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

2024-02
Vol.50 No.2

北京航空航天大学学报

第 50 卷　第 2 期　（总第 372 期）　2024 年 2 月

目　　次

视觉 SLAM 方法综述 …… 王朋，郝伟龙，倪翠，张广渊，巩慧（359）

基于多模态联合语义感知的零样本目标检测 …… 段立娟，袁鍾，王文健，梁芳芳（368）

基于顶点与主体区域同步检测的精准车牌定位 ……

…… 徐光柱，刘高飞，匡婉，万秋波，马国亮，雷帮军（376）

基于 EM 自注意力残差的图像超分辨率重建网络 …… 黄淑英，胡瀚洋，杨勇，万伟国，吴峥（388）

融合粗糙数据推理的离散麻雀搜索算法求解 HFSP 问题 …… 周宁，张嵩霖，张晨（398）

面向语法加权图文本的方面情感三元组抽取 …… 韩虎，孟甜甜（409）

结合空间注意力机制的实时鲁棒视觉跟踪 …… 马素刚，张子贤，蒲磊，侯志强（419）

基于跨模态近邻损失的可视-红外行人重识别 …… 赵三元，阿琪，高宇（433）

基于 Transformer 的深度条件视频压缩 …… 鲁国，钟天雄，耿晶（442）

基于优化布拉格结构的固态装配型谐振器 …… 张世峰，轩伟鹏，石林豪，董树荣，浦世亮（449）

基于差异化和空间约束的自动图像描述模型 …… 姜文晖，陈志亮，程一波，方玉明，左一帆（456）

融合卷积注意力和 Transformer 架构的行人重识别方法 …… 王静，李沛橦，赵容锋，张云，马振玲（466）

周期规律增强的多视角短期电力负荷预测 …… 苏伟，肖小龙，史明明，方鑫，司鑫尧（477）

基于深度强化学习的办公流程任务分配优化 …… 廖晨阳，于劲松，乐祥立（487）

改进 YOLOv5s 的弱光水下生物目标检测算法 …… 陈宇梁，董绍江，孙世政，闫凯波（499）

基于 K-means 聚类的多种群麻雀搜索算法 …… 闫少强，刘卫东，杨萍，吴丰轩，阎哲（508）

基于通信连通保持约束的多智能体覆盖控制 …… 张蕴霖，麻壮壮，石磊，邵晋梁（519）

基于变形图匹配的知识图谱多跳问答 …… 李香粤，方全，胡骏，钱胜胜，徐常胜（529）

基于 Dijkstra 算法的平滑路径规划方法 …… 巩慧，倪翠，王朋，程诺（535）

语言引导的多粒度特征融合目标分割方法 …… 谭荃戈，王蓉，吴澳（542）

基于多级语义对齐的图像-文本匹配算法 …… 李艺茹，姚涛，张林梁，孙玉娟，付海燕（551）

基于多级联合的图池化方法 …… 董晓龙，黄俊，秦锋，洪旭东（559）

联合方面注意力交互的图文方面类情感识别 …… 赵一成，王素格，廖健，何东欢（569）

基于多模态掩码 Transformer 网络的社会事件分类 …… 陈宏，钱胜胜，李章明，方全，徐常胜（579）

区域感知实时人像超分辨率重建网络 …… 龚柯存，周梦琳，唐东明（588）

基于多模态双向信息增强的 RGBT 跟踪网络 …… 赵伟，刘磊，王鲲鹏，涂铮铮，罗斌（596）

基于 Transformer 和动态 3D 卷积的多源遥感图像分类 …… 高峰，孟德森，解正源，亓林，董军宇（606）

自注意力相似度迁移跨模态哈希网络 …… 梁焕，王海荣，王栋（615）

多输入傅里叶神经网络及其麻雀搜索优化 …… 黎亮亮，张著洪，张永丹（623）
基于 Transformer 的点云几何有损压缩方法 …… 刘舸昕，章骏腾，丁丹丹（634）
膜内麻雀优化 ELM 的软件缺陷预测算法 …… 唐宇，代琪，杨梦园，陈丽芳（643）
基于注意力机制与条件卷积的行人重识别方法 …… 姬广凯，王蓉，彭舒凡（655）
基于损失平滑的对抗样本攻击方法 …… 黎妹红，金双，杜晔（663）
基于布局化-语义联合表征遥感图文检索方法 …… 张若愚，聂婕，宋宁，郑程予，魏志强（671）
图文数据的多级关系分析与挖掘方法 …… 郭瑞萍，王海荣，王栋（684）

期刊基本参数：CN 11-2625/V * 1956 * m * A4 * 336 * zh * P * ¥ 50.00 * 400 * 35 * 2024-02

（**编　辑**　张　嵘　孙　芳　苏　磊　卞欢欢　王　茜　李亚泰）

JOURNAL OF BEIJING UNIVERSITY OF
AERONAUTICS AND ASTRONAUTICS

Vol. 50 No. 2 （Sum 372） February 2024

CONTENTS

An overview of visual SLAM methods
…… WANG Peng，HAO Weilong，NI Cui，ZHANG Guangyuan，GONG Hui （359）
Zero-shot object detection based on multi-modal joint semantic perception
…… DUAN Lijuan，YUAN Ying，WANG Wenjian，LIANG Fangfang （368）
Accurate license plate location based on synchronous vertex and body region detection
…… XU Guangzhu，LIU Gaofei，KUANG Wan，WAN Qiubo，MA Guoliang，LEI Bangjun （376）
Image super-resolution reconstruction network based on expectation maximization self-attention residual
…… HUANG Shuying，HU Hanyang，YANG Yong，WAN Weiguo，WU Zheng （388）
Discrete sparrow search algorithm incorporating rough data-deduction for solving hybrid flow-shop scheduling problems
…… ZHOU Ning，ZHANG Songlin，ZHANG Chen （398）
Aspect sentiment triple extraction for grammar-weighted graph text
…… HAN Hu，MENG Tiantian （409）
Real-time robust visual tracking based on spatial attention mechanism
…… MA Sugang，ZHANG Zixian，PU Lei，HOU Zhiqiang （419）
Cross-modality nearest neighbor loss for visible-infrared person re-identification
…… ZHAO Sanyuan，A Qi，GAO Yu （433）
A Transformer based deep conditional video compression
…… LU Guo，ZHONG Tianxiong，GENG Jing （442）
Solidly mounted resonator based on optimized Bragg structure
…… ZHANG Shifeng，XUAN Weipeng，SHI Linhao，DONG Shurong，PU Shiliang （449）
Image captioning model based on divergence-based and spatial consistency constraints
…… JIANG Wenhui，CHEN Zhiliang，CHENG Yibo，FANG Yuming，ZUO Yifan （456）
A person re-identification method for fusing convolutional attention and Transformer architecture
…… WANG Jing，LI Peitong，ZHAO Rongfeng，ZHANG Yun，MA Zhenling （466）
Periodic pattern-enhanced multi-view short-term load prediction
…… SU Wei，XIAO Xiaolong，SHI Mingming，FANG Xin，SI Xinyao （477）
Optimization of office process task allocation based on deep reinforcement learning
…… LIAO Chenyang，YU Jinsong，LE Xiangli （487）
Improved YOLOv5s low-light underwater biological target detection algorithm
…… CHEN Yuliang，DONG Shaojiang，SUN Shizheng，YAN Kaibo （499）
Multi group sparrow search algorithm based on K-means clustering
…… YAN Shaoqiang，LIU Weidong，YANG Ping，WU Fengxuan，YAN Zhe （508）
Multi-agent coverage control based on communication connectivity maintenance constraints
…… ZHANG Yunlin，MA Zhuangzhuang，SHI Lei，SHAO Jinliang （519）
Multi-hop knowledge graph question answering based on deformed graph matching
…… LI Xiangyue，FANG Quan，HU Jun，QIAN Shengsheng，XU Changsheng （529）
A smooth path planning method based on Dijkstra algorithm
…… GONG Hui，NI Cui，WANG Peng，CHENG Nuo （535）
Language-guided target segmentation method based on multi-granularity feature fusion
…… TAN Quange，WANG Rong，WU Ao （542）
Image-text matching algorithm based on multi-level semantic alignment
…… LI Yiru，YAO Tao，ZHANG Linliang，SUN Yujuan，FU Haiyan （551）

Graph pooling method based on multilevel union
…… DONG Xiaolong，HUANG Jun，QIN Feng，HONG Xudong （559）
Image-text aspect emotion recognition based on joint aspect attention interaction
…… ZHAO Yicheng，WANG Suge，LIAO Jian，HE Donghuan （569）
Multi-modal mask Transformer network for social event classification
…… CHEN Hong，QIAN Shengsheng，LI Zhangming，FANG Quan，XU Changsheng （579）
Region-aware real-time portrait super resolution reconstruction network
…… GONG Kecun，ZHOU Menglin，TANG Dongming （588）
Multimodal bidirectional information enhancement network for RGBT tracking
…… ZHAO Wei，LIU Lei，WANG Kunpeng，TU Zhengzheng，LUO Bin （596）
Multi-source remote sensing image classification based on Transformer and dynamic 3D-convolution
…… GAO Feng，MENG Desen，XIE Zhengyuan，QI Lin，DONG Junyu （606）
Cross-modal hashing network based on self-attention similarity transfer
…… LIANG Huan，WANG Hairong，WANG Dong （615）
Multi-input Fourier neural network and its sparrow search optimization
…… LI Liangliang，ZHANG Zhuhong，ZHANG Yongdan （623）
Lossy point cloud geometry compression based on Transformer
…… LIU Gexin，ZHANG Junteng，DING Dandan （634）
Software defect prediction algorithm for intra-membrane sparrow optimizing ELM
…… TANG Yu，DAI Qi，YANG Mengyuan，CHEN Lifang （643）
Person re-identification method based on attention mechanism and CondConv
…… JI Guangkai，WANG Rong，PENG Shufan （655）
Adversarial attack method based on loss smoothing
…… LI Meihong，JIN Shuang，DU Ye （663）
Remote sensing image-text retrieval based on layout semantic joint representation
…… ZHANG Ruoyu，NIE Jie，SONG Ning，ZHENG Chengyu，WEI Zhiqiang （671）
Multilevel relation analysis and mining method of image-text
…… GUO Ruiping，WANG Hairong，WANG Dong （684）

http://bhxb.buaa.edu.cn jbuaa@buaa.edu.cn
DOI: 10.13700/j.bh.1001-5965.2022.0376

视觉 SLAM 方法综述

王朋[1,2]，郝伟龙[1]，倪翠[1,*]，张广渊[1]，巩慧[1]

(1. 山东交通学院 信息科学与电气工程学院，济南 250357；2. 山东省科学院自动化研究所，济南 250013)

摘　　要：实时定位与建图 (SLAM) 技术搭载特定传感器，使移动机器人在无任何环境先验条件下，在运动过程中自主建立环境模型来计算自身位姿，大幅提高其自主导航能力，以及对不同应用环境的适应性。视觉 SLAM 方法以相机作为外部传感器，通过采集周围环境信息来创建地图并实时估计机器人自身位姿。为此，介绍了具有代表性的经典视觉 SLAM 方法及与深度学习相结合的视觉 SLAM 方法，分析了视觉 SLAM 方法中采用的不同特征检测方法、后端优化、闭环检测，以及动态环境下视觉 SLAM 方法的应用，总结了视觉 SLAM 方法的问题，并探讨了视觉 SLAM 方法在未来的热点研究方向和发展前景。

关 键 词：视觉实时定位与建图；深度学习；特征检测；位姿估计；闭环检测

中图分类号：TP391.4

文献标志码：A　　**文章编号**：1001-5965(2024)02-0359-09

当移动机器人被放置在一个未探测过的环境中时，首先要确定所在的环境地图及在该环境地图中的位置，该问题于 1986 年被提出并命名为实时定位与建图(simultaneous localization and mapping, SLAM)问题，由于其具有非常重要的理论价值和应用价值，SLAM 迅速成为热点研究方向，被认为是实现全自主移动机器人的关键。早期的 SLAM 主要是激光 SLAM，拥有精度高、解决方案较为成熟的优点，但其主要适用于室内环境，且存在体积大、价格昂贵和感知信息少的缺点；而视觉 SLAM 使用相机作为主传感器，用拍摄的视频流作为输入，实现实时定位与建图，能够同时适用于室内和室外环境，且成本较低[1]。随着计算机视觉的快速发展，视觉 SLAM 以其信息获取及时、信息量丰富等优势逐渐成为研究热点。自 21 世纪以来，不断有学者提出新的视觉 SLAM 方法，扩充视觉 SLAM 领域中的相关技术，为该领域提供了更加广阔的研究思路。

本文以时间顺序回顾现有的视觉 SLAM 方法，重点关注其优势和局限性展开介绍，探讨未来视觉 SLAM 领域的挑战和前景，以促进更多创新研究，推动移动机器人自主导航技术的发展。

1　经典视觉 SLAM 方法

视觉 SLAM 方法的工作流程主要分为五大部分：信息采集、前端视觉里程计、后端优化、闭环检测、建图，如图 1 所示。前端视觉里程计通过提取、匹配图像数据中的特征点得到路标点，并估计出相机当前的位姿，实现实时定位。期间存在的累积误差由后端优化根据帧间的共视关系等方法进行消除，提高跟踪与建图的精度。位姿估计期间，如果有 2 帧图像的相似度达到一定标准，则认为形成了闭环，对 2 帧图像之间的所有路标点和位姿进行一次优化，经过优化处理后的路标点最终用于建图，建图的精度直接影响后续动态路径规划、实时避障等工作的效果。

经典视觉 SLAM 方法常用的传感器主要有单

收稿日期：2022-05-18；**录用日期**：2022-06-23；**网络出版时间**：2022-10-10 10:03
网络出版地址：link.cnki.net/urlid/11.2625.V.20221009.1040.009
基金项目：中国博士后科学基金 (2021M702030)；山东省交通运输厅科技计划 (2021B120)
* **通信作者**. E-mail：emilync@126.com

引用格式：王朋，郝伟龙，倪翠，等. 视觉 SLAM 方法综述 [J]. 北京航空航天大学学报，2024，50（2）：359-367.
WANG P，HAO W L，NI C，et al. An overview of visual SLAM methods [J]. Journal of Beijing University of Aeronautics and Astronautics，2024，50（2）：359-367（in Chinese）.

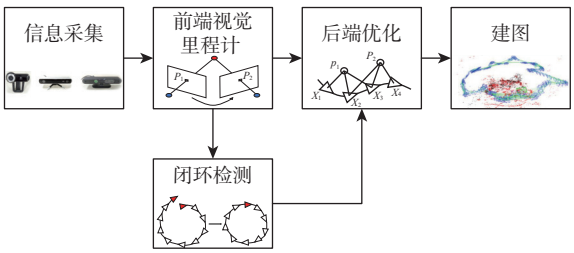


图1 视觉SLAM方法流程

Fig. 1 Flow of visual SLAM methods

目、双目和RGB-D相机。单目视觉SLAM方法基于单目相机完成,该相机结构简单、成本低,相机与物体进行相对运动时,与相机距离不同的物体在图像序列中产生的位移也不同,据此物体在相机视野中形成了视差,依据视差可以判断相机距离不同物体的远近关系,但是难以获得相机与物体的真实距离。为解决无法得到图像深度的问题,人们提出双目视觉SLAM方法和RGB-D SLAM方法。双目视觉SLAM方法由2个单目相机构成,2个相机之间的距离称为基线,基线是已知的,通过对2个相机同一场景的拍摄情况进行对比计算,可以得到相机到每一个像素之间的距离,再通过2个相机形成的视差计算得到任意像素间的距离,该方法能够同时适应室内和室外环境,受环境因素的影响较小,但是测量到的深度范围大小与基线大小呈正相关,通常情况下传感器体积较大,计算视差的过程较为复杂,导致实时性较低。因此,降低计算量是双目视觉SLAM方法的重点问题之一。RGB-D SLAM方法使用激光传感器,利用红外结构光原理或飞行时间(time of flight, TOF)法,向物体投射激光并接收返回的激光来测量相机到物体的距离,由于该距离直接由激光传感器物理测量得到,不需要像双目视觉SLAM方法一样消耗大量计算资源,但是存在成本高、测量范围小、噪声大、易受光照强度等外界环境影响的问题。

1.1 MonoSLAM

2007年,Davison等提出的MonoSLAM[2]通过在每帧图像中提取稀疏的特征点来进行位姿估计,可以在概率框架内在线运行,这是第1个使用单目相机完成的实时视觉SLAM系统,是基于扩展卡尔曼滤波器(extended Kalman filter, EKF)的视觉SLAM方法,具有良好的实时性。实验表明,在小规模运动中,MonoSLAM能将误差限制在1~2 cm之间。MonoSLAM的诞生标志着视觉SLAM方法的研究由理论转向了实际,具有里程碑的意义。但是,MonoSLAM仅用了单个线程,特征点的提取和匹配、相机位姿的估计和建图工作都是逐帧进行,而且更新的计算复杂度很高,为了提高SLAM速度,每帧图像中只能处理约10个特征点,且特征点容易丢失。

1.2 PTAM

2007年,Klein和Murray提出了平面跟踪与地图构建[3](parallel tracking and mapping, PTAM)。为解决MonoSLAM计算复杂度高的问题,PTAM对视觉里程计做了优化,不再逐帧提取特征点,而是在一部分图像序列中选出最具有代表性的一帧作为关键帧,只处理关键帧中的特征点,降低了计算复杂度,首次提出Tracking、Mapping的前后端双线程架构,如图2所示。Tracking线程使用基于加速段测试的特征[4](features from accelerated segment test, FAST)作为特征提取方法,只需更新相机位姿,不受建图线程计算成本的影响,易于实现实时计算。为提高计算速度,PTAM还在该线程中使用粗细2轮求解来估计相机位姿。第1轮粗求解只从图像金字塔中的第1层图像提取少量特征点,将估计出的相机位姿作为初始值传入第2轮的细求解中,第2轮细求解在整个图像金字塔中匹配特征点,有效保证了特征点的尺度不变性,提高了跟踪精度。Mapping线程使用光束法平差(bundle adjustment, BA)进行优化,与滤波法相比,BA是一种基于非线性优化的算法,其在单位时间内可处理的特征点数量远大于滤波法,相比于滤波法具有更高的精度。

PTAM是第1个使用非线性优化的方法,也是第1个将BA与视觉SLAM相结合的方法。在PTAM中,首次出现了前后端的概念,为后来出现的视觉SLAM方法提供了良好思路。但是,PTAM缺少闭环检测模块,无法对相机位姿和地图进行更好的优化。

1.3 DTAM

2011年,Newcombe等提出了单目视觉SLAM系统稠密跟踪与建图[5](dense tracking and mapping, DTAM)。该方法属于直接稠密法,其延续了处理关键帧的思想,但是传统的视觉SLAM方法在关键帧上提取特征点需要消耗大量计算资源,有利于提高系统鲁棒性,但不利于系统的实时性,而DTAM直接对关键帧中的每一个像素的信息进行分析计算,通过最小化全局空间规范能量函数来计算关键帧,实时重建三维模型,根据图像灰度信息、最小化光度误差来优化相机运动。DTAM不需要单独对特征点和描述子进行详细的计算,既减少了特征点的计算时间,也避免了弱纹理场景中特征点缺失导致的跟踪失败问题。DTAM具有稳定的定位与建图效果,但其在默认光照环境不变的条件下进行,

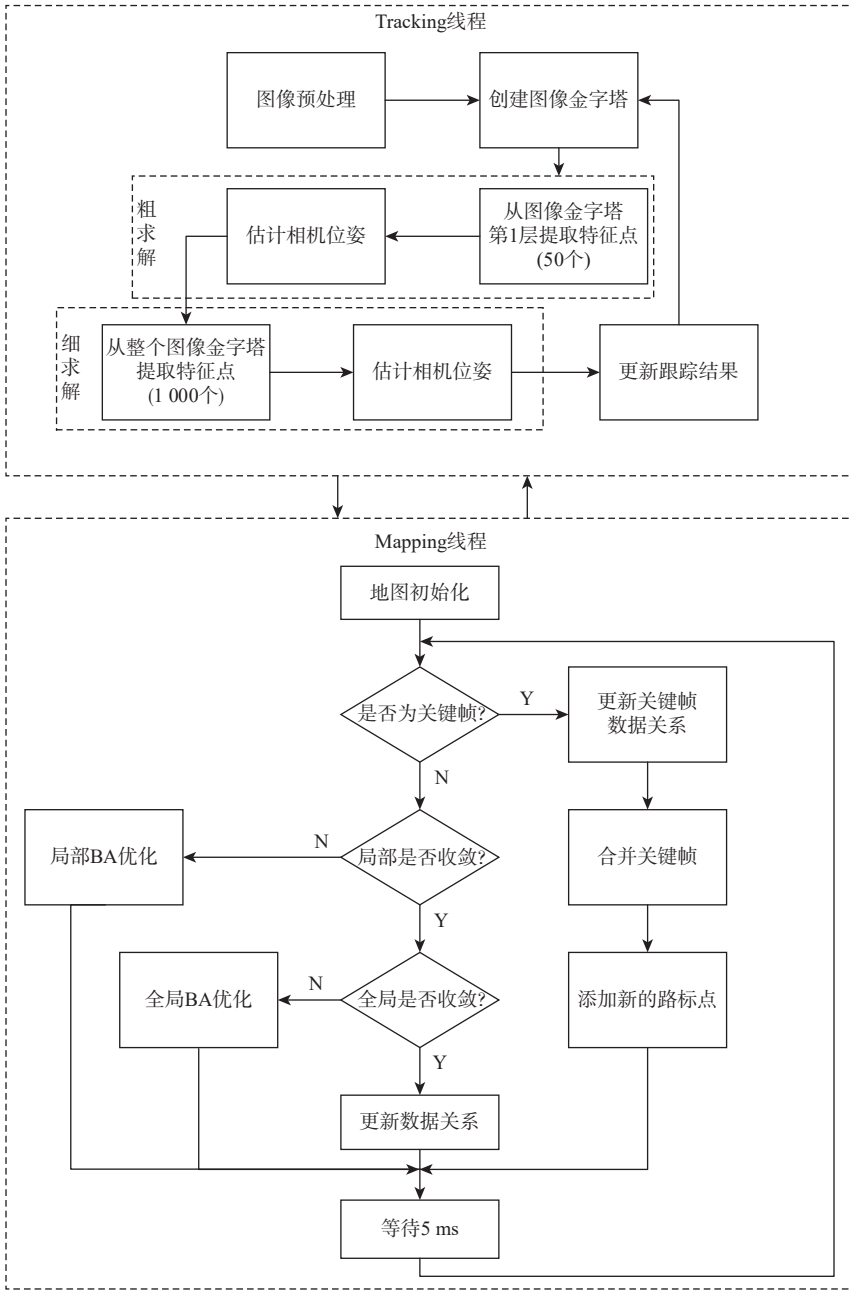


图 2　PTAM 框架

Fig. 2　PTAM framework

对光照强度变化比较敏感。

1.4　LSD-SLAM

2014 年, Engel 等提出了大规模直接单目 SLAM[6] (large-scale direct monocular SLAM, LSD-SLAM), 是半稠密的单目视觉 SLAM 方法, 之后又将其成功移植到了双目和 RGB-D 相机上。该方法不使用特征点, 将关注点放到了像素梯度与直接法的关系上, 提取梯度较为明显的像素(通常情况为物体的边缘线条), 使用方差归一化后的光度误差对图像的梯度进行跟踪, 生成半稠密深度图。LSD-SLAM 通过对当前帧和当前关键帧进行匹配得到当前帧的位姿, 判断当前帧与当前关键帧之间的距离是否大于某一阈值来决定是否将当前帧设置为当前关键帧, 如果当前帧被设置为新的关键帧, 则将上一个关键帧的深度图投影到当前帧来为新关键帧初始化深度图, 通过闭环检测进行位姿和地图的优化。实验表明, 在稳定的运动环境下, LSD-SLAM 可以将相机轨迹的绝对误差限制在 2 cm 内。但目前来说, 半稠密法获得的点云数量仍然不足, 存在尺度漂移、对相机曝光参数非常敏感、相机快速移动时容易丢失目标等问题, 且无基于直接法的闭环检测, 仍然需要以特征点法的方式进行闭环检测。

1.5　SVO

2014 年, Forster 等提出了半直接单目视觉里程

计[7](semi-direct monocular visual odometry, SVO),其结合特征点法和直接法的优点,是稀疏直接法视觉里程计,并使用提出的深度滤波器估计特征点的位置。该方法在图像中找到稀疏的特征点,将特征点周围一定范围内的像素用直接法进行特征点匹配,估计相机的运动轨迹及位姿。由于只寻找了稀疏的特征点,且不需要对特征描述子进行计算,SVO的速度很快,能够达到100 Hz,2016年提出的SVO2.0能够达到400 Hz,因此适用于计算平台受限的情况(如无人机和低端计算平台)。但是,SVO也存在明显的局限性。由于该方法的目标应用平台是无人机的俯视相机,而相机的主要运动方式是水平平移,且没有考虑相机视角旋转的情况,在其他常见的应用场景中表现不佳。由于缺少后端优化和闭环检测功能,对相机位姿的估计具有较为明显的累积误差,非常依赖于位姿估计的准确性,一旦跟踪失败很难重新定位。因此,SVO的建图功能非常有限,不算完整的SLAM。

1.6 ORB-SLAM

2015年,Mur-Artal等提出了基于单目相机的视觉SLAM系统ORB-SLAM[8],2017年提出了ORB-SLAM2[9],在上一代的基础上增加了对双目和RGB-D相机的支持,具有广泛的适用性。2021年,Campos等提出了ORB-SLAM3[10],实现了相机与惯性测量单元(inertial measurement unit, IMU)的结合,通过外加传感器提升了视觉SLAM方法的性能。ORB-SLAM在传统PTAM框架下添加了闭环检测模块,使用了Tracking、Local Mapping、Loop Closing三线程并行的方案。Tracking线程负责提取图像中的ORB[11]特征点,与上一关键帧进行比较,计算特征点的位置来估计相机当前的位姿,并适时地根据一定规则选取新的关键帧,每个ORB特征点都有可能对应一个路标点,而同一个路标点也可能会对应多个帧中不同的ORB特征点,关键帧和路标点组成了完整的地图。Local Mapping线程负责接收Tracking线程选出的关键帧并进行处理,将冗余的局部关键帧和质量较差的路标点删除,并通过三角化生成新的路标点,进行BA优化。Loop Closing线程负责接收Local Mapping线程筛选后的关键帧,与其他所有关键帧进行闭环检测,更新相机位姿和地图。ORB-SLAM围绕ORB特征进行计算,大大提高了计算效率,具有良好的尺度不变性和旋转不变性,ORB特征点中包括描述子,在相机进行大范围运动时能较好地完成闭环检测和重定位。不同于单目视觉SLAM方法,ORB-SLAM在单目条件下能够自动选择合适的初始化方案,而在双目和RGB-D相机条件下初始化速度很快,且原地不动就能完成初始化。闭环检测大大提高了定位和建图的准确性。在关键帧选取时,采用较为宽松的选取方案,提高了跟踪定位过程的鲁棒性,再删除效果不佳的关键帧,以提高闭环检测精度和BA优化效率。在静态环境下,定位准确且系统鲁棒性强,能够达到实时效果。经过多组实验证明,ORB-SLAM的平均关键帧轨迹均方根误差为1.57 cm,且在相机快速移动的状态下仍然可以保持良好的跟踪状态,鲁棒性远大于PTAM和LSD-SLAM,但是,单目跟踪在缺少闭环条件的大运动场景中会产生较大的漂移量[12]。ORB-SLAM对每帧图像都进行ORB特征点提取,非常耗时,地图由稀疏特征点组成,精度较低,且对动态环境较为敏感,在动态环境下容易跟踪失败。

1.7 DSO

2017年,Engel等提出了直接稀疏里程计[13](direct sparse odometry, DSO),其是基于单目相机的稀疏直接法视觉里程计,无论是鲁棒性、准确性还是速度上,都优于LSD-SLAM和ORB-SLAM。不同于特征点法需要找到某一特征点与哪些帧中的点相匹配,DSO将每个点投影到所有帧中,计算其在各帧中的残差,将残差在同一合理范围内的点视作同一个点的投影,而并不关心这些点之间的对应关系,通常情况下,系统的运行速度很快,可以生成非常稠密的点云。后端是由多个关键帧组成的滑动窗口,需要对滑动窗口和与优化相关的结构进行维护。为降低光照对视觉SLAM方法性能造成的影响,DSO提出光度标定,可以降低光照对跟踪的影响,有效提高了系统的鲁棒性。但是,DSO虽然在速度上相比之前的方法有很大的优势,但是与SVO相同,其没有闭环检测,跟踪失败后难以重新定位。

2 结合深度学习的视觉SLAM方法

深度学习通过模拟人脑神经元的结构构建复杂的神经网络模型,利用大量数据进行训练,模拟人的学习过程。早期,大部分视觉SLAM方法都是围绕静态环境进行工作,动态场景下使用相当困难。但是,通过深度学习可以实现计算机对物体的识别,还可以完成像素级的语义分割,对SLAM识别动态物体有极大的帮助,使得视觉SLAM方法在动态场景中也能正常高效的工作。通过深度学习,计算机甚至可以估计相机位姿等更为复杂的问题。因此,视觉SLAM方法与深度学习的结合成为了近年来热门的研究方向。

2.1 UnDeepVO

2018 年, Li 等提出了无监督深度学习的单目视觉里程计[14](monocular visual odometry through unsupervised deep learning, UnDeepVO), 其基于无监督深度学习设计。系统主体由位姿估计器和深度估计器 2 部分构成。位姿估计器本质上是基于 VGG 架构的卷积神经网络, 深度估计器主要基于编码-解码得到图像深度, 相比利用视差计算深度的方法, 该方法可使系统更易收敛。首先, 通过立体影像及其尺度来训练 UnDeepVO; 然后, 利用连续的单目影像进行测试, 不仅可以估计相机的位姿, 而且还可得到真实尺度的深度图。通过多组实验可知, 在 100～800 m 之间, UnDeepVO 的平均平移均方根误差为 4.07%, 平均旋转均方根误差为 2.02(°)/100 m, 而作为对照组, ORB-SLAM 在不使用闭环检测功能时, 2 项数据分别为 27.05% 和 10.23(°)/100 m。

2.2 DynaSLAM

2018 年, Bescos 等提出了 DynaSLAM[15], 该系统以 ORB-SLAM2 为基础, 利用实例分割和多视图几何的方法识别动态区域并提取, 有效提升了 SLAM 在动态场景下的性能。DynaSLAM 是 SLAM 与深度学习相结合的经典动态视觉 SLAM 系统, 使用 Mask R-CNN 网络进行实例分割, 检测出有潜在运动可能的物体, 利用多视图几何方法检测动态特征点, 通过区域生长法进一步得到动态区域, 将二者融合后检测出有潜在运动可能的物体[16], 剔除动态区域, 只使用静态区域的 ORB 特征点进行跟踪与建图, 并修复背景, 将相邻关键帧的图像投影到当前关键帧缺失的动态区域上进行修补。但是, DynaSLAM 将所有具有潜在运动可能的物体剔除(如静止的汽车、椅子等), 当图像中的优质特征点过少时, 将影响位姿估计的效果。

2.3 DSP-SLAM

2021 年, Wang 等提出了具有深度形状先验的面向对象 SLAM[17](object oriented SLAM with deep shape priors, DSP-SLAM), 其是以 ORB-SLAM 为骨架的面向对象的 SLAM 系统, 在单目、双目、RGB-D 相机下可以达到 10 Hz 的工作速度。该方法的创新点在于: 将地图分为实例对象和背景 2 部分, 通过深度学习的语义分割将实例对象检测出来, 并通过新的二阶优化算法, 以特定类别的深度形状作为先验估计目标实例的形状和位姿, 即使对新见到的物体也可以精确而完整的重建, 背景环境通过稀疏特征点构建, 在联合因子图中对二者进行优化, 不仅保证了环境地图建立的实时性, 还为前景目标对象构建了丰富而精确的稠密 3D 模型。

3 视觉 SLAM 方法的主要问题

3.1 特征提取

相比激光 SLAM, 虽然视觉 SLAM 方法具有成本低、获取信息量大、适用性广等天然优势, 但是仍存在不可忽略的问题。MonoSLAM 的提出首次解决了视觉 SLAM 方法的实时性问题, 使实际应用成为可能, 但是其使用的特征点法无法保证特征点跟踪的准确性和稳定性。因此, 对特征点的优化成为特征点法的关键问题。在视觉 SLAM 方法发展的早期, 许多特征点提取算法相继被提出, 通常提取图像中的角点作为特征点, 如 Harris 角点、FAST 角点、GFTT[18] 角点等。但是, 在实际应用中, 单纯的角点可能无法满足定位与建图的要求, 由于相机分辨率的限制, 当相机到角点的距离发生改变时, 远处的角点可能无法被提取到, 或者当相机视角旋转时, 同一个角点在不同视角下可能会发生很大的变化, 导致 2 个相同的角点无法匹配, 降低了 SLAM 的精度。随着计算机视觉领域技术的快速发展, 越来越多的特征点概念被相继提出, 如具有代表性的尺度不变特征[19](scale variant feature transform, SIFT)、加速稳健特征[20](speeded up robust features, SURF)、ORB 等, 都具有良好的性质, 如通过构建图像金字塔, 在每层不同尺度的金字塔图像中提取特征点来保证特征点具有尺度不变性, 使用灰度质心法, 连接图像块质心与图像块几何中心得到特征点的主方向, 保证特征点具有旋转不变性[21-22]。除此之外, 相比于普通的角点, 人工设计的特征点具有更好的可重复性、可区别性、本地性和高效性。其中, SIFT 充分考虑了相机在运动过程中图像出现的尺度、旋转、光照等变化, 提取出的特征点具有尺度不变性和旋转不变性, 对光照变化也有较好的鲁棒性。但是, SIFT 的描述子维度高达 128 维, 导致计算的时间复杂度高, 因此, 使用 SIFT 进行实时定位与建图工作很难保证系统的实时性。与 SIFT 相比, SURF 也具有尺度不变性和旋转不变性, 在计算时间复杂度方面也有所优化, 是与 SIFT 相类似的特征点。相比于前两者, ORB 特征在具有尺度不变性和旋转不变性的同时, 结合了 FAST 计算速度快的优势, 同时又改进了 FAST 不具有方向性的缺点, 并且使用计算速度极快的二进制描述子 BRIEF, 使得整个提取图像特征的环节速度大大提高, 有效降低了计算的时间复杂度。因此, ORB 特征能够在满足 SLAM 系统良好鲁棒性的前提下, 同时满足实时性要求, 是目前视觉 SLAM 方法中最常用的一种特征点。

虽然ORB可以满足许多视觉SLAM场景的使用,但要得到更好的表达效果,仍然需要进一步优化特征点。当前,处于研究热点的语义SLAM采用了深度学习的方法,但其本质仍然是优化提取图像特征。因此,图像特征的提取与表示仍然是视觉SLAM领域亟待解决的重要问题之一[23]。

3.2 位姿与地图的优化

SLAM的后端优化任务主要是解决前端视觉里程计位姿估计和建图带来的误差,防止形成较大的误差累积。前端视觉里程计中,采用计算相邻2帧之间关系的方法来估计相机的位姿和建图,这种链式计算方法随着时间的推移会产生较大的累积误差。因此,后端优化的思路是拓宽约束帧之间的时间和空间跨度,如在全局范围内选取具有代表性的关键帧,利用这些关键帧建立一个全局约束来优化之前通过相邻帧计算得到的不够准确的位姿和地图。

在相机运动过程中,有时会重复观测到之前观测过的场景,而由于前端链式计算带来的累积误差,此时的建图结果可能与之前的建图结果并不重合,要想建出较为精确的地图,需要添加闭环检测功能。实现闭环检测有多种方式,如对2帧之间的特征点进行匹配,但此类方法受光照等外界环境变化的影响较大。实现闭环检测最常用的算法是词袋模型,其将图像中包含的实体抽象出来,对应于词袋模型中的“单词”,由许多单词组合成“字典”,通过查询字典的方式,使用单词出现的情况来描述整幅图像,这样就把一幅图像转换成了一个向量来描述,通过向量的相似度即可估计图像的相似度。无监督机器学习中的聚类问题与字典的生成问题非常相似,通过K-means算法可以得到有效解决。当然,将深度学习与闭环检测联系起来也是当下的研究热点,在不同关键帧中进行语义分割,通过特定算法比较当前帧和关键帧的相似度来判断是否形成闭环也是比较常用的方法。

3.3 动态环境下的视觉SLAM方法

早期的视觉SLAM方法基本都基于静态环境,环境中动态的物体往往会对SLAM系统的定位与建图工作产生较大的影响,因此,要使视觉SLAM方法能够更稳定地应用于实际生活中,提高系统在动态环境下的鲁棒性是必须要解决的重大问题。当前,动态环境下的视觉SLAM问题的主流解决思路是将影响位姿估计和建图的动态部分剔除,仅使用环境中剩余的静态部分进行位姿估计和建图。使用运动向量来表示图像特征点的运动状态,当相机进行水平方向移动时,大多数特征点的运动向量是相同的,不同的运动向量表示的特征点很可能是正在移动的点,但是在相机视角旋转,尤其是相机视角绕光轴旋转时,环境中属于静态区域的特征点会产生许多方向不同的运动向量,这样判断动态区域的问题就会变得非常复杂。利用Delaunay三角剖分法,将图像序列中的特征点相连接形成多个三角形区域,比较相邻2帧中空间位置相同的2个特征点相连的几条边,如果边的长度发生较大变化,说明这2个特征点不属于同一个目标,将所有符合条件的边全部剔除后,计算由三角形组成的连通区域的面积,其中面积最大的部分即为静态区域。在特定的环境中,还可以使用深度学习,利用具有先验知识的语义信息来提取环境中主要的移动物体,如行人和车辆。当然,直接将动态区域剔除可能会导致图像中的特征点过少,不利于SLAM工作,而增加传感器能够解决这一问题,如IMU可以在短时间内较为准确地测量相机自身的角速度和加速度,在有动态物体干扰的情况下也可以比较准确地估计相机的位姿。

4 视觉SLAM方法的未来发展

相比激光SLAM,视觉SLAM方法除了具有成本低、获取信息量大、适用性广等特点外,还有一个关键性的优势,其工作方式更加接近人类认识世界的习惯,这也注定了其未来在智能机器人领域内将具有重要的地位。另外,虽然实际生产中使用更多的是激光SLAM,但是在未来多机器人合作领域,激光SLAM由于存在缺乏闭环检测手段、难以消除累积误差等关键问题,很难得到进一步的快速发展与应用,而视觉SLAM在消除累积误差方面的技术更加完善,且依靠深度学习和多传感器融合的帮助,越来越多视觉SLAM本身难以攻克的难题也得到了解决,未来有更大的发展空间。

4.1 与深度学习结合

近年来,计算机视觉领域技术的飞速发展很大程度上推动了视觉SLAM方法的发展,尤其是深度学习,在图像处理方面取得了优异成绩,在图像识别精度上超过了以往的其他方法。深度学习能够自动提取图像中难以观测到的有效特征信息,而视觉SLAM方法也需要提取图像中有效的特征点,因此,越来越多的学者使用深度学习来完成视觉SLAM的工作。2015年,Konda和Memisevic[24]将视觉里程计描述为分类问题,并使用卷积神经网络以图像序列作为输入,预测相机位姿和速度的离散变化,结果表明,就估计精度而言,该方法还无法与主流的视觉里程计方法相比,但证明了将深度学习与视

觉 SLAM 方法结合推向实际应用的可能性, 为以后视觉 SLAM 方法与深度学习的结合提供了宝贵的思路。经过众多学者的实践证明, 基于深度学习的 SLAM 比传统的视觉 SLAM 方法具有更好的鲁棒性。深度学习与闭环检测相结合可以获得很好的效果[25-26], 相关实验结果表明, 基于深度学习的闭环检测方法能够为系统带来更高的准确性和鲁棒性。此外, 深度学习还可以与建图环节相结合来创建语义地图。语义地图可以使机器人突破对环境地图单纯的几何理解, 有利于提高机器人的自主性和功能扩展性。例如, 在无人驾驶领域, 应用语义地图可以使汽车应对更加复杂的情况, 提高汽车行驶时的安全性和稳定性; 在家用机器人中, 使用语义地图可使机器人更加精确地分辨和寻找物体, 完成其他更加高级的工作。如果能够通过深度学习赋予环境更多的深层信息, 未来还可使机器人的工作方式更加接近人类, 使机器人更加智能化。

4.2　多传感器融合

相机能够捕捉环境中丰富的特征信息, 但是在剧烈运动、光照条件不佳等特殊场景下, 相机能捕捉到的特征信息变少, 导致 SLAM 难以正常工作。在目前计算机性能和算法上很难有重大突破的情况下, 可以添加传感器克服类似困难场景。2019 年, 香港科技大学开源了视觉-惯性 SLAM 系统 VINS-Fusion[27], 使用单目、双目、IMU、全球定位系统(global positioning system, GPS)的组合, IMU 能够在短时间内对相机的移动状态进行较为准确的估计, 在无法很好提取足够特征点的情况下发挥了关键作用。同时, 视觉 SLAM 方法中的闭环检测又能很好地解决 IMU 累积误差的问题, 起到了良好的补充作用, GPS 也大大增加了整个 SLAM 系统的鲁棒性。试验结果表明, 在相同条件下, 多传感器融合的 SLAM 系统无论精度还是鲁棒性都优于纯视觉 SLAM。虽然视觉 SLAM 方法向多传感器融合方向发展已经成为共识, 但是多传感器融合会导致移动机器人的体积增大, 需要根据具体应用情况人为寻找一个合适的度。此外, 多传感器融合还会造成信息冗余, 如何将短时间内接收到的大量信息集中在一起进行合理处理, 并做出正确决策也是一个重要的问题[28]。因此, 多传感器融合是 SLAM 领域未来的研究重点。

5　结 束 语

早期的视觉 SLAM 方法通过较为简单的特征点来实现定位与建图。为了提高 SLAM 系统的实时性, 提出了关键帧的处理方式。同时, 为了解决特征点难以匹配的困难, 设计出许多优秀的人造特征点, 如 SIFT、SURF、ORB 等。为了解决累积误差问题, 在后端优化的基础上添加了闭环检测模块来优化移动机器人的位姿及地图。但是, 目前仍存在若干难以彻底解决的问题, 如动态场景下的 SLAM 等。随着深度学习的飞速发展, 视觉 SLAM 方法所面对的种种困难出现了新的解决思路, 传统的功能模块中加入了深度学习, 无论是系统质量还是效率都得到了较大提升。然而, 视觉 SLAM 方法的研究还需继续提高系统的准确性和鲁棒性, 来适应更多、更复杂的应用场景。未来, 视觉 SLAM 方法将会在无人驾驶、智能机器人等领域得到快速发展与广泛应用。

参考文献（References）

[1] 权美香, 朴松昊, 李国. 视觉 SLAM 综述[J]. 智能系统学报, 2016, 11(6): 768-776.
QUAN M X, PIAO S H, LI G. An overview of visual SLAM[J]. CAAI Transactions on Intelligent Systems, 2016, 11(6): 768-776(in Chinese).

[2] 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.

[3] KLEIN G, MURRAY D. Parallel tracking and mapping for small AR workspaces[C]//Proceedings of the 6th IEEE and ACM International Symposium on Mixed and Augmented Reality. Piscataway: IEEE Press, 2007: 225-234.

[4] ROSTEN E. Machine learning for very high-speed corner detection[C]//Proceedings of the European Conference on Computer Vision. Berlin: Springer, 2006.

[5] NEWCOMBE R A, LOVEGROVE S J, DAVISON A J. DTAM: Dense tracking and mapping in real-time[C]//Proceedings of the International Conference on Computer Vision. Piscataway: IEEE Press, 2011: 2320-2327.

[6] ENGEL J, SCHÖPS T, CREMERS D. LSD-SLAM: Large-scale direct monocular SLAM[C]//Proceedings of the European Conference on Computer Vision. Berlin: Springer, 2014: 834-849.

[7] FORSTER C, PIZZOLI M, SCARAMUZZA D. SVO: Fast semi-direct monocular visual odometry[C]//Proceedings of the IEEE International Conference on Robotics and Automation. Piscataway: IEEE Press, 2014: 15-22.

[8] MUR-ARTAL R, MONTIEL J M M, TARDOS J D. ORB-SLAM: A versatile and accurate monocular SLAM system[J]. IEEE Transactions on Robotics, 2015, 31(5): 1147-1163.

[9] 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.

[10] 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.

[11] 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.

[12] ŞIMŞEK B, SATIR S, BILGE H Ş. Performance comparison of direct and feature based vSLAM algorithms[C]//Proceedings of the 29th Signal Processing and Communications Applications Conference. Piscataway: IEEE Press, 2021: 1-4.

[13] ENGEL J, KOLTUN V, CREMERS D. Direct sparse odometry[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018, 40(3): 611-625.

[14] LI R, WANG S, LONG Z, et al. UnDeepVO: Monocular visual odometry through unsupervised deep learning[C]//Proceedings of the IEEE International Conference on Robotics and Automation. Piscataway: IEEE Press, 2018: 7286-7291.

[15] 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.

[16] KANG S, GAO Y, LI K, et al. A visual SLAM algorithm based on dynamic feature point filtering[C]//Proceedings of the IEEE International Conference on Robotics and Biomimetics. Piscataway: IEEE Press, 2021: 1968-1973.

[17] WANG J, RÜNZ M, AGAPITO L. DSP-SLAM: Object oriented SLAM with deep shape priors[C]//Proceedings of the International Conference on 3D Vision. Piscataway: IEEE Presss, 2021: 1362-1371.

[18] SHI J. Good features to track[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 1994: 593-600.

[19] LOWE D G. Object recognition from local scale-invariant features[C]//Proceedings of the International Conference on Computer Vision. Piscataway: IEEE Press, 1999, 2: 1150-1157.

[20] BAY H, TUYTELAARS T, GOOL L V. SURF: Speeded up robust features[C]//Proceedings of the European Conference on Computer Vision. Berlin: Springer, 2006: 404-417.

[21] SUN C Z, ZHANG B, WANG J K, et al. A review of visual SLAM based on unmanned systems[C]//Proceedings of the 2nd International Conference on Artificial Intelligence and Education. Piscataway: IEEE Press, 2021: 226-234.

[22] JIAN M, WANG J, YU H, et al. Visual saliency detection by integrating spatial position prior of object with background cues[J]. Expert Systems with Applications, 2021, 168: 114219.

[23] 汤一明, 刘玉菲, 黄鸿. 视觉单目标跟踪算法综述[J]. 测控技术, 2020, 39(8): 21-34.
TANG Y M, LIU Y F, HUANG H. Overview of visual single target tracking algorithm[J]. Measurement and Control Technology, 2020, 39(8): 21-34(in Chinese).

[24] KONDA K R, MEMISEVIC R. Learning visual odometry with a convolutional network[C]//Proceedings of the International Conference on Computer Vision Theory and Application. Setubal: Science and Technology Publications, 2015: 486-490.

[25] HOU Y, ZHANG H, ZHOU S. Convolutional neural network-based image representation for visual loop closure detection[C]//Proceedings of the IEEE International Conference on Information and Automation. Piscataway: IEEE Press, 2015: 2238-2245.

[26] ZHANG X, SU Y, ZHU X. Loop closure detection for visual SLAM systems using convolutional neural network[C]//Proceedings of the 23rd International Conference on Automation and Computing. Piscataway: IEEE Press, 2017: 1-6.

[27] QIN T, CAO S, PAN J, et al. A general optimization-based framework for global pose estimation with multiple sensors[EB/OL]. (2019-01-11)[2022-05-01].https://arxiv.org/abs/1901.03642.

[28] GAUTAM A, MAHANGADE S, GUPTA V I, et al. An experimental comparison of visual SLAM systems[C]//Proceedings of the International Symposium of Asian Control Association on Intelligent Robotics and Industrial Automation. Piscataway: IEEE Press, 2021: 13-18.

An overview of visual SLAM methods

WANG Peng[1, 2]，HAO Weilong[1]，NI Cui[1, *]，ZHANG Guangyuan[1]，GONG Hui[1]

(1. School of Information Science and Electrical Engineering，Shandong Jiaotong University，Jinan 250357，China；
2. Institute of Automation，Shandong Academy of Sciences，Jinan 250013，China)

Abstract：Simultaneous localization and mapping (SLAM) enables mobile robots to calculate their position and pose by independently building an environment model during movement without any environmental prior conditions by carrying specific sensors. It can greatly improve the autonomous navigation ability of mobile robots and their adaptability to different application environments, and contribute to the subsequent implementation of dynamic path planning, real-time obstacle avoidance and multi-robot collaboration. Visual SLAM refers to using the camera as an external sensor to collect ambient information to create a map and estimate the robot's own position in real-time. The study describes and examines the various feature detection approaches, back-end optimization, loop closure detection, and the application of visual SLAM in a dynamic environment in addition to introducing the standard classical visual SLAM methods and the visual SLAM methods mixed with deep learning. This study addresses the current state-of-the-art in research and the potential growth of visual SLAM in the future before summarizing the issues with visual SLAM raised here.

Keywords：visual simultaneous localization and mapping; deep learning; feature detection; pose estimation; loop closure detection

Received：2022-05-18；**Accepted**：2022-06-23；**Published Online**：2022-10-10 10：03

URL：link.cnki.net/urlid/11.2625.V.20221009.1040.009

Foundation items：China Postdoctoral Science Foundation (2021M702030); Science and Technology Programe of Transportation Department of Shandong Province (2021B120)

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

http://bhxb.buaa.edu.cn jbuaa@buaa.edu.cn
DOI: 10.13700/j.bh.1001-5965.2022.0392

基于多模态联合语义感知的零样本目标检测

段立娟[1, 2, 3, *]，袁莹[1, 2, 3]，王文健[1, 2, 3]，梁芳芳[4, 5]

(1. 北京工业大学 信息学部，北京 100124；2. 可信计算北京市重点实验室，北京 100124；
3. 信息安全等级保护关键技术国家工程实验室，北京 100124；4. 河北农业大学 信息科学与技术学院，保定 071001；
5. 河北省农业大数据重点实验室，保定 071001)

摘 要：零样本目标检测借助语义嵌入作为引导信息，将未见过的物体的视觉特征与类别语义嵌入映射到同一空间，根据其在映射空间的距离进行分类，但由于语义信息获取的单一性，视觉信息缺乏可靠表示，易混淆背景信息和未见过对象信息，使得视觉和语义之间很难无差别对齐。基于此，借助视觉上下文模块捕捉视觉特征的上下文信息，并通过语义优化模块对文本上下文和视觉上下文信息进行交互融合，增加视觉表达的多样化，使模型感知到前景的辨别性语义，从而有效地实现零样本目标检测。在MS-COCO的2个划分数据集上进行实验，在零样本目标检测和广义零样本目标检测的准确率和召回率上取得了提升，结果证明了所提方法的有效性。

关 键 词：目标检测；零样本目标检测；多模态；上下文感知；语义优化
中图分类号：TP391
文献标志码：A **文章编号**：1001-5965(2024)02-0368-08

人类通常根据既往经验学习新的概念，能够归纳推断和演绎创新。如何使机器能够“类人概念学习”，不依赖大数据，不需要大规模长时间的训练也能够学会学习，即零样本学习，该问题受到了越来越多研究学者的重视，成为国内外研究的热点和难点。

零样本目标检测(zero-shot object detection, ZSD)的目的是使机器根据检测可见类别的经验并结合不可见类别的语义信息来识别出未见过的物体[1]。对于零样本目标检测中语义信息的获取，现有的研究思路是将词向量作为物体的语义信息，在训练时通过将分类器中可见类的视觉特征与词向量在同一嵌入空间对齐，学习视觉与语义的对应关系。因此，在测试不可见类别时，直接计算视觉特征与不可见类词向量在嵌入空间的相似度来得到类别标签[2-5]。但零样本目标检测涉及的语义信息仅来源于语义空间的标签词向量，往往忽视了视觉图像中存在的丰富的语义信息。此外，由于零样本独特的情景导致视觉信息缺乏可靠表示，很容易混淆背景信息和未见过对象信息，使得视觉和语义之间很难无差别对齐。

为解决上述问题，本文引入图像描述[6-7]提取文本上下文嵌入来弥补语义信息的单一性。针对零样本场景下传统目标检测方法提取到的视觉特征前背景区分能力不强问题，本文提出视觉上下文模块，旨在从图像中提取出视觉特征的上下文信息。为进一步增强文本和视觉信息的交互，本文还设计了语义优化模块，将文本上下文嵌入和视觉上下文信息进行深度融合。这种融合方式能够丰富视觉表达的多样性，并提高模型对前景的辨别性语义的感知能力，从而有效实现零样本目标检测。本文主要创新点为：①提出了一种新的多模态联合语义感知的零样本目标检测方法，突破了以往零样本目标

收稿日期：2022-05-19；**录用日期**：2022-06-23；**网络出版时间**：2022-10-18 18:26
网络出版地址：link.cnki.net/urlid/11.2625.V.20221017.1657.011
基金项目：国家自然科学基金 (62176009,62106065)；北京市教育委员会科学研究计划 (KZ201910005008)
* **通信作者**. E-mail：ljduan@bjut.edu.cn

引用格式：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).

检测语义的单一性, 引入图像描述弥补视觉上下文信息可能出现的误导性, 克服了传统零样本目标检测中的类别混淆问题。②提出语义优化模块, 通过视觉语义和文本语义交互融合来增加辨别性语义信息, 使得视觉与语义嵌入的对齐更精准。③在 MS-COCO[8] 数据集的 2 个划分数据集上进行实验, 证明了本文方法的有效性。

1　相关工作

自零样本学习[9] 概念被提出后, 人们发现借助语义信息能有效解决零样本分类问题, 进一步引发了对零样本目标检测[1] 的讨论。

现有的零样本目标检测主要分为基于生成器的方法和基于分类器的方法。基于生成器的方法希望仅用可见的类训练生成模型, 以此为不可见类生成视觉特征。Zhu 等[10] 基于构造生成器来实现零样本目标检测, 并通过视觉一致性检查模块确保生成的假视觉特征与原始的真实视觉特征一致。

基于分类器的方法是在目标检测的分类器中添加语义信息。早期, Bansal 等[2] 提出了零样本目标检测问题的设置, 并使用类别标签语义嵌入解决问题。随后, 文献 [3-5] 围绕如何实现视觉与语义的对齐来展开工作。针对零样本目标检测中背景类别与不可见类别的混淆问题, Rahman 等[11-12] 提出一种新的损失函数, 即极性损失来区分零样本目标检测中目标类别与背景类别, 以实现正负预测之间的最大分离, 并促进视觉和语义之间更好的对齐。在此基础上, Zheng 等[13] 提出级联的背景可学习 (background learnable cascade, BLC) 模型, 设计了背景可学习的区域提取网络 (background learnable region proposal network, BLRPN) , 学习更合适的背景类语义词向量, 减少了背景和不可见类之间的混淆, 并借助 Cascade R-CNN[14] 的级联结构设计了独特的分类器来提高检测准确率。Gu 等[15] 提出一种通过视觉和语言知识蒸馏的训练方法, 将目标检测模型提取的区域特征与基于图文对比学习的多模态预训练 (contrastive language-image pre-training, CLIP) 模型[16] 中提取的知识对齐。Li 等[17] 提出基于上下文引导 (context-guided, CG) 的 ZSD, 利用图像中对象周围的上下文信息进行零样本目标检测任务, 并在检测阶段使用这些信息。

早期的零样本目标检测工作[2-5] 围绕词向量与视觉特征的对齐来展开, 其语义信息只来源于标签词向量。Rahman[11-12] 和 Zheng[13] 等虽然进一步研究了区分背景与目标类别, 但仍然只依靠词向量来进行分类。Li 等[17] 提出的 CG 将上下文信息与零样本目标检测任务结合起来, 但只使用视觉的上下文信息, 在多数情况下不可靠, 有时会遇到与自然相反的上下文, 且同类物体有时看起来非常不同, 如果物体出现遮挡情况, 上下文可能会产生误导信息。本文使用多模态联合语义感知通过视觉语义与文本语义来增强特征的辨别性, 使得视觉与语义的对齐更精准, 有效提高了零样本目标检测的准确率。

2　零样本目标检测模型

为增强目标检测模型提取辨别性视觉特征的能力, 本文先通过视觉上下文模块获取视觉上下文特征, 再采用语义优化模块通过图像描述优化视觉上下文特征, 并将优化后的上下文特征融入视觉特征, 使视觉特征更具辨别性, 弥补零样本场景下视觉信息可靠性不高的问题。

2.1　整体结构

为使零样本任务下视觉信息与语义信息更好的对齐, 本文提出一种基于多模态联合语义感知的零样本目标检测方法, 整体结构如图 1 所示。

网络主要分为目标检测模块、视觉上下文模块和语义优化模块。其中, 训练过程包含目标检测模块、视觉上下文模块和语义优化模块, 测试过程包含目标检测模块和视觉上下文模块。输入图像经过目标检测网络的视觉特征提取器生成视觉特征向量 $\boldsymbol{P}\in\mathbf{R}^{H\times W\times d}$, 其中, H和W分别为不同视觉特征的高度和宽度, d为特征的通道维度。由于目标类别只出现在测试阶段, 强化模型对样本的特征提取能力尤为重要, 本文为更好地提取视觉信息, 将 $\boldsymbol{P}$ 输入到 2 个分支中。1 个分支是目标检测网络的区域特征提取器 (region proposal network, RPN), 为图像生成候选区域特征 $\boldsymbol{F}\in\mathbf{R}^{n\times d\times 7\times 7}$, 其中, n为候选区域个数; 另 1 个分支经过视觉上下文模块得到视觉上下文特征 $\boldsymbol{F}_{\mathrm{v}}\in\mathbf{R}^{H\times W\times d}$, 使模型提取的特征更具有辨别性。

由于视觉的上下文信息在多数情况下有误导性[18], 本文提出的语义优化模块引入了更直观的图像描述来优化 $\boldsymbol{F}_{\mathrm{v}}$。语义优化模块借鉴自然语言领域的图像描述模型, 将输入图像生成图像描述, 使用文本特征提取器, 通过图像描述生成文本上下文嵌入 $\boldsymbol{F}_{\mathrm{s}}\in\mathbf{R}^{1\times d_{\mathrm{s}}}$, 其中, d_{s}为文本上下文嵌入的维度。对 $\boldsymbol{F}_{\mathrm{v}}$和 $\boldsymbol{F}_{\mathrm{s}}$进行多模态上下文优化, 通过视觉与语义的交互作用得到经过语义优化的融合特征 $\boldsymbol{F}_{\mathrm{vs}}\in\mathbf{R}^{H\times W\times d}$。

为将经过语义优化的视觉上下文信息融合到区域特征中, 对 $\boldsymbol{F}_{\mathrm{vs}}$进行重塑操作, 并与 $\boldsymbol{F}$ 拼接得到

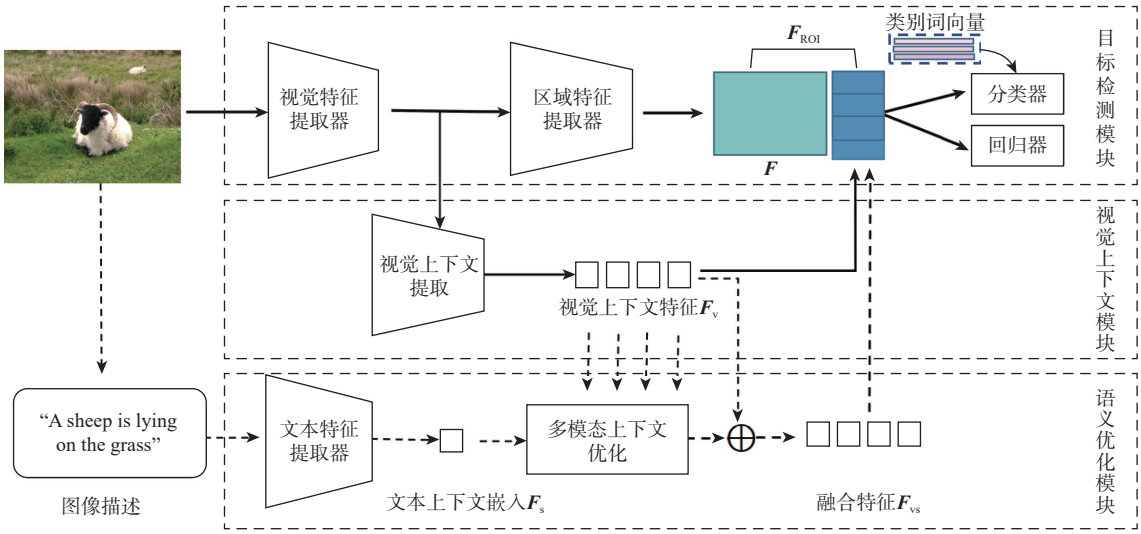


图1 基于多模态联合语义感知的零样本目标检测整体框架

Fig. 1 Overall framework of zero-shot object detection based on multi-modal joint semantic perception

增强后的区域特征$\boldsymbol{F}_{\mathrm{ROI}}\in\mathbf{R}^{n\times d\times 7\times 7}$。$\boldsymbol{F}_{\mathrm{ROI}}$与$\boldsymbol{F}$相比更具有辨别性，将$\boldsymbol{F}_{\mathrm{ROI}}$输入到目标检测的回归器与分类器中，并在分类器中按照BLC[13]的方法，将词向量与增强的区域特征计算相似性得到分类结果。

2.2 视觉上下文模块

视觉的上下文信息是由其周围像素（如背景信息、位置关系、不同类别对象的交互等）推断出的信息，已经在文献[17-19]中证明了上下文信息有助于提高各项任务的性能。因此，本文引入了视觉上下文模块提取视觉上下文特征，视觉上下文特征不仅能根据前背景存在的依赖关系丰富可见类上下文，还能将其迁移到不可见类的上下文中。

为收集视觉特征中有效的上下文信息，将$\boldsymbol{P}$作为输入，先通过3个空洞卷积[20]生成相同分辨率的上下文特征$\{\boldsymbol{h}_j\}_{j=1}^3\in\mathbf{R}^{H\times W\times d}$，$\{\boldsymbol{h}_j\}_{j=1}^3$包含更多感受野的上下文信息，使可以选择的有效上下文特征更多。为自适应地根据图像背景信息在$\{\boldsymbol{h}_j\}_{j=1}^3$中进行选择，使用一个卷积层将拼接的$\{\boldsymbol{h}_j\}_{j=1}^3$映射为权重矩阵$\boldsymbol{A}\in\mathbf{R}^{H\times W\times 3}$，其中，每个$\boldsymbol{A}_j\in\mathbf{R}^{H\times W}$对应$\boldsymbol{h}_j$的权重，再将$\boldsymbol{A}_j$与$\boldsymbol{h}_j$相乘，调整每个像素的上下文信息权重得到$\boldsymbol{M}\in\mathbf{R}^{H\times W\times d}$，通过这种注意力机制来选择多尺度的上下文特征。为增加特征多样性，提高模型的泛化性能，使用双线性插值对$\boldsymbol{M}$随机采样得到$\boldsymbol{\sigma}\in\mathbf{R}^{H\times W\times d}$和$\boldsymbol{\mu}\in\mathbf{R}^{H\times W\times d}$，通过式(1)得到采样后的上下文特征$\boldsymbol{N}\in\mathbf{R}^{H\times W\times d}$：

$$\boldsymbol{N}=\boldsymbol{\mu}+\varepsilon\boldsymbol{\sigma} \tag{1}$$

式中：ε为根据$\boldsymbol{\sigma}$生成的服从正态分布的随机标量。从$\boldsymbol{N}$中随机采样得到最终的上下文特征$\boldsymbol{Z}\in\mathbf{R}^{H\times W\times d}$，为对$\boldsymbol{Z}$进行后续优化，通过重塑操作将$\boldsymbol{Z}$重塑为视觉上下文特征$\boldsymbol{F}_{\mathrm{v}}\in\mathbf{R}^{H\times W\times d}$：

$$\boldsymbol{F}_{\mathrm{v}}=\mathrm{reshape}(\boldsymbol{Z}) \tag{2}$$

2.3 语义优化模块

语义优化模块通过图像描述来优化视觉的上下文特征，并将优化后的上下文特征融入视觉特征，弥补零样本场景下视觉信息不可靠的问题。

2.3.1 文本上下文提取

虽然视觉的上下文信息能够增强辨别性，但也可能存在误导性。一方面，图像中的上下文信息可能与直观感知相差甚远，如CG[17]中由于游戏手柄放在了一罐可乐前，机器视觉系统将其判定为叉子。另一方面，如果物体出现遮挡情况，视觉产生的上下文信息可能会产生不可靠引导。为优化视觉上下文特征，通过图像描述模型提取图像描述，并使用语言模型处理图像描述得到文本上下文嵌入。

具体流程如图2所示。图像描述模型由编码器和解码器构成[21]，对于输入图像$\boldsymbol{x}$，首先通过编码器为图像生成图像特征，然后经过解码器将图像特征转化为图像描述$\boldsymbol{C}$，其中，编码器使用VGG网络[22]，解码器由attention机制[23]和长短期记忆（long short-term memory, LSTM）[24]模型组成。为将图像描述转化为文本上下文嵌入，使用预训练的基于Transformer的双向编码表示（bidirectional encoder representation from Transformer, BERT）模型[25]作为文本特征提取器，$\boldsymbol{C}$经过BERT中的多头注意力模块得到文本上下文嵌入$\boldsymbol{F}_{\mathrm{s}}\in\mathbf{R}^{1\times d_{\mathrm{s}}}$。

2.3.2 多模态上下文优化

为借助文本上下文嵌入来优化视觉上下文特征，引入Transformer的Decoder模块[26]，其结构由多头注意力层和前馈神经网络层组成。文本上下文嵌入$\boldsymbol{F}_{\mathrm{s}}$通过映射层$\phi_{\mathrm{s}}$映射到视觉空间得到

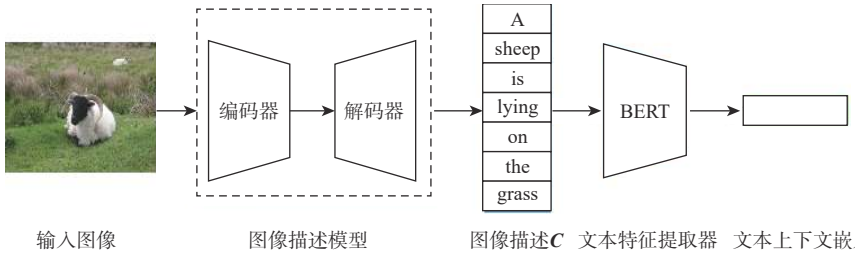


图 2　文本上下文提取流程

Fig. 2　Text context extraction process

$F_{se} \in \mathbf{R}^{1\times d}$, 视觉上下文特征$F_v$先通过自注意力层将输出结果作为查询向量$q$, F_{se}通过变换得到键向量k和值向量v, 再将其输入到多头注意力层和前馈神经网络中, 得到输出$F_{output} \in \mathbf{R}^{H\times W\times d}$。$F_{output}$和$F_v$根据式 (3) 得到最终文本优化的视觉上下文特征$F_{vs} \in \mathbf{R}^{H\times W\times d}$。

$$F_{vs} = F_v + \lambda F_{output} \tag{3}$$

式中: λ为可学习的参数, 用来控制残差连接[27]的范围。

为使用F_{vs}来补充区域特征的上下文信息, 通过全局平均池化改变其尺寸得到$F_{vsp} \in \mathbf{R}^{1\times d}$, 将$F_{vsp}$与区域特征$F$融合, 如下:

$$F_{ROI} = \mathrm{Conv}(\mathrm{Concat}(F, \mathrm{unsqueeze}(F_{vsp}))) \tag{4}$$

式中: unsqueeze(F_{vsp})表示增加F_{vsp}的维度并扩展其形状与F相同。将F_{vsp}和F拼接起来, 通过卷积得到$F_{ROI} \in \mathbf{R}^{n\times d\times 7\times 7}$。$F_{ROI}$代替$F$进行后续的分类和回归操作, 在测试阶段, 为避免语义优化过程中的图像描述模型见过不可见类而发生信息泄露, 实验时不使用语义优化模块, 而直接使用视觉上下文模块得到的视觉上下文特征F_v通过全局平均池化操作得到F_{vp}, 再与F融合, 操作过程如下:

$$F_{ROI} = \mathrm{Conv}(\mathrm{Concat}(F, \mathrm{unsqueeze}(F_{vp}))) \tag{5}$$

算法 1　语义优化算法。

输入: 图像x, 视觉上下文特征F_v。

输出: 增强的区域特征F_{ROI}。

步骤 1　x经过图像描述模型得到图像描述C。

$C = \mathrm{Image_caption}(x)$

步骤 2　C经过语言模型得到文本上下文嵌入F_s。

$F_s = \mathrm{BERT}(C)$

步骤 3　将F_s映射到视觉空间。

$F_{se} = \phi_s(F_s)$

步骤 4　进行多模态上下文优化。

$F_{output} = \mathrm{Transformer}(F_v, F_{se})$

步骤 5　对F_v和F_{output}做残差连接得到F_{vs}。

$F_{vs} = F_v + \lambda F_{output}$

步骤 6　将F_{vs}与F融合得到增强后的区域特征。

$F_{ROI} = \mathrm{Conv}(\mathrm{Concat}(F, \mathrm{unsqueeze}(\mathrm{GAP}(F_{vs}))))$

2.4　损失函数

目标检测模型的分类和回归共有t个阶段, 本文损失函数L_{total}可以表示为

$$L_{total} = \sum_{t=1}^{3} \alpha_t (L_{cls}^t + L_{reg}^t) \tag{6}$$

式中: L_{cls}^t为t阶段的分类损失, 使用多分类的交叉熵损失函数计算得到; L_{reg}^t为t阶段的回归损失, 使用 smothL1 损失函数计算得到; α_t为t阶段的权重, 参照 Cascade R-CNN[14]分别设置为 1,0.5,0.25。

3　实　验

3.1　实验设置

3.1.1　数据集设置

选择 MS-COCO 数据集[8]进行实验, COCO2014 数据集包含 80 个类别, 有 82 783 张训练图像和 40 504 张验证图像, 可以满足构建复杂的零样本情景。本文遵循文献 [2,12] 中数据集的设置, 将 MS-COCO 数据集划分为 2 种方式: 第 1 种由 48 个可见类别和 17 个不可见类别组成, 第 2 种由 65 个可见类别和 15 个不可见类别组成。其中, 训练集由可见类组成, 零样本目标检测的测试集由不可见类组成, 广义零样本目标检测 (generalized zero-shot object detection, GZSD) 的测试集由可见类和不可见类共同组成, 为防止训练时有不可见类别的物体出现, 将训练集中含有不可见类别的训练图像删除。

3.1.2　参数设置

目标检测模块基于 Cascade R-CNN[14]网络架构, 使用 ResNet-50[27]和特征金字塔网络 (feature pyramid networks, FPN)[28]作为特征提取网络。模型的分类器根据 BLC[13]方法对齐词向量与区域特征, 其中, 标签词向量使用 word2vec[29]得到, 而背景词向量使用 BLC[13]中的 BLRPN 方法得到。在语义优化模块中, 映射层ϕ_s通过一个全连接层实现, λ初始化为 0.000 1, 多模态上下文优化中, 多头注意力

的 head 设置为 4, layer 为 3。实验在 2 个 GPU 上进行, 每个 GPU 每次加载的图像数量为 1, 学习率为 0.002 5, 动量为 0.9, 重量衰减为 0.000 1。实验共进行 12 epoch, 并在第 8 个 epoch 后将学习率下降到初始的 0.1 倍。

3.2 实验结果

3.2.1 对比实验

分别在 48/17 和 65/15 的划分数据集上进行零样本目标检测实验, 零样本目标检测在测试时以不可见类的平均准确率作为评估标准, 并分别得出在阈值为 0.4、0.5、0.6 时的召回率, 取 5 次实验结果的平均值作为本文方法的实验结果。由表 1 可知, 本文方法相较于静态分配背景 (statically assigned background, SB)模型[2]、密集采样嵌入空间(densely sampled embedding space, DSES)[2]、CG[17]、带有文本描述的零样本目标检测(zero-shot object detection with textual descriptions, TD)[5] 和基于改进的视觉语义对齐的零样本目标检测(improved visual-semantic alignment for zero-shot object detection, PL)[12], 无论是准确率还是召回率都有一定的提升, 与基准模型 BLC[13] 相比, 平均准确率分别提高了 1.5% 和 0.2%。

表 1 48/17 和 65/15 数据集上零样本目标检测的实验结果

Table 1 Experimental results of ZSD on 48/17 and 65/16 datasets

方法	可见类/不可见类数量	召回率/%			平均准确率(阈值0.5)/%
		阈值0.4	阈值0.5	阈值0.6	
SB[2]	48/17	34.46	22.14	11.31	0.32
DSES[2]	48/17	40.23	27.19	13.63	0.54
TD[5]	48/17	45.50	34.30	18.10	
CG[17]	48/17				7.2
	65/15				10.9
PL[12]	48/17		43.59		10.10
	65/15		37.72		12.40
BLC[13]	48/17	49.63	46.39	41.86	9.90
	65/15	54.18	51.65	47.86	13.10
BLC(ms)[13]	48/17	51.33	48.87	45.03	10.60
	65/15	57.23	54.68	51.22	14.70
本文方法	48/17	51.40	48.69	45.29	12.10
	65/15	57.26	55.03	51.50	14.90

注: ms表示多尺度训练和测试。

分别在 48/17 和 65/15 的划分数据集上进行广义零样本目标检测实验, 广义零样本目标检测在测试时以可见类和不可见类的平均准确率和召回率作为评估标准。由表 2 可知, 在 48/17 的划分数据集上, 本文方法在可见类和不可见类的平均准确率与 BLC[13] 相比都有提升, 分别提高了 2% 和 0.6%, 在可见类和不可见类的召回率分别提高了 3.67% 和 2.79%, HM 中平均准确率和召回率分别提高了 0.9% 和 3.18%。在 65/15 的划分数据集上, 本文方法相比 BLC[13] 在可见类和不可见类的平均准确率分别提高了 0.3% 和 0.4%, 在可见类和不可见类的召回率分别提高了 1.14% 和 4.89%, HM 中平均准确率和召回率分别提高了 1.9% 和 3.11%。

表 3 和表 4 分别显示了本文方法与 BLC[13] 在 48/17 和 65/15 的划分数据集上阈值为 0.5 时的不可见类别的召回率。可以看到, 本文方法在大多数的不可见类上都有了显著的改进, 但是在“umbrella”这一类上的召回率很低, 在“tie”类上的召回率为 0。主要原因在于: 在分类器中, 计算可见类与不可见类的词向量的相似性时, 可见类中与这 2 个类相似的类较少, 使得检测存在难度。

表 2 48/17 和 65/15 数据集上广义零样本目标检测的实验结果

Table 2 Experimental results of GZSD on 48/17 and 65/15 datasets

方法	可见类/不可见类数量	平均准确率/%			召回率/%		
		可见类	不可见类	HM	可见类	不可见类	HM
DSES[2]	48/17				15.02	15.32	15.17
PL[12]	48/17	35.92	4.12	7.39	38.24	26.32	31.18
	65/15	34.07	12.40	18.18	36.38	37.16	36.76
BLC[13]	48/17	42.10	4.50	8.20	57.56	46.39	51.37
	65/15	36.00	13.10	19.20	56.39	51.65	53.92
本文方法	48/17	44.10	5.10	9.10	61.23	49.18	54.55
	65/15	36.30	13.50	21.10	57.53	56.54	57.03

注: HM表示可见类和不可见类的调和平均值。

表 3 48/17 数据集上不可见类别的召回率

Table 3 Recall of invisible categories on 48/17 dataset %

方法	总体召回率	召回率																
		bus	dog	cow	elephant	umbrella	tie	skateboard	cup	knife	cake	couch	keyboard	sink	scissors	airplane	cat	snowboard
基准(BLC[13])	46.4	77.4	88.4	71.9	77.2	0	0	41.7	38.0	45.6	34.3	65.2	23.8	14.1	20.8	48.3	79.9	61.8
本文方法	48.7	77.7	93.8	78.6	78.3	0.3	0	64.9	44.3	39.9	39.9	70.2	19.2	17.0	12.5	58.0	85.3	74.0

表 4　65/15 数据集上不可见类别的召回率

Table 4　Recall of invisible categories on 65/15 dataset　　%

方法	总体召回率	召回率														
		airplane	train	parking meter	cat	bear	suitcase	frisbee	snowboard	fork	sandwich	hot dog	toilet	mouse	toaster	hair drier
基准(BLC[13])	51.3	58.7	72.0	10.2	96.1	91.6	46.9	44.1	65.4	37.9	82.5	73.6	43.8	7.9	35.9	2.7
本文方法	55.0	67.0	77.4	9.6	96.9	91.8	53.7	49.2	69.0	41.1	86.1	77.6	50.1	10.9	42.3	2.7

3.2.2　消融实验

本节通过在 48/17 划分数据集上进行实验, 验证相同参数下不同模块对本文方法的贡献, 通过实验说明选择图像描述中的文本上下文嵌入来优化视觉上下文的优势。

1) 各模块性能。分别比较实验中各模块对实验结果的影响, 以 BLC[13] 为基准, 表 5 展示了在 48/17 划分数据集上添加不同模块时的平均准确率。可知, 当只添加语义优化模块时, 即将语义优化模块得到的文本上下文嵌入 $\boldsymbol{F}_s$ 与 $\boldsymbol{F}$ 进行融合, 实验结果降低。当只添加视觉上下文模块时, 即按照式(5)进行计算, 模型的平均准确率提高了 0.6%。当同时添加视觉上下文模块和语义优化模块后, 模型的平均准确率提升了 1.5%。

2) 语义信息的选取。在选择语义信息对视觉上下文特征进行优化时, 分别先后选择了词向量和通过图像描述得到的文本上下文嵌入作为语义指导, 根据表 6 中的实验结果可知, 使用词向量来优化视觉上下文特征对结果没有明显提升, 而使用图像描述得到的文本上下文嵌入优化视觉上下文特征时, 平均准确率得到明显提升。

表 5　各模块性能

Table 5　Performance of each module

基准(BLC[13])	视觉上下文模块	语义优化模块	平均准确率/%
√			10.6
√		√	10.0
√	√		11.2
√	√	√	12.1

表 6　语义信息的选取

Table 6　Selection of semantic information

语义信息	平均准确率/%
无	11.2
词向量	11.3
图像描述	12.1

3.3　结果可视化

对 65/15 的划分数据集进行可视化, 如图 3 所示。图中, 前 2 幅图是广义零样本目标检测的可视



(a) BLC[13]



(b) 本文方法



(c) 真实框

图 3　结果可视化

Fig. 3　Results visualisation

化结果，后2幅图是零样本目标检测的可视化结果。

可以看到，在广义零样本目标检测情境下，BLC[13]将不可见类的“飞机”识别为可见类的“风筝”，一方面是因为迁移策略很容易对可见类过拟合，另一方面，在图像视觉相似的情况下，缺少辨别性语义信息导致类别混淆问题，本文经过多模态联合语义感知有效增加了类间差异，正确识别出“飞机”。此外，BLC[13]在第2幅图中没有检测出不可见类“滑雪板”，本文方法正确检测出来。在零样本目标检测情境下，BLC[13]更容易出现类别混淆问题，特别是在外观相似的情况下，如将“熊”分类成“猫”，将“热狗”分类成“三明治”。本文方法引入图像描述优化视觉上下文信息，避免了类别混淆问题，将“熊”、“热狗”正确分类。

4 结 论

本文提出一种新的多模态联合语义感知的零样本目标检测方法。

1）通过零样本目标检测实验，验证了本文方法在平均准确率和召回率上都有一定的提升，在48/17划分数据集上平均准确率达到了12.10%，在65/15划分数据集上平均准确率达到了14.90%。

2）通过广义零样本目标检测实验，验证了本文方法在平均准确率和召回率上都有一定的提升，在48/17划分数据集上的调和平均值平均准确率达到了9.10%，在65/15划分数据集上的调和平均值平均准确率达到了21.10%。

参考文献（References）

[1] TAN C, XU X, SHEN F. A survey of zero shot detection: Methods and applications[J]. Cognitive Robotics, 2021, 1: 159-167.

[2] BANSAL A, SIKKA K, SHARMA G, et al. Zero-shot object detection[C]//Proceedings of the European Conference on Computer Vision. Berlin: Springer, 2018: 384-400.

[3] ZHANG L, WANG X, YAO L, et al. Zero-shot object detection via learning an embedding from semantic space to visual space[C]//Proceedings of the 29th International Joint Conference on Artificial Intelligence and 17th Pacific Rim International Conference on Artificial Intelligence. [S. l.]: IJCAI, 2020: 906-912.

[4] GUPTA D, ANANTHARAMAN A, MAMGAIN N, et al. A multi-space approach to zero-shot object detection[C]//Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. Piscataway: IEEE Press, 2020: 1209-1217.

[5] LI Z, YAO L, ZHANG X, et al. Zero-shot object detection with textual descriptions[C]//Proceedings of the AAAI Conference on Artificial Intelligence. Palo Alto : AAAI Press, 2019: 8690-8697.

[6] YANG X, TANG K, ZHANG H, et al. Auto-encoding scene graphs for image captioning[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2019: 10685-10694.

[7] FENG Y, MA L, LIU W, et al. Unsupervised image captioning [C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2019: 4125-4134.

[8] 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.

[9] WANG W, ZHENG V W, YU H, et al. A survey of zero-shot learning: Settings, methods, and applications[J]. ACM Transactions on Intelligent Systems and Technology, 2019, 10(2): 1-37.

[10] ZHU P, WANG H, SALIGRAMA V. Don’t even look once: Synthesizing features for zero-shot detection[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2020: 11693-11702.

[11] RAHMAN S, KHAN S, BARNES N. Polarity loss for zero-shot object detection[EB/OL]. (2020-04-02)[2022-05-01]. https://arxiv.org/abs/1811.08982v2.

[12] RAHMAN S, KHAN S, BARNES N. Improved visual-semantic alignment for zero-shot object detection[C]//Proceedings of the AAAI Conference on Artificial Intelligence. Palo Alto: AAAI Press, 2020: 11932-11939.

[13] ZHENG Y, HUANG R, HAN C, et al. Background learnable cascade for zero-shot object detection[C]//Proceedings of the Asian Conference on Computer Vision. Berlin: Springer, 2020: 107-123.

[14] CAI Z, VASCONCELOS N. Cascade R-CNN: Delving into high quality object detection[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2018: 6154-6162.

[15] GU X, LIN T Y, KUO W, et al. Zero-shot detection via vision and language knowledge distillation[EB/OL]. (2022-05-12)[2022-05-15]. https://arxiv.org/abs/2104.13921v1.

[16] RADFORD A, KIM J W, HALLACY C, et al. Learning transferable visual models from natural language supervision[C]//Proceedings of the International Conference on Machine Learning. [S. l.]: PMLR, 2021: 8748-8763.

[17] LI Y, SHAO Y, WANG D. Context-guided super-class inference for zero-shot detection[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2020: 944-945.

[18] GU Z, ZHOU S, NIU L, et al. Context-aware feature generation for zero-shot semantic segmentation[C]//Proceedings of the 28th ACM International Conference on Multimedia. New York: ACM, 2020: 1921-1929.

[19] YANG Z, WANG Y, CHEN X, et al. Context-Transformer: Tackling object confusion for few-shot detection[C]//Proceedings of the AAAI Conference on Artificial Intelligence. Palo Alto: AAAI Press, 2020: 12653-12660.

[20] YU F, KOLTUN V. Multi-scale context aggregation by dilated convolutions[EB/OL]. (2016-04-30)[2022-05-01]. https://arxiv.org/abs/1511.07122.

[21] XU K, BA J, KIROS R, et al. Show, attend and tell: Neural image caption generation with visual attention[C]//Proceedings of the International Conference on Machine Learning. [S. l.]: PMLR, 2015: 2048-2057.

[22] SIMONYAN K, ZISSERMAN A. Very deep convolutional net-

works for large-scale image recognition[EB/OL]. (2015-04-10)[2022-05-01]. https://arxiv.org/abs/1409.1556.

[23] BAHDANAU D, CHO K, BENGIO Y. Neural machine translation by jointly learning to align and translate[EB/OL]. (2016-05-19)[2022-05-01]. https://arxiv.org/abs/1409.0473v5.

[24] HOCHREITER S, SCHMIDHUBER J. Long short-term memory [J]. Neural Computation, 1997, 9(8): 1735-1780.

[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)[2022-05-01]. https://arxiv.org/abs/1810.04805v2.

[26] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]//Proceedings of the 31st International Conference on Neural Information Processing Systems. New York: ACM, 2017: 6000-6010.

[27] HE K, ZHANG X, REN S, 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.

[28] 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: 2117-2125.

[29] MIKOLOV T, SUTSKEVER I, CHEN K, et al. Distributed representations of words and phrases and their compositionality[C]//Proceedings of the 26th International Conference on Neural Information Processing Systems. New York: ACM, 2013: 3111-3119.

Zero-shot object detection based on multi-modal joint semantic perception

DUAN Lijuan[1, 2, 3, *], YUAN Ying[1, 2, 3], WANG Wenjian[1, 2, 3], LIANG Fangfang[4, 5]

(1. Faculty of Information Technology, Beijing University of Technology, Beijing 100124, China;
2. Beijing Key Laboratory of Trusted Computing, Beijing 100124, China;
3. National Engineering Laboratory for Critical Technologies of Information Security Classified Protection, Beijing 100124, China;
4. Faculty of Information Science and Technology, Hebei Agricultural University, Baoding 071001, China;
5. Hebei Key Laboratory of Agricultural Big Data, Baoding 071001, China)

Abstract: Existing zero-shot object detection maps visual features and category semantic embeddings of unseen items to the same space using semantic embeddings as guiding information, and then classifies the objects based on how close together the visual features and semantic embeddings are in the mapped space. However, due to the singleness of semantic information acquisition, the lack of reliable representation of visual information can easily confuse background information and unseen object information, making it difficult to indiscriminately align visual and semantic information. In order to effectively achieve zero-shot object detection, this paper uses the visual context module to capture the context information of visual features and the semantic optimization module to interactively fuse the text context and visual context information. By increasing the diversity of visual expressions, the model is able to perceive the discriminative semantics of the foreground. Experiments were conducted on two divided datasets of MS-COCO, and a certain improvement was achieved in the accuracy and recall rate of zero-shot target detection and generalized zero-shot target detection. The results proved the effectiveness of the proposed method.

Keywords: object-detection; zero-shot object detection; multi-modal; context perception; semantic optimization

Received: 2022-05-19; **Accepted**: 2022-06-23; **Published Online**: 2022-10-18 18:26

URL: link.cnki.net/urlid/11.2625.V.20221017.1657.011

Foundation items: National Natural Science Foundation of China (62176009,62106065); Program of Beijing Municipal Education Commission (KZ201910005008)

* **Corresponding author**. E-mail: ljduan@bjut.edu.cn

http://bhxb.buaa.edu.cn jbuaa@buaa.edu.cn
DOI: 10.13700/j.bh.1001-5965.2022.0396

基于顶点与主体区域同步检测的精准车牌定位

徐光柱[1,2]，刘高飞[2]，匡婉[2]，万秋波[2]，马国亮[3]，雷帮军[1,2,*]

(1. 湖北省水电工程智能视觉监测重点实验室（三峡大学），宜昌 443002；2. 三峡大学 计算机与信息学院，宜昌 443002；
3. 宜昌市公安局交通警察支队，宜昌 443002)

摘　　要：为应对非约束环境下的车牌精定位问题，提出一种基于顶点局部区域与主体区域同步检测策略的非约束性车牌定位算法。通过删减YOLOv5网络的输出结构，训练得到可同步检测车牌及顶点区域的车牌检测网络，在兼顾精度与计算速度的前提下，实现车牌顶点和主体区域的同步定位。针对一幅图中存在多个车牌区域及顶点区域存在少量漏检和误检的情况，分别设计了车牌顶点归类和单一缺失顶点预测后处理算法，借助顶点间的空间位置关系进行漏检目标预测和误检目标排查，有效改善了因场景复杂导致的个别顶点目标检测效果差的问题。所提算法在中国城市停车场数据集(CCPD)上的测试结果显示，平均精准率达99.25%，平均召回率达98.70%。所提算法不仅能够准确预测出车牌的4个顶点坐标，而且在中端GPU硬件平台上处理速度可达121帧/s，具有较好的应用价值。

关 键 词：深度学习；卷积网络；视觉目标检测；非约束车牌定位；车牌顶点检测
中图分类号：TP391
文献标志码：A　　**文章编号**：1001-5965(2024)02-0376-12

精准的车牌定位技术是车牌识别系统成功的关键，其精准程度直接影响后续车牌识别的准确性。因此，该问题受到了相关研究者的广泛关注[1-3]。

在停车场出入口、高速公路路口收费站等受约束性场景中，常设立道闸以使车辆停稳并面对摄像头，再采用带光源的摄像头拍照后对车牌进行定位和识别。因此，约束环境下得到的车牌图像清晰度高，车牌目标在图像中的特征明显。但在非约束场景中，由于没有良好的拍摄环境，受到恶劣天气、外物遮挡、随机拍摄角度和距离等不确定因素的影响，使得图像存在轮廓模糊、倾斜角度大、车牌部分受遮挡等现象。目前，约束环境下的车牌定位技术已趋于成熟。但在非约束环境下，由于环境光照、拍摄角度、遮挡等情况具有较大随机性，对车牌定位算法的稳定性造成较大干扰[4-6]，仍是充满挑战的开放性问题。

在非约束场景下车牌定位任务的研究中，深层卷积神经网络(deep convolution neural network, DCNN)被广泛应用。Xu等[5]提出RPNet，利用DCNN解决了端到端的车牌检测和识别问题，有效避免了传统方法采用多级处理产生的误差累加问题。RPNet的车牌检测原理是：在输出特征图上对车牌的中心点及宽高坐标进行回归，获得车牌外接矩形框。因此，RPNet只能解决一张图像只有一副车牌的车牌检测问题，应用场景受到限制。为应对图像中存在多副车牌和不存在车牌的情况，Selmi等[7]利用掩膜基于区域的卷积神经网络(mask region-based convolutional neural network, Mask R-CNN)[8]结构对车牌进行检测，借鉴GoogLeNet[9]构建了一个强大的特征提取模块，并采用该模块提取图像中的特征，将特征提取模块中输出的特征图输入到不进

收稿日期：2022-05-20；**录用日期**：2022-07-02；**网络出版时间**：2022-11-02 17:15
网络出版地址：link.cnki.net/urlid/11.2625.V.20221102.1538.003
基金项目：湖北省中央引导地方科技发展专项(2019ZYYD007)；湖北省水电工程智能视觉监测重点实验室开放基金(2022SDSJ03)
* **通信作者**. E-mail：Bangjun.Lei@ieee.org

引用格式：徐光柱，刘高飞，匡婉，等. 基于顶点与主体区域同步检测的精准车牌定位[J]. 北京航空航天大学学报，2024，50(2)：376-387.
XU G Z，LIU G F，KUANG W，et al. Accurate license plate location based on synchronous vertex and body region detection[J]. Journal of Beijing University of Aeronautics and Astronautics，2024，50(2)：376-387 (in Chinese).

行分割步骤的 Mask R-CNN 网络中进行车牌检测, 获取图像中的车牌候选区域, 利用全连接层输出置信度和车牌位置 2 个分支的信息, 该方法对图像中的非车牌背景区域和车牌区域进行了有效区分, 使车牌检测算法能应对图像中不存在车牌和存在多副车牌的情况, 但仍用矩形框对车牌进行粗定位, 未能解决车牌精定位的问题。

为得到车牌顶点坐标, Chen 等[10] 提出一种回归 4 个顶点坐标的方法, 考虑到图像中小车牌检测困难, 先检测车辆, 确定车辆的中心, 再根据车辆中心和车牌中心间的偏移量, 得到车牌在局部区域的矩形检测框, 将预选区域映射量化到预选区域的特征图上, 得到较准确的车牌锚框, 使用该锚框回归出车牌的 4 个顶点坐标, 得到精定位结果, 但采用先检测车辆再检测车牌的两阶段方法会导致车牌检测速度较慢, 无法满足实时性检测要求。为能直接检测车牌并获得车牌的顶点坐标, Wang 等[11] 受到人脸关键点检测任务的启发, 将多任务级联卷积神经网络 (multi-task cascaded convolutional neural network, MTCNN)[12] 的输出进行改造, 采用区域建议网络 (propose network, P-Net)、区域筛选网络 (region network, R-Net)、输出网络 (output network, O-Net) 3 种网络来处理, 利用 P-Net 将可能是车牌区域的目标过滤出来, 将过滤的目标框输入到 R-Net 中, R-Net 由卷积层和全连接层组成, 可对输入的感兴趣区域进行进一步调整和过滤, 将得到的车牌区域输入到 O-Net 中进行回归, 实现了网络输出车牌位置的同时, 也输出车牌内的 4 个顶点坐标, 由于该精定位算法通过直接检测车牌实现, 检测速度有所提升, 但由于边框筛选和顶点检测分开进行, 如果检测框偏离实际车牌区域, 顶点检测结果会受到较大影响。

为将边框结果和顶点坐标一起输出, Wang 等[13] 在车牌检测网络的检测头输出的特征通道中加入 4 个顶点坐标, 使网络可以同时输出车牌信息和顶点信息, 该方法采用特征金字塔[14](feature pyramid network, FPN) 的特征融合方式使车牌检测网络能够提取多种尺度的车牌信息, 但在计算精度时未计算车牌顶点的检测精度, 顶点分支仅辅助粗定位, 使得初定位结果更为准确, 并不能真正用于车牌区域的精确提取。

实际上, 由于车牌的 4 个顶点在图像中只占有很少的像素点, 网络输出端输出的特征图中, 车牌顶点这类像素级目标的特征很难得到有效保留, 通过坐标点回归方法很难获得高精准度的车牌顶点坐标。鉴于此, 许多研究者[10,13] 并未计算顶点检测的精度, 而是将顶点检测作为增加粗定位精度的一个辅助手段。

为实现非约束环境下车牌目标的快速且精准定位, 本文采用兼顾速度与精度的目标检测框架 YOLOv5[15], 并在其基础上删减不必要模块提升速度, 同时将顶点坐标的回归问题转化为包含顶点的局部车牌区域检测问题, 回避了直接进行点坐标回归方法无法保证精度的困难。通过充分利用车牌顶点附近的特征信息和上下文信息, 提升车牌顶点估计的鲁棒性。同时, 通过设计后处理算法, 结合检测到的顶点区域和车牌几何关系, 推理出由于特殊情况漏检的其他顶点区域, 克服了极端情况下车牌顶点区域漏检带来的定位困难问题, 进一步提升了复杂场景中的车牌检测精度和召回率。

1　相关网络模型

YOLOv5 是一种精确且快速的目标检测算法, 其在 YOLOv3[16] 和 YOLOv4[17] 的基础上集成了新的优化方法和模型策略, 实现了精度和速度上的双重提升。YOLOv5 的网络结构分为主干网络、颈部、头部 3 部分。主干网络由跨阶段局部网络 (cross stage partial network, CSPNet)[18] 模块、Focus 模块和空间金字塔池化 (spatial pyramid pooling, SPP)[19] 模块组成, 其结构如图 1 所示, 其中, 半透明区域为删减部分。CSPNet 能够消除网络反向优化时梯度信息冗余的现象, 增强卷积网络学习能力。Focus 模块对图像进行下采样, 将图像中空间信息转化到通道中, 降低了运算量。SPP 模块采用不同尺寸的池化核对特征图进行池化并拼接, 用于提升网络的感受野。颈部采用了路径聚合网络 (path aggregation network, PANet)[20] 结构, PANet 添加了一个自下而上的路径增强结构, 使得更多浅层特征得以保留, 改善了前向传播过程浅层特征丢失的问题。

YOLOv5 头部使用锚框及多尺度融合机制, 包括 C1、C2、C3 等 3 个检测头, 分别输出大、中、小尺寸特征图(见图 1)。特征图尺寸越大, 所关注的图像中的细节信息越丰富, 对尺寸较小的目标检测效果也越好, 但同时会造成大尺寸目标的漏检。因此, YOLOv5 的输出结果中综合了 3 个尺度的特征图上预测得到的结果, 使得 YOLOv5 对各类场景中不同角度、大小、光照条件下的车牌都能达到较好的检测效果, 保证了检测算法能准确且快速地检测车牌及其局部区域。经计算, C1 检测头的感受野约为 8 像素×8 像素大小的像素区域, 而中国城市停车场数据集 (Chinese city parking dataset, CCPD) 的训练集中车牌最小宽高为 70 像素×23 像素。在 YOLOv5 网络中, 输入尺寸为 640 像素×640 像素时, 最小的车牌尺寸宽高约为 62 像素×13 像素, 超

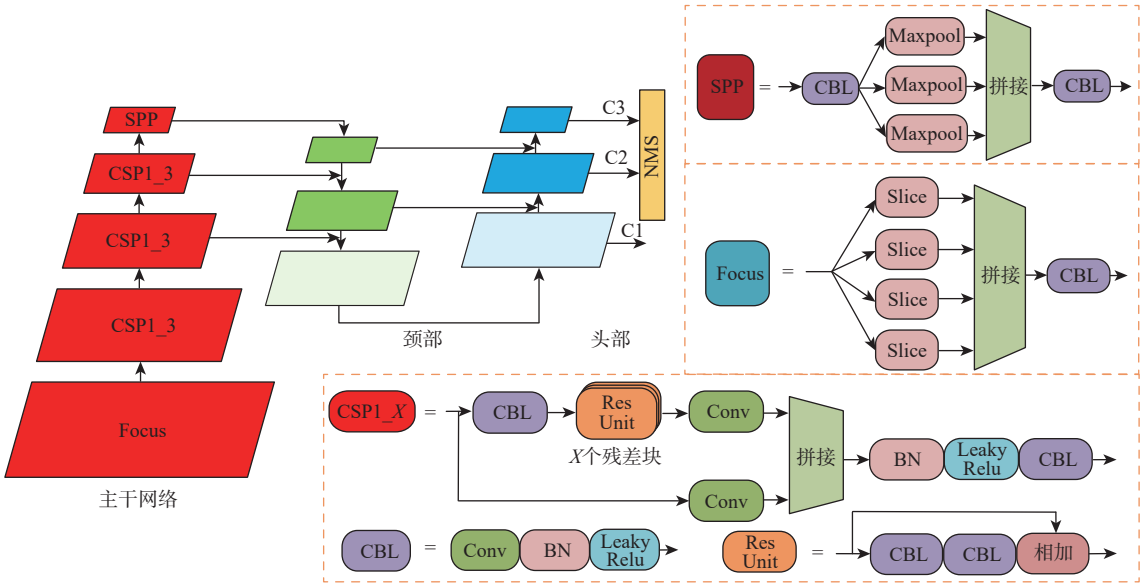


图 1 本文采用的双检测头 YOLOv5 网络结构

Fig. 1 Structure of dual-head YOLOv5 network used in this paper

出了 C1 检测头的感受范围。因此，C1 检测头在 CCPD 数据集上只发挥了少量的作用。此外，由于过小的车牌信息损失严重难以辨识，实际应用中不会要求检测和识别过小的车牌。为此，删除用于小目标检测的 C1 检测头，进一步减少网络参数，使得单张图像处理时间由 10.30 ms 减少到 8.25 ms。

2 顶点与主体目标检测相结合的车牌定位

与其他主流目标检测网络类似，YOLOv5 使用矩形检测框定位目标；用于车牌定位时，所获取的车牌区域中常包含部分背景区域，特别是在倾斜角度较大的情况下，无法准确有效地提取车牌区域，增加了后续车牌字符识别的难度。为解决该问题，本文通过将顶点坐标估计问题转换为以顶点为中心的局部区域检测问题，设计了一种主体与顶点目标检测相结合的车牌定位算法，其主体系统示意图如图 2 所示。首先，利用数据集中的车牌和顶点位置构建车牌和顶点区域检测数据集；然后，设置网络输入尺寸为 640 像素×640 像素，训练顶点与车牌区域同步检测的 YOLOv5 网络，将检测到的所有区域视作同一类别，进行无类别的非极大抑制 (non maximum suppression, NMS) 操作；最后，结合车牌顶点归类、单一缺失顶点预测的后处理过程获取精准的车牌区域。

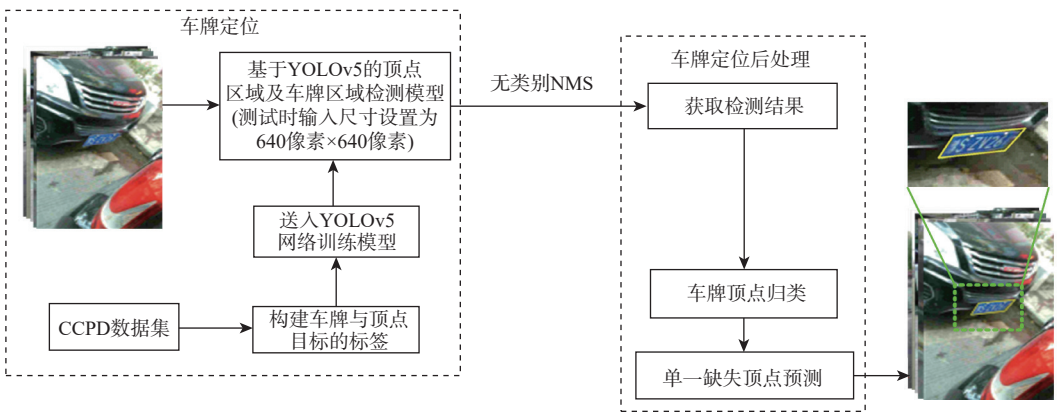


图 2 系统示意图

Fig. 2 System schematic diagram

2.1 基于 YOLOv5 的车牌顶点及主体区域同步检测

由于图像中的车牌大小不一，为统一顶点区域的特征，本文使用一种与车牌尺寸正相关的方式设定车牌 4 个顶点目标区域。图 3 为车牌顶点区域与实际车牌区域的位置关系。图中：A、B、C、D 为车牌 4 个顶点，以车牌顶点为中心的 4 个正方形框为模型需检测的顶点区域。顶点区域的边长与车牌尺寸正相关，顶点 C 与顶点 B 对应正方形区域边

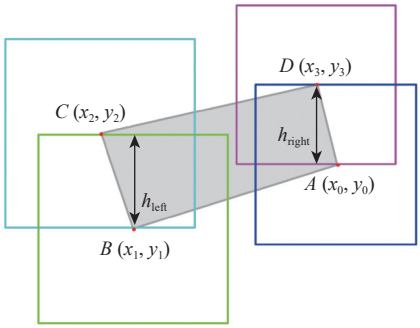


图 3　车牌顶点区域与车牌区域的空间位置关系

Fig. 3　Spatial position relationship between vertex regions and license plate

为 $2h_{\text{left}}$, 顶点 D 与顶点 A 对应正方形区域边长为 $2h_{\text{right}}$, h_{left}、h_{right} 分别为车牌左、右两侧顶点的高度差。使用上述方式将车牌的 4 个顶点区域截取出来, 如图 4 所示。考虑到顶点区域和车牌区域的样本平衡问题, 将顶点区域分为 4 类, 经实验证明, 将每个顶点区域分为单独的一类, 模型能更好地收敛, 有效避免了因训练集样本不平衡导致的性能损失, 提高了车牌定位的精确率。

$$h_{\text{left}} = y_1 - y_2 \tag{1}$$

$$h_{\text{right}} = y_0 - y_3 \tag{2}$$

2.2　后处理算法

本文设计的基于顶点与主体区域同步检测算法充分利用了神经网络输出层的特征信息, 将车牌顶点和顶点附近的像素视作一体, 一同作为待检测目标, 通过检测到的顶点区域间接计算顶点坐标。检测顶点区域的思想使得神经网络更易提取到鲁棒的特征信息, 进而更准确地得到顶点的坐标值。由实验结果可知, 根据顶点区域的中心点间接计算出顶点坐标比直接回归顶点坐标准确率更高且速度更快。由于车牌图像经过顶点与车牌区域同步检测后, 只能获得车牌顶点区域, 无法获得车牌 4 个具体顶点的坐标, 需对获得的车牌及顶点区域进行后处理才能得到具体的顶点坐标值, 图 5 为车牌及顶点区域进行处理后的流程。

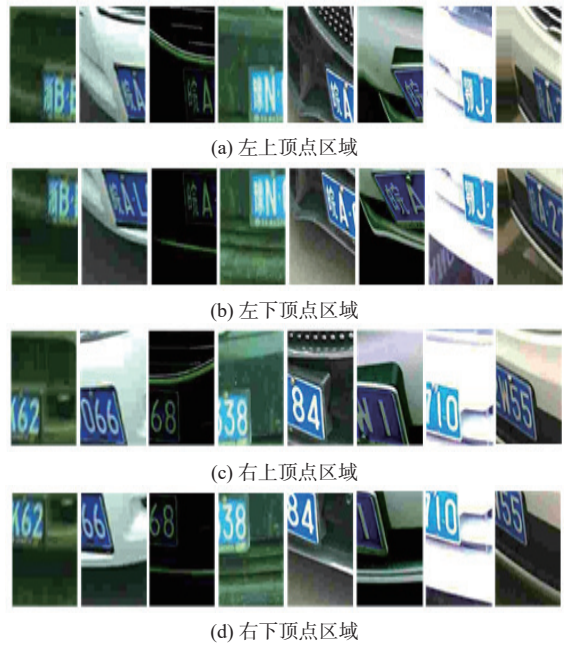


图 4　车牌顶点区域图例

Fig. 4　Instances of license plate vertex region

首先, 将图像送入训练好的 YOLOv5 网络中, 进行顶点与车牌区域同步检测; 其次, 通过顶点区

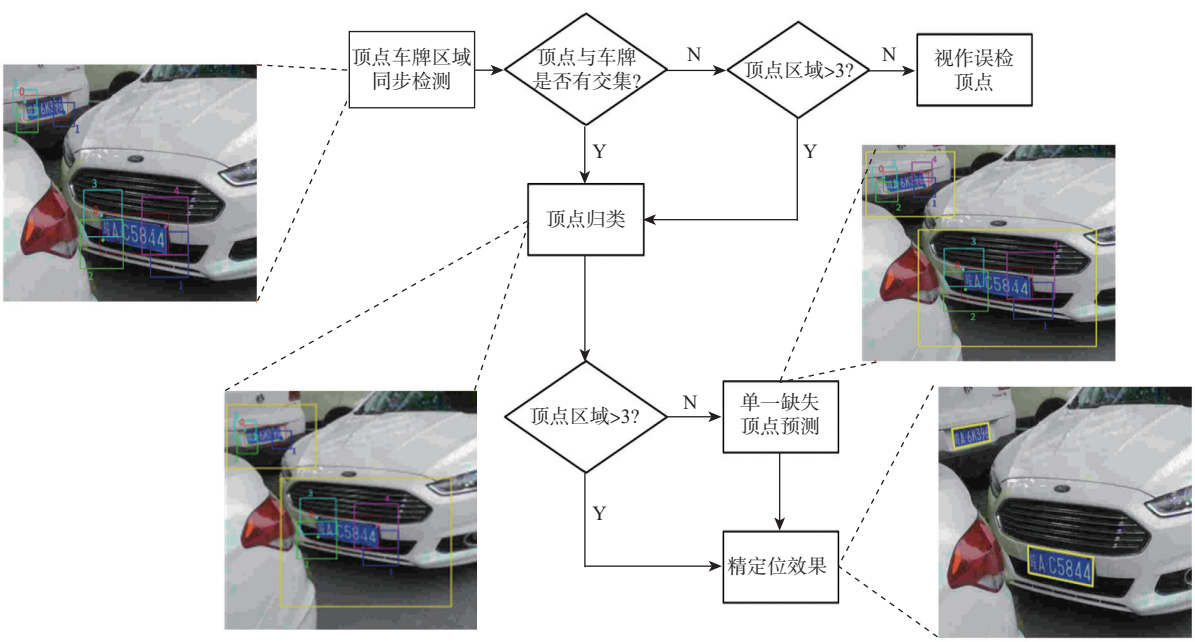


图 5　后处理流程

Fig. 5　Post-processing flow

域与车牌区域是否有交集对车牌顶点进行归类；然后，对单一缺少顶点区域进行补充，对少量误检到的顶点区域进行删除；最后，取所有顶点区域的中心点作为顶点坐标，按照空间关系顺序相连，即可获得车牌的精定位结果。本文提出的后处理步骤效率较高，在硬件设备下，同步检测顶点区域和车牌区域平均消耗 7.94 ms，后处理步骤平均消耗 0.30 ms，消耗的时间很少，保证了车牌定位的实时性。

2.2.1 车牌顶点归类算法

在实际应用场景中，无法确保相机拍摄的图像中只含有一副车牌，而且目标检测算法可能对一张图像中类似顶点的区域产生误检。因此，使用一种基于车牌区域的顶点归类算法处理出现多个顶点目标后的车牌顶点归类问题。

当某个车牌顶点区域与某个车牌区域存在交集时，该顶点区域中心点为其对应车牌区域中的车牌顶点。算法实现时，先遍历图中的车牌区域目标，每次获取车牌区域后，检测所有与之存在交集的检测框，将二者对应保存到集合 R 中；再在 R 中只保留车牌区域内顶点检测框大于 3 个的情况；最终针对极少量车牌区域发生漏检但顶点被成功检测的情况，判断剩余顶点数目是否大于 3，若是则保存到 R 中。车牌顶点归类处理效果如图 6 所示，经过顶点归类可有效解决图中有多个车牌和顶点误检的情况，图中黄色矩形框标示了归类结果。根据归类后的信息，可实现一张图像中多副车牌的同步检测，同时也可以实现误检顶点区域的排除。算法流程如图 7 所示。

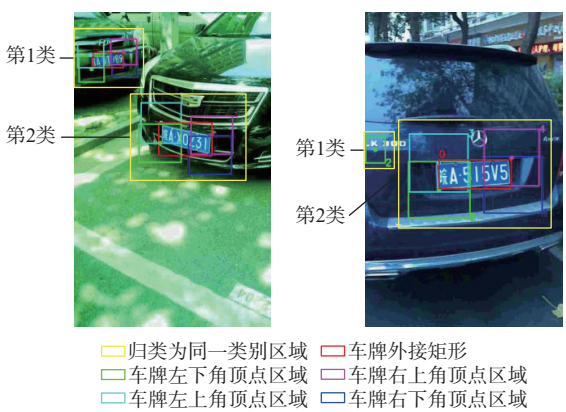


图 6 车牌顶点归类图例

Fig. 6 License plate vertexes grouping instances

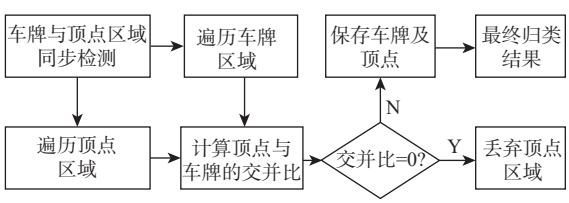


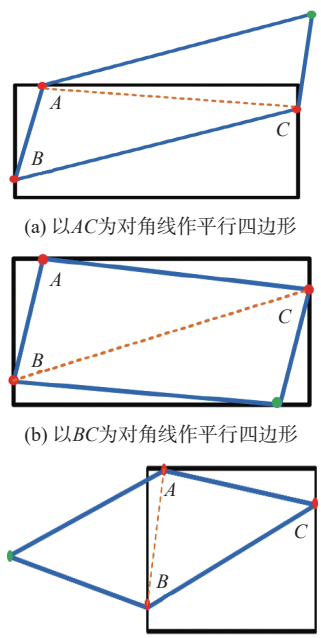
图 7 车牌顶点归类流程

Fig. 7 License plate vertexes grouping flow

2.2.2 单一缺失顶点预测

在极少数极端场景下，即使采用性能强大的目标网络也会出现个别顶点漏检的现象，主要有 2 种情况：①车牌图像在摄像头与车牌夹角较大的情况下拍摄，产生了较强的透视畸变，导致单一顶点区域在图像中呈现向内偏折的视觉效果，造成顶点区域特征与正面视图差异较大，使得网络对内折顶点漏检；②单个顶点受到其他异物遮挡，造成该顶点区域特征丢失，使得网络对受遮挡的顶点区域漏检。一般情况下，缺失顶点只会出现一个，对多个顶点都缺失的情况，很难对其进行精定位。因此，本文在只考虑单一顶点漏检的情况下，提出了一种根据车牌区域与余下 3 个顶点区域结合的单一缺少顶点预测算法，有效解决了单一顶点缺失的问题。

本文采用的单一缺失顶点预测算法如图 8 所示。假设获得的 3 个顶点分别记为 A、B、C，并得到 3 个顶点的外接矩形框，如图 8 中的黑色外框所示。首先，分别以 AC、BC、AB 为对角线作平行四边形；然后，分别计算获得的平行四边形的外接矩形框与 3 个顶点外接矩形框之间的交并比 (intersection over union, IoU)；最后，可知以 BC 为对角线预测得到的平行四边形与黑色外框的交并比最大，将该平行四边形的第 4 个顶点记为预测的缺失点。



(a) 以AC为对角线作平行四边形

(b) 以BC为对角线作平行四边形

(c) 以AB为对角线作平行四边形

图 8 单一缺失顶点预测图示

Fig. 8 Prediction illustration for single vertex missing cases

3 实 验

3.1 数据处理

为验证本文算法性能，使用 CCPD[5] 作为实验

数据集。CCPD 是目前规模较大的非约束环境下拍摄的国内车牌公开数据集, 共约 35 万张图像, 包含不同位置和拍摄角度、不同光照、不同天气环境下的车牌图像, 内含 9 个子集: 综合 (Base) 、过亮或过暗 (DB)、距离远 (FN)、倾斜 (Rotate)、较大倾斜 (Tilt)、不同天气 (Weather)、困难 (Challenge)、模糊 (Blur)、无牌 (NP)。CCPD 数据集中的标签数据包含车牌的外接矩形框、4 个顶点位置、车牌号码等多个标记信息。整体而言, CCPD 是目前仍有挑战性的非约束环境下的车牌定位数据集, 因此, 本文主要采用该数据集进行算法实验。

采用 CCPD 数据集中推荐的数据划分方式, 保证对比实验的公平性。划分时, 随机将 Base 子集划分为数量相等的 2 部分, 一部分作为训练集和验证集, 剩余部分与除 NP 子集之外的其他所有子集一起作为测试集。训练网络前, 先利用 CCPD 标签中的顶点标注信息和车牌标注信息结合 2.1 节的生成方式, 构建可供 YOLOv5 训练的目标检测数据集, 再随机抽取训练集中 20% 的图像作为验证集。

数据集构建过程中, 对顶点目标检测框超出边缘的情况, 会出现边缘信息丢失导致顶点区域检测不准的问题, 需要对图像进行填充。通过图像边界的背景像素对扩展部分进行填充。计算顶点区域的 4 个极限边界位置, 如下:

$$\begin{cases} x_{\min} = \min(x_{\mathrm{lu}1}, x_{\mathrm{lu}2}, \cdots, x_{\mathrm{lu}n}) \\ y_{\min} = \min(y_{\mathrm{lu}1}, y_{\mathrm{lu}2}, \cdots, y_{\mathrm{lu}n}) \\ x_{\max} = \max(x_{\mathrm{rd}1}, x_{\mathrm{rd}2}, \cdots, x_{\mathrm{rd}n}) \\ y_{\max} = \max(y_{\mathrm{rd}1}, y_{\mathrm{rd}2}, \cdots, y_{\mathrm{rd}n}) \end{cases} \tag{3}$$

假设图像中共有 n 个顶点区域, 每个车牌区域的左上角坐标和右下角坐标分别记为: $(x_{\mathrm{lu}i}, y_{\mathrm{lu}i})$、$(x_{\mathrm{rd}i}, y_{\mathrm{rd}i})$, 经过式 (3) 的运算, 遍历找到了顶点区域中最容易超过图像边界的坐标值。如图 9 所示, 假设图中有 4 个顶点区域, 且检测框的边界均超出图像边缘, 需按图 9 所示的方式对图像边缘进行扩展。图中, 4 个正方形框分别为 4 个极限边界框的位置, 图像外的蓝色条状矩形框和四角上的橙色矩形框为需扩展的区域。经过实验证明, 对检测框超出的边界处填充与图像边界相同的像素值能获得较好的顶点检测效果, 因此, 在超出部分的基础上额外扩大了与背景像素相同的 20 个像素。图 9 中扩展的蓝色区域主要利用原始图像边缘的像素点进行填充, 上方和下方按行填充, 左侧和右侧按列填充, 对于橙色矩形区域, 则随机使用相邻蓝色边界按行或按列进行像素填充。图像扩展后得到的新图像尺寸计算方式如下:

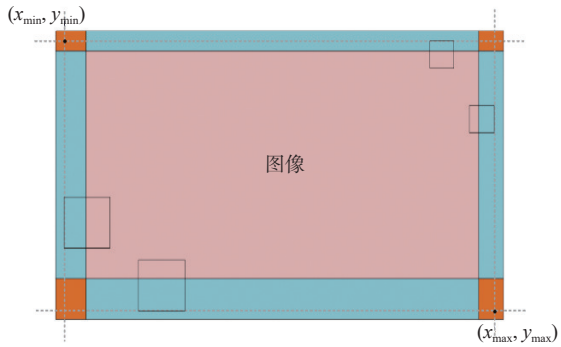


图 9　图像扩展方式

Fig. 9　Image expansion method

$$\begin{cases} W_{\mathrm{new}} = W + a_{\mathrm{left}} + a_{\mathrm{right}} \\ H_{\mathrm{new}} = H + a_{\mathrm{up}} + a_{\mathrm{down}} \end{cases} \tag{4}$$

$$a_{\mathrm{left}} = \begin{cases} -x_{\min} + 20 & x_{\min} \leqslant 0 \\ 0 & x_{\min} > 0 \end{cases} \tag{5}$$

$$a_{\mathrm{right}} = \begin{cases} x_{\max} - W + 20 & x_{\max} \geqslant W \\ 0 & x_{\min} < W \end{cases} \tag{6}$$

$$a_{\mathrm{up}} = \begin{cases} -y_{\min} + 20 & y_{\min} \leqslant 0 \\ 0 & y_{\min} > 0 \end{cases} \tag{7}$$

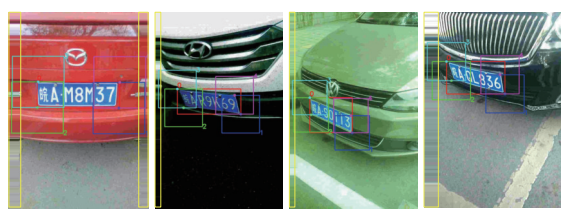
$$a_{\mathrm{down}} = \begin{cases} y_{\max} - H + 20 & y_{\max} \geqslant H \\ 0 & y_{\max} < H \end{cases} \tag{8}$$

式中: $(W_{\mathrm{new}}, H_{\mathrm{new}})$、$(W, H)$ 分别为扩展后、扩展前的图像宽高; a_{left}、a_{right}、a_{up}、a_{down} 分别为左侧、右侧、上方、下方需扩展的尺寸大小。

图 10 为车牌区域及顶点区域的标签信息可视化图像, 黄色矩形框为扩展区域。



(a) 不需要扩展的图像



(b) 需要扩展的图像

车牌外接矩形　车牌左下角顶点区域
车牌右上角顶点区域　车牌左上角顶点区域
车牌右下角顶点区域

图 10　车牌标签可视化示例

Fig. 10　Visualization instances of license plate labels

3.2　模型训练所用软硬件平台及参数设定

本文实验所用的软件平台为 Ubuntu16.04,

Pytorch 1.6, Cuda10.2, Cudnn10.2; 硬件环境为 GTX 1080Ti GPU, Intel Xeno ES-2460 CPU, 16 GB 内存。选择表 1 中的超参数进行网络训练, 采用随机梯度下降 (stochastic gradient descent, SGD) 作为优化器。在图像增强方面, 超参数的设置与 YOLOv5 在 COCO 数据集上的设置保持一致。为克服 SGD 优化器在训练过程中存在鞍点振荡的问题, 采用余弦退火算法[21] 动态调整学习率。余弦退火算法的优势是: 学习率在训练过程中存在瞬时跃变点, 有助于网络在优化过程中跳出局部最优点, 找到模型参数的全局最优解。在锚框的选择上, 由于本文删减了 YOLOv5 的一个检测头, 只需得到 6 类锚框的宽高即可。本文使用 K-means 算法对标注边界框的尺寸进行重新聚类分析, 获取表 1 中的 6 组锚框值。训练结束后, 从每个迭代次数表现最好的模型中选择最优模型作为最终训练结果。

表 1 模型训练参数

Table 1 Model training parameters

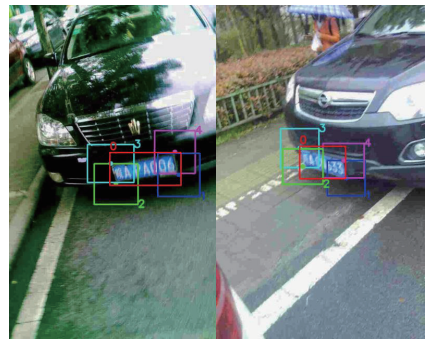
参数	数值
批次	16
类别数	5
学习率	0.01
动量	0.937
迭代次数	150
输入宽高	(640,640)
锚框宽高(6个)	(66,64), (123,49), (84,83), (100,99), (176,68), (124,123)

3.3 实验结果及分析

经过车牌顶点归类处理和单一缺失顶点预测处理后, 3 种情况下可以精确定位车牌, 如图 11 所示, 分别为检测到车牌区域及其对应的 4 个顶点区域、检测到车牌 4 个顶点区域、检测到车牌区域及其对应的 3 个顶点区域。如图 11(a)、(b) 所示, 可以直接将 4 个车牌顶点区域的中心按一定顺序连接得到精准的车牌区域, 而对于图 11(c), 则可以利用 2.2.2 节所述的单一缺失顶点预测算法预测缺失顶点的位置。

图 12 为单一缺失顶点预测前后效果。图 12(a) 为经过车牌与顶点同步检测后的效果, 因图中车牌过小且具有较大的倾斜角度, 所以图像在顶点定位时出现了单一顶点检测缺失。经过后处理步骤后, 其检测效果如图 12(b) 所示。可以看出, 利用平行四边形的性质结合检出的车牌区域预测出缺失顶点位置比较准确, 本文算法可有效提升单一顶点缺失情况下的检测准确率。

鉴于目前主流车牌定位算法采用矩形框间的



(a) 车牌区域和4顶点区域同时被检测到的情况



(b) 仅检测出4个顶点区域的情况



(c) 车牌区域和3顶点区域同时被检测到的情况

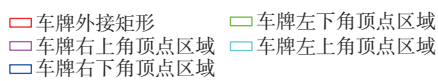
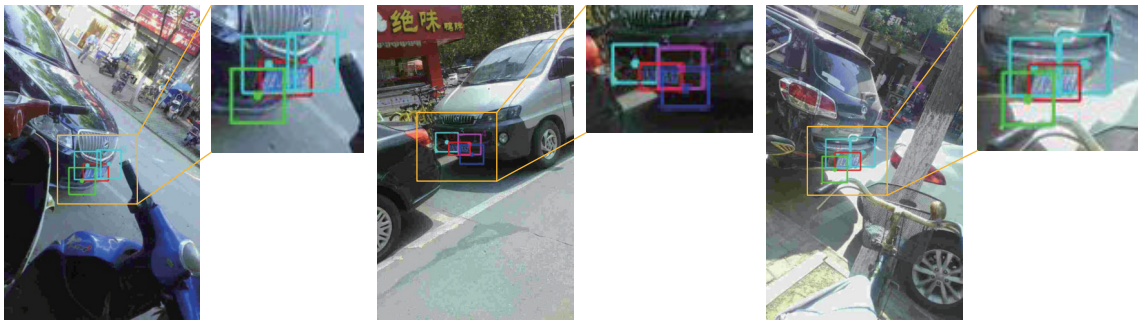


图 11 精准定位的车牌示例

Fig. 11 Accurate license plate location instances

交并比作为评估标准, 为与其他文献评估标准保持一致, 本文采用通用目标检测评判标准。由于车牌精定位获取的不是矩形框, 而是一个四边形框, 本文以检测车牌区域的外接矩形框与实际车牌区域外接矩形框的交并比作为评判指标[22]。在 CCPD 数据集中, 当检测框与标记框之间交并比>0.7 时, 认为检测正确, 否则认为检测错误。本文计算交并比的方式如图 13 所示。采用该方法计算交并比在判别车牌检测是否准确时更加具备合理性, 同时也更为严苛。鉴于 CCPD 数据集的标签中每张图像只标注了一张车牌, 众多使用 CCPD 数据集的文献[5,23~25] 均只采用精准度作为评估指标。为便于算法定位结果对比, 先统计了本文算法在各个子集中所检测到的车牌交并比>0.7 的比例作为准确



(a) 预测前



(b) 预测后

图 12　单一缺失顶点时的预测示例

Fig. 12　Prediction instances in cases of single vertex missing

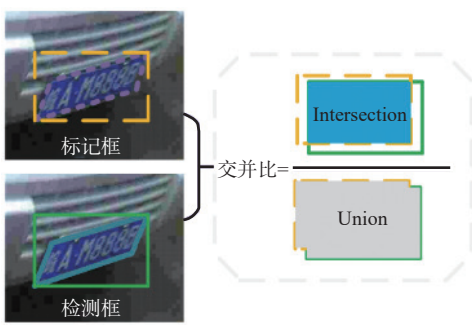


图 13　交并比计算方式

Fig. 13　IOU calculation method

率[5], 如表 2 所示。虽然其他文献因每张图像只有一副车牌的原因而忽略了召回率, 但本文研究发现存在少许漏检的情况, 因此同时统计了本文算法的召回率, 发现漏检率仅为 1.3%, 即平均召回率达到了 98.70%, 这也验证了 CCPD 数据集及其他相关文献仅采用精准率的合理性。

为评估本文算法的性能, 对除 NP 外的场景中的车牌定位准确率进行了统计, 并计算模型在整个测试集上的平均精准率。本文在实验过中分成了 E、F、G、H 等 4 种情况, E 表示使用 3 个检测头无

表 2　不同算法在各子集下的定位准确率

Table 2　Location accuracy of different algorithms under various subset

算法	平均精准率/%	检测速度/(帧·s^{-1})	准确率/%							
			Base	Challenge	DB	FN	Rotate	Tilt	Weather	Blur
YOLO9000[26]	93.10	42	98.80	88.60	89.60	77.30	93.30	91.80	84.20	
RPNet[5]	94.50	61	99.30	92.80	89.50	85.30	94.70	93.20	84.10	
YOLOv3[16]	98.24	52	99.94	97.72	94.35	91.04	97.87	99.00	99.73	98.60
ResNet-50[25]	97.80	26	99.70	93.50	92.70	94.50	98.50	96.20	99.10	
ILPRNET+YOLOv3[27]	96.00	52	97.10	90.50	97.20	93.30	91.60	94.60	97.90	
LPDR-RSCNet[23]	98.00	34	99.00	90.40	98.80	94.30	98.90	99.10	99.40	
MTLPR[11]	97.70	65								
YOLOv3(结合后处理)	99.20	49	99.99	98.70	96.91	97.44	99.29	99.37	99.96	98.83
本文算法(E)	99.16	100	99.99	98.60	96.61	97.42	99.31	99.36	99.96	98.69
本文算法(F)	99.26	97	99.99	98.78	97.30	97.49	99.32	99.42	99.97	98.96
本文算法(G)	99.10	126	99.98	98.56	96.43	97.29	99.30	99.28	99.96	98.43
本文算法(H)	99.25	121	99.99	98.75	97.10	97.64	99.33	99.39	99.98	98.88

后处理步骤，F表示3个检测头结合后处理步骤，G表示双检测头无后处理步骤，H表示双检测头结合后处理步骤。实验结果如表2所示。基于删减检测头的YOLOv5的本文算法在CCPD数据集上实现了99.25%的平均精准率，且在各个场景中都有较高的定位准确率。同时，考虑到YOLOv5算法具备高准确率和实时性的特点，为测试本文算法的运行效率，在3.2节所述的硬件条件下，测试得到单张图像所用时间为8.25 ms，达到较优水平。为证实本文提出的后处理算法的有效性，还训练了YOLOv3作为检测网络。如表3所示，加入后处理算法后，统计各子集上检测车牌的准确率可以发现，各子集上的检测准确率都高于原始YOLOv3，证明了本文算法的有效性。

表3 不同算法在PKU数据集上的定位准确率

Table 3 Location accuracy of different algorithms on PKU dataset

算法	平均精准率/%	检测速度/(帧·s^{-1})	准确率/%				
			G1	G2	G3	G4	G5
文献[28]	91.52	672	98.89	98.42	95.83	81.17	83.31
文献[29]	97.69	42	98.76	98.42	97.72	96.23	97.32
文献[25]	99.80	21	99.75	100	99.73	99.65	99.65
本文算法	99.84	100	100	100	100	100	99.23

为证实本文算法的泛化性能，在PKU[28]数据集上的G1～G5子集进行验证，并统计了PKU数据集下各子集的准确率。由于PKU数据集中尺寸较小的车牌数目较多，极端情况下的模糊车牌数目较少，负责小尺度检测的检测头起作用的场景相对较多。此外，PKU数据集中的车牌清晰度较高，车牌检测过程中几乎没有漏检顶点和误检顶点的现象。因此，本文选用实验E的模型，即采用3个检测头，但是不进行后处理步骤。在该模型下，与目前在PKU数据集上表现较好的车牌检测算法进行对比，如表3所示，平均精准率达到99.84%。可以看出，本文算法具有较好的泛化性，不仅能在CCPD数据集上取得优异效果，在PKU这类道路监控数据集上的效果也优于其他车牌检测算法。检测效果可视化如图14所示。在道路监控场景，本文算法依然能取得不错的检测效果，进一步证明了算法的有效性。



图14 PKU数据集车牌定位示例

Fig. 14 License plate location instances in PKU dataset

本文采用的删减检测头的算法虽然降低0.01%的平均精准率，但提升了24.7%的检测速度，提出的后处理算法在不同场景下都能起到提升平均精准率的作用，且对处理速度的影响较小。虽然本文算法通过检测车牌的顶点区域间接获得车牌的顶点坐标，能有效解决车牌的精定位问题，提出的后处理算法能有效应对图像中存在多副车牌的情况，但其在DB子集的准确率仍有提升空间。从表2中可以看出，DB场景下的测试准确率只有97.10%，低于文献[23,26]算法，即本文算法在光照变化较大的场景中检测效果一般。这主要是由2方面原因引起：①Base子集中极端光照场景下的车牌图像很少，而本文算法需要同时检测车牌和顶点区域，导致用Base子集作为训练集训练的模型对极端光照条件下的车牌顶点区域的检测效果不好。如图15(a)所示，光照过强的场景，顶点区域特征变化较大，神经网络很难提取图中车牌左上角和左下角的顶点区域特征，因此，顶点检测易出现漏检与误检的情况，进而造成精定位准确率降低。②DB场景中，有部分光照很弱场景中拍摄的车牌图像，如图15(b)所示(图15(c)为直方图均衡后的效果，仅为了易于对比)，光照过弱导致车牌轮廓清晰度较低，车牌及顶点区域都易出现漏检现象。图15(b)中所有顶点区域和车牌区域全部漏检，说明光照过暗对车牌检测及顶点检测影响很大，需要在后续工作中寻找解决方案。图16～图19为使用本文算法精定位正确的效果，主要包括车牌颜色与背景相似、光照不佳、车牌模糊、车牌倾斜角度大等场景下的图像。



(a) 光照过强, 顶点特征差异较大　　(b) 光照过弱导致轮廓不清晰　　(c) 对(b)图进行直方图均衡后的结果

图 15　DB 场景中的漏检与误检示例

Fig. 15　False and missing detection instances in DB sub-dataset



图 16　车牌与背景颜色相似情况的检测示例

Fig. 16　Instances of detection of license plate with similar background color



图 17　光照不佳情况的检测示例

Fig. 17　Instances with poor illumination



图 18　车牌模糊情况的检测示例

Fig. 18　Instances with fuzzy license plate images



图 19　车牌倾斜角度大时的检测示例

Fig. 19　Instances with tilted license plate images

可以看出, 本文算法在各种非约束条件下获取的车牌图像上均能取得不错的精定位效果。

为更好地解释算法失效的情况, 将检测错误的情况进行了图像对比。失效情况主要有 2 类: ①拍摄图像中的车牌轮廓模糊, 如图 20 所示。②CCPD 数据集的标记框与车牌的实际位置不符合, 如图 21 所示, 图中的黄色框是本文的检测框, 蓝色框是标签。对于情形①, 主要是由于光照太亮或太暗, 以及车牌出现在倾斜、模糊、低光照等极端情况, 造成车牌的部分轮廓不清晰。情形②是因为数据集本身的标签不准, 即使本文算法检测结果正确, 也被误判为检测错误。对于情形①的情况, 常造成车牌只识别一部分, 在实际应用中, 可利用车牌识别结果, 进行多次拍摄来避免这一问题。对于情形②的情况, 可对标签进行纠正, 消除计算准确率的过程中出现误判的情况。



图 20　极端条件导致的误检情况示例

Fig. 20　False detection instances under some extreme scenarios



图 21　标签错误导致的误判示例

Fig. 21　Misjudgment instances caused by false labels

4　结　论

针对矩形检测框不能准确拟合非约束环境下车牌区域, 以及目前非约束环境下车牌定位算法精度与效率偏低的问题, 本文提出一种主体与顶点目标检测相结合的车牌定位算法, 有效提升了非约束环境下的车牌检测效果。

1）通过缩减检测头的YOLOv5目标检测算法对以车牌顶点为中心的区域及车牌区域进行同步检测，在很小代价的前提下实现快速车牌检测。

2）本文算法充分利用了车牌局部信息和整体信息，相比于其他车牌精定位算法，在不引入额外参数量的前提下进一步提升了车牌检测效果，能有效应用于各种复杂的场景中。

本文算法虽在一定程度上提升了非约束环境下车牌定位的准确率，但在DB子集上的准确率还不是很高，在光照条件不理想的环境下，对顶点区域进行检测时存在着检测不准确和漏检的情况，说明算法抗光照变化的性能还有待提升。因此，后续研究可重点从3方面着手：①对训练集中DB这类样本进行有效的数据增广，增加模型对不理想光照环境下的适应能力；②针对DB场景下车牌轮廓模糊的情况，可引入图像增强算法对图像进行预处理，使车牌轮廓更清晰；③通过优化网络结构和提升训练集的多样性等方式进一步提升车牌定位整体性能。

参考文献（References）

[1] JAMTSHO Y, RIYAMONGKOL P, WARANUSAST R. Real-time license plate detection for non-helmeted motorcyclist using YOLO [J]. ICT Express, 2021, 7(1): 104-109.

[2] HUANG Q, CAI Z, LAN T. A single neural network for mixed style license plate detection and recognition[J]. IEEE Access, 2021, 9: 21777-21785.

[3] 曾丽华, 李超, 熊璋. 基于边缘与颜色信息的车牌精确定位算法[J]. 北京航空航天大学学报, 2007, 33(9): 1112-1116.
ZENG L H, LI C, XIONG Z. Precisely license plate locating algorithm based on edge and color information[J]. Journal of Beijing University of Aeronautics and Astronautics, 2007, 33(9): 1112-1116 (in Chinese).

[4] CHEN J I Z, ZONG J I. Automatic vehicle license plate detection using K-means clustering algorithm and CNN[J]. Journal of Electrical Engineering and Automation, 2021, 3(1): 15-23.

[5] XU Z B, YANG W, MENG A J, et al. Towards end-to-end license plate detection and recognition: A large dataset and baseline[C]// Proceedings of the European Conference on Computer Vision. Berlin: Springer, 2018: 255-271.

[6] SILVA S M, JUNG C R. License plate detection and recognition in unconstrained scenarios[C]//Proceedings of the European Conference on Computer Vision. Berlin: Springer, 2018: 580-596.

[7] SELMI Z, HALIMA M B, PAL U, et al. DELP-DAR system for license plate detection and recognition[J]. Pattern Recognition Letters, 2019, 129: 213-223.

[8] HE K M, GKIOXARI G, DOLLÁR P, et al. Mask R-CNN[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020, 42(2): 386-397.

[9] 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.

[10] CHEN L S, TIAN S, MA W J, et al. End-to-end trainable network for degraded license plate detection via vehicle-plate relation mining[J]. Neurocomputing, 2021, 446: 1-10.

[11] WANG W, YANG J, CHEN M, et al. A light CNN for end-to-end car license plates detection and recognition[J]. IEEE Access, 2019, 7: 173875-173883.

[12] YIN X, LIU X. Multi-task convolutional neural network for pose-invariant face recognition[J]. IEEE Transactions on Image Processing, 2017, 27(2): 964-975.

[13] WANG Y, BIAN Z P, ZHOU Y, et al. Rethinking and designing a high-performing automatic license Plate recognition approach[EB/OL]. (2021-06-17) [2022-05-20]. https://arxiv.org/abs/2011.14936.

[14] 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: 2117-2125.

[15] Ultralytics. YOLOv5[EB/OL]. (2022-05-12) [2022-05-20]. https://github.com/ultralytics/yolov5.

[16] REDMON J, FARHADI A. YOLOv3: An incremental improvement [EB/OL]. (2018-04-08) [2022-05-20]. https://arxiv.org/abs/1804.02767.

[17] ALEXEY B, WANG C Y, LIAO H. YOLOv4: Optimal speed and accuracy of object detection[EB/OL]. (2020-04-23) [2022-05-20]. https://arxiv.org/abs/2004.10934v1.

[18] WANG C Y, LIAO H, YEH I H, et al. CSPNet: A new backbone that can enhance learning capability of CNN[EB/OL]. (2019-11-27) [2022-05-20].https://arxiv.org/abs/1911.11929.

[19] HE K M, 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.

[20] LIU S, QI L, QIN H, et al. Path aggregation network for instance segmentation[EB/OL]. (2018-03-05) [2022-05-20]. https://arxiv.org/abs/1803.01534.

[21] LOSHCHILOV I, HUTTER F. SGDR: Stochastic gradient descent with warm restarts[EB/OL]. (2016-08-13) [2022-05-20]. https://arxiv.org/abs/1608.03983.

[22] 徐光柱, 匡婉, 李兴维, 等. YOLOv3与顶点偏移估计相结合的车牌定位[J]. 计算机辅助设计与图形学学报, 2021, 33(4): 569-579.
XU G Z, KUANG W, LI X W, et al. License plate location based on YOLOv3 and vertex offset estimation[J]. Journal of Computer-Aided Design & Computer Graphics, 2021, 33(4): 569-579(in Chinese).

[23] 唐倩, 贺伟, 张林江. 基于空间转换网络的端到端车牌检测与识别[J]. 光电子·激光, 2021, 32(5): 524-531.
TANG Q, HE W, ZHANG L J. End to end license plate detection and recognition based on spatial transformer network[J]. Journal of Optoelectronics·Laser, 2021, 32(5): 524-531(in Chinese).

[24] QIN S S, LIU S J. Towards end-to-end car license plate location and recognition in unconstrained scenarios[EB/OL]. (2020-08-25) [2022-05-20]. https://arxiv.org/abs/2008.10916.

[25] SILVA S M, JUNG C R. A flexible approach for automatic license plate recognition in unconstrained scenarios[J]. IEEE Transactions on Intelligent Transportation Systems, 2021, 23(6): 5693-5703.

[26] REDMON J, FARHADI A. YOLO9000: Better, faster, stronger[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2017: 7263-7271.

[27] ZOU Y, ZHANG Y, YAN J, et al. License plate detection and recognition based on YOLOv3 and ILPRNET[J]. Signal Image and Video Processing, 2021, 9: 312-320.

[28] LI B, TIAN B, LI Y, et al. Component-based license plate detection using conditional random field model[J]. IEEE Transactions on Intelligent Transportation Systems, 2013, 14(4): 1690-1699.

[29] YUAN Y, ZOU W, ZHAO Y, et al. A robust and efficient approach to license plate detection[J]. IEEE Transactions on Image Processing, 2016, 26(3): 1102-1114.

Accurate license plate location based on synchronous vertex and body region detection

XU Guangzhu[1, 2]，LIU Gaofei[2]，KUANG Wan[2]，WAN Qiubo[2]，MA Guoliang[3]，LEI Bangjun[1, 2, *]

(1. Hubei Key Laboratory of Intelligent Vision Monitoring for Hydropower Engineering (China Three Gorges University)，Yichang 443002，China；
2. College of Computer and Information Technology，China Three Gorges University，Yichang 443002，China；
3. Traffic Police Detachment，Public Security Bureau of Yichang City，Yichang 443002，China)

Abstract：A novel unconstrained license plate accurate location algorithm is designed by simultaneously detecting the four local vertex regions and the body of a license plate and fusing the results to address the issue that the widely used rectangular bounding boxes in mainstream target detection methods cannot meet the license plate location accuracy requirement in many unconstrained environments where the license plate images are not commonly rectangle. At first, the four local rectangular sub-regions with centers on four vertices of a license plate were annotated as vertex-region objects according to the size of the plate's contour-rectangle and the vertex coordinates. Then, a multi-class image dataset is built up in which the contour-rectangle region covering the whole license plate body is a class and the four kinds of vertex-region construct other four classes. In order to locate these five object classes efficiently, the output structure of the YOLOv5 network is modified by taking accuracy and efficiency into consideration and trained with the newly constructed multi-class dataset.Finally, vertex region grouping and single missing vertex forecasting are carried out as the post-processing to address the issue that there are multiple candidate license plates in an image and a few vertices region false or missing detection errors will happen in some unique instances.By exploiting the relationship among the vertexes, the post-processing can effectively recognize missing and false detection errors in some special complex scenarios and improve the whole system's performance greatly. The proposed algorithm is evaluated on the Chinese city parking dataset (CCPD), and reaches an average positioning accuracy of 99.25% and an average recall rate of 98.70%. The performance certificates our method not only can accurately predict the coordinates of the four vertices but also can run at 121 frame/s on a moderate GPU hardware platform, which has great application potential.

Keywords：deep learning；convolutional network；visual target detection；unconstrained license plate location；license plate vertex detection

Received：2022-05-20；**Accepted**：2022-07-02；**Published Online**：2022-11-02 17：15

URL：link.cnki.net/urlid/11.2625.V.20221102.1538.003

Foundation items：Hubei Provincial Central Leading Local Science and Technology Development Special Project (2019ZYYD007); Funds of Hubei Key Laboratory of Intelligent Vision Monitoring for Hydropower Engineering (2022SDSJ03)

*** Corresponding author**. E-mail：Bangjun.Lei@ieee.org

http://bhxb.buaa.edu.cn jbuaa@buaa.edu.cn
DOI: 10.13700/j.bh.1001-5965.2022.0401

基于 EM 自注意力残差的图像超分辨率重建网络

黄淑英[1]，胡瀚洋[2]，杨勇[3,*]，万伟国[4]，吴峥[2]

(1. 天津工业大学 软件学院，天津 300387；2. 江西财经大学 信息管理学院，南昌 330032；
3. 天津工业大学 计算机科学与技术学院，天津 300387；4. 江西财经大学 软件与物联网工程学院，南昌 330032)

摘 要：基于深度学习的图像超分辨率 (SR) 重建方法主要通过增加模型的深度来提升图像重建的质量，但同时增加了模型的计算代价，很多网络利用注意力机制来提高特征提取能力，但难以充分学习到不同区域的特征。为此，提出一种基于期望最大化 (EM) 自注意力残差的图像超分辨率重建网络。该网络通过改进基础残差块，构建特征增强残差块，以更好地复用残差块中所提取的特征。为增加特征信息在空间上的相关性，引入 EM 自注意力机制，构建 EM 自注意力残差模块来增强模型中每个模块的特征提取能力，并通过级联 EM 自注意力残差模块来构建整个模型的特征提取结构。所获得的特征图通过上采样的图像重建模块获得重建的高分辨率图像。将所提方法与主流方法进行实验对比，结果表明：所提方法在 5 个流行的 SR 测试集上能够取得较好的主观视觉效果和更优的性能指标。

关 键 词：超分辨率重建；注意力机制；期望最大化；特征增强残差块；EM 自注意力残差模块
中图分类号：TP391.41
文献标志码：A **文章编号**：1001-5965(2024)02-0388-10

图像超分辨率(super resolution, SR)重建技术是将退化的低分辨率(low resolution, LR)图像重建获得相对应的高分辨率(high resolution, HR)图像，该技术作为计算机视觉领域的一项基础性的底层任务，一直是研究的重点和热点。通常，图像的分辨率越高，所包含的纹理细节越丰富，视觉效果越清晰，更加适用于图像的后续处理任务。目前，很多设备能够胜任现实场景中对于清晰度的要求，但由于设备成本及拍摄环境的限制，很多场合获得的图像难以达到高分辨率的要求[1]。SR 重建技术在软件层面上解决了上述难题，一方面，其能有效降低硬件的成本，使设备的扩展维护更加灵活；另一方面，软件设计不会受到平台的限制，适用范围更广。因此，SR 重建技术在实际应用中具有巨大的前景，如公共安全[2]、医学诊断[3]、遥感观测[4]、视频信息增强等。

多年来，许多研究成果相继被提出用于解决 SR 重建任务中的不适定问题，包括基于插值的方法[5]、基于重建的方法[6-7]和基于样例学习的方法[8-11]等。近年来，随着深度卷积神经网络 (convolutional neural network, CNN) 研究的爆发式进展，基于深度学习的方法因其出色的学习能力及实时处理的能力而受到广泛的关注。

Dong 等[12]最早提出了利用 CNN 重建 SR 图像，提出的超分辨率卷积神经网络 (super-resolution convolutional neural network, SRCNN) 通过堆叠卷积层学习 LR 到 HR 的非线性映射关系来实现 HR 图像的重建，该方法的性能远超大部分基于样例学习的方法，但由于网络结构过于简单，性能受到一定的限制，虽然不断堆叠网络的层数可以有效提升网络性能，但随着网络层数的加深，会带来梯度消失/梯度爆炸的问题，使网络的训练难以进行。因此，

收稿日期：2022-05-21；**录用日期**：2022-07-02；**网络出版时间**：2022-10-18 11:11
网络出版地址：link.cnki.net/urlid/11.2625.V.20221017.1436.005
基金项目：国家自然科学基金 (61862030,62072218); 江西省自然科学基金 (20192ACB20002,20192ACBL21008)
* **通信作者**. E-mail：greatyangy@126.com

引用格式：黄淑英，胡瀚洋，杨勇，等. 基于 EM 自注意力残差的图像超分辨率重建网络 [J]. 北京航空航天大学学报，2024，50（2）：388-397. HUANG S Y，HU H Y，YANG Y，et al. Image super-resolution reconstruction network based on expectation maximization self-attention residual [J]. Journal of Beijing University of Aeronautics and Astronautics，2024，50（2）：388-397（in Chinese）.

受 ResNet[13] 的启发, Kim 等[14] 首次将残差网络引入 SR 重建任务中, 利用残差学习的方式增加网络层数, 设计了更深卷积网络的精确图像超分辨率(accurate image super-resolution using very deep convolutional network, VDSR)。随后, Kim 等[15] 利用循环残差的方式加深网络结构, 设计了深度循环卷积图像超分辨率网络(deeply-recursive convolutional network for image super-resolution, DRCN)。这 2 种方法在加深网络深度的前提下, 重建出的图像相比于 SRCNN 方法得到了不同程度的提升。受到 VDSR 的启发, 部分学者致力于开发出更深更宽的网络以提高算法的性能。Lim 等[16] 设计了增强的深度残差单幅图像超分辨网络(enhanced deep residual network for single image super-resolution, EDSR), 通过将残差模块中的批归一化层去除, 同时增加网络的深度和宽度, 使重建出的 HR 图像无论是视觉效果还是客观指标都达到更好。在 EDSR 的基础上, Zhang 等[17] 通过引入注意力机制设计了残差通道注意力网络(residual channel attention network, RCAN), 该网络的 SR 重建性能得到了进一步的提升。随后, Dai 等[18] 提出了二阶注意力单幅图像超分辨率网络(second-order attention network for single image super-resolution, SAN), 该网络在一阶注意力的基础上引入了二阶注意力机制, 相比于 RCAN, 性能再次得到了提升。

尽管增加网络的深度能提高图像 SR 重建的性能, 但与此同时也增加了网络结构的复杂性, 而过多的参数需要占用更大的计算资源, 导致算法很难在实际应用领域中得到推广。此外, 网络结构的增大也会造成模型的过拟合等问题。因此, 为解决上述问题, 提出了计算复杂度低、参数少的轻量级网络, 如基于信息蒸馏的快速准确的单图像超分辨率网络[19](fast and accurate single image super-resolution via information distillation network, IDN), 多尺度残差图像超分辨率网络[20](multi-scale residual network for image super-resolution, MSRN)和快速、精确、轻量级的级连残差超分辨网络[21](fast, accurate, and lightweight super-resolution with cascading residual network, CARN)等。虽然减小了网络规模, 但会降低 SR 重建图像的质量。另外, 一些学者的研究更加注重网络是否具有能够重建出满足人眼视觉的效果、包含更丰富纹理图像的能力。Ledig 等[22] 基于生成式对抗网络(generative adversarial network, GAN), 设计了第一个使用 GAN 的真实感单幅图像超分辨率网络(photo-realistic single image super-resolution using a generative adversarial network, SRGAN), 该网络获得的重建图像在性能指标上尽管未得到明显提升, 但拥有更丰富的纹理, 更符合人眼的视觉感知。Wang 等[23] 在优化 SRGAN 的基础上提出了增强型 GAN 网络结构, 在生成器中引入密集残差模块, 移除残差结构中的批归一化层, 将对抗损失改为相对判别器损失来提升网络的性能[24]。基于 GAN 的图像 SR 重建网络相比于其他深度学习网络能够获得更好的视觉效果, 但重建出的细节信息存在失真的现象, 因此, 该类方法不能满足某些图像领域的应用, 如医学图像的 SR 重建。

目前, 大部分的 SR 重建方法主要使用尺寸大小为 3×3 的卷积核, 虽然可以在一定程度上减少计算量, 但由于感受野较小, 无法充分学习到不同区域的特征; 同时, 现有的 SR 重建模型主要探索图像在通道间及空间上的相关性, 并不能取得很好的效果。近年来, 自注意力机制[25] 在自然语言处理领域[26] 取得显著成效, 并被成功引入 SR 重建中。例如, 基于非局部的 SR 重建网络[27] 取得了较好的重建效果, 该方法中的非局部自注意力机制通过学习一组过完备的基对像素特征进行重构, 使得每个像素充分捕获全局信息, 但是这组基的数目巨大, 且存在大量信息冗余。因此, 在保证计算量的情况下, 有效提取图像空间信息是本文研究的重点。

目前, 期望最大化 (expectation maximization, EM) 自注意力机制已成功应用于语义分割中, 能在很好地保持类间差异的同时, 缩小类内的差异。该自注意力机制放弃了在整幅图像上计算注意力图的做法, 利用 EM 算法迭代出一组紧凑的基, 在这组紧凑的基上运行注意力机制, 降低了网络的复杂度。基于此, 本文引入 EM 自注意力机制模型, 构建特征增强残差块与 EM 自注意力机制相结合的模块, 即 EM 自注意力残差模块, 增强网络的空间特征提取能力。在此基础上, 将多个 EM 自注意力残差模块级联来提取图像的深度空间特征, 利用全局残差的方式将初始的浅层信息与级联的 EM 自注意力残差模块学习到的深度空间特征进行融合, 通过重建模块获得最终的 SR 重建图像。本文主要贡献如下:

1) 基于 EM 自注意力机制, 构建特征增强残差块和 EM 自注意力机制相结合的特征提取模块, 即 EM 自注意力残差模块, 该模块能够能有效提取图像的空间特征。

2) 通过级联多个 EM 自注意力残差模块, 提出一种 EM 自注意力残差网络来实现图像的 SR 重建, 该网络能更好地融合图像的深度空间特征, 重建出纹理丰富的 HR 图像。

3）通过大量的实验及对各模块的消融实验，验证了本文提出的SR重建网络性能优于目前先进的SR重建模型。

1 SR重建网络

1.1 网络框架

目前，基于深度学习的图像SR重建方法中，大都采用了标准残差模块或密集残差块来提取图像的特征。标准残差模块虽然能够有效缓解深度网络的退化问题，但是难以充分利用网络层提取到的特征；密集残差块能够有效利用网络特征，但会增加网络结构的复杂度，易造成网络的过拟合问题。为更加有效地提取图像的空间信息，本文提出一种基于EM自注意力残差的图像SR重建网络，其主体结构如图1所示。该网络首先由单层卷积来实现浅层特征的提取，其输入通道数为3，输出通道数为64，采用3×3的卷积核。该层操作可以表示为

$$\boldsymbol{F}_0 = \mathrm{Conv_0}(\boldsymbol{I}_{\mathrm{LR}}) \tag{1}$$

式中：$\boldsymbol{F}_0$为卷积得到的初始浅层特征；Conv_0(·)表示首层卷积操作；$\boldsymbol{I}_{\mathrm{LR}}$为输入的原始LR图像。

然后，级联多个EM自注意力残差模块来提取图像的深度空间特征，该部分包含16个EM自注意力残差模块。最后，通过残差跳转连接将浅层特征与深度空间特征进行整合，输入到重建模块中获得最终的HR图像。

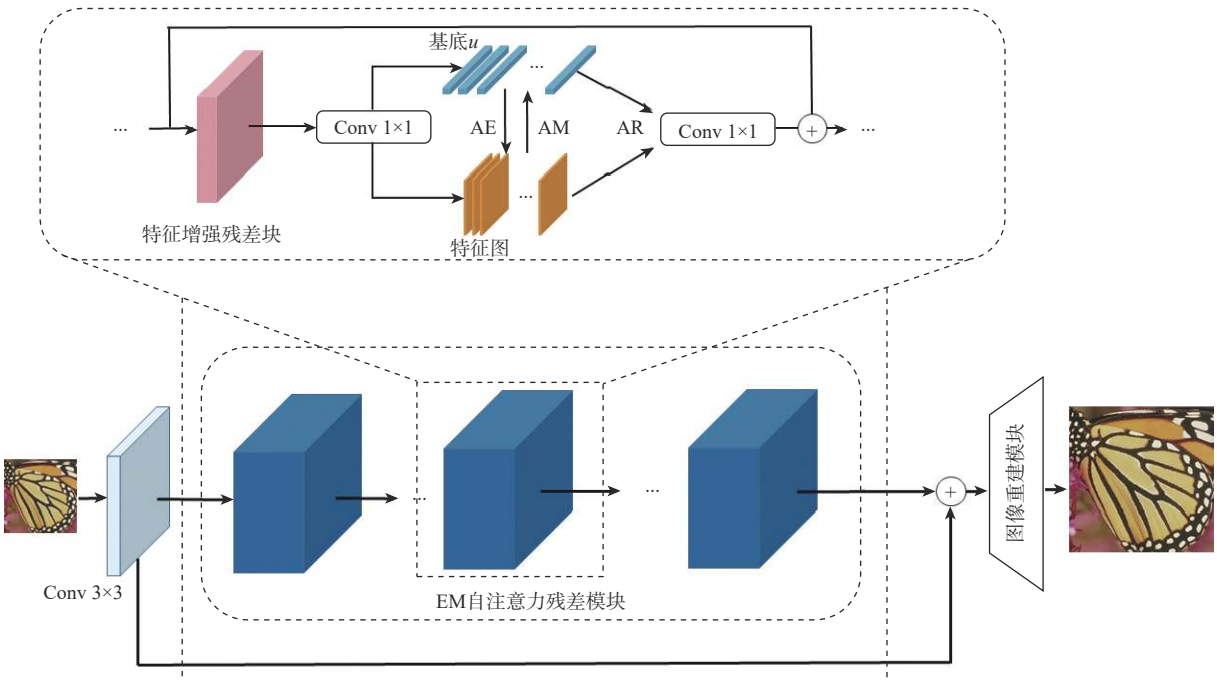


图1 基于EM自注意力残差网络的结构

Fig. 1 Network structure based on EM self-attention residual

1.1.1 EM自注意力残差模块

EM自注意力机制具有空间特征捕获的能力和降低计算复杂性的特点，为此，本文构建了一种EM自注意力残差模块来提取图像的空间特征。该模块主要由一个特征增强残差块和一个EM自注意力残差块构成，其结构如图1上部分所示。

1）特征增强残差块。原始的基础残差块（见图2(a)）在经过多次卷积的过程中，图像特征会存在丢失的现象。为更好地利用残差块中每层所提取的特征，构建了特征增强残差块（见图2(b)），用于每一个EM自注意力残差模块中来提取空间特征。该残差块通过在前2个特征层之间增加1个跳转连接和拼接操作来实现特征的复用，并增强网络的特征表达能力。每个卷积层的具体操作如下：第1个卷积层采用的卷积核大小为3×3，输入的通道数为64，输出的通道数为32，将其向后传递，与第2个卷积层的输出执行拼接操作；第2个卷积层同样采用3×3的卷积核，输入的通道数和输出通道数均为32；第3个和第4个卷积层分别采用1×1和3×3的卷积核，通道数均为64，来实现特征的提取和整合。另外，在第1个和第3个卷积层后分别添加ReLU函数来增加卷积层之间的非线性关系。

2）EM自注意力残差块。近年来，非局部网络[27]被提出后，自注意力机制在计算机视觉领域[28-29]得到了广泛的关注，并被许多研究者证明了其在语义分割中的有效性。然而，该自注意力机制需要生成一个巨大的注意力图，易产生巨大的空间复杂度和时间复杂度。EM自注意力机制能够利用EM算法

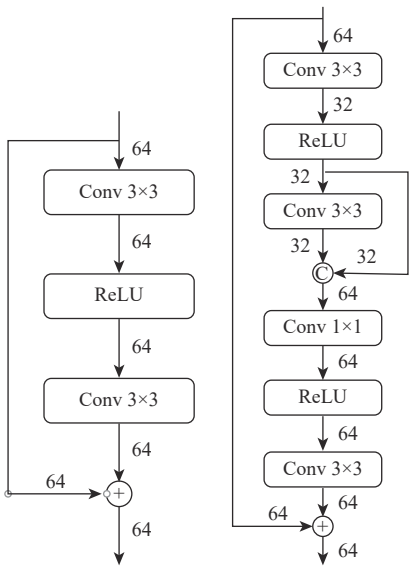


图 2　残差块结构

Fig. 2　Residual block structure

迭代出简单紧凑的基底来估算特征图在不同区域的关联性, 这种方式能够有效降低计算的成本和网络的整体参数量。

① EM 自注意力机制。EM 算法[30] 的目的在于寻找隐含变量模型的最大似然解。对于现有样本 $\boldsymbol{X}=[x_1,x_2,\cdots,x_n]$, 每个样本点 x_i 都有对应的隐藏变量 z_i。将 $\{\boldsymbol{X},\boldsymbol{Z}\}$ 构成完整的数据集, 其似然函数为 $\ln P(\boldsymbol{X},\boldsymbol{Z}|u)$, u 为模型参数。E 步为: 根据当前的参数 u^{old} 计算隐藏变量 $\boldsymbol{Z}$ 的后验分布, 并以此寻找完整数据集的似然 $Q\left(u,u^{\text{old}}\right)$, 其公式为

$$Q\left(u,u^{\text{old}}\right)=\sum_{\boldsymbol{Z}}P\left(\boldsymbol{Z}|\boldsymbol{X},u^{\text{old}}\right)\ln P(\boldsymbol{X},\boldsymbol{Z}|u) \tag{2}$$

M 步为: 通过求解最大似然函数来更新网络参数得到 u^{new}:

$$u^{\text{new}}=\arg\max_{u}Q\left(u,u^{\text{old}}\right) \tag{3}$$

EM 自注意力机制[31] 主要由注意力估计 (attention estimation, AE)、注意力最大化 (attention maximization, AM) 和注意力再估计 (attention re-estimation, AR) 3 部分组成, 如图 3 所示。AE 与 AM 分别对应 EM 算法的 E 步和 M 步。AE 用来估计隐含变量 $\boldsymbol{Z}$, AM 的作用为利用最大似然算法更新基底 u。AE 和 AM 交替进行, 经过若干次迭代后得到近似收敛的基底 u 和隐含变量 $\boldsymbol{Z}$, 就可以利用 AR 得到 $\boldsymbol{X}$ 的近似评估 $\widetilde{\boldsymbol{X}}$:

$$\widetilde{\boldsymbol{X}}=u\boldsymbol{Z} \tag{4}$$

② 残差块结构。EM 自注意力残差模块由特征增强残差块和 EM 自注意力机制构成, 通过残差方式将特征增强残差块与提取到的注意力特征进行融合, 得到增强的深度空间信息, 如图 1 上部分所示。首先, 构建特征增强残差块来提取空间特征; 然后, 在 EM 自注意力机制前后增加一个 1×1 的卷积层, 分别实现对输入特征图的压缩, 并将经过 EM 自注意力机制增强的特征重新映射回残差空间中; 最后, 将输入的特征与该模块得到的残差特征进行融合得到输出特征图。EM 自注意力残差模块的数学公式为

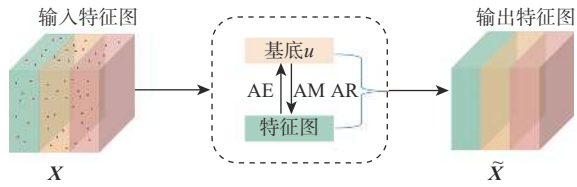


图 3　EM 自注意力机制

Fig. 3　EM self-attention mechanism

$$\boldsymbol{F}_i=H_{\text{EMA}}\left(H_{\text{df}}\left(\boldsymbol{F}_{i-1}\right)\right)+\boldsymbol{F}_{i-1} \tag{5}$$

$$H_{\text{df}}\left(\boldsymbol{F}_{i-1}\right)=\text{Conv}_{3\times3}\left(\text{ReLU}\left(\text{Conv}_{1\times1}\left(\text{concat}\left(\text{Conv}_{3\times3}\left(\boldsymbol{F}_{\text{c1}}\right),\right.\right.\right.\right.$$
$$\left.\left.\left.\left.\text{half}\left(\boldsymbol{F}_{\text{c1}}\right)\right)\right)\right)\right)+\boldsymbol{F}_{i-1} \tag{6}$$

$$\boldsymbol{F}_{\text{c1}}=\text{ReLU}\left(\text{Conv}_{3\times3}\left(\boldsymbol{F}_{i-1}\right)\right) \tag{7}$$

式中: $\boldsymbol{F}_i(i=1,\cdots,n)$ 为第 i 个 EM 自注意力残差模块获得的特征图; $H_{\text{EMA}}(\cdot)$ 表示 EM 自注意力机制操作; $H_{\text{df}}(\cdot)$ 表示特征增强残差块操作; ReLU(·) 表示 ReLU 激活函数; $\boldsymbol{F}_{\text{c1}}$ 为第 1 个卷积层和 ReLU 函数输出的特征图; $\text{Conv}_{i\times i}(\cdot)$ 表示卷积核尺寸为 $i\times i$ 的卷积操作; concat(·) 表示合并操作; half(·) 表示取一半通道数的特征图。

1.1.2　图像重建模块

针对模块学习到的特征信息, 利用全局残差的方式将初始的浅层特征与级联的 EM 自注意力残差模块学习到的深度空间特征融合, 融合后的特征图经过重建模块获得最终的 HR 图像。该重建模块由 2 个 3×3 卷积层、1 个 ReLU 激活函数和 PixelShuffle 函数构成, 输入通道数为 64, 输出通道数为 3, 具体结构如图 4 所示。图像重建模块结构可以表示为

$$\boldsymbol{I}_{\text{SR}}=\text{Conv}_{3\times3}\left(\text{up}\left(\text{ReLU}\left(\text{Conv}_{3\times3}\left(\boldsymbol{F}_0+\boldsymbol{F}_n\right)\right)\right)\right) \tag{8}$$

式中: $\boldsymbol{I}_{\text{SR}}$ 为重建得到的 HR 图像; $\boldsymbol{F}_n$ 为最后一个 EM 自注意力残差模块的输出; up(·) 表示 PixelShuffle 上采样操作。

1.2　网络设置

整个网络共有 1 个卷积层用于提取浅层特征, 以及 16 个 EM 自注意力残差模块, 最后用 1 个图像重建模块获得最终的 SR 重建图像。在所有卷积层中, 除了 3×3 大小卷积核外, 还采用了 1×1 大小的卷积核, 来增强模型的非线性, 减少参数量, 提高计

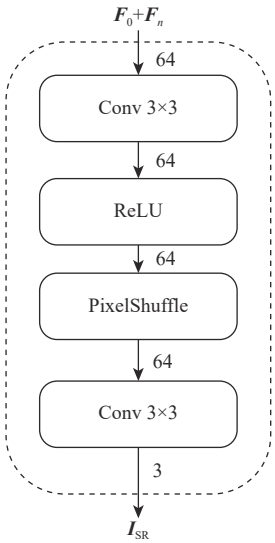


图4 图像重建模块结构

Fig. 4 Image reconstructed module structure

算速度。

2 实 验

本节在DIV2K_900数据集[32]上验证本文方法的有效性，并与其他8种较先进的SR重建方法进行比对，包括SRCNN[12]、VDSR[14]、DRCN[15]、MS-LapSRN[33]、SRMDNF[34]、LESRCNN[35]、PAN[36]、SMSR[37]等。双三次插值(Bicubic)方法的结果作为比较的基准图像。

为更好地比较不同方法的性能，展示了对比方法的主观效果和客观指标，并对本文提出模型中的各模块进行了消融实验，来验证各模块的有效性。

2.1 实验设置

本文模型采用ADAM优化函数来优化网络，初始学习率设为3×10^{-4}，每迭代2×10^5次，学习率减半。本文方法训练与测试的硬件环境搭配Intel酷睿i5 9400F、内存32 GB、配置NVIDIA GeForce RTX 2060 GPU的计算机，软件环境为64位Win10操作系统，代码书写采用PyTorch框架。

该网络使用DIV2K_900数据集[32]作为训练集，包含900组HR图像和LR图像对。为增加网络的泛化性，对训练数据集进行扩充，扩充方式参考EDSR方法的数据增强方式。本文方法的测试集为5个标准数据集：Set5[38]、Set14[9]、BSD100[39]、Urban100[40]和Manga109[41]。测试集的客观评价指标采用峰值信噪比PSNR和结构相似度SSIM。

本文模型采用L_1损失函数来训练，L_1损失函数相较于其他损失函数能更好地保持重建出的图像细节。

2.2 实验结果及分析

为客观评估方法的性能，表1给出了本文方法和对比方法在各数据集上的平均PSNR与SSIM值。可以看出，简单的Bicubic方法的平均PSNR与SSIM指标是最差的，因为该方法仅利用周围的像素点简单地计算要插入像素点的值，所以很难重建出更多的纹理细节，并且放大因子越大，重建结果越模糊。较早的深度学习的SR重建方法虽然网络比较简单，但也表现出了网络的强大特征表达能力，如SRCNN在不同放大因子的情况下，其平均PSNR和SSIM的值都存在较大的提升。随着后期网络模型的加深，SR重建结果在2个指标上都获得了较高的数值。从结果中可以看出，本文方法与其他几种先进的SR重建方法相比，在3种放大因子下获得的2个指标值几乎都是最大的，只有Manga109数据集上，2个指标值部分小于其他方法，但都接近最优值。

为更直观地观察各种方法的重建结果，图5和图6给出了本文方法与其他对比方法在BSD100数据集中2幅图像放大因子为2和4时的结果，并对局部的细节区域进行放大。可以看出，Bicubic方法重建出的结果比较模糊，不能获得清晰的纹理和边缘信息。其他深度学习的方法虽然重建出了更多的细节信息，能够观察到图像的边缘，但是边缘处仍然存在模糊现象。从局部放大区域可以看到，本文方法的结果获得的边缘更锐化，视觉效果也最好，与原始参考图像GT最接近。

同时，为比较不同SR重建方法的计算复杂度，测试了各方法的平均运行时间。表2给出了放大因子为2时不同方法在Set5数据集上SR重建的平均运行时间结果。综合考虑表1的客观评价指标和表2的运行时间可以看出，本文方法在获得最优的PSNR和SSIM指标值的同时，运行时间也在毫秒级别，可以达到实时的图像SR重建。

2.3 感知性能评估

随着图像超分辨重建研究的不断深入，常用于定量衡量图像重建质量的2种指标(PSNR和SSIM)不再满足人类的视觉需求，会存在重建图像的2种指标比较高反而视觉效果比较差的情况。因此，本文引入学习感知图像块相似度LPIPS来定量衡量本文方法的感知性能，LPIPS越小，表示性能越好。表3给出了放大因子为4时不同方法在Set5数据集上的LPIPS对比结果。为公平起见，选择的对比方法均未使用感知损失进行模型训练。可以看出，本文方法在所有对比方法中表现出了最好的感知性能。

2.4 消融实验

为验证本文网络中的EM自注意力残差模块

表 1　放大因子分别为 2、3、4 时，不同方法在各数据集上 SR 重建图像的 PSNR 和 SSIM 平均结果

Table 1　Average PSNR and SSIM results of SR reconstructed images obtained by different methods under different magnification factors of 2, 3, and 4 on each dataset

方法	放大因子	平均PSNR/dB					平均SSIM				
		Set5	Set14	BSD100	Urban100	Manga109	Set5	Set14	BSD100	Urban100	Manga109
Bicubic	2	33.66	30.24	29.56	26.88	31.05	0.929 9	0.868 8	0.843 1	0.840 3	0.935
SRCNN[12]	2	36.66	32.45	31.36	29.50	35.72	0.954 2	0.906 7	0.887 9	0.894 6	0.968
VDSR[14]	2	37.53	33.03	31.90	30.76	37.16	0.958 7	0.912 4	0.896 0	0.914 0	0.974
DRCN[15]	2	37.63	33.04	31.85	30.75	37.57	0.958 8	0.911 8	0.894 2	0.913 3	0.973
MS-LapSRN[33]	2	37.78	33.28	32.05	31.15	37.78	0.960 0	0.915 0	0.898 0	0.919 0	0.976
SRMDNF[34]	2	37.79	33.32	32.05	31.33	38.07	0.960 1	0.915 9	0.898 5	0.920 4	0.976
LESRCNN[35]	2	37.65	33.32	31.95	31.45	38.49	0.958 6	0.914 8	0.896 4	0.920 6	0.977 7
PAN[36]	2	38.00	33.59	32.18	32.01	38.70	0.960 5	0.918 1	0.899 7	0.927 3	0.977 3
SMSR[37]	2	38.00	33.64	32.17	32.19	38.76	0.960 1	0.917 9	0.899 0	0.928 4	0.977 1
本文方法	2	38.00	33.69	32.18	32.18	38.57	0.960 6	0.920 2	0.899 7	0.929 3	0.977 4
方法	放大因子	平均PSNR/dB					平均SSIM				
		Set5	Set14	BSD100	Urban100	Manga109	Set5	Set14	BSD100	Urban100	Manga109
Bicubic	3	30.39	27.55	27.21	24.46	26.95	0.868 2	0.774 2	0.738 5	0.734 9	0.856
SRCNN[12]	3	32.75	29.30	28.41	26.24	30.48	0.909 0	0.821 5	0.786 3	0.798 9	0.912
VDSR[14]	3	33.66	29.77	28.82	27.14	32.01	0.921 3	0.831 4	0.797 6	0.827 9	0.934
DRCN[15]	3	33.82	29.76	28.80	27.15	32.31	0.922 6	0.831 1	0.796 3	0.827 6	0.936
MS-LapSRN[33]	3	34.06	29.97	28.93	27.47	32.68	0.924 0	0.836 0	0.802 0	0.837 0	0.939
SRMDNF[34]	3	34.12	30.04	28.97	27.57	33.00	0.925 4	0.838 2	0.802 5	0.839 8	0.940 3
LESRCNN[35]	3	33.93	30.12	28.91	27.70	33.15	0.923 1	0.838 0	0.800 5	0.841 5	0.943 3
PAN[36]	3	34.40	30.36	29.11	28.11	33.61	0.927 1	0.842 3	0.805 0	0.851 1	0.944 8
SMSR[37]	3	34.40	30.33	29.10	28.25	33.68	0.927 0	0.841 2	0.805 0	0.853 6	0.944 5
本文方法	3	34.45	30.41	29.12	28.31	33.62	0.927 8	0.844 2	0.806 0	0.855 4	0.945 0
方法	放大因子	平均PSNR/dB					平均SSIM				
		Set5	Set14	BSD100	Urban100	Manga109	Set5	Set14	BSD100	Urban100	Manga109
Bicubic	4	28.42	26.00	25.96	23.14	25.15	0.810 4	0.702 7	0.667 5	0.657 7	0.789
SRCNN[12]	4	30.48	27.50	26.90	24.52	27.66	0.862 8	0.751 3	0.710 1	0.722 1	0.858
VDSR[14]	4	31.35	28.01	27.29	25.18	28.82	0.883 8	0.767 4	0.725 1	0.752 4	0.886
DRCN[15]	4	31.53	28.02	27.23	25.14	28.97	0.885 4	0.767 0	0.723 3	0.751 0	0.886
MS-LapSRN[33]	4	31.74	28.26	27.43	25.51	29.54	0.889 0	0.774 0	0.731 0	0.768 0	0.897
SRMDNF[34]	4	31.96	28.35	27.49	25.68	30.09	0.892 5	0.778 7	0.733 7	0.773 1	0.902
LESRCNN[35]	4	31.88	28.44	27.45	25.77	30.49	0.890 3	0.777 2	0.731 3	0.773 2	0.909 1
PAN[36]	4	32.13	28.61	27.59	26.11	30.51	0.894 8	0.782 2	0.736 3	0.785 4	0.909 5
SMSR[37]	4	32.12	28.55	27.55	26.11	30.54	0.893 2	0.780 8	0.735 1	0.786 8	0.908 5
本文方法	4	32.25	28.69	27.59	26.20	30.50	0.895 3	0.784 1	0.737 0	0.789 1	0.909 2



图 5　放大因子为 2 时 BSD100_37073 数据集中一幅图像重建结果的可视化比较

Fig. 5　Visualization comparison of reconstruction results of an image in BSD100_37373 dataset under a magnification factor of 2

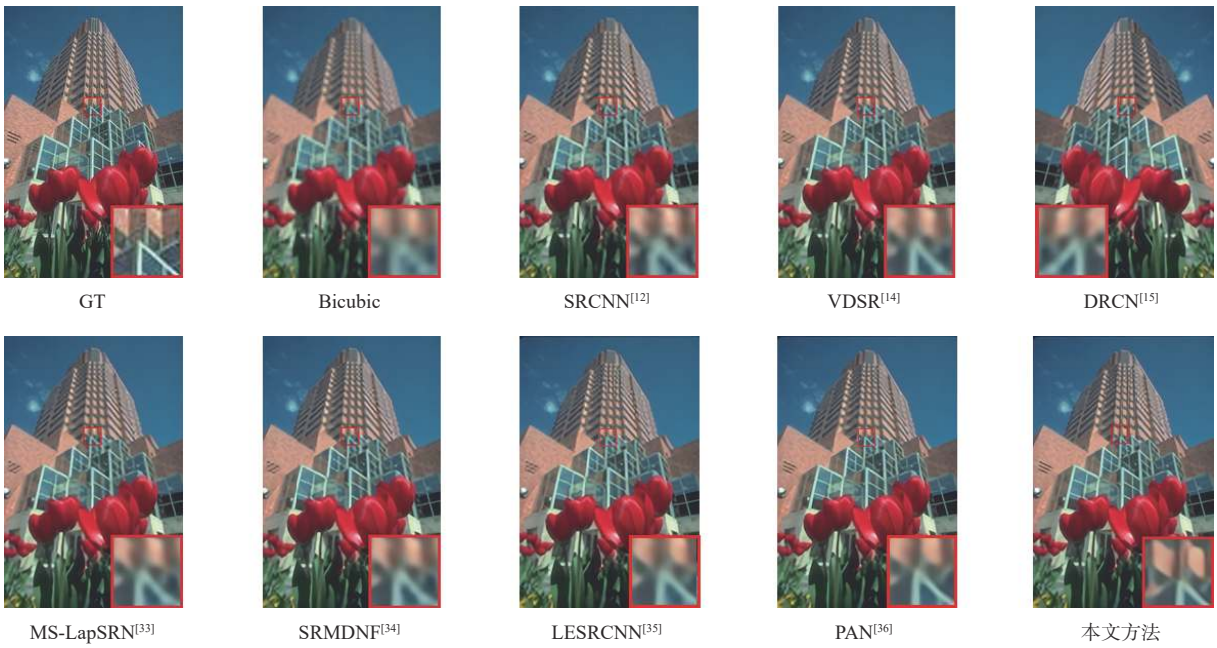


图 6 放大因子为 4 时 BSD100_86000 数据集中一幅图像重建结果的可视化比较

Fig. 6 Visualization comparison of reconstruction results of an image in BSD100_86000 dataset under a magnification factor of 4

表 2 放大因子为 2 时不同方法在 Set5 数据集上的平均运行时间

Table 2 Average running time of different methods on Set5 dataset under a magnification factor of 2

方法	平均运行时间/s
SRCNN[12]	0.002 4
VDSR[14]	0.005 2
DRCN[15]	0.103 6
MS-LapSRN[33]	0.038 4
SRMDNF[34]	0.187 3
LESRCNN[35]	0.002 6
PAN[36]	0.035 4
本文方法	0.090 2

表 3 放大因子为 4 时不同方法在 Set5 数据集上的 LPIPS 结果

Table 3 LPIPS results of different methods on Set5 dataset under a magnification factor of 4

方法	LPIPS
Bicubic	0.340 7
SRCNN[12]	0.242 1
VDSR[14]	0.224 6
DRCN[15]	0.228 9
LESRCNN[35]	0.218 8
PAN[36]	0.216 5
本文方法	0.214 2

和特征增强残差块的有效性，对不同结构进行了消融实验，并对其结果进行了对比分析。考虑实验的时间成本，不同结构下的模型都采用 Set5 数据集，放大因子为 4，实验结果如表 4 所示。表中，

表 4 放大因子为 4 时不同模块在 Set5 数据集上的比较结果

Table 4 Comparsion results of different blocks on Set5 dataset under a magnification factor of 4

模型	PSNR/dB	SSIM	参数量/10^3	计算量/10^9 FLOP
DARN_res	32.20	0.895 1	1 649	1 952
DARN_dr	32.08	0.894 0	1 436	1 756
DARN_ema	31.79	0.893 0	467	863
DARN_ca	32.11	0.894 2	1 445	1 757
DARN	32.25	0.895 3	1 568	1 878

注：FLOP为浮点运算次数，使用尺寸为1 280×720的图像来测试。

DARN_dr 表示模型中仅包含特征增强残差模块，DARN_ema 表示网络仅包含 EM 自注意力残差模块，DARN_ res 表示网络包含 EM 自注意力残差模块和基础残差块，DARN_ ca 表示网络包含通道注意力模块和特征增强残差块，DARN 为本文网络，包含 EM 自注意力残差模块和特征增强残差块。可以看出，当网络中只包含 1 种模块时，即特征增强块或 EM 自注意力残差模块，所获得的 PSNR 和 SSIM 值要低于同时包含 2 种模块的网络结果。将本文特征增强残差块换成基础残差块后，得到的 PSNR 和 SSIM 结果与本文方法的结果相比明显降低，尤其是 PSNR 值降低了 0.05 dB。同时，在未引入 EM 自注意力残差模块的情况下，参数量和计算量并无太大的变化，但 PSNR 和 SSIM 的结果大幅下降，而将 EM 自注意力残差模块换成通道注意力模块后，参数量与计算量和本文方法大体相等，但 PSNR 值降低了 0.14 dB。因此，从实验结果可以看出，本文方法构建的特征增强残差块和 EM 自注意

力残差模块具有良好的性能。

3　结　论

针对目前深度学习中网络结构过于复杂及空间信息提取计算代价大的问题，本文通过引入 EM 自注意力机制，提出了一种基于 EM 自注意力残差的图像 SR 重建网络。

1）通过改进基础残差块构建了一种特征增强残差块，来提升特征层间的特征复用能力；并通过引入自注意力机制，增加空间特征的相关性，构建了一种 EM 自注意力残差模块，来增强网络的空间特征提取能力。

2）为降低网络在深度空间信息提取过程中造成的信息丢失，采用全局残差的方式，将初始层得到的浅层特征与级联的 EM 自注意力残差模块提取的深度空间信息进行融合，经过上采样的图像重建模块得到重建的 HR 图像。

3）将本文模型与先进的 SR 重建模型进行对比，并对模型中构建的模块做了消融实验比较，来验证各模块的有效性。实验结果表明，本文方法能够重建出细节更丰富、边缘纹理更清晰的 HR 图像，获得了更好的视觉效果，并且取得了最优的客观评价指标值。

未来计划考虑视频图像的特性，将提出的 SR 重建方法应用于实际的 LR 视频重建中，使观察者能够体验到更高分辨率的视频效果。

参考文献（References）

[1] TAKEDA H, MILANFAR P, PROTTER M, et al. Super-resolution without explicit subpixel motion estimation[J]. IEEE Transactions on Image Processing, 2009, 18(9): 1958-1975.

[2] ZOU W W W, YUEN P C. Very low resolution face recognition problem[J]. IEEE Transactions on Image Processing, 2012, 21(1): 327-340.

[3] SHI W, CABALLERO J, LEDIG C, et al. Cardiac image super-resolution with global correspondence using multi-atlas PatchMatch [C]//Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention. Berlin: Springer, 2013, 8151: 9-16.

[4] ARUN P V, BUDDHIRAJU K M, PORWAL A, et al. CNN based spectral super-resolution of remote sensing images[J]. Signal Processing, 2020, 169: 107394.

[5] KEYS R. Cubic convolution interpolation for digital image processing[J]. IEEE Transactions on Acoustics, Speech, and Signal Processing, 1981, 29(6): 1153-1160.

[6] MARQUIAN A, OSHER S J. Image super-resolution by TV-regularization and Bregman iteration[J]. Journal of Scientific Computing, 2008, 37(3): 367-382.

[7] DONG W, ZHANG L, SHI G, et al. Image deblurring and super-resolution by adaptive sparse domain selection and adaptive regularization[J]. IEEE Transactions on Image Processing, 2011, 20(7): 1838-1857.

[8] YANG J, WRIGHT J, HUANG T S, et al. Image super-resolution via sparse representation[J]. IEEE Transactions on Image Processing, 2010, 19(11): 2861-2873.

[9] 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, 2010: 711-730.

[10] HU Y, WANG N, TAO D, et al. SERF: A simple, effective, robust, and fast image super-resolver from cascaded linear regression[J]. IEEE Transactions on Image Processing, 2016, 25(9): 4091-4102.

[11] HUANG Y, LI J, GAO X, et al. Single image super-resolution via multiple mixture prior models[J]. IEEE Transactions on Image Processing, 2018, 27(12): 5904-5917.

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

[13] HE K, ZHANG X, REN S, 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.

[14] 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.

[15] KIM J, LEE J K, LEE K M. Deeply-recursive convolutional network for image super-resolution[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2016: 1637-1645.

[16] 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. Piscataway: IEEE Press, 2017: 1132-1140.

[17] ZHANG Y L, LI K P, LI K, et al. Image super-resolution using very deep residual channel attention networks[C]//Proceedings of the European Conference on Computer Vision. Berlin: Springer, 2018, 11211: 286-310.

[18] DAI T, CAI J, ZHANG Y, et al. Second-order attention network for single image super-resolution[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2019: 11057-11066.

[19] HUI Z, WANG X, GAO X. Fast and accurate single image super-resolution via information distillation network[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2018: 723-731.

[20] LI J, FANG F, MEI K, et al. Multi-scale residual network for image super-resolution[C]//Proceedings of the European Conference on Computer Vision. Berlin: Springer, 2018, 11212: 527-542.

[21] 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, 11214: 256-272.

[22] LEDIG C, THEIS L, HUSZAR F, et al. Photo-realistic single image super-resolution using a generative adversarial network[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Re-

cognition. Piscataway: IEEE Press, 2017: 105-114.

[23] WANG X, YU K, WU S, et al. ESRGAN: Enhanced super-resolution generative adversarial networks[C]//Proceedings of the European Conference on Computer Vision. Berlin: Springer, 2018, 11133: 63-79.

[24] YADAV S, CHEN C, ROSS A. Relativistic discriminator: A one-class classifier for generalized iris presentation attack detection [C]//Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. Piscataway: IEEE Press, 2020: 2624-2633.

[25] LEI J, LIU C, JIANG D. Fault diagnosis of wind turbine based on long short-term memory networks[J]. Renewable Energy, 2019, 133: 422-432.

[26] TAN Z, WANG M, XIE J, et al. Deep semantic role labeling with self-attention[C]//Proceedings of the AAAI Conference on Artificial Intelligence. Palo Alto: AAAI, 2018.

[27] WANG X, GIRSHICK R, GUPTA A, et al. Non-local neural networks[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2018: 7794-7803.

[28] 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, 11211: 3-19.

[29] SANG H, ZHOU Q, ZHAO Y. PCANet: Pyramid convolutional attention network for semantic segmentation[J]. Image and Vision Computing, 2020, 103: 103997.

[30] 赵杨璐, 段丹丹, 胡饶敏, 等. 基于EM算法的混合模型中子总体个数的研究[J]. 数理统计与管理, 2020, 39(1): 35-50.
ZHAO Y L, DUAN D D, HU R M, et al. On the number of components in mixture model based on EM algorithm[J]. Journal of Applied Statistics and Management, 2020, 39(1): 35-50(in Chinese).

[31] LI X, ZHONG Z, WU J, et al. Expectation-maximization attention networks for semantic segmentation[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE Press, 2019: 9166-9175.

[32] AGUSTSSON E, TIMOFTE R. NTIRE 2017 challenge on sngle image super-resolution: Dataset and study[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2017: 1122-1131.

[33] LAI W S, HUANG J B, AHUJA N, et al. Deep Laplacian pyramid networks for fast and accurate super-resolution[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2017: 5835-5843.

[34] ZHANG K, ZOU W, ZHANG L. Learning a single convolutional super-resolution network for multiple degradations[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2018: 3262-3271.

[35] TIAN C, ZHUGE R, WU Z, et al. Lightweight image super-resolution with enhanced CNN[J]. Knowledge-Based Systems, 2020, 205: 106235.

[36] ZHAO H, KONG X, HE J, et al. Efficient image super-resolution using pixel attention[C]//Proceedings of the European Conference on Computer Vision. Berlin: Springer, 2020, 12537: 56-72.

[37] WANG L G, DONG X Y, WANG Y Q, et al. Exploring sparsity in image super-resolution for efficient inference[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2021: 4915-4924.

[38] BEVILACQUA M, ROUMY A, GUILLEMOT C, et al. Low-complexity single-image super-resolution based on nonnegative neighbor embedding[C]//Procedings of the British Machine Vision Conference. Berlin: Springer, 2012: 135.1-135.10.

[39] 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, 2: 416-423.

[40] 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.

[41] MATSUI Y, ITO K, ARAMAKI Y, et al. Sketch-based manga retrieval using Manga109 dataset[J]. Multimedia Tools and Applications, 2017, 76(20): 21811-21838.

Image super-resolution reconstruction network based on expectation maximization self-attention residual

HUANG Shuying[1]，HU Hanyang[2]，YANG Yong[3,*]，WAN Weiguo[4]，WU Zheng[2]

(1. School of Software，Tiangong University，Tianjin 300387，China；
2. School of Information Management，Jiangxi University of Finance and Economics，Nanchang 330032，China；
3. School of Computer Science and Technology，Tiangong University，Tianjin 300387，China；
4. School of Software and Internet of Things Engineering，Jiangxi University of Finance and Economics，Nanchang 330032，China)

Abstract: In recent years, most deep learning-based image super-resolution (SR) reconstruction methods mainly improve the quality of image reconstruction by increasing the depth of the model, while also increasing the computational cost of the model. Additionally, a lot of networks have implemented the attention mechanism to enhance their capacity for feature extraction, but it is still challenging to properly understand the properties of various regions. In response to the above problems, this paper proposes a novel SR reconstruction network based on expectation maximization (EM) self-attention residual. The network constructs a feature-enhanced residual block by improving the basic residual block to better reuse the features extracted from the residual block. In order to increase the spatial correlation of the feature information, an EM self-attention residual block is constructed by introducing the EM self-attention mechanism, which is used to enhance the feature extraction capability of each module in the deep network model. Moreover, the feature extraction structure of the entire model is constructed by cascading EM self-attention residual blocks. Finally, a reconstructed high-resolution image is obtained through an up-sampling image reconstruction module.In order to verify the effectiveness of the proposed method, this paper has carried out comparison experiments with some mainstream methods. The experimental results show that the proposed method can achieve better subjective visual effects and better objective evaluation indicators on five popular widely used SR test datasets.

Keywords: super-resolution reconstruction; attention mechanism; expectation maximization; feature-enhanced residual block; EM self-attention residual block

Received：2022-05-21；**Accepted**：2022-07-02；**Published Online**：2022-10-18 11：11

URL：link.cnki.net/urlid/11.2625.V.20221017.1436.005

Foundation items：National Natural Science Foundation of China (61862030,62072218); Natural Science Foundation of Jiangxi, China (20192ACB20002, 20192ACBL21008)

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

http://bhxb.buaa.edu.cn jbuaa@buaa.edu.cn
DOI: 10.13700/j.bh.1001-5965.2022.0424

融合粗糙数据推理的离散麻雀搜索算法求解HFSP问题

周宁*，张嵩霖，张晨
(兰州交通大学 电子与信息工程学院，兰州 730070)

摘 要：针对麻雀搜索算法(SSA)易陷入局部最优、无法求解离散优化问题等不足，提出了一种改进离散麻雀搜索算法(IDSSA)。抽象原始麻雀搜索算法的位置更新公式，针对个体的不同身份设计新的离散化启发式位置更新策略，并针对混合流水车间调度问题(HFSP)设计了编码与解码方式；引入粗糙数据推理理论，通过数学证明解释了引入理论的可行性与合理性，为算法提供理论支撑，提高可解释性；利用上近似的性质扩大搜索空间，提高种群多样性，避免算法早熟，结合划分及粗糙数据推理提出3种策略，促进种群间信息共享，调节种群的开发能力与探索能力，降低算法陷入局部最优的概率；使用改进离散麻雀搜索算法求解混合流水车间调度问题，对3个小规模实例与10个Liao经典测试集进行仿真实验，验证了改进离散麻雀搜索算法求解混合流水车间调度问题的可行性，通过与遗传算法、差分进化算法等经典算法的对比实验，证明了所提算法的优越性与改进策略的有效性。

关 键 词：麻雀搜索算法；离散化算法；粗糙集理论；全局优化；近似算法；数据关联；混合流水车间调度
中图分类号：TP301.6; TB181
文献标志码：A **文章编号**：1001-5965(2024)02-0398-11

麻雀搜索算法(sparrow search algorithm, SSA)由Xue和Shen等[1]提出，模拟了自然界中麻雀种群觅食与躲避天敌的行为，收敛速度快、求解精度高，但在迭代后期存在种群多样性减少、易陷入局部最优的问题。为解决这些问题，提升算法的综合寻优能力，吴新忠等[2]引入了混沌算子、自适应步长、变异算子与贪婪选择策略对SSA进行优化，并将其应用于矿井风流调控。付华和刘昊[3]引入精英混沌反向学习策略优化了SSA的初始化过程，结合鸡群算法优化了位置更新公式，引入变异算子提高了算法跳出局部最优的能力，并将其应用于煤与瓦斯突出危险性辨识。韩统等[4]引入对数螺旋与自适应步长策略，提升了SSA的寻优性能，并将其应用于无人机协同航迹规划。熊智等[5]引入混沌序列对SSA的初始化策略进行改进，引入自适应变异算子提高了算法全局寻优能力，并将其应用于无人机集群置信传播协同定位。SSA在连续优化问题中的良好表现使其受到广泛研究，但无法求解离散问题的不足却严重阻碍了其发展。目前，对SSA离散化的研究较少，熊国文等[6]设计了新的自然数编码与启发式策略，将SSA进行离散化，并用离散化后的算法求解了混合车辆路径优化问题。

混合流水车间调度问题(hybrid flow-shop scheduling problem, HFSP)是智能制造及工艺自动化领域的子问题，其求解方法大致分为确定性算法与群体智能算法。群体智能算法因其轻量化、成本低、求

收稿日期：2022-05-27；**录用日期**：2022-08-01；**网络出版时间**：2022-09-06 10:16
网络出版地址：link.cnki.net/urlid/11.2625.V.20220905.1326.002
基金项目：国家自然科学基金(61650207,61963023)；兰州交通大学天佑创新团队(TY202003)
* **通信作者**. E-mail：zhouning@lzjtu.edu.cn

引用格式：*周宁，张嵩霖，张晨. 融合粗糙数据推理的离散麻雀搜索算法求解HFSP问题[J]. 北京航空航天大学学报，2024，50(2)：398-408. ZHOU N，ZHANG S L，ZHANG C. Discrete sparrow search algorithm incorporating rough data-deduction for solving hybrid flow-shop scheduling problems[J]. Journal of Beijing University of Aeronautics and Astronautics，2024，50(2)：398-408 (in Chinese).*

解速度快等特点被广泛研究。Wu 和 Cao[7] 提出了贪婪模拟退火算法用于求解 HFSP 问题, 并将其应用于无缝钢管的制作中。Liang 等[8] 结合邻域搜索与跳蛙算法提出了混合跳蛙算法, 改善了算法易陷入局部最优的不足。Zhang 等[9] 提出了一种迭代贪婪算法用于求解 HFSP 问题, 并对铜带冷轧问题进行了建模和求解。Sun 等[10] 提出了一种分布式混合遗传算法用于求解 HFSP 问题, 引入新的下界规则与禁忌条件, 提升了算法的收敛速度与局部搜索的运行速度。Dong[11] 提出了一种自适应花授粉算法求解 HFSP 问题, 通过设置自适应参数帮助算法跳出局部最优, 提高了算法的收敛速度。Wang 等[12] 将鸟群算法离散化, 提出了改进鸟群优化算法, 引入复位机制, 帮助算法跳出局部最优, 并对低碳 HFSP 问题进行建模与求解。Schumacher 等[13] 扩展 HFSP 问题的约束条件, 使其更符合实际, 并结合局部搜索与禁忌搜索对问题进行求解。

HFSP 作为 NP 难问题[14], 其具有较大的解空间, 且往往包含多个局部最优解, 这些特征使部分群体智能算法无法得出可接受解, 而“没有免费午餐”定理[15] 也指出, 针对优化问题特性设计算法是非常必要的。SSA 具有较快的收敛速度与求解精度, 在面对解空间较大的高维问题时也有良好表现, 但其全局搜索能力不足, 易陷入局部最优, 无法求解包括 HFSP 在内的离散优化问题, 因此, 本文提出了一种改进的离散麻雀搜索算法 (improved discrete sparrow search algorithm, IDSSA)。首先, 对 SSA 的位置更新公式进行抽象, 结合其内在逻辑, 设计新的离散位置更新策略, 将 SSA 改造为离散麻雀搜索算法(discrete sparrow search algorithm, DSSA); 然后, 引入逆序变异初始化与粗糙数据推理理论对 DSSA 进行改进与优化, 扩大算法的搜索空间, 补充搜索策略, 增强算法跳出局部最优的能力; 最后, 使用 3 个小规模实例与 10 个 Liao 经典测试集对 IDSSA 进行仿真实验, 结果表明, 在求解 HFSP 问题时, 相较于其他算法, IDSSA 在求解精度与解的质量上有所提升。本文设计了全新的 DSSA, 并将其与粗糙数据推理理论有机融合, 补充了算法在可解释性方面的不足, 对 SSA 的发展具有一定的理论和实际意义。

1　传统麻雀搜索算法

SSA 是一种受自然界麻雀觅食与躲避天敌启发而提出的群体优化算法。算法使用发现者-加入者模型, 适应度较高的个体称为发现者, 引领种群的搜索方向, 加入者则跟随发现者进行搜索。麻雀种群除了按照发现者与加入者的位置更新公式进行移动外, 占种群个数 10%～20% 的警戒者个体还会进行侦察预警, 这些个体在种群中随机产生, 可以是发现者也可以是加入者。对于 D 维的寻优问题, 麻雀个体表示为: $\boldsymbol{x}_i=[x_{i,1},x_{i,2},\cdots,x_{i,D}]$。

发现者:

$$x_{i,j}^{t+1}=\begin{cases}x_{i,j}^{t}\exp\left(\dfrac{-i}{\alpha T}\right) & r<S_{\mathrm{T}}\\ x_{i,j}^{t}+Q & r\geqslant S_{\mathrm{T}}\end{cases}\tag{1}$$

加入者:

$$x_{i,j}^{t+1}=\begin{cases}Q\exp\left(\dfrac{w_j^t-x_{i,j}^t}{i^2}\right) & i>\dfrac{N}{2}\\ p_j^{t+1}+\dfrac{1}{D}\displaystyle\sum_{j=1}^{D}\mathrm{rand}\{-1,1\}\cdot\left(\left|x_{i,j}^t-p_j^{t+1}\right|\right) & \text{其他}\end{cases}\tag{2}$$

侦察预警行为:

$$x_{i,j}^{t+1}=\begin{cases}b_j^t+\beta\left|x_{i,j}^t-b_j^t\right| & f_i\neq f_{\mathrm{g}}\\ x_{i,j}^t+k\left[\dfrac{\left|x_{i,j}^t-w_j^t\right|}{(f_i-f_{\mathrm{w}})+\varepsilon}\right] & f_i=f_{\mathrm{g}}\end{cases}\tag{3}$$

式中: t 为当前迭代次数; $x_{i,j}$ 为第 i 只个体第 j 维的值; $\alpha\in(0,1]$ 为均匀分布的随机数; T 为最大迭代次数; Q 为正态分布的随机数; $r\in(0,1]$ 和 $S_{\mathrm{T}}\in[0.5,1]$ 分别为预警值和安全值; w_j 为全体种群适应度最差个体第 j 维的值; N 为种群个数; p_j 为发现者群体中适应度最佳个体第 j 维的值; D 为问题的维度; b_j 为全体种群适应度最佳个体第 j 维的值; β 为均值 0、方差 1 的正态分布随机数, 用于控制移动的步长; $k\in[-1,1]$ 为随机数, 既表示麻雀的运动方向, 也控制麻雀运动的步长; f_i、f_{w} 和 f_{g} 分别为第 i 只个体、全体适应度最差个体和适应度最佳个体的适应度; ε 为较小数, 用于防止 f_i 与 f_{w} 相同时分母为 0。

2　麻雀搜索算法离散化

目前, 常见的算法离散化方式是: 利用 Sigmoid 函数将个体值压缩至 [0,1] 范围内, 该类方法实现简单, 但存在无法保留原始算法特征的问题, 且离散后的算法个体值只能为 0 或 1, 在面对解码较为复杂的问题时通常无法解决。SSA 因为其收敛速度快、精度高的特点, 在连续优化问题中有很好的表现, 将其离散化时应保留 SSA 的发现者-加入者模型, 并对原始算法进行合理的抽象, 结合原算法的特点设计离散化的位置更新策略, 进而提出 DSSA。

1) 发现者位置更新策略。在 SSA 中, 发现者的位置更新方式可抽象为: 当预警值小于安全值

时，发现者个体在自身中等范围内移动，反之，则在自身小范围内移动。对于连续优化问题而言，由于解空间是连续的，上述更新策略可以由式 (1) 实现，而离散的解空间中此类解不一定存在，为了实现该策略，根据对个体信息改变的多少来定义中等范围与小范围移动。中等范围移动的操作示意图如图 1 所示。

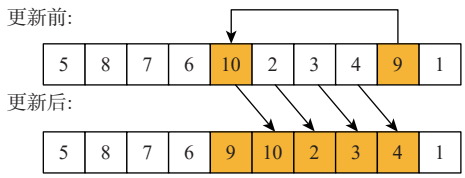


图 1 中等范围移动

Fig. 1 Medium range movement

中等范围移动的具体方式为：随机选择个体某 2 个维度的值，按照选取的顺序进行插入，并将其他维度的值进行后移或前移。中等范围移动操作大量改变了个体对应维度的值，但对个体维度值的顺序改变较小，总体而言对个体信息的改变能力中等。

小范围移动对个体信息的改变较小，具体操作为：选取个体某几个维度上的值进行重新排序，保留原个体与新个体中适应度最好的一个，如适应度相同，则优先在新个体中随机选择一个，操作示意图如图 2 所示。

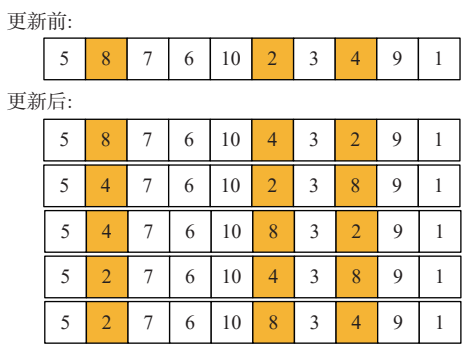


图 2 小范围移动

Fig. 2 Small range movement

2）加入者位置更新策略。在 SSA 中，加入者的更新策略可以抽象为：适应度靠后的 $N/2$ 个个体，在自身与全体种群适应度最差个体之间运动，其余个体会向发现者中适应度最佳个体靠近。同理，需要抽象定义出这 2 种位置更新策略的具体操作。

在自身 $\boldsymbol{x}_i$ 与全体种群适应度最差个体 $\boldsymbol{w}$ 之间运动的具体方式为：首先，在 $\boldsymbol{w}$ 中随机选择几个维度的值，将同维度的值插入新个体 $\boldsymbol{x}_{\text{new1}}$ 中；然后，将 $\boldsymbol{x}_i$ 中的其他值顺序插入到新个体 $\boldsymbol{x}_{\text{new1}}$ 中，交换 $\boldsymbol{x}_i$ 与 $\boldsymbol{w}$ 的先后地位，产生第 2 个新个体 $\boldsymbol{x}_{\text{new2}}$，在 2 个新个体中随机选择一个。操作示意图如图 3 所示。

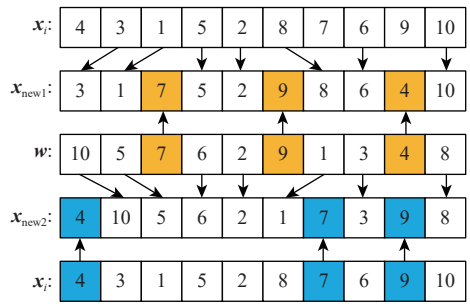


图 3 在最差个体与自身之间移动

Fig. 3 Moving between the worst individual and oneself

靠近发现者中适应度最佳个体的具体操作为：删除 $\boldsymbol{x}_i$ 中与 $\boldsymbol{w}$ 同维度相同的值，产生 $\boldsymbol{x}_i'$，用发现者中适应度最佳个体 $\boldsymbol{p}$ 同维度的值填补 $\boldsymbol{x}_i'$ 的空缺，如果 $\boldsymbol{x}_i'$ 中有重复的值则删去，用 $\boldsymbol{p}$ 的值顺序填补，产生新个体 $\boldsymbol{x}_{\text{new}}$。操作示意图如图 4 所示。

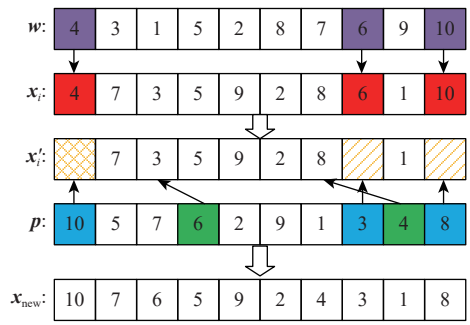


图 4 靠近最优个体

Fig. 4 Approaching the optimal individual

3）侦察预警行为位置更新策略。在传统 SSA 中，侦察预警行为的更新策略可以抽象为：若进行侦察预警行为的个体与全体种群适应度最佳个体的适应度相同，则其进行小范围移动；否则，其将靠近全体种群适应度最佳个体。

3 优化离散麻雀搜索算法

通过抽象 SSA 的位置更新策略提出的 DSSA，最大程度上保留了传统算法收敛速度快、精度高等特点，但传统 SSA 容易陷入局部最优的问题也被保留下来。为保证 DSSA 的种群多样性，提高算法跳出局部最优的能力，做如下改进。

3.1 逆序变异初始化

离散群体智能算法常被用来求解大规模的 NP 难问题，该类问题搜索空间大，对初始种群的要求较高，如果仅使用随机生成个体的方式初始化，种群的质量将无法保证，为花费较少的时间成本扩大搜索空间，也为了后续使用粗糙数据推理时论域更加完整，在初始化过程中使用逆序变异策略。逆序变异初始化种群步骤为：首先，通过随机数生成器生成种群；然后，计算个体适应度，对适应度后

30% 的个体使用逆序变异产生新个体, 对比新个体与原始个体的适应度, 选择适应度较好的个体加入种群。

逆序变异示意图如图 5 所示。

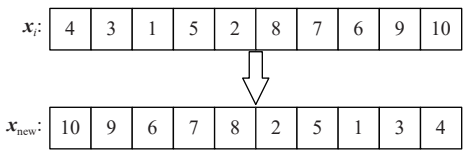


图 5　逆序变异

Fig. 5　Reverse sequence mutation

3.2　融合粗糙数据推理

3.2.1　理论基础

粗糙数据推理理论源于近似理论, 依托于等价关系和近似空间建立, 下面对相关理论与定义进行介绍。

定义 1　设 $\boldsymbol{U}$ 为任意一个数据集, $\boldsymbol{R}$ 为 $\boldsymbol{U}$ 上的等价关系, 令 $\boldsymbol{U}/\boldsymbol{R}=\{[x]_{\boldsymbol{R}}|x\in\boldsymbol{U}\}$, 即 $\boldsymbol{U}/\boldsymbol{R}$ 为 $\boldsymbol{U}$ 中由数据 x 确定的 $\boldsymbol{R}$ 等价类的集合, 称 $\boldsymbol{U}/\boldsymbol{R}$ 为 $\boldsymbol{U}$ 相对于 $\boldsymbol{R}$ 的划分, 其中, $[x]_{\boldsymbol{R}}$ 为 x 确定的 $\boldsymbol{R}$ 等价类。

定义 1 表明 $\boldsymbol{U}$ 上的等价关系 $\boldsymbol{R}$ 和 $\boldsymbol{U}$ 的划分 $\boldsymbol{U}/\boldsymbol{R}$ 之间一一对应, 而建立划分与等价关系的“桥梁”便是等价类, 划分的依据是等价类。

引入上近似的概念, 首先定义近似空间, 设 $\boldsymbol{U}$ 为数据集, $\boldsymbol{R}$ 为 $\boldsymbol{U}$ 上的等价关系, 将 $\boldsymbol{U}$ 和 $\boldsymbol{R}$ 构成的结构记为 $\boldsymbol{M}=(\boldsymbol{U},\boldsymbol{R})$, 称为近似空间。近似空间中 $\boldsymbol{U}$ 称为论域, 由于近似空间中 $\boldsymbol{R}$ 的存在可知, 对于论域 $\boldsymbol{U}$, 根据等价关系进行多少种划分, 就可得到多少个近似空间。

上下近似的定义建立在近似空间上, 设 $\boldsymbol{M}=(\boldsymbol{U},\boldsymbol{R})$ 为近似空间, 则 $\boldsymbol{U}$ 的任意子集 $\boldsymbol{X}$ 的上近似与下近似定义如下:

$$\boldsymbol{R}^{*}(\boldsymbol{X})=\bigcup\{[a]_{\boldsymbol{R}}|[a]_{\boldsymbol{R}}\in\boldsymbol{U}/\boldsymbol{R}\&[a]_{\boldsymbol{R}}\cap\boldsymbol{X}\neq 0\}\tag{4}$$

$$\boldsymbol{R}_{*}(\boldsymbol{X})=\bigcup\{[a]_{\boldsymbol{R}}|[a]_{\boldsymbol{R}}\in\boldsymbol{U}/\boldsymbol{R}\&[a]_{\boldsymbol{R}}\subseteq\boldsymbol{X}\}\tag{5}$$

子集 $\boldsymbol{X}$ 的上近似是所有和 $\boldsymbol{X}$ 的交集不是空集的划分的并, 子集 $\boldsymbol{X}$ 的下近似是所有包含在 $\boldsymbol{X}$ 中的划分的并。下近似从 $\boldsymbol{X}$ 内部逼近 $\boldsymbol{X}$, 上近似从 $\boldsymbol{X}$ 外部逼近, 如果 $\boldsymbol{X}$ 视为精确信息, 则上近似为粗糙信息, 包含与 $\boldsymbol{X}$ 相关的其他信息, 粗糙数据推理的粗糙便体现在此。

3.2.2　粗糙数据推理

在介绍粗糙数据推理时, 需引入粗糙推理空间的概念。粗糙推理空间是对近似空间 $\boldsymbol{M}=(\boldsymbol{U},\boldsymbol{R})$ 的补充, 设 $\boldsymbol{U}$ 为数据集, 称为论域, 其元素称为数据; 令 $\boldsymbol{K}=\{\boldsymbol{R}_1,\boldsymbol{R}_2,\cdots,\boldsymbol{R}_n\}(n\geqslant 1)$, $\boldsymbol{R}_1,\boldsymbol{R}_2,\cdots,\boldsymbol{R}_n$ 为 $\boldsymbol{U}$ 上 n 个不同的等价关系; 给定 $\boldsymbol{U}$ 上的二元关系 $\boldsymbol{S}\subseteq\boldsymbol{U}\times\boldsymbol{U}$, 称 $\boldsymbol{S}$ 为推理关系。将 $\boldsymbol{U}$、$\boldsymbol{K}$ 和 $\boldsymbol{S}$ 构成的结构记作 $\boldsymbol{W}=(\boldsymbol{U},\boldsymbol{K},\boldsymbol{S})$, 称为粗糙推理空间。可以看出, 粗糙推理空间相较于近似空间, 引入明确的推理关系 $\boldsymbol{S}$, 该推理关系一般是明确的事实。例如, DSSA 中某个体向其他个体靠近的操作便在 2 个个体间建立了明确的推理关系。

定义粗糙数据推理时, 设 $\boldsymbol{W}=(\boldsymbol{U},\boldsymbol{K},\boldsymbol{S})$ 为粗糙推理空间, 对于 $<a,b>\in\boldsymbol{S}$ 及 $\boldsymbol{R}\in\boldsymbol{K}$, 存在以下 3 个定义与 2 个引理。

定义 2　设 $b\in\boldsymbol{U}$, 如果 $b\in\boldsymbol{R}^{*}([a-\boldsymbol{R}])$, 则 a 关于 $\boldsymbol{R}$ 直接粗糙推出 b, 记作 $a\Rightarrow_{\boldsymbol{R}}b$。其中, $[a-\boldsymbol{R}]$ 称为后继集, 其定义为: $[a-\boldsymbol{R}]=\{x|x\in\boldsymbol{U}$ 且存在 $z\in[a]_{\boldsymbol{R}}$, 使得 $<z,x>\in\boldsymbol{S}\}$。

定义 3　$b_1,b_2,\cdots,b_n\in\boldsymbol{U}$, 如果 $a\Rightarrow_{\boldsymbol{R}}b_1$, $b_1\Rightarrow_{\boldsymbol{R}}b_2$, $b_2\Rightarrow_{\boldsymbol{R}}b_3,\cdots,b_{n-1}\Rightarrow_{\boldsymbol{R}}b_n(n\geqslant 0)$, 则称 a 关于 $\boldsymbol{R}$ 粗糙推出 b, 记作 $a|=_{\boldsymbol{R}}b$, 容易看出, $a\Rightarrow_{\boldsymbol{R}}b$ 是 $a|=_{\boldsymbol{R}}b$ 的特殊形式。

定义 4　对于 $\boldsymbol{R}\in\boldsymbol{K}$, a 关于 $\boldsymbol{R}$ 直接粗糙推出或粗糙推出 b 的推理称为 $\boldsymbol{W}=(\boldsymbol{U},\boldsymbol{K},\boldsymbol{S})$ 中关于 $\boldsymbol{R}$ 的粗糙数据推理, 简称为粗糙数据推理。

引理 1[16]　设 $\boldsymbol{W}=(\boldsymbol{U},\boldsymbol{K},\boldsymbol{S})$ 为粗糙推理空间, $a,b\in\boldsymbol{U}$ 且 $\boldsymbol{R}\in\boldsymbol{K}$, 则 $a\Rightarrow_{\boldsymbol{R}}b$ 当且仅当 $[b]_{\boldsymbol{R}}\cap[a-\boldsymbol{R}]\neq\varnothing$ 当且仅当 $[a]_{\boldsymbol{R}}\cap[a-\boldsymbol{R}]\neq\varnothing$。

引理 2[16]　设 $\boldsymbol{R}$ 为 $\boldsymbol{U}$ 上的等价关系, 如果 $a,b\in\boldsymbol{U}$, 则 $a\in[b]_{\boldsymbol{R}}$ 当且仅当 $[a]_{\boldsymbol{R}}=[b]_{\boldsymbol{R}}$ 当且仅当 $b\in[a]_{\boldsymbol{R}}$。

以上 3 个定义确定了基于划分的粗糙数据推理, 同理, 可定义出多种划分下的粗糙数据推理, 如 $a|=_{\boldsymbol{R}_1}b$ 与 $a|=_{\boldsymbol{R}_2}b$。在考虑多种划分下的数据推理时, 可引入嵌入算法[16], 该算法的作用为:假设 $a|=_{\boldsymbol{R}_1}b$ 与 $a|=_{\boldsymbol{R}_2}b$ 都成立, 若存在等价关系 $\boldsymbol{R}=\boldsymbol{R}_1\cap\boldsymbol{R}_2$, 嵌入算法可以判断 $a|=_{\boldsymbol{R}}b$ 是否成立。

粗糙数据推理理论能发掘数据间不明确、近似存在的粗糙数据联系, 给出相应的描述方法与一些联系的规则, 使得这些粗糙数据联系有依据的表示, 与传统一阶谓词逻辑不同的是: 粗糙数据推理可以在数据间进行, 保证了其可以与群体智能算法相结合。

3.2.3　改进策略及数学证明

在使用离散群体智能算法求解 NP 难问题时, 如何提高搜索空间大小, 降低算法落入局部最优的可能是设计与优化算法的核心思路。如果将粗糙数据推理建立在算法上, 利用上近似包含粗糙信息的能力, 通过确定的位置更新公式, 推理与演绎出其他可行的位置更新方法, 并且使用该方法进行位置更新, 就可以增加种群移动的选择, 扩大搜索空

间，帮助算法跳出局部最优解，并增大个体收敛到理论最优位置的可能。

DSSA由传统SSA抽象设计而来，从更新策略中可以看出，其存在大量靠近适应度最佳个体的操作，这些操作易导致算法陷入局部最优，且在算法迭代后期种群多样性严重减少，结合粗糙数据推理可以很大程度改善这些问题。粗糙数据推理中的上近似集能引入粗糙信息，扩大算法的搜索空间，进而提高算法跳出局部最优解的能力，在此证明上近似集具有包含粗糙信息的能力。

证明 1 在粗糙推理空间下，对任意的$x\in\boldsymbol{X}$，由等价类的性质可知，$x\in[x]_{\boldsymbol{R}}$，因此，$[x]_{\boldsymbol{R}}\cap\boldsymbol{X}\neq\varnothing$。

通过证明1可得到如下结论。

结论 1 假设$\boldsymbol{R}^*(\boldsymbol{X})$为子集$\boldsymbol{X}$的上近似，由上近似的定义可知，$[x]_{\boldsymbol{R}}\subseteq\boldsymbol{R}^*(\boldsymbol{X})$，则$x\in\boldsymbol{R}^*(\boldsymbol{X})$，因此，$\boldsymbol{X}\subseteq\boldsymbol{R}^*(\boldsymbol{X})$。

由结论1可知，集合的上近似集由外部逼近该集合，包含更多信息，在DSSA中加入者与预警机制的离散更新策略包含靠近局部最优的操作，降低了种群多样性，使用基于上近似的粗糙数据推理理论，可以包含更多模糊信息，扩大搜索空间，是粗糙数据推理理论能帮助DSSA跳出局部最优解的有力证明。

对DSSA与粗糙数据推理理论结合的合理性做出如下证明。首先，证明粗糙数据推理理论能保留算法确定的收敛操作信息。

证明 2 在粗糙推理空间下，对于任意由划分确立的等价关系$\boldsymbol{R}\in\boldsymbol{K}$，以及算法中的个体$a,b\in\boldsymbol{U}$。由等价类的定义可知$a\in[a]_{\boldsymbol{R}}$。若存在明确的收敛信息$<a,b>\in\boldsymbol{S}$，由后继集的定义可知，$b\in[a-\boldsymbol{R}]$。又因为$[a-\boldsymbol{R}]\subseteq\boldsymbol{R}^*([a-\boldsymbol{R}])$，所以$b\in\boldsymbol{R}^*([a-\boldsymbol{R}])$，因此可得$a\Rightarrow_{\boldsymbol{R}}b$或$a|=_{\boldsymbol{R}}b$。

通过证明2可得到如下结论。

结论 2 对于算法个体$a,b\in\boldsymbol{U}$，若存在确定的收敛信息$<a,b>\in\boldsymbol{S}$，则对任意基于划分的等价关系$\boldsymbol{R}\in\boldsymbol{K}$，有$a\Rightarrow_{\boldsymbol{R}}b$或$a|=_{\boldsymbol{R}}b$。

该结论阐明了粗糙数据推理能保证算法确定的收敛信息不丢失，是粗糙数据推理理论可与算法相结合的先决条件。

其次，证明粗糙数据推理理论能引入模糊信息，扩大算法的搜索空间。

证明 3 在粗糙推理空间下，对于任意由划分确立的等价关系$\boldsymbol{R}\in\boldsymbol{K}$，以及算法中的个体$a,b,u,v\in\boldsymbol{U}$。若存在确定的$<a,b>\in\boldsymbol{S}$，则$a\Rightarrow_{\boldsymbol{R}}b$，故$[b]_{\boldsymbol{R}}\cap[a-\boldsymbol{R}]\neq\varnothing$。当$u\in[b]_{\boldsymbol{R}}$时，由于$[u]_{\boldsymbol{R}}\in[b]_{\boldsymbol{R}}$，$[u]_{\boldsymbol{R}}\cap[a-\boldsymbol{R}]\neq\varnothing$，可得$a\Rightarrow_{\boldsymbol{R}}u$。同理，当$v\in[a]_{\boldsymbol{R}}$时，易知$v\Rightarrow_{\boldsymbol{R}}b$。

通过证明3可得到如下结论。

结论 3 设$\boldsymbol{W}=(\boldsymbol{U},\boldsymbol{K},\boldsymbol{S})$为粗糙推理空间，若$\boldsymbol{R}\in\boldsymbol{K}$且$a,b,u,v\in\boldsymbol{U}$，则有：①若$a\Rightarrow_{\boldsymbol{R}}b$，则当$u\in[b]_{\boldsymbol{R}}$时，有$a\Rightarrow_{\boldsymbol{R}}u$；②若$a\Rightarrow_{\boldsymbol{R}}b$，则当$v\in[a]_{\boldsymbol{R}}$时，有$v\Rightarrow_{\boldsymbol{R}}b$。

该结论表明，粗糙数据推理挖掘出了模糊信息，通过已有确定的收敛信息，合理地将具有粗糙联系的u、v引入，扩大算法的搜索空间，帮助算法跳出局部最优解，是粗糙数据推理能实现预想优化目标的有力数学证明。

证明2与证明3从理论上证明了粗糙数据推理理论能与DSSA有机结合，且能完成引入该理论的预期目标，同时也一定程度上解决了群体智能算法可解释性差的问题。

von Lücken等[17]将聚类的思想与进化算法相结合，并通过实验证明了对算法种群进行合理划分能有效提高算法的寻优能力。定义1阐述了等价关系与划分的一一对应关系，粗糙数据推理在等价类中建立也可称为划分之间建立，为将粗糙数据推理理论与算法融合，要先对算法的种群进行划分，采取3种划分方式。首先，根据适应度划分，DSSA的发现者-加入者模型本身可以看作依照适应度划分，与划分对应的等价关系记为$\boldsymbol{R}_1$。在此划分下，当算法运行中发生了加入者个体$\boldsymbol{x}_i$向适应度最佳个体$\boldsymbol{p}$靠近的情况，即出现了确定的收敛信息$<\boldsymbol{x}_i,\boldsymbol{p}>\in\boldsymbol{S}$，由结论2可知$\boldsymbol{x}_i\Rightarrow_{\boldsymbol{R}_1}\boldsymbol{p}$，再由结论3可得如图6(a)所示的推理关系。其次，根据汉明距离划分，以个体总数的10%为数量，分别取与$\boldsymbol{x}_i$和$\boldsymbol{p}$汉明距离最近的个体，且保证所取个体互不相同，对应的等价关系记为$\boldsymbol{R}_2$。同样的，可以得到如图6(b)所示的推理关系。

最后，引入一种结合汉明距离和适应度的划分，建立新的推理关系，由如下函数推出：

$$y(\boldsymbol{x}_i,\boldsymbol{x}_j)=\frac{H(\boldsymbol{x}_i,\boldsymbol{x}_j)}{\left|f(\boldsymbol{x}_i)-f(\boldsymbol{x}_j)\right|+\varepsilon}\tag{6}$$

式中：$H(\boldsymbol{x}_i,\boldsymbol{x}_j)$为个体$\boldsymbol{x}_i$和$\boldsymbol{x}_j$的汉明距离；$\left|f(\boldsymbol{x}_i)-f(\boldsymbol{x}_j)\right|$为个体$\boldsymbol{x}_i$和$\boldsymbol{x}_j$适应度差的绝对值；$\varepsilon$为很小的值，保证分母不为0。该函数同时考虑了个体差异与适应度差异，以该函数取较大值为标准，按照与汉明距

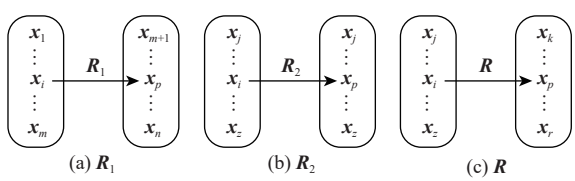


图6 推理关系

Fig. 6 Reasoning relationship

离划分同样的步骤进行划分, 其对应的等价关系记为 $\boldsymbol{R}$, 显然有 $\boldsymbol{R}=\boldsymbol{R}_1\cap\boldsymbol{R}_2$。同理可得如图 6(c) 所示的推理关系。

综上, 已在算法中建立了 3 种推理关系, 根据推理关系可以推导出以下 3 个更新策略, 用来提高算法的搜索空间, 降低算法陷入局部最优的可能, 具体操作为: 当个体 $\boldsymbol{x}_i$ 向 $\boldsymbol{p}$ 靠近, 即 $<\boldsymbol{x}_i,\boldsymbol{p}>\in\boldsymbol{S}$ 成立时, 使 $\boldsymbol{x}_i$ 在一组参数的控制下采用以下 3 种策略更新位置:

1) 当个体准备向 $\boldsymbol{p}$ 靠近时, 在 $\boldsymbol{p}$ 的 $\boldsymbol{R}_2$ 等价类中随机选择一个个体靠近。

2) 当个体向 $\boldsymbol{p}$ 靠近的同时, 在 $\boldsymbol{p}$ 的 $\boldsymbol{R}_1$ 等价类中随机选择一个个体靠近, 并且依据适应度贪心选择个体。

3) 当个体向 $\boldsymbol{p}$ 靠近的同时, 在 $\boldsymbol{p}$ 的 $\boldsymbol{R}$ 等价类中随机选择一个个体靠近, 并且依据适应度贪心选择个体。

3 种策略受迭代次数与一组参数的控制。迭代初期主要使用策略 1, 保证种群的多样性, 扩大搜索空间; 迭代中期主要使用策略 2, 提高收敛速度; 迭代后期主要使用策略 3, 提高收敛精度。3 种策略在算法迭代的全过程中都会进行, 为算法提供了更多位置更新策略, 能有效降低算法陷入局部最优的可能, 帮助算法收敛至一个更好的解。

3.3　算法步骤

步骤 1　初始化种群规模 N、最大迭代次数 T、当前迭代次数 t、发现者比例 P_{num}、警戒者比例 S_{num}、预警值 r 和安全值 S_{T} 等参数。

步骤 2　利用 3.1 节逆序变异初始化策略初始化种群。

步骤 3　根据所求问题计算每个麻雀个体的适应度 f_i, 并进行排序, 记录适应度最优个体 $\boldsymbol{g}$ 与最差个体 $\boldsymbol{w}$ 的位置信息及适应度 f_{g} 与 f_{w}。

步骤 4　依照发现者比例 P_{num} 与适应度将种群划分为发现者与加入者, 依照汉明距离划分种群, 根据式 (6) 划分种群, 在不同的划分下建立粗糙数据推理。

步骤 5　根据 DSSA 与粗糙数据推理推导出的更新策略更新发现者和加入者的位置, 依照警戒者比例 S_{num} 进行侦察预警行为。

步骤 6　若达到最大迭代次数, 输出最优解, 算法结束; 否则, 返回步骤 2。

3.4　时间复杂度分析

时间复杂度是评价算法执行效率的重要指标。假设所求解问题维度为 D, 麻雀的种群数量为 N, 最大迭代次数为 T, 求解问题适应度的时间为 $f(D)$, 初始化参数耗时为 t_0。

在 IDSSA 中, 进行逆序变异初始化需要利用随机数生成器生成种群, 计算种群适应度并进行排序, 选取适应度靠后 30% 的个体进行逆序变异, 贪心选择适应度较高个体。设随机生成每个个体耗时为 t_1, 对适应度排序耗时 t_2, 逆序变异是对个体维度信息的反转, 耗时为 t_3D, 依照适应度贪心选择个体耗时 $f(D)+t_4$, 则逆序变异初始化的时间复杂度为

$$T_1=O(N(t_1+f(D))+t_2+0.3N(t_3D+f(D)+t_4))=O(f(D)+D) \tag{7}$$

发现者数量为 $P_{\text{mum}}\times N$, 发现者更新模式为: 当 $r<S_{\text{T}}$ 时, 在中等范围内移动, 设中等范围移动耗时为 t_5D; $r\geqslant S_{\text{T}}$ 时, 在小范围内移动, 小范围移动需要交换维度值产生新个体, 设耗时为 t_6D, 并依照适应度进行贪心选择, 耗时为 $Cf(D)+t_7$, C 为产生新个体数量。则发现者位置更新策略的时间复杂度为

$$T_2=O(S_{\text{T}}\times P_{\text{mum}}\times N(t_5D)+(1-S_{\text{T}})\times P_{\text{mum}}\times N(t_6D+Cf(D)+t_7))=O(f(D)+D) \tag{8}$$

加入者数量为 $(1-P_{\text{mum}})\times N$, 加入者中适应度靠后的 $N/2$ 个个体在自身与当前适应度最差个体间运动, 该操作会生成 2 个个体并随机选择, 设耗时为 t_8D+t_9; 其余加入者会靠近位置最优的发现者, 设耗时为 $t_{10}D$。则加入者位置更新策略的时间复杂度为

$$T_3=O\left(\frac{N}{2}(t_8D+t_9)+\left(1-P_{\text{mum}}-\frac{1}{2}\right)N(t_{10}D)\right)=O(D) \tag{9}$$

侦察预警行为包括小范围移动与靠近全局最优的操作, 可得时间复杂度为

$$T_4=O(f(D)+D) \tag{10}$$

策略 1 需要对算法依照汉明距离进行划分, 设计算汉明距离耗时为 $t_{11}D$, 需计算 $N-1$ 次, 对汉明距离排序耗时为 t_{12}, 选择靠近个体耗时为 t_{13}, 靠近操作与离散位置更新策略中定义操作耗时相同, 耗时为 $t_{14}D$。因此, 策略 1 的时间复杂度为

$$T_5=O((N-1)t_{11}D+t_{12}+t_{13}+t_{14}D)=O(D) \tag{11}$$

由于策略 2 与策略 3 的划分及贪心选择需要计算适应度, 其余操作与策略 1 类似, 策略 2 与策略 3 的时间复杂度 T_6、T_7 都为 $O(f(D)+D)$。

故 IDSSA 的时间复杂度为

$$T_{\text{O}}=t_0+T(T_1+T_2+T_3+T_4+T_5+T_6+T_7)=O(f(D)+D) \tag{12}$$

文献 [3] 指出, 传统 SSA 的时间复杂度为 $O(f(D)+D)$,

与IDSSA时间复杂度相同，综上可知，本文算法IDSSA及对其改进策略并未提高算法的时间复杂度。

4 HFSP问题

4.1 问题描述

HSFP是经典的组合优化问题，已被证明是NP难问题。通常情况下，HSF问题被描述为：有J个工件需依次经过M道工序的连续加工，工序j有$o_j(o_j \geqslant 1; j=1,2,\cdots,M)$台机器，且至少有一道工序包含多台机器，工件可在当前工序的任意一台机器上加工。图7为HFSP问题的示意图。

为对问题的范围进行限定，有以下假设：

1）一台机器同一时刻只能加工一个工件，且一个工件同一时刻只能在一台机器上加工。

2）所有机器都可使用，不考虑机器损坏及损耗。

3）所有工件机器独立加工，且加工开始后不可中断。

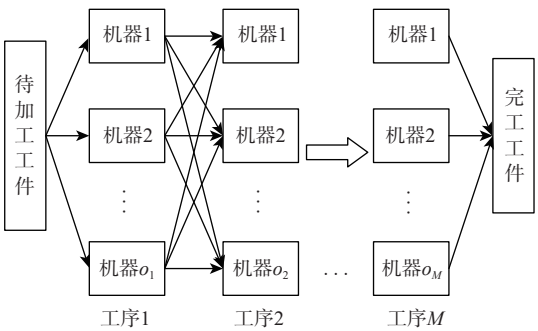


图7 HFSP问题示意图

Fig. 7 Schematic diagram of HFSP

4.2 数学模型

HFSP问题的数学模型以最小化所有工件的最大完成时间为优化目标：

$$\min C_{\max} = \min\{\max\{c_1, c_1, \cdots, c_n\}\} \tag{13}$$

具体约束如下：

$$0 \leqslant S_{i,j} < E_{i,j} \tag{14}$$

$$E_{i,j-1} \leqslant S_{i,j} \tag{15}$$

$$C_i = E_{i,M} \tag{16}$$

$$\sum_{k=1}^{o_j} X_{i,j,k} = 1 \tag{17}$$

$$T_{i,j} = \sum_{k=1}^{o_j} X_{i,j,k} F_{i,j,k} \tag{18}$$

$$E_{i,j} = S_{i,j} + T_{i,j} \tag{19}$$

式(13)表示优化目标，即最小化最大完工时间$C_{\max}$。式(14)为时间约束，工件在某一工序的开始时间$S_{i,j}$不能为负，且结束时间$E_{i,j}$一定大于开始时间。式(15)为工序约束，工件必须在前一工序加工结束后才能进行下一工序的加工。式(16)说明工件的完工时间C_i等于最后一道工序的加工结束时间。式(17)为机器约束，说明工件在每个阶段只能由一台机器加工。式(18)为工件某一阶段实际加工耗时的计算方法，$F_{i,j,k}$为工件i在第j道工序的机器k上加工所需时间。式(19)为工件在某工序的结束加工时刻计算方法，$T_{i,j}$为工件i在第j道工序的实际加工耗时。

4.3 编码与解码

群体智能算法求解HFSP问题的常见编码方式有：机器选择和工件排序(machine selection & operation sequence, MSOS)编码、工序排列编码。MSOS编码是对工件的机器选择与工序选择都进行编码，且在更新个体时同时更新这2部分，往往会导致算法的随机性过强且运算代价增大。工序排列编码是以工件的制造序号作为一个个体，且在第一工序时按照个体信息进行加工，依次选择并行机。例如，有10个工件时，假设个体为$\boldsymbol{x}_i$=[6,9,7,4,8,2,5,1,3,10]，则在第一工序时首先加工工件6，其次加工工件9，依次加工直到工件10。工序排列实现简单且稳定性高，可以很大程度上保证算法在迭代过程中稳定接近最优解。解码方式反映了算法的调度策略，为增强算法求解精度，采取如下调度策略：

1）先完工的工件优先进入下一工序进行加工。

2）单工件等待且单机空闲时，加工该工件。

3）多工件等待且单机空闲时，选择在该并行机上加工耗时最短的工件加工。

4）单工件等待且多并行机空闲时，选择在耗时最少的并行机上加工。

5）多工件等待且多并行机空闲时，选择总耗时最少的调度方法。

5 改进离散麻雀搜索算法求解HFSP问题

为验证IDSSA的性能及优越性，本文选取了3个小规模实例[18]与10个Liao经典测试集[19]，将IDSSA用于求解HFSP问题。实例1：某工厂需加工6个工件，工序分别为铣、车和磨，共有3台铣床、2台车床和3台磨床。实例2：某个钢铁生产企业的生产过程由炼钢、精炼、连钢和轧制4个步骤组成，企业共有3台炼钢炉、3台精炼炉、2台铸机和2台轧机。实例3：某工厂需要加工12个工件，工序分别为车、刨和磨，共有3台车床、2台刨床和4台磨床。10个Liao经典测试集[19](j30c5e1～j30c5e10)有30个工件与5道工序，j30代表工件数

为 30, c5 代表工序数为 5, e1～e10 为测试集编号。IDSSA 种群规模 N 为 150, 发现者比例 P_{num} 为 0.2, 警戒者比例 S_{num} 为 0.2, 安全值 S_T 为 0.8。

5.1　参数分析

融合粗糙数据推理对 IDSSA 的改进中, 形式化推导出了 3 种策略, 这 3 种策略的执行概率受到迭代次数的影响。为保证算法的寻优精度与速度, 通过正交实验来确定 3 种策略的执行概率。依照当前迭代次数与最大迭代次数的比值将算法划分为前、中、后期, 对应关系为前期 [0, 0.15]、中期 (0.15, 0.8]、后期 (0.8, 1]。最大迭代次数 T 为 300, 进行 10 次对比实验, 记录每次实验的最小完工时间, 考察其平均值。小规模实例 1 的正交实验如表 1 所示。

表 1　小规模实例 1 结果

Table 1　Results of small-scale example 1

组合	前期执行概率			中期执行概率			后期执行概率			最小完工时间平均值/s
	策略1	策略2	策略3	策略1	策略2	策略3	策略1	策略2	策略3	
1	0.3	0.4	0.3	0.2	0.7	0.1	0.1	0.2	0.7	13.5
2	0.3	0.4	0.3	0.2	0.7	0.1	0.2	0.2	0.6	13.7
3	0.3	0.4	0.3	0.3	0.6	0.1	0.1	0.2	0.7	13.6
4	0.3	0.4	0.3	0.3	0.6	0.1	0.2	0.2	0.6	13.7
5	0.4	0.3	0.3	0.2	0.7	0.1	0.1	0.2	0.7	13.7
6	0.4	0.3	0.3	0.2	0.7	0.1	0.2	0.2	0.6	13.7
7	0.4	0.3	0.3	0.3	0.6	0.1	0.1	0.2	0.7	13.5
8	0.4	0.3	0.3	0.3	0.6	0.1	0.2	0.2	0.6	13.7

观察表 1 可得, 小规模实例 1 更适合组合 1 所示执行概率, 剩余测试集的实验结果与概率组合选取结果如表 2 所示。

表 2　正交实验结果

Table 2　Orthogonal experimental results

算例	组合	最小完工时间平均值/s
小规模实例2	3	297
小规模实例3	1	24
j30c5e1	1	477.2
j30c5e2	6	619.8
j30c5e3	3	613.3
j30c5e4	2	577.6
j30c5e5	1	606.8
j30c5e6	6	613.7
j30c5e7	5	633.4
j30c5e8	3	691.7
j30c5e9	1	655.1
j30c5e10	1	599.3

5.2　小规模实验结果与分析

将 IDSSA 与遗传算法 (genetic algorithm, GA)、差分进化算法 (differential evolution algorithm, DEA)、果蝇优化算法 (fruit-fly optimization algorithm, FOA)、DSSA 进行 10 次对比实验, 记录加工时间的最小值与平均值。限制 IDSSA 与 DSSA 的迭代次数为 500 次, 其余算法不限制迭代次数。实验结果如表 3 所示。

表 3　小规模实例结果比较

Table 3　Comparison of results of small-scale examples

算法	加工时间最小值/s			加工时间平均值/s		
	实例1	实例2	实例3	实例1	实例2	实例3
GA	14.0		26	14.9		27.3
DEA		299	25		309.6	25.3
FOA	13.5	297		13.5	297.4	
DSSA	14	297	24	14.2	297.6	24.2
IDSSA	13.5	297	24	13.5	297	24

小规模实验表明, IDSSA 在 3 个小规模实例中都能收敛到理论最优解, 平均值与最小值相同, 即标准差为 0, 可见收敛至理论最优的情况非常稳定。对于实例 2 而言, IDSSA 与 FOA 虽然都收敛至理论最优的 297 s, 但 IDSSA 仅仅迭代 50 次以下就收敛, 而 FOA 迭代 18 000 次左右才收敛[20]。同时可以得出, 对 DSSA 改进后, 算法的寻优能力与稳定性都有提升。

5.3　大规模实验结果与分析

大规模算例为 10 个 Liao 算例[19]。将本文提出的 IDSSA 与 GA、DEA、生物地理学优化算法 (biogeography-based optimization algorithm, BBOA)、改进生物地理学优化算法 (improved biogeography-based optimization algorithm, IBBOA) 及 DSSA 进行 10 次对比实验, 记录加工时间的最小值和平均值, 实验结果如表 4 所示, 算法的迭代次数为 500 次。j30c5e1 算例的最优调度甘特图如图 8 所示, 图 8 右

表 4 大规模测试集结果比较

Table 4 Comparison of large-scale test set results

s

算法	加工时间最小值									
	j30c5e1	j30c5e2	j30c5e3	j30c5e4	j30c5e5	j30c5e6	j30c5e7	j30c5e8	j30c5e9	j30c5e10
GA	482.0	619.0	628.0	588.0	619.0	627	638	700.0	669	611
DEA	481.0	620.0	620.0	586.0	621.0	625	637	699.0	671	610
BBOA	484.0	621.0	623.0	585.0	618.0	628	639	698.0	670	612
IBBOA	474.0	616.0	610.0	577.0	609.0	615	629	685.0	654	596
DSSA	474.0	619.0	612.0	577.0	607.0	616	630	691.0	656	598
IDSSA	472.0	616.0	609.0	575.0	604.0	608	629	685.0	652	594

算法	加工时间平均值									
	j30c5e1	j30c5e2	j30c5e3	j30c5e4	j30c5e5	j30c5e6	j30c5e7	j30c5e8	j30c5e9	j30c5e10
GA	495.0	627.0	635.2	599.2	627.6	636.2	644.8	712.3	677.0	622.5
DEA	490.5	624.2	631.8	597.5	627.6	635.5	643.0	709.7	680.3	619.6
BBOA	491.8	625.2	636.4	594.4	625.9	640.1	643.7	710.2	681.4	621.9
IBBOA	479.2	617.6	614.1	582.0	616.9	620.4	635.5	693.1	662.5	605.4
DSSA	479.8	622.8	616.9	583.7	615.1	618.8	637.2	697.4	660.2	606.5
IDSSA	477.2	619.8	613.3	577.6	606.8	613.7	633.4	691.7	655.1	599.3

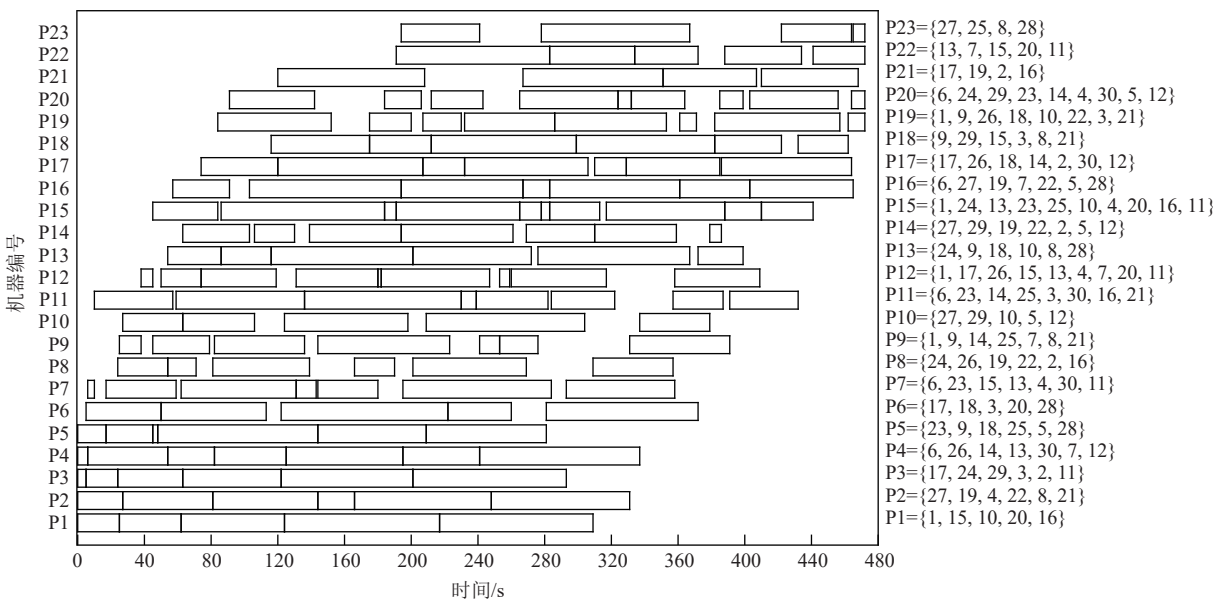


图 8 j30c5e1 算例的最优调度甘特图

Fig. 8 Optimal scheduling Gantt chart of j30c5e1

侧为该并行机上加工工件的序列。

观察表 4 可知，在求解 j30c5e2 算例时，IDSSA 在最小值相同的情况下，平均值高于 IBBOA，即在 10 次对比实验中整体效果略差于 IBBOA，除此以外，在其他算例的求解中与其他算法比较，IDSSA 在寻优精度上存在优势，在 j30c5e8 算例中，对比寻优精度相同的 IBBOA，IDSSA 所求解的整体质量更优。对比 DSSA，算法改进的优势在大规模实验中得到充分体现，说明在复杂的寻优情况下，结合粗糙数据推理引入的跳出局部最优的策略能指导算法收敛至全局最优。综上所述，对 SSA 的离散化是可行的，且提出的改进策略也是合理的，能有效提高 DSSA 的寻优能力。

6 结 论

本文提出了一种 IDSSA，针对 SSA 无法处理离散优化问题的不足，结合 SSA 的特点将 SSA 离散化，为提高 DSSA 跳出局部最优的可能，扩大算法的搜索空间，做出如下改进：

1）结合逆序变异策略初始化种群，提高初始种群质量。

2）结合粗糙数据推理，合理的扩充搜索策略，

提高迭代中的种群多样性, 扩展搜索空间, 改善算法易于陷入局部最优的缺陷, 提高算法的收敛速度与求解精度。

通过 HFSP 问题的 3 个小规模实例与 10 个 Liao 经典算例的测试, 与多种算法进行比较, 验证了算法的优越性与 2 种改进策略的有效性。

DSSA 对其他群体智能算法的离散化有一定的指导意义, 之后可以从以下方面做进一步研究: ①将 DSSA 应用于带其他约束的 HFSP 问题中, 提出更具针对性的高性能算法。②构建更符合实际情况的模型, 以 IDSSA 轻量化的特点, 为企业提供低成本的调度策略。

参考文献（References）

[1] XUE J K, SHEN B. A novel swarm intelligence optimization approach: Sparrow search algorithm[J]. Systems Science & Control Engineering, 2020, 8(1): 22-34.

[2] 吴新忠, 韩正化, 魏连江, 等. 矿井风流智能按需调控算法与关键技术[J]. 中国矿业大学学报, 2021, 50(4): 725-734.
WU X Z, HAN Z H, WEI L J, et al. Intelligent on-demand adjustment algorithm and key technology of mine air flow[J]. Journal of China University of Mining & Technology, 2021, 50(4): 725-734(in Chinese).

[3] 付华, 刘昊. 多策略融合的改进麻雀搜索算法及其应用[J]. 控制与决策, 2022, 37(1): 87-96.
FU H, LIU H. Improved sparrow search algorithm with multi-strategy integration and its application[J]. Control and Decision, 2022, 37(1): 87-96(in Chinese).

[4] 韩统, 汤安迪, 周欢, 等. 基于 LASSA 算法的多无人机协同航迹规划方法[J]. 系统工程与电子技术, 2022, 44(1): 233-241.
HAN T, TANG A D, ZHOU H, et al. Multiple UAV cooperative path planning based on LASSA method[J]. Systems Engineering and Electronics, 2022, 44(1): 233-241(in Chinese).

[5] 熊智, 李欣童, 熊骏, 等. 基于改进麻雀搜索算法的无人机集群置信传播协同定位方法[J]. 中国惯性技术学报, 2021, 29(2): 171-177.
XIONG Z, LI X T, XIONG J, et al. Cooperative positioning method for UAV swarm belief propagation based on improved sparrow search algorithm[J]. Journal of Chinese Inertial Technology, 2021, 29(2): 171-177(in Chinese).

[6] 熊国文, 张敏, 许文鑫. 基于众包模式的两级开闭混合车辆路径优化[J]. 浙江大学学报 (工学版), 2021, 55(12): 2397-2408.
XIONG G W, ZHANG M, XU W X. Vehicle routing optimization of two-echelon opening and closing hybrid based on crowd sourcing mode[J]. Journal of Zhejiang University (Engineering Science), 2021, 55(12): 2397-2408(in Chinese).

[7] WU X L, CAO Z. Greedy simulated annealing algorithm for solving hybrid flow shop scheduling problem with re-entrant batch processing machine[C]//Proceedings of the Chinese Automation Congress. Piscataway: IEEE Press, 2020: 4907-4911.

[8] LIANG X, WANG P, HUANG M. Flowshop scheduling problem with limited buffer based on hybrid shuffled frog leaping algorithm[C]//Proceedings of the IEEE 7th International Conference on Computer Science and Network Technology. Piscataway: IEEE Press, 2019: 87-93.

[9] ZHANG Q, TIAN Z, WANG S, et al. Iterated greedy algorithm for solving a hybrid flow shop scheduling problem with reentrant jobs[C]//Proceedings of the Chinese Control and Decision Conference. Piscataway: IEEE Press, 2020: 5636-5641.

[10] SUN X, SHEN W, SUN B. A modified genetic algorithm for distributed hybrid flowshop scheduling problem[C]//Proceedings of the IEEE 24th International Conference on Computer Supported Cooperative Work in Design. Piscataway: IEEE Press, 2021: 981-986.

[11] DONG X. An self-adaptive flower pollination algorithm for hybrid flowshop scheduling problem[C]//Proceedings of the IEEE 3rd Information Technology, Networking, Electronic and Automation Control Conference. Piscataway: IEEE Press, 2019: 2176-2179.

[12] WANG P, SANG H Y, TAO Q Y, et al. Improved migratory birds optimisation algorithm to solve low-carbon hybrid lot-streaming flowshop scheduling problem[C]//Proceedings of the World Conference on Computing and Communication Technologies. Piscataway: IEEE Press, 2020: 33-36.

[13] SCHUMACHER C, BUCHHOLZ P, FIEDLER K, et al. Local search and Tabu search algorithms for machine scheduling of a hybrid flow shop under uncertainty[C]//Proceedings of the Winter Simulation Conference. Piscataway: IEEE Press, 2020: 1456-1467.

[14] GAREY M R, JOHNSON D S, SETHI R. The complexity of flowshop and jobshop scheduling[J]. Mathematics of Operations Research, 1976, 1(2): 330-348.

[15] JOYCE T, HERRMANN J M. A review of no free lunch theorems, and their implications for metaheuristic optimization[M]//YANG X S. Nature-inspired algorithms and applied optimization. Berlin: Springer, 2018: 27-51.

[16] 闫硕. 基于上近似的粗糙数据推理研究及应用[D]. 北京: 北京交通大学, 2017: 32-45.
YAN S. Rough data-deduction based on the upper approximation and its applications[D]. Beijing: Beijing Jiaotong University, 2017: 32-45(in Chinese).

[17] VON LÜCKEN C, BRIZUELA C A, BARÁN B. Clustering-based multipopulation approaches in MOEA/D for many-objective problems[J]. Computational Optimization and Applications, 2022, 81(3): 789-828.

[18] 杜利珍, 王震, 柯善富, 等. 混合流水车间调度问题的果蝇优化算法求解[J]. 中国机械工程, 2019, 30(12): 1480-1485.
DU L Z, WANG Z, KE S F, et al. Fruit fly optimization algorithm for solving hybrid flow-shop scheduling problems[J]. China Mechanical Engineering, 2019, 30(12): 1480-1485(in Chinese).

[19] LIAO C J, TJANDRADJAJA E, CHUNG T P. An approach using particle swarm optimization and bottleneck heuristic to solve hybrid flow shop scheduling problem[J]. Applied Soft Computing, 2012, 12(6): 1755-1764.

[20] 张源, 陶翼飞, 王加冕. 改进差分进化算法求解混合流水车间调度问题[J]. 中国机械工程, 2021, 32(6): 714-720.
ZHANG Y, TAO Y F, WANG J M. An improved DE algorithm for solving hybrid flow-shop scheduling problems[J]. China Mechanical Engineering, 2021, 32(6): 714-720(in Chinese).

Discrete sparrow search algorithm incorporating rough data-deduction for solving hybrid flow-shop scheduling problems

ZHOU Ning*，ZHANG Songlin，ZHANG Chen

(School of Electronic and Information Engineering，Lanzhou Jiaotong University，Lanzhou 730070，China)

Abstract：To address the shortcomings of the sparrow search algorithm (SSA), such as easy fall into local optimum and inability to solve discrete optimization problems, an improved discrete sparrow search algorithm (IDSSA) is proposed. Firstly, the position update formula of the original sparrow search algorithm is abstracted, with a new discrete heuristic position update strategy designed according to the different identities of individuals, and with the encoding and decoding methods designed for the hybrid flow-shop scheduling problem (HFSP). Secondly, the rough data-deduction theory is introduced, and the feasibility and rationality of the above theory are explained by mathematical proofs, providing theoretical support for the algorithm and improving the interpretability. Then, the nature of upper approximation is adopted to expand the search space, improve the population diversity, and avoid prematurity of the algorithm. Division and rough data-deduction are combined to propose three strategies to promote information sharing among populations, regulate the exploitation ability and exploration ability of populations, and reduce the probability of the algorithm falling into local optimum. Finally, the improved discrete sparrow search algorithm is used to solve the hybrid flow shop scheduling problem. Simulation experiments are carried out on three small-scale practical examples and ten Liao's classic test sets to verify the feasibility of the improved algorithm. Results show the superiority of the proposed algorithm and the effectiveness of the improved strategy through comparison with classical algorithms such as genetic algorithm and differential evolutionary algorithm.

Keywords：sparrow search algorithm；discretization algorithm；rough set theory；global optimization；approximation algorithm；data association；hybrid flow-shop scheduling

Received：2022-05-27；**Accepted**：2022-08-01；**Published Online**：2022-09-06 10：16

URL：link.cnki.net/urlid/11.2625.V.20220905.1326.002

Foundation items：National Natural Science Foundation of China (61650207,61963023); Tianyou Innovation Team of Lanzhou Jiaotong University (TY202003)

* **Corresponding author**. E-mail：zhouning@lzjtu.edu.cn

http://bhxb.buaa.edu.cn jbuaa@buaa.edu.cn
DOI: 10.13700/j.bh.1001-5965.2022.0443

面向语法加权图文本的方面情感三元组抽取

韩虎*，孟甜甜
(兰州交通大学 电子与信息工程学院，兰州 730070)

摘 要：方面情感三元组抽取包括方面抽取、意见抽取和方面情感分类3项任务，以管道方式解决该任务的研究方法无法利用元素之间的交互信息，同时也会造成错误传播和冗余训练。基于此，提出一种基于门控注意力和加权图文本的方面情感三元组抽取方法。采用双向长短时记忆网络学习句子的序列特征表示；利用门控注意力单元学习单词之间的线性联系；利用语法距离加权图卷积网络增强三元组元素之间的交互；利用网格标记推理策略预测三元组。在4个公开数据集上进行实验，结果表明：所提方法可以有效增强三元组元素之间的交互，提高三元组抽取的准确率；同时，所提方法的 F_1 值分别为57.94%、70.54%、61.95%和67.66%，与基准模型相比均有所提高。

关 键 词：三元组抽取；门控注意力；加权图文本；双向长短时记忆网络；网格标记
中图分类号：TP391
文献标志码：A **文章编号**：1001-5965(2024)02-0409-10

方面情感分析 (aspect-based sentiment analysis, ABSA) 是对评论语句中特定评价方面给出情感极性的研究[1]，近年来引起较多研究人员的广泛关注。方面情感分类 (aspect sentiment classification, ASC) 是ABSA最深入研究的任务之一，旨在预测给定方面的情感极性[2]。方面术语抽取 (aspect term extraction, ATE) 致力于提取方面词[3]。意见术语抽取 (opinion term extraction, OTE) 旨在提取意见词，可以在很大程度上决定评论语句或相应方面术语的情感极性[4]。方面情感三元组抽取 (aspect sentiment triplet extraction, ASTE) 是ABSA近期提出的子任务，其通过三元组 (方面目标词、相应的意见词及表达的情感极性) 形成情感信息更完整的画面[5]。例如，在句子“Foods are delicious but the price is expensive”中，包含2个三元组：(Foods,delicious,POS) 和 (price,expensive,NEG)。POS、NEG和NEU分别代表积极、消极和中立。区别于ABSA任务提取的二元组 (Foods,POS) 和 (price,NEG)，ASTE任务能更好地反映用户评论中的多重情感因素 (方面，意见，情感)，也更适合于评论文本语义的建模。

近年来，Peng等[6]利用两阶段流水线方式解决ASTE任务，第1阶段通过联合标记方案预测模型中的方面和意见，第2阶段将第1阶段预测的方面和意见配对输出三元组，然而，这种流水线方法破坏了三元组结构内的交互，通常会遇到错误传播问题。Xu等[7]提出位置感知标记方案，通过计算ASTE任务中结构信息的分解特征，更好地捕获了三元组中元素之间的交互信息，但该模型无法捕捉方面和意见之间一对多的情况。Zhang等[8]提出多任务学习框架联合提取方面术语和意见术语，并使用双仿射评分器解析其对应的情感依赖关系，但该模型独立预测每个词对的情感极性，在生成三元组时不能保证情感一致性。Wu等[9]提出一种网格标记方案 (grid tagging scheme, GTS)，以端到端的方式解决面向方面的细粒度意见提取 (aspect-oriented fine-grained opinion extraction, AFOE) 任务，同时设

收稿日期：2022-05-31；**录用日期**：2022-11-04；**网络出版时间**：2023-01-09 13:22
网络出版地址：link.cnki.net/urlid/11.2625.V.20230106.1751.003
基金项目：国家自然科学基金 (62166024)
* **通信作者**. E-mail：hanhu_lzjtu@mail.lzjtu.cn

引用格式：*韩虎，孟甜甜. 面向语法加权图文本的方面情感三元组抽取 [J]. 北京航空航天大学学报，2024，50（2）：409-418.*
HAN H，MENG T T. Aspect sentiment triple extraction for grammar-weighted graph text[J]. Journal of Beijing University of Aeronautics and Astronautics，2024，50（2）：409-418（in Chinese）.

计了一种有效的推理策略，利用不同意见词之间的指示作用扩展到三元组提取，但该模型忽略了词对之间的语义关系，对每个数据集的推理策略轮数不一致，可能会导致推理不稳定、时间复杂度过高等问题。

以上方法都侧重于设计新的标记方式，将ASTE转化为统一的任务，不能有效建立单词之间的联系。同时，三元组之间存在各种关系，但却忽略了其语义相似性和句法依赖性等关系。例如，单词“price”和“is”存在主题依赖关系，即“nsubj”，表明存在一个方面。句法中的2个意见词“delicious”和“expensive”都被标注为“acomp”，表明其具有相同的属性。因此，引入语法依赖是非常必要的，如何将语义信息和语法距离反馈到图文本是重点问题。本文利用线性形式中的语义联系和图结构中的句法依存关系构建ASTE模型。对于每条评论语句，将其转换为唯一的图表示，其中，每个节点都是一个单词，并且基于单词本身、相邻关系和句法依赖关系构建边。由此，可以获得每个单词的精确含义，并深入了解它们之间的关系。本文主要贡献如下：

1）结合线性形式中的序列特征和门控注意力单元(gated attention unit, GAU)区分词的属性和词对之间的关系。

2）基于语法距离加权图文本增强单词和三元组元素之间的交互。

3）在4个公开数据集上的实验结果表明，该模型对于不同数据集的F_1值均有所提升。

1 相关工作

1.1 三元组抽取

近年来，ASTE任务备受关注，并取得了一定进展。Xu等[5]提出一种双通道跨度剪枝策略，通过对ATE和OTE任务的显式监督，缓解了跨度枚举引起的高计算成本，并提高了候选方面词和意见词的配对机会。Wu等[9]通过GTS推理策略标记所有词对之间的关系，将AFOE任务的所有方面及意见因素纳入统一的GTS中。Huang等[10]提出利用两阶段的流水线方法增强方面词和意见词之间的相关性，通过序列标记提取方面和意见，将一组名为Perceivable Pair的人工标签插入序列中，该标签表示方面词和意见词的跨度，可以建立每个候选方面-意见对的相关性。Mukherjee等[11]提出一种基于指针网络的编码器-解码器架构，指针网络在检测方面及意见各自的跨度时，可以有效捕获它们之间的相互依赖性。

1.2 标注方法

针对ASTE任务，本文采用标签集$M=\{\text{A},\text{O},\text{POS},\text{NEU},\text{NEG},\text{N}\}$描述句子中任意2个词$(w_i,w_j)$之间的关系。其中，A和O分别表示单词$w_i$、$w_j$是相同的方面词和意见词；当$(w_i,w_j)$由一个方面词和一个意见词组成时，标签POS、NEU和NEG分别表示积极、中立和消极的情绪类别；标签N表示词对(w_i,w_j)存在其他关系。如图1所示，在由句子“Foods are delicious but the price is expensive”生成的网格标记中，坐标(0,0)、(5,5)分别构成词对(Foods,Foods)和(price,price)，且被标记为A，同样的逻辑也适用于意见词。此外，坐标(2,0)被标记为POS，因为其构成了一个正确的三元组(Foods, delicious,POS)，包含了方面词、意见词和情感类别。简单起见，使用上三角网格对标记情况进行表示。

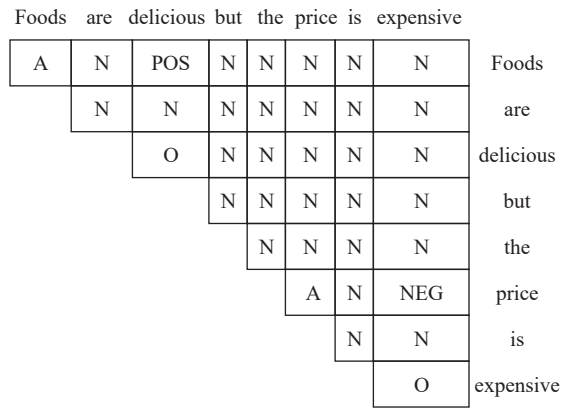


图1 GTS标记示例

Fig. 1 GTS tagging example

2 本文模型

本文提出了一种基于门控注意力和距离加权图文本的ASTE模型(a model based on gated attention and distance-weighted graph text for ASTE, GADWGT)，整体架构如图2所示，主要由嵌入层与隐藏层、门控注意力层、加权图序列层、推理层、解码层和输出层6部分组成。

2.1 问题定义

给定包含n个单词的句子$\boldsymbol{X}=[w_1,w_2,\cdots,w_n]$。首先，为每个单词分配2个标签：方面标签和意见标签。然后，利用$Y=\{\mathcal{B},\mathcal{I},\mathcal{O}\}$进行标注，标签$\mathcal{B}$和$\mathcal{I}$分别表示方面术语或意见术语组块的开始和内部，$\mathcal{O}$表示组块外的单词。同时，对每个方面术语用情感极性标签$S=\{\text{POS},\text{NEU},\text{NEG}\}$进行标注。本文旨在从句子中提取三元组集合$\gamma=\{(a,o,s)_m\}_{m=1}^{|\gamma|}$，$(a,o,s)_m$表示句子$\boldsymbol{X}$中的一个三元组，$a$、$o$和$s$分别表示方面、意见和其对应的情感极性，$|\gamma|$为三元组

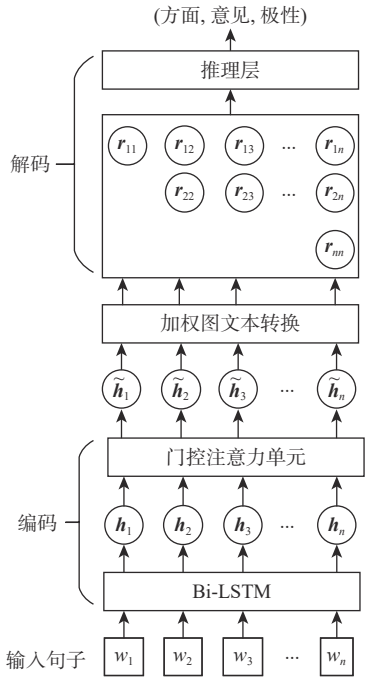


图 2　模型整体架构

Fig. 2　Overall architecture of model

的总数。

2.2　嵌入层和隐藏层

嵌入层将每个单词或词语通过嵌入矩阵映射到低维实值向量空间中。针对输入向量, 本文有 2 种表示方式: ①使用 Glove[12] 和 fastText[13] 进行双嵌入; ②使用 BERT[14] 预训练语言模型。通过双向长短时记忆网络 (bidirectional long-short-term memory network, Bi-LSTM) 学习句子中单词的隐藏信息 $\boldsymbol{H}=[\boldsymbol{h}_1,\boldsymbol{h}_2,\cdots,\boldsymbol{h}_n]$, 其中, $\boldsymbol{h}_t$为 Bi-LSTM 处于t时刻的隐藏状态向量。

2.3　门控注意力层

原始的 Transformer[15] 中, 自注意力机制未加入位置编码, 只在传入时加入绝对位置编码, 将 2 个位置上的“相同单词”看做同一个表示, 从语义上看是不合理的, 且绝对位置只能集中处理训练集中句子的最大长度, 若超过则无法处理。因此, 本文利用旋转式位置编码方式, 使得模型关注到单词之间的相对位置, 并将其作为重要指示信息引入 ABSA 模型中。计算方式定义如下:

$$\begin{cases} f(\boldsymbol{h},t)=\boldsymbol{R}_t\boldsymbol{h} \\ \boldsymbol{R}_t=\begin{bmatrix}\cos\alpha_t & -\sin\alpha_t \\ \sin\alpha_t & \cos\alpha_t\end{bmatrix} \end{cases} \tag{1}$$

$$f(\boldsymbol{q},m)^{\mathrm{T}}f(\boldsymbol{k},n)=g(\boldsymbol{q},\boldsymbol{k},m-n) \tag{2}$$

式中: $f(\cdot,t)$表示计算绝对位置信息的函数; α_t为单词t的位置信息; $\boldsymbol{q}$和$\boldsymbol{k}$分别为句中第 m 个单词隐藏层表示的查询向量和第 n 个单词隐藏层表示的键向量; m和n分别为评论文本中m位置和n位置; $g(\boldsymbol{q},\boldsymbol{k},m-n)$表示单词位置间的内积运算。

现有注意力机制受限于较短的上下文, 为加强单词之间的联系, 加入 GAU[16] 获得更高级的表示。门控注意力结构如图 3 所示。

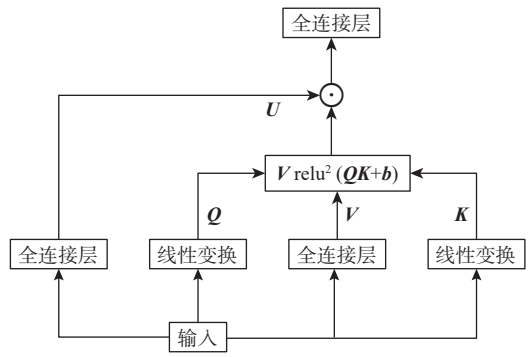


图 3　门控注意力结构

Fig. 3　Architecture of gated attention

门控注意力将门控线性单元 (gated linear unit, GLU) 与注意力作为统一的层, 尽可能多地共享二者之间的计算参数, 提高注意力分数的计算效率。定义如下:

$$\begin{cases} \tilde{\boldsymbol{H}}=(\boldsymbol{U}\odot\boldsymbol{A}\hat{\boldsymbol{V}})\boldsymbol{W}_h \\ \hat{\boldsymbol{V}}=\boldsymbol{A}\boldsymbol{V} \end{cases} \tag{3}$$

式中: $\tilde{\boldsymbol{H}}$为最终的门控单元输出; $\hat{\boldsymbol{V}}$为经过注意力计算的全连接层特征表示; $\boldsymbol{W}_h$为参数矩阵; $\boldsymbol{U}$为 GLU 的输出特征; $\boldsymbol{V}$为隐藏状态输入全连接层的特征表示; $\boldsymbol{A}$为包含单词注意力权重的矩阵, 定义为

$$\boldsymbol{A}=\mathrm{relu}^2(\boldsymbol{Q}(\boldsymbol{Z})\boldsymbol{K}(\boldsymbol{Z})^{\mathrm{T}})+\boldsymbol{b} \tag{4}$$

其中: relu 表示非线性激活函数; $\boldsymbol{Z}$为共享的特征表示; $\boldsymbol{Q}(\cdot)$和$\boldsymbol{K}(\cdot)$分别为对$\boldsymbol{Z}$每个维度和偏移量的线性变换; $\boldsymbol{b}$为相对位置偏量。

2.4　加权图序列层

图序列层主要由句法解析和图卷积网络 (graph convolutional network, GCN) 2 部分组成。前者结合依存关系树计算每个单词的相对权重, 并将其融入每层网络的输出中; 后者融合语法距离权重到每层网络, 对 GCN 的邻接矩阵进行加权。

2.4.1　句法解析

评论语句中表达的情感往往构成一个较为复杂的语法结构, 包括否定和对比等。因此, 单凭 GAU 学习单词特征难以客观全面地体现句子中上下文对判定 ASTE 的影响。本文将上下文与方面术语间的距离定义为二者在句法依赖树中的路径长度, 即语法距离。在句法依赖树中, 单词之间通过依赖关系进行关联, 与单词在评论语句中的具体位置无关, 在句法依赖树中, 越是相关的单词, 其语法距离越小。以图 4 为例, 方面“Foods”和“price”的依赖树语法距离分别如图 5 和图 6 所示。

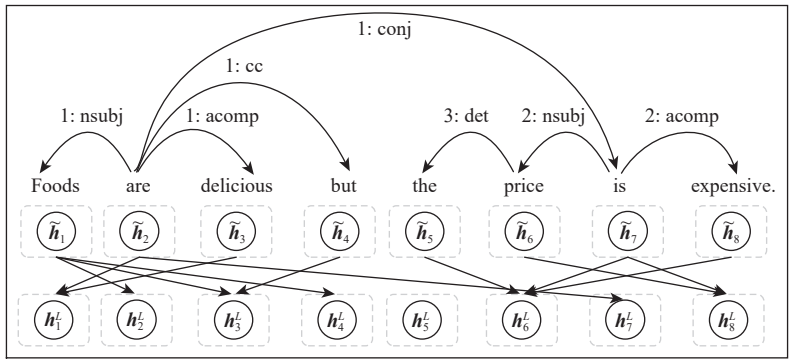


图4 加权图文本架构

Fig. 4 Architecture of weighted graph text

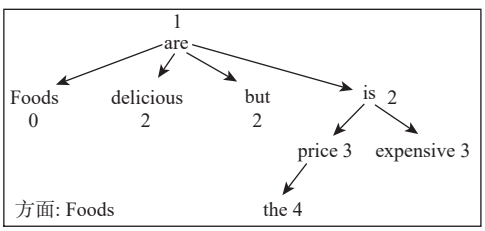


图5 方面“Foods”的依赖树语法距离

Fig. 5 Syntax distance of dependence tree for aspect “Foods”

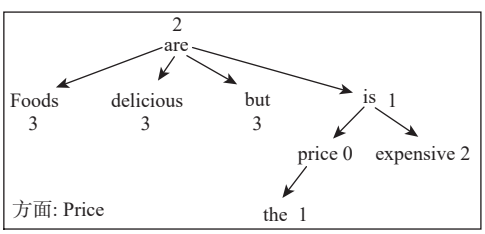


图6 方面“price”的依赖树语法距离

Fig. 6 Syntax distance of dependence tree for aspect “price”

基于句法依赖树计算语法距离时，离方面越近的上下文权重越高。对于每一个方面，由文献[17]算法可得到对应的语法距离向量$\boldsymbol{D}=(d_1,d_2,\cdots,d_n)$，则评论语句中各单词$w_i$的权重为

$$W_e(i)=1-\frac{d_i}{d_{\max}} \tag{5}$$

式中：$d_{\max}$为$\boldsymbol{D}$中的最大值，通过$W_e(i)$加权邻接矩阵$\bar{\boldsymbol{A}}$。

2.4.2 图卷积网络

对于图$\boldsymbol{G}=(\bar{\boldsymbol{V}}^w,\boldsymbol{E})$，$\bar{\boldsymbol{V}}^w$为节点集合，$\boldsymbol{E}$为边的集合。若节点$\bar{V}_i^w$和$\bar{V}_j^w$在句法上有依赖关系，则对应其邻接矩阵中$E_{ij}=1$。将当前节点$\bar{V}_i^w$第$l$层GCN输出记为$\bar{\boldsymbol{h}}_i^l$：

$$\bar{\boldsymbol{h}}_i^l=\mathrm{relu}\left(\sum_{j=1}^{n}E_{ij}\bar{\boldsymbol{h}}_i^{l-1}\boldsymbol{W}^l+\boldsymbol{b}^l\right) \tag{6}$$

式中：$\bar{\boldsymbol{h}}_i^{l-1}$为$l-1$层GCN的输出结果；权重$\boldsymbol{W}^l$和偏置项$\boldsymbol{b}^l$为可训练参数。

本文利用语法距离特征加权图卷积网络，定义如下：

$$\bar{\boldsymbol{h}}_i^l=\sigma\left[\left(\sum_{j=1}^{n}\bar{\boldsymbol{A}}'_{ij}\boldsymbol{g}_i^{l-1}+\boldsymbol{b}^l\right)\Big/\left(d_i^w+1\right)\right] \tag{7}$$

式中：$\sigma(\cdot)$为激活函数；$\bar{\boldsymbol{A}}'_{ij}$为通过距离加权后的邻接矩阵；$d_i^w$为单词$w_i$的度；$\boldsymbol{g}_i^{l-1}$为$l-1$层GCN的输出结果。本文将每层网络都通过语法距离加权，降低了GCN学习中的噪声信息，则L层加权GCN的最终输出为$\bar{\boldsymbol{h}}^L=[\bar{\boldsymbol{h}}_1^L,\bar{\boldsymbol{h}}_2^L,\cdots,\bar{\boldsymbol{h}}_n^L]$。

2.5 推理层

推理层旨在借助推理策略来生成词与词之间的关系特征表示。首先，将单词w_i和单词w_j的高层表示进行拼接，得到所有词对之间的初始特征表示$\boldsymbol{r}_{ij}$；其次，利用池化操作推断单词之间的对应关系，由于词对之间的关系对应于上三角网格中的每个元素，将其输入到线性全连接层可计算初始概率分布$\boldsymbol{z}_{ij}$；然后，通过拼接操作增强词对的关系特征表示$\tilde{\boldsymbol{r}}_{ij}$；最后，生成新的特征表示$\boldsymbol{e}_{ij}$。定义如下：

$$\boldsymbol{z}_i=\mathrm{maxpooling}(\boldsymbol{z}_i,:) \tag{8}$$

$$\boldsymbol{z}_j=\mathrm{maxpooling}(\boldsymbol{z}_j,:) \tag{9}$$

$$\tilde{\boldsymbol{r}}_{ij}=[\boldsymbol{r}_{ij};\boldsymbol{z}_i;\boldsymbol{z}_j;\boldsymbol{z}_{ij}] \tag{10}$$

$$\boldsymbol{e}_{ij}=\boldsymbol{W}_g\tilde{\boldsymbol{r}}_{ij}+\boldsymbol{b}_g \tag{11}$$

式中：$\boldsymbol{W}_g$和$\boldsymbol{b}_g$分别为权重矩阵和偏置项；$[\cdot;\cdot]$表示拼接操作；$\boldsymbol{z}_i$和$\boldsymbol{z}_j$用于捕获w_i、w_j与其他词之间的关系特征表示。在推理层中，词与词之间的关系用上三角矩阵来表示，因此，模型最终预测的是上三角矩阵。

2.6 解码层

本文根据所有的标记标签进行解码，使用标签及神经网络进行预测并训练，得到最终的方面-意见-情感三元组。当获得句子的GTS标注结果后，通过严格的匹配，从图1所示的标签中提取三元组。但是，由于GTS标注中含有大量无用的N标

签, 会引起三元组整体召回率下降, 必须降低三元组匹配的限制。

首先, 在上三角矩阵中找到标签 A、O 组成完整的方面术语或意见术语。然后, 遍历每个方面术语和意见术语交点处的标签, 若存在情感标签, 则将该方面-意见对的情感作为最多标签的情感, 若标签数量相同, 选择顺序为消极—>积极—>中立。

算法 1　GA-DWGT 解码算法。

输入: 句子$\boldsymbol{X}$的标注结果$R(w_i,w_j)$。

输出: 方面-意见-情感三元组γ。

1. 初始化方面集M_A、意见集M_O、三元组集γ为空集。

2. while 单词片段的起始端索引 $l \leqslant n$ and 单词片段的结束端索引 $r \leqslant n$ do

3.　　if $R(w_i,w_j)=M_A$　　where $l \leqslant i \leqslant r$

4.　　if $R(w_{l-1},w_{l-1}) \neq M_A$ and $R(w_{r+1},w_{r+1}) \neq M_A$ then

5.　　将单词$\{w_l,w_r\}$作为一个方面术语a, $M_A \leftarrow M_A \cup \{M_A\}$

6.　　end if

7.　　end if

8. end while

9. while $a \in M_A$ and $o \in M_O$ do

10.　　while $w_i \in M_A$ and $w_j \in M_O$ do

11.　　初始化s为None

12.　　if　num(NEG) ⩾ num(POS)　and num(NEG) ⩾ num(NEU)

13.　　then $s \leftarrow$ NEG

14.　　elif　num(POS) ⩾ num(NEU)　and num(POS) ⩾ num(NEG)

15.　　then $s \leftarrow$ POS

16.　　elif　num(NEU) ⩾ num(NEG)　and num(NEU) ⩾ num(POS)

17.　　then $s \leftarrow$ NEU

18.　　end if

19.　　if $s \neq$ None then

20.　　$\gamma \leftarrow \gamma \cup \{a,o,s\}$

21.　　end if

22.　　end while

23. end while

24. return 三元组集 γ

算法时间复杂度为$O(n^2)$, 第 2 行到第 8 行为方面术语提取伪代码, 意见术语提取同理可得。

2.7　输 出 层

输出层将词对之间最终的特征表示$\boldsymbol{e}_{ij}$输出到线性层, 利用 softmax 进行分类:

$$y_{ij}=\mathrm{softmax}(\boldsymbol{W}_y\boldsymbol{e}_{ij}+\boldsymbol{b}_y) \tag{12}$$

式中: y_{ij}为预测标签; $\boldsymbol{W}_y$和$\boldsymbol{b}_y$为可训练参数。

使用交叉熵损失函数衡量模型, 以最小化所有词对(w_i,w_j)的交叉熵误差为目标。损失函数定义如下:

$$L_{\mathrm{oss}}=-\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k\in C}\xi\left(\hat{y}_{ij}=k\right)\ln y_{ij} \tag{13}$$

式中: $\xi(\cdot)$表示指示函数; $\hat{y}_{ij}$为词对(w_i,w_j)的基础真实标签。

3　实　验

3.1　数 据 集

本文实验采用 4 个公开数据集进行验证, 分别为 SemEval-2014 Task 4(包括 Laptop 和 Restaurant 2 个领域)[1]、SemEval-2015 Task 12[18] 和 SemEval-2016 Task 5[19]。数据集被划分为训练集、测试集和验证集。具体数据统计结果如表 1 所示。

表 1　数据集统计结果

Table 1　Dataset statistics results

数据集		句子数量	三元组数量	情感极性积极的三元组数量	情感极性中立的三元组数量	情感极性消极的三元组数量
lap14	训练集	899	1 452	808	111	533
	验证集	225	383	199	48	136
	测试集	332	547	364	67	116
res14	训练集	1 259	2 356	1 693	172	491
	验证集	315	580	427	46	107
	测试集	493	1 008	784	68	156
res15	训练集	603	1 038	799	29	210
	验证集	151	239	181	9	49
	测试集	325	493	324	25	144
res16	训练集	863	1 421	1 036	55	330
	验证集	216	348	263	8	77
	测试集	328	525	416	30	79

3.2　超参数设置

模型在词嵌入层使用 2 种嵌入方式: ①使用 Glove 训练的 300 维通用嵌入矩阵和 fastText 训练的 100 维特定领域嵌入矩阵进行双嵌入, 隐藏状态向量维度设置为 100; ②使用 BERT 预训练语言模型进行嵌入, 词向量维度、隐藏状态向量维度均设置为 768。GCN 的层数设置为 2, 此时模型的性能表现最好。使用 Adam 作为优化器, 学习率设置为 0.001, Batch size 设置为 32, 训练迭代次数为 600。为避免过拟合, 将丢弃率设置为 0.5。

3.3　评价指标

与基准模型对比中, 采用精准率 P、召回率

R及F_1值作为评价指标。对于抽取的三元组(方面,意见,情感),当且仅当三元组中每个元素与测试数据集中的元素对应时视为正确。

$$P=\frac{T_P}{T_P+F_P} \tag{14}$$

$$R=\frac{T_P}{T_P+F_N} \tag{15}$$

$$F_1=\frac{2PR}{P+R} \tag{16}$$

式中:T_P为正确预测的三元组数量;F_P为预测不正确的三元组数量;F_N为未能预测出的正确三元组数量。

3.4 对比实验

为评估本文模型的性能,将其与一系列基准模型进行比较。

1)Peng-unified-R+PD。Peng等[6]提出一种两阶段的管道方法。第1阶段采用BIESO标签机制提取句子中的方面词和对应的意见词,利用二者之间的交互性;第2阶段生成所有的候选三元组,并基于分类器确定其是否有效。

2)Li-unified-R+PD。结合Peng等[6]提出的管道方法,第1阶段使用Li-unified-R[20]模型提取句子中的方面、意见及情感极性,第2阶段应用分类器获得有效三元组。

3)Peng-unified-R+IOG。采用Peng-unified-R模型[6]提取句子中的方面和情感,使用IOG模型[21]编码给定方面的信息以提取对应的意见词,从而生成最终的三元组。

4)IMN+IOG。采用交互式多任务学习网络(interactive multi-task learning network, IMN)[22]提取句子中显式的目标方面,再采用IOG模型[21]提取对应的意见词,从而生成最终的三元组。

5)PASTE。采用基于指针网络的解码框架调整编码器-解码器架构,考虑评论语句的整体检测范围,有效捕获方面和意见之间的交互,并预测其指向的情感极性[11]。

6)GTS。Wu等[9]设计了GTS方案,以端到端的方式解决三元组抽取问题,其采用推理策略充分利用了不同意见元素之间的指示作用。

本文模型借助加权图文本增强单词之间的交互信息,结合GTS推理策略输出预测结果,设计解码算法生成三元组。

3.5 实验结果与分析

表2给出了GA-DWGT模型在4个数据集上与相关基准模型的对比结果。可以看出,在未使用BERT的模型中,本文提出的GA-DWGT模型F_1性能均优于其他相关模型,表明其能够有效处理ASTE任务。进一步分析实验数据可以看出,GA-DWGT模型在res16数据集上的F_1性能与基准模型中表现最好的GTS-LSTM模型的水平相当,在另外3个数据集上的性能均优于所有基准模型,表明GA-DWGT模型的泛化能力最好。深度学习中,实验数据规模和模型深度对算法性能影响较大,且词向量在经过神经元非线性处理后得到的特征向量对于构建情感语义有积极意义。可以看到,IMN+IOG模型在res14和res15数据集上明显优于Peng-unified-R+IOG模型,这是因为IMN具有多层网络结构,且使用多领域文档级情感分类数据作为辅助任务,表明对词特征进行多次抽象,可以有效提高模型从复杂语句中识别方面词及相关情感词。然而,上述对比模型均是管道方法,对比现有端到端的基准模型GTS-LSTM, GA-DWGT在lap14和res14上的F_1值提升了1.3%和1.72%,在res15和res16上的F_1值提升了1.15%和0.73%,证明了本文采用门控注意

表2 GA-DWGT模型与基准模型结果对比

Table 2 Comparison of GA-DWGT model and benchmark model results %

模型	P				R				F_1			
	lap14	res14	res15	res16	lap14	res14	res15	res16	lap14	res14	res15	res16
Li-unified-R+PD	42.25	41.44	43.34	38.19	42.78	68.79	50.73	53.47	42.47	51.68	46.69	44.51
Peng-unified-R+PD	40.40	44.18	40.97	46.76	47.24	62.99	54.68	62.97	43.50	51.89	46.79	53.62
Peng-unified-R+IOG	48.62	58.89	51.70	59.25	45.52	60.41	46.04	58.09	47.02	59.64	48.71	58.67
IMN+IOG	49.21	59.57	55.24		46.23	63.88	52.33		47.68	61.65	53.75	
PASTE	52.10	63.40	54.80	62.30	48.10	61.90	52.60	63.60	50.00	62.60	53.70	62.90
GTS-CNN	55.93	70.79	60.09	62.63	47.52	61.71	53.57	66.98	51.38	65.94	56.64	64.73
GTS-LSTM	59.42	67.28	63.26	66.07	45.13	61.91	50.71	65.05	51.30	64.49	56.29	65.56
GA-DWGT	59.09	70.67	65.87	67.80	47.34	62.42	50.92	65.39	52.60	66.21	57.44	66.29
PASTE-BERT	59.70	68.70	63.60	68.00	55.30	63.80	59.80	67.70	57.40	66.10	61.60	67.80
GTS-BERT	57.52	70.92	59.29	68.58	51.92	69.49	58.07	66.60	54.58	70.20	58.67	67.58
GA-DWGT-BERT	63.80	70.74	64.11	65.41	51.74	68.38	59.92	70.08	57.94	70.54	61.95	67.66

力机制与加权图文本增强 ASTE 任务的强有力优势。在基于 BERT 的模型中, 所有模型性能都有显著提高。其中, GTS-BERT 模型在 res14 数据集上显著高于 PASTE-BERT 模型, 在 res16 数据集上表现相当, 这是因为 GTS-BERT 模型的网格标签预测任务和 BERT 模型的预训练任务本质上都属于判别式任务, 在数据语法结构较强的情况下表现更加优异。相较于 GTS-BERT 模型, 本文提出的 GA-DWGT-BERT 模型 F_1 性能进一步提高, 这是因为对于语法结构较弱的评论语句, 本文可以有效融入语义和语法知识来丰富表示, 从而增强三元组抽取的准确率。

本文统计了对比模型的可训练参数数量, 如表 3 所示。Peng-unified-R+PD 模型的参数数量最少, 虽然其性能优于 Li-unified-R+PD, 但是仍不及其他模型; Peng-unified-R+IOG 模型借鉴了 Peng-unified-R+PD 模型的管道方法, 其模型计算复杂度与 Peng-unified-R+PD 相比略高, 但是性能整体上有很大的提升; PASTE 模型基于指针网络捕获方面跨度和意见跨度的相互依赖性, 虽然其在 4 个数据集上的性能均高于 Li[7] 和 Peng[6] 等提出的方法, 但 PASTE 模型的计算复杂度较高; GTS-LSTM 模型的可训练参数规模比 GTS-CNN 小, 但二者性能相当, 因此, 若要应用到实际场景中, GTS-LSTM 模型效率高于 GTS-CNN 模型; GA-DWGT 模型的参数规模与 GTS-LSTM 接近, 但是性能整体上优于 GTS-LSTM 和其他模型, 表明 GA-DWGT 模型的训练效率远高于其他模型。

表 3　对比模型的可训练参数数量

Table 3　Number of trainable parameters of comparison models

模型	可训练参数数量/10^6
Li-unified-R+PD	0.31
Peng-unified-R+PD	0.21
Peng-unified-R+IOG	0.36
IMN+IOG	0.45
PASTE	0.71
GTS-CNN	0.47
GTS-LSTM	0.33
GA-DWGT	0.33

3.6　消融实验

本文 GA-DWGT 模型均基于 Glove 和 fastText 双嵌入, 且 GA-DWGT-BERT 显示出类似的趋势。

为检查 GA-DWGT 中不同组件的有效性, 对其进行消融研究。

1) GA-DWGT-GAU-WG。模型由 Bi-LSTM 层和推理层组成, Bi-LSTM 层用来学习输入语句的序列特征, 推理层利用网格标记推理策略预测三元组。

2) GA-DWGT-WG。模型在 Bi-LSTM 层后添加了 GAU, 以学习单词之间的语义联系。

3) GA-DWGT-GAU。模型在 Bi-LSTM 层后添加了加权图序列表示, 以增强三元组元素之间的交互, 提高抽取准确性。

4) GA-DWGT。本文完整的模型, 既考虑了单词之间的语义和语法关系, 也考虑了三元组元素之间的交互信息。

如表 4 所示, 可以看到, 对比 GA-DWGT 模型, 其他模型 F_1 值急剧下降, 说明门控注意力机制和加权图文本可以准确把握三元组元素之间的关系, 增强了三元组抽取的准确度。GA-DWGT-GAU-WG 模型在 4 个数据集上的 F_1 值均不及其他模型, 表明虽然推理策略的提出有力推动了 ASTE 任务的解决, 但是单词间语义和语法等信息仍不可忽略。GA-DWGT-WG 模型整体上性能优于 GA-DWGT-GAU 模型, 尤其表现在 res15 和 res16 数据集上, 但对比 GA-DWGT 模型仍有差距, 分析其原因在于: 相较于句子中的句法依存关系, res15 和 res16 数据集对单词间的语义信息更加敏感, 而 lap14 和 res14 数据集更依赖于上下文的语法特征, 表明利用门控注意力机制对 ASTE 任务虽然有利, 但也不可忽略句子中上下文的句法依赖信息。由此可见, 虽然利用图文本建模上下文可以很好地感知关系语义并区分单词的特征, 但推理策略和序列特征学习也不可或缺。

3.7　GCN 层数分析

为验证 GCN 层数对 GA-DWGT 模型性能的影响, 设置 GCN 层数从 1 到 10, 实验结果如图 7 所示。可以看到, 当 GCN 层数为 2 时, 本文模型在

表 4　消融实验结果

Table 4　Ablation study results　　%

模型	P				R				F_1			
	lap14	res14	res15	res16	lap14	res14	res15	res16	lap14	res14	res15	res16
GA-DWGT-GAU-WG	52.29	67.75	58.79	66.46	43.85	59.79	49.89	61.19	47.71	63.18	53.98	63.72
GA-DWGT-WG	57.07	67.78	59.65	67.68	42.20	61.01	55.01	64.29	48.52	64.22	57.23	65.94
GA-DWGT-GAU	55.71	69.57	62.89	65.59	43.85	59.59	49.89	62.55	49.08	64.20	55.65	64.03
GA-DWGT	59.09	70.67	65.87	67.80	47.34	62.42	50.92	65.39	52.60	66.21	57.44	66.29

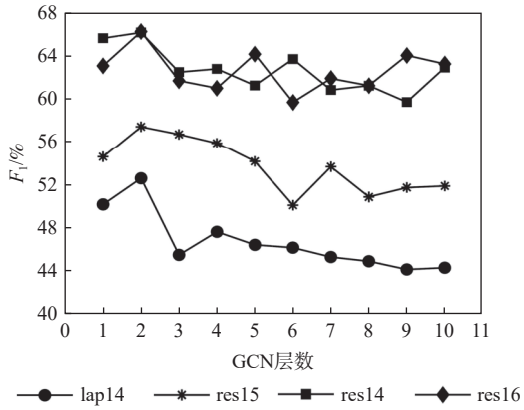


图7 GCN层数与 F_1 的关系

Fig. 7 Relationship between number of GCN layers and F_1

4个数据集上均取得了最优结果，其性能分别比平均值高出5.89%、3.49%、3.74%和3.62%；同时可以看到，随着GCN层数的增加，模型性能并未提高，并出现了不同程度的波动，总体呈下降趋势。结果表明，图神经网络融合特征的能力与网络层数有关，确定合理的GCN层数可以改善模型性能。结合实验结果，本文设置GCN层数为2。

3.8 案例分析

表5给出了5个典型的案例。例句1～2取自lap14测试集，例句3～5取自res15测试集。其中，例句3是不附带复杂词序的简单语句，所有模型都能够准确预测。例句1表达了一种负面情绪，可以看到，GTS-CNN和GTS-LSTM都错误地将“maintain”预测为一个方面词，且对情绪类别做出了错误的判断。例句2包含2个三元组，其他方法只提取到了方面词“screen”对应的三元组，本文提出的GA-DWGT成功识别了方面词“speed”对应的三元组。针对例句1和例句2，GA-DWGT都做出了正确的判断，表明其可以更好地理解上下文并区分单词类别。例句4委婉地表达了职工的消极态度，GTS-CNN和GTS-LSTM都预测职工态度为积极的，且GTS-CNN错误地将“should be”预测为一个意见词，而GA-DWGT仅在预测“should”为意见词时出了差错，对比之下，GA-DWGT的识别能力更强。例句5是一个复杂语句，包含3个三元组，其中，方面“food”对应2个意见词，可以看到，GTS-LSTM错误地将“food”和“excellent”匹配进一个三元组，GTS-CNN和GA-DWGT都做出了正确的预测。因此，进一步证明了GA-DWGT可以更好地理解上下文中的语义联系，并准确地识别三元组中的元素关系。

表5 案例分析

Table 5 Case study

例句	正确的三元组	模型预测三元组		
		GTS-CNN	GTS-LSTM	GA-DWGT
Made interneting difficult to maintain	(interneting, difficult, NEG)	(maintain, difficult, POS)	(maintain, difficult, POS)	(interneting, difficult, NEG)
Made interneting difficult to maintain and the screen is very sharp	(speed, much more, POS) (screen, sharp, POS)	(screen, sharp, POS)	(screen, sharp, POS)	(speed, much more, POS) (screen, sharp, POS)
The bread is top notch as well	(bread, top notch, POS)	(bread, top notch, POS)	(bread, top notch, POS)	(bread, top notch, POS)
The staff should be a bit more friendly	(staff, friendly, NEG)	(staff, more friendly, POS)	(staff, more friendly, POS) (staff, should be, POS)	(staff, friendly, NEG) (staff, should, NEG)
The food was extremely tasty, creatively presented and the wine excellent	(food, tasty, POS) (food, creatively presented, POS) (wine, excellent, POS)	(food, tasty, POS) (food, creatively presented, POS) (wine, excellent, POS)	(food, tasty, POS) (food, creatively presented, POS) (wine, excellent, POS) (food, excellent, POS)	(food, tasty, POS) (food, creatively presented, POS) (wine, excellent, POS)

4 结 论

本文提出了一种基于门控注意力与加权图文本的方面-意见-情感三元组抽取模型。

1）采用双嵌入技术弥补了传统词嵌入在语义表达能力上的缺失，并且用Bi-LSTM作为基础编码器，强化了评论语句中单词之间的双向序列特征融合。

2）通过距离加权图文本增强了评论语句中单词之间的句法依赖关系，用以提高三元组的分类准确性。

3）通过GTS推理策略生成包含语义和语法信息的关系特征表示，用以预测单词之间的成对关系。

4）通过实验发现，GA-DWGT在4个公开数据集上的性能表现良好，在 P、R、F_1 3项指标上均取得了较好的效果。

随着预训练语言模型的推出，未来工作中可以尝试使用生成式方法构建方面情感三元组抽取的新范式。

参考文献（References）

[1] PONTIKI M, GALANIS D, PAVLOPOULOS J, et al. SemEval-

2014 Task 4: Aspect based sentiment analysis[C]//Proceedings of the 8th International Workshop on Semantic Evaluation. Stroudsburg: ACL, 2014: 27-35.

[2] DONG L, WEI F R, TAN C Q. Adaptive recursive neural network for target-dependent twitter sentiment classification[C]//Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics. Stroudsburg: ACL, 2014, 2: 49-54.

[3] YIN Y C, WEI F R, DONG L, et al. Unsupervised word and dependency path embeddings for aspect term extraction[C]//Proceedings of the 25th International Joint Conference on Artificial Intelligence. Palo Alto: AAAI, 2016: 2979-2985.

[4] YANG B S, CARDIE C. Extracting opinion expressions with semi-Markov conditional random fields[C]//Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. Stroudsburg: ACL, 2012: 1335-1345.

[5] XU L, CHIA Y K, BING L D. Learning span-level interactions for aspect sentiment triplet extraction[C]//Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing. Stroudsburg: ACL, 2021: 4755-4766.

[6] PENG H Y, XU L, BING L D, et al. Knowing what, how and why: A near complete solution for aspect-based sentiment analysis[C]//Proceedings of the AAAI Conference on Artificial Intelligence. Palo Alto: AAAI, 2020, 34: 8600-8607.

[7] XU L, LI H, LU W, et al. Position-aware tagging for aspect sentiment triplet extraction[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing. Stroudsburg: ACL, 2020: 2339-2349.

[8] ZHANG C, LI Q C, SONG D W, et al. A multi-task learning framework for opinion triplet extraction[C]//Proceedings of the Findings of the Association for Computational Linguistics. Stroudsburg: ACL, 2020: 819-828.

[9] WU Z, YING C C, ZHAO F, et al. Grid tagging scheme for aspect-oriented fine-grained opinion extraction[C]//Proceedings of the Findings of the Association for Computational Linguistics. Stroudsburg: ACL, 2020: 2576-2585.

[10] HUANG L Z, WANG P Y, LI S J, et al. First target and opinion then polarity: Enhancing target-opinion correlation for aspect sentiment triplet extraction[EB/OL]. (2021-11-01)[2022-05-01]. https://arxiv.org/abs/2102.08549v1.

[11] MUKHERJEE R, NAYAK T, BUTALA Y, et al. PASTE: A tagging-free decoding framework using pointer networks for aspect sentiment triplet extraction[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing. Stroudsburg: ACL, 2021: 9279-9291.

[12] JEFFREY P, RICHARD S, CHRISTOPHER M. Glove: Global vectors for word representation[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing. Stroudsburg: ACL, 2014: 1532-1543.

[13] BOJANOWSKI P, GRAVE E, JOULIN A, et al. Enriching word vectors with subword information[J]. Transactions of the Association for Computational Linguistics, 2017, 5(1): 135-146.

[14] 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: ACL, 2019, 1: 4171-4186.

[15] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]//Proceedings of the 31st Annual Conference on Neural Information Processing Systems. La Jolla: NIPS, 2017: 5999-6009.

[16] HUA W Z, DAI Z H, LIU H X, et al. Transformer quality in linear time[EB/OL]. (2022-02-21)[2022-05-01]. https://arxiv.org/abs/2202.10447.

[17] 苏锦钿, 欧阳志凡, 余珊珊. 基于依存树及距离注意力的句子属性情感分类[J]. 计算机研究与发展, 2019, 56(8): 1731-1745.
SU J D, OUYANG Z F, YU S S. Aspect-level sentiment classification for sentences based on dependency tree and distance attention [J]. Journal of Computer Research and Development, 2019, 56(8): 1731-1745(in Chinese).

[18] PONTIKI M, GALANIS D, PAPAGEORGIOU H, et al. SemEval-2015 Task 12: Aspect based sentiment analysis[C]//Proceedings of the 9th International Workshop on Semantic Evaluation. Stroudsburg: ACL, 2015: 486-495.

[19] PATERIA S, CHOUBEY P K. SemEval-2016 Task 5: Aspect based sentiment analysis[C]//Proceedings of the 10th International Workshop on Semantic Evaluation. Stroudsburg: ACL, 2016: 318-324.

[20] LI X, BING L D, LI P J, et al. A unified model for opinion target extraction and target sentiment prediction[C]//Proceedings of the 33rd AAAI Conference on Artificial Intelligence. Palo Alto: AAAI, 2019: 6714-6721.

[21] FAN Z F, WU Z, DAI X Y, et al. Target-oriented opinion words extraction with target-fused neural sequence labeling[C]//Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Stroudsburg: ACL, 2019, 1: 2509-2518.

[22] HE R D, LEE W S, NG H T, et al. An interactive multi-task learning network for end-to-end aspect-based sentiment analysis[C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: ACL, 2020: 504-515.

Aspect sentiment triple extraction for grammar-weighted graph text

HAN Hu*, MENG Tiantian

(School of Electronic and Information Engineering, Lanzhou Jiaotong University, Lanzhou 730070, China)

Abstract: Aspect sentiment triple extraction includes three tasks: aspect term extraction, opinion term extraction, and aspect sentiment classification. However, research methods that solve this task in a pipeline way cannot utilize the interaction information between elements, and will also cause error propagation and redundant training. To solve the above problems, an aspect sentiment triple extraction method based on gated attention and weighted graph text is proposed, which makes full use of the semantic and grammatical relationships between triple elements to enhance element interaction. Firstly, the model uses a bidirectional long-short-term memory network to learn the sequence feature representation of sentences. Secondly, a gated attention unit is used to learn linear connections between words. Thirdly, a grammatical distance-weighted graph convolutional network is employed to enhance the interactions between triplet elements. Finally, a grid tagging inference strategy is applied to predict triples. Experimental results on four public datasets show that the proposed method can effectively enhance the interaction between triple elements and improve the accuracy of triple extraction. Moreover, the F_1 values of the proposed method are 57.94%, 70.54%, 61.95% and 67.66%, respectively, which are all improved compared to the baseline model.

Keywords: triple extraction; gated attention; weighted graph text; bidirectional long-short-term memory network; grid tagging

Received: 2022-05-31; **Accepted**: 2022-11-04; **Published Online**: 2023-01-09 13:22

URL: link.cnki.net/urlid/11.2625.V.20230106.1751.003

Foundation item: National Natural Science Foundation of China (62166024)

* **Corresponding author**. E-mail: hanhu_lzjtu@mail.lzjtu.cn

http://bhxb.buaa.edu.cn jbuaa@buaa.edu.cn
DOI: 10.13700/j.bh.1001-5965.2022.0329

结合空间注意力机制的实时鲁棒视觉跟踪

马素刚[1,2,*]，张子贤[1]，蒲磊[3]，侯志强[1,4]

(1. 西安邮电大学 计算机学院，西安 710121；2. 长安大学 信息工程学院，西安 710064；
3. 火箭军工程大学 作战保障学院，西安 710025；4. 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)

摘 要：为提高全卷积孪生网络(SiamFC)跟踪器在复杂场景下的跟踪能力，缓解跟踪器在跟踪过程中出现的目标漂移问题，提出一种结合空间注意力机制的实时目标跟踪算法。在SiamFC框架基础上，将改进的视觉几何组(VGG)网络作为主干网络，增强跟踪器对于目标深度特征的建模能力。对自注意力机制进行优化，提出一种即插即用的轻量级单卷积注意力模块(SCAM)，将空间注意力分解为2个并行的一维特征编码过程，减少空间注意力的计算复杂度。保留跟踪过程中的初始目标模板作为第1模板，通过分析连通域在跟踪结果响应图的变化动态选择第2模板，融合2个模板后对目标进行定位。实验结果表明：在OTB100、LaSOT和UAV123数据集上，所提算法相比于SiamFC跟踪成功率分别提高了0.082、0.045和0.045，跟踪精度分别提高了0.118、0.051和0.062；在VOT2018数据集上，所提算法相比于SiamFC在跟踪准确率、鲁棒性和期望平均重叠率上分别提高了0.029、0.276和0.134；跟踪速度达到了70帧/s，能够满足实时跟踪的需求。

关 键 词：目标跟踪；孪生网络；注意力机制；模板更新；深度学习
中图分类号：TB391.4
文献标志码：A **文章编号**：1001-5965(2024)02-0419-14

目标跟踪是计算机视觉领域中的一个重要研究课题，主要目的是在连续图像序列中的第1帧给定任意位置的感兴趣目标，通过一定方法估计后续图像中目标的位置。随着目标跟踪技术的不断发展，其已被广泛应用于无人驾驶[1]、机器交互[2]、外科手术[3]和海洋探索[4]等诸多领域。由于目标跟踪任务具有较少的先验信息，且跟踪对象的运动状态具有较高的不确定性，尽管目标跟踪已经过了数十年的发展，但在诸如背景杂乱、运动模糊和目标遮挡等复杂场景下仍具有较大挑战。

当前，主流的目标跟踪算法可以分为2类，分别为基于相关滤波[5]的目标跟踪算法和基于深度学习[6]的目标跟踪算法。基于相关滤波的目标跟踪算法通过提取目标特征训练滤波器，并对下一帧图像作滤波运算，以结果响应峰值所在位置作为目标位置。在特征构建方面，早期的相关滤波算法主要使用手工特征构建目标模型，如使用梯度直方图特征的核化相关滤波器(kernelized correlation filters, KCF)算法[7]、使用灰度特征的时空上下文(spatio-temporal context, STC)算法[8]和使用多通道颜色特征的颜色命名(color names, CN)算法[9]等。手工特征的构建算法相对简单，仅由统计计算完成，虽然使用此类特征的跟踪器的跟踪速度较快，但对目标特征建模能力较弱，跟踪器鲁棒性不强，难以应对复杂场景下的跟踪任务。随着深度学习的不断发展，依托大量数据训练的卷积神经网络(convolutional neural network, CNN)对目标特征有着强大的表征能力，吸引越来越多研究者的注意。Ma等[10]设计了使用深度特征的分层卷积特征(hierarchical convolutional features, HCF)跟踪器，Wang等[11]

收稿日期：2022-05-07；**录用日期**：2022-08-21；**网络出版时间**：2022-09-14 11:30
网络出版地址：link.cnki.net/urlid/11.2625.V.20220913.1335.001
基金项目：国家自然科学基金(62072370)；陕西省重点研发计划(2018ZDCXL-GY-04-02)；西安邮电大学研究生创新基金(CXJJZL2021011)
*通信作者. E-mail：msg@xupt.edu.cn

引用格式：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).

设计了浅层深度特征和深层深度特征自适应选择的基于全卷积网络的跟踪器 (fully convolutional network based tracker, FCNT), Bhat 等[12]设计了手工特征与深度特征融合的揭示深度跟踪力量跟踪器(unveiling the power of deep tracking, UPDT)。然而,基于相关滤波的跟踪器所使用的 CNN 最初是为图像分类任务设计的,意味着使用离线训练的 CNN 所提取的目标特征表示并不完全适合视觉跟踪任务。此外,高维度深度特征与复杂的相关滤波策略结合使得跟踪速度大幅降低,难以满足实时跟踪的需求。

近年来, Bertinetto 等[13]提出的全卷积孪生网络(fully convolutional Siamese network, SiamFC)跟踪算法对于跟踪速度和精度有着良好的平衡,其使用模板匹配思想训练的特征网络更适合目标跟踪任务的需要,受到广泛关注。SiamFC 算法将目标跟踪问题转化为特征块匹配问题,通过在离线训练阶段使用大量不同时间点的成对目标图像充分训练特征网络,使得提取到的特征能够更好地适应目标的外观变化;在线跟踪阶段,将模板图像特征与搜索图像特征进行互相关运算,获得跟踪结果响应图,再利用卷积的平移不变性定位目标所在位置。SiamFC 算法去除了传统跟踪算法架构中的模板更新阶段,相较与 CNN 结合的相关滤波算法,其在跟踪速度上有明显提升。

但是, SiamFC 算法存在如下问题:①使用层数较浅的 AlexNet 主干网络,难以获得目标的强大特征表示,缺乏对场景中干扰物体的辨识能力;②始终以第 1 帧目标作为模板与搜索区域做相似度计算,缺乏模板更新策略,固定的模板特征在复杂场景下的跟踪效果不理想。为进一步增强孪生网络框架在目标跟踪中的深度特征表示能力, He 等[14]提出双重连体孪生网络算法,通过构造语义分支和外观分支的双孪生网络结构来提高跟踪器的泛化能力。Pu 等[15]提出双注意力孪生网络算法,在孪生网络的双分支级联非局部空间注意力和通道注意力来实现对目标区域和通道的重点关注。Gupta 等[16]设计旋转变化模块,提高孪生网络对平面旋转对象的特征识别能力。针对孪生网络的模板更新问题,端到端学习的相关滤波跟踪 (correlation filter based tracking, CFNet) 算法[17]通过在模板分支引入相关滤波器在线调节模板特征。更新网络(UpdateNet)算法[18]通过累积模板和当前模板来估计下一帧模板。光流相关跟踪算法[19]设计历史序列质量判断网络,在搜索下一帧之前对所有历史序列特征进行自适应加权求和来更新模板特征。

为实现鲁棒跟踪,本文针对特征增强和模板更新问题对孪生网络跟踪器进行改进,提出一种结合轻量级空间注意力机制的实时目标跟踪算法。

1 本文算法

本文算法的总体框架如图 1 所示。CNN 为改进的 16 层视觉几何组网络[20](visual geometry group network with 16 layers, VGG16),用来提取模板图像和搜索区域图像的深度特征,同时使用单卷积注意力模块(single convolution attention moudule, SCAM)对模板特征进行增强,提高模板的抗干扰能力。为避免单一模板在跟踪过程中受到相似物体干扰及背景干扰,选取初始图像中的目标作为第 1 模板,采用连通域作为判断标准动态选择第 2 模板,与第 1

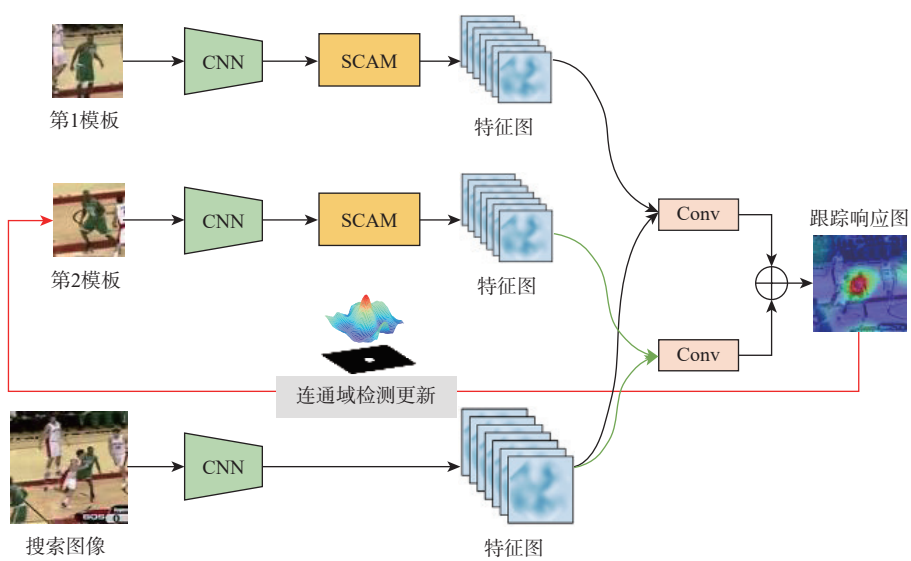


图 1 本文算法的总体框架

Fig. 1 Overall framework of proposed algorithm

模板进行融合, 共同决定跟踪响应结果, 使得跟踪效果更加鲁棒。

1.1　全卷积孪生网络

SiamFC 借助 CNN 强大的特征表达能力, 对模板与搜索图像特征之间的相似度进行比较, 相似度最大值的位置为目标中心。孪生网络的上下分支分别为目标模板和搜索图像, 将目标模板图像记为 $\boldsymbol{z}$, 搜索图像记为 $\boldsymbol{g}$, CNN 特征提取操作记为 $\phi(\cdot)$, 则 CNN 提取到的特征分别为 $\phi(\boldsymbol{z})$和 $\phi(\boldsymbol{g})$。通过卷积计算即可得到关于目标的跟踪响应图 $f(\boldsymbol{z},\boldsymbol{g})$, 表示为

$$f(\boldsymbol{z},\boldsymbol{g})=\phi(\boldsymbol{z})*\phi(\boldsymbol{g})+\boldsymbol{b} \tag{1}$$

式中: "$*$" 表示卷积运算; $\boldsymbol{b}$为特征偏置矩阵。由于卷积操作得到的响应图远小于搜索图像大小, 将峰值点在特征上的位置坐标乘以卷积步长即为目标在搜索图像上的具体位置。

1.2　轻量级空间注意力模块

增强特征点之间的远程依赖关系有利于视觉场景的全局理解, 如不断加深的深度神经网络和自然语言处理中的自注意力机制。但是, 简单地堆叠卷积层会使网络难以优化, 对于大尺寸特征图来说, 自注意力机制对特征点关系的建模计算过于复杂, 影响网络的建模速度。为快速建立特征点之间的空间依赖关系, 提高模型对不同区域对象的判别能力, 本文基于自注意力机制设计了一种轻量级的 SCAM。

自然语言处理中的自注意力机制本质上可以看作一个查询 (query) 与多个键 (key) 值 (value) 对之间的映射, 即当要查询某个单词在句中的含义时, 需要比较查询词与其他各单词的相似度, 再将相似度与句子中每个单词加权得到结果。卷积操作只能建立卷积核大小区域内的特征关系, 而自注意力机制可以捕获远程特征之间的关系。非局部注意力模块[21] 是自注意力机制在卷积神经网络中的应用。假设输入特征图表示为 $\boldsymbol{x}=(\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_{h\times w})\in\mathbf{R}^{h\times w}$, $\boldsymbol{x}_i$为特征图的特定空间位置特征值, 特征点总数 $N=h\times w$, $i=1,2,\cdots,N$。非局部注意力模块的计算可以表示为

$$\boldsymbol{y}_i=\boldsymbol{x}_i+\sum_{j=1}^{N}\frac{S(\boldsymbol{x}_i,\boldsymbol{x}_j)}{C(\boldsymbol{x})}(\boldsymbol{W}_{\text{value}}\boldsymbol{x}_j) \tag{2}$$

式中: $\boldsymbol{x}_j$为特征图的特定空间位置特征值, $j=1,2,\cdots,N$; $S(\cdot)$为相似度计算函数, 用于计算各空间位置之间的相似度; $C(\cdot)$为权重归一化函数; $\boldsymbol{W}_{\text{value}}$为特征映射矩阵。

$S(\boldsymbol{x}_i,\boldsymbol{x}_j)$可以表示为

$$S(\boldsymbol{x}_i,\boldsymbol{x}_j)=\mathrm{e}^{(\boldsymbol{W}_{\text{query}}\boldsymbol{x}_i)^{\mathrm{T}}(\boldsymbol{W}_{\text{key}}\boldsymbol{x}_j)} \tag{3}$$

式中: $\boldsymbol{W}_{\text{query}}$、$\boldsymbol{W}_{\text{key}}$为特征映射矩阵。

$C(\boldsymbol{x})$可以表示为

$$C(\boldsymbol{x})=\sum_{m=1}^{N}S(\boldsymbol{x}_i,\boldsymbol{x}_m) \tag{4}$$

式中: $\boldsymbol{x}_m$为特征图的特定空间位置特征值, $m=1,2,\cdots,N$。

全局上下文网络(global context network, GCNet)通过共享一个全局注意映射系数对自注意力机制进行简化计算[22], 使用一个全局线性映射$\mathrm{e}^{\boldsymbol{W}_{\text{key}}\boldsymbol{x}_j}$也能实现相同的相似度计算效果, 即将$S(\boldsymbol{x}_i,\boldsymbol{x}_j)$简化为$S(\boldsymbol{x}_i,\boldsymbol{x}_j)=\mathrm{e}^{\boldsymbol{W}_{\text{key}}\boldsymbol{x}_j}$, 简化后的自注意力机制可以表示为

$$\boldsymbol{y}_i=\boldsymbol{x}_i+\boldsymbol{W}_{\text{value}}\sum_{j=1}^{N}\frac{\mathrm{e}^{\boldsymbol{W}_{\text{key}}\boldsymbol{x}_j}}{C(\boldsymbol{x})}\boldsymbol{x}_j \tag{5}$$

此时, $C(\boldsymbol{x})=\sum_{m=1}^{N}\mathrm{e}^{\boldsymbol{W}_{\text{key}}\boldsymbol{x}_m}$, 对应简化的非局部网络结构如图 2 所示, $\mathrm{e}^{\boldsymbol{W}_{\text{key}}\boldsymbol{x}_j}/C(\boldsymbol{x})$对应 Softmax 函数。

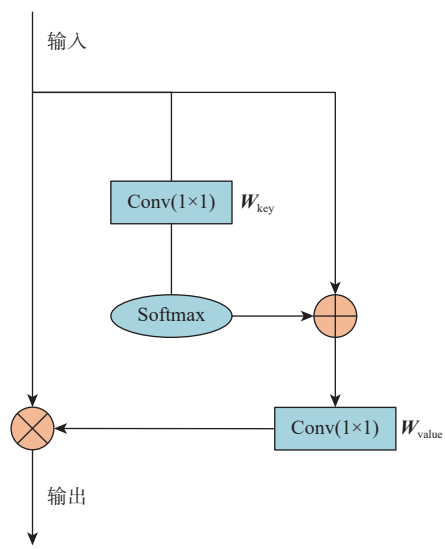


图 2　简化的非局部结构

Fig. 2　Simplified structure of non-local module

在上述基础上进一步简化注意力模块。神经网络中的池化计算是简化特征的直接操作, 但简单的全局池化计算会丢失特征的位置信息。考虑特征空间结构对于目标定位较为敏感, 本文设计了水平和垂直方向的条形池化将空间注意力计算分解为 2 个并行的一维特征编码过程。设计水平条形池化核σ^h高为h、宽为 1, 垂直条形池化核σ^w高为 1、宽为w, 分别沿着水平方向和垂直方向对通道进行聚合编码, 生成一对具有方向感知能力的特征矩

阵$\boldsymbol{F}^h \in \mathbf{R}^{c\times 1\times w}$和$\boldsymbol{F}^w \in \mathbf{R}^{c\times h\times 1}$。

假定输入特征为$\boldsymbol{X} \in \mathbf{R}^{c\times h\times w}$，则$\boldsymbol{X}$中第$c$通道$(i,j)$位置的特征点$\boldsymbol{X}(c,i,j)$经过水平条形池化后的输出可以表示为

$$\boldsymbol{F}^h(c,i,j)=\frac{1}{h}\sum_{j=1}^{h}\boldsymbol{X}(c,i,j) \tag{6}$$

经过垂直条形池化后的输出可以表示为

$$\boldsymbol{F}^w(c,i,j)=\frac{1}{w}\sum_{i=1}^{w}\boldsymbol{X}(c,i,j) \tag{7}$$

将生成的特征矩阵进行拼接，以便进行并行处理，处理后得到拼接，矩阵$\boldsymbol{F}_1=\xi(\boldsymbol{F}^h,\boldsymbol{F}^w)\in \mathbf{R}^{c\times 1\times(h+w)}$，定义$\xi(\cdot,\cdot)$为矩阵拼接操作。

将$\boldsymbol{F}_1$代替$\boldsymbol{W}_{\text{value}}$以减少网络模型参数量，计算特征空间的自注意力权重系数矩阵$\boldsymbol{F}_2 \in \mathbf{R}^{c\times 1\times(h+w)}$。根据式(5)，自注意力权重系数矩阵$\boldsymbol{F}_2$可以表示为

$$\boldsymbol{F}_2(c,1,j)=\boldsymbol{F}_1\sum_{j=1}^{h+w}\frac{\mathrm{e}^{\boldsymbol{W}_{\text{key}}\boldsymbol{F}_1(c,1,j)}}{\sum_{m=1}^{h+w}\mathrm{e}^{\boldsymbol{W}_{\text{key}}\boldsymbol{F}_1(c,1,m)}}\boldsymbol{F}_1(c,1,j) \tag{8}$$

将自注意力权重系数矩阵$\boldsymbol{F}_2$沿着$\boldsymbol{F}_1$拼接方向分离为$\boldsymbol{F}_2^h \in \mathbf{R}^{c\times 1\times w}$和$\boldsymbol{F}_2^w \in \mathbf{R}^{c\times h\times 1}$，再将$\boldsymbol{F}_2^h$和$\boldsymbol{F}_2^w$分别沿垂直方向和水平方向进行复制扩张，得到与原始特征维度相同的矩阵$\boldsymbol{F}_3 \in \mathbf{R}^{c\times h\times w}$和$\boldsymbol{F}_4 \in \mathbf{R}^{c\times h\times w}$，再与原始特征加权，最终使用残差连接避免训练中出现梯度爆炸和梯度消失现象，其输出$\boldsymbol{Y} \in \mathbf{R}^{c\times h\times w}$可以表示为

$$\boldsymbol{Y}=\boldsymbol{X}+\boldsymbol{X}\odot\boldsymbol{F}_3\odot\boldsymbol{F}_4 \tag{9}$$

式中：“$\odot$”表示矩阵点乘。最终得到本文提出的SCAM，图3为SCAM结构。如图4所示，本文可视化了使用SCAM前后网络模型对于图像关注程度的热力图，其中，红色区域为网络模型的重点关注区域，使用SCAM模块后网络更关注目标本身，同时抑制了干扰区域的信息。

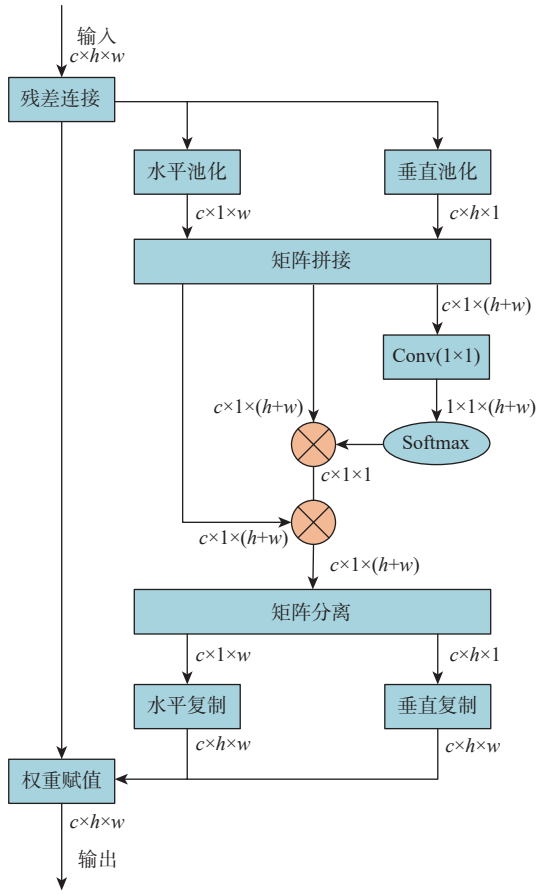
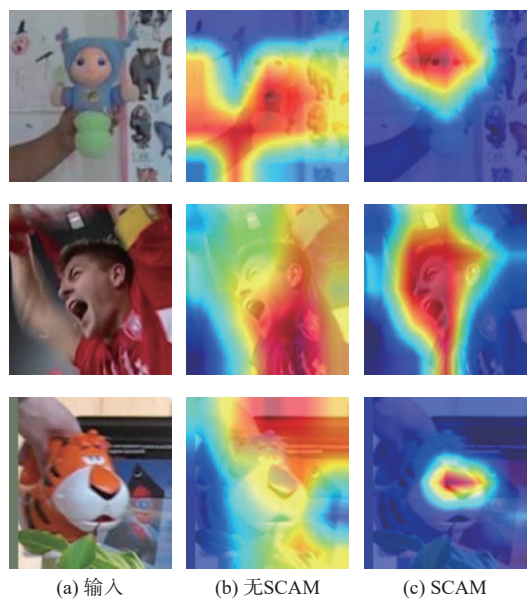


图3 SCAM结构

Fig. 3 Structure of SCAM



(a) 输入 (b) 无SCAM (c) SCAM

图4 热力图可视化

Fig. 4 Visualization of heatmap

1.3 模板更新策略

早期的跟踪器在跟踪过程中仅以跟踪响应图的峰值变化来决定跟踪响应图的质量，基于循环特征映射的大边缘目标跟踪[23](large margin object tracking with circulant feature maps, LMCF)算法使用响应峰值的变化趋势和平均峰值相关能量(average peak-to-correlation energy, APCE)来共同判断跟踪响应图质量，并作为模板更新的标准，取得了较好的跟踪性能提升。图5为在OTB100数据集中的Basketball序列上，SiamFC跟踪响应图峰值$\max(\boldsymbol{Q})$和APCE值ϖ经过标准化后的变化。可以看出，ϖ的变化趋势与响应峰值的变化趋势极其相似。当响应峰值呈上升趋势时，ϖ往往也为上升趋势，两者共同决定跟踪响应图质量，相当于仍使用单一峰值的变化作为依据。ϖ的计算公式为

$$\varpi = \frac{|\max(\boldsymbol{Q}) - \min(\boldsymbol{Q})|^2}{\mathrm{mean}\left(\sum_{i=1}^{h\times w}(\boldsymbol{Q}(i) - \min(\boldsymbol{Q}))^2\right)} \tag{10}$$

式中: $\boldsymbol{Q} \in \mathbf{R}^{h\times w}$为跟踪响应图矩阵; $\boldsymbol{Q}(i)$为特定位置的响应值; max(·)为求最大值函数; min(·)为求最小值函数; mean(·)求平均值函数。

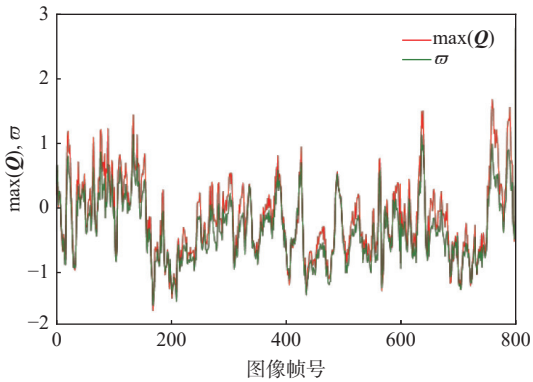
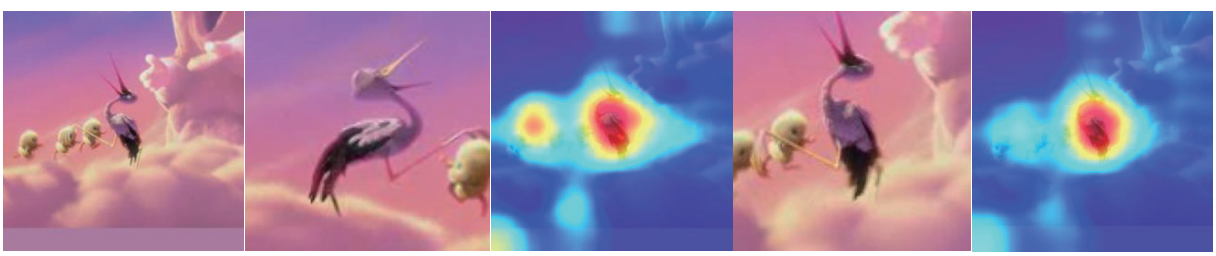


图 5　SiamFC 上响应峰值和 APCE 值的变化

Fig. 5　Variation of response peak value and APCE value on SiamFC

如图 6 所示, 当跟踪场景中出现相似物体时, 仅使用首帧目标作为模板的孪生网络跟踪器, 跟踪响应图呈现多峰响应的情况, 观察发现, 跟踪过程中第 1 次出现多峰响应时目标一般尚未丢失, 此时更新模板可以避免后续出现相似物体干扰影响跟踪器, 因此, 本文使用连通域作为判断标准, 设计一个新的模板更新策略。

假定跟踪序列的第 1 帧图像中跟踪目标的中心坐标为 P_1,跟踪目标宽为 w_1, 高为 h_1。以 P_1为中心裁剪出宽为 $w_1+(w_1+h_1)/2$、高为 $h_1+(w_1+h_1)/2$ 的图像, 再放缩为 127像素×127像素大小, 即为首帧模板 M_1。在跟踪当前帧 P_t时, 以上一帧预测目标中心 P_{t-1}为裁剪中心, 裁剪出 255像素×255像素大小的图像为搜索图像 z。当裁剪区域不足时, 使用红绿蓝均值填充的方式对图像进行填充。



(a) 搜索图像　(b) 首帧模板　(c) 首帧模板跟踪响应图　(d) 第2模板　(e) 双模板跟踪响应图

图 6　不同模板的跟踪响应图

Fig. 6　Tracking response map of different templates

复制首帧模板 $\boldsymbol{M}_1$作为第 2 模板 $\boldsymbol{M}_t$, 对模板 $\boldsymbol{M}_1$、模板 $\boldsymbol{M}_t$和搜索图像 $\boldsymbol{q}$进行特征提取, 分别得到深度特征 $\phi(\boldsymbol{M}_1)$、$\phi(\boldsymbol{M}_t)$和 $\phi(\boldsymbol{q})$。将 $\phi(\boldsymbol{M}_1)$与 $\phi(\boldsymbol{q})$进行互相关运算, 得到跟踪响应图 $f(\boldsymbol{q},\boldsymbol{M}_1)$, 将 $\phi(\boldsymbol{M}_t)$与 $\phi(\boldsymbol{q})$进行互相关运算, 得到跟踪响应图 $f(\boldsymbol{q},\boldsymbol{M}_t)$。融合 2 个响应图得最终跟踪响应图 $f(\boldsymbol{q};\boldsymbol{M}_1,\boldsymbol{M}_t)$, 可以表示为

$$f(\boldsymbol{q};\boldsymbol{M}_1,\boldsymbol{M}_t) = \alpha f(\boldsymbol{q},\boldsymbol{M}_1) + (1-\alpha) f(\boldsymbol{q},\boldsymbol{M}_t) \tag{11}$$

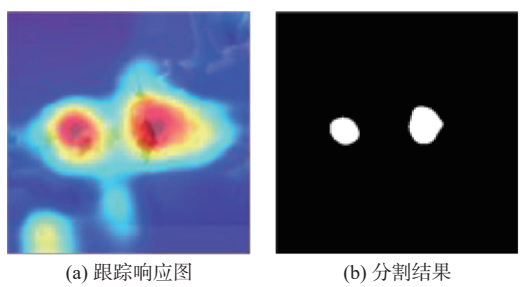
式中: α为融合参数。

跟踪响应图分割阈值 T表示为

$$T = (\max(f(\boldsymbol{q};\boldsymbol{M}_1,\boldsymbol{M}_t)) - \min(f(\boldsymbol{q};\boldsymbol{M}_1,\boldsymbol{M}_t)))\gamma \tag{12}$$

式中: γ为阈值参数。

图 7 为跟踪响应图分割结果。如图 7(a)所示, 跟踪响应图红色区域为跟踪器重点关注的区域, 对分割结果进行实验观察, 为分割出跟踪响应图红色区域所在的连通域图像, 最终确定在 OTB100 数据集上取分割阈值参数 γ为 0.85。随后继续进行实验确定响应图融合参数 α的取值, 如表 1 所示,



(a) 跟踪响应图　(b) 分割结果

图 7　跟踪响应图分割结果

Fig. 7　Segmentation results of tracking response map

最终取 α为 0.55。

将跟踪响应图灰度化得到响应灰度图 $\boldsymbol{G}$, 对 $\boldsymbol{G}$进行阈值处理, 使得

$$\boldsymbol{G}(i,j) = \begin{cases} 255 & \boldsymbol{G}(i,j) \geqslant T \\ 0 & \boldsymbol{G}(i,j) < T \end{cases} \tag{13}$$

图 7(b)为阈值分割处理结果, 计算分割结果中的连通域数目, 当其大于 1 时, 代表首帧模板难以应对当前跟踪场景中的干扰问题, 此时以 P_{t-1}为裁剪中心, 裁剪出宽为 $w_1+(w_1+h_1)/2$、高为 $h_1+(w_1+h_1)/2$

表1 在OTB100数据集上α取不同值时的实验结果

Table 1 Experimental result of different α values on OTB100 dataset

α	成功率	精度
0.50	0.655	0.870
0.55	0.669	0.890
0.60	0.651	0.869
0.65	0.658	0.878
0.70	0.659	0.876
0.75	0.660	0.870
0.80	0.655	0.875
0.85	0.659	0.875
0.90	0.659	0.880
0.95	0.653	0.866

大小的图像，再放缩为127像素×127像素大小，并将放缩后的图像作为第2模板$\boldsymbol{M}_t$，进行后续跟踪。当未出现多峰响应时，需要将峰值点在特征上的位置坐标乘以卷积步长即为目标在搜索图像上的具体位置。

2 实验分析

为验证本文算法的有效性，在Ubuntu18.04操作系统下采用Pytorch1.7.0框架编程实现了本文算法，并在短时跟踪数据集OTB100和VOT2018、长时跟踪数据集LaSOT、无人机数据集UAV123上与多个先进的跟踪器进行性能比较。实验所使用的主要硬件环境为：Intel 6136处理器，NVIDIA RTX 6000显卡和512 GB内存。

在网络训练阶段，本文选取GOT-10k作为训练数据集，数据集含有66 GB图像序列，包含了563种目标类别，训练过程中的网络优化参数设置与数据增强策略均与SiamFC[13]相同。对于响应图融合参数α和分割阈值参数γ，OTB100数据集上分别设置为0.55和0.85，VOT2018数据集上分别设置为0.65和0.8，LaSOT数据集上分别设置为0.6和0.8，UAV123数据集上均设置为0.95。

2.1 OTB100实验

OTB100数据集是目标跟踪算法研究中具有代表性的数据集，包含100个视频序列，11种挑战属性，分别为背景干扰、形变、快速运动、超出视野、图像外旋转、光照变化、运动模糊、遮挡、尺度变化、低分辨率和图像内旋转，每个视频包含一个或多个挑战属性。

为展示本文算法的性能，选取13个主流跟踪算法在OTB100数据集上进行对比实验，分别为：①基于手工特征的相关滤波算法。快速判别尺度跟踪器[24](fast discriminative scale space tracker, FDSST)算法、模板与像素级学习和(sum of template and pixel-wise learners, Staple)算法[25]、时空正则化判别式相关滤波器[26](spatially regularized discriminative correlation filters, SRDCF)、使用方向梯度直方图(histogram of gradient, HOG)特征和颜色特征的高效卷积算子[27](efficient convolution operators with HOG and CN, ECO-HC)。②基于深度特征的相关滤波算法。使用深度特征的时空正则化判别式相关滤波器[28](spatially regularized discriminative correlation filters with deep features, DeepSRDCF)。③基于深度特征的孪生网络类跟踪算法。SiamFC[13]、更深更宽的孪生网络[29](deeper and wider Siamese networks, SiamDW)、孪生区域候选网络[30](Siamese region proposal network, SiamRPN)、使用18层深度残差网络的判别模型预测(discriminative model prediction with an eighteen-layer ResNet network, DiMP-18)算法[31]、DaSiamRPN[32]、使用18层深度残差网络的概率回归判别模型预测(probabilistic discriminative model prediction with an eighteen-layer ResNet network, PrDiMP-18)算法[33]、通过重叠最大化实现精确跟踪(accurate tracking by overlap maximization, ATOM)算法[34]和梯度引导网络(gradient-guided network, GradNet)算法[35]。

2.1.1 定性分析

图8为OTB100数据集上本文算法与其他4种孪生网络跟踪器在5个复杂视频序列场景下的部分跟踪结果。为展示本文算法的优势，从以下5个方面进行定性分析：①背景干扰。在背景干扰场景下，基于模板匹配思想的孪生网络跟踪器容易跟踪到相似物体上，主要有2个原因。一个原因是深度特征的建模不足，对于相似物体的特征表达较为相似，使得模板匹配时在多个位置有着较高的响应峰值导致定位失败；另一个原因是缺乏模板更新，目标在运动过程中不断变化，缺乏模板更新的跟踪器无法捕捉到目标变化后的特征导致跟踪失败。以Bolt2序列为例，Bolt2序列中存在多个相似跟踪目标，只有本文算法和带有模板更新的DiMP-18算法能够实现稳定跟踪，使用深层网络的SiamDW算法虽然能够提取到足够深度的特征，但由于不能及时更新使得在相似物体接近时丢失目标。②光照变化。光照的剧烈变化往往使得目标外观发生较大变化，如Ironman视频序列的跟踪场景较暗，且伴随背景光线的剧烈变化，目标与背景难以区分。由于本文算法不但能够提取到判别度较高的特征，且能

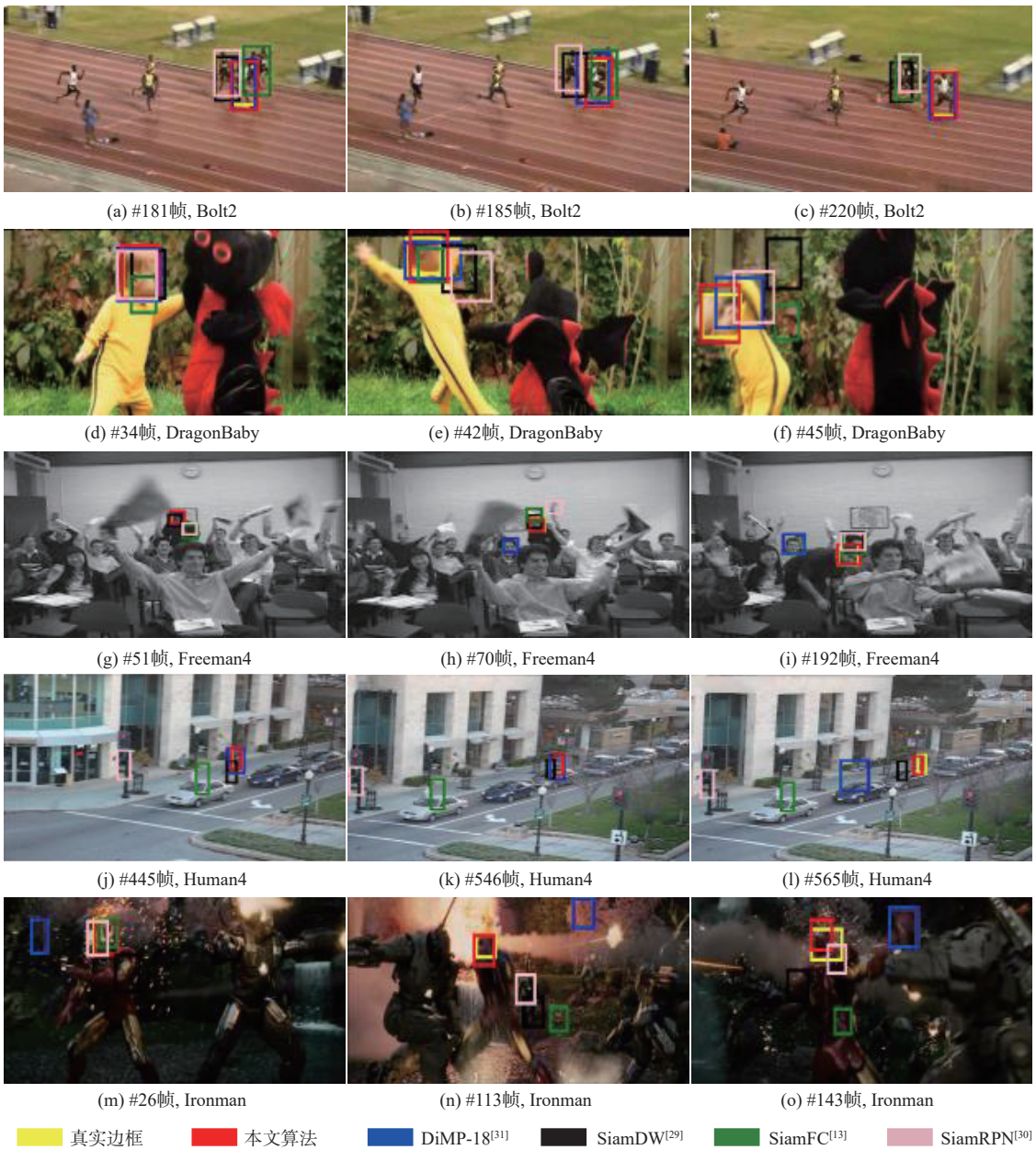


图 8　OTB100 数据集上部分跟踪结果比较

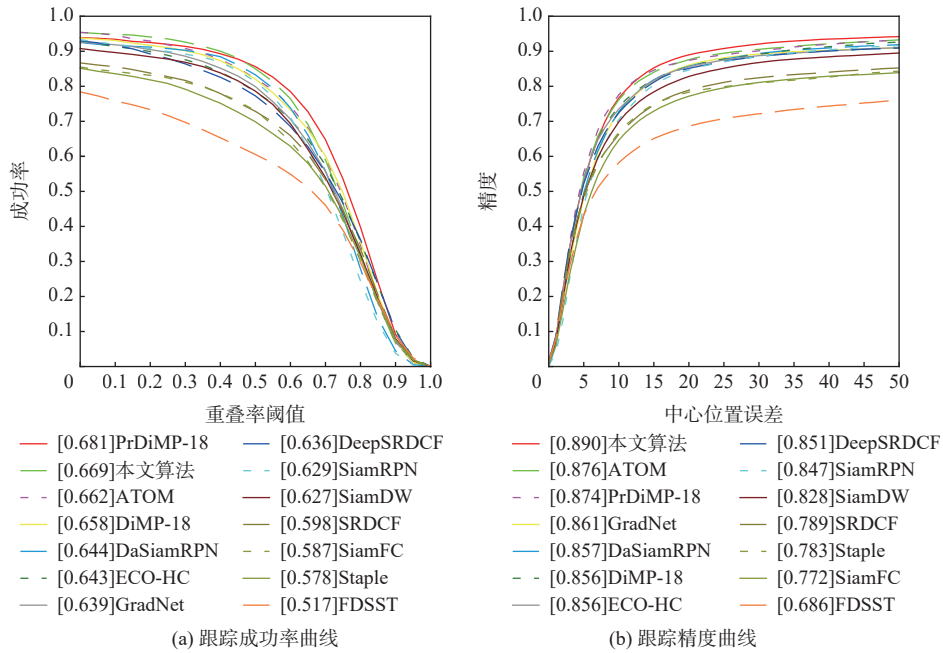
Fig. 8　Comparison of some tracking results

实现重点区域特征的关注，使得在#113 帧中只有本文算法能够稳定跟踪目标，可得本文算法优于仅使用深度网络的 SiamDW 算法和 DiMP-18 算法。③运动模糊。运动模糊主要是由于目标短时间内进行较大位移造成，所选视频序列均有不同程度的运动模糊现象出现，多数情况下，SiamFC 和 SiamRPN 在此场景下是难以跟上目标的，这是由于 AlexNet 提取的特征鲁棒性不够，经过本文对 SiamFC 的改进使得算法能够快速跟踪到目标。④目标旋转。目标在跟踪过程容易出现图像内旋转和图像外旋转 2 种情况，旋转过程中往往会使目标原有信息发生变化，要求跟踪器所提取的特征具有高度的旋转不变鲁棒性。例如，在 DragonBaby 序列中，目标在跳跃中出现剧烈旋转，尽管 SiamRPN 算法配备了高效的尺度变化模块，但是由于深度信息的缺失，使得 SiamRPN 算法在跟踪过程中出现目标漂移，而本文算法能够实现稳定跟踪。⑤遮挡。遮挡分为部分遮挡和完全遮挡 2 种情况，在遮挡情况下也易出现多峰响应的情况，如在 Freeman4 和 Human4 序列的跟踪中，所对比的多数跟踪器在连续遮挡情况下丢失目标，但是由于本文使用了双模板更新的算法，视频第 1 帧的模板与遮挡情况下所更新的模板联合定位目标，使得本文算法在遮挡情况下也能拥有较好的跟踪结果。

2.1.2　定量分析

图 9 为本文算法与对比算法在 OTB100 数据集

上的跟踪成功率和跟踪精度的对比结果。图 9(a)图例中的数值表示成功率曲线下方的面积,可以更好地评价跟踪效果,图 9(b)图例中数值表示中心位置误差阈值取 20 像素时各跟踪器的精度。相对于基准算法 SiamFC,本文算法的跟踪成功率提高了 0.082,跟踪精度提高了 0.118,相比于使用更深层网络与复杂模板更新策略的 PrDiMP-18、DiMP-18 和 ATOM 算法,本文算法在跟踪精度方面具有明显优势。



(a) 跟踪成功率曲线 (b) 跟踪精度曲线

图 9 不同算法在 OTB100 数据集上的评估结果

Fig. 9 Evaluation results of different algorithms on OTB100 dataset

跟踪速度影响跟踪的实际应用,图 10 为各算法在 OTB100 数据集上的速度对比实验。其中,Baseline 为基准算法 SiamFC,其跟踪速度为 101 帧/s,跟踪成功率为 0.587;Offline 为不带双模板更新时本文算法所达到的跟踪性能,跟踪速度为 79 帧/s,跟踪成功率为 0.657;Online 指带有模板更新策略时本文算法所达到的跟踪性能,跟踪速度为 70 帧/s,跟踪成功率为 0.669。虽然相较于基准算法,本文算法跟踪速度稍微下降,但是跟踪性能却得到了大幅提升,且能够达到实时跟踪的需求。

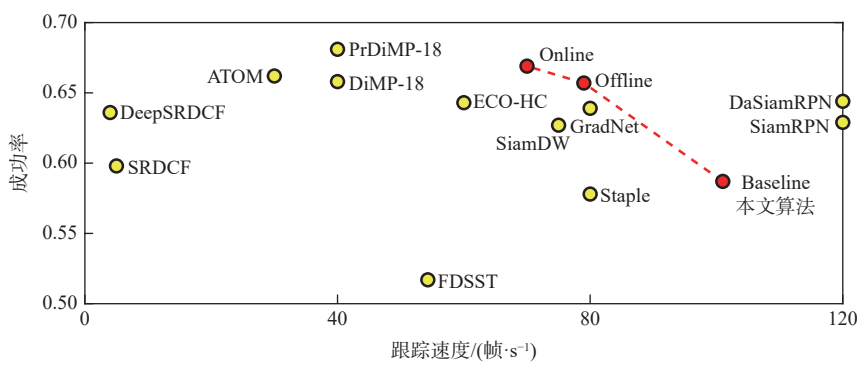


图 10 不同算法在 OTB100 数据集上的速度对比

Fig. 10 Speed comparison of different algorithms on OTB100 different

为验证本文算法在各跟踪属性上的性能,图 11 为 OTB100 数据集上各种挑战属性的跟踪成功率结果。本文算法在绝大多数属性上与同类算法相比具有竞争性,且在背景干扰、光照变化和遮挡属性上跟踪性能最优,说明本文算法能够大幅度改善孪生网络跟踪器性能,满足跟踪器在复杂场景下的跟踪需求。

2.2 VOT2018 实验

VOT2018 数据集是 2018 年单目标跟踪挑战赛所使用的比赛数据集,共有 60 个视频序列,采用准确率、鲁棒性和期望平均重叠率 EAO 来评估算法性能。其中,准确率指跟踪器在测试序列下的平均

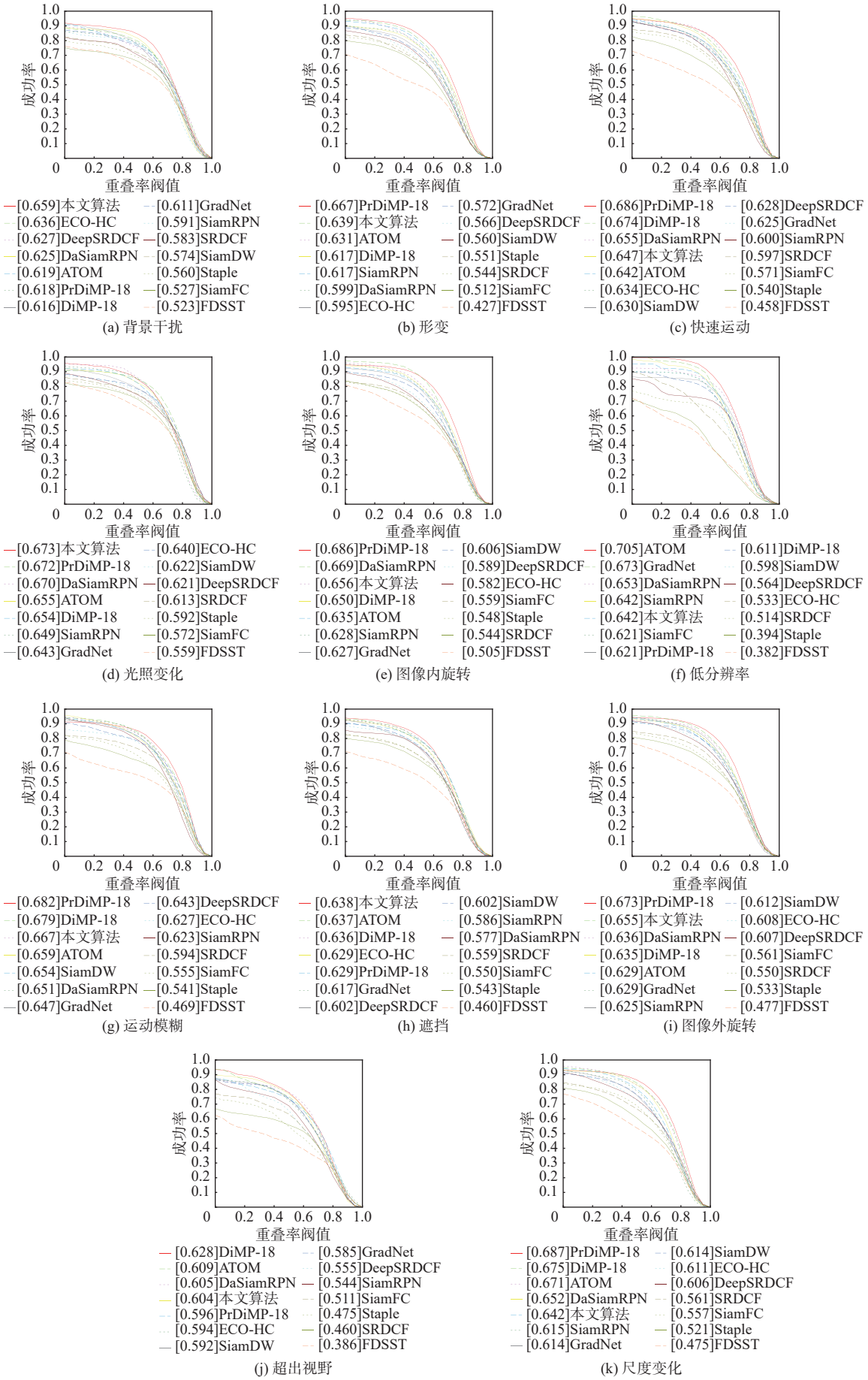


图 11　OTB100 数据集 上不同挑战属性的跟踪成功率

Fig. 11　Tracking success rate of different challenge attributes on OTB100 dateset

IoU重叠率;鲁棒性指跟踪器在测试序列下的跟踪失败次数除以跟踪序列数,预测目标区域与真实目标区域的IoU为0时判定为失败。准确率和EAO分数越高,鲁棒性越低,代表跟踪器性能越好,EAO为衡量算法性能的主要评价标准。实验对比算法有SiamFC[13]、动态孪生网络[36](dynamic Siamese network, DSiam)、UpdateNet算法[18]、连续空间域卷积跟踪(continuous convolution operators for visual tracking, CCOT)算法[37]、ECO[27]、基于VGG网络的孪生网络(Siamese network based on VGG network, SiamVGG)算法[38]、基于深度特征的具有信道和空间可靠性的判别相关滤波器[39](discriminative correlation filter with channel and spatial reliability based on deep features, DeepCSRDCF)。如表2所示,本文算法仅在鲁棒性方面低于ECO和DeepCSRDCF算法,在准确率和EAO方面均高于对比算法,与基准算法相比,本文算法的EAO提高了0.134,且优于同样使用VGG网络的跟踪器SiamVGG。

表2 不同算法在VOT2018数据集上的实验结果

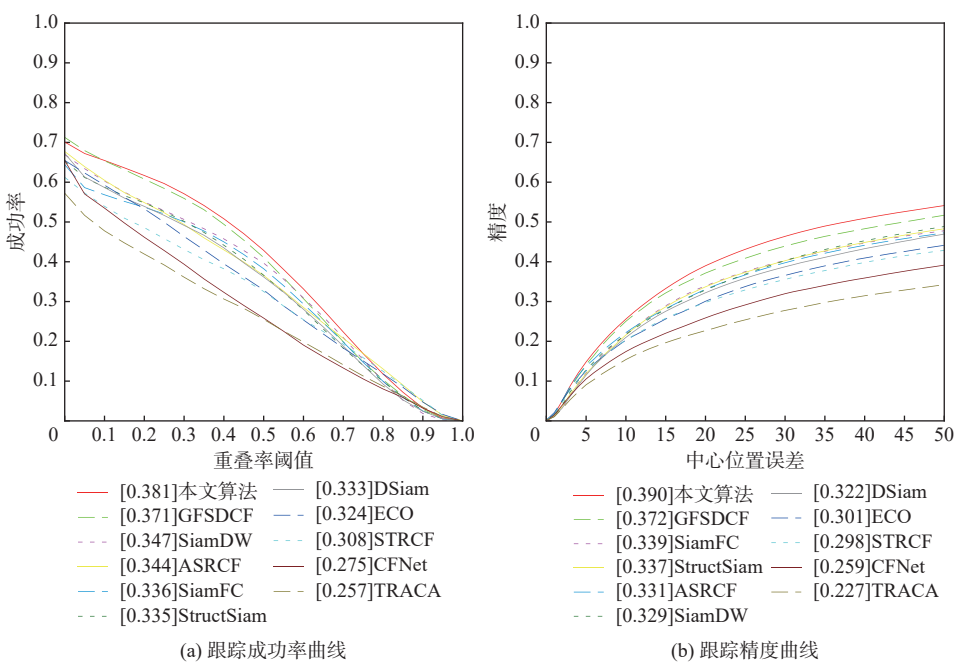
Table 2 Experimental results on of different algonithms VOT2018 dataset

算法	准确率	鲁棒性	EAO
SiamFC[13]	0.503	0.585	0.188
DSiam[36]	0.512	0.646	0.196
UpdateNet[18]	0.518	0.454	0.244
CCOT[37]	0.494	0.318	0.267
ECO[27]	0.484	0.276	0.280
SiamVGG[38]	0.531	0.318	0.286
DeepCSRDCF[39]	0.489	0.276	0.293
本文算法	0.532	0.309	0.322

2.3 LaSOT实验

LaSOT数据集是目标跟踪领域上的长时跟踪数据集,包含1 400个视频序列,而且每个视频平均有2 512帧,最长的视频序列高达11 397帧,包含14种跟踪挑战属性。在该数据集上的测试结果一定程度上反映了跟踪器长时间跟踪的性能,所采用的评价标准与OTB100数据集相同,均为跟踪精度和跟踪成功率。对比算法包括联合群特征选择与判别滤波器(joint group feature selection and discriminative filter learning, GFSDCF)算法[40]、SiamDW[29]、自适应空间正则相关滤波器(adaptive spatially-regularized correlation filters, ASRCF)算法[41]、SiamFC[13]、结构化孪生网络(structured Siamese network, StructSiam)算法[42]、Dsiam[36]、ECO[27]、时空正则相关滤波器(spatial-temporal regularized correlation filters, STRCF)算法[43]、CFNet[17]、基于上下文感知的多自动编码器深度特征压缩跟踪器(tracker based on context-aware deep feature compression with multiple autoencoders, TRACA)算法[44],整体性能对比结果如图12所示。相较于基准算法,本文算法的跟踪成功率提高了0.045,跟踪精度提高了0.051。

图13为各算法在LaSOT数据集上不同挑战属性跟踪成功率对比。相较于基准算法,本文算法在各属性的跟踪性能均有明显提升,但仍有部分属性



(a) 跟踪成功率曲线 (b) 跟踪精度曲线

图12 不同算法在LaSOT数据集上的评估结果

Fig. 12 Evaluation results of different algorithms on LaSOT dataset

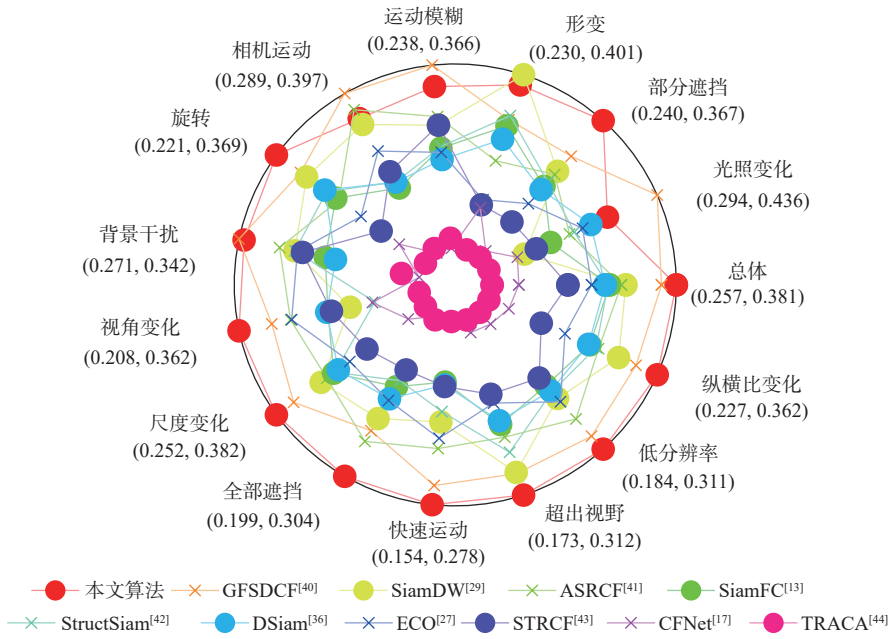


图 13　LaSOT 数据集上不同挑战属性跟踪成功率对比

Fig. 13　Comparison of tacking success rates of different challenge attributes on LaSOT dataset

上的跟踪性能不如相关滤波类算法 GFSDCF 和 ASRCF, 由于长时跟踪下, 滤波器在这些场景中能够学习到更好的模板更新策略。

2.4　UAV123 实验

UAV123 数据集是目标跟踪领域中的无人机跟踪数据集, 目标视角变化较多, 共包含 123 个视频序列, 均为无人机飞行状态下采集的视频序列。在此数据集下的测试结果能够反映出算法在无人机场景下的跟踪性能, 一定程度上推动了算法在无人机方向上的实际应用。所采用的评价标准与 OTB100 数据集相同, 均为跟踪精度和跟踪成功率。对比结果如图 14 所示, 本文算法的跟踪成功率和跟踪精度分别为 0.545 和 0.790。相较于基准算法, 本文算法的跟踪成功率提高了 0.045, 跟踪精度提高了 0.062。

2.5　消融实验

为分析算法各组件的有效性, 将算法进行拆解后与基准算法在 OTB100 数据集上进行比较, 共设计 4 组实验, 分别为: ①SiamFC, 基准算法; ②SiamFC-VGG(SiamFC-V), 使用改进过的 VGG 网络作为 CNN 的 SiamFC 跟踪器; ③SiamFC-VGG-Attention(SiamFC-V-A), 添加 SCAM 的 SiamFC-VGG 跟踪器; ④SiamFC-VGG-Attention-Update(SiamFC-V-A-U), 在实验③基础上使用本文模板更新策略的跟踪器。消融实验结果如表 3 所示, 所有组件均不同程度地提高了跟踪器的跟踪性能。

为展示 SCAM 的具体性能, 本文将添加 SCAM 的跟踪器与添加非局部模块的跟踪器在 OTB100

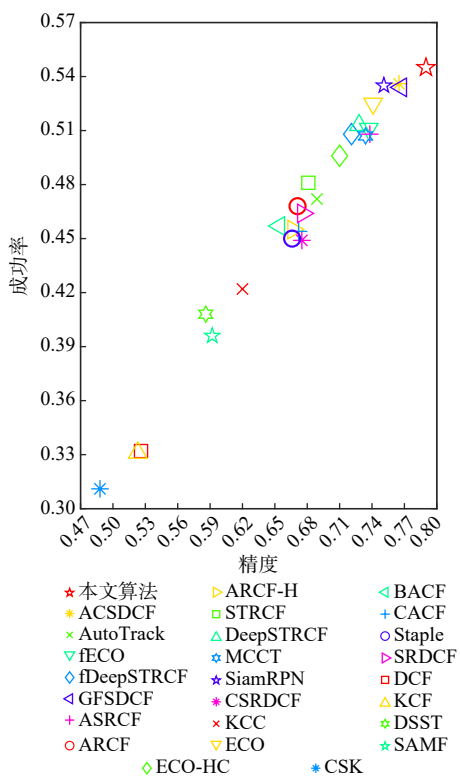


图 14　不同算法在 UAV123 数据集上的评估结果

Fig. 14　Evaluation results of different algorithms on UAV123 dataset

数据集上进行对比, 把 SiamFC-V-A 实验分为 2 组: ①SiamFC-VGG-SCAM, 在 SiamFC-VGG 基础上在模板分支上添加了 SCAM 的跟踪器; ②SiamFC-VGG-Non-local, 在 SiamFC-VGG 基础上在模板分支上添加了非局部模块的跟踪器。图 15 为 2 个跟踪器分别训练 50 轮次, 各轮次在 OTB100 数据集上的跟踪

表3 消融实验结果

Table 3 result of ablation study

实验	成功率	精度
SiamFC	0.587	0.772
SiamFC-V	0.624	0.833
SiamFC-V-A	0.657	0.875
SiamFC-V-A-U	0.669	0.890

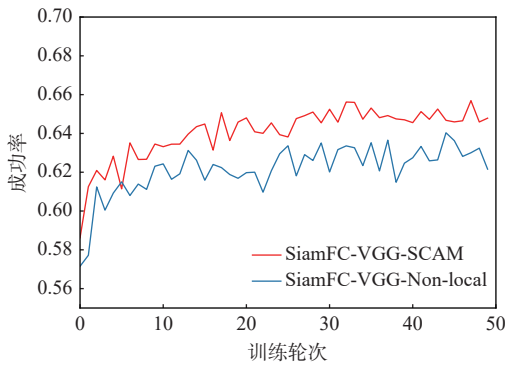


图15 SCAM与非局部模块对比实验结果

Fig. 15 Comparative experimental results of SCAM and non-local modules

器结果。实验结果表明,本文提出的SCAM在孪生网络跟踪器上的应用优势高于非局部模块。

3 结 论

1)本文提出一种轻量级的单卷积注意力模块,用于进一步关注重点区域的深度特征,抑制干扰物体所在区域的表达。

2)本文提出一种利用连通域作为判断标准的模板更新策略,对第2模板进行自适应更新,缓解多峰响应带来的目标漂移问题。

3)将本文算法网络模型在GOT-10k数据集上进行训练,在OTB100、VOT2018、LaSOT和UAV123数据集上进行性能测试,本文算法在背景干扰、光照变化、运动模糊、目标旋转和遮挡复杂场景下具有显著优势。

后续将考虑继续挖掘特征在通道上的关联,去除冗余特征对跟踪器的干扰,同时设计更加完善的模板更新机制,增强跟踪器在长时跟踪和无人机等复杂场景下的鲁棒跟踪。

参考文献(References)

[1] GAO M, JIN L S, JIANG Y Y, et al. Manifold Siamese network: A novel visual tracking ConvNet for autonomous vehicles[J]. IEEE Transactions on Intelligent Transportation Systems, 2020, 21(4): 1612-1623.

[2] 寇展, 吴健发, 王宏伦, 等. 基于深度学习的低空小型无人机障碍物视觉感知[J]. 中国科学:信息科学, 2020, 50(5): 692-703.
KOU Z, WU J F, WANG H L, et al. Obstacle visual sensing based on deep learning for low-altitude small unmanned aerial vehicles[J]. Scientia Sinica (Informationis), 2020, 50(5): 692-703(in Chinese).

[3] MARVASTI-ZADEH S M, CHENG L, GHANEI-YAKHDAN H, et al. Deep learning for visual tracking: A comprehensive survey[J]. IEEE Transactions on Intelligent Transportation Systems, 2022, 23(5): 3943-3968.

[4] LUO J H, HAN Y, FAN L Y. Underwater acoustic target tracking: A review[J]. Sensors, 2018, 18(2): 112.

[5] 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).

[6] 韩明, 王景芹, 王敬涛, 等. 基于孪生网络的目标跟踪研究综述[J]. 河北科技大学学报, 2022, 43(1): 27-41.
HAN M, WANG J Q, WANG J T, et al. Comprehensive survey on target tracking based on Siamese network[J]. Journal of Hebei University of Science and Technology, 2022, 43(1): 27-41(in Chinese).

[7] HENRIQUES J F, CASEIRO R, MARTINS P, et al. High-speed tracking with kernelized correlation filters[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015, 37(3): 583-596.

[8] ZHANG K H, ZHANG L, LIU Q S, et al. Fast visual tracking via dense spatio-temporal context learning[C]//Proceedings of the European Conference on Computer Vision. Berlin: Springer, 2014: 127-141.

[9] 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.

[10] 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, 2016: 3074-3082.

[11] WANG L J, OUYANG W L, WANG X G, et al. Visual tracking with fully convolutional networks[C]//Proceedings of the IEEE International Conference on Computer Vision. Piscataway: IEEE Press, 2016: 3119-3127.

[12] BHAT G, JOHNANDER J, DANELLJAN M, et al. Unveiling the power of deep tracking[C]//Proceedings of the European Conference on Computer Vision. Berlin: Springer, 2018: 493-509.

[13] BERTINETTO L, VALMADRE J, HENRIQUES J F, et al. Fully-convolutional Siamese networks for object tracking[C]//Eurpean Conference on computer Vision. Berlin: Springer, 2016: 850-865.

[14] HE A F, LUO C, TIAN X M, et al. A twofold Siamese network for real-time object tracking[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2018: 4834-4843.

[15] PU L, FENG X X, HOU Z Q, et al. SiamDA: Dual attention Siamese network for real-time visual tracking[J]. Signal Processing:Image Communication, 2021, 95: 116293.

[16] GUPTA D K, ARYA D, GAVVES E. Rotation equivariant Siamese networks for tracking[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2021: 12357-12366.

[17] VALMADRE J, BERTINETTO L, HENRIQUES J, et al. End-to-end representation learning for correlation filter based tracking[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2017: 5000-5008.

[18] ZHANG L C, GONZALEZ-GARCIA A, VAN DE WEIJER J, et al. Learning the model update for Siamese trackers[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE Press, 2020: 4009-4018.

[19] ZHU Z, WU W, ZOU W, et al. End-to-end flow correlation tracking with spatial-temporal attention[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2018: 548-557.

[20] ZHANG C Y, WANG H, WEN J W, et al. Deeper Siamese network with stronger feature representation for visual tracking[J]. IEEE Access, 2020, 8: 119094-119104.

[21] WANG X L, GIRSHICK R, GUPTA A, et al. Non-local neural networks[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2018: 7794-7803.

[22] CAO Y, XU J R, LIN S, et al. GCNet: Non-local networks meet squeeze-excitation networks and beyond[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision . Piscataway: IEEE Press, 2020: 1971-1980.

[23] 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.

[24] 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.

[25] 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.

[26] 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, 2016: 4310-4318.

[27] 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.

[28] DANELLJAN M, HAGER G, KHAN F S, et al. Convolutional features for correlation filter based visual tracking[C]//Proceedings of the IEEE International Conference on Computer Vision . Piscataway: IEEE Press, 2015: 58-66.

[29] ZHANG Z P, PENG H W. Deeper and wider Siamese networks for real-time visual tracking[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2020: 4586-4595.

[30] LI B, YAN J J, WU W, et al. High performance visual tracking with Siamese region proposal network[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2018: 8971-8980.

[31] 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, 2020: 6181-6190.

[32] ZHU Z, WANG Q, LI B, et al. Distractor-aware Siamese networks for visual object tracking[C]//Proceedings of the European Conference on Computer Vision. Berlin: Springer, 2018: 103-119.

[33] DANELLJAN M, VAN GOOL L, TIMOFTE R. Probabilistic regression for visual tracking[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2020: 7181-7190.

[34] 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, 2020: 4655-4664.

[35] LI P X, CHEN B Y, OUYANG W L, et al. GradNet: Gradient-guided network for visual object tracking[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE Press, 2020: 6161-6170.

[36] GUO Q, FENG W, ZHOU C, et al. Learning dynamic Siamese network for visual object tracking[C]//Proceedings of the IEEE International Conference on Computer Vision. Piscataway: IEEE Press, 2017: 1781-1789.

[37] DANELLJAN M, ROBINSON A, KHAN F S, et al. Beyond corre lation filters: Learning continuous convolution operators for visual tracking[C]//Proceedings of the European Conference on Computer Vision. Berlin: Springer, 2016: 472-488.

[38] YIN Z J, WEN C H, HUANG Z Y, et al. SiamVGG-LLC: Visual tracking using LLC and deeper Siamese networks[C]//Proceedings of the IEEE International Conference on Communication Techn ology. Piscataway: IEEE Press, 2020: 1683-1687.

[39] 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.

[40] XU T Y, FENG Z H, WU X J, et al. Joint group feature selection and discriminative filter learning for robust visual object tracking [C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE Press, 2020: 7949-7959.

[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, 2020: 4665-4674.

[42] ZHANG Y H, WANG L J, QI J Q, et al. Structured Siamese network for real-time visual tracking[C]//Proceedings of the Enropean conference on Computer Vision. Berlin: Springer, 2018: 355-370.

[43] 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.

[44] CHOI J, CHANG H J, FISCHER T, et al. Context-aware deep feature compression for high-speed visual tracking[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2018: 479-488.

Real-time robust visual tracking based on spatial attention mechanism

MA Sugang[1, 2, *], ZHANG Zixian[1], PU Lei[3], HOU Zhiqiang[1, 4]

(1. School of Computer Science and Technology, Xi'an University of Posts and Telecommunications, Xi'an 710121, China;
2. School of Information Engineering, Chang' an University, Xi'an 710064, China;
3. School of Operational Support, Rocket Force Engineering University, Xi'an 710025, China;
4. Shaanxi Key Laboratory of Network Data Analysis and Intelligent Processing,
Xi'an University of Posts and Telecommunications, Xi'an 710121, China)

Abstract: A real-time object tracking method coupled with a spatial attention mechanism is suggested in order to enhance the fully convolutional Siamese network (SiamFC) tracker's tracking capability in complex settings and alleviate the target drift problem in the tracking process. The improved visual geometry group (VGG) network is used as the backbone network to enhance the modeling ability of the tracker for the target deep feature. The self-attention mechanism is optimized, and a lightweight single convolution attention module (SCAM) is proposed. The spatial attention is decomposed into two parallel one-dimensional feature coding processes to reduce the computational complexity of spatial attention. The initial target template in the tracking process is retained as the first template, and the second template is dynamically selected by analyzing the variation of the connected domain in the tracking response map. The target is located after fusing the two templates. The experimental results show that, compared with SiamFC, the success rate of the proposed algorithm on OTB100, LaSOT, and UAV123 datasets is increased respectively by 0.082, 0.045, and 0.045, and the tracking accuracy by 0.118,0.051, and 0.062. On the VOT2018 dataset, the proposed algorithm improves the tracking accuracy, robustness, and expected average overlap by 0.029, 0.276, and 0.134, respectively, compared with SiamFC. Real-time tracking requirements can be satisfied by the tracking speed, which can approach 70 frames per second.

Keywords: object tracking; Siamese network; attention mechanism; model update; deep learning

Received: 2022-05-07; **Accepted**: 2022-08-21; **Published Online**: 2022-09-14 11:30

URL: link.cnki.net/urlid/11.2625.V.20220913.1335.001

Foundation items: National Natural Science Foundation of China (62072370); Key Research and Developement program of Shaanxi (2018ZDCXL-GY-04-02); Graduate Innovation Fund of Xi'an University of Posts and Telecommunications (CXJJZL2021011)

* **Corresponding author**. E-mail: msg@xupt.edu.cn

http://bhxb.buaa.edu.cn jbuaa@buaa.edu.cn
DOI: 10.13700/j.bh.1001-5965.2022.0422

基于跨模态近邻损失的可视-红外行人重识别

赵三元[1,2,*]，阿琪[1]，高宇[1]

(1. 北京理工大学 计算机学院，北京 100081；2. 北京理工大学 长三角研究院，嘉兴 314019)

摘 要：可视-红外跨模态行人重识别任务的目标是给定一个模态的特定人员图像，在其他不同模态摄像机所拍摄的图像集中进行检索，找出相同人员对应的图像。由于成像方式不同，不同模态的图像之间存在明显的模态差异。为此，从度量学习的角度出发，对损失函数进行改进以获取具有更加辨别性的信息。对图像特征内聚性进行理论分析，并在此基础上提出一种基于内聚性分析和跨模态近邻损失函数的重识别方法，以加强不同模态样本的内聚性。将跨模态困难样本的相似性度量问题转化为跨模态最近邻样本对和同模态样本对的相似性度量，使得网络对模态内聚性的优化更加高效和稳定。对所提方法在全局特征表示的基线网络和部分特征表示的基线网络上进行实验验证结果表明：所提方法对可视-红外行人重识别的预测结果相较于基线方法，平均准确度最高可提升8.44%，证明了方法在不同网络架构中的通用性；同时，以较小的模型复杂度和较低的计算量为代价，实现了可靠的跨模态行人重识别结果。

关 键 词：可视-红外行人重识别；度量学习；深度学习；跨模态学习；计算机视觉
中图分类号：TP391.4
文献标志码：A **文章编号**：1001-5965(2024)02-0433-09

给定一个感兴趣的查询人图像，行人重识别技术旨在确定此人是否被不同位置的摄像机所捕获[1-3]。行人重识别技术在公共安全，尤其是寻找失踪人员、行人分析、预防和惩治犯罪等众多方面都具有不可替代的作用。随着红外成像技术的发展，智能监控设备加强了夜间成像数据的捕获。因此，行人重识别技术不仅要求能够对可视图像进行检索，更需要在可视和红外图像之间进行行人图像的检索，这项工作被称为可视-红外行人重识别任务。

目前的行人重识别技术中，大多数方法用于解决单模态下的人员检索问题。但在夜间或光照条件较差时，可见光相机无法有效采集人员图像信息。目前，针对可视-红外图像进行跨模态检索的研究仍不充分。单模态行人重识别任务的主要难点在于样本间具有摄像机视角差异[4-5]、光照条件差异[6]、行人姿态差异[7-8]等模态内差异。可视-红外行人重识别在单模态行人重识别难点的基础上，又加入了模态间的差异。例如，可视图像与红外图像的成像原理不同、相机内部参数不同、图像通道数不同，使得这2个模态的图像之间具有显著的模态差异，给可视-红外人员重识别任务增加了难度，具有一定的挑战性。

目前，很多研究工作致力于减小可视-红外模态间特征差异来加强同一身份不同模态样本的内聚性，主要分为2种方法：第1种进行特征对齐，将2种不同模态的图像投影到一个特征空间中，以最大程度减少模态间的差异，加强同类别样本的内聚性；第2种使用生成对抗网络，将需要查询的模态转化成另一模态，以解决模态间的差异，增强样本的内聚性。这2种方法均是以间接的方式解决模态间差异，增加了网络的复杂度，也不易加强不同模态下同类样本的内聚性。跨模态同类样

收稿日期：2022-05-26；**录用日期**：2022-07-02；**网络出版时间**：2022-11-21 10:08
网络出版地址：link.cnki.net/urlid/11.2625.V.20221118.1553.001
基金项目：国家自然科学基金 (61902027)
*通信作者. E-mail：zhaosanyuan@bit.edu.cn

引用格式：赵三元，阿琪，高宇. 基于跨模态近邻损失的可视-红外行人重识别[J]. 北京航空航天大学学报，2024，50(2)：433-441.
ZHAO S Y，A Q，GAO Y. Cross-modality nearest neighbor loss for visible-infrared person re-identification[J]. Journal of Beijing University of Aeronautics and Astronautics，2024，50(2)：433-441 (in Chinese).

本的内聚性可以表示为直接计算所有不同模态样本对之间的距离之和。由于计算内聚性时会涉及到很多困难样本对，直接对其进行优化十分困难。原因有2方面：①该距离包括2种样本对之间的距离，分别为同模态样本对的距离和跨模态样本对的距离，由于网络优化函数的特性，跨模态样本对之间的优化要远比同模态样本对之间的优化困难，导致难以达到全局最优，甚至局部最优；②同模态样本对之间的优化与跨模态样本对之间的优化相互对抗，导致优化结果不收敛。

针对以上问题，本文提出一种基于内聚性分析和跨模态近邻损失函数的重识别方法近似了上述样本内聚性度量，直接加强了不同模态下同类样本的内聚性，大大降低了优化问题的难度。例如，给定可视行人图像，首先，寻找红外图像与可视图像的最近邻样本，然后，以该最近邻样本为中心，计算与其他样本的距离。本文将近邻样本作为连接2种模态样本的桥梁，将困难样本之间的联系转化为同类样本之间的联系和跨模态近邻样本之间的联系。本文方法比直接度量困难样本之间的联系更加有效，便于网络学习和推理。

1 相关工作

为解决外观差异和模态差异很难被特征集约束消除的问题，Wang等[9]使用图像级子网络将红外图像转化为可视图像，统一了具有不同模态的图像表示，结合特征级子网络进行外观差异的学习和特征嵌入，获得当时先进的结果。Hao等[10]学习了超球面流形嵌入并约束该超球面上的模态内差异和模态间差异。Ye等[11]在双流网络之上提出模态感知协作学习方法，处理特征级别和分类器级别的模态差异。Hao等[12]设计了一种空间对齐和模态对齐的特征嵌入方法来提取模态不变特征。Wang等[13]使用端到端的生成对抗网络以减少模态间和模态内的变化，用于学习身份一致的判别特征。Jia等[14]提出相似性推理度量，利用模态内样本相似性，规避最佳跨模态图像匹配的跨模态差异。Wei等[15]使用基于注意力的方法进行跨模态图像匹配。Wang等[16]生成跨模态配对图像，并执行数据集级和细粒度实例级对齐。Lu等[17]通过跨模态共享特定特征转移算法，探索模态共享信息和模态特定特征。Ye等[18]挖掘可视-红外模态内部分级和跨模态图级的上下文线索，提出动态双注意力聚合学习方法。Pu等[19]提出双高斯变分自动编码器。Li等[20]将红外-可视双模交叉模态学习重新表述为X-红外-可视三模态学习问题。Park等[21]利用跨模态人物图像之间的密集对应关系解决像素级的跨模态差异。Miao等[22]提出高度紧凑的模态感知风格适应框架，探索可视和红外模态之间的潜在关系。Ling等[23]提出一种多约束相似性学习方法，从实例到实例、中心到实例和中心到中心3个不同方面共同考虑跨模态关系。Zhao等[24]通过颜色无关一致性学习和身份感知模态适应方法对齐身份级别特征分布。Chen等[25]研究了神经特征搜索的通用范式，实现自动化特征选择。Wu等[26]提出一个联合模态和模式对齐的网络，探索不同模式中的模态间细微差别。Tian等[27]提出变分自蒸馏、变分跨模态蒸馏和变分互学习，以消除与任务无关的信息。

但是，上述研究都缺少对模态之间样本聚集性的分析，针对该问题，本文详细分析了跨模态问题的样本内聚性，并提出跨模态近邻损失函数。

如图1所示，传统方法的跨模态同类别样本内聚性度量可以表示为直接计算所有不同模态样本对之间的距离之和。在本文方法中，给定可视行人图像，首先寻找红外图像与可视图像的最近邻样本，然后以该最近邻样本为中心，计算与其他样本的距离。

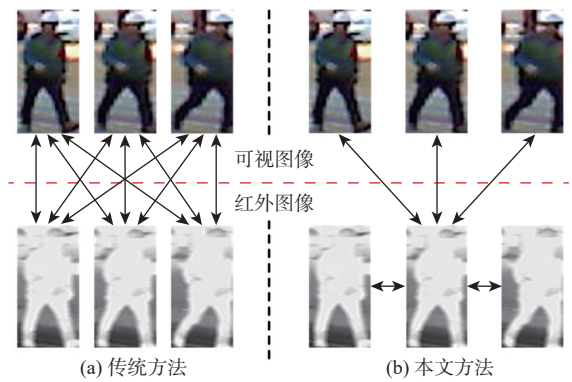


图1 不同方法的内聚性度量示意图

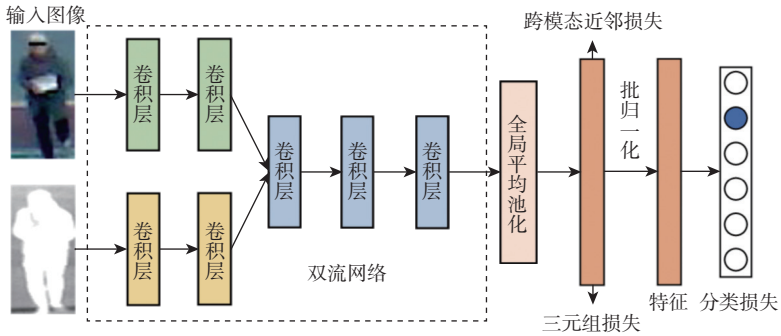
Fig. 1 Illustration of cohesion metrics for different methods

2 本文方法

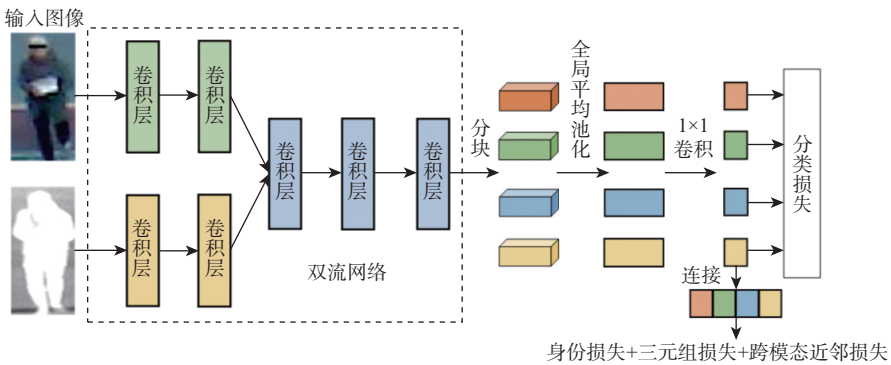
2.1 基线网络

考虑到跨模态近邻损失函数的通用性，本文选取2个基线网络，均为目前行人重识别任务中流行的基础网络结构，分别为全局特征表示的基线网络和部分特征表示的基线网络。基线网络结构示意图如图2所示。

如图2(a)所示，全局特征表示的基线网络结构主要由一个双流网络[28]、全局平均池化层、特征嵌入的瓶颈结构与损失计算组成。本文使用ResNet-50作为特征提取器。与文献[29]的实验结果保持一致，该双流网络不共享第1个浅层卷积块和第



(a) 全局特征表示的基线网络结构



(b) 部分特征表示的基线网络结构

图 2　基线网络结构示意图

Fig. 2　Schematic diagram of baseline network architecture

1 个残差块的参数, 而共享后 3 个残差块的参数。根据传统行人重识别模型, 损失函数结合身份损失 L_{ID} 和在线硬挖掘的三元组损失 L_{tri}, 则全局特征表示的基线网络的损失函数为

$$L_{\mathrm{b}} = L_{\mathrm{ID}} + \alpha L_{\mathrm{tri}} \tag{1}$$

式中: L_{ID} 为预测样本 ID 的交叉熵损失, 用于加强同身份样本之间的内聚性, 更好地学习身份相关的特征; L_{tri} 用于加强不同身份样本之间的分离性, 将与身份无关的特征忽略; α 为权重函数, 调节 L_{ID} 和 L_{tri} 的权重。

如图 2(b) 所示, 部分特征表示的基线网络同样包含 3 个部分: 双流网络的主干部分、部分特征提取模块和损失计算模块。将双流网络特征在水平上平均分为 P 个条带, 生成粗略的身体部分特征, 即生成 P 个三维特征图, 每个条带的特征都加以分类监督信息。网络将上述特征连接起来, 进行全局监督, 损失函数包括最终特征 $\boldsymbol{f}$ 的分类损失和各个部分特征 $\boldsymbol{f}^p$ 的三元组损失之和, 即

$$L = L_{\mathrm{ID}} + \alpha L_{\mathrm{tri}}(\boldsymbol{f}) + \gamma \sum_{p=1}^{P} L_{\mathrm{tri}}(\boldsymbol{f}^p) \tag{2}$$

式中: γ 为权重函数, 调节全局特征和部分特征损失函数之间的权重。

2.2　跨模态行人重识别中的内聚性

为更好地衡量建模跨模态行人重识别样本之间的距离特征, 本文提出内聚性概念, 用于衡量每组样本在特征空间中分布的聚集程度。

在可视-红外跨模态行人重识别任务中, 同类别的样本可以分为可视图像样本和红外图像样本。假设每个批次的训练样本中有 S 个类别, 每个类别有可视图像样本 K 个、红外图像样本 K 个, 则该批次中共有样本 $2SK$ 个。一个批次的样本特征可形式化表示为: $F_{K,S} = \{\boldsymbol{f}^{\mathrm{v}}_{k,s}, \boldsymbol{f}^{\mathrm{r}}_{k,s}\}$, $k = 1,2,\cdots,K, s = 1,2,\cdots,S$。本文使用一组样本中每张图像与其他图像的距离之和来定义内聚性, 则类别为 s 的样本特征的内聚性 C_s 可以表示为

$$C_s = \frac{1}{2K^2}\sum_{i=1}^{K}\sum_{j=1}^{K}\left(\left\|\boldsymbol{f}^{\mathrm{v}}_i - \boldsymbol{f}^{\mathrm{v}}_j\right\| + \left\|\boldsymbol{f}^{\mathrm{r}}_i - \boldsymbol{f}^{\mathrm{r}}_j\right\|\right) + \frac{1}{K^2}\sum_{i=1}^{K}\sum_{j=1}^{K}\left\|\boldsymbol{f}^{\mathrm{v}}_i - \boldsymbol{f}^{\mathrm{r}}_j\right\| \tag{3}$$

C_s 越小, 即类别 s 的内聚性越强, 代表类别 s 的样本在特征空间上的分布越紧密; 反之, C_s 越大, 则类别 s 的内聚性越弱, 其样本在特征空间上的分布越分散。记同模态样本内聚性 $C_s^{\mathrm{intra}} = \frac{1}{2K^2}\sum_{i=1}^{K}\sum_{j=1}^{K}\left(\left\|\boldsymbol{f}^{\mathrm{v}}_i - \boldsymbol{f}^{\mathrm{v}}_j\right\| + \left\|\boldsymbol{f}^{\mathrm{r}}_i - \boldsymbol{f}^{\mathrm{r}}_j\right\|\right)$, 跨模态样本内聚

性 $C_s^{\text{inter}}=\frac{1}{K^2}\sum_{i=1}^{K}\sum_{j=1}^{K}\left\|\boldsymbol{f}_i^{\text{v}}-\boldsymbol{f}_j^{\text{r}}\right\|$，则 C_s 可以表示为：$C_s=C_s^{\text{intra}}+C_s^{\text{inter}}$。对于可视-红外跨模态行人重识别任务来说，同模态样本之间的特征差异远小于跨模态样本之间的特征差异，即同模态样本之间的内聚性要远大于跨模态样本的内聚性，即：$C_s^{\text{inter}}\gg C_s^{\text{intra}}$。因此，$C_s$ 可以近似表示为：$C_s\approx C_s^{\text{inter}}$。综上所述，批次 $F_{K,S}$ 的内聚性可以表示为

$$C\approx\frac{1}{SK^2}\sum_{s=1}^{S}\sum_{i=1}^{K}\sum_{j=1}^{K}\left\|\boldsymbol{f}_{i,s}^{\text{v}}-\boldsymbol{f}_{j,s}^{\text{r}}\right\| \tag{4}$$

2.3 跨模态临近损失

跨模态样本对 $(\boldsymbol{f}_i^{\text{v}},\boldsymbol{f}_j^{\text{r}})$ 通常可以分为2种：简单的跨模态样本对和困难的跨模态样本对。简单的跨模态样本对之间的距离 d_{sim} 比困难的跨模态样本对之间的距离 d_{hard} 更加容易最小化，也可以形式化表示为

$$d_{\text{sim}}<d_{\text{hard}} \tag{5}$$

对于同一个类别，为方便表述，省略类别下标 s。对属于某一个模态的样本 $\boldsymbol{f}_i^{\text{v}}$ 或 $\boldsymbol{f}_j^{\text{r}}$，简单样本可以定义为在另一个模态中距离该样本的样本特征最小的样本，形式化表示为

$$\begin{cases}\boldsymbol{Q}_i^{\text{r}}=\boldsymbol{A}_{1\leqslant j\leqslant K}\left\|\boldsymbol{f}_i^{\text{v}}-\boldsymbol{f}_j^{\text{r}}\right\|\\ \boldsymbol{Q}_j^{\text{v}}=\boldsymbol{A}_{1\leqslant i\leqslant K}\left\|\boldsymbol{f}_i^{\text{v}}-\boldsymbol{f}_j^{\text{r}}\right\|\end{cases} \tag{6}$$

式中：$\boldsymbol{A}$ 为 argmin，即求出使范数达到最小值时样本特征的取值。

又有如下公式成立：

$$\left\|\boldsymbol{f}_i^{\text{v}}-\boldsymbol{f}_j^{\text{r}}\right\|\leqslant\left\|\boldsymbol{f}_i^{\text{v}}-\boldsymbol{Q}_i^{\text{r}}\right\|+\left\|\boldsymbol{Q}_i^{\text{r}}-\boldsymbol{f}_j^{\text{r}}\right\| \tag{7}$$

$$\left\|\boldsymbol{f}_j^{\text{r}}-\boldsymbol{f}_i^{\text{v}}\right\|\leqslant\left\|\boldsymbol{f}_j^{\text{r}}-\boldsymbol{Q}_j^{\text{v}}\right\|+\left\|\boldsymbol{Q}_j^{\text{v}}-\boldsymbol{f}_i^{\text{v}}\right\| \tag{8}$$

将式(7)与式(8)相加代入 C_s 表达式，整理可得

$$C_s\leqslant\frac{1}{2K}\sum_{i=1}^{K}\left(\left\|\boldsymbol{f}_i^{\text{v}}-\boldsymbol{Q}_i^{\text{r}}\right\|+\left\|\boldsymbol{f}_i^{\text{r}}-\boldsymbol{Q}_i^{\text{v}}\right\|\right)+\frac{1}{K^2}\sum_{i=1}^{K}\sum_{j=1}^{K}\left(\left\|\boldsymbol{f}_j^{\text{r}}-\boldsymbol{Q}_i^{\text{r}}\right\|+\left\|\boldsymbol{f}_j^{\text{v}}-\boldsymbol{Q}_i^{\text{r}}\right\|\right) \tag{9}$$

为方便表示，将式(9)不等号右边定义为 C_{near}，即：$C_s\leqslant C_{\text{near}}$。$C_{\text{near}}$ 的第1项是跨模态简单样本对距离之和，第2项为同模态困难样本对距离之和。因此，优化目标转化成为最小化 C_{near}，即最小化 C_s 的上界。与之前相比，跨模态困难样本对的距离度量被化简删除，因此，最小化 C_{near} 比最小化 C_s 更加有效、一致和稳定。跨模态近邻损失函数可以表示为

$$L_{\text{near}}=\frac{1}{SK}\sum_{s=1}^{S}\sum_{i}^{K}\left(\left\|\boldsymbol{f}_i^{\text{v}}-\boldsymbol{Q}_i^{\text{r}}\right\|+\left\|\boldsymbol{f}_i^{\text{r}}-\boldsymbol{Q}_i^{\text{v}}\right\|\right)+\frac{1}{SK^2}\sum_{s=1}^{K}\sum_{i=1}^{K}\sum_{j=1}^{K}\left(\left\|\boldsymbol{f}_j^{\text{r}}-\boldsymbol{Q}_i^{\text{r}}\right\|+\left\|\boldsymbol{f}_j^{\text{v}}-\boldsymbol{Q}_i^{\text{r}}\right\|\right) \tag{10}$$

在基线网络的基础上，损失函数加入跨模态近邻损失 L_{near}，总损失函数为

$$L_{\text{total}}=L_{\text{b}}+\beta L_{\text{near}} \tag{11}$$

式中：β 为权重，调节 L_{b} 和 L_{near} 的权重。若基线网络是部分特征表示的网络，为避免 P 个部分特征的跨模态近邻损失之间的相互对抗，本文的跨模态近邻损失函数 L_{near} 仅计算连接得到的最终特征表示。

从本节的理论分析可知，本文方法具有2个明显的优势：①度量简单跨模态样本和同模态样本，避免直接度量跨模态困难样本。例如，困难样本对之间相对位置出现较大局部范围调整，导致其他同类别样本间出现较大的度量差异时，本文方法能够更加合理地调整样本间的距离，防止出现较大的模态内和模态间的度量差异，保证了训练过程的效率和稳定性。②加强了模态样本间的内聚性，跨模态样本的内聚性在使用跨模态近邻损失进行度量时，继承了中心损失的优点，降低了优化难度。虽然跨模态样本度量具有很高的不确定度，但是最简单的跨模态样本对之间的可信度是高的，可以将跨模态简单样本对作为中心点进行拟合，大大加强模态样本间的内聚性。

3 实 验

3.1 数据集与评价指标

在可视-红外行人重识别任务中，常用的公开数据集为 SYSU-MM01[13] 和 RegDB[30] 数据集。SYSU-MM01 数据集使用6部相机收集了491个行人图像，其中，编号为1、2、4和5的相机为可见光相机，编号为3和6的相机为红外相机。该数据集共收集了287 628张可视图像和5 792张红外图像，其中，训练集包含296个行人，验证集包含99个行人，测试集包含96个行人。RegDB 数据集使用双相机系统收集可视图像和红外图像，共包含412个行人，每个行人收集了10张可视图像和10张红外图像，其测试环节包含2种模式，即可视-红外(查询图像为可视图像，检索集为红外图像)和红外-可视(查询图像为红外图像，检索集为可视图像)。

用于可视-红外行人重识别任务的评价指标有：①累积匹配特征 CMC；②平均准确度 mAP；③度量平均反向负惩罚 mINP。其中，累积匹配特征(CMC-n)为前 n 个检索图像中包含正确图像的概率，也可表示为 Rank-n。

3.2 实验设置

在图像的预处理阶段，将图像大小转化为228像素×144像素。主干网络 ResNet-50 在 ImageNet 数据集上进行预训练，并将最后的残差块的步幅由

2变为1。数据增强方式只进行水平翻转处理, 不进行随机擦除。对于每个批次的采样, 首先随机选择M个类别的行人图像, 然后对于每个类别的行人分别随机选取K个可视图像和红外图像, 共有$2MK$张图像。在 RegDB 数据集中, M设置为8, K设置为4, 共 64 张图像; 在 SYSU-MM01 数据集中, M设置为6, K设置为8, 共 96 张图像。采用随机梯度下降优化器进行优化, 动量参数设置为0.9, 初始学习率为0.1。同时将热身学习方法应用于训练过程。在前 10 轮训练中, 学习率 l 从 0.01 线性递增至 0.1, 然后保持 0.1 并在第 20 轮时降至 0.01, 再保持到第 50 轮, 降低至 0.001。

3.3　消融实验

为验证跨模态近邻损失的有效性和通用性, 设置了 2 组消融实验进行验证。第 1 组消融实验在全局特征表示的基础上与基线方法和基于中心损失的方法进行对比; 第 2 组消融实验在部分特征表示的基线网络结构上与基线方法进行对比。

3.3.1　全局特征表示的基线网络上的消融实验

分别设置了 4 组实验。第 1 组只使用基线网络损失函数L_b做为损失函数, 第 2 组使用L_b和中心损失L_c, 第 3 组使用L_b与每个模态分别添加中心损失L_c, 第 4 组使用L_b与跨模态近邻损失。在 RegDB 数据集的 2 种实验设置下进行消融实验, 实验结果如表 1 和表 2 所示。其中, CenL 1#表示使用第 2 组实验中的中心损失, CenL 2#表示使用第 3 组实验中的模态中心损失。

实验结果表明, 与基线方法相比, 2 种中心损失在可视-红外模式中有轻微提高, Rank-1 提高了约 0.64 % 和 0.52%, mAP 提高了 0.71% 和 0.68%, 而在红外-可视模式中却有小幅下降, Rank-1 下降了 1.16% 和 1.37%, mAP 提高了 0.13% 和 0。总体来说, 这 2 种中心损失的添加并未使该任务结果有显著提升。本文方法与基线方法相比在 2 个模式下均有所提高。在不经过参数调优之前, 在 2 种实验设置下, Rank-1 分别提高了 2.35% 和 1.41%, mAP 分别提高了 3.54% 和 2.15%。可以看出, 本文方法比中心损失对该任务结果的提升要显著。

3.3.2　部分特征表示的基线网络上的消融实验

为证明本文方法的通用性, 在使用部分特征表示的基线网络上设置了消融实验。由于各种类型的中心损失在部分特征表示的基线网络上对行人重识别的效果并无贡献, 仅设置 2 组实验: 第 1 组为无跨模态近邻损失的基线方法, 第 2 组为带有跨模态近邻损失的方法。第 1 组实验只采用身份损失和三元组损失进行损失度量; 第 2 组实验添加跨模态近邻损失。在 RegDB 数据集 2 种实验设置下进行消融实验, 实验结果如表 3 和表 4 所示。实验结果表明, 与基线方法相比, 本文方法在 2 种模式下均有所提高。在可视-红外模式下, 跨模态近邻损失相较于基线方法在 Rank-1 指标上提高了 2.89%, 在 mAP 指标上提高了 8.44%。在红外-可视模式下, 跨模态近邻损失相较于基线方法在 Rank-1 指标上提高了 5.13%, 在 mAP 指标上提高了 10.39%。可以看出本文方法对于该任务结果提升显著, 尤其在红外-可视模式下, mAP 指标提升超 10%。

表 1　RegDB 数据集中可视-红外模式下全局特征表示的消融实验结果

Table 1　Ablation experimental results of global feature representation under visible-infrared mode on RegDB dataset　%

方法	Rank-1	Rank-10	Rank-20	mAP	mINP
基线方法	82.76	91.91	94.57	80.64	73.44
CenL 1#	83.40	92.47	95.53	81.35	73.95
CenL 2#	83.28	92.58	95.16	81.32	74.17
本文方法	85.11	93.50	96.06	84.18	76.70

表 2　RegDB 数据集中红外-可视模式下全局特征表示的消融实验结果

Table 2　Ablation experimental results of global feature representation under infrared-visible mode on RegDB dataset　%

方法	Rank-1	Rank-10	Rank-20	mAP	mINP
基线方法	83.67	93.00	95.59	80.65	71.41
CenL 1#	82.51	92.75	95.67	80.78	72.62
CenL 2#	82.30	92.02	94.73	80.65	72.57
本文方法	85.08	93.00	95.87	82.80	75.40

表 3　RegDB 数据集中可视-红外模式下部分特征表示的消融实验结果

Table 3　Ablation experimental results of part feature representation under visible-infrared mode on RegDB dataset　%

方法	Rank-1	Rank-10	Rank-20	mAP	mINP
基线方法	91.05	97.16	98.57	83.28	68.84
本文方法	93.94	97.87	98.96	91.72	85.99

表 4　RegDB 数据集中红外-可视模式下部分特征表示的消融实验结果

Table 4　Ablation experimental results of part feature representation under infrared-visible mode on RegDB dataset　%

方法	Rank-1	Rank-10	Rank-20	mAP	mINP
基线方法	89.30	96.41	98.16	81.46	64.81
本文方法	94.43	97.80	98.55	91.85	85.76

在SYSU-MM01数据集中的单发-全局实验条件下进行检索，实验结果如表5所示。实验结果表明，与基线方法相比，本文方法在Rank-1指标上提高了1.79%，在mAP指标上提高了2.46%。可以看出，本文方法在SSYSU-MM01数据集十分有效，但与RegDB数据集上的提高程度相比，提升幅度不大，可能源于SYSU-MM01数据集中存在更复杂的背景信息和更难对齐的人物特征图，导致跨模态近邻损失在该数据集上难以充分发挥作用，但仍然足够证明本文方法在不同数据集上的通用性和有效性。

表5 SYSU-MM01数据集中部分特征表示的消融实验结果

Table 5 Ablation experimental results under part feature representation on SYSU-MM01 dataset %

方法	Rank-1	Rank-10	Rank-20	mAP	mINP
基线方法	58.18	90.49	95.34	55.25	39.54
本文方法	59.97	92.96	96.96	57.71	43.60

3.3.3 超参数设置

本文方法的超参数β度量了跨模态近邻损失对总损失的贡献程度。根据实验可得，β在0～0.1以内时，Rank-1和mAP随β的增长而增长，在0～0.2以内时，mINP随β的增长而增长。因此，本文取$\beta=0.2$。

3.4 可视化实验结果

本文对提出的跨模态近邻损失函数在红外-可视行人重识别任务中的结果进行可视化。图3为红外-可视和可视-红外模式下检索Rank-10结果示意图。其中，首列图像为查询图像，后10列图像为查询结果。可以看到，本文方法针对每张图像找到的10个相似度最高的结果均正确。

3.5 与先进方法的比较

将本文方法与目前先进的方法进行比较，包括多块匹配网络(multi-patch matching network, MPMN)[31]、神经特征搜索(neural feature search, NFS)[25]、模态和模式联合网络(modality and pattern alignment network, MPANet)[26]。在RegDB数据集和SYSU-MM01数据集上进行比较，RegDB数据集2种实验设置的实验结果如表6所示，在SYSU-MM01数据集中单发-全局实验条件下进行检索，实验结果如表7所示。

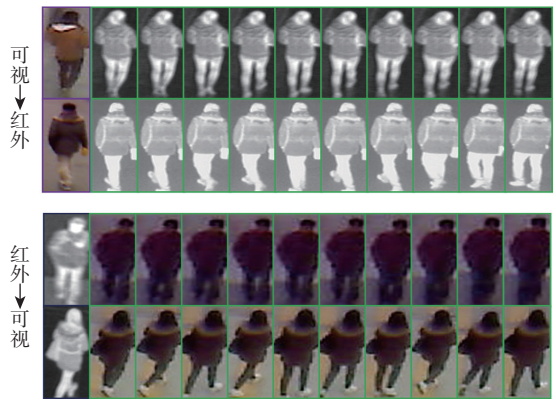


图3 红外-可视和可视-红外模式下检索Rank-10结果示意图

Fig. 3 Retrieval Rank-10 results for infrared-visible and visible-infrared modes

由表6可知，2个基线网络添加跨模态近邻损失函数后在2种模式中都获得了最佳性能，且与未添加相比差距明显。在可视-红外模式下，本文方法在全局特征表示的基线网络中，Rank-1、mAP和mINP分别达到了89.34%、86.49%和79.47%，在部分特征表示的基线网络中，Rank-1、mAP和mINP分别达到了93.94%、91.72%和85.99%；在红外-可视模式下，本文方法在全局特征表示的基线网络中，Rank-1、mAP和mINP分别达到了88.26%、85.34%和77.19%，在部分特征表示的基线网络中，Rank-1、mAP和mINP分别达到了94.43%、91.85%和85.76%。实验表明，跨模态近邻损失函数可以在不同粒度特征表示的网络结构上使用，结合传统的身份损失和三元组损失，更好地学习具有区分性的行人特征，显著提高了可视-红外跨模态行人重识别任务的表现。

由表7可知，与当前先进的方法相比，在部分特征表示的基线网络上添加跨模态近邻损失函数的本

表6 RegDB数据集中可视-红外和红外-可视模式下本文方法与先进方法的比较

Table 6 Comparison with proposed method and advanced mothods on RegDB datasets under visible-infrared and infrared-visual modes %

方法	Rank-1		mAP		mINP	
	可视-红外	红外-可视	可视-红外	红外-可视	可视-红外	红外-可视
MPMN[31]	86.56	84.62	82.91	79.49		
NFS[25]	80.54	77.95	72.10	69.79		
MPANet[26]	82.80	83.70	80.70	80.90		
本文方法(全局特征表示)	89.34	86.49	88.26	85.34	79.47	77.19
本文方法(部分特征表示)	93.94	94.43	91.72	91.85	85.99	85.76

表 7　SYSU-MM01 数据集单发-全局设置下较方法与先进方法的比较

Table 7　Comparison with proposed method and advanced methods on SYSU-MM01 datasets under one shot-global setting　%

方法	Rank-1	mAP	mINP
AGW[32]	47.50	47.64	35.30
NFS[25]	56.91	55.45	
MPANet[26]	70.58	68.24	
本文方法(全局特征表示)	53.30	52.50	39.32
本文方法(部分特征表示)	59.97	57.71	43.60

文方法, Rank-1、mAP 和 mINP 分别达到了 59.97%、57.71% 和 43.60%, 优于其他方法。

3.6　计算复杂度分析

本文方法仅在基线网络上添加了跨模态近邻损失, 未添加其他网络模块, 没有增加其他复杂的网络计算, 也没有增加任何的参数空间。由于损失函数的计算复杂度接近于 $O(1)$, 理论上本文方法带来的计算量增加可以忽略不计。

4　结　论

1) 本文在可视-红外跨模态行人重识别样本的内聚性分析基础上, 致力于减少不同模态下样本的模态差异, 提出跨模态近邻损失, 结合传统的身份损失和三元组损失, 将模态的中心点转化为跨模态简单样本对, 增强可视-红外小跨模态行人重识别任务中同类样本不同模态特征识别的内聚性, 同时保持不同类别特征的可辨别性。

2) 在全局特征表示和部分特征表示 2 个常用的基线网络上, 跨模态近邻损失函数都有出色的表现。与目前先进的方法相比, 本文方法在 2 个不同的基线网络上都取得了卓越的效果, 充分验证了本文方法的通用性和有效性。

3) 本文方法几乎不增加计算量, 就可以取得先进的效果。

参考文献（References）

[1] GONG S G, XIANG T. Person re-identification[M]. GONG S G, XIANG T. Visual analysis of behaviour. Berlin: Springer, 2011: 301-313.

[2] ZHENG L, YANG Y, HAUPTMANN A G. Person re-identification: Past, present and future[EB/OL]. (2016-10-10)[2020-03-15]. https://arxiv.org/abs/1610.02984.pdf.

[3] GHEISSARI N, SEBASTIAN T B, HARTLEY R. Person reidentification using spatiotemporal appearance[C]//Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2006: 1528-1535.

[4] KARANAM S, LI Y, RADKE R J. Person re-identification with discriminatively trained viewpoint invariant dictionaries[C]//Proceedings of the IEEE International Conference on Computer Vision. Piscataway: IEEE Press, 2016: 4516-4524.

[5] BAK S, ZAIDENBERG S, BOULAY B, et al. Improving person re-identification by viewpoint cues[C]//Proceedings of the IEEE International Conference on Advanced Video and Signal Based Surveillance. Piscataway: IEEE Press, 2014: 175-180.

[6] HUANG Y K, ZHA Z J, FU X Y, et al. Illumination-invariant person re-identification[C]//Proceedings of the 27th ACM International Conference on Multimedia. New York: ACM, 2019: 365-373.

[7] CHO Y J, YOON K J. Improving person re-identification via pose-aware multi-shot matching[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2016: 1354-1362.

[8] ZHAO H Y, TIAN M Q, SUN S Y, 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: 907-915.

[9] WANG Z X, WANG Z, ZHENG Y Q, et al. Learning to reduce dual-level discrepancy for infrared-visible person re-identification[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2020: 618-626.

[10] HAO Y, WANG N N, LI J E, et al. HSME: Hypersphere manifold embedding for visible thermal person re-identification[C]//Proceedings of the AAAI Conference on Artificial Intelligence. Palo Alto: AAAI Press, 2019, 33(1): 8385-8392.

[11] YE M, LAN X Y, LENG Q M. Modality-aware collaborative learning for visible thermal person re-identification[C]//Proceedings of the 27th ACM International Conference on Multimedia. New York: ACM, 2019: 347-355.

[12] HAO Y, WANG N N, GAO X B, et al. Dual-alignment feature embedding for cross-modality person re-identification[C]//Proceedings of the 27th ACM International Conference on Multimedia. New York: ACM, 2019: 57-65.

[13] WANG G A, ZHANG T Z, CHENG J, et al. RGB-infrared cross-modality person re-identification via joint pixel and feature alignment[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE Press, 2020: 3622-3631.

[14] JIA M X, ZHAI Y P, LU S J, et al. A similarity inference metric for RGB-infrared cross-modality person re-identification[EB/OL]. (2020-07-03)[2022-03-18].https://arxiv.org/abs/2007.01504.pdf.

[15] WEI X, LI D G, HONG X P, et al. Co-attentive lifting for infrared-visible person re-identification[C]//Proceedings of the 28th ACM International Conference on Multimedia. New York: ACM, 2020: 1028-1037.

[16] WANG G A, ZHANG T Z, YANG Y, et al. Cross-modality paired-images generation for RGB-infrared person re-identification[C]//Proceedings of the AAAI Conference on Artificial Intelligence. Palo Alto: AAAI Press, 2020, 34(7): 12144-12151.

[17] LU Y, WU Y, LIU B, et al. Cross-modality person re-identification with shared-specific feature transfer[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2020: 13376-13386.

[18] YE M, SHEN J B, CRANDALL D, et al. Dynamic dual-attentive aggregation learning for visible-infrared person re-identification[C]// Proceedings of the European Conference on Computer Vision. Berlin: Springer, 2020: 229-247.

[19] PU N, CHEN W, LIU Y, et al. Dual Gaussian-based variational subspace disentanglement for visible-infrared person re-identification[C]//Proceedings of the 28th ACM International Conference on Multimedia. New York: ACM, 2020: 2149-2158.

[20] LI D G, WEI X, HONG X P, et al. Infrared-visible cross-modal person re-identification with an X modality[C]//Proceedings of the AAAI Conference on Artificial Intelligence. Palo Alto: AAAI Press, 2020, 34(4): 4610-4617.

[21] PARK H, LEE S, LEE J, et al. Learning by aligning: Visible-infrared person re-identification using cross-modal correspondences [EB/OL]. (2021-08-17)[2022-03-20]. https://arxiv.org/abs/2108.07422.

[22] MIAO Z, LIU H, SHI W, et al. Modality-aware style adaptation for RGB-infrared person re-identification[C]//Proceedings of the International Joint Conference on Artificial Intelligence. New York: ACM, 2021: 916-922.

[23] LING Y G, LUO Z M, LIN Y J, et al. A multi-constraint similarity learning with adaptive weighting for visible-thermal person re-identification[C]//Proceedings of the International Joint Conference on Artificial Intelligence. New York: ACM, 2021: 845-851.

[24] ZHAO Z W, LIU B, CHU Q, et al. Joint color-irrelevant consistency learning and identity-aware modality adaptation for visible-infrared cross modality person re-identification[C]//Proceedings of the AAAI Conference on Artificial Intelligence. Palo Alto: AAAI Press, 2021, 35(4): 3520-3528.

[25] CHEN Y, WAN L, LI Z H, et al. Neural feature search for RGB-infrared person re-identification[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2021: 587-597.

[26] WU Q, DAI P Y, CHEN J, et al. Discover cross-modality nuances for visible-infrared person re-identification[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2021: 4328-4337.

[27] TIAN X D, ZHANG Z Z, LIN S H, et al. Farewell to mutual information: Variational distillation for cross-modal person re-identification[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2021: 1522-1531.

[28] YE M, WANG Z, LAN X, et al. Visible thermal person re-identification via dual-constrained top-ranking[C]//Proceedings of the International Joint Conference on Artificial Intelligence. New York: ACM, 2018: 1092-1099.

[29] LIU H J, TAN X H, ZHOU X C. Parameter sharing exploration and hetero-center based triplet loss for visible-thermal person re-identification[EB/OL]. (2020-14-04)[2022-03-20]. https://arxiv.org/abs/2008.06223.pdf.

[30] NGUYEN D T, HONG H G, KIM K W, et al. Person recognition system based on a combination of body images from visible light and thermal cameras[J]. Sensors, 2017, 17(3): 605.

[31] WANG P Y, ZHAO Z C, SU F, et al. Deep multi-patch matching network for visible thermal person re-identification[J]. IEEE Transactions on Multimedia, 2021, 23: 1474-1488.

[32] YE M, SHEN J B, LIN G J, et al. Deep learning for person re-identification: A survey and outlook[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 44(6): 2872-2893.

Cross-modality nearest neighbor loss for visible-infrared person re-identification

ZHAO Sanyuan[1, 2, *], A Qi[1], GAO Yu[1]

(1. School of Computer Science & Technology, Beijing Institute of Technology, Beijing 100081, China;
2. Yangtze Delta Region Academy, Beijing Institute of Technology, Jiaxing 314019, China)

Abstract: The goal of the visual-infrared person re-identification task is to search the image of a specific person in a given modality in the image set taken by other cameras in different modality to find out the corresponding image of the same person. Due to the different imaging methods, there are obvious modal differences between images of different modalities. Therefore, from the perspective of metric learning, the loss function is improved to obtain more discriminative information. The cohesiveness of image features is analyzed theoretically, and a re-recognition method based on cohesiveness analysis and cross-modal nearest neighbor loss function is proposed to strengthen the cohesiveness of different modal samples. The similarity measurement problem of cross-modal hard samples is transformed into the similarity measurement of cross-modal nearest neighbor sample pairs and the same modality sample pairs, which makes the optimization of modal cohesion of the network more efficient and stable. The proposed method is experimentally verified on the baseline networks of global feature representation and partial feature representation. Compared with the baseline method, the proposed method can improve the average accuracy of the visual and infrared person re-identification by up to 8.44%. The universality of the proposed method in different network architectures is proved. Moreover, at the cost of less model complexity and less computation, the reliable visual-infrared person re-identification results are achieved.

Keywords: visible-infrared person re-identification; metric learning; deep learning; cross-modality learning; computer vision

Received: 2022-05-26; **Accepted**: 2022-07-02; **Published Online**: 2022-11-21 10:08

URL: link.cnki.net/urlid/11.2625.V.20221118.1553.001

Foundation item: National Natural Science Foundation of China (61902027)

* **Corresponding author**. E-mail: zhaosanyuan@bit.edu.cn

http://bhxb.buaa.edu.cn jbuaa@buaa.edu.cn
DOI: 10.13700/j.bh.1001-5965.2022.0374

基于 Transformer 的深度条件视频压缩

鲁国，钟天雄，耿晶*
(北京理工大学 计算机学院，北京 100081)

摘　　要：近年来，基于深度学习的视频压缩技术主要基于卷积神经网络 (CNN) 且采用运动补偿-残差编码的架构，由于常见的 CNN 只能利用局部的相关性，以及预测残差本身的稀疏特性，难以取得最优压缩性能。因此，提出一种基于 Transformer 架构的条件视频压缩算法，以实现更优的压缩效果。所提算法基于前后帧之间的运动信息，利用可形变卷积得到对应的预测帧特征；将预测帧特征作为条件信息，对原始输入帧特征进行条件编码，避免了直接编码稀疏的残差信号；利用特征间的非局部相关性，提出一个基于 Transformer 的深度条件视频压缩编码算法，用来实现运动信息编码和条件编码，进一步提升压缩编码的性能。实验结果表明：所提算法在 HEVC、UVG 数据集上均超越了当前主流的基于深度学习的视频压缩算法。

关 键 词：视频压缩；Transformer；深度学习；神经网络；压缩算法
中图分类号：TP37
文献标志码：A　　**文章编号**：1001-5965(2024)02-0442-07

目前，视频数据是互联网流量中的重要组成部分，据相关报道，其已占互联网流量的 70%，并且该比例还在不断提升。视频压缩是通过利用视频时空间的信息冗余，从而节省存储空间和传输带宽的关键技术。在过去的几十年中，国内外提出一系列视频压缩编码标准，如 H.264[1]、H.265[2] 等。其中，目前主流的视频压缩算法使用手工设计的模块单元来减少视频中存在的时空信息冗余，如基于块的运动估计和离散余弦变换(discrete cosine transform, DCT)。虽然以上压缩编码标准已经取得了长足的进步，但是考虑到不断增加的视频数量和种类，设计更为高效的压缩编码算法仍然较为急迫。此外，现有算法不能充分利用大量的数据进行端到端的联合优化，因此，其性能也有待进一步提升。

最近几年基于深度学习的视频压缩技术不断发展，借助深度学习出色的表达能力，当前最新算法已经能够与 H.265[2] 实现相似甚至更好的压缩表现。当前主流的基于学习的视频压缩算法普遍采用运动补偿-残差压缩的混合编码架构。例如，Lu 等[3]提出可以完全端到端的深度视频压缩(end-to-end deep video compression, DVC)算法，采用光流估计网络进行运动估计，并且利用基于卷积神经网络(convolutional neural network, CNN) 的自编码器压缩运动信息和残差。但是，当前主流的基于学习的压缩编码算法主要面临以下 2 个挑战：①当前的自编码器架构普遍基于 CNN，使用卷积操作搭配非线性激活函数实现数据的非线性变换。由于卷积操作是简单局部信息的加权求和，且局限于较小范围，其学习到的变换是局部的、次优的。②为消除时间信息冗余，一般基于运动估计得到运动信息，并通过运动补偿模块得到当前帧的预测帧，仅需使用较低的码率编码当前帧和预测帧之间的像素差值或特征差值。由于残差信号的稀疏特性，直接对残差信号进行压缩，整体的压缩编码效率可能并不令人

收稿日期：2022-05-18；**录用日期**：2022-06-23；**网络出版时间**：2022-10-09 17:35
网络出版地址：link.cnki.net/urlid/11.2625.V.20221008.1705.008
基金项目：国家自然科学基金 (62102024)
* **通信作者**. E-mail：janegeng@bit.edu.cn

引用格式：*鲁国，钟天雄，耿晶. 基于 Transformer 的深度条件视频压缩 [J]. 北京航空航天大学学报，2024，50（2）：442-448.*
LU G，ZHONG T X，GENG J. A Transformer based deep conditional video compression [J]. Journal of Beijing University of Aeronautics and Astronautics，2024，50（2）：442-448 （in Chinese）.

满意。

为解决以上问题, 本文提出一种基于 Transformer 模型和条件编码的深度视频压缩(Transformer based deep conditional video compression, TVC)算法。该算法通过计算前后帧之间的运动信息, 借助可形变卷积生成预测帧特征。为避免直接编码稀疏的残差信号, 本文在编码端将预测帧特征作为条件信息, 从而提升对原始输入帧特征的编码效率。同时, 在解码端仅解码出输入帧特征和预测帧特征的残差, 使网络更易进行训练和提升重建效率。本文提出一个基于 Transformer 的自编码器架构, 用来实现运动信息编码和残差的条件编码算法, 从而充分利用 Transformer 的非局部的表达能力, 进一步增强压缩编码的性能。实验结果表明, 本文算法在标准数据集 HEVC 和 UVG 上能够超越当前主流的基于深度学习的视频压缩编码算法和传统压缩编码算法(如 H.265[2])。

1　相关工作

1.1　基于深度学习的图像压缩编码

传统的图像压缩标准(如 JPEG[4]、BPG[5] 等)是当前主流的图像压缩算法, 在各类场景中广泛应用。由于传统算法普遍基于手工设计的特征, 其性能有待进一步提升。

近年来, 基于神经网络的图像压缩算法得到越来越多的关注。其中, 基于循环神经网络(recurrent netural network, RNN)的图像压缩算法陆续被提出[6-8]。目前, 大部分主流算法均基于自编码器形式的 CNN 架构[9-11]。Ballé等[10] 提出的基于超先验模块的算法, 可以实现与 H.265[2] 的帧内编码相近的压缩性能。Minnen 等[11] 提出的自回归上下文模块的熵编码算法进一步提升了压缩性能, 使其超越了 H.265[2] 的帧内编码性能。文献 [12] 提出使用通道自回归熵模型, 改善了自回归模块的串行计算问题, 并进一步提升了编码性能。文献 [13-17] 将 Transformer 引入图像压缩任务。文献 [14] 将 Transformer 用于建模上下文模块, 取得了比基于卷积的上下文模块的模型[11] 更优的压缩性能。文献 [15] 提出一种基于 Transformer 的熵模型, 能够捕获概率分布估计中的长期依赖关系, 有效提升压缩性能。文献 [13,16-17] 使用 Transformer 作为图像压缩的骨架网络, 替代了编码器或解码器中使用卷积的结构, 取得了优于 CNN 的成果。

1.2　基于深度学习的视频压缩编码

随着基于神经网络的图像压缩编码的发展, 越来越多的基于神经网络的视频压缩编码算法被提出。大部分基于深度学习的视频压缩算法可分为 2 类: ①面向 P 帧编码, 即仅依靠前向参考帧。例如, Lu 等[3] 提出 DVC 算法替换了传统视频编码流程中的全部模块, 第一次实现了端到端优化的深度视频压缩网络。Hu 等[18] 提出自适应分辨率的光流压缩算法, 对不同帧的不同块分别设定合适的分辨率。Agustsson 等[19] 提出在光流中增加尺度参数, 从而更好地拟合运动信息的不确定性。以上算法均在像素域进行运动估计和运动补偿, Lu 等[20] 提出利用可形变卷积在特征空间进行更充分的相关操作。Li 等[21] 提出深度上下文视频压缩(deep contextual video compression, DCVC)算法, 通过定义并使用条件信息, 优化了网络的压缩性能, 取得了在多个数据集上的最优编码效率。②面向 B 帧编码, 可同时利用前后不同时刻的图像作为参考帧进行预测。例如, 文献 [22-24] 均同时使用待编码帧的前一帧和后一帧信息作为参考, 并对残差信息进行压缩。文献 [25-26] 采用更直接的方式对现有图像压缩算法进行拓展, 利用 3 维自编码器对图像成组压缩。

文献 [3,18-26] 算法虽然取得了一些进展, 但是绝大部分算法都依赖运动补偿和残差编码的架构, 对于稀疏的残差压缩, 其性能有待进一步提升。虽然 DCVC[21] 算法利用了条件信息, 但其在解码端直接预测原始图像, 给网络的训练和学习带来了很大挑战。Transformer 在众多领域均展现出超越 CNN 的表现, 然而相关研究在深度视频编码领域仍留有空白。本文算法通过使用 Transformer 作为骨架网络, 在特征空间进行运动估计和运动补偿, 并利用条件编码代替残差编码, 从而取得了超越基于 CNN 的深度视频编码算法的表现。

2　本文算法

2.1　算法概述

图 1 为本文算法结构。对于 t 时刻输入的图像 $\boldsymbol{x}_t$, $t-1$ 时刻的重建帧图像 $\hat{\boldsymbol{x}}_{t-1}$ 作为参考帧。首先, 通过特征提取器分别提取 $\boldsymbol{x}_t$ 和 $\hat{\boldsymbol{x}}_{t-1}$ 对应的特征 $\boldsymbol{f}_t$ 和 $\hat{\boldsymbol{f}}_{t-1}$。然后, 利用运动估计模块生成特征 $\boldsymbol{f}_t$ 和 $\hat{\boldsymbol{f}}_{t-1}$ 之间的运动信息 $\boldsymbol{m}_t$, 并且采用运动压缩模块进行压缩, 得到重建后的运动信息 $\hat{\boldsymbol{m}}_t$。采用基于可形变卷积的运动补偿模块将参考帧特征进行映射, 得到预测特征 $\bar{\boldsymbol{f}}_t$。在编码端, 以预测特征 $\bar{\boldsymbol{f}}_t$ 作为条件信息, 对原始输入帧特征 $\boldsymbol{f}_t$ 进行压缩。在解码端, 条件解码器仅输出残差信息 $\hat{\boldsymbol{R}}_t$, 将其与预测特征 $\bar{\boldsymbol{f}}_t$ 相加, 以重建更准确的输入帧特征 $\hat{\boldsymbol{f}}_t$。最后, 经过图像重

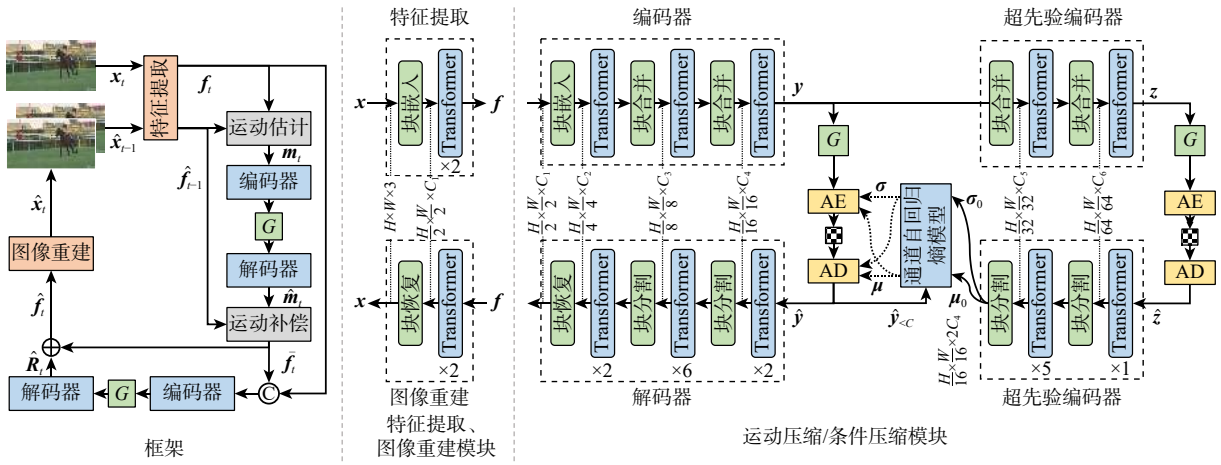


图 1 本文算法结构

Fig. 1 Structure of proposed algorithm

建模块，将$\hat{\boldsymbol{f}}_t$恢复为重建图像$\hat{\boldsymbol{x}}_t$。

为实现码率和失真的最优平衡，本文算法以式 (1) 进行优化：

$$L = R_{\mathrm{m}} + R_{\mathrm{c}} + \lambda d(\boldsymbol{x}_t, \hat{\boldsymbol{x}}_t) \tag{1}$$

式中：L为损失函数；R_{m}和R_{c}分别为运动编码和条件编码消耗的码率；$d(\boldsymbol{x}_t, \hat{\boldsymbol{x}}_t)$为重建帧与原始输入帧之间的损失；$\lambda$为超参数，用于控制优化过程中码率和损失之间的平衡。在运动编码和条件编码过程中，量化后的特征将通过熵编码转化为码流。类似的，在解码过程中，需要通过熵解码将码流恢复为特征。

如图 1 所示，为平衡压缩性能和编解码速度，本文引入基于 Transformer 的超先验网络和通道自回归熵模型[12]，快速、准确估计分布的均值和方差，用于上述熵编码和熵解码过程。其中，H 和 W 分别为图像的高和宽，C 为通道数，AE 和 AD 分别为隐含的熵编码和熵解码模块，G 为量化操作，编码器和解码器的结构对称，解码器下方标注了每层 Transformer 的数量，并使用虚线标注了图像或特征形状的变化。在训练阶段，使用一个码流估计网络来估计消耗的码率。在测试阶段，使用实际的熵编码算法。考虑到量化本身不可微分，采用在训练过程中加入均匀噪声近似量化的算法以保证端到端的优化。

2.2 条件编码

本文算法不再依赖于传统的残差压缩，而是基于高效的条件信息对原始图像（特征）进行直接编码。因此，如何得到高效的条件信息十分重要。本文将前一帧的特征$\hat{\boldsymbol{f}}_{t-1}$变换到当前时刻得到预测特征$\bar{\boldsymbol{f}}_t$，作为条件信息使用。

首先，基于一个双层 CNN 构成的运动估计模块来计算当前帧特征$\boldsymbol{f}_t$和前一帧特征$\hat{\boldsymbol{f}}_{t-1}$的运动信息$\boldsymbol{m}_t$；然后，将运动信息进行压缩编码，解码后的运动信息为$\hat{\boldsymbol{m}}_t$。本文算法借鉴了可形变卷积在运动补偿中的成功应用，同样基于可形变卷积将预测帧特征$\bar{\boldsymbol{f}}_t$对齐到当前时刻。卷积层将$\hat{\boldsymbol{m}}_t$转化为分组的偏移量信息，每组通道之间共享相同的偏移量。对于可形变卷积中的每个卷积核，存在与每个位置对应的运动信息，用于控制从参考帧特征中采样的位置。参考帧特征上对应位置的值被可形变卷积融合为预测特征上的一个值，如下：

$$\bar{\boldsymbol{f}}_t(p) = \sum_{p' \in K^2} \boldsymbol{w}(p') \cdot \hat{\boldsymbol{f}}_{t-1}(p + p' + \Delta p') \tag{2}$$

式中：K^2为 3×3 卷积核的全部位置，取为$\{-1,0,1\}^2$；$\boldsymbol{w}(p')$为对应位置的权重；p和$\Delta p'$分别为特征位置和对应的偏移量。

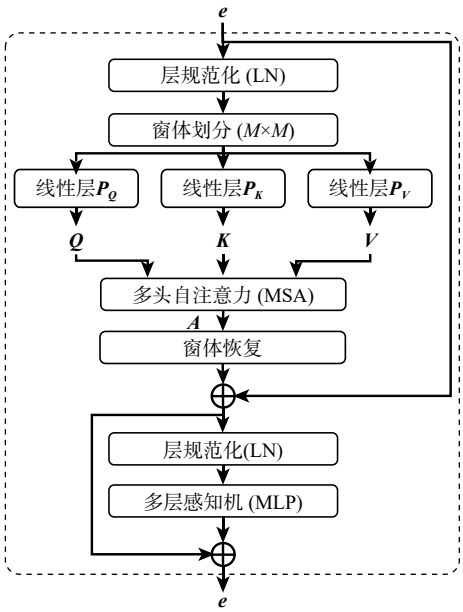
基于得到的预测特征$\bar{\boldsymbol{f}}_t$，将$\bar{\boldsymbol{f}}_t$和原始帧特征$\boldsymbol{f}_t$进行拼接，输入到编码器，对应的量化后的隐空间特征为$\hat{\boldsymbol{y}}_t$。在解码端，基于解码得到残差特征$\hat{\boldsymbol{R}}_t$和原始条件特征$\bar{\boldsymbol{f}}_t$，进行加法操作，得到最终的重建原始特征$\hat{\boldsymbol{f}}_t$，过程如下：

$$\begin{cases} \hat{\boldsymbol{y}}_t = G(F_{\mathrm{E}}([\boldsymbol{f}_t, \bar{\boldsymbol{f}}_t])) \\ \hat{\boldsymbol{f}}_t = F_{\mathrm{D}}(\hat{\boldsymbol{y}}_t) + \bar{\boldsymbol{f}}_t \end{cases} \tag{3}$$

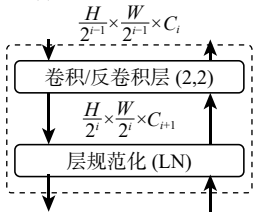
式中：F_{E}和F_{D}分别为编码操作和解码操作。

2.3 基于 Transformer 的压缩模型

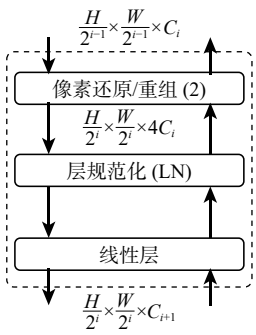
本文算法需要对运动信息$\boldsymbol{m}_t$及原始特征进行编码。为进一步提高压缩效率，引入 Transformer 作为基础模块构建自编码器，整体架构如图 1 所示。相比传统的 CNN 模型，Transformer 能够进一步利用特征的相关性，消除压缩冗余，获得更好的压缩效率。图 2 为 Transformer 块及相关模块结构，为进行特征的上采样和下采样，还引入块合并和块分割层（见图 2(c)），通过空间维度和通道维度的变换实现空间尺寸的变化。



(a) Swin-Transformer块



(b) 块嵌入/恢复



(c) 块合并/分割

图 2　Transformer 块及相关模块结构

Fig. 2　Structure of Transformer block and relevant modules

在确保 Transformer 长距离依赖特性的前提下, 降低其过高的计算代价, 本文提出的模块使用了基于滑动窗口的 Swin-Transformer[27]。如图 2(a) 所示, Swin-Transformer 块将嵌入特征分割为$M\times M$互不重叠的窗口, 在每个窗口内部运用自注意力机制进行计算, 从而实现了线性计算复杂度。将窗口重新恢复为特征图, 经过多层感知机对特征图进一步变换。在相邻的 Swin-Transformer 层之间, 会进行一次窗体向右下方的滑动, 距离为$M/2$, 保证了信息能够跨窗体流动, 保留了 Transformer 长距离依赖的优势。考虑传输到编(解)码器和超先验编(解)码器的特征图尺寸不同, 在实现中使用了不同的窗口大小, 分别设置为 8 和 4。对于第n个窗口$\boldsymbol{e}_n$, 其在自注意力机制中经过不同线性层生成对应的$\boldsymbol{Q}_n$、$\boldsymbol{K}_n$和$\boldsymbol{V}_n$, 如下:

$$\begin{cases}\boldsymbol{Q}_n=\boldsymbol{e}_n\boldsymbol{P}_Q\\ \boldsymbol{K}_n=\boldsymbol{e}_n\boldsymbol{P}_K\\ \boldsymbol{V}_n=\boldsymbol{e}_n\boldsymbol{P}_V\end{cases}\tag{4}$$

式中: $\boldsymbol{P}_Q$、$\boldsymbol{P}_K$和$\boldsymbol{P}_V$为映射矩阵, 在不同的窗口之间共享权重。所有窗体对应的$\boldsymbol{Q}_n$、$\boldsymbol{K}_n$和$\boldsymbol{V}_n$汇总为张量$\boldsymbol{Q}$、$\boldsymbol{K}$和$\boldsymbol{V}$。多头自注意力机制为

$$\boldsymbol{A}=\mathrm{SoftMax}(\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}/\sqrt{d}+\boldsymbol{B})\boldsymbol{V}\tag{5}$$

式中: $\boldsymbol{B}$为可学习的相对位置编码; d为多头自注意力机制中每个头的通道数, 本文设置为 32, 即第i组 Transformer 中的头数为$C_i/32$。

3　实　验

3.1　实验设置

3.1.1　训 练 集

本文使用 Vimeo-90K[28] 数据集作为训练集, 该数据集包含 89 800 个 7 帧、448×256 分辨率的视频序列。在训练中, 随机地将视频序列裁剪成大小为 256×256 的视频序列。

3.1.2　测 试 集

在 HEVC[2] 和 UVG[29] 数据集上评估本文算法的性能。其中, HEVC 被划分为 B、C、D、E 等 4 个类别, 共包含 16 个不同分辨率的视频序列。UVG 包含分辨率为 1 920×1 080 的 7 个高帧率视频序列。因此, 本文的测试集涵盖了具有多种分辨率和帧间差距的视频, 以全面评估框架的性能和鲁棒性。

3.1.3　评估指标

本文使用每个像素消耗的比特 Bpp 作为运动编码和条件编码过程中平均码率的衡量标准, 使用峰值信噪比 PSNR 来衡量重建图像与原始输入图像之间的损失。

3.1.4　实现细节

在训练过程中, 使用 4 种不同的λ取值($\lambda=256,512,1\,024,2\,048$), 并采用 6 个阶段的训练策略。每个阶段训练的轮数和损失函数如表 1 所示, 其中$d(\boldsymbol{x}_t,\bar{\boldsymbol{x}}_t)$表示计算原始输入图像与预测特征还原图像之间的损失, 在实验中使用均方误差 MSE 表示。

表 1　不同训练阶段的损失函数

Table 1　Loss function at different training stages

阶段	步数/10^4	损失函数
1	10	$\lambda d(\boldsymbol{x}_t,\bar{\boldsymbol{x}}_t)$
2	10	$R_{\mathrm{m}}+\lambda d(\boldsymbol{x}_t,\bar{\boldsymbol{x}}_t)$
3	20	$\lambda d(\boldsymbol{x}_t,\hat{\boldsymbol{x}}_t)$
4	10	$R_{\mathrm{c}}+\lambda d(\boldsymbol{x}_t,\hat{\boldsymbol{x}}_t)$
5	140	$R_{\mathrm{c}}+\lambda d(\boldsymbol{x}_t,\hat{\boldsymbol{x}}_t)$
6	10	$R_{\mathrm{c}}+\lambda d(\boldsymbol{x}_t,\hat{\boldsymbol{x}}_t)$

第1阶段，优化与运动信息相关的模块，使预测的条件信息尽可能与输入的当前帧接近。第2阶段，对运动信息编码的码率加以约束，从而平衡运动信息的码率代价和重建质量。第3阶段，冻结运动估计、运动压缩和运动补偿模块参数，优化网络的剩余参数，仅约束网络的重建质量。第4阶段，在第3阶段的基础上，加以对条件编码的码率约束。第5阶段，解冻运动模块的参数，整个网络端到端训练。第6阶段，冻结量化操作前的全部网络参数，并使用取整的量化方法对其他参数进一步训练。其中，前5个阶段使用5×10^{-5}的学习率，并在最后一个阶段降低为5×10^{-6}的学习率。训练共经过2×10^{6}步。

如图1所示，特征提取模块由1个块嵌入层和2层Transformer层构成，图像重建模块则由2层Transformer和1个块恢复构成，与编码器保持对称结构。其中，块嵌入使用步长、卷积核都为2的卷积层实现，将输入的图像分割成许多2×2互不重叠的块并分别编码为特征向量；块恢复是块嵌入的逆过程，使用同样配置的反卷积层。块嵌入与块恢复层的结构如图2(b)所示。此外，实验中，图1中自编码器使用的通道数设置为$(C_1,C_2,\cdots,C_6)=(64,128,160,192,96,128)$。

3.2 实验结果

本节对本文算法与传统算法H.265[2]、基于卷积和残差编码实现的视频压缩编码算法DVC[3]、文献[19]算法进行对比。为得到H.265[2]的压缩结果，与参考文献[3]保持一致，使用FFmpeg的默认模式(default)。为确保实验的公平性，基于深度学习的算法均使用基于MSE损失函数进行优化，所有算法在测试时都将图像组(group of pictures, GoP)的大小设置为10，并使用与H.265[2]相同的算法重建I帧。表2为相对码率节省BDBR结果，基线算法为H.264[1]。其中，TVC(Res)是基于Transformer并使用残差编码的框架，其仅将本文算法中基于条件信息的编码修改为对残差信息编码的算法。

图3为在HEVC、UVG数据集上不同算法的实

表2 HEVC和UVG数据集BDBR结果

Table 2 BDBR results on HEVC and UVG datasets

%

数据集	H.265[2]	DVC[3]	TVC	TVC(Res)
HEVC (B)	−21.95	−18.18	−33.94	−28.85
HEVC (C)	−14.48	7.07	−10.02	−7.94
HEVC (D)	−12.40	0.80	−11.40	−7.04
HEVC (E)	−30.81	−28.56	−30.56	−26.33
UVG	−26.07	−19.39	−35.83	−31.37

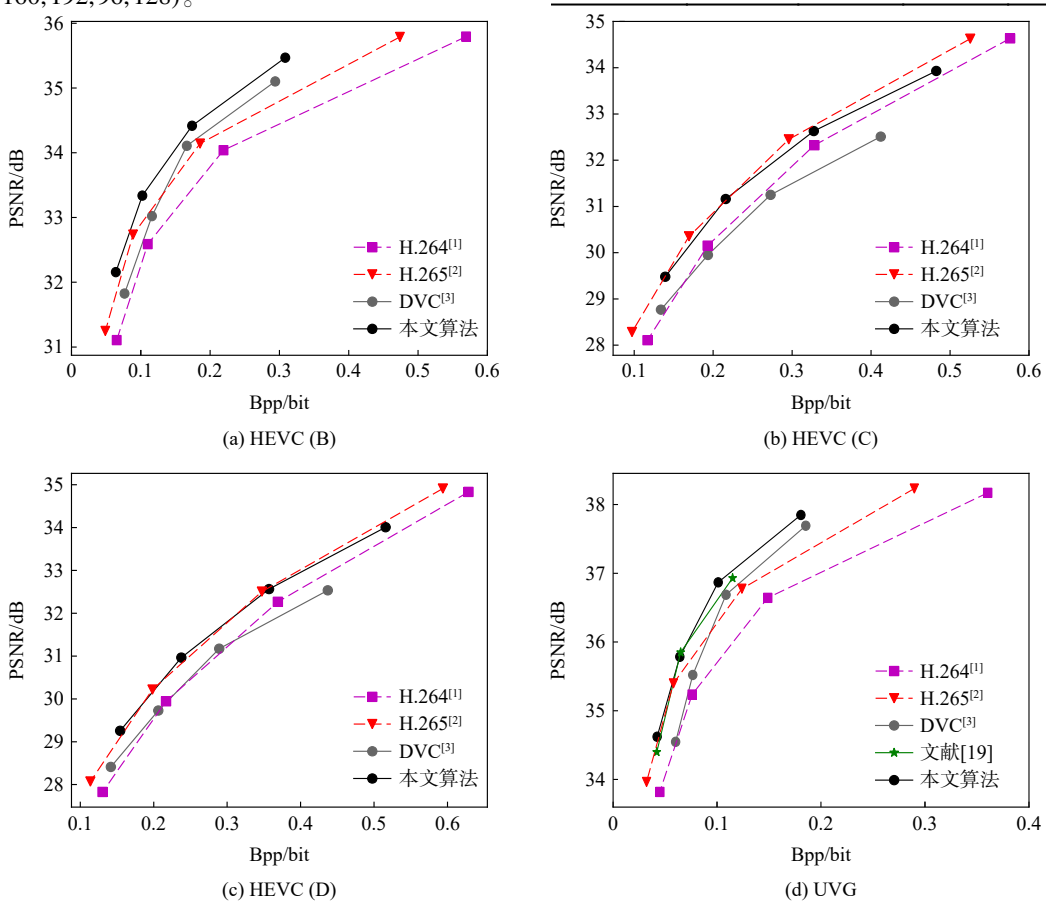


图3 HEVC、UVG数据集上不同算法的实验结果

Fig. 3 Experimental results for different algorithms on HEVC and UVG datasets

验结果。显然，本文算法在 HEVC[2] 的大部分子数据集上，均超越了现有的基于传统算法和基于深度学习的视频编码算法。特别是在 HEVC 的 B 类数据集上，本文算法相比 DVC[3] 和 H.265[2] 在 PSNR 指标上分别提升了约 0.40 dB 和 0.35 dB。另外，在 UVG 数据集上，本文算法相较 DVC[3]、文献 [19] 算法和 H.265[2] 分别提升了约 0.45 dB、0.10 dB 和 0.31 dB。

3.3　消融实验与模型分析

为验证本文算法使用 Transformer 代替卷积，以及使用条件编码代替残差编码的有效性，以 HEVC 数据集为例，进行消融实验。

如表 2 所示，在类似的重建质量下，本文算法相比基于残差编码的 TVC(Res) 算法平均能够节省约 4.0% 的码率，证明了条件编码相对于残差编码的优越性。本文还实现了 DVC[3] 算法作为典型的基于卷积和残差编码的深度视频编码算法。同样，使用残差编码的 TVC(Res) 算法相比 DVC 算法能够节省约 8.7% 的码率，验证了 Transformer 相较卷积层能够实现更有效的变换。Transformer 与条件编码的联合使用，使得本文算法相比 DVC 算法能够节省超过 12.7% 的码率。本文算法共含有约 45×10^6 的参数，在单张 V100（16 GB 内存）显卡上，编码 1 920×1 080 分辨率的视频帧约需 324 ms，解码约需 272 ms。

4　结　论

1）本文提出基于 Transformer 为主要构成模块的端到端的视频压缩算法，实现比卷积网络更充分的变换。

2）本文提出针对残差信息的条件编码方案，避免直接编码稀疏的残差信息，进一步提升残差压缩效率。

3）本文算法在多个标准数据集上进行了验证，其性能超越了当前主流的传统压缩算法和基于卷积和残差编码的深度压缩算法。

参考文献（References）

[1] Joint Video Team of ITU-T and ISO/IEC JTC 1. Advanced video coding:ISO/IEC 14496-10[S].JVT:Pattaya,2003.

[2] BROSS B.High efficiency video coding:ISO/IEC 23008-2[S]. Shanghai:ITU-T/ISO/IEC Joint Collaborative Team on Video Coding,2012.

[3] LU G, OUYANG W L, XU D, et al. DVC: An end-to-end deep video compression framework[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2020: 10998-11007.

[4] WALLACE G K. The JPEG still picture compression standard[J]. Communications of the ACM, 1991, 34(4): 30-44.

[5] BELLARD F. BPG image format[EB/OL]. (2015-02-01)[2022-05-10]. https://bellard.org/bpg.

[6] TODERICI G, O'MALLEY S M, HWANG S J, et al. Variable rate image compression with recurrent neural networks[EB/OL]. (2016-05-01)[2022-05-10]. https://arxiv.org/abs/1511.06085.pdf.

[7] TODERICI G, VINCENT D, JOHNSTON N, et al. Full resolution image compression with recurrent neural networks[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2017: 5435-5443.

[8] JOHNSTON N, VINCENT D, MINNEN D, et al. Improved lossy image compression with priming and spatially adaptive bit rates for recurrent networks[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2018: 4385-4393.

[9] BALLÉ J, LAPARRA V, SIMONCELLI E P. End-to-end optimized image compression[EB/OL]. (2017-05-03)[2022-05-10]. https://arxiv.org/abs/1611.01704.pdf.

[10] BALLÉ J, MINNEN D, SINGH S, et al. Variational image compression with a scale hyperprior[EB/OL]. (2018-05-01)[2022-05-10]. https://arxiv.org/abs/1802.01436.pdf.

[11] MINNEN D, BALLÉ J, TODERICI G. Joint autoregressive and hierarchical priors for learned image compression[C]//Proceedings of the 32nd International Conference on Neural Information Processing Systems. New York: ACM, 2018: 10794-10803.

[12] MINNEN D, SINGH S. Channel-wise autoregressive entropy models for learned image compression[C]//Proceedings of the IEEE International Conference on Image Processing. Piscataway: IEEE Press, 2020: 3339-3343.

[13] ZHU Y, YANG Y, COHEN T. Transformer-based transform coding[EB/OL]. (2022-01-29)[2022-05-10]. https://openreview.net/forum?id=IDwN6xjHnK8.

[14] KOYUNCU A B, GAO H, BOEV A, et al. Contextformer: A Transformer with spatio-channel attention for context modeling in learned image compression[EB/OL]. (2022-07-20)[2022-05-10]. https://arxiv.org/abs/2203.02452.pdf.

[15] QIAN Y C, LIN M, SUN X Y, et al. Entroformer: A Transformer-based entropy model for learned image compression[EB/OL]. (2022-05-14)[2022-05-10]. https://arxiv.org/abs/2202.05492.pdf.

[16] LU M, GUO P Y, SHI H Q, et al. Transformer-based image compression[EB/OL]. (2021-11-12)[2022-05-10]. https://arxiv.org/abs/2111.06707.pdf.

[17] BAI Y C, YANG X, LIU X M, et al. Towards end-to-end image compression and analysis with Transformers[EB/OL]. (2021-12-17)[2022-05-10]. https://arxiv.org/abs/2112.09300.pdf.

[18] HU Z H, CHEN Z H, XU D, et al. Improving deep video compression by resolution-adaptive flow coding[C]//Proceedings of the European Conference on Computer Vision. Berlin: Springer, 2020: 193-209.

[19] AGUSTSSON E, MINNEN D, JOHNSTON N, et al. Scale-space flow for end-to-end optimized video compression[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2020: 8500-8509.

[20] HU Z H, LU G, XU D. FVC: A new framework towards deep video compression in feature space[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscat-

away: IEEE Press, 2021: 1502-1511.

[21] LI J, LI B, LU Y. Deep contextual video compression[EB/OL]. (2021-12-14)[2022-05-01]https://arxiv.org/abs/2109.15047.

[22] WU C Y, SINGHAL N, KRÄHENBÜHL P. Video compression through image interpolation[C]//Proceedings of the European Conference on Computer Vision. Berlin: Springer, 2018: 425-440.

[23] DJELOUAH A, CAMPOS J, SCHAUB-MEYER S, et al. Neural inter-frame compression for video coding[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE Press, 2020: 6420-6428.

[24] YANG R, MENTZER F, VAN GOOL L, et al. Learning for video compression with hierarchical quality and recurrent enhancement[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2020: 6627-6636.

[25] PESSOA J, AIDOS H, TOMÁS P, et al. End-to-end learning of video compression using spatio-temporal autoencoders[C]//Proceedings of the IEEE Workshop on Signal Processing Systems. Piscataway: IEEE Press, 2020: 1-6.

[26] HABIBIAN A, VAN ROZENDAAL T, TOMCZAK J, et al. Video compression with rate-distortion autoencoders[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE Press, 2020: 7032-7041.

[27] LIU Z, LIN Y T, CAO Y, et al. Swin Transformer: Hierarchical vision Transformer using shifted windows[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE Press, 2022: 9992-10002.

[28] XUE T F, CHEN B A, WU J J, et al. Video enhancement with task-oriented flow[J]. International Journal of Computer Vision, 2019, 127(8): 1106-1125.

[29] MERCAT A, VIITANEN M, VANNE J. UVG dataset: 50/120fps 4K sequences for video codec analysis and development[C]//Proceedings of the 11th ACM Multimedia Systems Conference. New York: ACM, 2020: 297-302.

A Transformer based deep conditional video compression

LU Guo，ZHONG Tianxiong，GENG Jing*

(School of Computer Science and Engineering，Beijing Institute of Technology，Beijing 100081，China)

Abstract：Convolutional neural networks (CNN) are the foundation of most recent learning-based video compression algorithms, which also use residual coding and motion compensation architectures. It is difficult to attain the best compression performance given that typical CNN can only use local correlations and the sparsity of prediction residual. To solve the problems above, this paper proposed a Transformer-based deep conditional video compression algorithm, which can achieve better compression performance. The proposed algorithm uses deformable convolution to obtain the predicted frame feature based on the motion information between the front and rear frames. The predicted frame feature is used as conditional information to conditionally encode the original input frame feature which avoids the direct encoding of sparse residual signals. The proposed algorithm further utilizes the non-local correlation between the features and proposes a transformer-based autoencoder architecture to implement motion coding and conditional coding, which further improves the performance of compression. Experiments show that our Transformer based deep conditional video compression algorithm surpasses the current mainstream learning-based video compression algorithms in both HEVC and UVG datasets.

Keywords：video compression; Transformer; deep learning; neural network; compression algorithm

Received：2022-05-18；**Accepted**：2022-06-23；**Published Online**：2022-10-09 17：35

URL：link.cnki.net/urlid/11.2625.V.20221008.1705.008

Foundation item：National Natural Science Foundation of China (62102024)

* **Corresponding author**. E-mail: janegeng@bit.edu.cn

http://bhxb.buaa.edu.cn jbuaa@buaa.edu.cn
DOI: 10.13700/j.bh.1001-5965.2022.0436

基于优化布拉格结构的固态装配型谐振器

张世峰[1,2]，轩伟鹏[3]，石林豪[4]，董树荣[2]，浦世亮[1,*]

(1. 杭州海康威视数字技术股份有限公司，杭州 310051；2. 浙江大学 信息与电子工程学院，杭州 310058；
3. 杭州电子科技大学 电子信息学院，杭州 310018；4. 天通瑞宏科技有限公司，嘉兴 314499)

摘　　要：体声波谐振器的有效耦合系数和品质因数决定了体声波滤波器的整体性能。有效耦合系数依赖于叠层结构和压电材料。而品质因数高度依赖于损耗机制，主要为电学损耗和声学损耗。对于固态装配型谐振器（SMR），声学损耗主要为透过衬底的能量泄漏。为提高SMR的品质因数，改进了布拉格堆叠结构，使纵波和剪切波同时被约束在压电堆叠结构中，以减少声能对衬底的泄漏。同时，为抑制谐振腔的杂散模式，优化了布拉格结构顶层薄膜厚度，器件色散特性由Ⅱ型变为Ⅰ型。通过仿真和实验表明：基于优化布拉格结构的SMR性能得到了明显改善。

关　键　词：体声波；固体装配型谐振器；布拉格；剪切波；高品质因数
中图分类号：V221^{+}.3; TB553
文献标志码：A　　**文章编号**：1001-5965(2024)02-0449-07

随着移动技术的迅速发展与信息化进程的不断推进，从智能家居到汽车的自动化驾驶，再到虚拟现实的应用与发展，通信技术从最早期的人与人之间的通信演变为物与物之间的通信。随着5G通信时代的到来，智能手机与各种小型移动终端所支持的频段不断增加，其通信频率也逐渐升高，这些都对通信系统中的射频滤波器提出了更严格的要求。尤其是5G微型基站的出现，其对滤波器的功率容量也提出了更高的要求[1-3]。

薄膜体声波谐振器(thin film bulk acoustic resonator, FBAR)是近年来出现的一种采用压电薄膜加工的高频体声波谐振器。因其具有体积小、品质因数高且工作频率高的特点，基于FBAR的滤波器被广泛应用。目前，在通信设备中使用的主流体声波滤波器主要基于2种结构，即FBAR与固态装配型谐振器(solidly mounted resonator, SMR)[4]。然而，由于工作频率不断地升高，特别是5G通信技术的应用，对手机等小型通信设备的电池寿命与电路板集成空间都是巨大的挑战。因此，为了避免器件发热而产生的上述问题，需要提升滤波器的功率容量，这是一项有挑战性的任务。

SMR因采用布拉格衬底，具有较强的散热能力，使其拥有比FBAR更大的功率容量，在同样的工作频率下，工作温度更低，机械强度更可靠。然而，传统型SMR中的布拉格结构只能有效反射纵波，不能有效反射剪切波，导致其声学能量损耗较大，器件品质因数较低。因此，为了获得更高品质因数的SMR，有源区下方的布拉格反射结构需要能同时反射纵波和剪切波[5-6]。

本文提出一种新型的布拉格结构，能够有效反射纵波和剪切波，降低布拉格衬底对SMR的损耗，进而提高SMR的品质因数[7-9]。

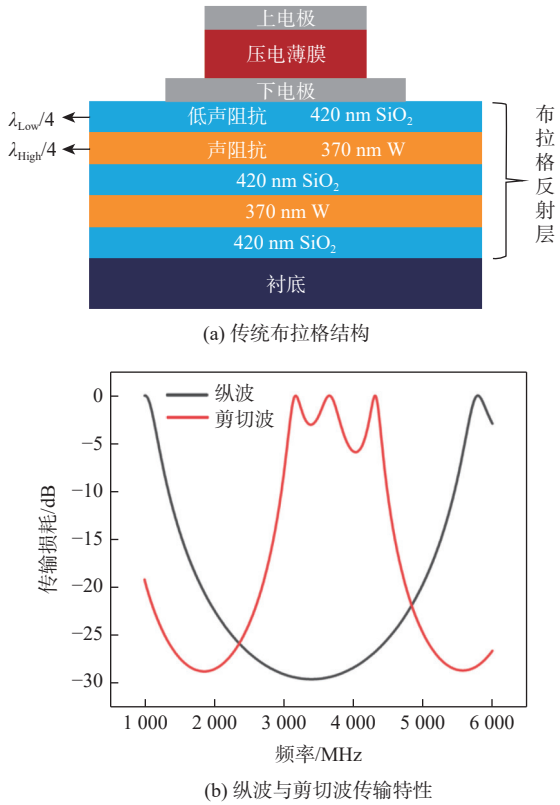
1　布拉格层结构损耗分析

电学损耗与声学损耗是SMR中最常见的损耗。电学损耗是由谐振器的电极与焊盘的欧姆损

收稿日期：2022-05-29；录用日期：2022-07-02；网络出版时间：2022-08-26 10:14
网络出版地址：link.cnki.net/urlid/11.2625.V.20220825.1648.002
*通信作者. E-mail：pushiliang@hikvision.com

*引用格式：张世峰，轩伟鹏，石林豪，等. 基于优化布拉格结构的固态装配型谐振器[J]. 北京航空航天大学学报，2024，50(2)：449-455.
ZHANG S F，XUAN W P，SHI L H，et al. Solidly mounted resonator based on optimized Bragg structure[J]. Journal of Beijing University of Aeronautics and Astronautics，2024，50(2)：449-455 (in Chinese).*

耗所造成的，可以通过对金属 Pad 连接层做加厚处理，或者是用电阻率较低的材料。而声学损耗主要横波平面泄露及横波通过布拉格层的能量泄露。研究人员已经使用干涉测量法对带谐振腔体声波（bulk acoustic wave, BAW）滤波器进行测试，发现横波损耗确实存在。虽然在有源区内，横波能量引起的振幅要比纵波能量引起的振幅小的多，并且横波泄露也并不是 SMR 的主要损耗机制，但其对反谐振频率处的品质因数却有极大的影响。一些研究人员已经用实验证明，可以通过对 SMR 中布拉格结构进行适当的改进，可以有效防止横波能量的泄露[10-13]。布拉格结构是典型的一维声子晶体，按 1/4 波长设计的布拉格结构只能产生一个特定方向的带隙，也就是只能对一个特定波长的声波具有较高的反射能力，也即对谐振频率下的纵波主模式的声波具有较好的反射能力，而不能有效的反射此频率下其他模式的声波，因此，会造成器件部分能量泄露。在一般情况下，压电薄膜中剪切波的波速约为纵波波速的一半，而传统 1/4 波长布拉格反射层只能有效的反射纵波，剪切波大部分能量会透过反射层泄露到衬底之中，显著降低谐振器的品质因数。



(a) 传统布拉格结构

(b) 纵波与剪切波传输特性

图 1 传统布拉格结构纵波和剪切波传输特性

Fig. 1 Transmission characteristics of longitudinal and shear waves of traditional Bragg structure

2 新型布拉格结构层设计

2.1 器件理论分析

本文的目标是实现一款具备高品质因数的 3.5GHz SMR。布拉格层的高声速材料选为钨（W），低声速材料选为二氧化硅（SiO_2），布拉格层数为 5。基于阻带理论和相位误差逼近理论，采用有限元法计算传统布拉格结构对纵波与剪切波的传输损耗，计算结果如图 1 所示，其中 λ_{Low} 为低声阻抗薄膜中的波长，λ_{High} 为高声阻抗薄膜中的波长。

在串联谐振频率附近，纵波传输损耗接近于 −29 dB，剪切波所对应的传输损耗仅为 −2.5 dB 左右，因此，剪切波没有得到有效反射。由图 1 可知，只要在压电薄膜中存在剪切波，则剪切波几乎无损耗的通过布拉格反射层并最终泄露到衬底。该现象是造成传统型 SMR 在并联谐振点的品质因数较低的主要原因。

根据薄膜中光学的阻带理论，在多层膜中，每一层的传输相位为所谓的基点 ϕ_n 时，多层膜结构会产生一个最大反射系数[14]：

$$\phi_n = \frac{n\pi}{1+c} \tag{1}$$

式中：n 为自然数，且 $n \leqslant c$，$c = \phi_H/\phi_L$，ϕ_H 和 ϕ_L 为高传输系数和低传输系数中的传输相位。当 c=1 时，即相位差为 π/2 的 1/4 波长布拉格反射镜设计。SMR 中使用的布拉格结构是根据光学中的布拉格反射镜延伸而来，当声波经过 2 个不同介质构成的薄膜时，会在界面处发生反射。根据布拉格方程可知，当波程差正好等于波长的整数倍时，反射波将会得到加强。当入射光由光疏介质到光密介质中时，将会存在半波损失，入射波与反射波在界面处产生了大小为 π 的相位差；而当入射光在高声阻抗层与低声阻抗层的光程差均为半个波长时，其相位差为 π/2 时，称之为布拉格反射镜。

$$\phi_{low} + \phi_{high} = n\pi \qquad n = 1,2,\cdots,N \tag{2}$$

式中：ϕ_{low} 为声波在低声阻抗材料中的相位差；ϕ_{high} 为声波在高声阻抗层中的相位差。由于相位差等于波数与波程差相乘，即

$$\phi_l = \frac{2\pi}{\lambda_l} t_l \tag{3}$$

式中：λ_l 为薄膜中的波长；t_l 为波程差或薄膜厚度。根据式(3)，进一步展开为

$$\phi_{low} + \phi_{high} = \frac{2\pi}{\lambda_{low}} t_{low} + \frac{2\pi}{\lambda_{high}} t_{high} = 2\pi f_r \left(\frac{t_{low}}{v_{low}} + \frac{t_{high}}{v_{high}} \right) = n\pi \tag{4}$$

又因为在每一层薄膜中, 其纵波与剪切波的声速都不同。因此, 可以将式(4)根据薄膜中的纵波与剪切波分为 2 组方程:

$$\begin{cases} 2\pi f_{\mathrm{r}}\left(\dfrac{t_{\mathrm{low}}}{v_{\mathrm{low_shear}}}+\dfrac{t_{\mathrm{high}}}{v_{\mathrm{high_shear}}}\right)=n_1\pi \\ 2\pi f_{\mathrm{r}}\left(\dfrac{t_{\mathrm{low}}}{v_{\mathrm{low_long}}}+\dfrac{t_{\mathrm{high}}}{v_{\mathrm{high_long}}}\right)=n_2\pi \end{cases} \tag{5}$$

式中: f_{r}为谐振器的串联谐振频率; $v_{\mathrm{low_shear}}$和$v_{\mathrm{high_shear}}$分别为低声阻抗材料和高声阻抗材料的剪切波波速; $v_{\mathrm{low_long}}$和$v_{\mathrm{high_long}}$分别为低声阻抗材料和高声阻抗材料的纵波波速; n_1和n_2为正整数; t_{low}和t_{high}为厚度。定义相位误差因子$\Delta\phi$, 使得其在式(2)中, 无论是纵波还是剪切波均为相等的值。此外, 保持较薄的层厚度, 可以令$n_1=1$, $n_2=2$。根据$t=\lambda_{\mathrm{f}}/4$, 声波在高声阻抗与低声阻抗材料中的相位差写为

$$\begin{cases} \phi_{\mathrm{low_long}}=\dfrac{n\pi}{1+c}\cdot\dfrac{1}{K_{\mathrm{L}}} \\ \phi_{\mathrm{high_long}}=\dfrac{cn\pi}{1+c}\cdot\dfrac{1}{K_{\mathrm{H}}} \end{cases} \tag{6}$$

式中: K_{L}和K_{H}分别为低声阻抗层和高声阻抗层中纵波和剪切波波速之间的比值。此时将式(5)代入到式(4)中可以得到关于薄膜厚度为未知数的方程组, 并且求解该方程组可以得到高低声阻抗厚度的表达式:

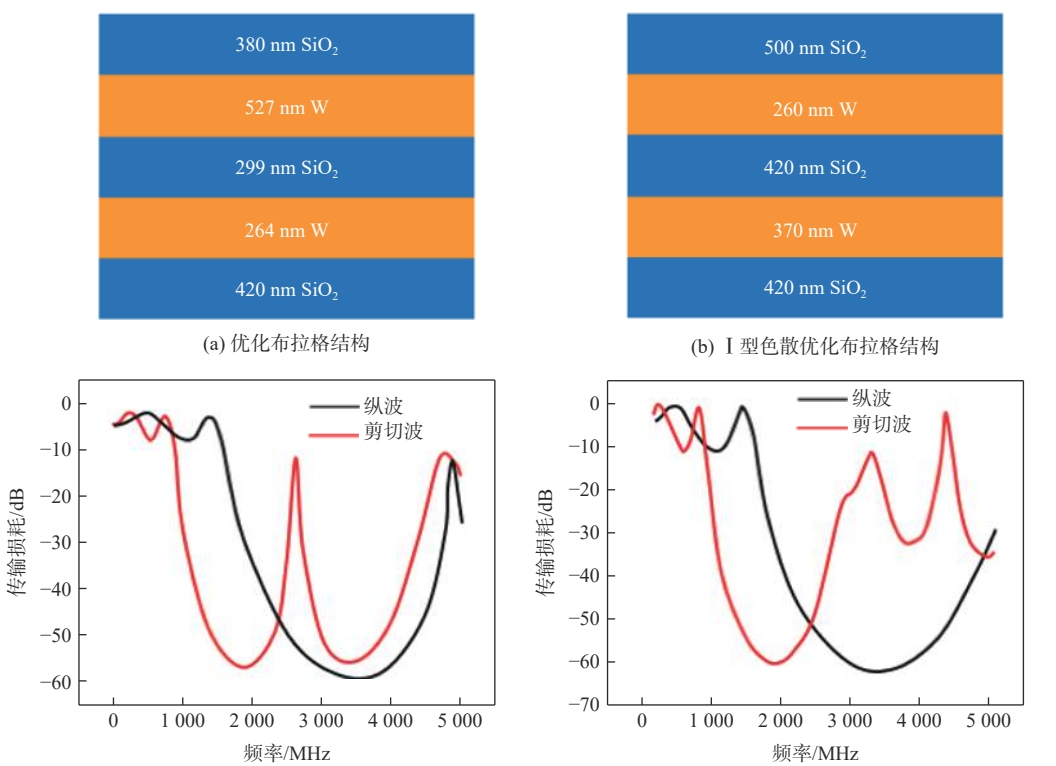
$$t_{\mathrm{low}}=\frac{3}{2}\lambda_{\mathrm{low_long}}/(1+c+K_{\mathrm{L}}+cK_{\mathrm{H}}) \tag{7}$$

$$t_{\mathrm{high}}=\frac{3}{2}c\lambda_{\mathrm{high_long}}/(1+c+K_{\mathrm{L}}+cK_{\mathrm{H}}) \tag{8}$$

因此, 只需确定常数 c, 即可得到具有同时反射纵波与剪切波能力的布拉格反射镜。在声学中, c 可以简化为高低声速阻抗材料中的纵波波速和剪切波波速之比, 通常 c 的值取 2。

2.2　器件特性仿真验证

基于阻带理论和相位误差逼近理论, 可以得到能够同时反射纵波与剪切波能力的布拉格反射结构, 并采用有限元方法进行仿真分析[15]。图 2 为优化的布拉格结构和 Ⅰ 型色散优化布拉格结构的纵波与剪切波传输特性, 从图中可以看到, 在 3.5 GHz 谐振频率下, 纵波和剪切波都具有较低的传输损耗分别为−55 dB 和−53 dB。另一方面, 考虑到基于氮化铝(AlN)的 SMR 为 Ⅱ 型色散, 其在串联谐振点上存在寄生振动。在上述设计的布拉格各层厚度的基础上, 对特定层的厚度进行变量扫描, 使谐振器的色散变为 Ⅰ 型, 并且保证在器件的谐振频率附近, 布拉格结构能够对纵



(a) 优化布拉格结构　　(b) Ⅰ型色散优化布拉格结构

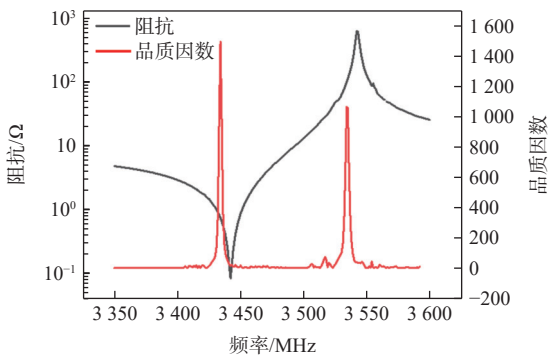
(c) 优化布拉格结构的纵波与剪切波传输特性　　(d) Ⅰ型色散优化布拉格结构的纵波与剪切波传输特性

图 2　优化的布拉格结构和 Ⅰ 型色散优化布拉格结构的纵波与剪切波传输特性

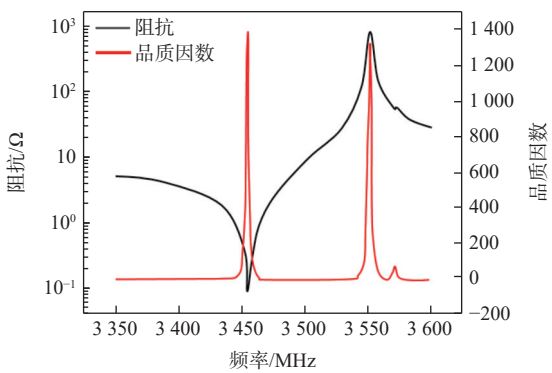
Fig. 2　Optimized Bragg structure, and transmission characteristics of longitudinal and shear waves of optimized Bragg structure for type Ⅰ dispersion

波与剪切波均具有较大的反射系数。在3.5 GHz谐振频率下，纵波和剪切波传输损耗分别为−60 dB和−28 dB。

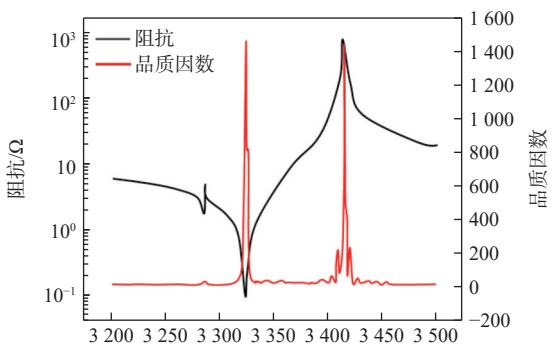
图3为3种布拉格结构的SMR阻抗和品质因数对比。由图3(a)可知，传统型布拉格结构串联谐振点的品质因数为1 495，并联谐振点的品质因数为1 021。由图3(b)可知，优化布拉格结构串联谐振点的品质因数为1 370，并联谐振点的品质因数为1 262。由图3(c)可知，Ⅰ型色散优化布拉格结构串联谐振点的品质因数为1 442，并联谐振点的品质因数为1 422。因此，优化后的SMR，其品质因数提高了39%。



(a) 传统型布拉格结构



(b) 优化布拉格结构



(c) Ⅰ型色散优化布拉格结构

图3 3种布拉格结构的SMR器件阻抗和品质因数对比

Fig. 3 Comparison of impedance and quality factor between SMR devices of three Bragg structures

3 实验结果与分析

3.1 固态装配型谐振器制备

针对Ⅰ型色散优化布拉格结构设计，进行SMR的制备，制备工艺流程如图4所示。采用蓝宝石作为衬底材料，在蓝宝石基片上使用磁控溅射沉积一层具有低声阻抗的SiO_2薄膜，随后在SiO_2薄膜之上使用磁控溅射法沉积一层具有高声阻抗W薄膜。考虑到W与SiO_2这2种材料之间的样式模量相差较大，2种材料之间的结合力也较差，在生长完SiO_2薄膜之后，先沉积一层钛镍合金，作为2种材料之间的黏附层。同时，在每层薄膜沉积完成之后，下次薄膜沉积之前，对器件表面进行清洗，避免颗粒污染导致薄膜质量生长较差。

布拉格多层薄膜的沉积会使得上层薄膜面临粗糙度过大的问题，而下电极与压电薄膜均是制备于布拉格反射结构之上。因此，为了保证AlN压电薄膜具有较好002择优取向生长，在布拉格反射结构制备完成之后，使用化学机械研磨(chemical-mechanical polishing, CMP)工艺实现布拉格反射结构的抛光与平坦化。图5为CMP工艺对薄膜表面平整度的影响，采用原子力显微镜(atomic force microscope, AFM)测试，使用CMP工艺后布拉格结构上层薄膜的粗糙度为2 nm，而未使用CMP工艺的器件薄膜粗糙度为4.13 nm。

采用磁控溅射法沉积金属Mo电极。金属Mo薄膜的晶体取向性能主要受到其下层基低的影响，由于Mo电极厚度较薄(薄膜厚度通常小于250 nm)，且布拉格结构最上层的SiO_2薄膜的晶格结构与Mo(110)晶格取向严重失配。因此，在沉积下电极之前，先制备厚度为30 nm的AlN薄膜作为诱导层，改善金属Mo的晶粒排列，利于Mo薄膜的择优生长，提高上层压电薄膜的质量。

压电薄膜是SMR最为关键的材料之一，其薄膜质量的好坏直接影响着SMR的频率、有效机电耦合系数、品质因数。在完成下电极的刻蚀之后，采用磁控溅射的方法制备AlN压电薄膜，控制溅射功率，关注AlN薄膜择优取向性，以及应力、厚度与表面粗糙度等性能。在完成AlN压电薄膜的图形化之后，采用金(Au)作为引出电极的材料。最终制备的SMR，下电极Mo的厚度为170 nm，压电层AlN的厚度为700 nm，上电极的厚度为150 nm。SMR俯视图如图6所示。

3.2 固态装配型谐振器性能测试

采用AglientE5071C矢量网络分析仪和探针台对SMR进行测试，图7为基于优化布拉格结构

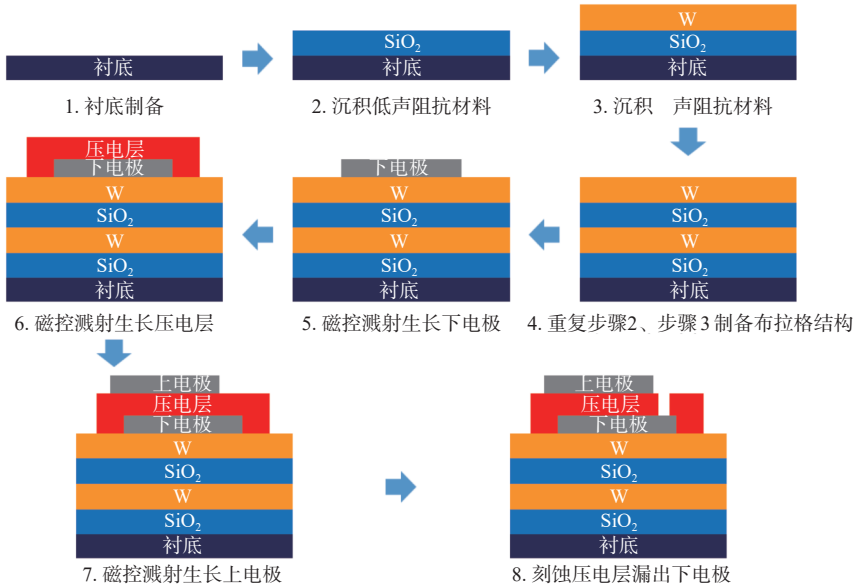
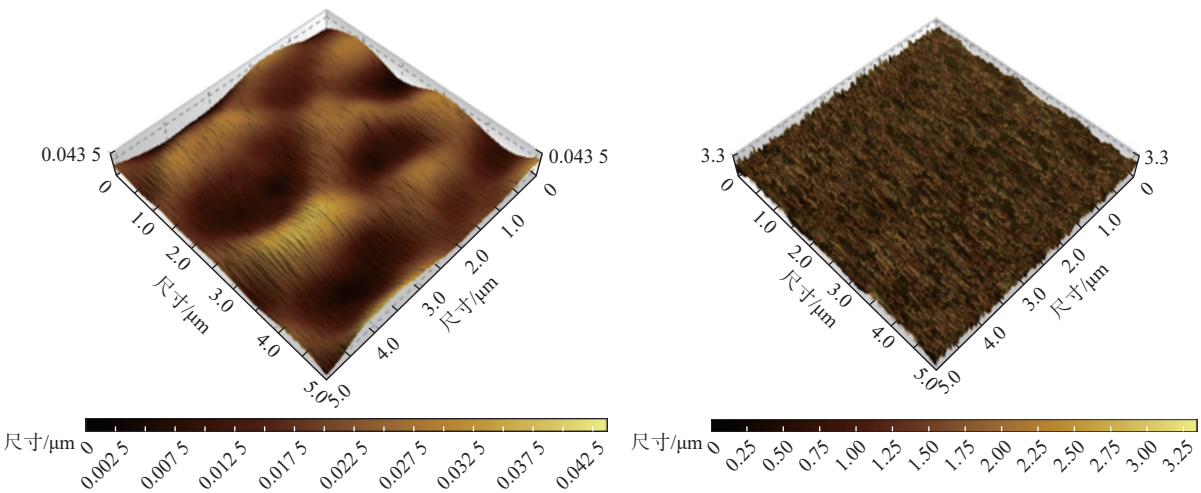


图 4　SMR 制备工艺流程

Fig. 4　Manufacturing process of SMR



(a) 未采用CMP处理的薄膜AFM测试图　　(b) 采用CMP处理后的薄膜AFM测试图

图 5　CMP 工艺对薄膜表面平整度的影响

Fig. 5　The CMP process influence for film surface smoothness

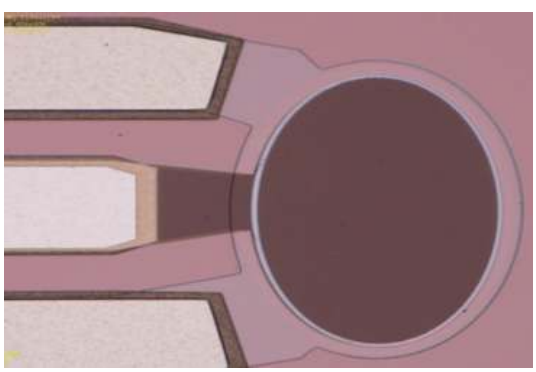


图 6　SMR 俯视图

Fig. 6　Top view of SMR

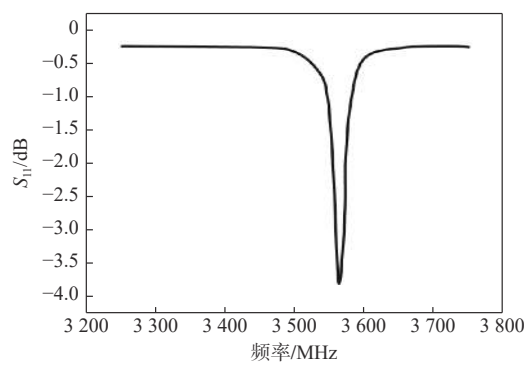


图 7　基于优化布拉格结构的 SMR 的实测 S_{11} 参数

Fig. 7　Measured S_{11} parameter of SMR based on optimized Bragg structure

SMR 的 S_{11} 参数(输入反射系数)测试结果。测试结果表明, 基于优化布拉格结构的 SMR 性能优于传统布拉格结构的 SMR, 并联谐振点的品质因数提高了 30%。

4 结 论

1）改进布拉格堆叠结构，将纵波和剪切波同时被约束在压电堆叠结构中，减少剪切波的衬底泄露，对提高器件品质因数至关重要；

2）优化布拉格结构顶层薄膜厚度，使器件色散特性由II型变为I型，可以抑制谐振腔的杂散模式，减少在谐振点上的寄生振动；

3）通过在W与SiO_2这2种材料间沉积钛镍合金黏附层，采用CMP工艺对布拉格反射结构抛光等工艺方法，可以提升SMR的生长质量。

4）通过SMR制备和测试表明，基于优化布拉格结构的SMR性能得到了明显改善。

参考文献（References）

[1] PIRINEN P, PENNANEN H, POUTTU A, et al. RF driven 5G system design for centimeter waves[J]. Wireless Communications and Mobile Computing, 2018, 2018: 1-9.

[2] PANARIELLO A, YU M, ERNST C. Ku-band high power dielectric resonator filters[J]. IEEE Transactions on Microwave Theory and Techniques, 2013, 61(1): 382-392.

[3] HAGELAUER A, FATTINGER G, RUPPEL C C W, et al. Microwave acoustic wave devices: Recent advances on architectures, modeling, materials, and packaging[J]. IEEE Transactions on Microwave Theory and Techniques, 2018, 66(10): 4548-4562.

[4] VILLA-LÓPEZ F H, THOMAS S, COLE M, et al. Finite element modelling of particle sensors based on solidly mounted resonators [C]//IEEE Sensors. Piscataway: IEEE Press, 2014: 574-577.

[5] FARINA M, ROZZI T. Electromagnetic modeling of thin-film bulk acoustic resonators[J]. IEEE Transactions on Microwave Theory and Techniques, 2004, 52(11): 2496-2502.

[6] SHEALY J B, VETURY R, GIBB S R, et al. Low loss, 3.7 GHz wideband BAW filters, using high power single crystal AIN-on-SiC resonators[C]//IEEE Mtt-S International Microwave Symposium. Piscataway: IEEE Press, 2017: 1476-1479.

[7] FATTINGER G G, MARKSTEINER S, KAITILA J, et al. Optimization of acoustic dispersion for high performance thin film BAW resonators[C]//IEEE Ultrasonics Symposium. Piscataway: IEEE Press, 2006: 1175-1178.

[8] THALHAMMER R, AIGNER R. Energy loss mechanisms in SMR-type BAW devices[C]//IEEE Mtt-S International Microwave Symposium Digest. Piscataway: IEEE Press, 2021: 225-228.

[9] TANSKAYA T N, ZIMA V N, KOZLOV A G. The influence of surface roughness of Bragg reflector layers on characteristics of microwave solidly mounted resonator[C]// Proceedings of the International Siberian Conference on Control and Communications. Piscataway: IEEE Press, 2015: 1-4.

[10] PARK H, SONG I, LEE M C, et al. Novel Bulk acoustic wave resonator structure for reducing electrical loss of electrodes[C]// Proceedings of the IEEE Mtt-S International Microwave Symposium Digest. Piscataway: IEEE Press, 2014: 1-4.

[11] LEE S H, YOON K H, LEE J K. Influence of electrode configurations on the quality factor and piezoelectric coupling constant of solidly mounted bulk acoustic wave resonators[J]. Journal of Applied Physics, 2002, 92(7): 4062-4069.

[12] KNITTL Z. Optics of thin films[M]. New York: Wiley, 1976: 66-67.

[13] VILLA-LÓPEZ F H, RUGHOOBUR G, THOMAS S, et al. Design and modelling of solidly mounted resonators for low-cost particle sensing[J]. Measurement Science and Technology, 2016, 27(2): 025101.

[14] GALIPEAU J, CHANG R E. Design considerations for high power BAW duplexers for base station applications[C]// Proceedings of the IEEE International Ultrasonics Symposium. Piscataway: IEEE Press, 2015: 1-4.

[15] KOZLOV A G. Effect of Bragg acoustic reflector on thermal processes in SMR under high power levels[C]// Proceedings of the 22nd International Conference on Thermal, Mechanical and Multi-Physics Simulation and Experiments in Microelectronics and Microsystems. Piscataway: IEEE Press, 2021: 1-6.

Solidly mounted resonator based on optimized Bragg structure

ZHANG Shifeng[1, 2], XUAN Weipeng[3], SHI Linhao[4], DONG Shurong[2], PU Shiliang[1, *]

(1. Hangzhou Hikvision Digital Technology Co., Ltd., Hangzhou 310051, China;
2. College of Information Science & Electronic Engineering, Zhejiang University, Hangzhou 310058, China;
3. College of Electronics & Information, Hangzhou Dianzi University, Hangzhou 310018, China;
4. Tiantong Ruihong Technology Co., Ltd., Jiaxing 314499, China)

Abstract: The effective coupling coefficient and the quality factor of the bulk acoustic wave resonators determine the overall performance of bulk acoustic wave filters. The effective coupling coefficient is dependent on the layer stack structure, especially the piezoelectric material, while the quality factor is highly dependent on the loss mechanism, including the electrical and acoustical loss. For solidly mounted resonator (SMR), the acoustic losses mainly include the acoustic leakage to the substrate. In this work, to enhance the quality factor of the SMR, the structure of the Bragg layer is optimized to reduce the acoustic energy leakage to the substrate. Through optimizing the Bragg stack, the longitudinal and shear waves can be confined in the piezoelectric stack simultaneously, and the quality factor at the anti-resonance frequency is highly improved. Moreover, to suppress the spurious mode of the resonator, the thickness of the top layer of the Bragg is optimized to change the device dispersion characteristic from type II to type I. Experimental results show that the performance of the SMR is improved greatly based on the optimized Bragg structure.

Keywords: bulk acoustic wave; solidly mounted resonator; Bragg; shear wave; high quality factor

Received: 2022-05-29; **Accepted**: 2022-07-02; **Published Online**: 2022-08-26 10:14
URL: link.cnki.net/urlid/11.2625.V.20220825.1648.002

* **Corresponding author**. E-mail: pushiliang@hikvision.com

http://bhxb.buaa.edu.cn jbuaa@buaa.edu.cn
DOI: 10.13700/j.bh.1001-5965.2022.0400

基于差异化和空间约束的自动图像描述模型

姜文晖，陈志亮，程一波，方玉明*，左一帆
(江西财经大学 信息管理学院，南昌 330032)

摘 要：多头注意力机制是图像描述模型的常用方法，该机制通过多分支结构构建关于输入特征的独特属性，以提高特征模型的区分性。然而，不同分支的独立性导致建模存在冗余性。同时，注意力机制会关注于不重要的图像区域，导致描述的文本不够准确。针对上述问题，提出一种损失函数作为训练目标的正则化项，以提高多头注意力机制的多样性和准确性。在多样性方面，提出一种多头注意力的差异化正则，鼓励多头注意力机制的不同分支关注于所描述目标的不同部件，使不同分支的建模目标变得简单。同时，不同分支相互融合，最后形成完整且更有区分性的视觉描述。在准确性方面，设计一种空间一致性正则。通过建模多头注意力机制的空间关联，鼓励注意力机制关注的图像区域尽可能集中，从而抑制背景区域的影响，提高注意力机制的准确性。提出差异化正则和空间一致性正则共同作用的方法，最终提升自动图像描述模型的准确性。所提方法在MS COCO数据集上对模型进行验证，并与多种代表性工作进行对比。实验结果表明：所提方法显著地提高了图像描述的准确性。

关 键 词：多头注意力机制；图像描述；差异性；空间约束；模态融合
中图分类号：TP391
文献标志码：A **文章编号**：1001-5965(2024)02-0456-10

图像描述任务是一项重要的跨学科挑战，其在车辆辅助驾驶、医疗影像报告生成和跨媒体数据管理等领域具有广泛应用。为了生成更为丰富的图像描述，需要深入挖掘图像中物体的类别、属性和关系等信息，并以流畅自然的语言准确描述这些内容。这一任务涉及计算机视觉和自然语言处理的交叉领域，推动人工智能技术在实际应用中取得更大突破，具有重要研究意义。

自动图像描述的主要实现框架是“编码器-解码器”结构[1-3]。其中，编码器分析输入图像的语义内容，形成一组图像特征向量；解码器输入这些特征向量，通过语言生成模型输出完整的图像描述，实现图像理解到语义表达。近期，受注意力机制在机器翻译成功应用的启发，大量研究者基于注意力机制设计自动图像描述模型[4]。该机制能够动态地改变输入特征的权重以指导不同时刻文本的生成。早期的研究工作采用加性的注意力模型，其主要采用单层或多层的长短期记忆(long short-term memory, LSTM)网络结构估计预测单词在输入特征上的权重分布[4-5]。随后，基于Transformer模型的结构[6]被广泛应用于自动图像描述任务中。该结构采用多头注意力机制，其每个分支先独立地预测输入特征的权重，后聚合不同分支的输出以指导文本生成。Transformer模型的提出极大地提高了图像描述模型的准确性。然而，现有研究工作仍然存在以下不足：①尽管多头注意力机制中的多分支结构可以构建关于输入特征的独特属性，然而，不同分支独立建模难以保证多头注意力机制的各个

收稿日期：2022-05-21；**录用日期**：2022-07-29；**网络出版时间**：2022-10-12 09:44
网络出版地址：link.cnki.net/urlid/11.2625.V.20221011.0935.001
基金项目：国家自然科学基金 (62161013,62162029)；江西省重点研发计划项目 (20203BBE53033)；江西省自然科学基金 (20224BAB212010, 20212BAB202011, 20224BAB212012,20232BAB202001)
* **通信作者**. E-mail：fa0001ng@e.ntu.edu.sg

引用格式：姜文晖，陈志亮，程一波，等. 基于差异化和空间约束的自动图像描述模型 [J]. 北京航空航天大学学报，2024，50（2）：456-465. JIANG W H，CHEN Z L，CHENG Y B，et al. Image captioning model based on divergence-based and spatial consistency constraints [J]. Journal of Beijing University of Aeronautics and Astronautics，2024，50（2）：456-465 （in Chinese）.

分支的多样性。潜在的冗余性难以最大化多头注意力机制的区分性。例如, Voita 等[7] 的研究发现, 删除多头注意力机制中的部分分支不会影响模型的准确性。当前, 已有一些工作尝试提高多头注意力在机器翻译任务中的多样性[8-9], 但这些工作主要针对语言特性展开研究, 面向图像描述任务的研究尚不充分。②部分研究工作通过可视化分析和量化分析, 发现注意力机制普遍存在“不聚焦”的问题[10]。具体地, 在生成描述的单词时, 注意力机制有时会关注在图像中不重要的区域(例如物体的背景), 导致“幻想”出与目标相关但未出现的物体[11]。已有部分学者研究提高注意力机制准确性的方法[12-13], 但是这些工作的研究对象是基于 LSTM 的注意力模型, 鲜有工作改进基于 Transformer 的注意力模型的准确性。

为解决以上问题, 本文提出一种目标函数作为训练目标的正则化项, 以同时提高多头注意力机制的多样性和准确性。在提高多头注意力机制的多样性方面, 本文受基于部件的模型(part-based models, PM)启发[14], 提出一种多头注意力的差异化正则(divergence-based regularization, DBR), 鼓励多头注意力机制的不同分支关注图像中所描述目标的不同部件, 从而使不同分支的建模目标变得简单。多头注意力机制中的不同分支互补, 融合后得到更加完整且更有区分性的建模, 从而生成更加准确的视觉描述。在提高多头注意力机制的准确性方面, 受图像描述任务中人类视觉注意力机制的研究成果启发[15], 人类的视觉注意力相对集中在被描述物体的周围。具体地, 注意力权重较大的图像区域, 其附近区域被关注的程度也相对较大。为此, 本文设计一种空间一致性正则(spatial consistency regularization, SCR)。通过建模多头注意力机制的空间关联性, 引入空间自约束提高多头注意力机制的准确性。本文提出 DBR 和 SCR 共同作用的方法, 最终提升自动图像描述模型的准确性。正则化项不包含任何参数, 仅在训练过程中起作用。因此, 本文方法不影响模型预测过程的时间复杂度和空间复杂度。

为验证本文方法的有效性, 本文在 MS COCO[16] 数据集上开展实验, 并与 15 种代表性研究工作进行对比。结果表明, 本文方法效果显著, 在 BLEU、CIDEr 等多个评价指标取得最优的性能。

1　相关工作

1.1　自动图像描述

自动图像描述任务主要以“编码器-解码器”为主要架构。编码器提取输入图像的语义特征, 解码器则对编码器的输出结果进行处理, 形成完整的文本描述[17]。鉴于深度神经网络的灵活性和较强的建模能力, 当前主要工作基于深度神经网络设计编码器和解码器的结构。其中, 编码器广泛采用卷积神经网络(convolutional neural network, CNN)进行图像的特征表示, 解码器则广泛采用循环神经网络(recurrent neural network, RNN)和 LSTM 对较长的文本序列进行关联建模[2-3]。相比上述结构 Transformer 不依赖于文本规则, 生成的图像描述语法灵活。

随着注意力机制在机器翻译领域的广泛应用, 越来越多学者将注意力机制引入“编码器-解码器”的结构中。Xu 等[4] 较早地将注意力机制引入自动图像描述任务中, 提出软注意力机制, 通过隐状态估算图像中不同空间特征的权重分布情况, 使每一时刻的文本预测都能自适应地关注图像中的不同区域, 从而提高下一时刻文本预测的准确性。然而, 注意力机制所学习的权重在模型中是隐变量, 缺少显式的监督信息指导, 导致注意力机制普遍存在“离焦”的问题[10]。为了解决该问题, Zhou 等[12] 提出基于图像和文本的匹配模型进行知识蒸馏以提高注意力机制的定位能力。Ma 等[13] 提出对预测单词重建作为对注意力机制的规则化, 以避免注意力机制关注于不相关的图像区域。Ji 等[18] 提出一种时空记忆力机制, 使注意力机制在序列预测的过程中同时建模时空关联。Zha 等[19] 通过建模视觉注意力在单词生成过程的上下文信息, 使注意力机制关注更加复杂的物体关系和细节。Zhang 等[20] 通过视觉图模型和语言图模型的对齐操作提高注意力机制的准确性。尽管这些方法都一定程度地改善了注意力机制, 然而, 其准确性仍远低于预期效果。

部分学者从其他角度改善图像描述的准确性。例如, Yang 等[21] 发现当前的图像描述模型因数据集偏置而错误构建了图像和文本信息的关联, 并从因果关系角度提出一种去混杂模型缓解该问题。Pan 等[22] 引入视觉和语言信息的高阶交互以进一步提高图像描述的准确性。Ji 等[23] 在描述图像的局部特征基础上引入图像的全局特征, 从而提高了编码器的区分性。Guo 等[24] 设计了一种多层级文本解码结构, 不断精化预测的文本信息, 提高了解码器的有效性。另外, Nguyen 等[25] 通过图像的场景图引入图像的结构化语义信息, 但对图像描述性能的改善不明显。不同于以上方法, 本文通过改善注意力机制的准确性以提高图像描述的效果。

1.2 多头注意力机制

基于Transformer的模型结构在自然语言处理相关的任务中得到广泛应用，近期也被成功应用于自动图像描述、图像分类等计算机视觉相关的研究任务中，并取得了引人注目的成绩。该结构的主要优势在于利用多头注意力机制，使多头注意力机制的不同分支关注于输入序列的不同表示子空间，从而获取关于输入的不同特性，最后集成为关于输入的鲁棒且全面的特征表达。然而，如图1所示，传统的多头注意力机制并未有效地建模不同分支的多样性。例如，Voita等[7]在研究中发现，在多数描述任务中，多头注意力机制仅有少数分支对最终形成的特征表示更重要，其他分支即使删除也不会影响最终模型的性能。为克服以上问题，一些研究改进多头注意力机制以提高模型的建模能力。Li等[9]提出差异性得分，通过余弦相似性测量2个分支输出特征表示的差异性，以期望得到最大化多分支的差异性，并在自然语言处理的相关任务中取得了较好的性能。然而，差异性得分的缺陷在于高维空间下任意向量之间的余弦距离接近于1，导致高维空间下该度量方式不具有区分性。Yang等[8]通过考虑多分支之间的交互信息从而提高多头注意力机制的区分性。Shen等[26]提出利用多头注意力机制显式地建模相同文本的不同依赖关系，从而提高多头注意力机制的多样性。Strubell等[27]利用多头注意力机制建模不同语言之间的特性以提高注意力机制的多样性，并在自然语言处理的相关任务验证了其方法的有效性。然而，以上研究主要集中在自然语言处理任务，鲜有研究多头注意力机制在自动图像描述任务上的区分性。

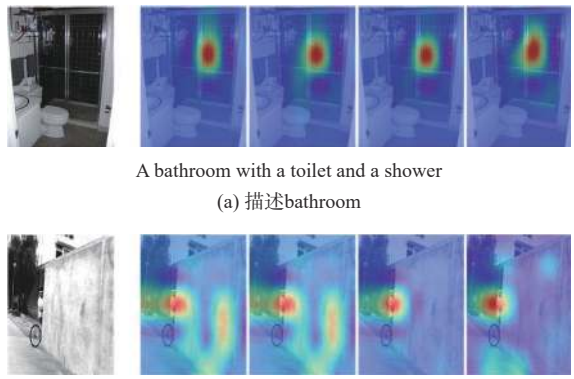


图1 多头注意力机制不同分支的可视化效果

Fig. 1 Visualization of the different branches of attention results from the multi-head attention mechanism.

2 本文方法设计

本文方法的整体结构如图2所示，其中，$\boldsymbol{V}$、$\boldsymbol{K}$和$\boldsymbol{Q}$分别为值矩阵、键矩阵和查询矩阵；i为第i个分支；$\tilde{y}_{1:t-1}$为输入的前t-1个单词；$\tilde{y}_t$为解码器预测输出的第t个单词。对于输入图像$\boldsymbol{I}$，其对应的语言描述为$\boldsymbol{y}_{1:T}$，其中T为文本描述的最大长度。$\boldsymbol{I}$经过卷积神经网络抽取图像的网格特征，其中每个特征向量对应于原始图像特定区域的高层语义表示。将网格特征扁平化排列后，通过自注意力机制进一步增强，最终得到图像的视觉特征编码$\boldsymbol{X}=[\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_N]$，其中$\boldsymbol{x}_i$为特征向量，$N$为网格特征的总数。解码器则基于$\boldsymbol{X}$预测描述图像内容的语句。其中，解码器的核心是跨模态注意力模块，该模块动态地改变输入特征的权重，从而对不同时间步形成不同的特征表示以指导文本生成。文本方法对跨模态注意力进行约束，使该模块更加准确地关注于具有区分性的输入特征。首先介绍多头注意力机制的结构，随后介绍提高多头注意力机制

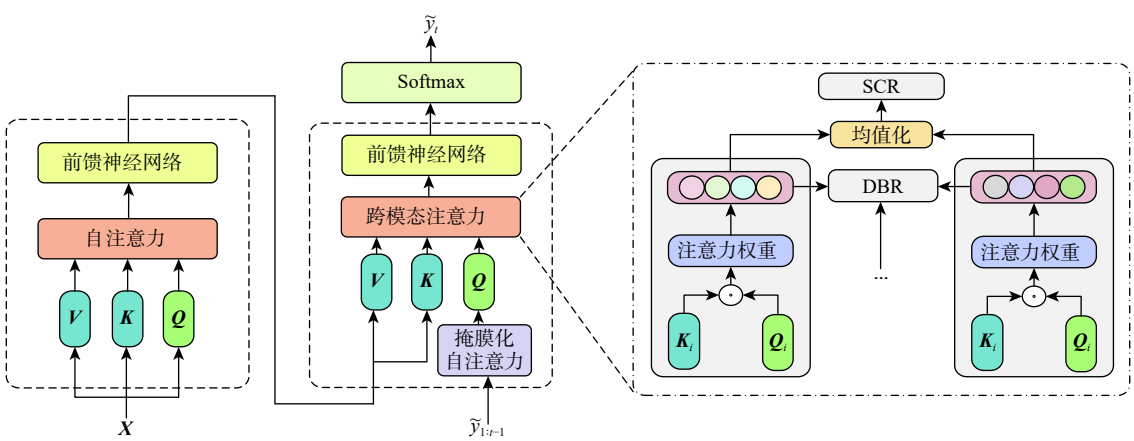


图2 本文方法的整体结构

Fig. 2 Overall framework of the proposed method

多样性和准确性的正则化方法, 最后介绍模型的学习策略。

2.1　多头注意力结构

本文提出的解码器的基本结构是标准的 Transformer 解码器[6]。其包含 1 个跨模态注意力模块和 1 个文本生成模块。其中, 跨模态注意模块通过基于点乘的注意力机制建模文本与图像之间的跨模态关联。该注意力机制以查询矩阵$\boldsymbol{Q}$、键矩阵$\boldsymbol{K}$和值矩阵$\boldsymbol{V}$为输入, 其中查询矩阵$\boldsymbol{Q}$由 M 个 d 维向量构成, 键矩阵$\boldsymbol{K}$和值矩阵$\boldsymbol{V}$由 N 个 d 维向量构成。在图像描述任务中, 以文本序列编码矩阵 $\boldsymbol{Y}$ 和视觉特征编码 $\boldsymbol{X}$ 为输入, 跨模态注意力模块首先将 $\boldsymbol{X}$ 和 $\boldsymbol{Y}$ 通过线性映射形成查询矩阵 $\boldsymbol{Q}$、键矩阵 $\boldsymbol{K}$ 和值矩阵 $\boldsymbol{V}$, 即

$$\begin{aligned}\boldsymbol{Q}_i &= \mathbf{W}_i^q \boldsymbol{Y}\\ \boldsymbol{K}_i &= \mathbf{W}_i^k \boldsymbol{X}\\ \boldsymbol{V}_i &= \mathbf{W}_i^v \boldsymbol{X}\end{aligned} \tag{1}$$

式中: $\boldsymbol{Q}_i \in \mathbf{R}^{M\times d}$ 为第 i 个分支的查询矩阵; $\boldsymbol{K}_i \in \mathbf{R}^{N\times d}$ 为第 i 个分支的键矩阵; $\boldsymbol{V}_i \in \mathbf{R}^{N\times d}$ 为第 i 个分支的值矩阵; $\boldsymbol{W}_i^{\boldsymbol{Q}}$、$\boldsymbol{W}_i^{\boldsymbol{K}}$和$\boldsymbol{W}_i^{\boldsymbol{V}}$为第 i 个分支的映射矩阵。

通过计算查询矩阵与键矩阵之间的相似性, 预测在 N 个不同的值向量上的权重矩阵, 计算式为

$$\boldsymbol{\alpha}_i = A(\boldsymbol{Q}_i, \boldsymbol{K}_i) = \mathrm{softmax}\left(\frac{\boldsymbol{Q}_i \boldsymbol{K}_i^{\mathrm{T}}}{\mathrm{sqrt}(d)}\right) \tag{2}$$

式中: sqrt 为求根函数; softmax 函数作用为将数值归一化为概率分布向量; $\boldsymbol{\alpha}_i = [a_{i,1}, a_{i,2}, \cdots, a_{i,N}]$为第 i 个分支的注意力权重, 描述了不同的值向量对应的注意力权重。较大的权重表示对应的值向量与查询的相关性更大。结合权重矩阵和值矩阵, 对不同的值向量加权融合, 得到经过注意力机制聚合后的向量:

$$\boldsymbol{Z}_i = A(\boldsymbol{Q}_i, \boldsymbol{K}_i)\boldsymbol{V}_i \tag{3}$$

多头注意力机制将 h 个不同的分支形成的向量表示拼接后, 投影形成最后的特征表示:

$$\boldsymbol{Z}_{\mathrm{att}} = \mathrm{concat}(\boldsymbol{Z}_1, \boldsymbol{Z}_2, \cdots, \boldsymbol{Z}_h)\boldsymbol{W}_h \tag{4}$$

该特征表示后续通过解码器(Dec)预测输出的第 t 个单词:

$$\tilde{y}_t = \mathrm{Dec}(\boldsymbol{Z}_{\mathrm{att}}) \tag{5}$$

式中: $\tilde{y}_t$表示解码器(Dec)预测输出第 t 个单词。

2.2　差异化正则

由式(1)～式(5)可知, Transformer 的结构中未显式地引入保证多头注意力多样性的机制。因此, 多头注意力中的不同分支可能关注于图像的相似区域, 造成分支的冗余, 从而降低了注意力机制的表达能力。因此, 本文设计了一种提高多头注意力机制差异性的损失函数, 以期提高模型的区分能力。受 PM[14] 的启发, 本文鼓励多头注意力机制将不同分支的注意力权重关注到所描述同一目标的不同部件上。一方面, 将每个分支的建模对象由“物体”降解为“部件”, 能使各分支的建模目标变得简单; 另一方面, 各个分支相互补充, 形成关于描述目标的更有区分性的特征表达。例如, 在描述名词“person”时, 各分支分别关注于图像中人的脸部、身体和四肢等不同部件, 从而使各分支具有独特性; 同时, 将不同分支建模的特征融合后形成关于名词“person”的完整特征表示, 从而特征更具区分性。为实现该目标, 本文基于 KL 散度提出一种多头注意力的 DBR, 鼓励多头注意力机制不同分支的注意力分布具有不一致性。如图 1 所示, 对于第 i 个分支和第 j 个分支形成的注意力向量$\boldsymbol{\alpha}_i$和$\boldsymbol{\alpha}_j$, 通过 KL 散度评估两者的不一致性:

$$D_{\mathrm{KL}}(\boldsymbol{\alpha}_i, \boldsymbol{\alpha}_j) = \boldsymbol{\alpha}_i(\ln \boldsymbol{\alpha}_i - \ln \boldsymbol{\alpha}_j) \tag{6}$$

本文通过计算所有注意力分支的注意力差异性, 并使其差异性最大化。因此, 本文设计的正则化项如下所示:

$$R_{\mathrm{DBR}} = -\frac{1}{h^2}\sum_{i=1}^{h}\sum_{j=1}^{h} D_{\mathrm{KL}}(\boldsymbol{\alpha}_i, \boldsymbol{\alpha}_j) \tag{7}$$

2.3　空间一致性正则

人类视觉注意力机制 AM 的多项研究证明, 人类在进行图像描述任务的过程中, 视觉注意力相对集中在被描述物体的周围。注意力权重较大的图像区域, 其附近区域被关注的程度也相应较大, 反之亦然。因此, 本文进一步提出 SCR, 使注意力机制关注的图像区域尽可能集中, 从而抑制其他无关区域(如背景)对图像描述结果的影响。

如图 1 所示, 对于多头注意力机制形成的注意力向量$\{\boldsymbol{\alpha}_i,\ i=1, 2, \cdots, h\}$, 本文对其均值化, 得到多头注意力机制最终输出的平均权重向量$\bar{\boldsymbol{a}}$:

$$\bar{\boldsymbol{a}} = \frac{1}{h}\sum_{i=1}^{h}\boldsymbol{\alpha}_i \tag{8}$$

其中, $\bar{\boldsymbol{a}}$的每个权重对应于输入图像特定空间区域的权重。将$\bar{\boldsymbol{a}}$中每个权重通过空间映射变换为 $w\times h$的注意力图, 记为$\bar{\boldsymbol{G}}$。随后, 定位注意力权重最大值点在$\bar{\boldsymbol{G}}$中的坐标:

$$(i', j') = \mathrm{argmax}\left(\bar{\boldsymbol{G}}\right) \tag{9}$$

本文期望空间注意力约束使 (i', j') 近邻处的注意力权重与$\bar{\boldsymbol{G}}(i', j')$接近, 因此, 本文定义如下的 SCR:

$$R_{\mathrm{SCR}}=\sum_{i=1}^{w}\sum_{j=i}^{h}\mathrm{e}^{-\mu * r}\|\bar{\boldsymbol{G}}(i',j')-\bar{\boldsymbol{G}}(i,j)\| \tag{10}$$

式中：r为(i',j')与(i,j)的空间距离；μ为控制空间注意力约束的影响范围，本文将μ设为1。由式(10)可见，R_{SCR}使邻近位置的注意力权重尽可能接近。同时，由于不同位置的注意力权重之间具有互斥性，促使较远位置的注意力权重更小。

2.4 学习策略

本文采用图像描述的常用训练方法[28]，即将训练过程分为2个阶段。第1阶段对每个时刻生成的单词采用交叉熵损失(cross-entropy loss, CL)函数进行训练，第2阶段采用强化学习对描述生成的模型调优。

在以CL函数为目标的训练阶段，模型输入真实文本$y_{1:t-1}$，预测与之对应的下一单词。记模型的参数为θ，采用如下最大似然估计，以最大化真实单词y_t的预测概率：

$$L_{\mathrm{XE}}(\theta)=-\sum_{t=1}^{T}\ln(p_\theta(y_t|y_{1:t-1})) \tag{11}$$

式中：L_{XE}为交叉熵损失函数；p_θ为参数θ下的概率。

在第1个训练阶段，本文引入正则化项指导注意力机制获取具有区分性的特征。最终训练的损失函数表示为

$$L_{\mathrm{all}}(\theta)=L_{\mathrm{XE}}(\theta)+\lambda(R_{\mathrm{DBR}}+R_{\mathrm{SCR}}) \tag{12}$$

值得注意的是，正则化项不包含任何参数，仅在训练过程中起作用，因此，不影响模型预测过程的复杂度。λ是正则化项系数，经验地，本文将λ设为1。

第2阶段，本文以第1阶段训练得到的θ作为初始值，以SCST强化学习[28]进一步优化文本描述的评价指标。强化学习的目标是最小化负期望奖励：

$$L_{\mathrm{RL}}(\theta)=-E_{\tilde{y}_{1:T}\sim p_\theta}[r(\tilde{y}_{1:T})] \tag{13}$$

式中：L_{RL}为强化学习损失函数；$r()$为奖励函数，本文用语言评价指标CIDEr描述；$\tilde{y}_{1:T}$为模型依据p_θ使用蒙特卡罗采样生成的句子描述。

在序列的预测过程中，本文采用集束搜索(beam search, BS)策略，即每个时刻从解码器的概率分布中采样概率最大的前k个单词，并在解码过程中始终保留置信度最高的前k个文本序列。最终将置信度最高的序列作为预测的文本描述。

3 实验结果与分析

3.1 数据集和评估指标

本文在MS COCO[16]公开数据集对图像描述模型进行评价。该数据集包含123 287张图片，每张图片均提供5条参考描述。本文采用文献[29]提出的训练集和测试集的划分方法进行模型的训练和评估。为评估模型生成图像描述的质量，本文采用标准的图像描述评估标准验证模型的效果，即：BLEU-1、BLEU-4、METEOR和CIDEr。

3.2 实施细节

本文基于深度学习框架Pytorch实现所述模型，模型的训练和测试均使用2080TI GPU。在图像的编码器部分，采用Jiang等[30]提出的方法 抽取图像的网格特征，其中网格的大小为7×7，每个特征表示为2 048维的向量。文本的编码采用标准的词嵌入模型。参照Transformer的一般设置，将维度d设为512，前馈神经网络(feedforward neural network, FNN)的隐藏层特征维度设为2 048，随机失活(dropout)的概率为0.1。本文采用ADAM优化器进行模型训练，批处理的大小(Batch Size)设为50。在交叉熵学习阶段，初始学习率设为0.000 5，学习率的变化过程参照模型训练的一般设置[31]。如果训练过程中，验证集的CIDEr连续下降5个周期(epoch)，则进入强化学习阶段。在强化学习阶段，学习率固定为5×10^{-6}。当验证集的CIDEr连续下降5个训练周期后，模型训练结束。在测试过程中，集束搜索中k值设为5。

3.3 消融实验与分析

为了验证DBR和SCR的有效性，本文设计了消融实验。本节对比的结果均为交叉熵学习阶段完成后模型的性能。

消融实验结果如表1所示。在MS COCO数据集上，DBR和SCR都有效地提高了图像描述的准确性。在MS COCO数据集上，DBR使CIDEr指标提高了1.0，SCR使CIDEr指标提高了0.6。同时，2种正则化方法互补，共同作用使CIDEr指标提高了1.4，显著高于基准Transformer方法。其他各项测评指标也有一定程度的提升。本文在Transformer的框架下也对比了其他提高多头注意力机制多样性的约束，包括余弦差异性(Cosine)[9]和谱有限等距性(spectral restricted isometry property, SRIP)[32]。不同于本文提出的多样性约束，余弦相似性和SRIP旨在最大化多分支模型形成的特征表示的差异性。由表1可见，余弦差异性和SRIP均一定程度地降低了图像描述的准确性。由此可见，最大化多分支特征表示的差异性不能保证同时优化图像描述的准确性。

本文还分析了多头注意力分支数量的变化对图像描述性能的(CIDEr)影响，其中维度d在实验中保持512不变。结果如表2所示。当分支数量

表 1　不同的正则化方法对图像描述性能的影响

Table 1　The effect of different regularization methods on image description performance

方法	BLEU-1	BLEU-4	CIDEr	METEOR
Transformer	76.4	36.4	116.9	28.3
+Cosine	76.0	36.4	115.4	28.0
+SRIP	76.4	36.7	116.5	28.3
+DBR	**76.5**	36.8	117.9	28.3
+SCR	76.4	**36.9**	117.5	**28.4**
本文方法	**76.5**	**36.9**	**118.4**	**28.4**

注: 加粗字体表示各列最优结果。

表 2　不同分支数量对图像描述性能的影响

Table 2　The effect of different number of branches on image description performance

h	性能	
	Transformer	+DBR
4	116.4	116.8
8	116.9	117.9
16	117	118.3
32	116.5	117.4

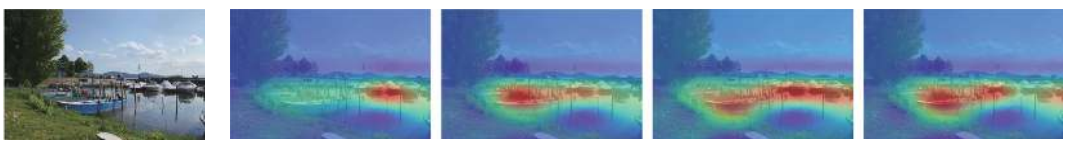
h 由 4 变化到 32 时, 基准模型的性能相对稳定, 意味着 h 增大后不同分支学习到的权重更有可能存在冗余性, 导致多头注意力机制没有学习到差异化的特征表示, 因而没有对图像描述带来性能提升。对比之下, 在增加了多样性损失函数之后, 图像描述的准确性显著提高。当 h 为 4 时, DBR 使 CIDEr 提高了 0.4; 当 h 为 16 时, DBR 使 CIDEr 提高了 1.3。这一结果充分验证了较大的 h 为多头注意力机制造成了冗余性, DBR 可以很好地降低冗余性, 因而在较大的 h 时带来的性能增益更加明显。

为进一步验证 DBR 对注意力机制的作用, 本文对多头注意力机制不同分支所关注的图像区域进行可视化分析。如图 3 所示, 标准的多头注意力机制的不同分支可能关注于图像的重复区域, 产生较大的冗余性, 导致输出的单词不够准确。当增加了 DBR 后, 不同分支关注于图像的不同区域, 最终聚合成为完整的对象表示, 提高了特征描述的完整性和区分性, 输出更加准确的描述。



输入图像　　生成描述: A boat sitting in the water

(a) Transformer方法



输入图像　　生成描述: A group of boats sitting in the water

(b) 增加DBR

图 3　多头注意力机制的可视化效果

Fig. 3　Visualization of multi-head attention mechanism

3.4　对比实验与分析

对比实验如表 3 所示, 本文挑选了 15 种业界具有代表性的方法与本文方法在 MS COCO 数据集上开展定量比较。这 15 种方法包括 Up-Down [5], Transformer [6], SCAN [12], AAT [33], RD [34], STMA [17], X-LAN [22], DIC [21], ORT [34], M2 [31], CaptionNet [35], TAA [36], GET [23], CAVG [19] 和 SG2Caps [25]。其中, Up-Down 和 Transformer 是基准方法; SCAN 和 X-LAN 通过修正注意力机制 AM 以提高图像描述的准确性; RD 和 AAT 引入解码器级联结构提高图像描述的性能; STMA 和 CAVP 引入时空约束改善注意力模型; ORT、M2 和 TAA 分别对 Transformer 的结构进行改进; GET 改善图像的特征表示提高图像描述的准确性; DIC 和 SG2Caps 分别是基于因果关联和场景图的代表性方法。为实现公平比较, 以上对比方法均采用 SCST 进行模型训练, 并仅对比单模型预测结果。

实验结果表明, 本文方法显著地改善了基于 Transformer 的基准模型, 也高于大部分业界具有代表性的对比方法。对于描述短语重叠率的评估指标, 本文方法 BLEU-1 指标比 GET 提高了 0.2, 比 X-LAN 提高了 0.9, 说明本文方法能精确地输出描述图像的单词; 对于描述语义相似性的指标 CIDEr, 本文方法相比 M2 提高了 0.8, 比 GET 提高了 0.4, 与 X-LAN 模型的性能一致, 说明本文方法能够很好地输出与人类主观描述一致的文本序列。对比 TAA、DIC、

表3 与主流方法在MS COCO测试集性能比较

Table 3 Comparisons of performance of this paper with mainstream methods on MS COCO test set

方法	BLEU-1	BLEU-4	METEOR	CIDEr
Up-Down	79.8	36.3	27.7	120.1
STMA	80.2	37.7	28.2	125.9
SCAN	80.2	38.0	28.5	126.1
AAT		38.7	28.6	128.6
RD		38.6	28.7	128.3
Transformer	80.1	38.8	28.7	127.2
ORT	80.5	38.6	28.7	128.3
M2	80.8	39.1	29.2	131.2
CaptionNet	80.4	38.5	28.8	127.6
TAA	78.6	37.1	27.5	119.6
X-LAN	80.8	39.7	29.5	132.0
DIC		38.7	28.4	128.2
GET	81.5	39.5	29.3	131.6
CAVP		38.6	28.3	126.3
SG2Caps		33.0	26.2	112.3
本文方法	81.7	39.3	29.2	132.0
ORT+DBR+SCR	81.0	39.3	29.1	131.3
M2 +DBR+SCR	81.1	39.5	29.9	131.8

SG2Caps、CAVP等不同方法，本文方法在各项指标上的性能也一致地高于对比方法，验证了本文方法的有效性。值得注意的是，M2、GET等方法改变了Transformer基准方法的结构，都在一定程度上增加了方法的计算复杂度。然而，本文方法没有引入额外的参数，也不会增加模型在预测时的时间复杂度。因此，与对比方法相比，本文方法计算效率更高。

为验证DBR和SCR的普适性，将本文方法应用于ORT和M2这2个基于Transformer的方法上。如表3所示，引入DBR和SCR后，ORT方法在CIDEr指标上提高了3.0，M2方法的CIDEr指标提高了0.6。BLEU-1、BLEU-4和METEOR指标也一致地高于对比方法，验证了本文方法的有效性。

3.5 可视化分析

图4为本文方法和Transformer基准方法在MS COCO部分测试集的测试效果对比。从描述结果中可以看出，基准方法错误预测了图像中的物体。引入DBR能够改善注意力机制关注更多有效的图像区域，从而使物体描述更加准确。例如，由图4(a)可知，DBR在描述人物时，会综合图像中人物不同部件的信息预测人物的性别，避免因数据集偏置[21]做出错误的预测。由图4(b)中可知，基准模型将银白色的车厢错判为“refrigerator”。引入DBR能够融合上下文信息从而正确预测“a subway train”。

进一步引入SCR，使注意力机制聚焦于图像的主体和目标物体的细节，从而生成更加细致的描述内容。例如，在图4(b)中，SCR能够突出图像的主体是镜头最前方的“a man”。在图4(c)中，引入SCR后，模型能够输出描述颜色的细节信息。以上可视化的分析结果从另一个角度验证了本文方法的有效性。



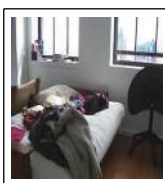
基准: A woman walking down a street holding an umbrella.
+DBR: A person in an orange vest holding an umbrella on the street.
+DBR+SCR: A person holding an umbrella in the street.
参考描述1: A man in orange garb carrying a umbrella and cell phone.
参考描述2: A monk holding an umbrella looking at a cellphone on a street.

(a) 本文方法与基准方法对比样例1



基准: A group of people sitting inside of a refrigerator.
+DBR: A group of people sitting on a subway train.
+DBR+SCR: A man standing in a subway train with people.
参考描述1: A People standing and sitting in a public transportation vehicle.
参考描述2: A group of people riding on a subway train.

(b) 本文方法与基准方法对比样例2



基准: A cat laying on a bed in a room.
+DBR: A dog laying on a bed in a room.
+DBR+SCR: A black dog laying on a bed in a room.
参考描述1: There is a dog sleeping on a futon.
参考描述2: A very cute dog laying on a big bed.

(c) 本文方法与基准方法对比样例3

图4 使用不同的正则化方法生成的图像描述效果对比

Fig. 4 Examples of captions generated by different regularization method

4　结　论

本文针对多头注意力机制潜在的冗余性和“不聚焦”等缺陷, 提出一种正则化方法, 同时提高多头注意力机制的多样性和准确性。

1) 本文方法鼓励多头注意力机制的不同分支关注于所描述目标的不同部件, 从而使不同分支的建模目标变得简单。不同分支相互融合, 最后形成完整且更有区分性的视觉描述。

2) SCR 通过建模多头注意力机制的空间关联, 引入空间自约束提高注意力机制的准确性。DBR 和 SCR 共同作用, 共同提升自动图像描述模型的准确性。

3) 在 MS COCO 数据集上对本文方法进行验证, 并与具有代表性的多种方法进行对比实验, 结果表明, 本文方法显著地提高了图像描述的准确性。

下一步的工作中, 将尝试改进图像的特征表示模型, 提出更有区分性的图像表示方法。同时, 构建物体之间的关联模型, 以提高自动图像描述模型对关系预测的鲁棒性。

参考文献 (References)

[1] WAN B Y, JIANG W H, FANG Y M, et al. Revisiting image captioning via maximum discrepancy competition[J]. Pattern Recognition, 2022, 122: 108358.

[2] 谭云兰, 汤鹏杰, 张丽, 等. 从图像到语言: 图像标题生成与描述[J]. 中国图象图形学报, 2021, 26(4): 727-750.
TAN Y L, TANG P J, ZHANG L, et al. From image to language: Image captioning and description[J]. Journal of Image and Graphics, 2021, 26(4): 727-750 (in Chinese).

[3] 石义乐, 杨文忠, 杜慧祥, 等. 基于深度学习的图像描述综述[J]. 电子学报, 2021, 49(10): 2048-2060.
SHI Y L, YANG W Z, DU H X, et al. Overview of image captions based on deep learning[J]. Acta Electronica Sinica, 2021, 49(10): 2048-2060 (in Chinese).

[4] XU K, BA J L, KIROS R, et al. Show, attend and tell: Neural image caption generation with visual attention[C]// Proceedings of the 32nd International Conference on International Conference on Machine Learning—Volume 37. New York: ACM, 2015: 2048-2057.

[5] ANDERSON P, HE X D, BUEHLER C, et al. Bottom-up and top-down attention for image captioning and visual question answering [C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2018: 6077-6086.

[6] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]// Proceedings of the 31st International Conference on Neural Information Processing Systems. New York: ACM, 2017: 6000-6010.

[7] VOITA E, TALBOT D, MOISEEV F, et al. Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned[C]// Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: Association for Computational Linguistics, 2019: 5797-5808.

[8] YANG B S, LI J A, WONG D F, et al. Context-aware self-attention networks[J]. Computer Science, 2019, 33(1): 387-394.

[9] LI J A, TU Z P, YANG B S, et al. Multi-head attention with disagreement regularization[C]// Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Stroudsburg: Association for Computational Linguistics, 2018: 2897-2903.

[10] LIU C X, MAO J H, SHA F, et al. Attention correctness in neural image captioning[J]. Computer Science, 2017, 31(1): 4176-4182.

[11] ROHRBACH A, HENDRICKS L A, BURNS K, et al. Object hallucination in image captioning[C]// Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Stroudsburg: Association for Computational Linguistics, 2018: 4035-4045.

[12] ZHOU Y E, WANG M, LIU D Q, et al. More grounded image captioning by distilling image-text matching model[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2020: 4777-4786.

[13] MA C Y, KALANTIDIS Y, ALREGIB G, et al. Learning to generate grounded visual captions without localization supervision[C]// Proceedings of the 16th European Conference, Glasgow. New York: ACM, 2020: 353-370.

[14] PAPANDREOU G, ZHU T, CHEN L C, et al. PersonLab: Person pose estimation and instance segmentation with a bottom-up, part-based, geometric embedding model[C]//Computer Vision – ECCV 2018. Berlin: Springer, 2018: 282-299.

[15] HE S, TAVAKOLI H R, BORJI A, et al. Human attention in image captioning: dataset and analysis[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE Press, 2020: 8528-8537.

[16] LIN T Y, MAIRE M, BELONGIE S, et al. Microsoft COCO: Common objects in context[C]//European Conference on Computer Vision. Berlin: Springer, 2014: 740-755.

[17] 毕健旗, 刘茂福, 胡慧君, 等. 基于依存句法的图像描述文本生成[J]. 北京航空航天大学学报, 2021, 47(3): 431-440.
BI J Q, LIU M F, HU H J, et al. Image captioning based on dependency syntax[J]. Journal of Beijing University of Aeronautics and Astronautics, 2021, 47(3): 431-440 (in Chinese).

[18] JI J Z, XU C, ZHANG X D, et al. Spatio-temporal memory attention for image captioning[J]. IEEE Transactions on Image Processing, 2020, 29: 7615-7628.

[19] ZHA Z J, LIU D Q, ZHANG H W, et al. Context-aware visual policy network for fine-grained image captioning[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 44(2): 710-722.

[20] ZHANG W Q, SHI H C, TANG S L, et al. Consensus graph representation learning for better grounded image captioning[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2021, 35(4): 3394-3402.

[21] YANG X, ZHANG H W, CAI J F. Deconfounded image captioning: A causal retrospect[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45(11): 12996-13010.

[22] PAN Y W, YAO T, LI Y H, et al. X-linear attention networks for

image captioning[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2020: 10968-10977.

[23] JI J Y, LUO Y P, SUN X S, et al. Improving image captioning by leveraging intra-and inter-layer global representation in transformer network[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2021, 35(2): 1655-1663.

[24] GUO L T, LIU J, LU S C, et al. Show, tell, and polish: Ruminant decoding for image captioning[J]. IEEE Transactions on Multimedia, 2020, 22(8): 2149-2162.

[25] NGUYEN K, TRIPATHI S, DU B, et al. In defense of scene graphs for image captioning[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE Press, 2022: 1387-1396.

[26] SHEN T, ZHOU T Y, LONG G D, et al. DiSAN: Directional self-attention network for RNN/CNN-free language understanding[C]// Proceedings of the AAAI Conference on Artificial Intelligence. Palo Alto: AAAI Press, 2018: 5446-5455.

[27] STRUBELL E, VERGA P, ANDOR D, et al. Linguistically-informed self-attention for semantic role labeling[C]// Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Stroudsburg: Association for Computational Linguistics, 2018: 5027–5038.

[28] RENNIE S J, MARCHERET E, MROUEH Y, et al. Self-critical sequence training for image captioning[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2017: 1179-1195.

[29] KARPATHY A, LI F F. Deep visual-semantic alignments for generating image descriptions[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2015: 3128-3137.

[30] JIANG H Z, MISRA I, ROHRBACH M, et al. In defense of grid features for visual question answering[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2020: 10264-10273.

[31] CORNIA M, STEFANINI M, BARALDI L, et al. Meshed-memory transformer for image captioning[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2020: 10575-10584.

[32] BANSAL N, CHEN X H, WANG Z Y. Can we gain more from orthogonality regularizations in training deep CNNs? [C]// Proceedings of the 32nd International Conference on Neural Information Processing Systems. New York: ACM, 2018: 4266-4276.

[33] HUANG L, WANG W M, XIA Y X, et al. Adaptively aligned image captioning via adaptive attention time[C]//Proceedings of the 33rd International Conference on Neural Information Processing Systems. New York: ACM,2019: 8942-8951.

[34] HERDADE S, KAPPELER A, BOAKYE K, et al. Image captioning: Transforming objects into words[C]//Proceedings of the 33rd International Conference on Neural Information Processing Systems. New York: ACM, 2019: 11137-11147.

[35] YANG L Y, WANG H L, TANG P J, et al. CaptionNet: A tailor-made recurrent neural network for generating image descriptions[J]. IEEE Transactions on Multimedia, 2021, 23: 835-845.

[36] YAN C G, HAO Y M, LI L, et al. Task-adaptive attention for image captioning[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2022, 32(1): 43-51.

Image captioning model based on divergence-based and spatial consistency constraints

JIANG Wenhui，CHEN Zhiliang，CHENG Yibo，FANG Yuming*，ZUO Yifan

(School of Information Management, Jiangxi University of Finance and Economics, Nanchang 330032, China)

Abstract：The multi-head attention mechanism has been widely adopted in image captioning. It is appealing for the ability to jointly attend to information from different representation subspaces. However, as each head captures distinct properties of the input individually, the diversity between heads' representations is not guaranteed. In the meanwhile, most existing attention models encounter the problem of "attention defocus", i.e., they fail to concentrate on correct image regions when generating the target words. Consequently, the generated sentences are not accurate enough. To address these problems, we propose a novel training objective that serves as an auxiliary regularization function to improve the diversity and accuracy of the multi-head attention mechanism. In the beginning, we present a divergence-based regularization that encourages each brain to concentrate on various areas of the goal. Partial representations are aggregated to produce distinct representations of the target. Secondly, we introduce a spatial consistency regularization that builds the spatial relationship among the attended regions. By encouraging the attended regions to be focussed, it enhances image captioning. We proposed a method for the joint action of divergence-based regularization and spatial consistency regularization. We compare the performance of the proposed method with state-of-the-art methods on challenging MS COCO datasets. The experimental results demonstrate the superior performance of the proposed method.

Keywords：multi-head attention mechanism; image captioning; diversity; spatial consistency; model fusion

Received：2022-05-21；**Accepted**：2022-07-29；**Published Online**：2022-10-12 09：44

URL：link.cnki.net/urlid/11.2625.V.20221011.0935.001

Foundation items：National Natural Science Foundation of China (62161013,62162029); Key R & D Program of Jiangxi Province (20203BBE53033); Natural Science Foundation of Jiangxi Province (20224BAB212010,20212BAB202011,20224BAB212012,20232BAB202001)

* **Corresponding author**. E-mail：fa0001ng@e.ntu.edu.sg

http://bhxb.buaa.edu.cn jbuaa@buaa.edu.cn
DOI: 10.13700/j.bh.1001-5965.2022.0456

融合卷积注意力和Transformer架构的行人重识别方法

王静，李沛橦，赵容锋，张云*，马振玲
(上海海洋大学 信息学院，上海 201306)

摘 要：行人重识别技术是智能安防系统中的重要方法之一，为构建一个适用各种复杂场景的行人重识别模型，基于现有的卷积神经网络和Transformer模型，提出一种融合卷积注意力和Transformer（FCAT）架构的方法，以增强Transformer对局部细节信息的关注。所提方法主要将卷积空间注意力和通道注意力嵌入Transformer架构中，分别加强对图像中重要区域的关注和对重要通道特征的关注，以进一步提高Transformer架构对局部细节特征的提取能力。在3个公开行人重识别数据集上的对比消融实验证明，所提方法在非遮挡数据集上取得了与现有方法相当的结果，在遮挡数据集上的性能得到显著提升。所提方法更加轻量化，在不增加额外计算量和模型参数的情况下，推理速度得到了提升。

关 键 词：行人重识别；深度学习；卷积神经网络；Transformer；注意力机制
中图分类号：TP3-05
文献标志码：A **文章编号**：1001-5965(2024)02-0466-11

随着信息化社会的不断发展，智能监控设施为社会治安系统提供了巨大的帮助。行人重识别(person re-identification, Re-ID)旨在利用计算机视觉相关技术来关联不同设备采集到的身份，完成对跨监控设备下目标行人的检测和跟踪任务，是一个典型的研究图像和视频的问题[1]。行人重识别技术的应用，大大提高了社会治安工作的效率，一定程度上推动了社会治安系统的变革。然而，现实应用中行人重识别还经常面临着光线变化、行人姿态变化、观察视角变化及局部遮挡等问题[2]，使得模型缺乏可辨别特征导致识别准确率下降。因此，针对这些挑战构建可靠的行人重识别模型是研究者们亟待解决的问题。

随着计算机运算能力的大幅提升，能够自动学习鲁棒性特征的深度学习方法被广泛应用于行人重识别领域中，使其不再使用复杂的手工设计算法，而是作为一个图像分类的问题，通过设计特征提取方法和度量损失函数，将特征学习和度量学习联合使用[3]。

卷积神经网络(convolutional neural network, CNN)中的ResNet因其独特的残差结构成为备受关注的主干网络，使网络模型可以向更深层次搭建，其应用占据了主流的行人重识别模型[4-9]。但卷积核的使用将CNN对特征的提取范围限制在了小区域内，且下采样模块往往会使行人的局部细节信息发生丢失[4]。另外，区域间的关联性也常被忽略[5]。有研究者为了增强CNN对重要局部信息的关注，将注意力机制引入行人重识别中，在注意力机制定位图像重要区域的帮助下，网络对特征信息利用的能力得到了改善[6-11]。

Transformer架构在自然语言处理(natural language processing, NLP)领域应用已经较为成熟[12-13]，

收稿日期：2022-06-02；**录用日期**：2022-07-25；**网络出版时间**：2022-08-11 17:29
网络出版地址：link.cnki.net/urlid/11.2625.V.20220810.1132.005
基金项目：国家自然科学基金(61806123,42101443)；国家重点研发计划(2019YFD0900805)
* **通信作者**. E-mail：y-zhang@shou.edu.cn

引用格式：王静，李沛橦，赵容锋，等. 融合卷积注意力和Transformer架构的行人重识别方法[J]. 北京航空航天大学学报，2024，50(2)：466-476. WANG J，LI P T，ZHAO R F，et al. A person re-identification method for fusing convolutional attention and Transformer architecture[J]. Journal of Beijing University of Aeronautics and Astronautics，2024，50(2)：466-476 (in Chinese).

受此影响, 研究者们将 Transformer 架构引入计算机视觉任务中, 并实现了与 CNN 模型相当的性能[14-15]。尽管 Transformer 架构利用多头注意力机制提取到序列向量间的隐含关系, 实现了先进的性能, 但仍然存在局部细节特征提取能力弱的问题。为了能够使 Transformer 架构更加注重于图像的局部区域, 增强局部特征的可辨别性, 受注意力机制可以增强 CNN 模型特征提取能力的启发, 本文提出一种融合卷积注意力和 Transformer 架构(fusing convolutional attention and transformer, FCAT)的方法用于行人重识别。本文首先介绍了关于行人重识别中 CNN 和 Transformer 的相关工作, 然后对本文进行理论分析, 最后通过实验证明本文方法的有效性。

1　相关工作

1.1　基于 CNN 的行人重识别

行人重识别技术已成为计算机视觉的一个热点问题, 其示意图如图 1 所示。深度学习模型因其强大的学习能力被研究者们广泛使用, 其中, CNN 因权值共享和稀疏连接成为图像处理领域中的主流模型[16], ResNet 网络因其残差连接的存在占据了行人重识别的重要地位[4]。基于 CNN 的行人重识别模型可以根据模型对特征的处理策略不同, 分为局部区域划分模型和注意力机制模型。

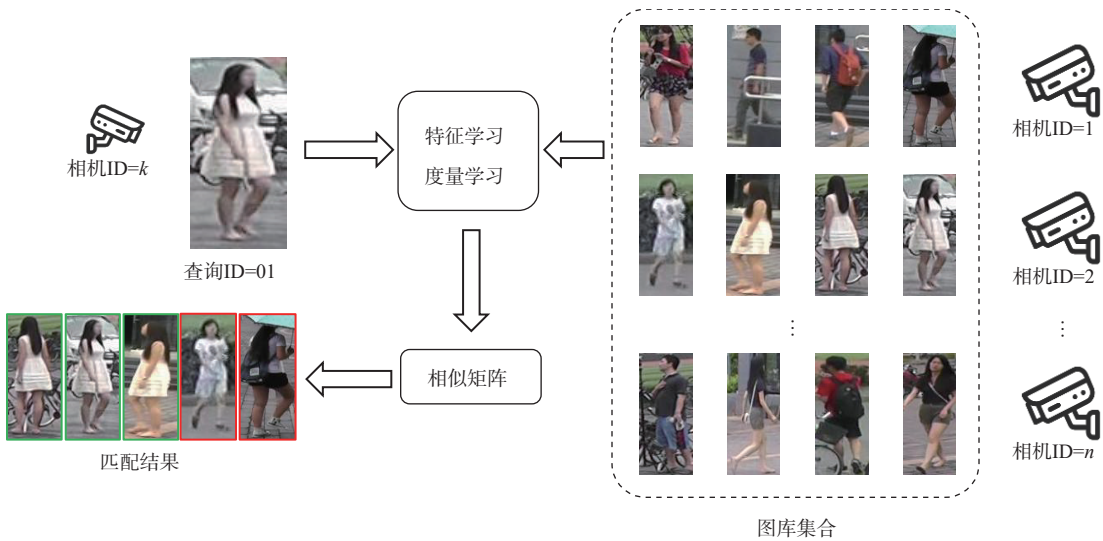


图 1　行人重识别示意图

Fig. 1　Schematic diagram of person re-identification

1.1.1　局部区域划分模型

局部区域划分模型旨在将查询图像和图库图像通过不同的局部划分方法分为不同局部区域, 进而在不同的局部区域中提取具有判别性的局部特征, 可通过获取不同局部特征间的依赖关系或联合局部特征和全局特征来提高身份识别的准确率。Song 等[17] 使用行人掩膜分割出行人区域来学习行人和背景的辨别性特征; Sun 等[18] 将行人图像从上到下划分为多个水平局部区域; Zhang 等[19] 通过计算 2 组局部特征之间的最短路径来解决局部区域的对齐问题; Su 等[20] 利用人类身体关键点线索构建局部区域来缓解姿态变化; Wang 等[21] 引入关键点估计模型将行人划分为多个局部区域; Xu 等[22] 利用姿态估计模型获得 14 个人体骨骼点构建了 7 个人体局部区域。但这些区域划分的策略要么使用了复杂的算法, 要么引入了额外的预训练模型, 最终模型的准确度会受划分方法和预训练模型的影响。

1.1.2　注意力机制模型

随着 CNN 的不断加深, 模型也变得更加复杂, 同时也带来了更大的计算量, 由于计算机的运算资源有限, 这就限制了神经网络的发展。于是注意力机制被引入深度学习网络中, 旨在从大量的信息中筛选出少量的重要信息, 并将模型的关注点聚焦于这些信息上, 忽略次重要信息, 起到了增强识别特征、抑制无关特征的作用[23]。已有大量的研究将其引入了行人重识别任务中并取得了优异的性能。Li 等[6] 提出的轻量级注意力网络 HACNN, 能同时学习行人图像中的硬区域级注意力和软像素级注意力, 得到了行人的不变特征表示。Cai 等[7] 提出一种多尺度的身体局部掩膜引导注意力网络, 通过联合学习全局和局部注意力提取行人图像中的可辨别行人特征。孙义博等[11] 将通道注意力 SE 模块嵌入 CNN 对关键特征信息进行加权强化, 提取更

鲁棒的行人表达特征。Zhang 等[9]提出一个高效的关系感知的全局注意力模块 RGA 来捕获全局结构信息以更好地进行注意力学习，显著增强了特征辨别能力。Chen 等[10]提出一个注意力金字塔网络，通过多尺度的方式在多个局部区域中进行注意力学习。He 等[4]构建了一基于视觉 Transformer 架构的目标重识别框架 TransReID，并在几个公开数据集上取得了先进的水平，为行人重识别提供了一个新的解决方案。

1.2 视觉 Transformer

受 Transformer 在 NLP 领域的影响，一些研究者尝试将 Transformer 迁移到计算机视觉任务中，这为计算机视觉领域带来了巨大的变革[15-16]。Carion 等[24]提出 DETR，将目标检测看作是一个直接的集合预测问题，首次将 Transformer 架构引入到高级视觉任务中。Dosovitskiy 等[25]提出一个纯 Transformer 的视觉 (vision transformer, ViT) 模型，将图像划分为多个图像块并编码形成序列向量，解决了 Transformer 在图像领域的输入问题，在图像分类基准上实现了先进的性能，并证明了当 Transformer 的预训练数据足够多时，Transformer 的表现会超过 CNN，为 Transformer 在计算机视觉领域打下了坚实的基础。但在图像块处理的过程中，ViT 使用卷积进行局部特征的提取，这种简单方式提取到的特征不具备较强的鲁棒性，这也成为 Transformer 架构的不足之处。因此，研究者们提出各种提升 Transformer 局部特征提取能力的方法。Liu 等[26]提出 Swin-Transformer 利用空间维度的移位窗口来建模全局和边界特征，增强了模型对局部特征提取能力。目前已有研究表明，CNN 结构可以通过 Transformer 架构实现更先进的性能 (如 VT[27]、BoT-Net[28])，Transformer 架构反过来也可以被 CNN 结构所加强 (如 DEIT[29]、ConViT[30]、ResT[31]) 提取更具辨别性的特征向量。He 等[4]提出 TransReID 基线模型，更进一步证明了 Transformer 架构在行人重识别任务中的巨大潜力。如今，Transformer 架构被广泛地应用在计算机视觉领域的各个任务中，包括图像分类、目标检测、图像分割等，并取得了显著的成效[14-15]。

2 融合 CNN 注意力和 Transformer 架构的方法

2.1 本文方法思路

融合卷积注意力和 Transformer 架构的网络 FCAT 由 4 个部分组成：区域编码模块、融合注意力模块、自注意力计算模块和度量损失模块。整体处理思路为：首先通过区域编码模块将图像编码为 N 个向量，再使用空间注意力调整 N 个向量的空间重要性分布，其次将 N 个向量与一个可学习的无明显语义的分类表示向量进行拼接操作得到 N+1 个向量，再将位置编码加到整个序列向量上作为 Transformer 编码器中自注意力计算模块的输入。经过自注意力机制的信息交互，再使用通道注意力增强信息交互后的分类表示，将其作为全局特征进行损失的计算。图 2 为 FCAT 架构方法的网络结构。

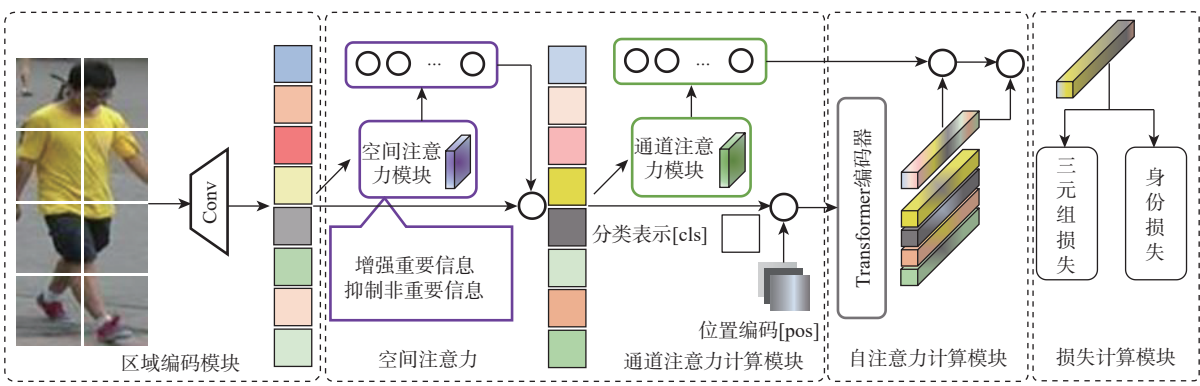


图 2 FCAT 网络结构

Fig. 2 FCAT network framework

2.2 主干网络

FCAT 的主干网络采用 ViT 模型，而 ViT 模型采用 Transformer 的编码器结构[25]。Transformer 是一个完全基于自注意力机制的深度学习架构，旨在捕捉序列向量之间的相互依赖性[12]，即自注意力的使用对象必须是一个序列向量，这导致自注意力机制不能单独完成现实中的各项任务。

2.2.1 区域编码

给定一张行人图像 $\boldsymbol{x} \in \mathbf{R}^{H\times W\times C}$，其中 H、W 和 C 分别为图像的高度、宽度和通道数，为解决 Transformer 在计算机视觉领域的输入问题，ViT 提出了图像块编码，即使用尺寸大小为 P，维度为 C' 的卷积核将图像划分为 $N=HW/P^2$ 个不重叠的图像块并编码成序列向量 $\boldsymbol{x}_r=[\boldsymbol{r}_1,\boldsymbol{r}_2,\cdots,\boldsymbol{r}_N]$。

2.2.2　分类表示

编码向量$\boldsymbol{r}_i$经过自注意力层会融合其他向量的部分信息, 最终在所有特征中选择一项作为全局特征放入多层感知机中产生分类结果, 但这会使模型对最终选择的向量产生偏置, 即发生了“不公平”的现象。为了避免这种情况的发生, Transformer 架构在序列向量中加入了一个与其他向量无关的代表整个序列向量的特征表示$\boldsymbol{r}_{\text{cls}}$, 即$\boldsymbol{x}'_r=[\boldsymbol{r}_{\text{cls}},\boldsymbol{r}_1,\boldsymbol{r}_2,\cdots,\boldsymbol{r}_N]$, 最终使用$\boldsymbol{r}_{\text{cls}}$作为全局特征产生分类结果, 如图 2 中的 [cls] 符号所示。

2.2.3　位置编码

在处理自然语言的过程中, 单词的位置和顺序定义了该句子的语法, 在计算机视觉的处理过程也有相同的概念。例如, 一张图片的像素的排列分布决定了图像的内容, 这对于图像的识别具有很大的帮助。但自注意力机制在处理序列向量的过程中具有位置不变的特性, 即自注意力机制无法捕获输入向量的位置关系, 因此 Transformer 在自注意力层前会在序列特征向量加入位置编码$\boldsymbol{v}_{\text{pos}}\in\mathbf{R}^{(N+1)\times C'}$, 即特征向量$\boldsymbol{f}=\boldsymbol{x}'_r+\boldsymbol{v}_{\text{pos}}=[\boldsymbol{f}_{\text{cls}},\boldsymbol{f}_1,\boldsymbol{f}_2,\cdots,\boldsymbol{f}_N]$, 位置编码如图 2 中的 [pos] 符号所示。

2.2.4　Transformer 编码器

Transformer 编码器由 N 个 Transformer 层组成, 而 Transformer 层由 2 部分组成: 多头自注意力层和 MLP 组成。经多头自注意力层会得到局部信息相互交融的全局特征向量集合$\boldsymbol{g}=M_{\text{MSA}}(\boldsymbol{f})=[\boldsymbol{g}_{\text{cls}},\boldsymbol{g}_1,\boldsymbol{g}_2,\cdots,\boldsymbol{g}_N]$, M_{MSA}表示多头自注意力模块, 最终选择$\boldsymbol{g}_{\text{cls}}$作为全局特征再经过 MLP 获得分类结果。另外, Transfromer 层在每个 MLP 层之前应用了 layernorm 层以加速模型的收敛速度。N 个 Transformer 层进行逐个串联操作以充分进行各局部特征的长程依赖关系的提取。

2.3　融合注意力模块

注意力机制倾向于选择性地将模型关注点集中于人类更为感兴趣的信息, 同时忽略其他感知内容, 因此, 该方法在行人重识别中可用于自适应地定位行人图像的辨别性局部区域[32-34]。受此影响, 本文提出 FCAT, 旨在增强 Transformer 模型对图像中重要信息的关注能力。

2.3.1　空间特征增强

对于一张行人图像$\boldsymbol{x}=[\boldsymbol{r}_1,\boldsymbol{r}_2,\cdots,\boldsymbol{r}_N]\in\mathbf{R}^{\text{H}\times\text{W}\times\text{C}}$, 可以将其视为 N 个区域向量的组合, 在进行身份识别的过程中, 每一个区域对识别的贡献并不相等。与任务相关的区域才是模型应该重点关注的区域。如对于行人身份识别主要取决于行人的像素部分, 而背景部分的像素相对具有较少贡献。由于空间注意力的本质就是定位重要区域并获取区域的权重分布向量, 因而本节通过使用空间注意力定位行人图像中的重要部分, 以间接增强 Transformer 架构对局部信息的提取能力。上述空间注意力获取区域权重向量并产生新的注意力特征的过程可表示为

$$\boldsymbol{x}_{\text{sa}}=F_{\text{sa}}(\boldsymbol{x})\times\boldsymbol{x} \tag{1}$$

式中: F_{sa}表示空间注意力函数, 其输出为模型对输入向量的重要性分布预测概率。空间注意力增强的 Transformer 特征的具体作用过程为: 区域编码模块将输入图像$\boldsymbol{x}$分块编码为序列向量$\boldsymbol{x}_{\text{r}}=[\boldsymbol{r}_1,\boldsymbol{r}_2,\cdots,\boldsymbol{r}_N]$, 再将$\boldsymbol{x}_{\text{r}}$输入空间注意力模块$F_{\text{sa}}$获取当前序列向量的空间权重分布$\boldsymbol{w}_{\text{s}}$, 最后将空间权重分布$\boldsymbol{w}_{\text{s}}$作用在序列向量$\boldsymbol{x}_{\text{r}}$上产生空间注意力编码向量$\boldsymbol{x}_{\text{sa}}$。为了保证模型对原始图像的信息保留, 受残差连接的启发, 本文引入了一个注意力权重因子α_{s}以控制模型对空间注意力向量的偏置程度, 同时也保留了原始图像的分布状态。上述过程可表示为

$$\boldsymbol{x}_{\text{sa}}=\boldsymbol{w}_{\text{s}}\times\boldsymbol{x}_{\text{r}}=F_{\text{sa}}(\boldsymbol{x}_{\text{r}})\times\boldsymbol{x}_{\text{r}} \tag{2}$$

$$\boldsymbol{x}'_{\text{sa}}=\alpha_{\text{s}}\boldsymbol{x}_{\text{r}}+(1-\alpha_{\text{s}})\boldsymbol{x}_{\text{sa}} \tag{3}$$

式中: α_{s}为空间注意力权重因子, 在实验过程中设置为 0.5; $\boldsymbol{x}'_{\text{sa}}$表示空间注意力向量和原始编码向量加权后的空间增强特征。

2.3.2　通道特征增强

空间注意力是一种在图像长宽方向进行像素重要性分布提取的方法, 是图像尺度特征的筛选。通道注意力则侧重于在所有的尺度特征集合中, 每个尺度特征的重要程度占比。与空间特征增强类似, 通道特征的增强表示在通道维度进行重要性的提取并加权至全局特征向量。

在拼接分类表示之前, 空间特征增强后的特征向量$\boldsymbol{x}'_{\text{sa}}$使用通道注意力获取当前特征向量的通道分布权重$\boldsymbol{w}_{\text{c}}$。但与空间特征增强不同的是, 通道分布权重$\boldsymbol{w}_{\text{c}}$表示的是 N 个特征向量的的整体的分布状态, 即通道分布权重$\boldsymbol{w}_{\text{c}}$对于每一个特征向量都是等价的, 为了减少模型的计算量, 本文将通道分布权重$\boldsymbol{w}_{\text{c}}$直接作用在分类表示的所有通道上。但考虑到自注意力层会使所有特征向量在通道维度上重新赋予原始权值, 本文选择将通道分布权重$\boldsymbol{w}_{\text{c}}$作用在自注意力层之后的全局特征$\boldsymbol{g}_{\text{cls}}$之上, 并引入通道权重因子$\alpha_{\text{c}}$以控制模型对通道注意力特征的偏置, 同时也保留了原始图像的通道分布状态, 通道注意力的整体作用过程表示为

$$\boldsymbol{g}_{\text{ca}}=\boldsymbol{w}_{\text{c}}\times\boldsymbol{g}_{\text{cls}}=F_{\text{ca}}(\boldsymbol{x}'_{\text{sa}})\times\boldsymbol{g}_{\text{cls}} \tag{4}$$

$$\boldsymbol{g}'_{\mathrm{ca}}=\alpha_{\mathrm{c}}\boldsymbol{g}_{\mathrm{cls}}+(1-\alpha_{\mathrm{c}})\boldsymbol{g}_{\mathrm{ca}} \tag{5}$$

式中：α_{c}为通道注意力权重因子，在实验过程中设置为0.5；F_{ca}为通道注意力计算函数；$\boldsymbol{g}'_{\mathrm{ca}}$表示原始全局特征$\boldsymbol{g}_{\mathrm{cls}}$与通道注意力加权特征$\boldsymbol{g}_{\mathrm{ca}}$融合后的通道增强特征。

2.3.3 注意力函数

为了使生成的注意力权重具有更好的自适应能力，本文使用卷积注意力模块CBAM[32]作为注意力权重的计算函数。但CBAM的注意力权重计算方式将通道注意力计算和空间注意力计算串联使用，不能够完成对本文的特征增强方案。故本文将CBAM的注意力权重计算方式拆分为2个独立的注意力计算函数，作为注意力改善模块的核心函数，具体运算过程如下：

$$F_{\mathrm{sa}}(\boldsymbol{x})=\sigma(\mathrm{Cov}^{7\times7}(\mathrm{cat}(F_{\mathrm{avg}}(\boldsymbol{x}),F_{\mathrm{max}}(\boldsymbol{x})))) \tag{6}$$

$$F_{\mathrm{ca}}(\boldsymbol{x})=\sigma(\mathrm{Cov}_1(\mathrm{Cov}_0(F_{\mathrm{avg}}(\boldsymbol{x})))+\mathrm{Cov}_1(\mathrm{Cov}_0(F_{\mathrm{max}}(\boldsymbol{x})))) \tag{7}$$

式中：$\sigma(\cdot)$表示sigmoid激活函数，将预测概率映射到[0,1]；$\mathrm{Cov}^{7\times7}$表示卷积核为7×7的卷积操作；$\mathrm{cat}(\cdot)$表示拼接操作；F_{avg}表示平均池化；F_{max}表示最大池化；Cov_0和Cov_1表示2个卷积核为1的卷积层，平均池化和最大池化共享Cov_0和Cov_1的权值。

2.4 损失函数

参考主流的基于深度学习的行人重识别模型[1,35-37]，本文将行人重识别视为图像分类任务，为融合卷积注意力后的Transformer全局特征构建身份损失L_{ID}和难样本三元组损失L_{Tri}指导模型的优化。其中，身份损失是不带有标签平滑的交叉熵损失，其用来计算当前模型对输出结果的概率分布与真实分布的距离，具体公式如下：

$$L_{\mathrm{ID}}=-\sum_i y_i\log P_i \tag{8}$$

$$P=M_{\mathrm{FC}}(\boldsymbol{g}'_{\mathrm{ca}}) \tag{9}$$

式中：$M_{\mathrm{FC}}(\cdot)$表示全连接层的作用过程，将注意力全局特征$\boldsymbol{g}'_{\mathrm{ca}}$映射为对身份的预测概率；$y_i$表示第$i$个身份的真实标签。$P_i$表示全局特征对图像身份的预测概率。$L_{\mathrm{ID}}$表示通过注意力全局特征$\boldsymbol{g}'_{\mathrm{ca}}$得到的身份损失。

带有软边界的难样本三元组损失的输入为3张图片：固定锚点图片A、正样本图片P和负样本图片N，该损失能够拉近正样本对的间距，增加负样本对间距，具构建过程如下：

$$L_{\mathrm{Tri}}=Ln(1+\exp(\|\boldsymbol{f}_{\mathrm{A}}-\boldsymbol{f}_{\mathrm{P}}\|_2^2-\|\boldsymbol{f}_{\mathrm{A}}-\boldsymbol{f}_{\mathrm{N}}\|_2^2)) \tag{10}$$

式中：$\boldsymbol{f}_{\mathrm{A}}$、$\boldsymbol{f}_{\mathrm{P}}$、$\boldsymbol{f}_{\mathrm{N}}$分别表示3种图片对应的注意力全局特征。

使用注意力全局特征构建的身份损失和难样本三元组损失可以更好地优化模型，其总体损失的计算方法为

$$L=\lambda_{\mathrm{ID}}L_{\mathrm{ID}}+\lambda_{\mathrm{tri}}L_{\mathrm{tri}} \tag{11}$$

式中：λ表示每种损失对应的权值，实验中均设置为0.5，本文通过最小化L端到端地训练网络。

3 实 验

3.1 数据集与评价指

在3个公开行人重识别数据集Market-1501[38]、DukeMTMC-ReID[39]、Occluded-Duke[40]上开展对比实验研究。其中，Occluded-Duke选自DukeMTMC-ReID，按照训练集、查询集、图库集中分别带有9%、100%、10%遮挡图像。数据集的详细信息如表1所示。

表1 数据集信息

Table 1 Datasets information

数据集	数据集整体		训练集		查询集		图库集	
	身份数	图像数	身份数	图像数	身份数	图像数	身份数	图像数
Market-1501	1 501	32 668	751	12 936	750	3 368	750	19 732
DukeMTMC-ReID	1 404	36 441	702	16 522	702	2 228	1 110	17 661
Occluded-Duke	1 404	36 441	702	15 618	519	2 210	1 110	17 661

遵循常规的行人重识别做法，本文使用平均匹配精度mAP[38]和准确率Rank-1[41]这2种评价指标来评估模型的性能。

3.2 实验设置

为了使模型能够更好地适配行人图像，本文将所有数据集的图像大小调整为256×128，训练集通过随机擦除、随机裁剪、水平翻转进行图像的增强，随机率设置为0.5。本文采用ViT作为主干网络，初始权重先在ImageNet-21K上进行预训练，然后在ImageNet-1K进行微调。在此基础上，ViT区域编码后加入了融合注意力模块，区域编码大小设置为16，区域编码维度设置为768，空间权重因子α_{s}和通道权重因子α_{c}均设置为0.5。训练迭代次数epoch为120，训练批量大小为64，测试批量大小为

256, 三元组损失的实例数设置为 4, 即每个批次中的每个行人身份都包含 4 张图像, 模型采用 SGD 作为优化器, 初始学习率设置为 0.008, 并以余弦选学习率衰减, 为了保证对比的公平, 所有模型均不使用重排序算法。实验使用一块 Nvidia Geforce 3090 GPU。

3.3　与现有方法的对比

在本节中列出了 FCAT 与现有主流行人重识别方法的对比结果, 包括基于局部区域划分的方法与基于注意力机制的方法 2 类。模型的主干网络分别为 CNN 和 Transformer。如表 2 所示。从对比结果中可以看出, 无论是一般的无遮挡的数据集(如 Market-1501、DukeMTMC-ReID)还是遮挡数据(如 Occluded-Duke), 本文所提方法都具有一定的优势, 表明本文方法可以较好地解决行人重识别领域的一些挑战。

表 2　FCAT 与其他方法的性能对比结果

Table 2　Performance results of FCAT compared to other methods

模型	主干网络	方法	mAP/%			Rank-1/%		
			Market-1501	DukeMTMC	Occluded-Duke	Market-1501	DukeMTMC	Occluded-Duke
局部区域划分模型	CNN	Spindle[42]				76.9		
	CNN	PDC[20]	63.4			84.1		
	CNN	MGCAM[17]	74.3			83.8		
	CNN	MGN[43]	78.4	78.4		88.7	88.7	
	CNN	AlignedReID[19]	79.3			91.8		
	CNN	Part Aligned[44]	63.4		20.2	81.0		28.8
	CNN	PGFA[40]	76.8	65.5	37.3	91.2	82.6	51.4
	CNN	PCB[18]	77.4	66.1	42.6	92.3	81.8	33.7
	CNN	PCB+RPP[18]	81.6	69.2		93.8	83.3	
	CNN	HOReID[21]	84.9	75.6	43.8	94.2	86.9	55.1
注意力机制模型	CNN	HACNN[6]	82.8	63.8		93.8	80.5	
	CNN	MMGA[7]	87.2	78.1		95.0	89.5	
	CNN	RGA-SC[9]	**88.4**	74.9		**96.1**	86.1	
	Transformer	DRL-Net[45]	86.9	76.6	50.8	94.7	88.1	**65.0**
	Transformer	TransReID/B[4]	86.8	79.3	53.1	94.7	88.8	60.5
	Transformer	TransReID/JPM[4]	88.2	80.6	55.7	95.0	89.6	64.2
	Transformer	FCAT	**88.4**	**80.8**	**55.9**	**95.1**	**90.2**	**64.5**

注: 其中粗体数字表示最优性能, 下划线数字代表次优性能。TransReID/B表示TransReID模型的基线模型, TransReID/JPM表示基线模型附加Jigsaw Patch Module操作。

从表 2 中 Market-1501 无遮挡数据集测试结果可以看出, 基于 Transformer 的行人重识别方法相对于基于 CNN 的模型有较大的优势, 本文所提模型实现了 88.2% 的 mAP 和 94.7% 的 Rank-1 准确度, 得到了与最先进方法相当的结果, 尤其是相比于基于局部划分的方法取得了较大的优势。尽管 TransReID、B 基线模型和 DRL-Net[45] 都是基于 Transformer 架构, 但其忽略了该模型对局部细节特征提取能力较弱的问题, 故在相同的网络架构上 FCAT 具有较大的优势, mAP 准确度提高了 1.4%, Rank-1 提高了 1.4%。另外, 从 TransReID 的 2 个模型中可以观察到, TransReID/JPM 模型因加入了局部划分的方法, 增强了 Transformer 对局部的关注能力, 模型性能也因此得到了提升, 这也说明增强 Transformer 对局部细节的关注可以提升模型的性能, 也证明本文所提方法的有效性。相比较本文所提方法来增强对局部的关注而言, 其使用额外的局部划分算法, 增加了模型的计算量。纵观局部划分的方法与注意力机制的方法, 可以观察到基于注意力的方法拥有较高的准确度, 这也间接证明注意力机制的有效性, 故本文所提方法具有一定的优势。

在 DukeMTMC-ReID 和 Occluded-Duke 数据集上, 由于 2 个数据集中有较多的遮挡图像存在, 所有模型的准确度都有下降, 而 Occluded-Duke 数据集中因为有更多的遮挡图像, 模型准确度相对下降地更多。

本文方法在 DukeMTMC-ReID 数据集上 mAP 为 80.8%, Rank-1 为 90.2%, 与 TransReID/B 相比, 本文方法 mAP 提高了 1.5%, Rank-1 提高了 1.4%。与 RGA 采用注意力增强卷积网络相比, 本文使用 Transformer 架构中融合注意力方法 mAP 提高了 5.9%, Rank-1 提高了 4.1%。此外, MMGA 模型证明

局部区域间的长程依赖可提升单个局部特征的分类能力[5]，从表2中可以看出其相比于局部区域划分具有一定的性能优势。而Transformer的全局特征在自注意力机制提取到的长程依赖关系的作用下，分类准确性相比于MMGA，mAP提高了1.2%，即使在遮挡的情况下也能够较好地进行身份识别。也证明了Transformer对较为复杂的环境挑战具有更强的适应性。

从Occluded-Duke遮挡数据上的对比结果可知，基于Transformer的模型对基于CNN的模型mAP提高了12.1%、Rank-1提高了9.4%，FCAT的mAP为55.9%、Rank-1为64.5%，与解决遮挡问题的DRL-Net模型相比mAP提高了5.1%。这些性能也体现了本文模型对解决遮挡问题具有一定的鲁棒性，也为行人重识别解决遮挡问题提供了思路。

3.4 消融实验及可视化

3.4.1 模型组件的消融研究

为了进一步验证FCAT的有效性及探索其解决行人重识别遮挡问题的能力，本节在遮挡数据集Occluded-Duke上对FCAT的组成部分进行了消融研究。所有实验均在同一环境下进行，如3.2节中所述，实验结果如表3所示。其中，HOReID代表基于CNN的方法，基线模型为不使用任何额外操作的基于ViT的模型（如TransReID、B[4]），+S表示在模型中使用空间注意力，+C表示在模型中使用通道注意力，+SC和+CS表示两者都使用。其中，+SC表示通道注意力处于空间注意力之后，+CS表示通道注意力位于空间注意力之前。

从表3前2行可以看出，基于Transformer的模型相对于基于卷积网络的模型在遮挡问题的处理上具有相当大的优势。基于局部特征的模型被证明对行人身份的识别是有益的。Transformer架构在区域编码的帮助下，将整张图像通过多个区域向量表示，相比基于CNN的模型HOReID通过使用姿态估计器进行局部划分具有更简便的划分优势，且具有更好的分类性能。这有效证明了Transformer架构对于行人重识别任务的有效性。

在区域编码的过程中，Transformer直接使用卷积进行局部特征的提取，这种简单方式提取到的局部特征对于复杂场景下的身份识别不具有较强的鲁棒性。为增强Transformer提取的局部特征鲁棒性，FCAT通过将空间注意力和通道注意力融入Transformer架构中，利用注意力机制提升重要区域和重要特征的模型关注度。其中，空间注意力通过提升图像中重要区域的重要权重获取Transformer对其的关注程度，通道注意力通过提升局部特征中

表3 模型遮挡数据集上的消融研究

Table 3 Ablation studies on Occluded-Duke %

方法	mAP	Rank-1
HOReID	43.8	55.1
基线	53.1	60.5
基线+S	55.8	63.3
基线+C	55.5	63.3
基线+CS	55.8	63.9
基线+ SC	55.9	64.5

辨别性强的细节特征的重要性，进一步增强重要区域重要特征对身份识别的影响，通过这种融合方式，以适应复杂的行人识别场景。在局部特征的基础上，再结合Transformer的自注意力机制提取到的长程依赖关系，最终获取到的全局特征具有较强的分类能力，进而提升模型在遮挡数据集上的识别性能。从表3的2～6行数据中可以观察到，在空间注意力或通道注意力的单独作用下，Transformer模型在遮挡数据集上的性能都得到了提升，利用空间注意力使mAP提高了2.7%，Rank-1提高了2.8%，通道注意使模型的mAP提高了2.4%，Rank-1提高了32.8%，但空间注意力提高的性能略微高于通道注意力的使用。在2种注意力的共同作用下，模型性能进一步得到了提升：+CS的mAP为55.8%，Rank-1为63.9%；+SC的mAP为55.9%，Rank-1为64.5%。结果表明，通过注意力机制提升模型对重要区域重要特征的关注可以在一定程度上提升模型在复杂场景下的识别性能。

为了进一步探究空间注意力和通道注意力位置对模型性能的影响，本文对通道注意力的位置进行了分析。由于空间注意力的特殊性，只能作用于序列向量之上，不能够直接作用于全局特征，这限制了空间注意力只能位于区域编码之后。而通道注意力最终得一组通道权重，将其作用于单个全局特征可以提升其分类能力，旨在提升模型对重要特征的关注。当使用空间注意力的情况下，通道注意力权重的提取位置可以位于空间注意力使用前后，实验中将通道注意力的提取在空间注意力之前表示为+CS，将通道注意力的提取在空间注意力之后表示为+SC。+CS从实际意义上可以理解为，通道注意力提取的是区域编码向量的通道重要性分布，该重要性分布会逐渐学习到整张图像的通道特征分布。相对的，+SC的通道注意力旨在提取空间注意力之后的序列向量的特征重要性分布，该分布会使模型逐渐向图像的重要区域学习。如表3所示，无论是+CS还是+SC模型的性能都要高于基线模型和单独注意力的使用。单独的空间注意力+S

与+CS 有相同的 mAP 准确度, 通道注意力位置的改变不会影响空间注意力的使用, 其只会影响经 Transformer 编码器后的全局特征, 进而影响模型性能。但经过通道注意力的作用, 模型的 Rank-1 精度得到了提升, 这证明了通道注意力可以提升全局特征分类能力的作用。当通道权重分布的提取位于空间注意力之后时, 模型会逐渐关注对重要区域重要特征的关注, 在 2 种注意力的共同作用下, 使得全局特征具有了更强的辨别性, mAP 和 Rank-1 精度都得到了进一步提升。

3.4.2　推理成本研究

为了进一步说明所提方法的轻量性, 本节对模型的推理时间、计算量和参数量做了相关的消融实验研究。实验运行环境与 3.2 节相同。另外, 为了减少由于设备内存差异而引起的推理时间误差, 每次输入模型的图像数量为 64, 用该方法以平均每个图像在设备中的 IO 误差。为了使对比更加地明显, 本文以 ResNet50 的推理成本为单位 1, 表达为 1*x*, 其他模型均以其为参考。实验结果如表 4 所示。

表 4　推理成本的消融研究

Table 4　Ablation studies of inference costs

方法	推理时间	计算量	模型参数量
ResNet50	1*x*	1*x*	1*x*
基线	0.433 8*x*	2.710 5*x*	3.641 1*x*
基线+JPM	0.459 5*x*	2.940 1*x*	3.942 7*x*
基线+S	0.421 2*x*	2.710 5*x*	3.641 1*x*
基线+C	0.395 5*x*	2.710 6*x*	3.644 2*x*
基线+ SC	0.415 9*x*	2.710 6*x*	3.644 2*x*

从实验结果可以看出, 在相同的运行条件下, Transformer 架构比 ResNet50 拥有 2.7 倍的计算量和 3.6 倍的模型参数量, 但模型的推理时间却大大优于 ResNet50, 在更复杂的计算下只用了 CNN 不到一半的时间, 这归功于 Transformer 内部多头自注意力机制的并行计算操作。另外, 为了说明 FCAT 融合注意力模型的轻量性, 本文将 TransReID 和 JPM 也引入推理成本的对比中, 从对比结果中可看出, 加入额外局部划分算法的 TransReID 和 JPM 相比于融合注意力模型的 FCAT 由更多的推理时间、计算量和模型参数量, 反观融合注意力模型的 FCAT, 在基线模型的基础上, 几乎不增加额外的计算量和模型参数, 但推理的时间却得到了提升, 这又进一步证明了基于注意力模型的优势。从表 4 的最后 3 行可以得出, 注意力模型在引入额外模块的前提下, 并没有明显增加模型的计算量和模型参数, 而且推理时间得到了提升, 这得益于注意力机制抑制冗余信息的优势。

3.4.3　可视化

本文进一步对消融研究的多种模型进行了可视化分析。在 Occluded-Duke 数据集上选出 4 张图像进行模型注意力的可视化, 分别包括未遮挡的行人 2 张、被遮挡的行人 2 张。图 3 为注意力热图的可视化结果。从可视化结果中可以观察到, 基于 CNN 的模型关注区域由于卷积感受野的原因大多分布在小的判别区域, 这使得模型提取到的特征在行人分类时不具备较强的辨别性。而基于 Transformer 的 ViT 模型在自注意力作用下, 模型捕获到了图像区域间的长程依赖关系, 使得其关注到了图像不同的局部区域。但其因在区域编码的过程中直接使用普通卷积提取局部特征, 这导致自注意力机制不能提取到准确的依赖关系, 模型关注度不能集中于图像中的可辨别区域。当在行人遮挡的情况下时, 行人局部信息发生缺失, 模型会存在较多的关注度在遮挡物上, 身份识别性能不具备说服力。从可视化结果中可以观察到, FCAT 相对于 ViT 基线模型而言在行人区域有更多的关注, 这是由于卷积注意力机制使用改变了模型关注方向。如图 3 中+S 所示, 实验中空间注意力的使用改变了区域编码向量的重要性分布, 经过不断的学习, 模型的关注会逐渐向与身份相关的区域优化, 故+S 相对于 ViT 有更多的关注在行人区域。但由于局部特征在通道维度具有多样性, 模型最终的分类能力不会集中于具有分辨性的通道特征, 还不足以使模型关注集中于行人区域, 致使大量关注处于遮挡物和背景之上。实验中通道注意力的使用会改变全局特征的通道分布状态, +C 实验组由于获取的全局特征是由直接区域编码的向量经自注意力机制得到, 即通道注意力是在整张图像的所有通道特征内进行重要性的提取, 故模型的关注会分布在图像的各个区域。

当通道注意力和空间注意力同时使用时, 由于通道注意力的提取位置不同, 产生了+CS 和+SC 这 2 组实验。当通道权重提取位置位于空间注意力之前时, 与+C 情况一致, 通道注意力权重是在所有区域的所有通道特征内进行重要性的提取。但在空间注意力的作用下, 区域编码向量会关注于重要区域得到全局特征, 在二者的共同作用下, 模型会向重要区域优化, 此时仍会有大量关注处于遮挡物和背景上。当通道权重提取位置位于空间注意力之后时, 通道权重的提取会逐渐关注于空间注意力作用后的重要区域, 将此通道权重作用在全局特征上会进一步加强模型对重要区域的辨别能力, 故在+SC 的可视化结果中, 大量的模型关注处于行人区域, 这进一步证明了 Transformer 融合卷积注意力

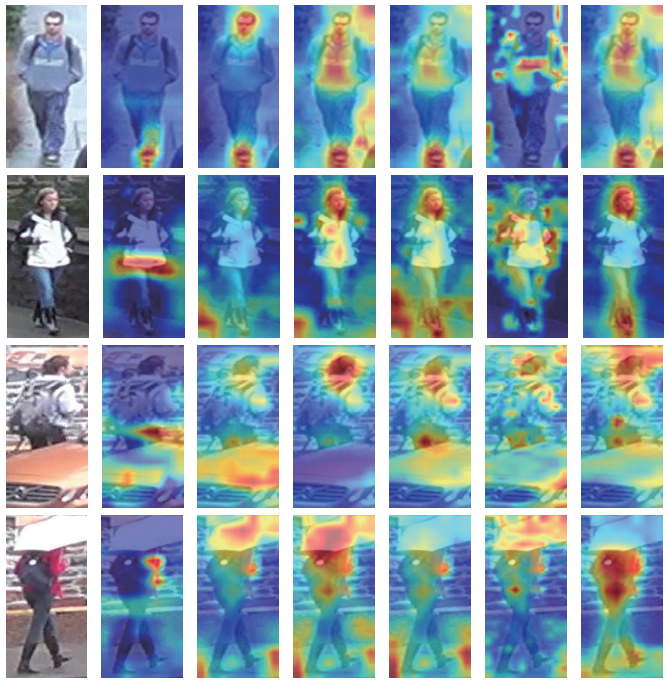


图3 模型注意力热图

Fig. 3 Model attention heatmap

的有效性，在一定程度上解决了Transformer局部特征提取能力弱的问题，使模型的判别更加集中于重要的局部区域。

4 结 论

本文提出的融合卷积注意力和Transformer架构方法，总结如下：

1）本文方法使用了一种较为新颖的结合卷积注意力和Transformer架构的方法。

2）本文方法在未遮挡数据集Market-1501上取得了与先进方法相媲美的成绩，mAP准确度为88.4%、Rank-1准确度为95.1%。在具有遮挡情况的DukeMTMC-ReID和Occluded-Duke数据集上取得了较为先进的水平，在DukeMTMC-ReID数据集上的mAP准确度为80.8%、Rank-1准确度为90.2%，在Occluded-Duke数据集上mAP准确度为55.9%、Rank-1准确度为64.5%。

3）尽管FCAT的计算量和参数量接近于ResNet50的3倍和4倍，但其拥有更短的推理时间。

4）FCAT相比于现有的方法更能够适用较为复杂的场景，模型注意力能够更集中于行人的区域。

5）相对于当前行人重识别领域研究进展，尽管FCAT取得较为先进的结果，但还不能够完全解决实际应用场景中的各种问题，模型仍有一定的提升空间。

参考文献（References）

[1] YE M, SHEN J B, LIN G J, et al. Deep learning for person re-identification: A survey and outlook[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 44(6): 2872-2893.

[2] ZHENG L, YANG Y, HAUPTMANN A G. Person re-identification: Past, present and future[EB/OL]. (2016-10-10) [2022-05-20]. https://arxiv.org/abs/1610.02984.pdf.

[3] LUO H, GU Y Z, LIAO X Y, et al. Bag of tricks and a strong baseline for deep person re-identification[C]// 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. Piscataway: IEEE Press, 2020: 1487-1495.

[4] HE S T, LUO H, WANG P C, et al. TransReID: Transformer-based object re-identification[C]// 2021 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway: IEEE Press, 2022: 14993-15002.

[5] PARK H, HAM B. Relation network for person re-identification[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2020, 34(7): 11839-11847.

[6] LI W, ZHU X T, GONG S G. Harmonious attention network for person re-identification[C]// 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2018: 2285-2294.

[7] CAI H L, WANG Z G, CHENG J X. Multi-scale body-part mask guided attention for person re-identification[C]// 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. Piscataway: IEEE Press, 2020: 1555-1564.

[8] 张晓伟, 吕明强, 李慧. 基于局部语义特征不变性的跨域行人重识别[J]. 北京航空航天大学学报, 2020, 46(9): 1682-1690.
ZHANG X W, LYU M Q, LI H. Cross-domain person re-identifica-

tion based on partial semantic feature invariance[J]. Journal of Beijing University of Aeronautics and Astronautics, 2020, 46(9): 1682-1690 (in Chinese).

[9] ZHANG Z Z, LAN C L, ZENG W J, et al. Relation-aware global attention for person re-identification[C]// 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2020: 3183-3192.

[10] CHEN G Y, GU T P, LU J W, et al. Person re-identification via attention pyramid[J]. IEEE Transactions on Image Processing, 2021, 30: 7663-7676.

[11] 孙义博, 张文靖, 王蓉, 等. 基于通道注意力机制的行人重识别方法[J]. 北京航空航天大学学报, 2022, 48(5): 881-889.
SUN Y B, ZHANG W J, WANG R, et al. Pedestrian re-identification method based on channel attention mechanism[J]. Journal of Beijing University of Aeronautics and Astronautics, 2022, 48(5): 881-889 (in Chinese).

[12] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all You need[C]// Proceedings of the 31st International Conference on Neural Information Processing Systems. New York: ACM, 2017: 6000-6010.

[13] GALASSI A, LIPPI M, TORRONI P. Attention in natural language processing[J]. IEEE Transactions on Neural Networks and Learning Systems, 2020, 32(10): 4291-4308.

[14] LIU Y, ZHANG Y, WANG Y X, et al. A survey of visual transformers[J]. IEEE Transaction on Neural Networks and Learning Systems, 2023.

[15] KHAN S, NASEER M, HAYAT M, et al. Transformers in vision: A survey[J]. ACM Computing Surveys, 2022, 54(10): 1-41.

[16] HADJI I, WILDES R P. What do we understand about convolutional networks? [EB/OL]. (2018-03-23) [2022-05-20]. https://arxiv.org/abs/1803.08834.pdf.

[17] SONG C F, HUANG Y, OUYANG W L, et al. Mask-guided contrastive attention model for person re-identification[C]// 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2018: 1179-1188.

[18] SUN Y F, ZHENG L, YANG Y, et al. Beyond part models: Person retrieval with refined part pooling (and A strong convolutional baseline)[C]// Computer Vision - ECCV 2018: 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part IV. New York: ACM, 2018: 501-518.

[19] ZHANG X, LUO H, FAN X, et al. AlignedReID: Surpassing human-level performance in person re-identification[EB/OL]. (2017-11-22) [2022-05-20].https://arxiv.org/abs/1711.08184.pdf.

[20] SU C, LI J N, ZHANG S L, et al. Pose-driven deep convolutional model for person re-identification[C]// 2017 IEEE International Conference on Computer Vision. Piscataway: IEEE Press, 2017: 3980-3989.

[21] WANG G A, YANG S, LIU H Y, et al. High-order information matters: learning relation and topology for occluded person re-identification[C]// 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2020: 6448-6457.

[22] XU J, ZHAO R, ZHU F, et al. Attention-aware compositional network for person re-identification[C]// 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2018: 2119-2128.

[23] GUO M H, XU T X, LIU J J, et al. Attention mechanisms in computer vision: a survey[J]. Computational Visual Media, 2022, 8(3): 331-368.

[24] CARION N, MASSA F, SYNNAEVE G, et al. End-to-end object detection with transformers[C]//European Conference on Computer Vision. Cham: Springer, 2020: 213-229.

[25] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, et al. An image is worth 16x16 words: Transformers for image recognition at scale[EB/OL]. (2020-10-22) [2022-05-20]. https://arxiv.org/abs/2010.11929.pdf.

[26] LIU Z, LIN Y T, CAO Y, et al. Swin transformer: hierarchical vision transformer using shifted windows[C]// 2021 IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE Press, 2022: 9992-10002.

[27] WU B C, XU C F, DAI X L, et al. Visual transformers: Token-based image representation and processing for computer vision[EB/OL]. (2020-06-05) [2022-05-21]. https://arxiv.org/abs/2006.03677.pdf.

[28] SRINIVAS A, LIN T Y, PARMAR N, et al. Bottleneck transformers for visual recognition[C]// 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2021: 16514-16524.

[29] TOUVRON H, CORD M, DOUZE M, et al. Training data-efficient image transformers & distillation through attention[EB/OL]. (2020-12-23) [2022-05-21].https://arxiv.org/abs/2012.12877.pdf.

[30] D'ASCOLI S, TOUVRON H, LEAVITT M L, et al. ConViT: Improving vision transformers with soft convolutional inductive biases[J]. Journal of Statistical Mechanics:Theory and Experiment, 2022, 2022(11): 114005.

[31] ZHANG Q L, YANG Y B. ResT: an efficient transformer for visual recognition[EB/OL]. (2021-05-28) [2022-05-21]. https://arxiv.org/abs/2105.13677.pdf.

[32] WOO S, PARK J, LEE J Y, et al. CBAM: Convolutional block attention module[C]//European Conference on Computer Vision. Cham: Springer, 2018: 3-19.

[33] PARK J, WOO S, LEE J Y, et al. BAM: Bottleneck attention module[EB/OL]. (2018-07-17) [2022-05-21]. https://arxiv.org/abs/1807.06514.pdf.

[34] LI X, WANG W H, HU X L, et al. Selective kernel networks[C]// 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2020: 510-519.

[35] YAO H T, ZHANG S L, HONG R C, et al. Deep representation learning with part loss for person re-identification[J]. IEEE Transactions on Image Processing, 2019, 28(6): 2860-2871.

[36] ZHANG Z Z, LAN C L, ZENG W J, et al. Densely semantically aligned person re-identification[C]// 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2020: 667-676.

[37] YE M, LAN X Y, WANG Z, et al. Bi-directional center-constrained top-ranking for visible thermal person re-identification[J]. IEEE Transactions on Information Forensics and Security, 2020, 15: 407-419.

[38] ZHENG L, SHEN L Y, TIAN L, et al. Scalable person re-identification: A benchmark[C]// 2015 IEEE International Conference on Computer Vision. Piscataway: IEEE Press, 2016: 1116-1124.

[39] ZHENG Z D, ZHENG L, YANG Y. Unlabeled samples generated by GAN improve the person re-identification baseline in vitro[C]// 2017 IEEE International Conference on Computer Vision. Piscataway: IEEE Press, 2017: 3774-3782.

[40] MIAO J X, WU Y, LIU P, et al. Pose-guided feature alignment for occluded person re-identification[C]// 2019 IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE Press, 2020: 542-551.

[41] WANG X G, DORETTO G, SEBASTIAN T, et al. Shape and appearance context modeling[C]// 2007 IEEE 11th International Conference on Computer Vision. Piscataway: IEEE Press, 2007: 1-8.

[42] ZHAO H Y, TIAN M Q, SUN S Y, et al. Spindle net: Person re-identification with human body region guided feature decomposition and fusion[C]// 2017 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2017: 907-915.

[43] WANG G S, YUAN Y F, CHEN X, et al. Learning discriminative features with multiple granularities for person re-identification[C]// Proceedings of the 26th ACM international conference on Multimedia. New York: ACM, 2018: 274-282.

[44] ZHAO L M, LI X, ZHUANG Y T, et al. Deeply-learned part-aligned representations for person re-identification[C]// 2017 IEEE International Conference on Computer Vision. Piscataway: IEEE Press, 2017: 3239-3248.

[45] JIA M X, CHENG X H, LU S J, et al. Learning disentangled representation implicitly via transformer for occluded person re-identification[J]. IEEE Transactions on Multimedia, 2023, 25: 1294-1305.

A person re-identification method for fusing convolutional attention and Transformer architecture

WANG Jing，LI Peitong，ZHAO Rongfeng，ZHANG Yun*，MA Zhenling

(College of Information Technology, Shanghai Ocean University, Shanghai 201306, China)

Abstract：Person Re-identification technology is one of the important methods in intelligent security systems. In order to build a person re-identification model suitable for various complex scenarios, this article proposed a method of Fusing Convolutional Attention and Transformer architecture (FCAT) based on existing convolutional neural networks and Transformer models to enhance the Transformer's attention to local detail information. This method mainly improves the transformer's ability to extract local detail features indirectly by embedding convolutional space attention and channel attention respectively to enhance the attention to important regions and important channel features in the image. Comparative ablation experiments on three publicly available pedestrian re-identification datasets demonstrate that the proposed method achieves comparable results on non-occluded datasets and significantly improves performance on occluded datasets. Additionally, the proposed model is more lightweight, leading to improved inference speed without increasing additional computational load or model parameters.

Keywords：person re-identification；deep learning；convolutional neural network；transformer；attention mechanisms

Received：2022-06-02；**Accepted**：2022-07-25；**Published Online**：2022-08-11 17：29

URL：link.cnki.net/urlid/11.2625.V.20220810.1132.005

Foundation items：National Natural Science Foundation of China (61806123,42101443); National Key R&D Program of China (2019YFD0900805)

* **Corresponding author**. E-mail：y-zhang@shou.edu.cn

http://bhxb.buaa.edu.cn jbuaa@buaa.edu.cn
DOI: 10.13700/j.bh.1001-5965.2022.0399

周期规律增强的多视角短期电力负荷预测

苏伟*，肖小龙，史明明，方鑫，司鑫尧
(国网江苏省电力有限公司 电力科学研究院，南京 210000)

摘 要：短期电力负荷预测对电力系统的可靠运行具有重要意义。现有方法存在如下问题：缺乏对特征之间依赖关系的挖掘；忽略了电力负荷变化的周期性规律。为此，提出一种周期规律增强的多视角短期电力负荷预测网络（EPISODE）方法。EPISODE方法主要包括2个核心组件：多视角特征学习组件和周期规律增强的电力负荷预测组件。前者旨在有效提取电力负荷数据中的静态特征与时序特征，以得到增强的特征表示；后者则是对电力负荷数据进行一般性时序挖掘和周期性时序挖掘，从而得到全面的电力负荷历史数据表征。基于后期融合的方式，实现短期电力负荷预测。在真实公开的电力负荷预测数据集上进行了大量实验。实验结果证明了所提方法相比现有基准方法的先进性。

关 键 词：周期规律增强；短期负荷预测；特征压缩与激励；门控循环单元；注意力神经网络
中图分类号：TP399
文献标志码：A **文章编号**：1001-5965(2024)02-0477-10

电力系统是中国经济发展的重要基石。其中，电力能源也是重要的能源之一。准确的电力负荷预测不仅可以指导电力生产部门制定合理的发电计划，降低发电成本，还有利于提高电力系统运行的安全稳定性[1]。根据负荷预测的时间跨度，电力负荷预测可大致分为：长期负荷预测、中长期负荷预测、短期负荷预测和超短期负荷预测。其中，长期负荷预测通常指未来若干年的负荷预测；中长期负荷预测是未来若干月或若干年的负荷预测；短期负荷预测是对未来若干日或若干周的负荷预测；超短期负荷预测是指对未来若干秒或若干分钟的电力预测。由于交流电不易存储的特点，电力系统需要及时调整生产调度。此时，调度的主要依据来源于短期电力负荷预测的结果。因此，短期电力负荷预测对于电力系统的重要性不言而喻。

传统的短期电力负荷预测方法主要分为基于统计理论的方法和基于机器学习的方法。基于统计理论的方法主要包括多元线性回归 (multiple linear regression, MLR)[2]、自回归移动平均值 (auto-regressive moving average, ARMA)[3]、差分整合移动平均自回归 (autoregressive integrated moving average, ARIMA)[4] 等。这类方法简单、效率高，但对非线性因素处理的能力较弱，很难取得好的实际应用效果。基于机器学习的方法主要包括人工神经网络 (artificial neural network, ANN) [5]、支持向量机 (support vector machine, SVM)[6-7]、随机森林 (random forest, RF)[8] 等。这类方法虽然考虑了负荷的时序性和其他的非线性因素，但其只能对少量数据及特征进行处理，而且训练时间长，对原始数据的平稳性要求也较高。

近年来，随着深度学习算法在诸多领域的成功应用，国内外许多学者都开始尝试利用深度学习算法解决短期电力负荷预测任务[9]。经典的深度学习算法包括卷积神经网络（convolutional neural net-

收稿日期：2022-05-21；**录用日期**：2022-07-02；**网络出版时间**：2022-12-14 15:58
网络出版地址：link.cnki.net/urlid/11.2625.V.20221213.1218.004
基金项目：国网江苏省电力有限公司科技项目 (J2021165)
* **通信作者**. E-mail：susw2011@163.com

引用格式：*苏伟，肖小龙，史明明，等. 周期规律增强的多视角短期电力负荷预测 [J]. 北京航空航天大学学报，2024，50（2）：477-486.*
SU W，XIAO X L，SHI M M，et al. Periodic pattern-enhanced multi-view short-term load prediction [J]. Journal of Beijing University of Aeronautics and Astronautics，2024，50（2）：477-486（in Chinese）.

work, CNN)[10]和循环神经网络(recurrent neural network, RNN)[11]等。其中,CNN凭借其优秀的局部特征提取能力,被广泛用于计算机视觉领域;RNN及其变种模型,如长短期记忆(long short-term memory, LSTM)[12]网络和门控循环单元(gated recurrent unit, GRU)[13]网络,则更加适用于挖掘序列数据,因而被广泛用于自然语言处理和时间序列预测等领域。为了增强模型对电力负荷历史的表征,研究学者联合多种深度神经网络来实现电力负荷的有效预测[14]。虽然这些基于深度学习算法的工作已经取得了显著的短期电力负荷预测效果,但是这些工作主要存在2个局限:①缺乏对模型所学各维度潜在特征之间的依赖关系的挖掘,难以达到最佳的特征提取效果;②忽略了电力负荷变化的周期性规律。比如,电力负荷往往在晚间或周末较为突出。

为了解决上述局限,本文提出了一种周期规律增强的多视角短期电力负荷预测网络(periodic pattern-enhanced multi-view short-term power load prediction networks, EPISODE)方法,该预测方法包括2个关键的核心组件:多视角特征学习组件和周期规律增强的电力负荷预测组件。其中,多视角特征学习组件旨在从多个视角提取电力数据中的特征信息。分别利用设计的1D-CNN网络与GRU网络,提取电力数据中包含的静态特征与时序特征,并利用上下文特征融合单元对2种特征进行融合,以增强特征的表示学习。而在周期规律增强的电力负荷预测组件中,除了利用GRU进行一般性时序历史编码外,还使用GRU-Skip来进行周期性时序历史编码,并利用后期融合的方式对二者进行融合,实现最终的电力负荷预测。为了验证模型的有效性,本文在公开数据集上开展了大量实验。结果表明相比现有方法,本文方法不仅在预测精度而且在预测效率上都具有明显的优越性。

1 相关工作

早期的电力负荷预测主要使用统计学理论或机器学习方法来进行电力负荷预测。例如,彭鹏和彭佳红[9]根据GDP、人口和全社会用电量等变量的历史数据建立多元线性回归模型,以进行电力负荷预测;李龙等[15]利用传统的人工神经网络对电力负荷进行预测。虽然这类方法取得了一定的预测效果,但这类方法普遍不能很好地处理非线性因素对电力负荷值的影响,导致其预测值与真实值的差异较大。

随着深度学习技术的蓬勃发展,越来越多学者开始采用深度学习作为主要技术路径,以解决电力负荷预测任务。陆继翔等[16]提出CNN-LSTM这一混合神经网络模型,其先通过CNN来提取可能影响电力负荷的特征,再将其作为LSTM网络的输入,从而实现短期电力负荷预测。李泽文等[17]提出多因素短期负荷预测模型,其先运用主成分分析法对原始多维输入变量的时间序列进行主成分提取,实现负荷数据的降维,然后通过双向LSTM网络,最终得到预测结果。杨洋和谷震浩[18]提出XGBoost-DNN,其首先利用XGBoost学习到从数据到目标的交叉特征,然后利用这份预训练学习到的交叉特征结合原数据通过DNN的特性去学习中期电力负荷的时序关系,最后得到预测结果。

虽然上述方法取得了一定的成功,但所提出的模型对时序特征和静态特征的利用仍然不充分,这导致模型挖掘数据中重要信息的能力下降。同时,上述工作忽略了电力负荷变化存在一定的周期性规律。因此,为了提升电力负荷预测的精度,本文方法引入了特征压缩与激励机制,增强对重要特征的筛选,利用GRU-Skip实现了对电力负荷数据中存在的周期性信息进行建模,同时基于后期融合方式进行电力负荷预测,进而实现了更加精准的电力负荷预测效果。

2 多视角短期电力负荷预测框架

2.1 问题定义

为能够及时地指导电力调度,节约能源,本文重点研究短期电力负荷预测,尤其是预测未来一天内的电力负荷。假设有一段时期(如几年)内某区域电力负荷的采样序列为$\{\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_Q\}$。该序列的采样间隔设为$D$,单位为min,其中$D$可以被60整除。这样每天的电力负荷一共有$60/D\times24$个采样点。与此同时,电力负荷采样序列各时刻相对应的天气相关数据序列,设为$\{\boldsymbol{x}'_1,\boldsymbol{x}'_2,\cdots,\boldsymbol{x}'_Q\}$,其中$\boldsymbol{x}'_i\in\mathbf{R}^d$ $(i=1,2,\cdots,Q)$表示每一时刻的天气相关多变量数据,包括温度和湿度等。d为该天气多变量的维度。由于天气相关因素大大影响电力负荷的大小,这些多变量数据可以用于辅助预测未来的电力负荷。

为了构建训练样本集合,对上述采样序列和天气相关数据按照滑动窗口的方式进行分割,其中,每个窗口覆盖M天的历史电力负荷数据和天气相关数据,即$M(60/D\times24)$个采样点数据。最终,假设获得一个包含N个训练数据样本的集合$S=\{(\boldsymbol{X}^1,\boldsymbol{y}^1),(\boldsymbol{X}^2,\boldsymbol{y}^2),\cdots,(\boldsymbol{X}^N,\boldsymbol{y}^N)\}$。其中,$\boldsymbol{X}^j=[\boldsymbol{x}_1^j,\boldsymbol{x}_2^j,\cdots,\boldsymbol{x}_T^j]\in\mathbf{R}^{(d+1)\times T}$ $(j=1,2,\cdots,N)$表示$T=M(60/D\times24)$个连续历史时刻的电力负荷和天气信息历史数据,

$\boldsymbol{x}_t^j=\left[X_{(j-1)*T+t};\boldsymbol{x}'_{(j-1)*T+t}\right]\in\mathbf{R}^{d+1}(j=1,2,\cdots,T)$表示第$t$个时刻观察到的电力负荷采样值和相关天气多变量数据。$\boldsymbol{y}^j=\left[X_{j*T+1},X_{j*T+2},\cdots,X_{j*T+\tau}\right]\in\mathbf{R}^{\tau}(j=1,2,\cdots,N)$表示对应的未来一天内$\tau$个时刻的真实电力负荷采样值, 其中, $\tau=60/D\times24$。本文旨在利用这些训练数据优化一个模型F, 使其可以捕捉未来短期电力负荷变化与历史负荷和天气等数据之间的潜在联系, 即能够根据历史T个时刻的负荷历史和天气信息即$\boldsymbol{X}$, 预测未来τ个时刻的电力负荷$\boldsymbol{y}$。模型的表达式为

$$F(\boldsymbol{X},\boldsymbol{\Theta})\rightarrow\boldsymbol{y} \tag{1}$$

式中: $\boldsymbol{\Theta}$为待学习的参数。

2.2　模型结构

如图 1 所示, EPISODE 方法包含 2 个组件: 多视角特征学习组件和周期规律增强的电力负荷预测组件。不同于传统的单一特征提取方式, 为了充分利用 CNN 的静态局部特征提取能力和门控循环单元的动态时序信息挖掘能力, 在特征学习组件中, 同时采用基于一维 CNN 和基于 GRU 的 2 种特征提取手段。其中, 将基于一维 CNN 提取的特征当作上下文信息, 用于增强基于 GRU 的特征提取。值得一提的是, 除了使用一维 CNN, 本文还引入了特征压缩与激励(squeeze-and-excitation)技术, 该技术在一定程度上弥补了 CNN 无法很好捕捉各特征之间的依赖关系的不足, 促进该组件的特征提取能力。同时, 为了深入挖掘电力负荷使用数据的变化趋势, 本文联合使用基于 GRU 的一般性时序历史编码和基于 GRU-Skip 的周期性时序历史编码, 以增强模型对电力负荷历史的表征能力。借助注意力机制, 对基于 GRU-Skip 挖掘到的周期性时序历史编码表示, 进行动态融合, 以实现更加准确的周期性时序历史表征。最后, 采用后期融合的方式, 实现对电力负荷的预测。

2.3　多视角特征学习

如图 1 所示, 本文设计的多视角特征学习组件主要由若干特征提取层构成。假设一共有K层特征提取层。每个特征提取层主要由 2 个模块构成: 基于 1D-CNN 的静态特征提取模块和基于 GRU 的时序特征提取模块。

2.3.1　基于 1D-CNN 的静态特征提取

如图 2 所示, 本文设计的基于 1D-CNN 的静态特征提取模块由一维卷积层、批量归一化层和压缩激励层组成。

该模块可定义为

$$\begin{cases}\widehat{\boldsymbol{X}}^0=\boldsymbol{X}\\ \widehat{\boldsymbol{X}}^{k+1}=G\left(\widehat{\boldsymbol{X}}^k\right)\qquad k=1,2,\cdots,K\end{cases} \tag{2}$$

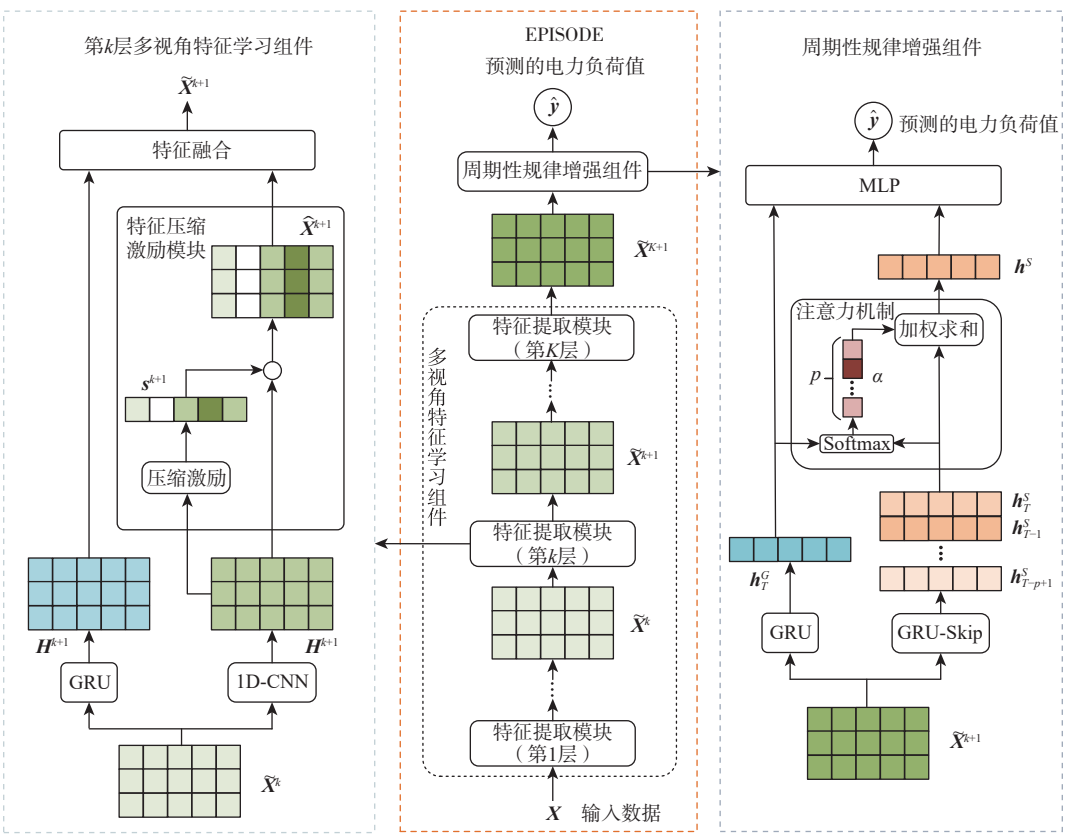


图 1　EPISODE 整体网络结构图

Fig. 1　Overall network structure diagram of EPISODE

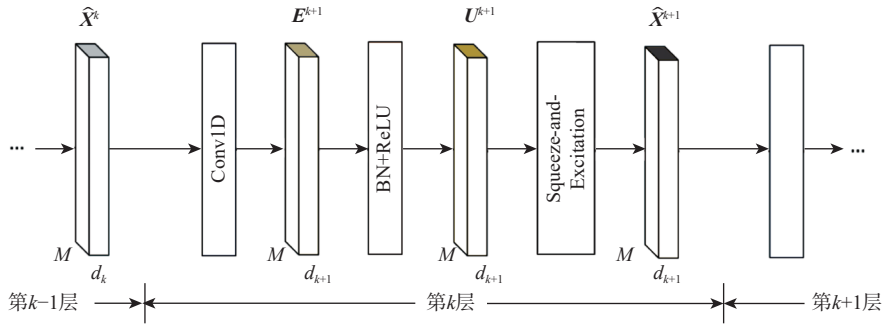


图2 基于一维卷积神经网络的静态特征提取模块

Fig. 2 1D-CNN based static feature extraction module

式中：$\widehat{\boldsymbol{X}}^k$表示第$k-1$层的静态特征提取操作的输出，也是第$k$层静态特征提取操作的输入。$\widehat{\boldsymbol{X}}^0=\boldsymbol{X}\in\mathbf{R}^{(d+1)\times T}$，表示原始的输入数据。$G$表示基于一维卷积的静态特征提取网络。该网络主要包含3个部分：一维卷积、批量归一化和特征的压缩与激励。

传统的CNN主要应用于图像等二维数据，其卷积核可以沿2个方向进行滑动，而一维卷积操作的卷积核则只按照一个维度进行滑动。在本任务背景下，将卷积核沿着时间的维度进行滑动，以此获得电力负荷及其相关天气变量的时序变化规律。为了方便后续说明，将第k层的一维卷积操作定义为

$$\boldsymbol{E}^{k+1}=\mathrm{CNN1}\left(\widehat{\boldsymbol{X}}^k\right)\qquad k=1,2,\cdots,K \tag{3}$$

式中：$\boldsymbol{E}^{k+1}=\left[\boldsymbol{e}_1^{k+1},\boldsymbol{e}_2^{k+1},\cdots,\boldsymbol{e}_T^{k+1}\right]\in\mathbf{R}^{d_{k+1}\times T}$为$\widehat{\boldsymbol{X}}^k$通过第$k$层1D-CNN得到的潜在表示特征。

本文引入一层批量归一化层以防止梯度爆炸，并将输出结果表示为$\boldsymbol{U}^{k+1}=\left[\boldsymbol{u}_1^{k+1},\boldsymbol{u}_2^{k+1},\cdots,\boldsymbol{u}_T^{k+1}\right]\in\mathbf{R}^{d_{k+1}\times T}$，具体计算如式(4)所示：

$$\boldsymbol{U}^{k+1}=\delta\left(\mathrm{BN}\left(\boldsymbol{E}^{k+1}\right)\right)\qquad k=1,2,\cdots,K \tag{4}$$

式中：BN表示批量归一化层；δ表示ReLU激活函数。

虽然上述运算一定程度上可以将特征的重要信息融合至局部空间中，但其不能很好地捕获特征之间的关联性，因而无法有选择地突出重要的信息特征和抑制不太重要的特征。因此，为了更好地捕捉数据特征，本文采用压缩激励技术[19]，如图3所示。本模块中的压缩激励操作主要涉及2个子操作：特征压缩(squeeze)操作F_{sq}和特征激励(excitation)操作F_{ex}。

特征压缩操作旨在捕获每维特征的全局性代表信息，其可通过使用全局平均池化等操作来实现。具体而言，对于通过批量归一化层获得的输出数据$\boldsymbol{U}^{k+1}$，将其在时间维度上进行压缩，以获得每维特征的全局性表示。假设该全局信息表示向量

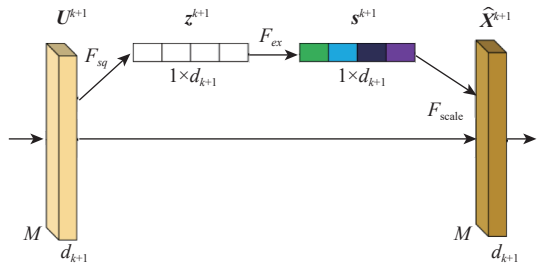


图3 压缩激励模块结构

Fig. 3 Squeeze-and-excitation module structure

为$\boldsymbol{z}^{k+1}=\left[z_1^{k+1},z_2^{k+1},\cdots,z_{d_{k+1}}^{k+1}\right]\in\mathbf{R}^{d_{k+1}}$，其中，$\boldsymbol{z}^{k+1}$的第$j$个元素$z_j^{k+1}$计算式为

$$z_j^{k+1}=F_{\mathrm{sq}}\left(\boldsymbol{U}^{k+1}\left[j,:\right]\right)=\mathrm{avg}\left(\boldsymbol{U}^{k+1}\left[j,:\right]\right)\qquad j=1,2,\cdots,d_{k+1} \tag{5}$$

特征激励操作通过挖掘特征之间的依赖关系，以更好地激励对电力负荷预测较重要的特征，同时抑制不太重要的特征。以第k个特征提取层为例，该特征激励操作可以通过使用如下的门控机制实现：

$$\boldsymbol{s}^{k+1}=F_{\mathrm{ex}}\left(\boldsymbol{z}^{k+1}\right)=\sigma\left(\boldsymbol{W}_2^k\delta\left(\boldsymbol{W}_1^k\boldsymbol{z}^{k+1}\right)\right) \tag{6}$$

式中：σ为Sigmoid激活函数；δ为ReLU激活函数；$\boldsymbol{W}_1^k\in\mathbf{R}^{\frac{d_{k+1}}{r_k}\times d_{k+1}}$和$\boldsymbol{W}_2^k\in\mathbf{R}^{d_{k+1}\times\frac{d_{k+1}}{r_k}}$为第$k$层特征提取单元中该操作的相关参数，$\boldsymbol{W}_1^k$和$\boldsymbol{W}_2^k$主要用于提升泛化能力，并使输出的向量$\boldsymbol{s}^{k+1}=[s_1^{k+1},s_2^{k+1},\cdots,s_{d_{k+1}}^{k+1}]\in\mathbf{R}^{d_{k+1}}$与$\boldsymbol{z}^{k+1}$的维度保持一致；$r_k$为第$k$层特征提取单元中该操作的超参数，代表缩减比例，主要用于减少模型的计算复杂度，取值范围是正整数；$\boldsymbol{s}^{k+1}\in\mathbf{R}^{d_{k+1}}$表示各潜在表示特征的重要性。最后，根据$\boldsymbol{s}^{k+1}$中的权重信息对潜在表示特征$\boldsymbol{U}^{k+1}$进行相应的激励，具体如式(7)所示：

$$\widehat{\boldsymbol{X}}^{k+1}\left[j,:\right]=F_{\mathrm{scale}}\left(\boldsymbol{U}^{k+1}\left[j,:\right],s_j^{k+1}\right)=s_j^{k+1}\boldsymbol{U}^{k+1}[j,:]\qquad j=1,2,\cdots,d_{k+1} \tag{7}$$

式中：s_j^{k+1}为向量$\boldsymbol{s}^{k+1}$的第j个元素；$\widehat{\boldsymbol{X}}^{k+1}=\left[\widehat{\boldsymbol{x}}_1^{k+1},\widehat{\boldsymbol{x}}_2^{k+1},\cdots,\widehat{\boldsymbol{x}}_T^{k+1}\right]\in\mathbf{R}^{d_{k+1}\times T}$为第$k$层特征提取单元对于输

入数据提取的最终特征表示。

2.3.2　基于 GRU 的时序特征提取

针对短期电力负荷预测任务, 本文选用 GRU 作为时序特征提取网络。其原因在于传统 RNN 在实际中, 较难捕获时间序列中长期时序依赖关系。如图 4 所示, 本文设计的基于 GRU 的时序特征提取模块由 GRU 单元和上下文特征融合单元 F_{cd} 组成。GRU 用于挖掘特征潜在的时序性依赖关系, 而上下文特征融合单元 F_{cd} 用于将基于 ID-CNN 捕获的静态特征作为上下文信息与 GRU 提取的时序特征进行融合, 进而增强特征的表示学习。

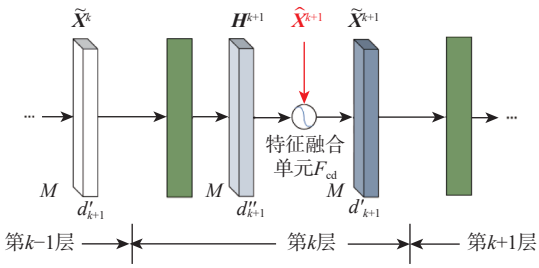


图 4　基于 GRU 的时序特征提取模块结构

Fig. 4　GRU based time series feature extraction module structure

整体而言, 第 k 层特征提取单元中该模块的定义, 如式(8)所示:

$$\begin{cases} \widetilde{\boldsymbol{X}}^0 = \boldsymbol{X} \\ \widetilde{\boldsymbol{X}}^{k+1} = H\left(\widetilde{\boldsymbol{X}}^k\right) \qquad k = 0,1,\cdots,K \end{cases} \tag{8}$$

式中: $\widetilde{\boldsymbol{X}}^k = \left[\widetilde{\boldsymbol{x}}_1^k, \widetilde{\boldsymbol{x}}_2^k, \cdots, \widetilde{\boldsymbol{x}}_T^k\right] \in \mathbf{R}^{d_k' \times T}$ 表示通过第 $k-1$ 层时序特征提取模块提取到的潜在特征表示; $\widetilde{\boldsymbol{X}}^0 = \boldsymbol{X} \in \mathbf{R}^{(d+1)\times T}$ 表示原始的输入数据; H 表示基于 GRU 的特征提取操作。

具体而言, 第 k 层特征提取单元中的 GRU 操作可定义为

$$\boldsymbol{H}^{k+1} = \mathrm{GRUf}\left(\widetilde{\boldsymbol{X}}^k\right) \qquad k = 1,2,\cdots,K \tag{9}$$

式中: $\boldsymbol{H}^{k+1} = \left[\boldsymbol{h}_1^{k+1}, \boldsymbol{h}_2^{k+1}, \cdots, \boldsymbol{h}_T^{k+1}\right] \in \mathbf{R}^{d_{k+1}'' \times T}$ 为第 k 层 GRU 输出的整体隐藏状态表示, 蕴含输入数据的时序特征; $\boldsymbol{h}_t^{k+1} \in \mathbf{R}^{d_{k+1}''}$ 为第 k 层 GRU 对应第 $t(t = 1,2,\cdots,T)$ 个时刻的隐藏状态输出。上述公式可进一步详细写为

$$\begin{cases} \boldsymbol{r}_t^{k+1} = \sigma\left(\widetilde{\boldsymbol{x}}_t^k \boldsymbol{W}_{\mathrm{xr}}^{k+1} + \boldsymbol{h}_{t-1}^{k+1} \boldsymbol{W}_{\mathrm{hr}}^{k+1} + \boldsymbol{b}_{\mathrm{r}}^{k+1}\right) \\ \boldsymbol{v}_t^{k+1} = \sigma\left(\widetilde{\boldsymbol{x}}_t^k \boldsymbol{W}_{\mathrm{xv}}^{k+1} + \boldsymbol{h}_{t-1}^{k+1} \boldsymbol{W}_{\mathrm{hv}}^{k+1} + \boldsymbol{b}_{\mathrm{v}}^{k+1}\right) \\ \boldsymbol{z}_t^{k+1} = \tanh\left(\widetilde{\boldsymbol{x}}_t^k \boldsymbol{W}_{\mathrm{xz}}^{k+1} + \boldsymbol{r}_t^{k+1} \odot \left(\boldsymbol{h}_{t-1}^{k+1} \boldsymbol{W}_{\mathrm{hz}}^{k+1}\right) + \boldsymbol{b}_{\mathrm{z}}^{k+1}\right) \\ \boldsymbol{h}_t^{k+1} = \left(1 - \boldsymbol{v}_t^{k+1}\right) \odot \boldsymbol{h}_{t-1}^{k+1} + \boldsymbol{v}_t^{k+1} \odot \boldsymbol{z}_t^{k+1} \end{cases} \tag{10}$$

式中: $\boldsymbol{W}_{\mathrm{xr}}^{k+1}$, $\boldsymbol{W}_{\mathrm{hr}}^{k+1}$, $\boldsymbol{W}_{\mathrm{xv}}^{k+1}$, $\boldsymbol{W}_{\mathrm{hv}}^{k+1}$, $\boldsymbol{W}_{\mathrm{xz}}^{k+1}$, $\boldsymbol{W}_{\mathrm{hz}}^{k+1}$, $\boldsymbol{b}_{\mathrm{r}}^{k+1}$, $\boldsymbol{b}_{\mathrm{v}}^{k+1}$, $\boldsymbol{b}_{\mathrm{z}}^{k+1}$ 为待学习的相关参数。

最后, 将每一层基于 GRU 提取的时序特征表示 $\boldsymbol{H}^{k+1}$ 与基于一维卷积神经网络提取的静态特征 $\widehat{\boldsymbol{X}}^{k+1}$ 通过特征融合单元 F_{cd}, 具体操作是将二者级联, 再通过一层全连接层:

$$\begin{aligned} \widetilde{\boldsymbol{X}}^{k+1} &= F_{\mathrm{cd}}\left(\boldsymbol{H}^{k+1}, \widehat{\boldsymbol{X}}^{k+1}\right) = \\ &\left[\boldsymbol{H}^{k+1}, \widehat{\boldsymbol{X}}^{k+1}\right] \boldsymbol{W}_{hx}^k + \boldsymbol{b}_{hx}^k \qquad k = 1,2,\cdots,K \end{aligned} \tag{11}$$

式中: $\boldsymbol{W}_{hx}^k \in \mathbf{R}^{(d_{k+1}+d_{k+1}'')\times T}$ 和 $\boldsymbol{b}_{hx}^k \in \mathbf{R}^T$ 为相关待学习参数, 主要用于控制特征连接后的维度, 并进一步提升泛化能力。

最后, $\widetilde{\boldsymbol{X}}^{K+1} = \left[\widetilde{\boldsymbol{x}}_1^{K+1}, \widetilde{\boldsymbol{x}}_2^{K+1}, \cdots, \widetilde{\boldsymbol{x}}_T^{K+1}\right] \in \mathbf{R}^{d_{K+1}' \times T}$ 即为通过 K 层特征提取单元得到的最终数据表示特征。

2.4　周期规律增强的电力负荷预测

本文设计的周期规律增强的电力负荷预测组件包含 3 个模块: 基于 GRU 的一般性时序历史编码、基于 GRU-Skip 的周期性时序历史编码和基于后期融合的电力负荷预测。

2.4.1　基于 GRU 的一般性时序历史编码

基于 GRU 的一般性时序历史编码旨在基于通过多视角特征学习组件得到的最终数据特征表示, 对已知的 T 个时刻的电力负荷数据进行一般性编码。具体而言, 该模块可定义为

$$\left[\boldsymbol{h}_1^{\mathrm{G}}, \boldsymbol{h}_2^{\mathrm{G}}, \cdots, \boldsymbol{h}_T^{\mathrm{G}}\right] = \mathrm{GRUg}\left(\widetilde{\boldsymbol{X}}^{K+1}\right) \tag{12}$$

式中: $\boldsymbol{h}_t^{\mathrm{G}} \in \mathbf{R}^{d_{\mathrm{G}}}$ 表示基于 GRUg 获得的第 t 个时刻的隐藏状态表示。本文采用最后一个时刻的隐藏状态表示来表征时序历史信息。

2.4.2　基于 GRU-Skip 的周期性时序历史编码

由于电力负荷往往具有周期性特点, 比如电力负荷往往在晚间或周末达到峰值。因此, 电力负荷数据之间存在周期性的依赖关系, 而这些关系难以被 GRU 捕获。为解决该问题, 本文引入了 GRU-Skip 模块, 该模块已被文献 [20] 证明了其对周期性规律编码的有效性。该模块的具体操作可以通过式(13)定义:

$$\begin{cases} \boldsymbol{r}_t^{\mathrm{S}} = \sigma\left(\widetilde{\boldsymbol{x}}_t^K \boldsymbol{W}_{\mathrm{xr}}^{\mathrm{S}} + \boldsymbol{h}_{t-p}^{\mathrm{S}} \boldsymbol{W}_{\mathrm{hr}}^{\mathrm{S}} + \boldsymbol{b}_{\mathrm{r}}^{\mathrm{S}}\right) \\ \boldsymbol{v}_t^{\mathrm{S}} = \sigma\left(\widetilde{\boldsymbol{x}}_t^K \boldsymbol{W}_{\mathrm{xv}}^{\mathrm{S}} + \boldsymbol{h}_{t-p}^{\mathrm{S}} \boldsymbol{W}_{\mathrm{hv}}^{\mathrm{S}} + \boldsymbol{b}_{\mathrm{v}}^{\mathrm{S}}\right) \\ \boldsymbol{z}_t^{\mathrm{S}} = \tanh\left(\widetilde{\boldsymbol{x}}_t^K \boldsymbol{W}_{\mathrm{xz}}^{\mathrm{S}} + \boldsymbol{r}_t^{\mathrm{S}} \odot \left(\boldsymbol{h}_{t-p}^{\mathrm{S}} \boldsymbol{W}_{\mathrm{hz}}^{\mathrm{S}}\right) + \boldsymbol{b}_{\mathrm{z}}^{\mathrm{S}}\right) \\ \boldsymbol{h}_t^{\mathrm{S}} = \left(1 - \boldsymbol{v}_t^{\mathrm{S}}\right) \odot \boldsymbol{h}_{t-p}^{\mathrm{S}} + \boldsymbol{v}_t^{\mathrm{S}} \odot \boldsymbol{z}_t^{\mathrm{S}} \end{cases} \tag{13}$$

式中: $\boldsymbol{h}_t^{\mathrm{S}} \in \mathbf{R}^{d_{\mathrm{S}}}$ 表示基于 GRU-Skip 获得的第 $t(t = p+1,\cdots,T)$ 个时刻的隐藏状态表示。从式 (13) 可以看出, 相较于传统的 GRU 网络, GRU-Skip 使用 $p(1 < p < T)$ 个时刻之前的隐藏状态, 而不是上一时刻的隐藏状态, 作为当前时刻的输入。p 为超参数, 表示电力负荷变化规律的周期。不同于一般性的时序历史编码(其默认时序变化周期为 1), 在周期性时序历史编码中, 选用最后 p 个时刻(最后一个周期)

的隐藏层状态$[\boldsymbol{h}^{\mathrm{S}}_{T-p+1},\boldsymbol{h}^{\mathrm{S}}_{T-p+2},\cdots,\boldsymbol{h}^{\mathrm{S}}_{T}]\in\mathbf{R}^{p\times d_{\mathrm{S}}}$，来全面表征整个历史（$T$个时刻）中包含的周期性信息。

由于这最后一个周期的每一时刻所包含的周期性信息对于整个电力负荷历史采样序列的表征，贡献程度不同，采用了注意力机制（attention mechanism, AM），将p个时刻的周期性历史信息进行动态聚合。如式（14）所示：

$$\begin{cases} f\left(\boldsymbol{h}^{\mathrm{S}}_{j},\boldsymbol{h}^{\mathrm{G}}_{T}\right)=\tanh\left(\boldsymbol{h}^{\mathrm{S}}_{j}\boldsymbol{W}_{\mathrm{SG}}\boldsymbol{h}^{\mathrm{G}}_{T}\right)\\ \alpha_j=\dfrac{\exp\left(f\left(\boldsymbol{h}^{\mathrm{S}}_{j},\boldsymbol{h}^{\mathrm{G}}_{T}\right)\right)}{\displaystyle\sum_{l=T-p+1}^{T}\exp\left(f\left(\boldsymbol{h}^{\mathrm{S}}_{\mathrm{l}},\boldsymbol{h}^{\mathrm{G}}_{T}\right)\right)}\\ j=T-p+1,T-p+2,\cdots,T \end{cases}\tag{14}$$

式中：$\boldsymbol{W}_{\mathrm{SG}}\in\mathbf{R}^{d_{\mathrm{S}}\times d_{\mathrm{G}}}$为待学习的参数。该设计的整体思想是参考一般性时序历史信息$\boldsymbol{h}^{\mathrm{G}}_{T}$，对各个时刻的周期性历史表征分配相应权重。接着，根据式(14)获得的注意力值α_j，对基于GRU-Skip获得的一系列周期性历史信息表征进行如下的加权融合：

$$\boldsymbol{h}^{\mathrm{S}}=\sum_{j=T-p+1}^{T}\alpha_j\boldsymbol{h}^{\mathrm{S}}_{j}\tag{15}$$

式中：$\boldsymbol{h}^{\mathrm{S}}\in\mathbf{R}^{d_{\mathrm{S}}}$表示基于GRU-Skip获得的最终表征向量。

2.4.3 基于后期融合的电力负荷预测

本文分别通过线性运算得到基于一般性时序历史和基于周期性时序历史的预测结果，然后采用后期融合的方式，对2种预测结果进行融合。具体如式（16）所示：

$$\hat{\boldsymbol{y}}=\boldsymbol{W}_{\mathrm{G}}\boldsymbol{h}^{\mathrm{G}}_{T}+\boldsymbol{W}_{\mathrm{S}}\boldsymbol{h}^{\mathrm{S}}\tag{16}$$

式中：$\boldsymbol{W}_{\mathrm{G}}\in\mathbf{R}^{\tau\times d_{\mathrm{G}}}$；$\boldsymbol{W}_{\mathrm{S}}\in\mathbf{R}^{\tau\times d_{\mathrm{S}}}$；$\hat{\boldsymbol{y}}\in\mathbf{R}^{\tau}$表示最终对于未来$\tau$个时刻的电力负荷预测结果。

针对模型优化，本文采用均方误差作为损失函数，即

$$L=\frac{1}{N}\sum_{i=1}^{N}\|\boldsymbol{y}_i-\hat{\boldsymbol{y}}_i\|^2+\frac{\lambda}{2N}\|\boldsymbol{w}\|^2\tag{17}$$

式中：$\boldsymbol{y}_i\in\mathbf{R}^{\tau}$为第$i$个样本未来电力负荷的真实值；$\hat{\boldsymbol{y}}_i\in\mathbf{R}^{\tau}$为模型针对第$i$个样本未来电力负荷的预测值；$N$为训练样本总个数；$\dfrac{\lambda}{2N}\|\boldsymbol{w}\|^2$为$L_2$正则化项，防止过拟合问题，提高模型的泛化能力，$\lambda(\lambda>0)$为超参数，$\boldsymbol{w}$为权重参数向量。

3 实验结果与分析

3.1 数据集

本文选取的数据集为第九届“中国电机工程学会杯”全国大学生电工数学建模竞赛提供的标准数据集。该数据集包含某一地区从2012年1月1日至2015年1月10日的电力负荷数据（每15 min采样1次，即$D=15$，每日共计96个时刻）、气象因素数据（包括日最高温度、日最低温度、日平均温度、日相对湿度及日降雨量）。与现有工作相同，本文采纳的具体任务为：根据一段时间的电力负荷和气象历史数据，预测未来1天共计96个时刻的电力负荷瞬时值。针对模型验证，本文通过随机采样的方式将该数据集进行如下分割：80%作为训练集，10%作为验证集，10%作为测试集。

3.2 实验设置

超参数设置：本文将多视角特征学习组件中的特征提取层数K设为5。因而，有5个一维卷积层，其卷积核数量分别设为256、128、256、256、128；卷积核大小分别设为5、5、3、3、3；对于5个SE模块，其中的缩减比例r_k均设为16。对于5层GRUf，其神经元个数分别设为64、128、128、128、64；而对于5层特征融合单元中全连接层的神经元个数，分别设为96、128、128、128、96。

周期规律增强的电力负荷预测组件中GRUg和GRU-Skip模块中的神经元个数均设为64，GRU-Skip模块中的超参数p设为7。式（16）中的超参数λ设为0.01。由于本文定义的电力负荷历史的时间窗口M是一个重要的超参数。本文采用网格搜索策略，将该超参数分别设置为7、14、21、28、30，并开展了不同超参数设置下的实验。根据实验结果，设M为21。本文基于Pytorch框架实现的EPISODE方法。具体使用RMSProp优化器优化网络参数。学习速率被设置为0.001。

评价指标：由于电力负荷预测本质上是回归任务。因此，本文采用回归任务中常用的评价指标：平均绝对百分比误差指标（mean absolute percentage error, MAPE）M_{MAPE}和均方根误差（root mean square error, RMSE）R_{MSE}来评价模型。这2个评价指标的定义分别为

$$M_{\mathrm{MAPE}}=\frac{1}{n}\sum_{i=1}^{n}\frac{\left|y_i-\hat{y}_i\right|}{y_i}\times100\%\tag{18}$$

$$R_{\mathrm{MSE}}=\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_i-\hat{y}_i\right)^2}\tag{19}$$

式中：n为待预测的电力负荷采样点总个数；y_i为每一时刻电力负荷的真实值；$\hat{y}_i$为每一时刻电力负荷的预测值。

3.3 针对时间窗口大小的实验

将用于数据划分的时间窗口大小M分别设置为7、14、21、28和30，进而训练模型，并在测试集

上得到相应的模型表现。这样设置的依据是考虑到以一周、周的倍数及一月的数据进行预测符合实际生产周期规律。如表 1 所示, 当时间窗口大小 M =21 时, 即根据前 21 天的电力负荷和气象历史数据预测未来一天电力负荷的方案预测效果最好, 因此, 接下来得实验均以 M=21 的时间窗口对数据进行划分条件, 进而训练模型。

表 1　EPISODE 在不同时间窗口的预测精度对比

Table 1　Comparison of prediction accuracy of EPISODE in different window

时间窗口	MAPE/%	RMSE/MW
M =7	1.523 4	165.724 9
M = 14	1.399 3	112.493 8
M = 21	1.011 5	81.246 0
M = 28	1.305 3	115.563 7
M = 30	1.140 0	100.872 8

为了直观显示不同时间窗口大小条件下的预测效果, 本文在测试集中随机选取一天作为待预测对象, 并根据相应的实验结果绘制图 5。根据图 5 中的曲线对比, 当时间窗口大小 M =21 时, 预测效果最佳。因此, 接下来的实验均以 M =21 的时间窗口对数据进行划分条件, 进而训练模型。

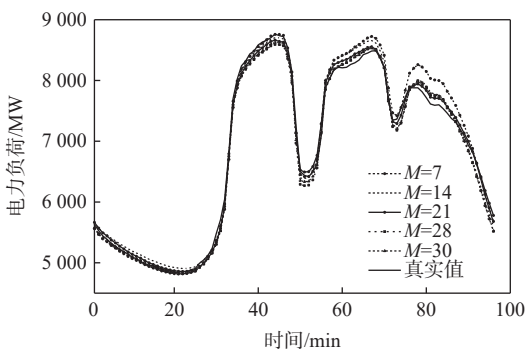


图 5　EPISODE 在不同时间窗口下的预测结果

Fig. 5　Prediction results of EPISODE in different windows

3.4　与基准方法对比试验

本文选择以下的基准方法进行比较, 来评估本文方法的有效性。本文选择的基准方法具体描述如下:

1) Attention-LSTM[14]: 该方法将电力负荷历史数据输入 2 层 LSTM, 1 层注意力机制网络和 1 个全连接层, 对电力负荷进行预测。

2) FOMNN[21]: 该方法分别使用多个卷积层和 GRU 模块来提取影响电力负荷的静态特征和时序特征, 然后通过注意力机制进行特征融合, 再通过 1 层 maxout 层和全连接层输出预测结果。

3) GRU-NN[22]: 该方法先单独将历史负荷通过一层 GRU 网络, 再与其他特征拼接, 再通过多层深度神经网络, 最后输出预测结果。

4) Dropout-ILSTM[23]: 该方法通过将 2 层 LSTM 的输出与输入进行加权求和, 得到输出结果, 并利用 Dropout 方法防止过拟合。

5) Attention-BiLSTM-LSTM[24]: 该方法先将所有输入依次通过 1 层双向 LSTM 和 1 层 LSTM, 然后将二者的输出通过注意力机制进行动态结合, 最后输出预测结果。

6) CNN-BiLSTM[25]: 该方法将所有输入依次通过 2 层一维卷积层、1 层池化层和 1 层扁平层, 以提取静态特征, 再通过 1 层双向 LSTM 和全连接层, 最后输出预测结果。

表 2 为 EPISODE 方法和 6 个基准方法对于电力负荷预测的精度和测试时间对比结果。表格中汇报的时间为方法在测试集所有样本上运行的结果, 测试集共 91 组样本。

表 2　EPISODE 与基准方法在数据集上的预测精度和测试时间对比

Table 2　The comparison of EPISODE and baselines methods on prediction accuracy and test time

方法	MAPE/%	RMSE/MW	时间/s
EPISODE	1.011 5	81.246 0	0.295 3
Attention-LSTM	2.787 7	233.002 2	0.717 8
FOMNN	2.196 7	194.704 7	0.305 0
GRU-NN	6.244 9	472.842 9	0.762 4
Dropout-ILSTM	5.549 7	392.077 4	0.922 2
Attention-BiLSTM-LSTM	2.198 4	194.757 7	1.509 4
CNN-BiLSTM	4.179 2	422.531 1	2.718 1

由表 2 和图 6 可以看出 EPISODE 方法在数据集上的表现都取得了比其他方法更好的预测效果, 验证了 EPISODE 方法的有效性。特别地, EPISODE 性能优于同样挖掘静态特征与时序特征的基准方法 FOMNN, 这表明引入特征压缩与激励模块与周期性时序历史编码的重要性。

3.5　模型消融实验

为了验证 EPISODE 方法各个关键模块的有效性, 引入 3 个衍生模型: EPISODE_ NoSE, EPISODE_ NoGRU_Skip 和 EPISODE_NoAttention。

1) 对于模型 EPISODE_NoSE, 在 EPISODE 方法的基础上去掉了特征压缩与激励模块。

2) 对于模型 EPISODE_NoSE, 在 EPISODE 方法的基础上去掉了用于周期性时序历史编码的 GRU_Skip 模块, 并用传统的 GRU 代替。

3) 对于模型 EPISODE_NoAttention, 在 EPISODE 方法的基础上去掉了用于周期性时序历史表征学习的注意力机制, 转而使用全连接层代替。具体公

式如下：

$$\widetilde{\boldsymbol{h}} = \boldsymbol{W}_h' \boldsymbol{h}_T^G + \left(\sum_{i=T-p+1}^{T} \boldsymbol{W}_i^{\mathrm{S}} \boldsymbol{h}_i^{\mathrm{S}} + \boldsymbol{b}_{\mathrm{h}} \right) \tag{20}$$

式中：$\boldsymbol{W}_h' \in \mathbf{R}^{\tau \times d_{\mathrm{G}}}$；$\boldsymbol{W}_i^{\mathrm{S}} \in \mathbf{R}^{\tau \times d_{\mathrm{S}}}$；$\boldsymbol{b}_h \in \mathbf{R}^{\tau}$；$\tau$为未来待预测的电力负荷采样点个数。

表3为EPISODE和3个衍生模型对电力负荷预测的精度对比结果。图7为EPISODE和3个衍生模型对测试集中某一天的电力负荷的预测结果。

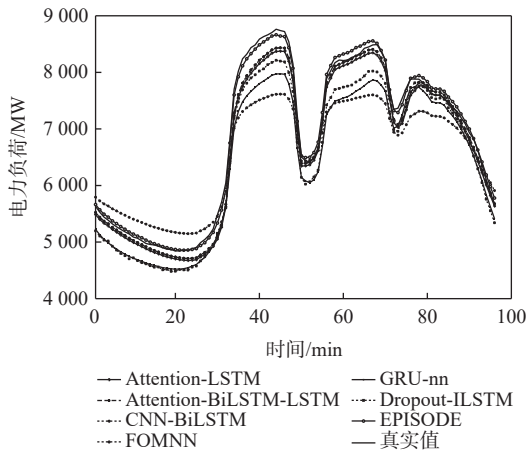


图6 EPISODE与基准方法的预测结果

Fig. 6 The load prediction results of EPISODE and baselines

表3 EPISODE与衍生模型的预测精度对比

Table 3 Comparison of prediction accuracy between EPISODE and derivations

模型	MAPE/%	RMSE/MW
EPISODE	1.011 5	81.246 0
EPISODE_ NoSE	1.454 2	123.208 9
EPISODE_NoGRU_Skip	1.622 3	168.620 0
EPISODE_NoAttention	1.642 7	188.648 5

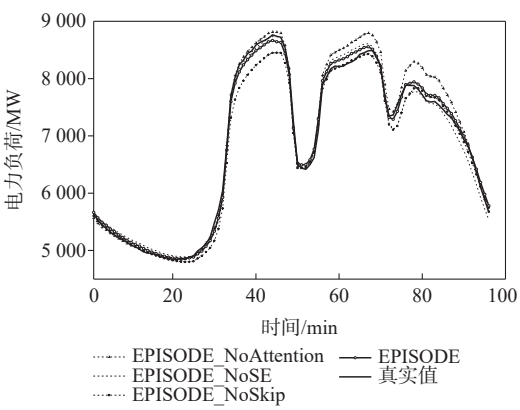


图7 EPISODE与衍生模型的预测结果

Fig. 7 The load prediction results of EPISODE and derivations

从表3和图7可以看出，EPISODE方法在数据集中的表现均高于3个衍生模型。这一实验结果说明添加特征压缩与激励模块，GRU-Skip模块，以及基于注意力机制的周期性历史编码方法的必要性。这些模块设计有效增强了模型对输入数据的特征提取能力，有助于实现更加准确的电力负荷预测。

3.6 敏感度分析实验

对EPISODE方法进行了敏感度分析，主要对其的关键超参数：多视角特征提取单元层数和GRU-Skip模块中表征周期的参数p，进行了相应的敏感度分析实验。

3.6.1 多视角特征提取单元层数K

本文将多视角特征提取单元层数K设置在[1,6]范围内，并以1为步长进行调节。如此设置是为了在保证模型尽可能轻量的前提下，实现精准的预测。图8为不同层数对电力负荷预测结果的影响。从图中可以看出，随着层数增加，预测精度也逐渐增加。但是当层数$K \geqslant 5$时，预测精度开始下降，即当$K=5$时，模型的预测精度最高。这说明，随着特征提取单元层数的增加，模型的泛化能力逐渐增强。但是当特征提取单元层数过大时，参数量过多，可能会发生过拟合现象。

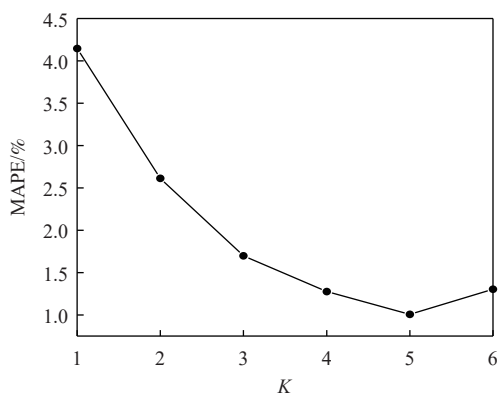


图8 特征提取单元层数的敏感性实验结果

Fig. 8 Sensitivity experiment results with different feature extraction layers

3.6.2 GRU-Skip模块中表征周期的参数p

在该项敏感度实验中，将GRU-Skip模块中表征周期的参数p（即跳过的时间步数）的取值分别设为5、7、10、14去训练模型。其中，取值为7和14是考虑到1周有7天，以7及7的倍数为1个周期符合现代生产规律。而取值为5和10则是分别对1周与2周天数的折中近似。图9为不同的p取值对电力负荷的预测结果的影响。从图中可以看出，模型的预测效果受p的取值影响较大，且在p为7时，模型预测值与真实值之间误差MAPE达到最小值，预测精度最高。这验证了结论：电力负荷数据具有一定的周期性特征，且利

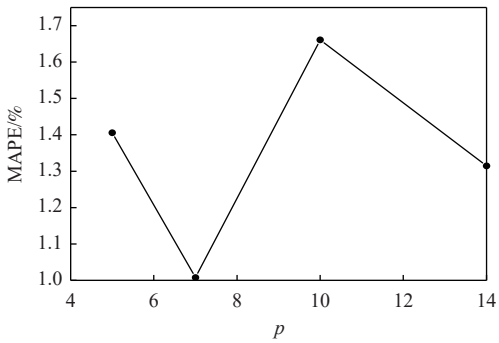


图 9　GRU-Skip 模块中p取值的敏感性实验结果

Fig. 9　Sensitivity experiments with different value of p in GRU-Skip module

用好周期性特征, 将有助于提升预测的精度。

4　结　论

1) 本文方法可实现优异的短期电力负荷预测效果, 可实现 1.0115% 的 MAPE 预测误差, 远高于同类的基准方法。

2) 消融实验证明本文方法中的各个组件对精准的电力负荷预测效果起到重要作用。

3) 本文方法目前只考虑了电力负荷的历史数据和相关气象数据, 未来拟引入更多影响因素, 例如节假日数据、当地经济数据、以及配电网拓扑结构等, 以实现更加准确的电力负荷预测。

参考文献（References）

[1] 康重庆, 夏清, 张伯明. 电力系统负荷预测研究综述与发展方向的探讨[J]. 电力系统自动化, 2004, 28(17): 1-11.
KANG C Q, XIA Q, ZHANG B M. Review of power system load forecasting and its development[J]. Automation of Electric Power Systems, 2004, 28(17): 1-11 (in Chinese).

[2] KEMP F. Applied multiple regression/correlation analysis for the behavioral sciences[J]. Journal of the Royal Statistical Society: Series D (the Statistician), 2003, 52(4): 691-691.

[3] DI CAPRIO U, GENESIO R, POZZI S, et al. Short term load forecasting in electric power systems: A comparison of ARMA models and extended Wiener filtering[J]. Journal of Forecasting, 1983, 2(1): 59-76.

[4] LEE Y S, TONG L I. Forecasting time series using a methodology based on autoregressive integrated moving average and genetic programming[J]. Knowledge-Based Systems, 2011, 24(1): 66-72.

[5] H ERNÁNDEZ L, BALADRÓN C, AGUIAR J, et al. Artificial neural network for short-term load forecasting in distribution systems[J]. Energies, 2014, 7(3): 1576-1598.

[6] LDIN W M, TU C S, YANG R F, et al. Particle swarm optimisation aided least-square support vector machine for load forecast with spikes[J]. IET Generation, Transmission & Distribution, 2016, 10(5): 1145-1153.

[7] 吴倩红, 高军, 侯广松, 等. 实现影响因素多源异构融合的短期负荷预测支持向量机算法[J]. 电力系统自动化 2016, 40(15): 67-72.
WU Q H, GAO J, HOU G S, et al. Short-term load forecasting support vector machine algorithm based on multi-source heterogeneous fusion of load factors[J]. Automation of Electric Power System, 2016, 40(15): 67-72 (in Chinese)

[8] 李婉华, 陈宏, 郭昆, 等. 基于随机森林算法的用电负荷预测研究[J]. 计算机工程与应用, 2016, 52(23): 236-243.
LI W H, CHEN H, GUO K, et al. Research on electrical load prediction based on random forest algorithm[J]. Computer Engineering and Applications, 2016, 52(23): 236-243(in Chinese).

[9] 彭鹏, 彭佳红. 基于多元线性回归模型的电力负荷预测研究[J]. 中国安全生产科学技术, 2011, 7(9): 158-161.
PENG P, PENG J H. Research on the prediction of power load based on multiple linear regression model[J]. Journal of Safety Science and Technology, 2011, 7(9): 158-161 (in Chinese).

[10] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. Imagenet classification with deep convolutional neural networks[J]. Communications of the ACM, 2017, 60(6): 84-90.

[11] RUMELHART D E, HINTON G E, WILLIAMS R J. Learning representations by back-propagating errors[J]. Nature, 1986, 323(6088), 533-536.

[12] HOCHREITER S, SCHMIDHUBER J. Long short-term memory [J]. Neural Computation, 1997, 9(8): 1735-1780.

[13] CHUNG J, GULCEHRE C, CHO K, et al. Gated feedback recurrent neural networks[C]//International Conference on Machine Learning. Lile: PMLR, 2015: 2067-2075.

[14] 彭文, 王金睿, 尹山青. 电力市场中基于 Attention-LSTM 的短期负荷预测模型[J]. 电网技术, 2019, 43(5): 1745-1751.
PENG W, WANG J R, YIN S Q. Short-term load forecasting model based on attention-LSTM in electricity market[J]. Power System Technology, 2019, 43(5): 1745-1751 (in Chinese).

[15] 李龙, 魏靖, 黎灿兵, 等. 基于人工神经网络的负荷模型预测[J]. 电工技术学报, 2015, 30(8): 225-230.
LI L, WEI J, LI C B, et al. Prediction of load model based on artificial neural network[J]. Transactions of China Electrotechnical Society, 2015, 30(8): 225-230 (in Chinese).

[16] 陆继翔, 张琪培, 杨志宏, 等. 基于 CNN-LSTM 混合神经网络模型的短期负荷预测方法[J]. 电力系统自动化, 2019, 43(8): 131-137.
LU J X, ZHANG Q P, YANG Z H, et al. Short-term load forecasting method based on CNN-LSTM hybrid neural network model[J]. Automation of Electric Power Systems, 2019, 43(8): 131-137 (in Chinese).

[17] 李泽文, 胡让, 刘湘, 等. 基于 PCA-DBILSTM 的多因素短期负荷预测模型[J]. 电力系统及其自动化学报, 2020, 32(12): 32-39.
LI Z W, HU R, LIU X, et al. Multi-factor short-term load prediction model based on PCA-DBILSTM[J]. Proceedings of the CSU-EPSA, 2020, 32(12): 32-39 (in Chinese).

[18] 杨洋, 谷震浩. 基于 XGBOOST-DNN 的中期电力负荷预测[J]. 计算机系统应用, 2021, 30(9): 186-191.
YANG Y, GU Z H. Mid-term power load forecasting based on XGBoost-DNN[J]. Computer Systems & Applications, 2021, 30(9): 186-191 (in Chinese).

[19] HU J, SHEN L, ALBANIE S, et al. Squeeze-and-excitation networks[EB/OL]. (2019-05-16) [2022-05-18]. https://arxiv.org/abs/1709.01507.pdf.

[20] LAI G K, CHANG W C, YANG Y M, et al. Modeling long- and short-term temporal patterns with deep neural networks[EB/OL].

(2018-04-18) [2022-05-18] .https://arxiv.org/abs/1703.07015.pdf.

[21] 庞昊, 高金峰, 杜耀恒. 基于多神经网络融合的短期负荷预测方法[J]. 电力自动化设备, 2020, 40(6): 37-43.
PANG H, GAO J F, DU Y H. Short-term load forecasting method based on fusion of multiple neural networks[J]. Electric Power Automation Equipment, 2020, 40(6): 37-43 (in Chinese).

[22] 王增平, 赵兵, 纪维佳, 等. 基于 GRU-NN 模型的短期负荷预测方法[J]. 电力系统自动化, 2019, 43(5): 53-58.
WANG Z P, ZHAO B, JI W J, et al. Short-term load forecasting method based on GRU-NN model[J]. Automation of Electric Power Systems, 2019, 43(5): 53-58 (in Chinese).

[23] 刘皓琪, 高飞, 王耀力, 等. 基于 Dropout-ILSTM 网络的短期电力负荷预测[J]. 电测与仪表, 2021, 58(5): 105-111.
LIU H Q, GAO F, WANG Y L, et al. Short-term power load forecasting based on Dropout-ILSTM network[J]. Electrical Measurement & Instrumentation, 2021, 58(5): 105-111 (in Chinese).

[24] 龚飘怡, 罗云峰, 方哲梅, 等. 基于 Attention-BiLSTM-LSTM 神经网络的短期电力负荷预测方法[J]. 计算机应用, 2021, 41(S1): 81-86.
GONG P Y, LUO Y F, FANG Z M, et al. Short-term power load forecasting method based on Attention-BiLSTM-LSTM neural network[J]. Journal of Computer Applications, 2021, 41(S1): 81-86 (in Chinese).

[25] 朱凌建, 荀子涵, 王裕鑫, 等. 基于 CNN-Bi LSTM 的短期电力负荷预测[J]. 电网技术, 2021, 45(11): 4532-4539.
ZHU L J, XUN Z H, WANG Y X, et al. Short-term power load forecasting based on CNN-BiLSTM[J]. Power System Technology, 2021, 45(11): 4532-4539 (in Chinese).

Periodic pattern-enhanced multi-view short-term load prediction

SU Wei*, XIAO Xiaolong, SHI Mingming, FANG Xin, SI Xinyao

(State Grid Jiangsu Electric Power Co., Ltd. Research Institute, Nanjing 210000, China)

Abstract: Short-term load prediction is essential to ensure the proper operation of the power system. The existing efforts have two limitations: lack of mining the dependencies between features and ignore the periodic pattern of load changes. To solve the above limitations, we propose periodic Pattern-enhanced MultI-view Short-term power load prediction nEtworks, dubbed EPISODE. The framework includes two core components: a multi-view feature learning component and a periodic pattern-enhanced load prediction component. The former aims to effectively extract static features and time series features to obtain enhanced feature representation; the latter is to perform general time series mining and periodic time series mining to obtain comprehensive historical feature representation. The combination of the two aforementioned qualities results in the realization of the short-term load forecast. Extensive experiments have been conducted on real-world datasets, and the experimental results demonstrate the superiority of our proposed method.

Keywords: periodic pattern-enhanced; short-term load prediction; feature squeeze and excitation; gated recurrent unit; attention mechanism

Received: 2022-05-21; **Accepted**: 2022-07-02; **Published Online**: 2022-12-14 15:58

URL: link.cnki.net/urlid/11.2625.V.20221213.1218.004

Foundation item: The Science and Technology Project of State Grid Jiangsu Electric Power Co., Ltd (J2021165)

* **Corresponding author**. E-mail: susw2011@163.com

http://bhxb.buaa.edu.cn jbuaa@buaa.edu.cn
DOI: 10.13700/j.bh.1001-5965.2022.0290

基于深度强化学习的办公流程任务分配优化

廖晨阳[1]，于劲松[1,*]，乐祥立[2]

(1. 北京航空航天大学 自动化科学与电气工程学院，北京 100191；2. 北京航空航天大学 高等理工学院，北京 100191)

摘 要：在办公平台中存在异构流程任务大量并行的情况，不仅需要任务执行者具有较强的能力,也对协同调度系统的性能提出了要求。采用强化学习(RL)算法，结合协作配合度、松弛度等定量分析，并基于马尔可夫博弈理论提出多智能体博弈模型，实现以总体流程配合度和最大完工时间为优化目标的优化调度系统，提高了总体执行效率。以真实的业务系统流程作为实验场景，在相同的优化目标下，对比D3QN等3种深度强化学习(DRL)算法和基于蚁群的元启发式算法，验证了所提方法的有效性。

关 键 词：工作流；任务调度；马尔可夫博弈；深度强化学习；协作度

中图分类号：TP301.6

文献标志码：A **文章编号**：1001-5965(2024)02-0487-12

当前，主流的办公平台多采用预置的静态日程安排和事务提醒系统对平台内多员工并行执行多条业务流程的情况实现调度[1]，管理者无法针对实际办公情况对调度方案和策略进行调整。如何对办公流程的状态进行分析，实现任务分配并动态生成更为合理的日程调度安排成为行业的迫切需要。

由于该问题属于典型的多项式复杂程度的NP难题，只能寻求对单一目标求次优解或对多目标求Pareto解集。

针对类似的调度优化问题，元启发式算法取得了良好效果。在针对云主机的有向无环图(directed acyclic graph, DAG)调度问题中，Gao等[2]考虑了云主机运行费用，并以任务完成时间为优化目标提出了混合遗传-人工蜂群算法(hybrid genetic algorithm & artificial bee colony algorithm, HGAABC)。Farahnakian等[3]提出了基于蚁群系统的虚拟机整合(ant colony system based virtual machine consolidation, ACS-VMC)模型。在办公领域，吕龙等[4]将蚁群算法与协作相容性和负载相结合，提出了基于蚁群算法的任务分配(maximizing compatibility based on ant colony optimization, ATO-MC)模型。Kumar等[5]基于贪婪算法，考虑了协作度，并将执行者工作负载作为优化目标，设计了基于贪婪算法的工作流任务分配(optimal work assignment, OWA)模型。许荣斌等[6]考虑了协作度，计算最大依赖度及最小冗余度设计了基于最大依赖度及最小冗余度的员工协作优化(maximal and minimal dependency redundancy, MDR)算法。在生产车间的任务调度领域，Piroozfard等[7]基于遗传算法，以低订单延误率和低碳排放为优化目标对车间作业调度问题(job shop scheduling problem, JSSP)设计了一种改进的多目标遗传算法(multi-objective genetic algorithm, MOGA)模型。

启发式算法可解释性强，可解决多目标优化问题，但吕龙等[4]的实验证明了先进的启发式算法的时间复杂度随问题规模增加呈快速增长趋势。在实际的企业级云平台中，问题规模往往较大，启发式算法很难在时间效率和优化效果上取得较好的平衡。

收稿日期：2022-04-28；**录用日期**：2022-05-06；**网络出版时间**：2022-05-27 15:48
网络出版地址：link.cnki.net/urlid/11.2625.V.20220526.2003.002
基金项目：国家重点研发计划(2018YFB1004100)
*通信作者. E-mail：yujs@buaa.edu.cn

引用格式：廖晨阳，于劲松，乐祥立. 基于深度强化学习的办公流程任务分配优化[J]. 北京航空航天大学学报，2024，50(2)：487-498.
LIAO C Y，YU J S，LE X L. Optimization of office process task allocation based on deep reinforcement learning[J]. Journal of Beijing University of Aeronautics and Astronautics，2024，50(2)：487-498 (in Chinese).

近年来，强化学习(reinforcement learning, RL)成为一种处理马尔可夫决策过程(Markov decision process, MDP)问题的新方法。由于RL能够在每个决策点学习最佳行动，并对动态变化做出快速反应，许多基于RL的方法已被应用于动态调度问题中。Cui等[8]提出了针对一般DAG调度问题的RL算法；Wu等[9]提出的云工作流任务调度的Q值学习算法(Q-learning algorithm for cloud workflow task scheduling, QL-CWTS)采用传统Q值(指当前状态动作价值)表，对单智能体调整reward函数的内部参数实现Pareto最优，与ATO-MC模型[4]调整权重融合不同优化目标的思路类似，随着优化目标的增加，人为调整均衡参数将更加困难；Wang[10]提出的基于集群和动态搜索的加权Q值学习算法(weighted Q-learning algorithm based on clustering and dynamic search, WQ-CDS)，针对普通Q值表维数爆炸这一问题，采用状态空间S聚类方法减少了盲目探索和计算复杂度，并设计了离散的奖励函数，通过最小化流程延误和流程松弛度减少总体延误率。深度强化学习(deep reinforcemewnt learning, DRL)可以更为有效地解决Q值表维数爆炸问题。Cao等[11]针对车间生产任务调度提出了相应的DRL算法，不同于通过均衡负载实现优化，Han和Yang[12]提出的DDDQNPR(属于D3QN算法)，通过提高机器的平均利用率间接优化完工时间。

综上，元启发式算法随着问题规模增大，面临挑战。目前，没有基于协作度和最大完工时间且采用RL算法实现的优化方法，其他相近领域的RL算法均设计了逻辑复杂的奖励函数，造成了模型的不稳定和收敛速度慢的问题。

本文采用RL领域的深度Q值网络(deep Q-network, DQN)算法原理，基于办公领域的协作配合度计算及工作流状态空间描述方法，实现了一种改进的多智能体双深度Q值网络(multi-agent double deep Q-network, MDDQN)，并构建了办公流程调度系统。

1) 相较元启发式算法，计算重调度时间与训练时间之和初始解和最终优化解的时间明显缩短。

2) 采用多智能体作为“棋手”进行马尔可夫博弈过程，自动化实现多目标均衡，避免繁复的手动调参，明显扩展了实用性。

3) 在最大完工时间和总体流程配合度2个优化指标上取得最优。

1 问题描述

办公系统的工作流实例由多个互为输入输出依赖的节点构成，如图1所示，实例1采用模板1，实例2、实例3采用模板2，实例4采用模板3。图1中圆圈表示节点，圆圈的颜色表示3种办理状态，黑色表示已分配办理人正在办理，圆圈中标记字母代表办理人，黑色圆圈越大表示当前在办节点的办理效率越高，预期用时越短。不同于车间调度、云主机调度问题，工作流的图结构存在环，本文通过等效建模将其转化为动态的DAG调度问题。

简化后的工作流任务节点(或称节点)的执行条件满足前提条件的约束。本文的特殊约束和简化如下：

1) 在工作流图中，若存在打回现象构成了环，则采用图2所示的等效建模，保证每个重调度时刻，输入调度系统的流程结构无环。

2) 若节点不设置前提条件，默认在执行该节点前，需完成其所有前继节点。

3) 存在多人都有权限办理某一项任务节点的情况，但每个节点同时只能有一个人执行。

4) 一个人某时刻只能做一项任务，直到这一任务节点完成。

5) 只要完成任务前提条件中的一项以上任务，就可以执行任务。

6) 一条流程可有多个结束节点，但起始节点只能有一个。

7) 不考虑前提条件，则流程结构退化为DAG图。

8) 现阶段的决策只依赖当前状态。

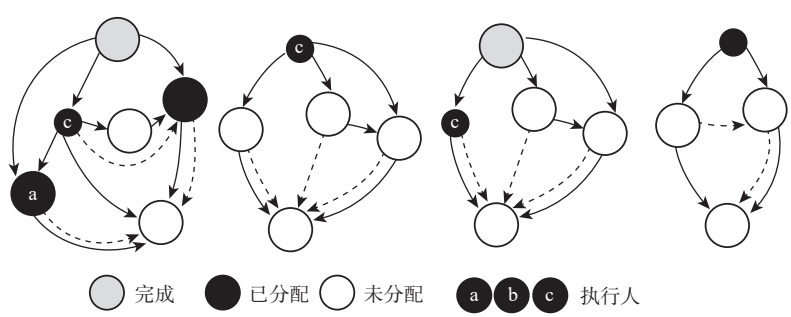


图1 工作流实例

Fig. 1 Workflow instance

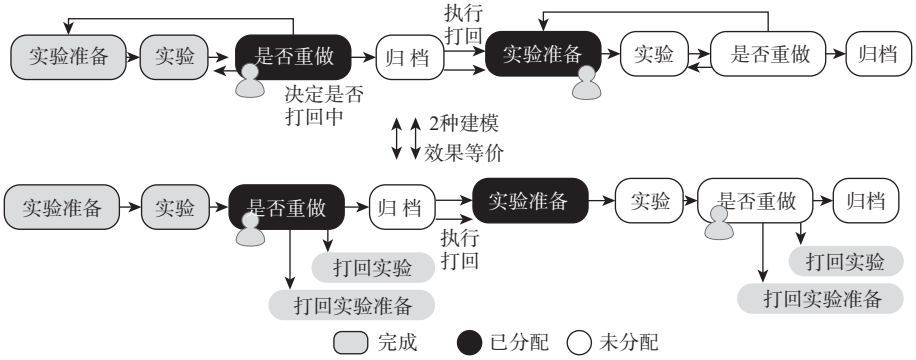


图 2　打回情况

Fig. 2　Call back situation

如图 3 所示, 要执行节点 2, 必先完成节点 1 和 6; 节点 3 或节点 4 完成, 就可以执行节点 5。上述建模符合 BPMN2.0 标准。

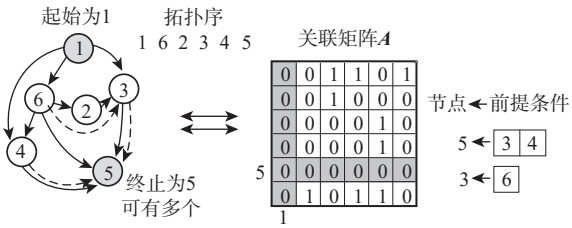


图 3　实例的数学表达

Fig. 3　Mathematical expression of instance

2　办公系统工作流调度问题建模

在重调度时刻 t, 将当前时刻设为参考时间轴原点。依靠有限的 m 名员工 $P=\{P_1,P_2,\cdots,P_m\}$, 对当前积压及未来陆续到达的 n 条流程 $J=\{J_1,J_2,\cdots,J_n\}$求解调度优化问题。每条流程由n_i个流程节点构成, $O_{i,j}$为流程J_i中的第 i 个节点。

以重调度时刻为坐标, 流程J_i的到达时间为A_i, $O_{i,0}$为流程J_i的入口节点, O_{i,e_i}为结束节点。设结束节点集合E_i, 则$\mathrm{e}_i\in E_i$。

设有权办理流程节点$O_{i,j}$的人员集合为所有员工的子集$P_{i,j}(P_{i,j}\in P)$。读取历史日志, 统计节点$O_{i,j}$由办理人P_k办理的平均时间和方差, 最终按正态分布构造统计量$t_{i,j,k}$。令$C_{i,j}$为任务节点$O_{i,j}$的完成时间, 则任务J_i的完成时间为其结束节点的完成时间C_{i,e_i}。A_i为J_i到达时间, 可设置流程在未来指定时间到达。D_i为人为设置的J_i截止时间, 若未人为设置, 则按历史日志计算平均值$\overline{C}_{i,\mathrm{e}_i}$乘以可调系数$K$得到截止办理时间。

调度系统将围绕最小化最终完工时间、最大化整体协作配合度 2 个优化目标实现序贯决策。

总体流程如图 4 所示, 每 1 类流程的各个节点对应了不同的执行人集合, 可分配给有权限办理的执行者办理。系统依据已有的执行日志计算各类流程的配合度矩阵, 作为当前环境的一部分, 配合度矩阵的参数非实时更新, 采用批量累进更新的方法。

对 RL 算法做如下定义:

1) 智能体 1 的局部优化目标为最小化完工时间, 智能体 2 的局部优化目标为最大化协作配合度, 两者通过 DDQN 的原理训练, 最终在功利规则的引导下自动实现全局均衡。

2) 2 个智能体均采用 DDQN 构建, 每个 DDQN 都由 Q 值网络 θ_{QN} 和目标 Q 值网络 θ_{TQN} 组成。

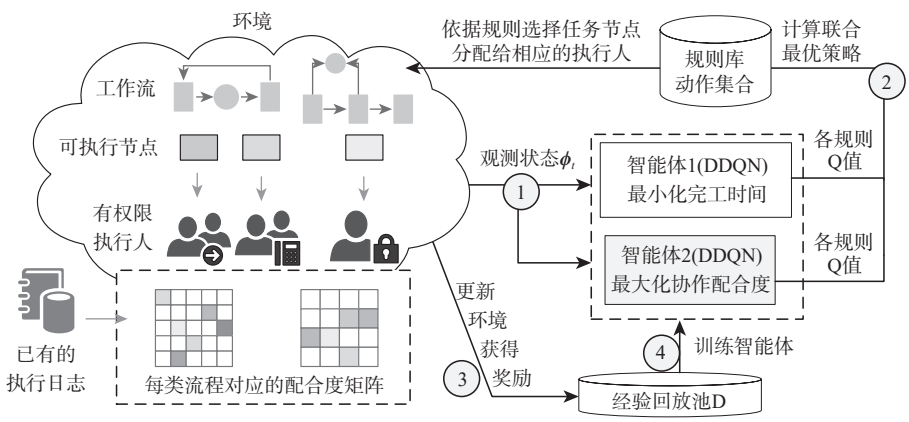


图 4　总体流程

Fig. 4　Overall process

3）动作集合$\{a_1,a_2,\cdots,a_6\}$对应结合了协作度的6种调度规则。

4）观测状态$\boldsymbol{\phi}_t$。时刻t输入智能体的状态为向量。由8个具有实际含义的归一化连续参数值组成，每一个参数表征当前环境的某方面特征。

在某重调度时刻t，2个智能体输出各规则的Q值。系统依据功利主义均衡规则计算联合最优策略，根据联合最优策略的概率分布，以ε-greedy从动作集合中推荐出当前时刻的调度规则。依据所选规则，从环境中的可执行节点进行选取，并分配到具体的执行人。其中，智能体的训练过程由经验回放实现。回放池D记录历史执行过程获得的奖励，以及执行的规则和规则执行前后环境状态的变化，每隔一定步长对智能体1和智能体2按DDQN的学习原理进行训练。

3 基于DDQN与马尔可夫博弈的任务分配策略

3.1 DDQN算法

3.1.1 强化学习与Q值学习理论

RL算法模拟了自然界生物智能体学习的过程，智能体基于环境状态S依据策略做出动作a，并获得相应奖励r。经典的RL基于MDP原理，经过多轮次训练，使智能体逐步学习得到如下知识：

1）对每一种状态s，执行动作a后获得奖励的期望值称为状态动作价值函数$Q(s,a)$。令γ为价值随时间的衰减率，则有

$$Q(s,a)=E_\pi(G_t|s_t=s,a_t=a)=E(r_{t+1}+\gamma r_{t+2}+\gamma^2 r_{t+3}+\cdots|s_t=s,a_t=a) \tag{1}$$

式中：G_t表示在t时刻若处于s状态并采用动作a时未来直到统程结束式+∞时刻所能获得的总的奖励的期望值。

2）每种状态下的最优策略π^*。式(2)描述了训练适应过程中常采用的ε-greedy贪婪策略。随着智能体不断适应环境，随机探索倾向减少，策略倾向于选择奖励最大的动作。

$$\pi(a|s)=\begin{cases}\varepsilon/l+1-\varepsilon & a^*=\arg\max\limits_{a\in(a_1,\cdots,a_l)}Q(s,a)\\ \varepsilon/l & \text{其他}\end{cases} \tag{2}$$

本文考虑多智能体的马尔可夫博弈过程，采用功利算法实现均衡，以ε概率选择使2个智能体获得单步奖励之和最大的动作a^*：

$$a^*=\arg\max_{a\in(a_1,\cdots,a_l)}(Q_1(s,a)+Q_2(s,a)) \tag{3}$$

式中：Q_1、Q_2函数输出值表示不同智能体的状态动作价值输出。在本文的调度问题中，假设如下场景：当任务流程中节点数为15、业务员数为8时，调度方案组合数可达35万亿个，现有计算机技术很难维护如此庞大的状态动作价值对，即$Q(s,a)$二维表。

3.1.2 深度Q值网络

为解决上述问题，Mnih等[13]首先提出了DQN算法。DQN将深度学习网络作为Q值函数的训练器，将原始的状态特征作为深度学习网络的输入，每个状态动作对作为深度学习网络的输出，避免了传统Q值网络在连续状态空间问题中难以维护庞大Q值表的问题。为消除训练过程中连续状态转换之间的相关性对网络参数θ迭代更新的影响，建立了容量为N的回放池D，其中存储了智能体Agent在重调度时刻t的各项参数(s_t,a_t,r_t,s_{t+1})。在DQN中使用的目标Q值y的计算方式如下：

$$y_i=\begin{cases}r_j & s_j\text{为终止状态}\\ r_j+\gamma\max\limits_{a'}Q(\boldsymbol{\phi}(s_{t+1}),a'_j,\theta) & s_j\text{不是终止状态}\end{cases} \tag{4}$$

目标Q值的计算需要用到当前的Q值网络A_{QN}的网络内部参数θ，同时又需要y_i更新网络参数以便进行下一次Q值计算，在训练机理上，两步之间的相关性太大，对reward奖励等指标的快速稳定收敛不利。

为解决上述问题，在NIPS 2015国际会议中研究者首次提出了DDQN网络结构[14]，在该算法中，一个Q值网络θ_{QN}用于选择动作并更新网络参数，另一个目标Q值网络θ_{TQN}(称为目标网络)，通过计算目标Q值y训练Q值网络θ_{QN}，y的计算方式如下：

$$y_i=\begin{cases}r_j & s_j\text{为终止状态}\\ r_j+\gamma\max\limits_{a'}\theta_{\mathrm{TQN}}(\boldsymbol{\phi}(s_{t+1}),a'_j,\theta') & \\ & s_j\text{不是终止状态}\end{cases} \tag{5}$$

随后，学者对该算法进行完善，以一定步长将参数迁移到目标Q值网络θ_{QN}，形成了DDQN算法[15-16]。

3.2 DDQN算法与规则库动作集合

本文依据DDQN算法原理，结合带有协作配合度分析的规则库动作集合，解决办公流程的调度问题，如图5所示。流程实例数据、实时人力资源利用率等环境状态参数由对应环境计算，环境如图5云朵所示。

将重调度时刻t_1从环境中观测到的当前状态向量$\boldsymbol{\phi}_{t_1}$输入深度神经网络，网络θ_{QN}的输出是神经网络依据当前状态参数计算的6个规则的Q值期望，从输出中计算使得Q值最大的规则，依据ε-

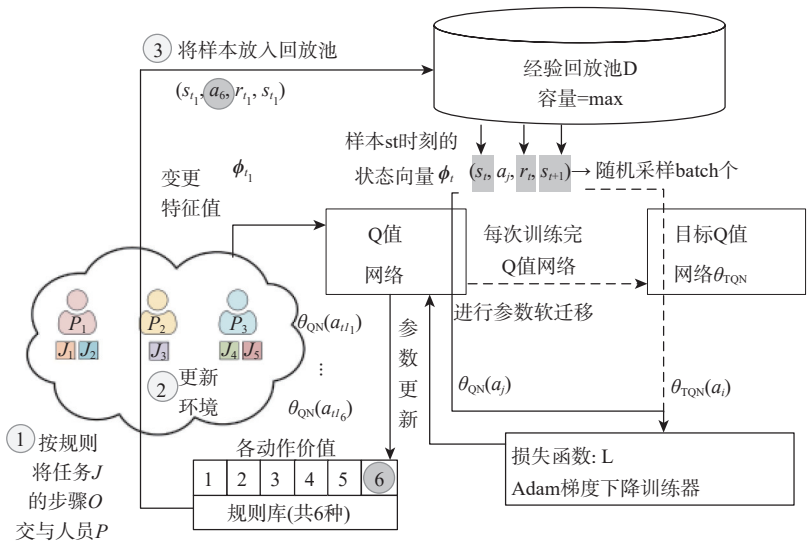


图 5　DDQN 与规则库的结合

Fig. 5　Combination of DDQN and rule base

greedy 贪婪算法选择规则, 图 5 示例中选择了规则 6。依据规则库中规则 6 算法计算单步调度方案, 将任务 J_i 的步骤 $O_{i,j}$ 交给人员 $P_{i,j,k}$ 办理, 根据单步调度方案, 更新环境并将此次调度的信息以四元组 $(s_{t_1}, a_6, r_{t_1}, s_{t_1+1})$ 的形式存入回放池 D 中。回放池 D 以队列的形式维护数据。每次训练过程, 单智能体从回放池 D 中随机选取 batch 个样本, 通过 θ_{QN} 和 θ_{TQN} 计算损失函数, 并通过梯度下降算法实现 Q 值网络参数的优化迭代。每次训练完成后, 以软迁移的方式将 Q 值网络 θ_{QN} 的参数迁移到目标 Q 值网络 θ_{TQN} 中。

在实际工程应用中, 平台内用户每操作一定步数启动训练流程, 模型会随着时间的推移, 不断适应实际办公场景的环境特点。

算法 1　DDQN 算法。

Begin

1: 将回放池 D 初始化为容量 max。

2: 用随机权重 θ 初始化 Q 值网络 θ_{QN}。

3: 用随机权重 $\theta^- = \theta$ 初始化目标 Q 值网络 θ_{TQN}。

4: for episode = 1 : L do

5:　观察初始环境状态 s_{t_1}, 计算状态的 8 大参数构成特征向量 $\boldsymbol{\phi}_{t_1}$。

6:　for $t = 1 : T\left(\sum_{i=1}^{n} n_i\right)$ do

7:　try 以概率 ε 选一个随机调度规则 a_{t_1}

8:　except $a_{t_1} = \arg\max_a \theta_{QN}(a, \boldsymbol{\phi}_{t_1}; \theta)$

9:　按计算得到的调度规则 a_{t_1} 计算调度方案并执行, 获得奖励 r_{t_1}, 并计算下一状态 s_{t_1+1}。

10:　计算状态 s_{t_1+1} 的 8 大参数, 构成特征向量 $\boldsymbol{\phi}_{t_1+1}$。

11:　在回放池 D 中存储此步日志 $(s_{t_1}, a_{t_1}, r_{t_1}, s_{t_1+1})$。

12:　在回放池 D 中随机选取 128 个训练样本。

13:　按式(5)计算目标 Q 值。

14:　计算损失函数, 对在线网络 Q 的参数 θ 进行梯度下降训练。

15:　将 θ_{QN} 的内部参数以软迁移方式迁移到 θ_{TQN} 的对应参数。

16:　end for

17: end for

End

由于存在前提条件, 并不一定需要完成全部的流程节点就可以完成该流程。将完成一个环境的全部流程作为一轮次的结束条件(L 轮), 记录总奖励、总体协作配合度, 从而分析算法有效性。

3.3　损失函数与深度 Q 值网络参数训练

单智能体的训练基于 Q 值网络 θ_{QN} 和目标 Q 值网络 θ_{TQN} 的当前参数 θ 与 θ^- 实现。

随机采样得到不连续的 128 个样本作为一个批次训练的输入。以 Adam 训练器对 Q 值网络 θ_{QN} 的参数实现梯度下降训练。对任意一个样本 (s_t, a_j, r_t, s_{t+1}), 损失函数如下:

$$L = E[(r_t + \gamma\theta_{TQN}(a_i) - \theta_{QN}(a_j))^2] \tag{6}$$

计算原理如图 6 所示。

1) 对状态 s_t 分析计算, 得到状态向量 $\boldsymbol{\phi}_t$ 输入 Q 值网络 θ_{QN} 的输出为 $(\theta_{QN}(a_1), \cdots, \theta_{QN}(a_6))$, 从中选出 a_j 对应 $\theta_{QN}(a_j)$。

2) 分析 (s_t, a_j, r_t, s_{t+1}) 中的 s_{t+1}, 计算状态向量 $\boldsymbol{\phi}_{t+1}$, 输入 Q 值网络 θ_{QN} 和目标 Q 值网络 θ_{TQN}。θ_{QN} 输出 $\theta_{QN}(a_{t+l_1}), \cdots, \theta_{QN}(a_{t+l_6})$ 6 个动作价值, 选取 $a_i =$

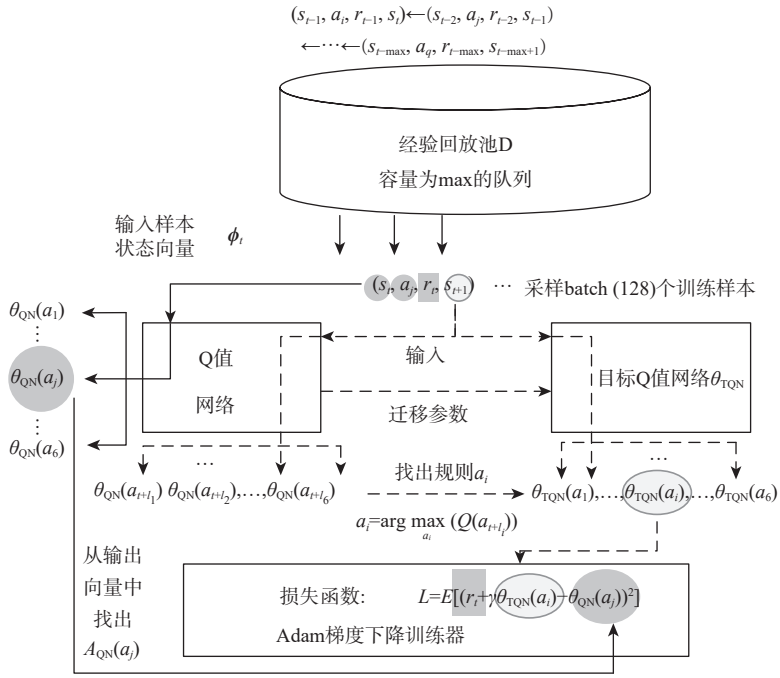


图 6 损失函数计算

Fig. 6 Loss function calculation

$\arg\max_{a_i}(\theta_{QN}(a_{t+l_i}))$, 从而在 θ_{TQN} 网络输出的$\theta_{TQN}(a_1)$, $\cdots$, $\theta_{TQN}(a_6)$6 个动作价值中选取角标 i 对应的$\theta_{TQN}(a_i)$, 进而按式(6)计算损失函数。

3.4 配合度计算

图 4～图 6 中的规则库共 6 种, 本文将用户间协作配合度作为重要依据, 结合传统调度算法的启发式规则设计 6 种规则。协作配合度计算依据 OA 业务数据库的统计进行。对任意流程模板, 近一段时间办结的实例$\{A_1, A_2, \cdots\}$, 令$t_{\min}$和$t_{\max}$分别为集合中办理过程最短和最长的时间, 业务人员P_i与P_j参与的流程平均办理时间为t_{ave}, 则P_i与P_j的协作配合度$c_{i,j}=1-\dfrac{t_{ij}-t_{\min}}{t_{\max}-t_{\min}}\omega$, ω 为可调节的超参数表征。本文中ω设置为 0.8, 当P_i和P_j 2 人参与的流程平均办理时间为 $1/2[t_{\min}+t_{\max}]$时, 协作配合度为 0.6。

对于该流程模板单个流程实例J_{n_1}的总体协作配合度 S, 计算如下:

$$S_{n_1}=\sum_{k=1}^{m}\sum_{v=1}^{m}\sum_{i=1}^{n}\sum_{j=1,j\neq i}^{n}A_{i,k}A_{j,v}\mathrm{DAG}(i,j)c_{k,v} \tag{7}$$

式中: $A_{i,k}$从 0,1 中取值, 表示J_{n_1}的节点$O_{n_1,i}$是否被分配给P_k; DAG(i,j) 表示$O_{n_1,i}$、$O_{n_1,j}$是否连通。

S 计算原理如图 7 所示。

在 S 计算基础上, 本文提出流程整体协作率, 将 S 值除以当前已分配节点连通边总数, 得到流程的整体协作率$\overline{S(t)}$。

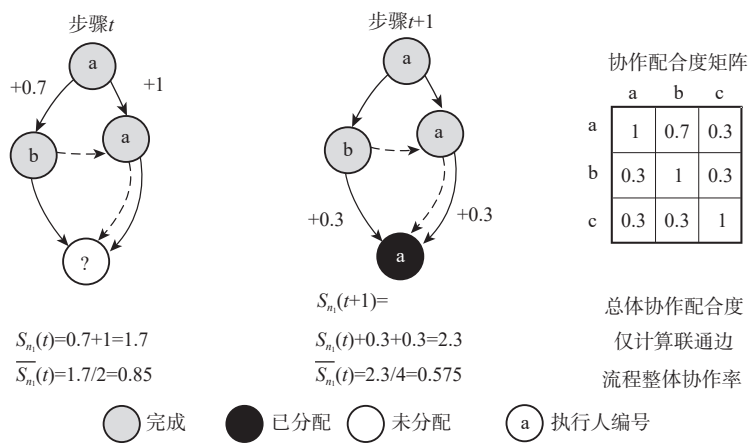


图 7 协作配合度与总体协作配合度

Fig. 7 Cooperation degree and overall cooperation degree

3.5　考虑配合度的规则库动作集合

6 种规则函数采用算法 2 的模板, 其中, 第 3 步与第 5 步涉及函数 $A()$、$A'()$、$B()$。根据不同的 $A()$、$A'()$、$B()$, 本文设计产生了 6 种规则。

算法 2　调度规则算法。

输入: 当前流程状态 s_t。

输出: 单步调度规则 $O_{i,j}$、$P_{i,j,k}$。

Begin

1: 根据 s_t, 求预计延误流程集 $J_{\text{delay}}(t)$和未完成流程集 $J_{\text{uc}}(t)$。

2: 计算各条流程可执行节点集 O。

3: 根据以下函数, 从 O 中选择流程 J_i的节点

$$O_{i,j}:\begin{cases}\text{如果}J_{\text{delay}}\text{ 预计延误集合为空，执行}A();\\ \text{如果已有实际的延误发生，执行}A'().\end{cases}$$

4: 根据 OA 权限系统得到办理人员集合 $P_{i,j}$。

5:在得到的集合 $P_{i,j}$中, 通过执行 $B_1()$ 或 $B_2()$ 函数选择办理人 $P_{i,j,k}$。

6: 返回 $O_{i,j}$、$P_{i,j,k}$。

End

设计 5 种调度函数$A_1()$、$A_2()$、$A_3()$、$B_1()$、$B_2()$。

1) 优先取剩余时间除以未完成节点数最小的流程$A_1()$。

$$J_i \leftarrow \arg\min i \in J_{\text{uc}}(t)\frac{D_i-\overline{T}_{\text{cur}}(t)}{n_i-F_i(t)}$$

式中: $\overline{T}_{\text{cur}}(t)$为各流程当前时刻已被安排执行的节点最晚完成时间的均值; $F_i(t)$为流程 J_i在 t 时刻完成的节点数量; $J_{\text{uc}}(t)$为未完成流程集。

2) 优先选择延误时间最短的流程$A_2()$。

$$J_i \leftarrow \arg\max i \in J_{\text{delay}}(t)\left(\overline{T}_{\text{cur}}(t)+\sum_{j\in n_i}\bar{t}_{i,j}-D_i\right)$$

式中: $J_{\text{delay}}(t)$为截止时间早于$T_{\text{cur}}(t)$的流程的集合; $\bar{t}_{i,j}$为办理流程节点$O_{i,j}$的平均办理时间。

3) 优先选剩余时间与完工率之积最小的流程$A_3()$。

$$J_i \leftarrow \arg\min i \in J_{\text{uc}}(t)\frac{F_i(t)}{n_i}(D_i-\overline{T}_{\text{cur}}(t))$$

4) 优先选择最早空闲的业务人员$B_1()$。

$$P_k \leftarrow \arg\min k \in P_{i,j}(\max(R_k(t),N_i(t),A_i))$$

式中: $R_k(t)$为 t 时刻, 业务人员 P_k最早可以开展$O_{i,j}$工作的预计时间。默认是 P_k当前排班列表的最晚完成时间, 如果当前任务 $O_{i,j}$可以插入员工 P_k排班列表的间隙中, 则$R_k(t)$另行计算; $N_i(t)$为 J_i正在执行的节点完工时间。

5)选择使总体协作相融度增加最大的人员$B_2()$。

$$P_k \leftarrow \arg\max k \in (S_i(t+1)-S_i(t))$$

6 种规则由 $A()$、$A'()$、$B()$构成, $A()$、$A'()$、$B()$分别从$A_1()$、$A_2()$、$A_3()$、$B_1()$、$B_2()$中选择。规则和调度函数对应关系如表 1 所示。

表 1　规则与调度函数对应关系

Table 1　Correspondence between rules and scheduling functions

规则编号	采用的 A(选取$O_{i,j}$)函数	采用的 A'(选取$O_{i,j}$)函数	采用的 B(选取$P_{i,j,k}$)函数
1	$A_1()$	$A_2()$	$B_1()$
2	$A_1()$	$A_1()$	$B_2()$
3	从$J_{\text{uc}}(t)$中随机选取流程J_i, $Q_{i,j}$的j由$B_1()$计算	从$J_{\text{uc}}(t)$中随机选取流程J_i, $Q_{i,j}$的j由$B_1()$计算	$B_1()$
4	$A_3()$	$A_2()$	$B_1()$
5	$A_2()$	$A_2()$	$B_1()$
6	$J_i,P_k \leftarrow \arg\max i \in J_{\text{uc}}(t)(\Delta S_i)$	$J_i,P_k \leftarrow \arg\max i \in J_{\text{uc}}(t)(\Delta S_i)$	$B_2()$

表 1 中, 规则 6 仅从各流程的总体协作配合度增加角度出发选择流程和办理人。

3.6　状态向量和奖励函数

奖励函数的设计对于网络参数和最终奖励随训练轮次的收敛性有着决定性影响。本文通过对比调度步骤$(O_{i,j},P_{i,j,k})$执行前后环境状态$\boldsymbol{\phi}_t$的改变设置奖励r_t。

环境状态向量$\boldsymbol{\phi}_t$由如下 8 个参数构成:

1) 人力资源利用率$U_{\text{ave}}(t)$, 指业务人员平均工作时间/流程总执行时间。

2) $U_{\text{std}}(t)$, 指所有业务人员时间利用率标准差。

3) $S_{\text{ave}}(t)$, 指以节点数为单位计算所有流程办理进度均值。

4) $S_{\text{std}}(t)$, 指各流程办理进度的标准差。

5) 名义延误率$T_{\text{e}}(t)$, 指$J_{\text{delay}}(t)$集元素个数/流程数 n。

6) $T_{\text{a}}(t)$, 指实际的延误率。

7) $C_{\text{ave}}(t)$, 指以流程为单位计算完工率。

8) $\overline{S}_{\text{ave}}(t)$, 指各流程整体协作率均值。

本文针对 2 个优化目标独立设计了如下 2 个奖励函数, 分别对应 2 个智能体:

$$r_{\text{make-span}}(t)=\left(\frac{\Delta T_{\text{cur}}(t)}{\max(\Delta T_{\text{cur}}(t))}\right)^3 \tag{8}$$

$$r_{\text{coop}}(t)=0.5+0.5\sqrt[3]{\Delta\overline{S}(t)} \tag{9}$$

式中: $r_{\text{make-span}}(t)$、$r_{\text{coop}}(t)$的最终计算取值均在 [0,1] 之间; $\Delta T_{\text{cur}}(t)$为采用调度步骤$(O_{i,j},P_{i,j,k})$后$T_{\text{cur}}(t)$的

增加值；$\Delta\overline{S}(t)$为采用调度步骤后$\overline{S}(t)$的增加值。

3.7 基于马尔可夫博弈的多目标优化

针对式(8)和式(9)这2个奖励函数，本文依据马尔可夫博弈理论构建了2个结构镜像但奖励函数不同的DDQN网络模型。将任务调度问题视作2个智能体的马尔可夫博弈。

基于单智能体DDQN模型的梯度下降算法，每个智能体可在训练过程中学习到各自的最优策略π，采用式(3)的选择机制将多目标优化调度问题的博弈映射到联合分布的集合。

如图8所示，智能体1与智能体2分别对最大完工时间和总体协作配合度S进行优化。网络输入共用状态向量$\boldsymbol{\phi}_t$。经验池记录的奖励包含2部分，$r(t)\leftarrow r_{\text{make-span}}(t),r_{\text{coop}}(t)$。

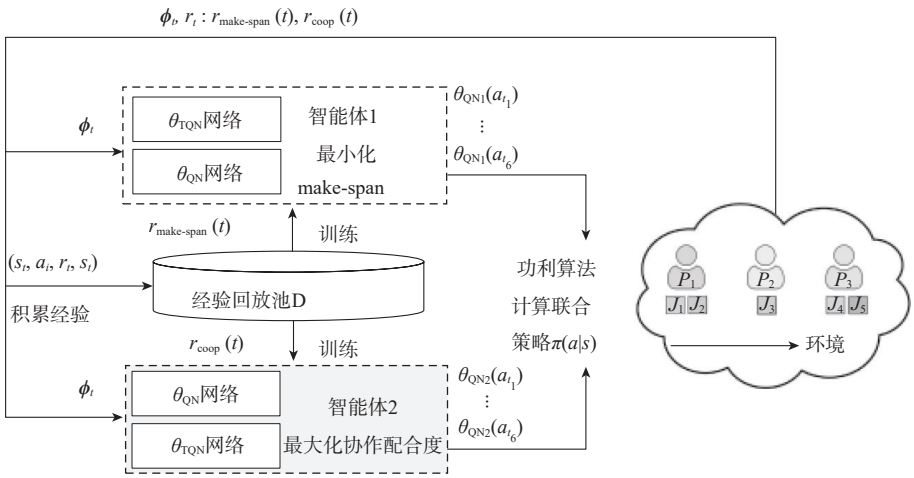


图8 多智能体博弈

Fig. 8 Multi agent game

3.8 目标深度Q值网络参数的软迁移方式

一般的DDQN网络采用N次训练后将θ_{QN}参数直接复制迁移到目标Q值网络θ_{TQN}的迁移方式，训练过程极不稳定。

本文采用Lillicrap等[17]提出的θ_{TQN}软迁移方式，迁移过程和网络参数如图9所示。

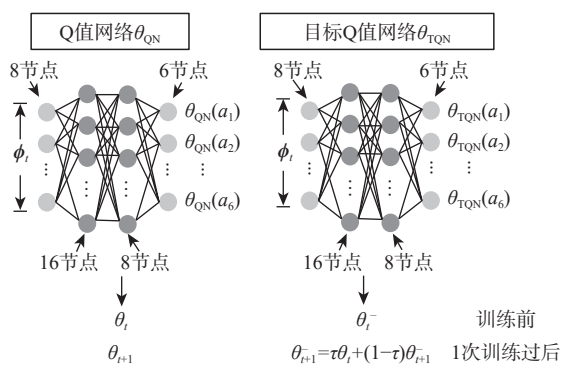


图9 θ_{TQN}内部参数的更新方法

Fig. 9 Methodology for updating θ_{TQN}

每次训练后，都对θ_{TQN}的内部参数θ^-以更新率τ更新，使其平滑地跟随θ_{QN}的内部参数θ，如下：

$$\theta_{t+1}^- = \tau\theta_t + (1-\tau)\theta_{t+1}^- \tag{10}$$

本文τ取0.125。参数包括神经网络每层的权重ω和偏置b。

4 模拟实验

4.1 仿真数据

实验采用的工作流图结构以国内某单位实际的办公流程拓扑结构作为参考蓝本，如图10所示。本文通过拓扑排序随机生成多类模板。每名员工可按系统自动生成的权限表办理不同种类的流程和不同种类的流程步骤，单员工可处理的种类

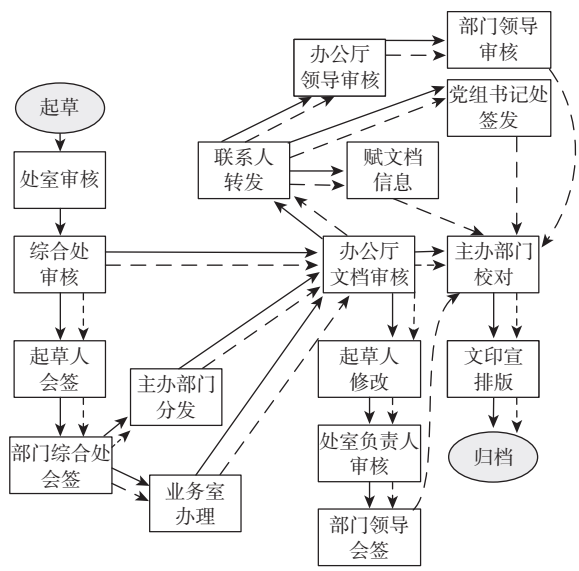


图10 原始拓扑结构

Fig. 10 Original topology structure

数没有限制。

为验证算法有效性, 本文对 6 类调度任务场景做了分类别的仿真测试。每个类别的调度任务场景包含 30 条流程、250～300 个流程节点数和 10～15 个业务人员, 每类流程由 5～10 个节点组成, 节点的拓扑结构从图 9 中截取部分得到。

对每个不同类别的场景采用算法, 进行 100 轮次训练, 每轮次的结束条件为 30 条流程(包含 250～300 个节点)都完成。每 128 步采样 128 个样本为一个训练批次, 观察前 32 批次训练的损失函数值与前 78 批次 θ_{QN} 参数 reward 奖励是否随训练轮次增加而收敛。

令当前重调度时刻 t 已积压的代办流程数为 n_1, 后续计划达到的流程数为 n_2, 对流程状态等输入数据的仿真按整体紧急程度 E_{deadline} 和业务到达时间分布分为表 2 所示的 6 类数据。

表 2　流程仿真数据类

Table 2　Process simulation data class

类别编号	E_{deadline}			n_1/n_2	
	5	5.6	6	1	0.5
1	√			√	
2	√				√
3		√		√	
4		√			√
5			√	√	
6			√		√

仿真系统给定业务员 $P_{i,j,k}(k|P_{i,j,k}\in P_{i,j})$办理 $O_{i,j}$ 的时间 $t_{i,j,k}$, 则流程 J_i 的平均办理时间仿真计算为 $\overline{C}_{\mathrm{i,e}_i}=\sum_k t_{i,j,k}$。设置 J_i 的截止时间 D_i 为 $\overline{C}_{\mathrm{i,e}_i}\times E_{\mathrm{deadline}}$。

仿真中不考虑人为等主客观因素。在每一调度步骤 $(O_{i,j},P_{i,j,k})$ 确定, 后按统计量 $t_{i,j,k}$ 的参数 $\bar{t}_{i,j,k}$(即 $P_{i,j,k}$ 办理 $O_{i,j}$ 的均值时长)取整执行。

4.2　算法有效性

本文算法本质是 RL 算法, 采用 RL 领域通用的 batch-reward 方法, 在一个完整流程执行过程中, 观察积分值是否随训练批次的增加而收敛。即在 128 步一批次训练后, 将当前智能体 1 和智能体 2 的参数冻结, 并重新代入到初始的数据场景中, 以冻结的参数序贯决策直到数据场景中所有流程都完成, 记录当前批次训练获得的最终奖励。

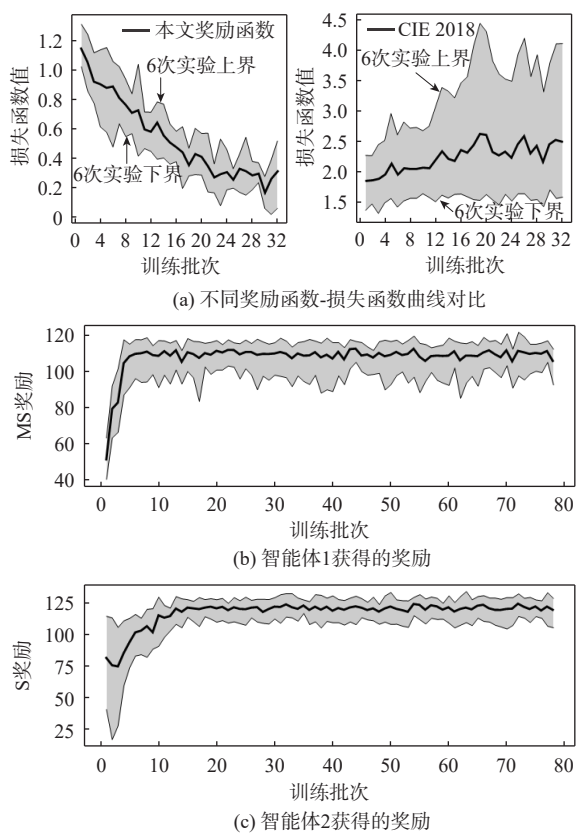
一般而言, 深度强化学习 Q 值网络损失函数的收敛性是判断算法稳定性的重要指标, 可以侧面反映算法的有效性, 本文对损失函数值的收敛性进行了对比。

针对奖励函数的设置, 经典的方法采用区间范围包含正负的奖励函数描述当前环境状态或调度步骤 $(O_{i,j},P_{i,j,k})$ 对当前状态 s_t 的影响[10,18], 或通过分析剩余业务等参数反向定义奖励函数的方法[11]。本文提供的 2 个奖励函数对应的多智能体在马尔可夫博弈的过程中实现了收敛, 且训练过程更为稳定。图 11 的实验结果为所述观点提供了依据。

在奖励函数方面, 本文采用相关领域引用较多的奖励函数[18], 对比不同算法对 Q 值网络参数收敛的影响。相较于目前多数研究采用的奖励函数, 本文模型的收敛性更优。

图 11(a) 的损失函数值为 2 个智能体的损失函数均值, 采用表 2 所示的 6 类数据测试 MDDQN 算法模型, 6 次失函数曲线的均值为加粗黑线。图 11(b)、(c) 分别展示了在 6 组数据中调用批训练冻结的参数在一个完整流程中得到的 2 部分奖励积分 $\int_t r_{\text{make-span}}(t)$ 和 $\int_t r_{\text{coop}}(t)$ 随训练轮次的增加而收敛的曲线。对每个智能体而言, 其参数都实现了局部收敛。

为进一步验证算法的有效性, 用 MDDQN 算法得到的调度方案对比表 1 中的 6 种单一规则。采用表 2 中类型 4 的数据进行测试, MDDQN 算法的最终完工时间为 492 h 与 6 种规则的最终完工时间如图 12 和表 3 所示。MDDQN 算法比最优的单一



(a) 不同奖励函数-损失函数曲线对比

(b) 智能体1获得的奖励

(c) 智能体2获得的奖励

图 11　收敛性验证

Fig. 11　Convergence verification

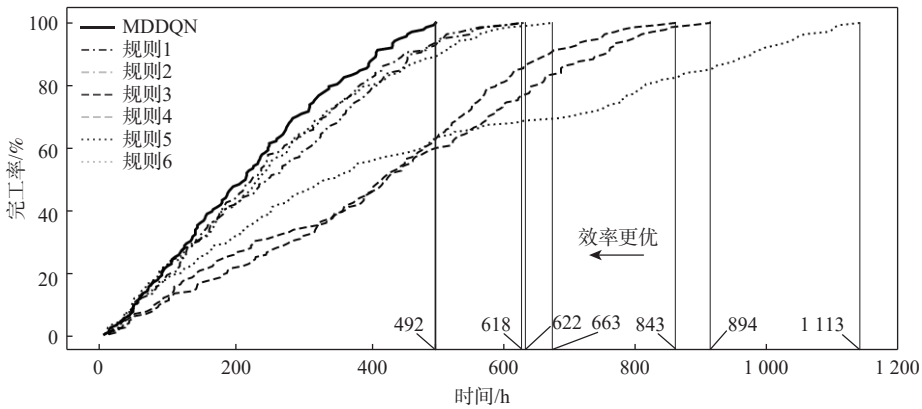


图 12 与单一规则比较

Fig. 12 Comparison with single rule

表 3 单一规则最终完工时间

Table 3 Single rule final completion

规则编号	最终完工时间消耗/h
1	622
2	618
3	894
4	843
5	663
6	1 113

规则在类型 4 数据的调度时间消耗上减少了 41.64%。

4.3 算法时间复杂度

为验证算法在时间复杂度上的优势，本文在表 2 类型 4 的参数仿真数据上与目前最优的元启发算法 ATO-MC 进行对比[4]。ATO-MC 采用执行者间信息素和启发式信息素结合手动调参方式实现双目标均衡。ATO-MC 进行一次完整重调度的时间复杂度为：$O(G_{\max}\times N_{\text{ant}}\times m^2\times N)$，$N=\sum_{i=1}^{n}n_i$为所有流程节点数，$m$ 为业务员数，代数为 $G_{\max}$，蚂蚁个数为 N_{ant}。

MDDQN 进行一次完整重调度的时间复杂度为：$O(N\times\left(1+\frac{N_{\text{batch}}}{I}\right)\times 2\times n_{\text{agents}}\times(i\times j+j\times k+k\times l))$，$i$ 为输入的特征向量$\boldsymbol{\phi}_t$的维数（本文取 8），j 和 k 为隐层神经元数（本文取 16 和 8），l为规则库规则数（本文取 6），n_{agents}为智能体数目（本文取 2），每隔 step 个重调度步骤后采样 batch 个样本训练一次（本文均为 128）。MDDQN 选取规则后，计算实际调度方案的时间复杂度为$\ln n+\ln m$。

取 250 节点数，m=15，蚁群设为 5 000，迭代数为 20 时，ATO-MC 需要约 750 亿次浮点计算，而本文算法仅需约 1 230 万次浮点计算。即使缩小蚁群规模，两者差距仍旧明显。

采用 30 条流程、250～300 个节点及 10～15 个执行人员时，元启发式算法运行时间过长，因此，对比实验基于类型 4 场景的参数缩减场景的规模，仿真得到 3 种规模的测试数据集，由小到大。10 次实验数据如图 13 所示。本文算法在时间复杂度上具有明显优势，且随问题规模的增加，算法时间复杂度并不明显增加，与算法复杂度分析的计算值吻合。

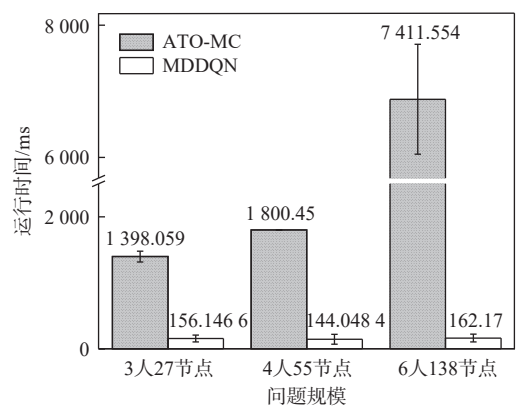


图 13 时间复杂度对比

Fig. 13 Time complexity comparison

4.4 最大完工时间与配合度

如果不考虑算法运行时间，元启发式算法在优化效果上与 MDDQN 算法接近。

本文在类型 4 仿真数据上进行了 10 次测试，与 ATO-MC 算法及 2 种典型的 RL 算法做对比实验，分别为 DRL 算法[11] 与 DDDQNPR（属于 D3QN 算法）算法[12]。

如图 14 和图 15 所示，本文算法在最大完工时间和协作配合度这 2 个优化指标上有一定优势。从稳定性角度出发，元启发算法相比 RL 算法有明显优势，而本文算法在解的稳定性方面略优于目前先进的单智能体强化学习算法。

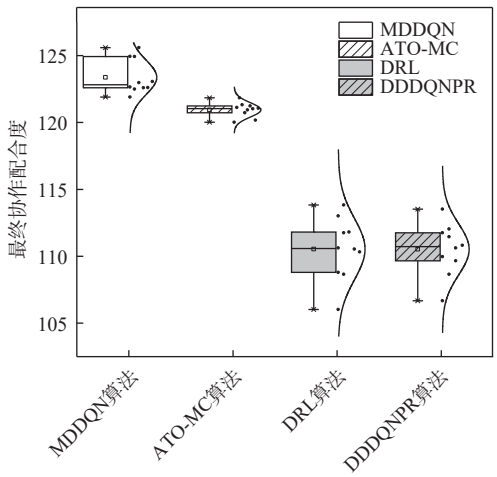


图 14　协作配合度对比箱型图

Fig. 14　Collaboration contrast box diagram

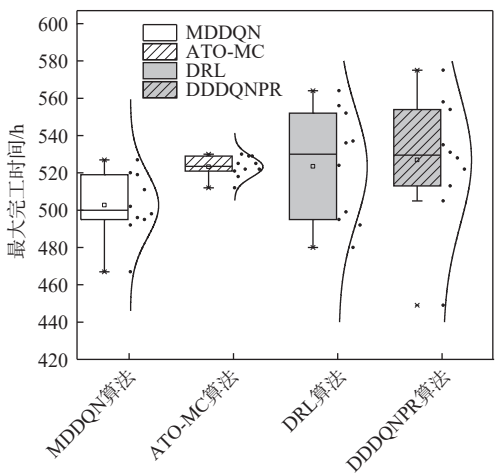


图 15　最大完工时间对比

Fig. 15　Make span comparison chart

5　结　论

本文将 RL 算法引入办公领域的工作流调度问题中, 提出一种可行的建模方法和改进的办公领域 DAG 调度优化算法。

1) 将协配合作度这一关键参数融入规则库和状态观测过程, 成功实现了一般 JSSP 的 RL 模型与本文具体问题对象的紧密结合。

2) 在相同的场景数据下获取多次测试结果, 在最大完工时间方面, 平均比 ATO-MC 减少了 3.94%, 相比目前同领域最优的单智能体强化学习算法减少了 3.97%。

3) 在总体协作配合度方面, 平均比 ATO-MC 增加了 2.01%, 相比目前同领域最优的单智能体强化学习算法增加了 11.61%。

4) 在时间复杂度方面, 在 6 名办理人 138 节点任务的数据场景中, 计算效率比元启发方式平均提高 45.7 倍。

虽然 RL 算法可以简便地实现线上训练, 但如果没有先验知识事先训练模型, 智能体就需要一段时间的线上数据积累来适应新的环境, 只有在一段时间的经验积累后, 其性能才能充分体现。在这段磨合期, 其解的质量不如启发式算法。如何进一步加快模型的收敛速度, 特别是提高模型在小样本下的效果是未来的重要研究方向。

参考文献（References）

[1] 关静静, 贺鹏涛, 张冉. 基于作业调度算法的医院 OA 系统的优化[J]. 中国医疗设备, 2018, 33(2): 155-157.
GUAN J J, HE P T, ZHANG R. Optimization of hospital OA system based on job scheduling algorithm[J]. China Medical Devices, 2018, 33(2): 155-157(in Chinese).

[2] GAO Y Q, ZHANG S Y, ZHOU J T. A hybrid algorithm for multi-objective scientific workflow scheduling in IaaS cloud[J]. IEEE Access, 2019, 7: 125783-125795.

[3] FARAHNAKIAN F, ASHRAF A, PAHIKKALA T, et al. Using ant colony system to consolidate VMs for green cloud computing[J]. IEEE Transactions on Services Computing, 2015, 8(2): 187-198.

[4] 吕龙, 胡海洋, 李忠金, 等. 基于蚁群算法的工作流系统优化任务分配[J]. 计算机集成制造系统, 2018, 24(7): 1723-1735.
LYU L, HU H Y, LI Z J, et al. Optimizing task allocation in workflow system based on ant colony optimization[J]. Computer Integrated Manufacturing Systems, 2018, 24(7): 1723-1735(in Chinese).

[5] KUMAR A, DIJKMAN R M, SONG M S. Optimal resource assignment in workflows for maximizing cooperation[C]//Proceedings of the 11th International Conference on Business Process. Berlin: Springer, 2013: 235-250.

[6] 许荣斌, 鲍广华, 杨培全, 等. 基于最大依赖度及最小冗余度的员工协作优化策略[J]. 计算机集成制造系统, 2017, 23(5): 1014-1019.
XU R B, BAO G H, YANG P Q, et al. Staff collective optimization strategy based on maximal dependency and minimal redundancy[J]. Computer Integrated Manufacturing Systems, 2017, 23(5): 1014-1019(in Chinese).

[7] PIROOZFARD H, WONG K Y, WONG W P. Minimizing total carbon footprint and total late work criterion in flexible job shop scheduling by using an improved multi-objective genetic algorithm[J]. Resources, Conservation and Recycling, 2018, 128: 267-283.

[8] CUI D L, KE W D, PENG Z P, et al. Multiple DAGs workflow scheduling algorithm based on reinforcement learning in cloud computing[C]//Proceedings of the International Symposium on Computational Intelligence and Intelligent Systems. Berlin: Springer, 2015: 305-311.

[9] WU J H, PENG Z P, CUI D L, et al. A multi-object optimization cloud workflow scheduling algorithm based on reinforcement learning[C]//Proceedings of the International Conference on Intelligent Computing. Berlin: Springer, 2018: 550-559.

[10] WANG Y F. Adaptive job shop scheduling strategy based on weighted Q-learning algorithm[J]. Journal of Intelligent Manufacturing, 2020, 31(2): 417-432.

[11] CAO Z, ZHANG H G, CAO Y, et al. A deep reinforcement learn-

ing approach to multi-component job scheduling in edge computing[C]//Proceedings of the 15th International Conference on Mobile Ad-Hoc and Sensor Networks. Piscataway: IEEE Press, 2020: 19-24.

[12] HAN B A, YANG J J. Research on adaptive job shop scheduling problems based on dueling double DQN[J]. IEEE Access, 2020, 8: 186474-186495.

[13] MNIH V, KAVUKCUOGLU K, SILVER D, et al. Human-level control through deep reinforcement learning[J]. Nature, 2015, 518 (7540): 529-533.

[14] VAN HASSELT H, GUEI A, SILLVER D. Deep reinforcement learning with double Q-Learning[C]//Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence. New York: ACM, 2016: 2094-2100.

[15] VAN HASSELT H. Double Q-learning[C]//Proceedings of the 23rd International Conference on Neural Infor- mation Processing Systems. New York: ACM, 2010: 2613-2621.

[16] VAN HASSELT H V, GUEZ A, SILVER D. Deep reinforcement learning with double Q-learning[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2016, 30(1): 1509.

[17] LILLICRAP T P, HUNT J J, PRITZEL A, et al. Continuous control with deep reinforcement learning[EB/OL]. (2019-07-05) [2022-04-01]. https://arxiv.org/abs/1509.02971.

[18] SHIUE Y R, LEE K C, SU C T. Real-time scheduling for a smart factory using a reinforcement learning approach[J]. Computers & Industrial Engineering, 2018, 125: 604-614.

Optimization of office process task allocation based on deep reinforcement learning

LIAO Chenyang[1], YU Jinsong[1,*], LE Xiangli[2]

(1. School of Automation Science and Electrical Engineering, Beihang University, Beijing 100191, China;
2. Shenyuan Honors College, Beihang University, Beijing 100191, China)

Abstract: In the office platform, we often need to face a large number of parallel heterogeneous process tasks. This not only tests the ability of task executors but also puts forward requirements for the performance of the scheduling system. The multi-agent game model based on Markov game theory is proposed in this paper, which adopts the reinforcement learning (RL) approach along with quantitative analysis of the degree of cooperation and relaxation. This model realizes the optimal scheduling system with the overall process degree and maximum completion time as the optimization objectives and enhances the overall execution efficiency. Finally, to confirm the efficacy of this approach, the meta-heuristic algorithm based on ant colony and the reinforcement learning algorithm based on D3QN and deep reinforcement learning (DRL) are contrasted using the real business system process as the experimental data and the identical optimization targets.

Keywords: work-flows; tasks scheduling; Markov games; deep reinforcement learning; cooperation degree

Received: 2022-04-28; **Accepted**: 2022-05-06; **Published Online**: 2022-05-27 15:48

URL: link.cnki.net/urlid/11.2625.V.20220526.2003.002

Foundation item: National Key R & D Program of China (2018YFB1004100)

* **Corresponding author**. E-mail: yujs@buaa.edu.cn

http://bhxb.buaa.edu.cn jbuaa@buaa.edu.cn
DOI: 10.13700/j.bh.1001-5965.2022.0322

改进 YOLOv5s 的弱光水下生物目标检测算法

陈宇梁，董绍江*，孙世政，闫凯波
(重庆交通大学 机电与车辆工程学院，重庆 400074)

摘　　要：针对水下光学图像目标检测过程中由于水中光线衰弱严重、图像环境复杂和拍摄设备移动等造成的生物识别精度低的问题，提出了基于改进 YOLOv5s 的弱光水下生物目标实时检测算法 YOLOv5s-underwater。针对弱光水下光线衰弱的问题，引入了限制对比度自适应直方图均衡 (CLAHE) 算法对输入图像进行预处理，解决了颜色失真和图像毛糙的问题。针对复杂的弱光水下图像环境，提出了快速空间金字塔池化 (SPPF) 模块，解决了水下物体区分度低和特征损失严重的问题。针对拍摄设备移动带来的场景和形态变化问题，提出了一种基于旋转窗口的 Swin-Transformer 模块，提高了模型的泛化能力。针对水下小目标，修改了网络模型结构，提高了小目标的检测能力。仿真和实验结果表明：所提算法相较于 YOLOv5s 检测精度提高 30.7%，证明了算法的有效性。

关 键 词：弱光水下生物目标；YOLOv5s；限制对比度自适应直方图均衡；快速空间金字塔池化；旋转窗口
中图分类号：TP391
文献标志码：A　　**文章编号**：1001-5965(2024)02-0499-09

近年来，随着中国渔业的发展，对水下渔业资源的准确勘测就显得尤为重要。相较于声学图像，光学图像具有分辨率高和信息丰富的特点，并且在获取方式上，光学图像成本也更低，是目前水下近距离环境感知的重要手段之一[1-3]。而水下光学图像受水中光线衰弱的影响，易出现图像特征模糊、目标物体与周围环境区分度较低的情况[4-5]，从而影响目标检测，且水下大部分目标由于拍摄位置较远，光学图像中多为小目标。通常，将目标尺寸为原始图像尺寸十分之一的目标认定为小目标[6]。由于搭载拍摄设备的载体在水下进行运动时受到来自环境的约束较小，难以固定观察距离和视角，且大多水下生物存在多种姿态，进一步加剧了水下光学图像目标检测的难度[7]。

传统的目标检测算法通常由人工进行目标特征提取后再进行分类，效果差、速度慢。自卷积神经网络被提出以来，基于深度学习的目标检测算法能够在同一个网络中进行不同尺度的特征提取，极大地提升了目标检测的效率。

目前，基于深度学习的目标检测算法主要分为2类：以 R-CNN 系列算法为代表的基于候选框的双阶段目标检测算法和以 YOLO 系列算法为代表的基于回归的单阶段目标检测算法。双阶段目标检测算法中具有代表性的是 Girshick 等提出的超快区域卷积神经网络(faster region convolutional noural network, Faster R-CNN)，该算法以速度和计算资源为代价，提高了目标检测与识别的精度[8]。为提高检测速度，降低资源消耗，提出了单阶段目标检测

收稿日期：2022-05-06；**录用日期**：2022-07-01；**网络出版时间**：2022-07-14 15:32
网络出版地址：link.cnki.net/urlid/11.2625.V.20220713.1906.001
基金项目：国家自然科学基金 (51775072)；重庆市科技创新领军人才支持计划 (CSTCCCXLJRC201920)；重庆市高校创新研究群体 (CXQT20019)；重庆市北碚区科学技术局技术创新与应用示范项目 (2020-6)；城市轨道交通车辆系统集成与控制实验室开放基金 (CKLURTSIC-KFKT-202007)
* **通信作者**. E-mail：dongshaojiang100@163.com

引用格式：陈宇梁，董绍江，孙世政，等. 改进 YOLOv5s 的弱光水下生物目标检测算法 [J]. 北京航空航天大学学报，2024，50(2)：499-507.
CHEN Y L，DONG S J，SUN S Z，et al. Improved YOLOv5s low-light underwater biological target detection algorithm [J]. Journal of Beijing University of Aeronautics and Astronautics，2024，50(2)：499-507 (in Chinese).

算法，其中具有代表性的是 Redmon 和 Farhadi 提出了基于深度学习的 YOLO 算法系列，该算法将目标框预测与识别 2 步进行融合，大幅提升检测速度和减少资源消耗的同时，降低了检测精度[9-10]。Liu 等基于使用不同特征图对不同尺寸的物体进行检测的思想，提出了单阶段多尺度目标检测（single shot multibox detecor, SSD）算法，但检测速度未得到提升，难以达到实时性要求[11]。YOLOv5 是目前 YOLO 系列算法中性能和通用性都极强的模型，其相较于以上算法在检测速度和精确度方面都更具优势[12]。其中 YOLOv5s 相较于 YOLOv5m、YOLOv5l、YOLOv5x 3 种不同版本的 YOLOv5 模型具有更小的模型体积和更快的检测速度，但精度相对较低。其余 3 种 YOLOv5 模型具有更高精度，但模型较大，且检测实时性较差，不利于现场的实时检测。因此，本文针对弱光水下生物目标实时检测的特定场景对 YOLOv5s 的网络结构进行改进。

首先，针对光线在水下传播时极易衰弱，且对不同颜色的光的衰减程度不同，从而导致色彩失真和轮廓毛糙的问题，本文将限制对比度自适应直方图均衡 (contrast limited adaptive histgram equalization, CLAHE) 算法加入输入端中，对输入图像进行预处理，对输入的弱光水下图像的 3 个颜色通道分别进行增强，提高图像的色彩及清晰度；其次，针对水下图像环境复杂、能见度较低等原因使得水下物体与周围环境间的区分度较低和图像纹理特征损失的问题，本文提出了快速空间金字塔池化（spatial pyramid pooling fast, SPPF）模块，以提高模型的纹理及轮廓特征描述能力；然后，针对大多数水下生物在图像中的形态和尺寸在不同的时间点与角度而差别较大的问题，本文提出了 Swin-Transformer（ST），模块以提高模型的泛化能力，加强对形态和尺寸变化较大物体的检测与识别能力；最后，针对水下图像中的小目标难以检测的问题，对颈部层结构进行修改，加强了网络对小目标特征的融合能力，提高了对小目标的检测和识别能力。

1 YOLOv5s-underwater 算法

本文将 YOLOv5s 与 CLAHE 算法、ST 模块和 SPPF 模块相结合，提出了 YOLOv5s-underwater 结构，如图 1 所示。

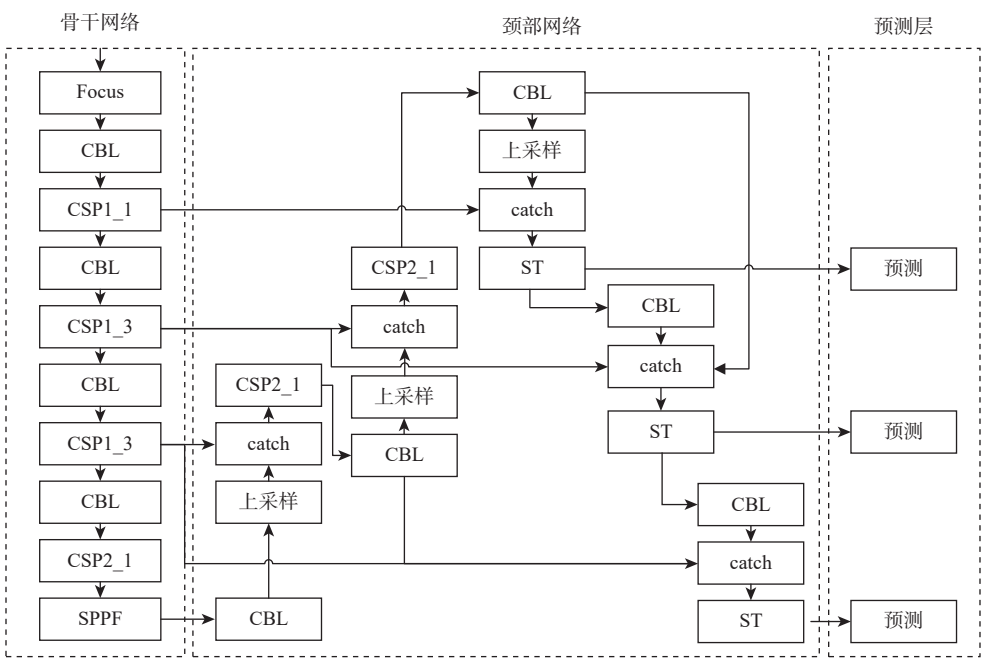


图 1 YOLOv5s-underwater 网络结构

Fig. 1 Network structure for YOLOv5s-underwater

1.1 CLAHE 算法

自适应直方图均衡化（adaptive histgram equalization, AHE）算法通过计算图像多个局部区域的直方图，并重新分布亮度，以此改变图像对比度，完成对图像的增强，但存在细节信息处理较差和放大图像噪声等问题。CLAHE 算法在图像的局部区域执行相应直方图均衡化算法，并对各小区域都引入对比度限幅，以此在优化图像对比度的同时，抑制其对噪声的过度放大。

CLAHE 算法先对图像进行图像分块填充等预处理，再对每个分块进行处理，使用对比度限制计算映射关系，最终使用插值方法得到增强

图像[13]。

本文针对光线在水下传播时极易衰弱从而导致色彩失真和轮廓毛糙的问题, 将 CLAHE 算法加入输入端中, 对输入的弱光水下图像的 3 个颜色通道分别进行增强后合成, 并与 AHE 算法进行对比, 结果如图 2 所示。

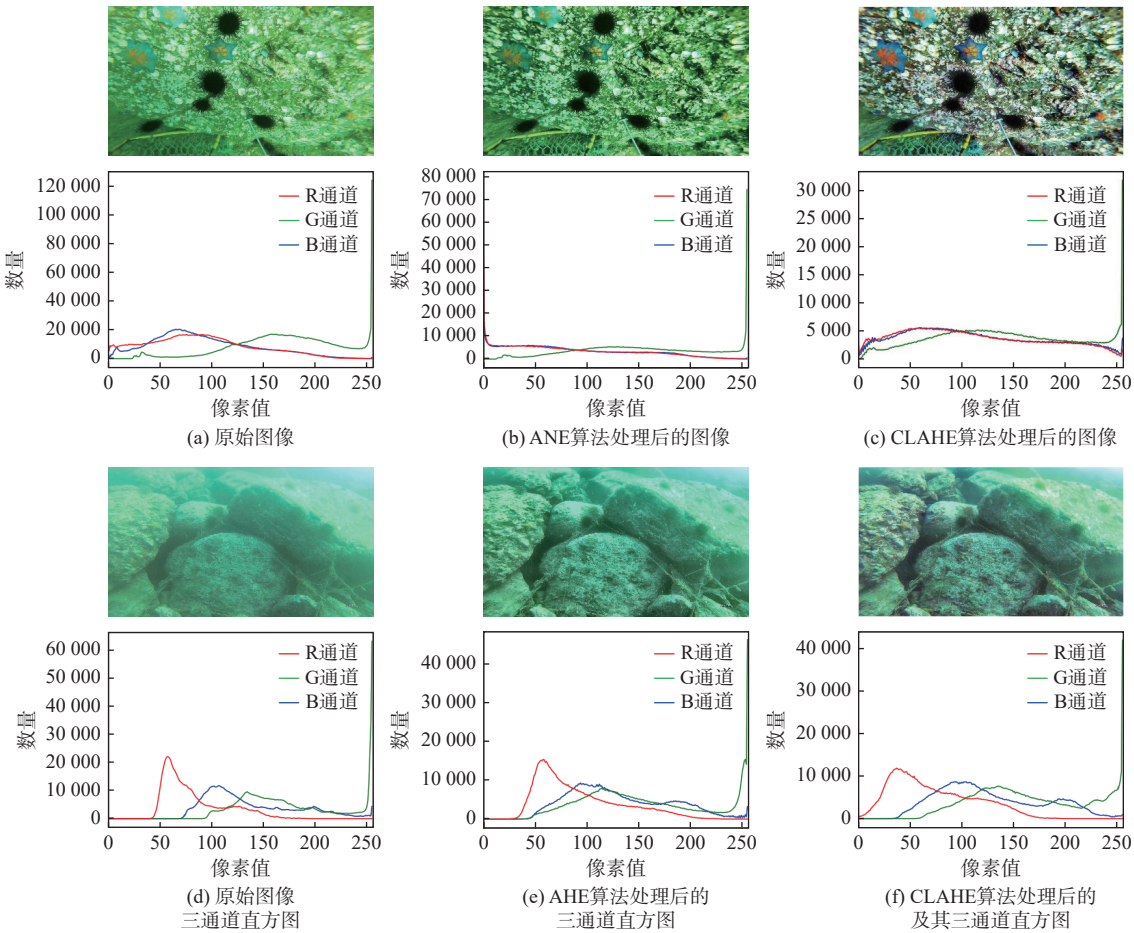


图 2　弱光水下图像经算法处理前后效果对比

Fig. 2　Comparison of effect of shallow sea underwater image before and after algorithm processing

1.2　SPPF 模块

由于水下图像环境复杂、能见度较低等, 图像纹理特征损失, 使得水下物体与周围环境间的区分度较低, 纹理不清晰。针对该问题, 本文提出了基于同尺寸的最大值池化层的 SPPF 模块, 以替代原 YOLOv5s 模型中由不同尺寸的最大值池化层组成的空间金字塔池化(spatial pyramid pooling , SPP)模块。SPPF 模块结构如图 3 所示。

SPPF 模块对输入特征图进行尺度均为 5×5 但程度不同的最大值池化, 并对各阶段池化层输出进行通道连接。其过程如下:

$$\begin{cases} \hat{z}^{l-1} = \mathrm{CBL}\left(z^{l-1}\right) \\ \hat{z}_1^{l-1} = \mathrm{MaxPool}\left(\hat{z}^{l-1}\right) \\ \hat{z}_2^{l-1} = \mathrm{MaxPool}\left(\hat{z}_1^{l-1}\right) \\ \hat{z}_3^{l-1} = \mathrm{MaxPool}\left(\hat{z}_2^{l-1}\right) \\ \hat{z}^{l-1} = \mathrm{CBL}\left(\left[\hat{z}^{l-1}; \hat{z}_1^{l-1}; \hat{z}_2^{l-1}; \hat{z}_3^{l-1}\right]\right) \end{cases} \tag{1}$$

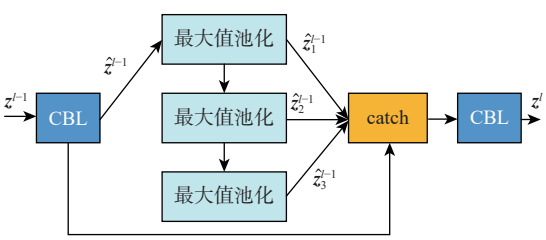


图 3　SPPF 模块结构

Fig. 3　Structure of SPPF module

在避免局部特征损失的同时, SPPF 模块能够减少冗余信息, 保留突出的纹理特征, 以使网络获得更强的纹理及轮廓特征描述能力, 从而降低由于水下图像轮廓及纹理特征不明显造成的影响。相较于 SPP 模块, SPPF 模块具有更快的前向传播速度, 从而缓解了由于融入 ST 模块带来的检测速度损失。

1.3　ST 模块

由于大多数水下生物在水下所受约束较少, 且图像中的姿态会随着运动或拍摄角度变化而发生

较大变化，存在同一生物的形态和尺寸在不同时间与角度差别较大的问题，从而对模型泛化性要求较高。本文参考ST模块[14]的特征提取及特征表征的方法，在YOLOv5s模型中融入ST模块基于滑动窗口技术的滑动窗口多头自注意力（shifted-window multi-head self-attention, SW-MSA）模型，该方法通过小图像片元和逐层进行邻域合并的方式构建层级特征表达，提高了模型对目标识别的泛化性，解决了姿态和尺寸的问题，增强了模型对特征的提取与表征能力。本文引入的ST模块由2个连续的Swin-Transformer结构组成，如图4所示，2个结构分别在一个窗口中计算窗口多头自注意力（window multi-head self-attention, W-MSA）和SW-MSA模型。图中，LN为layer norm模块，MLP为多层感知机模块。

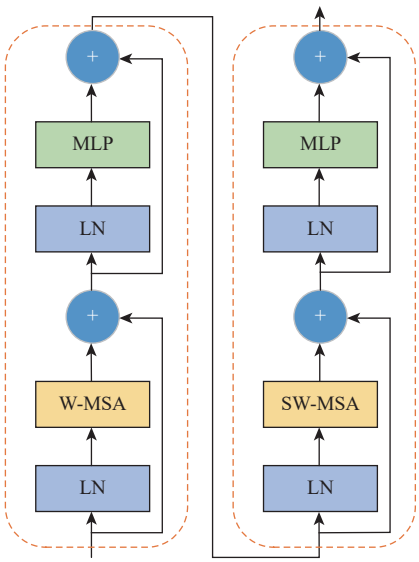


图4 Swin-Transformer模块

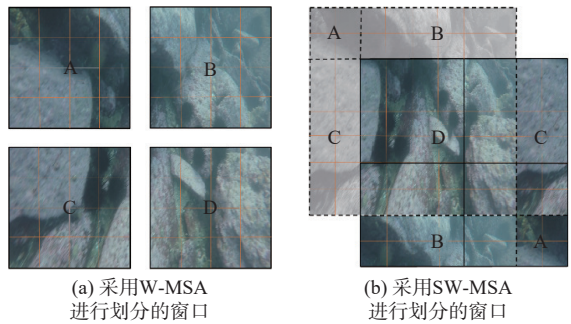
Fig. 4 Swin-Transformer module

假设输入图像分为8×8个块，一个窗口包含4×4个块。W-MSA方法如图5(a)所示，其将图层划分为不重叠的窗口，在各窗口中计算自注意力，但该方法使得各个窗口间缺乏信息交互，限制了特征表征能力。因此，SW-MSA引入了滑动窗口，使窗口进行滑移的同时，将区域进行移动填补，实现窗口间信息的交互，如图5(b)所示。

进行如图5(b)所示的变换后，去除了不同窗口中的特征图计算，大大降低了模型的计算延时。注意力计算为

$$\boldsymbol{Z}=\mathrm{Softmax}\left(\frac{\boldsymbol{QK}^{\mathrm{T}}}{\sqrt{d}}+\boldsymbol{B}\right)\boldsymbol{V} \tag{2}$$

式中：$\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}\in\mathbf{R}^{M^2\times d}$分别为查询（query）、键（key）、值（value），M^2为窗口中块的总数，d为键的维数；$\mathbf{B}\in\mathbf{R}^{M^2\times M^2}$为相对位置偏差，表征块间的相对位置。



(a) 采用W-MSA进行划分的窗口　(b) 采用SW-MSA进行划分的窗口

图5 窗口划分

Fig. 5 Windows partition

1.4 特征连接策略

为结合更丰富的图像信息，本文增加了颈部层的网络层数，并增强了模型对所提取多尺度特征的融合能力。输入图像经过多层特征提取后，底层的目标信息已非常模糊，因此，在增加上采样网络模块进行特征修复[15]的同时，增强与浅层特征的连接程度，进一步弥补并加强了定位信息。

颈部层是为了更好地利用骨干网络提取的特征，因此，本文结合了CSP2_X模块和ST模块对骨干网络提取的特征图进行不同阶段的再处理和合理使用。将模型17层后的结构进行修改，结构如图6所示。

为进一步结合模型所提取的低层次特征，本文增加了一次上采样，使特征图的尺寸恢复到与2层输出特征图相同的二维尺寸，并进行通道连接。后续也对4层和6层的输出进行通道连接，提升模型对较低层特征的融合能力，增强了模型对小目标物体的检测能力，并在特征输入预测层前通过ST模块进行特征提取。

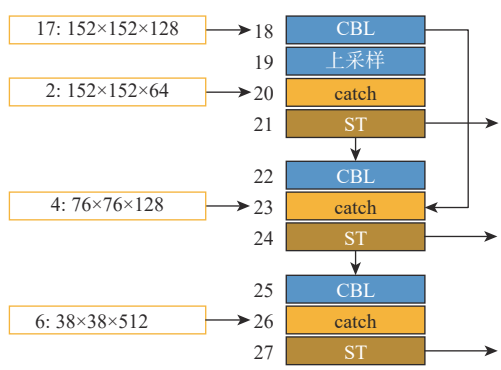


图6 颈部层部分网络结构

Fig. 6 Neck layer partial network structure

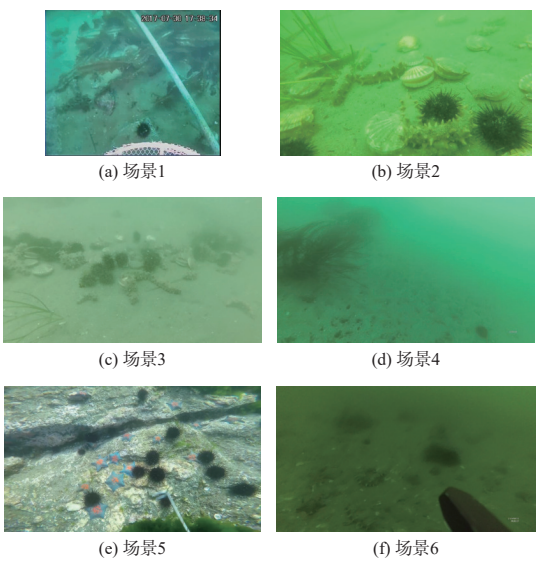
2 实验与分析

2.1 实验环境与数据集

实验环境为Windows操作系统，AMD Ryzen 5 5600X（CPU），16 GB随机存取内存（RAM），RTX

3060 Ti(GPU), 深度学习框架为 pytorch。

水下渔业资源勘探时, 光学图像主要通过水下机器人搭载摄像头的方式获取。因此, 通过鹏城实验室虎鲸开源计划的“underwater detection”数据集对本文算法的可靠性进行验证。该数据集混杂了多场景中通过水下机器人和潜水员携带摄像机所拍摄的图像, 能够充分模拟实际弱光中水下生物目标检测。数据集中包含 5 类水下生物目标 (海参、海胆、扇贝、海星和水草) 的多场景图像 7 543 幅, 共 56 563 个目标, 其中, 主要场景的部分图像如图 7 所示。取 755 幅图像(约 10%)作为验证集, 754 幅图像(约 10%)作为测试集, 剩余 6 034 幅图像(约 80%)作为训练集进行训练。将输入图像的尺寸统一处理为 608×608, 便于网络对图像特征的提取。



(a) 场景1　(b) 场景2　(c) 场景3　(d) 场景4　(e) 场景5　(f) 场景6

图 7　数据集图像

Fig. 7　Datasets images

2.2　评价指标

本文采用精确度P、召回率R、平均精确度P_m、量化指标F_1和检测速度 4 种评价指标对模型进行评价。

通过混淆矩阵对精确度、召回率和平均精确度进行计算, 混淆矩阵包含 4 个指标: 即真阳性T_P、真阴性T_N、假阳性F_P和假阴性F_N, 如图 8 所示。图中, 预测框与锚框之间的IoU值≥所设阈值的预测框数量为T_P, 相反, IoU<阈值的数目则为F_P, 未检测到的目标数量为F_N。

预测值 \ 真实值	阳性	阴性
阳性	T_P	F_P
阴性	F_N	T_N

图 8　二元分类的混淆矩阵

Fig. 8　Confusion matrix for binary classification

精确度指预测值为阳性的总数量中真阳性所占比例, 其直观反映模型误检程度的衡量值, 表示如下:

$$P=\frac{T_P}{T_P+F_P} \tag{3}$$

召回率指实际值为阳性的总数量中真阳性所占比例, 其直观反映模型漏检程度的衡量值, 表示如下:

$$R=\frac{T_P}{T_P+F_N} \tag{4}$$

平均精确度P_m为所有物种精确度的平均值, 表示如下:

$$P_m=\frac{1}{C}\sum_{i=1}^{C}P_i \tag{5}$$

式中: C为物种数量。

平均精确度中包含了P_m@0.5和P_m@0.5 : 0.95 2 种, 可对说明语序进行调换, 其中, P_m@0.5为IoU阈值取 0.5 时, 对于其中一个类别, 有n个正例样本, 对这n个样本的精确度求平均。将平均精确度的IoU阈值从 0.5 以 0.05 为步长增至 0.95, 并取其平均值, 计算式如下:

$$P_m@0.5=\frac{1}{C}\sum_{i=1}^{C}P@0.5_i \tag{6}$$

$$P_m@0.5:0.95=\frac{1}{10}(P_m@0.5+P_m@0.55+\cdots+P_m@0.95) \tag{7}$$

P_m@0.5表现精确度和召回率的变化趋势, 值越高, 越易使二者保持较高水平。

P_m@0.5 : 0.95是在不同IoU阈值取值情况下的综合表现, 相较于P_m@0.5更加考虑整体情况, 值越高, 模型高精度边界回归能力越强, 即预测框与锚框的拟合越精确。

平均精确度和召回率相互制约, 一方提升都可能导致另一方下降, 难以对比。为综合考虑平均精确度和召回率, 本文引入F_1作为量化评价指标, 表示如下:

$$F_1=2\frac{P_m\cdot R}{P_m+R} \tag{8}$$

2.3　实验结果与分析

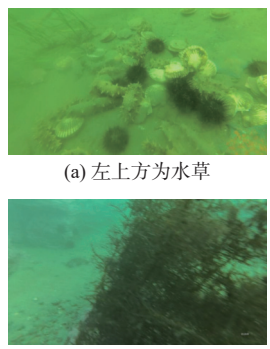
将 YOLOv5s、YOLOv5s+CLAHE、YOLOv5s+SPPF、YOLOv5s+CLAHE+SPPF、YOLOv5s+SPPF+ST、YOLOv5s+CLAHE+ SPPF+ST 与本文所提 YOLOv5s-underwater 进行对比实验, 精确度、平均精确度、召回率和F_1结果如表 1 所示。

表1 改进的YOLOv5s和YOLOv5s实验结果1

Table 1 Improved YOLOv5s and YOLOv5s experimental results 1 %

算法	精确度					平均精确度	召回率	F_1
	海参	海胆	扇贝	海星	水草			
YOLOv5s	61.5	67.7	65.3	69.9	20.0	56.9	68.1	62.00
YOLOv5s+CLAHE	73.0	73.4	72.4	74.2	89.7	76.5	68.4	72.22
YOLOv5s+SPPF	63.8	68.1	62.9	71.4	88.7	71.0	66.9	68.89
YOLOv5s+CLAHE+SPPF	74.3	76.9	72.1	76.6	94.2	78.8	67.1	72.48
YOLOv5s+SPPF+ST	69.3	74.6	71.4	76.6	97.2	77.8	66.0	71.42
YOLOv5s+CLAHE+SPPF+ST	81.9	84.2	79.7	80.5	99.1	85.1	68.1	75.66
YOLOv5s-underwater	84.8	85.6	83.0	84.8	99.8	87.6	67.4	76.18

通过表1可以看出，本文改进后的YOLOv5s-underwater算法相较于YOLOv5s算法对各目标物种的检测能力除R外均有所提高。其中，YOLOv5s+CLAHE通过对输入图像进行CLAHE算法处理，提高了算法的识别能力，使得平均精确度提升了19.6%，召回率提高了0.3%，F_1提高了10.22%。YOLOv5s+SPPF通过引入SPPF模块，增强了模型的纹理特征提取能力，使得算法对于该数据集的平均精确度提升了14.1%，召回率下降了1.2%，F_1提高了6.89%，对于尺寸相对较小、易与水下垃圾混淆的扇贝，其准确度下降了2.4%。YOLOv5s+CLAHE+SPPF通过引入SPPF模块后加入CLAHE算法对图像特征进行了增强，使得模型对于该数据集较YOLOv5s+CLAHE的平均精确度提升了2.3%，召回率降低了1.3%，F_1提高了0.26%；相较于YOLOv5s+SPPF，该模型的平均精确度提升了7.8%，召回率提升了0.2%，F_1提高了3.59%。YOLOv5s+SPPF+ST通过引入SPPF模块的同时将CSP2_X模块替换成ST模块，增强了模型的泛化能力和小目标的检测能力，使得算法相较YOLOv5s+SPPF对于该数据集的平均精确度提升了6.8%，召回率降低了0.9%，F_1提高了2.53%。YOLOv5s+CLAHE+SPPF+ST相较于YOLOv5s+SPPF+ST对于该数据集的平均精确度提升了7.3%，召回率提升了2.1%，F_1提高了4.24%；相较于YOLOv5s+CLAHE+SPPF，该算法的平均精确度提高了6.3%，召回率提升了1.0%，F_1提高了3.18%。YOLOv5s-underwater结合了以上特点，并对模型的颈部网络部分进行修改，使平均精确度提升了30.7%，F_1提高了14.18%，相较于YOLOv5s+CLAHE+SPPF+ST，对于该数据集的平均精确度提升了2.5%，F_1提高了0.52%。其中，该优化模型对于水草目标的检测效果提升极为明显，由于所使用数据集中水草目标的数量仅为87个，且外形差异较大(见图9)。证明该改进方法对于数据中的小样本、外表不固定的数据具有较好的优化效果。



(a) 左上方为水草

(b) 中部及右侧为水草

图9 部分包含水草的图像

Fig. 9 Section contains partially images of waterweeds

通过P_m@0.5、P_m@0.5:0.95和检测速度进一步验证算法的优化效果，结果如表2所示。

可以看出，YOLOv5s+CLAHE相较于YOLOv5s，mAP@0.5提升了9.6%，mAP@0.5 : 0.95提升了3.7%，检测速度降低了27.78帧/s；YOLOv5s+SPPF的mAP@0.5提升了7.5%，mAP@0.5 : 0.95提升了3.0%，SPPF模块在前向传播中速度显著，检测速度提高了81.06帧/s；YOLOv5s+CLAHE+SPPF相比于YOLOv5s+SPPF，检测速度降低了8.02帧/s，但P_m@0.5和P_m@0.5 : 0.95分别提升了3.6%和3.1%；相较YOLOv5s+

表2 改进的YOLOv5s和YOLOv5s实验结果2

Table 2 Improved YOLOv5s and YOLOv5s experimental results 2

算法	P_m@0.5/%	P_m@0.5 : 0.95/%	检测速度/(帧·s^{-1})
YOLOv5s	60.6	32.3	166.67
YOLOv5s+CLAHE	70.2	36.0	138.89
YOLOv5s+SPPF	68.1	35.3	247.73
YOLOv5s+CLAHE+SPPF	71.7	38.4	239.71
YOLOv5s+SPPF+ST	68.9	37.9	154.86
YOLOv5s+CLAHE+SPPF+ST	74.1	39.8	150.10
YOLOv5s-underwater	74.1	41.6	146.84

CLAHE, P_m@0.5和 P_m@0.5:0.95 分别提升了 1.5%和 2.4%, 检测速度提高了 100.82。YOLOv5s+SPPF+ST 相较于 YOLOv5s 和 YOLOv5s+ SPPF, P_m@0.5分别提升了 8.3% 和 0.8%, P_m@0.5 : 0.95分别降低了 5.6% 和 2.6%, 检测速度较 YOLOv5s 仅降低了 11.81 帧/s。进一步改进的 YOLOv5s+CLAHE+SPPF+ST 相较于 YOLOv5s+ SPPF+ST, P_m@0.5和 P_m@0.5 : 0.95分别提升了 5.2% 和 1.9%; 相较于 YOLOv5s+CLAHE+SPPF, P_m@0.5和 mAP@0.5 : 0.95分别提升了 2.4% 和 1.4%。本文算法 YOLOv5s-underwater 较 YOLOv5s+CLAHE+SPPF+ST, P_m@0.5 : 0.95提升了 1.8%, 检测速度降低了 3.26 帧/s; 相较于 YOLOv5s, P_m@0.5提升了 13.5%, P_m@0.5 : 0.95提升了 9.3%, 检测速度降低了 19.83 帧/s, 但对实时性影响较小。以上对比验证了 4 种优化方式均提高了网络的检测能力。

综合以上分析结果表明, 数据样本中光线主要来自于自然光, 以及多种使用场景进行切换, 因此, 存在多种色彩失真情况, 降低了网络对特征提取的能力, 在图像预处理中引入 CLAHE 算法极大改善了这一情况, 使图像能够通过增强方式改善关于不同色彩失真导致特征提取能力降低的问题。

同时, 拍摄设备随载具一同移动, 并伴随有晃动, 因此, 出现部分图像较为模糊和图像背景变换的情况, 如图 7(f)所示, YOLOv5s-underwater 模型中引入 SPPF 模块加强了目标轮廓和纹理特征提取能力。随着载具的运动和视角的切换, 同一生物的形态和尺寸也随之变换, 引入 ST 模块增强了网络对特征的提取和表征能力, 从而提高了模型对搭载的摄像头所拍摄到的弱光水下生物图像进行的目标检测能力。

为进一步验证本文提出的 YOLOv5s-underwater 算法在弱光水下生物目标检测识别上优于其他目标检测算法, 使用同样的弱光水下生物目标检测数据集对算法检测效果进行验证, 结果如表 3 和表 4 所示。

表 3　改进的 YOLOv5s 和其他目标检测算法实验结果 1 对比

Table 3　Comparison of experimental results 1 of improved YOLOv5s and other target detection algorithms

网络模型	平均精确度/%	召回率/%	F_1/%	检测速度/(帧·s^{-1})
Faster R-CNN	40.1	16.4	23.28	20.44
SSD	37.7	14.9	21.36	30.17
Mbv2-SSD	31.4	12.5	17.88	70.16

表 4　改进的 YOLOv5s 和其他目标检测算法实验结果对比

Table 4　Comparison of experimental results of improved YOLOv5s and other target detection algorithms

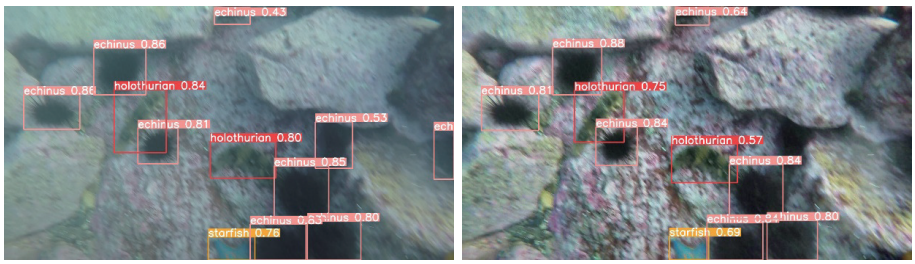
网络模型	平均精确度/%	召回率/%	F_1/%	检测速度/(帧·s^{-1})
CenterNet	41.9			93.38
YOLOv5m	67.1	56.7	61.46	83.33
YOLOv5l	77.7	64.0	70.19	52.63
YOLOv5x	83.1	70.9	76.52	32.26
YOLOv5s-underwater	87.6	67.4	76.18	140.84

通过表 4 可以看出, 本文算法在所使用的数据集上的平均精确度相较于其他 7 种算法至少提高了 4.5%, 召回率较 YOLOv5x 降低了 3.5%, 但较其余算法提升了至少 3.4%。F_1指标相较于除 YOLOv5x 外的其余算法至少提升 5.99%, 较 YOLOv5x 降低了 0.34%, 但在实时性方面, 检测速度提升了 3 倍, 且高于其余对比算法。以上对比进一步验证了本文算法的优越性。

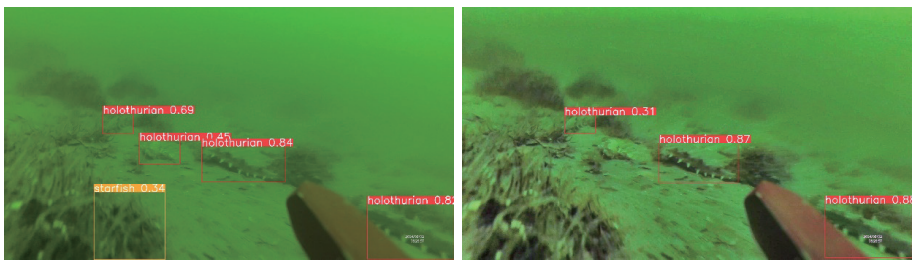
2.4　图像检测效果

随机抽取测试集中 3 张图像对本文算法改进方式进行测试, 结果如图 10 所示。

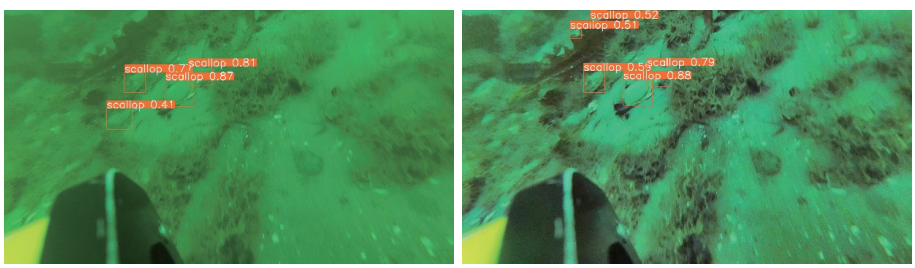
从图 10(a)中可看出, 左图中将石头间的深色空隙误检为海胆, 而右图并未发生误检和漏检; 从图 10(b)中可看出, 左图中将水下杂质误检为海星和海参, 而右图并未发生该类的误检与漏检; 从图 10(c)中可看出, 左图未检测出较远处目标较小的扇贝, 并将有泥沙覆盖的石头误检为扇贝, 而右图完成了正确的检测和识别。可以看出, 本文所改进模型的检测与识别效果优于 YOLOv5s 检测效果, 而且降低了对目标生物的误检率和漏检率。



(a) 图1的检测结果



(b) 图2的检测结果



(c) 图3的检测结果

图 10 YOLOv5s(左)和 YOLOv5s-underwater(右)的检测结果

Fig. 10 Test results of YOLOv5s (left) and YOLOv5S-Underwater (right)

3 结 论

1) 针对图像色彩失真、能见度低、目标与环境区分度较低和形态尺寸差异大等问题，本文提出了YOLOv5s-underwater算法，改进后的算法表现较好。其中，平均精确度为87.6%，召回率为67.4%，检测速度为140.84帧/s，检测识别效果都有所提高。

2) CLAHE预处理对输入的弱光水下图像进行了增强，极大提高了检测精度。

3) SPPR模块提高了模型的纹理及轮廓特征描述能力，并降低了正向计算的时间。

4) ST模块提高了模型的泛化能力。经实验结果分析，改进的网络模型对弱光水下生物目标有更高的检测和识别精度。

5) 特征连接策略增强了模型的多尺度特征融合能力，提高了对小目标的检测能力。

下一步工作将研究基于YOLOv5改进算法的智能水下机器人进行水下生物检测实际应用，以机器人搭载摄像头的方式代替潜水员进行水下生物的自主目视检测。

参考文献（References）

[1] 李宝奇, 黄海宁, 刘纪元, 等. 基于改进SSD的水下光学图像感兴趣目标检测算法研究[J]. 电子与信息学报, 2022, 44(10): 3372-3378.
LI B Q, HUANG H N, LIU J Y, et al. Underwater optical image interested object detection model based on improved SSD[J]. Journal of Electronics & Information Technology, 2022, 44(10): 3372-3378 (in Chinese).

[2] 金盛龙, 迟骋, 李宇, 等. 稀疏驱动自适应线谱增强的水下目标谱熵检测[J]. 声学学报, 2021, 46(6): 1059-1069.
JIN S L, CHI C, LI Y, et al. A supervised learning detection method with pre-processing of sparsity-based adaptive line enhancer[J]. Acta Acustica, 2021, 46(6): 1059-1069(in Chinese).

[3] 徐凤强, 董鹏, 王辉兵, 等. 基于水下机器人的海产品智能检测与自主抓取系统[J]. 北京航空航天大学学报, 2019, 45(12): 2393-2402.
XU F Q, DONG P, WANG H B, et al. Intelligent detection and autonomous capture system of seafood based on underwater robot[J]. Journal of Beijing University of Aeronautics and Astronautics, 2019, 45(12): 2393-2402(in Chinese).

[4] ZHOU J C, ZHANG D H, ZHANG W S. Classical and state-of-the-art approaches for underwater image defogging: A comprehensive survey[J]. Frontiers of Information Technology & Electronic Engineering, 2020, 21(12): 1745-1769.

[5] 董绍江, 刘伟, 蔡巍巍, 等. 基于分层精简双线性注意力网络的鱼类识别[J]. 计算机工程与应用, 2022, 58(5): 186-192.
DONG S J, LIU W, CAI W W, et al. Fish recognition based on hierarchical compact bilinear attention network[J]. Computer Engineering and Applications, 2022, 58(5): 186-192(in Chinese).

[6] 牛浩青, 欧鸥, 饶姗姗, 等. 改进 YOLOv3 的遥感影像小目标检测方法[J]. 计算机工程与应用, 2022, 58(13): 241-248.
NIU H Q, OU O, RAO S S, et al. Small object detection method based on improved YOLOv3 in remote sensing image[J]. Computer Engineering and Applications, 2022, 58(13): 241-248(in Chinese).

[7] WANG X H, ZHU Y G, LI D Y, et al. Underwater target detection based on reinforcement learning and ant colony optimization[J]. Journal of Ocean University of China, 2022, 21(2): 323-330.

[8] OKSUZ K, CAM B C, KALKAN S, et al. Imbalance problems in object detection: A review[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021, 43(10): 3388-3415.

[9] 周清松, 董绍江, 罗家元, 等. 改进 YOLOv3 的桥梁表观病害检测识别[J]. 重庆大学学报, 2022, 45(6): 121-130.
ZHOU Q S, DONG S J, LUO J Y, et al. Bridge apparent disease detection based on improved YOLOv3[J]. Journal of Chongqing University, 2022, 45(6): 121-130(in Chinese).

[10] REDMON J, FARHADI A. YOLOv3: An incremental improvement[EB/OL].(2018-04-06)[2022-05-01].https://arxiv.org/abs/1804.02767.

[11] AHMAD T, CHEN X N, SAQLAIN A S, et al. EDF-SSD: An improved feature fused SSD for object detection[C]//Proceedings of the IEEE 6th International Conference on Cloud Computing and Big Data Analytics. Piscataway: IEEE Press, 2021: 469-473.

[12] ZHANG Z D, TAN M L, LAN Z C, et al. CDNet: A real-time and robust crosswalk detection network on Jetson nano based on YOLOv5[J]. Neural Computing and Applications, 2022, 34(13): 10719-10730.

[13] WALIA I S, KUMAR D, SHARMA K, et al. An integrated approach for monitoring social distancing and face mask detection using stacked ResNet-50 and YOLOv5[J]. Electronics, 2021, 10(23): 2996.

[14] LIU Z, LIN Y, CAO Y, et al. Swin Transformer: Hierarchical vision transformer using shifted windows[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE Press, 2022: 9992-10002.

[15] 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.

Improved YOLOv5s low-light underwater biological target detection algorithm

CHEN Yuliang，DONG Shaojiang*，SUN Shizheng，YAN Kaibo

(School of Mechatronics and Vehicle Engineering，Chongqing Jiaotong University，Chongqing 400074，China)

Abstract：A real-time detection method of low-light underwater biological target based on improved YOLOv5s, known as YOLOv5s-underwater, was proposed to address the issue of low biometric recognition accuracy caused by the significant attenuation of light in water, the complex image environment, and the movement of shooting equipment in the process of underwater optical image target detection. Firstly, to solve the problem of weak underwater light attenuation, the contrast-limited adaptive histogram equalization (CLAHE) algorithm is introduced to preprocess the input image, which solves the problems of color distortion and image roughness. Secondly, the spatial pyramid pooling fast (SPPF) module is proposed to solve the problems of low discrimination and serious feature loss of underwater objects in the complex low-light underwater image environment. Thirdly, a Swin-Transformer module based on the spin window is proposed to improve the generalization ability of the model. Finally, the network model structure is modified to improve the detection ability of small underwater targets. Simulation and experiment prove that the proposed method improves the detection accuracy by 30.7% compared with YOLOv5s. Results from experiments support the method's efficacy.

Keywords：low-light underwater biological targets；YOLOv5s；contrast-limited adaptive histogram equalization；spatial pyramid pooling fast；spin window

Received：2022-05-06；**Accepted**：2022-07-01；**Published Online**：2022-07-14 15：32

URL：link.cnki.net/urlid/11.2625.V.20220713.1906.001

Foundation items：National Natural Science Foundation of China (51775072); Chongqing Science and Technology Innovation Leading Talents Support Program (CSTCCCXLJRC201920); Chongqing University Innovation Research Group (CXQT20019)；Technology Innovation and Application Demonstration Project of Chongqing Beibei Science and Technology Bureau (2020-6); Chongqing Key Laboratory of Urban Rail Transit System Integration and Control Open Fund (CKLURTSIC-KFKT-202007)

* **Corresponding author**. E-mail：dongshaojiang100@163.com

http://bhxb.buaa.edu.cn jbuaa@buaa.edu.cn
DOI: 10.13700/j.bh.1001-5965.2022.0328

基于K-means聚类的多种群麻雀搜索算法

闫少强，刘卫东，杨萍*，吴丰轩，阎哲
(火箭军工程大学 作战保障学院，西安 710025)

摘　　要：为改善麻雀搜索算法（SSA）在单种群搜索中收敛速度过快导致其收敛速度出现冗余，易忽略优质解而陷入局部最优的缺陷，提出一种基于K-means聚类的多种群麻雀搜索算法（KSSA）。将多种群机制引入SSA，减弱单种群的收敛能力，并减小陷入局部最优的概率；采用K-means聚类划分子种群，增加子种群间的差异性，同时使子种群内个体在小范围内专注搜索，提升前期搜索效率；借助加权重心交流策略改善种群间交流的质量，减少自身种群的干扰，同时消减因某一子种群陷入局部最优而导致所有子种群陷入局部最优的风险；引入动态反向学习到警戒者中，增强其反捕食行为，改善因子种群数量增加而带来的收敛速度变慢和收敛精度不足的缺陷。经测试函数仿真实验表明：较SSA等算法，KSSA具有更优的寻优性能。

关 键 词：麻雀搜索算法；优化算法；多种群；K-means聚类；种群交流
中图分类号：TP301.6
文献标志码：A　　**文章编号**：1001-5965(2024)02-0508-11

麻雀搜索算法 (sparrow search algorithm, SSA)是由Xue和Shen基于麻雀种群的觅食行为和反捕食行为在2020年提出的一种新型群体智能优化算法[1]。相较于粒子群(particle swarm optimization, PSO)算法[2]、遗传算法 (genetic algorithm, GA)[3]等经典算法的研究历程，SSA的研究和应用还处于初级阶段，凭借其相较于其他群智能算法收敛速度更快、稳定性更强、寻优精度更高的优点，具有很大的研究潜力和发展前景。

许多学者投入到SSA的研究中，文献[4]提出一种基于佳点集的麻雀搜索算法 (good point set sparrow search algorithm, GSSA)，利用佳点集改善了初始种群的质量，并提升了种群多样性。文献[5]提出一种混沌麻雀搜索算法 (chaos sparrow search algorithm, CSSA)，引入Tent混沌映射，增强了其跳出局部最优的能力。文献将加入者搜索阶段分成全局搜索和局部搜索2个阶段，加入可变螺旋因子增强了其全局搜索能力。

SSA具有收敛速度快的优点，但同样存在搜索速度过快导致其收敛速度冗余和易忽略优质解陷入局部的问题。为此，本文提出一种基于K-means聚类的多种群麻雀搜索算法 (K-means multi-group sparrow search algorithm, KSSA)，通过引入多种群机制减少由于单种群搜索而易陷入局部最优的概率，而后利用K-means聚类划分初始子种群提高前期搜索效率，随后通过加权重心交流策略提升种群间的交流质量，最后借助动态反向学习提高搜索精度和收敛速度。选取PSO算法[2]、鲸鱼优化算法 (whale optimization algorithm, WOA)[6]、灰狼优化算法 (grey wolf optimizer, GWO)[7]、SSA[1]、GSSA[4]、CSSA[5]共6种对比算法在8个测试函数上进行仿真实验，进一步验证KSSA的有效性。

1 麻雀搜索算法

麻雀作为群居鸟类，多活动在有人类居住的地方，性格活泼，聪明且记忆力强，警惕性非常高。在

收稿日期：2022-05-07；录用日期：2022-09-18；网络出版时间：2022-10-18 10:17
网络出版地址：link.cnki.net/urlid/11.2625.V.20221017.1733.012
基金项目：国家自然科学基金 (61703411)
*通信作者. E-mail：yyp_ing@163.com

引用格式：闫少强，刘卫东，杨萍，等. 基于K-means聚类的多种群麻雀搜索算法[J]. 北京航空航天大学学报，2024，50(2)：508-518.
YAN S Q，LIU W D，YANG P，et al. Multi group sparrow search algorithm based on K-means clustering[J]. Journal of Beijing University of Aeronautics and Astronautics，2024，50(2)：508-518 (in Chinese).

麻雀觅食的过程中, 种群个体分工明确, 可以根据对环境的适应程度把种群个体分为发现者和加入者。发现者的环境适应度高, 需要进行广泛搜索发现食物, 引导种群个体获取食物并掌握整个种群的搜索方向。加入者的环境适应度低于发现者, 为了提高自身适应度, 需跟随发现者获取食物。同时, 麻雀种群在觅食过程中势必会遭到天敌等外部自然环境的各种威胁, 为了提高生存概率, 麻雀种群会随机分配一部分个体作为警戒者, 时刻警惕周围环境, 一旦发现威胁便提醒种群逃离。

以麻雀种群在收割完的田地里搜寻遗落谷物为例, 发现者负责在广泛的田地里不断跳跃寻找遗落的谷物, 发现谷物后便停下进食。加入者察觉发现者寻找到了食物, 便径直跳跃到谷物所在地, 与发现者夺取食物, 竞相进食。但由于谷物数量有限, 加入者中并不能保证每只麻雀都能够摆脱饥饿的状态, 所以对于距离谷物所在地较远(即环境适应度较低)的麻雀, 就会放弃与种群争夺食物, 选择飞到其他地方独自觅食。同时, 为了保证种群的安全进食, 麻雀种群会随机在种群外围和内部安排一定数量的麻雀进行放哨侦查, 以便整个种群在应对突发情况时能够及时逃脱。以上便是麻雀种群觅食的整个过程。

设置当前的迭代次数为 t, 最大的迭代次数为 M, 当前第 i 个麻雀的位置为 X_i, 当前第 i 个麻雀在第 j 维中的位置为 $X_{i,j}$。发现者没有发现威胁时 ($R_2<S_\mathrm{T}$)负责引导种群觅食, 进行广泛地搜索; 当种群内个体已经发现了捕食者且发出了警报 ($R_2\geqslant S_\mathrm{T}$), 发现者引导种群前往安全区域的位置, 位置更新描述如下:

$$\boldsymbol{X}_i^{t+1}=\begin{cases}\boldsymbol{X}_i^t\exp\left(\dfrac{-i}{\alpha M}\right) & R_2<S_\mathrm{T}\\ \boldsymbol{X}_i^t+Q\boldsymbol{L} & R_2\geqslant S_\mathrm{T}\end{cases}\tag{1}$$

式中: α 为属于 (0,1] 的随机数; $R_2\in[0,1]$ 表示预警值; $S_\mathrm{T}\in[0.5,1]$ 表示当前环境的安全阈值; Q 为服从正态分布的随机数, 负责控制步长; $\boldsymbol{L}$ 为一个 $1\times d$ 的矩阵, 且元素全为 1, d 代表维度。

加入者为了获取食物, 跟随并监督发现者进行夺取食物($i\leqslant N/2$)或独自觅食($i>N/2$), 故加入者的位置更新描述如下:

$$\boldsymbol{X}_i^{t+1}=\begin{cases}Q\exp\left(\dfrac{\boldsymbol{X}_\mathrm{worst}^t-\boldsymbol{X}_i^t}{i^2}\right) & i>N/2\\ \boldsymbol{X}_\mathrm{P}^{t+1}+\left|\boldsymbol{X}_i^t-\boldsymbol{X}_\mathrm{P}^{t+1}\right|\boldsymbol{A}^+\boldsymbol{L} & \text{其他}\end{cases}\tag{2}$$

式中: $\boldsymbol{X}_\mathrm{worst}$ 表示当前种群的最坏位置; $\boldsymbol{X}_\mathrm{P}$ 为发现者当前所占据的最优位置; $\boldsymbol{A}$ 为一个负责控制方向的 $1\times d$ 的矩阵, 元素仅为 1 或−1, 且 $\boldsymbol{A}^+=\boldsymbol{A}^\mathrm{T}(\boldsymbol{A}\boldsymbol{A}^\mathrm{T})^{-1}$。

当意识到危险, 麻雀种群会做出反捕食行为, 当 $f_i\neq f_g$ 时, 表示当前处在种群边缘的麻雀意识到了危险, 需要向种群中心靠拢降低被捕食的风险。当时 $f_i=f_g$ 时, 这表明处于种群中心的麻雀个体意识到了危险, 需要逃离当前位置。其数学表达式如下:

$$\boldsymbol{X}_i^{t+1}=\begin{cases}\boldsymbol{X}_\mathrm{best}^t+\beta\left|\boldsymbol{X}_i^t-\boldsymbol{X}_\mathrm{best}^t\right| & f_i\neq f_\mathrm{g}\\ \boldsymbol{X}_i^t+K\left(\dfrac{\left|\boldsymbol{X}_i^t-\boldsymbol{X}_\mathrm{worst}^t\right|}{(f_i-f_\mathrm{w})+\varepsilon}\right) & f_i=f_\mathrm{g}\end{cases}\tag{3}$$

式中: $\boldsymbol{X}_\mathrm{best}$ 表示当前种群的最优位置; β 为服从标准正态分布的随机数负责控制步长; K 为 [−1,1] 的随机数, 用于控制麻雀移动的方向同时也可以控制移动步长; f_i、f_g 和 f_w 分别为第 i 个个体的适应度值、当前种群的最优和最差适应度值。为了防止分母为 0, ε 取一个极小的正实数。

2　基于 K-means 聚类的改进算法

2.1　多种群搜索机制

目前 SSA 采取单种群模式进行搜索, 所有的搜索过程都限制在一个种群内部, 一旦该种群陷入局部最优且不能逃脱局部最优陷阱时, 会降低整个算法的寻优性能。对于 SSA 来说, 由于其拥有“发现者-加入者”模型而导致收敛速度极快, 加入者会向发现者的位置聚拢争夺食物, 大量的个体同时向发现者当前占据的最优位置收敛, 会增加陷入局部最优的风险。并且对比单种群和多种群, 单种群内的个体数量是多种群的多倍, 会导致单种群麻雀算法收敛能力的冗余, 出现早熟停滞的现象。

传统的单种群搜索的求解效率较低, 在局部最优陷阱较多或多峰的搜索空间内易陷入局部最优的陷阱, 导致算法整体性能降低。而相较于单种群搜索, 多种群搜索具有以下优点[8-11]:

1) 可以将各个种群设置不同参数, 拿差分进化 (differential evolution, DE) 算法举例, 例如交叉概率与变异概率的设定, 可以使各个种群向着不同方向进行进化, 进而全面增强搜索能力。

2) 各个子种群之间可以进行信息的交流共享, 例如 DE 算法可以设置移民算子, 把最优个体引入到其他子种群之中, 实现当前子种群向其他子种群的学习, 有助于算法的收敛。

3) 种群不容易陷入局部最优, 不同的子种群进行各自地搜索, 搜索的方式更加灵活, 当一个种群陷入局部最优时, 可以借助与其他子种群的交流来逃脱局部最优的吸引。

多种群搜索机制是将种群个体 N 相对平均的随机分为 m 个子种群, 则每个子种群的个体为 round(N/m)=n, round 为向下取整数运算。如果

mod(N, m)≠0, mod 为求余运算, 则将余数 C 均匀分到前 C 个子种群。则搜索过程中, 第 t 次迭代中第 k 个子种群的第 i 个体的第 j 维表示为 $X^t_{k,i,j}$, 此时 $i \leqslant N_k$, N_k 为第 k 个子种群的个体数量, 且 $\sum_{k=1}^{m} N_k = N$, 如图 1 所示。

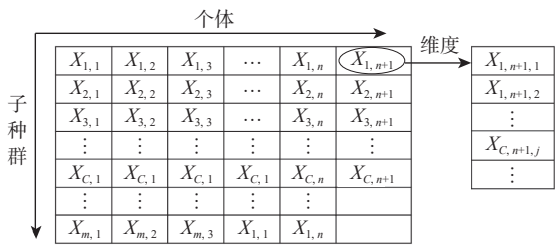


图 1 多种群的个体分布

Fig. 1 Individual distribution in multiple populations

2.2 K-means 聚类

目前的子种群划分主要是依靠上述随机性的均匀分配, 该方法划分的子种群进行搜索时可以理解为不同的子种群在同一空间范围内进行并行搜索, 其搜索范围较广泛。与单种群搜索相比尽管有更大概率防止其陷入局部最优, 但由于子种群中个体数量减少, 导致其前期搜索效率较低, 仍存在易忽略优质解的问题。为此, 本文提出一种 K-means 聚类划分子种群的方法,增加子种群间的差异性和前期搜索的效率。

K-means 聚类秉持“组内距离最近, 组间距离最远”的原则, 将一组 N 个样本的特征矩阵划分为 K 个无交集的簇, 直观上来看, 簇是一组一组聚集在一起的数据, 在一个簇中的数据就认为是同一类。簇就是聚类后的结果, 簇中所有数据的均值通常被称为这个簇的质心。其聚类的基本原理如下:

步骤 1 随机初始化 K 个聚类中心点, 并计算数据中每个点到 K 个聚类中心点的距离。

步骤 2 将每个数据点分到距离聚类中心点最近的聚类中心中。

步骤 3 针对每个类别重新计算聚类中心。

步骤 4 重复上面的步骤 2、步骤 3, 直到达到预先设置的停止条件(迭代次数、最小误差变化等)。

类推到本文算法中, 则是指将一组个体为 N 的位置矩阵划分为 K 个无交集的子种群。初始种群被 K-means 聚类后, 可以根据个体间距离相近程度被划分为同一簇, 也就是同一子种群。

由于前期不同子种群间距离较远, 同一子种群的个体可以在较小的范围内细致地搜索, 寻找优质解, 即各子种群在各自附近进行专注地搜索, 有助于前期寻找到优质解; 当某一子种群寻找到优质解后, 种群交流机制进一步帮助其他子种群进行收敛, 有助于算法效率的提升。其 2 种方法划分后的初始子种群搜索对比如图 2 所示, 值得注意的是虚线并不是搜索范围的划分, 仅是为了方便理解设置如此。

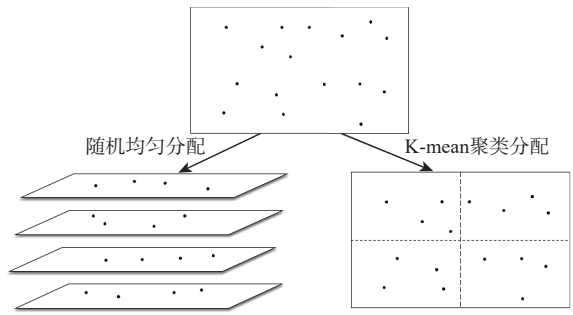


图 2 划分初始子种群原则

Fig. 2 Principle of dividing initial sub population

设初始种群个体为 100, 边界为 $[-1,1]\times \boldsymbol{E}_{1\times d}$, $\boldsymbol{E}_{1\times d}$ 为 $1\times d$ 元素全为 1 的矩阵, d 为维度。图 3 为在 2 维和 3 维下 K-means 聚类后初始子种群的示意图。

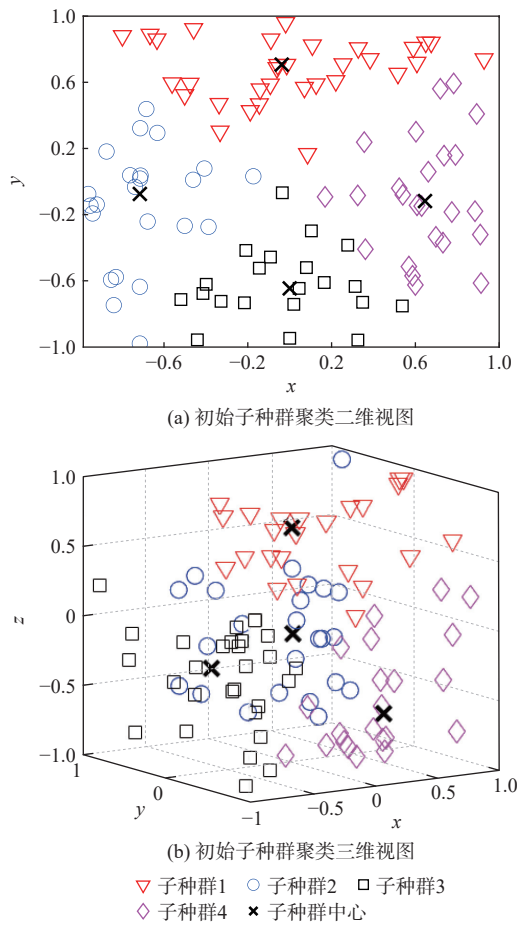


图 3 基于 K-means 聚类划分的初始子种群

Fig. 3 Initial sub-population based on K-means clustering

2.3 加权重心交流策略

一般的种群交流机制是将每个子种群中最优个体进行比较后[12], 选出其中最优的个体 X_{best}, 并将每个子种群中的最优个体或最差个体替换为该个体, 进而改善种群质量。但是一旦选出的最

优个体 X_{best} 为局部最优, 则会导致所有子种群陷入局部最优。还有另外一种比较常见的种群交流机制为随机重组策略[13], 就是指每隔一定迭代次数, 整个种群会进行随机重组, 种群中的每个个体被重新划分到新的子种群, 并开始新的搜索。该策略对种群之间的交流有极大的提升, 但是存在概率引起一种混乱的状态, 比如已经陷入局部最优的子种群个体可能导致更多新的子种群也陷入局部最优, 质量较差的个体被分到同一子种群导致该子种群质量偏低, 因此, 该类策略存在不稳定的缺陷。

为此本文提出一种加权重心交流策略, 即把每个子种群的最差位置替换为其他所有子种群中最优位置的加权重心, 权重为每个子种群最优位置的适应度占所有子种群最优位置的总适应度的比值。值得注意的是, 被替换最差位置的子种群并不参与该重心的选取。该策略作用如下:

1) 最差位置被替换为其他子种群的最优位置的加权重心, 首先最差位置被替换, 如果该子种群质量较差, 则有助于改善该种群; 如果该子种群质量较好, 则有助于帮助种群检验是否陷入局部最优。

2) 选取替换为不包括该子种群的其他所有子种群中最优位置的加权重心, 可以减少自身种群的干扰, 并且消减了某一子种群陷入局部最优而导致所有子种群陷入局部最优的风险, 同时可以防止向最优解收敛速度过快而引起其他优质解被忽略的情况。

以 4 个子种群为例, G_1、G_2、G_3、G_4 分别为 4 个种群, w_1、w_2、w_3、w_4 分别为 4 个子种群的最差个体经过加权重心交流后的位置, 黑色圆圈为全局最优, 如图 4 所示。在图 4(a)中, 由于 G_1、G_4 距离全局最优较近, 其权值较高, 所以 w_2、w_3 位置更靠近 G_1、G_4。在图 4(b)中, 由于 G_2 距离全局最优较近, 其权值较高, 所以 w_3 位置更靠近 G_2。2 种情况都会使加权重心交流后的个体更靠近全局最优, 但图 4(b)中的 w_2 则是远离全局最优, 这样有助于检验是否为局部最优。

则第 k 个子种群的最差位置公式更新如下:

$$\boldsymbol{X}_{\text{worst}}^{k}=\left(\sum_{j=1}^{n}\frac{f_{\text{g}}^{j}}{\sum_{i=1}^{n}f_{\text{g}}^{i}-f_{\text{g}}^{k}}\right)\boldsymbol{X}_{\text{best}}^{j}-\frac{f_{\text{g}}^{j}}{\sum_{i=1}^{n}(f_{\text{g}}^{i}-f_{\text{g}}^{k})}\cdot\boldsymbol{X}_{\text{best}}^{k} \tag{4}$$

式中: 上标表示为第 i, j, k 个子种群, 式 (4) 可进一步简化为

$$\boldsymbol{X}_{\text{worst}}^{k}=\sum_{j=1,j\neq k}^{n}\frac{f_{\text{g}}^{j}}{\sum_{i=1,i\neq k}^{n}f_{\text{g}}^{i}}\cdot\boldsymbol{X}_{\text{best}}^{j} \tag{5}$$

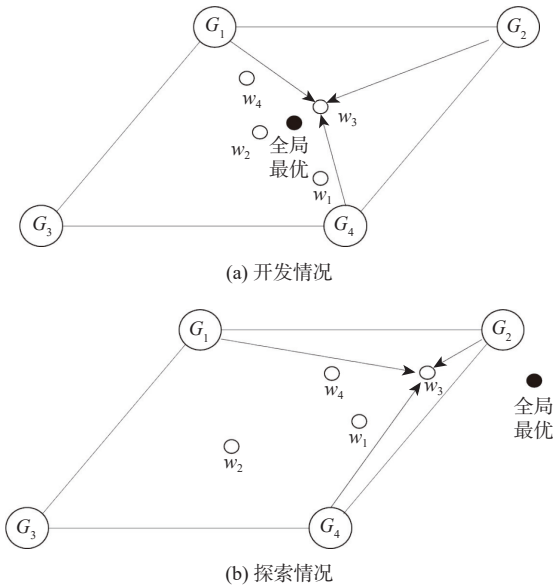


图 4　加权重心交流策略原理

Fig. 4　Principle of weighted center of gravity communication strategy

2.4　动态反向学习

由于多种群机制将种群分为了多个子种群, 子种群中的个体数量相较于原种群个体数量较少, 导致了收敛速度和寻优精度的降低, 这也是多种群机制的缺点。而反向学习策略有助于加快收敛速度和提升寻优精度, 且反向学习之后的解能更接近最优解。一般的反向学习只能在固定的空间搜索寻找最优解[14-15], 而动态反向学习具有比一般反向学习更好的寻优能力, 减少维度之间的干扰, 能在动态的空间内不断收敛到最优解。同时警戒者具有反捕食行为 (见式 (3)), 来帮助种群跳出局部最优和加速收敛, 为此本文将动态反向学习加入到警戒者的位置更新中, 增强反捕食行为, 弥补因子种群中个体数量减少而带来的收敛速度变慢和收敛精度不足的缺陷。

反向学习原理如下:

$$\boldsymbol{X}^{*}=\boldsymbol{a}+\boldsymbol{b}-\boldsymbol{X} \tag{6}$$

式中: $\boldsymbol{X}$ 为原始位置; $\boldsymbol{X}^{*}$ 为 $\boldsymbol{X}$ 在动态反向学习下的反向解; $\boldsymbol{a}$ 和 $\boldsymbol{b}$ 分别为下边界和上边界, 此时为一般的反向学习策略。

本文将反向学习扩展到每个维度, 对每个维度进行反向学习, 公式扩展如下:

$$X^*_{:,j}=a_j+b_j-X_{:,j} \tag{7}$$

式中：j为维度；a_j为第j维的下边界；b_j为第j维的上边界。

同时本文采取动态边界，即

$$a_j=\min\left(X_{:,j}\right)\ b_j=\max\left(X_{:,j}\right) \tag{8}$$

式中：$\min\left(X_{:,j}\right)$为所有个体中第j维的最小值；$\max\left(X_{:,j}\right)$为所有个体中第j维的最大值。

图5为第i个个体的动态反向学习示意图，黑色圆圈为全局最优解，方块分别为当前种群中每个维度中最大值、最小值，即每个维度的上边界、下边界。$|\boldsymbol{a}+\boldsymbol{b}|/2$为每一维边界的中点，值得注意的是实际中并不是一条直线，这里抽象为一条直线。可以看出动态反向学习同一般反向学习一样可以逼近最优解，能够加速收敛而且进一步排除了维度间的干扰，使得每个维度之间互不影响。

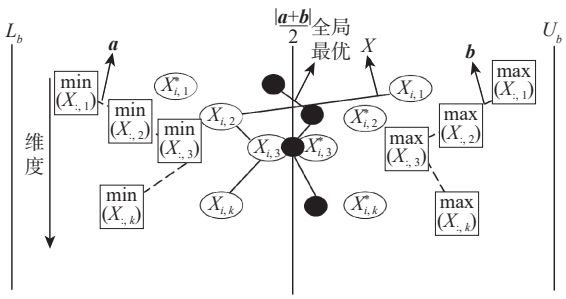


图5 动态反向学习原理

Fig. 5 Dynamic reverse learning principle

2.5 KSSA流程

KSSA采用多种群搜索机制，在种群初始化后，首先，采用K-means聚类划分子种群，增加子种群间的差异性，同时使子种群内个体在小范围内专注搜索，提升前期搜索效率；然后，在警戒者位置更新后，进行动态反向学习，加速子种群的收敛速度的同时具有更加细致的搜索精度；最后，在所有子种群更新完后采用加权重心交流策略，进行种群间的交流和优质解的传递，帮助改善种群的质量和跳出局部最优。

其具体实现步骤如下：

步骤1 设置初始化参数，包括最大迭代次数M，种群规模N，子种群数量m，发现者比例P_D，警戒者比例S_D，目标函数维度d，上下界u_b、l_b，最大迭代次数T，安全阈值S_T。

步骤2 初始化种群，并利用K-means聚类将种群划分为m个子种群，计算每个子种群的数量N_k。

步骤3 设置k=1。

步骤4 计算第k个子种群中每个个体的适应度值f_i，对适应度排序，选出当前最优适应度值f_g及其所对应的位置X_{best}，当前最差适应度值f_w及其所对应的位置X_{worst}。

步骤5 选取第k个子种群适应度值前P_DN_k的个体作为发现者，根据式(1)更新其位置，并确定发现者当前所占据的最优位置X_P。

步骤6 选取剩下的个体作为加入者，根据式(2)更新其位置。

步骤7 从第k个子种群中随机选取S_DN_k的个体作为警戒者，根据式(3)和式(7)更新其位置。

步骤8 计算适应度值，更新麻雀位置和f_g、X_{best}、f_w、X_{worst}，k=k+1。

步骤9 判断是否$k>m$，若是则根据式(5)更新每个子种群的最差位置，否则重复步骤4～步骤8。

步骤10 判断是否满足输出条件(达到最大迭代次数)，满足则输出结果，否则重复步骤3～步骤9。

2.6 时间复杂度分析

由于种群规模、迭代次数和维度分别为N，M和d，所以运行SSA的时间复杂度为$O(N\times M\times d)$，如2.5节算法步骤所示，KSSA较SSA主要多出以下操作：

1）种群个体数量N被K-means聚类分为m个子种群，这个任务的时间消耗为$O(N)$。

2）KSSA按m个子种群进行位置更新，会增加个体位置更新前排序次数，由于总体个数不变，所以个体更新的时间消耗和原SSA一样。

3）动态反向学习策略仅是对警戒者位置进行更新，其增加的时间消耗为$O(S_D\times N\times M\times d)$，显然$S_D$<1。

4）加权重心交流策略仅对m个子种群的最差个体进行更新，其增加的时间消耗为$O(m\times d)$。

总的来说，KSSA增加了时间的耗费，但并没有增加算法的时间复杂度，所以KSSA的时间复杂度仍是$O(N\times M\times d)$。

3 仿真实验和结果分析

3.1 测试函数和算法参数

本文选取了PSO、WOA、GWO、SSA、GSSA和CSSA作为对比算法，算法的参数设定如表1所示，为公平起见设定实验中每种算法的种群数量N为30，最大迭代次数M为500。本文实验环境均为AMD Ryzen 7 4800 U with Radeon Graphics

表 1　算法参数

Table 1　Algorithm parameter

算法	参数
PSO[2]	c_1=2,c_2=2,W_{mix}=0.2,W_{max}=0.9
WOA[6]	b=1
GWO[7]	a=(2→0)
SSA[1]	S_T=0.8,P_D=0.2,S_D=0.2
GSSA[4]	S_T=0.8,P_D=0.2,S_D=0.2,k=12 000
CSSA[5]	S_T=0.8,P_D=0.2,S_D=0.2
KSSA	m=4,S_T=0.8,P_D=0.2,S_D=0.2

@1.80 GHZ, 内存 16 GB, 装有 MATLAB 2020a 的 Windows 10 系统。本文选取 8 种基准测试函数进行仿真实验以验证 KPSSA 的性能，其中 F_1-F_4 为高维单峰函数，F_5-F_8 为高维多峰函数，函数信息如表 2 所示。

3.2　消融实验

消融实验是指通过设立对照组，去除某个模块的作用，来证明该模块的必要性，如果消融实验后得到结果不好或性能大幅下降，说明该模块起到了作用。因此，为了更好的验证每个策略或机制发挥的作用，本文 SSA、加入多种群机制的麻雀算法（SSA1）、加入 K-mean 聚类的多种群麻雀算法（SSA2）、在 SSA2 基础上加入加权重心交流机制的麻雀算法（SSA3）、在 SSA3 基础上加入动态反向学习的麻雀算法（即 KSSA）分别在 F_1（单峰）和 F_5（多峰）中进行测试仿真，如图 6 所示。其前后 2 种算法可分别作为对照组，检验新加入的机制或策略的必要性，可以看出，SSA、SSA1、SSA2、SSA3、KSSA 的性能呈现一种递增的趋势，说明每一种机制或策略的加入都对前一个算法性能有一定的提升。

图 7 为在前期迭代第 10 次时，SSA 和 KSSA 在 F_5 中的位置变化，图 7(c)中不同的形状代表不同的子种群。经过 10 次迭代，SSA 并没有表现出较好的搜索能力，且距离全局最优位置较远，陷入了局部最优陷阱；而 KSSA 经过 4 个子种群细致地搜索，种群率先发现较优位置且向其靠拢，并通过加权重心交流机制不断引导其他子种群向该方向聚拢，表明 KSSA 在前期具有较强的搜索能力且不易陷入局部最优。

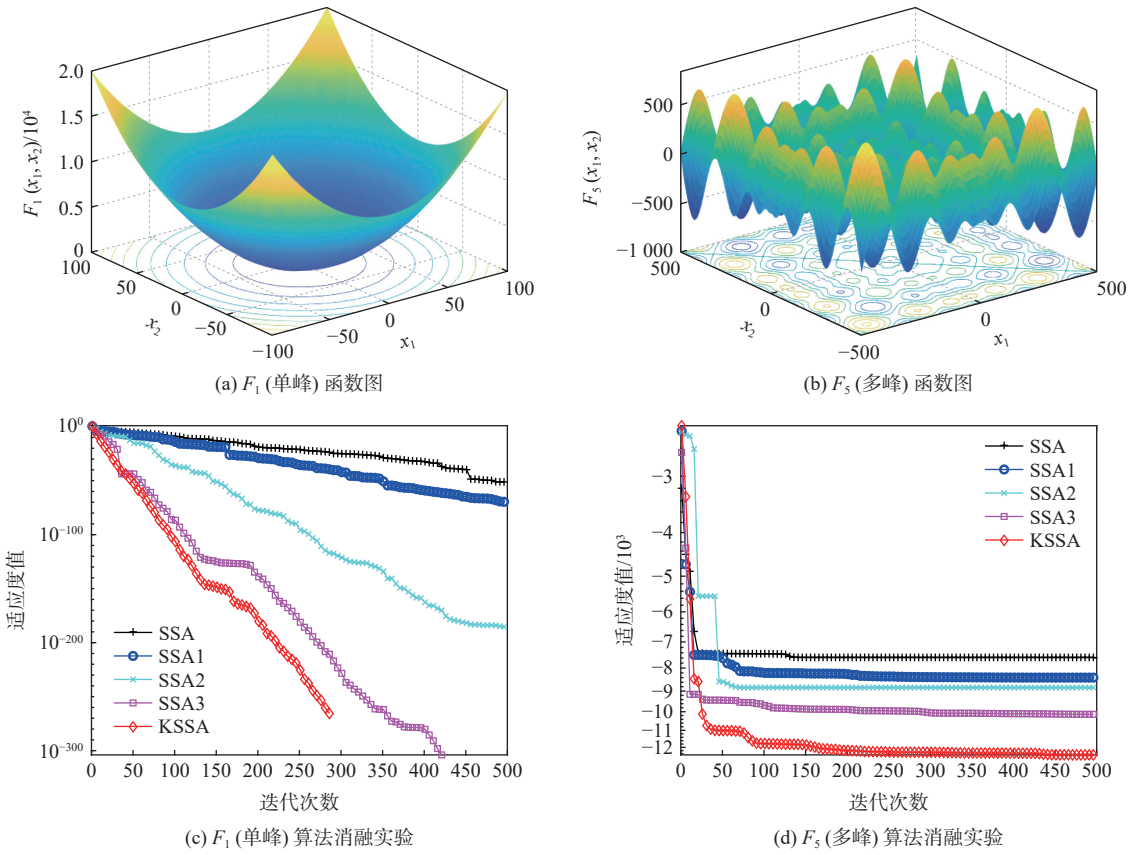
3.3　与其他算法对比

表 3 为上述 7 种算法独立运行 30 次后的结果，“I/A”指“指标/算法”，其中最优值、平均值、标准差这 3 项指标，每项指标中最好的结果进行了加粗处理。P 是指对 30 次运算结果进行 Wilcoxon 秩和检验的结果，当 P<0.05 时，可以认为 KSSA 与对比算法具有显著性差异。

表 2　测试函数

Table 2　Test function

类型	函数	维度	搜索区域	理论最优值
单峰	$F_1(x)=\sum_{i=1}^{n}x_i^2$	30	[−100,100]	0
	$F_2(x)=\sum_{i=1}^{n}\left(\sum_{j=1}^{i}x_j^2\right)$	30	[−100,100]	0
	$F_3(x)=\sum_{i=1}^{n-1}\left[100(x_{i+1}-x_i^2)^2+(x_i-1)^2\right]$	30	[−100,100]	0
	$F_4(x)=\sum_{i=1}^{n}([x_i+0.5])^2$	30	[−100,100]	0
多峰	$F_5(x)=\sum_{i=1}^{n}-x_i\sin\left(\sqrt{\lvert x_i\rvert}\right)$	30	[−500,500]	−418.98 n
	$F_6(x)=\sum_{i=1}^{n}\left[x_i^2-10\cos(2\pi x_i)+10\right]$	30	[−5.12,5.12]	0
	$F_7(x)=\frac{\pi}{n}\left\{10\sin(\pi y_1)+\sum_{i=1}^{n-1}(y_i-1)^2\left[1+10\sin^2(\pi y_{i+1})\right]+(y_n-1)^2\right\}+\sum_{i=1}^{n}u(x_i,10,100,4)\quad y_i=1+\frac{x_i+1}{4}$	30	[−50,50]	0
	$F_8(x)=0.1\left\{\sin^2(3\pi x_1)+\sum_{i=1}^{n}(x_i-1)^2\left[1+\sin^2(3\pi x_i+1)\right]+(x_n-1)^2\left[1+\sin^2(2\pi x_n)\right]\right\}+\sum_{i=1}^{n}u(x_i,5,100,4)$	30	[−50,50]	0

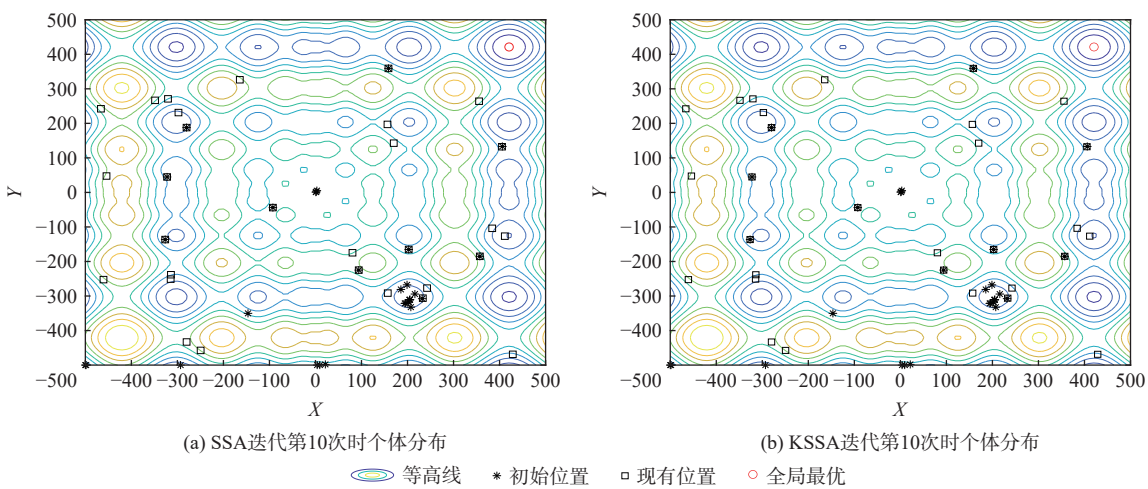


(a) F_1 (单峰) 函数图　(b) F_5 (多峰) 函数图

(c) F_1 (单峰) 算法消融实验　(d) F_5 (多峰) 算法消融实验

图 6　各改进策略对比

Fig. 6　Comparison of improvement strategies



(a) SSA迭代第10次时个体分布　(b) KSSA迭代第10次时个体分布

图 7　等高线位置分布图

Fig. 7　Contour line position distribution map

从表 3 可以看出，SSA 及其改进算法的寻优性能是强于 PSO、WOA、GWO 的，KSSA 仅在 F_4 中表现稍逊色于 SSA、GSSA、CSSA，但仅是平均值和标准差之间一个数量级的差别，且和 SSA 一样寻找了理论最优解。在其他函数中，KSSA 均排名第 1 且在多峰测试函数 F_6、F_7 中，其寻优精度的数量级远超 SSA 及其改进算法，表明了 KSSA 在多峰函数中的优越性。

图 8 为每个算法的收敛情况，在高维单峰函数中，KSSA 的线性收敛速度并不快，但是由于其优越的搜索能力，更不易陷入局部最优而进入停滞状态，甚至可以直接寻找到最优值，结束收敛；在高维多峰函数中，KSSA 的优势更加突显，并没有由于多峰的吸引而陷入局部最优，有很强的摆脱局部最优吸引的能力。

为更加直观地观察每个算法在所有函数下的

表 3　测试函数结果对比

Table 3　Comparison of test function results

算法	最优值							
	F_1	F_2	F_3	F_4	F_5	F_6	F_7	F_8
PSO	7.37×10^{-6}	4.50×10^{1}	1.76×10^{1}	1.10×10^{-5}	-7.99×10^{3}	3.23×10^{1}	4.47×10^{-8}	5.51×10^{-6}
WOA	3.80×10^{-84}	1.11×10^{4}	2.69×10^{1}	9.79×10^{-2}	-1.26×10^{4}	**0**	8.75×10^{-3}	6.90×10^{-2}
GWO	7.48×10^{-28}	5.52×10^{-8}	2.59×10^{1}	2.46×10^{-1}	-7.45×10^{3}	**0**	8.04×10^{-3}	1.97×10^{-1}
SSA	**0**	**0**	**0**	**0**	$\mathbf{-1.26\times10^{4}}$	**0**	1.57×10^{-32}	1.35×10^{-32}
GSSA	**0**	**0**	6.05×10^{-10}	1.36×10^{-12}	$\mathbf{-1.26\times10^{4}}$	**0**	2.24×10^{-13}	7.06×10^{-13}
CSSA	**0**	**0**	6.05×10^{-10}	1.36×10^{-12}	$\mathbf{-1.26\times10^{4}}$	**0**	2.24×10^{-13}	7.06×10^{-13}
KSSA	**0**	**0**	**0**	**0**	$\mathbf{-1.26\times10^{4}}$	**0**	$\mathbf{1.57\times10^{-32}}$	$\mathbf{1.35\times10^{-32}}$
算法	平均值							
	F_1	F_2	F_3	F_4	F_5	F_6	F_7	F_8
PSO	2.35×10^{-4}	9.21×10^{1}	8.62×10^{1}	2.62×10^{-4}	-5.13×10^{3}	5.69×10^{1}	6.91×10^{-3}	7.01×10^{-3}
WOA	9.85×10^{-75}	4.22×10^{4}	2.80×10^{1}	4.04×10^{-1}	-1.06×10^{4}	3.79×10^{-15}	2.50×10^{-2}	5.92×10^{-1}
GWO	3.17×10^{-26}	4.52×10^{-5}	2.69×10^{1}	5.90×10^{-1}	-5.45×10^{3}	8.49×10^{-1}	3.92×10^{-2}	5.58×10^{-1}
SSA	1.55×10^{-73}	3.28×10^{-61}	1.36×10^{-5}	1.10×10^{-8}	-9.06×10^{3}	**0**	1.30×10^{-9}	2.71×10^{-8}
GSSA	**0**	**0**	1.09×10^{-4}	$\mathbf{3.51\times10^{-9}}$	-1.11×10^{4}	**0**	7.55×10^{-10}	3.92×10^{-9}
CSSA	**0**	**0**	1.09×10^{-4}	3.51×10^{-9}	-1.11×10^{4}	**0**	7.55×10^{-10}	3.92×10^{-9}
KSSA	**0**	**0**	$\mathbf{4.05\times10^{-7}}$	4.45×10^{-8}	$\mathbf{-1.20\times10^{4}}$	**0**	$\mathbf{3.79\times10^{-27}}$	$\mathbf{1.35\times10^{-32}}$
算法	标准差							
	F_1	F_2	F_3	F_4	F_5	F_6	F_7	F_8
PSO	4.11×10^{-4}	4.13×10^{1}	5.72×10^{1}	6.59×10^{-4}	1.41×10^{3}	1.50×10^{1}	3.79×10^{-2}	9.64×10^{-3}
WOA	2.78×10^{-74}	1.64×10^{4}	5.38×10^{-1}	1.98×10^{-1}	1.85×10^{3}	2.08×10^{-14}	1.60×10^{-2}	3.52×10^{-1}
GWO	5.04×10^{-26}	9.41×10^{-5}	7.74×10^{-1}	3.16×10^{-1}	1.02×10^{3}	2.80×10^{0}	1.90×10^{-2}	2.09×10^{-1}
SSA	8.48×10^{-73}	1.49×10^{-60}	5.86×10^{-5}	2.34×10^{-8}	1.66×10^{3}	**0**	3.85×10^{-9}	1.14×10^{-7}
GSSA	**0**	**0**	2.58×10^{-4}	$\mathbf{9.68\times10^{-9}}$	1.02×10^{3}	**0**	1.56×10^{-9}	6.94×10^{-9}
CSSA	**0**	**0**	2.58×10^{-4}	9.68×10^{-9}	1.02×10^{3}	**0**	1.56×10^{-9}	6.94×10^{-9}
KSSA	**0**	**0**	$\mathbf{1.20\times10^{-6}}$	2.29×10^{-7}	$\mathbf{7.17\times10^{2}}$	**0**	$\mathbf{2.08\times10^{-26}}$	$\mathbf{5.57\times10^{-48}}$
算法	P							
	F_1	F_2	F_3	F_4	F_5	F_6	F_7	F_8
PSO	1.21×10^{-12}	1.21×10^{-12}	2.63×10^{-11}	1.79×10^{-11}	2.92×10^{-11}	1.21×10^{-12}	1.72×10^{-12}	1.21×10^{-12}
WOA	1.21×10^{-12}	1.21×10^{-12}	2.63×10^{-11}	1.79×10^{-11}	1.12×10^{-2}	3.34×10^{-1}	1.72×10^{-12}	1.21×10^{-12}
GWO	1.21×10^{-12}	1.21×10^{-12}	2.63×10^{-11}	1.79×10^{-11}	2.92×10^{-11}	4.39×10^{-12}	1.72×10^{-12}	1.21×10^{-12}
SSA	6.25×10^{-10}	1.66×10^{-11}	7.53×10^{-4}	3.47×10^{-3}	5.87×10^{-8}	Na	6.93×10^{-12}	4.57×10^{-12}
GSSA	Na	Na	1.31×10^{-1}	6.27×10^{-4}	2.30×10^{-10}	Na	2.67×10^{-11}	4.57×10^{-12}
CSSA	Na	Na	2.86×10^{-9}	4.28×10^{-4}	1.75×10^{-3}	Na	1.72×10^{-12}	1.21×10^{-12}
KSSA								

注: Na表示本文算法与其他算法相比几乎相同。

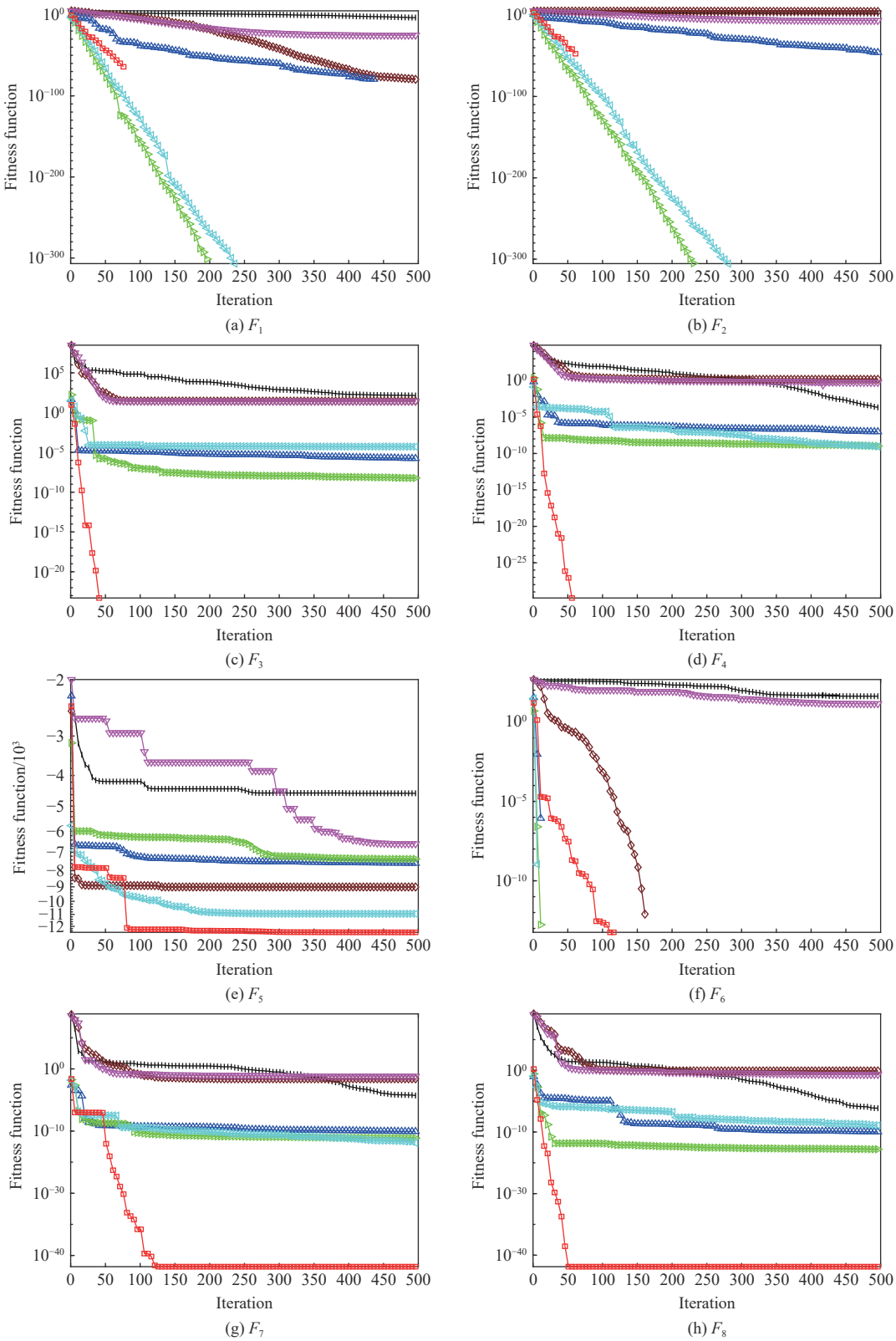


图 8 测试函数平均收敛曲线

Fig. 8 Average convergence curve of test function

效果，采用雷达图对排序结果进行绘图，如图 9 所示。图中正方形为 KSSA，其所围面积最小，说明 KSSA 整体上具有最好的寻优性能。综上来看，KSSA 在对比算法中综合排名第 1，具有较强的搜索能力，其次是 GSSA、CSSA、SSA。特别是在高维多峰的情况下，KSSA 表现突出。

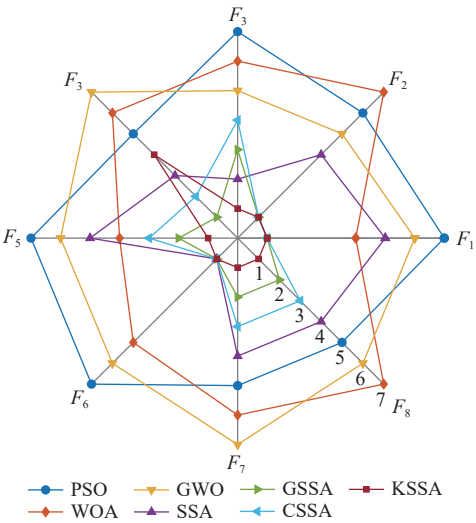


图 9　测试函数排序雷达图

Fig. 9　Radar diagram of test function sorting

4　结　论

1）KSSA 缓解了过快的收敛速度。利用 K-means 聚类划分初始子种群，借助多种群的机制减弱单种群的收敛能力，缓解了 SSA 单种群收敛速度过快导致收敛速度出现冗余，进而过早收陷入局部最优的缺点。

2）KSSA 加速了种群交流。通过加权重心交流策略，促进了种群间的个体交流同时提升了种群间的交流质量。

3）KSSA 算法搜索能力得到提升。引入动态反向学习策略，同时搜索方式更加灵活，使得更易跳出局部最优。

实验表明，改进的算法具有比 SSA 等算法更优的性能。下一步计划将 K-means 聚类与其他聚类算法进行对比，并检验在不同子种群设置不同参数的情况下 KSSA 的性能变化。

参考文献（References）

[1] XUE J K, SHEN B. A novel swarm intelligence optimization approach: Sparrow search algorithm[J]. Systems Science & Control Engineering an Open Access Journal, 2020, 8(1): 22-34.

[2] KENNEDY J, EBERHART J. Particle swarm optimization[C]//Proceedings of ICNN'95-International Conference on Neural Networks. Piscataway: IEEE Press, 2022: 1942-1948.

[3] WHITLEY D. A genetic algorithm tutorial[J]. Statistics and Computing, 1994, 4(2): 65-85.

[4] 闫少强, 杨萍, 朱东林, 等. 基于佳点集的改进麻雀搜索算法[J]. 北京航空航天大学学报, 2023, 49(10): 2790-2798.
YAN S Q, YANG P, ZHU D L, et al. Improved sparrow search algorithm based on good point set[J]. Journal of Beijing University of Aeronautics and Astronautics, 2023, 49(10): 2790-2798 (in Chinese).

[5] 吕鑫, 慕晓冬, 张钧, 等. 混沌麻雀搜索优化算法[J]. 北京航空航天大学学报, 2021, 47(8): 1712-1720.
LYU X, MU X D, ZHANG J, et al. Chaos sparrow search optimization algorithm[J]. Journal of Beijing University of Aeronautics and Astronautics, 2021, 47(8): 1712-1720 (in Chinese).

[6] NASIRI J, KHIYABANI F M. A whale optimization algorithm (WOA) approach for clustering[J]. Cogent Mathematics & Statistics, 2018, 5(1): 1483565.

[7] MIRJALILI S, MIRJALILI S M, LEWIS A. Grey wolf optimizer[J]. Advances in Engineering Software, 2014, 69: 46-61.

[8] ZHANG X X, CHEN W R, DAI C H, et al. Dynamic multi-group self-adaptive differential evolution algorithm for reactive power optimization[J]. International Journal of Electrical Power & Energy Systems, 2010, 32(5): 351-357.

[9] OUYANG J, YAN G R. A multi-group ant colony system algorithm for TSP[C]//Proceedings of 2004 International Conference on Machine Learning and Cybernetics. Piscataway: IEEE Press, 2005: 117-121.

[10] ZANDIEH M, KARIMI N. An adaptive multi-population genetic algorithm to solve the multi-objective group scheduling problem in hybrid flexible flowshop with sequence-dependent setup times[J]. Journal of Intelligent Manufacturing, 2011, 22(6): 979-989.

[11] MARINAKIS Y, MARINAKI M. A hybrid multi-swarm particle swarm optimization algorithm for the probabilistic traveling salesman problem[J]. Computers & Operations Research, 2010, 37(3): 432-442.

[12] WANG Y, CAI Z X. A hybrid multi-swarm particle swarm optimization to solve constrained optimization problems[J]. Frontiers of Computer Science in China, 2009, 3(1): 38-52.

[13] LIANG J J, SUGANTHAN P N. Dynamic multi-swarm particle swarm optimizer[C]//Proceedings 2005 IEEE Swarm Intelligence Symposium. Piscataway: IEEE Press, 2005: 124-129.

[14] XIA X W, LIU J N, LI Y X. Particle swarm optimization algorithm with reverse-learning and local-learning behavior[J]. Journal of Software, 2014, 9(2): 350-357.

[15] HE J, PENG Z, CUI D, et al. A teaching and learning optimization algorithm based on multi-inverse learning[J]. Engineering Science and Technology, 2019, 51(6): 159-167.

Multi group sparrow search algorithm based on K-means clustering

YAN Shaoqiang，LIU Weidong，YANG Ping*，WU Fengxuan，YAN Zhe

(School of Combat Support，Rocket Military Engineering University，Xi'an 710025，China)

Abstract：A K-means multi-group sparrow search algorithm (KSSA) based on K-means clustering is proposed in order to improve the convergence speed of the sparrow search algorithm (SSA) in single population search, which causes redundancy in its convergence speed and makes it simple to ignore the flaw that the high-quality solution falls into local optimization. Firstly, the multi-population mechanism is introduced into SSA to weaken the convergence ability of a single population and reduce the probability of falling into local optimization. Secondly, in order to boost the effectiveness of early search, the sub-population is divided, the differences between the sub-populations are increased, and the members of the sub-population are forced to concentrate on searching within a certain area Then, the weighted center of gravity communication strategy is used to improve the quality of population communication, reduce the interference of its own population, and reduce the risk of all sub populations falling into local optimization due to a sub population falling into local optimization. Finally, dynamic reverse learning is introduced into vigilant to enhance their back feeding behavior and improve the defects of slow convergence speed and insufficient convergence accuracy caused by the increase of factor population. Through the test function simulation experiment, it is proved that KSSA has better optimization performance than SSA and other algorithms.

Keywords：sparrow search algorithm；optimal algorithm；multiple groups；K-means clustering；population communication

Received：2022-05-07；**Accepted**：2022-09-18；**Published Online**：2022-10-18 10：17

URL：link.cnki.net/urlid/11.2625.V.20221017.1733.012

Foundation item：National Natural Science Foundation of China (61703411)

* **Corresponding author**. E-mail：yyp_ing@163.com

http://bhxb.buaa.edu.cn jbuaa@buaa.edu.cn
DOI: 10.13700/j.bh.1001-5965.2022.0340

基于通信连通保持约束的多智能体覆盖控制

张蕴霖[1]，麻壮壮[1]，石磊[1,2]，邵晋梁[1,2,*]

(1. 电子科技大学 自动化工程学院，成都 611731；2. 电磁空间认知与智能控制技术实验室，北京 100089)

摘　　要：覆盖控制会根据环境信息，尽量分散智能体以达到较好的空间覆盖效果，实现对于任务区域的最优监视。在这一过程中，智能体之间的协同依赖于连通的通信网络。受限于智能体在复杂电磁环境下有限的通信范围，覆盖控制中的分散行为可能造成通信网络的中断，导致任务失败。因此，将通信网络连通作为约束，基于梯度下降方法，提出了有界的分布式控制律，以保证在覆盖代价函数下降的同时，网络连通度不小于设定阈值。为减少通信连通保持对于覆盖效果的影响，提出了一种基于关键智能体辨识的分段控制方法。通过动态分配覆盖和通信连通保持的控制增益，以减少因二者运动趋势相反而产生的控制震荡和冗余。针对陷入局部最优的死锁现象，提出一种死锁消除控制，以及时消除死锁，提升覆盖性能。基于高频结构仿真（HFSS）软件生成的信号场的覆盖仿真实验表明了控制律的有效性。

关 键 词：多智能体系统；覆盖控制；通信连通性保持；维诺划分；死锁
中图分类号：TP13
文献标志码：A　　**文章编号**：1001-5965(2024)02-0519-10

覆盖控制描述了携带传感器的可移动的多智能体系统(multi-agent system, MAS)的最优部署问题，旨在驱动MAS从任意的初始位置出发，快速到达最优的部署位置，协作地实现对于侦测区域的最优监视。该问题有诸多现实应用，如机器人集群用于森林火灾的监测[1-2]、人员的搜索和救援[3]等。在该类场景中，为了分布式协同完成任务，智能体的决策需要邻居的状态，数据贡献依赖于连通的通信网络。连通的通信网络是传统覆盖控制算法[4-5]有效的重要前提。

然而，随着用频设备的增多，智能体往往在复杂的电磁环境中执行任务。辐射、传导等人为电磁干扰以及雷电等自然辐射源都将极大地降低通信链路的稳定性[6]。通信受限的智能体仅有有限的通信范围。由于覆盖控制会根据环境信息，尽量分散智能体以实现较好的空间覆盖效果，覆盖过程中的分散行为会中断部分通信链路。这可能会造成网络连通中断，导致任务失败。作为集群交互的基础，通信连通保持可以通过规划集群轨迹以限制通信网络的扩展，从而维持网络的连通，因此通信连通保持是复杂电磁环境下成功执行覆盖控制任务的关键。

在覆盖控制领域中，通信连通保持问题尚未得到广泛的研究。文献[7-8]采用始终维持初始通信链路的思路来保持覆盖过程中网络的连通。而文献[9-10]使用基于代数连通度梯度的方法，允许不影响网络全局连通的链路自由断开，其通信拓扑具有更好的灵活性和适应性[11]，是本文主要关注的方法。具体而言，文献[9]通过综合覆盖和通信连通保持控制律的方式实现通信连通保持下的覆盖控制，但其控制的设计是无界的；文献[10]提出一种梯度投影的方法，通过保留覆盖项在连通项的投影

收稿日期：2022-05-10；录用日期：2022-07-15；网络出版时间：2022-10-18 10:26
网络出版地址：link.cnki.net/urlid/11.2625.V.20221017.1412.004
基金项目：四川省自然科学基金(2021YFH0042)
*通信作者. E-mail：jinliangshao@uestc.edu.cn

引用格式：张蕴霖，麻壮壮，石磊，等. 基于通信连通保持约束的多智能体覆盖控制[J]. 北京航空航天大学学报，2024，50(2)：519-528.
ZHANG Y L，MA Z Z，SHI L，et al. Multi-agent coverage control based on communication connectivity maintenance constraints[J]. Journal of Beijing University of Aeronautics and Astronautics，2024，50(2)：519-528 (in Chinese).

的正交项来保持连通，但控制未补偿连通度的估计误差。借助强化学习方法，文献[12]采用几何方法约束智能体运动保持连通，但依赖于模型的训练。除此之外，覆盖和通信连通保持在运动趋势上具有相反的表现，如何减少通信连通保持对覆盖效果造成的影响也是上述文献中未能研究的。

启发于上述工作，本文基于代数连通度梯度的方法，考虑代数连通度的下界及估计误差，首先，对文献[9]中的梯度控制器进行拓展，实现了有界控制，并提供了理论证明。其次，在保持通信连通的基础上，进一步提出分段控制的方法，基于关键智能体的辨识结果进行控制增益的分配，以减少通信连通保持律对覆盖控制律的影响。最后，提出一种死锁消除控制以避免覆盖陷入局部最优。以基于HFSS软件仿真的信号场数据作为覆盖对象，通过仿真实验验证了所提控制律的有效性和性能表现。

1 模型建立与问题描述

1.1 覆盖控制任务

考虑一组由N个移动智能体组成的MAS在二维凸区域Q内执行覆盖任务，任务区域中分布着特定的侦测目标(覆盖对象)，侦测目标在空间中的分布由密度函数$\phi:Q\to\mathbf{R}_+$描述，其值衡量了任务区域Q内侦测目标的分布情况。MAS覆盖任务的目标是驱动智能体以最小的代价完成区域侦察和监视，其覆盖性能由如下的联合代价函数(损失函数)进行描述：

$$H(\boldsymbol{x})=\int_Q \min_{i\in\{1,\cdots,N\}} f(\|\boldsymbol{q}-\boldsymbol{x}_i\|)\phi(\boldsymbol{q})\mathrm{d}\boldsymbol{q} \tag{1}$$

式中：$\boldsymbol{q}\in\mathbf{R}^2$为区域$Q$内的任意一点位置的坐标；$\boldsymbol{x}_i\in\mathbf{R}^2$为第$i$个智能体的位置，$\boldsymbol{x}=(\boldsymbol{x}_1,\cdots,\boldsymbol{x}_N)$为所有$N$个智能体的位置；$f(\|\boldsymbol{q}-\boldsymbol{x}_i\|)\in\mathbf{R}$为智能体$i$对于$\boldsymbol{q}$点的侦测代价值，侦测代价的形式与覆盖对象有关，值往往与距离成正比，通常表示为$f(\|\boldsymbol{q}-\boldsymbol{x}_i\|)=\|\boldsymbol{q}-\boldsymbol{x}_i\|^2$，$\|\cdot\|$表示向量2范数。

1.2 通信网络模型

MAS的连通性是系统内所有个体的通信关系的整体度量，衡量了空间中不同部署位置的MAS通信网络的质量。建立通信模型是研究多智能体连通性的基础。为此，对于个体间的通信关系，需要建立合适的通信链路模型，以评估其通信质量，并进一步针对MAS，选择合适的通信连通性的量化指标。本文综合考虑小尺度衰落和路径衰减对于无线通信链路的影响，选择包错误率作为通信质量的衡量指标。首先给出结点$i\to j$的通信链路接收端处的信干噪比γ_{ij}模型为

$$\gamma_{ij}(d_{ij})=\frac{P^{i,j}d_{ij}^{-\beta}\left|h_{ij}\right|^2}{\sigma^2}, \tag{2}$$

式中：$P^{i,j}$为发送端i的发射功率；$d_{ij}^{-\beta}$为路径衰减，β为衰减因子；d_{ij}为传输距离；$\left|h_{ij}\right|^2$为小尺度衰落；σ^2为背景噪声。在给定信道参数β、h_{ij}、σ和发射功率$P^{i,j}$后，γ_{ij}与距离d_{ij}负相关。即距离越远，γ_{ij}越小，信道质量就相应越差。

在给定编码调制模式后，信道质量可由包错误率描述[13]，计算式为

$$p(\gamma_{ij})\approx\begin{cases}1 & 0<\gamma_{ij}<\gamma_p\\ a\mathrm{e}^{-g\gamma_{ij}} & \gamma_{ij}\geqslant\gamma_p\end{cases} \tag{3}$$

式中：$p(\gamma_{ij})$表示包错误率；a、g、γ_p为与编码和调制相关的参数。当γ_{ij}低于接收端通信系统的容忍阈值γ_p后，通信链路视为$i\to j$将无法在物理层建立，视为完全丢失。因此，边界处γ_p所对应的距离$R_{ij}=p^{-1}(\gamma_p)$可视为个体j相对于个体i的通信范围，即为通信感知域。当且仅当个体i位于个体j的通信感知域内，个体j才能接收到其发送的数据，二者才能建立起通信链路。由此可得到个体j的邻居集$\widetilde{N}_j$为

$$\widetilde{N}_j=\left\{i\in\{1,\cdots,N\}\,|\,d_{ij}\leqslant R_{ij}\right\} \tag{4}$$

基于上述通信链路模型，本文选择代数连通度作为系统连通度的度量指标。首先选择与距离负相关的边权函数$a(d_{ij}):\mathbf{R}_+\to\mathbf{R}_+$，其中边权0表示链路断开。然后计算通信图$G$对应的拉普拉斯矩阵$\boldsymbol{L}$。由于$\boldsymbol{L}$为半正定矩阵，因此拥有$N$个非负特征值，用$\lambda_2(G)$表示其第二小特征值，$\lambda_2(G)$即为代数连通度。当$G$为连通图时，$\lambda_2(G)>0$，反之$\lambda_2(G)=0$。$\lambda_2(G)$揭示了网络连通的代数条件，同时代数连通度越大，网络的连通程度越大，即结点之间的建立的通信链路就越多，质量就越好。

1.3 通信连通保持约束下的覆盖控制问题

在集群的分布式控制中，有时不仅需要网络保持连通，并且由于链路丢失或对收敛速度最低需求等因素还需设置连通度的下界。因此，结合上述的覆盖任务和通信模型，以$\lambda_2(G)\geqslant\varepsilon,\varepsilon>0$作为通信连通保持约束，$\varepsilon$为系统所能接受的最小连通度值。本文提出如下的MAS控制任务目标：给定任务区域Q和密度函数$\phi(\cdot)$，假设每个智能体具有相同的通信范围R，能自动与范围内个体建立通信链路(如广播通信)，通信拓扑构成无向图，并具有如下的一阶积分器动力学系统：

$$\dot{\boldsymbol{x}}_i=\boldsymbol{u}_i^{\mathrm{c}}+\boldsymbol{u}_i^{\mathrm{c}} \tag{5}$$

式中：$\boldsymbol{x}_i\in Q$为位置；$\boldsymbol{u}_i^{\mathrm{c}}\in\mathbf{R}^2$为覆盖控制律；$\boldsymbol{u}_i^{\mathrm{c}}\in\mathbf{R}^2$为通信连通保持控制律。设计分布式控制

使 MAS 在满足通信连通保持约束的同时, 使代价函数 (1) 最小化。本文研究的是密度函数已知的情况, 对于密度函数未知的场景, 可结合实际采集的数据, 采用插值、高斯估计等方法获取密度函数, 以此转化为本文研究的场景。

2　控制器设计

2.1　覆盖控制律

由于式 (1) 定义的联合全局代价函数难以直接优化求解, 因此本文首先利用维诺划分的方法将区域Q划分为互不重叠的N个子单元$V_1,\cdots,V_N$作为对应智能体的工作区域, 下面将介绍维诺划分的概念。

定义 1[14] 考虑对区域Q的一组划分$V_1,\cdots,V_N$, $V_i\cap V_j=\varnothing$, $\bigcup\limits_{i=1}^{N}V_i=Q$给定$\boldsymbol{x}=(\boldsymbol{x}_1,\cdots,\boldsymbol{x}_N)$, 若满足:

$$V_i=\left\{\boldsymbol{q}\in Q|f(\|\boldsymbol{q}-\boldsymbol{x}_i\|)\leqslant f\left(\|\boldsymbol{q}-\boldsymbol{x}_j\|\right),\forall j\neq i\right\},$$

则$V(\boldsymbol{x})=(V_1,\cdots,V_N)$是以$\boldsymbol{x}$为生成点的一组维诺划分。

维诺划分将根据每个智能体的当前位置, 分配不重叠的工作区域。根据维诺划分的定义, 显然, 如果点$\boldsymbol{q}$在维诺单元V_j中, 则侦测代价函数满足$\min\limits_{i\in\{1,\cdots,N\}}f(\|\boldsymbol{q}-\boldsymbol{x}_i\|)=f\left(\|\boldsymbol{q}-\boldsymbol{x}_j\|\right)$, 因此, 联合覆盖任务的损失函数$H(\boldsymbol{x})$可直接转化为

$$H(\boldsymbol{x})=\sum_{i=1}^{N}\int_{V_i}f(\|\boldsymbol{q}-\boldsymbol{x}_i\|)\phi(\boldsymbol{q})\mathrm{d}\boldsymbol{q}\tag{6}$$

由此实现了联合代价函数的解耦, 使得分布式算法的设计成为可能。同时, 下述引理描述了代价函数 (6) 的梯度性质。

引理 1[15]　代价函数 (6) 关于位置$\boldsymbol{x}_i$的梯度可计算为

$$\frac{\partial H}{\partial\boldsymbol{x}_i}=\int_{V_i}\frac{\partial}{\partial\boldsymbol{x}_i}f(\|\boldsymbol{q}-\boldsymbol{x}_i\|)\phi(\boldsymbol{q})\mathrm{d}\boldsymbol{q}\tag{7}$$

由引理 1, 结合$f(\|\boldsymbol{q}-\boldsymbol{x}_i\|)=\|\boldsymbol{q}-\boldsymbol{x}_i\|^2$, 本文研究的代价函数关于智能体位置的梯度可计算为

$$\frac{\partial H}{\partial\boldsymbol{x}_i}=2\boldsymbol{x}_i\int_{V_i}\phi(\boldsymbol{q})\mathrm{d}\boldsymbol{q}-2\int_{V_i}q\phi(\boldsymbol{q})\mathrm{d}\boldsymbol{q}\tag{8}$$

为了符号的简洁, 令

$$\begin{cases}M_{V_i}=\displaystyle\int_{V_i}\phi(\boldsymbol{q})\mathrm{d}\boldsymbol{q}\\ \boldsymbol{C}_{V_i}=\dfrac{1}{M_{V_i}}\displaystyle\int_{V_i}\boldsymbol{q}\phi(\boldsymbol{q})\mathrm{d}\boldsymbol{q}\end{cases}\tag{9}$$

M_{V_i}可视为维诺单元广义上的质量, $\boldsymbol{C}_{V_i}$可视为广义上的质心, 代价函数梯度 (7) 可表示为

$$\frac{\partial H}{\partial\boldsymbol{x}_i}=2M_{V_i}\left(\boldsymbol{x}_i-\boldsymbol{C}_{V_i}\right)\tag{10}$$

由此, 当每个智能体维诺质心$\boldsymbol{C}_{V_i}$和其生成点$\boldsymbol{x}_i$重合时, 称此维诺划分为 CVC(centroidal voronoi configuration), 代价函数 (1) 达到全局最小值。基于梯度下降的思想, 设计如下的有界的覆盖控制律$\boldsymbol{u}_i^{\mathrm{e}}$为

$$\boldsymbol{u}_i^{\mathrm{e}}=\begin{cases}\boldsymbol{C}_{V_i}-\boldsymbol{x}_i & \|\boldsymbol{u}_i^{\mathrm{e}}\|<k_{\mathrm{e}}\\ k_{\mathrm{e}}\dfrac{\boldsymbol{C}_{V_i}-\boldsymbol{x}_i}{\|\boldsymbol{C}_{V_i}-\boldsymbol{x}_i\|} & \text{其他}\end{cases}\tag{11}$$

式中: $k_{\mathrm{e}}>0$表示控制律限幅。

控制律 (11) 保证了个体始终朝着当前维诺单元对应的质心方向移动并且输出幅值不大于k_{e}, 经过多次迭代, 维诺单元不再变动时, 智能体将停留在维诺单元的质心, 完成对覆盖对象ϕ的联合最优覆盖, 实现 CVC。根据维诺划分定义, 智能体的维诺单元的计算仅需获取邻居智能体的位置, 因此控制律 (11) 可以分布式计算(不需要知道所有智能体的位置)。同时, 由于维诺单元V_i互不重叠, 且维诺质心始终在其维诺单元内, 控制律 (11) 通过限制智能体仅在各自维诺单元内移动, 避免了碰撞。

2.2　通信连通保持的区域覆盖控制

为实现通信连通保持的控制目标, 文献 [9] 基于梯度下降的思想, 提出基于梯度下降的控制律$\boldsymbol{u}_i^{\mathrm{c}}$为

$$\boldsymbol{u}_i^{\mathrm{c}}=\alpha\mathrm{csch}^2(\lambda_2-\varepsilon)\cdot\nabla_{\boldsymbol{x}_i}\lambda_2\tag{12}$$

式中: $\alpha>0$为控制增益; $\mathrm{csch}(\cdot)$为双曲余割函数, 用于连通度接近设定下界时迅速提升控制量; $\nabla_{\boldsymbol{x}_i}\lambda_2$为$\lambda_2$在个体$i$的位置$\boldsymbol{x}_i$上的梯度。其会使智能体拥有连通度增加的运动趋势, 从而保持了网络的连通。然而上述控制律是无界形式的, 无法应用于实际。因此本文对其进行限幅, 改进为有界形式:

$$\boldsymbol{u}_i^{\mathrm{c}}=\begin{cases}\alpha\mathrm{csch}^2(\lambda_2-\epsilon)\cdot\nabla_{\boldsymbol{x}_i}\lambda_2 & \|\boldsymbol{u}_i^{\mathrm{c}}\|<k_{\mathrm{c}}\\ k_{\mathrm{c}}\dfrac{\nabla_{\boldsymbol{x}_i}\lambda_2}{\|\nabla_{\boldsymbol{x}_i}\lambda_2\|} & \text{其他}\end{cases}\tag{13}$$

式中: $k_{\mathrm{c}}>k_{\mathrm{e}}$为控制律限幅。

由于代数连通度λ_2是全局指标, 其计算需要知道整个拉普拉斯矩阵信息, 不利于式 (13) 的分布式实施。为此, 需要对λ_2及其对应特征向量$\boldsymbol{v}_2$进行分布式估计。对于λ_2的估计问题, 现存的许多工作已经进行了充分的研究, 实现了误差有界的估计。如, 文献 [16-17] 在集中式的幂迭代法基础上, 结合共识估计器, 实现了λ_2、$\boldsymbol{v}_2$的联合分布式估计, 并理论上证明了期望值的渐进收敛性。这些方法中, 个体i的估计仅利用了拉普拉斯矩阵$\boldsymbol{L}$所在行的元素, 即一跳邻居的信息。通过与邻居交换数据, 便可实现λ_2及其特征向量对应分量v_2^i的分布式估计。通

过合理的设置估计器参数增益，可实现需求的收敛速度，并嵌入到控制律的执行当中。由于λ_2的分布式估计并非本文的研究重点，因此具体估计过程将在此省略。

在给定λ_2估计算法的基础上，现定义个体i估计的特征向量分量及其特征值分别为$\tilde{v}_2^i$和$\tilde{\lambda}_2^i$，所有个体的估计分量组成的特征向量及其对应特征值分别为$\tilde{\boldsymbol{v}}_2$和$\tilde{\lambda}_2$，估计系统误差有界且满足：

$$\left|\tilde{\lambda}_2^i-\lambda_2\right|\leqslant\psi_1 \tag{14}$$

$$\left|\tilde{\lambda}_2^i-\tilde{\lambda}_2\right|\leqslant\psi_2 \tag{15}$$

由此，通信连通保持控制$\boldsymbol{u}_i^{\mathrm{c}}$的分布式形式为

$$\boldsymbol{u}_i^{\mathrm{c}}=\begin{cases}\alpha\operatorname{csch}^2\left(\tilde{\lambda}_2^i-\tilde{\varepsilon}\right)\cdot\nabla_{x_i}\tilde{\lambda}_2 & \left\|\boldsymbol{u}_i^{\mathrm{c}}\right\|<k_{\mathrm{c}}\\ k_{\mathrm{c}}\dfrac{\nabla_{x_i}\tilde{\lambda}_2}{\left\|\nabla_{x_i}\tilde{\lambda}_2\right\|} & \text{其他}\end{cases} \tag{16}$$

式中：$\tilde{\varepsilon}=\varepsilon+\psi$，$\psi=\psi_1+\psi_2$。

假设$\|\boldsymbol{v}_2\|=1$，则$\lambda_2=\lambda_2\|\boldsymbol{v}_2\|^2=\boldsymbol{v}_2^{\mathrm{T}}\lambda_2\boldsymbol{v}_2=\boldsymbol{v}_2^{\mathrm{T}}\boldsymbol{L}\boldsymbol{v}_2$。因此梯度$\nabla_{\boldsymbol{x}_i}\tilde{\lambda}_2$可由式(17)逼近：

$$\frac{\partial\tilde{\lambda}_2}{\partial\boldsymbol{x}_i}=\tilde{\boldsymbol{v}}_2^{\mathrm{T}}\frac{\partial\boldsymbol{L}}{\partial\boldsymbol{x}_i}\tilde{\boldsymbol{v}}_2=\sum_{j\in\widetilde{N}_i}\frac{\partial a_{ij}}{\partial\boldsymbol{x}_i}\left(\tilde{v}_2^i-\tilde{v}_2^j\right)^2 \tag{17}$$

上述梯度方向与真实梯度方向相同，仅尺度有差异。下面的引理将描述$\nabla_{\boldsymbol{x}_i}\tilde{\lambda}_2$存在的条件。

引理2[18] 对于连通的拓扑图，若$\partial a_{ij}/\partial d_{ij}\neq 0, j\in\widetilde{N}_i$，且$\partial d_{ij}/\partial\boldsymbol{x}_i\neq 0,\forall i\neq j$，则存在至少一个$k\in\{1,\cdots,N\}$使得$\nabla_{\boldsymbol{x}_k}\tilde{\lambda}_2\neq 0$。

接下来将分析本文所提的通信连通保持控制的有效性，首先引入正向不变集(Forward Invariant Set)的概念。

定义2 正向不变集。设MAS符合动力学约束$\dot{\boldsymbol{x}}=f(\boldsymbol{x})$，如果集合$C\subset\mathbf{R}^n$满足$\boldsymbol{x}(t_0)\in C$，有$\boldsymbol{x}(t)\in C, t\geqslant t_0$，则集合$C$是一个正向不变集。

根据正向不变集的定义，直观上，本文考虑的通信连通控制的目标为将满足连通保持约束的状态集$C_1=\left\{\boldsymbol{x}\in\mathbf{R}^{Nn}:\lambda_2(G(\boldsymbol{x}))\geqslant\varepsilon\right\}$构造为正向不变集。但事实上，由于估计值$\tilde{\lambda}_2$存在误差$\psi$，导致计算的控制量也存在误差，直接将状态集$C_1$构造成正向不变集的目标难以实现。因此，考虑有误差的情况，连通目标可转化为将状态集$C_2=\left\{\boldsymbol{x}\in\mathbf{R}^{Nn}:\tilde{\lambda}_2(G(\boldsymbol{x}))\geqslant\tilde{\varepsilon}\right\}$构造为正向不变集。为此，下面的引理揭示了正向不变集的充分必要条件：

引理3[19] 假设MAS符合动力学约束$\dot{\boldsymbol{x}}=f(\boldsymbol{x})$，状态集$C$为一个平滑函数$h:\mathbf{R}^n\to\mathbf{R}$所对应的超水平集，即$C=\{\boldsymbol{x}\in\mathbf{R}^n|h(\boldsymbol{x})\geqslant 0\}$且当$h(\boldsymbol{x})=0,\nabla_{\boldsymbol{x}}h(\boldsymbol{x})\neq 0$。则$C$为正向不变集的充分必要条件是$\dot{h}(\boldsymbol{x})\geqslant 0, \forall\boldsymbol{x}\in\partial C$。

综上，本文给出如下定理证明通信连通保持控制(16)的有效性和合成的总控制(5)的有效性：

定理1 考虑由覆盖控制律(11)和通信连通保持律(16)组成的动力学系统(5)，给定边权函数满足$\partial a_{ij}/\partial d_{ij}\neq 0, j\in\widetilde{N}_i$以及$\varepsilon,\tilde{\varepsilon}$满足$\tilde{\varepsilon}=\varepsilon+\psi$。如果MAS的初始通信拓扑满足$\tilde{\lambda}_2(t_0)\geqslant\tilde{\varepsilon}$，则通信连通保持的目标$\lambda_2(G)\geqslant\varepsilon$将一直实现。

证明： 构造平滑函数$h(\tilde{\lambda}_2(\boldsymbol{x}))=\tilde{\lambda}_2(\boldsymbol{x})-\tilde{\varepsilon}$，其关于时间的导数为

$$\dot{h}(\tilde{\lambda}_2)=\nabla_x^{\mathrm{T}}h\left(\tilde{\lambda}_2(\boldsymbol{x})\right)\dot{\boldsymbol{x}}=\sum_{i=1}^{N}\left[\nabla_{\boldsymbol{x}_i}^{\mathrm{T}}h\left(\tilde{\lambda}_2(\boldsymbol{x})\right)\dot{\boldsymbol{x}}_i\right]=$$
$$\sum_{i=1}^{N}\left[\nabla_{\tilde{\lambda}_2}h\left(\tilde{\lambda}_2(\boldsymbol{x})\right)\nabla_{\boldsymbol{x}_i}^{\mathrm{T}}\tilde{\lambda}_2\cdot\left(\boldsymbol{u}_i^{\mathrm{e}}+\boldsymbol{u}_i^{\mathrm{c}}\right)\right]=$$
$$\sum_{i=1}^{N}\left[\nabla_{\boldsymbol{x}_i}^{\mathrm{T}}\tilde{\lambda}_2\cdot\left(\boldsymbol{u}_i^{\mathrm{e}}+\boldsymbol{u}_i^{\mathrm{c}}\right)\right]$$

由于覆盖避免了碰撞，即$\partial d_{ij}/\partial\boldsymbol{x}_i\neq 0,\forall i\neq j$，因此引理2成立。根据引理2，假设$\nabla_{\boldsymbol{x}_i}\tilde{\lambda}_2\neq 0, i=1,\cdots,M$，$\nabla_{\boldsymbol{x}_i}\tilde{\lambda}_2=0, i=M+1,\cdots,N$，$M\leqslant N$。考虑边界情况，即$\tilde{\lambda}_2\to\tilde{\varepsilon}^+$时，$\operatorname{csch}(\lambda_2^i)\to\infty$，因此总能找到某一点$\tilde{\lambda}_2'\geqslant\tilde{\varepsilon}$使得

$$\boldsymbol{u}_i^{\mathrm{c}}=k_{\mathrm{c}}\frac{\nabla_{\boldsymbol{x}_i}\tilde{\lambda}_2}{\left\|\nabla_{\boldsymbol{x}_i}\tilde{\lambda}_2\right\|}\qquad i=1,\cdots,M$$

此时有

$$\dot{h}(\tilde{\lambda}_2)=\sum_{i=1}^{M}\left[\nabla_{\boldsymbol{x}_i}^{\mathrm{T}}\tilde{\lambda}_2\cdot\left(\boldsymbol{u}_i^{\mathrm{e}}+k_{\mathrm{c}}\frac{\nabla_{\boldsymbol{x}_i}\tilde{\lambda}_2}{\left\|\nabla_{\boldsymbol{x}_i}\tilde{\lambda}_2\right\|}\right)\right]=$$
$$\sum_{i=1}^{M}\left[\left\|\nabla_{\boldsymbol{x}_i}^{\mathrm{T}}\tilde{\lambda}_2\right\|\cdot\left(\left\|\boldsymbol{u}_i^{\mathrm{e}}\right\|\cos\angle\left(\nabla_{\boldsymbol{x}_i}\tilde{\lambda}_2,\boldsymbol{u}_i^{\mathrm{e}}\right)+k_{\mathrm{c}}\right)\right]\geqslant$$
$$\sum_{i=1}^{M}\left[\left\|\nabla_{\boldsymbol{x}_i}^{\mathrm{T}}\tilde{\lambda}_2\right\|\cdot\left(-\left\|\boldsymbol{u}_i^{\mathrm{e}}\right\|+k_{\mathrm{c}}\right)\right]\geqslant$$
$$\sum_{i=1}^{M}\left[\left\|\nabla_{\boldsymbol{x}_i}^{\mathrm{T}}\tilde{\lambda}_2\right\|\cdot(-k_{\mathrm{e}}+k_{\mathrm{c}})\right]>0$$

根据引理3，状态集$C_2=\left\{\boldsymbol{x}\in\mathbf{R}^{Nn}:\tilde{\lambda}_2(G(\boldsymbol{x}))\geqslant\tilde{\varepsilon}\right\}$构成正向不变集，进一步，连通度存在下界，即$\lambda_2\geqslant\varepsilon$。证明完毕。

2.3 基于关键智能体辨识的分段控制

由于覆盖和通信连通保持两者在运动趋势上是相反的，具有很多冲突。这种冲突会使覆盖控制律和通信连通保持控制律在覆盖过程中交替产生主导，对覆盖效果造成影响：①造成控制的震荡，表现为智能体需要经过较多位置调整才能达到最终的稳定位置；②造成通信连通保持控制的冗余，增加功耗。因此，在保持通信连通的前提下，进一步考虑减少通信连通保持控制对于覆盖效果的影响，即合理分配控制增益，平稳控制输出，进而平稳整

个覆盖过程。显然, 一种可行的解决方案是直接减少覆盖控制项的增益来减少冲突, 但这种静态的增益选择是保守的, 会牺牲较多的覆盖效果。将提出一种基于关键智能体辨识的分段控制方法进行动态的增益分配。

分段控制的设计思路借用了关键智能体[20]的概念。关键智能体往往位于网络的脆弱位置, 表现为已有链路的断开将可能直接造成通信网络的不连通。在定义关键智能体之前, 首先将个体i的邻居根据距离的远近分为远邻$\widetilde{N}_i^{\mathrm{f}}$和近邻$\widetilde{N}_i^{\mathrm{c}}$, 满足$\widetilde{N}_i^{\mathrm{f}}=\left\{j\in\widetilde{N}_i:\left\|\boldsymbol{x}_i-\boldsymbol{x}_j\right\|>\delta\right\}$, $\widetilde{N}_i^{\mathrm{c}}=\widetilde{N}_i-\widetilde{N}_i^{\mathrm{f}}$, 其中$\delta\in(0,R)$为阈值, R为通信范围。显然, 远邻距离智能体个体较远, 相比近邻, 二者的通信链路更容易因运动而中断, 进而网络可能因此中断, 但如果智能体可以通过 2 跳的近邻访问到这个远邻, 网络仍保持连通。因此, 针对这一种情况, 提出关键智能体的概念:

定义 3 关键智能体[20]。若智能体i存在至少一个邻居$j\in\widetilde{N}_i^{\mathrm{f}}$且不满足$\exists k\in\widetilde{N}_i^{\mathrm{c}}$, $j\in\widetilde{N}_k^{\mathrm{c}}$, 则智能体$i$为关键智能体。

由此可见, 关键智能体对网络结构的稳定性至关重要。因此, 关键智能体的运动必须进行严格约束, 防止其链路中断, 而反之非关键智能体施加的运动约束可适当放宽。借助这个思想, 设计如下的分段控制:

$$\boldsymbol{u}_i=\begin{cases}\boldsymbol{u}_i^{\mathrm{e}} & \lambda_2^i>\lambda_2^{\mathrm{warn}}\\ \alpha_1\boldsymbol{u}_i^{\mathrm{e}}+\alpha_2\boldsymbol{u}_i^{\mathrm{c}} & \lambda_2^i\leqslant\lambda_2^{\mathrm{warn}},i\text{ 不为关键}\\ \alpha_3\boldsymbol{u}_i^{\mathrm{e}}+\boldsymbol{u}_i^{\mathrm{c}} & \lambda_2^i\leqslant\lambda_2^{\mathrm{warn}},i\text{ 为关键}\end{cases}\tag{18}$$

式中: α_1、α_2、α_3为增益用于削减对应控制, 满足$0\leqslant\alpha_3<\alpha_1\leqslant1$以及$\alpha_1(k_{\mathrm{e}}/k_{\mathrm{c}})\leqslant\alpha_2<1$; $\lambda_2^{\mathrm{warn}}$为警戒值。通过交换邻居集, 智能体$i$可分布式的辨识自己是否为关键智能体, 并合理分配两种控制的增益。当智能体为关键机器人时, 本文是对于覆盖控制项进行削减, 避免关键智能体的覆盖运动趋势使连通度发生骤降, 从而防止$\boldsymbol{u}_i^{\mathrm{c}}$骤增造成的控制震荡; 反之主要削减通信连通保持项, 以此减少通信连通保持控制的冗余。同时引入警戒值$\lambda_2^{\mathrm{warn}}$, 使得仅在连通度小于$\lambda_2^{\mathrm{warn}}$时开始进行通信连通保持控制, 进一步放宽运动自由度。

定理 2　考虑式 (18) 分段控制下的动力学系统 (5), 给定边权满足: $\partial a_{ij}/\partial d_{ij}\neq0, j\in\widetilde{N}_i$以及$\varepsilon,\tilde{\varepsilon}$满足$\tilde{\varepsilon}=\varepsilon+\psi$。则如果 MAS 的初始通信拓扑满足$\tilde{\lambda}_2(t_0)\geqslant\tilde{\varepsilon}$, 则通信连通保持的目标$\lambda_2(G)\geqslant\varepsilon$将一直实现。

证明: 由于$0\leqslant\alpha_1,\alpha_3\leqslant1,\alpha_1(k_{\mathrm{e}}/k_{\mathrm{c}})\leqslant\alpha_2<1$, 定理 2 的证明与定理 1 类似, 故在此省略。

2.4　死锁检测与消除控制

由上述可知, 在覆盖控制的求解过程中, 覆盖控制和通信连通保持控制具有相反的运动趋势。这可能使有通信连通保持的覆盖过程陷入局部最优解, 其表现为智能体的运动陷入停滞, 代价函数不再优化, 称之为死锁现象。从收益的角度出发, 如果在覆盖任务的前中期就出现死锁现象, 即 MAS 中的大多数智能体尚未覆盖到高密度区域便提前停滞在低密度区域, 那么 MAS 将维持在较差的覆盖配置下, 代价函数值较高。因此, 要致力于死锁消除机制, 使通信连通保持的 MAS 能够及时发现并消除死锁现象, 以获得更好的覆盖效果。

为应对死锁现象, 将首先提出死锁检测方法, 并在此基础上设计对应的死锁消除控制。定义代价函数的变化量函数$e(t)$为

$$e(t)=h(\boldsymbol{x}_{t_i})-h(\boldsymbol{x}_t),\quad t=nT,n\in\mathbf{N}^+\tag{19}$$

式中: t_i为上一个周期开始的时间; T为检测周期; 代价函数变化值$e(t)$描述了覆盖的侦测代价的周期性变化值, 量化了上一周期的覆盖效果。由于死锁现象发生时, 覆盖过程逐渐陷入停滞, 在收益上反映为侦测代价函数不再下降或者变化量减少。基于此, 可设定阈值$\Delta>0$, 当代价函数变化量小于该阈值时, 则可近似视为发生死锁现象。因此, 定义如下激活函数来周期性的检测死锁现象:

$$\varphi(t)=e(t)-\Delta\tag{20}$$

激活函数所需的侦测代价函数的计算可由式 (6) 分布式完成以加快计算速度。结合激活函数, 相应的死锁消除机制设计为

$$\boldsymbol{u}_i=\begin{cases}\dfrac{\sum_{k=1}^{n}\beta_k(\boldsymbol{C}_{V_k}-\boldsymbol{x}_k)}{\left\|\sum_{k=1}^{n}\beta_k(\boldsymbol{C}_{V_k}-\boldsymbol{x}_k)\right\|}v_{\mathrm{m}} & t=nT\leqslant t_{\mathrm{m}},\phi(t)\leqslant0\\ \boldsymbol{u}_i & \text{其他}\end{cases}\tag{21}$$

式中: $\beta_k=M_{V_k}\Big/\left(\sum M_{V_k}\right)$为权重, 权重与维诺单元的质量$M_{V_k}$正相关, 即区域内密度函数的区域积分值越大, 权重越大; v_m为智能体最大运动速度; t_m为死锁消除机制终止时间。

式 (21) 的核心思想在于: 周期性计算激活函数 (20) 来检测死锁现象。当死锁现象发生时, 即激活函数不大于零, 执行死锁消除控制, 控制 MAS 整体向着维诺质心$\boldsymbol{C}_{V_k}$的加权方向移动, 反之维持原控制。死锁消除控制一方面会使 MAS 主动离开死锁位置, 产生位置变化使维诺单元继续更新; 另一方

面能够使MAS向着高密度区域整体移动，加快覆盖过程。即便死锁现象再次发生，但由于MAS整体往高密度区域进行移动，覆盖效果是能一直得到优化的。结合死锁消除机制的覆盖过程中，智能体将周期性广播当前自身位置和维诺质心，并接受其他智能体的信息，然后分布式计算激活函数来检测死锁现象，再计算相应控制量并执行，从而能够及时跳出死锁的位置，以实现较好的覆盖效果。由于死锁消除控制会控制集群整体位移相同的距离，不会更改智能体间的相对距离，因此集群的连通度不会发生变化。

3 仿真验证

3.1 覆盖场强数据的生成

本文仿真实验考虑信号场上的覆盖控制，其中，空间中某点的电磁场强值将作为覆盖代价函数中的该点的密度函数值。信号场数据是通过三维电磁仿真软件HFSS进行获取，天线类型设置为广泛应用于射电望远镜、卫星通讯及电磁波雷达等民用和军用领域中的圆波导喇叭天线，天线长度为18.96 cm，天线结构图如图1所示。

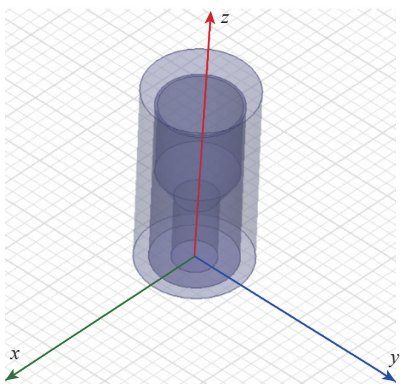


图1 天线结构

Fig. 1 Antenna structure

天线发射波段为5 GHz，得到天线方向图，如图2所示。

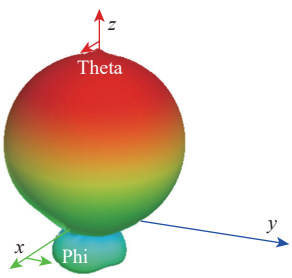


图2 天线方向图

Fig. 2 Antenna radiation pattern

3.2 通信连通保持的覆盖仿真实验验证

为验证本文所提控制律的性能，设置如下的实验场景：考虑20个智能体在边长100 m×68 m的矩形区域执行信号场覆盖任务，所覆盖的信号场图采用的是图2方向图的x-z轴的某一截面，如图3所示，每个位置的场强数值取10lg(·)dB并由不同的颜色进行表示。MAS其余参数设置如表1所示。

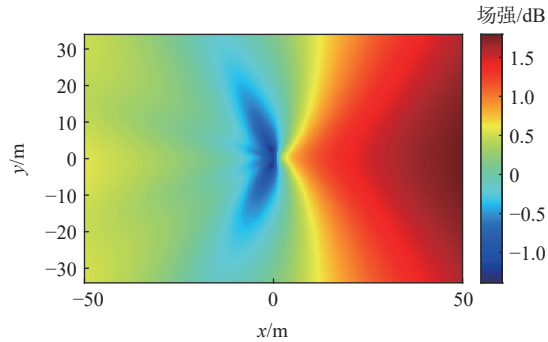


图3 信号场

Fig. 3 Signal field

表1 实验参数

Table 1 Experiment parameter

参数	数值
通信范围R/m	24
最小连通度约束	0.2
控制增益α	1.3
覆盖控制律限幅k_c	2
通信连通保持控制律限幅k_c	3
覆盖总时长/s	100

对应边权函数设置为

$$a_{ij}(d_{ij})=\begin{cases}\mathrm{e}^{-(d_{ij})^2/(250)} & d_{ij}\leqslant 24\\ 0 & \text{其他}\end{cases} \tag{22}$$

满足定理1的要求。MAS从区域的左下角出发，通过计算式(1)的代价函数，初始的覆盖代价值为$1\ 191\times10^3$。

作为Baseline，首先考虑无通信连通保持的信号场覆盖控制场景，实验中忽略了网络连通对于数据交换的影响。无通信连通保持的覆盖最终构型及轨迹如图4所示，其中黑色圆圈表示智能体(机器人)最终停留的位置，黑色虚线表示轨迹，时间为100 s。连通度随时间的变化曲线如图5所示。由图5可知，MAS的网络连通性随覆盖过程的进行而逐渐减小，最终连通度为0，网络断开。由于覆盖控制是基于邻居数据交换的分布式控制，其有效性和收敛性保证以通信网络的连通为前提，因此此处丢失连通性的控制是无法适用于实际的控制需要的。

作为对比，结合本文通信连通保持控制律(16)，通信连通保持的覆盖最终构型及轨迹如图6所示，黑色虚线表示轨迹，时间为100 s。而代价函数与连通度变化如图7所示。由图7可知，本文所提通信

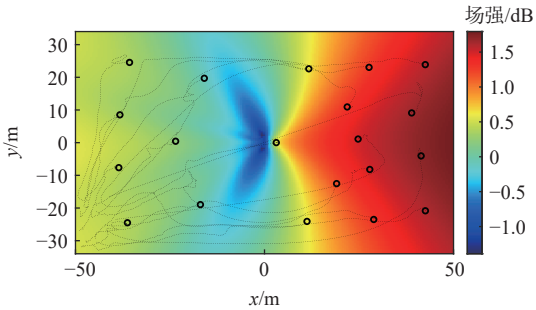


图 4　无通信连通保持的覆盖最终构型及轨迹

Fig. 4　Final configuration and trajectory of coverage without communication connectivity maintenance

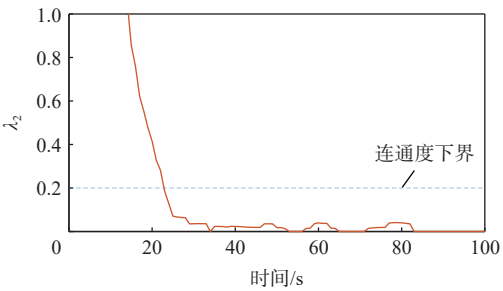


图 5　无通信连通保持的连通度曲线

Fig. 5　Connectivity curve without communication connectivity maintenance

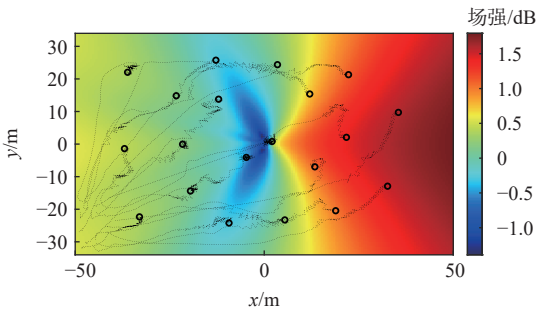


图 6　通信连通保持的覆盖最终构型及轨迹

Fig. 6　Final configuration and trajectory of coverage by communication connectivity maintenance

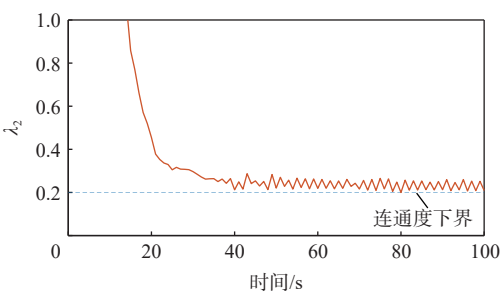


图 7　通信连通保持的连通度曲线

Fig. 7　Connectivity curve with communication connectivity maintenance

连通保持控制律有效, 能够保持连通度始终不低于设定阈值 0.2。

在性能方面, 图 4 对应相对侦测代价值为 108×10^3, 而图 6 对应为 183×10^3, 相比初始侦测代价 $1\ 191\times10^3$, 二者的覆盖性能都较好。为了从多维角度评估算法的性能, 仍选用表 1 中的参数设置, 无连通保持和有连通保持的控制在不同智能体数量设置下的侦测代价函数的性能对比如表 2 所示。每种设置均取 10 次实验的平均值, 并且每次实验中随机选择 MAS 的初始位置。其中侦测代价函数 $H(\boldsymbol{x}_T)$记录了用第 100 时刻的侦测代价。

表 2　有/无连通保持的覆盖性能对比

Table 2　Comparison of coverage performance with and without connectivity maintenance

智能体数量	无连通保持 $H(\boldsymbol{x}_T)/10^3$		有连通保持 $H(\boldsymbol{x}_T)/10^3$	
	均值	标准差	均值	标准差
10	162.3	7.227	346.3	73.66
15	128.7	1.939	255.6	19.04
20	109.5	1.054	176.5	6.62

由表 2 可知, 随着智能体数量的增多, 2 种设置的覆盖的侦测代价值均变少, MAS 整体覆盖性能将变好。然而, 无连通保持的侦测代价的均值和方差整体小于有连通保持的情况, 可见通信连通保持的覆盖控制牺牲了部分的覆盖性能, 并且最终实现的覆盖性能与 MAS 的初始位置有关。覆盖性能的稳定性将在智能体数量增多时($n=20$)有所改善。

3.3　分段控制的仿真实验验证

需要注意的是图 7 连通度曲线均存在较多震荡, 在对应的覆盖轨迹(图 6)上表现为存在较多的锯齿轨迹(黑色虚线重叠部分)。这表明了为了满足连通约束, 智能体需要在同一块区域反复调整位置。造成这种现象的原因是覆盖控制律与通信连通保持控制律存在较多的冲突, 进而导致控制输出数值上下震荡, 有通信连通保持的控制分量曲线如图 8 所示。其中, 覆盖控制量由 $\sum_{i=1}^{N}\left\|\boldsymbol{u}_i^{\mathrm{c}}\right\|$计算, 通信

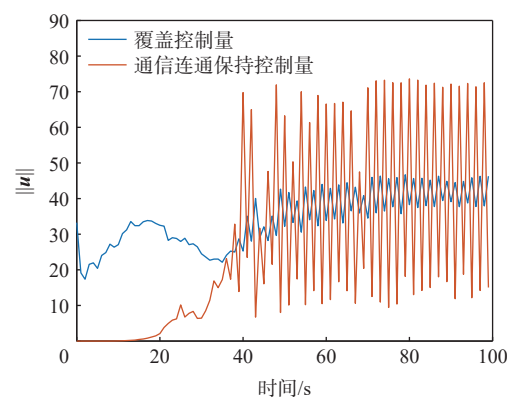


图 8　有通信连通保持的控制分量曲线

Fig. 8　Control component curves with communication connectivity maintaining coverage

连通保持控制量由$\sum_{i=1}^{N}\|\boldsymbol{u}_i^{\mathrm{c}}\|$计算。

为改善这一问题，对本文提出的基于关键智能体的分段控制(18)进行仿真实验，相关控制器参数设置为$\alpha_1=0.6,\alpha_2=0.5,\alpha_3=0.2,\delta=20$ m，警戒特征值$\lambda_2^{\mathrm{warn}}=0.5$。分段控制的连通度曲线如图9所示，控制分量的变化如图10所示。由图9和图10可知，分段控制合理分配覆盖控制律和通信连通保持项的控制占比，在连通度的变化上更为平滑稳定，并且在控制输出上，一方面减少了控制震荡，另一方面减少总体控制量。

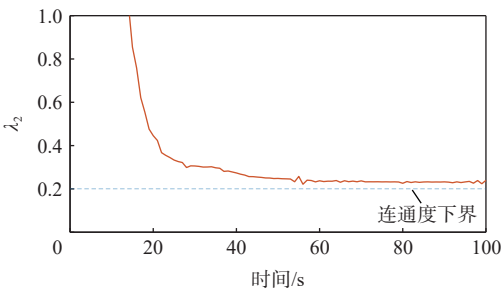


图9 分段控制的连通度曲线

Fig. 9 Connectivity curve of piecewise control

同时，为了从多维角度评估分段控制对于仅考虑通信连通保持的覆盖控制(不分段控制)的覆盖效果优化，不同智能体情况下连通保持控制与分段控制的各项性能指标如表3所示，其中均值$\bar{u}$以及方差σ取20～100 s区间的总控制量进行计算，即连通保持控制开始生效的区间，每种设置均取10次实验的平均值。

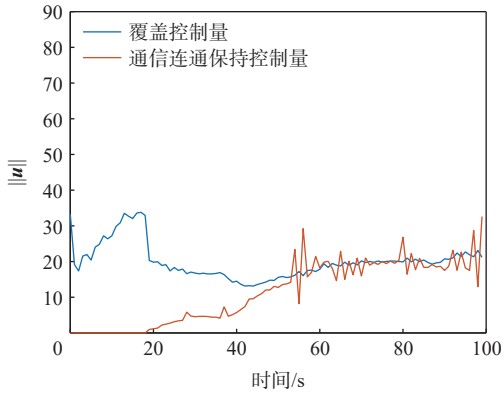


图10 分段控制的控制分量曲线

Fig. 10 Control component curves of piecewise control

由表3可知，相比于不分段的控制，分段控制的代价函数值稍低于不分段控制，表明其牺牲了部分的覆盖性能，但其总控制具有更低的均值和方差，表明其有效地减少了总控制地输出量，平稳控制输出。

表3 不分段控制与分段控制的覆盖效果对比

Table 3 Comparison of coverage performance between non piecewise control and piecewise control 10^3

智能体数量	$H(\boldsymbol{x}_T)$		$\bar{u}$		σ	
	不分段控制效果	分段控制效果	不分段控制效果	分段控制效果	不分段控制效果	分段控制效果
10	346.3	369.8	23.25	7.979	7.491	3.192
15	255.6	288.3	28.16	9.395	8.101	4.301
20	176.5	218.8	29.71	9.549	7.918	4.259

3.4 死锁消除控制的仿真实验验证

本文以10个智能体的覆盖场景为例，展示死锁消除控制对于覆盖效果的提升情况。取死锁检测周期$T=5$ s，阈值$\Delta=25\times10^3$，终止时间$t_{\mathrm{m}}=100$ s，最大速度$v_{\mathrm{m}}=2$ m/s。无死锁消除和有死锁消除的覆盖控制的代价函数优化情况如图11所示。由图11可知，随着覆盖过程的进行，在无死锁消除情况的覆盖中，代价函数会因连通保持控制律对覆盖控制律的抵消而放缓下降速度，最终陷入局部最优的覆盖构型；而在有死锁消除情况的覆盖中，智能体能够及时检测并消除死锁，使得代价函数能够进一步的优化，提升覆盖性能。

死锁消除机制在不同智能体数量设置下的覆盖代价函数值对比如表4所示，每种设置均取10次实验的平均值。由表4可知，死锁消除机制对

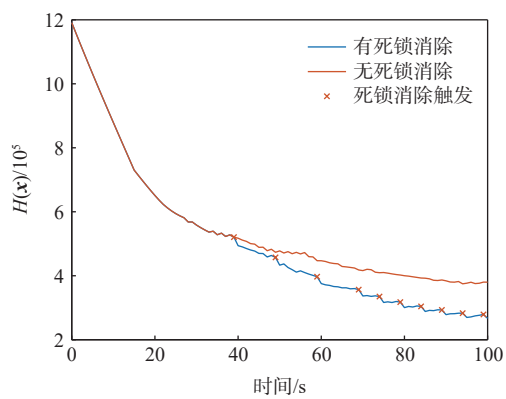


图11 有/无死锁消除控制的代价函数曲线

Fig. 11 Cost function curves with and without deadlock elimination control

表 4　有/无死锁消除的覆盖代价函数值对比

Table 4　Comparison of coverage cost function values with and without deadlock elimination　10^3

智能体数量	最优值	连通控制	分段控制	死锁消除
10	162.3	346.3	369.8	279.2
15	128.7	255.6	288.3	182.9
20	109.5	176.5	218.8	161.9

于覆盖效果均有一定程度的提升。随着智能体数量增大, MAS 覆盖时允许网络的拓展增多, 死锁对于效果的提升效果也逐渐减少。

4　结　论

1) 提出的综合覆盖控制和通信连通保持控制的控制律能够实现在满足代数连通度不低于设定阈值的同时, 降低覆盖的代价函数值。相比于不考虑通信连通约束的控制场景, 其代价函数值的均值和方差较高, 表明通信连通保持控制会牺牲部分的覆盖性能, 同时覆盖性能与初始位置有关。

2) 提出的分段控制律能够合理分配控制增益, 降低了通信连通保持的覆盖控制的总体控制量, 并平稳控制输出, 减少了通信连通保持控制对于覆盖效果的影响。

3) 提出的死锁消除控制律能够周期性检测并消除死锁现象, 使代价函数得到进一步地优化, 提升了最终的覆盖性能。

本文所研究的基于通信连通保持约束的覆盖控制算法并未考虑有障碍物的场景。因此, 未来的研究方向是将路径规划纳入本文的覆盖控制律的设计当中, 实现更为安全的覆盖控制。

参考文献（References）

[1] PHAM H X, LA H M, FEIL-SEIFER D, et al. A distributed control framework of multiple unmanned aerial vehicles for dynamic wildfire tracking[J]. IEEE Transactions on Systems, Man, and Cybernetics: Systems, 2020, 50(4): 1537-1548.

[2] SERAJ E, GOMBOLAY M. Coordinated control of UAVs for human-centered active sensing of wildfires[C]//2020 American Control Conference. Piscataway: IEEE Press, 2020: 1845-1832.

[3] KANTOR G, SINGH S, PETERSON R, et al. Distributed search and rescue with robot and sensor teams[M]//Springer Tracts in Advanced Robotics. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006: 529-538.

[4] CORTES J, MARTINEZ S, KARATAS T, et al. Coverage control for mobile sensing networks[J]. IEEE Transactions on Robotics and Automation, 2004, 20(2): 243-255.

[5] SCHWAGER M, RUS D, SLOTINE J J. Decentralized, adaptive coverage control for networked robots[J]. The International Journal of Robotics Research, 2009, 28(3): 357-375.

[6] WU Y L, REN X G, ZHOU H, et al. A survey on multi-robot coordination in electromagnetic adversarial environment: challenges and techniques[J]. IEEE Access, 2020, 8: 53484-53497.

[7] LUO W H, SYCARA K. Voronoi-based coverage control with connectivity maintenance for robotic sensor networks[C]//2019 International Symposium on Multi-Robot and Multi-Agent Systems. Piscataway: IEEE Press, 2019: 148-154.

[8] 董鑫, 李晓丽, 富锦芸. 具有盲区的有向视觉传感器网络连通保持覆盖控制[J]. 计算机应用研究, 2021, 38(10): 3137-3142.
DONG X, LI X L, FU J Y. Dcrected vision sensor network with blind spots keeps conneted and maintains coverage control[J]. Application Research of Computers, 2021, 38(10): 3137-3142(in Chinese).

[9] SILIGARDI L, PANERATI J, KAUFMANN M, et al. Robust area coverage with connectivity maintenance [C]//2019 International Conference on Robotics and Automation. New York: ACM, 2019: 2202-2208.

[10] KAWAJIRI S, HIRASHIMA K, SHIRAISHI M. Coverage control under connectivity constraints[C]//Proceedings of the 20th International Conference on Autonomous Agents and Multi Agent Systems. Richland: International Foundation for Autonomous Agents and Multiagent Systems, 2021: 1554-1556.

[11] KHATERI K, POURGHOLI M, MONTAZERI M, et al. A comparison between decentralized local and global methods for connectivity maintenance of multi-robot networks[J]. IEEE Robotics and Automation Letters, 2019, 4(2): 633-640.

[12] WU S G, PU Z Q, LIU Z, et al. Multi-target coverage with connectivity maintenance using knowledge-incorporated policy framework [C]//2021 IEEE International Conference on Robotics and Automation. Piscataway: IEEE Press, 2021: 8772-8778.

[13] LIU Q W, ZHOU S L, GIANNAKIS G B. Cross-layer combining of adaptive modulation and coding with truncated ARQ over wireless links[J]. IEEE Transactions on Wireless Communications, 2004, 3(5): 1746-1755.

[14] LIU Q C, YE M B, SUN Z Y, et al. Coverage control of unicycle agents under constant speed constraints[J]. IFAC-PapersOnLine, 2017, 50(1): 2471-2476.

[15] SCHWAGER M A. gradient optimization approach to adaptive multi-robot control[D]. Boston: Massachusetts Institute of Technology, 2009: 42-44.

[16] YANG P, FREEMAN R A, GORDON G J, et al. Decentralized estimation and control of graph connectivity for mobile sensor networks[J]. Automatica, 2010, 46(2): 390-396.

[17] SABATTINI L, CHOPRA N, SECCHI C. Decentralized connectivity maintenance for cooperative control of mobile robotic systems [J]. International Journal of Robotics Research, 2013, 32(12): 1411-1423.

[18] BLANCHINI F. Set invariance in control[J]. Automatica, 1999, 35(11): 1747-1767.

[19] CAPELLI B, SABATTINI L. Connectivity maintenance: Global and optimized approach through control barrier functions [C]//2020 IEEE International Conference on Robotics and Automation. Piscataway: IEEE Press, 2020: 5590-5596.

[20] HUNG P D, VINH T Q, NGO T D. Hierarchical distributed control for global network integrity preservation in multirobot systems[J]. IEEE Transactions on Cybernetics, 2020, 50(3): 1278-1291.

Multi-agent coverage control based on communication connectivity maintenance constraints

ZHANG Yunlin[1], MA Zhuangzhuang[1], SHI Lei[1, 2], SHAO Jinliang[1, 2, *]

(1. School of Automation Engineering, University of Electronic Science and Technology of China, Chengdu 611731, China;
2. Laboratory of Electromagnetic Space Cognition and Intelligent Control, Beijing 100089, China)

Abstract: Coverage control will disperse the agents as much as possible according to the environmental information to achieve a better spatial coverage effect and realize the optimal monitoring of the task area. In this process, the cooperation between agents depends on the connected communication network. Limited by the finite communication range of agents in complex electromagnetic environments, the decentralized behavior in coverage control may cause the interruption of the communication network and task failure. Therefore, to ensure that the coverage cost function lowers while the network connectivity does not fall below the predetermined threshold, this study uses the connectivity of the communication network as a constraint and offers a bounded distributed control law based on the gradient descent approach. A segmented control strategy based on the identification of critical agents is also proposed in order to lessen the impact of communication link maintenance on the coverage effect. By dynamically allocating the control gains of coverage and communication connectivity maintenance, the control oscillation and redundancy caused by the opposite movement trend of the two are reduced. Finally, aiming at the deadlock phenomenon of falling into local optimization, this paper proposes a deadlock elimination control, which can eliminate the deadlock in time and improve coverage performance. The coverage simulation experiment of the signal field generated by high-frequency structure simulation (HFSS) software shows the effectiveness of the proposed control laws.

Keywords: multi-agent system; coverage control; communication connectivity maintenance; Voronoi partition; deadlock

Received: 2022-05-10; **Accepted**: 2022-07-15; **Published Online**: 2022-10-18 10:26
URL: link.cnki.net/urlid/11.2625.V.20221017.1412.004
Foundation item: Sichuan Science and Technology Program (2021YFH0042)
* **Corresponding author**. E-mail: jinliangshao@uestc.edu.cn

http://bhxb.buaa.edu.cn jbuaa@buaa.edu.cn
DOI: 10.13700/j.bh.1001-5965.2022.0375

基于变形图匹配的知识图谱多跳问答

李香粤[1]，方全[2,*]，胡骏[2]，钱胜胜[2]，徐常胜[2]
(1. 郑州大学 河南先进技术研究院，郑州 450000；2. 中国科学院自动化研究所 模式识别国家重点实验室，北京 100190)

摘 要：知识图谱问答 (KGQA) 是给定自然语言问题，对问题进行语义理解和解析，进而利用知识图谱进行查询、推理得出答案的过程。但知识图谱通常是不完整的，链接缺失给多跳问答带来许多挑战。许多方法在利用知识图谱嵌入时忽略了重要的路径信息来评估路径和多关系问题之间的相关性；且使用文本语料库也会限制文本增强模型的可扩展性。针对这些现有方法的缺陷，提出了基于变形图匹配的知识图谱问答 (DGM-KGQA) 模型，该模型同时利用问题和主题实体构建语义子图，与知识图谱的局部结构匹配并找到正确答案。在基准数据集 MetaQA 上的实验结果验证了 DGM-KGQA 的有效性，该模型在完整知识图谱上检索到的答案准确率分别比 PullNet、EmbedKGQA 增加了 4.2%、0.8%；在完整度仅有一半的知识图谱上检索到的答案准确率分别比 PullNet、EmbedKGQA 增加了 11.1%、0.5%。实验证明提出的变形图匹配模型能够有效地增强知识图谱的关联性及多跳问答的答案准确率。

关 键 词：自然语言问题；链接缺失；文本语料库；多跳问答；变形图匹配
中图分类号：V221.3; TB553
文献标志码：A **文章编号**：1001-5965(2024)02-0529-06

知识图谱是具有有向图结构的语义知识库，基本组成单位是“主体–属性–客体”三元组，图中的节点代表实体，连接若干节点的边代表实体之间的语义关系，实体之间通过关系相互联结，构成网状的知识结构。通过使用具有定义良好的语法的正式查询语言，如 SQL、SPARQL 等，可以方便地访问存储在 KGs 中的知识，但这需要有实体的底层结构及其关系的知识储备，当普通用户使用自然语言语句访问知识图谱时，会出现了多跳困难的问题。知识图谱问答 (knowledge graph question answering, KGQA) 系统旨在为用户提供一个界面，用自然语言提问时，问答系统通过在知识图谱中查询，检索到尽可能匹配的答案，目前这些系统已经被集成到 Siri4 等流行的对话助手里。图匹配是指在两个或多个图之间建立有意义的结构对应关系[1]，在分子生物学、计算机视觉、社交网络分析等领域有深入的应用[2-6]。图匹配的方法往往被泛化为图编辑距离，或者寻找图之间的共同部分，或者可能是二次分配问题等问题，而这些方法本质上是 NP-Hard 问题[7-9]。

在知识图谱问答领域，一般从问题和主题实体出发，经过分析再映射或构造到知识图谱中寻找答案，这个过程本质上可以看成一种图匹配。但是，KGQA 本身的特点决定了该过程不是纯粹的组合方法，因为 Q&A 分析的结果和知识图谱的内容不仅可以用语义向量的形式表示，彼此存在直接或间接关联，这些关联和表示会大大减少组合的搜索空间，因此可以在单跳或多跳的 KGQA 应用场景中通过构造子图来解决困难。事实上，在非 KGQA 的应用场景中，许多研究人员最近通过如何缩小搜索空间等思想，通过构建神经网络或图相似度，将问题变得更易于解决[10-14]。

收稿日期：2022-05-18；**录用日期**：2022-06-23；**网络出版时间**：2022-08-15 12:26
网络出版地址：link.cnki.net/urlid/11.2625.V.20220812.1724.005
基金项目：国家自然科学基金 (62072456,62036012,62106262)；之江实验室开放课题 (2021KE0AB05)
* **通信作者**. E-mail：qfang@nlpr.ia.ac.cn

引用格式：李香粤，方全，胡骏，等. 基于变形图匹配的知识图谱多跳问答 [J]. 北京航空航天大学学报，2024，50（2）：529-534.
LI X Y，FANG Q，Hu J，et al. Multi-hop knowledge graph question answering based on deformed graph matching[J]. Journal of Beijing University of Aeronautics and Astronautics，2024，50（2）：529-534（in Chinese）.

本文提出了一个使用自然语言问题和主题实体构建语义子图的模型：基于变形图匹配的知识图谱 (knowledge graph question answering based on deformed graph matching, DGM-KGQA) 多跳回答模型。该模型将问答任务放在知识库和文本的结合上，既解决了知识图谱具有稀疏性的问题，又解决了需要使用的文本没有来源的困难。通过横向和纵向对比验证了该模型的创新部分，变形图匹配主要针对知识图谱的稀疏性进行设计，横向选取一些模型，如KV-Mem[15]、GraftNet[16]、PullNet[17]、EmbedKG-QA[18]，在公共基准数据集上进行比较；从纵向进行消融实验，验证线索向量化模块和局部匹配模块发挥了应有的作用。

变形图匹配模型的构成主要有三大板块：①信息传递方面。为了能够使知识图谱中的实体和关联之间不断地更新彼此的表示，融入一个线索的向量化模块，能够获取线索的向量表征，实现信息的传递；②线索匹配方面。提出了局部匹配层，利用初始匹配队列和相似度字典，组合不同类型的线索，并对其进行表征；③构建整体子图方面。完善了整体图神经网络层，用联结阈值评定不同实体的关联程度，确定所在范围。

1 方 法

1.1 问题描述

KGQA的过程是在知识图谱中寻找问题q(主题实体E_t)的答案a。知识图谱$K=(\mathcal{V},\mathcal{E},\mathcal{R})$,其中$\mathcal{V}$、$\mathcal{E}$、$\mathcal{R}$分别为知识图谱中所有的实体、关联、关系集合；关联$\mathcal{E}$是三元组$(h,r,t)$集合，其中$h,t\in\mathcal{V}$，$r\in\mathcal{R}$，而$(e_h,e_r,e_t)$表示三元组在知识图向量嵌入表示。而对于关联预测中，经常引入评分函数$s=\varphi(e_h,e_r,e_t)$,用于判断关联是否为真。

在KGQA中利用NLP的分析手段F而获得到的各种问答启发性的信息或知识，本文统称为问答线索C；限于篇幅，本文选取代表性的分析手段作为研究范围，包括命名实体识别f_N、关键字提取f_K和依赖语法分析f_D。相应的问答线索分别为

$$C_N=\{n_1,\cdots,n_{|C_N|}\} \tag{1}$$

$$C_K=\{k_1,\cdots,k_{|C_K|}\} \tag{2}$$

$$C_D=\left\{[d,r],d:d_1,\cdots,d_{|C_D|},r:(d_i,d_j),\quad i,j\in|C_D|\right\} \tag{3}$$

式中：C_N和C_K为线状线索，C_D为树线状线索；n_i为命名实体识别后的实体；k_j为关键字提取后的关键字；d、r分别为问题的部分成分、依赖关系。

从整体形式上，有如下定义：

$$C=F(q,E_t) \tag{4}$$

$$F=\{f_N,f_K,f_D\} \tag{5}$$

式中：F、C分别为问答分析手段、问答线索的集合；q为所提出的问题；E_t为主题实体，故

$$C=\{C_N,C_K,C_D\} \tag{6}$$

由此而形成的队列，分别称为点状线索队列和树状线索队列，形式定义分别为

$$L_{\text{linear}}=\{[l,r_{\text{linear}}],l_1,\cdots,l_{|C_N|+|C_K|},l_i\in C_N\ \text{ or }\ C_K,\ r_{\text{linear}}:\left(l_i,l_j\right),i\in|C_N|+|C_K|\} \tag{7}$$

$$L_{\text{tree}}=\{[d,r_{\text{tree}}],d:d_1,\cdots,d_{|C_D|},\ r_{\text{tree}}:(d_i,d_j),i,j\in|C_D|)\} \tag{8}$$

式中：L_{linear}、L_{tree}分别为点状线索队列、树状线索队列；l为线索队列。

在KGQA中，从问题和主题实体出发，经过语义分析并构建问题子图，再映射到知识图谱子图中以寻找答案，这个过程可以看作一种特殊的图匹配：①规模不平衡。问题子图与知识图谱的节点及关联的规模存在巨大差异；②范围不明确。知识图谱子图中可能是集合V中的任意一个实体集合；③构成不确定。知识库子图具体包含实体及关联是不确定的，也不是唯一的。

为此，定义了变形图匹配的概念，旨在形式化地描述这个过程，展开解决思路。在本文中，将由问答线索C与知识图谱K之间映射F_{DGM}的过程，称为变形图匹配。形式上，有如下定义：

$$F_{\text{DGM}}:C^+\to K^+ \tag{9}$$

$$C^+=(C,\boldsymbol{A}(C),\boldsymbol{E}_{\text{node}}(C),\boldsymbol{E}_{\text{edge}}(C)) \tag{10}$$

$$K^+=(\mathcal{V},\mathcal{E},\boldsymbol{E}_{\text{node}}(\mathcal{V}),\boldsymbol{E}_{\text{edge}}(\mathcal{E})) \tag{11}$$

式中：C^+和K^+分别是问答线索C与知识图谱K扩展；

$\boldsymbol{A}(C)$为问答线索C的邻接矩阵；$\boldsymbol{E}_{\text{node}}(C)$、$\boldsymbol{E}_{\text{edge}}(C)$分别为问答线索中节点向量、边向量；$\mathcal{V},\mathcal{E}$分别为知识图谱中所有的实体、关联；$\boldsymbol{E}_{\text{node}}(\mathcal{V})$，$\boldsymbol{E}_{\text{edge}}(\mathcal{E})$分别表示知识图谱的节点、关联的向量。

1.2 整体框架

利用变形图匹配实现知识图谱问答的过程，本文分为3个步骤：线索向量化、局部匹配和整体拼图。基于变形图匹配的知识图谱问答模型如图1所示。通过子图的变形匹配，知识图谱问答模型可以将问题线索映射到知识图谱子图中，再锁定更为精确的答案范围。主要由3个部分组成：线索向量化模块，局部匹配模块和整体拼图模块。

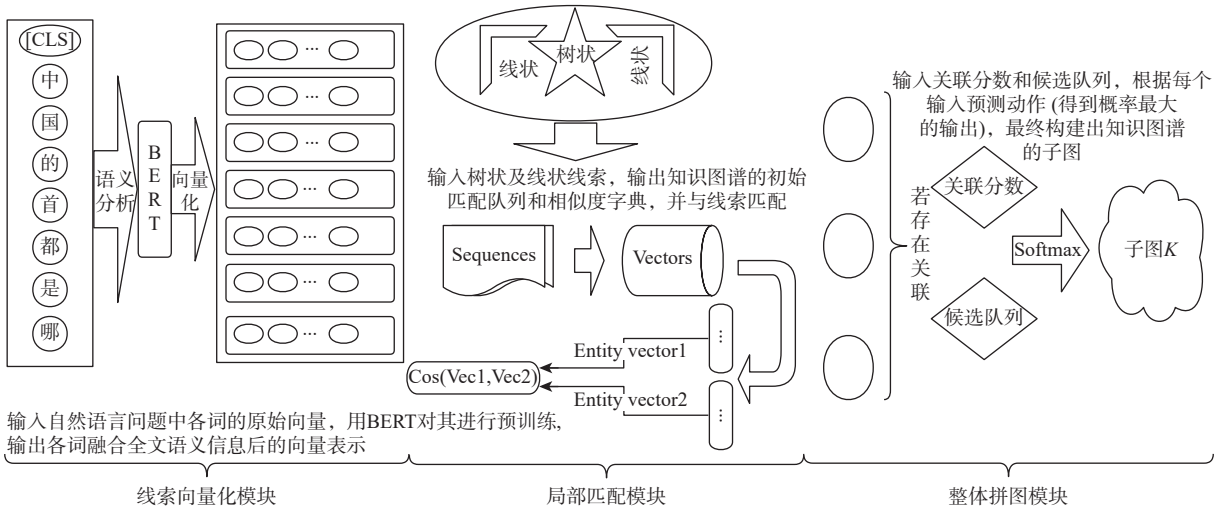


图 1　基于变形图匹配的知识图谱问答模型

Fig. 1　A knowledge graph Q&A model based on deformed graph matching

1.3　线索向量化模块

给定一个问题q(主题实体E_t), 知识图谱为$K=(\mathcal{V},\mathcal{E},\mathcal{R})$; 将问答线索扩展表示为$C^+$, 通过语义分析获得问答线索, 使用预训练语言模型 BERT 获得问题的向量表征, 再根据问题的向量表征获取问答线索相应的初始向量, 最后利用图神经网络 GNN 和注意力机制多次(n)更新问题线索的节点及关联, 实现信息传递的过程。

1.4　局部匹配模块

给定问题q(主题实体E_t), C^+和嵌入$(\mathcal{V},\mathcal{E})$; 匹配半径$\varnothing$, 需要输出知识图谱$K$的初始匹配队列mat和相似度字典sim, 合并线状和树状线索进行匹配, 匹配组合有 4 种 (n,n), (n,e), (e,n), (e,e),其中 n、e 分别为节点、关联(边), 将知识图谱中节点和关联(边)表征出来, 并使用线性函数和余弦函数进行相似度计算。最后, 在知识图谱中选择最匹配前$\varnothing$位的节点或关联, 即问题最相关的节点或关联。

1.5　整体拼图模块

在通过局部匹配找到匹配队列后, 需构建知识图谱K的子图。首先需要找出节点的类型判断 2 个节点(边)是否存在关联, 分别保存关联分数和关联候选队列, 然后根据节点间路径的长度, 使用关联预测方法计算相关性分数, 最大值大于联结阈值ϑ, 才有关联(此处自关联不在考虑范围内)。最后基于知识图谱构建子图, 以确定答案范围。

2　实验准备

2.1　数据集

MetaQA [19] 数据集如表 1 所示。该数据集包含 40 万个问题, 这些问题使用数十个模板生成, 最常见的问题是 3-hop, 其知识图谱来自电影领域, 将 wiki-movies 的问答从只有 1-hop 扩展到 2-hop, 再到 3-hop, 最终数据集包含问题超过 400 000 个, 其中包括 43 000 多个实体、9 个关系和 135 000 多个三元组。

表 1　MetaQA 数据集中不同跳数的问答对数量

Table 1　The number of question answering pairs with different hops in MetaQA dataset

MetaQA	训练集	验证集	测试集
1-hop	96 106	9 992	9 947
2-hop	118 948	14 872	14 872
3-hop	114 196	14 274	14 274

2.2　基线方法

为了证明本文模型的有效性, 在实验中选取了 4 种具有代表性的基线模型进行比较:

1) KV-Mem 基线模型[15]: 是一种基于 key-value 的记忆网络, 其使用 key 来查询相关的知识记忆, 迭代读取使用键值内存存储起来的知识, 然后进行多跳推理, 在回答问题时使用 value, 将所有答案映射到 embedding 空间, 与最后一次的问答作 softmax, 得到每一个答案的概率, 取概率最大的那个作为输出, 也可以在增加先验信息的时候更加灵活。

2) GraftNet 基线模型[16]: 通过启发式方法从一个基于特定问题的子图中提取答案, 然后使用图神经网络来推断答案。基于多源数据的问答, 特定的问题子图中包含: 文本、知识库、实体集合、关系, 以此实现早期融合的过程, 该模型的方法来源于图表示学习, 可以看作消息传递的神经网络框架, 使用无结构的文本数据作为结构化的知识图谱数据的补充; 对于无结构的文本数据, GraftNet 将每个文档看做一个节点, 并融入知识图谱结构中: 如果该文档中包含某个实体, 那么就在这个文档与实体之间建立连接。

3）PullNet基线模型[17]：通过学习用图CNN创建了一个特定于问题的子图，并不是用启发式的方法来检索子图，PullNet在模型GraftNet上做了改进，学习将问题中提取出来的实体和事实用于创建一个更相关的子图。

4）EmbedKGQA基线模型[18]：在3-hop任务上优于PullNet，将知识图谱问答作为链接预测任务，并合并知识图嵌入，并没有在解决KG稀疏性问题的时候依赖外部的文本数据，EmbedKGQA分别在两个数据集MetaQA和WebQuestionsSP上的实验表明，其可以在知识图谱缺失的时候更好地回答问题，在预测答案时获得了优秀的表现。

2.3 参数设置

本文选取的超参数范围如下：实体和关系嵌入维度{100,200}，学习率{0.000 5,0.001,0.01}，Batchsize{32,64,128}，子图对比约束超参数α{0.1,0.01,0.001}。

实验里，设置实体和关系维度为200，学习率为0.000 02，Epoch值为6～10，Dropout值为10%，Batchsize值为32，使用Adam优化器进行优化。

2.4 评估指标

本文使用总体性能作为评估指标，该指标能够对多跳问答的实验结果做出较为精确的判断，在一个基准数据集上用相同维度的实体和关系作为嵌入，同时评估其他现有模型的实验效果。答案选择需在子图里做出：

$$E_a = \text{argmax}_{x,r\in qa_\text{subgraph}}\phi(E_t,E_r,E_x) \tag{12}$$

3 实验分析

3.1 实验结果

在这样的知识图谱问答中，节点之间有完整的路径，没有遗漏的环节，这不符合现实，并且所提出的模型可以在不完整的知识图谱上工作。为了验证这个观点，本文模拟了一个不完整的知识图谱，即随机删除一半三元组（随机删除一个节点，概率为50%），这个设置在本文中称为KG-Half，而完整的知识图谱设置称为KG-Full，MetaQA数据集不同完整度的实验结果如表2所示。完整和不完整知识图谱中所有跳数如表3所示。由表3可知，本文模型优于过去的模型。

表2 MetaQA数据集不同完整度的实验结果

Table 2 Experimental results for MetaQA datasets with different degrees of completeness

模型	总体性能	
	KG-Full	KG-Half
GraftNet	89.1	57.9
PullNet	96.1	58.2
KV-Mem	74.0	45.8
EmbedKGQA	97.0	81.9
DGM-KGQA	98.2	90.0

表3 完整和不完整知识图谱中不同跳数的总体性能

Table 3 Overall performance of different hops in complete and incomplete knowledge graphs

模型	KG-Full			KG-Half		
	1-hop	2-hop	3-hop	1-hop	2-hop	3-hop
GraftNet	97.0	94.8	77.7	64.0	52.6	59.2
PullNet	97.0	99.9	91.4	65.1	52.1	59.7
KV-Mem	96.2	82.7	48.9	63.6	41.8	37.6
EmbedKGQA	97.5	98.8	94.8	83.9	91.8	70.3
DGM-KGQA	97.2	98.4	95.6	78.3	92.5	70.8

3.2 消融实验

线索向量化是通过图神经网络迭代得到知识图谱和与问题相关的线索向量[19]。这些向量可以直接映射到知识库的实体中，映射概率最高为答案。为了研究在多关系的知识图谱问答中引入局部匹配步骤的有效性，去除了局部匹配的步骤以进行消融研究。消融实验结果如表4所示。由表4可知，整体知识图谱优于局部性能，但局部匹配会增强节点的准确率，避免选择过大的搜索空间，这在多跳问答中起到关键作用。得到知识图谱中的

表4 本文模型的消融实验结果

Table 4 The ablation results of the model in this paper

Import	总体性能					
	KG-Full			KG-Half		
	1-hop	2-hop	3-hop	1-hop	2-hop	3-hop
-local matching	86.4	90.3	85.2	59.8	53.5	54.3
-Overall puzzle	93.2	95.7	91.5	72.4	85.8	64.5
DGM-KGQA(final)	97.2	98.4	95.6	78.3	92.5	70.8

注：-local matching, -Overall puzzle中的-为消融实验去掉该模块后的结果。

匹配队列后, 利用链接预测计算相似性分数, 根据知识图谱构造子图, 确定答案范围。加入整体拼图步骤后的实验结果也验证了这一步的重要性。

4　结　论

1) 本文提出了知识图谱多跳问答的变形图匹配, 这是一种利用问答线索和知识图谱之间映射的端到端问答模型, 其利用问题中的语义内涵, 构造了具有特征且相对较小的子图, 保证了数据量的可控性。

2) 模型融入了链接预测机制来处理知识图谱中缺少的链接, 利用评分函数对节点和节点间的关系做出更准确的判断。

3) 变形图匹配既有效地保证了知识库的完整性, 又解决了文本语料库不可用的难点。多跳问答的实验结果清楚地证明了本文模型相对于各种最新基准的优越性。

本文研究了三跳以内的问答, 关注了该模型在 MetaQA 数据集上的性能。在以后的实验中继续扩展到更多跳数及更多语言的问答任务中, 并在不同类型的数据集上对本文模型进行评估。

参考文献 (References)

[1] WANG R Z, YAN J C, YANG X K. Learning combinatorial embedding networks for deep graph matching [C]//2019 IEEE/CVF International Conference on Computer Vision. Piscatawa: IEEE Press, 2019: 3056-3065.

[2] KOCH O, KRIEGE N M, HUMBECK L. Chemical similarity and substructure searches [M]//Encyclopedia of Bioinformatics and Computational Biology. Amsterdam: Elsevier, 2019: 640-649.

[3] SHARAN R, IDEKER T. Modeling cellular machinery through biological network comparison[J]. Nature Biotechnology, 2006, 24: 427-433.

[4] SINGH R, XU J B, BERGER B. Global alignment of multiple protein interaction networks with application to functional orthology detection[J]. Proceedings of the National Academy of Sciences of the United States of America, 2008, 105(35): 12763-12768.

[5] ZHANG J W, YU P S. Multiple anonymized social networks alignment [C]//2015 IEEE International Conference on Data Mining. Piscataway: IEEE Press, 2015: 599-608.

[6] VENTO M, FOGGIA P. Graph matching techniques for computer vision[M]//BAI X, AHENG J, HANCOCK. Graph-based methods in computer vision. Hershey: IGI Global, 2012: 1-41.

[7] STAUFFER M, TSCHACHTLI T, FISCHER A, et al. A survey on applications of bipartite graph edit distance [C]// International Workshop on Graph-Based Representations in Pattern Recognition. Cham: Springer, 2017: 242-252.

[8] BUNKE H, SHEARER K. A graph distance metric based on the maximal common subgraph[J]. Pattern Recognition Letters, 1998, 19(3-4): 255-259.

[9] YAN J C, YIN X C, LIN W Y, et al. A short survey of recent advances in graph matching [C]//Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval. New York: ACM, 2016: 167-174.

[10] XU H T, LUO D X, ZHA H Y, et al. Gromov-wasserstein learning for graph matching and node embedding [EB/OL]. (2016-06-06)[2022-05-18]. http://arxiv.org/abs/1901.06003.

[11] ROTA BULÒ S, PELILLO M, BOMZE I M. Graph-based quadratic optimization: a fast evolutionary approach[J]. Computer Vision and Image Understanding, 2011, 115(7): 984-995.

[12] HEIMANN M, SHEN H M, SAFAVI T, et al. REGAL: representation learning-based graph alignment [C] //Proceedings of the 27th ACM International Conference on Information and Knowledge Management. New York: ACM, 2018: 117-126.

[13] DERR T, KARIMI H, LIU X R, et al. Deep adversarial network alignment [C]//Proceedings of the 30th ACM International Conference on Information & Knowledge Management. New York: ACM, 2021: 352-361.

[14] FEY M, LENSSEN J E, MORRIS C, et al. Deep graph matching consensus [EB/OL]. (2020-01-27)[2022-05-18]. https://arxiv.org/abs/2001.09621.

[15] MILLER A, FISCH A, DODGE J, et al. Key-value memory networks for directly reading documents [EB/OL]. (2016-10-10)[2022-05-18].https://arxiv.org/abs/1606.03126.

[16] SUN H T, DHINGRA B, ZAHEER M, et al. Open domain question answering using early fusion of knowledge bases and text [C]//Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Stroudsburg: Association for Computational Linguistics, 2018: 4231-4242.

[17] SUN H T, BEDRAX-WEISS T, COHEN W. PullNet: Open domain question answering with iterative retrieval on knowledge bases and text [C]//Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing. Stroudsburg: Association for Computational Linguistics, 2019: 2380-2390.

[18] SAXENA A, TRIPATHI A, TALUKDAR P. Improving multi-hop question answering over knowledge graphs using knowledge base embeddings [C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: Association for Computational Linguistics, 2020: 4498-4507.

[19] ZHANG Y Y, DAI H J, KOZAREVA Z, et al. Variational reasoning for question answering with knowledge graph [EB/OL]. (2017-11-27)[2022-05-18].https://arxiv.org/abs/1709.04071.

Multi-hop knowledge graph question answering based on deformed graph matching

LI Xiangyue[1], FANG Quan[2,*], HU Jun[2], QIAN Shengsheng[2], XU Changsheng[2]

(1. Zhengzhou University, Zhengzhou 450000, China;
2. National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China)

Abstract: Knowledge Graph Question Answering (KGQA) is a process in which a given natural language question is semantically understood and parsed, and then the knowledge graph is used to query and reason to get the answer. But knowledge graphs which lack links, bring many challenges to multi-hop question answering. Many methods ignore important path information to evaluate the correlation between paths and multi-relationship problems when using knowledge graph embeddings, and text corpora also limit the scalability of text-enhanced models. Due to the drawbacks of these existing approaches, the Multi-hop Knowledge Graph Question Answering Based on Deformed Graph Matching (DGM-KGQA) method is proposed. This method builds semantic subgraphs using both question and topic entities, which then match the local structure of the knowledge graph to determine the correct solution. The experimental results on the benchmark dataset MetaQA verify the effectiveness of DGM-KGQA. In comparison to PullNet and EmbedKGQA, the accuracy of the answers retrieved on the completed knowledge graph is 4.2% higher than that of PullNet and 0.8% higher than that of EmbedKGQA. The accuracy of the answers retrieved on half of the knowledge graphs is 11.1% higher than that of PullNet and 0.5% higher than that of EmbedKGQA. Experiments show that the proposed deformed graph matching model can effectively enhance the relevance of knowledge graphs and the answer accuracy of multi-hop question answering.

Keywords: natural language problems; lack of links; text corpora; multi-hop question answering; deformed graph matching

Received: 2022-05-18; **Accepted**: 2022-06-23; **Published Online**: 2022-08-15 12:26

URL: link.cnki.net/urlid/11.2625.V.20220812.1724.005

Foundation items: National Natural Science Foundation of China (62072456,62036012,62106262); Open Research Projects of Zhejiang Lab (2021KE0AB05)

* **Corresponding author**. E-mail: qfang@nlpr.ia.ac.cn

http://bhxb.buaa.edu.cn jbuaa@buaa.edu.cn
DOI: 10.13700/j.bh.1001-5965.2022.0377

基于 Dijkstra 算法的平滑路径规划方法

巩慧[1]，倪翠[1,*]，王朋[1,2]，程诺[1]

(1. 山东交通学院 信息科学与电气工程学院，济南 250357；2. 山东省科学院 自动化研究所，济南 250013)

摘 要：移动机器人在复杂环境下沿 Dijkstra 算法规划的路径运动时，由于所规划的路径存在转折点多、部分转折角度小等问题，导致移动机器人不得不频繁转向，甚至要暂停才能完成转向，严重影响机器人的工作效率。利用几何拓扑学方法，结合实际场景信息，提出一种基于 Dijkstra 算法的平滑路径规划方法。根据应用场景获取连续化地图，将连续化地图离散化后随机生成离散点阵，计算各点之间的欧氏距离，选取与各离散点距离较近、且连线不跨越障碍的多个点，将其连接并生成离散图。在离散图中利用 Dijkstra 算法搜索最优路径作为引导路径。当移动机器人沿引导路径运动时，结合实际场景信息，采用几何拓扑学计算出移动机器人每一时刻应该采取的最佳动作和运行路线。实验结果表明：所提方法能够有效减少移动机器人运动中的累计转弯角度，增大最小平均转折角度，提高所规划路径的平滑度，从而缩短移动机器人的运动时间，提升机器人的工作效率。

关 键 词：路径规划；Dijkstra 算法；引导路径；路径平滑度；最优路径
中图分类号：TP391.4
文献标志码：A **文章编号**：1001-5965(2024)02-0535-07

随着移动机器人技术的兴起与发展，移动机器人在许多领域都发挥着重要作用，如智能装配、物流运输等。在移动机器人应用中路径规划作为关键技术之一，其主要目标是使移动机器人在应用场景中找到一条从运动起点到目标点的安全无碰撞路径[1]。优秀的路径规划方法可以减少硬件磨损、节约资源、提高移动机器人的运动速度、提升作业效率，为移动机器人在各行业中的应用提供有利保证[2-4]。

路径规划的经典算法包括人工势场法、A*算法、蚁群算法、遗传算法和 Dijkstra 算法等[5-9]。人工势场法规划的路径一般较为平滑，收敛速度快，但不适用于复杂环境[10]。A*算法受到估计代价值的影响，空间需求大，并且很难找到最优路径[11]。蚁群算法具有较强的鲁棒性，但是存在计算量大、易陷入局部最优等问题[12]。遗传算法具有较强的全局搜索能力，鲁棒性高，但是后期搜索速度较为缓慢，训练时间长[13]。Dijkstra 算法作为一种贪心算法，以起始点为中心原点，逐层展开搜索，直到搜索出从中心原点到所有节点的最短路径[14]。其优点是算法简单方便，可以准确获得全局最优路径，但是偶尔会出现冗余的拐点，效率低[15]。上述算法普遍存在移动机器人运动时间长、工作效率低，且未考虑路径平滑问题。李全勇等[16]在 Dijkstra 算法的基础上引入了路径平滑度，通过时间权重优化目标求解最优路径，减少了移动机器人的运行时间，不过仍存在冗余拐点问题。Zhang 等[17]在 A*算法的基础上，提出一种结合人工势场法的新启发式函数，其中包括距离和障碍物信息，用于自动陆地车辆的路径规划，但是也没有考虑到实际转弯半径的大小。Luan 和 Thinh[18]在混合遗传算法的基础上结合三次贝塞尔曲线为移动机器人提供平滑路径，

收稿日期：2022-05-18；**录用日期**：2022-06-23；**网络出版时间**：2022-09-15 11:22
网络出版地址：link.cnki.net/urlid/11.2625.V.20220914.1950.002
基金项目：中国博士后科学基金 (2021M702030)；山东省交通运输厅科技计划项目 (2021B120)
* **通信作者**. E-mail：1149993471@qq.com

引用格式：巩慧，倪翠，王朋，等. 基于 Dijkstra 算法的平滑路径规划方法 [J]. 北京航空航天大学学报，2024，50（2）：535-541.
GONG H，NI C，WANG P，et al. A smooth path planning method based on Dijkstra algorithm [J]. Journal of Beijing University of Aeronautics and Astronautics，2024，50（2）：535-541（in Chinese）.

但是当移动机器人距离障碍物过近时，需要移动贝塞尔曲线中的相应控制点，这会导致规划好的整条路径都发生改变，无法保证得到最优路径。

本文在Dijkstra算法的基础上充分考虑路径平滑度对移动机器人运动的影响，提出一种基于Dijkstra算法的平滑路径规划方法，通过几何拓扑学方法计算移动机器人在每一时刻应采取的最佳动作和运行路线，提升移动机器人的工作效率。

1 相关工作

1.1 移动机器人与模拟地图的建立

移动机器人的底盘根据转向方式可以分为差速底盘、全向轮底盘、麦克纳姆轮底盘和阿克曼底盘等。本文所用的移动机器人和模拟环境下所用移动机器人底盘均为阿克曼底盘，其特点是移动机器人的车轮系统由4个车轮构成，前后各2个，后轮负责驱动，前轮负责转向。因此，阿克曼底盘的移动机器人在转向时存在最小旋转半径。

移动机器人进行运动规划时需要先采集环境信息并建立地图模型，然后在建立的地图模型上进行路径规划。建立的地图模型主要包括栅格地图和矢量地图2类。栅格地图是按照一定间距排列的一系列阵列数据，其内容和色彩要与实际地形图一致。矢量地图是将与地形图内容一致的空间进行矢量化而得到的地图。本文采用矢量地图，根据真实场景选用尺寸为9×5像素（随机生成尺寸）的迷宫网格，采用Prim算法生成模拟地图，如图1所示。其中，黑色线段代表障碍，白色区域为机器人的活动区域。地图生成表述式为

$$\boldsymbol{M}=\mathrm{Prim}(W,H) \tag{1}$$

式中：$\boldsymbol{M}$为迷宫矩阵；Prim为矩阵生成函数；W为模拟地图的宽度；H为模拟地图的高度。

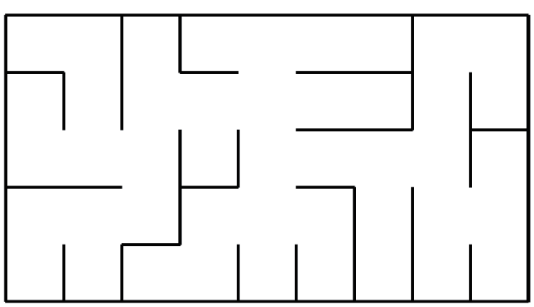


图1 Prim算法生成的模拟地图

Fig. 1 Simulation map generated by Prim's algorithm

1.2 基于Dijkstra算法的路径规划

Dijkstra算法是求解最短路径问题的经典算法。其核心思想是以起始点为原点，逐层向外扩展，直至到达终点。将Dijkstra算法应用到移动机器人路径规划中，为降低计算复杂度，将连续环境下的路径规划任务转化为离散环境下的路径规划任务，以简化路径规划过程。

首先，将模拟地图离散化，随机生成n个离散点，n的计算式如(2)所示，并保证各离散点之间的欧氏距离不小于32像素（根据地图尺寸计算），其目的是使点相对均匀地充满整个空间，避免出现一大堆点聚集在一起，无法连接到终点的现象。最终生成的离散点分布如图2所示。

$$n=\frac{4WH}{D^2} \tag{2}$$

式中：D为模拟地图中迷宫网格的个数。

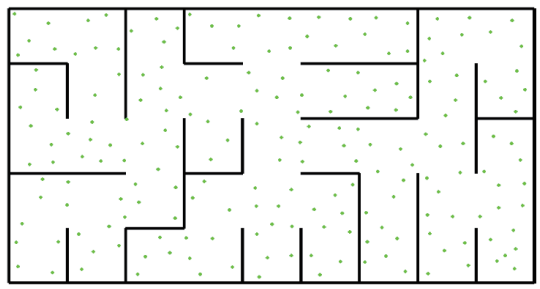


图2 模拟地图中随机生成的离散点示意图

Fig. 2 A sketch of randomly generated discrete points on a simulated map

然后，将n个离散点逐个与其距离最近、但不跨越障碍的m个点（本文取3个）连接，形成离散图。并将移动机器人的运动起点和终点添加到离散图中，如图3所示。其中，黑色“■”号代表移动机器人的起点，红色圆点代表终点。

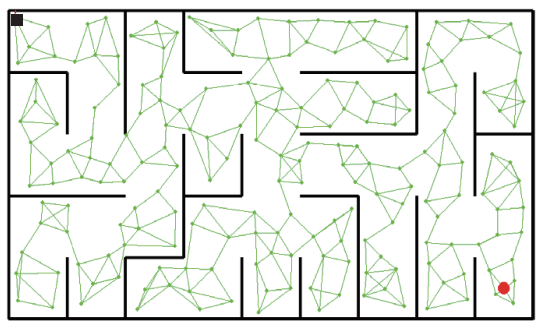


图3 由离散点生成的离散图

Fig. 3 A discrete graph generated from discrete points

最后，采用Dijkstra算法在离散图中搜索，以获得移动机器人从运动起点到终点的路径，如图4所示。

由图4可知，传统Dijkstra算法在离散图中进行路径规划时，生成的路径中转折点较多，累计转弯角度过大。实际应用时，移动机器人在沿着路径前进过程中需要频繁停下来转向，导致运行时间延长，降低了移动机器人的工作效率。并且路径中的

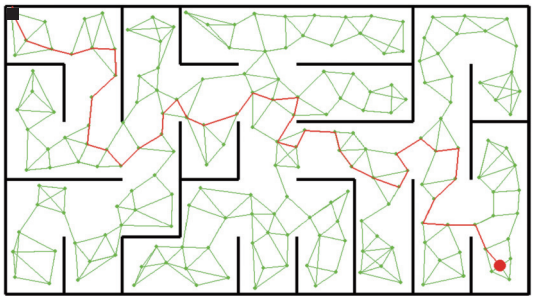


图 4　Dijkstra 算法在离散图中规划的路径

Fig. 4　A path planned by Dijkstra's algorithm in a discrete graph

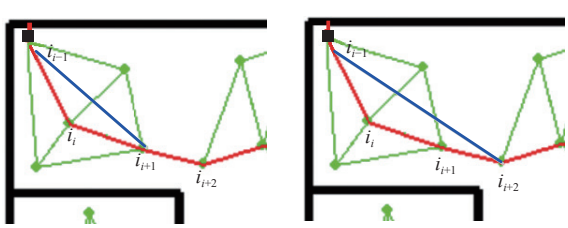
某些转折角度过小, 实际应用时, 可能无法满足移动机器人的转弯半径要求, 难以达到预期的转弯效果, 造成移动机器人无法按所规划的路径行走。

2　平滑路径规划方法

采用几何拓扑学方法, 通过计算移动机器人在每一时刻应采取的最佳动作, 对 Dijkstra 算法规划的路径进行优化。

获取引导路径。对 Dijkstra 算法规划的路径进行遍历。图 5 为引导路径获取方法, 设路径中的节点集合 $I=\{i_1,i_2,\cdots,i_n\}$, 当前节点为 i_i。先判断下一节点 i_{i+1} 是否为空, 若为空, 则表示该地图的引导路径构建完成; 若不为空, 则判断当前节点 i_i 是否需要删除, 如果 i_i 的前一个节点 i_{i-1} 与后一个节点 i_{i+1} 相连, 连线 (i_{i-1},i_{i+1}) 不会穿过障碍, 则删除当前节点 i_i, 将其前后节点 i_{i-1} 与 i_{i+1} 相连接, i_{i-1} 与 i_{i+1} 成为相邻节点, (i_{i-1},i_{i+1}) 成为现阶段的引导路径, 如图 5(a) 所示。然后, 将 i_{i+1} 作为当前节点, 连接 i_{i-1} 与 i_{i+2}, 若连线 (i_{i-1},i_{i+2}) 不会穿过障碍物, 则将节点 i_{i+1} 删除, 现阶段引导路径为 (i_{i-1},i_{i+2}), 如图 5(b) 所示。以此类推, 完成整个地图的引导路径构建, 如图 6 所示, 图中红色线表示 Dijkstra 算法规划的路径, 蓝色线表示获取的引导路径。

令移动机器人沿引导路径运动, 结合移动机器人周围实际场景信息, 通过几何拓扑学方法计算移动机器人在每一时刻应采取的最佳动作; 如图 7 所示, 假设移动机器人的运动路径为 $A\rightarrow B\rightarrow C$, 其中, A、B、C 这 3 点的坐标分别为 (x_1,y_1)、(x_2,y_2)、(x_3,y_3),



(a) 第1阶段引导路径　　(b) 第2阶段引导路径

图 5　引导路径获取方法

Fig. 5　Method for obtaining boot path

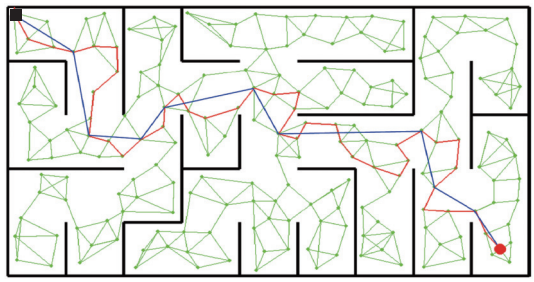


图 6　获取的引导路径示意图

Fig. 6　Acquired boot path diagram

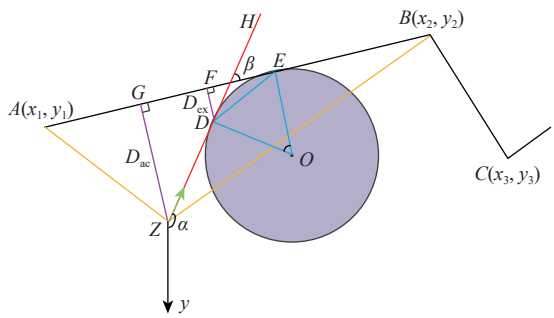


图 7　移动机器人行走轨迹示意图

Fig. 7　Mobile robot walking track diagram

移动机器人从当前位置 Z 点沿绿色箭头方向前进, 其前进方向所在直线 ZH 与目标路径 AB 之间的夹角 β 为

$$\beta=\arctan\frac{x_2-x_1}{y_2-y_1}-\alpha \tag{3}$$

式中: α 为机器人前进方向与 y 轴正方向的夹角。

存在一个圆心 O, 半径 R 的圆, 分别与机器人前进方向所在的直线 ZH 和目标路径 AB 相切。切点分别为 D 和 E。半径 R 即为移动机器人的最小转弯半径。过点 D 和点 Z 分别做直线 AB 的垂线, 交 AB 于点 F 和 G, 移动机器人转向时, 与目标路径之间的预期距离为线段 DF 的长度, 记为 D_{ex}, 计算式如 (4) 所示。与目标路径之间的实际距离为线段 ZG 的长度, 记为 D_{ac}, 计算式如 (5) 所示。

$$D_{ex}=2R\cos\left(90-\frac{\beta}{2}\right)\sin\frac{\beta}{2} \tag{4}$$

$$D_{ac}=\frac{x_1y_2+x_2y_0+x_0y_1-x_1y_0-x_2y_1-x_0y_2}{\sqrt{(x_1-x_2)^2+(y_1-y_2)^2}} \tag{5}$$

式中: x_0、y_0 为机器人所在点的横、纵坐标。

根据计算得出的 D_{ex} 和 D_{ac} 的值, 确定移动机器人的当前状态和应选择的最佳运动方案。

1) 当 $D_{ac}>D_{ex}$ 时, 表示移动机器人未到达预期的转向点, 此时应保持继续前进。

2) 当 $D_{ac}=D_{ex}$ 时, 表示移动机器人到达预期的转向点, 此时, 若 $\beta>0$, 则右转, 反之, 则左转。

3) 当 $D_{ac}<D_{ex}$ 时, 表示移动机器人已经超过预

期转向点，此时需要进行路径矫正。

矫正路径如图8所示。假设移动机器人的行走路径为$A\to B\to C$，A、B、C这3点的坐标分别为(x_1,y_1)、(x_2,y_2)、(x_3,y_3)，点Z表示移动机器人当前位置，绿色箭头表示移动机器人前进方向，点N为下一条路径的预期转向起点。为了保证机器人能够在下一条路径的预期转向起点之前完成矫正，就需要计算机器人目前矫正路径的终点X的坐标。移动机器人在矫正路径的时候会经过3个圆弧，分别对应图中的圆O_1、圆O_2(半径为$2R$)、圆O_3中的红色部分，最后通过圆O_4中的圆弧进入下一条路径。

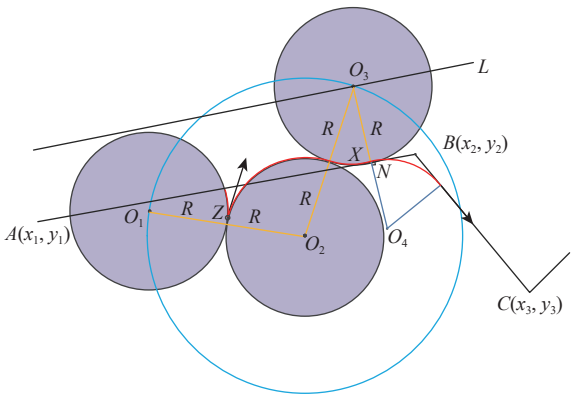


图8 移动机器人矫正路径示意图

Fig. 8 Path correction diagram of mobile robot

计算圆O_2的圆心坐标(x_{O_2},y_{O_2})，计算式为

$$x_{O_2}=x_0-R\sin\left(\alpha-\frac{\pi}{2}\right) \tag{6}$$

$$y_{O_2}=y_0-R\sin\left(\alpha-\frac{\pi}{2}\right) \tag{7}$$

过圆O_3的圆心做直线AB的平行线L，其函数形式如式(8)所示。以O_2为圆心，O_2O_3为半径作圆，如图8中所示圆O_2，函数表达如式(9)所示。

$$y=\frac{y_2-y_1}{x_2-x_1}(x-x_1-R\sin\alpha)+y_1+R\cos\alpha \tag{8}$$

$$(x-x_{O_2})^2+(y-y_{O_2})^2=(2R)^2 \tag{9}$$

通过式(8)和式(9)求O_3的圆心坐标。当$D_{ac}<D_{ex}$时，图8中的圆O_2与直线L存在2个交点，其横坐标函数表达式分别为

$$x_a=\frac{-u+\sqrt{u^2-4(1+k^2)v}}{2(1+k^2)} \tag{10}$$

$$x_b=\frac{-u-\sqrt{u^2-4(1+k^2)v}}{2(1+k^2)} \tag{11}$$

式中：x_a、x_b分别为2个交点的横坐标；k为直线L的斜率，u为一次项系数，v为常数项系数，其计算式为

$$k=\frac{y_2-y_1}{x_2-x_1} \tag{12}$$

$$u=-2k^2(x_1+R\sin\alpha)+2k(y_1+R\cos\alpha)-2y_0k-2x_0 \tag{13}$$

$$v=k^2x^2-2k(x_1+R\sin\alpha)(y_1+R\cos\alpha)+y^2+2y_0k(x_1+R\sin\alpha)-2y_0(y_1+R\cos\alpha)+y_0^2+x_0^2-4R^2 \tag{14}$$

根据式(10)和式(11)求出的x_a,x_b，可以计算得到O_3的圆心坐标(x_{O_3},y_{O_3})，计算式为

$$x_{O_3}=\begin{cases}x_a & \dfrac{x_a-x_b}{x_2-x_1}>0\\ x_b & \dfrac{x_a-x_b}{x_2-x_1}\leqslant 0\end{cases} \tag{15}$$

$$y_{O_3}=k(x_{O_3}-x_1-R\sin\alpha)+y_1+R\cos\alpha \tag{16}$$

根据圆O_3的圆心坐标计算移动机器人进入目标轨迹AB的位置点$X(x_X,y_X)$，计算式为

$$x_X=x_{O_3}+R\sin\alpha \tag{17}$$

$$y_X=y_{O_3}-R\cos\alpha \tag{18}$$

计算点X和点N到目标点A的距离，分别记为实际矫正距离d_1和预期矫正距离d_2，计算式为

$$d_1=\sqrt{(x_X-x_1)^2+(y_X-y_1)^2} \tag{19}$$

$$d_2=\sqrt{(x_N-x_1)^2+(y_N-y_1)^2} \tag{20}$$

式中：x_N和y_N分别表示点N的横坐标和纵坐标。

根据d_1和d_2即可判断机器人应该采取的动作，规则如下：

1) 当$\beta>0$，$d_1>d_2$时，移动机器人左转倒车；

2) 当$\beta<0$，$d_1>d_2$时，移动机器人右转倒车；

3) 当$\beta>0$，$d_1<d_2$时，移动机器人右转前进；

4) 当$\beta<0$，$d_1<d_2$时，移动机器人左转前进。

在离散图中运用Dijkstra算法进行路径规划时，虽然离散点的个数越多路径越趋于平滑，但当离散点的个数增多时，计算量也会显著增大，不利于实际应用。本文综合考虑路径平滑度和计算效率，选取n个离散点，将Dijkstra算法求解得到的路径作为引导路径，结合几何拓扑学方法，计算出移动机器人在每一时刻应采取的最佳动作，这样既能保证规划出的路径更加平滑，又能兼顾移动机器人的工作效率。

3 实验结果与分析

为了验证本文方法的有效性，实验选择Windows 10操作系统下的PyCharm平台，利用TensorFlow 2.1框架实现。

本实验根据真实环境随机生成20幅模拟地图，以及20组移动机器人的起始点坐标和目标点坐标，然后分别利用传统Dijkstra算法和本文方法进行路径规划，图9为部分路径规划的结果。其中红色线为传统Dijkstra算法规划的路径，蓝色线为

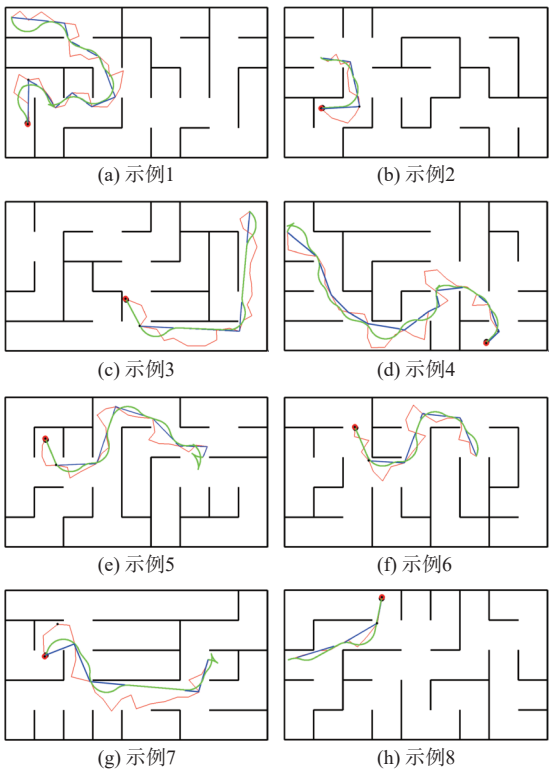


图 9　路径规划部分结果示意图

Fig. 9　Schematic diagram of partial result of path planning

优化后的引导路径, 绿色线为本文方法规划的路径。由图可知, 本文方法所规划的路径更加平滑, 能够有效降低移动机器人的转向频率, 避免移动机器人在运动过程中因转折角度过小导致暂停转向的问题, 保证了移动机器人运动的连续性和平稳性。

移动机器人运动过程中所转过的累计角度是评价一个路径规划方法所规划的路径是否平滑的重要指标。在上述随机生成的 20 幅模拟地图中, 分别对本文方法与传统 Dijkstra 算法、A*算法及文献 [17]、文献 [18] 改进的算法在累计转弯角度方面进行了比较, 结果如图 10 所示。由图可知, 本文方法累计转弯角度最小, 其次为文献 [18] 改进的算法, 再次为文献 [17] 改进的算法, Dijkstra 算法和

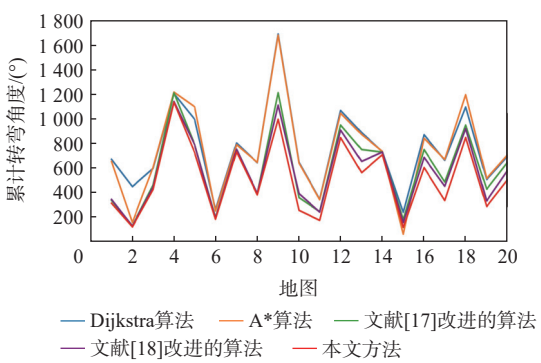


图 10　累计转弯角度比较

Fig. 10　Comparison of cumulative turning angles

A*算法的累计转弯角度最大。本文方法与传统 Dijkstra 算法相比, 平均累计转弯角度减少了 31.82%, 比 A*算法减少了 31.15%, 比文献 [17] 改进的算法减少了 14.15%, 比文献 [18] 改进的算法减少了 9.563%。结果证明本文方法所规划的路径最为平滑。

如果规划的路径中存在转折角度过小的情况, 可能无法满足移动机器人转弯半径要求, 造成机器人无法按照所规划的路径运动。将本文方法分别与传统 Dijkstra 算法、A*算法及文献 [17], 文献 [18] 改进的算法在最小转折角度方面进行了比较, 结果如图 11 所示。由图可知, 本文方法规划的路径中最小转折角度最大, 其次为文献 [17] 和文献 [18] 改进的算法, A*算法和传统 Dijkstra 算法的最小转折最小。本文方法与传统 Dijkstra 算法相比, 最小平均转折角度增大了 54.18%, 比 A*算法增大了 51.13%, 比文献 [17] 改进的算法增大了 17.63%, 比文献 [18] 改进的算法增大了 8.65%。结果证明本文方法规划的路径能有效避免因转折角度过小导致移动机器人无法按照规划路径行走的问题。

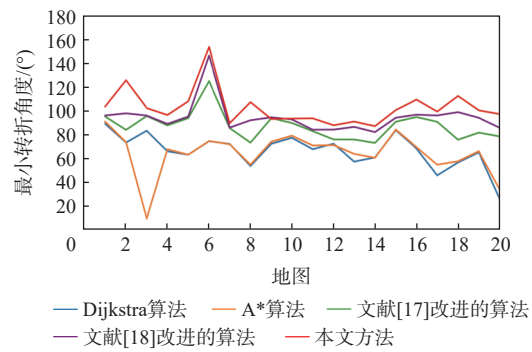


图 11　最小转折角度比较

Fig. 11　Comparison of minimum turning angles

在路径规划中, 移动机器人运动路程的总和称为总路程, 总路程与起点到终点的距离之比可以用于评价一个方法在路径规划方面的最优性, 以及该方法进行路径规划所带来的收益。将本文方法与传统 Dijkstra 算法、A*算法及文献 [17]、文献 [18] 改进的算法进行比较, 计算移动机器人运动的总路程与起点到终点的距离之比, 结果如图 12 所示。由图可知, 在 20 幅模拟地图中, 传统 Dijkstra 算法的平均距离比值为 1.192, A*算法的平均距离比值为 1.171 5, 文献 [17] 改进的算法平均距离比值为 1.178, 文献 [18] 改进的算法平均距离比值为 1.172 5, 本文方法平均距离比值为 1.174 5。本文方法所规划的总路程比传统 Dijkstra 算法、A*算法和文献 [18] 改进的算法长, 这是为了提高移动机器人运动的平滑性和连续性而增大转弯半径的结果。

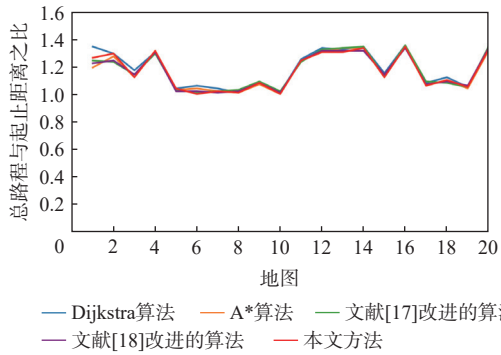


图 12 总路程与起止距离之比

Fig. 12 Ratio of total distance to start-stop distance

此外，移动机器人沿规划路径的运动时间也是评价一个路径规划方法好坏的重要指标。在移动机器人运动时间方面分别对本文方法、传统 Dijkstra 算法、A*算法及文献 [17]、文献 [18] 改进的算法进行了比较，结果如图 13 所示。由图可知，在 20 幅模拟地图中，本文方法移动机器人的运动时间最少，其次为文献 [17] 和文献 [18] 改进算法，A*算法和传统 Dijkstra 算法移动机器人的运动时间最多。与传统 Dijkstra 算法相比，本文方法移动机器人运动时间比传统 Dijkstra 算法缩短了 27.43%，比 A*算法缩短了 19.41%，比文献 [17] 改进的算法缩短了 7.75%，比文献 [18] 改进的算法缩短了 7.45%。由此可知，本文方法在移动机器人工作效率方面具有优势。

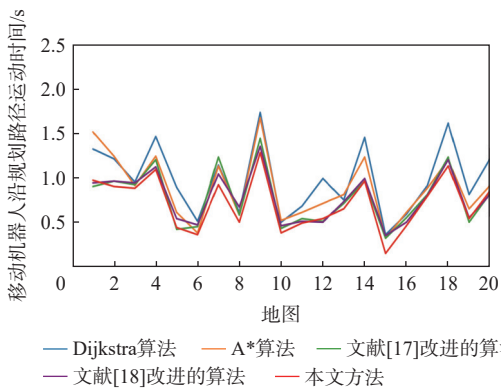


图 13 移动机器人沿规划路径运动时间比较

Fig. 13 Comparison of motion time of mobile robot along planned path

4 结 论

1）采用几何拓扑学方法，计算出移动机器人在每一时刻应该采取的最佳动作和运行路线，有效避免了路径规划过程中转折点过多导致移动机器人频繁暂停再转向，以及转折角度过小导致移动机器人无法按所规划的路径行走等问题。

2）实验结果表明，本文方法与其他算法相比，所规划的路径累计转弯角度更小，最小转折角度更大，移动机器人的运动路径和时间更短，能够有效改善移动机器人运动的平滑性和连续性，提升移动机器人的工作效率，更加符合实际生产生活的应用。

参考文献（References）

[1] LU H W. Artificial intelligence and robotics[M]. Berlin: Springer, 2021: 267-275.

[2] KO J H. An adaptive path-planning for intelligent AGV system[J]. Journal of the Institute of Electronics and Information Engineers, 2017, 54(4): 115-121.

[3] CHEN X, ZHAO M Y, YIN L Y. Dynamic path planning of the UAV avoiding static and moving obstacles[J]. Journal of Intelligent & Robotic Systems, 2020, 99(3-4): 909-931.

[4] 殷建军, 董文龙, 梁利华, 等. 复杂环境下农业机器人路径规划优化方法[J]. 农业机械学报, 2019, 50(5): 17-22.
YIN J J, DONG W L, LIANG L H, et al. Optimization method of agricultural robot path planning in complex environment[J]. Transactions of the Chinese Society for Agricultural Machinery, 2019, 50(5): 17-22(in Chinese).

[5] 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.

[6] ZHONG X Y, TIAN J, HU H S, et al. Hybrid path planning based on safe A* algorithm and adaptive window approach for mobile robot in large-scale dynamic environment[J]. Journal of Intelligent & Robotic Systems, 2020, 99(1): 65-77.

[7] TIAN X E, LIU L, LIU S A, et al. Path planning of mobile robot based on improved ant colony algorithm for logistics[J]. Mathematical Biosciences and Engineering, 2021, 18(4): 3034-3045.

[8] 魏彤, 龙琛. 基于改进遗传算法的移动机器人路径规划[J]. 北京航空航天大学学报, 2020, 46(4): 703-711.
WEI T, LONG C. Path planning for mobile robot based on an improved genetic algorithm[J]. Journal of Beijing University of Aeronautics and Astronautics, 2020, 46(4): 703-711(in Chinese).

[9] 车建涛, 高方玉, 解玉文, 等. 基于 Dijkstra 算法的水下机器人路径规划[J]. 机械设计与研究, 2020, 36(1): 44-48.
CHE J T, GAO F Y, XIE Y W, et al. Path planning of underwater robots based on the Dijkstra algorithm[J]. Mechanical Design and Research, 2020, 36(1): 44-48(in Chinese).

[10] 邓叶, 姜香菊. 基于改进人工势场法的四旋翼无人机航迹规划算法[J]. 传感器与微系统, 2021, 40(7): 130-133.
DENG Y, JIANG X J. Four-rotor UAV trark planning algorithm based on improved artificial potential field method[J]. Transducer and Microsystem Technologies, 2021, 40(7): 130-133(in Chinese).

[11] TANG G, TANG C Q, CLARAMUNT C, et al. Geometric A-star algorithm: An improved A-star algorithm for AGV path planning in a port environment[J]. IEEE Access, 2021, 9: 59196-59210.

[12] HE C W, MAO J A. AGV optimal path planning based on improved ant colony algorithm[J]. EDP Sciences, 2018, 232: 03052.

[13] LI Y H, HUANG Z H, XIE Y. Path planning of mobile robot based on improved genetic algorithm[C]//2020 3rd International Confer-

ence on Electron Device and Mechanical Engineering. Piscataway: IEEE Press, 2020: 691-695.

[14] SUN Y H, FANG M, SU Y X. AGV path planning based on improved Dijkstra algorithm[J]. Journal of Physics, 2021, 1746(1): 012052.

[15] ZHU D D, SUN J Q. A new algorithm based on Dijkstra for vehicle path planning considering intersection attribute[J]. IEEE Access, 2021, 9: 19761-19775.

[16] 李全勇, 李波, 张瑞, 等. 基于改进 Dijkstra 算法的 AGV 路径规划研究[J]. 机械工程与自动化, 2021(1): 23-25.

LI Q Y, LI B, ZHANG R, et al. Research on AGV path planning based on improved Dijkstra algorithm[J]. Mechanical Engineering and Automation, 2021(1): 23-25(in Chinese).

[17] ZHANG J, WU J, SHEN X, et al. Autonomous land vehicle path planning algorithm based on improved heuristic function of A-Star[J]. International Journal of Advanced Robotic Systems, 2021, 18(5): 172988142110427.

[18] LUAN P G, THINH N T. Hybrid genetic algorithm based smooth global-path planning for a mobile robot[J]. Mechanics Based Design of Structures and Machines, 2023, 51(3): 1758-1774.

A smooth path planning method based on Dijkstra algorithm

GONG Hui[1], NI Cui[1,*], WANG Peng[1,2], CHENG Nuo[1]

(1. College of Information Science and Electrical Engineering, Shandong Jiao Tong University, Jinan 250357, China;
2. Institute of Automation, Shandong Academy of Sciences, Jinan 250013, China)

Abstract: When the mobile robot moves along the path planned by the Dijkstra algorithm in a complex environment, due to the planned path having many turning points and some turning angles being small, the mobile robot has to turn frequently or even pause to complete the turning, which seriously affects the working efficiency of the robot. In this study, the mobile robot's actual scene data is combined with the geometric topology method to propose smooth path planning method based on Dijkstra algorithm. The continuous map is obtained according to the application scenario, and the discrete lattice is randomly generated after the discretization of the continuous map, and the Euclidean distance between the points is calculated. Multiple points which are close to the discrete points and whose connection does not cross the barrier are selected to connect them and generate the discrete graph. The Dijkstra algorithm is used to search the optimal path as the guidance path in the discrete graph. The geometric topology is utilized to determine the optimum action and the running path that the mobile robot should follow at each time as it proceeds along the guidance path in conjunction with the actual scene information. Experimental results show that the proposed method can effectively reduce the cumulative turning angles, increase the minimum average turning angle, and improve the smoothness of the planned path, thus shortening the movement time of the mobile robot and improving the working efficiency of the robot.

Keywords: path planning; Dijkstra algorithm; guide path; path smoothness; optimal path

Received: 2022-05-18; **Accepted**: 2022-06-23; **Published Online**: 2022-09-15 11:22

URL: link.cnki.net/urlid/11.2625.V.20220914.1950.002

Foundation items: China Postdoctoral Science Foundation (2021M702030); Science and Technology Plan Project of Transportation Department of Shandong (2021B120)

* **Corresponding author**. E-mail: 1149993471@qq.com

http://bhxb.buaa.edu.cn jbuaa@buaa.edu.cn
DOI: 10.13700/j.bh.1001-5965.2022.0384

语言引导的多粒度特征融合目标分割方法

谭荃戈，王蓉*，吴澳
(中国人民公安大学 信息网络安全学院，北京 100091)

摘　　要：语言引导的目标分割旨在将文本描述的目标与其所指代的实体进行匹配，从而实现对文本、实体之间关系的理解与指代目标的定位。该任务在信息抽取、文本分类、机器翻译等应用场景中具有重要的应用价值。基于Refvos模型提出一种语言引导的多粒度特征融合目标分割方法，能够对特定目标精准定位。利用Swin Transformer和Bert网络，分别提取多粒度的视觉特征和文本特征，提高对整体与细节的表征能力；将文本特征分别与不同粒度视觉特征进行融合，通过语言引导增强特定目标表达；通过卷积长短期记忆网络对多粒度融合特征进行优化，在不同粒度特征间进行信息交流，得到更精细化的分割结果。在UNC、UNC+、G-Ref、ReferIt数据集上进行训练并测试所提方法。实验结果表明：相比Refvos，所提方法在UNC数据集的val、testB子集中IoU结果分别提升0.92%、4.1%，在UNC+数据集的val、testA、testB子集中IoU结果分别提升1.83%、0.63%、1.75%。所提方法在G-Ref、ReferIt数据集的IoU结果分别为40.16%和64.37%，达到前沿水平，证明所提方法的有效性与先进性。

关　键　词：目标分割；指代分割；特征融合；跨模态；文本理解
中图分类号：V221+.3; TB553
文献标志码：A　　**文章编号**：1001-5965(2024)02-0542-09

目标分割是对图像中关注区域或具有特有属性的某类、某个目标进行分割定位的任务，主要包括语义分割、实例分割、语言引导的目标分割等具体技术。语义分割是对图像中的各类目标进行分割。实例分割是对图像中的各个目标进行分割。语言引导的目标分割，是对语言描述的特定目标进行分割，该任务需要同时对视觉模态和语言模态进行学习，在人物定位、人机交互、智能图像编辑等方面有良好的应用前景。语言引导的目标分割既需要关注整体特征，通过语言引导正确的目标，又需要关注局部特征，达到精确分割的目的。因此，如何提取不同粒度的特征并将视觉特征和文本特征进行有效合理的对齐与融合是该任务的主要难点。

Hu等[1]较早提出了语言引导的目标分割方法。利用卷积神经网络提取视觉特征，利用循环神经网络提取文本特征，然后将图像、文本特征进行拼接、对齐，再利用全卷积神经网络得到语言引导的目标分割结果。后续的目标分割方法围绕语言、视觉特征的提取方法和跨模态特征融合方法进行了改进。Liu等[2]利用记忆增强长短期记忆网络提取文本特征，并将各词汇的文本特征按顺序与视觉特征进行拼接融合来提高语言与视觉特征的融合效果。Margffoy-Tuay等[3]引入动态滤波器，将直接进行多模态拼接的特征融合方式改进为利用特征图进行匹配的方式，此外，还使用简单循环单元(simple recurrent unit, SRU)[4]提升文本特征提取效率，对文本特征与视觉特征进行了更合理的匹配。递归细化网络[5]利用卷积神经网络固有的特征金

收稿日期：2022-05-18；**录用日期**：2022-06-23；**网络出版时间**：2022-10-25 10:03
网络出版地址：link.cnki.net/urlid/11.2625.V.20221024.1456.001
基金项目：国家自然科学基金 (62076246)
*通信作者. E-mail：dbdxwangrong@163.com

引用格式：谭荃戈，王蓉，吴澳. 语言引导的多粒度特征融合目标分割方法[J]. 北京航空航天大学学报，2024，50(2)：542-550.
TAN Q G，WANG R，WU A. Language-guided target segmentation method based on multi-granularity feature fusion[J]. Journal of Beijing University of Aeronautics and Astronautics，2024，50(2)：542-550 (in Chinese).

字塔来获取不同尺度的图像语义信息，进而通过循环神经网络逐步细化分割掩模，从多尺度进行分割优化。此后很多前沿方法都采用了与递归细化网络相似的多尺度特征结构。Ye 等[6]提出了跨模态自注意力分割网络，运用跨模态注意力机制融合语言视觉特征并利用门控机制进行多尺度特征的融合。Chen 等[7]通过视觉文本匹配先得到一个可视文本像素级热力图，卷积循环神经网络通过将热力图转换为细化的热力图来进行自顶向下的特征优化，该框架能够逐步产生理想的语言引导的分割结果与合理的注意力分布。Huang 等[8-9]将语言中单词的词性进行分类，并通过图网络分析不同实体之间的关系，加强对目标分类的针对性，得到了良好的分类效果。Bellver 等[10]提出的 Refvos 使用 Deeplabv3[11]提取视觉特征并利用 Bert[12]提取文本特征，通过简单的特征相乘得到融合特征并经过维度变换得到分割结果，虽然分割准确率相较 Huang 等[8-9]提出的方法有所下降，但更为简洁，参数规模有所降低，推理速度有较好的提升，可以胜任语言引导的视频目标分割任务。上述方法大多使用卷积神经网络及其变体提取视觉特征，循环神经网络提取文本特征，并在特征融合阶段进行了手段不一的优化，取得了良好的分割结果，但同时也欠缺对图像全局特征与空间位置特征的表达。

本文基于 Refvos[10]模型提出一种语言引导的多粒度特征融合目标分类方法，通过 Swin Transformer[13]提取多粒度视觉特征，并与 Bert 提取的文本特征进行跨模态特征融合，增强对整体和细节特征的表达；最后，通过卷积长短期记忆网络优化多粒度特征，对分割结果进行逐步的细化，得到语言引导的分割结果。

1　相关工作

1.1　Refvos 模型

图 1 为 Refvos[10]语言引导的目标分割模型，其采用了 Deeplabv3[11]和 Bert[12]进行视觉文本特征的提取，进行对应相乘得到融合特征，再经过 1×1 卷积得到分割结果。模型简洁，参数量较小，并取得了较高的分割精度，因此，其具有较大的改进空间。

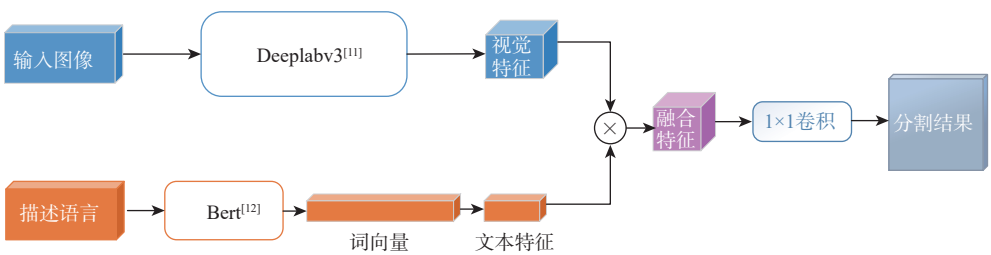


图 1　Refvos 模型

Fig. 1　Refvos model

1.2　Transformer

Transformer 是一种注意力编解码结构模型[14]。在自然语言处理领域取得显著性突破的 Bert[12]采用了 Transformer 的编码结构，将文本输入预训练好的 Bert[12]网络并进行输出优化即可得到适合相关任务的特征向量。

由此发展而来的视觉 Transformer 利用自注意力机制提取具有全局相关性的视觉特征[15]，相比于卷积神经网络擅长对局部特征进行提取，视觉 Transformer 可以有效反映图像的局部特征与全局特征。视觉 Transformer 模型如图 2 所示。

视觉 Transformer 首先对图像进行分块，并进行每个图像块的位置编码，随后，将线性处理后的图像块与位置编码输入 Transformer 编码器进行自注意力计算，最后，经过多层感知机得到分类结果。

视觉 Transformer 的局限性在于需要巨大的参数量支撑复杂的自注意力计算，对计算速度和存储量也有相对苛刻的要求，往往需要先将模型在大型数据集上进行预训练，再根据目标任务进行模型调整才能取得较好结果。本文方法采用 Swin Transformer 进行多粒度视觉特征提取，可以有效降低计算复杂度，充分提取图像整体与细节特征。

1.3　卷积长短期记忆网络

长短期记忆网络是一种特殊的循环神经网络[16]，其根据链式结构和门控机制可以有效地对长距离的信息进行分析，最初主要用于提取自然语言处理。由此发展出的卷积长短期记忆网络将原长短期记忆网络中的全连接层全部替换为卷积层[17]，这使得新的网络可以分析有序列关系的图像信息。卷积长短期记忆网络模型如图 3 所示，计算过程如下：

$$\boldsymbol{i}_t=\sigma(\mathrm{conv}(\boldsymbol{X}_t)+\mathrm{conv}(\boldsymbol{H}_{t-1})+\boldsymbol{W}_{\mathrm{ci}}\odot\boldsymbol{C}_{t-1}+\boldsymbol{b}_{\mathrm{i}}) \tag{1}$$

$$\boldsymbol{f}_t=\sigma(\mathrm{conv}(\boldsymbol{X}_t)+\mathrm{conv}(\boldsymbol{H}_{t-1})+\boldsymbol{W}_{\mathrm{cf}}\odot\boldsymbol{C}_{t-1}+\boldsymbol{b}_{\mathrm{f}}) \tag{2}$$

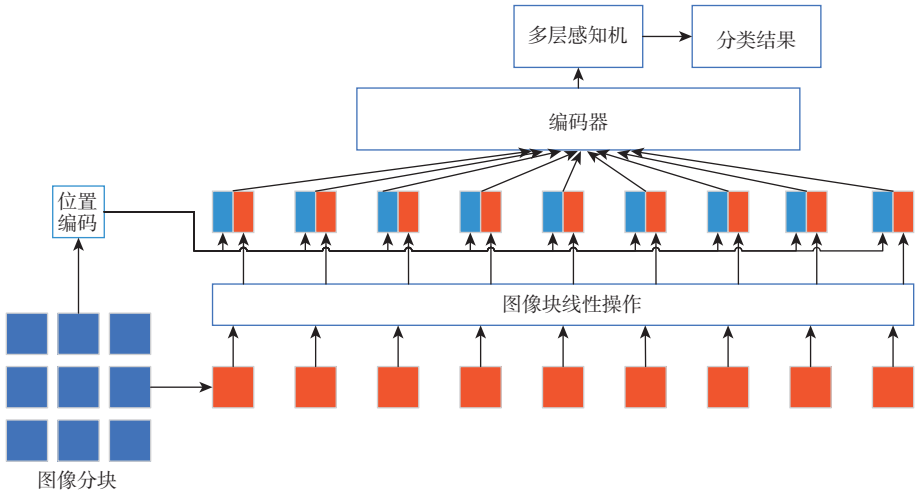


图2 视觉Transformer模型

Fig. 2 Visual Transformer model

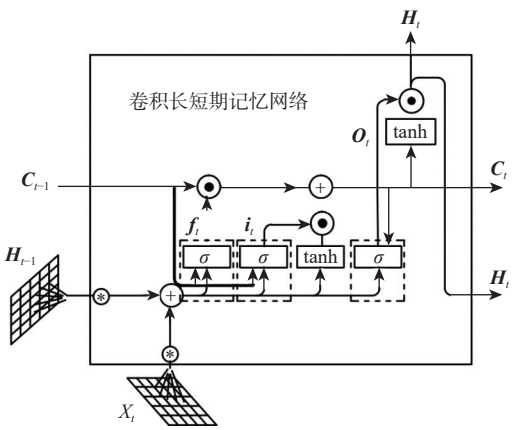


图3 卷积长短期记忆网络

Fig. 3 Convolutional long short-term memory network

$$\boldsymbol{C}_t=\boldsymbol{f}_t\odot\boldsymbol{C}_{t-1}+\boldsymbol{i}_t\odot\tanh(\mathrm{conv}(\boldsymbol{X}_t)+\mathrm{conv}(\boldsymbol{H}_{t-1})+\boldsymbol{b}_{\mathrm{c}}) \quad (3)$$

$$\boldsymbol{O}_t=\sigma(\mathrm{conv}(\boldsymbol{X}_t)+\mathrm{conv}(\boldsymbol{H}_{t-1})+\boldsymbol{W}_{\mathrm{co}}\odot\boldsymbol{C}_t+\boldsymbol{b}_{\mathrm{o}}) \quad (4)$$

$$\boldsymbol{H}_t=\boldsymbol{O}_t\odot\tanh(\boldsymbol{C}_t) \quad (5)$$

式中：σ为sigmoid激活函数；conv()为卷积计算；$\boldsymbol{X}_t$为时刻t的输入特征；$\boldsymbol{H}_{t-1}$为时刻$t-1$的输出特征；⊙为矩阵点乘计算；$\boldsymbol{W}_{\mathrm{ci}}$、$\boldsymbol{W}_{\mathrm{cf}}$、$\boldsymbol{W}_{\mathrm{co}}$为与记忆特征$\boldsymbol{C}_{t-1}$或$\boldsymbol{C}_t$点乘的权重；$\boldsymbol{b}_{\mathrm{i}}$、$\boldsymbol{b}_{\mathrm{f}}$、$\boldsymbol{b}_{\mathrm{c}}$、$\boldsymbol{b}_{\mathrm{o}}$为不同向量对应的偏置；tanh为激活函数；最后得到时刻t的输出$\boldsymbol{H}_t$。

2 本文方法

本文方法实现了在输入图像中精确分割输入语言所描述内容。本文方法的总体框架如图4所示，主要分为特征提取、特征融合、特征优化3个部分。

特征提取。为得到表达局部与整体的多粒度视觉特征，将输入图像顺序进行Swin Transformer的4阶段计算，每个阶段进行注意力计算提取特征以及窗口合并扩大感受野，逐渐得到粒度从小到大的视觉特征。输入描述语言经过Bert[12]网络提取文本特征。

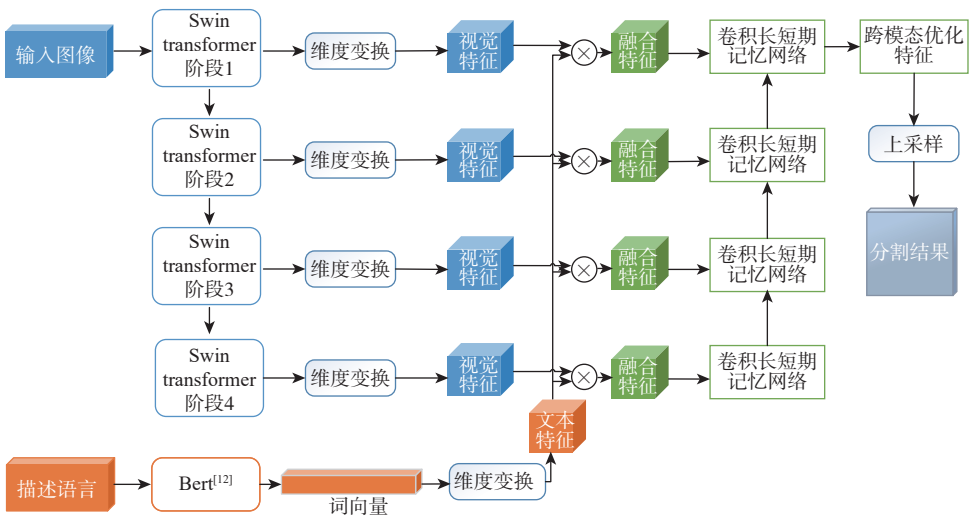


图4 本文方法总体框架

Fig. 4 General framework for proposed method

特征融合。进行多粒度的视觉文本特征融合，获取语言引导下的包含整体与细节信息的融合特征。

特征优化。利用经过卷积长短期记忆网络，对多粒度融合特征进行优化，逐步得到精细化的跨模态优化特征，最终上采样得出语言引导的图像分割结果。

2.1　视觉文本特征提取

2.1.1　视觉特征提取

语言引导的目标分割任务往往要从多个特征相似的实体中准确分割语言描述的对象，如分割出“左下角的水瓶”、“骑着自行车的人”等，需要从多个“水瓶”与“人”中准确定位语言描述的特殊个体，因此，更好地对图像中全局信息和位置信息进行提取可以有效提升语言引导目标分割的效果。为弥补卷积神经网络关注局部信息和视觉 Transformer 计算量大的缺陷，本文采用的视觉特征提取模型根据 Swin Transformer 进行改进。Swin Transformer 分为 4 个阶段提取不同粒度的视觉特征，每个阶段先对图像中局部窗口内部的像素进行自注意力计算，然后，通过窗口移动计算不同窗口之间的视觉特征的相关性，从而，在保证视觉特征全局相关的前提下，大大降低了计算的复杂度。Swin Transformer 结构如图 5 所示(以本文使用的 Swin-S 为例)。

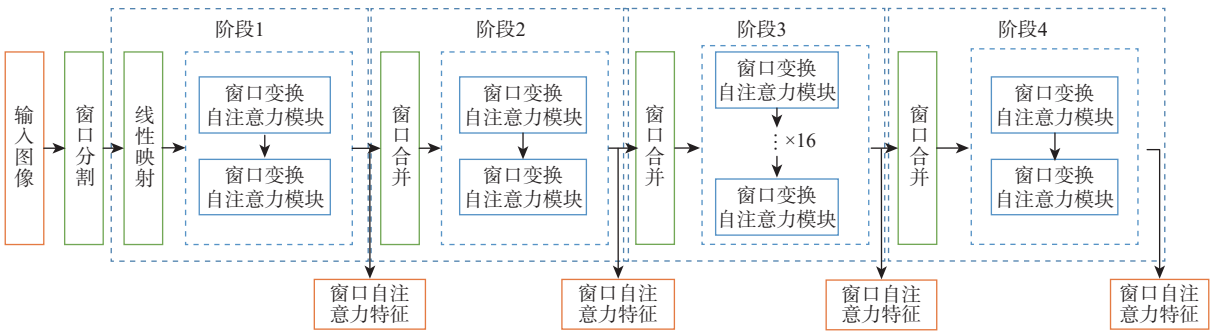


图 5　Swin Transformer 结构

Fig. 5　Swin Transformer structure

高H宽W的三维 RGB 输入图像经过窗口分割模块分为4×4像素一组的窗口区域，得到了尺寸为$\frac{H}{4}\times\frac{W}{4}$的 48 维特征图，随后线性映射模块将特征图维度转化为C，再经过 2 层窗口变换自注意力模块得到尺寸为$\frac{H}{4}\times\frac{W}{4}$的$C$维特征图。窗口变换自注意力在阶段 1 对图像中最小粒度的特征进行了提取。阶段 2、阶段 3 和阶段 4 中，首先，将特征图进行窗口合并，将2×2的 4 个窗口合并为 1 个窗口，从而减小窗口数量并扩大特征的粒度，再分别经过 2、18、2 个窗口变换自注意力模块提取粒度越来越大的视觉特征，分别得到尺寸维度为$\frac{H}{8}\times\frac{W}{8}\times 2C$、$\frac{H}{16}\times\frac{W}{16}\times 4C$和$\frac{H}{32}\times\frac{W}{32}\times 8C$的视觉特征。每个阶段的窗口变换自注意力模块数都为偶数，2 种不同结构的模块交替运算，如图 6 所示。

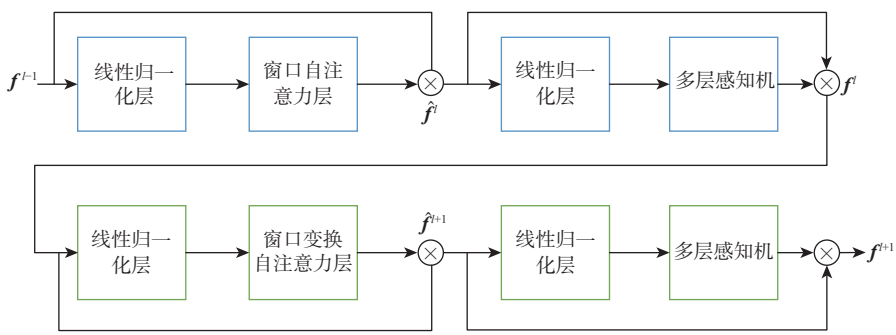


图 6　窗口变换自注意力模块

Fig. 6　Window transform self-attention module

上层计算得到的视觉特征图$\boldsymbol{f}^{l-1}$经过残差结构的线性归一化层和窗口注意力层得到$\hat{\boldsymbol{f}}^{l}$，再经过残差结构的线性归一化层与多层感知机得到视觉特征$\boldsymbol{f}^{l}$，此时，已经完成对窗口内的特征进行自注意力计算，类似卷积神经网络中利用卷积核的卷积运算，提取了局部的特征。接着，视觉特征图$\boldsymbol{f}^{l}$经过第 2 个窗口变换自注意力模块，该模块中使用了窗口变换自注意力层。窗口变换自注意力层先对窗口进行拆分与再结合，使得此前不同窗口的特征可以进行相互之间的自注意力计算，最后，经过残差

结构的线性归一化层和多层感知机得到输出视觉特征图$\boldsymbol{f}^{l+1}$，堆叠多层窗口变换自注意力模块，即可达到提取视觉特征全局相关性的目的。

图7为视觉特征提取过程，高和宽分别为H、W的三维输入图像$\boldsymbol{V}$通过Swin Transformer提取窗口注意力特征$\boldsymbol{V}_{an}(n=1,2,3,4)$，4个阶段计算得到的视觉特征$\boldsymbol{V}_{an}$的维度分别为$\frac{H}{4}\times\frac{W}{4}\times C$、$\frac{H}{8}\times\frac{W}{8}\times 2C$、$\frac{H}{16}\times\frac{W}{16}\times 4C$、$\frac{H}{32}\times\frac{W}{32}\times 8C$。随后利用$1\times1$卷积和双线性插值将不同粒度不同大小的视觉特征图进行特征归一化，得到尺寸$\frac{H}{4}\times\frac{W}{4}$的256维视觉特征$\boldsymbol{V}'_{an}$。

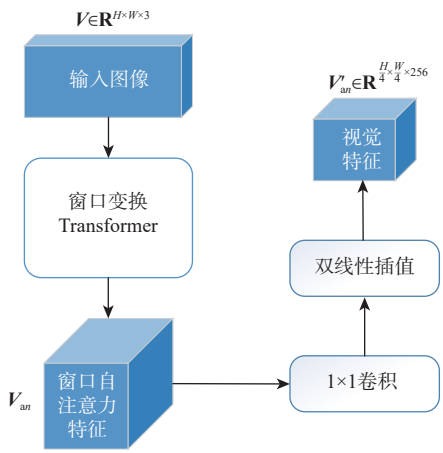


图7 视觉特征提取

Fig. 7 Visual feature extraction

2.1.2 文本特征提取

图8为文本特征提取过程，利用Bert[12]提取输入语言$\boldsymbol{L}$的词向量$\boldsymbol{L}_e\in\mathbf{R}^l$，进行特征归一化得到文本特征$\boldsymbol{L}''_e\in\mathbf{R}^{\frac{H}{4}\times\frac{W}{4}\times 256}$。归一化的视觉特征$\boldsymbol{V}'_{an}$和文本特征$\boldsymbol{L}''_e$进行矩阵相乘得到多粒度跨模态特征$\boldsymbol{F}\in\mathbf{R}^{\frac{H}{4}\times\frac{W}{4}\times 256}$。

2.2 特征融合

语言引导的目标分割方法需要将视觉特征和文本特征进行合理地对齐融合，进而实现对关注目标的准确分割。为使文本信息对图像整体和细节都具有引导作用，本文方法将归一化后的不同粒度视觉特征分别与文本特征进行矩阵相乘融合，得到

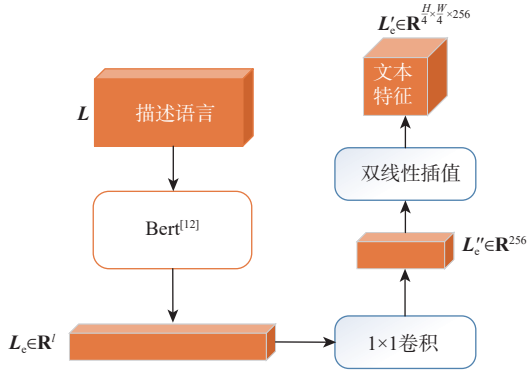


图8 文本特征提取

Fig. 8 Text feature extraction

多粒度的融合特征。多粒度融合特征计算过程为

$$\boldsymbol{V}'_{an}=\mathrm{Bi}(\mathrm{conv}(\mathrm{Tran}(\boldsymbol{V}))) \tag{6}$$

$$\boldsymbol{L}''_e=\mathrm{Bi}(\mathrm{conv}(\mathrm{Bert}(\boldsymbol{L}))) \tag{7}$$

$$\boldsymbol{F}_n=\boldsymbol{V}'_{an}\otimes\boldsymbol{L}''_e \tag{8}$$

式中：$n=1,2,3,4$；Tran为Swin Transformer第n阶段计算；conv表示1×1卷积计算；Bi表示双线性插值法；$\otimes$表示矩阵相乘计算。

2.3 多粒度特征优化

不同粒度的融合特征表达了语言与不同粒度视觉特征的对应关系，如果进行生硬的拼接，则会出现数据冗余，分割边界不清等情况。采用卷积长短期记忆网络可以有效对多粒度融合特征进行优化。如图1所示，以粒度最大到最小的融合特征顺序输入进行卷积长短期记忆网络计算，经过4个长短期记忆网络层，逐渐获取更加精细的特征图，最后经过上采样得到分割结果，多粒度特征优化具体计算流程如图9所示。

$\boldsymbol{F}_n(n=1,2,3,4)$表示输入的融合特征，$\boldsymbol{H}_n$表示经过卷积长短期记忆网络计算得到的各层输出，$\boldsymbol{C}_n$表示卷积长短期记忆网络的记忆信息。从粒度最大的融合特征$\boldsymbol{F}_4$作为输入开始优化计算，经过首个卷积长短期网络得到输出$\boldsymbol{H}_1$和记忆信息$\boldsymbol{C}_1$，第2层卷积长短期记忆网络由融合特征$\boldsymbol{F}_3$、$\boldsymbol{H}_1$、$\boldsymbol{C}_1$作为输入，以此类推，最后得到第4层网络的输出$\boldsymbol{H}_4$作为跨模态优化特征。

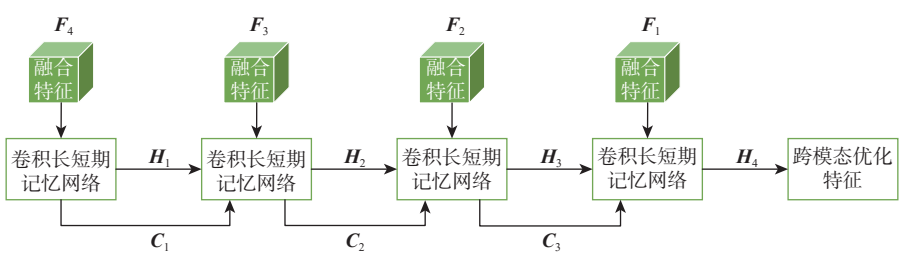


图9 多粒度特征优化计算流程

Fig. 9 Multi-granularity feature optimization calculation process

3　实验结果与分析

3.1　数据集

本实验利用UNC[18]、UNC+[18]、G-Ref[19]、ReferIt[20]数据集进行语言引导的目标分割实验。UNC[18]、UNC+[18]、G-Ref[19]都取自数据集MSCOCO[21], 3个数据集分别包含了19 994、19 992和26 711张图片, 并分别对超过50 000个目标进行了有142 209、141 564、104 560句的语言描述。UNC+[18]与UNC[18]数据集的不同之处在于UNC+[18]数据集中的语言描述不包含目标的位置信息, 而G-Ref[19]数据集的特点在于其句子较长, 对分割任务提出了更高的挑战。ReferIt[20]数据集取自数据集IAPR TC12[22], 其包含19 894张图片并对96 654个目标进行了13 525句语言描述。

3.2　实验设置

本实验在运用Pytorch深度学习框架NVIDIA-3090GPU, 并基于Refvos[10]语言引导的目标分割方法框架进行修改。模型对输入图像进行随机裁剪为448×448像素大小的图像, 采用Bert[12]预训练模型与Swin-transformer-S的在1K ImageNet[23]上的预训练模型进行训练, 采用计算预测分割结果与正确分割结果之间的交叉熵作为模型的损失函数, 采用随机梯度下降(stochastic gradient descent, SGD)算法进行优化。在UNC[18]、UNC+[18]、G-Ref[19]、ReferIt[20]数据集上进行训练, 初始学习率为0.06, 权重衰减为10^{-6}, 训练40轮, 并将初始学习率设为0.01, 权重衰减为10^{-6}, 继续训练10轮, 进行训练。

3.3　实验结果分析

本文方法在UNC[18]、UNC+[18]数据集上进行测试, 2个数据集都分别有val、testA、testB这3个测试子集。本文方法在2个数据集上分割总体交并比(intersection over union, IoU)与前沿方法的对比如表1所示。

表1　总体IoU实验结果对比

Table 1　Overall IoU comparison of experimental results　%

语言引导分割方法	UNC[18]			UNC+[18]			G-Ref[19]	ReferIt[20]
	val	testA	testB	val	testA	testB	val	test
RMI[2]	45.18	45.69	45.57	29.86	30.48	29.50	34.52	58.73
DMN[3]	49.78	54.83	45.13	38.88	44.22	32.29	36.76	52.81
RRN[5]	55.33	57.26	53.95	39.75	42.15	36.11	36.45	63.63
MattNet[24]	56.51	62.37	51.70	46.67	52.39	40.08		
CMSA[6]	58.32	60.61	55.09	43.76	47.60	37.89	39.98	63.80
Refvos[10]	59.45	63.19	54.17	44.71	49.73	36.17		
STEP[7]	60.04	63.46	58.97	48.18	52.33	40.41	46.40	64.13
BRINET[25]	61.35	63.37	59.57	48.57	52.87	42.13	47.57	
CMPC[8]	61.36	64.53	59.64	49.56	53.44	43.23	49.05	65.53
本文方法	60.37	62.31	58.27	46.54	50.36	37.92	40.16	64.37

实验结果表明本文方法在UNC[18]数据库的val、testB测试集中相较Refvos[10]分别提升了0.92%、4.1%, testA测试集下降了0.88%, 在UNC+[18]数据库val、testA、testB测试集中分别提升了1.83%、0.63%、1.75%。在UNC[18]中的分割精度全部大于UNC+, [18]表明描述空间位置的文本信息能对精准分割起到很大帮助。本文方法在G-Ref[19]和ReferIt[20]数据集上的分割精度达到了40.16%和64.37%, 达到了前沿水平。本文方法的精确度相较透视文本分组图像指代分割(see-through-text grouping for referring image segmentation, STEP)方法[7], 双向关系推断指代图像分割(bi-directional relationship inferring network for referring image segmentation, BRINET)方法[25]、跨模态渐进理解图像指代分割(referring image segmentation via cross-model progressive comprehension, CMPC)方法[8]等最前沿方法较低, 但上述方法相较本文方法模型更复杂, CMPC[18]除了视觉、文本特征提取模块, 还包括了图推理模块和语言引导的多特征融合模块。BRINET[22]包含一个交叉注意模块和一个双向模块。本文方法在结构相对简单的情况下仍然保持了较高的分割精度。

3.4　消融实验

本文进行了2个消融实验来证明本文方法的有效性。

为证明本文多粒度特征提取的有效性, 进行单一粒度的特征提取, 直接将Swin Transformer最后一阶段得到的特征与文本特征进行融合, 将该特征扩展为卷积长短期记忆网络的输入维度进行卷积长短期计算, 最后上采样得到分割结果, 并在UNC[18]数据集上进行训练并测试。IoU实验结果如表2所示。

实验结果表明, 在3种不同的测试集上本文方

表 2 消融实验结果总体 IoU 对比

Table 2 Overall IoU comparison of ablation experiment results %

方法	val	testA	testB
单一粒度	55.31	55.70	53.85
无特征优化	57.99	59.62	56.62
本文方法	60.37	62.31	58.27

法都比单一粒度的特征提取方法分割效果更好。多粒度特征提取模块有效地对图像不同粒度的特征进行了分析并得到了更为精确的分割结果。

为证明多粒度特征优化的有效性，进行无特征优化模块的实验，将卷积长短期记忆网络去掉，将不同阶段的 Swin Transformer 计算得到的特征分别与文本特征融合后直接将多粒度跨模态特征进行拼接与维度压缩，最后上采样得到分割结果。在 UNC[18] 数据集上进行训练并测试，实验 IoU 结果如表 2 所示。

实验结果表明，本文方法在 3 个不同测试集上的分割效果最优。不加入卷积长短期记忆网络，不同粒度的融合特征无法进行有效地整合，导致分割效果的下降，证明了多粒度特征融合模块的有效性。

3.5 定性结果

语言引导的图像分割效果如图 10～图 12 所示。图 10 和图 11 表示模型对语言所描述的对象进



图 10 “站在中间戴着围巾的女人”分割结果

Fig. 10 "The woman in the center with the scarf" split results



图 11 “最左边的男人”分割结果

Fig. 11 Segmentation results for "leftmost man" split results



图 12 “蛋糕第 2 层”分割结果

Fig. 12 "Second layer of the cake" split results

行了精准的分割。图 12 并没有对第 2 层蛋糕进行正确分割，而是几乎分割了整个蛋糕。表明本文方法在复杂目标与特有的语言描述对象的分割上还有待提高。

4 结 论

1）本文方法对 Refvos 语言引导的目标分割方法进行改进。在 UNC 与 UNC+数据集的实验结果表明，相比 Refvos，在 UNC 数据集的 val、testB 子集中 IoU 结果分别提升了 0.92%、4.1%，在 UNC+数据集的 val、testA、testB 中 IoU 测试结果分别提升了 1.83%、0.63%、1.75%。在 G-Ref 数据集 IoU 测试结果达到了 40.16%，在 ReferIt 数据集 IoU 测试结果达到了 64.37%。

2）本文方法使用 Swin Transformer 网络作为视觉特征提取器，并取得了较好的结果。

3）本文方法提出了利用多粒度的 Swin Transformer 视觉特征，并证明了其相较单粒度特征的优越性。

4）本文方法利用卷积长短期记忆网络对多粒度的融合特征进行优化，证明了其有效性。

参考文献（References）

[1] HU R H, ROHRBACH M, DARRELL T. Segmentation from natural language expressions[C]//Proceeding of the European Conference on Computer Vision. Berlin: Springer, 2016: 108-124.

[2] LIU C X, LIN Z, SHEN X H, et al. Recurrent multimodal interaction for referring image segmentation[C]//Proceedings of the 2017 IEEE International Conference on Computer Vision. Piscataway: IEEE Press, 2017: 1280-1289.

[3] MARGFFOY-TUAY E, PÉREZ J C, BOTERO E, et al. Dynamic multimodal instance segmentation guided by natural language queries[C]//Proceedings of the European Conference on Computer Vision. New York: ACM, 2018: 656–672.

[4] LEI T, ZHANG Y. Training RNNs as fast as CNNs[EB/OL]. (2017-09-12)[2022-01-05]. https://arxiv.org/abs/1709.02755v2.

[5] LI R Y, LI K C, KUO Y C, et al. Referring image segmentation via

recurrent refinement networks[C]//Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2018: 5745-5753.

[6] YE L W, ROCHAN M, LIU Z, et al. Cross-modal self-attention network for referring image segmentation[C]//Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2019: 10494-10503.

[7] CHEN D J, JIA S H, LO Y C, et al. See-through-text grouping for referring image segmentation[C]//Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE Press, 2019: 7453-7462.

[8] HUANG S F, HUI T R, LIU S, et al. Referring image segmentation via cross-modal progressive comprehension[C]//Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2020: 10485-10494.

[9] HUI T R, LIU S, HUANG S F, et al. Linguistic structure guided context modeling for referring image segmentation[C]//Proceeding of the European Conference on Computer Vision. Berlin: Springer, 2020: 59-75.

[10] BELLVER M, VENTURA C, SILBERER C, et al. A closer look at referring expressions for video object segmentation[J]. Multimedia Tools and Applications, 2023, 82(3): 4419-4438.

[11] CHEN L C, PAPANDREOU G, SCHROFF F, et al. Rethinking atrous convolution for semantic image segmentation[EB/OL]. (2017-12-05)[2022-01-12]. https://doi.org/10.48550/arXiv.1706.05587.

[12] DEVLIN J, CHANG M W, LEE K, et al. BERT: Pre-training of deep bidirectional transformers for language understanding[EB/OL] . (2018-11-11)[2022-01-12].https://arxiv.org/abs/1810.04805.

[13] 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.

[14] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]//Proceedings of the 31st International Conference on Neural Information Processing Systems. New York: ACM, 2017: 6000-6010.

[15] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, et al. An image is worth 16×16 words: Transformers for image recognition at scale[EB/OL]. (2021-06-03)[2021-12-11]. https://doi.org/10.48550/arXiv.2010.11929.

[16] HOCHREITER S, SCHMIDHUBER J. Long short-term memory [J]. Neural Computation, 1997, 9(8): 1735-1780.

[17] SHI X J, CHEN Z R, WANG H, et al. Convolutional LSTM Network: A machine learning approach for precipitation nowcasting [C]//Proceedings of the 28th International Conference on Neural Information Processing Systems. New York: ACM, 2015: 802–810.

[18] YU L C, POIRSON P, YANG S, et al. Modeling context in referring expressions[C]//Proceeding of the European Conference on Computer Vision. Berlin: Springer, 2016: 69-85.

[19] MAO J H, HUANG J, TOSHEV A, et al. Generation and comprehension of unambiguous object descriptions[C]//Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2016: 11-20.

[20] KAZEMZADEH S, ORDONEZ V, MATTEN M, et al. ReferItGame: Referring to objects in photographs of natural scenes[C]// Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing. Stroudsburg: Association for Computational Linguistics, 2014: 787-798.

[21] LIN T Y, MAIRE M, BELONGIE S, et al. Microsoft COCO: Common objects in context[C]//Proceeding of the European Conference on Computer Vision. Berlin: Springer, 2014: 740-755.

[22] ESCALANTE H J, HERNÁNDEZ C A, GONZALEZ J A, et al. The segmented and annotated IAPR TC-12 benchmark[J]. Computer Vision and Image Understanding, 2010, 114(4): 419-428.

[23] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. ImageNet classification with deep convolutional neural networks[J]. Communications of the ACM, 2017, 60(6): 84-90.

[24] YU L C, LIN Z, SHEN X H, et al. MAttNet: Modular attention network for referring expression comprehension[C]//Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2018: 1307-1315.

[25] HU Z W, FENG G, SUN J Y, et al. Bi-directional relationship inferring network for referring image segmentation[C]//Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2020: 4423-4432.

Language-guided target segmentation method based on multi-granularity feature fusion

TAN Quange，WANG Rong*，WU Ao

(School of Information and Cyber Security，People's Public Security University of China，Beijing 100091，China)

Abstract：The objective of language-guided target segmentation is to match the targets described in the text with the entities they refer to, thereby achieving an understanding of the relationships between text and entities, as well as the localization of the referred targets. This task has significant application value in scenarios such as information extraction, text classification, and machine translation. The paper proposes a language-guided multi-granularity feature fusion target segmentation method based on the Refvos model, which can accurately locate segment-specific targets. Using the Swin Transformer and Bert network to extract multi-granularity visual features and text features respectively, so as to obtain features that have strong expression ability to the whole and part. Through language direction, text features are combined with visual features of varying granularities to improve targeted expression. Ultimately, in order to achieve more precise segmentation results, we enhance multi-granularity fusion features using convolutional long and short-term memory networks to facilitate information flow across features of different granularities. The model was trained and tested on UNC and UNC+ datasets. Experimental results show that the proposed method compared with Refvos, IoU results in UNC dataset Val and testB are improved by 0.92% and 4.1% respectively, and IoU results in UNC+ dataset Val, testA and testB are improved by 1.83%, 0.63%, and 1.75% respectively. The proposed method IoU results of G-Ref and ReferIt data sets are 40.16% and 64.37%, reaching the frontier level. It is proved that the proposed method is effective and advanced.

Keywords：target segmentation; referring segmentation; feature fusion; cross-model; text understanding

Received：2022-05-18；**Accepted**：2022-06-23；**Published Online**：2022-10-25 10：03

URL：link.cnki.net/urlid/11.2625.V.20221024.1456.001

Foundation item：National Natural Science Foundation of China (62076246)

* **Corresponding author**. E-mail：dbdxwangrong@163.com

http://bhxb.buaa.edu.cn jbuaa@buaa.edu.cn
DOI: 10.13700/j.bh.1001-5965.2022.0385

基于多级语义对齐的图像-文本匹配算法

李艺茹[1]，姚涛[1,2,*]，张林梁[2]，孙玉娟[1]，付海燕[3]

(1. 鲁东大学 信息与电气工程学院，烟台 264000；2. 西南交通大学 烟台新一代信息技术研究院，烟台 264000；
3. 大连理工大学 信息与通信工程学院，大连 116024)

摘　　要：图像中的区域特征更关注于图像中的前景信息，背景信息往往被忽略，如何有效的联合局部特征和全局特征还没有得到充分地研究。为解决上述问题，加强全局概念和局部概念之间的关联得到更准确的视觉特征，提出一种基于多级语义对齐的图像-文本匹配算法。提取局部图像特征，得到图像中的细粒度信息；提取全局图像特征，将环境信息引入到网络的学习中，从而得到不同的视觉关系层次，为联合的视觉特征提供更多的信息；将全局-局部图像特征进行联合，将联合后的视觉特征和文本特征进行全局-局部对齐得到更加精准的相似度表示。通过大量的实验和分析表明：所提算法在2个公共数据集上具有有效性。

关 键 词：图像-文本匹配；跨模态信息处理；特征提取；神经网络；特征融合
中图分类号：TP311
文献标志码：A　　**文章编号**：1001-5965(2024)02-0551-08

随着信息技术的不断发展，社交网络上产生了大量多模态数据，因此，对多模态数据的处理越来越重要[1]。对于这些多模态数据而言，视觉和文本数据占据了主导地位，从而吸引了更多的研究学者专注于视觉和文本的跨模态任务[2-3]，例如，图像-文本匹配任务。虽然近年来图像-文本匹配任务取得了重大进展，但其仍是一个具有挑战性的问题[4]。因为图像和文本数据之间的异构特性导致其之间存在着巨大的语义鸿沟，使得图像和文本的匹配还远未解决[5-6]。

早期的图像-文本匹配的核心任务是学习一个共同的潜在空间，然后将图像特征或文本特征的实值量化为离散的哈希码[7]。随着深度学习的不断发展，图像-文本匹配的主流则变为特定模态的深度特征学习。传统的方法先利用基于卷积的神经网络来学习图像特征，利用基于循环神经网络(recurrent neural networks, RNN) (如长短时记忆网络(long short-term memory, LSTM) 或门控循环单元(gated recurrent unit, GRU)) 学习文本特征。然后分别将图像特征和文本特征量化为图像哈希码和文本哈希码，最后利用相似性函数来测量2个哈希码之间的距离[8]。然而，量化过程往往造成了较大的量化损失，导致了较低的匹配性能。同时，只涉及整张图像和整个句子的对齐往往是单一且不精准的，其缺乏更详细的匹配，即图像区域和文本中的单词之间的对齐[9-10]。

随着Faster R-CNN在目标检测任务中提取局部区域特征的优秀表现，R-CNN为区域卷积神经网络 (region convolution nettral networks)，堆叠交叉注意力网络 (stacked cross attention networks, SCAN) 算法[11]提出将Faster R-CNN用于图像-文本匹配任务中，首先，使用预先训练好的Faster R-CNN检测原始图像中的对象，然后，利用ResNet101从这些对象中提取图像的区域特征，使图像中的区域和文本中的单词进行对齐[12]。相似性推理和注意力过滤 (similaritve graph and attention filtration, SGRAF) 算法[13]

收稿日期：2022-05-19；**录用日期**：2022-06-23；**网络出版时间**：2022-10-09 15:47
网络出版地址：link.cnki.net/urlid/11.2625.V.20221009.1013.008
基金项目：国家自然科学基金 (61872170,62076052,61873117)；国家重点研发项目 (2021ZD0113002)
*通信作者. E-mail：yaotao@ldu.edu.cn

引用格式：李艺茹，姚涛，张林梁，等. 基于多级语义对齐的图像-文本匹配算法 [J]. 北京航空航天大学学报，2024，50（2）：551-558.
LI Y R，YAO T，ZHANG L L，et al. Image-text matching algorithm based on multi-level semantic alignment[J]. Journal of Beijing University of Aeronautics and Astronautics，2024，50（2）：551-558（in Chinese）.

将区域特征和文本特征进行全局和局部对齐。虽然这些算法相当成功，但还是存在以下缺点：区域和单词的对齐对于完成图像-文本匹配任务来说过于单一，应当考虑全局的匹配过程。因为 Faster R-CNN 在做目标检测任务时只是对图像的前景信息进行检测，然而对于图像-文本匹配任务来说，背景（环境）信息也固然重要[14-17]。当试图用区域特征去匹配相应的文本时，图像中的非对象元素是不能进行匹配的，由于丢失重要的图像信息，导致匹配的精度下降。

大多数算法通过将区域特征和文本特征进行对齐，从而学习其之间的相似性信息[18]。然而仅仅考虑图像区域和文本单词的对齐过于单一，同时可能丢失重要的环境（背景）等信息，需要引入图像的全局信息再进行匹配。

为了克服上述问题，本文提出了一种基于多级语义对齐的图像-文本匹配算法。本文算法将图像的区域特征和全局特征融合到一个统一的框架中。因此，不仅可以保持图像中的细粒度语义信息，又可以保持图像中的背景信息，以便于图像和文本的对齐过程。具体来说，首先，提取完整图像的区域特征，从而将图像中细粒度信息融入到网络的学习中；然后，提取图像的全局特征，在保证融入了图像的前景信息之后，又将重要的环境（背景）信息融入到网络的学习中；将全局-区域特征进行联合，得到联合后的视觉特征，从而找到全局-局部对象像素级概念之间的语义关系；对于文本来说，得到每个单词的特征表示；最后，将联合后的视觉特征和单词特征进行对齐得到相似性得分。本文贡献如下：

1）本文算法将图像的全局特征和区域特征进行融合构成融合后的图像特征，不仅考虑区域（对象）本身的关系，同时引入了图像的全局特征，从而加强了视觉表征中区域对象和全局概念之间的关系。

2）本文算法在将联合后的视觉特征与单词特征进行局部和全局对齐的过程中，不仅将区域对象和文本中描述前景物体的单词进行匹配，同时将全局概念和文本中的背景单词进行匹配，从而得到精准的匹配结果。

3）实验结果表明，与现有的几种算法相比，本文算法在各个方面都具有优越的性能。

1 相关工作

图像-文本匹配任务是近年来最流行的研究课题之一，本节将介绍近几年用于图像-文本匹配任务的优秀工作。图像-文本匹配大致可分为 2 种：全局匹配和局部匹配。

全局匹配算法。具体来说，全局匹配算法是以端到端的形式将视觉和文本嵌入到一个共享空间中，然后计算在共享空间中的视觉和文本特征的相似性。其中，全局匹配算法的难点在于如何学习一个映射函数将视觉和文本特征嵌入到共享空间中[19]。生成式跨模态特征学习框架（简称 GXN）算法[20]将生成目标和跨视图特征嵌入学习相结合，来学习对视觉和文本数据更有区别的表示。共识感知的视觉语义嵌入（consensus-aware visual-semantic embedding, CVSE）算法[21]尝试利用共识信息进行图像文本匹配，在共识水平上统一了 2 种模态的表示。通过一个图卷积神经网络学习共识感知概念表示。这些算法都可以被概括为全局图像-文本匹配算法[22-23]。因为其都直接将整个图像或文本编码为向量。但是，对于这些算法来说，只考虑了图像或文本的全局上下文的对齐，而忽略了图像区域和单词的对齐，因此，很难捕捉到图像中的细粒度信息，这可以通过区域图像-文本匹配算法来缓解。

局部匹配算法。局部匹配算法是图像-文本匹配算法中一种更精细的算法。其通过提取图像的区域特征和文本中单词的特征进行区域和单词的局部匹配。SCAN 算法[11]提出了一个堆叠交叉注意力模型，该模型将每个片段与来自其他模态的所有其他片段对齐。SGRAF 算法[13]以一种更全面的方式来描述图像和文本特征之间的局部对齐和全局对齐。虽然这些算法成功的对图像的区域特征和文本的单词特征进行一致性匹配，但是由于缺乏全局图像特征和文本特征之间的对齐，所以这些匹配算法的结果往往是不准确的。

注意力机制。由于注意力机制的有效性，该机制已被广泛应用，如图像字幕、知识问答等任务中。在计算机视觉中，视觉注意的目标是关注特定的图像或子区域。类似地，自然语言处理的注意方法是自适应地选择和聚合信息片段来推断结果。注意力机制的主要目的是突出重要的局部信息，或解决全局特征带来的无关信息或噪声信息的问题。最近，基于注意力的模型被提出用于图像-文本匹配问题。基于循环注意力记忆的迭代匹配（iterative matching with recurrent attention memory, IMRAR）算法[24]提出了一种与重复注意记忆的迭代匹配算法，来逐步探索区域-单词之间的内在关联。

尽管这些算法成功地学习了图像和文本的一致性表示，但其还是有缺陷[25-26]。仅仅考虑图像区域和文本单词的对齐过于单一，同时可能丢失重要的环境（背景）等信息。本文算法与前面提到的

2 种匹配算法不同。本文算法采用了具有区域-单词的视觉-语言注意力机制算法, 然后, 将全局图像特征和区域图像特征进行联合, 既引入了图像的细粒度信息, 又包含图像的背景(环境)信息, 得到更准确的视觉特征, 最后, 将联合后的视觉特征和文本特征进行全局和局部对齐。

2　本文算法

2.1　图像区域特征提取

本文算法网络框架如图 1 所示, 其中, Bi-GRU 为双向门控循环单元 (bidirectional gated recurrent unit)。

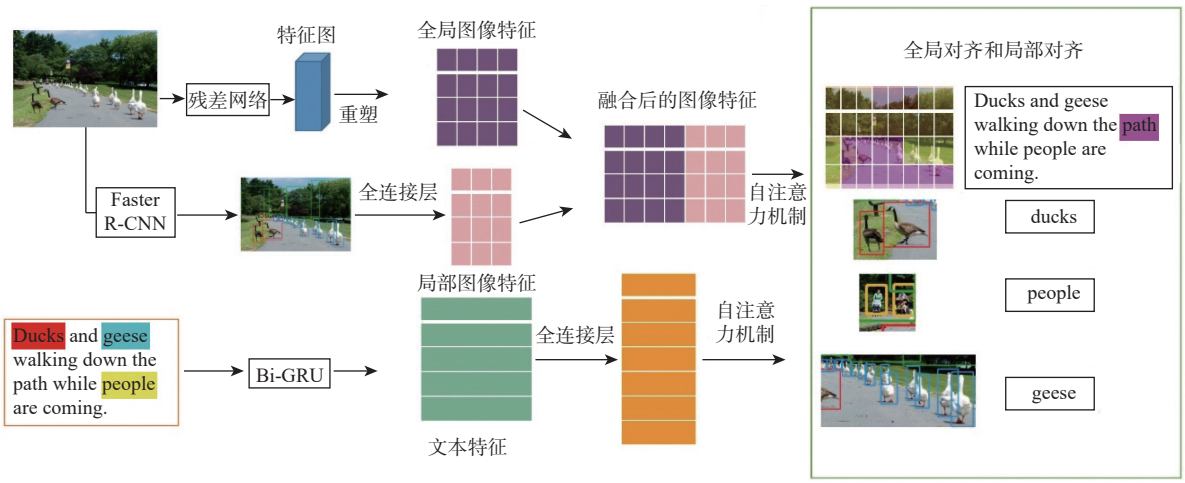


图 1　本文算法的网络框架

Fig. 1　Network framework of proposed algorithm

区域特征表示。对于每张输入图像$\boldsymbol{i}_m$来说, 其由多个区域组成, 检测对全局语义贡献最大的显著区域, 并将每个区域编码为特征向量。在这项工作中, 先利用预训练好的 Faster R-CNN 模型提取图像区域的特征, 可以预测对象的边界框, 并对其进行评分; 根据其分数选择前 K 个显著对象, 然后使用预训练好的 ResNet101 提取这些边界框的平均池化卷积特征; 最后, 利用一个全连接层将这些区域特征转换为 d 维特征向量, 作为图像的区域特征表示 $\boldsymbol{I}^1=[\boldsymbol{i}_1^1,\boldsymbol{i}_2^1,\cdots,\boldsymbol{i}_K^1]$。

全局特征表示。对于每张输入图像$\boldsymbol{i}_m$来说, 都调整为 256×256 像素的大小。通常, 使用预先训练好的 ResNet152, 去掉最后一个全连接层来提全局的图像特征。使用最后一层的特征向量进行线性映射, 并将其重塑为一组特征 $\boldsymbol{I}^2=[\boldsymbol{i}_1^2,\boldsymbol{i}_2^2,\cdots,\boldsymbol{i}_n^2]$, 其中, $\boldsymbol{i}_1^2\in\mathbf{R}^d$, n 为重塑后的特征图大小, d 为特征向量的维度。

图像特征融合。将图像的区域特征$\boldsymbol{I}^1$和图像的全局特征$\boldsymbol{I}^2$按照式 (1) 进行特征的连接:

$$\boldsymbol{I}=\mathrm{cat}(\boldsymbol{I}^1,\boldsymbol{I}^2) \tag{1}$$

将融合后的图像特征取平均得到平均图像融合特征$\overline{\boldsymbol{A}_{\boldsymbol{I}}}=\dfrac{1}{K+n}\sum_{i=1}^{K+n}\boldsymbol{i}_i$, 然后通过自注意力机制聚合所有融合图像特征得到全局融合图像特征$\bar{\boldsymbol{i}}$。

2.2　文本特征提取

每个文本都包含一组单词, 将每个单词编码为 d 维特征向量, 并将其合并为文本的全局特征。为此, 采用 Bi-GRU 将前馈和反馈上下文信息整合到单词表示中。具体来说, 利用标记化技术将每个句子划分为 L 个单词, 用 Bi-GRU 网络学习单词的特征中, 得到单词的表示 $\boldsymbol{T}=[\boldsymbol{t}_1,\boldsymbol{t}_2,\cdots,\boldsymbol{t}_L]$。同样地, 平均文本特征$\overline{\boldsymbol{A}_{\boldsymbol{T}}}=\dfrac{1}{L}\sum_{i=1}^{L}\boldsymbol{t}_i$, 通过自注意力机制聚合所有融合文本特征得到全局融合文本特征$\bar{\boldsymbol{t}}$。自注意力模块如图 2 所示。

2.3　全局相似性表示

全局融合图像特征$\bar{\boldsymbol{i}}$和全局融合文本特征$\bar{\boldsymbol{t}}$之间的相似性表示的计算式为

$$\boldsymbol{S}_{\mathrm{glo}}=S(\bar{\boldsymbol{t}},\bar{\boldsymbol{i}})=\frac{\left|\bar{\boldsymbol{t}}-\bar{\boldsymbol{i}}\right|^2}{\left\|\left|\bar{\boldsymbol{t}}-\bar{\boldsymbol{i}}\right|^2\right\|_2} \tag{2}$$

式中: $|\cdot|^2$和$\|\cdot\|_2$分别为元素的平方和l_2范数。以往的大多数方法利用余弦相似度来计算 2 个向量之间的距离, 但缺乏了元素和元素的对应。在本文中, 计算得到的相似度信息将不再是一个标量, 而是一个向量来表示图像特征和文本特征之间的对应关系。

2.4　局部相似性表示

为了计算图像特征和文本特征之间的局部相似度表示, 本文对联合后的视觉特征和文本特征进

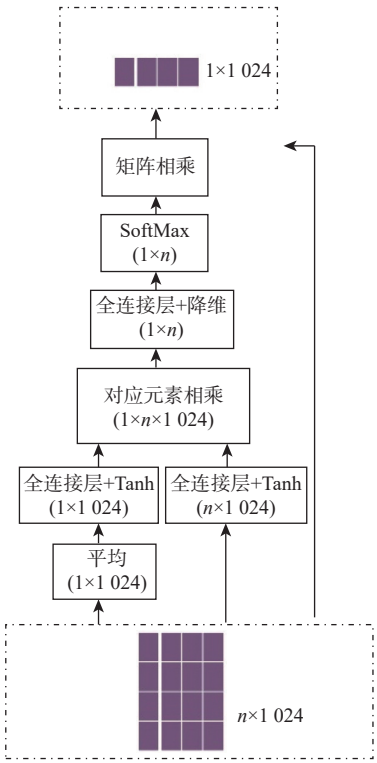


图2 自注意力模块

Fig. 2 Self-attention module

行重构，从而重点关注每个单词和每个区域。

首先，计算联合后的视觉特征 $\boldsymbol{I}$ 和文本特征 $\boldsymbol{T}$ 之间的余弦距离 $\boldsymbol{d}_{ij}$，同时对 $\boldsymbol{d}_{ij}$ 进行归一化得到 $\widetilde{\boldsymbol{d}_{ij}}$；然后，对联合后的图像特征和文本特征采用注意力机制，其中，权重计算式为

$$\boldsymbol{w}_{ij}=\frac{\exp(\alpha\widetilde{\boldsymbol{d}_{ij}})}{\sum_{i=1}^{K+n}\widetilde{\boldsymbol{d}_{ij}}\exp(\alpha\widetilde{\boldsymbol{d}_{ij}})} \tag{3}$$

式中：$\boldsymbol{w}_{ij}$ 为由具有尺度参数 α 的 softmax 函数计算出来的权重矩阵。

本文利用权重矩阵对融合后的图像特征进行重构得到重构图像特征，如下：

$$\boldsymbol{f}_j^i=\sum_{i=1}^{K+n}\boldsymbol{w}_{ij}\boldsymbol{i}_i \tag{4}$$

重构图像特征 $\boldsymbol{f}_j^i$ 和文本特征 $\bar{\boldsymbol{t}}$ 之间的相似性表示计算式为

$$\boldsymbol{S}_{\text{avg}}=S(\boldsymbol{f}_j^i,\bar{\boldsymbol{t}})=\frac{\left|\bar{\boldsymbol{t}}-\boldsymbol{f}_j^i\right|^2}{\left\|\left|\bar{\boldsymbol{t}}-\boldsymbol{f}_j^i\right|^2\right\|_2} \tag{5}$$

计算全局-局部相似度，既保留了图像中重要的前景信息，又对图像中的背景（环境）信息进行了保留。

2.5 损失函数

利用双向排名损失来训练本文算法，给定一个匹配的图像-文本对 $(\boldsymbol{i},\boldsymbol{t})$，以及在一个小批次中对应的最难匹配的图像 $\bar{\boldsymbol{i}}$ 和最难匹配的文本 $\bar{\boldsymbol{t}}$。利用式（6）来计算双向排名损失：

$$L(\boldsymbol{i},\boldsymbol{t})=\left[\beta-S(\boldsymbol{i},\boldsymbol{t})+S(\boldsymbol{i},\bar{\boldsymbol{t}})\right]_+ +\left[\beta-S(\boldsymbol{i},\boldsymbol{t})+S(\bar{\boldsymbol{i}},\boldsymbol{t})\right]_+ \tag{6}$$

式中：β 为范围参数；$S(\cdot)$ 为相似性预测函数；$[]_+$ 表示取0和括号内变量较大的一个。

3 实 验

为了验证本文算法的有效性，本文进行了大量的实验，其性能普遍高于现存算法。

3.1 数据集

本文在 MSCOCO 和 Flickr30K 数据集上评估本文算法的表现。MSCOCO 数据集包含 123 287 张图像，每张图像都有 5 个注释标题。该数据被分成 113 287 张图像用于训练，5 000 张图像用于验证，5 000 张图像用于测试。Flickr30K 数据集包含 31 783 张图像，每个图像有 5 个相应的标题。本文使用 1 000 张图像进行验证，1 000 张图像用于测试，其余的用于训练。

实验使用了 2 个性能评价指标：R_K 与 R_{sum}。R_K 表示检索结果取前 K 个实例时所获得的召回率，其值越高模型性能越好。R_{sum} 用来评估算法的整体性能，计算式为

$$R_{\text{sum}}=\{R_1+R_5+R_{10}\}_{T-I}+\{R_1+R_5+R_{10}\}_{I-T} \tag{7}$$

3.2 实现细节

本节为本文算法提供了更详细的训练设置和模型设置。对于全局图像特征提取来说，先将图像进行随机剪裁并调整为 256×256 像素。此外，利用 ResNet152 提取图像特征，其中第 4 层的特征图的维度为 7×7×2 048。对于区域特征提取来说，利用 Faster R-CNN 预测对象的边界框，并对其进行评分，根据分数选择前 K 个显著对象，其中对于每张图像来说 K=36，同时区域特征的维度为 2 048 维。对于文本来说，词嵌入的维度设置为 300。相似性表示的维度设置为 256，尺度参数 α 设置为 9，范围参数 β 设置为 0.2。采用 Adam 优化器来训练本文算法。表 1 为本文算法的网络细节。

3.3 与对比算法进行比较实验

将本文算法与现存算法进行了比较。

深度典型相关性分析 (deep canonical correlation analysis, DCCA) 算法[23]构建了多个堆叠的非线性转换层，并学习了视觉和文本表示的最大化相关性。

深度结构保存图像文本嵌入 (deep structure-

表 1　本文算法网络细节

Table 1　Details network of the proposed algolithm

步骤	输出编号	输入	操作	符号	输出
通用表示提取	①	图像	Faster R-CNN		36×2 048
	②	①	全连接层	$\boldsymbol{I}^1$	36×1 024
	③	图像	ResNet152,view		49×2 048
	④	③	全连接层	$\boldsymbol{I}^2$	49×1 024
	⑤	②④	特征拼接	$\boldsymbol{I}$	85×1 024
	⑥	⑤	自注意力模态	$\bar{\boldsymbol{i}}$	1×1 024
	⑦	文本	单词嵌入空间		L×300
	⑧	⑦	Bi-GRU	$\boldsymbol{T}$	L×1 024
	⑨	⑧	自注意力模态	$\bar{\boldsymbol{t}}$	1×1 024
相似性表示学习	⑩	⑥⑨	式(2)	$\boldsymbol{S}_{\text{glo}}$	1×256
	⑪	②⑧	式(3)～式(5)	$\boldsymbol{S}_{\text{avg}}$	L×256
	⑫	⑩⑪	特征拼接		(L+1)×256
	⑬	⑫	全连接层+sigmoid		1

presserving image-text embeedings, DSPE) 算法[22] 通过构建一个三重排名损失, 增强了跨模态编码特征之间的相关性学习。

GXN 算法[20] 使用学习到的文本或视觉特征生成图像或字幕, 从而减少跨模态信息差距。

改进视觉语义嵌入 (improving visual-semantic embedding, VSE++) 算法[19] 对用于多模态嵌入的常见损失函数进行了简单的修改, 然后与微调和数据增强相融合, 显著提高了检索的性能。

SCAN 算法[11] 提出了堆叠交叉注意力来发现区域和单词的潜在对齐。

场景图匹配 (scene graph matching, SGM) 算法[15] 可以通过聚集邻域信息来细化图上每个节点的表示, 从而获得对象层和关系层的跨模态特征。

全局表示和局部表示 (global representation and local representation, GRLR) 算法[27] 结合了全局和局部 2 种相似度, 可以更好地探索图像和文本之间的语义关系。

这些算法被分为 3 种类型: 1)全局匹配算法: DCCA 算法[23]、DSPE 算法[22]、GXN 算法[20]、VSE++ 算法[19]; 2)区域匹配算法: SCAN 算法[11], SGM 算法[15]; 3)双模态集成算法: SCAN 算法[11]; 4)多级匹配算法: GRLR 算法[27]。本文算法属于多级匹配方法。

3.4　实验结果

本节给出了在 2 个基准数据集上的匹配结果, 旨在证明本文算法的有效性, 如表 2 和表 3 所示, 最佳结果用粗体表示, 次优值用下划线表示。

在 MSCOCO 数据集上的召回率结果如表 2 所示, 由此可知, 本文算法在大多数情况下都优于其他算法。对于 1K 测试集 (1 000 张图像) 来说, 用于图像检测文本任务的 R_1=73.5%, 文本检测图像任务的 R_1=95.4%, 在图像检测文本任务上比次优结果提高了 0.1%, 在文本检测图像任务上比次优结果提高了 0.6%。用于图像检测文本任务的 R_5=98.6%, 文

表 2　在 MSCOCO 数据集上的结果

Table 2　Results on MSCOCO dataset　　%

算法类型	算法	R_1				R_5				R_{10}				R_{sum}	
		图像检测文本		文本检测图像		图像检测文本		文本检测图像		图像检测文本		文本检测图像		文本检测图像	
		1 K 测试集	5 K 测试集	1 K 测试集	5 K 测试集	1 K 测试集	5 K 测试集	1 K 测试集	5 K 测试集	1 K 测试集	5 K 测试集	1 K 测试集	5 K 测试集	1 K 测试集	5 K 测试集
全局匹配算法	DCCA[23]	22.5	6.9	34.6	21.1	45.5	31.8	19.2	6.6	30.4	20.9	41.3	32.2	193.5	119.5
	DSPE[22]	50.1		79.7		89.2		39.6		75.2		86.9		420.7	
	VSE++[19]	64.6	41.3	90.0	71.1	95.7	81.2	52.0	30.3	84.3	59.4	92.0	72.4	478.6	355.7
	GXN[20]	68.5	42.0			97.9	84.7	56.6	31.7			94.5	74.6		
区域匹配算法	SCAN[11]	70.9	46.4	94.5	77.4	97.8	87.4	56.4	34.4	87.0	63.7	93.9	75.7	500.5	384.0
	SGM[15]	73.4	**50.5**	93.8	79.3	97.8	87.9	57.5	35.5	87.3	64.9	94.3	76.5	504.1	393.9
双模态集成算法	SCAN[11]	72.7	50.4	94.8	82.2	98.4	90.0	58.5	**38.6**	88.4	69.3	94.8	80.4	507.9	**410.9**
多级匹配算法	GRLR[27]	68.9		94.1		98.0		58.6		88.2		94.9		320.4	
	本文算法	**73.5**	50.4	**95.4**	**82.4**	**98.6**	**90.2**	**59.0**	37.2	**89.1**	**69.6**	**95.6**	**80.7**	**511.2**	410.5

表3 在Flickr30K数据集上的结果

Table 3 Results on Flickr30K dataset %

算法类型	算法	R_1		R_5		R_{10}		R_{sum}
		图像检测文本	文本检测图像	图像检测文本	文本检测图像	图像检测文本	文本检测图像	文本检测图像
全局匹配算法	DCCA[23]	27.9	26.8	56.9	52.9	68.2	66.9	299.6
	DSPE[22]	40.3	29.7	68.9	60.1	79.9	72.1	351.0
	VSE++[19]	52.9	39.6	80.5	70.1	87.2	79.5	409.8
	GXN[20]	56.8	41.5			89.6	80.1	
区域匹配算法	SCAN[11]	67.9	43.9	89.0	74.2	94.4	82.8	452.2
	SGM[15]	**71.8**	91.7	53.5	79.6	95.5	86.5	**478.6**
双模态集成算法	SCAN[11]	67.4	48.6	90.3	77.7	**95.8**	85.2	465.0
多级匹配算法	GRLR[27]	68.2	43.3	89.1	73.5	94.5	82.5	288.6
	本文算法	68.4	**53.7**	**91.8**	**79.9**	95.3	**86.6**	475.7

本检测图像任务的 R_5=59.0%，在图像检测文本任务上比次优结果提高了0.2%，在文本检测图像任务上比次优结果提高了0.4%。用于图像检测文本任务的 R_{10}=89.1%，文本检测图像任务的 R_{10}=95.6%，在图像检测文本任务上比次优结果提高了0.7%，在Text-to-Image任务上比次优结果提高了0.7%。同时可以看出，这种优势在5K测试集上相似，这证明了本文算法的有效性，同时证明了将全局图像特征和局部图像特征进行融合的有效性。

在Flickr30K数据集上的召回率结果如表3所示，由此可知：本文算法的大多数情况下都优于其他算法，这证明了本文算法的有效性。跟SCAN[11]和SGM算法[15]比较可以发现，本文算法是优于这2种算法的。SCAN[11]和SGM算法[15]属于区域匹配算法，本文算法属于多级匹配算法。该算法不仅引入了区域图像特征，同时还结合了全局图像特征，并将2个特征进行联合，然后将联合后的视觉特征和文本特征进行全局和局部对齐。这证明了本文算法的有效性。和GRLR算法[27]比较可以发现，虽然2个算法都属于多级匹配算法从而进行图像-文本匹配，但本文算法的结果是优于这种算法的，由此可知，将图像的2个特征进行联合的做法是成功，同时将联合后的视觉特征和文本特征进行对齐的做法是成功的。

4 结 论

1）本文算法提取图像的全局特征和图像的区域特征。既保持了图像区域特征中的细粒度语义信息，又将全局的图像信息特别是背景信息融入到算法的学习中。

2）本文算法将2个特征进行联合，得到联合后的视觉特征。对于文本来说，将单词嵌入到GRU网络中，得到文本的表示。

3）本文算法将融合后的视觉特征和文本特征进行全局和局部对齐。大量的实验表明，本文算法有效性优于其他算法。

参考文献（References）

[1] ZHANG C Y, SONG J Y, ZHU X F, et al. HCMSL: Hybrid cross-modal similarity learning for cross-modal retrieval[J]. ACM Transactions on Multimedia Computing, Communications, and Applications, 2021, 17(1s): 1-22.

[2] CHUN S, OH S J, DE REZENDE R S, et al. Probabilistic embeddings for cross-modal retrieval[C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2021: 8411-8420.

[3] YU T, YANG Y, LI Y, et al. Heterogeneous attention network for effective and efficient cross-modal retrieval[C]//Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM, 2021: 1146-1156.

[4] DONG J F, LONG Z Z, MAO X F, et al. Multi-level alignment network for domain adaptive cross-modal retrieval[J]. Neurocomputing, 2021, 440: 207-219.

[5] WANG X, HU P, ZHEN L L, et al. DRSL: Deep relational similarity learning for cross-modal retrieval[J]. Information Sciences, 2021, 546: 298-311.

[6] ZHU L, ZHENG C Q, LU X, et al. Efficient multi-modal hashing with online query adaption for multimedia retrieval[J]. ACM Transactions on Information Systems, 2021, 40(2): 1-36.

[7] WU J L, XIE X X, NIE L Q, et al. Reconstruction regularized low-rank subspace learning for cross-modal retrieval[J]. Pattern Recognition, 2021, 113: 107813.

[8] JIAO F K, GUO Y Y, NIU Y L, et al. REPT: Bridging language models and machine reading comprehension via retrieval-based pre-training[C]//Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021. Stroudsburg: Association for Computational Linguistics, 2021: 150-163.

[9] LU X, ZHU L, LIU L, et al. Graph convolutional multi-modal hashing for flexible multimedia retrieval[C]//Proceedings of the 29th ACM International Conference on Multimedia. New York: ACM,

2021: 1414-1422.

[10] HU Y P, LIU M, SU X B, et al. Video moment localization via deep cross-modal hashing[J]. IEEE Transactions on Image Processing, 2021, 30: 4667-4677.

[11] LEE K H, CHEN X, HUA G, et al. Stacked cross attention for image-text matching[C]//European Conference on Computer Vision. Berlin: Springer, 2018: 212-228.

[12] WANG Y X, CHEN Z D, LUO X, et al. Fast cross-modal hashing with global and local similarity embedding[J]. IEEE Transactions on Cybernetics, 2022, 52(10): 10064-10077.

[13] DIAO H W, ZHANG Y, MA L, et al. Similarity reasoning and filtration for image-text matching[C]//Proceedings of the AAAI Conference on Artificial Intelligence. Washington, DC: AAAI, 2021, 35(2): 1218-1226.

[14] ZOU X T, WANG X Z, BAKKER E M, et al. Multi-label semantics preserving based deep cross-modal hashing[J]. Signal Processing:Image Communication, 2021, 93: 116131.

[15] WANG S J, WANG R P, YAO Z W, et al. Cross-modal scene graph matching for relationship-aware image-text retrieval[C]//2020 IEEE Winter Conference on Applications of Computer Vision. Piscataway: IEEE Press, 2020: 1497-1506.

[16] ZHAN Y W, WANG Y X, SUN Y, et al. Discrete online cross-modal hashing[J]. Pattern Recognition, 2022, 122: 108262.

[17] ZHU L P, TIAN G Y, WANG B Y, et al. Multi-attention based semantic deep hashing for cross-modal retrieval[J]. Applied Intelligence, 2021, 51(8): 5927-5939.

[18] WEN H K, SONG X M, YANG X, et al. Comprehensive linguistic-visual composition network for image retrieval[C]//Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM, 2021: 1369-1378.

[19] FAGHRI F, FLEET D J, KIROS J R, et al. VSE++: Improving visual-semantic embeddings with hard negatives[EB/OL]. (2017-07-08)[2022-01-25].https://arxiv.org/abs/1707.05612v4.

[20] GU J X, CAI J F, JOTY S, et al. Look, imagine and match: improving textual-visual cross-modal retrieval with generative models[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2018: 7181-7189.

[21] WANG H R, ZHANG Y, JI Z, et al. Consensus-aware visual-semantic embedding for image-text matching[C]//European Conference on Computer Vision. Berlin: Springer, 2020: 18-34.

[22] WANG L W, LI Y, LAZEBNIK S. Learning deep structure-preserving image-text embeddings[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2016: 5005-5013.

[23] ANDREW G, ARORA R, BILMES J, et al. Deep canonical correlation analysis[J]. Journal of Machine Learning Research, 2013, 28: 1247-1255.

[24] CHEN H, DING G G, LIU X D, et al. IMRAM: Iterative matching with recurrent attention memory for cross-modal image-text retrieval[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2020: 12652-12660.

[25] YAO T, LI Y R, GUAN W L, et al. Discrete robust matrix factorization hashing for large-scale cross-media retrieval[J]. IEEE Transactions on Knowledge and Data Engineering, 2023, 35(2): 1391-1401.

[26] YAO T, KONG X W, FU H Y, et al. Discrete semantic alignment hashing for cross-media retrieval[J]. IEEE Transactions on Cybernetics, 2020, 50(12): 4896-4907.

[27] 李志欣, 凌锋, 张灿龙, 等. 融合两级相似度的跨媒体图像文本检索[J]. 电子学报, 2021, 49(2): 268-274.
LI Z X, LING F, ZHANG C L, et al. Cross-media image-text retrieval with two level similarity[J]. Acta Electronica Sinica, 2021, 49(2): 268-274 (in Chinese).

Image-text matching algorithm based on multi-level semantic alignment

LI Yiru[1]，YAO Tao[1, 2, *]，ZHANG Linliang[2]，SUN Yujuan[1]，FU Haiyan[3]

(1. School of Information and Electrical Engineering，Ludong University，Yantai 264000，China；
2. Yantai Research Institute of New Generation Information Technology，Southwest Jiaotong University，Yantai 264000，China；
3. School of Information and Communication Engineering，Dalian University of Technology，Dalian 116024，China)

Abstract：The regional features in the image tend to pay more attention to the regional features in the image, and the environmental information is often ignored. How to effectively combine local features and global features has not been fully studied. A image-text maxching algorthm based on multi-level semantic alignment is proposed as a solution to this problem and to improve the association between global concepts and local concepts to provide more accurate visual characteristics. In order to obtain different visual relationship levels and provide more information for the joint visual features, this paper first extracts the local image features to obtain the fine-grained information in the image. It then extracts the global image features to introduce the environmental information into the network learning. To provide a more precise similarity representation, the picture characteristics are next integrated, and finally the combined visual and text features are aligned. Through a lot of experiments and analysis, the effectiveness of the proposed algorithm on two public datasets is proven.

Keywords：image-text matching；cross-modal information processing；feature extraction；neural network；feature fusion

Received：2022-05-19；**Accepted**：2022-06-23；**Published Online**：2022-10-09 15：47

URL：link.cnki.net/urlid/11.2625.V.20221009.1013.008

Foundation items：National Natural Science Foundation of China (61872170,62076052,61873117); National Key R & D Program of China (2021ZD-0113002)

* **Corresponding author**. E-mail：yaotao@ldu.edu.cn

http://bhxb.buaa.edu.cn jbuaa@buaa.edu.cn
DOI: 10.13700/j.bh.1001-5965.2022.0386

基于多级联合的图池化方法

董晓龙[1]，黄俊[1,2,*]，秦锋[1]，洪旭东[1]

(1. 安徽工业大学 计算机科学与技术学院，马鞍山 243000；2. 合肥综合性国家科学中心 人工智能研究院，合肥 230088)

摘 要：图池化方法已经在生物信息学、化学、社交网络、推荐系统等多个领域中得到广泛应用，但关于图池化方法大多没有很好的解决节点选择问题和池化带来的节点信息丢失问题。对此提出一种新的多级联合图池化(MUPool)方法。所提方法使用多视角模块从多个视角获取节点的特征，即通过多个卷积模块提取不同的特征。同时提出多级联合模块（级联），将不同池化层的输出串联，每一层都可以融合以往所有层的信息。提出使用后端融合模块，针对每个池化层建立一个分类器，对预测结果进行融合得到最终分类结果。所提方法在多个数据集上进行实验，准确度平均提高1.62%，所提方法可以与现有的分层池化方法相结合，结合后的方法准确度平均提高2.45%。

关 键 词：图卷积网络；图分类；图池化；深度学习；人工智能
中图分类号：TP37; TP183
文献标志码：A **文章编号**：1001-5965(2024)02-0559-10

在实际应用中，很多数据对象需用图结构来表示，如蛋白质、化学分子、社交网络和网页的推荐系统[1-3]，由此衍生出了图分类学习(graph classification learning, GCL)。

图卷积网络(graph convolution network, GCN)在图像分类方面取得了突破性的进展，研究人员由此将对于图片的卷积延伸到GCN中用于图和节点的分类，并且取得了显著的效果。Kipf和Wellind[4]提出GCN方法，该方法将特征从空间域转到谱域进行计算，复杂度较高，难以扩展。Hamilton和Ying[5]提出GraphSage，来学习图的聚合器，提升了GCN方法的灵活性和收敛速度。Namazi等[6]提出图注意力网络(graph attention network, GAN)方法，该方法使用注意力机制，为图中每个节点分配不同的权值，以提升模型的准确度。但由于这些方法都没有提取层次关系，没考虑到图的结构性信息。

为解决上述问题，研究人员开始使用池化方法。最初，池化层使用在图像卷积网络中。由于池化层有增加感受野、降低模型中参数数量、防止过拟合等特点，因此，研究人员将池化层迁移至图分类任务中。在图分类任务中，池化层可以降低模型的参数和捕捉图的层次结构信息，从而提高模型的准确度同时降低过拟合的风险。

图池化通常是选取或聚合一些节点信息，然后传送到下一层的网络，下一层的网络依旧会执行选取或聚合节点信息，最后，对提取的信息进行处理后再进行分类。最初，研究人员使用谱聚类算法对图节点进行粗化和聚合[7-9]，谱聚类算法虽然可以很好的聚合节点，但是特征分解复杂度较高。Ying等[10]通过学习节点的软分配矩阵从而对节点进行选择，在节点池化的过程中，该方法很难保持邻域的结构信息。目前，研究人员提出了根据图中节点和边的分数进行节点选择的图池化算法，这类算法是通过给定边或节点分数，将得分较高的节点或边聚合到一起，形成新的节点或边[11-14]。Lee等[13]提出基于节点选择的自注意力图池化(self-attention

收稿日期：2022-05-19；录用日期：2022-06-23；网络出版时间：2022-10-18 10:39
网络出版地址：link.cnki.net/urlid/11.2625.V.20221017.1449.006
基金项目：国家自然科学基金(61806005)；安徽高校协同创新项目(GXXT-2020-012)；安徽省教育委员会自然科学基金(KJ2021A0372, KJ2019A0064)
*通信作者. E-mail：huangjun.cs@ahut.edu.cn

引用格式：董晓龙，黄俊，秦锋，等. 基于多级联合的图池化方法[J]. 北京航空航天大学学报，2024，50(2)：559-568.
DONG X L，HUANG J，QIN F，et al. Graph pooling method based on multilevel union[J]. Journal of Beijing University of Aeronautics and Astronautics，2024，50(2)：559-568 (in Chinese).

graph pooling, SAGPool) 方法，该方法通过注意力机制学习节点的得分，同时可以解决图稀疏的问题。Ranjan 等[15] 提出了自注意力框架获取节点信息，同时提出稀疏分配矩阵，可以自适应的捕捉局部子图结构。图池化任务中主要存在的 2 个问题：①节点选择问题，比如仅仅考虑节点信息的重要性或图的结构重要性；②池化导致的节点信息丢失问题，图池化在池化的过程中抛弃节点，这一过程可能会将附带少量重要信息的节点抛弃掉。因此，本文提出了一种新的多级联合图池化 (简称 MUPool) 方法。为解决第 1 个问题，本文从多个视角对图进行特征提取，可以捕获更多有效信息，同时使用后端融合策略进一步保证信息的完整性。为解决第 2 个问题，本文提出使用级联模块，以往的分层池化算法多个层次相对独立，容易丢失信息。本文提出每一层都考虑到上一层节点特征，收集可能丢失的重要节点。

本文主要贡献总结如下：

1）本文提出了多视角信息融合机制，提升提取节点信息的完整性；其次，提出了 MUPool 方法，以克服因池化而造成的信息丢失的问题；最后，将后端融合策略加入 MUPool 方法，融合不同分类器结果，从而获得更准确的预测结果。

2）MUPool 方法中的级联和后端融合机制可以与现有分层池化网络相结合，实验结果表明，MUPool 方法可以大幅度的提升现有分层图池化方法的准确度。

1 相关工作

1.1 图神经网络

已有多种不同的基于图谱和基于空间域的图神经网络方法，图谱卷积是通过傅里叶转换将特征从空间域转到谱域进行计算[16-17]，但由于图谱技术需要进行拉普拉斯特征分解，所以复杂度很高。Defferrard 等[7] 使用切比雪夫多项式进行图卷积，从而减少了复杂度。由于拉普拉斯矩阵是密集的，图小波是稀疏的，文献 [18-22] 使用图小波变换进行图卷积。除了基于图谱的方法外，研究人员也提出了多种基于空间域的图神经网络方法[23-25]。Hamilton 和 Ying[5] 提出了一种能够利用顶点的属性信息产生未知顶点嵌入的一种归纳式图临近变样和特征融合 (graph sample and aggregate, GraphSAGE) 方法。Namazi 等[6] 提出扩展的多分辨率图表示学习 (scalable multi-resolution graph represention learning, SMGRL) 方法，该方法通过在粗化降维图上训练来降低复杂度，同时使用多分辨率图自相似性（self-similarity）提高分类准确度。文献 [26] 提出跳跃知识网络 JK-NET，将每一层的卷积结果都输出到最后一层进行聚合，解决了随着卷积层数增加而出现性能下降的问题。

1.2 图池化

图池化方法大致可以归纳为 3 类。

1）基于拓扑的池化方法。首先，根据图谱理论对图进行特征分解得到粗化图，随后，传输到下一层再进行相同的处理。Gao 等[27] 提出拓扑感知层，投票过程分为 2 个阶段，首先，根据拓扑信息执行本地投票，其次，基于节点的全局重要性执行全局投票。Atwood 和 Towsley[28] 提出对节点使用不同距离采样，随后聚合成一个特征堆叠矩阵，从而进行图分类。

2）全局池化方法。全局池化方法将图的特征信息全部聚合到一个节点上，Vinyals 等[29] 提出一种基于消息传递的图分类网络 Set2Set。Zhang 等[30] 提出 SortPool 方法，该方法通过对图节点的嵌入排序代替相加而最大程度保留节点信息。Takeshi 等[31] 提出多级注意力池化 (multi-level attention pooling, MLAP) 方法，该方法使用基于注意力机制的池化层和逐层表示的聚合机制传递信息。

3）分层池化方法。分层池化可以使用分层卷积池化的方式捕获图每一层的特征和结构信息。Zhang 等[32] 提出分层多视图池化 (hierarchical multi-view graph pooling, MVPool) 方法，该方法也使用了多视角模块，与 MUPool 方法不同的是，MVPool 方法根据不同的上下文信息使用注意力机制对不同视角的节点进行排名。Liu 等[33] 提出分层自适应池化 (hierarchical adaptive pooling, HAP) 方法，该方法使用跨级注意力机制捕获节点的高阶依赖信息，同时该方法提出 GCont 捕获图的属性，以此来保证粗化前后图内容的稳定。Diehl[34] 提出计算图中 2 个节点的边分数，通过聚合最高得分的节点而对图进行池化。Qin 等[35] 根据节点特征分数聚合分数空间中相邻的节点，提出生成粗化图的分层池化方法。Tang 等[36] 提出了基于社区可解释的图池化网络 CommPool 方法，该方法利用无监督学习方法检测图社区级潜在的节点表示。Pang 等[37] 提出将图中节点分为真样本和假样本，通过自注意力机制最大化真样本和假样本之间的互信息以提高方法准确度。Su 等[38] 提出局部胶囊池化网络以获取更多的节点表示信息，同时提出共享信息损失以解决图池化造成的信息丢失问题。Bacciu 等[39] 提出可扩展的 k-plexes 机制以优化特征在池化中上下文传播的方式，并且提出了后处理覆盖方法，处理稀疏无

标度图。

2　MUPool 方法框架

图 1 为 MUPool 方法的网络框架。首先, 使用图卷积对节点特征从多个视角进行提取, 随后, 对节点的特征进行重要性评分并且选择, 与此同时, 将节点的信息进行整合, 融合上一层的节点信息再进行分类, 最后, 将 3 层分类器得到的结果进行后端融合得到节点最终分类结果。该框架包含: 多视角模块、级联模块和后端融合模块。多视角和后端融合模块可以解决节点信息提取不完整的问题, 级联模块可以解决信息丢失的问题。

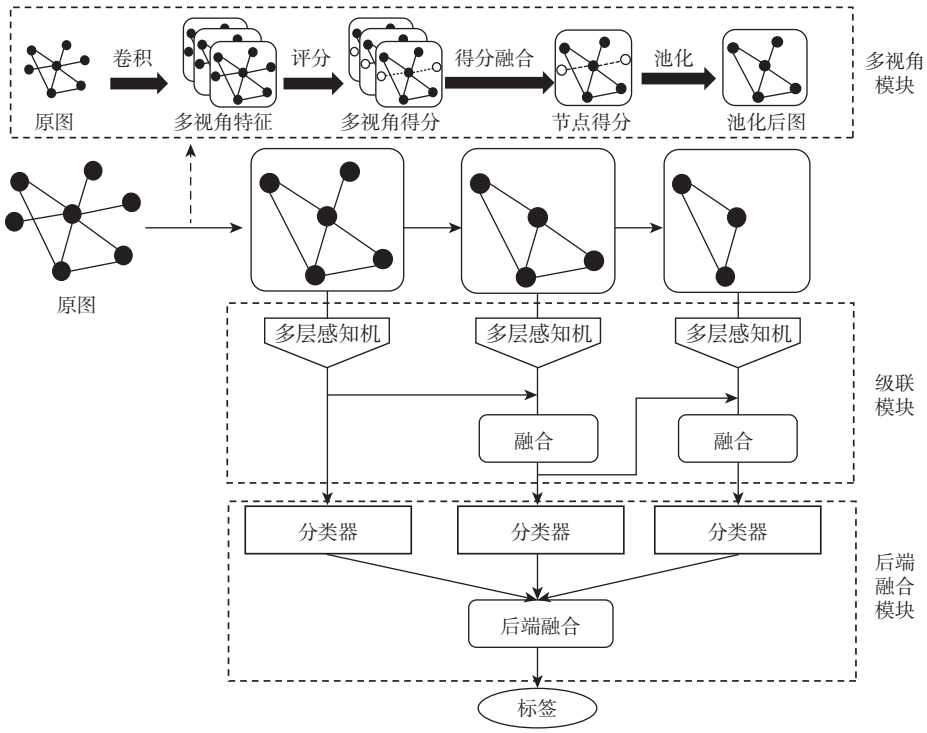


图 1　MUPool 方法的网络框架

Fig. 1　Network framework of MUPool method

2.1　基本定义

定义图$\boldsymbol{G}=(F,N,\boldsymbol{X},\boldsymbol{A})$, 其中$\boldsymbol{G}$表示某个图, F为节点特征维度, N为节点数量, $\boldsymbol{X}\in\mathbf{R}^{N\times F}$为$\boldsymbol{G}$的特征, $\boldsymbol{A}\in\mathbf{R}^{N\times N}$为$\boldsymbol{G}$的边矩阵, 如果 2 个节点$i$和$j$之间有相连的边, 则$A_{ij}$为 1, 否则为 0。将图和标签的集合定义为$C=\{(\boldsymbol{G}_1,y_1),(\boldsymbol{G}_2,y_2),\cdots,(\boldsymbol{G}_n,y_n)\}$, 其中$\boldsymbol{G}_n$为第$n$个图, y_n为对应图的标签。

2.2　多视角模块

多视角信息提取模块可以获取更多不同角度的信息, 提升提取特征的完整度, 特征提取方法为

$$\boldsymbol{X}_i^{l+1}=\sigma_i(\mathrm{GNN}i(\boldsymbol{X}^l,\boldsymbol{A}^l)) \tag{1}$$

式中: i为第i个视角; $\boldsymbol{X}^l\in\mathbf{R}^{N_l\times d}$为第$l$层节点的特征; $\boldsymbol{A}^l\in\mathbf{R}^{N_l\times N_l}$为第$l$层图的边矩阵; $\boldsymbol{X}_i^{l+1}\in\mathbf{R}^{N_{l+1}\times d}$为提取到的$l+1$层第$i$个视角的信息; N_l为第l层的节点数量; $d\in\mathbf{R}$为隐藏层维度; σ_i为激活函数。MUPool 方法中取 3 次卷积, 每个视角采用相同的卷积, 分别是 GNN1、GNN2和 GNN3, 此处的 GNN为 GraphSAGE[5]。

当得到多视角特征信息以后, 对节点进行重要性评分和节点得分融合, 具体如下:

$$\boldsymbol{S}_i^{l+1}=\sigma[\mathrm{Score}(\boldsymbol{X}_i^{l+1})] \tag{2}$$

$$\hat{\boldsymbol{S}}^{l+1}=\frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{\omega}_i\boldsymbol{S}_i^{l+1}) \tag{3}$$

式中: Score 为求特征重要性函数, 具体为GNN, 输出维度为 1, 作用是为每一个节点求一个分数; $\boldsymbol{S}_i^{l+1}\in\mathbf{R}^{N_{l+1}}$为第$l+1$层中第$i$个视角所有节点的分数。为了融合从各个视角提取的信息, 使用式(3)融合多视角分数信息。式(3)中: $\boldsymbol{\omega}_i$为可学习的参数; n为视角数量; $\hat{\boldsymbol{S}}^{l+1}\in\mathbf{R}^{N_{l+1}}$为第$l+1$层所有节点的分数。随后对重要节点进行筛选, 具体如下:

$$k=\mathrm{Max}[(\delta N_l),1] \tag{4}$$

$$\boldsymbol{I}_{\mathrm{idx}}=\mathrm{TopK}(\hat{\boldsymbol{S}}^l,k) \tag{5}$$

$$\boldsymbol{X}^{l+1}=\boldsymbol{X}^l[\boldsymbol{I}_{\mathrm{idx}}] \tag{6}$$

$$\boldsymbol{A}^{l+1}=\boldsymbol{A}^l[\boldsymbol{I}_{\mathrm{idx}}] \tag{7}$$

式中: 超参数δ为池化率; TopK为选择得分最高的k个节点索引的函数; $\boldsymbol{I}_{\mathrm{idx}}$为需要保留节点的索引。

式(6)作用是通过$\boldsymbol{I}_{\mathrm{idx}}$来获取需要保留的节点特征$\boldsymbol{X}^{l+1}\in\mathbf{R}^{N_{l+1}\times d}$，式(7)作用是通过$\boldsymbol{I}_{\mathrm{idx}}$来获取需要保留的边$\boldsymbol{A}^{l+1}\in\mathbf{R}^{N_{l+1}\times N_{l+1}}$，再将获取到的$\boldsymbol{X}^{l+1}$和$\boldsymbol{A}^{l+1}$送到下一层的多视角模块。

2.3 级联模块

研究发现，传统的池化方法提取的特征很容易丢失信息，本文通过级联模块收集池化过程中可能会丢失的信息。

图经过卷积池化后得到对应的重要节点特征$\boldsymbol{X}^{l}$，这些特征会进入多层感知机中生成$\boldsymbol{X}_l'$，如图2所示。再将$\boldsymbol{X}_l'$和上一层输出的结点特征融合得到$\boldsymbol{Z}_l$，即在分类时用到以往所有池化层的信息，具体如下：

$$\boldsymbol{X}_{\mathrm{fu}}^{l}=\sigma(\mathrm{Mean}(\mathrm{Gmp}(\boldsymbol{X}^{l}),\mathrm{Gap}(\boldsymbol{X}^{l}))) \tag{8}$$

$$\boldsymbol{X}_{l}'=\mathrm{MLP}(\boldsymbol{X}_{fu}^{l}) \tag{9}$$

$$\boldsymbol{Z}_l=\begin{cases}\sigma(\boldsymbol{X}_l') & l=1\\ \sigma\dfrac{1}{n_{\mathrm{M}}}\left(\displaystyle\sum_{i=1}^{l-1}\boldsymbol{X}_i'+\boldsymbol{\omega}_l\boldsymbol{X}_l'\right) & l\neq 1\end{cases} \tag{10}$$

式中：Gmp为图最大池化方法；Gap为图平均池化方法；Mean为求平均函数；MLP函数为多层感知机，作用是为了降低维度；$\boldsymbol{X}_{\mathrm{fu}}^{l}\in\mathbf{R}^{N_l\times d}$为级联前节点的特征；$\boldsymbol{X}_l'\in\mathbf{R}^{N_l\times(d/2)}$为降低维度后的节点特征；$\boldsymbol{\omega}_l$为可学习的参数；$n_{\mathrm{M}}$为MUPool方法的池化层数。式(10)是级联部分最主要的方法，由式(10)可知，$\boldsymbol{Z}_l$不仅有l层的信息，还融合了以往所有卷积池化层的信息。在3.5节消项实验中可以明显地看出该模块大幅度捕捉了丢失信息从而提高了MUPool方法的准确度。

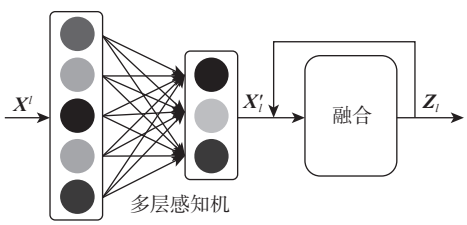


图2 级联模块示意图

Fig. 2 Schematic diagram of multilevel union module

2.4 后端融合模块

本节将详细介绍读出层的操作，因为池化过程中会减少节点数量，所以读出层需要处理不同尺度的图。在以往的方法中，很多方法使用求平均的方式来对读出层进行处理，但是该方式并不能很好的融合图的信息，且容易造成过拟合。由此，本文提出了后端融合机制，具体如下：

$$\boldsymbol{Y}_l=\mathrm{Classify}(\boldsymbol{Z}_l) \tag{11}$$

式中：Classify为分类器；$\boldsymbol{Z}_l$为第l层经过级联后的节点特征；$\boldsymbol{Y}_l\in\mathbf{R}^{N_l\times c}$为第$l$层分类结果；$c$为该数据集类别数。

$$\boldsymbol{Z}=\mathrm{softmax}\left(\frac{1}{h}\sum_{l=1}^{h}\boldsymbol{\omega}_l\boldsymbol{Y}_l\right) \tag{12}$$

式中：h为分类器的数量。将每层分类的结果通过后端融合的方法得到$\boldsymbol{Z}\in\mathbf{R}^{c}$，即图分类结果。在决策层将结果进行后端融合，使模型更加稳定，且不易出现过拟合。

3 实验与分析

本节先介绍实验环境和参数设置、数据集和对比方法，随后进行实验。实验过后将MUPool方法和多个其他方法进行对比，以证明MUPool方法的有效性。

3.1 实验环境和参数设置

使用PyTorch 1.9.0和pytorch-geometric 1.7.2编写代码。实验使用的操作系统为Ubuntu 18.04.5 LTS 64位，具有12核Intel i7-9750H CPU，NVIDIA GeForce RTX 2 060 6G显卡。

为了验证MUPool方法的有效性，在5个数据集上与10个图池化方法进行对比。在每个数据集中，将数据集随机分成3份，分别为训练数据集、验证数据集和测试数据集，比例为8∶1∶1。式(1)中σ_1的参数为0.2，σ_2的参数为0.15，σ_3的参数为0.1。式(4)中池化率δ设置为0.8，在所有数据集上的隐藏层大小d都设置为128维，该参数设定的原因在3.8节参数分析中给出。学习率为10^{-3}，权重衰减为10^{-4}，优化器为Adam，分类器使用的是多层感知机。在训练过程中，MUPool方法在验证数据集上的损失从第t轮之后100轮中一次都没有出现下降时则停止训练，并且取第t轮MUPool方法参数在测试集上进行测试，最后取20次实验的准确度平均值作为最终实验结果。

3.2 实验数据集

实验使用的5个数据集分别为IMDB-B[40]、IMDB-M[40]、PROTEINS[41]、D&D[42]和Mutagenicity[43]，具体参数见表1。IMDB-B[40]是影视数据集，包含爱情电影和动作电影，IMDB-M[40]是IMDB-B[40]扩展数据集，包含爱情、动作和科幻电影，二者皆为社交网络领域数据集。PROTEINS[41]和D&D[42]为生物信息学领域数据集，类别为酶或者非酶。Mutagenicity[43]为化学小分子领域数据集，类别为诱变剂或非诱变剂。

3.3 对比方法

由于基于拓扑池化方法提出较早且复杂度过

表 1　数据集和参数

Table 1　Datasets and parameters

数据集	图数	类别数	平均节点数	平均边数
IMDB-B[40]	1 000	2	19.77	96.53
IMDB-M[40]	1 500	3	13.00	65.54
PROTEINS[41]	1 113	2	39.06	72.82
D&D[42]	1 178	2	284.32	715.66
Mutagenicity[43]	4 337	2	30.32	30.77

大, 逐渐被全局池化和分层池化方法所替代, 本文中不再对比基于拓扑池化方法。以下方法中 Set2Set 方法[29] 和 SortPool 方法[30] 为全局池化方法, 其余皆为分层池化方法。

1) Set2Set 方法[29] 是一个典型的全局池化方法, 其先通过注意力机制得到节点信息, 接着使用长短期记忆网络 (long short-term memory networks, LSTM) 对所有节点进行编码从而生成下一层注意力信息。

2) SortPool 方法[30] 在空间域对图进行卷积从而提取多尺度局部图, 随后使用 SortPooling 层对节点排序, 并且将不重要的节点丢弃。

3) DiffPool 方法[10] 的每一层卷积, 都会生成一个可微分的软分配矩阵, 随后将节点进行聚类, 生成新的粗化图。

4) TopKPool 方法[11] 通过对节点进行排序, 选择排名靠前的 K 个节点, 丢弃排名靠后的节点。

5) SAGPool 方法[13] 是一种端到端的图表示学习方法, SAGPool 方法提出使用自注意力机制来对节点进行重要性评分, 同时考虑了节点特征和图的拓扑信息。

6) MinCutPool 方法[43] 提出了一种规范化快速学习的聚类函数, MinCutPool 可以在不用计算拉普拉斯算子分解的情况下进行图分类。

7) 自适应结构传感知池化 (adaptive structure aware pooling, ASAP) 方法[15] 提出了 Master2Token 自注意力机制, 可以更好的捕获附带重要信息的节点, 随后将相邻节点进行聚类, 删除不重要的类从而达到池化的目的。

8) EdgePool 方法[34] 提出融合节点特征和边的信息。在该方法中, 先计算 2 个相互连接的节点之间的互为相反的 2 条边的得分, 随后, 根据 2 条边的得分取最高分数融合到节点中。

9) GCIPool 方法[37] 提出了通过对正负样本的学习, 池化出真粗化图和假粗化图, 在真分类器上增加分数, 在假分类器上减少分数, 最后融合两者信息从而进行池化。

10) GMT(graph multiset transformer) 方法[44] 提出了基于多头注意力机制的全局池化, 根据节点的结构依赖性学习节点的分数, 保留得分高的节点, 丢弃得分低的节点。

3.4　实验结果

实验结果(准确度±标准差)如表 2 所示, 其中对比方法的准确度摘抄自原文, 最优结果加粗显示, 在多个数据集下, MUPool 方法平均可提高 1.62% 的准确度。

全局池化方法(如 Set2Set[29] 和 SortPool[30])的准确度明显低于分层池化方法, 主要原因是全局池化方法忽略了图的结构信息, 从而降低了方法的准确度。表 2 中 D&D[42] 数据集平均节点相对于其他数据集多, 即图结构复杂, MUPool 方法相较于其他方法有明显的提高, 这得益于 MUPool 方法中级联模块在不同层次之间收集丢失的节点特征, 保证了有效信息的完整性。

分层池化方法(如 SAGPool[13] 和 CGIPool[37])仅从单一角度去获取图的特征信息, 且多个层次之间关联度不足, 造成重要信息捕获不足。此外, 因为

表 2　MUPool 和对比方法在 5 个数据集上的准确度

Table 2　Accuracy of MUPool and comparison method on five datasets　%

方法	IMDB-B[40]	IMDB-M[40]	PROTEINS[41]	D&D[42]	Mutagenicity[43]
Set2Set[29]	72.90 ± 0.75	50.19 ± 0.39	71.46 ± 2.17	71.94 ± 0.56	77.69 ± 0.55
SortPool[30]	70.03 ± 0.86	47.83 ± 0.85	75.54 ± 0.94	79.37 ± 0.97	
DiffPool[10]	73.14 ± 0.70	51.31 ± 0.72	76.25 ± 0.88	80.64 ± 0.72	80.44 ± 0.82
TopKPool[11]	71.58 ± 0.95	48.59 ± 0.72	77.68 ± 2.23	82.43 ± 0.55	79.14 ± 0.76
SAGPool[13]			71.86 ± 0.97	76.45 ± 0.82	79.18 ± 0.82
MinCutPool[43]	72.65 ± 0.75	51.04 ± 0.70	76.05 ± 2.60	80.08 ± 2.3	
ASAP[15]	72.81 ± 0.50	50.78 ± 0.75	74.19 ± 0.97	76.87 ± 0.70	80.12 ± 0.88
EdgePool[34]	72.46 ± 0.74	50.79 ± 0.59	72.50 ± 3.20	75.85 ± 0.58	
CGIPool[37]	72.40 ± 0.87	51.45 ± 0.65	74.10 ± 2.31	73.11 ± 0.93	80.65 ± 0.79
GMT[44]	73.48 ± 0.76	50.66 ± 0.82	75.09 ± 0.59	78.72 ± 0.59	
MUPool	**76.05 ± 1.31**	**53.17 ± 0.62**	**81.65 ± 0.81**	**86.60 ± 0.73**	**81.55 ± 1.16**

每一层节点不同，所以的分类结果也会不同，相比于单个分类器，MUPool方法在每一层设置一个分类器，提高了方法的稳定性和准确度。

3.5 消项实验

为了检验各个模块的重要性，进行了消项实验。MUPool方法一共包含3个模块，即多视角、级联和后端融合模块。在实验过程中，减少其中1个模块而固定另外2个模块。

在多个数据集上实验20次，取准确度平均值，结果如表3所示。表3中从上往下依次包含4组实验，分别是完整的MUPool方法、缺少多视角模块的方法、缺少级联模块的方法和缺少后端融合模块的方法，最优结果加粗显示。实验结果表明，3个模块均是不可缺少的模块，对图分类起着非常重要作用。其次，在缺少级联模块时模型的准确度最差，即级联模块在整个模型中作用最大。

在池化过程中，信息会不可避免的随着层数增加而减少，同时信息也会浓缩，以往的方法虽然也会将每一层的信息加以使用，但是不同层次相对独立，这割裂了不同层次之间的关系。本文提出的级联模块在池化过程中可以重新将已经丢弃的信息通过上一层的信息复用从而加强了不同层次之间的关系，进而导致准确度相较于多视角模块和后端融合模块大幅度上升。

表3 MUPool方法去除某个模块后实验的准确度

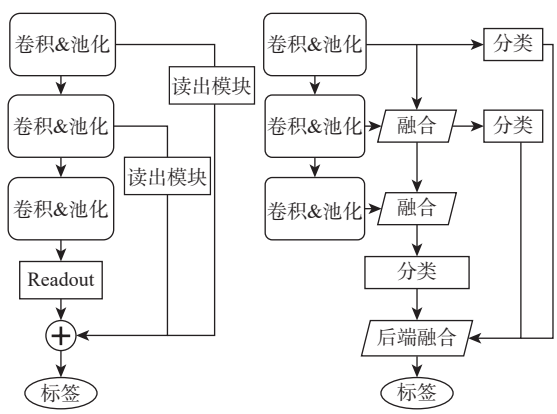
Table 3 Accuracy of MUPool method's experiment after removing a module %

方法	IMDB-B[40]	IMDB-M[40]	PROTEINS[41]	D&D[42]	Mutagenicity[43]
MUPool	**76.05**	**53.17**	**81.65**	**86.60**	**81.55**
缺少多视角模块	74.40	49.50	77.78	80.66	77.74
缺少级联模块	70.65	48.23	75.54	78.51	73.83
缺少后端融合模块	71.51	49.57	76.75	80.36	76.08

3.6 插件实验

由3.5节已知，级联模块和后端融合在MUPool方法中起了非常重要的作用，本节将验证这些模块对现有图池化方法的有效性。鉴于MUPool方法是分层池化方法，所以也选择2个分层池化方法：SAGPool[13]和CGIPool[37]方法，分别加入级联模块和后端融合模块。图3为插件实验，分层池化方法中添加级联模块和后端融合模块以后，在不改变原卷积和池化层代码的情况下，将由如图3(a)所示[13]变成如图3(b)所示框架，其中图3(a)为抽象出大多分层池化方法的共有框架，图3(b)为加入级联和后端融合后的分层池化框架，M&L为级联合后端融合框架。在每个池化层后添加级联模块，随后分类，将分类的结果进行后端融合。在多个数据集上进行了对比实验，结果如表4所示。

由表4可知，级联模块和后端融合模块加入到SAGPool[13]和CGIPool[37]分层池化方法后，其性能都得到显著提升，最优结果加粗表示，在多个数据集下，MUPool方法平均可提高2.45%的准确度。MUPool方法之所以会提升分层池化方法的准确度，是因为分层池化通用框架中不同层次之间相对松散独立，而加入级联模块和后端融合模块以后，



(a) 分层池化通用框架[13] (b) 融合M&L的框架

图3 插件实验

Fig. 3 Plug-in experiment

表4 原始方法和插入M&L模块的分层池化方法的准确度对比

Table 4 Comparison of accuracy between original method and the hierarchical pooling mothod with addition of M&L module %

方法	IMDB-B[40]	IMDB-M[40]	PROTEINS[41]	D&D[42]	Mutagenicity[43]
SAGPool[13]			71.86	76.45	**79.90**
SAGPool+M&L			**82.14**	**80.67**	78.85
CGIPool[37]	72.4	51.45	74.10	73.11	80.65
CGIPool+M&L	**72.6**	**53.43**	**74.94**	**75.63**	**81.29**

不同层次之间的联系显著加强, 这样的结构更加有利于遗漏信息的捕获, 从而提高分类准确度。由此可得, MUPool 方法中的级联模块和后端融合模块具有较强的扩展性, 可作为插件方式与现有的分层池化方法相结合。

3.7　性能分析

在 2 个数据集上, 使用不同的池化率测试不同方法所占用的显存大小。在 PROTEINS[41] 数据集上, 设置批量大小和隐藏层分别为 128、128。在 Mutagenicity[43] 数据集上, 批量大小和隐藏层分别为 256、128。

在运行前, 显存 60 s 稳定不变, 将此时的显存记为运行前显存 A(MByte)。在运行时, 显存 60 s 不发生变化, 将此时的显存记为运行时显存 B(MByte)。计算 B 和 A 的差值记为 1 次显存占用量。运行 10 次取显存占用平均值作为最终显存占用量, 结果显示如图 4 和图 5 所示。根据实验结果, 可以得出以下结论。

1) MUPool 方法在运行时占用显存量较少, 并且不随着池化率的大小而发生显著改变。

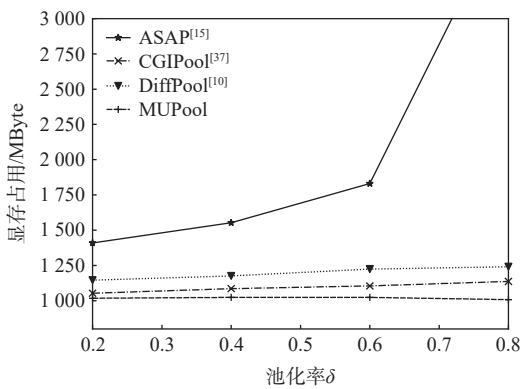


图 4　在 PROTEINS[41] 数据集上池化率对显存的影响

Fig. 4　Effect of pooling rate on video memory in PROTEINS[41] dataset

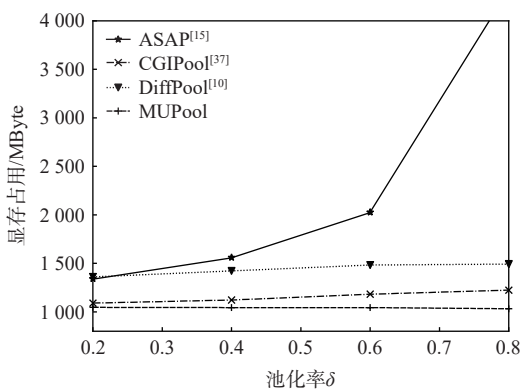


图 5　在 Mutagenicity[43] 数据集上池化率对显存的影响

Fig. 5　Effect of pooling rate on video memory in Mutagenicity[43] dataset

2) 在 PROTEINS[41] 和 Mutagenicity[43] 数据集中, 平均节点数接近, 但是 PROTEINS[41] 数据集的平均边数比 Mutagenicity[43] 数据集高一倍。在这 2 个数据集上, MUPool 方法所占用显存大小接近, 这主要得益于整个网络中都没有用到边的信息, 所以在稠密图上 MUPool 方法具有更大的优势。

对不同方法的运行时间做了对比实验, 在 MUPool 方法中参数设置与 3.1 节中描述一致, 对比方法则是使用原论文中的模型和参数, 如图 6 所示, 横轴为不同的数据集, 其中每个数据集中包括不同的方法, 纵轴为运行时间, 计算方式为计算从训练开始到方法在验证集上收敛的时间差, 其中“×”为时间超过表的上限。

由图 6 可知, MUPool 方法在多个数据集上运行时间均有一定的优势。这得益于 MUPool 方法使用级联处理特征信息而非大量的复杂计算, 同时 MUPool 方法没有使用到关于边的信息, 进一步降低了计算的复杂度。

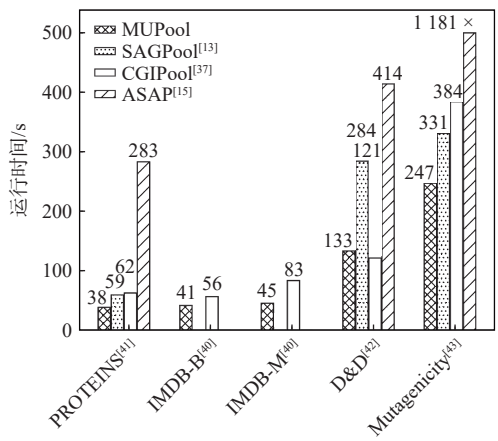


图 6　MUPool 方法与对比方法运行时间对比

Fig. 6　Comparison of running time between MUPool and comparison method

3.8　参数分析

MUPool 方法中有 2 个重要超参, 即池化率和隐藏层大小。本节分别设置了 2 组实验来分析参数敏感性。和 3.1 节中的实验设置一样, 随机划分数据集运行 20 次, 计算平均准确度, 并将其作为最终结果。

图 7 为第 1 组实验的结果, 横坐标为池化率 δ, 纵坐标为准确度。除了 PROTEINS[41] 数据集外, 其余数据集上 MUPool 方法的准确度随着池化率增长而提升。在低池化率时, MUPool 方法准确度较低, MUPool 方法丢弃了较多重要的信息。

图 8 为第 2 组的实验结果, 横坐标代表隐藏层大小 d。由图可知, 随着隐藏层大小 d 增加, MUPool 方法的性能会逐渐提升后下降, 并在 d=128 时取得

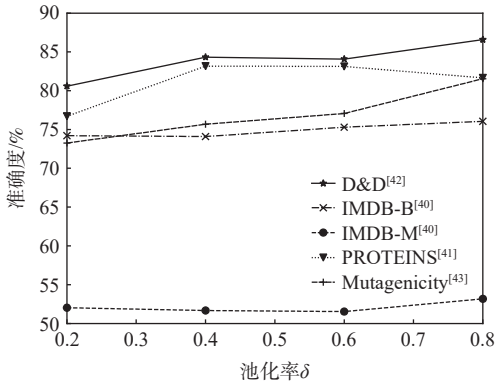


图7 不同池化率在不同数据集上对应的准确度

Fig. 7 Accuracy of different pooling rates on different datasets

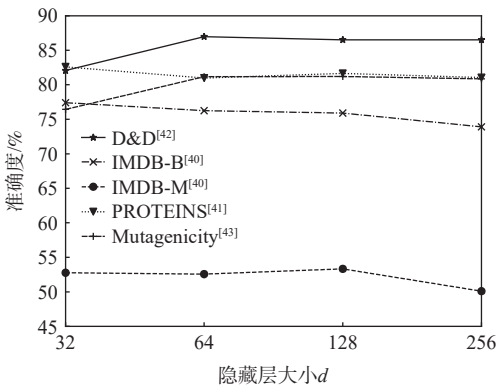


图8 不同隐藏层大小在不同数据集上对应的准确度

Fig. 8 Accuracy of different hidden size on different datasets

最优结果。因此，在本文实验中，隐藏层维度设置为128，池化率设置为0.8。

4 结 论

1）MUPool方法可提高图分类任务的准确度，在多个数据集下，平均可提高1.62%的准确度。

2）MUPool方法可以作为插件加入到分层池化模型中，在多个数据集下，平均可提高2.45%的准确度。

3）MUPool方法并未涉及到大量复杂计算，所以显存占用量低，且计算时间短。

参考文献（References）

[1] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large scale image recognition[C]//Proceedings of the 3th International Conference on Learning Representations. San Diego: ICLR, 2015.

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

[3] ZHANG X, ZHAO J B, LECUN Y. Character-level convolutional networks for text classification[C]//Proceedings of the 28th International Conference on Neural Information Processing Systems. New York: ACM, 2015: 649-657.

[4] KIPF T N, WELLING M. Semi-supervised classification with graph convolutional networks[EB/OL]. (2017-02-22) [2022-01-06]. http://arxiv.org/abs/1609.02907.

[5] HAMILTON W L, YING R, LESKOVEC J. Inductive representation learning on large graphs[C]//Proceedings of the 31st International Conference on Neural Information Processing Systems. New York: ACM, 2017: 1025-1035.

[6] NAMAZI R, GHALEBI E, WILLIAMSON S, et al. SMGRL: A scalable multi-resolution graph representation learning framework [EB/OL]. (2022-01-29) [2022-02-02]. https://arxiv.org/abs/2201.12670.

[7] DEFFERRARD M, BRESSON X, VANDERGHEYNST P. Convolutional neural networks on graphs with fast localized spectral filtering[C]//Proceedings of the 30th International Conference on Neural Information Processing Systems. New York: ACM, 2016: 3844-3852.

[8] RHEE S, SEO S, KIM S. Hybrid approach of relation network and localized graph convolutional filtering for breast cancer subtype classification[C]//Proceedings of the 27th International Joint Conference on Artificial Intelligence. California: International Joint Conferences on Artificial Intelligence Organization, 2018: 3527-3534.

[9] MA Y, WANG S H, AGGARWAL C C, et al. Graph convolutional networks with EigenPooling[C]//Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. New York: ACM, 2019: 723-731.

[10] YING R, YOU J X, MORRIS C, et al. Hierarchical graph representation learning with differentiable pooling[C]//Proceedings of the 32nd International Conference on Neural Information Processing Systems. New York: ACM, 2018: 4805-4815.

[11] GAO H Y, JI S W. Graph u-nets[C]//IEEE Transactions on Pattern Analysis and Machine Intelligence. Piscataway: IEEE Press, 2022: 4948-4960.

[12] CANGEA C, VELIČKOVIĆ P, JOVANOVIĆ N, et al. Towards sparse hierarchical graph classifiers[EB/OL]. (2018-11-03) [2022-02-03]. https://doi.org/10.48550/arXiv.1811.01287.

[13] LEE J, LEE I, KANG J. Self-attention graph pooling[C]//Proceedings of the International Conference on Machine Learning. New York: ACM, 2019: 3734-3743.

[14] DIEHL F, BRUNNER T, LE M T, et al. Towards graph pooling by edge contraction[C]//Proceedings of the International Conference on Machine Learning. New York: ACM, 2019.

[15] RANJAN E, SANYAL S, TALUKDAR P. ASAP: Adaptive structure aware pooling for learning hierarchical graph representations[C]//Proceedings of the AAAI Conference on Artificial Intelligence. Washton, D.C.: AAAI, 2020, 34(4): 5470-5477.

[16] BRUNA J, ZAREMBA W, SZLAM A, et al. Spectral networks and locally connected networks on graphs[EB/OL]. (2014-03-21) [2022-02-09]. https://arxiv.org/abs/1312.6203v3.

[17] CHUNG F R K, GRAHAM F C. Spectral graph theory[M]. Providence, R.I.: American Mathematical Society, 1997.

[18] XU B, SHEN H, CAO Q, et al. Graph wavelet neural network[C]//Proceedings of the 7th International Conference on Learning Representations. New Orleans: ICLR, 2019.

[19] LI M, MA Z, WANG Y G, et al. Fast haar transforms for graph neural networks[J]. Neural Networks, 2020, 128: 188-198.

[20] MA X, WU G, KIM W H Multi-resolution graph neural network for identifying disease-specific variations in brain connectivity[EB/OL]. (2019-12-03) [2022-02-03]. https://arxiv.ORG/ABS/1912.01181.

[21] WANG Y G, LI M, MA Z, et al. Haar graph pooling[C]//Proceedings of the 37th International Conference on Machine Learning. New York: ACM, 2020: 9952–9962.

[22] ZHENG X, ZHOU B, WANG Y G, et al. Decimated framelet system on graphs and fast g-framelet transforms[J]. Journal of Machine Learning Research, 2022, 23(18): 1-18.

[23] SCARSELLI F, GORI M, TSOI A C, et al. The graph neural network model[J]. IEEE Transactions on Neural Networks, 2009, 20(1): 61-80.

[24] TARLOW Y D, BROCKSCHMIDT M, ZEMEL R. Gated graph sequence neural networks[C]//Proceedings of the 4th International Conference on Learning Representations. San Juan: ICLR, 2016.

[25] GALLICCHIO C, MICHELI A. Fast and deep graph neural networks[C]//Proceedings of the AAAI Conference on Artificial Intelligence. Washton, D.C.: AAAI, 2020, 34(4): 3898-3905.

[26] CHEN W R, LUO C Y, WANG S K, et al. Representation learning with complete semantic description of knowledge graphs[C]//Proceedings of the 2017 International Conference on Machine Learning and Cybernetics. Piscataway: IEEE Press, 2017: 143-149.

[27] GAO H Y, LIU Y, JI S W. Topology-aware graph pooling networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021, 43(12): 4512-4518.

[28] ATWOOD J, TOWSLEY D. Diffusion-convolutional neural networks[C]//Proceedings of the 30th International Conference on Neural Information Processing Systems. New York: ACM, 2016: 2001-2009.

[29] VINYALS O, BENGIO S, KUDLUR M. Order matters: Sequence to sequence for sets[C]//Proceedings of the 4th International Conference on Learning Representations. San Juan: ICLR, 2016.

[30] ZHANG M, CUI Z, NEUMANN M, et al. An end-to-end deep learning architecture for graph classification[C]//Proceedings of the AAAI Conference on Artificial Intelligence. Menlo Park: AAAI, 2018: 4438-4445.

[31] ITOH T D, KUBO T, IKEDA K. Multi-level attention pooling for graph neural networks: Unifying graph representations with multiple localities[J]. Neural Networks, 2022, 145: 356-373.

[32] ZHANG Z, BU J J, ESTER M, et al. Hierarchical multi-view graph pooling with structure learning[J]. IEEE Transactions on Knowledge and Data Engineering, 2023, 35(1): 545-559.

[33] LIU N, JIAN S L, LI D S, et al. Hierarchical adaptive pooling by capturing high-order dependency for graph representation learning[J]. IEEE Transactions on Knowledge and Data Engineering, 2023, 35(4): 3952-3965.

[34] DIEHL F. Edge contraction pooling for graph neural networks [EB/OL]. (2019-05-27) [2022-03-09]. https://arxiv.org/abs/1905.10990.

[35] QIN J, LIU L, SHEN H, et al. Uniform pooling for graph networks[J]. Applied Sciences, 2020, 10(18): 6287-6301.

[36] TANG H T, MA G X, HE L F, et al. CommPOOL: An interpretable graph pooling framework for hierarchical graph representation learning[J]. Neural Networks, 2021, 143: 669-677.

[37] PANG Y S, ZHAO Y X, LI D S. Graph pooling via coarsened graph infomax[C]//Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM, 2021: 2177-2181.

[38] SU Z D, HU Z H, LI Y D. Hierarchical graph representation learning with local capsule pooling[C]//ACM Multimedia Asia. New York: ACM, 2021.

[39] BACCIU D, CONTE A, GROSSI R, et al. K-plex cover pooling for graph neural networks[J]. Data Mining and Knowledge Discovery, 2021, 35(5): 2200-2220.

[40] YANARDAG P, VISHWANATHAN S V N. A structural smoothing framework for Robust graph-comparison[C]//Proceedings of the 28th International Conference on Neural Information Processing Systems. New York: ACM, 2015: 2134-2142.

[41] FERAGEN A, KASENBURG N, PETERSEN J, et al. Scalable kernels for graphs with continuous attributes[C]//Proceedings of the 26th International Conference on Neural Information Processing Systems. New York: ACM, 2013: 216-224.

[42] DOBSON P D, DOIG A J. Distinguishing enzyme structures from non-enzymes without alignments[J]. Journal of Molecular Biology, 2003, 330(4): 771-783.

[43] KAZIUS J, MCGUIRE R, BURSI R. Derivation and validation of toxicophores for mutagenicity prediction[J]. Journal of Medicinal Chemistry, 2005, 48(1): 312-320.

[44] BAEK J, KANG M, HWANG S J. Accurate learning of graph representations with graph multiset pooling[C]//Proceedings of the 9th International Conference on Learning Representations. Vienna: ICLR, 2021.

Graph pooling method based on multilevel union

DONG Xiaolong[1], HUANG Jun[1, 2, *], QIN Feng[1], HONG Xudong[1]

(1. School of Computer Science and Technology, Anhui University of Technology, Maanshan 243000, China;
2. Institute of Artificial Intelligence, Hefei Comprehensive National Science Center, Hefei 230088, China)

Abstract: Graph pooling method has been widely used in bioinformatics, chemistry, social networks, recommendation systems and other fields. At present, the graph pooling method does not solve the problem of node selection and node information loss caused by pooling. A new graph pooling method is proposed, namely the graph pooling method based on multilevel union (MUPool). The suggested technique extracts distinct features from several convolution modules by using a multi-view module to obtain the properties of nodes from various angles. At the same time, a multilevel union module is proposed to concatenate the outputs of different pooling layers, each layer fusing information from all previous layers. The suggested approach builds a classifier based on each pooling layer using the late fusion module, then fuses the predicted results to obtain the final classification results. The proposed method is tested on multiple data sets, and the accuracy is improved by 1.62% on average, the proposed method can be combined with the existing hierarchical pooling method, the accuracy of the combined method is improved by 2.45% on average.

Keywords: graph convolutional network; graph classification; graph pooling; deep learning; artificial intelligence

Received: 2022-05-19; **Accepted**: 2022-06-23; **Published Online**: 2022-10-18 10:39

URL: link.cnki.net/urlid/11.2625.V.20221017.1449.006

Foundation items: National Natural Science Foundation of China (61806005); The University Synergy Innovation Program of Anhui Province (GXXT-2020-012); Natural Science Foundation of the Educational Commission of Anhui Province (KJ2021A0372,KJ2019A0064)

* **Corresponding author**. E-mail: huangjun.cs@ahut.edu.cn

http://bhxb.buaa.edu.cn jbuaa@buaa.edu.cn
DOI: 10.13700/j.bh.1001-5965.2022.0387

联合方面注意力交互的图文方面类情感识别

赵一成[1]，王素格[1,2,*]，廖健[1]，何东欢[1]
(1. 山西大学 计算机与信息技术学院，太原 030006；2. 山西大学 计算智能与中文信息处理教育部重点实验室，太原 030006)

摘 要：随着多媒体的快速发展，单纯采用文本的方面类情感分析，不能准确识别用户所表达的情感。而现有图文数据的方面类情感分析方法仅考虑图文模态间的交互，忽略图文数据的不一致性和相关性。因此，提出联合方面注意力交互网络(JAAIN)模型的图文方面类情感识别方法。所提方法针对图文数据的不一致性与相关性，通过多层次融合方面信息和图文信息，去除与给定方面无关的文本和图像，增强给定方面的图文模态数据的情感表示，将文本数据情感表示、图像数据情感表示及方面类情感表示进行拼接融合与全连接，实现图文方面类情感判别。在数据集Multi-ZOL上进行实验，实验结果表明：所提模型能够提升图文方面类情感判别的性能。

关 键 词：方面类情感分析；注意力机制；多模态情感分析；情感表示；多模态融合
中图分类号：TP391
文献标志码：A **文章编号**：1001-5965(2024)02-0569-10

随着Web 2.0技术的发展和手机拍照功能的日益升级，人们在用文字表达的同时会附上一张或多张相关的图片，使用户的情感表达方式以多模态形式呈现。方面级情感分析是文本情感分析领域的一个重要任务，一般分为2个子任务[1]：方面/属性类情感分析(aspect-category sentiment analysis, ACSA)和方面/属性项情感分析(aspect-term sentiment analysis, ATSA)。其中，方面类情感分析任务是给定一段文本和一个方面类，推断给定方面类的情感极性。由于方面类情感分析属于细粒度情感分析，在一个文本片段中可能连续出现多个方面的情感表达，而相应的图片也会附上多个，另一方面，还会出现文本或图片与方面类无关的情况。

根据Xu等[2]的统计显示，ZOL.com近40%的手机评论包含文本和图像，图文评论得到回复数是纯文本评论的3倍以上，说明人们更青睐于图文数据。因此，如何将文本和图像2种模态的数据进行有效融合，从而精确捕捉其情感逐渐受到研究者的关注[2-4]。

对于图文数据，研究者们[3-4]将图像融入经典的情感分析中，相比于单一文本内容数据可以得到更高的准确率。

现有多模态方面级情感分析方法，通常仅考虑对图文数据的交互，忽视了图文数据的不一致性和相关性[2]，本文提出了一种联合方面注意力交互网络(joint aspects attention interaction network, JAAIN)模型的方面类情感判别方法，从多个层次充分融合方面信息和图文信息，分别对图文数据的不一致性和相关性建模，使本文模型可以对图文数据与方面信息相互关联并进行有效表示。

本文的主要贡献如下：

1）通过分析给定的图文数据，发现文本与图像数据之间具有不一致性和相关性，为此，提出了一种JAAIN模型，从而增强给定方面的图文模态数据表示。

2）在模型建立时，采用多层次融合方面信息与

收稿日期：2022-05-19；录用日期：2022-07-02；网络出版时间：2022-10-18 10:42
网络出版地址：link.cnki.net/urlid/11.2625.V.20221017.1403.003
基金项目：国家自然科学基金(62076158,61906112)；山西省太原市小店区科技局项目(2020XDCXY05)
*通信作者. E-mail：wsg@sxu.edu.cn

引用格式：赵一成，王素格，廖健，等. 联合方面注意力交互的图文方面类情感识别[J]. 北京航空航天大学学报，2024，50(2)：569-578.
ZHAO Y C，WANG S G，LIAO J，et al. Image-text aspect emotion recognition based on joint aspect attention interaction[J]. Journal of Beijing University of Aeronautics and Astronautics，2024，50(2)：569-578 (in Chinese).

图文信息，去除与给定方面无关的上下文和图像，从而降低了图文信息中与方面无关部分对本文模型性能的影响，并利用 Transformer-encoder 的特性[5]，捕获图文数据内部信息。

1 相关工作

针对图文数据的方面级情感分析，与其密切相关的研究主要有文本方面级情感分析、图像情感分析和图文情感分析。

1.1 文本方面级情感分析

在 ACSA 任务中，循环神经网络和标准注意力机制被较多地使用，用于自动学习上下文和方面词的语义特征。注意力机制具有聚焦句子重要部分的能力。Gu 等[6] 使用注意力机制关注方面词和句子之间的语义关系。Xu 等[7] 使用了全局注意力和局部注意力去捕获方面和上下文之间不同粒度的交互信息。Wu 等[8] 提出了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。

1.2 图像情感分析

图像方面类情感分析的研究比较少，本文分别从与之相关的面部表情识别任务和视觉情感分析任务进行介绍。在面部表情识别任务中，许多研究者[9-10] 发现利用卷积神经网络可以有效提高表情识别准确率。Li 等[9] 提出了一个带有注意力机制的卷积神经网络，该网络将注意力主要关注于面部未被遮挡的部分，将全局特征与面部未被遮挡部分特征结合。Xie 等[10] 利用注意力对卷积神经网络进行了有针对性的改进，突出表情区域。视觉情感分析相关研究表明，人们对图像情感的认知会受到图像不同层次特征的影响[11]，当人们看到高质量的图像时，通常会产生积极的情感。Rao 等[12] 提出了多层深度表示网络，将卷积网络的不同卷积层进行外接，以提取各层的图像特征，再将这些特征进行拼接，有效地融合了不同层次的图像情感信息。

综合 2 个任务的研究进展，通过卷积网络提取图像特征和对图像使用注意力机制，在图像情感分析任务中取得较好的结果。

1.3 图文情感分析

对于图文方面类情感分析，相关的研究还比较少，而在图文情感分析中，图文的融合策略是解决图文情感分析任务的关键。根据图文数据的融合策略，可分为特征级融合[13]、混合融合[14-15] 和决策级融合[16]。特征级融合是在进行情感打分之前，将多个模态来源的数据整合成一个整体的特征向量[13]。特征级融合可能产生高维冗余的特征向量。决策级融合是发生在每种模态建模之后，将多个模态情感分类器的决策结果进行平均计算[15]。决策级融合忽略了模态之间的特征交互。现有的图文情感分析方法大多采用融合的方法，其通过共享神经网络表示层的特征向量实现模态间的特征交互，并设计给定的连接单元来实现多个模态特征的有效融合。Truong 和 Lauw[15] 认为图像信息有助于识别文本中的重要句子，利用注意力机制将图像信息作为指明文档中重要句子的对齐手段，提出了一种图像方面注意力网络。Xu 等 [16] 提出了一种图像的区域特征与文本的多个语义层次相互作用的双向多层次注意模型，模型使用了图像注意力网络和语义注意力网络。因此，针对性使用注意力机制可以使图文信息有效对齐。

图文方面类情感分析任务是细粒度图文情感分析，Xu 等 [3] 较早提出了图文方面类情感分析任务，并提出多交互记忆网络模型，捕获多模态间和单模态内的交互信息，但忽略了图文数据的特性。受文本方面类情感分析、图像情感分析和图文情感分析 3 种任务方法的启发，本文提出了 JAAIN 模型，通过多层次融合方面信息和图文信息，对给定方面类图文数据的不一致性与相关性建模，去除与给定方面无关的上下文和图像，增强给定方面的图文 2 种模态数据表示。

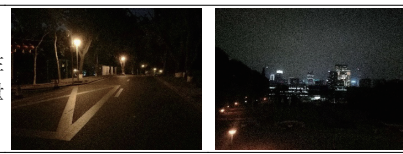
2 图文多模态数据的不一致性和相关性

通过分析现有图文数据，发现文本与图像数据之间具有不一致性和相关性，比如 Multi-ZOL 数据集中的 3 个有关手机“拍照效果”方面的实例，如表 1 所示。

实例 1 中的文本，用户并没有发表“拍照效果”方面的观点，而相应的照片中存在景色模糊不清，有大量噪点，因此，“拍照效果”方面的情感为负面极性，这说明了文本与图片数据之间具有不一致性。实例 2 中，文本对“拍照效果”方面并没有直接的情感表达，但本例伴随着一个高质量清晰的照片，因此，“拍照效果”方面的情感应为正面极性，说明文本数据和图像数据具有情感相关性。实例 3 中，文本对“拍照效果”方面明确表达了正向情感，并配了高质量清晰的照片，但图片的主题多样。由此可以看出，作为文本中的情感，表达细腻且语言表达风格多样，若结合图片数据可以增强给定方面类或项的情感表示的性能，而仅靠图片，难以确定其方面类型，需要文本中对方面类或项的描述。

表 1　图文数据不一致性和相关性样例

Table 1　Examples of image and text data inconsistencies and correlations

实例	文本	图像	方面	情感标签
1	手机挺好的, 黑色挺酷, 显示不错, 还有全程的护眼模式, 外观漂亮, 运行流畅, 就是输入法不太适应, 电池一天一冲充肯定要的, 就是在淘宝官方旗舰店买没有送耳机挺不开心的, 心理不平衡！电池容量小, 就这个缺点吧。		拍照效果	负面
2	1) 外观低调、沉稳、内敛、商务感十足, 深受男人喜爱; 2) 分屏、显示、色彩在这个价位上表现还是很不错的; 3) 提供摆拍姿势指导, 支持前置柔光灯; 4) 手机续航不错, 用一天妥妥的; 5) 支持多个知名品牌耳机音效。		拍照效果	正面
3	配置高, 性能强悍。顶级相机, 成像质量佳。不支持快充, 充电慢。后置摄像头是iPhoneSE的一大亮点, 延续了iPhone6s的顶级配置, 虽然没有光学防抖, 但依旧让我十分兴奋。		拍照效果	正面

针对图文数据中存在的不一致性和相关性特点, 本文提出了 JAAIN 模型, 通过多层次融合方面信息和图文信息增强给定方面的图文模态数据的情感表示, 去除与给定方面无关的图文信息。

3　联合方面注意力交互网络

图文方面类情感分析任务可以形式化定义为: 输入文本内容 $\boldsymbol{T}=(\boldsymbol{t}_1,\boldsymbol{t}_2,\cdots,\boldsymbol{t}_l)$ 和图像集 $\boldsymbol{V}_{\mathrm{I}}=(\boldsymbol{v}_1,\boldsymbol{v}_2,\cdots,\boldsymbol{v}_k)$, 通过建立模型, 预测给定方面词 $\boldsymbol{A}=(\boldsymbol{a}_1,\boldsymbol{a}_2,\cdots,\boldsymbol{a}_n)$ 的情感标签, 其中, l 为文本内容的长度, k 为图像的数量, n 为方面词的长度。JAAIN 模型的方面类情感判别模型的整体框架如图 1 所示, 包括图文特征抽取层、多层次方面交互层和图文特征融合分类层, 其中, $\boldsymbol{Q}$、$\boldsymbol{K}$、$\boldsymbol{V}$ 分别为注意力中的 Query、Key、Value, $\boldsymbol{M}^{\mathrm{ar}}$ 为图像注意力情感表示, $\boldsymbol{M}^{\mathrm{at}}$ 为文本注意力表示, $\boldsymbol{M}^{\mathrm{Ttar}}$ 为文本模态特征表示, n_{i} 为图片数量, DAIMA 表示方面直接交互的多头注意力(direct aspect information multi-head attention, DAIMA)机制, ADIMA 表示方面信息的深层交互(aspect deep interaction multi-head attention, ADIMA)机制。

3.1　Transformer-encoder 模型

Transformer-encoder 是一个基于注意力的序列模型, 其将句子中的单词转换成词向量, 然后, 通过多头自注意力模块, 残差连接和层归一化得到全局自注意力特征, 最后, 通过前馈网络, 残差连接和层归一化获得输出。该模型的自注意力机制能够挖掘全局依赖关系, 具备不同特征空间及全局范围的特征聚合能力的特点, 在多模态任务中已经证明了其有效性[17]。

3.2　图文特征的抽取

图文特征的抽取分为图像特征抽取、文本特征抽取和方面特征抽取。

3.2.1　图像特征嵌入表示

为了充分利用图像信息, 对于输入的图像集 $\boldsymbol{V}_{\mathrm{I}}=(\boldsymbol{v}_1,\boldsymbol{v}_2,\cdots,\boldsymbol{v}_k)$, 本文采用了卷积神经网络 ResNet-152 图像识别方法提取其图像特征。从图像集中依次取一张具体图片 $\boldsymbol{v}_i(i=1,2,\cdots,k)$, 将其重新调整为统一像素大小的图像 $\boldsymbol{v}_i'$。使用预先训练的 ResNet-152, 得到图像特征向量 $\hat{\boldsymbol{R}}$。

$$\hat{\boldsymbol{R}}=\mathrm{ResNet}(\boldsymbol{v}_i') \tag{1}$$

式中: $\hat{\boldsymbol{R}}$ 为三阶张量。

3.2.2　文本特征嵌入表示

BERT(bidirectional transformers) 模型[18] 可以从大型语料库中使用预先训练的模型参数产生动态的上下文嵌入表示。给定输入的文本内容 $\boldsymbol{T}=(\boldsymbol{t}_1,\boldsymbol{t}_2,\cdots,\boldsymbol{t}_l)$, 通过 BERT 预训练模型得到文本内容的嵌入表示 $\vec{\boldsymbol{T}}=(\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_l)$, 其中, 每个字的向量表示 $\boldsymbol{t}_i\in\mathbf{R}^d$, d 为向量的维度。

通过 ResNet-152 得到的图像特征层次远深于通过 BERT 模型得到的文本特征, 为使图像与文本的特征表示较为接近。本文利用 Transform-encoder 模型, 对 BERT 模型得到的文本特征 $\vec{\boldsymbol{T}}$ 进行文本重聚合, 从而使文本特征蕴含更深层的语义信息, 将文本重聚合的特征标记为 $\hat{\boldsymbol{T}}$, 其定义如下:

$$\hat{\boldsymbol{T}}=\text{Transform-encoder}(\vec{\boldsymbol{T}}) \tag{2}$$

3.2.3　方面特征嵌入表示

与文本特征嵌入表示类似, 本文仍然利用 BERT 模型学习方面词的特征嵌入表示。给出输入的方面词 $\boldsymbol{A}=(\boldsymbol{a}_1,\boldsymbol{a}_2,\cdots,\boldsymbol{a}_n)$, 通过 BERT 预训练模型得到方面特征的嵌入表示 $\vec{\boldsymbol{A}}=(\boldsymbol{d}_1,\boldsymbol{d}_2,\cdots,\boldsymbol{d}_n)$, 其中, 每个字的向量表示 $\boldsymbol{d}_i\in\mathbf{R}^c$, c 为向量的维度。

3.3　多层次方面的交互

由于图像数据和文本数据之间具有不一致性和相关性, 因此, 图像数据在为 ACSA 任务提供丰富的补充信息的同时, 也应避免引入与方面无关的噪音图文信息。针对这 2 个问题, 分别构建方面直

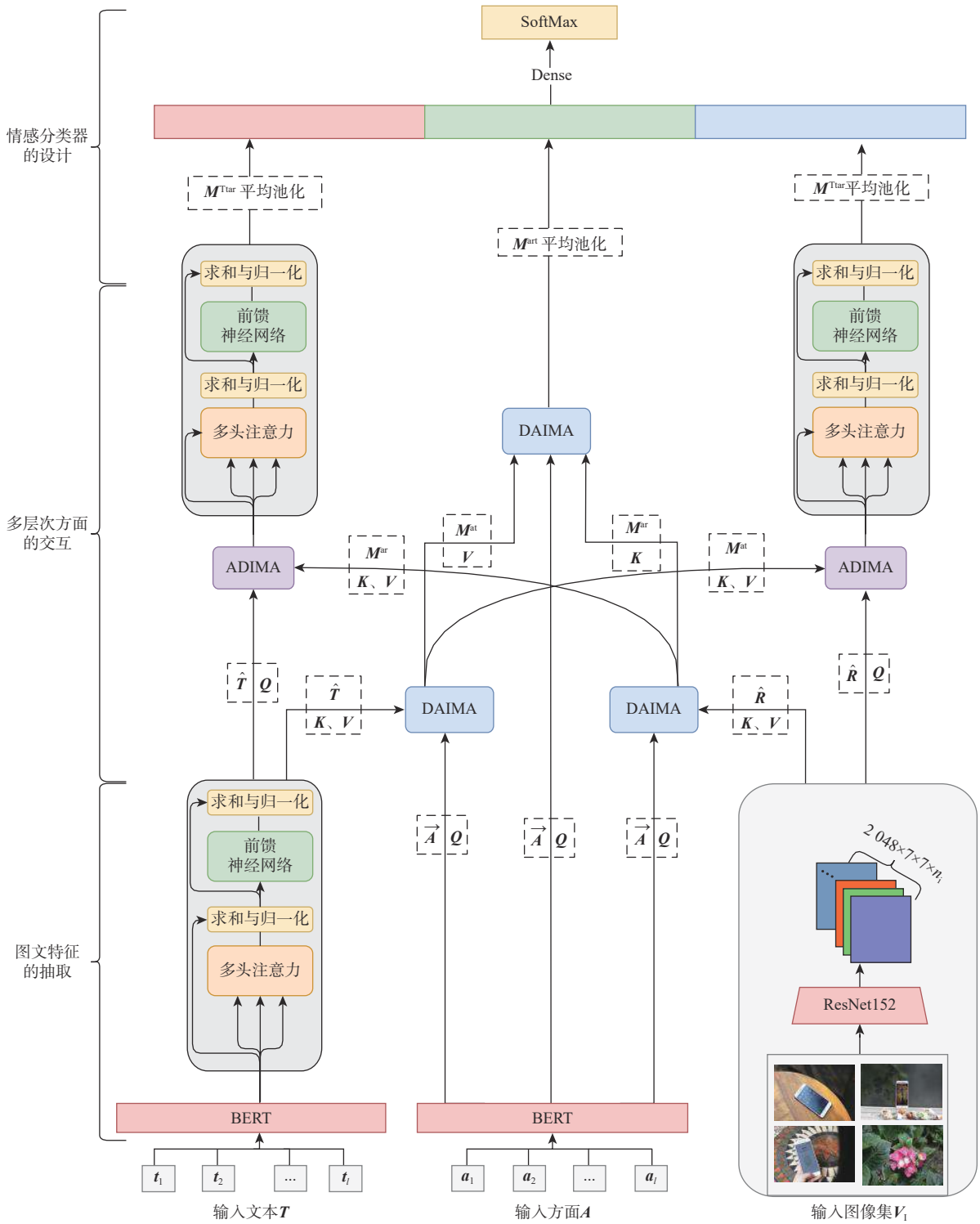


图 1 JAAIN 模型整体框架

Fig. 1 Overall framework of JAAIN model

接交互机制和方面深层交互机制。

3.3.1 方面直接交互机制的建立

针对图文模态数据的不一致性，建立 DAIMA 机制，利用 DAIMA 机制，去除与给定方面无关的文本和图像，并保留最相关的部分，再次利用 DAIMA 机制，将给定方面的图像表示和文本表示及方面信息三者直接交互，从而得到给定方面的图文注意力情感表示。

对于方面$\vec{A}$和文本内容$\hat{T}$，本文假设$\vec{A}$作为 Query，$\hat{T}$作为 Key 和 Value，给定方面与文本进行交互，获得给定方面的文本注意力情感表示M^{at}为

$$M^{at}=\text{MultiHead}\left(\vec{A},\hat{T},\hat{T}\right)=\text{Concat}\left(H_1^A,H_2^A,\cdots,H_h^A\right)W^{AT} \tag{3}$$

$$H_i^A = \text{Attention}\left(\vec{A}\boldsymbol{W}_i^{AQ}, \hat{T}\boldsymbol{W}_i^{AK}, \hat{T}\boldsymbol{W}_i^{AV}\right) \quad (4)$$

式中: H 为多头; h 为多头注意力头数; $d_k = d_v = d/h$; $\boldsymbol{W}^{AT} \in \mathbf{R}^{h\times d_v\times d}$, $\boldsymbol{W}_i^{AQ} \in \mathbf{R}^{d\times d_k}$, $\boldsymbol{W}_i^{AK} \in \mathbf{R}^{d\times d_k}$, $\boldsymbol{W}_i^{AV} \in \mathbf{R}^{d\times d_v}$, 均为可学习的参数。

同理, 对于方面 $\vec{A}$和图像 $\hat{\boldsymbol{R}}$, 本文假设 $\vec{A}$作为 Query, $\hat{\boldsymbol{R}}$作为 Key 和 Value, 给定方面与图像进行交互, 获得方面限定的图像注意力情感表示 $\boldsymbol{M}^{\text{ar}}$为

$$\boldsymbol{M}^{\text{ar}} = \text{MultiHead}\left(\vec{A}, \hat{\boldsymbol{R}}, \hat{\boldsymbol{R}}\right) = \text{Concat}\left(H_1^A, H_2^A, \cdots, H_h^A\right)\boldsymbol{W}^{AR} \quad (5)$$

$$H_j^A = \text{Attention}\left(\vec{A}\boldsymbol{W}_j^{AQ}, \hat{\boldsymbol{R}}\boldsymbol{W}_j^{AK}, \hat{\boldsymbol{R}}\boldsymbol{W}_j^{AV}\right) \quad (6)$$

式中: $\boldsymbol{W}^{AR} \in \mathbf{R}^{h\times d_v\times d}$可学习的参数。

为更好地融合不同模态方面限定的注意力情感表示, 捕获图文数据不一致性的特性, 得到给定方面的图文注意力情感表示, 本文再次利用 DAIMA 机制, 将方面 $\vec{A}$、给定方面的图像和文本模态注意力情感表示 $\boldsymbol{M}^{\text{ar}}$、$\boldsymbol{M}^{\text{at}}$分别作为 Query、Key、Value, 去除与给定方面无关图文信息, 获得给定方面的多模态注意力情感表示 $\boldsymbol{M}^{\text{art}}$为

$$\boldsymbol{M}^{\text{art}} = \text{MultiHead}\left(\vec{A}, \boldsymbol{M}^{\text{ar}}, \boldsymbol{M}^{\text{at}}\right) = \text{Concat}\left(H_1^A, H_2^A, \cdots, H_h^A\right)\boldsymbol{W}^{ART} \quad (7)$$

$$H_j^A = \text{Attention}\left(\vec{A}\boldsymbol{W}_j^{AQ}, \boldsymbol{M}^{\text{ar}}\boldsymbol{W}_j^{AK}, \boldsymbol{M}^{\text{at}}\boldsymbol{W}_j^{AV}\right) \quad (8)$$

式中: $\boldsymbol{W}^{ART} \in \mathbf{R}^{h\times d_v\times d}$, $\boldsymbol{W}_j^{AQ} \in \mathbf{R}^{d\times d_k}$, $\boldsymbol{W}_j^{AK} \in \mathbf{R}^{d\times d_k}$, $\boldsymbol{W}_j^{AV} \in \mathbf{R}^{d\times d_v}$, 均为可学习的参数。

对于图文不一致性对给定方面图文注意力情感表示的影响, 本文通过将其平均池化获得方面直接交互的图文特征表示, 表达式为

$$\boldsymbol{M}_{\text{avg}}^{\text{art}} = \sum_{j=1}^{n} \boldsymbol{M}_j^{\text{art}}/n \quad (9)$$

3.3.2　方面深层交互机制的建立

由于文本数据和图像数据具有相关性, 为了更好的让图像和文本之间进行双向交互, 增强给定方面的情感信息, 建立融合 ADIMA 机制, 以捕获给定方面的文本和图像信息的相关性, 进一步提取准确的语义和情感信息。在此基础上, 利用 Transformer-encoder 具有的不同特征空间及全局范围的特征聚合能力特点, 捕获模态数据内部信息。

本文使用文本特征嵌入表示 $\hat{\boldsymbol{T}}$去查询图像模态的给定方面的注意力情感表示 $\boldsymbol{M}^{\text{ar}}$, 使方面信息深层参与图像对文本的交互, 具体来说, 本文把文本特征嵌入表示 $\hat{\boldsymbol{T}}$作为 Query, 给定方面的图像注意力情感表示 $\boldsymbol{M}^{\text{ar}}$作为 Key 和 Value。针对图文相关性建模, 获得与图像相关的给定方面文本注意力情感表示 $\boldsymbol{M}^{\text{tar}}$为

$$\boldsymbol{M}^{\text{tar}} = \text{MultiHead}\left(\hat{\boldsymbol{T}}, \boldsymbol{M}^{\text{ar}}, \boldsymbol{M}^{\text{ar}}\right) = \text{Concat}\left(H_1^T, H_2^T, \cdots, H_h^T\right)\boldsymbol{W}^{TAR} \quad (10)$$

$$H_j^T = \text{Attention}\left(\hat{\boldsymbol{T}}\boldsymbol{W}_j^{TQ}, \boldsymbol{M}^{\text{ar}}\boldsymbol{W}_j^{TK}, \boldsymbol{M}^{\text{ar}}\boldsymbol{W}_j^{TV}\right) \quad (11)$$

式中: $\boldsymbol{W}^{TAR} \in \mathbf{R}^{h\times d_v\times d}$, $\boldsymbol{W}_j^{TQ} \in \mathbf{R}^{d\times d_k}$, $\boldsymbol{W}_j^{TK} \in \mathbf{R}^{d\times d_k}$, $\boldsymbol{W}_j^{TV} \in \mathbf{R}^{d\times d_v}$, 均为可学习的参数。

本文将给定方面与图像相关的文本注意力情感表示 $\boldsymbol{M}^{\text{tar}}$, 通过 Transformer-encoder 捕获文本单模态内部信息表达式为

$$\boldsymbol{M}^{\text{Ttar}} = \text{Transform-encoder}(\boldsymbol{M}^{\text{tar}}) \quad (12)$$

对于得到的文本特征信息 $\boldsymbol{M}^{\text{Ttar}}$, 本文将其平均池化获得方面深层交互的文本模态特征表示为

$$\boldsymbol{M}_{\text{avg}}^{\text{Ttar}} = \sum_{j=1}^{n} \boldsymbol{M}_j^{\text{Ttar}}/n \quad (13)$$

类似地, 使用图像特征嵌入表示 $\hat{\boldsymbol{R}}$查询给定方面的文本注意力情感表示 $\boldsymbol{M}^{\text{at}}$, 使方面信息深层参与文本对图像的交互, 具体来说, 针对图文相关性建模, 将图像特征嵌入表示 $\hat{\boldsymbol{R}}$作为 Query, 给定方面的文本模态的注意力情感表示 $\boldsymbol{M}^{\text{at}}$作为 Key 和 Value, 获得与文本相关的给定方面图像注意力情感表示 $\boldsymbol{M}^{\text{rat}}$如下:

$$\boldsymbol{M}^{\text{rat}} = \text{MultiHead}\left(\boldsymbol{M}^{\text{r}}, \boldsymbol{M}^{\text{at}}, \boldsymbol{M}^{\text{at}}\right) = \text{Concat}\left(H_1^R, H_2^R, \cdots, H_h^R\right)\boldsymbol{W}^{RAT} \quad (14)$$

$$H_j^R = \text{Attention}\left(\hat{\boldsymbol{R}}\boldsymbol{W}_j^{RQ}, \boldsymbol{M}^{\text{ai}}\boldsymbol{W}_j^{RK}, \boldsymbol{M}^{\text{ai}}\boldsymbol{W}_j^{RV}\right) \quad (15)$$

式中: $\boldsymbol{W}^{TAI} \in \mathbf{R}^{h\times d_v\times d}$, $\boldsymbol{W}_j^{IQ} \in \mathbf{R}^{d\times d_k}$, $\boldsymbol{W}_j^{IK} \in \mathbf{R}^{d\times d_k}$, $\boldsymbol{W}_j^{IV} \in \mathbf{R}^{d\times d_v}$, 均为可学习的参数。

同样, 将给定方面与文本相关的图像注意力情感表示 $\boldsymbol{M}^{\text{rat}}$, 通过 Transformer-encoder 捕获图像单模态内部信息, 表达式为

$$\boldsymbol{M}^{\text{Ttar}} = \text{Transform-encoder}(\boldsymbol{M}^{\text{tar}}) \quad (16)$$

对于得到的图像特征信息 $\boldsymbol{M}^{\text{Trat}}$进行平均池化, 获得方面深层交互的图像模态特征表示为

$$\boldsymbol{M}_{\text{avg}}^{\text{Trat}} = \sum_{j=1}^{n} \boldsymbol{M}_j^{\text{Trat}}/n \quad (17)$$

3.4　情感分类器的设计

为丰富图文模态融合表示的情感语义信息, 通过将给定方面的多模态注意力情感表示 $\boldsymbol{M}_{\text{avg}}^{\text{art}}$、文本模态特征表示 $\boldsymbol{M}_{\text{avg}}^{\text{Ttar}}$和图像模态特征表示 $\boldsymbol{M}_{\text{avg}}^{\text{Trat}}$进行拼接融合, 建立给定方面的图文信息表示 $\boldsymbol{S}$ 为

$$\boldsymbol{S} = \left[\boldsymbol{M}_{\text{avg}}^{\text{art}}, \boldsymbol{M}_{avg}^{\text{Ttar}}, \boldsymbol{M}_{\text{avg}}^{\text{Trat}}\right] \quad (18)$$

将其作为全连接层输入, 预测给定方面的情感

标签，如下：

$$\boldsymbol{x} = W_{\text{multi}}\boldsymbol{S} + \boldsymbol{b}_{\text{multi}} \tag{19}$$

$$P = \text{softmax}(\boldsymbol{x}) \tag{20}$$

式中：W_{multi}为权重；$\boldsymbol{b}_{\text{multi}}$为偏置项。

本文使用交叉熵损失优化模型，并在优化过程中，使用 dropout[19] 机制防止过拟合。

4 实 验

4.1 数据集

为评估本文模型的有效性，本文采用给定方面类的多模态情感分类数据集 Multi-ZOL。该数据集收集整理了 IT 信息和商业门户网站 ZOL.com 上的关于手机的评论。本文仅选取了原始数据中包含多模态评论的数据，共计 5 288 条。每条多模态数据包含 1 个文本内容、1 个图像集，以及至少 1 个但不超过 6 个评价方面。这 6 个方面分别为性价比、性能配置、电池寿命、外观与感觉、拍摄效果、屏幕。对于每个方面，都有一个从 1 到 10 的情感得分。把数据集按照训练集、测试集、验证集 8∶1∶1 的比例进行划分。表 2 显示了此数据集的详细统计信息。

表 2 Multi-ZOL 数据集统计

Table 2 Statistics of Multi-ZOL dataset

属性	数值
评论数	5 228
标签数	10
每个评论的平均词数	315.11
每个评论的最大词数	8 511
每个评论的最小词数	5
每个评论的平均图像数	4.5
每个评论的最大图像数	111
每个评论的最小图像数	1

4.2 参数设置

对于图像，本文将其大小参数调整为 224×224 像素，将图像利用预训练的 ResNet-152 以获得 7×7×2 048 维视觉特征向量。对于本文模型，将学习率设置为 10^{-5}，batch 大小设置为 4，dropout 设置为 0.3，注意力头数设置为 2，优化器使用 Adam。为了获得一个泛化能力较强的模型，防止过度拟合，在训练过程中运用了早停法，即当验证集的准确率和 F_1 值连续 10 轮没有发生变化时，训练过程将停止。

4.3 对比方法

为了验证本文模型的有效性，设计了如下模型进行对比实验。

长短期记忆 (long short-term memory, LSTM) 模型[20]。该模型获取了文本序列的上下文信息。

基于注意机制的 LSTM(简称，AEAT-LSTM) 模型[21]。该模型在每个词嵌入中增加方面嵌入，并将单词隐藏状态与注意力层的方面嵌入相结合，用于关注给定方面相关联的关键词。

MemNet(memory network) 模型[22]。一种记忆网络模型，将方面嵌入视为查询向量，使用多个注意力机制在输入词嵌入堆叠的记忆上生成深度记忆。

交互注意力网络（interactive attention networks, IAN）模型[23]。该模型是基于 LSTM 的模型，在方面和上下文中进行交互注意，用于方面级别情感分类。

RAM(recurrent attention network on memory) 模型[24]。一种基于记忆的模型，利用 Bi-LSTM 生成方面表示，再利用多个注意力层的输出与循环神经网络进行非线性结合。

TomBERT(target-oriented multimodal BERT) 模型[4]。一个多模态 BERT 架构，利用 BERT 获得文本词嵌入，再利用跨模态注意力，对目标词进行图像注意力，获得目标-图像的注意力嵌入，形成叠加的多层自注意力。

Co-Memory+Aspect 模型[3]。该模型除对文本和视觉记忆之间进行交互注意建模之外，还引入了方面嵌入的平均值作为文本和视觉记忆网络的输入。

MIMN(multi-channel user interest memory network) 模型[2]。一种多交互记忆网络，用于捕获多模态数据中的多重相关性，并进行方面级情感分析。

EF-CapTrBERT(EF-caption transformers BERT) 模型[25]。该模型利用图像生成文本的方法，为每个图像构建一个辅助句，使语言模型融入了更多的多模态信息。

4.4 实验结果与分析

本文实验使用的评估指标是精确率和 F_1 值。利用 JAAIN 模型，在 Multi-ZOL 数据集中进行对比实验，结果如表 3 所示，最优结果为加粗显示。

由表 3 实验结果可知：

1）本文模型在方面类情感分析任务的准确率和 F_1 值均高于单文本模态的模型。ATAE-LSTM[21]、IAN[23] 都是基于 LSTM[20] 的模型，MemNet[22] 和 RAM[24] 都是基于记忆网络的模型。LSTM[20] 模型表现最差，该模型没有利用方面词信息，而其他模型通过注意力机制获取给定方面文本信息。相比于单文本模态模型，本文模型通过引入图像数据，可以多层次地融合方面信息与图文信息，从而保留了更多和更重要的方面类情感信息。

2）本文模型结果优于其他引入图像数据的模

表 3　对比实验结果

Table 3　Comparative experimental results

数据类型	模型	精确率/%	F_1/%
文本	LSTM[20]	58.92	57.29
	MemNet[22]	59.51	58.73
	ATAE-LSTM[21]	59.58	58.95
	IAN[23]	60.08	59.47
	RAM[24]	60.18	59.68
文本+图像	TomBERT[4]	59.35	58.40
	Co-Memory+Aspect[3]	60.43	59.74
	MIMN[2]	61.59	60.51
	EF-CapTrBERT[25]	65.67	64.99
	MIMN+BERT+CNN152	68.77	68.53
	JAAIN	**74.57**	**74.48**

型。TomBERT[4] 模型主要用于方面术语情感分析任务, 并将目标词换为方面词, 但仅利用方面词进行图像注意力, 结果提升并不明显, 而本文模型同时将图像信息、文本信息、方面信息三者融合。Co-Memory+Aspect[3] 模型和 MIMN[2] 模型均与 MemNet[22] 模型类似, 但 MIMN[2] 模型使用方面指导注意力机制用于模型生成文本和图像的注意力向量,效果优于 Co-Memory+Aspect[3]。EF-CapTrBERT[25] 模型是通过图像生成文本, 增加了语言模型可用的文本数量, 并在图像中提取了对象级信息, 而 EF-CapTrBERT[25] 模型在图片生成文本时, 容易丢失图片本身具有的方面信息。因此, 本文模型利用了文本与图像数据之间的不一致性和相关性, 使方面信息多层次地与图文信息进行交互, 比已有的模型更能充分获得图像和文本中的方面信息, 并利用 Transformer-encoder 进一步捕获模态内部信息。

3）本文将 MIMN[2] 模型也使用了 BERT 和 CNN152 预训练模型进行了实验, 获得识别结果也有显著提升, 进一步说明了预训练模型是有效的。

综上所述, 本文 JAAIN 模型在所有基线模型中均获得了最佳性能。

4.5　消融实验

为了验证模型各部分的性能, 设计了如下的消融实验:

-JAAIN(text)。表示去掉模型中的文本部分。

-JAAIN(image)。表示去掉模型中的图像部分。

-DAIMA。表示去掉通过 DAI Multi-Head Attention 机制获得的给定方面的多模态注意力情感表示。

-ADIMA。表示去掉通过 ADI Multi-Head Attention 机制获得的文本模态特征表示部分和图像模态特征表示部分。

-Transformer-encoder。表示去掉通过 Transformer-encoder 捕获单模态信息部分。

上述 3 种模型与本文模型在 Multi-ZOL 数据中的比较结果如表 4 所示, 最优结果加粗显示。

表 4　消融实验对比结果

Table 4　Comparison results of ablation experiments　%

模型	精确率	F_1
JAAIN	**74.57**	**74.48**
-JAAIN(image)	68.04	67.80
-JAAIN(text)	39.32	30.36
-DAIMA	72.21	72.20
-ADIMA	73.20	73.09
-Transformer-encoder	73.27	73.28

由表 4 实验结果可知:

1）-JAAIN(image) 和 -JAAIN(text) 模型的实验结果表明, 仅使用单模态的图像数据和文本数据, 不能有效解决图文数据之间存在的不一致性, 导致方面类情感分析精度降低。

2）-DAIMA 模型去除针对图文数据不一致性的建模部分, 使图文交互过程中丢失了图文数据间不一致性的捕获, 同时缺少了方面信息的直接交互, 导致引入了与方面信息无关的噪音。-ADIMA 模型去除利用图文数据相关性的建模部分, 缺少了深层次方面信息与图文信息间的交互, 导致给定方面的图文信息之间交互不充分。上述 2 种方法都降低了方面类情感分析的精度。

3）-Transformer-encoder 模型的实验结果, 说明了不使用 Transformer-encoder 模型, 降低了特征间的聚合能力, 无法捕获单模态的内部信息。

4.6　Transformer-encoder 模型的层数影响

本文模型多次利用 Transformer-encoder 模型, 层数也是影响模型性能的重要参数。在获得文本表示中, 利用 Transform-encoder 模型对通过 BERT 模型得到的文本特征进行文本内容聚合, 获取更深层次的表示。对这部分使用的 Transform-encoder 模型的层数进行了实验, 实验中设定其他的 Transformer-encoder 层数为 4。实验结果如图 2 所示。

本文模型在 4 层达到了最好的结果, 表明使用 4 层的 Transformer-encoder 模型能获取与图像表示比较接近的文本深层次表示。

在本文模型中, 通过 Transformer-encoder 模型捕获单模态的内部信息, 并对这部分使用的 Transform-encoder 模型的层数进行了实验, 实验中设定其他的 Transformer-encoder 模型层数为 4。实验结果如图 3 所示。

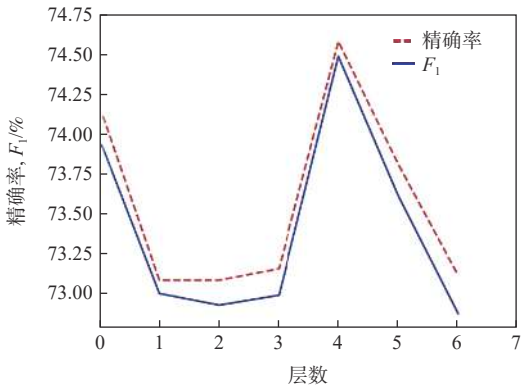


图2 文本表示中Transformer-encoder模型层数影响

Fig. 2 Influence of Transformer-encoder model layers in text representation

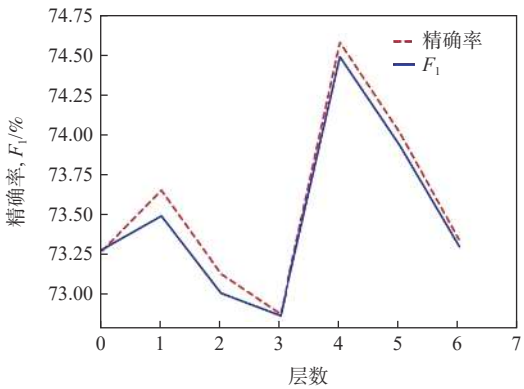


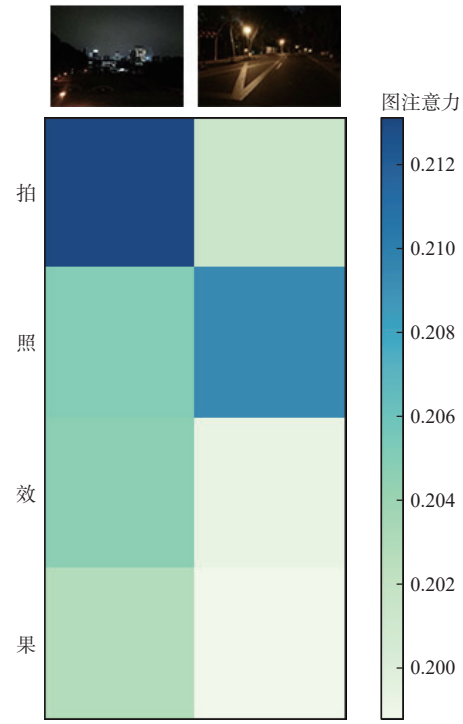
图3 捕获单模态的内部信息中Transformer-encoder层数影响

Fig. 3 Effect of number of Transformer-encoder model layers in capturing internal information of a single modality

由图3可知，此任务在4层时，Transformer-encoder模型的聚合能力达到最佳，充分捕获到模态内部信息。

4.7 可视化实验

为了进一步证明本文模型的有效性，分别对表1提出的3个实例进行了可视化实验。通过方面直接交互机制对实例1进行可视化，如图4所示。本文模型关注了图像在“拍照效果”方面的信息，忽略了文字中的无效信息，有效捕获图文数据不一致性的特性。通过方面深层交互机制对实例2和实例3进行可视化，如图5和图6所示，文本背景颜色越深表示注意力高。本文模型分别关注了文本和图像在“拍照效果”方面的信息，有效地捕获了给定方面的文本和图像信息的相关性。通过可视化实验，说明本文模型通过对图文数据的不一致性和相关性分别建模，可以有效地增强给定方面的图文模态数据的情感表示。



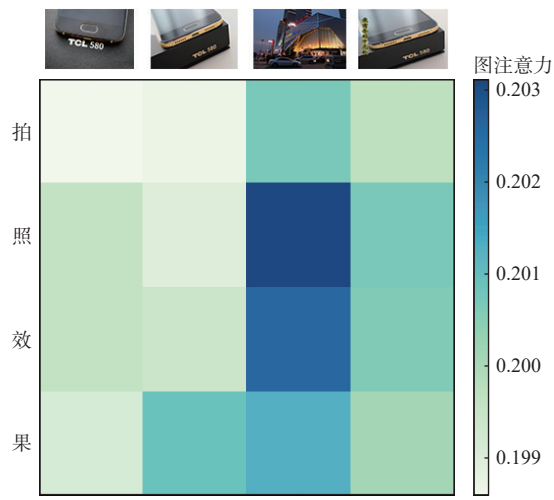
(a) 图像注意力可视化

手机挺好的，黑色挺酷，显示不错，还有全程的护眼模式，外观漂亮，运行流畅，就是输入法不太适应，电池一天一冲充肯定要的，就是在淘宝官方旗舰店买没有送耳机挺不开心的，心理不平衡！。电池容量小，就这个缺点吧。

(b) 文本注意力可视化

图4 方面直接交互机制对实例1可视化展示

Fig. 4 A visual display of aspect direct interaction mechanism for example 1



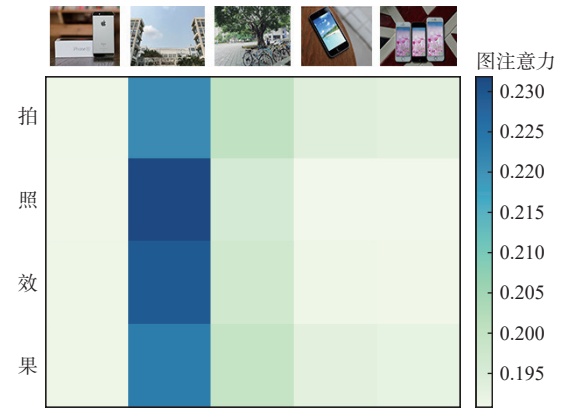
(a) 图像注意力可视化

1、外观低调、沉稳、内敛、商务感十足，深受男人喜爱；2、分屏、显示、色彩在这个价位上表现还是很不错的；3、提供摆拍姿势指导，支持前置柔光灯；4、手机续航不错，用一天妥妥的；5、支持多个知名品牌耳机音效；

(b) 文本注意力可视化

图5 方面深层交互机制对实例2可视化展示

Fig. 5 A visual display of aspect deep interaction mechanism for example 2



(a) 图像注意力可视化



(b) 文本注意力可视化

图 6　方面深层交互机制对实例 3 可视化展示

Fig. 6　A visual display of the aspect deep interaction mechanism for example 3

5　结　论

1）针对图文方面类情感分析任务，本文提出 JAAIN 模型，深入探索了方面与图文数据间的相互关系和影响，本文模型在图文交互过程中，通过多层次融和方面信息与图文信息，利用 Transformer-encoder 模型捕获了单模态内部信息，对给定方面的图文数据间的不一致性和相关性分别建模，增强给定方面的图文模态数据表示。

2）本文在多模态数据集 Multi-ZOL 上进行了实验，结果表明，本文模型性能优于其他比较模型，证明了本文模型可以有效地在图文交互过程中，分别对图文数据的不一致性和相关性建模，使方面信息和图文信息深度融合，并去除与给定方面无关的信息，从而得到更加准确的结果。

参考文献（References）

[1] PONTIKI M, GALANIS D, PAVLOPOULOS J, et al. SemEval-2014 task 4: Aspect based sentiment analysis[C]//Proceedings of the 8th International Workshop on Semantic Evaluation. Stroudsburg: Association for Computational Linguistics, 2014: 27-35.

[2] XU N, MAO W J, CHEN G D. Multi-interactive memory network for aspect based multimodal sentiment analysis[C]//Proceedings of the AAAI Conference on Artificial Intelligence. Washton, D.C.: AAAI, 2019, 33(1): 371-378.

[3] XU N, MAO W J, CHEN G D. A co-memory network for multimodal sentiment analysis[C]//Proceedings of the 41st International ACM SIGIR Conference on Research & Development in Information Retrieval. New York: ACM, 2018: 929-932.

[4] YU J F, JIANG J. Adapting BERT for target-oriented multimodal sentiment classification[C]//Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence. California: International Joint Conferences on Artificial Intelligence Organization, 2019: 5408-5414.

[5] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]//Proceedings of the 31st International Conference on Neural Information Processing Systems. New York: ACM, 2017: 6000–6010.

[6] GU S Q, ZHANG L P, HOU Y X, et al. A position-aware bidirectional attention network for aspect-level sentiment analysis [C]//Proceedings of the 27th International Conference on Computational Linguistics. Santa Fe: Curran Associates, Inc., 2018: 774–784.

[7] XU Q N, ZHU L, DAI T, et al. Aspect-based sentiment classification with multi-attention network[J]. Neurocomputing, 2020, 388: 135-143.

[8] 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.

[9] LI Y, ZENG J B, SHAN S G, et al. Occlusion aware facial expression recognition using CNN with attention mechanism[J]. IEEE Transactions on Image Processing, 2019, 28: 2439-2450.

[10] XIE S Y, HU H F, WU Y B. Deep multi-path convolutional neural network joint with salient region attention for facial expression recognition[J]. Pattern Recognition, 2019, 92: 177-191.

[11] ZHAO S C, GAO Y, JIANG X L, et al. Exploring principles-of-art features for image emotion recognition[C]//Proceedings of the 22nd ACM international conference on Multimedia. New York: ACM, 2014: 47-56.

[12] 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.

[13] PORIA S, CHATURVEDI I, CAMBRIA E, et al. Convolutional MKL based multimodal emotion recognition and sentiment analysis[C]//Proceedings of the 2016 IEEE 16th International Conference on Data Mining. Piscataway: IEEE Press, 2016: 439-448.

[14] CAO D L, JI R R, LIN D Z, et al. A cross-media public sentiment analysis system for microblog[J]. Multimedia Systems, 2016, 22(4): 479-486.

[15] TRUONG Q T, LAUW H W. VistaNet: Visual aspect attentionnetwork for multimodal sentiment analysis[C]//Proceedings of theAAAI Conference on Artificial Intelligence. Washton, D.C.: AAAI, 2019, 33(1): 305-312.

[16] XU J, HUANG F R, ZHANG X M, et al. Visual-textual sentiment classification with bi-directional multi-level attention networks[J]. Knowledge-Based Systems, 2019, 178: 61-73.

[17] LU J S, BATRA D, PARIKH D, et al. ViLBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks[EB/OL].(2019-08-06)[2022-01-03].https://arxiv.org/abs/1908.02265.

[18] DEVLIN J, CHANG M W, LEE K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding [C]//Proceedings of 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Minneapolis: Association for Computational Linguistics,

2019: 4171-4186.

[19] HINTON G E, SRIVASTAVA N, KRIZHEVSKY A, et al. Improving neural networks by preventing co-adaptation of feature detectors[J]. Computer Science, 2012, 3(4): 212-223.

[20] HOCHREITER S, SCHMIDHUBER J. Long short-term memory [J]. Neural Computation, 1997, 9(8): 1735-1780.

[21] WANG Y Q, HUANG M L, ZHU X Y, et al. Attention-based LSTM for aspect-level sentiment classification[C]//Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. Stroudsburg: Association for Computational Linguistics, 2016: 606-615.

[22] TANG D Y, QIN B, LIU T. Aspect level sentiment classification with deep memory network[C]//Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. Stroudsburg: Association for Computational Linguistics, 2016: 214-224.

[23] MA D H, LI S J, ZHANG X D, et al. Interactive attention networks for aspect-level sentiment classification[C]//Proceedings of the 26th International Joint Conference on Artificial Intelligence. New York: ACM, 2017: 4068-4074.

[24] CHEN P, SUN Z Q, BING L D, et al. Recurrent attention network on memory for aspect sentiment analysis[C]//Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. Stroudsburg: Association for Computational Linguistics, 2017: 452-461.

[25] 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: 3034-3042.

Image-text aspect emotion recognition based on joint aspect attention interaction

ZHAO Yicheng[1], WANG Suge[1, 2, *], LIAO Jian[1], HE Donghuan[1]

(1. School of Computer & Information Technology, Shanxi University, Taiyuan 030006, China;
2. Key Laboratory of Computational Intelligence and Chinese Information Processing of Ministry of Education, Shanxi University, Taiyuan 030006, China)

Abstract: Due to the quick development of social media, the sentiment conveyed by users cannot be reliably identified by an Aspect-Category Sentiment Analysis of the text alone. However, the existing Aspect-Category Sentiment Analysis methods for image and text data only consider the interaction between image and text modalities, ignoring the inconsistency and correlation of image and text data. Therefore, this paper proposes a joint aspect attention interaction network (JAAIN) model for aspect-category sentiment identification. The suggested technique improves the representation of image and text modalities in particular aspects by multi-level aspect, image, and text information fusion. It does this by removing the text and images that are unrelated to certain aspects. The text data sentiment representation, image data sentiment representation and aspect category sentiment representation are concatenated, fused and fully connected to realize sentiment discrimination of image and text aspects. The experimental results show that the proposed model can improve the performance of sentiment identification in images and text on the Multi-ZOL Dataset.

Keywords: aspect-category sentiment analysis; attention mechanism; multi-modal sentiment analysis; sentiment expression; multi-modal fusion

Received: 2022-05-19; **Accepted**: 2022-07-02; **Published Online**: 2022-10-18 10:42

URL: link.cnki.net/urlid/11.2625.V.20221017.1403.003

Foundation items: National Natural Science Foundation of China (62076158,61906112); Project of Science and Technology Bureau of Xiaodian District, Taiyuan City, Shanxi Province (2020XDCXY05)

* **Corresponding author**. E-mail: wsg@sxu.edu.cn

http://bhxb.buaa.edu.cn jbuaa@buaa.edu.cn
DOI: 10.13700/j.bh.1001-5965.2022.0388

基于多模态掩码 Transformer 网络的社会事件分类

陈宏[1]，钱胜胜[2]，李章明[2]，方全[2]，徐常胜[2,*]

(1. 郑州大学 河南先进技术研究院，郑州 450000；2. 中国科学院 自动化研究所，北京 100190)

摘　　要：多模态社会事件分类的关键是充分且准确地利用图像和文字2种模态的特征。然而，现有的大多数方法存在以下局限性：简单地将事件的图像特征和文本特征连接起来，不同模态之间存在不相关的上下文信息导致相互干扰。因此，仅仅考虑多模态数据模态间的关系是不够的，还要考虑模态之间不相关的上下文信息（即区域或单词）。为克服这些局限性，提出一种新颖的基于多模态掩码 Transformer 网络（MMTN）模型的社会事件分类方法。通过图-文编码网络来学习文本和图像的更好的表示。将获得的图像和文本表示输入多模态掩码 Transformer 网络来融合多模态信息，并通过计算多模态信息之间的相似性，对多模态信息的模态间的关系进行建模，掩盖模态之间的不相关上下文。在2个基准数据集上的大量实验表明：所提模型达到了最先进的性能。

关 键 词：多模态；社会事件分类；社交媒体；表示学习；多模态 Transformer 网络

中图分类号：TP391

文献标志码：A　　**文章编号**：1001-5965(2024)02-0579-09

随着社交媒体上各种社会事件的发布，人们会对其进行大量的转发和评论，这些包含用户意见和情感的内容可以反映许多政治、经济、安全、教育等问题。对于有需要的人和组织，社会事件分析能对他们提供非常大的帮助。比如当灾难发生时，相关应急机构和救援组织需要根据灾难的类别及严重程度来采取相应的措施。因此，从海量的社交媒体数据中自动分析其中的社会事件，挖掘出相关主题，对用户更好的浏览和搜索及政府的监管具有重要意义。

现有的各种用于社会事件分类的方法可以分为基于单模态方法[1-4]和基于多模态方法[5-7]。基于单模态方法是只使用图像或文本信息进行特征提取进而实现社会事件分类。例如，Kumar 等[1]提出了一个实时的推文跟踪系统，以帮助急救人员在灾难发生时获得态势感知。在早期，这类方法利用手工特征提取方法对社交事件进行建模，然而，手工制作的特征可能包含人的强烈地主观性，导致难以捕捉高度复杂的社会事件内容。随着深度学习的发展，一些方法开始用深度神经网络自动提取社会事件深层次的表示。尽管分类的精度得到了很大提升，但是现有方法大多只使用单一模态数据（图像或文本），导致社会事件分类受到了很大的限制。为此，一些研究者提出了基于多模态的方法，同时使用图像和文本信息。例如，Mouzannar 等[5]提出了一种多模态深度学习框架，该框架将图像级特征和文本级特征连接起来，将事件帖子分为5个类别，然而，其只是简单地将图像特征和文本特征连接起来。为了更好地利用图像和文本信息进行社交事件分类，Abavisani 等[7]提出了一种多模态融合方法，利用图像和文本作为输入，通过预训练模型从图像和文本中提取特征，并引入一个交叉注意力模块进行多模态信息融合，然而其只考虑了图像级和文本级特征，细粒度的图像和文本信息在很大

收稿日期：2022-05-19；录用日期：2022-11-04；网络出版时间：2023-01-12 11:57
网络出版地址：link.cnki.net/urlid/11.2625.V.20230111.1720.002
基金项目：国家自然科学基金 (61832002)
*通信作者. E-mail：csxu@nlpr.ia

引用格式：陈宏，钱胜胜，李章明，等. 基于多模态掩码 Transformer 网络的社会事件分类 [J]. 北京航空航天大学学报，2024，50（2）：579-587. CHEN H，QIAN S S，LI Z M，et al. Multi-modal mask Transformer network for social event classification [J]. Journal of Beijing University of Aeronautics and Astronautics，2024，50（2）：579-587（in Chinese）.

程度上被忽略了,但是这些细粒度的图像和文本信息非常重要并且包含了更丰富的信息。一般不同的模态之间存在着一种不平衡和互补的关系。但是不同的模态之间也存在着不相关的背景和相互干扰。所以在多模态融合过程中,不仅要考虑到多模态数据模态间的关系,还要考虑模态之间不相关的上下文信息。

因此,这些多模态方法虽然通过联合图像和文本特征的方式能提高社会事件分类的性能,但这些现有方法仍然无法有效解决以下2个挑战。

1)大多数现有方法只是将文本特征和图像特征简单拼接,忽略了不同模态之间的异质性及细粒度的特征,这可能会引起重要信息丢失并导致提取的特征不能很好地对事件进行表示。因此,如何捕捉文本和图像特征中的关系,充分利用细粒度的图像和文本信息并系统地融合文本和图像特征是社会事件分类的关键。

2)不同模态之间存在不相关的上下文信息导致相互干扰。因此,仅仅考虑多模态数据的模态间的关系是不够的,还要考虑模态之间不相关的上下文信息(即区域或单词)。

针对上述的局限性,本文提出了一种基于多模态掩码Transformer网络(multi-modal mask transformer network, MMTN)模型的社会事件分类方法,充分利用模态之间的相似性和差异性,对多模态信息的模态间关系进行建模,并掩盖掉模态之间不相关的上下文信息。1)针对挑战1,提出一种MMTN模型来更好的将图像和文本细粒度的信息融合成多模态事件表示。具体来说,首先,使用预训练的ResNet[8]模型来提取图像的细粒度特征表示,并使用BERT(bidirectional encoder representation transformers)[9]模型作为预训练语言模型来获取文本单词的表示。然后,将图像区域和文本单词的表示输入到多模态Transformer网络中。多模态Transformer网络利用Transformer编码器[10]可以聚合社会事件的细粒度图像和文本信息,捕获多模态数据模态间的关系,以相互补充和增强。2)针对挑战2,在网络模型中引入一个掩码模块,在多模态数据进行融合时,对2个模态之间的相似性及差异性进行建模,并利用其掩盖模态之间不相关的上下文信息,以加强相关上下文之间的相关性。在得到社会事件的统一表示之后,利用具有相应激活函数的全连接层产生的预测概率来确定危机事件的类别。

在2个公共基准数据集上评估了本文模型,实验结果表明,本文模型优于目前最好的模型。

总体而言,本文的工作有如下几个贡献点。

1)提出MMTN模型,其可以充分利用并融合图像和文本之间的全局特征及区域级和单词级的细粒度特征并捕捉到单词与区域之间的关系,更好的进行多模态特征融合。

2)该网络模型的掩码模块可以计算模态间的相关性,并掩盖模态之间不相关的信息,以帮助加强不同模态之间相关上下文之间的相关性。

3)在CrisisMMD[11]和PHEME[12]这2个公共基准数据集上的广泛实验结果表明,本文模型比目前最先进的社会事件分类模型表现更好。

1 相关工作

1.1 社会事件分类

随着互联网的快速发展,社交媒体上的社会事件数量大量增长,高效的浏览和监控社交事件变得具有挑战性。近些年研究人员一直致力于社会事件分析,并提出了许多不同的方法,一些方法利用人工提取特征。而人工提取特征通常需要大量的预处理和特征的手工制作工作,并且很容易受到制作人的情感影响,导致难以获取社会事件的深层表示。随着深度学习的发展,研究者们开始使用深度神经网络自动学习社会事件潜在的深层表征。但许多现有的方法都是基于单模态的。例如,Stowe等[3]提出了一个灾害相关的分类系统,来对相关推文进行识别,然后把其分类到所代表的更加细粒度的类别。To等[4]提出了一种基于匹配的方法来更好地识别与特定灾难类型(如地震、洪水)相关的推文。该方法首先通过匹配一些关键词,在哈希标签的集合中搜索候选哈希标签,同时从语料库中提取主题标签,然后,在候选哈希标签中进行筛选,最后,使用关键词和筛选后的主题标签来匹配相关推文。Li等[13]使用卷积神经网络来定位和量化灾难图像中的损害。Nalluru等[14]结合文本和图像特征对社交媒体帖子在紧急情况下的相关性进行分类。

虽然这些方法能够对社会事件进行有效的识别,但是在实际应用中,社会事件的信息都是以图像-文本对的形式出现的,仅仅依靠单一模态进行检测将会影响整体的检测质量。本文提出的MMTN模型能够捕获多模态数据的模态间关系,将图像和文本细粒度的信息融合成更完整的多模态事件表示。

1.2 多模态学习

随着多模态学习的发展,许多研究人员开始使用多模态学习方法来进行社会事件分类。例如,Mouzannar等[5]提出了一个多模态深度学习框架,

通过文本、图像和视频来识别社交媒体帖子上的灾害相关信息。Kelly 等[6]提出了一个系统, 用于从 Twitter 消息中的文本和图像内容中实时提取信息, 结合消息的时空元数据来过滤并可视化事件, 并通过分析发布的文本内容和图像来监控洪水事件。Abavisani 等[7]引入一个交叉注意力模块来对社会事件帖子的图像和文本级特征进行多模态融合, 进而对其分类。现有方法大多利用预训练模型对图片或文本进行特征提取, 并取最后一层的输出作为图片和文本特征的全局表示, 然后进行多模态特征融合。但是这样会丢失很多模态内特征的细节, 例如图片区域信息及单词级的信息。Nie 等[15]提出了一种端到端多尺度融合模型, 并将多尺度语义信息嵌入到哈希码中来进行跨模态检索任务。Wu 等[16]利用 3 种不同的子网络分别从文本、空间域和频域中提取特征, 然后通过叠加共同注意力层, 将这 3 个特征深度融合。Mao 等[17]提出一个多级跨模态融合网络模型, 在每个级别融合来自不同模态的层次上下文特征。但是并未考虑到模态间不相关的信息。Qi 等[18]提出通过实体增强的多模态融合模型并用于多模态谣言检测, 该模型利用从图片和文本中提取的实体计算模态间的一致性特征并与多模态特征拼接。

虽然大多数现有的多模态方法在社会事件分类任务上取得了良好的性能, 但是并没有充分的利用多模态数据细粒度的特征, 并且大部分方法在多模态融合的时候忽略掉了不同模态之间可能存在不相关上下文信息。本文提出基于 MMTN 的社会事件分类方法, 对不同模态之间的相似性和差异性进行建模, 并掩盖不同模态之间不相关的上下文信息。与 Qi 等[18]提出的方法相比, Qi 等[18]采用提取命名实体、BERT 和 VGG(Oxford Visual Geometry Group) 的方法来进行图片与文本的特征提取, 本文直接采用 BERT 和 Resnet 来进行区域细粒度特征及全局特征提取, 更好的保证了模型端到端。Qi 等[18]通过提取的文本实体和视觉实体信息进行相似度计算, 得到一致性特征并与其他特征拼接作为最终特征进而分类; 而本文方法通过掩码模块里的注意力机制, 计算文本与图片模态间的相似性, 并在多模态特征融合过程中直接剔除掉不相关的信息。

2　本文方法

2.1　任务描述

社会事件分类任务可以视为一个多分类问题, 任务目的是要识别贴子中反映的事件并将其按照所属事件类别进行分类。给定一组社会事件帖子 P, 将具有 M 个样本的训练集表示为

$$P=\{(p_1,y_1),(p_2,y_2),\cdots,(p_M,y_M)\}$$

式中: $p_m=(V_m,W_m)$ 表示第 m 个图-文对; $y_m\in\{1,2,\cdots,C\}$ 表示样本 p_m 的标签。C 为整个社会事件数据集类别的数量。本文的目标为训练一个网络模型, 根据帖子的图片 V 和文本 W 来对新的社会事件帖子进行分类。

2.2　整体框架

图 1 为本文方法的总体框架, 图中矢量均为不同特征的代称。引入了 MMTN, 通过充分利用图像和文本信息并掩盖模态之间不相关的上下文信息来提高危机事件分类任务的性能。整个框架由以下组件组成。

1) 图-文编码网络。对于给定的输入文本和图片, 分别使用 BERT 和 ResNet 来获取文本内容和视觉内容的细粒度嵌入表示。

2) MMTN。由于不同的模态信息可以相互补充, 提出了一种 MMTN, 用于学习融合图像区域和文本单词关系的有效方法。

3) 掩码网络。为了过滤掉图像区域和文本文字表示中的干扰, 引入一个掩码模块, 来掩盖在不同模态之间不相关的信息。

4) 危机事件检测层。利用具有相应激活函数的全连接层来检测危机事件的类别。

2.3　图-文编码网络

之前其他方法一般都是以预训练模型最后一层的输出作为特征表示, 其包含了文本或图片的全局信息, 但是却少了细致的局部信息, 比如每个单词的特征或图片每个区域的特征。为更好的获得社会事件的图像和文本特征表示, 不仅提取事件的全局特征, 也提取事件的局部信息。

2.3.1　文本特征提取器

对于文本内容, 采用基于 Transformer 的双向编码器表示 BERT 的最后一层的输出作为文本的全局特征, 此外, 为文本设置一个最大单词数, 用 BERT 提取长度为最大单词数的每个单词的特征作为文本单词级的细粒度特征。当某条文本单词数量超过了最大单词数, 截取并保留该文本前最大单词数长度的单词; 当某条文本单词数量小于最大单词数, 在编码时对其空余位置用 0 补齐。

给定一条文本内容 $\boldsymbol{W}$, 将 $\boldsymbol{W}$ 建模成一个单词序列 $\boldsymbol{W}=[w_1,w_2,\cdots,w_n]$, n 为该条文本的单词数。输入 BERT 网络得到转换后的特征表示为 $\boldsymbol{F}_{\boldsymbol{W}}=[\boldsymbol{t}_1,\boldsymbol{t}_2,\cdots,\boldsymbol{t}_n]$, 式中: $\boldsymbol{t}_i$ 对应于 w_i 转换后的特征。单

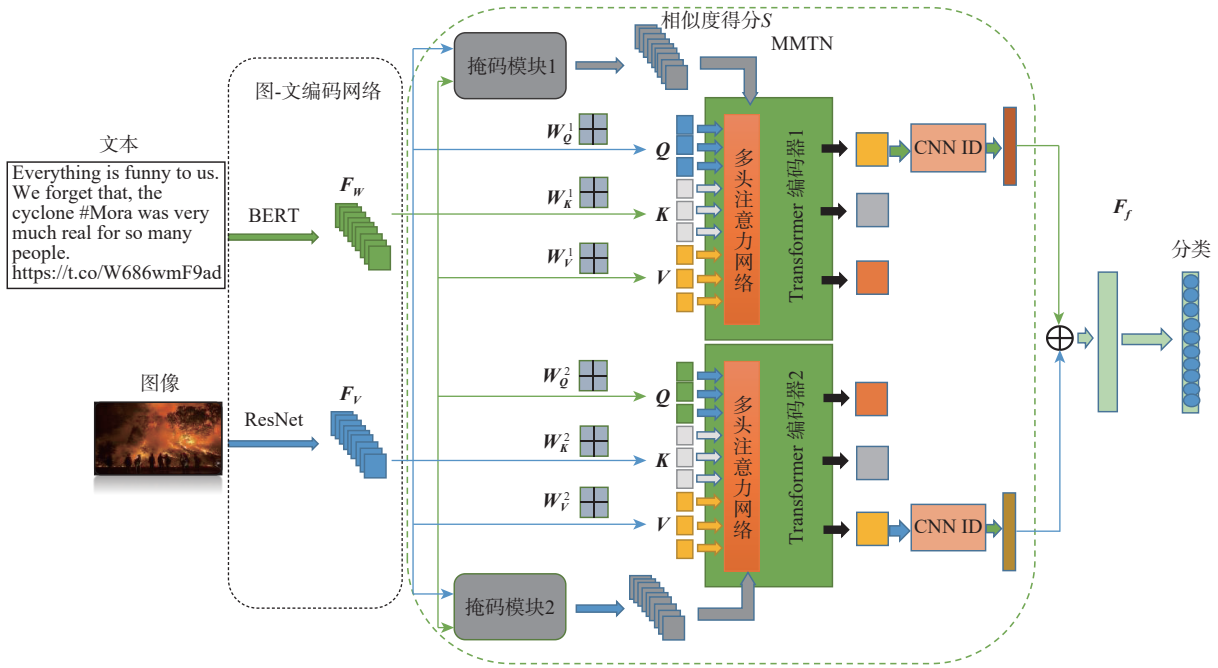


图 1 本文方法的总体框架

Fig. 1 Overall framework for the proposed method

词的表示t_i是由预训练模型 BERT 转换得到：

$$\boldsymbol{F}_W=\{\boldsymbol{t}_1,\boldsymbol{t}_2,\cdots,\boldsymbol{t}_n\}=\mathrm{BERT}(\boldsymbol{W}) \tag{1}$$

式中：$\boldsymbol{F}_W\in\mathbf{R}^{r\times d}$，$r$为文本中单词的个数，$d$为文本嵌入的维度。

2.3.2 视觉特征提取器

对于视觉内容V，使用自底向上的注意力预训练模型 ResNet 来提取图片的全局特征，并取倒数第 2 层隐藏层的输出作为区域级细粒度特征。即给定视觉内容V，视觉特征提取器的输出为

$$\boldsymbol{T}_V=[\boldsymbol{v}_1,\boldsymbol{v}_2,\cdots,\boldsymbol{v}_s]=\mathrm{RseNet}(V) \tag{2}$$

式中：$\boldsymbol{T}_V\in\mathbf{R}^{s\times 2\,048}$，$s$为图片区域的个数。

为了使其适应模型，采用全连接层调整$\boldsymbol{T}_V$的维度，即最后输入模型的视觉特征$\boldsymbol{F}_V\in\mathbf{R}^{s\times d}$。

2.4 多模态 Transformer 网络

为有效融合社会事件帖子的文本和视觉特征，提出了一个多模态 Transformer 网络来构建多模态上下文信息并从中提取高阶互补信息。

如图 1 所示，多模态 Transformer 网络由 2 个 Transformer 编码器[10]（编码器 1，编码器 2）组成。先将文本细粒度特征$\boldsymbol{F}_W$与 2 个参数矩阵$\boldsymbol{W}_K^1$和$\boldsymbol{W}_V^1$分别相乘得到的结果作为键$\boldsymbol{K}$-值$\boldsymbol{V}$对输入编码器 1 中，同时使用从 ResNet 模型获得的细粒度的图片区域特征$\boldsymbol{F}_V$与另一个参数矩阵$\boldsymbol{W}_Q^1$相乘得到的结果作为编码器 1 输入的查询$\boldsymbol{Q}$。于是可以得到以图片特征作为$\boldsymbol{Q}$的融合特征：

$$\boldsymbol{f}_{VW}=\mathrm{Encoder1}(\boldsymbol{F}_V,\boldsymbol{F}_W) \tag{3}$$

式中：$\boldsymbol{f}_{VW}\in\mathbf{R}^{s\times d}$；与$\boldsymbol{F}_V$的维度一致。

同理，将细粒度的图像区域特征$\boldsymbol{F}_V$与参数矩阵相乘结果作为$\boldsymbol{K}$和$\boldsymbol{V}$输入到编码器 2 中，同时使用细粒度的文本特征$\boldsymbol{F}_W$与另一个参数矩阵相乘结果作为编码器 2 输入的$\boldsymbol{Q}$。于是得到以文本特征作为$\boldsymbol{Q}$的融合特征：

$$\boldsymbol{f}_{WV}=\mathrm{Encoder2}(\boldsymbol{F}_W,\boldsymbol{F}_V) \tag{4}$$

式中：$\boldsymbol{f}_{WV}\in\mathbf{R}^{r\times d}$，与$\boldsymbol{F}_W$的维度一致。

一个区域信息会与文本所有单词级的细粒度特征进行共同注意力计算，同时，每个单词级特征也会与所有区域信息进行共同注意力计算，以捕捉单词与区域之间上下文的关系。

引入一维卷积对$\boldsymbol{f}_{VW}$和$\boldsymbol{f}_{WV}$进行局部信息聚合并降维。本文采用 3 种不同大小的卷积核（k=1,2,3）来获得 3 种尺度的特征，然后将其串联起来送入全连接层获得最终的融合特征，即

$$\boldsymbol{F}_{VW}=\mathrm{CNN1D}(\boldsymbol{f}_{VW}) \tag{5}$$

$$\boldsymbol{F}_{WV}=\mathrm{CNN1D}(\boldsymbol{f}_{WV}) \tag{6}$$

式中：$\boldsymbol{F}_{WV}\in\mathbf{R}^{1\times d}$；$\boldsymbol{F}_{VW}\in\mathbf{R}^{1\times d}$。

2.5 掩码网络

为了捕捉不同模态之间的关系，掩盖模态之间不相关的背景，掩码网络示意图如图 2 所示，在进行多模态特征融合的网络中引入一个掩码模块，通过计算不同模态之间的相关性，筛选出更相关的上下文信息。

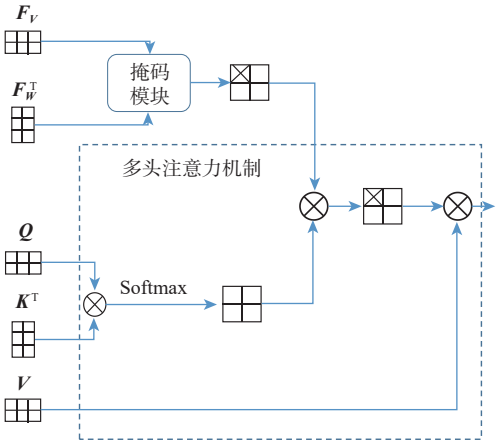


图 2　掩码网络示意图

Fig. 2　Schematic diagram of mask network

自注意力机制[10]是 MMTN 的核心。单头自注意力机制可以定义为

$$\text{Attention}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=\text{Softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{d_k}}\right)\boldsymbol{V} \tag{7}$$

虽然通过自注意力机制所有上下文的信息都能够得到充分的交互, 但是其中不相关的上下文会影响交互结果。于是, 为消除这些影响, 对自注意力机制进行改进。以第 1 个掩码模块为例, 首先, 对$\boldsymbol{F}_W$和$\boldsymbol{F}_V$计算相似度得分:

$$S(\boldsymbol{F}_V,\boldsymbol{F}_W)=\text{Softmax}\left(\frac{\boldsymbol{F}_V\boldsymbol{F}_W^{\mathrm{T}}}{\sqrt{d}}\right) \tag{8}$$

然后, 根据相似度得到掩码分数:

$$S_{i,j}=\begin{cases}0 & S(\boldsymbol{F}_V,\boldsymbol{F}_W)_{i,j}<l\\ 1 & S(\boldsymbol{F}_V,\boldsymbol{F}_W)_{i,j}\geqslant l\end{cases} \tag{9}$$

式中: $S(\boldsymbol{F}_V,\boldsymbol{F}_W)_{i,j}$表示$\boldsymbol{F}_W$中的第$i$个单词与$\boldsymbol{F}_V$中第$j$个区域的相似度得分; l为相似度阈值; 若相似度得分低于这个阈值, 该处的掩码分数置为 0, 否则为 1。

当获取到 2 个模态细粒度特征之间的掩码分数之后, 将其与多模态 Transformer 网络中的多头注意力网络结合。以编码器 1 为例, 首先, 由$\boldsymbol{F}_W$和$\boldsymbol{F}_V$得到输入的键、值和查询, 然后, 将掩码模块加入到自注意力机制的过程, 结合后的单头多模态自注意力机制定义为

$$\text{MAttention}=S(\boldsymbol{F}_V,\boldsymbol{F}_W)\cdot \text{Softmax}\left(\frac{\boldsymbol{W}_Q^1\boldsymbol{F}_V(\boldsymbol{W}_K^1\boldsymbol{F}_W)^{\mathrm{T}}}{\sqrt{d}}\right)\boldsymbol{W}_V^1\boldsymbol{F}_W \tag{10}$$

2.6　危机事件检测层

在得到掩码 Transformer 网络输出的 2 个多模态融合特征表示$\boldsymbol{F}_{VW}$和$\boldsymbol{F}_{WV}$后, 将其连接在一起得到最终的特征表示:

$$\boldsymbol{F}_f=\text{Concat}(\boldsymbol{F}_{VW},\boldsymbol{F}_{WV}) \tag{11}$$

使用带有 Softmax 激活函数的全连接层和 dropout 机制, 将多模态特征向量$\boldsymbol{F}_f$投影到事件类别的目标空间中:

$$\hat{\boldsymbol{Y}}=\text{Softmax}(\boldsymbol{W}_b\boldsymbol{F}_f+\boldsymbol{b}) \tag{12}$$

式中: $\boldsymbol{W}_b$为权重矩阵; $\boldsymbol{b}$为偏置项。

使用交叉熵损失作为分类损失:

$$\mathcal{L}(\boldsymbol{Y},\hat{\boldsymbol{Y}})=-\frac{1}{M}\sum_{i=1}^{M}\sum_{c=1}^{C}y_{i,c}\ln\hat{y}_{i,c} \tag{13}$$

式中: $\boldsymbol{Y}$为危机事件的真是类别标签; M为危机事件帖子的总数; C为危机事件的类别数; $\hat{\boldsymbol{Y}}$为模型预测的类别; $y_{i,c}$为测试集样本的真实标签; $\hat{y}_{i,c}$为测试集样本输出的预测标签。

3　实　验

3.1　实验准备

3.1.1　数据集

使用 CrisisMMD[11]和 PHEME[12]这 2 个数据集来验证本文模型的有效性。表 1 和表 2 分别为用于实验的 2 个数据集每个类别及其图片和文本的数量信息, 其中 CrisisMMD[11]数据集包含了 2017 年在全球发生的 7 个危机事件, 但是只采用了部分 PHEME[12]数据集。因为原始 PHEME[12]数据集所包含的 9 类事件中有 4 类样本量很小并且很多事件的图像信息不全, 所以将这 4 个类别的事件删除掉, 只保留剩下的 5 类事件。针对缺失图片的问题, 在本文中, 采用全白图片来作为该条文本对应

表 1　CrisisMMD[11]数据集

Table 1　CrisisMMD[11] dataset

事件	文本数量/条	图片/张
Hurricane Irma	4 041	4 525
Hurricane Harvey	4 000	4 443
Hurricane Maria	4 000	4 562
Mexico earthquake	1 239	1 382
California wildfires	1 486	1 589
Sri Lanka floods	832	1 025
Iraq-Iran earthquake	499	600
所有事件	16 097	18 126

表 2　PHEME[12]数据集

Table 2　PHEME[12] dataset

事件	文本数量/条	图片/张
Charlie Hebdo	2 079	911
Sydney siege	1 221	413
Ferguson	1 143	355
Ottawa shooting	890	231
Germanwings-crash	469	179
所有事件	5 802	2 089

的图片。鉴于用于本文实验的PHEME[12]数据集数量较小，同时受Abavisani等[7]的启发，采取与其类似的方法对PHEME[12]数据集进行数据增强。在实验中，训练数据占数据集的70%，其余用于测试。

3.1.2 评价指标

社会事件分类任务通常使用准确率(accuracy)作为评估指标。但是这2个数据集每个类别的样本数不平衡，只使用准确率作为指标会导致结果可靠性降低。因此，添加了宏F_1分数(Macro F_1-score)和加权F_1分数(Weighted F_1-score)2个评估指标。

3.1.3 参数设置

对于社会事件帖子的多模态嵌入，用在ImageNet[19]上进行预训练得到的Resnet来提取图像特征，图像嵌入的维度为2 048。利用预训练的BERT来提取文本特征，文本嵌入的维度为768。为了适应本文模型，通过一个全连接层，将图像嵌入的维度由2 048转换为768。本文模型使用自适应矩估计(Adam)优化器进行训练，一共训练了100个epoch，学习率为0.000 2，小批量设置为64。在CrisisMMD[11]和PHEME[12]这2个数据集上，掩码模块的阈值l分别设置为0.1和0.3。

3.2 比较模型

选取了7种最先进的模型进行比较。

1) 有监督的模态双变换器(简称MMBT)模型[20]。一种用于对图像和文本进行分类的有监督的多模态双变换器模型。

2) 紧凑型双线性池化(compact bilinear pooling, CBP)模型[21]。在视觉问答任务中提出的模型。也可以将其用于的社会事件分类任务。

3) 紧凑型双线性门控池化(compact bilinear gated pooling, CBGP)模型[22]是CBP[21]的扩展，在紧凑型双线性池化模块中加入了一个额外的感知门。

4) 用于灾害响应的多模态深度学习(multi-modal deep learning for disaster response, MDL-DR)模型[23]。使用VGG-16[24]提取图像特征并利用基于CNN的架构来获取文本特征，然后将这些特征连接起来执行分类任务。

5) Multi-RC模型[25]。使用门控循环单元(gated recurrent unit, GRU)[26]和CNN[27]来提取文本和图像特征。然后使用平均和连接的方法进行不同的模态融合。

6) SCBD(SSE-cross BERT denseNet)模型[7]。使用DenseNet[28]和BERT来提取图像和文本特征，并且应用交叉注意机制来进行不同模态特征的融合。

7) 多模态共同注意力网络(multimodal co-attention networks, MCAN)模型[16]。使用VGG-19和BERT提取图像和文本特征，同时使用离散余弦变换与CNN结合的方法提取图片的频域特征，再通过叠加共同注意力层，将这3个特征深度融合。

3.3 实验结果分析

3.3.1 定量分析

实验结果如表3和表4所示，通过观察表中的数据可以得到以下结果。

1) 有些模型虽然采用多模态的方法，但是效果并不是很理想，比如MDL-DR[23]和Multi-RC[25]模型。可以看出使用CNN、VGG和GRU来提取单一模态特征的方法不如使用BERT和Resnet的效果好。

表3 不同模型在CrisisMMD数据集上的实验结果

Table 3 Experimental results of different models on CrisisMMD dataset %

模型	准确率	宏F_1分数	加权F_1分数
MMBT[20]	87.82	84.78	87.62
CBP[21]	91.54	89.36	91.23
CBGP[22]	92.12	91.46	92.45
MDL-DR[23]	86.77	85.73	86.91
Multi-RC[25]	83.95	82.23	83.95
SCBD[7]	93.66	95.10	93.66
MCAN[16]	94.82	94.87	94.85
MMTN	96.72	97.09	96.72

表4 不同模型在PHEME数据集上的实验结果

Table 4 Experimental results of different models on PHEME dataset %

模型	准确率	宏F_1分数	加权F_1分数
MMBT[20]	85.92	84.27	88.66
CBP[21]	89.42	88.52	90.33
CBGP[22]	86.14	87.79	90.12
MDL-DR[23]	85.3	85.1	85.4
Multi-RC[25]	83.67	82.61	83.17
SCBD[7]	90.90	91.26	90.91
MCAN[16]	91.19	91.55	91.21
MMTN	94.65	94.89	94.65

2) 一些模型(例如，CBP[21]、CBGP[22]、SCBD[7]、MCAN[16])在2个数据集上表现良好，这是因为其使用DenseNet和BERT等预训练网络获取到了比较好的图像和文本特征。并且相比其余直接将2个模态特征拼接的模型，SCBD[7]和MCAN[16]模型引入了交叉注意力网络来进行不同模态特征的融合，效果更好。

3) 与所有模型相比，本文模型在2个数据集上实现了最佳性能并优于其他事件分类模型。本文模型的优越性归因于其3个特性：①本文模型使用

细粒度的图像信息和细粒度的文本信息来检测事件。②提出的多模态 Transformer 网络可以更有效地融合不同模态的信息。③提出的掩码网络可以有效的过滤掉多模态信息融合过程中不相关的信息。

3.3.2　消去实验

为了验证本文模型的有效性, 比较了 MMTN 模型的几个变式。

1) MMTN¬$\boldsymbol{F}_{VW}$表示去掉去掉 MMTN 模型中的编码器 1。

2) MMTN¬$\boldsymbol{F}_{WV}$表示去掉 MMTN 模型中的编码器 2。

3) MMTN¬M表示去掉掩码网络。

4) MMTN¬FI表示去掉细粒度特征, 采用预训练模型最后一层的输出作为图片和文本特征。

表 5 和表 6 为 MMTN 模型的不同变式在 2 个数据集上的实验结果。

表 5　MMTN 模型的不同变式在 CrisisMMD[11] 数据集上的检测性能比较

Table 5　Comparison of detection performance of different variants of MMTN model on CrisisMMD[11] dataset　%

模型	准确率	宏F_1分数	加权F_1分数
MMTN¬$\boldsymbol{F}_{VW}$	94.82	95.63	94.84
MMTN¬$\boldsymbol{F}_{WV}$	93.86	93.63	93.85
MMTN¬M	95.72	96.35	95.72
MMTN¬FI	94.76	94.75	94.77
MMTN	96.72	97.09	96.72

表 6　MMTN 模型的不同变式在 PHEME[12] 数据集上的检测性能比较

Table 6　Comparison of detection performance of different variants of MMTN model on PHEME[12] dataset　%

模型	准确率	宏F_1分数	加权F_1分数
MMTN¬$\boldsymbol{F}_{VW}$	91.31	91.61	91.29
MMTN¬$\boldsymbol{F}_{WV}$	91.71	92.42	91.71
MMTN¬M	93.44	93.65	93.44
MMTN¬FI	92.11	92.78	92.17
MMTN	94.65	94.89	94.65

由表 5 和表 6 可知:

1) MMTN-$\boldsymbol{F}_{VW}$和MMTN-$\boldsymbol{F}_{WV}$模型不仅有所差别而且都比完整的模型效果要差, 说明在 MMTN 模型中, 以文本特征为 $\boldsymbol{Q}$ 的编码器 2 和以图象特征为 $\boldsymbol{Q}$ 的编码器 1 学习到了不同的多模态之间的关系, 两者结合可以达到互补的效果。

2) MMTN¬FI模型比完整的模型效果要差很多, 说明细粒度的图像和文本信息非常重要并且包含了更丰富的信息。

3) 去掉掩码网络的MMTN-M模型检测精度相比于MMTN-$\boldsymbol{F}_{WV}$和MMTN-$\boldsymbol{F}_{VW}$模型虽然有所上升, 但相比于完整的模型还是低了一些, 这表明了掩码网络对模型的有效性, 也证明了本文模型能有效地融合图像区域和文本单词关系。

3.3.3　掩码模块阈值敏感性实验

为了分析掩码模块的阈值对实验结果的影响并验证掩码模块的有效性, 比较了不同的阈值在 2 个数据集上的准确率, 实验结果如图 3 所示。

由图可知, 准确率先随着阈值的增加而不断上升, 这证明了掩码模块的有效性。但当达到一定程度时, 又随着阈值的增加而下降, 说明当阈值取得过高时, 也会剔除掉一些有用的信息。

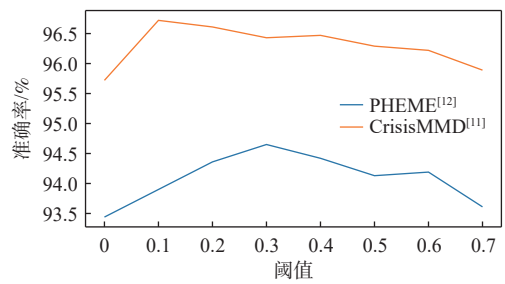


图 3　掩码模块不同阈值实验结果比较

Fig. 3　Comparison of experimental results with different thresholds of mask module

4　结　论

1)使用全局特征及区域级和单词级的细粒度特征来更好地对社会事件图像和文本进行表示, 并使用多模态 Transformer 网络来融合多模态细粒度特征信息, 其可以捕获不同模态之间的关系, 更好的进行多模态特征融合。

2)引入一种掩码网络, 通过计算不同模态之间上下文信息之间的相关性, 在进行多模态特征融合时, 过滤掉模态之间无关的上下文信息。通过在 2 个基准数据集上的实验表明本文模型更加有效。

参考文献 (References)

[1] KUMAR S, BARBIER G, ABBASI M, et al. TweetTracker: An analysis tool for humanitarian and disaster relief[C]//Proceedings of the International AAAI Conference on Web and Social Media. Washington, D. C. : AAAI, 2021, 5(1): 661-662.

[2] SHEKHAR H, SETTY S. Disaster analysis through tweets[C]//Proceedings of the 2015 International Conference on Advances in Computing, Communications and Informatics. Piscataway: IEEE Press, 2015: 1719-1723.

[3] STOWE K, PAUL M J, PALMER M, et al. Identifying and categor-

izing disaster-related tweets[C]//Proceedings of the Fourth International Workshop on Natural Language Processing for Social Media. Stroudsburg: Association for Computational Linguistics, 2016.

[4] TO H, AGRAWAL S, KIM S H, et al. On identifying disaster-related tweets: Matching-based or learning-based? [C]//Proceedings of the 2017 IEEE Third International Conference on Multimedia Big Data. Piscataway: IEEE Press, 2017: 330-337.

[5] MOUZANNAR H, RIZK Y, AWAD M. Damage identification in social media posts using multimodal deep learning[C]//Proceedings of the 15th International Conference on Information Systems for Crisis Response and Management. Rochester: ISCRAM, 2018.

[6] KELLY S, ZHANG X B, AHMAD K. Mining multimodal information on social media for increased situational awareness[C]//Proceedings of the 14th International Conference on Information Systems for Crisis Response and Management. Albi: ISCRAM, 2017.

[7] ABAVISANI M, WU L W, HU S L, et al. Multimodal categorization of crisis events in social media[C]//Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2020: 14667-14677.

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

[9] DEVLIN J, CHANG M W, LEE K, et al. BERT: Pre-training of deep bidirectional Transformers for language understanding [EB/OL]. (2019-05-24) [2022-02-16]. https://arxiv.org/abs/1810.04805.

[10] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]//Proceedings of the 31st International Conference on Neural Information Processing Systems. New York: ACM, 2017: 6000-6010.

[11] ALAM F, OFLI F, IMRAN M. CrisisMMD: Multimodal twitter datasets from natural disasters[C]//Proceedings of the International AAAI Conference on Web and Social Media. Washington, D. C. : AAAI, 2018, 12(1): 465-473.

[12] KOCHKINA E, LIAKATA M, ZUBIAGA A. All-in-one: Multi-task learning for rumour verification[C]//Proceedings of the 27th International Conference on Computational Linguistics. Santa Fe: Association for Computational Linguistics, 2018: 3402-3413.

[13] LI X K, CARAGEA D, ZHANG H Y, et al. Localizing and quantifying damage in social media images[C]//Proceedings of the 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining. Piscataway: IEEE Press, 2018: 194-201.

[14] NALLURU G, PANDEY R, PUROHIT H. Relevancy classification of multimodal social media streams for emergency services [C]//Proceedings of the 2019 IEEE International Conference on Smart Computing. Piscataway: IEEE Press, 2019: 121-125.

[15] NIE X S, WANG B W, LI J J, et al. Deep multiscale fusion hashing for cross-modal retrieval[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2021, 31(1): 401-410.

[16] WU Y, ZHAN P W, ZHANG Y J, et al. Multimodal fusion with co-attention networks for fake news detection[C]//Proceedings of the Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021. Stroudsburg: Association for Computational Linguistics, 2021: 2560-2569.

[17] MAO Y D, JIANG Q P, CONG R M, et al. Cross-modality fusion and progressive integration network for saliency prediction on stereoscopic 3D images[J]. IEEE Transactions on Multimedia, 2021, 24: 2435-2448.

[18] QI P, CAO J, LI X R, et al. Improving fake news detection by using an entity-enhanced framework to fuse diverse multimodal clues[C]//Proceedings of the 29th ACM International Conference on Multimedia. New York: ACM, 2021: 1212-1220.

[19] DENG J, DONG W, SOCHER R, et al. ImageNet: A large-scale hierarchical image database[C]//Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2009: 248-255.

[20] KIELA D, BHOOSHAN S, FIROOZ H, et al. Supervised multimodal bitransformers for classifying images and text[EB/OL]. (2019-09-06) [2022-02-18]. https://arxiv.org/abs/1909.02950.

[21] FUKUI A, PARK D H, YANG D, et al. Multimodal compact bilinear pooling for visual question answering and visual grounding[C]//Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. Stroudsburg: Association for Computational Linguistics, 2016: 457-468.

[22] KIELA D, GRAVE E, JOULIN A, et al. Efficient large-scale multimodal classification[EB/OL]. (2008-02-06) [2022-02-18]. https://arxiv.org/abs/1802.02892.

[23] OFLI F, ALAM F, IMRAN M. Analysis of social media data using multimodal deep learning for disaster response[EB/OL]. (2020-04-14)[2022-02-18] .https://arxiv.org/abs/2004.11838.

[24] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale visual recognition[EB/OL]. (2015-04-10) [2022-02-20]. https://arxiv.org/abs/1409.1556.

[25] LI X, CARAGEA D. Improving disaster-related tweet classification with a multimodal approach[C]//Proceedings of the 17th ISCRAM Conference. Blacksburg: ISCRAM, 2020: 893-902.

[26] CHO K, VAN MERRIENBOER B, GULCEHRE C, et al. Learning phrase representations using RNN Encoder–Decoder for statistical machine translation[C]//Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing. Stroudsburg: Association for Computational Linguistics, 2014: 1724-1734.

[27] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. ImageNet classification with deep convolutional neural networks[J]. Communications of the ACM, 2017, 60(6): 84-90.

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

Multi-modal mask Transformer network for social event classification

CHEN Hong[1]，QIAN Shengsheng[2]，LI Zhangming[2]，FANG Quan[2]，XU Changsheng[2,*]

(1. Henan Institute of Advanced Technology，Zhengzhou University，Zhengzhou 450000，China；
2. Institute of Automation，Chinese Academy of Sciences，Beijing 100190，China)

Abstract：Utilizing both the properties of the text and image modalities to the fullest extent possible is essential for multi-modal social event classification. However,most of the existing methods have the following limitations: They simply concatenate the image features and textual features of events. The existence of irrelevant contextual information between different modalities leads to mutual interference. Therefore,it is not enough to only consider the relationship between modalities of multimodal data,but also consider irrelevant contextual information between modalities (such as regions or words). To overcome these limitations,this paper proposes a novel social event classification method based on multimodal mask transformer network (MMTN) model. Specifically,the authors learn better representations of text and images through an image-text encoding network. To combine multimodal data,the resultant picture and word representations are input into a multimodal mask Transformer network. By calculating the similarity between the multimodal information,the relationship between the modalities of the multimodal information is modeled,and the irrelevant contexts between the modalities are masked. Extensive experiments on two benchmark datasets demonstrate that the proposed model achieves the state-of-the-art performance.

Keywords：multi-modal；crisis event classification；social media；representation learning；multi-model Transformer network

Received：2022-05-19；**Accepted**：2022-11-04；**Published Online**：2023-01-12 11：57
URL：link.cnki.net/urlid/11.2625.V.20230111.1720.002
Foundation item：National Natural Science Foundation of China (61832002)
* **Corresponding author**. E-mail：csxu@nlpr.ia

http://bhxb.buaa.edu.cn jbuaa@buaa.edu.cn
DOI: 10.13700/j.bh.1001-5965.2022.0394

区域感知实时人像超分辨率重建网络

龚柯存[1,2]，周梦琳[3]，唐东明[1,2,*]

(1. 西南民族大学 国家民族事务委员会计算机系统重点实验室，成都 610041；
2. 西南民族大学 计算机科学与工程学院，成都 610041；3. 成都信息工程大学 软件工程学院，成都 610225)

摘 要：在人像超分辨率重建领域，传统方法通常将整幅图像进行统一处理，导致效率低下。为降低模型的推理时延，提出了一种实时超分辨率重建模型RASR。该模型利用门控单元处理低分辨率图像，识别出人像边缘区域；采用分区重建策略，使用不同尺寸的子模型分别针对包含或不包含人像边缘的区域进行重建。实验结果表明：与现有方法相比，RASR模型在4倍上采样重建场景下的推理时延降低了88%，能够更有效地重建高分辨率人像图像。

关 键 词：区域感知；单图像超分辨率；门控单元；通道划分块；深度学习
中图分类号：TP391.4
文献标志码：A **文章编号**：1001-5965(2024)02-0588-08

作为一项经典的计算机视觉任务，单图像超分辨率 (single-image super-resolution, SISR) 是一种利用低分辨率图像恢复或重建至高分辨率图像的技术，本质上是通过机器学习算法学习低分辨率图像到高分辨率图像的特征映射，从而重建出相应的高分辨率图像。SISR可以在许多实际应用中实现，如监控系统、遥感、医学图像增强，还可作为图像识别和分类任务的先前工作。

为解决SISR问题，已有多种超分辨率 (super resolution, SR) 技术被提出，可以分为2个方向：基于插值的方法和基于学习的方法。三次卷积插值[1]为经典的基于插值的方法。近年来，随着备受关注的深度卷积神经网络 (convolution neural network, CNN) 的发展，基于深度学习的超分辨率重建方法以其优异的性能和实时性受到了广泛关注[2-4]。Dong等[5]设计了超分辨率重建卷积神经网络 (super-resolution convolution neural network, SRCNN)，其结合了双线性插值法，获得了当时最优的重建效果。Lim等[6]改进了ResNet[7]的结构，并设计了增强型深度超分辨率网络 (enhanced deep super-resolution network, EDSR)，显著提高了重建性能。Tong等[8]设计了SRDenseNet，使用块之间的密集跳跃连接，缓解了梯度消失问题，并加快了训练过程。更深的网络通常需要更多时间进行推理，为加速深层网络的推理速度，科研人员提出了多种方法。Szegedy等[9]指出，使用1×1的卷积核改变通道数，再使用3×3卷积核进行特征提取，能以微小的精度损失代价减少大量的计算，并提出了InceptionNet。Han等[10]结合特征图冗余度提出了GhostNet，通过引入Ghost模块，使用更少的参数提取图像特征，加速了模型的推理。在通道之间关系的处理上，科研人员也有相应的研究。Hu等[11]提出了SE(squeeze-and-excitation)模块，该模块为不同信道赋予不同的权重，增强了网络的表示能力。Wang等[12]引入非局部块，通过计算特征图2个位置之间的相关性来直接捕获远程依赖关系。

然而，以上轻量级模型没有针对特定场景做出优化，即对于人像超分辨率重建场景，存在以下缺

收稿日期：2022-05-20；**录用日期**：2022-07-02；**网络出版时间**：2022-09-20 11:55
网络出版地址：link.cnki.net/urlid/11.2625.V.20220919.1238.001
基金项目：国家自然科学基金 (61873217)；四川省科技计划 (2021JDRC0063)；西南民族大学中央高校基本科研业务费专项资金 (校2021118)
*通信作者. E-mail：tdm_2010@swun.edu.cn

引用格式：龚柯存，周梦琳，唐东明. 区域感知实时人像超分辨率重建网络 [J]. 北京航空航天大学学报，2024，50(2)：588-595.
GONG K C，ZHOU M L，TANG D M. Region-aware real-time portrait super resolution reconstruction network[J]. Journal of Beijing University of Aeronautics and Astronautics，2024，50(2)：588-595 (in Chinese).

陷: ①未结合区域感知信息来对不同区域做不同处理, 而是对全图像进行处理; ②对传递的全特征图进行了计算, 影响模型推理速度。对于实时人像超分辨率重建任务, 人像区域是兴趣区域 (area-of-interest, AoI), 而从特征图角度出发, Han 等[10] 指出, 特征图之间存在较多冗余。

为此, 本文提出了一个区域感知实时人像超分辨率重建网络 RASR。RASR 由 2 部分组成: 区域感知模块和超分辨率重建模块。受 U-Net[13]、BiSeNet[14] 和 Fan 等[15] 提出短期密集级联网络 (short-term dense concatenate, STDC) 的启发, 在训练传统语义分割网络的同时, 训练一个轻量门控单元, 指导其输出人像轮廓区域。超分辨率重建模块对特征图进行处理前, 会依据概率分布图进行区域划分, 人像边缘区域使用大规模超分辨率重建模型进行处理, 而对背景区域和其他人像区域, 根据计算设备算力使用小规模模型或传统插值方法进行处理。针对实时人像超分辨率重建场景, 本文的主要贡献如下:

1) 设计了一个轻量门控单元, 能准确、快速地对人像轮廓进行提取。

2) 设计了一种区域感知策略的超分辨率重建模型, 对于人像轮廓区域使用大尺寸模型进行处理。

3) 进一步改进了超分辨率重建模块, 提出通道划分特征提取器 (channel-wise split feature extractor, CSFE), 对特征图中的一部分进行处理, 结合注意力机制, 以较少的计算需求获得了较好的重建效果。

1　本文方法

1.1　区域感知模块

1.1.1　整体结构

区域感知模块整体结构如图 1(a) 所示。在编码阶段, 第 1、3、5 层为二维卷积层、批归一化层 (BN)、非线性激活层 (ReLU) 和注意力层的组合, 如图 1(b) 所示, 并在 CBR1 层和 CBR2 层后接入平均池化操作来对特征图进行下采样, 符合特征金字塔结构; 第 2 层和第 4 层由 He 等[16] 提出的多尺度特征集成块 (multi-scale feature integration block, MFIB) 堆叠而成。在解码阶段, 第 1 层为 2 倍上采样层, 上采样后的特征图与编码阶段第 2 层的输出特征图进行置信度计算, 并最终与自身相加; 第 2 层为上采样层、卷积层和批归一化层的组合, 负责输出最终的单通道分割图像; 可信特征提取器 (trusted feature extractor, TFE) 负责寻找解码阶段具有歧义的像素。使用轮廓损失指导 MFIB1 和 MFIB2 输出, 增强模型输出轮廓信息的能力。图 1 展示了一个完整的轻量语义分割模型。

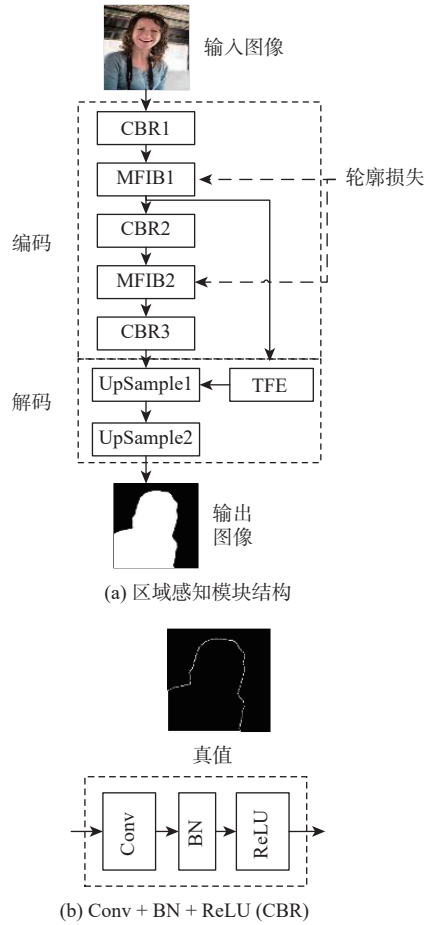


图 1　轻量语义分割模型网络结构

Fig. 1　Network architecture of lightweight semantic segmentation model

1.1.2　可信特征提取器

由于轻量语义分割模型在多个地方插入了平均池化操作, 平均池化操作虽然能较大地减少特征图的大小, 从而减少计算量, 但也不可避免地会造成空间分辨率的损失, 导致图像的全局一致性遭到破坏。然而, 颜色和纹理通常包含在相当低级的信息中, 即通常包含在高分辨率特征图中, 为提取颜色和纹理, 需要从低级语义信息中进行提取[17-18]。因此, 引入一种新的可信特征提取器, 从低级语义信息中提取出包含丰富色彩信息和纹理信息但置信度不同的特征, 并将其传递给解码阶段, 使解码器更加关注置信度低的特征。可信特征提取器结构如图 2 所示。

可信特征提取器从本质上来说是一个门控单元, 是一个“即插即用”模块。传统解码阶段直接使用低级语义信息会带来极大的噪声, 这些噪声会在编码阶段被消除, 但在低级语义信息中被保留下来。因此, 本文设计了一种门控系统, 将低级语义信息进行置信度计算, 只保留必要的轮廓信息, 去

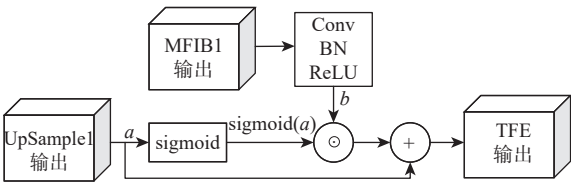


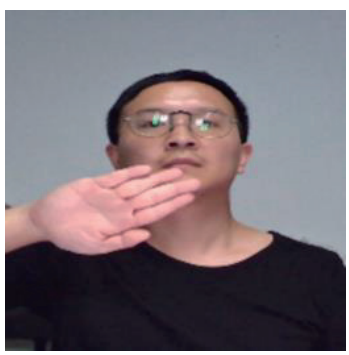
图2 可信特征提取器结构

Fig. 2 Trusted feature extractor structure

除噪声的干扰。低级语义信息即MFIB1的输出特征图通常包含丰富的色彩和纹理特征，通过卷积、批归一化和非线性激活后提取出具有高置信度的纹理，令处理后的特征图为 b。经过UpSample1层输出的高级语义信息 a 经过sigmoid激活函数后统一到0～1之间，可以看作是一个概率值，记为sigmoid(a)。因为要让解码器关注低置信度区域，所以计算1−sigmoid(a)后再与 b 逐像素点求乘积，即可得到解码器更需关注的、去除噪声后的低置信度高分辨率特征图，最终与高级语义特征图相加后输出，具体如下：

$$O_{\mathrm{TFE}} = a + \left(1 - \mathrm{sigmoid}(a)\right) \odot b \tag{1}$$

可信特征提取器输出的特征图可以当作一个人像轮廓特征图，因为其更关注特征图中具有不确定性的特征点，而这类特征点往往是人像边缘区域，从可视化特征图可以很明显地看出来，如图3所示。因此，本文将可信特征提取器用于人像边缘复杂区域的区域感知，配合超分辨率重建模块进行有选择的重建。



(a) 低分辨率图像



(b) 可信特征提取器输出特征图

图3 低分辨率图像与可信特征提取器输出特征图对比

Fig. 3 Comparison of low-resolution image and TFE output features

1.1.3 轮廓损失

首先，使用一个固定核（本文使用拉普拉斯核）对真实分割图的轮廓进行提取；然后，对MFIB1和MFIB2输出的特征图进行一次1×1的逐点卷积以调整通道数，调整为与轮廓图特征通道数相同；最后，使用交叉熵损失和Dice Coefficient损失对轮廓图与特征图之间的差异进行求解，并利用反向传播更新MFIB1和MFIB2的参数，引导模型的浅层学习空间细节特征，使其输出更多的轮廓信息，并使单分支语义分割模型也具有上下文旁路才有的提取空间细节特征的能力。损失函数如下：

$$l_{\mathrm{BCE}}(A,B) = \frac{1}{N}\sum_i L_i = \frac{1}{N}\sum_i -[y_i \log_2 p_i + (1-y_i)\log_2(1-p_i)] \tag{2}$$

$$l_{\mathrm{Dice}}(A,B) = \frac{2|A \cap B|}{|A| + |B|} \tag{3}$$

$$l_{\mathrm{Profile}}(A,B) = \omega_1 l_{\mathrm{BCE}}(A,B) + \omega_2 l_{\mathrm{Dice}}(A,B) \tag{4}$$

式中：A 为轮廓图；B 为MFIB1和MFIB2经过逐点卷积后输出的特征图；l_{BCE} 为交叉熵损失；N 为特征数；y_i 为轮廓图 A 中第 i 个特征；p_i 为特征图 B 中第 i 个特征；l_{Dice} 为Dice Coefficient损失；$|A \bigcap B|$ 为 A 与 B 的重叠区域像素个数；$|A|$、$|B|$ 分别为轮廓图 A、特征图 B 的像素个数；总损失 $l_{\mathrm{profile}}(A,B)$ 为 l_{BCE} 和 l_{Dice} 的加权求和；ω_1 和 ω_2 分别为 l_{BCE} 和 l_{Dice} 的权重，本文设置 ω_1=0.5，ω_2=1。

基于MFIB1可一定程度的输出轮廓信息，使用可信特征提取器可将该高可信的轮廓信息与解码器阶段拥有的较低置信度区域相结合，融合特征空间后输出给最后一层，从而有效提升模型的性能。

重要的是，轮廓损失仅在训练阶段对MFIB1和MFIB2进行指导，在推理阶段并不使用，因此，该单分支语义分割模型能对人像区域进行高效率的分割。

1.2 超分辨率重建模块

本文根据AoI特性及区域感知模块，设计了一种区域感知超分辨率重建网络。超分辨率重建模块整体结构如图4所示。

1.2.1 区域感知重建策略

首先，将低分辨率图像和可信特征进行切分，切分数量为8×8，低分辨率图像分辨率为224×224。然后，对切分后的低分辨率图像按照是否为AoI进

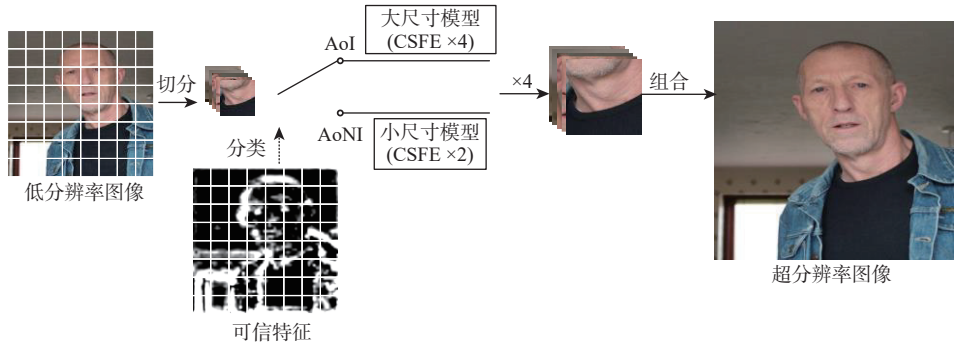


图 4　超分辨率重建模块整体框架示意图

Fig. 4　Schematic diagram of overall framework of super-resolution reconstruction module

行分类, 分类依据为区域感知模块输出的可信特征图。可信特征图与低分辨率图像统一被切割为 28×28 分辨率的小图, 设定一个简单的判别方法对 64 个区域依次进行判断分类: 若可信特征小图内有值为 1 的像素点, 则该区域为 AoI, 如图 4 中可信特征图里带有白色像素的小图; 反之则为 AoNI, 如图 4 中可信特征图里的全黑色小图。模型针对 AoI 与 AoNI 分别进行重建。由于 AoI 比 AoNI 更加重要, 而较大尺寸模型往往拥有更优秀的特征提取能力[19], 因此, 选择较大尺寸的模型对 AoI 进行特征提取, 如图 4 中大尺寸模型, 其采用 4 个通道划分特征提取器堆叠。而对于 AoNI, 则使用小尺寸模型提取特征, 本文采用 2 个通道划分特征提取器堆叠。大尺寸模型整体框架如图 5 所示, 小尺寸模型整体框架如图 6 所示。

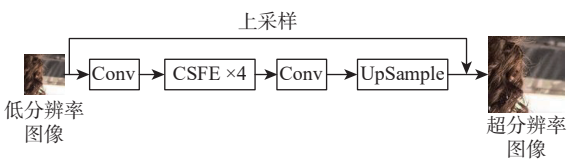


图 5　大尺寸模型框架

Fig. 5　Large scale model framework

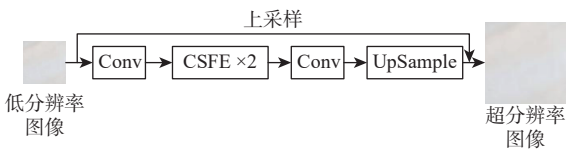


图 6　小尺寸模型框架

Fig. 6　Small scale model framework

切分后的低分辨率图像经过卷积核为 3×3、stride 为 1、padding 为 1 的卷积操作预处理后, 输入通道划分特征提取器。提取后的特征图再经过一个普通二维卷积后, 使用亚像素卷积技术进行上采样。

最终将 8×8 张超分辨率图像组装成一张分辨率为 896×896 的超分辨率重建图像。

1.2.2　通道划分特征提取器

经过多层卷积神经网络处理后, 特征图的分辨率普遍会下降, 但通道数会急剧上升[20], Han 等[10]指出, 通道之间具有一定的冗余性, 并不是每个通道的特征图对于超分辨率重建模型都同等重要。从通道注意力机制也能很好的证明, 通道注意力机制对通道间的依赖关系进行建模[11], 通过给予不同通道不同的权重, 对特征图进行通道级别的调整, 使模型更关注于一部分通道的特征, 并且抑制另一部分无用通道特征。基于此, 本文在通道注意力层前引入了一种新的通道划分特征提取块 (channel-wise split block, CSB)。

CSB 将输入的特征图从通道层面进行划分, 将特征图划分为 2 个分支, 划分比例设置为 ∂。设输入通道数为 M, 则用于进一步特征提取的特征图通道数为 $\partial \times M$。由于通道划分操作会给原特征图带来空间上的损失, 破坏空间一致性, 对该 $\partial \times M$ 通道的特征图进行处理时不能使用 1×1 的卷积核进行逐点卷积, 只能使用 3×3 大小的卷积核或 1×1 大小的卷积核。$\partial \times M$ 通道的特征图经过卷积层和激活层后, 与余下的 $(1-\partial) \times M$ 通道的特征图通道直接连接, 形成新的 M 通道特征图。实验时, 不同 ∂ 取得的效果与产生的处理时延如表 1 所示。经验证, ∂=0.25 时模型有最优组合性能。由于仅有较少特征图进行了处理, 计算量较小, 模组可以堆叠更多的 CSFE 以增加模型拟合训练数据的能力。

在堆叠多次残差块和 CSB 后, 加入通道方向上

表 1　CSFE 模块性能对比

Table 1　Performance comparison of CSFE module

∂	MAC/10^9	时延/ms	mPSNR/dB	mSSIM
0.25	28.29	20.87	32.94	0.938 8
0.50	29.01	22.78	32.97	0.940 1
0.75	29.12	24.92	33.01	0.940 4
1.00	29.21	27.74	33.03	0.940 5

注: MAC为前向传播每秒执行的乘累加计算量, mPSNR为平均峰值信噪比, mSSIM为平均结构相似度。

的注意力层，用于自适应调整各通道的响应值，使模型动态调整特征通道的信息量。

图7展示了由2个残差块、4个CSB、通道注意力层组合而成的通道划分特征提取器。

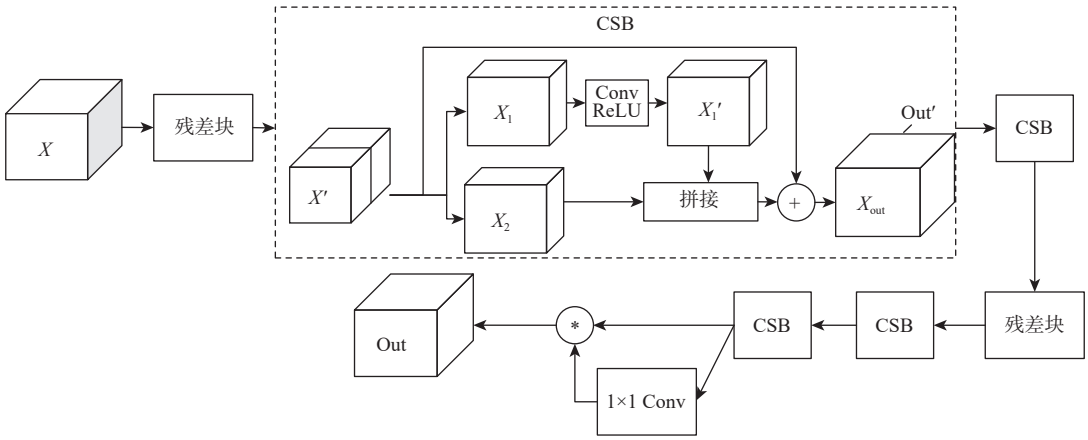


图7 通道划分特征提取器结构

Fig. 7 Structure of channel-wise split feature extractor

2 实验结果与分析

本节设计了区域感知策略的实时人像超分辨率重建实验来对本文提出的重建模型进行验证，与双三次插值法Bicubic及现有方法进行对比，并对实验数据进行分析。

2.1 数据集

对于分割模型网络的训练，本文使用的人像matting数据集aisegmentc_datasets是目前已知最大的人像matting数据集，包含34 427张图像和对应的matting结果图。针对在线会议场景对该数据集进行扩充处理，采集了真实在线会议场景下的视频截图，数量为1 080张，并对其使用开源标注软件LabelMe进行人像区域标注，合并进aisegmentc_datasets数据集后，彩色图像与二值分割图像数量均为35 507张，图像分辨率为896×896。

对于超分辨率重建模型的训练，本文使用扩充后的aisegmentc_datasets数据集图像作为输入重建模型的高分辨率图像，再对高分辨率图像使用OpenCV提供的resize()方法进行下采样处理，分别获得下采样2倍与下采样4倍的图像作为输入重建模型的低分辨率图像，使用的插值方法为双线性插值法。

2.2 训练参数设定

首先，训练语义分割网络。本文使用交叉熵损失函数计算模型输出分割图与真实分割图之间的差异；使用Adam作为优化器，优化器参数momentum1=0.5，momentum2=0.999；学习率设置为5×10^{-3}，以余弦函数调整学习率，周期为20；训练总轮数为200轮。

其次，训练重建倍率为2倍、4倍的超分辨率重建网络。与语义分割网络的训练方法相似，但本文使用的超分辨率网络由2部分组成：参数量多的大尺寸模型和参数量少的小尺寸模型。2种模型都使用数据集中所有图像进行训练。由于小尺寸模型参数量更少，更易产生过拟合现象，大、小模型学习率分别设置为1×10^{-4}、2×10^{-4}，训练总轮数分别为150轮、250轮。训练与测试所用的GPU为NVIDIA GTX 1080Ti。

2.3 评价标准

对于重建模型重建出的图像好坏的评价标准，本文采用平均峰值信噪比mPSNR和平均结构相似度mSSIM作为客观评价指标，将模型推理时延和前向传播每秒执行的乘累加计算量MAC作为评价标准。

2.4 区域感知模块结果分析

本节测试了区域感知模块，该模块主要用于提取编码阶段置信度低的特征，而低级语义信息通常包含丰富的色彩和纹理特征，具体信息为人像信息和人像轮廓边缘信息。

从图3可以看出，可信特征提取器很好地对图像中需要重点关注的人像边缘纹理信息进行了提取，而对背景区域则给予了高置信度。在解码阶段，解码器会对人像及其轮廓区域给予高权重，更有利于恢复出高可靠的人像区域。表2展示了加入可信特征提取器与不加入可信特征提取器的语义分割指标对比。

加入可信特征提取器后，分割模型的推理时延略微增加，mPA和mIoU指标大幅增加，证明了编码阶段高分辨率特征图中提取出的低级语义信息的重要性，并在解码阶段作为一种输入，用于指导

表 2　可信特征提取器应用前后语义分割模型性能对比

Table 2　Performance comparison of semantic segmentation models before and after application of TFE

有无TFE	参数量/10^3	MAC/10^6	时延/ms	mPA/%	mIoU
无	150.08	97.01	1.01	91.7	93.26
有	150.18	97.15	1.02	92.2	94.81

注: mPA为平均像素准确率, mIoU为平均交并比。

解码器恢复高可信度人像分割信息。

2.5　超分辨率重建模块结果分析

2.5.1　通道划分特征提取器结果

本节测试了图 7 所示的通道划分特征提取器, 包括不同划分比例∂对通道划分特征提取器的影响、计算量对比、推理时延、模型输出结果与原高分辨率图之间 PSNR 和 SSIM 对比。

表 1 展示了不同∂对通道划分特征提取器的影响。∂越大, 代表越多通道用于进一步的特征提取, 意味着能更好地拟合输入数据, 但是会导致计算量、时延等增加。实验结果表明, 当通道划分比例∂=0.25 时, 即模型对 25% 的特征图进行进一步特征提取, 而对剩下 75% 的特征图采用直接堆叠的方式, 能获得最低时延和相对较好的 mPSNR、mSSIM, ∂=1.00 时, 即可视为普通卷积模块。为在时延、mPSNR 和 mSSIM 之间获得平衡, 本文使用∂=0.25 作为通道划分特征提取器的参数进行模型的搭建。

2.5.2　与现有方法进行对比

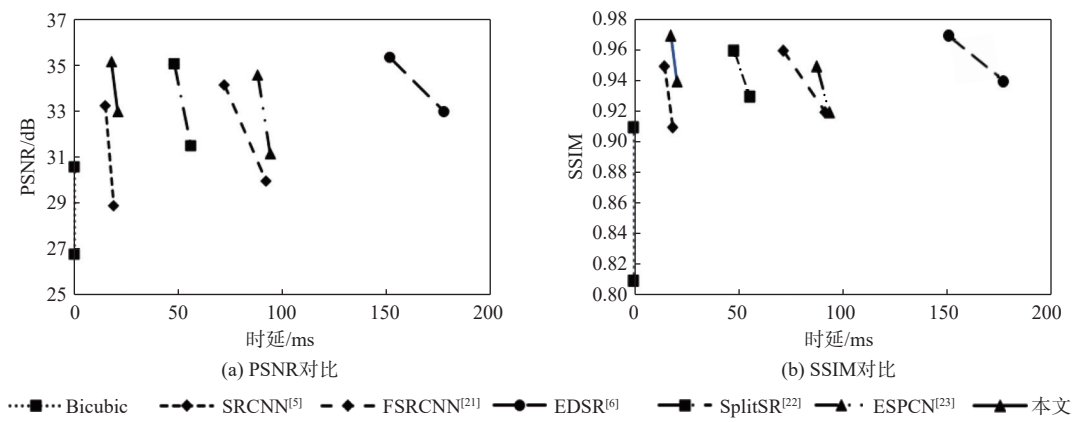
将本文提出的区域感知超分辨率重建模型与现有方法进行对比实验, 参与对比实验的方法有传统插值方法 Bicubic, 基于深度学习的超分辨率重建方法 SRCNN[5]、FSRCNN[21]、EDSR[6]、SplitSR[22] 和 ESPCN[23]。基于深度学习的超分辨率重建模型文件均为其官方实现。输入模型的低分辨率图像分辨率均为 224×224, 上采样倍率 s 为 4 倍。实验结果如表 3 和图 8、图 9 所示。

结果表明, 本文提出的基于分区重建策略的实时人像超分辨率重建模型在在线会议场景下的重建效果明显。在以 2 倍和 4 倍上采样倍率重建时, 相比于传统的双三次插值法 Bicubic, 本文模型在 PSNR 评价指标上有 4.57 dB 和 6.2 dB 的提升, 在 SSIM 评价指标上有 0.06 和 0.13 的提升, 其在高倍率重建时有着很大的优势。而对于经典的超分辨率重建方法, 如 FSRCNN[21]、EDSR[6] 等, PSNR 与 SSIM 指标虽然提升不大, 但在处理时延方面则有较大优势, 这是由于 FSRCNN[21]、EDSR[6] 的模型参数较多, 需要较长的推理时间。在 2 倍上采样倍率重建时, 本文模型比现有大模型 EDSR[6] 推理时延降低了 88.08%, 4 倍上采样倍率重建时降低了 88.14%, 并且 PSNR 与 SSIM 指标仅有略微降低。本文模型参数量并不是最少的, 其推理时延在对比方法中有优势, 这是因为其使用了区域感知重建策略, 在人像复杂区域使用大尺寸模型进行处理, 在背景区域则使用小尺寸模型进行处理, 避免了对整图进行复杂、密集的计算, 大大降低了推理时延。由图 8 可以看出, 本文模型在图中坐标点均为偏左上, 兼具了推理时延与 PSNR/SSIM 指标。相较于参数量更少的 SRCNN 模型, 虽然其有着更少的参

表 3　2 倍和 4 倍重建倍率下各模型的性能对比

Table 3　Performance comparison of various models under 2 and 4 reconstruction magnification

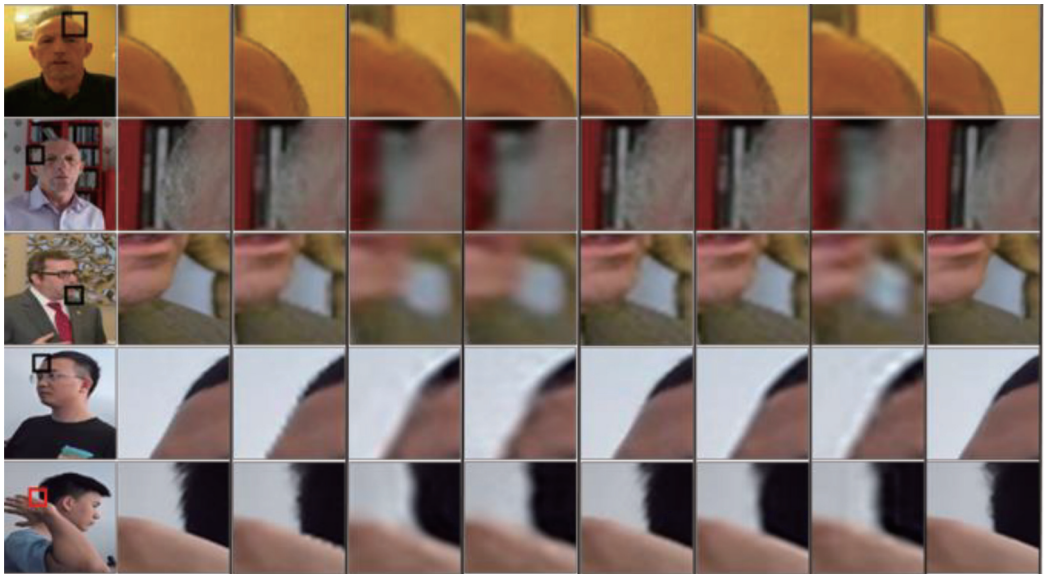
方法	PSNR/dB		SSIM		时延/ms	
	s=2	s=4	s=2	s=4	s=2	s=4
Bicubic	30.56	26.74	0.91	0.81	0.2	0.2
SRCNN[5]	33.21	28.86	0.95	0.91	15	19
FSRCNN[21]	34.12	29.93	0.96	0.92	72	92
EDSR[6]	35.31	32.96	0.97	0.94	151	177
SplitSR[22]	35.05	31.46	0.96	0.93	48	56
ESPCN[23]	34.54	31.11	0.95	0.92	88	94
本文	35.13	32.94	0.97	0.94	18	21



(a) PSNR对比　　(b) SSIM对比

图 8　不同模型性能对比

Fig. 8　Performance comparison of different models



(a) 高分辨率图像 (b) Bicubic (c) SRCNN[5] (d) FSRCNN[21] (e) EDSR[6] (f) SplitSR[22] (g) ESPCN[23] (h) 本文

图 9 不同模型重建效果对比

Fig. 9 Reconstruction comparison of different models

数量与更低的推理时延，但是从图 8 中所处点的位置可以看出，SRCNN 均不如本文模型。因此，综合考虑推理时延和 PSNR/SSIM，本文模型效能最优。

3 结 论

本文提出了一种区域感知超分辨率重建模型。在图像分割模型中加入可信特征提取模块，该模块的引入虽轻微增加了模型的推理时延，但在 mPA 和 mIoU 指标上实现了显著提升，证实了编码阶段提取的低级语义信息在解码阶段用于人像分割信息恢复的重要性。本文探索了不同通道划分比例对 CSFE 模块的影响，实验结果表明，当划分比例为 0.25 时，即仅对 25% 的特征图进行深度特征提取，模型能够在保持较低时延的同时，提供相对较好的 mPSNR 和 mSSIM 性能。本文提出的区域感知超分辨率重建模型在多种评价指标和在线会议场景下的表现均优于现有方法，尤其在推理时延上，相比于大尺寸模型，实现了显著提速。

未来研究将聚焦于以下方面：①参数优化与模型压缩；②算法泛化性研究；③实时处理能力提升；④更深入的特征提取策略研究。

参考文献（References）

[1] KEYS R. Cubic convolution interpolation for digital image processing[J]. IEEE Transactions on Acoustics, Speech, and Signal Processing, 1981, 29(6): 1153-1160.

[2] KIM J, LEE J K, LEE K M. Deeply-recursive convolutional network for image super-resolution[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2016: 1637-1645.

[3] 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.

[4] LIU F, YU Q, CHEN L, et al. Aerial image super-resolution based on deep recursive dense network for disaster area surveillance[J]. Personal Ubiquitous Computing, 2022, 26: 1205-1214.

[5] DONG C, LOY C C, HE K, et al. Learning a deep convolutional network for image super-resolution[C]//Proceedings of the European Conference on Computer Vision. Berlin: Springer, 2014: 184-199.

[6] 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. Piscataway: IEEE Press, 2017: 136-144.

[7] HE K, ZHANG X, REN S, 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.

[8] TONG T, LI G, LIU X, et al. Image super-resolution using dense skip connections[C]//Proceedings of the IEEE International Conference on Computer Vision. Piscataway: IEEE Press, 2017: 4799-4807.

[9] SZEGEDY C, LIU W, JIA Y, et al. Going deeper with convolutions[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2015: 1-9.

[10] HAN K, WANG Y, 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: 1580-1589.

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

[12] WANG X, GIRSHICK R, GUPTA A, et al. Non-local neural networks[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2018:

7794-7803.

[13] RONNEBERGER O, FISCHER P, BROX T. U-Net: Convolutional networks for biomedical image segmentation[C]//Proceedings of the 18th International Conference on Medical Image Computing and Computer-Assisted Intervention. Berlin: Springer, 2015: 234-241.

[14] YU C, WANG J, PENG C, et al. BiSeNet: Bilateral segmentation network for real-time semantic segmentation[C//Proceedings of the European Conference on Computer Vision. Berlin: Springer, 2018: 334-349.

[15] FAN M, LAI S, HUANG J, et al. Rethinking bisenet for real-time semantic segmentation[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2021: 9716-9725.

[16] HE Z, LIU K, LIU Z, et al. A lightweight multi-scale feature integration network for real-time single image super-resolution[J]. Journal of Real-Time Image Processing, 2021, 18(4): 1221-1234.

[17] ACUNA D, LING H, KAR A, et al. Efficient interactive annotation of segmentation datasets with Polygon-RNN++[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2018: 859-868.

[18] LING H, GAO J, KAR A, et al. Fast interactive object annotation with Curve-GCN[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2019: 5257-5266.

[19] BELKIN M, HSU D, MA S, et al. Reconciling modern machine-learning practice and the classical bias-variance trade-off[J]. Proceedings of the National Academy of Sciences, 2019, 116(32): 15849-15854.

[20] 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.

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

[22] LIU X, LI Y, FROMM J, et al. SplitSR: An end-to-end approach to super-resolution on mobile devices[C]//Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies.New York:ACM, 2021, 5(1): 1-20.

[23] SHI W, 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.

Region-aware real-time portrait super resolution reconstruction network

GONG Kecun[1, 2], ZHOU Menglin[3], TANG Dongming[1, 2, *]

(1. Key Laboratory for Computer Systems of National Ethnic Affairs Commission, Southwest Minzu University, Chengdu 610041, China;
2. School of Computer Science and Engineering, Southwest Minzu University, Chengdu 610041, China;
3. School of Software Engineering, Chengdu University of Information Technology, Chengdu 610225, China)

Abstract: Conventional techniques typically process the entire image uniformly, which leads to low efficiency in the field of portrait super-resolution reconstruction.To reduce the inference latency of the model, this research proposes a real-time super-resolution reconstruction model RASR. The model first uses gating unit to process the low-resolution images and identify the edge of the portrait. Then, a partition reconstruction strategy is adopted, and sub-models of different sizes are used to reconstruct the areas containing or not containing the portrait edge, respectively. The experimental results show that the RASR model is able to reconstruct high-resolution portrait images more efficiently by reducing the inference latency of the RASR model by 88% in a 4-foldsampling reconstruction scene compared to the existing methods.

Keywords: region aware; single-image super-resolution; gating unit; channel-wise split block; deep learning

Received: 2022-05-20; **Accepted**: 2022-07-02; **Published Online**: 2022-09-20 11:55

URL: link.cnki.net/urlid/11.2625.V.20220919.1238.001

Foundation items: National Natural Science Foundation of China (61873217); Sichuan Science and Technology Program (2021JDRC0063); The Fundamental Research Funds for the Central Universities, Southwest Minzu University (2021118)

* **Corresponding author**. E-mail: tdm_2010@swun.edu.cn

http://bhxb.buaa.edu.cn jbuaa@buaa.edu.cn
DOI: 10.13700/j.bh.1001-5965.2022.0395

基于多模态双向信息增强的RGBT跟踪网络

赵伟[1,2]，刘磊[1,2]，王鲲鹏[1,2]，涂铮铮[1,2,*]，罗斌[1,2]

(1. 多模态认知计算安徽省重点实验室，合肥 230601；2. 安徽大学 计算机科学与技术学院，合肥 230601)

摘 要：可见光-热红外(RGBT)目标跟踪旨在挖掘可见光和热红外数据的互补优势，实现鲁棒的目标跟踪。目前主流方法通常引入模态权重来实现多模态信息融合，但简单地为各个模态分配权重无法充分挖掘可见光和热红外模态的互补优势。基于此，提出了一种多模态双向信息增强的RGBT跟踪网络(MBIENet)。设计了一种特征聚合模块，用于聚合模态共享特征和模态特定特征以建模目标外观信息；提出了一种新的多模态双向调制融合模块，可有效融合模态互补信息，减少冗余特征和无用特征对跟踪器的影响；提出了一个轻量化的通道空间注意力模块，可自适应调整不同环境下不同模态的贡献。在GTOT、RGBT234和LasHeR数据集上的实验结果表明：所提跟踪算法的准确率和成功率优于当前主流的跟踪算法。

关 键 词：可见光-热红外；目标跟踪；深度学习；多模态信息融合；多模态信息交互

中图分类号：TP183

文献标志码：A **文章编号**：1001-5965(2024)02-0596-10

视觉目标跟踪在无人驾驶、医学图像处理和现代化军事等计算机视觉领域中扮演着重要角色。近年来，虽然该领域研究取得了巨大的进展，但仍然存在一些亟待解决的问题。例如，当照明条件较差或光照变化很大时，基于可见光图像的跟踪算法难以实现鲁棒的跟踪性能[1-3]；而热红外图像对光照变化等因素不敏感，研究者也尝试利用热红外图像进行目标跟踪，但热红外图像通常分辨率低，纹理差，在某些条件下不可靠[4-6]。因此，通过融合可见光-热红外(RGB-thermal infrared, RGBT)图像的互补信息，可以大大提高跟踪算法的鲁棒性。

RGBT视觉目标跟踪旨在利用可见光和热红外图像的互补优势，在给定第1帧目标初始状态的情况下，估计视频序列后续帧中目标的状态。由于可见光和热红外数据的强大互补优势，RGBT目标跟踪器可以在低照度、恶劣天气等不利条件下良好地工作。

可见光图像和热红外图像是在同一场景不同光谱下的成像，尽管成像原理不同，但仍具有共有的信息，如目标的边界信息、细粒度的纹理信息等，称为模态的共享特征。其中，可见光图像具有更加丰富的颜色和纹理信息，热红外图像则具有物体表面的热量分布信息，这些不一致的信息称为模态的特定特征。充分利用可见光和热红外图像的模态特定特征和模态共享信息，可以有效提升RGBT目标跟踪器的性能。

依据参数共享[7]的形式，现有的RGBT跟踪器主要分为2种结构。第1类为软共享结构，即通过为每个模态设计一个单独的特征提取网络以提取模态特定特征，并通过信息访问的方式传递不同模态的模态共享和模态互补特征。Gao等[8]提出了一种基于深度自适应融合网络的RGBT跟踪框架(deep adaptive fusion network, DAFNet)，设计了一个自适应融合模块，通过为不同模态分配权重来融合

收稿日期：2022-05-20；**录用日期**：2022-07-02；**网络出版时间**：2022-10-18 17:41

网络出版地址：link.cnki.net/urlid/11.2625.V.20221018.0824.001

基金项目：国家自然科学基金(62376005)；安徽省重点研发计划(202104d07020008,KJ2020A0033)；安徽省自然科学基金(2108085MF211)；安徽省高校协同创新项目(GXXT-2022-014)

* **通信作者**. E-mail：zhengzhengahu@163.com

引用格式：赵伟，刘磊，王鲲鹏，等. 基于多模态双向信息增强的RGBT跟踪网络[J]. 北京航空航天大学学报，2024，50(2)：596-605. ZHAO W，LIU L，WANG K P，et al. Multimodal bidirectional information enhancement network for RGBT tracking[J]. Journal of Beijing University of Aeronautics and Astronautics，2024，50(2)：596-605 (in Chinese).

互补信息。Li 等[9] 提出了一种新的挑战感知的 RGBT 跟踪器 (challenge-aware RGBT tracker, CAT), 处理 RGBT 跟踪的模态共享挑战和模态特定挑战, 并将挑战分支的特征信息聚合到主干网络中以增强模型性能。但值得注意的是, 基于挑战感知的 RGBT 跟踪网络 (CAT[9]、ADRNet[10]、APFNet[11]) 需要人工标注挑战标签或生成不同挑战的虚拟数据, 用来训练不同的挑战分支, 训练成本较高。

第 2 类为硬共享结构, 即只使用一个特征提取网络同时提取不同模态的模态共享特征, 并通过特征补充的方式添加额外分支去挖掘不同模态的特定特征与互补特征。Li 等[12] 分析了模态共享特征和模态特定特征对 RGBT 跟踪的重要性, 并提出了一种多适配器卷积网络 (multi-adapter convolutional network, MANet), 通过设计不同的适配器聚合模态共享特征和模态特定特征以建模目标外观信息, 从而进行鲁棒的 RGBT 目标跟踪。然而, MANet 等网络对模态共享、特定特征的利用采用了简单的特征求和操作[13], 忽略了单一模态的共享特征可能不可靠的情况, 限制了跟踪性能的进一步提升。

为解决上述问题, 本文利用软共享结构和硬共享结构的特性, 提出了一种新的多模态双向信息增强网络 (multimodal bidirectional information enhancement network, MBIENet)。该网络可充分利用模态共享特征与模态特定特征建模单模态外观特征, 并采用多模态双向调制融合模块自适应地传递模态互补特征, 提高跟踪器的性能。

1　网络模型

1.1　MBIENet 整体结构

如图 1 所示, 本文基于 MDNet[14] 构建网络, 使用在 ImageNet[15] 数据集上预训练的 VGG-M[16] 网络作为特征提取器, 为权衡跟踪器的速度和精度, 只保留前 3 层 (Conv1 ~ Conv3) 提取多模态特征, 卷积核大小分别为 7×7、5×5、3×3, 在第 1 层和第 2 层之后采用 ReLU 激活函数、局部响应归一化 (local response normalization, LRN) 和最大池化层操作, 第 3 层只采用 ReLU 激活函数。为提取模态特定特征, 网络为每个模态分配一个具有相同结构但并不共享参数的特征提取器 (见图 1 中蓝、橙色卷积层)。此外, 在每个卷积层, 本文采用一个与双流分支相同结构的特征提取器对模态共享特征进行提取 (见图 1 中绿色卷积层)。

如图 1 所示, 网络的输入为成对的可见光和热红外图像, 在每个卷积层上, 先利用特征提取器提取每个模态的模态共享特征和模态特定特征, 并利用特征聚合模块 (feature aggregation module, FAM) 对这 2 种特征进行聚合, 以对单个模态中目标的外观进行建模。再通过多模态双向调制融合模块 (multimodal bidirectional modulation fusion module, MBMF) 自适应融合多模态互补特征, 并聚合到双流分支上, 进行信息增强。在双流分支的第 3 个卷积层后, 网络拼接可见光模态和热红外模态的特征作为融合特征, 并利用通道空间注意力模块 (channel-spatial attention module, CSA) 对其加权调整。最终将加权调整后的融合特征送入 3 个全连接层 (FC4 ~ FC6), 结合 SoftMax 函数对当前帧的前景和背景进行判别, 从而定位目标位置。全连接层的输出维度分别为 512、512、2。其中, 第 3 层全连接层 FC6 与 MDNet[14] 类似, 其拥有 k 个分支, 每个分支对应一个特定的域, 即对应跟踪任务的每个视频序列。在跟踪过程中, 网络移除离线训练的多分支 FC6 层, 并初始化一个新的 FC6 层判别当前序列的前景和背景, 以适应该序列下目标外观的变化。

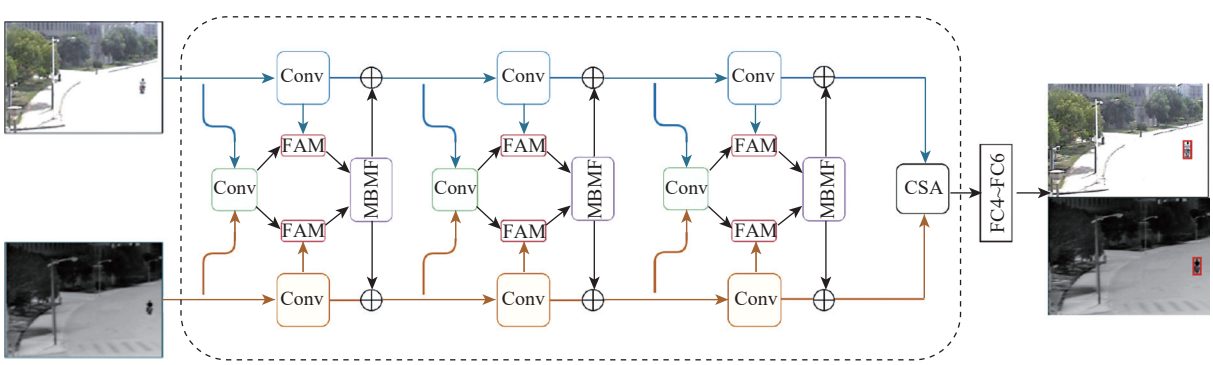


图 1　MBIENet 框架

Fig. 1　MBIENet framework

1.2　特征聚合模块

可见光图像和热红外图像是同一场景在不同光谱下的成像。可见光图像拥有更丰富的颜色、纹理细节特征; 热红外图像反映的是物体表面的温度分布。尽管二者有显著差别, 但仍具有共有信息, 如目标边界、细粒度纹理等。通过聚合模态的特定

特征和共享特征，可对每个模态中目标的外观进行建模，从而为目标跟踪提供更为丰富的判别信息。

现有的模态特定特征和模态共享特征的聚合方法通常对两者的特征进行求和[12]，却未考虑如下问题：①在复杂背景下，某一模态提供的特征可能是无效的。例如，在目标温度和背景相近时，热红外图像几乎无法提供可靠的判别信息（见图2(a)）；或在强光下，可见光图像几乎无法提供有效的信息（见图2(b)）。这些情况下，如果网络直接聚合其共享特征，会引入不必要的噪声。②模态共享特征、模态特定特征包含相似信息（如背景信息），应减少这些冗余信息对网络的影响。

为解决上述问题，本文提出了一个简单有效的FAM模块，其具体结构如图3所示。

为解决特定条件下某一模态不可靠的问题，本文设计了一个卷积层和Sigmoid激活函数作为门控机制，通过生成一组权重矩阵，来对输入的模态共享特征进行自适应调整，从而减轻特定条件下不可靠模态的噪声引入。本文采用特征求和的方式处理模态共享特征、模态特定特征，从而建模单个模态中更加鲁棒的外观特征，这些特征可以为模态融合提供有效的判别信息，其数学表示如下：

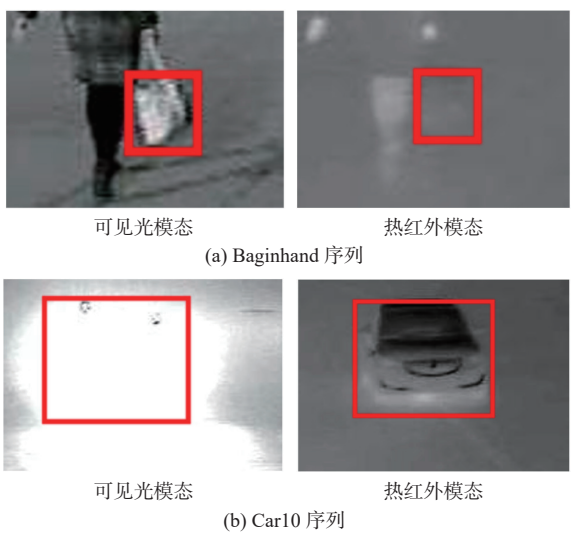


图2 模态不可靠示例

Fig. 2 Example of modal unreliability

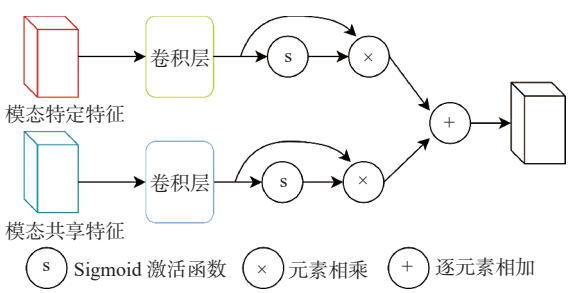


图3 特征聚合模块结构

Fig. 3 Structure of feature aggregation module

$$\begin{cases} \boldsymbol{f}_{\mathrm{c}}' = \mathrm{Conv}^{\mathrm{c}}(\boldsymbol{f}_{\mathrm{c}}) \\ \boldsymbol{f}_{\mathrm{u}}' = \mathrm{Conv}^{\mathrm{u}}(\boldsymbol{f}_{\mathrm{u}}) \\ \boldsymbol{f}_{\mathrm{com}} = \boldsymbol{f}_{\mathrm{c}}' \cdot \mathrm{Sigmoid}(\boldsymbol{f}_{\mathrm{c}}') + \boldsymbol{f}_{\mathrm{u}}' \cdot \mathrm{Sigmoid}(\boldsymbol{f}_{\mathrm{u}}') \end{cases} \tag{1}$$

式中：$\boldsymbol{f}_{\mathrm{c}}$和$\boldsymbol{f}_{\mathrm{u}}$分别为模态共享特征和模态特定特征；$\mathrm{Conv}^{\mathrm{c}}$、$\mathrm{Conv}^{\mathrm{u}}$代表卷积层，由大小为3×3的卷积核与ReLU激活函数组成，用来学习一组权重参数；$\boldsymbol{f}_{\mathrm{c}}'$、$\boldsymbol{f}_{\mathrm{u}}'$为提取后的模态特征；$\boldsymbol{f}_{\mathrm{com}}$为聚合模态共享特征、模态特定特征后的输出结果。

1.3 多模态双向调制融合模块

多模态特征融合是利用不同模态互补信息的一种有效手段。早期工作[17-19]基于贝叶斯框架中的协同稀疏表示，通过优化每个模态的可靠性权重进行融合，但当协同表示不足以反映模态可靠性时，将会失败。现有的深度学习RGBT跟踪器[8,13,20]通常为每个模态学习一组可靠的权重，从而自适应地对可见光模态和热红外模态进行特征融合。值得注意的是，为每个模态简单地分配权重无法充分挖掘RGBT的互补优势，且无法过滤冗余和无用特征带来的干扰。为弥补上述不足，本文提出一种新的MBMF模块以有效融合不同模态的信息，从而丰富特征图的特征表达，其具体结构如图4所示。

MBMF模块的输入特征来自FAM模块。为传递模态之间的互补信息，受FiLM[21]启发，本文利用一种模态的先验信息构建2个条件，分别用于缩放与移动特征，从而对另一个模态进行调制，同时引入Sigmoid激活函数生成门控机制来减少噪声的传播。

如图4所示，首先，使用卷积核大小为1×1的卷积层和ReLU激活函数提取每个模态的特征；其次，采用一组卷积核大小为3×3的卷积层提取特定模态的外观特征，并使用Sigmoid激活函数生成一组权值矩阵用于控制模态互补特征的传播，这些互补特征有利于增强目标在另一个模态中的特征表示；最后，以特征求和的方式将这些特征传递到另一个模态上。数学表示如下：

$$\begin{cases} \boldsymbol{f}_{\mathrm{R}}^{2} = \mathrm{Conv}_{\mathrm{R}}^{2}(\mathrm{Conv}_{\mathrm{R}}^{1}(\boldsymbol{f}_{\mathrm{RGB}})) \\ \boldsymbol{f}_{\mathrm{T}}^{\mathrm{S}} = \boldsymbol{f}_{\mathrm{T}} \cdot \mathrm{Sigmoid}(\boldsymbol{f}_{\mathrm{R}}^{2}) \end{cases} \tag{2}$$

$$\begin{cases} \boldsymbol{f}_{\mathrm{T}}^{2} = \mathrm{Conv}_{\mathrm{T}}^{2}(\mathrm{Conv}_{\mathrm{T}}^{1}(\boldsymbol{f}_{\mathrm{T}})) \\ \boldsymbol{f}_{\mathrm{R}}^{\mathrm{S}} = \boldsymbol{f}_{\mathrm{RGB}} \cdot \mathrm{Sigmoid}(\boldsymbol{f}_{\mathrm{T}}^{2}) \end{cases} \tag{3}$$

式中：$\boldsymbol{f}_{\mathrm{RGB}}$、$\boldsymbol{f}_{\mathrm{T}}$为输入MBMF模块的模态特征图；Conv表示卷积层，下标表示其作用的模态，上标表示其卷积层序号；$\boldsymbol{f}_{\mathrm{R}}^{2}$、$\boldsymbol{f}_{\mathrm{T}}^{2}$为第2层卷积层的输出；$\boldsymbol{f}_{\mathrm{R}}^{\mathrm{S}}$、$\boldsymbol{f}_{\mathrm{T}}^{\mathrm{S}}$为调制模态特征图。

虽然调制模态特征图包含另一个模态的互补信息，但会在一定程度上损失自身模态的判别特

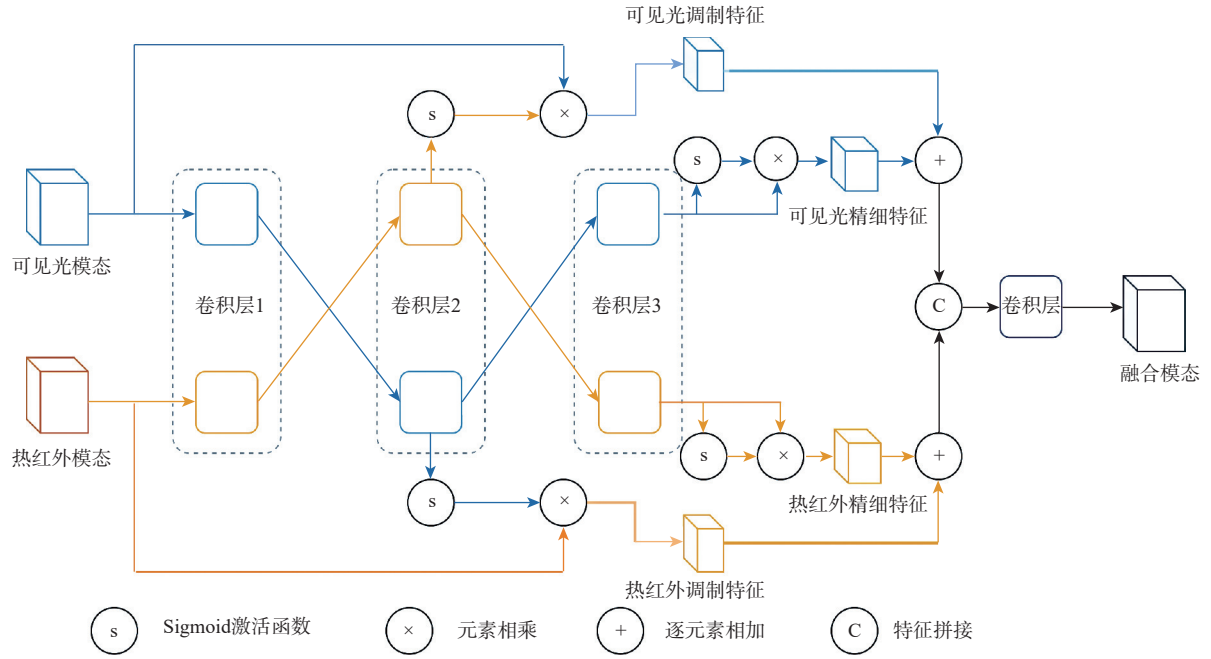


图 4　多模态双向调制融合模块结构

Fig. 4　Structure of multimode bidirectional modulation fusion module

征。为解决该问题, 本文设计了另一组分支去提取单模态的精细特征图, 如图 4 所示。具体来说, 为过滤冗余特征, 实现可靠的多模态融合, 采用大小为 3×3 的卷积核与 Sigmoid 门控机制对单模态特征图进行精细提取, 从而过滤冗余、无用特征对网络性能的干扰。将聚合调制模态特征和精细模态特征生成单模态的可靠特征表示, 拼接来自不同模态的可靠特征表示后, 采用一个卷积核大小为 3×3 的卷积层自适应地提取模态融合信息。

$$\begin{cases} \boldsymbol{f}_{\mathrm{R}}^{3}=\mathrm{Conv}_{\mathrm{R}}^{3}(\boldsymbol{f}_{\mathrm{R}}^{2}) \\ \boldsymbol{f}_{\mathrm{CPR}}=\boldsymbol{f}_{\mathrm{R}}^{\mathrm{S}}+\boldsymbol{f}_{\mathrm{R}}^{3}\cdot\mathrm{Sigmoid}(\boldsymbol{f}_{\mathrm{R}}^{3}) \end{cases} \tag{4}$$

$$\begin{cases} \boldsymbol{f}_{\mathrm{T}}^{3}=\mathrm{Conv}_{\mathrm{T}}^{3}(\boldsymbol{f}_{\mathrm{T}}^{2}) \\ \boldsymbol{f}_{\mathrm{CPT}}=\boldsymbol{f}_{\mathrm{T}}^{\mathrm{S}}+\boldsymbol{f}_{\mathrm{T}}^{3}\cdot\mathrm{Sigmoid}(\boldsymbol{f}_{\mathrm{T}}^{3}) \end{cases} \tag{5}$$

$$\boldsymbol{f}_{\mathrm{out}}=\mathrm{Conv}(\boldsymbol{f}_{\mathrm{CPR}}\oplus\boldsymbol{f}_{\mathrm{CPT}}) \tag{6}$$

式中: $\boldsymbol{f}_{\mathrm{R}}^{3}$、$\boldsymbol{f}_{\mathrm{T}}^{3}$为第 3 层卷积层的输出特征; $\boldsymbol{f}_{\mathrm{R}}^{\mathrm{S}}$、$\boldsymbol{f}_{\mathrm{T}}^{\mathrm{S}}$为调制模态特征图; $\boldsymbol{f}_{\mathrm{CPR}}$、$\boldsymbol{f}_{\mathrm{CPT}}$为单模态精细特征图; “$\oplus$”表示特征拼接; $\boldsymbol{f}_{\mathrm{out}}$为最终提取到的融合特征。

本文利用融合特征聚合到双流分支上, 用于对主干网络进行信息增强, 从而建模目标的鲁棒外观特征表示, 具体如图 4 所示。

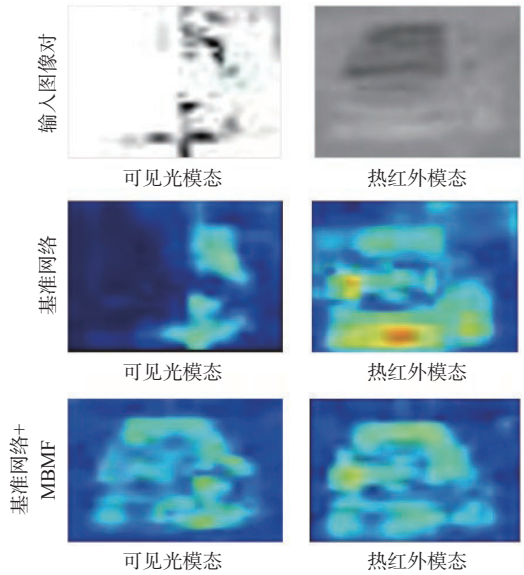
为验证本文所提模块的有效性, 在 2 个不同的序列上对模态特征图进行了可视化, 如图 5 所示。在图 5(a) 中, 由于强光影响, 可见光模态只能提取微弱的特征表示, 相反, 热红外模态可以提供大量的判别特征。在图 5(b) 中, 由于热交叉, 即目标热量分布与背景一致, 导致热红外模态无法提取有效的外观特征。可知, 加入 MBMF 模块可以有效传递模态的互补特征, 在抑制噪声的同时, 增强 2 种模态的外观表示。

1.4　通道空间注意力模块

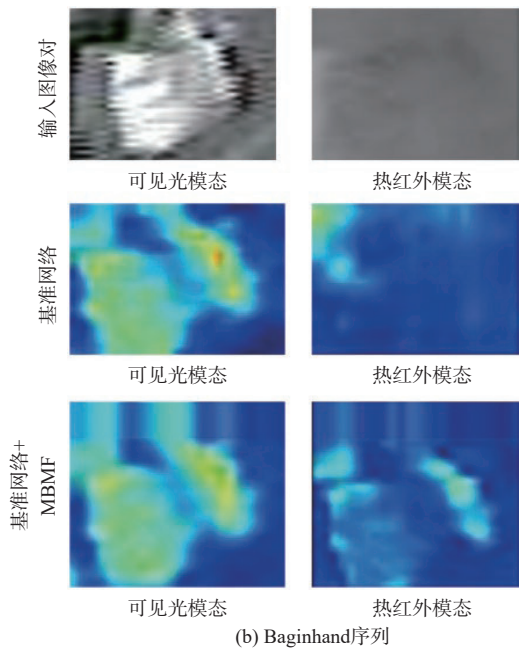
注意力机制在计算机视觉领域发挥着非常重要的作用, 通过模拟人脑对现实世界的认知, 使神经网络更加关注输入图像中更重要的信息。现有的注意力机制大多从输入特征的通道维度和空间维度学习一组注意力权重, 继而从这 2 个维度对输入特征进行加权。在 RGBT 目标跟踪领域, Xu 等[13] 利用全局平均池化和全局最大池化操作实现通道注意力机制, 并对所有卷积层进行自适应通道加权。Zhu 等[20] 使用全局平均池化和 3 个全连接层实现通道注意力机制, 并生成一个权重矩阵来平衡 2 种模态的贡献。为联合通道和空间注意力, Wang 等[22] 引入 CBAM[23] 模块对融合后的特征进行通道和空间上的加权, 有效提高了网络性能, 但其复杂的结构为网络带来了更多的计算开销。

为此, 本文提出了一种新的 CSA 模块, 在通道维度和空间维度上对融合特征进行调制, 从而平衡不同场景下不同模态的贡献, 并且其结构简单, 只采用了 2 个卷积层和 2 个全连接层实现操作, 节省了计算资源。CSA 模块的结构如图 6 所示。

CSA 模块分为 2 个分支, 分别对应通道注意力结构和空间注意力结构。为平滑特征, 采用 3×3 的卷积层对输入特征进行提取。为获取特征图通道注意力上的权重矩阵, 使用全局平均池化操作将特征的全局信息压缩到一维向量上, 并使用 2 个全连



(a) Car10序列



(b) Baginhand序列

图 5 特征图可视化

Fig. 5 Visualization of feature map

接层对向量进行非线性变换，从而得到通道上的特征向量。

$$\boldsymbol{f}' = \mathrm{Conv}(\boldsymbol{f}_{\mathrm{input}}) \tag{7}$$

$$\boldsymbol{V}_{\mathrm{channel}} = \mathrm{FC}_2(\mathrm{FC}_1(\mathrm{AVG}(\boldsymbol{f}'))) \tag{8}$$

式中：$\boldsymbol{f}_{\mathrm{input}}$为输入 CSA 模块的特征；$\boldsymbol{f}'$为提取的特征；FC表示全连接层；$\boldsymbol{V}_{\mathrm{channel}}$为输出的通道特征向量；AVG( ) 表示平均池化函数。

本文设计了一个简单的分支获取空间上的特征矩阵。采用 Max 函数获取每个像素点在其所有通道上的最大值，并利用 3×3 的卷积层提取其特征，从而生成一个单通道的特征图，该特征图反映了空间上每个像素点的贡献。

$$\boldsymbol{f}_{\mathrm{spatial}} = \mathrm{Conv}_{\mathrm{sp}}(\mathrm{Max}(\boldsymbol{f}')) \tag{9}$$

式中：$\mathrm{Conv}_{\mathrm{sp}}$表示提取空间特征的卷积层，卷积核大小为 3×3；$\boldsymbol{f}_{\mathrm{spatial}}$为空间特征图

将通道向量与空间特征图相乘，生成一个与 CSA 模块输入特征维度相同的特征矩阵，并且通过 Sigmoid 激活函数将其特征映射至 0～1 之间，即转化为权重特征图，从而对输入特征进行空间通道上的调制，具体公式如下：

$$\boldsymbol{f}_{\mathrm{out}} = \boldsymbol{f}_{\mathrm{input}} \cdot \mathrm{Sigmoid}(\boldsymbol{f}_{\mathrm{spatial}} \boldsymbol{V}_{\mathrm{channel}}) \tag{10}$$

结合通道注意力和空间注意力，对融合后的特征进行调制，使得网络在不同挑战的视频序列下自适应调整 2 种模态的贡献，实现鲁棒的 RGBT 跟踪。

1.5 网络训练细节

MBIENet 采用随机梯度下降法对网络进行训练。本文采用 GTOT[17] 和 RGBT234[24] 作为训练数据集，其中，在 GTOT[17] 上训练的网络模型用于测试 RGBT234[24] 和 LasHeR[25] 数据集，在 RGBT234[24] 上训练的网络模型用于测试 GTOT[17] 数据集。每次训练，网络从训练数据集视频序列中随机选取 8 帧作为一个批次，输入的可见光图像和热红外图像的大小先被调整为 107×107，高斯采样的包围框与真实包围框的交并比大于 0.7 的为正样本、小于 0.5 的为负样本，网络在目标周围提取 64 个正样本和 196 个负样本。训练分为 2 步进行：①对网络主干、FAM 模块、MBMF 模块及全连接层进行 200 次迭代训练，其中，全连接层的学习率为 10，其余模块学习率

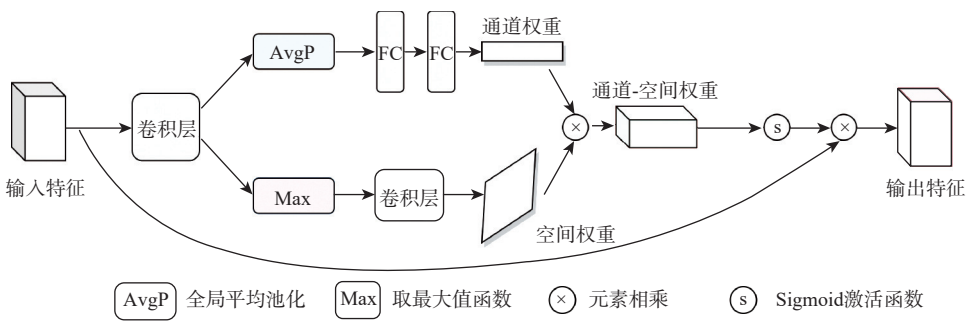


图 6 通道空间注意力模块结构

Fig. 6 Structure of channel-spatial attention module

均为 0.001; ②添加 CSA 模块, 进行 200 次迭代训练, 学习率设置与第①步一致。在训练过程中, 采用 Softmax 交叉熵损失函数对每个域(即不同的跟踪序列)进行二值分类(前景和背景), 损失函数如下:

$$L_{\mathrm{cls}}=-\frac{1}{N}\sum_{i=1}^{N}y_i\ln p_i+(1-y_i)\ln(1-p_i) \tag{11}$$

式中: N为训练阶段的样本个数; p_i为第i个样本的网络预测概率值; y_i为第i个样本的真实标签。

最终, 网络模型分别在 3 个 RGBT 跟踪数据集 GTOT[17]、RGBT234[24] 和 LasHeR[25] 上进行测试。

1.6　网络跟踪细节

RGBT 跟踪的典型过程为: 利用第 1 帧目标状态(即第 1 帧目标的包围框)对模型进行微调, 再在下一帧中定位目标的位置。在每个视频序列上, 网络冻结除全连接层(FC4 ~ FC6)外所有的模块, 并使用目标初始帧的状态对全接连层进行微调以适应目标外观的变化。具体来说, 网络从第 1 帧目标包围框周围高斯采样 500 个正样本和 5 000 个负样本作为训练样本, 利用其对全连接层进行 50 次迭代训练, 最后一个全连接层的学习率设置为 0.001, 其他层为 0.000 1。与 MDNet[14] 类似, 本文采用了短期和长期[14] 的原则来更新 3 个全连接层的参数, 并且利用边界框回归技术提高目标定位精度, 解决跟踪过程中目标尺度变化问题。为防止后续帧中潜在的不可靠性, 只在初始帧中训练一个边界框回归器, 并且只在跟踪成功的帧中使用它来调整目标状态。对于给定第t帧, 网络以$t-1$帧目标预测位置为中心高斯采样 256 个候选框, 将其输入到网络中得到预测得分, 选取得分最高的 5 个包围框, 将其平均位置作为当前帧的预测结果。重复上述过程, 直到整个视频序列结束, 完成对目标在给定视频序列中的定位。

2　实验评估

2.1　评测数据集与评测指标

1) GTOT[17] 数据集包括 50 对不同场景和条件下的 RGBT 视频序列, 共约 15×10^3 帧图像和 7 个视觉跟踪挑战属性。

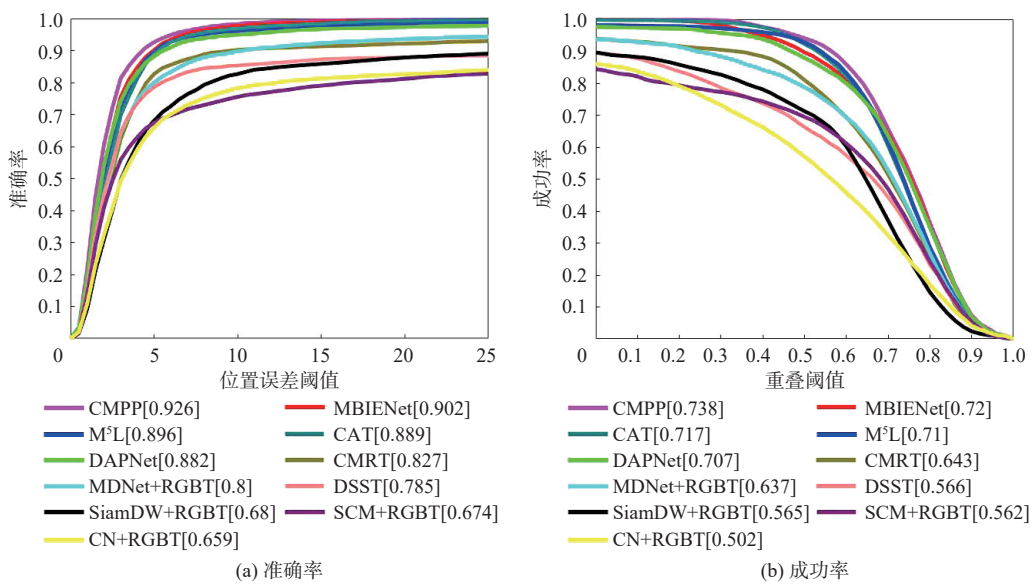
2) RGBT234[24] 是大型 RGBT 跟踪数据集, 包含 234 个高度对齐的 RGBT 视频对, 约 234×10^3 帧, 并提供了每个视频序列准确的注释, 充分考虑了各种环境下的挑战。RGBT234 数据集有 12 个属性注释, 用于不同跟踪算法对不同挑战的有效性分析。

3) LasHeR[25] 是目前较大的 RGBT 跟踪数据集, 其包含 1 224 个对齐的 RGBT 视频序列及更多样化的属性标注。其中, 245 个视频序列被单独划分为测试数据集, 其余视频序列被设计为训练数据集。

本文采用 RGBT234[24] 数据集中提出的评估协议, 采用准确率和成功率作为评价指标。其中, 准确率表示跟踪结果中心点与真实位置中心点距离小于阈值的所有帧的比例, 成功率衡量了成功跟踪的帧重叠大于阈值的比例。

2.2　GTOT 性能评估

如图 7 所示, 在 GTOT[17] 数据集上, 相较于 M^{5}L[26] 和 CAT[9] 跟踪器, 本文提出的跟踪器在准确率和成功率 2 个指标上分别有 0.006、0.01 和 0.013、0.003 的性能提升。相对的, 相比在 GTOT[17] 数据集上表现最佳的 CMPP[27] 跟踪器, 本文在跟踪性能上有所下降, 这是因为 CMPP[27] 算法采用大量的历史样本



(a) 准确率　　(b) 成功率

图 7　GTOT 评估曲线

Fig. 7　Evaluation curves of GTOT

作为时间信息来增强当前帧的特征表示，而本文所提网络只使用当前帧的信息。

2.3 RGBT234性能评估

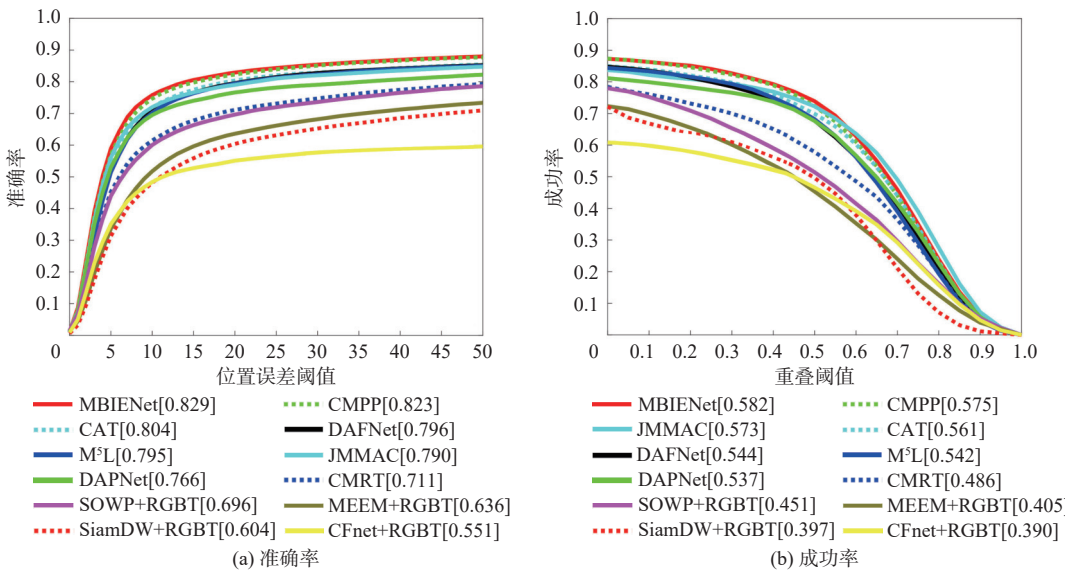
如图8所示，在RGBT234[24]数据集上，对比目前先进的RGBT跟踪器，本文提出的网络模型达到了最佳的跟踪性能。与高性能跟踪器CMPP[27]相比，本文提出的跟踪器在准确率和成功率2个指标上分别有0.006和0.007的性能优势。与CAT[9]相比，在准确率和成功率2个指标上则取得了0.025和0.021的较大的性能提升。此外，与DAFNet[8]和M^{5}L[26]相比，本文在准确率和成功率2个指标上分别有0.033、0.038和0.034、0.04的性能提升。这些结果充分说明了本文所提跟踪器的有效性。

2.4 LasHeR性能评估

本文采用LasHeR[25]单独划分为测试数据集的245个视频序列来评估跟踪器的性能。如图9所示，与先进的跟踪器相比，本文提出的跟踪器和目前LasHeR[25]上表现较好的MaCNet[28]在准确率和成功率2个指标上分别有0.002和0.005的性能提升。与CAT[9]相比，在准确率和成功率2个指标上分别有0.034和0.041的性能提升。与夺得VOT2019RGBT冠军的mfDiMP[29]相比，在准确率和成功率2个指标上分别有0.037和0.012的性能提升。

2.5 消融实验

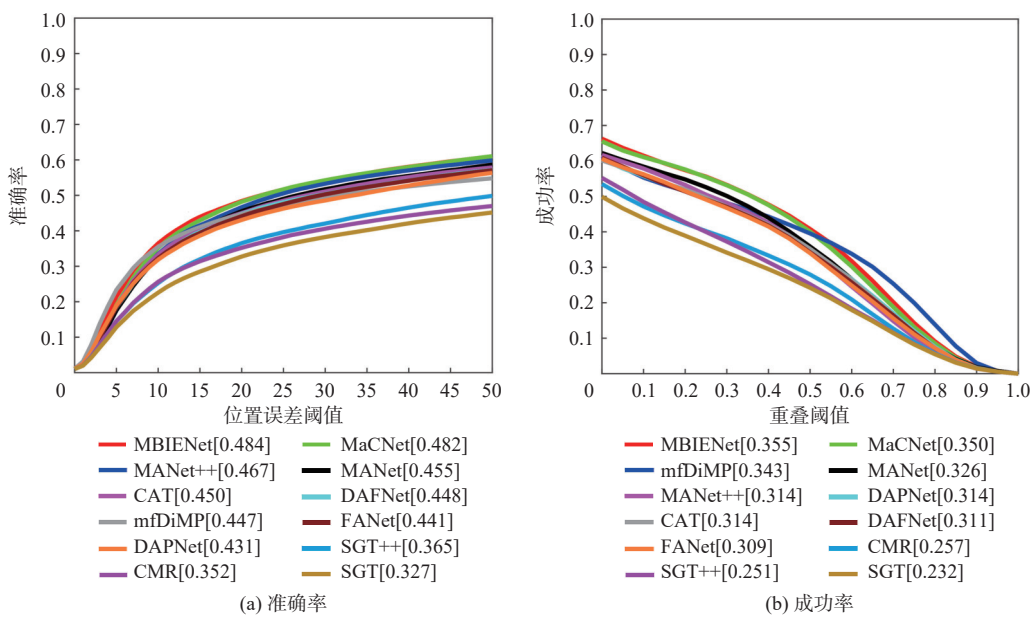
为证实本文所提模块的有效性，在测试数据集上做了消融实验。如表1所示，分别做了以下实



(a) 准确率　　(b) 成功率

图8　RGBT234评估曲线

Fig. 8　Evaluation curves of RGBT234



(a) 准确率　　(b) 成功率

图9　LasHeR评估曲线

Fig. 9　Evaluation curves of LasHeR

表 1　消融实验结果

Table 1　Ablation experiments results

模块	准确率			成功率		
	RGBT234	LasHeR	GTOT	RGBT234	LasHeR	GTOT
MBIENet-FAM	0.818	0.477	0.892	0.568	0.348	0.707
MBIENet-MBMF	0.813	0.471	0.886	0.560	0.342	0.701
MBIENet-CSA	0.822	0.479	0.899	0.574	0.347	0.718
MBIENet	0.829	0.484	0.902	0.582	0.355	0.720

验: ①MBIENet-FAM。去除 FAM 模块, 将模态共享特征、模态特定特征相加代替其输入 MBMF 模块做融合操作。②MBIENet-MBMF。去除 MBMF 模块, 拼接 FAM 模块的输出, 并使用一个卷积层提取模态拼接后的融合特征。③MBIENet-CSA。去除 CSA 模块, 将第 3 层卷积核提取的模态特征进行拼接, 输入到全连接层做二分类操作。

可以得知, 在 3 种不同的消融实验上, 模型性能都有显著下降, 从而证实了所提模块的有效性。

2.6　通道空间注意力模块有效性分析

为证明本文提出的 CSA 模块的有效性, 采用不同的注意力模块在 GTOT[17] 和 RGBT234[24] 数据集上进行对比实验, 注意力模块包括 SE[30]、CBAM[23] 和 CSA。实验结果如表 2 所示, 其中, Concat 表示在通道维度简单拼接来自不同模态的特征, 其余为在拼接模态特征的基础上利用不同的注意力机制对特征进行调制。值得注意的是, 为保证公平对比, 本文未采用 CBPNet[13] 类似的堆叠注意力机制 (即在每个卷积层都添加注意力) 的操作, 而是仅在第 3 个卷积层后使用一层注意力机制进行对比实验。

表 2　注意力机制比较

Table 2　Comparison of attention mechanisms

模块	准确率		成功率		检测速度/(帧·s^{-1})
	GTOT	RGBT234	GTOT	RGBT234	
MBIENet Concat	0.896	0.822	0.709	0.574	1.8
MBIENet CBAM	0.897	0.824	0.714	0.576	1.38
MBIENet SE	0.899	0.825	0.718	0.580	1.6
MBIENet CSA	0.902	0.829	0.720	0.582	1.54

由表 2 可知, 相较于简单拼接可见光和热红外图像的特征, 采用注意力机制更能在一定程度上提升网络的跟踪性能。相对的, 注意力机制则会为网络带来更大的计算资源消耗。SE[30] 通道注意力机制通过在单个空间下对高响应的通道进行增强来提高网络的性能, 实验结果说明了其有效性。CBAM[23] 通过串行通道-空间注意力机制对特征图进行增强, 但由于可见光和热红外图像存在空间分布上的成像差异, 单纯采用空间注意力仅能在单个模态内部选择空间上高响应的区域, 无法关注到 2 个模态整体的最优区域。相反, CSA 采用通道-空间注意力并行机制, 通过特征相乘操作将通道和空间注意力结合, 利用通道注意力选择 2 个模态中最优特征, 再与空间注意力相乘, 达到从整体选择出空间上最优区域的目的。实验结果显示, CSA 优于 CBAM[23]、SE[30] 注意力, 在 RGBT234[24]、GTOT[17] 数据集上的准确率和成功率分别提升了 0.007、0.008 和 0.006、0.011, 说明了本文所提 CSA 模块的有效性。

2.7　跟踪结果可视化对比

图 10 展示了不同序列下 MBIENet 和 CAT[9]、M^{5}L[26]、DAPNet[31] 的跟踪可视化结果对比。可知,

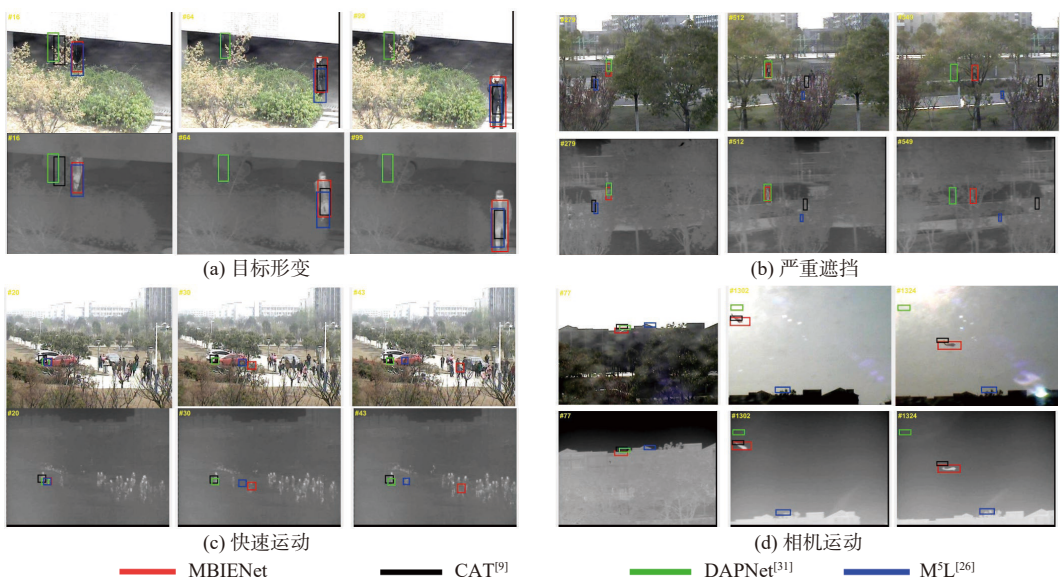


图 10　跟踪结果比较

Fig. 10　Tracking results comparison

在这些挑战视频序列上，相较于其他跟踪器，本文提出的MBIENet有着更加精准的跟踪结果，充分说明了本文提出的跟踪器在不同挑战下的有效性。

3 结 论

1）本文提出的多模态双向信息增强的RGBT跟踪网络可有效利用可见光图像与热红外图像的互补信息，进行可靠的RGBT视频目标定位。

2）充分考虑了模态共享特征与模态特定特征对目标外观表示的重要作用，并采用自适应的方式传递模态互补信息，以增强2个模态的特征表示。

3）在公开的3个数据集GTOT、RGBT234、LasHeR上，本文所提网络都达到了优秀的跟踪性能。同时，在不同挑战序列下，本文所提网络相较于CAT、M^{5}L等网络有着更加精确的目标定位能力。

在未来的工作中，将在模态互补方面继续研究RGBT目标跟踪，提高其效率和准确性。

参考文献（References）

[1] BERYINETTO L, VALMADRE J, HENRIQUES J, et al. Fully-convolutional Siamese networks for object tracking[C]//Proceedings of the European Conference on Computer Vision. Berlin: Springer, 2016: 850-865.

[2] GE S M, LUO Z, ZHANG C H, et al. Distilling channels for efficient deep tracking[J]. IEEE Transactions on Image Processing, 2020, 29: 2610-2621.

[3] ZHANG J M, JIN X K, SUN J, et al. Spatial and semantic convolutional features for robust visual object tracking[J]. Multimedia Tools and Applications, 2020, 79(21): 15095-15115.

[4] LIU Q, LU X H, HE Z Y, et al. Deep convolutional neural networks for thermal infrared object tracking[J]. Knowledge-Based Systems, 2017, 134: 189-198.

[5] ZHANG L C, GONZALEZ-GARCIA A, DANELLJAN M, et al. Synthetic data generation for end-to-end thermal infrared tracking [J]. IEEE Transactions on Image Processing, 2019, 28(4): 1837-1850.

[6] LIU Q, LI X, HE Z Y, et al. Learning deep multi-level similarity for thermal infrared object tracking[J]. IEEE Transactions on Multimedia, 2021, 23: 2114-2126.

[7] SUN T X, SHAO Y F, LI X N, et al. Learning sparse sharing architectures for multiple tasks[C]//Proceedings of the AAAI Conference on Artificial Intelligence. Palo Alto: AAAI Press, 2020, 34: 8936-8943.

[8] GAO Y, LI C L, ZHU Y B, et al. Deep adaptive fusion network for high performance RGBT tracking[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE Press, 2019: 91-99.

[9] LI C L, LIU L, LU A D, et al. Challenge-aware RGBT tracking [C]//Proceedings of the European Conference on Computer Vision. Berlin: Springer, 2020: 222-237.

[10] ZHANG P Y, WANG D, LU H C, et al. Learning adaptive attribute-driven representation for real-time RGB-T tracking[J]. International Journal of Computer Vision, 2021, 129(9): 2714-2729.

[11] XIAO Y, YANG M M, LI C L, et al. Attribute-based progressive fusion network for RGBT tracking[C]//Proceedings of the AAAI Conference on Artificial Intelligence. Palo Alto: AAAI Press, 2022: 2831-2838.

[12] LI C L, LU A D, ZHENG A H, et al. Multi-adapter RGBT tracking [C]//Proceedings of the IEEE/CVF International Comference on Computer Vision. Piscataway: IEEE Press, 2019: 2262-2270.

[13] XU Q, MEI Y M, LIU J P, et al. Multimodal cross-layer bilinear pooling for RGBT tracking[J]. IEEE Transactions on Multimedia, 2022, 24: 567-580.

[14] 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.

[15] 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.

[16] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2018.

[17] LI C L, CHENG H, HU S Y, et al. Learning collaborative sparse representation for grayscale-thermal tracking[J]. IEEE Transactions on Image Processing, 2016, 25(12): 5743-5756.

[18] LI C L, ZHAO N, LU Y J, et al. Weighted sparse representation regularized graph learning for RGB-T object tracking[C]//Proceedings of the 25th ACM International Conference on Multimedia. New York: ACM, 2017: 1856-1864.

[19] LI C L, SUN X, WANG X, et al. Grayscale-thermal object tracking via multitask Laplacian sparse representation[J]. IEEE Transactions on Systems, Man, and Cybernetics: Systems, 2017, 47(4): 673-681.

[20] ZHU Y B, LI C L, TANG J, et al. Quality-aware feature aggregation network for robust RGBT tracking[J]. IEEE Transactions on Intelligent Vehicles, 2020, 6(1): 121-130.

[21] PEREZ E, STRUB F, DE VRIES H, et al. FiLM: Visual reasoning with a general conditioning layer[C]//Proceedings of the AAAI Conference on Artificial Intelligence. Palo Alto: AAAI Press, 2018, 483: 3942-3951.

[22] WANG X, SHU X J, ZHANG S L, et al. MFGNet: Dynamic modality-aware filter generation for RGB-T tracking[J]. IEEE Transactions on Multimedia, 2023, 25: 4335-4348.

[23] 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.

[24] LI C L, LIANG X Y, LU Y J, et al. RGB-T object tracking: Benchmark and baseline[J]. Pattern Recognition, 2019, 96: 106977.

[25] LI C L, XUE W L, JIA Y Q, et al. LasHeR: A large-scale high-diversity benchmark for RGBT tracking[J]. IEEE Transactions on Image Processing, 2022, 31: 392-404.

[26] TU Z Z, LIN C, ZHAO W, et al. M^{5}L: Multi-modal multi-margin metric learning for RGBT tracking[J]. IEEE Transactions on Image Processing, 2022, 31: 85-98.

[27] WANG C Q, XU C Y, CUI Z, et al. Cross-modal pattern-propaga-

tion for RGB-T tracking[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2020: 7064-7073.

[28] ZHANG H, ZHANG L, ZHUO L, et al. Object tracking in RGB-T videos using modal-aware attention network and competitive learning[J]. Sensors, 2020, 20(2): 393.

[29] ZHANG L C, DANELLJAN M, GONZALEZ-GARCIA A, et al. Multi-modal fusion for end-to-end RGB-T tracking[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE Press, 2019: 2252-2261.

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

[31] ZHU Y B, LI C L, LUO B, et al. Dense feature aggregation and pruning for rgbt tracking[C]//Proceedings of the 27th ACM International Conference on Multimedia. New York: ACM, 2019: 465-472.

Multimodal bidirectional information enhancement network for RGBT tracking

ZHAO Wei[1, 2], LIU Lei[1, 2], WANG Kunpeng[1, 2], TU Zhengzheng[1, 2, *], LUO Bin[1, 2]

(1. Anhui Provincial Key Laboratory of Multimodal Cognitive Computation, Hefei 230601, China;
2. School of Computer Science and Technology, Anhui University, Hefei 230601, China)

Abstract: The goal of RGB-thermal infrared (RGBT) visual object tracking, which has drawn increasing interest in recent years, is to take advantage of the complimentary strengths of RGB and thermal infrared picture data to accomplish reliable visual tracking. For obtaining a robust appearance representation of an object, existing mainstream methods introduced the modal weight to fuse information of two modalities. Simply assigning weights to the individual modalities can't fully explore the complementary benefits of RGB and thermal infrared modalities. To solve these problems, propose a novel multimodal bidirectional information enhancement network for RGBT tracking (MBIENet). Specifically, design a feature aggregation module to aggregate modality-shared and modality-specific features for modeling the appearance information of an object. Further proposes a novel multimodal bidirectional modulation fusion module that can effectively fuse the complementary information of two modalities and alleviate the impact of redundant and useless features on the tracker. The contributions of various modalities in various situations are then adaptively adjusted using a lightweight channel-spatial attention module that is proposed. Experimental results on GTOT, RGBT234, and LasHeR datasets show that the accuracy rate and success rate of the proposed method are better than the existing mainstream trackers.

Keywords: RGB-thermal infrared; object tracking; deep learning; multimodal information fusion; multimodal information interaction

Received: 2022-05-20; **Accepted**: 2022-07-02; **Published Online**: 2022-10-18 17:41

URL: link.cnki.net/urlid/11.2625.V.20221018.0824.001

Foundation items: National Natural Science Foundation of China (62376005); Anhui Provincial Key Research and Development Program (202104d07020008,KJ2020A0033); Anhui Provincial Natural Science Foundation (2108085MF211); University Synergy Innovation Program of Anhui Province (GXXT-2022-014)

* **Corresponding author**. E-mail: zhengzhengahu@163.com

http://bhxb.buaa.edu.cn jbuaa@buaa.edu.cn
DOI: 10.13700/j.bh.1001-5965.2022.0397

基于 Transformer 和动态 3D 卷积的多源遥感图像分类

高峰，孟德森，解正源，亓林*，董军宇
(中国海洋大学 计算机科学与技术学院，青岛 266100)

摘 要：多源遥感数据具有互补性和协同性，近年来，基于深度学习的方法已经在多源遥感图像分类中取得了一定进展，但当前方法仍面临关键难题，如多源遥感图像特征表达不一致，融合困难，基于静态推理范式的神经网络缺乏对不同类别地物的适应性。为解决上述问题，提出了基于跨模态 Transformer 和多尺度动态 3D 卷积的多源遥感图像分类模型。为提高多源特征表达的一致性，设计了基于 Transformer 的融合模块，借助其强大的注意力建模能力挖掘高光谱和 LiDAR 数据特征之间的相互作用；为提高特征提取方法对不同地物类别的适应性，设计了多尺度动态 3D 卷积模块，将输入特征的多尺度信息融入卷积核的调制，提高卷积操作对不同地物的适应性。采用多源遥感数据集 Houston 和 Trento 对所提方法进行验证，实验结果表明：所提方法在 Houston 和 Trento 数据集上总体准确率分别达到 94.60% 和 98.21%，相比 MGA-MFN 等主流方法，总体准确率分别至少提升 0.97% 和 0.25%，验证了所提方法可有效提升多源遥感图像分类的准确率。

关 键 词：高光谱图像；激光雷达；Transformer；多源特征融合；动态卷积
中图分类号：TP753
文献标志码：A **文章编号**：1001-5965(2024)02-0606-09

近年来，随着对地观测平台、遥感传感器的日益丰富，针对同一地区的海量遥感图像(多光谱、高光谱、合成孔径雷达、激光雷达等)源源不断的产生，可以提供具有较强互补性的地物信息[1]。其中，高光谱图像 (hyperspectral image, HSI) 能够利用连续的光谱通道对地成像，光谱通道多达数十甚至数百个以上，能够探测到不同地物的丰富材质信息，已经被广泛应用于地物分类任务中。但是，高光谱图像面临着同物异谱、同谱异物或阴影遮挡等问题，影响地物分类的精度[2]。与高光谱图像不同，激光雷达(light detection and ranging, LiDAR)主动探测技术能够提供地物精确的高程信息及地球表面的三维空间信息，已经成为获取地形空间分布信息的有效方法，其获取的信息对高光谱图像来说是很好的补充。将 LiDAR 数据与高光谱图像联合用于地物分类能够有效提升分类精度。当前，如何综合利用同一地区的多源传感器数据提升地物分类、目标检测和分析的准确性，是遥感应用领域重要的研究方向之一。

针对高光谱和 LiDAR 联合地物分类，国内外学者提出了多种方法。Merentitis 等[3] 使用随机森林方法对多源遥感特征进行分类，缓解了特征维度过高的问题。Rasti 等[4] 提出了一种正交总变化分量分析方法对 2 类图像的信息进行融合。曹琼等[5] 提出了一种基于多级融合的多源遥感图像分类模型，利用 LiDAR 点云数据提取的建筑物掩膜对非建筑物部分进行分类，利用条件随机场模型对得到的分类结果进行优化处理。

随着深度学习技术的发展，卷积神经网络(convolutional neural network, CNN)在遥感图像分

收稿日期：2022-05-20；**录用日期**：2022-07-02；**网络出版时间**：2022-12-30 08:44
网络出版地址：link.cnki.net/urlid/11.2625.V.20221228.1616.001
基金项目：国家重点研发计划 (2018AAA0100602)；山东省自然科学基金 (ZR2019QD011)
* **通信作者**. E-mail：qilin2020@ouc.edu.cn

引用格式：高峰，孟德森，解正源，等. 基于 Transformer 和动态 3D 卷积的多源遥感图像分类[J]. 北京航空航天大学学报，2024，50(2)：606-614. GAO F，MENG D S，XIE Z Y，et al. Multi-source remote sensing image classification based on Transformer and dynamic 3D-convolution[J]. Journal of Beijing University of Aeronautics and Astronautics，2024，50(2)：606-614 (in Chinese).

类任务中体现出巨大的优势[6]。Hang 等[7] 提出了一种基于耦合 CNN 的方法, 使用 3 层卷积对 2 组图像分别进行特征提取, 将输出送入耦合融合模块, 通过决策级分类得到最终结果。Zhang 等[8] 提出一种基于互信息的注意力网络, 可有效增强多源特征的可用成分。Liao 等[9] 提出将多源遥感图像特征通过构建图模型来得到鲁棒的融合特征。Xue 等[10] 使用自校准卷积提取多尺度特征, 增大了 CNN 的感受野, 有效改善了多源遥感图像融合分类的效果。Zhang 等[11] 提出一种交叉感知神经网络, 增强高光谱和 LiDAR 数据间的特征交互, 有效提升多源遥感图像融合的效果。Zhao 等[12] 提出一种差分 Gabor 神经网络, 提取多源图像不同尺度、不同方向及不同语义的信息, 进一步提升了多源遥感图像融合分类的性能。

尽管当前多源遥感图像特征融合分类方法取得了非常好的性能, 但仍面临 2 个关键难题: ①多源遥感图像特征表达不一致, 融合困难。高光谱和 LiDAR 反映地物的特性存在差异, 提供的特征表达不统一, 当前方法难以充分挖掘多源图像特征间的关联关系。谷歌公司提出了 Vision Transformer(ViT)模型[13], 可有效挖掘图像像元之间的关联关系, 在完成图像分类任务时显著超越了主流的 CNN 模型, 借助 Transformer 强大的特征整合能力, 有望实现多源遥感图像特征的一致性表达。②遥感图像特征静态提取方式缺乏对不同类别地物的适应性。当前方法大多基于传统的静态推理模式, 对所有输入图像采用相同的网络结构和参数, 缺乏对不同类别地物的适应性, 一定程度上限制了神经网络的表征能力。近期, 出现了 CondConv[14]、DynamicConv[15] 等动态神经网络模型, 可以根据输入样本的不同自适应调节卷积核的参数。将动态网络的思路引入高光谱图像特征提取, 有望提升地物分类的性能。

为解决上述问题, 本文提出了基于 Transformer 特征融合的多源遥感图像分类模型(Transformer-based fusion network, TBFNet)。为提高多源特征表达的一致性, 设计了基于 Transformer 的融合模块, 借助其强大的注意力机制挖掘高光谱和 LiDAR 数据特征之间的相互作用, 从而提升地物分类效果。同时, 为提高特征提取对不同地物类别的适应性, 设计了多尺度动态 3D 卷积模块, 将输入特征的多尺度信息融入卷积核的调制, 提高卷积操作对不同地物的适应性。为验证模型的有效性, 在 Houston 和 Trento 2 个多源遥感数据集上进行实验, 结果表明, 本文模型在总体准确率和 Kappa 系数评价指标中均取得了最好的分类结果。

1　本文方法

1.1　总体框架

本文提出的 TBFNet 整体流程如图 1 所示。图中, 卷积层数字表示: (输出通道数, 卷积核尺寸, 填充值)。首先, 使用主成分分析法对高光谱图像进行预处理; 其次, 使用若干卷积层对高光谱和 LiDAR 数据进行特征提取, 其中, 对高光谱图像采用本文设计的多尺度动态 3D 卷积, 对 LiDAR 数据采用普通的 2D 卷积; 然后, 将获得的特征输入 Transformer 融合器, 得到融合特征; 最后, 将高光谱图像特征、LiDAR 数据特征、融合特征结合在一起用于分类。

1.2　图像预处理

高光谱图像包含多达数十甚至数百以上的光谱通道, 保存了地物大量的细节信息, 同时也存在大量冗余信息。因此, 利用主成分分析法对其进行通道降维(本文中将高光谱图像降维为 30 个通道), 可以更好地去除冗余, 保留主要信息。同时, 以每个像素点为中心选取尺寸大小为$k\times k$的图像块用于构建训练集和测试集。

1.3　多尺度动态 3D 卷积

由于 3D 卷积核能够同时对数据的光谱维度及空间维度进行处理, 在高光谱图像处理领域应用较广。本文对 3D 卷积进行了改进, 赋予其动态特性, 可以更有效地提高特征提取对不同地物的适应性。

CondConv[14] 是典型的 2D 动态卷积算法, 其主要思想是通过对输入进行变换计算得到卷积核参数。具体来说, 一个动态卷积层的卷积核是若干个专家知识(卷积核)的线性组合, 每个专家的加权系数由网络学习得到。受 CondConv 启发, 本文设计了多尺度动态 3D 卷积, 通过多尺度注意力模块计算卷积核权重, 该模块结构如图 2 所示。图中: AAP 为自适应均值池化。输入的高光谱特征为$\boldsymbol{X}\in\mathbf{R}^{N\times H\times W\times C}$, H、W为空间维度的尺寸, C为通道维度的尺寸, N为批处理大小。在计算注意力时, 首先, 将$\boldsymbol{X}$通过多尺度自适应均值池化得到1×1、2×2和4×4 3 个尺度的输出, 训练时可以保留输入特征的空间结构信息; 然后, 多尺度信息插值为相同尺寸并进行拼接, 输入类似 SENet[16] 的结构计算注意力权重; 最后, 得到形状为$N\times E$的矩阵, E为卷积核数量。通过上述操作, 完成了各个 3D 卷积核注意力分数的计算。

在神经网络推理过程中, 将得到的尺寸为$N\times E$的注意力分数矩阵与预先生成的卷积核权重

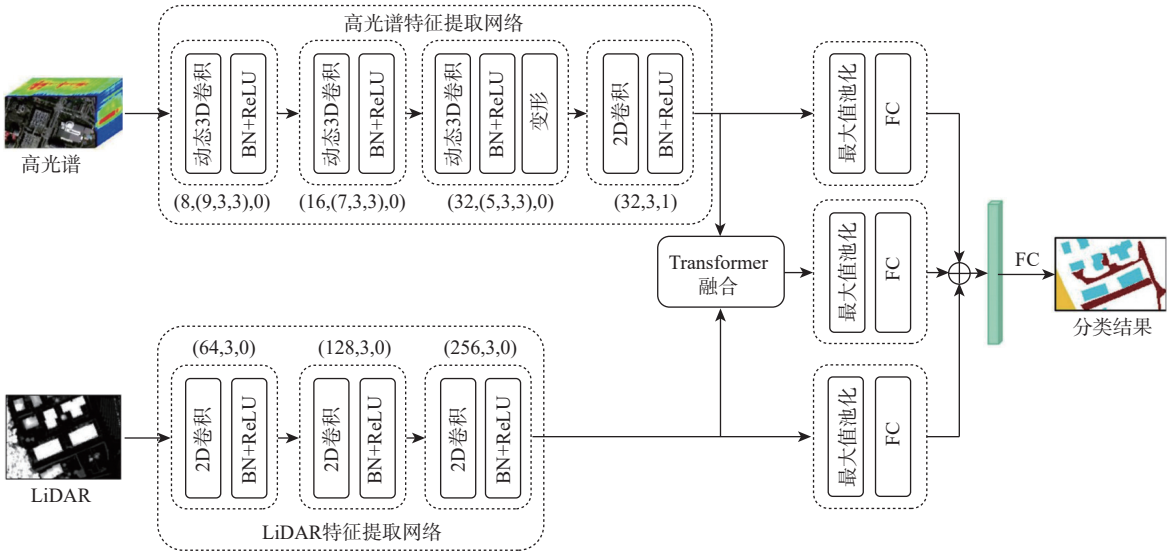


图 1 基于跨模态 Transformer 和多尺度动态 3D 卷积的多源遥感图像融合分类框架

Fig. 1 Multi-source remote sensing image fusion classification framework based on cross-modal Transformer and multi-scale dynamic 3D-convolution

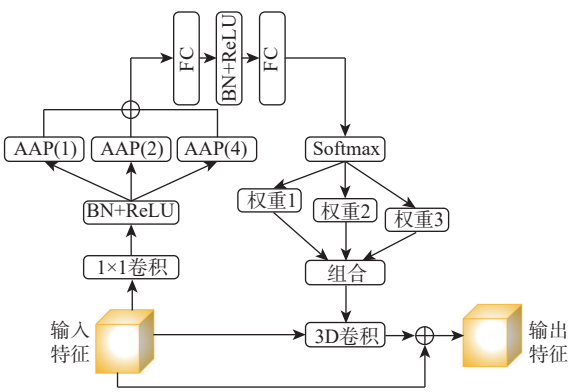


图 2 多尺度动态 3D 卷积模块结构

Fig. 2 Structure of multi-scale dynamic 3D-convolution module

矩阵进行矩阵乘法，动态生成 3D 卷积滤波器，再将其作用于输入得到最终的输出。

1.4 基于 Transformer 的融合模块

为利用高光谱图像和 LiDAR 数据的互补优势，充分挖掘多源特征间的关联关系，本文设计了基于 Transformer 的融合模块，总体结构如图 3 所示。

基于 Transformer 的融合模块的输入为高光谱图像特征 $\boldsymbol{X}\in\mathbf{R}^{H\times W\times C}$ 和 LiDAR 数据 $\boldsymbol{Y}\in\mathbf{R}^{H\times W\times C}$。首先，$\boldsymbol{X}$ 经过线性层处理得到特征 $\boldsymbol{Q}$，$\boldsymbol{Y}$ 经过线性层处理分别得到特征 $\boldsymbol{K}$ 和 $\boldsymbol{V}$。然后，三分支特征输入多头注意力模块进行融合，融合后的特征输入多层感知机（multilayer perceptron, MLP），再进行特征融合得到最终的特征 $\boldsymbol{U}$。需要指出的是，对于本文设计的 Transformer 融合器的输入特征 $\boldsymbol{X}$、$\boldsymbol{Y}$ 与进入多头注意力模块的 $\boldsymbol{Q}$、$\boldsymbol{K}$、$\boldsymbol{V}$ 输入特征密切相关，可通过实验探索输入的高光谱特征 $\boldsymbol{X}$、LiDAR 特征 $\boldsymbol{Y}$ 及 $\boldsymbol{Q}$、$\boldsymbol{K}$、$\boldsymbol{V}$ 的对应关系对分类结果的影响。经过实

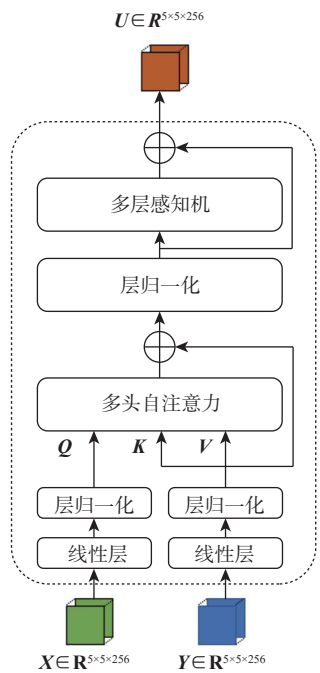


图 3 基于 Transformer 的融合模块结构

Fig. 3 Structure of fusion module based on Transformer

验对比，当 $\boldsymbol{Q}$ 来自于高光谱特征，$\boldsymbol{K}$ 和 $\boldsymbol{V}$ 来自于 LiDAR 特征时，分类模型在 2 个数据集上均能取得最好的分类结果。

缩放点乘注意力机制具有学习全局和局部特征的能力，且能有效捕捉特征元素之间的长距离依赖关系。将层归一化后的特征 $\boldsymbol{Q}$、$\boldsymbol{K}$、$\boldsymbol{V}$ 按照式（1）计算多源特征间的关联关系：

$$\boldsymbol{Z}=\mathrm{softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{d}}\right)\boldsymbol{V} \tag{1}$$

式中：d 为 $\boldsymbol{K}$ 的特征向量维度；$\sqrt{d}$ 为缩放因子。

本文采用基于多头计算的注意力机制, 在多个子空间中计算多源特征间的关联, 有效降低了计算复杂度, 同时提升了特征融合效率。首先, 采用h个注意力头, 将输入特征按通道维拆分为h个特征子集; 然后, 将各特征子集输入注意力计算模块; 最后, 所有注意力头的输出结果进行拼接, 得到特征$\boldsymbol{Z}$。将$\boldsymbol{Z}$输入线性层进行进一步加工, 同时, 利用残差连接缓解深度神经网络的退化问题。下面利用层归一化对计算结果进行归一化处理, 以加速收敛, 具体操作如下:

$$\boldsymbol{Z}' = \text{LayerNorm}(\boldsymbol{Z}\boldsymbol{W} + \boldsymbol{V}) \tag{2}$$

式中: $\boldsymbol{W}$为全连接层的权重参数。

$\boldsymbol{Z}'$输入含有一个隐层的多层感知机进行特征融合, 进行残差连接和归一化, 得到最终的融合特征$\boldsymbol{U}$, 具体如下:

$$\boldsymbol{U} = \text{FC2}(\text{ReLU}(\text{FC1}(\boldsymbol{Z}'))) + \boldsymbol{Z}' \tag{3}$$

式中: FC1表示第 1 个全连接层, FC2表示第 2 个全连接层。

基于 Transformer 的融合模块的主要思想是利用注意力机制来学习高光谱图像和 LiDAR 数据间的关联关系, 利用关系矩阵加权输入特征图的每个位置, 结合 2 种数据进一步捕捉不同地物的分布特征。将多模态特征分别输入到融合器中, 可同时进行模态内和模态间的信息融合, 并捕捉 2 种模态特征之间的潜在相互作用, 获得更好的融合特征进行分类。

2　实验结果与分析

2.1　实验平台与数据

本文实验使用平台具体为: Intel(R) Xeon(R) Platinum 8156 处理器, 配有 NVIDIA 2080Ti 显卡, 32 GB DDR4L 内存。实验使用 PyTorch 深度学习框架, 初始学习率设置为 0.04, 每 50 次迭代乘以 0.8 来减少学习率, 动量权重为 0.9, 训练时批处理大小设置为 128。

为验证本文方法的有效性, 使用 Houston 和 Trento 2 个数据集来验证模型用于多源遥感图像分类的性能。Houston 数据集由美国机载激光测绘中心获取, 其中, 高光谱图像在 380～1 050 nm 波长范围内有 144 个光谱波段, 数据集大小为 349 像素×1 905 像素, 包含 15 个地物类别, 对应的 LiDAR 数据和高光谱数据分辨率一致, 均为 2.5 m。Trento 数据集由特伦托大学科研人员在意大利南部的郊外区域采集而得, 其中, 高光谱图像包含 63 个光谱波段, 数据集大小为 166 像素×600 像素, 共有 6 个类别, 空间分辨率为 1 m。

2.2　实验参数分析

在多源遥感图像分类任务中, 需要先提取遥感数据中合适大小的图像块作为训练数据, 以保留每个像素点周围的空间邻域信息。本文中, 图像块的尺寸为k, 为探究k对模型性能的影响, 设置$k = 7, 9, 11, 13, 15$进行实验, 结果如图 4 所示。$k = 11$时, 本文模型在 2 个数据集上的总体准确率均达到了最大值。k值较小时, 图像块中包含的空间邻域信息也较少, 容易受到图像噪声的干扰; k值较大时, 空间邻域信息过多, 干扰了中心像素的特征, 容易影响分类的准确率, 并增加计算量。为平衡模型的性能和效率, 在后续实验中, 图像块大小设置为11×11。

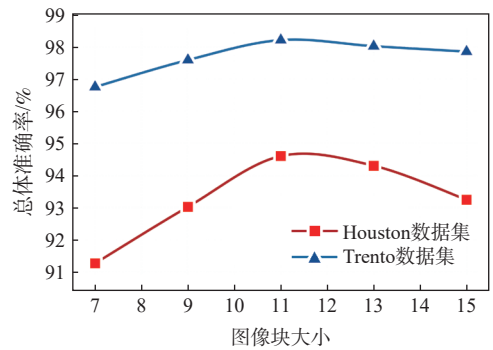


图 4　不同图像块大小与总体准确率的关系

Fig. 4　Relationship between different image block sizes and overall accuracy

2.3　消融实验

为验证基于 Transformer 的融合模块及多尺度动态 3D 卷积的有效性, 进行了消融实验, 以测试 2 个模块对分类准确率的提升效果。

在 Houston 和 Trento 数据集上的实验结果如表 1 所示。可以看出, 仅使用动态 3D 卷积和仅使用基于 Transformer 的融合模块相较于基准网络而言, 在 2 个数据集上的总体准确率都有较大的提升, 若同时使用动态 3D 卷积和基于 Transformer 的融合模块, 则可使分类精度进一步提升。这是因为本文提出的多尺度动态 3D 卷积可充分提取高光谱图像中的特征, 增强对不同地物类别的适应性; 基

表 1　基于 Transformer 的融合模块与多尺度动态 3D 卷积消融实验结果

Table 1　Ablation experiments results of Transformer-based fusion module and multi-scale dynamic 3D-convolution

方法	总体准确率/%	
	Houston数据集	Trento数据集
普通3D卷积	92.12	96.98
动态3D卷积	92.56	97.25
普通3D卷积+Transformer	93.43	97.89
动态3D卷积+Transformer	94.60	98.21

于Transformer的融合模块将来自高光谱和LiDAR 2种模态的特征进行有效融合，建立多源特征间的关联关系，显著提高了模型的分类准确率。

本文设计的Transformer融合器的输入特征$\boldsymbol{X}$和$\boldsymbol{Y}$与进入多头注意力模块的$\boldsymbol{Q}$、$\boldsymbol{K}$、$\boldsymbol{V}$ 3个输入特征密切相关，通过实验探究输入特征$\boldsymbol{X}$、$\boldsymbol{Y}$与多源特征对应关系对分类结果的影响。在2个数据集上的实验结果如表2所示。表中：输入特征$\boldsymbol{X}$和$\boldsymbol{Y}$分别代表不同遥感数据源的特征，输入特征为“无”表示未使用Transformer融合器。通过总体准确率的比较可知，当输入特征$\boldsymbol{X}$为高光谱特征、$\boldsymbol{Y}$为LiDAR特征时，分类模型在2个数据集上能取得最好的分类结果。这是因为相比于LiDAR数据，高光谱数据中拥有更加丰富的光谱信息，由其作为Query可以更好地将LiDAR数据中相关的特征进行关联，充分融合不同模态数据中的优势特征。后续实验中，都采用输入特征$\boldsymbol{X}$为高光谱特征、$\boldsymbol{Y}$为LiDAR特征的对应关系。

表2 Transformer融合器不同输入特征实验结果

Table 2 Experimental results of different input features for Transformer fusion crafting

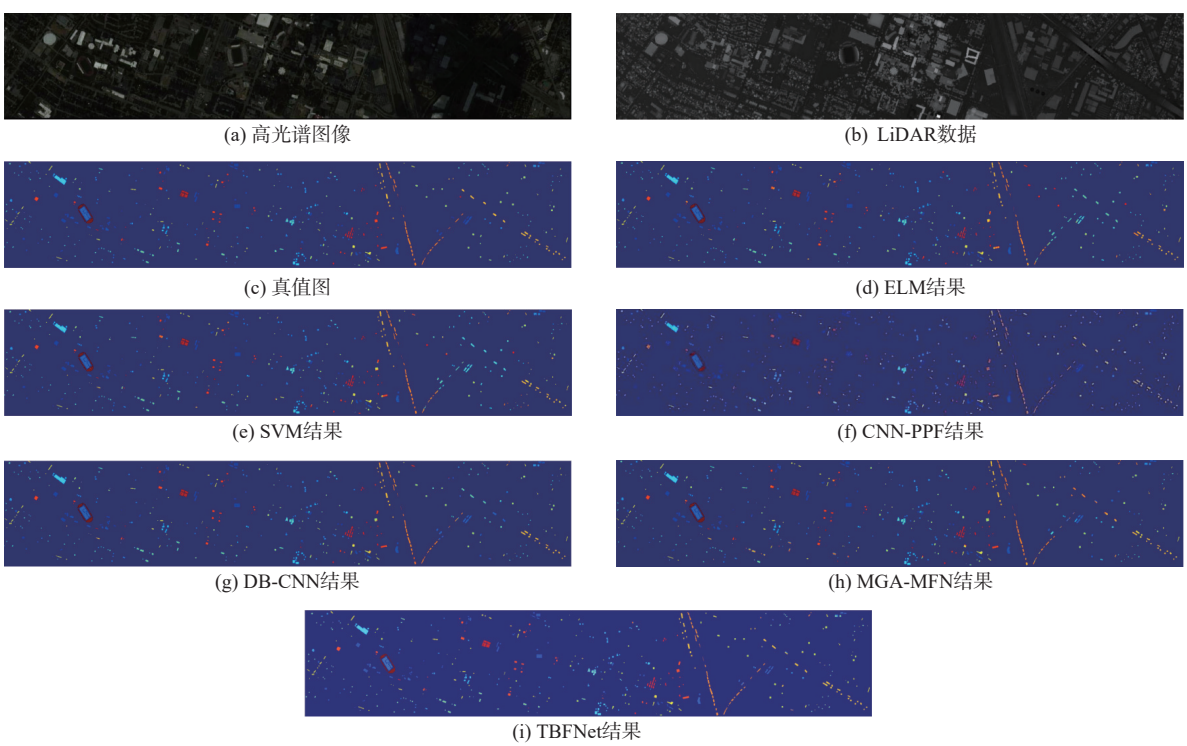
数据集	特征$\boldsymbol{X}$	特征$\boldsymbol{Y}$	总体准确率/%
Houston	无	无	90.20
	LiDAR	高光谱	92.56
	高光谱	LiDAR	94.60
Trento	无	无	97.25
	LiDAR	高光谱	98.06
	高光谱	LiDAR	98.21

2.4 对比实验分析

为验证本文提出的TBFNet的有效性，使用极限学习机(extreme learning machine, ELM)、支持向量机(support vector machine, SVM)、基于像素对特征的卷积神经网络(CNN with pixel-pair features, CNN-PPF)[17]、双分支卷积神经网络(dual branch CNN, DB-CNN)[18]、基于互指导注意力的多层次融合网络(mutual guidance attention-based multi-level fusion network, MGA-MFN)[8]5种分类方法进行对比实验。

为保证对比实验结果的公平性，所有分类模型均使用了相同的训练、测试样本集。针对实验结果，分别进行定性和定量分析，视觉上以图像的形式展示各分类方法的结果(见图5、图6)，同时以表格形式展示客观评价指标(见表3~表5)，包括总体准确率、平均准确率和Kappa系数。为排除实验中随机因素的影响，表格最终的实验结果取10次实验的平均值。

图5与图6分别展示了Houston和Trento数据集上不同分类方法的分类结果，表3~表5展示了不同方法在2个数据集上的3项客观指标结果。从可视化结果图中可以看出，本文模型在2个数据集上均取得了最好的分类效果。如表3和表4所示，在Houston数据集上，TBFNet在大多数类别地物的



(a) 高光谱图像 (b) LiDAR数据

(c) 真值图 (d) ELM结果

(e) SVM结果 (f) CNN-PPF结果

(g) DB-CNN结果 (h) MGA-MFN结果

(i) TBFNet结果

图5 Houston数据集上不同方法的分类结果

Fig. 5 Classification results of different classification methods on Houston dataset

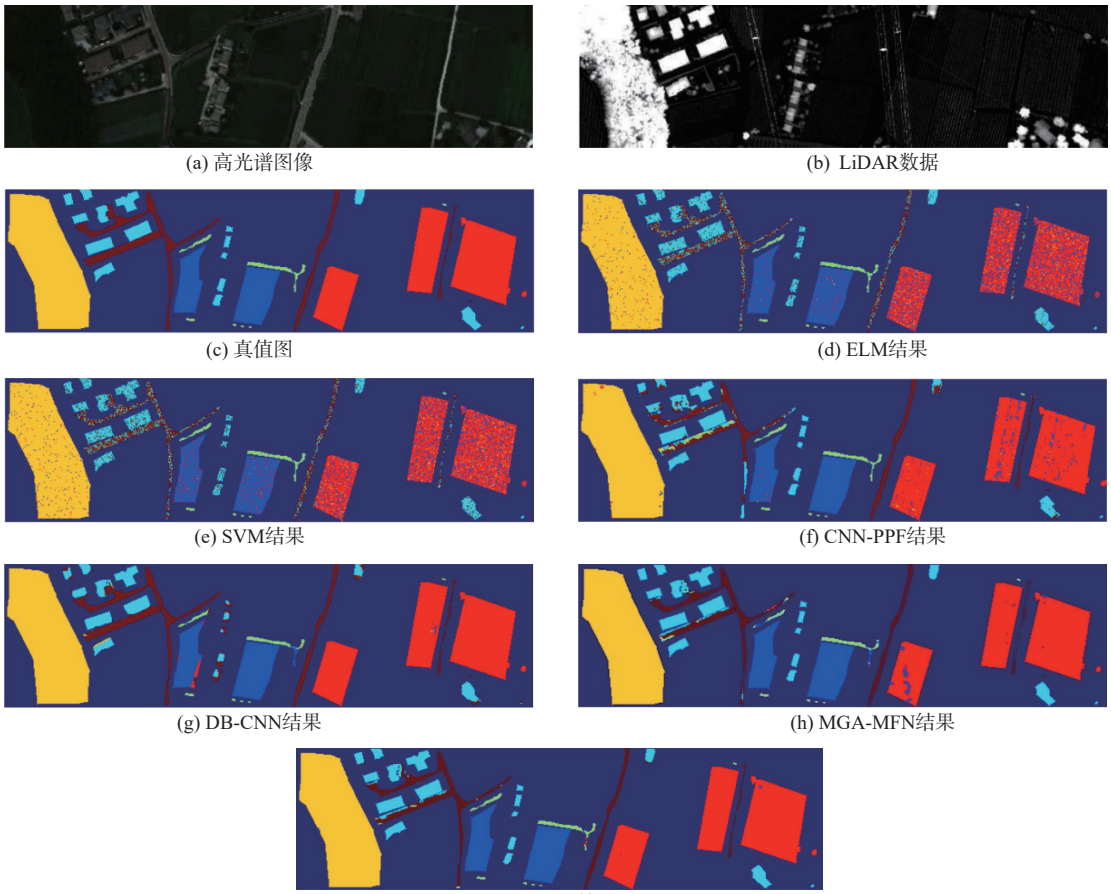


图 6　Trento 数据集上不同方法的分类结果

Fig. 6　Classification results of different classification methods on Trento dataset

分类准确率上表现更好, 并且在总体准确率、平均准确率、Kappa 系数 3 个指标上相比目前主流的分类方法至少分别提升 0.97%、0.39%、1.05%, 展现出良好的分类性能。在 Trento 数据集上, 如表 5 所示, TBFNet 对于 Woods、Vineyard、Road 样本的分类结果均取得了很好的效果, 由于 Ground 用于训练的样本较少, 且与 Road 具有相似性, 导致模型对 Ground 样本的分类结果较差。整体来看, TBFNet

表 3　Houston 数据集上不同分类方法的准确率

Table 3　Accuracy of different classification methods on Houston dataset　%

方法	Health grass (198/1 053)	Stressed grass (190/1 064)	Synthetic grass (192/505)	Trees (188/1 056)	Soil (186/1 056)	Water (182/143)	Residential (196/1 072)	Commercial (191/1 053)
ELM	83.10	83.70	100.0	91.86	98.86	95.10	80.04	68.47
SVM	82.43	82.05	99.80	92.80	98.48	95.10	75.47	46.91
CNN-PPF	83.57	98.21	98.42	97.73	96.50	97.20	85.82	56.51
DB-CNN	83.10	84.10	100.0	93.09	100.0	99.30	92.82	82.34
MGA-MFN	94.21	90.70	100.0	99.81	100.0	100.0	86.29	96.58
TBFNet	99.76	82.44	100.0	93.23	98.69	98.56	90.93	97.30

方法	Road (193/1 059)	Highway (191/1 036)	Railway (181/1 054)	Parking lot 1 (192/1 041)	Parking lot 2 (184/285)	Tennis court (181/247)	Running track (187/473)
ELM	84.80	49.13	80.27	79.06	71.58	99.60	98.52
SVM	77.53	60.04	81.02	85.49	75.09	100.0	98.31
CNN-PPF	71.20	57.12	80.55	62.82	63.86	100.0	98.10
DB-CNN	84.70	65.44	88.24	89.53	92.28	96.76	99.79
MGA-MFN	85.84	81.27	96.20	97.50	92.88	100.0	100.0
TBFNet	95.35	93.90	95.31	95.73	97.35	99.60	99.58

注: 类别后括号中数字表示训练样本数和测试样本数。

表4 Houston数据集上不同分类方法的性能比较

Table 4 Performance comparison of different classification methods on Houston dataset %

方法	总体准确率	平均准确率	Kappa系数
ELM	81.92	84.27	80.45
SVM	80.49	83.37	78.98
CNN-PPF	83.33	83.21	81.88
DB-CNN	87.98	90.11	86.98
MGA-MFN	93.63	94.71	93.08
TBFNet	94.60	95.10	94.13

在总体准确率和Kappa系数指标上均优于当前主流的多源遥感图像分类模型。这是因为TBFNet中的动态3D卷积可以有效提升模型的表达能力与特征提取能力，同时Transformer融合器能够捕获不同模态特征之间的关联关系，利用其内在的互补性进一步提升分类的准确率。2个数据集上的实验结果均显示出TBFNet取得了最优的分类效果，同时也具有较好的泛化能力。

为验证使用Transformer融合器的必要性，将其与拼接-卷积、分解双线性池化[19]特征融合方式进行对比，将主干网络中Transformer融合器分别替换为这2种融合方式进行实验。表6展示了不同特征融合方法在Houston和Trento数据集上的结果。

从表6中可以看出，本文设计的Transformer融合器在总体准确率上取得了最优的表现。相比于分解双线性池化融合方式，本文方法在参数量和乘法-累加运算次数2个指标上增加较少，在保证推理效率的同时，能够有效提升分类的总体准确率。

表5 Trento数据集上不同分类方法的性能比较

Table 5 Performance comparison of different classification methods on Trento dataset %

方法	准确率						总体准确率	平均准确率	Kappa系数
	Apple trees (129/3 905)	Buildings (125/2 778)	Ground (105/374)	Woods (154/8 969)	Vineyard (184/10 317)	Roads (122/3 052)			
ELM	95.81	96.97	96.66	99.39	82.24	86.52	91.32	92.93	90.42
SVM	88.62	94.04	93.53	98.90	88.96	91.75	92.77	92.63	95.85
CNN-PPF	90.11	83.34	71.13	99.04	99.37	89.73	94.76	88.97	93.04
DB-CNN	98.07	95.21	93.32	99.93	98.78	89.98	97.92	96.19	96.81
MGA-MFN	93.78	92.71	76.42	100	99.70	98.29	97.96	93.77	97.27
TBFNet	94.28	93.70	79.80	99.90	99.94	99.82	98.21	95.83	97.61

注：类别后括号中数字表示训练样本数和测试样本数。

表6 融合方式对比实验

Table 6 Comparative experiments of different fusion methods

特征融合方式	参数量/10^6	乘法-累加运算次数/10^6	总体准确率/%	
			Houston数据集	Trento数据集
Transformer融合	1.9	73.2	94.60	98.21
分解双线性池化	1.5	69.8	93.56	97.75
拼接-卷积	1.5	60.1	93.05	97.73

与简单的拼接-卷积融合方式相比，本文方法虽然带来了较多的参数量和运算次数，但分类的总体准确率显著提升。实验结果表明，本文设计的基于Transformer的特征融合方法能够有效提升地物分类的准确度，且能够保持较高的推理速度。

3 结 论

1）为解决多源遥感图像特征表达不一致问题，本文设计了基于Transformer的融合模块，借助其强大的注意力机制挖掘高光谱图像和LiDAR数据特征之间的关联关系，提高了多源特征表达的一致性。

2）为提高遥感图像特征提取对不同类别地物的适应性，本文提出多尺度动态3D卷积模块用于高光谱图像特征提取，将输入特征的多尺度信息融入卷积核的调制，显著改善了高光谱图像的特征提取效果，进而提高多源地物分类的准确率。

3）在Houston和Trento数据集上的实验结果表明，本文方法明显优于DB-CNN、MGA-MFN等主流方法；同时，也验证了本文方法在不同类别分布、复杂度的地物分类任务中具有较好的泛化与扩展能力。

参考文献（References）

[1] UEZATO T, FAUVEL M, DOBIGEON N. Hyperspectral image unmixing with LiDAR data-aided spatial regularization[J]. IEEE Transactions on Geoscience and Remote Sensing, 2018, 56(7): 4098-4108.

[2] WEI W, ZHANG J, ZHANG L, et al. Deep cube-pair network for hyperspectral imagery classification[J]. Remote Sensing, 2018, 10(5): 1-18.

[3] MERENTITIS A, DEBES C, HEREMANS R. Ensemble learning in hyperspectral image classification: Toward selecting a favorable

bias-variance tradeoff[J]. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2014, 7(4): 1089-1102.

[4] RASTI B, GHAMISI P, GLOAGUEN R. Hyperspectral and LiDAR fusion using extinction profiles and total variation component analysis[J]. IEEE Transactions on Geoscience and Remote Sensing, 2017, 55(7): 3997-4007.

[5] 曹琼, 马爱龙, 钟燕飞, 等. 高光谱-LiDAR 多级融合城区地表覆盖分类[J]. 遥感学报, 2019, 23(5): 892-903.
CAO Q, MA A L, ZHONG Y F, et al. Urban classification by multi-feature fusion of hyperspectral image and LiDAR data[J]. Journal of Remote Sensing, 2019, 23(5): 892-903(in Chinese).

[6] GE C, DU Q, LI W, et al. Hyperspectral and LiDAR data classification using kernel collaborative representation based residual fusion [J]. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2019, 12(6): 1963-1973.

[7] HANG R, LI Z, GHAMISI P, et al. Classification of hyperspectral and LiDAR data using coupled CNNs[J]. IEEE Transactions on Geoscience and Remote Sensing, 2020, 58(7): 4939-4950.

[8] ZHANG T, XIAO S, DONG W, et al. A mutual guidance attention-based multi-level fusion network for hyperspectral and LiDAR classification[J]. IEEE Geoscience and Remote Sensing Letters, 2022, 19: 1-5.

[9] LIAO W, PIŽURICA A, BELLENS R, et al. Generalized graph-based fusion of hyperspectral and LiDAR data using morphological features[J]. IEEE Geoscience and Remote Sensing Letters, 2015, 12(3): 552-556.

[10] XUE Z, YU X, TAN X, et al. Multiscale deep learning network with self-calibrated convolution for hyperspectral and LiDAR data collaborative classification[J]. IEEE Transactions on Geoscience and Remote Sensing, 2022, 60: 1-16.

[11] ZHANG M, LI W, TAO R, et al. Information fusion for classification of hyperspectral and LiDAR data using IP-CNN[J]. IEEE Transactions on Geoscience and Remote Sensing, 2022, 60: 1-12.

[12] ZHAO X, TAO R, LI W, et al. Fractional Gabor convolutional network for multisource remote sensing data classification[J]. IEEE Transactions on Geoscience and Remote Sensing, 2022, 60: 1-18.

[13] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, et al. An image is worth 16×16 words: Transformers for image recognition at scale[C]//Proceedings of the 9th Interbational Conference on Learning Representations. Schloss Dagstuhl: ICLR, 2021: 1-22.

[14] YANG B, BENDER G, LE Q V, et al. CondConv: Conditionally parameterized convolutions for efficient inference[C]//Proceedings of the 32nd International Conference on Neural Infomation Processing Systems. New York: ACM, 2019: 1-15.

[15] CHEN Y, DAI X, LIU M, et al. Dynamic convolution: Attention over convolution kernels[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2020: 11027-11036.

[16] HU J, SHEN L, ALBANIE S, et al. Squeeze-and-excitation networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020, 42(8): 2011-2023.

[17] LI W, WU G, ZHANG F, et al. Hyperspectral image classification using deep pixel-pair features[J]. IEEE Transactions on Geoscience and Remote Sensing, 2017, 55(2): 844-853.

[18] XU X, 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.

[19] YU Z, YU J, FAN J, et al. Multi-modal factorized bilinear pooling with co-attention learning for visual question answering[C]//Proceedings of the IEEE International Conference on Computer Vision. Piscataway: IEEE Press, 2017: 1821-1830.

Multi-source remote sensing image classification based on Transformer and dynamic 3D-convolution

GAO Feng，MENG Desen，XIE Zhengyuan，QI Lin*，DONG Junyu

(School of Computer Science and Technology，Ocean University of China，Qingdao 266100，China)

Abstract：Benefited from the complementarity and synergy of multi-source remote sensing data, deep learning-based methods have made significant progress in remote sensing image classification in recent years. Building a powerful multi-source data joint classification model is typically difficult for the following reasons: the feature fusion is hampered by the heterogeneous gap between HSI and LiDAR data; the representation power, efficiency, and interpretability are constrained by the current static inference paradigm.To solve both problems, we propose a Transformer-based fusion network. Specifically, to bridge the heterogeneous gap between HSI and LiDAR data, we design a feature fusion module based on Transformer to exploit the feature interactions between multi-source data. After that, we create a multi-scale dynamic 3D-convolution module to collect the information from different scales and use it to modulate the 3D-convolution kernel. The method was validated with Houston and Trento datasets. The overall accuracy of the proposed method reached 94.60% and 98.21% respectively. Compared with mainstream methods such as MGA-MFN, the overall accuracy of the two datasets was improved by at least 0.97% and 0.25% respectively. The experimental results demonstrate that our method can effectively improve the accuracy of multi-source remote sensing image classification.

Keywords：hyperspectral image；light detection and ranging；Transformer；multi-source feature fusion；dynamic convolution

Received：2022-05-20；**Accepted**：2022-07-02；**Published Online**：2022-12-30 08：44

URL：link.cnki.net/urlid/11.2625.V.20221228.1616.001

Foundation items：National Key R & D Program of China (2018AAA0100602); Shandong Provincial Natural Science Foundation (ZR2019QD011)

* **Corresponding author**. E-mail：qilin2020@ouc.edu.cn

http://bhxb.buaa.edu.cn jbuaa@buaa.edu.cn
DOI: 10.13700/j.bh.1001-5965.2022.0402

自注意力相似度迁移跨模态哈希网络

梁焕，王海荣*，王栋
(北方民族大学 计算机科学与工程学院，银川 750021)

摘　　要：为进一步提升跨模态检索性能，提出自注意力相似度迁移跨模态哈希网络模型。设计了一种通道空间混合自注意力机制强化关注图像的关键信息，并使用共同注意力方法加强模态信息交互，提高特征学习质量；为在哈希空间重构相似关系，采用迁移学习的方法利用实值空间相似度引导哈希码的生成。在3个常用的数据集MIRFLICKR-25K、IAPR TC-12和MSCOCO上与深度跨模态哈希（DCMH）、成对关系引导的深度哈希（PRDH）、跨模态汉明哈希（CMHH）等优秀方法进行对比实验，结果显示哈希码长度为64 bit的条件下，所提模型在3个数据集图像检索文本任务的平均精确度均值（MAP）达到72.3%，文本检索图像任务的MAP达到70%，高于对比方法。

关 键 词：跨模态检索；哈希学习；注意力机制；迁移学习；无监督学习
中图分类号：V221+.3; TB553
文献标志码：A　　**文章编号**：1001-5965(2024)02-0615-08

随着人工智能技术及自媒体的蓬勃发展，信息呈现多模态的形势，如文本、图像、视频、音频及3D模型等。海量多模态数据包含语义关联，跨模态检索日益成为研究者关注的热点问题。

跨模态检索处理过程一般是对多模态数据进行特征提取；建立模态间关联机制；对数据进行相似度度量及排序等操作。模态间关联机制的作用是获得多模态数据的公共表示[1]，据此跨模态研究可分为将多模态数据转化为实值公共表示的子空间方法和将多模态数据特征哈希变换为二值公共表示的哈希方法。

跨模态哈希方法在降低存储空间，减少检索耗费时间方面表现优越，取得了良好的检索效果。

跨模态哈希方法可进一步分为有监督哈希方法[2]与无监督哈希方法。

有监督哈希方法是利用已有标签提取语义信息，融入哈希码来提高检索精度[3]。例如，文献[4]为缩小异构数据间的异质性，提出自监督对抗哈希网络，利用多标签标注挖掘丰富的语义信息，使原始相关数据在哈希空间内保持模态相关性；文献[5]通过双重注意力机制提升模态特征质量，并利用标签语义监督弥合多模态数据的语义鸿沟，提升跨模态检索精度。注意力机制是机器学习模型中嵌入的一种特殊结构，用来自动学习和计算输入数据对输出数据的贡献大小[6-9]，文献[10]通过计算共有标签的比例细化相似度，并使用混合注意力模型提高图像关键信息的关注度。

无监督哈希方法侧重于模态内和模态间的关系学习，利用未标记数据训练样本挖掘数据的内在性质和规律，并将学习的多模态特征转换为二进制哈希码[11-12]。文献[13]提出一个无监督耦合循环对抗网络，外层循环对抗网络主要是使不同模态提取更有代表性的公共特征向量，内层循环对抗网络学习高质量的哈希编码；为充分挖掘模态间关联信息，对语义信息的重要性做出判断，文献[14]提出一种基于张量正则多态分解理论的空间汇集网络模型，辅助骨干网络捕获更重要的语义信息，利用全局协方差池化技术获取模态语义信息，并设计一

收稿日期：2022-05-21；**录用日期**：2022-07-02；**网络出版时间**：2022-10-21 13:24
网络出版地址：link.cnki.net/urlid/11.2625.V.20221021.0943.001
基金项目：宁夏回族自治区教育厅高等学校科学研究重点项目(NYG2022051)；宁夏自然科学基金项目(2023AAC03316)
*通信作者. E-mail：bmdwhr@163.com

引用格式：梁焕，王海荣，王栋. 自注意力相似度迁移跨模态哈希网络[J]. 北京航空航天大学学报，2024，50(2)：615-622.
LIANG H，WANG H R，WANG D. Cross-modal hashing network based on self-attention similarity transfer[J]. Journal of Beijing University of Aeronautics and Astronautics，2024，50(2)：615-622 (in Chinese).

种新的重构损失来优化模型参数；文献 [15] 采用迁移学习的方法，在哈希空间重构实值相似度值，充分保留各模态的特征信息，学习同时保留多模态数据的共享属性和单模态数据特定属性的哈希码。文献 [16] 采用一种联合矩阵分解学习统一的哈希码，将所有的模态联合分解到一个共享的潜在语义空间，并整合统一的哈希码和单独模态的哈希码，得到最终的哈希码。

综上所述，跨模态哈希检索研究已取得重大进展，但现有方法在分析模态信息重要程度方面的研究较少，为此，本文提出自注意力相似度迁移跨模态哈希网络模型，设计了通道空间混合注意力机制加强关注图像关键信息，使用多头注意力增强对文本关键信息的关注，并采用共同注意力机制加强模态信息交互。另外，大多数无监督哈希方法在重构相似矩阵时需要在实值空间构建相似矩阵，会产生大量参数，本文使用迁移学习的方法，利用实值空间不同实例相似度数值引导汉明码的生成，保留相似关系的同时降低了参数训练的复杂度。

1 本文模型

本文模型如图 1 所示，包含特征处理和哈希学习 2 部分。

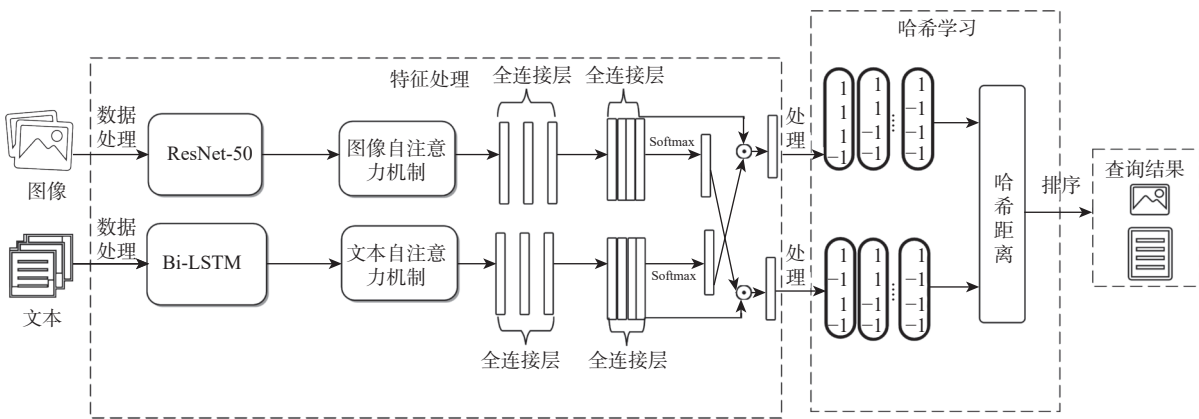


图 1 本文模型

Fig. 1 The proposed model

对于图像模态数据使用残差网络（residual network, ResNet）进行特征提取，设计并使用了一种通道空间注意力分辨图像信息的关键性，文本数据则利用双向长短期记忆（Bi-directional long short-term memory, Bi-LSTM）网络提取特征，多头注意力机制计算单词间的相互影响。为获得统一的表示将处理过的特征传入参数共享的全连接层，并使用共同注意力机制加强模态信息交互。同时，利用实值空间数据成对相似度值引导汉明码的生成，有效避免重构矩阵时产生的大量参数，提升模型训练速率。

1.1 特征处理

特征提取分为图像和文本 2 部分，使用 ResNet-50 作为图像特征提取网络的基础模型，考虑到深度神经网络在训练过程中会产生大量冗余信息和消耗大量训练时间，使用其预训练网络来替代。

为了加强对图像重点区域的关注，采用注意力机制对提取到的特征做进一步处理。通道注意力[17-18]通过探索特征图中不同通道之间的相关性来获取每个通道的相对重要性；空间注意力[19-20]旨在探索不同区域的相对重要性。为了同时获得通道注意力和空间注意力的优势[21-22]，结合现有的注意力机制，设计了混合注意力模型，网络结构如图 2 所示。

对于给定的特征图 $\boldsymbol{F}_{\text{map}}=\boldsymbol{R}^{C\times H\times W}$，其中 C、H、W 分别为特征图的通道数、高度、宽度，$\boldsymbol{R}$ 为特征图矩阵。在通道注意力子分支，首先，进行全局平均池化，然后，经过全连接层-激活函数（Rule）-全连接层-归一化层（batch normalization, BN）处理，将原来的输出通道每个通道用对应的权重进行加权，得到加权后的通道注意力特征掩码 $\boldsymbol{M}_{\text{c}}$，表达式为

$$\boldsymbol{M}_{\text{c}}=b(l(\text{Rule}\left(l\left(g\left(\boldsymbol{F}_{\text{map}}\right)\right)\right)) \tag{1}$$

式中：b 表示 BN 层；l 表示全连接层；g 表示全局平均池化层。

对于空间子分支，将原始特征图经过 1×1×1 卷积层组合成单通道的特征融合图，得到 $H\times W$ 的矩阵 $\boldsymbol{F}_0$，避免通道信息对空间注意力权重的影响；将 $\boldsymbol{F}_0$ 经 Softmax 函数计算获得全局相关性矩阵 $\boldsymbol{F}_1$，$\boldsymbol{F}_1$ 和 $\boldsymbol{F}_0$ 对应元素相乘，获得全局特征矩阵，输入到 BN 获得注意力掩码，表达式如下：

$$\boldsymbol{F}_0=\text{Conv}_{1\times1\times1}\left(\boldsymbol{F}_{\text{map}}\right) \tag{2}$$

$$\boldsymbol{F}_1=\text{Softmax}(\boldsymbol{F}_0) \tag{3}$$

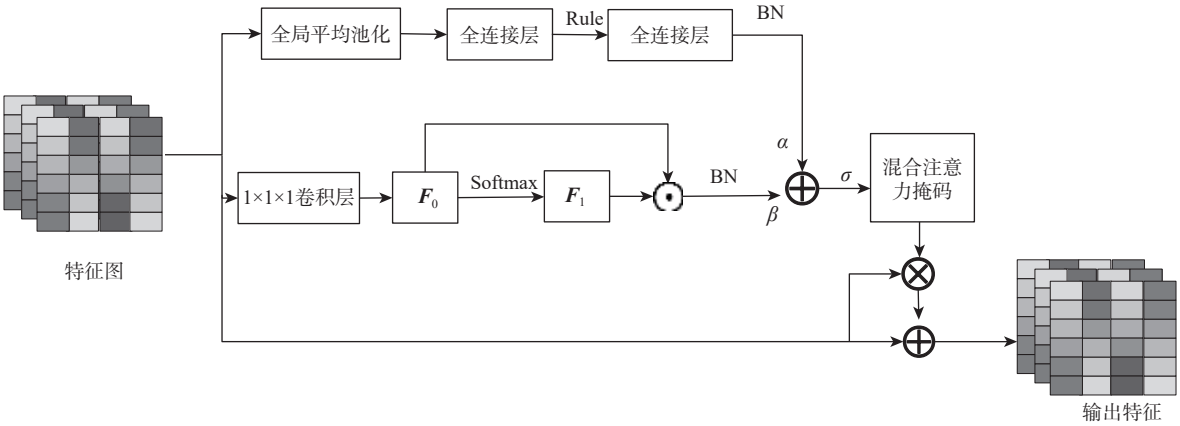


图 2　混合注意力机制

Fig. 2　Hybrid attention mechanism

$$\boldsymbol{M}_{\mathrm{s}} = b(\boldsymbol{F}_0 \times \boldsymbol{F}_1) \tag{4}$$

其中, $\mathrm{Conv}_{1\times1\times1}$ 表示 1×1×1 卷积。将通道注意力和空间注意力按比例分配得到混合注意力掩码, 与原始特征图结合得到最终特征 $\boldsymbol{F}$, 表达式为

$$\boldsymbol{F} = \boldsymbol{F}_{\mathrm{map}}(1 + \sigma(a \times \boldsymbol{M}_{\mathrm{c}} + \beta \times \boldsymbol{M}_{\mathrm{s}})) \tag{5}$$

式中: α 和 β 为学习得到的可变权重; σ 为 Sigmoid 函数。

文本模态数据采用 Bi-LSTM 作为提取特征的神经网络, 并使用多头注意力加强词与词之间的关联, 获得更精确的语义表达。多头注意力机制将每个注意力结果拼接并进行线性转换, 使得输入和输出结构对标, 由于每个注意力机制关注的特征层减少, 多头的计算代价和单头相似。具体实现过程为, 将文本输出特征变换为查询值 $\boldsymbol{Q}$、键值 $\boldsymbol{K}$、值向量 $\boldsymbol{V}$, 经过线性变换后进行多次缩放点积计算, 头与头之间参数不共享, 将多次的缩放点积计算结果进行拼接, 再进行线性变换得到的值就是多头注意力计算结果, 计算式为

$$\mathrm{Attention}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}) = \mathrm{Softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\psi}\right) \tag{6}$$

式中: ψ 为缩放因子。

多头注意力模型计算过程如图 3 所示, 其中, h 为 h 次缩放点积计算。

为获得统一的表示, 将注意力机制计算得到的图像和文本特征经过参数共享的全连接层, 表示为 512 维的特征向量 $\boldsymbol{F}_{\mathrm{p}}$ 和 $\boldsymbol{F}_{\mathrm{t}}$。为减少图像和文本通道之间的信息不平衡, 加强模态信息交互, 采用共同注意机制, 使用其中一种模态的学习特征用于指导另一种模态的特征训练。分别对图像特征 $\boldsymbol{F}_{\mathrm{p}}$ 和文本特征 $\boldsymbol{F}_{\mathrm{t}}$ 使用 Softmax 函数, 得到注意力分布 $\boldsymbol{A}_{\mathrm{p}}$=Softmax($\boldsymbol{F}_{\mathrm{p}}$) 和 $\boldsymbol{A}_{\mathrm{t}}$=Softmax($\boldsymbol{F}_{\mathrm{t}}$), 通过交替相乘不同模态的注意力分布和特征值得到新的图像特征 $\boldsymbol{W}_{\mathrm{p}} = \boldsymbol{F}_{\mathrm{p}} \circ \boldsymbol{A}_{\mathrm{t}}$ 和文本特征 $\boldsymbol{W}_{\mathrm{t}} = \boldsymbol{F}_{\mathrm{t}} \circ \boldsymbol{A}_{\mathrm{p}}$, 其中, ∘表示

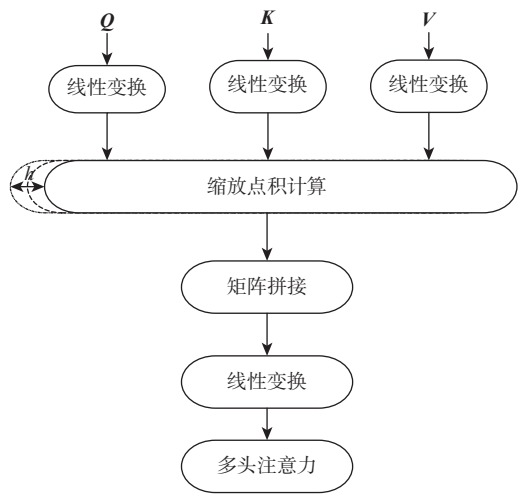


图 3　多头注意力机制

Fig. 3　Multi-headed self-attention

Hadamard 乘积运算, 并将其经过全连接层的处理, 得到 k 维的特征向量, k 即为哈希码的长度。

1.2　哈希学习

跨模态哈希算法需要在哈希空间重构相似关系, 现有方法大多需要在实值空间构建实值相似矩阵, 再对哈希空间的哈希相似矩阵进行优化, 在构建相似关系矩阵时会产生大量参数, 增加训练负担。因此, 实验采用成对相似度迁移的方法利用实值空间内相似值引导哈希码的生成, 如图 4 所示。

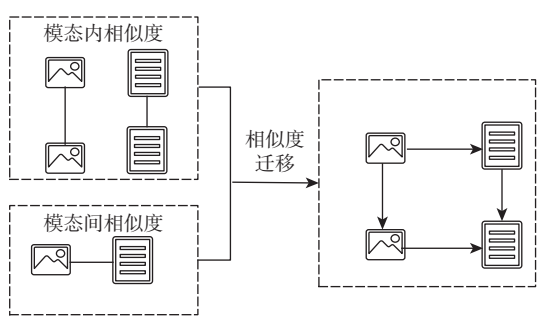


图 4　相似度迁移

Fig. 4　Similarity transfer

相似度迁移包括模态内相似度迁移和模态间相似度迁移2部分。对于模态内相似度，在实值空间内使用余弦距离$S=\cos(\boldsymbol{L}_i,\boldsymbol{L}_j)$表示实例相似度，$\boldsymbol{L}_i$为第$i$个实例的实值向量，$\boldsymbol{L}_j$为第$j$个实例的实值向量。哈希空间内用$D$表示哈希空间内的相似度，计算式为

$$D=\frac{\boldsymbol{B}_i\times\boldsymbol{B}_j^{\mathrm{T}}}{\sqrt{\boldsymbol{B}_i^2}\times\sqrt{\boldsymbol{B}_j^2}^{\mathrm{T}}} \tag{7}$$

式中：$\boldsymbol{B}_i$为第i个实例的哈希向量。由于离散的哈希码不利于神经网络的反向传播，采用式(8)作为哈希函数：

$$\boldsymbol{B}_v=\tanh(\rho\boldsymbol{W}_v)\in[-1,1]^{n\times k} \tag{8}$$

$\boldsymbol{W}_v$为经过特征处理的模态特征矩阵，v表示数据所属模态，在参数ρ趋于无限大时，tanh函数输出逐渐逼近函数sign的输出，从而使反向传播更容易实现。

根据不同模态在实值空间内的数据分布，采用相似度重建的方式保持相似关系，将各模态内成对的相似度值迁移至哈希空间，模态内相似度目标函数表示为

$$S_{\mathrm{intra}}=\min\omega|S_{\mathrm{p}}-D_{\mathrm{p}}|_F^2+\gamma|S_{\mathrm{t}}-D_{\mathrm{t}}|_F^2 \tag{9}$$

式中：S_{p}和D_{p}分别为公共实值空间内和哈希空间内图像实例的相似度；S_{t}和D_{t}分别为公共实值空间内和哈希空间内文本实例的相似度；ω、γ为网络参数。

模态间相似度迁移可利用成对多模态信息进行计算，由于相同模态的同语义实例的特征关系矩阵对角线为1[23]，而不同模态存在语义鸿沟，所以在此约束模态间最大相似度为1，模态间相似度迁移损失函数为

$$S_{\mathrm{inter}}=\min\|\boldsymbol{E}-\mathrm{diag}\left(D\left(\boldsymbol{B}_{\mathrm{p}},\boldsymbol{B}_{\mathrm{t}}\right)\right)\|_F^2 \tag{10}$$

式中：$\boldsymbol{E}$表示长度为k的单位向量，k为当前哈希码的长度；diag()表示矩阵的对角线元素组成的向量；$\boldsymbol{B}_{\mathrm{p}}$、$\boldsymbol{B}_{\mathrm{t}}$分别为不同模态实例的哈希向量。

自注意力相似度迁移跨模态哈希网络的最终目标函数L_{loss}为

$$L_{\mathrm{loss}}=\min\mu S_{\mathrm{inter}}+(1-\mu)S_{\mathrm{intra}} \tag{11}$$

式中：μ为网络参数。

2 实验及结果分析

2.1 数据集与评价指标

MIRFLICKR-25K数据集包含从Flickr收集的25 000个实例。每个图像都带有相关的文本标签。总共选择了20 015个图像-文本对，每一对的文本都表示为1 386维词袋向量；MSCOCO数据集包括82 785张训练图像和40 504张验证图像，每个图像有5个标题。在实验中，将训练图像和验证图像合并，并去除那些在文本模态中没有实例的样本。最后，训练数据集包含122 218个图文对，每个实例的文本都表示为2 026维词袋向量，每个实例至少用80个类中的一个进行标注；IAPR TC-12数据集包含来自275个类的200 000个图文对。对于每个实例，文本都转换为1 251维的词袋向量。在删除无标注的实例后，使用19 998个实例的子集进行实验。

为了验证实验的有效性，对数据集进一步划分，划分结果如表1所示。

表1 数据集配置表

Table 1 Configuration table of datasets

数据集	总数目/个	训练集/个	验证集/个
MIRFLICKR-25K	20 015	10 000	2 000
MSCOCO	122 218	10 000	5 000
IAPR TC-12	19 998	10 000	2 000

汉明排序是跨模态哈希算法常用的评估性能的方法，实验使用平均精确度均值(mean average precision，MAP)作为评价标准。查询第q个样本的精确度均值的计算式为

$$E_{\mathrm{AP}}(q)=\frac{1}{l_q}\sum_{r=1}^{R}P_q(r)\delta_q(r) \tag{12}$$

MAP的计算式为

$$E_{\mathrm{MAP}}=\frac{1}{Q}\sum_{q=1}^{Q}E_{\mathrm{AP}}(q) \tag{13}$$

式中：l_q为排名前R的检索结果的正确数目统计，R为整个数据库的样本数；$P_q(r)$表示前r个样本通过距离测量$\delta_q(r)$获得的搜索结果的精度；$\delta_q(r)$表示通过查询获得的第r个样本是否相关，$\delta_q(r)$=1表示相关数据，$\delta_q(r)$=0表示无关数据；Q为所有查询样本的数目。

2.2 实验设置

为验证本文模型，在NVIDIA 1080 Ti GPU的工作站上使用Python语言基于Pytorch框架搭建实验平台，在跨模态检索通用的MIRFLICKR-25K、MSCOCO和IAPR TC-12数据集上进行对比实验。

在实验中，将经过预处理的多模态数据传入网络G，得到各个模态特征矩阵，计算实值相似度值，通过哈希函数计算得到初始哈希码并计算目标函数，之后反向传播整体网络，利用随梯度下降法更

新网络参数及哈希码直至收敛。特别地, 激活函数中的ρ初始值设为 1, 并随迭代次数不断增大。训练的批次大小设置为 16, 网络参数α、β、μ、γ、ω初始值设置为 0.2, 采用网格搜索的方法, 每个值以 0.03 为步长, 寻找效果最佳的参数。

2.3　实验结果及对比分析

本文模型是一种无监督跨模态哈希检索方法。为验证模型的有效性, 使用 MAP 作为评价标准, 在 3 个常用数据集 MIRFLICKR-25K、MSCOCO 和 IAPR TC-12 上, 当哈希码长度为 16 bit、32 bit 和 64 bit 时, 分别与深度跨模态哈希 (deep cross-modal hashing, DCMH)[23] 方法、成对关系引导的深度哈希 (pairwise relationship guided deep hashing, PRDH)[24] 方法、跨模态汉明哈希 (cross-modal hamming hashing, CMHH)[25] 方法、基于自约束和注意力且比特位可扩展的跨模态哈希检索 (self-constraining and attention-based hashing network for bit-scalable cross-modal retrieval, SCAHN)[26] 方法、用于高效跨模态检索的相关哈希网络 (correlation hashing network for efficient cross-modal retrieval, CHN)[27]、用于跨模态检索的自监督对抗式哈希网络 (Self-supervised adversarial hashing networks for cross-modal retrieval, SSAH)[28]、用于无监督跨模态检索 (unsupervised cross-modal retrieval, UCH) 的深度语义保留重构哈希方法进行对比分析, 实验结果如表 2 所示, 其中, 加黑数值为最优结果。

表 2　在 3 个公共数据集上各方法的 MAP

Table 2　MAP of each method on three public datasets

检索任务	方法	MIRFLICKR-25K上的MAP			MSCOCO上的MAP			IAPR TC-12上的MAP		
		16 bit	32 bit	64 bit	16 bit	32 bit	64 bit	16 bit	32 bit	64 bit
图像检索文本	DCMH	0.71	0.72	0.72	0.49	0.55	0.53	0.45	0.47	0.49
	PRDH	0.64	0.66	0.65	0.51	0.52	0.51	0.45	0.43	0.46
	CMHH	0.72	0.72	0.74	0.52	0.53	0.54	0.41	0.42	0.43
	SCAHN	0.78	0.78	0.79	0.59	0.62	0.63	0.46	0.48	0.51
	CHN	0.73	0.71	0.71	0.54	0.55	0.53	0.42	0.45	0.49
	SSAH	0.75	0.76	0.78	0.48	0.51	0.51	0.52	0.54	0.55
	UCH	0.65	0.67	0.68	0.52	0.53	0.55	0.45	0.47	0.47
	本文模型	**0.85**	**0.85**	**0.88**	**0.65**	**0.70**	**0.69**	**0.54**	**0.57**	**0.60**
文本检索图像	DCMH	0.72	0.72	0.74	0.48	0.49	0.51	0.49	0.51	0.53
	PRDH	0.69	0.72	0.71	0.45	0.48	0.49	0.41	0.45	0.47
	CMHH	0.74	0.76	0.77	0.49	0.52	0.56	0.41	0.43	0.44
	SCAHN	0.78	0.79	0.81	0.61	0.63	0.65	0.51	0.53	0.56
	CHN	0.75	0.74	0.74	0.49	0.51	0.52	0.49	0.51	0.53
	SSAH	0.72	0.75	0.77	0.42	0.46	0.48	0.49	0.51	0.53
	UCH	0.66	0.67	0.67	0.50	0.52	0.55	0.45	0.47	0.49
	本文模型	**0.83**	**0.86**	**0.87**	**0.64**	**0.68**	**0.66**	**0.55**	**0.57**	**0.57**

由表 2 可得, 在 MIRFLICKR-25K、MSCOCO 和 IAPRTC-12 数据集上, 本文模型与对比方法相比, 在图像检索文本任务中, 哈希码长度为 16 bit、32 bit、64 bit 时 MAP 分别提升了 (0.07、0.06、0.02)、(0.07、0.08、0.03) 和 (0.09、0.06、0.05); 在文本索引图像的任务中 MAP 分别提升了 (0.05、0.03、0.04)、(0.07、0.05、0.04) 和 (0.06、0.01、0.01); 本文模型在 IAPRTC-12 数据集提升较小, 这可能是因为该数据集分类数目多且规模较小, 使得模型训练不够充分造成的。同时, 大多数方法随着比特数的增加, 检索效果更好, 这是因为随着比特数增加, 保存的语义信息更加丰富。

本文模型在哈希码长度为 64 bit 的条件下, 3 个数据集图像检索文本任务的 MAP 达到 0.723, 文本检索图像任务的 MAP 达到 0.700, 可以充分证明本文模型的有效性。通过对不同模态数据进行注意力计算, 能有效的分辨出重要的特征信息来提升获取特征的质量, 进而生成更具辨别性的哈希码, 从而实现检索性能的提升。使用共同注意力机制, 加强不同模态信息交互, 增强语义互补性。此外, 通过对成对相似值的迁移学习, 更精确的表示样本间的相似性, 减少了不必要的误差, 对提升检索性能有着重要的贡献。通过上述步骤, 可以充分挖掘模态间的相关性, 有效弥合模态间

的语义差距进而提高检索的准确率。

2.4 消融实验

为证明各个模块的效果，设哈希码长度为32 bit，在MSCOCO数据集上利用MAP作为评价标准进行实验，实验设置与2.2节在MSCOCO数据集上的设置一样，实验结果如表3所示。其中，SSMH-1是本文模型，SSMH-2在SSMH-1的基础上取消了图像混合注意力；SSMH-3取消了文本注意力；SSMH-4取消了共同注意力机制；SSMH-5在迁移学习过程中使用汉明距离计算哈希码距离。

表3 消融实验结果

Table 3 Results of ablation experiments

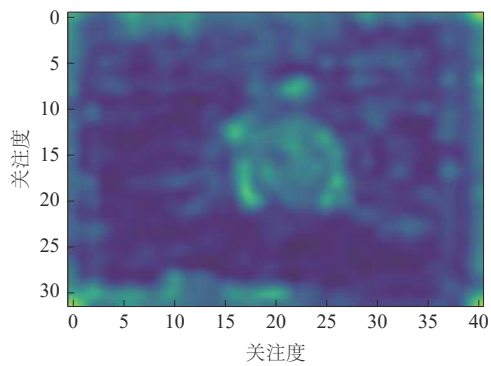
模型	MAP	
	图像检索文本	文本检索图像
SSMH-1	0.697 8	0.684 1
SSMH-2	0.687 6	0.675 5
SSMH-3	0.698 3	0.669 2
SSMH-4	0.685 7	0.677 2
SSMH-5	0.691 7	0.683 2

由表3可得，① SSMH-2实验数据验证了使用混合注意力，可以对图像的信息分配权重，对关键信息加强关注，使图像语义信息表达更加精确，从而提升检索性能；② 在图像检索文本任务中，SSMH-3的性能高于SSMH-1，但总体性能SSMH-1优于SSMH-3，证明多头注意力的使用对检索精度有一定的提高；③ SSMH-4实验数据验证了共同注意力模块可以利用不同模态存在语义互补性，进一步提升检索性能；④ SSMH-5在迭代过程中采用汉明距离计算哈希空间内的相似度，检索性能低于本文模型，证明式(7)的应用更合理。

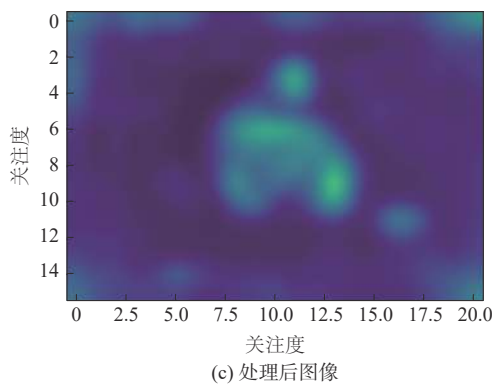
为证明混合注意模块的有效性，在实验中对该模块进行可视化，结果如图5所示。



(a) 原始图像



(b) 热力图



(c) 处理后图像

图5 混合注意力效果

Fig. 5 Effect of hybrid attention

图5中的原始图像是MACOCO数据集中的实例，可以看出在经过混合注意力处理后，重点信息“马”的图像部分被加强表示，其余部分被减弱，可证明混合注意力机制是有效的。

3 结 论

1）针对现有算法对各模态关键信息关注不足及在实值空间构建关系矩阵会增加模型训练难度的问题，本文提出自注意力相似度迁移跨模态哈希网络模型。

2）本文模型使用两级注意力机制关注多模态数据的重点信息，增强语义交互，提升语义表达性；并使用相似度值迁移的方法，降低模型训练难度。

3）本文模型可实现较为优异的检索性能，通过

在公开的 3 个多模态数据集 MIRFLICKR-25K、IAPR TC-12 和 MSCOCO 上的实验评估表明, 注意力机制和相似度迁移方法的使用对检索精度有一定的提升。在哈希码长度为 64 bit 的条件下, 3 个数据集图像检索文本任务的 MAP 达到 0.723, 文本检索图像任务的 MAP 达到 0.700。

参考文献（References）

[1] 邵杰. 基于深度学习的跨模态检索[D]. 北京: 北京邮电大学, 2017: 10-13.
SHAO J. Cross-modal retrieval based on deep learning [D]. Beijing University of Posts and Telecommunications, 2017: 10-13(in Chinese).

[2] ZHANG Y, OU W, ZHANG J, et al. Category supervised cross-modal hashing retrieval for chest x-ray and radiology reports [J]. Computers and Electrical Engineering, 2022,98: 107673.

[3] KAI L, QI G J, YE J, et al. Linear subspace ranking hashing for cross-modal retrieval[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2016, 39(9): 1825-1838.

[4] LI C, DENG C, LI N, et al. Self-supervised adversarial hashing networks for cross-modal retrieval[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2018: 4242-4251.

[5] WU J, WENG W, FU J, et al. Deep semantic hashing with dual attention for cross-modal retrieval[J]. Neural Computing and Applications, 2022, 34(7): 5397-5416.

[6] HUANG Z, WANG X, HUANG L, et al. CCNet: Criss-cross attention for semantic segmentation[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE Press, 2019: 603-612 .

[7] LI X, WANG W, HU X, et al. Selective kernel networks[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2019: 510-519.

[8] YU C, WANG J, PENG C, et al. BiSeNet: Bilateral segmentation network for real-time semantic segmentation[C]//Proceedings of the European Conference on Computer Vision. Piscataway: IEEE Press, 2018: 325-341.

[9] FU J, LIU J, TIAN H, 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: 3146-3154.

[10] 柳兴华, 曹桂涛, 林秋斌, 等. 自适应混合注意力深度跨模态哈希[J]. 计算机应用, 2022, 42(12): 3663-3670.
LIU X H, CAO G T, LIN Q B, et al. Adaptive hybrid attention hashing for deep cross-modal retrieval[J]. Journal of Computer Applications, 2022, 42(12): 3663-3670(in Chinese).

[11] WU G, LIN Z, HAN J, et al. Unsupervised deep hashing via binary latent factor models for large-scale cross-modal retrieval [C]//Proceedings of the 27th International Joint Conference on Artificial Intelligence. New York: ACM, 2018: 2854-2860.

[12] KAUR P, PANNU H S, MALHI A K. Comparative analysis on cross-modal information retrieval: A review[J]. Computer Science Review, 2021, 39(2): 100336.

[13] LI C, DENG C, WANG L, et al. Coupled CycleGAN: Unsupervised hashing network for cross-modal retrieval[C]//Proceedings of the 33rd AAAI Conference on Artificial Intelligence. Palo Alto: AAAI Press, 2019, 33(1): 176-183.

[14] CHENG S, WANG L, DU A. Deep semantic-preserving reconstruction hashing for unsupervised cross-modal retrieval[J]. Entropy, 2020, 22(11): 1266.

[15] 康培培, 林泽航, 杨振国, 等. 成对相似度迁移哈希用于无监督跨模态检索[J]. 计算机应用研究, 2021, 38(10): 3025-3029.
KANG P P, LIN Z H, YANG Z G, et al. Pairwise similarity transferring hash for unsupervised cross-modal retrieval[J]. Application Research of Computers, 2021, 38(10): 3025-3029 (in Chinese).

[16] WANG D, WANG Q , HE L. et al. Joint and individual matrix factorization hashing for large-scale cross-modal retrieval[J]. Pattern Recognition, 2020, 107: 107479.

[17] GUO M H, XU T X, LIU J J, et al. Attention mechanisms in computer vision: A survey[J]. Computational Visual Media, 2022, 8(3): 331-368.

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

[19] WANG X L, GIRSHICK R, GUPTA A, et al. Non-local neural networks [C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2018: 7794-7803.

[20] GUO J, MA X, SANSOM A, et al. Spanet: Spatial pyramid attention network for enhanced image recognition[C]//Proceedings of the IEEE International Conference on Multimedia and Expo. Piscataway: IEEE Press, 2020: 1-6.

[21] WOO S, PARK J, LEE J Y, et al. Cbam: Convolutional block attention module[C]//Proceedings of the European Conference on Computer Vision. Piscataway: IEEE Press, 2018: 3-19.

[22] HOU Q, ZHOU D, FENG J. Coordinate attention for efficient mobile network design[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2021: 13713-13722.

[23] JIANG Q Y, LI W J. Deep cross-modal hashing[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2017: 3232-3240.

[24] YANG E, DENG C, LIU W, et al. Pairwise relationship guided deep hashing for cross-modal retrieval[C]//Proceedings of the 31st AAAI Conference on Artificial Intelligence. New York: ACM, 2017: 1618-1625.

[25] CAO Y, LIU B, LONG M, et al. Cross-modal hamming hashing[C]//Proceedings of the European Conference on Computer Vision. Piscataway: IEEE Press, 2018: 202-218.

[26] WANG X, ZOU X, BAKKER E M, et al. Self-constraining and attention-based hashing network for bit-scalable cross-modal retrieval[J]. Neurocomputing, 2020, 400: 255-271.

[27] CAO Y , LONG M , WANG J , et al. Correlation hashing network for efficient cross-modal retrieval[C]//Proceedings of the British Machine Vision Conference 2017. London: British Machine Vision Association, 2017: 1-7.

[28] LI C, DENG C, LI N, et al. Self-supervised adversarial hashing networks for cross-modal retrieval[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2018: 4242-4251.

Cross-modal hashing network based on self-attention similarity transfer

LIANG Huan，WANG Hairong*，WANG Dong

(Department of Computer Science and Engineering, North Minzu University, Yinchuan 750021，China)

Abstract：To further improve the performance of cross-modal retrieval, a cross-modal hashing network model is proposed based on self-attention similarity transfer. A channel spatial hybrid self-attention mechanism is designed to strengthen the key information of the concerned image, and the common attention method is used to enhance the interaction of modal information, thus improving the quality of feature learning. To reconstruct the similarity relationship in the hash space, the transfer learning method is used to guide the generation of hash codes by using the real-valued space similarity. Comparative experiments are carried out on three commonly used datasets, MIRFLICKR-25K, IAPRTC-12 and MSCOCO, with excellent methods such as deep cross-modal hashing (DCMH), pairwise relationship guided deep hashing (PRDH) and cross-modal hamming hashing (CMHH). The results show that when the length of the hash code reaches 64 bit, the mean average precision (MAP) of the text retrieval task with image queries in three datasets is 72.3%, and that the MAP of the image retrieval task with text queries is 70%. These values are higher than those of other methods.

Keywords：cross-modal retrieval；hashing learning；attention mechanism；transfer learning；unsupervised learning

Received：2022-05-21；**Accepted**：2022-07-02；**Published Online**：2022-10-21 13：24

URL：link.cnki.net/urlid/11.2625.V.20221021.0943.001

Foundation items：Ningxia Hui Autonomous Region Department of Education Higher Education Key Project of Scientific Research (NYG2022051); Ningxia Natural Science Foundation Project (2023AAC03316)

* **Corresponding author**. E-mail：bmdwhr@163.com

http://bhxb.buaa.edu.cn jbuaa@buaa.edu.cn
DOI: 10.13700/j.bh.1001-5965.2022.0404

多输入傅里叶神经网络及其麻雀搜索优化

黎亮亮，张著洪*，张永丹

(贵州大学 大数据与信息工程学院，贵州省系统优化与科学计算特色重点实验室，贵阳 550025)

摘 要：鉴于反向传播（BP）神经网络存在灵敏度高但收敛速度慢，以及已有傅里叶神经网络不具备多输入数据特征提取能力，借助多个傅里叶神经网络构建能接收多维数据的堆叠神经网络，进而将其与多层感知器融合，获得基于梯度下降的多输入傅里叶神经网络。结合此神经网络获取全局最优参数值难的因素，通过在麻雀搜索算法中引入Cat混沌映射、动态种群规模调节机制及参数自适应调节方案，提出改进型麻雀搜索算法，并将其应用于多输入傅里叶神经网络的参数优化及高维函数优化问题的求解。理论分析可得，所提算法的计算复杂度主要由种群规模和优化问题的维度决定。比较性的数值实验表明，所获神经网络提取多源数据特征的能力和泛化能力强，同时所提算法处理高维优化问题具有明显优势且收敛速度快。

关 键 词：傅里叶神经网络；多层感知器；麻雀搜索；高维函数优化；多属性分类
中图分类号：TP301.6
文献标志码：A **文章编号**：1001-5965(2024)02-0623-11

人工神经网络已被广泛用于函数拟合、风险预测、系统控制、模式分类等[1]，但权值与阈值多且冗余神经元难以剔除，导致存在公认的收敛速度慢、易陷入局部搜索、稳定性差、抗噪能力弱、易梯度消失和梯度爆炸等不足[2]。目前，多输入神经网络主要有反向传播神经网络(back-propagation neural network, BPNN)、极限学习机(extreme learning machine, ELM)、径向基函数(radial basis function, RBF)神经网络、小波神经网络(wavelet neural network, WNN)、循环神经网络、卷积神经网络等。前3种神经网络不具备噪声抑制功能，且对输入特征的提取能力均偏弱；小波神经网络具有强噪声抑制能力，但不具备特征提取功能，且收敛速度慢及求解精度不高；循环神经网络具有记忆性、参数共享和特征提取功能，但噪声抑制能力较弱且计算复杂度高；卷积神经网络虽然有强的特征提取功能，但缺乏噪声抑制能力，致使数据分类效果受噪声干扰大，网络训练耗时长。因此，研究噪声抑制和特征提取能力强，且网络训练效果好和效率高的多输入神经网络，已成为神经网络研究面临的重要课题。

三层傅里叶神经网络(Fourier neural network, FNN)[3]提出后，因其结构简单、权值与阈值更新速度快，且具有特征表示、时间序列信号模拟和噪声抑制能力及数据特征提取能力强，已被应用于时间序列分析、扰动抑制、控制器的参数实时调节、预测等问题[3-7]，但因信号特征提取单一且仅适合于一维函数逼近、时间序列分析等，导致应用受限。于是，构建能同时提取多种输入信号的特征信息、求解精度高、不易陷入局部最优、参数更新速度快，且特征提取与噪声抑制能力强的多输入FNN(multi-input FNN, MIFNN)已受到重视。

现有神经网络的参数优化方法大致可概括为2类，即诸如梯度下降、最小二乘法、牛顿法等数值优化[8-9]和智能优化[10-11]。数值优化是基于最优性条件的数值迭代算法，其更新步长直接影响算法的全局收敛性，常常仅能获得局部最优解。例如，杨

收稿日期：2022-05-21；**录用日期**：2022-11-25；**网络出版时间**：2022-12-14 09:27
网络出版地址：link.cnki.net/urlid/11.2625.V.20221213.1513.006
基金项目：国家自然科学基金(62063002)
*通信作者. E-mail：zhzhang@gzu.edu.cn

引用格式：黎亮亮，张著洪，张永丹. 多输入傅里叶神经网络及其麻雀搜索优化[J]. 北京航空航天大学学报，2024，50(2)：623-633.
LI L L，ZHANG Z H，ZHANG Y D. Multi-input Fourier neural network and its sparrow search optimization[J]. Journal of Beijing University of Aeronautics and Astronautics，2024，50(2)：623-633 (in Chinese).

旭华[1]提出三层单输入、多输出FNN，其借助最小二乘法优化网络参数，获得的网络能有效降低噪声对样本特征提取的影响，但仅适用于小规模数据集的函数逼近；Lin等[9]利用牛顿法训练网络参数，能使网络的泛化能力相较于FNN有所提升，但算法的计算复杂度高。另外，智能优化是源于自然界中生物进化、觅食行为特性或物理现象启发的随机搜索算法，其因结构简单、收敛速度快[10-11]，已成为较为活跃的研究分支，并已有众多具有强竞争力的新型优化算法被提出，如麻雀搜索算法(sparrow search algorithm, SSA)[12]、哈里斯鹰优化(Harris hawks optimization, HHO)算法[13]、天鹰优化(aquila optimizer, AO)算法[14]、野狗优化算法(dingo optimization algorithm, DOA)[15]、鲸鱼优化算法(whale optimization algorithm, WOA)[16]等。这些算法均经由生物行为特性而提出，属于群智能优化算法，且求解中、低维优化问题具有强的竞争力。文献[17]和文献[18]分别将正余弦策略与柯西变异引入SSA，获得的SSA可有效平衡全局开采和局部勘测能力，且收敛速度快及寻优精度高，但搜索效果欠稳定且在应用于高维优化问题时，搜索性能急剧退化，以及早熟现象较为严重。此类算法常由已获得的最好个体引导种群进化或种群更新，算法在进化后期出现种群多样性不足，进而出现局部收敛现象。例如，文献[19]和文献[20]通过将种群划分思想和精英引导的学习策略引入到粒子群优化算法中，获得的算法能有效平衡全局开采和局部勘测能力，可提高算法在高维函数优化问题上的求解精度，但算法的计算复杂度较高。Zhang等将果蝇的视觉信息处理机制与智能优化算法中种群更新策略结合[21-22]，获得果蝇视觉进化神经网络，其相比于元启发式随机搜索算法，求解高维优化问题的性能有明显优势且效率高。尽管如此，全局收敛能力强、结构简单且收敛速度快的新型优化算法仍是智能优化研究关注的重点。

综上，本文通过将多层感知器(multi-layer perceptron, MLP)与FNN结合，提出基于梯度下降的MIFNN；进而，提出可求解高维优化问题的改进型麻雀搜索算法(improved sparrow search algorithm, ISSA)，并将其用于MIFNN的结构优化。

1 多输入傅里叶神经网络模型与设计

1.1 神经网络的拓扑结构

传统的FFN[1,3]是一种由$2L+2$个神经节点依次连接而成的单输入单输出三层神经网络，其中间层包含$2L$个神经节点，每个节点的激励函数在正交基集$\{1,\sin nx,\cos nx,1\leqslant n\leqslant L\}$中，且输出由傅里叶级数的有限项表示，即

$$f(x)\sim\frac{a_0}{2}+\sum_{n=1}^{L}(a_n\cos nx+b_n\sin nx) \tag{1}$$

其中，a_0、a_n、b_n为傅里叶系数，且在神经网络中为神经节点的连接权值。由于此神经网络存在待定的$2L+1$个权重，当其触及多源信号拟合、多维数据挖掘时，必然存在大量尚需确定的权值。为此，本文将MLP与基于正交基集$\{1,\cos nx,1\leqslant n\leqslant L\}$的堆叠FFN组合，建立MIFNN，如图1所示。

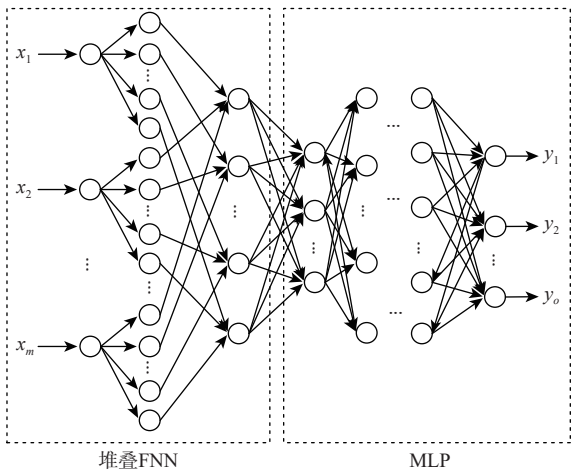


图1 MIFNN的结构示意图

Fig. 1 Schematic diagram of MIFNN structure

图1中，输入向量为$[x_1,x_2,\cdots,x_m]$，m为输入维度；输出向量为$[y_1,y_2,\cdots,y_o]$，o为输出维度；MIFNN的前馈结构由堆叠FNN与MLP顺次连接而成。堆叠FNN的主要作用是将多源输入信息中每个分量信息经由多通道进行特征提取，去除冗余噪声信息；其由m个结构相同的单输入FNN堆叠而成，包含m个输入神经元、mL个扩展神经元及L个输出神经元；每个子网络的扩展神经元个数为L；特别地，第k个输出神经元的输出是每个FNN的扩展层中第k个神经元的输出量的线性加权组合。MLP包含$K+1$个神经层，其输入层与堆叠FNN的输出层通过全连接方式连接，且第k层的神经元个数为s_k；每层的神经元激励函数均为$g(\cdot)$。由此可知，MIFNN首先以堆叠FNN的输入作为输入，进而以MLP的输出为输出。

结合单输入FNN的固有特性，堆叠FNN的主要任务是多通道提取多输入特征，消除冗余噪声对特征提取的影响，其每个输出神经元仅接收所有单输入FNN的扩展层中相同位置的神经元输出，并经由加权方式获得输出。换言之，其借助多通道对多源输入数据进行正交扩展，实现多通道特征提取

和噪声抑制, 继而融合多输入的属性特征; 其次, 每个单输入 FNN 的扩展层与堆叠 FNN 的输出层的神经元个数相等, 原因在于堆叠 FNN 的每个输出神经元仅接收各单输入 FNN 对应通道的信息, 此有助于利用多通道提取输入特征信息及提高神经网络的训练效率。另外, MIFNN 利用堆叠 FNN 提取多输入的特征, 进而利用 MLP 实现函数拟合或数据分类, 此二者的结合既可回避噪声对问题求解质量的影响, 又可提升问题求解的效果与效率。

1.2　误差损失函数

假定样本集 Ω 由 M 个样例构成, 即 $\{(\boldsymbol{x}(n), \boldsymbol{d}(n))|1\leqslant n\leqslant M\}$, 其中 $\boldsymbol{x}(n)$ 为第 n 时刻的输入向量, 维度为 m; $\boldsymbol{d}(n)$ 为该样例的标签或观测向量。MIFNN 接收输入样例 $\boldsymbol{x}(n)$ 后, 经由堆叠 FNN 中各个单输入 FNN 的余弦正交扩展, 提取样例的深度特征信息, 进而经由 MLP 产生输出信号。MIFNN 的详细设计如下:

1) 输入层: 由 m 个神经节点表示, 其节点 i 在第 n 时刻接收输入向量 $\boldsymbol{x}(n)$ 中第 i 个分量 $x_i(n)$; 每个神经节点与相应扩展层中各神经元的连接权值为 1。

2) 扩展层: 通过余弦正交扩展, 深度提取 $\boldsymbol{x}(n)$ 中每个维度的特征; 第 i 个输入分量 $x_i(n)$ 对应的 FNN 中扩展神经元 j 的输出表示如下:

$$y_{j+(i-1)L}^{(\mathrm{E})}(n)=\cos(jx_i(n)) \qquad 1\leqslant i\leqslant m, 1\leqslant j\leqslant L \tag{2}$$

3) 堆叠 FNN 的第 j 个输出神经元接收 m 个 FNN 中每个网络的第 j 个扩展神经元的输出, 并经由权值 w_{ij}^{F} 线性组合获得其输出, 即

$$y_j^{(\mathrm{F})}(n)=\sum_{i=1}^{m} w_{ij}^{\mathrm{F}}(n)y_{j+(i-1)L}^{(\mathrm{E})}(n) \tag{3}$$

4) 通过全连接方式将堆叠 FNN 的输出层中各神经元的输出映射为 MLP 的输入, 并经由激励函数 $g(\cdot)$ 获相应输出, 即 MLP 的输入层中第 j 个神经元的输出计算式为

$$y_j^{(0)}(n)=g(H_j^{(0)}(n))=g\left(\sum_{i=1}^{L} w_{ij}^{(0)}(n)y_i^{(\mathrm{F})}(n)+b_j^{(0)}(n)\right) \tag{4}$$

式中: $1\leqslant j\leqslant s_0$, $H_j^{(0)}(n)$ 为 MLP 的输入层中第 j 个神经元在第 n 时刻的诱导局部域; s_0 为 MLP 输入层的神经元个数; $b_j^{(0)}$ 为 MLP 的输入层中第 j 个神经元的阈值; $w_{ij}^{(0)}$ 为堆叠 FNN 的输出层中第 i 个神经元与 MLP 的输入层中第 j 个神经元的连接权值。

5) MLP 的第 l 个隐含层中第 j 个神经元的输出计算式为

$$y_j^{(l)}(n)=g(H_j^{(l)}(n))=g\left(\sum_{i=1}^{s_{l-1}} w_{ij}^{(l)}(n)y_i^{(l-1)}(n)+b_j^{(l)}(n)\right) \tag{5}$$

式中: $1\leqslant j\leqslant s_l, 1\leqslant l\leqslant K-1$, $H_j^{(l)}(n)$ 为第 l 层中第 j 个神经元在第 n 时刻的诱导局部域; $b_j^{(l)}$、$w_{ij}^{(l)}$ 和 $s_{l\text{-}1}$ 分别为第 l 层中第 j 个神经元的阈值、该神经元与第 l-1 层中第 i 个神经元的连接权值和第 l-1 层的神经元个数。

6) MIFNN 的输出层中第 j 个神经元的输出计算式为

$$o_j(n)=g\left(\sum_{i=1}^{s_{K-1}} w_{ij}^{(K)}(n)y_j^{(K-1)}(n)+b_j^{(K)}(n)\right) \tag{6}$$

式中: $1\leqslant j\leqslant s_K$, $b_j^{(K)}$、$w_{ij}^{(K)}$ 和 s_K 分别为输出层中第 j 个神经元的阈值、此神经元与第 K-1 个隐含层中第 i 个神经元的连接权值和此层的神经元个数。于是, 在线学习下第 n 时刻的损失函数可表示为

$$E_n(\boldsymbol{W},\boldsymbol{b})=\frac{1}{2}\sum_{j=1}^{s_K}\left(o_j(n)-d_j(n)\right)^2 \tag{7}$$

式中: $\boldsymbol{W}$ 和 $\boldsymbol{b}$ 分别为 MIFNN 的权值和阈值构成的向量。

1.3　权值与阈值更新

传统的 BP 方法是将前馈传播产生的误差量反向传递给各神经层, 并借助梯度下降法更新网络中的 $\boldsymbol{W}$ 和 $\boldsymbol{b}$, 此易于使网络参数的更新陷入局部最优且收敛速度慢。然而, 动量项的引入却有助于提高梯度下降法的收敛速度和增强抗噪能力及网络训练的稳定性[1-2]。为此, 建立含动量项的梯度下降 MIFNN(momentum and gradient descent based MIFNN, MGD-MIFNN), 具体如下:

1) 全连接层 (MLP) 的 BP: 此为熟知且含动量项的 BP。具体而言, 基于式(7), 在第 n 时刻下第 k 个隐含层中第 j 个神经节点的局部梯度信息计算式为

$$\delta_j^{(k)}(n)=-\frac{\partial E_n}{\partial H_j^{(k)}(n)}=-\frac{\partial E_n}{\partial y_j^{(k)}(n)}g'(H_j^{(k)}(n)) \qquad 1\leqslant j\leqslant s_K \tag{8}$$

因此, 第 n+1 时刻下 MLP 的第 k 个隐含层中第 j 个神经元节点与上一个隐含层中第 i 个神经元节点之间的连接权值由下式更新:

$$w_{ij}^{(k)}(n+1)=w_{ij}^{(k)}(n)-\eta\frac{\partial E_n}{\partial w_{ij}^{(k)}(n)}-\alpha\frac{\partial E_{n-1}}{\partial w_{ij}^{(k)}(n-1)} \tag{9}$$

$$\frac{\partial E_n}{\partial w_{ij}^{(k)}(n)}=\frac{\partial H_j^{(k)}(n)}{\partial w_{ij}^{(k)}(n)}\cdot\frac{\partial E_n}{\partial H_j^{(k)}(n)}=-y_i^{(k-1)}(n)\delta_j^{(k)}(n) \tag{10}$$

其中, η 和 α 分别为固定的学习率和动量因子。同理, 在第 n+1 时刻下第 k 个隐含层中第 j 个神经元

节点的阈值经由下式更新：

$$b_j^{(k)}(n+1)=b_j^{(k)}(n)-\eta\frac{\partial E_n}{\partial b_j^{(k)}(n)}-\alpha\frac{\partial E_{n-1}}{\partial b_j^{(k)}(n-1)} \tag{11}$$

$$\frac{\partial E_n}{\partial b_j^{(k)}(n)}=\frac{\partial H_j^{(k)}(n)}{\partial b_j^{(k)}(n)}\cdot\frac{\partial E_n}{\partial H_j^{(k)}(n)}=-\delta_j^{(k)}(n) \tag{12}$$

综上，MLP 中含动量项的权值和阈值更新为

$$w_{ij}^{(k)}(n+1)=\begin{cases}w_{ij}^{(k)}(n)+\eta y_i^{(k-1)}\delta_j^{(k)}(n)-\\ \quad\alpha\dfrac{\partial E_{n-1}}{\partial w_{ij}^{(k)}(n-1)} & 1\leqslant k\leqslant K\\ w_{ij}^{(k)}(n)+\eta y_i^{(F)}\delta_j^{(k)}(n)-\\ \quad\alpha\dfrac{\partial E_{n-1}}{\partial w_{ij}^{(k)}(n-1)} & k<1,k>K\end{cases} \tag{13}$$

式中：$1\leqslant j\leqslant s_k,1\leqslant i\leqslant s_{k-1}$。另外，

$$b_j^{(k)}(n+1)=b_j^{(k)}(n)+\eta\delta_j^{(k)}(n)-\alpha\frac{\partial E_{n-1}}{\partial b_j^{(k)}(n-1)}\quad 0\leqslant k\leqslant K,1\leqslant j\leqslant s_k \tag{14}$$

2）堆叠 FNN：输出层中第 j 个神经元与第 i 个 FNN 中第 j 个扩展神经元的连接权值更新如下：

$$w_{ij}^{\mathrm{F}}(n+1)=w_{ij}^{\mathrm{F}}(n)-\eta y_i^{(k-1)}\frac{\partial E_n}{\partial w_{ij}^{\mathrm{F}}(n)}-\alpha\frac{\partial E_{n-1}}{\partial w_{ij}^{\mathrm{F}}(n-1)} \tag{15}$$

其中，

$$\frac{\partial E_n}{w_{ij}^{\mathrm{F}}(n)}=\frac{\partial y_j^{(\mathrm{F})}(n)}{w_{ij}^{\mathrm{F}}(n)}\cdot\frac{\partial E_n}{\partial y_j^{(\mathrm{F})}(n)}=y_{j+(i-1)N}^{(\mathrm{E})}(n)\frac{\partial E_n}{\partial y_j^{(\mathrm{F})}(n)} \tag{16}$$

$$w_{ij}^{\mathrm{F}}(n+1)=w_{ij}^{\mathrm{F}}(n)-\eta y_{j+(i-1)N}^{(\mathrm{E})}(n)\frac{\partial E_n}{\partial y_j^{(\mathrm{F})}(n)}-\alpha\frac{\partial E_{n-1}}{\partial w_{ij}^{\mathrm{F}}(n-1)}\quad 1\leqslant i\leqslant m,1\leqslant j\leqslant L \tag{17}$$

MGD-MIFNN 在 MLP 基础上增加堆叠 FNN，可增强多属性特征提取的能力，提升神经网络的拟合能力。该神经网络与 FNN 相比，能处理多输入下信号拟合、数据分类等问题，应用范围广；与 MLP 相比，通过堆叠 FNN 对输入信号作平滑处理，消除隐含的噪声，使得模型的输出受噪声干扰小，函数逼近能力强。

MGD-MIFNN 是基于带动量项的梯度下降法获得网络的权值和阈值，此在超参数下，必然易于仅能获得局部最佳参数向量值，进而较难使得损失函数达到全局最小。因此，探讨求解高维优化问题的 ISSA，使损失函数达到全局最小。

2 改进型麻雀搜索算法

2.1 麻雀搜索算法

SSA 由 3 种类型麻雀组成[17]，即发现者、加入者和侦察者。发现者拥有较大的适应度值，负责为加入者提供觅食区域和方向；加入者为了发现食物源，会一直跟随发现者，且不断监视发现者并争夺食物；在发现者和加入者中随机选取部分麻雀作为侦察者，其作用是当发生危险时产生反捕食行为。

在第 t 时刻，SSA 选取规模为 N 的麻雀种群 $X(t)$ 中适应度值较大的前 20% 麻雀构成发现者种群 $Y(t)$，个体 i 的位置为 $\boldsymbol{y}_i(t)$，即 $[y_{i1}(t),y_{i2}(t),\cdots,y_{iD}(t)]$ D 为函数优化问题的决策向量维度；第 i 个发现者的位置更新描述如下：

$$\boldsymbol{y}_i(t+1)=\begin{cases}\boldsymbol{y}_i(t)\exp\left(-\dfrac{i}{\alpha T}\right) & R<G\\ \boldsymbol{y}_i(t)+Q\boldsymbol{L} & R\geqslant G\end{cases} \tag{18}$$

式中：T 为最大时间步长；α 和 R 分别为区间 [0,1] 上的随机数和预警值；G 为安全值；Q 为服从标准正态分布的非线性步长因子；$\boldsymbol{L}$ 为元素均为 1 的 D 维向量。

$X(t)$ 中剩下的 80% 麻雀构成加入者种群 $Z(t)$，加入者 i 的位置为 $\boldsymbol{z}_i(t)$，即 $[z_{i1}(t),z_{i2}(t),\cdots,z_{iD}(t)]$，其位置经由式 (19) 更新：

$$\boldsymbol{z}_i(t+1)=\begin{cases}Q\exp\left(\dfrac{\boldsymbol{x}_{\mathrm{w}}(t)-\boldsymbol{z}_i(t)}{i^2}\right) & i<\dfrac{N}{2}\\ \boldsymbol{x}_{\mathrm{p}}(t+1)+\left|\boldsymbol{z}_i(t)-\boldsymbol{x}_{\mathrm{p}}(t+1)\right|\boldsymbol{A}^{+}\boldsymbol{L} & i\geqslant\dfrac{N}{2}\end{cases} \tag{19}$$

式中：$\boldsymbol{x}_{\mathrm{w}}(t)$ 和 $\boldsymbol{x}_{\mathrm{p}}(t+1)$ 分别为 $X(t)$ 中适应度最低和最高的个体；$\boldsymbol{A}$ 表示一个与麻雀个体的位置同维度的列向量，其内部元素由 1 和-1 随机组成，$\boldsymbol{A}^{+}=\boldsymbol{A}^{\mathrm{T}}(\boldsymbol{A}\boldsymbol{A}^{\mathrm{T}})^{-1}$。

侦察者种群 $P(t)$ 由 $X(t)$ 中随机抽取的 20% 个麻雀构成；侦察者 i 的位置为 $\boldsymbol{p}_i(t)$，即 $[p_{i1}(t),p_{i2}(t),\cdots,p_{iD}(t)]$，经由式 (20) 进行位置更新：

$$\boldsymbol{p}_i(t+1)=\begin{cases}\boldsymbol{x}_{\mathrm{g}}(t)+\beta(\boldsymbol{p}_i(t)-\boldsymbol{x}_{\mathrm{g}}(t)) & f_i>f_{\mathrm{g}}\\ \boldsymbol{p}_i(t)+\gamma\dfrac{(\boldsymbol{p}_i(t)-\boldsymbol{x}_{\mathrm{w}}(t))}{f_i-f_{\mathrm{g}}} & f_i\leqslant f_{\mathrm{g}}\end{cases} \tag{20}$$

式中：$\boldsymbol{x}_{\mathrm{g}}(t)$ 为到第 t 时刻为止获得的全局最好个体，其适应度表示为 f_{g}；β 和 γ 为非线性步长因子；f_i 为 $Z(t)$ 中麻雀 i 的适应度；f_{w} 为当前种群中最差麻雀的适应度。

2.2 改进型麻雀搜索算法

虽然 SSA 具有求解精度高、稳定性好等优点，但面对高维或性能指标极为复杂的优化问题时，其搜索性能和效率急剧退化，存在的主要不足如下：

1）初始种群中麻雀的位置因分布不均匀而导致算法搜索过早出现适应度值远超过平均水平的“超级麻雀”，使其余的大量麻雀聚集，丧失种群多样性；

2) 发现者和加入者种群的规模不变性导致算法在迭代前期, 发现者的数量相对较少, 而在迭代后期, 发现者的数量又相对较多, 不利于有效地从全局开采向局部勘测转变, 进而导致算法的收敛速度慢;

3) 步长因子 Q、β 和 γ 的取值随机性直接影响 SSA 的寻优效率和效果, 导致 SSA 易于陷入局部搜索。

2.2.1　算法改进

针对 SSA 存在的不足, 围绕初始种群的生成方式, 发现者和加入者所占比例及算法参数进行如下改进。

1) 初始种群: 在 SSA 中, 种群随机初始化致使初始个体在解空间分布不均, 容易出现聚集现象, 从而导致种群多样性不足[17]。鉴于 Cat 混沌映射结构简单, 且在 [0,1] 间产生的混沌序列分布均匀[23], 在此借助式 (21) 产生混沌序列:

$$\begin{bmatrix} q_{n+1} \\ v_{n+1} \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 1 & 2 \end{bmatrix} \begin{bmatrix} q_n \\ v_n \end{bmatrix} \bmod 1 \tag{21}$$

式中: mod 1 表示取余运算符; q_n 和 v_n 为 0 到 1 之间的随机数。进而, 经由式 (22) 生成初始个体:

$$x_j = l_j + k_j(u_j - l_j) \qquad 1 \leqslant j \leqslant D \tag{22}$$

式中: u_j 和 l_j 分别为优化问题的候选解 $\boldsymbol{x}$ 中第 j 个分量的上、下界; k_j 由式(21)确定依次取 q_n 的值。

2)种群规模动态调节机制: 算法迭代初期, 发现者种群 $X(t)$ 的规模要求较大, 有利于增强处理高维优化问题的全局搜索能力。随着迭代次数的增加, 种群规模要求自适应减小, 加入者种群的规模自适应增大, 有助于增强局部搜索能力。具体而言, $X(t)$ 的规模由式 (23) 确定:

$$N_{\mathrm{p}} = \lfloor rN \rfloor \tag{23}$$

式中: $\lfloor \cdot \rfloor$ 表示向下取整; r 为发现者在种群中的占比, 其由式 (24) 确定:

$$r = 0.25 + 0.3\left(1 - \tan\frac{\pi t}{4T}\right) \tag{24}$$

$Y(t)$ 的规模由式 (25) 确定:

$$N_{\mathrm{s}} = N - N_{\mathrm{p}} \tag{25}$$

3) 非线性步长因子: Q、β 和 γ 易于导致 SSA 在迭代中、后期快速陷入局部搜索; 步长因子越大, 越容易发现全局最优解, 但收敛速度慢; 反之, 算法的局部勘测能力强, 收敛速度慢。为权衡高维优化问题求解情形下的寻优质量和搜索效率, 要求 Q、β 和 γ 随 t 的增大自适应地非线性递减, 即

$$Q = \frac{\exp\left(2\left(1-\frac{t}{T}\right)\right) - \exp\left(-2\left(1-\frac{t}{T}\right)\right)}{\exp\left(2\left(1-\frac{t}{T}\right)\right) + \exp\left(-2\left(1-\frac{t}{T}\right)\right)} \tag{26}$$

$$\beta = (f_{\mathrm{g}} - f_i)\sin\left(\pi\exp\left(-\frac{3}{2}\tan\left(\frac{t}{T}\right)^2\right)\right) \tag{27}$$

其中, γ 与 β 的取值相同。

2.2.2　算法描述

ISSA 的算法描述如下:

步骤 1　输入参数: 种群规模 N, 维度 D, 最大迭代数 T, 安全值 G;

步骤 2　置 t←1, 依据式(21)和式(22)初始化麻雀种群 $X(t)$, 适应度最高、最低的个体分别记作 $\boldsymbol{x}_{\mathrm{g}}$ 和 $\boldsymbol{x}_{\mathrm{w}}$;

步骤 3　经由式(23)和式(24)计算发现者、加入者种群的规模 N_{p} 及 N_{s};

步骤 4　$X(t)$ 中适应度较高的 N_{p} 个个体构成 $Y(t)$, 其余的个体构成 $Z(t)$;

步骤 5　由式(26)和式(27)计算非线性步长因子 Q、β 及 γ;

步骤 6　经由式(18)更新种群 $Y(t)$;

步骤 7　经由式(19)更新种群 $Z(t)$;

步骤 8　从 $X(t)$ 中随机抽取 20% 个体构成 $P(t)$, 并据式(20)更新 $P(t)$;

步骤 9　$Y(t)\cup Z(t)\cup P(t)$ 中适应度较高的 N 个个体构成 $X(t+1)$, 更新 $\boldsymbol{x}_{\mathrm{g}}$ 和 $\boldsymbol{x}_{\mathrm{w}}$;

步骤 10　置 t←t+1; 若 t <T, 则返回步骤 4; 否则, 结束, 且输出 $\boldsymbol{x}_{\mathrm{g}}$。

ISSA 的计算复杂度主要由步骤 6～步骤 8 确定。在一个迭代周期内, 步骤 6 共执行 $2N_{\mathrm{p}}D+N_{\mathrm{p}}$ 次算术运算; 步骤 7 共运行 $5N_{\mathrm{s}}D+N_{\mathrm{s}}$ 次; 步骤 8 共运行 $0.2N+0.8ND$ 次; 因此, 步骤 6～步骤 8 的计算次数为 $3N_{\mathrm{s}}D+1.2N+2.8ND$。由于 N_{s} 介于 $0.45N$～$0.75N$ 之间, 因此, ISSA 的计算复杂度主要由种群规模 N 和优化问题的维度 D 决定。

3　数值实验

在配置为 Windows10 (CPU/3.7 GHz, RAM/4.0 GB)/MATLAB 2018 环境下展开数值实验。首先, 借助具有强竞争力的元启发式随机搜索算法和基准测试问题, 检测 ISSA 求解高维函数优化问题的性能; 其次, 以开源数据集为例, 测试 MGD-MIFNN 和基于 ISSA 的 MIFNN 处理数据分类问题的性能; 最后, 测试 MIFNN 的结构变化对数据分类性能的影响。

3.1　ISSA 的性能测试

与 ISSA 比较的算法包括 SSA[12]、HHO 算法[13]

、AO算法[14]、DOA[15]及WOA[16]。对比算法的参数设置除种群规模和最大迭代数外均源于相应文献；设置各算法的种群规模为100，最大迭代次数为100；ISSA的安全值G取0.8。测试事例包括8种具有代表的基准函数优化问题F_1～F_8[24]，其中F_1～F_4是单峰值函数，F_5～F_8是含大量局部最优解的多峰值函数；每个事例的维度均取为2 000，因此，F_1～F_8为高维的单峰、多峰值函数，其相应的优化问题为高维函数优化问题。对于每个测试事例，每种算法均独立运行25次，且将各自获得的25个最好目标值的均值μ、均方差σ用于算法的比较分析，如表1和表2所示；以F_2、F_4、F_6、F_8为例，各算法的搜索曲线如图2所示。

经由表1和表2的均值可知，DOA和WOA求解以上8个测试事例的性能整体上劣于其他算法，同时表中的均方差值说明，WOA求解多模态优化问题的能力相对偏弱；HHO算法和AO算法在获得的解质量和搜索效果的稳定性方面整体上劣于SSA，且SSA和AO算法均不能有效求解F_5；SSA具有一定的优势。另外，ISSA能成功求解以上各测试问题，特别是求解F_5～F_8获得的解质量较为理想。因此，相比对比算法，ISSA求解高维优化问题具有明显优势。由图2可知，相比于对比算法，ISSA在初始时刻能产生多样且质量较好的个体，表明Cat混沌映射能增强初始种群的多样性和分布的均匀性；ISSA的收敛速度快且获得的解的精度高，此说明非线性步长因子能提升寻优效率，且种群规模动态调节有助于平衡算法的全局开采与局部勘测能力。

表1 各算法独立运行25次后获得的统计结果（F_1~F_4）

Table 1 Statistical results acquired by each algorithm after 25 runs per example(F_1-F_4)

算法	μ				σ			
	F_1	F_2	F_3	F_4	F_1	F_2	F_3	F_4
SSA	2.47×10	3.20×10^{-1}	5.14×10^{-4}	1.80×10^{-3}	1.91×10^{-1}	5.51×10^{-1}	1.10×10^{-3}	1.60×10^{-3}
HHO	6.45×10^{-1}	7.44×10^{0}	1.26×10^{-1}	5.79×10^{-4}	2.58×10^{-1}	9.25×10^{0}	1.22×10^{-1}	5.52×10^{-4}
AO	7.07×10^{-1}	7.05×10^{0}	2.96×10^{-2}	3.04×10^{-4}	1.24×10^{-1}	1.68×10	4.57×10^{-2}	3.42×10^{-4}
DOA	7.50×10^{-1}	2.00×10^{3}	4.97×10^{2}	1.90×10^{-3}	1.46×10^{-4}	5.80×10^{-2}	1.30×10^{0}	2.50×10^{-3}
WOA	7.50×10^{-1}	1.99×10^{3}	1.56×10^{2}	1.49×10^{-2}	1.49×10^{-5}	1.74×10^{0}	4.02×10	1.67×10^{-2}
ISSA	2.17×10^{-4}	6.90×10^{-4}	9.43×10^{-6}	6.35×10^{-4}	8.19×10^{-4}	2.80×10^{-3}	3.66×10^{-5}	5.99×10^{-4}

表2 各算法独立运行25次后获得的统计结果（F_5~F_8）

Table 2 Statistical results acquired by each algorithm after 25 runs per example(F_5-F_8)

算法	μ				σ			
	F_5	F_6	F_7	F_8	F_5	F_6	F_7	F_8
SSA	-2.10×10^{5}	6.72×10^{-7}	1.91×10^{-4}	1.55×10^{-2}	2.30×10^{4}	1.16×10^{-6}	1.98×10^{-4}	5.27×10^{-2}
HHO	-8.31×10^{5}	1.32×10^{-5}	1.42×10^{-2}	5.40×10^{-3}	3.46×10^{4}	2.12×10^{-5}	1.34×10^{-2}	9.10×10^{-3}
AO	-1.28×10^{5}	5.33×10^{-6}	1.77×10^{-2}	2.15×10	3.76×10^{4}	9.32×10^{-6}	4.21×10^{-2}	1.23×10
DOA	-4.68×10^{4}	1.17×10^{0}	2.00×10^{2}	4.53×10	7.49×10^{3}	6.90×10^{-3}	5.30×10^{-3}	2.05×10^{0}
WOA	-6.75×10^{5}	1.17×10^{-1}	7.76×10	1.05×10	1.18×10^{5}	4.34×10^{-2}	2.28×10	9.39×10^{0}
ISSA	-8.38×10^{5}	1.03×10^{-8}	6.26×10^{-6}	7.22×10^{-5}	2.59×10^{-2}	2.58×10^{-8}	1.87×10^{-5}	1.25×10^{-4}

3.2 数据分类

选取开源UCI数据集中11组具有代表性的数据集作为数据分类事例，即Iris、Wine、Haberman、Hayes、Immunotherapy、Cryotherapy、Heart、Breast-Cancer、Transfusion、Audit和Banknote。各数据集的属性个数、样本容量及类别数构成的向量依次为[4,150,3]、[13,178,3]、[3,306,2]、[5,132,3]、[7,90,2]、[6,90,2]、[12,299,2]、[9,116,2]、[4,748,2]、[17,773,2]和[4,1 372,2]。在每个数据集中随机抽取70%的样例构成训练集，其他样例构成测试集。

3.2.1 MIFNN的样本分类性能测试

为测试批量损失函数下基于梯度下降的MGD-MIFNN和ISSA的MIFNN（简记作ISSA-MIFNN）处理以上样本分类问题的性能，一方面选取4种经典神经网络（BPNN、ELM、RBF、WNN）参与MGD-MIFNN的性能比较；另一方面，将以上参与比较的5种智能优化算法在批量损失函数下寻找MIFNN的最优权值和阈值，记此时的算法为X-MIFNN，在此X意指SSA、HHO、AO、DOA、WOA或ISSA。

MIFNN与BP均包含2个隐含层，各自的第1、

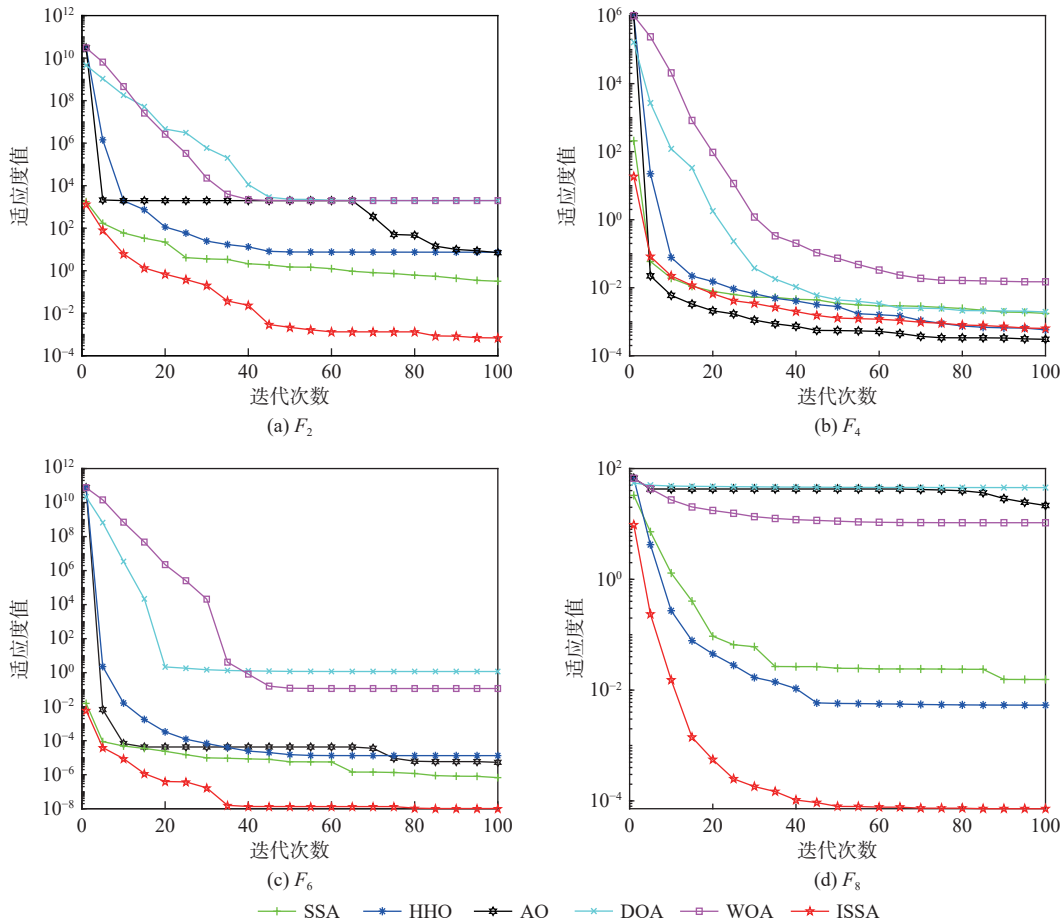


图 2　各算法的平均搜索曲线

Fig. 2　Average search curves of algorithms

第 2 隐含层均分别含 15 个和 5 个神经元, 同时 MIFNN 的每个扩展子层包含的神经元个数 L 取 15 且 MLP 的输入层神经元个数为 5; ELM、RBF 与 WNN 均仅含 1 个隐含层, 其神经元个数为 20; 设置梯度下降学习率为 0.08, 动量因子为 0.05, 非线性激活函数为 tanh 函数, 网络训练的最大迭代次数为 300。另外, ISSA-MIFNN、SSA-MIFNN、HHO-MIFNN、AO-MIFNN、DOA-MIFNN 及 WOA-MIFNN 的种群大小设置为 30, 网络的权值和阈值所在区间为 [−10,10], 其他参数设置同上。各基于梯度下降、智能优化的神经网络获得的训练分类准确率 TrAcc 及测试分类准确率 TeAcc 分别如表 3 和表 4 所示。以 Haberman、Immunotherapy、Cryotherapy 和Transfusion 为例, 各神经网络的损失函数值变化曲线如图 3 所示。

由表 3 和表 4 可知, MGD-MIFNN 与 BP、ELM、RBF、WNN 相比, 其对其中 9 种数据集的训练分类性能最好, 且对其中 8 种数据集的测试性能具有明显优势。然而, 其他神经网络至多求解 3 种数据集的分类问题有优势, 因此, MGD-MIFNN 的优势较为明显, 且适用范围较广。尽管 MGD-MIFNN 已展现出优势, 但其求解诸如 Wine、Immunotherapy、Transfusion 等分类问题的测试分类精度有待提高, 根源在于动量项并不能完全使梯度下降法摆脱陷入局部搜索的困境, 进而导致对特定的分类问题出现过拟合或欠拟合。

与 MGD-MIFNN 相比, ISSA-MIFNN 在数据集的训练分类准确率方面, 对于 Haberman、Heart、Transfusion 分类问题的训练分类精度得到显著提升, 且二者对于 Immunotherapy、Audit 的训练分类准确率均为 100%。ISSA-MIFNN 比 MGD-MIFNN 的测试分类性能有显著性的提升, 且对于 8 种数据集, 获得的测试分类准确率提升较大, 仅求解数据集 Banknote 测试集获得的测试分类准确率提升 1.95%。

由此可知, ISSA 能有效改善 MIFNN 的数据分类性能, 且 ISSA-MIFNN 能有效地回避 MGD-MIFNN 出现欠拟合或过拟合。ISSA-MIFNN 与以上其他基于智能优化的 MIFNN 相比, 数据分类能力较强, 因而其泛化能力较强。具有能求解高维优化问题的 ISSA 有助于提升神经网络的分类性能。由图 3 可知, MGD-MIFNN 的搜索速度快, 收敛精度高, 但容

易致使模型过拟合；BP、ELM、RBF及WNN收敛速度慢，易陷入局部最优；ISSA-MIFNN能有效克服梯度下降易陷入局部搜索的不足，具有平衡局部勘测与全局开采的能力，能快速地获得较为理想的分类效果。基于SSA、HHO、AO及WOA的MIFNN易陷入局部搜索，且局部搜索速度较慢；DOA-MIFNN的收敛速度快，但获得的数据分类效果不理想。由此可知，ISSA-MIFNN能有效兼顾全局与局部搜索，得到的分类效果较好，模型泛化能力强，数据分类能力强。

表3 各神经网络的训练分类准确率比较

Table 3 Comparison of training classification accuracy of neutral networks %

神经网络	TrAcc										
	Iris	Wine	Haberman	Hayes	Immunotherapy	Cryotherapy	Heart	BreastCancer	Transfusion	Audit	Banknote
MGD-MIFNN	100	100	81.30	100	100	100	85.17	100	67.50	100	74.27
BPNN	95.24	95.16	73.36	70.65	80.65	93.55	75.42	88.89	72.08	100	100
ELM	98.10	100	74.77	64.13	85.48	96.77	84.69	90.12	74.95	100	100
RBF	95.24	95.97	73.36	64.13	82.26	80.65	68.42	82.72	77.44	46.03	91.35
WNN	95.24	99.19	78.50	77.17	96.77	98.39	83.73	88.89	67.50	100	99.48
SSA-MIFNN	100	89.52	78.50	86.96	95.16	98.39	74.64	97.53	82.22	100	70.83
HHO-MIFNN	93.33	79.30	74.30	90.22	96.77	95.16	78.95	88.89	77.44	100	56.15
AO-MIFNN	96.19	87.10	82.71	90.22	95.16	93.55	77.51	91.36	74.28	100	66.15
DOA-MIFNN	100	100	83.18	98.91	100	100	86.6	100	86.23	100	70.52
WOA-MIFNN	100	91.13	80.84	94.57	98.39	100	83.73	86.42	84.32	100	67.71
ISSA-MIFNN	99.05	98.39	86.45	97.83	100	98.39	87.08	95.06	87.95	100	73.13

表4 各神经网络的测试分类准确率比较

Table 4 Comparison of test classification accuracy of neutral networks %

神经网络	TeAcc										
	Iris	Wine	Haberman	Hayes	Immunotherapy	Cryotherapy	Heart	BreastCancer	Transfusion	Audit	Banknote
MGD-MIFNN	100	87.04	82.60	82.50	71.43	89.29	85.56	82.86	48.00	100	56.55
BPNN	100	92.59	79.35	42.50	75.00	85.71	83.33	71.43	76.00	100	100
ELM	100	96.30	77.17	65.00	82.14	89.29	81.11	82.86	78.22	99.14	100
RBF	100	87.04	77.17	67.50	75.00	82.14	81.11	68.57	78.67	71.12	92.96
WNN	100	88.89	71.74	52.50	42.86	82.14	76.67	62.86	70.22	100	98.54
SSA-MIFNN	100	77.78	80.43	80.00	75.00	92.86	82.22	78.57	76.00	99.57	57.04
HHO-MIFNN	100	68.52	84.87	85.00	78.57	92.86	87.78	68.57	76.89	99.14	56.80
AO-MIFNN	97.78	83.33	80.43	80.00	78.57	85.71	75.56	64.19	80.44	99.14	55.34
DOA-MIFNN	97.78	88.89	83.70	77.50	82.14	96.43	85.56	82.86	73.78	100	56.07
WOA-MIFNN	97.78	68.52	81.52	75.00	75.00	96.43	78.89	65.71	76.00	100	57.04
ISSA-MIFNN	100	96.30	85.78	90.00	89.29	100	91.11	85.71	80.89	100	58.50

3.2.2 MIFNN的结构变化对分类性能的影响

设置MIFNN的MLP中第1、第2隐含层的神经元个数构成的结构向量为[10,5]、[15,10]或[20,15]。每种设置及其他参数设置不变，以Transfusion和Banknote为例，各神经网络获得的训练、测试分类准确率如表5所示。由表3～表5可知，在MGD-MIFNN中MLP的2个隐含层的神经元个数为15和5的前提下，若减少神经元个数，如10和5，则在Transfusion上的训练和测试分类准确率分别下降12.24%和10.67%，导致欠拟合现象；若神经元个数依次增加到20和15，则在Banknote的训练集上的分类准确率提高0.21%，但在测试集上的分类准确率却下降0.97%，且计算复杂度显著提高，以及出现过拟合现象。另外，随着隐含层的神经元数目逐渐增加，ISSA-MIFNN在Transfusion、Banknote上的训练和测试分类准确率均无明显提升，暗示随着神经元个数的不断增多，将存在冗余神经元，且导致计算成本高。因此，合理设置MIFNN中MLP的隐含层神经元个数，将有助于高效处理多属性数据分类问题。

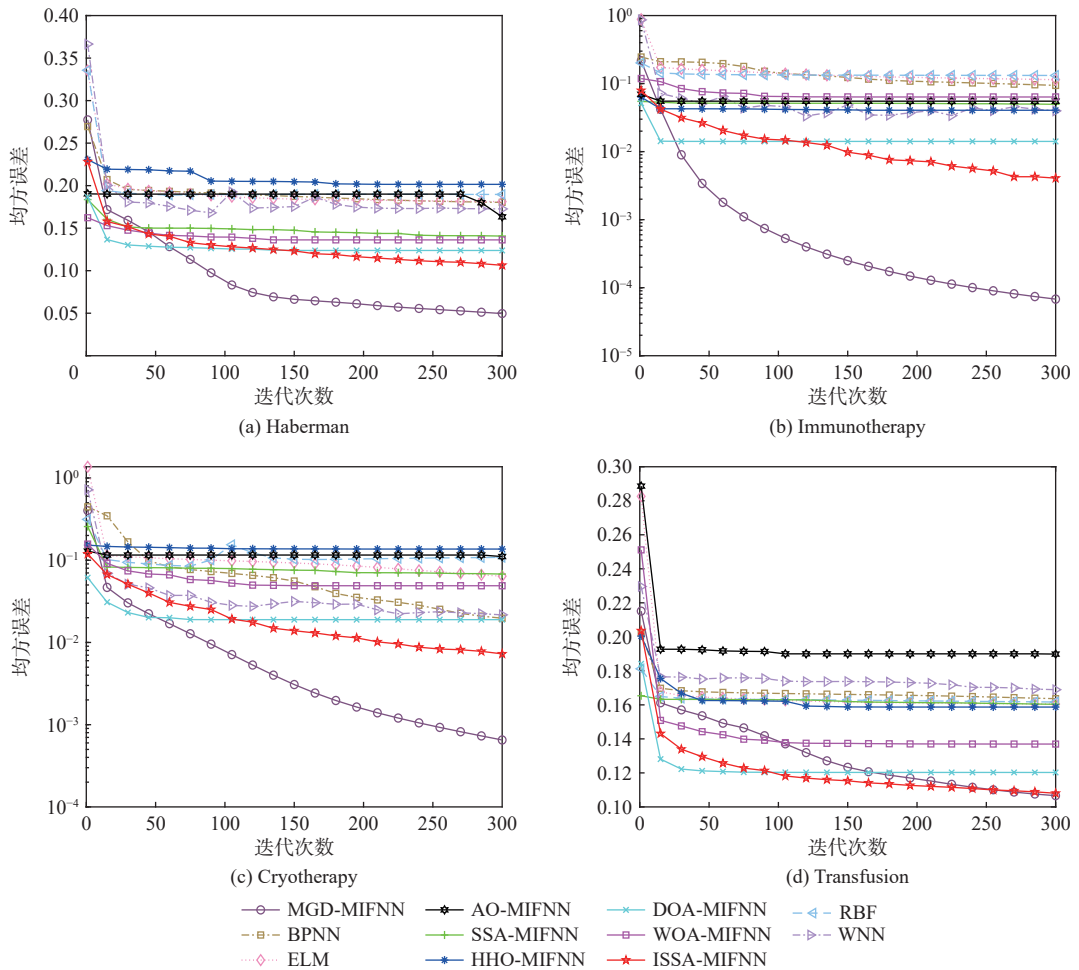


图 3　各神经网络的损失函数变化曲线

Fig. 3　Loss function variation curves of neural networks

表 5　各神经网络在不同网络结构下的分类准确率比较

Table 5　Comparison of neural networks' classification accuracy rates based on different network structures　%

神经网络	TrAcc				TeAcc			
	Transfusion		Banknote		Transfusion		Banknote	
	[10,5]	[20,15]	[10,5]	[20,15]	[10,5]	[20,15]	[10,5]	[20,15]
MGD-MIFNN	55.26	86.81	72.25	77.50	37.33	61.33	56.31	54.13
SSA-MIFNN	81.26	80.88	72.19	76.56	75.56	78.22	56.31	52.67
HHO-MIFNN	81.07	74.19	58.13	70.63	76.44	79.11	56.80	55.83
AO-MIFNN	79.35	73.80	69.17	57.40	75.56	76.89	56.55	57.28
DOA-MIFNN	80.88	87.38	73.33	73.85	76.00	74.22	56.80	57.52
WOA-MIFNN	80.31	83.37	73.75	71.04	75.56	77.33	53.16	59.47
ISSA-MIFNN	82.79	87.57	72.25	77.50	76.44	79.11	57.28	58.50

4　结　论

鉴于单输入 FNN 的应用范围窄及多层感知器对含噪声的输入数据的灵敏度高等问题, 本文通过设计能提取多源信息特征的堆叠 FNN, 并将其与多层感知器融合, 获得 MIFNN。进而, 采用含动量项的梯度下降法或 ISSA 优化网络的权值与阈值。理论分析和实验验证表明:

1) MGD-MIFNN 的搜索速度快, 收敛精度高, 在中小规模数据集上的分类准确率高;

2) ISSA 相较于参与比较的元启发式随机搜索算法, 处理高维优化问题具有明显优势, 其收敛速度快, 得到的解质量高;

3) ISSA-MIFNN 的模型泛化能力强且数据分类准确率高, 能有效回避过拟合和欠拟合。

总之, MGD-MIFNN 是一种具有明显竞争力的新型神经网络, 为数据分类问题的解决提供了又一新途径, 但对于大样本分类问题, 仍存在模型过拟

合且分类准确率不高的不足；ISSA求解高维优化问题已有明显优势，其与MIFNN结合获得的ISSA-MIFNN求解数据分类问题，已展现独特优势，但处理个别数据分类问题的能力有待提高。下一步将进一步研究MGD-MIFNN及ISSA-MIFNN处理大样本图像分类、高维曲面拟合的能力。

参考文献（References）

[1] 杨旭华. 神经网络及其在控制中的应用研究[D]. 杭州: 浙江大学, 2004: 24-32.
YANG X H. Study on neural networks machine and its application in control[D]. Hangzhou: Zhejiang University, 2004: 24-32(in Chinese).

[2] 杨晓帆, 陈廷槐. 人工神经网络固有的优点和缺点[J]. 计算机科学, 1994, 21(2): 23-26.
YANG X F, CHEN T H. Advantages and disadvantages inherent in artificial neural networks[J]. Computer Science, 1994, 21(2): 23-26(in Chinese).

[3] 邹阿金, 沈建中. 傅立叶神经网络建模研究[J]. 湘潭大学自然科学学报, 2001, 23(2): 23-26.
ZOU A J, SHEN J Z. A neural network model structure study based on Fourier[J]. Natural Science Journal of Xiangtan University, 2001, 23(2): 23-26(in Chinese).

[4] GASHLER M S, ASHMORE S C. Modeling time series data with deep Fourier neural networks[J]. Neurocomputing, 2016, 188(5): 3-11.

[5] 党选举, 潘登. 基于电流注入法的开关磁阻电机转矩脉动抑制[J]. 组合机床与自动化加工技术, 2021(5): 115-119.
DANG X J, PAN D. Torque ripple reduction of switched reluctance motor based on current injection method[J]. Modular Machine Tool & Automatic Manufacturing Technique, 2021(5): 115-119(in Chinese).

[6] BENRABAH M, KARA K, AITSAHED O, et al. Adaptive Fourier series neural network PID controller[J]. International Journal of Control, Automation and Systems, 2021, 19(10): 3388-3399.

[7] PENG P, XIE L, WEI H. A deep Fourier neural network for seizure prediction using convolutional neural network and ratios of spectral power[J]. International Journal of Neural Systems, 2021, 31(8): 2150022.

[8] 张雨浓, 李钧, 张智军, 等. SIMO傅里叶三角基神经网络的权值直接确定法和结构自确定算法[J]. 信息与控制, 2011, 40(4): 507-513.
ZHANG Y N, LI J, ZHANG Z J, et al. A weights-direct-determination method and structure-automatic-determination algorithm for SIMO trigonometrically-activated Fourier neural networks[J]. Information and Control, 2011, 40(4): 507-513(in Chinese).

[9] LIN L, WU X H, QI J, et al. Power quality disturbance classification based on a novel Fourier neural network and hyperbolic S-transform[J]. International Journal of Signal Processing, Image Processing and Pattern Recognition, 2016, 9(1): 111-124.

[10] 龚鹃, 邓方, 陈杰. 基于自适应差分进化算法和傅里叶神经网络的电子罗盘的标定[C]//中国自动化学会控制理论专业委员会B卷. 北京: 中国自动化学会控制理论专业委员会, 2011: 741-746.
GONG P, DENG F, CHEN J. Calibration of electronic compass based on adaptive differential evolution algorithm and Fourier neural network[C]// Technical Committee on Control Theory, Chinese Association of Automation, Vol. B. Beijing: Technical Committee on Control Theory, Chinese Association of Automation, 2011: 741-746(in Chinese).

[11] 陈诗雨, 李小勇, 杜杨杨, 等. Fourier神经网络非线性拟合性能优化研究[J]. 武汉大学学报 (工学版), 2020, 53(3): 277-282.
CHEN S Y, LI X Y, DU Y Y, et al. Optimization study of Fourier neural network nonlinear fitting performance[J]. Engineering Journal of Wuhan University, 2020, 53(3): 277-282(in Chinese).

[12] 薛建凯. 一种新型的群智能优化技术的研究与应用[D]. 上海: 东华大学, 2020.
XUE J K. Research and application of a novel swarm intelligence optimization technique[D]. Shanghai: Donghua University, 2020(in Chinese).

[13] HEIDARI A A, MIRJALILI S, FARIS H, et al. Harris hawks optimization: Algorithm and applications[J]. Future Generation Computer Systems, 2019, 97(8): 849-872.

[14] ABUALIGAH L, YOUSRI D, ABD ELAZIZ M, et al. Aquila optimizer: A novel meta-heuristic optimization algorithm[J]. Computers & Industrial Engineering, 2021, 157: 107250.

[15] BAIRWA A K, JOSHI S, SINGH D. Dingo optimizer: A nature-inspired metaheuristic approach for engineering problems[J]. Mathematical Problems in Engineering, 2021, 2021(26): 2571863.

[16] MIRJALILI S, LEWIS A. The whale optimization algorithm[J]. Advances in Engineering Software, 2016, 95: 51-67.

[17] 李爱莲, 全凌翔, 崔桂梅, 等. 融合正余弦和柯西变异的麻雀搜索算法[J]. 计算机工程与应用, 2022, 58(3): 91-99.
LI A L, QUAN L X, CUI G M, et al. Sparrow search algorithm combining sine-cosine and cauchy mutation[J]. Computer Engineering and Applications, 2022, 58(3): 91-99(in Chinese).

[18] 付华, 刘昊. 多策略融合的改进麻雀搜索算法及其应用[J]. 控制与决策, 2022, 37(1): 87-96.
FU H, LIU H. Improved sparrow search algorithm with multi-strategy integration and its application[J]. Control and Decision, 2022, 37(1): 87-96(in Chinese).

[19] YANG Q, CHEN W N, GU T, et al. A distributed swarm optimizer with adaptive communication for large-scale optimization[J]. IEEE Transactions on Cybernetics, 2020, 50(7): 3393-3408.

[20] LI D, GUO W, LERCH A, et al. An adaptive particle swarm optimizer with decoupled exploration and exploitation for large scale optimization[J]. Swarm and Evolutionary Computation, 2021, 60(7): 100789.

[21] ZHANG Z H, LI L, LU J X. Gradient-based fly immune visual recurrent neural network solving large-scale global optimization[J]. Neurocomputing, 2021, 454: 238-253.

[22] ZHANG Z H, XIAO T Y, QIN X C. Fly visual evolutionary neural network solving large-scale global optimization[J]. International Journal of Intelligent Systems, 2021, 36(11): 6680-6712.

[23] 徐辰华, 李成县, 喻昕, 等. 基于Cat混沌与高斯变异的改进灰狼优化算法[J]. 计算机工程与应用, 2017, 53(4): 1-9.
XU C H, LI C X, YU X, et al. Improved grey wolf optimization algorithm based on chaotic Cat mapping and Gaussian mutation[J]. Computer Engineering and Applications, 2017, 53(4): 1-9(in

Chinese).
[24] 杨彤. 改进鲸鱼算法在函数优化中的应用[D]. 西安: 西安电子科技大学, 2020: 15-17.
YANG T. Applications of improved whale optimization algorithm in function optimization[D]. Xi'an: Xidian University, 2020: 15-17(in Chinese).

Multi-input Fourier neural network and its sparrow search optimization

LI Liangliang, ZHANG Zhuhong*, ZHANG Yongdan

(Guizhou Provincial Characteristic Key Laboratory of System Optimization and Scientific Computation, College of Big Data and Information Engineering, Guizhou University, Guiyang 550025, China)

Abstract: In engineering applications, the back-propagation (BP) neural network often encounters many limitations due to its slow convergence and high noise sensitivity, and the reported Fourier neural networks cannot extract the features of multi-attribute input data. Hereby, this work proposes a gradient descent-based multi-input Fourier neural network after integrating the multi-layer perceptron with an overlapping Fourier neural network. Then to address the difficulty in deciding the global optimal parameter settings, the Cat chaotic map and the mechanisms of population-size adjustment and parameter adaptiveness are designed to promote the sparrow search algorithm's ability to balance global exploration and local exploitation. An improved sparrow search algorithm is thus developed, optimizing the parameter settings and solving high dimensional function optimization problems. The theoretical analysis shows that the improved algorithm's computational complexity is decided by its population size and the optimization problem dimension. Numerically comparative experiments have validated that the acquired Fourier neural network can effectively extract the features of multi-attribute data with strong generalization ability, and that the improved algorithm has significant advantages in coping with high dimensional function optimization problems.

Keywords: Fourier neural network; multi-layer perceptron; sparrow search; high dimensional function optimization; multi-attribute classification

Received: 2022-05-21; **Accepted**: 2022-11-25; **Published Online**: 2022-12-14 09:27
URL: link.cnki.net/urlid/11.2625.V.20221213.1513.006
Foundation item: National Natural Science Foundation of China (62063002)
* **Corresponding author**. E-mail: zhzhang@gzu.edu.cn

http://bhxb.buaa.edu.cn jbuaa@buaa.edu.cn
DOI: 10.13700/j.bh.1001-5965.2022.0412

基于 Transformer 的点云几何有损压缩方法

刘舸昕，章骏腾，丁丹丹*
(杭州师范大学 信息科学与技术学院，杭州 311121)

摘 要：点云被广泛地用于三维物体表达，不过真实世界采集到的点云往往数据庞大，不利于传输与储存，针对点云数据冗余性问题，引入基于注意力机制的 Transformer 模块，提出一种基于 Transformer 的端到端多尺度点云几何压缩方法。将点云进行体素化，在编码端利用稀疏卷积提取特征，进行多尺度的逐步下采样，结合 Transformer 模块加强点空间特征感知与提取；在解码端进行对应的多尺度上采样重建，同样采用 Transformer 模块对有用特征进行加强与恢复，逐步细化并重建点云。与 2 种点云标准编码方法对比，所提方法平均获得 80% 和 75% 的 BD-Rate 增益；与基于深度学习的点云压缩方法对比，平均获得 16% 的 BD-Rate 增益，在相同码率点有约 0.6 的 PSNR 提升。实验结果表明：Transformer 在点云压缩领域的可行性与有效性；在主观质量方面，所提方法也有明显的主观效果提升，重建的点云更接近原始点云。

关 键 词：点云几何压缩；Transformer；注意力机制；深度学习；稀疏卷积
中图分类号：TP391
文献标志码：A **文章编号**：1001-5965(2024)02-0634-09

点云是表示三维物体形状、颜色、反射强度等的数据集合，如今已经成为描述三维物体形状和结构的主流方式，被广泛应用于自动驾驶、虚拟现实等领域。现实应用中采集到的点云往往包含数量庞大的点，存在大量冗余，如果直接存储和传输未经压缩的点云需要消耗大量的内存和带宽，点云压缩能够有效解决上述问题。

传统点云压缩方法中，比较具有代表性的是文献 [1] 提出的基于几何的点云压缩(geometry-based point cloud compression, G-PCC)方法，基于视频的点云压缩(video-based point cloud compression, V-PCC)方法。其中，G-PCC 通常用于静态点云的压缩，V-PCC 则用于动态点云的压缩。以上 2 种方法利用八叉树分解、二维到三维的投影、三角曲面模型等传统方法，能够完成点云的无损和有损压缩。

随着深度学习的迅速发展，学者们提出基于深度学习的点云压缩方法并获得比传统方法更高的增益。文献 [2-5] 使用三维卷积基于自动编码器实现端到端的点云几何压缩，相比于 G-PCC 明显提升了压缩性能，但是由于体素化后的点云有大量空体素存在，在全部体素上使用三维卷积极大地增加了计算复杂度。Wang 等[6] 提出基于稀疏卷积和多尺度下采样的点云几何压缩(point cloud geometry compression, PCGCv2)方法，有效降低了计算复杂度，但仍然没有充分利用点云的几何关系，尚有巨大的性能提升潜力。同时，随着 Transformer 技术在图像、视频处理任务中崭露头角，学者们尝试将 Transformer 嵌入到点云任务中提升性能。文献 [7-8] 将 Transformer 运用在点云形状分类、语义场景分割等方面，取得了更加准确的结果。

受到上述工作启发，本文从密集点云的有损压缩出发，基于 PCGCv2 整体框架，引入一种 Trans-

收稿日期：2022-05-26；**录用日期**：2022-07-02；**网络出版时间**：2022-11-04 16:54
网络出版地址：link.cnki.net/urlid/11.2625.V.20221104.1438.003
基金项目：国家自然科学基金 (6217010731)；浙江省自然基金 (LY20F010013)
* **通信作者**. E-mail：DandanDing@hznu.edu.cn

引用格式：刘舸昕，章骏腾，丁丹丹. 基于 Transformer 的点云几何有损压缩方法 [J]. 北京航空航天大学学报，2024，50（2）：634-642.
LIU G X，ZHANG J T，DING D D. Lossy point cloud geometry compression based on Transformer[J]. Journal of Beijing University of Aeronautics and Astronautics，2024，50（2）：634-642（in Chinese）.

former 模块完成基于多尺度的点云压缩。具体地，将点云体素化后，编码端对输入的点云进行下采样，同时用 Transformer 和卷积网络逐层提取并加强点云的有用特征；然后，对下采样的坐标及特征进行编码，完成压缩；解码端通过对应的上采样，配合基于 Transformer 的特征重建来预测体素概率，从而逐步完成点云的重建。在动态图像专家组（moving pictures experts group, MPEG）标准的数据集上对本文方法进行测试，结果表明，本文方法在客观指标与主观质量上均显著高于当前 G-PCC（Octree）、G-PCC（Trisoup）、基于深度学习的自适应点云几何编码（adaptive deep learning-based point cloud geometry coding, ADL-PCC）[2]、PCGCv2[6] 等对比方法。

1　相关工作

1.1　点云压缩

点云几何压缩分为无损压缩和有损压缩。无损压缩旨在提高压缩率并减少编解码时间，而有损压缩旨在提高压缩率，提升点云的重建质量。

MPEG 提出的 G-PCC、V-PCC 编码器是目前主流的点云压缩方法。G-PCC 包含 2 种编码器，Trisoup 几何编码利用修剪的八叉树和表面模型表示点云，只能用于静态点云的有损压缩；Octree 几何编码用完整的八叉树表示点云，八叉树的最下层叶子节点只表示一个单独的体素，可用于静态点云的无损压缩和有损压缩。V-PCC 编码器将三维点云投影到二维平面上，再使用已经成熟的高效视频编码（high-efficiency video coding, HEVC）方法进行编码，适用于动态点云和静态点云的无损及有损压缩。

近年来，随着深度学习方法的迅速发展，深度学习逐渐被应用于点云压缩领域。基于深度学习方法的点云几何无损压缩通过神经网络准确预测点云中每一个点的概率，帮助熵编码压缩点云，预测越准确则压缩率越高。Huang 等[9] 提出 OctSqueeze 基于八叉树结构来表示大型雷达点云，首次运用深度学习方法训练树结构的熵模型，利用熵模型对点云进行压缩，减小了雷达点云的传输带宽。但是通过树结构构建的上下文信息只利用了父节点的信息，还没有充分利用邻域信息。Que 等[10] 提出 VoxelContext-Net 使用八叉树体素混合方式进行点云压缩，基于八叉树表示点云，并且使用体素化的方法获取每个体素周围的邻居信息，利用邻居信息预测当前点的占用概率，进一步提升了点云熵编码的压缩率。Nguyen 等[11] 提出的 Voxel-DNN 基于体素表示点云，用自回归先验的方式逐步预测每一个体素块的概率，提升熵编码过程中的概率预测准确性，提升了点云无损压缩率，但是编解码时间较长。Nguyen 等[12] 后续提出的 MSVoxel-DNN 方法，使用了多尺度的方法，把低尺度体素作为先验，帮助预测高尺度体素的概率，并且将体素的概率预测分为多组，有效缩短了编码时间过长的问题。

在点云几何有损压缩领域，基于深度学习方法通过神经网络提取点与点之间的特征，利用特征重建点云，特征提取越充分则重建效果越好。Yan 等[4] 提出基于点的特征提取方法，构建了自动编码器结构的网络结构。Huang 和 Liu[13] 使用多尺度损失函数，提升重建效果。Guo 等[8] 提出一个基于八叉树的自适应网络，能够处理大尺度点云。但是以上基于点的方法都没有充分提取特征信息，并且重建效果还有待提高。Wang 等[3] 提出基于体素的方法 PCGCv1，进行端到端的点云有损压缩，点云体素化后利用三维卷积提取点与点之间的特征信息，同时用自分编码器结构，作为熵编码先验，提升压缩率。Guarda 等[2] 提出 ADL-PCC 方法，创新了体素化点云自适应分割方式，让点云前处理过程更加合理，减少后续三维卷积的计算复杂度，但是编解码时间较长。Quach 等[14] 基于自分编码器结构，在解码端裁剪点云进行细化时设定了自适应阈值，提升重建效果。Wang 等[6] 提出 PCGCv2 多尺度的点云压缩方法，使用稀疏卷积在体素块上进行卷积，相比三维卷积降低了计算复杂度，减少了计算时间；此外，PCGCv2 方法还进行了多尺度的下采样和上采样，逐步精细的恢复重建点云，以提升重建效果。PCGCv2 方法通过稀疏卷积的方式获得特征信息，有效降低了复杂度，并获得显著性能提升，因此，本文以 PCGCv2 为基础，在其原有架构基础上继续探索有用特征的提取与加强方法，以获得更高的压缩性能。

1.2　Transformer

Transformer 自诞生以来，已经在自然语言处理、图像处理等领域取得了巨大的成果。Transformer 摒弃了传统的卷积神经网络（convolutional neural networks, CNN）和循环神经网络（recurrent neural network, RNN）的做法，输入被嵌入为特征，再进行线性变换得到 3 个矩阵，通过自注意力机制获取相关性，特征元素之间相关性高的被赋予更高的权重，反之赋予低的权重。

文献 [15] 提出 Transformer 结构，并运用在机器

翻译领域，通过自注意力机制获取未翻译文本和已翻译前文之间的关系，赋予不同权重，提升翻译准确性。文献 [16] 将 Transformer 运用在图像领域，图像被分割为多个块，每个块进行语义嵌入，利用注意力机制得到各个块之间的相关性，提升图像分类准确度。文献 [17] 提出基于图像的 Transformer，利用局部注意力机制进一步获得局部的特征信息，让 Transformer 同时提取到全局特征与局部特征信息，克服注意力机制缺少局部特征的不足。文献 [18] 提出基于窗口转移注意力机制的 Swin Transformer，获取局部特征信息的同时极大减少了计算复杂度，Transformer 在图像领域真正具备了可用性，同时可在不同尺度获取注意力图，获取全局特征信息。

在点云任务方面，文献 [7] 提出 Point Transformer，将 Transformer 用于点云领域，提出向量注意力，相对于自注意力减少计算复杂度，在点云分类、点云场景语义分割等任务上优于以往方法。文献 [8] 提出 Point Cloud Transformer 通过邻域特征提取模块获取点云内部的邻域关系，并且提出偏移注意力机制，其中包含自注意力模块和偏移模块，计算自注意力模块的输入和输出之间的偏移，通过偏移模块得到新的特征，在点云分类、语义分割上取得高准确度。文献 [19] 提出基于 Transformer 的点云上采样方法 PU-Transformer，使用移动通道多头自注意力机制，相比自注意力减少了计算复杂度，同时更好地提取点与点之间的特征，在点云补全可视化效果上取得了优秀的结果。文献 [20] 提出 Oct-Attention，通过八叉树的方法表示点云，利用注意力机制提取八叉树中的领域信息，针对大尺度的雷达点云进行无损压缩，提升了压缩性能，缺点是解码时间较长。

上述方法表明，基于神经网络的点云压缩方法相比传统方法有显著优势，而 Transformer 能够进一步提升网络对特征的感知能力，从而获得更好的效果，在点云压缩领域，目前有点云无损压缩方法结合了八叉树与 Transformer，Transformer 模块在其中发挥增强特征信息的作用，达到更高的压缩率。但是 Transformer 与体素相结合的压缩方法还有待探索。受此启发，本文将 Transformer 模块引入点云压缩任务中，并且结合体素化的方法，加强特征提取环节，探索如何利用注意力机制获得更高压缩效率。

2 本文方法

2.1 网络结构

如图 1 所示，本文方法的结构由编码器、解码器和熵编码组成，“$\boldsymbol{C}_X$”为点云的几何坐标信息，“ReLU”为激活函数，“Conv$c\times n^3$”表示输出通道为 c，卷积核为3×3×3的稀疏卷积，“IRN”表示 Inception-Residual 模块，“Q”表示量化，“AE”、“AD”为算术编解码。在编码端，对点云 $\boldsymbol{X}$进行稀疏卷积逐步下采样得到多尺度的点云$\{\boldsymbol{X}_1,\boldsymbol{X}_2,\boldsymbol{Y}\}$；在瓶颈层，$\boldsymbol{Y}\{\boldsymbol{C}_Y,\boldsymbol{F}_Y\}$的几何信息 $\boldsymbol{C}_Y$进行八叉树无损编码，经过量化的特征信息 $\hat{\boldsymbol{F}}_Y$由熵模型进行有损编码得到 $\hat{\boldsymbol{F}}'_Y$，在解码端 $\boldsymbol{C}_Y$、$\hat{\boldsymbol{F}}'_Y$进行对应的多尺度上采样，逐步细化重建点云最终得到$\boldsymbol{C}'_X$。

具体地，在前处理过程中，点云 $\boldsymbol{X}$用稀疏张量 $[\boldsymbol{C}_X,\boldsymbol{F}_X]$ 进行表示，其中 $\boldsymbol{C}_X$为点云的坐标，$\boldsymbol{F}_X$为坐标对应的体素占用信息，用稀疏卷积进行处理可以极大降低计算复杂度，提升计算速度。

在编码端，首先，通过步长为 1，卷积核大小为3×3×3的稀疏卷积对输入点云进行初步特征信息提取；然后，通过步长为 2，卷积核大小为3×3×3的

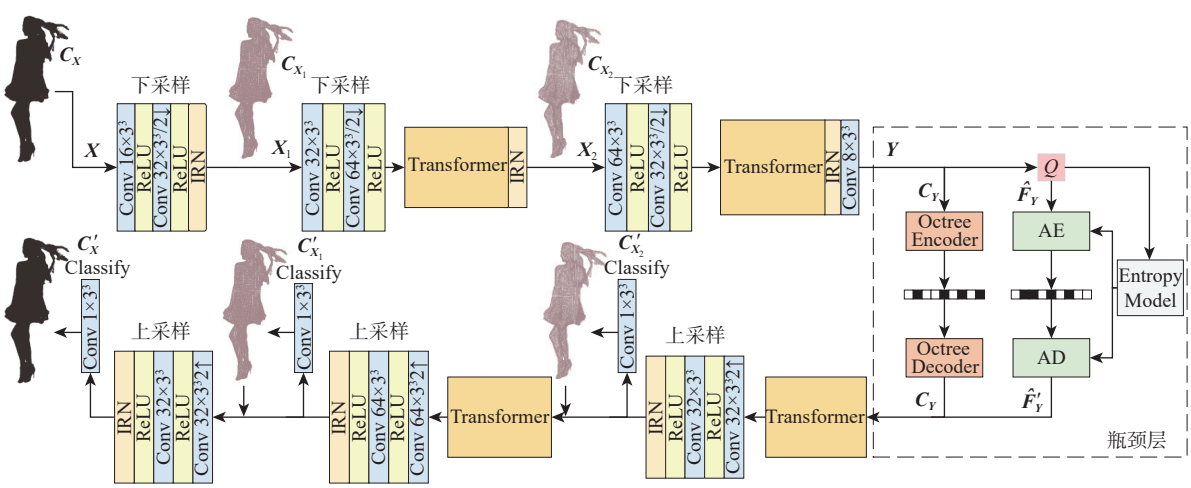


图 1 本文方法的整体结构

Fig. 1 Overall structure of proposed method

稀疏卷积进行第 1 次下采样, 并且通过 IRN 模块加强特征信息。其中, IRN 模块结构来自 PCGCv2 的设计, 包含 3 次连续的 Inception 模块和残差模块。对下采样结果 $\boldsymbol{X}_1=[\boldsymbol{C}_{X_1},\boldsymbol{F}_{X_1}]$ 进行第 2 次下采样, 不同的是在第 2 次下采样后使用 Transformer 模块和 IRN 模块共同提取特征信息, 得到 $\boldsymbol{X}_2=[\boldsymbol{C}_{X_2},\boldsymbol{F}_{X_2}]$, Transformer 包括 2 个连续的偏移注意力模块, 充分提取特征信息。随后进行第 3 次下采样, 结构与第 2 次相同, 得到 $\boldsymbol{Y}=[\boldsymbol{C}_Y,\boldsymbol{F}_Y]$。

在瓶颈层, 坐标信息 $\boldsymbol{C}_Y$ 通过 G-PCC 进行无损编码(坐标信息的码率约为 0.025), 在编解码过程中, 特征信息 $\boldsymbol{F}_Y$ 先经过量化得到 $\hat{\boldsymbol{F}}_Y$, 利用 $\hat{\boldsymbol{F}}_Y$ 构建一个正态分布的条件概率模型作为熵模型, 再由熵模型对其进行算术编解码。在训练过程中, 需要通过反向传播实现梯度下降, 优化网络模型中的权重。对 $\boldsymbol{F}_Y$ 进行量化会导致反向传播中的梯度为 0, 所以在训练时本文使用随机噪声代替量化过程, 以保证反向传播的可微性。

在解码端, 通过 Transformer 模块进行特征提取, 随后进行第 1 次上采样, 再通过 IRN 模块进一步提取特征, 上采样过后的点云需要进一步细化, 如图 1 Classify 模块所示, 提取的特征信息经过卷积核为 3、输出通道为 1 的卷积得到每个体素的概率, 选取概率最高的 K 个点保留, K 为编码端同尺度点云的点数。在训练时, 根据编码端的 Ground Truth 和解码端的概率共同选取体素, 在测试时, 只根据解码端的概率选取体素, 去除上采样过后概率低的体素。通过这种方式保证每一次重建点云的质量。随后第 2 次上采样与第 1 次相同。与编码端对应, 解码端的第 3 次上采样只包括一个 IRN 模块提取特征, 得到重建点云 $\boldsymbol{C}'_X$。

Transformer 模块包括 2 个连续的偏移注意力机制, 偏移注意力机制结构如图 2 所示。输入 $\boldsymbol{F}_X$ 为点云特征信息, 偏移注意力模块包括自注意力模块和偏移模块。自注意力模块中 $\boldsymbol{F}_X$ 线性变换得到 $\boldsymbol{Q}$、$\boldsymbol{K}$、$\boldsymbol{V}$ 这 3 个矩阵, $\boldsymbol{Q}$ 矩阵转置后与 $\boldsymbol{K}$ 矩阵叉乘得到注意力图(attention map, AM), 注意力图与矩阵 $\boldsymbol{V}$ 叉乘得到注意力机制增强后的注意力特征(attention features, AF); 偏移模块中注意力特征与输入作减法得到残差, 经过卷积、批归一化、ReLU 激活后与输入特征 $\boldsymbol{F}_X$ 相加, 得到输出 $\boldsymbol{F}'_X$。偏移注意力机制可以增强特征, 在编码端充分提取特征, 在解码端进行更准确的概率预测, 去除概率低的体素, 在重建阶段精确地进行点云细化, 最终获得更好的重建效果。

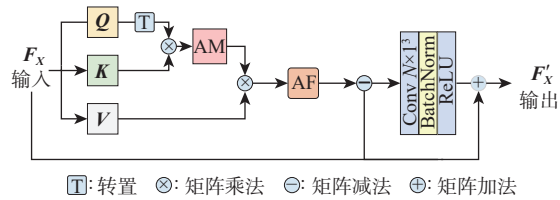


图 2　偏移注意力整体结构

Fig. 2　Overall structure of offset attention

2.2　损失函数

点云有损压缩可以提高压缩率。以提升点云重建质量为目的, 在训练时使用比特率-失真(R-D)权衡损失函数, 由二进制交叉熵(binary cross entropy, BCE)和压缩比特率组成, R-D 总损失函数为:

$$J_{\text{loss}}=R+\lambda D \tag{1}$$

J_{loss} 由 R 和 D 这 2 部分组成, R 为点云压缩比特率, D 为点云重建质量; 参数 λ 用来控制码率, 可以设置为 0.025 到 10 之间, 训练得到不同码率的多个模型。如式 (2) 所示, 其中 $\hat{\boldsymbol{F}}_Y$ 为瓶颈层中点云特征图, $p_{\hat{F}_Yi}$ 为每个体素对应的预测概率, $R_{\hat{F}_Y}$ 为利用特征图计算码率的大小。

$$R_{\hat{F}_Y}=\frac{1}{N}\sum_{i=1}^{N}-\log_2(p_{\hat{F}_Yi}) \tag{2}$$

式 (3) 为损失函数 L_{BCE}, 其中 x_i 表示当前体素的真实占用情况(1 表示占用, 0 表示未占用), p_i 为网络预测的体素占用概率, 每个体素预测的概率 p_i 越接近真实占用情况 x_i, 则 L_{BCE} 的值越小, 优化 L_{BCE} 可得到更高质量的重建点云。

$$L_{\text{BCE}}=\frac{1}{N}\sum_{i=1}^{N}-[x_i\log_2(p_i)+(1-x)\log_2(1-p_i)] \tag{3}$$

3　实验结果及设置

3.1　实验设置

本文实验的训练集使用 ShapeNet 制作, ShapeNet 数据集内包括了约 51 300 个 CAD 表面模型。训练集的制作过程如下: 从 ShapeNet 密集点云数据集中随机挑选 20 000 个进行采样, 将得到的点云经过随机旋转, 量化三维坐标至 7 bit。在训练过程中从 20 000 个训练集中随机选出 2 000 个作为验证集。如表 1 所示, 本文测试序列选用了 12 个密集点云包括 8i Voxelized Full Bodies (8iVFB)数据集中的 Longdress、Loot、Redandblack、Soldier; Microsoft Voxelized Upper Bodies (MVUB)数据集中的 Andrew、David、Phil、Sarah; Owlii dynamic human mesh (Owlii)数据集中的 Basketball_player、Dancer、Exercise、Model。其中 Soldier 经过下采样, 具体做法为

将坐标乘以0.7再四舍五入之后，去除其中相同的点。Basketball_player、Dancer、Exercise、Model下采样倍率为0.375。MPEG已将这些数据集作为点云压缩的标准测试序列，更符合点云压缩领域的实际应用。

表1 实验所用的测试序列及点数

Table 1 Test sequences and point counts used in experiment

序列集合	测试序列	下采样倍率	点数
MVUB	Andrew		279 664
	David		330 797
	Phil		356 258
	Sarah		299 363
8iVFB	Longdress		857 966
	Loot		805 285
	Redandblack		757 691
	Soldier	0.7	569 520
Owlii	Basketball_player	0.375	456 455
	Dancer	0.375	402 337
	Exercise	0.375	369 267
	Model	0.375	375 392

在上述数据集中进行测试，与G-PCC(Trisoup)编码器、G-PCC(Octree)编码器、ADL-PCC及PCGCv2对比，计算复杂度和主观效果。

所使用的实验平台为Intel Core i7-9700 CPU，32 GB内存，NVIDIA GeForce RTX 3090GPU。训练轮数为50，总时间为60 h，初始学习率为8×10^{-4}，逐轮衰减，每一轮的衰减系数为0.5，到10^{-5}为下限，批量大小为1。

3.2 实验对比

如表2和表3所示，其中，PSNR D1为点到点PSNR，PSNR D2为点到面PSNR，本文方法对比传统方法和深度学习方法有显著BD-Rate增益，相对G-PCC(Octree)和G-PCC(Trisoup)编码器在PSNR D1、PSNR D2上平均约80%和75%的BD-Rate增益，相对基于深度学习的方法如PCGCv2平均约16%的BD-Rate增益，在单个数据集上也有较多的BD-Rate增益。

率失真(rate distortion, RD)曲线如图3和图4所示，相比G-PCC编码方法，本文方法在PSNR D1与PSNR D2上均有显著提升。相比基于深度学习方法的ADL-PCC，本文方法在低码率点如点深度为0.3有约2 dB的PSNR提升，高码率点如点深度为0.6有约1 dB的PSNR提升。相比基于深度学习方法的PCGCv2，本文方法在低码率点如点深度为0.05，PSNR D1提升0.65 dB，PSNR D2提升0.24 dB，在略高的码率例如点深度为0.25、0.5、0.6，PSNR D1和PSNR D2都有约0.6 dB的提升。以上实验结果表明Transformer在点云有损压缩领域的可行性和有效性。

表2 不同对比方法在客观指标PSNR D1、PSNR D2上的平均BD-Rate增益

Table 2 Average BD-Rate gain on objective metrics PSNR D1, PSNR D2 for different comparison methods %

序列集合	BD-Rate增益					
	G-PCC(Octree)		G-PCC(Trisoup)		PCGCv2	
	PSNR D1	PSNR D2	PSNR D1	PSNR D2	PSNR D1	PSNR D2
8iVFB	−91.09	−86.90	−90.75	−90.71	−17.43	−18.11
MVUB	−91.59	−87.02	−91.47	−90.91	−14.21	−15.53
Owlii	−73.80	−53.85	−47.74	−43.89	−15.02	−15.42

表3 不同对比方法在单个数据集客观指标PSNR D1、PSNR D2上的BD-Rate增益

Table 3 BD-Rate gain on objective metrics PSNR D1, PSNR D2 for different comparison methods on a single dataset %

测试序列	BD-Rate增益					
	G-PCC(Octree)		G-PCC(Trisoup)		PCGCv2	
	PSNR D1	PSNR D2	PSNR D1	PSNR D2	PSNR D1	PSNR D2
Longdress	−91.61	−88.09	−91.34	−91.84	−19.13	−17.93
Loot	−92.07	−89.05	−93.20	−93.23	−27.32	−28.64
Redandblack	−90.36	−86.99	−89.45	−90.66	−13.82	−15.08
Soldier	−90.31	−83.45	−89.01	−87.09	−9.45	−10.78

如图5所示，选取MPEG和8iVFB测试序列中David、Longdress数据集进行不同方法的可视化效果对比。在相近的码率点下，本文方法相比G-PCC(Octree)和G-PCC(Trisoup)的可视化效果有显著的提升，相比PCGCv2和ADL-PCC保留了更多细节，同时整体结构更接近原始点云。实验表明，本文方法在整体轮廓、局部细节等方面都更接近原始点云，拥有更好的主观效果，充分证明了Transformer在点云有损压缩领域的有效性。

如表4所示，在MPEG标准测试序列中的MVUB数据集上对比各种方法的平均编解码时间，G-PCC(Octree)平均编解码时间最快，编码时间0.73 s，解码时间0.07 s。本文方法编码时间为1.08 s，解码时间为0.71 s，相比PCGCv2分别慢了0.55 s和0.53 s，相比G-PCC(Trisoup)快约2倍，ADL-PCC平均编解码时间最慢。实验证明Transformer模块的引入可提升客观指标并改善主观效果，但也会带来计算复杂度与编解码时间的增加。

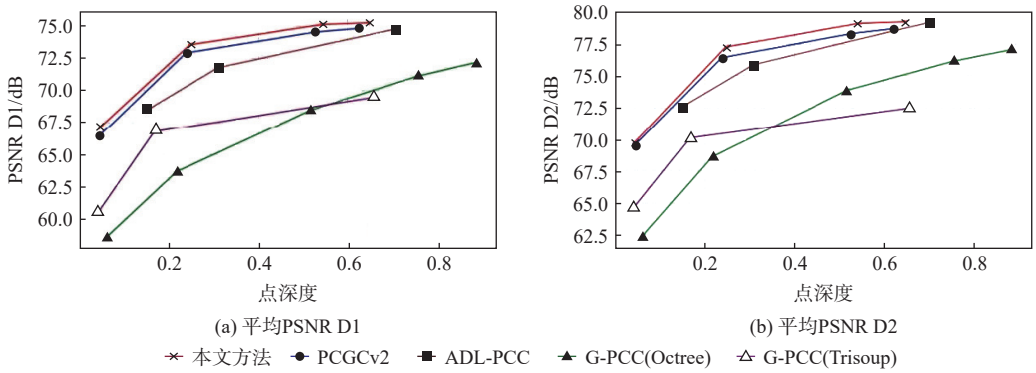


图 3　对比不同方法在客观指标 PSNR D1、PSNR D2 上的平均 RD 曲线

Fig. 3　Comparison of different methods with average RD curves on objective metrics PSNR D1, PSNR D2

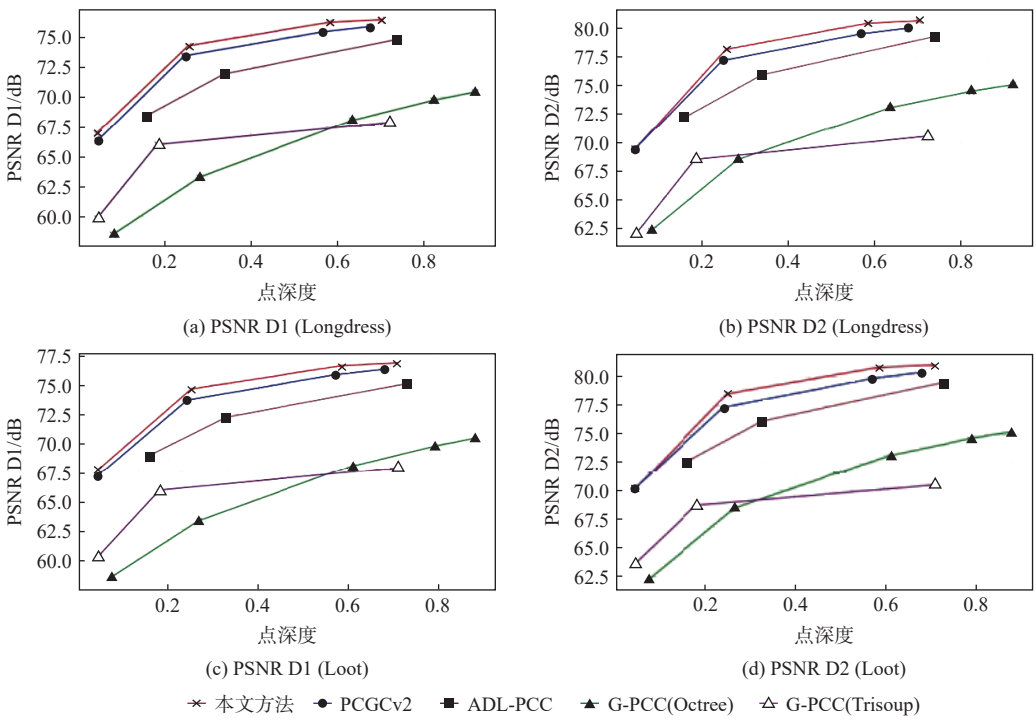


图 4　对比不同方法在单个数据集客观指标 PSNR D1、PSNR D2 上的 RD 曲线

Fig. 4　Comparison of different methods with RD curves on objective metrics PSNR D1, PSNR D2 on a single dataset

3.3　消融实验

为验证网络结构的有效性, 进行多组消融实验验证 Transformer 模块的数目及作用, 如表 5 所示。

IRN 中包含残差结构, 残差单元通过跳跃连接的方式实现, 输入信息从低维传输到高维, 可以加快网络收敛, 有效解决网络梯度消失、爆炸的问题, 同时 IRN 模块也可以更好地表达特征信息。关于 IRN 模块的有效性, 如表 5 所示, 其中 W/o IRN 为去除 IRN 结构, 由于网络难以收敛, 训练难度大, 控制码率点困难, 码率高于其他保留 IRN 的结构, 但是在客观指标 PSNR D1 和 PSNR D2 上反而明显低于其他结构, 该实验证明 IRN 模块可以加快网络收敛, 解决网络退化等问题。

Transformer 中 Attention 模块数量的消融实验如表 5 所示, Attention×1 为 Transformer 内包含 1 个偏置注意力模块, Attention×2 为本文方法 Transformer 内包含连续的 2 个偏置注意力模块, Attention×3 为 Transformer 内包含连续的 3 个偏置注意力模块。W/o Transformer 为去除 Transformer 结构, 此时结构与 PCGCv2 相同。在 0.25 的码率点下, 使用 Transformer 且内部包含 2 个偏移注意力模块的效果最好, 相比不使用 Transformer, PSNR D1 有平均 0.633 dB 的提升, PSNR D2 有平均 0.845 dB 的提升。可见, 在目前架构下使用 2 个注意力模块就可以达到最佳性能。目前, 由于显存限制, 实验无法在第 1 次下采样后使用 Transformer 模块进行特征增强, 这将在后期工作中通过改进 Transformer 模块来进一步实现。

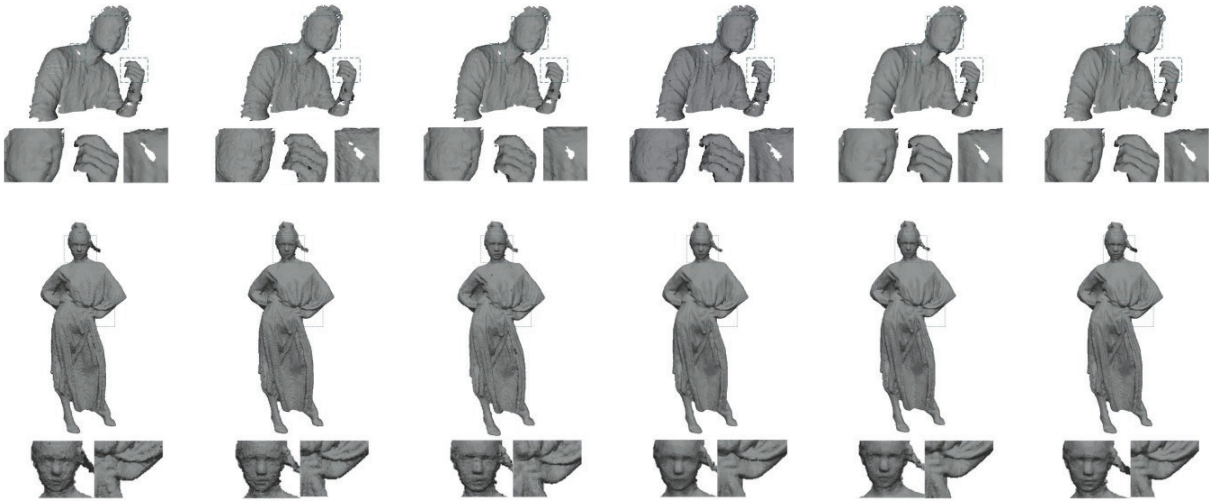


图 5 对比不同方法在单个数据集上的可视化效果

Fig. 5 Comparison of visual results of different methods on a single dataset

表 4 对比不同方法在 MVUB 测试集上的平均编解码时间

Table 4 Comparison of different methods in terms of average encoding and decoding time on MVUB test set s

方法	编码时间	解码时间
G-PCC(Octree)	0.73	0.07
G-PCC(Trisoup)	2.06	1.10
ADL-PCC	19.06	6.5
PCGCv2	0.53	0.18
本文方法	1.08	0.71

表 5 结构消融实验在测试序列上的平均客观指标对比

Table 5 Comparison of average objective metrics on test sequences for structural ablation experiments

网络结构	点深度	PSNR D1/dB	PSNR D2/dB
W/o IRN	0.253	72.623	76.147
W/o Transformer	0.239	72.903	76.389
Attention×1	0.235	73.086	76.065
Attention×2(本文方法)	0.248	73.536	77.234
Attention×3	0.246	73.238	76.228

4 结 论

1）本文尝试将 Transformer 模块与体素化方法结合，提出一种基于 Transformer 的端到端点云几何有损压缩方法。Transformer 模块将网络的感受视野扩展到全局，提取特征信息的全局相关性，从而增强特征信息，更准确地对体素块进行概率预测，逐步细化达到更好的重建效果。

2）本文方法在客观指标、BD-Rate、主观效果上都显著超过对比方法 G-PCC（Octree）、G-PCC（Trisoup）。对比深度学习方法 ADL-PCC，本文方法在编解码时间上有明显优势，客观指标也高于该方法。相比 PCGCv2，增加的 Transformer 结构带来了可观的客观指标增益，在主观效果方面相比 PCGCv2 保留了更多细节，整体结构更加完整。

3）在 PCGCv2 基础上增加 Transformer 后，网络的计算复杂度和编解码时间有所增加。在未来的工作中，将探索更高效的网络结构以减少计算复杂度和编解码时间，并且将基于 Transformer 的点云有损压缩扩展到稀疏点云领域中，进一步探索 Transformer 在三维点云领域的作用。

参考文献（References）

[1] SCHWARZ S, PREDA M, BARONCINI V, et al. Emerging MPEG standards for point cloud compression[J]. IEEE Journal on Emerging and Selected Topics in Circuits and Systems, 2019, 9(1): 133-148.

[2] GUARDA A F R, RODRIGUES N M M, PEREIRA F. Point cloud coding: Adopting a deep learning-based approach[C]//Proceedings of the Picture Coding Symposium. Piscataway: IEEE Press, 2019: 1-5.

[3] WANG J, ZHU H, LIU H J, et al. Lossy Point Cloud Geometry Compression via End-to-End Learning[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2021, 31(12): 4909-4923.

[4] YAN W, SHAO Y T, LIU S, et al. Deep autoencoder-based lossy geometry compression for point clouds[J/OL]. Computer Vision and Pattern Recognition, 2019. (2019-4-18)[2022-1-22]. https://doi.org/10.48550/arXiv.1905.03691.

[5] WEN X Z, WANG X, HOU J H, et al. Lossy geometry compres-

sion of 3d point cloud data via an adaptive octree-guided network[C]//Proceedings of the IEEE International Conference on Multimedia and Expo. Piscataway: IEEE Press, 2020: 1-6.

[6] WANG J Q, DING D D, LI Z, et al. Multiscale point cloud geometry compression[C]//Proceedings of the Data Compression Conference. Piscataway: IEEE Press, 2021: 73-82.

[7] ZHAO H S, JIANG L, JIA J Y, et al. Point Transformer[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE Press, 2021: 16239-16248.

[8] GUO M H, CAI J X, LIU Z N, et al. PCT: Point cloud Transformer[J]. Computional Visual Media, 2021, 7(2): 187-199.

[9] HUANG L L, WANG S L, WONG K, et al. Octsqueeze: Octree-structured entropy model for lidar compression[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2020: 1310-1320.

[10] QUE Z Z, LU G, XU D. Voxelcontext-net: An Octree based framework for point cloud compression[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2021: 6038-6047.

[11] NGUYEN D, QUACH M, VALENZISE G, et al. Lossless coding of point cloud geometry using a deep generative model[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2021, 31(12): 4617-4629.

[12] NGUYEN D, QUACH M, VALENZISE G, et al. Multiscale deep context modeling for lossless point cloud geometry compression[C]//Proceedings of the IEEE International Conference on Multimedia & Expo Workshops. Piscataway: IEEE Press, 2021: 1-6.

[13] HUANG T X, LIU Y. 3D point cloud geometry compression on deep learning[C]//Proceedings of the 27th ACM International Conference on Multimedia. New York: ACM, 2019: 890-898.

[14] QUACH M, VALENZISE G, DUFAUX F. Learning convolutional Transforms for lossy point cloud geometry compression[C]// Proceedings of the IEEE International Conference on Image Processing. Piscataway: IEEE Press, 2019: 4320-4324.

[15] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]// Proceedings of the 31st International Conference on Neural Information Processing Systems. New York: ACM, 2017: 6000-6010.

[16] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, et al. An image is worth 16×16 words: Transformers for image recognition at scale[J/OL]. Computer Vision and Pattern Recognition, 2021. (2021-6-3)[2022-1-22]. https://doi.org/10.48550/arXiv.2010.11929.

[17] PARMAR N, VASWANI A , USZKOREIT J, et al. Image transformer[EB/OL]. (2018-6-15)[2022-1-22]. https://arxiv.org/abs/1802.05751.

[18] LIU Z, LIN Y T, CAO Y, et al. Swin Transformer: Hierarchical vision transformer using shifted windows[C]//Proceedings of the International Conference on Computer Vision. Piscataway: IEEE Press, 2021: 10012-10022.

[19] QIU S, ANWARS, BARNES N. PU-Transformer: Point cloud upsampling transformer[C]//Proceedings of the Asian Conference on Computer Vision. Berlin: Springer, 2022: 2475-2493.

[20] FU C Y, LI G, SONG R, et al. OctAttentioN: Octree-based large-scale contexts model for point cloud compression[C]//Prodeedings of the 36th AAAI conference on Artificial Intelligence. Palo Alto: AAAI Press, 2022: 625-633.

Lossy point cloud geometry compression based on Transformer

LIU Gexin，ZHANG Junteng，DING Dandan*

(School of Information Science and Technology，Hangzhou Normal University，Hangzhou 311121，China)

Abstract：Point clouds are widely used for 3D object representation, however, real-world captured point clouds often have huge data, which is unfavorable for transmission and storage. To address the redundancy problem of point cloud data, an end-to-end Transformer-based multiscale point cloud geometry compression method is proposed by introducing the Transformer module based on the attention mechanism. The point cloud is voxelized, features are extracted using sparse convolution at the encoder, multi-scale gradual downsampling is performed, and the Transformer module is combined to enhance the point-space feature perception and extraction; at the decoder, the corresponding multi-scale up-sampling is performed for reconstruction, and the Transformer module is also used to enhance and recover the useful features, and the point cloud is progressively refined and reconstructed. Compared with two standard point cloud coding methods, the proposed method obtains 80% and 75% BD-Rate gain on average; compared with the deep learning-based point cloud compression method, it obtains 16% BD-Rate gain on average, and there is about 0.6 PSNR enhancement at the same bit rate. The experimental results demonstrate the feasibility and effectiveness of Transformer in the field of point cloud compression. In terms of subjective quality, the proposed method also has significant subjective effect improvement, and the reconstructed point cloud is closer to the original point cloud.

Keywords：point cloud geometry compression；Transformer；attention mechanism；deep learning；sparse convolution

Received：2022-05-26；**Accepted**：2022-07-02；**Published Online**：2022-11-04 16：54

URL：link.cnki.net/urlid/11.2625.V.20221104.1438.003

Foundation items：National Natural Science Foundation of China (6217010731); Zhejiang Provincial Nature Fund (LY20F010013)

* **Corresponding author**. E-mail：DandanDing@hznu.edu.cn

http://bhxb.buaa.edu.cn jbuaa@buaa.edu.cn
DOI: 10.13700/j.bh.1001-5965.2022.0438

膜内麻雀优化ELM的软件缺陷预测算法

唐宇[1]，代琪[2]，杨梦园[1]，陈丽芳[1, 3, *]

(1. 华北理工大学 理学院，唐山 063210；2. 中国石油大学 信息科学与工程学院，北京 102249；
3. 河北省数据科学与应用重点实验室，唐山 063210)

摘 要：原始麻雀搜索算法存在寻优精度低、迭代后期容易陷入局部极值的问题，结合高效寻优性能的改进麻雀搜索算法和具有并行计算能力的膜计算，提出一种膜内麻雀优化算法(IMSSA)。在10个CEC2017测试函数上的实验结果表明，IMSSA具有更高的寻优精度。为进一步验证IMSSA的性能，使用IMSSA优化极限学习机(ELM)参数，提出一种膜内麻雀优化ELM(IMSSA-ELM)算法，并将其应用于软件缺陷预测领域。实验结果表明：在15个公开的软件缺陷数据集中，IMSSA-ELM算法预测性能在G-mean、MCC这2个评价指标下明显优于其他4种先进的对比算法，表明IMSSA-ELM算法具有更好的预测精度和稳定性，其实验结果在Friedman ranking和Holm's post-hoc test非参数检验中具有明显的统计显著性。

关 键 词：改进麻雀搜索算法；膜计算；极限学习机；优化算法；软件缺陷预测
中图分类号：TP393.08
文献标志码：A **文章编号**：1001-5965(2024)02-0643-12

生活中存在各种各样的应用软件，为人们的生活方式提供极大的便利，但软件质量问题成为人们关注的焦点。大多数研究人员认为软件缺陷是影响软件质量的主要问题，存在泄露用户信息的风险[1]，造成严重的信息安全问题[2]。因此，提出一种高效的软件缺陷预测算法提高软件质量、降低软件开发和维护成本成为关键问题。

近年来，极限学习机(extreme learning machine, ELM)作为一种机器学习模型，具有高效的预测能力，被广泛应用于预测领域。董浩等[3]通过核主成分分析(kernel principle component analysis, KPCA)改进ELM的神经网络输入量，降低模型训练的不确定性，提高ELM的预测精度。陈恒志等[4]通过建立ELM的连铸坯预测模型，并与其他2种预测算法进行比较，突出ELM的预测性能。以上2种方法都体现出ELM高效的预测能力，但ELM性能容易受到输入权值和隐含层神经元阈值的影响。为解决此问题，已有研究人员采用群智能优化算法优化ELM参数。Liu等[5]使用改进布谷鸟搜索(improved cuckoo search, ICS)算法优化ELM(ICS-ELM)参数提高绝缘栅双极晶体管(insulated-gate bipolar transistor, IGBT)结温预测的精度。

Ding等[6]使用粒子群优化(particle swarm optimization, PSO)算法优化ELM，提高临床癌症诊断结果的准确性。Li等[7]使用鲸鱼优化算法优化ELM，该算法具有更高的准确性和泛化能力。

上述算法虽然在ELM稳定性和预测精度上得到提升，但各算法的寻优能力和计算速度还可以进一步提升。膜计算已经广泛应用于群智能优化算法中，有效地提高算法的全局搜索能力。

孙远等[8]将膜计算和PSO算法结合应用到混合信号分离中，膜计算保证PSO算法的计算效率及全局搜索的多样性。谢佩军和张育斌[9]使用膜计算和PSO算法改进ELM，融合膜计算的并行计算

收稿日期：2022-05-29；**录用日期**：2022-06-24；**网络出版时间**：2023-07-04 13:44
网络出版地址：link.cnki.net/urlid/11.2625.V.20220701.1556.001
基金项目：国家自然科学基金(52074126)
* **通信作者**. E-mail：hblg_clf@163.com

引用格式：唐宇，代琪，杨梦园，等. 膜内麻雀优化ELM的软件缺陷预测算法[J]. 北京航空航天大学学报，2024，50(2)：643-654.
TANG Y，DAI Q，YANG M Y，et al. Software defect prediction algorithm for intra-membrane sparrow optimizing ELM[J]. Journal of Beijing University of Aeronautics and Astronautics，2024，50(2)：643-654 (in Chinese).

能力和粒子群算法的快速收敛能力改进ELM,有效地提高ELM的预测能力和稳定性。

传统ELM算法是一种单隐含层前馈神经网络、具有学习速度快的特点[10],已经广泛应用于软件缺陷预测领域。Xu等[11]结合KPCA和加权极限学习机(weight extreme learning machine, WELM),提出一种软件缺陷预测框架(kernel principle component analysis and weight extreme learning machine, KPWE),并将其应用于软件缺陷预测领域,提高软件缺陷预测精度,但WELM存在参数随机选取的问题。

文献[12]采用合成少数类过采样技术(synthetic minority over-sampling technique, SMOTE)算法处理数据集并且使用加权ELM算法,建立一种用于软件缺陷的预测模型,与传统ELM算法相比有效提高模型预测精度,但未对WELM参数进行优化。综上所述,ELM算法应用于软件缺陷预测领域存在参数难以选取、稳定性差的问题。

针对以上问题,已有相关研究人员使用麻雀搜索算法(sparrow search algorithm, SSA)优化ELM参数,增加ELM算法的预测精度和稳定性。曾亮等[13]使用SSA为ELM提供最优参数,但算法的寻优精度低。刘栋等[14]将SSA优化ELM参数并应用于功率预测中,有效地提高算法的稳定性,但麻雀种群在迭代后期容易陷入局部极值。膜计算具有高效并行计算的能力,广泛应用于盲源分离[8]、水肥预测[9]、路径规划[15]等领域。兰世豪等[15]将膜计算和粒子群算法用于路径规划中,解决路径规划不合理和规划速度慢的问题,但粒子群算法存在全局寻优能力弱的问题。

针对以上问题,本文将膜计算应用于软件缺陷预测领域,并尝试将具有高效全局搜索能力的改进麻雀搜索算法(improved sparrow search algorithm, ISSA)与具有优秀并行计算能力的膜计算相结合,提出一种膜内麻雀优化(intra-membrane sparrow optimization algorithm, IMSSA)ELM(IMSSA-ELM)算法,最后将该算法应用于软件缺陷预测领域。

1 IMSSA设计

1.1 研究目的

SSA中发现者在迭代后期容易陷入局部最优解[16],本文使用自适应超参数更新发现者位置公式,有效地扩大发现者的搜索范围、增强麻雀种群在迭代后期跳出局部极值的能力。麻雀种群中加入者根据发现者进行更新位置,具有一定的盲目性,使用混合变异策略,在加入者位置更新后使用混合变异策略生成加入者的候选解,将加入者和候选解的适应度值进行比较,选取适应度值小的作为当前迭代次数的加入者。该策略扩大加入者的搜索范围,增加解的多样性。由于警戒者数量较少,活动范围容易受到限制,最后使用翻筋斗觅食策略生成警戒者的候选解,将警戒者和候选解的适应度值进行比较,选取适应度值小的作为当前迭代次数的警戒者,该策略增大警戒者的活动范围,同时增加警戒者搜索最优解的活力。

由于麻雀种群初始化的随机性使得麻雀个体对局部最优解的搜索不够精细,本文将具有强大并行计算能力的膜计算与ISSA进行结合,将麻雀均匀放到3个基本膜中,使麻雀个体在解空间中搜索更加精细,将每个基本膜的最优适应度值传入表层膜中进行比较,在表层膜中得到当前迭代次数中的最优适应度值,并传回各个基本膜中,使得ISSA具有优异全局寻优能力的同时,进一步增强ISSA的局部搜索能力。

为进一步验证IMSSA优化性能,本文使用IMSSA优化ELM参数,提高ELM算法的预测精度和稳定性。为验证IMSSA-ELM算法的效果,本文将其应用于软件缺陷预测领域,实验结果表明,IMSSA-ELM算法具有更高的预测精度和更好的稳定性。

1.2 改进麻雀搜索算法

1.2.1 自适应超参数

在麻雀种群中,发现者具有较好的适应度值,可以为种群提供食物方向,但发现者在位置更新时受$\exp\left(\frac{-i}{\alpha T_{\max}}\right)$影响较大。根据文献[17]可知,$\exp\left(\frac{-i}{\alpha T_{\max}}\right)$取值范围可能会从(0,1)逐渐减小到(0,0.4),在迭代后期会降低发现者的全局搜索能力。因此,本文使用自适应超参数扩大发现者的搜索范围,提高算法搜索能力,使用自适应超参数的发现者位置更新如下:

$$\boldsymbol{X}_i^{t+1}=\begin{cases}\boldsymbol{X}_i^t\cdot\exp\left(\dfrac{-i}{\omega\alpha T_{\max}}\right) & R_2<G\\ \boldsymbol{X}_i^t+Q\cdot\boldsymbol{L} & R_2\geqslant G\end{cases}\tag{1}$$

$$\omega=\sigma c^{t^2}\tag{2}$$

式中:$\sigma=1$常量;$c=0.9$[17]为ω的自适应因子,有助于提高发现者的全局搜索能力;t为当前迭代次数;$\boldsymbol{X}_i^t$表示第t次迭代时第i个麻雀的位置;$T_{\max}$表示最大的迭代次数;α为(0,1)范围内的一个随机数;R_2为预警值,若达到预警值则说明麻雀群体已经遇到危险,需要采取措施($R_2\in[0,1]$);G为安全值,若在安全值范围内麻雀群体可以正常活动($G\in[0.5,1]$);Q为服从标准正态分布的随机数;$\boldsymbol{L}$表示一个

$1\times d$的矩阵且矩阵中每个元素都为 1。

1.2.2　混合变异策略

加入者跟随发现者寻找食物, 通过发现者更新自己的位置, 因此, 加入者的搜索具有盲目性。本文引入混合变异策略更新加入者位置。麻雀种群在迭代前期需要大范围进行寻优, 而在迭代后期多样性会有所降低, 导致算法过早收敛, 容易陷入局部极值。付华和刘昊[18]引入柯西高斯变异策略, 在每次迭代后选取适应度值最好的麻雀个体进行变异, 该策略有效地提高麻雀种群跳出局部最优解的能力。本文提出混合变异策略使用标准柯西分布函数和标准高斯分布函数, 并将其应用到加入者的位置更新中, 增强麻雀种群的多样性, 使加入者具有更强的跳出局部最优解能力[19]。混合变异策略根据迭代次数引入动态变化参数λ_1、λ_2。

$$\left(\boldsymbol{X}_{\text{best}}^{t+1}\right)' = \boldsymbol{X}_{\text{best}}^{t+1}\left[1+\lambda_1\text{Cauchy}(0,1)+\lambda_2\text{Gauss}(0,1)\right] \tag{3}$$

$$\begin{cases}\lambda_1 = 1-\dfrac{t^2}{T^2}\\ \lambda_2 = \dfrac{t^2}{T^2}\end{cases} \tag{4}$$

式中: T为最大迭代次数。标准高斯分布函数和标准柯西分布函数分别为

$$\begin{cases}\text{Gauss}(x) = \dfrac{1}{\sqrt{2\pi}}\exp\left(-\dfrac{x^2}{2}\right) & -\infty < x < +\infty\\ \text{Cauchy}(x) = \dfrac{1}{\pi(1+x^2)} & -\infty < x < +\infty\end{cases} \tag{5}$$

混合变异策略是在每次迭代后, 根据当前迭代中加入者位置, 产生一个变异后的候选位置, 将 2 个位置的适应度值进行比较, 若$\boldsymbol{X}_i^{t+1}$的适应度值更好, 则继续更新$\left(\boldsymbol{X}_i^{t+1}\right)'$, 若$\left(\boldsymbol{X}_i^{t+1}\right)'$的适应度值更好, 将$\left(\boldsymbol{X}_i^{t+1}\right)'$的位置和适应度值分配给$\boldsymbol{X}_i^{t+1}$。在迭代过程中, 参数$\lambda_1$逐渐减小, 参数$\lambda_2$逐渐增大, 增强算法跳出局部极值和全局搜索的能力。

1.2.3　翻筋斗觅食策略

警戒者由发现者和加入者随机选取生成, 在麻雀种群中警戒者适应度值相对较差, 并且容易集中在麻雀群体边缘。王正通等[20]将翻筋斗觅食策略应用于灰狼捕猎的位置更新中, 该策略可以增加灰狼的搜索范围, 提高灰狼算法的收敛性。本文将翻筋斗觅食策略应用到麻雀种群中警戒者的位置更新中, 该策略可以使警戒者更加高效地进行位置更新, 还可以增强解的多样性。翻筋斗觅食策略位置更新如下:

$$\boldsymbol{X}_i^{t+1} = \boldsymbol{X}_i^t + S(r_1\boldsymbol{X}_{\text{b}}^t - r_2\boldsymbol{X}_i^t) \qquad i = 1,2,\cdots,n \tag{6}$$

式中: S为空翻因子, 决定进行翻筋斗觅食策略后麻雀的位置, 根据文献 [21] 取 $S=2$; n为麻雀数量; r_1、r_2为[0,1]随机数; $\boldsymbol{X}_{\text{b}}^t$为第$t$次迭代最优位置; $\boldsymbol{X}_i^t$为第i只麻雀第t次迭代的位置。麻雀翻筋斗示意图如图 1 所示。

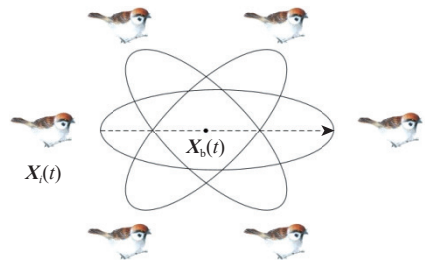


图 1　翻筋斗觅食策略

Fig. 1　Somersault foraging strategy

在警戒者位置更新中引入翻筋斗觅食策略, 有效地提高警戒者的寻优能力, 翻筋斗觅食策略主要是围绕每次迭代最优解进行的, 这也增加警戒者在空间中的搜索效率。在迭代后期麻雀种群搜索范围逐渐较小, 使用翻筋斗觅食策略可以有效增加跳出局部最优解的能力。

1.3　膜计算

膜计算是由 Păun[21]提出的一种新型自然计算分支, 旨在生命细胞的结构与功能中抽象出计算模型[22]。膜系统主要由膜结构、对象和进化规则 3 部分构成。膜计算具有分布式、并行性和非确定性的特点[9]。细胞型膜系统结构示意图如图 2 所示。

在图 2 中, 0 表示表层膜, 1 和 2 表示基本膜。a和b表示对象, $a\to a_{\text{s}}$, $b\to b_{\text{s}}$表示区域内对象与基本膜内进行信息传递的进化规则。细胞型膜系统的多元组表示:

$$\Pi = (V,H,\mu,R_1,R_2) \tag{7}$$

式中: $V=\{a,b\}$表示一个字母表, 其元素称为对象(对象a和b); $H=\{1,2\}$表示膜标号集合 (以图 2 中 2 个基本膜为例); μ表示由m个膜组成的膜结构; $R_i, 1\leqslant i\leqslant m$表示膜区域的进化规则。

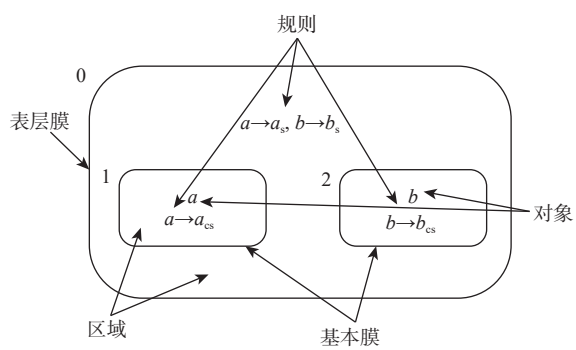


图 2　细胞型膜系统结构示意图

Fig. 2　Schematic diagram of cellular membrane system

1.4 IMSSA-ELM 算法

IMSSA 采用细胞型膜系统的单层膜结构。本文将N个麻雀均匀放入m个基本膜内进行精细化的局部搜索。各基本膜每迭代一次,都需要对膜内麻雀的适应度值进行排序,并将各细胞膜内最优适应度值的麻雀个体输送到表层膜内进行当前迭代次数的最后排序,选出当前迭代次数的全局最优解输送回各基本膜内用于麻雀种群下一次迭代的位置更新。

以第 1 次迭代为例:首先,将各个基本膜内麻雀个体按照适应度值进行排序,选取适应度值最小的麻雀个体对应位置作为该膜内的最优位置;然后,将m个基本膜的最优位置$(\boldsymbol{g}'_1,\boldsymbol{g}'_2,\cdots,\boldsymbol{g}'_m)$通过膜内规则$R$输出到表层膜中进行比较,在$m$个局部最优位置中选取适应度值最小的麻雀作为当前迭代次数的全局最优位置$\boldsymbol{g}$。

最后,通过膜内规则R将$\boldsymbol{g}$送到各基本膜用于下一次迭代更新麻雀位置。由于膜计算具有强大的并行处理能力,麻雀群体在m个基本膜内同时进行局部搜索相当于麻雀群体在解空间中进行全局搜索。将膜计算和 ISSA 相结合有效提高 SSA 的局部搜索能力,增强麻雀种群跳出局部最优解的能力。

目前 ELM 在软件缺陷预测中存在随机选取参数、预测精度低和稳定性差的问题。本文 IMSSA-ELM 算法,结合膜计算并行计算能力和改进麻雀高效寻优能力,将麻雀个体均匀地分布到各个基本膜中,让每个麻雀先在基本膜中进行搜索,再将最优位置和适应度值传到表层膜中进行比较,有效地提高算法的搜索范围及全局搜索能力,同时膜计算具有强大的并行计算能力,麻雀种群可以在各个基本膜中同时进行精细化搜索。算法中麻雀个体由位置和适应度值表示,软件缺陷预测准确率作为 IMSSA-ELM 算法目标函数。IMSSA-ELM 算法具体步骤如下:

步骤 1 初始化膜系统结构,其中包括一个表层膜 0 和m个基本膜,初始化麻雀种群及相关参数;

步骤 2 将麻雀个体均匀分布到m个基本膜中,并计算每个麻雀个体的适应度值;

步骤 3 在麻雀种群中选取适应度值较小的麻雀作为发现者,根据式 (1) 更新发现者位置;

步骤 4 除发现者之外的麻雀作为加入者,更新加入者的位置;

步骤 5 根据式 (3) 生成加入者的候选解;

步骤 6 将加入者和候选解的适应度值进行比较,选取适应度值较小的作为加入者;

步骤 7 在发现者和加入者中随机选取麻雀作为警戒者,更新警戒者位置;

步骤 8 根据式 (6) 生成警戒者的候选解;

步骤 9 将警戒者和候选解的适应度值进行比较,选取适应度值较小的作为警戒者;

步骤 10 将每个麻雀个体的最优位置与当前基本膜内群体的最优位置$\boldsymbol{g}'$作比较,选择最优适应度值最小的麻雀作为基本膜内的最优位置;

步骤 11 将m个基本膜内的最优位置根据规则R输送到表层膜中,通过对比m个麻雀个体的适应度值,选出当前迭代次数的最优位置,并将最优麻雀个体输送到m个基本膜中;

步骤 12 继续更新麻雀群体位置;

步骤 13 判断算法运行是否达到最大迭代次数,若达到循环则结束,若未达到则返回步骤 3;

步骤 14 输出全局最优位置和最优适应度值,得到 ELM 最优参数。IMSSA-ELM 算法流程如图 3 所示。

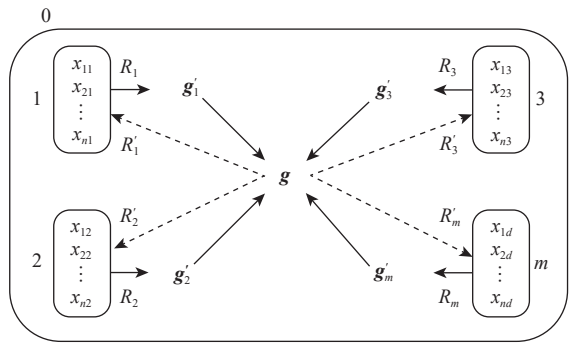


图 3 IMSSA-ELM 算法示意图

Fig. 3 Schematic diagram of IMSSA-ELM algorithm

IMSSA-ELM 算法如下所示:

输入 最大迭代次数 M,警戒值 R_2,麻雀种群数量 N,基本膜 k,基本膜数量 m

输出 $\boldsymbol{B}_X$ 和 $\boldsymbol{B}_F$

开始

初始化麻雀种群数量 N 和基本膜 k

for t = 1 to M do

根据式 (1), 式 (3), 式 (6) 得到 ISSA 算法

end for

$k_m \leftarrow N/m$ //将麻雀种群均匀放入 m 个基本膜

for j=1 to M do

for i=1 to m do

$\boldsymbol{B}_{\text{pos1}} \leftarrow \text{ISSA}(\boldsymbol{B}_X)$//记录基本膜中最优麻雀位置

$\boldsymbol{B}_f(i) \leftarrow \text{fob}j(\boldsymbol{B}_{\text{pos1}}(i,:))$ //记录基本膜中最优麻雀位置的适应度值

end for

$\boldsymbol{B}_F \leftarrow \min(\text{sort}(\boldsymbol{B}_f(i)))$ //将适应度值排序,选取适应度值最小的麻雀为最优位置

13　　if $\boldsymbol{B}_F<\boldsymbol{f}_{\text{Min}}$

14　　$\boldsymbol{B}_{\text{pos}}\leftarrow\boldsymbol{X}_{\text{new}}(i)$ //出现适应度值更小的麻雀时, 对其位置进行更新

15　　end if

16 end for

17 Return $\boldsymbol{B}_X$ and $\boldsymbol{B}_F$

2　IMSSA 性能实验

实验环境: 使用软件 MATLAB; 操作系统: win10 家庭中文版; CPU: AMD R7-4800; 内存 16 GB。

2.1　IMSSA 对比实验

本文为验证 IMSSA 的优化能力, 选取 4 种先进优化算法作为对比算法。Xue 和 Shen[16] 根据麻雀觅食和反捕食行为, 提出一种具有高效寻优性能的麻雀搜索算法 SSA。Sallam 等[22] 提出的一种改进的多算子差分进行 (improved multi-operator differential evolution, IMODE) 算法, 该算法根据解的质量和子种群的多样性选择最佳算子, 并在进化过程后期使用序列二次规划法 (sequential quadratic programming, SQP) 进行局部搜索, 有效地提高算法的优化能力和适用性。Mohamed 等[23] 提出的一种改进的获取共享知识 (adaptive gaining sharing knowledge, AGSK) 算法, 该算法将原始获取共享知识 (gaining sharing knowledge, GSK) 算法中知识因子和知识 2 个控制参数进行自适应设置, 使算法具有更强的优化性能。江妍等[24] 提出一种基于分数阶麻雀搜索算法 (adaptivesparrow search algorithm, ASSA), 有效地提高 SSA 的全局搜索能力。本文将 IMSSA、SSA、IMODE、AGSK、ASSA 这 5 种优化算法在 CEC2017 测试函数中进行对比, 实验设置 IMSSA、ASSA 和 SSA 种群规模为 30, 最大迭代次数设为 500[19], 其他算法参数根据参考文献 [23] 设置。

本文在优化算法常用 CEC2017 函数中随机选取 10 个函数进行实验[23], CEC2017 函数信息如表 1 所示。为突出本文算法的性能与稳定性, 将 5 种优化算法分别在 10 个 CEC2017 函数上独立运行 30 次, 优化结果如表 2 所示, 每个函数中最优值加粗表示, 为方便观察每种算法在各 CEC2017 函数中的寻优性能, 表 2 的实验结果是由 5 种优化算法评价指标减去 CEC2017 函数最优值得出[23]。

表 1　CEC2017 函数基本信息

Table 1　Characteristics of CEC2017 benchmark functions

函数	序号	维度	范围	最优值
平移和旋转 Rastrigin 函数	C01	10	[−100,100]	500
平移和旋转 Lunacek Bi_Rastrigin 函数	C02	10	[−100,100]	700
平移和旋转非连续 Rastrigin 函数	C03	10	[−100,100]	800
混合函数 4 (*N*=4)	C04	10	[−100,100]	1 400
混合函数 6 (*N*=4)	C05	10	[−100,100]	1 600
混合函数 6 (*N*=5)	C06	10	[−100,100]	1 700
混合函数 6 (*N*=6)	C07	10	[−100,100]	2 000
复合函数 1 (*N*=3)	C08	10	[−100,100]	2 100
复合函数 1 (*N*=4)	C09	10	[−100,100]	2 300
复合函数 4 (*N*=4)	C10	10	[−100,100]	2 400

分析表 2 可知, IMSSA 在大部分 CEC2017 函数中均取得最优适应度值和最优平均适应度值, 并且 IMSSA 标准差值相比于其他 4 种优化算法稳定性接近。IMSSA 在 C01、C02、C05、C07、C08、C09、C10 这 7 个 CEC2017 函数中取得最优适应度值, 表明 IMSSA 在多峰函数、混合函数和复合函数中具有很强的寻优性能。在函数 C04 中, AGSK 取得最优适应度值, 但 IMSSA 仅与最优适应度值相差 53.832, 充分说明 IMSSA 具有高效的寻优能力。在 C06 测试函数中, ASSA 取得最优适应度值, 但 IMSSA 为次优适应度值, 并且 2 种算法的适应度值接近。

AGSK 在 7 个 CEC2017 函数中获得最优标准差值, 充分说明 AGSK 优化算法相比于其他 4 种优化算法具有更加优秀的稳定性, IMODE 在 C08 取得最优标准差值, 并且 IMODE 在其他 9 个 CEC2017 函数中寻优能力较强, 表明 IMODE 算法同样具备高效的优化能力。综上所述, 5 种先进的优化算法虽然各有优势, 但 IMSSA 具有更高的寻优精度, AGSK 具有更好的稳定性。

2.2　非参数统计检验

本文使用非参数统计检验对 5 种先进的优化算法的寻优结果进行分析, 使用 Friedman 检验来检验每一组结果之间的统计差异, 并使用 Holm 检验来找出 IMSSA-ELM 算法与哪种对比算法存在显著性差异。非参数统计分析需要考虑算法的平均排

表2 各算法优化结果

Table 2 3Optimization results of each algorithm

算法	C01				C02			
	最优值	最差值	平均值	标准差	最优值	最差值	平均值	标准差
IMSSA	**33.914**	117.73	**78.579**	19.647	**96.636**	158.07	133.03	16.493
ASSA	39.103	158.02	126.51	31.707	102.40	172.78	133.79	13.606
SSA	326.94	470.74	412.07	33.216	642.36	773.56	727.59	28.529
IMODE	215.01	356.19	266.06	44.849	621.82	1 252.6	861.21	250.12
AGSK	76.287	**103.06**	88.230	**8.297 8**	110.59	**136.81**	**124.19**	**8.688 3**

算法	C03				C04			
	最优值	最差值	平均值	标准差	最优值	最差值	平均值	标准差
IMSSA	**31.211**	**77.441**	62.421	**9.138 0**	96.697	1 124.3	285.88	227.83
ASSA	32.685	99.041	**58.822**	14.050	109.51	27 414	3 631.9	6 926.8
SSA	275.84	382.17	331.99	31.277	97.980	19 223	4 102.0	4 774.3
IMODE	176.10	300.47	237.55	38.684	103.57	260.38	167.53	53.626
AGSK	71.912	112.95	98.216	11.242	**42.865**	**56.807**	**47.673**	**4.849 9**

算法	C05				C06			
	最优值	最差值	平均值	标准差	最优值	最差值	平均值	标准差
IMSSA	**90.223**	**682.57**	**428.26**	138.95	78.836	**195.85**	**133.41**	**35.781**
ASSA	158.57	914.91	580.22	186.11	**71.840**	532.16	250.55	126.47
SSA	2 415.5	7 295.7	4 168.8	1 127.5	938.01	9 739.8	3 140.3	2 490.5
IMODE	538.56	2 204.6	1 498.1	544.34	587.23	1 456.8	967.73	255.26
AGSK	550.84	907.42	738.77	**104.88**	99.017	243.21	146.12	45.266

算法	C07				C08			
	最优值	最差值	平均值	标准差	最优值	最差值	平均值	标准差
IMSSA	**71.457**	352.03	**201.48**	73.075	**160.58**	317.37	268.29	32.983
ASSA	75.892	510.55	292.15	129.06	190.55	323.88	279.25	40.194
SSA	732.58	1 486.9	1 090.4	187.73	438.64	735.10	633.05	61.259
IMODE	488.72	979.70	766.05	155.53	384.49	450.60	417.83	**22.045**
AGSK	212.50	**412.20**	281.83	**60.994**	177.14	**307.28**	**264.04**	43.527

算法	C09				C010			
	最优值	最差值	平均值	标准差	最优值	最差值	平均值	标准差
IMSSA	**355.26**	**441.35**	**394.00**	19.205	**286.69**	537.23	**450.48**	65.565
ASSA	442.70	647.13	506.88	37.550	527.92	779.56	681.47	59.241
SSA	829.88	1 979.9	1 286.9	213.85	1 138.1	1 798.2	1 457.1	138.83
IMODE	671.16	888.74	776.26	83.061	802.47	1 112.9	955.99	97.061
AGSK	41 629	45 327	43 461	**11 806**	474.97	**525.60**	501.22	**13.510**

序，平均排序用来衡量一种算法与其他对比算法的好坏。平均排序是根据每种算法在不同数据集上的性能进行排序，当运行完所有数据集后，最小平均排序值即为最优算法。APVs表示调整后的p值，其表示拒绝假设的最低显著水平[25]，p<0.05表示2种算法之间存在显著性差异，p>0.05表示2种算法之间差异不明显，即算法性能接近。本实验使用最小值进行排序。表3和表4分别是5种优化算法在最优值和标准差2种评价指标下的非参数统计分析结果。

表 3　各算法最优值 Friedman 平均排序和 APVs 值

Table 3　Friedman average ranking and APV value of best values of each algorithm

算法	Friedman排序	*p*值
IMSSA	1.200 0	
ASSA	2.300 0	0.050 0
AGSK	2.600 0	0.025 0
IMODE	4.000 0	0.016 7
SSA	4.900 0	0.012 5

表 4　各算法标准差 Friedman 平均排序和 APVs 值

Table 4　Friedman mean ranking and APVs of standard deviations for each algorithm

算法	Friedman排序	*p*值
AGSK	1.500 0	
IMSSA	2.100 0	0.050 0
ASSA	3.000 0	0.025 0
IMODE	3.800 0	0.016 7
SSA	4.600 0	0.012 5

分析表 3 可知, IMSSA 与 ASSA 性能接近, 但 IMSSA 排名最高, IMSSA 与 IMODE 和 SSA 存在显著性差异, 充分说明 IMSSA 具有高效的寻优能力。分析表 4 可知, AGSK 的标准差值排名最高, 充分说明 AGSK 优化算法具有更好的稳定性, 但 IMSSA 的 p 值为 0.05, 表明 IMSSA 和 AGSK 不存在显著性差异, 并且 IMSSA 标准差值排名第 2, 充分说明 IMSSA 具有更好的稳定性。

3　软件缺陷预测算法对比实验

3.1　软件缺陷预测算法实验结果

IMSSA 选取麻雀数量为 30[26], 隐含层神经元个数为 100, 激活函数为 Sigmoid[7]。根据文献 [8] 实验结果可知, 本文实验设置基本膜为 3 个。其他对比算法参数根据参考文献 [16, 22-24] 设置。本文选用公开软件缺陷数据集 NASA、SOFTLAB、AEEEM、MORPH、JIRA 这 5 组中的 15 个数据集进行仿真实验, 实验使用五折交叉验证。软件缺陷数据集中包括软件设计问题、软件编码问题、软件测试问题、软件操作问题。由于软件缺陷数据集存在不平衡比大的问题, 本文使用 SMOTE 算法处理数据集不平衡问题, 使用主成分分析 (principle component analysis,PCA) 进行降维, 处理软件缺陷数据集中维度较高问题, 降维后的主成分个数如表 5 所示。

为进一步验证 IMSSA 的优越性, 本文选取 4 种先进的软件缺陷预测算法作为对比算法。Ding 等[6] 提出一种使用粒子群算法为 ELM 提供最优参数 (PSO-ELM), 在模型分类预测取得高效性能。Xu 等[11] 提出一种 KPWE 软件缺陷预测框架, 该框架使用 KPCA 提取具有代表性的数据特征, 再使用 WELM 算法进行预测, 对数据集进行分层抽样, 50% 作为训练集, 50% 作为测试集, 有效保证训练集和测试集缺陷率相同。江妍等[24] 提出一种基于 ASSA, 有效地提高 SSA 的寻优能力。实验选用 ELM、KPWE、PSO-ELM、ASSA-ELM 这 4 种先进的对比算法。为充分验证 IMSSA 的优化能力, 本节实验设置保持

表 5　数据集基本信息

Table 5　Basic information of dataset

组	特征数	数据集	样本数	属性数	缺陷数	缺陷率	主成分数
NASA	37	PC1	735	2	61	0.083 0	11
	37	CM1	344	2	42	0.122 1	10
	39	MC2	125	2	44	0.352 0	10
	37	MW1	263	2	27	0.102 7	10
SOFTLAB	29	ar6	101	2	15	0.148 5	7
AEEEM	61	EQ	324	2	129	0.398 1	20
	61	JDT	997	2	206	0.206 6	23
	61	PDE	1 497	2	209	0.140 3	22
	61	ML	1862	2	245	0.132 1	26
MORPH	20	arc	234	2	27	0.115 4	9
	20	camel-1.0	339	2	13	0.038 3	11
	20	poi-1.5	237	2	136	0.573 8	11
	20	redktor	176	2	27	0.153 4	10
	20	velocity-1.4	196	2	146	0.744 9	11
JIRA	65	Hive-0.9.0	1 416	2	283	0.199 8	21

与2.1节相同，IMSSA、ASSA-ELM、PSO-ELM最大迭代次数统一设置为500。本文使用G-mean、MCC这2种评价指标评估IMSSA-ELM算法与4种先进预测算法的实验结果[27]，如表6和表7所示。评价指标值越高说明模型预测性能越好，各评价指标的最优结果已进行加粗处理。

表6为5种软件缺陷预测算法分别在15个数据集实验的G-mean值。表7为5种软件缺陷预测算法分别在15个数据集实验的MCC值。分析表6和表7，IMSSA-ELM算法在15个软件缺陷数据集中分别与ELM、KPWE、PSO-ELM、ASSA-ELM这4种先进的预测算法进行对比，在G-mean、MCC这2个评价指标下，IMSSA-ELM算法在大部分数据集中均可取得最优值，并且在2个评价指标中IMSSA-ELM算法均取得最优平均值0.8648和0.6800，充分说明IMSSA-ELM算法的高效性。分析表6可知，IMSSA-ELM算法在PC1、CM1、MC2、MW1、ar6、PDE、arc、camel-1.0、poi-1.5、redktor、velocity-1.4、hive-0.9.0这12个数据集中取得最优G-mean值，在大规模数据集PDE、JDT、ML中同样

表6 各算法预测G-mean值

Table 6 G-mean value predicted by each algorithm

数据集	G-mean值				
	ELM	KPWE	PSO-ELM	ASSA-ELM	IMSSA-ELM
PC1	0.710 3	0.582 8	0.901 9	0.831 6	**0.913 4**
CM1	0.493 3	0.524 8	0.748 0	0.789 4	**0.812 2**
MC2	0.374 2	0.543 2	0.867 5	0.918 1	**0.928 9**
MW1	0.439 0	0.660 5	0.809 0	0.685 4	**0.850 5**
ar6	0.594 1	0.452 9	0.936 0	0.908 0	**0.984 1**
EQ	0.670 7	0.655 2	**0.879 5**	0.848 4	0.851 1
JDT	0.695 8	0.622 7	**0.866 7**	0.798 8	0.824 8
PDE	0.561 6	0.568 9	0.733 2	0.680 3	**0.743 7**
ML	0.702 9	0.480 8	0.769 4	**0.775 1**	0.771 7
arc	0.573 8	0.582 8	0.838 5	0.781 8	**0.889 8**
camel-1.0	0.558 1	0.449 3	0.627 4	0.703 4	**0.917 7**
poi-1.5	0.574 2	0.681 6	0.773 2	0.790 0	**0.857 6**
redktor	0.678 2	0.645 3	0.802 3	0.864 3	**0.883 2**
velocity-1.4	0.680 4	0.716 5	0.866 0	0.892 9	**0.932 8**
hive-0.9.0	0.707 9	0.712 0	0.772 1	0.788 7	**0.810 1**

表7 算法预测MCC值

Table 7 8MCC value predicted by each algorithm

数据集	MCC值				
	ELM	KPWE	PSO-ELM	ASSA-ELM	IMSSA-ELM
PC1	0.331 2	0.217 3	0.626 2	0.483 0	**0.637 7**
CM1	0.040 1	0.174 6	0.475 4	0.481 1	**0.550 0**
MC2	0.089 1	0.305 7	0.816 2	0.845 8	**0.853 5**
MW1	0.013 2	0.334 9	**0.576 4**	0.414 8	0.568 2
ar6	0.171 5	0.204 8	0.904 4	0.807 9	**0.930 5**
EQ	0.356 4	0.416 8	**0.773 5**	0.709 7	0.697 8
JDT	0.379 8	0.444 8	**0.694 6**	0.588 5	0.616 0
PDE	0.126 6	0.290 5	**0.424 3**	0.338 1	0.416 9
ML	0.296 4	0.223 7	**0.451 8**	0.429 8	0.435 4
arc	0.110 1	0.253 7	0.596 0	0.478 0	**0.734 5**
camel-1.0	0.267 8	0.120 0	0.488 2	0.378 1	**0.772 5**
poi-1.5	0.147 5	0.377 6	0.547 5	0.580 4	**0.710 2**
redktor	0.284 0	0.349 7	0.680 8	0.655 0	**0.782 8**
velocity-1.4	0.375 2	0.400 4	0.839 3	0.786 6	**0.913 7**
hive-0.9.0	0.395 5	0.475 2	0.493 1	0.523 9	0.580 6

取得较好的 G-mean 值。IMSSA-ELM 算法在其余 3 个数据集中虽未取得最优值, 但 IMSSA-ELM 算法的 G-mean 与最优值接近, 并且 IMSSA-ELM 算法在 15 个数据集中取得最优平均值, 充分说明 IMSSA-ELM 算法具有高效的预测性能。分析表 7 可得, 在 PC1、CM1、arc、camel-1.0 这 4 个不平衡比大的数据集中, IMSSA-ELM 算法明显优于其他 4 种对比算法的 MCC 值。IMSSA-ELM 算法虽然在 5 个数据集中未取得最优评价指标值, 但 IMSSA-ELM 算法的评价指标值大多为次优值, 并且与最优值相差较小。在 MW1、EQ、JDT、PDE、ML 这 5 个数据集中, IMSSA-ELM 算法均取得次优值, 但在 15 个数据集中, IMSSA-ELM 算法取得最优平均 MCC 值, 充分说明 IMSSA-ELM 算法的预测能力。

标准差是衡量算法效果和稳定性的重要指标, 在实验过程中, 标准差越低算法稳定性越好, 反之稳定性越差[27]。如图 4 和图 5 所示, 在评价指标 G-mean 下, IMSSA-ELM 算法的标准差均小于其他 4 种预测算法, 在评价指标 MCC 中 KPWE 取得最优标准差, 由于 KPWE 算法的 MCC 值相对于 IMSSA-ELM 算法普遍较低, 并且 MCC 值比较接近。因此, KPWE 取得最优 MCC 值, 但 IMSSA-ELM 算法的标准差值与最优标准差值相差较小。综上所述, 充分表明 IMSSA-ELM 算法在 15 个软件缺陷数据集上具有更好的稳定性。

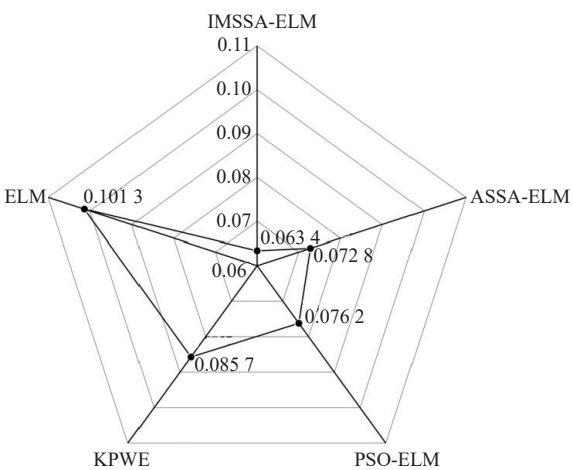


图 4　各算法预测 G-mean 标准差

Fig. 4　Standard deviation of G-mean for each algorithm

3.2　非参数统计检验

本文使用非参数统计检验对评价指标结果进行统计分析, 使用 Friedman 检验来检验每一组结果之间的统计差异, 并使用 Holm 检验来找出 IMSSA-ELM 算法与哪种对比算法存在显著性差异, 实验使用最大值进行排序。

表 8 和表 9 为 5 种软件缺陷预测算法在 G-mean 和 MCC 这 2 种评价指标下的非参数统计分析结果。分析表 8 可知, IMSSA-ELM 算法与 ELM、KPWE 这 2 种算法存在显著的统计差异性, 但 IMSSA-ELM 算法与 ASSA-ELM 和 PSO-ELM 性能相近。由表 8 和表 9 可知, IMSSA-ELM 算法与 ASSA-ELM 和 PSO-ELM 不存在显著性差异, 但 IMSSA-ELM 算法的 Friedman 排名第 1, 充分说明 IMSSA-ELM 算法的优越性。

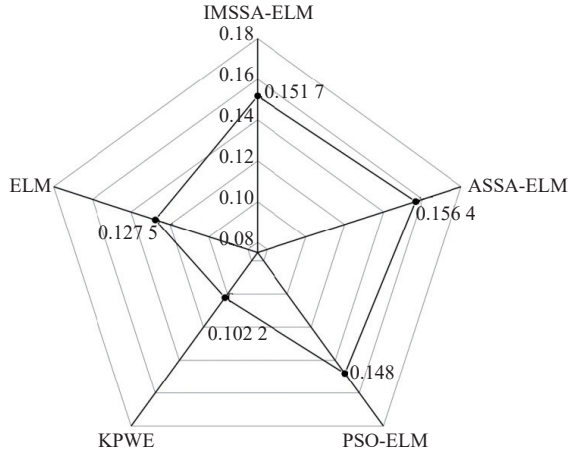


图 5　各算法预测 MCC 标准差

Fig. 5　Standard deviation of MCC for each algorithm

表 8　F-measure 的 G-mean 平均排序和 APVs 值

Table 8　Friedman ranking and APVs of G-mean

算法	Friedman排序	*p*值
IMSSA-ELM	1.200 0	
ASSA-ELM	1.400 0	0.075 3
PSO-ELM	1.400 0	0.075 3
KPWE	4.466 7	0.000 0
ELM	4.533 3	0.000 0

表 9　MCC 的 Friedman 平均排序和 APVs 值

Table 9　Friedman ranking and APVs of MCC

算法	Friedman排序	*p*值
IMSSA-ELM	1.400 0	
ASSA-ELM	2.666 7	0.056 5
PSO-ELM	1.933 3	0.355 6
KPWE	4.200 0	0.000 0
ELM	4.800 0	0.000 0

3.3　分析与讨论

传统 ELM 算法在小规模数据集中预测精度普通较高, 但在大规模数据集中, ELM 算法的预测能力还有很大的提升空间。ELM 算法在不平衡比较大的数据集中, 预测性能大幅度下降, 说明单隐含层前馈神经网络的 ELM 算法稳定性差。为进一步优化 ELM 算法参数, 本文结合具有高效寻优能力的 ISSA 和具有并行计算能力的膜计算, 使麻雀种群在解空间中可以进行更加精细的搜索, 提高

SSA的全局搜索能力，达到更加高效的预测性能。本文IMSSA-ELM算法在PDE、ML、hive-0.9.0这3个大规模数据集中，能够获得更好的评价指标数值，说明IMSSA-ELM算法在大规模数据集中也能体现强大的预测能力。在PC1、camel-1.0这2个不平衡比大的数据集中，IMSSA-ELM算法同样表现出更优的预测效果，充分说明IMSSA-ELM算法相对比其他预测算法可以提升不平衡数据的预测效果，同时也表明IMSSA-ELM算法在大规模数据集上的预测性能。IMSSA-ELM算法在G-mean、MCC这2个评价指标下，IMSSA-ELM预测算法的均值分别优于ELM算法（0.263 8和0.460 1），充分说明IMSSA-ELM算法预测效果更为显著，虽然在少数数据集中对比预测算法各有优势，但综合IMSSA-ELM算法在15个数据集中的评价指标分析，充分说明IMSSA-ELM算法的预测能力和稳定性要优于其他4种预测算法。

4 结 论

1）本文提出IMSSA，IMSSA不仅增强SSA在迭代后期跳出局部极值的能力，还增加麻雀种群搜索最优解的活力。实验结果表明，IMSSA相比于其他4种优化算法在10个CEC2017测试函数中具有更强的寻优精度。

2）为进一步提高ISSA的局部搜索能力，本文将膜计算与ISSA进行结合，把麻雀群体均匀地放入各个基本膜进行局部寻优，利用膜计算的并行计算能力，各个基本膜中的麻雀可以同时进行精细化的局部搜索，然后通过膜计算的规则将局部最优麻雀传入表层膜中进行全局寻优，进而得到最优ELM参数。在15个公开的软件缺陷数据集中，IMSSA-ELM预测算法的2个评价指标明显优于其他4种预测算法，并且算法的最优平均评价指标值最大，标准差最小，说明IMSSA-ELM算法的预测精度更高、稳定性更好。使用Friedman检验来检验每一组结果之间的统计差异，使用Holm检验IMSSA-ELM算法在评价指标G-mean和MCC下与其他4种对比算法均存在显著性差异，表明IMSSA-ELM算法的优越性。

未来研究工作的重点在于探索更加快速、高效的群智能优化算法，并将其应用于软件缺陷预测或数据挖掘等相关领域。

参考文献（References）

[1] XU Z, LI L, YAN M, et al. A comprehensive comparative study of clustering-based unsupervised defect prediction models[J]. Journal of Systems and Software, 2021, 172: 110862.

[2] JAYANTHI R, FLORENCE L. Software defect prediction techniques using metrics based on neural network classifier[J]. Cluster Computing, 2019, 22(1): 77-88.

[3] 董浩, 李明星, 张淑清, 等. 基于核主成分分析和极限学习机的短期电力负荷预测[J]. 电子测量与仪器学报, 2018, 32(1): 188-193.
DONG H, LI M X, ZHANG S Q, et al. Short-term power load forecasting based on kernel principal component analysis and extreme learning machine[J]. Journal of Electronic Measurement and Instrumentation, 2018, 32(1): 188-193(in Chinese).

[4] 陈恒志, 杨建平, 卢新春, 等. 基于极限学习机(ELM)的连铸坯质量预测[J]. 工程科学学报, 2018, 40(7): 815-821.
CHEN H Z, YANG J P, LU X C, et al. Quality prediction of the continuous casting bloom based on the extreme learning machine (ELM)[J]. Chinese Journal of Engineering Science, 2018, 40(7): 815-821(in Chinese).

[5] LIU B Y, CHEN G L, LIN H C, et al. Prediction of IGBT junction temperature using improved cuckoo search-based extreme learning machine[J]. Microelectronics Reliability, 2021, 124: 114267.

[6] DING L, ZHANG X Y, WU D Y, et al. Application of an extreme learning machine network with particle swarm optimization in syndrome classification of primary liver cancer[J]. Journal of Integrative Medicine, 2021, 19(5): 395-407.

[7] LI L L, SUN J, TSENG M L, et al. Extreme learning machine optimized by whale optimization algorithm using insulated gate bipolar transistor module aging degree evaluation[J]. Expert Systems with Applications, 2019, 127: 58-67.

[8] 孙远, 杨峰, 郑晶, 等. 基于膜计算与粒子群算法的盲源分离方法[J]. 振动与冲击, 2018, 37(17): 63-71.
SUN Y, YANG F, ZHENG J, et al. Blind source separation method based on membrane computing and particle swarm optimization[J]. Journal of Vibration and Shock, 2018, 37(17): 63-71(in Chinese).

[9] 谢佩军, 张育斌. 膜计算粒子群算法改进极限学习机的水肥预测模型研究[J]. 中国农机化学报, 2021, 42(4): 142-149.
XIE P J, ZHANG Y B. Research on water and fertilizer prediction model of improved extreme learning machine by membrane computing particle swarm optimization[J]. Chinese Journal of Chinese Agricultural Mechanization, 2021, 42(4): 142-149(in Chinese).

[10] ADNAN R M, MOSTAFA R R, KISI O, et al. Improving streamflow prediction using a new hybrid ELM model combined with hybrid particle swarm optimization and grey wolf optimization[J]. Knowledge-Based Systems, 2021, 230: 107379.

[11] XU Z, LIU J, LUO X P, et al. Software defect prediction based on kernel PCA and weighted extreme learning machine[J]. Information and Software Technology, 2019, 106: 182-200.

[12] KHADIJAH K, SASONGKO P S. Software defect prediction using synthetic minority over-sampling technique and extreme learning machine[J]. Kinetik Game Technology Information System Computer Network Computing Electronics and Control, 2020, 5(3): 203-210.

[13] 曾亮, 雷舒敏, 王珊珊, 等. 基于OVMD-SSA-DELM-GM模型的超短期风电功率预测方法[J]. 电网技术, 2021, 45(12): 4701-4712.
ZENG L, LEI S M, WANG S S, et al. Ultra-short-term wind power prediction method based on OVMD-SSA-DELM-GM model[J]. Power System Technology, 2021, 45(12): 4701-4712 (in Chinese).

[14] 刘栋, 魏霞, 王维庆, 等. 基于SSA-ELM的短期风电功率预测[J].

智慧电力, 2021, 49(6): 53-59.
LIU D, WEI X, WANG W Q, et al. Short-term wind power prediction based on SSA-ELM[J]. Smart Power, 2021, 49(6): 53-59(in Chinese).

[15] 兰世豪, 韩涛, 黄友锐, 等. 基于膜计算和粒子群的煤矿移动机器人动态窗口算法研究[J]. 工矿自动化, 2020, 46(11): 46-53.
LAN S H, HAN T, HUANG Y R, et al. Research on dynamic window algorithm of coal mine mobile robot based on membrane computing and particle swarm[J]. Industry and Mine Automation, 2020, 46(11): 46-53(in Chinese).

[16] XUE J K, SHEN B. A novel swarm intelligence optimization approach: sparrow search algorithm[J]. Systems Science & Control Engineering, 2020, 8(1): 22-34.

[17] ZHANG C, DING S. A stochastic configuration network based on chaotic sparrow search algorithm[J]. Knowledge-Based Systems, 2021, 220(10): 106924.

[18] 付华, 刘昊. 多策略融合的改进麻雀搜索算法及其应用[J]. 控制与决策, 2022, 37(1): 87-96.
FU H, LIU H. Improved sparrow search algorithm based on multi-strategy fusion and its application[J]. Control and Decision, 2022, 37(1): 87-96(in Chinese).

[19] WANG W C, XU L, CHAU K W, et al. Yin-Yang firefly algorithm based on dimensionally Cauchy mutation[J]. Expert Systems with Applications, 2020, 150: 113216.

[20] 王正通, 程凤芹, 尤文, 等. 基于翻筋斗觅食策略的灰狼优化算法[J]. 计算机应用研究, 2021, 38(5): 1434-1437.
WANG Z T, CHENG F Q, YOU W, et al. Optimization algorithm of gray wolf based on somersault foraging strategy[J]. Application Research of Computers, 2021, 38(5): 1434-1437(in Chinese).

[21] PĂUN G. Computing with membranes[J]. Journal of Computer and System Sciences, 2000, 61(1): 108-143.

[22] SALLAM K M, ELSAYED S M, CHAKRABORTTY R K, et al. Improved multi-operator differential evolution algorithm for solving unconstrained problems[C]//Proceedings of the IEEE Congress on Evolutionary Computation. Piscataway: IEEE Press, 2020: 1-8.

[23] MOHAMED A W, HADI A A, MOHAMED A K, et al. Evaluating the performance of adaptive gainingsharing knowledge based algorithm on CEC 2020 benchmark problems[C]//Proceedings of the IEEE Congress on Evolutionary Computation. Piscataway: IEEE Press, 2020: 1-8.

[24] 江妍, 马瑜, 梁远哲, 等. 基于分数阶麻雀搜索优化 OTSU 肺组织分割算法[J]. 计算机科学, 2021, 48(S1): 28-32.
JIANG Y, MA Y, LIANG Y Z, et al. Optimization of OTSU lung tissue segmentation algorithm based on fractional sparrow search[J]. Computer Science, 2021, 48(S1): 28-32(in Chinese).

[25] GARCI S, TRIGUERO I, DERRAC J, et al. Evolutionary-based selection of generalized instances for imbalanced classification[J]. Knowledge-Based Systems, 2012, 25(1): 3-12.

[26] 吕鑫, 慕晓冬, 张钧, 等. 混沌麻雀搜索优化算法[J]. 北京航空航天大学学报, 2021, 47(8): 1712-1720.
LYU X, MU X D, ZHANG J, et al. Chaossparrow search optimization algorithm[J]. Journal of Beijing University of Aeronautics and Astronautics, 2021, 47(8): 1712-1720(in Chinese).

[27] DAI Q, LIU J W, LIU Y. Multi-granularity relabeled undersampling algorithm for imbalanced data[J]. Applied Soft Computing, 2022, 124: 109083.

Software defect prediction algorithm for intra-membrane sparrow optimizing ELM

TANG Yu[1]，DAI Qi[2]，YANG Mengyuan[1]，CHEN Lifang[1, 3, *]

(1. College of Science，North China University of Science and Technology，Tangshan 063210，China；
2. College of Information Science and Engineering，China University of Petroleum，Beijing 102249，China；
3. Hebei Key Laboratory of Data Science and Application，Tangshan 063210，China)

Abstract：The original sparrow search algorithm,easy to fall into local extremum in the later stage of iteration, has the problems of low optimization accuracy. Combining the improved sparrow search algorithm with efficient optimization performance and the membrane computing with parallel computing capability, an intra-membrane sparrow optimization algorithm is proposed (IMSSA). The experimental results on ten CEC2017 test functions show that IMSSA has higher optimization accuracy. In addition, to further verify the performance of IMSSA, the extreme learning machine(ELM) parameters are optimized using IMSSA. An intra-membrane sparrow optimal ELM algorithm to be used in software defect predictionis proposed (IMSSA-ELM). The experimental results show that in the 15 public software defect datasets, the prediction performance of the IMSSA-ELM algorithm is significantly better than the other fourcompared algorithms under the two evaluation indicators of G-mean and MCC. The results also show that the IMSSA-ELM algorithm has better prediction accuracy and stability,and have obvious statistical significance in Friedman ranking and Holm's post-hoc test nonparametric tests.

Keywords：improved sparrow search algorithm; membrane calculation; extreme learning machine; optimization algorithm; software defect prediction

Received：2022-05-29；**Accepted**：2022-06-24；**Published Online**：2023-07-04 13：44
URL：link.cnki.net/urlid/11.2625.V.20220701.1556.001
Foundation item：National Natural Science Foundation of China (52074126)
* **Corresponding author**. E-mail：hblg_clf@163.com

http://bhxb.buaa.edu.cn jbuaa@buaa.edu.cn
DOI: 10.13700/j.bh.1001-5965.2022.0454

基于注意力机制与条件卷积的行人重识别方法

姬广凯，王蓉*，彭舒凡
(中国人民公安大学 信息网络安全学院，北京 100038)

摘　　要：行人重识别是计算机视觉领域的一个重要部分，但是容易受到行人图片实际采集环境的影响，导致行人特征表达不充分，进一步导致模型精度不高。提出一种基于注意力机制和条件卷积改进的行人重识别方法，使行人特征得到更充分的表达。将注意力机制引入特征提取网络ResNet50中，对输入图像空间和通道上的关键信息进行加权强化，同时抑制可能的噪声；将条件卷积模块引入主干网络，动态调整卷积核参数，使模型能够在保持高效推理的同时提高容量和性能；利用Market1501、MSMT17和DukeMTMC-ReID主流数据集对改进方法进行评估，Rank1分别提升1.1%、2.4%、1.3%，mAP分别提升0.5%、2.3%、1.3%，结果表明：改进方法能够使行人特征得到更好的表达，识别精度得到提升。

关　键　词：注意力机制；条件卷积；ResNet50；行人重识别；深度学习
中图分类号：V221+.3; TB553
文献标志码：A　　**文章编号**：1001-5965(2024)02-0655-08

行人重识别是实现在不同摄像机视野下检索和匹配特定行人的技术，广泛应用于视频监控和刑事侦查领域。行人重识别的关键在于学习相同行人和不同行人的视觉特征，并加以辨别和区分。然而，实际采集的行人图像容易受到视图、姿势、照明和分辨率变化的影响，使得行人的特征鉴别具有一定的难度[1]。因此，提升识别精度的关键在于如何更加充分地提取和表达行人特征[2]。

依据不同的目的和动机，可以将行人重识别方法划分为基于度量学习、基于局部特征和基于序列特征方法[2]。基于度量学习的方法目的在于学习2个行人客体之间的相似性和差别，从而学习从原始影像到目标特征的对应关系，以便使用特征空间上的距离函数，缩短相同ID行人特征间距，拉长不同ID行人特征间距。Wieczorek等[3]提出了一种新的基于类内质心三元组损失(centroid triplet loss, CTL)模型，这样的聚合表示方法对异常值更具鲁棒性，缩小搜索空间，节约内存，缩短检索时间。Li等[4]提出一种新的搜索空间组合深度空间(combined depth space, CDS)模型，使得网络能够显式地学习组合模式特征，平衡CTL和交叉熵损失，充分发挥了其优点。Yan等[5]提出了一种新的基于成对关系的损失函数—细粒度差异感知损失(fine-grained difference-aware pairwise loss, FIDI)模型，这种损失可以使模型有效地从外观差异中学习表达特征，并且能够更有效地利用图像样本，与CTL模型相比，数据利用率更高。基于局部特征的方法旨在学习行人的局部特征，在特征对齐的基础上进行学习，解决了遮挡、边界检测、视图姿势变化等条件下现有方法存在的缺陷，常用的方法有预定义的条纹分割、多尺度融合、软注意、行人语义提取和全局局部特征学习。Zhou等[6]提出了一个新的行人重识别网络(omni-scale network, OSNet)架构，利用不同空间尺度的特征并协同多尺度的组合特征，以及融合多尺度特征与信道权重动态；张晓伟等[7]利用局部语义特征不变性，将全局特征与局部特征对齐，

收稿日期：2022-06-02；**录用日期**：2022-10-06；**网络出版时间**：2022-10-10 14:38
网络出版地址：link.cnki.net/urlid/11.2625.V.20221009.0706.012
基金项目：国家自然科学基金(62076246)
*通信作者. E-mail：dbdxwangrong@163.com

引用格式：姬广凯，王蓉，彭舒凡. 基于注意力机制与条件卷积的行人重识别方法[J]. 北京航空航天大学学报，2024，50(2)：655-662.
JI G K，WANG R，PENG S F. Person re-identification method based on attention mechanism and CondConv[J]. Journal of Beijing University of Aeronautics and Astronautics，2024，50(2)：655-662 (in Chinese).

提高了行人重识别跨域适应能力。基于序列特征的方法利用视频的序列信息进行行人识别，以视频的格式作为模型的输入，同时使用空间和时间上的互补线索，可以克服图像特征的局限性。Liu 等[8]提出了一种新的时空相关性和拓扑学习框架，通过综合整体的上下文信息，并结合人体生理特征，提出了一种利用上下文信息增强的拓扑图进行多尺度结构的构建，并能有效挖掘出与外观信息互补的线索，从而提高特征表示能力。

本文对 CTL 网络模型[3]进行改进，提出一种基于注意力机制与条件卷积改进的行人重识别方法。以 ResNet50 为主干网络，融合空间通道注意力模块 (convolutional block attention module, CBAM)[9]，使改进模型对特征在空间和通道上的关键信息加以强化和关注；将条件卷积模块(conditionally parameterized convolutions, CondConv)[10]引入主干网络中，动态地调整卷积核的参数，使改进模型能够在保持高效推理的同时提高容量和性能。

1 CTL 网络模型

CTL 网络模型在行人重识别强基线 (ReID strong baseline, RST) 模型[11]的基础上进行改进得到，如图 1 所示。首先，将输入图片通过主干网络 ResNet50 进行特征提取；然后，经过全局平均池化得到特征向量用于计算 CTL、中心损失和三元组损失；再经过批量归一化层得到归一化后的特征用于计算推理阶段的类内质心；最后，经过全连接层进行图像分类并计算分类损失。

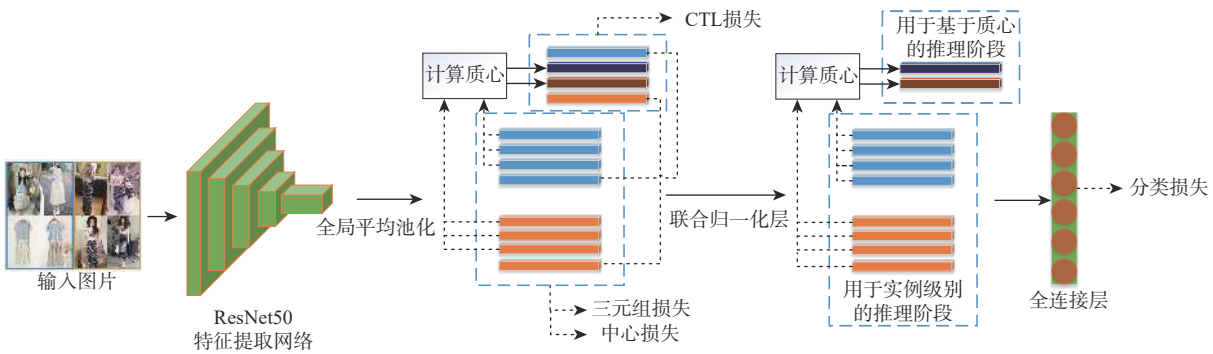


图 1 CTL 网络模型

Fig. 1 CTL network model

三元组损失是行人重识别方法中常用的损失函数，使相同 ID 的行人输入图像在特征空间产生聚类，表达式为[12]

$$L_{\text{triplet}} = [\|f(\boldsymbol{A}) - f(\boldsymbol{P})\|_2^2 - \|f(\boldsymbol{A}) - f(\boldsymbol{N})\|_2^2 + \alpha]_+^{[12]} \quad (1)$$

式中：$[]_+$表示取 0 和括号内变量较大的一个；f为训练阶段的样本图像嵌入函数；$\boldsymbol{A}$为锚图像；$\boldsymbol{P}$为正实例样本；$\boldsymbol{N}$为负实例样本；α为边距参数，在此基础上提出了 CTL 损失，计算式为

$$L_{\text{triplet}} = \left[\|f(\boldsymbol{A}) - \boldsymbol{c}_{\boldsymbol{P}}\|_2^2 - \|f(\boldsymbol{A}) - \boldsymbol{c}_{\boldsymbol{N}}\|_2^2 + \alpha\right]_+^{[12]} \quad (2)$$

CTL 损失测量的是锚图像$\boldsymbol{A}$与类内质心$\boldsymbol{c}_{\boldsymbol{P}}$和不同类质心$\boldsymbol{c}_{\boldsymbol{N}}$的距离，而不是比较锚图像$A$与正负实例样本之间的距离，使得模型训练过程抗干扰能力增强，不易受到单一图像虚假匹配。计算类内质心的方法如下。

在模型的训练阶段，每小批数据包含P类样本，每类样本含有M个实例，故每批数据量大小为$P \times M$。$S_k = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_i, \cdots, \boldsymbol{x}_M\}$表示第$k$类的一组样本；$\boldsymbol{x}_i$表示第$k$类中的第$i$个样本。并在训练时将$S_k$中的样本经过计算得到查询向量$\boldsymbol{q}_k$，其余$M-1$个样本用于计算类内质心原型，计算式为

$$\boldsymbol{c}_{k_P} = \frac{1}{|S_k\{\boldsymbol{q}_k\}|} \sum_{\boldsymbol{x}_i \in S_k \setminus \{\boldsymbol{q}_k\}} f(\boldsymbol{x}_i) \quad (3)$$

在模型的推理阶段，质心被计算为给定类别质心的平均值，计算式为

$$\boldsymbol{c}_k = \frac{1}{|G_k|} \sum_{\boldsymbol{x}_i \in G_k} f(\boldsymbol{x}_i) \quad (4)$$

式中：G_k为第k类行人所有特征向量的集合。

CTL 网络模型在训练和检索过程中将同一 ID 下的行人图片均使用平均质心表示。这样的聚合表示方法对异常值更具鲁棒性，确保提取的行人特征更加稳定。因为每一类行人都以类内质心代表，减少了待查询目标向量的搜索空间，检索时间和存储需求都显著下降，加快了模型的训练速度。

2 本文的改进方法

本文对 CTL 网络模型的主干网络进行改进，将空间通道注意力模块引入 ResNet50 中的每个 Layer 层之间，对输入图像的信息在空间和通道维度上进

行强化和关注；然后将条件卷积引入主干网络，动态地调整卷积核的参数，使改进模型能够在保持高效推理的同时提高容量和性能，整体框架如图 2 所示，其中，σ 为 Sigmoid 函数。

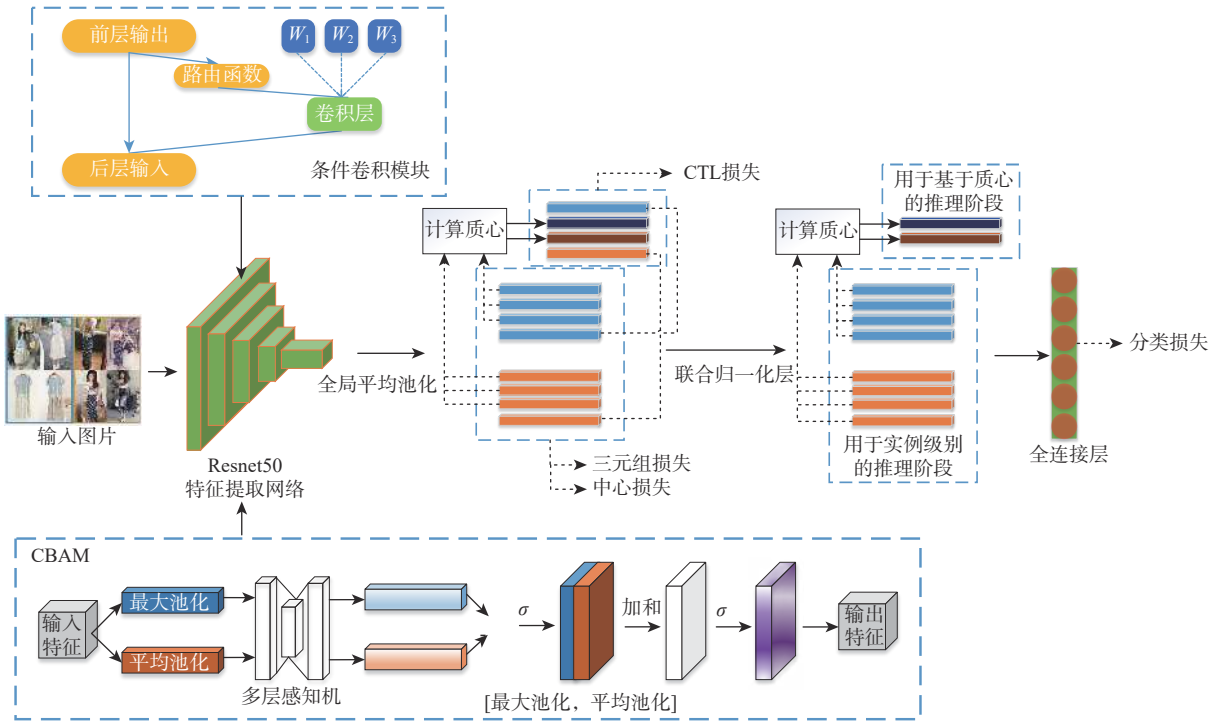


图 2　改进后的 CTL 网络模型

Fig. 2　Improved CTL network model

2.1　CBAM

为了使改进后的 CTL 网络模型关注行人图像在空间和通道上的关键信息，本文在 ResNet50 的 4 个 Layer 之间嵌入 CBAM，使行人特征表达更充分，并抑制可能的噪声。加入 CBAM 的 ResNet50 网络结构如图 3 所示。

CBAM 由通道注意力模块 (channel attention module, CAM) 与空间注意力模块 (spatial attention module, SAM) 这 2 类注意力机制串行连接构成，与单独采用 SAM 或 CAM 相比，特征图中的关键信息更容易被关注。模块中，特征图 $\boldsymbol{F}$ 先经过 CAM 再经过 SAM 得到 $\boldsymbol{F}_{\text{out}}$，如图 4 所示。

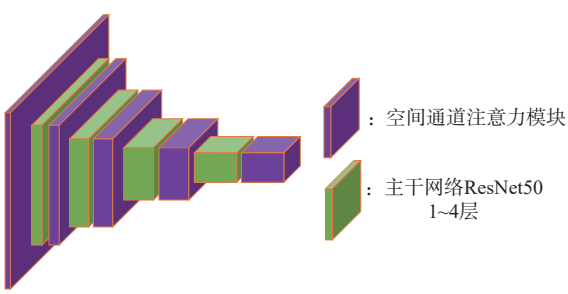


图 3　CBAM-ResNet50 网络结构

Fig. 3　CBAM-ResNet50 network structure

在 CAM 中，特征图 $\boldsymbol{F}_i$ 分别通过最大池化及平均池化得到 MaxPool($\boldsymbol{F}_i$) 和 AvgPool($\boldsymbol{F}_i$)，然后经过多层感知机 (multi layer perceptrons, MLP)，将得到的输出加和生成特征图 $\boldsymbol{F}_{\text{CAM}}$，最后，将其与输入特征图 $\boldsymbol{F}_i$ 相乘得到 $\boldsymbol{F}_{\text{CAMout}}$，计算式为

$$\text{MLP}(\boldsymbol{F}) = W_2(\text{ReLU}(W_1\boldsymbol{F} + \boldsymbol{B}_1)) + \boldsymbol{B}_2 \tag{5}$$

$$\boldsymbol{F}_{\text{max}}^{\text{C}} = \text{MaxPool}(\boldsymbol{F}_i)_{\text{max}} \tag{6}$$

$$\boldsymbol{F}_{\text{avg}}^{\text{C}} = \text{AvgPool}(\boldsymbol{F}_i) \tag{7}$$

$$\boldsymbol{F}_{\text{CAM}} = \sigma\left(\text{MLP}\left(\boldsymbol{F}_{\text{max}}^{\text{C}}\right) + \text{MLP}\left(\boldsymbol{F}_{\text{avg}}^{\text{C}}\right)\right) \tag{8}$$

$$\boldsymbol{F}_{\text{CAMout}} = \boldsymbol{F}_{\text{CAM}} \otimes \boldsymbol{F}_i \tag{9}$$

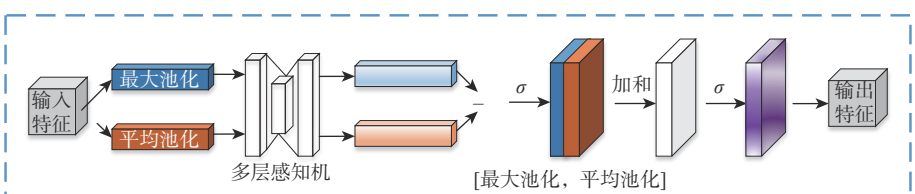


图 4　CBAM 结构

Fig. 4　Structure of CBAM

式中：⊗表示逐点相乘。在输入空间注意力后，特征图$\boldsymbol{F}_{\mathrm{CAMout}}$分别经过最大池化和平均池化，得到2个特征图并将其按通道进行拼接，再经过一层卷积进行降维，生成注意力特征图，最后，将其与输入特征图相乘，从而获得特征加权后的最终结果。计算式如式(10)～(13)所示：

$$\boldsymbol{F}_{\mathrm{max}}^{\mathrm{S}}=\mathrm{MaxPool}(\boldsymbol{F}_{\mathrm{CAMout}}) \tag{10}$$

$$\boldsymbol{F}_{\mathrm{avg}}^{\mathrm{S}}=\mathrm{AvgPool}(\boldsymbol{F}_{\mathrm{CAMout}}) \tag{11}$$

$$\boldsymbol{F}_{\mathrm{SAM}}=\sigma\left(\mathrm{Conv}^{k\times k}\left(\boldsymbol{F}_{\mathrm{max}}^{\mathrm{S}}\oplus\boldsymbol{F}_{\mathrm{avg}}^{\mathrm{S}}\right)\right) \tag{12}$$

$$\boldsymbol{F}_{\mathrm{SAMout}}=\boldsymbol{F}_{\mathrm{SAM}}\otimes\boldsymbol{F}_{\mathrm{CAMout}} \tag{13}$$

式中：⊕表示级联，总计算式如下：

$$\boldsymbol{F}_{\mathrm{out}}=\boldsymbol{F}_{\mathrm{SAMout}}[\boldsymbol{F}_{\mathrm{CAMout}}(\boldsymbol{F}_i)] \tag{14}$$

2.2 条件卷积模块

在网络模型训练过程中往往采用增大数据量及加深网络层数的方式来提升性能，此类方式容易造成计算资源消耗及过拟合等问题。为提升模型容量和精度的同时减少模型复杂度从而节省训练时间，本文引入条件卷积模块，结构如图5所示。

在常规卷积运算中，同一个卷积核在运算时会与每一个输入样本进行卷积运算，其参数只受模型训练时反向传播来决定。而条件卷积模块中的卷积核参数由输入样本决定，其核心思想是让每一个卷积层都拥有多套权重，对卷积核进行加权从而减少卷积次数。可将卷积过程计算式化为

$$\mathrm{Output}(x)=\sigma[(\alpha_1W_1+\ldots+\alpha_iW_i+\ldots+\alpha_nW_n)*x] \tag{15}$$

式中：α_i为根据输入x得到的权重标量，$\alpha_i=r_i(x)$，$r_i(x)$为路由函数；n为α_i的维度。在数学上，常规的卷积运算$\alpha_1W_1*x+\alpha_2W_2*x+\cdots+\alpha_nW_n*x$与条件卷积运算$(\alpha_1W_1+\alpha_2W_2+\cdots+\alpha_nW_n)*x$等价。而条件卷积产生的额外计算量更小，只需要付出一次额外的线性组合运算的代价，条件卷积就可以将计算量缩减至常规卷积的$\frac{1}{n}$。路由函数定义为

$$r(x)=\sigma\left[\mathrm{AvgPool}(x)\boldsymbol{R}\right] \tag{16}$$

式中：$\boldsymbol{R}$为将输入x映射到n个权重标量α_i的学习路由权重矩阵。因此，条件卷积通过路由函数为多套卷积核产生了对应权重，最终加权求和，克服了常规卷积运算只拥有局部感受野的缺陷。可以利用动态的方式更好地覆盖、表达和利用样本之间的关系以提高模型性能。

为定量分析引入条件卷积模块后对于模型复杂度的影响，本文在保持模型容量和性能不变的条件下，对引入条件卷积前后，主干网络ResNet50中Layer3及Layer4的参数量和计算量进行对比，令输入的batchsize=1。模型对比结果如表1所示。

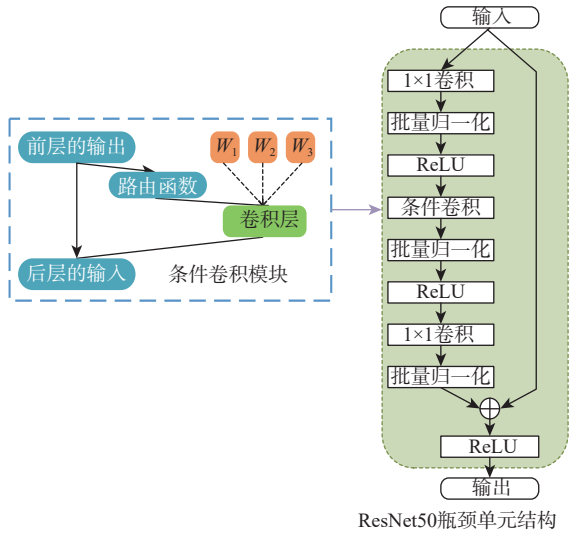


图5 条件卷积模块结构图

Fig. 5 Structure of CondConv model

表1 引入条件卷积前后参数量与计算量对比

Table 1 Comparison of number of parameters and flops before and after introduction of CondConv

模块	参数量	计算量
Basic Layer3	7 098 368	959 119 360
CondConv Layer3	3 564 050	506 139 136
Basic Layer4	14 964 736	529 301 504
CondConv Layer4	7 891 465	302 813 696

可以看到在保持容量和性能不变的条件下，引入条件卷积模块后，模型具有更少的参数量与计算量。并且与标准卷积相比，条件卷积可以为每一个实例计算一个卷积核参数，从而减少卷积次数，提升学习效率。

3 仿真实验结果与分析

3.1 数据集

实验使用数据集Market1501[13]、MSMT17[14]和DukeMTMC-ReID[15]，如表2所示。Market1501[13]数据集中有1 501个行人ID、32 668张图像；MSMT17[14]数据集有4 101个行人ID、126 441张图像。DukeMTMC-ReID[15]数据集中有1 404个行人ID、36 411张图像。DukeMTMC-ReID[15]和Market150[13]数据集中训练集和测试集的比例为1∶1。MSMT17[14]训练集的行人ID数为1 041，图像数为32 621；测试集的行人ID数为3 060，图像数为93 820。

3.2 实验设置

本实验在NVIDIA TITAN V上进行实现，pytorch训练框架版本为1.10。模型的主干网络为ResNet50。

表 2　数据集简介

Table 2　Introduction of datasets

数据集	公开年份	摄像机数/个	ID数	图片数量/张	评测
Market1501[13]	2015	6	1 501	32 668	CMC+mAP
MSMT17[14]	2018	15	4 101	126 441	CMC+mAP
DukeMTMC-ReID[15]	2017	8	1 404	36 411	CMC+mAP

训练 epoch 为 120, batchsize 为 64, 实验设置初始学习率为 0.000 35, 在训练迭代到第 40 和第 70 个 epoch 时学习率衰减为原始的 1/10, 最后一个卷积层的步长设为 1, 设置中心损失的权重为 0.000 5, 其余损失的权重均为 1; 在推理阶段使用标准化后的向量进行余弦相似性的距离度量, 并使用交叉视图匹配设置, 确保在检索过程中, 对于每个查询, 排除同一摄像头捕获的样本。

3.3　仿真实验

在所使用的 3 个数据集上对改进模型进行训练与测试, 评价指标为平均精度均值(mean avetage precision, mAP)和累积匹配曲线(cummulative match characteristic, CMC)。Rank-k 表示模型推理阶段前 k 位检索结果中包含查询目标的比例。

实验 1: 嵌入 CBAM。由表 3 可知, 将 CBAM 嵌入 ResNet50 网络后, 模型在 3 个数据集上的表现性能均有所提升。在 Market1501[13] 数据集上, mAP 和 Rank-1 分别提升了 0.3% 和 0.8%; 在 MSMT17[14] 数据集上, mAP 和 Rank-1 分别提升了 1.6% 和 1.5%; 在 DukeMTMC-ReID[15] 数据集上, mAP 和 Rank-1 分别提升了 0.9% 和 1.1%。

本实验分别测试了 CBAM 中空间卷积核(spatialkernel)的大小对实验结果的影响, 设置了 3、5、7 3 种不同大小的卷积核。实验结果表明, 前 2 层注意力模块卷积核大小设置为 5, 后 3 层注意力模块卷积核设置为 3 时, 实验效果达到最佳。

实验 2: 引入 Cond Conv。由表 4 可知, 引入 Cond Conv 之后, 模型性能提升。在 Market1501[13] 数据集上, mAP 和 Rank-1 分别提升了 0.4% 和 1.0%; 在 MSMT17[14] 数据集上, mAP 和 Rank-1 分别提升了 1.3% 和 1.1%; 在 DukeMTMC-ReID[15] 数据集上, mAP 和 Rank-1 分别提升了 0.7% 和 0.8%。

表 3　基于空间通道注意力机制改进后的 CTL 模型在不同数据集上的性能对比

Table 3　Performance comparison of improved CTL model based on CBAM attention mechanism on different datasets　%

模型	Rank-1			mAP		
	Market1501[13]	MSMT17[14]	DukeMTMC-ReID[15]	Market1501[13]	MSMT17[14]	DukeMTMC-ReID[15]
CTL	97.5	89.5	95.4	98.3	91.3	96.1
引入注意力机制后	98.3	91.0	96.5	98.6	92.9	97.0

表 4　基于条件卷积模块改进后的 CTL 模型在不同数据集上的性能对比

Table 4　Performance comparison of improved CTL model based on CondConv module on different datasets　%

模型	Rank-1			mAP		
	Market1501[13]	MSMT17[14]	DukeMTMC-ReID[15]	Market1501[13]	MSMT17[14]	DukeMTMC-ReID[15]
CTL	97.5	89.5	95.4	98.3	91.3	96.1
引入条件卷积后	98.5	90.6	96.2	98.7	92.6	96.8

本实验测试了将条件卷积模块嵌入主干网络 ResNet50 不同 Layer 中, 结果表明在每一个 Layer 中均嵌入条件卷积模块效果最佳, 卷积核大小设置为 3。

实验 3: 同时嵌入 CBAM 和条件卷积模快。由表 5 可知, 模型精度在 MSMT17[14] 数据集上的提升最为显著, 其次是 DukeMTMC-ReID[15] 数据集。在 Market1501[13] 数据集上提升最小, 原因可能是原始模型识别精度已足够高。在 Market1501[13] 数据集上, mAP 和 Rank-1 分别提升了 0.5% 和 1.1%; 在 MSMT17[14] 数据集上, mAP 和 Rank-1 分别提升了 2.3% 和 2.4%; 在 DukeMTMC-ReID[15] 数据集上, mAP 和 Rank-1 均提升了 1.3%。

图 6～图 8 给出了原模型及加入不同模块后

表5 基于注意力机制和条件卷积改进后的CTL模型在不同数据集上的性能对比

Table 5 Performance comparison of improved CTL model based on attention mechanism and CondConv on different datasets %

模型	Rank-1			mAP		
	Market1501[13]	MSMT17[14]	DukeMTMC-ReID[15]	Market1501[13]	MSMT17[14]	DukeMTMC-ReID[15]
CTL	97.5	89.5	95.4	98.3	91.3	96.1
同时引入注意力机制和条件卷积后	98.6	91.9	96.7	98.8	93.6	97.4

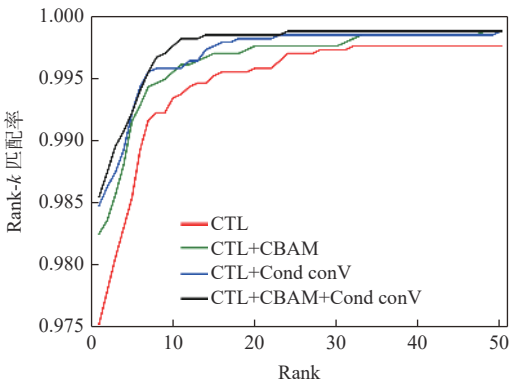


图6 数据集Market1501[13]上的CMC曲线对比

Fig. 6 CMC curve comparison chart on dataset Market1501[13]

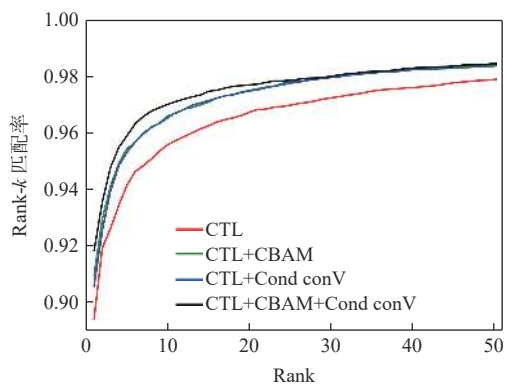


图7 数据集MSMT17[14]上的CMC曲线对比

Fig. 7 CMC curve comparison chart on dataset MSMT17[14]

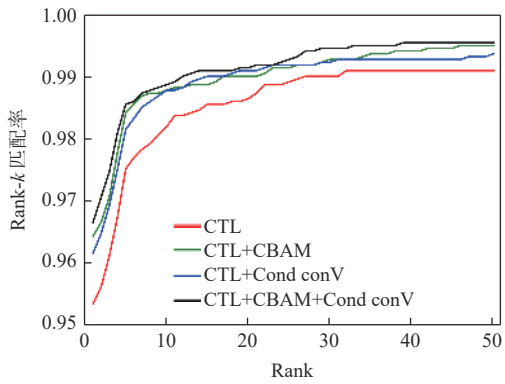


图8 数据集DukeMTMC-ReID[15]上的CMC曲线对比

Fig. 8 CMC curve comparison chart on dataset DukeMTMC-ReID[15]

CMC评价曲线对比。可以看出，与原模型相比，改进后的模型在识别精度上得到提升。

表6为改进方法与最新方法的对比。与改进前的方法相比，改进方法在Market1501[13]、MSMT-17[14]和DukeMTMC-ReID[15]这3个数据集上均提升了识别精度。与其他方法相比，改进方法在识别精度上也有较为明显的优势，证明了改进方法是有效的。

表6 改进方法与最新方法的性能比较

Table 6 Performance comparison between improved method and latest methods %

方法	Rank-1			mAP		
	Market1501[13]	MSMT17[14]	DukeMTMC-ReID[15]	Market1501[13]	MSMT17[14]	DukeMTMC-ReID[15]
CDNet[4]	95.1	78.9	88.6	86.0	54.7	76.8
OSNet[6]	94.8	79.1	88.7	86.7	55.1	76.6
PAT[16]	95.4		88.8	88.0		78.2
HLGAT[17]	97.5		92.7	93.4		87.3
CTL	97.5	89.5	95.4	98.3	91.3	96.1
CTL +CBAM	98.3	91.0	96.5	98.6	92.9	97.0
CTL +CondConv	98.5	90.6	96.2	98.7	92.6	96.8
CTL+CBAM+CondConv	98.6	91.9	96.7	98.8	93.6	97.4

4 结 论

1）本文改进方法可实现较为优异的识别精度，在Market1501、MSMT17和DukeMTMC-ReID这3个主流数据集上进行测试，Rank1分别提升1.1%、2.4%、1.3%，mAP分别提升0.5%、2.3%、1.3%。

2）经过消融实验后发现，注意力机制和条件卷积分别对模型性能有所增强。其中注意力机制是

对行人图片中的关键特征进行加权强化, 有助于抑制噪声干扰。

3) 经过参数量和计算量的对比, 可以发现条件卷积模块对于模型容量有所提升, 进而提高识别精度。

本文改进方法简洁有效、易于实现, 使行人特征得到更充分的表达。为进一步提升性能, 仍需优化各项参数。

参考文献（References）

[1] MING Z Q, ZHU M K, WANG X K, et al. Deep learning-based person re-identification methods: A survey and outlook of recent works[J]. Image and Vision Computing, 2022, 119: 104394.

[2] 孙义博, 张文靖, 王蓉, 等. 基于通道注意力机制的行人重识别方法[J]. 北京航空航天大学学报, 2022, 48(5): 881-889.
SUN Y B, ZHANG W J, WANG R, et al. Pedestrian re-identification method based on channel attention mechanism[J]. Journal of Beijing University of Aeronautics and Astronautics, 2022, 48(5): 881-889 (in Chinese).

[3] WIECZOREK M, RYCHALSKA B, DąBROWSKI J. On the unreasonable effectiveness of centroids in image retrieval[C]//Neural Information Processing: 28th International Conference. Berlin: Springer, 2021: 212-223.

[4] LI H J, WU G J, ZHENG W S. Combined depth space based architecture search for person re-identification[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2021: 6729-6738.

[5] YAN C, PANG G S, BAI X, et al. Beyond triplet loss: Person re-identification with fine-grained difference-aware pairwise loss[J]. IEEE Transactions on Multimedia, 2021, 24: 1665-1677.

[6] ZHOU K Y, YANG Y X, CAVALLARO A, et al. Learning generalisable omni-scale representations for person re-identification[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021, 44(9): 5056-5069.

[7] 张晓伟, 吕明强, 李慧. 基于局部语义特征不变性的跨域行人重识别[J]. 北京航空航天大学学报, 2020, 46(9): 1682-1690.
ZHANG X W, LYU M Q, LI H. Cross-domain person re-identification based on partial semantic feature invariance[J]. Journal of Beijing University of Aeronautics and Astronautics, 2020, 46(9): 1682-1690 (in Chinese).

[8] LIU J W, ZHA Z J, WU W, et al. Spatial-temporal correlation and topology learning for person re-identification in videos[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2021: 4370-4379.

[9] WOO S, PARK J, LEE J Y, et al. Cbam: Convolutional block attention module[C]//Proceedings of the European conference on computer vision. Munich: ECCV, 2018: 3-19.

[10] YANG B, BENDER G, LE Q V, et al. Condconv: Conditionally parameterized convolutions for efficient inference[EB/OL]. (2019-01-10) [2022-02-19]. https://arxiv.org/abs/1904.04071.html.

[11] WIECZOREK M, MICHALOWSKI A, WROBLEWSKA A, et al. A strong baseline for fashion retrieval with person re-identification models[C]//International Conference on Neural Information Processing. Berlin: Springer, 2020: 294-301.

[12] SCHROFF F, KALENICHENKO D, PHILBIN J. Facenet: A unified embedding for face recognition and clustering[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2015: 815-823.

[13] ZHENG L, SHEN L, TIAN L Y, et al. Scalable person re-identification: A benchmark[C]//Proceedings of the IEEE International Conference on Computer Vision. Piscataway: IEEE Press, 2015: 1116-1124.

[14] WEI L H, ZHANG S L, GAO W, et al. Person transfer GAN to bridge domain gap for person re-identification[C]//Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2018 : 79-88.

[15] ZHENG Z D, ZHENG L, YANG Y. Unlabeled samples generated by gan improve the person re-identification baseline in vitro[C]// Proceedings of the IEEE International Conference on Computer Vision. Piscataway: IEEE Press, 2017: 3754-3762.

[16] LI Y L, HE J F, ZHANG T Z, et al. Diverse part discovery: Occluded person re-identification with part-aware trans-former[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2021: 2898–2907.

[17] ZHANG Z, ZHANG H J, LIU S. Person re-identification using hetero-geneous local graph attention networks[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2021: 12136–12145.

Person re-identification method based on attention mechanism and CondConv

JI Guangkai，WANG Rong*，PENG Shufan

(Police Information Engineering and Network Secuvity College, People's Public Security University of China，Beijing 100038，China)

Abstract：Person Re-identification is an important part of the field of computer vision, but it is easily affected by the actual collection environment of person images, resulting in insufficient expression of person features and further leading to low model accuracy. An improved person re-identification method based on attention mechanism and CondConv is proposed to fully express pedestrian features. The attention mechanism is introduced into the feature extraction network ResNet50, and the key information in the input image space and channel is weighted, while suppressing possible noise. The CondConv is introduced into the backbone network and the convolution kernel parameters are dynamically adjusted to improve the capacity and performance of the model while maintaining efficient reasoning. Mainstream data sets such as Market1501, MSMT17 and DukeMTMC-ReID are used to evaluate the improved method. Rank-1 is increased by 1.1%, 2.4% and 1.3% respectively, and mAP is increased by 0.5%, 2.3% and 1.3%; respectively. The results show that the improved method can better express person features and improve recognition accuracy.

Keywords：attention mechanism; CondConv; ResNet50; person re-identification; deep learning

Received：2022-06-02；**Accepted**：2022-10-06；**Published Online**：2022-10-10 14：38

URL：link.cnki.net/urlid/11.2625.V.20221009.0706.012

Foundation item：National Natural Science Foundation of China (62076246)

* **Corresponding author**. E-mail： dbdxwangrong@163.com

http://bhxb.buaa.edu.cn jbuaa@buaa.edu.cn
DOI: 10.13700/j.bh.1001-5965.2022.0478

基于损失平滑的对抗样本攻击方法

黎妹红[1,2,*]，金双[1,2]，杜晔[1,2]

(1. 北京交通大学 智能交通数据安全与隐私保护技术北京市重点实验室，北京 100044；
2. 北京交通大学 计算机与信息技术学院，北京 100044)

摘 要：深度神经网络(DNNs)容易受到对抗样本的攻击，现有基于动量的对抗样本生成方法虽然可以达到接近100%的白盒攻击成功率，但是在攻击其他模型时效果仍不理想，黑盒攻击成功率较低。针对此，提出一种基于损失平滑的对抗样本攻击方法来提高对抗样本的可迁移性。在每一步计算梯度的迭代过程中，不直接使用当前梯度，而是使用局部平均梯度来累积动量，以此来抑制损失函数曲面存在的局部振荡现象，从而稳定更新方向，逃离局部极值点。在ImageNet数据集上的大量实验结果表明：所提方法与现有基于动量的方法相比，在单个模型攻击实验中的平均黑盒攻击成功率分别提升了38.07%和27.77%，在集成模型攻击实验中的平均黑盒攻击成功率分别提升了32.50%和28.63%。

关 键 词：深度神经网络；对抗样本；黑盒攻击；损失平滑；人工智能安全
中图分类号：V221+.3; TP309; TP391.41
文献标志码：A **文章编号**：1001-5965(2024)02-0663-08

尽管深度神经网络(deep neural networks, DNNs)在各类计算机视觉任务上获得了先进的性能，如图像分类[1]、目标检测[2]、语义分割[3]等，但最近的研究表明其非常容易受到攻击者在正常样本上恶意添加人眼难以察觉的细微扰动，从而误导模型产生错误结果对抗样本的攻击[4]。对抗样本的存在引发了对安全敏感性应用的广泛担忧，如人脸识别[5]、自动驾驶[6]等。

近年来，对抗样本的生成算法成为了广泛的研究热点，因为其不仅可以帮助发现模型存在的漏洞[7]，还可以用来提高模型的鲁棒性[8]。此外，一些研究还表明对抗样本具有可迁移性[9]，这意味着利用某一个特定模型生成的对抗样本也可以成功地攻击其他模型。对抗样本可迁移性的存在使得现实世界中攻击者不知道模型结构和参数的黑盒攻击成为可能，引发了严重的安全问题。

在攻击者可以获取目标模型结构和参数的白盒设置中，很多现有的攻击方法能够成功地生成对抗样本，例如快速梯度符号法(fast gradient sign method, FGSM)[10]、迭代快速梯度符号法(iterative fast gradient sign method, I-FGSM)[11]、投影梯度下降(projected gradient descent, PGD)法[12]等，并且能够达到接近100%的攻击成功率。然而，采用这些方法生成的对抗样本可迁移性较差，黑盒攻击成功率通常很低，特别是攻击具有防御机制(如对抗训练[13]、输入去噪[14]等)的模型。为解决这个问题，研究人员提出可以通过更好的优化算法(如动量梯度[15]、Nesterov加速梯度[16]等)或攻击集成模型[17]等方法来提高黑盒攻击成功率。尽管如此，白盒攻击和基于迁移的黑盒攻击在攻击性能(攻击成功率)上仍然存在很大的差距。

受到现有基于动量攻击方法的启发，本文提出了一种基于损失平滑的对抗样本攻击方法来提高对抗样本的可迁移性。与之不同的是，现有方法利

收稿日期：2022-06-11；**录用日期**：2022-09-06；**网络出版时间**：2022-09-15 11:12
网络出版地址：link.cnki.net/urlid/11.2625.V.20220914.1950.002
基金项目：国家重点研发计划项目(2020YFB2103800,2020YFB2103802)
*通信作者. E-mail：mhli1@bjtu.edu.cn

引用格式：*黎妹红，金双，杜晔. 基于损失平滑的对抗样本攻击方法[J]. 北京航空航天大学学报，2024，50(2)：663-670.*
LI M H，JIN S，DU Y. Adversarial attack method based on loss smoothing[J]. Journal of Beijing University of Aeronautics and Astronautics，2024，50(2)：663-670 (in Chinese).

用当前数据点的梯度来累积动量，而本文方法不仅考虑当前的数据点，而且充分利用其邻域。具体地说，使用当前数据点某个邻域范围内的局部平均梯度值来替代其单个数据点梯度值，从而有效平滑损失函数曲面的局部振荡，稳定更新方向，避免陷入局部极值点，生成迁移性更强的对抗样本。

在ImageNet数据集[18]上的实验结果表明，与现有基于动量的攻击方法—动量迭代快速梯度符号法（momentum iterative fast gradient sign method, MI-FGSM）[19]和Nesterov迭代快速梯度符号法（nesterov iterative fast gradient sign method, NI-FGSM）[20]相比，本文方法显著提高了黑盒攻击成功率。例如，同样是基于Inception-v3（Inc-v3）模型生成对抗样本，本文方法与现有方法相比，攻击成功率平均提升了约40%。

综上所述，本文方法可以与现有基于动量的方法很好地集成，具有良好的可扩展性。大量实验结果验证了本文方法的有效性，显著提高了现有方法的黑盒攻击成功率和对抗样本的可迁移性。

1 相关工作

1.1 对抗攻击

定义$\boldsymbol{x}$为原始输入图像，$\boldsymbol{y}$为其正确标签，$\boldsymbol{\theta}$为模型的参数，$f(\boldsymbol{x};\boldsymbol{\theta})$为模型的输出，$J(\boldsymbol{x},\boldsymbol{y};\boldsymbol{\theta})$为模型的损失函数，通常为交叉熵损失函数。对于对抗样本生成过程，通常会使用无穷范数L_∞对添加的最大扰动量ε进行限制，目标是在原始样本$\boldsymbol{x}$的邻域内寻找对抗样本$\boldsymbol{x}^{\text{adv}}$，使其在满足$L_\infty$范数约束的限制下有$\|\boldsymbol{x}^{\text{adv}}-\boldsymbol{x}\|_\infty<\varepsilon$，尽可能使模型的损失函数最大化，让模型产生错误的预测结果，即$f(\boldsymbol{x};\boldsymbol{\theta})\neq f\left(\boldsymbol{x}^{\text{adv}};\boldsymbol{\theta}\right)$。过程如下：

$$\begin{aligned}&\arg\max_{\boldsymbol{x}^{\text{adv}}} J(\boldsymbol{x},\boldsymbol{y};\boldsymbol{\theta})\\&\text{s.t.}\ \|\boldsymbol{x}^{\text{adv}}-\boldsymbol{x}\|_\infty<\varepsilon\end{aligned}\tag{1}$$

近年来提出了许多对抗攻击方法，包括基于梯度的方法、基于优化的方法、基于决策的方法等，其中基于梯度的方法最为典型、研究最为广泛。而本文研究内容也是基于梯度的方法，因此，本节将首先介绍基于梯度的对抗攻击方法。

1.1.1 快速梯度符号法

FGSM[10]是第1个基于梯度的对抗攻击方法，也是最为经典的一个方法。方法实现过程如式(2)所示，首先，计算模型损失函数$J(\boldsymbol{x},\boldsymbol{y};\boldsymbol{\theta})$对于输入图像$\boldsymbol{x}$的梯度；然后，通过符号函数$\text{sign}(\cdot)$对梯度进行归一化，使得计算得到的对抗扰动满足$L_\infty$范数的限制；最后，将对抗扰动添加到输入图像上即可生成对抗样本$\boldsymbol{x}^{\text{adv}}$。然而FGSM由于只计算了一次梯度，不够精确，因此攻击能力有限。

$$\boldsymbol{x}^{\text{adv}}=\boldsymbol{x}+\varepsilon\text{sign}(\nabla_{\boldsymbol{x}}J(\boldsymbol{x},\boldsymbol{y};\boldsymbol{\theta}))\tag{2}$$

1.1.2 迭代快速梯度符号法

FGSM[10]由于只进行了一步梯度更新，计算得到的对抗扰动并不精确，因此，其攻击效果较差。针对该问题，I-FGSM[11]将向输入图像添加的单步对抗扰动分为多步迭代完成，每一轮迭代只添加较小的扰动，实现过程如下：

$$\boldsymbol{x}_{t+1}^{\text{adv}}=\boldsymbol{x}_t^{\text{adv}}+\alpha\text{sign}\left[\nabla_{\boldsymbol{x}}J\left(\boldsymbol{x}_t^{\text{adv}},\boldsymbol{y};\boldsymbol{\theta}\right)\right]\tag{3}$$

式中：α为迭代步长；t为迭代轮数；$\boldsymbol{x}_t^{\text{adv}}$为第$t$轮迭代生成的对抗样本。

结果表明，与FGSM[10]相比，I-FGSM[11]具有更强的白盒攻击能力，但代价是生成的对抗样本可迁移性降低，其黑盒攻击成功率相对FGSM[10]有所下降；出现这种现象的主要原因是I-FGSM[11]在每一轮迭代中都会将对抗样本沿梯度符号的方向贪婪地移动，导致损失函数容易陷入较差的局部极大值，出现“过拟合”现象。

1.1.3 动量迭代快速梯度符号法

为解决I-FGSM[11]在迭代过程中损失函数容易陷入局部极大值的问题，Dong等[19]提出了一种基于动量的迭代方法MI-FGSM。实现过程如下：

$$\boldsymbol{g}_{t+1}=\mu\boldsymbol{g}_t+\frac{\nabla_{\boldsymbol{x}}J\left(\boldsymbol{x}_t^{\text{adv}},\boldsymbol{y};\boldsymbol{\theta}\right)}{\|\nabla_{\boldsymbol{x}}J\left(\boldsymbol{x}_t^{\text{adv}},\boldsymbol{y};\boldsymbol{\theta}\right)\|_1}\tag{4}$$

$$\boldsymbol{x}_{t+1}^{\text{adv}}=\boldsymbol{x}_t^{\text{adv}}+\alpha\text{sign}(\boldsymbol{g}_{t+1})\tag{5}$$

式中：$\boldsymbol{g}_t$为前t轮迭代累积的梯度值；μ为动量项的衰减因子；初始状态下令$\boldsymbol{x}_0^{\text{adv}}=\boldsymbol{x}$，$\boldsymbol{g}_0=\boldsymbol{0}$。

通过将动量项整合到对抗攻击的迭代过程中，该方法可以有效稳定更新方向，避免陷入局部极大值，从而生成更具迁移性的对抗样本。结果表明，MI-FGSM[19]不仅保持了优秀的白盒攻击能力，而且有效提高了黑盒攻击成功率和对抗样本的可迁移性。

1.1.4 Nesterov迭代快速梯度符号法

为进一步提高黑盒攻击成功率，Lin等[20]提出了一种基于Nesterov加速梯度的迭代方法NI-FGSM。实现过程如式(6)～式(8)所示。该方法在每一步迭代过程中，不直接使用当前位置计算梯度值，而是在之前累积的梯度方向上向前“看一步”，根据下一步位置计算梯度值。

通过将Nesterov加速梯度整合到对抗攻击的迭代过程中，NI-FGSM[20]进一步提高了黑盒攻击成功率和对抗样本的可迁移性。

$$\boldsymbol{x}_t^{\text{nes}} = \boldsymbol{x}_t^{\text{adv}} + \alpha\mu\boldsymbol{g}_t \tag{6}$$

$$\boldsymbol{g}_{t+1} = \mu\boldsymbol{g}_t + \frac{\nabla_{\boldsymbol{x}} J\left(\boldsymbol{x}_t^{\text{nes}}, \boldsymbol{y}; \boldsymbol{\theta}\right)}{\| \nabla_{\boldsymbol{x}} J\left(\boldsymbol{x}_t^{\text{nes}}, \boldsymbol{y}; \boldsymbol{\theta}\right) \|_1} \tag{7}$$

$$\boldsymbol{x}_{t+1}^{\text{adv}} = \boldsymbol{x}_t^{\text{adv}} + \alpha \text{sign}(\boldsymbol{g}_{t+1}) \tag{8}$$

式中: 初始状态下令 $\boldsymbol{x}_0^{\text{adv}} = \boldsymbol{x}$, $\boldsymbol{g}_0 = \boldsymbol{0}$。

1.2　对抗防御

面对对抗样本的威胁, 近几年各种对抗样本防御方法被提出。其中, 对抗训练[21] 防御方法应用最为广泛, 该方法将生成的对抗样本加入训练数据集中和正常样本一起参与模型的训练, 使模型学习到对抗扰动的特征, 从而增强模型的鲁棒性。Tramer 等[22] 利用其他模型转移来的对抗扰动来进一步增强训练数据, 这种方法被称为集成对抗训练。然而, 这种对抗训练方法作为最强大、研究最广泛的防御方法之一, 计算开销通常较大, 并且难以扩展到大规模数据集和复杂的神经网络[23]。

Guo 等[24] 使用如 JPEG 压缩、总方差最小化等图像变换消除图像中的对抗性扰动。Xie 等[25] 对输入图像采用随机调整大小和填充来消除对抗性扰动。Liao 等[26] 提出了高级表征引导去噪器(high-level representation guided denoiser, HGD)来净化对抗样本。但这些基于图像变换和输入去噪的方法并不能保证完全消除对抗性扰动。

2　本文方法

在本节中, 首先详细描述本文方法, 然后分析本文方法和现有攻击方法之间的区别。

2.1　基于损失平滑的对抗攻击方法

正如第 1 节的内容中提到: I-FGSM[11] 在迭代过程中损失函数容易陷入局部极大值; MI-FGSM[19] 将动量项集成到 I-FGSM[11] 中以稳定更新方向、逃离局部极值点; NI-FGSM[20] 进一步引入 Nesterov 加速梯度, 通过利用其前瞻性来提高对抗样本的可迁移性。

本文观察到上述方法都只考虑了对抗样本优化路径上数据点的梯度信息, 即仅仅考虑了 $\boldsymbol{x}_0^{\text{adv}} = \boldsymbol{x}, \boldsymbol{x}_1^{\text{adv}}, \cdots, \boldsymbol{x}_{t-1}^{\text{adv}}, \boldsymbol{x}_t^{\text{adv}}, \cdots, \boldsymbol{x}_T^{\text{adv}} = \boldsymbol{x}^{\text{adv}}$, 而没有利用这些数据点附近的梯度信息, 由于损失函数曲面存在局部振荡现象, 所以曲面上某一数据点的梯度信息可能并不稳定可靠。考虑一个简单的情况, 如图 1 所示。

假设模型A是用来生成对抗样本的源模型, 模型B是对抗样本攻击的目标模型, $\boldsymbol{g}_{\text{A}}$ 和 $\boldsymbol{g}_{\text{B}}$ 分别为模型A和模型B损失函数曲面上某一数据点的梯度, 由图 1 可见, 由于模型A的损失函数曲面存在振荡现象, 局部非光滑, 这导致 $\boldsymbol{g}_{\text{A}}$ 的方向和 $\boldsymbol{g}_{\text{B}}$ 相差较大, 内积 $\langle \boldsymbol{g}_{\text{A}}, \boldsymbol{g}_{\text{B}} \rangle$ 较小, 使得利用 $\boldsymbol{g}_{\text{A}}$ 生成的对抗扰动并不能很好地迁移到模型B上, 对抗样本的可迁移性大大降低。一种有效的解决方式是对模型A的损失曲面作局部平滑处理, 使用局部平滑处理之后的数据点梯度 $\boldsymbol{G}_{\text{A}}$ 替代原来的 $\boldsymbol{g}_{\text{A}}$ 去生成对抗扰动, 由于 $\boldsymbol{G}_{\text{A}}$ 的方向和 $\boldsymbol{g}_{\text{B}}$ 更加接近, 内积 $\langle \boldsymbol{G}_{\text{A}}, \boldsymbol{g}_{\text{B}} \rangle$ 相较于 $\langle \boldsymbol{g}_{\text{A}}, \boldsymbol{g}_{\text{B}} \rangle$ 更大, 所以使用 $\boldsymbol{G}_{\text{A}}$ 生成的对抗扰动能更好地迁移到模型B上, 对抗样本的可迁移性得到有效提升。

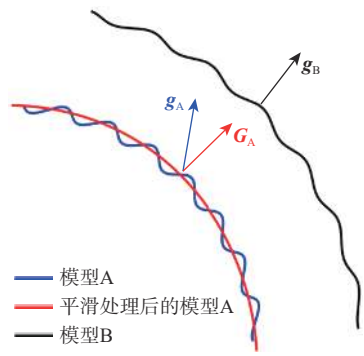


图 1　非平滑损失曲面示例

Fig. 1　Example of non-smooth loss function surface

基于这种思想, 本文提出了基于损失平滑的对抗攻击方法, 其总体思路如图 2 所示。

具体地, 本文方法使用某个数据点邻域内梯度的数学期望 $E_{\varepsilon'}(\boldsymbol{x})$ 来代替这个点的梯度值, 如式 (9) 所示:

$$E_{\varepsilon'}(\boldsymbol{x}) = E_{\|\boldsymbol{x}'-\boldsymbol{x}\|_\infty < \varepsilon'}\left[\nabla_{\boldsymbol{x}} J(\boldsymbol{x}', \boldsymbol{y}; \boldsymbol{\theta})\right] \tag{9}$$

式中: $E_{\varepsilon'}(\boldsymbol{x})$ 为 $\boldsymbol{x}$ 邻域内梯度的数学期望, ε' 为 $\boldsymbol{x}$ 邻域范围的上界, 本文设置为 $\varepsilon' = \beta\varepsilon$, 其中, β 为超参数。

但实际上, 由于输入空间的连续性, 无法直接计算 $E_{\|\boldsymbol{x}'-\boldsymbol{x}\|_\infty < \varepsilon'}\left[\nabla_{\boldsymbol{x}} J(\boldsymbol{x}', \boldsymbol{y}; \boldsymbol{\theta})\right]$, 因此, 本文通过在 $\boldsymbol{x}$ 的邻域范围内采样 N 个样本来近似该表达式的值, 即

$$E_{\varepsilon'}(\boldsymbol{x}) \approx \frac{1}{N}\sum_{i=1}^{n} \nabla_{\boldsymbol{x}} J\left(\boldsymbol{x}^i, \boldsymbol{y}; \boldsymbol{\theta}\right) \tag{10}$$

式中: $\boldsymbol{x}^i = \boldsymbol{x} + \boldsymbol{r}_i$, $\boldsymbol{r}_i \sim U\left(-(\beta\cdot\varepsilon)^d, (\beta\cdot\varepsilon)^d\right)$, $U\left(a^d, b^d\right)$ 表示 d 维均匀分布。

计算得到 $E_{\varepsilon'}(\boldsymbol{x})$ 后, 通过其值来累积动量, 分别替代 MI-FGSM[19] 计算式 (4) 中的 $\nabla_{\boldsymbol{x}} J\left(\boldsymbol{x}_t^{\text{adv}}, \boldsymbol{y}; \boldsymbol{\theta}\right)$ 和 NI-FGSM[20] 计算式 (7) 中的 $\nabla_{\boldsymbol{x}} J\left(\boldsymbol{x}_t^{\text{nes}}, \boldsymbol{y}; \boldsymbol{\theta}\right)$, 如下:

$$\boldsymbol{g}_{t+1} = \mu\boldsymbol{g}_t + \frac{E_{\varepsilon'}(\boldsymbol{x})}{\| E_{\varepsilon'}(\boldsymbol{x}) \|_1} \tag{11}$$

然后更新 $\boldsymbol{x}_{t+1}^{\text{adv}}$:

$$\boldsymbol{x}_{t+1}^{\text{adv}} = \boldsymbol{x}_t^{\text{adv}} + \alpha \text{sign}(\boldsymbol{g}_{t+1}) \tag{12}$$

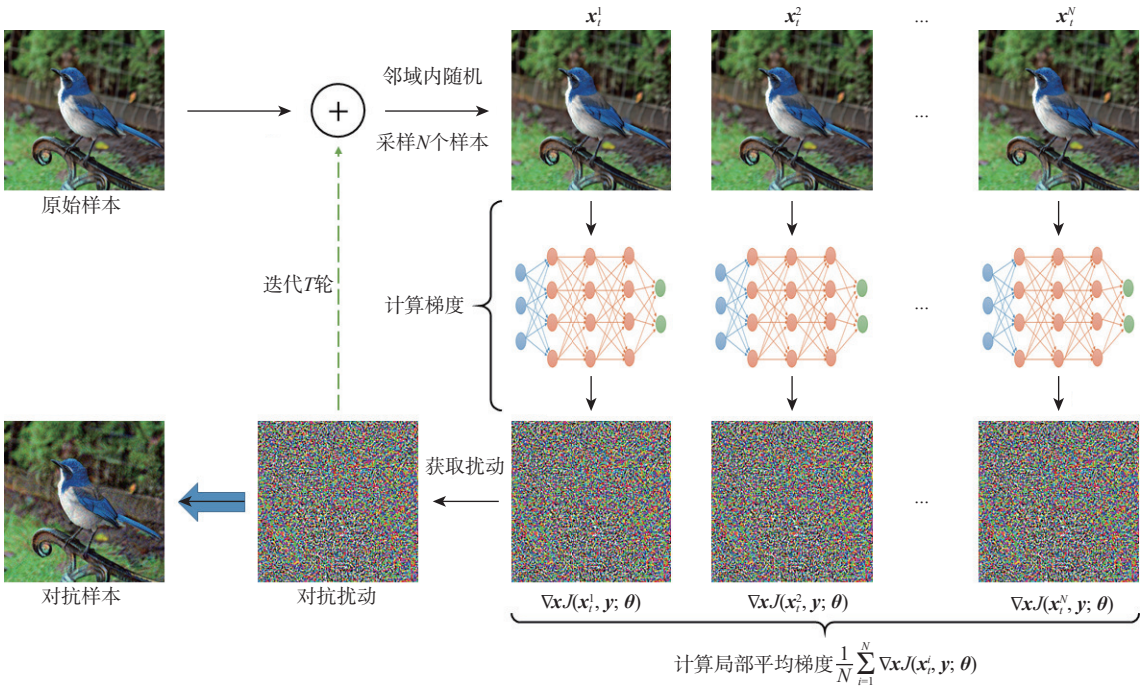


图2 本文方法框架图

Fig. 2 Frame diagram of the proposed method

综上所述，本文方法使用局部平均梯度来生成对抗样本，以此来抑制损失函数曲面存在的局部振荡现象，稳定更新方向，避免陷入局部极值点，进而提高黑盒攻击成功率和对抗样本的可迁移性。本文方法同时具有良好的可扩展性，可以与现有基于动量的方法（MI-FGSM[19]、NI-FGSM[20]）很好地集成。将本文方法引入MI-FGSM[19]中可得到对应的改进方法LS-MI-FGSM（loss-smoothed MI-FGSM），改进方法中的算法伪代码如下：

输入 神经网络f和损失函数J；原始输入图像$\boldsymbol{x}$和对应的真实标签$\boldsymbol{y}$；最大扰动量ε；迭代轮数T；衰减因子μ；邻域范围大小β；邻域内采样样本数量N

输出 对抗样本$\boldsymbol{x}^{\text{adv}}$

1) 计算迭代步长$\alpha=\varepsilon/T$

2) 计算邻域范围上界$\varepsilon'=\beta\varepsilon$

3) 初始化参数$\boldsymbol{x}_0^{\text{adv}}=\boldsymbol{x},\boldsymbol{g}_0=\boldsymbol{0}$

4) For $t=0$ to $T-1$ do

5) 在$\boldsymbol{x}_t^{\text{adv}}$的$\varepsilon'$邻域范围内随机采样$N$个样本

6) 根据式(10)计算$E_{\varepsilon'}(\boldsymbol{x})$

7) 根据式(11)更新$\boldsymbol{g}_{t+1}$

8) 根据式(12)更新$\boldsymbol{x}_{t+1}^{\text{adv}}$

9) 得到对抗样本$\boldsymbol{x}^{\text{adv}}=\boldsymbol{x}_T^{\text{adv}}$

类似地，扩展NI-FGSM[20]后也可以得到对应的改进方法LS-NI-FGSM（loss-smoothed NI-FGSM）。

2.2 与已有攻击方法的区别

目前已有的对抗样本攻击方法，如I-FGSM[11]、MI-FGSM[19]、NI-FGSM[20]等方法，都只使用了对抗样本优化路径上的数据点梯度值，而本文LS-M(N)I-FGSM为解决损失函数曲面存在局部振荡现象从而导致单个数据点梯度值不可靠的问题，充分利用这些数据点邻域内的梯度信息，对损失函数曲面作局部平滑处理，使用平滑处理后的梯度-局部平均梯度替代单点梯度来生成对抗样本。与已有方法相比，本文方法生成的对抗样本可迁移性更强，黑盒攻击成功率进一步提升。

图3总结了上述提到的几种攻击方法之间的转换关系。如果控制邻域范围上界的超参数β设置为0，本文LS-MI-FGSM和本文LS-NI-FGSM分别转换为MI-FGSM[19]和NI-FGSM[20]。如果动量衰减因子μ设置为0，MI-FGSM[19]和NI-FGSM[20]转换为I-FGSM[11]。如果迭代轮数T设置为1，I-FGSM[11]转

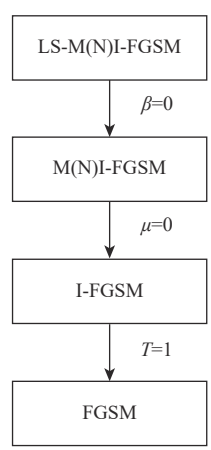


图3 不同对抗攻击方法关系

Fig. 3 Relationship between different adversarial attack methods

换为 FGSM[10]。

3 实　验

为了验证本文方法的有效性, 在 ImageNet 数据集上进行了大量的实验。在本节中, 首先详细说明实验设置, 然后分别进行单个模型攻击实验和集成模型攻击实验, 比较本文方法和现有方法的攻击成功率, 最后是关于超参数β和N的研究。

3.1 实验设置

数据集。从 ImageNet 验证集的 1 000 个类别中随机选择 1 000 张图像(即一个类别选取一张图像), 其均能被测试的模型正确分类。

网络模型。共使用 7 个神经网络模型, 包括 4 个经过正常训练的模型, 分别为 Inception-v3 (Inc-v3)[27]、Inception-v4(Inc-v4)[28]、Inception-Resnet-v2(IncRes-v2)[28]、Resnet-v2-101(Res-101)[29] 及 3 个经过对抗训练的模型[30], 分别为 Inc-v3$_{ens3}$、Inc-v3$_{ens4}$、IncRes-v2$_{ens}$。

基线方法。现有基于动量的 2 种攻击方法 MI-FGSM[19] 和 NI-FGSM[20]。

超参数。最大扰动量ε设置为 16, 迭代轮数T设置为 10, 步长α设置为 1.6, 对于 MI-FGSM[19] 和 NI-FGSM[20], 动量衰减因子μ设置为1.0, 对于本文方法 LS-MI-FGSM 和 LS-NI-FGSM, 设置$\beta = 3.0$, $N = 20$。

3.2 攻击单个模型

首先, 进行单个模型攻击实验, 分别使用 MI-FGSM[19]、NI-FGSM[20]、LS-MI-FGSM 和 LS-NI-FGSM 在 4 个正常模型(Inc-v3[27]、Inc-v4[28]、IncRes-v2[30]、Res-101[29])上生成对抗样本, 然后, 用生成的对抗样本去攻击所有模型, 攻击成功率(模型分类错误的图像数量占图像总数的比例), 如表 1 所示。根据表 1, 在黑盒设置下, 与 2 种基线方法 MI-FGSM[19] 和 NI-FGSM[20] 相比, 本文改进后的 2 种方法 (LS-MI-FGSM 和 LS-NI-FGSM) 的攻击成功率显著提高, 并且保持接近 100% 的白盒攻击成功率。例如, 如果使用 Inc-v3[27] 模型生成对抗样本, 表 1 中 4 个攻击方法都可以在白盒设置下达到接近 100% 的攻击成功率; 但是在黑盒设置下, 用基线方法 MI-FGSM[19] 生成的对抗样本攻击 Inc-v4[28] 模型的成功率仅为 45.1%, 攻击 Inc-v3$_{ens3}$[30] 模型的成功率仅为 13.5%, 而用对应的本文 LS-MI-FGSM 生成的对抗样本攻击这 2 个模型, 攻击成功率可以达到 85.8% 和 56.7%, 分别提升了 40.7% 和 43.2%, 综合比较, 对所有的黑盒攻击成功率进行平均, 本文改进点后的方法平均黑盒攻击成功率分别提升了 38.07% 和 27.77%。实验结果充分验证了本文方法的有效性, 显著提高了现有方法的黑盒攻击成功率和对抗样本的可迁移性。

表 1　单个模型攻击实验中不同方法的攻击成功率

Table 1　Success rates in single model setting by various adversarial attack methods　%

模型	攻击方法	Inc-v3[27]	Inc-v4[28]	IncRes-v2[28]	Res-101[29]	Inc-v3$_{ens3}$[30]	Inc-v3$_{ens4}$[30]	IncRes-v2$_{ens}$[30]
Inc-v3[27]	MI-FGSM[19]	100.0*	45.1	41.9	35.5	13.5	12.9	6.3
	LS-MI-FGSM	99.8*	85.8	84.7	77.3	56.7	54.2	36.1
	NI-FGSM[20]	100.0*	51.7	48.4	41.0	12.5	13.5	5.7
	LS-NI-FGSM	100.0*	79.2	77.5	69.8	38.9	39.1	23.8
Inc-v4[28]	MI-FGSM[19]	56.8	99.8*	46.3	41.4	16.9	14.5	7.8
	LS-MI-FGSM	88.2	98.1*	84.0	76.8	60.6	59.9	46.5
	NI-FGSM[20]	65.0	100.0*	52.4	46.1	16.4	13.1	7.1
	LS-NI-FGSM	86.3	99.8*	80.2	72.2	48.3	47.2	33.8
IncRes-v2[28]	MI-FGSM[19]	59.2	51.9	97.8*	45.4	22.3	16.6	12.3
	LS-MI-FGSM	84.5	82.8	94.8*	79.6	66.6	62.0	60.1
	NI-FGSM[20]	62.5	54.4	98.9*	46.4	20.7	15.9	10.1
	LS-NI-FGSM	85.4	83.5	99.1*	76.2	54.9	50.5	43.4
Res-101[29]	MI-FGSM[19]	59.1	51.0	49.2	99.2*	25.0	21.6	13.2
	LS-MI-FGSM	84.7	80.8	79.9	99.5*	68.7	64.6	54.3
	NI-FGSM[20]	64.7	57.9	56.5	99.4*	23.5	21.0	11.6
	LS-NI-FGSM	82.2	77.4	78.2	99.7*	59.6	53.6	43.4

注: *表示白盒攻击, 其余为黑盒攻击。

3.3 攻击集成模型

Liu 等[31] 研究发现攻击集成模型可以有效增强对抗攻击的可迁移性, 于是本文采用文献 [19] 中融合不同模型 logits 输出的集成攻击方法, 以证明本文方法可以进一步提高多模型环境下对抗攻击的可迁移性。

具体地, 分别使用 MI-FGSM[19]、NI-FGSM[20]、LS-MI-FGSM、LS-NI-FGSM 方法在由 4 个正常模型(Inc-v3[27]、Inc-v4[28]、IncRes-v2[28]、Res-101[29])构成的集成模型上生成对抗样本, 然后用生成的对抗

样本去攻击所有模型，攻击成功率如表2所示。

根据表2，在白盒设置下，4种方法都能达到接近100%的攻击成功率；但是在黑盒设置下，也就是在攻击3个对抗训练模型时，本文方法LS-MI-FGSM和LS-NI-FGSM相比于基线方法MI-FGSM[19]和NI-FGSM[20]的平均攻击成功率分别提高了32.50%和28.63%，充分验证了本文方法的有效性和优越性。

表2 集成模型攻击实验中不同方法的攻击成功率

Table 2 Success rates in multi-model setting by various adversarial attack methods %

攻击方法	Inc-v3[27]	Inc-v4[28]	IncRes-v2[28]	Res-101[29]	Inc-v3$_{ens3}$[30]	Inc-v3$_{ens4}$[30]	IncRes-v2$_{ens}$[30]
MI-FGSM[19]	100.0*	99.7*	99.4*	99.9*	48.2	42.8	28.5
LS-MI-FGSM	99.8*	99.4*	98.8*	99.7*	82.5	78.3	56.2
NI-FGSM[20]	100.0*	99.9*	99.8*	100.0*	45.5	41.0	25.4
LS-NI-FGSM	100.0*	100.0*	100.0*	99.9*	74.3	70.6	52.9

注：*表示白盒攻击，其余为黑盒攻击。

3.4 超参数研究

本节研究了提出的攻击方法中2个超参数β和N对攻击成功率的影响。

3.4.1 邻域范围上界β

研究决定邻域范围大小的超参数β对攻击成功率的影响(N固定为20)，实验结果如图4所示。

由图4可知，当$\beta=0$时，LS-MI-FGSM和LS-NI-FGSM分别转换为MI-FGSM[19]和NI-FGSM[20]，攻击成功率最低；当$\beta=0.5$时，虽然邻域范围很小，但本文方法有效提高了攻击成功率；继续增加β，可以观察到攻击成功率先逐渐升高到达峰值然后再逐渐降低，可见$\beta=3.0$时攻击性能最佳，因此本文在实验中设置$\beta=3.0$。

3.4.2 邻域内采样的样本数量N

研究决定邻域内采样的样本数量的超参数N对攻击成功率的影响(β固定为3.0)，实验结果如图5

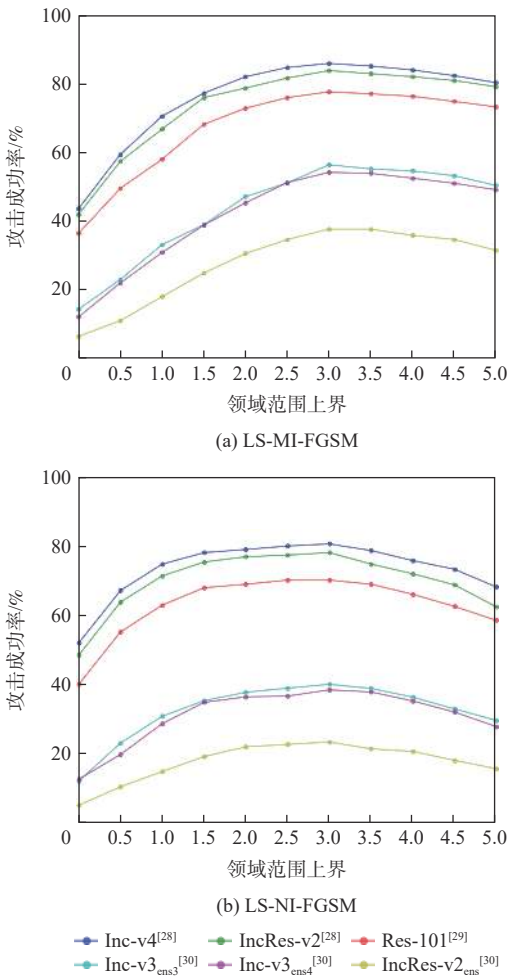


图4 邻域范围上界与攻击成功率关系

Fig. 4 Relationship between upper bound of neighborhood and attack success rate

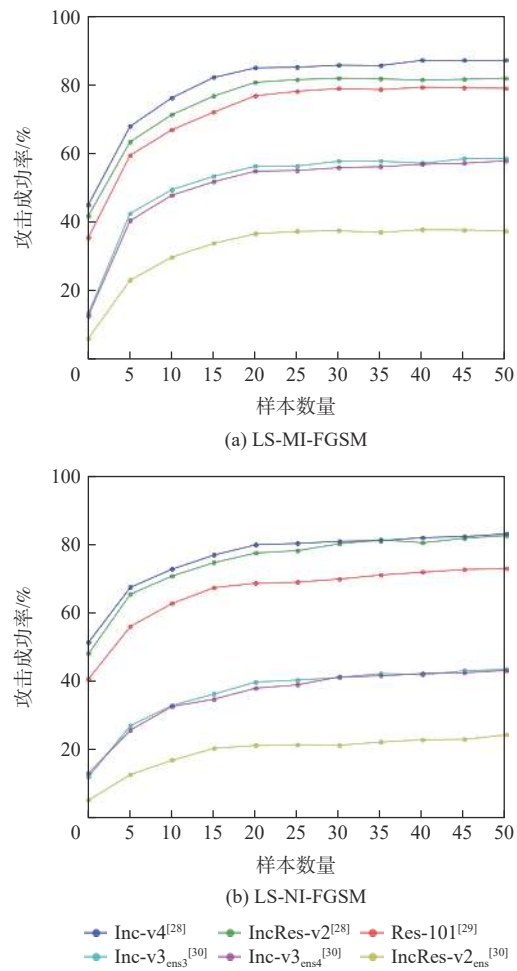


图5 邻域内采样的样本数量与攻击成功率关系

Fig. 5 Relationship between number of sampled in neighborhood and attack success rate

所示。

由图 5 可知，当 $N=0$时，LS-MI-FGSM 和 LS-NI-FGSM 分别转换为 MI-FGSM[19] 和 NI-FGSM[20]，攻击成功率最低，随着N的值从 0 逐渐增加到 20，攻击成功率快速提高；N的值超过 20 以后，攻击成功率趋于稳定。

注意到，N的值表示在邻域范围内采样的样本数量，所以在每一步迭代过程需要N次前向传播和反向传播来计算这N个样本的梯度，因此，更大的N值意味着更大的计算开销，为了平衡攻击成功率和计算开销，在实验中设置$N=20$。

综上所述，在本文的实验中，设置$\beta=3.0$和$N=20$。

4　结　论

1）本文方法可以与现有基于动量的方法 MI-FGSM 和 NI-FGSM 集成得到改进后的 2 种方法 LS-MI-FGSM 和 LS-NI-FGSM。

2）与 MI-FGSM 和 NI-FGSM 相比，LS-MI-FGSM 和 LS-NI-FGSM 在单个模型攻击实验中的平均黑盒攻击成功率分别提升了 38.07% 和 27.77%。

3）与 MI-FGSM 和 NI-FGSM 相比，LS-MI-FGSM 和 LS-NI-FGSM 在集成模型攻击实验中的平均黑盒攻击成功率分别提升了 32.50% 和 28.63%。

4）研究了超参数β和N对黑盒攻击成功率和对抗样本可迁移性的影响。

下一步研究方向是在保证对抗攻击性能的基础上，提高对抗样本的隐蔽性。

参考文献（References）

[1] CHEN L Y, LI S B, BAI Q, et al. Review of image classification algorithms based on convolutional neural networks[J]. Remote Sensing, 2021, 13(22): 4712-4730.

[2] ZAIDI S S A, ANSARI M S, ASLAM A, et al. A survey of modern deep learning based object detection models[J]. Digital Signal Processing, 2021, 126: 512-523.

[3] YUAN X H, SHI J F, GU L C. A review of deep learning methods for semantic segmentation of remote sensing imagery[J]. Expert Systems with Applications, 2021, 169: 114417.

[4] CHAKRABORTY A, ALAM M, DEY V, et al. A survey on adversarial attacks and defences[J]. CAAI Transactions on Intelligence Technology, 2021, 6(1): 25-45.

[5] GUO Y J, WEI X X, WANG G Q, et al. Meaningful adversarial stickers for face recognition in physical world[J]. IEEE Transaction on Pattern Analysis and Machine Intelligence, 2023, 45(3): 2711-2725.

[6] LIU A S, LIU X L, FAN J X, et al. Perceptual-sensitive GAN for generating adversarial patches[C]//Proceedings of the AAAI Conference on Artificial Intelligence. Washton, D.C.: AAAI, 2019, 33(1): 1028-1035.

[7] SHU M, LIU C, QIU W, et al. Identifying model weakness with adversarial examiner[C]//Proceedings of the 34th AAAI Conference on Artificial Intelligence. New York: AAAI, 2022: 11998-12006.

[8] WANG Y, MA X, BAILEY J, et al. On the convergence and robustness of adversarial training[C]//Proceedings of the 36rd International Conference on Machine Learning. New York: ICML, 2019: 11426-11438.

[9] DEMONTIS A, MELIS M, PINTOR M, et al. Why do adversarial attacks transfer? Explaining transferability of evasion and poisoning attacks[C]//Proceedings of the 28th USENIX Conference on Security Symposium. New York: ACM, 2018: 321-338.

[10] GOODFELLOW I J, SHLENS J, SZEGEDY C . Explaining and harnessing adversarial examples[C]//Proceedings of the 3rd International Conference on Learning Representations. Washington, D.C.: ICLR, 2015: 1-11.

[11] KURAKIN A, GOODFELLOW I, BENGIO S . Adversarial examples in the physical world[C]//Proceedings of the 5th International Conference on Learning Representations. Washington, D.C.: ICLR, 2017: 99-112.

[12] MADRY A, MAKELOV A, SCHMIDT L, et al. Towards deep learning models resistant to adversarial attacks[C]//Proceedings of Machine Learning Research. Baltimore: MD, 2017, 162: 52-65.

[13] BAI T, LUO J Q, ZHAO J, et al. Recent advances in adversarial training for adversarial robustness[C]//Proceedings of the 30th International Joint Conference on Artificial Intelligence. Freiburg: IJCAI, 2021: 4312-4321.

[14] CHEN J X, FENG X, JIANG L, et al. State of charge estimation of lithium-ion battery using denoising autoencoder and gated recurrent unit recurrent neural network[J]. Energy, 2021, 227(9): 1-8.

[15] LIU Y L, GAO Y, YIN W . An improved analysis of stochastic gradient descent with momentum[C]//Proceedings of the 34th Conference on Neural Information Processing Systems. Cambridge: NeurIPS, 2020: 18261-18271.

[16] QU G N, LI N . Accelerated distributed nesterov gradient descent[J]. IEEE Transactions on Automatic Control, 2020, 65(6): 2566-2581.

[17] HANG J, HAN K J, CHEN H, et al. Ensemble adversarial black-box attacks against deep learning systems[J]. Pattern Recognition, 2019, 101: 107184.

[18] YANG K, YAU J, FEI-FEI L, et al. A study of face obfuscation in imageNet[C]//Proceedings of the 39rd International Conference on Machine Learning. New York: ICML, 2022.

[19] DONG Y, LIAO F, PANG T, et al. Boosting adversarial attacks with mommtum[C]//Proceedings of the 2018 Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2018: 9185-9193.

[20] LIN J D, SONG C B, HE K, et al. Nesterov accelerated gradient and scale invariance for improving transferability of adversarial examples[C]//Proceedings of the 8th International Conference on Learning Representations. Washington, D.C.: ICLR, 2019: 1-12.

[21] STUTZ D, HEIN M, SCHIELE B. Disentangling adversarial robustness and generalization[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake: CVPR, 2019: 6976-6987.

[22] TRAMER F, KURAKIN A, PAPERNOT N, et al. Ensemble ad-

versarial training: Attacks and defenses[C]//Proceedings of the 6th International Conference on Learning Representations. Washington, D.C.: ICLR, 2018: 1-22.

[23] KURAKIN A, GOODFELLOW I, BENGIO S. Adversarial machine learning at scale[C]//Proceedings of the International Conference on Learning Representations. Washington, D.C.: ICLR, 2016: 1-17

[24] GUO C, RANA M, CISSE M, et al. Countering adversarial images using input transformations[C]//Proceedings of the International Conference on Learning Representations. Washington, D.C.: ICLR, 2018: 1-12.

[25] XIE C H, WANG J Y, ZHANG Z S, et al. Mitigating adversarial effects through randomization[C]//Proceedings of the 6th International Conference on Learning Representations. Washington, D.C.: ICLR, 2018: 1-16.

[26] LIAO F Z, LIANG M, DONG Y P, et al. Defense against adversarial attacks using high-Level representation guided denoiser[C]//Proceedings of the 31th IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake: CVPR, 2018: 1778-1787.

[27] SZEGEDY C, VANHOUCKE V, IOFFE S, et al. Rethinking the inception architecture for computer vision[C]//Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2016: 2818-2826.

[28] SZEGEDY C, IOFFE S, VANHOUCKE V, et al. Inception-v4, inception-resNet and the impact of residual connections on learning [J]. CoRR, 2016, 31(4): 4278-4284.

[29] 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. Salt Lake:CVPR, 2016: 770-778.

[30] WANG H J,WANG Y S. Self-ensemble adversarial training for improved robustness[C]//Proceedings of the International Conference on Learning Representations. Washington, D.C.: ICLR, 2021:1-18.

[31] LIU Y P, CHEN X Y, LIU C, et al. Delving into transferable adversarial examples and black-box attacks[C]//Proceedings of the 5th International Conference on Learning Representations. Washington, D.C.: ICLR, 2017: 1-24.

Adversarial attack method based on loss smoothing

LI Meihong[1, 2, *], JIN Shuang[1, 2], DU Ye[1, 2]

(1. Beijing Key Laboratory of Security and Privacy in Intelligent Transportation, Beijing Jiaotong University, Beijing 100044, China;
2. School of Computer and Information Technology, Beijing Jiaotong University, Beijing 100044, China)

Abstract: Deep neural networks (DNNs) are susceptible to attacks from adversairial samples. Most existing momentum-based adversarial attack methods achieve nearly 100% attack success rates under the white-box setting, but only achieve relatively low attack success rates under the black-box setting. An adversarial attack method based on loss smoothing is proposed, which can further improve the adversarial transferability. By integrating the locally averaged gradient term into the iterative process for attacks, our methods can suppress the local oscillation of the loss surface, stabilize the update direction and escape from poor local maxima. Empirical results on the standard ImageNet dataset demonstrate that the proposed method could significantly improve the adversarial transferability by 38.07% and 27.77% under single-model setting, and 32.50% and 28.63% under ensemble-model setting than the existing methods.

Keywords: deep neural network; adversarial example; black-box attack; loss-smoothing; artificial intelligence security

Received: 2022-06-11; **Accepted**: 2022-09-06; **Published Online**: 2022-09-15 11:12

URL: link.cnki.net/urlid/11.2625.V.20220914.1950.002

Foundation items: National Key R&D Program of China (2020YFB2103800,2020YFB2103802)

* **Corresponding author**. E-mail: mhli1@bjtu.edu.cn

http://bhxb.buaa.edu.cn jbuaa@buaa.edu.cn
DOI: 10.13700/j.bh.1001-5965.2022.0527

基于布局化-语义联合表征遥感图文检索方法

张若愚，聂婕*，宋宁，郑程予，魏志强
(中国海洋大学 信息科学与工程学部，青岛 266101)

摘　　要：遥感图文检索可以从类别繁多、内容复杂的遥感数据中检索到有价值的信息，对环境评估、城市规划以及灾害预测具有重要意义。但是，遥感图文跨模态检索存在一个关键问题，即遥感图像的空间布局信息被忽略。其主要体现在2个方面：①遥感目标的远距离建模困难；②遥感相邻次要目标被淹没。基于以上问题，提出了一种基于布局化-语义联合表征的跨模态遥感图像文本检索(SL-SJR)，主要包括主导语义监督的布局化视觉特征提取(DSSL)模块、布局化视觉-全局语义交叉指导(LV-GSCG)模块和多视角匹配(MVM)模块。DSSL模块实现主导语义类别特征监督下图像的布局化建模。LV-GSCG模块计算布局化视觉特征与文本中提取的全局语义特征的相似度来实现不同模态特征的交互。MVM模块建立跨模态特征指导的多视角度量匹配机制以消除跨模态数据之间的语义鸿沟。在4个基线遥感图像文本数据集上的实验验证，结果表明所提方法在大多数跨模态遥感图像文本检索任务中可以达到最先进的性能。

关　键　词：遥感图像；跨模态检索；空间布局信息；主导语义监督；类监督机制
中图分类号：P 407.8
文献标志码：A　　**文章编号**：1001-5965(2024)02-0671-13

“航空航天科技是20世纪以来发展最为迅速、对人类生产生活影响最大的科技领域之一。”近年来，随着航空航天事业的蓬勃发展，中国遥感卫星和传感器技术在不断进步，海量化遥感数据飞速膨胀(如NASA地球观测卫星每天返回约1 T图像数据)。高效、系统、合理地利用海量遥感数据对实现环境评估[1]、城市规划[2]以及灾难预测[3]等研究提供更可靠的数据支撑具有重要意义。作为遥感数据处理的重要手段之一，检索方法依靠深度学习、度量学习等手段从海量数据中挖掘有价值的信息，极大地提高了遥感数据的利用率。其中，图文跨模态检索能够探索图像和文本数据的潜在关联，实现2种模态数据的互补，引发了学者们的关注。

近年来，在遥感图像领域有少量的图像文本检索研究。为了细化遥感图像和文本之间的对应关系，Cheng等[4]设计了一种基于卷积神经网络(convolutional neural network, CNN)的深层语义对齐网络(deep semantic alignment network, DSAN)，通过更新视觉和文本的特征，来充分挖掘遥感图像和文本之间潜在的语义关系。虽然DSAN具有较好的检索精度，但是遥感图像中多尺度上下文信息被忽略，导致图像中关键信息丢失。因此，Yuan等[5]提出了一种新的非对称多模态特征匹配网络(asymmetric multimodal feature matching network, AMFMN)来提取具有多尺度特征和动态滤波功能的显著特征。葛芸等[6]提出一种基于多尺度池化和范数注意力机制的深度CNN，主要在通道层面与空间层面自适应地给显著特征加权关注了不同尺度的显著特征，结合了空间信息以提高了检索性能。

通过对遥感图像、文本数据以及跨模态检索框架深入研究与分析，遥感图文跨模态检索存在一个关键问题，即遥感图像的空间布局信息被忽略，其

收稿日期：2022-06-22；**录用日期**：2022-07-08；**网络出版时间**：2023-01-12 14:17
网络出版地址：link.cnki.net/urlid/11.2625.V.20230111.1851.005
基金项目：国家重点研发计划(2021YFF070400)；国家自然科学基金(62072418,62172376)；中央高校基本科研业务费专项资金(202042008)
*通信作者. E-mail：niejie@ouc.edu.cn

引用格式：张若愚，聂婕，宋宁，等. 基于布局化-语义联合表征遥感图文检索方法[J]. 北京航空航天大学学报，2024，50(2)：671-683.
ZHANG R Y，NIE J，SONG N，et al. Remote sensing image-text retrieval based on layout semantic joint representation[J]. Journal of Beijing University of Aeronautics and Astronautics，2024，50(2)：671-683 (in Chinese).

主要体现在2个方面：一方面，遥感目标的远距离建模困难。遥感目标分散在全局图像各处，全局目标的相关性建模能够挖掘不同遥感对象的相互依赖，增强遥感图像的宏观特征。在遥感图像特征提取方面，基于CNN的算法框架取得一系列成就，但是由于CNN自身具有局部感知特点，这会导致每个神经元每次只能感知到距离中心位置相当近的像素[7]，无法对遥感目标远距离建模。尽管基于注意力的方法通过计算相关性权重实现远距离特征交互，但该方法在相关性挖掘过程中将遥感目标分解与多个像素级特征表示，从而增加了不必要的噪声。另一方面，遥感相邻次要目标被淹没。在利用CNN进行特征建模时，相邻特征被卷积为一个高维特征，造成次要目标在训练过程中被主要目标淹没，失去了图像中的重要信息。在遥感图像中的小目标在卷积过程中生成为一个特征表示，降低了跨模态检索的准确性。

因此，如何通过视觉的空间布局化建模以实现遥感对象的远距离依赖和次要目标的保留是遥感图文检索方法是亟需解决的问题。针对遥感目标分散而导致远距离建模困难的问题，Yuan等[8]设计了一个多级信息动态融合模块，从不同层次提取遥感图像特征，并利用局部的信息补充全局信息，使用全局信息纠正局部信息。Rong等[9]为了克服遥感相邻次要目标被淹没的问题提出了特征全局感知模块和标签重建2个历史信息引导模块，特征全局感知模块能够使得当前模型在学习新的类别时更加关注与历史信息识别的旧类相关的部分，同时后者可以从背景中检索到属于学习到的类的特征。这些方法取得一系列成就，但是仍存在改进的空间。在图像分类领域中，视觉Transformer模型将输入图像分解为序列块，并利用Transformer模型构建序列块之间的联系，并取得了令人瞩目的成就。因此，受视觉Transformer模型启发，本文提出了基于空间布局-语义联合特征的遥感图文检索（spatial layout-semantic joint representation in remote sensing image-text retrieval, SL-SJR）方法。SL-SJR方法遵循了经典图文检索模式，首先将查询图像（文本）和数据库数据输入至检索模型生成图像特征（文本特征），再将生成的图像特征（文本特征）与数据库数据计算相似性得分，通过排序来达到检索目的。特别地，本文提出了主导语义监督下的布局化视觉特征提取（dominant semantic supervision layout visual feature extraction, DSSL）模块将遥感图像序列化为图像块来实现图像的布局化的建模。值得注意力的是，DSSL模块利用分类网络提取图片的主导语义类别特征，并借助Transformer模型生成布局化视觉特征并挖掘每个特征的显著信息。此外，本文提出布局化视觉-全局语义交叉指导（layout visual-global semantic cross guidance, LV-GSCG）模块，通过计算布局化视觉特征与文本中提取的全局语义特征的相似度，并利用掩码过滤不相关特征或弱相关联系来实现不同模态特征的交互，从而生成全局语义指导的视觉特征与布局化视觉指导的语义特征，达到跨模态的特征的相互补充。最后，本文构建了多视角匹配（multi-view matching, MVM）模块，除了经典的图像特征文本特征的匹配外，还将生成的全局语义指导的视觉特征、布局化视觉指导的语义特征与未经过指导的特征进行匹配，缓解了跨模态检索的语义鸿沟。

1 相关工作

本文主要从Transformer模型和跨模态图文检索2个方面对当前的研究现状进行相关总结。

1.1 Transformer

Transformer模型最早出现在自然语言处理领域，其通过使用多头自注意力机制和跨模态注意力机制，不仅可以自适应地探索上下文内容，也可以从不同角度提取特征。Zhang等[10]在Transformer的基础上将知识图谱融入到双向Transformer模型的编码器（bidirectional encoder representations from Transformer, BERT）中，利用信息化实体增强语言的表达。Dosovitskiy等[11]将图像转换为图像块序列并应用Transformer模型，在图像分类任务中表现出了极高的准确性。近年来，受Transformer模型在文本领域取得一系列成就的启发，该方法被逐渐应用于计算机视觉领域。Carion等[12]提出了一种基于Transformer模型的端到端目标检测方法。Srinivas等[13]在最后一个ResNet Bottleneck层[14]中利用多头自注意力（muti-head self attention, MHSA）而不是空间卷积来捕获更有效的全局上下文信息。Query2Label[15]利用Transformer解码器来查询类标签的存在性，这比多标签分类更简单有效。

随着自然语言处理和计算机视觉等领域的不断深入，Transformer模型被逐渐应用到跨模态检索领域中[16]。Messina等[16]提出了一个关系感知自注意力的体系结构，从而将视觉和文本2个模态的数据映射到同一公共空间，实现了高效的图像文本检索。为了提升检索的准确性，Valentin等[17]利用多模态的Transformer对视频中的不同模态进行联合编码来实现文本视频检索。Maleki等[18]利用Transformer方法构建一种双重注意力网络，实现了组织

病理学档案的跨模态检索。

本文同样使用了 Transformer 模型, 但与大部分的检索工作不同, 本文在检索网络中引入了类标签监督机制, 实现主导语义指导的布局化视觉特征提取。在类标签监督下, 抽取到的布局块可以更有效地挖掘遥感图像中显著语义信息。

1.2　跨模态图文检索

跨模态遥感图文检索方法在最开始时主要是指图像字幕技术, 这种技术一般使用深度学习的网络自动获取到每幅图片的字幕, 然后在检索过程中计算生成字幕和标记文本之间的相似度并对结果进行排序。Shi 等[19] 从多级图像理解和语言生成这 2 个阶段介绍了 RS 图像检索框架。Lu 等[20] 制作了用于生成遥感图像标题的 RSICD 数据集。Hoxha 等[21] 利用波束搜索为目标图像生成多个字幕, 然后根据词汇相似性选择最佳字幕。为了解决在 RSICD 数据集中因交叉熵损失函数造成的过拟合问题, Li 等[22] 提出了截断交叉熵损失并取得较好的结果。Lu 等[23] 将声音作为输入并通过一个注意力框架生成特定的句子。尽管基于句子的遥感图像检索取得了一系列成就, 但该过程仍然存在一些难以克服的缺点。比如在机器学习中生成的句子较为粗糙, 这严重阻碍了在其他领域的应用。

近年来, 随着深度学习和标注技术的不断进步, 成熟的图像文本跨模式检索方法逐渐形成。在输入一张查询图片后, 跨模态图文检索网络根据彼此之间的相似性在数据库中找到对应的文本, 然后再将该文本作为最终搜索的结果。Faghri 等[24] 使用了一个三元组损失函数来最小化相似图像和文本之间的距离。Lee 等[25] 将图像中的区域与标题中的单词对齐来计算其相似度。Wang 等[26] 设计了一个相似度函数来精确测量输入图像和句子之间的距离, 并用一个排名损失函数来训练这个网络。Devlin 等[27] 提出了一种基于深度学习的图像文本检索模型, 该模型使用 CNN 来提取查询图像的视觉特征, 并在视觉空间中检索相似的图像。之后, 利用 RNN 对候选文本描述进行编码, 并根据文本的语义特征重新排序, 从而得到最终的检索结果。2019 年, 一个利用深层神经网络解决遥感图像文本检索问题的深层双向三重网络(deep bidirectional triple network, DBTN)[28] 被提出, 并发布了一个新的 TextRS 数据集。为了优化遥感图像和文本之间的对应关系, DSAN[5] 利用语义对齐模块挖掘遥感图像和文本之间的潜在语义关系。然而, DSAN 忽略了对遥感图像非常重要的多尺度背景信息。因此, 提出了一种新的非对称多模态特征匹配网络[5], 用于捕获具有多尺度特征的显著特征, 并结合关键词属性, 以进一步提高检索精度。

虽然上述检索网络均取得了突破性的进展, 但都未能高效解决遥感目标远距离建模困难以及次要目标难以保留的问题。本文通过视觉的空间布局化建模, 构建主导语义监督的布局化视觉特征提取 (dominant semantic supervision layout visual feature extraction, DSSLVFE)、LV-GSCG 和 MVM 模块, 最终实现了遥感对象的远距离依赖和次要目标的保留。

2　SL-SJR 方法介绍

本文提出了一种基于空间布局化-语义联合表征的跨模态遥感图文检索方法, 主要包括 DSSLVDE、LV-GSCG、MVM 3 个模块, 整体模型如图 1 所示, 图中: $\boldsymbol{v}^{\mathrm{cnn}}$为通过 CNN 提取到的遥感图像特征; $\boldsymbol{e}_t$为文本单词嵌入特征; $\boldsymbol{T}$为一个句子中每个单词特征向量组成的矩阵; $\boldsymbol{F}_t'$为一个句子的特征; $\boldsymbol{Z}_l$为经过 Transformer 编码器中的多层感知机得

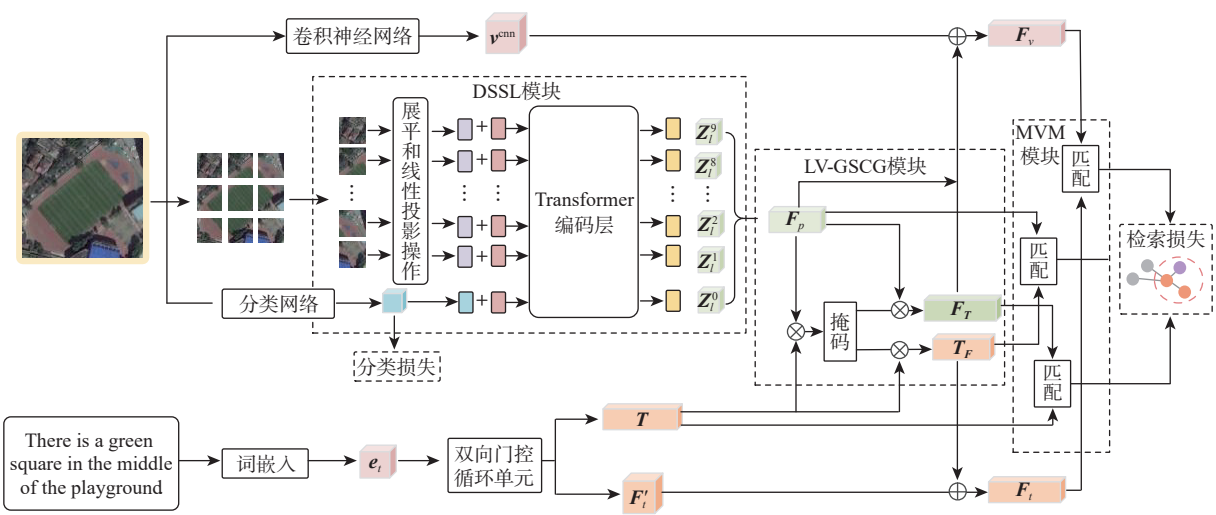


图 1　模型整体架构图

Fig. 1　Overall structure diagram of the model

到的特征向量；$\boldsymbol{F}_p$为N个布局块特征向量和类别嵌入向量组成的矩阵；$\boldsymbol{F}_T$为遥感文本指导的全局图像特征；$\boldsymbol{T}_F$为遥感图像指导的全局文本特征；$\boldsymbol{F}_v$为全局图像特征；$\boldsymbol{F}_t$为全局文本特征，L为Transformer编码层的层数。在主导语义类别特征的监督下，DSSL模块将图像序列为图像块实现布局化建模，同时利用Transformer挖掘每个布局化视觉特征的显著信息。LV-GSCG模块计算布局化视觉特征与文本中提取的全局语义特征的相似度来实现不同模态特征的交互，达到跨模态的特征的相互补充。MVM模块建立跨模态特征指导的多视角度量匹配机制来精确地测量输入遥感图像和文本之间的距离，以消除跨模态数据之间的语义鸿沟。本文遵循经典的跨模态检索的方法，分为图像特征提取、文本特征提取以及相似度匹配3个部分。特别地，本文设计DSSLVFE、LV-GSCG、MVM 3个模块。因此，本文将分5个部分来进行介绍，分别是特征表示、DSSLVFE模块、LV-GSCG模块、MVM模块、损失函数。

2.1 特征表示

2.1.1 图像特征表示

对于输入的遥感图像$I\in\mathbf{R}^{H\times W\times C}$，其中$(H, W)$为输入遥感图像的分辨率，$C$为输入图像的通道数。本文利用在ImageNet数据集上的经过预训练的CNN中ResNet18网络作为主干网络，并将预训练好的参数传输到任务中。

$$\boldsymbol{v}^{\text{cnn}}=f_{\text{CNN}}(I,\theta) \tag{1}$$

式中：θ为f_{CNN}中的一系列参数。

2.1.2 文本特征表示

为了更好地实现图像特征和文本特征的交互，本文使用双向门控循环单元(Bi-GRU)分别从向前、向后这2个方向提取句子和单词级别的特征，提取句子的上下文相关信息。

对于一个句子S，假定其由m个单词组成$\{\boldsymbol{w}_t\}_{t=1}^{m}$，每一个单词本文可以通过一个嵌入矩阵$\boldsymbol{W}_e$得到嵌入向量$\boldsymbol{e}_t$，即$\boldsymbol{e}_t=\boldsymbol{W}_e(\boldsymbol{w}_t)$，其中$t\in[1,m]$。随后，本文采用Bi-GRU从向前、向后2个方向来总结句子的上下文信息，具体过程如式(2)和式(3)所示。第1个正向GRU依次读入单词$\boldsymbol{w}_1$-$\boldsymbol{w}_m$，第2个后向GRU依次读入单词$\boldsymbol{w}_m$-$\boldsymbol{w}_1$。$\overrightarrow{\boldsymbol{h}}_t$代表隐藏状态是由前向GRU生成，$\overleftarrow{\boldsymbol{h}}_t$代表隐藏状态是由反向GRU生成。

$$\overrightarrow{\boldsymbol{h}}_t=\overrightarrow{f_{\text{GRU}}}(\boldsymbol{e}_t,\overrightarrow{\boldsymbol{h}}_{t-1}) \tag{2}$$

$$\overleftarrow{\boldsymbol{h}}_t=\overleftarrow{f_{\text{GRU}}}(\boldsymbol{e}_t,\overleftarrow{\boldsymbol{h}}_{t-1}) \tag{3}$$

本文对Bi-GRU的这2种隐藏状态进行平均，以获得每一步的单词特征$\boldsymbol{s}_t\in\mathbf{R}^d$，其中$d$表示维度，具体如式(4)～式(6)所示：

$$\boldsymbol{s}_t=(\overrightarrow{\boldsymbol{h}}_t+\overleftarrow{\boldsymbol{h}}_t)/2 \tag{4}$$

$$\boldsymbol{T}=(\boldsymbol{s}_1,\boldsymbol{s}_2,\cdots,\boldsymbol{s}_m) \tag{5}$$

$$\boldsymbol{F}_t'=\frac{1}{m}\sum_{t=1}^{m}\boldsymbol{s}_t \tag{6}$$

式中：$\boldsymbol{T}\in\mathbf{R}^{m\times d}$；$\boldsymbol{F}_t'\in\mathbf{R}^d$。

2.2 DSSLVFE模块

为了更好构建空间布局化的表示并抽取布局块的主导语义，本文构建了DSSLVFE模块，如图2所示。主导语义的关键在于在训练过程中引入类标签的监督，从而使抽取的布局块更有效地挖掘显著特征。这一模块通过实现布局化建模来实现遥感对象的远距离依赖和次要目标的保留，又能利用类监督机制实现对每个布局块的显著特征挖掘。

具体来说，将输入的遥感图像$I\in\mathbf{R}^{H\times W\times C}$切成许多大小相同的布局块，将其展平后得到一系列的展平后的二维块$\boldsymbol{X}_P\in\mathbf{R}^{N\times(P^2C)}$，其中$(P, P)$代表输入图像经过切割后每个布局块的分辨率，$N=(H\times W)/(P\times P)$代表布局块代表布局块的个数并决定了输入到Transformer的序列长度。由于Transformer在其每一层输入均接收向量的大小为d的恒定向量，所以本文将遥感图像分块处理后将每个布局块展平成一维向量并使用线性投影变换层映射成d维大小向量。DSSLVFE追加了一个类监督机制构建显著类别特征，并与布局块向量一起输入到Transformer编码层中进行图像的显著特征挖掘：

$$\boldsymbol{Z}_0=(\boldsymbol{X}_{\text{class}},\boldsymbol{X}_P^1\boldsymbol{E},\boldsymbol{X}_P^2\boldsymbol{E},\cdots,\boldsymbol{X}_P^N\boldsymbol{E})+\boldsymbol{E}_{\text{pos}} \tag{7}$$

式中：$\boldsymbol{X}_{\text{class}}\in\mathbf{R}^{(P^2C)}$表示经过类监督分类网络得到的类别向量；$\boldsymbol{X}_P^N\in\mathbf{R}^{(P^2C)}$表示切块展平后第$N$个布局块的特征向量；$\boldsymbol{E}\in\mathbf{R}^{(P^2C)\times d}$代表线性投影变换层；$\boldsymbol{E}_{\text{pos}}\in\mathbf{R}^{(N+1)\times d}$表示位置编码向量；$\boldsymbol{Z}_0\in\mathbf{R}^{(N+1)\times d}$作为下一步Transformer编码层的输入。

在完成视觉布局构建后，本文运用Transformer编码层进行后续操作，通常情况下Transformer编码层包含$l\in[1,L]$个Transformer编码层，每个Transformer编码层由多头自注意力(MSA)、多层感知机(MLP)、归一化层(LN)组成，具体过程如下：

$$\boldsymbol{Z}_l'=f_{\text{MSA}}(f_{\text{LN}}(\boldsymbol{Z}_{l-1}))+\boldsymbol{Z}_{l-1} \tag{8}$$

$$\boldsymbol{Z}_l=f_{\text{MLP}}(f_{\text{LN}}(\boldsymbol{Z}_l'))+\boldsymbol{Z}_l' \tag{9}$$

$$\boldsymbol{F}_p=(\boldsymbol{Z}_L^0,\boldsymbol{Z}_L^1,\cdots,\boldsymbol{Z}_L^N) \tag{10}$$

式中：LN表示Transformer编码层中的归一化操作；$\boldsymbol{Z}_l'$表示经过第l个Transformer编码层中多头自注意

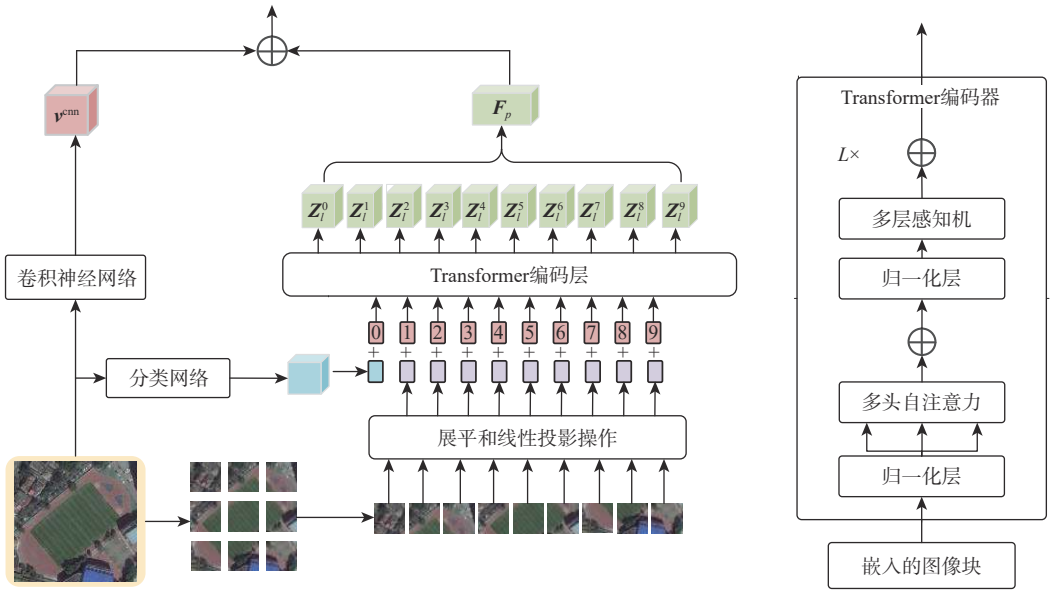


图 2　主导语义监督下的布局化视觉特征提取模块

Fig. 2　Dominant semantic supervision under layout visual feature extraction module

力 (MSA) 后得到的特征向量; $\boldsymbol{Z}_l$表示经过第 l 个 Transformer 编码层中的 MLP 得到的特征向量; $\boldsymbol{F}_p \in \mathbf{R}^{(N+1)\times d}$表示 N 个布局块特征向量和类别嵌入向量组成的矩阵。

本文将 CNN 提取的遥感图像特征向量$\boldsymbol{v}^{\text{cnn}} \in \mathbf{R}^d$和利用类监督机制得到的显著遥感图像的特征向量$\boldsymbol{F}_p \in \mathbf{R}^{(N+1)\times d}$这 2 部分进行融合得到遥感图像特征向量$\boldsymbol{F}_m \in \mathbf{R}^{(N+1)\times d}$:

$$\boldsymbol{F}_m = \boldsymbol{v}^{\text{cnn}} \oplus \boldsymbol{F}_p \tag{11}$$

式中: $\oplus$表示向量加法; 在$\boldsymbol{F}_p \in \mathbf{R}^{(N+1)\times d}$中的 $(N+1)$ 个 d 维向量中分别加入$\boldsymbol{v}^{\text{cnn}} \in \mathbf{R}^d$。

2.3　LV-GSCG 模块

为了进一步挖掘遥感图像全局语义信息, 本文构建了 LV-GSCG 模块, 如图 3 所示, 该模块能够获取遥感图像指导的全局文本特征和遥感文本指导的全局图像特征, 实现了布局化视觉信息和全局语义信息的交互。

LV-GSCC 模块利用掩码机制设计了一种全新的针对图像块和词向量深层次交互的方法来充分挖掘遥感图像和文本 2 种模态之间更细微的语义相关性, 对于遥感图像中的每一个图像块都能与文本中的每一个单词进行交互, 这样就能获得遥感图像指导的全局文本特征和遥感文本指导的全局图像特征。

在 2.1 和 2.2 节中已经得到所有布局块特征向量组成的全布局块特征向量$\boldsymbol{F}_p \in \mathbf{R}^{(N+1)\times d}$ 和所有单词特征向量组成的全单词特征向量$\boldsymbol{T} \in \mathbf{R}^{m\times d}$, 为了得到布局块和单词的关联性, 本文将$\boldsymbol{F}_p$与$\boldsymbol{T}$的转置矩阵$\boldsymbol{T}^{\text{T}}$叉乘得到每个布局块对于每个单词的原始相关矩阵$\boldsymbol{S}_{\text{origin}}$, 然后将获得的布局块和单词的相关性矩阵采用掩码机制得到掩码相关性矩阵$\boldsymbol{S}_{\text{mask}}$, 以达到屏蔽和突出特定元素的目的, 在本文, 当布局块和单词的相关性矩阵中元素值大于 0.5 可以理解为关联性强。本文将其元素值设置为 1, 反之当布局块和单词的相关性矩阵中元素值小于 0.5 可以理解为关联性较弱本文将其元素值设置 0。过程如下:

$$\boldsymbol{S}_{\text{origin}} = \boldsymbol{F}_p \otimes \boldsymbol{T}^{\text{T}} \tag{12}$$

$$\boldsymbol{S}_{\text{mask}} = \text{mask}(\boldsymbol{S}_{\text{origin}}) \tag{13}$$

式中: $\boldsymbol{S}_{\text{origin}} \in \mathbf{R}^{(N+1)\times m}$; $\boldsymbol{S}_{\text{mask}} \in \mathbf{R}^{(N+1)\times m}$; $\otimes$表示矩阵乘法。

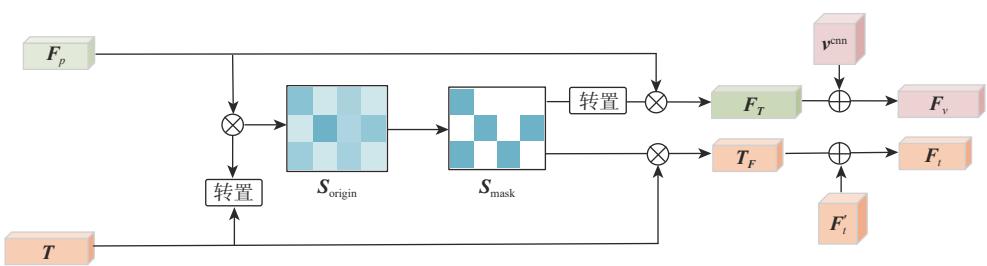


图 3　布局化视觉-全局语义交叉指导模块

Fig. 3　Layout vision global semantic cross guidance module

本文通过上一步得到的掩码相关性矩阵$\boldsymbol{S}_{\text{mask}}$来分别获取遥感文本指导的全局图像特征$\boldsymbol{F}_{\boldsymbol{T}}$和遥感图像指导的全局文本特征$\boldsymbol{T}_{\boldsymbol{F}}$以实现布局化视觉信息和全局语义信息的交互,过程如下:

$$\boldsymbol{F}_{\boldsymbol{T}}=(\boldsymbol{S}_{\text{mask}})^{\text{T}}\otimes\boldsymbol{F}_{p} \tag{14}$$

$$\boldsymbol{T}_{\boldsymbol{F}}=\boldsymbol{S}_{\text{mask}}\otimes\boldsymbol{T} \tag{15}$$

式中:$\boldsymbol{F}_{\boldsymbol{T}}\in\mathbf{R}^{m\times d}$;$\boldsymbol{T}_{\boldsymbol{F}}\in\mathbf{R}^{(N+1)\times d}$。

经过这一过程后,全局文本特征$\boldsymbol{T}_{\boldsymbol{F}}$的维度与全布局块特征向量$\boldsymbol{F}_{p}$维度相同,即$\boldsymbol{T}_{\boldsymbol{F}}$的每一维都是受特定布局块监督的文本特征向量。同理,全局图像特征$\boldsymbol{F}_{\boldsymbol{T}}$的维度与所有单词特征向量组成的全单词特征向量$\boldsymbol{T}^{\text{T}}$维度相同,即$\boldsymbol{F}_{\boldsymbol{T}}$的每一维都是受特定词向量监督的图像特征向量。本文认为$\boldsymbol{T}_{\boldsymbol{F}}$是极其有效的全局文本特征,其每一维文本特征向量与监督它的特定布局块具有极强相关性,且通过掩码相关性矩阵已过滤掉部分无效的词向量。$\boldsymbol{F}_{\boldsymbol{T}}$是极其有效的全局图像特征,其每一维图像特征向量与监督它的特定词向量具有极强相关性,且通过掩码相关性矩阵已过滤掉部分无效的布局块向量。

本文将$\boldsymbol{F}_{\boldsymbol{T}}$和$\boldsymbol{T}_{\boldsymbol{F}}$分别与之前得到的图像特征$\boldsymbol{F}_{m}$和文本特征$\boldsymbol{F}_{t}'$进行融合,得到全局图像特征$\boldsymbol{F}_{v}$和全局文本特征$\boldsymbol{F}_{t}$:

$$\boldsymbol{F}_{v}=\boldsymbol{v}^{\text{cnn}}\oplus\boldsymbol{F}_{\boldsymbol{T}} \tag{16}$$

$$\boldsymbol{F}_{t}=\boldsymbol{F}_{t}'\oplus\boldsymbol{T}_{\boldsymbol{F}} \tag{17}$$

式中:$\boldsymbol{F}_{v}\in\mathbf{R}^{m\times d}$;$\boldsymbol{F}_{t}\in\mathbf{R}^{(N+1)\times d}$。

2.4 MVM模块

为了消除跨模态数据之间的语义鸿沟,本文设计了一个多视角匹配模块。该模块设计了一种跨模态布局特征与语义特征交互的相似度度量方法来精确地测量输入遥感图像和文本之间的距离。多视角相似度度量方法主要分为3个部分:①相似度度量S_1为将遥感图像特征$\boldsymbol{F}_{v}$和句子特征$\boldsymbol{F}_{t}$进行相似度计算;②相似度度量S_2为将遥感图像指导的针对各布局块的全局文本特征$\boldsymbol{T}_{\boldsymbol{F}}$与布局块特征向量$\boldsymbol{F}_{p}$进行相似度计算并求均值;③相似度度量$S_3$为将遥感文本指导的针对各单词向量的全局图像特征$\boldsymbol{F}_{\boldsymbol{T}}$与全单词特征向量$\boldsymbol{T}$进行相似度计算并求均值,最终将3部分加和得到最终的相似度度量S,具体过程如下:

$$S_1=\phi(\boldsymbol{F}_{v},\boldsymbol{F}_{t}) \tag{18}$$

$$S_2=\phi(\boldsymbol{F}_{p},\boldsymbol{T}_{\boldsymbol{F}})/s_{\text{patch}} \tag{19}$$

$$S_3=\phi(\boldsymbol{F}_{\boldsymbol{T}},\boldsymbol{T})/s_{\text{word}} \tag{20}$$

$$S=S_1+\alpha_1S_2+\alpha_2S_3 \tag{21}$$

式中:ϕ为计算余弦相似度;s_{patch}为布局块数量;s_{word}为词向量数量;α_1和α_2为超参数。

2.5 损失函数

在跨模态特征匹配领域的主流损失函数是三元组损失,三元组损失函数的目的是增加样本与其对应的负样本之间的距离,同时使样本与其正样本的距离尽可能接近,在跨模态匹配方面的损失函数为

$$\begin{aligned}L_1=&[\alpha-S(I_1,T_1)+S(I_1,\widehat{T_1})]_{\widehat{T}}+\\&[\alpha-S(I_1,T_1)+S(\widehat{I_1},T_1)]_{\widehat{I}}\end{aligned} \tag{22}$$

式中:α为边界参数;I_1为输入的图像特征;T_1为输入的文本特征;$\widehat{I_1}$为负样本对应的图像特征;$\widehat{T_1}$表示负样本对应的文本特征。式(22)的第1部分是基于图像寻找文本的情况,$S(I_1,T_1)$表示图像和与之对应的正样本的相似度,$S(I_1,\widehat{T_1})$表示图像和与负样本的相似度;式(22)的第2部分是基于文本寻找图像的情况,$S(I_1,T_1)$表示文本和与之对应正样本之间的相似度,$S(\widehat{I_1},T_1)$表示文本和与负样本的相似度。

除此之外,为了在DSSL模块更好地学习遥感图像的类别以获取图像中主导信息,本文采用交叉熵损失函数来训练这个分类过程,如式(23)所示:

$$L_2=(1/N)\sum_{i=1}^{N}(-Y_i\lg(\widehat{Y}_i)-(1-Y_i)\lg(1-\widehat{Y}_i)) \tag{23}$$

式中:Y为视觉Transformer中类嵌入得到的类别预测结果;$\widehat{Y}$为对应的类别标签。

本文利用一个加权方案将2个损失函数融合作为本模型的损失函数如式(24)所示:

$$L=\beta L_1+(1-\beta)L_2 \tag{24}$$

式中:β为超参数。

3 实验结果与分析

为了验证本文方法的有效性,本文在搭载图形处理器(GPU)的服务器上,并在4个公开的遥感图像文本数据集来进行实验验证来评估模型性能。

3.1 数据集介绍

1)Sydney数据集。Sydney数据集包含7个场景的555张遥感图像,其中场景包括住宅、草原、河流、海洋、工业区和跑道,在这个数据集中的每张图像由5个不同的文本来进行描述。

2)UCM数据集。UCM数据集包含21个场景的2 100张遥感图像,每张遥感图像大小为256×256像素,地面采样距离是0.300 48 m。UCM数据集来自美国地质调查局国家地图区域图像,在这个数据集中的每张图像由5种不同的文本描述。

3) RSICD 数据集。RSICD 数据集是一个大规模的遥感图像文本数据集, 每张遥感图像大小为 224×224 像素, RSICD 包含 30 个场景的 10 921 张图像, 数据集中的大部分图像来自谷歌地球和百度地图, 数据集中的每张图像由 5 种不同的文本描述。

4) RSITMD 数据集。RSITMD 数据集是一个细粒度的数据集, 是从 RSICD 数据集和谷歌地球数据集中精选出来的 4 743 张遥感图像, 与上面的 3 个数据集相比, RSITMD 数据集提供了更细粒度和更少重复的描述, 并且每张图像依然用 5 个句子来进行描述。RSITMD 数据集为每幅图像分配 1～5 个关键词来对目标对象进行补充。

5) 为了满足损失函数 L_2 中类别标签监督的需求, 本文需要得到 4 类数据集的类别标签。其中 RSITMD 数据集包含完整类别标签, RSICD 数据集包含部分类别标签, 因此, 对于 UCM、Sydney 2 类数据集和 RSICD 数据集的部分数据, 本文参考 RSITMD 的类别, 对其进行人工标注类别标签。由于 RSITMD 数据集中的图像文件名自带标签, 举例来说通常文件名是由标签和序号组成, 本文直接取依次遍历图像文件获取标签和序号存到字典中。类别标签包括{棒球场, 船只, 工业区, 交叉路口, 飞机, 游乐场, 池塘, 体育场, 机场, 荒地, 海滩, 桥梁, 中心, 教堂, 商业区, 密集住宅, 沙漠, 农田, 森林, 草地, 中型住宅, 山, 公园, 停车场, 港口, 火车站, 度假村, 河流, 学校, 稀疏住宅, 广场, 储罐, 高架桥}共 33 类。最终, 本文得到 4 类数据集的类别标签, 满足损失函数 L_2 的需求。

3.2　基线介绍

1) SCAN 模型[29]。SCAN 模型是基于 VSE 的一种图文检索方法, 使用 Faster RCNN 提取图像特征, 并尝试将图像中的目标与文本中的目标对齐。

2) CAMP 模型[30]。CAMP 模型提出了一种自适应消息传递方法, 该方法自适应地控制跨模态信息传输的流程, 并使用融合特征计算图像和文本的匹配分数。

3) MTFN 模型[26]。MTFN 模型基于秩分解的思想设计了一个多模态融合网络来计算嵌入特征的距离。

4) AMFMN 模型[5]。AMFMN 模型中设计了一个相似度函数来精确测量输入图像和文本之间的距离, 并设计了一个动态可变边界的损失函数来解决了样本对匹配边界问题。

3.3　实验细节

本文中的所有实验都是基于 PyTorch 深度学习框架开展的, 对于初始遥感图像本文首先将其缩放成 278×278 像素, 然后随机裁剪和旋转图像以增强训练样本, 输入图像的最终大小为 256×256 像素。

为了保证实验的一致性, 在保证模型不受深度网络过度拟合影响的同时, 本文使用预训练 ResNet-18 和视觉 Transformer 模型作为视觉特征提取工具, 视觉嵌入维度大小设置为 512, 多头个数设置为 64 个, 图像块的大小设置为 32×32。单词嵌入维度设置为 300, 双向 GRU 的隐藏层设置为 512。本文使用 Adam 优化器对网络进行迭代次数(epoch)为 150 的训练, 最小批量大小(batch_size)设置为 128。在训练过程中, 将初始学习率设置为 0.000 2, 每 15 次迭代进行一次学习率衰减, 衰减因子为 0.1, 多视角匹配模块中相似度度量计算公式中超参数 α_1 与 α_2 初始设置为 0.1 和 0.1, 损失函数中超参数的 β 初始设置为 0.5。

3.4　评估指标

为了充分说明本文方法的有效性, 采用召回率 R_K 和平均召回率 R_{m}(R_{m} 表示 R_K 的平均值)作为评价指标来对模型进行评估, R_K 表示检索结果取 K 个实例时所获得的召回率, 这 2 个评价指标的值越高表明模型的性能越好, 其过程如式 (25) 和式 (26) 所示。其中 N 表示实例的个数, R_{LK}=0 表示第 K 个返回结果与查询实例无关, R_{LK}=1 表示第 K 个返回结果与查询实例相关。

$$R_K = (1/N)\sum_{K=1}^{N} R_{LK} \tag{25}$$

$$R_{\mathrm{m}} = (1/N)\sum_{K=1}^{N} R_K \tag{26}$$

3.5　对比实验和分析

为了评估本文方法的有效性, 将 SL-SJR 方法与现有的遥感图像文本检索方法(如 SCAN t2i、SCANi2t、CAMP-triplet、CAMP-bce、MTFN、AMFMN)进行比较, 如表 1～表 4 所示(表中加粗数值表示最优结果)。在 4 个数据集(RSICD、RSITMD、UCM、Sydney), 图像检索文本以及文本检索图像 2 个检索任务分别测试了 R_1、R_5、R_{10}、R_{m} 这 4 个性能指标。

将本文方法分别与 4 种图像文本检索的基线方法 (SCAN t2i、SCAN i2t、CAMP-triplet、CAMP-bce) 在 4 种遥感数据集上进行比较, 从表 1 可以看出, 与 4 种基线方法相比, 本文方法在遥感图像检索文本和遥感文本检索图像 2 个任务上的 R_1、R_5、R_{10} 这些指标均有明显的提升。与当前存在最先进的遥感图像检索方法 AMFMN 在 RSITMD 数据集上的实验结果表明, 在遥感图像检索文本任务中的 R_1 提升了 27.12%, R_{10} 提升了 1.67%; 在遥感文本检

表1 RSICD数据集对比实验结果

Table 1 Comparative experimental results on RSICD

方法	R_1		R_5		R_{10}		R_m
	图像检索文本	文本检索图像	图像检索文本	文本检索图像	图像检索文本	文本检索图像	
SCAN t2i	4.42	4.02	11.20	11.54	17.68	18.60	11.24
SCAN i2t	**5.90**	3.86	13.21	16.83	19.96	26.49	14.38
CAMP-triplet	5.22	4.30	13.05	17.10	21.02	27.54	14.71
CAMP-bce	4.50	3.03	10.08	15.12	16.48	23.05	12.04
MTFN	5.01	4.96	12.86	12.24	21.30	29.12	14.24
AMFMN	5.39	5.05	15.32	18.19	28.82	29.70	17.07
本文	5.31	**5.32**	**19.12**	**19.69**	**31.47**	**32.50**	**18.90**

表2 RSITMD数据集对比实验结果

Table 2 Comparative experimental results on RSITMD

方法	R_1		R_5		R_{10}		R_m
	图像检索文本	文本检索图像	图像检索文本	文本检索图像	图像检索文本	文本检索图像	
SCAN t2i	10.59	10.04	28.72	29.54	38.41	42.91	26.70
SCAN i2t	10.84	9.62	25.86	29.80	37.26	41.06	25.74
CAMP-triplet	11.82	8.69	27.40	27.17	38.12	43.60	26.13
CAMP-bce	9.21	6.81	22.56	25.65	35.73	40.05	23.34
MTFN	10.80	9.82	27.68	30.28	36.40	48.27	27.21
AMFMN	11.37	9.04	27.70	32.46	39.29	49.68	28.25
本文	**13.05**	**9.82**	**29.42**	**35.39**	**39.82**	**52.65**	**30.03**

表3 UCM数据集对比实验结果

Table 3 Comparative experimental results on UCM

方法	R_1		R_5		R_{10}		R_m
	图像检索文本	文本检索图像	图像检索文本	文本检索图像	图像检索文本	文本检索图像	
SCAN t2i	13.68	11.06	40.92	48.28	63.28	69.40	41.10
SCAN i2t	12.54	11.24	42.15	46.26	**65.45**	74.93	42.10
CAMP-triplet	10.37	8.30	38.71	45.13	62.78	70.50	39.30
CAMP-bce	14.22	10.91	39.02	47.95	62.98	72.56	41.27
MTFN	10.89	**14.06**	38.16	49.38	60.05	76.22	41.46
AMFMN	11.90	13.81	41.42	45.52	61.90	70.67	40.87
本文	**14.76**	11.33	**39.52**	**50.86**	61.43	**82.29**	**43.37**

表4 Sydney数据集对比实验结果

Table 4 Comparative experimental results on Sydney

方法	R_1		R_5		R_{10}		R_m
	图像检索文本	文本检索图像	图像检索文本	文本检索图像	图像检索文本	文本检索图像	
SCAN t2i	17.26	16.83	40.60	56.04	58.05	71.24	43.34
SCAN i2t	**19.48**	15.62	44.81	**58.20**	55.28	73.53	44.49
CAMP-triplet	17.80	14.50	45.28	45.63	61.44	69.55	42.37
CAMP-bce	14.26	13.41	42.53	50.32	57.35	73.86	41.96
MTFN	16.23	14.05	42.09	56.95	52.68	77.18	43.20
AMFMN	15.52	14.14	43.10	56.21	60.34	75.52	44.14
本文	15.52	**16.90**	**50.00**	55.17	**67.24**	**80.69**	**47.59**

索图像任务中的 R_1 提升了17.65%, R_5 提升了6.06%, R_{10} 提升了1.03%; 2个任务指标的平均值 R_m 总体提升了4.27%。AMFMN方法分别在浅层和深层提取特征进行多尺度特征融合后并进行冗余处理得到

的特征向量作为遥感图像的特征, 而文本嵌入是对句子进行嵌入并没有关注到文本中的每一个单词, 因此 AMFMN 方法的图像和文本的交互性不强。而本文方法采用全局语义和布局化视觉交叉指导方法, 将遥感图像中的每一个布局块和句子中的单词分别交互, 既实现了布局块和语义块的相互补充, 又解决了跨模态数据的鸿沟, 因而取得了较好的效果。

与 MTFN 方法在 RSITMD 数据集上的实验结果表明, 本文方法在图像检索文本任务中 R_1 提升了 20.83%, R_5 提升了 6.29%, R_{10} 提升了 9.40%; 在文本检索图像任务中 R_5 提升 16.88%, R_{10} 提升了 9.07%; 在 2 个检索任务中所有指标的平均值 R_m 提升了 10.36%。

3.6　消融实验

本文进行了 4 组消融实验:第 1 组实验验证 DSSLVFE 模块和 LV-GSCG 模块这 2 个子模块的性能, 假设 DSSLVFE 模块设为 A 模块, LV-GSCG 模块设为 B 模块; 第 2 组实验验证 DSSLVFE 模块中布局块的大小的对模型的影响; 第 3 组实验 MVM 模块中相似度度量计算公式中超参数 α_1 与 α_2 的不同标量值对模型性能的影响; 第 4 组实验验证损失函数中不同的 β 对模型性能的影响。

3.6.1　DSSLVFE 和 LV-GSCG 消融实验

对于第 1 组消融实验的具体结果如表 5 所示, 在遥感图像检索任务的 R_m 指标中, 在 AMFMN 的基础上添加 DSSLVFE 模块后, R_m 较基线提高了 4.65%, 添加 V-GSCG 模块和 DSSLVFE 模块后, R_m 较基线提高了 5.07%; 在图像检索文本任务中的 R_1、R_5、R_{10} 指标中, 在 AMFMN 的基础上添加 DSSLVFE 模块后所有的指标均成上升趋势, 分别提升 4.14%、0.80%、5.54%, 在 AMFMN 的基础上添加 DSSLVFE 模块和 LV-GSCG 模块后 R_1、R_5 分别提升了 22.89%、6.40%; 在文本检索图像任务中的 R_1、R_5、R_{10} 指标中, 在 AMFMN 基础上添加 DSSLVFE 模块后 3 个指标分别提升 7.76%、5.36%、5.23%, 在 AMFMN 基础上添加 DSSLVFE 模块和 LV-GSCG 模块后 R_1、R_5、R_{10} 分别提升了 15.52%、7.21%、2.13%。这表明 DSSLVFE 模块和 LV-GSCG 模块对于遥感图像文本检索有着重要的贡献。

3.6.2　DSSLVFE 模块中布局块的大小的消融实验

第 2 组实验验证 DSSLVFE 模块中布局块的大小对模型的影响, 对于第 2 组消融实验, 本文验证该模块中不同布局块个数对模型的影响。本文模型图像块大小从表 6 可以看出, 随着图像块大小的增加, R_m 呈现上升趋势, 但是当布局块大小增加到 64 的时候, R_m 呈现下降趋势。经分析可得, 相比于布局块大小为 64, 在布局块大小为 32 即布局块较小时, 图片总大小不变, 因此布局块数量变多。由于布局块较小, 目标捕捉不完全, 无法获得部分目标的完整语义; 而布局块数量太多, 导致模型计算复杂, 容易忽略有效信息。在布局块大小为 128 即布局块较大时, 图片总大小不变, 因此布局块数量变少。由于布局块过大, 同一布局块内存在多目标信息, 容易错误提取目标信息; 而布局块数量太少会导致布局块信息无法有效交互。因此, 经第 2 组消融实验证明, 布局块大小为 64 时为最优空间布局化表示, 最有效的抽取布局块的主导语义, 最终得到最优预测结果。

表 5　消融实验结果

Table 5　Ablation experimental results

方法	R_1		R_5		R_{10}		R_m
	图像检索文本	文本检索图像	图像检索文本	文本检索图像	图像检索文本	文本检索图像	
AMFMN	10.62	8.50	27.65	33.01	40.04	51.64	28.58
AMFMN+A	11.06	9.16	27.87	34.78	**42.26**	**54.34**	29.91
AMFMN+A+B	**13.05**	**9.82**	**29.42**	**35.39**	39.82	52.65	**30.03**

注: 加粗数值表示最优结果。

表 6　图像块个数的消融实验结果

Table 6　Ablation experimental results of number of image blocks

图像块个数	图像块大小	R_1		R_5		R_{10}		R_m
		图像检索文本	文本检索图像	图像检索文本	文本检索图像	图像检索文本	文本检索图像	
64	32	8.85	**9.42**	27.21	34.07	42.04	54.38	29.33
16	64	**12.83**	8.19	27.88	**34.12**	**42.26**	**55.53**	**30.13**
4	128	11.73	8.72	**28.98**	33.89	40.93	54.03	29.71

注: 加粗数值表示最优结果。

3.6.3 MVM中相似度度量计算公式中超参数α_1与α_2的消融实验

对于第3组消融实验，本文验证多视角匹配模块中相似度度量计算公式中超参数α_1与α_2的不同标量值对模型性能的影响，对于超参数α_1与α_2的不同标量值的性能，如表7所示，在4个评价指标下，加入超参数α_1与α_2并进行权重调整，从表7中可以看出，当随着α_1与α_2的不断降低R_m呈现上升趋势，当α_1与α_2均降至0.01时模型性能达到最优，之后随着α_1与α_2降低R_m呈现下降趋势。由于S_2代表了针对单一布局块的全局文本特征与单一文本块的相似度，S_3代表了针对单一词向量的全局图像特征与单一词向量的相似度，因此S_2和S_3相似度计算作用的特征区域比S_1作用区域小，其收敛速度更快。所以当α_1与α_2均为1时，S_1的影响被忽略。但当α_1与α_2非常小时，S_2和S_3的影响同样被忽略。因为，本文最终选取α_1与α_2均为0.01时，得到模型最优性能。

表7 相似度度量中参数α_1与α_2消融实验结果

Table 7 Parameters in similarity measurement α_1 and α_2 ablation experimental results

(α_1,α_2)	R_1		R_5		R_{10}		R_m
	图像检索文本	文本检索图像	图像检索文本	文本检索图像	图像检索文本	文本检索图像	
(1,1)	0.44	0.40	2.43	2.21	4.65	3.81	2.32
(0.1,0.1)	10.62	8.89	**28.10**	31.02	41.15	47.70	27.91
(0.05,0.05)	10.62	**9.96**	**28.10**	**34.34**	41.37	49.87	29.04
(0.01,0.01)	**12.83**	8.19	27.88	34.12	**42.26**	**55.53**	**30.13**
(0.000 5,0.000 5)	10.37	8.27	26.55	34.25	39.38	55.00	28.97
(0.000 1,0.000 1)	9.96	7.92	23.67	30.49	38.05	52.30	27.06

注：加粗数值表示最优结果。

3.6.4 损失函数中超参数β的消融实验

对于第4组消融实验，本文验证损失函数中超参数β的不同标量值对模型性能的影响，如表8所示，在4个评价指标下，加入超参数β进行权重调整后，当2个损失函数L_1和L_2占比相同时，性能最好。当β=0时，损失函数中只剩下了L_2，此时R_m下降了0.94%。当β=1时，损失函数只剩下L_1，R_m下降了0.37%。此外，增加L_1比增加L_2的权重更有利于训练模型。当β从0逐渐增加到1的过程中，R_m逐渐增大，但是当β=0.5时达到最大值，然后呈下降趋势。因此，盲目增加损失函数的权重是不可行的，当2种损失达到平衡时，得到的特征最准确。

表8 损失函数中β消融实验结果

Table 8 In the loss function β ablation experimental results

$(\beta,1-\beta)$	R_1		R_5		R_{10}		R_m
	图像检索文本	文本检索图像	图像检索文本	文本检索图像	图像检索文本	文本检索图像	
(1,0)	**13.05**	9.82	**29.42**	**35.39**	39.82	52.65	30.025
(0.5,0.5)	12.83	8.19	27.88	34.12	42.26	**55.53**	**30.135**
(0,1)	11.28	**11.37**	28.32	34.69	**43.58**	49.87	29.85

3.7 可视化展示

通过上述实验，得到如下结果。图4为图像生成文本结果的3个示例。通过训练完成的本文模型中各模块的计算，本文得到图像与不同文本的相似度大小，选出3个与图像相似度值最大的文本作为最终的跨模态检索结果。其中DSSL模块构建了图像的空间布局化表示并抽取布局块的主导语义，显著挖掘布局块特征。之后LV-GSCG模块进一步挖掘遥感图像全局语义信息，获取遥感图像指导的全局文本特征和遥感文本指导的全局图像特征，实现了布局化视觉信息和全局语义信息的交互。最后多视角匹配模块消除跨模态数据之间的语义鸿沟，精确地测量输入遥感图像和文本之间的距离。

从图4(a)和图4(b)可得，经过模型筛选出的文本，均与该图像的文本标签集合相匹配，展示了本文各个模块的有效性。从图4(c)可得，相似度排名第3位的文本与原数据集中图像的文本标签不匹配。经分析，应为“green land”单词产生的文本特征对结果产生过大影响，之后作者还会以此展开新的研究。

图5为文本生成图像结果的3个示例。通过本文模型中各个模块的计算，本文得到文本与不同图像的相似度大小，最终选出3个与文本相似度值最大的图像作为最终的跨模态检索结果。从图5(a)和图5(b)可得，经过模型筛选出的文本，均与该文本的图像标签集合相匹配，展示了本文各个模块的



Top 1: A white building with two domes in the center of the earth. (√)
Top 2: There are two white domed buildings in the center of the earth. (√)
Top 3: A white building with two domes in the middle. (√)

(a) 示例1



Top 1: Two planes were parked next to a rectangular gray building. (√)
Top 2: Four planes were parked in the open air. (√)
Top 3: The two planes stopped beside a rectangular, gray building. (√)

(b) 示例2



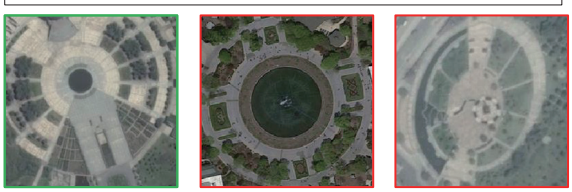
Top 1: There are many houses on one side of the viaduct. (√)
Top 2: There are many houses over the viaduct. (√)
Top 3: The high and low buildings are located on the green land. (×)

(c) 示例3

图 4　图像生成文本结果

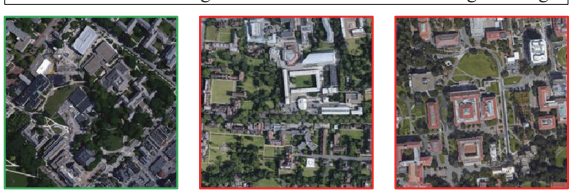
Fig. 4　Image generated text result

A straight road passes through the central ellipse.



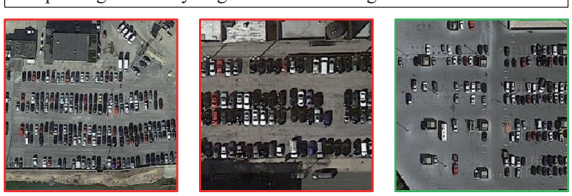
(a) 示例1

Lush trees are decorating this school with several teaching buildings.



(b) 示例2

The parking lot is very large and covers a large area.



(c) 示例3

图 5　文本生成图像结果

Fig. 5　Text generated image result

有效性。从图 5(c)可得, 相似度排名前两位的图像与原数据集中文本的图像标签不匹配。经分析, 应为遥感数据集中存在某些图像样本过于相似, 导致产生的预测结果存在较大误差。

4　结　论

1) 提出了 SL-SJR 方法, 将遥感图像序列化为图像块来实现图像的布局化的建模, 来实现遥感对象的远距离依赖和次要目标的保留, 极大地提高了遥感图文检索的准确度。

2) 设计了 DSSL 模块, 在主导语义类别特征的监督下, 融合 Transformer 方法挖掘每个布局化视觉特征的显著信息。通过将布局块特征和主导语义类别特征融合 Transformer 方法解决了遥感图像中目标分散而导致远距离建模困难以及遥感相邻次要目标被淹没的问题。此外, 提出了 LV-GSCG 模块, 通过计算布局化视觉特征与文本中提取的全局语义特征的相似度来实现不同模态特征的交互, 达到跨模态的特征的相互补充。最后, 构建了 MVM 模块, 建立跨模态特征指导的多视角度量匹配机制, 缓解了跨模态检索的语义鸿沟。

3) 在 4 个公开遥感图文检索数据集上来评估 SL-SJR 方法, 通过对比实验和全面的消融实验验证了所提出方法的有效性。

参考文献 (References)

[1] LIU Y J, LI X F, REN Y B. A deep learning model for oceanic mesoscale eddy detection based on multi-source remote sensing imagery[C]// IGARSS 2020 - 2020 IEEE International Geoscience and Remote Sensing Symposium. Piscataway: IEEE Press, 2020: 6762-6765.

[2] ZHANG Q L, SETO K C. Mapping urbanization dynamics at regional and global scales using multi-temporal DMSP/OLS nighttime light data[J]. Remote Sensing of Environment, 2011, 115(9): 2320-2329.

[3] NOGUEIRA K, FADEL S G, DOURADO I C, et al. Exploiting ConvNet diversity for flooding identification[J]. IEEE Geoscience and Remote Sensing Letters, 2018, 15(9): 1446-1450.

[4] CHENG Q M, ZHOU Y Z, FU P, et al. A deep semantic alignment network for cross-modal image-text retrieval in remote sensing[J]. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2021, 14: 4284-4297.

[5] YUAN Z Q, ZHANG W K, FU K, et al. Exploring a fine-grained multiscale method for cross-modal remote sensing image retrieval[J]. IEEE Transactions on Geoscience and Remote Sensing, 2022, 60: 4404119.

[6] 葛芸, 马琳, 叶发茂, 等. 基于多尺度池化和范数注意力机制的遥感图像检索[J]. 电子与信息学报, 2022, 44(2): 543-551.
GE Y, MA L, YE F M, et al. Remote sensing image retrieval based on multi-scale pooling and norm attention mechanism[J]. Journal of Electronics & Information Technology, 2022, 44(2): 543-551(in Chinese).

[7] 李彦甫, 范习健, 杨绪兵, 等. 基于自注意力卷积网络的遥感图像分类[J]. 北京林业大学学报, 2021, 43(10): 81-88.
LI Y F, FAN X J, YANG X B, et al. Remote sensing image classification farmework based on self-attention convolutional neural network[J]. Journal of Beijing Forestry University, 2021, 43(10): 81-88(in Chinese).

[8] YUAN Z Q, ZHANG W K, TIAN C Y, et al. Remote sensing cross-

modal text-image retrieval based on global and local information[J]. IEEE Transactions on Geoscience and Remote Sensing, 2022, 60: 3163706.

[9] RONG X E, SUN X, DIAO W H, et al. Historical information-guided class-incremental semantic segmentation in remote sensing images[J]. IEEE Transactions on Geoscience and Remote Sensing, 2022, 60: 5622618.

[10] ZHANG Z Y, HAN X, LIU Z Y, et al. ERNIE: Enhanced language representation with informative entities[C]//Annual Meeting of the Association for Computational Linguistics. Stroudsburg: ACL, 2019: 1441-1451.

[11] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, et al. An image is worth 16x16 words: transformers for image recognition at scale [C]//International Conference on Learning Representations. Washington DC: ICLR, 2020: 16-28.

[12] CARION N, MASSA F, SYNNAEVE G, et al. End-to-end object detection with transformers [C]// In European Conference on Computer Vision. Berlin: Springer, 2020: 213-229.

[13] SRINIVAS A, LIN T Y, PARMAR N, et al. Bottleneck transformers for visual recognition[C]// /IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2021: 16514–16524.

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

[15] LIU S L, ZHANG L, YANG X, et al. Query2Label: A simple transformer way to multi-label classification[C]//IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2021: 661-670.

[16] MESSINA N, AMATO G, FALCHI F, et al. Towards efficient cross-modal visual textual retrieval using transformer-encoder deep features[C]// 2021 International Conference on Content-Based Multimedia Indexing. Piscataway: IEEE Press, 2021: 1-6.

[17] GABEUR V, SUN C, ALAHARI K, et al. Multi-modal transformer for video retrieval[C]// Conference on Computer Vision. Berlin: Springer, 2020: 214-229.

[18] MALEKI D, TIZHOOSH H R. LILE: look in-depth before looking elsewhere -- A dual attention network using transformers for cross-modal information retrieval in histopathology archives[J]. Proceedings of Machine Learning Research. Virtual: PMLR, 2022: 3002-3013.

[19] SHI Z W, ZOU Z X. Can a machine generate humanlike language descriptions for a remote sensing image[J]. IEEE Transactions on Geoscience and Remote Sensing, 2017, 55(6): 3623-3634.

[20] LU X Q, WANG B Q, ZHENG X T, et al. Exploring models and data for remote sensing image caption generation[J]. IEEE Transactions on Geoscience and Remote Sensing, 2018, 56(4): 2183-2195.

[21] HOXHA G, MELGANI F, SLAGHENAUFFI J. A new CNN-RNN framework for remote sensing image captioning[C]//2020 Mediterranean and Middle-East Geoscience and Remote Sensing Symposium. Piscataway: IEEE Press, 2020: 1-4.

[22] LI X L, ZHANG X T, HUANG W, et al. Truncation cross entropy loss for remote sensing image captioning[J]. IEEE Transactions on Geoscience and Remote Sensing, 2021, 59(6): 5246-5257.

[23] LU X Q, WANG B Q, ZHENG X T. Sound active attention framework for remote sensing image captioning[J]. IEEE Transactions on Geoscience and Remote Sensing, 2020, 58(3): 1985-2000.

[24] FAGHRI F, FLEET D J, KIROS J R , et al. VSE++: improving visual-semantic embeddings with hard negatives[C]// British Machine Vision Conference. London : British Machine Vision Association, 2017: 1707-1717.

[25] LEE K H, CHEN X, HUA G, et al. Stacked cross attention for image-text matching[C]//European Conference on Computer Vision. Berlin: Springer, 2018: 212-228.

[26] WANG T, XU X, YANG Y, et al. Matching images and text with multi-modal tensor fusion and re-ranking[C]// the 27th ACM International Conference. New York: ACM, 2019, 1: 12-20.

[27] DEVLIN J, CHENG H, FANG H, et al. Language models for image captioning: the quirks and what works[J]. Computer Science, 2015, 2(53): 100-105.

[28] ABDULLAH, BAZI, RAHHAL A, et al. TextRS: Deep bidirectional triplet network for matching text to remote sensing images[J]. Remote Sensing, 2020, 12(3): 405.

[29] QU B, LI X L, TAO D L, et al. Deep semantic understanding of high resolution remote sensing image[C]// International Conference on Computer. Piscataway: IEEE Press, 2016: 124-128.

[30] WANG Z H, LIU X H, LI H S, et al. CAMP: Cross-modal adaptive message passing for text-image retrieval[C]//2019 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway: IEEE Press, 2010: 5763-5772.

Remote sensing image-text retrieval based on layout semantic joint representation

ZHANG Ruoyu，NIE Jie*，SONG Ning，ZHENG Chengyu，WEI Zhiqiang

(Department of Information Science and Engineering, Ocean University of China, Qingdao 266101, China)

Abstract: Remote sensing image-text retrieval can retrieve valuable information from remote sensing data. It is of great significance to environmental assessment, urban planning and disaster prediction. However, there is a key problem that the spatial layout information of remote sensing images is ignored, which is mainly reflected in two aspects: one is the difficulty of long-distance modeling of remote sensing targets; the other, the submerge of the remote sensing adjacent secondary targets. Based on the above problems, this paper proposes a cross modal remote sensing image-text retrieval model based on layout semantic joint representation, which includes the dominant semantic supervison layout visual feature extraction module (DSSL), Layout visual-global semantic cross guidance (LV-GSCG) and multi-view matching (MVM). The DSSL module realizes the layout modeling of images under the supervision of dominant semantic category features. The LV-GSCG module calculates the similarity between the layout visual features and the global semantic features extracted from text to realize the interaction of different modal features. The MVM module establishes a cross-modal feature-guided multi-view metric matching mechanism to eliminate the semantic gap between the cross-modal data. Experimental validation on four baseline remote sensing image text datasets shows that the model can achieve state-of-the-art performance in most cross-modal remote sensing image text retrieval tasks.

Keywords: remote sensing image; cross modal retrieval; spatial layout information; leading semantic supervision; class supervision mechanism

Received：2022-06-22；**Accepted**：2022-07-08；**Published Online**：2023-01-12 14：17

URL：link.cnki.net/urlid/11.2625.V.20230111.1851.005

Foundation items：National Key Research and Development Program of China (2021YFF070400); National Natural Science Foundation of China (62072418,62172376); the Fundamental Research Funds for the Central Universities (202042008)

* **Corresponding author**. E-mail：niejie@ouc.edu.cn

http://bhxb.buaa.edu.cn jbuaa@buaa.edu.cn
DOI: 10.13700/j.bh.1001-5965.2022.0599

图文数据的多级关系分析与挖掘方法

郭瑞萍，王海荣*，王栋
(北方民族大学 计算机科学与工程学院，银川 750021)

摘　　要：如何高效挖掘多模态数据间隐藏的语义关联是当前多模态知识抽取的重点任务之一，为更细粒度地挖掘图像与文本数据间关系，提出了一种多级关系分析与挖掘（MRAM）方法，引入BERT-Large模型，提取文本特征构建文本连接图，利用Faster-RCNN网络提取图像特征来学习空间位置关系和语义关系并构建图像连接图，进而完成单模态内部语义关系计算，在此基础上，使用节点切分方法和带多头注意力机制的图卷积网络(GCN-MA)进行局部和全局的图文关系融合。此外，为提升关系挖掘效率，采用了基于注意力机制的连边权重剪枝策略，用以增强重要分支表示，减少冗余信息干扰。在公开的Flickr30K、MSCOCO-1K、MSCOCO-5K数据集上进行方法实验，并与11种方法进行实验结果的对比分析，所提方法在Flickr30K上的平均召回率提高了0.97%和0.57%，在MSCOCO-1K上的平均召回率提高了0.93%和0.63%，在MSCOCO-5K上的平均召回率提高了0.37%和0.93%，实验结果验证了所提方法的有效性。

关　键　词：关系挖掘；多级关系；注意力机制；图卷积网络；图文数据
中图分类号：V 221^{+}.3; TB 553
文献标志码：A　　**文章编号**：1001-5965(2024)02-0684-11

随着文本、图像、视频等多模态数据急剧增涨，对数据的获取、分析与处理逐步由单模态转化为多模态形式。大量多模态数据之间包含语义关联，如何挖掘到语义相近的多模态数据进行知识抽取成为研究热点，关系挖掘作为多模态知识抽取中的重要任务也备受关注。现有研究主要聚焦于异构数据相似性来学习对应关系，可大致分为全局表示方法和区域词对应方法。

全局表示方法分别提取图像和句子的全局特征，文献[1]介绍了深度图像标签嵌入模型，采用图卷积网络从图像中提取视觉特征，并使用SkipGram提取语义标签，与匹配图像标签对的结构化目标相关联；文献[2]提出用于图像句子联合表示的网络，利用图文间的最终分数进行融合和匹配；文献[3]设计了高效的字符接收模块，该模块在不同粒度级别卷积原始字符来学习文本语义嵌入，并将图文表示编码到多模态空间中，通过对比成对损失函数关联图文间语义信息；文献[4]提出一种基于秩的张量融合网络来学习图文数据全局特征下的相似性；文献[5]通过语言引导的图注意网络来捕获图像间的关联；文献[6]学习图像中显著对象和文本之间的公共语义空间来获得关系增强的特征；文献[7]提出了联合模态相似矩阵的无监督哈希方法来保持2种模态间相关性，利用抽样和加权方法，使具有相似语义的样本距离更近，不同语义样本距离更远；文献[8]采用图卷积网络方法使每个样本的语义信息更完整，利用生成式对抗网络在公共空间中获得模态不变表示。这些方法使显著对象在图文数据整体表示中起主导作用，但涉及到次要对象时往往被忽略。

收稿日期：2022-07-09；**录用日期**：2022-08-14；**网络出版时间**：2022-10-18 11:00
网络出版地址：link.cnki.net/urlid/11.2625.V.20221017.1553.007
基金项目：宁夏自然科学基金项目(2023AAC03316)；宁夏回族自治区教育厅高等学校科学研究重点项目(NYG2022051)；北方民族大学科研项目(2021XYZJK06)
* **通信作者**. E-mail：bmdwhr@163.com

引用格式：*郭瑞萍，王海荣，王栋. 图文数据的多级关系分析与挖掘方法[J]. 北京航空航天大学学报，2024，50(2)：684-694.*
GUO R P，WANG H R，WANG D. Multilevel relation analysis and mining method of image-text[J]. Journal of Beijing University of Aeronautics and Astronautics，2024，50(2)：684-694 (in Chinese).

区域词对应方法侧重于学习图文数据的局部区域词对应, 文献 [9] 在网络迭代学习过程中计算区域和单词的相似性, 但只能优化最相似的部分, 效果也不够明显; 文献 [10] 设计了交叉注意力机制来加权较重要的区域, 以图像区域和句子中单词作为上下文计算图文相似度; 文献 [11] 利用语义概念结合正确语义顺序来改进图像表示, 设计出有利于语义增强的相似度匹配模型; 文献 [12] 提出 GCN-LSTM(graph convolutional networks-long short term memory) 架构将语义和空间对象关系集成到图像编码器中学习区域级表示; 文献 [13] 将外部先验知识嵌入到区域语义空间中来指导推理图像视频中的对象关系; 文献 [14] 设计了位置聚焦注意网络来探索图文数据间相关性, 将位置信息和视觉特征集成到图像表示中, 使视觉表达更加完整; 文献 [15] 提出了一种迭代匹配与重复注意记忆方法来选择突出特征, 使用多步对齐来捕获图文数据间对应关系, 通过一种内存蒸馏单元来融合多模态特征。

上述方法虽然学到了突出对象的对应关系, 但没有明确揭示出图文对象间的语义关系, 因此, 一些研究者提出将整体与区域结合起来进一步提升图文关系捕获效果。文献 [16] 设计了视觉语义推理 (visual semantic reasoning, VSRN) 模型通过区域关系和全局语义利用图卷积神经网络 (graph convolutional networks, GCN) 推理生成可视化关系信息; 文献 [17] 结合全面和细粒度的多模态交互方法, 并采用自适应门控制方案合理处理冗余信息; 文献 [18] 构建了视觉对象和关系表示的场景图, 并利用 GCN 模块对场景图进行关系识别与渐进推理; 文献 [19] 利用多注意力机制与剩余学习策略将全局上下文和局部引导特征相结合进而计算多实例表示; 文献 [20] 提出上下文感知注意网络, 通过聚合全局上下文选择性地关注关键局部片段; 文献 [21] 介绍了场景图匹配方式捕获图文数据中的对象和关系, 进一步提高了性能; 文献 [22] 利用分类损失把每个多模态数据视为一个类来学习图文数据模态间相似性; 文献 [23] 将高阶语义信息进行视觉语义匹配, 通过图卷积网络来处理由每个节点表示的图像和文本场景图, 以交叉图注意机制的方式来计算图文数据间相似度; 文献 [24] 采用图像级 GCN 来捕获相似对象提供的特征信息, 使用语言转换器理解图像中不同对象间的关系。

综上, 基于图文数据关系挖掘方法的研究在计算异构数据相似性、学习对应关系方面已经取得了一定的成果, 但仍然存在文本和图像特征有效信息易丢失和语义关系挖掘不充分的问题。为此, 本文提出一种图文数据的多级关系分析与挖掘(multi-level relationship analysis and mining of image-text, MRAM)方法, 聚焦于多模态数据间隐藏的高级语义关联, 融合 BERT-Large(bidirectional encoder representation from transformers-large) 模型与 Bi-GRU (bidirection gated recurrent unit) 挖掘文本关系, 利用空间位置和语义交互 2 种方式学习图像关系, 分别计算局部和全局的图文节点间相似度, 基于注意力机制的连边权重剪枝策略处理无关信息。

1　方法模型

MRAM 方法 (如图 1 所示) 使用 BERT-Large 预训练模型与 Bi-GRU 网络提取文本信息, 调用空间关系编码器进行空间关系提取、建立图像连接图学习区域之间相关性, 计算局部相似性时提出“节点切分方法”减少计算复杂度, 实现并行处理, 基于多

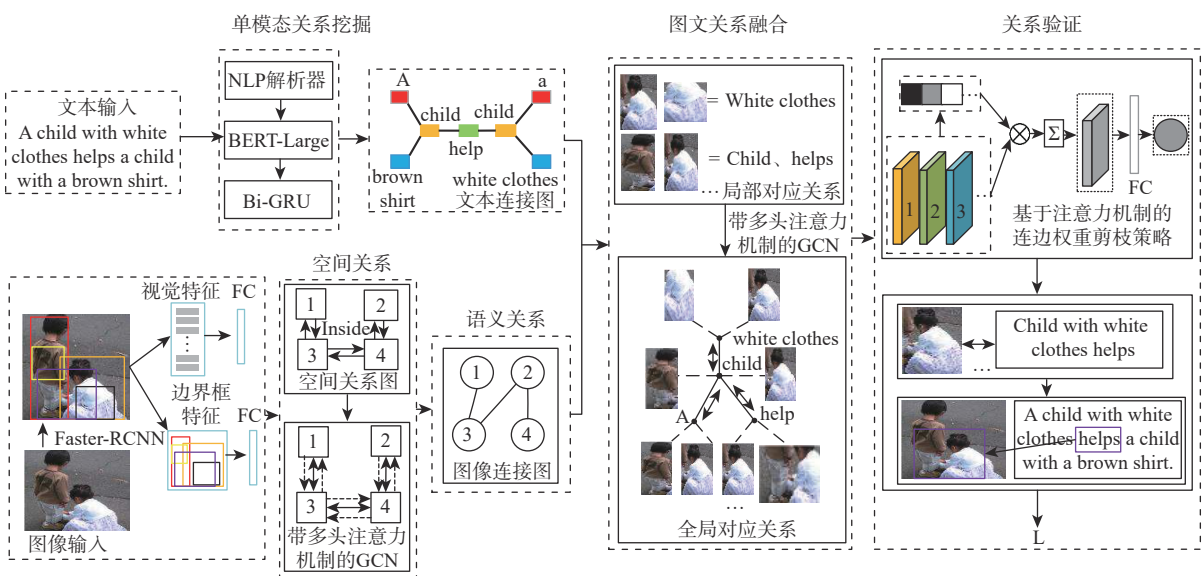


图 1　多级关系分析与挖掘模型图

Fig. 1　Multi-level relationship analysis and mining model diagram

头注意力机制的GCN进行局部对应关系补全与传播，形成全局对应关系，采用融合注意力机制的连边权重剪枝策略优化方法模型。具体处理流程如图1所示。

图1展示了多级关系分析与挖掘模型图，其主要包含单模态关系挖掘、图文关系融合、关系验证3个核心部分。

2 单模态关系挖掘

由于图文数据表现特征具有底层特征异构、高层语义相关性，为了便于关系挖掘，MRAM方法第1步工作即是分模态进行统一特征表示，在其基础上，分别进行单模态内部关系挖掘。

2.1 特征表示

2.1.1 文本表示

给定一个句子$\boldsymbol{E}$，采用NLP(neuro linguistic programming)句法分析工具进行预处理后，添加2个特殊的标记：[CLS]和[SEP]，并将其输入预训练的BERT-Large模型[25]作为整个模型的文本编码层，经过嵌入层和24层transformer编码器块来学习文本中的内部语义信息，设隐藏层大小为1 024，自注意力头数为16，可训练参数约340 M。通过Transformer编码器和自注意力机制结合可以捕获文本的上下文信息。每个单词的最终嵌入是标记嵌入、位置嵌入、语句分块嵌入的组合，具体标记公式为

$$\boldsymbol{E}=[\boldsymbol{e}_1,\boldsymbol{e}_2,\cdots,\boldsymbol{e}_z] \tag{1}$$

式中：$\boldsymbol{e}_z$表示句子$\boldsymbol{E}$中的第z个单词，将其输入预训练的BERT-Large模型作为整个模型的文本编码层，计算公式为

$$\boldsymbol{D}_x=\text{BERT-Large}'(\boldsymbol{E}_x)\qquad x\in[1,z] \tag{2}$$

式中：$\boldsymbol{D}_x$表示将句子$\boldsymbol{E}_x$输入到BERT-Large模型后的结果向量，将结果向量$\boldsymbol{D}_x$通过双向GRU网络映射到h维嵌入空间中生成文本表示，计算公式为

$$\boldsymbol{T}_x=\text{Bi-GRU}''(\boldsymbol{D}_x)\qquad x\in[1,z] \tag{3}$$

式中：$\boldsymbol{T}_x=\{\boldsymbol{t}_1,\boldsymbol{t}_2,\cdots,\boldsymbol{t}_z\}$表示利用前向GRU和后向GRU读取$\boldsymbol{D}_x$后的结果，最后一层隐藏状态是整个句子的文本语义表示，包含了高度集中的语义信息。

2.1.2 图像表示

输入一张图像$\boldsymbol{I}$，使用Faster-RCNN网络结合预训练的ResNet-101提取区域级图像特征，ResNet-101[26]在VG数据集[27]上进行了预训练。采用区域级特征编码器预测实例类和空间坐标来进行特征表示。对每个类别用非最大抑制法，联合阈值设为0.65，置信度阈值设为0.35，选择置信度得分最高的前48个感兴趣区域。给定在图像中检测到的区域i，平均池化层后选择2 048维特征$\boldsymbol{e}_i'$，利用全连接层将$\boldsymbol{e}_i'$转换为d维嵌入空间，转换公式为

$$\boldsymbol{v}_i=\boldsymbol{W}_1\boldsymbol{e}_i'+\boldsymbol{b}_1 \tag{4}$$

式中：$\boldsymbol{W}_1$和$\boldsymbol{b}_1$分别为权重矩阵和偏差。

图像区域特征提取的边界框用一个四维空间坐标$\boldsymbol{s}_i=[\boldsymbol{A}_i,\boldsymbol{B}_i,\boldsymbol{G}_i,\boldsymbol{H}_i]$表示，其中$(\boldsymbol{A}_i,\boldsymbol{B}_i)$表示框中左上点的横坐标与纵坐标，$(\boldsymbol{G}_i,\boldsymbol{H}_i)$为边界框的宽度和高度。每张图像用一组对象$\boldsymbol{O}=\{\boldsymbol{o}_1,\boldsymbol{o}_2,\cdots,\boldsymbol{o}_a\}$来表示，$a$代表图像区域数。本文实验中，将$a$设为48，每个对象$\{\boldsymbol{o}_i=\boldsymbol{v}_i\|\boldsymbol{s}_i\}$都与图像特征向量$\boldsymbol{v}_i$、边界框特征向量$\boldsymbol{s}_i$相关联，整个图像可以表示为$\boldsymbol{O}=[\boldsymbol{v}_1\|\boldsymbol{s}_1,\boldsymbol{v}_2\|\boldsymbol{s}_2,\cdots,\boldsymbol{v}_i\|\boldsymbol{s}_i,\cdots,\boldsymbol{v}_a\|\boldsymbol{s}_a]$。

2.2 关系挖掘

图像文本分别进行特征表示完后，对其进行内部关系挖掘。文本通过构建文本连接图来计算节点间的相似度，相似度大小即为连边权重。对于图像，基于空间关系编码器来确定空间关系，采用基于多头注意力机制的GCN加权重要关系，将区域特征设置为图节点，参考空间关系给区域特征节点建立连边构建图像连接图，计算图节点间的相似度来挖掘图像内部语义关系。

2.2.1 文本关系挖掘

使用NLP句法分析工具识别句子中的语义依赖关系，采用预训练BERT-Large模型结合Bi-GRU网络从输入句子中提取特征向量$\boldsymbol{t}$。构建文本连接图时，将每个单词设置为节点，根据句法分析决定词节点间是否连边，将连接图设为$J=\{X,F\}$，$\mathbb{X}$是单词节点的集合，$\mathbb{F}$是边的集合，记单词节点数为c。如图2展示了文本关系挖掘过程图。

图2为用NLP工具识别语义依赖关系、采用预训练BERT-Large模型和Bi-GRU网络提取句子特征来改进文本表示挖掘文本关系的过程，最终构建起文本连接图。通过内积公式计算文本连接图中两两节点间相似性，从而获得连边权重：

$$\boldsymbol{Q}_{x,y}=\frac{\exp\left(\lambda\boldsymbol{t}_x^{\mathrm{T}}\boldsymbol{t}_y\right)}{\sum_{y=0}^{c}\exp\left(\lambda\boldsymbol{t}_x^{\mathrm{T}}\boldsymbol{t}_y\right)} \tag{5}$$

式中：设所求单词表示$\boldsymbol{t}$的相似矩阵为$\boldsymbol{Q}_{x,y}$，$\boldsymbol{Q}_{x,y}$代表第x个和第y个单词节点间的相似性；λ是一个比例因子，使用参数λ乘以内积结果进行运算，利用softmax归一化得$\boldsymbol{Q}_{x,y}$。单词节点间的语义依赖性即连边权重，设为$\boldsymbol{W}_3$，具体公式为

$$\boldsymbol{W}_3=\left\|\boldsymbol{Q}_{x,y}\cdot\boldsymbol{K}\right\|_2 \tag{6}$$

式中：矩阵$\boldsymbol{K}$表示节点的邻接矩阵，连边权重$\boldsymbol{W}_3$通

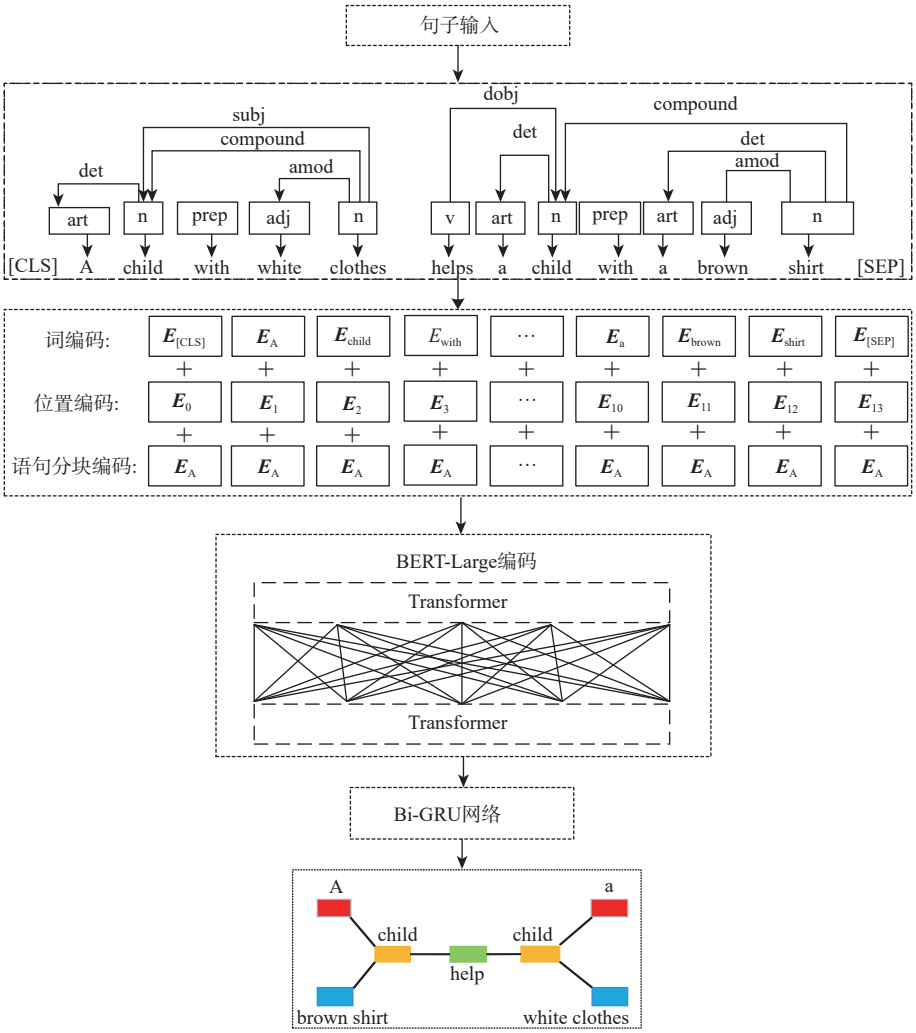


图 2　文本关系挖掘图

Fig. 2　Text relationship mining graph

过相似度矩阵$\boldsymbol{Q}_{x,y}$和邻接矩阵$\boldsymbol{K}$相乘得到。最后经过L_2正则, 从而归一化权重、减少梯度爆炸等问题。

2.2.2　图像关系挖掘

在进行图像内部关系挖掘时, 从空间关系和语义关系 2 个层面进行分析, 其分析挖掘过程如图 3 所示。

图 3 展示了将区域特征提取完后图像关系挖掘的 2 个阶段, 即空间关系和语义关系挖掘。空间关系引入空间关系编码器[28], 其通过连接孤立区域来反映物体的空间形式和对象之间的关系。给定 2 个区域对象, 通过三元组$\boldsymbol{q}_{u,v}=(\boldsymbol{o}_u,\boldsymbol{r},\boldsymbol{o}_v)$来表示对象间相对几何位置的空间关系, $\boldsymbol{o}_u$表示图像区域节点 u 的对象特征向量, $\boldsymbol{r}$表示关系特征向量, $\boldsymbol{o}_v$表示邻居区域节点 v 的对象特征向量, 边缘和相应标签由 IoU(目标检测中预测边框和真实边框的交并比)、相对距离和角度决定, 当 2 个区域距离很远或者重叠区域较少时, 他们之间存在关系的可能性也就很小。将空间关系图设为$\boldsymbol{Z}_{u,v}=(\boldsymbol{V}',\boldsymbol{q}_{u,v})$, $\boldsymbol{q}_{u,v}$和$\boldsymbol{q}_{v,u}$包含对称方向的边, $\boldsymbol{V}'$为图像空间特征节点的集合, 原始 GCN 不包含有向边和标签信息, 但空间图是有方向性的, 并且包含了标签信息, 因此, 图的不同方向和边标签要分别通过变换矩阵和偏差进行转换, 这使得空间编码器对方向和标签都很敏感。随后采用带多头注意力机制的 GCN 推理空间图, 把每个注意头的输出特征连接起来, 通过聚焦重要边来挖掘空间关系, 得到相应的区域图像表示, 计算公式为

$$\boldsymbol{v}_i'=\Big\|_{n=1}^{N'}\mathrm{ReLU}\left(\sum_{v_j\in N(\boldsymbol{v}_i)}\boldsymbol{W}^n_{\mathrm{dir}(i,j)}\cdot\left(\alpha^n_{i,j}\boldsymbol{v}_j\right)+\boldsymbol{b}^n_{L(i,j)}\right)\tag{7}$$

式中: N'为头注意力的数量; $N(\boldsymbol{v}_i)$为$\boldsymbol{v}_i$的邻居集合; $\boldsymbol{W}^n_{\mathrm{dir}(i,j)}$为一个相关权重矩阵; $\boldsymbol{b}^n_{L(i,j)}$为偏置项; $\boldsymbol{d}'(i,j)$为每条边方向性敏感的变换矩阵; $L(i,j)$为连边$(\boldsymbol{v}_i\text{-}\boldsymbol{v}_j)$的标签; $\alpha^n_{i,j}$为多头注意力系数, 注意力系数计算公式为

$$\alpha_{i,j}=\frac{\exp\left((\boldsymbol{W}_4\boldsymbol{v}_i)^{\mathrm{T}}\cdot\boldsymbol{V}'_{\mathrm{dir}(i,j)}\boldsymbol{v}_j\right)+\boldsymbol{b}^4_{L(i,j)}}{\sum_{\boldsymbol{v}_j\in N(\boldsymbol{v}_i)}\exp\left((\boldsymbol{W}_4\boldsymbol{v}_i)^{\mathrm{T}}\cdot\boldsymbol{V}'_{\mathrm{dir}(i,j)}\boldsymbol{v}_j\right)+\boldsymbol{b}^4_{L(i,j)}}\tag{8}$$

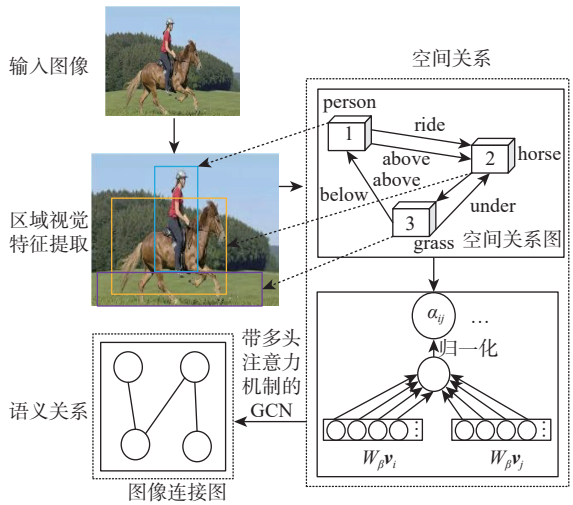


图3 图像关系挖掘图

Fig. 3 Image relation mining graph

式中：$\boldsymbol{W}_4$和$\boldsymbol{V}'_{\mathrm{dir}(i,j)}$为相关变换矩阵；$\boldsymbol{b}'_{L(i,j)}$为偏置项，通过基于注意力的空间关系编码器对区域进行编码后，区域特征学习到了对象间更精准的空间关系。

语义关系则在获得具有空间关系的区域特征$\boldsymbol{v}'_i$后，将区域特征连接构成图像连接图$M=(V,H)$，V表示图节点的集合，H表示由n个图节点组成的$n\times(n-1)$个节点对构建的连边的集合，计算每个节点对的相似度，计算公式为

$$\boldsymbol{P}_{i,j}=\frac{\exp\left(\lambda \boldsymbol{v}_i'^{\mathrm{T}}\boldsymbol{v}_j\right)}{\sum_{j=0}^{n}\exp\left(\lambda \boldsymbol{v}_i'^{\mathrm{T}}\boldsymbol{v}_j\right)} \tag{9}$$

式中：$\boldsymbol{P}_{i,j}$表示第i个和第j个图像区域节点间的相似性，之后通过计算连边权重来度量图节点间的相似性。

3 图文关系融合

单模态内部关系挖掘完后进行图文关系融合，即在前面基础上进一步学习图文数据间局部以及全局对应关系。将文本连接图中的节点用$\boldsymbol{J}'_\alpha\in\mathbf{R}^{c\times d}$表示，图像连接图中的节点用$\boldsymbol{M}'_\beta\in\mathbf{R}^{n\times d}$表示，$c$为文本图中的节点数，$d$为维数，$n$为图像连接图中的节点数。为融合具有语义相似性的多模态数据，需计算局部某个区域节点与另一模态节点间的相似度，并采用“节点切分方法”实施相似性操作来建立局部关联。为捕获全局对应关系，采用基于多头注意力机制的GCN通过相邻节点来进行对应关系补全与传播。

3.1 局部对应关系

参考前面的图像连接图和文本连接图，输入句子后，需要从文本连接图中找到和图像节点相似的文本节点，即单词嵌入节点$\boldsymbol{J}'_\alpha$，图像中节点$\boldsymbol{M}'_\beta$，然后计算内积，再乘λ权重进行归一化操作，这样每个文本都得到一个相似度权重，再把权重和原文本嵌入相乘得出结果，即$\boldsymbol{U}_{m\to j}$表示句子中有哪些词是和图像中节点对应。计算公式为

$$\boldsymbol{U}_{m\to j}=\mathrm{softmax}_\alpha\left(\lambda \boldsymbol{M}'_\beta \boldsymbol{J}'^{\mathrm{T}}_\alpha\right)\boldsymbol{J}'_\alpha \tag{10}$$

类似地，$\boldsymbol{U}_{j\to m}$表示图像连接图中有哪些节点和句子中的单词节点对应，计算公式为

$$\boldsymbol{U}_{j\to m}=\mathrm{softmax}_\beta\left(\lambda \boldsymbol{J}'_\alpha \boldsymbol{M}'^{\mathrm{T}}_\beta\right)\boldsymbol{M}'_\beta \tag{11}$$

为更好地计算相似性，需要每个节点与其对应的另一模态节点采用“节点切分方法”进一步实施局部相似性操作。将文本节点的第m个特征切分为等值大小的j个小节点：$[\boldsymbol{t}_{m,1},\boldsymbol{t}_{m,2},\cdots,\boldsymbol{t}_{m,j}]$，其对应节点也划分为等值大小的$j$个小节点：$[\boldsymbol{v}_{m,1},\boldsymbol{v}_{m,2},\cdots,\boldsymbol{v}_{m,j}]$，如$\boldsymbol{t}_{m,j}$表示相关句子向量中单词的集合，将整个向量划分为$j$个小节点，第1个表示为$\boldsymbol{t}_{m,1}$，其所对应的就是图像中的某个小节点$\boldsymbol{v}_{m,1}$，小节点和小节点之间计算相似性，将$\boldsymbol{t}_{m,1}$和$\boldsymbol{v}_{m,1}$进行余弦相似度计算，$\boldsymbol{t}_{m,2}$和$\boldsymbol{v}_{m,2}$进行余弦相似度计算，计算公式为

$$\boldsymbol{h}_{m,e''}=\cos\left(\boldsymbol{v}_{m,e''},\boldsymbol{t}_{m,e''}\right)\qquad e''\in[1,j] \tag{12}$$

式中：$\cos(\cdot)$为余弦相似度函数；$\boldsymbol{h}_{m,e''}$为一个标量值。再将每个小节点计算出来的余弦相似性拼接，计算公式为

$$\boldsymbol{h}_m=\boldsymbol{h}_{m,1}\|\boldsymbol{h}_{m,2}\|\cdots\|\boldsymbol{h}_{m,j} \tag{13}$$

式中：“||”表示串联计算；$\boldsymbol{h}_m$为第m个文本节点的对应向量，最后连接得出所有节点相似性。通过节点切分方法，可以将句子中的单词节点与其相似度较高的图像节点相关联，该方法可以减少计算复杂度，通过分成多个小节点可实现更好的并行化运行，也可以在后续找对应关系的时候，使来自不同位置的向量起到不同的作用。通过该方法计算相似性后，可以将对应关系保留并表示为一个向量，如果不切分，计算完相似性之后就为一个标量，不方便之后计算。

3.2 全局对应关系

完成局部对应关系后，将所得结果输入，对于文本连接图，将每个单词节点和局部相似度较高的图像节点建立对应关系。为了对全局模式下的多模态数据特征进一步融合，采用一层GCN应用k个内核来进行局部对应关系的传播与推理，关系的补全与传播通过相邻局部聚合信息来更新。GCN具体计算公式为

$$\boldsymbol{h}'_m=\Big\|_{k=1}^{K}\sigma\left(\sum_{i\in N'_m}\boldsymbol{W}_3\boldsymbol{W}_k\boldsymbol{h}_i+\boldsymbol{D}'\right) \tag{14}$$

式中：$\sigma(\cdot)$为tanh激活函数；N'_m为第m个节点的邻

域; $\boldsymbol{W}_k$和$\boldsymbol{D}'$为 k 核需要学习的参数; $\boldsymbol{h}_i$表示第 i 个节点对应的局部节点相似性。为稳定训练过程并丰富模型能力, 采用多头注意力机制增强上述 GCN, 计算公式为

$$\hat{\boldsymbol{h}}'_m = \Big\|_{p=1}^{P} \sigma\left(\sum_{h'_i \in N(h'_m)} \boldsymbol{W}^p \cdot \left(\alpha_{m,i}^p \boldsymbol{h}'_i\right)\right) \tag{15}$$

式中: P 为头注意力的个数; $N(\boldsymbol{h}'_m)$表示$\boldsymbol{h}'_m$的邻居集合; $\boldsymbol{W}^P$为头注意力个数为 P 的投影矩阵; $\alpha_{m,i}^p$为多头注意力系数; $\boldsymbol{h}'_i$为通过一层 GCN 应用 k 个内核进行局部对应关系传播与推理后第 e 个节点对应的节点相似度。通过带多头注意力机制的 GCN 给图文数据传播完局部相邻节点的对应关系后, 每个单词都会找到相关图像区域进行对应关系的建立, 然后沿文本内部传播与补全。传播完后将卷积向量送到感知机中, 用多层 MLP 计算。综合所有局部对应关系做推理, 逐步学习到全局对应关系。从文本到图像、从图像到文本的相似性计算公式为

$$\boldsymbol{S}_{j\to m} = \frac{1}{n}\sum_m \boldsymbol{W}_6^j\left(\tanh\left(\boldsymbol{W}_5^j \hat{\boldsymbol{h}}'_m + \boldsymbol{b}_5^j\right)\right) + \boldsymbol{b}_6^j \tag{16}$$

$$\boldsymbol{S}_{m\to j} = \frac{1}{c}\sum_e \boldsymbol{W}_6^m\left(\tanh\left(\boldsymbol{W}_5^m \hat{\boldsymbol{h}}'_e + \boldsymbol{b}_5^m\right)\right) + \boldsymbol{b}_6^m \tag{17}$$

式中: $\boldsymbol{W}_6^j$和$\boldsymbol{b}_6^j$为文本 MLP 参数; $\boldsymbol{W}_6^m$和$\boldsymbol{b}_6^m$为图像 MLP 参数; tanh(·)为激活函数。计算图文数据整体相似性时, 需要把句子到图像、图像到句子的全局相似性相加, 计算公式为

$$\boldsymbol{S}(\boldsymbol{J}', \boldsymbol{M}') = \boldsymbol{S}_{j\to m} + \boldsymbol{S}_{m\to j} \tag{18}$$

式中: $\boldsymbol{J}'$表示文本连接图; $\boldsymbol{M}'$表示图像连接图。通过联合考虑从文本到图像、从图像到文本的相似性来学习到全局相似性, 进而建立起全局对应关系, 实现图文关系融合。

在图文关系融合中采用节点切分方法计算余弦相似度进一步建模局部关系, 并采用带多头注意力机制的图卷积网络推理全局关系。为进一步提升语义关系挖掘效率, 采用基于注意力机制的连边权重剪枝策略和三元组损失函数进行关系验证。

4　关系验证

随着模型运行与数据增长, 无意义关系会影响到模型空间复杂度与时间复杂度甚至影响到模型运行效率。为此, 使用基于注意力机制的连边权重剪枝策略、三元组损失函数来增强重要关系信息, 抑制无效关系, 以保证挖掘关系的有效性。

4.1　基于注意力机制的连边权重剪枝策略

为增强重要分支展示, 去除无用信息, 加入基于注意力机制的连边权重剪枝策略。给相似度$\hat{\boldsymbol{h}}'_m$计算聚合权重α_g, 其公式为

$$\alpha_g = \frac{\delta\left(\mathrm{BN}\left(\boldsymbol{W}_7 \hat{\boldsymbol{h}}'_m\right)\right)}{\sum_{\hat{\boldsymbol{h}}'_i \in N(\hat{\boldsymbol{h}}'_m)} \delta\left(\mathrm{BN}\left(\boldsymbol{W}_7 \hat{\boldsymbol{h}}'_i\right)\right)} \tag{19}$$

式中: $N(\hat{\boldsymbol{h}}'_m)$为$\hat{\boldsymbol{h}}'_m$的邻居集合; $\delta(\cdot)$层表示激活函数; BN(·)层表示进行批量归一化处理; $\boldsymbol{W}_7 \in \mathbf{R}^{d\times 1}$为一个 d 维相似度向量的线性变换矩阵; $\hat{\boldsymbol{h}}'_i$为经过关系挖掘后的第 i 个节点对应的节点相似度, 为$\hat{\boldsymbol{h}}'_m$聚合相似性表示:

$$\boldsymbol{h}_H = \sum_{\hat{\boldsymbol{h}}'_m \in N(\hat{\boldsymbol{h}}'_i)} \alpha_g \hat{\boldsymbol{h}}'_m \tag{20}$$

将阈值设定为 0.75, 若相似度聚合权重乘以相似度大于等于 0.75, 则保留其连边, 小于则剪掉, 采用此策略剪掉相似度低的连边, 最终将$\boldsymbol{h}_H$送入完全连接层来预测图文之间最终相似性。基于注意力机制的连边权重剪枝策略算法的具体处理过程如下所示:

输入: 节点相似度$\hat{\boldsymbol{h}}'_m$、线性变换矩阵$\boldsymbol{W}_7$、相似度邻居集合$N(\cdot)$、阈值ε。

输出: 聚合相似度$\boldsymbol{h}_H$。

1: 初始化α_g, $\boldsymbol{h}_H$; //初始化聚合权重与聚合相似度。

2: for m from 1 to M do。

3:　利用式(19)更新α_g; //给经过关系挖掘后的第 m 个文本节点对应的节点相似度$\hat{\boldsymbol{h}}'_m$计算聚合权重。

4:　使用式(20)更新$\boldsymbol{h}_H$; //通过相似度聚合权重α_g乘以相似度$\hat{\boldsymbol{h}}'_m$来聚合相似性表示即计算出聚合相似度大小。

5: end for。

6: 根据$\boldsymbol{h}_H$具体大小来设置阈值ε, 此处设置为 0.75, 判断是剪枝还是保留连边。

7: if $\boldsymbol{h}_H$<ε then //若相似度聚合权重乘以相似度的结果$\boldsymbol{h}_H$小于阈值ε。

8:　剪掉连边。

9: else。

10:　保留相似度连边。

11: end if。

4.2　损失函数

本文方法使用三元组损失函数来衡量相似性关系, 使相匹配的多模态数据相似度比不匹配的相似度高γ。为了提高计算效率, 只进行小批量计算, 因此将损失函数定义为

$$\begin{aligned}\boldsymbol{L}(m,j) = &\max\left(\gamma - \boldsymbol{S}_{m\to j} + \boldsymbol{S}_{m\to \hat{j}}, 0\right) + \\ &\max\left(\gamma - \boldsymbol{S}_{m\to j} + \boldsymbol{S}_{\hat{m}\to j}, 0\right)\end{aligned} \tag{21}$$

式中: γ为设定的阈值。图像 m 和句子 j 为小批量

计算中的对应对，图像 m 和句子 $\hat{j}$、图像 $\hat{m}$ 和句子 j 为非对应对。其中 $\hat{j}=\mathrm{argmax}_{p\neq j}\boldsymbol{S}_{m\to p}$ 且 $\hat{m}=\mathrm{argmax}_{t\neq m}\boldsymbol{S}_{t\to j}$ 为小批量中最难的负样本。通过三元组损失函数可拉近同类样本间的距离，推远异类样本间距，进而优化能够产生最高损失的难样本，从而得到更精准的图文数据间关系应用于多模态学习与检索。

5 实验及结果分析

5.1 实验设置与评价指标

为验证本文方法，基于 Ubuntu 操作系统，使用 python、pytorch、cuda、gcc 等语言工具，搭建了实验平台，调用 Flickr30K[29]、MSCOCO[30] 数据集进行实验。

1）数据集。Flickr30K 包含从 Flickr 网站收集的 31 783 张图片，每张图像对应 5 种句子描述，MSCOCO 由 123 287 张图像组成，每张图像与 5 个文本描述相关联，将数据集拆分成训练集、验证集、测试集来评估模型，如表 1 所示。

表 1 数据集统计表

Table 1 Statistical table of dataset

数据集	划分	图像	文本
Flickr30K	训练集	29 783	148 915
	验证集	1 000	5 000
	测试集	1 000	5 000
MSCOCO-1K	训练集	121 287	606 435
	验证集	1 000	5 000
	测试集	1 000	5 000
MSCOCO-5K	训练集	113 287	566 435
	验证集	5 000	25 000
	测试集	5 000	25 000

2）评价指标。使用召回率 R_q 进行方法评价，表示在排名前 q 个样本的返回结果中，检索到符合事实的正确项占总测试项的比例，本文衡量的是 R_1,R_5,R_{10}，具体公式为

$$R_q=\frac{1}{Z}\sum_{x'=1}^{Z}\upsilon\left(y'_{x'},q\right) \tag{22}$$

$$\upsilon\left(y'_{x'},q\right)=\begin{cases}1 & y_{x'}\in\left\{p_1^{x'},p_2^{x'},\cdots,p_q^{x'}\right\}\\ 0 & y_{x'}\notin\left\{p_1^{x'},p_2^{x'},\cdots,p_q^{x'}\right\}\end{cases} \tag{23}$$

式中：Z 为检索实例数量。如果 $\upsilon\left(y'_{x'},q\right)=1$ 表明所查询的例子 $y'_{x'}$ 和得到的前 q 个候选结果有关系，结果正确；如果 $\upsilon\left(y'_{x'},q\right)=0$ 则表示所查询的例子 $y'_{x'}$ 不在前 q 个候选结果范围内，结果错误。

3）参数设置。将预训练 BERT-Large 模型学习率设为 0.000 05，将句子最大长度设为 90，隐藏层嵌入大小设为 778；提取每张图像的前 48 个区域，获得每个 2 048 维的区域特征，通过全连接层将其转换为 1 024 维特征表示。将 GCN 层数设置为 1，每层 GCN 都由 8 个注意头组成，每个注意头计算 12 个隐藏单元，设置 GCN 核个数等于 10，每个核都为 32 维。此外，将损失函数中的参数 γ 设置为 0.3。对于 Flickr30K 数据集，将 λ 设为 25，初始学习率设为 0.000 3，每 10 个周期降低 0.1；对于 MSCOCO 数据集，将 λ 设为 15，初始学习率设为 0.000 5，每 5 个周期降低 0.1。通过多次实验得出，上述参数设定可以在 3 个数据集的 2 个任务上得出最优的检索率。

5.2 结果分析

基于搭建的环境，使用对应的评价指标，对本文提出来的 MRAM 方法进行了有效性验证，并与 VSE++[31]、GXN[32]、SCO[33]、MMCA[34]、多层语义对齐[35] 等多种方法比较，进行实验结果对比分析，具体如表 2 所示。

本文方法旨在深层次提取文本、图像特征，充分挖掘图文数据内和数据间语义关系。将本文方法应用在多模态检索任务中，取得了较好的检索效果。从表 2 对比结果可看出，本文方法在 Flickr30K 数据集的句子检索任务下，R_1 与当前方法中最好的结果相比有较大提升，提升 2%，R_5 有小幅提升，提升 0.9%，在图像检索任务下 R_5 和 R_{10} 也有小幅提升，分别提升 1.1%、0.6%。而 MSCOCO-1K 数据集中，对于以文索图任务，其 R_1、R_5、R_{10} 分别提升了 1.2%、1.3%、0.3%，以图索文中任务中 R_1、R_5 分别提升了 1.2%、0.7%。相比之下，MSCOCO-5K 检索性能低于 MSCOCO-1K 测试集的结果，可能因为 MSCOCO-5K 是较广泛的目标集，包含更多数量的实体关系与噪声信息。与多种方法比较，图像到文本检索任务中 R_1、R_{10} 各提升了 1%、0.1%，文本到图像检索任务中 R_1、R_{10} 各提升了 0.9%、1.9%。虽然上述模型在某些项得分略低于其他模型，但本文方法成果验证了在多模态检索中考虑隐藏关系的必要性。

5.3 消融实验分析

为验证 MRAM 模型中关键部分的有效性，采用控制变量法，通过去掉模型中的部分模块来查看模型性能是否发生变化，如表 3 所示，划分文本关系挖掘、图像关系挖掘、图文关系融合、关系验证 4 个模块在 Flickr30K 和 MSCOCO-1K 数据集的 2 个任务上进行消融实验以及结果分析。

1）w/o 预训练的 BERT-Large 模型。评估了将预训练的 BERT-Large 引入文本编码层的影响，由于其可以捕捉到 Bi-GRU 网络易忽视的上下文关系信息，因此没有经过预训练 BERT-Large 编码的模型尤其图像检索方面损失约 1.88%。

表 2　实验结果分析

Table 2　Analysis of experimental results　　%

方法	R_1						R_5						R_{10}					
	以文索图			以图索文			以文索图			以图索文			以文索图			以图索文		
	Flickr-30K	MS-COCO-1K	MS-COCO-5K	Flickr-30K	MS-COCO-1K	MS-COCO-5K	Flickr-30K	MS-COCO-1K	MS-COCO-5K	Flickr-30K	MS-COCO-1K	MS-COCO-5K	Flickr-30K	MS-COCO-1K	MS-COCO-5K	Flickr-30K	MS-COCO-1K	MS-COCO-5K
m-CNN[2]	33.6	42.8		26.2	32.6		64.1	73.1		56.3	68.6		74.9	84.1		69.6	82.8	
VSE++[31]	52.9	64.6	41.3	39.6	52.0	30.3	80.5	90.0	71.1	70.1	84.3	59.4	87.2	95.7	81.2	79.5	92.0	72.4
GXN[32]		68.5	42.0		56.6	31.7								97.9	84.7		94.3	74.6
SCO[33]	55.5	69.9	42.8	41.1	56.7	33.1	82.0	92.9	72.3	70.5	87.5	62.9	89.3	97.5	83.0	80.1	94.8	75.5
SCAN[10]	67.4	72.7	50.4	48.6	58.8	38.6	90.3	94.8	82.2	77.7	88.4	69.3	95.8	98.4	90.0	85.2	94.8	80.4
VSRN[16]	70.4	74.0	53.0	53.0	60.8	40.5	89.2	94.3	81.1	77.9	88.4	70.6	93.7	97.8	89.4	85.7	94.1	81.1
CAMP[17]	68.1	72.3	50.1	51.5	58.5	39.0	89.7	94.8	82.1	77.1	87.9	68.9	95.2	98.3	89.7	85.3	95.0	80.2
SGM[21]	71.8	73.4	50.0	53.5	57.5	35.3	91.7	93.8	79.3	79.6	87.3	64.9	95.5	97.8	87.9	86.5	94.3	76.5
多层语义[35]	69.6	74.5		61.5	60.1		91.2	95.1		78.8	87.4		97.1	97.8		88.4	96.5	
CANN[20]	70.1	75.5	52.5	52.8	61.3	41.2	91.6	95.4	83.3	79.0	89.7	70.3	97.2	98.5	90.9	87.9	95.2	82.9
MMCA[34]	74.2	74.8	54.0	54.8	61.6	38.7	92.8	95.6	82.5	81.4	89.8	69.7	96.4	97.7	90.7	87.8	95.0	80.8
MRAM	76.2	76.7	55.0	59.7	62.8	42.1	93.7	96.9	83.2	82.5	90.5	72.5	97.1	98.8	91.0	89.0	96.3	82.9

表 3　消融实验表

Table 3　Table of ablation experiments　　%

方法	R_1				R_{10}			
	以文索图		以图索文		以文索图		以图索文	
	Flickr30K	MSCOCO-1K	Flickr30K	MSCOCO-1K	Flickr30K	MSCOCO-1K	Flickr30K	MSCOCO-1K
w/o预训练的BERT-Large模型	74.9	75.9	58.1	60.4	95.4	98.1	87.6	94.2
w/o基于多头注意力的空间关系	74.5	74.8	58.7	61.4	95.7	97.1	88.1	95.1
w/o全局对应关系	68.6	68.6	55.4	58.1	95.7	97.2	86.8	93.4
w/o局部对应关系中的节点切分方法	75.8	76.1	57.0	61.9	96.7	97.9	88.8	95.6
w/o全局对应关系中的多头注意力	75.2	75.2	58.7	61.3	96.3	97.5	88.1	94.9
全局对应关系w/2GCN	72.1	73.3	55.6	59.2	92.6	95.9	85.1	93.2
w/o基于注意力机制的连边权重剪枝策略	74.6	75.4	57.3	61.4	95.0	97.3	86.9	95.1
MRAM方法	76.2	76.7	59.7	62.8	97.1	98.8	89.0	96.3

2) w/o 基于多头注意力的空间关系。利用基于注意力的空间关系编码器来学到具有空间关系感知的区域级图像表示。去除带多头注意力的空间关系后, 文本检索图像和图像检索文本性能分别下降了 1.68% 和 1.13%, 说明通过参考学习到的空间位置关系, 可以更好地获得图像在语义和空间层次上目标间的多关系视觉特征, 进而弥补后期图像特征参与相似度计算时易丢失有效信息等问题, 有助于图像关系挖掘。

3) 在图文关系融合模块中, 证明了其每个组件的重要性。在 w/o 全局对应关系中试图只考虑局部对应关系挖掘, 然而性能显著下降, 因为局部对应关系部分只从小范围节点方面计算图文相似性, 而全局对应关系则进一步实现了整体范围的图文关系传播、推理与实现。w/o 全局对应关系中的节点切分方法。计算图文局部相似性时没有用节点切分方法, 模型性能虽然轻微下降, 但是通过该方法却可以优化相似度计算, 降低其运算过程中的时间复杂度, 推动整个模型实现, 方便后续操作。在 w/o 全局对应关系中的多头注意力中删掉全局对应关系中的多头注意力后, 检索性能下降了约 1.18%, 这证实了多头注意力机制可有效加权重要关系信息, 从而捕获隐藏在多模态数据间深层次的语义关联。w/o 全局对应关系 w/2GGN。尝试在全局对应关系中用 2 层 GCN, 结果表明更深的网络会降低性能, 主要原因可能是使用 2 层 GCN 后, 即使有些节

点不能构成局部对应，但其也会在间接相邻的节点之间互相传播关系，会被其他节点对应关系干扰，因此最后学习到的关系就不那么准确，检索效率随之下降。

4）w/o 基于注意力机制的连边权重剪枝策略。移除基于注意力机制的连边权重剪枝策略之后，模型效果也一定程度下降了，这是由于注意力机制能将注意力关注到最相关的地方，从而可以参考阈值有效剪枝，提高模型效率。

6 结 论

1）MRAM 方法分别提取特征构建文本连接图和图像连接图。局部阶段学习节点级别的小范围关系，在全局范围内传播局部关系到邻居节点，将实体、关系对应融合，推断出多模态数据间深层次的语义关联。

2）与当前方法在 Flickr30K、MSCOCO-1K 和 MSCOCO-5K 3 个数据集上进行实验结果对比，本文方法可达到较为优异的检索效果，在以文索图任务上平均召回率提升 0.76%，以图索文任务上平均召回率提升 0.71%。

3）实验结果表明本文方法有效挖掘到了关系，将其用于复杂场景下多模态检索中的效率也明显提升，但仍然存在模型训练过慢、参数过于复杂、参数量庞大等问题。

在未来的研究中，将考虑采用蒸馏方法来改善模型训练问题，同时考虑构建基于常识关系的多模态知识图谱，针对本文需要根据权重判断的关系，借鉴外部知识图谱中的信息来确定，进一步提升多模态语义关系挖掘的准确性和有效性。

参考文献（References）

[1] FROME A, CORRADO G S, SHLENS J, et al. Devise: A deep visual-semantic embedding model//Proceeding of the 26th International Conference on Advances in Neural Information Proceeding Systems. New York: ACM, 2013, 26: 2121-2129.

[2] MA L, LU Z D, SHANG L F, et al. Multimodal convolutional neural networks for matching image and sentence[C]//Proceedings of the IEEE International Conference on Computer Vision. Piscataway, NJ: IEEE Press, 2015: 2623-2631.

[3] WEHRMANN J, BARROS R C. Bidirectional retrieval made simple[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2018: 7718-7726.

[4] WANG T, XU X, YANG Y, et al. Matching images and text with multi-modal tensor fusion and re-ranking[C]//Proceedings of the 27th ACM International Conference on Multimedia. New York: ACM, 2019: 12-20.

[5] WANG P, WU Q, CAO J W, et al. Neighbourhood watch: Referring expression comprehension via language-guided graph attention networks[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2019: 1960-1968.

[6] MAFLA A, DEY S, BITEN A F, et al. Multi-modal reasoning graph for scene-text based fine-grained image classification and retrieval[C]//Proceedings of the IEEE Winter Conference on Applications of Computer Vision. Piscataway: IEEE Press, 2021: 4022-4032.

[7] LIU S, QIAN S S, GUAN Y, et al. Joint-modal distribution-based similarity hashing for large-scale unsupervised deep cross-modal retrieval[C]//Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM, 2020: 1379-1388.

[8] DONG X F, ZHANG H X, DONG X, et al. Iterative graph attention memory network for cross-modal retrieval[J]. Knowledge-Based Systems, 2021, 226: 107138.

[9] KARPATHY A, LI F F. Deep visual-semantic alignments for generating image descriptions[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2015: 3128-3137.

[10] LEE K H, CHEN X, HUA G, et al. Stacked cross attention for image-text matching[C]//Proceedings of the European Conference on Computer Vision. Berkin: Springer, 2018: 212-228.

[11] HUANG Y, WANG W, WANG L. Instance-aware image and sentence matching with selective multimodal LSTM[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2017: 7254-7262.

[12] YAO T, PAN Y W, LI Y H, et al. Exploring visual relationship for image captioning[C]//Proceedings of the European Conference on Computer Vision. Berlin: Springer, 2018: 711-727.

[13] HOU J Y, WU X X, QI Y Y, et al. Relational reasoning using prior knowledge for visual captioning [EB/OL]. (2019-06-04)[2022-08-20]. http://arxiv.org/abs/1906.01290.

[14] WANG Y X, YANG H, QIAN X M, et al. Position focused attention network for image-text matching [EB/OL]. (2019-07-23)[2022-08-20]. http://arxiv.org/abs/1907.09748.

[15] CHEN H, DING G G, LIU X D, et al. IMRAM: Iterative matching with recurrent attention memory for cross-modal image-text retrieval[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscatawa: IEEE Press, 2020: 12652-12660.

[16] LI K P, ZHANG Y L, Li K, et al. Visual semantic reasoning for image-text matching[C]//Proceedings of the IEEE International Conference on Computer Vision. Piscataway: IEEE Press, 2019: 4653-4661.

[17] WANG Z H, LIU X H, LI H S, et al. CAMP: Cross-modal adaptive message passing for text-image retrieval[C]//Proceedings of the IEEE International Conference on Computer Vision. Piscataway: IEEE Press, 2019: 5764-5773.

[18] YANG Z Q, QIN Z C, YU J, et al. Scene graph reasoning with prior visual relationship for visual question answering[EB/OL]. (2019-08-21)[2020-08-20]. http://arxiv.org/abs/1812.09681.

[19] SONG Y, SOLEYMANI M. Polysemous visual-semantic embedding for cross-modal retrieval[C]//Proceedings of the IEEE Confer-

ence on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2019: 1979-1988.

[20] ZHANG Q, LEI Z, ZHANG Z, et al. Context-aware attention network for image-text retrieval[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2020: 3533-3542.

[21] WANG S J, WANG R P, YAO Z W, et al. Cross-modal scene graph matching for relationship-aware image-text retrieval[C]//Proceedings of the IEEE Winter Conference on Applications of Computer Vision. Piscataway: IEEE Press, 2020: 1497-1506.

[22] ZHENG Z D, ZHENG L, GARRETT M, et al. Dual-path convolutional image-text embeddings with instance loss[J]. ACM Transactions on Multimedia Computing, Communications, and Applications, 2020, 16(2): 51.

[23] LI Y, ZHANG D, MU Y. Visual-semantic matching by exploring high-order attention and distraction[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2020: 12783-12792.

[24] DONG X Z, LONG C J, XU W J, et al. Dual graph convolutional networks with transformer and curriculum learning for image captioning[C]//Proceedings of the 29th ACM International Conference on Multimedia. New York: ACM, 2021: 2615-2624.

[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)[2022-08-20]. http://arxiv.org/abs/1810.04805.

[26] 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.

[27] KRISHNA R, ZHU Y K, GROTH O, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations[J]. International Journal of Computer Vision, 2017, 123(1): 32-73.

[28] LI L J, GAN Z, CHENG Y, et al. Relation-aware graph attention network for visual question answering[C]//Proceedings of the IEEE International Conference on Computer Vision. Piscataway: IEEE Press, 2019: 10312-10321.

[29] YOUNG P, LAI A, HODOSH M, et al. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions[J]. Transactions of the Association for Computational Linguistics, 2014, 2: 67-78.

[30] LIN T Y, MAIRE M, BELONGIE S, et al. Microsoft COCO: Common objects in context[C]//European Conference on Computer Vision. Berlin: Springer, 2014: 740-755.

[31] FAGHRI F, FLEET D J, KIROS J R, et al. Vse++: Improving visual-semantic embeddings with hard negatives[EB/OL]. (2017-07-18)[2022-08-20]. http://arxiv.org/abs/1707.05612.

[32] GU J X, CAI J F, JOTY S R, et al. Look, imagine and match: Improving textual-visual cross-modal retrieval with generative models [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2018: 7181-7189.

[33] HUANG Y, WU Q, SONG C F, et al. Learning semantic concepts and order for image and sentence matching[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2018: 6163-6171.

[34] WEI X, ZHANG T Z, LI Y, et al. Multi-modality cross attention network for image and sentence matching[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2020: 10938-10947.

[35] 杜锦丰, 王海荣, 李明亮, 等. 多层语义对齐的跨模态检索方法研究[J]. 郑州大学学报 (理学版), 2021, 53(4): 83-88.
DU J F, WANG H R, LI M L, et al. Research on cross-modal retrieval method based on multi-layer semantic alignment[J]. Journal of Zhengzhou University (Natural Science Edition), 2021, 53(4): 83-88(in Chinese).

Multilevel relation analysis and mining method of image-text

GUO Ruiping，WANG Hairong*，WANG Dong

(School of Computer Science and Engineering, North Minzu University, Yinchuan 750021, China)

Abstract：How to efficiently mine the hidden semantic association between multi-modal data is one of the key tasks of multi-modal knowledge extraction. In order to mine fine-grained relation between image and text, multilevel relation analysis and mining method of image-text（MRAM）was proposed. BERT-Large (bidirectional encoder representation from transformers-large) extracted text feature and constructed text connection graphs, while the Faster-RCNN network extracted image feature to learn spatial position relation and semantic relation, then constructed image connection graphs, so as to complete the calculation of single-modal internal semantic relation. The node segmentation method and graph convolutional network with multi-head attention (GCN-MA) fused local and global relation of text and image. To improve the efficiency of relation mining, edge weight pruning strategy based on the attention mechanism strengthened the representation of important branches, and reduced the interference of redundant information. The proposed method was tested on Flickr30K, MSCOCO-1K and MSCOCO-5K datasets, and was compared with 11 methods. The average recall rate on Flickr30K was increased by 0.97% and 0.57%, the average recall rate on MSCOCO-1K was increased by 0.93% and 0.63%, and the average recall rate on MSCOCO-5K was increased by 0.37% and 0.93%. Experimental results verify the effectiveness of the proposed method.

Keywords：relation mining；multilevel relation；attention mechanism；graph convolutional network；image-text data

Received：2022-07-09；**Accepted**：2022-08-14；**Published Online**：2022-10-18 11：00

URL：link.cnki.net/urlid/11.2625.V.20221017.1553.007

Foundation items：The Ningxia Natural Science Foundation Project (2023AAC03316); The Key Scientific Research Project of Colleges and Universities of Ningxia Hui Autononous Region Education Department (NYG2022051); The North Minzu University Research Project (2021XYZJK06)

* **Corresponding author**. E-mail：bmdwhr@163.com

《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 405,407 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

《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

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50卷 第2期　2024年2月

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

JOURNAL OF BEIJING UNIVERSITY OF AERONAUTICS AND ASTRONAUTICS (JBUAA)

(Monthly, Started in 1956)

Vol.50 No.2　February 2024

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

ISSN 1001-5965

中国标准连续出版物号：ISSN 1001-5965 / CN 11-2625/V　　国内定价：50.00元／期

