{ij}(\hat{\boldsymbol{x}})^*\hat{\boldsymbol{y}}_{ij}\end{cases}\tag{9}$$

2) 子问题$\boldsymbol{w}^{k+1}$

第2个子问题可以通过将导数设为零在时域中直接求解，即

$$\frac{\partial}{\partial \boldsymbol{w}}\left\{\frac{1}{2}\|\boldsymbol{s}_2\odot\boldsymbol{w}\|_2^2+\frac{\rho}{2}\|\boldsymbol{g}-\boldsymbol{s}_1\odot\boldsymbol{w}+\boldsymbol{h}\|_2^2+\frac{\mu}{2}\|\boldsymbol{w}-\boldsymbol{w}_{t-1}\|_2^2\right\}=0 \tag{10}$$

得到$\boldsymbol{w}$的解如下:

$$\boldsymbol{w}=\frac{\rho\cdot\boldsymbol{s}_1\odot(\boldsymbol{g}+\boldsymbol{h})+\mu\boldsymbol{w}_{t-1}}{\rho(\boldsymbol{s}_1\odot\boldsymbol{s}_1)+(\boldsymbol{s}_2\odot\boldsymbol{s}_2)+\mu} \tag{11}$$

3) 迭代更新

在每次的迭代更新中, $\boldsymbol{h}$使用式(5)的最后 1 行进行更新。步长参数ρ使用文献 [15] 中相同的方案用尺度β进行更新, 即

$$\rho^{k+1}=\min\left(\rho_{\max},\beta\rho^k\right) \tag{12}$$

1.4.2　TRCF 目标跟踪算法描述

输入: 第t帧图像、目标位置$\boldsymbol{p}_{t-1}$、前一帧$t-1$上的尺度$\boldsymbol{\iota}_{t-1}$、外观模型$\boldsymbol{x}_{t-1}^{\text{model}}$、正则化器$\boldsymbol{s}_1$、$\boldsymbol{s}_2$和滤波器$\boldsymbol{w}_{t-1}$。

输出: t帧处的预测位置$\boldsymbol{p}_t$和尺度$\boldsymbol{\iota}_t$。

在跟踪过程中, 对于每一帧, 以前一帧中的目标位置为中心裁剪相应尺度的样本区域来提取其手工特征和深度特征, 并采用式(13)将 2 种特征进行线性组合以得到最终的目标特征。

$$\boldsymbol{\chi}_{\text{total}}=v\boldsymbol{\chi}_{\text{m}}+(1-v)\boldsymbol{\chi}_{\text{n}} \tag{13}$$

式中: $\boldsymbol{\chi}_{\text{m}}$为手工特征; $\boldsymbol{\chi}_{\text{n}}$为深度特征; $\boldsymbol{\chi}_{\text{total}}$为总特征; v为平衡 2 种特征的参数。

本文采用文献 [26] 中提出的线性插值方法更新外观模型$\boldsymbol{\chi}_t^{\text{model}}$和滤波器$\boldsymbol{w}_t$, 学习率为$\eta$, $\boldsymbol{\chi}_t$为从第t帧提取的特征。如下:

$$\boldsymbol{\chi}_t^{\text{model}}=(1-\eta)\boldsymbol{\chi}_{t-1}^{\text{model}}+\eta\boldsymbol{\chi}_t \tag{14}$$

$$\boldsymbol{w}_t=(1-\eta)\boldsymbol{w}_{t-1}+\eta\boldsymbol{w}_t \tag{15}$$

将相关滤波器的响应函数定义为$\dot{y}$, 如下:

$$\dot{y}=\boldsymbol{F}^{-1}\left\{\sum_{d=1}^{D}\hat{\boldsymbol{x}}_d\odot\hat{\boldsymbol{w}}_d\right\} \tag{16}$$

响应的最大值位置即为目标的估计位置。

对于序列的第 2 帧至最后 1 帧:

1) 在以目标位置$\boldsymbol{p}_{t-1}$为中心的第t帧上裁剪图像补丁, 确定搜索区域。

2) 提取手工特征及深度特征使用式(14)进行线性融合, 构建目标描述。

3) 搜索式(16)$\dot{y}$中的最高值, 作为目标位置估计$\boldsymbol{p}_t$。

4) 使用文献 [35] 中的尺度空间相关滤波器来估计尺度$\boldsymbol{\iota}_t$。

5) 生成新空间正则化器$\boldsymbol{s}_1$、$\boldsymbol{s}_2$, 并放弃旧的空间正则化器。

6) 提取$\boldsymbol{p}_t$和$\boldsymbol{\iota}_t$处的特征$\boldsymbol{\chi}_t$, 并使用式(14)将外观模型更新为$\boldsymbol{\chi}_t^{\text{model}}$。

7) 使用外观模型$\boldsymbol{\chi}_t^{\text{model}}$和正则化器$\boldsymbol{s}_1$、$\boldsymbol{s}_2$及第$t-1$帧滤波器$\boldsymbol{w}_{t-1}$优化滤波器。

8) 使用式(15)将滤波器更新为$\boldsymbol{w}_t$。

2　实验结果与分析

2.1　实验细节

本文使用 MATLAB 2018a 实现算法, 所有实验均在 64 位 Windows10 平台下完成。计算机处理器参数为 Intel(R) Core (TM) i7-9700 3.00 GHz CPU, 内存为 48 GB。跟踪器的搜索区域被设置为大小为$5XY/2$的正方形, 其中, X和Y分别为目标的高度和宽度。使用$v=0.85$来平衡手工特征与深度特征。在使用显著性目标检测算法构建正则化因子过程中, 尺度$\lambda=2$, 重映射阈值$\varepsilon=0.1$。对于 ADMM 迭代, $\boldsymbol{g}$、$\boldsymbol{w}$和 $\boldsymbol{h}$的初始值都是空矩阵。ADMM 迭代次数、初始步长ρ、超参数μ、最大值$\rho_{\max}$和尺度β分别为 2、1、16、100 和 10。用于模型更新的学习率设置为$\eta=0.0192$。

将本文算法与其他 CF 目标跟踪算法采用一遍评估法(one pass evaluation, OPE)进行评估, 即使用精确度和成功率[36] 对算法进行测评。

在每一帧中, 精确度由跟踪目标物体的中心位置与手动标记的目标物体真实值之间的中心位置误差(center location error, CLE)确定, 用精确度图中 CLE 低于 20 像素的帧的百分比对跟踪器进行排名。成功率由边界框和真实边界框之间的交并比来衡量, 定义为

$$I_{\text{oU}}=\frac{\boldsymbol{C}_p\cap\boldsymbol{C}_q}{\boldsymbol{C}_p\cup\boldsymbol{C}_q} \tag{17}$$

式中: $\boldsymbol{C}_p$为跟踪预测的边界框; $\boldsymbol{C}_q$为标定的真实边界框; $\cap$和$\cup$分别表示这 2 个区域的交集和并集。当交并比大于给定阈值时, 定义为跟踪成功, 跟踪成功的帧数占视频序列总帧数的百分比即为成功率。使用成功图的曲线下面积(area under curve, AUC)对跟踪器进行排名。

2.2　定量分析

本节对提出的跟踪器进行全面的实验评估, 采用 5 个公开的 UAV 目标数据集 UAV123[21]、VisDrone-SOT2018[22]、 UAV20L[21]、 UAVTrack112[23]、UAVDT[24] 与多个先进的 CF 目标跟踪算法进行对比实验, 包括 AutoTrack[37]、ADTrack_v2[38]、MRCF[39]、MSCF[40]、ARCF[41]、SRECF[42]、DRCF[27]、SRDCF[15]、STRCF[35]、Staple[43]、BACF[13]、DSST[34]、fDSST[44]、SAMF[45]。

UAV123数据集中包含UAV拍摄的123个彩色序列。VisDrone-SOT2018数据集由35个专注于复杂场景下面临众多挑战的视频序列组成。UAV20L包含20个长时跟踪序列，由超过54 000帧的各种具有挑战性的场景序列组成。UAVDT是采用UAV拍摄的车辆交通内容的数据集，共50个视频序列，包含天气条件、飞行高度、车辆类别等14种属性。UAVTrack112是基于真实世界生活场景构建的空中跟踪的数据集，共有112个视频序列。

2.2.1 UAV123数据集

OPE实验结果如图4所示。在精确度方面，TRCF(70.3%)目标跟踪算法排名第1，比第2名的ADTrack_v2算法高出0.2%，比排名第3的基准算法DRCF性能提升0.7%。在成功率方面，TRCF目标跟踪算法以48.8%的得分排名第1，比第2名的MRCF算法高0.3%，比第3名的MSCF算法高0.5%。ADTrack_v2算法和基准算法DRCF以相同的分数分别排在第5位和第6位。

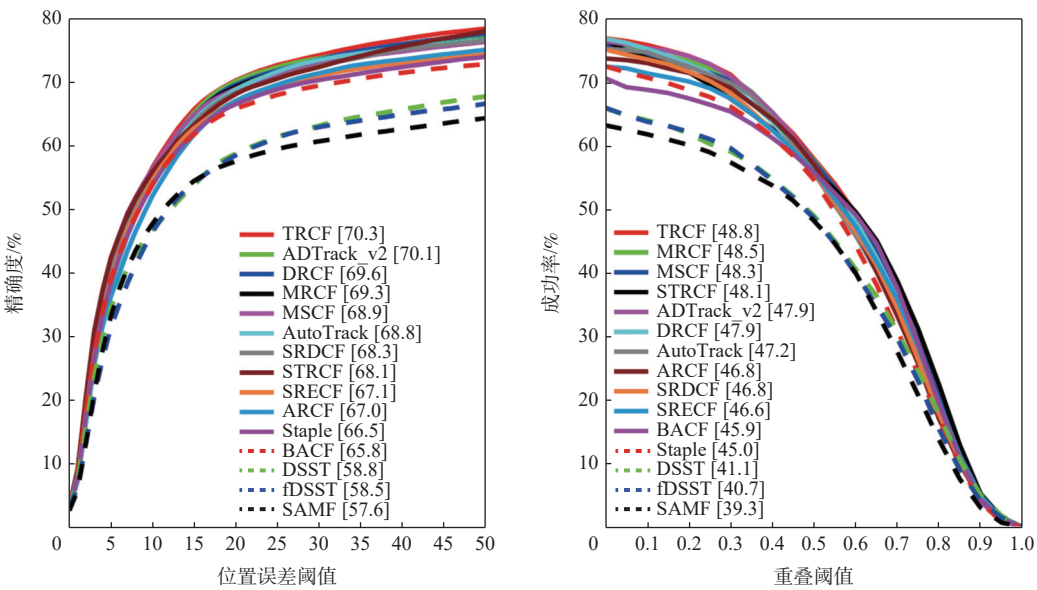


图4 UAV123数据集上的精确度和成功率曲线

Fig. 4 Precision and success rate plots on UAV123 dataset

为评估TRCF目标跟踪算法在不同挑战属性上的表现，将15种CF目标跟踪算法在UAV123数据集对于不同挑战属性上的性能进行汇总。图5为12种挑战属性上的成功率曲线。TRCF目标跟踪算法在6个挑战属性达到了最好性能，分别为纵横比变换(41.6%)、相机运动(49.0%)、超出视野(42.0%)、尺度变化(45.7%)、部分遮挡(41.0%)及视角变化(43.5%)。在小物体和快速运动2种挑战属性中，TRCF目标跟踪算法分别以46.6%和37.0%排进了前3。上述结果表明，TRCF目标跟踪算法对于目标发生较大外观变化和位置变化的挑战属性性能表现优异，加入时间正则化模块与高效的显著性目标检测算法模块相结合共同优化滤波器后，可以更加精准地跟踪目标，并构建更加鲁棒的外观模型，从而提高跟踪性能。

2.2.2 VisDrone-SOT2018数据集

图6显示了在VisDrone-SOT2018数据集上的评估结果。在精确度方面，TRCF目标跟踪算法以79.9%排名第2，比排名第3的ARCF算法高0.2%，SRECF算法则以79.1%排名第4，与本文算法相差0.8%。在成功率方面，TRCF目标跟踪算法与第1名相差1.6%排在第2位，比SRECF算法(57.4%)、AutoTrack算法(57.3%)、DRCF算法(57.3%)、MSCF算法(57.0%)分别提高了1%、1.1%、1.1%、1.4%。

2.2.3 UAV20L数据集

目标跟踪算法在UAV20L数据集上的测评结果如图7所示。在精确度方面，本文算法以60.6%的得分排在第1名，比第2名的STRCF算法高3.1%。MSCF算法和MRCF算法分别以57.0%和56.8%排在第3位和第4位。本文算法比排在第8位的基准算法DRCF提高了6.4%。在成功率方面，TRCF目标跟踪算法以42.9%的得分领先其他先进算法排在第1名，与排名第2～3名的STRCF算法和MRCF算法相比，分别提高了1.9%和3.2%，而与BACF算法、MSCF算法、ADTrack_v2算法、ARCF算法及本文基准算法DRCF相比，TRCF目标跟踪算法在成功率上分别提高了3.2%、3.6%、4.4%、4.8%和5.2%。实验结果表明，本文设计的CF目标跟踪框架提升了CF目标跟踪算法对长时UAV目标跟踪的处理能力。

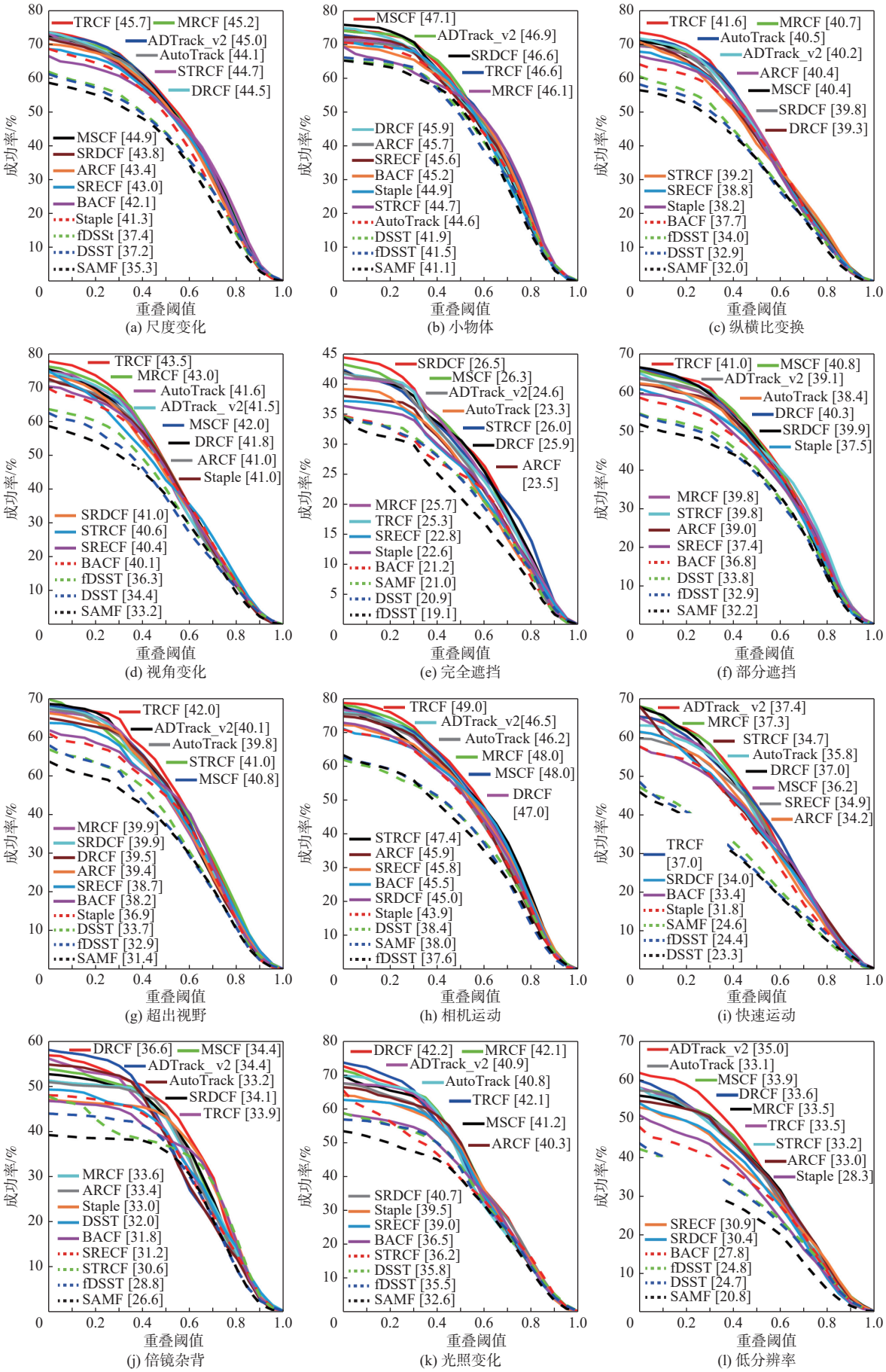


图 5　UAV123 数据集上的 12 个挑战属性下的成功率曲线

Fig. 5　Success rate plots of 12 challenging attributes on UAV123 dataset

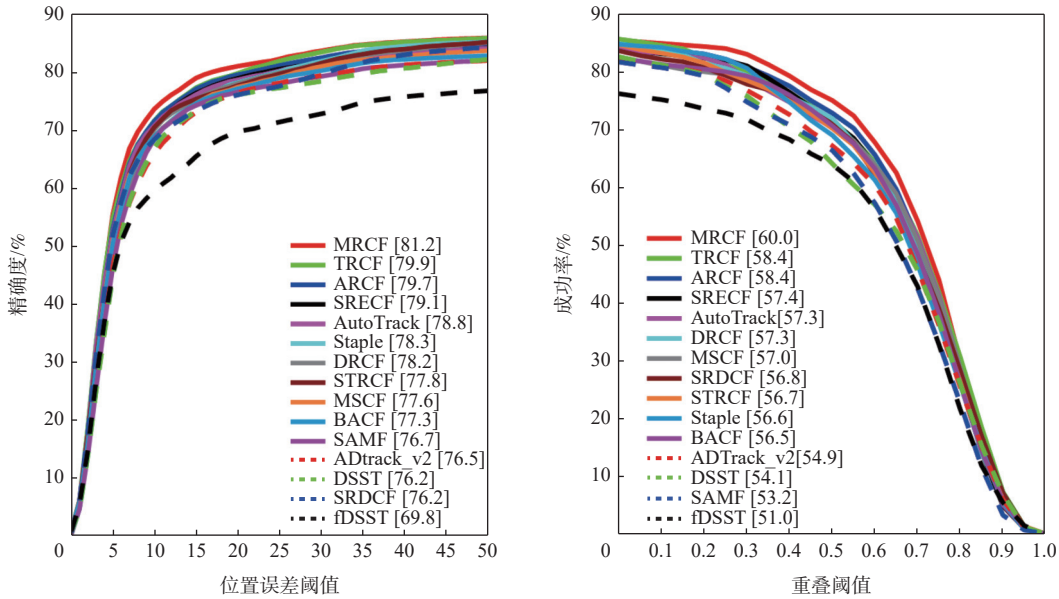


图 6 VisDrone-SOT2018 数据集上的精确度和成功率曲线

Fig. 6 Precision and success rate plots on VisDrone-SOT2018 dataset

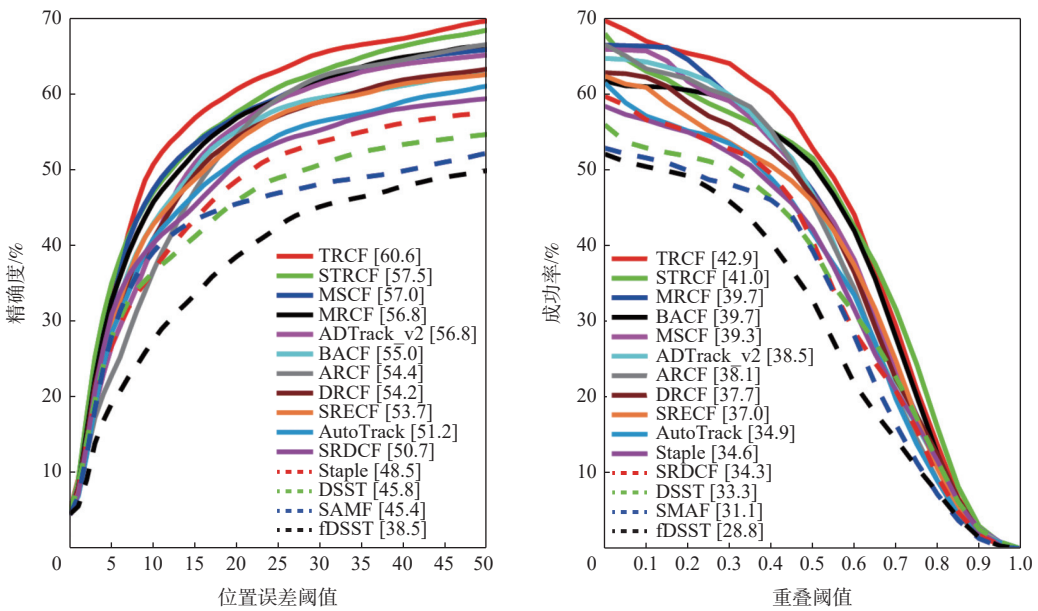


图 7 UAV20L 数据集上的精确度和成功率曲线

Fig. 7 Precision and success rate plots on UAV20L dataset

2.2.4 UAVTrack112 数据集

图 8 为算法在 UAVTrack112 数据集上的评测结果。TRCF 目标跟踪算法的精确度为 68.0%，排在第 5 位。TRCF 目标跟踪算法的成功率为 46.1%，比 ARCF 算法高出 0.4%，比 SRECF 算法高出 0.9%，与第 1 名算法差距仅为 0.7%。由于 UAVTrack112 数据集主要关注小物体、光照变化及快速运动等挑战因素，是面向全天场景构建的 UAV 目标序列。对于目标占比极小的小物体，深度网络能提取到的语义信息有限，显著性特征图的提取也易受到周围较大物体的影响而导致不精准显著性图的生成，使跟踪结果受到影响。但 TRCF 目标跟踪算法相比于基准算法 DRCF 精确度高出 0.5%，成功率高出 0.3%，改进后的本文算法跟踪性能整体有明显提升。

2.2.5 UAVDT 数据集

各算法在 UAVDT 数据集上的评估结果如图 9 所示。TRCF 目标跟踪算法的精确度和成功率分别为 73.2% 和 45.4%，排在第 6 名。在精确度方面，TRCF 目标跟踪算法与排名第 1～5 位的 ARCF 算法、MRCF 算法、AutoTrack 算法、MSCF 算法、SRECF 算法的精确度相差均未超过 1%，但比基准算法 DRCF 的精确度提高了 1.3%。成功率方面，TRCF 目标跟踪算法比基准算法 DRCF 的成功率高出 0.6%。UAVDT 数据集是面向 UAV 拍摄的车辆

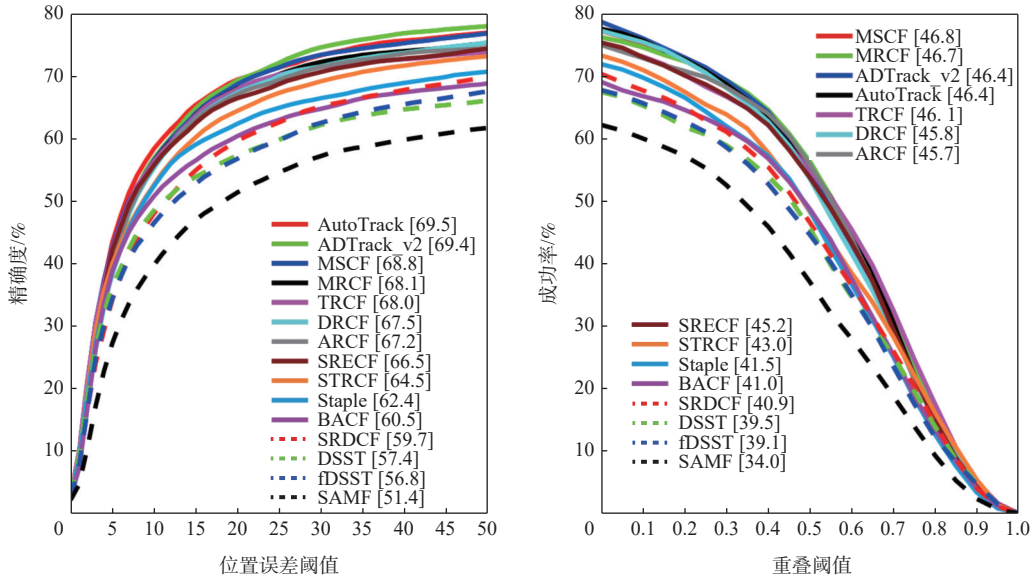


图 8　UAVTrack112 数据集上的精确度和成功率曲线

Fig. 8　Precision and success rate plots on UAVTrack112 dataset

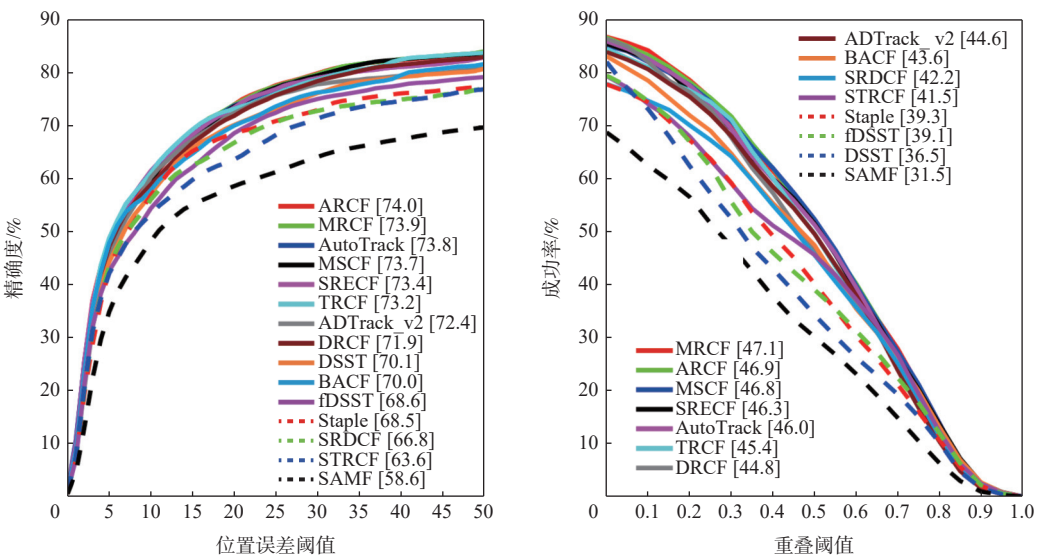


图 9　UAVDT 数据集上的精确度和成功率曲线

Fig. 9　Precision and success rate plots on UAVDT dataset

目标, 多数为距离较远的小目标车辆及包含数个夜间和雾天场景的目标序列。在照明条件有利的情况下, 显著性目标检测算法能够较好地检测到目标轮廓, 但在暗夜场景目标不清晰, 影响显著性特征图的结果, 使跟踪性能受到影响。相比于基准算法 DRCF, 本文算法无论是精确度还是成功率都有明显提升, 证明本文算法能起到提升算法整体跟踪性能的作用。

为直观地展示本文算法相较于基准算法在不同数据集上的提升幅度, 表 1 列出了本文算法与基准算法在 5 个不同基准数据集上的性能数值量化。

表 1　本文算法与基准算法的性能比较

Table 1　Comparisons between the proposed method and baseline method　%

基准数据集	精确度		成功率	
	DRCF	TRCF	DRCF	TRCF
UAV123[21]	69.6	**70.3**	47.9	**48.8**
VisDrone-SOT2018[22]	78.2	**79.9**	57.3	**58.4**
UAVTrack112[23]	67.5	**68.0**	45.8	**46.1**
UAV20L[21]	54.2	**60.6**	37.7	**42.9**
UAVDT[24]	71.9	**73.2**	44.8	**45.4**

注: 粗体表示与基准算法相比性能增强。

2.2.6　时间正则化参数的影响

进一步分析不同时间正则化参数对滤波器跟踪性能的影响, 正则化参数 μ 决定了用当前帧中的新样本替换从前一帧中学习到的滤波器的速率。

从图 10 可以看出，μ对 TRCF 目标跟踪器的精度有明显影响，当$\mu=16$时，TRCF 目标跟踪器的性能达到最佳。

2.3 消融实验

2.3.1 跟踪性能对比

为了验证 TRCF 目标跟踪算法中显著性特征检测模块、轻量级深度特征和时间正则化对本文算法性能的影响，在 3 个 UAV 基准数据集上进行了消融实验，结果如表 2 所示。

在UAV123 数据集中，成功率方面，加入 CF-VGG 和加入 CF-VGG 与时间正则化的改进算法高于基准算法，加入 3 个模块后，本文算法的成功率比基准算法提高 0.9%。在精确度方面，加入 CF-VGG、加入 CF-VGG 与时间正则化及本文算法得分均高于基准算法，分别提高了 0.5%、0.2% 和 0.7%。

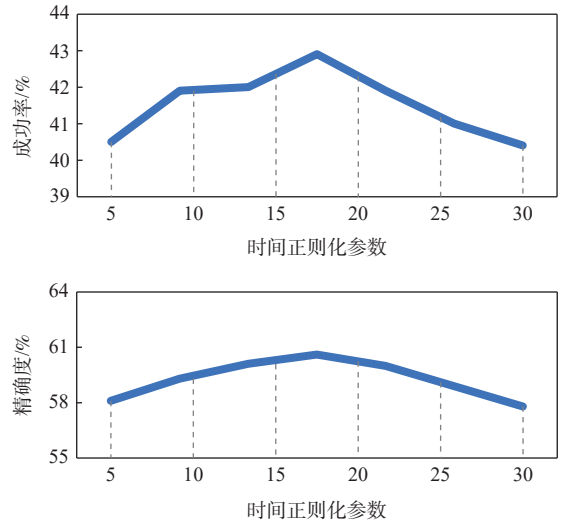


图 10 UAV20L 数据集上超参数分析

Fig. 10 Hyperparameter analysis on UAV20L dataset

表 2 消融实验结果对比

Table 2 Ablation study results %

基准数据集	精确度						成功率					
	DRCF	DRCF+Tr	DRCF+FT	DRCF+CF-VGG	DRCF+CF-VGG+Tr	DRCF+CF-VGG+FT+Tr	DRCF	DRCF+Tr	DRCF+FT	DRCF+CF-VGG	DRCF+CF-VGG+Tr	DRCF+CF-VGG+FT +Tr
UAV123[21]	69.6	66.3	67.9	70.1	69.8	**70.3**	47.9	46.4	47.7	48.3	48.4	**48.8**
UAVDT[24]	71.9	71.0	73.2	70.2	73.3	**73.2**	44.8	44.3	46.0	44.7	45.4	**45.4**
VisDrone-SOT2018[22]	78.2	80.6	78.3	76.2	79.9	**79.9**	57.3	57.6	56.8	55.6	58.5	**58.4**

注：粗体表示与基准算法相比性能增强；Base+FT、Base+CF-VGG、Base+Tr、Base+CF-VGG+Tr分别表示在基准算法中加入FT显著性目标检测模块、轻量级深度特征、时间正则化、同时加入轻量级深度特征和时间正则化；Base+CF-VGG+FT+Tr为加入所有模块的TRCF目标跟踪算法。

在 UAVDT 数据集中，包含较多飞行高度较高拍摄的小目标物体及夜间、雾天等光线原因使得目标外观模糊的序列。结果显示，只加入时间正则化和只加入 CF-VGG 的算法性能有轻微下降，但其他改进算法的跟踪性能均高于基准算法。在精确度方面，本文算法提高了 1.3%，成功率方面提高了 0.6%。

在 VisDrone-SOT2018 数据集中，依次加入时间正则化、CF-VGG 和 FT 显著性目标检测模块后，相比于基准算法，在跟踪性能上都得到了不同程度的提升。加入时间正则化模块在精确度方面提升最明显，提高了 2.4%，加入 CF-VGG 和时间正则化在成功率方面提升显著，提高了 1.2%。本文算法相较于基准算法，精确度和成功率分别提高了 1.7% 和 1.1%。

由于 UAV 数据集构建时就有针对性的集中面向某几种主要挑战或是面向某种特殊场景，例如，UAVDT 是基于 UAV 摄制的车辆交通内容的数据集，其中由于拍摄高度较高小目标对象序列较多，导致单独加入轻量级深度特征模块时，能提取到的目标语义信息有限，跟踪性能有所下降。加入 FT 显著性目标检测模块后在 UAV123 数据集上跟踪性能略有下降。这是由于显著性目标检测算法对于分辨率高、轮廓清楚的目标具有较好的提取效果，对于 UAV123 数据集存在的因为像素等原因导致的分辨率低的目标，检测能力会相对较弱。但从表 2 可知，当各模块逐层递增加入基准算法后，相较于基准算法，无论是成功率还是精确度都有不同程度的提升。不同 UAV 目标跟踪数据集所面向的主要挑战在 3 个模块的协同处理下都得到了相较于基准算法更好的解决。

2.3.2 跟踪速度分析

对于 UAV 目标跟踪问题，帧率（frames per second，FPS）是至关重要的评价指标。表 3 为各算法在 UAV 基准数据集上的帧率对比。为分析不同模块对计算复杂度及跟踪速度的影响，在 3 个 UAV 目标跟踪数据集上加入不同模块的算法进行帧率对比，实验结果如表 4 所示。由于时间正则化是对多个样本的合理近似，几乎不额外增加计算复杂度，因此单独加入时间正则化模块，帧率几乎不受影响。使用显著性目标检测算法构建正则化器后，算法整体复杂度略有上升，在 3 个数据集上平均下降了 4.3 帧/s，对于算法的整体速率影响相对较小。使用 CF-VGG 第 10 层的输出作为深度特征来描述

表 3　各算法在 UAV 数据集上的帧率比较

Table 3　FPS comparison of different algorithms on UAV datasets　　帧/s

算法	帧率			平均值
	UAVDT	VisDrone-SOT2018	UAV123	
MRCF	46.931 6	33.364 1	35.062 8	38.452 8
MSCF	28.691	24.818 9	23.393 7	25.634 5
BACF	58.749	37.088 9	39.502 7	45.113 5
DSST	*140.946*	*61.745 7*	*84.639 1*	*95.776 9*
fDSST	**201.106 6**	**146.093**	**164.268 4**	**170.489**
SAMF	14.722	7.553 3	9.955 9	10.743 7
SRDCF	15.505 8	8.074 5	10.292 3	11.290 9
STRCF	28.712	20.912 9	21.209 5	23.611 5
AutoTrack	35.213 8	41.812 4	40.865 2	39.297 1
ARCF	22.156 7	21.146 3	22.44	21.914 3
DRCF	38.362 7	30.208	31.977 4	33.516 0
TRCF	24.115 7	16.790 8	23.022 3	21.309 6

注: 粗体、斜体、下划线分别表示第1、第2、第3名。

表 4　消融实验帧率对比

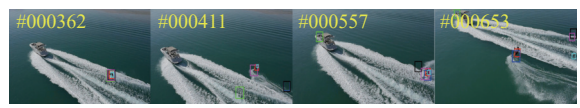
Table 4　FPS results of ablation study　　帧/s

模块	帧率			平均值
	VisDrone-SOT2018	UAV123	UAVDT	
DRCF	30.208	39.078 3	38.362 7	35.883
DRCF+Tr	27.464 7	33.701 1	41.445 6	34.203 8
DRCF+FT	25.242 4	30.748 7	38.730 3	31.573 8
DRCF+CF-VGG	19.629 8	23.057 5	29.153 9	23.947 1
DRCF+CF-VGG+FT+Tr	16.790 8	23.022 3	24.115 7	21.309 6

目标的语义信息, 该层数据输出的计算量大, 耗费时间长, 对帧率影响明显, 在 5 个数据集上的平均帧率下降了 11.935 7 帧/s。引入 3 个模块的 TRCF 目标跟踪算法计算复杂度最高, 在 VisDrone-SOT-2018 数据集上的帧率为 16.790 8 帧/s, 在 UAV123 数据集上的帧率为 23.022 3 帧/s, 在 UAVDT 数据集上的帧率为 24.115 7 帧/s, 平均帧率约为 21 帧/s, 整体仍可满足实时性的要求。需要强调的是, 本文实验均是在仅配备单 CPU 的计算机上进行, 以最大程度地模拟 UAV 运行环境。

2.4　定性分析

为了更加直观地展现 TRCF 目标跟踪算法相对于其他 UAV 目标跟踪算法的优势, 在 UAV123 数据集上进行可视化的定性分析, 结果如图 11 所示。图 11(a)为 wakeboard5 序列, 其中包括形变、部分遮挡、平面外旋转的挑战因素, 从第 362～411 帧目标被海浪部分遮挡并且伴有平面外旋转, ADTrack_v2、MSCF、MRCF 算法完全偏离目标, STRCF 及 AutoTrack 算法也发生不同程度的漂移。第 557～653 帧, 目标在运动的过程中发生旋转和形变, 只有 TRCF 目标跟踪算法能够精确跟踪目标, MRCF 算法相比于本文算法更加偏离中心, 其他 4 种算法已经远离目标中心。图 11(b)为 bird1_1 序列。从第 92～108 帧, 由于视频中数字和画质波动导致目标外观被严重干扰, 并且在目标自身也发生了平面外旋转和形变, 明显看到只有 TRCF 目标跟踪算法能够持续跟踪目标, 其他算法均发生偏离。在 uav5 序列中(见图 11(c)), 主要的挑战因素是小物体、低分辨率、尺度变化、光照变化及视角变化。从第 42 帧开始, 其他算法均已偏离目标, 并且在后续帧中目标受到剧烈的光照变化影响, 自身尺度也发生变化, TRCF 目标跟踪算法能够持续稳定地跟踪目标。图 11(d)为 person19_2 序列, 从 1 890～1 942 帧目标从视野中消失。在第 1 988 帧, 目标的一部分重新出现在视野中, 只有 TRCF 目标跟踪算法能够继续对目标进行跟踪。直到第 2001 帧, 目标再次完整出现在视野中, 其他 5 种算法已经完全丢失跟踪目标。



(a) wakeboard5 序列



(b) bird1_1 序列



(c) uav5 序列



(d) person19_2 序列

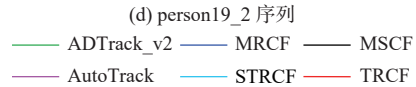


图 11　6 个先进的目标跟踪算法的跟踪结果在部分序列上的可视化对比

Fig. 11　Visualization of tracking results of six advanced object tracking algorithms on selected sequences

3　结　论

1) 引入的高效显著性目标检测算法, 有效提高了滤波器描述目标对象轮廓的能力。

2) 本文算法使用轻量级深度网络 CF-VGG 进行特征提取, 有效提高了跟踪精确度。

3) 本文算法将自适应空间正则化与时间正则化融合, 在不提高算法复杂度的前提下提高了应对复杂跟踪环境的能力。

4) 本文算法在 5 个 UAV 数据集上实现了较为优异的跟踪性能。在 UAV20L 数据集上成功率为

42.9%, 精确度为 60.6%; 在 UAV123 数据集上成功率为 48.8%, 精确度为 70.3%; 在 UAVDT 数据集上成功率为 45.4%, 精确度为 73.2%; 在 VisDrone-SOT-2018 数据集上成功率为 58.4%, 精确度为 79.9%; 在 UAVTrack112 数据集上成功率为 46.1%, 精确度为 68.0%。

在下一步工作中, 将考虑在本文所设计的框架中加入图像增强模块, 增强算法在黑暗环境跟踪目标的能力。

参考文献（References）

[1] TAHIR A, BÖLING J, HAGHBAYAN M H, et al. Swarms of unmanned aerial vehicles: a survey[J]. Journal of Industrial Information Integration, 2019, 16: 100106.

[2] YUAN C, LIU Z X, ZHANG Y M. Aerial images-based forest fire detection for firefighting using optical remote sensing techniques and unmanned aerial vehicles[J]. Journal of Intelligent & Robotic Systems, 2017, 88(2): 635-654.

[3] KANELLAKIS C, NIKOLAKOPOULOS G. Survey on computer vision for UAVs: current developments and trends[J]. Journal of Intelligent & Robotic Systems, 2017, 87(1): 141-168.

[4] JIAO L C, WANG D, BAI Y D, et al. Deep learning in visual tracking: a review[J]. IEEE Transactions on Neural Networks and Learning Systems, 2023, 34(9): 5497-5516.

[5] FU C H, LI B W, DING F Q, et al. Correlation filters for unmanned aerial vehicle-based aerial tracking: a review and experimental evaluation[J]. IEEE Geoscience and Remote Sensing Magazine, 2022, 10(1): 125-160.

[6] BOLME D S, BEVERIDGE J R, DRAPER B A, et al. Visual object tracking using adaptive correlation filters[C]//Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2010: 2544-2550.

[7] DANELLJAN M, KHAN F S, FELSBERG M, et al. Adaptive color attributes for real-time visual tracking[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2014: 1090-1097.

[8] MA C, HUANG J B, YANG X K, et al. Hierarchical convolutional features for visual tracking[C]//Proceedings of the IEEE International Conference on Computer Vision. Piscataway: IEEE Press, 2015: 3074-3082.

[9] DANELLJAN M, ROBINSON A, SHAHBAZ KHAN F, et al. Beyond correlation filters: learning continuous convolution operators for visual tracking[C]//European Conference on Computer Vision. Berlin: Springer, 2016: 472-488.

[10] DANELLJAN M, BHAT G, KHAN F S, et al. ECO: efficient convolution operators for tracking[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2017: 6931-6939.

[11] BHAT G, JOHNANDER J, DANELLJAN M, et al. Unveiling the power of deep tracking[C]//European Conference on Computer Vision. Berlin: Springer , 2018: 493-509.

[12] GALOOGAHI H K, SIM T, LUCEY S. Correlation filters with limited boundaries[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2015: 4630-4638.

[13] GALOOGAHI H K, FAGG A, LUCEY S. Learning background-aware correlation filters for visual tracking[C]//Proceedings of the IEEE International Conference on Computer Vision. Piscataway: IEEE Press, 2017: 1144-1152.

[14] YUAN Y, CHEN Y M, JING Y Q, et al. FRATCF: feature-residue real-time UAV tracking based on automatic spatio-temporal regularization correlation filter[C]//Proceedings of the IEEE International Conference on Multimedia and Expo. Piscataway: IEEE Press, 2022: 1-6.

[15] DANELLJAN M, HÄGER G, KHAN F S, et al. Learning spatially regularized correlation filters for visual tracking[C]//Proceedings of the IEEE International Conference on Computer Vision. Piscataway: IEEE Press, 2015: 4310-4318.

[16] LUKEŽIC A, VOJÍR T, ZAJC L C, et al. Discriminative correlation filter with channel and spatial reliability[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2017: 4847-4856.

[17] 王法胜, 尹双双, 贺冰, 等. 引入高斯掩膜的相关滤波目标跟踪算法[J]. 中国图象图形学报, 2023, 28(10): 3092-3106.
WANG F S, YIN S S, HE B, et al. A Gaussian mask-based correlation filter tracking algorithm[J]. Journal of Image and Graphics, 2023, 28(10): 3092-3106(in Chinese).

[18] LIN F L, FU C H, HE Y J, et al. ReCF: exploiting response reasoning for correlation filters in real-time UAV tracking[J]. IEEE Transactions on Intelligent Transportation Systems, 2022, 23(8): 10469-10480.

[19] HAN R Z, FENG W, WANG S. Fast learning of spatially regularized and content aware correlation filter for visual tracking[J]. IEEE Transactions on Image Processing, 2020, 29: 7128-7140.

[20] LI Y M, FU C H, HUANG Z Y, et al. Intermittent contextual learning for keyfilter-aware UAV object tracking using deep convolutional feature[J]. IEEE Transactions on Multimedia, 2020, 23: 810-822.

[21] MUELLER M, SMITH N, GHANEM B. A benchmark and simulator for UAV tracking[C]//European Conference on Computer Vision. Berlin: Springer, 2016: 445-461.

[22] WEN L Y, ZHU P F, DU D W, et al. VisDrone-SOT2018: the vision meets drone single-object tracking challenge results[C]//European Conference on Computer Vision. Berlin: Springer, 2019: 469-495.

[23] FU C H, CAO Z A, LI Y M, et al. Onboard real-time aerial tracking with efficient Siamese anchor proposal network[J]. IEEE Transactions on Geoscience and Remote Sensing, 2021, 60: 5606913.

[24] DU D W, QI Y K, YU H Y, et al. The unmanned aerial vehicle benchmark: object detection and tracking[C]//European Conference on Computer Vision. Berlin: Springer, 2018: 375-391.

[25] WANG Y B, WANG F S, WANG C, et al. Learning saliency-aware correlation filters for visual tracking[J]. The Computer Journal, 2022, 65(7): 1846-1859.

[26] ZHU G B, WANG J Q, WU Y, et al. MC-HOG correlation tracking with saliency proposal[C]//Proceedings of the AAAI Conference on Artificial Intelligence. Palo Alto: AAAI Press, 2016, 30(1): 3690-3696.

[27] FU C H, XU J T, LIN F L, et al. Object saliency-aware dual regularized correlation filter for real-time aerial tracking[J]. IEEE Transactions on Geoscience and Remote Sensing, 2020, 58(12): 8940-8951.

[28] HOU X D, ZHANG L Q. Saliency detection: a spectral residual approach[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2007: 1-8.

[29] ACHANTA R, HEMAMI S, ESTRADA F, et al. Frequency-tuned salient region detection[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2009: 1597-1604.

[30] WANG N, ZHOU W G, SONG Y B, et al. Real-time correlation tracking via joint model compression and transfer[J]. IEEE Transactions on Image Processing, 2020, 29: 6123-6135.

[31] CHATFIELD K, SIMONYAN K, VEDALDI A, et al. Return of the devil in the details: delving deep into convolutional nets[C]//Proceedings of British Machine Vision Conference. Nottingham: BMVA Press, 2014: 1-11.

[32] HINTON G, VINYALS O, DEAN J. Distilling the knowledge in a neural network[C]//Proceedings of IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE Press, 2021: 10387-10396.

[33] ROMERO A, BALLAS N, KAHOU S E, et al. FitNets: hints for thin deep nets[EB/OL].(2015-03-27)[2023-06-01]. http://arxiv.org/abs/1412.6550.

[34] DANELLJAN M, HÄGER G, KHAN F S, et al. Accurate scale estimation for robust visual tracking[C]//Proceedings of the British Machine Vision Conference. Nottingham: BMVA Press, 2014: 1-11.

[35] LI F, TIAN C, ZUO W M, et al. Learning spatial-temporal regularized correlation filters for visual tracking[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2018: 4904-4913.

[36] WU Y, LIM J, YANG M H. Online object tracking: a benchmark [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2013: 2411-2418.

[37] LI Y M, FU C H, DING F Q, et al. AutoTrack: towards high-performance visual tracking for UAV with automatic spatio-temporal regularization[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2020: 11920-11929.

[38] LI B W, FU C H, DING F Q, et al. All-day object tracking for unmanned aerial vehicle[J]. IEEE Transactions on Mobile Computing, 2023, 22(8): 4515-4529.

[39] YE J J, FU C H, LIN F L, et al. Multi-regularized correlation filter for UAV tracking and self-localization[J]. IEEE Transactions on Industrial Electronics, 2022, 69(6): 6004-6014.

[40] ZHENG G Z, FU C H, YE J J, et al. Mutation sensitive correlation filter for real-time UAV tracking with adaptive hybrid label[C]//Proceedings of the IEEE International Conference on Robotics and Automation. Piscataway: IEEE Press, 2021: 503-509.

[41] DAI K N, WANG D, LU H C, et al. Visual tracking via adaptive spatially-regularized correlation filters[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2019: 4665-4674.

[42] FU C H, JIN J, DING F Q, et al. Spatial reliability enhanced correlation filter: an efficient approach for real-time UAV tracking[J]. IEEE Transactions on Multimedia, 2021, 26: 4123-4137.

[43] BERTINETTO L, VALMADRE J, GOLODETZ S, et al. Staple: complementary learners for real-time tracking[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2016: 1401-1409.

[44] DANELLJAN M, HÄGER G, KHAN F S, et al. Discriminative scale space tracking[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(8): 1561-1575.

[45] LI Y, ZHU J. A scale adaptive kernel correlation filter tracker with feature integration[C]//European Conference on Computer Vision. Berlin: Springer, 2015: 254-265.

Saliency-aware triple-regularized correlation filter algorithm for UAV object tracking

HE Bing，WANG Fasheng*，WANG Xing，SUN Fuming

(School of Information and Communication Engineering，Dalian Minzu University，Dalian 116600，China)

Abstract：Object tracking in unmanned aerial vehicle (UAV) scenarios has been widely applied in many real-world tasks. Different from general object tracking, UAV object tracking is more easily affected by complex environmental interferences and computational limitations. In this paper, a saliency-aware triple-regularized correlation filter (TRCF) for UAV object tracking was proposed. An efficient saliency detection algorithm was used to dynamically generate a dual-spatial regularizer to suppress boundary effects and apply a penalty to irrelevant background noise coefficients. Temporal regularization was introduced to address the filter degradation problem caused by appearance variation, providing a more robust appearance model. In addition, a lightweight deep network CF-VGG was employed to extract deep features, which were linearly fused with hand-crafted features to describe the object's semantic information with improved tracking accuracy. Experiments were conducted on five publicly available UAV benchmark datasets. Results show that compared to the baseline methods, the proposed method demonstrates improvements on the five datasets, proving its effectiveness and robustness. The method also shows a real-time tracking speed of about 21 frames per second, making it suitable for UAV object-tracking tasks.

Keywords：unmanned aerial vehicle object tracking；correlation filter；saliency feature map；temporal regularization；lightweight deep network

Received：2023-06-15；**Accepted**：2023-09-15；**Published Online**：2023-12-07 07：00

URL：link.cnki.net/urlid/11.2625.V.20231206.1417.005

Foundation items：National Natural Science Foundation of China (61972068,61976042); Liaoning Revitalization Talents Program (XLYC2007023)

* **Corresponding author**. E-mail：wangfasheng@dlnu.edu.cn

http://bhxb.buaa.edu.cn jbuaa@buaa.edu.cn
DOI: 10.13700/j.bh.1001-5965.2023.0366

基于无监督深度学习的航拍图像拼接算法

梁镇锋[1,2]，夏海英[1,2,3,*]，谭玉枚[1,2]，宋树祥[1,2]

(1. 广西师范大学 广西类脑计算与智能芯片重点实验室，桂林 541004；2. 广西师范大学 广西高校集成电路与微系统重点实验室，桂林 541004；3. 广西师范大学 光电信息技术广西高校工程研究中心，桂林 541004)

摘　　要：传统的图像拼接算法过度依赖特征的准确定位或分布，导致在复杂的航拍场景下鲁棒性差。因此，提出了一个完整的无监督深度学习航拍图像拼接框架，其由无监督深度单应性估计网络和无监督图像融合网络组成。无监督深度单应性估计网络旨在通过学习参考图像和目标图像之间的单应性变换，为后续的拼接工作提供准确的对齐信息；无监督图像融合网络用于学习航拍图像拼接的变形规则，生成最终的拼接结果。为了训练所提学习框架，提供了一个用于无监督航拍图像拼接的真实数据集，比较了尺度不变特征变换(SIFT)+Ransac、加速非线型扩散特征检测与匹配(AKAZE)+增强型高效二进制局部图像描述符(BEBLID)、基于BRIEF算法的快速二值特征向量(ORB)+Ransac和基于深度学习的图像拼接算法，实验结果表明，结构相似性指数(SSIM)提高了39.94%，峰值信噪比(PSNR)提高了36.55%，均方根误差(RMSE)降低了66.09%。此外，所提算法在真实的航拍场景下相较于现有的基于深度学习和传统的图像拼接算法具有更好的视觉拼接效果和鲁棒性。

关　键　词：航拍图像；深度单应性估计；图像拼接；无监督；图像融合

中图分类号：TP391

文献标志码：A　　**文章编号**：1001-5965(2025)07-2437-13

近年来，无人机技术得到了快速的发展。无人机的高效、灵活、低成本和高分辨率等特性使其成为航拍的主流方式[1]。然而，在航拍过程中，由于飞行高度和相机焦距等因素的限制，单张航拍图像的视野范围通常较小，往往无法完整地覆盖需要观察的整个区域。因此，为了获取目标区域的全局信息，需要将多张航拍图像拼接成一幅全景图像[2]。

图像拼接技术是计算机视觉领域的重要研究课题，广泛应用于多个领域，如航空航天、地质勘探、虚拟现实等领域。然而，目前已有的图像拼接方法主要针对自然图像设计，而对于航拍图像的拼接仍然存在许多挑战。首先，由于特殊的拍摄角度和复杂的环境条件，航拍图像与传统自然图像存在显著差异。航拍图像通常具有高度重复的纹理和结构[3]，如建筑物、树木、道路等，这使得传统的特征点匹配方法容易出现误匹配，导致拼接失败。其次，航拍图像在拍摄过程中普遍存在姿态、尺度和光照等变化，这些变化会导致不同图像之间的颜色和亮度差异较大，从而影响图像拼接的质量。此外，由于飞行高度和相机视角的限制，相邻图像之间通常存在显著的视角变化和重叠区域变形，这也对图像拼接算法提出了更高的要求。

因此，本文针对航拍场景下的图像拼接问题，提出了一种基于无监督深度学习的航拍图像拼接算法，主要贡献如下：

1）鉴于目前缺乏用于航拍图像拼接的标准数

收稿日期：2023-06-15；录用日期：2024-03-29；网络出版时间：2024-04-22 21:15
网络出版地址：link.cnki.net/urlid/11.2625.V.20240419.1511.004
基金项目：广西揭榜制科技项目(桂科AB23026103)；国家自然科学基金(62106054)；广西创新驱动重大专项(桂科AA20302003)
*通信作者. E-mail：xhy22@mailbox.gxnu.edu.cn

引用格式：梁镇锋，夏海英，谭玉枚，等. 基于无监督深度学习的航拍图像拼接算法[J]. 北京航空航天大学学报，2025，51(7)：2437-2449.
LIANG Z F，XIA H Y，TAN Y M，et al. Aerial image stitching algorithm based on unsupervised deep learning[J]. Journal of Beijing University of Aeronautics and Astronautics，2025，51(7)：2437-2449 (in Chinese).

据集，提出了一个适用于无监督航拍图像拼接研究的航拍图像数据集。

2）设计了一个级联的双分支深度单应性估计网络，通过自下而上的学习策略来精确估计待拼接航拍图像之间的单应性变换。

3）设计了一个无监督图像融合网络，采用无监督学习的方式，学习图像拼接的变形规则，同时结合深度单应性估计网络，构建了一个完整的无监督端到端的图像拼接框架。

1 相关工作

1.1 传统的图像拼接算法

传统的基于特征的图像拼接算法一直是图像拼接的主流方法。然而，这些算法对于特征点的精确定位存在严重依赖，当图像纹理复杂或具有较大的尺度变化时，图像拼接很可能会失败。Chen等[4]提出了一种基于最优接缝算法和半投影变换的无人机图像拼接算法，对拼接图像的形状进行校正，以保持更多区域的原始形状，但当无人机图像中的结构物体非常密集时，会增加寻找合适接缝的难度，可能导致图像拼接不完美。Jong等[5]提出了一种基于Log-Gabor特征变换的有效特征检测方法，该方法不依赖于特征点检测的图像强度，而是利用Log-Gabor函数构建多个缩放和定向的Log-Gabor尺度空间，有效提取更独特和匹配的特征，在处理近乎均匀场景时，该方法具有更好的特征匹配性能，但处理非平面的航拍场景图像时效果不佳。Zhang等[6]提出了一种基于人类视觉系统(human visual system, HVS)和尺度不变特征变换(scale-invariant feature transform, SIFT)算法的拼接策略，结合SIFT算法和HVS对拼接图像的重叠区域进行划分，建立属性关系模型，但采用SIFT算法作为拼接过程中的特征提取方法，存在计算量大、速度慢等问题。宋飞等[7]针对丘陵山区耕地小型无人机航拍图像中的尺度变化、几何畸变、图像重叠等问题，提出了基于双特征的丘陵山区耕地低空遥感图像配准算法，采用加速鲁棒特征(speeded up robust features, SURF)算法提取图像的特征点，并构建了能够稳健描述航拍图像几何特征的双特征描述子。宗慧琳等[8]针对在泥石流区灾害应急中使用无人机高分辨率影像特征匹配时时效性较低的问题，提出了一种改进加速非线性扩散特征检测与匹配(accelerated-nonlinear diffusion-based feature detection and matching, AKAZE)无人机影像特征匹配的算法，该算法使用AKAZE特征点检测算法提取局部稳定不变特征，用增强型高效二进制局部图像描述符(boosted efficient binary local image descriptor, BEBLID)描述检测到的特征点，采用最近邻次近邻距离比完成初步匹配。Zaragoza等[9]提出的尽可能投影变换(as-projective-as-possible, APAP)算法旨在实现图像的投影变换，使得变换后的图像尽可能地保持原始图像的投影属性。Lin等[10]提出的自适应尽可能自然(adaptive as-natural-as-possible, AANAP)算法则旨在保持图像变换的自然性和真实性。Chang等[11]提出的保形半投影变换算法是一种半投影变换算法，其目的是在图像变换的同时尽量保持图像中物体的形状和结构。Chen等[12]提出的全局相似性先验算法旨在利用图像的全局特征和结构信息，以先验知识的形式指导后续的图像配准任务。然而，这些算法都需要将图像划分为密集网格，同时每个网格都用一个单应性矩阵对齐，这将导致图像拼接的时间大幅度增加，不适用于具有高分辨率的航拍图像[13]。事实上，在航拍场景下，由于飞行高度和角度的变化，图像的景深会发生明显变化，使得以上的传统图像拼接算法难以精确地完成拼接任务。此外，由于航拍图像中的场景复杂多变，存在遮挡、树木枝干等场景结构重复的情况，进一步增加了图像拼接的难度。

1.2 基于深度学习的图像拼接算法

随着深度学习技术的发展，基于深度学习的图像拼接也逐渐受到研究者们的关注，并取得了一些令人瞩目的成果。然而，一些深度学习拼接算法[14-15]中只有部分步骤(如特征提取或特征匹配)是通过深度学习实现的，因此，不能算作完整的深度图像拼接模型。为了解决这一问题，Nie等提出了无视角深度图像拼接算法[16-17]，但这些算法只能在“无视差”的合成数据集上进行训练，限制了其真实场景中的应用效果，随后，又提出了一种无监督的深度图像拼接算法[18]，其主要思想是通过无监督的图像重建网络，消除从特征到像素的伪影，从而提高拼接效果，但对于复杂的航拍场景，该算法的拼接效果并不理想，仍需要进一步改进和优化。Zhu等[19]提出了一种改进的VGG16孪生网络的无人机遥感图像拼接模型，其中，网络的前13层被截取形成共享权重特征提取网络，并引入改进的压缩-激励模块来有效提取图像重叠区域的特征。Huang等[20]构建了一个深度感知像素级单应性估计网络(depth-aware pixel-specific homography estimation network, DPH-Net)，通过估算特定像素的单应性矩阵参数，可以轻松计算候选图像之间的像素对应关系以进行配准。马腾宇等[21]针对传统方法对医学图像配准计算成本高、运行时间长的问题，提出了

一种基于无监督学习的深度多模态可变形图像配准框架, 但该方法适用于医学图像配准, 对于航拍场景下的图像拼接存在一定的困难。为此, 本文提出了一个完整的无监督深度学习航拍图像拼接框架。与传统图像拼接算法相比, 本文算法利用卷积神经网络自动提取图像特征, 并通过学习得到参考图像和目标图像之间的变换矩阵, 从而实现图像的无缝拼接。与基于深度学习的有监督图像拼接算法相比, 本文算法采用无监督方式, 无需手动标注训练数据。所提出的双分支深度单应性估计网络能够更好地捕捉复杂的图像结构, 从而准确地估计单应性。同时, 本文提出的无监督图像融合网络引入了密集连接和残差路径, 以加强特征的深度传播, 进一步提高模型对图像语义信息的理解能力。这种设计使模型能够更加准确地对齐图像, 从而在真实的航拍场景中获得更佳的效果。

2　网络模型

本文提出了一个完整的无监督深度学习航拍图像拼接框架, 该框架包括无监督深度单应性估计网络和无监督图像融合网络 2 个阶段: ①通过双分支的深度单应性估计网络得到参考图像和目标图像之间的单应性变换, 为后续的图像融合提供准确的对齐信息; ②通过无监督图像融合网络得到最终的航拍图像拼接全景图。基于无监督深度学习的航拍图像拼接总体架构如图 1 所示。

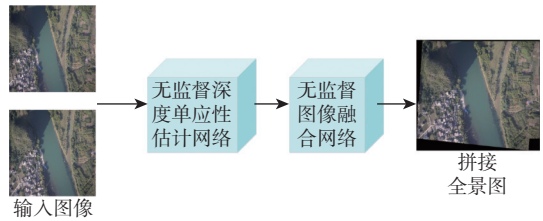


图 1　无监督深度学习航拍图像拼接框架

Fig. 1　Unsupervised deep learning framework for aerial image stitching

2.1　无监督深度单应性估计网络

在航拍图像拼接任务中, 传统算法需依次执行特征检测、特征匹配、单应性估计、图像配准和图像融合等紧密关联的步骤。直接将此类复杂流程整合至单一深度学习网络非常具有挑战性。为简化网络结构, 本文提出了一个预训练的级联双分支深度单应性估计网络。该网络的核心功能在于计算图像间的投影变换, 为后续的拼接工作提供准确的对齐信息。如图 2 所示, 该网络由低分辨率分支(见图 2 顶部)和高分辨细化分支(见图 2 底部)组成, 可通过自下而上的方式实现对图像 $\boldsymbol{I}^{\mathrm{A}}$ 和 $\boldsymbol{I}^{\mathrm{B}}$ 之间单应性矩阵 $\boldsymbol{H}$ 的估计和细化。

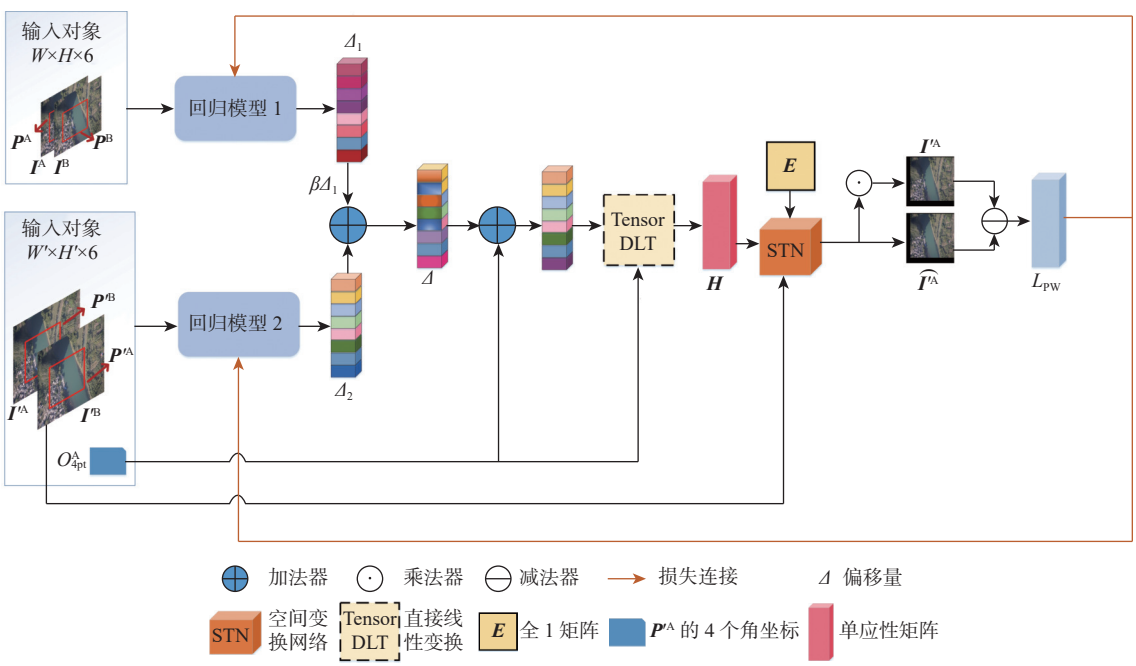


图 2　无监督深度单应性估计网络

Fig. 2　Unsupervised deep homography estimation network

2.1.1　低分辨率分支

尽管已有基于深度学习的单应性估计方法, 但准确性不足以解决航拍图像的拼接问题。为此, 本文提出的单应性估计网络包含一个低分辨率分支, 使得该网络可以精确地估计单应性, 以适用于航拍图像拼接。

首先, 在低分辨率分支中, 回归模型 1 选用 SqueezeNet 作为主干网络。因其轻量级架构和高效性能,

适用于对大规模数据进行训练。由于低分辨率分支主要负责处理图像的全局信息，因此采用轻量级模型可以降低模型复杂度和计算成本，同时保持良好的性能。SqueezeNet[22]是一种由多个Fire模块组成的深度卷积网络，包括卷积层、池化层、全连接层等多个层级。每个Fire模块由Squeeze部分和Expand部分组成，Squeeze部分通过1×1卷积将输入数据压缩成较少的通道数，以降低计算成本，Expand部分则采用1组连续的1×1卷积和3×3卷积拼接的方式来增加通道数，以扩展特征表达能力。Fire模块的结构如图3所示。网络输出能够唯一确定单应性的8个坐标偏移量Δ，而不是传统的3×3单应性矩阵$\boldsymbol{H}$。

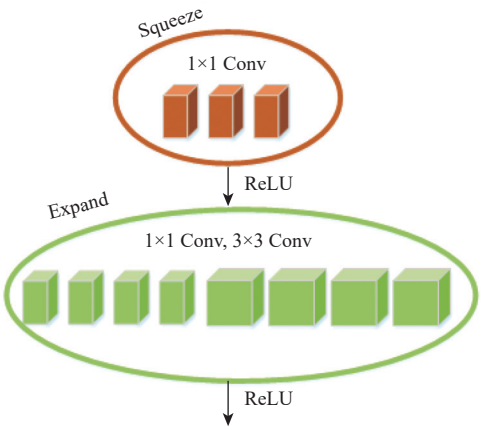


图3 Fire模块示意图

Fig. 3 Fire module

在具体的实现中，首先，待拼接的航拍图像被下采样到低分辨率。然后，以堆叠的图像块$\boldsymbol{P}^{\mathrm{A}}$和$\boldsymbol{P}^{\mathrm{B}}$（如图2顶部矩形框所示）作为网络的输入，图像块分别从图像$\boldsymbol{I}^{\mathrm{A}}$和$\boldsymbol{I}^{\mathrm{B}}$的中心位置裁剪，大小为128×128。其次，图像被馈送到网络，通过1个7×7的卷积层，使用8个Fire模块（Fire2～Fire9），每个模块中卷积核数量逐渐增加，并且在Fire4、Fire8这几层之后使用步长为2的最大池化。为了保留更多的信息，网络的下采样操作被延后，使得卷积层可以得到更大的激活图。最后，通过2个全连接层得到偏移量Δ，将预测的偏移量通过直接线性变换（direct linear transformation，DLT）算法[23]转换为相应的张量单应性矩阵$\boldsymbol{H}$。DLT算法可以作为一种基本方法，用于估计从不同视角捕获的2幅图像之间的几何关系。在获得图像对的单应性矩阵$\boldsymbol{H}$后，使用空间变换网络（spatial transformer network，STN）[24]得到变换后的图像，其过程如图4所示。其中，θ为变换参数，$T_\theta(G)$为对目标图像以参数θ为变换参数对图像进行变换。

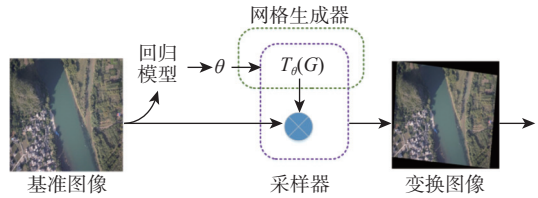


图4 空间变换过程示意图

Fig. 4 Spatial transformation process

2.1.2 高分辨率细化分支

针对航拍图像所具有的高分辨率和广泛覆盖范围的特点，本文引入了一个高分辨率细化分支，旨在进一步提升预测偏移量的精度。在回归模型2中，采用ResNet-34[25]结构作为该分支的骨干网络，因其具有足够的深度和多层残差连接，能够有效地学习图像的抽象特征和语义信息，可以从特征层面提取更多的细节信息，适合处理细节丰富的高分辨率图像。网络的输入包含2部分：①大小为512×512×6的堆叠图像块$\boldsymbol{P}'^{\mathrm{A}}$和$\boldsymbol{P}'^{\mathrm{B}}$。其中，2个图像块是从图像$\boldsymbol{I}'^{\mathrm{A}}$和$\boldsymbol{I}'^{\mathrm{B}}$的中心位置裁剪而来。②图像块$\boldsymbol{P}'^{\mathrm{A}}$的4个角坐标，表示为$O_{4\mathrm{pt}}^{\mathrm{A}}$。表1给出了网络的具体参数，该网络生成固定大小的8维特征向量，表示为8个坐标偏移量。

表1 网络各层参数及特征图大小

Table 1 Parameters of each layer of network and feature map size

层名称	层操作	特征图大小/像素
Con1_x	7×7, 64, 步长=2	256×256
最大池化	3×3, 步长=2	128×128
Con2_x	$\begin{bmatrix}3\times3,64\\3\times3,64\end{bmatrix}\times3$	128×128
Con3_x	$\begin{bmatrix}3\times3,128\\3\times3,128\end{bmatrix}\times4$	64×64
Con4_x	$\begin{bmatrix}3\times3,256\\3\times3,256\end{bmatrix}\times6$	32×32
Con5_x	$\begin{bmatrix}3\times3,512\\3\times3,512\end{bmatrix}\times3$	16×16
平均池化	2×2, 步长=2	8×8
全连接层	8个偏移量	1×1

2.1.3 损失函数

由于有监督方式的深度单应性估计需要标签，因此在真实的航拍图像应用中受到限制，导致有监督的方法不可靠。本文以图像块作为网络的输入，当强制变换目标图像靠近参考图像时，不再填充变换图像中的无效像素。给定图像对$\boldsymbol{P}^{\mathrm{A}}$和$\boldsymbol{P}^{\mathrm{B}}$，其离散像素位置用齐次坐标$\boldsymbol{x}_i=(x,y,1)^{\mathrm{T}}$表示，本文希望网络输出最小化平均$L_1$像素光度损失。无监督单应性估计网络的目标函数如下：

$$L_{\mathrm{PW}}=\frac{1}{|\boldsymbol{x}_i|}\sum_{\boldsymbol{x}_i}\left|\boldsymbol{H}(\boldsymbol{E})\odot\boldsymbol{P}^{\mathrm{A}}(\boldsymbol{x}_i)-\boldsymbol{P}^{\mathrm{B}}\boldsymbol{H}(\boldsymbol{x}_i)\right| \tag{1}$$

式中: ⊙表示逐像素乘法; $\boldsymbol{P}^{\mathrm{A}}$、$\boldsymbol{P}^{\mathrm{B}}$分别为从原始图像$\boldsymbol{I}^{\mathrm{A}}$、$\boldsymbol{I}^{\mathrm{B}}$中心位置处裁剪出的图像块; $\boldsymbol{E}$为与P^{A}大小相同的全 1 矩阵; $\boldsymbol{H}(\boldsymbol{E})$表示对矩阵$\boldsymbol{E}$做单应性变换。本文选择 l_1 损失, 因为之前的工作观察到其更适合图像配准问题[26]。

2.2　无监督图像融合网络

从深度单应性估计网络中获得参考图像和目标图像之间的单应性矩阵$\boldsymbol{H}$后, 需要将变换后的图像进行拼接处理以获得最终的拼接全景图。为此, 本文设计了一个无监督图像融合网络, 利用无监督学习的方式, 学习航拍图像拼接的变形规则。图像融合网络与深度单应性估计网络相结合, 构成了一个完整的无监督端到端的图像拼接框架。无监督图像融合网络整体框架如图 5 所示。

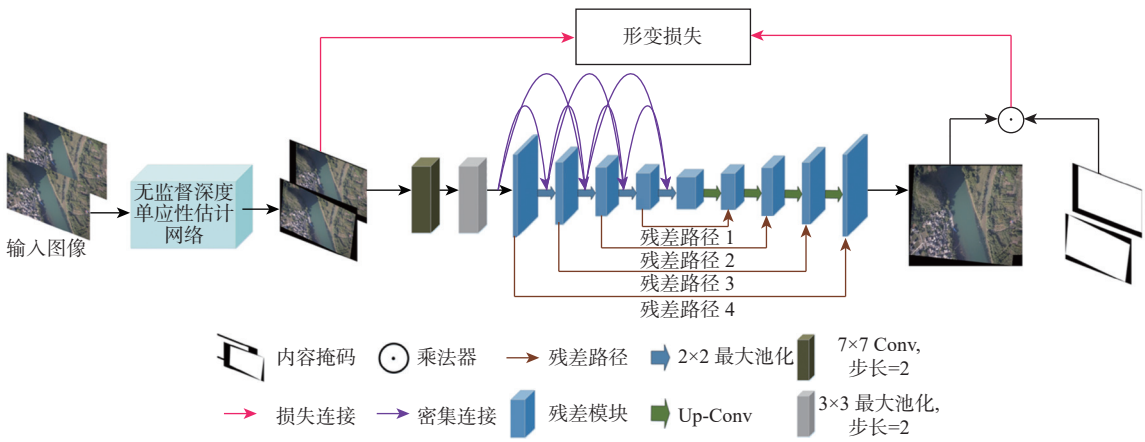


图 5　无监督图像融合网络整体框架

Fig. 5　Overall framework of image fusion network

无监督图像融合网络模型采用编码器-解码器结构, 以实现端到端的训练。编码器和解码器均由残差模块组成, 如图 6 所示。

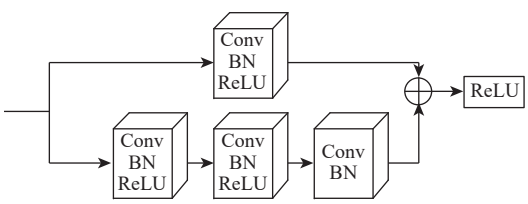


图 6　残差模块示意图

Fig. 6　Residual module

2.2.1　编码网络

无监督图像融合网络采用 ResNet50[25] 作为编码器, 编码器部分由 4 个残差模块组成, 共实现了 4 次下采样。首先, 对输入图像进行了卷积核大小为 7×7、通道数为 64 的卷积操作, 并通过 3×3 的最大池化层进行下采样。然后, 通过引入 4 组不同的残差模块对输入特征向量进行操作以有效避免深度网络中出现的梯度消失和网络退化问题。本文通过在残差模块之间插入 6 个密集连接, 并使用最大池化操作统一特征向量的维度, 构建了一种稠密的神经网络结构。该结构能够充分利用浅层特征, 增强特征的深度延伸, 进而提高网络的特征提取和表示能力。

2.2.2　解码网络

解码器部分包含 4 个上采样和 4 个残差模块, 残差模块如图 6 所示。由于编码器中卷积层较浅, 其提取的特征较为浅层, 而解码器中对应的特征卷积层较深, 其提取的特征则更深层次。因此, 在语义上存在较大差距。为了解决这一问题, 本文提出了一种由多个残差块构成的残差路径。每个残差块都由 2 个 3×3 卷积和 1 个 1×1 卷积组成, 形成 1 个固定的残差块结构。

残差路径中的残差块数量逐渐减少, 图 5 中的 4 个残差路径分别使用 4、3、2、1 个残差块。这样的设计可以弥补编码器卷积层空间特征和解码器卷积层的语义特征之间的差距, 促进信息传播, 并在拼接前尽可能地使编码器和解码器的深度保持一致。此外, 为了保留图像特征信息, 残差路径不使用批归一化(batch normalization, BN)和 ReLU 激活函数。本文选取第 1 条残差路径结构进行展示, 如图 7 所示。

本文提出的无监督图像融合网络能够提高信息提取完整性, 并增强模型对复杂环境的适应性。采用 ResNet 作为编码器, 具有较强的特征提取能力和高效的计算速度, 能够显著减少模型训练时间。同时, 通过跨层连接简化了网络训练, 缓解了梯度爆炸问题, 还可以建立更深层的网络结构。此外, 本文采用的编解码网络具有多尺度结合和特征拼接的特点, 可以提高网络的融合精度, 使网络更加注重细节信息的捕获。

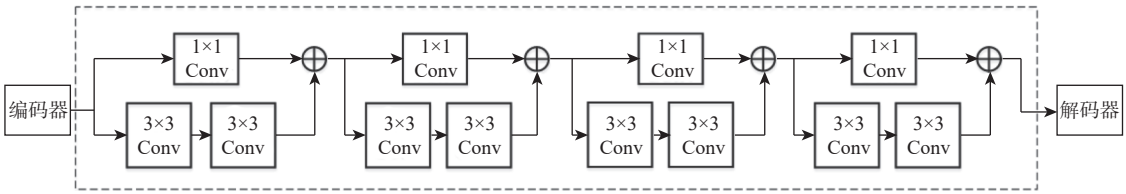


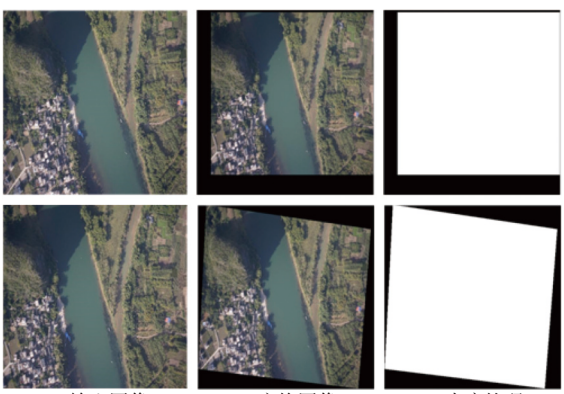
图7 残差路径示意图

Fig. 7 Residual path

图像融合网络通过内容掩码学习图像拼接的变形规则，使用内容掩码来约束拼接图像的特征，以使其接近于变换后的图像。网络的输入包括经过第1阶段单应性估计网络变换后的图像$\boldsymbol{I}^{\mathrm{AT}}$、$\boldsymbol{I}^{\mathrm{BT}}$以及内容掩码$\boldsymbol{M}^{\mathrm{A}}$、$\boldsymbol{M}^{\mathrm{B}}$，如图8所示。本文通过式(2)获得变换后的图像$\boldsymbol{I}^{\mathrm{AT}}$和$\boldsymbol{I}^{\mathrm{BT}}$，并且通过将全1矩阵$\boldsymbol{E}$代替等式中的$\boldsymbol{I}^{\mathrm{A}}$和$\boldsymbol{I}^{\mathrm{B}}$，得到图像$\boldsymbol{I}^{\mathrm{A}}$和$\boldsymbol{I}^{\mathrm{B}}$的内容掩码$\boldsymbol{M}^{\mathrm{A}}$、$\boldsymbol{M}^{\mathrm{B}}$。

$$\begin{cases}\boldsymbol{I}^{\mathrm{AT}}=\boldsymbol{T}(\boldsymbol{I}^{\mathrm{A}},\boldsymbol{X})\\ \boldsymbol{I}^{\mathrm{BT}}=\boldsymbol{T}(\boldsymbol{I}^{\mathrm{B}},\boldsymbol{H})\end{cases} \tag{2}$$

式中：$\boldsymbol{X}$为单位矩阵；$\boldsymbol{T}$为使用3×3单应性矩阵对图像变换。



(a) 输入图像 (b) 变换图像 (c) 内容掩码

图8 输入图像、变换图像和内容掩码示意图

Fig. 8 Input image, transformed image, and content mask

2.2.3 损失函数

无监督图像融合网络的目标函数如下：

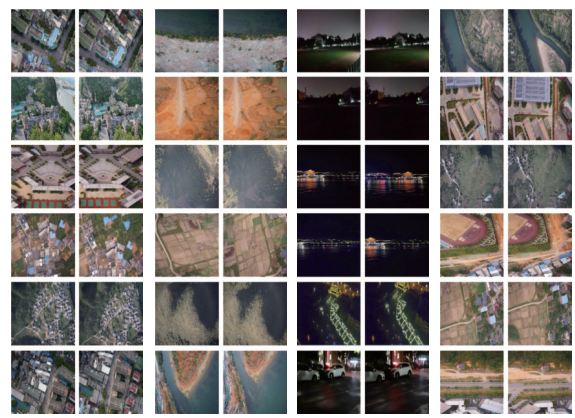
$$L_{\mathrm{Fusion}}=l_{\mathrm{p}}(\boldsymbol{S}_{\mathrm{Result}}\odot\boldsymbol{M}^{\mathrm{A}},\boldsymbol{I}^{\mathrm{AT}})+l_{\mathrm{p}}(\boldsymbol{S}_{\mathrm{Result}}\odot\boldsymbol{M}^{\mathrm{B}},\boldsymbol{I}^{\mathrm{BT}})。 \tag{3}$$

式中：$\boldsymbol{S}_{\mathrm{Result}}$为拼接后的图像；$l_{\mathrm{p}}$为感知损失[27]。为了使拼接图像的特征尽可能接近于变换目标图像的特征，本文采用VGG-19[28]网络中的“Conv4_3”层计算感知损失，该层具有足够的深度，可以缩小变换图像之间的特征差异。

3 实验和结果对比分析

3.1 航拍图像数据集

鉴于目前缺乏用于航拍图像拼接的标准数据集，为了训练图像拼接网络，本文提出了一个适用于无监督深度学习航拍图像拼接的数据集。该数据集由DJ Min2无人机拍摄的航拍图像组成，总计包含5万对图像。考虑到无人机在航拍过程中容易受到气流、恶劣天气等干扰因素的影响，图像可能存在尺度差异、旋转角度大、质量差等挑战。为了客观描述该数据集，将其分为低纹理图像、夜晚图像、复杂纹理图像和常规图像4个类别，每个类别约有12 000对图像。此外，从所有类别中选取4万对图像用于训练，随机选择3 000对图像用于测试。部分数据集示例如图9所示。



(a) 复杂纹理图像 (b) 低纹理图像 (c) 夜晚图像 (d) 常规图像

图9 航拍图像数据集示例

Fig. 9 Aerial image dataset

3.2 实验环境

采用搭载英伟达RTX3090显卡的计算机设备作为实验硬件环境，操作系统为Window10，CPU为Inter i7-12700K，GPU为RTX 3090 24 GB，深度学习框架为Pytorch，实验数据评估平台为Pycharm，并在训练和测试过程中保持环境的一致性。

3.3 实验细节

训练过程分为2个阶段：无监督深度单应性估计网络和无监督图像融合网络。首先，无监督深度单应性估计网络使用Adam优化器进行了150个epoch的训练，初始学习率为0.000 1，学习率按指数衰减策略进行调整。值得强调的是，所有的训练过程均为无监督学习方式，这意味着本框架仅使用待拼接的航拍图像作为输入，无需标签信息进行监督。在完成单应性估计网络的训练后，对本文的无监督图像融合网络进行训练，其训练策略与单应性估计网络相同，但最大训练周期设置为50。此外，

2 个训练步骤的批次大小设置为 16。

3.4　单应性估计的比较

为了客观评估所提出的无监督深度单应性估计网络的性能, 在真实的无人机航拍数据集上与 3×3 的单位矩阵 $\boldsymbol{X}_{3\times3}$、SIFT+Ransac[29]、AKAZE+BEBLID[8]、基于 BRIEF 算法的快速二值特征向量(oriented brief, ORB)+Ransac[30]、深度单应性网络(deep homography network, DHN)[31]、文献 [19] 算法、DPH-Net[20]、无监督深度单应性网络(unsupervised deep homography network, UDHN)[32] 和无监督深度图像拼接(unsupervised deep image stitching, UDIS)[18] 这 9 种算法进行了比较。本文采用图像重叠区域的峰值信噪比(peak signal-to-noise ratio, PSNR)[33]、结构相似性指数(structural similarity index measure, SSIM)[34] 和均方根误差(root mean square error, RMSE)[35] 作为性能评价指标, 定量评价结果如表 2 所示。PSNR 和 SSIM 值越大, 代表配准性能越好。RMSE 值越小, 图像拼接效果越好。

表 2　不同算法的单应性比较

Table 2　Homography estimation results of different methods

算法		平均PSNR/dB	平均SSIM	平均RMSE
传统单应性估计	$\boldsymbol{X}_{3\times3}$	14.233 6	0.243 4	16.796 3
	SIFT+Ransac[29]	23.832 5	0.748 5	7.267 5
	AKAZE+BEBLID[8]	22.132 3	0.624 1	7.410 3
	ORB+Ransac[30]	21.632 5	0.598 6	8.887 5
深度单应性估计(有监督)	DHN[31]	19.956 3	0.613 1	5.223 6
	文献[19]算法	23.972 3	0.754 3	4.921 6
	DPH-Net[20]	22.835 6	0.741 2	6.623 5
深度单应性估计(无监督)	UDHN[32]	21.852 6	0.672 1	5.120 3
	UDIS[18]	25.052 1	0.802 3	4.265 1
	本文	27.251 3	0.837 7	3.013 6

其中, $\boldsymbol{X}_{3\times3}$ 用于对参考图像进行无变换的单应性操作。SIFT+Ransac、AKAZE+BEBLID 和 ORB+Ransac 作为传统单应性算法的代表。DHN、文献 [19] 方法、DPH-Net、UDHN 和 UDIS 为基于深度学习的算法, 其中, DHN、文献 [19] 算法、DPH-Net 为有监督算法, UDHN 和 UDIS 为无监督算法。

本文测试了 10 组不同的图像对, 从表 2 可以看出, 与 SIFT+Ransac、AKAZE+BEBLID、ORB+Ransac 算法相比, 本文算法的 SSIM 值分别增大了 11.92%、34.23% 和 39.94%; PSNR 值分别增大了 14.35%、23.13% 和 25.97%; RMSE 值分别减小了 58.53%、59.33% 和 66.09%。与 DHN、文献 [19] 算法、DPH-Net 相比, 本文算法的 SSIM 值分别增大了 36.63%、11.06% 和 13.02%; PSNR 值分别增大了 36.55%、13.68% 和 19.34%; RMSE 值分别减小了 42.3%、38.77% 和 54.5%。与 UDHN、UDIS 算法相比, 本文算法的 SSIM 值分别增大了 24.64%、4.41%; PSNR 值分别增大了 24.71%、8.78%; RMSE 值分别减小了 41.14%、29.34%。实验表明, 通过对 SSIM、PSNR、RMSE 这 3 个客观评价指标与其他算法的比较, 证明了本文算法对高质量无人机航拍图像拼接的有效性和可靠性。

3.5　拼接结果的可视化

为了验证本文算法在航拍图像中的优越性, 与传统的图像拼接算法进行了比较, 并与最近的基于深度学习的图像拼接算法进行了比较。

3.5.1　与传统的图像拼接算法的比较

选择 APAP[9]、鲁棒性误差等级分析(robust error level analysis, Robust ELA)[36]、AANAP[10]、SURF+Ransac[37]、SIFT+Ransac[29]、AKAZE+BEBLID[8] 这 6 种基于传统的图像拼接算法与本文算法进行比较, 不同算法的图像拼接结果如图 10 所示。将严重失真和拼接图像存在严重错位的情况定义为“拼接失败”。选取了 5 组具有挑战性的图像。从图 10 可以看到, 传统的图像拼接算法容易受到特征点数量和分布的影响, 在复杂的航拍图像中表现不佳。如图 10 中第 1 行所示, Robust ELA 算法拼接失败, 而其他算法拼接结果产生了较大的模糊和伪影, 并且图像不能很好地对齐。观察发现, 传统的图像拼接算法在处理纹理复杂区域时容易产生伪影。本文在图 10 中展示了更多的拼接结果, 其中, 输入和输出一起显示, 图像中各矩形框指示严重模糊及错位严重的区域。

另外, 为了更加直观地评估本文算法的性能, 选取了 5 组极具挑战性的航拍图像进行拼接。这 2 幅图像均包含尺度变化、低重叠率及旋转角度等复杂信息, 与传统的图像拼接算法进行了对比。拼接结果如图 11 所示, 从图 11 的第 1 行和第 2 行可

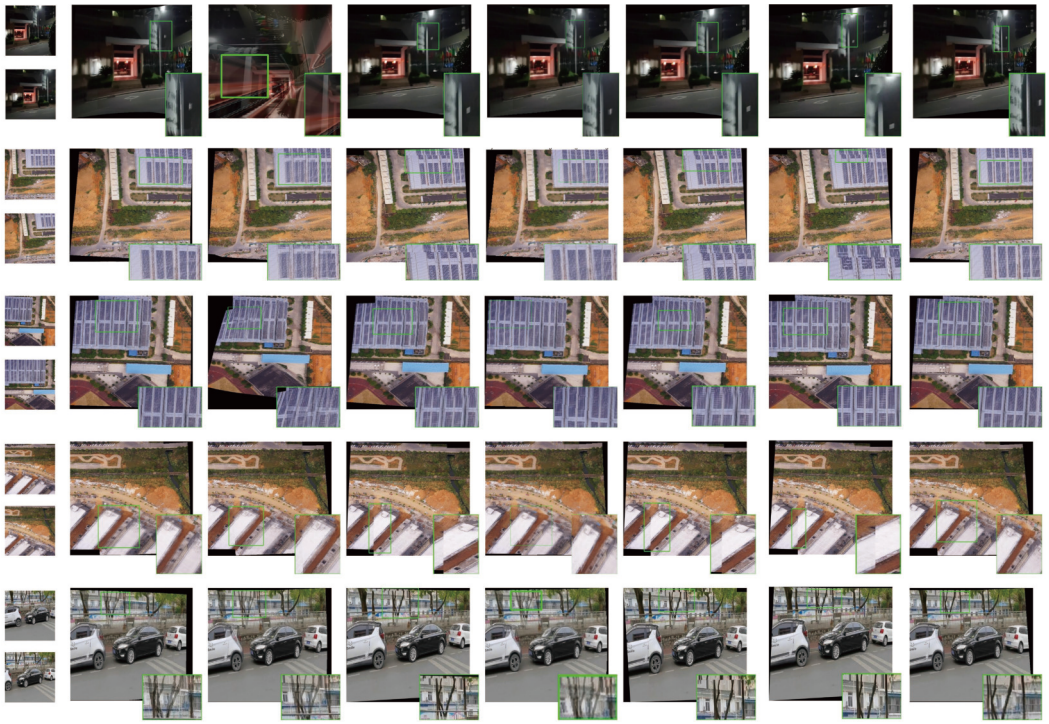


图 10 传统图像拼接算法的航拍图像拼接结果

Fig. 10 Aerial image stitching results of traditional image stitching methods

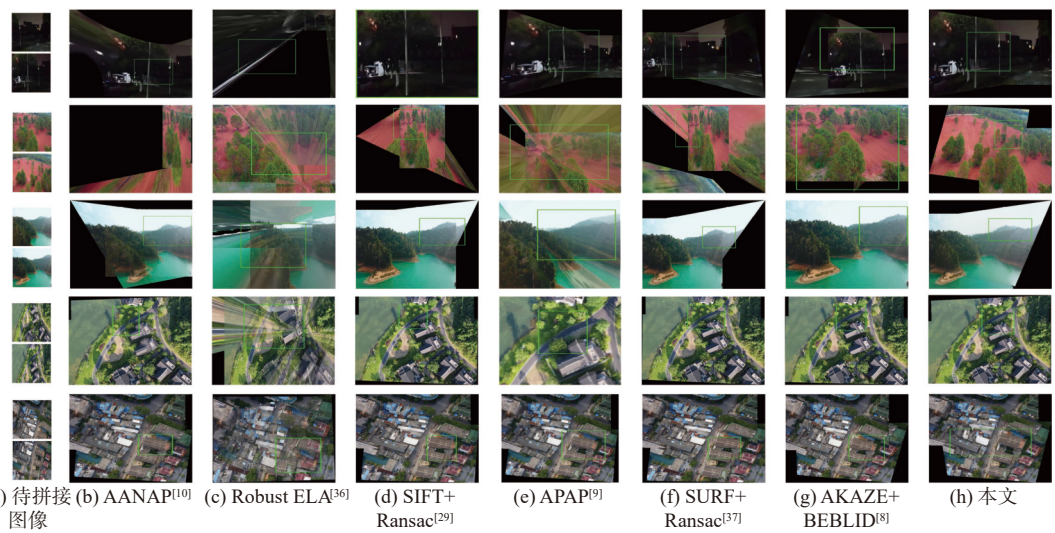


图 11 传统图像拼接算法在复杂场景下的拼接结果

Fig. 11 Stitching results of traditional image stitching methods in complex scenarios

以看到，AANAP、Robust ELA、SIFT+Ransac、APAP、AKAZE+BEBLID 及 SURF+Ransac 算法在航拍图像拼接任务中均拼接失败，这是由于这些算法对于图像复杂性和特征点稀疏性较为敏感，难以处理航拍图像中的细节特征。从第 3 行、第 4 行和第 5 行的拼接结果可以看到，对于 SIFT+Ransac、Robust ELA、AKAZE+BEBLID 算法，住宅建筑部分的缝合边界存在显著的错位，APAP 算法和 AANAP 算法采用全局优化策略，在一般的图像拼接任务中能达到较高的配准精度，然而，对于无人机航拍影像，仅采用全局优化策略无法满足细节特征的精确匹配要求，因此拼接结果图中存在明显的伪影。相比之下，本文提出的无监督深度学习的航拍图像拼接算法成功克服了这一问题，能够精确地对齐图像并消除伪影。即使在传统算法无法检测到足够多的特征点的情况下，本文算法仍具有相当可靠的鲁棒性。

3.5.2 与基于深度学习的图像拼接算法的比较

目前，基于深度学习的图像拼接算法仍在不断

发展中。本文选择了基于变分特征的图像拼接网络(variational feature-based image stitching network, VFISNet)[16]、文献 [19] 算法、DPH-Net[20] 和 UDIS[18] 作为有监督和无监督深度图像拼接的代表, 与本文算法进行比较。其中, VFISNet 是能够在完整的深度学习框架下拼接来自任意视角的图像的算法, 但其存在一个明显的缺点, 即只能拼接大小为 128 像素×128 像素的图像。为了解决该问题, 本文对其采用了双立方插值方法, 以产生任意大小的拼接结果。

图 12 中, 第 1 列展示了待拼接的 2 幅航拍图像, 第 2 列展示了采用 VFISNet 算法进行拼接的结果。可以看出, 该算法拼接结果存在较大程度的失真, 甚至出现了拼接失败的情况。这是因为 VFISNet 算法是一种有监督学习算法, 其在合成数据集上训练, 因此, VFISNet 算法在真实的航拍数据集上的鲁棒性较差。图 12 第 3 列展示了采用 UDIS 算法对航拍图像进行拼接的结果, 观察发现, 该算法在纹理复杂区域的图像对齐方面表现不佳, 出现了部分错位的情况。图 12 第 4 列和第 5 列展示了采用文献 [19] 算法和 DPH-Net 算法对航拍图像进行拼接的结果, 相比之下, 可以看到 DPH-Net 算法的建筑和树木部分的缝合边界存在明显的错位和伪影, 文献 [19] 算法则存在较为明显的伪影。

相比之下, 本文算法在图像拼接方面呈现出更好的表现, 能够实现完全对齐且无伪影的拼接效果, 如图 12 第 6 列所示。此外, 使用基于深度学习的 VFISNet 算法、UDIS 算法、文献 [19] 算法及 DPH-Net 算法对 3.5.1 节提到的 5 组具有挑战性的航拍图像进行拼接, 结果如图 13 所示。实验结果显示, VFISNet 算法在处理这些复杂图像时出现了拼接失败的问题, 与传统的图像拼接算法相比, UDIS 算法、文献 [19] 算法及 DPH-Net 算法虽然能够完成航拍图像拼接任务, 但也出现了一些错位及伪影的情况。相比之下, 本文算法能够精确地对齐图像并消除伪影, 在深度图像拼接方面具有优越性, 并且在航拍场景下能够更好地适应实际需求。

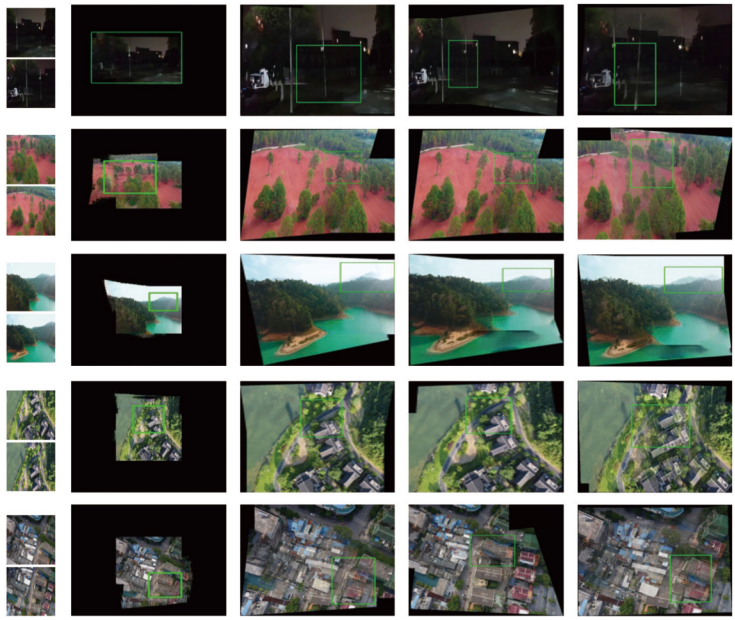


图 12　基于深度学习算法的航拍图像拼接结果

Fig. 12　Aerial image stitching results of deep learning-based methods

3.6　运行效率的比较

为验证本文算法在运行效率方面的有效性, 对 SIFT+Ransac[29]、AKAZE+BEBLID[8]、ORB+Ransac[30]、DPH-Net[20]、DHN[31]、UDHN[32]、UDIS[18] 这 7 种算法的图像拼接时间进行了比较, 并将结果列于表 3 中, 其中, 待拼接的图像对的分辨率为 4 000×6 000。从表 3 可以看出, SIFT+Ransac 算法需要 14.05 s 来完成图像的拼接, AKAZE+BEBLID 算法需要 7.66 s, ORB+Ransac 算法需要 6.01 s, 而本文算法仅需 3.63 s。传统的图像拼接算法较耗时, 因为它们包括特征提取、匹配、单应性估计、图像变形和融合等多个步骤。与传统算法相比, 本文算法通过无监督的端到



(a) 待拼接图像 (b) VFISNet[16] (c) UDIS[18] (d) 文献[19]算法 (e) DPH-Net[20]

图 13 基于深度学习算法在复杂场景下的拼接结果

Fig. 13 Stitching results of deep-learning-based methods in complex scenarios

表 3 不同算法的图像拼接时间比较

Table 3 Comparison of image stitching time of different methods s

算法		拼接时间
传统的图像拼接方法	SIFT+Ransac[29]	14.05
	AKAZE+BEBLID[8]	7.66
	ORB+Ransac[30]	6.01
基于深度学习的图像拼接方法	DPH-Net[20]	4.62
	DHN[31]	4.24
	UDHN[32]	3.69
	UDIS[18]	2.39
	本文	3.63

端深度学习算法来估计单应性，因此具有更快的推理时间。基于深度学习的图像拼接算法中，DPH-Net需要4.62 s来完成图像的拼接，DHN算法需要4.24 s，UDHN算法需要3.69 s，然而，这些算法仅包含图像拼接的第1阶段，即单应性估计，不能称为完整的深度图像拼接模型。为了更全面地对比图像拼接所需时间，在使用这些算法估计单应性的基础上，使用本文提出的图像融合网络以得到拼接结果。尽管本文算法在运行效率上稍差于无监督算法UDIS，但本文算法在拼接的准确性和鲁棒性上具有优势。从图13可以看出，UDIS算法在纹理复杂区域的图像对齐方面表现不佳，出现了部分错位的情况，相比之下，本文算法在图像拼接方面表现更佳。实验结果证明了本文算法在保持运行效率的同时，确保了拼接的准确性和鲁棒性。

3.7 消融实验

为了验证本文提出的基于无监督深度学习的航拍图像拼接框架中各个部分的必要性，进行了广泛的消融实验。具体来说，本文验证了双分支单应性估计网络在无监督航拍图像拼接框架中的作用，该网络包括低分辨率分支和高分辨率细化分支。同时，对于无监督图像融合网络，本文验证了所提出的编码模块的密集连接和编解码网络之间的残差路径对网络性能的影响。所有研究的结构列于表4中。表中，单分支、双分支和编解码网络分别对应于图2中的低分辨率分支、低分辨率分支+高分辨率细化分支和图5中的编码器和解码器部分。图14展示了不同结构的消融实验结果。可以得出以下结论：

1）当单应性估计网络中的低分辨率分支和编解码网络直接组合时，可以完成图像拼接任务，如表4中的V1框架。从图14的第1列可以看到，该结构无法完成准确的图像对齐，图像中的柱子拼接边界存在明显的错位，且图像的灯光区域产生了严重的伪影。

表 4 消融研究框架

Table 4 Ablation research framework

框架	单分支	双分支	编解码网络	编解码网络+密集连接	编解码网络+残差路径
V1	√		√		
V2		√	√		
V3		√		√	
V4		√			√
V5		√		√	√



(a1) V1　(a2) V2　(a3) V3　(a4) V4　(a5) V5

(a) 不同结构的拼接结果图



(b1) V1　(b2) V2　(b3) V3　(b4) V4　(b5) V5

(b) 局部放大图

图 14　消融实验结果

Fig. 14　Ablation experiment results

2) 与 V1 相比, 双分支单应性估计网络可以更准确地生成变换矩阵, 这直接影响到后续图像融合的效果。如图 14 第 2 列所示, 可以看到, 图像中的柱子拼接边界只有一点错位。然而, 由于高分辨率细化分支卷积核的感受野对于高分辨率的航拍图像来说过小, 因此, 会在图像中引入一些伪影。

3) 与 V2 相比, V3 在编解码网络中加入了密集连接, 从而形成了一种稠密的网络结构, 这种连接方式能够重用浅层特征, 并加强特征的深度传播, 进一步提高模型对图像语义信息的理解能力。这种设计能够使模型更加准确地对齐图像。如图 14 第 3 列所示, V3 在图像中灯光区域的对齐方面表现出更好的效果。

4) 与 V2 相比, V4 在编码器和解码器之间添加了 4 个残差路径, 用于拼接同维度的特征, 这种残差路径的设计可以有效地增加网络深度, 提高特征的抽象层次, 从而更好地捕捉图像的语义信息。如图 14 中第 4 列所示, V4 结构能够进一步减少图像中的伪影现象, 从而提高图像拼接的质量。

5) 与 V3 和 V4 相比, V5 在编解码网络中同时引入了密集连接和残差路径, V5 结构能够最大程度地提高网络的表达能力, 从而获得最佳结果, 如图 14 第 5 列所示, 图像中的柱子拼接边界能够完全对齐, 并且消除了图像中灯光区域的伪影。

因此, 只有包括双分支单应性估计网络和无监督图像融合网络在内的完整无监督图像拼接结构, 才能实现对自然航拍图像的高质量无缝拼接。

本文通过大量的实验, 充分验证了无监督深度单应性估计网络和图像融合网络的有效性, 并揭示了本文算法相对于现有技术的优越性。此外, 消融实验结果进一步证明了本文提出的基于无监督深度学习的航拍图像拼接框架中各个部分的有效性和必要性。

3.8　视觉质量的用户研究

为了与其他算法进行定量比较, 采用用户研究来评估视觉质量。在研究中, 邀请了 20 名参与者, 其中, 包括 10 名具有计算机视觉背景的专业人员及 10 名无相关背景的学生。这些研究在测试集中进行, 每个参与者需要在随机选取的 50 张图像中, 逐一比较每种算法与本文算法的拼接结果, 并且用户可以自由放大图像以获取更清晰的细节。每位参与者需要回答哪种算法的拼接结果最优, 数字以百分比的形式展示, 统计结果如图 15 所示。结果显示, 对本文算法表示更偏好的人数超过对其他方法表示更偏好的人数, 这意味着, 在用户的评价中, 本文算法的航拍图像拼接结果具有更高的视觉质量。

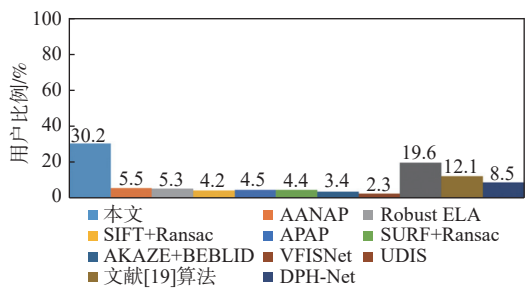


图 15　视觉质量的用户研究示意图

Fig. 15　User study on visual quality

4　结　论

本文将卷积神经网络的算法和理念引入到图

像拼接领域，利用深度学习的算法学习航拍图像的特征，通过多次实验及分析，得出以下结论：

1）提出了一个完整的无监督航拍图像拼接框架，不需要任何的标签数据。该框架包括无监督深度单应性估计网络和无监督图像融合网络2个阶段，实现了航拍图像端到端的拼接。

2）建立了一个真实的航拍图像数据集，为无监督图像拼接提供了可靠的数据基础。

3）大量的实验结果证明，相比于其他的图像拼接算法，本文算法定量评价、视觉拼接效果和鲁棒性方面都具有更好的表现。

4）未来的研究工作将深入研究三维航拍图像的拼接任务。与二维拼接任务相比，三维航拍图像拼接的计算成本更高，但同时也提供了更高质量的地理信息。

参考文献（References）

[1] 蒲良, 张学军. 基于深度学习的无人机视觉目标检测与跟踪[J]. 北京航空航天大学学报, 2022, 48(5): 872-880.
PU L, ZHANG X J. Deep learning based UAV vision object detection and tracking[J]. Journal of Beijing University of Aeronautics and Astronautics, 2022, 48(5): 872-880(in Chinese).

[2] YANG C, LIU X, ZHOU H, et al. Towards accurate image stitching for drone-based wind turbine blade inspection[J]. Renewable Energy, 2023, 203: 267-279.

[3] XIE W H. Research on target extraction system of UAV remote sensing image based on artificial intelligence[C]//Proceedings of the IEEE International Conference on Integrated Circuits and Communication Systems. Piscataway: IEEE Press, 2023: 1-5.

[4] CHEN J, LI Z X, PENG C L, et al. UAV image stitching based on optimal seam and half-projective warp[J]. Remote Sensing, 2022, 14(5): 1068.

[5] JONG T K, BONG D B L. An effective feature detection approach for image stitching of near-uniform scenes[J]. Signal Processing: Image Communication, 2023, 110: 116872.

[6] ZHANG J D, XIU Y. Image stitching based on human visual system and SIFT algorithm[J]. The Visual Computer, 2024, 40(1): 427-439.

[7] 宋飞, 杨扬, 杨昆, 等. 基于双特征的丘陵山区耕地低空遥感图像配准算法[J]. 北京航空航天大学学报, 2018, 44(9): 1952-1963.
SONG F, YANG Y, YANG K, et al. Low-altitude remote sensing image registration algorithm based on dual-feature for arable land in hills and mountains[J]. Journal of Beijing University of Aeronautics and Astronautics, 2018, 44(9): 1952-1963(in Chinese).

[8] 宗慧琳, 袁希平, 甘淑, 等. 改进AKAZE算法的泥石流区无人机影像特征匹配[J]. 测绘通报, 2023(2): 91-96.
ZONG H L, YUAN X P, GAN S, et al. An improved AKAZE algorithm for UAV image feature matching in debris flow area[J]. Bulletin of Surveying and Mapping, 2023(2): 91-96(in Chinese).

[9] ZARAGOZA J, CHIN T J, BROWN M S, et al. As-projective-as-possible image stitching with moving DLT[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2013: 2339-2346.

[10] LIN C C, PANKANTI S U, RAMAMURTHY K N, et al. Adaptive as-natural-as-possible image stitching[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2015: 1155-1163.

[11] CHANG C H, SATO Y, CHUANG Y Y. Shape-preserving half-projective warps for image stitching[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2014: 3254-3261.

[12] CHEN Y S, CHUANG Y Y. Natural image stitching with the global similarity prior[C]//European Conference on Computer Vision. Berlin: Springer, 2016: 186-201.

[13] 梁镇锋, 夏海英. 一种面向无人机航拍图像的快速拼接算法[J]. 广西师范大学学报 (自然科学版), 2023, 41(3): 41-52.
LIANG Z F, XIA H Y. A fast stitching algorithm for UAV aerial images[J]. Journal of Guangxi Normal University (Natural Science Edition), 2023, 41(3): 41-52(in Chinese).

[14] HOANG V D, TRAN D P, NHU N G, et al. Deep feature extraction for panoramic image stitching[C]//Asian Conference on Intelligent Information and Database Systems. Berlin: Springer, 2020: 141-151.

[15] YAN M, YIN Q, GUO P. Image stitching with single-hidden layer feedforward neural networks[C]//Proceedings of the International Joint Conference on Neural Networks. Piscataway: IEEE Press, 2016: 4162-4169.

[16] NIE L, LIN C Y, LIAO K, et al. A view-free image stitching network based on global homography[J]. Journal of Visual Communication and Image Representation, 2020, 73: 102950.

[17] NIE L, LIN C Y, LIAO K, et al. Learning edge-preserved image stitching from large-baseline deep homography[EB/OL]. (2020-12-11)[2023-06-01]. http://arxiv.org/abs/2012.06194.

[18] NIE L, LIN C Y, LIAO K, et al. Unsupervised deep image stitching: reconstructing stitched features to images[J]. IEEE Transactions on Image Processing, 2021, 30: 6184-6197.

[19] ZHU F Z, LI J C, ZHU B, et al. UAV remote sensing image stitching via improved VGG16 Siamese feature extraction network[J]. Expert Systems with Applications, 2023, 229: 120525.

[20] HUANG C W, PAN X, CHENG J C, et al. Deep image registration with depth-aware homography estimation[J]. IEEE Signal Processing Letters, 2023, 30: 6-10.

[21] 马腾宇, 李孜, 刘日升, 等. 基于无监督学习的多模态可变形配准[J]. 北京航空航天大学学报, 2021, 47(3): 658-664.
MA T Y, LI Z, LIU R S, et al. Multimodal deformable registration based on unsupervised learning[J]. Journal of Beijing University of Aeronautics and Astronautics, 2021, 47(3): 658-664(in Chinese).

[22] IANDOLA F N, HAN S, MOSKEWICZ M W, et al. SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <0.5 MB model size[EB/OL]. (2016-11-04)[2023-06-01]. http://arxiv.org/abs/1602.07360.

[23] HARTLEY R, ZISSERMAN A. Multiple view geometry in computer vision[M]. 2nd ed. Cambridge: Cambridge University Press, 2004.

[24] JADERBERG M, SIMONYAN K, ZISSERMAN A. Spatial Transformer networks[C]//Proceedings of the 21th International Conference on Neural Information Processing Systems. Cambridge: MIT Press, 2015: 2017-2025.

[25] HE K M, ZHANG X Y, REN S Q, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2016: 770-778.

[26] ZHAO H, GALLO O, FROSIO I, et al. Loss functions for image restoration with neural networks[J]. IEEE Transactions on Computational Imaging, 2017, 3(1): 47-57.

[27] JOHNSON J, ALAHI A, LI F F. Perceptual losses for real-time style transfer and super-resolution[C]//European Conference on Computer Vision. Berlin: Springer, 2016: 694-711.

[28] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition[EB/OL]. (2015-04-10)[2023-06-01]. http://arxiv.org/abs/1409.1556.

[29] BROWN M, LOWE D G. Automatic panoramic image stitching using invariant features[J]. International Journal of Computer Vision, 2007, 74(1): 59-73.

[30] RUBLEE E, RABAUD V, KONOLIGE K, et al. ORB: an efficient alternative to SIFT or SURF[C]//Proceedings of the International Conference on Computer Vision. Piscataway: IEEE Press, 2011: 2564-2571.

[31] DETONE D, MALISIEWICZ T, RABINOVICH A. Deep image homography estimation[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2016: 5668-5676.

[32] NGUYEN T, CHEN S W, SHIVAKUMAR S S, et al. Unsupervised deep homography: a fast and robust homography estimation model[J]. IEEE Robotics and Automation Letters, 2018, 3(3): 2346-2353.

[33] WINKLER S, MOHANDAS P. The evolution of video quality measurement: from PSNR to hybrid metrics[J]. IEEE Transactions on Broadcasting, 2008, 54(3): 660-668.

[34] WANG Z, BOVIK A C, SHEIKH H R, et al. Image quality assessment: from error visibility to structural similarity[J]. IEEE Transactions on Image Processing, 2004, 13(4): 600-612.

[35] HODSON T O. Root-mean-square error (RMSE) or mean absolute error (MAE): when to use them or not[J]. Geoscientific Model Development, 2022, 15(14): 5481-5487.

[36] LI J, WANG Z M, LAI S M, et al. Parallax-tolerant image stitching based on robust elastic warping[J]. IEEE Transactions on Multimedia, 2018, 20(7): 1672-1687.

[37] BAY H, TUYTELAARS T, VAN GOOL L. SURF: speeded up robust features[C]//European Conference on Computer Vision. Berlin: Springer, 2006: 404-417.

Aerial image stitching algorithm based on unsupervised deep learning

LIANG Zhenfeng[1, 2], XIA Haiying[1, 2, 3, *], TAN Yumei[1, 2], SONG Shuxiang[1, 2]

(1. Guangxi Key Laboratory of Brain-inspired Computing and Intelligent Chips, Guangxi Normal University, Guilin 541004, China;
2. Key Laboratory of Integrated Circuits and Microsystems, Guangxi Normal University, Guilin 541004, China;
3. Guangxi Higher Education Engineering Research Center of Optoelectronic Information Technology, Guangxi Normal University, Guilin 541004, China)

Abstract: Traditional image stitching approaches predominantly depend on accurate feature localization and distribution, which leads to suboptimal robustness in intricate aerial photography contexts. Consequently, a comprehensive unsupervised deep learning framework for aerial image stitching was devised, encompassing an unsupervised deep homography estimation network and an unsupervised image fusion network. First, the deep homography estimation network was employed to provide precise alignment data for subsequent stitching by ascertaining the homographic transformation between reference and target images. Subsequently, the image fusion network was utilized to learn deformation patterns of aerial image stitching, generating the final stitched output. Additionally, a real dataset for unsupervised aerial image stitching was introduced to facilitate the training of the learning framework. Comparative analysis was conducted on the suggested unmanned aerial vehicle aerial image dataset, incorporating scale-invariant feature transform (SIFT) + Ransac, accelerated-nonlinear diffusion-based feature detection and matching (AKAZE) + boosted efficient binary local image descriptor (BEBLID), oriented brief (ORB) + Ransac, and deep-learning-based image stitching algorithms. Experiments show that the value of structural similarity (SSIM) is increased by 39.94%; the peak signal-to-noise ratio (PSNR) is increased by 36.55%, and the root mean square error (RMSE) is reduced by 66.09%. Moreover, the proposed method demonstrates superior visual stitching performance and robustness in authentic aerial scenarios compared to existing deep-learning-based and traditional image stitching methods.

Keywords: aerial image; deep homography estimation; image stitching; unsupervised; image fusion

Received: 2023-06-15; **Accepted**: 2024-03-29; **Published Online**: 2024-04-22 21:15

URL: link.cnki.net/urlid/11.2625.V.20240419.1511.004

Foundation items: Guangxi Leaderboard Technology Project (Guike JB23023006); Guangxi Key Research and Development Project (Guike AB23026103); National Natural Science Foundation of China (62106054); Major Special Projects of Guangxi Science and Technology (Guike AA20302003)

* **Corresponding author**. E-mail: xhy22@mailbox.gxnu.edu.cn

http://bhxb.buaa.edu.cn jbuaa@buaa.edu.cn
DOI: 10.13700/j.bh.1001-5965.2023.0367

基于语义规则的自适应情感词典自动构建算法

卫青蓝[1]，何雨[2]，宋金宝[1,*]
(1. 中国传媒大学 数据科学与智能媒体学院，北京 100024；2. 中国传媒大学 信息与通信工程学院，北京 100024)

摘 要：使用词典进行文本情感分析的方法虽然快捷无监督，但其准确性受到词典质量的约束。现有中文通用词典往往都由手动构建，无法自动发现新词且存在情感歧义词，因此在跨域应用时，现有词典质量有待提高。针对上述问题，提出了一种基于语义规则的领域自适应中文情感词典自动构建算法。构建了中文情感固定词典，有效消除了情感歧义性；提出了新的领域自适应的中文新词发现方法，实现了对通用领域中文词典的自动扩充；提出了融合词性筛选和语义规则嵌入的情感词汇倾向无监督计算方案，有效提高了精度。实验证明：在常用计算机语料库上，采用情感固定词典的情感分析方法比使用其他中文通用词典准确率平均提高9.31%，精确率平均提高12.77%，精确率和召回率的调和平均数 F_1 值平均提高7.43%。在酒店、中文情感分析语料库2个数据集上，提出的情感词典自动构建算法较先进算法准确率平均提高了7.41%,召回率平均提高了12.23%，F_1 值平均提高了9.08%。

关 键 词：情感词典；情感分析；语义规则；无监督；领域自适应
中图分类号：V221+.3; TB553
文献标志码：A **文章编号**：1001-5965(2025)07-2450-10

互联网的普及带来了海量网络文本，其中蕴含着丰富的信息，网络文本的情感倾向分析是自然语言处理的一个热点研究。中文相较英文具有基数大且语义信息丰富的特点，尤其具有挑战性。基于情感词典的方法无需人工标记数据，但其最大难点是准确率非常依赖于情感词典的质量。

中文词典质量决定于词典自动构建的3个步骤：新词发现、情感词汇识别和情感倾向判断。新词发现是中文词典构建的重要步骤之一，其通过对语料库进行分析，从中挖掘出未录入的新词。常用方法包括基于互信息或其改进形式[1-3]，但这些方法往往只适用于单个领域的语料库，跨域应用时新词发现效果不佳。情感词汇识别的目的是找出句子中表达积极或消极情感的词汇，从而判断句子的整体情感倾向。目前常用的方法包括基于词间相似性[4-6]、基于词间依赖关系[7-8]和基于词性标记[9-10]的情感词汇识别方法。这些方法存在着明显的局限性，如忽略了词句间的语义规则，或只使用粗糙的词性信息，影响了情感词识别的准确性。情感倾向判断的目的是确定情感词汇的情感倾向，常见方法包括基于局部词汇和基于全局语篇的方法。前者根据已知的情感词汇来判断未知词汇的情感倾向，但这种方法要求词汇具有固定的倾向，而实际上很多词汇的倾向会随着语境变化。后者根据已标注的语篇来判断其中词汇的情感倾向，这种方法需要大量人工标注，因此非常耗时。

针对上述问题，本文提出了一个基于语义规则的领域自适应中文情感词典自动构建算法。主要创新点包括：①构建了中文情感固定词典，在常见的计算机语料库上比使用其他中文通用词典准确

收稿日期：2023-06-15；**录用日期**：2023-12-01；**网络出版时间**：2024-03-07 18:12
网络出版地址：link.cnki.net/urlid/11.2625.V.20240305.1537.001
基金项目：国家自然科学基金 (72474198,62301510)；中国传媒大学亚洲传媒研究中心科研资助 (AMRC2023-4)
*通信作者. E-mail：songjinbao@cuc.edu.cn

引用格式：卫青蓝，何雨，宋金宝. 基于语义规则的自适应情感词典自动构建算法 [J]. 北京航空航天大学学报，2025，51（7）：2450-2459. WEI Q L，HE Y，SONG J B. An adaptive automatic construction algorithm for sentiment dictionaries based on semantic rules [J]. Journal of Beijing University of Aeronautics and Astronautics，2025，51（7）：2450-2459（in Chinese）.

率平均提高 9.31%, 精确率平均提高 12.77%, 精确率和召回率的调和平均数 F_1 值平均提高 7.43%, 有效消除了跨域情感歧义性; ②提出了新的领域自适应的中文新词发现方法, 实现了通用领域中文词典的自动扩充; ③创新性地提出语义规则嵌入的情感词汇倾向无监督计算方法, 在 2 个数据集上较最先进算法准确率平均提高了 7.41%, 召回率平均提高了 12.23%, F_1 值平均提高了 9.08%。

1　相关工作

1.1　种子词典

Kanayama 等[11] 提出了种子词汇的概念, 指具有明显情感倾向的极性原子, 可以用来推测未知词汇的情感倾向。常用的中文情感种子词典最早是台湾大学简体中文情感词典(national Taiwan university sentiment dictionary, NTUSD)[12], 包含积极和消极 2 类情感词汇。后续还出现了大连理工大学情感词汇本体库(Dalian university of technology sentiment ontology, DU)[13]、知网情感分析用词语集(HowNet sentiment dictionary, HowNet)[14]、清华大学李军中文褒贬义词典(TsingHua university Li Jun Chinese sentiment dictionary, THU)[15] 等。这些情感词典虽然覆盖了大部分情感词汇并标注了情感倾向, 但存在以下问题: ①忽略了领域歧义词汇的影响; ②词典之间的情感倾向标注存在冲突; ③都是通过人工构建, 不能及时更新之后出现的词汇。这些问题导致了基于词典的情感分析准确率较低。为了解决上述问题, Ma[8] 和 Liu[16] 等提出了一种基于多个词典的种子词典构建算法, 筛选出情感倾向一致的词汇作为种子词典, 但没有考虑歧义词汇的处理, 且词典的覆盖范围不够全面。

1.2　基于语义规则的情感倾向判断

领域情感词典的构造主要关注领域新词的识别、新词情感倾向判断和歧义词汇的情感倾向矫正。现有工作大多只关注其中一点而忽略了另一方面, 如 Yu 等[2] 提出基于左右熵和点互信息(semantic orientation from pointwise mutual information with left-right entropy, SO-PMI)的新词发现方法, 对数据集的新词发现结果进行情感倾向分析, 但该方法只关注了语料库的新词情感倾向, 忽略了歧义词汇的情感倾向矫正问题; Ma 等[8] 提出了一个面向中文的情感极性检测框架, 解决了歧义词汇情感极性判断问题, 但没有讨论新词的发现及识别问题。

在进行词汇情感倾向判断之前, 情感词汇的识别与提取是一个非常重要的步骤。目前, 情感词汇识别的方法主要有依赖关系法[2,6,10-11]、词性筛选法[9-10]、词汇上下文相似性法[4] 和互信息或其改进形式结合左右熵法[1-3,16]。其中, 互信息法被广泛地运用在中文词汇的发现和情感识别中。然而, 使用点互信息(pointwise mutual information, PMI)公式提取情感词汇时, 自动调整阈值是个难点, 这是因为阈值与语料库大小相关。而现有文献中阈值的选择或是缺乏可解释性说明, 或是采用穷举法导致效率低下, 同时不具有普适性。

语义规则法是根据词汇的语义关系和情感倾向来判断情感极性的方法, 常被用来判断情感词汇的倾向。语义规则包括词汇层面[2,5,7-8,10-11,16] 和句子层面的语义规则[4,9], 其中, 词汇层面的语义规则主要用于无标注的语料库。考虑句子层面的语义规则[3,6] 使用标注的语料库, 但标注的语料库的规模和质量有限, 不能覆盖所有领域, 因此不能快速生成特定领域情感词典。

1.3　情感词典的领域自适应性

情感词典自动构建通常采用知识库和语料库结合的方法, 主要包括以下 3 种方法: ①关系图半监督法[7,17-18], 该方法利用词与词之间的相似关系构建词间关系图, 再根据已知极性的词汇, 运用图算法来推测其他词汇的极性。例如, Tan 等[18] 提出了一种随机游走算法, 利用旧领域和目标领域的情感词和语篇构建面向领域的情感词典, 但该方法的前提是 2 个领域必须包含极性相同的情感词汇。②自举半监督法, 为了解决标注语料不足的问题, 该方法先利用少量标注词确定文本片段的极性, 再结合抽取结果, 判断未知情感的文本片段。例如, Almatarneh 等[9] 提出的从电影评论语料库中自动构建域特定情感词典的算法, 考虑了情感词汇与句子总体情感之间的联系, 根据句子的情感标签推断词汇的情感极性和权重。③深度表示法[5,19], 该方法根据上下文训练词向量, 使得语义相近的词在向量空间上距离较近, 以此来判断词的情感极性。然而这些方法只针对特定领域, 没有考虑领域适应性。

2　本文方法

本文提出了一个基于语义规则的领域自适应中文情感词典自动构建算法框架, 如图 1 所示。算法框架主要包括 3 部分: 中文情感固定词典构建、新词发现及情感词汇倾向判断。首先, 基于 4 个中文通用情感词典, 筛选出常用且无歧义的情感词汇, 来构建一个中文情感固定词典。其次, 在传统 PMI 公式基础上, 提出了一个新的阈值公式, 用于在特定领域的文本中自适应地发现新词。最后, 融

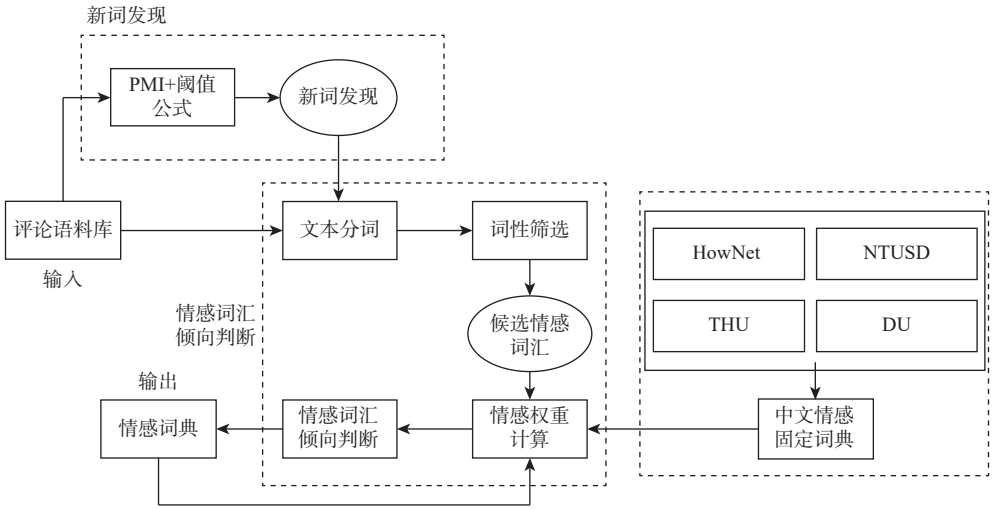


图1 情感词典自动构建算法框架

Fig. 1 Framework of automatic construction algorithm of sentiment dictionary

合词性筛选和情感权重计算来判断词汇的情感倾向。其中,词性筛选的作用是过滤大部分非情感词汇,情感权重计算则是根据本文提出的语义规则,综合考虑词汇本身和上下文的情感影响,计算词汇最终的情感权重。

2.1 中文情感固定词典构建

中文情感固定词典构建的一个重要步骤是判断词汇的情感倾向,即从已知词汇的情感倾向来推断未知词汇的情感倾向。为了提高推断的准确性,已知词汇的情感倾向应该是不随语境变化而改变的,这类词汇称为情感固定词汇,它们的集合构成情感固定词典。本文以现有的中文通用情感词典为基础,构造情感固定词典。现有的中文情感词典包括NTUSD[12]、DU[13]、HowNet[14]和THU[15]。将每个词典分为积极词典P_i和消极词典N_i;对所有积极(消极)词典取并集,得到积极(消极)词汇集合;取积极词汇集合和消极词汇集合的交集,得到歧义词汇集合;将积极(消极)词汇集合减去歧义词汇集合,得到初步的情感固定词典(参见式(1)和式(2))。对初步词典进行人工筛选,剔除人们在日常中不常使用的词汇,如“同文共规”、“摊书拥百城”等类似的成语和俗语,以及极性不明确的词汇,如“烟花”、“回顾”等类似不含有情感的词汇,最终得到积极情感固定词典p和消极固定词典n:

$$p=\bigcup_{i=1}^{4}P_i-\left(\bigcup_{i=1}^{4}P_i\cap\bigcup_{i=1}^{4}N_i\right) \tag{1}$$

$$n=\bigcup_{i=1}^{4}N_i-\left(\bigcup_{i=1}^{4}P_i\cap\bigcup_{i=1}^{4}N_i\right) \tag{2}$$

2.2 新词发现

现有新词发现方法普遍使用传统的PMI公式(参见式(3))计算词汇之间的相关性,通过固定阈值筛选得到候选词汇。

$$\mathrm{PMI}(x,y)=\log_2\frac{p(x,y)}{p(x)p(y)} \tag{3}$$

式中:$p(x,y)$为2个单词同时出现的频率;$p(x)$和$p(y)$分别为单词x和单词y出现的频率。然而,传统的PMI公式存在一些问题,如阈值的选择与语料库的大小相关,不同领域的语料库需要不同的阈值,而穷举法寻找最佳阈值的效率很低。此外,传统的PMI公式没有考虑新词的出现频率和比例,可能导致偶然组合的词汇被误认为新词。为了解决这些问题,本文提出了一种语料库自适应的阈值公式,如下:

$$t=\begin{cases}\min\left(\log_2\dfrac{1}{p(w_i)},\log_2\dfrac{1}{p(w_{i+1})}\right)-1 & g=1\\ \min\left(\log_2\dfrac{1}{p(w_i)},\log_2\dfrac{1}{p(w_{i+1})}\right)+\log_2\dfrac{g-1}{g} & \text{其他}\end{cases} \tag{4}$$

式中:t为不同语料库对应的阈值。考虑到新词是由旧词组合而来,新词的出现频率一定小于旧词,例如,给定一个字“树”,会想到“树枝”、“树叶”等,显然“树叶”的出现频率会低于“树”,与此同时,新词的出现次数较旧词来说还应当在语料库占据一定的比例,以此来排除偶然组合,如“颗树”、“树里”会出现在句子中但不属于词语的排列组合。因此引入缩放因子g,含义为相邻词汇的出现频率之比:

$$g=\max\left(\frac{N(w_i)}{N(w_{i+1})},\frac{N(w_{i+1})}{N(w_i)}\right)\qquad g\in[1,+\infty) \tag{5}$$

式中:$N(w_i)$为单词w_i在语料库中出现的次数。当g取1时,表示相邻词汇出现频率相当,此时新词出现频率应当至少为旧词的1/2;而当g值越大时,表明相邻词汇的出现频率相差很大,出现频率大的词汇大概率为介词,此时是否组成新词应当由频率较

小的词汇决定, 若该词汇经常与另一词汇共同出现, 即组合词汇的出现频率与较小频率相近, 则判断其应当是新词。由于阈值公式应当与 PMI 公式对齐, 因此在计算时将词汇的出现频率先取倒数再取对数。

本文认为, 新词一般由若干个旧词组合而成, 如“社恐”是“社交”与“恐惧”的组合, 因此本文先对句子进行分词, 再计算相邻词之间的 PMI。新词发现的具体方法流程如图 2 所示。

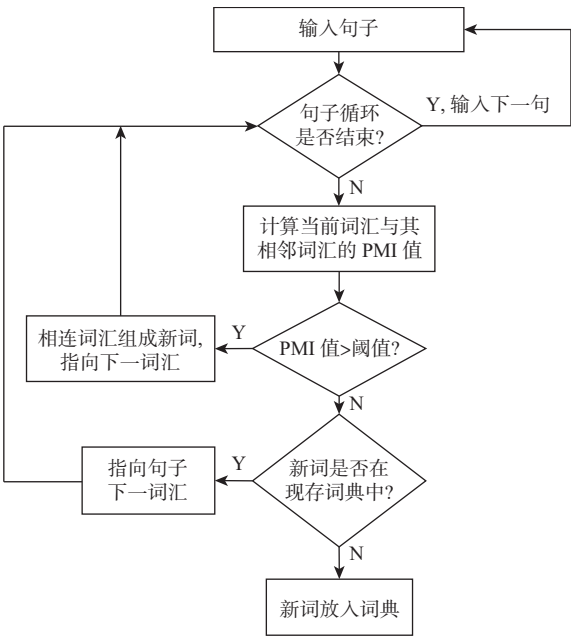


图 2　PMI 公式新词发现方法流程

Fig. 2　Flowchart of new word discovery algorithm based on PMI formula

2.3　词汇的情感倾向判别

情感分析是自然语言处理的一个重要任务, 其目的是识别和提取文本中的情感态度和倾向。为了实现这一目的, 需要从语料库中找出携带情感的词汇, 并判断它们的情感倾向, 主要包括词性筛选和情感权重计算。

2.3.1　词性筛选

本文发现句子情感主要取决于主观词汇, 而客观词汇如“果汁”、“电脑”等对句子情感无影响。若给无情感词汇赋予情感, 会干扰整体语料库的情感判断。如, 若判断“苹果”为积极词汇, 则含“苹果”的句子都可能被误判为积极, 导致整体语料库的情感计算准确率降低。一般而言, 形容词、名词和动词都能表达情感, 其中形容词尤为突出。因此, 本文在情感权重计算之前先进行词性过滤, 既能排除无情感词汇, 又能提高方法效率。

2.3.2　情感权重计算

本文考虑了词汇与句子之间的 4 种语义规则来进行情感权重计算: ①词汇之间的关系, 即大概率出现在某极性词汇旁边的词汇也是对应极性; ②词汇到句子的关系, 即当表示某极性的词汇出现在句子中时, 该句情感倾向大概率为对应极性; ③句子之间的关系, 即一般情况下评论表达的情感不会突变, 因此相邻句子之间的情感倾向相近; ④句子到词汇的关系, 即若句子整体情感倾向确定, 易得到句中词汇实体的情感倾向。本文还考虑了否定词汇和转折词汇对句子情感转变的影响。按照上述 4 种规则和影响因素, 本文情感权重计算主要分为以下 4 个部分:

1) 根据句内位置分析词汇情感倾向。在该模块中, 统计候选情感词汇 w_i 分别与积极词汇的共现次数 $P_1(i)$、与消极词汇的共现次数 $N_1(i)$ 及 $N(w_i)$, 计算其情感权重 $W_1(i)$:

$$W_1(i)=\frac{P_1(i)-N_1(i)}{N(w_i)} \tag{6}$$

当绝对值大于阈值时, 本文将候选词汇添加到对应情感倾向的情感固定词典中。

2) 计算句子情感倾向。采用步骤 1)得到的情感词典来计算句子的情感倾向, 具体方法流程如图 3 所示。图中的句子是以分隔符号结束的分句, 分隔符号包括逗号、分号、问号、感叹号等。当句子中有转折词时, 预处理阶段会将其分成 2 个分句。

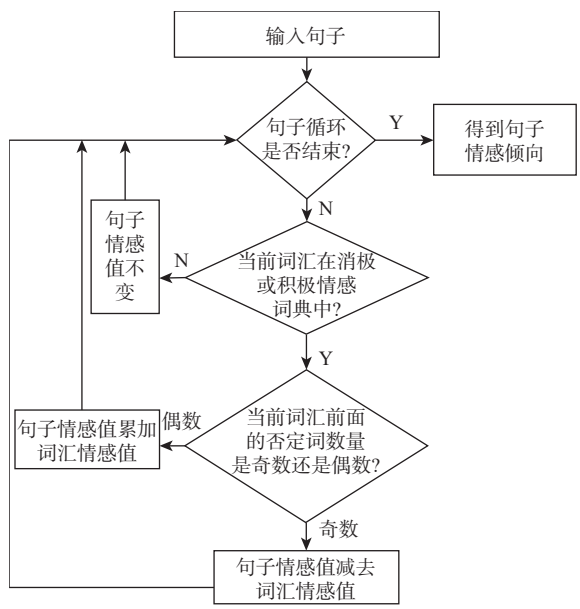


图 3　句子情感计算流程

Fig. 3　Flowchart of sentence sentiment calculation

3) 根据句间关系分析句子情感倾向。得到所有分句的情感值后, 用规则 3)来推断中性句子的情感倾向。例如, 若句子前后情感倾向都为积极且没有转折词, 则该句情感也应为正向。

4) 根据句子情感推断词汇情感倾向。从句子

的整体情感倾向可以判断句中情感词汇的倾向。统计w_i在积极句子中的出现次数$P_2(i)$、在消极句子中出现次数$N_2(i)$及$N(w_i)$来计算情感权重$W_2(i)$:

$$W_2(i)=\frac{P_2(i)-N_2(i)}{N(w_i)} \tag{7}$$

综合$W_1(i)$和$W_2(i)$,即可得到图1中的情感权重。

3 实 验

3.1 实验环境

使用Windows10操作系统,Python版本为Python 3.6及以上,并使用Pycharm进行代码编译。

3.2 数据集

基于情感权重计算提出了情感词汇倾向判断的新方法,考虑到权重数值仅可识别词汇的正负倾向,暂不适用于不同情绪之间的差异,如“喜”、“怒”、“哀”、“恨”,因此在实验中使用仅包含正负2种情感类别的数据集。本文使用了4个中文情感分析语料库,分别为外卖领域的中文短文本情感分析语料,计算机、手机、书籍领域的中文情感分析语料,衣服、水果、平板、酒店、洗发水领域的中文情感分析语料,以及谭松波博士整理的中文酒店评论语料[20]。

3.3 实验细节

实验主要分为新词发现部分和词汇情感判别部分。新词发现部分设置新词最大长度为6。为确保分词更加准确,设定用户词典为本文构造的基础情感固定词典、转折词典、否定词典、停用词词典及副词词典的集合。词汇情感判别部分需要判断情感词汇前面是否存在否定词汇和转折词汇,设置距离为3。实验采用精确率、准确率、召回率和F_1值作为客观评价指标。

精确率P针对预测结果,表示正样本预测正确的精确度,P越大说明误检的越少。

$$P=\frac{P_{\mathrm{T}}}{P_{\mathrm{T}}+P_{\mathrm{F}}} \tag{8}$$

准确率A为预测正确的结果占总样本的比例。

$$A=\frac{P_{\mathrm{T}}+N_{\mathrm{T}}}{P_{\mathrm{T}}+P_{\mathrm{F}}+N_{\mathrm{T}}+N_{\mathrm{F}}} \tag{9}$$

召回率R针对原样本,表示在实际为正的样本中被预测为正样本的覆盖率,R越大说明漏检的越少。

$$R=\frac{P_{\mathrm{T}}}{P_{\mathrm{T}}+N_{\mathrm{F}}} \tag{10}$$

F_1值同时考虑精确率和召回率,使两者同时达到最高,取得平衡。

$$F_1=\frac{2PR}{P+R} \tag{11}$$

式中:P_{T}表示实际为正样本且预测正确的样本数量;P_{F}表示实际为负样本但被预测错误为正样本的样本数量;N_{T}表示实际为负样本且预测正确的样本数量;N_{F}表示实际为正样本但被预测错误为负样本的样本数量。

3.4 实验结果分析

3.4.1 自动阈值法有效性验证

1)自动阈值法与穷举法对比。将2个方法得到的不同新词输入到图1中的情感词汇倾向判断模块中,得到不同的情感词典,再用基于词典的情感计算方法对原语料库进行句子级情感倾向对比分析。本文在6类数据集上进行实验,包括外卖、计算机、水果、平板、书籍、酒店主题。表1展示了不同方法下的综合指标结果。

表1 穷举法与自动阈值法得到的新词质量比较

Table 1 Quality of new words obtained using exhaustive and adaptive threshold methods

数据集	P		A		R		F_1	
	穷举法	自动阈值法	穷举法	自动阈值法	穷举法	自动阈值法	穷举法	自动阈值法
外卖	0.802 0	0.826 8	0.773 7	0.787 2	0.725 2	0.726 3	0.761 1	0.773 3
计算机	0.763 4	0.774 9	0.793 8	0.811 1	0.852 1	0.876 7	0.805 2	0.822 7
水果	0.877 8	0.875 7	0.846 6	0.867 2	0.804 4	0.855 8	0.838 7	0.865 6
平板	0.865 1	0.867 5	0.818 0	0.830 0	0.753 3	0.779 0	0.805 0	0.820 8
书籍	0.763 8	0.766 0	0.755 6	0.759 8	0.798 9	0.805 7	0.780 9	0.785 3
酒店	0.813 5	0.827 9	0.817 3	0.823 1	0.823 2	0.815 8	0.818 3	0.821 8

表1显示,本文提出的自动阈值法在大多数情况下都优于穷举法,只有水果数据集精确率和酒店数据集召回率稍差。分析原因为:酒店数据集上自动阈值法生成的新词相较穷举法较少,被判断为积极的词汇较少,积极情感词典词汇不足,导致一些正样本没有被正确识别,从而降低了召回率。而水果数据集上自动阈值法生成的被判断为积极的新词较多,导致被误判为正向的负样本比例高于正确判断的正样本,从而降低了精确率。并且可以看到,当自动阈值法的精确率或召回率略低于穷举法时,其对应的召回率或精确率会显著高于穷举法。总的来说,自动阈值法的F_1值和准确率均高于穷

举法, 自动阈值法的综合表现优于穷举法。

2) 穷举法与自动阈值法输出的新词结果对比。为了深入分析表 1 结果, 以计算机领域语料为例, 进一步分析了自动阈值法在新词发现中的优势。表 2 对比了不同方法在新词发现中的结果。可以看出, 穷举法存在 2 个问题: 阈值过小时, 新词质量低下, 无意义的词汇组合过多, 如“送杀毒软件”、“过程中”等; 阈值过大时, 新词数量减少, 高质量新词被遗漏。而自动阈值法则能够较好地平衡新词的数量和质量, 有效提取出大部分的高质量新词。

3) 基于不同语料库的新词发现结果。表 3 给出了自动阈值法在 6 类数据集上发现的新词。结果显示, 阈值公式能够有效提取语料库中的新词, 并且这些新词大都是领域相关词汇。由此可见, 自动阈值法具有良好的领域自适应性。

表 2　计算机语料库上不同方法新词发现结果对比

Table 2　New word discovery results using different methods on a computer corpus

方法		新词发现结果	发现总数
穷举法	2	外观漂亮, 键盘很, 速度还可以, LED屏, 还不错, 送杀毒软件, 很实用, 回家, 驱动安装, 出现问题……	839
	4	外观漂亮, 速度还可以, LED屏, 还不错, 送杀毒软件, 很实用, 回家, 安装盘, 刚买, 出现问题……	639
	6	外观漂亮, LED屏, 送杀毒软件, 回家, 刚买, 提点, 出现问题, 很多人, 独显, 售后……	366
	8	LED屏, 杀毒软件, 回家, 独显, 小黑, 过程中, 售后, 第二天, 取货, 找不到, 帮朋友买, 光驱……	198
	10	杀毒, 回家, 独显, 光驱, 工作人员, 适合女孩子, 内存条, 蓝牙, 值得购买, 钢琴烤漆……	70
	12	杀毒, 质保, 微星, 双面胶, 散热, 便携, 散热量, 西数, 魔兽, 刻录, 磨砂, 双通道, 分区, 芯片组, 触控, 附赠, 2G	17
	14	散热, 魔兽	2
自动阈值法		独显, 售后, 杀毒软件, 续航能力, 微星, 客服, 双面胶, 质保, 便携, 散热量, 性价高, 蓝牙……	42

表 3　不同语料库新词发现结果

Table 3　New word discovery results on different corpora

语料库	新词发现结果
外卖	送餐, 菜品, 小哥, 百度, 店里, 尖椒, 给力, 宫保鸡丁, 土豆丝, 卷饼, 性价, 鸡腿, 京酱肉丝, 皮蛋瘦肉粥……
书籍	小熊, 到底, 职场, 育儿, 杜拉拉升职记, 神奇校车, 当当网, 亲子, 快递员, 男人, 文字, 提到, 小鸡……
计算机	独显, 售后, 杀毒软件, 续航能力, 客服, 双面胶, 质保, 便携, 显卡, 散热量, 性价高, 蓝牙, 触摸板, 钢琴烤漆……
水果	第一次, 给力, 下单, 直采, 网购, 甘肃天水, 售后, 客服, 冷链, 品相, 物有所值, 性价高, 火龙果, 次日达……
酒店	地理位置, 隔音效果, 补充点评, 入住, 携程, 门童, 前台, 总台, 退房, 结账, 身体健康, 四星级, 旺角, 网速……
平板	续航能力, 下单, 售后, 国人, 生日礼物, 手机壳, 做工精细, 官网, 王者荣耀, 运存, 皮套, 数据线, 性价高……

3.4.2　词汇的情感判别消融实验

进行情感倾向判别之前, 需要先确定词汇是否为情感词汇。情感词汇是能够表达或传递情感的词汇, 如“美”、“丑”等。而非情感词汇是没有情感色彩的词汇, 如“天气”、“吃饭”等。在实际应用中, 可以根据经验主观判断, 但在实验中, 需要一个客观的标准。为了确定情感词汇, 本文采用融合词性筛选和情感权重计算的方法。具体来说, 根据词性过滤掉一些非显著情感词汇, 再根据情感词典计算每个词的情感权重, 如果权重大于某个阈值, 则认为是情感词汇。下文分别讨论词性和情感权重对句子情感分析的影响。

1) 词性对情感计算结果的影响。本文的目的是找出对情感分析影响最大的词性组合。为此, 在计算机数据集上进行了实验。实验方法是使用自动阈值法确定情感词汇, 并随机设定情感权重阈值为 $\pm k$。实验结果如表 4 所示。可以得出: ①在所有词性组合中, [a, uw] 的组合取得了最高的准确率和 F_1 值。这说明新词词典、形容词对情感分析有重要影响。其中, 新词词典是本文创新点之一, 其能够有效捕捉网络语言中的新兴情感表达方式。②单独比较形容词、动词和名词, 可以发现形容词的准确率和 F_1 值都明显高于动词和名词, 这也符合常识, 因为形容词往往能够直接表达情感, 而动词和名词的情感则更加隐晦和依赖于上下文。③当名词加入词性组合时, 召回率会提高, 但准确率会降低。这可能是因为名词经常与某些情感词共现, 导致被误判为情感词, 从而影响了情感判断的正确性。然而, 表 2 显示了大部分语料库中的新词都是具有特定领域和情感倾向的名词, 它们不仅数量较少, 而且情感性较强。因此, 当形容词和新词组合时, F_1 值和准确率都有所提升。

2) 情感权重对情感计算结果的影响。词性筛选后得到候选情感词汇, 采用情感权重法判定其是

表4 不同词性组合的情感计算结果

Table 4 Sentiment calculation results for different part-of-speech combinations

词性组合	P	A	R	F_1
[a]	0.779	0.806	0.856	0.815
[a,uw]	0.775	0.811	0.877	0.823
[v]	0.723	0.776	0.894	0.800
[v,uw]	0.717	0.775	0.908	0.801
[n]	0.717	0.777	0.915	0.804
[n,uw]	0.712	0.775	0.923	0.804
[a,n]	0.719	0.779	0.916	0.806
[a,n,uw]	0.715	0.777	0.921	0.805
[v,a]	0.720	0.777	0.905	0.802
[v,a,uw]	0.722	0.776	0.897	0.800
[v,n]	0.711	0.773	0.918	0.801
[v,n,uw]	0.708	0.771	0.921	0.800
[v,a,n]	0.718	0.777	0.912	0.803
[v,a,n,uw]	0.721	0.779	0.910	0.804

注: uw表示新词词典; a表示形容词; v表示动词; n表示名词。

否为情感词汇。情感权重指一个词汇在表达情感时的重要程度, 可以通过式(6)和式(7)计算得到。确定情感权重阈值 k, 在此基础上对候选词进行分类: 只有当候选词的情感权重超过 k 时, 才认定为情感词汇; 否则认为不具有明显的情感色彩。为了寻找最佳阈值 k, 在不同数据集上进行实验, 固定词性为 [a,uw], 实验结果如表5所示。可以看出, 在不同数据集上, 当阈值为0.3时, 准确率和 F_1 值都较高, 因此, 本文最终确定最佳情感权重阈值 k 为0.3。

3.4.3 对比实验

1) 中文情感固定词典与通用情感词典比较。为了验证本文构建的中文情感固定词典的有效性, 将其与其他中文通用词典及所有通用词典的简单合并集合进行比较, 并采用基于词典的情感计算方法[21]对计算机评论语料进行了情感倾向分析, 实验结果如表6所示。可以看出, 中文情感固定词典各项评

表5 6类数据集上不同情感权重阈值下的情感计算结果

Table 5 Sentiment calculation results on six datasets with different sentiment weight thresholds

语料库	P							
	k=0	k=0.1	k=0.2	k=0.3	k=0.4	k=0.5	k=0.6	k=0.7
外卖	0.715 5	0.782 4	0.823 8	0.826 8	0.825 8	0.826 5	0.828 7	0.829 4
计算机	0.696 9	0.694 1	0.724 6	0.774 9	0.780 7	0.783 5	0.785 7	0.789 1
书籍	0.718 0	0.718 5	0.730 6	0.739 7	0.758 0	0.766 0	0.764 4	0.766 3
衣服	0.844 5	0.868 2	0.867 5	0.889 2	0.896 6	0.899 3	0.903 5	0.903 6
酒店	0.761 3	0.758 3	0.795 6	0.816 1	0.822 3	0.828 0	0.828 5	0.828 3
水果	0.845 5	0.874 3	0.868 4	0.875 7	0.889 0	0.890 6	0.890 4	0.890 4
语料库	A							
	k=0	k=0.1	k=0.2	k=0.3	k=0.4	k=0.5	k=0.6	k=0.7
外卖	0.760 3	0.800 1	0.786 0	0.787 2	0.783 5	0.782 7	0.780 4	0.780 7
计算机	0.757 7	0.756 2	0.781 3	0.811 1	0.803 6	0.804 1	0.804 1	0.805 8
书籍	0.733 3	0.735 7	0.746 3	0.752 3	0.757 5	0.759 8	0.755 4	0.755 9
衣服	0.883 2	0.897 3	0.897 4	0.902 1	0.904 2	0.894 9	0.892 3	0.892 1
酒店	0.796 3	0.799 3	0.812 1	0.819 4	0.818 0	0.820 2	0.820 3	0.819 8
水果	0.856 9	0.868 9	0.865 7	0.867 2	0.860 5	0.855 7	0.854 9	0.854 9
语料库	R							
	k=0	k=0.1	k=0.2	k=0.3	k=0.4	k=0.5	k=0.6	k=0.7
外卖	0.863 9	0.831 1	0.727 5	0.726 3	0.718 5	0.715 3	0.706 8	0.706 5
计算机	0.911 8	0.915 8	0.907 3	0.876 7	0.844 1	0.840 1	0.836 1	0.834 6
书籍	0.841 4	0.847 1	0.847 1	0.841 9	0.815 7	0.805 7	0.797 1	0.794 8
衣服	0.939 4	0.936 8	0.937 9	0.918 6	0.913 8	0.889 6	0.878 4	0.877 8
酒店	0.863 2	0.878 6	0.840 0	0.824 6	0.811 2	0.808 2	0.807 8	0.806 8
水果	0.873 8	0.861 6	0.862 0	0.855 8	0.823 8	0.811 0	0.809 4	0.809 4
语料库	F_1							
	k=0	k=0.1	k=0.2	k=0.3	k=0.4	k=0.5	k=0.6	k=0.7
外卖	0.782 7	0.806 0	0.772 7	0.773 3	0.768 4	0.766 9	0.762 9	0.763 0
计算机	0.790 0	0.789 7	0.805 7	0.822 7	0.811 2	0.810 8	0.810 1	0.811 2
书籍	0.774 8	0.777 5	0.784 6	0.787 5	0.785 8	0.785 3	0.780 4	0.780 3
衣服	0.889 4	0.901 2	0.901 4	0.903 7	0.905 1	0.894 4	0.890 8	0.890 5
酒店	0.809 0	0.814 0	0.817 2	0.820 3	0.816 7	0.818 0	0.818 0	0.817 4
水果	0.859 2	0.867 9	0.865 2	0.865 6	0.855 2	0.848 9	0.848 0	0.848 0

表 6　中文情感固定词典与通用情感词典比较

Table 6　Comparison of Chinese sentiment-fixed and general sentiment dictionaries

语料库	中文情感固定词典	NTUSD[12]	THU[15]	HowNet[14]	DU[13]	ALL
			P			
计算机	0.788 0	0.775 2	0.682 1	0.597 8	0.755 8	0.626 3
水果	0.886 2	0.817 5	0.753 2	0.701 6	0.799 9	0.695 0
手机	0.874 8	0.836 7	0.795 4	0.697 9	0.819 4	0.720 1
酒店	0.821 5	0.793 8	0.697 2	0.692 8	0.749 8	0.696 1
			A			
计算机	0.806 6	0.740 7	0.732 4	0.650 0	0.765 2	0.688 0
水果	0.836 7	0.715 1	0.758 9	0.749 6	0.742 3	0.751 9
手机	0.859 6	0.782 9	0.820 8	0.760 6	0.814 0	0.789 8
酒店	0.811 3	0.735 7	0.758 0	0.752 6	0.759 5	0.762 0
			R			
计算机	0.838 6	0.677 7	0.870 2	0.916 3	0.783 5	0.931 8
水果	0.772 6	0.553 7	0.770 0	0.868 4	0.646 1	0.897 8
手机	0.840 2	0.704 5	0.865 1	0.921 0	0.806 7	0.950 2
酒店	0.795 4	0.636 7	0.912 0	0.907 4	0.778 8	0.930 0
			F_1			
计算机	0.812 5	0.723 2	0.764 8	0.723 5	0.769 4	0.749 1
水果	0.825 5	0.660 2	0.761 5	0.776 1	0.714 8	0.783 5
手机	0.857 1	0.764 9	0.828 8	0.794 1	0.813 0	0.819 3
酒店	0.808 2	0.706 6	0.790 3	0.785 7	0.764 0	0.796 2

注：ALL为NTUSD、THU、HowNet、DU这几部词典对应指标值的并集结果。

价指标明显优于其他词典。精确率平均提高 12.77%，准确率平均提高 9.31%，F_1 值平均提高 7.43%。表明相较于通用情感词典，中文情感固定词典质量更高。可以看到，本文构建的中文情感固定词典在召回率上表现不是最佳，即在情感分析时，一些正样本没有被识别，同时可以发现，召回率高的一些词典在计算精确率时比较低。分析原因为：一些中性词汇既可能出现在正向句子中，也可能出现在负向句子中，但词典将其判断为积极词汇，从而导致一些正向句子被正确识别但也会将负向句子判断为正向，因此出现召回率高但精确率低的情况。

2）本文方法与其他方法比较。在中文情感固定词典的基础上，基于本文提出的领域自适应中文情感词典自动构建算法，可以得到领域情感词典。为了验证该方法的有效性，在中文情感分析和酒店语料库上将其与 SO-PMI 的新词发现方法[2]、基于互信息和邻接熵的新词发现方法[1]进行对比实验，结果如表 7 所示。可知，本文方法在 2 个数据集上准确率、召回率和 F_1 值都达到了最佳，准确率平均提高了 7.41%，召回率平均提高了 12.23%，F_1 值平均提高了 9.08%。还可以看到，本文方法在 2 个数据集上精确率都低于基于 SO-PMI 的方法，但召回率也相应地显著高于基于 SO-PMI 的方法。这是因为本文方法在预测时会将一些负样本预测为正样本，从而导致精确率的降低，但同时也会将更多的正样本预测为正样本，从而导致召回率的提高。

表 7　不同情感词典自动构建算法比较

Table 7　Comparison of different automatic construction algorithms for sentiment dictionary

语料库	本文方法	SO-PMI(最佳情感权重阈值为0.3)[2]	SO-PMI(最佳情感权重阈值为0)[2]	TwoSim[1]	Conj-TwoSim[1]
			P		
酒店	0.827 9	0.894 6	0.886 9	0.711 1	0.711 7
中文情感分析	0.864 9	0.870 6	0.901 8	0.758 9	0.768 8
			A		
酒店	0.823 1	0.781 6	0.786 8	0.748 7	0.749 4
中文情感分析	0.825 4	0.799 6	0.808 7	0.714 7	0.715 7
			R		
酒店	0.815 8	0.638 3	0.657 3	0.806 3	0.807 0
中文情感分析	0.771 2	0.703 7	0.692 7	0.635 7	0.633 2
			F_1		
酒店	0.821 8	0.745 0	0.755 1	0.755 7	0.756 4
中文情感分析	0.815 4	0.778 3	0.783 6	0.691 9	0.688 4

4　结　论

1）提出了一个领域自适应的中文情感词典自动构建算法。创新性地提出了一个中文情感固定词典，解决现有中文通用情感词典的情感歧义问题；基于 PMI 公式设计了自动阈值法，实现在不同领域自适应地提取新词；通过词性过滤和语义规则计算情感权重来判断词汇的情感倾向，最终得到领域情感词典。

2）本文提出的中文情感固定词典在计算机语料库上相较于其他中文通用情感词典，情感计算精确率平均提高 12.77%。

3）使用本文算法构造的领域情感词典较目前最新中文情感词典构建算法得到的情感词典，在2个数据库上情感计算结果召回率平均提高了12.23%。

参考文献（References）

[1] 刘伟童, 刘培玉, 刘文锋, 等. 基于互信息和邻接熵的新词发现算法[J]. 计算机应用研究, 2019, 36(5): 1293-1296.
LIU W T, LIU P Y, LIU W F, et al. New word discovery algorithm based on mutual information and branch entropy[J]. Application Research of Computers, 2019, 36(5): 1293-1296(in Chinese).

[2] YU S J, WANG B Y, LU T. Sentiment lexicon construction based on improved left-right entropy algorithm[J]. Journal of Donghua University (English Edition), 2022, 39(1): 65-71.

[3] CHEN Z, LI X, WANG M, et al. Domain sentiment dictionary construction and optimization based on multi-source information fusion[J]. Intelligent Data Analysis, 2020, 24(2): 229-251.

[4] ZHANG S X, XU H Q, ZHU G L, et al. A data processing method based on sequence labeling and syntactic analysis for extracting new sentiment words from product reviews[J]. Soft Computing, 2022, 26(2): 853-866.

[5] ZHANG W, ZHU Y C, WANG J P. An intelligent textual corpus big data computing approach for lexicons construction and sentiment classification of public emergency events[J]. Multimedia Tools and Applications, 2019, 78(21): 30159-30174.

[6] WU S X, WU F Z, CHANG Y, et al. Automatic construction of target-specific sentiment lexicon[J]. Expert Systems with Applications, 2019, 116: 285-298.

[7] HUANG S, NIU Z D, SHI C Y. Automatic construction of domain-specific sentiment lexicon based on constrained label propagation [J]. Knowledge-Based Systems, 2014, 56: 191-200.

[8] MA T H, RONG H, HAO Y S, et al. A novel sentiment polarity detection framework for Chinese[J]. IEEE Transactions on Affective Computing, 2022, 13(1): 60-74.

[9] ALMATARNEH S, GAMALLO P. Automatic construction of domain-specific sentiment lexicons for polarity classification[C]// International Conference on Practical Applications of Agents and Multi-Agent Systems. Berlin: Springer, 2018: 175-182.

[10] ZHANG B, XU D, ZHANG H, et al. STCS lexicon: spectral-clustering-based topic-specific Chinese sentiment lexicon construction for social networks[J]. IEEE Transactions on Computational Social Systems, 2019, 6(6): 1180-1189.

[11] KANAYAMA H, NASUKAWA T. Fully automatic lexicon expansion for domain-oriented sentiment analysis[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing. Sydney: Association for Computational Linguistics, 2006: 355-363.

[12] KU L W, LIANG Y T, CHEN H H. Opinion extraction, summarization and tracking in news and blog corpora[C]//Proceedings of the Spring Symposia on Computational Approaches to Analyzing Weblogs. Palo Alto: AAAI Press, 2006: 100-107.

[13] 徐琳宏, 林鸿飞, 潘宇, 等. 情感词汇本体的构造[J]. 情报学报, 2008, 27(2): 180-185.
XU L H, LIN H F, PAN Y, et al. Constructing the affective lexicon ontology[J]. Journal of the China Society for Scientific and Technical Information, 2008, 27(2): 180-185(in Chinese).

[14] FU X H, LIU G, GUO Y Y, et al. Multi-aspect sentiment analysis for Chinese online social reviews based on topic modeling and HowNet lexicon[J]. Knowledge-Based Systems, 2013, 37: 186-195.

[15] ZENG X K, YANG C, TU C C, et al. Chinese LIWC lexicon expansion via hierarchical classification of word embeddings with sememe attention[C]//Proceedings of the AAAI Conference on Artificial Intelligence. Palo Alto, USA: AAAI Press, 2018: 5650-5657.

[16] LIU J Y, YAN M J, LUO J. Research on the construction of sentiment lexicon based on Chinese microblog[C]//Proceedings of the 8th International Conference on Intelligent Human-Machine Systems and Cybernetics. Piscataway: IEEE Press, 2016: 56-59.

[17] ESULI A, SEBASTIANI F. Pageranking wordnet synsets: an application to opinion mining[C]//Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics. Prague: Association for Computational Linguistics, 2007: 424-431.

[18] TAN S B, WU Q. A random walk algorithm for automatic construction of domain-oriented sentiment lexicon[J]. Expert Systems with Applications, 2011, 38(10): 12094-12100.

[19] WANG L Y, RUI X. Sentiment lexicon construction with representation learning based on hierarchical sentiment supervision[C]//Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. Copenhagen: Association for Computational Linguistics, 2017: 502-510.

[20] SophonPlus. ChineseNlpCorpus[DS/OL]. https://github.com/SophonPlus/ChineseNlpCorpus/tree/master.

[21] 张成功, 刘培玉, 朱振方, 等. 一种基于极性词典的情感分析方法[J]. 山东大学学报 (理学版), 2012, 47(3): 47-50.
ZHANG C G, LIU P Y, ZHU Z F, et al. A sentiment analysis method based on a polarity lexicon[J]. Journal of Shandong University (Natural Science), 2012, 47(3): 47-50(in Chinese).

An adaptive automatic construction algorithm for sentiment dictionaries based on semantic rules

WEI Qinglan[1]，HE Yu[2]，SONG Jinbao[1,*]

(1. School of Data Science and Intelligent Media，Communication University of China，Beijing 100024，China；
2. School of Information and Communication Engineering，Communication University of China，Beijing 100024，China)

Abstract: Although text sentiment analyses using dictionaries are efficient and unsupervised, their accuracy relies heavily on the dictionary quality. The quality of existing Chinese sentiment dictionaries in cross-domain applications needs to be improved, as the manually constructed Chinese dictionaries fail to automatically discover new words and include emotionally ambiguous ones. In this paper, a domain-adaptive automatic construction algorithm for Chinese sentiment dictionaries was proposed by adopting semantic rules. The Chinese sentiment-fixed dictionary was constructed. It eliminated sentimental ambiguity effectively. A novel domain-adaptive method for discovering new Chinese words was proposed, which enabled the automatic expansion of the Chinese dictionary for general domains. An innovative unsupervised framework for recognizing words' sentiments based on part-of-speech filtering and semantic rules helps realize higher sentiment detection accuracy. The experiment proves that in the common computer corpus, sentiment analysis using the sentiment-fixed dictionary achieves an average accuracy improvement of 9.31%, precision improvement of 12.77%, and F_1 value improvement of 7.43%, compared to other general Chinese dictionaries. Meanwhile, on two datasets, the hotel and Chinese sentiment analysis corpora, the proposed algorithm for the automatic construction of sentiment dictionaries improves the accuracy by an average of 7.41%, the recall rate by 12.23%, and the F_1 value by 9.08%, compared to the advanced methods.

Keywords: sentiment dictionary; sentiment analysis; semantic rule; unsupervised; domain adaptation

Received：2023-06-15；**Accepted**：2023-12-01；**Published Online**：2024-03-07 18：12

URL：link.cnki.net/urlid/11.2625.V.20240305.1537.001

Foundation items：National Natural Science Foundation of China (72474198,62301510); Supported by "Asia Media Research Center, Communication University of China" (AMRC2023-4)

* **Corresponding author**. E-mail：songjinbao@cuc.edu.cn

http://bhxb.buaa.edu.cn jbuaa@buaa.edu.cn
DOI: 10.13700/j.bh.1001-5965.2023.0369

基于差异性特征蒸馏的多模态连续学习方法

贺驰原，程少旭，许林峰*，孟凡满，吴庆波
(电子科技大学 信息与通信工程学院，成都 611731)

摘　　要：近年来连续学习成为一个新的研究热点，但在多模态架构的连续学习任务中，数据不能被完全利用，导致了严重的灾难性遗忘和学习受阻问题。因此，提出了基于特征蒸馏的多模态连续学习方法。该方法重点考虑不同模态在任务表现方面的差异性，选择较多或较少地保留模态旧知识，以激发各模态从整体角度挖掘具有判别性特征的潜力。在多模态行为识别数据集 UESTC-MMEA-CL 上的实验验证了所提方法的有效性。在进行到第 8 个任务时，所提方法的平均准确率在微调基础上提升了 22.0%，在不遗忘学习 (LwF) 的基础上提升了 20.1%。与经典的知识蒸馏方法相比，提出的差异性特征蒸馏方法显著提高了传感器模态的利用率，从而更显著地缓解了多模态网络的灾难性遗忘问题。

关 键 词：机器学习；连续学习；多模态；行为识别；特征蒸馏
中图分类号：TP391
文献标志码：A　　**文章编号**：1001-5965(2025)07-2460-08

传统深度学习算法不能在非独立同分布的数据集上学习良好的效果，只有当训练数据与测试数据相似时，才能发挥出算法的性能。但是在当前时代，信息呈指数式爆炸增长，人工智能应当具备自适应学习的能力，如何让机器像人类一样具备学习新知识并且不忘旧知识的能力至关重要。在这一契机下，连续学习(增量学习、终生学习)这一研究方向被提出。连续学习旨在督促神经网络学习新知识的同时不忘记旧知识。具体而言，深度神经网络需要分阶段式地进行训练，而在后续任务中无法重新访问到之前任务的训练数据。在这一过程中伴随着灾难性遗忘的问题，即网络在学习新的知识后无法保有对于旧知识的有效性。目前，通过合适的方法可以使网络学习过的任务保持精度的同时，在新任务的训练中又得到较好的学习效果，即算法应同时具有可塑性和稳定性是研究的重点。许多工作对网络在连续学习过程中的灾难性遗忘问题加以改善，其主要分为基于参数保护、基于知识蒸馏及基于参数扩展的连续学习策略。

多模态学习是另一个重要的研究方向，其要求神经网络通过来自不同模态的信息综合判断给出预测，并且预测精度相较单模态应有一定的提高。当前，多模态学习的概念还很广泛，从输入数据信息的角度和网络架构一般可以分为以下几种形式：①输入数据同构、网络同构(如 RGB 图像、合成孔径雷达图像模态)；②输入数据异构、网络同构 (如 RGB 图像、音频模态)；③输入数据和网络均异构 (如 RGB 图像、惯性测量单元传感器模态)。文献 [1] 考虑使用多模态网络进行连续学习，提出了多模态第一人称视角的大规模行为识别数据集 UESTC-MMEA-CL，并进行了详细地任务划分，在多模态网络 Time Binding Network[2] 的基础上，使用视频图像、加速度计传感器、陀螺仪传感器 3 种模态的

收稿日期：2023-06-15；**录用日期**：2023-12-01；**网络出版时间**：2024-02-26 11:14
网络出版地址：link.cnki.net/urlid/11.2625.V.20240221.1715.001
基金项目：新一代人工智能国家科技重大专项 (2021ZD0112001)；国家自然科学基金 (62071086)；四川省自然科学基金创新研究群体项目 (2023NSFSC1972)
* **通信作者**. E-mail：lfxu@uestc.edu.cn

引用格式：*贺驰原，程少旭，许林峰，等. 基于差异性特征蒸馏的多模态连续学习方法 [J]. 北京航空航天大学学报，2025，51（7）：2460-2467.*
HE C Y，CHENG S X，XU L F，et al. Continual learning method based on differential feature distillation for multimodal network[J]. Journal of Beijing University of Aeronautics and Astronautics，2025，51（7）：2460-2467（in Chinese）.

数据进行人类日常行为识别任务, 并严格制定了多模态场景下的连续学习任务, 在不同的模态组合与经典的连续学习方法下评测了基线模型的表现。

出于直观的印象, 多个模态的数据带来了更为广泛的信息空间, 使神经网络对于某一事物的判别有了更多的依据。理论上, 在连续学习的过程中, 不同模态可以通过互补的方式减少灾难性遗忘, 但事实上, 在自适应训练的过程中 (使用经典的连续学习方法施加约束), 多模态网络并没有“灵活”地学习到一种策略来进行这种互补。至今, 多模态网络在几种代表性连续学习方法中的表现均很难超过其中表现最好的单模态网络[1]。为此, 本文为网络设计一种灵活的连续学习策略来提升多模态网络在连续学习中的各模态利用率。本文方法基于模型蒸馏, 在当前任务中的蒸馏强度取决于各模态网络在上一任务中的贡献度。实验证实了本文方法相对基线模型的显著提升。

1　相关工作

1.1　多模态学习

1.1.1　相关数据集

为了补充视觉数据无法表达的且有助于识别的关键信息, 一些工作提出了多模态形式的数据集, 其中最具代表性的是关于人类行为活动的多模态数据集 (包含 2 种或更多种模态信号, 如视频、音频、传感器等)。这些多模态的数据可以反映更真实和细致的人类行为信息, 有助于提高行为识别的准确率。

根据场景的不同, 可以将多模态行为活动数据集分为厨房环境、日常生活环境和其他环境 3 种类型。厨房环境最为常见, 因为厨房中存在着丰富和复杂的物体操作和交互行为。EPIC-KITCHENS[3] 是一个基于第一人称视角的大规模多模态数据集, 由 32 名参与者在厨房环境中收集, 每位参与者都被要求使用头戴式相机记录下他们进入厨房后的每 1 秒。该数据集包含 RGB、光流和音频等多种模态信号, 并标注了 11 万个粗粒度和 39 万个细粒度级别的行为标签。CMU-MMAC[4] 介绍了 29 个厨房活动, 这些数据利用第一人称视角相机和惯性运动传感器进行采集。日常生活环境是另一种常见的场景类型, 因为日常生活涉及到各种各样且具有挑战性的活动, UTD-MHAD[5] 由 Kinect 摄像头和可穿戴惯性传感器收集, 由 RGB 视频、深度视频、骨骼位置和 27 种不同人类动作 (如右臂投掷、胸前交叉手臂、篮球射击等) 的惯性信号组成。如今, 由于生活方式分析和健康监测等有趣的生活日志应用, 基于第一人称视角的行为识别已成为人们广泛关注的话题。然而, 由于从智能眼镜等可穿戴设备中捕获多模态数据并不容易, 多模态的行为识别进展相对缓慢。

文献 [1] 综合考量了上述各个数据集的局限性, 提出了 UESTC-MMEA-CL 数据集。不同于现有的多模态自我中心活动数据集[3-5], UESTC-MMEA-CL 数据集[1] 由自主开发的独立的相机和传感器集成设备采集, 适用于生活记录的可穿戴设备。收集的人类活动视觉数据和传感器数据同步良好, 共包含 32 个日常活动类别, 每个样本剪辑由视频、加速度和陀螺仪信号组成, 可以提供丰富的物体和运动信息。

1.1.2　多模态行为识别方法

对于复杂情况下的人类活动识别, 仅使用视觉数据可能会由于场景和物体的噪声而缺乏对行为的识别。为了更稳定地分析人类活动, 一些研究者将视觉信号与其他模态的数据相结合, 以提高识别的准确性和鲁棒性。例如, Song 等[6] 将视觉信号分解为单帧、光流和稳定光流 3 种输入形式, 借助陀螺仪和加速度计数据对活动进行分类。Kazakos 等[2] 提出了一个多模态时间绑定框架用以结合来自视频、光流和音频 3 种模态信号, 以及 1 个中级融合时间绑定网络。与传统的融合方法不同, 多模态信号在时间融合前聚合, 并随时间共享权重。Nakamura 等[7] 采用了一个堆叠的长短期记忆网络 (long short-term memory, LSTM) 来处理视觉和加速度的融合特征, 并在心率传感器数据的帮助下进行活动识别。此外, 还有一些研究致力于通过分析中级特征 (如人们的脸、目光和手) 来研究人类的意图。

1.2　连续学习

1.2.1　连续学习任务

连续学习任务强调模型能在学习新知识的同时不忘或尽可能地少忘记旧知识。具体而言, 给出划分好的多个任务, 要求模型依次学习每个任务, 并且在下一个任务到来时, 之前任务的数据不可以复习。每学习完一个新的任务, 对模型进行评测, 评测的范围包括新任务与之前的所有任务。类增量学习 (class-incremental learning, CIL) 是连续学习最具代表性的任务之一, 网络需要依次进行每个任务的类别的学习, 且不再复习, 并完成在当前时刻对于之前所有出现过的类别的分类任务, 这非常具有挑战性。

1.2.2　连续学习方法

为了提高网络在 CIL 任务上的性能, 已经有许

多工作对网络在连续学习过程中的灾难性遗忘问题加以改善。现有工作主要分为基于参数保护、基于知识蒸馏及基于参数扩展3类。

基于参数保护的方法旨在评估参数的重要性并保护重要参数。最早提出的弹性权重巩固(elastic weight consolidation, EWC)[8]提出了对旧任务关键参数施加二次惩罚的方法,并利用Fisher信息矩阵来选择关键参数。Liu等[9]通过旋转参数空间获得了更好的Fisher信息矩阵近似,但可能会高估或低估。为了解决此问题,突触智能[10]积累了参数学习的变化,通过这些变化来估计重要性。

基于知识蒸馏的方法也十分常见,这类方法的关键是利用旧模型或老师模型来指导新模型或学生模型学习新任务和旧任务。最具代表性的不遗忘学习(learning without forgetting, LwF)[11]同时训练新任务和旧任务输出层,并将新模型与旧模型对齐从而保持对旧任务输出层分布的一致性。增量分类器和表示学习(incremental classifier and representation learning, iCaRL)[12]则结合特征提取器和分类器更新,并使用记忆单元存储部分数据样本以进行重放和蒸馏训练。为了进行更细粒度的蒸馏,池化输出蒸馏[13]使用层间特征间的多尺度池化蒸馏方法,其更看重新旧特征间的差异性而不仅是输出逻辑值之间的差异。适配特征巩固[14]提出一种评估网络每层对于当前任务知识重要性程度的方式,并进行具有选择性的蒸馏。

基于参数扩展的方法的关键是动态地增加网络容量以适应新任务而不影响旧任务。最具代表性的方法是动态扩展表示[15],其根据每个新任务需要动态地添加或剪枝神经元并使用选择性重新训练策略来减少灾难遗忘。为了克服随着任务增多而导致的网络参数量增长,提出了内存高效可扩展模型(memory efficient expandable model, MEMO)[16]、特征增强与压缩方法(feature boosting and compression, FOSTER)[17]、自维持表征扩展机制(self-sustaining representation expansion, SSRE)[18]等策略。MEMO通过扩展最后一个线性输出层来适应新任务。FOSTER则通过扩展新的骨干网络并结合知识蒸馏的思想将新旧骨干网络压缩为单体骨干网络,以达到参数量与旧网络持平的目的。SSRE通过网络参数重构的方式将新训练的网络层参数叠加回旧的网络参数上,以使网络适应新旧任务。

2 本文方法

2.1 多模态连续学习任务

文献[1]提出了一个多模态连续学习任务的实际需求场景,即一个多模态行为识别模型在部分基本源数据上进行了良好的离线训练,并有非常高的预测精度。但随着一系列新场景中的行为活动被纳入到行为识别的需求中,新数据增量式输入。因受限于隐私问题、存储限制或计算开销,旧场景中的行为数据已无法纳入与新数据的联合对比训练中,此情形可以概括为图1。因此,Xu等[1]认为一个多模态的网络应具有可以连续学习或终生学习的能力,即在学习新知识的同时不忘记或尽可能少忘记旧知识,并试图解决其中普遍存在的灾难性遗忘问题。Xu等[1]尝试利用一些经典的连续学习方法以改善多模态网络连续学习过程中的灾难性遗忘问题,其中包括基于参数保护的正则化方法EWC[8]、基于知识蒸馏方法LwF[11]及基于样本重放的方法iCaRL[12]。Xu等[1]发现,对比微调(不使用任何连续学习方法进行任务),3种方法均能一定程度地缓解多模态网络的灾难性遗忘问题,但多模态网络在联合预测的准确率超过单模态的情况下,存在比单模态更严重的遗忘,如图像模态加上加速度传感器模态的模型在EWC和LwF的方法下无法比肩单独的图像模态。显然,这与引言中提到的直观印象不符。基于此,本文提出基于模态重要性评估的差异性特征蒸馏方法解决多模态场景下严重的灾难性遗忘问题。

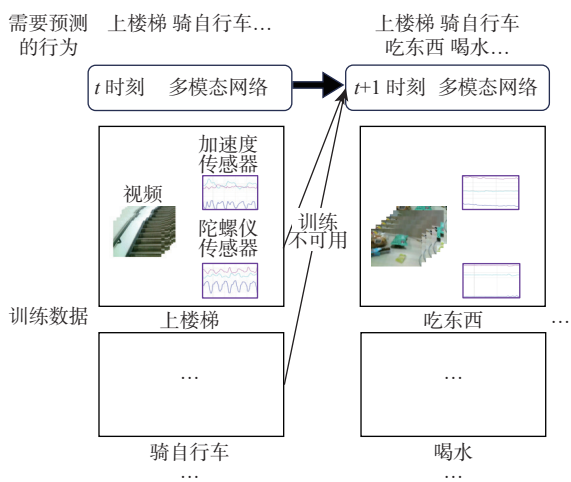


图1 多模态连续行为识别示意图

Fig. 1 Multimodal continual behavior recognition

2.2 方法架构

本文方法架构如图2所示,核心理论方法主要集中在模态重要性评估和差异性特征蒸馏2部分。主要思想为:通过模态重要性评估的模块得到不同模态分支对于该任务知识的敏感程度,并对模型蒸馏量进行动态调控,以达到在该任务上保护“对以往任务重要的模态”,激励“尚未表达或较少表达的模态”的目的。

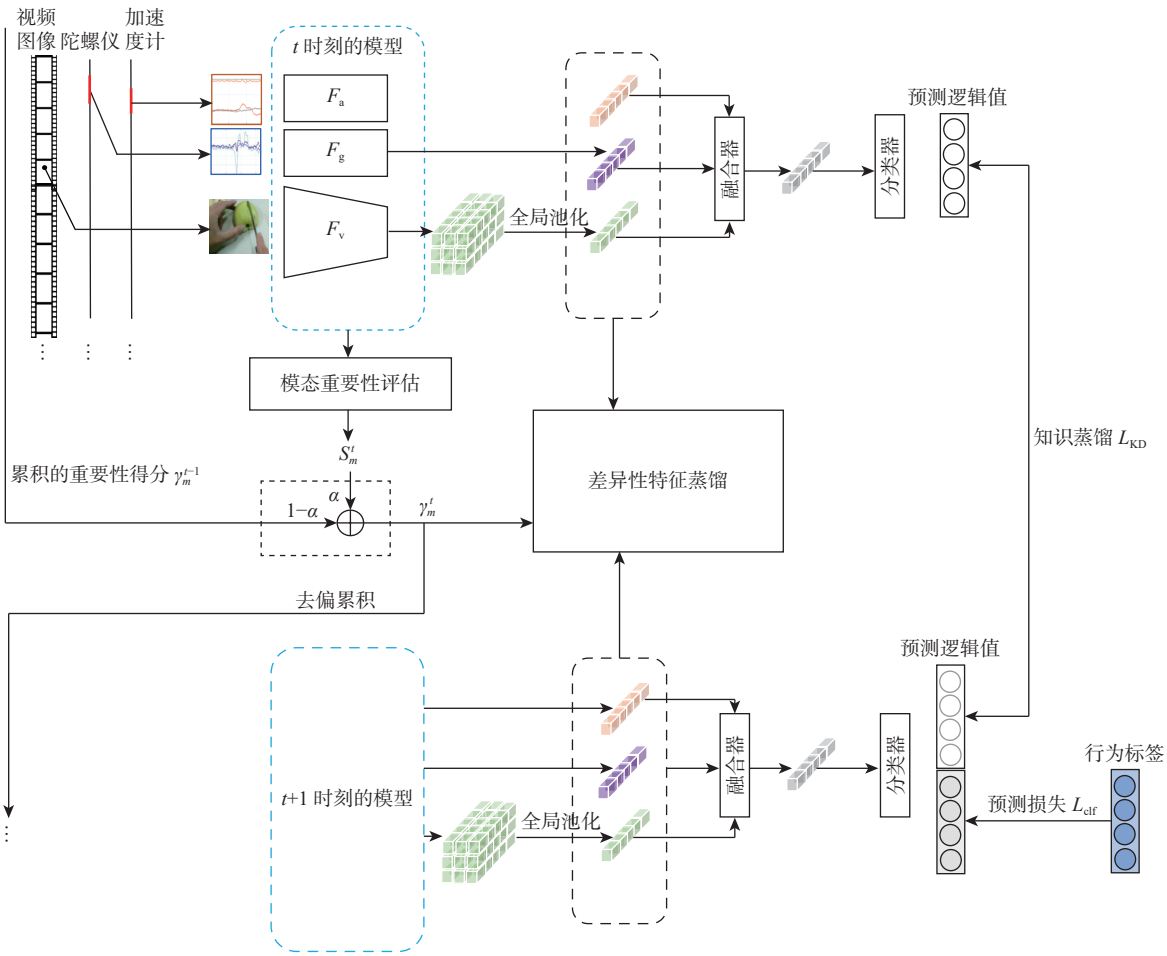


图 2　差异性特征蒸馏示意图

Fig. 2　Diagram of differential feature distillation

2.2.1　模态重要性评估

在一个任务的训练中, 模态会自适应地根据上述 2 个目标督促每个模态学习新知识并保存旧知识。为了了解每个模态在每个任务中表达的情况, 本文提出了模态重要性评估的方法。根据网络参数的费雪信息量推测网络对于某一任务中知识的响应程度, 这一思想最早被应用在文献 [8] 中, 其用来评估参数对于该批量数据的响应程度。具体而言, 分别保留在任务 $t(t>1)$模型训练完成的各模态特征提取器骨干的可训练参数 $\theta(F_m^t), m\in\{v,a,g\}$($v$、$a$、$g$ 分别表示视频图像、加速度计、陀螺仪), 对模型进行 1 个训练轮次的前向传播, 并根据该任务的所有行为类别预测结果与真实标签的交叉熵计算损失函数, 进行反向传播产生各部分参数的梯度。其中, p_m^t为 t 任务时刻模态 m 特征提取器网络骨干的可训练参数, $p_m^t\in\theta(F_m^t)$。对于当前任务 t, 利用对于该任务所有训练数据 D_t作为激活输入, 其为多模态数据集合, 包含视频图像、加速度计和陀螺仪数据。通过反向传播算法得到网络参数的梯度 ∇p_m^t。因为模态网络的异构性, 采用 m 模态梯度模 (平方项) 的平均值作为对 m 模态的重要性程度。具体地, m模态重要性程度得分可以表达为

$$Q_m^t=\frac{1}{\left|\theta(F_m^t)\right|}\sum_{p_m^t\in\theta(F_m^t)}\left|\nabla p_m^t\right|^2 \tag{1}$$

将其在各模态上归一化得到重要性程度系数为

$$S_m^t=\frac{1}{|m|}\sum_m Q_m^t \tag{2}$$

受文献 [8] 的启发, 本文使用去偏累积的方式累计当前任务和之前所有任务的模态重要性程度得分结果, 因为随着任务数的增多, 越来越偏向于保留旧知识, 不重视它们会导致更高昂的代价。t时刻 m 模态累积的重要性程度为

$$\gamma_m^t=\alpha S_m^t+(1-\alpha)\gamma_m^{t-1} \tag{3}$$

初始任务时刻, 设定累积量为 0, 即 $\gamma_m^0=0$。去偏系数 $\alpha=N_t\Big/\sum_{i=1}^{t}N_i$, 其中, N_t为当前任务所包含的行为类别数, $\sum_{i=1}^{t}N_i$为之前任务的总行为类别数。因为在连续学习的过程中, 网络总有偏向新任务类别知识的趋向, 而这种趋向可以通过系数进行修正, 这意味着随着任务数的增加, 保护旧任务的知识比学习新知识更重要。

2.2.2 差异性特征蒸馏

特征蒸馏是一种利用特权特征(如交叉特征、后事件类特征等)指导学习模型的方法,其可以提高模型的泛化能力和效率。特征蒸馏通常应用于“教师-学生”模型中,其通过让一个性能优异的教师模型向另一个学生模型传递有效且高度抽象的特征表示,从而提升学生模型的能力。这种传递可以发生在深层或浅层。本文希望多模态模型能够根据不同任务合理地分配各个模态的学习权重,并充分利用多源信息。例如,在当前任务中,如果某个模态A比较重要或对其知识比较依赖,本文尝试保持该模态的知识不被遗忘;反之,如果某个模态B比较次要或对其知识比较鲁棒,本文就会降低该模态的保护程度,并鼓励其在下一个任务中更加活跃。

基于此,本文利用在上一个任务中暂存的旧模型来指导当前任务的新模型,使学习到重要知识的模型参数得以被保留。记当前任务的多模态网络分支为$\boldsymbol{F}_m^t(\cdot)$,上一任务暂存的旧模型为$\boldsymbol{F}_m^{t-1}(\cdot)$,多模态输入为$x_m^t=\{x_{\mathrm{v}}^t,x_{\mathrm{a}}^t,x_{\mathrm{g}}^t\}$,可得到$\boldsymbol{F}_m^t(x_m^t)$和$\boldsymbol{F}_m^{t-1}(x_m^t)$。利用Norm-L2范数在得到的2个特征之间进行约束,从而约束新旧模型之间的变化幅度:

$$L_{\mathrm{D}}^{m,t}=\|\boldsymbol{F}_m^t(x_m^t)-\boldsymbol{F}_m^{t-1}(x_m^t)\| \tag{4}$$

结合上一任务t-1时刻的模态重要性评估结果S_m^{t-1},将其作为线性加权的权重以获得差异性特征蒸馏损失:

$$L_{\mathrm{D}}^t=\sum_m\gamma_m^{t-1}L_{\mathrm{D}}^{m,t} \tag{5}$$

2.2.3 连续学习过程训练损失函数

在一次完整的任务5的训练中,监督多模态网络所使用的损失函数为

$$L^t=L_{\mathrm{D}}^t+L_{\mathrm{clf}}^t+L_{\mathrm{KD}}^t \tag{6}$$

式中:L_{clf}^t为预测当前任务行为类别的交叉熵损失;L_{KD}^t为面向最后预测逻辑值的知识蒸馏损失[11]。

3 实 验

3.1 实验环境

3.1.1 数据集

实验基于UESTC-MMEA-CL[1]多模态行为识别数据集,该数据集包含32个不同场景的人类日常行为活动,如喝水、骑自行车、上楼梯等。每个数据样本包括1组同步的视频图像、加速度计传感数据与陀螺仪传感数据。每个类别共包含约200段数据样本,每段时长约为17 s。为确保实验的公平性,使用的训练、测试数据划分均与文献[1]中的基线一致。

3.1.2 传感器数据预处理

由于陀螺仪是测量角速度的传感器,其受到温度、零点漂移等因素的影响,需要进行校准和滤波处理。本文对于数据的处理方法与文献[1]保持一致。对加速度和陀螺仪信号分别进行内核大小为5的中值滤波,以去除异常值。由于陀螺仪信号存在累积误差,采用积分将滤波后的角速度转换为角度变化量。另外,为了消除陀螺仪信号中由于温度等因素引起的零点漂移问题,在积分之前对每个轴向上减去其平均值作为去偏处理。

3.1.3 多模态网络参数

使用的网络与文献[1]保持一致,其详细信息与参数设置如表1所示。

表1 网络参数设置

Table 1 Network parameter setting

模态	输入维度	特征提取器	融合器	分类器
视频图像	[8,3,224,224]	BN-Inception[19]	Concat+FC+ReLU	FC+Softmax
加速度计	[8,3,64]	DeepConvLSTM[20]		
陀螺仪	[8,3,64]	DeepConvLSTM[20]		

注:FC为全连接层。

3.1.4 多模态训练配置

视频流分支由随机梯度下降训练优化器[21],动量为0.9,批量大小为8,丢弃率为0.5,学习率为0.001。加速度计和陀螺仪网络分支由均方根传播训练优化器,丢弃率为0.5,学习率为0.001。本文初始化来自ImageNet的预训练模型的RGB网络。全部网络训练了50个迭代轮次,学习率在轮次为10和20处衰减了10倍。模型的连续学习任务设定与文献[1]保持一致。

3.2 实验结果

3.2.1 对比基线

对比了所有多模态组合下(视频图像+加速度计+陀螺仪、视频图像+加速度计、视频图像+陀螺仪)基线模型和使用差异性特征蒸馏模型的遗忘曲线,如图3所示。可以发现,在微调和知识蒸馏方法下,加入差异性特征蒸馏均能大幅优化多模态网络在连续学习中的表现。如图3(a)所示,在视频图像、加速度计、陀螺仪三模态网络进行到第8个任务时,使用差异性特征蒸馏方法的模型准确率比原有模型提高超过20%(在微调方法下提高22.0%、在LwF方法下提高20.1%),在其他模态组合方式下也有大幅度提高。

3.2.2 消融实验

在原有的基线模型中,由于传感器模态没有被合理和充分地利用,并且在连续学习过程中存在优化复杂性与不平衡性,导致多模态网络的性能无法

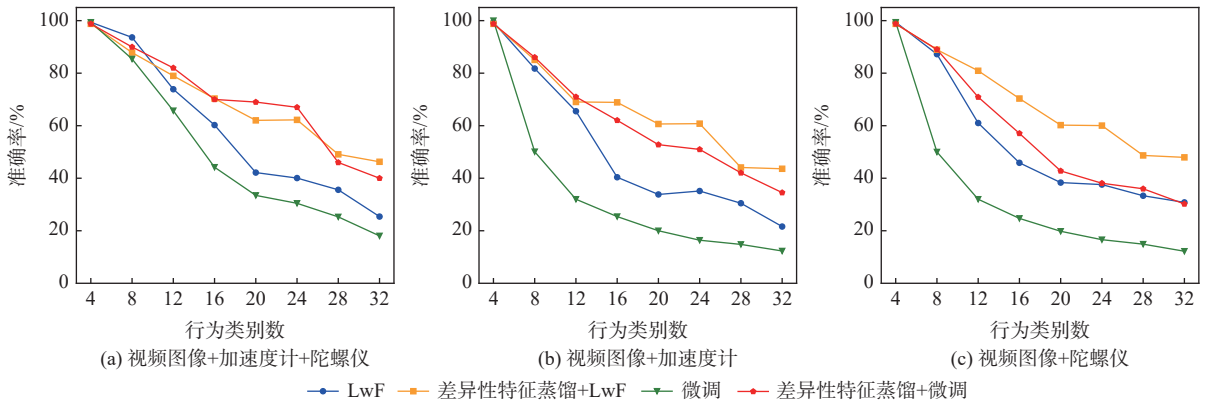


图 3　遗忘曲线对比

Fig. 3　Comparison of forgetting curves

高于最优的单模态网络。利用差异性特征蒸馏方法后, 这种情况得到显著改善, 使用最后 1 个任务时的类平均准确率来评价网络的连续学习性能, 如表 2 所示。可见, 差异性特征蒸馏方法能够更合理地利用各模态网络, 发挥其知识互补性, 从而表现优于最好的单模态网络 (视频图像模态)。

表 2　消融对比

Table 2　Ablation comparison

模型+连续学习策略	平均准确率/%	
	8×4	4×8
视频图像单模态+LwF	29.3	53.2
三模态+LwF	25.8	39.0
三模态+LwF+特征蒸馏	30.6	52.5
三模态+LwF+差异性特征蒸馏	46.3	58.9

注: “8×4” 代表将全32个行为类别分成8个任务, 每个任务中有4个类别; “4×8” 代表将全32个行为类别分成4个任务, 每个任务中有8个类别。

用, 在 “8×4” 的任务设置下, 将网络对于第 1 个任务 (初始任务) 随着任务数目的增加所保持的准确率作为衡量指标, 对比如图 4 所示。在仅使用微调的方法下, 单模态网络对于初始任务知识的灾难性遗忘较严重, 去偏的特征蒸馏能够对其进行一定程度的缓解, 如在视频图像单模态的任务下, 当进行到第 5 个任务时 (行为类别数为 20), 微调方法准确率仅为 16.9%, 而特征蒸馏方法的准确率可以达到 68.0%, 性能提升至微调的 4 倍以上。另外, 微调方法下, 多模态网络无法记忆任何长期知识, 初始任务所学习到的知识全部被遗忘, 而本文方法对其有显著改善。

3.2.3　模态利用率对比

为了直观说明本文方法在每个任务的学习中能够更合理地利用各模态, 将每个模态的重要性程度得分 (如式(1)所示) 随着任务的增加累积起来, 以表示每个模态的激活响应程度。图 5 对比了进行和不进行差异性特征蒸馏时, 各个模态的激活响

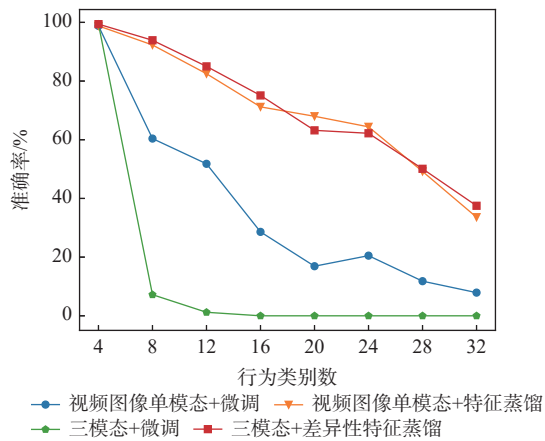


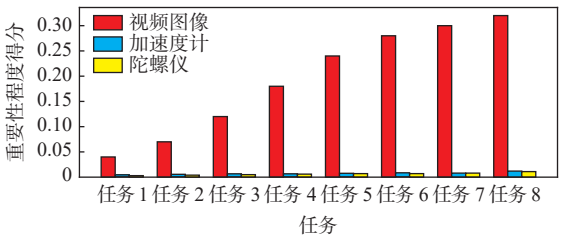
图 4　初始任务的遗忘情况对比

Fig. 4　Comparison of forgetting status of initial task

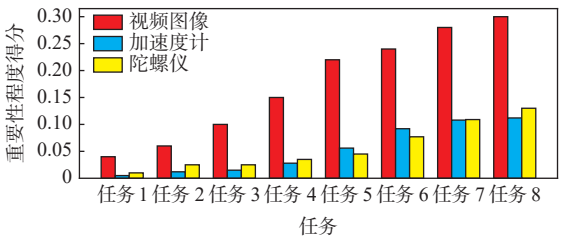
应程度。可以观察到, 通过普通的全局蒸馏策略, 随着任务的增加, 传感器模态 (加速度计和陀螺仪) 发挥非常有限的作用 (随着任务的增加没有显著的增长), 而网络对于知识的响应集中于视频图像模态; 相反地, 差异性特征蒸馏策略通过鼓励先前任务中贡献度较低的模态学习后续任务的知识并保护重要性程度高的模态, 使传感器模态的响应显著提高, 这也再次佐证了本文的观点。

3.2.4　与现有多模态融合方法对比

现有的针对多模态信息融合方法对于行为识别有着一定的提升, 但其在设计时都缺乏针对可持续任务的考量。参照文献 [1], 比较了本文使用的融合方法与 2 种先进的多模态融合方法, 包括多层次注意力融合网络 (multi-level attention fusion network, MAFnet)[22] 中使用的时间-模态注意力及多模态图卷积神经网络 (multimodal graph convolutional neural network, MM-GCN)[23] 中使用的图卷积融合。如表 3 所示, 比较了方法在所有 32 个行为类别上联合训练的准确率和在 “8×4” 连续任务最后 1 个任务训练结束后的准确率。可以发现, 这些细粒度的多模态融合方法并不能很好地适用于连续学习任务。



(a) 基线模型+LwF



(b) 基线模型+LwF+差异性特征蒸馏

图 5 模态利用率对比

Fig. 5 Modality utilization comparison

表 3 融合方法对比

Table 3 Fusion method comparison

模型+连续学习策略	准确率/%	
	32类	8×4
三模态-MAFnet[22]	95.4	12.2
三模态-MM-GCN[23]	96.0	16.1
三模态-Concat + FC + ReLU	95.6	58.9

3.2.5 去偏系数的有效性

在 2.2.1 节中，采用去偏累加的方式累加模态重要性程度。为了验证其合理性，在“4×8”的连续任务下对比了去偏累加和直接累加的任务准确率情况，如表 4 所示。可以发现，随着任务进行到 3 和 4，注重旧知识上的模态重要性程度比新知识上的模态重要性程度显得更重要。

表 4 去偏系数的有效性 (三模态)

Table 4 Effectiveness of debiasing coefficient (triple modalities)

累加方式	准确率/%			
	任务1	任务2	任务3	任务4
直接累加	98.3	82.5	62.2	53.0
去偏累加	98.3	82.5	68.2(+6.0)	58.9(+5.9)

注：括号内数值为相对于“直接累加”的准确率相对增益。

4 结 论

1）针对多模态网络在连续学习过程中遇到的问题，即在现有方法下，无法合理并充分利用各模态之间的信息互补性而导致相较于单模态网络更严重的灾难性遗忘和学习受阻问题，提出基于模型特征蒸馏的差异性保护方法，动态平衡各模态网络在连续学习过程中如何有效地保护旧知识并学习新知识。

2）在 UESTC-MMEA-CL 数据集上的实验结果展示了本文方法在基线模型的基础上有大幅度提高，模态利用率可视化再次佐证了本文的观点。

3）本文方法是灵活可迁移的，可以被运用在各种经典的连续学习方法与多模态网络结构上。

参考文献（References）

[1] XU L F, WU Q B, PAN L L, et al. Towards continual egocentric activity recognition: a multi-modal egocentric activity dataset for continual learning[EB/OL].(2023-01-26)[2023-06-01]. https://arxiv.org/abs/2301.10931v1.

[2] KAZAKOS E, NAGRANI A, ZISSERMAN A, et al. EPIC-fusion: audio-visual temporal binding for egocentric action recognition [C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE Press, 2019: 5491-5500.

[3] DAMEN D M, DOUGHTY H, FARINELLA G M, et al. Rescaling egocentric vision: collection, pipeline and challenges for EPIC-KITCHENS-100[J]. International Journal of Computer Vision, 2022, 130(1): 33-55.

[4] SPRIGGS E H, DE LA TORRE F, HEBERT M. Temporal segmentation and activity classification from first-person sensing[C]// Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops. Piscataway: IEEE Press, 2009: 17-24.

[5] CHEN C, JAFARI R, KEHTARNAVAZ N. UTD-MHAD: a multimodal dataset for human action recognition utilizing a depth camera and a wearable inertial sensor[C]//Proceedings of the IEEE International Conference on Image Processing. Piscataway: IEEE Press, 2015: 168-172.

[6] SONG S B, CHANDRASEKHAR V, MANDAL B, et al. Multimodal multi-stream deep learning for egocentric activity recognition[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. Piscataway: IEEE Press, 2016: 378-385.

[7] NAKAMURA K, YEUNG S, ALAHI A, et al. Jointly learning energy expenditures and activities using egocentric multimodal signals[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2017: 6817-6826.

[8] KIRKPATRICK J, PASCANU R, RABINOWITZ N, et al. Overcoming catastrophic forgetting in neural networks[J]. Proceedings of the National Academy of Sciences of the United States of America, 2017, 114(13): 3521-3526.

[9] LIU X L, MASANA M, HERRANZ L, et al. Rotate your networks: better weight consolidation and less catastrophic forgetting[C]// Proceedings of the 24th International Conference on Pattern Recognition. Piscataway: IEEE Press, 2018: 2262-2268.

[10] ZENKE F, POOLE B, GANGULI S. Continual learning through synaptic intelligence[C]//Proceedings of the 34th International Conference on Machine Learning. Sydney: PMLR, 2017, 70: 3987-3995.

[11] LI Z Z, HOIEM D. Learning without forgetting[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018, 40(12):

2935-2947.

[12] REBUFFI S A, KOLESNIKOV A, SPERL G, et al. iCaRL: incremental classifier and representation learning[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2017: 5533-5542.

[13] DOUILLARD A, CORD M, OLLION C, et al. PODNet: pooled outputs distillation for small-tasks incremental learning[C]// European Conference on Computer Vision. Berlin: Springer, 2020: 86-102.

[14] KANG M, PARK J, HAN B. Class-incremental learning by knowledge distillation with adaptive feature consolidation[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2022: 16050-16059.

[15] YAN S P, XIE J W, HE X M. DER: dynamically expandable representation for class incremental learning[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, IEEE Press, 2021: 3013-3022.

[16] ZHOU D W, WANG Q W, YE H J, et al. A model or 603 exemplars: towards memory-efficient class-incremental learning[C]//Proceedings of the 11th International Conference on Learning Representations. Kigali: ICLR, 2023.

[17] WANG F Y, ZHOU D W, YE H J, et al. FOSTER: feature boosting and compression for class-incremental learning[C]//European Conference on Computer Vision. Berlin: Springer, 2022: 398-414.

[18] ZHU K, ZHAI W, CAO Y, et al. Self-sustaining representation expansion for non-exemplar class-incremental learning[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2022: 9286-9295.

[19] IOFFE S, SZEGEDY C. Batch normalization: accelerating deep network training by reducing internal covariate shift[C]//Proceedings of the 32nd International Conference on Machine Learning. Lille: PMLR, 2015, 37: 448-456.

[20] ORDÓÑEZ F J, ROGGEN D. Deep convolutional and LSTM recurrent neural networks for multimodal wearable activity recognition[J]. Sensors, 2016, 16(1): 115.

[21] QIAN N. On the momentum term in gradient descent learning algorithms[J]. Neural Networks, 1999, 12(1): 145-151.

[22] BROUSMICHE M, ROUAT J, DUPONT S. Multi-level attention fusion network for audio-visual event recognition[EB/OL]. (2021-06-12)[2023-06-01]. https://arxiv.org/abs/2106.06736v1.

[23] SHI Z S, LIANG J, LI Q Q, et al. Multi-modal multi-action video recognition[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE Press, 2021: 13658-13667.

Continual learning method based on differential feature distillation for multimodal network

HE Chiyuan，CHENG Shaoxu，XU Linfeng*，MENG Fanman，WU Qingbo

(School of Information and Communication Engineering，University of Electronic Science and Technology of China，Chengdu 611731，China)

Abstract：Continual learning has become a new research hotspot in recent years. However, in the continual learning of multimodal architecture, the data are generally not fully utilized, resulting in catastrophic forgetting and learning obstruction. To address these issues, a multimodal continual learning method based on feature distillation was proposed. By focusing on differences in task performance between different modalities, this method chose to retain more or less old knowledge of the modality, so as to stimulate each modality's potential in exploring discriminative features from an overall perspective. Experiments on the multimodal behavior recognition dataset UESTC-MMEA-CL validated the effectiveness of this method. When approaching the eighth task, the proposed method achieved an improved accuracy by an average of 22.0% and 20.1% based on fine-tuning and learning without forgetting (LwF), respectively. Compared with the classic knowledge distillation method, the proposed method better utilized sensor modalities, thereby significantly alleviating the catastrophic forgetting issue of multimodal networks.

Keywords：machine learning; continual learning; multimodality; behavior recognition; feature distillation

Received：2023-06-15；**Accepted**：2023-12-01；**Published Online**：2024-02-26 11：14

URL：link.cnki.net/urlid/11.2625.V.20240221.1715.001

Foundation items：National Science and Technology Major Project (2021ZD0112001); National Natural Science Foundation of China (62071086); Natural Science Foundation of Sichuan Province (2023NSFSC1972)

* **Corresponding author**. E-mail：lfxu@uestc.edu.cn

http://bhxb.buaa.edu.cn jbuaa@buaa.edu.cn
DOI: 10.13700/j.bh.1001-5965.2023.0380

基于多元优化的批量图像隐写载体选择方法

王仰光[1]，姚远志[2,*]，俞能海[1]

(1. 中国科学技术大学 网络空间安全学院，合肥 230027；2. 合肥工业大学 计算机与信息学院，合肥 230601)

摘 要：批量图像隐写通过载体选择在多张载体图像中嵌入秘密信息，为社交网络中的隐蔽通信提供有效的途径。与传统的图像隐写相比，批量图像隐写的关键科学问题是在保证抗检测性能的情况下设计有效的载体选择方法。基于此，提出一种基于多元优化的批量图像隐写载体选择方法，通过联合分析嵌入失真、图像相关性、嵌入容量三元因素，将批量图像隐写中的载体选择建模为多元优化问题。同时，针对社交网络可能对载密图像进行压缩的情况，在批量图像隐写载体选择时设计了秘密信息分片与重组策略，提升了批量图像隐写的鲁棒性。充分的实验结果表明：所提方法在抗检测性能、嵌入容量和鲁棒性上取得了令人满意的效果，为基于社交网络的隐蔽通信提供了技术支撑。

关 键 词：多元优化；批量图像隐写；载体选择；抗检测性；鲁棒性

中图分类号：TP309.2

文献标志码：A **文章编号**：1001-5965(2025)07-2468-10

图像隐写是信息隐藏研究领域的重要分支，旨在以数字图像为载体构建隐蔽通信信道，从而保证嵌入信息/数据安全传输。与传统的加密技术不同，隐蔽通信强调通信的隐蔽性，即不被攻击者察觉到秘密信息传输这一行为的发生，从而实现安全的通信。近年来数字图像的广泛传播为隐蔽通信提供了极大的便利，发送者可以从社交网络上获得大量的图像[1]，这些图像可以为隐蔽通信提供丰富载体。联合图像专家组(joint photographic experts group, JPEG)图像是社交网络上最流行的载体之一，为实施基于图像隐写的隐蔽通信提供了极大便利。

图像隐写的研究通常基于由美国学者 Simmons 在 1984 年提出的“囚徒问题”模型[2]。近年来，以单张 JPEG 图像为载体的隐写算法研究发展日益成熟[3-10]，其核心思路都是通过考虑 JPEG 图像的某些载体特征来辅助定义失真，再通过调用自适应隐写编码以追求整体失真最小，如校验网格编码[3]。在嵌入秘密信息时，合适的隐写代码可以最大限度地减少研究人员定义的失真函数，最新的研究集中在定义有效的失真函数上。已有的 JPEG 图像隐写失真函数有基于通用小波相对失真的 JPEG 图像隐写方法 (universal wavelet relative distortion for JPEG domains, J-UNIWARD)[4]、基于保持图像块间和块内系数差统计特性的 JPEG 图像隐写方法[5] 和最小化隐写检测概率的 JPEG 图像隐写方法[6]，它们的主要目的是将较大的失真分配给纹理复杂的区域，将较小的失真分配给平滑的区域。

上述以 JPEG 图像为载体的隐写方法仅考虑了单张 JPEG 图像，隐写的嵌入容量常常受到限制。如果在单张 JPEG 图像中嵌入大量秘密信息，很容易导致嵌入的秘密信息被检测出来，从而威胁隐写安全性。随着当前互联网及社交网络的飞速发展，隐写的候选载体图像规模十分庞大，批量图像隐写应运而生。在文献 [11-12] 中，Ker 证明了在使用多张载体图像进行隐写时，隐写安全性正比于图像数目的平方根，侧面验证了批量图像隐写的优势。

收稿日期：2023-06-16；**录用日期**：2023-09-02；**网络出版时间**：2023-11-08 12:03

网络出版地址：link.cnki.net/urlid/11.2625.V.20231107.1103.004

基金项目：国家重点研发计划基金 (2018YFB0804102)；国家自然科学基金 (61802357)；中央高校基本科研业务费专项资金 (WK3480000009)

* **通信作者**. E-mail：yaoyz@hfut.edu.cn

引用格式：王仰光，姚远志，俞能海. 基于多元优化的批量图像隐写载体选择方法 [J]. 北京航空航天大学学报，2025，51（7）：2468-2477.
WANG Y G，YAO Y Z，YU N H. Cover selection method for batch image steganography based on multivariate optimization [J]. Journal of Beijing University of Aeronautics and Astronautics，2025，51（7）：2468-2477 （in Chinese）.

现有的载体选择方法, 主要针对单张图像的某些固有属性。以 Liao 等[13] 的方法为例, 该方法将总秘密信息按图像纹理自适应地分配到各载体图像中。然而, 批量图像并非是单张图像的简单堆加, 图像之间的相关性也是进行载体选择时需要参考的因素。Wang 等[14] 验证了选择相似度高的图像进行隐写会导致载密图像集的抗检测性能下降, 进而提出一种基于图像相似度的载体选择方法。但是该方法仅考虑了图像嵌入失真和图像相似度, 忽略了嵌入容量在批量图像隐写中的重要性。此外, 在实际应用中, 社交网络可能对载密图像进行压缩, 因此, 载密图像压缩后的秘密信息恢复也是实际应用中批量图像隐写需要关注的重要问题, 文献 [15] 采用了重压缩方案来提高隐写的鲁棒性, 同时将嵌入域移到离散余弦变换 (discrete cosine transform, DCT) 系数的低频区域, 进而提高了自适应隐写的鲁棒性和安全性。文献 [16] 提出一种基于嵌入域选择和自适应纠错的抗有损 JPEG 压缩的鲁棒图像隐写方法, 通过自适应地选择嵌入域及自适应地调整纠错码的纠错能力, 提高了隐写的鲁棒性并降低对抗隐写分析性能的影响。

针对上述问题, 本文提出一种基于多元优化的批量图像隐写载体选择方法, 通过联合分析嵌入失真、图像相关性、嵌入容量三元因素, 将批量图像隐写中的载体选择建模为多元优化问题。同时, 针对社交网络可能对载密图像进行压缩的情况, 在批量图像隐写载体选择时设计了秘密信息分片与重组策略, 提升了批量图像隐写的鲁棒性。充分的实验结果表明, 本文提出的基于多元优化的批量图像隐写载体选择方法在抗检测性能、嵌入容量和鲁棒性上取得了令人满意的效果, 为基于社交网络的隐蔽通信提供技术支撑。

1　相关工作

1.1　载体选择方法

迄今为止, 学术界已经提出很多用于批量图像隐写的载体选择方法。对于多幅相同尺寸的图像, 文献 [17] 证明了安全隐写容量正比于图像数量的平方根。文献 [18] 提出的载体选择方法是基于经验知识设计的, 在抗检测性方面的效果有待加强。文献 [19] 采用深度神经网络, 通过学习各载体像素修改概率来联合载体选择与嵌入容量分配问题。文献 [13] 将总体待嵌入秘密信息按图像纹理自适应地分配到各载体图像中。文献 [20] 将载体数据集选择和载荷分配结合在一起考虑, 利用图像代价函数一阶导数的单调递增性, 得到载体数据集和载荷分配的最优化结果。文献 [21] 考虑隐写者持有的候选载体图像可能已经过某种预处理的情况, 如对比度增强、图像去噪等, 进而提出一种规避预处理的载体选择方法, 同时考虑预处理与嵌入操作引起的失真。然而, 批量图像并非是单张图像的简单堆加, 图像之间的相关性也是进行载体选择时需要参考的因素。文献 [14] 通过实验验证了选择相似度高的图像进行隐写会导致载密图像集的抗检测性能下降, 同时联合图像的嵌入失真, 提出一种基于图像相似度的批量图像隐写载体选择方法。但是该方法仅考虑了图像嵌入失真和图像相似度, 忽略了嵌入容量在批量图像隐写中的重要性。此外, 在实际应用中, 社交网络可能对载密图像进行压缩, 因此, 载密图像压缩后的秘密信息恢复也是实际应用中批量图像隐写需要关注的重要问题。

1.2　影响抗检测性的三元因素

根据现有的相关文献可以总结出, 嵌入失真、图像相关性、嵌入容量三元因素均会影响到批量图像隐写的抗检测性能。

文献 [22] 提供了一种计算图像嵌入失真的思路, 将嵌入失真与图像嵌入代价进行结合。计算图像相关性的方法有很多, 常见的有: 均方误差、峰值信噪比、结构相似性指数、相关系数、互信息等。Wang 等[14] 提出使用奇异值分解 (singular value decomposition, SVD) 来计算不同空域图像之间的相似度, 同时结合空域图像上隐写的特点, 创新性地提出使用小奇异值的变化来衡量不同图像之间的相似度。本文采用的单张图像隐写方法为 J-UNIWARD, 该方法是在 JPEG 图像的非零 DCT 系数上进行嵌入, 在 JPEG 图像隐写的抗检测性上表现优秀。

1.3　抗检测性分析方法

目前, 主流的批量图像隐写的抗检测性分析方法是利用载体图像集和载密图像集的特征进行有监督训练, 本文采用隐写分析领域广泛使用的集成分类器[23]。对不同载体选择方法获得的成对载体图像和载密图像进行特征提取(如 JPEG 图像富模型隐写分析, 特征维度 7 850 维 (cfstar)[23] 和特征维度 22 510 维 (ccJRM)[24]), 使用集成分类器进行训练和分类, 其流程示意图如图 1 所示。

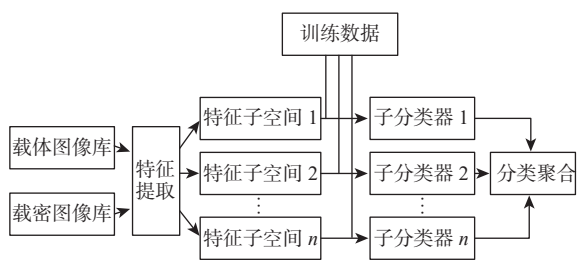


图 1　集成分类器流程示意图

Fig. 1　Flowchart of ensemble classifier process

与经典的分类方法支持向量机相比，集成分类器是一种更有效的隐写分析方法，可以提高检测准确率，改善鲁棒性和提高可解释性，从而实现更加准确和可靠的隐写分析。用错误检测概率P_{E}衡量不同载体选择方法的抗检测性，定义如下：

$$P_{\mathrm{E}}=\min_{P_{\mathrm{FA}}}\frac{1}{2}(P_{\mathrm{FA}}+P_{\mathrm{MD}}) \tag{1}$$

式中：P_{FA}为虚警率（将正常图像判断为载密图像的概率）；P_{MD}为漏检率（将载密图像判断为正常图像的概率）。P_{E}越大表示抗检测性能越好，通常$0\leqslant P_{\mathrm{E}}\leqslant 0.5$，$P_{\mathrm{E}}=0.5$表示分类器完全随机猜测，即完全无法区分正常图像与载密图像。

2 基于多元优化的载体选择方法

本文提出的基于多元优化的批量图像隐写载体选择方法如图2所示，通过联合分析嵌入失真、图像相关性、嵌入容量三元因素，将批量图像隐写中的载体选择建模为多元优化问题。

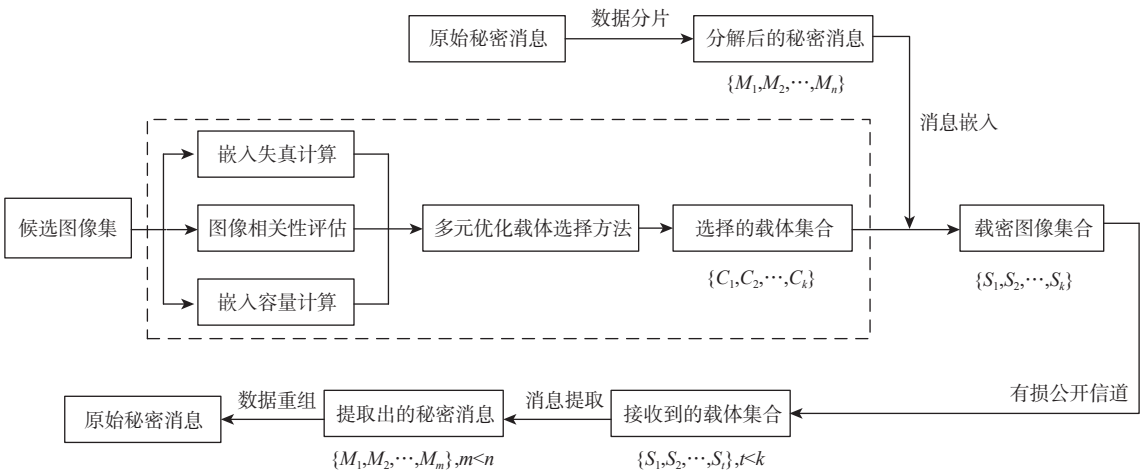


图2 基于多元优化的批量图像隐写载体选择方法

Fig. 2 Cover selection method for batch image steganography based on multivariate optimization

2.1 秘密信息预处理

批量图像隐写中，隐写者是将待嵌入秘密信息嵌入到指定的载体图像集中，在秘密信息提取阶段，接收方需要接收到所有隐写后的载密图像，才可以把秘密信息完整地提取出来。然而，社交网络并不是完全可靠的，载密图像集在社交网络上有损公开信道的传输过程中可能会面临图像丢失、图像压缩等问题，导致嵌入的秘密信息无法被正常提取恢复。

为解决上述问题，本文提出秘密信息的分片与重组策略，基于非方阵范德蒙矩阵的特殊性质，将范德蒙矩阵与秘密信息相结合，将秘密信息分片后嵌入到不同的载体图像中，这样即使在传输过程中部分载密图像被破坏，仍可以从剩余部分的载密图像中恢复出完整的秘密信息，无论待嵌入的是什么格式的秘密信息，都可以将其视为二进制流，因此，不同格式的秘密信息均可以按照下述方法进行分片。

1）将待嵌入秘密信息序列分成若干块，每块包括L_1位信息，然后，分别将各块转化为q进制的序列（q为奇素数），转换后每块包括L_2位：

$$L_1=\left\lfloor L_2\log_2 q\right\rfloor \tag{2}$$

2）将转换后的q进制数字序列分为K块，其中，每块包括m位，分别是$\{d_{k,1},d_{k,2},\cdots,d_{k,m}\}$，$k\in[1,K]$。

3）构造范德蒙矩阵$\boldsymbol{A}$：

$$\boldsymbol{A}=\begin{bmatrix}1 & 1 & \cdots & 1\\ a_1 & a_2 & \cdots & a_n\\ \vdots & \vdots & & \vdots\\ a_1^{m-1} & a_2^{m-1} & \cdots & a_n^{m-1}\end{bmatrix}\bmod q \tag{3}$$

式中：$a_1,a_2,\cdots,a_n\in[0,q-1]$，确保彼此各不相同，且满足条件$m\leqslant n\leqslant q$。

4）数据分片。如式(4)所示，此时m份的秘密信息被扩展成了n份，这n份秘密信息分别对应范德蒙矩阵的参数$a_1,a_2,\cdots,a_n\in[0,q-1]$，将二者联合起来，当作新的秘密信息分片进行嵌入。

$$[t_{k,1},t_{k,2},\cdots,t_{k,n}]=[d_{k,1},d_{k,2},\cdots,d_{k,m}]\cdot\boldsymbol{A} \tag{4}$$

5）数据重组。当接收方接收到大于等于m份秘密信息分片后（此处以m为例），根据所获得的m份秘密信息分片$[t'_{k,1},t'_{k,2},\cdots,t'_{k,m}]$及对应的范德蒙矩阵的参数$a'_i$，可以重构范德蒙矩阵$\boldsymbol{A}_1$，参考式(3)。然后，将接收到的$m$份秘密信息乘以新构造的范德蒙矩阵$\boldsymbol{A}_1$的逆矩阵（广义逆矩阵），即可恢复出原始秘密信息，如下：

$$[d_{k,1},d_{k,2},\cdots,d_{k,m}]=[t'_{k,1},t'_{k,2},\cdots,t'_{k,m}]\cdot\boldsymbol{A}_1^{-1} \tag{5}$$

2.2　三元因素的计算

2.2.1　图像嵌入失真计算

受文献 [22] 的启发, 提出一种衡量图像嵌入失真的思路, 把图像的嵌入失真和嵌入代价联系起来。首先, 根据一维低通小波分解滤波器$\boldsymbol{h}$和一维高通小波分解滤波器$\boldsymbol{g}$, 可以构建所需要的 3 个线性平移不变滤波器, 它们的核用$B=\{\boldsymbol{K}^{(1)},\boldsymbol{K}^{(2)},\boldsymbol{K}^{(3)}\}$表示, 如式 (6) 所示, 然后, 使用二维 LH(low-high, 包含了图像的水平边缘信息)、HL(high-low, 包含了图像的垂直边缘信息)、HH(high-high, 包含了图像对角线方向的高频信息)小波方向的高通滤波器计算出载体图像$\boldsymbol{X}_i(1\leqslant i\leqslant n)$在 LH、HL 和 HH 方向分解的残差。

$$\begin{cases}\boldsymbol{K}^{(1)}=\boldsymbol{h}\cdot\boldsymbol{g}^{\mathrm{T}}\\ \boldsymbol{K}^{(2)}=\boldsymbol{g}\cdot\boldsymbol{h}^{\mathrm{T}}\\ \boldsymbol{K}^{(3)}=\boldsymbol{g}\cdot\boldsymbol{g}^{\mathrm{T}}\end{cases}\tag{6}$$

给定 1 组载体图像$\boldsymbol{X}_i$和嵌入数据后对应的载密图像$\boldsymbol{Y}_i$, 嵌入代价ω_i为所有小波系数$W_{uv}^{(k)}(\boldsymbol{Y}_i)$相对于载体图像$\boldsymbol{X}_i$的小波系数$W_{uv}^{(k)}(\boldsymbol{X}_i)$相对变化之和, 如式 (7) 所示, 其中, σ是一个稳定数值计算的常数, ω_i越小代表该图像的嵌入代价越小, 在嵌入数据方面具有更高的优先级, 即嵌入失真越小。图 3 是 BOSSbase[25] 数据库里的 10 000 张图像经过计算后的嵌入失真分布图(质量因子(quality factor, QF)=75)。为便后续讨论, 嵌入失真被归一化到区间 0～1 之间。

$$\omega_i\triangleq\sum_{k=1}^{3}\sum_{u=1}^{n_1}\sum_{v=1}^{n_2}\frac{\left|W_{uv}^{(k)}(\boldsymbol{X}_i)-W_{uv}^{(k)}(\boldsymbol{Y}_i)\right|}{\sigma+\left|W_{uv}^{(k)}(\boldsymbol{X}_i)\right|}\tag{7}$$

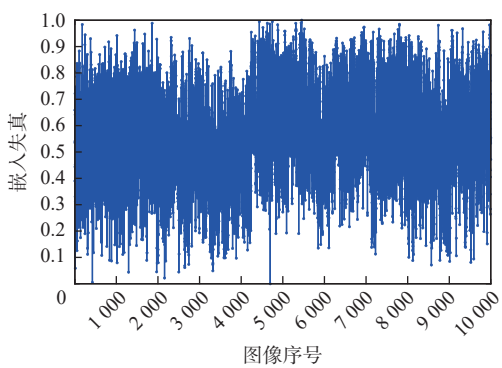


图 3　10 000 张 JPEG 图像的嵌入失真分布

Fig. 3　Embedding distortion for 10 000 JPEG images

2.2.2　图像相关性评估

由于批量图像不是单张载体图像的简单堆加, 因此, 设计批量图像隐写的载体选择方法时, 需要关注不同图像之间的相关性, 如图像之间的相似度。当载体图像相似时, 会为隐写分析提供大量相关样本, 这不利于隐写的抗检测性能。文献 [14] 提出相似度过高的图像不能同时用来进行隐写, 否则抗检测性会降低, 并提出在空域图像上利用小奇异值(而不是大奇异值)进行载体选择的方法。该方法更适合空域图像隐写术的特性。JPEG 图像与空域图像有所不同, 但该方法仍具有参考价值。

本文设计的图像相似度计算方法如下: 假设图像为$\boldsymbol{X}$, 尺寸为$r_1\times r_2$, 将图像的矩阵表示进行 SVD 分解, 如下:

$$\boldsymbol{X}=\boldsymbol{U}\cdot\boldsymbol{D}\cdot\boldsymbol{V}^{\mathrm{T}}\tag{8}$$

式中: $\boldsymbol{U}$的尺寸为$r_1\times r_1$; $\boldsymbol{V}$的尺寸为$r_2\times r_2$; $\boldsymbol{D}$为$\boldsymbol{X}$的奇异值矩阵。给定载体图像$\boldsymbol{C}$及对应的载密图像$\boldsymbol{S}$可以得到二者的奇异值矩阵$\boldsymbol{D}_{\boldsymbol{C}}$和$\boldsymbol{D}_{\boldsymbol{S}}$, 然后, 采用式 (9) 方法衡量每个奇异值在嵌入前后的变化, 结果如图 4 所示。

$$\Delta=|\boldsymbol{D}_{\boldsymbol{C}}-\boldsymbol{D}_{\boldsymbol{S}}|/\boldsymbol{D}_{\boldsymbol{C}}\tag{9}$$

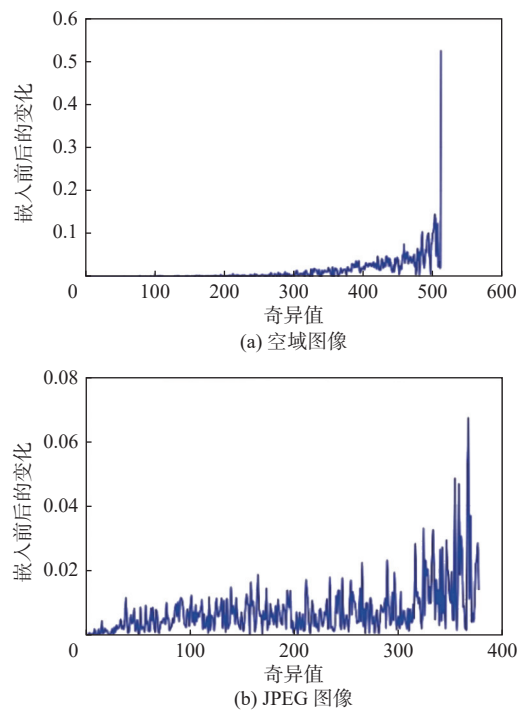


图 4　空域图像和 JPEG 图像隐写后的 SVD 变化幅度展示

Fig. 4　Change in SVD of spatial domain and JPEG images after steganography

从图 4 可以看出, 在 JPEG 图像上进行隐写嵌入后, 不同大小的 SVD 值均有改变, 这点与空域图像不同, 因此, 在考虑图像相似度时, 所有奇异值均需考虑在内。定义D_{dif}为不同图像间奇异值的差距, 计算方法如下:

$$D_{\mathrm{dif}}=\sqrt{\sum_{i=1}^{r}[\boldsymbol{D}_{\boldsymbol{C}}(i)-\boldsymbol{D}_{\boldsymbol{S}}(i)]^2}\tag{10}$$

式中: $r=\min\{\mathrm{rank}(\boldsymbol{D}_{\boldsymbol{C}}),\mathrm{rank}(\boldsymbol{D}_{\boldsymbol{S}})\}$。

图像$\boldsymbol{C}$和图像$\boldsymbol{S}$的图像相似度定义如下:

$$\Phi(\boldsymbol{C},\boldsymbol{S})=\frac{1}{D_{\mathrm{dif}}+\varepsilon} \tag{11}$$

式中：ε为稳定计算的常数。

2.2.3 嵌入容量计算

本文采用的嵌入方法为适用于JPEG图像上的J-UNIWARD，该方法是在JPEG图像的非零DCT系数上进行信息嵌入，在JPEG图像隐写的抗检测性上表现优秀，图像$\boldsymbol{X}_i$的嵌入容量E_i计算方法如式(12)所示。在真实场景下选择载体时，嵌入容量也是一个需要考虑的因素。图5是BOSSbase[25]数据库里的10 000张图像经过计算后得到的嵌入容量分布图(QF=75)。

$$E_i=\sum_{r=1}^{r_1}\sum_{c=1}^{r_2}\mathrm{sign}(r,c) \tag{12}$$

式中：sign(·)表示函数，定义如下：

$$\mathrm{sign}(x)=\begin{cases}1 & C_{\mathrm{DCT}}(r,c)>0\\ 0 & C_{\mathrm{DCT}}(r,c)=0\end{cases} \tag{13}$$

其中：C_{DCT}是图像$\boldsymbol{X}_i$的DCT系数。

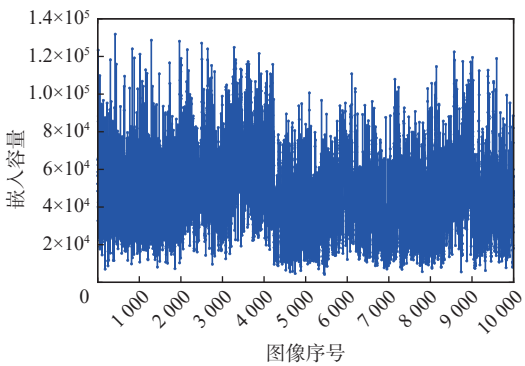


图5 10 000张JPEG图像的嵌入容量分布

Fig. 5 Embedding capacity for 10 000 JPEG images

2.3 实施步骤

2.3.1 单因素下的抗检测性分析

为验证嵌入失真、图像相关性和嵌入容量对批量图像隐写抗检测性的影响，设计了3种单因素的载体选择方法。

1) 基于嵌入失真的载体选择方法(Dis)

计算出候选载体图像库中所有图片的嵌入失真，将其进行升序排序，从小到大依次选择载体图像。

2) 基于图像相似度的载体选择方法(IS)

将候选载体图像集记为Dataset，需要载体图像集的大小为l_{total}，将已选图像集记为Select，大小为l_1，剩余图像集记为Rest，大小为l_2。

初始化：Select图像集为空，$l_1=0$；Rest图像集等于Dataset图像集。

步骤1 计算Rest图像集中每张图像与其他图像的总相似度之和，选择总相似度之和最小的图像当作第1张图像；更新Select图像集和Rest图像集；

步骤2 计算Rest图像集中每张图像的相对总相似度，方法如下：

$$T_{\mathrm{IS}}(i)=\sum_{j=1}^{l_1}\Phi(i,j)\qquad 1\leqslant i\leqslant l_2 \tag{14}$$

将其进行升序排列，选择相对总相似度最小的图片，更新Select图像集和Rest图像集；

步骤3 重复步骤2，直到$l_1=l_{\mathrm{total}}$。

3) 基于嵌入容量的载体选择方法(EC)

计算出候选载体图像库中所有图片的嵌入容量，将其进行降序排序，从大到小依次选择载体图像。

抗检测性分析实验设置如下：

步骤1 将BOSSbase[25]数据库中的所有图像压缩为QF=75的候选JPEG图像集，记为Dataset。

步骤2 根据不同的单因素载体选择方法，选择k(k=1 000, 2000, 3 000, 4 000, 5 000)张图像当作所选择的载体图像集Cover_select；作为对比，使用MATLAB自带的伪随机函数，从Dataset随机选择出相同数目的载体图像集，记为Cover_Rand。

步骤3 使用J-UNIWARD对所选择的载体图像集Cover_select和Cover_Rand进行模拟嵌入，嵌入容量为0.5，得到对应的载密图像集Stego_select和Stego_Rand。

步骤4 使用ccJRM提取2种载体选择方法对应的载体载密对的特征，采用集成分类器进行抗检测性分析，得到错误检测概率P_{E}。

步骤5 汇总P_{E}关于k的表格数据，结果如表1所示。可以看出，相比于随机载体选择方法，以嵌入失真、图像相关性和嵌入容量作为衡量指标的载体选择方法在抗检测性上更具优势，在相同的条件下，具有更高的抗检测性。

表1 单因素下的抗检测性分析

Table 1 Undetectability analysis under single-factor conditions

方法	P_{E}				
	k=1 000	k=2 000	k=3 000	k=4 000	k=5 000
Rand	0.256 4	0.226 6	0.212 2	0.201 2	0.197 4
Dis	0.341 0	0.306 5	0.287 2	0.267 9	0.252 0
IS	0.344 2	0.304 9	0.278 6	0.269 0	0.252 0
EC	0.352 2	0.316 3	0.290 0	0.272 9	0.255 3

2.3.2 载体选择方法

综上所述，影响批量图像隐写抗检测性能的因素有嵌入失真、图像相关性和嵌入容量，考虑到嵌入失真和嵌入容量为单张图像的固有属性，而图像

相似度为图像之间的参考属性。因此, 设计图像固有属性指标H_i, 计算方法如下:

$$H_i = \alpha D_i + \beta E_i^{-1} \tag{15}$$

式中: $\alpha+\beta=1$, α 和 β 分别为嵌入失真和嵌入容量的影响权重, 二者的具体计算方法在后续章节中将进行详细阐述。

假设候选载体图像集为$\{\boldsymbol{X}_1,\boldsymbol{X}_2,\cdots,\boldsymbol{X}_n\}$, 其对应的嵌入失真为$\{D_1,D_2,\cdots,D_n\}$, 对应的嵌入容量为$\{E_1,E_2,\cdots,E_n\}$, 计算出所有图像的固有属性指标$\{H_1,H_2,\cdots,H_n\}$, 将其从小到大重新排列, 得到新的下标顺序, 记为$\sigma_1,\sigma_2,\cdots,\sigma_n$, 满足$H_{\sigma(1)}\leqslant H_{\sigma(2)}\leqslant\cdots\leqslant H_{\sigma(n)}$。选取固有属性指标 H 最小的图像$\sigma(1)$作为所选择的第 1 张载体图像。

对于图像$\boldsymbol{X}_i$和$\boldsymbol{X}_j$, 定义$f(\boldsymbol{X}_j/\boldsymbol{X}_i)$表示在已选择图像$\boldsymbol{X}_i$的情况下, 选择图像$\boldsymbol{X}_j$作为下一张载体图像的代价, 计算方法如式 (16) 所示。其逻辑为, $f(\boldsymbol{X}_j/\boldsymbol{X}_i)$应该与图像相似度$\Phi(i,j)$成比例, 因为相似的图像不适合隐写术。同时, 固有属性指标H_i对隐写术也很重要。因此, $f(\boldsymbol{X}_j/\boldsymbol{X}_i)$被定义为图像相似度和固有属性指标的乘积:

$$f(\boldsymbol{X}_j/\boldsymbol{X}_i) = \Phi(i,j)\cdot(\alpha D_j + \beta E_j^{-1}) \tag{16}$$

假如已选图像$\boldsymbol{X}_i$不止 1 张, 定义图像$\boldsymbol{X}_j$的待选总代价为$F(\boldsymbol{X}_j)$, 计算方法如下:

$$F(\boldsymbol{X}_j) = \sum_i f(\boldsymbol{X}_j/\boldsymbol{X}_i) \tag{17}$$

经过计算后, 选择待选总代价$F(\boldsymbol{X}_j)$最小的图片为下 1 张载体图像, 更新相关参数, 重复上述过程, 从剩余的$n-1$张图像中选择出其他载体图像。

具体载体选择方法如下:

假设候选载体图像集为 Dataset, 所需载体图像集大小为 l_{total}, 已选图像集记为 Select, 大小为 l_1; 剩余图像集记为 Rest, 大小为 l_2。

步骤 1　初始化 Select 图像集为空, l_1=0; Rest 图像集等于 Dataset 图像集。

步骤 2　计算 Rest 图像集中所有图像的固有属性指标H_i, 选择H_i最小的图像作为第 1 张图像, 更新 Select 图像集和 Rest 图像集。

步骤 3　计算 Rest 图像集中每张图像的待选总代价F, 将F进行升序排列, 选择待选总代价最小的图片, 更新 Select 图像集和 Rest 图像集。

步骤 4　重复步骤 3, 直到 $l_1=l_{\text{total}}$。

2.3.3　参数优化

在式 (15) 中, α、β 通过以下方法计算不同取值下的最大P_{E}来确认:

$$[\alpha,\beta] = \max P_{\text{E}} \qquad \alpha+\beta=1 \tag{18}$$

步骤 1　初始化$\alpha\leftarrow 0, \beta\leftarrow 1-\alpha$。

步骤 2　根据上述载体选择方法, 从 BOSSbase[25] 数据库中选择 3 000 张图像当作载体图像集。

步骤 3　使用 J-UNIWARD 进行模拟嵌入, 嵌入容量为 0.5, 得到载密图像集。

步骤 4　使用 ccJRM 提取对应的特征对, 采用集成分类器进行抗检测性分析, 得到错误检测概率P_{E}并记录。

步骤 5　若$\alpha\leqslant 1$, 则$\alpha\leftarrow\alpha+0.01, \beta\leftarrow 1-\alpha$, 回到步骤 2 重复执行; 若$\alpha=1$, 则执行步骤 6。

步骤 6　画出P_{E}关于α和β的函数图像, 确定α、β 的具体数值。

实验结果如图 6 所示, 可以看出, 当$\alpha=0.93$、$\beta=0.07$时, P_{E}最大, 如图中标注的极值点, 后续实验也将以此数值进行验证。

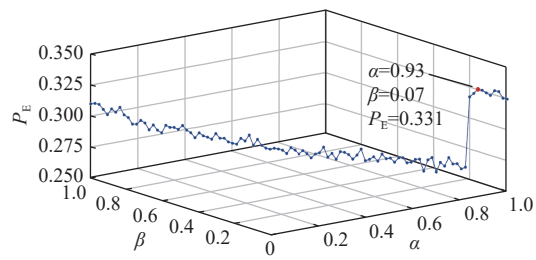


图 6　不同权值下的错误检测概率分布

Fig. 6　Probability distributions of false detections for different weights

3　实　验

为验证本文方法的有效性, 进行如下实验。

3.1　环境设置

采用 Windows 10 专业版 21H2 操作系统, CPU 为 AMD R5 3600, 3.95 GHz, 内存为 16 GB, 软件采用 MATLAB R2021a。

为增加实验的可信度, 本实验选取图像处理中常用的数据集 BOSSbase[25] 里的全部图片, 图像尺寸均为 512 像素×512 像素, 图像总数为 10 000 张。实验开始前, 使用不同的 QF(60、75、90)对原始数据集 BOSSbase[25] 里的图像进行 JPEG 压缩, 得到原始载体图像库 cover60、cover75 和 cover90, 之后在每个原始载体图像库上的操作都完全相同, 后续步骤以 cover75 为样例载体图像库 Dataset 进行描述。

为使实验结果更具说服力, 采用 4 种载体选择方法, 分别为:

1) 随机载体选择方法(记为 Rand)

使用 MATLAB 自带的伪随机函数, 从 10 000 张图像中随机选择指定数目的载体图像组成载体图像集。

2) Liao 等的方法[13](记为 Liao)

将总秘密信息按图像纹理特征自适应地分配到各载体图像中。

3) Wang 等的方法[14](记为 Wang)

将图像的嵌入失真和图像相似度相结合,进行载体选择。

4) 本文方法

分别使用上述 4 种载体选择方法,从载体数据库中选择数目为 k 的载体图像集(k=1 000, 2 000, 3 000, 4 000, 5 000),使用 J-UNIWARD 进行秘密信息嵌入,嵌入容量分别为 0.4 和 0.5,从而得到共 5×2=10 组载体/载密图像集。

得到载体/载密图像集后,分别使用 cfstar 和 ccJRM 提取对应的载体/载密图像集的特征,使用集成分类器对其抗检测性进行评价,得到 P_E 值。为了提高 P_E 的准确性,每轮隐写分析均重复进行 10 次,用 10 次随机实验得到的抗检测性 P_E 的均值作为最终结果,如表 2～表 4 所示。

表 2 不同条件下的错误检测概率(QF=60)

Table 2 Probability of false detections under different conditions (QF = 60)

嵌入率	特征集	方法	P_E				
			k=1 000	k=2 000	k=3 000	k=4 000	k=5 000
0.4	cfstar	Rand	0.305 8	0.283 5	0.273 2	0.260 0	0.260 0
		Liao[13]	0.369 6	0.331 1	0.309 5	0.295 4	0.280 1
		Wang[14]	0.376 0	0.347 8	0.323 9	0.304 9	0.298 9
		本文方法	0.380 2	0.348 2	0.328 5	0.305 3	0.296 6
	ccJRM	Rand	0.290 0	0.268 6	0.258 5	0.249 0	0.248 8
		Liao[13]	0.350 8	0.325 3	0.300 9	0.286 5	0.267 7
		Wang[14]	0.359 8	0.336 0	0.313 8	0.298 5	0.288 2
		本文方法	0.363 0	0.338 1	0.316 8	0.298 9	0.289 5
0.5	cfstar	Rand	0.250 2	0.220 3	0.202 3	0.190 5	0.188 5
		Liao[13]	0.311 5	0.263 0	0.238 8	0.224 1	0.208 5
		Wang[14]	0.316 7	0.280 3	0.254 0	0.235 9	0.224 4
		本文方法	0.322 1	0.286 9	0.255 7	0.235 3	0.223 6
	ccJRM	Rand	0.229 8	0.202 2	0.188 7	0.174 7	0.172 5
		Liao[13]	0.284 0	0.248 5	0.230 0	0.210 8	0.191 3
		Wang[14]	0.293 6	0.264 0	0.240 6	0.221 3	0.202 2
		本文方法	0.299 0	0.267 5	0.243 0	0.219 9	0.205 5

3.2 实验分析

3.2.1 抗检测性分析

从表 2～表 4 可以看出,与随机载体选择方法相比,另外 3 种载体选择方法对批量图像隐写的抗检测性均有不同程度的提升,证明要实现安全的批量图像隐写,载体选择是不容忽视的问题。对比 Liao 等[13] 和 Wang 等[14] 的方法,本文方法在抗检测性方面更具优势,验证了本文方法的有效性。

表 3 不同条件下的错误检测概率(QF=75)

Table 3 Probability of false detections under different conditions (QF = 75)

嵌入率	特征集	方法	P_E				
			k=1 000	k=2 000	k=3 000	k=4 000	k=5 000
0.4	cfstar	Rand	0.328 2	0.306 7	0.297 1	0.286 6	0.283 4
		Liao[13]	0.420 3	0.387 9	0.366 5	0.356 6	0.337 3
		Wang[14]	0.430 7	0.394 8	0.378 1	0.363 2	0.351 5
		本文方法	0.427 5	0.397 5	0.384 7	0.364 8	0.352 5
	ccJRM	Rand	0.312 8	0.287 5	0.281 3	0.271 3	0.268 9
		Liao[13]	0.400 2	0.376 4	0.354 2	0.336 4	0.323 4
		Wang[14]	0.403 6	0.383 1	0.363 9	0.346 3	0.335 8
		本文方法	0.407 1	0.384 6	0.365 5	0.351 2	0.335 3
0.5	cfstar	Rand	0.280 0	0.252 1	0.235 8	0.222 0	0.215 9
		Liao[13]	0.380 0	0.337 9	0.312 7	0.289 6	0.277 5
		Wang[14]	0.384 7	0.347 4	0.324 4	0.303 1	0.282 6
		本文方法	0.394 3	0.348 5	0.328 5	0.303 2	0.289 0
	ccJRM	Rand	0.256 4	0.226 6	0.212 2	0.201 2	0.197 4
		Liao[13]	0.341 0	0.306 5	0.287 2	0.267 9	0.252 0
		Wang[14]	0.351 0	0.317 4	0.291 9	0.275 4	0.266 0
		本文方法	0.353 3	0.320 3	0.293 3	0.277 7	0.266 1

表 4 不同条件下的错误检测概率(QF=90)

Table 4 Probability of false detections under different conditions (QF = 90)

嵌入率	特征集	方法	P_E				
			k=1 000	k=2 000	k=3 000	k=4 000	k=5 000
0.4	cfstar	Rand	0.324 9	0.313 2	0.309 8	0.301 6	0.299 4
		Liao[13]	0.401 3	0.393 8	0.384 7	0.381 0	0.374 1
		Wang[14]	0.413 4	0.406 0	0.396 3	0.382 7	0.370 0
		本文方法	0.421 3	0.410 5	0.404 7	0.391 3	0.378 4
	ccJRM	Rand	0.316 8	0.298 3	0.291 9	0.287 1	0.284 6
		Liao[13]	0.400 9	0.391 7	0.377 8	0.370 2	0.361 2
		Wang[14]	0.414 6	0.395 6	0.384 1	0.376 9	0.362 0
		本文方法	0.419 5	0.400 4	0.386 5	0.378 0	0.363 9
0.5	cfstar	Rand	0.280 4	0.259 9	0.252 1	0.240 7	0.237 3
		Liao[13]	0.366 0	0.357 6	0.340 6	0.332 0	0.317 5
		Wang[14]	0.375 8	0.365 5	0.352 9	0.332 6	0.316 5
		本文方法	0.382 4	0.366 0	0.355 1	0.335 5	0.321 0
	ccJRM	Rand	0.267 1	0.246 9	0.237 0	0.229 0	0.227 7
		Liao[13]	0.364 8	0.346 7	0.327 5	0.318 2	0.305 6
		Wang[14]	0.366 2	0.353 5	0.339 6	0.325 5	0.305 3
		本文方法	0.373 5	0.356 8	0.339 6	0.328 5	0.315 9

当选择的载体图像集规模较小时,批量图像隐写的抗检测性较好,随着选择的载体图像集规模变大,抗检测性开始下降。这符合理论分析,因为选

择的载体图像集规模越小, 越可以体现批量图像隐写的优势。

3.2.2　嵌入容量分析

Wang 等的方法[14] 与本文方法嵌入容量的对比如表 5 所示。可以看出, 本文方法相比于 Wang 等的方法[14], 在嵌入容量上有不同程度的提升, 更符合实际应用的需求。

3.2.3　鲁棒性分析

利用 2.1 节中的秘密信息预处理方法, 测试嵌入方法为 J-UNIWARD, 嵌入率分别为 0.15 和 0.25, 2 组测试参数分别为 m=5, n=9, q=11 和 m=5, n=31, q=37。2 组参数所对应的数据冗余率分别为 (9−5)/9=44.44% 和 (31−5)/31=83.87%, 各自独立重复进行 1 000 次实验, 模拟社交网络上载密图像传输过程中遭受压缩处理导致的数据丢失。秘密信息恢复的正确率定义为恢复原始数据的次数除以总测试次数。表 6 为不同平均数据丢失率下的秘密信息恢复正确率。实验发现, 只要数据的平均丢失率小于数据冗余率, 就可以成功恢复秘密信息。

表 5　Wang 等的方法[14] 和本文方法嵌入容量的对比

Table 5　Comparison of embedding capacity between Wang et al.'s method[14] and proposed method

QF	方法	嵌入容量/bit				
		k=1 000	k=2 000	k=3 000	k=4 000	k=5 000
60	Wang[14]	69 254 497	122 784 610	168 444 548	207 175 886	235 671 773
	本文方法	69 264 640	122 781 191	168 444 548	207 293 584	235 880 610
75	Wang[14]	88 527 118	157 237 348	216 089 517	267 350 520	309 605 320
	本文方法	88 540 196	157 251 458	216 143 376	267 350 520	309 758 006
90	Wang[14]	121 091 974	221 886 393	310 810 927	389 066 645	461 304 521
	本文方法	121 203 657	222 234 100	311 097 664	389 620 084	462 032 474

表 6　不同平均数据丢失率下的秘密信息恢复正确率

Table 6　Accuracy of secret information recovery under different average data loss rates　%

数据冗余率	平均数据丢失率	恢复正确率	
		嵌入率为0.15	嵌入率为0.25
44.44	33.49	72	65
	51.72	0	0
83.87	81.27	56	43
	91.24	0	0

3.2.4　时间复杂度分析

在上述实验环境配置下, 4 种载体选择方法的时间复杂度如表 7 所示。相比于 Liao 等[13] 和 Wang 等[14] 的方法, 本文方法的时间复杂度略高, 这主要是因为本文方法涉及了嵌入容量的计算。但相比于 Wang 等的方法[14], 本文方法的时间复杂度相差不大, 同时, 批量图像隐写的主要目的是保证秘密信息传递的隐蔽性, 抗检测性更加重要, 因此, 本文方法的时间复杂度是可以接受的。

表 7　4 种载体选择方法的时间复杂度

Table 7　Time complexity of four cover selection methods　s

方法	时间复杂度
Rand	0.001 5
Liao等[13]	6 244.25
Wang等[14]	10 687.36
本文方法	10 836.71

4　结　论

批量图像隐写通过载体选择在多张载体图像中嵌入秘密信息, 将秘密信息通过不可察觉的方式进行传输, 为社交网络中的隐蔽通信提供有效的途径, 主要结论如下:

1) 本文聚焦于批量图像隐写中十分关键的载体选择问题, 通过联合分析嵌入失真、图像相关性、嵌入容量三元因素, 将批量图像隐写中的载体选择建模为多元优化问题。

2) 为解决社交网络可能对载密图像进行压缩的情况, 进一步设计了秘密信息的分片与重组策略, 提升了批量图像隐写的鲁棒性。

3) 充分的实验结果表明, 本文方法在抗检测性能和鲁棒性上取得了令人满意的效果。

本文方法中, 关于嵌入失真、图像相关性和嵌入容量对批量图像隐写抗检测性的影响分析, 是启发式的定义, 并没有在理论上证明三者之间的具体关系。在今后的研究中, 需要定量地分析三者之间的关系。

参考文献 (References)

[1] SALMAN A. Facebook by the numbers[EB/OL]. (2022-12-08)[2023-04-13]. http://omnicoreagency.com/-facebook-statistics/.

[2] SIMMONS G J. The prisoners' problem and the subliminal channel[C]//Advances in Cryptology. Berlin: Springer, 1984: 51-67.

[3] FILLER T, JUDAS J, FRIDRICH J. Minimizing additive distortion in steganography using syndrome-trellis codes[J]. IEEE Transactions on Information Forensics and Security, 2011, 6(3): 920-935.

[4] HOLUB V, FRIDRICH J. Digital image steganography using universal distortion[C]//Proceedings of the First ACM Workshop on Information Hiding and Multimedia Security. New York: ACM, 2013: 59-68.

[5] GUO L J, NI J Q, SHI Y Q. Uniform embedding for efficient JPEG steganography[J]. IEEE Transactions on Information Forensics and Security, 2014, 9(5): 814-825.

[6] COGRANNE R, GIBOULOT Q, BAS P. Steganography by minimizing statistical detectability: the cases of JPEG and color images[C]//Proceedings of the ACM Workshop on Information Hiding and Multimedia Security. New York: ACM, 2020: 161-167.

[7] PAN Y F, NI J Q, SU W K. Improved uniform embedding for efficient JPEG steganography[C]//Cloud Computing and Security. Berlin: Springer, 2016: 125-133.

[8] WANG Z C, ZHANG X P, YIN Z X. Hybrid distortion function for JPEG steganography[J]. Journal of Electronic Imaging, 2016, 25: 050501.

[9] WEI Q D, YIN Z X, WANG Z C, et al. Distortion function based on residual blocks for JPEG steganography[J]. Multimedia Tools and Applications, 2018, 77(14): 17875-17888.

[10] 张卫明, 王宏霞, 李斌, 等. 多媒体隐写研究进展[J]. 中国图象图形学报, 2022, 27(6): 1918-1943.
ZHANG W M, WANG H X, LI B, et al. Overview of steganography on multimedia[J]. Journal of Image and Graphics, 2022, 27(6): 1918-1943(in Chinese).

[11] KER A D. Batch steganography and pooled steganalysis[C]//Information Hiding. Berlin: Springer, 2007: 265-281.

[12] KER A D. Perturbation hiding and the batch steganography problem[C]//Information Hiding. Berlin: Springer, 2008: 45-59.

[13] LIAO X, YIN J J, CHEN M L, et al. Adaptive payload distribution in multiple images steganography based on image texture features[J]. IEEE Transactions on Dependable and Secure Computing, 2022, 19(2): 897-911.

[14] WANG Z C, FENG G R, SHEN L Q, et al. Cover selection for steganography using image similarity[J]. IEEE Transactions on Dependable and Secure Computing, 2023, 20(3): 2328-2340.

[15] YIN Z X, KE L F. Robust adaptive steganography based on dither modulation and modification with re-compression[J]. IEEE Transactions on Signal and Information Processing Over Networks, 2021, 7: 336-345.

[16] DUAN X L, LI B, YIN Z X, et al. Robust image steganography against lossy JPEG compression based on embedding domain selection and adaptive error correction[J]. Expert Systems with Applications, 2023, 229: 120416.

[17] KER A D. A capacity result for batch steganography[J]. IEEE Signal Processing Letters, 2007, 14(8): 525-528.

[18] EVSUTIN O, KOKURINA A, MESHCHERYAKOV R. Approach to the selection of the best cover image for information embedding in JPEG images based on the principles of the optimality[J]. Journal of Decision Systems, 2018, 27(S1): 256-264.

[19] ZHONG N, QIAN Z X, WANG Z C, et al. Batch steganography via generative network[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2021, 31(1): 88-97.

[20] WANG Z C, ZHANG X P, QIAN Z X. Practical cover selection for steganography[J]. IEEE Signal Processing Letters, 2019, 27: 71-75.

[21] WANG Z C, ZHANG X P, YIN Z X. Joint cover-selection and payload-allocation by steganographic distortion optimization[J]. IEEE Signal Processing Letters, 2018, 25(10): 1530-1534.

[22] HOLUB V, FRIDRICH J, DENEMARK T. Universal distortion function for steganography in an arbitrary domain[J]. EURASIP Journal on Information Security, 2014, 2014(1): 1.

[23] KODOVSKY J, FRIDRICH J, HOLUB V. Ensemble classifiers for steganalysis of digital media[J]. IEEE Transactions on Information Forensics and Security, 2012, 7(2): 432-444.

[24] KODOVSKÝ J, FRIDRICH J. Steganalysis of JPEG images using rich models[C]//Media Watermarking, Security, and Forensics, Bellingham: SPIE, 2012, 8303: 7.

[25] JESSICA F. BOSSbase[EB/OL]. (2018-10-23)[2023-04-13]. http://dde.binghamton.edu/download/.

Cover selection method for batch image steganography based on multivariate optimization

WANG Yangguang[1]，YAO Yuanzhi[2,*]，YU Nenghai[1]

(1. School of Cyber Science and Technology，University of Science and Technology of China，Hefei 230027，China；
2. School of Computer Science and Information Engineering，Hefei University of Technology，Hefei 230601，China)

Abstract：Batch image steganography provides an effective means for covert communication on social networks by embedding secret messages into multiple cover images through cover selection. Compared with traditional image steganography, a key challenge of batch image steganography lies in designing an effective cover selection method while ensuring good undetectability performance. In this paper, a multivariate optimization-based cover selection method for batch image steganography was proposed. By jointly analyzing embedding distortion, image correlation, and embedding capacity, this method modelled the cover selection for batch image steganography as a multivariate optimization problem. Additionally, to enhance the batch image steganography robustness against the possible compression of images after steganography by social networks, a secret message fragmentation and reassembly strategy was designed. Experimental results demonstrate that the proposed method for batch image steganography achieves satisfactory performance in undetectability, embedding capacity, and robustness, providing technical support for covert communication on social networks.

Keywords：multivariate optimization; batch image steganography; cover selection; undetectability; robustness

Received：2023-06-16；**Accepted**：2023-09-02；**Published Online**：2023-11-08 12：03

URL：link.cnki.net/urlid/11.2625.V.20231107.1103.004

Foundation items：National Key Research and Development Program of China (2018YFB0804102); National Natural Science Foundation of China (61802357); The Fundamental Research Funds for the Central Universities (WK3480000009)

* **Corresponding author**. E-mail：yaoyz@hfut.edu.cn

http://bhxb.buaa.edu.cn jbuaa@buaa.edu.cn
DOI: 10.13700/j.bh.1001-5965.2023.0386

基于多标签对抗领域自适应的行人属性识别算法

胡强梁[1,2]，陈琳[1,2,*]，尚明生[1,2]

(1. 中国科学院 重庆绿色智能技术研究院 电子信息所，重庆 400714；2. 中国科学院大学重庆学院 人工智能学院，重庆 400714)

摘 要：针对无监督领域自适应算法通常局限于单标签学习问题，难以适配针对行人属性的多标签分类任务，提出一种多标签对抗领域自适应的行人属性识别算法。为适应行人属性多标签领域迁移任务，基于多标签特征分离模块，利用特定类别语义对主干网络提取的深度特征进行属性分离，有效提取特定属性的表征信息。针对不同领域属性特征分布差异较大的难点，提出基于分类器复用的多标签领域鉴别模块，同时实现多标签领域对齐和多标签分类，有效利用预测的鉴别信息捕获特征分布的多模式结构。实验结果表明：所提算法对比基准模型有明显提升，在平均准确率、准确率、召回率和F1指标上分别提升了4.49%、5.5%、11.44%和5.89%；所提算法为多标签领域自适应学习提供了新思路。

关 键 词：行人属性识别；多标签学习；多标签领域自适应；多属性特征提取；多标签分类
中图分类号：V351; TB391
文献标志码：A **文章编号**：1001-5965(2025)07-2478-10

行人属性识别通过提取给定人物图像中的年龄、性别等结构化属性信息，为视频智能化分析提供了结构化信息。与常规的单标签图像识别任务不同，行人属性识别通常需要对单个行人图像中数十个属性同时进行分析，因此，行人属性识别是更复杂的多标签分类任务。传统的行人属性识别方法通常侧重于从手工特征、属性分类器或属性关系的角度开发鲁棒的特征表示。随着深度学习技术的不断发展，基于深度学习的行人属性识别方法得到了广泛的研究和应用，其中包括基于全局特征、基于视觉注意力、基于序列预测、基于损失函数改进和基于图神经网络等诸多研究方向[1]。但基于深度学习的行人属性方法通常聚焦于有监督的学习任务，需要大量的标注数据实现模型的充分训练。

然而，构建大规模的行人图像数据集需要手动标记行人属性的多标签信息，相比单标签数据集，需要数倍的人力和时间成本。在处理未标记行人属性数据集时，如何充分利用已有标记行人属性数据集进行无监督训练，仍是一个重要且亟待解决的研究课题。

由于不同数据集使用不同图像传感器采集，并且受到光线、气候、角度等因素影响，在成像模态、分辨率和图片比例等方面都会存在差异，这就使得这些数据集之间存在显著的分布差异，这种差异被称为域偏移[2-3]。在存在域偏移的数据上直接进行无监督训练得到的模型通常难以保证其有效的识别精度。

在行人分析任务中，无监督领域自适应(unsupervised domain adaption, UDA)方法是有效解决域偏移的方法之一[2]，现有UDA研究主要通过探索领域不变特征表示方法，将领域知识迁移到目标任务中。例如，文献[4]将已标注的行人辅助身份标签转换为可学习的多标签软标记问题，挖掘潜在的标签信息提高无监督行人重识别任务精度。文献[5]

收稿日期：2023-06-16；**录用日期**：2023-09-23；**网络出版时间**：2023-11-29 09:20
网络出版地址：link.cnki.net/urlid/11.2625.V.20231128.1326.003
基金项目：国家自然科学基金（82030066,61902370）；重庆市教委在渝高校与中科院所合作重点项目（HZ2021008,HZ2021017）
*通信作者. E-mail：chenlin@cigit.ac.cn

引用格式：胡强梁，陈琳，尚明生. 基于多标签对抗领域自适应的行人属性识别算法[J]. 北京航空航天大学学报，2025，51(7)：2478-2487.
HU Q L，CHEN L，SHANG M S. Pedestrian attribute recognition algorithm based on multi-label adversarial domain adaptation[J]. Journal of Beijing University of Aeronautics and Astronautics，2025，51(7)：2478-2487 (in Chinese).

将单标签的无监督行人重识别转换为多标签学习任务, 利用同一身份来约束行人在不同图像中的特征一致性。然而, 直接将现有 UDA 方法直接套用到多标签的行人属性识别领域时, 由于特定属性关注不同的区域特征, 如帽子、眼镜等仅关注头部区域特征, 如果进行统一的领域特征学习, 可能在转换后的领域中混淆特有的属性特征, 导致实际应用效果欠缺。

针对上述问题, 本文提出基于无监督的多标签对抗域自适应行人属性识别算法, 利用对抗学习范式实现源域和目标域之间的多标签特征对齐和多属性预测。该算法由 2 个主要模块组成: 多标签特征分离 (multi-label feature disentanglement, MFD) 模块和多标签领域鉴别 (multi-label domain discriminator, MDD) 模块。其中, MFD 模块利用类别语义特征来指导学习与特定属性类别相关的图像特征, 并分离与该属性无关的图像特征, 使得分离后的属性特征更加关注相应的语义区域, 减少多标签领域迁移中的特征混淆问题。MDD 模块通过统一优化目标函数, 实现多标签领域对齐和分类, 利用预测的鉴别信息捕获多属性特征的复杂分布, 并有效地缓解负迁移现象。

本文工作主要有以下创新点:

1) 提出基于多标签对抗域自适应的行人属性识别算法, 将端到端的领域自适应方法引入多标签行人属性识别问题中, 用于提高无监督行人属性识别准确率。

2) 结合基于多标签语义解耦的多属性特征分离模块, 对主干网络提取的深度特征进行属性分离, 使领域自适应算法适应行人属性的多标签分类任务。同时引入基于多标签分类器的领域鉴别模块, 使对抗性领域自适应的训练过程更加稳定, 有效提高多标签领域迁移效果。

3) 建立了跨领域无监督行人属性识别训练方案, 在多个真实场景的行人属性数据集上进行了对比实验, 结果证明本文算法可以有效提高目标域缺失标注下的属性识别性能。

1　相关工作

1.1　行人属性识别

随着深度学习技术的不断发展, 基于深度学习的属性识别方法得到了广泛的研究和应用, 并取得了显著进展。文献 [6] 提出基于多属性联合学习的 DeepMAR 网络, 首次将神经网络引入到行人属性识别中。在此基础上, 许多学者发现行人属性识别任务与某些任务存在一定相关性, 有助于属性识别。如文献 [7] 提出基于多任务学习的属性识别, 该方法将属性识别与视角预测相结合。 文献 [8] 提出一种结合弱监督和定位方法的属性识别方法。文献 [9] 为减少与属性识别任务无关的额外特征的干扰, 提出结合定位引导任务的行人属性识别方法。文献 [10] 则认为姿态信息有助于属性识别任务, 引入姿态预测和人体部件定位任务。文献 [11] 提出用分割任务进一步提高对属性敏感区域的定位能力。文献 [12] 提出通过利用人体关键点精准捕获人体部件特征的方法。

近年来, 注意力机制也是深度学习领域应用较为广泛的组件之一, 许多学者也提出各种基于注意力机制的行人属性识别方法。文献 [13] 针对行人细粒度特征问题, 引入了注意力特征分支网络, 通过多个分支网络提取不同尺度的特征。文献 [14] 提出基于增广样本注意力一致的方法, 该方法建立在同一个样本图片经过不同的数据增广方法之后注意力区域应保持一致的假设上。文献 [15] 提出基于属性图片互相指导的注意力网络方法, 通过结合图片引导的特征和属性引导的特征, 为不同的属性定制不同综合特征的学习。文献 [16] 结合不同维度的注意力网络, 从多个角度出发捕获有助于属性识别的行人特征。

另外, 许多研究者发现行人属性之间存在着语义和空间关系, 可以利用这些关系提升属性识别效果。文献 [17] 提出可以根据属性之间的关系对所有属性进行分组, 然后使用长短时记忆网络学习组内组间属性的关系用以提升属性识别性能。文献 [18] 提出基于图卷积的属性识别方法, 首次将图卷积应用于该领域, 该方法的输入端除了行人图片额外增加了行人属性的语义词向量。文献 [19] 针对空间-语义关系、空间上下文及语义联系等 3 种关系进行建模, 提出基于序列上下文关系的学习模型。

上述基于深度学习的方法虽然已取得了较大的进展, 然而模型进行充分训练通常依赖于大量已标注数据, 当行人属性识别目标域缺失标注时, 如何基于领域自适应的行人属性识别方法来解决标注缺失的情况, 仍是当前行人属性识别中被忽视的一个难点问题。

1.2　无监督领域自适应

UDA 是迁移学习领域中备受关注的方法之一。主流的 UDA 方法存在几个大的方向, 其中, 基于度量标准和对抗学习的研究成果较为丰硕。在基于度量标准的研究上, 文献 [20] 提出深度适配网络 (deep adaptation network, DAN), 该模型引入了统

计学习中的多核最大均值差异，适配多个高层网络。文献 [21] 进一步提出了联合自适应网络，利用特征和标签的联合概率分布来减少不同域之间的分布差异。

在基于对抗学习的研究上，提出域对抗网络 (domain adversarial neural network, DANN)[22]。该方法关注于如何在不同域之间选择可供迁移的特征，把迁移任务看作域混淆问题，为此引入域判别器。文献 [23] 借鉴了这种思想，但与 DANN 不同的是利用重建目标域图像的方式来学习共享的编码表示。文献 [24] 将特征表示分为共享表示和私有表示，其中，共享表示可以用来进行迁移。 文献 [25] 提出对抗性区分域自适应模型，该模型通过在不同的训练阶段学习到 2 个独立的特征提取网络，但分类器是共享的，从而实现域适配。文献 [26] 认为只用一个判别器不能很好地捕捉到多模式结构，提出用多个判别器来捕捉细粒度特征。文献 [27] 提出循环一致性对抗域适应模型，使用生成性图像空间对齐和潜在表征空间对齐来适应不同领域。文献 [28] 提出的条件对抗域适应 (conditional adversarial domain adaptation, CDAN) 模型，利用分类器预测中所传递的判别信息来辅助对抗性适应。文献 [29] 提出的最大分类器差异 (maximum classifier discrepancy, MCD) 模型，使用 1 个特征提取器和 2 个任务特定的分类器，通过考虑特定任务的决策边界来调整目标域的分布。

以上所述的领域自适应算法通常仅考虑单标签迁移任务，而行人属性识别是典型的多标签多任务学习过程，如何利用多标签数据进行有效的领域知识迁移仍有待进一步的研究。

2 本文工作

本文提出的基于 UDA 的多标签对抗领域自适应 (multi-label adversarial domain adaptation, MLADA) 行人属性识别算法的模型框架如图 1 所示。该算法通过特征提取器 G，从带标注的源域数据和未标注的目标域数据中提取卷积神经网络（convolutional neural network, CNN）特征。这些特征输入到多属性特征分离模块 MFD 中，通过 MFD 将主干特征分离成基于特定属性的多标签特征向量。最终，分离后的特征向量进入多标签分类器产生分类损失。同时，多属性分类器 C 结合核范数 Wasserstein 距离 (nuclear-norm Wasserstein discrepancy, NWD) 差异和梯度反转层实现与特征提取器 G 的对抗学习，使得整个模型在实现属性识别的同时完成了领域对齐。

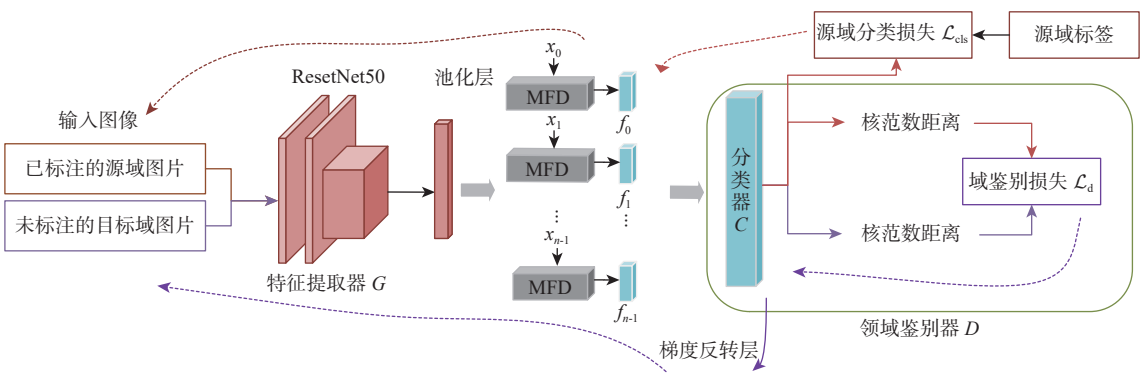


图 1 本文算法框架

Fig. 1 Framework of the proposed algorithm

2.1 符号定义

本文针对 UDA 多标签行人属性识别问题，源域中有标签的行人属性数据集定义如下：$\boldsymbol{D}_s=\{x_i^s,y_i^s\}_{i=1}^{n_s}$，其中，$x_i^s$ 表示源域的第 i 个行人图像样本，y_i^s 表示源域第 i 个图像样本对应的多属性标记，n_s 表示来自源域的样本数量。定义目标域的未标注行人属性数据集 $\boldsymbol{D}_t=\{x_i^t\}_{i=1}^{n_t}$，其中，$x_i^t$ 表示目标域的第 i 个行人图像样本，n_t 表示目标域的样本数量，目标域的样本标签未知。

2.2 模型概述

本文算法框架如图 1 所示，框架中红色部分表示的是特征提取器，使用 ResNet50 作为主干网络。灰色部分表示用于对主干特征进行特征解耦的多属性特征分离模块，蓝色部分表示行人属性识别的多标签分类器，由多个全连接网络构成。本文算法的整体损失函数由 2 个损失组成，分别为源域样本的分类损失 $\mathcal{L}_{cls}$，源域特征和目标域特征输入隐式多标签领域鉴别模块后产生的域鉴别损失 $\mathcal{L}_d$。

2.2.1 多标签特征分离模块

多标签特征分离模块以标签语义的注意机制为指导，学习语义特定的特征表示[30]。给定输入图像 $\boldsymbol{A}$，首先，通过特征提取器提取其特征图 $\boldsymbol{F}\in$

$\mathbf{R}^{W\times H\times N}$, 其中, W、H 和 N 分别为特征图的宽度、高度和通道数, 如式 (1) 所示:

$$\boldsymbol{F}^{\boldsymbol{A}} = f_{\text{cnn}}(\boldsymbol{A}) \tag{1}$$

式中: $f_{\text{cnn}}(\cdot)$为一个特征提取器, 由卷积网络实现, 本文使用的是 ResNet50 作为主干基础网络。对于每个标签 l, 本文算法使用预先训练的 Glove 模型提取 d 维语义嵌入向量, 具体实现时 d 设为 300, 如式 (2) 所示:

$$\boldsymbol{x}_l = f_{\text{g}}(w_l) \tag{2}$$

式中: w_l 为标签 l 的语义词; $f_{\text{g}}(\cdot)$为已训练好的 Glove 模型。得到语义特征向量 $\boldsymbol{x}_l$, 之后引入了一种语义引导的注意机制, 该机制结合了 $\boldsymbol{x}_l$来引导更多地关注语义感知区域, 从而学习与该标签相对应的特征向量。更具体地, 对图像中每个位置 (w,h), 首先, 使用低秩双线性方法融合相应的图像特征 $\boldsymbol{f}_{wh}^{\boldsymbol{A}}$和 $\boldsymbol{x}_l$, 如式 (3) 所示:

$$\tilde{\boldsymbol{f}}_{l,wh}^{\boldsymbol{A}} = \boldsymbol{P}^{\text{T}}\left(\tanh\left(\left(\boldsymbol{U}^{\text{T}}\boldsymbol{f}_{wh}^{\boldsymbol{A}}\right)\odot\left(\boldsymbol{V}^{\text{T}}\boldsymbol{x}_l\right)\right)\right)+\boldsymbol{b} \tag{3}$$

式中: tanh(·)为双曲正切函数; $\boldsymbol{U}\in\mathbf{R}^{N\times d_1}$、$\boldsymbol{V}\in\mathbf{R}^{d_{\text{s}}\times d_1}$、$\boldsymbol{P}\in\mathbf{R}^{d_1\times d_2}$、$\boldsymbol{b}\in\mathbf{R}^{1\times d_2}$为可学习的参数, $\odot$为逐元素乘法运算, d_1 和 d_2 分别为联合嵌入和输出特征的尺寸, d_{s} 为语义向量维度。然后, 在 $\boldsymbol{x}_l$的指导下计算注意力系数$\tilde{a}_{l,wh}$, 该注意力表示位置 (w,h) 的重要程度, 计算过程为

$$\tilde{a}_{l,wh} = f_{\text{att}}\left(\tilde{\boldsymbol{f}}_{l,wh}^{\boldsymbol{A}}\right) \tag{4}$$

式中: $f_{\text{att}}(\cdot)$为注意力函数, 由一个全连接网络实现。对所有位置重复该过程, 为了使不同样本的系数易于比较, 使用 softmax 函数对所有位置进行归一化:

$$a_{l,wh} = \frac{\text{e}^{\left(\tilde{a}_{l,wh}\right)}}{\sum\limits_{w',h'}\text{e}^{\left(\tilde{a}_{l,w'h'}\right)}} \tag{5}$$

最后, 对所有位置执行加权平均池化以获得编码与标签l相关信息的特征向量:

$$\boldsymbol{F}_l = \sum_{w,h} a_{l,wh}\boldsymbol{f}_{wh}^{\boldsymbol{A}} \tag{6}$$

2.2.2　多标签领域鉴别模块

现有的基于对抗训练的无监督域适应方法通常需要额外的鉴别器, 通过优化源域数据分类损失和领域鉴别损失来实现跨领域的知识迁移。然而这些方法忽视了用于源域数据分类的分类器, 其领域对抗性损失是由特征提取器和额外的领域鉴别器对抗产生的, 而分类器只考虑源域数据的分类准确性。事实上, 由于分类器是通过对源域数据进行监督学习得到的, 在源域数据上具有更高的准确性, 而在无标注的目标域上分类器识别率较低, 即: 针对源域中的相似数据, 分类器具有相似的分类结果, 而对于目标域的相似数据, 分类器表现得较为迟钝。

因此, 鉴于多标签分类器具有隐式辨别能力, 可结合核范数用于鉴别源域和目标域, 即基于复用分类器的领域鉴别模块[31], 本文将其扩展到多标签领域自适应问题, 提出 MDD 模块, 以提高多标签对抗学习能力。通过多标签分类器 C 可以得到预测矩阵 $C(\boldsymbol{F})\in\mathbf{R}^{b\times k}$, 其包含 b 个样本的 k 个属性的预测值, $C(\cdot)$表示分类函数。结合能提高预测多样性的核范数, 可以得出 2 个域属性特征分布的 NWD 差异 W, 该值可显示出分类器对 2 个域的鉴别能力, 如式 (7) 所示:

$$W = \sup_{\|\|C\|_*\|_L\leqslant K} E_{\tilde{\boldsymbol{D}}_{\text{s}}}[\|C(\boldsymbol{F})\|_*] - E_{\tilde{\boldsymbol{D}}_{\text{t}}}[\|C(\boldsymbol{F})\|_*] \tag{7}$$

式中: $\|\cdot\|_*$为核范数; $\|\cdot\|_L$为 Lipschitz 半范数; K 为 Lipschitz 常数; E 为期望; $\tilde{\boldsymbol{D}}_{\text{s}}$和 $\tilde{\boldsymbol{D}}_{\text{t}}$分别为 $\boldsymbol{D}_{\text{s}}$和 $\boldsymbol{D}_{\text{t}}$的特征分布。其中, 鉴别器可表述为$D=\|C\|_*$, 即结合了核范数的分类器。当分类器 C 用于分类时, 它可以实现对多标签图片的分类, 与此同时, 当 C 用作鉴别器时, 又能实现领域之间特征级别的对齐域鉴别损失的计算:

$$\mathcal{L}_{\text{d}}\left(\boldsymbol{x}^{\text{s}},\boldsymbol{x}^{\text{t}}\right) = \frac{1}{N_{\text{s}}}\sum_{i=1}^{N_{\text{s}}} D\left(Y\left(G\left(\boldsymbol{x}_i^{\text{s}}\right)\right)\right) - \frac{1}{N_{\text{t}}}\sum_{i=1}^{N_{\text{t}}} D\left(Y\left(G\left(\boldsymbol{x}_i^{\text{t}}\right)\right)\right) \tag{8}$$

式中: $Y(\cdot)$为多属性特征分离模块; $D(\cdot)$为判别器模块。

通过最大化域鉴别损失 $\mathcal{L}_{\text{d}}$使领域鉴别误差的最大化, 即特征提取器 G 生成的特征要混淆源域数据集和目标域数据集:

$$\hat{W} = \max_{D}\mathcal{L}_d\left(\boldsymbol{x}^{\text{s}},\boldsymbol{x}^{\text{t}}\right) \tag{9}$$

通过对抗训练, 当特征提取器 G 生成的特征使 C 无法鉴别时, 此时通过 G 提取的不同领域属性特征实现了一定程度的对齐。

2.2.3　多标签对抗训练损失函数

行人属性识别通过在源域 $\boldsymbol{D}_{\text{s}}$ 中有监督地训练得到多标签分类器 C, 将 C 用于无标签的源域 $\boldsymbol{D}_{\text{t}}$ 中得到预测结果。行人属性多标签分类损失表达式为

$$\mathcal{L}_{\text{cls}} = \frac{1}{N}\sum_{i=1}^{N}\boldsymbol{y}_i\ln(\sigma(C(\boldsymbol{F}))) + (1-y_i)\ln(1-\sigma(C(\boldsymbol{F}))) \tag{10}$$

式中: $\boldsymbol{F}$ 为经特征提取器和多属性特征分离模块处理得到的所有属性图像特征; $\sigma(\cdot)$为 sigmoid 函数。

目前被广泛采用的行人属性公开数据集中，都存在严重的不平衡问题，即在数据集中对应某种属性通常为负样本，只存在少数正样本（如是否光头、戴帽子等）。属性分布不均衡的现象导致行人属性识别模型在进行训练时接收大量负样本学习，模型识别更倾向于负例，导致模型对正样本的识别正确率较低。为缓解这一问题，通过引入样本权重 $\boldsymbol{w}$ 进行平衡[32]，其中，$\boldsymbol{w}=\{w_j\}_{j=1}^{K}$，$K$ 为数据集的属性数，w_j 的定义如下：

$$w_j=\begin{cases}\mathrm{e}^{m_j} & y_{ij}=0\\ \mathrm{e}^{1-m_j} & y_{ij}=1\end{cases} \tag{11}$$

式中：m_j 为第 j 个属性的正样本率。因此，基于样本权重的源域已标注数据的分类损失计算式为

$$\mathcal{L}_{\mathrm{cls}}=\frac{1}{N_{\mathrm{s}}}\sum_{i=1}^{N_{\mathrm{s}}}\boldsymbol{w}(\boldsymbol{y}_i^{\mathrm{s}}\ln(\sigma(C(\boldsymbol{F}_{\mathrm{s}})))+\left(1-\boldsymbol{y}_i^{\mathrm{s}}\right)\ln(1-\sigma(C(\boldsymbol{F}_{\mathrm{s}}))) \tag{12}$$

对抗学习过程如图 1 所示，对抗训练存在于特征提取器和分类器之间，其中，多标签分类器 C 同时完成多属性分类与领域鉴别，特征提取器 G 基于 MFD 模块生成具有区分度的属性特征，同时基于共享的主干网络，使源域与目标域上生成的特征一致来混淆分类器 C，从而实现多标签领域自适应任务。为避免模型的单调交替更新，使用梯度反转层来实现在一次反向传播内完成所有参数的更新，最终利用 min-max 模式完成领域自适应的对抗学习：

$$\mathcal{L}_{\mathrm{d}}=\min_{G}\max_{C}\mathcal{L}_{\mathrm{d}}\left(\boldsymbol{x}^{\mathrm{s}},\boldsymbol{x}^{\mathrm{t}}\right) \tag{13}$$

式中：特征提取器的优化方向为最小化 $\mathcal{L}_{\mathrm{d}}$ 实现不同领域之间的特征对齐，而分类器的优化方向为最大化 $\mathcal{L}_{\mathrm{d}}$ 实现。

综上所述，MLADA 模型的总损失函数 $\mathcal{L}_{\mathrm{total}}$ 由 2 部分组成，一是源域多标签分类损失 $\mathcal{L}_{\mathrm{cls}}$，二是用于领域自适应的鉴别损失 $\mathcal{L}_{\mathrm{d}}$：

$$\mathcal{L}_{\mathrm{total}}=\min_{G}\{\min_{C}\mathcal{L}_{\mathrm{cls}}(\boldsymbol{x}^{\mathrm{s}},\boldsymbol{y}^{\mathrm{s}})+\lambda\max_{C}\mathcal{L}_{\mathrm{d}}(\boldsymbol{x}^{\mathrm{s}},\boldsymbol{x}^{\mathrm{t}}))\} \tag{14}$$

式中：λ 为超参数用于平衡 $\mathcal{L}_{\mathrm{cls}}$ 和 $\mathcal{L}_{\mathrm{d}}$，文中设置为 1。通过对上述损失目标的优化，MLADA 可实现源域到目标域的多标签 UDA，有效提高行人属性识别模型在目标域中的识别准确率。

3 实验设置

3.1 数据集

RAPv1[33] 数据集是一个采集于室内监控场景的数据集，来源于 26 台不同位置的摄像机。包含 41 585 张图像，分辨率范围从 36×92 到 344×554 不等。其中，33 268 张图像被用于训练算法模型，剩余的图像则被用于测试模型的准确性。

RAPv2[33] 在 RAPv1 的基础上进行扩充，包含 84 928 张图像，这些图像被分为 3 个部分：训练集、验证集和测试集，分别包含 50 957 张、16 986 张和 16 985 张图像。

PA100K[13] 数据集是由 598 个真实户外监控摄像头捕获到的图像构建而成，包含 100 000 张行人图像，是目前为止最大的行人属性识别数据集。整个数据集分为训练集、验证集和测试集 3 个部分，比例为 8∶1∶1。

3.2 评估指标

本文使用 2 种评价方法来衡量属性识别算法的性能，分别是基于标签的方法和基于实例的方法。对于基于标签的方法，先计算每个属性的准确率，再取所有属性的平均准确率 (mean accuracy, mA) 作为评价指标，记为 E_{mA}，具体定义如下：

$$E_{\mathrm{mA}}=\frac{1}{2N}\sum_{i=1}^{K}\left(\frac{T_i^{\mathrm{P}}}{P_i}+\frac{T_i^{\mathrm{N}}}{N_i}\right) \tag{15}$$

式中：K 为行人属性个数；T_i^{P} 和 T_i^{N} 分别为第 i 个属性正确预测的正样本和负样本个数；P_i 和 N_i 分别为第 i 个属性的总的正样本和负样本的数量；N 为数据集中所有样本总数。

从基于实例的角度来看，评价指标通常包括准确率（accuracy, Acc）、精准度（precision, Prec）、召回率（recall, Rec）和 F1 值这 4 个常用的指标，分别记为 E_{Acc}、E_{Prec}、E_{Rec} 和 E_{F1}，具体定义如下：

$$E_{\mathrm{Acc}}=\frac{1}{N}\sum_{i=1}^{N}\frac{|y_i\cap\hat{y}_i|}{|y_i\cup\hat{y}_i|} \tag{16}$$

$$E_{\mathrm{Prec}}=\frac{1}{N}\sum_{i=1}^{N}\frac{|y_i\cap\hat{y}_i|}{|\hat{y}_i|} \tag{17}$$

$$E_{\mathrm{Rec}}=\frac{1}{N}\sum_{i=1}^{N}\frac{|y_i\cap\hat{y}_i|}{|y_i|} \tag{18}$$

$$E_{\mathrm{F1}}=\frac{2E_{\mathrm{Prec}}E_{\mathrm{Rec}}}{E_{\mathrm{Prec}}+E_{\mathrm{Rec}}} \tag{19}$$

式中：y_i 为测试样本中第 i 个属性的真实标签；$\hat{y}_i$ 为模型输出的第 i 个属性的预测结果。

3.3 实验设置

本文算法与其他 UDA 最新的先进算法进行了比较，为确保实验结果的准确性，对所有对比算法都进行了官方代码复现实验，并保证了数据集预处理方法和计算评价指标的方法完全相同。在实验中，本文对图片应用了一些数据增强技术（如随机擦除、随机裁剪、水平翻转），以提高数据的多样性

和实验的鲁棒性。基础网络采用 ResNet50 作为主干网络, 优化器采用随机梯度下降(stochastic gradient descent, SGD) 算法, 训练时实验软硬件环境和模型的具体参数如表 1 和表 2 所示。

表 1　实验软硬件环境

Table 1　Experimental hardware and software environment

实验环境	配置
CPU	Intel Xeon E5-2680 v4
GPU	NVIDIA GeForce RTX 1080Ti
操作系统	Ubuntu 18.04.2
深度学习框架	Pytorch 1.9
计算机语言	Python 3.8

表 2　模型参数设置

Table 2　Model parameters setting

参数	数值
图片预处理大小	256×192
批处理大小	64
学习率	0.01
动量大小	0.9
训练轮次	30

为更准确地评估本文算法的性能, 本文在 3 个常用行人属性识别数据集 PA100k、RAPv1 和 RAPv2 上进行实验, 并选取其公共 10 个属性进行 2 组 UDA 实验。第 1 组: RAPv1⇛PA100k。第 2 组: RAPv2⇛PA100k。定义 A⇛B 表示, 以 B 为目标域, A 为源域, 进行 UDA 实验。

4　实验结果与分析

4.1　对比实验分析

为评估 MLADA 在 UDA 行人属性识别任务上的有效性, 将本文算方法与 UDA 中的先进算法进行对比实验, 包括基于域鉴别器对抗的 DANN[22]、基于多核最大均值差异的 DAN[20]、基于最大化双分类器差异的 MCD[29]、基于双调节策略的 CDAN[28] 及基于复用分类器的(Discriminator-free Adversarial Learning Network, DALN)[31]。实验中所有算法的基准模型使用行人属性领域中基于改进损失函数的 StrongBaseline[32], 该基准模型直接在源域上进行训练并在目标域上进行预测得到属性识别结果。

图 2 展示了上述 5 种算法与本文算法的效果对比, 可见本文算法在 RAPv1⇛PA100k 和 RAPv2⇛PA100k 这 2 种实验场景下, 性能都优于其他对比算法。图 2(a)使用 RAPv1 数据集全部 41 585 张图片作为源域已标注数据, 80 000 张 PA100k 训练集数据作为目标域未标注数据, 本文算法在 E_{mA}、

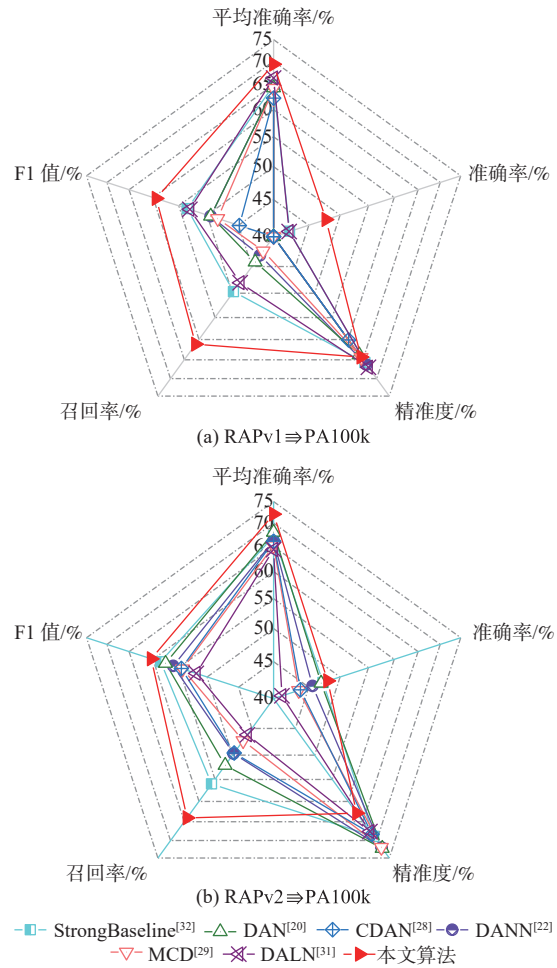


图 2　对比实验效果

Fig. 2　Comparison of experiment results

E_{Acc} 和 E_{Rec} 指标上提升明显, 均超过了基线算法及对比的领域自适应算法, 例如, 对比最近的 UDA 算法 DALN[28], 本文算法在 E_{mA}、E_{Acc} 和 E_{Rec} 提升了 7.7%、6.85% 和 17.86%, 虽然由于召回率大幅提升导致 E_{Prec} 指标下降(−4.71%), 但在综合识别率 E_{F1} 指标上提升了 8.45%。

图 2(b)使用 RAPv2 数据集全部 84 928 张图片作为源域已标注数据, 本文算法在 E_{mA}、E_{Acc}、E_{Rec} 和 E_{F1} 指标上分别比次优效果提升了 4.03%、0.75%、8.01% 和 1.86%。从上述分析可得, 本文算法通过结合 MFD 模块和 MDD 模块, 在应对多标签无监督领域属性识别自适应任务上, 有着明显的优势, 可以有效提升标签缺失情况下的行人属性识别性能。

图 3 展示了第 1 组 RAPv1⇛PA100k 实验中 6 种测试算法的行人属性准确度。在大多数属性中, 本文算法都优于其他对比算法, 只有性别、短袖和靴子不是最优。其中, 本文算法的裤子属性比次优高出了 8.81%。

图 4 展示了第 2 组 RAPv2⇛PA100k 实验中,

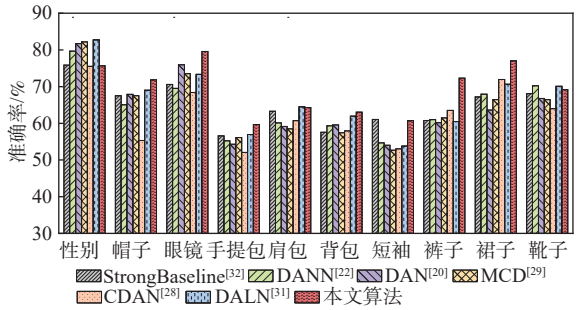


图3 RAPv1⇛PA100k的各属性准确率

Fig. 3 Accuracy of each attribute on RAPv1⇛PA100k

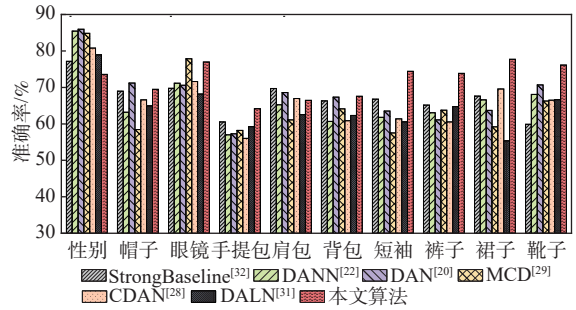


图4 RAPv2⇛PA100k的各属性准确率

Fig. 4 Accuracy of each attribute on RAPv2⇛PA100k

6种算法的各属性准确度对比。在多数属性中，本文算法均优于其他对比算法，尤其是短袖、裤子和裙子属性，比次优算法分别高出7.64%、8.74%和8.17%，其他属性识别结果虽不是最优但也和最优相差无几。

为进一步分析本文算法相较于常规的基线行人属性识别性能差异，使用StrongBaseline[32]和本文算法在PA100k测试集上进行属性识别效果分析，这2种模型都是在RAPv2⇛PA100k实验下使用对应的方法训练而得，如图5所示。相比于未使用领域自适应的基线模型StrongBaseline，本文算法能够更加准确识别多个属性，例如：背包属性，StrongBaseline识别成了手提包或肩包。另外，StrongBaseline更容易造成漏识别，如眼镜、短袖等，上述结果表明本文算法有效降低源域与目标的领域偏移从而得到更好的识别结果。



图5 属性识别效果分析(RAPv2⇛PA100k)

Fig. 5 Performance of attribute recognition on RAPv2⇛PA100k

综上所述，本文算法相较于现有的对比算法，在多标签行人属性识别领域自适应效果上有着更好的表现。

4.2 消融实验分析

为评估不同模块对实验效果的贡献，在不加入对抗损失情况下，以ResNet50为基线模型，仅使用行人属性多标签分类损失$\mathcal{L}_{\text{cls}}$，以及在$\mathcal{L}_{\text{cls}}$中加入多标签分离模块MFD(ResNet50+MFD)进行对比分析。其次，在加入领域自适应鉴别损失$\mathcal{L}_{\text{d}}$中，包括了加入使用MFD及传统的DANN[22]模型进行对抗训练(ResNet50+MFD+DA)，单独加入多标签领域鉴别模块MDD(ResNet50+MDD)，以及加入MFD后(ResNet50+MFD+MDD，即本文算法)的结果进行对比分析。

在RAPv1⇛PA100k和RAPv2⇛PA100k这2种实验场景下进行消融实验，实验结果如表3和表4所示，最优结果用加粗表示。

由表3和表4可知，在加入MFD模块后，在2组实验下E_{mA}值都有所提高，分别提升了0.47%和1.63%，说明多标签特征分离有助于主干网络提取更有利于行人属性识别的特征。在加入对抗损失后，单独加入MDD模块时，第1组实验下E_{mA}值

表 3　在 RAPv1⇒PA100k 上的消融实验

Table 3　Ablation experiments on RAPv1⇒PA100k

算法	E_{mA}/%	E_{Acc}/%	E_{Prec}/%	E_{Rec}/%	E_{F1}/%
ResNet50	64.85	42.22	65.75	49.71	53.44
ResNet50+MFD	65.32	43.04	**67.17**	51.93	54.99
ResNet50+MFD+DA	65.83	39.41	63.43	45.86	50.01
ResNet50+MDD	66.34	42.14	66.95	47.99	52.99
本文算法	**69.34**	**47.72**	64.33	**61.15**	**59.33**

表 4　在 RAPv2⇒PA100k 上的消融实验

Table 4　Ablation experiments on RAPv2⇒PA100k

算法	E_{mA}/%	E_{Acc}/%	E_{Prec}/%	E_{Rec}/%	E_{F1}/%
ResNet50	67.19	47.24	68.93	56.01	58.47
ResNet50+MFD	68.82	47.92	**73.15**	55.28	59.57
ResNet50+MFD+DA	67.74	44.51	66.22	53.26	55.66
ResNet50+MDD	64.34	41.14	67.56	46.15	51.88
本文算法	**72.04**	**47.99**	62.85	**64.01**	**60.33**

提升了 1.49%, 而在第 2 组实验 E_{mA} 值反而下降, 这说明 MDD 模块在没有属性分离时, 对于特定的属性特征提取困难, 导致模型的性能提升并不稳定。值得注意的是, 使用传统对抗训练结合 MFD 模块(ResNet50+MFD+DA)的结果, 对比没有对抗学习的 ResNet50+MFD 模型在 2 个数据集上效果均有所下降, 说明如不使用 MDD 模块复用分类器进行多属性对抗, 导致了一定的领域混淆, 从而导致精度下降。

当同时加入 MFD 模块和 MDD 模块时, 在 2 组实验下 E_{mA}、E_{Acc}、E_{Rec} 和 E_{F1} 值都达到了最高。在第 1 组实验中, 对比基线 ResNet50 模型, 本文算法的 E_{mA}、E_{Acc}、E_{Rec} 和 E_{F1} 值分别提高了 4.49%、5.5%、11.44% 和 5.89%。值得注意的是, 由于召回率的显著性提升(+11.44%), 导致识别精度 E_{Prec} 有所下降(−6.08%), 但综合性能 E_{F1} 值保持了增长(+5.89%)。另外, 考虑实际应用场景中召回率相比识别精度更为重要(如检出背包、特定鞋子等属性), 因此, 本文算法更具有实用价值。由上述分析表明, 在 MFD 和 MDD 模块的共同作用下, 可以有效提升 UDA 多标签属性识别性能。

5　结　论

本文提出一种基于 UDA 的多标签行人属性识别算法, 该算法包括 2 个主要模块: 多属性特征分离模块和基于分类器复用的多标签领域鉴别模块。

1) 多属性特征分离模块通过使用类别的语义特征来指导图像特征的学习, 实现了特定属性特征分离提取。基于分类器复用的多标签领域鉴别模块通过使用对抗范式, 实现了多标签领域对齐和标签分类。

2) 本文在多个公开数据集上进行了实验, 结果表明, 本文算法在各项评价指标上都优于现有算法, 证明了其有效性和可行性。

在下一步工作中, 拟基于半监督自适应的方法, 利用目标域中的少量标签数据对多标签分类器进行增强, 进一步提高行人属性领域自适应学习的精度。

参考文献 (References)

[1] WANG X, ZHENG S F, YANG R, et al. Pedestrian attribute recognition: a survey[J]. Pattern Recognition, 2022, 121: 108220.

[2] WILSON G, COOK D J. A survey of unsupervised deep domain adaptation[J]. ACM Transactions on Intelligent Systems and Technology, 2020, 11(5): 1-46.

[3] STACKE K, EILERTSEN G, UNGER J, et al. Measuring domain shift for deep learning in histopathology[J]. IEEE Journal of Biomedical and Health Informatics, 2021, 25(2): 325-336.

[4] YU H X, ZHENG W S, WU A, et al. Unsupervised person re-identification by soft multilabel learning[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2019: 2148-2157.

[5] WANG D, ZHANG S. Unsupervised person re-identification via multi-label classification[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2020: 10978-10987.

[6] LI D W, CHEN X T, HUANG K Q. Multi-attribute learning for pedestrian attribute recognition in surveillance scenarios[C]//Proceedings of 3rd IAPR Asian Conference on Pattern Recognition. Piscataway: IEEE Press, 2015: 111-115.

[7] SARFRAZ M S, SCHUMANN A, WANG Y, et al. Deep view-sensitive pedestrian attribute inference in an end-to-end model[C]//Proceedings of the British Machine Vision Conference. London: BMVA Press, 2017, 134: 1-13.

[8] YU K, LENG B, ZHANG Z, et al. Weakly-supervised learning of mid-level features for pedestrian attribute recognition and localization[C]//Proceedings of the British Machine Vision Conference. London: BMVA Press, 2017: 1-11.

[9] LIU P, LIU X, YAN J, et al. Localization guided learning for pedestrian attribute recognition[C]//Proceedings of the British Machine Vision Conference. London: BMVA Press, 2018: 142-152.

[10] LI D, CHEN X, ZHANG Z, et al. Pose guided deep model for pedestrian attribute recognition in surveillance scenarios[C]//Proceedings of the IEEE International Conference on Multimedia and Expo. Piscataway: IEEE Press, 2018: 1-6.

[11] WU M D, HUANG D, GUO Y F, et al. Distraction-aware feature learning for human attribute recognition via coarse-to-fine attention mechanism[C]//Proceedings of the AAAI Conference on Artificial Intelligence. Palo Alto: AAAI Press, 2020: 12394-12401.

[12] ZHANG J, REN P, LI J. Deep template matching for pedestrian attribute recognition with the auxiliary supervision of attribute-wise keypoints[EB/OL]. (2020-11-13)[2020-05-26]. https://doi.org/10.48550/arXiv.2011.06798.

[13] LIU X, ZHAO H, TIAN M, et al. Hydraplus-net: attentive deep features for pedestrian analysis[C]//Proceedings of the IEEE International Conference on Computer Vision. Piscataway: IEEE Press, 2017: 350-359.

[14] GUO H, ZHENG K, FAN X, et al. Visual attention consistency under image transforms for multi-label image classification[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2019: 729-739.

[15] JI Z, HE E, WANG H, et al. Image-attribute reciprocally guided attention network for pedestrian attribute recognition[J]. Pattern Recognition Letters, 2019, 120: 89-95.

[16] TAN Z, YANG Y, WAN J, et al. Attention-based pedestrian attribute analysis[J]. IEEE Transactions on Image Processing, 2019, 28 (12): 6126-6140.

[17] ZHAO H, TIAN M, SUN S, et al. Spindle net: person re-identification with human body region guided feature decomposition and fusion[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2017: 1077-1085.

[18] LI Q, ZHAO X, HE R, et al. Visual-semantic graph reasoning for pedestrian attribute recognition[C]//Proceedings of the AAAI Conference on Artificial Intelligence. Palo Alto: AAAI Press, 2019, 33(1): 8634-8641.

[19] WU J, LIU H, JIANG J, et al. Person attribute recognition by sequence contextual relation learning[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2020, 30(10): 3398-3412.

[20] LONG M, CAO Y, WANG J, et al. Learning transferable features with deep adaptation networks[C]//Proceedings of the International Conference on Machine Learning. Piscataway: IEEE Press, 2015: 97-105.

[21] LONG M, ZHU H, WANG J, et al. Deep transfer learning with joint adaptation networks[C]//Proceedings of the International Conference on Machine Learning. Piscataway: IEEE Press, 2017: 2208-2217.

[22] GANIN Y, LEMPITSKY V. Unsupervised domain adaptation by backpropagation[C]//Proceedings of the International Conference on Machine Learning. Piscataway: IEEE Press, 2015, 9908: 1180-1189.

[23] GHIFARY M, KLEIJN W B, ZHANG M, et al. Deep reconstruction-classification networks for unsupervised domain adaptation[C]// Proceedings of the European Conference on Computer Vision. Berlin: Springer, 2016: 597-613.

[24] BOUSMALIS K, TRIGEORGIS G, SILBERMAN N, et al. Domain separation networks[C]//Proceedings of the Annual Conference on Neural Information Processing Systems. Piscataway: IEEE Press, 2016, 29: 343-351.

[25] TZENG E, HOFFMAN J, SAENKO K, et al. Adversarial discriminative domain adaptation[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2017: 7167-7176.

[26] PEI Z, CAO Z, LONG M, et al. Multi-adversarial domain adaptation[C]//Proceedings of the AAAI Conference on Artificial Intelligence. Palo Alto: AAAI Press, 2018, 32(1): 3934-3941.

[27] HOFFMAN J, TZENG E, PARK T, et al. Cycada: cycle-consistent adversarial domain adaptation[C]//Proceedings of the International Conference on Machine Learning. Piscataway: IEEE Press, 2018: 1989-1998.

[28] LONG M, CAO Z, WANG J, et al. Conditional adversarial domain adaptation[C]//Proceedings of the Annual Conference on Neural Information Processing Systems. Piscataway: IEEE Press, 2018, 31: 1-11.

[29] SAITO K, WATANABE K, USHIKU Y, et al. Maximum classifier discrepancy for unsupervised domain adaptation[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2018: 3723-3732.

[30] CHEN T, XU M, HUI X, et al. Learning semantic-specific graph representation for multi-label image recognition[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE Press, 2019: 522-531.

[31] CHEN L, CHEN H, WEI Z, et al. Reusing the task-specific classifier as a discriminator: discriminator-free adversarial domain adaptation[C]//Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2022: 7181-7190.

[32] JIA J, HUANG H, YANG W, et al. Rethinking of pedestrian attribute recognition: realistic datasets with efficient method[EB/OL]. (2020-05-25)[2020-05-26]. https://doi.org/10.48550/arXiv.2005.11909.

[33] LI D, ZHANG Z, CHEN X, et al. A richly annotated pedestrian dataset for person retrieval in real surveillance scenarios[J]. IEEE Transactions on Image Processing, 2018, 28(4): 1575-1590.

Pedestrian attribute recognition algorithm based on multi-label adversarial domain adaptation

HU Qiangliang[1, 2], CHEN Lin[1, 2, *], SHANG Mingsheng[1, 2]

(1. Chongqing Institute of Green and Intelligent Technology, Chinese Academy of Sciences, Chongqing 400714, China;
2. School of Artificial Intelligence, Chongqing School of University of Chinese Academy of Sciences, Chongqing 400714, China)

Abstract: Current unsupervised domain adaption algorithms usually consider only single-label learning, failing to adapt to multi-label classification tasks in pedestrian attribute recognition. To address this issue, a multi-label adversarial domain adaptation algorithm was proposed for pedestrian attribute recognition. First, to adapt to the domain transfer task of multiple attribute labels, a multi-label feature disentanglement module was employed to effectively disentangle attributes of deep features extracted from the backbone network with specific attributes based on category-specific semantics. Secondly, to reduce the gap of attribute feature distributions in different domains, a multi-label domain discrimination module based on classifier reuse was proposed to achieve both multi-attribute domain alignment and multi-label classification, which effectively exploited the predicted discrimination information to capture the multi-mode structures of feature distribution. Experimental results show that the proposed algorithm achieves superior results than the baseline model, with improvements of 4.49%, 5.5%, 11.44% and 5.89% on mean accuracy, accuracy, recall and F1 values, respectively. The proposed algorithm provides a new insight for multi-label domain adaptation learning.

Keywords: pedestrian attribute recognition; multi-label learning; multi-label domain adaptation; multi-label feature extraction; multi-label classification

Received: 2023-06-16; **Accepted**: 2023-09-23; **Published Online**: 2023-11-29 09:20

URL: link.cnki.net/urlid/11.2625.V.20231128.1326.003

Foundation items: National Natural Science Foundation of China (82030066,61902370); The Key Project of Chongqing Municipal Education Commission based on Cooperation of Chongqing Local Universities and the Chinese Academy of Sciences (HZ2021008,HZ2021017)

*** Corresponding author**. E-mail: chenlin@cigit.ac.cn

http://bhxb.buaa.edu.cn jbuaa@buaa.edu.cn
DOI: 10.13700/j.bh.1001-5965.2023.0392

基于双路视觉 Transformer 的图像风格迁移

纪宗杏，贝佳*，刘润泽，任桐炜
(南京大学 计算机软件新技术国家重点实验室，南京 210093)

摘　　要：图像风格迁移旨在根据风格图像调整内容图像的视觉属性，使其保留原始内容的同时呈现出特定风格样式，从而生成具有视觉吸引力的风格化图像。针对现有代表性方法大多未考虑不同图像域间的编码差异，专注提取图像局部特征而忽视了全局上下文信息的重要性，提出一种新型的基于双路视觉 Transformer 的图像风格迁移方法 Bi-Trans，对内容图像域和风格图像域进行独立编码，提取风格参数向量以离散化表征图像风格，通过交叉注意力机制与条件实例归一化（CIN）将内容图像标定至目标域风格，从而生成风格化图像。实验结果表明，该方法无论是内容保留度还是风格还原度均优于现有方法。

关 键 词：图像风格迁移；视觉 Transformer；任意风格化；条件实例归一化；注意力机制
中图分类号：TP301
文献标志码：A　　**文章编号**：1001-5965(2025)07-2488-10

图像风格迁移旨在将图像渲染为特定风格，同时保留其原始内容语义。虽然，生成式人工智能 (artifical intelligence generated content, AIGC) 也能够利用内容图像与若干个文本提示词自动化生成具有特定视觉语义的精美图像，满足个性化服务与智能性创作需求。然而，不同于图像风格迁移任务定义，即仅调整内容图像的视觉属性而非生成全新图像，图像风格迁移相关的 AIGC 研究[1-3] 通常基于隐式扩散模型[4]，存在大规模修改内容图像的问题。此外，由于文本提示词难以组织且文本语义的细微变化将生成截然不同的图像内容，AIGC 无法实现稳定可控的图像生成。相比之下，图像风格迁移以内容图像为生成基准并以风格图像为样式参考，其生成图像的可预期性更强，同时也更加符合人们美化既有图像的基本需求。因此，图像风格迁移任务仍然具有较大的发展意义与研究价值。

随着深度学习技术在计算机视觉领域的渗透应用，大量基于卷积神经网络 (convolutional neural network, CNN) 的图像风格迁移方法被提出并取得可观成效，可大致划分为基于优化的方法和基于前馈的方法[5]。

基于优化的方法通过不断迭代更新噪声图像，持续融合内容特征与风格特征，以实现风格化。文献 [6] 采用 Gram 矩阵表征图像风格，基于预训练视觉几何组 19 层（visual geometry group 19-layer, VGG19）网络优化噪声图像，使其呈现内容图像的高层语义，同时具有风格图像的特征相关性，推动了基于优化的图像风格迁移技术研究。然而，低效的迭代过程使此类方法无法满足实际应用需求。

基于前馈的方法通过训练图像转换网络，从而在推理阶段经一次前馈即可生成风格化图像。文献 [7] 引入感知损失，并以无监督方式训练了一个转换网络，实现了实时图像风格迁移。 文献 [8] 提出了一个多尺度纹理生成网络，以完全前馈方式迁移艺术风格，并进一步验证了实例归一化对于提升风格化质量的有效性。 文献 [9] 设计了一个 2 阶段

收稿日期：2023-06-19；**录用日期**：2024-01-19；**网络出版时间**：2024-03-09 07:30
网络出版地址：link.cnki.net/urlid/11.2625.V.20240305.1538.002
基金项目：国家自然科学基金 (62072232)；中央高校基本科研业务费专项资金 (021714380026)；软件新技术与产业化协同创新中心
* **通信作者**. E-mail：beijia@nju.edu.cn

引用格式：纪宗杏，贝佳，刘润泽，等. 基于双路视觉 Transformer 的图像风格迁移 [J]. 北京航空航天大学学报，2025，51（7）：2488-2497. JI Z X，BEI J，LIU R Z，et al. Dual-channel vision Transformer-based image style transfer [J]. Journal of Beijing University of Aeronautics and Astronautics，2025，51（7）：2488-2497 （in Chinese）.

图像风格化过程, 即先迁移全局样式生成一张风格化草图, 再在高分辨率下修补其局部纹理细节, 从而能实时生成高质量迁移图像。上述方法虽能高效执行图像风格化, 但不具备泛化能力, 即需要重新训练模型以适配新风格图像。文献 [10] 通过训练一个预测网络提取面向特定风格的特征缩放参数, 并通过条件实例归一化(conditional instance normalization, CIN)方式将内容特征映射至风格图像域, 得到风格化图像。文献 [11] 以无参数方式调整内容特征的整体分布, 使其匹配风格特征的通道均值与标准差。同样地, 文献 [12] 对内容特征进行白化与着色, 以使其协方差匹配至风格特征。文献 [13] 通过风格注意力机制将局部风格样式灵活融合至内容特征图中, 从而达到风格化效果。文献 [14] 提出了一个新的注意力和归一化模块, 能够在每个点的基础上自适应地执行注意力归一化, 同时考虑了深层特征与浅层特征, 从而在保持内容结构的同时更好地转移风格样式。文献 [15] 通过在风格转移过程中预测卷积核和偏置传递风格图像的统计特征和结构信息, 使风格迁移不仅仅局限于全局特征, 而是能够更细致地处理局结构。

随着视觉 Transformer 的提出, 其全局信息整合能力有效弥补了 CNN 局部表征的短板, 使模型更加关注内容语义的形状与结构。文献 [16] 首次将视觉 Transformer 引入到图像风格迁移任务中, 使用 2 个 Transformer 图像编码器表征内容语义和纹理样式, 并通过 Transformer 图像解码器交叉融合不同图像域特征, 进而转换生成风格化图像。文献 [17] 引入条形窗口注意力用于整合局部信息, 并集成水平与垂直方向上的长距离依赖关系, 从而消除风格化图像中的网格伪影。文献 [18] 引入 2 种注意力机制, 基于自注意力编码图像块序列间的相关性, 并通过交叉注意力将风格样式自适应合成至内容图像中。文献 [19] 引入了一种新的边缘损失, 用于增强内容的细节, 在风格特征过度渲染导致的模糊情况下改善图片的清晰度。文献 [20] 提出了一种新颖的组合式 Transformer 自编码器, 分别处理高相关性和低相关性的特征块, 能够有效地处理和保留内容图像的结构和语义信息。然而, 现有基于视觉 Transformer 的图像风格迁移方法大多存在形状偏向, 忽视了对于目标纹理及色彩分布的精准刻画。

为提升风格迁移质量, 本文提出了一种基于双路视觉 Transformer 的图像风格迁移方法, 利用 Transformer 架构捕获图像长距离依赖关系以建模全局上下文信息, 并引入 Transformer 风格参数提取器对图像风格进行离散化表示, 最终在解码架构中将内容嵌入配准对齐至风格特征, 从而实现图像风格化。本文主要贡献包括 3 个方面: ①提出了一种基于双路视觉 Transformer 的图像风格迁移方法, 针对内容图像域和风格图像域进行独立编码, 捕获图像长距离依赖关系以建模全局上下文信息, 兼具较高的内容保真度与风格还原度; ②引入了一个 Transformer 风格参数提取器, 通过交叉注意力机制提取风格参数, 将图像风格离散化表示为一组关键特征, 提升了风格表征刻画的丰富性和准确性; ③在 Transformer 图像解码阶段显式度量风格分布特征, 基于预测得到的风格特征分布参数反归一化网络中间嵌入, 从而基于特征分布匹配实现特定风格渲染。

1　卷积神经网络和视觉 Transformer

1.1　卷积神经网络

文献 [21] 在 1998 年提出了用于手写数字识别的卷积神经网络 LeNet-5。文献 [22] 提出的 AlexNet 在 ImageNet 竞赛中取得的出色表现, 奠定了 CNN 在计算机视觉领域的地位。

得益于 CNN 强大的非线性表征学习能力, 图像风格迁移技术得到蓬勃发展。基于 CNN 的图像风格迁移方法通常按编码-解码架构划分网络逻辑, 其中编码器大多基于预训练图像分类网络, 其表征能力直接决定了风格呈现质量。这类方法需堆叠大量卷积池化层以获取充足的感受野, 从而将低层视觉特征重组为高层抽象语义。与此同时, 网络量级和复杂度得到同步提升, 且特征分辨率不断降低, 进而无法精准刻画内容语义及纹理样式[16]。此外, 这类方法未考虑不同图像域间的编码差异, 将基于自然图像数据集训练所得的编码器直接应用于编码风格图像, 从而引入表征误差。如文献 [23-24] 中所述, CNN 还存在明显的纹理偏向, 不利于保留内容图像的原始语义结构。

1.2　视觉 Transformer

文献 [25] 在 2020 年提出视觉 Transformer, 将图片裁剪为 16 个切块, 并在切块序列的每个切块上直接应用 Transformer, 在图像分类任务中取得优越性能的同时, 还具有很强的扩展性。文献 [25] 证明了当拥有足够多的数据时, 视觉 Transformer 的表现会超过 CNN, 推动了后续视觉 Transformer 的研究。

随着 Transformer 在计算机视觉领域的成功应用, 图像风格迁移任务被重新赋予生命力。与 CNN

不同，视觉 Transformer 依赖注意力机制在网络浅层即可捕获图像全局信息，建模图像块间的长距离依赖关系，与人类通过形状辨认物体的视觉感知特性相似。然而，视觉 Transformer 无法有效表征各图像块内部的像素相关性，即图像局部特征刻画能力较差，存在形状偏向[24]。现有代表性方法大致遵循原始视觉 Transformer 结构设计，即先基于自注意力机制进行图像编码，再通过计算内容编码与风格编码间的交叉注意力实现特征融合，从而将内容图像映射至目标风格域。这类方法无法准确塑造风格图像中的色彩分布与显著样式，当参考风格中存在精细纹理时，风格整体呈现质量可能不及最先进的 CNN 方法。

1.3 不同风格迁移方法对比

作为图像风格迁移任务中的基本构成要素，形状与纹理分别对应内容和风格，任一偏向类型均不利于实现高质量图像渲染。图 1 对比了本文方法与 3 种现有代表性图像风格迁移方法的风格化效果。其中，AdaIN[11] 和 WCT[12] 基于 CNN，而 StyTr²[16] 基于视觉 Transformer。由此可见，AdaIN[11] 由于忽视内容图像域与风格图像域间的编码差异，其风格表征能力较差，参考样式几乎不可感知；WCT[12] 存在严重的纹理偏向，注重局部纹理刻画而削弱了风格化图像的整体性；StyTr²[16] 具有较强的形状偏向，缺乏纹理感，其色彩分布也与参考风格间存在细微偏差。

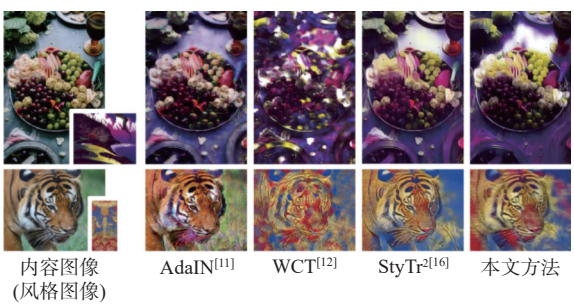


图 1 本文方法与代表性方法的风格化效果对比

Fig. 1 Comparison of stylization results between proposed method and representative methods

2 本文算法

本文提出了一种基于双路视觉 Transformer 的图像风格迁移方法 Bi-Trans，主体架构依托于现有基线方法 StyTr²[16]。Bi-Trans 将图像风格迁移任务流程划分为风格参数预测和风格化图像块序列生成任务。给定内容图像 $\boldsymbol{I}_{\mathrm{c}} \in \mathbf{R}^{H\times W\times 3}$ 和风格图像 $\boldsymbol{I}_{\mathrm{s}} \in \mathbf{R}^{H\times W\times 3}$，最终生成风格化图像 $\boldsymbol{I}_{\mathrm{cs}} \in \mathbf{R}^{H\times W\times 3}$。式中：$H$、$W$ 分别为图像高度和图像宽度。文献 [24] 表明 CNN 存在纹理偏向，而视觉 Transformer 存在形状偏向。为增强图像编码的纹理偏向，Bi-Trans 没有采用 StyTr²[16] 图像划分编码阶段中的线性投影方式，而是将其更换为预训练 VGG19 网络，分别提取 $\boldsymbol{I}_{\mathrm{c}}$、$\boldsymbol{I}_{\mathrm{s}}$ 在 relu4_1 层上的特征编码 $\boldsymbol{F}_{\mathrm{c}}^{4_1} \in \mathbf{R}^{(H/m)\times(W/m)\times C}$ 和 $\boldsymbol{F}_{\mathrm{s}}^{4_1} \in \mathbf{R}^{(H/m)\times(W/m)\times C}$，从而将输入图像划分为 $L=(H/m)\times(W/m)$ 个 $m\times m$ 大小的图像块，得到相应的图像块特征序列 $\boldsymbol{\varepsilon}_{\mathrm{c}}=\{\boldsymbol{\varepsilon}_{\mathrm{c}}^1,\boldsymbol{\varepsilon}_{\mathrm{c}}^2,\cdots,\boldsymbol{\varepsilon}_{\mathrm{c}}^L\}\in\mathbf{R}^{L\times C}$ 和 $\boldsymbol{\varepsilon}_{\mathrm{s}}=\{\boldsymbol{\varepsilon}_{\mathrm{s}}^1,\boldsymbol{\varepsilon}_{\mathrm{s}}^2,\cdots,\boldsymbol{\varepsilon}_{\mathrm{s}}^L\}\in\mathbf{R}^{L\times C}$。其中：$m=8$ 为图像块尺寸；$C=512$ 为特征维度。图 2 展示了本文方法的整体流程。

2.1 可学习的位置编码

为编码图像块间的空间分布关系，将内容图像块特征序列 $\boldsymbol{\varepsilon}_{\mathrm{c}}$ 输入到 Transformer 内容图像编码器之前，先对各内容图像块进行位置编码。本文希望在具有相似语义的内容图像块上施加一致的风格化效果，且同一图像区域的位置编码应不随图像尺度变化而改变。如文献 [16] 中所述，图像块是基于语义上下文和内容相关性进行排列的，而非类似于

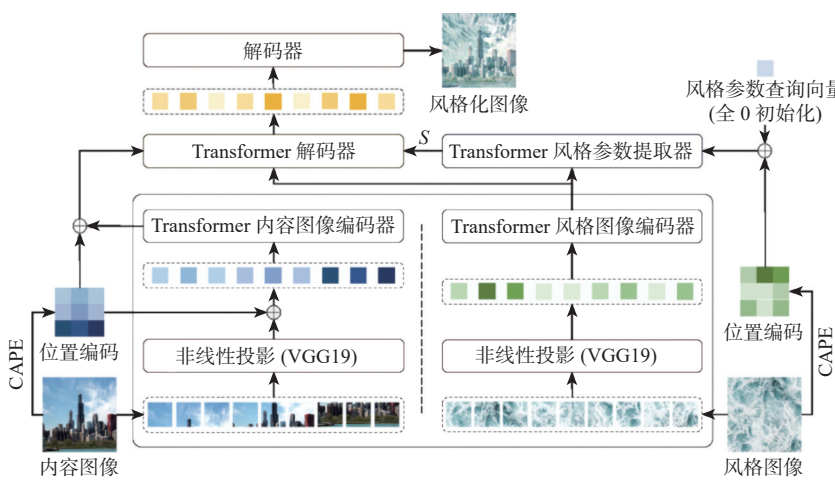


图 2 基于双路视觉 Transformer 的图像风格迁移方法流程图

Fig. 2 Flowchart of image style transfer method based on dual-channel vision Transformer

文本所蕴含的自然逻辑。因此, 本文选择采用文献[16]提出的基于内容感知的位置编码 (content-aware positional embedding, CAPE), 计算过程为

$$\hat{\boldsymbol{P}}_{\mathrm{c}}=\mathrm{Conv1bd}\left(\mathrm{AvgPool}_n\times n\left(\boldsymbol{\varepsilon}_{\mathrm{c}}\right)\right),\tag{1}$$

式中: Conv1×1表示点卷积运算; AvgPool表示平均池化操作; $n=18$为平均池化后的特征图分辨率; $\hat{\boldsymbol{P}}_{\mathrm{c}}\in\mathbf{R}^{n\times n\times C}$为内容图像位置编码。具体, 先通过平均池化将$\boldsymbol{\varepsilon}_{\mathrm{c}}$分辨率降低至$n\times n$; 再通过点卷积操作实现跨通道语义信息融合; 最后, 通过双线性差值将$\hat{\boldsymbol{P}}_{\mathrm{c}}$上采样至与$\boldsymbol{\varepsilon}_{\mathrm{c}}$尺寸相同, 得到内容图像块位置编码序列$\boldsymbol{P}_{\mathrm{c}}=\left\{\boldsymbol{P}_{\mathrm{c}}^{1},\boldsymbol{P}_{\mathrm{c}}^{2},\cdots,\boldsymbol{P}_{\mathrm{c}}^{L}\right\}\in\mathbf{R}^{L\times C}$。该位置编码方式具有语义一致性和尺度不变性等优点, 更适用于图像风格迁移任务。

对于风格图像而言, 本文希望在具有相似语义的图像块上提取出一致的风格参数, 能有效反映其结构特性, 即纹理排列与色调变化。类似地, 对风格图像进行位置编码, 得到$\hat{\boldsymbol{P}}_{\mathrm{s}}\in\mathbf{R}^{n\times n\times D}$。其中, $D=768$为点卷积运算的输出通道数。

2.2　Transformer 图像编码器

文献[24]表明基于 ImageNet 数据集预训练的视觉 Transformer, 因存在特定于分类任务和模型结构的形状偏向, 不适合用作图像风格迁移任务的特征提取器。此外, 如引言中所述, 内容图像和风格图像分属不同的图像域, 若采用通用编码器将引入特征误差。因此, 分别为内容图像域和风格图像域重新设计训练 Transformer 编码器以捕获图像长距离依赖关系, 提取领域特定的图像编码。

Transformer 内容图像编码器共堆叠了N个编码层, 每层包含 1 个多头自注意力 (multi-head self-attention, MSA) 模块和 1 个多层感知 (multi-layer perceptron, MLP) 模块, 其结构如图 3(a) 所示。为融合内容图像块的空间位置信息, 将$\boldsymbol{\varepsilon}_{\mathrm{c}}$与$\boldsymbol{P}_{\mathrm{c}}$相加, 得到内容输入序列$\boldsymbol{E}_{\mathrm{c}}=\left\{\boldsymbol{\varepsilon}_{\mathrm{c}}^{1}+\boldsymbol{P}_{\mathrm{c}}^{1},\boldsymbol{\varepsilon}_{\mathrm{c}}^{2}+\boldsymbol{P}_{\mathrm{c}}^{2},\cdots,\boldsymbol{\varepsilon}_{\mathrm{c}}^{L}+\boldsymbol{P}_{\mathrm{c}}^{L}\right\}\in\mathbf{R}^{L\times C}$, 并计算其 MSA:

$$\begin{aligned}\mathrm{MSA}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=\mathrm{Concat}&\left(\mathrm{Attn}\left(\boldsymbol{Q}\boldsymbol{W}_{1}^{\mathrm{q}},\boldsymbol{K}\boldsymbol{W}_{1}^{\mathrm{k}},\boldsymbol{V}\boldsymbol{W}_{1}^{\mathrm{v}}\right),\cdots,\right.\\&\left.\mathrm{Attn}\left(\boldsymbol{Q}\boldsymbol{W}_{h}^{\mathrm{q}},\boldsymbol{K}\boldsymbol{W}_{h}^{\mathrm{k}},\boldsymbol{V}\boldsymbol{W}_{h}^{\mathrm{v}}\right)\right)\boldsymbol{W}^{\mathrm{o}}\end{aligned}\tag{2}$$

式中: $\boldsymbol{Q}=\boldsymbol{K}=\boldsymbol{E}_{\mathrm{c}}$、$\boldsymbol{V}=\boldsymbol{\varepsilon}_{\mathrm{c}}$分别为查询向量、键向量和值向量; h为多头注意力头个数; $\boldsymbol{W}_{i}^{\mathrm{q}},\boldsymbol{W}_{i}^{\mathrm{k}},\boldsymbol{W}_{i}^{\mathrm{v}}\in\mathbf{R}^{C\times(C/h)}$为第$i$个注意力头对应的参数矩阵; $\boldsymbol{W}^{\mathrm{o}}\in\mathbf{R}^{C\times C}$为输出投影矩阵; Attn表示文献[26]中的注意力操作; Concat表示维度拼接。进一步通过 MLP 模块挖掘内容图像特征的非线性关系, 定义了 2 层全连接, 将注意力输出先升维再降维。对各模块输入输出进行残差连接, 经层归一化 (layer normalization, LN) 后再输入至下一模块, 最终生成内容编码$\boldsymbol{\Phi}_{\mathrm{c}}\in\mathbf{R}^{L\times C}$。因此, Transformer 内容图像编码器每层的计算过程为

$$\begin{cases}\hat{\boldsymbol{\Phi}}_{\mathrm{c}}=\mathrm{LN}(\mathrm{MSA}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})+\boldsymbol{Q})\\ \boldsymbol{\Phi}_{\mathrm{c}}=\mathrm{LN}\left(\mathrm{MLP}\left(\hat{\boldsymbol{\Phi}}_{\mathrm{c}}\right)+\hat{\boldsymbol{\Phi}}_{\mathrm{c}}\right)\end{cases}\tag{3}$$

类似地, 构建 Transformer 风格图像编码器, 其结构与内容图像编码器相同, 以提取风格输入序列$\boldsymbol{E}_{\mathrm{s}}=\left\{\boldsymbol{\varepsilon}_{\mathrm{s}}^{1},\boldsymbol{\varepsilon}_{\mathrm{s}}^{2},\cdots,\boldsymbol{\varepsilon}_{\mathrm{s}}^{L}\right\}\in\mathbf{R}^{L\times C}$特征, 得到风格编码

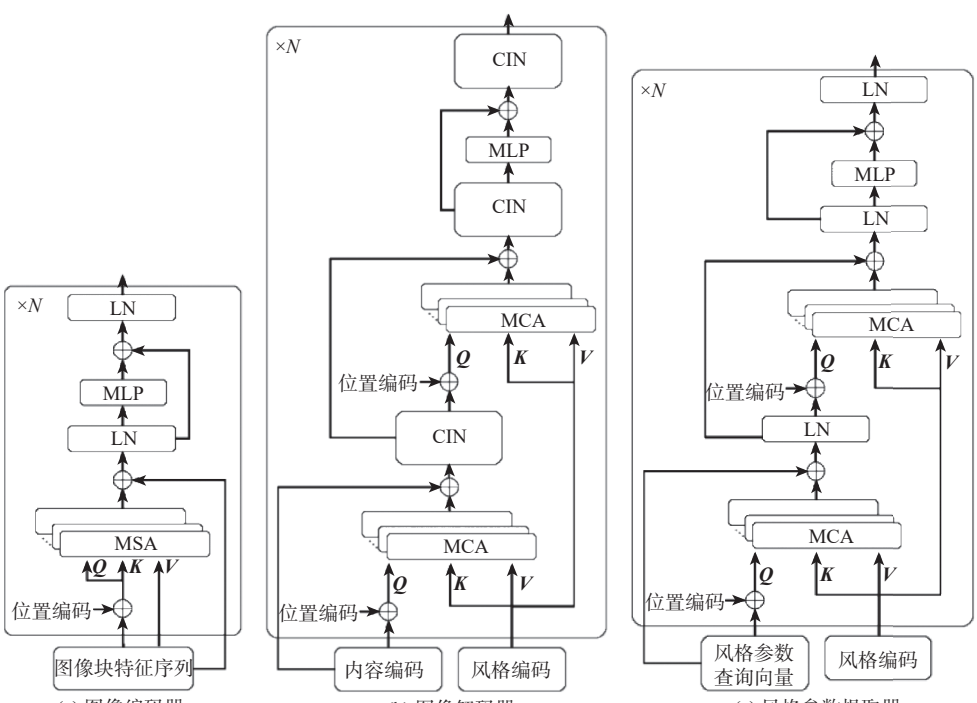


图 3　视觉 Transformer 编解码架构图

Fig. 3　Encoder-decoder architecture of vision Transformer

$\boldsymbol{\Phi}_{\mathrm{s}} \in \mathbf{R}^{L\times C}$。由于风格化图像中无需保留风格图像的语义结构，$\boldsymbol{E}_{\mathrm{s}}$中无需添加风格图像块位置编码。进一步对$\boldsymbol{\Phi}_{\mathrm{s}}$及逆行维度变换并计算其通道均值，得到$\boldsymbol{\Phi}'_{\mathrm{s}} \in \mathbf{R}^{1\times D}$，以作为Transformer风格参数提取器输入。

2.3 Transformer 风格参数提取器

受文献[10]的启发，为实现任意风格迁移，额外设计训练了一个Transformer风格参数提取器，其在隐空间中学习各参数分量与风格编码间的相关性，从而使各参数分量侧重反映目标风格的不同特征，实现图像风格拆解并将其离散化表示为若干个关键特征的组合，如颜色特征、形状特征、纹理特征等。

将风格参数定义为一个D维特征向量$\boldsymbol{S} \in \mathbf{R}^{1\times D}$并将其初始化为全0向量$\boldsymbol{O}$。由于风格参数需反映风格图像的结构信息，先对2.1节中的风格图像位置编码$\hat{\boldsymbol{P}}_{\mathrm{s}}$进行通道均值计算，得到$\boldsymbol{P}_{\mathrm{s}} \in \mathbf{R}^{1\times D}$；再将其与$\boldsymbol{O}$相加作为风格参数提取器的查询向量输入，即$\boldsymbol{Q}=\boldsymbol{O}+\boldsymbol{P}_{\mathrm{s}}$。如图3(c)所示，Transformer风格参数提取器共堆叠了N个提取层，每层包含2个多头交叉注意力(multi-head cross-attention, MCA)模块和1个MLP模块。其中，MCA模块的键向量和值向量均为风格编码$\boldsymbol{\Phi}_{\mathrm{s}}$，即$\boldsymbol{K}=\boldsymbol{V}=\boldsymbol{\Phi}_{\mathrm{s}}$，其计算方式与MSA相同。风格参数提取过程为

$$\begin{cases} \boldsymbol{S}'' = \mathrm{LN}(\mathrm{MCA}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})+\boldsymbol{Q}) \\ \boldsymbol{S}' = \mathrm{LN}(\mathrm{MCA}(\boldsymbol{S}''+\boldsymbol{P}_{\mathrm{s}},\boldsymbol{K},\boldsymbol{V})+\boldsymbol{S}'') \\ \boldsymbol{S} = \mathrm{LN}(\mathrm{MLP}(\boldsymbol{S}')+\boldsymbol{S}') \end{cases} \tag{4}$$

进一步对提取出的风格参数$\boldsymbol{S}$进行降维，以减少冗余特征，增强可理解性，从而将其精简为$\boldsymbol{S} \in \mathbf{R}^{1\times U}$。其中，$U=100$为降维后风格参数向量维度。

2.4 Transformer 图像解码器

Transformer图像解码器基于风格编码$\boldsymbol{\Phi}_{\mathrm{s}}$和风格参数$\hat{\boldsymbol{S}}$，对内容编码$\boldsymbol{\Phi}_{\mathrm{c}}$进行分解与重组，使其保留语义结构特征并融合风格样式特征。与风格参数提取器相同，Transformer图像解码器共堆叠了N个解码层，每层包含2个MCA模块和1个MLP模块，最终生成风格化编码$\boldsymbol{\Phi}_{\mathrm{cs}} \in \mathbf{R}^{L\times C}$，其结构如图3(b)所示，计算过程为

$$\begin{cases} \boldsymbol{\Phi}''_{\mathrm{cs}} = \mathrm{CIN}(\mathrm{MCA}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})+\boldsymbol{Q};\hat{\boldsymbol{S}}) \\ \boldsymbol{\Phi}'_{\mathrm{cs}} = \mathrm{CIN}\left(\mathrm{MCA}\left(\boldsymbol{\Phi}''_{\mathrm{cs}}+\boldsymbol{P}_{\mathrm{c}},\boldsymbol{K},\boldsymbol{V}\right)+\boldsymbol{\Phi}''_{\mathrm{cs}};\hat{\boldsymbol{S}}\right) \\ \boldsymbol{\Phi}_{\mathrm{cs}} = \mathrm{CIN}\left(\mathrm{MCA}\left(\boldsymbol{\Phi}'_{\mathrm{cs}}\right)+\boldsymbol{\Phi}'_{\mathrm{cs}};\hat{\boldsymbol{S}}\right) \end{cases} \tag{5}$$

式中：$\boldsymbol{Q}=\boldsymbol{\Phi}_{\mathrm{c}}+\boldsymbol{P}_{\mathrm{c}}, \boldsymbol{K}=\boldsymbol{V}=\boldsymbol{\Phi}_{\mathrm{s}}$。Transformer图像解码过程即为计算内容图像块与各风格图像块间的相关性，从而根据内容图像块的语义特征选取最适合施加的显著纹理样式。交叉注意力计算有助于建立内容图像域中的语义概念与风格图像域中的纹理样式之间的关联关系，能根据注意力大小决定施加的纹理类型以及纹理施加的强弱程度，从而实现跨图像域风格纹理迁移。

为利用提取到的风格参数$\hat{\boldsymbol{S}}$，将各模块后的LN替换为CIN，通过学习2个缩放平移参数$\boldsymbol{\gamma}_{\hat{S}}$、$\boldsymbol{\beta}_{\hat{S}} \in \mathbf{R}^{1\times C}$来调整特征编码的整体分布，从而将内容特征配准对齐至风格特征，将风格化过程定义为

$$\begin{cases} \mathrm{CIN}(\boldsymbol{x};\hat{\boldsymbol{S}}) = \boldsymbol{\gamma}_{\hat{S}}\left(\dfrac{\boldsymbol{x}-\mu(\boldsymbol{x})}{\sigma(\boldsymbol{x})}\right)+\boldsymbol{\beta}_{\hat{S}} \\ \boldsymbol{\gamma}_{\hat{S}} = \hat{\boldsymbol{S}}\boldsymbol{W}_{\gamma}+\boldsymbol{b}_{\gamma} \\ \boldsymbol{\beta}_{\hat{S}} = \hat{\boldsymbol{S}}\boldsymbol{W}_{\beta}+\boldsymbol{b}_{\beta} \end{cases} \tag{6}$$

式中：$\mu(\boldsymbol{x})$、$\sigma(\boldsymbol{x})$分别是特征编码$\boldsymbol{x}$的通道均值和通道标准差；$\boldsymbol{W}_{\gamma}, \boldsymbol{W}_{\beta} \in \mathbf{R}^{100\times C}$为投影矩阵；$\boldsymbol{b}_{\gamma}$、$\boldsymbol{b}_{\beta}$为偏置项。

进一步对Transformer解码输出进行维度变换，将其由特征空间重新映射至像素空间。受文献[24]和文献[16]的启发，为避免引入感知不一致性与边缘伪影，并未直接上采样Transformer解码输出，而是定义了一个3层CNN解码器，对其进行特征提炼并中和其形状偏向，最终生成风格化图像$\boldsymbol{I}_{\mathrm{cs}}$。其中，CNN解码器每层通过一个2倍最近邻上采样来扩大编码尺寸，并通过若干个3×3卷积来压缩特征通道。

2.5 损失网络

在模型训练阶段，为专注比较风格化图像与内容图像以及与风格图像的感知差异，仍采用预训练VGG19网络提取图像特征，利用CNN的纹理偏向特性促使模型捕获刻画更加精细的风格纹理，从而对图像编码器、风格参数提取器和图像解码器进行参数优化，以使风格化图像尽可能保留原始内容语义，同时呈现出特定纹理样式。本文遵循文献[24]的观点，使用感知损失来评估$\boldsymbol{I}_{\mathrm{cs}}$与$\boldsymbol{I}_{\mathrm{c}}$的语义结构相似性，同时衡量其与$\boldsymbol{I}_{\mathrm{s}}$的视觉外观相似性，从而弱化形状偏向，生成更加真实多样的局部纹理细节。

定义relu4_2为内容损失计算层，提取$\boldsymbol{I}_{\mathrm{cs}}$、$\boldsymbol{I}_{\mathrm{c}}$在其上的特征编码$\boldsymbol{F}^{4_2}_{\mathrm{cs}}$和$\boldsymbol{F}^{4_2}_{\mathrm{c}}$，将内容损失定义为图像特征编码间的均方差：

$$\mathscr{L}_{\mathrm{con}} = \frac{1}{|\boldsymbol{\Psi}_{\mathrm{c}}|}\sum_{l\in\boldsymbol{\Psi}_{\mathrm{c}}}\frac{1}{H^l W^l}\sum_{i=1}^{H^l}\sum_{j=1}^{W^l}\left(F^l_{\mathrm{cs},ij}-F^l_{\mathrm{c},ij}\right)^2 \tag{7}$$

式中：$\boldsymbol{\Psi}_{\mathrm{c}}=\{4_2\}$为内容层；$H^l$、$W^l$分别为$l$层上的

特征图的高和宽; $F^l_{cs,ij}$为特征图$\boldsymbol{F}^l_{cs}$中(i,j)位置的特征值; $|\cdot|$为集合基数。

在 relu1_1 到 relu4_1 层上将$\boldsymbol{I}_{cs}$、$\boldsymbol{I}_s$编码为$\boldsymbol{F}_{cs}$、$\boldsymbol{F}_s$, 并分别计算得到图像编码对应的分布特征$\{\mu_{cs},\sigma_{cs}\}$和$\{\mu_s,\sigma_s\}$, 将分割损失定义为 2 对分布参数的均方差之和:

$$\mathcal{L}^{\mu,\sigma}_{sty}=\frac{1}{|\boldsymbol{\Psi}_s|}\sum_{l\in\boldsymbol{\Psi}_s}\left(\mu^l_{cs}-\mu^l_s\right)^2+\left(\sigma^l_{cs}-\sigma^l_s\right)^2 \tag{8}$$

式中: $\boldsymbol{\Psi}_s=\{1_1,2_1,3_1,4_1\}$为风格层; μ^l_{ts}、σ^2_{cs}分别为特征图$\boldsymbol{F}^l_{cs}$的通道均值和通道标准差。

为提升图像表征性能, 额外引入 2 个恒等损失项[13], 其能有效限制不必要的图像变动, 并要求当内容图像与风格图像相同时, 风格化图像应为输入图像本身。固定输入为 2 张内容图像 (风格图像), 经双路视觉 Transformer 编解码, 最终生成风格化图像$\boldsymbol{I}_{cc}$($\boldsymbol{I}_{ss}$)。计算风格化图像与输入图像间的像素差异以评估其视觉语义相似性:

$$\mathcal{L}^1_{id}=\frac{1}{HW}\sum_{i=1}^{H}\sum_{j=1}^{W}\left(I^{ij}_{cc}-I^{ij}_c\right)^2+\left(I^{ij}_{ss}-I^{ij}_s\right)^2 \tag{9}$$

式中: I^{ij}_{cc}表示图像$\boldsymbol{I}_{cc}$在(i,j)处的像素值。

由于像素损失无法精准衡量图像间的视觉语义相似性, 即当存在像素位移时, 视觉语义相差较小的图像可能对应较大的像素损失, 从而造成模型误判。因此, 进一步在特征空间中计算风格化图像和输入图像间的编码差异:

$$\mathcal{L}^2_{id}=\frac{1}{|\boldsymbol{\Psi}_{id}|}\sum_{l\in\boldsymbol{\Psi}_{id}}\frac{1}{H^lW^l}\sum_{i=1}^{H^l}\sum_{j=1}^{W^l}\left(F^l_{cc,ij}-F^l_{c,ij}\right)^2+\left(F^l_{ss,ij}-F^l_{s,ij}\right)^2 \tag{10}$$

式中: $\boldsymbol{\Psi}_{id}=\{1_1,2_1,3_1,4_1\}$为恒等损失层。

最终, 将 4 个分支损失加权求和, 得到总训练损失:

$$\mathcal{L}_{tot}=\lambda_c\mathcal{L}_{con}+\lambda_s\mathcal{L}^{\mu,\sigma}_{sty}+\lambda^1_{id}\mathcal{L}^1_{id}+\lambda^2_{id}\mathcal{L}^2_{id}, \tag{11}$$

式中: λ_c、λ_s、λ^1_{id}和λ^2_{id}为各分支损失权重(具体取值详见 3.1 节)。

3　实验与结果分析

3.1　数据集与实验设置

使用 MSCOCO 数据集[27]和 WikiArt 数据集[28]提供训练所需的内容图像及风格图像。在训练阶段, 先将输入图像统一缩放至 280×280, 再对其随机裁剪, 从而控制输入分辨率为 256×256。堆叠层数N、注意力头个数h分别设定为 3 和 8, 各分支损失权重λ_c、λ_s、λ^1_{id}和λ^2_{id}设置为 7、10、70 和 1。以 4 为批尺寸训练模型共 16 万次迭代, 指定 Adam 为优化器并设置其初始学习率为 5×10^{-4}。针对前 10 000 次训练迭代, 采用学习率预热策略, 随后每轮迭代均进行学习率衰减, 衰减系数为 1×10^{-5}。实验基于一台配置 Xeon@2.40 GHz 处理器、64 GB 内存和 Nvidia RTX 3090 24 GB 显卡的计算机。

3.2　风格任意性实验

经充分训练, Transformer 风格参数提取器能从任意风格编码中提取出足以表征目标样式的参数向量, 并将其应用于 CIN, 以标定内容特征。从 WikiArt 官网随机下载 8 张绘画作品, 以测试本文方法对于未知风格的泛化性能。如图 4 所示, 可见: ① 本文方法能精准刻画任意参考风格, 能有效提取并重塑风格图像中的色彩分布以及纹理样式; ②得益于内容感知位置编码, 本文方法能在相同语义上产生一致的风格化效果, 有效避免了边缘伪影与过度风格化问题; ③本文方法具有较高的内容保真度, 风格化图像中原始内容结构与语义信息依旧清晰可辨认。

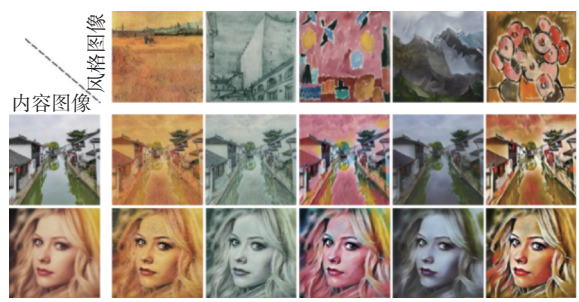


图 4　本文方法任意风格迁移效果示例

Fig. 4　Examples of arbitrary style transfer using proposed method

3.3　对比实验

将 Bi-Trans 与 6 个具有代表性的图像风格迁移方法作对比。其中, 基于 CNN 的方法包括 Ghiasi[10]、AdaIN[11]和 SANet[13], 而基于视觉 Transformer 的方法包括 StyTr^2[16]、S2WAT[17]和 STTR[18]。

3.3.1　定性对比

图 5 展示了本文方法与 6 个对比方法的风格迁移效果。

1) StyTr^2[16]由于采用简单线性投影划分表征图像块, 且仅在像素空间中编码图像区域相关性, 其风格表现力不及本文方法 (如第 4 行)。此外, StyTr^2[16]还倾向于将纹理样式集中迁移至显著物体, 而使图像背景呈现弱风格化效果 (如第 2 行、第 3 行)。

2) S2WAT[17]在不同输入组合下的迁移质量差异较大, 由于将注意力计算限制在横向、纵向及局部窗口内, 存在感受野不足问题, 整体内容结构及语义边缘因施加纹理样式而被弱化 (如第 4 行、第 5 行)。

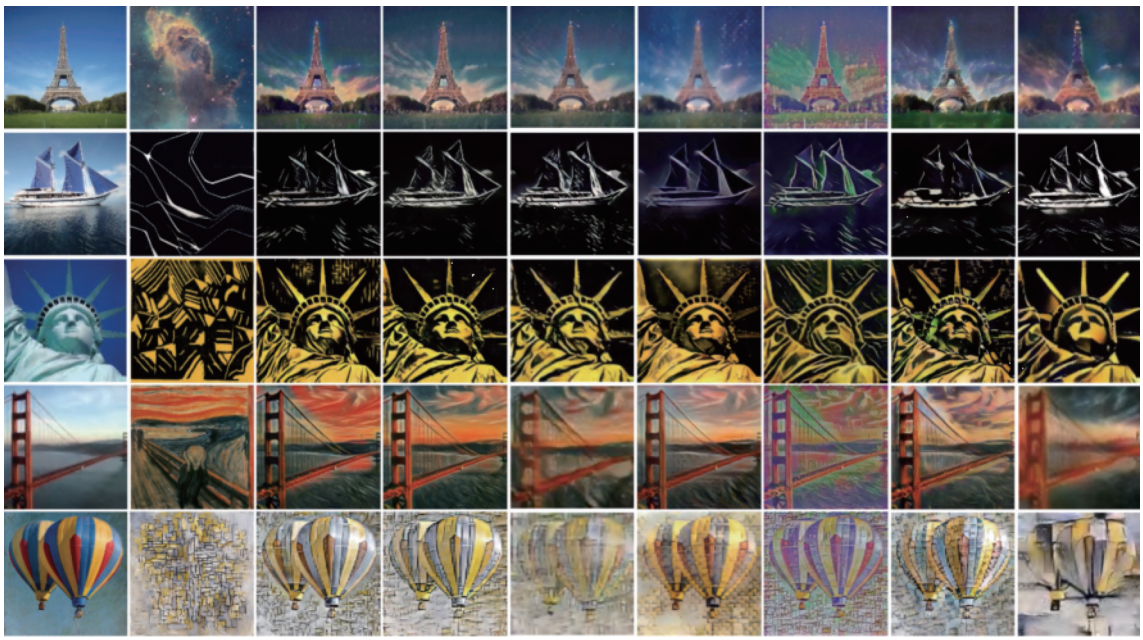


图5 本文方法与现有方法的风格迁移效果对比

Fig. 5 Comparison of style transfer results between proposed method and existing methods

3) STTR[18]倾向于保留精细的轮廓结构以尽可能避免内容失真,其风格化效果主要体现在色彩分布,缺乏显著纹理(如第2行、第4行)。此外,STTR[18]对于目标样式的刻画偏离真实纹理(如第5行),使其风格化质量整体落后于其他基于视觉Transformer的方法。

4) Ghiasi[10]的风格表征能力最差,仅适用于迁移纹理色彩相对单一的简单样式(如第2行、第3行),对于复杂风格则存在严重的色彩偏差。

5) 由于AdaIN[11]直接使用风格特征的分布参数对内容特征进行全局缩放,其存在风格化不完全不充分问题,风格化图像中残留少部分内容伪影(如第2行、第3行),且对于复杂纹理的重塑能力欠佳(如第4行)。

6) SANet[13]存在纹理样式施加的合理性问题,目标风格被非均匀地迁移至语义显著物上而产生割裂效果(如第1行),且同一纹理具有语义异质性(如第4行中将天空纹理迁移至桥梁)。此外,SANet[13]无法准确迁移复杂风格(如第5行),特别是当源域和目标域存在较大的视觉结构差异时。

7) 本文方法与同样基于视觉Transformer的方法相比,无论是风格可感知性还是内容可辨识度,均取得可比甚至更优的风格化效果,其能精准刻画显著样式特征,施加均匀且一致的风格化效果;本文方法与基于CNN的方法相比,得益于基础架构的表征能力,更加忠实于原始内容与参考风格,语义纹理间过渡也更加平滑自然。

3.3.2 定量对比

与文献[16]相同,采用感知损失作为风格化质量评价指标,分别使用式(7)和式(8)来衡量生成图像与输入图像间的内容差异及风格差异,值越小表示迁移质量越高。从MSCOCO测试集[27]中随机选取了20张内容图像,从WikiArt官网下载了20张不同风格不同流派的艺术画作,从而构成400个内容-风格图像对,并计算各方法在所有图像对上的平均内容和平均风格损失。如表1、表2前3列所示,本文方法在2个评价指标上的平均损失值均最小,进而验证了本文方法相对于现有代表性方法的先进性,表明本文方法能同时兼顾内容语义保留与纹理样式重塑。

此外,与上述定性对比结果一致,SANet[13]因执行异质化风格迁移,其平均内容损失最大。

STTR[18]因避免内容失真而执行弱风格化,其平均风格损失最大;StyTr²[16]在2个评价指标上均仅次于本文方法,表明其作为基于视觉Transformer的图像风格迁移基线方法的有效性。

除使用均值、标准差等统计学概念来表征图像风格外,本文还引入风格迁移任务中常用的Gram矩阵,其主对角线元素表示各响应层的特征强度,其他元素表示各响应层间的相关系数,从而将图像风格描述为底层特征间的组合关系。同样地,在风格层$\boldsymbol{\Psi}_{\mathrm{s}}$上计算基于Gram矩阵的风格损失:

$$\mathscr{L}_{\mathrm{sty}}^{\mathrm{Gram}}=\frac{1}{|\boldsymbol{\Psi}_{\mathrm{s}}|}\sum_{l\in\boldsymbol{\Psi}_{\mathrm{s}}}\frac{1}{(C^l)^2}\sum_{i=1}^{C^l}\sum_{j=1}^{C^l}\left(G_{\mathrm{cs},ij}^l-G_{\mathrm{s},ij}^l\right)^2 \tag{12}$$

表 1　本文方法与基于 CNN 的方法平均风格化损失对比

Table 1　Comparison of average stylization loss between proposed method and CNN-based methods

方法	$\mathcal{L}_{con}$	$\mathcal{L}_{sty}^{\mu,\sigma}$	$\mathcal{L}_{sty}^{Gram}$
Ghiasi[10]	0.94	2.81	12.51
AdaIN[11]	0.91	1.39	11.63
SANet[13]	0.97	2.88	16.18
本文方法	0.69	1.12	7.31

表 2　本文方法与基于视觉 Transformer 的方法平均风格化损失对比

Table 2　Comparison of average stylization loss between proposed method and vision Transformer based methods

方法	$\mathcal{L}_{con}$	$\mathcal{L}_{sty}^{\mu,\sigma}$	$\mathcal{L}_{sty}^{Gram}$
StyTr^2[16]	0.79	1.35	9.86
S2WAT[17]	0.93	2.64	13.47
STTR[18]	0.79	3.86	27.58
本文方法	0.69	1.12	7.31

式中：C^l为 l层上的特征通道数；$G^l_{cs,ij}=\boldsymbol{F}^l_{cs,i}\boldsymbol{F}^l_{cs,j}$；$\boldsymbol{F}^l_{cs,i}$为特征图$\boldsymbol{F}^l_{cs}$的第$i$个响应层。如表 1、表 2 最后 1 列所示，各方法在新风格度量指标上的定量计算结果大致遵循其在原有指标上的表现，即本文方法风格差异最小，STTR[18] 风格损失值最大，进一步佐证了本文方法的有效性。

3.4　消融实验

开展消融实验对方法不同组件进行有效性验证，具体包括风格参数提取器、非线性投影、CNN 解码器及恒等损失，相应风格化结果如图 6 后 4 列所示。此外，基于 3.3 节中构建的内容-风格图像对和 3 项评价指标定量对比了不同消融设置下的方法。如表 3 所示，完整方法 2 项差异均最小，可见，本文方法各组成部分缺一不可。此外，CNN 解码器与恒等损失对于实施准确稳定的风格化影响程度较大。

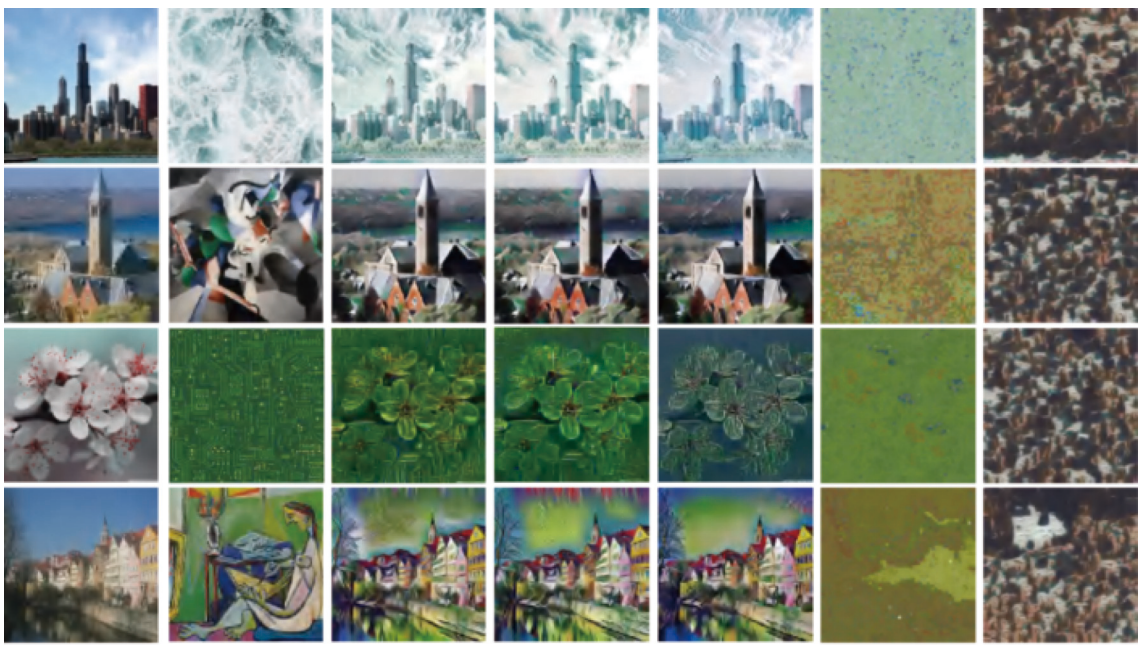


图 6　不同消融设置下的风格迁移效果对比

Fig. 6　Comparison of style transfer results under different ablation settings

表 3　不同消融设置下的平均风格化损失对比

Table 3　Comparison of average stylization loss under different ablation settings

方法	$\mathcal{L}_{con}$	$\mathcal{L}_{sty}^{\mu,\sigma}$	$\mathcal{L}_{sty}^{Gram}$
消融(a)	0.70	1.02	3.15
消融(b)	0.64	1.37	4.71
消融(c)	1.88	11.36	18.61
消融(d)	1.93	6.75	16.05
完整方法	0.69	0.94	3.02

3.4.1　风格参数提取器

为验证 Transformer 风格参数提取器的表征能力及风格参数的有效性，直接对风格编码$\boldsymbol{\Phi}_s$进行通道降维，以作为 Transformer 图像解码器中使用的风格参数$\hat{\boldsymbol{S}}$，即移除 Transformer 风格参数提取器，作为消融 (a) 方法。如图 6 消融 (a) 所示，消融 (a) 方法的风格化结果仍能有效保留内容语义并呈现特定风格，但整体纹理感不及完整方法，特别是图像背景样式被削弱 (如第 1 行、第 3 行)，与 StyTr^2[16] 效果相似。然而，这也侧面印证了将 LN 替换为 CIN 的可行性以及通过特征配准对齐实现图像风格迁移的有效性。

3.4.2　非线性投影

采用预训练 VGG19 网络进行图像块划分与编

码，以中和视觉 Transformer 的形状偏向。为进行有效性验证，将非线性投影替换为原简单线性投影方式，作为消融 (b) 方法，即对输入图像进行一次卷积操作，将卷积核大小、卷积步长和输出通道数分别设置为 m、m、C，从而得到图像块特征序列 $\boldsymbol{\varepsilon}_c$ 和 $\boldsymbol{\varepsilon}_s$。如图 6(b) 所示，消融 (b) 方法的风格化结果无论是色彩分布还是纹理样式均偏离参考风格 (如第 1 行、第 3 行)，可见，非线性投影能够提升模型对纹理色彩的学习能力，更加契合风格迁移任务目标。

3.4.3 CNN 解码器

如文献 [11] 中所述，直接将 Transformer 解码输出上采样至图像空间，会产生严重的棋盘效应，使图像块间的样式区分度较低。本文将 CNN 解码器替换为由全连接层、ReLU 激活与上采样层构成的多层感知机，作为消融 (c) 方法，其风格化结果如图 6(c) 所示，验证了 CNN 解码器相对于直接上采样的优越性。

3.4.4 恒等损失

本文方法在训练阶段共包含 4 项分支损失，其中 $\mathcal{L}_{con}$ 和 $\mathcal{L}_{sty}$ 为图像风格迁移任务中的必要损失，分别用于保留内容与融合样式。本文移除 2 个恒等损失项 $\mathcal{L}_{id}^1$ 和 $\mathcal{L}_{id}^2$，作为消融 (d) 方法，以验证其对风格化效果的影响，如图 6(d) 所示，当移除恒等损失时，消融 (d) 方法的迁移图像存在内容失真与风格漂移问题，模型无法学习到施加纹理样式的合理位置，引入大量冗余变动而导致内容语义紊乱。

4 结 论

1) 结合 CIN 在任意风格迁移中的成功经验，并借助视觉 Transformer 强大的表达能力，提出了一种基于双路视觉 Transformer 的任意图像风格迁移方法 Bi-Trans，相较于基准方法，增强图像块编码的纹理偏向，更加准确地塑造目标风格样式的色彩分布与显著纹理，解决了现有基于 CNN 的图像风格迁移方法中存在的感受野有限、混淆图像域及纹理偏向等问题，兼具较高的内容保真度与风格还原度。

2) 实验结果表明了本文方法相较于现有代表性方法的先进性与优越性，验证了本文方法各组成部分对于实现高质量图像风格迁移的不可或缺性与有效性。

3) 尽管本文方法堆叠较少的网络层即可有效获得全局感受野，但相较于基于 CNN 的图像风格迁移方法，存在计算复杂度高等问题，其在 Transformer 内容图像编码阶段、Transformer 风格图像编码阶段及 Transformer 图像解码阶段均涉及到两两图像块间的相关性计算，在图像渲染效率上有待进一步提升。

参考文献（References）

[1] ZHANG Y, HUANG N, TANG F, et al. Inversion-based style transfer with diffusion models[C]//IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2023: 10146-10156.

[2] ZHANG Y X, DONG W M, TANG F, et al. ProSpect: prompt spectrum for attribute-aware personalization of diffusion models[J]. ACM Transactions on Graphics, 2023, 42(6): 1-14.

[3] WANG Z, ZHAO L, XING W. Stylediffusion: controllable disentangled style transfer via diffusion models[C]//IEEE International Conference on Computer Vision. Piscataway: IEEE Press, 2023: 7677-7689.

[4] ROMBACH R, BLATTMANN A, LORENZ D, et al. High-resolution image synthesis with latent diffusion models[C]//IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2022: 10674-10685.

[5] JING Y C, YANG Y Z, FENG Z L, et al. Neural style transfer: a review[J]. IEEE Transactions on Visualization and Computer Graphics, 2020, 26(11): 3365-3385.

[6] GATYS L A, ECKER A S, BETHGE M. Image style transfer using convolutional neural network [C]//IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2016: 2414-2423.

[7] JOHNSON J, ALAHI A, LI F F. Perceptual losses for real-time style transfer and super-resolution[C]//Computer Vision-ECCV . Cham: Springer, 2016: 694-711.

[8] ULYANOV D, LEBEDEV V, VEDALDI A, et al. Texture networks: feed-forward synthesis of textures and stylized images [EB/OL]. (2016-03-10)[2023-01-10]. https://arxiv.org/abs/1603.03417v1.

[9] LIN T W, MA Z Q, LI F, et al. Drafting and revision: Laplacian pyramid network for fast high-quality artistic style transfer[C]//IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2021: 5137-5146.

[10] GHIASI G, LEE H, KUDLUR M, et al. Exploring the structure of a real-time arbitrary neural artistic stylization network[C]//British Machine Vision Conference. Great Britain: BMVA, 2017: 1-27.

[11] HUANG X, BELONGIE S. Arbitrary style transfer in real-time with adaptive instance normalization[C]//IEEE International Conference on Computer Vision. Piscataway: IEEE Press, 2017: 1510-1519.

[12] LI Y, FANG C, YANG J, et al. Universal style transfer via feature transforms[C]//Annual Conference on Neural Information Processing Systems. La Jolla: NIPS, 2017: 1-11.

[13] PARK D Y, LEE K H. Arbitrary style transfer with style-attentional networks[C]//IEEE/CVF Conference on Computer Vision and Pattern Recognition . Piscataway: IEEE Press, 2019: 5873-5881.

[14] LIU S H, LIN T W, HE D L, et al. AdaAttN: revisit attention mechanism in arbitrary neural style transfer[C]//IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE Press, 2021: 6629-6638.

[15] CHANDRAN P, ZOSS G, GOTARDO P, et al. Adaptive convolu-

tions for structure-aware style transfer [C]//IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2021: 7968-7977.

[16] DENG Y Y, TANG F, DONG W M, et al. StyTr2: image style transfer with transformers[C]//IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2022: 11316-11326.

[17] ZHANG C, YANG J, WANG L, et al. S2WAT: image style transfer via hierarchical vision transformer using Strips Window Attention[EB/OL]. (2022-11-07)[2023-06-19]. https://arxiv.org/abs/2210.12381.

[18] WANG J B, YANG H, FU J L, et al. Fine-grained image style transfer with visual transformers[C]//Computer Vision-ACCV. Cham: Springer, 2023: 427-443.

[19] ZHANG C Y, DAI Z Y, CAO P, et al. Edge enhanced image style transfer *via* transformers[C]//Proceedings of the ACM International Conference on Multimedia Retrieval. New York: ACM, 2023: 105-114.

[20] FENG J X, ZHANG G, LI X H, et al. A compositional transformer based autoencoder for image style transfer[J]. Electronics, 2023, 12(5): 1184.

[21] LECUN Y, BOTTOU L, BENGIO Y, et al. Gradient-based learning applied to document recognition[J]. Proceedings of the IEEE, 1998, 86(11): 2278-2324.

[22] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. ImageNet classification with deep convolutional neural networks[J]. Communications of the ACM, 2017, 60(6): 84-90.

[23] GEIHOS R, RUBISCH P, MICHALIS C, et al. ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness[C]//Proceedings of the International Conference on Learning Representations. Washington DC: ICLR, 2019: 1-22.

[24] WEI H P, DENG Y Y, TANG F, et al. A comparative study of CNN- and transformer-based visual style transfer[J]. Journal of Computer Science and Technology, 2022, 37(3): 601-614.

[25] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, et al. An image is worth 16x16 words: transformers for image recognition at scale [EB/OL]. (2021-06-03)[2023-06-19]. https://arxiv.org/abs/2010.11929.

[26] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]//Advances in Neural Information Processing Systems. La Jolla: NIPS, 2017: 5998-6008.

[27] LIN T Y, MAIRE M, BELONGIE S, et al. Microsoft COCO: common objects in context[C]//Computer Vision-ECCV. Cham: Springer, 2014: 740-755.

[28] PHILLIPS F, MACKINTOSH B. Wiki art gallery, inc. a case for critical thinking[J]. Issues in Accounting Education, 2011, 26(3): 593-608.

Dual-channel vision Transformer-based image style transfer

JI Zongxing，BEI Jia*，LIU Runze，REN Tongwei

(State Key Laboratory for Novel Software Technology, Nanjing University，Nanjing 210093，China)

Abstract：Image style transfer aims to adjust the visual properties of a content image based on a style reference image, preserving the original content while presenting specific styles to generate visually appealing stylized images. Most existing representative methods focus on extracting local image features without considering the encoding differences between different image domains or the importance of global contextual information. To address this issue, Bi-Trans, a novel image style transfer method based on a dual-channel vision Transformer was proposed. This method encoded the content and style image domains independently, extracting style parameter vectors to discretely represent the image style. By using a cross-attention mechanism and conditional instance normalization (CIN), the content image was calibrated to the target style domain, generating the stylized image. Experimental results demonstrate that the proposed method is superior to existing methods in terms of both content preservation and style restoration.

Keywords：image style transfer; vision Transformer; arbitrary stylization; conditional instance normalization; attention mechanism

Received：2023-06-19；**Accepted**：2024-01-19；**Published Online**：2024-03-09 07：30

URL：link.cnki.net/urlid/11.2625.V.20240305.1538.002

Foundation items：National Science Foundation of China (62072232)；The Fundamental Research Funds for the Central Universities (021714380026)；The Collaborative Innovation Center of Novel Software Technology and Industrialization

* **Corresponding author**. E-mail：beijia@nju.edu.cn

http://bhxb.buaa.edu.cn jbuaa@buaa.edu.cn
DOI: 10.13700/j.bh.1001-5965.2023.0394

基于旋转目标感知网络的SAR船舶检测方法

王梓懿[1,2]，尹嘉豪[1,2]，黄博斌[1,2]，高峰[1,2,*]

(1. 中国海洋大学 计算机科学与技术学院，青岛 266100；
2. 中国海洋大学 青岛市海洋信息挖掘与推理工程研究中心，青岛 266100)

摘 要：目标尺寸变化多样且干扰因素多，目标有多种方向且训练样本数据量有限是当前合成孔径雷达（SAR）船舶检测方法主要面临的2个难题。为此，提出了一种用于SAR图像船舶检测的旋转目标感知网络RCAR-Net。主干网络使用基于多尺度Transformer架构的PVTv2，可以更好地保留特征图的局部连续性，同时更好地融合图像的多尺度特征；将旋转边界框与RetinaNet结合，有效减少了背景冗余以及噪声的干扰；引入Cutout方法进行数据增强，用现有样本的部分遮挡来扩大数据集，提高模型的鲁棒性和泛化能力；为了在保证检测精度的同时节省计算和内存开销，使用高效的CARAFE算子对低分辨率的特征图进行上采样，提高多尺度融合效果。RCAR-Net在SSDD和HRSID这2个SAR船舶检测数据集的平均精度分别达到93.63%和90.37%，明显优于DPAN、PANet等方法，对于目标尺寸变化和噪声干扰具有较强的适应性。

关 键 词：合成孔径雷达；船舶检测；旋转边界框；深度学习；注意力机制
中图分类号：TP391
文献标志码：A **文章编号**：1001-5965(2025)07-2498-08

合成孔径雷达(synthetic aperture radar, SAR)由于其具有全天候和全天时工作的特性，在地球观测中得到了广泛应用[1]。由于SAR图像中的船舶检测对军事目标分析、海上救援、海上运输和海洋渔业等都有重要意义[2]，因此设计高效鲁棒的船舶目标检测算法意义重大。传统的SAR图像目标检测方法主要依赖于研究人员通过先验知识手动设置特征，但一直存在检测效率低、算法复杂、鲁棒性差和难以迁移等缺点[3]。

近年来，因基于深度学习的SAR图像船舶目标检测方法具有较强的鲁棒性和泛化能力，无需手动设计特征，受人为因素的影响较小，已成为SAR图像分析与应用领域的主流方法。在大量数据训练下，可以挖掘出传统方法难以得到的性质。2020年，Zhang等[4]发布了SAR船舶目标检测的数据集，引起了遥感领域科研人员的广泛关注。随后，Wei等[5]公布了更大规模的HRSID数据集，该数据集包含5 604张高分辨率SAR图像和16 951个船舶目标。在深度神经网络结构方面，目前主流的目标检测算法分为以R-CNN[6]为代表的二阶段目标检测算法和以YOLO[7]、RetinaNet[8]为代表的单阶段目标检测算法两类。二阶段目标检测算法用到了候选区域的概念，第一步利用滤波窗口在图像上滑动提取可能包含目标的区域，然后利用神经网络对提取区域的目标进行分类和定位。Li等[9]引入Faster R-CNN，用于检测多种复杂环境中的船舶。但是，二阶段目标检测算法的速度相对较慢，难以实现实时检测，因此，当前更多的模型采用了单阶段目标检测算法。单阶段目标检测算法基于回归的思想，直接将待检测图片输入网络，不需要提取候选区域，在输出层得到待检测图片中的边界框及类别。基于RetinaNet，Wang等[10]实现了船舶目标的自动实

收稿日期：2023-06-19；**录用日期**：2023-09-23；**网络出版时间**：2023-11-28 14:26
网络出版地址：link.cnki.net/urlid/11.2625.V.20231127.1912.001
基金项目：科技创新2030-新一代人工智能重大项目(2022ZD0117202)；青岛市自然科学基金(23-2-1-222-zyyd-jch)
*通信作者. E-mail：gaofeng@ouc.edu.cn

引用格式：王梓懿，尹嘉豪，黄博斌，等. 基于旋转目标感知网络的SAR船舶检测方法[J]. 北京航空航天大学学报，2025，51(7)：2498-2505. WANG Z Y，YIN J H，HUANG B B，et al. A rotated content-aware retina network for SAR ship detection[J]. Journal of Beijing University of Aeronautics and Astronautics，2025，51(7)：2498-2505 (in Chinese).

时检测。Zhou 等[11] 构建了一个轻量级的混合特征提取网络, 能够高效定位船舶目标。

尽管当前的 SAR 船舶检测方法已经取得了良好的性能, 但仍然面临一些有挑战性的问题:

1) 目标尺寸变化多样且干扰因素多。不同类型船舶的尺寸有很大的差异, 是影响探测精度的一个重要因素, 为此, 如何更好地提取并融合

图像的多尺度信息成为影响目标检测算法性能的关键问题。SAR 图像中存在的相干斑点噪声和海岸区域对船舶目标的准确识别造成了明显的干扰[12], 这对模型的鲁棒性提出了更高的要求。

2) 目标有多种方向且训练样本数量有限。由于船舶目标具有密集分布、方向任意和横纵比例大的特点, 水平的边界框难以精确地匹配, 且容易引入大量背景像素及噪声。同时, 卫星传感器获取的图像尺寸一般比普通相机产生的图像尺寸大很多, 需要考虑模型的计算和储存成本。

为了解决上述问题, 本文提出了一种用于 SAR 图像船舶目标检测的旋转目标感知网络(rotated content-aware retina network, RCAR-Net)。具体而言, 为了解决目标尺寸变化多样的问题, 主干网络使用基于多尺度 Transformer 架构的 PVTv2, 可以更好地保留特征图的局部连续性, 同时更好地融合图像的多尺度特征。然后, 为了解决目标有多种方向且训练样本数量有限的问题, 本文将旋转边界框与 RetinaNet 结合, 有效减少了背景冗余以及噪声的干扰。同时, 本文引入 Cutout 方法进行数据增强, 用现有样本的部分遮挡来扩大数据集, 提高模型的鲁棒性和泛化能力。最后, 为了在保证检测精度的同时节省计算和内存开销, 本文使用高效的 CARAFE 算子对低分辨率的特征图进行 2 倍的上采样, 提高多尺度融合效果。

本文的主要贡献总结如下:

1) 将基于多尺度 Transformer 的 PVTv2 作为主干网络, 能够有效地提取高分辨率图像的多尺度特征, 并有效保留局部特征。使用 CARAFE 算子优化 FPN 中的上采样方法, 在扩大模型感受野的同时引入很少的计算开销。

2) 使用旋转边界框以解决传统水平边界框难以分离和会引入大量噪声的问题。数据增强方面使用 Cutout, 能使模型更好地利用图像的全部上下文, 增强模型的鲁棒性和泛化能力。

3) 本文方法在 SSDD 和 HRSID 这 2 个 SAR 目标检测数据集上分别取得了 93.63% 和 90.37% 的平均准确率, 优于 DAPN、PANet 等当前主流方法。

1　本文方法

1.1　总体框架

本文提出的 RCAR-Net 主要由 4 个模块组成: 预处理模块, 基于 PVTv2 的主干网络, 基于 CRFPN 的多尺度特征融合模块, 旋转自适应的目标检测模块。总体架构如图 1 所示。

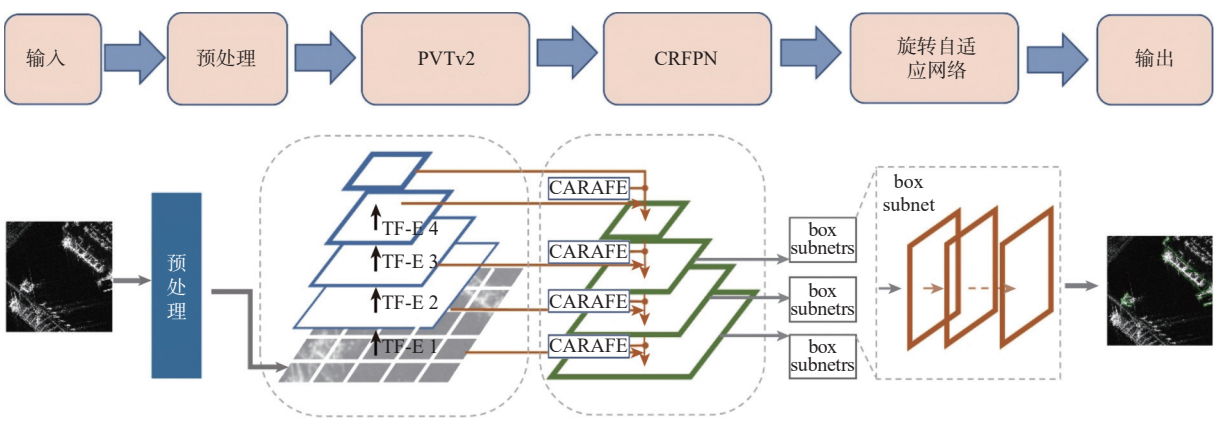


图 1　RCAR-Net 总体框架

Fig. 1　RCAR-Net overall framework

1.2　训练数据预处理

当前方法为了避免模型过拟合, 训练数据预处理时, 通常会使用正则化技术, 同时, 简单的图像变换(裁剪、翻转等)也可以提高模型的准确率。本文提出的 RCAR-Net 采用 Cutout[13] 实现正则化, 具体而言, 在训练样本上进行随机位置裁剪, 用现有样本的部分遮挡版本来扩大数据集。Cutout 能使模型更好地利用图像的上下文信息, 而不是依赖于一小组特定的视觉特征, 增强了模型的鲁棒性和泛化能力。

1.3　基于 PVTv2 的主干网络

为了解决目标尺寸变化多样的问题, 本文使用金字塔视觉(Transformer pyramid vision transformer, PVT)[14] 作为模型的主干网络, 相比卷积神经网络

(convolutional neural network, CNN)和常规的视觉(Transformer vision transformer, ViT)有显著优势。

图2展示了CNN、ViT与PVTv2的对比。随着网络深度的增加,CNN的感受野逐渐增大,但仍然缺少全局的感知信息。ViT[15]通过计算全局各个区域之间的相似性,具有较大的感受野,能够更好地捕获图像的语义信息。2021年,Wang等[16]提出了PVT,克服了将Transformer移植到密集预测任务中的困难,证明了在检测和分割等密集预测任务中,纯Transformer骨干网络可以取得优异的性能。如图2(c)所示,PVT可利用渐进式缩小特征的金字塔结构,减少计算复杂度。实验表明,将PVT作为RetinaNet的主干网络有显著优势。本文采用的PVTv2主要有如下3个独特的模块:

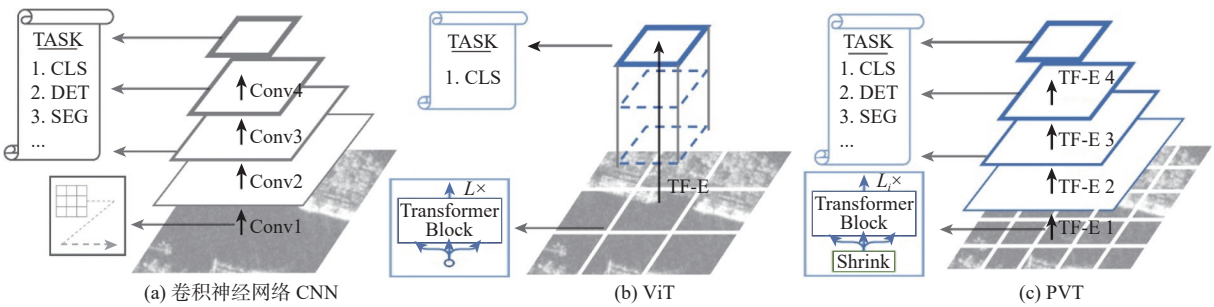


图2 图2卷积神经网络CNN、ViT与PVT的架构对比

Fig. 2 Convolutional neural networkNN, ViT and PVT architecture comparison

线性空间注意力。为解决PVT[16]处理高分辨率图像时,空间注意力操作带来的计算复杂度较大的问题,PVTv2[14]提出了线性空间注意力。不同于PVT使用卷积操作,PVTv2在多头注意操作之前使用平均池化将空间维数由$h\times w$降至固定大小$P\times P$,这使得PVTv2具有较低的计算复杂度。当输入大小为$h\times w\times c$,线性空间注意力的复杂度为

$$\Omega(A)=\frac{(2h^2w^2c)}{R^2}+hwc^2R^2 \tag{1}$$

$$\Omega(A_l)=2hwP^2c \tag{2}$$

式中:A为PVT中使用的空间注意力;A_l为PVTv2中使用的线性空间注意力;R为空间注意力的空间还原比;P为线性空间注意力的池大小,设置为7。

重叠分块嵌入。PVT将图像视作非重叠块序列,而这种处理方式会在一定程度上破坏图像的局部连续性。因此,在块嵌入方面,PVTv2对图像进行序列化时采用重叠块嵌入方法,采用带填充的卷积实现。该方法扩大了块窗口,使近邻窗口重叠一半面积。当输入大小为$h\times w\times c$,使用步长为S,尺寸为$2S-1$,数量为C,填充为$S-1$的卷积进行处理,输出的尺寸为$\frac{h}{S}\times\frac{w}{S}\times C$。

卷积前馈网络。PVT中采用了定长位置编码,这对于多尺度图像特征提取不够灵活。而PVTv2去掉了定长的位置编码,将带零填充的位置编码[17-19]引入,并在前馈网络的全连接层以及高斯误差线性单元(Gaussian error linear unit, GELU)之间添加了一个padding为1的3×3卷积,这使得其能更好地处理任意尺度图像。

1.4 内容感知的特征金字塔网络

特征上采样是目标检测、实例分割和场景解析等任务中的关键操作。在RCAR-Net中使用特征上采样操作将低分辨率特征图与高分辨率特征图融合[8]。最广泛使用的特征上采样方法为最近邻插值和双线性插值,仅通过像素点的空间位置来决定上采样核,并没有利用到特征图的语义信息,且感受野通常较小。Deconvolution[20]通过卷积完成上采样操作,但由于对特征图的每个位置都使用了相同的上采样核,难以捕捉特征图的语义信息,此外还会引入大量的参数和计算量。CARAFE(content-aware reassembly of features)[21]具有较大的感受野,能更好地利用周围的信息。此外CARAFE可以动态地生成自适应的卷积核,支持特定于实例的内容感知处理。最后,CARAFE是轻量级的,能引入很少的计算开销,并且可以轻松集成到现有的网络架构中。CARAFE分为2个主要模块,分别是核预测模块和内容感知重组模块。

核预测模块。若输入特征图形状为$h\times w\times c$,首先使用1×1卷积将通道压缩至c_m,以减小后续步骤的计算量。之后进行内容编码,假设采用上采样核尺寸为$k_{up}\times k_{up}$,希望得到形状为$\sigma h\times\sigma w\times C$的输出特征图,如果希望对输出特征图的每个位置使用不同的上采样核,那么需要预测的上采样核形状为$\sigma H\times\sigma W\times k_{up}\times k_{up}$。CARAFE利用一个$k_{encoder}\times k_{encoder}$的卷积层来预测上采样核,输入通道数为$c_m$,输出通道数为$\sigma^2k_{up}^2$。将通道维在空间维展开,得到上采样核的形状为$\sigma H\times\sigma W\times k_{up}^2$。实验表明[21],$k_{encoder}=$

$k_{up}-2$ 是性能和效率之间的一个平衡。最后用 softmax 对上采样核进行归一化。

内容感知重组模块。找到每个输入特征图中对应输出特征图中的位置，取出以之为中心的 $k_{up}\times k_{up}$ 的区域，和预测出的该点的上采样核作点积，得到输出值。相同位置的不同通道共享同一个上采样核。

特征金字塔网络(feature pyramid network, FPN)[14]通过自顶向下的路径和横向的连接来构造特征金字塔，广泛应用于目标检测领域[21]。在自上而下的路径中，FPN 通过最近邻插值对低分辨率的特征图进行 2 倍的上采样，将结果与高分辨率的特征图融合。本文使用 CARAFE 替换掉了原有的最近邻插值，构建内容感知的特征金字塔网络(content-aware reassembly of features pyramid network, CRFPN)，如图 3 所示。

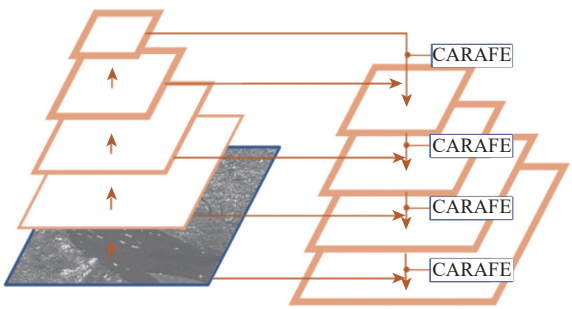


图 3　内容感知的特征金字塔网络

Fig. 3　Content perception feature pyramid network

1.5　旋转自适应的 RetinaNet

基于单阶段的目标检测算法的速度较快，实时性好。一般来说，遥感系统产生的图像尺寸比普通相机传统传感器大得多，这为 SAR 图像船舶目标检测带来了巨大的计算负担。为满足船舶航行监控等具有实时性的任务，本文选择单阶段的 RetinaNet 作为基本框架。由 1 个主干网络和 2 个任务专用子网络组成。采用 FPN[22] 作为骨干网络来提取多尺度特征。分类和回归子网络是若干个堆叠卷积层的完全卷积网络。RetinaNet 针对现有单阶段目标检测模型中正样本和负样本的类别不平衡问题，使用焦点损失(Focal loss)来降低大量简单负样本在标准交叉熵中所占权重，以提高困难负样本所占权重。Focal loss 的公式为

$$L(p_t)=-(1-p_t)^{\gamma}\ln(p_t) \tag{3}$$

$$p_t=\begin{cases}p & y=1\\ 1-p & \text{其他}\end{cases} \tag{4}$$

式中：y 的取值为 1 和−1，分别代表前景和背景；p 的取值范围为 [0,1]，表示模型预测属于前景的概率；γ 为常数，范围是 [0,5]，当其为 0 时，Focal loss 和普通的交叉熵损失函数一致。

此外，为解决传统水平边界框会包含大量的背景像素等问题，本文在 RetinaNet 的基础上，在每个 FPN 的输出层添加了旋转自适应模块来生成一系列候选旋转框。RetinaNet 用于通用目标检测时，输出水平边界框格式(如图 4 (a) 所示)为 (x,y,w,h)，其中 x 和 y 为边界框的中心，w 和 h 为目标长宽。为了与普通目标检测兼容，本文将旋转目标的表示方式定义(如图 4(b) 所示)为 (x,y,w,h,θ)，其中 $\theta\in[-\pi/2,\pi/2]$，表示 x 轴的方向到宽度 w 方向的夹角。所有其他设置和原始的 RetinaNet 相同。

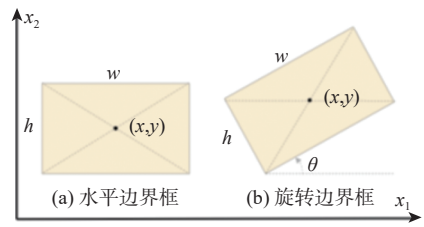


图 4　2 种边界框示意

Fig. 4　Two bounding box diagrams

网络的参数化回归目标如下：

$$\Delta x_g=(x_g-x)R(\theta)\cdot\left(\frac{1}{w},\frac{1}{h}\right) \tag{5}$$

$$\Delta y_g=(y_g-y)R(\theta)\cdot\left(\frac{1}{w},\frac{1}{h}\right) \tag{6}$$

$$(\Delta w_g,\Delta h_g)=\ln(w_g,h_g)-\log(w,h) \tag{7}$$

$$\Delta\theta_g=\frac{1}{\pi}(\theta_g-\theta+k\pi) \tag{8}$$

式中：x_g 和 x 分别表示标记的真实检测框和锚点框。k 为整数，以确保 $(\theta_g-\theta+k\pi)\in[-\pi/2,\pi/2]$，水平锚框用 $\theta=0$ 来表示。在旋转自适应 RetinaNet 中，首先对 CRFPN 的输出进行解码，然后重新计算回归目标。损失函数方面，本文使用了 Smooth L1 Loss 作为回归损失。如图 5 所示，上方为采用水平边界框的 Pow-FAN[23] 的检测结果，下方为本文提出的

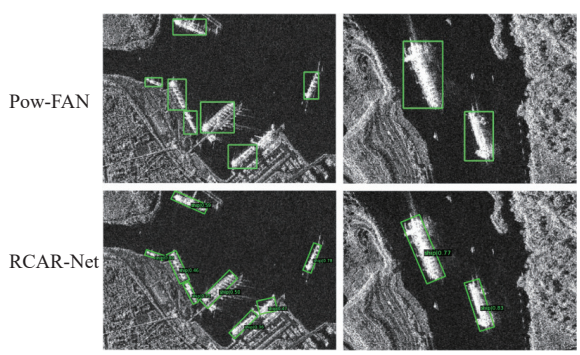


图 5　水平边界框与旋转边界框对比

Fig. 5　Compare the horizontal bounding box with the rotating bounding box

RCAR-Net的检测结果。可以看到，水平边界框会包含大量的背景像素，引入噪声。

2 实验结果与分析

本文通过实验来评估所提出的SAR船舶目标检测模型RCAR-Net的性能。首先介绍实验中使用的2个数据集，然后对实验细节和评价指标进行介绍。在本节的第3部分，本文在SAR Ship Detection Dataset(SSDD)数据集上构建了详细的消融实验，以确定模型各个模块的贡献。最后，在2个数据集上与9个当前最先进的方法进行对比，说明本文提出的RCAR-Net的优势。

2.1 数据集介绍

本文实验中使用的两个数据集为SSDD[4]和High-Resolution SAR Images Dataset(HRSID)[5]。SSDD是一个广泛用于SAR图像船舶目标检测的数据集[4]，在数据格式上兼容Pascal VOC[24]，包含沿海和近海场景的2 540艘不同尺寸的船舶目标和1 160张图像，平均图像大小为500像素×500像素，平均每张图像包含2.12艘船舶。本文实验将图像的大小调整为608像素×608像素，训练集和测试集分别包含928张和232张图像。此外，为了验证本文提出RCAR-Net的泛化能力，还在HRSID数据集上进行了大量实验。HRSID是规模更大的SAR图像船舶目标数据集，包含从Sentinel-1和TerraSAR-X收集的，不同分辨率、海况、海域和沿海港口的5 604张高分辨率SAR图像和16 951个船舶目标[5]。在HRSID数据集上，本文将图像的大小调整为800像素×800像素，训练集和测试集分别包含3 623张和1 981张图像。

2.2 实验细节与评估指标

本文实验在硬件配置为Intel(R) Xeon(R) Silver 4110@2.10 GHz的CPU，16 GB内存的服务器上进行，使用的GPU为NVIDIA GeForce RTX 2 080 Ti，显存大小为12 GB。软件环境为python3.8，Cuda10.2，PyTorch 1.6.0。本文在网络模型训练过程中使用Adam[25]为优化器进行学习和网络参数更新，以提升网络收敛的速度。初始学习率设置为0.000 1，热身迭代次数为500个epoch。参考Pascal VOC[24]的目标检测任务，使用平均精度(mAP)作为评价指标以验证本文模型的检测性能。

2.3 消融实验

本文进行了一系列的消融实验以验证RCAR-Net各个模块的有效性，包括PVTv2、CRFPN和Cutout。本文使用Pascal VOC2007中计算mAP的方法(11点插值法)：在[0,1]之间等差地采样10个值作为置信度阈值计算精确度和召回率，评测了交并比阈值为0.5下的mAP值。

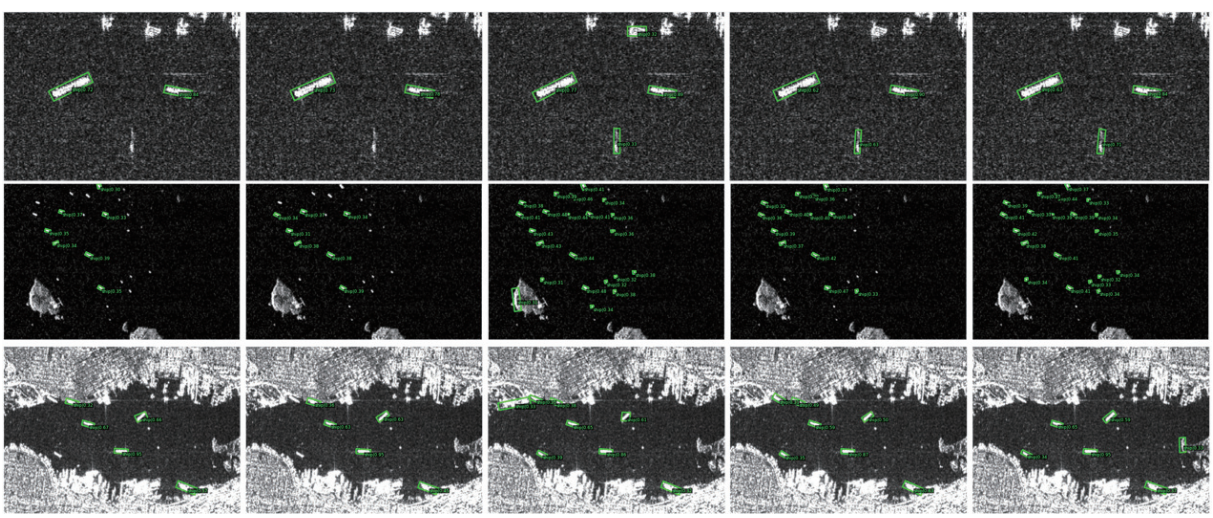
表1定量地显示了本文方法各个模块对于目标检测性能的影响，从结果可以看出，随着各个模块的加入，网络的检测精度逐渐提高。Baseline由于存在难以提取并融合图像的多尺度特征等缺点，仅实现了69.42%的mAP。本文提出的RCAR-Net，mAP达到76.39%，相比Baseline增长了6.97%，证明了本文方法的有效性。

图6对本节的检测结果进行了可视化展示，从左到

表1 RCAR-Net消融实验

Table 1 RCAR-Net ablation experiment

Baseline	PVTv2	CARAFE	Cutout	mAP/%
实验设置				69.42
	√			72.07
		√		73.09
	√		√	73.26
	√	√	√	**76.39**



(a) Base line (b) PVTv2 (c) CRFPN (d) PVTv2+Cutout (e) 本文方法

图6 消融实验中检测结果的可视化展示

Fig. 6 Visualization of measurement results in ablation experiments

右依次为 Baseline、PVTv2、CRFPN、PVTv2+Cutout、本文方法。

PVTv2 的有效性：如表 1 所示，当 Baseline 使用 PVTv2 作为主干网络后，mAP 相比 Baseline 提高了 2.65%，说明基于 Transformer 架构的 PVTv2 能更好地提取图像的多尺度特征，并保留特征图的局部连续性。

CRFPN 的有效性：如表 1 所示，当 Baseline 使用 CRFPN 后，mAP 相比 Baseline 提高了 3.67%。这表明，CARAFE 算子能出色地完成上采样工作，融合图像的多尺度特征。

Cutout 的有效性：如表 1 所示，在 PVTv2 作为主干网络的基础上使用 Cutout 进行数据增强后，mAP 得到了进一步提高，说明该方法可以进一步增强模型学习的抗噪声的能力。

2.4　对比实验

为了充分验证本文所提出的 RCAR-Net 在 SAR 船舶目标检测方面的性能优势，设计了一系列对比实验，以 RCAR-Net 与当前行业内主流的 SAR 船舶目标检测方法为对象。这些对比实验的目标是在 SSDD 和 HRSID 数据集上进行检测，以实际的效果和数据为基础进行性能对比与分析，最终结果如表 2 所示。

为了保证比较的公平性和准确性，本文在计算 mAP 的方法上，采用了 Pascal VOC2012 中给出的公式，而非消融实验中使用的 11 点插值法。通过对比结果分析，可以发现，RCAR-Net 在 SSDD 和 HRSID 这 2 个数据集上的检测性能都表现出色，甚至超越了现有的最先进的方法。具体来说，其

表 2　本文方法与当前方法在 SSDD 和 HRSID 数据集上的对比

Table 2　Comparison of the proposed method with current methods on SSDD and HRSID datasets

方法	mAP/%	
	SSDD	HRSID
Faster R-CNN[26]	92.07	81.80
Cascade RCNN[27]	91.61	81.89
YOLOv4[28]	92.16	83.23
SSD300[29]	87.06	79.05
SSD512[29]	89.19	82.50
DAPN[17]	90.60	88.20
Double-Head R-CNN[18]	91.17	80.41
PANet[19]	91.73	80.11
RetinaNet[8]	86.37	77.32
Quad-FPN[30]	92.84	86.12
LFG-Net[31]	93.01	88.50
RCAR-Net (本文)	93.63	90.37

mAP 值分别超过了第 2 名 0.62% 和 1.87%，这充分证明了本文方法在检测精度上的优越性和鲁棒性，同时也证实了其对于不同数据集的定性。

为了更直观地展示实验结果，图 7 给出了在 HRSID 数据集上，使用 Baseline（RetinaNet）和 RCAR-Net 进行船舶检测结果的可视化对比。从对比结果中可以看出，Baseline 在处理复杂海况中的船舶目标时存在漏检问题，而 RCAR-Net 能够准确地检测出复杂海况下以及靠近海岸的船舶。因此，可以得出结论，本文方法对于目标尺寸变化和噪声干扰具有较强的适应性，充分体现了其在实际应用场景中的强大实力和潜力。

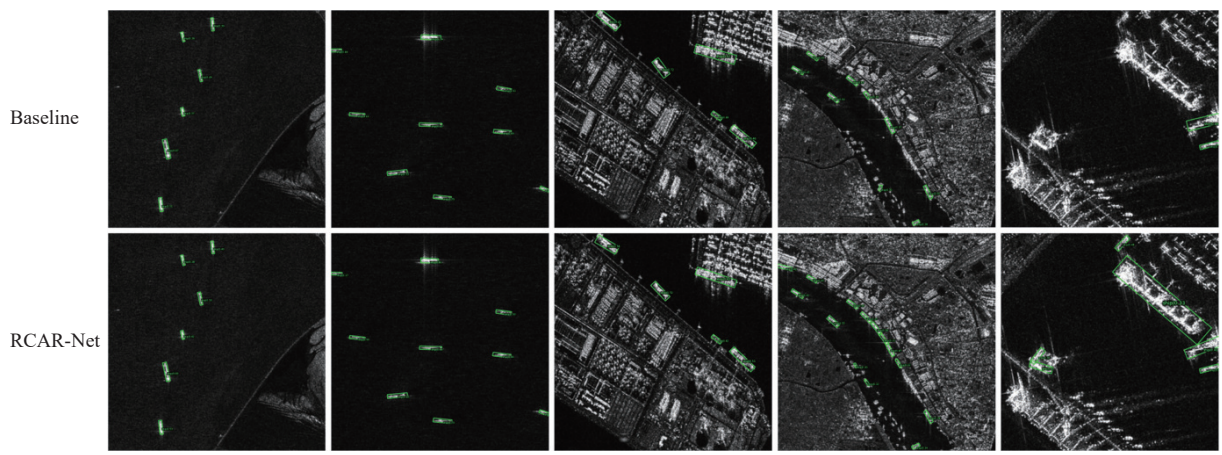


图 7　在 HRSID 数据集上的船舶目标检测结果展示

Fig. 7　Presentation of ship target detection results on the HRSID dataset

3　结　论

1）本文提出的 RCAR-Net 算法在 SAR 船舶检测任务上性能卓越。在 SSDD 与 HRSID 数据集上，其平均精度（mAP）分别达到 93.63% 和 90.37%，显著优于多种主流方法，验证了算法的有效性与鲁棒性。

2）采用旋转边界框是提升精度的关键策略。

该方法能紧密贴合任意方向及大长宽比的船舶，有效减少背景噪声干扰，并改善对密集目标的检测效果。

3）PVTv2主干网络与CARAFE上采样算子的结合被证明是成功的。消融实验显示，二者分别在多尺度特征捕获和低开销特征融合上作出显著贡献，共同增强了模型对不同尺寸目标的感知能力。

尽管RCAR-Net精度领先，但其较高的计算复杂度是未来应用需解决的问题。后续研究可聚焦于模型轻量化，如采用剪枝或知识蒸馏技术，以在保证精度的前提下满足实时处理与低成本部署的需求。

参考文献（References）

[1] MOREIRA A, PRATS-IRAOLA P, YOUNIS M, et al. A tutorial on synthetic aperture radar[J]. IEEE Geoscience and Remote Sensing Magazine, 2013, 1(1): 6-43.

[2] REIGBER A, SCHEIBER R, JAGER M, et al. Very-high-resolution airborne synthetic aperture radar imaging: signal processing and applications[J]. Proceedings of the IEEE, 2013, 101(3): 759-783.

[3] CUI J Y, JIA H C, WANG H P, et al. A fast threshold neural network for ship detection in large-scene SAR images[J]. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2022, 15: 6016-6032.

[4] ZHANG T W, ZHANG X L, LI J W, et al. SAR ship detection dataset (SSDD): official release and comprehensive data analysis[J]. Remote Sensing, 2021, 13(18): 3690.

[5] WEI S J, ZENG X F, QU Q Z, et al. HRSID: a high-resolution SAR images dataset for ship detection and instance segmentation[J]. IEEE Access, 2020, 8: 120234-120254.

[6] GIRSHICK R, DONAHUE J, DARRELL T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation[C]// Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2014: 580-587.

[7] REDMON J, DIVVALA S, GIRSHICK R, et al. You only look once: unified, real-time object detection[C]//Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2016: 779-788.

[8] LIN T Y, GOYAL P, GIRSHICK R, et al. Focal loss for dense object detection[C]//2017 IEEE International Conference on Computer Vision(ICCV). Piscataway: IEEE Press, 2017: 2999-3007.

[9] LI J W, QU C W, SHAO J Q. Ship detection in SAR images based on an improved faster R-CNN[C]//Proceedings of the 2017 SAR in Big Data Era: Models, Methods and Applications. Piscataway: IEEE Press, 2017: 1-6.

[10] WANG Y Y, WANG C, ZHANG H, et al. Automatic ship detection based on RetinaNet using multi-resolution Gaofen-3 imagery [J]. Remote Sensing, 2019, 11(5): 531.

[11] ZHOU Z, CHEN J, HUANG Z X, et al. HRLE-SARDet: a lightweight SAR target detection algorithm based on hybrid representation learning enhancement[J]. IEEE Transactions on Geoscience and Remote Sensing, 2023, 61: 5203922.

[12] CHEN S W, CUI X C, WANG X S, et al. Speckle-free SAR image ship detection[J]. IEEE Transactions on Image Processing, 2021, 30: 5969-5983.

[13] DEVRIES T, TAYLOR G W, ASSIRI Y. Improved regularization of convolutional neural networks with cutout[EB/OL]. (2017-11-29)[2023-06-01]. https://arXiv.org/abs/1708.04552v2.

[14] WANG W H, XIE E Z, LI X, et al. PVT v2: improved baselines with pyramid vision transformer[J]. Computational Visual Media, 2022, 8(3): 415-424.

[15] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, et al. An image is worth 16x16 words: transformers for image recognition at scale [EB/OL]. (2021-06-03)[2023-06-02]. https://doi.org/10.48550/arXiv.1708.04552.

[16] WANG W, XIE E, LI X, et al. Pyramid vision transformer: a versatile backbone for dense prediction without convolutions[C]// 2021 IEEE/CVF International Conference on Computer Vision(ICCV). Piscataway: IEEE Press, 2021: 548- 558.

[17] CUI Z Y, LI Q, CAO Z J, et al. Dense attention pyramid networks for multi- scale ship detection in SAR images[J]. IEEE Transactions on Geoscience and Remote Sensing, 2019, 57(11): 8983-8997.

[18] WU Y, CHEN Y P, YUAN L, et al. Rethinking classification and localization for object detection[C]//Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2020: 10183-10192.

[19] LIU S, QI L, QIN H F, et al. Path aggregation network for instance segmentation[C]//Proceedings of the 2018/CVF IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2018: 8759-8768.

[20] NOH H, HONG S, HAN B. Learning deconvolution network for semantic segmentation[C]//Proceedings of the 2015 IEEE International Conference on Computer Vision. Piscataway: IEEE Press, 2015: 1520-1528.

[21] WANG J, CHEN K, XU R, et al. CARAFE: content-aware ReAssembly of FEatures[C]//Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE Press, 2019: 3007-3016.

[22] LIN T Y, DOLLÁR P, GIRSHICK R, et al. Feature pyramid networks for object detection[C]//Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2017: 2117-2125.

[23] XIAO M, HE Z, LI X Y, et al. Power transformations and feature alignment guided network for SAR ship detection[J]. IEEE Geoscience and Remote Sensing Letters, 2022, 19: 4509405.

[24] EVERINGHAM M, VAN GOOL L, WILLIAMS C K, et al. The pascal visual object classes (VOC) challenge[J]. International Journal of Computer Vision, 2010, 88(2): 303-338.

[25] LOSHCHILOV I, HUTTER F. Decoupled weight decay regularization[EB/OL]. (2019-01-04)[2023-06-07]. https://doi.org/10.48550/arXiv.1711.05101.

[26] REN S Q, HE K M, GIRSHICK R, et al. Faster R-CNN: towards real-time object detection with region proposal networks[C]//Proceedings of the IEEE Transactions on Pattern Analysis and Machine Intelligence. Piscataway: IEEE Press, 2017: 1137-1149.

[27] CAI Z W, VASCONCELOS N. Cascade R-CNN: delving into high quality object detection[C]//Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2018: 6154-6162.

[28] WANG C Y, BOCHKOVSKIY A, LIAO H M. Scaled-YOLOv4: scaling cross stage partial network[C]//Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2021: 13024-13033.

[29] LIU W, ANGUELOV D, ERHAN D, et al. SSD: single shot multi-box detector[C]//Proceedings of the European Conference on Computer Vision. Berlin: Springer, 2016: 21-37.

[30] ZHANG T W, ZHANG X L, KE X. Quad-FPN: a novel quad feature pyramid network for SAR ship detection[J]. Remote Sensing, 2021, 13(14): 2771.

[31] WEI S J, ZENG X F, ZHANG H, et al. LFG-net: low-level feature guided network for precise ship instance segmentation in SAR images[J]. IEEE Transactions on Geoscience and Remote Sensing, 2022, 60: 5231017.

A rotated content-aware retina network for SAR ship detection

WANG Ziyi[1, 2], YIN Jiahao[1, 2], HUANG Bobin[1, 2], GAO Feng[1, 2, *]

(1. School of Computer Science and Technology, Ocean University of China, Qingdao 266100, China;
2. Qingdao Marine Information Mining and Inference Engineering Research Center, Ocean University of China, Qingdao 266100, China)

Abstract: Current synthetic aperture radar (SAR) ship detection methods primarily encounter two challenges: 1) the variability of target sizes and the abundance of interfering factors; 2) multiple orientations of targets and a limited quantity of training samples. To address these issues, this paper introduces a Rotating Target-Aware Network for SAR image ship detection, RCAR-Net. Firstly, the backbone network employs the PVTv2 based on a multi-scale Transformer architecture, which better preserves the local continuity of feature maps while enhancing the integration of multi-scale image features. In conjunction, a combination of rotating bounding boxes with RetinaNet effectively reduces background redundancy and noise interference. The model's generalizability and robustness are further enhanced by the introduction of the Cutout data augmentation technique, which uses partial occlusion of current samples to enlarge the dataset. Finally, the effective CARAFE operator is used to upsample low-resolution feature maps, improving the multi-scale fusion effect, reducing computational and memory costs while maintaining detection accuracy. RCAR-Net achieves an average precision of 93.63% and 90.37% on the SSDD and HRSID SAR ship detection datasets, respectively, significantly outperforming current methods such as DPAN and PANet, demonstrating strong adaptability to changes in target size and noise interference.

Keywords: synthetic aperture radar; ship detection; rotated bounding box; deep learning; attention mechanism

Received: 2023-06-19; **Accepted**: 2023-09-23; **Published Online**: 2023-11-28 14:26

URL: link.cnki.net/urlid/11.2625.V.20231127.1912.001

Foundation items: National Science and Technology Major Project of China Under Grant (2022ZD0117202); National Science Fundation of Qingdao (23-2-1-222-zyyd-jch)

* **Corresponding author**. E-mail: gaofeng@ouc.edu.cn

http://bhxb.buaa.edu.cn jbuaa@buaa.edu.cn
DOI: 10.13700/j.bh.1001-5965.2023.0411

针对航拍小目标检测的YOLOv7改进方法

刘一诺，张琪*，王蓉，李冲
(中国人民公安大学 信息网络安全学院，北京 102600)

摘　　要：针对目前检测技术在航拍小目标检测任务中存在的漏检率和误检率较高的问题，提出一种基于改进YOLOv7的航拍小目标检测方法。在主干网络中加入CBAM融合注意力机制，将特征图在空间和通道两方面合理分配网络权重，抑制背景干扰，提升检测精度；引入一种用于低分辨率图像和小目标细化检测的SPD-Conv模块，消除原有卷积模块的跨卷积层和池化层，解决了原始卷积模块中存在的细粒度信息丢失以及对于特征表示学习效率较低的问题；在处理后的DOTA航拍数据集上进行性能评估。实验结果表明：改进的YOLOv7算法在处理后的DOTA航拍数据集上准确率P达到83.7%，召回率R达到78.2%，均值平均精度mAP50达到81.5%，比原始YOLOv7算法精度提升了3.1%。说明所提算法可以有效降低漏检和错检率，具有良好性能。

关　键　词：YOLOv7；小目标检测；注意力机制；卷积神经网络；计算机视觉
中图分类号：V221+.3; TB553
文献标志码：A　　**文章编号**：1001-5965(2025)07-2506-07

目标检测作为计算机视觉领域的核心问题之一，广泛应用于人脸检测[1]、行人检测[2]、动作识别[3]等研究方向。无人机可以实现高分辨率影像的采集，进行监测和监视；也可以代替人工去不同场所进行物资、信息的输送，完成救援赈灾等任务。在公共安全方面，无人机因其经济实惠、体积较小的优点，可以收集特定地点(森林、海洋等)的准确信息，方便研究人员进行安全领域的排查。在航拍领域中，随着无人机等航拍设备的发展，针对航拍小目标图像的目标检测成为新挑战，相较于大目标和中目标，航拍小目标分辨率低导致其信息量较少、样本不均衡、数据集较少、对定位的精度要求高；但在实际应用中，航拍小目标图像包含的目标多，可以采集到人力难以抵达场所的图像和视频，获取图像方便且成本低，广泛应用于各个领域。常用的目标检测算法在面对小目标检测时效果会骤降，基于上述问题，针对航拍小目标的检测研究有很高的研究价值。

在传统目标检测算法中，算法根据目标不同分配不同的特征提取方法进行检测。基于卷积神经网络技术的目标检测将算法分为以YOLO系列算法为典型代表的单阶段算法和以R-CNN[4]算法为主要代表方法的两阶段算法2种，两阶段算法比起单阶段算法利用增加的产生候选框进行特征提取的操作，提高了检测的准确性但是检测速度相应变慢。单阶段算法中YOLO结合目标检测中的分类与回归方法，将整张图片作为输入，使用端到端的Darknet网络进行检测，完整地使用了上下文的关联性，极大地降低了误检率，Redmon等[5]提出基于单阶段目标检测的YOLO算法；Redmon等[6]引入了锚框机制并提出了YOLOv2算法；其继续引入darknet-53网络[7]，使用sigmoid函数提出YOLOv3算法；Bochkovskiy等[8]更改YOLOv3算法中激活函数为Mish，提升了检测的速度和精度，提出了YOLOv4

收稿日期：2022-06-28；**录用日期**：2023-06-28；**网络出版时间**：2023-12-06 16:03
网络出版地址：link.cnki.net/urlid/11.2625.V.20231206.0926.001
基金项目：国家自然科学基金(62076246)；中国人民公安大学安全防范工程双一流专项(2023SYL08)
* **通信作者**. E-mail：qi.zhang@ppsuc.edu.cn

引用格式：*刘一诺，张琪，王蓉，等. 针对航拍小目标检测的YOLOv7改进方法[J]. 北京航空航天大学学报，2025，51(7)：2506-2512.*
LIU Y N，ZHANG Q，WANG R，et al. Improved YOLOv7 method for aerial small target detection in aerial photography[J]. Journal of Beijing University of Aeronautics and Astronautics，2025，51(7)：2506-2512 (in Chinese).

算法; YOLOv5 将之前版本进行合理融合, 一次性完成目标定位和分类, 得到了更加快速准确的结果; 之后, YOLOv6 使用更加简洁高效的 Efficient-Rep 基线网络和 Rep-PAN 瓶颈网络, 并采用无锚范式进一步提升检测精度; 随着 YOLO 模型的快速发展, Wang 等[9] 提出 YOLOv7 模型, 该模型在速度和精度上远远超过了所有已知的检测器。其在模型结构上改进了原始的 ELAN 结构, 使用扩展、换序、融合这 3 种方式提出 E-ELAN 结构, 在模型缩放方面, 采用深度与宽度结合并针对串联类型的模型改进使添加卷积后模型的输入输出通道数不变, 在重参数模块使用重参数卷积, 在提升检测速度时保证精度不会下降。由此可见, YOLOv7 可以很好地应用于常规目标检测。但在航拍小目标检测方面, 因其对航拍特征图的不同特征赋予相同权重, 以及原始卷积模块的跨卷积层和池化层使得细粒度信息丢失以及对于特征表示的学习效率较低, 因此在面向小目标这种分辨率低、信息量较少、对定位的精度要求高等对象时, YOLOv7 检测方法存在目标检测精度不高的问题。

因此, 本文基于 YOLOv7 模型针对小目标检测场景进行改进, 提出一种针对航拍小目标的检测方法。首先, 在 YOLOv7 网络模型中加入卷积注意力模块(convolutional block attention module, CBAM)融合通道注意力机制, 合理分配网络权重提高模型对深层次的重要特征的提取能力; 其次, 添加用于低分辨率图像和小目标的 SPD-Conv 模块, 解决原始卷积模块中存在的细粒度信息丢失以及对于特征表示的学习效率较低的问题; 最后, 对航拍数据集进行处理, 选取有效样本进行检测, 在不影响训练效果的情况下减少参数量和训练时间。实验结果表明, 本文方法可以有效提升对于航拍小目标的检测精度。

1　YOLOv7 网络模型

YOLO 是一种基于深度神经网络进行对象识别和定位的算法。YOLOv7 作为现阶段最先进的 YOLO 模型, 其主要改进技术有以下 4 个方面。

1.1　E-ELAN 高效聚合网络

YOLOv7 使用的高效聚合网络如图 1 所示。E-ELAN 利用群卷积扩展计算块的通道和基数, 并对一个计算层的所有计算块应用相同的算法得到特征, 之后利用换序和融合的基数方式将不同组的特征进行组合, 最后添加特征映射执行合并基数, 由此增强不同特征映射学到的特征, 提高参数和计算的使用。

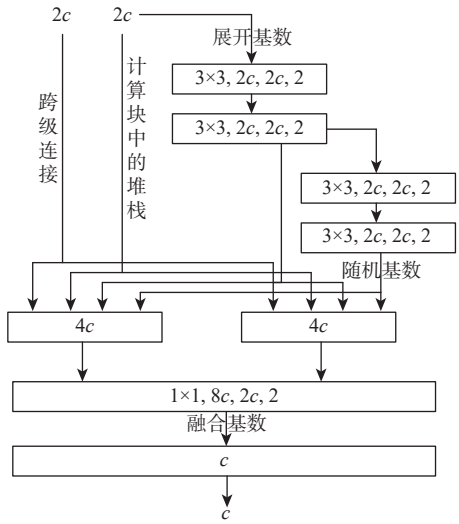


图 1　E-ELAN

Fig. 1　Efficient aggregation networks

1.2　模型缩放

模型缩放的结构如图 2 所示, 为解决传统模型缩放在串联模型上对输入输出宽度的劣势, YOLOv7 使用一个基于级联的模型进行模型深度缩放, 即模型只能够在计算块层中的深度方面进行长度缩放, 对于传输层内的模型其余部分无法进行相应程度的模型宽度缩放。其目的是利用模型的不同属性生成不同尺度的模型, 可以适用于不同推理速度的需求。

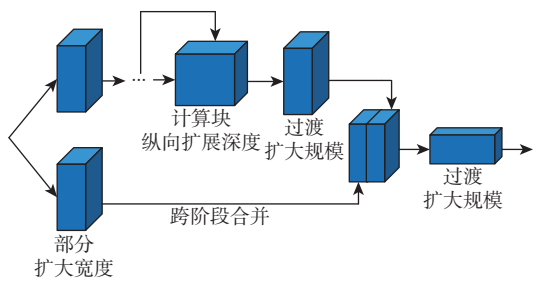


图 2　模型缩放

Fig. 2　Model scaling

1.3　重参数化模型

YOLOv7 的重参数化模型如图 3 所示, 原始的 RepConv 的身份连接破坏了 ResNet 中的残差和 DenseNet 中的连接, 为解决在有残差连接或串联接层中 RepConv 不应该有标识连接的问题, 使用无标识连接的 RepConvN 设计规划重参数化卷积

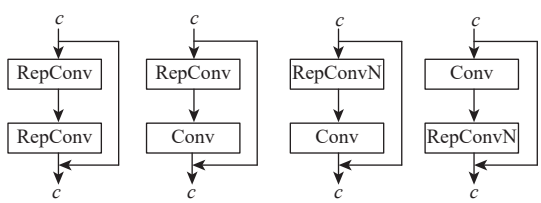


图 3　重参数化模型

Fig. 3　Reparametric models

的体系结构，当一个带有残差或拼接的卷积层被重新参数化卷积代替时，不存在恒等连接。该方法丰富了训练资源，使得网络获得更多信息，在提升网络模型的推理速度基础上获得了更加精确的训练结果。

1.4 辅助头检测

YOLOv7网络在辅助检测头方面，结构如图4所示，在Head部分YOLOv7使用引导头和辅助头2种检测头，分别用于负责检测获取网络信息中包含的深层信息和更浅层次的网络信息，负责对检测好的网络正误样本数据进行匹配与检测；在标签分配方面通过对引线头的预测生成层次标签同时引导辅助头和引线头学习，可以有效地提高召回率。

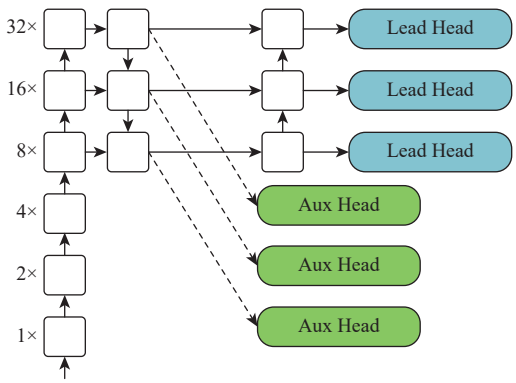


图4 辅助头检测

Fig. 4 Auxiliary head detection

2 改进方法

2.1 模型框架

本文考虑到对于小目标检测中存在的诸多问题，基于YOLOv7模型进行改进，使在小目标场景中，检测的速度效率与实际检测数据精度能达到合理平衡，降低各种小目标场景检测工作中的漏检率与误检率，提升实际检测精度，改进后的网络结构如图5所示。

改进具体操作是在YOLOv7的Backbone中将第2卷积层和第4卷积层替换为CBAM融合注意力机制，在目标输入后首先通过CBAM分配合理权重，提高模型对深层次的重要特征的提取能力，在后续处理中节省参数量以及提升检测精度；在YOLOv7的Head部分101层后面添加一个SPD-Conv模块和Conv模块联合模块，依靠SPD-Conv模块对于小目标特有的检测性能，对小目标进行聚焦从而增加识别的速度与精度。改进的YOLOv7分别在头和尾修改和添加了相应模块以提升性能。改进后的网络结构工作流程如下，在Backbone中，输入图像经过一层卷积（卷积+BN+SiLU）得到采样后的特征图，之后输入到添加的CBAM融合注意力机制中改善模型合理分配权重的能力，输出一个2倍降采样的特征图；再次通过一层卷积与一层CBAM通道输出一个4倍降采样的特征图，经过一系列新提出的E-ELAN模块处理被降采样的特征图，保证在E-ELAN模块的每步处理中，输入通道与输出通道保持一致，以此来提高网络的效率；将网络分为2个分支，其中一个分支采用最大池化实现空间降采样再进行卷积压缩，另一分支先进行卷积压缩再利用卷积完成降采样。将Backbone最后一次输出的32倍降采样特征图输入到Head中，使用SPPCSPC缩小该特征图的通道数，将经

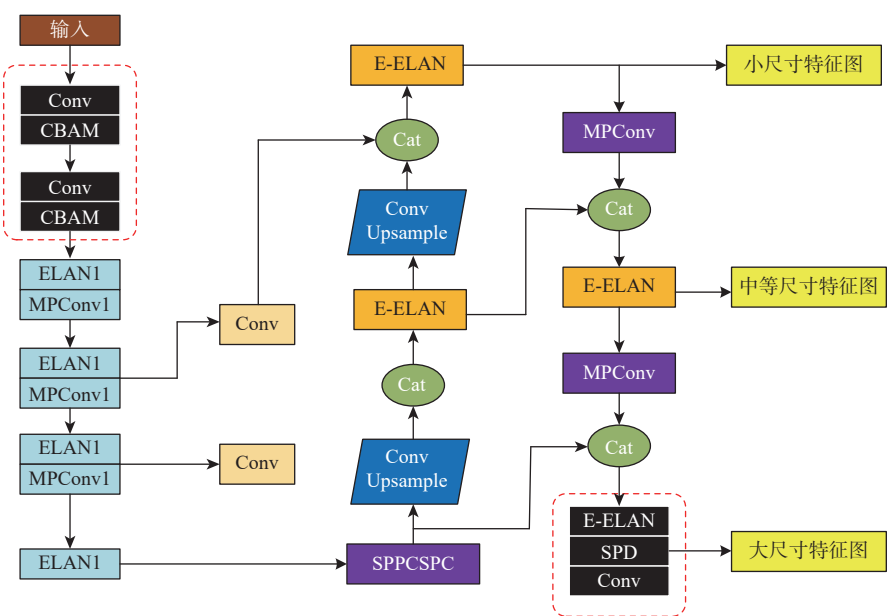


图5 改进YOLOv7模型框架

Fig. 5 Improvements to the YOLOv7 model framework

SPPCSPC 缩小的特征图使用从上到下的策略与先前通过 2 个分支得到的 2 个特征图进行融合, 按照从下到上的策略分层融合, 之后通过 E-ELAN 模块, 以及引入的一种改进 SPD-Conv 模块, 对小目标进行聚焦, 提高小目标检测的针对性, 最后进行三部分的输出, 得到最终的检测结果。

2.2　特征提取

CBAM[10] 是一种融合注意力机制模块, 结构原理如图 6 所示。

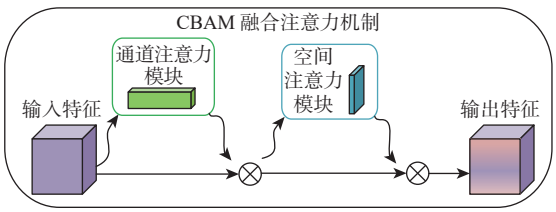


图 6　CBAM 模块

Fig. 6　Convolutional block attention module

本文将 CBAM 插入到 YOLOv7 的 Backbone 中, 替代原始卷积层, 先对输入特征图进行注意力的训练, 可以有效提升模型对权重的合理分配能力, 对于深层次的特征, 也有着更好的识别能力和提取能力。

CBAM 包含通道注意力模块(channel attention module, CAM)和空间注意力模块(spatial attention module, SAM)2 个互相独立的注意力子模块。此种方法开销几乎可以忽略, 不仅解决掉了参数多和计算量太大造成的问题, 还可以即插接即用直接集成到现有的网络架构系统中。经实验验证, 该模块几乎可以添加到所有的常规卷积层中进行使用, 并与常规卷积网络一起进行端到端的训练, 具有广泛的适用性。

在本文的网络结构中, 经过卷积网络处理后的特征图首先输入通道模块。通道注意力模块利用特征的通道注意力之间的关系来生成通道注意力图(见图 7), 主要用于确定输入图像关注点。计算公式如下:

$$\boldsymbol{M}_{\mathrm{c}}(\boldsymbol{F})=\sigma\left(\mathrm{MLP}\left(\mathrm{AvgPool}(\boldsymbol{F})\right)+\mathrm{MLP}\left(\mathrm{MaxPool}(\boldsymbol{F})\right)\right)=\sigma\left(\boldsymbol{W}_1\left(\boldsymbol{W}_0\left(\boldsymbol{F}_{\mathrm{avg}}^{\mathrm{c}}\right)\right)+\boldsymbol{W}_1\left(\boldsymbol{W}_0\left(\boldsymbol{F}_{\mathrm{max}}^{\mathrm{c}}\right)\right)\right)\tag{1}$$

式中: σ表示 sigmoid 函数; $\boldsymbol{W}_0 \in \boldsymbol{R}^{C/r\times 1}$, $\boldsymbol{W}_1 \in \boldsymbol{R}^{C\times C/r}$, $\boldsymbol{W}_0$和$\boldsymbol{W}_1$表示共享多层感知器(multilayer perceptron, MLP)的权值; AvgPool 和 MaxPool 分别表示平均和最大池化操作。

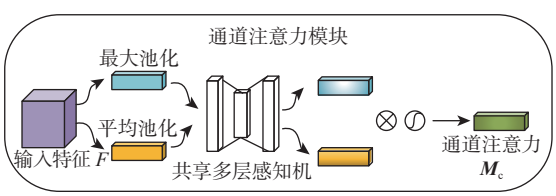


图 7　通道注意力模块

Fig. 7　Channel attention module

在进行计算时, 为了整合使用平均池化和最大池化生成的平均池化特征$\boldsymbol{F}_{\mathrm{avg}}^{\mathrm{c}}$和最大池化特征$\boldsymbol{F}_{\mathrm{max}}^{\mathrm{c}}$, 然后转发到由 MLP 层和隐藏层所共同组成的信息共享层网络上以进行信息的集成, 生成通道注意力映射$\boldsymbol{M}_{\mathrm{c}} \in \boldsymbol{R}^{C\times 1\times 1}$, 并在通过共享网络处理完成每个通道特征的描述符后, 特征向量可以使用 SUM 运算进行合并计算并获得输出。

经过通道注意力模块处理后的特征图进入空间注意力模块。空间注意力模块利用特征点的空间关系生成空间注意图(见图 8), 其主要关注通道注意力模块判定的关注点, 作为信息与通道注意互补。计算公式如下:

$$\boldsymbol{M}_{\mathrm{s}}(\boldsymbol{F})=\sigma\left(\boldsymbol{f}^{7\times 7}\left(\left[\mathrm{AvgPool}(\boldsymbol{F});\mathrm{MaxPool}(\boldsymbol{F})\right]\right)\right)=\sigma\left(\boldsymbol{f}^{7\times 7}\left[\boldsymbol{F}_{\mathrm{avg}}^{\mathrm{s}};\boldsymbol{F}_{\mathrm{max}}^{\mathrm{s}}\right]\right)\tag{2}$$

式中: $\boldsymbol{f}^{7\times 7}$为滤波器大小为7×7的卷积运算。

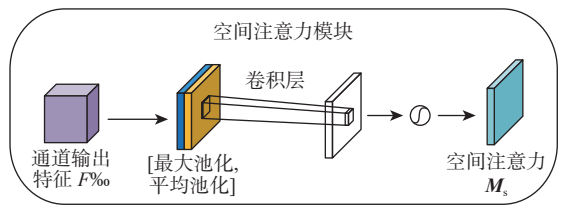


图 8　空间注意力模块

Fig. 8　Spatial attention module

在空间计算的过程中, 空间注意图中的模块作用原理是首先用沿通道轴进行池化操作生成$\boldsymbol{F}_{\mathrm{avg}}^{\mathrm{s}} \in \boldsymbol{R}^{1\times H\times W}$和$\boldsymbol{F}_{\mathrm{max}}^{\mathrm{s}} \in \boldsymbol{R}^{1\times H\times W}$这 2 个池化特征, 将二者进行连接处理生成一个有效的特征描述符, 之后再使用一个标准的卷积运算生成空间注意图$\boldsymbol{M}_{\mathrm{s}}(F) \in \boldsymbol{R}^{H\times W}$, 其包含的信息为标记或忽略的位置。

2.3　空间-深度卷积

在面对低分辨率图像和小目标的检测中, 卷积神经网络的良好性能会大大下降, 这是由于在现有卷积神经网络中使用了跨卷积层和池化层, 这样就会导致细粒度信息丢失以及对于特征表示的学习效率较低的问题。为了解决上述情况, 本文引入一种卷积神经模块 SPD-Conv[11], 该模块是由一个空间到深度(SPD)层和一个非跨步卷积层组成, 取代了每个跨卷积层和池化层, 消除了原有两层之间的交替, 可以应用到大多数卷积神经网络中。SPD-Conv 结构如图 9 所示。

在本文网络中, 最深层次的特征图经 E-ELAN 模块输出后, 进入 SPD-Conv 模块, 首先, 将输入特征图进行下采样, 然后, 将下采样得到的子映射沿信道维度连接起来, 得到一个空间维度减少了一个

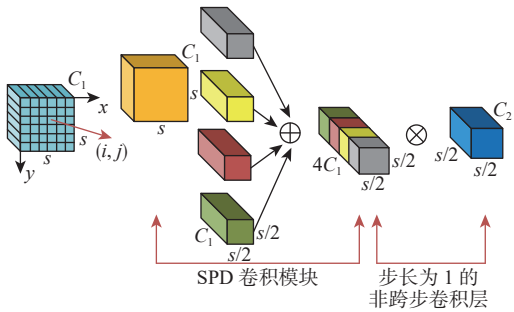


图 9 SPD-Conv 模块

Fig. 9 Space-to-depth conv module

比例因子，信道维度增加了一个比例因子的特征映射；最后，在 SPD 特征转换层之后，添加一个带有 C_2 滤波器的非跨步（步长=1）卷积层尽可能地保留所有的判别特征信息，$C_2 < k^2C_1$。将该过程用映射关系表示如下：将任意大小 $X(S,S,C_1)$ 的中间特征图映射为 $X\left(\frac{S}{k},\frac{S}{k},k^2C_1\right)$。

3 实验与结果分析

3.1 数据集

本文采用公开的 DOTA 数据集 V1.0 版本[12]，DOTA 图像数据集是专门被用于对航拍图像系统中随机出现过的小尺寸范围目标进行的图像检测或分析工作的大型图像数据集。其包含 2 806 幅航拍图，拥有飞机、棒球场、港口等 15 个类别详细标注，完全注释的 DOTA 图像包含 188 282 个实例，可用于发现和评估航拍图像中的小目标物体。无论从数量还是质量上来说，在同类型数据集中都具有很大优势。数据集部分样本示例如图 10 所示。

本文实验中将 DOTA 数据集分为训练集和验证集 2 个部分，按照 3∶1 的比例以及样本中目标是否清晰、数据情况是否良好的标准，选取训练集 8 533 张样本，验证集 2 726 张样本，共有 15 类目标样本，共计 11 259 张。按照 YOLO 的标注格式进行数据集处理。



图 10 数据集部分样本示例

Fig. 10 Example of a partial sample of a dataset

3.2 实验环境及模型训练

本文实验使用的环境配置为 Ubuntu20.04 操作系统，使用 A100 显卡，32 核 CPU，16 GB 内存，3 096 GB 系统盘，CUDA 版本为 11.6.0，Python 语言环境为 3.10。

本文实验总迭代次数为 300 次，batch size 设置为 16，特征通道数为 8。

3.3 评估指标与实验结果分析

评价指标采用目标检测任务中常用的 P、R、mAP50 和参数量。其中，P 表示模型检测结果的平均准确率，R 表示模型召回率，mAP50[13] 表示模型均值和平均精度，mAP50 的数值越大，模型检测数据的总体性能就越好，参数量越少，则模型体积越小。

为充分验证每个模块在实验操作中应用的有效性，进行多次消融实验分析和对比实验，分别对比只添加 CBAM 融合注意力机制和只添加 SPD-Conv 模块的实验结果。

3.3.1 消融实验

为减少实验赘述，本文将 YOLOv7 模型简称为 Y，将 YOLOv7+SPD-Conv 卷积的模型简称为 YS，将 YOLOv7+CBAM 注意力的模型简称为 YC，将 YOLOv7+SPD-Conv 卷积+CBAM 注意力的模型简称为 YSC。

实验结果如表 1 所示，添加了 SPD-Conv 模块后，准确率 P 提升了 0.2%，召回率 R 提升了 1.2%，mAP50 提升了 1.4%；添加了 CBAM 模块后，准确度 P 提升了 0.6%，召回率 R 提升了 2.6%，mAP50 提升了 2.3%；将两者均加入到网络模型中，准确率 P 提升了 3% 达到 83.7%，召回率 R 提升了 2.8% 达到 78.2%，mAP50 提升了 3.1%，各评价指标均有所提升，因此该消融实验可以证明单一模块与融合模块对于改进的 YOLOv7 网络模型的有效性。

3.3.2 对比实验

在保证配置环境及初始训练参数一致的情况下，本文将改进的 YOLOv7 网络模型与其他网络模型进行实验来验证改进网络模型的有效性，结果如

表 1 实验评估指标对比

Table 1 Comparison of experimental evaluation indicators %

模型	P	R	mAP50
Y	83.0	75.4	78.4
YS	83.2	76.6	79.8
YC	83.6	78.0	80.7
YSC	83.7	78.2	81.5

表 2 所示。

从表 2 可以看出, 改进后的 YOLOv7 网络模型在输入相同尺寸图片的情况下, mAP 值超过了其他经典网络模型, 更适合小目标检测场景。

表 2　实验评估指标对比

Table 2　Comparison of experimental evaluation indicators

模型	输入尺寸/像素	mAP50/%
Fast RCNN	720×720	54.5
YOLOv5	720×720	67.4
YOLOv7	720×720	78.4
PFP-Net[14]	720×720	79.3
SSD[15]	720×720	80.6
OURS	720×720	81.5

3.4　实验结果展示

实验对比结果展示如图 11 和图 12 所示, 在同一张检测样本中, 原始算法与改进算法进行对比, 证明在实际场景中, 通过上述方法改进的 YOLOv7 算法有效减少了漏检率, 并且在多数标签中, 检测置信度均有提升。



图 11　原始算法下的目标检测

Fig. 11　Object detection under the original algorithm



图 12　改进算法下的目标检测

Fig. 12　Improved object detection under the algorithm

从表 3 可以看出, 在飞机、篮球场、直升机、海港等几类类别目标上, 改进的 YOLOv7 算法获得了更高的检测精度, 在轮船等几类类别目标上, 检测精度基本保持一致。

表 3　DOTA 数据集上主要类别目标检测精度

Table 3　Accuracy of target detection for each category on the DOTA dataset　%

模型	Plane	Baseball-diamond	Helicopter	Roundabout	Tennis-court	Harbor	Large-vehicle	Basketball-court	Ship
YOLOv7	92.4	86.2	72.6	64.9	92.4	83.7	89.4	87.2	89.0
OURS	96.7	88.9	77.8	70.2	95.8	87.7	91.3	88.1	88.8

4　结　论

1) 本文方法可实现较为优异的小目标检测性能, 在处理后的 DOTA 航拍数据集上 P 达到 83.7%, R 达到 78.2%, 均值平均精度 mAP 达到 81.5%。

2) 本文方法在 Yolov7 算法中引入卷积块注意力模块, 通过空间通道注意力提升模型的特征提取能力, 引入了 SPD-Conv 模块, 实现了无信息损失的下采样。

3) 针对航拍场景中普遍存在的微小目标(如车辆、小型设施等), 本方法在 DOTA 航拍数据集上 mAP 较基准模型提升 3.1%, 显著改善了小目标的定位与分类准确性。

参考文献 (References)

[1] NAJIBI M, SAMANGOUEI P, CHELLAPPA R, et al. SSH: single stage headless face detector[C]//Proceedings of the 2017 IEEE International Conference on Computer Vision. Piscataway: IEEE Press, 2017: 4885-4894.

[2] ZHANG L L, LIN L, LIANG X D, et al. Is faster R-CNN doing well for pedestrian detection?[C]//Proceedings of the European Conference on Computer Vision– ECCV 2016. Berlin: Springer, 2016: 443-457.

[3] RAGHUNANDAN A, Mohana, RAGHAV P, et al. Object detection algorithms for video surveillance applications[C]//Proceedings of the 2018 International Conference on Communication and Signal Processing. Piscataway: IEEE Press, 2018: 563-568.

[4] UIJLINGS J R R, VAN DE SANDE K E A, GEVERS T, et al. Selective search for object recognition[J]. International Journal of Computer Vision, 2013, 104(2): 154-171.

[5] REDMON J, DIVVALA S, GIRSHICK R, et al. You only look once: unified, real-time object detection[C]//Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2016: 779-788.

[6] REDMON J, FARHADI A. YOLO9000: better, faster, stronger[C]//Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2017: 6517-6525.

[7] REDMON J, FARHADI A. Yolov3: an incremental improvement [EB/OL]. (2018-04-08)[2021-03-25]. https://arxiv.org/10.48550/arxiv.1804.02767.

[8] BOCHKOVSKIY A, WANG C Y, LIAO H Y M. Yolov4: optimal speed and accuracy of object detection[EB/OL]. (2020-04-23)[2021-04-15]. https://arxiv.org/abs/2004.10934.

[9] WANG C Y, BOCHKOVSKIY A, LIAO H M. YOLOv7: trainable bag-of-freebies sets new state-of-the-art for real-time object detectors[C]//Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2023: 7464-7475.

[10] WOO S, PARK J, LEE J Y, et al. CBAM: convolutional block attention module[C]//Proceedings of the Computer Vision – ECCV 2018. Berlin: Springer, 2018: 3-19.

[11] SUNKARA R, LUO T. No more strided convolutions or pooling: a new CNN building block for low-resolution images and small objects[EB/OL]. (2022-08-07)[2022-08-21]. https://arxiv.org/abs/2208.03641v1.

[12] XIA G S, BAI X, DING J, et al. DOTA: a large-scale dataset for object detection in aerial images[C]//Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2018: 3974-3983.

[13] CHEN Y C, ZHENG W S, LAI J H, et al. An asymmetric distance model for cross-view feature mapping in person reidentification[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2017, 27(8): 1661-1675.

[14] Zhao H, Shi J, Qi X, et al. Pyramid scene parsing network[EB/OL]. (2014-04-27)[2022-08-22]. https://doi.org/10.48550/arxiv.1612.01105.

[15] BERG A C, FU C Y, SZEGEDY C, et al. SSD: single shot MultiBox detector[EB/OL]. (2015-03-30)[2023-09-16]. https://doi.org/10.1007/978-3-319-46448-0_2.

Improved YOLOv7 method for aerial small target detection in aerial photography

LIU Yinuo，ZHANG Qi*，WANG Rong，LI Chong

(College of Information and Cyber Security，People's Public Security University of China，Beijing 102600，China)

Abstract：This paper proposes an improved YOLOv7-based aerial small target detection method to address the high rates of missed and false detections in current detection technologies for aerial small target detection tasks. First, a CBAM fusion attention mechanism is incorporated into the backbone network, allocates weights reasonably in both spatial and channel-wise of the feature map, suppresses background interference and improves detection accuracy. The second is the SPD-Conv module, which removes the original convolutional module's cross-convolutional and pooling layers, improves feature representation learning efficiency, and mitigates fine-grained information loss in low-resolution images and small targets refinement detection. Finally, the improved YOLOv7 is evaluated on a processed DOTA aerial dataset. According to the results, it outperforms the original YOLOv7 by 3.1%, achieving 83.7% precision, 78.2% recall, and 81.5% average accuracy on the dataset. The improved algorithm effectively reduces missed and false detections, demonstrating a strong performance.

Keywords：YOLOv7; small target detection; attention mechanism; convolutional neural network; computer vis

Received：2022-06-28；**Accepted**：2023-06-28；**Published Online**：2023-12-06 16：03

URL：link.cnki.net/urlid/11.2625.V.20231206.0926.001

Foundation items：National Natural Science Foundation of China (62076246); Double First-Class Special Project in Security Engineering at People's Public Security University of China (2023SYL08)

* **Corresponding author**. E-mail：qi.zhang@ppsuc.edu.cn

http://bhxb.buaa.edu.cn jbuaa@buaa.edu.cn
DOI: 10.13700/j.bh.1001-5965.2023.0416

多目视觉下基于融合特征的密集行人跟踪方法

黄煜杰，陈凯*，王子源，王紫腾
(南京航空航天大学 机电学院，南京 210016)

摘　　要：针对当前大部分计算机视觉跟踪方法仍不能有效解决目标受遮挡以及在摄像机视角中消失后重现等问题，基于融合特征相关性对多目标行人跟踪方法进行了研究：基于高斯混合模型（GMM）更新行人特征池以减少人员密集所导致的特征污染；基于K-means算法动态计算目标特征相似性阈值；利用融合特征相似性关联行人特征，加入单应性约束校验以判定行人的新增与重现。在公开数据集Shelf上进行实验，结果显示所提方法平均精确度相较其他算法分别提升16.05%、7.39%，平均成功率分别提升16.04%、4.16%。完整视频流下的平均错跟率为10.11%，在控制错跟数量方面取得显著效果之外还能够在行人重现后有效关联至原目标。

关 键 词：目标跟踪；特征提取；多目视觉；动态阈值；融合特征；高斯混合模型；相关性匹配

中图分类号：TP391.4

文献标志码：A　　**文章编号**：1001-5965(2025)07-2513-13

随着机器学习相关领域的高速发展，神经网络和深度学习算法在效率等方面的不断更新优化对计算机视觉领域的研究产生了极大的推进作用。同时，近年来随着城镇人口密度不断扩大，人与人之间的联系不断增强，出于对社会治安管理和城市信息化发展需求，衍生出了对多目标行人进行监控和轨迹跟踪的需求。因此，利用计算机视觉进行多目标行人跟踪任务具有重要的现实意义。

多目标跟踪，一般指在目标数量未知的情况下，对摄像机视角下的行人、物体等多目标进行检测并对不同目标赋予不同ID，以便对其进行轨迹分析、预测等工作。而在多目视觉下，多目标跟踪任务则不仅要在时间流中进行，还需要在同一时刻下各个视角之间完成。由于各视角图像中物体、人员所处的相对位置、背景等存在差异，尤其在出现2个相似目标发生重叠、遮挡、背景失真等复杂环境时，多目标的跟踪难度提高，相应地对跟踪算法也提出了更高要求。现有的多目标跟踪算法主要集中在基于单目视觉的研究上，主要从单目视觉下跟踪的准确率、目标ID识别率等方面进行提升和优化[1-5]。当前针对多目标跟踪任务的相关研究相对较少，现有的算法虽然在跟踪效率、成功率等方面做出了不同程度的贡献[6-7]，但仍然不能很好地解决被跟踪目标受其他目标或环境遮挡、目标从摄像机视角中完全消失后又重新出现等问题[8]。多数情况下会出现被跟踪目标受遮挡后跟踪中断、跟踪对象漂移、出现跟踪误差等问题[9-11]；而目标消失后重现所导致的跟踪ID错误、漏跟、无法对应至消失前的目标ID等情况更难以解决。

1 相关工作

近年来，计算机视觉在多目标跟踪方面获得了广泛的发展。在相关领域产生了许多优秀的算法，

收稿日期：2023-06-28；**录用日期**：2023-09-01；**网络出版时间**：2023-10-11 13:00
网络出版地址：link.cnki.net/urlid/11.2625.V.20231010.0958.001
基金项目：国家自然科学基金(52202417)；中国博士后科学基金(2022TQ0155,2022M721605)；虚拟现实技术与系统全国重点实验室(北京航空航天大学)开放课题基金(VRLAB2023A02)
* **通信作者**. E-mail：chen_kai@nuaa.edu.cn

引用格式：*黄煜杰，陈凯，王子源，等. 多目视觉下基于融合特征的密集行人跟踪方法[J]. 北京航空航天大学学报，2025，51(7)：2513-2525. HUANG Y J，CHEN K，WANG Z Y，et al. A dense pedestrian tracking method based on fusion features under multi-vision[J]. Journal of Beijing University of Aeronautics and Astronautics，2025，51(7)：2513-2525 (in Chinese).*

如 DeepSORT[10]、MOTDT[11]、ByteTrack[12] 等，这些算法在多目标跟踪领域从计算效率、跟踪准确率、解决遮挡问题等方面做出了突破性的创新。

文献 [10] 基于 SORT[9] 目标跟踪框架进行改进，其延用 SORT 中的核心跟踪框架，利用卡尔曼滤波器和匈牙利算法预测当前时刻的位置并进行数据关联，为每个对象分配目标 ID。引入深度学习模型，采用一个简单的卷积神经网络（convolutional neural network, CNN）在每次跟踪结束后提取并保存被检测目标的外观特征。深度神经网络和卡尔曼滤波器的结合可以使跟踪具有较高的精度以及较强的鲁棒性，但卡尔曼滤波器无法对长时间受遮挡的目标进行准确的状态预测。此外，作者巧妙地将目标的运动特征与外观特征进行融合以实现目标的关联匹配，这有助于改善目标受遮挡时的跟踪效果，减少目标 ID 的跳变情况。但由于滤波算法消耗的状态空间较大，当目标超过最大保存时间而未被关联时，该目标的融合特征将被删除，无法从根源上解决目标长时间消失后重现的问题。

文献 [11] 在实时性的基础上将检测和跟踪结果结合起来，并基于深层神经网络选择最佳目标候选项，以此解决在线跟踪中的不可靠检测问题；提出分层数据关联策略，使用不同的特征分层地将目标 ID 与不同的候选框相关联，以避免其他不需要的对象和背景污染。同时，设计了轨迹段置信度计算模块来评估使用时序信息的卡尔曼滤波器的准确率，结合分层关联方法，使得 MOTDT 模型可以以实时速度运行，但由于卡尔曼滤波不适用于长期跟踪，因此该模型对目标遮挡、消失后重现等情况存在很大的跟踪不稳定性。

文献 [12] 是一种基于 Tracking-by-Detection 范式的跟踪方法。目前，大多数多目标跟踪框架仅通过关联相似度高于阈值的检测框来获取 ID，这会带来大量由于遮挡而产生的漏检和碎片化轨迹。文献 [12] 提出的一种新的数据关联模型 BYTE，不仅对高置信度框进行关联，还会将低置信度框与未匹配的跟踪对象进行二次关联计算与背景过滤，利用其与轨迹的相似性来恢复真实目标。该方法对部分遮挡而导致的低置信度检测框具有非常好的跟踪效果，但其不仅对检测器提出了较高的要求，在单目视觉下也无法处理行人被完全遮挡后的关联匹配问题。

经过实际的实验表明，大多数多目标跟踪算法对目标完全遮挡、长时间消失后重现等情况存在较严重的跟踪误差、漏跟、目标 ID 跳变等问题。为了解决这些问题，本文在融合特征提取与更新、特征相关性计算、行人新增与重现判定等几个方面进行了创新性研究。

2 行人跟踪框架介绍

本文的多目标行人跟踪框架主要分为行人目标检测、融合特征提取与更新、特征相关性计算、行人新增与重现判定四个模块。该跟踪框架总体实现流程如图 1 所示，本文首先采用 YOLOv5[13] 算法对视频流下的多视角图像进行行人检测，获取每一帧下每个视角中每一个行人的检测框；其次，基于分区特征相关性进行目标特征融合，提取每一帧中所有待关联行人特征；选取查询项和匹配项计算特征相似度矩阵，基于 K-means[14] 确定动态阈值，并对所有目标进行特征匹配；此外，结合基于特征相关的行人新增与重现判定以及基于单应性约束的关联结果校验，实现多目标行人关联；将已关联到的所有特征存入特征池中，基于高斯混合模型（Gaussian mixture model, GMM）更新特征池，并在多目视觉下选出每个目标 ID 最具代表性的特征作为下一次特征匹配的模板，淘汰重复、简单特征，完成多目视觉下的多目标行人跟踪任务。

2.1 融合特征提取与更新

通过 YOLO 算法获取所有行人检测框后，需对图像中相应的待关联行人目标进行特征提取。目前相关领域对图像特征提取的算法有很多，但大多无法解决部分/完全遮挡、消失并重现后的 ID 对应等问题。本节从特征相关性出发，基于目标图像的分区特征相关性提取特征；此外，提出了一种基于 GMM 的特征池更新策略，分别更新所有目标的特征模板，使得特征池在每一帧下均能维持代表性。为了更好地促进多目视觉下密集行人的跟踪研究，本文提供了开源代码：https://github.com/HYJtooo/Multivisual-fusion-features-tracking。

2.1.1 基于分区特征相关性的融合特征提取

普通的行人局部特征提取方法只以单一、固定的顺序连接单个个体的局部特征，而忽略身体部位之间的内在关系，使得许多部位相互孤立，进而导致对应部位具有相似属性的不同个体身份发生混淆。本文基于分区特征相关性的特征提取方法，则是一种考虑了个体身体部位与其余部位之间关系的特征网络。即使在行人图像缺少身体部位的情况下，也能提供较为准确的特征表示。

本文所提取的图像对象均为所检测到的行人检测框，对正常站立的行人而言，主要身体部位包括头部、肩部、腰部（肘部）、臀部（手部）、膝部、脚部六部分。首先提取所有行人检测框所对应的目

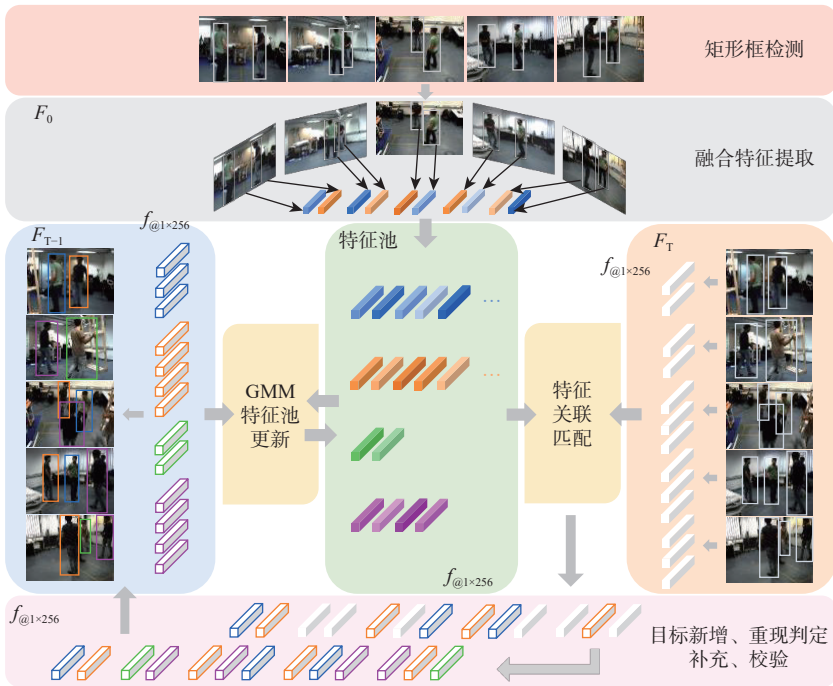


图 1　融合特征相关多目标跟踪方法总体流程图

Fig. 1　Overall process of multi-target tracking method based on feature correlation fusion

标行人图像, 基于 ResNet-50 网络从图像中提取一个大小为 $H\times W\times C$ 的特征图, 将其平均划分为 6 个水平分区[15], 使其能够有效区分上述几个主要的身体区域, 如图 2 所示。所提取的分区级特征能够更有效地表达出所代表身体部位包含的信息, 有利于后续的特征融合与关联匹配。

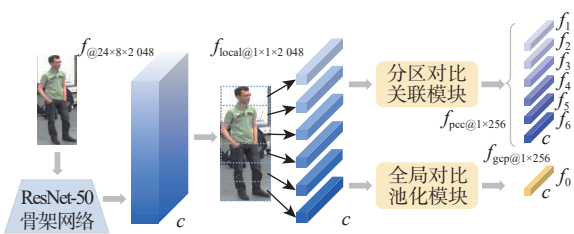


图 2　基于特征相关性的融合特征提取

Fig. 2　Fusion feature extraction based on feature correlation

对每个分区特征进行最大全局池化处理, 得到 6 个尺寸为 $1\times1\times C$ 的分区级特征; 提出分区对比关联模块 PCC 和全局对比池化模块 GCP, 将分区级特征分别输入 2 个模块中, 得到每一个分区特征与其他分区的相关性特征 $f_1\sim f_6$ 以及全局对比特征 f_0。

在 PCC 模块中, 需要得到每个身体部位与除该部位之外的其他部位之间的关系信息。如图 3(a) 所示, 分别以每一个分区级特征作为主特征 $\bar{p}_i$, 通过聚合除 $\bar{p}_i$ 之外其他分区级特征以得到该分区 i 的相关性特征 f_i。

$$\begin{cases} f_i=\bar{p}_i+R(\overline{C}(\bar{p}_i,\bar{r}_i)) & i=1,\cdots,6 \\ \bar{r}_i=\bar{A}(\overline{C}(\bar{p}_j)) & j=1,\cdots,6,\ j\neq i \end{cases} \tag{1}$$

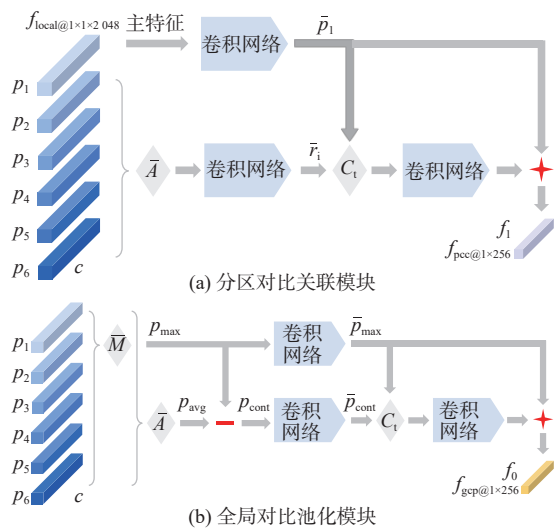


图 3　融合特征提取流程

Fig. 3　Fusion feature extraction process

式中: R 为一个由 1×1 卷积、批量归一化[16] 和 ReLU[17] 层组成的子网络; $\bar{A}$ 为对特征进行的平均池化操作; $\overline{C}$ 为特征与特征的串联; $\bar{r}_i$ 为除分区 i 外所有分区级特征串联后进行平均池化得到的聚合特征, 其包含除主特征之外的其余所有分区级特征信息。

式(1)中, $R(\overline{C}(\bar{p}_i,\bar{r}_i))$ 为一个与主特征部位相关联, 又同时包含其他部位信息的分区级融合特征。若 $\bar{p}_i$ 存在缺陷(如遮挡引起特征污染), 但与其他身体部位的内在关联性是客观存在的。加入分区级融合特征不仅使 f_i 表达 $\bar{p}_i$ 本身的信息, 还包含与其他身体部位的关联信息, 弥补 $\bar{p}_i$ 自身问题, 进而使 $f_1\sim f_6$ 更利于表达目标特征, 在密集人群环境中尤

其使遮挡问题更具鲁棒性。

在GCP模块中，主要目的在于获取一个最能体现与其他特征差异的全局特征f_0。如图3(b)所示，在获取f_0的过程中，对比特征p_{cont}为最大池化特征$p_{\max}$与平均池化特征p_{avg}的差值，因此其包含所有水平分区中更具有区分性的一部分特征。

$$f_0 = \bar{p}_{\max} + R(\overline{C}(\bar{p}_{\max}, \bar{p}_{\text{cont}})) \tag{2}$$

式中：R为一个由1×1卷积、批量归一化[16]和ReLU[17]层组成的子网络；$\overline{C}$为特征与特征的串联；对比特征p_{cont}通过卷积操作得到$\bar{p}_{\text{cont}}$。

按式(2)进行全局对比特征计算，获得f_0不仅有整体特征中最突出的一部分，还包含剩余所有分区中相对更突出的一部分，使整体特征更具有代表性。在f_0中加入$\bar{p}_{\text{cont}}$信息，能够减少背景杂波对图像特征产生影响[18]，使特征更具识别力。

通过融合全局对比特征f_0和分区相关性特征$f_1 \sim f_6$，最终可以依据目标检测框提取到多视角下基于分区特征相关性的所有待关联行人特征。

2.1.2 特征池更新策略

传统的单目视觉目标跟踪器中，行人的模型集由一个连续的样本序列组成。但外观的缓慢变化会导致大量的冗余特征，同时之前的特征将会随着时间流而被遗忘[10,19]。在多目视觉下进行多目标跟踪时，目标行人的动作、姿态等会随时间发生动态变化，而在不同视角下，摄像机所拍摄到的同一个目标行人图像也会存在不同程度的遮挡、背景变换、相对位置变化等情况。为了保证多视角下行人跟踪的准确性与成功率，本文提出基于GMM的特征池更新策略，每次更新时将不同视角、不同时间点下的行人特征收集为高斯分量的混合物[20](特征模板)，每个分量都从不同方面表达行人特征，使从特征池中选出的所有特征模板持续适用于当前环境和状态下的多视角多目标跟踪，保持其泛化能力。

GMM对普通高斯模型进行扩展，使用多个高斯分布组合来刻画数据分布，提高了模型的自由度及鲁棒性。在计算上GMM是基于EM算法的推导，其主要分为求期望(E步)和目标最大化(M步)，如下：

$$Q(\theta, \theta^{(i)}) = \sum_{Z_H} P(Z_H | X_{T,C}, \theta^{(i)}) \ln P(X_{T,C}, Z_H | \theta^{(i)}) \tag{3}$$

$$\theta^{(i+1)} = A \max_{\theta} Q(\theta, \theta^{(i)}) \tag{4}$$

式中：$X_{T,C}$为特征池中某行人特征集合；T为特征所处时间；C为特征所处的视角编号；隐变量Z_H为特征所属高斯分布的类别；H为特征对应的行人ID；$\theta^{(i)}$为第i次迭代后参数的估计值。

利用Q函数计算各个高斯的均值与协方差矩阵。初始参数：$X_{T,C} = (x_{t_1,c_1}, x_{t_2,c_2}, \cdots, x_{t_N,c_N})$，$Z_H = (z_{h_1}, z_{h_2}, \cdots, z_{h_N})$，$\theta = (p, \mu, \Sigma)$。$x_{t_N,c_N}$为特征池中第$N$个特征所处的时间与视角。各高斯分布、均值与协方差矩阵可初始化为$p = (p_1, p_2, \cdots, p_K)$，$\mu = (\mu_1, \mu_2, \cdots, \mu_K)$，$\Sigma = (\Sigma_1, \Sigma_2, \cdots, \Sigma_K)$。参数$\theta$包含隐变量$Z_p$的概率分布，$N$为该目标在特征池中存储的特征数量，$K$为高斯分布数量。将设定好的参数代入并化简后可得到新的Q函数，如下：

$$Q(\theta, \theta^{(i)}) = \sum_{i=1}^{N} \sum_{k=1}^{K} (\ln p_k + \ln \phi(x_{t_i,c_i} | \mu_k, \Sigma_k)) P(z_{h_i} = C'_k | x_{t_i,c_i}, \theta^{(i)}) \tag{5}$$

$P(z_{h_i} = C'_k | x_{t_i,c_i}, \theta^{(i)})$为隐变量在当前参数下的后验概率，其内部所有参数都已确定，因此将其用新变量γ_{ij}表示如下：

$$\gamma_{ij} = \frac{p_j \phi(x_{t_i,c_i} | \mu_j, \Sigma_j)}{\sum_{k=1}^{K} p_k \phi(x_{t_i,c_i} | \mu_k, \Sigma_k)} \tag{6}$$

将新的Q函数代入式(4)中求解$\theta^{(i+1)}$，由此得到下一时刻的各高斯分布、均值与协方差矩阵：

$$p_j^{(i+1)} = \frac{1}{N} \sum_{i=1}^{N} \gamma_{ij} \tag{7}$$

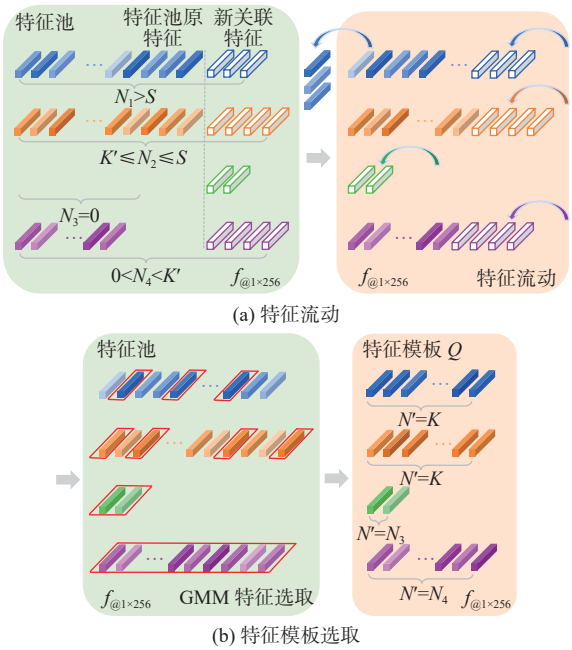
$$\mu_j^{(i+1)} = \frac{\sum_{i=1}^{N} x_{t_i,c_i} \gamma_{ij}}{\sum_{i=1}^{N} \gamma_{ij}} \tag{8}$$

$$\Sigma_j^{(i+1)} = \frac{\sum_{i=1}^{N} (x_{t_i,c_i} - \mu_j)^{\mathrm{T}} (x_{t_i,c_i} - \mu_j) \gamma_{ij}}{\sum_{i=1}^{N} \gamma_{ij}} \tag{9}$$

通过式(7)所得各高斯分布概率密度函数$p_j^{(t+1)}$，即可获取所提取出的该行人所有代表性特征$x_{t,c}$，以此完成特征模板的选取。

基于GMM的特征池整体更新策略分为两步：如图4(a)，每一帧的关联匹配结束后，将该帧下每个目标ID的所有新关联特征添加进特征池中；单个目标在特征池中的存储上限为S；若特征池中目标ID的特征数量超过S，则在添加新特征前淘汰最早添加的特征，使特征数量维持在存储上限。若当前帧未出现新关联特征，则维持现有特征。特征流动可表示为

$$X_{T,C} = (x_{t_1,c_1}, x_{t_2,c_2}, \cdots, x_{t_N,c_N}) \qquad N \leqslant S \tag{10}$$



(a) 特征流动

(b) 特征模板选取

图 4　基于 GMM 的特征池更新过程

Fig. 4　Feature pool updating process based on GMM

$$X_{\text{new}}=(x'_{t'_1,c'_1},x'_{t'_2,c'_2},\cdots,x'_{t'_W,c'_W}) \tag{11}$$

$$X_{\text{del}}=(x_{t_1,c_1},x_{t_2,c_2},\cdots,x_{t_D,c_D}) \qquad D=0 \text{ 或 } D=N+W-S \tag{12}$$

$$X_{T,C}^{(t+1)}=X_{T,C}-X_{\text{del}}+X_{\text{new}} \tag{13}$$

式中: $X_{T,C}$为特征池中某行人特征集合; X_{new}为新增特征; X_{del}为待删除旧特征; N为现有特征数; W为新增特征数; D为需删除特征数; $X_{T,C}^{(t+1)}$为特征流动完成后特征池中该行人特征集合。

如图 4(b) 所示, 当特征池中的目标 ID 特征超过K'个时, 利用 GMM 在该目标 ID 的所有特征中选取K个最具代表性的特征作为下一帧关联匹配的特征模板Q, 否则直接取现有全部K'个特征。

如图 5 所示, 单个目标在特征池中最大存储数量S会影响关联成功率。设置$S=50$能在减少计算次数的情况下使特征池保持适宜的特征更新速率; 此外, GMM 模型的组件数量K也在很大程度上影响算法效率。为在选取到足够代表性特征的同时尽量减少特征相似度计算次数, 设置$K=10$; 而在新增目标中, 目标特征较为相似, 需要存储更多不同因素影响下的特征。设置$K'=15$适当增加初始阶段存储数量以保持足够的泛化能力。

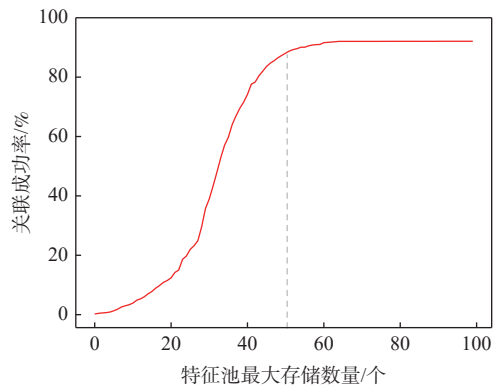


图 5　特征池最大存储数量选取

Fig. 5　Maximum number of feature pool storage selection

如图 6 所示, 由于行人姿态随着时间流不断发生变化, 在利用 GMM 选取特征模板时可将不同高斯分量对应的行人图像分别提取出来。这些模板包含不同角度、光照、背景和遮挡等影响下所形成的特征, 故在下一次特征匹配中, 各视角下的行人特征都能够在特征模板中被匹配到; 同时, 由于特征池更新过程中不断加入新关联特征并删除旧特征, 模板中的行人姿态与所处环境等始终与当前状态相近, 使其更适用于当前状态下的跟踪[21]。

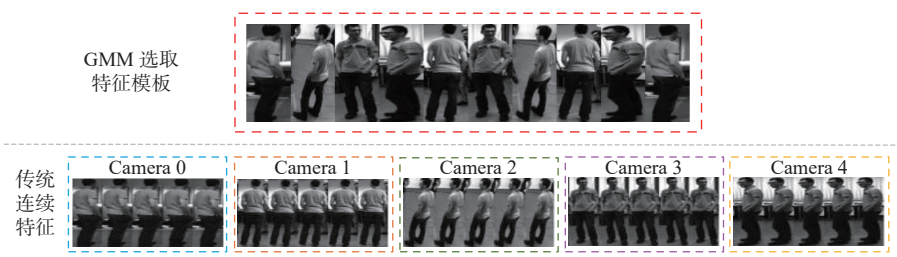


图 6　GMM 更新特征池与传统连续特征图对比

Fig. 6　Using GMM update feature pool and traditional continuous feature maps

此外, 对于已关联到的目标 ID, 当出现目标在各视角中完全消失、完全受遮挡等情况时, 特征池中的存储特征不会随时间和环境的变化而被污染, 也不会因状态空间问题而对后续的关联匹配产生影响[22-24]。因此, 当该目标重现时, 该框架能够重新跟踪并关联到之前的 ID 中, 有效避免 ID 跳变、跟踪漂移等问题。

2.2　特征相关性计算

2.2.1　特征相关相似度矩阵计算

在获取当前帧所有视角中的每个待关联行人特征后, 需要利用特征池选出的特征模板对所有待关联特征进行匹配: 如图 7 所示, 以基于 GMM

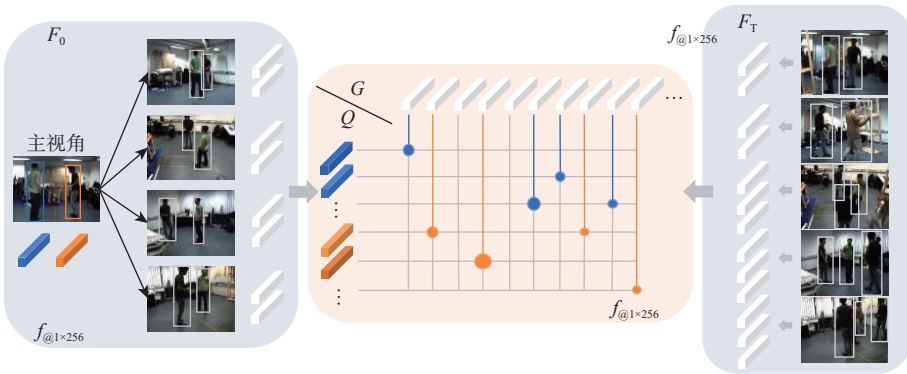


图7 特征相关的相似度矩阵计算

Fig. 7 Similarity matrix calculation of feature correlation

更新特征池选取的Q作为查询项，一个目标ID经GMM选出的10个特征为一组；当前帧下所有待关联特征为匹配项G，通过Q与之间的距离计算，可以获取所有G在每一组Q下的基于特征相关的相似度矩阵，每个Q与G的计算结果越小，两特征越相近。

由于多目视觉下图像之间的背景、视角存在差异，可能会导致一组查询项中的某些特征与匹配项存在较大不同。正常情况下，取某匹配项与一组Q的所有计算结果的均值作为其相似度计算结果；若某匹配项与一组Q中的相似度得分存在较大差异，则利用聚类算法进行相似度矩阵计算和判定。该方法有效排除了由于少量特征相关性误差而导致的特征关联错误，使跟踪结构具有非常好的鲁棒性。同时，以特征而非图像作为输入，有效提高了相似度矩阵计算效率[25-26]。

2.2.2 动态阈值确定

确定动态阈值，是对特征相关性计算结果进行划分，判断该结果是否达到关联成功的标准。但该标准在不同时间、视角下的每一组关联中会由于光照、行人姿态变化、背景干扰等原因[27]而不同。当该标准存在差异时，普通的人为固定阈值策略就会存在跟踪漂移、关联误差及成功率降低等情况，累积误差也会持续升高。K-means作为一种对异常值敏感的无监督聚类算法，其特点十分贴合本文所需。本文提出基于K-means[14]的动态阈值策略，对特征相似度矩阵进行2次聚类，确定每一帧行人特征关联匹配的动态阈值，以适应当前环境下的多视角跟踪。

在获取相似度矩阵时，错误对象之间的关联无法避免，即使目标正确，也会出现由于相机角度不同引起的计算结果偏差。因此在利用相似度结果进行关联前，需要先排除这些干扰特征。如图7所示，以一组Q为例，其相似度结果为一个$m\times n$矩阵$\boldsymbol{M}_{p,m\times n}^{(t)}$，其中，$p$为目标ID，$t$为当前时刻，$m$为该组$Q$数量，$n$为当前帧所有$G$数量。进行第一次聚类，所需类别数量为2。如式(14)所示，利用K-means对$\boldsymbol{M}_{p,m\times n}^{(t)}$进行列操作，即对每一个$G$在一组$Q$中的所有结果聚类，取聚类中心较小一类所有元素的均值作为该匹配项在该组查询项当中的相似度结果，由此获得一个$1\times n$的向量$\boldsymbol{L}_{p,1\times n}^{(t)}$。此时利用$\boldsymbol{L}_{p,1\times n}^{(t)}$中的$n$个元素按式(15)进行动态阈值计算，在该范围内确定关联成功的标准。之后进行第二次聚类，在所有G中确定与该组Q匹配的正确特征，利用K-means对$\boldsymbol{L}_{p,1\times n}^{(t)}$进行行操作，如式(16)所示，找到$m'$个聚类中心较小一类的元素，同时判断各元素是否满足动态阈值，完成该组Q的匹配过程。

$$\boldsymbol{L}_{p,1\times n}^{(t)} = A\left(K_{\min}(\boldsymbol{M}_{p,m\times n}^{(t)}[,:])\right) \tag{14}$$

$$D_y = \min(L_{p,1\times n}^{(t)}) + (\max(L_{p,1\times n}^{(t)}) - \min(L_{p,1\times n}^{(t)}))T_{\text{ori}} \tag{15}$$

$$\boldsymbol{L}_{p,m'}^{'(t)} = K_{\min}(\boldsymbol{L}_{p,1\times n}^{(t)}[:]) \tag{16}$$

式中：[,:]为列向量操作；[:]为行向量操作；$K_{\min}$为取K-means后聚类中心最小一类的所有元素；A为取平均值操作；T_{ori}为初始固定阈值；$\boldsymbol{L}_{p,1\times n}^{(t)}$为最终选出的成功关联项。

如图8所示，匹配项的浅色特征点是在第一次聚类中被排除的项，直径较大点即为第一次聚类结果$\boldsymbol{L}_{p,1\times n}^{(t)}$，灰色平面为该次关联中由$\boldsymbol{L}_{p,1\times n}^{(t)}$得到的动态阈值。不难发现，在成功关联组(0,2,5,7,8)中，若干扰特征存在，其相似度结果总是大于或远大于正常值。因此，干扰特征与正确特征的数量比例不会影响聚类结果，仅需找到聚类中心较小一类的元素即为所需分类，可有效降低特征干扰。

基于K-means的动态阈值确定方法可以适应不同环境下的特征相关相似度计算结果，尤其是在多目视觉下的动态多目标跟踪中，有效减少了跟踪误差以及由于背景动态变化而引起的特征关联问题，提高了目标特征跟踪的准确性。

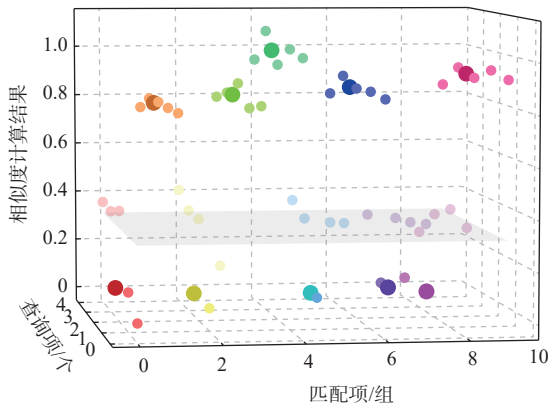


图 8　K-means 动态阈值确定示例

Fig. 8　Example of K-means dynamic threshold determination

2.2.3　行人特征匹配与跟踪

多目标行人特征的关联匹配是将特征池中基于 GMM 选出的特征模板以及当前帧下的所有待关联特征进行特征相关的相似度计算得到的[28]。而在初始帧特征池为空时, 则需要单独在各视角中进行行人关联匹配。在初始帧F_0中, 通过 YOLO[13] 算法可得到每个视角下所有的行人检测框, 选取待关联检测框数量最多的视角为主视角; 在多视角间进行目标特征的关联匹配: 将主视角中所有特征作为Q_0, 其余视角的所有特征作为G_0, 进行相似度矩阵计算。将其余视角的目标特征与主视角关联, 实现初始帧内不同视角间的行人跟踪, 如图 6 所示; 之后, 将初始帧跟踪关联到的所有目标特征分别加入特征池, 开始多目视觉下的多目标跟踪过程。

基于 K-means[14] 的动态阈值确定以及目标特征关联任务虽能够动态地修正阈值以实现不同环境的适应性, 但鉴于 K-means 本身所存在的对初始簇中心敏感、不适合离散分类等缺陷, 在实际的目标特征关联中仍会存在目标漏跟等情况。为了解决该问题同时进一步提高跟踪准确率, 本文采用交并比 I 与中心偏移率 C_x 两项指标对行人检测框进行突变检测。对于已关联到的目标行人检测框, 将其与上一帧同视角下同 ID 检测框进行交并比 I 与中心偏移率 C_x 计算, 如下:

$$I=\frac{S_{t-1}\cap S_t}{S_{t-1}\cup S_t} \tag{17}$$

$$C_x=\frac{|x_t-x_{t-1}|}{x_{t-1}} \tag{18}$$

式中: S_{t-1},S_t,x_{t-1},x_t分别为上一帧与当前帧同视角中同 ID 检测框面积以及检测框中心点的 x 坐标。

当 I 与中心偏移率 C_x 均小于所设阈值时, 认为该目标特征关联匹配成功。此外, 在当前帧的目标特征关联结束后, 若某视角中前一帧出现的目标 ID 在当前帧的同视角下未出现时, 将会触发漏跟检测: 对上一帧的目标 ID 检测框与当前帧下的所有未关联检测框进行 I 与中心偏移率 C_x 计算, 若能同时满足两项指标组成的突变检测, 可认为该检测框与该目标 ID 成功关联。通过突变检测作为多目标关联匹配的补充, 有效减少了多目标跟踪的失误率, 并大幅降低了漏跟情况。

2.3　行人新增与重现判定

在多目视觉下的多目标动态跟踪中时常出现行人新增以及从视角中消失后重现的情况, 而这也是大部分多目标跟踪算法所存在的问题[29-30]。本节提出基于特征相关的行人新增与重现判定方法, 同时提出基于单应性约束的关联结果校验, 进一步对行人检测框进行特征关联补充。

2.3.1　特征相关的行人新增与重现判定

在完成目标特征关联与突变检测后, 若当前帧仍然在 60% 以上的视角中出现 2 个以上的未关联特征, 将会触发行人新增判定: 如图 9 所示, 以 5 个视角中分别出现一个未关联特征为例, 将未关联的所有特征同时作为Q_n和G_n进行相似度矩阵计算。对所得到的n阶方阵, 取其上三角矩阵进行目标特征关联。结合基于 K-means 的动态阈值确定方法, 当出现至少 2 个特征存在于至少 2 个视角且相似度矩阵计算结果低于动态阈值时, 判定出现新增行人。判定出现新增目标后, 将对应的目标信息添加进Q当中, 并在下一次目标特征关联匹配时进行该新增目标的跟踪。

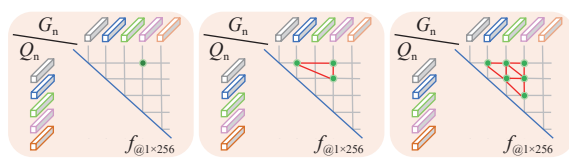


图 9　特征相关的行人新增判定示意图

Fig. 9　Pedestrian addition decision diagram based on feature correlation

结合特征池更新策略, 若已关联的目标 ID 在所有视角中完全消失, 其特征池中的特征及特征模板仍会保留, 因此当完全消失后的目标在任意视角重新出现时, 其特征仍会被关联至原 ID 中, 解决了目标重现的 ID 关联匹配问题。

2.3.2　基于单应性约束的关联结果校验

由于单应性约束算法不具有关联匹配的主动性, 因此本文在特征关联完成后利用该方法对关联结果进行验证工作。此时各相机矩阵参数以及每个视角下的目标 ID 位置等信息都成为已知量。利用相机矩阵便可以将各视角下目标 ID 的二维中心坐标投影并转换为三维空间位置坐标。各视角中

的同一个目标在投影至三维空间时其中心坐标应集中在同一点附近。

基于单应性约束的校验过程对所有行人检测框进行错跟、漏跟校验[31]。通过对当前帧目标ID在各视角中的位置信息进行两两投影计算，获取其投影后的中心位置坐标。如图10所示，输入一组具有同ID编号的检测框二维坐标，若其在三维空间的投影距离均在较小范围内，则认为该组关联均成功；反之，若某视角出现错跟等情况，则与该点对应的三维坐标距离计算均会出现较大偏差，则很容易判别并校正错误项。此外，如图10虚线所示，对于部分应关联而未关联到的漏跟现象，该方法也能够展现出色的效果。

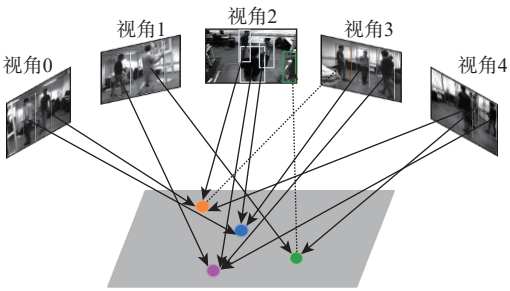


图10 基于单应性约束的关联结果校验

Fig. 10 Verification of corresponding results based on homography constraints

该校验方法在基于融合特征相关性的多目标跟踪结果的基础上进一步减少了多目视觉跟踪过程中偶然出现的错跟、漏跟现象，使跟踪准确率和成功率进一步提升。

3 实验结果与分析

3.1 公开数据集与评价指标介绍

本文的跟踪方法不仅需在多视角下进行多目标跟踪，还需要保证一定的实时性，因此仅基于图片集进行测试具有一定的局限性，需要选择在一定空间范围内动态行人的视频流作为数据集对模型进行测试。本文选择公开数据集Shelf[32]作为模型的测试集。Shelf数据集是由室内的4个人近距离搬运货架的场景组成。周围有5台经过校准的摄像机，但每个视图都会出现不同程度的遮挡，同时会出现行人的新增、完全消失、重现等多种复杂情况，有利于算法的测试。该公开数据集能较好地展现实际过程中行人的动作姿态以及可能出现的对检测、跟踪和特征关联有影响的情况。

本文所采用的评价指标包括：中心位置误差C、跟踪精确度、跟踪成功率、错跟数、误跟数等。C为跟踪目标中心到真实目标中心之间的欧氏距离，以像素为单位，(x,y)表示跟踪目标中心位置，(x_0,y_0)表示真实目标中心位置。此时C可表述为

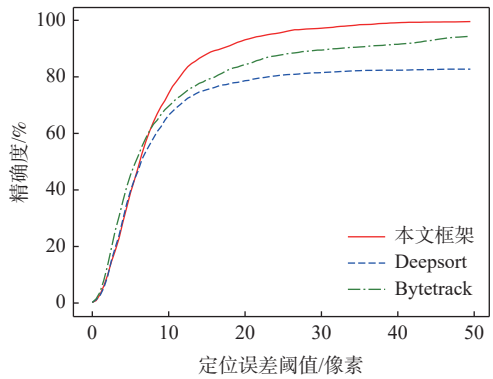
$$C=\sqrt{(x-x_0)^2+(y-y_0)^2} \tag{19}$$

由于本文所研究的多目视觉下的多目标行人在视频帧中是连续的且尺寸相差不大，因此目标的跟踪精确度可以表示为在给定范围内跟踪预测框与真实目标框的C。本文采用Wu等[33]提出的成功率评估方法，通过计算跟踪预测框与真实目标框间的交并比I来判断两框的重合程度，进而判定跟踪成功率。错跟表示在一定的交并比阈值下，同一目标ID匹配时，跟踪预测框不满足阈值以及跟踪预测框满足阈值时，目标ID不匹配等情况。误跟表示在一定的交并比阈值下，真实目标ID与跟踪预测ID匹配不上的所有情况，包括跟踪预测框不满足阈值、跟踪预测框满足阈值但预测ID与真实ID不符、真实目标框在阈值范围内无跟踪预测框且图像内无同ID的其他预测框等。在多目视觉下，错跟数和误跟数分别表示跟踪算法在所有视角下的错跟总数和误跟总数。

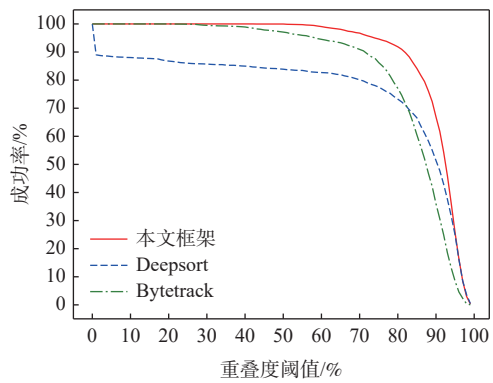
本文将所提出的跟踪框架与相关工作提到的优秀跟踪算法进行比较。由于本文主要解决的是在多视角下存在遮挡的密集人群跟踪问题以及行人消失后重现等问题，同时聚焦于多目视觉下目标出现在有效区域后的多视角关联匹配问题。而本文所对比的主要是基于单目视觉的多目标跟踪算法，无法同时对不同视角下的同一目标进行关联匹配。因而在实验过程中，在本文的算法框架中进行多目视觉下的多目标跟踪，在对比算法中分别对数据集的各个视角进行单独跟踪后整合，最后利用C和I等评价指标对比多视角下的综合跟踪结果。

3.2 实 验

将本文跟踪框架与经典算法Deepsort[10]以及最新的多目标跟踪算法Bytetrack[12]进行对比。为了充分分析算法性能，在图11(a)中，将本文框架与2种跟踪算法进行目标跟踪偏移距离精度分析，横坐标表示跟踪中心位置偏移阈值，单位为像素，纵坐标表示跟踪精度。这里只评估C小于50的，当C大于50时认为未跟踪到目标。此时曲线表示检测到的目标偏移距离小于不同偏移阈值的数量占总帧数的比例。经计算，定位误差阈值在20～50像素时，本文算法平均精确度为97.73%，相比Deepsort、Bytetrack算法分别提升16.05%、7.39%。



(a) 不同算法跟踪精确度对比



(b) 不同算法跟踪成功率对比

图 11　不同算法下多视角跟踪结果对比

Fig. 11　Comparison of multi-view tracking results under different algorithms

$$S' = \frac{|R_i \cap R_0|}{|R_i \cup R_0|} \tag{20}$$

式中: R_i为跟踪预测框面积; R_0为真实目标框面积。

在图 11(b) 中, 对比了不同算法下的跟踪成功率, 横坐标表示跟踪预测矩形框与真实矩形框的交并比阈值, 如式(20)所示。纵坐标表示跟踪成功率。曲线表示所有检测目标的交并比大于不同阈值的数量占总帧数的比例, 实线代表本文算法。经计算, 在重叠度阈值为 30%～70% 时, 本文算法平均跟踪成功率为 98.37%, 相比其他算法分别提升 16.04%、4.16%。从图 11 中可以看出, 本文提出的跟踪框架在多视角下存在行人新增、长时间遮挡、消失后重现等复杂情况下不仅跟踪的精度高且跟踪目标的成功率也优于其他跟踪算法。

图 11 中的 2 种评价指标主要用于单目视觉下的多目标检测, 精确度和成功率的阈值比对主要是基于跟踪预测框与真实目标框能够在视频流中准确对应的情况下。而在多目视觉中, 不同视角间的目标有相对位置差异较大、遮挡方式和遮挡位置不同等特点, 跟踪算法出现目标 ID 匹配错误、未出现跟踪预测框等误跟情况较为频繁。仅靠跟踪精确度和成功率不能地很好地表现跟踪算法在错跟、漏跟等方面的表现。

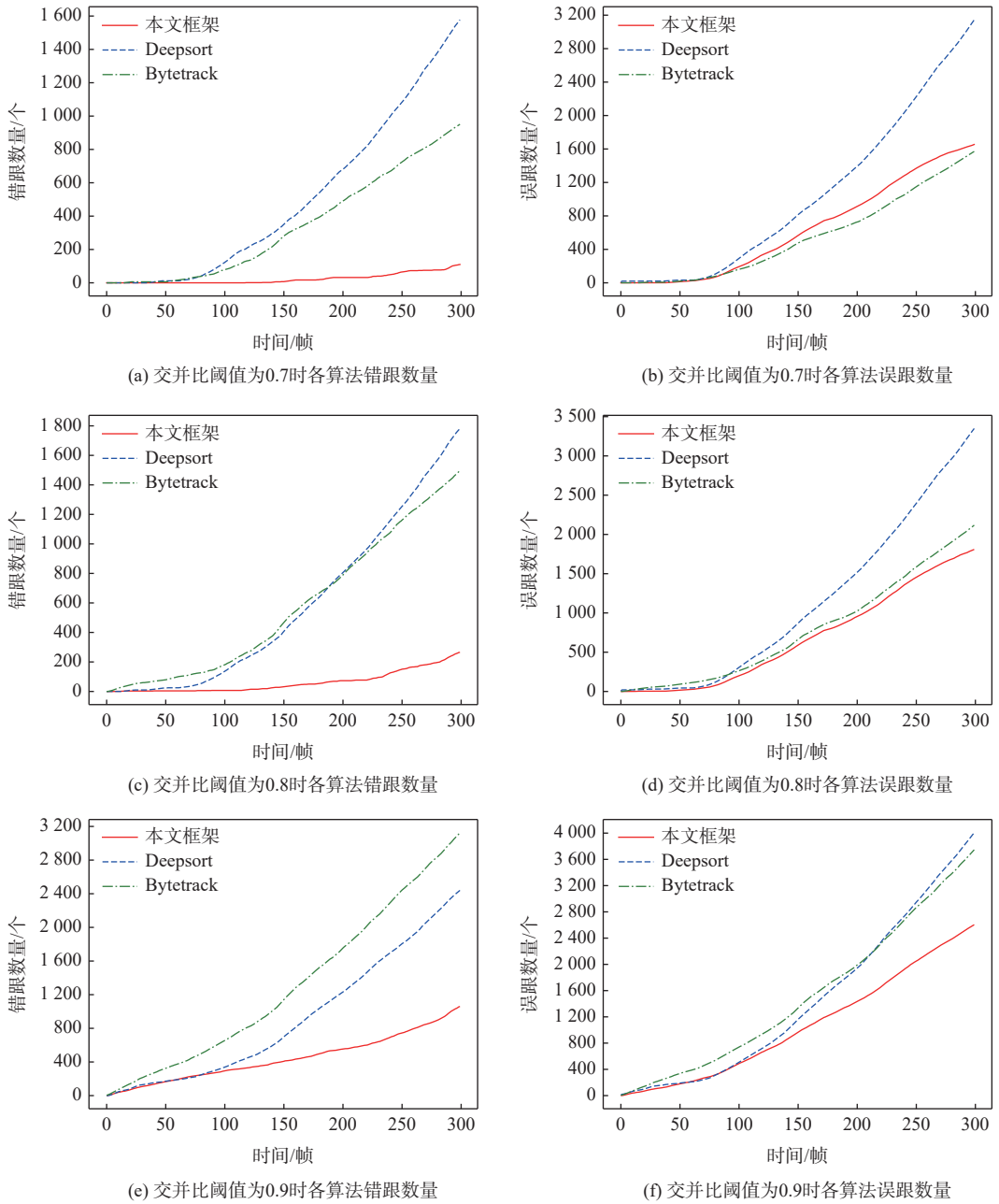
图 12(a)～图 12(f) 对比了不同重叠度阈值下各跟踪算法随着时间流所出现的错跟/误跟情况。横坐标表示视频流帧数, 纵坐标表示错跟/误跟总数量。曲线表示在多目视觉下, 到某一帧为止各算法出现的错跟/误跟数量总和, 实线表示本文算法。

从图 12 可以看出, 本文提出的跟踪框架在不同阈值下均能够展现比其他算法更优秀的跟踪稳定性及控制失误的性能, 尤其是在错跟数量的控制上展现出了非常好的效果。相比之下, 其他算法由于行人消失后重现、不同视角受不同程度遮挡等原因, 在后期会出现持续的 ID 不匹配、跟踪错误等情况。

同时, 在误跟方面, 虽然本文的算法框架由于聚焦有效区域的多目标行人跟踪而主动忽略前期的部分行人跟踪框, 但在总体的误跟数量控制上仍然相比其他算法有较为明显的优势。各算法在完整视频流不同阈值下所产生的错跟和误跟数量如表 1 和表 2 所示。由完整视频流的真实行人框总数 4 733, 可计算出本文算法平均错跟率为 10.11%, 相比其他 2 个算法分别提升了 40.84% 和 39.26%; 本文的算法平均误跟率相较其他 2 个算法也分别提升了 31.22% 和 9.63%。

在图 13 中展现了不同跟踪算法下某行人从视角中消失后重现的跟踪情况。横坐标表示时间, 纵坐标 0/1 表示目标行人是否跟踪到。实线表示本文跟踪框架。从图中可以发现, 在该目标消失并重现后, 只有本文算法能够在重现后的大部分时间重新跟踪到该目标并赋予相同 ID。相比之下, 其他算法几乎无法解决目标消失后重现的问题, 在目标消失后均丢失了其特征。

在图 14 中, 展示出了一些关键帧的跟踪结果。每一行是一帧下不同视角的图像。第 1 行是 2 号行人新增时的场景, 只有前两个视角检测到了新增目标出现, 该帧下各算法均取得了不错的跟踪效果。第 2 行是 2 号行人从所有视角中完全消失后重现的场景, 其中本文算法能够跟踪到同一目标并预测出与真实值相同的 ID, 而其他算法会出现 ID 匹配错误等情况。第 3 行是多视角下各行人均处于有效跟踪区域当中的场景。可以发现, 其他算法在目标出现遮挡、消失并重现后会出现 ID 匹配错误、跟踪漂移等误跟现象。相比之下, 本文算法能够在多目视觉下有效地跟踪并关联出所有目标并预测出正确 ID。



(a) 交并比阈值为0.7时各算法错跟数量　(b) 交并比阈值为0.7时各算法误跟数量

(c) 交并比阈值为0.8时各算法错跟数量　(d) 交并比阈值为0.8时各算法误跟数量

(e) 交并比阈值为0.9时各算法错跟数量　(f) 交并比阈值为0.9时各算法误跟数量

图 12　不同阈值下各算法随时间产生的错跟/漏跟数量

Fig. 12　The number of error-tracking or mismatches occurring over time by each algorithm under different thresholds

表 1　各算法在不同阈值下的错跟数量

Table 1　The number of error-tracking at different thresholds for each algorithm

算法	阈值		
	0.7	0.8	0.9
本文框架	110	265	1 060
Deepsort	1 577	1 782	2 440
Bytetrack	951	1 498	3 126

表 2　各算法在不同阈值下的误跟数量

Table 2　The number of mismatches at different thresholds for each algorithm

算法	阈值		
	0.7	0.8	0.9
本文框架	1 654	1 809	2 604
Deepsort	3 144	3 349	4 007
Bytetrack	1 571	2 118	3 746

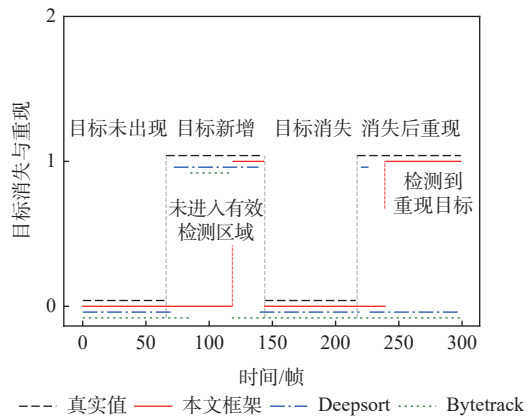


图 13　算法对目标消失后重现的跟踪性能对比

Fig. 13　Comparison of tracking performance of each algorithm after disappearance and reappearance

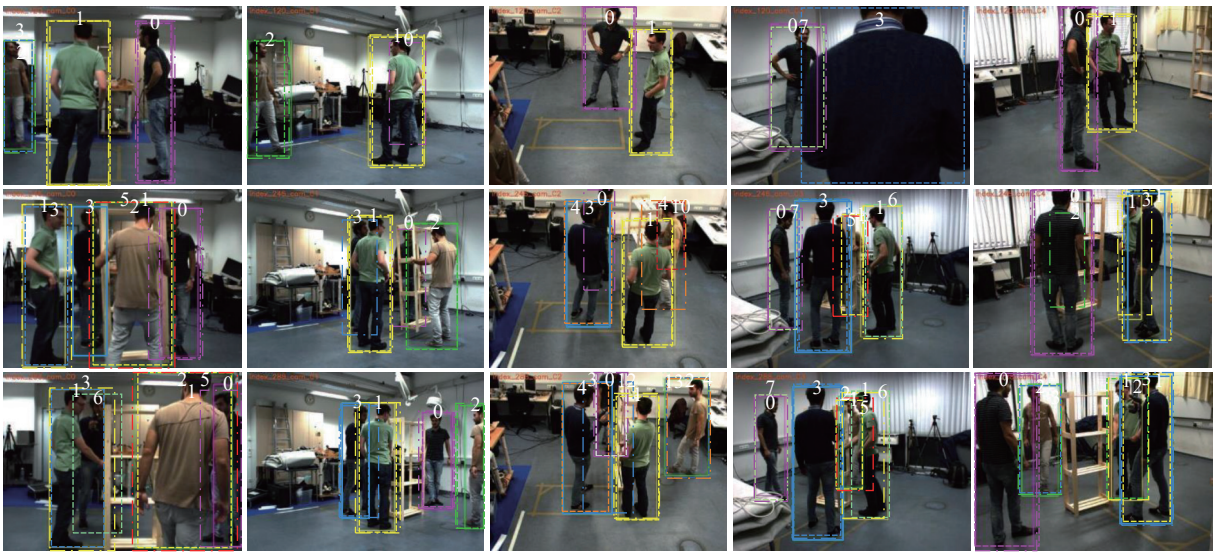


图 14　一些关键帧的跟踪结果对比

Fig. 14　Comparison of tracking results for some key frames

4　结　论

1) 本文算法可实现较为优异的跟踪精确度及跟踪成功率。如定位误差阈值为 15 时, 跟踪精确度为 88.45%; 阈值为 35 时, 跟踪精确度为 98.48%。重叠度阈值为 50% 时, 跟踪成功率为 99.97%; 阈值为 85% 时, 跟踪成功率为 84.13%。

2) 本文算法能够在多目视觉下有效地控制跟踪的错跟及误跟数量。如阈值为 0.7 时, 错跟数量为 110/4 733, 错跟率仅为 2.3%; 在不同阈值下的平均错跟率为 10.11%。相比其他算法有大幅提升。

3) 本文算法提出的利用 GMM 更新特征池并选取目标特征模板有效解决了行人受长时间遮挡/从视角中消失后重现等问题。

4) 本文算法可以有效进行行人的新增及重现判定, 结合动态阈值方法, 使算法在不同环境和背景下不需要人为调整固定参数且能产生较好的跟踪效果, 保证了其泛化能力。

5) 本文算法聚焦于提高多目视觉下有效检测区域中的多目标跟踪效果, 在少数视角上的新增行人跟踪具有局限性。

参考文献（References）

[1] ZHUANG B H, LU H C, XIAO Z Y, et al. Visual tracking via discriminative sparse similarity map[J]. IEEE Transactions on Image Processing, 2014, 23(4): 1872-1881.

[2] BERCLAZ J, FLEURET F, TÜRETKEN E, et al. Multiple object tracking using K-shortest paths optimization[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2011, 33(9): 1806-1819.

[3] WU T F, LU Y, ZHU S C. Online object tracking, learning and parsing with and-or graphs[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(12): 2465-2480.

[4] HE K M, ZHANG X Y, REN S Q, et al. Spatial pyramid pooling in deep convolutional networks for visual recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015, 37(9): 1904-1916.

[5] SIKDAR A, CHATTERJEE D, BHOWMIK A, et al. Open-set metric learning for person re-identification in the wild[C]//Proceedings of the 2020 IEEE International Conference on Image Processing. Piscataway: IEEE Press, 2020: 2356-2360.

[6] HAGER G D, BELHUMEUR P N. Efficient region tracking with parametric models of geometry and illumination[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1998, 20(10): 1025-1039.

[7] REN S Q, HE K M, GIRSHICK R, et al. Faster R-CNN: towards real-time object detection with region proposal networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(6): 1137-1149.

[8] KWON O H, TANKE J, GALL J. Recursive Bayesian filtering for multiple human pose tracking from multiple cameras[C]// Computer Vision – ACCV 2020. Berlin: Springer, 2021: 438-453.

[9] BEWLEY A, GE Z Y, OTT L, et al. Simple online and realtime tracking[C]//Proceedings of the 2016 IEEE International Conference on Image Processing. Piscataway: IEEE Press, 2016: 3464-3468.

[10] WOJKE N, BEWLEY A, PAULUS D. Simple online and realtime tracking with a deep association metric[C]//Proceedings of the 2017 IEEE International Conference on Image Processing. Piscataway: IEEE Press, 2017: 3645-3649.

[11] COMANICIU D, RAMESH V, MEER P. Real-time tracking of non-rigid objects using mean shift[C]//Proceedings of the 2000 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2000: 142-149.

[12] ZHANG Y F, SUN P Z, JIANG Y, et al. ByteTrack: multi-object

tracking byAssociating every detection box[C]//Proceedings of the Computer Vision – ECCV 2022. Berlin: Springer, 2022: 1-21.

[13] ZHU X K, LYU S C, WANG X, et al. TPH-YOLOv5: improved YOLOv5 based on transformer prediction head for object detection on drone-captured scenarios[EB/OL]. (2021-08-26)[2023-05-25]. https://doi.org/10.48550/arXiv.21081.1539.

[14] HARTIGAN J A, WONG M A. Algorithm AS 136: a K-means clustering algorithm[J]. Applied Statistics, 1979, 28(1): 100.

[15] LIU Z, LIN Y T, CAO Y, et al. Swin transformer: hierarchical vision transformer using shifted windows[C]//Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE Press, 2021: 9992-10002.

[16] IOFFE S , SZEGEDY C. Batch normalization: accelerating deep network training by reducing internal covariate shift[C]// Proceedings of the 32nd International Conference on International Conference on Machine Learning. Lille: ACM, 2015, 37: 448-456.

[17] HINTON G E, SRIVASTAVA N, KRIZHEVSKY A, et al. Improving neural networks by preventing co-adaptation of feature detectors[EB/OL]. (2012-07-03)[2023-05-28]. https://doi.org/10.48550/arXiv.1207.0580.

[18] YU E, LI Z L, HAN S D, et al. RelationTrack: relation-aware multiple object tracking with decoupled representation[J]. IEEE Transactions on Multimedia, 2022, 25: 2686-2697.

[19] LI W, XIONG Y J, YANG S, et al. Semi-TCL: semi-supervised track contrastive representation learning[EB/OL]. (2021-07-06)[2023-05-29]. https://doi.org/10.48550/arXiv.2107.02396.

[20] CHEN K, SONG X, ZHAI X, et al. An integrated deep learning framework for occluded pedestrian tracking[J]. IEEE Access, 2019, 7: 26060-26072.

[21] BAE S H, YOON K J. Robust online multiobject tracking with data association and track management[J]. IEEE Transactions on Image Processing, 2014, 23(7): 2820-2833.

[22] BRASO G, LEAL-TAIXE L. Learning a neural solver for multiple object tracking[C]//Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2020: 6246-6256.

[23] HORNAKOVA A, HENSCHEL R, ROSENHAHN B, et al. Lifted disjoint paths with application in multiple object tracking[EB/OL]. (2020-06-25)[2023-06-01]. https://doi.org/10.48550/arXiv.2006.14550.

[24] ZENG F G, DONG B, ZHANG Y A, et al. MOTR: end-to-end multiple-object tracking withTransformer[C]//Proceedings of the Computer Vision – ECCV 2022. Berlin: Springer, 2022: 659-675.

[25] DAI P, WENG R L, CHOI W, et al. Learning a proposal classifier for multiple object tracking[C]//Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2021: 2443-2452.

[26] WANG Z D, ZHENG L, LIU Y X, et al. Towards real-time multi-object tracking[C]//Proceedings of the Computer Vision – ECCV 2020. Berlin: Springer, 2020: 107-122.

[27] WANG Q, ZHENG Y, PAN P, et al. Multiple object tracking with correlation learning[C]//Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2021: 3875-3885.

[28] MEINHARDT T, KIRILLOV A, LEAL-TAIXÉ L, et al. TrackFormer: multi-object tracking with transformers[C]//Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2022: 8834-8844.

[29] STADLER D, BEYERER J. Improving multiple pedestrian tracking by track management and occlusion handling[C]//Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2021: 10953-10962.

[30] NIGAM J, RAMESHAN R M. EgoTracker: pedestrian tracking with re-identification in egocentric videos[C]//Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops. Piscataway: IEEE Press, 2017: 980-987.

[31] KHAN S M, SHAH M. A multiview approach to tracking people in crowded scenes using a planar homography constraint[C]//Proceedings of the Computer Vision – ECCV 2006. Berlin: Springer, 2006: 133-146.

[32] BELAGIANNIS V, AMIN S, ANDRILUKA M, et al. 3D pictorial structures for multiple human pose estimation[C]//Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2014: 1669-1676.

[33] WU Y, LIM J, YANG M H. Online object tracking: a benchmark[C]//Proceedings of the 2013 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2013: 2411-2418.

A dense pedestrian tracking method based on fusion features under multi-vision

HUANG Yujie，CHEN Kai*，WANG Ziyuan，WANG Ziteng

(College of Mechanical and Electrical Engineering，Nanhang University，Nanjing 210016，China)

Abstract：Many multi-object pedestrian tracking algorithms have been proposed in computer vision, and great progress has been made in tracking efficiency and accuracy recently. Practical applications are severely hampered by the fact that the majority of tracking techniques now in use are still unable to address the issues of object occlusion and reappearance in camera perspectives. To tackle the above problems in dense crowds under multi-vision, the multi-target pedestrian tracking method is based on fusion feature correlation. The feature pool was updated based on GMM to reduce feature pollution caused by dense people. To ensure the tracking universality, the similarity threshold of target features was calculated dynamically based on K-means. The similarity of fused features is used to associate the pedestrian features, with the homography constraint check to determine the addition and reappearance of pedestrians, which reduces error and miss tracking. The results of experiments using several algorithms on the public dataset Shelf indicate that the suggested method's average accuracy is 16.05% and 7.39% higher than that of other methods, while its average success rate is 16.04% and 4.16% higher. The average error tracking rate under the complete video is 10.11%, which achieves significant results in controlling mistracking and effectively associates with the original ID after the pedestrian’s reappearance.

Keywords：object tracking; feature extraction; multiple vision; dynamic threshold; fusion feature; gaussian mixture model; correlation matching

Received：2023-06-28；**Accepted**：2023-09-01；**Published Online**：2023-10-11 13：00

URL：link.cnki.net/urlid/11.2625.V.20231010.0958.001

Foundation items：National Natural Science Foundation of China (52202417); China Postdoctoral Science Foundation (2022TQ0155,2022M721605); Open Project Program of State Key Laboratory of Virtual Reality Technology and Systems, Beihang University (VRLAB2023A02)

* **Corresponding author**. E-mail：chen_kai@nuaa.edu.cn

http://bhxb.buaa.edu.cn jbuaa@buaa.edu.cn
DOI: 10.13700/j.bh.1001-5965.2023.0451

基于三元组哈希损失的半监督图像检索

邵伟志，熊思宇，潘丽丽*
(中南林业科技大学 计算机与数学学院，长沙 410004)

摘　　要：目前大多数基于深度学习的图像检索方法是在有监督条件下进行的，需要大量的标签数据，但实际应用中获取大量标签数据困难且成本高昂。此外，现有基于欧氏距离的三元组损失计算不够精确，使模型对图像相似性学习的能力欠佳。采用熵最小化伪标签、三元组损失和半监督学习技术，提出了一种新的半监督哈希图像检索模型（SSITL）。应用多阶段模型联合与锐化技术为未标记数据生成伪标签，并通过熵最小化处理以提高伪标签的置信度。同时，利用标记数据和未标记数据的聚类结果选择三元组，并采用基于通道权重矩阵的三元组哈希损失（CWT loss）帮助SSITL学习图像相似性。为了生成更好的哈希码，在2个汉明嵌入间使用MixUp进行混洗得到新的汉明嵌入以改善图像检索性能。实验结果表明：相较于其他方法，SSITL在相仿的时间开销下，在CIFAR-10和NUS-WIDE数据集上的检索平均准确率分别提高了1.2%和0.7%，强有力地验证了SSITL是一种优秀的半监督哈希图像检索模型。

关　键　词：图像检索；三元组哈希损失；半监督学习；伪标签；深度学习
中图分类号：TP391
文献标志码：A　　**文章编号**：1001-5965(2025)07-2526-12

在图像处理领域中，图像检索充当着举足轻重的作用。基于文本的图像检索方法[1]最先提出并应用于图像检索，其本质是将对图像的检索转化为对图像标注的检索。但是人工标注成本高昂并且具有主观性，文本标注无法全面描述对象，对图像内容的描述存在严重缺失。基于内容的图像检索方法[2]主要依赖对图像视觉特征的提取，在视觉特征与图像数据之间建立关联并基于此进行图像检索，但高水平语义概念之间的鸿沟依然存在并影响图像检索的准确率。

近年来，基于卷积神经网络(convolutional neural network, CNN)获取的图像特征因其具有较强判别能力、紧凑性表示和较高的搜索效率成为图像检索的主流方法[3]。深度神经网络可以进行有效的特征表示学习，并基于深度特征得到二进制哈希编码，通过对二进制哈希编码进行汉明排序可以实现高效的图像检索[4]。现有的哈希方法可以分为无监督哈希和有监督哈希两类。无监督方法可以在没有语义标签等相关信息的情况下完成检索任务，但语义差距困境，即数据的高级语义描述和低级特征描述存在差距的缺点导致无监督方法的检索精度不够理想[5-6]。

而监督方法中的相似度集合可以根据数据的标签或相关性反馈来构建，这代表监督方法获取的哈希码质量更高。但是监督方法的训练对标记数据的依赖很严重，使用有限的标记数据进行训练很容易导致模型过拟合。半监督算法可以帮助模型从标记数据的分布中学习未标记数据的分布，从而减少了深度学习对昂贵标记数据的依赖[7]。在图像检索领域，人脸识别任务中提出的三元组损失函数[8]

收稿日期：2023-07-10；**录用日期**：2023-09-14；**网络出版时间**：2023-10-30 21:35
网络出版地址：link.cnki.net/urlid/11.2625.V.20231030.1403.004
基金项目：湖南省自然科学基金面上项目(2021JJ31164)；湖南省教育厅科学研究重点项目(22A0195)
*通信作者. E-mail：lily_pan@163.com

引用格式：邵伟志，熊思宇，潘丽丽. 基于三元组哈希损失的半监督图像检索 [J]. 北京航空航天大学学报，2025，51（7）：2526-2537.
SHAO W Z，XIONG S Y，PAN L L. Semi-supervised image retrieval based on triplet hash loss [J]. Journal of Beijing University of Aeronautics and Astronautics，2025，51（7）：2526-2537（in Chinese）.

具有很强的图像相似性保持能力, 在许多检索任务中取得了很好的效果。现有的三元组损失中通常使用欧氏距离计算, 其存在 2 个问题: ①三元组的选取质量直接影响网络学习的有效性; ②欧氏距离的计算难以描绘图像之间的区分性, 三元组损失在训练网络的过程中容易忽略图像结构性、完整性的信息。

本文针对以上问题, 将三元组损失与半监督算法相结合, 提出一种基于三元组哈希损失的半监督图像检索模型(semi-supervised image retrieval based on triplet hash loss, SSITL)。在 SSITL 中, 提出一种熵最小化伪标签的构建方法, 为未标记数据生成高置信度伪标签用于模型训练提取深度特征, 然后对深度特征进行聚类并依据聚类中心选取三元组, 设计了一种基于通道权重矩阵的三元组哈希损失(channel weight triplet hash loss, CWT loss), 为图像特征生成哈希码, 降低存储压力和计算成本的同时实现高效的图像检索。

1　相关工作

1.1　三元组损失函数

Schroff 团队于 2015 年提出三元组损失函数, 构建了 FaceNet 并将三元组损失应用于人脸识别领域[8]。在三元组损失中定义了锚样本、正样本和负样本, 其中锚样本和正样本来自同一类别, 负样本来自其他类别。由三元组损失训练的网络通过拉近特征空间中锚样本与正样本之间的距离, 拉远锚样本与负样本之间的距离, 提高模型的识别率和准确率。为了进一步挖掘三元组之间的潜在联系, Song 等[9] 通过将小批量内成对距离向量提升为成对距离矩阵, 优化结构化预测目标。郑大刚等[10] 提出了一种基于三元组损失的深度哈希方法, 通过优化三元组损失函数提高人脸识别任务的精确度。杜雨佳等[11] 利用三元组网络实现跨域数据检索, 通过特征间距离度量不同模态数据的相似性。此外, 还有许多研究将三元组损失应用于图像视觉任务。刘晗煜等[12] 提出了一种基于视角一致性的三元组损失函数并应用于车辆重识别技术。Liao 等[13] 提出一种小批量图采样方法, 通过构造所有类的最近邻图用于信息采样, 降低了深度度量学习的内存和计算成本。Yang 等[14] 通过动态更新的原型来纠正三元组损失中的离群锚点, 加强了三元组损失的稳健性。然而, 在训练集中选取到高质量三元组对网络学习的有效性影响较大, 且使用欧氏距离计算的三元组损失函数难以描述图像的区分性, 导致网络的图像检索准确率降低。

1.2　半监督学习

在半监督学习中, 同时使用标记数据和未标记数据训练模型, 减少了对标记数据的依赖, 更符合实际应用场景。大多数半监督学习的关键思想是平滑性的先验假设, 即彼此接近的数据点或数据点相同的结构可能具有相同的标签。在 pseudo-label[15] 中, 通过为未标记数据添加伪标签实现半监督学习, 其中伪标签是对未标记数据的高置信度预测。MixMatch 算法[16] 通过“锐化函数”最小化熵来调整预测分布的温度, 以鼓励模型输出对未标记数据的可靠预测。近年来, 大部分半监督学习工作围绕具有一致性正则化的半监督深度学习模型, 它倾向于对输入进行稳定估计。对于相同样本的目标预测, 一致性正则化鼓励模型对于数据的预测尽量平滑, 使模型可以忽略小扰动的影响。为进一步优化未标记数据的学习, Zhang 等[17] 提出一种半监督深度哈希方法(semi-supervised deep hashing for large scale image retrieval, SSDH), 最小化标记数据的经验误差以及标记和未标记数据的嵌入误差。Guo 等[18] 提出中心乘积量化网络(central product quantization network for semi-supervised image retrieval, CPQN), 设计中心相似度策略以保持标记数据中的语义相似度和底层数据结构, 并通过一致性正则化将其推广到未标记数据中。Wang 等[19] 使用对抗学习策略将未标记数据的知识传递给判别模型, 学习更广义的语义特征。魏翔等[20] 提出一种可靠标签选择与学习的算法, 应用标签传染算法并综合考虑了样本输出均值和一致性。现有的半监督学习在不同程度上改善了对标签数据的依赖性, 但是在与图像检索算法结合的过程中, 面对大规模图像模型训练时间过长; 同时, 图像检索常用的哈希码并没有在现有的半监督学习算法中得到优化, 影响了图像检索的质量。

2　半监督图像检索模型 SSITL

本文提出基于三元组哈希损失的半监督图像检索模型 SSITL, 如图 1 所示。为了缓解由于标记数据不足带来的性能下降, SSITL 从多角度图像的预测中学习数据的互补信息, 提出熵最小化伪标签构建方法, 通过熵最小化处理生成高置信度伪标签; 基于熵最小化伪标签构建相似性集合, 利用相似性集合计算模型的哈希损失; 在 CNN 中, 引入聚类空间并根据聚类结果生成三元组, 有效利用未标记数据的信息, 并提出 CWT loss 以改善网络对于图像相似性的学习能力; 最后, 为生成高效、紧凑的图像哈希码, 使用 MixUp[21] 在 2 个汉明嵌入之间进

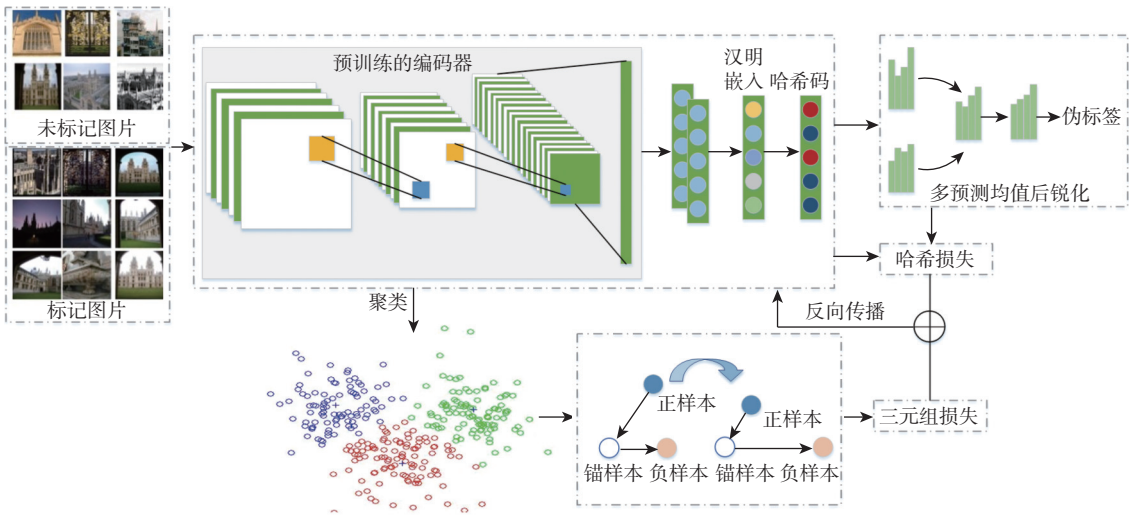


图1 基于三元组哈希损失的半监督图像检索模型

Fig. 1 A semi-supervised image retrieval model based on triplet hash loss

行混洗以产生新的汉明嵌入，对新汉明嵌入进行量化得到最终的哈希码，并基于哈希码的比对排序得出检索结果。

2.1 熵最小化伪标签

2.1.1 生成伪标签

为了有效降低图像标注成本并提高模型检索精度，本文提出熵最小化伪标签生成方法，其流程如图2所示。首先，通过数据扩充技术增加未标记图片的多样性，使模型学习图像的多角度互补信息；然后，集成多阶段模型的投票生成粗略伪标签；最后，对粗略伪标签进行预测平均并锐化以获得高置信度的低熵伪标签。

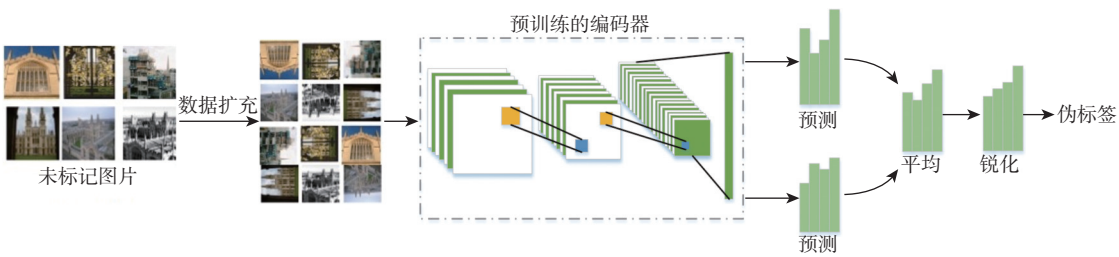


图2 伪标签生成流程

Fig. 2 Pseudo-label generation process

给定一个图像数据集I，其中包含m个标记数据$X=\{x_1,\cdots,x_m\}$和n个未标记数据$U=\{u_1,\cdots,u_n\}$，图像类别为$B=\{B_1,\cdots,B_i\}$。对于未标记图像U，首先对图像进行翻转、平移和旋转等数据扩充方法得到$\hat{U}$并对扩充后数据的预测平均，以此将对未标记数据的预测进行一次修正。

$$\begin{cases}\hat{U}=O_{\text{augment}}(U)\\ \hat{\boldsymbol{q}}=\dfrac{1}{k}\displaystyle\sum_{i=1}^{k}P_{\text{model}}(u_i;\theta)\end{cases}\tag{1}$$

式中：O_{augment}为数据扩充操作，$P_{\text{model}}(u_i;\theta)$表示模型的预测，$\theta$为网络的参数。为了从不同的角度获取信息，对于未标记数据$(u_1,\hat{\boldsymbol{q}})$，通过时间集成的思想，使用不同训练期间的网络对伪标签进行投票。

$$\boldsymbol{q}=\frac{1}{2}(\hat{\boldsymbol{q}}+P_{n-\text{model}}(u_1;\theta))\tag{2}$$

式中：$P_{n-\text{model}}(u_1;\theta)$表示经过$n$个epoch后的模型的预测；$\boldsymbol{q}$为粗略伪标签。对于投票得到的$\boldsymbol{q}$使用锐化函数$E_{\text{sharpen}}$，锐化函数有利于生成低熵伪标签$\boldsymbol{P}$，锐化函数计算如下：

$$\boldsymbol{P}=E_{\text{sharpen}}(\boldsymbol{q},T)=q_i^{\frac{1}{T}}\Big/\sum_{j=1}^{L}q_j^{\frac{1}{T}}\tag{3}$$

式中：T为超参数。随着$T\to 0$，伪标签将趋向于“One-hot”。因此，最终的伪标签会随着T的减小趋近低熵。

2.1.2 哈希损失

监督方法根据数据的标签或相关性反馈构建相似性集合$S=(s_{ij})$，存在图像对x_i和x_j，其相似性标签$s_{ij}=1$表示x_i和x_j相似，$s_{ij}=0$表示x_i和x_j不相似。由于未标记图像缺少标签信息，SSITL从多角度图像的预测中学习数据的互补信息并通过熵最小化处理生成高置信度的伪标签以帮助构建相似性集合S，并基于S产生哈希码损失函数。为了减

小构建的难度和计算资源的开销, 本文仅在每个 batch 中构建相似性集合。

$$s_{ij}=\begin{cases}1 & (I_i,I_j)\in B_i\\ 0 & (I_i,I_j)\notin B_i\end{cases} \tag{4}$$

式中: $(I_i,I_j)\in B_i$表示图像I_i和I_j属于同一类别, $(I_i,I_j)\notin B_i$表示图像I_i和I_j属于不同类别, 通过标签或伪标签是否相等来判断。现有的哈希方法通常使用全连接层作为哈希函数生成哈希码。然而, 在半监督环境中(即标记数据不足), 缺少标记数据可能导致汉明空间的嵌入无法完整表示数据集, 从而对未标记数据的泛化能力较差。因此对汉明嵌入进行混洗, 使用 tanh 函数作为全连接层f_{γ_μ}和f_{γ_ε}的激活函数。

$$\begin{cases}\boldsymbol{\mu}=\tanh(f_{\gamma_\mu}(f_{\gamma_{\text{cnn}}}(I)))\\ \boldsymbol{\varepsilon}=\tanh(f_{\gamma_\varepsilon}(f_{\gamma_{\text{cnn}}}(I)))\end{cases} \tag{5}$$

将f_{γ_μ}和f_{γ_ε}的激活值$\boldsymbol{\mu}$和$\boldsymbol{\varepsilon}$经过修改后的 MixUp 方法得到的h作为汉明嵌入:

$$\boldsymbol{h}\sim F_{\text{mixup}}(\mu,\varepsilon) \tag{6}$$

其中, F_{mixup}混洗方法计算如下:

$$F_{\text{mixup}}=\begin{cases}\lambda\sim\text{Beta}(\alpha,\alpha)\\ \lambda'=\max(\lambda,1-\lambda)\\ \boldsymbol{h}=\lambda'\boldsymbol{\mu}+(1-\lambda')\boldsymbol{\varepsilon}\end{cases} \tag{7}$$

式中: α为超参数。最后用符号函数将汉明嵌入h量化为二进制码$\boldsymbol{b}$。

$$\boldsymbol{b}=\text{sign}(\boldsymbol{h}) \tag{8}$$

式中, 通常使用 sigmoid 函数$\sigma(x)=1/(1+\text{e}^{-\alpha x})$, 但是 sigmoid 函数对于汉明距离间微小差异的区分能力较差, 影响了哈希码的质量。为了优化哈希码, 本文遵循 Wang 等[22]使用 Weibull 分布, Weibull 分布计算如下:

$$\sigma(x)=\text{e}^{-\left(\frac{2x}{b}\right)^{\gamma}} \tag{9}$$

Weibull 分布同样满足汉明距离越小, $P(1|\boldsymbol{h}_i,\boldsymbol{h}_j)$越大, 这表明图像对$P(\boldsymbol{h}_i,\boldsymbol{h}_j)$应该被分类为相似; 否则, $P(0|\boldsymbol{h}_i,\boldsymbol{h}_j)$越小表明$P(\boldsymbol{h}_i,\boldsymbol{h}_j)$应该被分类为不相似, 其中, $\boldsymbol{h}_i$和$\boldsymbol{h}_j$分别表示图像I_i和I_j的汉明嵌入。考虑到式(9)中目标函数是离散的, SSITL 将二进制哈希码替换为实值特征h的内积距离, 并添加量化损失。哈希损失计算如下:

$$L_{\text{hash}}=\min_{\theta}\sum_{s_{ij}\in S}-\text{In}P(s_{ij}|\boldsymbol{h}_i,\boldsymbol{h}_j)+\lambda_1\sum\|\boldsymbol{b}_i-\boldsymbol{h}_i\|_2^2 \tag{10}$$

2.2　基于通道权重矩阵的三元组哈希损失

2.2.1　构建三元组

现有方法中三元组损失通常使用三元组样本间的欧氏距离来计算, 这种方法会使得网络丢失图像的空间结构和底层特征信息。基于相似图像应具有相似的激活通道权重的思想, 本文提出了 CWT loss。在 CNN 中, 每个卷积层的输出都可以看作一个三维张量, 其中第三维表示该层输出的通道数。通道权重矩阵编码了 CNN 中每个激活通道的重要性。CWT loss 通过学习通道权重矩阵的相似性来改善网络对于图像相似性学习的能力。

首先, 使用调整后的聚类方法对深度特征进行聚类以选取三元组。训练开始前, 对初始的m个标签数据经过中间卷积层提取的特征进行聚类, 得到k个初始聚类中心$\boldsymbol{C}^h=\{\boldsymbol{C}_1^h,\cdots,\boldsymbol{C}_k^h\}$。根据下式计算数据$\boldsymbol{x}_i$与$\boldsymbol{C}_i$的距离。$\boldsymbol{x}_i^h$的聚类类别由$\boldsymbol{C}_i^h$决定, $\boldsymbol{C}_i^h$表示$\boldsymbol{x}_i^h$距离最近的聚类中心。

$$V_{\text{cluster}}=\begin{cases}d_{ij}=\|\boldsymbol{x}_i-\boldsymbol{C}_i\|_2^2\\ \boldsymbol{C}_i=W_{\text{argmin}_{i\in\{1,2,\cdots,k\}}d_{ij}}\end{cases} \tag{11}$$

训练开始后, 为了让未标记样本加入聚类更新且聚类结果可以随着网络迭代而优化, 在聚类中心的更新方法中加入了一个权重项, 该权重项根据一次网络迭代后 CNN 编码器提取的特征计算如下:

$$\boldsymbol{C}_i^h=\frac{1}{2}\left(\frac{1}{|\boldsymbol{C}_i|}\sum_{I\in\boldsymbol{C}_i}\boldsymbol{x}_i^h+\frac{1}{|\boldsymbol{C}_i|}\sum_{I\in\boldsymbol{C}_i}\boldsymbol{x}_i^{h(1)}\right) \tag{12}$$

式中: $\boldsymbol{x}_i^{h(1)}$表示$\boldsymbol{x}_i$在经过一次迭代后的 CNN 中得到的汉明嵌入。特征的距离采用欧氏距离计算如下:

$$d_{ij}=\left\|\boldsymbol{x}_i^h-\boldsymbol{x}_j^h\right\|_2^2 \tag{13}$$

然后, 根据聚类结果选取三元组样本对(I_1,I_2,I_3), 其中I_1为每个 batch 中标记的数据, I_2为聚类空间中与I_1同类中距离最远的数据, I_3为聚类空间与I_1异类中距离最近的数据, B_i表示相同的聚类类别。

$$V_{\text{cluster2}}=\begin{cases}I_2=W_{\text{argmax}_{(I_1,I_2)\in B_i}d(I_1,I_2)}\\ I_3=W_{\text{argmin}_{(I_1,I_3)\notin B_i}d(I_1,I_3)}\end{cases} \tag{14}$$

2.2.2　三元组哈希损失

根据聚类结果选取样本对(I_1,I_2,I_3)后, CWT loss 为每个样本生成一个通道权重矩阵以计算相似性, 损失函数计算流程如图 3 所示。

给定三元组(I_1,I_2,I_3), 令$\boldsymbol{A}_{I_1}^{(l)}\in\mathbf{R}^{c\times h\times w}$表示$I_1$在 CNN 中第$l$层的激活通道, 其中$c$表示通道数, h和w表示空间大小。将$\boldsymbol{Q}_{I_1}^{(l)}\in\mathbf{R}^{c\times h\times w}$作为$\boldsymbol{A}_{I_1}^{(l)}$二维重塑矩阵, 以便后续通道权重矩阵的计算。

$$\begin{cases}\tilde{\boldsymbol{G}}_{I_1}^{(l)}=\boldsymbol{Q}_{I_1}^{(l)}\times\boldsymbol{Q}_{I_1}^{(l)T}\\ \tilde{\boldsymbol{G}}_{I_1}^{(l)}\in\mathbf{R}^{c\times c}\end{cases} \tag{15}$$

式中: $\tilde{\boldsymbol{G}}_{I_1}^{(l)}$的第$i$行第$j$列表示对网络提取信息的通道

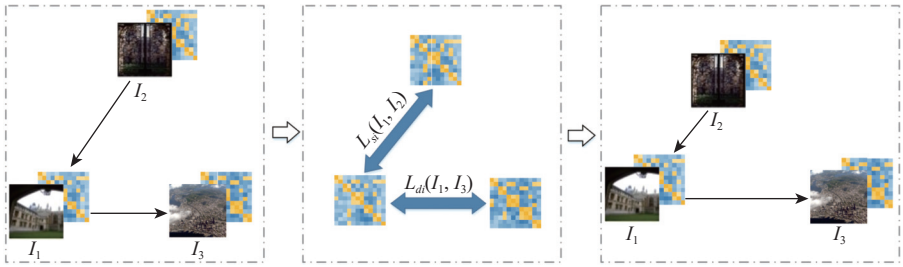


图3 损失函数计算图示

Fig. 3 Graphical representation of loss function calculation

激活模式进行了编码，对$\tilde{\boldsymbol{G}}_{I_1}^{(l)}$实施逐行归一化得到$\boldsymbol{G}_{I_1}^{(l)}$。$\boldsymbol{G}_{I_1}^{(l)}$表示为$I_1$的通道权重矩阵。通过同样的方法，可以得到$I_2$和$I_3$的通道权重矩阵$\boldsymbol{G}_{I_2}^{(l)}$和$\boldsymbol{G}_{I_3}^{(l)}$。三元组数据间的相似性度量定义为通道权重矩阵间的差值。

$$L_{si}(I_1,I_2)=\frac{1}{B_i^2}\sum_{(I_1,I_2)\in B_i}\left\|\boldsymbol{G}_{I_1}^{(l)}-\boldsymbol{G}_{I_2}^{(l)}\right\|_F^2 \tag{16}$$

式中：$(I_1,I_2)\in B_i$表示在同一批次中选取与标记数据I_1最远的同类样本I_2；$\|.\|_{\mathrm{F}}^2$为弗罗贝尼乌斯(Frobenius)范数。同理，差异损失函数计算如下：

$$L_{di}(I_1,I_3)=\frac{1}{B_i^2}\sum_{(I_1,I_3)\notin B_i}\left\|\boldsymbol{G}_{I_1}^{(l)}-\boldsymbol{G}_{I_3}^{(l)}\right\|_{\mathrm{F}}^2 \tag{17}$$

式中：$(I_1,I_3)\notin B_i$表示在同一批次中选取与标记数据I_1最近的不同类样本I_3。三元组损失L_T的计算如式(18)所示，最大化同类样本间相似度，最小化异类样本间差异度。

$$L_T=\sum_i^N\left(L_{si}(I_1,I_2)-L_{di}(I_1,I_3)\right) \tag{18}$$

融合哈希损失L_{hash}与三元组损失L_T，可以得到最终半监督损失-基于通道权重矩阵的三元组哈希损失L_{CWT}。

$$L_{\mathrm{CWT}}=L_{\mathrm{hash}}+\lambda_2 L_T \tag{19}$$

其中，λ_2作为超参数控制三元组损失在三元组哈希损失中的权重。

3 实验分析

3.1 实验设置

本文在CIFAR-10和NUS-WIDE这2个广泛使用的数据集上进行了实验。CIFAR-10[23]是一个包括60 000个32×32彩色图像的小型数据集。NUS-WIDE[24]包含从Flickr收集的近270 000张图像。在NUS-WIDE中，选取了21个常见类别，每个类别包含5 000张图像；从每个类中随机选取100张图像构建查询集，其他图像作为基本集。在训练期间，本文从数据集中随机抽取500张图像作为标记数据，其他图像作为未标记数据。

为了实验的公平，采用ImageNet上预训练的ResNet152进行特征学习。在TensorFlow框架上实现SSITL，使用Adam算法训练，学习率为0.000 1，批量大小为64，训练轮次为200。超参数通过交叉验证确定，其中$T=0.5$，$\alpha=0.75$，$\lambda_1=5$，$\lambda_2=0.5$，本文将展示不同超参数对实验结果的影响。所有实验均在服务器(16 GB内存、NVIDIA RTX2070)上进行。为了保证实验结果的准确度，采用5次实验的平均值。

3.2 评价指标及对比模型

现有的图像检索精度常用的评价指标是AP与MAP。平均准确率(average precision, AP)是衡量检索每个类别图像准确度的指标。假设模型返回N张与查询图像相关的图像，则单个类别的平均准确率Z_{AP}^i为：

$$Z_{\mathrm{AP}}^i=\left(\frac{1}{x_1}+\frac{1}{x_2}+\cdots+\frac{1}{x_N}\right)/N \tag{20}$$

平均检索精度(mean average precision, MAP)则是衡量检索所有类别图像准确度的指标，即AP的平均值。

$$Z_{\mathrm{MAP}}=\frac{1}{M}\sum_{i=1}^{M}Z_{\mathrm{AP}}^i \tag{21}$$

将SSITL与最先进的哈希方法进行比较，包括2种传统方法(SDH[25]和ITQ[26])、3种深度监督哈希方法(DPSH[27]、DRSCH[28]和DSDH[29])，以及5种深度半监督哈希方法(SSDH[17]、BGDH[30]、DSH-GAN[31]、CPQN[18]和ABML[19])。

3.3 半监督图像检索

本文在CIFAR-10数据集上进行了不同长度哈希码(12 bits、24 bits、32 bits和48 bits)的检索实验，结果如表1所示。从表1中可以看出，随着比特位的增加，监督哈希方法(DPSH、DRSCH和DSDH)的MAP值达到61%～81%，比传统方法ITQ的21%～25%高出3倍多，这表明通过CNN同时学习深度特征与哈希码的方法卓有成效。此外，半监督哈希方法的MAP值均比监督哈希方法高，这说明利用未标记数据可以提高哈希码的质量。

表 1　CIFAR-10 数据集上不同方法的平均精度

Table 1　MAP scores of different methods on CIFAR-10

方法	CIFAR-10			
	12 bits	24 bits	32 bits	48 bits
SSITL	0.818	0.838	0.853	0.854
ABML[19]	0.815	0.832	0.850	0.851
CPQN[18]	0.817	0.830	0.848	0.852
BGDH[30]	0.805	0.824	0.826	0.833
DSH-GAN[31]	0.751	0.801	0.807	0.811
SSDH[17]	0.802	0.810	0.816	0.819
DPSH[27]	0.737	0.775	0.801	0.798
DSDH[29]	0.738	0.784	0.795	0.818
DRSCH[28]	0.616	0.625	0.630	0.629
SDH[25]	0.438	0.520	0.558	0.587
ITQ[26]	0.219	0.242	0.250	0.252

从表 1 中可见, SSITL 方法的 MAP 值均高于其他方法, 这证明采用熵最小化伪标签和 CWT loss 的半监督检索模型能够生成高质量的哈希码。表 1 中不同哈希码长度的实验结果显示, 12 bits 时 SSITL 的效果略弱于 48 bits 时。这是因为 48 bits 哈希码包含的信息较低比特位哈希码更丰富。与其他方法相比, SSITL 在低比特位上的检索性能差距较小, 而随着比特位增加, 检索精度差距拉大, 优势更为明显, 这表明增加比特位有助于 SSITL 传递更多有效的信息以生成高质量的哈希码。

本文在 NUS-WIDE 数据集上进行了与 CIFAR-10 数据集上相同的实验, 结果如表 2 所示。从表 2 可知, 与其他方法相比, SSITL 方法在不同哈希码长度上均取得了最高的检索精度, 在 32 bits 时的检索 MAP 值达到了 88%。实验在半监督环境中展开, 监督哈希方法(DPSH、DRSCH 和 DSDH)的表现不如半监督方法。而与半监督哈希方法(SSDH、BGDH 和 DSH-GAN)相比, SSITL 的检索精度最佳。综合表 1 和表 2, 发现在 CIFAR-10 和 NUS-WIDE 数据集上, SSITL 在高比特位的检索优势较低比特位的优势更明显。总体而言, 在 CIFAR-10 和 NUS-WIDE 数据集上, 与其他哈希方法相比, SSITL 都表现出最好的检索精度, 48 bits 下的 MAP 值分别为 85.4% 和 87.3%, 相比较次优的 CPQN 方法分别提升了 0.2% 和 0.3%, 有效地证明了 SSITL 在图像检索中的有效性。

表 2　NUS-WIDE 数据集上不同方法的平均精度

Table 2　MAP scores of different methods on NUS-WIDE dataset

方法	NUS-WIDE			
	12 bits	24 bits	32 bits	48 bits
SSITL	0.838	0.857	0.880	0.873
ABML[19]	0.835	0.851	0.872	0.869
CPQN[18]	0.833	0.849	0.869	0.870
BGDH[30]	0.805	0.824	0.826	0.833
DSH-GAN[31]	0.828	0.843	0.848	0.851
SSDH[17]	0.803	0.808	0.826	0.833
DPSH[27]	0.767	0.778	0.795	0.798
DSDH[29]	0.772	0.804	0.821	0.831
DRSCH[28]	0.616	0.623	0.627	0.627
SDH[25]	0.541	0.548	0.579	0.621
ITQ[26]	0.663	0.700	0.707	0.723

3.4　消融实验

为了进一步验证和分析 SSITL 检索的有效性, 本文进行了消融实验, 将 SSITL 与其 2 种变体进行了比较。第 1 种变体记为 Nmid SSITL, 采用结合熵最小化伪标签和哈希损失 L_{hash} 的模型。第 2 种变体记为 Nsim SSITL, 采用结合聚类和基于通道权重的三元组哈希损失 L_{CWT} 的模型。由于图像数据仍含有少量标签信息, 因此在第 2 个变体中, 网络将调整损失函数中的比重系数 λ_2 使得网络主要通过三元组损失来学习, 实验结果如表 3 所示。

表 3　CIFAR-10 和 NUS-WIDE 数据集上 SSITL 不同模型的平均精度分数

Table 3　MAP scores of different SSITLs on CIFAR-10 and NUS-WIDE datasets

方法	CIFAR-10			
	12 bits	24 bits	32 bits	48 bits
SSITL	0.818	0.838	0.853	0.854
Nmid SSITL	0.781	0.812	0.828	0.831
Nsim SSITL	0.737	0.764	0.785	0.778
方法	NUS-WIDE			
	12 bits	24 bits	32 bits	48 bits
SSITL	0.838	0.857	0.880	0.873
Nmid SSITL	0.807	0.823	0.844	0.857
Nsim SSITL	0.775	0.788	0.815	0.820

从表 3 中可以观察到几种现象: ①在 CIFAR-10 中, SSITL 在不同比特位时比 Nmid SSITL 分别高出 3.7%、2.6%、2.5% 和 2.3%, 这表明本文提出的 L_{CWT} 有助于网络更好地学习图像相似性, 增强了网络对图像的区分能力, 从而生成更好的哈希码。②相比 Nsim SSITL, SSITL 在不同比特位时的优势分别为 8.1%、7.4%、6.8% 和 7.6%, 这表明熵最小化伪标签构建方法成功帮助网络从未标记数据中进行学习, 并能生成有效的哈希码。③Nmid SSITL 与 Nsim SSITL 相比, 在不同比特位时的优势分别为 4.4%、4.8%、4.3% 和 5.3%。这表明在 SSITL 中, 熵最小化伪标签构建方法的重要性较 L_{CWT} 更加显

著，可以有效地帮助网络在半监督环境下生成高质量哈希码，并为三元组哈希损失对网络的学习与优化奠定了基础。

在表3中，NUS-WIDE数据集上，呈现了与CIFAR-10上类似的结果。相比Nmid SSITL和Nsim SSITL，SSITL在不同比特位下表现更佳，且Nmid SSITL比Nsim SSITL表现更好。这再次证明了熵最小化伪标签方法与基于通道权重的三元组哈希损失的有效性，同时也表明所提出的伪标签构建与相似性学习是兼容的，能共同提升半监督环境中哈希码的生成质量。

从表1～表3中可以看出，二进制码的长度对检索精度影响很大。为了验证SSITL是否改善了哈希码，本文将其与变体方法进行了比较，在CIFAR-10和NUS-WIDE数据集上使用不同数量的标记和未标记数据，哈希码长度设为48 bits，实验结果如图4所示，从图中可以观察到3种现象。

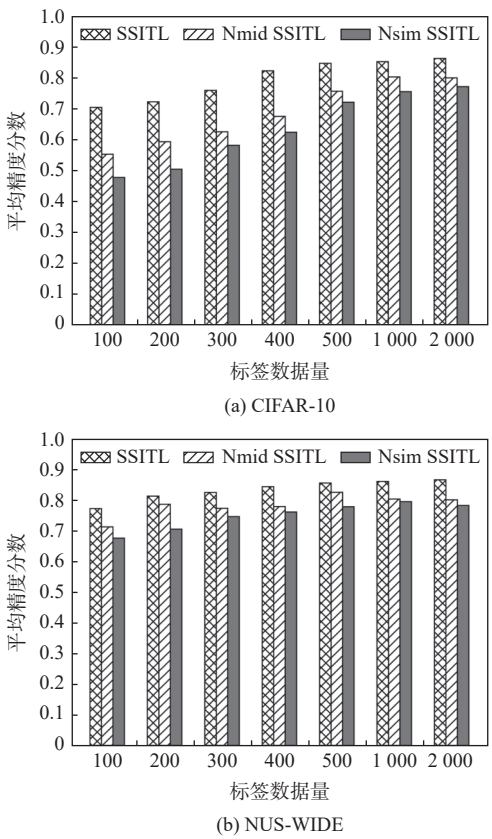


图4 不同数据集上不同标签数据量下的平均精度值

Fig. 4 MAP scores for different amounts of labeled data on different datasets

1）随着标记数据数量的增加，SSITL及其变体方法的检索精度稳步提升，标签信息的增多能有效改善模型的检索性能。在CIFAR-10数据集中，标记数据数量从100增长到2 000时，SSITL的检索性能从71.4%提高到85.8%，改善幅度达到14.4%；Nmid SSITL和Nsim SSITL的改善幅度分别达到了25.2%和29.8%。这表明随着标签数据的增多，Nmid SSITL和Nsim SSITL的收益更大。SSITL能够利用伪标签和聚类方法学习未标记数据的图像相似性，因此改进相对较少。而与Nsim SSITL相比，Nsim SSITL的改进效果更显著，这是因为Nmid SSITL采用了熵最小化伪标签构建方法，对未标记数据的学习更有效。

2）SSITL稳定地优于其变体方法，表明本文提出的融合熵最小化伪标签和基于通道权重的三元组哈希损失的图像检索模型有效地优化了网络的学习能力，提升了哈希码的生成质量。在CIFAR-10和NUS-WIDE数据集上，标记数据量从100增长到2 000时，SSITL的MAP分数的改善幅度最低，分别为6%～15.2%和6.1%～6.4%。

3）SSITL与其变体方法的检索精度差距随未标记数据的增多（即标签数据的减少）而增大，表明未标记数据越多，SSITL的检索优势越显著。

这3种现象证明SSITL成功地利用了未标记数据来改进网络的学习能力，提升了哈希码的生成质量，从而提高了图像检索精度。

3.5 不同长度哈希码实验

为了进一步评估哈希码长度对SSITL检索效果的影响，本文分别在CIFAR-10和NUS-WIDE数据集上使用（8,16,32,64,128,256）长度哈希码对不同的方法进行检索实验，如图5所示。

在CIFAR-10数据集上的实验结果如图5(a)所示。对于非深度哈希方法SDH，随着哈希码长度的增加，检索效果不断提升。同样，深度哈希方法SSDH、DPSH和DSH-GAN的检索精度也随着哈希码长度的增加而提高，这表明增加哈希码长度能改善检索精度。然而，从128 bits开始，哈希码长度的增加带来的检索精度提升趋势缓慢。这是因为其他深度哈希方法未能有效利用半监督环境中的信息，即便哈希码长度增加，也难以准确区分图像。此外，其余深度哈希方法对于图像间汉明距离辨别力不足，网络无法学习到有效的汉明嵌入，因此增加哈希码长度无法显著提高性能。相比之下，SSITL的检索准确率随着哈希码长度的增加而稳步提升，且始终优于其他方法。这是因为SSITL能够构建有效的伪标签，充分利用未标记数据和三元组哈希损失进行哈希码学习，并通过采样后混洗的汉明嵌入增强了网络对于汉明距离的区分能力。最后，特征相似性的学习也辅助提升了生成哈希码的质量。

随后，在NUS-WIDE数据集上进行了相同的实验，如图5(b)所示。由图5(b)可以发现，非深度哈希方法SDH的检索精度随哈希码长度的增加而提升；哈希码长度达到256 bits时，深度半监督哈希方

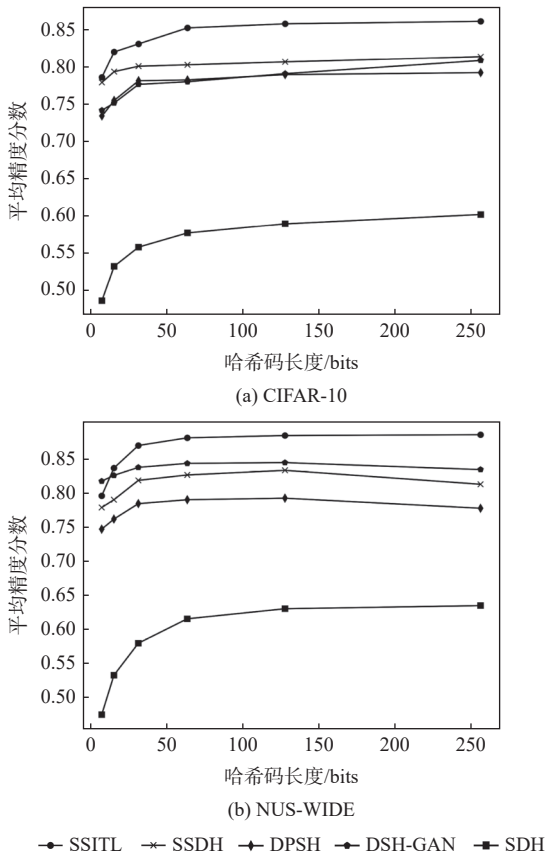


图 5　不同数据集上不同哈希码长度下的平均精度值

Fig. 5　MAP scores with different hash code lengths on different datasets

法 SSDH 和深度监督哈希方法 DPSH 的检索精度反而出现下降现象。此外, 在 CIFAR-10 与 NUS-WIDE 数据集上, 当哈希码长度小于 48 bits 时, 所有方法的检索性能都较低。总的来说, 图 5 中的实验结果表明在哈希码长度大于 32 bits 时, SSITL 的性能优于其他方法, 并且随着哈希码长度的增加, SSITL 的检索优势更加明显。

3.6　不可见类的检索

除了基本的评估实验, 本文还针对不可见类进行了检索实验以进一步探索 SSITL 的性能。在不可见类的实验中, 仅使用数据集中 75% 的类别的数据来训练网络, 其余 25% 的类别的数据用于网络的评估。具体地说, 在不可见类的实验中数据集被分为 4 个部分: train1/test1 和 train2/test2。其中, set1=train1+test2 包含 75% 的类别的数据, set2=train2+test1 包含 25% 的类别的数据, 且训练集与测试集中的数据量大小相同。

本文使用 train1 作为训练集, test2 作为查询集, train1 与 test1 作为库集。详细的实验设置遵循 SSDH[19]。

需要指出, CIFAR-10 数据集的 set1 由 7 个类组成, 而 NUS-WIDE 数据集包含 15 个类。因此, NUS-WIDE 数据集中的检索精度通过 5 个不同随机分割的平均值计算, MAP 分数基于所有返回的图像计算。使用预训练后的 ResNet 的最后一个全连接层提取的特征作为非深度哈希方法的输入。

表 4 展示了不同方法在 CIFAR-10 数据集上对不可见类的检索结果。从表 4 中可以看出, 相较于标准实验, 各种方法在不可见类的检索精度整体下降。这是由于在网络训练期间, 无法有效获取不可见类的信息。而 SSITL 在检索不可见类时仍表现出色, 在不同比特位设置下, 其 MAP 均取得最高分, 比次优的 ABML 分别高出 0.3%、0.2%、1.2% 和 1.0%。这表明, 结合熵最小化伪标签与 CWT loss 的 SSITL 有效改善了图像相似性学习能力, 提升了生成的哈希码质量。因此, 相比其他方法, SSITL 对不可见类检索的泛化性更好, 鲁棒性更强。

在 NUS-WIDE 数据集上的不可见类检索实验结果见表 5, 表 5 与表 4 呈现的结果类似。在 NUS-

表 4　CIFAR-10 数据集上不同方法对不可见类的检索精度

Table 4　MAP scores of unseen class retrieval with different methods with on CIFAR-10 datasets

方法	CIFAR-10			
	12 bits	24 bits	32 bits	48 bits
SSITL	0.319	0.327	0.349	0.358
ABML[19]	0.316	0.325	0.337	0.348
BGDH[30]	0.267	0.279	0.284	0.294
DSH-GAN[31]	0.281	0.288	0.299	0.310
SSDH[17]	0.287	0.291	0.309	0.317
DPSH[27]	0.264	0.275	0.281	0.293
DSDH[29]	0.255	0.263	0.278	0.288
DRSCH[28]	0.217	0.218	0.233	0.251
SDH[25]	0.190	0.192	0.197	0.206
ITQ[26]	0.153	0.162	0.192	0.199

表 5　NUS-WIDE 数据集上不同方法对不可见类检索的平均精度

Table 5　MAP scores of unseen class retrieval with different methods on NUS-WIDE datasets

方法	NUS-WIDE			
	12 bits	24 bits	32 bits	48 bits
SSITL	0.537	0.553	0.581	0.584
ABML[19]	0.532	0.550	0.574	0.582
BGDH[30]	0.511	0.529	0.545	0.538
DSH-GAN[31]	0.508	0.539	0.542	0.541
SSDH[17]	0.514	0.534	0.538	0.549
DPSH[27]	0.487	0.512	0.514	0.527
DSDH[29]	0.255	0.263	0.278	0.288
DRSCH[28]	0.458	0.463	0.471	0.468
SDH[25]	0.468	0.489	0.491	0.505
ITQ[26]	0.490	0.486	0.493	0.507

WIDE数据集上，SSITL进行不可见类的检索精度最高，比次优方法ABML分别高出0.5%、0.3%、0.7%和0.2%。表4和表5的结果表明，SSITL在不可见类检索方面的优势是泛化的。

为了直观地呈现SSITL与其他方法的检索结果，本文进行了统计分析，结果如图6和图7所示。图6展示了在CIFAR-10和NUS-WIDE数据集上对可见类进行检索的实验结果，所有结果均是取五次实验的均值，并通过误差线的形式描绘了多次实验的偏差。通过图6可以发现，SSITL不仅在精度上优于其他方法，偏差值也是最小的。类似地，图7展示了在CIFAR-10和NUS-WIDE数据集上对不可见类进行检索的实验结果。图7的呈现方式与图6类似，但由于是对不可见类的检索，各种方法的检索精度都有所降低。由图7可知，SSITL相比其他方法在不可见类检索的精度也更高。综合图6和图7，再次验证了SSITL的检索精度的稳定性。

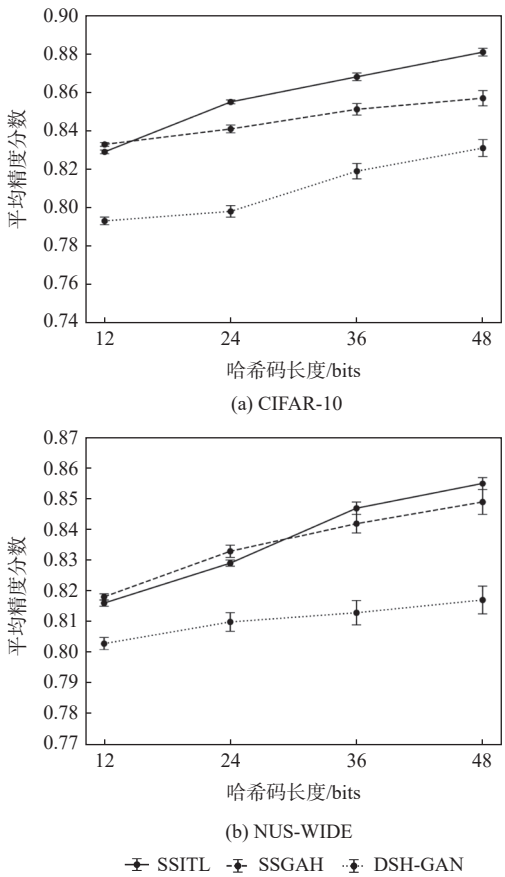


图6 不同数据集上的可见类检索

Fig. 6 Seen class retrieval on different datasets

3.7 不同层激活通道的对比实验

为了直观地分析不同激活通道层对检索的影响，本文选取了ResNet152的4个通道层：第5层、第30层、第100层和第150层的激活通道来计算三元组哈希损失。同时，为了观察不同比特位哈希

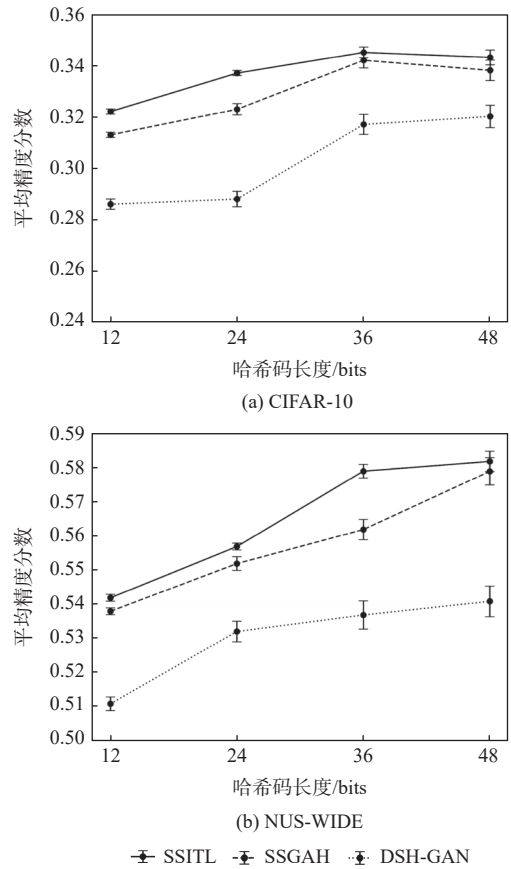


图7 不同数据集上的不可见类检索

Fig. 7 Unseen class retrieval on different datasets

码长度(12 bits、24 bits、32 bits和48 bits)以及在不同数据集上的结果，本文在CIFAR-10和NUS-WIDE数据集上进行了对比实验，结果如表6和表7所示。

表6 CIFAR-10数据集上使用不同层激活通道的不可见类检索

Table 6 Unseen class retrieval using activation channels of different layers on CIFAR-10 datasets

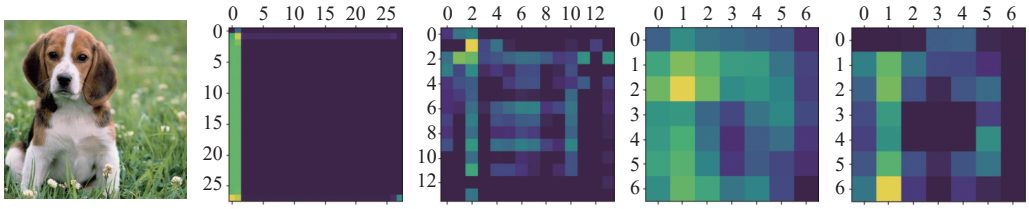
不同层激活通道	CIFAR-10			
	12 bits	24 bits	32 bits	48 bits
第5层	0.226	0.233	0.251	0.273
第30层	0.286	0.305	0.329	0.343
第100层	0.319	0.327	0.349	0.358
第150层	0.308	0.319	0.346	0.354

表7 NUS-WIDE数据集上使用不同层激活通道的不可见类检索

Table 7 Unseen class retrieval using activation channels of different layers on NUS-WIDE datasets

不同层激活通道	NUS-WIDE			
	12 bits	24 bits	32 bits	48 bits
第5层	0.462	0.473	0.481	0.483
第30层	0.504	0.511	0.539	0.542
第100层	0.537	0.553	0.581	0.584
第150层	0.518	0.532	0.574	0.579

从表 6 和表 7 中可以看出, 在 CIFAR-10 与 NUS-WIDE 这 2 个数据集中, SSITL 选取第 100 层作为激活通道的检索效果均是最好的。不同层的激活值与原图的可视化如图 8 所示。对于较浅的第 5 层和第 30 层, 由于神经网络在浅层时提取的图像信息相对不完整, 基于其激活通道计算的损失无法有效地区分同类与异类, 因此导致检索精度较低。而观察第 150 层的检索效果可以看到, 当网络过深时, 检索效果反而会下降。这是因为深层网络得到的激活通道值虽然包含更多信息, 但经过采样层与 ReLU 层的压缩, 无法有效地在损失计算中得到体现, 检索精度反而下降。



(a) 原图　(b) 第 5 层　(c) 第 30 层　(d) 第 100 层　(e) 第 150 层

图 8　原图与不同层激活值的可视化对比

Fig. 8　Visualization of the original image and activation values of different layers

3.8　不同权重的三元组损失的对比实验

为了进一步分析三元组损失 L_T 对于实验结果的影响, 本文进行了不同权重 λ_2 下, 使用 12 bits、24 bits、32 bits 和 48 bits 不同哈希码长度的实验, 实验结果如表 8 和表 9 所示。

表 8　CIFAR-10 数据集上不同权重的不可见类检索

Table 8　Unseen class retrieval using different weight values on CIFAR-10 datasets

权重值	CIFAR-10			
	12 bits	24 bits	32 bits	48 bits
$\lambda_2=0$	0.275	0.286	0.305	0.311
$\lambda_2=0.2$	0.302	0.311	0.332	0.343
$\lambda_2=0.5$	0.319	0.327	0.349	0.358
$\lambda_2=1$	0.284	0.306	0.329	0.338
$\lambda_2=2$	0.265	0.281	0.297	0.301

表 9　NUS-WIDE 数据集上不同权重值的不可见类检索

Table 9　Unseen class retrieval using different weight values on NUS-WIDE datasets

权重值	NUS-WIDE			
	12 bits	24 bits	32 bits	48 bits
$\lambda_2=0$	0.508	0.517	0.526	0.539
$\lambda_2=0.2$	0.519	0.542	0.559	0.571
$\lambda_2=0.5$	0.537	0.553	0.581	0.584
$\lambda_2=1$	0.513	0.537	0.541	0.564
$\lambda_2=2$	0.497	0.511	0.530	0.543

从表 8 和表 9 中可以看到, 当权重为 0 时, SSITL 的检索效果最差, 这是可以预见的。因为三元组损失能有效地改善网络对于图像的区分能力从而提升检索效果。当权重提升至 0.2 时, SSITL 的检索精度明显提高, 但不及权重为 0.5 时的表现, 这是因为三元组损失需要在训练过程中占据一定的分量才能提升模型的检索能力。最后, 权重为 1 和为 2 时的检索结果表明, 权重并非越高越好。这是因为在 SSITL 中, CWT loss 依赖聚类信息, 随着权重值 λ_2 的增加, 哈希损失在三元组哈希损失中的影响程度下降, 降低了生成的哈希码质量。

3.9　检索时间

为了综合评估 SSITL 的检索性能, 本文进行了 SSITL 与其他方法的检索时间对比实验。所有的实验都在同一台 PC(NVIDIA Titan 2080ti GPU, Intel Core i9-9900k 3.6-5.0 GHz CPU 和 64-GB 内存) 上进行, 不同方法的检索时间如表 10 所示。从表 10 可以看出, 所有方法的检索时间开销几乎相近, 在 CIFAR-10 和 NUS-WIDE 数据集上所有方法的时间开销平均值分别为 3.76 ms 和 12.07 ms; SSITL 在 CIFAR-10 和 NUS-WIDE 上的时间开销分别为 3.58 ms 和 12.04 ms, 均低于平均时间开销。综合考虑检索精度与检索时间的实验结果表明, SSITL 在与对比方法相近的时间开销下有效地提升了检索精度。这说明, 使用熵最小化伪标签和三元组哈希损失的 SSITL 虽然在模型训练阶段增加了一定难度, 但在测试阶段, SSITL 具有更好的学习图像相似性的能力, 从而在检索中发挥优势。

表 10　不同方法的检索时间对比

Table 10　Retrieval time comparison of different methods　ms

方法	CIFAR-10	NUS-WIDE
SSITL	3.58	12.04
SSGAH[32]	3.51	12.01
SSDH[17]	3.44	11.63
DSH-GAN[31]	3.45	12.07
DRSCH[28]	4.49	12.64
NINH[3]	4.01	12.04
CNNH[4]	3.83	12.07

4 结 论

1）本文提出一种基于三元组哈希损失函数的半监督图像检索模型SSITL，通过熵最小化伪标签和CWT loss有效减少了对标记数据的依赖以及改善了图像相似性学习能力。实验结果表明，在相仿时间开销下，SSITL在CIFAR-10和NUS-WIDE数据集上的平均检索准确率分别达到85.4%和88.0%。

2）不可见类检索的实验进一步证明SSITL具有优越的泛化性和鲁棒性。本文提出的融合熵最小化伪标签和CWT loss的半监督检索模型能提升半监督环境中哈希码的质量，从而提高了图像检索效果。

在未来的工作中将探索更高效的哈希算法，在得到精简哈希码的同时，尽可能减少特征信息损失并降低算法的复杂度。

参考文献（References）

[1] LI W, DUAN L X, XU D, et al. Text-based image retrieval using progressive multi-instance learning[C]//Proceedings of the 2011 International Conference on Computer Vision. Piscataway: IEEE Press, 2011: 2049-2055.

[2] LIU Y, ZHANG D S, LU G J, et al. A survey of content-based image retrieval with high-level semantics[J]. Pattern Recognition, 2007, 40(1): 262-282.

[3] CHEN R Y, PAN L L, LI C, et al. An improved deep fusion CNN for image recognition[J]. Computers, Materials & Continua, 2020, 65(2): 1691-1706.

[4] LAI H J, PAN Y, YE L, et al. Simultaneous feature learning and hash coding with deep neural networks[C]//Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2015: 3270-3278.

[5] CHEN Y B, MANCINI M, ZHU X T, et al. Semi-supervised and unsupervised deep visual learning: a survey[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2024, 46(3): 1327-1347.

[6] 刘颖, 程美, 王富平, 等. 深度哈希图像检索方法综述[J]. 中国图象图形学报, 2020, 25(7): 1296-1317.
LIU Y, CHENG M, WANG F P, et al. Deep Hashing image retrieval methods[J]. Journal of Image and Graphics, 2020, 25(7): 1296-1317(in Chinese).

[7] ZHU X, GOLDBERG A B. Introduction to semi-supervised learning[J]. Synthesis Lectures on Artificial Intelligence and Machine Learning, 2009, 3(1): 1-130.

[8] SCHROFF F, KALENICHENKO D, PHILBIN J. FaceNet: a unified embedding for face recognition and clustering[C]//Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2015: 815-823.

[9] SONG H O, XIANG Y, JEGELKA S, et al. Deep metric learning via lifted structured feature embedding[C]//Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2016: 4004-4012.

[10] 郑大刚, 刘光杰, 茅耀斌, 等. 基于三元组损失函数的深度人脸哈希方法[J]. 太赫兹科学与电子信息学报, 2021, 19(2): 313-318.
ZHENG D G, LIU G J, MAO Y B, et al. Deep face Hashing based on ternary-group loss function[J]. Journal of Terahertz Science and Electronic Information Technology, 2021, 19(2): 313-318(in Chinese).

[11] 杜雨佳, 李海生, 姚春莲, 等. 基于三元组网络的单图三维模型检索[J]. 北京航空航天大学学报, 2020, 46(9): 1691-1700.
DU Y J, LI H S, YAO C L, et al. Monocular image based 3D model retrieval using triplet network[J]. Journal of Beijing University of Aeronautics and Astronautics, 2020, 46(9): 1691-1700(in Chinese).

[12] 刘晗煜, 黄宏恩, 郑世宝. 基于视角一致性三元组损失的车辆重识别技术[J]. 测控技术, 2021, 40(8): 47-53,63.
LIU H Y, HUANG H E, ZHENG S B. View consistency triplet loss for vehicle re-identification[J]. Measurement & Control Technology, 2021, 40(8): 47-53,63 (in Chinese).

[13] LIAO S C, SHAO L. Graph sampling based deep metric learning for generalizable person re-identification[C]//Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2022: 7349-7358.

[14] YANG S, ZHANG Y F, ZHAO Q H, et al. Prototype-based support example miner and triplet loss for deep metric learning[J]. Electronics, 2023, 12(15): 3315.

[15] LI Z, KO B, CHOI H J. Naive semi-supervised deep learning using pseudo-label[J]. Peer-to-Peer Networking and Applications, 2019, 12(5): 1358-1368.

[16] BERTHELOT D, CARLINI N, GOODFELLOW I, et al. MixMatch: a holistic approach to semi-supervised learning[EB/OL]. (2019-10-23)[2023-05-23]. https://doi.org/10.48550/arXiv.1905.02249.

[17] ZHANG J, PENG Y X. SSDH: Semi-supervised deep hashing for large scale image retrieval[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2017, 29(1): 212-225.

[18] GUO Z T, HONG C Q, ZHUANG W W, et al. CPQN: central product quantization network for semi-supervised image retrieval [C]//Proceedings of the 2021 IEEE International Conference on Big Data. Piscataway: IEEE Press, 2021: 3183-3190.

[19] WANG G A, HU Q H, YANG Y, et al. Adversarial binary mutual learning for semi-supervised deep hashing[J]. IEEE Transactions on Neural Networks and Learning Systems, 2022, 33(8): 4110-4124.

[20] 魏翔, 王靖杰, 张顺利, 等. ReLSL: 基于可靠标签选择与学习的半监督学习算法[J]. 计算机学报, 2022, 45(6): 1147-1160.
WEI X, WANG J J, ZHANG S L, et al. ReLSL: reliable label selection and learning based algorithm for semi-supervised learning[J]. Chinese Journal of Computers, 2022, 45(6): 1147-1160(in Chinese).

[21] ZHANG H Y, CISSE M, DAUPHIN Y N, et al. Mixup: beyond empirical risk minimization[EB/OL]. (2018-04-27)[2023-05-25]. https://doi.org/10.48550/arXiv.1710.09412.

[22] WANG G A, HU Q H, YANG Y, et al. Adversarial binary mutual learning for semi-supervised deep hashing[J]. IEEE Transactions on Neural Networks and Learning Systems, 2021, 33(8): 4110-4124.

[23] KRIZHEVSKY A, HINTON G. Convolutional deep belief networks on cifar-10[J]. Unpublished Manuscript, 2010, 40(7): 1-9.

[24] CHUA T S, TANG J H, HONG R C, et al. NUS-WIDE: a real-world web image database from National University of Singapore [C]// Proceedings of the ACM International Conference on Image and Video Retrieval. New York: ACM, 2009: 1-9.

[25] SHEN F M, SHEN C H, LIU W, et al. Supervised discrete hashing[C]//Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2015: 37-45.

[26] GONG Y C, LAZEBNIK S, GORDO A, et al. Iterative quantization: a Procrustean approach to learning binary codes for large-scale image retrieval[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2012, 35(12): 2916-2929.

[27] LI W J, WANG S, KANG W C. Feature learning based deep supervised hashing with pairwise labels[EB/OL]. (2016-04-21)[2023-05-27]. https://doi.org/10.48550/arXiv.1511.03855.

[28] ZHANG R M, LIN L, ZHANG R, et al. Bit-scalable deep hashing with regularized similarity learning for image retrieval and person re-identification[J]. IEEE Transactions on Image Processing, 2015, 24(12): 4766-4779.

[29] LI Q, SUN Z, HE R, et al. Deep supervised discrete hashing[C]// Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach: Curran Associates, 2017: 2479-2488.

[30] YAN X, ZHANG L, LI W J. Semi-supervised deep Hashing with a bipartite graph[C]//Proceedings of the 26th International Joint Conference on Artificial Intelligence. Melbourne: IJCAJ, 2017: 3238-3244.

[31] QIU Z F, PAN Y W, YAO T, et al. Deep semantic hashing with generative adversarial networks[C]// Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM, 2017: 225-234.

[32] WANG G A, HU Q H, CHENG J, et al. Semi-supervised generative adversarial hashing for image retrieval[C]// Computer Vision – ECCV 2018. Berlin: Springer, 2018: 491-507.

Semi-supervised image retrieval based on triplet hash loss

SHAO Weizhi，XIONG Siyu，PAN Lili*

(School of Computer and Mathematics，Central South University of Forestry and Technology，Changsha 410004，China)

Abstract: Currently, most of the image retrieval methods based on deep learning are supervised techniques, which require massive labeled data. However, it is very difficult and expensive to label so much data in real applications. Furthermore, the network learned picture similarity poorly since the triple loss functions that were in place were computed using Euclidean distance. In this work, a novel semi-supervised hash image retrieval model (SSITL) is proposed that mixes the pseudo-labels with entropy minimization, triplet hash loss and semi-supervised learning. The multi-stage model union and sharpening technique are used to generate pseudo-labels, and the pseudo-labels are processed with entropy minimization to improve their confidence. The triplet hash loss based on the channel weight matrix is utilized to assist SSITL in learning the similarity of images, while the triples are chosen concurrently depending on the clustering outcomes of labeled and unlabeled data. In order to generate a better hash code, Mix Up is used to shuffle between two Hamming embeddings to obtain a new Hamming embedding for image retrieval. The abundant experimental results show that compared with other methods, SSITL improves the average retrieval accuracy by 1.2% and 0.7% respectively on CIFAR-10 and NUS-WIDE datasets under similar time cost, which strongly demonstrates that SSITL is an excellent semi-supervised hash framework for image retrieval.

Keywords: image retrieval; triplet hash loss; semi-supervised learning; pseudo-labels; deep learning

Received：2023-07-10；**Accepted**：2023-09-14；**Published Online**：2023-10-30 21：35

URL：link.cnki.net/urlid/11.2625.V.20231030.1403.004

Foundation items：The General Program of Natural Science Foundation of Hunan Province (2021JJ31164); The Key Program of Science Research Foundation of Education Department of Hunan Province (22A0195)

* **Corresponding author**. E-mail：lily_pan@163.com

http://bhxb.buaa.edu.cn jbuaa@buaa.edu.cn
DOI: 10.13700/j.bh.1001-5965.2023.0475

面向基于知识图谱个性化推荐的诱导信息识别

倪文锴，彭舒凡，杜彦辉*
(中国人民公安大学 信息技术与网络安全学院，北京 102600)

摘　　要：互联网信息服务算法推荐管理，是构建智能信息时代国家互联网治理体系的重要手段。个性化推荐算法是互联网信息服务算法推荐的重要技术之一，知识图谱在个性化推荐算法中有广泛应用，同时知识图谱和推荐算法容易受到攻击者的数据投毒攻击，进而影响推荐结果，造成诱导信息传播。当前，针对此类诱导信息识别缺少有效的模型，基于此开展诱导信息识别模型研究，在对用户历史行为记录及用户偏好的演化过程进行分析的基础上，研究基于用户兴趣与群体感知的诱导信息检测方法，对相似用户群体历史偏好进行群体偏好建模，对具有共性特征的群体内异常曝光的信息进行离群点分析，构建集 node2vec-side 物品表示、高斯混合模型（GMM）群体划分和 LUNAR 异常检测的诱导信息识别模型 NGL，从用户偏好变化与推荐结果演变推理实现诱导信息识别。在 RippleNet 和 MKR 推荐系统上进行诱导信息识别实验，结果表明：NGL 模型优于现有的异常检测模型。

关　键　词：诱导传播；知识图谱；异常检测；群体划分；个性化推荐
中图分类号：TP391
文献标志码：A　　**文章编号**：1001-5965(2025)07-2538-15

互联网信息推荐算法作为新兴的社会资源配置权力和数字经济时代的重要底层架构，逐步成为影响内容分发、服务提供、资源配置的基础性力量，但在其提高信息生产力、追求经济效益的同时，在一定程度上带来了技术、社会和安全风险。目前互联网推荐算法仍存在影响网络舆论、诱导用户沉迷或过度消费等问题，迫切需要加强对互联网推荐算法的规范和管理，进行诱导信息的有效识别检测，以提升防范化解算法推荐安全风险的能力，促进算法相关行业健康有序发展。

当前的推荐系统诱导识别方法大多是针对某一种诱导策略的特性，通过考虑该攻击下用户的某些共有特征，从而判定其中一类用户为恶意用户，即现有检测的应用场景均是针对特定的恶意用户，从攻击特征的角度来检测异常用户，而不能有效地检测未知类型的异常诱导信息或者多攻击混合攻击的情形，同时缺少对诱导信息本身的识别检测，在诱导信息传播时不能很好识别。此外，当前主流的一些异常检测模型并不能很好地适用于推荐算法领域，特别是基于知识图谱的个性化推荐算法中。

基于以上分析，本文针对个性化推荐算法中存在的易受数据投毒攻击并造成信息诱导传播的问题，开展面向基于知识图谱的个性化推荐算法诱导信息识别研究。在对用户历史行为记录及用户偏好的演化过程进行分析的基础上，研究知识图谱信息异常检测和基于用户兴趣与群体感知的诱导信息识别方法，通过提取用户特征向量，对相似用户群体历史偏好进行群体偏好建模，对具有共性特征的群体内异常曝光的信息进行离群点分析，提出面向知识图谱个性化推荐算法的诱导识别模型 NGL，

收稿日期：2023-07-19；**录用日期**：2023-09-12；**网络出版时间**：2023-09-22 09:09
网络出版地址：link.cnki.net/urlid/11.2625.V.20230921.0909.001
基金项目：中国人民公安大学网络空间安全执法技术双一流专项 (2023SYL07)
* **通信作者**. E-mail：duyanhui@ppsuc.edu.cn

引用格式：倪文锴，彭舒凡，杜彦辉. 面向基于知识图谱个性化推荐的诱导信息识别 [J]. 北京航空航天大学学报，2025，51（7）：2538-2552.
NI W K，PENG S F，DU Y H. Identification of induced information for personalized recommendations based on knowledge graph [J]. Journal of Beijing University of Aeronautics and Astronautics，2025，51（7）：2538-2552（in Chinese）.

并在基于知识图谱的推荐系统 RippleNet 和 MKR 上进行实验。

1　相关工作

1.1　面向推荐算法的数据投毒攻击

当前推荐系统对大多数用户都是公开的, 系统会根据用户行为和物品特征进行推荐, 这使得一些攻击者能够轻易攻击系统, 影响推荐结果(信息诱导传播过程如图 1 所示)。针对推荐系统的信息诱导传播主要分为 2 类: ①通用类型数据投毒攻击, 即通过构造虚假用户评分的方式进行信息诱导传播, 然而, 该攻击效用较低, 需要一定数量的虚假用户才能达到攻击目的, 这不仅使攻击成本大幅提高, 同时也使此类攻击更容易被检测; ②针对特定推荐算法进行数据投毒攻击, 攻击者通过在推荐系统中注入一些虚假的、有误导性的数据, 从而干扰推荐算法的学习过程, 达到诱导传播目标信息的目的。

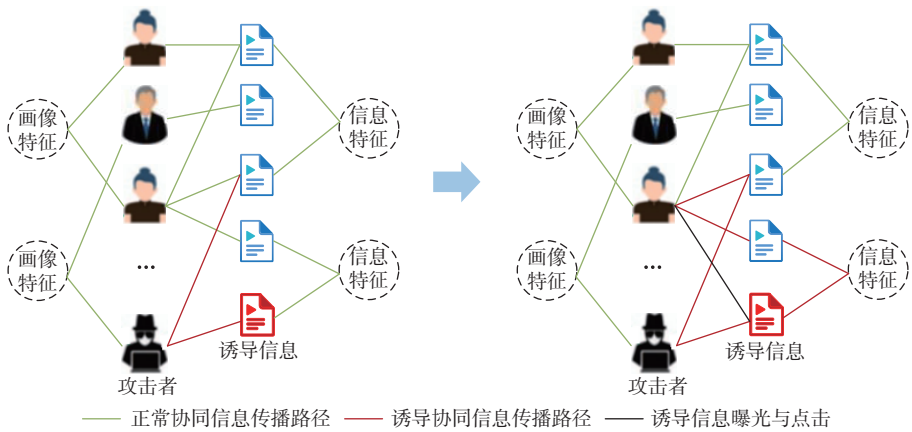


图 1　信息诱导传播模型

Fig. 1　Information induced propagation model

1.1.1　通用数据投毒

通用数据投毒是推荐系统中传统的信息诱导方式, 攻击者通过模拟正常用户的评分行为利用固定规则生成伪造用户画像的伪造物品评分, 并将其注入到推荐系统中从而造成信息诱导传播。虚假画像通常包括选择物品、填充物品、未评分物品和目标物品 4 部分(如图 2 所示)。

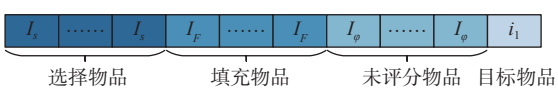


图 2　伪造用户画像

Fig. 2　Fake user portraits

在此类攻击过程中, 攻击者首先会选择目标物品, 旨在推广或打压该物品。然后攻击者会根据不同策略筛选出选择物品, 以使其攻击更加有效, 同时, 攻击者还会选择填充物品, 以增加攻击难度并且更好地隐藏自身。一旦攻击者确定了选择物品和填充物品, 他们将让批量产生的伪造用户按照一定的规则对这些物品进行评分, 创建大量伪造用户画像, 并伪装成与被欺骗的正常用户兴趣相同的人, 最终对推荐结果造成影响。

其中包括: ①平均攻击选定固定的攻击物品后, 选择物品为空, 填充物品随机选择并赋全局的均值分布值; ②流行性攻击选择选取流行的物品作为选择物品项, 并给目标物品和流行物品高分; ③随机攻击选定固定的攻击物品后, 选择物品为空, 填充物品随机选择并随机赋值, 该攻击不需要任何系统知识, 成本较低; ④无组织攻击指来自多个攻击者的分散、无组织的攻击行为。

本文实验部分综合上述 4 种攻击, 按照不同的攻击比例构造混合攻击模式, 采用分散的方式对推荐系统进行攻击, 达到干扰推荐结果的目的。

1.1.2　特殊数据投毒

除了传统的面向推荐系统的通用攻击方式, 攻击者利用基于知识图谱的推荐系统结构和数据特点, 设计更为隐蔽的数据投毒方式。

1) 使用不完整和扰动数据攻击。Zhang 等[1] 针对攻击者受数据收集有限和对训练数据的访问受限等问题, 提出数据缺失攻击和噪声数据攻击 2 种攻击方式, 在数据缺失攻击中, 论文随机地将一些评分设为缺失值, 并将缺失值注入到推荐系统中。在噪声数据攻击中, 作者向一些评分添加了噪声, 并将带有噪声的评分注入到推荐系统中。论文提出了结合概率生成模型的双层优化框架 RAPU-G, 寻找包含较多交互数据且未受到显著干扰的用户和项目, 并据此构造虚假用户-项目交互, 此外, 还提出结合特定上下文的启发式规则来处理数据不完整性和扰动(RAPU-R)并进行数据投毒攻击。

2）面向知识图谱的数据投毒攻击。Zhang 等[2]提出一组针对知识图谱的数据投毒模型，通过在知识图谱上添加或者删除事件来有效操作知识图谱中嵌入模型。Bhardwaj P 等[3]设计对抗性事件添加的中毒攻击，提高对诱导事实的置信度。Bhardwaj 等[4]还提出一种实例归因的方式对知识嵌入进行对抗攻击，采用启发式方式替换或者删除在知识图谱嵌入过程中影响力较大的三元组实现投毒攻击。Wu 等[5]通过研究虚假链接的影响，提出一种基于知识图谱的推荐系统投毒攻击策略，设计相应的攻击模型。该模型能够分析当前推荐状态和特定项目以评估不同攻击策略（添加或者删除事实）的影响，从而找到最佳的攻击组合。

1.2 面向推荐算法的攻击检测技术

目前针对推荐算法中攻击检测技术的研究主要集中在恶意用户的发现上，通过分析恶意用户与正常用户之间的差异来发现恶意攻击的存在，主要包括以下 2 方面角度：①通过用户的异常行为判定恶意用户，Aktukmak 等[6]提出概率分解模型和顺序攻击检测算法。Cai 和 Zhang[7]提出一种基于目标项目与用户评级行为分析的无监督欺诈攻击检测模型（DSA-AURB）。Zhang 等[8]提出一种基于图卷积网络（graph convolutional network, GCN）的用户表示学习框架（GraphRfi）进行欺诈者检测。Yang 和 Niu[9]提出类型信任度（GTD）的概念，提出一种基于 GTD 改进信任矩阵的算法来防御欺诈攻击。Chen 等[10]提出基于信任攻击行为特征的检测模型。Zhang 等[11]基于隐马尔克夫模型和层次聚类，提出无监督攻击检测方法（UD-HMM）。②从组群攻击检测的角度实现恶意用户群体发现。Zhang 等[12]提出针对协同过滤推荐算法的基于图嵌入的群体欺诈攻击检测模型，通过聚类得到攻击群体。Hao 等[13]提出基于 GCN 的混合检测模型，能够同时检测基于模型生成的欺诈攻击和基于群体的欺诈攻击。Jeny 等[14]提出一种基于密度的聚类算法 (density-based spatial clustering of applications with noise, DBSCAN++) 的检测群体攻击的方法，计算并利用可疑度来聚类攻击组群，该方法不仅可以识别个体攻击者，还可以识别群体攻击。

当前对于个性化推荐中的攻击检测主要集中于攻击的发现，即发现恶意用户组群，而由于知识图谱本身的脆弱性，一些特定针对节点进行投毒的攻击更具有隐蔽性。此外，当前的检测方式主要集中于特定的攻击场景，而对于多种类型混合或者新类型攻击普适性不佳。本文从推荐结果入手，旨在从推荐信息的角度从用户终端发现攻击并及时阻止诱导信息传播，实用性更广泛。

2 诱导识别模型 NGL

针对上述分析提出面向知识图谱推荐系统的诱导信息识别模型 NGL（如图 3 所示）主要包括 2 部分：①知识图谱异常信息检测。对推荐系统中使用的知识图谱进行异常节点信息检测，旨在发现针对推荐系统知识图谱本身数据的投毒攻击。其中，面向知识图谱的异常识别引擎是指一种用于在知识图谱中检测异常信息节点的算法，通过对知识图谱的分析初步筛选异常信息节点。②基于用户兴趣与群体感知的异常曝光信息检测。首先，根据历史偏好对用户进行画像表示，针对相似用户群体历史偏好进行细粒度群体划分，然后，通过异常检测模型 LUNAR 从各群体偏好变化与推荐结果演变推理实现异常信息检测。综合 2 部分模块形成诱导信息识别模型，并对基于知识图谱的个性化推荐算法中产生的信息进行识别，及时发现并阻止诱导信息传播。

2.1 知识图谱异常信息检测

本节提出一种基于 node2vec[15] 的 node2vec-side 知识图谱异常信息检测方式。在传统的 node2vec 算法中，节点的向量表示是通过随机游走生成的节点序列来学习的。然而，仅仅考虑节点之间的上下文关系可能无法充分捕捉知识图谱推荐信息间的语义信息。因此，为了更好地表示节点和关系，node2vec-side 引入了节点的关系信息。将推荐领域以三元组保存的知识图谱进行重构并同质化为带权有向图，然后，在该有向图中采用多元随机游走策略（广度优先游走（breadth-first search, BFS）偏向和深度优先游走（depth-first-search, DFS）偏向）分别在节点上进行游走得到节点表示，同时根据生成的游走序列聚合节点间的关系，最终综合各游走策略下的表示结果获得知识图谱的知识表示。

其中，假设 $f(x)$ 为节点 x 在某一偏好策略游走下得到的向量表示，则对于知识图谱中的关系集合 $R=\{r_1,r_2,r_3,\cdots,r_n\}$，假设任意关系 $r_i\in R$，可通过下式得到关系的向量表示形式：

$$\boldsymbol{f}(r_i)=\frac{\sum_i \boldsymbol{f}_i(r_i)}{N_{r_i}}=\frac{\sum_{(h,r_i,t)=G\vee S}(\boldsymbol{f}(t)-\boldsymbol{f}(h))}{\min\left(\text{count}\left(r_i\right),Z\right)} \tag{1}$$

式中：$Z=\lceil\alpha\cdot\text{sum}(G)\rceil$，$Z$ 表示平均化归值，α 表示比例系数且 $\alpha\in(0,1)$，其中 G 表示知识图谱三元组集，即 $G=\{(h,r,t)\}$，$\text{sum}(G)$ 表示知识图谱中三元组的总数；$\mathcal{S}$ 表示游走序列集，即 $\mathcal{S}=\{S_v\mid v\in V\}$，$V$ 为节点集；N_{r_i} 表示关系 r_i 权重值；$\boldsymbol{f}(t)$、$\boldsymbol{f}(h)$ 表示基于

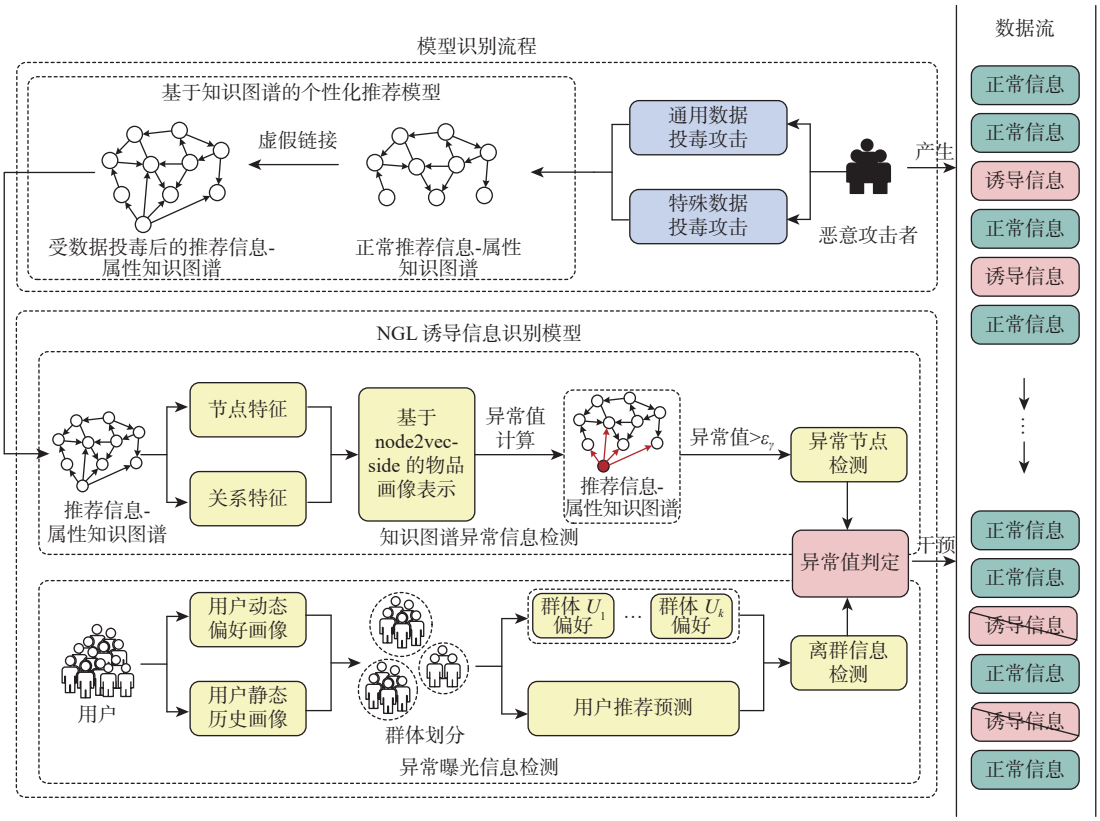


图 3　NGL 诱导识别模型

Fig. 3　NGL induced recognition model

node2vec 表示的头尾节点向量表示; count(r_i) 表示关系 r_i 在知识图谱中出现的次数。单一游走策略下的节点和关系向量集 $\boldsymbol{f}_i(h,r,t)$ 为

$$\boldsymbol{f}_i(h,r,t)=\{\{\boldsymbol{f}_i(h),\boldsymbol{f}_i(r),\boldsymbol{f}_i(t)\}\mid(h,r,t)\in G\} \tag{2}$$

为获得更全面的物品画像, 可采用多种游走策略优化的方式, 通过 DFS 和 BFS 策略获得知识表示 $\boldsymbol{f}_1(h,r,t)$ 和 $\boldsymbol{f}_2(h,r,t)$ 后, 再结合获得最终该知识图谱下的知识表示 $\boldsymbol{f}(h,r,t)$, 其中 $\boldsymbol{f}_1(h,r,t)$ 和 $\boldsymbol{f}_2(h,r,t)$ 表示采用 2 种不同游走策略得到的向量集, $\boldsymbol{f}(h,r,t)$ 表示将两者结合后得到的最终知识表示, 下式中的∪表示向量合并操作, 本文采用向量维度叠加的方式。

$$\begin{aligned}\boldsymbol{f}(h,r,t)&=\boldsymbol{f}_1(h,r,t)\cup\boldsymbol{f}_2(h,r,t)=\\&\{\boldsymbol{f}_1(h)\cup\boldsymbol{f}_2(h),\boldsymbol{f}_1(r)\cup\boldsymbol{f}_2(r),\boldsymbol{f}_1(t)\cup\boldsymbol{f}_2(t)\}\end{aligned} \tag{3}$$

node2vec-side 知识表示的得分函数 $G(h,r,t)$ 如下所示:

$$G(h,r,t)=h(g_{1r}(h,r,t),g_{2r}(h,r,t)) \tag{4}$$

$$\begin{aligned}g_{ir}(\boldsymbol{f}(h,r,t))&=\mathrm{sim}(\boldsymbol{f}(h)+\boldsymbol{f}(r),\boldsymbol{f}(t))=\\\cos(\boldsymbol{f}(h)+\boldsymbol{f}(r),\boldsymbol{f}(t))&=\frac{(\boldsymbol{f}(h)+\boldsymbol{f}(r))\boldsymbol{f}(t)}{\|\boldsymbol{f}(h)+\boldsymbol{f}(r)\|\|\boldsymbol{f}(t)\|}\qquad i=1,2\end{aligned} \tag{5}$$

式中: sim 表示相似度函数, 本文采用余弦相似度作为相似度计算函数。$h(X_1,X_2)$表示融合函数, 在本文中取为平均值函数 $h(X_1,X_2)=(X_1+X_2)/2$。

利用知识图谱中所有涉及某个节点的三元组 (h,r,t) 通过得分函数 $G(h,r,t)$ 计算该节点的异常得分, 然后通过计算该节点周围节点和关系节点的向量表示与该节点向量表示之间的差异来评估该节点的异常程度。如果某个节点的节点异常值超过阈值, 则该节点可能也是异常节点。物品 v 异常值 ε_v 的计算公式如下:

$$\varepsilon_v=\frac{\sum\limits_{(h,r,t)\in V}G(h,r,t)}{\mathrm{count}(V)} \tag{6}$$

式中: V 表示知识图谱中涉及物品 v 的三元组集; count(V) 表示集合 V 中三元组的个数。当物品 v 异常值 ε_v 大于某一阈值 ε 时, 可以认定为该物品 v 属于异常集。算法 1 如下所示:

算法 1　异常节点检测。

输入: 知识图谱三元组集 N, 推荐信息集 V, 异常节点集 Z。

输出: 异常节点集合 Z。

初始化: $Z=\varnothing$, 游走策略集合 $P(p,q)=\{\mathrm{BFS},\mathrm{DFS}\}$、比例系数 $\alpha(0<\alpha<1)$、阈值 value; 利用知识图谱中的三元组 $(h,r,t)\in N$ 得到新的图 $G=(V,E)$, 其中 V 为节点集合, E 为边集合, 初始化 r_emb= $\{e_r|r\in E\}$。

1. edge_weight = Count(r) /*每个三元组 $(h,r,t)\in$

E, 计算边权重 /

2. G.add_weighted_edges_from(edge_weight) / 增加边权重完成知识图谱重构化 /

3. 在 G 上为每个节点 $v \in V$ 进行 r 次随机游走，得到游走序列集 $S_v = \{\boldsymbol{s}_{v,1}, \boldsymbol{s}_{v,2}, \cdots, \boldsymbol{s}_{v,r}\}$

4. node_emb=node2vec(G, dim, walk_length, num_walks, p, q, weight_key) / 通过设定 p,q 进行随机游走得到节点向量 */

5. for $\boldsymbol{s}_{v,i}$ in S_v://获取每个游走序列 $\boldsymbol{s}_{v,i} \in S_v$;

for (h, r, t) in $\boldsymbol{s}_{v,i}$: //获取每一步的头尾节点

$e'_r = e'_t - e'_h$;//e'_h 头节点 h 的向量，e'_t 为尾节点 t 的向量

6. $\boldsymbol{r}_{\text{emb}} = \left\{ e_r | e_r = \dfrac{\sum\limits_{(h,r_i,t)\in E} e'_r}{|\min(\text{edge_weight}, Z)|} \right\}$//作平均处理

7. 执行步骤 d), 选择不同 p,q 进行多元化游走

8. 得到节点向量列表 emb = $\{e_v \mid e_v \in \text{node_emb1} \cup \text{node_emb2}\}$ 和所有关系表示列表 emb$r$ = $\{e_r \mid e_r \in r_\text{emb1} \cup r_\text{emb2}\}$

9. for v in V:

$$\text{score}(v) = \sum_{(h,r,t)\in v} \text{sim}(f_\text{score}(h,r,t))/\text{count}(v)$$//计算推荐信息集中每个信息的得分，其中 f_score 表示得分函数

If score(v) ≥ value:

Z.append(v)

10. 输出 Z

2.2 异常曝光信息检测

本节在对用户行为记录及用户偏好的演化过程进行分析的基础上，引入用户静态历史画像，并与推荐系统中提取的用户画像相结合，形成新的用户画像。同时，结合群体内部用户间的协同行为分析，应用 GMM 进行群体划分，最后对具有共性特征的群体内异常曝光的信息进行离群点分析，并从用户偏好变化与推荐结果演变推理实现诱导信息识别。

2.2.1 用户画像提取

本节根据 node2vec-side 构建的物品画像和用户点击记录构建用户历史静态画像，并与推荐系统生成的画像形成用户 u 的最终画像，研究的目标推荐系统为 RippleNet[16]。

在用户静态历史画像构建部分，将采用 2.1 节构建的物品画像获得物品向量 $\boldsymbol{v}$，通过分析用户 $\boldsymbol{u}$ 的历史点击记录集，将历史点击物品 $\boldsymbol{v}_u$ 的向量和作为用户 $\boldsymbol{u}$ 的静态历史画像。有研究表明[17]，新用户更容易选择那些“热门”推荐的物品，故在用户静态历史画像表示时，对于新用户（即没有点击历史的用户）考虑通过其他用户历史点击次数较多的物品（热门物品）与点击次数较少的物品（冷门物品）计算所得作为新用户静态历史画像，即

$$u_{\text{new}} = \sum_{i=1}^{n} v'_{\text{hoti}} - \beta \cdot \sum_{i=1}^{n} v'_{\text{coldi}} \tag{7}$$

式中：v'_{hoti} 为按点击热度由高到低排序所得第 i 位的物品向量；v'_{coldi} 为按点击热度从低到高排序所得第 i 位的物品向量；β 表示权重值，根据实际情况选择 β 值，默认为 1，旨在平衡热门物品与冷门物品的占比；n 表示选择物品排名前 n，可根据推荐系统中物品总数来确定。

在 RippleNet 的推荐训练过程中提取推荐系统生成的用户画像，综合推荐系统中动态生成的偏好画像和本节构建的用户静态历史画像，得到推荐系统 RippleNet 的用户画像。算法 2 如下所示：

算法 2 用户画像构建。

输入：算法 1 中所得的节点向量列表 emb 和所有关系表示列表 emb$r$；用户历史点击记录字典 user_clicks；权重参数 β。选择的物品数量 n；推荐系统 RippleNet。

输出：推荐系统用户画像向量集 User_portraits。

初始化：User_portraits={}

1. 针对每个用户 u，根据其历史点击记录集 user_clicks_u，计算用户的静态历史画像 static_vector_u

if user_clicks_u=∅: //采用式（7）中的方法计算静态历史画像

hot_items = v[:n]//按照热度由高到低排序的前 n 个物品向量

cold_items = v[-n:]//按照热度由低到高排序的后 n 个物品向量

static_vector_u=sum(hot_items)+βsum(cold_items)

else:

clicked_items_u = []

for item_id in user_clicks_u:

clicked_items_u.append(emb [item_id])

static_vector_u=sum(clicked_items_u) / len(clicked_items_u)

2. 在 RippleNet 的推荐训练过程中，结合用户的动态生成的偏好画像，得到最终的用户画像 user_profile_u。

user_profile_u = static_vector_u + ripple_net_profile_u

User_portraits.append(user_profile_u)

3. 输出 User_portraits

2.2.2　群体划分

本节采用 GMM 对用户画像进行聚类, 进行群体划分, 以满足复杂场景下的推荐需求和安全需求, 通过对用户数据进行聚合和分析, 识别用户的行为模式和兴趣偏好。

群体划分可以更好地达到高内聚低耦合的效果, 即同一个群体内的用户具有高度相似的属性或特征, 而不同群体之间的数据点具有明显的区分度。GMM 是常见一种的原型聚类, 与 k 均值基于距离进行聚类不同, GMM 基于概率分布来进行聚类[18]。大部分自然界中的随机变量都符合高斯分布, GMM 采用最大化后验概率的成分选择方法来实现聚类, 每个数据点的后验概率反映了其属于不同类别的可能性, 而非单一类别的确定性。通过对提取的用户画像进行 GMM 聚类, 将用户划分到不同的群体, 每个群体代表一组具有相似兴趣偏好的用户。在群体划分基础上可进行群体偏好演化分析。群体偏好演化分析旨在观察不同群体在兴趣和偏好上的变化趋势, 通过监测群体中的推荐物品情况来分析群体偏好的演化情况。

图 4 为 Book-Crossing 数据集在 RippleNet 推荐模型中面对各种攻击后的群体划分结果。其中, 每个子图表示一种具体攻击条件下的处理结果。通过子图的比较可以发现, 在不同攻击条件下, 采用 GMM 进行用户群体划分的结果是不同的, 但都存在一些大群体和小群体。这表明, 不同攻击对用户行为模式产生的影响并非完全一致, 这为后续针对各群体间分析提供了基础。

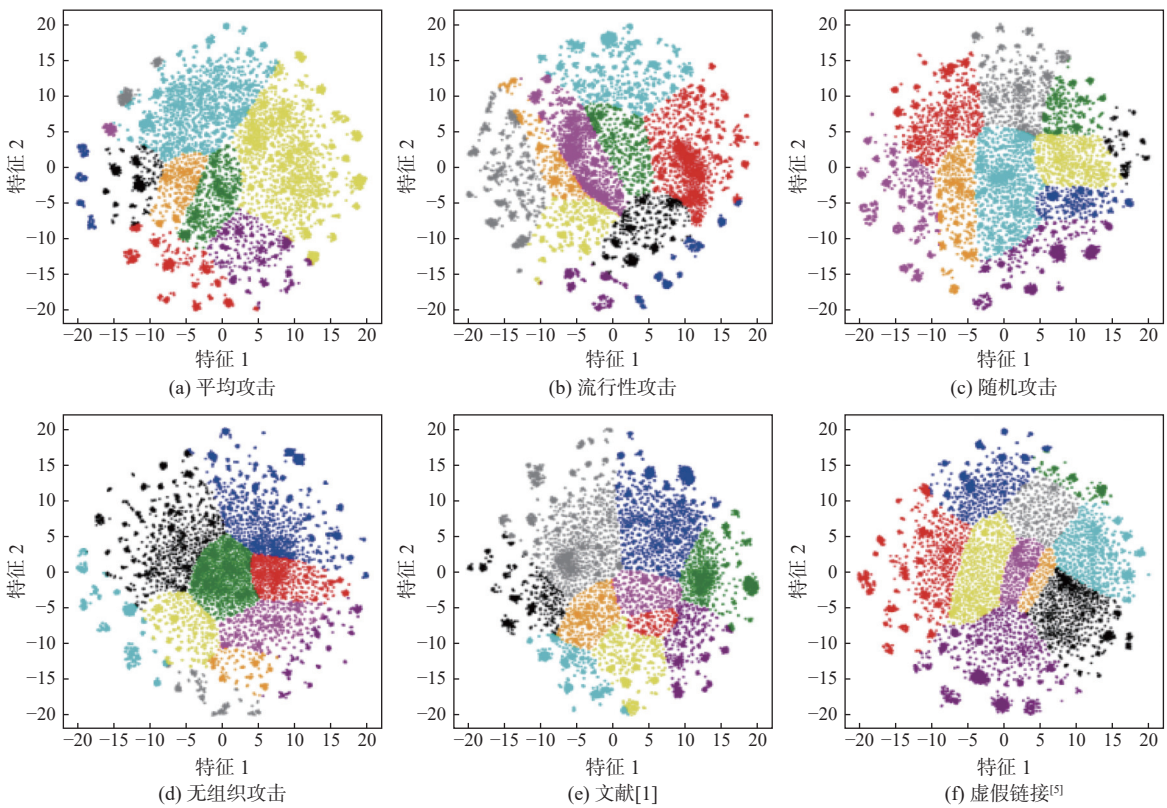


图 4　用户群体划分

Fig. 4　User group division

2.2.3　离群信息检测

不同于群体间的流行偏好信息, 经过数据投毒的诱导信息虽然在群体间能达到一定的推广度, 但与其他信息在属性、特征等方面的相似度上存在差异, 而这种差异能通过有效的离群检测模型进行识别, 从而筛选出某个群体间的诱导信息。基于此, 本节基于 LUNAR 模型[19] 采用 GCN 的方式, 聚合邻居节点, 对具有共性特征的群体内异常曝光的信息进行离群点分析, 从用户偏好变化与推荐结果演变推理实现诱导信息传播识别。

针对该推荐系统中的用户, 可以通过推荐算法获得其个性化的推荐列表, 选取推荐结果的前 N 个作为推荐列表 top_N; 然后分别构建各群体的推荐信息集 $N=\{v_1,v_2,v_3,\cdots,v_n\}$; 最后在每个群体间针对推荐信息集 N 进行异常检测, 经过检测后模型通过将类似推荐信息聚类进行分组(如图 5 中虚线圆形围成节点), 而未能进入任何分组的节点将被判定为离群点, 也即异常节点, 如果在某群体间某推荐信息判断为异常点(如图 5 中红色节点), 则该信息的异常值增加 1。为了判断某一异

常信息是否属于推荐系统中的诱导信息，可以设定一个异常值阈值ξ，如果推荐信息的异常值超过该阈值，则可以判断为诱导信息。如图5所示，通过该异常信息检测方法可以识别和过滤可能存在的诱导信息，诱导信息集合呈现出一种“金字塔”型分布，即自底向上，异常值逐层增加，越接近塔尖的位置，异常值越高，相应集合中属于诱导信息的可能性也越大。在应用中，根据具体场景设定不同的阈值ξ，模型将$\varepsilon \geqslant \xi$的信息集合（如图中$V_{n1}$、$V_{n7}$）判定为诱导信息集。

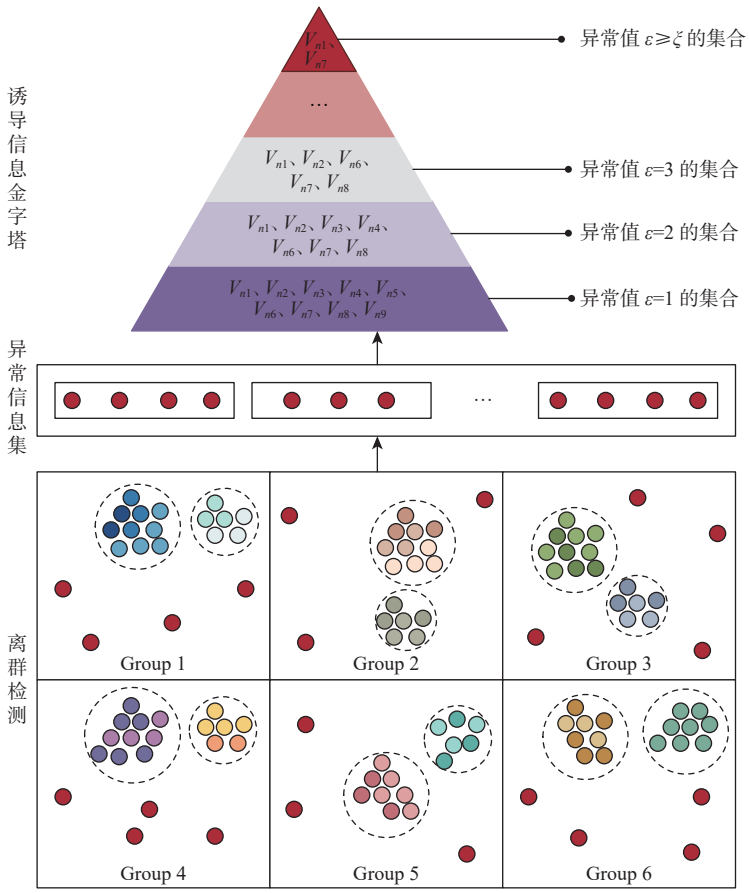


图5 诱导信息检测过程

Fig. 5 Induced information detection process

3 实验设计及分析

本节为验证NGL模型的诱导信息识别能力，首先在推荐系统常用数据集中选择若干目标信息采用不同的数据投毒方式实现诱导信息传播，然后在该场景中实现对目标信息的诱导进行识别，并与当前主流的异常检测模型进行实验对比。

3.1 数据集分析

MovieLens-1m是一个电影推荐公开数据集，其中包括6 040个用户和3 706部电影的1 000 209条评分数据。相应的知识图谱中有182 011个实体、12个关系。

Book-Crossing是一个书籍推荐的公开数据集，包括105 283个用户和340 555部书的1 149 780条评分数据，相应知识图谱中有77 903个实体、25个关系。

Last.FM是一个音乐推荐的公开数据集，包括1 892个用户和17 632位艺术家的92 834条权重数据，相应的知识图谱中有9 366个实体、60个关系。具体各数据集信息如表1所示。

3.2 实验设置与分析

本次实验选择基于知识图谱的推荐系统RippleNet[16]和MKR[20]作为推荐应用场景。

3.2.1 目标信息诱导传播

本节实现各数据集下特定目标信息在两种推荐系统中的诱导传播，选择若干目标信息构建诱导信息集R进行传播，采用当前主流的通用数据投毒攻击，包括平均攻击、流行性攻击、随机攻击和混合攻击以及2类特殊数据投毒攻击，包括Zhang等[1]和虚假链接[5]，虚假链接将通过向知识图谱中添加虚假三元组的方式实现诱导传播。表2展示通用数据投毒攻击的各参数设置。攻击规模、填充规模和选择规模是通用诱导攻击中的基本参数，其中，攻击规模反映恶意用户的数量与系统正常用户数

表 1　推荐系统数据集统计信息

Table 1　Statistics for recommended system datasets

数据集	用户数	项目数	交互数	三元组数
MovieLens-1M	6 036	2 445	753 772	1 241 995
Book-Crossing	17 860	14 967	139 746	151 500
Last.FM	1 872	3 846	42 346	15 518

表 2　通用数据投毒攻击参数

Table 2　General data poisoning attack parameter

数据集	攻击规模	填充规模	选择规模
MovieLens-1M	0.15	0.015	0.05
Book-Crossing	0.2	0.015	0.1
Last.FM	0.2	0.015	0.1

的比例。攻击规模、填充规模和选择规模将决定诱导攻击强度。

本文将平均命中率 $R_{\text{hit@}K}$ 作为评价诱导传播策略的主要指标。该指标定义为在用户推荐列表前 K 个中包含目标信息的用户所占比例:

$$\frac{1}{|I|}\sum_{i\in I}R_{\text{hit@}K}=\frac{\sum_{u\in U_i}H_{ui}}{|I||U_i|}\qquad H_{ui}=\begin{cases}1 & rk_i\leqslant K\\ 0 & rk_i>K\end{cases}\tag{8}$$

式中: i 表示目标信息; I表示目标信息集; $|I|$表示目标项的数量; U_i表示用户集; $|U_i|$表示用户的数量; rk_i表示目标信息 i 在用户推荐列表的位置。3 个数据集在 RippleNet 和 MKR 中进行数据投毒后的结果如表 3 所示, 其中, None 表示不存在任何数据投毒的情况下目标信息的受众度; 当攻击后的命中率高于攻击前时, 说明目标信息被推荐给更多的用户, 从而实现了目标信息诱导传播目的。

表 3　推荐信息诱导传播对比

Table 3　Comparison of recommendation message induced propagation　%

数据集	攻击	RippleNet				MKR			
		$R_{\text{hit@1}}$	$R_{\text{hit@3}}$	$R_{\text{hit@5}}$	$R_{\text{hit@10}}$	$R_{\text{hit@1}}$	$R_{\text{hit@3}}$	$R_{\text{hit@5}}$	$R_{\text{hit@10}}$
MovieLens-1M	无攻击	0.014 4	0.044 1	0.070 6	0.137 1	0.003 5	0.025 9	0.055 7	0.138 9
	平均攻击	0.043 5	0.145 2	0.251 3	0.512 5	0.324 2	0.576 0	0.654 2	0.963 7
	流行性攻击	0.031 1	0.102 7	0.190 9	0.401 6	0.330 7	0.822 0	1.293 6	2.220 5
	随机攻击	0.024 8	0.104 5	0.197 7	0.414 3	0.259 2	0.835 3	1.482 8	2.738 6
	无组织攻击	0.049 8	0.153 9	0.248 5	0.438 6	0.229 4	0.689 3	1.123 3	2.180 3
	文献[1]	0.029 8	0.126 1	0.240 3	0.468 5	0.152 1	0.476 8	0.715 2	1.267 5
	虚假链接[5]	0.098 5	0.349 4	0.582 4	1.146 1	0.006 3	0.026 3	0.090 3	0.290 0
Book-Crossing	无攻击	0.004 9	0.016 6	0.028 5	0.058 5	0.000 0	0.000 0	0.000 0	0.000 0
	平均攻击	0.043 3	0.074 7	0.118 0	0.202 3	0.024 8	0.605 5	0.866 1	1.167 3
	流行性攻击	0.004 8	0.033 7	0.046 4	0.084 4	0.000 0	0.040 3	0.037 4	0.371 1
	随机攻击	0.029 1	0.067 9	0.089 9	0.162 1	0.626 9	1.587 1	2.247 3	3.366 2
	无组织攻击	0.011 2	0.041 8	0.063 0	0.119 8	0.048 5	0.083 4	0.142 6	2.116 4
	文献[1]	0.006 7	0.029 5	0.051 2	0.117 1	0.131 7	0.716 2	1.650 3	3.405 9
	虚假链接[5]	0.007 9	0.019 6	0.033 8	0.092 9	0.000 0	0.024 8	0.023 0	0.065 3
Last.FM	无攻击	0.018 2	0.089 7	0.130 3	0.242 5	0.000 0	0.000 0	0.000 0	0.027 5
	平均攻击	0.056 5	0.165 2	0.312 7	0.731 3	0.138 3	0.418 2	0.720 9	1.488 7
	流行性攻击	0.046 9	0.207 6	0.280 1	0.572 5	0.051 8	0.103 6	0.145 5	0.345 2
	随机攻击	0.031 4	0.081 2	0.157 0	0.392 5	0.244 6	0.971 0	1.287 0	2.655 4
	无组织攻击	0.035 2	0.094 4	0.158 6	0.361 6	0.313 5	0.884 7	1.455 6	2.947 7
	文献[1]	0.051 6	0.185 8	0.364 1	0.827 9	0.039 0	0.099 3	0.192 7	0.459 9
	虚假链接[5]	0.061 4	0.147 6	0.239 0	0.537 8	0.008 2	0.032 1	0.188 6	0.418 7

3.2.2　诱导信息识别

本节实验采用白盒测试和黑盒测试 2 种方法来评估模型的性能和识别诱导信息的能力。

1) 白盒测试。2.2.1 节在用户画像提取阶段对 RippleNet 的推荐过程进行了分析, 向其引入 node2vec-side 物品画像进行优化, 因此在白盒测试中选择 RippleNet 作为目标推荐系统。

白盒测试中采用 NGL 模型对 3.2.1 节中实现的诱导信息进行识别, 由于当前针对个性化推荐算法的攻击检测模型主要用于发现恶意用户上, 不具备识别诱导信息传播的能力, 故本次实验采用的对比模型包括当前主流的异常检测模型: 基于角度的异常检测模型 ABOD[21], 基于聚类的异常检测模型 KNN[22], 基于孤立森林的异常检测模型 INNE[23]、I

Forest[24]，基于距离的异常检测模型MCD[25]、ECOD[26]、ROD[27]，基于图的异常检测模型LUNAR[19]，基于密度的异常检测模型COPOD[28]、HBOS[29]，基于深度学习的异常检测模型ALAD[30]。对于NGL的参数设置如表4所示，其中，α分别取为0.1、0.2和0.1，β取为0.1，n设置为10，表中n_cluster表示群体划分的个数，n表示节点聚合的邻居个数，d表示向量维度，top_N表示获取用户的推荐列表前N项，并根据不同的应用场景选取阈值ξ。

表4 白盒测试中NGL参数设置

Table 4 NGL parameter settings in white-box testing

数据集	n_cluster	n	epsilon	d	top_N
MovieLens-1M	15	30	0.01	48	10
Book-Crossing	10	20	0.1	64	10
Last.FM	15	20	0.1	48	25

对不同攻击策略下的识别检测结果如表5～表7所示，表中的缺失值表示未能完成异常检测任务。实验采用的评价指标为精确率P和AUC(ROC曲线下与坐标轴围成的面积)，其精确率P如下所示：

$$P = T_P/(F_N + F_P) \tag{9}$$

式中：T_P表示正确识别的诱导信息；F_N表示诱导信息误判为正常信息的数目；F_P表示正常信息误判为诱导信息的数目。

实验结果表明，NGL模型和其他异常检测模型的检测效果与推荐系统中诱导信息传播的程度相关，即诱导信息传播程度越高，相应的检测效果越好。原因是即使推荐系统受到攻击，如果没有在用户群体中造成诱导传播，各模型也很难将其判定为诱导信息。同时，对不同攻击方式的模型精度也不同，一般来说攻击越有效，NGL模型识别的精度也

表5 RippleNet中MovieLens-1m诱导识别结果

Table 5 Induced recognition results of MovieLens-1m in RippleNet

模型	P/%						AUC					
	平均攻击	流行性攻击	随机攻击	无组织攻击	文献[1]	虚假链接[5]	平均攻击	流行性攻击	随机攻击	无组织攻击	文献[1]	虚假链接[5]
ABOD	6.45	8.33	14.15	9.26	9.28	5.28	61.69	57.66	65.43	63.27	55.63	59.40
KNN	10.08	8.85	14.58	13.79	7.03	8.06	59.68	59.05	62.72	62.14	59.45	60.75
INNE	6.68	7.14	11.94	8.91	7.41	9.15	57.31	61.95	63.79	61.28	56.11	58.61
MCD	6.48	6.04	19.25	14.29	8.35	7.58	61.73	55.40	65.93	65.78	58.67	60.90
COPOD	3.18	6.61	16.34	9.51	9.88	12.51	56.73	59.26	60.66	57.49	68.89	72.75
LUNAR	7.69	8.27	22.49	12.13	7.79	10.81	63.31	61.81	66.82	63.51	61.19	62.45
ECOD	3.26	4.35	14.82	10.53	10.53	16.56	58.43	63.57	65.44	58.19	65.31	71.84
I Forest	2.42	6.25	11.48	8.97	7.69	13.09	57.80	59.64	61.97	58.41	62.25	66.20
HBOS	3.50	7.18	9.69	8.11	7.41	14.16	57.96	60.28	62.81	58.71	63.27	68.00
ALAD	3.29	6.12	6.82	8.57	11.84	15.93	59.69	55.88	59.12	56.47	62.57	69.73
ROD	1.85	5.71	6.77	4.86	12.71	14.29		55.27	56.71	53.55	61.76	72.46
NGL	18.16	10.53	23.91	17.24	25.56	38.81	71.65	65.13	71.04	68.08	73.90	79.89

表6 RippleNet中Book-Crossing诱导识别结果

Table 6 Induced recognition results of Book-Crossing in RippleNet

模型	P/%						AUC					
	平均攻击	流行性攻击	随机攻击	无组织攻击	文献[1]	虚假链接[5]	平均攻击	流行性攻击	随机攻击	无组织攻击	文献[1]	虚假链接[5]
ABOD	21.40	1.72	16.24	5.12	8.80		82.68	62.04	77.18	62.11	64.34	
KNN	27.09	3.30	16.65	17.84	9.83	0.35	83.55	57.96	71.89	76.57	66.44	52.98
INNE	16.13	7.52	9.28	14.28	7.91	1.76	78.49	61.20	65.20	70.70	60.26	57.18
MCD	21.94	3.08	34.25	4.49	0.99	3.74	72.81	66.84	84.39	59.32	56.29	55.16
COPOD	8.82	1.36	2.31	7.24	6.13	3.08	63.46	56.75		65.88	73.81	58.01
LUNAR	22.80	3.33	18.45	8.22	7.67	6.64	72.23	66.19	75.32	71.65	69.13	59.54
ECOD	15.49	2.15	4.87	8.35	7.06	0.86	73.50	58.94	63.80	66.31	65.80	58.29
I Forest	15.58	2.45	11.31	15.25	8.23	0.27	64.84	57.63	67.99	73.88	71.06	55.95
HBOS	14.84	2.04	5.41	8.65	6.25	1.44	63.09	54.70	64.86	66.28	72.92	55.22
ALAD	2.62	3.11	3.95	3.22	1.65	2.81	59.19	56.01	57.51	57.51	56.51	60.87
ROD	14.24	4.47	5.26	7.16	7.59	2.50	61.33	57.51	58.02	62.63	64.38	56.99
NGL	36.32	11.58	39.36	25.43	14.23	12.05	88.64	74.23	89.41	84.79	78.22	75.11

表 7　RippleNet 中 Last.FM 诱导识别结果

Table 7　Induced recognition results of Last.FM in RippleNet

模型	*P*/%						AUC					
	平均攻击	流行性攻击	随机攻击	无组织攻击	文献[1]	虚假链接[5]	平均攻击	流行性攻击	随机攻击	无组织攻击	文献[1]	虚假链接[5]
ABOD	10.94	7.66	7.33	4.43	6.02	2.48	57.51	63.18	65.90	57.62	54.58	
KNN	14.60	10.68	14.21	20.34	6.03	8.27	60.18	63.41	68.42	66.61	57.62	58.16
INNE	7.79	7.12	12.64	8.32	9.82	6.14	64.88	62.55	61.17	57.29	72.18	54.95
MCD	9.55	6.32	6.06	6.89	4.83	7.40	55.07	59.80	64.19	57.41	55.63	55.71
COPOD	21.10	3.29	6.58	4.95	20.55	13.09	70.13	55.75	62.05		76.69	61.58
LUNAR	11.42	9.08	17.91	12.05	9.33	7.84	67.72	63.16	71.10	63.19	62.15	54.06
ECOD	11.11	10.91	8.07	13.55	12.22	8.29	65.82	59.15	62.76	60.49	75.01	55.62
I Forest	14.38	10.57	7.20	12.31	11.44	8.04	65.57	61.88	61.84	58.43	74.59	56.32
HBOS	22.89	9.34	7.81	10.61	13.78	7.68	69.16	58.91	61.17	58.02	72.03	56.18
ALAD	9.85	3.21	8.40	8.91	5.17	5.82	69.23	54.03	60.83	57.09	55.89	54.84
ROD	14.15	8.24	6.19	6.90	14.18	3.70	72.23	59.03	61.90	56.49	69.63	
NGL	38.51	13.94	25.51	22.73	26.81	18.23	75.98	72.38	76.47	73.79	79.41	74.59

越高。相比其他模型, NGL 模型考虑了信息诱导传播的影响, 通过对用户行为序列的建模和分析, 可以更好地识别异常信息的诱导传播, 提高检测的准确性。NGL 采用 LUNAR 作为异常检测的核心模块, LUNAR 中采用 GCN 对图数据的学习和表示, 更好地捕捉推荐信息不同层次的特征, 具有更好的异常诱导识别泛化能力, 而基于聚类或者基于距离的异常检测模型通常只考虑数据点之间的距离关系, 忽略了数据之间的拓扑结构关系。与此同时, NGL 模型融入对知识图谱的异常检测模块, 在知识图谱上发现异常节点, 对虚假链接数据投毒攻击的检测也更有针对性, 这也是现有针对恶意用户检测方式不具有的优势。

2) 黑盒检测。该部分将 MKR 作为黑盒检测的目标个性化推荐系统, MKR 是华为云推出的一种基于知识图谱的深度学习推荐模型, 其使用知识图谱来提高推荐系统的效果, 并在多任务学习的框架下进行训练。实验已知信息包括 MKR 个性化推荐系统采用的知识图谱和用户历史行为记录, 对 MKR 内部推荐过程未知, 针对 MKR 的推荐结果进行诱导信息识别, NGL 参数设置如表 8 所示, 其他参数与白盒设置保持一致。

表 8　黑盒测试中 NGL 参数设置

Table 8　NGL parameter settings in black-box testing

数据集	*n*_cluster	*n*	epsilon	*d*	top_*N*
MovieLens-1M	10	30	0.05	16	10
Book-Crossing	10	20	0.01	16	25
Last.FM	10	20	0.1	16	25

对不同攻击策略下的识别检测结果如表 9～表 11 所示, 实验指标与白盒测试相同。

3.3　识别结果分析

经过上述实验结果分析, 可以得出结论: 无论是黑盒测试还是白盒测试, NGL 模型在异常诱导信息的检测方面表现均优于现有的异常检测模型。在相同规模的攻击下, 结合表 3 可知相同类型攻击下 MKR 推荐系统中目标信息诱导传播程度更大, 由图 6 可知 NGL 模型在 MKR 推荐系统中的识别精确度更高, 同时相较于其他攻击, NGL 对随机攻击的检测精确度也更高, 原因可能是攻击者采用随机赋值评分的方式对目标信息进行诱导传播, 在用户群体中相较于其他正常信息异常程度更大, 更容易被模型检测。此外, 模型对虚假链接造成传播的诱导信息识别能力较低, 主要原因是该类型的投毒攻击主要毒化知识图谱结构, 使得诱导传播更隐蔽, 单纯的离群分析难以发现异常点, 需结合知识图谱本身的异常检测进一步识别诱导信息节点。

3.4　消融实验

本节研究群体划分和知识图谱异常节点检测对模型识别的结果, 主要采用的个性化推荐模型为 RippleNet, 选择的代表诱导攻击方式为无组织攻击、文献 [1]、虚假链接[5], 实验对比情况如表 12 所示。

由表 12 可知, 群体划分和知识图谱异常节点检测对模型的识别能力均有一定影响: 在引入群体划分后, 面对无组织攻击、文献 [1] 的诱导方式, 模型能够在群体内部根据群体偏好对诱导信息进行

表9 MKR中MovieLens-1m诱导识别结果

Table 9 Induced recognition results of MovieLens-1m in MKR

模型	*P*/%						AUC					
	平均攻击	流行性攻击	随机攻击	无组织攻击	文献[1]	虚假链接[5]	平均攻击	流行性攻击	随机攻击	无组织攻击	文献[1]	虚假链接[5]
ABOD	14.29	34.91	7.14	13.80	14.71	3.85	60.90	78.44	53.41	62.76	66.49	55.20
KNN	29.41	35.27	4.76	24.96	23.61	4.95	59.45	79.60	53.10	62.83	68.12	56.73
INNE	15.79	39.57	33.91	23.22	28.92	3.70	71.58	82.02	76.31	74.89	75.97	58.08
MCD	6.40	34.86	4.08	6.81	22.65	1.63	56.63	79.08	56.11	63.95	72.86	
COPOD	5.04	5.10	3.27	4.77	19.26	3.13	60.72	56.12	57.15	62.00	67.34	55.01
LUNAR	32.74	38.42	14.29	35.71	29.63	3.86	66.84	81.93	58.68	62.75	71.39	57.39
ECOD	7.41	15.79	1.82	2.17	23.08	7.42	65.83	63.51		60.21	65.30	58.26
I Forest	5.44	37.21	34.49	7.14	18.37	2.44	62.24	77.18	77.72	72.70	74.07	55.64
HBOS	7.69	22.71	6.46	1.99	21.43	4.08	57.65	69.28	63.26	69.12	66.73	56.40
ALAD	11.23	17.24	19.59	12.84	23.03	3.76	69.91	67.55	71.60	67.32	69.40	56.91
ROD	7.14	3.27	2.56	5.40	15.79	2.04	70.24		55.10	65.05	68.88	
NGL	35.71	40.55	43.19	41.22	37.54	18.79	77.95	82.47	81.97	83.14	78.84	70.35

表10 MKR中Book-Crossing诱导识别结果

Table 10 Induced recognition results of Book-Crossing in MKR

模型	*P*/%						AUC					
	平均攻击	流行性攻击	随机攻击	无组织攻击	文献[1]	虚假链接[5]	平均攻击	流行性攻击	随机攻击	无组织攻击	文献[1]	虚假链接[5]
ABOD	29.76	9.67	29.15	9.43	10.74	2.31	75.56	64.12	79.54	65.10	62.71	58.24
KNN	27.49	24.77	54.72	29.58	14.93	5.44	75.42	75.57	88.89	80.69	62.26	54.86
INNE	22.07	17.49	34.72	24.10	41.26	1.96	67.35	71.48	85.07	79.32	83.23	
MCD	16.13	6.00	57.81	10.48	16.05	9.81	61.40	58.90	89.44	64.37	80.21	61.91
COPOD	14.90	20.97	47.78	9.09	51.01	4.05	64.66	72.85	86.84	69.53	85.59	53.43
LUNAR	29.42	27.46	46.67	34.03	13.62	5.99	74.69	79.11	82.72	86.95	73.20	59.87
ECOD	8.46	20.76	36.11	7.23	48.27	3.63	62.00	70.38	80.97	73.32	77.74	55.78
I Forest	18.23	14.72	49.81	36.29	50.50	4.45	66.03	65.37	86.67	85.05	82.65	54.39
HBOS	14.05	13.84	43.59	2.67	44.76	1.44	63.44	65.74	82.04	63.71	83.43	
ALAD	3.43	28.89	18.78	2.38	6.44	5.89	59.71	82.39	73.48	61.81	60.02	57.51
ROD	4.22	11.08	31.11	6.27	1.87	1.03	68.53	71.23	82.75	77.54		
NGL	33.69	35.48	71.45	43.55	55.47	11.46	86.69	87.64	96.88	91.08	89.27	82.58

表11 MKR中Last.FM诱导识别结果

Table 11 Induced recognition results of Last.FM in MKR

模型	*P*/%						AUC					
	平均攻击	流行性攻击	随机攻击	无组织攻击	文献[1]	虚假链接[5]	平均攻击	流行性攻击	随机攻击	无组织攻击	文献[1]	虚假链接[5]
ABOD	12.99	26.67	45.15	16.59	25.81	17.86	69.46	71.09	74.30	67.82	67.30	71.99
KNN	20.18	30.21	55.82	29.64	36.82	14.29	75.30	66.28	79.13	72.15	64.63	72.69
INNE	22.91	32.91	46.41	14.48	51.73	13.64	74.68	65.59	73.79	66.38	78.82	68.02
MCD	9.09	29.72	59.47	17.20	46.81	6.90	56.33	70.26	87.93	74.26	74.62	61.04
COPOD	17.10	8.71	55.34	5.35	20.93	4.77	68.99	64.58	86.89	62.23	59.47	62.54
LUNAR	27.53	32.81	57.17	30.33	48.15	17.24	79.38	77.61	85.59	76.48	75.63	71.69
ECOD	10.73	17.59	49.93	4.04	19.15	4.17	71.64	68.48	82.60	59.61	58.46	63.44
I Forest	37.04	24.54	61.95	54.26	49.46	7.90	78.01	71.47	87.48	79.63	77.80	62.85
HBOS	33.91	9.09	60.90	53.25	28.29	4.57	79.27	63.42	87.31	78.83	59.57	65.48
ALAD	18.72	18.89	29.17	41.67	42.55	9.92	72.29	68.82	66.83	73.28	77.26	62.14
ROD	14.81	7.11	10.37	8.46	30.37	3.63	71.85	67.23	69.75	65.44	72.54	66.97
NGL	42.62	41.57	70.52	61.35	57.46	33.68	81.26	79.34	92.06	85.49	83.95	95.12

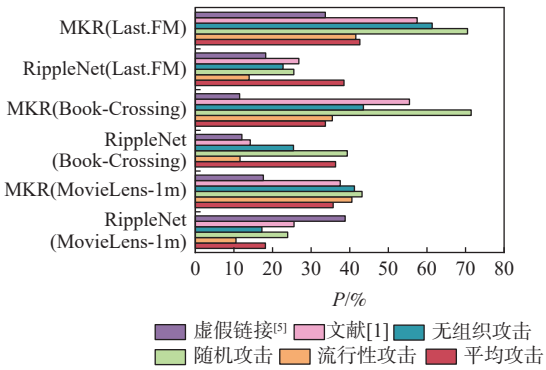


图 6　NGL 模型对各攻击的诱导识别结果

Fig. 6　The induced recognition results of NGL model for each attack

有效识别, 在 3 个数据集上的准确度和 AUC 上均有明显提高, 而相对于虚假链接[5] 提高并不明显, 可能原因在于后者采用的是面向知识图谱本身的数据投毒, 系统中并不存在大量恶意用户, 使得这种方式的诱导信息更为隐蔽, 但是在引入知识图谱异常节点检测模块后, 模型的识别能力有了较大的提升, 而其他 2 种诱导攻击方式的提升效果不明显, 因此通过知识图谱异常节点检测模块主要针对的是面向知识图谱本身的数据投毒攻击。在应用场景中, 攻击者采用的诱导方式并不确定, 因此通过采用 2 个模块结合的方式能够有效提升模型的识别能力, 应对多种攻击场景。

3.5　群体划分对模型的影响

本节研究 NGL 模型在个性化推荐系统中面对不同用户群体时, 群体划分对于诱导信息识别结果的影响, 其中图 7、图 8 分别表示 3 个数据集下

表 12　不同部分对模型识别结果的影响

Table 12　Impact of different parts on model recognition results

数据集	模块		P/%			AUC		
	群体划分	异常节点检测	无组织攻击	文献[1]	虚假链接[5]	无组织攻击	文献[1]	虚假链接[5]
MovieLens-1M	–	–	12.91	7.91	11.08	63.88	61.87	62.51
	√	–	16.63	25.40	13.11	68.05	73.03	66.67
	√	√	17.25	25.23	38.52	68.93	73.47	79.78
Book-Crossing	–	–	8.47	8.62	7.32	72.56	69.72	59.97
	√	–	24.75	13.92	7.99	84.31	77.23	60.67
	√	√	25.43	14.23	12.05	84.79	78.22	75.11
Last.FM	–	–	12.05	9.33	7.84	63.19	62.15	54.06
	√	–	22.38	26.08	9.90	73.59	78.46	58.76
	√	√	22.73	26.81	18.23	73.79	79.41	74.59

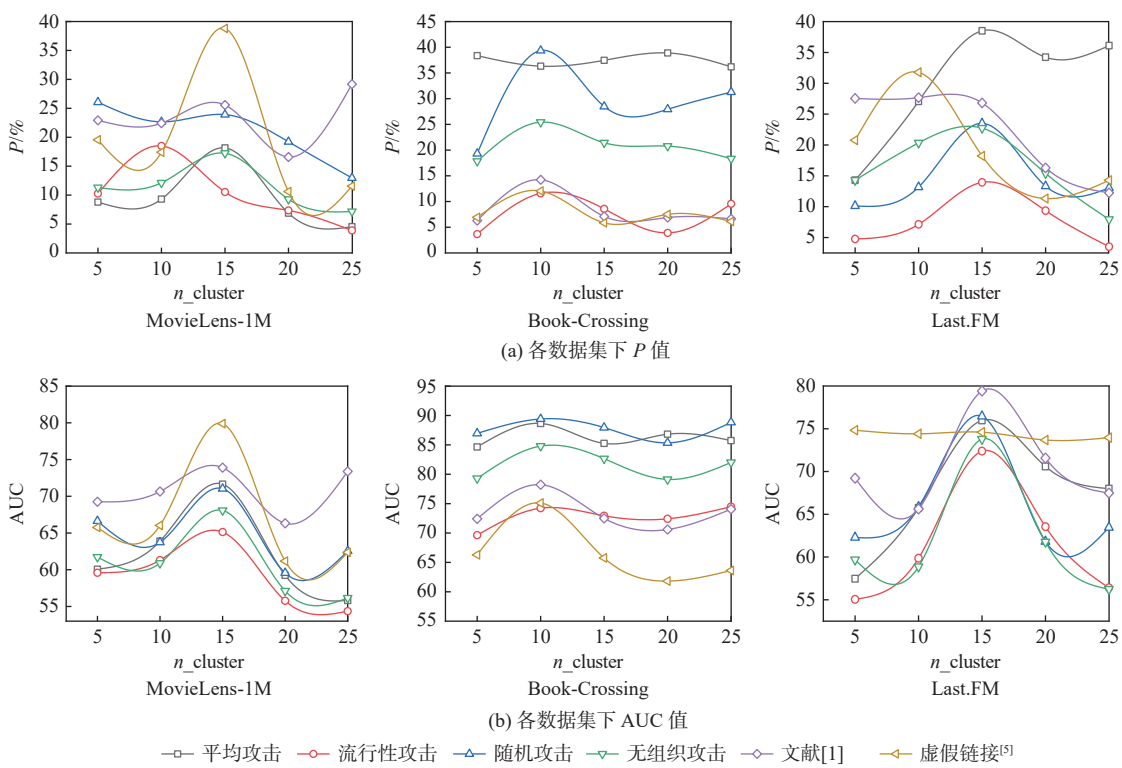
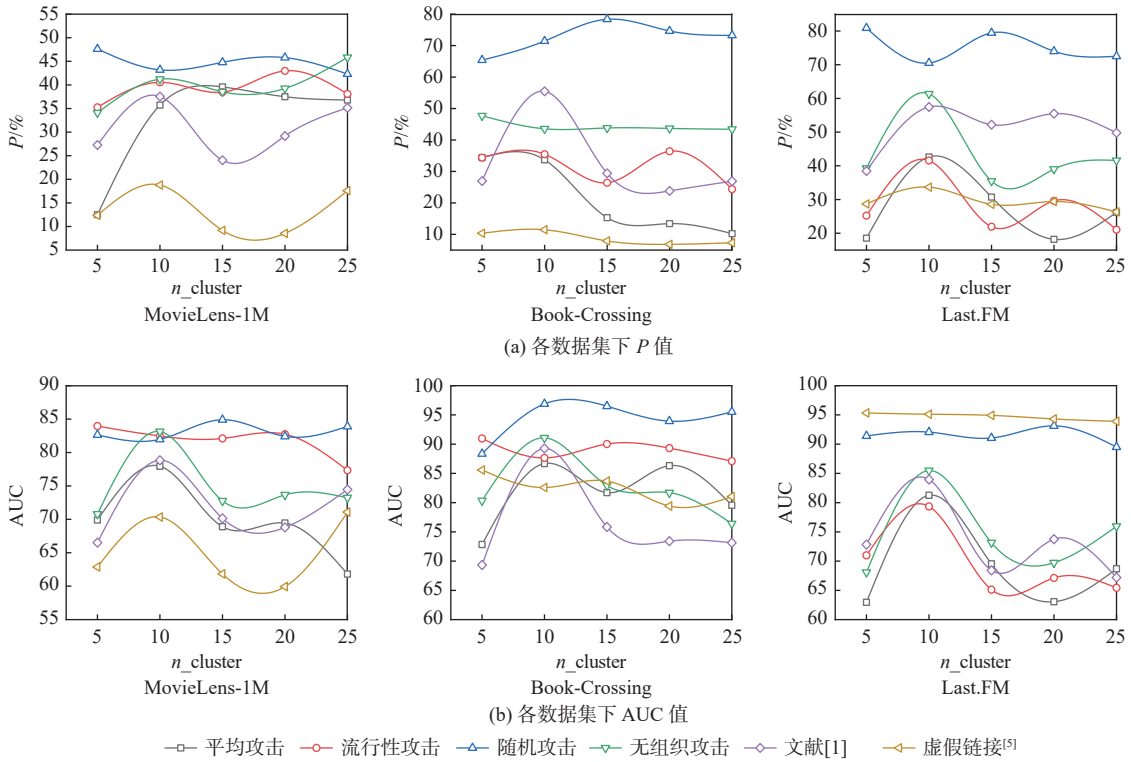


图 7　RippleNet 中不同群体划分的结果

Fig. 7　Results of different group divisions in RippleNet



(a) 各数据集下 P 值

(b) 各数据集下 AUC 值

图 8 MKR 中不同群体划分的结果

Fig. 8 Results of different group divisions in MKR

RippleNet、MKR 中群体划分对模型识别的影响。通过曲线变化可知，在面对不同的数据投毒攻击时，NGL 选择的划分策略不同，对诱导信息的识别结果波动变化，但对用户群体只划为 5 类时，模型对各攻击的识别结果普遍较低，原因在于群体过少可能会造成群体间偏好差异较大，一些小群体间偏好的信息可能会被误判为诱导信息，对模型的识别造成干扰；而群体数量过多会造成群体数量不平衡，可能存在一些小群体，同时由于小群体中用户数量较少，容易将正常的多样化推荐信息误判为诱导信息，同样影响模型的性能。以 RippleNet 为例，对各投毒攻击下目标诱导信息的识别中，总体而言 MovieLens-1M、Last.FM 数据集在群体数为 15 时模型性能最好，Book-Crossing 数据集则在 10 个群体时模型表现最佳，可能是因为 Book-Crossing 中用户基数大、用户兴趣分布较为平均。因此，模型群体划分的策略主要受个性化推荐系统和推荐数据集的影响。虽然在特定情形下，模型对不同数据投毒攻击所采取的最佳划分策略不完全相同，但是在同一情形下，模型同样可尝试从其他数据投毒攻击中学习到的划分策略应用于未知攻击的识别。

4 结 论

本文针对当前个性化推荐算法易受异常信息诱导攻击的情况，以基于知识图谱的个性化推荐算法为基础，研究知识图谱信息异常检测和基于用户兴趣与群体感知的诱导信息检测方法，提出诱导信息识别模型 NGL，从知识图谱节点信息的异常检测和推荐结果对用户产生影响 2 个方面来进行信息识别。实验结果表明，相较于当前主流的异常检测模型，NGL 在基于知识图谱的个性化推荐算法中效果更好、泛化能力更强。但是 NGL 模型仍存在一定的不足：在异常诱导信息检测方面，对于在用户群体间形成传播诱导的异常信息识别程度高，对还尚未形成传播的信息识别能力较低。在未来的工作中可以考虑进一步融合对恶意用户的识别并且增加上线前对推荐内容的识别，多维度加强模型的诱导信息识别能力。

参考文献（References）

[1] ZHANG H T, TIAN C X, LI Y L, et al. Data poisoning attack against recommender system using incomplete and perturbed data[C]// Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining. New York: ACM, 2021: 2154-2164.

[2] ZHANG H T, ZHENG T H, GAO J, et al. Data poisoning attack against knowledge graph embedding[EB/OL]. (2019-06-24)[2023-05-04]. https://doi.org/10.48550/arXiv.1904.12052.

[3] BHARDWAJ P, KELLEHER J, COSTABELLO L, et al. Poisoning knowledge graph embeddings via relation inference patterns [EB/OL]. (2021-11-11)[2023-05-04]. https://doi.org/48550/arXiv.

2111.06345.

[4] BHARDWAJ P, KELLEHER J, COSTABELLO L, et al. Adversarial attacks on knowledge graph embeddings via instance attribution methods[EB/OL]. (2021-11-04)[2023-05-04]. https://doi.org/10.48550/arXiv.2111.03120.

[5] WU Z W, CHEN C T, HUANG S H. Poisoning attacks against knowledge graph-based recommendation systems using deep reinforcement learning[J]. Neural Computing and Applications, 2022, 34(4): 3097-3115.

[6] AKTUKMAK M, YILMAZ Y, UYSAL I. Quick and accurate attack detection in recommender systems through user attributes[C]// Proceedings of the 13th ACM Conference on Recommender Systems. New York: ACM, 2019: 348-352.

[7] CAI H Y, ZHANG F Z. Detecting shilling attacks in recommender systems based on analysis of user rating behavior[J]. Knowledge-Based Systems, 2019, 177: 22-43.

[8] ZHANG S J, YIN H Z, CHEN T, et al. GCN-based user representation learning for unifying robust recommendation and fraudster detection[C]// Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM, 2020: 689-698.

[9] YANG L, NIU X X. A genre trust model for defending shilling attacks in recommender systems[J]. Complex & Intelligent Systems, 2023, 9(3): 2929-2942.

[10] CHEN X, DENG X, HUANG C S, et al. Detection of trust shilling attacks in recommender systems[J]. IEICE Transactions on Information and Systems, 2022, E105.D(6): 1239-1242.

[11] ZHANG F Z, ZHANG Z N, ZHANG P, et al. UD-HMM: an unsupervised method for shilling attack detection based on hidden Markov model and hierarchical clustering[J]. Knowledge-Based Systems, 2018, 148: 146-166.

[12] ZHANG F Z, QU Y Q, XU Y S, et al. Graph embedding-based approach for detecting group shilling attacks in collaborative recommender systems[J]. Knowledge-Based Systems, 2020, 199: 105984.

[13] HAO Y J, MENG G Y, WANG J, et al. A detection method for hybrid attacks in recommender systems[J]. Information Systems, 2023, 114: 102154.

[14] JENY J R V, SOWMYA R, KIRAN G S, et al. Shilling attack detection system for online recommenders[C]//Proceedings of the 2022 International Conference on Inventive Computation Technologies. Piscataway: IEEE Press, 2022: 988-992.

[15] GROVER A, LESKOVEC J. node2vec: scalable feature learning for networks[EB/OL]. (2016-07-03)[2023-06-25]. https://doi.org/10.48550/arXiv.1607.00653.

[16] WANG H, ZHANG F, WANG J, et al. RippleNet: propagating user preferences on the knowledge graph for recommender systems[C]// Proceedings of the 27th ACM Interna-tional Conference on Information and Knowledge Manage-ment. New York: ACM, 2018: 417-426.

[17] ZHANG C J, ZENG A. Behavior patterns of online users and the effect on information filtering[J]. Physica A: Statistical Mechanics and its Applications, 2012, 391(4): 1822-1830.

[18] AZMEDROUB B, OUARZEDDINE M. Polarimetric SAR images clustering with Gaussian mixtures model[C]//Proceedings of the 2015 3rd International Conference on Control, Engineering & Information Technology. Piscataway: IEEE Press, 2015: 1-5.

[19] GOODGE A, HOOI B, NG S K, et al. LUNAR: unifying local outlier detection methods via graph neural networks[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2022, 36(6): 6737-6745.

[20] WANG H W, ZHANG F Z, ZHAO M, et al. Multi-task feature learning for knowledge graph enhanced recommendation[C]//Proceedings of the The World Wide Web Conference. New York: ACM, 2019: 2000-2010.

[21] KRIEGEL H P, SCHUBERT M, ZIMEK A. Angle-based outlier detection in high-dimensional data[C]// Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM, 2008: 444-452.

[22] RAMASWAMY S, RASTOGI R, SHIM K. Efficient algorithms for mining outliers from large data sets[C]// Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data. New York: ACM, 2000: 427-438.

[23] BANDARAGODA T R, TING K M, ALBRECHT D, et al. Isolation-based anomaly detection using nearest-neighbor ensembles[J]. Computational Intelligence, 2018, 34(4): 968-998.

[24] LIU F T, TING K M, ZHOU Z H. Isolation forest[C]//Processings of the 2008 Eighth IEEE International Conference on Data Mining. Piscataway: IEEE Press, 2008: 413-422.

[25] HARDIN J, ROCKE D M. Outlier detection in the multiple cluster setting using the minimum covariance determinant estimator[J]. Computational Statistics & Data Analysis, 2004, 44(4): 625-638.

[26] LI Z, ZHAO Y, HU X Y, et al. ECOD: unsupervised outlier detection using empirical cumulative distribution functions[J]. IEEE Transactions on Knowledge and Data Engineering, 2022, 35(12): 12181-12193.

[27] ALMARDENY Y, BOUJNAH N, CLEARY F. A novel outlier detection method for multivariate data[J]. IEEE Transactions on Knowledge and Data Engineering, 2022, 34(9): 4052-4062.

[28] LI Z, ZHAO Y, BOTTA N, et al. COPOD: copula-based outlier detection[C]//Proceedings of the 2020 IEEE International Conference on Data Mining. Piscataway: IEEE Press, 2020: 1118-1123.

[29] GOLDSTEIN M, DENGEL A. Histogram-based outlier score (HBOS): a fast unsupervised anomaly detection algorithm[C]//Proceedings of the KI-2012 35th Genman Conference on Artifical Intelligence. Saarbrücken: Springer, 2012.

[30] ZENATI H, ROMAIN M, FOO C S, et al. Adversarially learned anomaly detection[C]//Proceedings of the 2018 IEEE International Conference on Data Mining. Piscataway: IEEE Press, 2018: 727-736.

Identification of induced information for personalized recommendations based on knowledge graph

NI Wenkai，PENG Shufan，DU Yanhui*

(College of Information and Cyber Security，People's Public Security University of China，Beijing 102600，China)

Abstract：In the age of intelligent information, managing the recommendations of Internet information service algorithms is a crucial step in creating a national Internet governance framework. Personalized recommendation algorithm is one of the important technologies for Internet information service algorithm recommendation. The knowledge graph is widely used in personalized recommendation algorithms. At the same time, the knowledge graph and recommendation algorithm are vulnerable to data poisoning attacks by attackers, which in turn affects the recommendation results and induces information dissemination. There is a lack of effective models for identifying this type of induced information. Based on this, this article conducts research on the induced information identification model. Based on the analysis of user historical behavior records and the evolution process of user preferences, we study induction based on user interests and group perception. Information detection method, perform group preference modeling on the historical preferences of similar user groups, perform outlier analysis on abnormally exposed information within groups with common characteristics, and construct set node2vec-side item representation, GMM group division, and LUNAR induction for anomaly detection User preference modifications and recommendation result evolution reasoning are the basis for the realization of induced information recognition by the information recognition model NGL (Induced information detection model that incorporates node2vec-side item representation, GMM group division, and LUNAR, NGL). Induced information recognition experiments were conducted on RippleNet and MKR recommendation systems. The results show that the NGL model proposed in this article is better than the existing anomaly detection model.

Keywords：induced propagation；knowledge graphs；anomaly detection；group division；personalized recommendations

Received：2023-07-19；**Accepted**：2023-09-12；**Published Online**：2023-09-22 09：09

URL：link.cnki.net/urlid/11.2625.V.20230921.0909.001

Foundation item：People's Public Security University of China Cyberspace Security law enforcement technology double first-class project (2023SYL07)

* **Corresponding author**. E-mail：duyanhui@ppsuc.edu.cn

http://bhxb.buaa.edu.cn jbuaa@buaa.edu.cn
DOI: 10.13700/j.bh.1001-5965.2023.0511

基于人眼视觉机制的伪装目标检测网络

张冬冬，王春平，付强*
(陆军工程大学石家庄校区，石家庄 050003)

摘 要：伪装目标检测是一项新兴的视觉检测任务，旨在识别出完美隐藏在周围环境中的伪装目标，在多个领域中具有广泛应用。针对当前伪装目标检测算法无法准确、完整地识别目标结构和边界的问题，基于人类在观察伪装图像时的视觉感知过程，设计了一种生物启发式框架，并命名为定位和细化网络（PRNet）。利用Res2Net提取图像的原始特征，从多层级信息中挖掘目标的边缘线索；特别设计特征增强模块，在丰富全局上下文信息的同时能够扩大感受野；定位模块利用双注意力机制从通道和空间2个维度来定位目标的大致位置；细化模块同时关注前景和背景中的目标线索，利用多类型信息进一步细化目标的结构和边缘。在3个广泛使用的伪装目标检测基准数据集上的大量实验结果表明，所提网络的整体性能明显优于14种比较算法，在多种复杂场景中表现优异。

关 键 词：伪装目标检测；卷积神经网络；视觉感知；特征增强；注意力机制
中图分类号：TP753
文献标志码：A **文章编号**：1001-5965(2025)07-2553-09

伪装目标检测(camouflaged object detection, COD)旨在检测隐藏在背景中的伪装目标，近年来在计算机视觉领域引起了越来越多的关注。得益于发现并找出潜在伪装对象的能力，COD技术被广泛应用于多种领域中，包括农业领域(如蝗虫预警[1]、果实熟度检测[2])、工业领域(如缺陷检测[3])、医学领域(如息肉分割[4]、肺部感染分割[5])和艺术领域(如艺术拼贴[6]、人体彩绘[7])。

传统的COD方法[8-10]主要依赖手工设计的低级特征(如颜色、纹理和梯度)来描述目标，以区分图像中的伪装目标和背景。然而，这些方法对伪装目标的表征能力有限，当场景非常复杂或伪装目标特征与背景极为相似时，通常会失效。近年来，随着卷积神经网络(convolutional neural network, CNN)的发展成熟，研究人员提出了多种基于CNN的COD算法。Fan等[11]开发了一个简单且高效的伪装目标搜索识别网络(search identification network, SINet)，该网络主要由搜索和识别2个模块组成：搜索模块通过扩大感受野来获取更多具有鉴别性的特征，从而可以从复杂环境中搜索到目标；识别模块则是利用注意力机制来消除不相关特征的干扰，进而能够准确判别目标类型。但是该网络在目标搜索过程中引入了大量特征，可能会造成特征冗余，并对识别过程产生一定影响。因此，Mei等[12]构建了由定位模块和聚焦模块组成的COD网络PFNet，该网络从全局角度定位潜在目标，通过聚焦于模糊区域来逐步完善粗略预测。Fan等[13]在SINet的基础上提出了SINetV2模型，将COD划分为搜索和识别2个阶段，在搜索阶段使用纹理增强模块来捕获具有放大上下文提示的细粒度纹理特征，在识别阶段构建了级联的群体翻转注意力模块，通过不同的特征金字塔逐步完善粗略预测。Pang等[14]设计了一种混合比例的三重网络ZoomNet，该网络使用三重结构提取不同“缩放”比例的差异化伪装

收稿日期：2023-08-04；**录用日期**：2023-10-05；**网络出版时间**：2023-11-01 08:43
网络出版地址：link.cnki.net/urlid/11.2625.V.20231031.1708.001
*通信作者. E-mail：1418748495@qq.com

引用格式：张冬冬，王春平，付强. 基于人眼视觉机制的伪装目标检测网络[J]. 北京航空航天大学学报，2025，51(7)：2553-2561.
ZHANG D D，WANG C P，FU Q. Camouflaged object detection network based on human visual mechanisms[J]. Journal of Beijing University of Aeronautics and Astronautics，2025，51(7)：2553-2561 (in Chinese).

目标特征，然后利用尺度集成单元来筛选和聚合特定比例的特征，并利用分层混合尺度单元进一步重组和增强混合比例特征。Ren 等[15]通过放大伪装目标与周围环境之间的纹理差异性来提升模型从背景中分离伪装目标的能力。Zhou 等[16]通过构建边界制导模块对目标边界特征进行显示建模，增强边界的特征表达，从而提升模型性能。尽管上述方法在 COD 任务中展现出了强大的检测能力，但由于伪装目标外观与周围环境高度相似，并且在边界上模糊不清，现有方法往往难以有效和完全识别伪装目标。此外，在有额外挑战存在的情况下，例如复杂背景、多尺度、遮挡等，现有方法无法完整地识别目标的结构和边缘。

生物视觉研究[8]表明，为了找出当前视觉场景中的伪装对象，观察者首先会形成潜在的区域建议，利用周围线索排除不合理区域定位伪装对象的粗略位置；然后，观察者会将注意力聚集在这些粗略区域，并借助边缘信息来刻画目标整体轮廓；最后，成功从背景中分离出伪装对象。针对上述问题，受人类观察伪装图像时的视觉行为启发，本文提出了一种基于人眼视觉机制的伪装目标检测网络（positioning refinement network, PRNet），设计特征增强模块（feature enhancement module, FEM）以提升特征多样性以及网络的特征表达能力，提高网络对复杂场景的应对能力；设计定位模块（positioning module, PM）以模仿人类视觉行为中的定位过程，借助通道注意力和空间注意力从全局角度定位不同尺度的伪装目标；设计细化模块（refinement module, RM）以模仿视觉行为中的细化过程，在融合边缘先验、区域先验、领域先验以及语义先验的基础上，深入挖掘前景和背景中的目标线索，进一步完善目标的结构和边缘，同时规避遮挡对检测的影响，进而实现伪装目标的精确检测。

1 算法设计

所提 PRNet 的整体框架如图 1 所示，网络具体步骤为：①特征提取，Res2Net-50[17]作为主干网络，从输入图像中提取多级特征，即 $f_i(i=1,2,\cdots,5)$；②特征增强，FEM 对主干网络的输出特征进行增强，在丰富上下文信息的同时，通过扩张感受野提升特征的多样性；③粗略定位，PM 从全局角度挖掘目标的位置信息，粗略定位目标区域，并形成区域先验；④捕获边缘先验，使用边缘感知模块（edge-aware module, EAM）来捕获与目标相关的边界语义，进而获得边缘先验，EAM 源自文献[18]，并且所有参数设置与原文献相同；⑤细化目标区域，RM 从前景和背景 2 个角度探索目标线索，并通过综合多种先验知识进一步完善目标边界和区域。

1.1 特征增强模块

COD 所面临的最大困难是伪装目标外观的变化，比如尺度变化、遮挡和边界模糊。针对上述问题，上下文信息可以提供丰富的线索，有助于区分

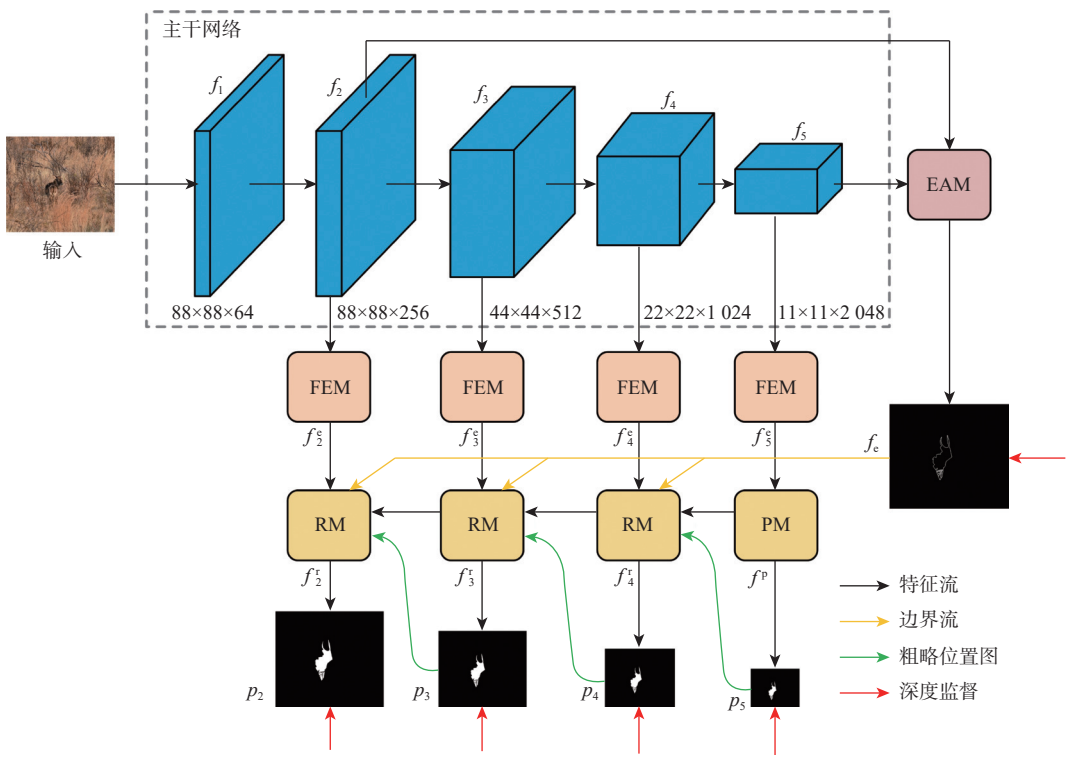


图 1 PRNet 整体框架

Fig. 1 Overall framework of PRNet

伪装目标和背景, 利于定位伪装区域。许多先前的研究工作, 例如 SINet[11] 和 D2C-Net[19], 已经证明增加感受野可以使提取的特征更加多样, 能够提升网络的检测性能。因此, 本文构建了 FEM, 该模块在扩张感受野的同时能够挖掘上下文信息, 为应对伪装目标外观变化提供可行方案。FEM 的结构如图 2 所示, 其由 2 个分支组成, 即全局上下文分支和感受野分支。这 2 个分支并行运行, 2 个分支的输出特征与输入特征通过像素相加的方式进行融合。

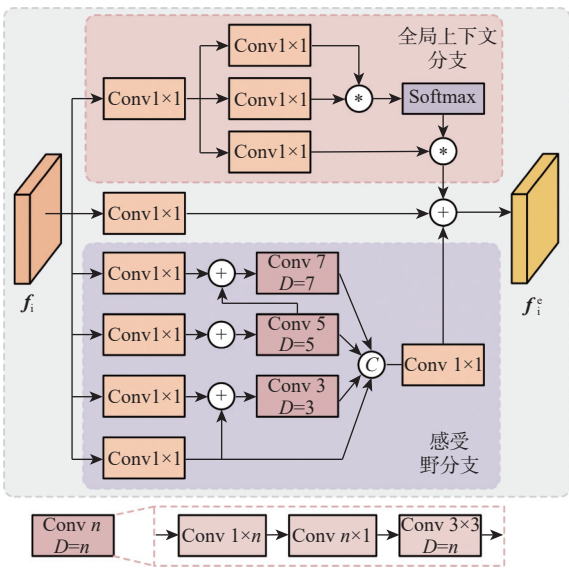


图 2　特征增强模块

Fig. 2　Feature enhancement module

在全局上下文分支中, 输入特征 $\boldsymbol{f}_i \in \mathbf{R}^{C\times H\times W}(i=2,3,4,5)$首先经过 1×1 的卷积进行通道缩减(变为 $C/4$), 紧接着通过 3 个 1×1 的卷积生成关键特征 $\boldsymbol{K} \in \mathbf{R}^{\frac{C}{16}\times H\times W}$、查询特征 $\boldsymbol{Q} \in \mathbf{R}^{\frac{C}{16}\times H\times W}$和值特征 $\boldsymbol{V} \in \mathbf{R}^{\frac{C}{16}\times H\times W}$。然后, $\boldsymbol{Q}$的转置和$\boldsymbol{K}$之间进行矩阵乘法运算, 随后利用 Softmax 函数来生成空间注意力图 $\boldsymbol{S} \in \mathbf{R}^{(H\times W)\times(H\times W)}$。最后, 通过矩阵乘法运算将 $\boldsymbol{V}$ 与 $\boldsymbol{S}$ 进行组合来获得输出特征。全局上下文分支的处理过程可表示为

$$\begin{cases} \boldsymbol{f}_i' = f^{1\times1}(\boldsymbol{f}_i) \\ \boldsymbol{Q} = f^{1\times1}(\boldsymbol{f}_i'), \boldsymbol{K} = f^{1\times1}(\boldsymbol{f}_i'), \boldsymbol{V} = f^{1\times1}(\boldsymbol{f}_i') \\ \boldsymbol{S} = \mathrm{Softmax}(\boldsymbol{Q}^{\mathrm{T}} \odot \boldsymbol{K}) \\ \boldsymbol{f}_i^{\mathrm{g}} = \boldsymbol{V} \odot \boldsymbol{S} \end{cases} \tag{1}$$

式中: $f^{1\times1}$表示 1×1 卷积; $\odot$表示矩阵乘法运算; $\boldsymbol{f}_i^{\mathrm{g}}$表示输入特征$\boldsymbol{f}_i(i=2,3,4,5)$经过全局上下文分支所生成的输出特征。

在感受野分支中, 共有 4 个支路, 每个支路均包含用于减少通道数的 1×1 卷积, 用于局部特征感知的级联卷积(1×n 和 n×1), 以及用于上下文特征提取的 3×3 膨胀卷积, 膨胀率为 n。上一个支路的输出将会被馈送到下一个支路中, 以在更大的感受野中做进一步处理。最后, 所有支路的输出按照通道相加的方式进行融合, 并通过 1×1 的卷积进行通道恢复。感受野分支的处理过程可表示为

$$\begin{cases} \boldsymbol{K}_1 = f^{1\times1}(\boldsymbol{f}_i) \\ \boldsymbol{K}_2 = f_3^{3\times3}(f^{3\times1}(f^{1\times3}(f^{1\times1}(\boldsymbol{f}_i)+\boldsymbol{K}_1))) \\ \boldsymbol{K}_3 = f_5^{3\times3}(f^{5\times1}(f^{1\times5}(f^{1\times1}(\boldsymbol{f}_i)+\boldsymbol{K}_2))) \\ \boldsymbol{K}_4 = f_7^{3\times3}(f^{7\times1}(f^{1\times7}(f^{1\times1}(\boldsymbol{f}_i)+\boldsymbol{K}_3))) \\ \boldsymbol{f}_i^{\mathrm{r}} = f^{1\times1}(\boldsymbol{K}_1 \oplus \boldsymbol{K}_2 \oplus \boldsymbol{K}_3 \oplus \boldsymbol{K}_4) \end{cases} \tag{2}$$

式中: $\boldsymbol{K}_i(i=1,2,3,4)$表示不同支路的输出; $f_n^{3\times3}$表示膨胀率为 n 的 3×3 膨胀卷积; $\oplus$表示通道相加; $\boldsymbol{f}_i^{\mathrm{r}}$表示输入特征$\boldsymbol{f}_i(i=2,3,4,5)$经过感受野分支所生成的输出特征。

将输入特征与全局上下文分支和感受野分支的输出按照像素相加的方式进行融合, 进而得到包含丰富上下文信息和多样化的增强型特征。特征增强模块的输出$\boldsymbol{f}_i^{\mathrm{e}}(i=2,3,4,5)$可表示为

$$\boldsymbol{f}_i^{\mathrm{e}} = \boldsymbol{f}_i^{\mathrm{g}} + \boldsymbol{f}_i^{\mathrm{r}} + f^{1\times1}(\boldsymbol{f}_i) \tag{3}$$

1.2　定位模块

高层特征包含丰富的语义信息, 这是定位目标的关键。对于 COD 任务, 伪装目标通常隐藏在背景中或与背景混合在一起, 这极大地增加了定位目标的难度, 并且这一特性往往会导致伪装目标定位不准确或检测结果不完整。针对上述问题, 本文提出了 PM, 该模块旨在利用高层语义信息, 从全局角度定位伪装目标。PM 的结构如图 3 所示, 其主要由通道注意力分支和空间注意力分支组成, 2 个分支能够并行运行, 可以提高运算效率。2 种注意力均以非本地的方式实现, 以捕获通道和空间的长期依赖关系, 从而从全局角度增强最高层特征的语义表达能力, 使 PM 能够有效地推断伪装目标的初始位置。

PM 借助高层语义信息来实现伪装目标的初步定位, 其输入特征是 EFM 的输出特征 $\boldsymbol{f}_5^{\mathrm{g}}$。在给定

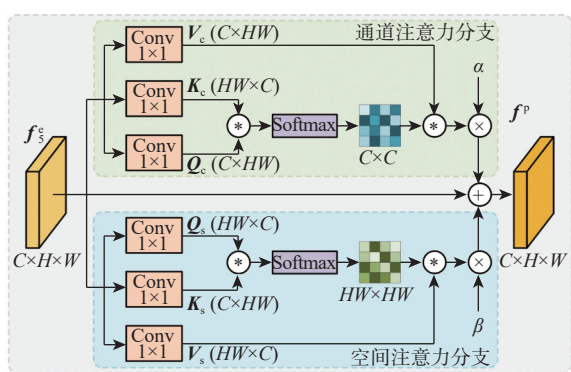


图 3　定位模块

Fig. 3　Positioning module

输入$\boldsymbol{f}_5^{\mathrm{g}} \in \mathbf{R}^{C\times H\times W}$的情况下，对于通道注意力分支，输入特征通过1×1的卷积和整形操作生成查询特征$\boldsymbol{Q}_{\mathrm{c}} \in \mathbf{R}^{C\times HW}$、关键特征$\boldsymbol{K}_{\mathrm{c}} \in \mathbf{R}^{HW\times C}$和值特征$\boldsymbol{K}_{\mathrm{c}} \in \mathbf{R}^{C\times HW}$。对于空间注意力分支，同样可以获得对应的查询特征$\boldsymbol{Q}_{\mathrm{s}} \in \mathbf{R}^{HW\times C}$、关键特征$\boldsymbol{K}_{\mathrm{s}} \in \mathbf{R}^{C\times HW}$和值特征$\boldsymbol{V}_{\mathrm{s}} \in \mathbf{R}^{HW\times C}$。2个分支中的查询特征和关键特征分别执行矩阵乘法操作，并使用Softmax归一化分别生成通道注意力$\boldsymbol{\omega}_{\mathrm{c}} \in \mathbf{R}^{C\times C}$和空间注意力$\boldsymbol{\omega}_{\mathrm{s}} \in \mathbf{R}^{HW\times HW}$：

$$\begin{cases} \boldsymbol{\omega}_{\mathrm{c}} = \mathrm{Softmax}(\boldsymbol{Q}_{\mathrm{c}} \odot \boldsymbol{K}_{\mathrm{c}}) \\ \boldsymbol{\omega}_{\mathrm{s}} = \mathrm{Softmax}(\boldsymbol{Q}_{\mathrm{s}} \odot \boldsymbol{K}_{\mathrm{s}}) \end{cases} \tag{4}$$

然后，通道注意力和空间注意力分别与对应的值特征进行矩阵相乘，并将汇聚的注意力特征整形为$\mathbf{R}^{C\times H\times W}$。为增强容错能力，本文在2个分支中分别引入了可学习参数α和β。最后，2个分支的输出与输入特征之间进行融合获得最终输出$\boldsymbol{f}^{\mathrm{p}}$：

$$\boldsymbol{f}^{\mathrm{p}} = \boldsymbol{f}_5^{\mathrm{g}} + \alpha \otimes (\boldsymbol{\omega}_{\mathrm{c}} \odot \boldsymbol{V}_{\mathrm{c}}) + \beta \otimes (\boldsymbol{\omega}_{\mathrm{s}} \odot \boldsymbol{V}_{\mathrm{s}}) \tag{5}$$

式中：⊗表示逐元素相乘。

1.3 细化模块

在COD任务中，由于伪装对象隐藏在周围环境中，与背景之间的边界模糊不清，使得获取边界准确的伪装检测结果非常困难。在实践中，已有的一些COD算法[18,20-21]已经证明了边界信息在COD检测任务中至关重要，其可以提供有用的约束来指导网络进行特征提取。同时，由于伪装目标的外观与背景非常相似，使初步预测中常常出现假阳性和假阴性的结果。受先前工作[12,22-23]的启发，在细化检测的过程中，不仅要关注前景中的线索，还要挖掘背景中的线索。另外，不同层的特征往往包含不同的语义信息，有效地融合不同层特征可以提升网络的特征表征能力，势必会提升模型的性能。综上分析，本文设计了FM，该模块在综合利用多种先验知识的同时，从前景和背景2个方向来探索目标线索，进一步细化目标边界和区域，其结构如图4所示。在给定输入$\boldsymbol{f}_i^{\mathrm{e}}, \boldsymbol{f}_{i+1}^{\mathrm{r}}, \boldsymbol{p}_{i+1}(i=2,3,4)$和$\boldsymbol{f}_{\mathrm{e}}$的情况下，FM的工作过程可以分为两步。需要注意的是，当时$i=4$，$\boldsymbol{f}_5^{\mathrm{r}}$即为$\boldsymbol{f}^{\mathrm{p}}$。

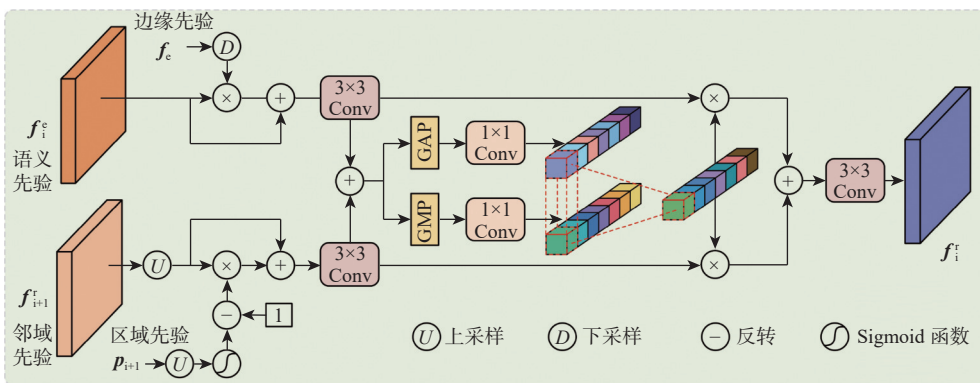


图4 细化模块

Fig. 4 Refinement module

第①步，将边缘先验$\boldsymbol{f}_{\mathrm{e}}$注入输入特征$\boldsymbol{f}_i^{\mathrm{e}}$中，从前景角度出发获得边缘增强特征$\boldsymbol{f}_i^{\mathrm{ee}}$；将区域先验$\boldsymbol{p}_{i+1}$注入输入特征$\boldsymbol{f}_{i+1}^{\mathrm{r}}$中，从背景角度出发获得区域增强特征$\boldsymbol{f}_{i+1}^{\mathrm{re}}$。

$$\begin{cases} \boldsymbol{f}_i^{\mathrm{ee}} = f^{3\times3}(\boldsymbol{f}_i^{\mathrm{e}} \otimes \boldsymbol{D}(\boldsymbol{f}_{\mathrm{e}})) \oplus \boldsymbol{f}_i^{\mathrm{e}}) \\ \boldsymbol{f}_{i+1}^{\mathrm{re}} = f^{3\times3}((U(\boldsymbol{f}_{i+1}^{\mathrm{r}}) \otimes (1-\sigma(U(\boldsymbol{p}_{i+1})))) \oplus (U(\boldsymbol{f}_{i+1}^{\mathrm{r}})) \end{cases} \tag{6}$$

式中：D表示下采样；U表示上采样；σ表示Sigmoid函数。

第②步，首先利用逐像素相加的方式融合边缘增强特征和区域增强特征：

$$\boldsymbol{f}_i^{\mathrm{a}} = \boldsymbol{f}_i^{\mathrm{ee}} \oplus \boldsymbol{f}_{i+1}^{\mathrm{re}} \tag{7}$$

然后，$\boldsymbol{f}_i^{\mathrm{a}}$被馈送到一个双分支结构的注意力引导模块中，以获得注意力引导权重。具体来说，$\boldsymbol{f}_i^{\mathrm{a}}$在2个分支分别进行全局平均池化和全局最大池化操作，而后经过1×1的卷积获得相应的注意力权重，随后通过逐元素相加将2个注意力权重结合起来，并通过Sigmoid函数来获得最终的注意力引导权重。整个过程可以用下式表示：

$$\omega = \sigma(f_k^{1\times1}(\mathrm{GAP}(\boldsymbol{f}_i^{\mathrm{a}})) \oplus f_k^{1\times1}(\mathrm{GMP}(\boldsymbol{f}_i^{\mathrm{a}}))) \tag{8}$$

式中：GAP(·)表示全局平均池化操作；GMP(·)表示全局最大池化操作；$f_k^{1\times1}$表示卷积核大小为k的1×1卷积，卷积核大小k可以根据的$\boldsymbol{f}_i^{\mathrm{c}}$通道数$C$进行自主设置，其计算公式如下：

$$k = \left\lfloor (1+\log_2 C)/2 \right\rfloor \tag{9}$$

式中：⌊·⌋表示向下取整操作。

最后，注意力引导权重ω分别与边界增强特征$\boldsymbol{f}_i^{\mathrm{ee}}$和区域增强特征$\boldsymbol{f}_{i+1}^{\mathrm{re}}$相乘，并通过逐元素相加的方式进行整合，以获得目标线索丰富的融合特征。

而后, 经过 3×3 的卷积获得最终输出, 即 $\boldsymbol{f}_i^{\mathrm{r}}$。

$$\boldsymbol{f}_i^{\mathrm{r}} = f^{3\times3}((\boldsymbol{f}_i^{\mathrm{ee}} \otimes \omega) \oplus (\boldsymbol{f}_{i+1}^{\mathrm{re}} \otimes (1-\omega))) \tag{10}$$

1.4　损失函数

PRNet 中有 5 个输出, 即 4 个伪装目标掩膜输出和 1 个伪装目标边缘输出。对于伪装目标掩膜输出, 对其施加加权二进制交叉熵损失($L_{\mathrm{BCE}}^{\omega}$)和加权交并比(intersection over union, IoU)损失($L_{\mathrm{IOU}}^{\omega}$)[24], 以更加关注硬像素并突出较硬像素。对于伪装目标边缘输出, 对其施加 Dice 损失(L_{Dice})[25], 以应对正负样本之间的强烈失衡。因此, 总损失函数为

$$L_{\mathrm{total}} = \sum_{i=2}^{5}(L_{\mathrm{BCE}}^{\omega}(\boldsymbol{P}_i, \boldsymbol{G}_{\mathrm{o}}) + L_{\mathrm{IOU}}^{\omega}(\boldsymbol{P}_i, \boldsymbol{G}_{\mathrm{o}})) + \lambda L_{\mathrm{Dice}}(\boldsymbol{P}_{\mathrm{e}}, \boldsymbol{G}_{\mathrm{e}}) \tag{11}$$

式中: $\boldsymbol{P}_i$和$\boldsymbol{P}_{\mathrm{e}}$分别表示伪装对象掩膜预测和边缘预测; $\boldsymbol{G}_{\mathrm{o}}$和$\boldsymbol{G}_{\mathrm{e}}$分别表示伪装对象掩膜真值和边缘真值; λ是平衡参数, 在实验中设置为 4。

2　实　验

所有实验均在配置为 E5-2630v4 CPU、RTX 3 090 GPU(24 GB video memory)、64 GB RMB 的图像工作站上进行, 以深度学习框架 Pytorch 为编译工具完成。在训练阶段, 本文将所有输入图像的大小调整为 352×352, 使用在 ImageNet 上预训练的 Res2Net-50 权重初始化主干网络, 其他层随机初始化。此外, 使用 Adam 优化器, 学习率为 1e-4, 总训练轮数为 50, 批量大小设置为 16。在测试阶段, 图像大小被调整为 352×352 后输入网络中, 无需任何后处理即可得到预测结果。

2.1　数据集

当前, COD 领域应用最为广泛的基准数据集有: CAMO[26]、COD10K[11] 和 NC4K[27]。CAMO 包含 1 250 张图像, 其中 1 000 张用于训练, 250 张用于测试。COD10K 是目前最大且具有高质量像素级注释的伪装对象数据集, 共包含 5 066 张伪装图像, 其中 3 040 张用于训练, 2 026 张用于测试。NC4K 是当前数据量最大的伪装目标测试集, 由 4 121 张从互联网上下载的图像组成。遵循以前的工作[2], 本文使用 CAMO 和 COD10K 的训练集作为训练集, 使用其测试集和 NC4K 作为测试集。

2.2　评价指标

为了全面比较本文模型和其他方法, 本文利用 5 个广泛使用的指标来评估 COD 的性能。PR 曲线[16]度量模型的查全率和查准率。F-measure(F_β)[28] 综合度量查全率和查准率。S-measure(S_α)[29] 评估预测图和真实图之间的结构相似性。E-measure(E_ϕ)[30] 基于人类视觉感知机制评估 COD 的整体和局部精度。平均绝对误差(MAE, M)[31] 评估预测图与地面实况之间的像素级误差。

2.3　实验结果及分析

2.3.1　消融实验

为了评估本文所提模块对提升网络检测能力的贡献度, 以 Res2Net-50 和 4 个检测头构成的模型作为模型(Baseline), 然后分别增加 FEM、PM 和 RM 构成新的模型, 将加入单个模块的模型预测结果与 PRNet 进行比较。值得注意的是, 在增加 RM 时, 用于边缘检测的 EAM 模块也会一并引入。消融实验的定量结果如表 1 所示。

表 1　消融实验定量结果(“↑”/“↓”分别表示值越大/小越好)

Table 1　Quantitative results of ablation experiments(“↑”/“↓” indicates that larger/smaller is better.)

模型	CAMO				COD10K				NC4K			
	S↑	E↑	F↑	M↓	S↑	E↑	F↑	M↓	S↑	E↑	F↑	M↓
Baseline	0.533	0.523	0.304	0.197	0.584	0.583	0.307	0.105	0.585	0.572	0.367	0.153
Baseline+FEM	0.758	0.81	0.704	0.087	0.781	0.849	0.690	0.041	0.812	0.868	0.767	0.058
Baseline+PM	0.639	0.664	0.507	0.145	0.666	0.709	0.481	0.078	0.69	0.719	0.575	0.109
Baseline+RM	0.778	0.834	0.741	0.085	0.802	0.871	0.709	0.037	0.830	0.885	0.794	0.052
PRNet	0.837	0.893	0.799	0.051	0.834	0.904	0.756	0.032	0.875	0.897	0.820	0.041

从表 1 可以看出, 综合多个模块与使用单个模块相比, 在 3 个数据集上的 4 个指标得到了大幅提升; 使用单个模块的模型相较于基线模型性能均有不同程度的提升, 值得注意的是, 增加 RM 模块使得模型性能得到了极大的提高, 相较于 Baseline, 在 3 个数据集上平均将 4 个度量(S_α、E_ϕ、F_β、M)分别提高了 4.5%、3.5%、4.4% 和 1.7%。

为直观地说明 EFM、PM 和 RM 的有效性, 图 5 给出了 EFM 的输入和输出、PM 的输出以及 RM 的输出对应的部分特征图。对比图 5 中的第 3 列和第 4 列, 可以清楚地发现, 引入 EFM 使特征的多样性更加丰富, 增强了模型的特征提取和表征能力。对比第 4 列和第 5 列, 可以清楚地发现, 引入 PM 使 PRNet 能够从复杂背景中粗略地定位出伪装区

域。但是，伪装目标的边缘仍然非常模糊，整个区域需要进一步细化。对比第 5 列和第 6 列，可以发现，经过 RM 处理，伪装目标的边缘得到了突显，整个伪装区域从视觉上能够被辨识，这验证了所提出的 RM 在细化伪装目标的边缘和轮廓方面是有效的。

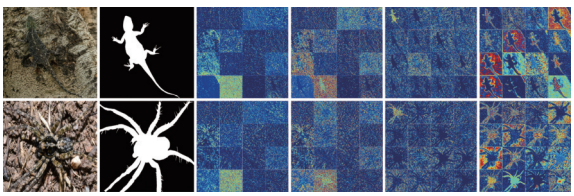


图 5 特征图可视化实例

Fig. 5 Example of feature map visualization

2.3.2 对比实验

为了证明所提模型 PRNet 的有效性，将其与 14 种最先进的方法进行比较，包括 SINetV2[13]、JCSOD[32]、Rank-Net[27]、MGL[33]、PFNet[12]、OCENet[34]、ERRNet[22]、BgNet[23]、ZoomNet[14]、BSANet[20]、C2-FNet[35]、FAPNet[16]、MFFN[36] 和 ASBI[37]。为保证比较的公平性，上述方法的所有预测结果均由开源代码在同一平台上重新训练的模型来生成，并使用相同代码 https://github.com/lartpang/PySODEvalToolkit 进行评估。

表 2 给出了 PRNet 和其他 14 种最先进方法在 4 个基准数据集上的定量结果。从表中可以清楚地看出，PRNet 的性能优于其他方法，在 4 个评价指标上有很大的优势。具体地，与最佳竞争方法 ASBI 相比，PRNet 在 3 个数据集上平均将四个度量（S_α、E_ϕ、F_β、M）分别提高了 1.3%、0.6%、1.1% 和 0.9%。除了使用上述 4 个度量进行整体定量比较外，本文还在图 6 中展示了 PR 曲线和 F-measure 曲线。从图 6 可以看出，本文所提的 PRNet 模型优于其他先进方法。

表 2 不同模型的定量检测结果

Table 2 Quantitative detection results of different models

方法	CAMO				COD10K				NC4K			
	$S\uparrow$	$E\uparrow$	$F\uparrow$	$M\downarrow$	$S\uparrow$	$E\uparrow$	$F\uparrow$	$M\downarrow$	$S\uparrow$	$E\uparrow$	$F\uparrow$	$M\downarrow$
SINetV2[13]	0.815	0.870	0.783	0.074	0.813	0.886	0.713	0.037	0.845	0.901	0.802	0.048
JCSOD[32]	0.767	0.810	0.729	0.086	0.800	0.872	0.718	0.036	0.835	0.886	0.805	0.049
Rank-Net[27]	0.785	0.842	0.736	0.081	0.786	0.863	0.672	0.043	0.823	0.883	0.771	0.054
MGL[33]	0.570	0.499	0.302	0.182	0.635	0.584	0.341	0.111	0.662	0.596	0.428	0.136
PFNet[12]	0.776	0.832	0.738	0.087	0.799	0.875	0.700	0.039	0.828	0.886	0.786	0.053
OCENet[34]	0.775	0.818	0.709	0.092	0.789	0.854	0.680	0.044	0.822	0.871	0.766	0.057
ERRNet[22]	0.767	0.801	0.671	0.104	0.744	0.801	0.578	0.063	0.792	0.833	0.692	0.077
BgNet[23]	0.652	0.684	0.534	0.137	0.657	0.718	0.475	0.082	0.699	0.749	0.594	0.101
ZoomNet[14]	0.797	0.842	0.765	0.076	0.819	0.864	0.742	0.032	0.838	0.878	0.803	0.047
BSANet[20]	0.800	0.852	0.770	0.076	0.812	0.876	0.733	0.035	0.836	0.887	0.803	0.050
C2FNet[35]	0.768	0.823	0.726	0.089	0.808	0.883	0.724	0.036	0.834	0.890	0.798	0.049
FAPNet[16]	0.798	0.849	0.758	0.082	0.820	0.885	0.727	0.036	0.849	0.898	0.805	0.048
MFFN[36]	0.790	0.837	0.752	0.081	0.826	0.872	0.753	0.037	0.837	0.877	0.800	0.052
ASBI[37]	0.821	0.874	0.782	0.073	0.829	0.896	0.742	0.033	0.857	0.907	0.818	0.044
本文	0.837	0.893	0.799	0.051	0.834	0.904	0.756	0.032	0.875	0.897	0.820	0.041

图 7 显示了所提模型和 5 种具有代表性的 COD 模型在 6 种类型伪装目标下的视觉检测结果，从左至右依次为：①多目标、②大目标、③小目标、④遮挡、⑤边缘复杂、⑥“完美”隐藏。从整体检测效果来看，所提方法能更准确、更完整地分割伪装对象。具体来说，从第 1 列可以看出所提模型能够同时检测出多个伪装对象，而其他模型容易漏检。在第 2 列和第 3 列中，伪装目标具有不同尺度，在这种情况下，所提模型表现良好，能够准确定位不同大小的伪装目标，且能够完整地检测出大尺度目标。在第 4 列中，目标被周围环境部分遮挡，而所提模型仍然能成功地检测出伪装目标。在第 5 列中，检测伪装目标复杂的边界结构是极具挑战性的。正如图中所展示的那样，所提模型能够较为准确地检测出对象的边界，而其他方法对于复杂边界的检测效果较差。在第 6 列中，目标在视觉上“完美”地隐藏在背景环境中，所提模型依然能够准确地捕获伪装目标。总之，所提模型能够在不同的挑战因素下生成高质量的预测图。

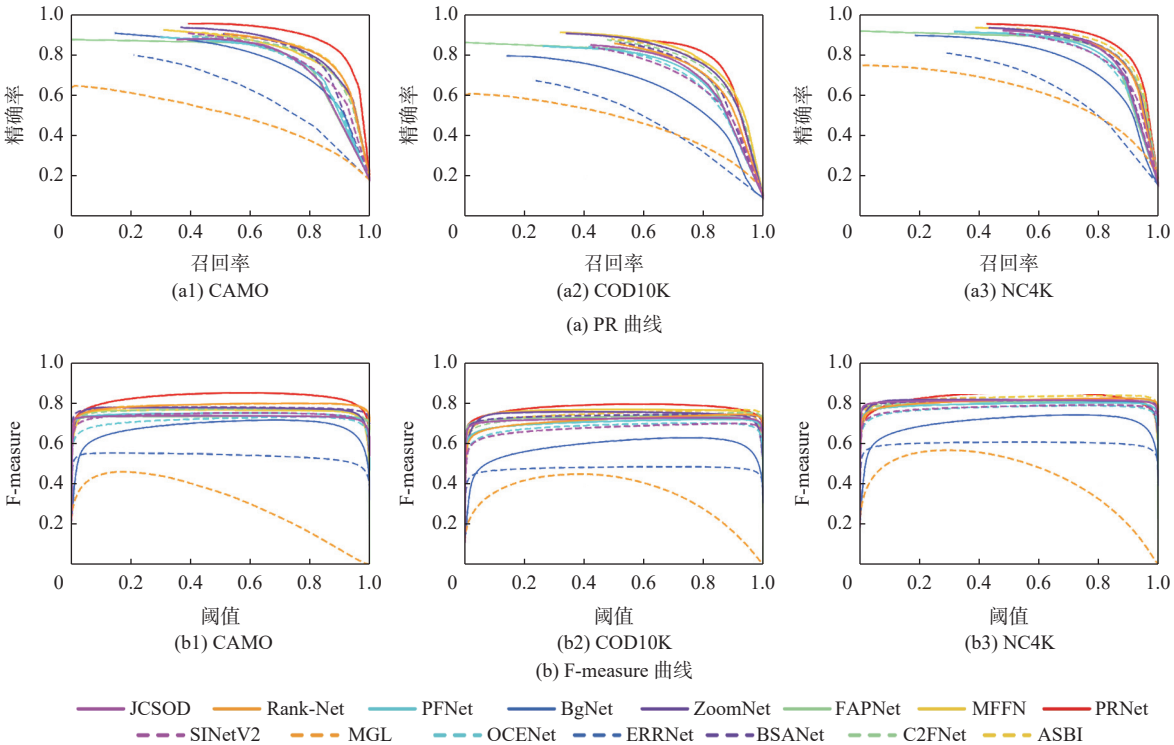


图 6　不同模型的 PR 曲线和 F-measure 曲线

Fig. 6　PR curves and F-measure curves for different models

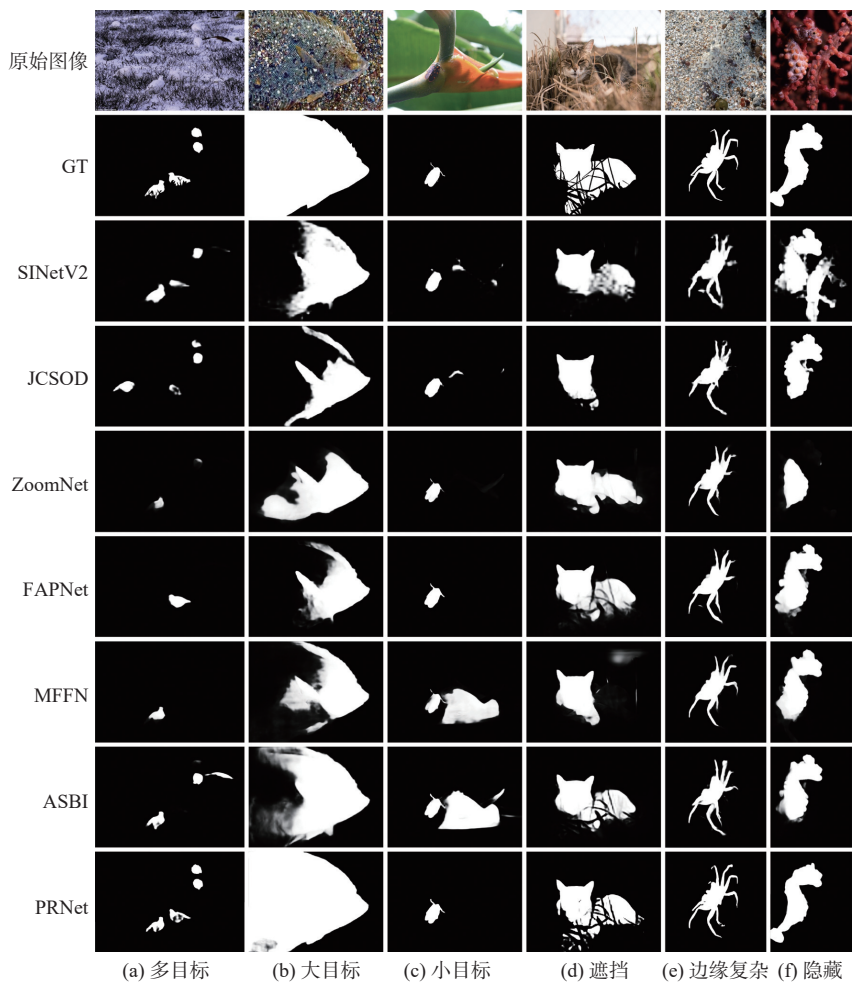


图 7　不同模型的检测结果

Fig. 7　Detection results of different models

3 结 论

1）本文针对伪装目标检测中边缘不准确、轮廓不完整的问题，提出了一种基于人眼视觉机制的伪装目标检测网络PRNet，通过模拟人类观察伪装图像的视觉感知过程，显著提升了目标检测的准确性和完整性。

2）通过在3个基准数据集（CAMO、COD10K、NC4K）上的对比实验，本方法相比最佳竞争方法ASBI，在4个核心评价指标（S-measure、E-measure、F-measure、MAE）上平均提升了1.3%、0.6%、1.1%和0.9%，整体性能明显优于14种先进的COD方法。

3）通过消融实验验证了各模块的有效性，其中细化模块（RM）对性能提升贡献最大，相较于基线模型在3个数据集上平均将四个度量指标分别提升了4.5%、3.5%、4.4%和1.7%，证明了多先验知识融合和前景-背景双向探索机制的有效性。

4）通过特征图可视化分析表明，特征增强模块显著丰富了特征多样性，定位模块能够从复杂背景中粗略定位伪装区域，细化模块有效突显了目标边缘并完善了轮廓结构，验证了生物启发式框架设计的合理性。

5）鉴于伪装目标检测在多个应用领域的重要性，未来的研究将探索将本文提出的视觉机制模拟方法扩展到其他视觉检测任务中，如显著性目标检测、医学图像分割等领域，进一步验证人眼视觉机制在计算机视觉任务中的通用性和有效性。

参考文献（References）

[1] CHUDZIK P, MITCHELL A, ALKASEEM M, et al. Mobile real-time grasshopper detection and data aggregation framework[J]. Scientific Reports, 2020, 10(1): 1150.

[2] CHEN S M, XIONG J T, JIAO J M, et al. Citrus fruits maturity detection in natural environments based on convolutional neural networks and visual saliency map[J]. Precision Agriculture, 2022, 23(5): 1515-1531.

[3] LIN Z S, YE H X, ZHAN B, et al. An efficient network for surface defect detection[J]. Applied Sciences, 2020, 10(17): 6085.

[4] JI G P, CHOU Y C, FAN D P, et al. Progressively normalized self-attention network for video polyp segmentation[C]//Proceedings of the Medical Image Computing and Computer Assisted Intervention – MICCAI 2021. Cham: Springer, 2021: 142-152.

[5] WU Y H, GAO S H, MEI J, et al. JCS: an explainable COVID-19 diagnosis system by joint classification and segmentation[J]. IEEE Transactions on Image Processing, 2021, 30: 3113-3126.

[6] YANG Z Y, DAI Q H, ZHANG J S. Visual perception driven collage synthesis[J]. Computational Visual Media, 2022, 8(1): 79-91.

[7] HORVÁTH G, PERESZLÉNYI Á, ÅKESSON S, et al. Striped bodypainting protects against horseflies[J]. Royal Society Open Science, 2019, 6(1): 181325.

[8] MERILAITA S, SCOTT-SAMUEL N E, CUTHILL I C. How camouflage works[J]. Philosophical Transactions of the Royal Society B: Biological Sciences, 2017, 372(1724): 20160341.

[9] SENGOTTUVELAN P, WAHI A, SHANMUGAM A. Performance of decamouflaging through exploratory image analysis[C]// 2008 First International Conference on Emerging Trends in Engineering and Technology. Piscataway: IEEE Press, 2008: 6-10.

[10] YIN J Q, HAN Y B, HOU W D, et al. Detection of the mobile object with camouflage color under dynamic background based on optical flow[J]. Procedia Engineering, 2011, 15: 2201-2205.

[11] FAN D P, JI G P, SUN G L, et al. Camouflaged object detection[C]//Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2020: 2774-2784.

[12] MEI H Y, JI G P, WEI Z Q, et al. Camouflaged object segmentation with distraction mining[C]//Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2021: 8768-8777.

[13] FAN D P, JI G P, CHENG M M, et al. Concealed object detection[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 44(10): 6024-6042.

[14] PANG Y W, ZHAO X Q, XIANG T Z, et al. Zoom in and out: a mixed-scale triplet network for camouflaged object detection[C]// Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2022: 2150-2160.

[15] REN J J, HU X W, ZHU L, et al. Deep texture-aware features for camouflaged object detection[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2023, 33(3): 1157-1167.

[16] ZHOU T, ZHOU Y, GONG C, et al. Feature aggregation and propagation network for camouflaged object detection[J]. IEEE Transactions on Image Processing, 2022, 31: 7036-7047.

[17] GAO S H, CHENG M M, ZHAO K, et al. Res2Net: a new multi-scale backbone architecture[J]. IEEE Trans Pattern Anal Mach Intell, 2021, 43(2): 652-662.

[18] SUN Y J, WANG S, CHEN C, et al. Boundary-guided camouflaged object detection[EB/OL]. (2022-07-02)[2023-07-25]. https://doi.org/10.48550/arXiv.2207.00794.

[19] WANG K, BI H B, ZHANG Y, et al. D^2C-Net: a dual-branch, dual-guidance and cross-refine network for camouflaged object detection[J]. IEEE Transactions on Industrial Electronics, 2022, 69(5): 5364-5374.

[20] ZHU H W, LI P, XIE H R, et al. I can find you! boundary-guided separated attention network for camouflaged object detection[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2022, 36(3): 3608-3616.

[21] QIN X B, ZHANG Z C, HUANG C Y, et al. BASNet: boundary-aware salient object detection[C]//Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2019: 7471-7481.

[22] JI G P, ZHU L, ZHUGE M C, et al. Fast camouflaged object detection via edge-based reversible re-calibration network[J]. Pattern Recognition, 2022, 123: 108414.

[23] CHEN T Y, XIAO J, HU X G, et al. Boundary-guided network for

camouflaged object detection[J]. Knowledge-Based Systems, 2022, 248: 108901.

[24] WEI J, WANG S H, HUANG Q M. F³Net: fusion, feedback and focus for salient object detection[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2020, 34(7): 12321-12328.

[25] XIE E Z, WANG W J, WANG W H, et al. Segmenting transparent objects in the wild[C]//Proceedings of the Computer Vision – ECCV 2020. Berlin: Springer, 2020: 696-711.

[26] LE T N, NGUYEN T V, NIE Z L, et al. Anabranch network for camouflaged object segmentation[J]. Computer Vision and Image Understanding, 2019, 184: 45-56.

[27] LV Y Q, ZHANG J, DAI Y C, et al. Simultaneously localize, segment and rank the camouflaged objects[C]//Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2021: 11586-11596.

[28] ACHANTA R, HEMAMI S, ESTRADA F, et al. Frequency-tuned salient region detection[C]//Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2009: 1597-1604.

[29] CHENG M M, FAN D P. Structure-measure: a new way to evaluate foreground maps[J]. International Journal of Computer Vision, 2021, 129(9): 2622-2638.

[30] FAN D P, JI G P, QIN X B, et al. Cognitive vision inspired object segmentation metric and loss function[J]. Scientia Sinica Informationis, 2021, 51(9): 1475.

[31] PERAZZI F, KRÄHENBÜHL P, PRITCH Y, et al. Saliency filters: Contrast based filtering for salient region detection[C]//Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2012: 733-740.

[32] LI A X, ZHANG J, LV Y Q, et al. Uncertainty-aware joint salient object and camouflaged object detection[C]//Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2021: 10066-10076.

[33] ZHAI Q, LI X, YANG F, et al. Mutual graph learning for camouflaged object detection[C]//Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2021: 12992-13002.

[34] LIU J W, ZHANG J, BARNES N. Confidence-aware learning for camouflaged object detection[EB/OL]. (2021-06-22)[2023-07-13]. https://doi.org/10.48550/arXiv.2106.11641.

[35] CHEN G, LIU S J, SUN Y J, et al. Camouflaged object detection via context-aware cross-level fusion[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2022, 32(10): 6981-6993.

[36] ZHENG D H, ZHENG X C, YANG L T, et al. MFFN: multi-view feature fusion network for camouflaged object detection[EB/OL]. (2021-06-22)[2023-07-15]. https://doi.org/10.48550/arXiv.2210.06361.

[37] ZHANG Q, SUN X X, CHEN Y R, et al. Attention-induced semantic and boundary interaction network for camouflaged object detection[J]. Computer Vision and Image Understanding, 2023, 233(8): 103719.

Camouflaged object detection network based on human visual mechanisms

ZHANG Dongdong，WANG Chunping，FU Qiang*

(Shijiazhuang Campus，Army Engineering University of PLA，Shijiazhuang 050003，China)

Abstract：Camouflaged object identification is a new visual detection job that has many applications in several fields. Its goal is to identify camouflaged targets that are completely disguised in their environment. To address the problem of current camouflaged object detection algorithms failing to accurately and completely identify the object's structure and boundaries, this paper designs a bio-heuristic framework based on the human visual perception process when observing camouflaged images, and names its Positioning and Refinement Network (PRNet). Res2Net is used to extract the original features of the image and mine the edge cues of the target from multi-level information. A feature enhancement module is specially designed to expand the perceptual field while enriching the global contextual information. Then, the positioning module utilizes the dual-attention mechanism to locate the approximate position of the target from both channel and spatial dimensions. Lastly, the refinement module leverages multi-type information to further improve the target structure and edges by concentrating on target cues in both the foreground and background. Extensive experimental results on three widely used benchmark datasets for camouflage target detection demonstrate that the overall performance of the proposed network significantly outperforms 14 comparative algorithms and performs well in a variety of complex scenarios.

Keywords：camouflaged object detection；convolutional neural network；visual perception；feature enhancement；attention mechanism

Received：2023-08-04；**Accepted**：2023-10-05；**Published Online**：2023-11-01 08：43

URL：link.cnki.net/urlid/11.2625.V.20231031.1708.001

* **Corresponding author**. E-mail：1418748495@qq.com

http://bhxb.buaa.edu.cn jbuaa@buaa.edu.cn
DOI: 10.13700/j.bh.1001-5965.2023.0552

基于自适应阈值和速度优化的轻量化语义VSLAM方法

齐浩[1]，付悦欣[1]，胡祝华[1,*]，吴佳琪[1]，赵瑶池[2]

(1. 海南大学 信息与通信工程学院，海口 570228；2. 海南大学 网络空间安全学院，海口 570228)

摘 要：视觉同步定位与地图构建（VSLAM）是一种利用视觉等传感器来获取未知环境信息的技术，广泛应用于无人驾驶、机器人、增强现实等领域。然而，室内场景下的VSLAM对动态对象进行像素级的语义分割存在较高的计算开销，并且光照变化使得动态物体的外观也发生变化，导致其与静态环境产生遮挡或混淆。针对以上问题，提出了一种基于自适应阈值和速度优化的轻量化语义VSLAM模型。采用了轻量化的一阶段目标检测网络YOLOv7-tiny，结合光流算法，有效地检测了图像的动态区域，并对不稳定特征点进行了剔除。同时，特征点提取算法基于输入图像的对比度信息，自适应地调整阈值。结合二进制词袋与局部建图线程精简的优化方法，加快了加载和匹配速度，提高了系统在室内动态场景下的运行速度。实验结果表明：所提算法在室内高动态场景下能够有效地剔除动态特征点，提高了相机的定位精度。在运行速率方面平均处理速度达到了19.8 FPS，在实际场景下可以满足实时性的需求。

关 键 词：VSLAM；动态场景；YOLOv7-tiny；自适应阈值；特征点

中图分类号：V221+.3; TB553

文献标志码：A **文章编号**：1001-5965(2025)07-2562-11

视觉同步定位与地图构建(visual simultaneous localization and mapping, VSLAM)由于其低成本传感器,易于融合其他传感器以及更丰富的环境信息的优势而发展迅速。其不仅能提供自我定位、位姿估计和环境地图构建的方法,而且能在未知且复杂的环境下提供丰富的场景信息[1]。目前的VSLAM方法存在诸多不足,具体如下:①现有的VSLAM系统通常假设环境是静态的,而实际生活与应用场景都是非静止的,VSLAM算法的精度不够理想[2];②多数VSLAM算法没有充分利用场景信息进行语义分析,并且部分算法时间复杂度过高,难以满足实时性的需求;③目前动态场景下特征点的误匹配概率较高,整体系统的稳定性较差;④在环境光照变化明显的场景中,现有VSLAM系统不能随着环境改变而调整阈值参数,导致系统的鲁棒性不高[3]。

现有的VSLAM系统主要分为依靠直接法的VSLAM和特征点法的VSLAM。关于直接法,Newcombe等[4]提出了稠密跟踪与地图构建(Dense Tracking and Mapping, DTAM),其利用不同帧深度图直接匹配来估计位姿,但在光照变化场景下精度较差。Forster等[5]提出了半直接单目视觉里程计(semi-direct monocular visual odometry, SVO),其通过角点提取和利用周围像素信息进行位姿估计,实现了高速度和较高准确性。然而,缺乏后端优化和闭环检测线程使SVO容易累积误差,对长时间位姿估计

收稿日期：2023-08-28；**录用日期**：2024-03-22；**网络出版时间**：2024-05-08 10:50

网络出版地址：link.cnki.net/urlid/11.2625.V.20240507.1513.002

基金项目：国家重点研发项目 (2022YFD2400504)；国家自然科学基金 (62161010)；海南省重点研发计划项目 (ZDYF2022SHFZ039, ZDYF2022GXJS348)

* **通信作者**. E-mail：eagler_hu@hainanu.edu.cn

引用格式：齐浩，付悦欣，胡祝华，等. 基于自适应阈值和速度优化的轻量化语义VSLAM方法[J]. 北京航空航天大学学报，2025，51(7)：2562-2572. QI H，FU Y X，HU Z H，et al. A lightweight semantic VSLAM approach based on adaptive thresholding and speed optimization[J]. Journal of Beijing University of Aeronautics and Astronautics，2025，51(7)：2562-2572 (in Chinese).

的精度构成挑战。关于特征点法，Davision 等[6] 提出了单目即时定位与地图构建（monocular simultaneous localization and mapping, Mono SLAM），其采用特征点法和扩展卡尔曼滤波算法，能够在小范围环境中完成自身位姿估计，但是在大环境中会出现大范围漂移且精度较低的问题。Klein 等[7] 的并行跟踪与建图（parallel tracking and mapping, PTAM）系统引入了基于非线性优化的方法，通过并行化跟踪线程和建图线程，采用关键帧联合优化的思路，摒弃了逐帧处理的方式。在定向 FAST 与旋转 BRIEF SLAM2（oriented FAST and rotated BRIEF SLAM2, ORB-SLAM2）[8] 的基础上，Campos 等[9] 进一步提出了 ORB-SLAM3，通过引入惯性测量单元（inertial measurement unit, IMU）数据进行辅助，实现了多传感器融合的视觉惯性 SLAM 系统。这一创新为多传感器融合 VSLAM 系统提供了新的思路和解决方案。为高效利用场景图像信息，Shen 等[10] 基于 ORB-SLAM3 算法，将图像特征与离线的词袋模型相结合实现一种实时在线更新词袋的闭环检测模型。

上述 VSLAM 算法都假设机器人所处的环境为静止状态，因此无法对复杂的动态场景进行识别。动态场景中的 VSLAM 算法主要可以分为以下 2 类。

第 1 类主要是借助传统的方法利用静态特征点进行位姿估计和地图构建。基于传统的几何方法如帧差法和背景扣除法[11-12]，都是通过对输入图像的像素进行操作来检测环境中存在的动态对象。除此以外，还可以利用光流法来判别场景中的运动物体。Klappstein 等[13] 利用光流法计算动态对象与场景中光流运动的偏离程度，从而实现动态物体的分割。Derome 等[14] 利用不同图像之间的残差计算光流，之后再通过残差场中的点来判定动态对象。光流法不仅能够准确表现运动物体的信息，还可以提供丰富的场景数据用于三维重建。但是光流法的前提是亮度恒定，系统极易受光照变化的影响。

第 2 类借助图像中的语义信息，对输入图像中的动态对象进行检测并将图像中包含的动态区域剔除。动态语义 SLAM（dynamic semantic SLAM，DS-SLAM）[15] 在 VSLAM 的三大线程之外建立了并行的语义分割线程，利用 SegNet[16] 网络对动态对象进) 行检测分割。这种多线程处理虽然可以提高定位效果，但也增加了系统的复杂性，导致系统更加冗余庞大。Bescos 等[17] 提出了 DynaSLAM 系统，通过采用多视图几何和像素级的语义分割网络来处理动态对象，并借助静态图对动态对象遮挡的背景进行修复。然而像素级的语义分割网络通常需要进行大量的计算，无法满足实时性要求。Fu 等[18] 和 Cai 等[19] 将注意力模块融入掩码区域卷积神经网络（mask region-based convolutional neural network, Mask R-CNN）[20] 的特征提取网络，以提升动态目标的特征提取和分割精度。但受限于 Mask R-CNN 神经网络的特性，该系统仍然难以满足实时性的要求。Qi 等[21] 和 Chang 等[22] 均采用 YOLO 网络结合光流法对动态对象进行剔除，但是没有考虑到光流法的提取效果易受不同环境下光照变化的影响。

通过对以上研究现状的分析可知，尽管 VSLAM 系统在动态环境中表现出了一定的适用性，但仍然存在系统冗余和复杂性较高的问题，以及未充分考虑环境光照变化对动态 VSLAM 的影响。针对目前 VSLAM 在动态场景下存在的挑战，本文进行了以下工作。

1）为了提高对动态目标的检测速度，本文在对 VSLAM 算法改进的基础上，构建了基于 YOLOv7-tiny 的轻量化语义 SLAM 算法模型。通过将 YOLOv7-tiny 模型与卢卡斯-卡纳德（Lucas-Kanade, LK）光流算法相结合，本文算法不仅实现了对图像动态区域的精准检测，还高效剔除了动态特征点，从而充分满足了 VSLAM 系统实时运行的需求。

2）本文提出了基于自适应阈值的特征点提取算法，并对 VSLAM 系统进行速度优化。在传统的特征点提取算法中，采用固定的阈值常常使得系统难以适应各种环境变化，从而导致其鲁棒性相对较弱。为此，本文基于输入图像的对比度信息实时对提取阈值的大小进行调整，从而提高 SLAM 系统的稳定性和特征点提取效率。此外，通过二进制词袋与局部建图线程简化相结合的优化方法，进一步提升了 SLAM 系统的速度。

1　本文方法

本文将性能优异的 YOLOv7-tiny 网络结合光流法作为动态物体的剔除模块，融合到基于自适应阈值和速度优化的 ORB-SLAM2 的系统中，改进后的视觉系统整体框架如图 1 所示。在初始化阶段，系统根据时间戳加载输入图像和二进制词袋模型，读取相机标定等参数，完成 VSLAM 系统的初始化。之后系统用 YOLOv7-tiny 网络提取结果，并用光流法进行动态物体判断和剔除。与此同时，系统根据每张图像的对比度自适应的确定特征点提取过程中的阈值，跟踪线程依据语义信息提取静态物体上的特征点，并利用特征点进行局部地图跟踪。

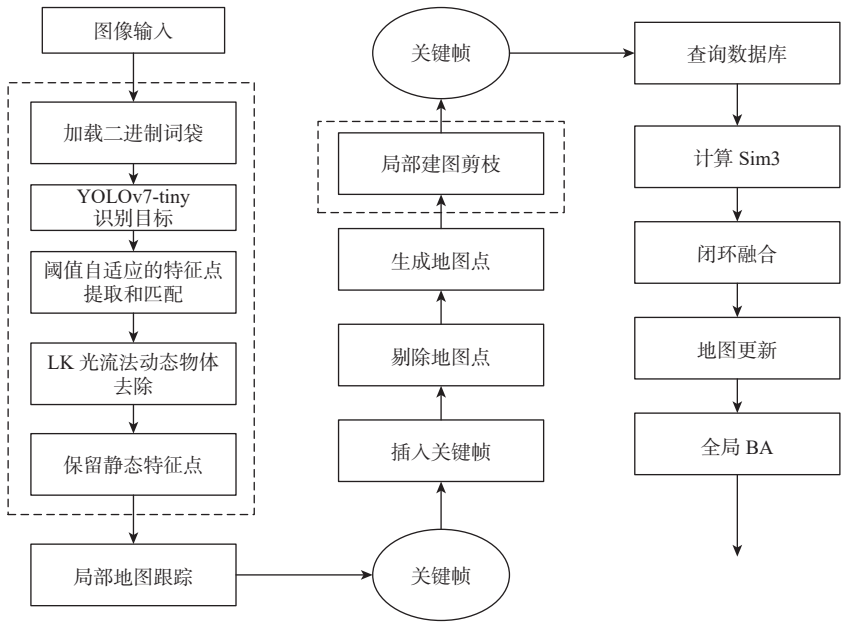


图1 系统整体框

Fig. 1 System structure

在局部地图构建线程中，系统主要完成了关键帧插入、地图点的剔除和生成。

1.1 动态特征点检测

YOLOv7[23]是近年来提出的实时目标检测模型之一，其在5～160帧/s的目标检测模型中都具有较高的速度和准确度。在GPU V100上，当实现30帧/s的实时检测速度时，其能够达到56.8%AP的准确率。YOLOv7-tiny是专门针对边缘GPU设备而设计的模型，在具有较高准确度的同时，也能够满足语义VSLAM系统对于实时运行的速率要求。相较于YOLOv7模型，YOLOv7-tiny模型在规模上进行了缩减，实现了高效且紧凑的设计。同时，YOLOv7-tiny也仍然延续了YOLOv7网络的模型重参数化和标签分配策略，对新型的高效层聚合网络(efficient layer aggregation networks, ELAN)结构进行精简，并对级联型的网络进行模型的缩放以此来保持网络模型初始设计时的最佳结构和性能。

在使用YOLOv7-tiny模型进行物体检测之前，首先根据物体的动态特性将室内物体划分为2类：一级动态物体和二级动态物体。一级动态物体包括人和动物，以及被人移动的物体，如书本和杯子。二级动态物体基本不会移动，如冰箱、洗衣机和电视。根据YOLOv7-tiny模型的检测结果，若为一级动态物体，则利用LK光流算法进一步进行判定。若为二级动态物体，则直接视为静态物体。相比于将所有特征点进行动态检测的方法，该策略在高动态数据集下能够提供更好的定位精度，同时也减少了系统的实时性影响。

光流法通过比较2幅连续图像中对应像素的亮度值变化，能够精确地推断出这些像素在图像中的运动方向和大小。假设将图像的灰度视为时间的函数，在时刻t，位于图像坐标(x,y)处的ORB特征点的灰度表示为

$$I(x,y,t)=I(x+\mathrm{d}x,y+\mathrm{d}y,t+\mathrm{d}t) \tag{1}$$

式中：t和$t+\mathrm{d}t$分别为相邻图像帧对应的时间；$I(x,y,t)$和$I(x+\mathrm{d}x,y+\mathrm{d}y,t+\mathrm{d}t)$则为对应的灰度。基于光流法的灰度不变假设，推导出了相关结果[22]。

$$\left[\boldsymbol{I}_x\boldsymbol{I}_y\right]_k\begin{bmatrix}\boldsymbol{u}\\ \boldsymbol{v}\end{bmatrix}=-\boldsymbol{I}_k \qquad k=1,\cdots,k^2 \tag{2}$$

式中：$\boldsymbol{u}$为特征点在X轴上的运动速度；$\boldsymbol{v}$为在y轴上的速度；$\boldsymbol{I}_x$为图像在该点处x轴方向的梯度；I_y为图像在该点处y轴方向的梯度；$\boldsymbol{I}_k$为特征点灰度对时间的变化量。在LK光流中，以特征点为中心假设6×6的窗口，即$k=6$，内部36个像素具有相同的运动速度和方向。式(2)为一个超定方程，常用最小二乘法求解，如果一次迭代效果不佳，可以多迭代几次即可求解。

当相在拍摄过程中发生移动时，静态部分由于相机的运动也会产生光流矢量。

$$\begin{bmatrix}\boldsymbol{U}\\ \boldsymbol{V}\end{bmatrix}=\frac{1}{N}\sum_{i=1}^{N}\begin{bmatrix}\boldsymbol{u}_i\\ \boldsymbol{v}_i\end{bmatrix} \tag{3}$$

式中：$\boldsymbol{U}$为静态部分的特征点在x轴上的运动速度；$\boldsymbol{V}$为静态部分的特征点在y轴上的运动速度。接下来，利用式(2)计算得到的动态光流速度$\boldsymbol{u}$和$\boldsymbol{v}$，与静态部分的速度$\boldsymbol{U}$和$\boldsymbol{V}$进行计算，以判断特征点是否为动态特征点。

$$\sqrt{(\boldsymbol{U}-\boldsymbol{u})^2+(\boldsymbol{V}-\boldsymbol{v})^2}>\varepsilon_{\mathrm{th}} \tag{4}$$

ε_{th}由实验结果所得，为静态平均光流矢量的 2 倍。最后根据求解结果，如果特征点的光流大于阈值ε_{th}，则该点被认为是动态点。

1.2　FAST 图像特征点提取

基于加速分段测试的特征提取（features from accelerated segment test, FAST）是一种用于快速目标检测和识别的角点检测器。在光照不足或者光线过强的情况下，图像细节的信息缺失较多，相应图像的对比度也大幅下降。但是在 VSLAM 系统运行过程中，FAST 特征点提取过程中的阈值参数是固定不变的，不能根据环境的变化动态调整或者做出改变。在这样的情况下，如果仍然沿用和正常光照条件下同样的阈值参数，就会导致 VSLAM 系统不容易提取到图像中的特征点，甚至因初始特征点较少而难以完成 VSLAM 系统的初始化，进而影响后续特征匹配、关键帧的生成等过程。

针对以上问题，本文采用了一种基于自适应阈值的改进 FAST 特征点提取方法，根据图像的对比度信息自适应调整原本系统中设置的固定阈值。一方面，能够提高系统在不同光照条件下的鲁棒性和稳定性。另一方面，在光照正常且图像纹理结构特征丰富、特征点信息充足的环境中，能够在保持系统性能和准确性的同时，适度减少提取的特征点数目，进一步降低了系统整体的算力消耗。

图像对比度是指像素点和周围临近像素之间黑与白灰度层级的反差，视为图像中黑色与白色的比值[24]。一般情况下图像的对比度越高，图像表现的层次越丰富，包含的信息与细节也就越丰富。图像对比的计算公式如式（6）所示，C代表图像对比度，i和j分别表示当前像素和其邻域像素的灰度值，$\mu(i,j)$表示当前像素与邻域像素之间的灰度差，$P_\mu(i,j)$表示相邻像素间灰度差为μ的概率。在图像中像素邻域通常有四邻域、对角邻域和八邻域 3 种，其中四邻域和对角邻域叠加可得八邻域，在本文中使用的是四邻域。

$$\mu(i,j)=|i-j| \tag{5}$$

$$C=\sum_{\mu}\mu(i,j)^2P_\mu(i,j) \tag{6}$$

FAST 的阈值根据文献 [8] 的推荐设置为 20 和 7，该推荐基于其在 ORB-SLAM2 上对 EuRoc[25] 数据集的多次实验。本文根据求出的图像对比度C，设计自适应阈值λ：

$$\lambda=\alpha\ln C+\beta \tag{7}$$

式中：α和β为自适应参数，α范围为 0.5～1，β范围为 10～15，取值根据实验数据确定。自适应阈值λ能够使系统更好地适应不同的图像和场景。对数变换可以使得对比度C在较小的范围内取值时，阈值λ有一个较大的变化；而在C取值较大时，λ的变化相对较小。当图像对比度较低时，算法能够更加敏感地调整阈值；而当对比度较高时，阈值的调整相对较为平缓，从而更好地适应不同对比度的图像。

接下来根据本文设置的自适应阈值λ，进行特征点的提取。如图 2 所示，FAST 特征点的提取以检测点P为圆心，以半径为 3 的像素点构成一个检测圆，在检测圆上选取 16 个连续像素点，比较圆上像素点的灰度值与检测点P的灰度值的大小。

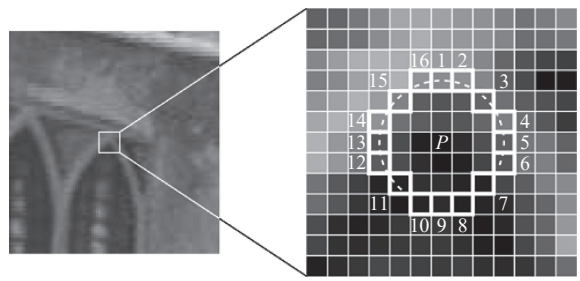


图 2　FAST 特征点提取示意[26]

Fig. 2　Schematic diagram of FAST feature point extraction[26]

特征点提取分为以下 4 个步骤：

步骤 1　初步筛选：首先，计算圆周上的像素点P_1、P_9与中心像素点的像素差。如果这 2 个像素差的绝对值都小于设定的阈值，则该点不可能是一个特征点。否则，将其加入候选特征点的列表中。

步骤 2　进一步的筛选：对于候选点P，计算P_1、P_5、P_9和P_{13}与P点的像素差。如果这些像素差的绝对值中，小于阈值的个数不超过 1 个，则P点是一个候选特征点。否则，P 点不是一个特征点。

步骤 3　确定特征点：对于候选特征点P，如果其像素圆上有连续 12 个像素与中心像素的像素差绝对值大于设定的阈值，则P点是一个特征点。否则，P 点不是一个特征点。

步骤 4　遍历并检查所有像素：最后，遍历选择区域内的每一个像素点，对每个像素执行上述 3 个步骤。基于自适应阈值的特征提取算法流程如图 3 所示。算法 1 为自适应阈值算法流程。

算法　自适应阈值算法。

输入： 图像 images。

输出： 特征点集合 keypoints。

1.计算自适应阈值：$\lambda=\alpha\ln C+\beta$

2.以点 P 为圆心选择 16 个点

for P in image do

if ($|P_1-P|<\lambda$) or ($|P_9-P|<\lambda$) then

P 不是特征点

break

end if

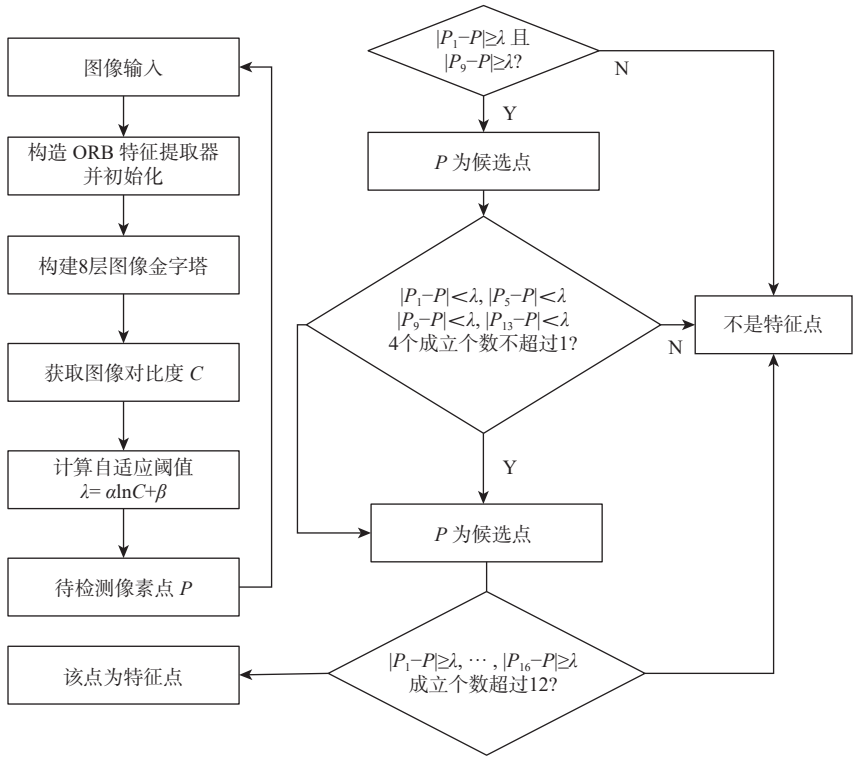


图 3 基于自适应阈值的特征点提取算法流程

Fig. 3 Based on adaptive threshold feature point extraction algorithm flow chat

if 满足$|P-P_i|\geqslant\lambda, i=1,5,9,13$的个数大于 1 then

P 不是特征点

break

end if

if 有 12 个连续点满足$|P-P_j|\geqslant\lambda, j=1,\cdots,16$ then

P 是特征点

else

break

end if

end for

1.3 VSLAM 系统的速度优化

本文系统中的字典和词袋模型都是基于词袋库 DBoW2[27]。在原算法中预先放置的词袋格式为文本，文件体积大，加载读取时间过长，对 SLAM 系统的运行效率具有较大的影响。因此，本文将文本格式的词袋转化为二进制词袋文件，并修改相应代码重新编译。相较于文本格式，二进制词袋文件不仅能确保内容不易被修改或损坏，而且文件体积更小且存在进一步压缩的可能。更为重要的是，二进制词袋文件更契合计算机的读取方式，从而减少了原有文本格式转化和数据处理的时间。同时，二进制词袋不仅能够在 VSLAM 系统的初始化阶段加快地图和词袋模型加载时间，而且也能够加速回环检测时匹配检测速度。视觉词袋模型流程如图 4 所示。

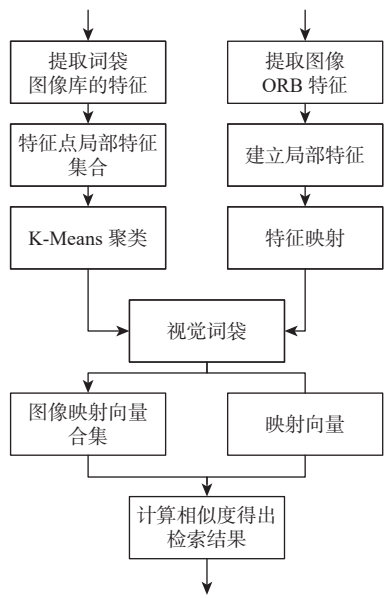


图 4 视觉词袋模型流程

Fig. 4 Visual bags of word structure diagram

局部建图线程主要是对关键帧进行处理，完成地图点的剔除和创建，与此同时对地图进行局部最优处理，最后将关键帧加入闭环检测队列。在这个过程中耗时最长的为局部光束法平差（bundle adjustment, BA）优化，其主要是对共视图中的关键帧进行局部求解最优。在回环检测线程中，系统对整体的关键帧和地图点进行全局 BA 优化。因此，删减局部 BA 部分的优化策略能够减少时间消耗，并且对算法精度的影响较小。优化后的局部地图构建线程如图 5 所示，其中虚线部分表示已被删减的内容。

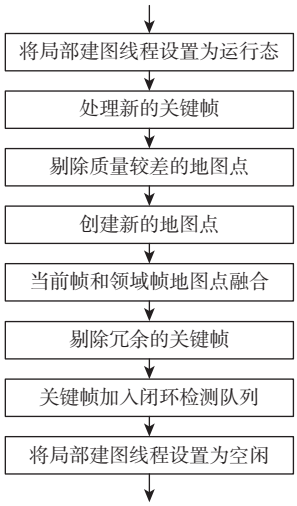


图 5　优化后的局部地图构建线程流程

Fig. 5　Optimized local mapping flow chart

2　实　验

2.1　数据集与实验环境

本文实验中使用的是 TUM 公开数据集[28]。TUM RGB-D 数据集是由 Microsoft Kinect 摄像头与深度传感器所记录的一系列多样化场景序列组成，包含动态场景、3D 物体重建、机器人 SLAM、结构与低纹理等众多应用和研究场景。该数据集除了提供包含深度信息的彩色图片序列外，还包括图片序列对应的时间戳和外部设备捕捉到的相机位姿，能够为评估算法的准确度提供很好的依据。本文实验采用 TUM 数据集中的 8 组序列，包括 4 个低动态数据序列 fr3_sitting 和 4 个高动态数据序列 fr3_walking。

在对实验结果进行定量评估时，本文使用 SLAM 领域通用的评价标准，即相对位姿误差（relative pose error, RPE）和绝对轨迹误差（absolute trajectory error, ATE）[28]。在实验中，分别计算了动态场景下每个数据集 ATE 和 RPE 的均方根误差（root mean squared error, RMSE）。文中相对提升率 η 的计算公式为

$$\eta = \frac{E_0 - E_M}{E_0} \times 100\% \tag{8}$$

实验中 VSLAM 部分均采用 C++开发，深度学习部分为 python 语言，其他实验环境参数如表 1 所示。

2.2　速度优化实验

二进制词袋和原有词袋占用空间和时间效率的对比如表 2 所示。从表中可以得出，在对原有的词袋模型进行优化和格式转换后，无论是储存空间的占用还是模型的加载时间都优于原有词袋模型。这主要因为二进制词袋在被计算设备读取或者查找时，不需要再进行格式转换和数据处理，从而提升了时间效率。

表 1　实验环境

Table 1　Experimental environment

配置名称	配置情况
处理器	Intel(R) Core(TM) i9-10900x CPU@3.70 GHz
显卡	NVIDIA GeForce RTX 3080
操作系统	Ubuntu 18.04
深度学习框架	Keras 2.0.9、TensorFlow-gpu 1.14.0

表 2　不同词袋性能对比

Table 2　Performance comparison of different BOW

性能	词袋空间大小/MB	平均加载时间/ms
原有词袋	145.3	8 289.01
二进制词袋	42.3	267.92

注：二进制词袋比原有词袋在词袋空间和平均加载时间分别提升了 70.89%、96.77%。

在对系统进行优化前，首先利用 C++标准库的时钟函数对跟踪线程和局部地图构建线程的时间消耗进行统计，最终统计结果如表 3 所示。从表中可以看出，局部地图构建线程平均耗时为 609.71 ms，其中局部 BA 优化部分耗时最长，达到 458.52 ms。不难发现，对局部建图线程进行优化是较为高效的方式。通过对局部 BA 部分的删减，能够将局部地图构建部分的时间缩短为原来的四分之一。

表 3　不同线程各部分耗时对比

Table 3　Time-consuming comparison of various parts of different threads

线程	模块	用时/ms	标准差/ms
跟踪线程	特征提取	17.78	3.85
	位姿估计	2.87	1.31
	整体耗时	33.22	13.99
局部地图构建线程	关键帧插入	17.96	8.11
	地图点创建	91.17	30.59
	局部BA优化	458.52	319.41
	总体耗时	609.71	368.44

为了进一步研究实验的效果，本文统计了各个算法在跟踪线程中处理单帧图片的用时，实验结果如表 4 所示。将 ORB-SLAM2 算法与 YOLOv3[29] 和 YOLOv4[30] 算法结合，形成了 YOLOv3-SLAM 和 YOLOv4-SLAM 系统。从表中可以看出，基于 YOLO 目标检测网络的语义 SLAM 系统在运行效率上要明显高于基于像素级实例分割 SLAM 系统，避免了复杂度高和计算量大的图像分割。本文算法与 YOLOv3-SLAM 相比，处理单帧图片时间平均减少了 33.44%，与 YOLOv4-SLAM 相比平均减少了 36.35%。这是因为本文算法采用了高效的 YOLOv7-tiny 网

表4 跟踪线程处理单帧图片用时对比

Table 4 Time-consuming comparison of tracking thread processing single frame

数据集名称[28]	不同模型处理单帧图片耗时/ms			
	YOLOv3-SLAM[29]	YOLOv4-SLAM[30]	DynaSLAM[17]	本文算法
fr3_s_static	55.96	55.54	1 444.07	25.57
fr3_s_xyz	61.36	61.90	1 616.43	42.36
fr3_s_half	70.43	66.75	1 534.07	55.78
fr3_s_rpy	63.69	58.57	1 491.33	41.89
fr3_w_static	83.42	82.97	1 676.27	62.32
fr3_w_xyz	89.47	87.60	1 718.56	69.37
fr3_w_half	95.11	90.74	1 683.11	50.36
fr3_w_rpy	82.57	72.72	1 519.81	54.78

络作为动态特征检测网络，以及对VSLAM系统进行了优化。在局部建图线程中优化处理关键帧的策略以提升计算效率，并采用二进制词袋模型对特征描述子进行压缩编码，显著加速特征点匹配与地图数据的加载过程。同时，基于自适应阈值的特征点提取算法能在对比度高、纹理特征丰富的场景下少提取一定数量的特征点进而减少计算量提高效率。由此可以看出本文算法具有更好的运行效率，更能满足动态场景下VSLAM系统对于实时性的需求。

2.3 自适应阈值实验

基于自适应阈值的特征点提取算法在曝光不足和过曝情况下的提取效果如图6所示。从图中能够看出，在不同的环境光照下基于阈值自适应的提取算法均能较多地提取图像中的特征点。

为了深入研究自适应阈值的效果，进一步将实验分为2组：一组增加了自适应阈值算法的模块，另一组则没有增加该模块。2组实验在其余模块均进行了相应的优化和改进，以确保唯一的变量为自适应阈值算法的引入与否。同时，对2个动态数据集进行了对比度和亮度的调整，并使用绝对轨迹误差的RMSE值作为评价指标。从表5和表6中可以看出，在不同的环境下，加入自适应阈值部分的算法有着更低的绝对轨迹误差。



(a) 原算法特征点提取效果 (b) 原算法特征点提取效果

(c) 改进后特征点提取效果 (d) 改进后特征点提取效果

图6 改进前后特征点提取效果对比图

Fig. 6 Comparison of feature point extraction effect before and after improvement

表5 对比度降低和亮度减少的定位效果

Table 5 Positioning effect of contrast reduction and brightness reduction m

数据集名称	未加入阈值法	加入阈值
fr3_walking_static	0.184 179	0.179 079
fr3_walking_xyz	0.023 520	0.021 872

表6 对比度降低和亮度增加的定位效果

Table 6 Positioning effect of contrast reduction and brightness increase m

数据集名称	未加入阈值	加入阈值
fr3_walking_static	0.209 476	0.188 325
fr3_walking_xyz	0.028 880	0.023 741

2.4 动态特征点剔除的精度实验

在8个包含高动态和低动态场景的数据集序列上进行了综合比较实验，包含ORB-SLAM2、YOLOv3-SLAM、YOLOv4-SLAM、DynaSLAM以及本文算法。评估结果主要基于绝对轨迹误差和相对位姿误差，并通过计算RMSE值进行量化分析。实验结果如表7和表8所示。从实验结果来看，在低动态环境下，由于物体运动速度相对较慢，传感器噪声较低，ORB-SLAM2算法已经能够较好地跟踪并定位相机位姿。而在高动态环境中，本文算法的绝对轨迹误差和相对位姿误差比ORB-SLAM2算法有显著的降低。

与其他动态SLAM算法相比，本文算法比以YOLOv3和YOLOv4为主的轻量化模型的定位误差更低。与DynaSLAM相比，在不同的数据集下均有最小的定位误差。在实际应用中，系统一般需要考虑准确性和实时性两方面。由于本文算法使用了轻量级的目标检测网络YOLOv7-tiny，在需要快速响应或资源有限的场景中，既可以保证准确性的同时，又可以保证实时性。综上，本文算法性能更优。

为了更进一步直观地展示本文算法的效果，本文展示了在不同动态环境中轨迹图的表现。图7展示了在高动态场景下，ORB-SLAM2算法、Dyna-SLAM与本文算法的估计轨迹与误差分布。其中，真实轨迹以黑色线条表示，估计轨迹用蓝色线条表示，红色线条则代表真实轨迹与估计轨迹之间的差异。仔细观察可见，在高动态环境中，ORB-SLAM2算法的估计轨迹和真实轨迹存在较大的差距，而本

表 7　绝对轨迹误差的 RMSE 对比

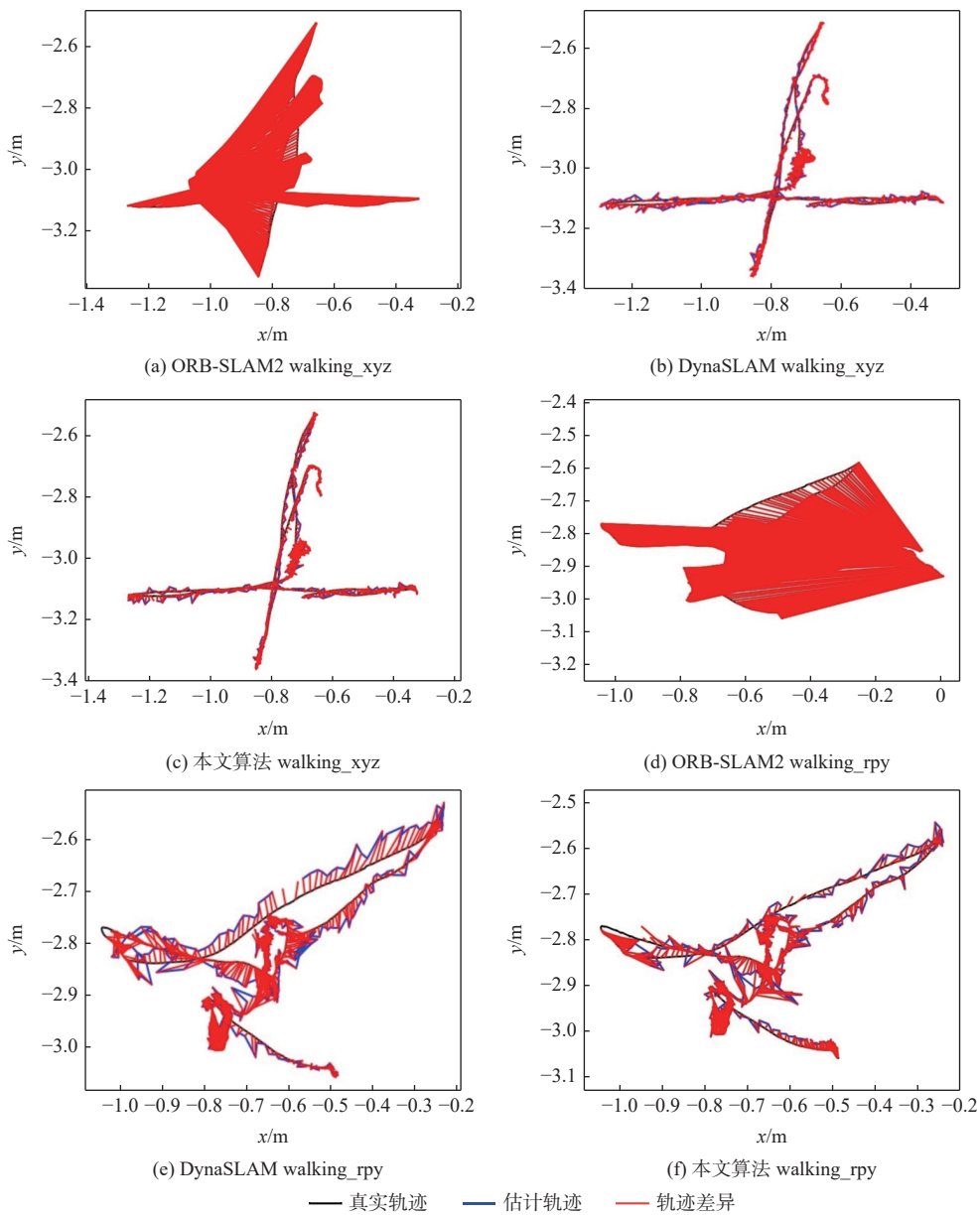
Table 7　RMSE comparison of absolute trajectory error

数据集名称	ORB-SLAM2[8]	YOLOv3-SLAM[29]	YOLOv4-SLAM[30]	DynaSLAM[17]	本文算法
fr3_sitting_static	0.149 578	0.357 965	0.343 385	0.078 521	0.210 825
fr3_sitting_xyz	0.017 513	0.020 901	0.023 578	0.022 394	0.017 827
fr3_sitting_halfsphere	0.032 686	0.029 224	0.025 969	0.025 076	0.025 955
fr3_sitting_rpy	0.144 598	0.276 151	0.299 480	0.253 892	0.328 201
fr3_walking_static	2.757 299	0.261 943	0.220 086	0.124 637	0.136 923
fr3_walking_xyz	1.440 868	0.017 448	0.019 748	0.020 895	0.017 363
fr3_walking_halfsphere	0.977 844	0.041 687	0.052 667	0.030 216	0.033 667
fr3_walking_rpy	2.283 019	0.101 932	0.205 203	0.079 428	0.076 706

表 8　相对位姿误差的 RMSE 的对比

Table 8　RMSE comparison of relative pose error

数据集名称	ORB-SLAM2[8]	YOLOv3-SLAM[29]	YOLOv4-SLAM[30]	DynaSLAM[17]	本文算法
fr3_sitting_static	0.005 814	0.006 678	0.006 538	0.006 379	0.006 153
fr3_sitting_xyz	0.011 067	0.012 211	0.012 060	0.012 741	0.011 950
fr3_sitting_halfsphere	0.011 109	0.028 343	0.030 124	0.018 423	0.025 955
fr3_sitting_rpy	0.016 804	0.016 104	0.016 180	0.020 151	0.016 260
fr3_walking_static	0.025 154	0.009 937	0.009 911	0.008 539	0.009 858
fr3_walking_xyz	0.032 623	0.015 785	0.015 911	0.014 958	0.015 203
fr3_walking_halfsphere	0.068 258	0.016 998	0.017 734	0.015 907	0.016 233
fr3_walking_rpy	0.029 868	0.022 164	0.022 131	0.024 993	0.022 125



(a) ORB-SLAM2 walking_xyz　(b) DynaSLAM walking_xyz

(c) 本文算法 walking_xyz　(d) ORB-SLAM2 walking_rpy

(e) DynaSLAM walking_rpy　(f) 本文算法 walking_rpy

图 7　高动态场景下估计轨迹和真实轨迹对比

Fig. 7　Comparison of real and estimated trajectories in high dynamic scenes

文算法和 DynaSLAM 算法都能够较为精确地估计和跟踪物体的运动轨迹。在图 8 中，可以观察到在低动态场景下，3 种算法的相机位姿估计轨迹与真实轨迹大致相符。综上，改进算法针对高动态场景下的优化更为显著，而在低动态场景中，其提升效果并不明显。

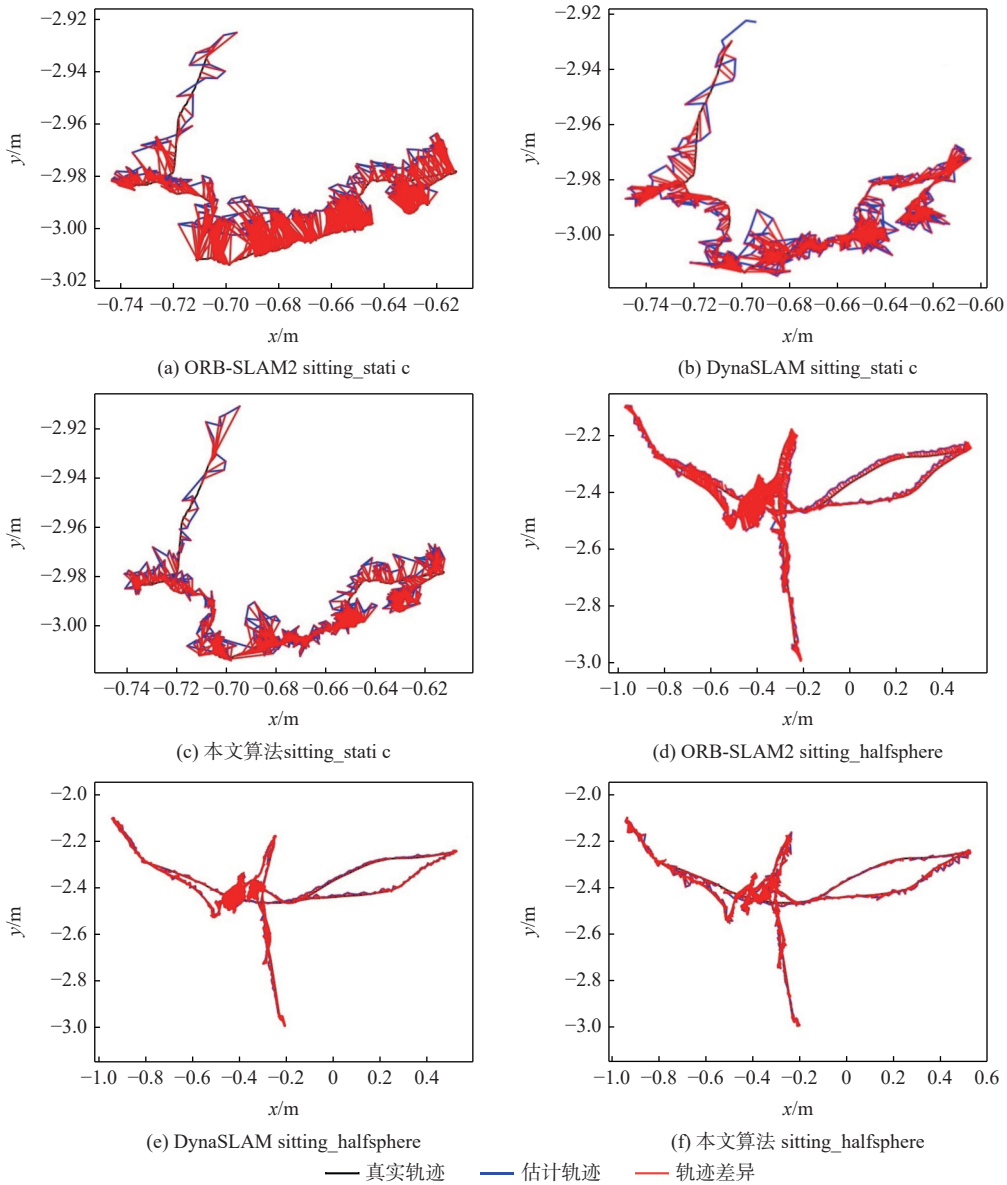


图 8 低动态场景下估计轨迹和真实轨迹对比

Fig. 8 Comparison of real and estimated trajectories in low dynamic scene

2.5 讨 论

在室内动态场景中，基于特征的 ORB-SLAM2 算法依赖于连续帧之间的特征匹配，而动态物体会影响相机姿态估计和地图生成的可靠性。同时，传统的固定阈值方法在光照变化较大时容易产生误检测或漏检测。针对以上问题，本文算法根据图像对比度自适应地调整阈值，在曝光不足和过曝的室内环境下提取效果显著提升，提高了动态物体检测的鲁棒性。相较于 YOLOv3-SLAM 和 YOLOv4-SLAM 算法，本文算法处理单帧图片的时间平均减少了 33.44% 和 36.35%。本文更高的运行效率可以满足动态场景下系统对于实时性的需求。但是，在室内极端光照条件下或者低动态场景中仍然存在很大的进步空间，未来可以结合图像增强算法或使用多传感器信息进行光照估计，以提供更准确的光照补偿，进而改善动态物体检测的性能。

3 结 论

本文研究动态场景下的 VSLAM 算法，旨在充分利用图像中的高层次语义信息，提高在动态环境中的 SLAM 算法的精度和鲁棒性。

1）在处理单帧图片耗时上，与 YOLOv3-SLAM 和 YOLOv4-SLAM 相比，本文算法单帧处理时间分别平均降低了 33.4% 和 36.4%，显著提升了系统运

行效率。

2）定位精度方面，在高动态环境下，本文算法在动态特征点剔除精度上优于 ORB-SLAM2，其绝对轨迹误差和相对位姿误差均显著降低。相较于基于 YOLOv3/YOLOv4 轻量化模型的动 SLAM 方法以及 DynaSLAM，本文算法在不同数据集下展现出更低的定位误差。得益于采用的轻量级目标检测网络 YOLOv7-tiny，本文算法在保证动态场景下定位准确性的同时，有效满足了系统实时性要求，综合性能更优。

未来本课题将会从以下方面开展研究：一方面，在之后的研究中可以通过融合激光雷达、IMU 传感器、轮速计等多种传感器信号，采用多模态传感器融合的方法来增强算法的鲁棒性和适用性。另一方面，轻量化语义 SLAM 模型还需要进一步优化，以保证在算力受限制的嵌入式平台上保持实时性。

致谢　感谢深圳市云视机器人有限公司对本工作在资金和实验设备上的大力支持 (HD-KYH-2 021 307)。

参考文献（References）

[1] CHEN W F, SHANG G T, JI A H, et al. An overview on visual SLAM: from tradition to semantic[J]. Remote Sensing, 2022, 14(13): 3010.

[2] AI Y B, RUI T, LU M, et al. DDL-SLAM: a robust RGB-D SLAM in dynamic environments combined with deep learning[J]. IEEE Access, 2020, 8: 162335-162342.

[3] YU L J, YANG E F, YANG B Y. AFE-ORB-SLAM: robust monocular VSLAM based on adaptive FAST threshold and image enhancement for complex lighting environments[J]. Journal of Intelligent & Robotic Systems, 2022, 105(2): 26.

[4] NEWCOMBE R A, LOVEGROVE S J, DAVISON A J. DTAM: dense tracking and mapping in real-time[C]//Proceedings of the 2011 International Conference on Computer Vision. Piscataway: IEEE Press, 2011: 2320-2327.

[5] FORSTER C, PIZZOLI M, SCARAMUZZA D. SVO: fast semi-direct monocular visual odometry[C]//Proceedings of the 2014 IEEE International Conference on Robotics and Automation. Piscataway: IEEE Press, 2014: 15-22.

[6] DAVISON A J, REID I D, MOLTON N D, et al. MonoSLAM: real-time single camera SLAM[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2007, 29(6): 1052-1067.

[7] KLEIN G, MURRAY D. Parallel tracking and mapping for small AR workspaces[C]//Proceedings of the 2007 6th IEEE and ACM International Symposium on Mixed and Augmented Reality. Piscataway: IEEE Press, 2007: 225-234.

[8] MUR-ARTAL R, TARDÓS J D. ORB-SLAM2: an open-source SLAM system for monocular, stereo, and RGB-D cameras[J]. IEEE Transactions on Robotics, 2017, 33(5): 1255-1262.

[9] CAMPOS C, ELVIRA R, RODRÍGUEZ J J G, et al. ORB-SLAM3: an accurate open-source library for visual, visual–inertial, and multimap SLAM[J]. IEEE Transactions on Robotics, 2021, 37(6): 1874-1890.

[10] SHEN X Q, CHEN L H, HU Z H, et al. A closed-loop detection algorithm for online updating of bag-of-words model[C]// Proceedings of the 2023 9th International Conference on Computing and Data Engineering. New York: ACM, 2023: 34-40.

[11] CHENG Y H, WANG J. A motion image detection method based on the inter-frame difference method[J]. Applied Mechanics and Materials, 2014, 490-491: 1283-1286.

[12] CUTLER R, DAVIS L S. Robust real-time periodic motion detection, analysis, and applications[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020, 22(8): 781-796.

[13] KLAPPSTEIN J, VAUDREY T, RABE C, et al. Moving object segmentation using optical flow and depth information[C]//Proceedings of the Advances in Image and Video Technology. Berlin: Springer, 2009: 611-623.

[14] DEROME M, PLYER A, SANFOURCHE M, et al. Moving object detection in real-time using stereo from a mobile platform[J]. Unmanned Systems, 2015, 3(4): 253-266.

[15] YU C, LIU Z X, LIU X J, et al. DS-SLAM: a semantic visual SLAM towards dynamic environments[C]//Proceedings of the 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems. Piscataway: IEEE Press, 2018: 1168-1174.

[16] BADRINARAYANAN V, KENDALL A, CIPOLLA R. SegNet: a deep convolutional encoder-decoder architecture for image segmentation[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(12): 2481-2495.

[17] BESCOS B, FÁCIL J M, CIVERA J, et al. DynaSLAM: tracking, mapping, and inpainting in dynamic scenes[J]. IEEE Robotics and Automation Letters, 2018, 3(4): 4076-4083.

[18] FU Y X, HAN B, HU Z H, et al. CBAM-SLAM: a semantic SLAM based on attention module in dynamic environment[C]//Proceedings of the 2022 6th Asian Conference on Artificial Intelligence Technology. Piscataway: IEEE Press, 2022: 1-6.

[19] CAI D P, HU Z H, LI R Q, et al. AGAM-SLAM: an adaptive dynamic scene semantic SLAM method based on GAM[C]//Proceedings of the Advanced Intelligent Computing Technology and Applications. Berlin: Springer, 2023: 27-39.

[20] HE K M, GKIOXARI G, DOLLÁR P, et al. Mask R-CNN[C]//Proceedings of the 2017 IEEE International Conference on Computer Vision. Piscataway: IEEE Press, 2017: 2980-2988.

[21] QI H, HU Z H, XIANG Y F, et al. ATY-SLAM: a visual semantic SLAM for dynamic indoor environments[C]//Proceedings of the Advanced Intelligent Computing Technology and Applications. Berlin: Springer, 2023: 3-14.

[22] CHANG Z Y, WU H L, SUN Y L, et al. RGB-D visual SLAM based on Yolov4-tiny in indoor dynamic environment[J]. Micromachines, 2022, 13(2): 230.

[23] WANG C Y, BOCHKOVSKIY A, LIAO H M. YOLOv7: trainable bag-of-freebies sets new state-of-the-art for real-time object detectors[C]//Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press,

2023: 7464-7475.

[24] GONZALES R C, WINTZ P. Digital image processing[M]. Reading: Addison-Wesley Longman Publishing Co., Inc., 1987.

[25] BURRI M, NIKOLIC J, GOHL P, et al. The EuRoC micro aerial vehicle datasets[J]. The International Journal of Robotics Research, 2016, 35(10): 1157-1163.

[26] VISWANATHAN D G. Features from accelerated segment test (fast)[C]// Proceedings of the 10th Workshop on Image Analysis for Multimedia Interactive Services. Piscataway: IEEE Press, 2009: 6-8.

[27] QADER W A, AMEEN M M, AHMED B I. An overview of bag of Words;Importance, implementation, applications, and challenges[C]// Proceedings of the 2019 International Engineering Conference. Piscataway: IEEE Press, 2019: 200-204.

[28] STURM J, ENGELHARD N, ENDRES F, et al. A benchmark for the evaluation of RGB-D SLAM systems[C]//Proceedings of the 2012 IEEE/RSJ International Conference on Intelligent Robots and Systems. Piscataway: IEEE Press, 2012: 573-580.

[29] REDMON J, FARHADI A. YOLOv3: an incremental improvement [EB/OL]. (2018-04-08)[2023-07-25]. https://doi.org/10.48550/arXiv.1804.02767.

[30] BOCHKOVSKIY A, WANG C Y, LIAO H M. YOLOv4: optimal speed and accuracy of object detection[EB/OL]. (2020-04-23)[2023-07-25]. https://doi.org/10.48550/arXiv.2004.10934.

A lightweight semantic VSLAM approach based on adaptive thresholding and speed optimization

QI Hao[1], FU Yuexin[1], HU Zhuhua[1,*], WU Jiaqi[1], ZHAO Yaochi[2]

(1. School of Information and Communication Engineering, Hainan University, Haikou 570228, China;
2. School of Cyberspace Security (School of Cryptology), Hainan University, Haikou 570228, China)

Abstract: Visual simultaneous localization and mapping (VSLAM) is a technology that utilizes visual and other sensory sensors to acquire information about unknown environments. It is widely applied in fields such as autonomous driving, robotics, augmented reality, and more. However, pixel-level semantic segmentation of dynamic objects entails high computing costs for indoor visual SLAM, and variations in lighting make dynamic items appear more difficult to see, potentially causing occlusions or confusion with the static surroundings. To address these challenges, a lightweight semantic VSLAM model is proposed, which is based on adaptive thresholding and velocity optimization. Initially, a lightweight one-stage object detection network, YOLOv7-tiny, is utilized in conjunction with the optical flow algorithm to effectively detect dynamic regions within images and filter out unstable feature points. Additionally, the feature point extraction algorithm dynamically adjusts the threshold based on the contrast information of the input images. Moreover, the combination of a binary bag-of-words method with a simplified optimization technique for local mapping threads improves the system's loading and matching speed in indoor dynamic scenarios. Experimental results show that the proposed algorithm can effectively eliminate dynamic feature points in indoor high-dynamic scenes, improving the positioning accuracy of the camera. The average processing speed reaches 19.8 frames per second (FPS), meeting real-time requirements in practical scenarios.

Keywords: visual simultaneous localization and mapping; dynamic scenes; YOLOv7-tiny; adaptive threshold; feature points

Received: 2023-08-28; **Accepted**: 2024-03-22; **Published Online**: 2024-05-08 10:50

URL: link.cnki.net/urlid/11.2625.V.20240507.1513.002

Foundation items: National Key Research and Development Program of China (2022YFD2400504); National Natural Science Foundation of China (62161010); Key Research and Development Project of Hainan Province (ZDYF2022SHFZ039,ZDYF2022GXJS348)

* **Corresponding author**. E-mail: eagler_hu@hainanu.edu.cn

《北京航空航天大学学报》征稿简则

《北京航空航天大学学报》是北京航空航天大学主办的以航空航天科学技术为特色的综合性自然科学学术期刊(月刊)。本刊以反映航空航天领域研究成果与动态、促进学术交流、培养科技人才和推动科技成果向社会生产力转化为办刊宗旨。本刊为中国自然科学技术核心期刊,并被 Ei Compendex 等国内外权威文献检索系统收录。本刊向国内外公开发行,为进一步提高办刊质量和所刊出文章的学术水平,特制定本简则。

1 论文作者及内容

1.1 本刊面向海内外所有学者。

1.2 主要刊载与航空航天科学技术有关的材料科学及工程、飞行器设计与制造、宇航科学与工程、信息与电子技术、控制技术和自动化工程、流体力学和动力工程、计算机科学及应用技术、可靠性工程与失效分析等领域的研究文章。航空航天科学技术民用方面以及具有航空航天工程背景的应用数学、应用物理、应用力学和工程管理等方面的文章也在本刊优先考虑之列。

2 来稿要求

2.1 论文应具有创新性、科学性、学术性和可读性。

2.2 论文为原创作品,尚未公开发表过,并且不涉及泄密问题。若发生侵权或泄密问题,一切责任由作者承担。

2.3 主题明确,数据可靠,图表清晰,逻辑严谨,文字精练,公式符号标准规范,标点符号正确。

2.4 文稿撰写顺序:中文题名(一般不超过 20 个汉字),作者中文姓名、单位、所在城市、邮政编码(作者顺序始终不得更改),中文摘要(包括目的、方法、结果及结论),中文关键词(5 ~ 8 个),中图分类号,引言,正文,参考文献,英文题名,作者英文姓名、单位、所在城市、邮政编码、国别,英文摘要,英文关键词。首、末页下角注明基金项目名称及编号,作者信息。

2.5 作者请登录本刊网页进行在线投稿。

3 稿件的审核、录用与版权

3.1 来稿须经专家两审和主编、编委讨论后决定刊用与否。

3.2 若来稿经审查后认定不宜在本刊发表,将及时告知作者。如果在投稿满 3 个月后仍未收到本刊任何通知,作者有权改投其他刊。在此之前,请勿一稿多投,否则一切后果自负。

3.3 来稿一经作者签字并在本刊刊出,即表明所有作者都已经认可其版权转至本刊编辑部。本刊在与国内外文献数据库或检索系统进行交流及合作时,不再征询作者意见。

邮寄地址:100191 北京市海淀区学院路 37 号 《北京航空航天大学学报》编辑部

办公地点:北京航空航天大学办公楼西配楼412~416房间

电　　话:(010)82316698,82316981,82315499,82317448

E-mail: jbuaa@ buaa. edu. cn

http://bhxb. buaa. edu. cn

http://www. buaa. edu. cn

IISTJA 工信科技期刊集群 集群成员期刊

《北京航空航天大学学报》第五届编辑委员会

主　任（主　编）：赵沁平

（以下按姓氏笔画为序）

副主任（副主编）：丁希仑　王少萍　孙志梅　李秋实　李焕喜　杨嘉陵
苗俊刚　相　艳　徐立军　钱德沛　曹晋滨

编　委：马殿富　王　琪　王　聪　邓小燕　王青云　王荣明　刘　宇
刘　红　江　洁　刘　强　闫　鹏　朱天乐　刘铁钢　齐铂金
陈万春　邹正平　苏东林　杨世春　沈成平　邱志平　宋知人
杨树斌　张晓林　杨晓奕　杨继萍　李惠峰　吴新开　张瑞丰
杨照华　宋凝芳　周　锐　林宇震　林贵平　战　强　姚仰平
胡庆雷　赵秋红　段海滨　赵巍胜　席　平　郭　宏　徐　洁
徐世杰　郭洪波　康　锐　翟　锦　熊华钢

编辑部主任：张　嵘

编　　辑：孙　芳　万玉敏　王　茜　卞欢欢　李亚泰

北京航空航天大学学报
Beijing Hangkong Hangtian Daxue Xuebao
（原《北京航空学院学报》）
（月刊　1956年创刊）
第51卷 第7期　2025年7月

JOURNAL OF BEIJING UNIVERSITY OF AERONAUTICS AND ASTRONAUTICS (JBUAA)
(Monthly, Started in 1956)
Vol.51 No.7　July 2025

主管单位　中华人民共和国工业和信息化部
主办单位　北京航空航天大学
主　　编　赵沁平
编辑出版　《北京航空航天大学学报》编辑部
邮　　编　100191
地　　址　北京市海淀区学院路37号
印　　刷　北京科信印刷有限公司
发　　行　《北京航空航天大学学报》编辑部
发行范围　国内外发行
联系电话　(010) 82316698 82316981
82315499 82317448
电子信箱　jbuaa@buaa.edu.cn

Administrated by Ministry of Industry and Information Technology of the People's Republic of China
Sponsored by Beijing University of Aeronautics and Astronautics (BUAA)
(Beijing 100191, P. R. China)
Chief Editor ZHAO Qinping
Edited and Published by Editorial Board of JBUAA
Printed by Beijing Kexin Printing Co., Ltd.
Distributed by Editorial Board of JBUAA
Telephone (010) 82316698 82316981
82315499 82317448
E-mail jbuaa@buaa.edu.cn
http://bhxb.buaa.edu.cn

中国标准连续出版物号：ISSN 1001-5965
CN 11-2625/V

ISSN 1001-5965
9 771001 596250 07>

微信公众号　超星电子书　QQ群

国内定价：50.00元／期