ISSN 1001-5965
CODEN BHHDE8

# 北京航空航天大學学报

# JOURNAL OF BEIJING UNIVERSITY OF AERONAUTICS AND ASTRONAUTICS

2022-08
Vol.48 No.8

北京航空航天大学学报

第 48 卷　第 8 期　（总第 354 期）　2022 年 8 月

# 目　　次



**期刊基本参数：** CN 11-2625/V * 1956 * m * A4 * 210 * zh * P * ￥50.00 * 400 * 23 * 2022-08

（**编　辑**　张　嵘　李艳霞　苏　磊　孙　芳　卞欢欢　王　茜）

JOURNAL OF BEIJING UNIVERSITY OF AERONAUTICS AND ASTRONAUTICS

Vol. 48 No. 8 (Sum 354) August 2022

# CONTENTS

http://bhxb.buaa.edu.cn jbuaa@buaa.edu.cn
**DOI**: 10.13700/j.bh.1001-5965.2021.0517

# 面向目标检测的双驱自适应遥感图像超分重建方法

成科扬[1,2,3,*], 荣兰[1], 蒋森林[1], 詹永照[1]

(1. 江苏大学 计算机科学与通信工程学院, 镇江 212013;
2. 江苏省大数据泛在感知与智能农业应用工程研究中心, 镇江 212013; 3. 镇江昭远智能科技有限公司, 镇江 212013)

**摘 要**: 现有光学遥感图像超分重建方法主要是生成视觉上令人满意的图像,并未考虑后续目标检测任务的特殊性,不能有效地应用到目标检测中。基于此,提出了面向目标检测的双驱动自适应多尺度光学遥感图像超分重建方法,将超分重建网络和目标检测网络结合起来,进行联合优化。针对光学遥感图像的特点设计了自适应多尺度遥感图像超分重建网络,集成选择性内核网络和自适应特征门控单元来特征提取和融合,重建出初步遥感图像。通过提出的双驱动模块,将特征先验驱动损失和任务驱动损失传到超分重建网络中,提高目标检测的性能。在 UCAS-AOD 和 NWPU VHR-10 数据集上进行实验,并与 5 种主流方法进行比较,所提方法的峰值信噪比和平均准确率相较于 FDSR 方法分别提高了 1.86 dB 和 3.73%。实验结果表明,所提方法和光学遥感图像目标检测结合可以取得更好的效果,综合性能更佳。

**关 键 词**: 遥感图像超分重建; 目标检测; 多尺度; 特征先验驱动; 任务驱动
**中图分类号**: V221+.3; TB553
**文献标志码**: A **文章编号**: 1001-5965(2022)08-1343-10

光学遥感图像超分辨率重建(简称超分重建)是对一幅或者多幅具有互补信息的低分辨率光学遥感图像进行处理,来获得高分辨率光学遥感图像的技术。光学遥感图像是遥感图像目标检测的数据支撑和应用基础,为监视地球表面提供了丰富的信息,在灾害监控、城市经济水平评估、资源勘探等领域具有广泛的应用[1]。因此,提高遥感图像的分辨率意义重大。光学遥感图像超分重建方法一般分为 2 类:①以人为中心的方法;②以机器为中心的方法。以人为中心的方法,常以峰值信噪比(peak signal-to-noise ratio, PSNR)、结构相似度(structural similarity, SSIM)作为评价指标,生成视觉上令人满意的图像,供人们观看和识别,通常该方法忽略了后续计算机视觉任务(如目标检测、分类)的特殊性,生成的图像可以"愚弄"人类感知,但不能"愚弄"机器感知。以机器为中心的方法把计算机视觉任务的执行结果作为优化指标,通过任务来评估方法的重建性能,考虑到光学遥感图像超分重建后续任务的特殊性,将超分重建任务视为预处理步骤,设计原则着重于学习特殊任务的分辨率不变性,来处理一张遥感图像中的多尺度目标,以便有利于更高级别的计算机视觉任务处理和分析。

在低分辨率光学遥感图像目标检测中,感兴

**收稿日期**: 2021-09-06; **录用日期**: 2021-09-17; **网络出版时间**: 2021-10-29 10:13
**网络出版地址**: kns.cnki.net/kcms/detail/11.2625.V.20211028.1656.002.html
**基金项目**: 国家自然科学基金(61972183); 江苏省科技项目(BE2022781); 镇江市"金山英才"高层次领军人才培养计划培养对象科研项目
* **通信作者**. E-mail: kycheng@ujs.edu.cn

**引用格式**: *成科扬, 荣兰, 蒋森林, 等. 面向目标检测的双驱自适应遥感图像超分重建方法[J]. 北京航空航天大学学报, 2022, 48(8): 1343-1352. CHENG K Y, RONG L, JIANG S L, et al. Double drive adaptive super-resolution reconstruction method of remote sensing images for object detection[J]. Journal of Beijing University of Aeronautics and Astronautics, 2022, 48(8): 1343-1352 (in Chinese).*

趣的对象相对于场景来说所占的像素值很小,因此,通常需要遥感图像超分重建技术来对光学遥感图像先进行预处理,再提高目标检测应用的性能。但是目前,很多遥感图像超分重建方法都是针对自然图像,没有考虑到光学遥感图像中目标的尺度多样性,重建出的光学遥感图像效果一般,而且大部分成熟的图像超分重建方法都是以人为中心的方法,只生成视觉上令人满意的图像,并没有考虑到目标检测任务的特殊性,同时遥感图像超分重建和目标检测框架之间都是独立优化,它们之间的相互作用没有得到充分探索,导致遥感图像超分重建技术没有更好地为目标检测任务服务。

为了解决上述问题,本文设计了一种灵活的、面向目标检测的双驱动自适应多尺度遥感图像超分重建方法,使超分重建方法更好地为目标检测任务服务,以获得更好的检测性能。

# 1 相关工作

## 1.1 光学遥感图像超分重建

光学遥感图像退化后,会丢失部分高频信息,因此用于目标检测的关键信息很可能会丢失,导致无法准确定位,光学遥感图像超分重建方法的关键就是重建出这些丢失的高频信息。对于光学遥感图像超分重建而言,一些针对一般图像设计的深度学习超分重建方法没有考虑到光学遥感图像的尺度多样性,重建效果不好。Lei 等[2]提出一种多分支结构来提取遥感图像的局部和全局信息,并取得了不错的重建效果。Jiang 等[3]提出一种深度蒸馏递归网络,在网络中添加了多个交互式的连接,可以有效共享从多个并行卷积层提取的特征。Xu 等[4]提出深度记忆连接网络,利用多个跳跃连接将光学遥感图像的细节和环境信息相结合。Gu 等[5]提出一种多尺度残差网络,在网络中引入挤压和激励模块,为光学遥感图像重建出精确的高频信息。Lu 等[6]采用大、中、小尺度的深度残差神经网络模拟不同大小的感受野,获取全局、上下文和局部信息,融合不同尺度的信息,重建出高分辨率的光学遥感图像。Wang 等[7]采用多个自适应多尺度特征提取模块、挤压和激励模块及自适应门控机制来进行特征提取和融合。但是上述方法均存在以下 3 点不足:①由于遥感图像涉及范围广,地物之间的尺度差异很大,上述方法所采用的都是固定多尺度的方法,很难有效提取各种不同尺度的特征,从而影响重建效果。②上述网络中,用于特征提取和融合的网络结构往往是固定的。由于图像退化和图像内容多样性等复杂因素,自适应特征信息提取和融合更有利于光学遥感图像的超分重建。③上述遥感图像超分重建方法的研究主要集中在以人为中心的方法上,为了获得较好的视觉,忽略了后续计算机视觉任务的具体需求。

## 1.2 光学遥感图像超分重建和目标检测

超分重建是一个成熟的研究课题,并且被广泛应用到目标检测领域,具体来说,将图像超分重建作为遥感图像目标检测的预处理阶段可以提高目标检测的性能,特别是对于目标较小的场景。文献[8]中提到超分重建是未来遥感图像目标检测的一个有价值的预处理步骤。Cao 等[9]使用高分辨率的航空图像和低分辨率的卫星图像,结合字典学习方法来增强车辆,并使用 SSD(single shot multi box detector)来检测,结果表明,当使用超分重建网络做为预处理步骤时,与原始低分辨率的图像相比,目标检测的性能有所提高。Haris 等[10]使用深度反向投影网络[11](deep back-projection networks, DBPN)来做图像超分重建,并使用 SSD 目标检测,设计了一个损失函数来优化超分重建网络以提高目标检测的性能,实验结果表明,该算法端到端训练,提高了目标检测的性能。文献[12]利用边缘增强型生成对抗网络,重建出图像遗漏的边缘信息,并且与目标检测网络结合起来,与先进的目标检测方法相比,具有更好的性能。文献[13]由于遥感图像中车辆的尺寸相对较小,缺乏足够的细节来区分车辆和相似的目标,采用生成对抗网络框架,以端到端的方式实现超分辨率重建卷积网络(super resolution convolutional network, SRCNN)和车辆检测同时进行,并且在训练过程中将检测损失反向传播到 SRCNN 便于检测,实验结果表明,该算法比先进的检测器性能更好。但是以上方法只是简单地将超分重建网络和目标检测网络结合起来,没有充分发挥超分重建网络的特殊性,本文进一步利用后续目标检测任务,将特征先验驱动模块和任务驱动模块结合起来,并端到端地训练整个网络,使超分重建网络生成目标检测网络更适合检测的超分图像,从而实现更好的目标检测性能。

# 2 双驱动自适应多尺度光学遥感图像超分重建网络

本文提出了一种面向目标检测的双驱动自适

应多尺度超分重建方法,主要包括自适应多尺度超分重建模块和双驱动模块,具体结构如图 1 所示。低分辨率遥感图像 $I_{LR}$ 首先通过为遥感图像专门设计的自适应多尺度超分重建模块,得到重建后的超分辨率图像 $I_{SR}$,该模块包含了自适应多尺度特征提取块,集成可选择的多尺度特征提取和特征门控单元,可以灵活地融合遥感图像的多尺度特征并增强目标特征。然后将超分辨率图像 $I_{SR}$ 和原始的高分辨率图像 $I_{HR}$ 送入特征先验驱动模块中进行特征对齐,并将特征先验驱动损失传入到超分重建网络中,以指导生成更适合目标检测的超分辨率遥感图像。考虑到后续目标任务的特殊性,将超分辨率后的光学遥感图像送入任务驱动模块中,即目标检测模块,并将任务驱动损失传递到超分重建网络,获得最终的遥感图像的检测结果。

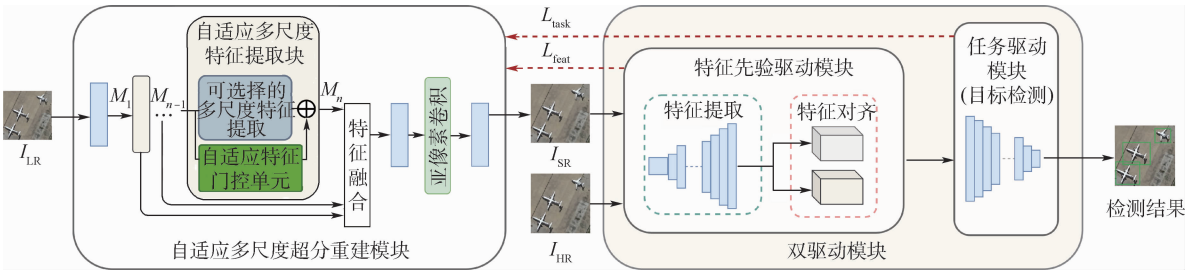



图 1　整体网络结构

Fig. 1　Overall network structure

## 2.1　自适应多尺度超分重建网络

针对光学遥感图像尺度多样性的特点,设计了自适应多尺度光学遥感图像超分重建网络(adaptive multi-scale neural network, AMNN)结构,主要分为 4 个部分:原始特征提取、自适应多尺度特征提取(adaptive multi-scale block, AMB)、特征融合和图像重建。其中,AMB 是整个模型的关键部分,包括可选择的多尺度特征提取(selective multi-scale feature extraction, SMFE)模块和自适应特征门控单元(adaptive feature gate unit, AFG)模块。

### 2.1.1　可选择的多尺度特征提取

常见的 GoogLeNet、Inception-ResNet 等常规固定多尺度特征提取网络,每层的尺度固定,感受野大小相同,这对于具有尺度多样性和尺度跨度较大的光学遥感图像来说是远远不够的。固定的多尺度很难有效提取光学遥感图像的特征,影响重建效果,并且会造成光学遥感图像处理过程中识别率降低和部分细节信息丢失等问题,从而影响遥感图像目标检测精度的提升。因此,受 SKNet[14] 的启发,针对光学遥感图像的特点,将常规的固定多尺度特征提取换成可选择的多尺度特征提取方法,有效、灵活地提取光学遥感图像的多尺度特征。

可选择的多尺度特征提取如图 2 所示。首先,对前一层 AMB 模块传递过来的特征图 $M_{i-1}$ 进行 3×3 卷积处理得到特征图 $M_i$。然后,对特征图 $M_i$ 进行 3 次不同尺度的卷积,获得不同尺度的特征图,本文选择的是卷积核大小分别为 1×1、3×3、5×5 的 3 个卷积。本文根据遥感图像的大小和遥感图像中目标物体的大小选择合适的尺度个数和卷积核的大小,以充分提取光学遥感图

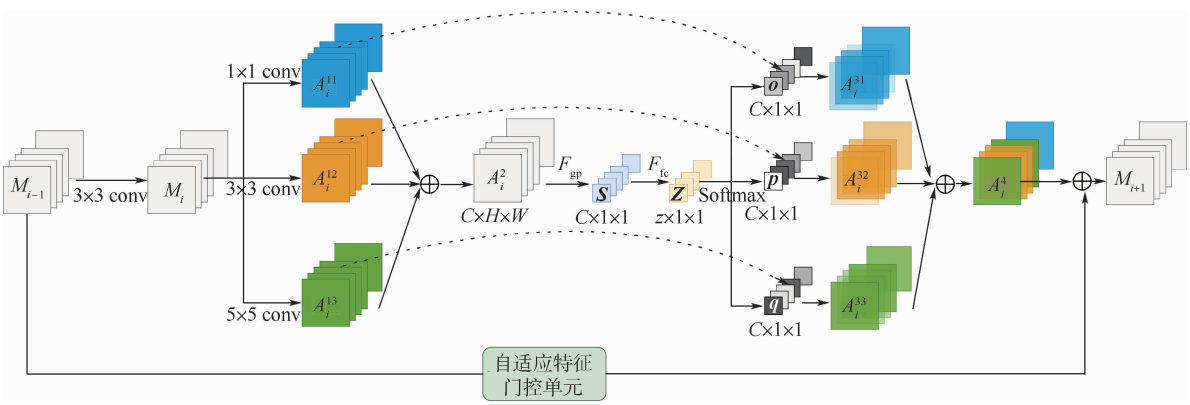



图 2　自适应多尺度特征提取块结构

Fig. 2　Adaptive multi-scale feature extraction block structure

像上的特征信息,而不需要重新设计多尺度特征提取网络。第 $j$ 层的特征图 $A_i^{1j}$ 定义为

$$A_i^{1j} = \varphi(w_i^{1j}A_i^0 + b_i^{1j}) \quad j = 1,2,3 \tag{1}$$

式中:$\varphi$ 表示卷积操作;$A_i^0$ 为第 $i$ 个 AMB 模块中的初始特征;$w_i^{1j}$ 为第 $j$ 个卷积核的权重;$b_i^{1j}$ 为第 $j$ 个卷积层的偏置。

为了多尺度的特征图信息能更好地流向下一层,提高特征信息的灵敏度,首先通过元素求和获得的多尺度特征图得到 $A_i^2$,其满足:

$$A_i^2 = A_i^{11} + A_i^{12} + A_i^{13} \tag{2}$$

然后使用全局平均池化得到每个通道的全局信息 $\boldsymbol{S} \in \mathbf{R}^c$。$\boldsymbol{S}$ 的第 $c$ 个元素 $\boldsymbol{S}_c$ 为

$$\boldsymbol{S}_c = F_{gp}(A_{ic}^2) = \frac{1}{H \times W}\sum_{m=1}^{H}\sum_{n=1}^{W}A_{ic}^2(m,n) \tag{3}$$

式中:$F_{gp}$ 表示全局平均池化操作;$A_{ic}^2$ 为 $A_i^2$ 特征图中的第 $c$ 个特征图;$H$ 为特征图的高度;$W$ 为特征图的宽度。

进一步,为了网络自适应的选择和提高计算效率,对 $\boldsymbol{S}$ 进行压缩得到 $\boldsymbol{Z}$:

$$\boldsymbol{Z} = \delta(B(\boldsymbol{K}_S)) \tag{4}$$

式中:$\delta$ 表示 ReLu 操作;$B$ 表示批量标准化操作;$\boldsymbol{K}_S$ 为每个通道全局信息 $\boldsymbol{S}$ 的向量表示,$\boldsymbol{K}_S \in \mathbf{R}^{d\times c}$,设置 $d$ 的值为 32。

采用软注意跨通道的方式来自适应地选择不同空间尺度的信息,也就是使用 Softmax 操作得到多尺度特征图 $A_i^{11}$、$A_i^{12}$、$A_i^{13}$ 的软注意向量 $\boldsymbol{o}$、$\boldsymbol{p}$、$\boldsymbol{q}$:

$$\boldsymbol{o}_c = \frac{e^{\boldsymbol{O}_c\boldsymbol{z}}}{e^{\boldsymbol{O}_c\boldsymbol{z}} + e^{\boldsymbol{P}_c\boldsymbol{z}} + e^{\boldsymbol{Q}_c\boldsymbol{z}}} \tag{5}$$

$$\boldsymbol{p}_c = \frac{e^{\boldsymbol{P}_c\boldsymbol{z}}}{e^{\boldsymbol{O}_c\boldsymbol{z}} + e^{\boldsymbol{P}_c\boldsymbol{z}} + e^{\boldsymbol{Q}_c\boldsymbol{z}}} \tag{6}$$

$$\boldsymbol{q}_c = \frac{e^{\boldsymbol{Q}_c\boldsymbol{z}}}{e^{\boldsymbol{O}_c\boldsymbol{z}} + e^{\boldsymbol{P}_c\boldsymbol{z}} + e^{\boldsymbol{Q}_c\boldsymbol{z}}} \tag{7}$$

式中:$\boldsymbol{O},\boldsymbol{P},\boldsymbol{Q} \in \mathbf{R}^{C\times d}$,$C$ 表示通道数;$\boldsymbol{z}$ 为上文压缩后的特征,$\boldsymbol{O}_c\boldsymbol{z} \in \mathbf{R}^{1\times d}$ 表示 $\boldsymbol{z}$ 的软注意力向量 $\boldsymbol{O}$ 的第 $c$ 行;$\boldsymbol{o}_c$ 为 $\boldsymbol{o}$ 的第 $c$ 个元素;$\boldsymbol{P}_c\boldsymbol{z}$,$\boldsymbol{Q}_c\boldsymbol{z}$,$\boldsymbol{p}_c$,$\boldsymbol{q}_c$ 同理,并且 $\boldsymbol{o}_c$,$\boldsymbol{p}_c$,$\boldsymbol{q}_c$ 满足 $\boldsymbol{o}_c + \boldsymbol{p}_c + \boldsymbol{q}_c = 1$。

最后将软注意向量 $\boldsymbol{o}$、$\boldsymbol{p}$、$\boldsymbol{q}$ 和特征图 $A_i^{11}$、$A_i^{12}$、$A_i^{13}$ 进行点乘分别得到特征特图 $A_i^{31}$、$A_i^{32}$、$A_i^{33}$,最终将这 3 个特征图融合了多个尺度的特征图 $A_i^4$。

### 2.1.2 自适应特征门控单元

在 2.1.1 节中已经提取了光学遥感图像的多尺度信息,但是这些多尺度的特征信息中包含了大量的冗余信息,如果直接将这些特征信息传递到下一层,会大大降低图像的重建效果,增加计算开销。因此,为了得到更好的重建效果和减少计算,需要在可选择的多尺度特征提取层之间添加自适应的特征门控单元,以此来适应遥感图像重建时复杂的非线性映射关系,减少冗余信息,使得网络更加灵活。因此,在特征传递的过程中采用一种简单的自适应门控机制,以解决特征传递过程中的冗余信息及增加网络的灵活性。自适应的特征门控单元如图 3 所示。

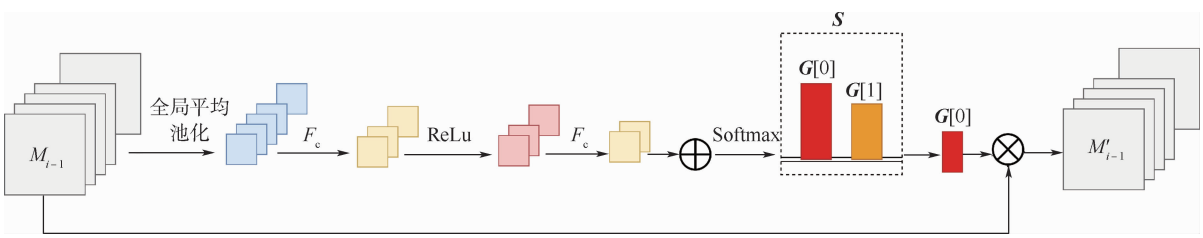



图 3 自适应特征门控单元结构

Fig. 3 Adaptive feature gating structure

自适应特征门控的关键在于自适应获得输入特征图 $M_{i-1}$ 的门控得分,当确定门控得分 $\mathrm{Score}(M_{i-1})$ 时,也就是确定需要保留多少比例的特征信息时,被保留下来的特征信息 $M'_{i-1}$ 为

$$M'_{i-1} = M_{i-1} \cdot \mathrm{Score}(M_{i-1}) \tag{8}$$

为了计算门控得分,首先使用全局平均池化操作对特征图进行降维操作,然后添加 2 个与 BN 连接的全连接作为简单非线性映射函数和 1 个 ReLu 函数来捕捉通道之间的依赖性,最后经过 Softmax 操作输出包含 2 个元素的向量 $\boldsymbol{G}$,$\boldsymbol{G}$ 的 2 个元素满足:

$$\boldsymbol{G}[0] + \boldsymbol{G}[1] = 1 \tag{9}$$

值较大的元素被记为特征图 $M_{i-1}$ 的门控得分。

## 2.2 双驱动模块

2.1 节针对光学遥感图像的尺度多样性的特点进行了自适应多尺度的超分重建,以提取光学遥感图像中的多尺度信息,减少部分信息丢失,但

是重建模型没有考虑到视觉任务目标检测的具体需求,重建出的图像并不能进行很好的检测,而且目标检测任务和超分重建模型之间都是独立优化的。光学遥感图像目标检测结果好坏很大程度依赖图像清晰及足够的纹理信息来提取特定的特征信息。因此,提出双驱动模块(double driven module, DDM),加入特征先验驱动(feature priority driven, FPD)和任务驱动(task driven, TD),减少超分辨率图像和真实高分辨率图像之间的特征差距,并将目标检测网络和超分重建网络结合起来联合训练,使超分重建模型更适合目标检测,为面向目标检测的遥感图像超分重建方法提供一种解决办法。

DDM 包含 FPD 和 TD,网络结构如图 1 中双驱动模块所示。为了减少超分辨率图像和真实高分辨率图像之间的特征差距,首先加入特征先验驱动,使用训练好的、带有 ResNet50-C4 的 Mask R-CNN[15]作为特征提取器 $F_{ext}$,因为 Mask R-CNN 引入了掩码反射,其主干提取的特征图比 Faster R-CNN(FR-CNN)[16]精细,而且没有和后续检测器之间过度耦合,有助于提高生成图像在其他检测网络中的可用性。特征对齐之后,将损失传递到超分重建网络,以约束超分重建图像的特征和真实图像的特征尽可能的相似。然后,发现虽然减少了超分辨率遥感图像和真实高分辨率图像之间的特征差距,但是特征先验驱动是一种依靠经验选择的结果,缺乏灵活性和适应性。因此,为了充分探索超分重建网络和目标检测网络之间的相互作用,取得更好的检测结果,同时加入任务驱动,将目标检测网络 FR-CNN 和自适应多尺度超分重建网络进行联合训练,显式地将任务驱动损失 $L_{task}$,也就是检测损失包含在自适应多尺度超分重建网络训练中。

### 2.3　损失函数

本文中,自适应多尺度超分重建损失 $L_{rec}$ 采用 $L_1$ 损失平均绝对误差,表达式为

$$L_{rec} = \frac{1}{N}\sum_{i=1}^{N} \left\| \mathrm{AMNN}(I_{LR}^{i}) - I_{HR}^{i} \right\| \tag{10}$$

式中:$N$ 为样本总数;$i$ 为训练集样本的索引;$I_{LR}^{i}$ 和 $I_{HR}^{i}$ 分别为第 $i$ 个低分辨率和高分辨率图像;$\mathrm{AMNN}(I_{LR})$ 为自适应多尺度超分重建模型 AMNN 重建出的图像。

用特征对齐损失 $L_{feat}$ 描述自适应多尺度重建超分辨率图像和原始图像之间的特征差异,特征对齐损失 $L_{feat}$ 使用均方误差来表示:

$$L_{feat} = \frac{1}{N}\sum_{i=1}^{N} \left\| (\mathrm{Fex}(\mathrm{AMNN}(I_{LR}^{i})) - \mathrm{Fex}(I_{HR}^{i})) \right\|_F \tag{11}$$

式中:$\mathrm{Fex}(\mathrm{AMNN}(I_{LR}^{i}))$ 为重建光学遥感图像的特征;$\mathrm{Fex}(I_{HR}^{i})$ 为原始高分辨率图像的特征。

任务驱动损失 $L_{task}$ 为

$$L_{task} = L_{cls} + \lambda L_{reg} \tag{12}$$

式中:

$$L_{cls} = E_{I_{LR}}[-\ln(\mathrm{Det}_{cls}(\mathrm{AMNN}(I_{LR})))] \tag{13}$$

$$L_{reg} = E_{I_{LR}}[\mathrm{smooth}_{L_1}(\mathrm{Det}_{reg}(\mathrm{AMNN}(I_{LR})), t_*)] \tag{14}$$

式中:$E_{I_{LR}}$ 表示训练数据来自原始低分辨率图像;$\mathrm{smooth}_{L_1}$ 表示平滑函数;$\lambda$ 用来平衡损失,本文设置为 1;$\mathrm{Det}_{cls}$ 和 $\mathrm{Det}_{reg}$ 分别为 FR-CNN 的分类和回归损失;$t_*$ 为真实的边界框坐标。

自适应多尺度重建损失可以有效保持图像的结构信息和纹理信息,特征先验驱动损失和任务驱动损失可以驱动超分重建网络更好地学习下游的任务。最终,设置框架的总体损失为

$$L_{overall} = L_{rec} + \lambda_1 L_{feat} + \lambda_2 L_{task} \tag{15}$$

式中:$\lambda_1$ 和 $\lambda_2$ 分别为特征先验驱动损失和任务驱动损失的相对强度的权重。

## 3　实验结果和分析

### 3.1　数据集

本文在 2 个广泛使用的 UCAS-AOD[17]和 NWPU VHR-10[18]数据集上进行评估,并与目前先进的方法比较。UCAS-AOD 数据集用于飞机和车辆检测,飞机数据集中包括 600 张图像和 3 210 架飞机,车辆数据集包括 310 张图像和2 819辆车辆,所有图像都经过精心挑选,并且图像种类少,目标分辨率低,适合对本文方法进行有效性验证。NWPU VHR-10 数据集包含 800 张高分辨率卫星图像,其中包含目标图像 650 张,背景图像 150 张,目标种类有 10 类,包括飞机、舰船、油罐、棒球场、网球场、篮球场、田径场、港口、桥梁和汽车,该数据集中目标的种类丰富,并且目标的尺度跨度大,对本文提出的自适应多尺度超分重建方法能做出很好的验证。

### 3.2　实验环境和参数设置

本文的实验环境为 24 GB NVIDIA TATIAN

GPUs,选用的深度学习框架为 Pytorch。以端到端的方式单独训练网络结构进行权重初始化。在单独训练时,先训练自适应多尺度超分重建网络至收敛,再训练 FPD 模块,最终训练目标检测网络。初始化权重后,对自适应多尺度超分重建模块、FPD 模块和目标检测模块进行联合训练,即将特征对齐损失和目标检测损失传递到超分重建网络中。

对原始高分辨率图像进行降采样,生成相应的低分辨率图像进行训练,并通过翻转和旋转变换来增强训练图像。在进行训练时,将学习率设置为0.000 1,每迭代 40 000 次,衰减率为 0.9,批处理大小设置为 5,使用 ADAM 作为优化器。通过实验,AMB 模块合适的值是 16 块。当 AMB 模块大于 16 时,收敛的速度开始放慢出现过拟合的情况,并且花费时间较长;当 AMB 模块小于 16 时,光学遥感图像的多尺度特征不能被充分提取。综合考虑选取 16 块 AMB 模块。该模型采用 $\lambda_1$ 和 $\lambda_2$ 来控制特征对齐损失和重建损失的贡献。通过实验,如图 4 所示,当 $\lambda_1$ 取 0.75,$\lambda_2$ 取 1 时,本文方法检测性能最好。由于该方法是面向目标检测的,选用峰值信噪比 PSNR 和平均检测精度 mAP 作为评价指标。图中:括号内数字表示 $\lambda_1$ 和 $\lambda_2$ 的取值。

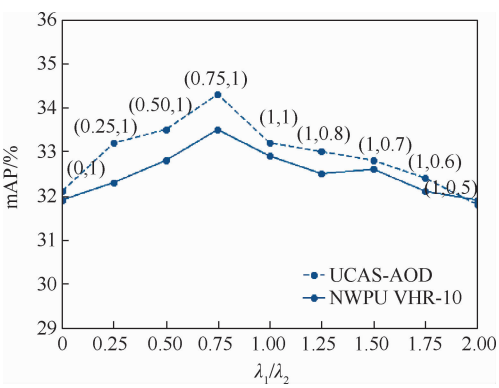



图 4 不同 $\lambda_1$ 和 $\lambda_2$ 下本文方法的检测性能

Fig. 4 Proposed method detects performance under different $\lambda_1$ and $\lambda_2$

## 3.3 实验分析

在 UCAS-AOD 和 NWPU VHR-10 数据集上分别测试了本文方法。为了客观分析本文方法的性能,在 UCAS-AOD 数据集上,对方法中不同的模块进行消融实验,更好地评估了不同模块对方法的影响。此外,还引入 5 种主流具有代表性的方法 Bicubicu[19]、MSRN[20]、AMFFN[7]、TDSR[10] 和 FDSR[21] 进行对比实验,验证了本文方法的有效性。

### 3.3.1 消融实验

为了验证本文提出的自适应多尺度超分重建模块及 DDM 的有效性,在 UCAS-AOD 数据集上进行消融实验,实验结果如表 1 所示。首先利用 FR-CNN 得到原始高分辨率图像和经过处理后的低分辨率图像的 mAP,分别为 70.43% 和 40.32%,可以看出分辨率对目标检测结果的影响巨大。当使用自适应多尺度超分重建模块 AMNN 先预处理低分辨率图像,再送入目标检测网络 FR-CNN 中检测,没有加入 DDM,没有进行联合优化,可以看到检测的精度比没处理的低分辨率图像高出 15.02%,这说明了 AMNN 模块的有效性,可以大大提高图像的分辨率,同时提高检测精度。当加入 TD 进行联合优化时,可以发现 PSNR 提升了 0.31 dB,mAP 提升了 9.99%,这说明了 TD 对超分重建和目标检测性能的有效影响,同时也说明了联合优化带来的好处。最后加入 FPD 将2 个驱动和 AMNN 结合起来联合训练,可以发现该方法大大提高了低分辨率图像的检测精度,并且非常接近原始高分辨率图像的检测性能,仅仅只差了1.30%,这有力地说明了 DDM 的有效性。

**表 1 消融实验**

**Table 1 Ablation test**

| 实验 | PSRN/dB | mAP/% |
|---|---|---|
| FR-CNN(HR) | | 70.43 |
| FR-CNN(LR) | | 40.32 |
| AMNN + FR-CNN (无联合训练) | 27.75 | 55.34 |
| AMNN + FR-CNN + TD (联合训练) | 28.06 | 65.33 |
| AMNN + FR-CNN + TD + FPD (联合训练) | 28.79 | 69.13 |

### 3.3.2 对比实验

对比实验选择了几种主流的具有代表性的超分重建方法并且将图像放大 2 倍进行对比。这些超分重建后的图像的检测性能在 UCAS-AOD 数据集飞机类上进行测试,因为飞机的尺度跨度较大,适合进行有效性验证。对比方法选用的检测网络与本文方法相同,均采用带有 FPN 的 FR-CNN 网络。表 2 给出了不同方法重建出的光学遥感图像的 PSNR 值和图像检测性能 AP 的结果,$AP_S$、$AP_M$、$AP_L$ 分别代表小、中、大尺度目标的检测性能。如表 2 所示,当 2 倍下采样时,AP 由 47.6%

**表 2　不同方法在 UCAS-AOD 数据集飞机类上的实验效果比较**

**Table 2　Experimental results compared with different methods on UCAS-AOD dataset**

| 方法 | PSNR/dB | AP/% | $AP_{0.5}$/% | $AP_{0.75}$/% | $AP_S$/% | $AP_M$/% | $AP_L$/% |
|---|---|---|---|---|---|---|---|
| 原始图像 | | 47.6 | 59.2 | 41.1 | 21.5 | 48.5 | 58.7 |
| Bicubicu[19] | 25.89 | 22.14 | 37.9 | 23.01 | 6.71 | 23.46 | 38.15 |
| MSRN[20] | 28.15 | 23.45 | 44.5 | 24.78 | 8.83 | 25.67 | 43.63 |
| TDSR[10] | 27.34 | 26.34 | 46.98 | 27.78 | 9.04 | 28.84 | 45.96 |
| AMFFN[7] | 28.56 | 24.78 | 44.54 | 25.01 | 8.92 | 25.65 | 43.98 |
| FDSR[21] | 27.56 | 26.33 | 46.78 | 27.66 | 9.15 | 29.03 | 45.97 |
| 本文方法 | **28.97** | **44.89** | **55.02** | **37.32** | **20.3** | **38.45** | **57.87** |

注:黑体数据表示最优结果。

降至 22.14%,可以看出超分重建网络的性能对目标检测网络 FR-CNN 的检测结果有很大影响,其中小尺度和中尺度目标受影响较大,$AP_S$ 由 21.5% 降到 6.71%,$AP_M$ 由 48.5% 降到 23.46%,根据分析,这是由于多尺度信息的丢失和下游目标检测任务的限制造成的。采用自适应多尺度超分重建模块和 DDM,不仅可以重建出图像的多尺度信息,还可以显著提高遥感图像目标检测的性能,该方法 AP 值达到了44.89%,与原始高分辨率图像的检测结果仅仅差了 2.71%,这说明了该方法的有效性,其中,小尺度目标检测效果提升的更明显,$AP_S$ 由 2 倍下采样的 6.71% 提高至 20.3%。

由表 3 和表 4 可知,MSRN 和 AMFFN 都是采用多尺度来重建超分辨率遥感图像的方法,但是其多尺度网络都是固定的,不能灵活提取光学遥感图像的多尺度信息,而且没有考虑到后续目标检测任务的特殊性,因此无论是重建效果还是目标检测效果都存在严重不足,在 UCAS-AOD 和 NWPU VHR-10 数据集上,本文方法与之相比,PSRN 分别平均提升 1.38 dB 和 1.67 dB,mAP 分别平均提升了 10.67% 和10.3%,这说明了自适应多尺度超分重建模块和 DDM 的有效性。TDSR 利用检测网络的损失,对超分重建网络 D-DBPN 进行优化,提高了目标检测的性能,但是其网络结构很深,可能会导致梯度消失等问题。FDSR 则仅仅利用特征提取器将原始图像特征和重建图像特征进行对齐,再将对齐损失传递到 D-DBPN 网络,该方法的局限性也很大。上述 2 个方法在检测精度上比常规的超分重建方法高,但是未考虑光学遥感图像的特点,重建效果一般。在 UCAS-AOD 和 NWPU VHR-10 数据集上,TDSR 和 FDSR 方法的 mAP 分别为 62.96% 和63.91%,但 PSNR 只有 26.62 dB 和26.80 dB。考虑到这 2 个方法的

**表 3　不同方法在 UCAS-AOD 数据集上的实验效果比较**

**Table 3　Experimental results compared with different methods on UCAS-AOD dataset**

| 方法 | PSNR/dB | mAP/% |
|---|---|---|
| Bicubicu[19] | 25.95 | 48.85 |
| MSRN[20] | 27.55 | 58.97 |
| AMFFN[7] | 27.56 | 59.23 |
| TDSR[10] | 26.43 | 62.96 |
| FDSR[21] | 26.65 | 63.98 |
| 本文方法 | **28.75** | **69.67** |

注:黑体数据表示最优结果。

**表 4　不同方法在 NWPU VHR-10 数据集上的实验效果比较**

**Table 4　Experimental results compared with different methods on NWPU VHR-10 dataset**

| 方法 | PSNR/dB | mAP/% |
|---|---|---|
| Bicubicu[19] | 24.86 | 47.56 |
| MSRN[20] | 27.01 | 57.98 |
| AMFFN[7] | 27.12 | 58.32 |
| TDSR[10] | 26.82 | 62.97 |
| FDSR[21] | 26.96 | 63.84 |
| 本文方法 | **28.58** | **68.61** |

注:黑体数据表示最优结果。

优缺点,引入 DDM,结合 FPD 和 TD,在 UCAS-AOD 数据集和 NWPU VHR-10 数据集上,PSNR 分别达到了 28.75 dB 和 28.58 dB,比TDSR和 FDSR 平均高了 2 dB,而目标 mAP 提升的更明显,达到了 69.67% 和68.61%,这说明该方法在重建效果和检测精度上都有了较大的提升。

为了更好地验证本文方法的优越性,挑选了 UCAS-AOD 和 NWPU VHR-10 数据集上具有代表性的检测结果进行可视化展示,具体如图 5 所示,实线方框表示正确的检测结果,虚线方框表示漏检,点虚线方框表示错误的检测结果。从图中可以看出,其他方法在不同程度上存在错检和漏检的情况,而本文方法检测结果良好。

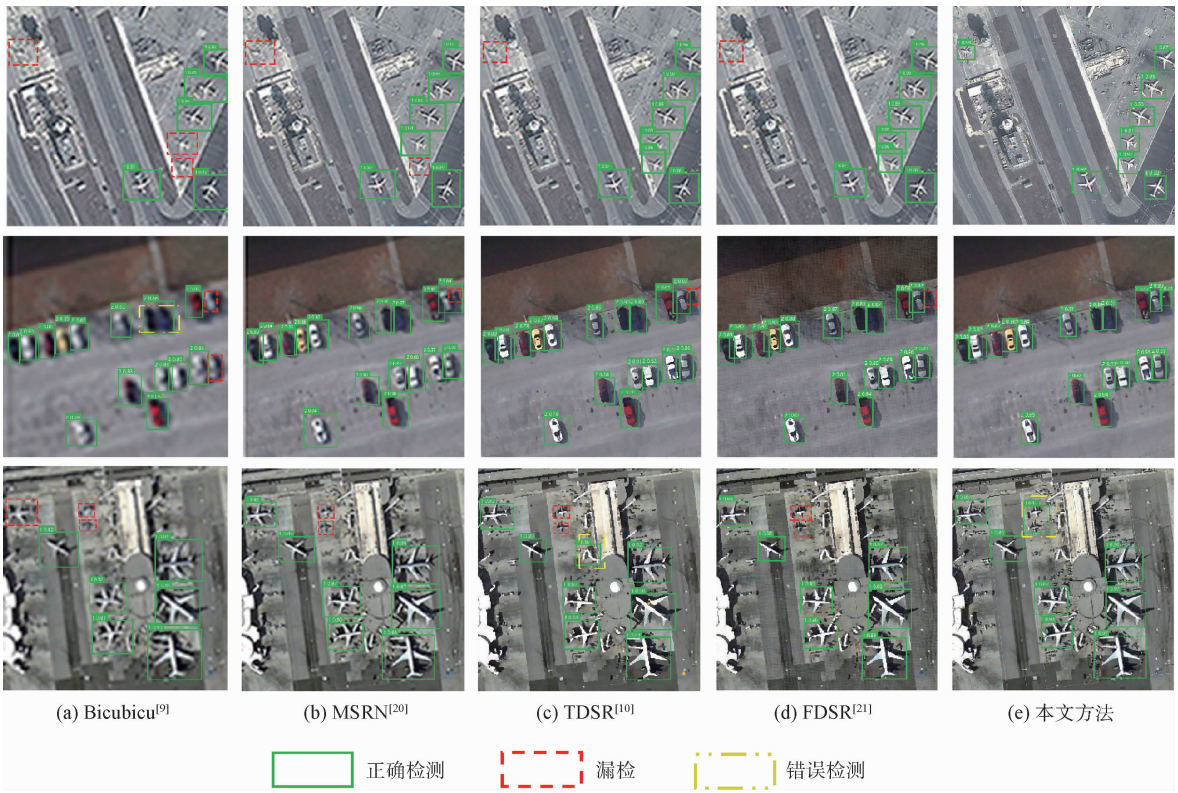



图 5 检测效果对比示例

Fig. 5 Example of detection effect comparison

综上所述,本文方法的综合性能最好,在尺度多样性的光学遥感图像上,不仅有较好的重建效果,而且在检测精度上也有很大的提升。

## 4 结 论

针对光学遥感图像超分重建在目标检测中的应用问题,根据光学遥感图像的多尺度特点和后续目标检测任务的特殊性,提出了一个面向目标检测的双驱动自适应多尺度的光学遥感图像超分重建方法,该方法的优势有 2 个:

1) 构建了自适应多尺度光学遥感图像超分重建网络,提高了遥感图像目标的特征表达能力。

2) 设计了双驱动模块,提高了光学遥感图像超分重建网络在目标检测上的应用效果。

本文主要工作有以下 3 个方面:

1) 提出一种面向目标检测的双驱动自适应多尺度遥感图像超分重建方法,将遥感图像超分重建任务和目标检测任务结合起来,做联合优化。

2) 针对光学遥感图像目标检测任务的特殊性,提出双驱动模块,一方面约束超分重建遥感图像的特征和真实图像的特征尽可能相似,另一方面考虑到遥感图像目标检测任务的特殊性,使超分网络更好地为目标检测任务服务,提高目标检测的性能。

3) 针对光学遥感图像的特点设计了自适应多尺度遥感图像超分重建网络,集成选择性内核网络和自适应特征门控单元来进行特征提取和融合,可以灵活地适应光学遥感图像的多尺度特征,增强目标特征,减少多尺度特征中的冗余信息,提高遥感图像的重建效果。

在未来的研究中,将进一步研究遥感图像超分在遥感图像目标检测中的应用,尤其是小目标的检测应用问题,希望可以得到更好的检测性能。

## 参考文献 (References)

[1] TSAGKATAKIS G, AIDINI A, FOTIADOU K, et al. Survey of deep-learning approaches for remote sensing observation enhancement[J]. Sensors, 2019, 19(18): 3929.

[2] LEI S, SHI Z W, ZOU Z X. Super-resolution for remote sensing images via local-global combined network[J]. IEEE Geoscience and Remote Sensing Letters, 2017, 14(8): 1243-1247.

[3] JIANG K, WANG Z, YI P, et al. Deep distillation recursive network for remote sensing imagery super-resolution[J]. Remote Sensing, 2018, 10(11): 1700.

[4] XU W J, XU G L, WANG Y, et al. High quality remote sensing image super-resolution using deep memory connected network [C]// IEEE International Geoscience and Remote Sensing Symposium. Piscataway: IEEE Press, 2018: 8889-8892.

[5] GU J, SUN X, ZHANG Y, et al. Deep residual squeeze and excitation network for remote sensing image super-resolution[J]. Remote Sensing, 2019, 11(15): 1817.

[6] LU T, WANG J M, ZHANG Y D, et al. Satellite image super-resolution via multi-scale residual deep neural network[J]. Remote Sensing, 2019, 11(13): 1588.

[7] WANG X Y, WU Y D, MING Y, et al. Remote sensing imagery super resolution based on adaptive multi-scale feature fusion network[J]. Sensors, 2020, 20(4): 1142.

[8] KOESTER E, SAHIN C S. A comparison of super-resolution and nearest neighbors interpolation applied to object detection on satellite data[EB/OL]. (2019-07-08)[2021-09-01]. https://arxiv.org/abs/1907.05283.

[9] CAO L J, WANG C, LI J. Vehicle detection from highway satellite images via transfer learning[J]. Information Sciences, 2016, 366: 177-187.

[10] HARIS M, SHAKHNAROVICH G, UKITA N. Task-driven super resolution: Object detection in low-resolution images[C]// International Conference on Neural Information Processing. Berlin: Springer, 2021: 387-395.

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

[12] RABBI J, RAY N, SCHUBERT M, et al. Small-object detection in remote sensing images with end-to-end edge-enhanced GAN and object detector network[EB/OL]. (2020-04-28)[2021-09-01]. https://arxiv.org/abs/2003.09085.

[13] JI H, GAO Z, MEI T C, et al. Vehicle detection in remote sensing images leveraging on simultaneous super-resolution[J]. IEEE Geoscience and Remote Sensing Letters, 2020, 17(4): 676-680.

[14] 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, 2019: 510-519.

[15] HE K, GKIOXARI G, DOLLÁR P, et al. Mask R-CNN[C]// Proceedings of the IEEE International Conference on Computer Vision. Piscataway: IEEE Press, 2017: 2961-2969.

[16] REN S Q, HE K M, GIRSHICK R, et al. Faster R-CNN: Towards real-time object detection with region proposal networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(6): 1137-1149.

[17] ZHU H G, CHEN X G, DAI W Q, et al. Orientation robust object detection in aerial images using deep convolutional neural network[C]// 2015 IEEE International Conference on Image Processing. Piscataway: IEEE Press, 2015: 3735-3739.

[18] CHENG G, ZHOU P C, HAN J W. Learning rotation-invariant convolutional neural networks for object detection in VHR optical remote sensing images[J]. IEEE Transactions on Geoscience and Remote Sensing, 2016, 54(12): 7405-7415.

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

[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 (ECCV). Berlin: Springer, 2018: 517-532.

[21] WANG B, LU T, ZHANG Y D. Feature-driven super-resolution for object detection[C]// 2020 5th International Conference on Control, Robotics and Cybernetics (CRC). Piscataway: IEEE Press, 2020: 211-215.

# Double drive adaptive super-resolution reconstruction method of remote sensing images for object detection

CHENG Keyang[1,2,3,*], RONG Lan[1], JIANG Senlin[1], ZHAN Yongzhao[1]

(1. School of Computer Science and Communication Engineering, Jiangsu University, Zhenjiang 212013, China;
2. Jiangsu Province Big Data Ubiquitous Perception and Intelligent Agricultural Application Engineering Research Center, Zhenjiang 212013, China; 3. Zhenjiang Zhaoyuan Intelligent Technology Co., Ltd., Zhenjiang 212013, China)

**Abstract**: The existing optical remote sensing image super-resolution reconstruction method is mainly to generate visually satisfactory images, and does not take into account the particularity of the subsequent target detection task, so it cannot be effectively applied to target detection. Therefore, a double drive adaptive multi-scale optical remote sensing image super-resolution reconstruction method for target detection is proposed. The super-resolution reconstruction network and target detection network are combined for joint optimization. According to the characteristics of optical remote sensing image, an adaptive multi-scale remote sensing image super-partition reconstruction network is designed. The selective kernel network and adaptive gating unit are integrated to extract and fuse features, and the primary remote sensing image is reconstructed. Through the double drive module, and task driven will feature a priori driver loses to the above points in the network, on the one hand, improve the performance of target detection. The proposed method was tested on UCAS-AOD and NWPU VHR-10 datasets, and compared with the five mainstream algorithms, peak signal-to-noise ratio and average accuracy improved by 1.86 dB and 3.73%, respectively, compared with the FDSR algorithm. Experimental results show that compared with other methods, the combination of the proposed algorithm and optical remote sensing image target detection can achieve better results and the comprehensive performance is the best.

**Keywords**: super-resolution reconstruction of remote sensing images; object detection; multi-scale; features priori driven; task driven

**Received**: 2021-09-06; **Accepted**: 2021-09-17; **Published online**: 2021-10-29 10:13

**URL**: kns.cnki.net/kcms/detail/11.2625.V.20211028.1656.002.html

**Foundation items**: National Natural Science Foundation of China (61972183); Jiangsu Science and Technology Project (BE2022781); Zhenjiang Jinshan High-level Leading Talents Training Plan Scientific Research Project

* **Corresponding author**. E-mail: kycheng@ujs.edu.cn

http://bhxb.buaa.edu.cn jbuaa@buaa.edu.cn
**DOI**: 10.13700/j.bh.1001-5965.2021.0529

# 基于深度强化学习与扩展卡尔曼滤波相结合的交通信号灯配时方法

吴兰[1,*], 吴元明[1], 孔凡士[2], 李斌全[1]

(1. 河南工业大学 电气工程学院, 郑州 450001; 2. 郑州铁路职业技术学院 电气工程学院, 郑州 450001)

**摘 要**: 深度Q学习网络(DQN)因具有强大的感知能力和决策能力而成为解决交通信号灯配时问题的有效方法,然而外部环境扰动和内部参数波动等原因导致的参数不确定性问题限制了其在交通信号灯配时系统领域的进一步发展。基于此,提出了一种DQN与扩展卡尔曼滤波(EKF)相结合(DQN-EKF)的交通信号灯配时方法。以估计网络的不确定性参数值作为状态变量,包含不确定性参数的目标网络值作为观测变量,结合过程噪声、包含不确定性参数的估计网络值和系统观测噪声构造EKF系统方程,通过EKF的迭代更新求解,得到DQN模型中的最优真实参数估计值,解决DQN模型中的参数不确定性问题。实验结果表明:DQN-EKF配时方法适用于不同的交通环境,并能够有效提高车辆的通行效率。

**关 键 词**: 深度Q学习网络(DQN); 感知能力; 决策能力; 交通信号灯配时系统; 参数不确定性; 扩展卡尔曼滤波(EKF)

**中图分类号**: V221$^+$.3; TB553

**文献标志码**: A **文章编号**: 1001-5965(2022)08-1353-11

随着社会的快速发展和人们物质生活水平的提高,日益增长的汽车数量会造成交通拥堵问题。交通拥堵会带来很多不必要的资源损失和环境污染,并且会增加交通事故发生的概率。现有的交通信号灯配时方法有2种,即定时算法和最长队列优先配时算法。定时算法在高度动态的交通环境下,对于缓解交通拥堵效率低下;最长队列优先配时算法根据车辆队列长度调整交通信号灯的配时方案,仍然不能解决交叉口的拥堵问题。人工指挥交通有时是有效的,但浪费人力和时间。因此,需要交通信号灯能够学会与环境互动得到合理的配时方案。

随着通信技术和计算机技术的发展,交通信号灯可根据当前的交通拥堵情况合理更换配时方案。早期的许多研究者提出了各种各样的自适应交通信号灯配时系统,如SCOOT(split cycle offset optimizing technique)和SCATS(Sydney coordinated adaptive traffic system)等。Robertson和Bretherton[1]介绍了SCOOT中的在线车流量模型和实时信号优化器的关键技术,证明了SCOOT在城市交通信号灯配时系统中的有效性。Lowrie[2]阐述了SCATS配时系统的主要原理和算法。这些交通配时系统已经在世界上数百个城市得到应用。还有一些研究主要集中在模糊神经网络和遗传算法来解决交通信号灯的配时问题。王史春[3]提出了一种基于模糊神经网络的交通信号灯配时系统,能根据实时交通状况调整交通信号灯的配时方案。胡智鹏[4]将遗传算法应用于交通信号灯

**收稿日期**: 2021-09-06; **录用日期**: 2021-09-17; **网络出版时间**: 2021-10-12 13:38
**网络出版地址**: kns.cnki.net/kcms/detail/11.2625.V.20211012.1106.001.html
**基金项目**: 国家自然科学基金(61973103); 河南省软科学研究计划(212400410005)
* **通信作者**. E-mail: wulan@haut.edu.cn

**引用格式**: *吴兰, 吴元明, 孔凡士, 等. 基于深度强化学习与扩展卡尔曼滤波相结合的交通信号灯配时方法[J]. 北京航空航天大学学报, 2022, 48(8): 1353-1363. WU L, WU Y M, KONG F S, et al. Traffic signal timing method based on deep reinforcement learning and extended Kalman filter[J]. Journal of Beijing University of Aeronautics and Astronautics, 2022, 48(8): 1353-1363 (in Chinese).*

的配时系统中,使各车道滞留汽车数量大量减少。但是模糊神经网络和遗传算法都需要大量计算,并存在收敛问题,已经不适用于复杂的交通环境。对于现代的交通环境,强化学习算法[5-7]可以更有效地解决交通拥堵问题。根据强化学习的算法性质可将其分为基于概率策略的算法和基于价值函数的算法。基于概率策略的算法主要包括 Policy gradient 和 Actor-Critic 等。Garg 等[8]提出了一种基于 Policy gradient 的强化学习算法来调整交通信号灯的配时方案。Genders 等[9]采用 Actor-Critic 自适应算法,并利用 2 个动作空间的神经网络函数设计了交通信号灯配时系统。但是基于概率策略的算法容易陷入局部最优,因此无法在复杂的交通环境中应用。基于价值函数的算法主要包括 Q-learning 和 SARSA 等。Zhou 等[10]提出了基于 Sarsa(λ)的交通信号灯配时模型,该模型考虑了延误时间和等待车辆的数量,并从其综合经验中学习得到最优的配时方案。Yin 等[11]在 Q-learning 算法的基础上综合考虑了相位差和拥堵因子的影响,从而调整交通信号灯的配时方案。然而,在 Q-learning 和 SARSA 算法中,所有的价值都存储在表格 Q-table 中,当状态-动作空间较大时,表格就会占用非常大的计算机内存,同时对表格中数值的提取也非常困难,导致 Q-learning 和 SARSA 算法无法适应高度动态的交通环境。随着深度学习[12-14]的快速发展,许多研究者将深度强化学习算法[15-16]应用到交通信号灯配时系统中。Tan 等[17]提出了将合作式深度 Q 学习网络(deep Q-learning network,DQN)应用于交通信号灯配时系统中,该模型将困难的任务分解为若干相对简单的子任务,再将各子任务的结果分层聚合得到价值函数,最终根据价值函数获得交通信号灯的配时方案。Zeng 等[18]提出了将 DQN 与递归神经网络相结合的算法,利用经验回放对智能体进行训练,生成交通信号灯的配时方案。Liao 等[19]提出了一种基于时差惩罚的 DQN 应用到交通信号灯配时系统中,以保证交通控制系统的安全性和吞吐能力。Liang 等[20]提出了将 Dueling DQN 和 Double DQN 应用于交通信号灯配时领域中,Dueling DQN 和 Double DQN 都是在值函数的结果上改进,以此解决 DQN 中存在的过估计和不准确问题。DQN 在交通信号灯配时领域取得了广泛应用,但是外部环境的干扰或内部参数的波动都将导致模型中出现参数不确定性问题,限制了 DQN 在交通信号灯配时系统领域的进一步发展。针对上述问题,本文提出了一种基于 DQN 与扩展卡尔曼滤波(extended Kalman filter,EKF)[21]相结合(DQN-EKF)的交通信号灯配时系统。该方法的主要思想是:以估计网络的不确定性参数值作为状态变量并与过程噪声共同构造系统状态方程;以包含不确定性参数的目标网络值作为观测变量,并与观测噪声和包含不确定性参数的估计网络值共同构造系统观测方程;通过 EKF 的迭代更新求解,得到 DQN 模型中的最优真实参数估计值,以解决 DQN 模型中的参数不确定性问题。在仿真实验中模拟了不同交通仿真环境,并通过对比实验证明了 DQN-EKF 配时方法的有效性。

## 1 相关工作

### 1.1 深度强化学习

强化学习又称再励学习、评价学习或增强学习,在智能控制和预测分析等领域有着广泛应用。强化学习可以看作试探评价的学习过程。首先,智能体选择动作用于环境,环境接受该动作后状态发生变化,同时产生一个强化信号(奖励或惩罚)反馈给智能体;然后,智能体根据强化信号和当前环境的状态选择下一个动作,选择动作的原则是使受到奖励强化信号的概率增大。虽然传统的强化学习具有优秀的决策能力,但是无法处理状态-动作空间过大的情况,使得传统的强化学习很难应用到真实的交通环境中。

深度强化学习是将深度学习的感知能力与强化学习的决策能力相结合的学习算法,并且能够通过端对端的方式实现从原始输入到输出的直接控制。深度强化学习与强化学习最大的区别是:深度强化学习利用深度神经网络强大的表征能力拟合强化学习存储信息的 Q-table 以解决状态-动作空间过大的问题,使深度强化学习能够应用于交通信号灯配时系统中。深度强化学习中的 2 个关键技术如下:

1)经验池。经验池的主要功能是解决样本相关性和非静态分布问题。具体内容是:将智能体与环境交互得到的样本储存到回放记忆单元,并随机选择一些样本用于对深度强化学习网络的训练。这种处理方式打破了样本间的关联,使样本间相互独立。

2)固定目标网络参数。目标网络与估计网络结构相同,但目标网络的参数不会迭代训练更新,而是每隔一定时间步将当前估计网络的参数完全复制过来。这种方法可以减少目标网络值函

数与估计网络值函数的相关性，提高了训练的稳定性。

### 1.2　扩展卡尔曼滤波

标准卡尔曼滤波（Kalman filter，KF）是一种对系统状态进行最优估计的算法，适用于线性、离散和有限维系统。每个有外部变量的自回归移动平均系统或可用有理传递函数表示的系统都可以转换为用状态空间表示的系统，从而用卡尔曼滤波进行计算。因此，自从卡尔曼滤波理论问世以来，在通信系统、电力系统和工业控制等领域得到了广泛应用。

EKF 是标准卡尔曼滤波在非线性情形下的一种拓展形式。EKF 的基本思想是：利用泰勒级数展开式将非线性系统局部线性化，采用标准卡尔曼滤波框架对信号进行滤波处理。EKF 处理非线性系统的能力使其在航天系统、雷达系统和军事领域取得了广泛应用。

## 2　基于 DQN-EKF 的交通信号灯配时方法

基于 DQN-EKF 算法的交通配时系统中，将交通信号灯配时系统抽象为智能体，十字路口的交通环境抽象为被控对象，交通信号灯的配时方案抽象为动作，车辆累计等待时间的变化抽象为奖励。首先，智能体根据被控对象提供的状态信息选择动作；然后，执行动作后，被控对象将当前状态和奖励反馈给智能体，并且不断重复此过程；最后，智能体以获取最大奖励值为目标不断更新参数直至得到最优动作。经验池用来存储训练样本，并用于 DQN 网络训练。具体的模型框架如图 1所示。图中：$\boldsymbol{s}_t$ 为 $t$ 时刻的状态；$\boldsymbol{a}_t$ 为 $t$ 时刻的动作；$\boldsymbol{r}_t$ 为 $t$ 时刻的奖励；$\boldsymbol{D}_t$ 为 $t$ 时刻的回放记忆单元。

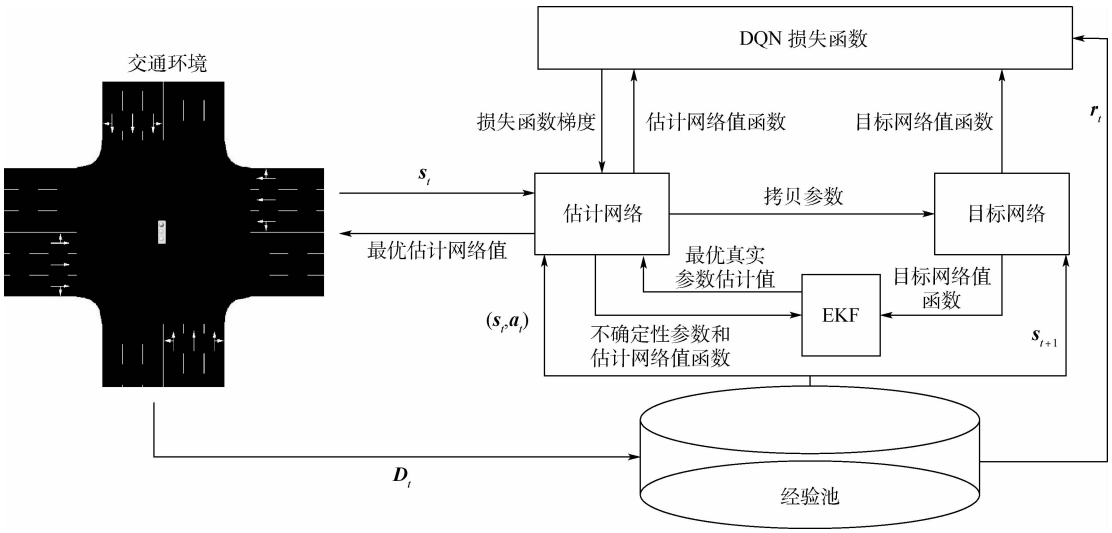



图 1　模型总体框架

Fig. 1　Overall framework of the model

### 2.1　状态空间

为了准确定义十字路口的交通状态信息，本文采用离散交通状态编码（discrete traffic state encoding，DTSE）方法将整个十字路口网格化，并构建交通状态信息矩阵。该方法能够减少数据维度，加快训练速度，并有助于智能体做出有效决策。正常车辆长度为 4.5 ~ 5 m，为保证不将2 辆车放在同一个网格里，并略大于普通车辆的长度以减少计算量，则设网格大小为 5 m。在每个网格中，状态值为 2 个值向量 < 位置，速度 > 。位置矩阵表示网格中是否有车辆，如果有车辆，取值为 1，否则为 0。速度矩阵表示车辆当前速度，单位为 m/s。图 2 为交通状态信息表示。

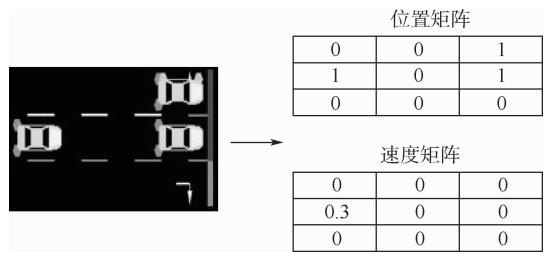



图 2　交通状态信息表示

Fig. 2　Traffic state information representation

### 2.2　动作空间

交通信号灯需要根据当前的交通状态选择合适的动作来引导十字路口的车辆。本文中交通信号灯的相位变化如图 3 和图 4 所示。十字路口中，相位 1 表示南北方向直行（同时北向西右转，

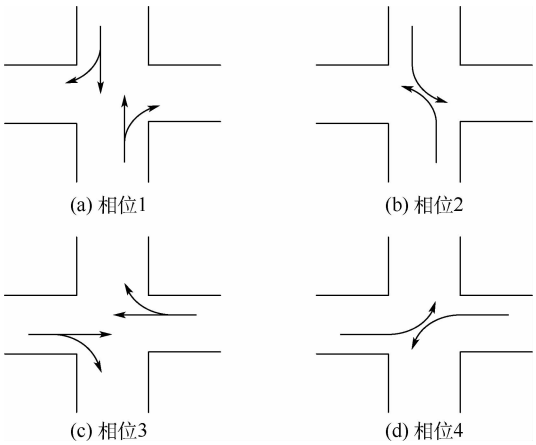



图 3 十字路口相位变化

Fig. 3 Phase change diagram of crossroads

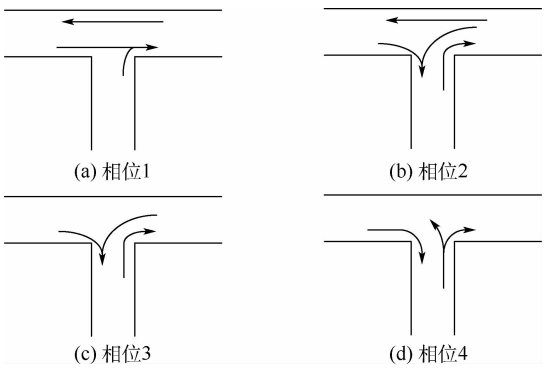



图 4 丁字路口相位变化

Fig. 4 Phase change diagram of T-junction

南向东右转),相位 2 表示南北方向左转,相位 3 表示东西方向直行(同时东向北右转,西向南右转),相位 4 表示东西方向左转;丁字路口中,相位 1 表示东西直行(同时南向东右转),相位 2 表示东向直行且左转(同时西向南右转,南向东右转),相位 3 表示东向左转(同时西向南右转,南向东右转),相位 4 表示南向左转(同时西向南右转,南向东右转)。为保证交通系统的稳定性,每个相位都设置了红绿灯的最大持续时间和变化范围,并且在红灯与绿灯之间设置持续时间为 3 s 的黄色信号灯。

### 2.3 奖励函数

奖励函数是区分强化学习和其他学习算法的重要因素。奖励的作用是就先前动作的表现向强化学习模型提供反馈。因此,奖励的定义方式对指导强化学习过程非常重要,有助于智能体采取最优的行动策略。在交通信号灯配时系统中,主要目标是提高十字路口的车辆通行效率。车辆通行效率的主要衡量标准是车辆的等待时间,因此,将奖励定义为相邻 2 个周期之间累计等待时间之差,具体公式如下:

$$\boldsymbol{r}_t = \boldsymbol{W}_t - \boldsymbol{W}_{t+1} \tag{1}$$

式中:$\boldsymbol{r}_t$ 为 $t$ 时刻所有执行动作的奖励值;$\boldsymbol{W}_t$ 为 $t$ 时刻执行不同动作的累计等待时间。

### 2.4 Q-learning 算法

Q-learning 是基于价值函数的强化学习算法,其主要思想是:用全部状态$(\boldsymbol{s}_1,\boldsymbol{s}_2,\cdots,\boldsymbol{s}_t,\cdots)\in\boldsymbol{S}$ 与动作$(\boldsymbol{a}_1,\boldsymbol{a}_2,\cdots,\boldsymbol{a}_t,\cdots)\in\boldsymbol{A}$ 构建一张表格来存储期望值,根据期望值选择获得回报最大的动作。$Q(\boldsymbol{s}_t,\boldsymbol{a}_t)$ 为在某一时刻的状态 $\boldsymbol{s}_t$ 下采取动作 $\boldsymbol{a}_t$ 的估计网络值函数,环境会根据智能体执行的动作反馈相应的奖励 $\boldsymbol{r}_t$,其中,$(\boldsymbol{r}_1,\boldsymbol{r}_2,\cdots,\boldsymbol{r}_t,\cdots)\in\boldsymbol{R}$。Q-learning 的更新公式如下:

$$Q(\boldsymbol{s}_t,\boldsymbol{a}_t) = Q(\boldsymbol{s}_t,\boldsymbol{a}_t) + \alpha(\boldsymbol{r}_t + \gamma\max_{\boldsymbol{a}_{t+1}} Q(\boldsymbol{s}_{t+1},\boldsymbol{a}_{t+1}) - Q(\boldsymbol{s}_t,\boldsymbol{a}_t)) \tag{2}$$

式中:$\alpha\in(0,1)$ 为学习率;$\gamma\in(0,1)$ 为衰减因子。

### 2.5 DQN 算法

DQN 在 Q-learning 的基础上增加了神经网络,并采用对偶网络结构。其中,一个估计网络用来选择动作并更新参数;另一个目标网络只用来计算函数值,主要用于估计网络的更新。目标网络的参数不会进行迭代更新,而是每隔一段时间从主网络中将参数复制过来。因此,2 个网络的结构相同,但是参数不同。DQN 利用这种网络结构最小化损失函数并更新网络结构。DQN 的损失函数如下:

$$\mathrm{Loss}(\boldsymbol{\theta}_t) = (\boldsymbol{r}_t + \gamma\max_{\boldsymbol{a}_{t+1}} Q(\boldsymbol{s}_{t+1},\boldsymbol{a}_{t+1}\mid\tilde{\boldsymbol{\theta}}) - Q(\boldsymbol{s}_t,\boldsymbol{a}_t\mid\boldsymbol{\theta}_t))^2 \tag{3}$$

式中:$\boldsymbol{Z}_t = \boldsymbol{r}_t + \gamma\max_{\boldsymbol{a}_{t+1}} Q(\boldsymbol{s}_{t+1},\boldsymbol{a}_{t+1}\mid\tilde{\boldsymbol{\theta}})$ 为 $t$ 时刻目标网络值函数,$\tilde{\boldsymbol{\theta}}$ 为目标网络的固定权重;$Q(\boldsymbol{s}_t,\boldsymbol{a}_t\mid\boldsymbol{\theta}_t)$ 为 $t$ 时刻估计网络值函数,$\boldsymbol{\theta}_t$ 为 $t$ 时刻估计网络的不确定性参数。

为使损失函数最小化,确定了损失函数相对于函数参数 $\boldsymbol{\theta}_t$ 的梯度:

$$\frac{\mathrm{dLoss}(\boldsymbol{\theta}_t)}{\mathrm{d}\boldsymbol{\theta}_t} = (\boldsymbol{r}_t + \gamma\max_{\boldsymbol{a}_{t+1}} Q(\boldsymbol{s}_{t+1},\boldsymbol{a}_{t+1}\mid\tilde{\boldsymbol{\theta}}))\cdot\frac{\mathrm{d}Q(\boldsymbol{s}_t,\boldsymbol{a}_t\mid\boldsymbol{\theta}_t)}{\mathrm{d}\boldsymbol{\theta}_t} - Q(\boldsymbol{s}_t,\boldsymbol{a}_t\mid\boldsymbol{\theta}_t)\frac{\mathrm{d}Q(\boldsymbol{s}_t,\boldsymbol{a}_t\mid\boldsymbol{\theta}_t)}{\mathrm{d}\boldsymbol{\theta}_t} \tag{4}$$

权重更新如下:

$$\boldsymbol{\theta}_{t+1}\leftarrow\boldsymbol{\theta}_t + \alpha\left((\boldsymbol{Z}_t - Q(\boldsymbol{s}_t,\boldsymbol{a}_t\mid\boldsymbol{\theta}_t))\frac{\mathrm{d}Q(\boldsymbol{s}_t,\boldsymbol{a}_t\mid\boldsymbol{\theta}_t)}{\mathrm{d}\boldsymbol{\theta}_t}\right) \tag{5}$$

### 2.6 DQN 与 EKF 相结合的控制算法

本文将 EKF 用于解决深度强化学习模型中的参数不确定性问题,结构如图 5 所示。具体结

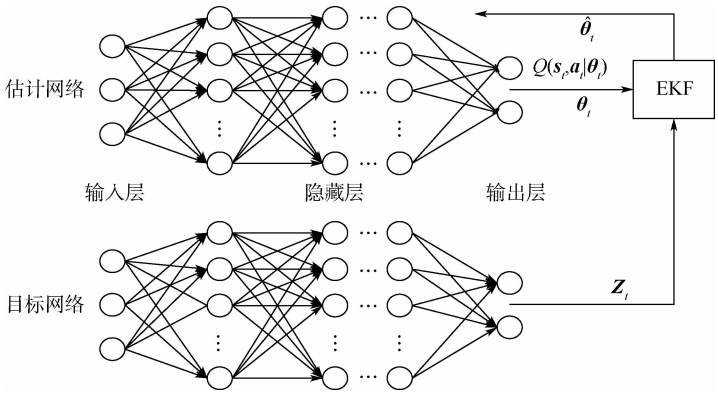



图 5　DQN 与 EKF 相结合的结构

Fig. 5　Structure of combination of DQN and EKF

合过程如下:

首先,明确 EKF 的状态方程和观测方程,具体公式如下:

$$\boldsymbol{x}_i = f(\boldsymbol{x}_{i-1}) + \boldsymbol{w}_{i-1} \tag{6}$$

$$\boldsymbol{z}_i = h(\boldsymbol{x}_i) + \boldsymbol{v}_i \tag{7}$$

式中:$\boldsymbol{x}_i$ 为状态变量;$\boldsymbol{z}_i$ 为观测变量;$f(\cdot)$ 和 $h(\cdot)$ 分别为非线性系统的状态函数和观测函数;$\boldsymbol{w}_i$ 为过程噪声,$\boldsymbol{v}_i$ 为观测噪声,协方差分别为 $\boldsymbol{O}_i$ 和 $\boldsymbol{R}_i$。由于过程噪声和观测噪声满足高斯分布,设 $E[\boldsymbol{w}_i]=0$,$E[\boldsymbol{v}_i]=0$,$E[\boldsymbol{w}_i\boldsymbol{w}_i^{\mathrm{T}}]=\boldsymbol{O}_i$,$E[\boldsymbol{v}_i\boldsymbol{v}_i^{\mathrm{T}}]=\boldsymbol{R}_i$。

然后,将 $t$ 时刻的不确定性参数 $\boldsymbol{\theta}_t$ 代入 EKF 状态方程,$\boldsymbol{Z}_t$ 和 $Q(\boldsymbol{s}_t,\boldsymbol{a}_t\mid\boldsymbol{\theta}_t)$ 代入 EKF 观测方程。具体公式如下:

$$\boldsymbol{\theta}_i = \boldsymbol{\theta}_{i-1} + \boldsymbol{w}_{i-1} \tag{8}$$

$$\boldsymbol{Z}_i = Q(\boldsymbol{s}_t,\boldsymbol{a}_t\mid\boldsymbol{\theta}_i) + \boldsymbol{v}_i \tag{9}$$

式中:状态变量 $\boldsymbol{\theta}_i$ 表示 $t$ 时刻不确定性参数 $\boldsymbol{\theta}_t$ 迭代 $i$ 次的参数值;观测变量 $\boldsymbol{Z}_i$ 表示 $t$ 时刻的目标网络值函数 $\boldsymbol{Z}_t$ 迭代 $i$ 次的函数值。非线性状态函数 $f(\boldsymbol{\theta}_i)=\boldsymbol{\theta}_i$,非线性观测函数 $h(\boldsymbol{\theta}_i)=Q(\boldsymbol{s}_t,\boldsymbol{a}_t\mid\boldsymbol{\theta}_i)$。

分别将 $f(\boldsymbol{\theta}_i)$ 在 $\hat{\boldsymbol{\theta}}_{i|i}$ 处泰勒一阶展开,$h(\boldsymbol{\theta}_i)$ 在 $\breve{\boldsymbol{\theta}}_{i|i-1}$ 处泰勒一阶展开,具体公式如下:

$$f(\boldsymbol{\theta}_i) = f(\hat{\boldsymbol{\theta}}_{i|i}) + \left.\frac{\partial f}{\partial\boldsymbol{\theta}_i}\right|_{\boldsymbol{\theta}_i=\hat{\boldsymbol{\theta}}_{i|i}} (\boldsymbol{\theta}_i - \hat{\boldsymbol{\theta}}_{i|i}) + o(\boldsymbol{\theta}_i - \hat{\boldsymbol{\theta}}_{i|i}) \tag{10}$$

$$h(\boldsymbol{\theta}_i) = h(\breve{\boldsymbol{\theta}}_{i|i-1}) + \left.\frac{\partial h}{\partial\boldsymbol{\theta}_i}\right|_{\boldsymbol{\theta}_i=\breve{\boldsymbol{\theta}}_{i|i-1}} (\boldsymbol{\theta}_i - \breve{\boldsymbol{\theta}}_{i|i-1}) + o(\boldsymbol{\theta}_i - \breve{\boldsymbol{\theta}}_{i|i-1}) \tag{11}$$

式中:$\hat{\boldsymbol{\theta}}_{i|i}$ 为迭代 $i$ 次时的真实参数估计值;$\breve{\boldsymbol{\theta}}_{i|i-1}$ 为迭代 $i$ 次时的真实参数预测值;$o(\boldsymbol{\theta}_i-\hat{\boldsymbol{\theta}}_{i|i})$ 和 $o(\boldsymbol{\theta}_i-\breve{\boldsymbol{\theta}}_{i|i-1})$ 为高阶项可忽略。状态转移矩阵 $\boldsymbol{F}_i = \left.\frac{\partial f}{\partial\boldsymbol{\theta}_i}\right|_{\boldsymbol{\theta}_i=\hat{\boldsymbol{\theta}}_{i|i}} = \boldsymbol{E}$,观测转移矩阵 $\boldsymbol{H}_i = \left.\frac{\partial h}{\partial\boldsymbol{\theta}_i}\right|_{\boldsymbol{\theta}_i=\breve{\boldsymbol{\theta}}_{i|i-1}} = \frac{\mathrm{d}Q(\boldsymbol{s}_t,\boldsymbol{a}_t\mid\breve{\boldsymbol{\theta}}_{i|i-1})}{\mathrm{d}\boldsymbol{\theta}_i}$。则状态方程和观测方程为

$$\boldsymbol{\theta}_i = f(\hat{\boldsymbol{\theta}}_{i-1|i-1}) + \boldsymbol{F}_{i-1}(\boldsymbol{\theta}_{i-1} - \hat{\boldsymbol{\theta}}_{i-1|i-1}) + \boldsymbol{w}_{i-1} \tag{12}$$

$$\boldsymbol{Z}_i = h(\breve{\boldsymbol{\theta}}_{i|i-1}) + \boldsymbol{H}_i(\boldsymbol{\theta}_i - \breve{\boldsymbol{\theta}}_{i|i-1}) + \boldsymbol{v}_i \tag{13}$$

迭代 $i$ 次时真实参数预测值 $\breve{\boldsymbol{\theta}}_{i|i-1}$ 由迭代 $i-1$ 次时真实参数估计值 $\hat{\boldsymbol{\theta}}_{i-1|i-1}$ 代替。具体推导公式如下:

$$\breve{\boldsymbol{\theta}}_{i|i-1} = E[f(\hat{\boldsymbol{\theta}}_{i-1|i-1}) + \boldsymbol{F}_{i-1}(\boldsymbol{\theta}_{i-1} - \hat{\boldsymbol{\theta}}_{i-1|i-1}) + \boldsymbol{w}_{i-1}] = \hat{\boldsymbol{\theta}}_{i-1|i-1} \tag{14}$$

由状态变量 $\boldsymbol{\theta}_i$ 与真实参数预测值 $\breve{\boldsymbol{\theta}}_{i|i-1}$ 的误差计算得到状态预测误差协方差 $\breve{\boldsymbol{P}}_{i|i-1}$。结果表明,迭代 $i$ 次时的 $\breve{\boldsymbol{P}}_{i|i-1}$ 与迭代 $i-1$ 次时的状态估计误差协方差 $\hat{\boldsymbol{P}}_{i-1|i-1}$ 有关,具体推导公式如下:

$$\begin{aligned}\breve{\boldsymbol{P}}_{i|i-1} &= E[(\boldsymbol{\theta}_i - \breve{\boldsymbol{\theta}}_{i|i-1})(\boldsymbol{\theta}_i - \breve{\boldsymbol{\theta}}_{i|i-1})^{\mathrm{T}}] = \\ &E\{[\boldsymbol{F}_{i-1}(\boldsymbol{\theta}_{i-1} - \hat{\boldsymbol{\theta}}_{i-1|i-1}) + \boldsymbol{w}_{i-1}]\cdot \\ &[\boldsymbol{F}_{i-1}(\boldsymbol{\theta}_{i-1} - \hat{\boldsymbol{\theta}}_{i-1|i-1}) + \boldsymbol{w}_{i-1}]^{\mathrm{T}}\} = \\ &\boldsymbol{F}_{i-1}\hat{\boldsymbol{P}}_{i-1|i-1}\boldsymbol{F}_{i-1}^{\mathrm{T}} + \boldsymbol{O}_{i-1}\end{aligned} \tag{15}$$

迭代 $i$ 次时的真实观测变量预测值 $\breve{\boldsymbol{Z}}_{i|i-1}$ 等于 $h(\breve{\boldsymbol{\theta}}_{i|i-1})$,具体推导公式如下:

$$\breve{\boldsymbol{Z}}_{i|i-1} = E[h(\breve{\boldsymbol{\theta}}_{i|i-1}) + \boldsymbol{H}_i(\boldsymbol{\theta}_i - \breve{\boldsymbol{\theta}}_{i|i-1}) + \boldsymbol{v}_i] = h(\breve{\boldsymbol{\theta}}_{i|i-1}) \tag{16}$$

由观测变量 $\boldsymbol{Z}_i$ 与真实观测变量预测值 $\breve{\boldsymbol{Z}}_{i|i-1}$ 的误差计算得到观测预测误差协方差 $\ddot{\boldsymbol{P}}_{i|i-1}$,具体推导公式如下:

$$\begin{aligned}\ddot{\boldsymbol{P}}_{i|i-1} &= E[(\boldsymbol{Z}_i - \breve{\boldsymbol{Z}}_{i|i-1})(\boldsymbol{Z}_i - \breve{\boldsymbol{Z}}_{i|i-1})^{\mathrm{T}}] = \\ &E\{[\boldsymbol{H}_i(\boldsymbol{\theta}_i - \breve{\boldsymbol{\theta}}_{i|i-1}) + \boldsymbol{v}_i][\boldsymbol{H}_i(\boldsymbol{\theta}_i - \breve{\boldsymbol{\theta}}_{i|i-1}) + \boldsymbol{v}_i]^{\mathrm{T}}\} = \\ &\boldsymbol{H}_i\breve{\boldsymbol{P}}_{i|i-1}\boldsymbol{H}_i^{\mathrm{T}} + \boldsymbol{R}_i\end{aligned} \tag{17}$$

同理可得,观测预测误差与状态预测误差之间的协方差矩阵表示为状态观测预测误差协方差 $\overset{\cdots}{\boldsymbol{P}}_{i\mid i-1}$,具体推导公式如下:

$$\begin{aligned}\overset{\cdots}{\boldsymbol{P}}_{i\mid i-1} &= E[(\boldsymbol{\theta}_i-\breve{\boldsymbol{\theta}}_{i\mid i-1})(\boldsymbol{Z}_i-\breve{\boldsymbol{Z}}_{i\mid i-1})^{\mathrm{T}}] = \\ &E\{(\boldsymbol{\theta}_i-\breve{\boldsymbol{\theta}}_{i\mid i-1})[\boldsymbol{H}_i(\boldsymbol{\theta}_i-\breve{\boldsymbol{\theta}}_{i\mid i-1})+\boldsymbol{v}_i]^{\mathrm{T}}\} = \\ &\breve{\boldsymbol{P}}_{i\mid i-1}\boldsymbol{H}_i^{\mathrm{T}}\end{aligned} \tag{18}$$

EKF 增益 $\boldsymbol{K}_i$ 表示状态观测预测误差协方差与观测预测误差协方差的比重,具体公式如下:

$$\begin{aligned}\boldsymbol{K}_i &= \overset{\cdots}{\boldsymbol{P}}_{i\mid i-1}(\ddot{\boldsymbol{P}}_{i\mid i-1})^{-1} = \\ &\breve{\boldsymbol{P}}_{i\mid i-1}\boldsymbol{H}_i^{\mathrm{T}}(\boldsymbol{H}_i\breve{\boldsymbol{P}}_{i\mid i-1}\boldsymbol{H}_i^{\mathrm{T}}+\boldsymbol{R}_i)^{-1}\end{aligned} \tag{19}$$

迭代 $i$ 次时的真实参数估计值 $\hat{\boldsymbol{\theta}}_{i\mid i}$ 通过不断迭代更新得到最优真实参数估计值 $\hat{\boldsymbol{\theta}}_t$,并返回至估计网络中用于计算值函数,具体公式如下:

$$\hat{\boldsymbol{\theta}}_{i\mid i} = \breve{\boldsymbol{\theta}}_{i\mid i-1}+\boldsymbol{K}_i[\boldsymbol{Z}_i-\breve{\boldsymbol{Z}}_{i\mid i-1}] \tag{20}$$

迭代 $i$ 次时的状态变量 $\boldsymbol{\theta}_i$ 与真实参数估计值 $\hat{\boldsymbol{\theta}}_{i\mid i}$ 的误差协方差表示为状态估计误差协方差 $\hat{\boldsymbol{P}}_{i\mid i}$,具体推导公式如下:

$$\begin{aligned}\hat{\boldsymbol{P}}_{i\mid i} &= E[(\boldsymbol{\theta}_i-\hat{\boldsymbol{\theta}}_{i\mid i})(\boldsymbol{\theta}_i-\hat{\boldsymbol{\theta}}_{i\mid i})^{\mathrm{T}}] = \\ &E\{[\boldsymbol{\theta}_i-\breve{\boldsymbol{\theta}}_{i\mid i-1}-\boldsymbol{K}_i(\boldsymbol{Z}_i-\breve{\boldsymbol{Z}}_{i\mid i-1})]\cdot \\ &[\boldsymbol{\theta}_i-\breve{\boldsymbol{\theta}}_{i\mid i-1}-\boldsymbol{K}_i(\boldsymbol{Z}_i-\breve{\boldsymbol{Z}}_{i\mid i-1})]^{\mathrm{T}}\} = \\ &(\boldsymbol{I}-\boldsymbol{K}_i\boldsymbol{H}_i)\breve{\boldsymbol{P}}_{i\mid i-1}(\boldsymbol{I}-\boldsymbol{K}_i\boldsymbol{H}_i)^{\mathrm{T}}+\boldsymbol{K}_i\boldsymbol{R}_i\boldsymbol{K}_i^{\mathrm{T}}\end{aligned} \tag{21}$$

综上所述,DQN 与 EKF 相结合的主要过程为:①将 DQN 模型 $t$ 时刻的 $\boldsymbol{\theta}_t$、$Q(\boldsymbol{s}_t,\boldsymbol{a}_t\mid\boldsymbol{\theta}_t)$ 和 $\boldsymbol{Z}_t$ 作为输入构建 EKF 的状态方程和观测方程,通过计算误差协方差得到 EKF 增益;②迭代更新使 EKF 增益不断减小真实参数估计值的误差至收敛,得到最优真实参数估计值 $\hat{\boldsymbol{\theta}}_t$;③将最优真实参数估计值 $\hat{\boldsymbol{\theta}}_t$ 传输到估计网络中得到 $Q(\boldsymbol{s}_t,\boldsymbol{a}_t\mid\hat{\boldsymbol{\theta}}_t)$,根据估计网络值 $Q(\boldsymbol{s}_t,\boldsymbol{a}_t\mid\hat{\boldsymbol{\theta}}_t)$ 中的最大值选择最优的配时方案用于交通环境,提高交通路口的通行效率。同时得到 $t+1$ 时刻的 $\boldsymbol{\theta}_{t+1}$、$\boldsymbol{Z}_{t+1}$ 和 $Q(\boldsymbol{s}_{t+1},\boldsymbol{a}_{t+1}\mid\boldsymbol{\theta}_{t+1})$,重复上述过程得到 $t+1$ 时刻的最优配时方案,以此类推直至全局模型收敛。

# 3 仿真实验设计及实验结果分析

## 3.1 仿真实验设计

城市交通仿真(simulation of urban mobility, SUMO)是一个微观交通模拟软件,可以准确模拟城市交通场景,其还提供了一个可视化的图形界面,支持多种网格格式的输入和各种道路网络设计。利用 SUMO 中提供的 Traci 接口模块与仿真平台在线交互,使交通信号灯配时系统能够获取实时的交通状态信息,并有效管理交通路况。在 SUMO 平台上,模拟 3 种不同的交通仿真场景,并设置了正常交通流和高峰交通流。场景 1 为相同路段长度的十字路口,每条路段长度为 300 m;场景 2 为多种路段长度的十字路口,路段长度分别为 200 m、300 m、400 m 和 500 m;场景 3 为相同路段长度的丁字路口,每条路段长度为 300 m。正常交通流中,每条道路的车辆到达率为 489 veh/h(veh/h 表示每小时通过的车辆数),高峰交通流中,每条道路的车辆到达率为 727 veh/h。为保证交通安全,车辆的最大速度为 13.89 m/s,最大加速度为 2 m/s$^2$。仿真实验中,深度强化学习网络的超参数如表 1 所示。

**表 1 参数设置**

**Table 1 Parameter setting**

| 参数 | 数值 |
|---|---|
| 批量大小 $B$ | 32 |
| 衰减因子 $\gamma$ | 0.9 |
| 学习率 $\alpha$ | 0.001 |
| 经验池大小 $N$ | 10 000 |
| 探索率 $\varepsilon$ | 1.0→0.01 |

## 3.2 仿真实验结果分析

由于初始阶段的探索机制和经验不足问题,使智能体不能学习到正确的配时方案,导致 DQN、Dueling DQN、Double DQN 和 DQN-EKF 的平均等待时间和平均队列长度较长。随着探索率的降低和经验累积,DQN、Dueling DQN、Double DQN 和 DQN-EKF 的平均等待时间和平均队列长度开始逐渐下降至收敛。

在场景 1 中,正常交通流的情况下,由图 6、图 7 和表 2 可知,在平均等待时间方面,DQN-EKF

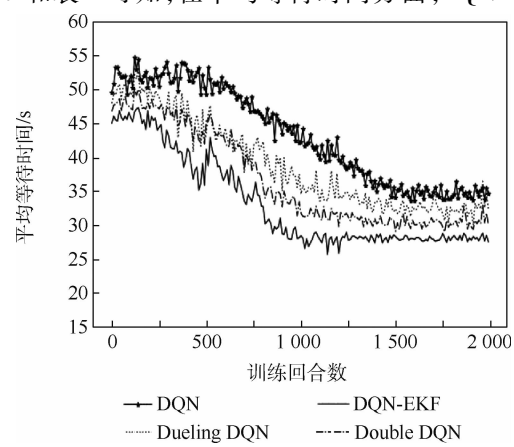



图 6 场景 1 中正常交通流下的平均等待时间变化

Fig. 6 Variation of average waiting time under normal traffic flow in Scenario 1

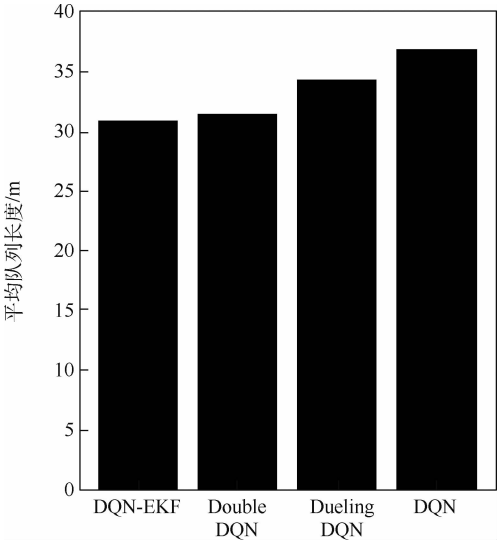



图 7　场景 1 中正常交通流下的平均队列长度

Fig. 7　Average queue length under normal traffic flow in Scenario 1

**表 2　场景 1 中正常交通流下的算法性能对比**

**Table 2　Comparison of algorithm performance under normal traffic flow in Scenario 1**

| 算法 | 平均等待时间/s | 平均队列长度/m |
|---|---|---|
| DQN | 33.70 | 36.88 |
| Dueling DQN | 30.87 | 34.31 |
| Double DQN | 29.02 | 31.47 |
| DQN-EKF | 25.81 | 30.87 |

分别比 Double DQN、Dueling DQN 和 DQN 降低了 11.06%、16.39% 和 23.41%;在平均队列长度方面,DQN-EKF 分别比 Double DQN、Dueling DQN 和 DQN 降低了 1.91%、10.03% 和 16.30%。

在场景 1 中,高峰交通流的情况下,由图 8、图 9 和表 3 可知,在平均等待时间方面,DQN-EKF 分别比 Double DQN、Dueling DQN 和 DQN 降低了 13.03%、18.57% 和 22.22%;在平均队列长度方面,DQN-EKF 分别比 Double DQN、Dueling DQN 和 DQN 降低了 10.81%、15.75% 和 20.23%。

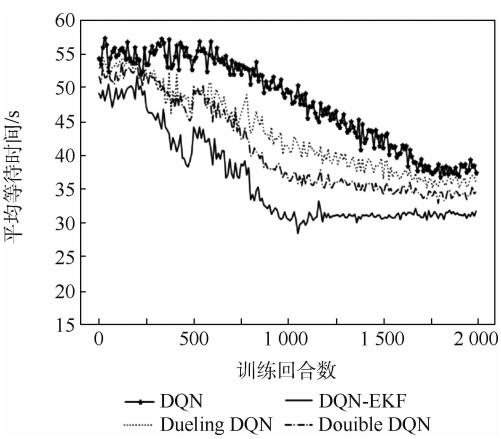



图 8　场景 1 中高峰交通流下的平均等待时间变化

Fig. 8　Variation of average waiting time under peak traffic flow in Scenario 1

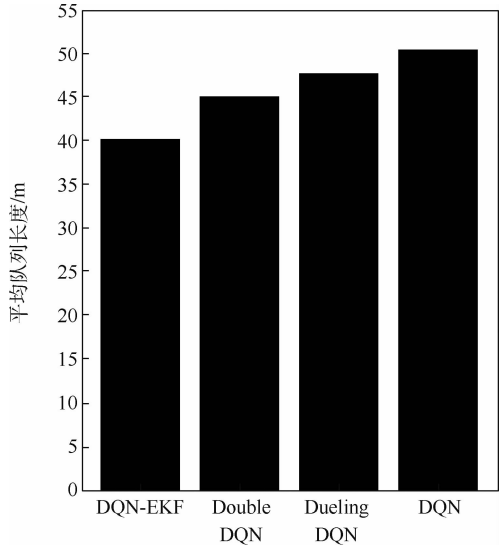



图 9　场景 1 中高峰交通流下的平均队列长度

Fig. 9　Average queue length under peak traffic flow in Scenario 1

**表 3　场景 1 中高峰交通流下的算法性能对比**

**Table 3　Comparison of algorithm performance under peak traffic flow in Scenario 1**

| 算法 | 平均等待时间/s | 平均队列长度/m |
|---|---|---|
| DQN | 36.81 | 50.36 |
| Dueling DQN | 35.16 | 47.68 |
| Double DQN | 32.92 | 45.04 |
| DQN-EKF | 28.63 | 40.17 |

在场景 2 中,正常交通流的情况下,由图 10、图 11 和表 4 可知,在平均等待时间方面,DQN-EKF 分别比 Double DQN、Dueling DQN 和 DQN 降低了 13.51%、18.89% 和 23.79%;在平均队列长度方面,DQN-EKF 分别比 Double DQN、Dueling DQN 和 DQN 降低了 10.85%、18.45% 和 27.24%。

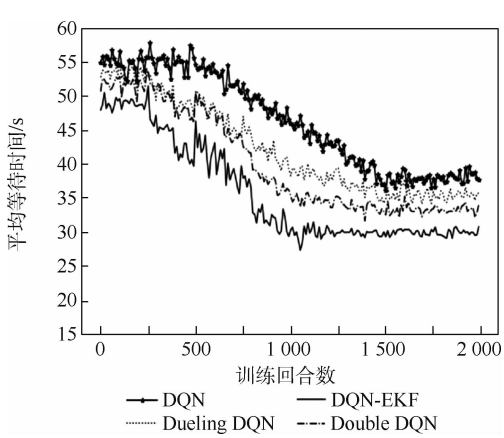



图 10　场景 2 中正常交通流下的平均等待时间变化

Fig. 10　Variation of average waiting time under normal traffic flow in Scenario 2

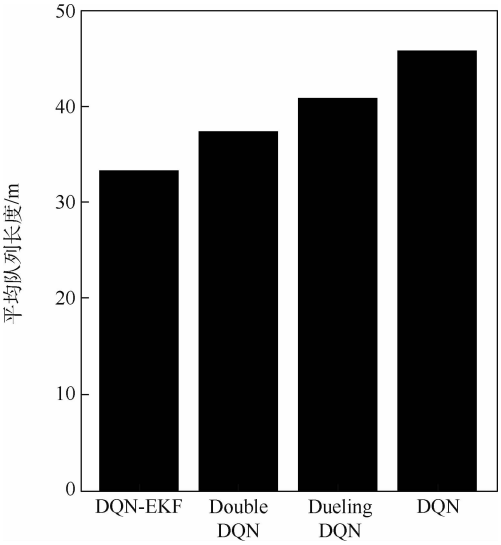



图 11 场景 2 中正常交通流下的平均队列长度

Fig. 11 Average queue length under normal traffic flow in Scenario 2

**表 4 场景 2 中正常交通流下的算法性能对比**

**Table 4 Comparison of algorithm performance under normal traffic flow in Scenario 2**

| 算法 | 平均等待时间/s | 平均队列长度/m |
|---|---|---|
| DQN | 36.03 | 45.74 |
| Dueling DQN | 34.28 | 40.81 |
| Double DQN | 31.75 | 37.33 |
| DQN-EKF | 27.46 | 33.28 |

在场景 2 中,高峰交通流的情况下,由图 12、图 13 和表 5 可知,在平均等待时间方面,DQN-EKF 分别比 Double DQN、Dueling DQN 和 DQN 降低了9.85%、15.23% 和 20.01%;在平均队列长度方面,DQN-EKF 分别比 Double DQN、Dueling DQN 和 DQN 降低了 8.38%、13.08% 和 20.26%。

在场景 3 中,正常交通流的情况下,由图 14、图 15 和表 6 可知,在平均等待时间方面,DQN-EKF

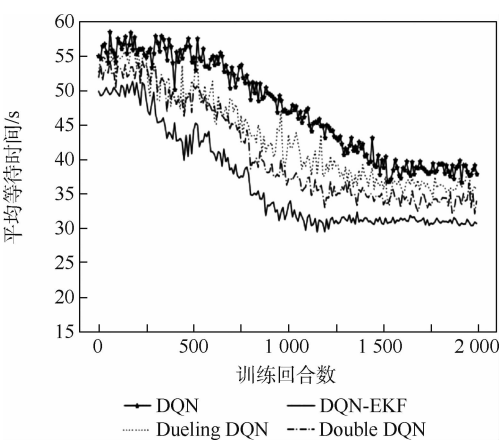



图 12 场景 2 中高峰交通流下的平均等待时间变化

Fig. 12 Variation of average waiting time under peak traffic flow in Scenario 2

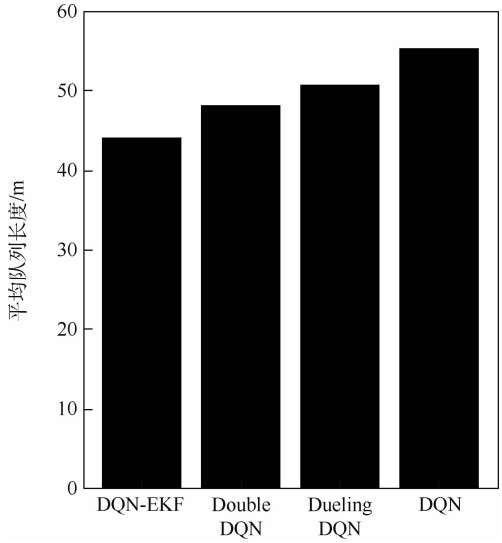



图 13 场景 2 中高峰交通流下的平均队列长度

Fig. 13 Average queue length under peak traffic flow in Scenario 2

**表 5 场景 2 中高峰交通流下的算法性能对比**

**Table 5 Comparison of algorithm performance under peak traffic flow in Scenario 2**

| 算法 | 平均等待时间/s | 平均队列长度/m |
|---|---|---|
| DQN | 37.98 | 55.42 |
| Dueling DQN | 35.84 | 50.84 |
| Double DQN | 33.70 | 48.23 |
| DQN-EKF | 30.38 | 44.19 |

分别比 Double DQN、Dueling DQN 和 DQN 降低了 11.29%、18.66% 和 25.19%;在平均队列长度方面,DQN-EKF 分别比 Double DQN、Dueling DQN 和 DQN 降低了 16.56%、19.94% 和 29.24%。

在场景 3 中,高峰交通流的情况下,由图 16、图 17 和表 7 可知,在平均等待时间方面,DQN-EKF 分别比 Double DQN、Dueling DQN 和 DQN 降低了 11.02%、20.87% 和 27.47%;在平均队列长

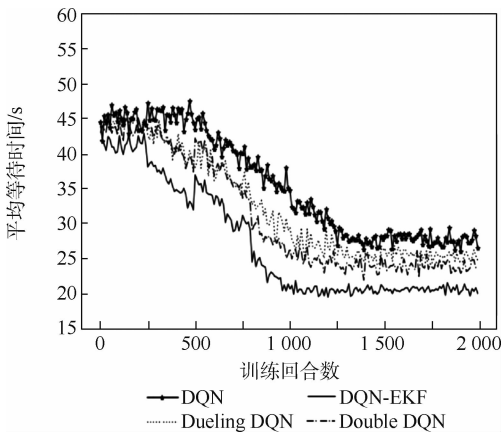



图 14 场景 3 中正常交通流下的平均等待时间变化

Fig. 14 Variation of average waiting time under normal traffic flow in Scenario 3

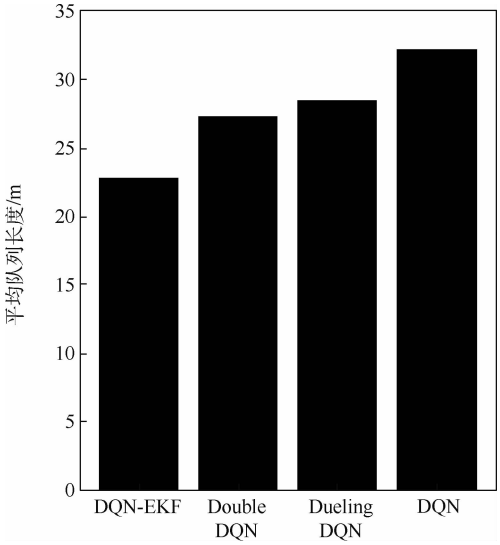



图 15　场景 3 中正常交通流下的平均队列长度

Fig. 15　Average queue length under normal traffic flow in Scenario 3

**表 6　场景 3 中正常交通流下的算法性能对比**

**Table 6　Comparison of algorithm performance under normal traffic flow in Scenario 3**

| 算法 | 平均等待时间/s | 平均队列长度/m |
|---|---|---|
| DQN | 26.68 | 32.18 |
| Dueling DQN | 24.54 | 28.44 |
| Double DQN | 22.50 | 27.29 |
| DQN-EKF | 19.96 | 22.77 |

度方面,DQN-EKF 分别比 Double DQN、Dueling DQN 和 DQN 降低了 14.69%、19.28% 和 27.61%。

通过在不同交通仿真环境下进行对比实验,充分证明 DQN-EKF 能够有效解决 DQN 中的参数不确定性问题,并且 DQN-EKF 的实验结果同样优于 Dueling DQN 和 Double DQN。主要原因为:Dueling DQN 和 Double DQN 都只是在 DQN 的值函数结果上做出改进,并没有从根本上解决由参数不确定性导致的值函数不准确问题,因此在对

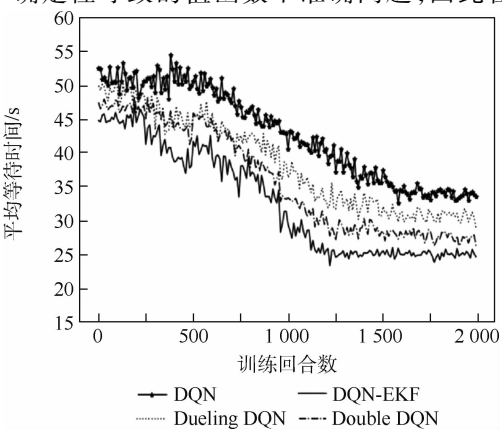



图 16　场景 3 中高峰交通流下的平均等待时间变化

Fig. 16　Variation of average waiting time under peak traffic flow in Scenario 3

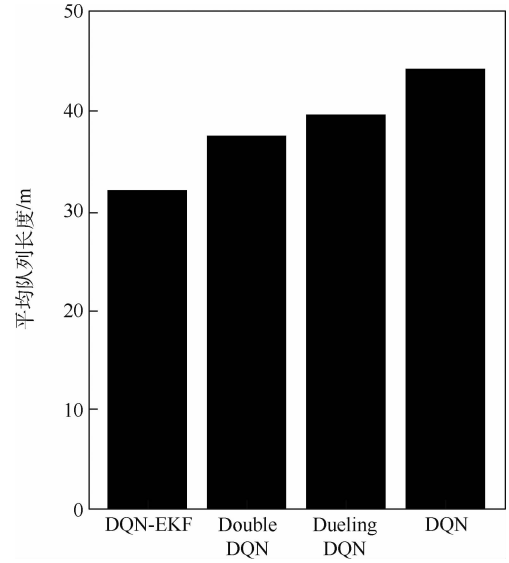



图 17　场景 3 中高峰交通流下的平均队列长度

Fig. 17　Average queue length under peak traffic flow in Scenario 3

**表 7　场景 3 中高峰交通流下的算法性能对比**

**Table 7　Comparison of algorithm performance under peak traffic flow in Scenario 3**

| 算法 | 平均等待时间/s | 平均队列长度/m |
|---|---|---|
| DQN | 32.62 | 44.36 |
| Dueling DQN | 29.90 | 39.78 |
| Double DQN | 26.59 | 37.64 |
| DQN-EKF | 23.66 | 32.11 |

比实验中没有 DQN-EKF 的实验效果好。Dueling DQN 与 DQN 的区别在于:Dueling DQN 不直接训练得到估计网络值函数,而是通过训练得到的状态函数与优势函数相加得到值函数,以此解决 DQN 中估计网络值函数不准确的问题;Double DQN 与 DQN 的区别在于:Double DQN 目标网络值函数中的动作由 DQN 估计网络参数计算得到动作代替,以此解决 DQN 目标网络值函数过估计的问题。综上所述,Dueling DQN、Double DQN 与 DQN 的网络结构基本相同且同样存在参数不确定性问题,因此 Double DQN 和 Dueling DQN 具有与 EKF 相结合的可能性。

## 4　结　论

1)针对 DQN 中存在的参数不确定性问题,提出将不确定性参数值作为 EKF 模型的输入,通过不断迭代更新得到最优真实参数估计值,利用最优真实参数估计值计算准确的值函数,并根据值函数选择最优的配时方案。

2)实验中设置了 3 种不同的交通仿真环境,并分别在正常交通流和高峰交通流的情况下进行了对比实验。实验证明,DQN-EKF 适用于不同的

交通环境,并能够有效提高交通路口的通行效率,同时算法性能优于 Dueling DQN、Double DQN 和 DQN。

3)分析说明 DQN-EKF 的实验结果优于 Dueling DQN、Double DQN 和 DQN 的主要原因,并讨论了 Dueling DQN 和 Double DQN 与 EKF 结合的可能性。

接下来的研究重点将关注多路口交通信号灯的配时方法、各路口之间的联合配时机制和处理复杂交通状态信息的方法等。

## 参考文献(References)


[1] ROBERTSON D I, BRETHERTON R D. Optimizing networks of traffic signals in real time-The SCOOT method [J]. IEEE Transactions on Vehicular Technology, 1991, 40(1): 11-15.

[2] LOWRIE P R. SCATS: The Sydney coordinated adaptive traffic system principles, methodology, algorithms [C]// International Conference on Road Traffic Signalling, 1982.

[3] 王史春. 基于 BP 模糊神经网络的交通信号灯控制器设计[J]. 云南民族大学学报(自然科学版), 2011, 20(6): 511-514.
WANG S C. Design of traffic signal controller based on BP fuzzy neural network [J]. Journal of Yunnan University for Nationalities(Natural Science Edition), 2011, 20(6): 511-514(in Chinese).

[4] 胡智鹏. 基于遗传算法的一种改进交叉路口信号灯实时控制优化方法[J]. 山东工业技术, 2015, 206(24): 110-111.
HU Z P. An improved optimization method for real-time control of intersection lights based on genetic algorithm [J]. Shandong Industrial Technology, 2015, 206(24): 110-111(in Chinese).

[5] BOUDERBA S I, MOUSSA N. Reinforcement learning (Q-LEARNING) traffic light controller within intersection traffic system[C]// Proceedings of the 4th International Conference on Big Data and Internet of Things. New York: ACM, 2019: 1-6.

[6] BUSCH J, LATZKO V, REISSLEIN M, et al. Optimised traffic light management through reinforcement learning: Traffic state agnostic agent vs. holistic agent with current V2I traffic state knowledge[J]. IEEE Open Journal of Intelligent Transportation Systems, 2020, 1: 201-216.

[7] GUO J, HARMATI I. Comparison of game theoretical strategy and reinforcement learning in traffic light control[J]. Periodica Polytechnica Transportation Engineering, 2020, 48(4): 313-319.

[8] GARG D, CHLI M, VOGIATZIS G. Deep reinforcement learning for autonomous traffic light control[C]// 2018 3rd IEEE International Conference on Intelligent Transportation Engineering (ICITE). Piscataway: IEEE Press, 2018: 214-218.

[9] GENDERS W, RAZAVI S, ASCE A M. Policy analysis of adaptive traffic signal control using reinforcement learning[J]. Journal of Computing in Civil Engineering, 2019, 34(1): 04019046.

[10] ZHOU X, FEI Z, QUAN L, et al. A Sarsa(λ)-based control model for real-time traffic light coordination[J]. The Scientific World Journal, 2014, 2014: 759097.

[11] YIN M, WANG Y, LI Z. Optimization of multi-intersection traffic signal timing model based on improved Q-learning[J]. IOP Conference Series: Materials Science and Engineering, 2020, 768(7): 072100.

[12] SHABESTARY S, ABDULHAI B. Deep learning vs. discrete reinforcement learning for adaptive traffic signal control[C]// 2018 IEEE International Conference on Intelligent Transportation Systems (ITSC). Piscataway: IEEE Press, 2018: 18308952.

[13] WANG J G, ZHOU L B. Traffic light recognition with high dynamic range imaging and deep learning[J]. IEEE Transactions on Intelligent Transportation Systems, 2019, 20(4): 1341-1352.

[14] SHARMA M, BANSAL A, KASHYAP V, et al. Intelligent traffic light control system based on traffic environment using deep learning[J]. IOP Conference Series: Materials Science and Engineering, 2021, 1022(1): 012122.

[15] LI H, KUMAR N, CHEN R, et al. Deep reinforcement learning [C]//2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Piscataway: IEEE Press, 2018: 18096630.

[16] KUMAR N, RAHMAN S S, DHAKAD N. Fuzzy inference enabled deep reinforcement learning-based traffic light control for intelligent transportation system[J]. IEEE Transactions on Intelligent Transportation Systems, 2021, 22(8): 4919-4928.

[17] TAN T, BAO F, DENG Y, et al. Cooperative deep reinforcement learning for large-scale traffic grid signal control[J]. IEEE Transactions on Cybernetics, 2020, 50(6): 2687-2700.

[18] ZENG J, HU J, ZHANG Y. Adaptive traffic signal control with deep recurrent Q-learning[C]//2018 IEEE Intelligent Vehicles Symposium(IV). Piscataway: IEEE Press, 2018: 18167960.

[19] LIAO L, LIU J, WU X, et al. Time difference penalized traffic signal timing by LSTM Q-network to balance safety and capacity at intersections[J]. IEEE Access, 2020, 8: 80086-80096.

[20] LIANG X, DU X, WANG G, et al. Deep reinforcement learning for traffic light control in vehicular networks[EB/OL]. (2018-03-29)[2021-09-01]. https://arxiv.org/abs/1803.11115.

[21] ZHOU Y, OZBAY K, CHOLETTE M, et al. A mode switching extended Kalman filter for real-time traffic state and parameter estimation[C]//2020 IEEE International Conyerence on Intelligent Transportation Systems (ITSC). Piscataway: IEEE Press, 2020: 20303008.

# Traffic signal timing method based on deep reinforcement learning and extended Kalman filter

WU Lan[1,*], WU Yuanming[1], KONG Fanshi[2], LI Binquan[1]

(1. College of Electrical Engineering, Henan University of Technology, Zhengzhou 450001, China;
2. College of Electrical Engineering, Zhengzhou Railway Vocationaland Technical College, Zhengzhou 450001, China)

**Abstract**: The deep Q-learning network (DQN) has become an effective method to solve the traffic signal timing problem because of its strong perception and decision-making ability. However, in the field of traffic signal timing systems, the problem of parameter uncertainty caused by external environment disturbance and internal parameter fluctuation limits its further development. Based on this, a traffic signal timing method combining DQN and extended Kalman filter (DQN-EKF) is proposed. In this method, the uncertain parameters of the estimated network are taken as the state variables, and the target network values with uncertain parameters are taken as the observed variables. The EKF system equation is constructed by combining the process noise, the estimated network values with uncertain parameters and the system observation noise. The optimal estimation of the parameters in the DQN model is obtained through the iterative updating of the EKF Uncertainty. The experimental results show that the DQN-EKF timing algorithm is suitable for different traffic environments and can effectively improve the traffic efficiency of vehicles.

**Keywords**: deep Q-learning network (DQN); perception ability; decision making ability; traffic signal timing system; parameter uncertainty; extended Kalman filter (EKF)

**Received**: 2021-09-06; **Accepted**: 2021-09-17; **Published online**: 2021-10-12 13:38

**URL**: kns.cnki.net/kcms/detail/11.2625.V.20211012.1106.001.html

**Foundation items**: National Natural Science Foundation of China (61973103); Soft Science Research Program of Henan Province (212400410005)

* **Corresponding author**. E-mail: wulan@haut.edu.cn

http://bhxb.buaa.edu.cn jbuaa@buaa.edu.cn
**DOI**: 10.13700/j.bh.1001-5965.2021.0532

# 基于改进大气散射模型的单幅图像去雾方法

杨勇[1], 邱根莹[1], 黄淑英[2,*], 万伟国[3], 胡威[1]

(1. 江西财经大学 信息管理学院, 南昌 330032; 2. 天津工业大学 软件学院, 天津 300387;
3. 江西财经大学 软件与物联网工程学院, 南昌 330032)

**摘 要**: 雾天情况下获得的图像通常会出现对比度低、色彩丢失及噪声等问题,传统的去雾方法主要着眼于解决对比度低、色彩损失等问题,而没有考虑空气中灰尘颗粒散射隐藏的噪声光,导致去雾结果中易出现大量的噪声。针对该问题,提出了一种基于改进大气散射模型的单幅图像去雾方法。结合雾霾天气的特点,通过增加空气中介质散射的噪声光对传统雾天成像的大气散射模型进行改进;针对暗通道先验计算透射率不准确的问题,根据改进的模型构建一种透射率精细化的求取方法;结合全变分模型保边抑噪的思想,构造一种新的目标函数,迭代求解获得去雾图像。实验结果和对比分析表明:所提方法能有效去除图像中的雾,减少去雾结果中的噪声,同时也能保留图像中丰富的纹理信息。

**关 键 词**: 图像去雾; 大气散射模型; 暗通道先验; 目标函数; 自适应权重
**中图分类号**: TP391.41
**文献标志码**: A **文章编号**: 1001-5965(2022)08-1364-12

雾霾天气下拍摄的图像通常认为由 2 部分光成像所得,即经过大气层后衰减的场景反射光和随场景深度变化的空气光。场景反射光为由场景反射到摄像头的光束,经过大气层时一部分光被空气中的粒子散射、吸收,导致这部分光线被衰减。空气光为飘浮的大气颗粒(如灰尘、薄雾和烟雾)对大气光进行散射被摄像头接收的光束。因此,雾霾天收集到的图像容易出现模糊、对比度降低及噪声等现象[1],严重影响了后续高级计算机视觉任务的精确性,如自动车辆导航、户外监控、遥感及物体识别等[2]。因此,恶劣环境下图像的处理工作对后续计算机视觉任务具有重要的意义。

近年来,研究者已经提出许多经典的去雾方法,并取得了较好的效果。现有的去雾方法大致可分为以下 2 类:

1) 传统的去雾方法[3-8]。传统的去雾方法包含基于图像增强的方法和基于物理模型的方法。基于图像增强的方法通常使用直方图均衡、Retinex 理论及伽马矫正等方法来实现图像去雾。例如,Galdran[3]提出了一种基于人工多曝光图像融合的去雾方法,利用伽马矫正及多尺度拉普拉斯混合方法得到清晰的无雾图像,该方法对噪声并不敏感,但应用在饱和度较高的图像上时容易导致偏色。目前,基于图像增强的方法由于没有考虑雾天条件下的成像原因,获得的结果通常会出现局部增强过度或增强不足及偏色等问题。基于物理模型的方法建立在大气散射模型的基础上,通过估计模型中的参数(如大气光值、透射率)反推物理模型,求解得到去雾图像。但在单幅图像去雾任务中,由于一些关键信息(如深度、颜色等)缺失严重,导致透射率估计困难,需要引入先验信息来解决信息缺失的问题。

**收稿日期**: 2021-09-06; **录用日期**: 2021-09-17; **网络出版时间**: 2021-09-28 19:12
**网络出版地址**: kns.cnki.net/kcms/detail/11.2625.V.20210928.1143.001.html
**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: shuyinghuang2010@126.com

**引用格式**: *杨勇, 邱根莹, 黄淑英, 等. 基于改进大气散射模型的单幅图像去雾方法[J]. 北京航空航天大学学报, 2022, 48(8): 1364-1375. YANG Y, QIU G Y, HUANG S Y, et al. Single image dehazing method based on improved atmospheric scattering model[J]. Journal of Beijing University of Aeronautics and Astronautics, 2022, 48(8): 1364-1375 (in Chinese).*

He 等[4]通过分析大量室外无雾图像,提出了一种暗通道先验去雾方法来估计透射率及大气光值,但该方法应用在图像中灰度强度与大气光值相似的区域时通常会低估这部分场景的透射率,从而导致恢复的场景出现过饱和及色彩偏暗的现象。针对这些问题,Lu 等[5]从 HSV 空间探索暗通道先验,发现了暗通道和饱和度、亮度之间的关系,提出了饱和度迭代的方法来求得合适的透射率,可以有效防止透射率被低估的情况。另外,针对暗通道先验方法中使用的最小值滤波可能导致去雾图像出现伪边缘的问题,Yang 和 Wang[6]提出了一种新的分段函数代替最小值滤波,来防止由于使用最小值滤波出现的伪边缘问题,由于暗通道先验假定无雾图像的暗通道为 0,会导致透射率被低估,该方法直接对无雾图像的暗通道进行估计避免了这一问题。Meng 等[7]基于对传输函数固有边界约束的探索,提出了一种边界约束与上下文正则化相结合的单幅图像去雾方法。Berman 和 Avidan[8]提出了一种基于非局部先验的去雾方法,利用雾线得到了清晰的无雾图像。

上述基于物理模型的方法考虑了已有大气散射模型中光线衰减对图像的影响,虽然能实现一定程度的图像去雾,但是针对雾浓度较大的图像得到的结果通常会存在一定程度的噪声和色偏现象[4-8],尤其是在天空区域。这是因为对大气散射模型中的透射率评估不准确,以及雾天成像没有考虑到雾气浓度较高时,空气中的灰尘颗粒引起的散射在成像过程中会给图像带来一些隐含的噪声。这就导致去雾后图像细节被还原的同时噪声也被放大,使得去雾结果主观效果不佳。

2) 基于深度学习的去雾方法[9-14]。近年来,由于深度学习强大的特征学习能力,许多研究者纷纷将深度学习用于处理和解决计算机视觉领域中的问题。Liu 等[9]提出了一种端到端的网格去雾网络(GridDehazeNet)用于单幅图像去雾,该网络由预处理模块、主干模块及后处理模块组成,可以有效提取和融合不同尺度的特征信息。Chen 等[10]提出了端到端的门控上下文聚合网络(GCANet),使用平滑膨胀卷积有效去除膨胀卷积导致的网格伪影,并提出了门控子网络融合不同层次的特征。Hong 等[11]提出了利用异质任务模拟的知识提取去雾网络(KDDN),采用过程导向学习机制实现图像重建任务。Dong 等[12]提出了一种基于密集特征融合的多尺度增强去雾网络(MSBDN),通过基于 U-Net 的编码-解码结构实现单幅图像的去雾。Huang 等[13]提出了一种简单而有效的去雾网络,通过学习有雾图像与无雾图像之间的残差图来实现图像去雾。Yang 等[14]提出了一个聚合多尺度特征图的去雾网络(Y-net)来重建清晰的图像。Dong 等[15]构建了一个端到端的有融合鉴别器的生成对抗网络(FD-GAN),以频率信息为附加先验信息来得到更自然的去雾图像。Wu 等[16]提出了一个基于自动编码器框架的紧凑去雾网络(AECR-Net),以及基于对比学习的正则化方法来更好地利用有雾图像和无雾图像的信息。虽然基于深度学习的去雾方法在大多数情况下能取得较好的去雾效果,但依赖于大量的训练数据。然而,由于真实的训练集往往难以获得,目前基于深度学习的去雾方法在合成雾图上表现良好,但在真实有雾图像上的表现不能获得令人满意的效果[17]。另外,目前使用的合成训练集大都利用大气散射模型生成,没有考虑隐含噪声光的影响,因此在处理浓雾图像时不能达到满意的效果。

基于上述分析,本文针对传统方法去雾不彻底及图像中噪声被放大的问题,提出了一种基于改进大气散射模型的单幅图像去雾方法。首先,考虑雾天灰尘颗粒产生的散射光影响,对传统的大气散射模型进行改进;然后,为了获取更精确的透射率,基于暗通道先验,构建一种精细化透射率的求取方法;最后,根据改进的大气散射模型,构建一个新的目标函数,求解得到清晰的去雾图像,在函数的求解过程中,结合图像纹理区域和雾气浓度的判断,定义自适应权值,具有保留边缘和抑制噪声的作用。通过与 Galdran[3]、He[4]、Yang[6]、Meng[7]、Berman[8]等提出的方法进行对比,验证了本文方法在有效去雾的同时能够更好地抑制噪声和保留边缘,所恢复的图像更加清晰和自然。

## 1　本文方法

本文针对图像去雾不彻底及去雾后图像存在噪声被放大的问题,提出了一种基于改进大气散射模型的去雾方法,具体流程如下:首先,针对现有大气散射模型存在的问题进行改进;然后,针对模型中的参数透射率进行估计,利用暗通道的求解方法及雾的感知密度评估,进一步细化透射率的求解;最后,根据改进的大气散射模型构建新的目标函数,并利用求解得到的去雾图像初始值、雾的感知密度及透射率来估算目标函数中的参数,求解得到清晰的去雾图像。图 1 展示了本文去雾方法的整体流程。

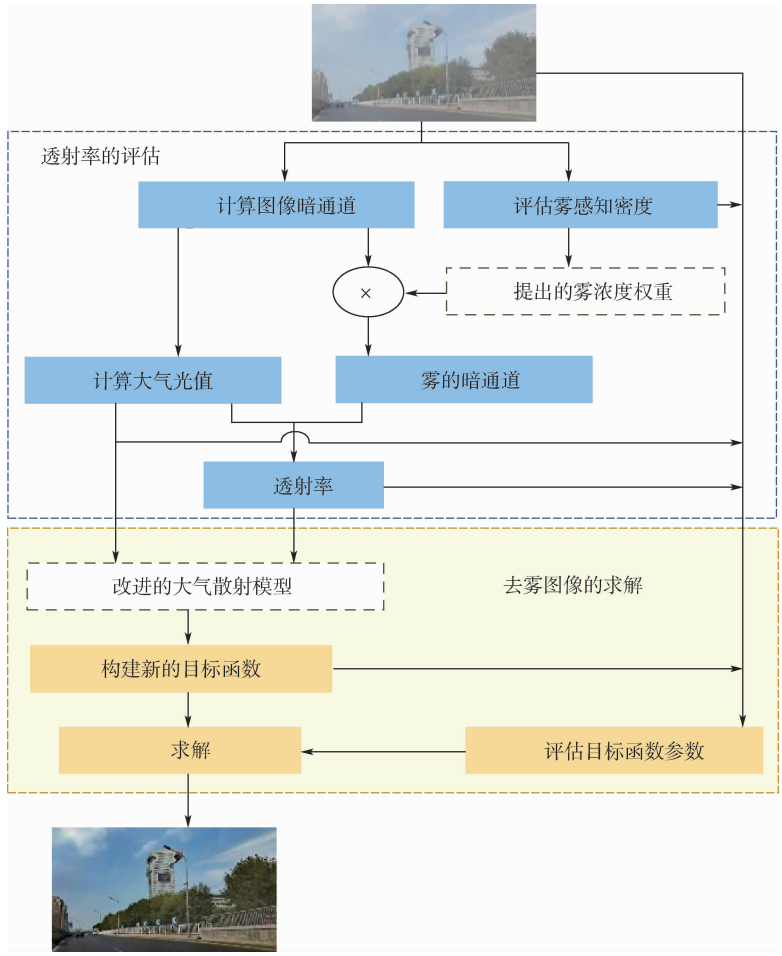



图 1 本文去雾方法流程

Fig. 1 Flow chart of the proposed dehazing method

### 1.1 改进的大气散射模型

大气中的雾气或微尘颗粒对光的吸收和散射是造成雾霾产生的主要原因[18],因此,雾天拍摄的图像通常能见度都较低,许多信息被隐藏。1999 年,Nayar 和 Narasimhan[19] 建立了大气散射模型来解释雾天场景的成像过程,其定义可被简化为

$$\boldsymbol{I} = \boldsymbol{J}t + (1 - t)\boldsymbol{A} \tag{1}$$

式中:$\boldsymbol{I}$ 为获得的有雾图像;$\boldsymbol{J}$ 为原始清晰的去雾图像;$\boldsymbol{A}$ 为大气光值;$t$ 为透射率,其表达式为

$$t = \mathrm{e}^{-\beta d} \tag{2}$$

式中:$\beta$ 为与波长相关的参数;$d$ 为深度。

根据式(1)可知,大气散射模型将有雾图像分解为 2 部分,分别可看作场景光和空气光的成像。传统的去雾方法一般通过求解模型中的透射率及大气光值来获得去雾后的图像。而大气光值通常假设为常数,透射率则认为与场景的深度相关,这样获得的空气光会比真实的更简单,因为真实的空气光中还包含灰尘等各种粒子引起的噪声光。这部分噪声光在成像过程中往往被忽略,而在去雾过程中就会被放大,使得恢复的去雾图像存在大量的噪声点,导致主观效果不佳。针对这一问题,本文对传统的大气散射模型进行了改进,将空气中灰尘等介质散射的噪声光加入到模型中。改进后的模型可以表示为

$$\boldsymbol{I} = \boldsymbol{J}t + (1 - t)\boldsymbol{A} + (1 - t)\boldsymbol{N} \tag{3}$$

式中:$\boldsymbol{N}$ 为噪声矩阵,$(1-t)\boldsymbol{N}$ 为噪声光项。

根据改进的大气散射模型(3),一幅有雾图像接收的光可以简单地分为 3 个部分,即衰减的场景反射光 $\boldsymbol{J}t$、大气光的散射部分 $(1-t)\boldsymbol{A}$、图像中的噪声光部分 $(1-t)\boldsymbol{N}$。第 2、3 项可以认为共同构成了图像中的“雾”,其中第 2 项为雾的主要部分,记作雾层,第 3 项可以看作隐藏的噪声。

### 1.2 透射率估计

暗通道先验方法假定无雾图像暗通道值为 0,导致透射率被低估,因此,本文增加了一项雾气权重来解决由于透射率被低估产生的伪影问题。

#### 1.2.1 图像暗通道计算

暗通道先验方法中提到在非天空区域的局部区域中存在暗像素,这些暗像素的强度主要由空气光组成,因此,本文利用暗通道先验来初始估计

一幅暗通道像素图。初始估计的暗通道计算公式为

$$I_{\min} = \mathrm{RF}(\min(I_R, I_G, I_B)) \tag{4}$$

式中:RF(·)为对图像使用递归滤波[20]进行的平滑操作;$I_R$、$I_G$、$I_B$分别表示图像的红绿蓝三通道;$I_{\min}$为暗通道图。

原始的暗通道先验方法使用最小值滤波来求图像的暗通道,但最小值滤波会导致伪边缘出现,因此,本文直接对图像三通道的最小值使用保边性能好、快速且不需要引导图像的 RF 进行滤波,去除图像中的细节部分。

1.2.2　雾气权重计算

暗像素的强度主要由空气光形成,但实际成像中其也包含一定比例的场景光。因此,本文使用雾感知密度评估器[21](FADE)估计有雾图像 $\boldsymbol{I}$ 中的雾气浓度评估图,并根据此图评估暗通道 $I_{\min}$ 的权重图。

$$[\mathrm{density}, D_{\mathrm{map}}] = \mathrm{FADE}(\boldsymbol{I}) \tag{5}$$

式中:density 为图像雾浓度评估值;$D_{\mathrm{map}}$为图像 1.562 5% 尺寸的雾浓度评估图。

由 FADE 方法生成的 $D_{\mathrm{map}}$中包含缺失值,因此先对这些缺失值使用距离最近的非缺失值进行填充。图 2 展示了 2 幅有雾图像获得精细化透射率过程的中间结果。图 2(b)为有雾图像 $\boldsymbol{I}$ 所对应的雾浓度评估图 $D_{\mathrm{map}}$,图 2(c)为 $D_{\mathrm{map}}$被放大并被分段阈值处理的结果,记作 $D_{\mathrm{map1}}$。从图 2(b)可观察到,$D_{\mathrm{map}}$在整体上保持雾的浓度分布,浓雾区域的值较大,薄雾区域处的值较小,但是包含很多噪声。因此,本文对 $D_{\mathrm{map}}$进行两步细化处理得到更准确的权重 weight。

对 $D_{\mathrm{map}}$值进行分段阈值处理,因为雾的存在是人类感知深度的基本线索,所以去雾过程中需要保留部分远处的雾以获得更好的视觉效果。因此,本文选择一个较大的阈值 $\delta$ 来对 $D_{\mathrm{map}}$值做分段阈值处理。设置一个固定值 $\delta$,大于 $\delta$ 的部分为浓雾区域,$D_{\mathrm{map}}$的值置为 $\delta$;小于 $\delta$ 的部分则认为 $D_{\mathrm{map}}$随场景的深度变化,$D_{\mathrm{map}}$的值保持不变。因此,$D_{\mathrm{map}}$映射到 $D_{\mathrm{map1}}$的操作可定义为

$$D_{\mathrm{map1}} = \begin{cases} D_{\mathrm{map}} & D_{\mathrm{map}} < \delta \\ \delta & D_{\mathrm{map}} \geqslant \delta \end{cases} \tag{6}$$

式中:$\delta = 0.85$ 为经验值,其确定过程将在后续实验中给出。$D_{\mathrm{map}}$的尺寸为有雾图像的 1.562 5%,因此,使用最近邻插值将 $D_{\mathrm{map}}$放大到与有雾图像相同的尺寸,并进行分段阈值处理,结果如图 2(c)所示的 $D_{\mathrm{map1}}$。从图 2(c)可以观察到,$D_{\mathrm{map1}}$还包含较多的场景信息,因此使用 RF 滤波对 $D_{\mathrm{map1}}$进行平滑,获得与深度相关的权重信息:

$$\mathrm{weight} = \mathrm{RF}(D_{\mathrm{map1}}) \tag{7}$$

式中:weight 为求得的 $I_{\min}$中空气光的权重。

1.2.3　透射率计算

经过计算得到图像的暗通道 $I_{\min}$和雾气权重 weight,通过 weight 对 $I_{\min}$进行加权,计算得到雾层暗通道 $\mathrm{fog}_{\min}$:

$$\mathrm{fog}_{\min} = \mathrm{weight} \cdot I_{\min} \tag{8}$$

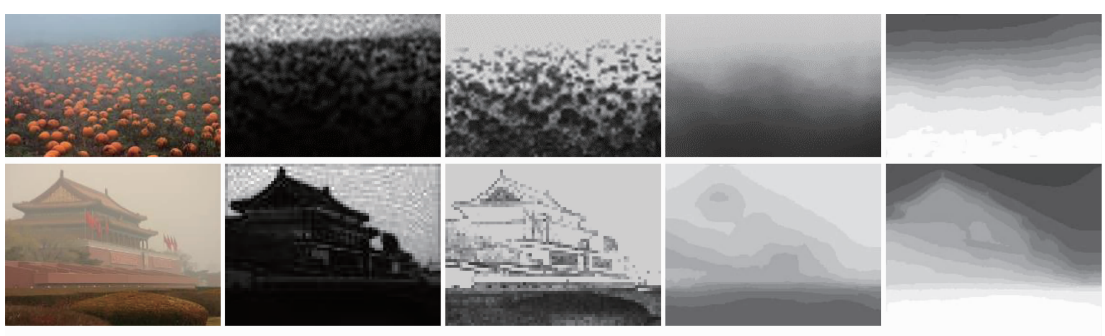
对大气光值 $A$ 进行评估。找到 $I_{\min}$前 0.1% 的像素,并分别找到有雾图像中对应位置的像素值,将 3 个通道的像素均值作为大气光值 $\boldsymbol{A}$。计算大气光值的最小值 $A_{\min}$,将其代入改进模型(3)中的第2 项,可得式(9)。联合式(8)已获取的 $\mathrm{fog}_{\min}$,对式(9)进行变形,即可求得更精细的透射率,如式(10)所示。

$$\mathrm{fog}_{\min} = A_{\min}(1 - t) \tag{9}$$

$$t = 1 - \frac{\mathrm{fog}_{\min}}{A_{\min}} \tag{10}$$

图 2(e)展示了细化的透射率 $t$。

从改进的大气散射模型(3)中可知,要求取去雾图像 $\boldsymbol{J}$,还需求得另外 3 个未知量,即透射率 $t$、大气光值 $\boldsymbol{A}$ 及噪声项 $\boldsymbol{N}$。通过上述过程已求得了 $t$、$\boldsymbol{A}$,下面将构造一个新的目标函数来同时求解去雾图像 $\boldsymbol{J}$ 和噪声项 $\boldsymbol{N}$。



(a) 有雾图像$\boldsymbol{I}$　(b) $D_{\mathrm{map}}$　(c) $D_{\mathrm{map1}}$　(d) weight　(e) $t$

图 2　透射率及中间结果

Fig. 2　Transmittance and intermediate results

### 1.3 目标函数的构造

在 1.2 节已获得精细化的透射率 $t$ 和大气光值 $\boldsymbol{A}$,针对模型(3)中还存在的去雾图像 $\boldsymbol{J}$ 和噪声 $\boldsymbol{N}$ 两个未知数,构造了一个新的目标函数,利用交替优化及全变分[22]的求解方法得到清晰的无雾图像。对模型(3)进行变形,得到

$$\boldsymbol{J}+\frac{(1-t)\boldsymbol{N}}{t}-\frac{\boldsymbol{I}-(1-t)\boldsymbol{A}}{t}=0 \tag{11}$$

因为 $t$、$\boldsymbol{I}$ 和 $\boldsymbol{A}$ 是已知的,所以式(11)中的第 3 项是一个常数,式(11)又可以写为

$$\boldsymbol{J}+\frac{(1-t)\boldsymbol{N}}{t}-\boldsymbol{J}_0=0 \tag{12}$$

式中:$\boldsymbol{J}_0=\dfrac{\boldsymbol{I}-(1-t)\boldsymbol{A}}{t}$。

根据式(12),通过增加 2 个未知参数 $\boldsymbol{J}$、$\boldsymbol{N}$ 的约束项,可以构建一个新的关于未知参数的目标函数,其定义如下:

$$\min_{\boldsymbol{J},\boldsymbol{N}}\frac{1}{2}\iint\left\|\boldsymbol{J}+\frac{(1-t)\boldsymbol{N}}{t}-\boldsymbol{J}_0\right\|_2^2\mathrm{d}x\mathrm{d}y+w\iint|\nabla\boldsymbol{J}|\mathrm{d}x\mathrm{d}y+\frac{1}{2}\iint|\boldsymbol{N}|^2\mathrm{d}x\mathrm{d}y \tag{13}$$

式中:第 1 项为数据保真项;第 2 项为对 $\boldsymbol{J}$ 的约束项,具有保留边缘、抑制噪声的作用,$w$ 为权重,平衡该项抑制噪声的程度,$\nabla\boldsymbol{J}$ 为去雾图像的梯度;第 3 项为对 $\boldsymbol{N}$ 的约束项。从式(13)可知,除了需求的变量 $\boldsymbol{J}$、$\boldsymbol{N}$ 以外,目标函数中还存在未知参数 $w$,下面将给出 $w$ 的评估。

#### 1.3.1 参数 $w$ 的求取

参数 $w$ 起着约束第 2 项抑制噪声程度的作用,如果 $w$ 取固定值,则说明该项对图像中的每个像素点处理强度是相同的。但是,图像中包含平滑区域、纹理区域及噪声区域,如果对不同的区域采用同样的权值,则会造成图像纹理丢失或噪声抑制不彻底等问题。因此,本文针对不同的区域提出一种自适应的权值求解方法。参数 $w$ 的定义过程如下。

由于差分曲率[23]可以反映图像的不同区域,图像纹理区域的差分曲率值较大,平滑区域及孤立噪声点的差分曲率值较小。因此,可以利用差分曲率的值来判断图像的不同区域。差分曲率的计算公式如下:

$$\begin{cases}J^i_{0_{\eta\eta}}=\dfrac{(J^i_{0_x})^2J^i_{0_{xx}}+2J^i_{0_x}J^i_{0_y}J^i_{0_{xy}}+(J^i_{0_y})^2J^i_{0_{yy}}}{(J^i_{0_x})^2+(J^i_{0_y})^2}\\ J^i_{0_{\xi\xi}}=\dfrac{(J^i_{0_y})^2J^i_{0_{xx}}-2J^i_{0_x}J^i_{0_y}J^i_{0_{xy}}+(J^i_{0_x})^2J^i_{0_{yy}}}{(J^i_{0_x})^2+(J^i_{0_y})^2}\end{cases} \tag{14}$$

$$d_i=\left|\left|J^i_{0_{\eta\eta}}\right|-\left|J^i_{0_{\xi\xi}}\right|\right| \tag{15}$$

式中:$d_i$ 为 $\boldsymbol{J}_0$ 中像素点 $i$ 处的曲率差分值;$J^i_{0_x}$、$J^i_{0_y}$ 和 $J^i_{0_{xx}}$、$J^i_{0_{yy}}$、$J^i_{0_{xy}}$ 分别为 $\boldsymbol{J}_0$ 中像素点 $i$ 处对 $x$、$y$ 方向的一阶和二阶偏导;$J^i_{0_{\eta\eta}}$ 和 $J^i_{0_{\xi\xi}}$ 分别为 $\boldsymbol{J}_0$ 在梯度方向和垂直于梯度方向上的二阶导数。

根据差分曲率特性,可以利用差分曲率值来控制权值的变化,有效检测图像纹理区域,减小约束项对该区域的作用,对平滑区域及孤立的噪声点则增大该约束项的作用。因此,根据差分曲率值的变化趋势,权值可以定义为关于差分曲率值单调递减的函数,再结合雾的浓度变化,权重 $w$ 的公式为

$$w_i=\left[(1-t)\mathrm{e}^{-d_i t}\right]^{|\gamma-\mathrm{density}|}\qquad w_i\in(0,1) \tag{16}$$

式中:$w_i$ 为像素 $i$ 处的权重值;$\gamma$ 的值根据经验选取为 5,其具体选取过程将在实验部分给出。

由于图像的噪声与纹理都属于高频成分,在抑制噪声的同时容易造成纹理信息的丢失,通过参数 $w$ 来划分噪声严重的天空区域,在目标函数求解过程中加强对天空区域的去噪,削弱纹理部分的去噪。式(16)中,$t$ 及 $1-t$ 控制在深度大的区域,$w$ 值更大,$d_i$ 控制在图像纹理区域,减小 $w$ 的值,$\gamma-\mathrm{density}$ 控制 $w$ 值在图像整体雾气越浓时 $w$ 值越大。通过对式(16)分析可知,当透射率 $t$ 较小、曲率差分值较小、density 值较大时,即场景的深度就大,包含的纹理较少,雾气就较浓,此时的 $w_i$ 值也就大。

#### 1.3.2 求解目标函数

由式(12)可知,目标函数包含变量 $\boldsymbol{N}$、$\boldsymbol{J}$,因此对该目标函数采用交替优化的方法求解 2 个参数。首先,固定 $\boldsymbol{N}$ 值,求解关于 $\boldsymbol{J}$ 的函数,式(17)可看成简单的全变分去噪模型。然后,再固定 $\boldsymbol{J}$,求解关于 $\boldsymbol{N}$ 的函数。关于 $\boldsymbol{J}$ 的函数定义如下:

$$F_1=\frac{1}{2}\iint\left\|\boldsymbol{J}+\frac{(1-t)\boldsymbol{N}}{t}-\boldsymbol{J}_0\right\|_2^2\mathrm{d}x\mathrm{d}y+w\iint|\nabla\boldsymbol{J}|\mathrm{d}x\mathrm{d}y \tag{17}$$

式(17)为泛函,该泛函的核为

$$f_1=\frac{1}{2}\left[\boldsymbol{J}+\frac{(1-t)\boldsymbol{N}}{t}-\boldsymbol{J}_0\right]^2+w(J_x^2+J_y^2)^{\frac{1}{2}} \tag{18}$$

式中:$J_x$、$J_y$ 分别为 $\boldsymbol{J}$ 对 $x$、$y$ 的偏导。

泛函取得极值的必要条件为满足以下欧拉-拉格朗日方程:

$$\frac{\partial f_1}{\partial\boldsymbol{J}}=\frac{\partial\left(\dfrac{\partial f_1}{\partial J_x}\right)}{\partial x}+\frac{\partial\left(\dfrac{\partial f_1}{\partial J_y}\right)}{\partial y} \tag{19}$$

通过式(19)得到 $\boldsymbol{J}$ 的迭代公式为

$$J_t=\frac{\partial\left(\dfrac{\partial f_1}{\partial J_x}\right)}{\partial x}+\frac{\partial\left(\dfrac{\partial f_1}{\partial J_y}\right)}{\partial y}-\frac{\partial f_1}{\partial\boldsymbol{J}} \tag{20}$$

$$J_t = w\left[\frac{(J_x^{k-1})^2 J_{yy}^{k-1} + (J_y^{k-1})^2 J_{xx}^{k-1} - 2J_x^{k-1}J_y^{k-1}J_{xy}^{k-1}}{[(J_x^{k-1})^2 + (J_y^{k-1})^2]^{\frac{3}{2}}}\right] - \left[J^{k-1} + \frac{(1-t)N^{k-1}}{t} - \boldsymbol{J}_0\right] \tag{21}$$

$$J^k = J^{k-1} + \mathrm{st} \cdot J_t \tag{22}$$

式中：$J_y^{k-1}$、$J_x^{k-1}$ 和 $J_{yy}^{k-1}$、$J_{xx}^{k-1}$、$J_{xy}^{k-1}$ 分别为 $J^{k-1}$ 对 $y$、$x$ 的一阶和二阶偏导；$J_t$ 为梯度下降流；st 为步长，其值为 1。

式(22)为 $\boldsymbol{J}$ 的迭代公式，$k$ 为迭代次数。将 $\boldsymbol{J}$ 看作常数，最小化式(23)来求解 $\boldsymbol{N}$：

$$F_2 = \frac{1}{2}\iint\left\|\boldsymbol{J} + \frac{(1-t)\boldsymbol{N}}{t} - \boldsymbol{J}_0\right\|_2^2 \mathrm{d}x\mathrm{d}y + \frac{1}{2}\iint|\boldsymbol{N}|^2\mathrm{d}x\mathrm{d}y \tag{23}$$

泛函(23)的核为

$$f_2 = \frac{1}{2}\left\|\boldsymbol{J} + \frac{(1-t)\boldsymbol{N}}{t} - \boldsymbol{J}_0\right\|_2^2 + \frac{1}{2}|\boldsymbol{N}|^2 \tag{24}$$

从式(24)可知，$f_2$ 恒大于 0。因此，当 $f_2$ 取最小值时，$F_2$ 也取极小值。式(24)的解可以看作是式(25)一系列解的组合：

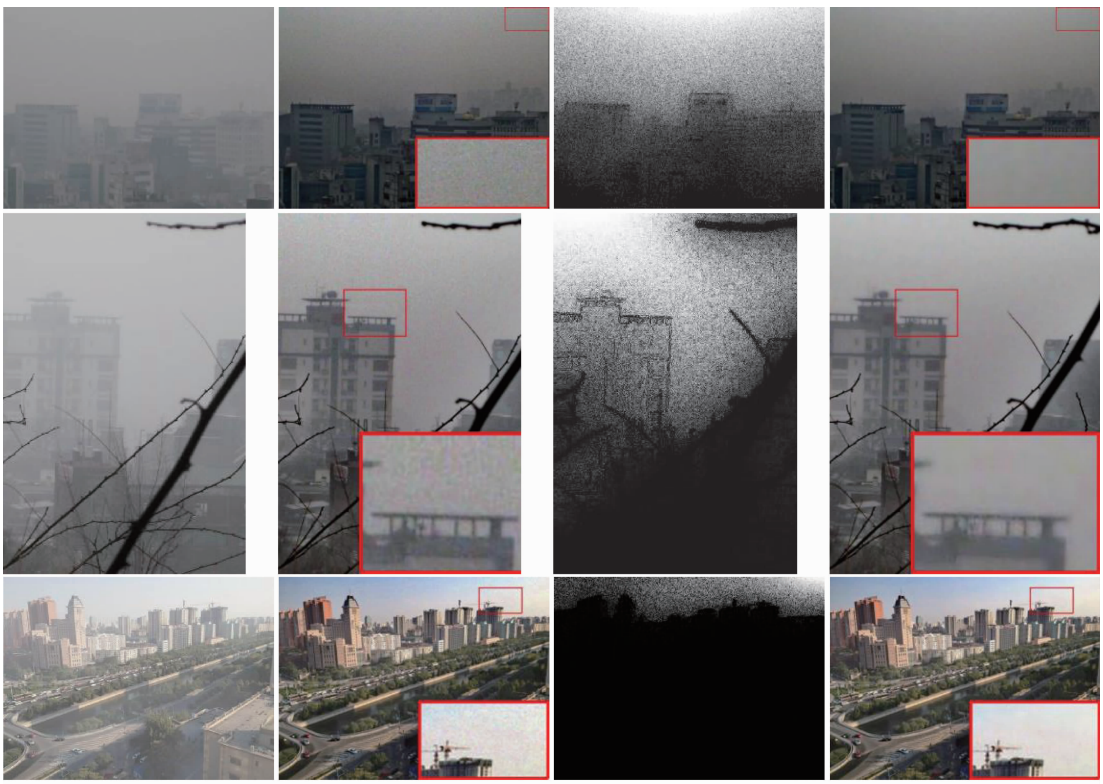
$$\min_n \frac{1}{2}\left(J_i + \frac{(1-t_i)n}{t_i} - J_{0_i}\right)^2 + \frac{1}{2}n^2 \tag{25}$$

式中：$i$ 为某个像素；$n$ 为 $i$ 位置对应的噪声。

通过式(25)可以得到求解 $n$ 的迭代公式：

$$n^k = \frac{\frac{1-t_i}{t_i}(J_{0_i} - J_i^k)}{\frac{(1-t_i)^2}{t_i^2} + 1} \tag{26}$$

求解目标函数的过程中，需要对式(22)、式(26)进行多次迭代，每次迭代得到第 $k$ 次的值 $J^k$、$N^k$。当 $J_t$ 趋于 0 时，目标函数(13)趋向得到极小值，但耗时较长。因此，为了提高方法运行速度，本文在保证方法效果的情况下设定一个较小的迭代次数 5。为了验证本文方法的有效性，图 3 展示了本文提出的改进大气散射模型及传统大气散射模型得到的去雾结果。其中，第 1、2 行为真实有雾图像结果，第 3 行为加入噪声的合成有雾图像结果。图 3(a)、(b)分别展示了有雾图像及其对应的原始大气散射模型去雾结果，图 3(c)、(d)分别展示了本文方法计算的权值 $w$ 和改进模型的去雾结果。从图中可观察到，直接利用透射率 $t$ 和大气光值 $\boldsymbol{A}$ 恢复的有雾图像结果中出现了噪声放大的现象，而本文方法可以在有效去雾的同时抑制噪声，证明了本文方法的有效性。



(a) 有雾图像　(b) 原始模型的去雾结果　(c) 权重$w$　(d) 改进模型的去雾结果

图 3　噪声图像去雾结果及中间结果

Fig. 3　Noise image dehazing results and intermediate results

## 2　实验结果与分析

本文方法在 3.19 GHz CPU，8 GB 内存的计算机上，MATLAB 2018b 的环境下运行。先讨论方法中参数值的选取，再对本文方法及文献[3-4, 6-8, 24]所提方法的主观效果进行比较，最后使用

Reside 数据集[25]中的 SOTS 数据及 HSTS 数据对本文方法及其他方法进行客观评估。

### 2.1 参数值选取

对于本文方法中式(6)和式(16)中 2 个参数 $\delta$、$\gamma$ 值的选取,本文采用实验的方法进行确定。先固定 $\gamma=5$,分别对参数 $\delta=0.7,0.75,0.8,0.85,0.9,0.95$ 时的结果进行对比。图 4 展示了 2 幅真实有雾图像在不同阈值 $\delta$ 时的去雾结果。可以看出,2 幅图像在 $\delta=0.7,0.75,0.8$ 时图像去雾结果都不彻底,边缘存在模糊现象。如第 1 行图像中放大区域所示,广告牌边缘及字体边缘模糊,对比度较低。第 2 行图像中屋檐下的立柱之间能够观察到明显有薄雾存在。

当 $\delta=0.9,0.95$ 时,第 1 行图像广告牌边缘包括整体建筑物部分亮度都变得更暗,第 2 组图像中放大区域所示屋檐边缘处明显出现颜色过饱和现象。尤其当 $\delta=0.95$ 时,亮度过暗导致第 1 行图像中广告牌字体开始模糊,同时第 2 行图像中天空区域出现明显的伪影。因此,本文选择 0.85 作为阈值 $\delta$。

固定 $\delta=0.85$,分别对 $\gamma=3,4,5,6$ 时的去雾效果进行主观效果的对比。图 5 展示了不同 $\gamma$ 值

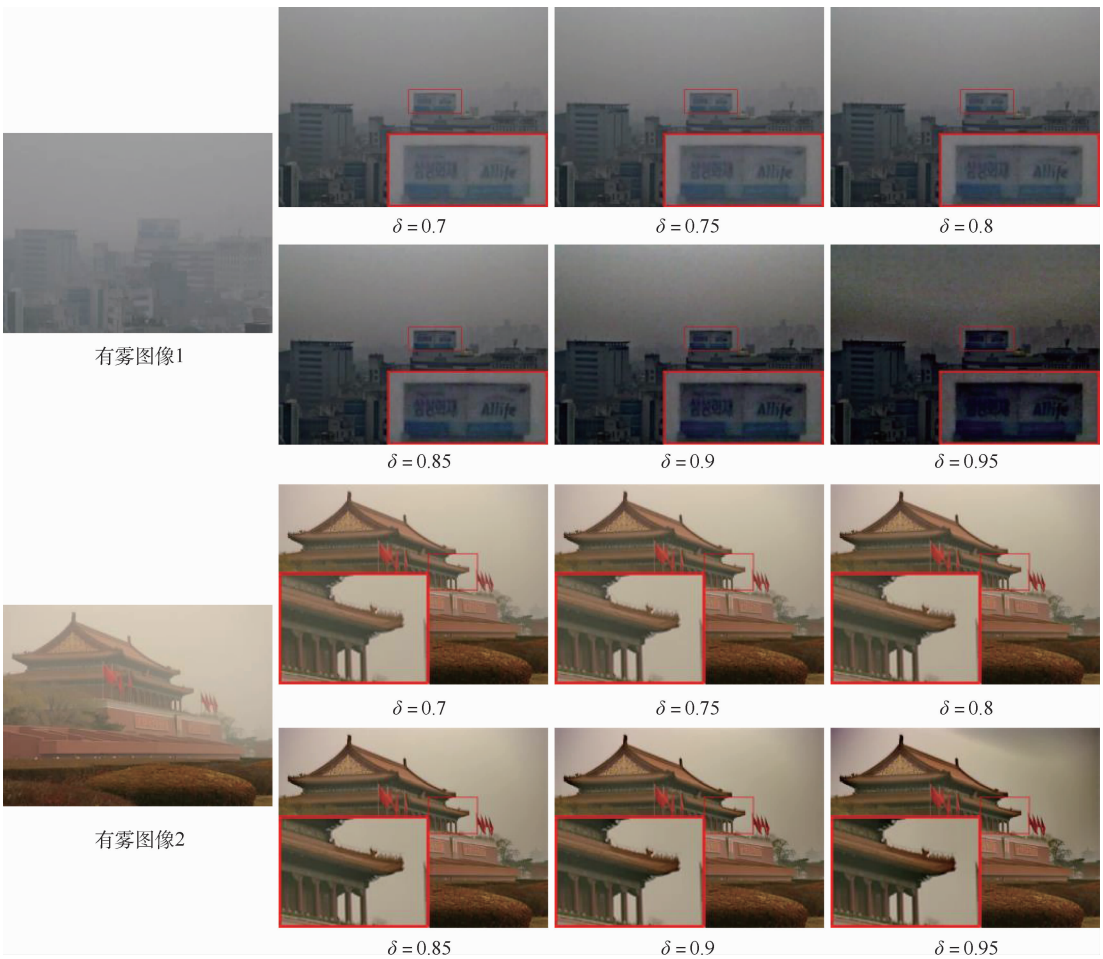



图 4 不同 $\delta$ 值的去雾结果

Fig. 4 Dehazing results with different $\delta$

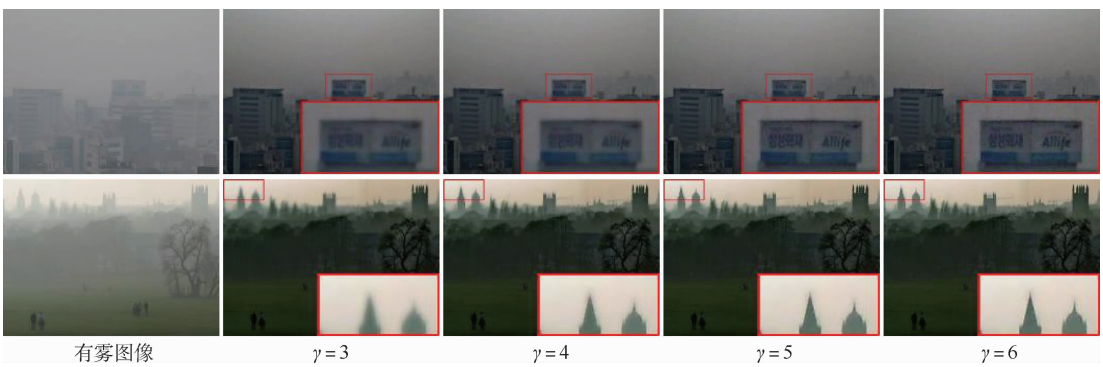



图 5 不同 $\gamma$ 值的去雾结果

Fig. 5 Dehazing results with different $\gamma$

的去雾效果。可以观察到，当 $\gamma=3,4$ 时，第 1 行图像中放大区域的广告牌边缘及广告牌中的字都比 $\gamma=5$ 时模糊，第 2 行图像中建筑物的顶端边缘都有不同程度的模糊。当 $\gamma=6$ 时，边缘保留的效果比 $\gamma=3,4,5$ 时效果好，但从图中可以观察到，噪声也比 $\gamma=3,4,5$ 时多，并且 $\gamma$ 值越大，计算量也越大。因此，从主观效果及计算量考虑，本文选择 5 作为 $\gamma$ 值。

## 2.2　去雾结果的主观分析

为了验证本文方法的有效性，展示了在真实有雾图像及合成有雾图像上应用本文方法及各种比较方法的结果，如图 6 和图 7 所示。可以观察到，Galdran[3] 方法在真实有雾图像 1 的前景处及真实有雾图像 2 的草地上的去雾结果出现色偏；He[4] 方法所获得真实有雾图像 1 的去雾结果中出现了过饱和的情况，对真实有雾图像 2 及合成有雾图像 2 的去雾结果在远景区域出现了严重的噪声；Yang[6] 方法对真实有雾图像 1 的去雾结果出现了色偏的问题，对真实有雾图像 2 的去雾结果中显示草地及南瓜的颜色也都存在色偏的问题；

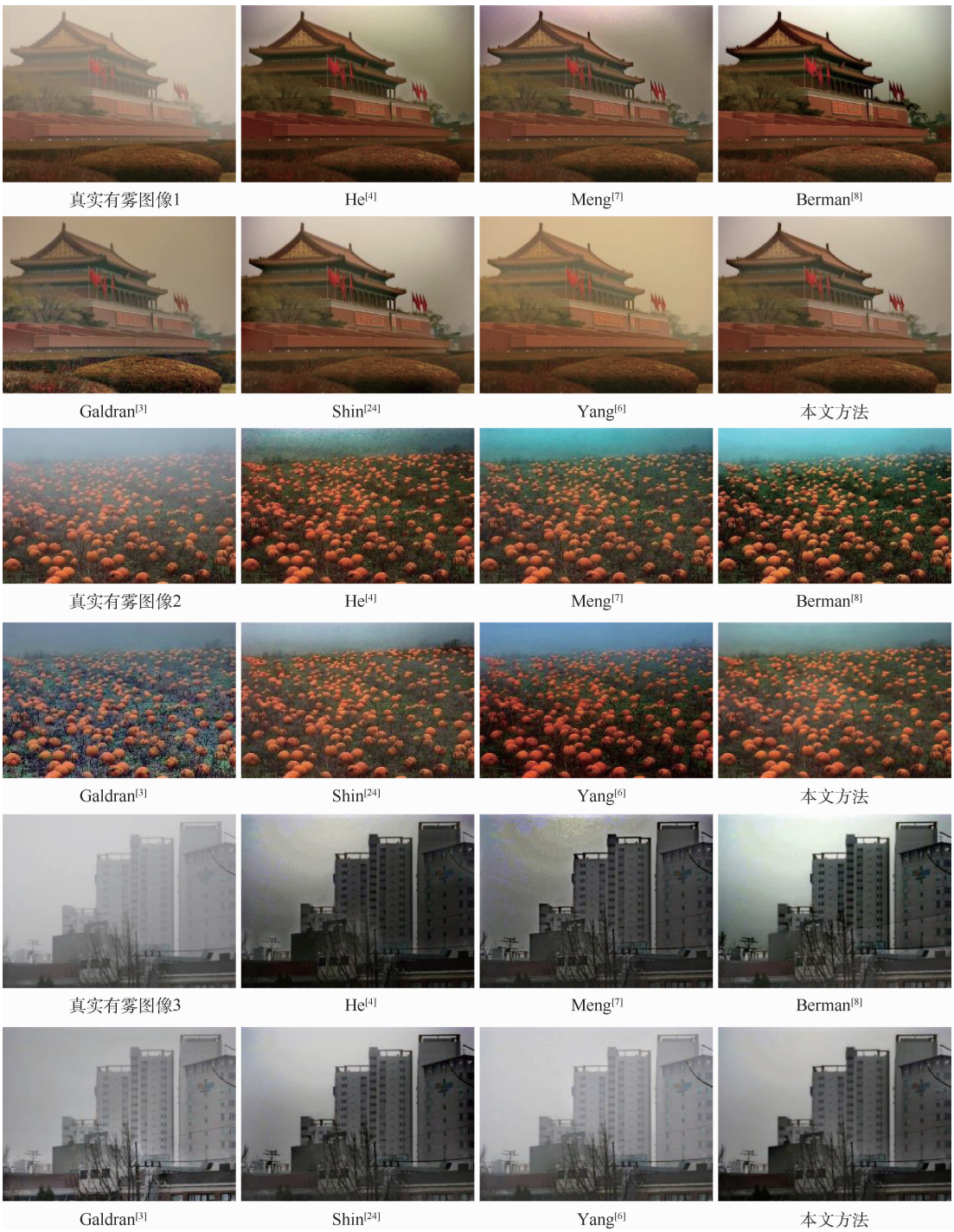



图 6　真实有雾图像去雾结果对比

Fig. 6　Comparison of dehazing results of real hazy images


图7 合成有雾图像去雾结果对比

Fig. 7 Comparison of dehazing results of synthetic hazy images

Meng[7]方法所获得的真实有雾图像1及合成有雾图像2的去雾结果中天空区域都有比较严重的噪声;Berman[8]方法在真实有雾图像1的去雾结果中非天空区域出现过饱和的情况,在真实有雾图像2上的去雾结果中浓雾处出现色偏问题,在合成有雾图像2上的结果中天空区域存在噪声;Shin[24]方法在合成有雾图像1上的去雾结果中前景处存在明显的色偏,在合成有雾图像2上的去雾结果中天空区域的左上角部分存在噪声。通过上述分析及图6、图7中的结果对比可知,本文方法可有效防止去雾过程中噪声被放大的情况。

### 2.3 去雾结果的客观评价

为了更客观地评价各种方法的性能,本文方法与其他比较方法分别在HSTS数据集、SOTS室外数据集及SOTS室内数据集上进行了实验。其中,HSTS数据集包含10幅合成有雾图像,SOTS室外数据集包含500幅室外合成有雾图像,SOTS室内数据集包含500幅室内合成有雾图像。本文采用结构相似性(SSIM)、峰值信噪比(PSNR)及色差(CIEDE2000)值对去雾结果进行评估。表1~表3分别展示了本文方法及其他去雾方法在HSTS、SOTS室外及室内数据集上的去雾结果的平均指标。

从表1~表3可以看出,本文方法在HSTS数据集上的SSIM及PSNR指标为最优,在SOTS室外、室内数据集上的3个指标均为最优,表明本文方法得到的结果更好地保持了原始图像的结构细节,同时与原始图像相比颜色失真最小。Yang[6]方法在HSTS数据集及SOTS室外数据集上多个指标为第二,但Yang[6]方法在SOTS室内数据集上却表现较差,这可能是由于该方法去雾后图像

**表 1　不同方法在 HSTS 数据集上的 SSIM、PSNR、CIEDE2000 值对比**

**Table 1　Comparison of SSIM, PSNR and CIEDE2000 values of different methods on HSTS dataset**

| 方法 | SSIM | PSNR | CIEDE2000 |
|---|---|---|---|
| He[4] | 0.738 4 | 14.86 | 13.00 |
| Meng[7] | 0.716 0 | 15.10 | 12.96 |
| Berman[8] | 0.768 8 | 17.51 | 10.68 |
| Galdran[3] | 0.788 9 | 17.04 | 11.41 |
| Shin[24] | 0.769 3 | 17.15 | 10.65 |
| Yang[6] | 0.800 7 | 18.21 | **8.99** |
| 本文方法 | **0.810 3** | **18.49** | 9.42 |

注:黑体数据表示最优结果。

**表 2　不同方法在 SOTS 室外数据集上的 SSIM、PSNR、CIEDE2000 值对比**

**Table 2　Comparison of SSIM, PSNR and CIEDE2000 values of different methods on SOTS outdoor dataset**

| 方法 | SSIM | PSNR | CIEDE2000 |
|---|---|---|---|
| He[4] | 0.753 7 | 14.65 | 13.90 |
| Meng[7] | 0.781 9 | 15.55 | 12.21 |
| Berman[8] | 0.802 2 | 18.06 | 10.25 |
| Galdran[3] | 0.832 0 | 17.93 | 10.33 |
| Shin[24] | 0.817 9 | 17.65 | 10.07 |
| Yang[6] | 0.827 2 | 18.59 | 9.40 |
| 本文方法 | **0.873 0** | **19.39** | **8.26** |

注:黑体数据表示最优结果。

**表 3　不同方法在 SOTS 室内数据集上的 SSIM、PSNR、CIEDE2000 值对比**

**Table 3　Comparison of SSIM, PSNR and CIEDE2000 values of different methods on SOTS indoor dataset**

| 方法 | SSIM | PSNR | CIEDE2000 |
|---|---|---|---|
| He[4] | 0.821 3 | 16.66 | 10.73 |
| Meng[7] | 0.793 8 | 17.04 | 10.41 |
| Berman[8] | 0.748 8 | 17.29 | 11.34 |
| Galdran[3] | 0.781 4 | 17.52 | 10.86 |
| Shin[24] | 0.808 9 | 18.46 | 9.17 |
| Yang[6] | 0.773 0 | 16.12 | 12.40 |
| 本文方法 | **0.884 8** | **21.13** | **6.40** |

注:黑体数据表示最优结果。

存在较重的偏色。He[4]方法在 3 个数据集上的 CIEDE2000 及 PSNR 指标值均较差,这可能与透射率被低估相关。Meng[7]方法在 HSTS 数据集上得到的 SSIM 值最差,在其余 2 个数据集上得到的 SSIM 值也较差,原因是该方法去雾后可能产生的噪声影响到了图像质量。通过上述客观对比实验,可以证明本文方法比其他去雾方法具有更好的性能。

### 2.4　消融实验

为了验证本文提出的透射率求取方法和新的目标函数在去雾方法中的有效性,使用 SOTS 室外数据集测试这 2 部分对去雾任务的作用,表 4 展示了消融实验结果。实验主要测试 4 个方法:基准、方法 1、方法 2、本文方法。基准方法为 He[4]方法;方法 1 为在 He[4]方法的基础上将透射率的求取方法替换为本文提出的透射率求取方法,目标函数不变;方法 2 为在 He[4]方法的基础上使用本文提出的目标函数及其优化求解方法,透射率求取方法不变。从表 4 中可看到,方法 1 和方法 2 相对于基准方法获得的 3 个指标值都得到了明显的提升,而本文方法获得的指标值都能达到最优,表明本文改进的透射率求取方法及新的目标函数都能对基准方法的效果起到提升作用。

**表 4　SOTS 室外数据集上的消融实验**

**Table 4　Ablation experiments on SOTS outdoor dataset**

| 方法 | 改进透射率求取 | 新的目标函数 | SSIM | PSNR | CIEDE2000 |
|---|---|---|---|---|---|
| 基准 | × | × | 0.753 7 | 14.65 | 13.90 |
| 方法 1 | √ | × | 0.868 5 | 19.38 | 8.29 |
| 方法 2 | × | √ | 0.813 3 | 15.76 | 11.59 |
| 本文方法 | √ | √ | **0.873 0** | **19.39** | **8.26** |

注:黑体数据表示最优结果。

## 3　结　论

1)本文针对雾天环境下大气散射模型所存在的问题,提出一种基于改进大气散射模型的单幅图像去雾方法。

2)对于模型中的透射率参数,通过评估雾的感知密度计算雾气权重值,并结合暗通道方法得到的大气光值来精细化求解透射率参数。

3)根据改进的大气散射模型,结合已获得大气光值和透射率参数,构建一个新的目标函数来求取去雾图像。

4)对于目标函数中的权重参数,本文针对图像的不同区域定义一种自适应的权值求解方法。

5)通过对来自 Reside 数据集中的 SOTS 数据和 HSTS 数据进行去雾实验,并将本文方法与先进的去雾方法进行对比,结果表明,本文方法在有效去雾的同时,能很好地抑制图像中存在的噪声。

### 参考文献(References)

[1] ZHANG J, TAO D C. FAMED-Net: A fast and accurate multi-scale end-to-end dehazing network[J]. IEEE Transactions on Image Processing, 2019, 29: 72-84.

[2] SALAZAR-COLORES S, CABAL-YEPEZ E, RAMOS-ARRE-

GUIN J M, et al. A fast image dehazing algorithm using morphological reconstruction[J]. IEEE Transactions on Image Processing, 2019, 28(5): 2357-2366.

[3] GALDRAN A. Image dehazing by artificial multiple-exposure image fusion[J]. Signal Processing, 2018, 149: 135-147.

[4] HE K M, SUN J, TANG X O. Single image haze removal using dark channel prior[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2010, 33(12): 2341-2353.

[5] LU Z W, LONG B Y, YANG S Q. Saturation based iterative approach for single image dehazing[J]. IEEE Signal Processing Letters, 2020, 27: 665-669.

[6] YANG Y, WANG Z W. Haze removal: Push DCP at the edge [J]. IEEE Signal Processing Letters, 2020, 27: 1405-1409.

[7] MENG G F, WANG Y, DUAN J Y, et al. Efficient image dehazing with boundary constraint and contextual regularization[C]// Proceedings of the IEEE International Conference on Computer Vision. Piscataway: IEEE Press, 2013: 617-624.

[8] BERMAN D, AVIDAN S. Non-local image dehazing[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2016: 1674-1682.

[9] LIU X H, MA Y R, SHI Z H, et al. GridDehazeNet: Attention-based multi-scale network for image dehazing[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE Press, 2019: 7314-7323.

[10] CHEN D D, HE M M, FAN Q N, et al. Gated context aggregation network for image dehazing and deraining[C]// 2019 IEEE Winter Conference on Applications of Computer Vision. Piscataway: IEEE Press, 2019: 1375-1383.

[11] HONG M, XIE Y, LI C H, et al. Distilling image dehazing with heterogeneous task imitation[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2020: 3462-3471.

[12] DONG H, PAN J S, XIANG L, et al. Multi-scale boosted dehazing network with dense feature fusion[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2020: 2157-2167.

[13] HUANG S Y, LI H X, YANG Y, et al. An end-to-end dehazing network with transitional convolution layer[J]. Multidimensional Systems and Signal Processing, 2020, 31(4): 1603-1623.

[14] YANG H H, YANG C H H, TSAI Y C J. Y-net: Multi-scale feature aggregation network with wavelet structure similarity loss function for single image dehazing[C]// IEEE International Conference on Acoustics, Speech and Signal Processing. Piscataway: IEEE Press, 2020: 2628-2632.

[15] DONG Y, LIU Y H, ZHANG H, et al. FD-GAN: Generative adversarial networks with fusion-discriminator for single image dehazing[C]// Proceedings of the AAAI Conference on Artificial Intelligence. Palo Alto: AAAI, 2020, 34(7): 10729-10736.

[16] WU H, QU Y, LIN S, et al. Contrastive learning for compact single image dehazing[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2021: 10551-10560.

[17] SHAO Y, LI L, REN W, et al. Domain adaptation for image dehazing[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2020: 2808-2817.

[18] WU Q B, ZHANG J G, REN W Q, et al. Accurate transmission estimation for removing haze and noise from a single image[J]. IEEE Transactions on Image Processing, 2019, 29: 2583-2597.

[19] NAYAR S K, NARASIMHAN S G. Vision in bad weather[C]// Proceedings of the IEEE International Conference on Computer Vision. Piscataway: IEEE Press, 1999, 2: 820-827.

[20] GASTAL E S L, OLIVEIRA M M. Domain transform for edge-aware image and video processing[M]. New York: ACM, 2011: 1-12.

[21] CHOI L K, YOU J, BOVIK A C. Referenceless prediction of perceptual fog density and perceptual image defogging[J]. IEEE Transactions on Image Processing, 2015, 24(11): 3888-3901.

[22] RUDIN L I, OSHER S, FATEMI E. Nonlinear total variation based noise removal algorithms[J]. Physica D: Nonlinear Phenomena, 1992, 60(1-4): 259-268.

[23] CHEN Q, MONTESINOS P, SUN Q S, et al. Adaptive total variation denoising based on difference curvature[J]. Image and Vision Computing, 2010, 28(3): 298-306.

[24] SHIN J, KIM M, PAIK J, et al. Radiance-reflectance combined optimization and structure-guided $l_0$-norm for single image dehazing[J]. IEEE Transactions on Multimedia, 2019, 22(1): 30-44.

[25] LI B Y, REN W Q, FU D P, et al. Benchmarking single-image dehazing and beyond[J]. IEEE Transactions on Image Processing, 2018, 28(1): 492-505.

# Single image dehazing method based on improved atmospheric scattering model

YANG Yong[1], QIU Genying[1], HUANG Shuying[2,*], WAN Weiguo[3], HU Wei[1]

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

**Abstract**: Images obtained in foggy conditions often suffer from low contrast, color loss, and noise. At present, many traditional dehazing methods mainly focus on solving problems such as low contrast and color loss, but do not consider the hidden noise light scattered by dust particles in the air, resulting in a large amount of noise in the dehazing results. This work provides an image dehazing algorithm based on an enhanced atmospheric scattering model to address the mentioned problems. Firstly, according to the characteristics of haze, the traditional atmospheric scattering model of hazy imaging is improved by adding the noise light reflected by the medium in the air. Then, in order to address the transmission calculation inaccuracy problem for the dark channel prior, a refined calculation method of transmission is constructed according to the improved model. Finally, combined with the idea of edge preservation and noise suppression of the total variation model, a new objective function is constructed and solved iteratively to obtain the final defogging image. A large number of experimental results and comparative analyses show that the proposed method can effectively remove the haze in the image, reduce the noise in the dehazing results, and retain the rich texture information in the image.

**Keywords**: image dehazing; atmospheric scattering model; dark channel prior; objective function; adaptive weight

**Received**: 2021-09-06; **Accepted**: 2021-09-17; **Published online**: 2021-09-28 19:12

**URL**: kns.cnki.net/kcms/detail/11.2625.V.20210928.1143.001.html

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

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

http://bhxb.buaa.edu.cn jbuaa@buaa.edu.cn
**DOI**: 10.13700/j.bh.1001-5965.2021.0511

# 面向量化分块压缩感知的区域层次化预测编码

刘浩*, 郑浩然, 黄荣
(东华大学 信息科学与技术学院, 上海 201620)

**摘 要**: 在量化分块压缩感知的预测编码中,低参考价值的候选者将导致较差的率失真性能。为了高效地降低编码失真,提出了一种基于螺旋逐块扫描的区域层次化预测编码方法。在以同一采样率进行观测后,各块按由内向外的扫描次序进行预测与量化。当前观测矢量从上下文感知候选集中选取与之具有最小误差的反量化矢量,作为其预测矢量;根据层次相关性,所有块被划分到3种区域之一,通过块编码模型为不同区域设定自适应的质量因子,关键区域被赋予较大的质量因子。与现有的预测编码方法相比,所提方法综合利用了矢量之间的空域相关性和层次相关性,实验结果获得了至少0.12 dB的率失真增益。

**关 键 词**: 量化分块压缩感知; 预测编码; 层次相关性; 关键区域; 质量因子
**中图分类号**: TN919.8
**文献标志码**: A **文章编号**: 1001-5965(2022)08-1376-07

近年来,压缩感知成为一种新兴的稀疏信号处理技术。对于目标图像,随机投影获得少量的观测矢量,非线性算法进行高概率的重构[1-2]。为降低测量端的资源消耗,压缩成像采用分块压缩感知(BCS)来获取观测矢量,并通过平滑投影算法恢复图像[3-4]。观测矢量是实数值,随机投影不会带来真正的压缩,无法产生用于数据传输的二进制比特流[5]。若缺少量化和熵编码,图像压缩感知就无法评估率失真性能[6]。因此,研究人员对量化分块压缩感知(QBCS)给予了极大的关注,配套方案是执行逐块的标量量化[7]。

QBCS测量端首先对目标图像进行分块,采用块级观测矩阵对各块进行随机投影,获得各块的观测矢量,然后执行预测和量化。如果需要改变整幅图像的压缩比,测量端可调整采样率或质量因子[8]。分块过程虽然减轻了测量端的资源消耗,却忽略了各块之间的相关性。若当前块的原始像素与相邻块的原始像素具有较强的空域相关性,那么投影至低维空间后所得观测矢量之间的相关性依然较强。若能在量化之前消除观测矢量之间的相关性,那么率失真性能将得到提升。预测编码是求取当前观测矢量与相邻反量化矢量之间的最小残差,并针对残差执行量化与熵编码,以节省码流。

观测矩阵应与目标图像无关。Tran等[9]提出了一种双向预测方法,部分地固定了观测矩阵,同时需要对图像进行像素域的采集。为了获得信号的统计特征,Shi等[10]提出了一种利用卷积神经网络联合优化观测网络和重构网络的图像压缩感知框架,从训练图像中自适应地构建观测矩阵。然而,上述2种方法难以兼容面向QBCS的常用图像重构算法。

QBCS测量端的预测编码需要执行观测域的预测和逐块量化。Mun和Fowler[11]采用差分脉冲编码调制(DPCM)对自然图像进行预测编码,获取前一反量化矢量与当前观测矢量的差值。通

**收稿日期**: 2021-09-03; **录用日期**: 2021-09-17; **网络出版时间**: 2021-09-28 20:03
**网络出版地址**: kns.cnki.net/kcms/detail/11.2625.V.20210928.1203.002.html
**基金项目**: 国家自然科学基金(62001099); 中央高校基本科研业务费专项资金(2232021G-09)
* **通信作者**. E-mail: liuhao@dhu.edu.cn

**引用格式**: *刘浩, 郑浩然, 黄荣. 面向量化分块压缩感知的区域层次化预测编码[J]. 北京航空航天大学学报, 2022, 48(8): 1376-1382. LIU H, ZHENG H R, HUANG R. Region-hierarchical predictive coding for quantized block compressive sensing[J]. Journal of Beijing University of Aeronautics and Astronautics, 2022, 48(8): 1376-1382 (in Chinese).*

过扩展 DPCM 机制，Zhang 等[12]提出了空域方向预测编码（SDPC）方法，更多地利用了相邻块之间的空域相关性，观测矢量从 3 种方向模式的候选者中选择预测矢量。与 SDPC 不同，Li 等[13]引入了中值滤波预测量化（MFPQ）方法，当前块的预测矢量是相邻 3 个反量化矢量的中值加权，虽然重建质量略低，但 MFPQ 方法的计算复杂度远低于 SDPC 方法，且具有良好的抗误码能力。与传统的光栅扫描顺序相比，螺旋预测编码（SPC）方法可以减少图像边缘的影响[14]。近期，Chen 等[15]借鉴 MFPQ 的思想，根据渐近谱分析理论提出了自适应加权预测（AWP）方法，通过对 4 个相邻块进行线性加权来增加候选者的数量，取得了一定的率失真增益。这 4 种预测编码方法均执行基于相同质量因子的逐块标量量化。

现有的预测编码方法往往采用固定方向模式的预测机制，由于复杂度限制，难以继续增加候选者或方向模式的数量；层次相关性表征了逐块预测编码过程中不同区域的参考价值，目前尚无文献予以考虑。为了克服现有方法的不足，本文将预测编码中的空域相关性和层次相关性结合起来，在螺旋逐块扫描中执行上下文感知的预测机制，并通过块编码模型为不同区域分配合适的质量因子，以提高预测编码的率失真性能。

## 1　测量端的功能模块及矢量

在 QBCS 测量端，图 1 给出了主要的功能模块及矢量。BCS 观测模块将目标图像划分为不重叠的块，所有块采用同一采样率进行观测，获得各块的观测矢量；量化模块和反量化模块仅占用了少量的复杂度，生成预测矢量的预测模块消耗了主要的计算资源；熵编码模块用于生成二进制比特流，以供数据传输或存储；预测模块需要缓存反量化矢量。在收到比特流之后，重建端可执行典型的图像重构算法。

QBCS 测量端将目标图像 $x$ 划分为 $N$ 个不重叠的块，块尺寸是 $B^2$ 像素。块级观测矩阵 $\boldsymbol{\Phi}_B$ 是一个大小为 $M_B \times B^2$ 的高斯随机矩阵，$M_B$ 比 $B^2$ 小得多。图像级观测矩阵 $\boldsymbol{\Phi}$ 由沿对角线的多个 $\boldsymbol{\Phi}_B$ 组成，即 $\boldsymbol{\Phi} = \mathrm{diag}([\boldsymbol{\Phi}_B, \boldsymbol{\Phi}_B, \cdots, \boldsymbol{\Phi}_B])$。所有块按一定的次序编号，$i$ 表示块号，$i = 1, 2, \cdots, N$。$\boldsymbol{x}_i$ 表示图像中第 $i$ 块的原始信号。BCS 观测模块通过块级观测矩阵 $\boldsymbol{\Phi}_B$ 获取所有块的观测矢量，$\boldsymbol{y}_i$ 表示第 $i$ 块的观测矢量。当采样率 $S = M_B/B^2$ 时，$\boldsymbol{x}_i$ 由块级观测矩阵 $\boldsymbol{\Phi}_B$ 进行观测，从而产生第 $i$ 块的观测矢量 $\boldsymbol{y}_i$。块级观测过程可以表示为

$$\boldsymbol{y}_i = \boldsymbol{\Phi}_B \boldsymbol{x}_i \tag{1}$$

测量端只需存储 $\boldsymbol{\Phi}_B$，而不需要存储整个 $\boldsymbol{\Phi}$。在所有块以同一采样率进行观测后，测量端开始执行逐块的预测编码。由于相邻块之间的相关性，预测编码技术可以减少观测矢量的冗余。在测量端，$\tilde{\boldsymbol{y}}_i$ 表示图像 $x$ 中第 $i$ 块的反量化矢量；针对观测矢量 $\boldsymbol{y}_i$ 的预测，第 $i$ 块的候选集由其邻域中的一个或多个反量化矢量组成；$\tilde{\boldsymbol{y}}_i^{\mathrm{p}}$ 表示第 $i$ 块的预测矢量，其是从第 $i$ 块的候选集中选择的反量化矢量。因此，第 $i$ 块的预测残差 $\boldsymbol{d}_i$ 是观测矢量 $\boldsymbol{y}_i$ 与预测矢量 $\tilde{\boldsymbol{y}}_i^{\mathrm{p}}$ 的差值，其计算公式为

$$\boldsymbol{d}_i = \boldsymbol{y}_i - \tilde{\boldsymbol{y}}_i^{\mathrm{p}} \tag{2}$$

式中：$\boldsymbol{d}_i$ 为实数值，需要执行量化操作。由于随机投影，压缩感知的质量因子应该是细粒度的[16]。测量端根据质量因子对预测残差 $\boldsymbol{d}_i$ 进行量化，得到第 $i$ 块的量化值 $s_i$，然后对 $s_i$ 进行熵编码。对于后续的预测编码，$\tilde{\boldsymbol{d}}_i$ 表示第 $i$ 块的反量化残差。通过将 $\tilde{\boldsymbol{d}}_i$ 与预测矢量 $\tilde{\boldsymbol{y}}_i^{\mathrm{p}}$ 相加，反量化矢量 $\tilde{\boldsymbol{y}}_i$ 可表示为

$$\tilde{\boldsymbol{y}}_i = \tilde{\boldsymbol{d}}_i + \tilde{\boldsymbol{y}}_i^{\mathrm{p}} \tag{3}$$

如果当前块具有更好的候选集，其预测矢量将更接近当前观测矢量。基于候选集中的反量化矢量，预测编码需要计算当前观测矢量与每个反量化矢量之间的误差，预测矢量是候选集中具有最小误差的反量化矢量。

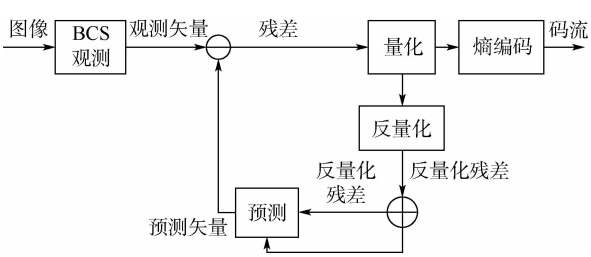



图 1　QBCS 测量端的功能模块及矢量

Fig. 1　Modules and vectors of QBCS measurement end

## 2　区域层次化预测编码

在所有块以同一采样率进行观测后，本节提出了一种基于螺旋逐块扫描的区域层次化预测编码（RHPC）方法，综合利用矢量之间的空域相关性和层次相关性，以取得更好的率失真性能。

### 2.1　上下文感知候选集

RHPC 方法从图像的中心块开始，按由内向外的扫描顺序对后续块进行预测编码。顺时针或逆时针的逐块扫描过程在统计意义上是等价的。

上下文感知候选集将从相邻块的反量化矢量中选择至多 2 个较重要的反量化矢量,其重要性准则包括:在逐块预测编码中,某块被编码得越早,对后续块的影响面更大,其就越相对重要;空域上更接近的相邻块更重要。当前块的邻域包括 8 个可能参考的相邻块,图 2 给出了上下文感知候选集的示意图。根据上述重要性准则,当前块的 8 个相邻块从 1 到 8 排序,其中标号越小,对于预测编码越重要。按照图 2 中标号递增和实际可选情况,第 $i$ 块从邻域的反量化矢量中选择至多 2 个较重要的反量化矢量,组成第 $i$ 块的上下文感知候选集 $P_i$。基于上下文感知和双方向模式,RH-PC 方法提高了候选者的匹配度。

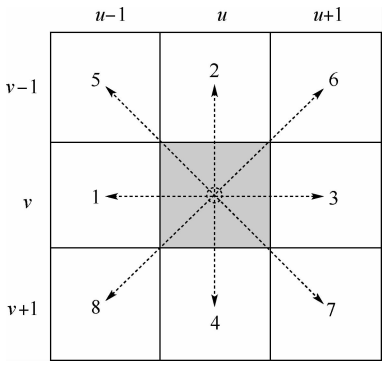



图 2 上下文感知候选集的示意图

Fig. 2 Schematic diagram of context-aware candidate set

通过最小化观测矢量 $\boldsymbol{y}_i$ 与上下文感知候选集 $P_i$ 中的反量化矢量 $\tilde{\boldsymbol{y}}_{P_i}$ 之间的误差,第 $i$ 块的预测矢量 $\tilde{\boldsymbol{y}}_i^{\mathrm{p}}$ 由式(4)确定:

$$\tilde{\boldsymbol{y}}_i^{\mathrm{p}} = \arg\min_{\tilde{\boldsymbol{y}}_{P_i}\in P_i} \|\boldsymbol{y}_i - \tilde{\boldsymbol{y}}_{P_i}\|_1 \tag{4}$$

式中:$\|*\|_1$ 为 $\ell_1$ 范式,其将矢量中所有条目的绝对值累加。第 $i$ 块的观测矢量 $\boldsymbol{y}_i$ 减去其预测矢量 $\tilde{\boldsymbol{y}}_i^{\mathrm{p}}$,获得预测残差 $\boldsymbol{d}_i$,其量化值为 $s_i$。当前块的预测矢量越匹配,则预测残差越小。量化值 $s_i$ 执行熵编码,生成比特流;同时,量化值 $s_i$ 被反量化,获得反量化残差 $\tilde{\boldsymbol{d}}_i$。

## 2.2 层次相关性

预测编码可以消除矢量之间的冗余。首先,相邻矢量在观测域具有空域相关性;其次,各块在图像中具有不同的层次相关性。在逐块预测过程中,一个反量化矢量可能被 0 ~ 8 个观测矢量所参考。图 3 给出了层次相关性的一个示例,其中 6 个观测矢量(块号分别为 4、11、12、13、14、15)的预测可利用块号为第 3 块的反量化矢量,5 个观测矢量(块号分别为 5、6、14、15、16)的预测可利用块号为第 4 块的反量化矢量。因此,图 3 中的第 3 块比第 4 块具有更强的层次相关性。

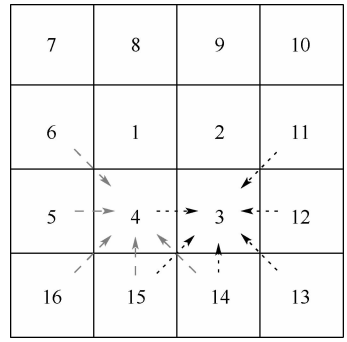



图 3 逐块预测过程中的层次相关性

Fig. 3 Hierarchical correlation during block-by-block measurement prediction

## 2.3 基于区域划分的质量因子

对于目标图像的所有观测矢量,螺旋逐块扫描的预测编码顺序呈现出由内向外的趋势。正如图 3 所示,第 3 块的反量化矢量可能将被 6 个观测矢量作为预测参考,而第 4 块的反量化矢量可能将被 5 个观测矢量作为预测参考,依此类推。当相同的反量化矢量可能预测更多的观测矢量时,反量化矢量所在的块就具有更强的层次相关性。因此,根据层次相关性,所有块被划分到 3 种区域之一:关键区域、非关键区域、分散区域。如果一个反量化矢量可能被至少 6 个观测矢量所预测参考,则该反量化矢量所在的块将被分类到关键区域;在这 6 个对应的观测矢量中,不属于关键区域的那些块将被分类到非关键区域;其他块属于分散区域。关键区域的块大多位于在预测编码顺序中更为关键的角点处。图 4 给出了 3 种区域的示例图,关键区域(黑色)、非关键区域(对角线)、分散区域(白色),其中块号 $i$ 按由内向外的螺旋扫描顺序从 1 到 100,较小的块号表示该块是较早编码的。

图像的不同区域包含差异化的层次信息,可以为各区域自适应地分配质量因子,更大的质量因子将导致较大的码率和较小的编码失真。关键区域具有很强的参考价值,有助于提高对应非关

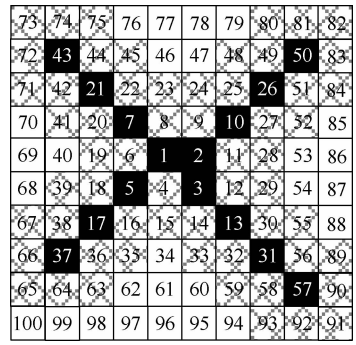



图 4 三种区域的示例

Fig. 4 Diagram of three regions

键区域的预测精度。因此，层次相关性较强的关键区域应赋予较大的质量因子。质量因子是整数值，$C_i$表示第 $i$ 块的观测矢量中所有分量的标准差，RHPC 方法不依赖于特定的比特率估计模型，使用式(5)所示的块编码模型来预测比特数[17]：

$$R_i^{\mathrm{m}}(Q) = \alpha C_i Q^{\beta} \tag{5}$$

式中：$R_i^{\mathrm{m}}(Q)$ 为第 $i$ 块的预测比特数；模型参数 $\alpha>0$ 和 $\beta<0$ 依赖于图像内容和缺省质量因子 $Q$。参数拟合所需的 $J$ 个样本($J\approx 0.2N$)来自于 $N$ 个观测矢量的随机抽取，这些样本采用缺省质量因子 $Q$ 进行量化，结合式(5)，模型参数 $\alpha$ 和 $\beta$ 通过式(6)求解：

$$\{\alpha,\beta\} = \operatorname{argmin}\sum_{j=1}^{J}\left(R_j^{\mathrm{a}} - R_j^{\mathrm{m}}(Q)\right)^2 \tag{6}$$

式中：$R_j^{\mathrm{a}}$ 和 $R_j^{\mathrm{m}}(Q)$ 分别表示第 $j$ 个样本($1\leqslant j\leqslant J$)的实际比特数和基于块编码模型的预测比特数。$Q^{\mathrm{n}}$ 表示非关键区域的质量因子，通过求解式(7)获得：

$$\sum_{i\in \Pi_{\mathrm{k}}\cup\Pi_{\mathrm{n}}} R_i^{\mathrm{m}}(Q) \approx \sum_{i_{\mathrm{k}}\in\Pi_{\mathrm{k}}} R_{i_{\mathrm{k}}}^{\mathrm{m}}([\tau Q]) + \sum_{i_{\mathrm{n}}\in\Pi_{\mathrm{n}}} R_{i_{\mathrm{n}}}^{\mathrm{m}}(Q^{\mathrm{n}}) \tag{7}$$

式中：$\Pi_{\mathrm{k}}$ 和 $\Pi_{\mathrm{n}}$ 分别表示关键区域和非关键区域；$i_{\mathrm{k}}$ 和 $i_{\mathrm{n}}$ 为对应区域的块号；由于量化因子是整数值，符号“[ ]”表示选取最接近的整数；$\tau$ 为经验值。至此，关键区域、非关键区域、分散区域分别采用质量因子$[\tau Q]$、$Q^{\mathrm{n}}$、$Q$ 逐块地执行量化和熵编码。由于 RHPC 方法将更大的质量因子分配给更重要的区域，因此可以保持图像的平均采样率，同时提高目标图像的重建质量。重建端在收到比特流后，将执行图像重构算法。

### 2.4　复杂度分析

本节分析各种预测编码方法的计算复杂度。在现有方法中，计算开销主要取决于式(4)求解最小误差的次数。由于均用到 3 种方向模式，SPC 与 SDPC 方法具有基本相当的复杂度。目标图像 $x$ 分为 $N$ 块，每一观测矢量的条目数量为 $M_B$，候选集中反量化矢量的数量为 $a$。预测编码的计算复杂度可表述为 $O(a\times N\times M_B)$，其中，MFPQ、SDPC、SPC、AWP 方法对应的 $a$ 值分别为 1、3、3、1，这 4 种方法的 $N\times M_B$ 是相同的。为了确定线性加权系数，AWP 方法需要执行大规模的离线训练过程，这种预处理的复杂度要远高于后续的预测编码过程[15]。式(6)中参数拟合需要 $J\approx 0.2N$ 个样本，其迭代次数为 $h$，计算复杂度可表述为 $O(h\times J)$；令 $g=0.2h$，参数拟合的计算复杂度为 $O(g\times N)$，其中 $g$ 值相对较低，RHPC 方法的计算复杂度可表述为 $O(a\times N\times M_B+g\times N)$，其中 $a=2$。通常，$M_B$ 远大于 $g$，因此 RHPC 方法的复杂度低于 SDPC 与 SPC 方法，但高于 MFPQ 与 AWP 方法。

## 3　实验结果

大量的实验用于比较 RHPC 方法和其他预测编码方法，如 SDPC[12]、MFPQ[13]、SPC[14] 和 AWP[15] 方法。12 幅标准测试图像来自 USC-SIPI 数据集，如图 5 所示，从左至右、从上至下依次是 Aerial、Boat、Couple、Elaine、House、Lena、Mandrill、Peppers、Barbara、Lighthouse、Crowd、Bridge，这些测试图像被统一成 512×512 的灰度图像，涵盖人脸、动物、植物和房屋等多种类型。在 QBCS 测量端，块级观测矩阵 $\boldsymbol{\Phi}_B$ 为高斯随机矩阵，块尺寸 $B^2$ 通常为16×16，每种方法均选择采样率 $S$ 和缺省质量因子 $Q$ 的典型参数组合，RHPC 方法需要在此基础上进一步确定 3 种区域的质量因子。与 JPEG 类似，质量因子的变化范围是从 1 到 100。

图 5　实验中使用的测试图像

Fig. 5　Test images used in the experiments

基于某($S$,$Q$)组合，测量端执行测试图像的观测与编码。典型的($S$,$Q$)组合包括(0.3,36)、(0.4,48)、(0.5,60)、(0.6,72)和(0.7,84)，对应的编码失真具有递减趋势。面向 QBCS 的平滑投影 Landweber(SPL)算法是最具代表性的图像重构算法[18]，该算法采用 5 级离散二级小波变换(DDWT)作为稀疏基，小波域的收敛控制率是 25，实验使用 SPL 算法来恢复每幅图像。编码失真采用峰值信噪比(PSNR)来评价。

在求取非关键区域的质量因子时，不同的经验值 $\tau$ 会在一定程度上影响 RHPC 方法的性能，可通过实验选择较为合适的 $\tau$ 值。对于 2 幅图像，图 6 显示了在执行 RHPC 方法时 PSNR 与 $\tau$ 的取值对应关系。可以看出，$\tau$ 通常在 1.05 和 1.15之间时性能较好，典型地，$\tau=1.10$。

当($S$,$Q$)组合分别为(0.3,36)、(0.5,60)和

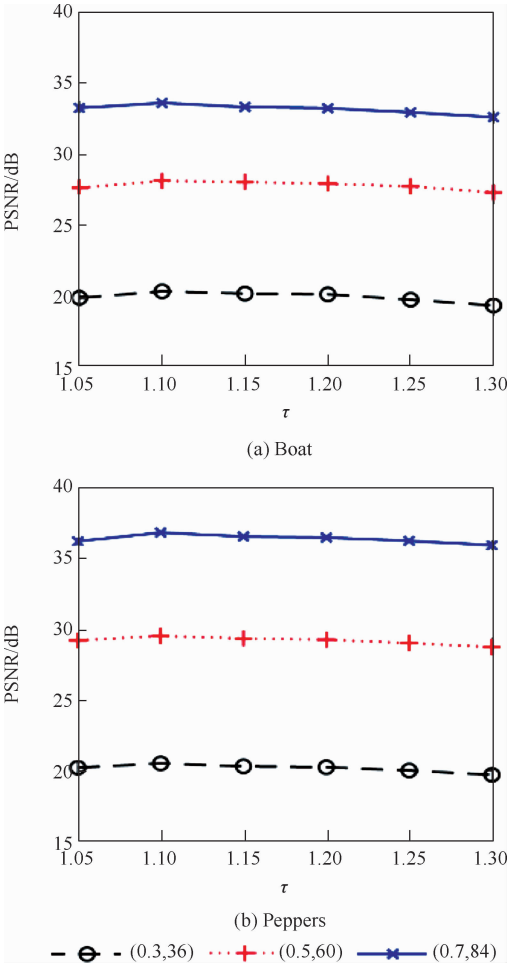



图 6 经验值 $\tau$ 对 RHPC 方法的性能影响

Fig. 6 Influence of empirical value $\tau$ on RHPC performance

(0.7,84)时,图 7 给出了使用 RHPC 方法的重构图像。这些例子表明,较大的采样率和质量因子可以产生更令人满意的重构图像。当$(S,Q)=(0.3,36)$时,重构图像中的某些区域模糊;当$(S,Q)=(0.7,84)$时,重构图像在主观质量上具有更好的层次性、对比度和细节层次。实验所选的典型$(S,Q)$组合具有良好的代表性。

比特率采用每像素比特数(bpp)进行评估。在相同的 SPL 重构算法下,图 8 给出了 5 种预测编码方法下所有测试图像的平均率失真曲线。在用每一$(S,Q)$组合测量完 12 幅测试图像之后,可求得其平均比特率和平均 PSNR 值,所得到的平均率失真曲线越高表示性能越好。在 5 种预测编码方法中,MFPQ 方法的实时性最好,但率失真性能最低;SPC 方法的复杂度最高,但率失真性能居中。相较于率失真曲线最低的 MFPQ 方法,其他的 SDPC、SPC、AWP 和 RHPC 方法分别取得了 0.57 dB、0.72 dB、0.86 dB、0.98 dB 的率失真增益。在 5 种预测编码方法中,RHPC 方法相较现有预测编码方法进一步提升了率失真性能,获得

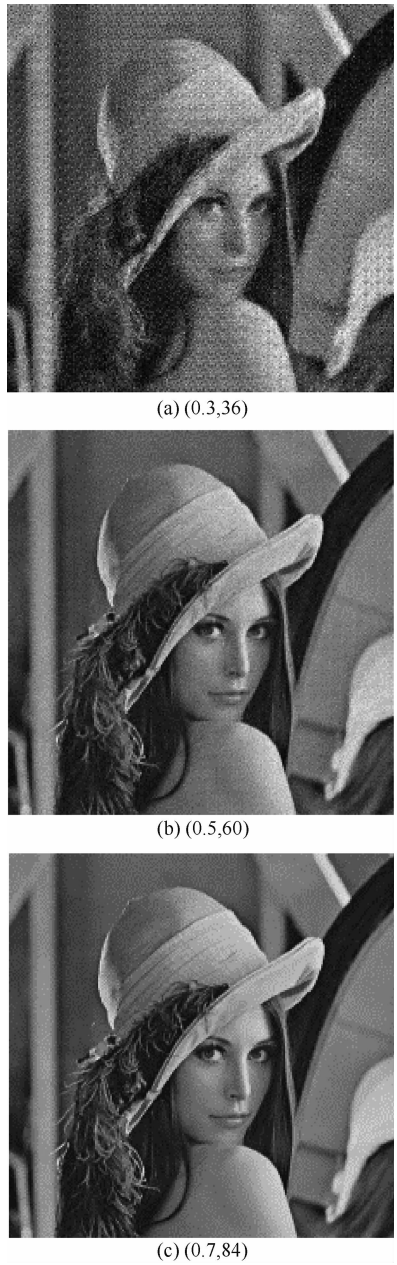



图 7 基于 RHPC 方法的 Lena 重构图像(使用典型的$(S,Q)$组合)

Fig. 7 Reconstructed images of Lena by RHPC with typical $(S,Q)$ combinations

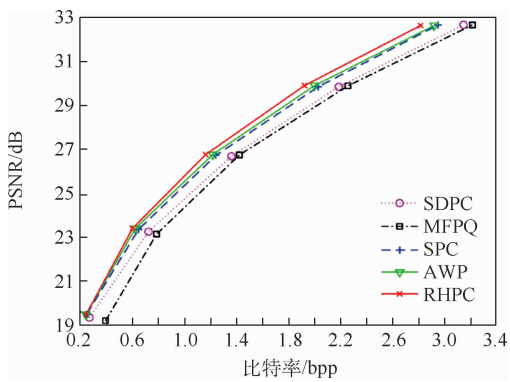



图 8 不同预测编码方法的平均率失真曲线

Fig. 8 Average rate-distortion curves of different predictive coding methods

了至少 0.12 dB 的率失真增益。在保持平均比特率的前提下,RHPC 方法可以减少测量端的编码失真,以居中的计算复杂度获得了更好的重建质量。

## 4　结　论

预测编码是量化分块压缩感知的重要机制之一,为了高效地降低编码失真,本文提出了一种基于螺旋逐块扫描的区域层次化预测编码方法。本文的主要贡献如下:

1) 所有块按由内向外的扫描顺序逐块地进行预测与量化,上下文感知候选集被用于获取相对更优的反量化矢量,提高了候选者的匹配度。

2) 通过块编码模型为不同区域自适应地分配质量因子,具有较强参考价值的区域被分配较大的质量因子,利用层次相关性进一步降低了整幅图像的编码失真。

3) 所提方法不需要将信号分布作为先验知识,能够综合利用矢量之间的空域相关性和层次相关性,在预测编码中获得更好的率失真性能。

### 参考文献 (References)


[1] CHEN Z, HOU X S, SHAO L, et al. Compressive sensing multi-layer residual coefficients for image coding[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2020, 30(4): 1109-1120.

[2] PEETAKUL J, ZHOU J J, WADA K. A measurement coding system for block-based compressive sensing images by using pixel-domain features[C]// 2019 Data Compression Conference (DCC). Piscataway: IEEE Press, 2019: 599.

[3] UNDE A S, DEEPTHI P P. Rate-distortion analysis of structured sensing matrices for block compressive sensing of images[J]. Signal Processing: Image Communication, 2018, 65: 115-127.

[4] FOWLER J E, MUN S, TRAMEL E W. Block-based compressed sensing of images and video[J]. Foundations and Trends in Signal Processing, 2012, 4(4): 297-416.

[5] WANG L J, WU X L, SHI G M. Binned progressive quantization for compressive sensing[J]. IEEE Transactions on Image Processing, 2012, 21(6): 2980-2990.

[6] PUDI V, CHATTOPADHYAY A, LAM K Y. Efficient and light-weight quantized compressive sensing using μ-law[C]// 2018 IEEE International Symposium on Circuits and Systems. Piscataway: IEEE Press, 2018: 1-5.

[7] RAPP J, DAWSON R M A, GOYAL V K. Estimation from quantized Gaussian measurements: When and how to use dither[J]. IEEE Transactions on Signal Processing, 2019, 67(13): 3424-3438.

[8] WANG X Q, LI G, QUAN C, et al. Distributed detection of sparse stochastic signals with quantized measurements: The generalized Gaussian case[J]. IEEE Transactions on Signal Processing, 2019, 67(18): 4886-4898.

[9] TRAN T T T, PEETAKUL J, PHAM C D K, et al. Bi-directional intra prediction based measurement coding for compressive sensing images[C]// 2020 IEEE 22nd International Workshop on Multimedia Signal Processing. Piscataway: IEEE Press, 2020: 1-6.

[10] SHI W, JIANG F, LIU S, et al. Image compressed sensing using convolutional neural network[J]. IEEE Transactions on Image Processing, 2019, 29: 375-388.

[11] MUN S, FOWLER J E. DPCM for quantized block-based compressed sensing of images[C]// 2012 Proceedings of the 20th European Signal Processing Conference (EUSIPCO). Piscataway: IEEE Press, 2012: 1424-1428.

[12] ZHANG J, ZHAO D B, JIANG F. Spatially directional predictive coding for block-based compressive sensing of natural images[C]// 2013 IEEE International Conference on Image Processing. Piscataway: IEEE Press, 2013: 1021-1025.

[13] LI R, LIU H B, HE W, et al. Space-time quantization and motion-aligned reconstruction for block-based compressive video sensing[J]. KSII Transactions on Internet and Information Systems, 2016, 10(1): 321-340.

[14] TIAN W, LIU H. Measurement-domain spiral predictive coding for block-based image compressive sensing[C]// Proceedings of 10th International Conference on Image and Graphics. Piscataway: IEEE Press, 2019: 3-12.

[15] CHEN Q L, CHEN D R, GONG J L. Weighted predictive coding methods for block-based compressive sensing of images[C]// 2020 3rd International Conference on Unmanned Systems (ICUS). Piscataway: IEEE Press, 2020: 587-591.

[16] YUAN X, HAIMI-COHEN R. Image compression based on compressive sensing: End-to-end comparison with JPEG[J]. IEEE Transactions on Multimedia, 2020, 22(11): 2889-2904.

[17] ZHANG Z, FANG R, LIN J, et al. A novel rate control method for still image coding[C]// Proceedings of 5th International Conference on Computer and Communications. Piscataway: IEEE Press, 2019: 1777-1781.

[18] TREVISI M, AKBARI A, TROCAN M, et al. Compressive imaging using RIP-compliant CMOS imager architecture and land-weber reconstruction[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2020, 30(2): 387-399.

# Region-hierarchical predictive coding for quantized block compressive sensing

LIU Hao*, ZHENG Haoran, HUANG Rong

(College of Information Science and Technology, Donghua University, Shanghai 201620, China)

**Abstract**: During the predictive coding of quantized block compressive sensing, a large quantity of inefficient candidates will lead to low rate-distortion performance. To efficiently reduce the encoding distortion, this paper proposes a region-hierarchical predictive coding method for quantized block compressive sensing, which is based on the block-by-block spiral scan. After all blocks are measured at a subrate, the measurement vector of each block is numbered and encoded in spiral scan order. For the current measurement vector, its prediction vector is the inverse quantization vector with maximum similarity from its context-aware candidate set. According to its hierarchical correlation, each measurement vector is classified into one of three regions. The block coding model is used to determine adaptive quality factors for different regions, where the key region is assigned a larger quality factor. As compared with the existing predictive coding methods, the proposed method jointly utilizes the local correlation and hierarchical correlation among these vectors, and the experimental results show that at least 0.12 dB rate-distortion gain is obtained.

**Keywords**: quantized block compressive sensing; predictive coding; hierarchical correlation; key region; quality factor

**Received**: 2021-09-03; **Accepted**: 2021-09-17; **Published online**: 2021-09-28 20:03

**URL**: kns.cnki.net/kcms/detail/11.2625.V.20210928.1203.002.html

**Foundation items**: National Natural Science Foundation of China (62001099); the Fundamental Research Funds for the Central Universities (2232021G-09)

* **Corresponding author**. E-mail: liuhao@dhu.edu.cn

http://bhxb.buaa.edu.cn jbuaa@buaa.edu.cn
**DOI**: 10.13700/j.bh.1001-5965.2021.0528

# HEVC 对偶编码单元划分优化算法

刘美琴[1,2], 徐晨铭[1,2], 姚超[3,*], 林春雨[1,2], 赵耀[1,2]

(1. 北京交通大学 信息科学研究所, 北京 100044; 2. 现代信息科学与网络技术北京市重点实验室, 北京 100044;
3. 北京科技大学 计算机与通信工程学院, 北京 100083)

**摘 要**: 为了解决视频数据量日益增长与用户享受高质量视频体验需求之间的矛盾,HEVC 在 H.264/AVC 标准的基础上通过引入新型的编码结构和算法进一步将编码效率提升了 50%,但是也极大地提升了编码复杂度。基于此,提出对偶编码单元(CU)划分网络 DualNet,来降低 HEVC 中帧内编码复杂度。该网络由预测网络和目标网络 2 个部分组成,其中,预测网络通过分析图像统计特征实现编码单元划分决策,从而跳过四叉树的遍历搜索,提高编码单元划分决策的时间效率;目标网络基于率失真代价评价和优化决策模型提升编码单元划分性能,实现模型互补和最优率失真估计。实验结果表明:与 HEVC 标准对比,所提算法在实现相近的压缩效果的前提下能够节省 64.06% 的编码时间。

**关 键 词**: 视频编码; H.265/HEVC; 编码单元(CU)划分; 深度学习; 对偶神经网络
**中图分类号**: TP391
**文献标志码**: A **文章编号**: 1001-5965(2022)08-1383-07

ITU-T 与 ISO/IEC 标准化组织制定了一系列视频编码标准[1],如 H.262[2]、H.264/AVC[3]、H.265/HEVC[4]、H.266/VVC[5] 等。与 H.264/AVC 标准相比,H.265/HEVC 标准提高了帧内预测、帧间预测的性能,在获得相当解码重建质量的情况下,可以提升 50% 以上的编码效率。其中,HEVC 使用编码单元(coding unit, CU)替代了 H.264/AVC 中的宏块单元(macro block, MB),CU 结构的划分占据了大量的编码时间,在提高编码效率的同时,也急剧增加了编码复杂度,如基于四叉树的 CU 递归划分方式占用了 HEVC 编码时间的 80%。因此,如何优化 CU 划分方式、降低 HEVC 的编码复杂度成为了当前的研究热点。

HEVC 将每帧图像划分为编码树单元(coding tree unit, CTU)[6],CTU 采用四叉树结构可以进一步划分为多个 CU。为寻找 CU 的最佳划分方式,HEVC 需遍历 CTU 中所有从 64×64 到 8×8 不同尺度的 CU,并分别计算率失真代价[7]。另外,在确定 CU 最佳划分方式的过程中,需通过计算编码失真和码率,拟合最小率失真代价。因此,在 HEVC 中实现 CU 最优划分所采用的递归搜索算法具有较高的时间复杂度,需要消耗大量的编码时间。因此,本文提出基于对偶网络的 CU 划分算法 DualNet。在 CU 划分过程中,构建基于卷积神经网络(convolutional neural networks, CNN)的预测网络,先提取图像统计性特征,预测当前编码帧可能的 CU 划分模式;再构建目标网络,通过基于率失真函数的强化学习方法决策最优的 CU 划分模式反馈训练预测模型,实现 CU 划分的最优率失真估计,避免遍历所有 CU 划分模式,以加

**收稿日期**: 2021-09-06; **录用日期**: 2021-09-17; **网络出版时间**: 2021-10-11 16:35
**网络出版地址**: kns.cnki.net/kcms/detail/11.2625.V.20211011.1531.001.html
**基金项目**: 国家自然科学基金(61972028,61902022,62120106009);中央高校基本科研业务费专项资金(2019JBM018,FRF-TP-19-015A1)
* **通信作者**. E-mail: yaochao@ustb.edu.cn

**引用格式**: *刘美琴, 徐晨铭, 姚超, 等. HEVC 对偶编码单元划分优化算法[J]. 北京航空航天大学学报, 2022, 48(8): 1383-1389.*
*LIU M Q, XU C M, YAO C, et al. Dual coding unit partition optimization algorithm of HEVC[J]. Journal of Beijing University of Aeronautics and Astronautics, 2022, 48(8): 1383-1389 (in Chinese).*

快 CU 划分速度,提升 HEVC 的编码性能。

综上,本文提出基于率失真约束的 DualNet 对偶优化网络,用于在降低 HEVC 帧内模式 CU 划分决策复杂度的同时提高编码质量;设计 CU 划分阈值动态调整优化策略,实现网络训练的快速优化,用于提高模型对数据及 HEVC 逆行 CU 划分决策的普适性。所提 DualNet 与 HEVC 标准编码软件(HM16.5)相比,编码时间平均降低 64.06%,而平均比特率增量 BD-BR 和平均视频客观质量提高值 BD-PSNR 仅为 2.876% 和 -0.120 dB。

# 1 相关工作

## 1.1 HEVC 编码单元划分优化算法

降低 CU 划分方式计算复杂度的方法包括启发式和基于机器学习的方式。在启发式划分算法中,文献[8-13]分别通过改进预测模式、CU 存储数据结构和图像特征优化决策模式提高编码效率。在改进预测模式方面,Zhang 等[8]采用简化预测模式改进了粗略率失真代价的搜索策略,加快了 CU 的划分速度;Lu 和 Li[9]提出了利用帧内相邻块预测当前 CU 划分的方法。在改进数据结构方面,Wang 等[10]将原有四叉树形式改进为扩展四叉树结构,以存储划分结果。在改进图像特征方面,Qing 等[11]利用图像显著性特征作为 CU 划分的判断标准;Kibeya 等[12]利用在划分过程中产生的零星图像块处理矩阵特征,以提前判断 CU 的划分性能;朱蕾琦等[13]利用小波变化后的低频子图做帧内模式选择,以减少 CU 划分过程中的率失真计算次数。在上述工作中,基于图像特征的算法效果最佳,大部分后续研究工作致力于通过分析图像特征来跳过率失真搜索过程。

基于机器学习的划分方式是利用机器学习模型,手动[14-16]或自动[17-20]提取图像特征辅助 CU 划分以代替 HEVC 中的 CU 划分方式,减少 CU 划分所需的计算时间。Zhang 等[14]利用决策树离线训练编码信息,加快编码速度;Kim 和 Park[15]提出了在线与离线相结合的训练模式,利用贝叶斯决策模型训练损失函数划分 CU;Fu 等[16]利用支持向量机提取图像特征,离线训练划分模型。上述手动提取特征法难以全面获取图像特征信息,限制了其编码准确度。自动提取图像特征法则利用 CNN,自动挖掘大规模图像数据中与 CU 划分相关的特征。Liu 等[17]首次提出了利用深度 CNN 辅助 CU 划分决策;Zhang 等[18]利用深度 CNN 学习图像纹理辅助 CU 划分;易清明等[19]提出了基于 Inception 模块的 CNN 结构对 CU 划分进行提前预测;Xu 等[20]提出了端到端的 CU 划分结构,通过三通道 CNN 综合提取图像特征。上述工作均将 CU 的划分看作二元分类问题,忽略了率失真函数编码标准,间接降低了网络的编码性能,且传统深度 CNN 对全局信息敏感度不高,也会在一定程度上限制 CU 的特征提取精度[21]。

## 1.2 对偶网络

对偶网络是一种常用于知识蒸馏和模型强化的深度网络。在知识蒸馏的研究中,文献[22-24]运用了对偶网络将复杂模型转变为相对简单的模型,实现了模型的轻量化。Bae 等[22]设计了一种多重对偶网络结构进行密集知识转移,提升了图像分类的准确度,并降低了其计算复杂度;Abbasi 等[23]提出了通用对偶师生网络模型,运用于不同场景;Xiao 等[24]利用对偶网络将知识转移关系转换为竞争关系,通过无监督方式提升了模型的适应性。基于上述工作,对偶网络也被运用于加强网络模型之间的相关性,提升单网络的训练效果,实现网络模型的互补。Lu 等[25]提出了基于双域融合的图像质量评估算法,联合频率域和空间域双网络,实现模型互补;Zhou 等[26]通过对偶网络实现了对磁共振成像的二次重建;Wang 等[27]利用对偶蒸馏网络,丰富图像重建过程中图像细节并互相监督特征提取过程;苏志雄等[28]构建了对偶模型,在网络计划模型计算中分别体现机动时间和路差,使网络计划更具针对性和有效性。上述工作均从对偶网络出发,利用并行网络模块解决多维问题,实现模型的互补和强化。

为了简化 HEVC 编码算法中 CU 划分算法并融合率失真函数作为 CU 划分决策的优化函数,本文利用对偶网络结构,构建 CU 划分决策模型,在保证编码性能的前提下提高 HEVC 帧内模式中 CU 的划分效率。其中,对偶网络融合了用于预测 CU 划分的预测网络和用于评价划分结果的目标网络。预测网络使用 CNN 自动提取图像的统计性像素特征,实现像素级 CU 的划分预测,端到端地输出划分的预测结果。目标网络以率失真函数约束损失函数,采用动态规划方式获得最佳划分阈值,优化网络模型参数,提高 CU 划分的精度。

# 2 编码单元划分对偶网络

本文构建的 CU 划分对偶网络 DualNet 结构如图 1 所示。其中,预测网络用于预测 CU 的划分,目标网络结构与预测网络相似,仅参数更新速度慢于预测网络,采用率失真代价评估预测网络

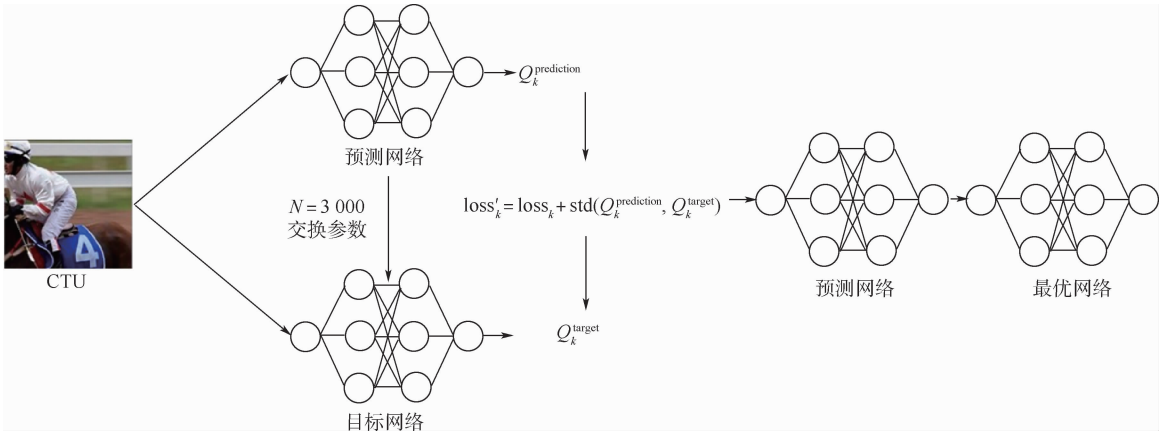



图 1　DualNet 结构示意图

Fig. 1　Structure of DualNet

的参数更新趋势。对偶网络预测模型评价值的标准差将作为损失函数的一部分返回预测网络进行拟合,最终得到具有最优参数的网络模型。

## 2.1　预测网络结构

预测网络用于替代 HEVC 标准编码软件(HEVC test model, HM)[29]中 CU 递归划分过程,在编码过程中读取由 CNN 训练得到的划分预测模型。具体地,采用分层 CU 划分图(hierarchical CU partition map, HCPM)[20]存储划分真值和预测值。与 HEVC 编码单元划分中自顶向下逐层计算率失真代价不同,该结构可将 CTU 作为输入,直接得到 1 + 4 + 16 = 21 个划分单元的整体输出。

CU 的划分方式与其图像特征的复杂程度密切相关。因此,预测网络借助 CNN 分析图像特征复杂度,实现 CU 的划分,如图 2 所示。具体地,预测网络首先对输入的 CTU 亮度矩阵进行去均值和降采样操作,加快网络的训练速度;其次,利用由 3 层 CNN 构成的 3 条分支 $B_1$、$B_2$、$B_3$ 提取图像特征,以对应 HCPM 中的 CU 划分层级;然后,不同层级的特征图归并为一个特征向量,在全连接层分别由 3 条支路处理,以符合 HCPM 的 3 层输出模式和量化参数(quantization parameters, QP)对 CU 划分的影响;最后,采用图像真值与对应特征向量间的交叉熵损失函数评价特征提取准确度。

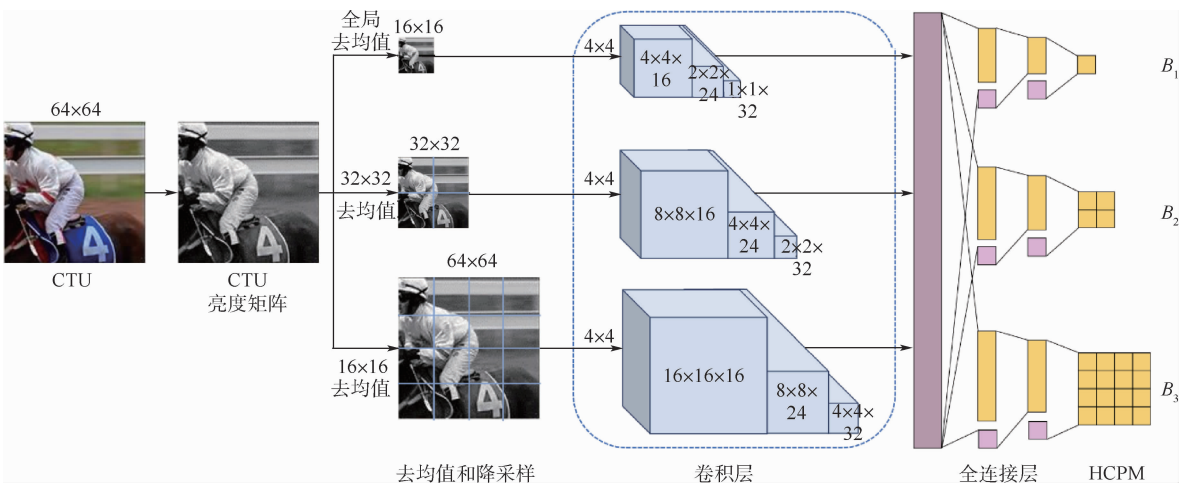



图 2　预测网络结构示意图

Fig. 2　Structure of prediction network

## 2.2　对偶编码单元划分网络

总体模型采用对偶网络优化结构,参考率失真代价建立评价体系,对预测网络进行评价和监督,不断优化预测模型参数,其结构如图 3 所示。

本文针对单一 CNN 模型无法全面考虑 CU 划分影响因素的问题,参考 HEVC 中率失真代价原理,设置包括失真代价 $Q_{kd}$ 和时间代价 $Q_{kt}$ 的目标评价体系 $Q_k$($k$ = 64,32,16)。$Q_{kd}$ 由真值和模型预测结果不相同的情况组成,用 $matrix_k[1][0]$ 与 $matrix_k[0][1]$ 之和表示。具体地,本文采用二

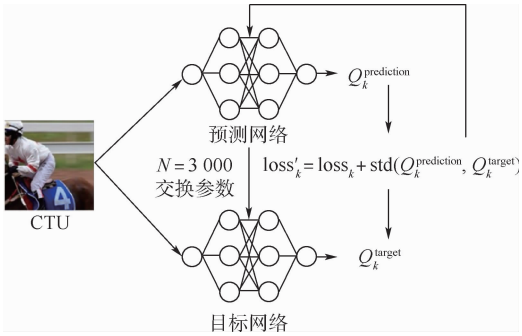



图 3 对偶优化网络结构

Fig. 3 Structure of dual optimal networks

维数组 $\text{matrix}_k[m][n]$($k=64,32,16$; $m=0,1$; $n=0,1$)记录划分结果。其中,$k$ 代表当前 CU 的大小,$m$ 代表划分真值(值为 0 表示不划分,1 表示划分),$n$ 代表由预测模型得到的划分结果,即 $\text{matrix}_k[m][n]$ 记录了 CU 大小为 $k$ 时划分真值为 $m$ 且预测为 $n$ 的结果数量。$Q_{kt}$ 由预测模型决定划分的情况数之和表示,即 $\text{matrix}_k[0][1]$ 与 $\text{matrix}_k[1][1]$ 之和。总代价 $Q_k$ 由失真代价和时间代价和表示:

$$Q_k = \alpha Q_{kt} + \beta Q_{kd} \tag{1}$$

式中:$\alpha$ 和 $\beta$ 分别为 $Q_{kt}$ 和 $Q_{kd}$ 的比例系数。一般而言,失真代价对编码效果影响大于时间代价。

此外,由于单一监督信息会导致模型决策朝某一固定的方向发展,导致最终模型出现负优化或无法适应其他可能出现的情况。因此,本文建立对偶网络结构平衡图像特征和率失真代价之间的关系,将原有用于训练预测划分模型的网络视为预测网络,并采用目标网络监督其训练过程。目标网络的结构与预测网络相同,在训练迭代的过程中不实时更新参数。对于一组数据,2 个网络同时对其划分方式输出预测模型,当预测网络优于目标网络的评价值累计 $N$ 次时,将预测网络的参数赋给目标网络。该结构有效地对预测网络的训练实现了强监督操作,以免因损失函数不合理而导致模型的负优化。

为了控制预测网络参数向对偶网络中具有较好效果的网络方向发展,预测网络只在其评价结果 $Q_k^{\text{prediction}}$ 亚于目标网络评价结果 $Q_k^{\text{target}}$ 时拟合二者之间的损失。因此,为了平衡率失真与图像失真,本文设计的损失函数 $\text{loss}'_k$ 由原有卷积损失 $\text{loss}_k$ 与 2 个网络评价值的标准差 loss′组成,如下:

$$\text{loss}'_k = \text{loss}_k + \text{std}(Q_k^{\text{prediction}}, Q_k^{\text{target}}) \tag{2}$$

式中:std 表示求解 $Q_k^{\text{prediction}}$ 和 $Q_k^{\text{target}}$ 标准差,$k=64,32,16$。

通常,二分决策方法取 0.5 作为划分阈值。DualNet 网络使用率失真影响了决策方向,因此,DualNet 的划分阈值不能采用 0.5,而是采用基于率失真的动态阈值策略来计算划分阈值。阈值调整公式如下:

$$Q = \text{matrix}_k[0][1]\cdot\gamma - \text{matrix}_k[1][0]\cdot\delta \tag{3}$$

式中:$Q$ 为最终的划分阈值;$\gamma$ 和 $\delta$ 为比例系数。$\text{matrix}_k[0][1]$ 对模型预测结果影响更大。

## 3 实 验

### 3.1 实验数据及设置

1)实验数据。本文的训练数据采用文献[20]中的 CPH-Intra 数据集[30],包括 2 000 张无损图像。每个样本由 CU 的亮度矩阵和一个代表是否划分的二分类标签组成。测试数据使用 5 种不同类别的 JCT-VC[26] 序列,序列参数如表 1 所示。

**表 1 JCT-VC 测试序列参数**

**Table 1 JCT-VC test sequence parameters**

| 类别 | 序列名称 | 分辨率 | 帧数 | 帧率/fps |
|---|---|---|---|---|
| A | Traffic | 2 560×1 600 | 150 | 30 |
| B | BasketballDrive | 1 920×1 080 | 500 | 50 |
| C | BasketballDrill | 832×480 | 500 | 50 |
| D | BasketballPass | 416×240 | 500 | 50 |
| E | Johnny | 1 280×720 | 600 | 60 |

注:fps 为帧/s。

2)评价指标。本文使用提案 VCEG-M33 中的 BD-BR 方法[31],包含 BD-BR、BD-PSNR 和 $\Delta T$ 三个评价指标。其中,BD-BR 表示在视频客观质量相同的情况下,优化算法相比原始算法的比特率增量;BD-PSNR 表示在相同码率的情况下,优化算法相比原始算法的视频客观质量提高值;$\Delta T$ 表示在编码相同帧数的情况下,优化算法节省的时间,计算如下:

$$\Delta T = \frac{T' - T}{T} \times 100\% \tag{4}$$

式中:$T$ 和 $T'$ 分别为 HEVC 标准和优化算法消耗的编码时间。

3)参数设置。本文采用 Momentum 优化器,学习率设置为 0.01。式(1)中的关键参数 $\beta$ 设置为 0.005,$\alpha$ 设置为 0.01;式(3)中的关键参数 $\gamma$ 和 $\delta$ 分别设置为 0.001 和 0.0005。DualNet 网络训练使用一张 NVIDIA GeForce GTX 1080Ti 显卡,Tensorflow 版本为 1.14,采用 HM16.5 帧内模式默认参数配置文件 encoder_intra_main.cfg 实现编码过程。

### 3.2　消融实验

为了评估对偶网络结构和动态阈值的性能,本文完成了 4 组实验:第 1 组实验(记为 CNN)采用 2.1 节 CNN 训练模型,划分阈值设置为[0.5,0.5,0.5];第 2 组实验(记为 Thr-CNN)采用带有动态阈值的 CNN 训练模型,训练后动态阈值结果为[0.49,0.55,0.63];第 3 组实验(记为 DualNet-$E_1$)采用 2.2 节的对偶 CNN 训练模型,划分阈值设置为[0.5,0.5,0.5];第 4 组实验(记为 DualNet-$E_2$)采用具有动态阈值的对偶 CNN 进行训练,训练后动态阈值结果为[0.48,0.55,0.63]。训练所得模型经过 4 种 QP(QP = 22,27,32,37)测试,计算每个序列的前 20 帧。测试结果使用 JCT-VC 提供的 VCEG-AE07.xls[29]计算 BD-BR 和 BD-PSNR,消融实验结果如表 2 所示。

由于本文算法跳过了递归计算率失真损失的过程,以上算法相比 HM16.5 在大规模降低计算复杂度的情况下,均对视频解码重建质量产生了影响。表 2 中:BD-BR 值均为正,代表 HM16.5 的编码效率优于上述算法,但是其值越低表示码率越小,编码压缩效果越接近 HM16.5;BD-PSNR 值均为负,代表压缩后视频质量相比 HM16.5 均有损失,其值越高代表损失越小,视频质量越接近 HM16.5。由表 2 可知,Thr-CNN 仅使用动态划分阈值,没有目标网络监督训练,无法确定阈值变化方向,导致在部分数据集下测试效果不佳。DualNet-$E_1$使用对偶网络结构,预测网络参数模型受目标网络影响,仍然使用 0.5 作为划分阈值,限制了模型预测效果。采用动态阈值的对偶网络 DualNet-$E_2$的平均 BD-PSNR 为 −0.120 dB、平均 BD-BR 为 2.876%、平均 $\Delta T$ 为 −64.06%,性能优于 CNN、Thr-CNN、DualNet-$E_1$ 三个网络,表明该网络不仅适用于不同分辨率的测试序列的 CU 划分过程,而且获得最佳编码质量的同时,大大节约了 HM16.5 帧内编码时间。

### 3.3　实验结果及分析

为了评估本文所提对偶网络对 CNN 处理 CU 划分的优化效果,与单神经网络结构相似但无率失真监督学习的 PPMAC[17]和 ETH-CNN[20]算法进行对比,实验结果如表 3 所示。

从表 3 可知,本文算法训练得到的模型对大部分数据集的压缩效果和效率优于对比算法,且具有较小的离散程度。由于本文算法将率失真代价作为评价方法,与对比算法中编码效果较优的 ETH-CNN 相比,BD-PSNR 平均提升了0.008 dB,BD-BR 降低了 0.192%,提高了重建质量。动态

**表 2　消融实验结果(JCT-VC)**

**Table 2　Results of ablation study (JCT-VC)**

| 类别 | 算法 | BD-PSNR/dB | BD-BR/% | $\Delta T$/% |
|---|---|---|---|---|
| A | CNN | −0.149 | 2.771 | −63.19 |
| | Thr-CNN | −0.133 | 2.480 | −66.79 |
| | DualNet-$E_1$ | −0.148 | 2.757 | −66.01 |
| | DualNet-$E_2$ | −0.131 | 2.429 | −63.55 |
| B | CNN | −0.119 | 4.981 | −72.29 |
| | Thr-CNN | −0.094 | 3.904 | −77.10 |
| | DualNet-$E_1$ | −0.120 | 4.967 | −75.31 |
| | DualNet-$E_2$ | −0.094 | 3.941 | −74.29 |
| C | CNN | −0.141 | 2.934 | −43.77 |
| | Thr-CNN | −0.134 | 2.796 | −51.98 |
| | DualNet-$E_1$ | −0.142 | 2.969 | −50.03 |
| | DualNet-$E_2$ | −0.130 | 2.738 | −47.87 |
| D | CNN | −0.138 | 2.412 | −48.32 |
| | Thr-CNN | −0.116 | 2.029 | −50.53 |
| | DualNet-$E_1$ | −0.135 | 2.359 | −44.75 |
| | DualNet-$E_2$ | −0.107 | 1.853 | −57.06 |
| E | CNN | −0.146 | 3.636 | −75.04 |
| | Thr-CNN | −0.136 | 3.355 | −77.51 |
| | DualNet-$E_1$ | −0.141 | 3.501 | −78.48 |
| | DualNet-$E_2$ | −0.138 | 3.421 | −77.55 |
| 标准差 | CNN | 0.011 | 1.016 | 14.02 |
| | Thr-CNN | 0.018 | 0.735 | 13.08 |
| | DualNet-$E_1$ | 0.011 | 1.013 | 15.01 |
| | DualNet-$E_2$ | 0.019 | 0.821 | 12.23 |
| 平均值 | CNN | −0.139 | 3.347 | −60.52 |
| | Thr-CNN | −0.123 | 2.913 | −64.78 |
| | DualNet-$E_1$ | −0.137 | 3.311 | −62.92 |
| | DualNet-$E_2$ | −0.120 | 2.876 | −64.06 |

**表 3　编码单元划分对比实验结果**

**Table 3　Experimental results of CU partition**

| 类别 | 算法 | BD-PSNR/dB | BD-BR/% | $\Delta T$/% |
|---|---|---|---|---|
| A | PPMAC | −0.240 | 4.945 | −60.84 |
| | ETH-CNN | −0.125 | 2.550 | −61.01 |
| | DualNet-$E_2$ | −0.131 | 2.429 | −63.55 |
| B | PPMAC | −0.141 | 6.018 | −69.51 |
| | ETH-CNN | −0.121 | 4.265 | −76.32 |
| | DualNet-$E_2$ | −0.094 | 3.941 | −74.29 |
| C | PPMAC | −0.538 | 12.205 | −63.58 |
| | ETH-CNN | −0.133 | 2.863 | −52.98 |
| | DualNet-$E_2$ | −0.130 | 2.738 | −47.87 |
| D | PPMAC | −0.457 | 8.401 | −63.53 |
| | ETH-CNN | −0.106 | 1.842 | −56.42 |
| | DualNet-$E_2$ | −0.107 | 1.853 | −57.06 |
| E | PPMAC | −0.307 | 7.956 | −66.55 |
| | ETH-CNN | −0.153 | 3.822 | −70.68 |
| | DualNet-$E_2$ | −0.138 | 3.421 | −77.55 |
| 标准差 | PPMAC | 0.160 | 2.787 | 3.32 |
| | ETH-CNN | 0.017 | 0.977 | 9.78 |
| | DualNet-$E_2$ | 0.019 | 0.821 | 12.23 |
| 平均值 | PPMAC | −0.337 | 7.905 | −64.80 |
| | ETH-CNN | −0.128 | 3.068 | −63.48 |
| | DualNet-$E_2$ | −0.120 | 2.876 | −64.06 |

阈值使得 CU 划分模型能在各种情况下做出合理的判断。同时,本文算法比 ETH-CNN 算法平均节省了 0.58% 的编码时间,提升了 HEVC 编码效率。本文算法参考了 HEVC 标准中 CU 划分原则,并利用对偶网络实现监督学习,相比于不使用率失真监督信息和对偶网络的深度模型,算法模型对压缩效果和时间均有优化,验证了对偶 CU 划分网络 DualNet 的有效性。

## 4 结 论

本文提出了基于对偶网络结构的 CU 划分决策网络模型,分别构建用于决策的预测网络和用于率失真优化的目标网络,并且反馈优化预测网络对 CU 划分的阈值,提高模型的预测准确度。实验结果表明,本文算法可以提高视频编码的质量,降低了编码复杂度。

在后续的研究中,将在此基础上,进一步探索帧间预测模式对 CU 划分的影响,实现帧内模式和帧间模式 CU 划分的优化,以进一步提升 HEVC 的性能。

### 参考文献(References)


[1] LIU D, LI Y, LIN J, et al. Deep learning-based video coding: A review and a case study[J]. ACM Computing Surveys, 2020, 53(1): 1-35.

[2] TUDOR P. MPEG-2 video compression[J]. Electronics & Communication Engineering Journal, 1995, 7(6): 257-264.

[3] WIEGAND T, SULLIVAN G J, BJONTEGAARD G, et al. Overview of the H.264/AVC video coding standard[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2003, 13(7): 560-576.

[4] SULLIVAN G J, OHM J R, HAN W J, et al. Overview of the high efficiency video coding (HEVC) standard[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2012, 22(12): 1649-1668.

[5] ZHANG Y, ZHAO Y, LIN C, et al. Block partitioning decision based on content complexity for future video coding[C]// International Conference on Image and Graphics. Berlin: Springer, 2019: 70-80.

[6] GUO H, ZHU C, XU M, et al. Inter-block dependency-based CTU level rate control for HEVC[J]. IEEE Transactions on Broadcasting, 2019, 66(1): 113-126.

[7] JAMALI M, COULOMBE S. Fast HEVC intra mode decision based on RDO cost prediction[J]. IEEE Transactions on Broadcasting, 2018, 65(1): 109-122.

[8] ZHANG M, ZHAI X, LIU Z, et al. Fast algorithm for HEVC intra prediction based on adaptive mode decision and early termination of CU partition[C]// 2018 Data Compression Conference. Piscataway: IEEE Press, 2018: 434-434.

[9] LU J, LI Y. Fast algorithm for CU partitioning and mode selection in HEVC intra prediction[C]// 2019 12th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics. Piscataway: IEEE Press, 2019: 1-5.

[10] WANG M, LI J, ZHANG L, et al. Extended quad-tree partitioning for future video coding[C]// 2019 Data Compression Conference. Piscataway: IEEE Press, 2019: 300-309.

[11] QING A, ZHOU W, WEI H, et al. A fast CU partitioning algorithm in HEVC inter prediction for HD/UHD video[C]// 2016 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference. Piscataway: IEEE Press, 2016: 1-5.

[12] KIBEYA H, BELGHITH F, AYED M A B, et al. A fast CU partitionning algorithm based on early detection of zero block quantified transform coefficients for HEVC standard[C]// International Image Processing, Applications and Systems Conference. Piscataway: IEEE Press, 2014: 1-5.

[13] 朱蕾琦,张其善,杨东凯,等. 改进的帧内帧间模式选择快速算法[J]. 北京航空航天大学学报, 2008, 34(12): 1411-1414.
ZHU L Q, ZHANG Q S, YANG D K, et al. Fast mode selection for intra and inter prediction[J]. Journal of Beijing University of Aeronautics and Astronautics, 2008, 34(12): 1411-1414(in Chinese).

[14] ZHANG D, DUAN X, ZANG D. Decision tree based fast CU partition for HEVC lossless compression of medical image sequences[C]// 2017 9th International Conference on Wireless Communications and Signal Processing. Piscataway: IEEE Press, 2017: 1-6.

[15] KIM H S, PARK R H. Fast CU partitioning algorithm for HEVC using an online-learning-based bayesian decision rule[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2016, 26(1): 130-138.

[16] FU B, ZHANG Q, HU J. Fast prediction mode selection and CU partition for HEVC intra coding[J]. IET Image Processing, 2020, 14(9): 1892-1900.

[17] LIU Z, YU X, GAO Y, et al. CU partition mode decision for HEVC hardwired intra encoder using convolution neural network[J]. IEEE Transactions on Image Processing, 2016, 25(11): 5088-5103.

[18] ZHANG Y, WANG G, TIAN R, et al. Texture-classification accelerated CNN scheme for fast intra CU partition in HEVC[C]// 2019 Data Compression Conference. Piscataway: IEEE Press, 2019: 241-249.

[19] 易清明,林成思,石敏. 利用深度学习的 HEVC 帧内编码单元快速划分算法[J]. 小型微型计算机系统, 2021, 42(2): 368-373.
YI Q M, LIN C S, SHI M. Fast HEVC coding units partitioning algorithm based on deep learning[J]. Journal of Computer Systems, 2021, 42(2): 368-373(in Chinese).

[20] XU M, LI T, WANG Z, et al. Reducing complexity of HEVC: A deep learning approach[J]. IEEE Transactions on Image Processing, 2018, 27(10): 5044-5059.

[21] CHUNG C H, PENG W H, HU J H. HEVC/H.265 coding unit split decision using deep reinforcement learning[C]// 2017 International Symposium on Intelligent Signal Processing and Communication Systems. Piscataway: IEEE Press, 2017: 570-

575.

[22] BAE J H, YEO D, YIM J, et al. Densely distilled flow-based knowledge transfer in teacher-student framework for image classification[J]. IEEE Transactions on Image Processing, 2020, 29:5698-5710.

[23] ABBASI S, HAJABDOLLAHI M, KARIMI N, et al. Modeling teacher-student techniques in deep neural networks for knowledge distillation[C]// 2020 International Conference on Machine Vision and Image Processing. Piscataway: IEEE Press, 2020:1-6.

[24] XIAO R, LIU Z, WU B. Teacher-student competition for unsupervised domain adaptation[C]//2020 25th International Conference on Pattern Recognition. Piscataway: IEEE Press, 2021: 8291-8298.

[25] LU Y, LI W, NING X, et al. Image quality assessment based on dual domains fusion[C]// 2020 International Conference on High Performance Big Data and Intelligent Systems. Piscataway: IEEE Press, 2020:1-6.

[26] ZHOU B, ZHOU S K. DuDoRNet: Learning a dual-domain recurrent network for fast MRI reconstruction with deep T1 prior[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2020: 4273-4282.

[27] WANG H, TIAN Q, LI L, et al. Image demoiréing with a dual-domain distilling network[C]// 2021 IEEE International Conference on Multimedia and Exposition. Piscataway: IEEE Press, 2021:1-6.

[28] 苏志雄,李星梅,乞建勋. 网络计划中构建对偶网络模型的理论和方法[J]. 北京航空航天大学学报, 2012, 38(2): 257-262.
SU Z X, LI X M, QI J X. Theory and method of creating dual network model in network planning[J]. Journal of Beijing University of Aeronautics and Astronautics, 2012, 38(2): 257-262 (in Chinese).

[29] HM software [CP/OL]. [2021-08-28]. https://hevc.hhi.fraunhofer.de/svn/svnHEVCSoftware/tags/HM-16.5/.

[30] CPH-Intra[DS/OL]. [2021-08-28]. https://github.com/Projects/CPH.

[31] GRELLERT M, BAMPI S, CORREA G, et al. Learning-based complexity reduction and scaling for HEVC encoders[C]// 2018 IEEE International Conference on Acoustics, Speech and Signal Processing. Piscataway: IEEE Press, 2018:1208-1212.

# Dual coding unit partition optimization algorithm of HEVC

LIU Meiqin[1,2], XU Chenming[1,2], YAO Chao[3,*], LIN Chunyu[1,2], ZHAO Yao[1,2]

(1. Institute of Information Science, Beijing Jiaotong University, Beijing 100044, China;
2. Beijing Key Laboratory of Modern Information Science and Network Technology, Beijing 100044, China;
3. School of Computer and Communication Engineering, University of Science and Technology Beijing, Beijing 100083, China)

**Abstract**: To resolve the conflict between the increasing amount of video data and the demand for high-quality video experience, HEVC has boosted the compression performance by 50% based on dramatically increased complexity of H.264/AVC. In this paper, a fast coding unit (CU) partition algorithm is proposed to reduce the computational complexity of HEVC intra coding. To define the partition criteria, we design a convolutional neural network, named dual neural networks (DualNet). DualNet consists of two subnetworks, a prediction network and a target network. The prediction network is used to determine the partition actions by extracting images statistical features for skipping the traversal search of quadtree and improving the time efficiency of the CU partition. And the target network is to optimize the performance of the CU partition based on rate-distortion for achieving model complementarity. Experimental results show that the proposed algorithms can save 64.06% of the compression time with similar compression performance to HEVC.

**Keywords**: video coding; H.265/HEVC; coding unit (CU) partition; deep learning; dual neural networks

**Received**: 2021-09-06; **Accepted**: 2021-09-17; **Published online**: 2021-10-11 16:35

**URL**: kns.cnki.net/kcms/detail/11.2625.V.20211011.1531.001.html

**Foundation items**: National Natural Science Foundation of China (61972028, 61902022, 62120106009); the Fundamental Research Funds for the Central Universities (2019JBM018, FRF-TP-19-015A1)

* **Corresponding author**. E-mail: yaochao@ustb.edu.cn

http://bhxb.buaa.edu.cn jbuaa@buaa.edu.cn
**DOI**: 10.13700/j.bh.1001-5965.2021.0533

# 基于 IoU 约束的孪生网络目标跟踪方法

周丽芳[1,2,3,*], 刘金兰[1,3], 李伟生[2,3], 雷帮军[4], 何宇[1,3], 王一涵[1]

(1. 重庆邮电大学 软件工程学院, 重庆 400065; 2. 重庆邮电大学 计算机科学与技术学院, 重庆 400065;
3. 重庆邮电大学 图像认知重庆市重点实验室, 重庆 400065;
4. 三峡大学 水电工程智能视觉监测湖北省重点实验室, 宜昌 443002)

**摘 要**: 基于孪生网络的跟踪方法通过离线训练跟踪模型,不需要对跟踪模型进行在线更新,兼顾了跟踪精度和速度。现有孪生网络目标跟踪方法使用固定阈值选择正负训练样本易造成训练样本漏选问题,且训练时分类分支和回归分支之间存在低相关性问题,不利于训练出高精度的跟踪模型。为此,提出了一种基于交并比(IoU)约束的孪生网络目标跟踪方法。通过使用动态阈值策略根据预定义锚框与目标真实框的相关统计特征,动态调整正负训练样本的界定阈值,提升跟踪精度。所提方法使用 IoU 质量评估分支代替分类分支,通过锚框与目标真实框之间的 IoU 反映目标位置,提升跟踪精度,降低模型的参数量。在数据集 VOT2016、OTB-100、VOT2019、UAV123 上进行了对比实验,所提方法均有较好的表现。在 VOT2016 数据集上,所提方法的跟踪精度比 SiamRPN 方法高 0.017,期望平均重叠率为0.463,与 SiamRPN + + 相比仅差 0.001,实时运行速度可达 220 帧/s。

**关 键 词**: 目标跟踪; 深度学习; 孪生网络; 交并比(IoU)约束; 动态阈值
**中图分类号**: TP183; TP391
**文献标志码**: A **文章编号**: 1001-5965(2022)08-1390-09

目标跟踪(object tracking)作为计算机视觉领域的基本问题之一,无论是在军事国防还是民用安全方面都取得了成功应用,主要包括智慧医疗、模式识别和人机交互等方面[1]。

目前,目标跟踪方法可以分为传统方法、基于相关滤波的方法和基于深度网络的方法[2]。在基于深度网络的方法中,基于孪生网络的目标跟踪方法受到广泛的关注。该方法使用大量训练数据集离线训练全卷积的孪生网络,得到相似性度量函数,在线跟踪阶段利用训练得到的函数来解决一般的通用目标跟踪问题,离线训练跟踪模型后,该模型在执行跟踪任务时将不更新参数。因此,基于孪生网络的目标跟踪方法具有较高的跟踪精度,同时能以超实时的速度运行。

但是,目前基于区域建议的孪生网络目标跟踪方法都使用固定阈值来匹配正负训练样本。虽然这种方式能够较好地匹配大部分训练样本,但对一些不规则样本的匹配程度较低,将会损害模型的跟踪精度。在目标检测领域中也存在此类问题。Zhang 等[3]提出了一种自动根据目标统计特征选择正负训练样本的方法,提高了小目标的检测率,从而有效提升了检测精度。

此外,孪生网络在训练时分类分支与回归分支相对独立,这种低相关性也会损害跟踪器的精

**收稿日期**: 2021-09-06; **录用日期**: 2021-09-17; **网络出版时间**: 2021-11-01 14:56
**网络出版地址**: kns.cnki.net/kcms/detail/11.2625.V.20211029.1727.003.html
**基金项目**: 重庆市教育委员会科学技术研究计划(KJZD-K201900601); 重庆市自然科学基金(cstc2019jcyj-msxmX0461); 水电工程智能视觉监测湖北省重点实验室(三峡大学)开放基金(2020SDSJ01); 国家级大学生创新创业训练计划(202110617009)
* **通信作者**. E-mail: zhoulf@cqupt.edu.cn

**引用格式**: *周丽芳, 刘金兰, 李伟生, 等. 基于IoU约束的孪生网络目标跟踪方法[J]. 北京航空航天大学学报, 2022, 48(8): 1390-1398. ZHOU L F, LIU J L, LI W S, et al. Object tracking method based on IoU-constrained Siamese network[J]. Journal of Beijing University of Aeronautics and Astronautics, 2022, 48(8): 1390-1398 (in Chinese).*

度。SiamFC + +[4]通过引入质量评估分支适当缓解了该问题,但质量评估分支和分类分支在整个训练过程中相互独立,这种低相关性同样会损害模型的跟踪性能。因此,模型的跟踪精度和速度依然存在提升的空间。

受以上工作启发,本文提出了一种基于交并比(intersection over union,IoU)约束的孪生网络目标跟踪方法。设计了一种通过动态阈值决定正负训练样本的方法,根据预定义锚框与目标真实框的相关统计特征,动态调整正负训练样本的界定阈值,从而更适合于训练样本的选择,可以在几乎不带来额外计算量和参数的情况下提升模型的跟踪精度。在 SiamRPN[5]的基础上提出使用 IoU 质量评估分支代替传统的分类分支,在保证跟踪器精度的同时,降低了模型参数,提升了跟踪速度,实时运行速度可达 220 帧/s,相比 SiamRPN 提升了 22%。

## 1　相关工作

近年来,基于孪生网络的目标跟踪方法,因其在跟踪精度与实时性方面保持了较好的平衡,迅速吸引国内外研究者不断研究与探索。SINT[6]首次使用孪生网络进行目标跟踪,其结构相对简单,跟踪过程中不对模板更新。SiamFC[7]使用一种全卷积孪生网络,可以更加有效地利用数据,其在 ILSVRC15[8]数据集上端到端训练,取得了较高的跟踪精度。

由于 SiamFC 的结构简单且跟踪性能较好,基于该模型的改进方法也层出不穷。CFNet[9]在 SiamFC 结构中引入相关滤波层,使得网络可进行端到端的训练;RASNet[10]通过引入残差注意机制和通道注意机制来缓解孪生网络中的过拟合问题;SA-Siam[11]设计了一种双重孪生网络结构,主要由语义分支和外观分支组成,利用语义特征与外观特征互补进行目标跟踪,有效提高了跟踪性能。

以上对于 SiamFC 的改进均取得了不错的成绩,但在对目标尺度估计时大多通过多尺度测试的方式,不利于自适应跟踪过程中目标的复杂外观变化,大大降低了跟踪速度。为解决该问题,SiamRPN[5]借鉴 Faster R-CNN[12]中的区域建议网络,通过区域建议网络中的回归分支直接给出目标大小,避免了多尺度测试,在提高跟踪精度的同时也提升了跟踪速度。C-RPN[13]通过使用级联区域建议网络框架来缓解跟踪器遇到与目标相似的干扰物和目标大尺度变化时性能退化的问题。

上述方法在不同程度上对采用孪生网络的跟踪方法进行了改进,并取得了较好的效果,但使用仅有 5 层深度的 AlexNet[14]作为特征提取网络,没有充分利用现代深度神经网络的优势。SiamDW[15]通过设计裁剪残差单元,利用更深的网络结构进行目标跟踪。SiamRPN + +[16]使用均匀分布的采样方式使目标在中心点附近进行偏移,缓解网络因为破坏了严格平移不变性带来的影响。得益于 ResNet50[17]作为特征提取网络,SiamRPN + + 在跟踪性能上取得了很大的提升。

为了更好地利用上下文信息或背景信息进行在线跟踪,研究人员对现有采用孪生网络的跟踪方法进行了改进,增强了跟踪模型的鲁棒性。SiamCAR[18]将目标跟踪任务分解为像素分类和该像素处对目标边界框的回归,提出了一种新颖的全卷积孪生网络,以逐像素的方式解决端到端的视觉跟踪问题,提升了跟踪性能,但在区域建议网络部分,与之前的目标跟踪方法一样,SiamCAR 仍采用固定阈值来选择正负训练样本,容易导致训练样本漏选,在一定程度上影响了跟踪性能的进一步提升。

## 2　基于 IoU 约束的孪生网络框架

### 2.1　IoU 约束的动态阈值

一些基于孪生网络的目标跟踪方法通过训练分类分支来对正负训练样本进行分类,训练回归分支对目标状态进行回归。SiamRPN[5]在确定正负训练样本时,通常采用固定阈值对预定义锚框进行正负训练样本划分。当预定义锚框与目标真实框之间的 IoU 值高于 0.6 时,该锚框被视作正训练样本,其对应分类标签为 1;当 IoU 值低于 0.3时,该锚框被视作负训练样本,其对应分类标签为 0;当 IoU 值处于 0.3 ~ 0.6 之间时,该锚框在训练过程中会被忽略,其对应的分类标签为 - 1。尽管 SiamRPN 使用固定阈值作为筛选正负训练样本的条件提升了跟踪精度,但是对于众多形状大小不一的训练目标,固定阈值始终难以适合每个目标。对于不规则的样本,当使用固定阈值策略时,可能不存在与其真实框的 IoU 值大于 0.6 的预定义锚框,这些目标在训练时将会被忽略。这种情况下,即便有大量的训练数据来训练跟踪模型,很多训练样本也并没有被有效利用,损害了跟踪模型的精度。

为使正负训练样本的选取更加合理，受Zhang等[3]的启发，本文提出了一种通过动态阈值决定正负训练样本的方法。该方法根据预定义锚框与目标真实框的相关统计特征，动态调整正负训练样本的界定阈值，更适合于训练样本的选择。动态阈值选择方法的具体计算步骤如算法1所示。

**算法1** 动态阈值选择方法。

**输入**：真值框(ground-truth box) $g$、锚框集合 $A$、超参数 $n$。

**输出**：正训练样本 $P$、负训练样本 $N$。

1：为 $g$ 的候选正样本构建一个空集 $C_g \leftarrow \varnothing$

2：计算每个预定义锚框的中心点到目标中心点的距离

3：$S \leftarrow$ 选取中心点距离 $g$ 的中心点最近的 $n$ 个样本作为候选正样本

4：$C_g = C_g \cup S$

5：计算 $C_g$ 和 $g$ 之间的IoU值：$D_g = \mathrm{IoU}(C_g, g)$

6：计算 $D_g$ 的均值：$M = \mathrm{Mean}(D_g)$

7：计算ground-truth的IoU阈值：$T = M$

8：for $c \in C_g$ do

9： if IoU$(c,g) > T$，且 $c$ 的中心点在 $g$ 中 then

10： $P = P \cup c$

11： end if

12：end for

13：$N = A - P$

14：return $P, N$

当锚框与目标真实框之间的IoU值高于阈值 $T$ 时，该锚框被视作正训练样本；反之，则被视作负训练样本。一般来说，锚框的中心越接近目标中心，IoU值也越大，这将产生更高质量的锚框，因此选取距离目标最近的样本作为候选正样本。文献[3]在目标检测领域提出了一种自适应选择正负训练样本的算法ATSS，根据对象的统计特征自动选择正负训练样本，计算出候选正样本与目标真实框之间的IoU的均值和标准差，再计算两者之和，即为正负训练样本的界定阈值，其中，计算标准差是为了选择一个最合适的特征金字塔层。

本文使用AlexNet作为特征提取方式，仅使用第5层提取到的特征，因此只计算每个候选正样本与目标真实框之间的IoU的均值作为界定正负训练样本的阈值。均值 $M$ 可以反映预定义锚框与目标真实框的匹配程度。若均值高，代表该目标与预定义锚框匹配程度高，应适当调高阈值来调整正训练样本；反之，均值低，代表匹配程度低，应适当降低阈值。这样动态地调整阈值可以适应大多数的训练样本，达到提升跟踪性能的效果。

### 2.2 IoU约束的质量评估

一些基于孪生网络的目标跟踪方法把目标跟踪任务看作是一次性目标检测任务，通过对视频序列的每一帧图像进行目标检测来达到跟踪的目的。例如，SiamRPN通过借鉴目标检测方法Faster R-CNN中的区域建议网络来对跟踪目标进行检测，具体为：输入某一帧视频图像时，使用分类分支对视频图像中的目标进行分类，利用回归分支对目标的位置进行精确回归。但是，该方法在训练时，分类分支与回归分支相对独立，然而在跟踪阶段却又直接利用分类得分去选择对应回归分支的回归框作为最终目标位置，分类分支与回归分支的低相关性将损害跟踪器的精度。

SiamFC++[4]通过引入质量评估分支来缓解该问题，其对每一个预定义锚框进行评估，并将质量评估得分与分类得分的乘积作为最终得分，实验结果表明，该方法提高了分类分数与定位精度的相关性。然而，质量评估分支和分类分支在整个训练过程中相互独立，却直接将质量评估得分与分类得分的乘积作为最终得分，这会造成质量评估分支和分类分支之间的低相关性，从而影响跟踪性能。除此之外，目标跟踪任务不需要跟踪模型预测多个类别的目标，只需知道被跟踪目标的位置即可。目标的位置信息可以通过锚框与目标真实框之间的IoU来反映。因此，本文使用IoU质量评估分支代替传统的分类分支。IoU质量评估分支不仅有效反映了目标的位置，还降低了模型的参数量。

如图1所示，本文提出的基于IoU约束的孪生网络目标跟踪框架的孪生网络部分与SiamRPN相同，区域建议网络中的分类分支被IoU质量评估分支取代。孪生网络部分将预处理好的模板图像(记作 $z$)与搜索图像(记作 $x$)输入到特征提取网络AlexNet中，得到模板图像特征与搜索图像特征，分别记作 $\boldsymbol{\varphi}(z)$ 和 $\boldsymbol{\varphi}(x)$。

区域建议网络由互相关部分和监督部分组成。其中，监督部分有2个分支，分别为IoU质量评估分支和回归分支。如果有 $k$ 个锚框，则网络需要输出 $k$ 个通道用于IoU质量评估和 $4k$ 个通道用于回归。因此，互相关操作部分需要先通过 $3 \times 3$ 的卷积操作将 $\boldsymbol{\varphi}(z)$ 分成 $[\boldsymbol{\varphi}(z)]_{\mathrm{lq}}$ 和 $[\boldsymbol{\varphi}(z)]_{\mathrm{reg}}$

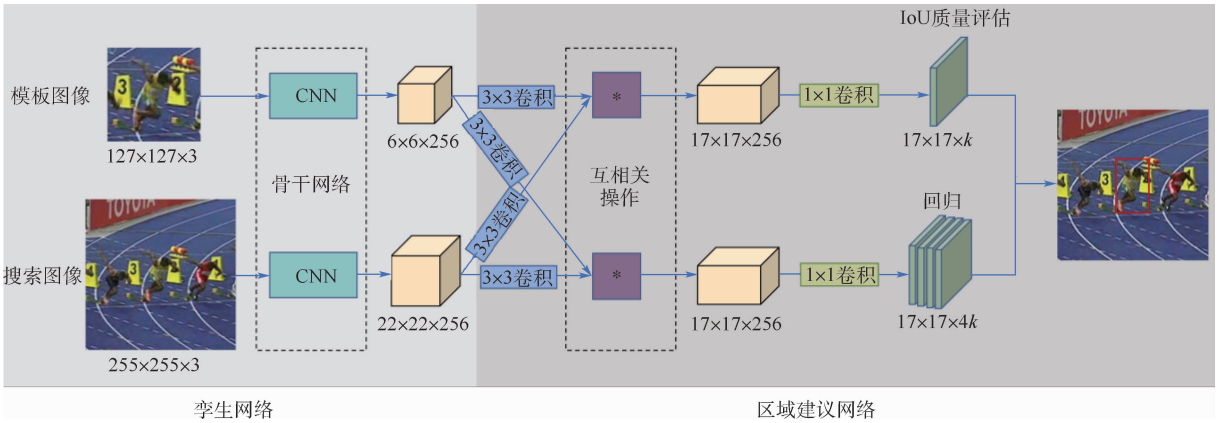



图 1　基于 IoU 约束的孪生网络目标跟踪框架

Fig. 1　Object tracking framework based on IoU-constrained Siamese network

2 个分支。同样,$\boldsymbol{\varphi}(x)$也被 3×3 的卷积操作分为 2 个分支$[\boldsymbol{\varphi}(x)]_{\text{lq}}$和$[\boldsymbol{\varphi}(x)]_{\text{reg}}$。卷积后的特征经过互相关操作计算相似性,如下:

$$\boldsymbol{p}^{\text{lq}}_{w\times h\times k} = [\boldsymbol{\varphi}(x)]_{\text{lq}} * [\boldsymbol{\varphi}(z)]_{\text{lq}} \tag{1}$$

$$\boldsymbol{p}^{\text{reg}}_{w\times h\times 4k} = [\boldsymbol{\varphi}(x)]_{\text{reg}} * [\boldsymbol{\varphi}(z)]_{\text{reg}} \tag{2}$$

式中:$k$ 为每个位置预定义锚框的个数;"*"表示卷积操作;$[\boldsymbol{\varphi}(x)]_{\text{lq}}$和$[\boldsymbol{\varphi}(x)]_{\text{reg}}$为卷积操作的卷积核;$\boldsymbol{p}^{\text{lq}}_{w\times h\times k}$为 IoU 质量评估分支上经过互相关操作之后得到的特征向量;$\boldsymbol{p}^{\text{reg}}_{w\times h\times 4k}$为回归分支上经过互相关操作之后得到的特征。

对于 IoU 质量评估分支,将互相关后的特征$\boldsymbol{p}^{\text{lq}}_{w\times h\times k}$经过一个 1×1 的卷积操作后输入到 IoU 质量评估头中,输出 $k$ 维 IoU 质量评估得分图。回归分支的操作与 IoU 质量评估分支类似,互相关后的特征$\boldsymbol{p}^{\text{reg}}_{w\times h\times 4k}$经过一个 1×1 的卷积操作后输入到回归头中,输出 $4k$ 维坐标偏移量。这样,由 IoU 质量评估分支与回归分支共同作用,能够更加精确地定位目标。

采用 Li 等[19]提出的损失函数来训练 IoU 质量评估分支,具体函数表达式为

$$\text{Lquality} = -|y-\sigma|^{\beta}[(1-y)\ln(1-\sigma)+y\ln(\sigma)] \tag{3}$$

式中:$y$ 为质量评估分支标签;$\sigma$ 为质量评估分支预测结果;$\beta$ 为控制比例常数。

回归损失采用 IoU 损失函数,如式(4)所示。在跟踪阶段,先选择 IoU 质量评估分支输出的最大 IoU 的候选框,再根据最大 IoU 的候选框选择对应的回归框作为最终的跟踪目标位置。

$$\text{loss}_{\text{IoU}} = -\ln\left(\frac{\text{Intersection}(A,B)}{\text{Union}(A,B)}\right) \tag{4}$$

式中:$\text{loss}_{\text{IoU}}$为回归损失函数;$\text{Intersection}(A,B)$为预测框 $A$ 与目标真实框 $B$ 的交集;$\text{Union}(A,B)$为预测框 $A$ 与目标真实框 $B$ 的并集。

## 3　数据集与实验设置

### 3.1　数据集及评价指标

为保证实验结果的可靠性,本文选用目标跟踪领域的常见数据集进行了实验评估。

1) OTB 数据集[20]。2013 年提出,是用于评估跟踪器性能的视频数据集。OTB 数据集将目标跟踪中存在的各种挑战总结归纳为 11 种,分别对应视频的不同属性,数据集中的每个视频对应至少 2 种属性。另外,OTB 数据集中视频的每一帧图像都有经过手工标注的跟踪框及跟踪目标的中心点坐标。OTB 数据集使用一次性通过评估模式得到精度图和成功率图 2 个评价指标。

2) VOT 数据集。最初是因每年举办的 VOT 目标跟踪挑战赛而提出,现已成为评估跟踪器性能的必测数据集之一。为了能更好地评估跟踪器的性能,VOT 数据集每年都会更新并提高挑战难度。VOT2016[21]与 VOT2019[22]数据集都包含了 60 个不同的挑战视频,视频序列的每一帧都用手工进行详细标注。VOT 挑战赛主要采用 3 个衡量标准来分析跟踪性能,即精度、鲁棒性和期望平均重叠率。

① 精度的计算方法为:首先,计算每一帧的平均精度,计算方法如下:

$$\boldsymbol{\Phi}_t(i) = \frac{1}{N_{\text{rep}}}\sum_{k=1}^{N_{\text{rep}}}\boldsymbol{\Phi}_t(i,k) \tag{5}$$

式中:$N_{\text{rep}}$为跟踪器在一个视频序列上的重复次数;$\boldsymbol{\Phi}_t(i,k)$为第 $i$ 个跟踪器在第 $t$ 帧重复次数为 $k$ 时的精度。

然后,计算整个视频序列的平均精度,如下:

$$\rho_{\text{A}}(i) = \frac{1}{N_{\text{valid}}}\sum_{j=1}^{N_{\text{valid}}}\boldsymbol{\Phi}_j(i) \tag{6}$$

式中:$N_{\text{valid}}$为有效视频帧的数量。

② 鲁棒性的值可以反映出跟踪器在整个跟踪过程中失败的次数。当跟踪器跟踪目标时,如果某一视频帧的重叠率为 0 时,则表示跟踪失败。鲁棒性的计算方法为

$$\rho_{\text{R}}(i) = \frac{1}{N_{\text{rep}}}\sum_{k=1}^{N_{\text{rep}}} F(i,k) \tag{7}$$

式中:$F(i,k)$为第 $i$ 个跟踪器在重复次数 $k$ 时的失败次数。

③ 期望平均重叠率是 VOT 官方在 VOT2015 数据集后新增的评价指标,其能同时反映跟踪器的精度与鲁棒性。期望平均重叠率越大,代表跟踪器的性能越好。其计算过程为:首先,计算帧数为 $N_s$ 的视频中每帧的平均重叠率,计算式如式(8)所示;然后,计算每个长度为 $N_s$ 的视频序列的平均精度 $\hat{\boldsymbol{\Phi}}_{N_s}$;最后,将整个视频集的长度划分为 $[N_{\text{hi}}, N_{\text{lo}}]$ 区间,求得区间内 $\hat{\boldsymbol{\Phi}}_{N_s}$的平均精度,即为最终的期望平均重叠率值,计算式如式(9)所示。

$$\hat{\boldsymbol{\Phi}}_{N_s} = \frac{1}{N_s}\sum_{s=1}^{N_s} \boldsymbol{\Phi}_s \tag{8}$$

$$\hat{\boldsymbol{\Phi}} = \frac{1}{N_{\text{hi}} - N_{\text{lo}}}\sum_{N_s=N_{\text{lo}}}^{N_{\text{hi}}} \hat{\boldsymbol{\Phi}}_{N_s} \tag{9}$$

式中:$N_s$ 为帧数。

3) UAV123 数据集。是第一个航空视频数据集[23],主要用于评估跟踪器在实时场景下是否适合部署在无人机上,共包含 123 个高清视频序列,评价指标与 OTB 数据集类似。

### 3.2 实验设置

#### 3.2.1 实验环境

实验使用 Ubuntu 18.04 系统,CPU 型号为 Intel 酷睿 i9-9900K,GPU 为 2 张 Nvidia GeForce RTX 2070 显卡,内存大小为 64 GB。深度学习框架为 Pytorch 1.1.0,Python 版本为 3.7。

#### 3.2.2 训 练

先将 AlexNet 在 ImageNet 数据集上进行预训练,再将预训练后的 AlexNet 在 ImageNet VID[8]、Youtube-BB[24]、COCO[25]、ImageNet DET[8]、GOT-10k[26]与 LaSOT[27]6 个数据集上经过 50 个 Epoch 的迭代训练。其中,批尺寸大小设置为 32,模板图像大小为 127 × 127 像素,搜索图像大小为 255 × 255 像素。使用随机梯度下降算法优化模型。前 5 轮中,学习率由 0.001 线性增长至0.01,后 45 轮中,学习率按照对数方式逐步降低至 0.000 1。在前 10 轮的训练中,特征提取网络 AlexNet 的参数被固定,仅训练区域建议子网络。自 11 轮开始,AlexNet 的最后 2 层加入训练。

#### 3.2.3 测 试

使用训练好的模型分别在数据集 VOT2016、OTB-100、VOT2019、UAV123 上进行测试。

## 4 实验结果与分析

### 4.1 VOT2016 数据集上的实验结果

利用精度、鲁棒性和期望平均重叠率 3 个评价指标与较为先进的 8 个跟踪器(SiamBAN[28]、SiamMask[29]、SiamFC + +[4]、SiamRPN + +[16]、SiamRPN[5]、DaSiamRPN[30]、ATOM[31]、SiamFC[7])进行了对比实验。同时,使用参数量来衡量模型的尺寸大小,以 MB 为单位,使用速度来表示跟踪视频的帧率,以帧/s 为单位。实验结果如表 1 所示。

**表 1 不同方法在 VOT2016 数据集上的实验结果**

**Table 1 Experimental results of different methods on VOT2016 dataset**

| 方法 | 精度 | 鲁棒性 | 期望平均重叠率 | 参数量/MB | 速度/(帧·s$^{-1}$) |
|---|---|---|---|---|---|
| 本文方法 | 0.635 | 0.200 | 0.463 | 41.8 | 220 |
| SiamBAN[28] | 0.666 | 0.144 | 0.505 | 410 | 54.53 |
| SiamMask[29] | 0.643 | 0.219 | 0.455 | 82.1 | 55 |
| SiamFC + +[4] | 0.612 | 0.266 | 0.357 | 71.24 | 90 |
| SiamRPN + +[16] | 0.640 | 0.200 | 0.464 | 206 | 35 |
| SiamRPN[5] | 0.618 | 0.238 | 0.393 | 23.8 | 180 |
| DaSiamRPN[30] | 0.610 | 0.220 | 0.411 | 86.3 | 160 |
| ATOM[31] | 0.610 | 0.187 | 0.430 | 108 | 30 |
| SiamFC[7] | 0.530 | 0.460 | 0.235 | 8.92 | 86 |

从表 1 中可以看出,本文方法在精度、鲁棒性、期望平均重叠率和速度指标中均取得了较好的成绩。由于使用 IoU 质量评估分支代替了分类分支,缓解了分类分支与回归分支的低相关性问题,本文方法在精度方面比 SiamFC + + 高了 0.023,比 SiamRPN 高 0.017。此外,SiamBAN 虽然在精度上取得了最高得分,但是其跟踪速度仅为 54.53 帧/s,低于本文方法。动态阈值方法能根据锚框与目标真实框的匹配程度动态调整正负训练样本的界定阈值,来更好地匹配训练样本,使得训练出的模型更具有判别性,因此本文方法相较于 DaSiamRPN 在鲁棒性上降低了 0.02。本文方法与 SiamRPN + + 在期望平均重叠率上仅相差 0.001,说明本文方法能使基于区域建议网络的跟踪器在 AlexNet 上取得较好的性能。

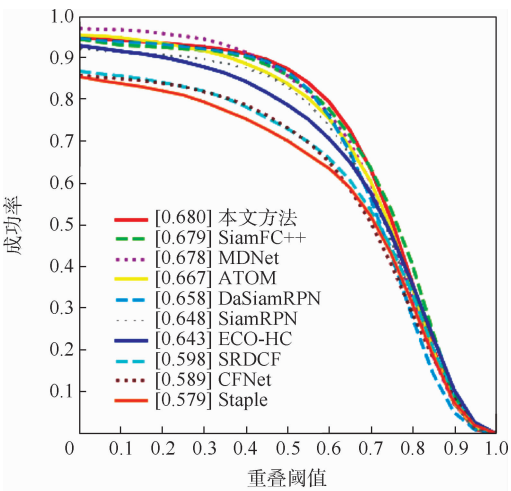
在与其他跟踪器精度相差不大的情况下,本文模型的参数量为 41.8 MB,实时运行速度为 220 帧/s,比大多数参与评估的跟踪器更具有优势。其中,SiamFC + + 引入了质量评估分支,但是

由于其区域建议网络中包含分类、质量评估、回归3个分支,参数量多,运行速度仅为90帧/s。而SiamBAN、SiamMask和SiamRPN++都使用ResNet50作为骨干网络,增大了跟踪器的模型,从而增加了参数量,影响了跟踪的速度。而本文提出的跟踪器的骨干网络为AlexNet,其参数量小于ResNet50,并使用IoU质量评估分支代替原有的分类分支,进一步减少了参数量,提高了实时跟踪速度。

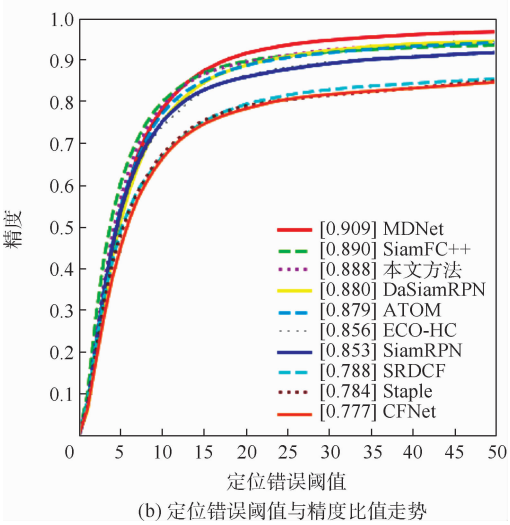
### 4.2　OTB-100数据集上的实验结果

OTB-100数据集是目标跟踪领域中比较经典的评估测试集。本节采用一次性通过评估模式,使用成功率和精度2个评价指标评估本文方法,同时与9个比较先进的跟踪器(CFNet[9]、MDNet[32]、DaSiamRPN[30]、SiamRPN[5]、ECO-HC[33]、Staple[34]、SiamFC++[4]、ATOM[31]、SRDCF[35])做了对比。

如图2所示,本文方法的成功率得分为0.680,高于其他跟踪器;精度得分为0.888,仅次于MDNet和SiamFC++。MDNet通过在线模板更新,获得了最高的跟踪精度,但其跟踪速度仅为1.52帧/s,无法完成实时性需求。本文方法不需要在线微调模型,能以220帧/s的超实时性速度运行。SiamFC++使用GoogleNet作为骨干网络,获得了较高的跟踪精度,但其模型复杂,参数量大,运行速度仅为90帧/s,在实时性上不如本文方法。




(a) 重叠阈值与成功率比值走势




(b) 定位错误阈值与精度比值走势

图2　不同方法在OTB-100数据集上的实验结果

Fig. 2　Experimental results of different methods on OTB-100 dataset

### 4.3　VOT2019数据集上的实验结果

VOT2019数据集是在2019年视觉跟踪挑战赛上提出的,相较于VOT2016数据集,VOT2019数据集具有更高的挑战性。本节使用与VOT2016数据集相同的评价指标,比较了本文方法与其他6个跟踪方法(SiamBAN[28]、SiamRPN++[16]、SiamRPN[5]、SPM[36]、SA-Siam-R[22]、MemDTC[22])在VOT2019数据集上的性能表现。

由表2可知,本文方法在精度、鲁棒性与期望平均重叠率得分分别为0.597、0.522和0.289,相较于基础网络SiamRPN均有较大提升。与其他先进的跟踪方法相比,本文方法也保持了相近的结果。虽然本文方法在部分评价指标上与SiamBAN和SiamRPN++相比具有一定差距,但由于SiamRPN++采用深层网络,在提高精度和鲁棒性的同时降低了实时性;而SiamBAN也因为使用深层网络,需要计算不同层的分类和回归而导致参数量增加,造成跟踪速度下降。

表2　不同方法在数据集VOT2019上的实验结果

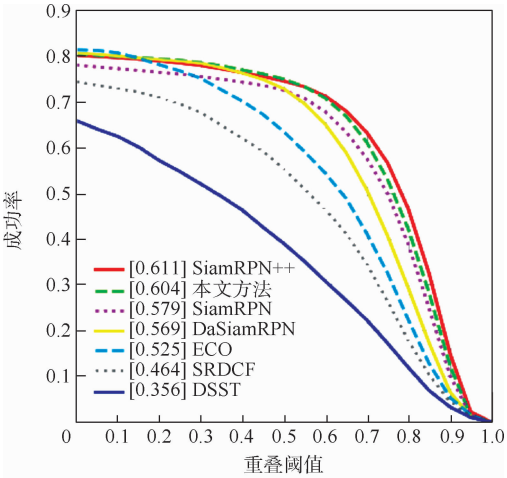
Table 2　Experimental results of different methods on VOT2019 dataset

| 方法 | 精度 | 鲁棒性 | 期望平均重叠率 |
|---|---|---|---|
| 本文方法 | 0.597 | 0.522 | 0.289 |
| SiamBAN[28] | 0.602 | 0.396 | 0.327 |
| SiamRPN++[16] | 0.599 | 0.482 | 0.285 |
| SiamRPN[5] | 0.573 | 0.547 | 0.260 |
| SPM[36] | 0.577 | 0.507 | 0.275 |
| SA-Siam-R[22] | 0.559 | 0.492 | 0.253 |
| MemDTC[22] | 0.485 | 0.587 | 0.228 |

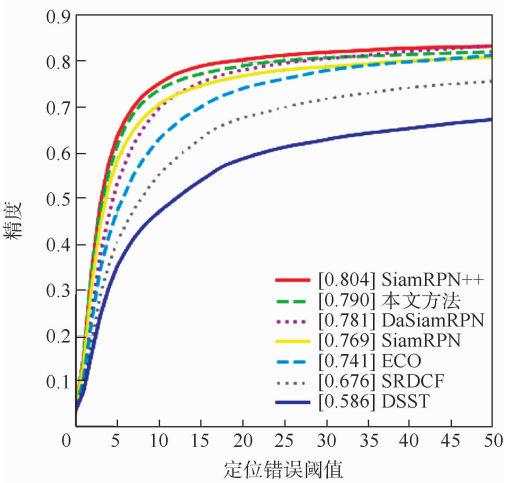
### 4.4　UAV123数据集上的实验结果

UAV123数据集中目标的主要特点为运动速度快、尺度变化大、光照变化大和目标遮挡等,使得跟踪该数据集上的目标具有较强的挑战性。本节采用一次通过评估模式与其他6个跟踪方法(SiamRPN++[16]、SiamRPN[5]、DaSiamRPN[30]、ECO[33]、SRDCF[35]、DSST[37])进行了比较。

如图3所示,本文方法的成功率得分为0.604,精度得分为0.790,成功率和精度得分都仅次于使用了ResNet50作为特征提取网络的SiamRPN++。ResNet50是较深层的骨干网络,




(a) 重叠阈值与成功率比值走势




(b) 定位错误阈值与精度比值走势

图 3 不同方法在 UAV123 数据集上的实验结果

Fig. 3 Experimental results of different methods on UAV123 dataset

能够提取到更具有判别性的特征,因此在这种存在较多小尺寸高速运动的目标的数据集上取得了最好的精度。但是,考虑到目标跟踪任务不仅需要高精度,同时也需要保持较高的实时性。本文方法采用 AlexNet 作为特征提取网络,具有较高的实时性。

## 5 结 论

1) 本文提出了一种基于 IoU 约束的孪生网络目标跟踪方法。通过使用动态阈值代替固定阈值来匹配样本,使得正负训练样本的选择更加合理,提升了跟踪精度。同时,使用 IoU 质量评估分支来解决分类分支与回归分支的低相关性问题,降低了模型参数,提升了跟踪速度和精度。

2) 在数据集 VOT2016、OTB-100、VOT2019、UAV123 上与多个跟踪模型进行了对比实验。相较于其他方法,本文模型在速度上有了较大提升,精度上也有所提升。在 VOT2016 数据集上,跟踪速度为 220 帧/s,明显优于其他跟踪方法,跟踪精度为 0.635,优于大部分跟踪方法。

## 参考文献 (References)

[1] 周千里,张文靖,赵路平,等. 面向个体人员特征的跨模态目标跟踪方法[J]. 北京航空航天大学学报,2020,46(9):1635-1642.
ZHOU Q L,ZHANG W J,ZHAO L P,et al. Cross-modal object tracking algorithm based on pedestrian attribute[J]. Journal of Beijing University of Aeronautics and Astronautics,2020,46(9):1635-1642(in Chinese).

[2] 罗元,肖航,欧俊雄. 基于深度学习的目标跟踪技术的研究综述[J]. 半导体光电,2020,41(6):757.
LUO Y,XIAO H,OU J X. Research on target tracking technology based on deep learning[J]. Semiconductor Optoelectronics,2020,41(6):757(in Chinese).

[3] ZHANG S F,CHI C,YAO Y Q,et al. Bridging the gap between anchor-based and anchor-free detection via adaptive training sample selection[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press,2020:9759-9768.

[4] XU Y D,WANG Z Y,LI Z X,et al. SiamFC + +:Towards robust and accurate visual tracking with target estimation guidelines[C]// Proceedings of the AAAI Conference on Artificial Intelligence. Palo Alto:AAAI,2020:12549-12556.

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

[6] TAO R,GAVVES E,SMEULDERS A W M. Siamese instance search for tracking[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway:IEEE Press,2016:1420-1429.

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

[8] RUSSAKOVSKY O,DENG J,SU H,et al. ImageNet large scale visual recognition challenge[J]. International Journal of Computer Vision,2015,115(3):211-252.

[9] 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,2017:2805-2813.

[10] WANG Q,TENG Z,XING J L,et al. Learning attentions:Residual attentional Siamese network for high performance online visual tracking[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press,2018:4854-4863.

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

[12] REN S,HE K,GIRSHICK R,et al. Faster R-CNN:Towards real-time object detection with region proposal networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,2017,39(6):1137-1149.

[13] FAN H,LING H B. Siamese cascaded region proposal networks for real-time visual tracking[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway:IEEE Press,2019:7952-7961.

[14] KRIZHEVSKY A,SUTSKEVER I,HINTON G E. ImageNet classification with deep convolutional neural networks[C]// Advances in Neural Information Processing Systems,2012:1097-1105.

[15] ZHANG Z P,PENG H W. Deeper and wider Siamese networks for real-time visual tracking[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway:IEEE Press,2019:4591-4600.

[16] LI B,WU W,WANG Q,et al. SiamRPN++:Evolution of Siamese visual tracking with very deep networks[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway:IEEE Press,2019:4282-4291.

[17] 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.

[18] GUO D Y,WANG J,CUI Y,et al. SiamCAR:Siamese fully convolutional classification and regression for visual tracking[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway:IEEE Press,2020:6269-6277.

[19] LI X,WANG W H,WU L J,et al. Generalized focal loss:Learning qualified and distributed bounding boxes for dense object detection[EB/OL]. (2020-06-08)[2021-09-01]. https://arxiv.org/abs/2006.04388.

[20] WU Y,LIM J,YANG M H. Object tracking benchmark[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,2015,37(9):1834-1848.

[21] HADFIELD S J,BOWDEN R,LEBDA K. The visual object tracking VOT2016 challenge results[C]// European Conference on Computer Vision. Berlin:Springer,2016:777-823.

[22] KRISTAN M,MATAS J,LEONARDIS A,et al. The seventh visual object tracking VOT2019 challenge results[C]// Proceedings of the IEEE International Conference on Computer Vision Workshops. Piscataway:IEEE Press,2019:2206-2241.

[23] MUELLER M,SMITH N,GHANEM B. A benchmark and simulator for UAV tracking[C]// European Conference on Computer Vision. Berlin:Springer,2016:445-461.

[24] REAL E,SHLENS J,MAZZOCCHI S,et al. YouTube-BoundingBoxes:A large high-precision human-annotated data set for object detection in video[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway:IEEE Press,2017:5296-5305.

[25] 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.

[26] HUANG L H,ZHAO X,HUANG K Q. GOT-10k:A large high-diversity benchmark for generic object tracking in the wild[EB/OL]. (2019-11-20)[2021-09-01]. https://arxiv.org/abs/1810.11981v2.

[27] FAN H,LIN L T,YANG F,et al. LaSOT:A high-quality benchmark for large-scale single object tracking[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway:IEEE Press,2019:5374-5383.

[28] ZEDU C,BINENG Z,GUORONG L,et al. Siamese box adaptive network for visual tracking[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway:IEEE Press,2020:6668-6677.

[29] WANG Q,ZHANG L,BERTINETTO L,et al. Fast online object tracking and segmentation:A unifying approach[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway:IEEE Press,2019:1328-1338.

[30] ZHU Z,WANG Q,LI B,et al. Distractor-aware Siamese networks for visual object tracking[C]// European Conference on Computer Vision. Berlin:Springer,2018:101-117.

[31] DANELLJAN M,BHAT G,KHAN F S,et al. ATOM:Accurate tracking by overlap maximization[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway:IEEE Press,2019:4660-4669.

[32] 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.

[33] DANELLJAN M,BHAT G,SHAHBAZ K F,et al. ECO:Efficient convolution operators for tracking[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway:IEEE Press,2017:6638-6646.

[34] BERTINETTO L,VALMADRE J,GOLODETZ S,et al. Staple:Complementary learners for real-time tracking[C]// Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. Piscataway:IEEE Press,2016:1401-1409.

[35] DANELLJAN M,HAGER G,SHAHBAZ K F,et al. Learning spatially regularized correlation filters for visual tracking[C]// Proceedings of the IEEE International Conference on Computer Vision. Piscataway:IEEE Press,2015:4310-4318.

[36] WANG G T,LUO C,XIONG Z W,et al. SPM-tracker:Series-parallel matching for real-time visual object tracking[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway:IEEE Press,2019:3643-3652.

[37] DANELLJAN M,HAGER G,KHAN F,et al. Accurate scale estimation for robust visual tracking[C]// British Machine Vision Conference. Berlin:Springer,2014:1-11.

# Object tracking method based on IoU-constrained Siamese network

ZHOU Lifang[1,2,3,*], LIU Jinlan[1,3], LI Weisheng[2,3], LEI Bangjun[4], HE Yu[1,3], WANG Yihan[1]

(1. School of Software Engineering, Chongqing University of Posts and Telecommunications, Chongqing 400065, China;
2. School of Computer Science and Technology, Chongqing University of Posts and Telecommunications, Chongqing 400065, China;
3. Chongqing Key Laboratory of Image Cognition, Chongqing University of Posts and Telecommunications, Chongqing 400065, China;
4. Hubei Key Laboratory of Intelligent Vision Based Monitoring for Hydroelectric Engineering, China Three Gorges University, Yichang 443002, China)

**Abstract**: The tracking method based on the Siamese network trains the tracking model offline. Therefore, it maintains a good balance between tracking accuracy and speed, which attracts the interest of a growing number of researchers recently. The existing Siamese network object tracking method uses a fixed threshold to select positive and negative training samples, which is easy to cause the problem of missing training samples, and such methods have low correlation between the classification branch and the regression branch during training, which is not conducive to training a high-precision tracking model. To this end, an object tracking method based on intersection over union (IoU)-constrained siamese network is proposed. By using a dynamic threshold strategy, the thresholds of positive and negative training samples are dynamically adjusted according to the relevant statistical characteristics of the predefined anchor boxes and the real boxes. Thereby improving the tracking accuracy. In addition, the proposed method uses the IoU quality assessment branch to replace the classification branch, and reflects the position of the target through the IoU between the anchor box and the target ground-truth frame, which improves the tracking accuracy and reduces the amount of model parameters. The proposed object tracking method based on the IoU-constrained Siamese network has been compared and tested on four datasets: VOT2016, OTB-100, VOT2019, and UAV123. Ideal results have been achieved in these datasets. The tracking accuracy of the proposed method in this paper is 0.017 higher than SiamRPN on the VOT2016 dataset. And with a real-time running speed at 220 frame/s, the expected average overlap rate is 0.463, which is only 0.001 worse than SiamRPN++.

**Keywords**: object tracking; deep learning; Siamese network; intersection over union (IoU)-constrained; dynamic threshold

**Received**: 2021-09-06; **Accepted**: 2021-09-17; **Published online**: 2021-11-01 14:56

**URL**: kns.cnki.net/kcms/detail/11.2625.V.20211029.1727.003.html

**Foundation items**: Science and Technology Research Program of Chongqing Education Commission of China (KJZD-K201900601); Natural Science Foundation of Chongqing, China (cstc2019jcyj-msxmX0461); Open Found of Hubei Key Laboratory of Intelligent Vision Based Monitoring for Hydroelectric Engineering (China Three Gorges University) (2020SDSJ01); National College Student Innovation Entrepreneurship Training Program (202110617009)

* **Corresponding author**. E-mail: zhoulf@cqupt.edu.cn

http://bhxb.buaa.edu.cn jbuaa@buaa.edu.cn
**DOI**: 10.13700/j.bh.1001-5965.2021.0518

# 基于动态语义记忆网络的长尾图像描述生成

刘昊, 杨小汕, 徐常胜*
(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**: 图像描述生成任务旨在基于输入图像生成对应的自然语言描述。现有任务数据集中大部分图像的描述语句通常包含少量常见词和大量罕见词,呈现出长尾分布。已有研究专注于提升模型在整个数据集上的描述语句准确性,忽视了对大量罕见词的准确描述,限制了在实际场景中的应用。针对这一问题,提出了基于动态语义记忆网络(DSMN)的长尾图像描述生成模型,旨在保证模型对常见名词准确描述的同时,提升模型对罕见名词的描述效果。DSMN 模型能够动态挖掘罕见词与常见词的全局语义关系,实现从常见词到罕见词的语义知识迁移,通过协同考虑全局单词语义关系信息及当前输入图像和已生成单词的局部语义信息提升罕见词的语义特征表示能力和预测性能。为了有效评价长尾图像描述生成方法,基于 MS COCO Captioning 数据集定义了长尾图像描述生成任务专用测试集 Few-COCO。在 MS COCO Captioning 和 Few-COCO 数据集上的多个量化实验表明,DSMN 模型在 Few-COCO 数据集上的罕见词描述准确率为 0.602 8%,召回率为 0.323 4%,F-1 值为0.356 7%,相较于基准方法提升明显。

**关 键 词**: 深度学习; 图像理解; 图像描述生成; 长尾分布; 记忆网络
**中图分类号**: TP391
**文献标志码**: A **文章编号**: 1001-5965(2022)08-1399-10

图像描述生成(image captioning)任务旨在基于输入图像生成对应的自然语言描述。该任务要求模型具备图像视觉信息的理解能力和自然语言序列的生成能力,并且实现从视觉域到文本域的语义信息转换[1]。近年来,随着深度学习方法的引入,图像描述生成任务得到了快速发展。目前,对于图像描述任务的大部分研究工作都采用“编码器-解码器”(encoder-decoder)架构,即用卷积神经网络(convolutional neural network,CNN)作为输入图像的编码器得到视觉域语义信息特征,再使用循环神经网络(recurrent neural network,RNN)或深度自注意力转换网络(transformer)作为特征的文本域解码器用于生成对应的描述文本序列[2]。基于这一架构,近年来提出了许多改进方法,如注意力机制(attention)[3]、视觉空间特征注意力(spatial feature attention)[4]、视觉哨兵(visual sentinel)[5]、外部知识(external knowledge)[6]、场景图分析(scene graph analysis)[7]等,有效提升了模型的生成效果。

作为一种跨模态生成任务,图像描述生成任务的训练样本是由图像和对应的文本序列组成的。现有数据集中的大部分图像都是用户在日常生活中拍摄的,因此图像的描述语句通常包含少量使用频率极高的词(常见词)和大量使用频率较低的词(罕见词),呈现出长尾分布(long-tail distribution)。以 MS COCO Captioning[8]数据集为例,“狗(dog)”和“车(car)”等日常生活中常见的对象出现频率较高,覆盖样本量在 3 万以上;而

**收稿日期**: 2021-09-06; **录用日期**: 2021-10-01; **网络出版时间**: 2021-11-16 11:37
**网络出版地址**: kns.cnki.net/kcms/detail/11.2625.V.20211115.1939.002.html
**基金项目**: 国家重点研发计划(2018AAA0100604); 国家自然科学基金(61720106006,62036012,62072455,61721004,U1836220,U1705262); 中国科学院前沿科学重点研究计划(QYZDJ-SSW-JSC039); 北京市自然科学基金(L2010011)
* **通信作者**. E-mail: csxu@nlpr.ia.ac.cn

**引用格式**: *刘昊, 杨小汕, 徐常胜. 基于动态语义记忆网络的长尾图像描述生成[J]. 北京航空航天大学学报, 2022, 48(8): 1399-1408. LIU H, YANG X S, XU C S. Long-tail image captioning with dynamic semantic memory network[J]. Journal of Beijing University of Aeronautics and Astronautics, 2022, 48(8): 1399-1408 (in Chinese).*

“白鼬鼠(ferret)”等罕见动物的出现频率很低,覆盖样本量则小于 10。已有的大多数研究工作[2-4,7,9-12]专注于提升模型在整个数据集上的描述语句准确性,并且已取得了较好的性能。但现有方法依赖 CIDEr[13]、SPICE[10]、ROGUE[14]、METEOR[15]及 BLEU[16]等人工定义的语义准确性评价指标来优化模型,因此大多数已有模型往往受到部分常见词的支配,从而忽视了对大量罕见词的描述准确性,限制了已有图像描述生成方法在实际中的应用。

针对上述问题,本文提出基于动态语义记忆网络(dynamic semantic memory network,DSMN)的长尾图像描述生成模型,旨在保证模型对常见词准确描述的同时,提升模型对罕见词,尤其是罕见名词的描述效果。本文提出的 DSMN 模型能够挖掘单词的全局语义关系信息,并与输入图像和已生成单词的局部语义信息相结合进行单词生成。模型主要由 4 个模块组成:①视觉特征编码器,主要功能是基于预训练的深度卷积神经网络提取图像的视觉特征表示;②语义特征解码器,主要功能是基于输入图像特征和已生成单词的局部上下文信息预测下一个单词的局部语义特征;③动态语义记忆单元,主要功能是存储常见词与罕见词的判别性语义表示并进行动态更新,实现常见词与罕见词的全局关系挖掘和语义知识迁移,最终得到要预测单词的全局语义关系信息;④局部-全局语义相结合的单词生成模块,主要功能是融合局部语义特征和全局语义关系信息来协同生成下一个单词。

## 1 相关工作

### 1.1 图像描述生成任务

图像描述生成任务是一种将图像视觉内容转换为自然语言描述的跨模态任务,要求生成的结果能够准确描述图像主要内容(语义准确性),同时保证文本序列本身流利性(语法准确性)[1]。基于传统机器学习方法的研究大多采用基于模版或基于检索的方法。前者利用目标识别模型将图像中的显著目标识别出来并将对应语义标签填入预定义的描述模版中作为描述结果[17-19];后者则是在一个既有的描述候选池中选择一个特征层面最为相似的描述项作为描述结果[20-22]。这 2 类方法的结果因为来自预定义内容而具有很高的语法准确性,但是预定义内容难以完全匹配图像信息,因而存在明显的语义偏差[1]。

近年来,随着深度学习方法的引入,利用深度网络可以更准确地提取输入图像的主要语义内容[23],再通过深度语言模型可以生成语法流利而语义相关的描述语句[24]。2015 年,Vinyals 等[11]使用 Inception 网络作为提取图像语义特征的编码器,使用长短期记忆(LSTM)网络作为生成描述文本的解码器,模型取得了明显优于传统方法的性能,同时接近人类的水平。此后,Xu 等[3]使用注意力机制实现生成过程中的视觉上下文信息建模方法,并从软注意力和硬注意力 2 种机制评估模型的效果。在此基础上,Anderson 等[4]利用预训练的物体检测模型提取图像中的目标空间位置信息和视觉特征,并提出自顶向下和自底向上相结合的注意力机制增强模型性能。Rennie 等[25]提出了基于强化学习思路的自评价序列训练(SCST)方法,通过图像描述生成模型的输出语句构建自评价约束对模型进行更新,有效缓解了已有描述生成模型训练优化目标和测试评价指标不一致的问题。随着深度自注意力转换网络 transformer 在自然语言处理(NLP)领域的广泛采用,Herdade[26]、Guo[27]、He[28]、Cornia[2]等提出了利用 transformer 作为解码器的描述生成模型。这些方法分别从视觉区域位置编码、视觉几何相关的自注意力机制、空间图自注意力机制、分层特征自注意力机制等方面将 transformer 解码过程与视觉语义信息建立关联,并取得了明显的性能提升。还有一些方法[29-30]基于 transformer 的自注意力机制进一步挖掘特征之间的深层语义关系以提升描述效果。此外,还有一些基于跨模态预训练语言模型的描述生成模型在近年被提出,如 ViLBERT[31]和 ERNIE[32],二者采用不同的跨模态 transformer 结构设计,并使用不同的跨模态任务进行预训练,最终将图像描述生成作为下游任务进行模型更新。

上述图像描述生成方法主要研究如何提升模型的描述性能,为了增强图像描述生成方法的实用性,近年来还出现了多个新的研究分支,如多样化描述生成[33-34](diverse image captioning)、风格化描述生成[35-36](stylized image captioning)及新目标描述生成[9,37-38](novel object image captioning)。其中,新目标描述生成任务针对常用图像描述生成数据集覆盖目标类别较少的问题,利用零样本识别的思路,将新目标的外部识别信息融入经典描述生成过程中,实现新目标对应单词的正确生成[19]。此后有多个新目标描述的改进模型被提出,如 Yao 等[39]提出的拷贝机制、Li 等[9]提出的指针机制和 Wu 等[38]提出的动态模版生成方法等。不同于已有的新目标描述生成方法,

本文提出的长尾图像描述生成方法主要针对常用数据集中不同单词词频分布不平衡的问题,重点提升模型对罕见词的描述效果。

### 1.2　记忆网络

记忆网络(memory network)最早是由 Sukhbaatar 等[40]提出,其核心部分是由多个不同质心(centroid)特征表示构成的特征集合,相比于 RNN 和 LSTM 等网络中的隐藏状态向量,其可以学习到更多更复杂的有效信息。训练过程中,记忆网络可以将有效知识信息写入(write-in)特征集合,形成特征层面的知识表示;在测试过程中,网络可以根据输入信息读出(read-out)需要的知识特征表示,从而实现知识利用。记忆网络在自然语言处理领域已经有很多重要应用。例如,在问答、对话生成等任务中,记忆网络可以积累和学习更多的历史信息来提升模型性能[40-41]。同时在一些跨模态任务中,记忆网络也可以用于保存跨模态知识,有效提升模型的效果[42-43]。本文则是在图像描述任务中,通过记忆网络动态学习单词的语义表示,在训练过程中学习和挖掘常见词和罕见词之间的关系,在生成过程中增强罕见词的语义表示,从而提升模型对罕见词的描述准确性。

## 2　模型结构和训练方法

### 2.1　问题定义

图像描述生成任务的数据集由图像和文本样本对$\{\boldsymbol{I}^m, X^m\}$组成,其中,$m \in 1,2,\cdots,M$,$M$ 为数据集中的样本个数,$\boldsymbol{I}^m$ 为输入图像,$X^m = \{x_1, x_2, \cdots, x_T\}$ 为对应的文本描述,$x_t$ 为构成文本描述的单词,$T$ 为单词个数。图像描述生成任务的目标是学习以图像作为输入的跨模态文本生成模型,使其可以为任意输入图像 $\boldsymbol{I}$ 生成能够准确描述图像内容并且语法通顺的自然语句。

为便于介绍,将数据集中的单词集合定义为 $W$,将出现频次小于固定阈值 $\tau$ 的罕见词集合定义为 $W_{\text{rare}} \subset W$。以 MS COCO Captioning 为例,以 50 为频次阈值的罕见词分布情况如图 1 所示。不同于传统的文本描述生成模型,本文提出的长尾图像描述生成任务,旨在保证常见词准确描述的同时,提升模型对罕见词中的罕见名词的描述效果。为了完成这一目标,本文提出 DSMN 模型,其结构如图 2 所示,下面对 DSMN 模型的4 个主要模块及模型的训练方法进行详细介绍。

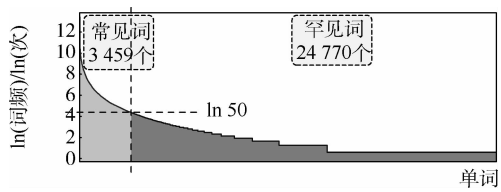



图 1　MS COCO Captioning 数据集中频次小于 50 的罕见词分布示意图

Fig. 1　Distribution diagram of rare words with less than 50 occurrences in MS COCO Captioning dataset

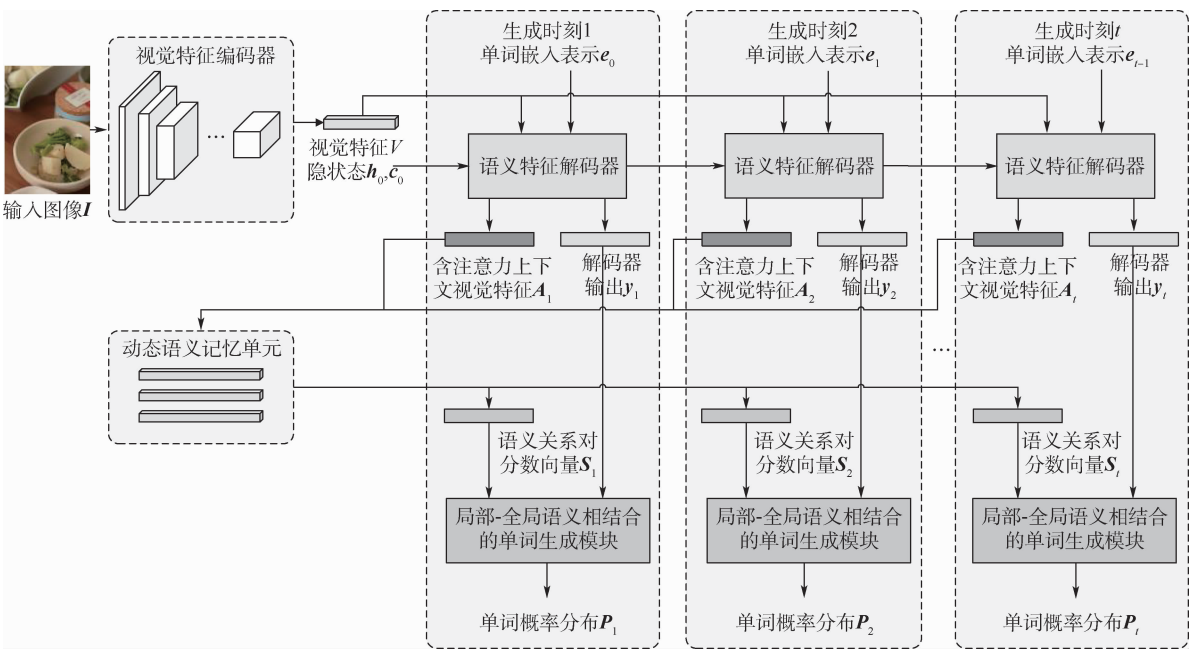



图 2　DSMN 模型结构

Fig. 2　Model architecture of dynamic semantic memory network

### 2.2 视觉特征编码器

与已有的常见图像描述生成方法[2-3]类似，本文采用的视觉编码器由在 ILSVRC-2015 数据集上预训练的 ResNet-101 网络构成。提取最后一个卷积层的特征图(feature map)作为图像的视觉特征表示 $V=\{\boldsymbol{V}_1,\boldsymbol{V}_2\cdots,\boldsymbol{V}_D\}$，$D=14\times14$ 表示局部视觉特征的个数。

### 2.3 语义特征解码器

本文采用以 LSTM 为基础单元的语义解码器。解码器在每一个生成时刻 $t$ 的输入包括前一时刻单词的嵌入表示 $\boldsymbol{e}_{t-1}$、视觉特征集合 $V$ 的线性映射集合 $\hat{V}$ 及 LSTM 前一个时刻的隐状态 $\boldsymbol{h}_{t-1}$ 和 $\boldsymbol{c}_{t-1}$。为提取图像中与当前时刻要预测的单词更为相关的视觉信息，先根据视觉特征 $\hat{\boldsymbol{V}}_i$ 和历史状态 $\boldsymbol{h}_{t-1}$ 计算包含注意力上下文的视觉特征 $\boldsymbol{A}_t$：

$$a_{ti}=f_{\text{att}}(\boldsymbol{h}_{t-1},\hat{\boldsymbol{V}}_i) \tag{1}$$

$$\boldsymbol{\alpha}_t=\text{softmax}(\boldsymbol{a}_t) \tag{2}$$

$$\boldsymbol{A}_t=\sum_i\alpha_{ti}\hat{\boldsymbol{V}}_i \tag{3}$$

式中：$a_{ti}$、$\alpha_{ti}$ 分别为 $\boldsymbol{a}_t$、$\boldsymbol{\alpha}_t$ 的第 $i$ 个元素；$f_{\text{att}}$ 为线性变换函数；$\boldsymbol{a}_t$ 为注意力模块的输出权重；$\boldsymbol{\alpha}_t$ 为 softmax 变换后的注意力权重。

本文将注意力上下文表示作为面向输入样本的局部语义特征，用于动态语义记忆表示的积累以及全局语义特征的提取。

在此之后，把单词嵌入表示 $\boldsymbol{e}_{t-1}$、包含注意力上下文的视觉特征 $\boldsymbol{A}_t$ 和前一个时刻的隐状态 $\boldsymbol{h}_{t-1}$ 输入 LSTM 预测当前时刻隐状态 $\boldsymbol{h}_t$ 和 $\boldsymbol{c}_t$：

$$\boldsymbol{h}_t=\text{LSTM}(\boldsymbol{e}_t,\boldsymbol{h}_{t-1},\boldsymbol{A}_t) \tag{4}$$

最终基于隐状态 $\boldsymbol{h}_t$ 得到解码器输出 $\boldsymbol{y}_t$ 为

$$\boldsymbol{y}_t=\boldsymbol{W}_y\boldsymbol{h}_t \tag{5}$$

式中：$\boldsymbol{W}_y$ 为生成解码器输出的映射变换参数。

### 2.4 动态语义记忆单元

通过 2.3 节介绍的语义解码器，基于图像的视觉特征和已生成的单词迭代生成文本描述。由于罕见词的样本数量较少，这种模型对于罕见词的预测任务无法进行充分的学习，影响了描述模型的性能。本文提出动态语义记忆单元来存储单词的语义信息，并在模型训练中挖掘常见词和罕见词之间的关联，进而增强模型对罕见词的语义表示能力和预测能力。动态语义记忆单元包含 2 种工作模式：写入模式和读出模式。前者利用解码器中生成的注意力上下文特征动态积累单词的语义表示；后者则是通过记忆单元中存储的语义表示计算当前要预测的单词的全局语义特征。

#### 2.4.1 写入模式

在模型训练过程中，动态语义记忆单元的写入模式，主要是对单词语义特征表示进行存储和动态更新。在记忆单元中，每一个单词对应一个语义记忆表示。在模型的每一个生成时刻 $t$，记忆单元将 2.3 节计算得到的包含注意力上下文的视觉特征 $\boldsymbol{A}_t$ 收集到对应单词 $w_n$ 的记忆表示中。具体来说，在每次写入操作中，计算当前单词的表示 $\boldsymbol{A}_t$ 与其对应的记忆单元中的第 $n$ 个单词的记忆表示 $\boldsymbol{M}_n$ 的平均值，并用其更新记忆单元中对应的单词记忆表示。在模型训练结束后，记忆单元实际上存储了数据集中每个单词在不同图像中所代表的关键语义特征。

#### 2.4.2 读出模式

记忆单元的读出模式如图 3 所示。模型可以利用记忆单元存储的单词语义特征表示计算当前要预测单词的全局语义特征。具体来说，在每一个生成时刻 $t$，把包含注意力上下文的视觉特征 $\boldsymbol{A}_t$ 和每一个单词的动态语义记忆表示 $\boldsymbol{M}_n$ 映射到同一个嵌入表示空间：

$$\boldsymbol{A}'_t=\text{ReLU}(\boldsymbol{W}_1,\boldsymbol{A}_t) \tag{6}$$

$$\boldsymbol{M}'_n=\text{ReLU}(\boldsymbol{W}_1,\boldsymbol{M}_n) \tag{7}$$

式中：$\boldsymbol{W}_1$ 为 2 个映射变换采用的共享参数。

将映射后的语义特征 $\boldsymbol{A}'_t$ 和每个单词的记忆表示 $\boldsymbol{M}'_n$ 拼接得到语义关系表示 $\boldsymbol{R}_{tn}$：

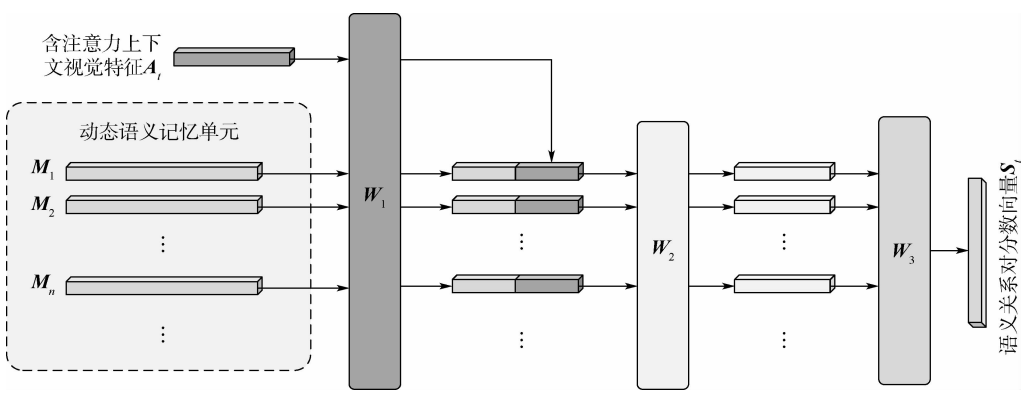



图 3 动态语义记忆单元的读出过程

Fig. 3 Read-out process of dynamic semantic memory unit

$$\boldsymbol{R}_{tn} = \boldsymbol{A}'_t \| \boldsymbol{M}'_n \qquad n = 1,2,\cdots,N \tag{8}$$

式中:"‖"表示特征向量的拼接操作。

在此之后,对语义关系表示进行变换计算当前时刻的单词语义特征与记忆单元中存储的所有单词的记忆表示的关系分数 $S_{tn}$:

$$\boldsymbol{R}'_{tn} = \mathrm{ReLU}(\boldsymbol{W}_2, \boldsymbol{R}_{tn}) \qquad n = 1,2,\cdots,N \tag{9}$$

$$S_{tn} = \sigma(\boldsymbol{W}_3 \boldsymbol{R}'_{tn}) \qquad n = 1,2,\cdots,N \tag{10}$$

式中:$\boldsymbol{W}_2$、$\boldsymbol{W}_3$ 为映射变化的参数。

该语义关系分数表示了当前单词的语义特征与记忆单元中各个单词的记忆表示之间的关联性。记忆单元中的所有单词与当前要预测单词的语义关系分数组成了当前单词的 $N$ 维全局语义关系向量 $\boldsymbol{S}_t$。

### 2.5　局部-全局语义单词生成模块

在 2.3 节中,基于 LSTM 的隐状态 $\boldsymbol{h}_t$ 来预测当前时刻要生成的单词 $\boldsymbol{y}_t$,主要用到了输入图像和已生成单词的上下文信息,将这种生成方式称为局部语义单词生成。这种方法对于常见单词通常有较好的预测效果,但对于罕见词,模型无法得到充分训练。因此,采用局部与全局语义结合的单词生成方法。具体来说,本文通过线性融合 2.3 节计算得到的解码器输出 $\boldsymbol{y}_t$ 和 2.4节得到的全局语义关系分数 $\boldsymbol{S}_t$ 来计算下一时刻要生成单词的概率分布:

$$P_t(w_n) = \begin{cases} \mathrm{softmax}(K y_{tn} + (1-K) S_{tn}) & w_n \in W_{\mathrm{rare}} \\ \mathrm{softmax}(y_{tn}) & w_n \notin W_{\mathrm{rare}} \end{cases} \tag{11}$$

式中:$y_{tn}$为通过局部语义预测得到的下一时刻第 $n$ 个单词的概率;$S_{tn}$为通过全局语义预测得到的下一时刻第 $n$ 个单词的概率;$K$ 为线性融合超参数,可以决定全局语义信息对最终预测的影响,本文将在实验部分进行详细分析。对于常见词,仅通过局部语义就可以得到较好的预测性能。因此在模型训练阶段,只针对罕见词采用局部与全局语义相结合的方式来融合预测。而在模型测试阶段,无法提前知道要预测的下一个单词是常见词或罕见词,因此统一采用局部与全局语义相结合的方式进行融合预测。

### 2.6　优化目标

为了优化本文提出的模型,需要分别针对单词生成任务和记忆单元更新设计优化目标。采用图像描述生成任务中常用的负对数似然(negative likelihood,NLL)损失来训练单词生成模块:

$$\mathcal{L}_{\mathrm{dec}}(\boldsymbol{I}, X) = -\sum_{t=1}^{T} \log_2 p(x_t \mid \boldsymbol{\Theta}) \tag{12}$$

式中:$\boldsymbol{I}$ 和 $X$ 分别为图像和文本描述;$T$ 为文本描述包含的单词个数;$p(x_t \mid \boldsymbol{\Theta})$ 为在时刻 $t$ 生成训练样本中对应单词 $x_t$ 的概率;$\boldsymbol{\Theta}$ 代表模型可学习的参数。

记忆单元的优化目标使用均方误差:

$$\mathcal{L}_{\mathrm{mem}}(\boldsymbol{I}, S, X) = -\frac{1}{T}\sum_{t=1}^{T} (\boldsymbol{S}_t - \hat{\boldsymbol{y}}_t)^2 \tag{13}$$

式中:$\boldsymbol{S}_t$ 表示在时刻 $t$ 基于记忆单元得到的全局语义关系分数向量;$\hat{\boldsymbol{y}}_t$ 为训练样本标注描述中应单词的 one-hot 向量。

最终综合考虑上述 2 种损失函数来优化整个模型:

$$\mathcal{L}(\boldsymbol{I}, S, X) = \mathcal{L}_{\mathrm{dec}}(\boldsymbol{I}, X) + \mathcal{L}_{\mathrm{mem}}(\boldsymbol{I}, S, X) \tag{14}$$

### 2.7　模型训练

在初始训练阶段,模型尚不能产生具有足够语义关联性的注意力上下文特征,因此本文将模型的训练过程分成 3 个阶段,对模型不同模块进行分阶段训练。

1)对单词生成模块的训练。仅使用负对数似然损失 $\mathcal{L}_{\mathrm{dec}}(\boldsymbol{I}, X)$ 来训练模型。动态语义记忆单元不参与训练,单词生成模块的融合超参数 $K$ 设为 1。

2)在学到一个性能稳定的单词生成模型后,利用解码器产生的注意力上下文特征动态更新所有单词的语义记忆表示,并采用 $L_{\mathrm{mem}}$ 损失来约束模型训练。动态语义记忆单元以"写入模式"参与模型训练,单词生成模块的融合超参数 $K$ 设为 1。解码器和单词生成模块不参与优化。

3)基于动态语义记忆单元提升模型对罕见词的语义表示能力和单词生成能力。采用完整的约束函数 $\mathcal{L}_{\mathrm{mem}}(\boldsymbol{I}, S, X)$ 同时训练解码器、单词生成模块和动态语义记忆单元。动态语义记忆单元以"读出模式"参与模型训练,单词生成模块的融合超参数 $K$ 将根据验证集的性能被设置为 0 ~ 1 之间的值。

## 3　实验设计与结果分析

### 3.1　数 据 集

本文实验中所使用的数据集为图像描述任务常用数据集 MS COCO Captioning[8]。该数据集基于 MS COCO 2014 图像数据集构造。每一张图像都由人工标注得到 5 个文本描述。由于官方测试集无法进行线下验证,本文采用文献[44]中给出的基于训练集和验证集的线下划分方案,其中共有 113 287 张图像用于训练,5 000 张图像用于验证,以及 5 000 张图像用于测试。本文使用了与

大量已有工作[3-4,9,39]中相同的预处理方法,对数据集所有文本描述中的词汇做词频过滤,在整个数据集出现总次数少于5次的词汇将会替换为“UNK”标记,最终形成了包含9 487个单词的有效词汇表;设定了最大句子长度为15,超出的部分将会被截断。

此外,为了有效评价长尾图像描述生成方法,本文基于MS COCO Captioning数据集定义了罕见词集合。考虑到名词通常与图像内容更相关,更容易通过挖掘常见词和罕见词的关系来提升预测性能,在整个数据集中出现总次数小于阈值$\tau$的名词定义为罕见词,其余的词汇为常见词。本文使用NLTK工具库[45]中提供的PoS tagging工具判定名词,对MS COCO Captioning数据集中描述文本的名词频率进行了统计,分别以100、80、50、20、10作为罕见词的最高频次进行了统计,如表1所示。在本实验中,设置阈值$\tau$为50。为了提高包含罕见词的样本占比,将MS COCO Captioning Karpathy测试集中包含至少1个罕见词的样本整理成长尾图像描述生成任务专用测试集Few-COCO。

**表1 MS COCO Captioning数据集的词频分布**

**Table 1 Word frequency distribution of MS COCO Captioning dataset**

| 统计项 | 统计值 |
|---|---|
| 全部词汇 | 9 487 |
| 全部名词 | 7 669 |
| 出现100次以下名词 | 6 089 |
| 出现80次以下名词 | 5 859 |
| 出现50次以下名词 | 5 420 |
| 出现20次以下名词 | 4 156 |
| 出现10次以下名词 | 2 888 |

### 3.2 评价指标

在评价指标上,本文遵循已有研究[3-4,9,39]常用的5个性能指标进行生成性能评估,即CIDEr[13]、BLEU[16]、METEOR[15]、ROGUE[14]和SPICE[10]。这5个指标从词频、语料相关性、语义丰富程度等方面衡量生成描述的质量,将其称为生成文本相似性指标。此外,本文还采用精确率(precision)、召回率(recall)和F-1值3个分类任务中常用的指标评估罕见词描述准确性。为了便于计算,定义真阳性(true positive,TP)样本为“出现在真实描述语句中且出现在模型生成语句中”的罕见词;定义假阳性(false positive,FP)样本为“未出现在真实描述语句中但出现在模型生成语句中”的罕见词;定义真阴性(true negative,TN)样本为“出现在真实描述语句中但未出现在模型生成语句中”的罕见词。然后根据上述定义的TP、FP、TN来计算精确率、召回率和F-1这3个指标。

### 3.3 实现细节

本文提出的DSMN模型使用PyTorch 1.6作为模型实现框架,编码器部分的ResNet-101网络使用了TorchVision 0.4.2工具包中的实现,并使用预训练参数来提取视觉特征。具体来说,使用ResNet-101网络的最后一个卷积层的输出作为图像的视觉特征,局部特征向量的个数为$14\times14$,局部特征向量的维度为2 048。在LSTM解码器中,特征维数为512,单词嵌入表示维数为512,注意力上下文特征维数为512。动态语义记忆表示维度为512。在全局语义关系分数计算中,3次映射后的特征向量维度分别为64、32和9 487。

训练过程中,模型使用学习率为$4\times10^{-4}$的Adam优化器[46],dropout概率设为0.5。3个训练阶段的epoch总数分别设置为30、1和25。每训练3个epoch之后将学习率减小为之前的0.8倍。每训练5个epoch之后将规划采样[47](scheduled sampling)系数增加0.05直到0.25。在单词生成模块中,把融合系数设置为0.25。

### 3.4 对比方法

选取2个已有的采用与本文方法类似的CNN编码器和LSTM解码器的图像文本描述方法进行对比。①NIC模型[11]是由一个基于预训练Inception网络的图像视觉特征编码器和一个基于LSTM的语义解码器组成,是最早的基于深度学习的描述生成模型。②Att2in模型[25]采用预训练的ResNet网络作为图像视觉特征编码器并使用LSTM作为解码器。不同于NIC模型,Att2in模型使用注意力机制考虑视觉特征的上下文信息,并选择与当前要预测单词最为相关的局部视觉特征参与单词生成。

### 3.5 量化实验

#### 3.5.1 生成文本相似性评估

在MS COCO Captioning Karpathy测试集上的生成文本相似性结果如表2所示。可以看出,在原始测试集上,本文提出的模型在图像描述生成任务中常用的评价指标上有一定的提升。相比于采用了相同的CNN编码器和带有注意力的LSTM解码器的Att2in模型,本文模型的性能提升较小。但考虑到本文方法的主要目标是在保证常见词描述性能的前提下提升罕见词的描述准确性,这一结果足以说明DSMN模型能够保持在常见词上的描述性能。

为了进一步分析包含罕见词的图像描述生成的结果,在 3.1 节提出的 Few-COCO 测试集上计算生成本文相似性指标,结果如表 3 所示。可以看到,相比表 2 中的结果,本文的 DSMN 模型和已有模型的结果都有所下降。这是由于 Few-COCO 测试集中的所有样本都包含罕见词,相比 MS COCO Captioning Karpathy 测试集具有更大的挑战。比已有模型有明显的提升,表明 DSMN 模型在包含罕见词的样本上具有更稳定的描述生成性能。

**表 2　MS COCO Captioning Karpathy 测试集上的生成文本相似性指标结果**

**Table 2　Results of similarity metrics of generated captions in MS COCO Captioning Karpathy test split**

| 模型 | CIDEr | BLEU | METEOR | ROGUE | SPICE |
|---|---|---|---|---|---|
| NIC[11] | 91.6 | 28.6 | 24.2 | 52.1 | 17.5 |
| Att2in[25] | 99.0 | 30.6 | 25.2 | 53.8 | 18.2 |
| DSMN | 99.3 | 31.1 | 25.4 | 53.8 | 18.8 |

**表 3　Few-COCO 测试集上的生成文本相似性指标结果**

**Table 3　Results of similarity metrics for generated captions in Few-COCO test split**

| 模型 | CIDEr | BLEU | METEOR | ROGUE | SPICE |
|---|---|---|---|---|---|
| NIC[11] | 89.8 | 27.8 | 23.5 | 51.7 | 17.0 |
| Att2in[25] | 96.9 | 30.4 | 25.1 | 53.3 | 18.1 |
| DSMN | 97.2 | 30.4 | 25.2 | 53.2 | 18.4 |

### 3.5.2　罕见词描述准确性评估

在 Few-COCO 测试集上的罕见词描述准确性指标结果如表 4 所示。可以看出,本文提出的 DSMN 模型在罕见词的描述准确性上相比于已有模型有明显提升。这表明 DSMN 模型在保证常见词描述性能的同时,可以显著提升对罕见词的描述效果。

**表 4　Few-COCO 测试集上的罕见词描述准确性结果**

**Table 4　Results of rare word accuracy metrics for generated captions in Few-COCO test split**

| 模型 | 精确率/% | 召回率/% | F-1 值/% |
|---|---|---|---|
| NIC[11] | 0.000 1 | 0.074 8 | 0.001 4 |
| Att2in[25] | 0.373 3 | 0.129 2 | 0.166 2 |
| DSMN | 0.602 8 | 0.323 4 | 0.356 7 |

### 3.5.3　消融实验

为了进一步分析本文提出的 DSMN 模型中的动态记忆单元的作用,对 DSMN 进行模块消融试验。将仅包含编码器和解码器的模型作为基础模型 DSMN-Base,并将记忆单元中记忆表示随机初始化后直接进行第 3 阶段训练的模型作为消融模型 DSMN-w/o MEM。在 Few-COCO 测试集上的罕见词描述准确性指标结果如表 5 所示。可以看到,相比于 DSMN 模型,基础模型 DSMN-Base 和消融模型 DSMN-w/o MEM 的描述准确性有着非常明显的下降。随机初始化记忆表示的消融模型 DSMN-w/o MEM 尽管没有初始的语义特征积累,但在第 3 段训练的优化使其仍然在训练中积累了单词之间的有效关系知识,从而在一定程度上提升了模型对罕见词的描述效果。

**表 5　Few-COCO 测试集上的罕见词描述准确性消融实验结果**

**Table 5　Ablation results of rare word accuracy metrics for generated captions in Few-COCO test split**

| 模型 | 精确率/% | 召回率/% | F-1 值/% |
|---|---|---|---|
| DSMN-Base | 0.373 3 | 0.129 2 | 0.166 2 |
| DSMN-w/o MEM | 0.535 6 | 0.224 5 | 0.266 2 |
| DSMN | 0.602 8 | 0.323 4 | 0.356 7 |

### 3.5.4　超参数分析实验

根据本文的模型设计,单词生成模块中的融合参数 $K$ 决定了模型主要依赖于局部语义特征还是全局语义关系来生成下一个单词。在 Few-COCO 测试集上分析融合参数 $K$ 对 DSMN 模型的性能影响,结果如表 6 所示。可以看出,当 $K$ 取 0.25 时模型具有最佳的性能,增大或减小 $K$ 会降低模型的整体性能。当 $K$ 取过小的值时,单词的预测完全由记忆单元得到的全局关系分数决定,这会导致模型无法充分利用输入图像和已生成单词中包含的上下文语义信息而降低性能;而当 $K$ 取过大的值时,模型将无法充分利用记忆单元中包含的全局语义关系知识,因而影响对罕见词的预测效果。

**表 6　Few-COCO 测试集上融合参数 $K$ 的分析结果**

**Table 6　Results of analyzing combination coefficient $K$ in Few-COCO test split**

| $K$ | 精确率/% | 召回率/% | F-1 值/% |
|---|---|---|---|
| 0.50 | 0.424 3 | 0.137 9 | 0.180 8 |
| 0.25 | 0.602 8 | 0.323 4 | 0.356 7 |
| 0 | 0.400 1 | 0.169 8 | 0.201 4 |

## 3.6　结果可视化分析

为了更加直观地分析本文提出的 DSMN 模型性能,在图 4 中展示了 Att2in 模型和 DSMN 模型的描述生成结果(其中 GT 为数据集中真实描述,单词 broccoli 和 racquet 为罕见词)。可以看出,本文提出的 DSMN 模型准确地生成了“花椰菜(broccoli)”和“网球拍(racquet)”这 2 个罕见词,同时描述文本语句语法正确。此外本文 DSMN 模型也准确描述了图像中的常见词“碗(bowl)”和“网球(tennis)”。Att2in 模型生成的描述只用到了常用词,尽管内容和语法正确,但由于无法准确描述罕见词而丢失了细节信息。这些结果表明和已有方法相比,DSMN 模型在保持良好的常见词

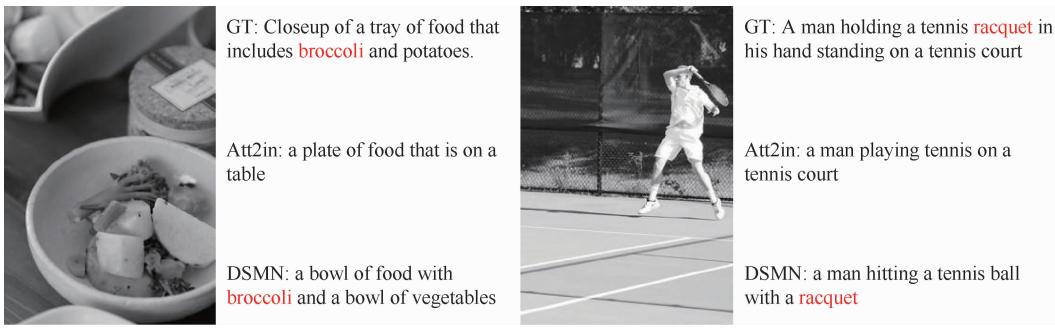



图 4 描述生成结果展示

Fig. 4 Examples of captioning results

描述能力的同时,也能够更准确地描述罕见词。

## 4 结 论

1) 本文提出的 DSMN 模型能够动态挖掘罕见词与常见词的全局语义关系,通过协同考虑全局单词语义关系信息及当前输入图像和已生成单词的局部语义信息提升罕见词的语义特征表示性能。

2) 为了有效评估模型对罕见词的描述性能,定义了罕见词描述专用测试集 Few-COCO,引入分类任务中常用的精确率、召回率和 F-1 指标评估模型对罕见词的描述效果。

3) 相较于基准方法,DSMN 模型在生成文本相似性指标上有 0.3 的提升,罕见词描述准确率提升了约 63%。模型能够在保持常见词描述效果的同时,有效提升对罕见词的描述效果。

DSMN 模型已经证明对数据集中的罕见名词的描述准确性有明显提升,仍需要提升对其他类型罕见词的描述能力,需要改进模型记忆网络的特征提取和读取方法设计及超参数调整,以更好地挖掘单词之间的全局语义关系知识。

### 参考文献 (References)

[1] HOSSAIN M Z, SOHEL F, SHIRATUDDIN M F, et al. A comprehensive survey of deep learning for image captioning [EB/OL]. (2018-10-14) [2021-09-01]. https://arxiv.org/abs/1810.04020.

[2] CORNIA M, STEFANINI M, BARALDI L, et al. Meshed-memory transformer for image captioning [C]// 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2020: 10575-10584.

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

[4] ANDERSON P, HE X, BUEHLER C, et al. Bottom-up and top-down attention for image captioning and visual question answering [C]// 2018 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2018: 6077-6086.

[5] LU J S, XIONG C M, PARIKH D, et al. Knowing when to look: Adaptive attention via a visual sentinel for image captioning [C]// 2017 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2017: 3242-3250.

[6] LU D, WHITEHEAD S, HUANG L F, et al. Entity-aware image caption generation [EB/OL]. (2018-11-07) [2021-09-01]. https://arxiv.org/abs/1804.07889.

[7] YANG X, TANG K H, ZHANG H W, et al. Auto-encoding scene graphs for image captioning [C]// 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2019: 10677-10686.

[8] CHEN X L, FANG H, LIN T Y, et al. Microsoft COCO captions: Data collection and evaluation server [EB/OL]. (2015-04-03) [2021-09-01]. https://arxiv.org/abs/1504.00325.

[9] LI Y H, YAO T, PAN Y W, et al. Pointing novel objects in image captioning [C]// 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2019: 12489-12498.

[10] ANDERSON P, FERNANDO B, JOHNSON M, et al. SPICE: Semantic propositional image caption evaluation [C]// European Conference on Computer Vision. Berlin: Springer, 2016: 382-398.

[11] VINYALS O, TOSHEV A, BENGIO S, et al. Show and tell: A neural image caption generator [C]// 2015 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2015: 3156-3164.

[12] YOU Q Z, JIN H L, WANG Z W, et al. Image captioning with semantic attention [C]// 2016 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2016: 4651-4659.

[13] VEDANTAM R, ZITNICK C L, PARIKH D. CIDEr: Consensus-based image description evaluation [C]// 2015 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2015: 4566-4575.

[14] LIN C Y. ROUGE: A package for automatic evaluation of summaries [C]// Proceedings of the Workshop on Text Summarization Branches Out, 2004: 74-81.

[15] BANERJEE S, LAVIE A. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments [C]// Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization, 2005: 65-72.

[16] PAPINENI K, ROUKOS S, WARD T, et al. BLEU: A method for

automatic evaluation of machine translation[C]// Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics. New York:ACM,2002:311-318.

[17] DEVLIN J, CHENG H, FANG H, et al. Language models for image captioning:The quirks and what works[EB/OL]. (2015-10-14)[2021-09-01]. https://arxiv.org/abs/1505.01809v2.

[18] KULKARNI G, PREMRAJ V, DHAR S, et al. Baby talk:Understanding and generating image descriptions[C]// 2011 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway:IEEE Press,2011:1601-1608.

[19] LI S, KULKARNI G, BERG T, et al. Composing simple image descriptions using web-scale n-grams[C]// Proceedings of the Fifteenth Conference on Computational Natural Language Learning. New York:ACM,2011:220-228.

[20] GONG Y C, WANG L W, HODOSH M, et al. Improving image-sentence embeddings using large weakly annotated photo collections[C]// European Conference on Computer Vision. Berlin:Springer,2014:529-545.

[21] ORDONEZ V, KULKARNI G, BERG T. Im2Text:Describing images using 1 million captioned photographs[J]. Advances in Neural Information Processing Systems,2011,24:1143-1151.

[22] HODOSH M, YOUNG P, HOCKENMAIER J. Framing image description as a ranking task:Data, models and evaluation metrics (extended abstract)[J]. Journal of Artificial Intelligence Research,2013,47:853-899.

[23] REN S Q, HE K M, GIRSHICK R, et al. Faster R-CNN:Towards real-time object detection with region proposal networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,2017,39(6):1137-1149.

[24] GERS F A, SCHMIDHUBER J, CUMMINS F. Learning to forget:Continual prediction with LSTM[J]. Neural Computation, 2000,12(10):2451-2471.

[25] RENNIE S J, MARCHERET E, MROUEH Y, et al. Self-critical sequence training for image captioning[C]// 2017 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway:IEEE Press,2017:1179-1195.

[26] HERDADE S, KAPPELER A, BOAKYE K, et al. Image captioning:Transforming objects into words[EB/OL]. (2020-01-11)[2021-09-01]. https://arxiv.org/abs/1906.05963v1.

[27] GUO L T, LIU J, ZHU X X, et al. Normalized and geometry-aware self-attention network for image captioning[C]// 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway:IEEE Press,2020:10324-10333.

[28] HE S, LIAO W, TAVAKOLI H R, et al. Image captioning through image transformer[C]// Proceedings of the Asian Conference on Computer Vision,2020.

[29] 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.

[30] 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[C]// Proceedings of the AAAI Conference on Artificial Intelligence. Palo Alto: AAAI, 2021, 35(2): 1655-1663.

[31] 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)[2021-09-01]. https://arxiv.org/abs/1908.02265.

[32] SUN Y, WANG S H, LI Y K, et al. ERNIE:Enhanced representation through knowledge integration[EB/OL]. (2019-04-19)[2021-09-01]. https://arxiv.org/abs/1904.09223.

[33] VENUGOPALAN S, HENDRICKS L A, ROHRBACH M, et al. Captioning images with diverse objects[C]// 2017 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway:IEEE Press,2017:1170-1178.

[34] CHEN S Z, JIN Q, WANG P, et al. Say as you wish: Fine-grained control of image caption generation with abstract scene graphs[C]// 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2020:9959-9968.

[35] CHEN T L, ZHANG Z P, YOU Q Z, et al. "Factual" or "emotional":Stylized image captioning with adaptive learning and attention[C]// European Conference on Computer Vision. Berlin:Springer,2018:527-543.

[36] ZHAO W T, WU X X, ZHANG X X. MemCap:Memorizing style knowledge for image captioning[C]// Proceedings of the AAAI Conference on Artificial Intelligence. Palo Alto:AAAI,2020,34(7):12984-12992.

[37] AGRAWAL H, DESAI K R, WANG Y F, et al. Nocaps:Novel object captioning at scale[C]// 2019 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway: IEEE Press,2019:8947-8956.

[38] WU Y, ZHU L, JIANG L, et al. Decoupled novel object captioner[C]// Proceedings of the 26th ACM International Conference on Multimedia. New York:ACM,2018:1029-1037.

[39] YAO T, PAN Y W, LI Y H, et al. Incorporating copying mechanism in image captioning for learning novel objects[C]// 2017 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway:IEEE Press,2017:5263-5271.

[40] SUKHBAATAR S, SZLAM A, WESTON J, et al. End-to-end memory networks[EB/OL]. (2015-11-24)[2021-09-01]. https://arxiv.org/abs/1503.08895v5.

[41] CHEN H, REN Z, TANG J, et al. Hierarchical variational memory network for dialogue generation[C]// Proceedings of the 2018 World Wide Web Conference,2018:1653-1662.

[42] HUANG Y, WANG L. ACMM:Aligned cross-modal memory for few-shot image and sentence matching[C]// 2019 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway:IEEE Press,2019:5773-5782.

[43] HAN J W, NGAN K N, LI M J, et al. A memory learning framework for effective image retrieval[J]. IEEE Transactions on Image,2005,14(4):511-524.

[44] JOHNSON J, KARPATHY A, LI F F. DenseCap:Fully convolutional localization networks for dense captioning[C]// 2016 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway:IEEE Press,2016:4565-4574.

[45] LOPER E, BIRD S. NLTK:The natural language toolkit[C]// Proceedings of the COLING/ACL on Interactive Presentation Sessions,2006:69-72.

[46] KINGMA D P, BA J. Adam:A method for stochastic optimiza-

tion[EB/OL]. (2017-01-30)[2021-09-01]. https://arxiv.org/abs/1412.6980v8? hl:ja.

[47] BENGIO S, VINYALS O, JAITLY N, et al. Scheduled sampling for sequence prediction with recurrent neural networks[EB/OL]. (2015-09-23)[2021-09-01]. https://arxiv.org/abs/1506.03099v3.

# Long-tail image captioning with dynamic semantic memory network

LIU Hao, YANG Xiaoshan, XU Changsheng*

(National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China)

**Abstract**: Image captioning takes image as input and outputs a text sequence. Nowadays, most images included in image captioning datasets are captured from daily life of internet users. Captions of these images are consequently composed of a few common words and many rare words. Most existing studies focus on improving performance of captioning in the whole dataset, regardless of captioning performance among rare words. To solve this problem, we introduce long-tail image captioning with dynamic semantic memory network (DSMN). Long-tail image captioning requires model improving performance of rare words generation, while maintaining good performance of common words generation. DSMN model dynamically mining the global semantic relationship between rare words and common words, enabling knowledge transfer from common words to rare words. Result shows DSMN improves performance of semantic representation of rare words by collaborating global words semantic relation and local semantic information of the input picture and generated words. For better evaluation on long-tail image captioning, we organized a task-specified test split Few-COCO from original MS COCO Captioning dataset. By conducting quantitative and qualitative experiments, the rare words description precision of DSMN model on Few-COCO dataset is 0.602 8%, the recall is 0.323 4%, and the F-1 value is 0.356 7%, showing significant improvement compared with baseline methods.

**Keywords**: deep learning; image understanding; image captioning; long-tail distribution; memory network

**Received**: 2021-09-06; **Accepted**: 2021-10-01; **Published online**: 2021-11-16 11:37

**URL**: kns.cnki.net/kcms/detail/11.2625.V.20211115.1939.002.html

**Foundation items**: National Key R & D Program of China (2018AAA0100604); National Natural Science Foundation of China (61720106006, 62036012, 62072455, 61721004, U1836220, U1705262); Key Research Program of Frontier Sciences, CAS (QYZDJ-SSW-JSC039); Beijing Natural Science Foundation (L2010011)

* **Corresponding author**. E-mail: csxu@nlpr.ia.ac.cn

http://bhxb.buaa.edu.cn jbuaa@buaa.edu.cn
**DOI**: 10.13700/j.bh.1001-5965.2021.0504

# 结合多层特征及空间信息蒸馏的医学影像分割

郑宇祥[1], 郝鹏翼[1,2,*], 吴冬恩[1], 白琮[1,2]

(1. 浙江工业大学 计算机科学与技术学院, 杭州 310023;
2. 浙江省可视媒体智能处理技术研究重点实验室, 杭州 310023)

**摘 要**: U-Net 在医学影像分割领域是目前应用最广泛的分割模型,其"编码-解码"结构也成为了构建医学影像分割模型最常用的结构。尽管 U-Net 在许多领域实现了非常高的分割准确度,但是存在着计算复杂度高、推理速度慢、运行消耗内存大等问题,导致其难以在移动应用平台部署。为解决这一问题,提出了一种结合多层特征及空间信息蒸馏的医学影像分割方法 TinyUnet。该方法使用轻量化的 U-Net 作为学生网络。考虑到小模型没有足够的学习能力,通过选择合适的蒸馏位置,对多层教师特征图进行蒸馏;同时加强教师网络深层特征图的边缘,并构建边缘关键点图结构,采用图卷积网络对学生网络进行空间信息蒸馏,从而补充重要的边缘信息和空间信息。实验表明:在 3 个医学影像数据集上,TinyUnet能够达到 U-Net 98.3% ~99.7%的分割准确度,但是将 U-Net 的参数量平均降低了99.6%,运算速度提高了约 110 倍;同时,与其他轻量化医学影像分割模型相比,TinyUnet 不仅具有较高的分割准确度,而且占用内存更少,运行速度更快。

**关 键 词**: 医学影像分割; 特征蒸馏; 深度学习; 图神经网络; 空间信息
**中图分类号**: TP391
**文献标志码**: A **文章编号**: 1001-5965(2022)08-1409-09

U-Net[1]是一种对称的端到端的卷积神经网络,自 2015 年被提出后,U-Net 凭借在医学影像分割的众多领域(如脑肿瘤、脑白质高信号、超声图像神经分割)上取得的优异表现,迅速成为了医学图像分割的基准模型。尽管 U-Net 及其他衍生而来 U 型结构网络(如 UNet + +[2])在分割精度上不断提高,但随之而来的问题是模型的参数量骤增,导致模型计算复杂度很高,运算消耗的内存巨大,而且必须有强大的 GPU 提供支持才能完成运算,这给移动设备上部署医学影像分割模型造成了极大的困难。

为了压缩模型,研究者通常从模型量化、模型剪枝、低秩分解、知识蒸馏 4 个方面进行研究。文献[3-6]在模型训练中用少量比特的权重替代浮点权重,但是当量化到特殊位宽时,很多现有的训练方法和硬件平台尚未适用,需要设计专用的系统架构,灵活性不高。文献[7]通过特定的剪枝规则,去除网络中参数冗余或不重要的分支,来对模型进行压缩。文献[8]在 U-Net 中引入深度可分离卷积,将三维卷积张量分解为二维滤波器和一维向量。然而这些算法普遍存在压缩后的模型难以降低运算时间、计算成本较高和神经网络难以收敛需要通过对模型反复微调等问题。知识蒸馏通过将教师模型中的知识转移到紧凑小巧的学

**收稿日期**: 2021-08-31; **录用日期**: 2021-09-17; **网络出版时间**: 2021-10-29 14:42
**网络出版地址**: kns.cnki.net/kcms/detail/11.2625.V.20211028.1726.003.html
**基金项目**: 国家自然科学基金(61801428,U20A20196,U1908210); 浙江省自然科学基金(LR21F020002)
* **通信作者**. E-mail: haopy@zjut.edu.cn

**引用格式**: *郑宇祥, 郝鹏翼, 吴冬恩, 等. 结合多层特征及空间信息蒸馏的医学影像分割[J]. 北京航空航天大学学报, 2022, 48(8): 1409-1417. ZHENG Y X, HAO P Y, WU D E, et al. Medical image segmentation based on multi-layer features and spatial information distillation[J]. Journal of Beijing University of Aeronautics and Astronautics, 2022, 48(8): 1409-1417 (in Chinese).*

生模型中,实现模型压缩。与其他方法相比,可使深层网络变浅,极大地降低了计算成本[9],此外由于知识蒸馏的过程就是学生网络学习的过程,因此不需要对学生网络进行多次微调。

在医学影像分割领域,知识蒸馏也有一些应用。例如,文献[10]将 U-Net、nnU-Net[11]、级联 U-Net[12]等模型通过知识蒸馏的方式进行集成学习以提高分割准确率;文献[13]将分割不同病灶的模型通过知识蒸馏的方式进行集成学习,以此让学生模型获得多任务分割的能力;文献[14]将 2 个相互学习模型进行集成,每个模型不仅显式地从相应的标注中提取一种模态知识,而且还以相互引导的方式隐式地从对应的标注中探索另一种模态。这些研究侧重于将多个网络中的知识传递给一个网络来提升分割的准确性,或者使模型能进行多任务分割,但是最终所得到的模型依然是庞大的。

另外一些研究用知识蒸馏的方式得到参数量少的学生模型,以达到模型压缩的效果,同时也引入一些方法来提升学生模型的分割性能。例如,文献[15]在知识蒸馏中引入批量归一化和类别重新加权等方法改善蒸馏效果;文献[16]在模型每一次采样之前使用 L2 距离计算蒸馏损失进行知识传递。但是,这些方法对教师、学生特征图直接进行蒸馏的方式忽略了教师特征图中部分负值特征的作用,导致教师网络传递的知识有限;而且,由于医学影像的边界往往存在采样伪影、空间混叠和噪声等问题,导致参数量较少的学生网络在医学影像分割过程中对边界的分割能力不足。此外,空间信息是医学影像中的一个重要特征,具体表现为直接邻域中的像素具有相似的灰度值,而由于学生网络受模型尺寸的限制,在医学影像分割过程中,会因为感受野过小而提取不到足够的空间信息,导致分割效果不理想。

因此,为了在极大地压缩 U-Net 模型的同时获得更高的分割准确度,本文提出了结合多层特征及空间信息蒸馏的医学影像分割方法 TinyUnet。该方法能够让学生网络在训练过程中基于多层教师特征图从设定的有效信息区间中学习更大范围的特征知识,并由此形成多层蒸馏损失。同时,该方法对教师网络的最深层特征图进行边缘加强,并构建图卷积网络,从而对学生网络进行空间信息蒸馏,形成空间蒸馏损失,使得学生网络对边界的识别能力及提取空间信息的能力有所提升。此外,学生网络根据掩模进行学习,形成分割损失。因此,本文提出的 TinyUnet 总体损失函数由多层蒸馏损失、空间蒸馏损失和分割损失3 部分加权求和组成,通过反向传播来更新参数,得到了占用内存更小、运算速度更快的轻量化学生网络来对医学影像进行分割。TinyUnet 在 NIH[17]和 EM[18]2 个公开数据集及口腔全景片数据集上进行了验证,结果表明,相比于原始 U-Net,TinyUnet 在 3 个数据集上能够保持 98.3% ~99.7% 的分割准确度,但是 TinyUnet 将 U-Net 的参数量平均降低了 99.6%,将运算速度提高了约 110 倍。

## 1 本文方法

本文提出的 TinyUnet 框架如图 1 所示,其主要由教师网络、学生网络、多层特征蒸馏、空间信息蒸馏 4 部分组成。教师网络和学生网络具有一样的结构,但是学生网络的滤波器更少。对于多层特征蒸馏,本文探索了合适的蒸馏位置,并通过一种边缘激活函数,从教师特征图中获取更多有用的特征将其传递给学生网络。对于空间信息蒸馏,本文对教师特征图加强边缘信息,并基于图卷积网络对各结点和相邻结点进行蒸馏,使学生网络学习重要的空间信息。此外,学生网络通过对掩模的学习,进一步更新自身的参数。由此,TinyUnet通过对多层蒸馏损失、空间蒸馏损失和分割损失加权求和得到的总体损失进行反向传播,最终获得参数量小但具有较强分割能力的学生网络模型。

### 1.1 教师网络和学生网络

本文将 U-Net 网络中连续 2 次的卷积操作(conv)、批量归一化操作(batch normlization)表示为一个卷积块 conv_block($c_1$,$c_2$),其中,$c_1$ 表示输入特征的通道数,$c_2$ 表示输出特征的通道数。用 $\vartheta_1 \sim \vartheta_9$ 表示 U-Net 中的 9 个卷积块,用 in_c 表示输入图像的通道数,$N$ 表示第 1 个卷积块的过滤器个数,out_c 表示输出图像的通道数。U-Net 的结构如表 1 所示。

由表 1 可以看出,参数 $N$ 决定着 U-Net 中过滤器的数量,从而影响网络整体的参数量。本文将 $N=64$ 时的 U-Net 作为教师网络,将 $N$ 取较小值时的 U-Net 作为学生网络。

### 1.2 多层特征蒸馏

对于特征蒸馏而言,选择合适位置进行蒸馏至关重要。如果选取单一位置进行特征蒸馏,学生网络会因为学习不到有效的特征而导致其分割性能不佳;相反,如果进行过于频繁的蒸馏,学生网络则会因为失去明确的学习目标而导致性能下降。

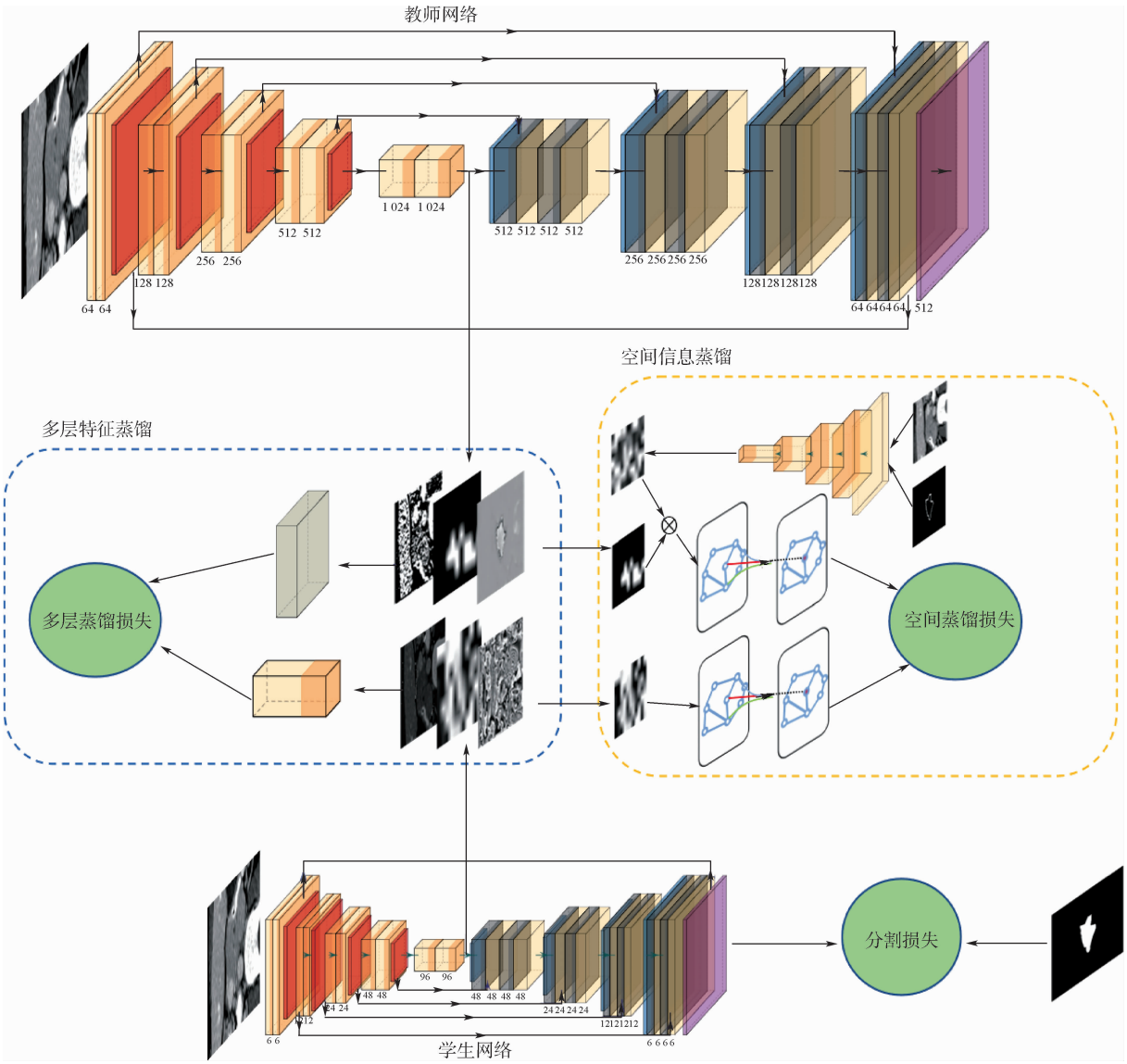



图 1　本文提出的 TinyUnet 框架

Fig. 1　Framework of the proposed TinyUnet

**表 1　U-Net 结构**

**Table 1　Structure of U-Net**

| 卷积块 | 结构 |
|---|---|
| $\vartheta_1$ | conv_block(in_c, $N$) |
| $\vartheta_2$ | MaxPooling→conv_block($N$, $N\times2$) |
| $\vartheta_3$ | MaxPooling→conv_block($N\times2$, $N\times4$) |
| $\vartheta_4$ | MaxPooling→conv_block($N\times4$, $N\times8$) |
| $\vartheta_5$ | MaxPooling→conv_block($N\times8$, $N\times8$) |
| $\vartheta_6$ | UpSampling→conv_block($N\times16$, $N\times4$) |
| $\vartheta_7$ | UpSampling→conv_block($N\times8$, $N\times2$) |
| $\vartheta_8$ | UpSampling→conv_block($N\times4$, $N$) |
| $\vartheta_9$ | UpSampling→conv_block($N$, out_c) |

特征蒸馏中，教师网络的推理过程主要体现在神经元的激活状态。如果经过激活之前的教师特征图中的值为正，说明该值是有效信息，但是如果是负值，根据文献[19]，则需要在激活边界附近寻找阈值，将阈值作为神经元是否被激活的标志。本文将特征图内每个通道的期望作为该阈值，以此让学生网络学习教师网络神经元的激活状态。

对于假定的蒸馏位置 $\boldsymbol{\Phi}_{\{x,y,z\}}$，本文提取教师网络中经过 $\vartheta_x$、$\vartheta_y$、$\vartheta_z$ 块但未被 Relu 函数激活的教师特征图集合 $\boldsymbol{F}^{\mathrm{T}}=\{\boldsymbol{F}_x^{\mathrm{T}},\boldsymbol{F}_y^{\mathrm{T}},\boldsymbol{F}_z^{\mathrm{T}}\}$。对 $\boldsymbol{F}^{\mathrm{T}}$ 中的每个特征图 $\boldsymbol{F}_i^{\mathrm{T}},i\in\{x,y,z\}$，计算 $\boldsymbol{F}_i^{\mathrm{T}}$ 中各通道的期望，得到期望向量$\boldsymbol{\theta}_i\in\mathbf{R}^{1\times C}$，计算方法为

$$\boldsymbol{\theta}_i=E[\boldsymbol{F}_i^{\mathrm{T}j}\mid\boldsymbol{F}_i^{\mathrm{T}j}<0,i\in\{x,y,z\},j\in C] \tag{1}$$

式中：$C$ 为 $\boldsymbol{F}_i^{\mathrm{T}}$ 中的通道数；$\boldsymbol{F}_i^{\mathrm{T}j}\in\mathbf{R}^{1\times H\times W}$ 为 $\boldsymbol{F}_i^{\mathrm{T}}$ 中第 $j$ 个通道上的特征。$\boldsymbol{\theta}_i$ 作为特征图中有效信息的阈值，对 $\boldsymbol{F}_i^{\mathrm{T}}$ 中的特征进行信息筛选，去除小于阈值的特征信息，由此得到了边缘激活函数 $f$：

$$f(\boldsymbol{F}_i^{\mathrm{T}})=\max(\boldsymbol{F}_i^{\mathrm{T}},\boldsymbol{\theta}_i) \tag{2}$$

学生网络经过 $\vartheta_x$、$\vartheta_y$、$\vartheta_z$ 块后，同样得到了未被 Relu 函数激活的学生特征图集合 $\boldsymbol{F}^{\mathrm{S}}=\{\boldsymbol{F}_x^{\mathrm{S}},\boldsymbol{F}_y^{\mathrm{S}},\boldsymbol{F}_z^{\mathrm{S}}\}$。由于教师特征图和学生特征图的通道数存在差异，先对学生特征图进行一个 1×1 的卷积操作，对齐通道数，再进行批量归一化，最终与 $f(\boldsymbol{F}_i^{\mathrm{T}})$ 计算多层蒸馏损失 $\mathcal{L}_{\mathrm{feat}}$：

$$\mathcal{L}_{\mathrm{feat}}=\sum_{i\in\{x,y,z\}}w_i\mathcal{L}_2(f(\boldsymbol{F}_i^{\mathrm{T}}),\mathrm{Bn}(\phi_{\mathrm{s}}(\boldsymbol{F}_i^{\mathrm{S}}))) \tag{3}$$

式中：$\mathcal{L}_2$ 表示欧氏距离；Bn 表示批量归一化操作；$\phi_{\mathrm{s}}$ 表示 1×1 的卷积操作。本文根据特征图的大小设定超参数 $w_i$ 进行加权求和。

### 1.3 空间信息蒸馏

大多数医学影像分割模型往往需要深层的编码器，即堆叠更多的局部操作才能提取到广泛的空间信息。在 U-Net 结构中，下采样通过多个过滤器进行特征映射丢失了部分空间信息，而上采样的反卷积、反池化操作也同样都是局部操作，很难恢复全局空间信息。由于学生网络的通道数减少，获取空间信息的能力进一步减弱，为提升学生网络获取空间信息的能力，受文献[20]的启发，对经过 $\vartheta_5$ 后的特征图上进行了空间信息蒸馏。

#### 1.3.1 边缘信息提取

采用 canny 边缘算子提取掩模的边缘轮廓，用预训练的 ResNet-34 根据边缘轮廓对输入图像进行边缘特征提取，得到了通道数为 $C'$ 的边缘特征 $\boldsymbol{F}_{\mathrm{edge}}\in\mathbf{R}^{C'\times H'\times W'}$。教师网络经过第 5 个卷积块 $\vartheta_5$ 后得到了通道数为 $C^{\mathrm{T}}$ 的特征图 $\boldsymbol{F}_{\mathrm{deep}}^{\mathrm{T}}\in\mathbf{R}^{C^{\mathrm{T}}\times H^{\mathrm{T}}\times W^{\mathrm{T}}}$，$\boldsymbol{F}_{\mathrm{deep}}^{\mathrm{T}}$ 和 $\boldsymbol{F}_{\mathrm{edge}}$ 进行哈达玛积运算得到 $\boldsymbol{F}_{\mathrm{edge}}^{\mathrm{T}'}$，$\boldsymbol{F}_{\mathrm{deep}}^{\mathrm{T}}$ 和 $\boldsymbol{F}_{\mathrm{edge}}^{\mathrm{T}'}$ 进行矩阵乘法加强了边缘信息的教师特征图 $\boldsymbol{F}_{\mathrm{edge}}^{\mathrm{T}}$，即

$$\boldsymbol{F}_{\mathrm{edge}}^{\mathrm{T}}=\boldsymbol{F}_{\mathrm{deep}}^{\mathrm{T}}\otimes(\phi_{\mathrm{T}}(\boldsymbol{F}_{\mathrm{edge}})\odot\mathrm{UpSampling}(\boldsymbol{F}_{\mathrm{deep}}^{\mathrm{T}})) \tag{4}$$

式中：$\phi_{\mathrm{T}}$ 表示 1×1 的卷积操作，使得 $\boldsymbol{F}_{\mathrm{edge}}$ 和 $\boldsymbol{F}_{\mathrm{deep}}^{\mathrm{T}}$ 具有相同的通道数；UpSampling 表示上采样操作，使得 $\boldsymbol{F}_{\mathrm{edge}}$ 和 $\boldsymbol{F}_{\mathrm{deep}}^{\mathrm{T}}$ 具有相同的图像大小；“⊙”表示哈达玛积运算；“⊗”表示矩阵乘法。

学生网络经过 $\vartheta_5$ 后得到通道数为 $C^{\mathrm{S}}$ 的特征图 $\boldsymbol{F}_{\mathrm{deep}}^{\mathrm{S}}\in\mathbf{R}^{C^{\mathrm{S}}\times H^{\mathrm{T}}\times W^{\mathrm{T}}}$。$\boldsymbol{F}_{\mathrm{deep}}^{\mathrm{S}}$ 经过 1×1 卷积后特征图 $\boldsymbol{F}_{\mathrm{edge}}^{\mathrm{S}}\in\mathbf{R}^{C^{\mathrm{T}}\times H^{\mathrm{T}}\times W^{\mathrm{T}}}$。

#### 1.3.2 空间特征提取

对 $\boldsymbol{F}_{\mathrm{edge}}^{\mathrm{S}}$ 构造图 $\boldsymbol{\mathcal{G}}^{\mathrm{S}}=[\boldsymbol{V}^{\mathrm{S}},\boldsymbol{E}^{\mathrm{S}}]$，对 $\boldsymbol{F}_{\mathrm{edge}}^{\mathrm{T}}\in\mathbf{R}^{C^{\mathrm{T}}\times H^{\mathrm{T}}\times W^{\mathrm{T}}}$ 构造图 $\boldsymbol{\mathcal{G}}^{\mathrm{T}}=[\boldsymbol{V}^{\mathrm{T}},\boldsymbol{E}^{\mathrm{T}}]$。$\boldsymbol{\mathcal{G}}^{\mathrm{T}}$ 和 $\boldsymbol{\mathcal{G}}^{\mathrm{S}}$ 表示 $\boldsymbol{F}_{\mathrm{edge}}^{\mathrm{T}}$、$\boldsymbol{F}_{\mathrm{edge}}^{\mathrm{S}}$ 中像素关键点之间的相对关系，$\boldsymbol{V}$ 表示 $\boldsymbol{\mathcal{G}}$ 的结点集合，$\boldsymbol{E}$ 表示 $\boldsymbol{\mathcal{G}}$ 的边的集合。

对于 $\boldsymbol{\mathcal{G}}^{\mathrm{T}}$，将 $\boldsymbol{F}_{\mathrm{edge}}^{\mathrm{T}}$ 经过模糊均值聚类[21]得到中心点集合作为 $\boldsymbol{V}^{\mathrm{T}}\in\mathbf{R}^{|K|\times C^{\mathrm{T}}}$，$K$ 为聚类中心的数量，同时也表示 $\boldsymbol{V}^{\mathrm{T}}$ 中结点数量。每个结点的特征表示为 $\mathbf{vector}^{\mathrm{T}}\in\mathbf{R}^{1\times C^{\mathrm{T}}}$ 的向量。通过随机初始化邻接矩阵 $\boldsymbol{A}^{\mathrm{T}}\in\mathbf{R}^{|K|\times|K|}$ 来表示结点的连通性，作为 $\boldsymbol{E}^{\mathrm{T}}$。同理，对于 $\boldsymbol{\mathcal{G}}^{\mathrm{S}}$，将 $\boldsymbol{F}_{\mathrm{edge}}^{\mathrm{S}}$ 经过模糊均聚类得到关键点，作为结点集合 $\boldsymbol{V}^{\mathrm{S}}\in\mathbf{R}^{|K|\times C^{\mathrm{T}}}$，通过随机初始化 $\boldsymbol{A}^{\mathrm{S}}\in\mathbf{R}^{|K|\times|K|}$ 的邻接矩阵来表示结点的连通性，作为 $\boldsymbol{E}^{\mathrm{S}}$。

在训练阶段，使用一个预训练单层的 GCN 网络来提取 $\boldsymbol{F}_{\mathrm{edge}}^{\mathrm{T}}$ 的空间特征，用另一个随机初始化的单层 GCN 网络来提取 $\boldsymbol{F}_{\mathrm{edge}}^{\mathrm{S}}$ 的空间特征。GCN 网络将结点特征和相邻结点特征相加后进行特征提取，得到新的结点特征 $\boldsymbol{V}^{\mathrm{T}'}\in\mathbf{R}^{|K|\times C^{\mathrm{T}}}$，$\boldsymbol{V}^{\mathrm{S}'}\in\mathbf{R}^{|K|\times C^{\mathrm{T}}}$，即

$$\boldsymbol{V}^{\mathrm{T}'}=\sigma((\tilde{\boldsymbol{D}}^{\mathrm{T}})^{-\frac{1}{2}}(\boldsymbol{A}^{\mathrm{T}}+\boldsymbol{I})(\tilde{\boldsymbol{D}}^{\mathrm{T}})^{-\frac{1}{2}}\boldsymbol{V}^{\mathrm{T}}\boldsymbol{W}_{\mathrm{G}}^{\mathrm{T}}) \tag{5}$$

$$\boldsymbol{V}^{\mathrm{S}'}=\sigma((\tilde{\boldsymbol{D}}^{\mathrm{S}})^{-\frac{1}{2}}(\boldsymbol{A}^{\mathrm{S}}+\boldsymbol{I})(\tilde{\boldsymbol{D}}^{\mathrm{S}})^{-\frac{1}{2}}\boldsymbol{V}^{\mathrm{S}}\boldsymbol{W}_{\mathrm{G}}^{\mathrm{S}}) \tag{6}$$

式中：$\sigma$ 为非线性的激活函数；$\boldsymbol{I}$ 为单位矩阵；$\boldsymbol{A}$ 为邻接矩阵；$\tilde{\boldsymbol{D}}$ 为 $\boldsymbol{A}+\boldsymbol{I}$ 的度矩阵；$\boldsymbol{W}_{\mathrm{G}}^{\mathrm{T}}$ 和 $\boldsymbol{W}_{\mathrm{G}}^{\mathrm{S}}$ 为 GCN 网络的权重。

由于 GCN 网络推理得到的结点特征 $\boldsymbol{V}^{\mathrm{T}'}$、$\boldsymbol{V}^{\mathrm{S}'}$ 难以直接进行比较计算，将 $\boldsymbol{V}^{\mathrm{T}'}$、$\boldsymbol{V}^{\mathrm{S}'}$ 与 $\boldsymbol{F}_{\mathrm{deep}}^{\mathrm{T}}$、$\boldsymbol{F}_{\mathrm{deep}}^{\mathrm{S}}$ 进行矩阵乘法运算，得到 $\boldsymbol{F}_2^{\mathrm{T}'}$ 和 $\boldsymbol{F}_2^{\mathrm{S}'}$。通过计算 $\boldsymbol{F}_2^{\mathrm{T}'}$ 和 $\boldsymbol{F}_2^{\mathrm{S}'}$ 之间的 L2 距离，作为空间蒸馏损失 $\mathcal{L}_{\mathrm{spa}}$：

$$\mathcal{L}_{\mathrm{spa}}=\mathcal{L}_2((\boldsymbol{V}^{\mathrm{T}'}\otimes\phi_{\mathrm{T}}(\boldsymbol{F}_{\mathrm{deep}}^{\mathrm{T}})),(\boldsymbol{V}^{\mathrm{S}'}\otimes\phi_{\mathrm{S}}(\boldsymbol{F}_{\mathrm{deep}}^{\mathrm{S}}))) \tag{7}$$

### 1.4 总体损失函数

在学生网络进行图像分割时，需要对分割结果进行约束，从而使学生网络根据真实标签进行学习。本文采用二分类交叉熵函数作为分割损失 $L_{\mathrm{seg}}$：

$$L_{\mathrm{seg}}=-[(\mathrm{Mask})\log_2(\mathrm{Out}_{\mathrm{seg}}^{\mathrm{S}}))+(1-\mathrm{Mask})\log_2(1-\mathrm{Out}_{\mathrm{seg}}^{\mathrm{S}})] \tag{8}$$

式中：Mask 表示掩模；$\mathrm{Out}_{\mathrm{seg}}^{\mathrm{S}}$ 为学生网络的分割结果。

综上，该模型的总损失函数为

$$\mathcal{L}_{\mathrm{total}}=\alpha\mathcal{L}_{\mathrm{feat}}+\beta\mathcal{L}_{\mathrm{spa}}+\gamma\mathcal{L}_{\mathrm{seg}} \tag{9}$$

式中：$\alpha$、$\beta$ 和 $\gamma$ 为超参数，采用 Adam 优化算法对损失函数进行优化。本文提出的 TinyUnet 方法整体流程如算法 1 所示。

**算法 1** 本文提出的 TinyUnet 方法。

**输入**：图像数据集 Image，掩模集合 Mask，超参数 $N$、$K$、$\Phi$。

**输出**：医学影像分割模型 TinyUnet。

1. 初始化神经网络参数,预处理 Image 中的图像。

2. Repeat

对 $\Phi$ 中各蒸馏位置提取教师特征图集合 $\boldsymbol{F}^{\mathrm{T}}$ 和学生特征图集合 $\boldsymbol{F}^{\mathrm{S}}$。

使用式(1)~式(3),计算得到 $\mathcal{L}_{\mathrm{feat}}$。

根据 Mask 使用 canny 算子生成边缘轮廓。

使用式(4)得到教师边缘特征图 $\boldsymbol{F}^{\mathrm{T}}_{\mathrm{edge}}$,对 $\boldsymbol{F}^{\mathrm{S}}_{\mathrm{edge}}$ 进行卷积操作得到 $\boldsymbol{F}^{\mathrm{S}}_{\mathrm{edge}}$。

对 $\boldsymbol{F}^{\mathrm{T}}_{\mathrm{edge}}$ 和 $\boldsymbol{F}^{\mathrm{S}}_{\mathrm{edge}}$ 构建图 $\boldsymbol{\mathcal{G}}^{\mathrm{T}}$ 和 $\boldsymbol{\mathcal{G}}^{\mathrm{S}}$。

使用式(5)~式(7),得到 $\mathcal{L}_{\mathrm{spa}}$。

使用式(8)得到 $\mathcal{L}_{\mathrm{seg}}$。

使用式(9)得到 $\mathcal{L}_{\mathrm{total}}$。

3. 直到达到总损失函数的收敛阈值或迭代次数上限。

4. 获得轻量模型 TinyUnet。

## 2　实验与分析

### 2.1　数据集

为了验证本文方法的有效性,在一个口腔全景片数据集和 2 个公共数据集上进行了实验。口腔全景片数据集由浙江中医药大学口腔医院提供,共有 916 张大小为 2 440 × 1 280 的口腔全景片。本文采用高斯模糊进行数据增强,采用对比度限制自适应直方图均衡化来增强全景片的局部对比度。按照 60%、20%、20% 的比例将数据集划分为训练集、验证集和测试集。

NIH 数据集[17]由 82 个 3D 腹部 CT 扫描和相应的胰腺分割图像组成。经过感兴趣区域裁剪,最终获得了一个包含 7 059 张 176 × 112 的二维图像的数据集,其中 60% 的图像用于训练模型,20% 的图像用于验证,另外 20% 用于测试。EM 数据集[18]由 EM 细胞分割挑战赛提供,是一组来自果蝇第一龄幼虫腹侧腹神经索的连续透射电镜切片的 90 张图像,其图像大小为 200 × 200,每个图像都带有一个对应的标注图,包含细胞(白色)和膜(黑色),由于该数据集较小,本文将其中 78 张图像作为训练集,剩余的 12 张图像用作测试。

### 2.2　实现细节和评价指标

本文使用 Pytorch 来构建模型,并在 GTX2080 Ti GPU 上进行训练和测试。对于本文提出的 TinyUnet,先对教师网络、边缘特征生成网络和用于对教师特征图提取空间信息 GCN 进行预训练,这些网络在学生网络的训练中不参与反向传播。对于口腔全景片数据集、NIH 数据集、EM 数据集,本文分别设定学习率为 0.005、0.01、0.001,优化器使用了 Adam 优化器,损失函数超参数 $\alpha=\beta=\gamma=1$。

在医学影像分割领域,最常用的分割标准是 Dice 系数得分,用分割的面积占标签图像中真实面积的百分比来衡量分割的准确度,不受数据不平衡的影响,Dice 得分计算公式如式(10)所示,范围为 0~1,值越大,说明分割准确度越高。

$$\mathrm{Dice}=\frac{2\sum_{n=1}^{N}p_n r_n+\epsilon}{\sum_{n=1}^{N}p_n+\sum_{n=1}^{N}r_n+\epsilon} \tag{10}$$

式中:$p_n\in[0,1]$ 表示预测像素;$r_n\in\{0,1\}$ 表示标签像素;$\epsilon$ 是为了保证计算稳定加入的噪声。

此外,用模型的参数量(# Parameter)和 Pytorch框架下模型实际占用的物理空间(Size)来衡量模型的大小。采用同一并行度下计算浮点运算数(floating point operations, FLOPs)衡量模型的实际运行速度。

### 2.3　本文方法的参数选择

U-Net 中第 1 个卷积块的过滤器数量 $N$ 控制着网络的模型大小,能够影响网络的分割准确率。本文在 NIH 数据集上对 $N$ 的选择进行了分析,此处仅仅采用 U-Net,同时将 $N$ 分别设置为 4,6,8,16,32,其分割 Dice 值如表 2 所示。当 $N=64$ 时,等同于原始的 U-Net 模型,不做任何压缩。$N$ 越小,学生网络的参数量和占用的内存空间也就越小,运算速度会加快,但是分割效果也会变差。综合考虑,选择第 1 个卷积块的过滤器数量 $N=6$ 的 U-Net 作为本文方法中的学生网络。

确定了学生网络之后,需要从教师网络中进行多层特征蒸馏,而蒸馏位置的选择至关重要。由表 1 的 U-Net 结构中可以看出,教师网络和学生网络都含有 9 个卷积块,为了更好地确定蒸馏位置,此处暂不使用空间信息蒸馏。本文在 3 个数据集上进行了实验,不同蒸馏位置下的分割 Dice

**表 2　不同参数 $N$ 的实验结果**

**Table 2　Experimental results of different parameters $N$**

| $N$ | Dice | #Parameter/$10^3$ | Size/MB | GFLOPs |
|---|---|---|---|---|
| 4 | 0.677 | 58.557 | 0.271 | 0.455 |
| 6 | 0.716 | 131.119 | 0.547 | 1.003 |
| 8 | 0.718 | 232.537 | 0.938 | 1.766 |
| 16 | 0.723 | 926.769 | 3.59 | 6.960 |
| 32 | 0.741 | 3 700 | 14.18 | 27.630 |

值如表 3 所示。其中:$\Phi_{\{0\}}$ 表示不选取任意蒸馏位置,即不进行蒸馏,与其他蒸馏位置对比可视为消融实验,证明多层特征蒸馏的有效性。可以看出,如果仅对单层特征图进行蒸馏,选择 $\Phi_{\{9\}}$ 作为蒸馏位置有较高的分割精度,当对多层特征图进行蒸馏时,选择 $\Phi_{\{1,5,9\}}$ 作为蒸馏位置能得到最高的分割精度,但是蒸馏的次数大于 3 时,则会因为过于频繁的特征蒸馏导致模型的准确率降低。

在确定了学生网络和多层特征蒸馏的位置之后,对于本文提出的 TinyUnet 中的空间信息蒸馏,需要选择合适的聚类个数 $K$。$K$ 值越大,表示能获得更多的边缘关键点,可以在分割网络的解码器中还原更丰富的边缘细节;$K$ 值越小,表示考虑了更大范围内的像素点,因此在构建 GCN 的过程中能够使用更广的空间信息。本文在 3 个数据集上对 $K=\{0,4,8,16,32\}$ 进行了实验,分割 Dice 值如表 4 所示。当 $K=0$ 与 $K\in\{4,8,16,32\}$ 对比时,可视为消融实验,证明了所提方法的有效性。此外,对于边缘较清晰的全景片数据集,采用较多的聚类中心能得到较高的分割精度,当 $K=32$ 时,Dice 值比其他 3 种 $K$ 得到的 Dice 值高出了 0.02 ~ 0.08;对于边缘较模糊的 NIH 数据集和边缘不明显的 EM 数据集,当 $K=8$ 时,分割准确度较高,这是由于较少的聚类中心提供了更大的感受野来补充空间信息。

**表 3 不同蒸馏位置的 Dice 值**

**Table 3 Dice score of different distillation locations**

| 蒸馏位置 | Dice | | |
|---|---|---|---|
| | 口腔全景片数据集 | NIH 数据集 | EM 数据集 |
| $\Phi_{\{0\}}$ | 0.886 | 0.716 | 0.911 |
| $\Phi_{\{1\}}$ | 0.897 | 0.717 | 0.924 |
| $\Phi_{\{5\}}$ | 0.891 | 0.721 | 0.922 |
| $\Phi_{\{9\}}$ | 0.900 | 0.724 | 0.924 |
| $\Phi_{\{1,5,9\}}$ | 0.903 | 0.728 | 0.929 |
| $\Phi_{\{1,3,5,7,9\}}$ | 0.911 | 0.726 | 0.923 |
| $\Phi_{\{1,2,3,4,5,6,7,8,9\}}$ | 0.893 | 0.722 | 0.920 |

**表 4 不同聚类个数 $K$ 的 Dice 值**

**Table 4 Dice score of different cluster number $K$**

| $K$ | Dice | | |
|---|---|---|---|
| | 口腔全景片数据集 | NIH 数据集 | EM 数据集 |
| 0 | 0.903 | 0.728 | 0.929 |
| 4 | 0.903 | 0.730 | 0.931 |
| 8 | 0.904 | 0.744 | 0.932 |
| 16 | 0.909 | 0.744 | 0.930 |
| 32 | 0.911 | 0.745 | 0.930 |

## 2.4 在口腔全景片数据集上的结果和分析

本文将提出的 TinyUnet 与 U-Net、未蒸馏的学生网络 Unet-6、只进行多层特征蒸馏的学生网络 Unet-6-dis、Unet-fixed[5]、LightUnet[16] 和 EMKD[22] 进行了对比。其中,采用文献[5,22]提供的代码对 Unet-fixed 和 EMKD 进行实验;根据文献[16]在 Pytorch 框架下实现了 LightUnet。这些方法在口腔全景片数据集中的实验结果如表 5 所示。可以看出,TinyUnet 的 Dice 值仅仅比 U-Net 降低了 0.003,但是 TinyUnet 的参数量仅是 U-Net 的 0.4%,TinyUnet 的模型大小仅是 U-Net 的 0.97%。对比其他方法,TinyUnet 的 Dice 值比 EMKD[22] 仅低了 0.001,比 LightUnet[16] 高出了 0.005。参数量最小的是 LightUnet,但是 TinyUnet 的运行速度在所有方法中最快,根据 cudnn 平台计算模型运行的 FLOPs 可以看出,TinyUnet 的运行速度是 LightUnet 的 4.6 倍,是 Unet-fixed 的 117 倍,是 EMKD 的 2 倍。图 2 给出了不同方法在同一实例上的分割结果。通过将真实标签的边缘附在各个方法的分割图上(如图 2 蓝色边缘所示),可以清晰地看出仅仅使用较少过滤器的 U-Net(Unet-6)的分割效果最不理想,经过了多层特征蒸馏之后的 Unet-6(Unet-6-dis)表现稍好,Unet-fixed、LightUnet 对边界的处理能力稍弱,而 TinyUnet 对牙齿边界的处理非常接近于 U-Net,由此说明 TinyUnet 在采用较少参数量的同时,通过多层特征蒸馏和空间信息蒸馏,能够获得较强的边缘识别能力。

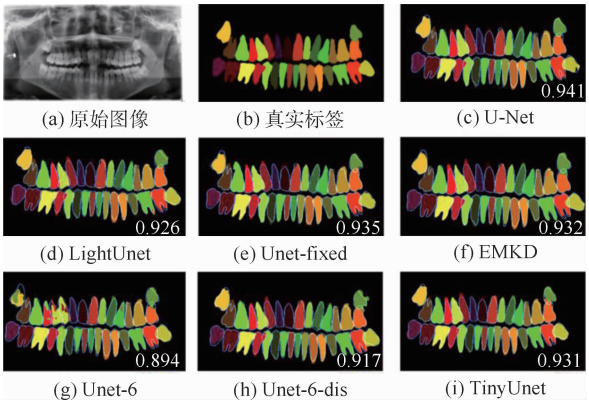
**表 5 不同方法在口腔全景片数据集上的结果**

**Table 5 Results of different methods on oral panoramic film dataset**

| 方法 | Dice | #Parameter/$10^6$ | Size/MB | GFLOPs |
|---|---|---|---|---|
| U-Net | 0.914 | 34.5 | 56.6 | 110.5 |
| Unet-fixed | 0.67 | 4.84 | 9.23 | 117.5 |
| LightUnet | 0.906 | 0.066 7 | 0.473 | 4.633 |
| EMKD | 0.912 | 0.353 | 1.59 | 2.031 |
| Unet-6 | 0.889 | 0.131 | 0.547 | 1.003 |
| Unet-6-dis | 0.903 | 0.131 | 0.547 | 1.003 |
| TinyUnet | 0.911 | 0.131 | 0.547 | 1.003 |

## 2.5 在 NIH 和 EM 数据集上的结果和分析

对于公开的 NIH 胰脏分割数据集和 EM 细胞分割数据,U-Net、Unet-fixed、LightUnet、EMKD、Unet-6、Unet-6-dis、TinyUnet 的实验结果如表 6 所示。




(a) 原始图像　(b) 真实标签　(c) U-Net　(d) LightUnet　(e) Unet-fixed　(f) EMKD　(g) Unet-6　(h) Unet-6-dis　(i) TinyUnet

图 2　不同方法对同一口腔全景片的分割结果

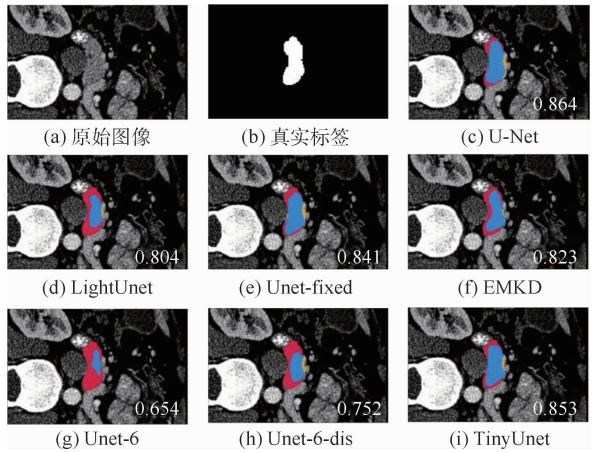
Fig. 2　Segmentation results of the same panoramic radiograph by different methods

**表 6　不同方法在 NIH 和 EM 数据集上的结果**

**Table 6　Results of different methods on NIH and EM datasets**

| 方法 | Dice | | #Parameter/$10^6$ | Size/MB | GFLOPs |
|---|---|---|---|---|---|
| | NIH 数据集 | EM 数据集 | | | |
| U-Net | 0.757 | 0.936 | 34.5 | 56.6 | 110.5 |
| Unet-fixed | 0.746 | 0.920 | 4.84 | 9.23 | 117.5 |
| LightUnet | 0.741 | 0.926 | 0.066 7 | 0.473 | 4.633 |
| EMKD | 0.748 | 0.932 | 0.353 | 1.59 | 2.031 |
| Unet-6 | 0.716 | 0.928 | 0.131 | 0.547 | 1.003 |
| Unet-6-dis | 0.728 | 0.929 | 0.131 | 0.547 | 1.003 |
| TinyUnet | 0.744 | 0.932 | 0.131 | 0.547 | 1.003 |

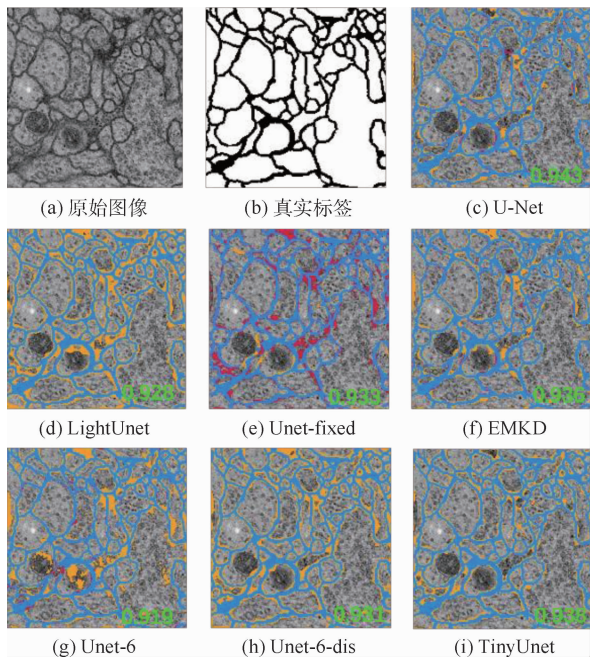
从表 6 中可以看出，TinyUnet 能够在尽可能保持高分割准确度的同时极大地降低模型参数量和内存，并极大程度地加快运行速度。在 NIH 数据集中，TinyUnet 保持了 U-Net 98.3% 的准确度，比 LightUnet 的分割准确度提升了 0.003；对于 EM 数据集，TinyUnet 保持了 U-Net 99.6% 的准确度，比 LightUnet 的分割准确度提升了 0.006。在 2 个数据集上，对比 U-Net 和 Unet-fixed 方法，TinyUnet 将模型内存大小降低了 94.1% ~99.03%。尽管 EMKD 在 NIH 数据集上的分割准确度比 TinyUnet 高 0.004，但是 TinyUnet 的参数量仅为 EMKD 的 37%，运行速度是 EMKD 的 2 倍。此外，虽然 LightUnet 的参数量比 TinyUnet 小，但是 TinyUnet 的运行速度是 LightUnet 的 4.7 倍。图 3 给出了对于同一胰脏实例的分割结果，图 4 给出了对于同一细胞的分割结果，其中红色区域表示假阴性区域，蓝色区域表示真阳性区域，橙色区域表示假阳性区域。可以看出，使用多层特征蒸馏的学生网络尽管比未蒸馏的学生网络表现优异，但是同时进行多层特征蒸馏和空间信息蒸馏的学生网络表现更加优异，在 NIH 和 EM 这样目标边




(a) 原始图像　(b) 真实标签　(c) U-Net　(d) LightUnet　(e) Unet-fixed　(f) EMKD　(g) Unet-6　(h) Unet-6-dis　(i) TinyUnet

图 3　不同方法对同一胰脏实例的分割结果

Fig. 3　Segmentation results of the same pancreas instance by different methods



(a) 原始图像　(b) 真实标签　(c) U-Net　(d) LightUnet　(e) Unet-fixed　(f) EMKD　(g) Unet-6　(h) Unet-6-dis　(i) TinyUnet

图 4　不同方法对同一细胞实例的分割结果

Fig. 4　Segmentation results of the same cell instance by different methods

缘模糊或者目标边缘不明显的数据集上，TinyUnet 通过从教师网络中学习隐含特征和空间信息，提升学生网络对边缘的识别能力。

## 3　结　论

对于小模型由于模型参数少往往导致医学影像分割精度不佳的问题，本文提出了一种有效的结合多层特征及空间信息蒸馏的医学影像分割方法 TinyUnet。主要结论如下：

1）通过多层特征蒸馏监督学生网络学习教师网络神经元的激活状态，以此获取更多的隐含信息。

2）通过加强教师网络最深层特征图的边缘，

并构建图卷积网络来蒸馏教师网络特征图中的空间信息,从而使得学生网络有效地学习图像中邻近像素间的空间信息,并监督学生网络关注识别难度较大的目标边缘。

3)在 3 个不同类型的医学影像分割数据集上与多个先进的小型的医学影像分割模型进行了对比和分析并对提出的 TinyUnet 进行了参数选取,结果表明,TinyUnet 可以在高分割准确度的前提下得到比现有方法更小更轻运行速度更快的网络模型。

## 参考文献(References)


[1] RONNEBERGER O, FISCHER P, BROX T. U-Net: Convolutional networks for biomedical image segmentation[C]// Medical Image Computing and Computer-Assisted Intervention, 2015:234-241.

[2] ZHOU Z W, SIDDIQUEE M M R, TAJBAKHSH N, et al. UNet++:A nested U-Net architecture for medical image segmentation[EB/OL]. (2018-07-18)[2021-08-30]. https://arxiv.org/abs/1807.10165.

[3] HUBARA I, COURBARIAUX M, SOUDRY D, et al. Binarized neural networks[C]// Proceedings of the 30th International Conference on Neural Information Processing Systems. New York ACM, 2016:4114-4122.

[4] 饶川,陈靓影,徐如意,等.一种基于动态量化编码的深度神经网络压缩方法[J].自动化学报,2019,45(10):1960-1968.
RAO C, CHEN J Y, XU R Y, et al. A dynamic quantization coding based deep neural network compression method[J]. Acta Automatica Sinica, 2019, 45(10):1960-1968(in Chinese).

[5] ASKARIHEMMAT M, HONARI S, ROUHIER L, et al. U-Net fixed-point quantization for medical image segmentation[C]// LABELS 2019, HALMICCAI 2019, CuRIOUS 2019. Berlin: Springer, 2019:115-124.

[6] SON S, NAH S, LEE K M. Clustering convolutional kernels to compress deep neural networks[C]// European Conference on Computer Vision. Berlin: Springer, 2018:225-240.

[7] HE Y, LIU P, WANG Z W, et al. Filter pruning via geometric median for deep convolutional neural networks acceleration[C]// 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2019:4335-4344.

[8] VAZE S, NAMBURETE A. Segmentation of fetal adipose tissue using efficient CNNs for portable ultrasound[C]// PIPPI 2018, DATRA 2018. Berlin: Springer, 2018:55-65.

[9] 董晓,刘雷,李晶,等.面向稀疏卷积神经网络的 GPU 性能优化方法[J].软件学报,2020,31(9):2944-2964.
DONG X, LIU L, LI J, et al. Performance optimizing method for sparse convolutional neural networks on GPU[J]. Journal of Software, 2020, 31(9):2944-2964(in Chinese).

[10] LACHINOV D, SHIPUNOVA E, TURLAPOV V. Knowledge distillation for brain tumor segmentation[C]// International MICCAI Brainlesion Workshop. Berlin: Springer, 2019: 324-332.

[11] ISENSEE F, JAEGER P F, KOHL S A A, et al. nnU-Net: A self-configuring method for deep learning-based biomedical image segmentation[J]. Nature Methods, 2021, 18(2):203-211.

[12] LACHINOV D, VASILIEV E, TURLAPOV V. Glioma segmentation with cascaded U-Net[C]// International MICCAI Brainlesion Workshop. Berlin: Springer, 2018:189-198.

[13] HUANG Z C, WANG Z X, CHEN J, et al. Real-time colonoscopy image segmentation based on ensemble knowledge distillation[C]// 2020 5th International Conference on Advanced Robotics and Mechatronics(ICARM). Piscataway: IEEE Press, 2020:454-459.

[14] LI K, YU L Q, WANG S J, et al. Towards cross-modality medical image segmentation with online mutual knowledge distillation[EB/OL]. (2020-10-04)[2021-08-30]. https://arxiv.org/abs/2010.01532.

[15] MANGALAM K, SALZAMANN M. On compressing U-net using knowledge distillation[EB/OL]. (2016-12-01)[2021-08-30]. https://arxiv.org/abs/1812.00249.

[16] VAZE S, XIE W, NAMBURETE A. Low-memory CNNs enabling real-time ultrasound segmentation towards mobile deployment[J]. IEEE Journal of Biomedical and Health Informatics, 2020:20(4):1059-1069.

[17] ROTH H R, LU L, FARAG A, et al. DeepOrgan: Multi-level deep convolutional networks for automated pancreas segmentation[C]// International Conference on Medical Image Computing and Computer-Assisted Intervention. Berlin: Springer, 2015:556-564.

[18] CARDONA A, SAALFELD S, PREIBISCH S, et al. An integrated micro- and macroarchitectural analysis of the Drosophila brain by computer-assisted serial section electron microscopy [J]. PLoS Biology, 2010, 8(10):e1000502.

[19] HEO B, LEE M, YUN S, et al. Knowledge transfer via distillation of activation boundaries formed by hidden neurons[C]// Proceedings of the AAAI Conference on Artificial Intelligence. Palo Alto: AAAI, 2019, 33:3779-3787.

[20] TE G S, LIU Y L, HU W, et al. Edge-aware graph representation learning and reasoning for face parsing[C]// European Conference on Computer Vision. Berlin: Springer, 2020: 258-274.

[21] BEZDEK J C, EHRLICH R, FULL W. FCM: The fuzzy $c$-means clustering algorithm[J]. Computers & Geosciences, 1984, 10(2-3):191-203.

[22] QIN D, BU J J, LIU Z, et al. Efficient medical image segmentation based on knowledge distillation[J]. IEEE Transactions on Medical Imaging, 2021, 40(12):3820-3831.

# Medical image segmentation based on multi-layer features and spatial information distillation

ZHENG Yuxiang[1], HAO Pengyi[1,2,*], WU Dong'en[1], BAI Cong[1,2]

(1. College of Computer Science and Technology, Zhejiang University of Technology, Hangzhou 310023, China;
2. Key Laboratory of Visual Media Intelligent Processing Technology of Zhejiang Province, Hangzhou 310023, China)

**Abstract**: U-Net is currently the most widely used segmentation model, and its "coding-decoding" structure has also become the most commonly used structure for building medical image segmentation models. Although U-Net has achieved very high segmentation accuracy in many fields, but there are problems such as highcomputational complexity, slow reasoning speed, and high memory consumption, which makes it difficult to deploy on mobile application platforms. To solve this problem, a medical image segmentation method combining multi-layer features and spatial information distillation, named as TinyUnet, is proposedin this paper. This method uses the U-Net with fewer parameters as the student network, which is smaller and lighter than the original U-Net. Considering that the small model does not have enough learning ability, this method distils the multi-layer teacher feature maps by selecting the appropriate distillation position; at the same time, this method strengthens the edge of the deep feature map of the teacher network, constructs the edge key point map structure, and uses the graph convolution network to distil the spatial information of the student network, so as to guide the student network to obtain more effective edge information and spatial information. Experiments show that TinyUnet can maintain the segmentation accuracy of U-Net from 98.3% to 99.7% on the three medical datasets, but reduces the parameters of U-Net by 99.6% on average and increases the computing speed by about 110 times. Meanwhile, compared with other advanced compact medical image segmentation models, TinyUnet not only achieves good segmentation accuracy but also occupies less memory and runs faster.

**Keywords**: medical image segmentation; feature distillation; deep learning; graph neural network; spatial information

**Received**: 2021-08-31; **Accepted**: 2021-09-17; **Published online**: 2021-10-29 14:42

**URL**: kns.cnki.net/kcms/detail/11.2625.V.20211028.1726.003.html

**Foundation items**: National Natural Science Foundation of China (61801428, U20A20196, U1908210); Zhejiang Provincial Natural Science Foundation of China (LR21F020002)

* **Corresponding author**. E-mail: haopy@zjut.edu.cn

http://bhxb.buaa.edu.cn jbuaa@buaa.edu.cn
**DOI**: 10.13700/j.bh.1001-5965.2021.0509

# 基于彩色三要素的无参考对比度失真图像质量评价方法

丁盈秋[1], 杨杨[1,2,*], 成茗[1], 张卫明[3]

(1. 安徽大学 电子信息工程学院, 合肥 230601; 2. 合肥综合性国家科学中心人工智能研究院, 合肥 230088;
3. 中国科学技术大学 网络空间安全学院, 合肥 230027)

**摘 要**: 图像质量评价是图像处理领域中基本且具有挑战性的问题。对比度失真对图像质量的感知影响较大,目前针对对比度失真图像的无参考图像质量评价研究相对较少。基于此,提出了基于彩色三要素的无参考对比度失真图像质量评价方法,利用彩色三要素的亮度、色调和饱和度3个参数实现了对比度失真图像的质量评价方法。在亮度方面,提取矩特征及图像直方图与均匀分布之间的Kullback-Leibler散度特征。在色调和饱和度方面,分别在HSV空间的H和S通道中提取颜色加权局部二值模式(LBP)直方图特征。利用AdaBoosting BP神经网络训练预测模型。在5个标准图像数据库中进行广泛的实验分析和交叉验证,结果表明,所提方法与现有的对比度失真图像质量评价方法相比,性能有明显的提升。

**关 键 词**: 图像质量评价; 对比度失真; HSV颜色空间; 无参考; 彩色三要素; BP神经网络

**中图分类号**: TP391

**文献标志码**: A **文章编号**: 1001-5965(2022)08-1418-10

近年来,随着网络与通信技术的发展,生活中图像信息大幅度增长。但是,图像在采集、压缩、处理及传播等过程中,由于各种原因经常会出现不同程度和种类的失真,从而影响人们对信息的获取。因此,对图像质量评价的研究也越来越受到关注[1]。图像的评估是通过人类视觉系统(human visual system, HVS)进行的,但是主观评价在时间和经济成本上有局限性,为了克服主观评价的不足,科研人员提出了大量的客观图像质量评价(image quality assessment, IQA)方法。根据参考图像的可用性,将现有的IQA方法划分为全参考(full reference, FR)[1-3]、半参考(reduced reference, RR)[4-6]及无参考(no reference, NR)[7-13]方法。FR方法需要依赖全部原始信息,RR方法只需要部分原始信息,而NR方法完全不依赖于原始信息。

目前的IQA方法主要针对模糊、JPEG压缩及噪声等失真图像[12],鲜有针对对比度失真图像的质量评价研究。另外,在实际应用场合中,由于较难获得原始图像信息,建立符合人类视觉系统的、稳定且针对对比度失真图像的NR-IQA方法就显得尤为重要。目前,针对对比度失真的NR质量评价方法有:文献[7]从图像中提取矩特征和熵特征,并利用图像数据库对其进行自然场景

**收稿日期**: 2021-09-02; **录用日期**: 2021-09-17; **网络出版时间**: 2021-11-02 16:04
**网络出版地址**: kns.cnki.net/kcms/detail/11.2625.V.20211101.1707.014.html
**基金项目**: 安徽省高等学校自然科学基金(KJ2021A0016); 国家自然科学基金(61502007,61871411)
* **通信作者**. E-mail: sky_yang@ahu.edu.cn

**引用格式**: *丁盈秋, 杨杨, 成茗, 等. 基于彩色三要素的无参考对比度失真图像质量评价方法[J]. 北京航空航天大学学报, 2022, 48(8): 1418-1427. DING Y Q, YANG Y, CHENG M, et al. No reference quality assessment method for contrast-distorted images based on three elements of color[J]. Journal of Beijing University of Aeronautics and Astronautics, 2022, 48(8): 1418-1427 (in Chinese).*

统计建模,在自然场景统计模型的基础上,根据图像的自然程度来衡量每幅对比度失真图像的失真程度(CDIQA)方法。文献[8]在 CDIQA 的基础上做了改进,提出了基于改进学习方法的无参考对比度失真图像质量评价(ICDIQA)。文献[9]通过结合局部和全局特征提出了一个 NR 图像质量度量(NIQMC),根据视觉显著性检测技术寻找具有最大信息的最优区域,将最优区域的熵作为局部质量度量,通过计算对比度失真图像和原始图像的均匀分布的直方图之间的 Jensen-Shannon(JS)散度来测量全局对比度失真。文献[10]利用相位一致性的熵、边缘强度和 3 个基于对比度能量的特征进行对比失真评估,虽然其是一种 NR-IQA 模型的通用度量,但是对于对比度失真的图像表现出很高的性能。文献[11]寻求对比度失真与相关图像直方图特征之间的关系,定义了特征直方图,其是图像块直方图集的特征向量,利用图像特征直方图的随机性和相应特征值的幅值反映图像对比度的变化。文献[12]提出了一种基于结构的多面统计表示的对比失真图像的无参考质量度量,从 S-CIELAB 颜色空间三通道中提取空间强度、分布和方向 3 个质量度量。文献[13]提出一种基于梯度域和 HSV 空间的无参考对比度失真图像质量度量,通过将梯度域上的局部二值模式(local binary pattern,LBP)算子与 HSV 颜色空间上的颜色矩相结合得到综合质量度量。

尽管上述针对对比度失真图像的 NR-IQA 方法已取得了一定的进展,但仍存在一些不足。一方面,文献[7-10]只计算了亮度信息的失真,忽略了对比度失真对色度的影响。另一方面,文献[11-13]虽然考虑了色度失真,但是并没有考虑全面。本文提出了基于彩色三要素的无参考对比度失真图像质量评价方法,从亮度、色调和饱和度 3 个方面计算对比度失真。首先,对于亮度失真,在灰度上提取 2 组特征:①矩特征,包括均值、方差、偏度和峰度;②图像直方图与均匀分布之间的 KL(Kullback-Leibler)散度和逆 KL 散度。然后,对于图像的饱和度和色调方面,分别在 HSV 空间的 H 和 S 通道中计算颜色加权 LBP 直方图特征。最后,利用 AdaBoosting BP 神经网络[12]对质量模型进行训练。通过在 5 个标准图像数据库中进行广泛的实验分析和交叉验证,结果表明,本文方法与现有方法相比较性能具有明显提升。

## 1　基于彩色三要素的对比度失真图像质量评价

色度学早已表明,通常用亮度、色调和饱和度三要素来表征彩色图像特征。亮度即颜色的亮度,色调描述的是一种纯色(纯黄色、纯红色或纯蓝色)的属性,而饱和度是指一种纯色被白光稀释程度的度量[14]。如图 1 所示,图 1(a)为来自 CID2013 数据库[15]的原图,图 1(b)、(c)为对应的具有不同对比度变化的图像,图 1(d)~(f)为图 1(a)~(c)图像对应的亮度图,图 1(g)~(i)为对应的色调图,图 1(j)~(l)为对应的饱和度图,MOS 为测试图像的主观得分(mean opinion score),$\overline{Y}$ 为亮度均值参数,$\overline{H}$ 为色调均值参数,$\overline{S}$ 为饱和度均值参数。MOS 值的范围为 1~5,值越小代表图像质量越差,对比度失真程度越大。从图 1 中可以主观看出,随着对比度失真程度增加,亮度越来越低,色调变化范围不大,饱和度越来越大(因为白光越来越少),对应的客观参数表示亮度均值越来越小,色调均值在小范围内越来越小,饱和度均值越来越大。这表明主观视觉与客观参数可以对应,三要素参数是表征彩色对比度失真的有效因素。据此,本文提出了基于彩色三要素的无参考对比度失真图像质量评价方法。

本节详细介绍本文方法的实现步骤,分别从亮度、色调及饱和度 3 个方面来提取对应的表征特征,方法流程如图 2 所示。

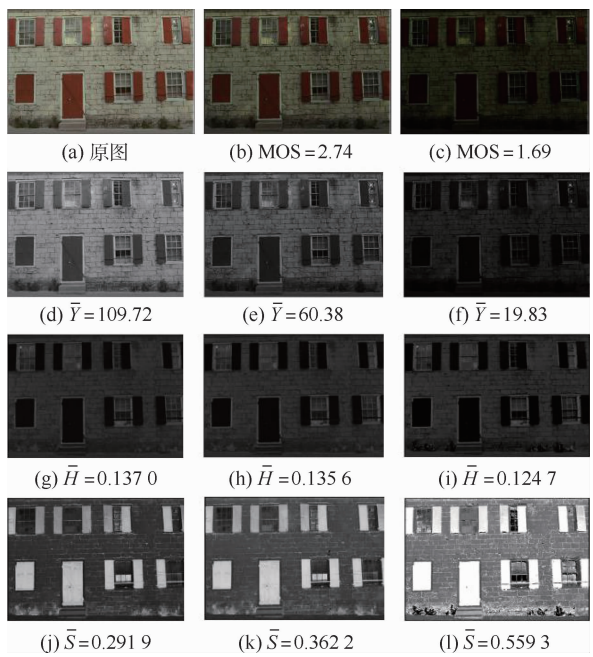



图 1　原图及其对比度失真版本图像的相关参数

Fig. 1　Related parameters of original image and its contrast-distorted version

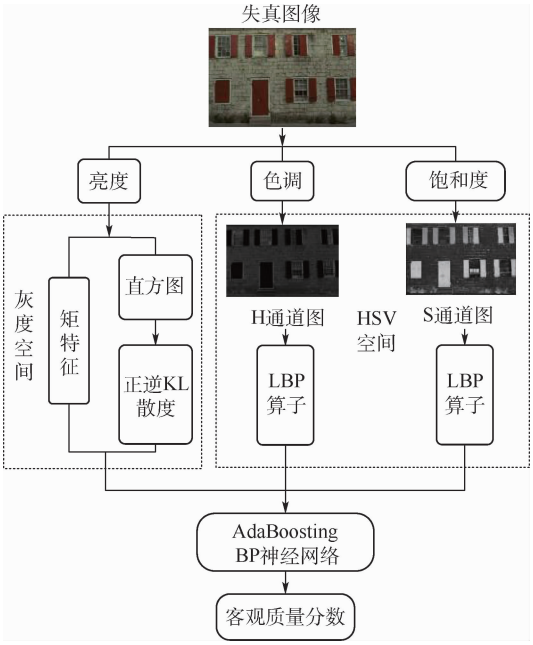



图 2 本文方法流程

Fig. 2 Flow chart of the proposed method

## 1.1 基于亮度的特征提取

### 1.1.1 矩特征

图像的矩特征包括均值、方差、偏度和峰度，在众多针对对比度失真图像质量评价的研究中得到了广泛的应用[7]。图像的均值表示图像的整体亮度，方差可以用来计算图像对比度，偏度反映了图像像素值的对称性，峰度用来测量图像的正态分布。

设 $\boldsymbol{I}_{\text{gray}}$ 表示图像 $\boldsymbol{I}$ 的灰度版本。图像 $\boldsymbol{I}$ 的均值 $\text{Me}(\boldsymbol{I})$、方差 $\text{Va}(\boldsymbol{I})$、偏度 $\text{Sk}(\boldsymbol{I})$ 和峰度 $\text{Ku}(\boldsymbol{I})$ 可表示为

$$\text{Me}(\boldsymbol{I}) = M(\boldsymbol{I}_{\text{gray}}) \tag{1}$$

$$\text{Va}(\boldsymbol{I}) = \sqrt{M(\boldsymbol{I}_{\text{gray}} - M(\boldsymbol{I}_{\text{gray}})^2)} \tag{2}$$

$$\text{Sk}(\boldsymbol{I}) = \sqrt[3]{M(\boldsymbol{I}_{\text{gray}} - M(\boldsymbol{I}_{\text{gray}})^3)} \tag{3}$$

$$\text{Ku}(\boldsymbol{I}) = \sqrt[4]{M(\boldsymbol{I}_{\text{gray}} - M(\boldsymbol{I}_{\text{gray}})^4)} \tag{4}$$

式中：$M(\cdot)$ 表示均值算子。

图 3 描述了图 1(a) ~ (c) 所示图像的矩特征。可见，均值和方差随着对比度失真程度的增大而减小，相反，偏度和峰度随之增大。这表明矩特征能够有效地表示不同对比度失真的图。

### 1.1.2 KL 散度特征

对于一幅图像来说，直方图集中在左边时，图像较暗，集中在右边时，则较亮，当直方图分布较均匀时，图像的对比度也会比较明显。图 4 绘制了均匀分布及图 1(a) ~ (c) 所示图像直方图的概率分布。可以看出，随着对比度失真程度的增加，概率密度曲线越来越偏离均匀分布。

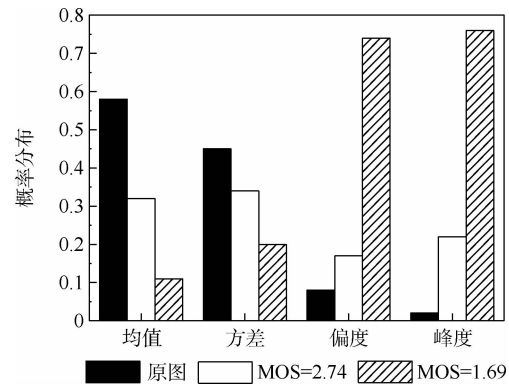



图 3 图 1(a) ~ (c) 所示图像的矩特征

Fig. 3 Moment features of images shown in Fig. 1(a) – Fig. 1(c)

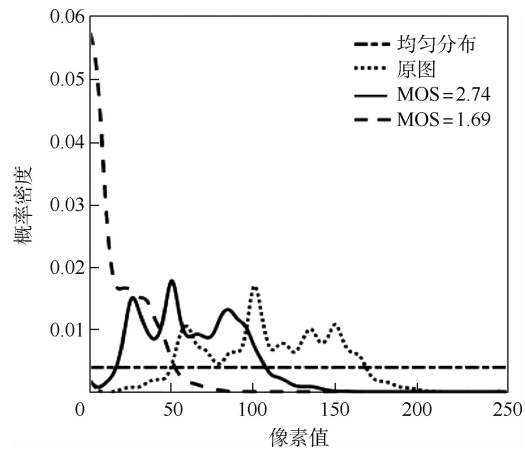



图 4 均匀分布、原图及不同程度对比度失真图像的概率密度

Fig. 4 Probability density of uniform distribution, original image and different degree of contrast distortion image

因此，图像直方图与均匀分布的偏离度为对比度失真的图像质量提供了一个很好的度量。假设 $\boldsymbol{h}_1$ 和 $\boldsymbol{h}_2$ 分别为对比度失真的图像直方图和均匀分布直方图，用 KL 散度来测量 $\boldsymbol{h}_1$ 和 $\boldsymbol{h}_2$ 之间的距离，KL 散度通常被用来评估 2 个分布之间的距离，定义为

$$D_{\text{KL}}(\boldsymbol{h}_1 \| \boldsymbol{h}_2) = -\int \boldsymbol{h}_1(t) \ln \boldsymbol{h}_2(t) \mathrm{d}t + \int \boldsymbol{h}_1(t) \ln \boldsymbol{h}_1(t) \mathrm{d}t \tag{5}$$

然而 KL 散度是不对称的，可能会引起一些不稳定性。因此，本文同时使用 KL 散度和逆 KL 散度作为特征。逆 KL 散度描述为 $D_{\text{KL}}(\boldsymbol{h}_2 \| \boldsymbol{h}_1)$。在 2.2 节中，用实验验证了使用 KL 散度和逆 KL 散度的结合使方法性能大大提升。

## 1.2 基于色调与饱和度的特征提取

为了计算测试图像在色调和饱和度方面的失真，引入与人类视觉系统特点一致的 HSV 颜色空

间。HSV 模型的3 个组成部分分别为:①色调,用角度度量,取值范围为 0° ~360°;②饱和度,取值范围为 0% ~100% ;③明度,即亮度,取值范围为 0% ~100% 。HSV 颜色空间是 RGB 颜色空间的一种非线性变换,转化过程为

$$H = \begin{cases} 0^\circ & \max = \min \\ 60^\circ\left(\dfrac{G-B}{\max-\min}+0\right) & \max = R \\ 60^\circ\left(\dfrac{B-R}{\max-\min}+2\right) & \max = G \\ 60^\circ\left(\dfrac{R-G}{\max-\min}+4\right) & \max = B \end{cases} \tag{6}$$

$$S = \begin{cases} \dfrac{\max-\min}{\max} & \max \neq 0 \\ 0 & \text{其他} \end{cases} \tag{7}$$

$$V = \max \tag{8}$$

式中:$H$、$S$、$V$ 分别为色调值、饱和度值和亮度值;$R$、$G$、$B$ 分别为图像每个像素 RGB 3 通道的值;min 和 max 分别为$(R,G,B)$中的最小值和最大值。

因为 1.1 节已经提取出合适的亮度特征,所以本节从 HSV 颜色空间的 H 和 S 通道中计算图像的颜色失真。LBP 算子常用来描述局部区域中心和相邻像素之间的关系,因此本节计算 H 和 S 通道图中每个像素的 LBP 值,计算公式为

$$\mathrm{LBP}_{n,r} = \begin{cases} \sum_{t=0}^{n-1} F(U_t - U_c) & \mathrm{Ti}(\mathrm{LBP}_{n,r}) \leqslant 2 \\ n+1 & \text{其他} \end{cases} \tag{9}$$

式中:$n$ 为像素邻域,设置为 8;$r$ 为邻域的半径,设置为 1;$U_t$ 和 $U_c$ 分别为 H 或 S 通道图像中的邻域像素和中心像素;函数 $F(\cdot)$的定义为

$$F(U_t - U_c) = \begin{cases} 1 & U_t - U_c \geqslant 0 \\ 0 & \text{其他} \end{cases} \tag{10}$$

Ti 为用于计算按位转换次数的度量,其定义为

$$\mathrm{Ti}(\mathrm{LBP}_{n,r}) = \left\| F(U_{n-1} - U_c) - F(U_0 - U_c) \right\| + \sum_{t=0}^{n-1} \left\| F(U_t - U_c) - F(U_{t-1} - U_c) \right\| \tag{11}$$

研究发现,当均匀 LBP 算子的位转换次数不超过 2 时,可以增强 LBP 算子的区分能力[12]。因此,旋转不变的均匀 LBP 算子具有 $n+2$ 个模式。基于色调和饱和度的 LBP 映射可描述为

$$\mathrm{CW}^U = \mathrm{LBP}_{n,r}(U) \tag{12}$$

因此,可以得到 2 个 LBP 图,即 $\mathrm{CW}^H$ 和 $\mathrm{CW}^S$。

然而, LBP 算子只编码相邻像素之间的差异,不能捕获准确的幅度信息。为了解决这一问题,通过累积 $\mathrm{CW}^U$ 中具有相同 LBP 模式的像素来计算加权 LBP 直方图,则颜色加权 LBP 直方图定义为

$$L^U(k) = \sum_{Y=1}^{M}\sum_{X=1}^{N} U(X,Y) f(\mathrm{CW}^U(X,Y),k) \tag{13}$$

$$f(x,y) = \begin{cases} 1 & x = y \\ 0 & \text{其他} \end{cases} \tag{14}$$

式中:$M$ 和 $N$ 分别为输入图像的行数和列数;$k$ 为可能的 LBP 模式,$k \in [0,9]$;$U(X,Y)$ 为分配给 LBP 值的权重。本文使用 H 和 S 通道图作为每个像素的 LBP 权重。对于测试图像,可以产生 20 个色调和饱和度特征。

为了验证提取特征的有效性,图 5(a)、(b)分别展示了图 1(a) ~(c)中在 H 和 S 通道的颜色加权 LBP 直方图。图中,横坐标表示对应 LBP 图中的不同模式,纵坐标表示具有相同模式的所有像素的总和。通过观察图 5(a)、(b)可以看出,H 和 S 通道中每个 LBP 模式的累积值均随着对比度的变化而变化,即不同对比度的图像具有不同的统计特征,这表明本节所提出的组合特征可以有效地表示对比度的变化。

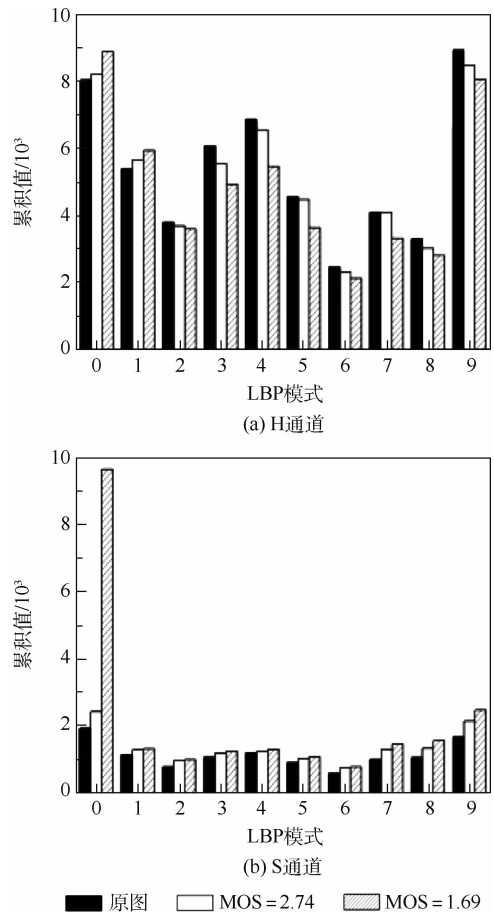



图 5　H 和 S 通道下 LBP 各模式的累积值

Fig. 5　Accumulated amplitude of each mode of LBP under H and S channels

### 1.3 AdaBoosting BP 神经网络评价模型

提取到所有特征之后,本文通过 AdaBoosting BP 神经网络建立了一个从图像特征到图像质量的映射。AdaBoosting 算法是学习弱学习算法的强回归算法[12],本文利用 BP 神经网络作为弱学习算法。AdaBoosting BP 神经网络结构的原理如图 6所示。图上方的虚线框描述了 BP 神经网络的结构,在 BP 神经网络中,输入层的神经元个数与图像特征的维数相同。输出层只有一个神经元节点,输出结果即被预测的图像质量结果。2 个隐藏层的节点数和输入层的节点数目相同。第 1 层隐藏层使用 tangent sigmoid 函数作为激活函数。第 2 层隐藏层使用径向基函数(radial basis function, RBF)作为激活函数。

AdaBoosting BP 神经网络算法步骤如下:

1)确定 BP 神经网络的个数 $T$,将 $T$ 设置为 10[12],并将训练集 $X_{\mathrm{tr}}$的主观分数 $Y_{\mathrm{tr}}$映射到[0,1]。

2)对于第 $i$ 个 BP 神经网络,使用 $X_{\mathrm{tr}}$和 $Y_{\mathrm{tr}}$进行训练,并计算 $X_{\mathrm{tr}}$和测试集 $X_{\mathrm{te}}$的预测结果,分别为 $\hat{Y}^i_{\mathrm{tr}}$和 $\hat{Y}^i_{\mathrm{te}}$。

3)使用第 $i$ 个 BP 神经网络的训练集分布 $D_i$ 代表每个训练集对计算训练误差的权重,计算公式为

$$D_{i,j}=\begin{cases}\dfrac{1}{Z} & i=1\\ D_{i-1,j}(1+\sigma\cdot\partial(Y^j_{\mathrm{tr}}-\hat{Y}^{i-1,j}_{\mathrm{tr}})) & i=2,3,\cdots,T\end{cases} \tag{15}$$

式中:$D_{i,j}$为 $D_i$ 的第 $j$ 个元素;$\sigma$ 设置为0.1[12];$Z$ 为训练集中图像的数量;函数$\partial$的定义为

$$\partial(x)=\begin{cases}1 & x>\mathrm{threshold}\\ 0 & \text{其他}\end{cases} \tag{16}$$

4)每个 BP 神经网络的权重定义为

$$\omega_i=\frac{1}{\mathrm{e}^{-b(|\mathrm{Error}_i|-c)}} \tag{17}$$

式中:由于本文计算的是预测器权重,设置 $b=-1$,$c=-\ln 2$[12];$\mathrm{Error}_i$ 为第 $i$ 个 BP 神经网络的误差,计算公式为

$$\mathrm{Error}_i=\sum_{j=1}^{M}D_{i,j}\cdot\partial(Y^j_{\mathrm{tr}}-\hat{Y}^{i,j}_{\mathrm{tr}}) \tag{18}$$

5)将 $i$ 个 BP 神经网络预测结果线性组合,可得最终预测结果为

$$\hat{Y}=\sum_{i=1}^{T}\omega_i\hat{Y}^i_{\mathrm{te}} \tag{19}$$

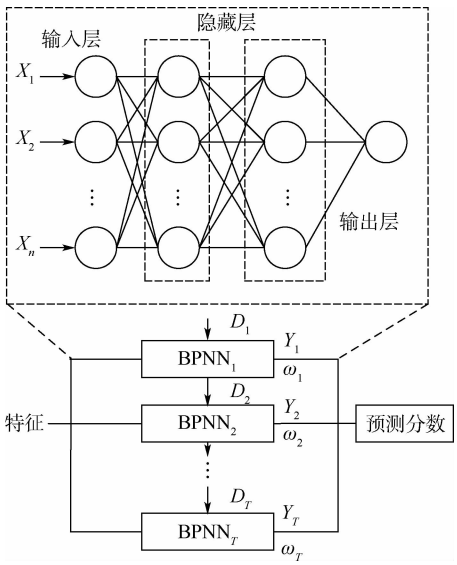



图 6 AdaBoosting BP 神经网络模型

Fig. 6 AdaBoosting BP neural network model

## 2 实验结果与分析

### 2.1 实验设置

为了验证本文方法的性能,引入了 5 个用于图像对比度失真评价的公共图像数据库,分别为 CID2013[15]、CCID2014[4]、CSIQ[16]、TID2008[17]和 TID2013[18]数据库。其中,CSIQ 数据库的主观评分由差分平均意见得分(differential mean opinion core, DMOS)表示,DMOS 值越小代表图像质量越好,而其他数据库用 MOS 表示,MOS 值越大表示图像质量越好。表 1 描述了这 5 个图像数据库的主要特征。

在实验中,本文采用 Pearson 秩相关系数(PLCC)、Spearman 秩相关系数(SRCC)和 Kendalls 秩相关系数(KRCC)3 个准则对方法的性能进行

表 1 五个图像质量数据库的特征

Table 1 Features of five image quality databases

| 数据库 | 原始图像数量 | 对比度失真图像数量 | 失真类型 | 图像尺寸 | 分数类型 | 主观分数范围 |
|---|---|---|---|---|---|---|
| CID2013[15] | 15 | 400 | 1 | 768×512 | MOS | [1~5] |
| CCID2014[4] | 15 | 655 | 1 | 768×512 | MOS | [1~5] |
| CSIQ[16] | 30 | 116 | 6 | 512×512 | DMOS | [0~1] |
| TID2008[17] | 25 | 200 | 17 | 512×384 | MOS | [0~9] |
| TID2013[18] | 25 | 250 | 24 | 512×384 | MOS | [0~9] |

了测试。其中,预测精度采用 PLCC,而预测单调性使用 SRCC 和 KRCC。PLCC、SRCC 和 KRCC 的值越接近 1,表示方法性能越好。在计算 PLCC 之前,需要在主观和客观得分之间提供一个非线性的逻辑映射函数[10]:

$$A(x) = \alpha_1\left[\frac{1}{2} - \frac{1}{1 + e^{\alpha_2(x-\alpha_3)}}\right] + \alpha_4 x + \alpha_5 \quad (20)$$

式中:$x$ 为原始 IQA 得分;$A(x)$ 为映射的 IQA 得分;$\alpha_1$、$\alpha_2$、$\alpha_3$、$\alpha_4$、$\alpha_5$ 为回归参数。

## 2.2　定量分析

为了验证亮度方面本文所使用的 KL 散度特征和逆 KL 散度特征结合后对性能的影响,在本文步骤中将散度特征分别用常用的 KL 散度、JS 散度替换并比较 3 种特征的性能。在 3 个数据库中的结果如表 2 所示。可以看出,使用 KL 散度或 JS 散度时性能相差不大,但同时使用 KL 散度和逆 KL 散度时,性能提升较大,因此本文选用 KL 散度和逆 KL 散度结合的方法。

另外,为了验证 AdaBoosting BP 神经网络的有效性,本节将回归模型替换成常用的支持向量回归(support vector regression, SVR)[19]和随机森林(random forest, RF)[20]进行了比较。实验结果如表 3 所示。可以看出,AdaBoosting BP 神经网络作为回归模型的方法与使用 SVR 或 RF 的方法相比,在 3 个数据库中性能明显提升。因此,本文使用 AdaBoosting BP 神经网络训练质量模型。

**表 2　使用不同散度的性能比较**

**Table 2　Performance comparison using different divergence**

| 散度类型 | CID2013[15] | | | CSIQ[16] | | | TID2008[17] | | |
|---|---|---|---|---|---|---|---|---|---|
| | PLCC | SRCC | KRCC | PLCC | SRCC | KRCC | PLCC | SRCC | KRCC |
| KL 散度 | 0.930 | 0.926 | 0.768 | 0.949 | 0.915 | 0.767 | 0.897 | 0.860 | 0.686 |
| JS 散度 | 0.935 | 0.931 | 0.777 | 0.947 | 0.931 | 0.777 | 0.896 | 0.862 | 0.683 |
| 正逆 KL 散度 | **0.969** | **0.966** | **0.848** | **0.966** | **0.945** | **0.818** | **0.927** | **0.911** | **0.749** |

注:黑体数据表示最好结果。

**表 3　使用不同回归模型的性能比较**

**Table 3　Performance comparison using different regression models**

| 回归模型 | CID2013[15] | | | CSIQ[16] | | | TID2008[17] | | |
|---|---|---|---|---|---|---|---|---|---|
| | PLCC | SRCC | KRCC | PLCC | SRCC | KRCC | PLCC | SRCC | KRCC |
| SVR | 0.906 | 0.895 | 0.718 | 0.905 | 0.851 | 0.708 | 0.912 | 0.867 | 0.693 |
| RF | 0.921 | 0.905 | 0.739 | 0.920 | 0.871 | 0.712 | 0.922 | 0.881 | 0.703 |
| AdaBoosting BP 神经网络 | **0.969** | **0.966** | **0.848** | **0.966** | **0.945** | **0.818** | **0.927** | **0.911** | **0.749** |

注:黑体数据表示最好结果。

## 2.3　不同数据库方法性能比较

为了验证本文方法的性能,表 4 展示了在 CID2013、CCID2014、TID2013、TID2008 和 CSIQ 数据库上本文方法的性能评价结果,以及与 13 种针对对比度失真图像评价方法的比较结果。为了更好的比较,本文将 IQA 方法分为 3 组:FR、RR 和 NR 度量。FR-IQA 指标有 QMC[1]、PCQI[2]、QCCI[3];RR-IQA 指标涉及 RIQMC[4]、RCIQM[5]、CIQM[6];NR-IQA 方法由 CDIQA[7]、ICDIQA[8]、NIQMC[9]、BIQME[10]、HEFCS[11]、文献[12]和文献[13]组成。此外,表中还列出了性能标准的数据库数量加权平均值。对于每个数据库,随机选择 80% 的图像进行模型训练,20% 的图像用于性能测试。为避免数据精度误差带来的影响,训练-测试过程重复 1 000 次,并在表 4 中显示中值。

从表 4 中可以看出,本文方法明显优于其他 NR-IQA 方法,甚至比 FR 和 RR 方法更具竞争力。在 NR 方法中,本文方法结果明显优于 5 个数据库中的所有 NR 模型。仅仅在 CSIQ 数据库中略低于 RIQMC 和 RCIQM 方法,但是 RIQMC 和 RCIQM 均属于半参考质量度量方法,需要原始图像的部分信息。从表 4 中的平均值结果来看,本文方法的结果明显优于其他方法。

## 2.4　泛化性能实验

为了证明本文方法的泛化能力,本节进行了跨数据库验证实验。通过对每个数据库的所有对比度失真图像进行训练来获得训练模型,使用这些训练模型测试其他 4 个数据库的失真图像。性能结果如表 5 所示,其中行代表训练集,列代表测试集。对于相同的训练和测试集,表中表示为 0。从表 5 中可以看出,用 CSIQ 数据库测试而在其他数据库上训练时,取得了最好的性能结果。在 CID2013 和 CCID2014 数据库上测试而在其他数

表4 五个对比度失真图像数据库上本文方法和其他方法的性能比较

Table 4 Performance comparison of the proposed method and other methods on five contrast-distorted image databases

| 数据库 | 秩相关系数 | FR | | | RR | | |
|---|---|---|---|---|---|---|---|
| | | QMC[1] | PCQI[2] | QCCI[3] | RIQMC[4] | RCIQM[5] | CIQM[6] |
| CID2013[15] | PLCC | 0.806 | 0.924 6 | 0.934 5 | 0.899 5 | 0.918 7 | 0.913 9 |
| | SRCC | 0.767 4 | 0.923 2 | 0.929 3 | 0.900 5 | 0.920 3 | 0.920 6 |
| | KRCC | 0.578 5 | 0.758 0 | 0.762 0 | 0.716 2 | 0.754 3 | 0.724 0 |
| CCID2014[4] | PLCC | 0.895 2 | 0.872 1 | 0.888 0 | 0.872 6 | 0.884 5 | 0.885 3 |
| | SRCC | 0.870 5 | 0.886 9 | 0.895 7 | 0.846 5 | 0.856 5 | 0.869 7 |
| | KRCC | 0.684 6 | 0.682 0 | 0.702 1 | 0.650 7 | 0.669 5 | 0.685 4 |
| CSIQ[16] | PLCC | 0.960 5 | 0.948 2 | 0.946 6 | 0.965 2 | 0.964 5 | 0.946 2 |
| | SRCC | 0.953 2 | 0.948 8 | 0.951 2 | **0.957 9** | 0.956 9 | 0.949 6 |
| | KRCC | 0.816 5 | 0.814 4 | 0.798 2 | **0.827 9** | 0.819 8 | 0.810 5 |
| TID2008[17] | PLCC | 0.803 6 | 0.882 1 | 0.881 4 | 0.858 5 | 0.880 7 | 0.892 2 |
| | SRCC | 0.752 9 | 0.900 2 | 0.898 9 | 0.809 5 | 0.857 8 | 0.868 1 |
| | KRCC | 0.571 9 | 0.722 6 | 0.711 9 | 0.622 4 | 0.670 5 | 0.689 0 |
| TID2013[18] | PLCC | 0.797 2 | 0.873 8 | 0.873 3 | 0.865 1 | 0.886 6 | 0.897 0 |
| | SRCC | 0.733 6 | 0.917 5 | 0.912 6 | 0.804 4 | 0.854 1 | 0.862 1 |
| | KRCC | 0.551 3 | 0.709 3 | 0.685 4 | 0.617 8 | 0.667 5 | 0.687 3 |
| 加权平均值 | PLCC | 0.851 4 | 0.892 0 | 0.900 6 | 0.883 0 | 0.898 5 | 0.899 4 |
| | SRCC | 0.815 3 | 0.906 6 | 0.911 0 | 0.856 7 | 0.879 2 | 0.886 6 |
| | KRCC | 0.633 4 | 0.719 4 | 0.722 4 | 0.671 0 | 0.701 0 | 0.704 6 |

| 数据库 | 秩相关系数 | NR | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | CDIQA[7] | ICDIQA[8] | NIQMC[9] | BIQME[10] | HEFCS[11] | 文献[12] | 文献[13] | 本文方法 |
| CID2013[15] | PLCC | 0.866 8 | 0.912 9 | 0.869 1 | 0.900 4 | 0.897 3 | 0.943 5 | 0.964 6 | **0.968 9** |
| | SRCC | 0.850 0 | 0.908 1 | 0.866 8 | 0.902 3 | 0.877 7 | 0.933 8 | 0.960 3 | **0.966 2** |
| | KRCC | 0.658 8 | 0.703 5 | 0.669 0 | 0.722 3 | 0.690 6 | 0.781 4 | 0.835 4 | **0.848 0** |
| CCID2014[4] | PLCC | 0.837 1 | 0.877 9 | 0.843 8 | 0.858 8 | 0.865 0 | 0.923 5 | 0.910 9 | **0.925 4** |
| | SRCC | 0.802 6 | 0.851 2 | 0.811 3 | 0.830 9 | 0.842 6 | 0.911 8 | 0.902 3 | **0.914 5** |
| | KRCC | 0.603 6 | 0.659 8 | 0.605 2 | 0.630 5 | 0.639 5 | 0.723 6 | 0.728 5 | **0.748 3** |
| CSIQ[16] | PLCC | 0.666 3 | 0.881 7 | 0.874 7 | 0.810 6 | 0.941 7 | 0.936 8 | 0.926 9 | **0.965 5** |
| | SRCC | 0.585 6 | 0.814 5 | 0.853 3 | 0.784 8 | 0.903 9 | 0.887 6 | 0.895 3 | 0.944 7 |
| | KRCC | 0.439 0 | 0.690 3 | 0.668 9 | 0.598 3 | 0.752 4 | 0.729 0 | 0.731 2 | 0.818 2 |
| TID2008[17] | PLCC | 0.632 0 | 0.756 8 | 0.776 7 | 0.899 3 | 0.865 0 | 0.865 4 | 0.876 3 | **0.926 5** |
| | SRCC | 0.572 3 | 0.703 6 | 0.732 4 | 0.848 8 | 0.804 2 | 0.800 3 | 0.817 6 | **0.911 3** |
| | KRCC | 0.425 3 | 0.498 9 | 0.541 9 | 0.646 0 | 0.630 2 | 0.609 8 | 0.638 9 | **0.748 7** |
| TID2013[18] | PLCC | 0.579 8 | 0.696 3 | 0.722 5 | 0.852 4 | 0.844 3 | 0.895 7 | 0.911 1 | **0.951 0** |
| | SRCC | 0.508 2 | 0.642 9 | 0.645 8 | 0.814 9 | 0.749 9 | 0.840 1 | 0.853 0 | **0.924 0** |
| | KRCC | 0.362 8 | 0.453 6 | 0.468 7 | 0.610 9 | 0.568 7 | 0.659 5 | 0.688 7 | **0.768 0** |
| 加权平均值 | PLCC | 0.767 2 | 0.843 9 | 0.825 3 | 0.869 6 | 0.875 3 | 0.917 9 | 0.921 1 | **0.943 1** |
| | SRCC | 0.724 9 | 0.812 3 | 0.792 7 | 0.845 0 | 0.836 6 | 0.890 7 | 0.898 1 | **0.930 5** |
| | KRCC | 0.546 3 | 0.621 1 | 0.596 6 | 0.649 7 | 0.648 1 | 0.714 3 | 0.737 9 | **0.781 0** |

注:黑体数据表示最好结果。

据库上训练时,性能较差。在CSIQ和TID2008及TID2013互相作为训练集和测试集时,性能结果都较好。这是因为CSIQ数据库包含的对比度失真方面最少,其次是TID2008和TID2013,而CID2013和CCID2014数据库中涉及的对比度失真方面最多。另外,本文方法的特征共有27个,属于低维特征,这就可能导致在追求计算效率的同时鲁棒性较差。这也表明未来的研究应该考虑更多的属性。

## 2.5 一致性实验

为了分析在质量回归模型中训练集和测试集的比例对性能的影响,对本文模型的训练集和测试集在3个比例下进行实验,分别为80%、50%和20%的失真图像用于训练,剩余的20%、50%和80%用于测试。每个数据库在3个分割比例下的SRCC、PLCC和KRCC值如表6所示。通过表6中的结果得出,本文方法的预测性能随着训练集在数据库上比例的减少而下降。然而,在有大量失真图像的数据库上,如CID2013和CCID2014,分割比例对性能的影响相对较小。这表明训练数据不足会限制本文方法的泛化能力。

表 5　跨数据库验证的性能

Table 5　Performance of cross-database verification

| 数据库 | 秩相关系数 | CID2013[15] | CCID2014[4] | CSIQ[16] | TID2008[17] | TID2013[18] |
|---|---|---|---|---|---|---|
| CID2013[15] | PLCC | | 0.922 | 0.663 | 0.580 | 0.540 |
| | SRCC | 0 | 0.902 | 0.653 | 0.556 | 0.503 |
| | KRCC | | 0.733 | 0.485 | 0.398 | 0.361 |
| CCID2014[4] | PLCC[16] | 0.966 | | 0.649 | 0.541 | 0.503 |
| | SRCC | 0.965 | 0 | 0.626 | 0.500 | 0.458 |
| | KRCC | 0.839 | | 0.434 | 0.350 | 0.317 |
| CSIQ[16] | PLCC | 0.588 | 0.540 | | 0.771 | 0.769 |
| | SRCC | 0.586 | 0.443 | 0 | 0.701 | 0.670 |
| | KRCC | 0.401 | 0.302 | | 0.520 | 0.490 |
| TID2008[17] | PLCC | 0.475 | 0.468 | 0.858 | | 0.955 |
| | SRCC | 0.387 | 0.364 | 0.853 | 0 | 0.932 |
| | KRCC | 0.277 | 0.260 | 0.641 | | 0.787 |
| TID2013[18] | PLCC | 0.503 | 0.502 | 0.806 | 0.956 | |
| | SRCC | 0.364 | 0.323 | 0.779 | 0.948 | 0 |
| | KRCC | 0.261 | 0.248 | 0.553 | 0.809 | |

表 6　三个数据库上不同训练集和测试集比例性能比较

Table 6　Performance comparison of different training set and test set ratios on three databases

| 测试集 | CID2013[15] | | | CSIQ[16] | | | TID2008[17] | | |
|---|---|---|---|---|---|---|---|---|---|
| 比例/% | PLCC | SRCC | KRCC | PLCC | SRCC | KRCC | PLCC | SRCC | KRCC |
| 20 | 0.903 | 0.904 | 0.727 | 0.766 | 0.755 | 0.553 | 0.828 | 0.789 | 0.601 |
| 50 | 0.951 | 0.952 | 0.808 | 0.905 | 0.893 | 0.712 | 0.897 | 0.872 | 0.694 |
| 80 | **0.969** | **0.966** | **0.848** | **0.966** | **0.945** | **0.818** | **0.927** | **0.911** | **0.749** |

注:黑体数据表示最好结果。

## 2.6　消融实验

本文采用了 3 种特征,分别从亮度、色调和饱和度方面提取统计特征来预测图像质量。为了验证各部分特征对性能的影响,本节在 CID2013 数据库上进行了消融实验。总体结果如图 7 所示。图中,V1 表示从灰度图中提取的矩特征,V2 表示正逆 KL 散度特征,H 和 S 分别表示从 HSV 颜色空间提取的色调和饱和度特征。随着各部分特征的增加,在 CID2013 数据库上,PLCC、SRCC 和 KRCC 的值不断提高,这表示性能也越来越高。

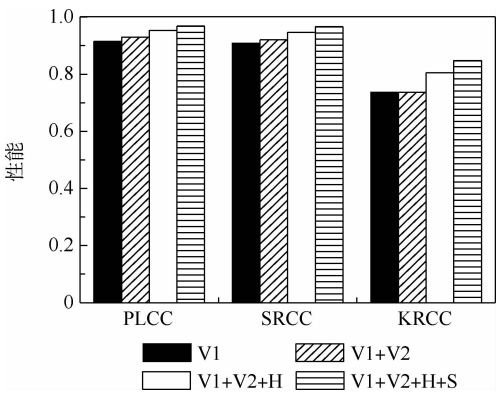



图 7　CID2013 数据库上的消融结果

Fig. 7　Ablation results on CID2013 database

## 2.7　通用性评估实验

为了检验本文方法对于其他图像失真类型的效果,本节在 CSIQ 数据库中对于各种失真类型图像进行了评价实验。CSIQ 数据库中包含 JPEG 压缩失真、JPEG2K 压缩失真、高斯模糊失真(Gaussian blur, GB)、高斯白噪声失真(white Gaussian noise, WGN)、高斯粉红噪声失真(pink Gaussian noise, PGN)和全局对比度失真(global contrast decrements, GCD)。本文方法针对每种失真类型图像评估的 SRCC、PLCC 和 KRCC 值如表 7所示。从结果中可以看出,本文方法针对每种失真类型图像的评价都具有一定效果,但是在针对对比度失真图像上具有最佳性能。究其原因,不同的失真类型图像具有不同的特性,而本文从彩色三要素方面提取的各种特征是针对对比度失真图像特性的。

表 7　在 CSIQ 数据库上不同失真类型图像的性能比较

Table 7　Performance comparison of different distortion types in CSIQ database

| 秩相关系数 | JPEG | JPEG2K | GB | WGN | PGN | GCD |
|---|---|---|---|---|---|---|
| PLCC | 0.913 | 0.891 | 0.908 | 0.942 | 0.931 | **0.966** |
| SRCC | 0.862 | 0.852 | 0.867 | 0.934 | 0.907 | **0.945** |
| KRCC | 0.688 | 0.684 | 0.702 | 0.801 | 0.773 | **0.818** |

注:黑体数据表示最好结果。

## 3 结 论

1）本文提出了基于彩色三要素的无参考对比度失真图像质量评价方法。从彩色三要素的亮度、色调和饱和度出发，提取 3 方面的特征参数。在亮度方面，在灰度上提取了矩特征及对比度失真图像直方图与均匀分布之间的散度特征。在饱和度和色调方面，分别在 HSV 空间的 H 和 S 通道中计算颜色加权 LBP 直方图特征。基于这些特征，采用 AdaBoosting BP 神经网络对质量模型进行训练。

2）在 5 个公开数据库上的实验结果可以看出，本文方法对对比度失真图像的评价性能优于其他相关的 IQA 方法。

在未来的工作中，可以考虑更多属性的特征，并探索更好的神经网络用于质量模型训练。

### 参考文献（References）


[1] GU K, ZHAI G T, YANG X K, et al. Automatic contrast enhancement technology with saliency preservation [J]. IEEE Transactions on Circuits and Systems for Video Technology, 2015, 25(9): 1480-1494.

[2] WANG S Q, MA K D, YEGANEH H, et al. A patch-structure representation method for quality assessment of contrast changed images[J]. IEEE Signal Processing Letters, 2015, 22(12): 2387-2390.

[3] SUN W, YANG W M, ZHOU F, et al. Full-reference quality assessment of contrast changed images based on local linear model [C] // 2018 IEEE International Conference on Acoustics, Speech and Signal Processing. Piscataway: IEEE Press, 2018: 1228-1232.

[4] GU K, ZHAI G T, LIN W S, et al. The analysis of image contrast: From quality assessment to automatic enhancement [J]. IEEE Transactions on Cybernetics, 2016, 46(1): 284-297.

[5] LIU M, GU K, ZHAI G T, et al. Perceptual reduced-reference visual quality assessment for contrast alteration [J]. IEEE Transactions on Broadcasting, 2017, 63(1): 71-81.

[6] KIM D, LEE S, KIM C. Contextual information based quality assessment for contrast-changed images[J]. IEEE Signal Processing Letters, 2019, 26(1): 109-113.

[7] FANG Y M, MA K D, WANG Z, et al. No-reference quality assessment of contrast-distorted images based on natural scene statistics[J]. IEEE Signal Processing Letters, 2015, 22(7): 838-842.

[8] WU Y J, ZHU Y H, YANG Y, et al. A no-reference quality assessment for contrast-distorted image based on improved learning method[J]. Multimedia Tools and Applications, 2019, 78(8): 10057-10076.

[9] GU K, LIN W S, ZHAI G T, et al. No-reference quality metric of contrast-distorted images based on information maximization [J]. IEEE Transactions on Cybernetics, 2017, 47(12): 4559-4565.

[10] GU K, TAO D C, QIAO J F, et al. Learning a no-reference quality assessment model of enhanced images with big data [J]. IEEE Transactions on Neural Networks and Learning Systems, 2018, 29(4): 1301-1313.

[11] KHOSRAVI M H, HASSANPOUR H. Blind quality metric for contrast-distorted images based on eigen decomposition of color histograms[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2020, 30(1): 48-58.

[12] ZHOU Y, LI L D, ZHU H C, et al. No-reference quality assessment for contrast-distorted images based on multifaceted statistical representation of structure[J]. Journal of Visual Communication and Image Representation, 2019, 60: 158-169.

[13] LYU W J, LU W, MA M. No-reference quality metric for contrast-distorted image based on gradient domain and HSV space [J]. Journal of Visual Communication and Image Representation, 2020, 69: 102797.

[14] GONZALEZ R C, WOODS R E. Digital image processing[M]. 3rd ed. Beijing: Publishing House of Electronics Industry, 2011.

[15] GU K, ZHAI G T, YANG X K, et al. Subjective and objective quality assessment for images with contrast change[C] // 2013 IEEE International Conference on Image Processing. Piscataway: IEEE Press, 2013: 383-387.

[16] LARSON E C, CHANDLER D M. Most apparent distortion: Full-reference image quality assessment and the role of strategy [J]. Journal of Electronic Imaging, 2010, 19(1): 011006.

[17] PONOMARENKO N, LUKIN V, ZELENSKY A, et al. TID2008-A database for evaluation of full-reference visual quality assessment metrics[J]. Advances of Modern Radio electronics, 2009, 10: 30-45.

[18] PONOMARENKO N, IEREMEIEV O, LUKIN V, et al. Color image database TID2013: Peculiarities and preliminary results [C] // European Workshop on Visual Information Processing (EUVIP). Piscataway: IEEE Press, 2013: 106-111.

[19] WILLIAMS C K I. Learning with kernels: Support vector machines, regularization, optimization, and beyond [J]. Journal of the American Statistical Association, 2003, 98(462): 489-490.

[20] ANTONIO C, ENDER K, JAMIE S. Decision forests: A unified framework for classification, regression, density estimation, manifold learning and semi-supervised learning [J]. Foundations and Trends in Computer Graphics and Vision, 2011, 7(2-3): 81-227.

# No reference quality assessment method for contrast-distorted images based on three elements of color

DING Yingqiu[1], YANG Yang[1,2,*], CHENG Ming[1], ZHANG Weiming[3]

(1. School of Electronic and Information Engineering, Anhui University, Hefei 230601, China;
2. Institute of Artificial Intelligence, Hefei Comprehensive National Science Center, Hefei 230088, China;
3. School of Cyber Science and Technology, University of Science and Technology of China, Hefei 230027, China)

**Abstract**: Image quality assessment is a basic and challenging problem in the field of image processing, among which the contrast distortion has a greater impact on the perception of image quality. However, there is relatively little research on the no-reference image quality assessment of contrast-distorted images. This paper proposes a no-reference contrast-distorted image quality assessment method based on the three elements of color. The three parameters of brightness, hue and saturation of the three elements of color are used to realize the assessment of contrast-distorted images. First, in terms of brightness, the moment feature and the Kullback-Leibler divergence between the image histogram and the uniform distribution are extracted. Secondly, in terms of hue and saturation, the color-weighted local binary patterns (LBP) histogram features are extracted from the H and S channels of the HSV space, respectively. Finally, the AdaBoosting BP neural network is used to train the prediction model. Through extensive experimental analysis and cross-validation in five standard image databases, the experimental results show that the performance of this method is significantly improved compared with the existing contrast-distorted image quality assessment methods.

**Keywords**: image quality assessment; contrast distortion; HSV color space; no reference; three elements of color; BP neural network

**Received**: 2021-09-02; **Accepted**: 2021-09-17; **Published online**: 2021-11-02 16:04

**URL**: kns.cnki.net/kcms/detail/11.2625.V.20211101.1707.014.html

**Foundation items**: Natural Science Foundation of the Anhui Higher Education Institutions of China (KJ2021A0016); National Natural Science Foundation of China (61502007,61871411)

* **Corresponding author**. E-mail: sky_yang@ahu.edu.cn

http://bhxb.buaa.edu.cn jbuaa@buaa.edu.cn
**DOI**: 10.13700/j.bh.1001-5965.2021.0523

# 基于图对比注意力网络的知识图谱补全

刘丹阳[1], 方全[2,*], 张晓伟[1], 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(KG)补全旨在通过知识库中已知三元组来预测缺失的链接。由于大多数方法都是独立地处理三元组,而忽略了知识图谱所具有的异质结构和相邻节点中固有的丰富的信息,导致不能充分挖掘三元组的特征。考虑基于端到端的知识图谱补全任务,提出了一种图对比注意力网络(GCAT),通过注意力机制同时捕获局部邻域内实体和关系的特征,并封装实体邻域上下文信息。为了有效封装三元组特征,引入一个子图级别的对比训练对象用于增强生成的实体嵌入的质量。为了验证 GCAT 的有效性,在链接预测任务上评估了所提方法,实验结果表明,在数据集 FB15k-237 中,MRR 比 InteractE 提高 0.005,比 A2N 模型提高 0.042;在数据集 WN18RR 中,MRR 比 InteractE 提高 0.019,比 A2N 模型提高 0.032。实验证明提出的 GCAT 模型能够有效预测知识图谱中缺失的链接。

**关 键 词**: 知识图谱(KG); 注意力机制; 对比学习; 知识图谱补全; 链接预测
**中图分类号**: TP391.1
**文献标志码**: A **文章编号**: 1001-5965(2022)08-1428-08

知识图谱(knowledge graph, KG)中,结构化的知识通常被组织为事实三元组(头实体,关系,尾实体)或者是$(h,r,t)$,近年来,一些知识图谱如 Freebase[1] 和 DBpedia[2],已被广泛应用到人工智能行业和领域中。然而,大多数知识图谱通常是稀疏的,并且具有不完整性,即缺少大量客观存在的事实三元组,这引起了对于知识图谱补全任务的研究。知识图谱补全的一种主流方法是基于知识图谱嵌入的方法。一般来说,该方法基于现有的知识图谱中的三元组,将实体和关系的表示嵌入到低维向量空间中,并通过实体嵌入和关系嵌入来评估每个事实三元组的合理性。这些模型都是利用结构化的实体关系交互来学习知识图谱中低维向量表示。

具有代表性的知识图谱嵌入模型包括 TransE[3]、DistMult[4] 和 ComplEx[5] 等。TransE[3] 模型将查找有效三元组的过程视为实体的翻译操作,通过最小化损失函数以学习实体和关系的表示。DistMult[4] 模型通过将关系矩阵 $\boldsymbol{M}_r$ 分解为对角矩阵,降低了模型的参数数量和复杂度,显著提高了知识图谱中潜在信息的挖掘效果。ComplEx[5] 模型是 DistMult[4] 模型的扩展,主要对非对称关系建模,即将实体和关系嵌入复杂空间中,该模型能够很好地挖掘潜在的语义关联。但是上述模型都是独立地处理三元组,具有较弱的特征学习能力。

为了增加模型的表达能力,基于卷积神经网络(convolutional neural network, CNN)的模型被提出应用于知识图谱补全任务。例如,Dettmers 等提出的具有多层卷积的 ConvE[6] 模型,使用 2D 卷

**收稿日期**: 2021-09-06; **录用日期**: 2021-09-17; **网络出版时间**: 2021-11-02 10:15
**网络出版地址**: kns.cnki.net/kcms/detail/11.2625.V.20211101.1526.008.html
**基金项目**: 国家自然科学基金 (62072456,62036012); 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**: *刘丹阳, 方全, 张晓伟, 等. 基于图对比注意力网络的知识图谱补全[J]. 北京航空航天大学学报, 2022, 48(8): 1428-1435. LIU D Y, FANG Q, ZHANG X W, et al. Knowledge graph completion based on graph contrastive attention network[J]. Journal of Beijing University of Aeronautics and Astronautics, 2022, 48(8): 1428-1435 (in Chinese).*

积和多层非线性特征建模知识图谱,然而 ConvE 可以捕获的特征交互数量有限。InteractE[7] 模型通过特征排列、"方格"特征重塑和循环卷积方法来捕获实体和关系特征之间的最大异构交互,进一步增强模型的表达能力,提高了 ConvE 模型的性能。更多基于 CNN 的模型包括 ConvR[8]、ConvKB[9]等。

图神经网络(graph neural network,GNN)近来被广泛应用于知识图谱嵌入,进一步提高了链接预测的质量。例如,R-GCN[10] 引入了 GCN 以建模知识图谱中的多关系数据。基于 GNN 的模型还包括 A2N[11]、CompGCN[12] 等。这些模型可以同时对节点特征和结构信息进行端到端的学习,通常使用基于 GNN 的模型充当编码器用于捕获知识图谱的图结构信息,基于 CNN 的模型被用作解码器以预测三元组的合理性得分。

基于 GNN 的模型虽然可以有效地提取深层次的特征,然而,绝大多数方法都是独立地处理知识图谱中的三元组,忽略了实体局部邻域的异质结构信息。知识图谱本身所具有的这种异质结构也可能包含丰富的、有价值的信息。同时,GCN 在进行邻域聚合时,所有的邻居都共享相同的权重。为了解决以上问题,本文引入了图注意网络(graph attention network,GAT)[13],通过注意力网络对每个实体的邻域结构进行建模,邻近节点特征的权重完全取决于节点特征的重要性,从而全面捕获实体的局部上下文特征。

对比学习[14] 近来引起了科研人员广泛的研究兴趣,其主要目的是通过训练编码器,使其在遵循数据观察的成对样本之间形成对比,从而捕获正样本和负样本的统计相关性,目前,已被广泛应用于图表示学习方法[15-17]中。

Velickovic 等[18] 提出的图对比学习方法 DGI(deep graph infomax)通过训练编码器使得节点表示和全局图表示间的互信息最大化,并在无监督模式下取得了较强的性能。Peng 等[19] 提出的图形互信息(graphical mutual information,GMI)可以最大限度地提高节点表示与其 1-hop 邻居的原始特征之间的互信息,在归纳节点分类和链接预测任务上获得了最优结果。

到目前为止,很少有研究对比学习在知识图谱补全中的作用。本文考虑了中心实体与其上下文关系子图之间的强相关性,对比了来自实体和子图的编码,有效封装三元组的邻域上下文信息和学习高质量的实体表示,进而提高知识图谱补全的预测性能。

本文提出了一种图对比注意力网络(graph contrastive attention network, GCAT),采用对比学习的方法,充分利用知识图谱中的异质结构信息,同时捕获局部邻域内实体和关系的特征。具体地说,为了处理局部邻域三元组的异质性,设计了 GCAT 来捕获实体特征信息。①子图内邻域聚合。利用注意力网络来更新特定关系子图内的实体特征。②子图间邻域聚合。通过基于注意力机制的池化操作实现不同关系子图间的特征聚合。③子图对比约束。通过对比来自每个中心实体及其特定关系子图的编码增强实体嵌入。该局部对比学习模块使学习到的实体嵌入能够更好地捕获知识图谱中的局部邻域上下文信息。④利用基于 ConvE[6] 的解码器模块学习知识图谱中的最终实体和关系表示。本文中对于关系子图的邻域聚合均是对子图中不同类型的一阶邻居上下文特征进行聚合。

# 1　本文方法

## 1.1　基本符号

知识图谱由主体-属性-客体三元组事实组成。在形式上,其被表示为$(h,r,t)$,其中每个三元组都描述了从头实体 $h$ 到尾实体 $t$ 之间的关系 $r$。在知识图谱补全任务中,目的是给定关系和其中一个实体来预测缺失的实体,如给定$(h,r)$推断 $t$ 或者给定$(r,t)$推断 $h$。

本文定义了初始化的实体特征集和关系特征集,分别表示为

$$e = \{\boldsymbol{e}_1, \boldsymbol{e}_2, \cdots, \boldsymbol{e}_{N_e}\}$$

$$r = \{\boldsymbol{r}_1, \boldsymbol{r}_2, \cdots, \boldsymbol{r}_{N_r}\}$$

式中:$\boldsymbol{e}_i \in \mathbf{R}^{N_e \times d}$,$\boldsymbol{r}_i \in \mathbf{R}^{N_r \times d}$,$N_e$ 和 $N_r$ 分别为实体和关系的数量,$d$ 为实体和关系的初始化嵌入维度。

编码器将实体嵌入和关系嵌入输出到解码器中以预测三元组的合理性。本文选择 ConvE[6] 作为解码器。

## 1.2　整体框架

本节以自上而下的方式介绍提出的模型及对于模型的理论讨论。GCAT 的方法框架如图 1 所示。

图 1 左边为知识图谱按关系类型划分的不同关系子图,$\boldsymbol{e}_i$ 为实体,$\boldsymbol{r}_i$ 为关系类型。直观上,实体与其局部邻居更相关,由局部邻域组成的子图为学习邻域结构上下文信息起着关键作用。本文通过提取相同关系类型的一阶邻居,得到实体的

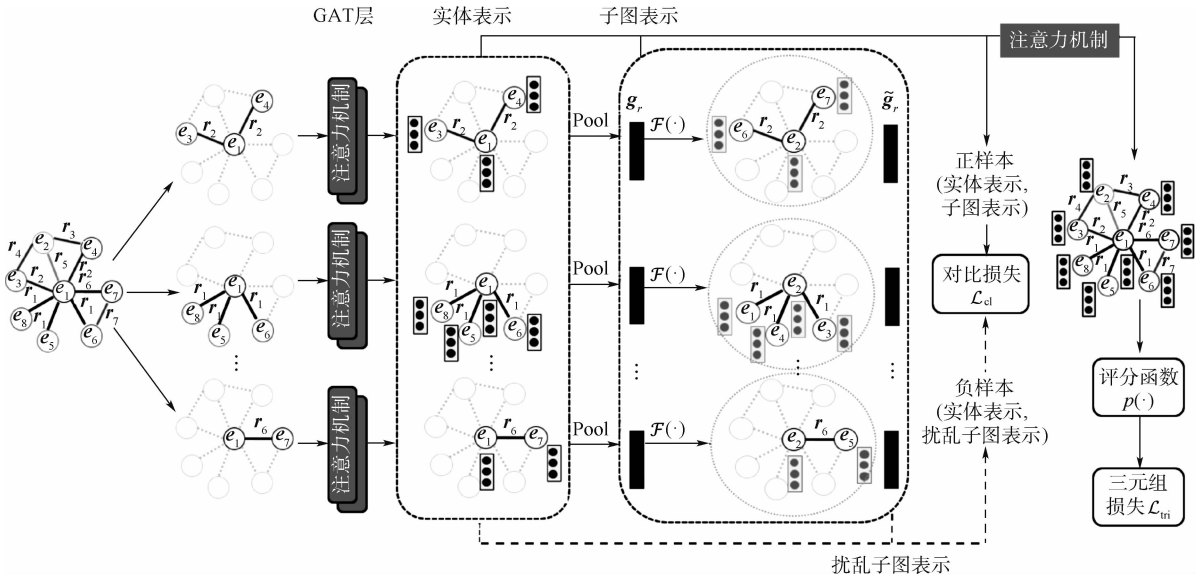



图 1 图对比注意力网络模型架构

Fig. 1 Model architecture of graph contrastive attention network

关系子图,从而在编码器阶段能够封装更多的实体邻域信息及关系信息。

1) 模型通过 2 层 GAT[13] 获得子图内的实体特征。

2) 通过基于注意力机制的池化操作实现不同关系子图间的特征聚合,从而捕获池化后的关系子图的图表示和中心实体邻域的全局信息(Pool 为池化操作)。

3) 子图对比约束。$\mathcal{F}(\cdot)$为混淆函数,用于生成关系子图负样本,通过将真实关系子图表示 $\boldsymbol{g}_r$ 与错误的关系子图表示 $\tilde{\boldsymbol{g}}_r$ 进行对比进一步增强实体表示,$\mathcal{L}_{\mathrm{cl}}$为子图对比损失函数。

图 1 右边为基于 ConvE[6] 模型的评分函数,$p(\cdot)$为三元组评分函数,$\mathcal{L}_{\mathrm{tri}}$为三元组损失函数。

## 1.3 编码器

### 1.3.1 子图内邻域聚合

首先,在关系子图中的每个实体应用线性变换,将中心实体的邻接实体转换为与其相同的低维向量空间,用于邻域聚合。然后,为了充分考虑子图中节点特征,对实体执行自注意力机制,计算其注意力系数,表达式为

$$e_{ij} = f_r(\boldsymbol{W}\boldsymbol{e}_i, \boldsymbol{W}\boldsymbol{e}_j) \tag{1}$$

式中:$e_{ij}$为实体 $j$ 的特征对于实体 $i$ 的重要性;$\boldsymbol{W}$为可学习的线性变换权重矩阵;$f_r(\cdot)$为关系子图中特定于关系 $r \in \mathbf{R}^{N_r \times d}$ 的实体重要性评分函数。

对于中心实体 $i$,计算每条边 $e_{ij}$的注意力系数,形式为

$$I(e_{ij}) = \sigma(f_r(\boldsymbol{W}\boldsymbol{e}_i, \boldsymbol{W}\boldsymbol{e}_j)) \tag{2}$$

式中:注意力系数 $I(\cdot)$为边 $e_{ij}$对中心实体 $i$ 的重要性;$\sigma(\cdot)$为由 LeakyReLU$(\cdot)$实现的非线性激活函数。

本文中,对于所有的实体特征,均采用相同形式的注意力机制,但注意力参数不同。实体注意力评分函数 $f_r(\cdot)$采用 GAT[13] 定义的形式为

$$f_r(e_{ij}) = [\boldsymbol{W}\boldsymbol{e}_i \| \boldsymbol{W}\boldsymbol{e}_j]\boldsymbol{a}_r \tag{3}$$

式中:"$\|$"表示向量拼接操作;$\boldsymbol{a}_r$ 为同关系类型下三元组共享的可训练的注意力参数。

因此,在中心实体 $i$ 的邻域上,实现注意力系数的归一化表示为

$$\alpha_{ij} = \mathrm{softmax}(I(e_{ij})) = \frac{\exp(I(e_{ij}))}{\sum_{k \in N_i} \exp(I(e_{ik}))} \tag{4}$$

式中:$k \in N_i$,$k$ 为实体中心实体 $i$ 的邻居实体,$N_i$ 为中心实体 $i$ 的邻域。

为了稳定注意力机制的学习过程及进一步丰富模型,本文使用了多头注意力机制。将 $K$ 个注意力头的低维向量拼接起来更新第一层模型的实体嵌入向量,如下:

$$\boldsymbol{e}'_i = \Big\|_{k=1}^{K} \sigma\Big(\sum_{j \in N_i} \alpha_{ij}^k \boldsymbol{W}^k \boldsymbol{e}_j\Big) \tag{5}$$

考虑到参数的大小,使用平均而不是拼接 $K$ 独立注意头的嵌入,因此,模型的最后一层中输出的实体嵌入向量如下:

$$\boldsymbol{e}'_i = \sigma\Big(\frac{1}{K}\sum_{k=1}^{K}\sum_{j \in N_i} \alpha_{ij}^k \boldsymbol{W}^k \boldsymbol{e}_j\Big) \tag{6}$$

### 1.3.2 子图间邻域聚合

根据式(1)~式(6),得到了关系子图中实体的特征表示。为了捕获中心实体 $i$ 邻域的全局信

息,采用平均池化来处理每个子图,得到关系子图的向量表示:$\boldsymbol{g}_r \in \mathbf{R}^{N_r \times d'}$,$d'$为通过多层注意力机制输出的实体表示的维度。由于不同关系子图对中心实体的贡献度不同,对于每个子图,计算其注意力值 $O_r \in \mathbf{R}^{N_r}$,如下:

$$O_r = \boldsymbol{q}^{\mathrm{T}} \cdot \tanh(\boldsymbol{W}_r \cdot (\boldsymbol{g}_r)^{\mathrm{T}} + \boldsymbol{b}_r) \tag{7}$$

式中:$\boldsymbol{q} \in \mathbf{R}^{d''}$为子图共享的注意力向量;$\boldsymbol{W}_r \in \mathbf{R}^{d'' \times d'}$为特定于子图表示 $\boldsymbol{g}_r$ 的权重矩阵;$\boldsymbol{b}_r \in \mathbf{R}^{d'' \times N_r}$为偏置向量;tanh 为非线性激活函数。

使用 softmax 函数归一化注意力值得到子图的权重系数 $\alpha_r$ 为

$$\alpha_r = \frac{\exp(\mathrm{LeakyReLU}(O_r))}{\sum_{r' \in N_r} \exp(\mathrm{LeakyReLU}(O_{r'}))} \tag{8}$$

较大的 $\alpha_r$ 意味着相应的关系子图嵌入 $\boldsymbol{g}_r$ 更为重要。对于所有的子图嵌入,聚合这些嵌入得到最终的实体嵌入向量 $\boldsymbol{e}''_i = \sum_{r \in N_r} \alpha_r \cdot \boldsymbol{g}_r$。

### 1.4　子图对比约束

在知识图谱中,实体的表示依赖于其局部邻域信息,不同的实体有不同的邻域上下文信息。为了有效地封装三元组的特征和学习高质量的实体表示,本文考虑了中心实体与其关系子图之间的强相关性,采用子图对比约束来保证实体嵌入的质量。

模型依赖于关系子图中特定的中心实体,将其真实关系子图表示与一个错误的关系子图表示进行对比。具体地说,对于每个特定关系子图中的实体表示 $\boldsymbol{e}''_i$,将其关系子图表示 $\boldsymbol{g}_r$ 视为正样本。对于子图表示 $\boldsymbol{g}_r$,设计了一个混淆函数 $\mathcal{F}(\cdot)$,为了方便计算,给定一组上下文子图表示,混淆函数 $\mathcal{F}(\cdot)$通过随机扰乱子图中节点以生成负样本子图表示 $\tilde{\boldsymbol{g}}_r$,使得中心实体与其上下文子图强相关,而与其他负样本子图弱相关:

$$\mathcal{F}(\{\boldsymbol{g}_{r_1}, \boldsymbol{g}_{r_2}, \cdots, \boldsymbol{g}_{r_N}\}) \rightarrow \{\tilde{\boldsymbol{g}}_{r_1}, \tilde{\boldsymbol{g}}_{r_2}, \cdots, \tilde{\boldsymbol{g}}_{r_N}\}$$

本文利用间隔三元组损失[15]对模型进行优化,以便在一定程度上可以得到很好的区分,并得到高质量的实体表示。子图对比损失表示为

$$\mathcal{L}_{\mathrm{cl}} = \frac{1}{N_r} \sum_{i=1}^{N_r} (-\max(\sigma(\boldsymbol{e}''_i \boldsymbol{g}_r) - \sigma(\boldsymbol{e}''_i \tilde{\boldsymbol{g}}_r) + \tau, 0)) \tag{9}$$

式中:$\sigma(x) = 1/[1 + \exp(-x)]$为 softmax 激活函数;$\tau$ 为间隔值;$\boldsymbol{e}''_i$为编码器输出的实体向量;$\boldsymbol{g}_r$ 为正样本子图表示;$\tilde{\boldsymbol{g}}_r$ 为负样本子图表示。在子图对比约束下,增强了关系子图上下文中实体嵌入的质量。

### 1.5　解 码 器

解码器模块可以采用现有的知识图谱嵌入模型。本文使用 ConvE[6]作为解码器来验证 GCAT 模型的有效性。ConvE[6]模型是评估三元组合理性最常用的解码器之一,主要通过卷积层和全连接层来建模输入实体和关系之间的交互作用。给定$(h, r, t)$三元组,ConvE[6]首先将 $h$ 和 $t$ 重新嵌入到二维张量中,然后对被重塑的张量应用标准的卷积运算来计算三元组的分数。在 ConvE[6]中,三元组的评分函数定义为

$$p(h, r, t) = \mathrm{RELU}(\mathrm{vec}(\mathrm{RELU}([\boldsymbol{e}_h; \boldsymbol{e}_r] * \omega)\boldsymbol{W}))\boldsymbol{e}_t \tag{10}$$

式中:$\omega$ 为一组滤波器;"$*$"为卷积算子;vec$(\cdot)$为将张量转换为向量;RELU 为激活函数。

为了训练模型,使用带有标签平滑效果的标准交叉熵损失,定义如下:

$$\mathcal{L}_{\mathrm{tri}} = -\frac{1}{N} \sum_i (t_i \cdot \ln p_i + (1 - t_i) \cdot \ln(1 - p_i)) \tag{11}$$

式中:$t_i$ 为三元组 $i$ 的标签;$p_i$ 为三元组相应的得分。

结合知识图谱嵌入和子图对比约束,总的目标函数定义如下:

$$\mathcal{L} = \mathcal{L}_{\mathrm{tri}} + \alpha \mathcal{L}_{\mathrm{cl}} \tag{12}$$

式中:$\alpha$ 为子图对比约束的超参数。本文共同训练 GCAT 的所有组件。

## 2　实验准备

### 2.1　数 据 集

本文在 2 个基准数据集上进行了实验以评估知识图谱补全任务的性能,数据集的详细统计数据如表 1 所示,$N_e$ 为实体数量,$N_r$ 为关系数量,Train、Valid 和 Test 分别表示训练集、验证集和测试集的三元组数。FB15k-237[20]为数据集 FB15k[3]的子集,删除了反转关系以解决可逆关系问题。WN18RR[6]为数据集 WN18[3]的子集,同样删除了反转事实三元组,以确保评估数据集不会由于冗余反关系而出现测试泄露。

**表 1　数据集统计数据**

**Table 1　Statistics of datasets**

| 数据集 | $N_e$ | $N_r$ | Train | Valid | Test |
|---|---|---|---|---|---|
| FB15k-237 | 14 541 | 237 | 272 115 | 17 535 | 20 466 |
| WN18RR | 40 943 | 11 | 86 835 | 3 034 | 3 134 |

### 2.2 基线方法

为了证明模型的有效性,本文选取了 11 种基线模型进行比较。

1) TransE[3]。一种标准的几何模型,根据所选的距离函数,要求尾实体嵌入约为头实体和关系嵌入的总和。

2) DistMult[4]。将关系矩阵简化为对角线矩阵的标准双线性模型。

3) ComplEx[5]。通过引入复数值嵌入来扩展 DistMult,以更好地建立模型的非对称关系。

4) ConvE[6]。一种先进的 CNN 模型,其中 CNN 用于定义评分功能。

5) RotatE[21]。一种流行的旋转模型,该模型将关系视为复杂矢量空间中从头实体到尾实体的旋转。

6) R-GCN[10]。通过权重共享和系数约束有效地建模多关系网络。

7) HypER[22]。一种超网络体系结构,可生成简化的特定于关系的卷积过滤器。

8) TuckER[23]。一个具有完全表达能力的基于张量分解的模型。

9) A2N[11]。一种基于 GNN 的模型,通过注意力机制学习实体的查询相关表示。

10) CompGCN[12]。一种建模多关系有向图的 GNN 模型,能够同时学习实体和关系的向量表示。

11) InteractE[7]。一种通过增加实体和关系的异构特征交互的知识图谱嵌入模型。

### 2.3 参数设置

模型的最终参数设置根据验证集上的平均倒数排名(MRR)确定。本文选择的超参数范围如下:实体和关系嵌入维度{100,200},学习率{0.000 5,0.001,0.01},批量大小{128,256,512,1 024},标签平滑{0,0.1,0.2},子图对比约束超参数 $\alpha$ {0.1,0.01,0.001}。

实验中,将实体和关系维度设置为 200,学习率为 0.01,标签平滑为 0.1,批量大小为 256,2 层 GAT,第 1 层采用 8 个注意力头,第 2 层采用 1 个注意力头,对于数据集 FB15k-237 和数据集 WN18RR,子图对比约束超参数 $\alpha$ 设置为 0.001。根据验证集的 MRR 使用提前停止策略,每 20 个 epochs 对其进行一次评估,并使用了 Adam 优化器进行优化。

### 2.4 评估指标

在知识图谱补全任务中,对于每个测试三元组$(h,r,t)$,通过将 $h$ 和 $t$ 替换为数据集中的所有其他实体以计算三元组得分,根据得分进行降序排序。类似于文献[24]的工作,在测试阶段,应用过滤器设置,在训练集、验证集和测试集中已经存在的有效三元组进行排序之前将被过滤掉。本文使用 2 个标准度量指标来评估知识图谱补全性能、MRR 及所有测试三元组的排名分数的比例(Hits@$N$),其中,$N$ = 1,3,10。

## 3 实验分析

### 3.1 实验结果

模型在数据集 FB15k-237 和 WN18RR 使用相同的实验环境:Pytorch1.7.0,CUDA10.1。将本文模型与基线模型进行实验比较,结果如表 2 和表 3 所示。

数据集 FB15k-237 上,GCAT 获得了最好的 MRR 和最高的 Hits@3、Hits@1。具体分析如下:①本文方法优于 TransE、DistMult 等基准知识图谱

**表 2 数据集 FB15k-237 的实验结果**

**Table 2 Experimental results on FB15k-237 dataset**

| 模型 | MRR | Hit@1 | Hit@3 | Hit@10 |
|---|---|---|---|---|
| TransE[3] | 0.294 | | | 0.465 |
| DistMult[4] | 0.241 | 0.155 | 0.263 | 0.419 |
| ComplEx[5] | 0.247 | 0.158 | 0.276 | 0.428 |
| ConvE[6] | 0.325 | 0.237 | 0.356 | 0.501 |
| RotatE[21] | 0.338 | 0.241 | 0.375 | 0.533 |
| R-GCN[10] | 0.248 | | | 0.417 |
| HypER[22] | 0.341 | 0.252 | 0.376 | 0.520 |
| TuckER[23] | 0.358 | 0.266 | 0.394 | 0.544 |
| A2N[11] | 0.317 | 0.232 | 0.348 | 0.486 |
| CompGCN[12] | 0.355 | 0.264 | 0.390 | 0.535 |
| InteractE[7] | 0.354 | 0.263 | | 0.535 |
| GCAT | 0.359 | 0.269 | 0.395 | 0.540 |

**表 3 数据集 WN18RR 的实验结果**

**Table 3 Experimental results on WN18RR dataset**

| 模型 | MRR | Hit@1 | Hit@3 | Hit@10 |
|---|---|---|---|---|
| TransE[3] | 0.226 | | | 0.501 |
| DistMult[4] | 0.430 | 0.390 | 0.440 | 0.490 |
| ComplEx[5] | 0.440 | 0.410 | 0.460 | 0.510 |
| ConvE[6] | 0.430 | 0.400 | 0.440 | 0.520 |
| RotatE[21] | 0.476 | 0.428 | 0.492 | 0.571 |
| R-GCN[10] | | | 0.137 | |
| HypER[22] | 0.465 | 0.436 | 0.477 | 0.522 |
| TuckER[23] | 0.470 | 0.443 | 0.482 | 0.526 |
| A2N[11] | 0.450 | 0.420 | 0.460 | 0.510 |
| CompGCN[12] | 0.479 | 0.443 | 0.494 | 0.546 |
| InteractE[7] | 0.463 | 0.430 | | 0.528 |
| GCAT | 0.482 | 0.447 | 0.495 | 0.546 |

嵌入模型,主要原因是 GCAT 充分考虑了实体邻域结构信息,而不是对独立的三元组建模。②本文方法较基于 CNN 方法的 ConvE 在 Hit@10 上提高了 0.039,MRR 上提升了 0.034。由此可知,与 CNN 模型相比,GCAT 可以捕获更多的实体和关系的上下文信息。

在数据集 WN18RR 上,GCAT 获得了最好的 MRR 和最高的 Hits@3、Hits@1。具体分析如下:①与 TuckER 模型相比,本文模型在 MRR 上获得了 0.012 的提升,在 Hits@10 上获得了 0.020 的提升。这表明了本文模型的有效性。②RotatE 模型建模和推断了关系模式,与 ComplEx、ConvE 等模型相比性能得到了显著提升。由于 RotatE 模型使用了 1 000 维初始化嵌入,比 GCAT 模型 200 维嵌入要大,RotatE 模型使用了自对抗负抽样方法,GCAT 在 Hit@10 指标上结果比 RotatE 模型差。③与基于 GNN 的方法 R-GCN、A2N 和 CompGCN 相比,GCAT 模型能够编码更多的三元组特征,进一步提高了知识图谱补全性能。

### 3.2　消去测试

为了验证对比学习机制在 GCAT 中的影响,本文通过考虑其变体进行消去实验。GCAT-wo 为不加入对比学习机制的 GCAT 变体。本文在数据集 FB15k-237 和 WN18RR 上进行了实验,结果如表 4 和表 5 所示。可以得出,GCAT 模型始终优于其变体。一个可能的原因就是 GCAT-wo 变体没有完全对邻域局部结构进行建模,这说明引入对比学习机制能够充分挖掘数据之间的特征差异,有效地封装实体局部上下文结构信息,使得相似的实体在特征空间中距离更近,从而更准确地预测实体之间缺失的链接关系,进一步提高知识图谱补全模型的预测性能。

**表 4　GCAT 和其变体方法在数据集 FB15k-237 上的实验结果**

**Table 4　Experimental results of GCAT and its variant on FB15k-237 dataset**

| 方法 | MRR | Hit@1 | Hit@3 | Hit@10 |
|---|---|---|---|---|
| GCAT-wo | 0.357 | 0.266 | 0.392 | 0.540 |
| GCAT | 0.359 | 0.269 | 0.395 | 0.540 |

**表 5　GCAT 和其变体方法在数据集 WN18RR 上的实验结果**

**Table 5　Experimental results of GCAT and its variant on WN18RR dataset**

| 方法 | MRR | Hit@1 | Hit@3 | Hit@10 |
|---|---|---|---|---|
| GCAT-wo | 0.475 | 0.441 | 0.489 | 0.541 |
| GCAT | 0.482 | 0.447 | 0.495 | 0.546 |

## 4　结　论

1)本文提出了一个端到端的用于知识图谱补全的图对比注意力网络,在统一的框架中使用图注意力网络,同时建模了局部邻域内实体和关系特征,并融合了更多实体的邻域上下文结构信息。

2)模型中融入了对比学习机制,通过对比实体和其子图的编码信息,有效捕获了三元组的特征,充分利用了知识图谱中的异质特性和保证实体嵌入的质量。

3)图对比注意力网络模型可以有效地补全知识图谱中缺失的三元组。链接预测的实验结果表明,邻域的上下文信息和异质结构具有重要意义,并证明了本文模型相对于基线模型的优越性。

### 参考文献(References)


[1] BOLLACKER K, EVANS C, PARITOSH P, et al. Freebase: A collaboratively created graph database for structuring human knowledge[C]// Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data. New York: ACM, 2008: 1247-1250.

[2] LEHMANN J, ISELE R, JAKOB M, et al. DBpedia-A large-scale, multilingual knowledge base extracted from Wikipedia [C]// The Semantic Web, 2015: 167-195.

[3] BORDES A, USUNIER N, GARCIA-DURÁN A, et al. Translating embeddings for modeling multi-relational data[C]// Proceedings of the 26th International Conference on Neural Information Processing Systems. New York: ACM, 2013: 2787-2795.

[4] YANG B, YIH W, HE X, et al. Embedding entities and relations for learning and inference in knowledge bases[C]// Processings of the International Conference on Learning Representations (ICLR), 2015: 1-12.

[5] TROUILLON T, WELBL J, RIEDEL S, et al. Complex embeddings for simple link prediction[C]// Proceedings of the 33rd International Conference on Machine Learning (ICML), 2016: 2071-2080.

[6] DETTMERS T, MINERVINI P, STENETORP P, et al. Convolutional 2D knowledge graph embeddings[C]// Proceedings of the AAAI Conference on Artificial Intelligence, 2018, 32: 1811-1818.

[7] VASHISHTH S, SANYAL S, NITIN V, et al. InteractE: Improving convolution-based knowledge graph embeddings by increasing feature interactions[C]// Proceedings of the AAAI Conference on Artificial Intelligence, 2020, 34(3): 3009-3016.

[8] JIANG X T, WANG Q, WANG B. Adaptive convolution for multi-relational learning[C]// Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL), 2019: 978-987.

[9] NGUYEN D Q, NGUYEN T D, NGUYEN D Q, et al. A novel embedding model for knowledge base completion based on con-

volutional neural network[C]// Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL),2018:327-333.

[10] SCHLICHTKRULL M, KIPF T N, BLOEM P, et al. Modeling relational data with graph convolutional networks[C]// The Semantic Web,2018:593-607.

[11] BANSAL T, JUAN D C, RAVI S, et al. A2N: Attending to neighbors for knowledge graph inference[C]// Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics,2019:4387-4392.

[12] VASHISHTH S, SANYAL S, NITIN N, et al. Composition-based multi-relational graph convolutional networks[C]// Proceedings of the International Conference on Learning Representations (ICLR),2020:1-15.

[13] VELICKOVIC P, CUCURULL G, CASANOVA A, et al. Graph attention networks[C]// Proceeding of the International Conference on Learning Representations(ICLR),2018:1-12.

[14] JAISWAL A, BABU A R, ZADEH M Z, et al. A survey on contrastive self-supervised learning [J]. Technologies, 2020, 9(1):2.

[15] JIAO Y Z, XIONG Y, ZHANG J W, et al. Sub-graph contrast for scalable self-supervised graph representation learning[C]// IEEE International Conference on Data Mining. Piscataway: IEEE Press,2020:222-231.

[16] SUN F Y, HOFFMAN J, VERMA V, et al. InfoGraph: Unsupervised and semi-supervised graph-level representation learning via mutual information maximization[C]// Proceeding of the International Conference on Learning Representations(ICLR), 2020:1-16.

[17] YOU Y N, CHEN T L, SUI Y D, et al. Graph contrastive learning with augmentations[C]// Proceedings of the 34th Conference on Neural Information Processing Systems (NeurIPS), 2020.

[18] VELICKOVIC P, FEDUS W, HAMILTON W L, et al. Deep graph infomax[C]// Proceeding of the International Conference on Learning Representations(ICLR),2019:1-17.

[19] PENG Z, HUANG W B, LUO M N, et al. Graph representation learning via graphical mutual information maximization[C]// Proceedings of the Web Conference 2020. New York: ACM, 2020:259-270.

[20] TOUTANOVA K, CHEN D Q, PANTEL P, et al. Representing text for joint embedding of text and knowledge bases[C]// Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing,2015:1499-1509.

[21] ZHOU X H, YI Y H, JIA G. Path-RotatE: Knowledge graph embedding by relational rotation of path in complex space[C]// 2021 IEEE/CIC International Conference on Communications in China (ICCC). Piscataway: IEEE Press,2021:905-910.

[22] BALAZEVIC I, ALLEN C, HOSPEDALES T. Hypernetwork knowledge graph embeddings[C]// Proceeding of the 28th International Conference on Artificial Neural Networks (ICANN), 2019:553-565.

[23] BALAZEVIC I, ALLEN C, HOSPEDALES T. TuckER: Tensor factorization for knowledge graph completion[C]// Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP),2019:5185-5194.

[24] SUN Z Q, VASHISHTH S, SANYAL S, et al. A re-evaluation of knowledge graph completion methods[C]// Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics,2020:5516-5522.

# Knowledge graph completion based on graph contrastive attention network

LIU Danyang[1], FANG Quan[2,*], ZHANG Xiaowei[1], HU Jun[2], QIAN Shengsheng[2], XU Changsheng[2]

(1. Henan Institute of Advanced Technology, Zhengzhou University, Zhengzhou 450000, China;
2. National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China)

**Abstract**: Knowledge graph (KG) completion aims to predict missing links based on the known triples in a knowledge base. Since most KG completion methods dealt with triples independently without capture the heterogeneous structure of KG and the rich information that was inherent the in neighbor nodes, which resulted in incomplete mining of triple features. This study revisits the end-to-end KG completion task, and proposes a novel graph contrastive attention network (GCAT), which can capture latent representations of entities and relations simultaneously through attention mechanism, and encapsulate more neighborhood context information from the entity. Specifically, to effectively encapsulate the features of triples, a subgraph-level contrastive training object is introduced, enhancing the quality of generated entity representation. To justify the effectiveness of GCAT, the proposed model is evaluated on link prediction tasks. Experimental results show that on the dataset FB15k-237, MRR of the model is 0.005 and 0.042 higher than that of InteractE and A2N, respectively, and that on the dataset WN18RR, MRR is 0.019 and 0.032 higher than that of InteractE and A2N, respectively. Experiments prove that the proposed model can effectively predict the missing links in KGs.

**Keywords**: knowledge graph (KG); attention mechanism; contrastive learning; knowledge graph completion; link prediction

**Received**: 2021-09-06; **Accepted**: 2021-09-17; **Published online**: 2021-11-02 10:15

**URL**: kns.cnki.net/kcms/detail/11.2625.V.20211101.1526.008.html

**Foundation items**: National Natural Science Foundation of China (62072456,62036012); 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.1001-5965.2021.0526

# 文本信息辅助图像差异描述生成

陈玮婧, 王维莹, 金琴*
(中国人民大学 信息学院, 北京 100872)

**摘 要**: 图像描述生成任务要求机器自动生成自然语言文本来描述图像所呈现的语义内容,从而将视觉信息转化为文本描述,便于对图像进行管理、检索、分类等工作。图像差异描述生成是图像描述生成任务的延伸,其难点在于如何确定2张图像之间的视觉语义差别,并将视觉差异信息转换成对应的文本描述。基于此,提出了一种引入文本信息辅助训练的模型框架TA-IDC。采取多任务学习的方法,在传统的编码器-解码器结构上增加文本编码器,在训练阶段通过文本辅助解码和混合解码2种方法引入文本信息,建模视觉和文本2个模态间的语义关联,以获得高质量的图像差别描述。实验证明,TA-IDC模型在3个图像差异描述数据集上的主要指标分别超越已有模型最佳结果12%、2%和3%。

**关 键 词**: 图像差异描述; 模态融合; 图像描述; 计算机视觉; 自然语言处理
**中图分类号**: TP37
**文献标志码**: A **文章编号**: 1001-5965(2022)08-1436-09

图像内容描述是计算机视觉和人工智能领域一个热门的研究课题,其要求计算机生成自然语言语句来描述图像呈现的视觉内容。图像描述生成模型可以自动将图像转化为语言描述,更容易对图像进行检索、分类等处理。

现有的图像描述生成工作多聚焦于自动生成对单张图像的描述。描述2张相似图像之间差别的图像差异描述则是该领域近年来新兴起的任务。在实际应用方面,该任务可用于辅助物种识别、医学影像观察及监视和追踪多媒体的变化等。且由于需要描述差异的2张图像十分相似,模型需要学会捕捉和描述图像间的细粒度差别,因此,该任务也能帮助理解细粒度的图像信息。

相比一般的图像描述,图像差异描述不仅要求模型能够正确理解图像所呈现的视觉语义信息,还要求模型具有比较视觉语义从而识别出差异的能力,并要将这种差异用自然语言准确描述出来。因此,解决图像差异描述任务的关键在于:如何准确地捕捉到相似图像之间的差异,以及如何更好地建模视觉和语言之间的跨模态语义关联。

本文调研并总结了前人在图像差异描述领域的工作,发现现存模型主要基于编码器-解码器架构。这些模型先通过预训练模型提取2张图像的各区域局部特征。在编码图像差异时,对2张图像的局部特征进行简单的运算,再通过注意力机制[1-3]计算局部区域之间的相似度提取视觉差异。最终编码后的特征被输入到解码器中解码得到差别描述语句。在模型训练过程中,文本信息只在整个模型的输出端参与损失函数的计算,提供监督信息,没有被充分利用于跨模态语义信息的建模。

因此,本文提出了一种在训练阶段引入文本信息辅助跨模态语义信息建模的框架TA-IDC。

**收稿日期**: 2021-09-06; **录用日期**: 2021-09-17; **网络出版时间**: 2021-10-15 14:24
**网络出版地址**: kns.cnki.net/kcms/detail/11.2625.V.20211015.0739.001.html
**基金项目**: 国家自然科学基金(61772535,62072462); 北京市自然科学基金(4192028)
* **通信作者**. E-mail: qjin@ruc.edu.cn

**引用格式**: 陈玮婧, 王维莹, 金琴. 文本信息辅助图像差异描述生成[J]. 北京航空航天大学学报, 2022, 48(8): 1436-1444.
CHEN W J, WANG W Y, JIN Q. Image difference caption generation with text information assistance [J]. Journal of Beijing University of Aeronautics and Astronautics, 2022, 48(8): 1436-1444 (in Chinese).

该框架在图像差异编码器-解码器的传统模型基础上,增加一个文本的编码器,在训练阶段将文本信息引入模型中,从而通过多任务学习的方式使文本以多种形式参与到跨模态语义关联的建模中。

本文提出了一种文本辅助解码的方法,在训练阶段,使解码器分别基于编码的视觉信息和文本信息生成差异描述语句。由于解码器要基于来自不同模态的信息生成相同的描述语句,模型需要使编码后的视觉信息与文本信息在语义空间中相似,从而协助跨模态信息之间的关联。

更进一步的,本文还提出了一种混合解码的方法。在训练阶段,将视觉信息和文本信息混合作为解码器的输入信号,以促进不同模态信息的融合,更好地建模视觉和文本 2 种模态信息之间的语义关联。通过对比本文模型和现有模型在多个数据集上生成描述的各项指标,可以证明本文模型的有效性和鲁棒性。

## 1　相关工作

图像描述任务涉及计算机视觉和自然语言处理两大领域,涉及视觉模态和文本模态的跨模态语义理解,具有很高的研究价值。同时,相关技术在视觉辅助、基于文本的图像搜索、社交媒体上的图像理解等方面有着广泛的实际应用[2],在现实世界中有广泛的应用场景。区别于图像物体识别任务,图像描述不仅要求计算机能够识别出图像中的物体,还要求其能理解图像中包含的语义信息,如图像中物体的状态、属性、物体之间的关系等,并用自然语言将其表达出来[4-5]。如何弥补 2 种模态之间的语义鸿沟,将图像中的视觉信息和文本信息对应起来,是该项任务的一大难点,也是关键所在。

图像差异描述是图像描述任务的延伸。目前提到的图像描述研究,大多是指对单张图像生成自然语言描述,而图像差异描述任务则要对 2 张相似的图像生成描述它们之间差异的句子,如图 1所示。由于图像之间存在一定的相似性,模型需要能排除其他无关因素(如拍摄角度的变化)的干扰,准确找到图像间的差异并准确描述出来。这要求模型学会理解和描述图像间细粒度的差别,因此也可以推广到单张图像的细粒度描述任务上,在未来可进一步用于协助普通图像描述模型生成更细粒度的图像描述。

近年来,图像差异描述的研究引起了学术界的广泛关注。图像描述中经典的编码器-解码器结构[6-8],还有基于注意力机制[9-12]方法被迁移应用到了图像差异描述任务中(见表 1)。Jhamtani 和 Berg-Kirkpatrick[1] 使用 VIRAT 监控视频数据集,按照一定的方法从中抽取图像帧组对进行人工标注,构造出一个监控图像差异描述数据集 Spot-the-diff,还提出了一种基于预先提取的物体

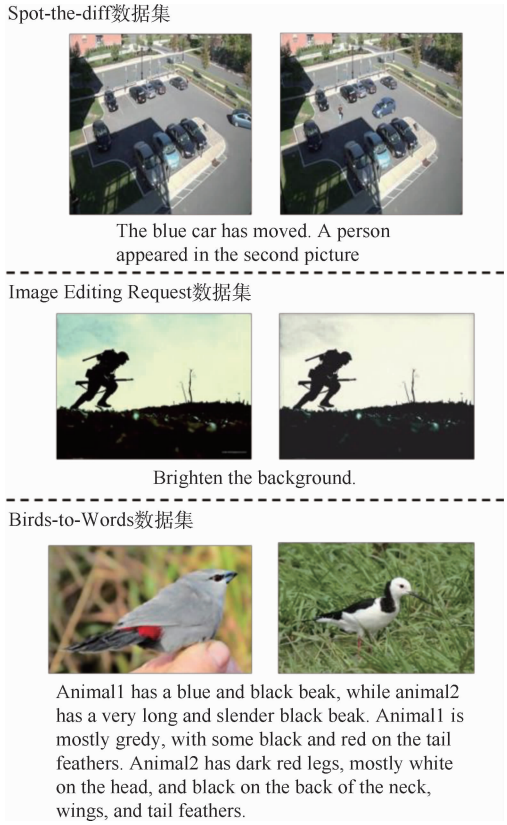



图 1　图像差异描述示例

Fig. 1　Examples of image difference captioning

**表 1　图像差异描述相关工作**

**Table 1　Related work of image difference captioning**

| 模型 | 数据集 | 模型介绍 |
|---|---|---|
| [1] | Spot-the-diff | 提出一个基于预先提取的物体级别差异特征的模型,其对复杂变化、细微变化和多处变化的描述效果欠佳 |
| [2] | Spot-the-diff、Image Editing Request | 提出一个基于动态相关注意力机制的模型,其对复杂变化描述效果欠佳,且存在描述语句不流畅的问题 |
| [3] | Birds-to-Words | 提出基于 Transformer 结构的差异描述模型,其存在描述不存在的差异、描述差异错误的问题 |
| [13] | Birds-to-Words | 结合使用语义分割模型和图卷积神经网络描述图像差异,并引入额外的单图描述数据进行增强,整体模型结构复杂,参数量大训练时间长,且只在一个图像差异数据集上进行了实验 |

级别差异特征的图像描述模型,适用于 Spot-the-diff 数据集中仅涉及局部物体变化的相同场景图像对,即可以通过将 2 张图像相减来确定差异特征的场景。Tan 等[2]通过图像编辑网站收集数据并进行筛选标注,构造了一个语义更丰富多变的图像编辑差异描述数据集 Image Editing Request,并提出了一个基于动态相关注意力机制的模型,该模型在 Spot-the-diff 数据集和 Image Editing Request 数据集上均取得了更好的效果。Forbes 等[3]设计了一个涉及更细粒度差异的鸟类图像差异描述数据集 Birds-to-Words,每个样例包含 2 张在野外环境中拍摄的鸟类照片,存在拍摄角度、距离、光线和背景等差异,但描述文本仅关注鸟类外形的差异,即羽毛、喙等身体部位的大小、长度及颜色等特性的细粒度差别,这对模型的鲁棒性和准确性提出了更高的要求,还提出了基于 Transformer 结构的差异描述模型,并在 Birds-to-Words 数据集上进行了实验。针对该数据集,Yan 等[13]使用语义分割模型预先分割出图像中鸟类的区域,再用图卷积神经网络编码该部分特征,提取鸟类外形差异,还使用额外的单张图像描述数据来增强模型对图像的语义理解,在 Birds-to-Words 数据集上的表现超过了已有的模型。

总体来说,目前已有的图像差异描述生成模型大多在训练阶段没有充分利用文本信息进行训练。本文模型针对这一问题进行了改进,通过在基线模型上增加文本编码器引入文本信息辅助训练的方法,显著提升了描述的总体质量。

## 2 模型改进

图像差异描述任务可以定义为:给定 2 张图像$(i_1,i_2)$作为输入,要求生成一段自然语言描述$s=(w_1,w_2,\cdots,w_n)$来形容 2 张图像之间的区别。

本节将详细介绍针对图像差异描述任务设计的基线模型,以及本文提出的使用文本信息辅助训练的多任务学习框架,即文本辅助解码和混合解码 2 种改进方法。

### 2.1 基线模型

本文的基线模型采用传统的编码器-解码器结构,如图 2 所示。基线模型可分为 2 个模块:图像编码模块和解码模块。使用 Transformer 模型作为图像的编码器,使用带有注意力机制的门控循环单元(gate recurrent unit, GRU)[14]模型作为解码器。基线模型的输入为 2 张相似的图像,输出为描述其差异的自然语言语句。

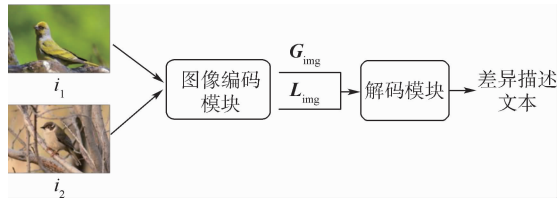



图 2 基线模型整体结构示意图

Fig. 2 Framework of baseline model

#### 2.1.1 图像编码模块

输入图像$i_1$、$i_2$首先分别经过一个预训练好的卷积神经网络[15](convolutional neural network, CNN),将图像切割成$d\times d$个区域,分别提取出每个区域的局部特征,在最后一层卷积层输出一个形状为$(d,d,f)$的特征向量。然后,将其压平成形如$(d^2,f)$的矩阵$\boldsymbol{E}$,作为图像的初级特征表示。

定义图像$i_1$、$i_2$的初级特征表示为$\boldsymbol{E}^1$、$\boldsymbol{E}^2$,$\boldsymbol{e}^1_{x,y}$表示第 1 个图像中第$(x,y)$区域的局部特征($x,y\in[1,d]$),则有:$\boldsymbol{E}^1=\mathrm{CNN}(i_1)=\langle \boldsymbol{e}^1_{1,1},\cdots,\boldsymbol{e}^1_{d,d}\rangle$,$\boldsymbol{E}^2=\mathrm{CNN}(i_2)=\langle \boldsymbol{e}^2_{1,1},\cdots,\boldsymbol{e}^2_{d,d}\rangle$。

然后需要使 2 张图像的特征进行交互。考虑到不同图像的特征交互方式可能对模型产生不同的影响,在初步尝试了多种特征交互方式后,选择相连、相减和混合这 3 种比较有效的特征交互方式作为候选,如图 3 所示。定义 2 张图像的综合特征为$\boldsymbol{J}$,则有$\boldsymbol{J}\in\{\boldsymbol{E}^1\circ\boldsymbol{E}^2,\boldsymbol{E}^1-\boldsymbol{E}^2,\boldsymbol{E}^1\circ\boldsymbol{E}^2\circ(\boldsymbol{E}^1-\boldsymbol{E}^2)\}$。其中,“$\circ$”表示将 2 个向量直接进行拼接,即:$\boldsymbol{E}^1\circ\boldsymbol{E}^2=\langle \boldsymbol{e}^1_{1,1},\cdots,\boldsymbol{e}^1_{d,d},\boldsymbol{e}^2_{1,1},\cdots,\boldsymbol{e}^2_{d,d}\rangle$。

给定图像综合特征$\boldsymbol{J}$,基线模型使用一个$N$层的 Transformer 模型作为图像的编码器,其中每层通过注意力机制编码局部特征之间的关联。

定义 Transformer 中带层归一化的多头自注意

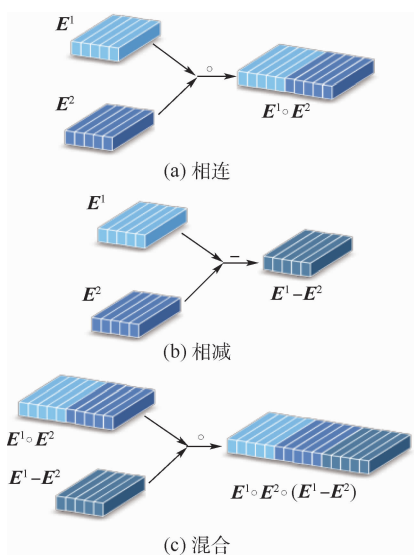



图 3 三种特征交互方式

Fig. 3 Three types of feature interaction

力层(multi-head self-attention)为$\mathrm{ATTN}_{\mathrm{MH}}$,带层归一化的前馈网络(feed forward network)为 FFN。编码器第 $i$ 层的输入为第 $i-1$ 层的输出 $\boldsymbol{L}^{i-1}$,经过自注意力层与前馈网络编码后第 $i$ 层的输出为 $\boldsymbol{L}^{i}$:

$$\boldsymbol{L}^{i} = \mathrm{FF}(\mathrm{ATTN}_{\mathrm{MH}}(\boldsymbol{L}^{i-1})) \tag{1}$$

编码器第 1 层的输入初始 $\boldsymbol{L}^{0}=\boldsymbol{J}$,最终输出 $\boldsymbol{L}=\boldsymbol{L}^{N}$ 为编码器编码得到的图像特征。

### 2.1.2　解码模块

解码模块以编码后的图像特征 $\boldsymbol{L}$ 为输入,解码生成差异描述文本。使用 $M$ 层带注意力机制的 GRU 模型作为解码器,其结构如图 4 所示。

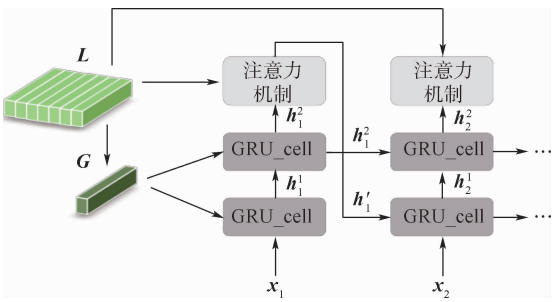



图 4　解码模块的结构示意图

Fig. 4　Structure of decoding module

对一般 GRU 模型,令 $\mathrm{GRU}_{\mathrm{D}}$ 表示解码器,$\boldsymbol{h}_t$ 为时间 $t$ 解码器隐藏层的输出向量,$\boldsymbol{x}_t$ 为时间 $t$ 的输入:

$$\boldsymbol{h}_t = \mathrm{GRU}_{\mathrm{D}}(\boldsymbol{x}_t, \boldsymbol{h}_{t-1}) \tag{2}$$

输出预测结果时,对每个时间点 $t$,将 GRU 模型的最后一层输出 $\boldsymbol{h}_t^M$ 依次经过一个线性层和一个 softmax 层,最终得到对应的预测词语 $w_t$:

$$p_\theta(w_t \mid w_{<t}, \boldsymbol{L}) = \mathrm{softmax}(\boldsymbol{W}_1 \boldsymbol{h}_t^M + \boldsymbol{b}_1) \tag{3}$$

使用交叉熵损失函数进行目标优化:

$$\mathrm{Loss}_{\mathrm{I}} = -\sum_{s} \log_2 p_\theta(s \mid \boldsymbol{L}) \tag{4}$$

本文的解码器在一般 GRU 模型上进行了 2 点改进,即使用全局特征初始化和引入注意力机制。

注意到,GRU 模型需要初始化隐藏层向量。为了更好地编码图像中的变化信息,可以从图像局部特征 $\boldsymbol{L}$ 中提取图像全局特征 $\boldsymbol{G}$,用于初始化解码器的隐向量 $\boldsymbol{h}_0$:

$$\boldsymbol{G} = \mathrm{Global}(\boldsymbol{L}) \tag{5}$$

$$\boldsymbol{h}_0 = \boldsymbol{G} \tag{6}$$

式中:Global()表示从图像局部特征 $\boldsymbol{L}$ 中提取出图像全局特征 $\boldsymbol{G}$。本文采用的是简单将局部特征取平均的方法,得到图像的全局特征 $\boldsymbol{G}$。

另一个改进是引入注意力机制,促进跨模态语义信息交互。针对生成的第 $t$ 个时间步,利用 GRU 最后一层隐藏层的输出向量 $\boldsymbol{h}_t^M$,通过注意力机制从图像局部特征 $\boldsymbol{L}$ 中提取信息(式(7)~式(8)),即以 $\boldsymbol{h}_t^M$ 为 Query,以 $\boldsymbol{L}$ 为 Key 和 Value,$d_{\mathrm{h}}$ 为 $\boldsymbol{h}_t^M$ 的维度,计算得到包含图像信息的特征。将结果与 $\boldsymbol{h}_t^M$ 拼起来,依次经过线性层和 tanh 层,作为解码器该时间步的输出 $\boldsymbol{h}_t'$,并以此更新下一时间点的第 1 层隐向量 $\boldsymbol{h}_{t+1}^1$:

$$\boldsymbol{a}_t = \frac{(\boldsymbol{h}_t^M)^{\mathrm{T}} \boldsymbol{L}}{\sqrt{d_{\mathrm{h}}}} \tag{7}$$

$$\boldsymbol{\alpha}_t = \mathrm{softmax}(\boldsymbol{a}_t) \tag{8}$$

$$\boldsymbol{h}_{t+1}^1 = \boldsymbol{h}_t' = \tanh(\boldsymbol{W}_2[\boldsymbol{h}_t^M; \boldsymbol{\alpha}_t \boldsymbol{L}] + \boldsymbol{b}_2) \tag{9}$$

输出预测结果和目标优化过程同式(3)和式(4)。

## 2.2　TA-IDC 模型

受到 Ma 等[16]在文本摘要生成任务的相关工作的启发,本文提出了一种在训练阶段引入文本信息辅助跨模态语义信息建模的框架 TA-IDC,整体结构如图 5 所示(图中:训练阶段同时进行①②或①③,测试阶段只进行①)。该框架在基线

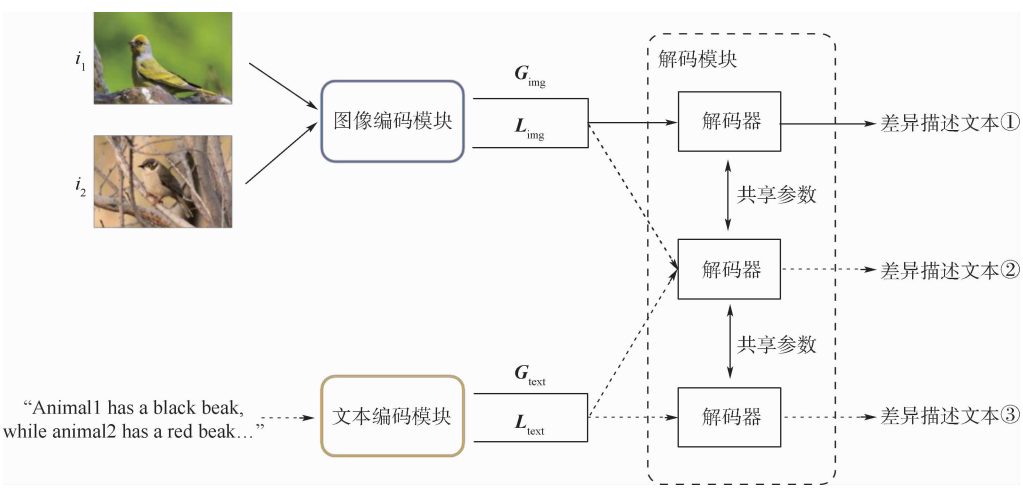



图 5　TA-IDC 模型整体结构示意图

Fig. 5　Framework of TA-IDC model

模型之上增加一个文本编码器,在训练阶段通过多任务学习的方式,引入额外的文本信息辅助训练,使文本以多种形式参与到跨模态语义建模中,最终得到一个更强的差异描述生成模型。在测试阶段只进行主任务,根据图像的视觉信息生成差异描述。在该框架之下,本文提出了文本辅助解码机制和混合解码机制 2 种引入文本信息的方法。

2.2.1 文本辅助解码机制

文本辅助解码机制的具体思路为:在训练阶段,同时将图像对应的人工标注输入文本编码器中进行编码,并将其与编码后的图像信息输入到同一个解码器中,分别生成差异描述,叠加两者损失进行训练。

本文采用 $K$ 层 GRU 模型作为文本编码器。与 GRU 模型作为解码器的方法类似,有

$$\boldsymbol{h}_t = \mathrm{GRU}_{\mathrm{E}}(\boldsymbol{x}_t, \boldsymbol{h}_{t-1}) \tag{10}$$

将模型最后一层除最后时刻外的所有时刻隐藏层输出向量($\boldsymbol{h}_1^K, \boldsymbol{h}_2^K, \cdots, \boldsymbol{h}_{n-1}^K$)作为文本的局部特征 $\boldsymbol{L}_{\text{text}}$,将最后时刻 $n$ 的最后一层隐藏层输出向量 $\boldsymbol{h}_n^K$ 作为文本的全局特征 $\boldsymbol{G}_{\text{text}}$:

$$\boldsymbol{L}_{\text{text}} = (\boldsymbol{h}_1^K, \boldsymbol{h}_2^K, \cdots, \boldsymbol{h}_{n-1}^K) \tag{11}$$

$$\boldsymbol{G}_{\text{text}} = \boldsymbol{h}_n^K \tag{12}$$

将图像编码器的输出 $\boldsymbol{G}_{\text{img}}$、$\boldsymbol{L}_{\text{img}}$和文本编码器的输出 $\boldsymbol{G}_{\text{text}}$、$\boldsymbol{L}_{\text{text}}$分别放入解码器中,按式(2)~式(9)生成最终的描述。

叠加生成损失,将损失函数修改为

$$\mathrm{Loss} = \mathrm{Loss}_{\mathrm{I}} + \mathrm{Loss}_{\mathrm{T}} = \mathrm{Loss}(\boldsymbol{G}_{\text{img}}, \boldsymbol{L}_{\text{img}}) + \mathrm{Loss}(\boldsymbol{G}_{\text{text}}, \boldsymbol{L}_{\text{text}}) \tag{13}$$

2.2.2 混合解码机制

文本辅助解码的方法虽然可以将文本信息引入进来辅助训练,但在生成描述时文本信息和图像信息并没有进行直接的交互,而是分别输入到解码器中进行训练,引导生成图像差异描述,再合计损失。因此,本文提出了另一种引入文本信息的思路:在训练阶段进行解码时,除了输入图像的视觉信息,还将图像和文本的信息混合后输入到解码器中,分别生成差异描述,叠加损失进行优化,从而加强模态间的信息交互与融合。

具体做法为:使用图像全局特征 $\boldsymbol{G}_{\text{img}}$初始化隐藏层向量 $\boldsymbol{h}_0$,通过注意力机制融合文本局部特征 $\boldsymbol{L}_{\text{text}}$:

$$\boldsymbol{h}_0 = \boldsymbol{G}_{\text{img}} \tag{14}$$

$$\boldsymbol{L} = \boldsymbol{L}_{\text{text}} \tag{15}$$

或者使用文本全局特征 $\boldsymbol{G}_{\text{text}}$初始化隐藏层向量 $\boldsymbol{h}_0$,通过注意力机制融合图像局部特征 $\boldsymbol{L}_{\text{img}}$:

$$\boldsymbol{h}_0 = \boldsymbol{G}_{\text{text}} \tag{16}$$

$$\boldsymbol{L} = \boldsymbol{L}_{\text{img}} \tag{17}$$

则最终的损失函数为

$$\mathrm{Loss} = \mathrm{Loss}_{\mathrm{I}} + \mathrm{Loss}_{\mathrm{IT}} = \mathrm{Loss}(\boldsymbol{G}_{\text{img}}, \boldsymbol{L}_{\text{img}}) + \mathrm{Loss}(\boldsymbol{G}_{\text{img}}, \boldsymbol{L}_{\text{text}}) \tag{18}$$

或

$$\mathrm{Loss} = \mathrm{Loss}_{\mathrm{I}} + \mathrm{Loss}_{\mathrm{TI}} = \mathrm{Loss}(\boldsymbol{G}_{\text{img}}, \boldsymbol{L}_{\text{img}}) + \mathrm{Loss}(\boldsymbol{G}_{\text{text}}, \boldsymbol{L}_{\text{img}}) \tag{19}$$

## 3 实验结果与分析

为了更好地检验模型的实际效果,本文在 3 个不同的数据集上进行了多次实验,并选取了合适的指标对模型生成的图像差异描述进行了评估。本节将介绍实验用到的数据集、评价指标和模型具体细节与参数设置,并通过分析实验结果,证明本文提出的改进模型相比已有模型,可以生成更好的图像差异描述。

### 3.1 数据集

本文选择了 3 个源于现实应用场景的图像差异描述数据集进行实验:监控图像差异描述数据集 Spot-the-diff[1]、图像编辑差异描述数据集 Image Editing Request[2] 和鸟类图像差异描述数据集 Birds-to-Words[3]。

1)监控图像差异描述数据集 Spot-the-diff。该数据集是在 VIRAT 监控视频数据集的基础上进行抽帧构建的。由于监控视频一般较少有角度变换,该数据集的样例中两图之间差异较小,主要是局部上的变化,如停车场或街道上一些车辆或行人位置的移动、出现或消失。

2)图像编辑差异描述数据集 Image Editing Request。该数据集中的图像主要来自图像编辑网站 Reddit(https://www.reddit.com/r/photoshoprequest)和 Zhopped(http://zhopped.com),上面有用户发布的源图像和修改请求,其他用户可以按需求对源图像进行编辑。该数据集将源图像、编辑后的图像作为一组,进行了人工筛选和标注。该数据集包含局部和全局上的差异,差异类型更丰富,涉及图中物体的增删修改及整体光线、颜色的变化。

3)鸟类图像差异描述数据集 Birds-to-Words。该数据集中的图像来自研究级别的野生动植物观测网站 iNaturalist(https://www.inaturalist.org),主要选取了其中鸟类的图像,以保证两图之间有一定的相似性。该数据集中两图之间差异相较前 2 个数据集要更丰富,而且存在背景、照片拍摄角度、拍摄光线等干扰因素,鸟类之间的差别更加细粒度。

综合来看,3 个数据集所展现的图像差异层次各不相同,涉及全局差异、局部物体级别差异和更细粒度的局部差异。本文选择这 3 个包含多样化视觉差异的数据集,可以更好地检验模型的有效性和鲁棒性。数据集的具体信息如表 2 所示。

**表 2　数据集统计信息**

**Table 2　The statistics of datasets**

| 数据集参数 | Spot-the-diff | Image Editing Request | Birds-to-Words |
|---|---|---|---|
| 训练集大小 | 17 676 | 3 053 | 12 805 |
| 验证集大小 | 3 310 | 381 | 1 556 |
| 测试集大小 | 1 270 | 493 | 337 |
| 图片数量 | 39 232 | 11 121 | 3 520 |
| 词汇量 | 2 060 | 2 460 | 2 634 |
| 词语/句子 | 10.96 | 7.50 | 32.10 |
| 句子/标注 | | | 2.60 |
| 图像差异类型 | 局部 | 全局、局部 | 局部、细粒度 |

### 3.2　评价指标

本文主要使用图像描述领域常用的自动评价指标 BLEU-4[17]、ROUGE-L[18]、METEOR[19]和 CIDEr[20]对模型进行评估。这些指标可以在一定程度上反映出模型生成的描述的流畅性和准确性。

为了能更客观地将本文模型与已有模型的生成效果进行对比,在不同数据集上训练时,以已有模型所选择的主指标作为该数据集的主指标:对 Spot-the-diff 数据集,以 CIDEr 作为主指标;对 Image Editing Request 数据集,以 METEOR 作为主指标;对 Birds-to-Words 数据集,以 ROUGE-L 作为主指标。

### 3.3　实验参数设置

图像编码模块中的卷积神经网络模型具体使用的是在 ImageNet[21]等图像数据集上预训练好的 ResNet-101[15]模型。取其全连接层之前的最后 1 个卷积层的输出作为图像的特征表示,其是大小为(7,7,2 048)的二维网格状向量,压平后为(49,2 048)的向量,即图像被划分为 49 个区域,每个区域的特征向量大小为 2 048。

图像编码模块和文本编码模块的编码器的隐藏层大小均为 512。为防止过拟合,设置丢弃率 dropout 为 0.5,学习率 lr 为 0.000 1,批大小 batch size 为 32。训练时采用梯度下降法优化损失,每 50 个 batch 在验证集上进行 1 次评估,使用每个数据集的主评价指标作为评估标准,从中选择更优的模型。当模型连续 50 轮没有继续优化时,停止训练。因数据集之间存在差异,各数据集中不同模块使用到的 Transformer 模型和 GRU 模型的层数略有不同。在 Spot-the-diff 数据集和 Image Editing Request 数据集上,图像编码器、文本编码器和解码器的层数分别为 1、2、2,在 Birds-to-Words 数据集上则为 2、2、2。

### 3.4　实验结果

#### 3.4.1　图像特征交互方式实验

本节实验主要探究了不同的图像特征交互方式对生成图像差异描述结果的影响。在 3 个数据集上分别使用不同的特征交互方式训练基线模型,结果如表 3 所示。

**表 3　图像特征交互方式实验结果**

**Table 3　Experimental results of image feature interaction**

| 数据集 | 特征交互方式 | BLEU-4 | METEOR | ROUGE-L | CIDEr |
|---|---|---|---|---|---|
| Spot-the-diff | 相连 | 0.092 | 0.129 | 0.319 | 0.383 |
| | 相减 | 0.070 | 0.117 | 0.278 | 0.271 |
| | 混合 | 0.090 | 0.130 | 0.327 | 0.366 |
| Image Editing Request | 相连 | 0.042 | 0.129 | 0.355 | 0.181 |
| | 相减 | 0.060 | 0.128 | 0.355 | 0.234 |
| | 混合 | 0.056 | 0.142 | 0.401 | 0.222 |
| Birds-to-Words | 相连 | 0.242 | 0.228 | 0.468 | 0.133 |
| | 相减 | 0.279 | 0.216 | 0.467 | 0.102 |
| | 混合 | 0.230 | 0.229 | 0.468 | 0.156 |

由此可见,对于图像初级特征 $\boldsymbol{E}^1$、$\boldsymbol{E}^2$,分别使用相连 $\boldsymbol{E}^1 \circ \boldsymbol{E}^2$、相减 $\boldsymbol{E}^1 - \boldsymbol{E}^2$ 和混合 $\boldsymbol{E}^1 \circ \boldsymbol{E}^2 \circ (\boldsymbol{E}^1 - \boldsymbol{E}^2)$ 这 3 种特征交互方式,在不同数据集上的表现效果不完全相同。在 Spot-the-diff 数据集上以 CIDEr 作为主指标,使用相连方式的效果最佳;而在 Image Editing Request 数据集和 Birds-to-Words 数据集上分别以 METEOR 和 ROUGE-L 作为主指标,使用混合方式的效果最佳。

出现这一结果的可能原因是:Spot-the-diff 数据集中的图像对之间主要是局部物体级别的差异,基本不存在角度变换等干扰因素。因此,使用相连这种简单的特征交互方式就可以达到很好的效果,复杂的交互方式反而可能影响模型的效果。后 2 个数据集上全局变化和局部变化都更复杂,且存在较多的干扰因素,如 Birds-to-Words 数据集中存在拍摄角度、背景等干扰因素,在这种情况下,单独使用相连或相减的特征交互方式的效果就不如两者结合起来的混合方式好。后者可以为模型提供更丰富的视觉信息,帮助模型在进行图像编码时更好地将变化的区域对应起来,排除掉干扰因素带来的变化,从而更准确地描述出图像间的差异。

之后的实验均在此实验的基础上进行,即在 Spot-the-diff 数据集上使用相连的特征交互方式,在 Image Editing Request 数据集和 Birds-to-Words 数据集上使用混合的特征交互方式。

3.4.2 与 SOTA 对比实验

将本文提出的引入文本信息辅助训练的模型 TA-IDC 和前人提出的 SOTA 模型在 3 个数据集的测试集上的实验结果进行对比,如表 4 所示。

**表 4 与 SOTA 对比实验结果**

**Table 4 Experimental results of comparison with SOTA**

| 数据集 | 模型 | BLEU-4 | METEOR | ROUGE-L | CIDEr |
|---|---|---|---|---|---|
| Spot-the-diff | DDLA-single | 0.085 | 0.120 | 0.286 | 0.328 |
| | DDLA-multi | 0.062 | 0.108 | 0.260 | 0.297 |
| | Relational Speaker | 0.081 | 0.122 | 0.314 | 0.353 |
| | TA-IDC | 0.106 | 0.129 | 0.339 | 0.475 |
| 数据集 | 模型 | BLEU-4 | METEOR | ROUGE-L | CIDEr |
| Image Editing Request | Relational Speaker | 0.067 | 0.128 | 0.373 | 0.264 |
| | TA-IDC | 0.035 | 0.150 | 0.426 | 0.216 |
| 数据集 | 模型 | BLEU-4 | METEOR | ROUGE-L | CIDEr |
| Birds-to-Words | Neural Naturalist | 0.220 | | 0.430 | 0.250 |
| | L2C | 0.318 | | 0.456 | 0.163 |
| | TA-IDC | 0.278 | 0.231 | 0.481 | 0.223 |

可以看到,本文模型生成的图像差异描述,能在 3 个图像差异描述数据集上主要指标分别超越已有模型最佳结果 12%、2% 和 3%,说明本文模型具有较好的准确性和鲁棒性,能为各种具有不同层次差异的图像对均生成更优质量的图像差异描述。

3.4.3 消融实验

本文通过消融实验,对比了基线模型、文本辅助解码模型和进行混合解码模型的实验,结果如表 5 所示。

对比表 5 中第 1、2 行的结果可以看到,使用文本辅助解码的模型在 3 个数据集上生成的描述在主指标上均有提高,其中在 Spot-the-diff 数据集和 Birds-to-Words 数据集上提升较多,而在 Image Editing Request 数据集上提升不明显。本文认为导致这一结果的可能原因是:通过这种方法引入的文本信息没有和图像信息进行直接的交互,主要是通过共享解码器隐式地要求图像和文本生成近似的特征,从而提升描述模型的整体能力。而 Image Editing Request 数据集规模较小,图像差异类型和描述语义比较丰富,模型在面对这样比较复杂的数据集时,实际提升效果就可能低于预期。

对比表 5 中第 1 行和第 3、4、5 行结果可以看到,引入文本信息并进行混合解码,在 3 个数据集上的主指标都有了较明显的提升。这说明本文提出的混合解码的方法,确实可以更好地建模跨模态之间的语义关联,有效提升模型生成的图像差异描述质量。

**表 5 消融实验结果**

**Table 5 Experimental results of Ablation study**

| 模型 | Spot-the-diff | | | | Image Editing Request | | | | Birds-to-Words | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | BLEU-4 | METEOR | ROUGE-L | CIDEr | BLEU-4 | METEOR | ROUGE-L | CIDEr | BLEU-4 | METEOR | ROUGE-L | CIDEr |
| 基线 | 0.092 | 0.129 | 0.319 | 0.383 | 0.056 | 0.142 | 0.401 | 0.222 | 0.230 | 0.229 | 0.468 | 0.156 |
| TA-IDC (T/T) | 0.105 | 0.131 | 0.327 | 0.416 | 0.047 | 0.147 | 0.418 | 0.221 | 0.245 | 0.224 | 0.472 | 0.112 |
| TA-IDC (I/T) | 0.110 | 0.133 | 0.335 | 0.425 | 0.035 | 0.150 | 0.426 | 0.216 | 0.260 | 0.228 | 0.477 | 0.172 |
| TA-IDC (T/I) | 0.106 | 0.129 | 0.339 | 0.475 | 0.070 | 0.145 | 0.376 | 0.216 | 0.278 | 0.231 | 0.481 | 0.223 |
| TA-IDC (I/T + T/I) | 0.102 | 0.123 | 0.326 | 0.430 | 0.042 | 0.149 | 0.413 | 0.208 | 0.253 | 0.219 | 0.470 | 0.147 |

## 3.5 实验结果可视化

为了更清楚地展现模型结果,本节分别选取了 3 个数据集中的部分例子进行展示,如图 6 所示。图中的模型生成描述为本文提出的最优模型生成的描述。描述文本中,标蓝的部分是模型正确描述的图像间差异,标绿的部分是人工标注中提到而模型没有描述到的图像间的差异,标红的部分是模型描述的差异在图像间不存在或有错误。

由图 6 可较为直观地了解本文模型在 3 个数据集上图像差异描述的实际生成情况。综合来看,在 Spot-the-diff 数据集和 Image Editing Request 数据集上,模型生成的差异描述的准确性相对更高,且模型能关注到图像上的细节,描述的语法结构也近似人工标注;对 Birds-to-Words 数据集,因为图像对之间的差异更复杂、更细粒度,需要描述的不同点也更多,所以生成的描述虽然在表达上也很接近人工标注,但准确性相对本文差。

值得注意的是,本文在分析实验结果时发现,生成的图像差异描述依然存在一些问题。例如,描述的差异变化前后顺序颠倒;模型可能会为了


图 6　三个数据集上部分实例展示

Fig. 6　Partial examples demonstration on three datasets

描述差异而凭空想象 2 张图像之间存在一些差异,但事实上并不存在这样的差异等。未来可以针对这些问题对模型继续进行改进和提升。

## 4　结　论

1）本文针对图像差异描述任务,提出了一种在训练阶段引入文本信息进行辅助训练的模型框架 TA-IDC,在以编码器-解码器为主要架构的模型之上,增加一个文本的编码器,通过引入文本辅助解码和混合解码机制,将图像信息和文本信息进行了更充分的特征融合,从而更好地建模视觉和文本 2 种跨模态信息之间的语义关联。

2）实验证明,本文模型可以在多个不同差异类型的数据集上,生成比已有模型更高质量的图像差异描述,具有较好的鲁棒性和有效性。

未来可针对模型生成结果中存在的一些问题继续改进,并进一步探索如何将该模型用于细粒度的图像理解问题,如用于辅助一般图像描述模型,生成更细粒度的单张图像描述。

### 参考文献（References）

[1] JHAMTANI H, BERG-KIRKPATRICK T. Learning to describe differences between pairs of similar images[C]// Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, 2018: 4024-4034.

[2] TAN H, DERNONCOURT F, LIN Z. Expressing visual relationships via language[C]// Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019: 1873-1883.

[3] FORBES M, KAESER-CHEN C, SHARMA P, et al. Neural naturalist: Generating fine-grained image comparisons[C]// Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, 2019: 708-717.

[4] 苗益,赵增顺,杨雨露,等. 图像描述技术综述[J]. 计算机科学,2020,47(12):149-160.
MIAO Y, ZHAO Z S, YANG Y L, et al. Survey of image captioning methods[J]. Computer Science, 2020, 47(12): 149-160 (in Chinese).

[5] HOSSAIN M Z, SOHEL F, SHIRATUDDIN M F, et al. A comprehensive survey of deep learning for image captioning[J]. ACM Computing Surveys, 2019, 51(6): 1-36.

[6] VINYALS O, TOSHEV A, BENGIO S, et al. Show and tell: A neural image caption generator[C]// Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2015: 3156-3164.

[7] JIA X, GAVVES E, FERNANDO B, et al. Guiding the long-short term memory model for image caption generation[C]// Proceedings of 2016 IEEE International Conference on Computer Vision. Piscataway: IEEE Press, 2016: 2407-2415.

[8] WU J, CHEN T, WU H, et al. Fine-grained image captioning with global-local discriminative objective[J]. IEEE Transactions on Multimedia, 2020, 23: 2413-2427.

[9] XU K, BA J, KIROS R, et al. Show, attend and tell: Neural image caption generation with visual attention[C]// Proceedings of the 32nd International Conference on Machine Learning. New York: ACM, 2015: 2048-2057.

[10] LU J, XIONG C, PARIKH D, et al. Knowing when to look: Adaptive attention via a visual sentinel for image captioning[C]// Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2017: 3242-3250.

[11] MING J, HUANG S, DUAN J, et al. SALICON: Saliency in context[C]// Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2015: 1072-1080.

[12] PEDERSOLI M, LUCAS T, SCHMID C, et al. Areas of attention for image captioning[C]// Proceedings of 2017 IEEE Interna-

tional Conference on Computer Vision. Piscataway: IEEE Press, 2017: 1251-1259.

[13] YAN A, WANG X, FU T, et al. L2C: Describing visual differences needs semantic understanding of individuals[C]// Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics, 2021: 2315-2320.

[14] CHO K, MERRIËNBOER 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, 2014: 1724-1734.

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

[16] MA S, SUN X, LIN J, et al. Autoencoder as assistant supervisor: Improving text representation for Chinese social media text summarization[C]// Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, 2018: 725-731.

[17] PAPINENI K, ROUKOS S, WARD T, et al. BLEU: A method for automatic evaluation of machine translation[C]// Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, 2002: 311-318.

[18] LIN C Y. ROUGE: A package for automatic evaluation of summaries[C]// Proceedings of the Workshop on Text Summarization Branches Out, 2004: 74-81.

[19] BANERJEE S, LAVIE A. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments[C]// Proceedings of the ACL workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization, 2005: 65-72.

[20] VEDANTAM R, ZITNICK C L, PARIKH D. CIDEr: Consensus-based image description evaluation[C]// Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2015: 4566-4575.

[21] DENG J, DONG W, SOCHER R, et al. ImageNet: A large-scale hierarchical image database// Proceedings of 2009 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2009: 248-255.

# Image difference caption generation with text information assistance

CHEN Weijing, WANG Weiying, JIN Qin*

(School of Information, Renmin University of China, Beijing 100872, China)

**Abstract**: The image captioning task requires the machine to automatically generate natural language text to describe the semantic content of the image, thus transforming visual information into textual descriptions that facilitate image management, retrieval, classification, and other tasks. Image difference captioning is an extension of the image captioning task, which requires generating natural language sentences to describe the differences between two similar images. The difficulty of this task is how to determine the visual semantic difference between two images and convert the visual difference information into the corresponding textual descriptions. Previous studies do not make full use of textual information in the training stage to model cross-modal semantic associations between visual difference information and text. In this regard, the proposed framework named TA-IDC uses textual information to assist training. It adopts a multi-task learning method, adding a text encoder to the encoder-decoder structure and introducing textual information by text-assisted decoding and mixed decoding during the training stage. This aids in the modeling of semantic relationships between visual and text modalities, resulting in more accurate picture difference captions. Experimentally, TA-IDC outperforms the best results of existing models on main metricsby 12%, 2%, and 3% on three image difference caption datasets, respectively.

**Keywords**: image difference captioning; modal fusion; image captioning; computer vision; natural language processing

**Received**: 2021-09-06; **Accepted**: 2021-09-17; **Published online**: 2021-10-15 14:24

**URL**: kns.cnki.net/kcms/detail/11.2625.V.20211015.0739.001.html

**Foundation items**: National Natural Science Foundation of China (61772535, 62072462); Beijing Natural Science Foundation (4192028)

* **Corresponding author**. E-mail: qjin@ruc.edu.cn

http://bhxb.buaa.edu.cn jbuaa@buaa.edu.cn
**DOI**: 10.13700/j.bh.1001-5965.2021.0537

# 一种傅里叶域海量数据高速谱聚类方法

张熳, 徐兆瑞, 沈项军*

(江苏大学 计算机科学与通信工程学院, 镇江 212013)

**摘 要**: 谱聚类方法广泛应用于数据挖掘和模式识别等领域,但大规模数据上高计算代价的特征向量求解及大数据带来的巨大内存需求,使得其应用于大规模数据时受到了极大的限制。为此,研究了基于傅里叶域的海量数据高速谱聚类方法。利用数据模式的重复性特点在傅里叶域建模,将耗时的特征向量计算转化为对预先确定的傅里叶域判别基进行选择来确定最终的特征向量,计算过程只需进行简单的乘法和加法运算,计算量得到极大的约减;分批次训练样本,使用部分样本即可估计出整体数据的特征向量分布,确定最终的特征向量,压缩了计算时间和内存需求。在 Ijcnn1、RCV1、Covtype-mult、Poker 及 MNIST-8M 等大规模数据上的实验结果表明,所提方法在聚类精度等各项指标基本保持的前提下,训练时间相比 FastESC、LSSHC、SC_RB、SSEIGS 及 USPEC 等方法最高快了 810.58 倍,证明了所提方法在处理大规模聚类数据方面具有显著优势。

**关 键 词**: 谱聚类; 傅里叶域; 海量数据; 高速计算; 低内存需求

**中图分类号**: TP37

**文献标志码**: A **文章编号**: 1001-5965(2022)08-1445-10

聚类作为机器学习的基础研究方法,可以应用于数据挖掘[1-2]、模式识别[3-4]等领域。而谱聚类作为聚类方法的一种,能够描述具有高度复杂非线性结构的真实数据,使得其在处理非线性可分数据集方面取得了良好的表现,受到了学术界越来越多的关注。例如,谢娟英等[4]提出了基于谱聚类的无监督特征选择方法,用于处理具有高维小样本特点且包含大量与疾病无关的基因样本数据。朱光辉等[5]基于 Apache Spark 分布式并行计算框架研究并实现了大规模并行谱聚类算法 SCoS,成功减少了谱聚类算法的时间和内存开销。李玉等[6]提出了一种可变类谱聚类算法,实现了遥感影像分割中类别数的准确、自动判别。

然而,由于谱聚类需要在全体样本数据上构造拉普拉斯图[7],并根据构造的拉普拉斯图得到数据的特征向量,在样本数据增大后,其构造拉普拉斯矩阵所需的空间急剧增长,且特征向量的求解特点同样导致了计算量的巨大增长,阻碍了谱聚类方法应用于大规模数据。近年来,众多学者和研究人员对大规模谱聚类方法的优化和改进做了大量的研究与努力。多数改进方法针对大规模矩阵进行近似,其中较有代表性的是 Nyström 方法[8-9],该方法通过局部数据的特征向量推导出整个数据集的特征向量的近似值。例如,Du 和 Tsang[10]将 Nyström 方法应用于海洋盐度的高精度计算。薛丽霞等[11]利用 Nyström 方法提出了一种基于密度峰值优化的谱聚类算法,能够自动确定聚类数目,同时解决了谱聚类算法处理海量数据效率低的问题。Bouneffouf[12]提出了一种基于 Nyström 的可扩展聚类算法。相较于 Nyström 方

**收稿日期**: 2021-09-08; **录用日期**: 2021-10-17; **网络出版时间**: 2022-05-19 13:20
**网络出版地址**: kns.cnki.net/kcms/detail/11.2625.V.20220518.1910.002.html
**基金项目**: 国家自然科学基金 (61572240)
* **通信作者**. E-mail: xjshen@ujs.edu.cn

**引用格式**: *张熳, 徐兆瑞, 沈项军. 一种傅里叶域海量数据高速谱聚类方法[J]. 北京航空航天大学学报, 2022, 48(8): 1445-1454.*
*ZHANG M, XU Z R, SHEN X J. A high-speed spectral clustering method in Fourier domain for massive data[J]. Journal of Beijing University of Aeronautics and Astronautics, 2022, 48(8): 1445-1454 (in Chinese).*

法,Wu 等[13]提出的基于随机面元特征的谱聚类(spectral clustering random binning, SC_RB)算法使用随机面元特征生成大型稀疏特征矩阵的内积,隐式地近似图相似度矩阵。上述方法有效解决了内存不足的问题。为了进一步提升谱聚类的准确率,Yang 等[14]引入了超图的概念,超图可以描述数据间更为复杂的相互关系,从而大大提升了聚类的准确率,但同时也增加了算法的时间负担。

在近期的研究中,傅里叶方法被广泛应用于提高算法的运算效率和降低存储量级。受文献[15-17]启发,本文提出了一种基于快速傅里叶变换的谱聚类(spectral clustering via fast Fourier transform, SC-FFT)方法。该方法利用聚类数据具有模式重复性的特点,通过傅里叶方法对大规模数据进行建模。本文模型的显著优点是训练速度得到极大提升,这是由于傅里叶域的谱特性,本文证明谱聚类方法的特征向量求解可以通过预先设定的傅里叶基来获得。因此,传统谱聚类中复杂的特征向量计算,可以转化为在傅里叶域中寻找一定数量的判别基;在傅里叶域中,复杂的大规模矩阵特征向量计算可以被傅里叶域矩阵上的简单加法和乘法运算所描述,使得本文模型的训练速度得到极大提升。本文方法的另一个显著优点是使用的样本急剧减少。这是由于傅里叶域谱特征的计算实际是多个样本的傅里叶域特征的累加,这种特性使得本文方法仅需采用部分样本就可以估计出全部数据的特征向量。因此,本文模型所需的内存空间得到了极大的压缩,同时由于只计算了部分样本的傅里叶域谱特征,模型的训练速度得到了进一步提升。综上,本文方法训练过程只需将少量的训练样本分批次输入,直到得出稳定的傅里叶判别基即可,从而最大程度地提升了谱聚类方法在大规模数据上的实用性。因此,本文方法可以在较快的运算速度、较低的内存需求下,对大规模数据进行聚类。在 Ijcnn1、RCV1、Covtype-mult、Poker 及 MNIST-8M 等多个大规模数据上的实验结果表明,本文方法的训练时间比现有大规模谱聚类方法最高快了 810.58 倍,聚类的准确率、召回率等指标能够达到现有大规模谱聚类方法的水平。

## 1 相关工作

### 1.1 大规模谱聚类

在众多的大规模谱聚类方法中,代表性的方法是使用 Nyström 方法[8-9]来降低求解特征向量的计算成本。其主要思想为:给定数据矩阵,利用 Nyström 方法对数据进行随机采样,并计算子矩阵的特征向量,用计算出的特征向量估计原始大规模矩阵的特征向量。例如,Fowlkes 等[8]将 Nyström 方法应用到排列组合的分组问题中,基于真实场景中目标所占像素远少于全部像素这一事实,与 Nyström 方法随机抽样数据相符合,成功解决了大规模数据上的分组问题。Li 等[18]提出了 FastESC 算法,使用随机傅里叶特征显式表示核空间中的数据,通过构造远小于数据规模的核矩阵来解决内存不足的问题,算法的运行速度明显加快。

另一类大规模谱聚类方法是基于特征映射[19-22]的方法。例如,Hansen 和 Mahoney[23]提出了 SSEIGS 算法,该算法是基于特征向量不适用于对数据局部特征感兴趣的情况而提出的,其提供了一种构造拉普拉斯半监督特征向量的方法来解决该问题,使用局部偏置特征向量执行局部偏置机器学习,克服了谱聚类应用于大规模数据的难题。Wu 等[13]提出了 SC_RB 算法,使用随机面元特征生成大型稀疏特征矩阵的内积,隐式近似图相似度矩阵,并引入了奇异值分解求解器用于计算大规模矩阵的特征向量,加速相似图的构造和特征分解。Rahman 和 Bouguila[19]将基于直方图的特征和基于分布的密度特征相结合,开发出了一种灵活的特征映射技术,该技术包含数据的先验知识,为其提供了灵活的表示,提高了分类器的识别能力。Vázquez-Martín 和 Bandera[20]提出了一种关键帧检测器,通过建立基于局部特征性的辅助映射,利用谱聚类方法得到图划分,该方法不需要估计摄像机的运动,且在该检测器内定义的相似性度量可用于任何类型的特征。万月等[21]提出了基于稀疏自编码的局部谱聚类映射算法,通过对数据进行预处理,利用稀疏自编码提取能反映原始数据本质的深层次特征,以此代替原始数据,对每个数据利用其邻域进行线性重构,以重构权值代替高斯核函数建立邻接矩阵。

此外,还有不基于以上 2 种代表性方法的大规模谱聚类方法。例如,Huang 等[24]提出了 USPEC 算法,采取混合的代表选择策略,设计了 k-最接近代表的快速逼近方法,用于构造稀疏亲和子矩阵,将稀疏子矩阵解释为二部图,利用转移切割对图进行有效划分,从而得到聚类结果。Yang 等[14]将超图的概念运用到谱聚类方法的改进中,提出了 GraphLSHC 算法,将超图扩展为通用的格式用于捕获复杂的高阶关系,同时提出了“特征

技巧”的概念用于降低计算复杂度以加速特征问题的求解过程,该算法通过只存储超图实例矩阵,而不存储拉普拉斯矩阵来克服大规模数据上内存不足的问题。

### 1.2　基于傅里叶域的相关应用

近期的研究[15-17]表明,可以通过循环矩阵转换数据样本到傅里叶域,利用离散傅里叶变换(DFT)[25-26]对数据矩阵进行近似对角化,该方法极大提升了计算效率。例如,Henriques 等[16]证明了具有模式重复性的大规模数据矩阵是可循环的,并推导了一种基于傅里叶变换的数据转换,该转换可以将数据矩阵分块对角化,从而将存储量和计算量减少一个数量级。该方法可以解决很多经典的机器学习算法在大规模数据上的应用问题。Gao[27]在文献[15]的基础上提出了数据的增量聚类概念,即每次有新数据到达时无需重新开始聚类,将计算复杂度限制为只考虑新的数据,大大降低了算法的空间复杂度,打破了大规模数据聚类内存不足的限制。Bibi 等[28]提出了一种新的套索循环重组方案,即通过在傅里叶域中优化原来的套索对偶形式,将问题提升到更高的维度,许多大型线性系统的求解便可以由一维数据的快速傅里叶变换及元素向量积操作替代,大大提升了算法的运算效率。

## 2　本文方法

### 2.1　傅里叶域谱聚类的性质分析

由于数据具有模式上的重复性[29],可以对单一数据进行循环建模。对于单一数据元素 $\boldsymbol{x}_i=[x_{i1},x_{i2},\cdots,x_{im}]$,$m$ 为特征维度,其对应的循环矩阵为

$$\boldsymbol{X}=\begin{bmatrix} x_{i1} & x_{i2} & \cdots & x_{im} \\ x_{im} & x_{i1} & \cdots & x_{i(m-1)} \\ \vdots & \vdots & & \vdots \\ x_{i2} & x_{i3} & \cdots & x_{i1} \end{bmatrix} \tag{1}$$

循环矩阵具有重要的数学性质,如 $\boldsymbol{X}$ 可由 $\boldsymbol{x}_i$ 经过离散傅里叶变换生成的对角矩阵进行表示,即

$$\boldsymbol{X}=\boldsymbol{F}\mathrm{diag}(\hat{\boldsymbol{x}})\boldsymbol{F}^{\mathrm{H}} \tag{2}$$

式中:$\boldsymbol{F}$ 为傅里叶矩阵,在 $\boldsymbol{X}$ 维度确定的情况下,可以将其视作一个常量矩阵;$\hat{\boldsymbol{x}}$ 为 $\boldsymbol{x}$ 经过离散傅里叶变换后的数据;$\mathrm{diag}(\hat{\boldsymbol{x}})$ 表示对角矩阵,其对角元素依次对应 $\hat{\boldsymbol{x}}$ 中的样本元素。

一个 $m\times m$ 维的傅里叶矩阵表述如下:

$$\boldsymbol{F}=\begin{bmatrix} 1 & 1 & 1 & \cdots & 1 \\ 1 & \omega & \omega^2 & \cdots & \omega^{m-1} \\ 1 & \omega^2 & \omega^4 & \cdots & \omega^{2(m-1)} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & \omega^{m-1} & \omega^{2(m-1)} & \cdots & \omega^{(m-1)^2} \end{bmatrix} \tag{3}$$

式中:$\omega=\mathrm{e}^{2\pi m/i}$。傅里叶矩阵具有 $\boldsymbol{F}\boldsymbol{F}^{\mathrm{H}}=\boldsymbol{I}$ 的性质,$\boldsymbol{F}^{\mathrm{H}}$ 为傅里叶矩阵 $\boldsymbol{F}$ 的共轭转置矩阵。

**定理 1**　谱聚类方法的特征向量可以转换为傅里叶域基的一个真子集。

**证明**　对应到每一个具体的样本,一般的谱聚类方法模型可以表示为

$$\begin{cases}\mathrm{argmin}\sum \boldsymbol{v}^{\mathrm{T}}\boldsymbol{x}_i L_{ij}\boldsymbol{x}_j^{\mathrm{T}} \\ \text{s.t.}\quad \sum \boldsymbol{v}^{\mathrm{T}}\boldsymbol{x}_i D_{ii}\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{v}=1\end{cases} \tag{4}$$

式中:$L_{ij}$ 为拉普拉斯矩阵第 $i$ 行第 $j$ 列的元素值;$D_{ii}$ 为度矩阵主对角线上的第 $i$ 个元素;$\boldsymbol{v}$ 为最终要求取的特征向量。

将以上建模方法中的数据转换为循环数据表达,则相应的谱聚类方法模型可以描述为

$$\begin{cases}\mathrm{argmin}\sum \boldsymbol{v}^{\mathrm{T}}\boldsymbol{f}(\boldsymbol{x}_i)L_{ij}\boldsymbol{f}^{\mathrm{T}}(\boldsymbol{x}_j) \\ \text{s.t.}\quad \sum \boldsymbol{v}^{\mathrm{T}}\boldsymbol{f}(\boldsymbol{x}_i)D_{ii}\boldsymbol{f}^{\mathrm{T}}(\boldsymbol{x}_i)\boldsymbol{v}=1\end{cases} \tag{5}$$

式中:$\boldsymbol{f}(\boldsymbol{x}_i)$ 为第 $i$ 个样本构成的循环矩阵。

根据式(5)构造拉格朗日函数,并求解其特征向量 $\boldsymbol{v}$,得到

$$\sum \boldsymbol{f}(\boldsymbol{x}_i)L_{ij}\boldsymbol{f}^{\mathrm{T}}(\boldsymbol{x}_j)\boldsymbol{v}=\sum \boldsymbol{\Lambda}\boldsymbol{f}(\boldsymbol{x}_i)D_{ii}\boldsymbol{f}^{\mathrm{T}}(\boldsymbol{x}_i)\boldsymbol{v} \tag{6}$$

式中:$\boldsymbol{\Lambda}$ 为特征值矩阵。根据循环矩阵的重要数学性质 $\boldsymbol{f}(\boldsymbol{x}_i)=\boldsymbol{F}\mathrm{diag}(\hat{\boldsymbol{x}}_i)\boldsymbol{F}^{\mathrm{H}}$,代入到式(6)中可以得到

$$\boldsymbol{F}\sum \mathrm{diag}(\hat{\boldsymbol{x}}_i\odot\hat{\boldsymbol{x}}_i^{*}\odot D_{ii})^{-1}\cdot \sum \mathrm{diag}(\hat{\boldsymbol{x}}_i\odot\hat{\boldsymbol{x}}_j^{*}\odot L_{ij})\boldsymbol{F}^{\mathrm{H}}\hat{\boldsymbol{v}}=\boldsymbol{\Lambda}\hat{\boldsymbol{v}} \tag{7}$$

根据 $\boldsymbol{F}\boldsymbol{F}^{\mathrm{H}}=\boldsymbol{I}$ 的性质,可以对式(7)进一步化简,得到

$$\sum \mathrm{diag}(\hat{\boldsymbol{x}}_i\odot\hat{\boldsymbol{x}}_i^{*}\odot D_{ii})^{-1}\cdot \sum \mathrm{diag}(\hat{\boldsymbol{x}}_i\odot\hat{\boldsymbol{x}}_j^{*}\odot L_{ij})\boldsymbol{F}^{\mathrm{H}}\hat{\boldsymbol{v}}=\boldsymbol{\Lambda}\boldsymbol{F}^{\mathrm{H}}\hat{\boldsymbol{v}} \tag{8}$$

不失一般性,此时可令 $\hat{\boldsymbol{v}}=\boldsymbol{F}$,则对应的特征值为

$$\boldsymbol{\Lambda}=\sum \frac{1}{\mathrm{diag}(\hat{\boldsymbol{x}}_i\odot\hat{\boldsymbol{x}}_i^{*}\odot D_{ii})}\sum \mathrm{diag}(\hat{\boldsymbol{x}}_i\odot\hat{\boldsymbol{x}}_j^{*}\odot L_{ij}) \tag{9}$$

**证毕**

通过以上证明可得,若对原始数据进行循环建模,傅里叶矩阵可以作为谱聚类方法中一组预先确定的特征向量判别基。若对原始数据进行循环傅里叶建模,则会由于频域的变化产生一定的

误差,这一点 Henriques 等[16]已经给出了详尽的证明。此外,式(9)中的 $\mathrm{diag}(\hat{\boldsymbol{x}}_i \odot \hat{\boldsymbol{x}}_i^* \odot \boldsymbol{D}_{ii})$ 和 $\mathrm{diag}(\hat{\boldsymbol{x}}_i \odot \hat{\boldsymbol{x}}_j^* \odot \boldsymbol{L}_{ij})$ 非常容易通过计算得到,这是因为运算对象为对角矩阵的点积运算,只涉及简单的算数乘法操作。

综上,传统的谱聚类方法计算并选取前 $k$ 小个特征值所对应的特征向量,该运算过程包含了费时的矩阵特征向量求解。而本文定理 1 得出只需将数据转换到傅里叶域,并根据数据在傅里叶域的累加和乘法计算,确定前 $k$ 小个特征值,并输出其所对应的傅里叶基即可。因此,该方法避免了复杂的矩阵特征向量求解过程,极大提高了海量数据谱聚类训练时特征向量的求解速度;此外,由于傅里叶域特征向量求解的特点,只需要储存对角化特征值矩阵,以及预先设定的傅里叶基,不需要储存所有样本的拉普拉斯图矩阵,其需要的内存也得到了极大的压缩。

**定理 2** 傅里叶域内部分样本及其特征向量判别基可近似估计出整体数据样本的分布。

**证明** 对于全部样本数据矩阵 $\boldsymbol{X} \in \mathbf{R}^{m \times n}$,存在正交矩阵 $\boldsymbol{U} = [\boldsymbol{u}^1, \boldsymbol{u}^2, \cdots, \boldsymbol{u}^m]$ 和 $\boldsymbol{V} = [\boldsymbol{v}^1, \boldsymbol{v}^2, \cdots, \boldsymbol{v}^n]$,使得

$$\boldsymbol{U}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{V} = \boldsymbol{\Sigma} = \mathrm{diag}(\sigma_1, \sigma_2, \cdots, \sigma_\rho) \tag{10}$$

即 $\boldsymbol{X} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^{\mathrm{T}}$,其中,$\boldsymbol{u}^i$ 和 $\boldsymbol{v}^i$ 分别为矩阵 $\boldsymbol{U}$ 和矩阵 $\boldsymbol{V}$ 的第 $i$ 列向量,$\rho = \min\{m, n\}$,且 $\sigma_1 \geqslant \sigma_2 \geqslant \cdots \geqslant \sigma_\rho \geqslant 0$。同时,$\boldsymbol{U}$ 中的列向量是 $\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}$ 的特征向量,$\boldsymbol{V}$ 中的列向量是 $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$ 的特征向量。

若 $\boldsymbol{X}$ 的秩为 $r$,对于 $k < r$ 且 $k \in \boldsymbol{Z}^+$,有 $\boldsymbol{X}_k = \boldsymbol{U}_k\boldsymbol{\Sigma}_k\boldsymbol{V}_k^{\mathrm{T}} = \sum_{t=1}^{k} \sigma_t \boldsymbol{u}^t \boldsymbol{v}^{t\mathrm{T}}$ 为 $\boldsymbol{X}$ 的 $k$ 秩近似,其中,$\boldsymbol{U}_k$ 和 $\boldsymbol{V}_k$ 分别为矩阵 $\boldsymbol{U}$ 和矩阵 $\boldsymbol{V}$ 的前 $k$ 列,$\boldsymbol{\Sigma}_k$ 为由 $\boldsymbol{X}$ 的前 $k$ 个奇异值组成的对角矩阵。

若用 $\boldsymbol{X}_k = \boldsymbol{U}_k\boldsymbol{U}_k^{\mathrm{T}}\boldsymbol{X}$ 近似 $\boldsymbol{X}$,会产生 $\boldsymbol{X}$ 与 $\boldsymbol{X}_k$ 之间的误差 $\|\boldsymbol{X} - \boldsymbol{X}_k\|_F^2$。此外,还有 $\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}$ 与 $\boldsymbol{X}_B\boldsymbol{X}_B^{\mathrm{T}}$ 之间的误差 $\|\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}} - \boldsymbol{X}_B\boldsymbol{X}_B^{\mathrm{T}}\|_F$,其中,$\boldsymbol{X}_B$ 为从全局样本 $\boldsymbol{X}$ 中随机选取 $B$ 个样本组成的局部样本数据矩阵。根据文献[30],对于每一个正整数 $k$($k \leqslant \mathrm{rank}(\boldsymbol{X}_B)$)有

$$\|\boldsymbol{X} - \boldsymbol{H}_k\boldsymbol{H}_k^{\mathrm{T}}\boldsymbol{X}\|_F^2 \leqslant \|\boldsymbol{X} - \boldsymbol{X}_k\|_F^2 + 2\sqrt{k}\|\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}} - \boldsymbol{X}_B\boldsymbol{X}_B^{\mathrm{T}}\|_F \tag{11}$$

式中:$\boldsymbol{H}_k$ 为 $\boldsymbol{X}$ 的前 $k$ 个特征值对应的特征向量组成的矩阵,且 $\boldsymbol{H}_k^{\mathrm{T}}\boldsymbol{H}_k = \boldsymbol{I}_k$。

根据文献[31]可知,$\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}$ 与 $\boldsymbol{X}_B\boldsymbol{X}_B^{\mathrm{T}}$ 之间的误差 $\|\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}} - \boldsymbol{X}_B\boldsymbol{X}_B^{\mathrm{T}}\|_F$ 存在如下关系:

$$E(\|\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}} - \boldsymbol{X}_B\boldsymbol{X}_B^{\mathrm{T}}\|_F) \leqslant \frac{1}{\sqrt{c}}\|\boldsymbol{X}\|_F^2 \tag{12}$$

式中:$c$ 为可调节的超参数。假设数据服从正态分布,即 $\delta \in (0,1)$,同时令 $\eta = 1 + \sqrt{8\log_2(1/\delta)}$,则在概率至少为 $1-\delta$ 的情况下有

$$\|\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}} - \boldsymbol{X}_B\boldsymbol{X}_B^{\mathrm{T}}\|_F \leqslant \frac{\eta}{\sqrt{c}}\|\boldsymbol{X}\|_F^2 \tag{13}$$

根据定理 1 可知,$\boldsymbol{X}_B$ 的特征向量为傅里叶域预先确定的判别基 $\boldsymbol{F}_k$,结合式(10)与式(12),傅里叶域的误差可以描述为

$$\|\hat{\boldsymbol{X}} - \boldsymbol{F}_k\boldsymbol{F}_k^{\mathrm{T}}\hat{\boldsymbol{X}}\|_F^2 \leqslant \|\boldsymbol{X} - \boldsymbol{X}_k\|_F^2 + 2\frac{\eta\sqrt{k}}{c}\|\boldsymbol{X}\|_F^2 \tag{14}$$

式中:$\hat{\boldsymbol{X}} = \boldsymbol{F}\boldsymbol{X}$ 为原始数据在傅里叶域的投影。

**证毕**

由定理 2 可知,整体数据集合 $\boldsymbol{X}$ 可以由局部样本数据 $\boldsymbol{X}_B$ 及其上的部分傅里叶域特征向量判别基 $\boldsymbol{F}_k$ 近似表示,近似误差取决于采样数 $B$ 和数据在傅里叶域用到的判别基个数 $k$。因此,在误差可接受的范围内,用部分样本代替全部样本是可行的。同时,由于傅里叶域内的计算特性,可以在仅使用部分样本的基础上,进一步通过分批次训练对谱聚类方法进行优化,在很大程度上减小了谱聚类方法的运行时间和所需的内存空间,使得海量数据的谱聚类得以在低内存的普通机器上快速实现。

## 2.2 傅里叶域谱聚类方法

基于定理 1 和定理 2 的分析,本文针对海量数据的谱聚类方法,通过随机采样部分样本,计算其傅里叶域特征向量判别基得到所求的谱聚类特征向量。为了得到足够随机样本的傅里叶域特征向量的稳定输出,本文所提谱聚类方法步骤如下:

**步骤 1** 将规模为 $n$ 的全体数据划分为 $P = \lceil n/b \rceil$ 个批次,$\lceil \cdot \rceil$ 为向上取整符号,且 $b \ll n$。标记每个批次的数据为 $\boldsymbol{X}_p \in \mathbf{R}^{m \times b}$,其中 $p = 1, \cdots, P$,并在傅里叶域,依据 $\hat{\boldsymbol{X}}_p = \boldsymbol{F}\boldsymbol{X}_p$ 对每个批次的数据进行傅里叶的数据转换。

**步骤 2** 根据式(9)计算 $\hat{\boldsymbol{X}}_p$ 的特征值,并与之前批次累加的特征值累加,得到相应的特征向量集合 $\boldsymbol{F}_k$。

根据式(9)可知,每个批次对应的特征值为

$$\boldsymbol{\Lambda}_p = \mathrm{diag}\left(\sum_{i=b(p-1)}^{bp} \frac{1}{\hat{\boldsymbol{x}}_i \odot \hat{\boldsymbol{x}}_i^* \odot \boldsymbol{D}_{ii}} \sum_{j=b(p-1)}^{bp} \hat{\boldsymbol{x}}_i \odot \hat{\boldsymbol{x}}_j^* \odot \boldsymbol{L}_{ij}\right) \tag{15}$$

利用傅里叶域内数据可累加的特点,用每个批次所求得的特征值通过累加操作近似全局特征值。由定理 1 可知,本文方法并不是求解特征值的精确值,而是寻求前 $k$ 小个特征值的分布,继而挑选出其对应的特征向量。因此,在每个批次训

练过后不断更新估计特征值 $\tilde{\boldsymbol{\Lambda}}$,直到其满足近似精度。具体的累加更新规则如下:

$$\tilde{\boldsymbol{\Lambda}}_p = \frac{(p-1)\tilde{\boldsymbol{\Lambda}}_{p-1} + \boldsymbol{\Lambda}_p}{p} \tag{16}$$

式中:$\tilde{\boldsymbol{\Lambda}}_p$ 为前 $p$ 个批次数据特征值的累积。当前$(p-1)$个批次所对应的特征值 $\tilde{\boldsymbol{\Lambda}}_{p-1}$中前 $k$ 小的特征值的分布与前 $p$ 个批次所对应的特征值 $\tilde{\boldsymbol{\Lambda}}_p$ 中前 $k$ 小的特征值的分布一致时,认为模型已经稳定。

**步骤 3**　重复步骤 1 和步骤 2,直到前 $k$ 小的特征值的分布稳定,停止输入批量样本到模型中,模型训练终止。换言之,当前采样数 $B$ 已经足够大,使得经过部分样本拟合得到的全局样本的近似误差已经控制在可接受的范围内。此时,相应的特征向量集合为

$$\boldsymbol{F}_k = \{\boldsymbol{a}_1, \boldsymbol{a}_2, \cdots, \boldsymbol{a}_k\} \tag{17}$$

式中:$\boldsymbol{a}_i(1 \leqslant i \leqslant k)$为第 $i$ 小特征值对应的傅里叶域内预先确定的特征向量。

本文方法过程示意如图 1 所示。图中:$\hat{\boldsymbol{X}}_p = \boldsymbol{F}\boldsymbol{X}_p$ 为原始批次数据在傅里叶域的投影,$\tilde{\boldsymbol{\Lambda}}_p^{(k)}$ 表示 $p$ 次样本输入后前 $k$ 小个特征值的分布。根据本文方法,最终挑选出傅里叶域特征向量。图 1 中黑色的圆圈和矩形分别表示选择出的前 $k$ 小个特征值及其对应的特征向量。

分析本文方法的时间复杂度和空间复杂度。由于每个批次用到 $b$ 个样本,结合其傅里叶域维度 $m$,考虑一个批次的样本数据,其计算拉普拉斯矩阵的时间复杂度为 $O(b^2m)$,将一个样本数据转换到傅里叶域的时间复杂度为 $O(m\log_2 m)$,因此,根据式(15)得到的时间复杂度为 $O(b(m\log_2 m + bm))$,空间复杂度为 $O(b^2m)$。由于本文方法分批次输入到模型中,假设共用了 $p$ 个批次,则总的时间复杂度为 $O(pb^2m)$,空间复杂度仍为 $O(b^2)$。Nyström 方法的时间复杂度为 $O(ngd + g^3 + ng^2)$,其中,$n$ 为总样本数,$g$ 为模型训练用到的样本数,$d$ 为 Nyström 方法对应的输入维度,空间复杂度为 $O(m^2)$。$b \ll g < n$,可见,本文方法在时间和空间复杂度上均优于 Nyström 方法。

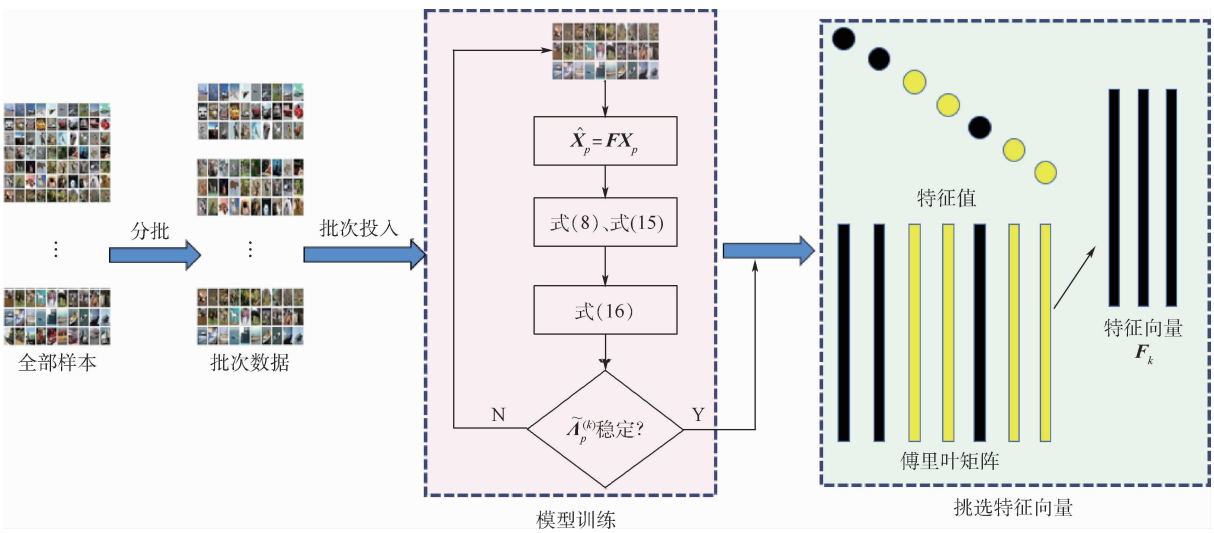



图 1　本文方法流程示意图

Fig. 1　Schematic diagram of the proposed method

## 3　实验结果分析

为了评估本文方法的有效性,在具有代表性的大规模数据集上,将本文方法与其他较优越的大规模谱聚类方法进行了性能比较。实验使用 MATLAB 2020b 软件在 Linux 服务器中进行,服务器版本为 Linux 5.8.0-45-generic,CPU 主频为 2 068 MHz,总内存为 64 GB。

### 3.1　数据集、对比方法和评价标准

表 1 为本文选取的 5 个大规模数据集(https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/)及其基本信息,如下:

1)Ijcnn1。来自 IJCNN 2001 的时间序列的数据集合,包含 22 个特征属性。

2)RCV1。新的文本分类测试集合,一个样本包含 47 236 个特征属性。

3)Covtype-mult。一组根据地图数据预测森林覆盖类型的样本,每个样本具有 54 个特征属性。

4)Poker。一组扑克记录样本,模拟一个人的

**表 1 大规模数据集**

**Table 1 Large-scale datasets**

| 数据集名称 | 样本数 | 特征维度 | 类别数 |
|---|---|---|---|
| Ijcnn1 | 126 701 | 22 | 2 |
| RCV1 | 534 135 | 47 236 | 52 |
| Covtype-mult | 581 012 | 54 | 7 |
| Poker | 1 025 010 | 10 | 10 |
| MNIST-8M | 8 000 000 | 784 | 10 |

2 只手从 52 张标准扑克牌中分别各抽出 5 张,生成一个具有 10 个特征属性的样本。

5) MNIST-8M。通过为每个手写数字生成 134 个随机变形,Loosli 等[32]将 MNIST 数据集从 6 万个扩增到 800 万个。

将本文方法与如下大规模谱聚类方法进行比较:①FastESC[18],基于显示特征映射的快速大规模谱聚类;②LSSHC[14],基于地标表示的大规模谱聚类;③SC_RB[13],利用随机分集特征的可扩展谱聚类;④SSEIGS[23],基于大规模数据的局部偏倚特征向量的谱聚类方法;⑤USPEC[24],超可扩展谱聚类。

为了评价不同方法的聚类结果,采用准确率、召回率、聚类精度、$F_1$、聚类纯度和运行时间 6 个评价指标进行聚类分析。为了增强结果的可靠性,在每一个数据集上,每一种测试方法都运行 10 次,取其均值和方差。

1) 准确率。

$$\text{accuracy} = \frac{\text{TN} + \text{TP}}{\text{FN} + \text{FP} + \text{TN} + \text{TP}} \tag{18}$$

式中:FN 为将 $i$ 类的样本分成 $j$ 类的数目;FP 为将不属于 $i$ 类的分为 $i$ 类的数目;TN 为将不属于 $i$ 类的样本未分类为 $i$ 类的数目;TP 为将属于 $i$ 类的样本分成 $i$ 类的数目。

2) 召回率。

$$\text{recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{19}$$

3) 精度。

$$\text{precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{20}$$

4) $F_1$。

$$F_1 = \frac{2 \times \text{reacall} \times \text{precision}}{\text{recall} + \text{precision}} \tag{21}$$

5) 纯度。

$$\text{purity} = \sum_{i=1}^{K} \frac{m_i}{m} P_i \tag{22}$$

式中:$K$ 为聚类数目;$m_i$ 为聚类 $i$ 中所有成员的个数;$m$ 为聚类划分涉及的成员个数;$P_i = \max(P_{ij})$ 为聚类 $i$ 的纯度;$P_{ij} = m_{ij}/m_i$ 为聚类 $i$ 中成员属于类 $j$ 的概率;$m_{ij}$ 为聚类 $i$ 中成员属于类 $j$ 的个数。

### 3.2 实验结果

为了更好地比较各方法的性能优势,本文统一了各数据集最终的维度。Ijcnn1 数据集原特征维度为 22 维,统一降低到 3 维;RCV1 数据集原特征维度为 47 236 维,统一降低到 400 维;Covtype-mult 数据集原特征维度为 54 维,统一降低到 8 维;Poker 数据集原特征维度为 10 维,统一降低到 2 维;MNIST-8M 数据集原特征维度为 784 维,统一降低到 50 维。

不同方法在本文使用的所有数据集上的运行时间对比如图 2 所示。为了更直观地体现本文方法在运行时间上的优势,将基准线设置在 $10^1$ s 位置。观察图 2,在 Poker、Covtype-mult、RCV1 和 Ijcnn1 数据集上,本文方法的运行时间皆在基准线左侧,说明在这 4 个大规模数据集上,模型训练的时间均不超过 10 s。而在规模最大的 MNIST-8M

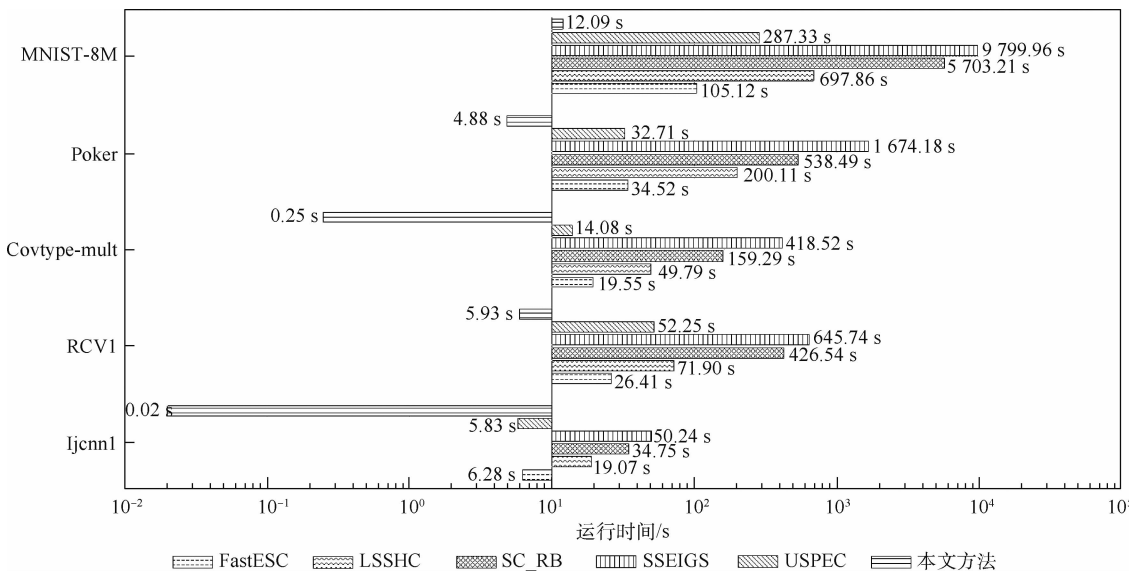



图 2 各方法在不同数据集上的运行时间对比

Fig. 2 Running time comparison of each method on different datasets

数据集上,本文方法仅需 12.09 s 便可完成计算,训练速度是 SC_RB 方法的 471.73 倍、SSEIGS 方法的 810.58 倍。在规模次大的 Poker 数据集上,本文方法仅需 4.88 s 便可以完成谱聚类的训练,而 USPEC 方法需要 32.71 s,FastESC 方法需要 34.52 s,SC_RB 方法需要 538.49 s,SSEIGS 方法需要 1 674.18 s。

综上所述,本文方法在运行时间方面具有显著优势,在大规模数据集上的训练速度提高明显,因此,本文方法可满足海量数据的聚类应用需求。

表 2 和表 3 展示了本文方法与 5 种对比方法在平均准确率和平均聚类纯度指标上的得分情况。由表 2 可知,在 Poker 数据集上,本文方法平均准确率为 0.728 2,相较准确率最低的 FastESC 方法高了 0.257 2。在 MNIST-8M 数据集上,本文方法平均准确率为 0.779 2,比其他方法高出了 0.000 7 ~ 0.219 9。由此可见,本文方法足以保证聚类结果的质量。由表 3 可知,在 MNIST-8M 数据集上,本文方法平均纯度为 0.626 6,比其他方法高出了 0.002 5 ~ 0.154 5。进一步验证了本文方法聚类结果的优越性。

表 4 ~ 表 6 展示了本文方法与对比方法在平均精度、平均召回率和平均 $F_1$ 指标上的得分情况。由表 4 可知,本文方法的平均精度相比 FastESC 方法,在 MNIST-8M 数据集上高出 0.104 5,在 RCV1 数据集高出 0.266 0;相比 LSSHC 方法,在 MNIST-8M 数据集上高出 0.089 4,在 RCV1 数据集上高出 0.360 4。由表 5 可知,在 RCV1 数据集上,本文方法的平均召回率高于其他对比方法 0.004 7 ~ 0.374 7。

由表 6 可知,平均 $F_1$ 指标得分也体现了本文方法的优越性,其在 Ijcnn1、RCV1 和 MNIST-8M 数据集上的平均 $F_1$ 得分均高于其他对比方法。由此可见,本文方法在聚类精度、召回率和 $F_1$ 这 3 个指标上的表现很好。其中,本文方法在 Poker 数据集上的各项指标与其对比方法存在一定的差距,这是由于 Poker 数据集特征维度不高,仅有 10 维,降维效果欠佳。

综上可知,本文方法能够取得与现存大规模谱聚类方法中性能较好的方法相当的聚类结果。

**表 2　不同方法的平均准确率**

**Table 2　Average accuracy of various methods**

| 方法 | Ijcnn1 | RCV1 | Covtype-mult | Poker | MNIST-8M |
|---|---|---|---|---|---|
| FastESC | 0.875 0 ± 0.008 5 | 0.111 0 ± 0 | 0.472 0 ± 0.001 2 | 0.471 0 ± 0.004 3 | 0.559 3 ± 0.021 6 |
| LSSHC | 0.576 9 ± 0.010 4 | 0.207 3 ± 0.000 2 | 0.233 9 ± 0.000 1 | 0.666 7 ± 0.014 7 | 0.778 5 ± 0.010 6 |
| SC_RB | 0.900 4 ± 0 | 0.300 1 ± 0.005 3 | 0.443 3 ± 0.006 7 | 0.677 9 ± 0.024 8 | 0.600 7 ± 0.017 5 |
| SSEIGS | 0.869 1 ± 0 | 0.299 9 ± 0 | 0.338 8 ± 0.000 2 | 0.703 1 ± 0.034 5 | 0.599 0 ± 0.002 1 |
| USPEC | 0.890 3 ± 0.000 7 | 0.246 5 ± 0.019 2 | 0.417 6 ± 0.016 2 | 0.722 8 ± 0.020 1 | 0.743 1 ± 0.010 1 |
| 本文方法 | 0.902 4 ± 0 | 0.338 4 ± 0.001 5 | 0.487 6 ± 0 | 0.728 2 ± 0.010 4 | 0.779 2 ± 0.020 6 |

**表 3　不同方法的平均纯度**

**Table 3　Average purity of various methods**

| 方法 | Ijcnn1 | RCV1 | Covtype-mult | Poker | MNIST-8M |
|---|---|---|---|---|---|
| FastESC | 0.904 3 ± 0 | 0.357 9 ± 0 | 0.488 1 ± 0 | 0.534 7 ± 0 | 0.472 1 ± 0.001 2 |
| LSSHC | 0.904 3 ± 0 | 0.432 4 ± 0.000 1 | 0.510 5 ± 0.000 1 | 0.679 1 ± 0.021 3 | 0.624 1 ± 0.019 9 |
| SC_RB | 0.904 3 ± 0 | 0.381 2 ± 0.001 1 | 0.496 2 ± 0.000 2 | 0.679 2 ± 0.021 5 | 0.549 9 ± 0.001 3 |
| SSEIGS | 0.904 3 ± 0 | 0.381 2 ± 0.001 2 | 0.511 4 ± 0.002 1 | 0.679 2 ± 0.012 2 | 0.549 9 ± 0.022 4 |
| USPEC | 0.904 3 ± 0.000 7 | 0.446 4 ± 0.072 6 | 0.496 7 ± 0.022 5 | 0.679 3 ± 0.005 0 | 0.572 8 ± 0.001 5 |
| 本文方法 | 0.904 4 ± 0 | 0.452 3 ± 0.000 2 | 0.521 3 ± 0 | 0.662 5 ± 0.010 3 | 0.626 6 ± 0.012 2 |

**表 4　不同方法的平均精度**

**Table 4　Average precision of various methods**

| 方法 | Ijcnn1 | RCV1 | Covtype-mult | Poker | MNIST-8M |
|---|---|---|---|---|---|
| FastESC | 0.826 9 ± 0 | 0.200 7 ± 0 | 0.378 3 ± 0 | 0.668 0 ± 0 | 0.319 5 ± 0.002 |
| LSSHC | 0.826 9 ± 0 | 0.106 3 ± 0.000 1 | 0.413 2 ± 0.000 3 | 0.667 6 ± 0.003 2 | 0.334 6 ± 0.002 3 |
| SC_RB | 0.826 9 ± 0 | 0.465 4 ± 0.012 0 | 0.393 9 ± 0.000 1 | 0.664 3 ± 0.026 4 | 0.319 9 ± 0.004 6 |
| SSEIGS | 0.827 0 ± 0 | 0.466 5 ± 0.010 2 | 0.529 1 ± 0.010 1 | 0.663 4 ± 0.021 3 | 0.376 1 ± 0.015 6 |
| USPEC | 0.826 7 ± 0.000 6 | 0.297 6 ± 0.090 7 | 0.373 0 ± 0.012 7 | 0.667 8 ± 0.021 5 | 0.379 4 ± 0.011 2 |
| 本文方法 | 0.827 0 ± 0 | 0.466 7 ± 0.000 5 | 0.386 5 ± 0 | 0.668 1 ± 0 | 0.424 0 ± 0 |

表 5 不同方法的平均召回率

Table 5 Average recall of various methods

| 方法 | Ijcnn1 | RCV1 | Covtype-mult | Poker | MNIST-8M |
|---|---|---|---|---|---|
| FastESC | 0.999 9 ±0 | 0.040 4 ±0 | 0.753 4 ±0.021 6 | 0.677 6 ±0 | 0.332 9 ±0.001 2 |
| LSSHC | 0.538 2 ±0.000 9 | 0.181 6 ±0.000 6 | 0.172 9 ±0.000 1 | 0.708 4 ±0.000 2 | 0.377 7 ±0.000 3 |
| SC_RB | 0.999 9 ±0.000 4 | 0.410 4 ±0.010 2 | 0.786 0 ±0.038 5 | 0.702 5 ±0.012 3 | 0.330 1 ±0.001 1 |
| SSEIGS | 0.969 1 ±0 | 0.399 9 ±0.012 6 | 0.599 7 ±0.000 9 | 0.705 5 ±0.004 3 | 0.481 5 ±0.001 1 |
| USPEC | 0.970 2 ±0.000 9 | 0.111 7 ±0 | 0.706 7 ±0.047 4 | 0.710 0 ±0.001 2 | 0.486 8 ±0.021 3 |
| 本文方法 | 0.999 9 ±0 | 0.415 1 ±0 | 0.786 3 ±0.024 8 | 0.705 1 ±0.010 2 | 0.486 9 ±0.002 3 |

表 6 不同方法的平均 $F_1$

Table 6 Average $F_1$ of various methods

| 方法 | Ijcnn1 | RCV1 | Covtype-mult | Poker | MNIST-8M |
|---|---|---|---|---|---|
| FastESC | 0.905 2 ±0 | 0.067 3 ±0 | 0.492 2 ±0 | 0.500 1 ±0.000 1 | 0.326 1 ±0.000 2 |
| LSSHC | 0.651 6 ±0.000 5 | 0.133 5 ±0.000 1 | 0.243 7 ±0.000 1 | 0.657 7 ±0.005 0 | 0.553 9 ±0.000 4 |
| SC_RB | 0.905 0 ±0.000 1 | 0.350 1 ±0.001 2 | 0.445 4 ±0.001 0 | 0.662 8 ±0.001 2 | 0.427 0 ±0.021 2 |
| SSEIGS | 0.892 1 ±0 | 0.359 8 ±0.001 2 | 0.493 2 ±0.000 3 | 0.657 8 ±0.016 1 | 0.496 1 ±0.010 1 |
| USPEC | 0.892 7 ±0.000 7 | 0.162 1 ±0.003 8 | 0.487 6 ±0.021 7 | 0.678 4 ±0.001 2 | 0.424 8 ±0.002 3 |
| 本文方法 | 0.905 2 ±0 | 0.365 5 ±0 | 0.498 8 ±0.002 0 | 0.644 4 ±0.000 1 | 0.554 8 ±0.000 1 |

为了观察和验证本文方法在不同维度下的性能，绘制了在 RCV1 和 Covtype-mult 这 2 个中等规模数据集上不同维度下数据聚类准确率变化的曲线，如图 3 所示。

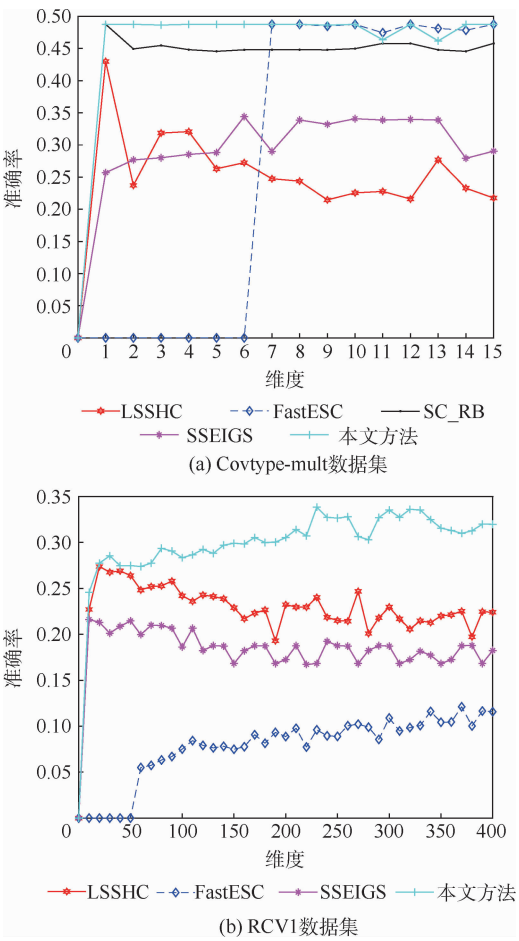



图 3 聚类准确率随维度变化的曲线

Fig. 3 Variation curves of clustering accuracy with dimensionality

图 3 中，由于 FastESC 方法的设定为最终降低到的维度不得低于聚类数目，在小于聚类数目的维度上将其准确率标为 0。观察图 3(a)可以发现，本文方法在取很低的维度时，如取维度为 1 时，就能达到高于其他所有对比方法的准确率，且随着选取维度的增大，其准确率相对稳定在较高的水平上。例如，LSSHC 方法在取维度为 1 时，同样可以达到相对较高的准确率，但是随着所取维度的增加，其准确率反而下降，且稳定性较差；SC_RB 方法在 1 维度时得到了较高的准确率，且在 2 维及更高维度上准确率相对稳定，但是其准确率在较高的维度上低于本文方法。可见，本文方法在 Covtype-mult 数据集上各个维度时均取得了较高的准确率。图 3(b)为 RCV1 数据集上各方法准确率随维度变化的曲线。因为所选的 RCV1 数据集的特征维度为 47 236，且 SC_RB 方法的空间复杂度相对较高，在升至 400 维以上时，运行时间过长甚至会出现内存不足的问题，所以本文在该数据集上选取的最大特征维度为 400 维。此外，RCV1 数据集的聚类数目为 52 个，再结合其 4 万量级的高特征维度，使得各方法在该数据集上的聚类正确率均较低。观察图 3(b)可以发现，本文方法的准确率高于其他所有对比方法。综上，本文方法可以在较低维度下取得较高且稳定的准确率。

## 4 结 论

1）结合大规模聚类数据具有模式重复性的特点和傅里叶域变换的性质，选择在傅里叶域对

数据进行建模,改变了传统特征向量的求解过程,成功地避免矩阵的复杂求逆运算,使得本文方法在计算效率方面具有显著优势。

2) 本文方法利用傅里叶域中特征向量求解的特点,实现了只从大规模数据上采样少量样本便可以训练好模型,在压缩了内存空间的同时进一步加快了方法的运算速度。

3) 本文在 Ijcnn1、RCV1、Covtype-mult、Poker 及 MNIST-8M 等大规模数据集上验证了本文方法的优越性。本文方法在保证精度的情况下,运行速度远快于 FastESC、LSSHC、SC_RB、SSEIGS 及 USPEC 等方法。

本文方法在处理大规模数据方面具有显著优势。未来将研究基于核空间及深度学习的傅里叶域大规模数据谱聚类方法。

## 参考文献 (References)


[1] PAN P, YOSHIDA Y. Average sensitivity of spectral clustering [C] // Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. New York: ACM, 2020: 1132-1140.

[2] XING L Q. Intelligent multimedia urban planning construction based on spectral clustering algorithms of large data mining[J]. Multimedia Tools and Applications, 2020, 79: 35183-35194.

[3] ALSHAMMARI M, STAVRAKAKIS J, TAKATSUKA M. Refining a *k*-nearest neighbor graph for a computationally efficient spectral clustering[J]. Pattern Recognition, 2021, 114: 107869.

[4] 谢娟英,丁丽娟,王明钊. 基于谱聚类的无监督特征选择算法[J]. 软件学报, 2020, 31(4): 1009-1024.
XIE J Y, DING L J, WANG M Z. Spectral clustering based unsupervised feature selection algorithms[J]. Journal of Software, 2020, 31(4): 1009-1024(in Chinese).

[5] 朱光辉,黄圣彬,袁春风,等. SCoS: 基于 Spark 的并行谱聚类算法设计与实现[J]. 计算机学报, 2018, 41(4): 868-885.
ZHU G H, HUANG S B, YUAN C F, et al. SCoS: The design and implementation of parallel spectral clustering algorithm based on Spark[J]. Chinese Journal of Computers, 2018, 41(4): 868-885(in Chinese).

[6] 李玉,袁永华,赵雪梅. 可变类谱聚类遥感影像分割[J]. 电子学报, 2018, 46(12): 3021-3028.
LI Y, YUAN Y H, ZHAO X M. Spectral clustering of variable class for remote sensing image segmentation[J]. Acta Electronica Sinica, 2018, 46(12): 3021-3028(in Chinese).

[7] ARRIETA J M, NAKASATO J C, PEREIRA M C. The *p*-Laplacian equation in thin domains: The unfolding approach[J]. Journal of Differential Equations, 2021, 274: 1-34.

[8] FOWLKES C, BELONGIE S, CHUNG F, et al. Spectral grouping using the Nyström method[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2004, 26(2): 214-225.

[9] DAKULAGI V. A new Nyström approximation based efficient coherent DOA estimator for radar application[J]. AEU-International Journal of Electronics and Communications, 2020, 124: 153328.

[10] DU Y, TSANG L. Accurate calculations of emissivities of polar ocean surfaces between 0.5 and 2 GHz using an NTBC/Nyström/SMCG method[J]. IEEE Transactions on Geosciences and Remote Sensing, 2020, 58(4): 2732-2744.

[11] 薛丽霞,孙伟,汪荣贵,等. 基于密度峰值优化的谱聚类算法[J]. 计算机应用研究, 2019, 36(7): 1948-1950.
XUE L X, SUN W, WANG R G, et al. Spectral clustering based on density peak value optimization[J]. Application Research of Computers, 2019, 36(7): 1948-1950(in Chinese).

[12] BOUNEFFOUF D. Spectral clustering using eigenspectrum shape based Nyström sampling [EB/OL]. (2020-07-21) [2021-09-01]. https://arxiv.org/abs/2007.11416.

[13] WU L F, CHEN P Y, YEN I E, et al. Scalable spectral clustering using random binning features[C] // Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. New York: ACM, 2018: 2506-2515.

[14] YANG Y, DENG S, LU J, et al. GraphLSHC: Towards large scale spectral hypergraph clustering[J]. Information Sciences, 2021, 544: 117-134.

[15] 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.

[16] HENRIQUES J F, CARREIRA J, CASEIRO R, et al. Beyond hard negative mining: Efficient detector learning via block-circulant decomposition[C] // Proceedings of the IEEE International Conference on Computer Vision. Piscataway: IEEE Press, 2014: 14144978.

[17] BODDETI V N, KANADE T, KUMAR B V K V. Correlation filters for object alignment[C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2013: 2291-2298.

[18] LI H, RAY N, GUAN Y, et al. Fast large-scale spectral clustering via explicit feature mapping[J]. IEEE Transactions on Cybernetics, 2018, 49(3): 1058-1071.

[19] RAHMAN M H, BOUGUILA N. Efficient feature mapping in classifying proportional data[J]. IEEE Access, 2020, 9: 3712-3724.

[20] VÁZQUEZ-MARTÍN R, BANDERA A. Spatio-temporal feature-based keyframe detection from video shots using spectral clustering[J]. Pattern Recognition Letters, 2013, 34(7): 770-779.

[21] 万月,陈秀宏,何佳佳. 利用稀疏自编码的局部谱聚类映射方法[J]. 传感器与微系统, 2018, 37(1): 145-148.
WAN Y, CHEN X H, HE J J. Local spectral clustering mapping algorithm using sparse autoencoders[J]. Transducer and Microsystem Technologies, 2018, 37(1): 145-148(in Chinese).

[22] BARNHILL E, HOLLIS L, SACK I, et al. Nonlinear multiscale regularisation in MR elastography: Towards fine feature mapping [J]. Medical Image Analysis, 2017, 35: 133-145.

[23] HANSEN T J, MAHONEY M W. Semi-supervised eigenvectors for large-scale locally-biased learning[J]. The Journal of Machine Learning Research, 2014, 15(1): 3691-3734.

[24] HUANG D, WANG C D, WU J S, et al. Ultra-scalable spectral clustering and ensemble clustering[J]. IEEE Transactions on Knowledge and Data Engineering, 2020, 32(6): 1212-1226.

[25] DONATELLI M, NOVARA P, ROMANI L, et al. A merged tuning of binary and ternary Loop's subdivision[J]. Computer Aided Geometric Design, 2019, 69: 27-44.

[26] GODI P K, KRISHNA B T, KOTIPALLI P. Design optimization of multiplier-free parallel pipelined FFT on field programmable gate array[J]. IET Circuits Devices & Systems, 2020, 14(7): 995-1000.

[27] GAO S J. Fast incremental spectral clustering in titanate application via graph Fourier transform[J]. IEEE Access, 2020, 8: 57252-57259.

[28] BIBI A, ITANI H, GHANEM B. FFTLasso: Large-scale LASSO in the Fourier domain[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2017: 17355224.

[29] ANTOGNINI J M, SOHL-DICKSTEIN J. PCA of high dimensional random walks with comparison to neural network training [EB/OL]. (2018-06-22)[2021-09-01]. https://arxiv.org/abs/1806.08805v1.

[30] DRINEAS P, KANNAN R, MAHONEY M W. Fast Monte Carlo algorithms for matrices II: Computing a low-rank approximation to a matrix[J]. SIAM Journal on Computing, 2006, 36(1): 158-183.

[31] DRINEAS P, KANNAN R. Fast Monte-Carlo algorithms for approximate matrix multiplication[C]// Proceedings on the 42nd IEEE Symposium on Foundations of Computer Science. Piscataway: IEEE Press, 2001: 452-459.

[32] LOOSLI G, CANU S, BOTTOU L. Training invariant support vector machines using selective sampling[M]// BOTTOU L, CHAPELLE O, WESTON J. Large scale kernel machines. Cambridge: MIT Press, 2007.

# A high-speed spectral clustering method in Fourier domain for massive data

ZHANG Man, XU Zhaorui, SHEN Xiangjun*

(School of Computer Science and Communication Engineering, Jiangsu University, Zhenjiang 212013, China)

**Abstract**: Spectral clustering is widely used in data mining and pattern recognition. However, due to the high computational cost of eigenvector solutions and the huge memory requirements brought by big data, spectral clustering algorithm is greatly limited when it is applied to large-scale data. Therefore, this paper studies a high-speed spectral clustering method for massive data in the Fourier domain. This method makes full use of the repeatability of data pattern, and uses this characteristic to model in the Fourier domain. To get final eigenvectors, the time-consuming eigenvector pursuitcan be transformed into the selection of the pre-determined discriminant basis in the Fourier domain. The calculation process only needs simple multiplication and addition, so the amount of time for calculation is greatly reduced. On the other hand, due to the characteristics of calculation in the Fourier domain, another advantage of this method is that it can train the samples in batches, that is, only using part of the samples can well estimate eigenvector distribution in the whole data. The experimental results on large-scale data such as Ijcnn1, RCV1, Covtype-mult, Poker and MNIST-8M show that the training time of the proposed method is at most 810.58 times faster than that of algorithms FastESC, LSSHC, SC_RB, SSEIGS and USPEC, on the premise that the clustering accuracy and other indicators are basically maintained, which proves that the proposed method has significant advantages in processing large-scale data.

**Keywords**: spectral clustering; Fourier domain; large-scale data; high-speed computation; low memory requirement

**Received**: 2021-09-08; **Accepted**: 2021-10-17; **Published online**: 2022-05-19 13:20

**URL**: kns.cnki.net/kcms/detail/11.2625.V.20220518.1910.002.html

**Foundation item**: National Natural Science Foundation of China (61572240)

* **Corresponding author**. E-mail: xjshen@ujs.edu.cn

http://bhxb.buaa.edu.cn jbuaa@buaa.edu.cn
**DOI**: 10.13700/j.bh.1001-5965.2021.0520

# 面向鱼眼图像的人群密度估计

杨家林, 林春雨*, 聂浪, 刘美琴, 赵耀
(北京交通大学 信息科学研究所, 北京 100044)

**摘 要**: 针对传统人群密度估计方法在鱼眼图像畸变下不适用的问题,提出了一个面向鱼眼图像的人群密度估计方法,实现了在鱼眼镜头场景下对人流量的监控。在模型结构方面,引入了可变形卷积,提高了模型对鱼眼畸变的适应能力。在生成目标数据方面,利用鱼眼图像的畸变特点,基于高斯变换,对人群标注转换的密度图进行符合鱼眼畸变的分布匹配。在训练方面,对损失函数的计算进行了优化,避免了模型在训练中陷入局部最优解的问题。由于鱼眼人群计数的数据集比较匮乏,采集并标注了相应的数据集。通过主客观实验与经典方法进行了对比,所提方法在测试集中的平均绝对误差达 3.78,低于对比方法,证明了面向鱼眼图像的人群密度估计方法的优越性。

**关 键 词**: 鱼眼图像; 人群统计; 畸变处理; 分布匹配; 鱼眼图像数据集
**中图分类号**: TP391
**文献标志码**: A **文章编号**: 1001-5965(2022)08-1455-09

国内外的大型活动中频发踩踏事件,已经造成了不小的伤亡。因此,在安防背景下,使用图像/视频对敏感区域(如车站、楼梯口、商场)的人流量监控是一个重要的研究课题,该成果可以防止人群骚乱、踩踏现象的发生,对突发情况下的人群汇集现象做出预警等。然而,以往的方案都是在针孔相机模型下进行的,随着鱼眼镜头的日益推广,其广阔视角能够监控到更广泛的人群。因此,研究鱼眼镜头下的人群计数统计具有十分重要的意义。此外,国内外大多数研究集中在校正鱼眼图像的畸变上,而对于鱼眼镜头下的人群密度估计问题,相关研究较少,绝大部分方案都是直接套用针孔相机模型的方法,而相较于针孔相机图像,鱼眼图像存在一定的畸变,直接套用上述方案会导致训练出的模型不准确。

针对上述问题,本文的主要工作如下:

1) 使用鱼眼镜头采集了 1 150 张人群照片,包括许多场景(如车站、商场、街道等)、不同密集程度人数的情况。同时,经过镜像翻转、随机光照变化、堆积对比度变化等数据增强方式,将数据集扩大至 5 000 张。该数据集已开放(https://download.csdn.net/download/Megurine_Luka_/19194736?spm=1001.2014.3001.5503)。

2) 提出了面向鱼眼图像的人群密度估计方法,并着眼于鱼眼图像的畸变问题,提出了解决方案。由于鱼眼镜头会使原图像中目标变得狭长或扁平,且目标越远离图像中心,这种畸变就越明显。本文通过调整二维高斯变换的协方差矩阵的参数,实现了在形状和角度上对密度图进行符合鱼眼畸变的分布匹配,进而使得模型在训练中更加适应鱼眼畸变。同时,在模型中引入了可变形卷积层[1],提高了模型的建模能力。

3) 在训练人群统计模型的过程中,会出现不收敛的情况。这是因为模型为减小损失,将结果全部输出为接近 0 的数,这样就陷入了一个局部最优解。本文使用对号函数计算损失函数,避免

**收稿日期**: 2021-09-06; **录用日期**: 2021-10-01; **网络出版时间**: 2021-11-17 10:01
**网络出版地址**: kns.cnki.net/kcms/detail/11.2625.V.20211116.1635.004.html
**基金项目**: 国家自然科学基金 (62172032,61972028)
* **通信作者**. E-mail: cylin@bjtu.edu.cn

**引用格式**: *杨家林, 林春雨, 聂浪, 等. 面向鱼眼图像的人群密度估计[J]. 北京航空航天大学学报, 2022, 48(8): 1455-1463.*
*YANG J L, LIN C Y, NIE L, et al. Crowd density estimation for fisheye images[J]. Journal of Beijing University of Aeronautics and Astronautics, 2022, 48(8): 1455-1463 (in Chinese).*

了模型训练时陷入局部最优解的情况。

## 1 相关工作

### 1.1 鱼眼图像下的检测问题

有关鱼眼图像的研究重点,在于畸变校正方面,如文献[2-3],而关于检测的研究较少。一些研究[4]虽然涉及了鱼眼图像,但思路是先校正图像,再按照一般图像进行处理。这种处理方式存在 2 个问题:①校正算法通常不太完美,处理过的图像仍存在畸变,甚至造成新的失真;②校正算法步骤繁琐,耗费的时间太长,不利于实时进行。

### 1.2 人群统计

早期的人群计数方法主要基于检测或者基于回归,但面临透视失真及目标之间的遮挡问题,导致目标可能只有一部分会出现在图像中,这将大大影响计数方法的准确性。随着深度学习的发展,卷积神经网络开始大量应用于图像识别中,包括人群估计问题。目前,通过卷积神经网络模型识别原图像,生成密度图,再基于密度图估计人群数量,已经成为处理此类问题的一个经典方案。

有关生成密度图的研究成果很多。比较有代表性的有 Zhang 等[5]提出的多列卷积神经网络(MCNN),该模型是一种 3 列卷积神经网络结构,通过 3 种不同尺寸滤波器的感受野,每列卷积网络能识别不同尺寸大小的人头目标。不过,MCNN 存在模型难以训练、结构冗余、分支结构效率不高、结构精度不高等缺点,对此,Li 等[6]提出了高度拥挤场景下的空洞卷积网络(CSRNet),该结构的前半部分是去掉全连接层的 VGG-16[7],后半部分是由一些空洞卷积层堆叠而成,其目的是逐步增大网络的感受野,避免池化造成的图像失真。Guo 等[8]通过空洞卷积,使得神经网络获得不同感受野,再利用可变形卷积,以适应不同的人头轮廓,从而起到更好的定位效果。

最近,随着注意力机制在自然语言处理领域的成功应用,其思想也被引入到了人群计数的研究中[9-11]。

此外,Zhang 等[12]将条件随机场应用于人群统计中,通过计算各个特征图之间的注意力,构造条件随机场,利用条件随机场的迭代更新机制进行特征融合。Gao 等[13]通过结合风格迁移,提高模型在不同场景下的适应性。

### 1.3 目标密度图的生成与损失函数计算

神经网络模型的训练需要相应的训练数据,即原始图像与作为目标的密度图,同时设计损失函数指导训练过程。密度图的生成通常基于手工标注的散点图,经过高斯变换等方式,转化为密度图,通过计算神经网络的输出与该密度图的差异作为损失函数,完成训练。

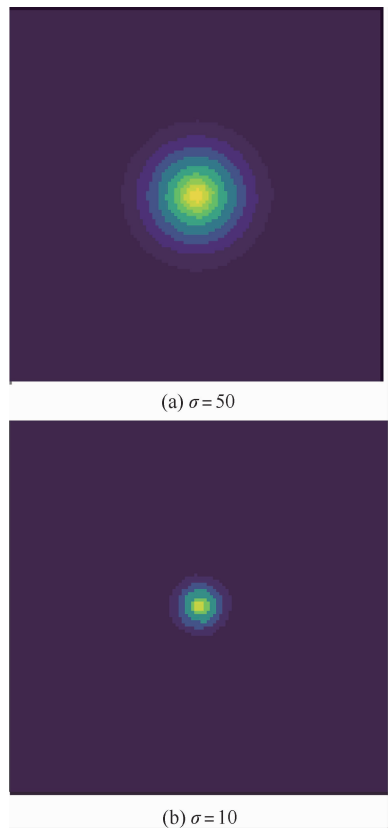
目前,最常用的方法是利用高斯核函数来处理点标注图:

$$D(\boldsymbol{x}) = \sum_{m=1}^{M} \frac{1}{\sqrt{2\pi}\sigma} \exp\left( - \frac{\|\boldsymbol{x} - \boldsymbol{x}_m\|_2^2}{2\sigma^2} \right) \tag{1}$$

式中:$\boldsymbol{x}$ 为要计算的像素;$D(\boldsymbol{x})$ 为对应的高斯函数值;$\boldsymbol{x}_m$ 为标注点坐标;$\sigma$ 为高斯函数的方差,其值越大,则该高斯核所覆盖的面积就越大;$M$ 为图像像素的总数。

由图 1 可见,该思路最大的不确定因素就是 $\sigma$ 的选取。早期的方案[5-6]中,$\sigma$ 取一个固定的值。显然,较大的目标(在本任务中,目标就是人头)对应较大的 $\sigma$ 值,较小的目标对应较小的 $\sigma$ 值,于是,就有了如下解决方案:令 $l$ 表示像素 $\boldsymbol{x}$ 与其最近的若干(通常取 3)标注点的平均距离,令 $\sigma = \eta l$($\eta$ 通常取 0.3),这样处理后的 $\sigma$ 值会自适应地根据具体情景调整大小。

不过,这样的处理方法仍有不足之处,因为高斯函数不能保证能够将目标人头精确覆盖。在鱼眼图像中,无论是不同图像间,还是在同一图像中,人头大小的差异都非常大。如果高斯函数不能比较精确地覆盖,很容易造成误判。



(a) $\sigma = 50$

(b) $\sigma = 10$

图 1 $\sigma$ 值对高斯分布的影响

Fig. 1 Influence of $\sigma$ on Gaussian distribution

对此,Wan 和 Chan[14] 提出在对模型进行训练的同时,对生成的密度图也做出一定的微调。首先,用现有的人群预测模型,如 MCNN[5] 或者 CSRNet[6],预训练一个人群密度估计器。同时,设计一个密度图生成器,将其生成的密度图与估计器协同训练,最终得到一个最优的结果。然而,从实验结果看,算法的效果一般,且算法模型较为复杂。

Ma 等[15] 提出了一种基于贝叶斯公式的损失函数计算方法。先根据点标注的人头位置,通过代入贝叶斯公式,计算出每个像素属于每个点标注的期望,再将神经网络的预测结果与期望相乘,得到人群估计的总人数,将估计人数与样本的真实人数作比较,得到损失函数。该方案可以有效解决上述高斯函数无法精确覆盖的问题,且实现起来较为简便,因此有很大的参考价值。

## 2　鱼眼镜头下的人群统计

### 2.1　鱼眼镜头与鱼眼图像

鱼眼镜头属于一种超宽视场镜头,其超大范围的视场是以牺牲图像的直观性换来的。鱼眼镜头拍摄的图像通常或多或少地存在一些“畸变”。第 1 节提到的人群算法,都是基于常规图像下的检测。如果不针对这些畸变做出相应的调整,必然会对检测效果造成一定的影响。

### 2.2　预测模型

人群统计系统的构建主要分为 2 个步骤:预测模型的选取与损失函数的计算。预测模型通过识别输入图像,输出密度图。损失函数则会指导预测模型的训练,使其生成质量更高的密度图。本文的预测模型以 CSRNet[6] 为基准模型,并在此基础上改进,将其中的 2 层空洞卷积层修改为可变形卷积层,提高网络对畸变的建模能力。

本文中网络结构主要分为 2 部分。第 1 部分结构与 VGG-16 类似,只是没有全连接层。因为全连接层的存在会导致模型只能处理大小固定的图像,而去掉全连接层则能体现模型的灵活性。第 2 部分由 2 层可变卷积层与若干个空洞卷积层组成,其中空洞卷积层的目的是扩大模型的感受野,并尽量减小图像信息的失真,以生成更高质量的人群密度图。与 CSRNet 相比,本文将其中的前 2 个空洞卷积层替换为可变形卷积层。相较于空洞卷积,可变形卷积更为灵活,虽然空洞卷积能够扩大感受野,但其卷积核仍是方形。而可变形卷积通过给每个卷积核参数添加方向向量(通过一个额外卷积层训练得到),使得卷积核可以改变为任意形状。

图 2 为预测模型的完整结构图,其输入为 3 通道的 RGB 图像,输出为单通道密度图,长宽为输入图像的 1/8。

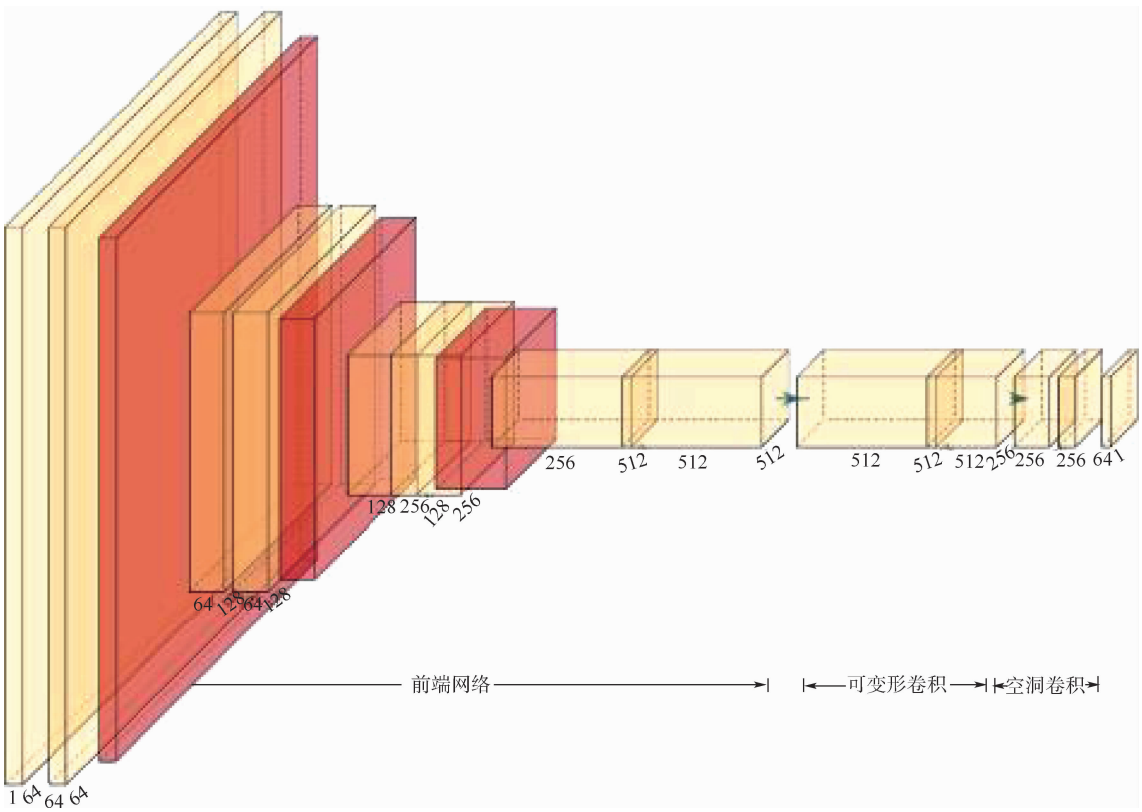



图 2　预测模型结构

Fig. 2　Structure of prediction model

## 2.3 基于鱼眼图像畸变的贝叶斯损失函数优化

### 2.3.1 训练数据的准备

在模型训练之前,需要将每个点标注通过高斯变换,生成一个个大致圆形的区域,生成供训练的目标密度图。由于鱼眼图像所造成的畸变,需要对密度图进行调整,才能使高斯函数更好地覆盖人群的头部。

考虑到鱼眼图像的畸变特点是:位于图像中心的目标无畸变,但随着目标远离图像中心,其沿图像中心方向的宽度会逐渐缩小,进而变得狭长(或扁平)。因此,高斯函数图像必须根据标注点到图像中心的方向、距离,做出相应的变换(如拉伸、旋转),才能精确地覆盖目标。具体步骤如下:

对于多维高斯分布而言,其表达式为

$$f(\boldsymbol{x}) = \frac{1}{(2\pi)^{\frac{d}{2}}|\boldsymbol{\Sigma}|^{\frac{1}{2}}}\exp\left[-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right] \tag{2}$$

式中:$d$ 为高斯分布维数;$\boldsymbol{\mu}$ 为均值;$\boldsymbol{\Sigma}$ 为协方差矩阵,其对角线上的元素表示该分布在各个维度上分量的方差。

对于二维高斯分布,其协方差矩阵 $\boldsymbol{\Sigma}$ 是一个 $2\times2$ 的矩阵,如式(3)所示,其主对角线上的值 $\alpha$ 与 $\beta$ 分别表示 $x$ 和 $y$ 轴的方差,如果 $\alpha$ 减小,$\beta$ 不变,则会导致分布图像 $x$ 轴方向变得扁平,如图 3 所示。矩阵斜对角线上的值是二维高斯分布 $x$ 与 $y$ 的协方差,在本文中,协方差均视为 0。

$$\boldsymbol{\Sigma} = \begin{bmatrix}\alpha & 0\\ 0 & \beta\end{bmatrix} \tag{3}$$

利用该性质,固定协方差矩阵的 $\beta$ 值,令 $\alpha$ 值按照式(4)、式(5)变化:

$$\beta = \alpha = 20 \tag{4}$$

$$\alpha = \sigma - 0.95\sigma(\mathrm{dis}/\mathrm{mdis}) \tag{5}$$

式中:dis 为标注点到图像中心的距离;mdis 为鱼眼图像的半径。由此距离图像中心较远的目标变将会被拉长。

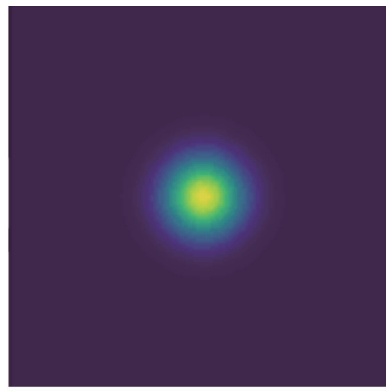
同时,根据标注点与中心的相对位置,对该高斯核做一定的旋转。如图 4 所示,$P$ 为图像中心点,$Q$ 为标注点,$\theta$ 为 $PQ$ 与横轴的夹角,整个高斯核将绕纵轴旋转角度 $\theta$。

旋转变换公式为

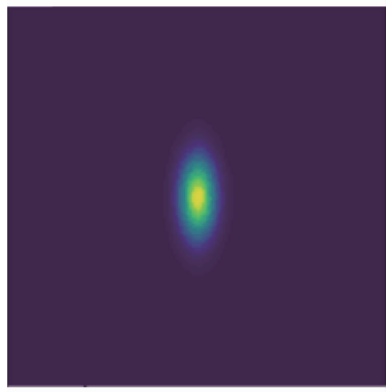
$$x' = (x-q_x)\cos\theta + (y-q_y)\sin\theta + q_x \tag{6}$$

$$y' = -(x-q_x)\sin\theta + (y-q_y)\cos\theta + q_y \tag{7}$$

式中:$(x',y')$ 为旋转变换后的坐标;$(q_x,q_y)$ 为 $Q$ 点坐标。



(a) $\Sigma=\begin{bmatrix}50 & 0\\ 0 & 50\end{bmatrix}$



(b) $\Sigma=\begin{bmatrix}10 & 0\\ 0 & 50\end{bmatrix}$

图 3 方差值的变化对高斯分布的影响

Fig. 3 Influence of covariance matrix on Gaussian distribution

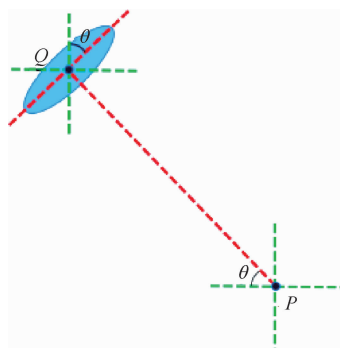



图 4 旋转效果

Fig. 4 Effects of rotation

通过上述拉伸、旋转变换,由高斯函数图像生成的目标密度图能够更好地契合鱼眼图像,如图 5所示。

### 2.3.2 鱼眼图像计数优化的损失函数

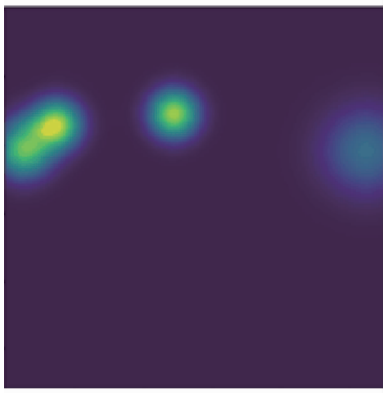
损失函数用于指导预测模型的训练,对模型预测结果的精确与否至关重要,本文在此方面也做出一定的改进。损失函数分为 2 部分,即计数回归损失与环境损失。

1) 计数回归损失

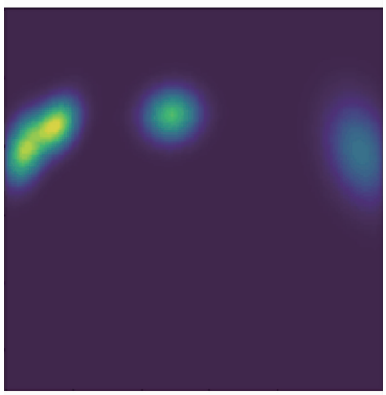
该损失函数表示人群数目预测值与真实值之间的误差。

具体来说,先按照 2.3.1 节的步骤生成密度

(a) 原图像



(b) 未经拉伸旋转变换的目标密度



(c) 经过拉伸旋转变换的目标密度

图 5　变换效果

Fig. 5　Transformation effect

图,并按式(8)、式(9)计算。

令

$$N(\boldsymbol{x}_m;\boldsymbol{z}_n,\sigma^2\boldsymbol{1}_{2\times2}) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{\|\boldsymbol{x}_m-\boldsymbol{z}_n\|_2^2}{2\sigma^2}\right) \tag{8}$$

有

$$P(y_n\mid\boldsymbol{x}_m) = \frac{N(\boldsymbol{x}_m;\boldsymbol{z}_n,\sigma^2\boldsymbol{1}_{2\times2})}{\sum_{i=1}^{N}N(\boldsymbol{x}_m;\boldsymbol{z}_i,\sigma^2\boldsymbol{1}_{2\times2})} \tag{9}$$

式中:$n$ 为标注点的序号;$m$ 为像素点的序号;$\boldsymbol{z}_n$ 为标注点的坐标向量;$y_n$ 为第 $n$ 个标注点;条件概率 $P(y_n\mid\boldsymbol{x}_m)$ 表示第 $m$ 个像素属于第 $y_n$ 个标注点所代表的人头的概率。

一个像素点既可能位于目标人头上,也可能不属于任何目标,或者说属于“环境”。事实上,一张图像中大部分的像素都属于“环境”。

像素 $\boldsymbol{x}_m$ 属于“环境”的概率计算式为

$$P(y_n\mid\boldsymbol{x}_m) = N(\boldsymbol{x}_m;\boldsymbol{z}_0,\sigma^2\boldsymbol{1}_{2\times2}) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left[-\frac{(\delta-\|\boldsymbol{x}_m-\boldsymbol{z}_n^m\|_2)^2}{2\sigma^2}\right] \tag{10}$$

式中:$\boldsymbol{z}_n^m$ 为离像素 $\boldsymbol{x}_m$ 最近的标注点的位置;$\delta$ 为一个常数。

式(9)就改写成

$$P(y_n\mid\boldsymbol{x}_m) = \frac{N(\boldsymbol{x}_m;\boldsymbol{z}_n,\sigma^2\boldsymbol{1}_{2\times2})}{\sum_{i=0}^{N}N(\boldsymbol{x}_m;\boldsymbol{z}_i,\sigma^2\boldsymbol{1}_{2\times2})} \tag{11}$$

将式(11)的计算结果代入式(12):

$$E[c_n] = \sum_{m=1}^{M}P(y_n\mid\boldsymbol{x}_m)\boldsymbol{Dest}(\boldsymbol{x}_m) \tag{12}$$

式中:$E[c_n]$ 为第 $n$ 个标注点(第 $n$ 个人头)在整张图像上的数学期望;**Dest** 为神经网络所输出的密度图,$\boldsymbol{Dest}(\boldsymbol{x}_m)$ 表示像素 $\boldsymbol{x}_m$ 在密度图上所对应的数值。

第 $n$ 个标注点的数目显然是 1,因此损失函数为

$$\text{Loss} = \sum_{n=1}^{N}F(1-E[c_n]) \tag{13}$$

式中:$F$ 为距离函数。

2) 环境损失

神经网络也可能会将本来属于环境的点判断为标注点。损失函数也应对此作出惩罚:通过原始标注点生成一个图像,凡是与标注点的距离大于 $r$ 的都被视为环境像素,否则是标注点像素。如图 6 所示,假设深灰色为样本点,黑色圆的半径是 $r$,浅灰色为环境区域。

具体来说,圆心为标注点,黑色圆周的是“有可能是人头”的区域,剩下的就是“环境”(负样本),如果神经网络将“环境”判断成“人群”,就会计入到损失函数内。

定义如下矩阵:

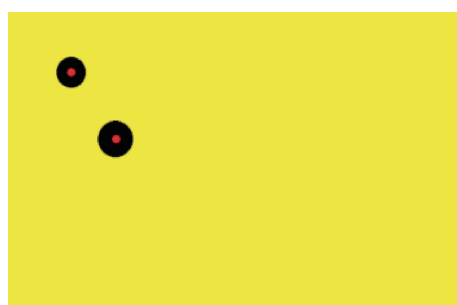


图 6　标注点与环境区域

Fig. 6　Labeled points and environment areas

$$\mathbf{Mask} = (\mathrm{mask}_{ij}) \tag{14}$$

$$\mathrm{mask}_{ij} = \begin{cases} 1 & \text{坐标}(i,j)\text{ 位于圆外} \\ 0 & \text{坐标}(i,j)\text{ 位于圆内} \end{cases} \tag{15}$$

则环境损失为

$$\mathrm{Loss} = \mathrm{sum}(\mathbf{Mask} \circ \mathbf{Dest}_{\mathrm{pre}}) \tag{16}$$

式中:sum 为对矩阵所有元素的值求和;$\mathbf{Dest}_{\mathrm{pre}}$为预测模型所估计的密度图;“∘”表示 Hadamard 积。

3) 总体优化

损失函数为

$$\mathrm{Loss} = \sum_{n=1}^{N} F(1 - E[c_n]) + \mathrm{sum}(\mathbf{Mask} \circ \mathbf{Dest}_{\mathrm{pre}}) \tag{17}$$

距离函数 $F$ 的选择会影响到模型的训练。Bayesian Loss 在文献[15]中 $F$ 采用的是绝对值函数,但这样会造成一个问题,对于“未将标注点判断成人头”与“将环境区域判断成了人头”,其所占权重是一样的。然而,图像中的“环境区域”通常远远比“标注区域”广大,神经网络为了减小损失,会将所有 **Dest** 中的点都判断为“0”以减小损失,这样就陷入了一个局部最优解。

对此,本文将损失函数式(17)的第 1 项换成了对号函数 $1/E[c_n] + E[c_n]$,对号函数的极小值点是 1,当 $E[c_n]$趋于 0 时,函数的导数值将急剧上升。因此,选用对号函数可以减少模型对人数估计过低的情况。

损失函数如下:

$$\mathrm{Loss} = \frac{\gamma}{N}\sum_{n=1}^{N}\left(\frac{1}{E[c_n]} + E[c_n]\right) + \mathrm{sum}(\mathbf{Mask} \circ \mathbf{Dest}_{\mathrm{pre}}) \tag{18}$$

式中:系数 $\gamma$ 用于配置权重,本文中取 1;$N$ 为该样本图像中的目标总数。值得注意的是,在数据集中,各样本之间的目标数目差距很大,有的图像上的目标数少,其产生的损失函数也很小,从而难以对模型训练产生作用,因此需要对损失函数除以目标数,使得数据样本更加平衡。

此外,预测模型需要对其输出结果做一定的限制。因为只有可能出现一个人头占许多个像素点的情况,而一个像素点上不可能出现 1 个以上的人头,所以必须将预测模型的输出结果限制在 0 ~ 1 之间(本实验中,将输出结果限制在 0 ~ 0.25之间,达到的效果是最好的),这大大降低了模型训练的难度,且其他文献并没有提到这一点。

# 3 实 验

## 3.1 实验设置

### 3.1.1 数据集

由于目前没有适用于鱼眼图像人群检测的公开数据集,本文采集了一个新的鱼眼镜头下的人群数据集。具体的,使用带有鱼眼镜头手机拍摄人群图像,并使用 labelme 插件对图像中的人头进行点标注。数据集共有 1 150 张照片,涵盖了各个场景(如学生放学、商场及车站等)及各种人数情况下的人群。将 1 150 张图像的数据集分为 2 部分:1 000 张作为训练集,150 张作为测试集。使用时,通过程序读取图像与存有图像标注结果的 json 文件,计算每个像素属于每个标注点的概率,作为神经网络的训练样本。

### 3.1.2 评价指标

评价指标为平均绝对误差(MAE)与误差(Bias)。MAE 的计算式为

$$\mathrm{MAE} = \frac{1}{m}\sum_{i=1}^{m}|\mathrm{estimate}_i - \mathrm{gt}_i| \tag{19}$$

式中:$m$ 为测试集中样本的数目;$\mathrm{estimate}_i$ 为模型对第 $i$ 个测试样本所估计的人数;$\mathrm{gt}_i$ 为第 $i$ 个样本的真实人数。

误差 Bias 表示模型平均估计一人所产生的偏差,计算公式为

$$\mathrm{Bias} = \frac{\sum_{i=1}^{m}|\mathrm{estimate}_i - \mathrm{gt}_i|}{\sum_{i=1}^{m}\mathrm{gt}_i} \tag{20}$$

### 3.1.3 实验细节

考虑到当前数据集规模不大,为防止训练 epoch数量过多导致过拟合,将训练的 epoch 数量设置为 12,学习率设置为 $10^{-8}$。

## 3.2 实验结果及分析

### 3.2.1 对比方法

本节实验以 MCNN[5]、CSRNet[6]、Bayesian Loss[15]及最近提出的 D2CNet[16]作为对照组,给出其在测试集中的 MAE 与误差。

### 3.2.2 定量结果

表 1 记录了各个模型在迭代训练中 MAE 出现的最低值(最优结果)。图 7 显示了各个模型在迭代训练次数增加时 MAE 的变化趋势。

表 1 及图 7 表明,本文方法有最低的 MAE,以及较快的收敛速度,Bayesian Loss 其次,之后是 CSRNet 与 D2CNet,而较早提出的 MCNN 模型效果较差。

**表 1　不同方法的最优结果**

**Tabel 1　The best result of different methods**

| 方法 | MAE | Bias |
|---|---|---|
| D2CNet | 6.18 | 0.44 |
| MCNN | 12.21 | 0.88 |
| CSRNet | 5.66 | 0.41 |
| Bayesian Loss | 3.96 | 0.29 |
| 本文方法 | 3.78 | 0.27 |

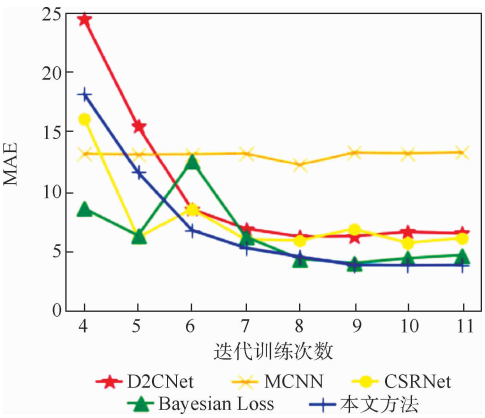



图 7　MAE 对比

Fig. 7　Comparison of MAE

### 3.2.3　消融实验

本文主要的改进有 3 点:①使用变形高斯核函数生成目标密度图;②使用对号函数等方法优化损失函数的计算;③使用可变形卷积优化模型结构。本文对这 3 个条件做消融实验探究了其对模型效果的影响:

**方案 1**　使用绝对值函数计算损失函数,其余条件同本文方法一致。

**方案 2**　不使用变形高斯卷积核处理目标密度图,其余条件一致。

**方案 3**　使用空洞卷积代替可变形卷积,其余条件一致。

同时,将本文方法不做任何改变,作为实验的对照组。

表 2 为各方案在迭代训练中,MAE 与 Bias 出现的最低值(最优结果)。表 2 显示,使用本文方法计算损失函数对模型的提升很大,使用变形卷积核可进一步提升模型的准确性,而可变形卷积的引入对模型也有一定的提升。

**表 2　消融实验结果**

**Tabel 2　Ablation results**

| 对比方法 | MAE | Bias |
|---|---|---|
| 方案 1 | 5.59 | 0.40 |
| 方案 2 | 4.22 | 0.30 |
| 方案 3 | 3.89 | 0.28 |
| 对照组 | 3.78 | 0.27 |

### 3.2.4　可视化结果

密度图生成效果如图 8 所示,左起第 1 列为原始图像,第 2 列为由真实标注生成的密度图,第 3 列为模型生成的密度图。

可见,本文方法生成密度图的效果不错,能够较为准确地反映人群密度的密集与稀疏,人群数目的估计也较为准确。

### 3.2.5　分　析

通过分析实验结果得知,本文方法能够较为准确地生成密度图,且相比其他方法有着更低的 MAE,达 3.78。影响人群密度估计结果的原因大致如下:

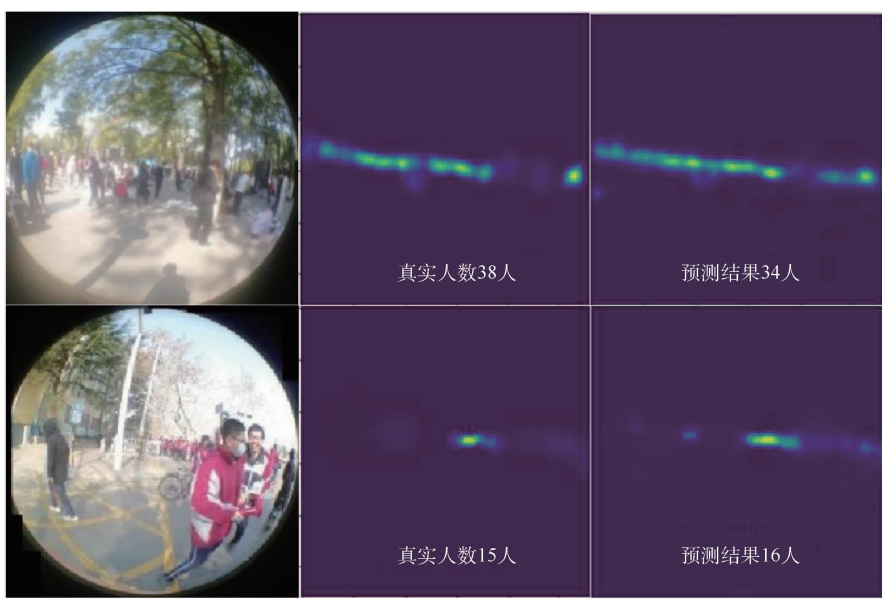



图 8　人群密度图生成效果

Fig. 8　Display of crowd density maps

1）数据集较为模糊，人群的特征并不是很明显。数据集在训练、预测时难以做到特别精确。但在实际应用中，监控摄像头拍摄的图像也不是很清晰，因此使用较为模糊的数据集更接近实际情况。

2）应用场景较为复杂，如在商场货架上悬挂的衣服，会对模型的训练产生干扰。

3）数据集中样本人群的疏密不均衡也会对模型的效果造成影响。

## 4 结 论

本文提出了一个面向鱼眼图像的人群密度估计方法，通过引入可变形卷积，以及对目标数据密度图进行符合鱼眼畸变的分布匹配，提高了模型对鱼眼畸变的适应能力；通过优化损失函数，提高了模型的鲁棒性；通过与其他方法的对比实验，证明了本文方法的优越性。同时，本文采集并标注了鱼眼图像下的人群数据集，为工作的开展提供了数据支撑。

后续将针对影响人群密度估计结果的原因，做进一步的研究。

### 参考文献（References）


[1] DAI J F, QI H Z, XIONG Y W, et al. Deformable convolutional networks[C]// 2017 IEEE International Conference on Computer Vision. Piscataway: IEEE Press, 2017: 764-773.

[2] XUE Z C, XUE N, XIA G S, et al. Learning to calibrate straight lines for fisheye image rectification[C]// 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2019: 1643-1651.

[3] LIAO K, LIN C, Y ZHAO. A deep ordinal distortion estimation approach for distortion rectification[J]. IEEE Transactions on Image Processing, 2021, 30: 3362-3375.

[4] 徐佳，杨鸿波，宋阳，等. 基于鱼眼摄像头的一种人脸识别技术[J]. 信息通信，2018，31(1)：131-132.
XU J, YANG H B, SONG Y, et al. A face recognition technology based on fisheye camera[J]. Information & Communications, 2018, 31(1): 131-132(in Chinese).

[5] ZHANG Y Y, ZHOU D S, CHEN S Q, et al. Single-image crowd counting via multi-column convolutional neural network[C]// 2016 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2016: 589-597.

[6] LI Y H, ZHANG X F, CHEN D M. CSRNet: Dilated convolutional neural networks for understanding the highly congested scenes[C]// 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2018: 1091-1100.

[7] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition[EB/OL]. (2015-04-10)[2021-09-01]. https://arxiv.org/abs/1409.1556.

[8] GUO D, LI K, ZHA Z J, et al. DADNet: Dilated-attention-deformable ConvNet for crowd counting[C]// Proceedings of the 27th ACM International Conference on Multimedia. New York: ACM, 2019: 1823-1832.

[9] ZHANG A R, SHEN J Y, XIAO Z H, et al. Relational attention network for crowd counting[C]// 2019 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway: IEEE Press, 2019: 6787-6796.

[10] WANG Q, BRECKON T P. Crowd counting via segmentation guided attention networks and curriculum loss[EB/OL]. (2020-08-03)[2021-09-01]. https://arxiv.org/abs/1911.07990.

[11] DAS S S S, RASHID S M M, ALI M E. CCCNet: An attention based deep learning framework for categorized crowd counting[EB/OL]. (2019-11-12)[2021-09-01]. https://avxiv.org/abs/1912.05765.

[12] ZHANG A R, YUE L, SHEN J Y, et al. Attentional neural fields for crowd counting[C]// 2019 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway: IEEE Press, 2019: 5713-5722.

[13] GAO J Y, HAN T, WANG Q, et al. Domain-adaptive crowd counting via inter-domain features segregation and Gaussian-prior reconstruction[EB/OL]. (2019-11-08)[2021-09-01]. https://arxiv.org/abs/1912.03677.

[14] WAN J, CHAN A. Adaptive density map generation for crowd counting[C]// 2019 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway: IEEE Press, 2019: 1130-1139.

[15] MA Z H, WEI X, HONG X P, et al. Bayesian loss for crowd count estimation with point supervision[C]// 2019 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway: IEEE Press, 2019: 6141-6150.

[16] CHENG J, XIONG H P, CAO Z G, et al. Decoupled two-stage crowd counting and beyond[J]. IEEE Transactions on Image Processing, 2021, 30: 2862-2875.

# Crowd density estimation for fisheye images

YANG Jialin, LIN Chunyu*, NIE Lang, LIU Meiqin, ZHAO Yao

(Institute of Information Science, Beijing Jiaotong University, Beijing 100044, China)

**Abstract**: Aiming at the problem that the traditional crowd density estimation methods are not applicable under the distortion of fisheye images, this paper presents a crowd density estimation method for fisheye images, which realizes the monitoring of human traffic in scene of using fisheye lens. For model structure, we introduced deformable convolution to improve the adaptability of the model to fisheye distortion. For generating the training targets, we used Gaussian transform to perform a distribution match on the density maps of annotations, which depends on the features of fisheye distortion. For training, we optimized the loss function to avoid the model from falling into local optimal solutions. In addition, we collected and labeled the corresponding dataset due to the lack of dataset for fisheye crowd estimation. At last, by comparing the subjective and objective experiments with classical algorithms, we proved the superiority of the crowd estimation method for fisheye images in this paper with the mean absolute error of 3.78 in the test dataset, which is lower than others.

**Keywords**: fisheye image; crowd estimation; distortion processing; distribution matching; fisheye image dataset

**Received**: 2021-09-06; **Accepted**: 2021-10-01; **Published online**: 2021-11-17 10:01

**URL**: kns.cnki.net/kcms/detail/11.2625.V.20211116.1635.004.html

**Foundation items**: National Natural Science Foundation of China (62172032,61972028)

* **Corresponding author**. E-mail: cylin@bjtu.edu.cn

http://bhxb.buaa.edu.cn jbuaa@buaa.edu.cn
**DOI**: 10.13700/j.bh.1001-5965.2021.0522

# 用于遥感图像变化检测的全尺度特征聚合网络

刘国强, 房胜*, 李哲
(山东科技大学 计算机科学与工程学院, 青岛 266590)

**摘 要**: 变化检测(CD)是遥感的一项重要任务,通常面临许多伪变化和较大的尺度变化。目前的方法主要侧重于对差异特征的建模,忽略了从原始图像中提取足够的信息,影响了特征的识别能力,难以稳定地区分出变化区域。针对以上问题,提出了一种全尺度特征聚合网络(FFANet)来更充分地利用原始图像特征,促使生成的特征表示在语义上更丰富、在空间上更准确,从而提高了网络对小目标和目标边缘的检测性能。同时,拓展了深监督来结合多尺度的预测图,以促使不同对象在更合适的尺度上进行检测,从而提升了网络对对象尺度变化的鲁棒性。在 CDD 数据集上,相比于基线网络,所提方法仅增加了 $1.01\times10^6$ 的参数量,就将 $F_1$ 分数提升了 0.034。

**关 键 词**: 变化检测(CD); 深监督; 全尺度特征聚合; 多尺度预测; 遥感图像
**中图分类号**: TP751
**文献标志码**: A **文章编号**: 1001-5965(2022)08-1464-07

变化检测(change detection, CD)的目的是识别不同时间采集的同一区域的多时相遥感图像间的差异,在城市扩张[1]、农田制图[2]、灾害监测[3]等诸多领域有着广泛的应用。作为一项特殊的遥感任务,变化检测很容易受到外界环境因素的干扰,如光照变化、季节变化[4]、噪声干扰[5]等,会导致具有相同语义概念的物体在不同的时间和空间位置可能表现出不同的光谱行为。此外,由于感兴趣的对象可能存在较大的尺度变化,要求所提方法能够鲁棒地检测出不同尺度的物体。

传统的变化检测方法主要关注遥感图像的光谱值、纹理和形状,而忽略了对空间上下文的利用。例如,变化向量分析(CVA)[6]先计算图像间的变化向量,再结合变化方向和幅度来判断变化类型。主成分分析(PCA)[7]常被用来减少冗余数据,而缨帽变换(KT)可产生稳定的光谱成分,为长期研究提供基础的光谱信息。此外,人工神经网络(ANN)和支持向量机(SVM)[8]等机器学习方法能够处理更大的数据集,可避免维度爆炸。

近年来,深度学习算法,尤其是卷积神经网络(CNN)在变化检测领域取得了很好的效果。FC-EF[9]、FC-Siam-conc[9]、FC-Siam-diff[9]是早期的3种全卷积神经网络,实现了端到端的训练。其中,后2种网络应用了孪生架构,对目前问题进行了深入研究。IFN[10]通过深监督和注意力机制提高了输出变化图中对象边界的完整性和内部的紧凑性。进一步地,SNUNet[11]采用孪生的子网络作为编码器,并将 NestedUNet[12]作为解码器,通过在编码器和解码器间及解码器和解码器间进行密集的信息传输来更充分地利用低层次细节特征,确保深层特征定位准确。为了生成更具辨识性的特征图,STANet[13]、DASNet[14]引入了自注意力机制来捕获远程依赖。

基于深度特征的方法能够提取到图像的深层语义,因此对伪变化的鲁棒性较高。然而,这些方法大多是基于 U-Net[15]或 FCN[16]架构实现的,其

**收稿日期**: 2021-09-06; **录用日期**: 2021-10-01; **网络出版时间**: 2021-10-29 16:46
**网络出版地址**: kns.cnki.net/kcms/detail/11.2625.V.20211028.1950.004.html
**基金项目**: 山东省自然科学基金(ZR2020MF132)
* **通信作者**. E-mail: fangsheng@tsinghua.org.cn

**引用格式**: *刘国强, 房胜, 李哲. 用于遥感图像变化检测的全尺度特征聚合网络[J]. 北京航空航天大学学报, 2022, 48(8): 1464-1470. LIU G Q, FANG S, LI Z. A full-scale feature aggregation network for remote sensing image change detection[J]. Journal of Beijing University of Aeronautics and Astronautics, 2022, 48(8): 1464-1470 (in Chinese).*

利用普通的编码器来提取原始图像的特征存在2方面的问题：①低层次特征图缺乏足够的语义，当特征由编码器传向解码器时存在较大的语义鸿沟；②随着逐层的进行下采样，深层特征的空间定位变得不太准确，影响了对小目标和目标边缘的检测。

许多研究[15-20]表明，不同尺度的特征图发挥着不同的作用。低层次的小尺度特征图探索了丰富的空间信息，能够突出物体的边界；而高层次的大尺度特征图提取了物体的深层语义，能更鲁棒地识别出伪变化。因此，不同尺度的特征图之间存在一定的互补性。本文提出了一种全尺度特征聚合的编码器来最大限度地发挥不同尺度特征图的优势，以确保编码器生成的所有尺度的特征图都语义丰富且定位准确。此外，为了适应目标的多尺度变化，本文拓展了深监督以在解码器的每个尺度上都生成相应的预测图，结合这些具有不同尺度表示的预测图来生成最终的预测结果。

# 1　网络结构

## 1.1　网络主体架构

本文设计的全尺度特征聚合网络（FFANet）包括双流编码器、解码器和分类器3部分。如图1所示，编码器由共享权值的双流分支组成，用以独立地编码双时相图像的全尺度特征，这些特征被传向解码器以进一步提取差异信息。分类器结合解码器生成的所有尺度的特征图来生成最终的预测结果。

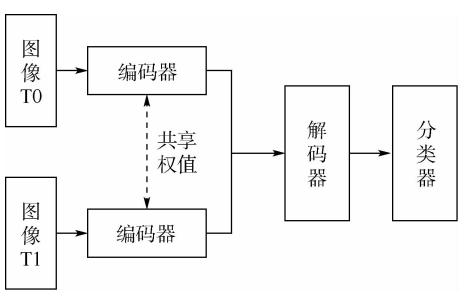



图1　FFANet的网络结构

Fig. 1　Network architecture of FFANet

## 1.2　编 码 器

编码器是共享权值的双流结构，每一条流都包含自上而下和自下而上2条分支，如图2所示，2条分支相加以合并为每条流的输出。

### 1.2.1　自上而下的分支

作为网络的主干，自上而下的分支被用来提取双时相图像的有效特征。由于低层次特征图包含了丰富的空间信息，突出了对象边界，通过密集

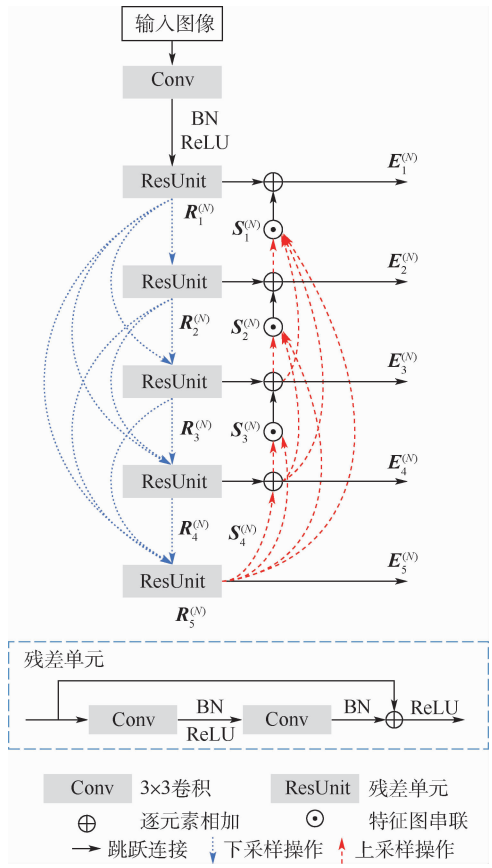



图2　用于全尺度特征融合的编码器

Fig. 2　Encoders for full-scale feature fusion

的跳跃连接使深层能直接接收所有较浅层的特征图以确保深层特征具有准确的位置表示。$\{\boldsymbol{R}_i^{(N)} \mid i=1,2,3,4,5; N=1,2\}$表示自上而下的分支所生成的特征图的集合，$i$为层索引，$N$代表编码器的2个流，$\boldsymbol{R}_i^{(N)}$可表示为

$$\boldsymbol{R}_i^{(N)} = \begin{cases} C_{\mathrm{R}}(C_{3\times3}(\boldsymbol{R}_0^{(N)})) & i=1 \\ C_{\mathrm{R}}([D(\boldsymbol{R}_k^{(N)})_{k=1}^{i-1}]) & i=2,3,4,5 \end{cases} \tag{1}$$

式中：$\boldsymbol{R}_0^{(N)}$为原始图像；$C_{3\times3}(\cdot)$为3×3卷积；$C_{\mathrm{R}}(\cdot)$为残差单元；$D(\cdot)$为下采样操作；$[\cdot]$为特征图的串联。

### 1.2.2　自下而上的分支

深层特征包含了更丰富的语义，有利于困难样本的识别。如图2所示，自下而上的连接将深层特征的高层次语义传递到了浅层，显著提高了浅层特征的识别能力，从而缓解了编码器和解码器特征图间的语义鸿沟。相比于FPN[17]，本文方法通过密集的跳跃连接，避免了深层特征逐层进行上采样时语义信息可能会发生的衰减问题。

为避免网络参数量显著增长，在实验中，本文先通过1×1卷积减少深层特征图的通道数，再利用双线性插值法来上采样该大尺度特征图，最终

将映射到同一层的特征图串联起来，并用一个额外的 1×1 卷积再次减少自下而上的分支特征图数量，以使 2 个分支对应的特征图的通道数相同。$\{\boldsymbol{S}_i^{(N)} \mid i=1,2,3,4;N=1,2\}$ 表示自下而上的分支所生成的特征图的集合，$i$ 为层索引，$N$ 为编码器的 2 个流，$\boldsymbol{S}_i^{(N)}$ 可表示为

$$\boldsymbol{S}_i^{(N)} = \begin{cases} U(C(\boldsymbol{R}_5^{(N)})) & i=4 \\ C([U(C(\boldsymbol{R}_5^{(N)})), U(C(\boldsymbol{S}_k^{(N)}))_{k=i+1}^{4}]) & i=1,2,3 \end{cases} \tag{2}$$

式中：$C(\cdot)$ 为 1×1 卷积；$U(\cdot)$ 为上采样操作（双线性插值）。

如图 3 所示，以 $\boldsymbol{S}_3^{(N)}$ 为例来描述如何构建自下而上的分支。首先，通过 1×1 卷积减少 $\boldsymbol{R}_5^{(N)}$ 的通道数量，再对其进行上采样得到 $\boldsymbol{S}_4^{(N)}$。然后，$\boldsymbol{S}_4^{(N)}$ 和 $\boldsymbol{R}_5^{(N)}$ 都通过 1×1 卷积进行降维，再上采样到与 $\boldsymbol{R}_3^{(N)}$ 相同的尺度并进行串联。最后，串联后的特征图通过 1×1 卷积再次降维，使 $\boldsymbol{S}_3^{(N)}$ 的通道数量与对应的 $\boldsymbol{R}_3^{(N)}$ 相同。

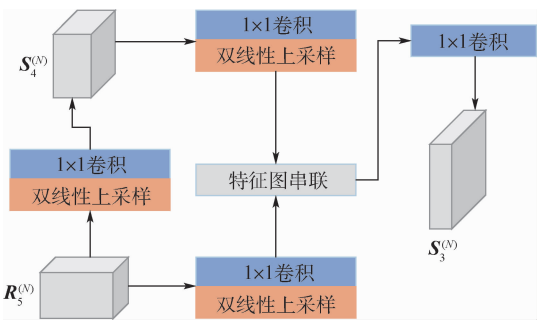



图 3 特征图 $\boldsymbol{S}_3^{(N)}$ 的构建

Fig. 3 Construction of feature map $\boldsymbol{S}_3^{(N)}$

### 1.2.3 横向连接

自上而下的分支提取了原始图像空间定位准确的特征，自下而上的分支则将高层次的语义传递到了浅层。为了融合这 2 个分支，本文将其对应的特征图逐元素的进行相加来作为编码器的每一条流的输出。$\{\boldsymbol{E}_i^{(N)} \mid i=1,2,3,4,5;N=1,2\}$ 表示编码器的每一条流所输出的特征图的集合，$i$ 为层索引，$N$ 为编码器的 2 个流，$\boldsymbol{E}_i^{(N)}$ 可表示为

$$\boldsymbol{E}_i^{(N)} = \begin{cases} \boldsymbol{R}_i^{(N)} + \boldsymbol{S}_i^{(N)} & i=1,2,3,4 \\ \boldsymbol{R}_5^{(N)} & i=5 \end{cases} \tag{3}$$

式中：$\boldsymbol{R}_i^{(N)}$ 和 $\boldsymbol{S}_i^{(N)}$ 分别为这 2 个分支所生成的对应的特征图。

通过横向连接，本文在所有尺度上都构建了具有精确位置表示的高层次语义特征图，从而促进了网络对小目标和目标边缘的检测，也提高了网络对伪变化的检测性能。

## 1.3 解码器

借鉴于孪生网络[9-11,18]，本文将编码器的2 个流中尺度相同的特征图串联在一起来提取差异信息。$\{\boldsymbol{E}_1,\boldsymbol{E}_2,\boldsymbol{E}_3,\boldsymbol{E}_4,\boldsymbol{E}_5\}$ 表示串联后的特征图集合，则 $\boldsymbol{E}_i$ 可表示为

$$\boldsymbol{E}_i = C([\boldsymbol{E}_i^{(1)},\boldsymbol{E}_i^{(2)}]) \quad i=1,2,3,4,5 \tag{4}$$

式中：$\boldsymbol{E}_i^{(1)}$ 和 $\boldsymbol{E}_i^{(2)}$ 分别为编码器的 2 个流生成的特征图；$i$ 为层索引；$C(\cdot)$ 为 1×1 卷积，用来减少串联后特征图的通道数量（实验中，本文将串联后特征图的通道数减少为原来的一半）。

解码器如图 4 所示，通过卷积操作，网络提取到了双时相图像间的差异信息，而上采样（反卷积）则逐层恢复了语义特征图的分辨率。$\boldsymbol{G}_1$、$\boldsymbol{G}_2$、$\boldsymbol{G}_3$、$\boldsymbol{G}_4$、$\boldsymbol{G}_5$ 分别表示解码器生成的不同尺度的特征图，其具有差异信息的多层次表示。

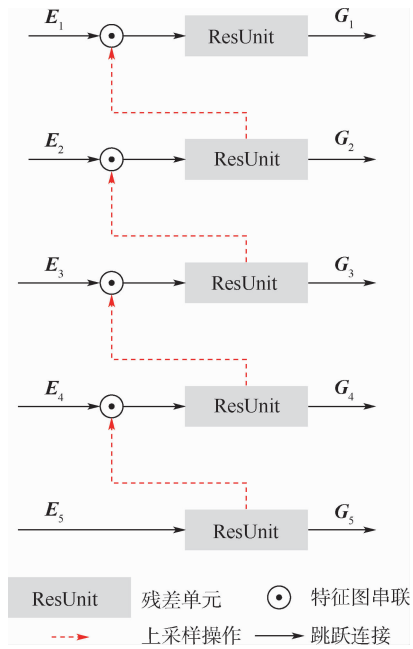



图 4 用于生成多尺度差异特征的解码器

Fig. 4 Decoder for generating multi-scale difference feature

## 1.4 分类器

不同对象适合被检测的尺度并不是完全相同的，小尺度特征图中包含了更多的细节信息，有利于道路、车辆等小块地物的检测；而大尺度特征图虽然边缘纹理等细节信息损失的比较严重，但有利于抑制对象内部的白斑和空洞现象，更适合于检测大片的农田或建筑物等目标。

为了提高网络对对象尺度变化的鲁棒性，本文设计了多尺度预测方法。如图 5 所示，解码器的每一层特征（$\boldsymbol{G}_1,\boldsymbol{G}_2,\boldsymbol{G}_3,\boldsymbol{G}_4,\boldsymbol{G}_5$）都通过 1×1

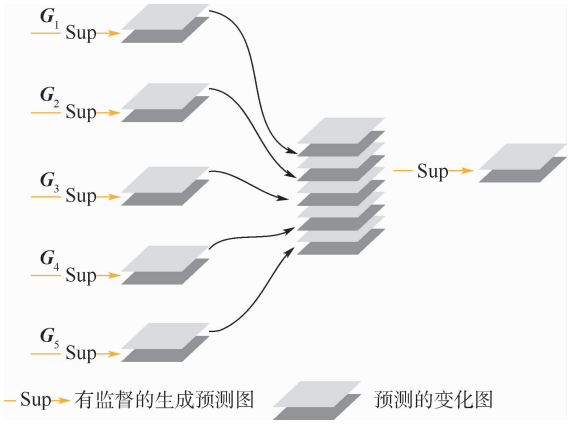



图 5　多尺度分类器

Fig. 5　Multi-scale classifier

卷积降到二维（变化或不变化）以生成不同尺度的预测图，再利用双线性插值法将所有大尺度的预测图上采样到跟原始图像相同（$\boldsymbol{G}_1$ 不进行上采样），并通过真值图进行有监督的训练。最终这些预测图被串联在一起，再额外利用一个 3×3 卷积压缩到二维，并利用真值图监督生成最终的预测图。受深监督的影响，分类器还加速了网络收敛，并促使解码器生成的特征图更具辨识性。获取预测图 **Map** 的流程为

$$\boldsymbol{Map} = \mathrm{Sup}([\mathrm{Sup}(\boldsymbol{G}_i)_{i=1}^{5}]) \tag{5}$$

式中：Sup(·) 表示将特征图压缩至二维并上采样（双线性插值）到与原始图像相同的尺度，通过真值图进行有监督的训练。

## 2　实验与分析

为了合理评价 FFANet，本文在 2 个公开的大尺度数据集上评估了其性能增益和参数量、计算量的额外开销，并将 FFANet 与其他先进方法作了对比。

### 2.1　实验数据

1）CDD[21]。通过裁剪和旋转 7 对随季节变化的图像生成了 10 000 对训练样本和 3 000 对验证及测试样本。分割后的图像大小均为 256×256 像素，空间分辨率为 3～100 cm/像素。

2）LEVIR[11]。共包含 637 对超高分辨率（50 cm/像素）的遥感图像，大小为 1 024×1 024 像素。受 GPU 内存容量限制，将每张图像不重叠地切分成 16 张 256×256 像素的小图块。

### 2.2　实验参数和评价指标

为了验证网络性能，FFANet 未经过预训练，并使用普通的交叉熵损失函数。具体参数为：学习率为 0.05，优化器 Adam，批大小为 16。在 NVIDIA Tesla v100 上训练了 100 轮，并最终使模型达到收敛。本文使用 3 个评价指标：精确率 $P$、回归率 $R$ 和 $F_1$ 分数，表达式分别为

$$P = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{6}$$

$$R = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{7}$$

$$F_1 = \frac{2PR}{P + R} \tag{8}$$

式中：TP 为真阳性的样本数量；FP 为假阳性的样本数量；FN 为假阴性的样本数量。

### 2.3　对比实验

本文与其他先进的变化检测方法作了对比，FC-EF[9]、FC-Siam-conc[9]、FC-Siam-diff[9] 是 3 种 U 型的全卷积神经网络，其中后 2 种是 U-Net 的孪生拓展。IFN[10] 在 U 型结构的基础上引入了注意力机制和多层次深监督来促进编码器和解码器的特征图更好的融合。SNUNet[11] 结合了孪生网络和 UNet++，并通过在编码器和解码器间及解码器和解码器间进行更紧凑的信息传输来增强深层特征的空间定位。DASNet[14] 采用了孪生网络加对比损失的架构，并利用双重自注意力机制来提升特征辨识性，以更鲁棒性的区分出变化。

在表 1 中，对比了 FFANet 与上述方法在 CDD 和 LEVIR 数据集上的参数量、计算量和性能指标。其中在 CDD 数据集上，FFANet 的精确率 $P$、回归率 $R$ 和 $F_1$ 分数分别为 0.962、0.957 和 0.960，相比于 SNUNet 分别提升了 0.006、0.008、0.007。在 LEVIR 数据集上，FFANet 的这 3 个指标分别为0.925、0.892 和 0.908，同样取得了最先进的效果。

此外，FFANet 的参数量和浮点运算次数分别为 $8.64\times10^6$ 和 28.81 GFLOPs，低于大部分的变化检测网络，有 3 方面的原因：①通过大量的1×1 卷积严格控制参数数量；②减少了网络整体的通道数量；③FFANet 基于 U 型架构，要比基于 UNet++ 的 SNUNet 更轻量。

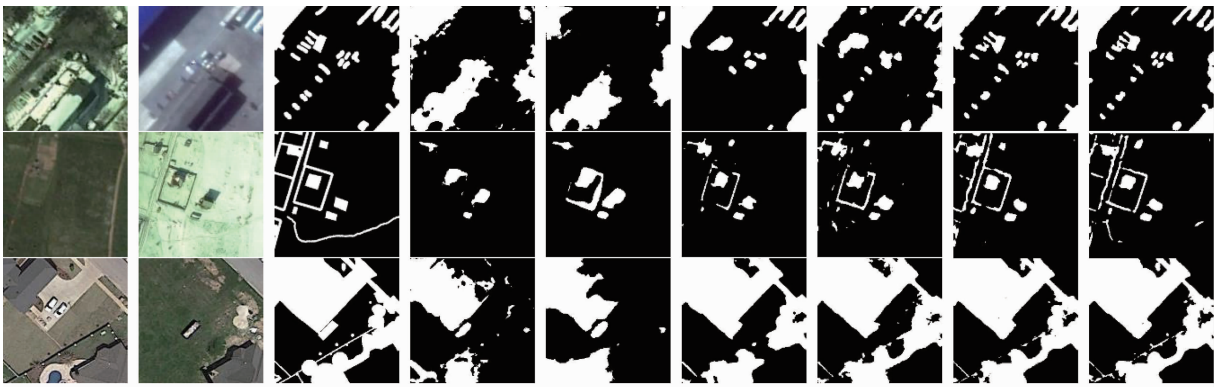
为了更直观地评估 FFANet，在图 6 中可视化了 FFANet、FC-Siam-conc、FC-Siam-diff、IFN、DASNet 和 SNUNet 在 CDD 测试集上的结果。可以观察到，FFANet 能更完整地分割出小目标和目标边缘，这是因为本文所提出的编码器和分类器提高了网络性能。编码器通过全尺度的特征聚合确保生成的所有尺度的特征图都语义丰富又定位准确，从而促进了对小目标和目标边缘的检测。分类器则通过结合不同尺度的预测图，进一步提高了网络对对象尺度变化的鲁棒性。

表 1 CDD 和 LEVIR 数据集上 FFANet 与其他方法的对比

Table 1 Comparison of FFANet with other methods on CDD and LEVIR datasets

| 方法 | 参数量/$10^6$ | 计算量/GFLOPs | CDD | | | LEVIR | | |
|---|---|---|---|---|---|---|---|---|
| | | | $P$ | $R$ | $F_1$ | $P$ | $R$ | $F_1$ |
| FC-EF | 1.35 | 7.14 | 0.749 | 0.494 | 0.595 | 0.754 | 0.730 | 0.742 |
| FC-Siam-conc | 1.55 | 10.64 | 0.779 | 0.622 | 0.692 | 0.852 | 0.736 | 0.790 |
| FC-Siam-diff | 1.35 | 9.44 | 0.786 | 0.588 | 0.673 | 0.861 | 0.687 | 0.764 |
| IFN | 35.72 | 164.53 | 0.950 | 0.861 | 0.903 | 0.903 | 0.876 | 0.889 |
| DASNet | 16.25 | 113.09 | 0.914 | 0.925 | 0.919 | 0.811 | 0.788 | 0.799 |
| SNUNet | 12.03 | 109.62 | 0.956 | 0.949 | 0.953 | 0.889 | 0.874 | 0.881 |
| FFANet | 8.64 | 28.81 | 0.962 | 0.957 | 0.960 | 0.925 | 0.892 | 0.908 |

注:GFLOPs 指 $10^9$ 次浮点运算。



(a) T0 (b) T1 (c) 真值图 (d) FC-Siam-conc (e) FC-Siam-diff (f) IFN (g) DASNet (h) SNUNet (i) FFANet

图 6 CDD 测试集上不同方法的可视化

Fig. 6 Visualization of different methods on CDD test set

## 2.4 模块间的消融实验

本节对所提出的编码器和分类器作了消融实验,所有实验的超参数设置完全相同。如表 2 所示,表中编码器和分类器指本文所提出的编码器和分类器。当编码器所在列没有“√”时,使用普通编码器(见图 7)代替全尺度特征聚合的编码器。当分类器所在列没有“√”时,只对解码器特征 $\boldsymbol{G}_1$ 进行有监督的训练来获取最终的预测图。

从表 2 可以看出,本文所提的编码器仅增加了 $1.01\times10^6$ 参数量,就将 $F_1$ 分数提升了 0.025,这说明充分利用原始图像特征能显著提升网络性能。此外,本文所提出的分类器在几乎不增加参数量的情况下,将 $F_1$ 分数提升了 0.024。这可归结于 2 个因素:①多层次的深监督促使解码器生成了更具辨识性的特征表示;②多尺度预测图的结合提高了网络对对象尺度变化的鲁棒性。

表 2 CDD 数据集上的消融实验

Table 2 Ablation experiments on CDD data set

| 序号 | 编码器 | 分类器 | 参数量/$10^6$ | $P$ | $R$ | $F_1$ |
|---|---|---|---|---|---|---|
| ① | × | × | 7.63 | 0.955 | 0.900 | 0.926 |
| ② | √ | × | 8.64 | 0.960 | 0.942 | 0.951 |
| ③ | × | √ | 7.63 | 0.957 | 0.943 | 0.950 |
| ④ | √ | √ | 8.64 | 0.962 | 0.957 | 0.960 |

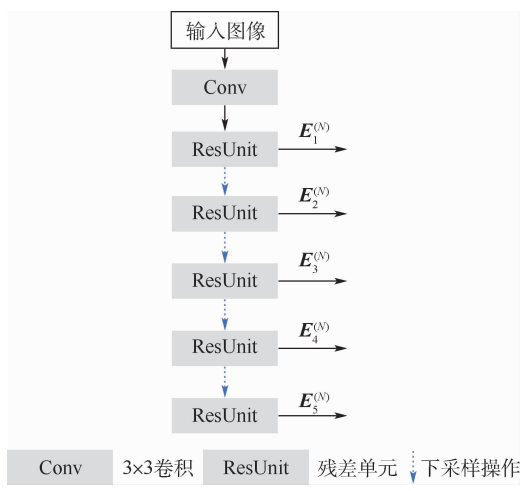



图 7 普通的编码器

Fig. 7 Plain encoder

图 8 可视化了表 2 所对应的测试集的实验结果。可以观察到,全尺度特征聚合的编码器和多尺度预测的分类器都有利于生成更清晰的边缘,并促进了对小目标的检测。然而,从图 8 中第 1 列和第 3 列可看出,尽管全尺度聚合的编码器整体上提升了检测性能,但预测结果中多出了一些“假阳性”样本。经过比对发现,这些虚假变化主要出现在变化区域边缘或在图像边缘,这可能是由于变化区域和非变化区域相邻像素的特征在

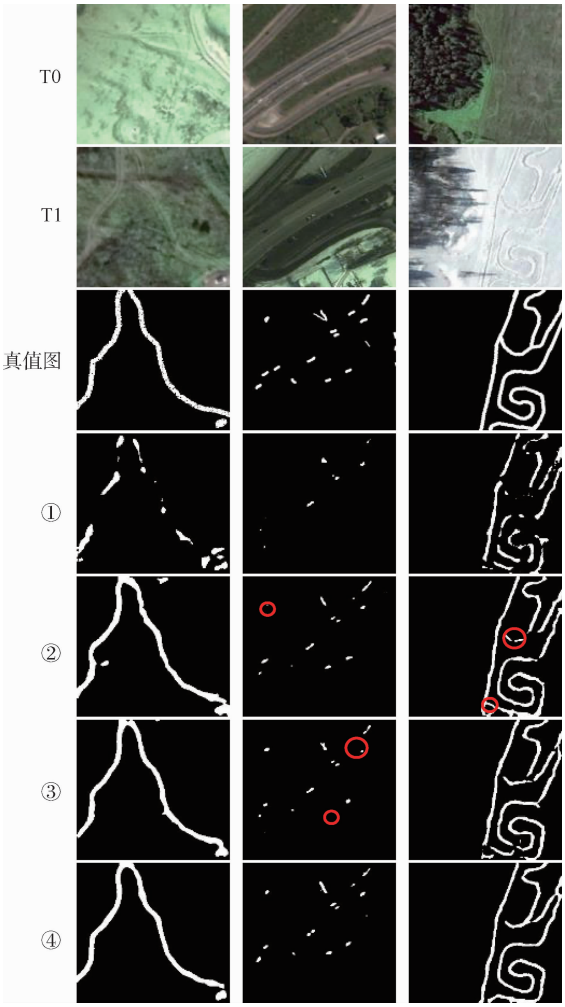



图 8　编码器和分类器的对比实验

Fig. 8　Comparison experiments of encoder and classifier

深层中被混合到了一起，编码器自下而上的分支将这些特征传递向浅层时，变化像素为非变化像素贡献了一些描述错误的语义信息。

在图 8 中，错误的分类用圆圈作了标记。从第 2 列、第 3 列中可以观察到，所提出的编码器和分类器在结合使用后，一些彼此预测错误的目标能同时被纠正回来，说明这 2 个模块在网络中所补充的信息并不是完全相同的，两者间存在着很强的互补性，因此建议这 2 个模块结合在一起来使用。

## 3　结　论

1）提出了一个全尺度特征聚合网络，仅用 $8.64\times10^6$ 的参数量和 28.81 GFLOPs 浮点运算次数就在 CDD 和 LEVIR 数据集上取得了最先进的性能，相比于其他方法有较大优势。

2）所提网络通过对特征图的充分利用，显著提高了对小目标和目标边缘的检测性能，并通过结合多尺度预测图提升了对对象尺度变化的鲁棒性。

3）单独使用改进的编码器会产生了一些“假阳性”样本，影响网络的精确率，但将改进的编码器和分类器结合在一起使用将有利于消除这一负面影响。

## 参考文献（References）

[1] LEICHTLE T, GEIß C, LAKES T, et al. Class imbalance in unsupervised change detection-A diagnostic analysis from urban remote sensing[J]. International Journal of Applied Earth Observation and Geoinformation, 2017, 60: 83-98.

[2] USEYA J, CHEN S B, MUREFU M. Cropland mapping and change detection: Toward zimbabwean cropland inventory[J]. IEEE Access, 2019, 7: 53603-53620.

[3] QIAO H J, WAN X, WAN Y C, et al. A novel change detection method for natural disaster detection and segmentation from video sequence[J]. Sensors (Basel), 2020, 20(18): 5076.

[4] RONG K, FANG B, CHEN G, et al. Progressive domain adaptation for change detection using season-varying remote sensing images[J]. Remote Sensing, 2020, 12(22): 3815.

[5] VU V T, PETTERSSON M I, MACHADO R, et al. False alarm reduction in wavelength-resolution SAR change detection using adaptive noise canceler[J]. IEEE Transactions on Geoscience and Remote Sensing, 2017, 55(1): 591-599.

[6] BOVOLO F, BRUZZONE L. A novel theoretical framework for unsupervised change detection based on CVA in polar domain [C]// 2006 IEEE International Symposium on Geoscience and Remote Sensing. Piscataway: IEEE Press, 2006: 379-382.

[7] DENG J S, WANG K, DENG Y H, et al. PCA-based land-use change detection and analysis using multitemporal and multisensor satellite data[J]. International Journal of Remote Sensing, 2008, 29(15-16): 4823-4838.

[8] LI W, LU M, CHEN X W. Automatic change detection of urban land-cover based on SVM classification[C]// 2015 IEEE International Symposium on Geoscience and Remote Sensing. Piscataway: IEEE Press, 2015: 1686-1689.

[9] DAUDT R C, SAUX B L, BOULCH A. Fully convolutional Siamese networks for change detection[C]// IEEE International Conference on Image Processing (ICIP). Piscataway: IEEE Press, 2018: 4063-4067.

[10] ZHANG C Z, PENG Y, TAPETE D, et al. A deeply supervised image fusion network for change detection in high resolution bitemporal remote sensing images[J]. ISPRS Journal of Photogrammetry and Remote Sensing, 2020, 166: 183-200.

[11] FANG S, LI K Y, SHAO J Y, et al. SNUNet-CD: A densely connected Siamese network for change detection of VHR images [J]. IEEE Geoscience and Remote Sensing Letters, 2021, 19: 1-5.

[12] ZHOU Z, SIDDIQUEE M, TAJBAKHSH N, et al. UNet + +: A nested U-Net architecture for medical image segmentation[C]// Deep Learning in Medical Image Analysis (DLMIA) Workshop, 2018: 3-11.

[13] CHEN H, SHI Z W. A spatial-temporal attention-based method

and a new dataset for remote sensing image change detection [J]. Remote Sensing, 2020, 12(10): 1662.

[14] CHEN J, YUAN Z Y, PENG J, et al. DASNet: Dual attentive fully convolutional siamese networks for change detection of high resolution satellite images [J]. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2020, 14: 1194-1206.

[15] RONNEBERGER O, FISCHER P, BROX T. U-Net: Convolutional networks for biomedical image segmentation [C]// Medical Image Computing and Compwter-Assisted Intervertion-MICCAI 2015, 2015.

[16] SHELHAMER E, LONG J, DARRELL T. Fully convolutional networks for semantic segmentation [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(4): 640-651.

[17] LIN T Y, DOLLAR P, GIRSHICK R, et al. Feature pyramid networks for object detection [C]// 2017 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2017: 936-944.

[18] BAO T E, FU C Q, FANG T, et al. PPCNET: A combined patch-level and pixel-level end-to-end deep network for high-resolution remote sensing image change detection [J]. IEEE Geoscience and Remote Sensing Letters, 2020, 17(10): 1797-1801.

[19] HUAN R, ZHOU M, XING Y, et al. Change detection with various combinations of fluid pyramid integration networks [J]. Neurocomputing, 2021, 437: 84-94.

[20] YANG K, LIU Z, LU Q, et al. Multi-scale weighted branch network for remote sensing image classification [C]// 2010 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2019.

[21] LEBEDEV M A, VIZILTER Y V, VYGOLOV OLEG, et al. Change detection in remote sensing images using conditional adversarial networks [J]. The International Archives of the photogrammetry, Rewote Sensing and Spatial Information Sciences, 2018, 48(2): 565-571.

# A full-scale feature aggregation network for remote sensing image change detection

LIU Guoqiang, FANG Sheng*, LI Zhe

(College of Computer Science and Engineering, Shandong University of Science and Technology, Qingdao 266590, China)

**Abstract**: Change detection (CD) is an important task of remote sensing, always facing many pseudo changes and large scale variations. However, existing methods mainly focus on modeling difference features and neglect extracting sufficient information from the original images, which affects feature discrimination and makes it difficult to distinguish change regions stably. To address these problems, a full-scale feature aggregation network (FFANet) is proposed to make fuller use of the original image features, which drives the generated feature representations to be semantically richer and spatially more precise, thus improving the network's detection performance for small targets and target edges. Deep supervision is also extended to combine multi-scale prediction maps to drive the detection of different objects at more appropriate scales, thus improving the robustness of the network to object scale variations. On the CDD dataset, our proposed method improves the $F_1$-score by 0.034 compared to the baseline network by increasing the number of parameters by only $1.01 \times 10^6$.

**Keywords**: change detection (CD); deep supervision; full-scale feature aggregation; multi-scale prediction; remote sensing images

**Received**: 2021-09-06; **Accepted**: 2021-10-01; **Published online**: 2021-10-29 16:46

**URL**: kns.cnki.net/kcms/detail/11.2625.V.20211028.1950.004.html

**Foundation item**: Shandong Provincial Natural Science Foundation (ZR2020MF132)

* **Corresponding author**. E-mail: fangsheng@tsinghua.org.cn

http://bhxb.buaa.edu.cn jbuaa@buaa.edu.cn
**DOI**: 10.13700/j.bh.1001-5965.2021.0549

# 基于改进空间通道信息的全局烟雾注意网络

董泽舒[1], 袁非牛[1,*], 夏雪[2]
(1. 上海师范大学 信息与机电工程学院, 上海 201418; 2. 江西财经大学 信息管理学院, 南昌 330032)

**摘 要**: 针对烟雾因半透明、形状不规则和边界模糊造成分割困难的问题,提出了基于注意力机制的长距离信息建模方法,以提取长距离像素间的依赖和连续性关系。通过注意力机制作用原理,解决孤立小块区域误分类问题,减少非连续区域的烟雾误判。为避免注意力网络大尺寸矩阵运算造成的内存和计算负担,对空间和通道2种注意力方式进行改进,分别设计了双向定位空间注意力(BDA)模块和多尺度通道注意力(MSCA)融合模块,弥补现有注意力全局池化操作导致的大量空间信息丢失。将所提注意力模块和残差深度网络合并,构建面向图像烟雾分割的全局烟雾注意网络,在尽可能不丢失全局信息相关性的同时减少内存消耗。实验结果表明:所提网络在DS01、DS02、DS03合成烟雾测试集上,取得的平均交并比分别为73.13%、73.81%、74.25%,总体上优于对比算法。

**关 键 词**: 烟雾分割; 双向定位; 空间注意力; 多尺度融合; 通道注意力
**中图分类号**: TP391
**文献标志码**: A **文章编号**: 1001-5965(2022)08-1471-09

在各种灾害中,火灾是最普遍威胁公众安全和社会发展的主要灾害之一。及时准确地检测火灾的发生,能够极大程度地降低其对人们生产生活可能造成的危害。对火灾的检测一般分为烟雾检测和火焰检测2种。在实际场景中,火焰的燃烧会产生大量烟雾,这些烟雾相较于火焰具有更大的流动性,其以火源为中心向外扩散,浓度随气压和风向等多种因素影响,有更大的可观测面积;烟雾在扩散过程中会遮盖住部分火焰,影响火焰检测的效果。因此,烟雾检测相较于火焰检测对于火灾的控制更加有效[1-3]。

随着深度学习的广泛发展,越来越多的神经网络结构被用于图像识别、目标检测或语义分割。烟雾分割是一种密集的二分类任务,基于深度学习的烟雾语义分割能够逐像素地识别烟雾,即可精确指示其位置,这对于辅助消防员或消防机器人分析安全与危险区、判断火灾大小、评估救援方案有着重要意义。

烟雾的分割应致力于在尽可能保证精确度的同时,最大限度地减少网络参数和内存消耗,从而在实时视频烟雾分割中保证最小的延时。目前,已经有很多网络[4-5]实现了端到端的像素级烟雾分割。

由于烟雾具有不透明度,越靠近边缘的烟雾像素分割的难度越大。如何捕捉用于分类的语义信息和定位的上下文信息,成为需要着重考虑的问题。网络深度的增加会捕捉到更多语义信息,但同时也丢失了局部细节信息。解决两者间矛盾的方法主要分为以下3种:①采用空洞卷积的方式[6]捕捉更大尺度的信息;②用多尺度融合[7]的方式对不同尺度的信息进行整合;③采用深浅层级之间的跳跃连接(skip-connection)结构和逐步上采样的方式[8]避免特征的丢失。

此外,由于烟雾形状不规则,且具有扩散和连

**收稿日期**: 2021-09-14; **录用日期**: 2021-10-01; **网络出版时间**: 2021-10-28 15:31
**网络出版地址**: kns.cnki.net/kcms/detail/11.2625.V.20211027.2052.002.html
**基金项目**: 国家自然科学基金(61862029,62062038); 江西省教育厅课题(GJJ201117)
* **通信作者**. E-mail: yfn@ustc.edu

**引用格式**: *董泽舒, 袁非牛, 夏雪. 基于改进空间通道信息的全局烟雾注意网络[J]. 北京航空航天大学学报, 2022, 48(8): 1471-1479. DONG Z S, YUAN F N, XIA X. Improved spatial and channel information based global smoke attention network [J]. Journal of Beijing University of Aeronautics and Astronautics, 2022, 48(8): 1471-1479 (in Chinese).*

续的特性,分割结果图中与主体烟雾区域距离较远的烟雾孤立小区域多为误判。为避免此类误判,需建模长距离像素之间的依赖和连续性关系。

近年来,随着注意力机制[9]的发展,图像的语义分割被赋予了新的生机。空间注意力模块[10]通过对全局所有像素进行相关性建模,能够获得全局像素点的相关性,在分割任务过程中,对当前像素点的分类能够建模全局像素信息,从而有效解决距离主体烟雾区域距离较远的孤立区域被误判的问题。通道注意力模块能够利用空间信息聚合判断通道的重要性。两者结合能够互相补充缺失信息,进一步优化分割效果。通道注意力[11-13]往往伴随着全局池化或者矩阵运算。前者舍弃了绝大多数的空间信息,用此时的通道权重对特征图进行注意力提取可能不够合理,因为丢失的空间信息可能对通道信息的收集有负面影响;后者涉及较大维度的矩阵运算,不利于实时性分割。

为解决以上注意力的信息丢失与内存消耗问题,本文提出了一种新型的注意力网络结构——全局烟雾注意网络(global smoke attention network,GSA-Net)。该网络包含 2 个模块:沿图像长、宽方向捕获空间信息的双向定位空间注意力(BDA)模块和减少空间信息丢失的多尺度通道注意力(MSCA)融合模块。实验表明,本文网络能够在快速分割的条件下有效避免分割结果中出现主体烟雾区域距离较远的噪点区域,并减少烟雾像素的误分类。

# 1 相关工作

## 1.1 注意力机制

注意力机制的首次提出是在自然语言处理领域,被应用于机器翻译,用于捕获长距离的词向量依赖。Wang 等[10]将其迁移到图像领域,设计了非局部操作(non-local)模块,有效捕捉了空间信息全局特征依赖。Fu 等[14]在此基础上考虑到通道信息对图像分割的影响,设计了双注意力网络(dual attention network),捕获了空间维度和通道维度的注意力并进行特征融合。与双注意力模块相似,Yuan 等[15]设计了目标上下文网络(object context network),捕获空间和通道维度的注意力。虽然以上方法能够有效建模全局特征,但随着矩阵运算产生的内存消耗是巨大的。一些方法致力于减少注意力机制中矩阵运算所造成的内存负担,如十字交叉注意力网络[16](criss-cross attention network)利用覆盖整图的“十字型”小区域注意力计算替代全局注意力计算,大幅减小了计算消耗;SENet[11](squeeze-and-excitation network)、BAM[13](bottleneck attention module)、CBAM[12](convolutional block attention module)避免了大尺寸的矩阵运算。然而 SENet 网络只嵌入了通道重要性权重,忽略了空间信息的重要性。BAM 和 CBAM 通过在通道上进行全局池化来引入位置信息,但只能捕获局部信息,无法获取长范围依赖的信息。Hou 等[17]提出了一种为轻量级网络设计的注意力机制,通过将位置信息嵌入通道注意力中形成坐标注意力(coordinate attention),为注意力模块的设计提供了新的思路。

## 1.2 烟雾分割

传统的烟雾分割采用机器学习算法,需要手动提取部分特征,张娜等[18]计算像素在颜色上的上近似与下近似,获得粗糙度直方图,并利用直方图中波峰分布的信息自适应地选取阈值进行粗分割,结合运动检测与颜色统计规则获得最终烟雾分割图。Lin 等[19]提取烟雾序列上的体积纹理特征(volume LBP),提出基于采样块的烟雾检测方法。Filonenko 等[20]采用基于颜色的二次烟雾分割,并加入边缘粗糙度特征来区分与烟雾具有相同颜色的非烟物体,以此减少误报。

深度学习避免了手动提取特征,能够更加高效地训练出高准确度模型。Tao 等[21]采用 AlexNet[22]进行了烟雾识别。Yin 等[23]提出了一种 DNCNN 网络实现烟雾识别。Yuan 等[5]提出一种波浪形编解码结构,实现了烟雾的硬分割和浓度检测。在最新的研究中,Yuan 等[24]将注意力嵌入门控循环单元(GRU),用于长距离特征关系获取,并设计了密集金字塔池化模块为上采样提供多尺度上下文信息,该网络在烟雾语义分割上取得了良好效果。

# 2 烟雾分割网络

针对烟雾检测过程中对精度和实时性的要求、烟雾区域连续成片状的特性、边缘像素点不易分割的特性,本文通过改进空间注意力和通道注意力,分别设计了双向定位空间注意力模块和多尺度通道注意力融合模块,并将两者并联,提出一种能够在尽可能不丢失全局信息相关性的同时减少参数量和内存消耗的新型注意力结构,结合 ResNet50 网络[25]设计了能够有效分割烟雾图像的全局烟雾注意网络。

## 2.1 双向定位空间注意力模块

空间上下文的嵌入在语义理解中扮演着重要

的角色。其中,注意力机制的提出无疑为空间信息的提取开辟了新的方向。目前的空间注意力机制大多采用矩阵运算,用于捕捉全局作用域中任意 2 个像素点的彼此关系,但随之而来的是内存的巨大消耗和高计算量。本文设计了一种双向定位空间注意力模块,如图 1 所示。$H$、$W$、$C$ 分别表示图片的高度、宽度、通道数,首先,对输入尺寸为 $H\times W\times C$ 的特征图,沿着高、宽 2 个方向进行全局平均池化,分别获得尺寸为 $H\times 1\times 1$ 和 $1\times W\times 1$ 的 2 个特征向量,2 个向量通过矩阵乘法生成 $H\times W\times 1$ 的特征图。分别得到沿高度方向和宽度方向的 2 个一维注意力向量和 1 个高宽之间相关信息的二维注意力面特征图。然后,采用 Sigmoid 激活函数获取这 2 个一维特征向量和二维特征图的相关性权重,依次作为原图的高、宽、双向注意力权重向量,并通过逐点相乘的方式加权输入特征图。最后,3 幅经过加权得到的特征图通过逐点相加生成双向定位空间注意图。

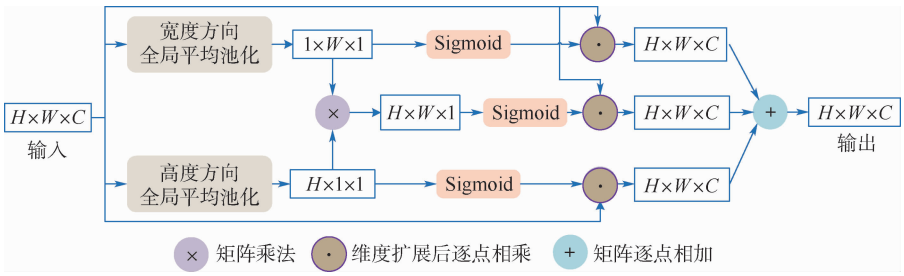



图 1　双向注意力模块

Fig. 1　Bi-direction attention model

不同方向的 2 个向量的加权注意力能够分别捕捉到特征图沿着一个空间方向的长距离依赖,并保存沿着另一个空间方向的精确位置信息,通过横纵坐标的方式定位到特征图的空间像素点。二维特征图的加权能够整合 2 个空间方向的相关信息,从而建立空间关系映射。本文通过 2 个方向的全局平均池化分别得到长维度和宽维度的一维权重向量,并通过矩阵运算得到平面维度的二维权重图,分别用得到的 2 个权重向量和 1 个权重图对原图进行加权。通过将加权过后的 3 个特征图融合,能够在不经过复杂矩阵运算的条件下实现全局尺度的空间信息捕捉。通过注意力机制,对当前像素的类别判定依赖于全局像素的相关性。因此,能够避免分割结果烟雾区域不连续、距离大片烟雾区域较远的区域存在孤立噪点的情况。

### 2.2　注意力复杂度分析

本文改进的注意力模块避免了大尺寸矩阵运算导致的内存和计算负担,本节进一步对内存减少情况进行量化分析。

传统注意力机制采用矩阵加权为全局相关性建模。对输入尺寸为 $H\times W\times C$ 的特征图通过 reshape 得到 $HW\times C$ 的基准图,并通过与其自身转置进行矩阵乘法和 softmax 激活层,得到尺寸为 $HW\times HW$ 的权重矩阵。再经过与基准图进行矩阵乘和加权,得到最终尺寸为 $H\times W\times C$ 的特征图。本文通过不同方向的全局池化后再加权,将内存消耗从 $HW\times HW$ 降低到 $H\times W$,因此大幅减少了内存和计算复杂度的负担。

### 2.3　多尺度通道注意力融合模块

现有通道注意力的获取依赖于全局池化后的一维特征向量,该向量在每个通道上只保留了一个值,丢失了大部分输入特征图的信息,用其来建模整张特征图的通道重要性不够全面。本文提出了一种多尺度通道注意力融合模块,以在不同空间尺度范围内对通道进行加权,提升注意力模块的有效性。

首先,将输入的特征图通过空间金字塔池化模块[26]提取不同尺度的特征,经过池化后的特征图空间高和宽分别为 $1\times 1$、$2\times 2$、$3\times 3$、$6\times 6$。同时,将输入特征图经过全局平均池化、全连接层和 Sigmoid 激活层,生成通道注意力权重向量。然后,分别对 4 个尺度的特征图采用注意力权重矢量进行加权,这种空间尺寸接近的加权能够更好发挥通道权重的作用。经过加权的多尺度特征图通过卷积、ReLU 激活和双线性插值恢复到输入特征图的尺寸,并在通道维度连接后通过卷积、BN、ReLU 激活降低通道数量,生成与输入尺寸相同的特征图。最后,与权重向量进行二次加权,得到输出特征图。整体流程如图 2 所示。可以看出,注意力权重向量被充分共享多次。相比于传统的注意力机制,本文注意力机制通过对不同尺度的特征图进行通道加权和共享,能够更好地捕获通道维度的依赖关系,从而辅助优化分割效果。

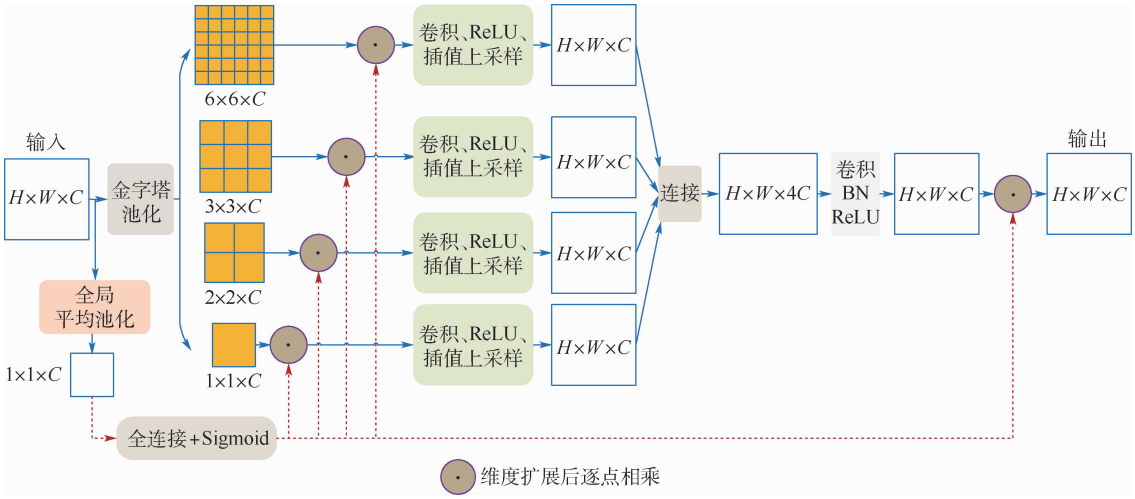



图 2 多尺度通道注意力融合模块

Fig. 2 Multi-scale channel attention fusion model

### 2.4 整体网络结构

图 3 为本文网络结构,采用 ResNet50 残差结构[26]作为骨干网络。骨干网络对输入的 RGB 图像进行特征提取,将提取的特征图分别送入双向定位空间注意力模块和多尺度通道注意力融合模块,将这 2 个模块输出结果以并联方式融合,从而整合空间注意力和通道注意力信息,融合结果经过卷积进一步提取信息并上采样至原图大小。

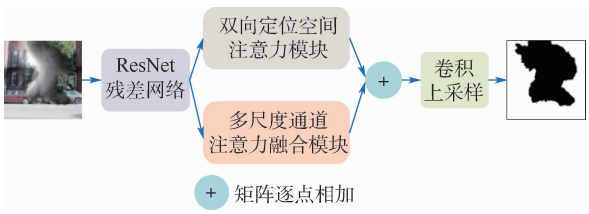



图 3 全局烟雾注意力网络

Fig. 3 Global smoke attention network

## 3 实验及结果分析

### 3.1 数据集及实验细节

本文用于训练和测试的数据集与文献[5]相同,包括 1 个含 70 632 张图像的虚拟合成烟雾训练集和 3 个含 1 000 张图像的虚拟合成烟雾测试集(DS01、DS02、DS03)。针对烟雾的透明性、随机性及流体特性,由计算机图形学原理生成 8 162 张形状随机的纯烟雾图像。每一个纯烟雾样本都是一张 256×256 像素的 RGBA 图像,分别包含表示颜色的 RGB 通道 $\boldsymbol{S}$ 和一个表示不透明度的 $\alpha$ 通道,其中 $\alpha$ 的取值范围为[0,1]。按照线性合成的规则,纯烟雾图像 $\boldsymbol{S}$ 与背景 $\boldsymbol{B}$ 的组合生成观察图像 $\boldsymbol{I}$,数学表示为

$$\boldsymbol{I} = \alpha\boldsymbol{S} + (1-\alpha)\boldsymbol{B} \tag{1}$$

通过上述方法,能够构造大数目的训练集而不用繁琐地制作标签。虚拟烟雾与真实场景背景图像的随机线性合成如图 4 所示。这些生成的虚拟数据在烟雾的动态特征、颜色、大小、背景复杂程度等方面具有多样性,能模拟出大多数真实烟雾场景。

采用 Python 和 Tensorflow 实现本文网络模型,采用在 ImageNet 上预训练的 ResNet50 作为骨干网络,最后两阶段空洞率为 2 和 4,输出特征图尺寸为原图的 1/8。采用随机梯度下降算法在英伟达 GeForce GTX2080Ti 显卡上进行训练,学习率设置为 0.002 5,动量设置为 0.9,权重衰减设置为 0.001。

图 4 虚拟合成数据集图例

Fig. 4 Samples from virtually synthesized datasets

### 3.2 对比实验

为了证明本文网络结构的有效性,在 3 个虚

拟测试集上和 1 个真实场景测试集上进行测试,并与已有的深度学习语义分割算法在烟雾分割上进行对比。出于公平考虑,本文未与文献[18-21]所提到的传统烟雾识别算法对比。由于目前利用深度学习实现烟雾语义分割的论文较少,本文除了对比烟雾分割网络之外,还对比了在其他数据集上取得良好效果的经典网络模型。对比网络包括 FCN-8S[27]、SegNet[28]、SMD[29]、TBFCN[7]、DeepLab v1[30]、ESPNet[31]、LRN[32]、DSS[4]、HG-Net2[33]、HG-Net8[33]、MS-Net[3] 和 W-Net[5]。所有模型都采用合成虚拟烟雾训练集进行训练。

本文模型在 3 个虚拟烟雾测试集上的表现如表 1 所示,评价指标采用平均交并比 mIoU。在 DS02 测试集中,本文模型的表现取得了与 W-Net 接近的效果,在 DS01 和 DS03 测试集上达到了最好精度。本文网络为一次编解码结构,相较于经过 2 次编解码和多次通道连接的 W-Net 模型的复杂度更低,能够达到更好的实时检测效果。

测试的实际效果如图 5 所示。图中第 1 列和第 2 列为原始图像和标签图像,其余列为不同算法的分割效果。通过对实际效果的分析可以看出,在大范围烟雾图像中,本文方法将烟雾按照单块状分割,极少出现不与主要区域连接的误分类噪点,说明本文模型能够有效捕捉长距离依赖,优化边缘的分割效果,通过建模通道间的相互关系避免误分类。

在真实数据集上进行的进一步实验如图 6 所示。图中第 1 列为真实图像,其余列为不同算法的分割结果,这些真实图像没有对应的标签图,可

**表 1　不同算法对比结果**

**Table 1　Comparison for different algorithms**

| 算法 | mIoU/% | | |
|---|---|---|---|
| | DS01 | DS02 | DS03 |
| FCN-8S[27] | 64.03 | 63.28 | 64.38 |
| SegNet[28] | 56.94 | 56.77 | 57.18 |
| SMD[29] | 62.88 | 61.50 | 62.09 |
| TBFCN[7] | 66.67 | 65.85 | 66.20 |
| DeepLab v1[30] | 68.41 | 68.97 | 68.71 |
| ESPNet[31] | 61.85 | 61.90 | 62.77 |
| LRN[32] | 66.43 | 67.71 | 67.46 |
| DSS[4] | 71.04 | 70.01 | 69.81 |
| HG-Net2[33] | 63.58 | 62.40 | 63.61 |
| HG-Net8[33] | 63.85 | 63.27 | 64.46 |
| W-Net[5] | 73.06 | 73.97 | 73.36 |
| 本文 | 73.13 | 73.81 | 74.25 |

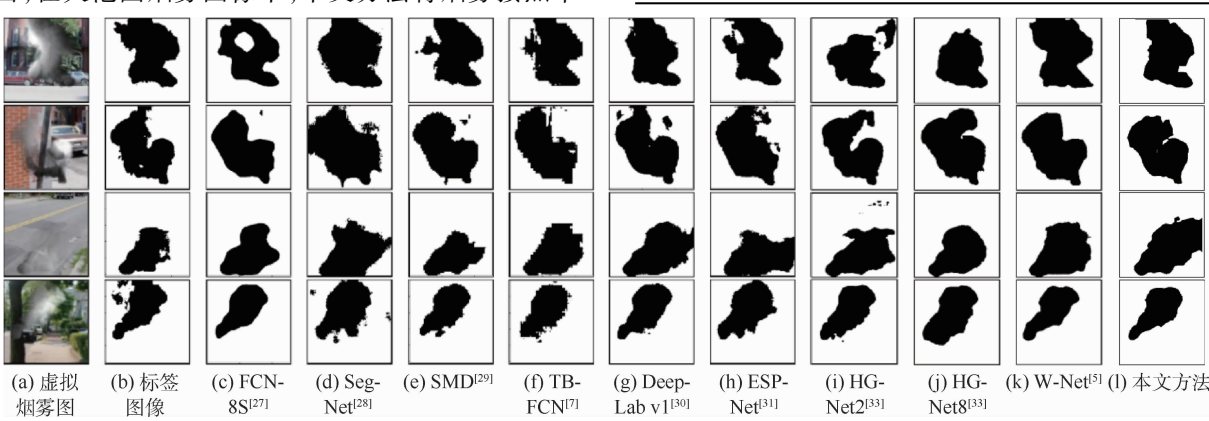



图 5　虚拟烟雾测试集分割结果

Fig. 5　Segmented results of virtual smoke test datasets

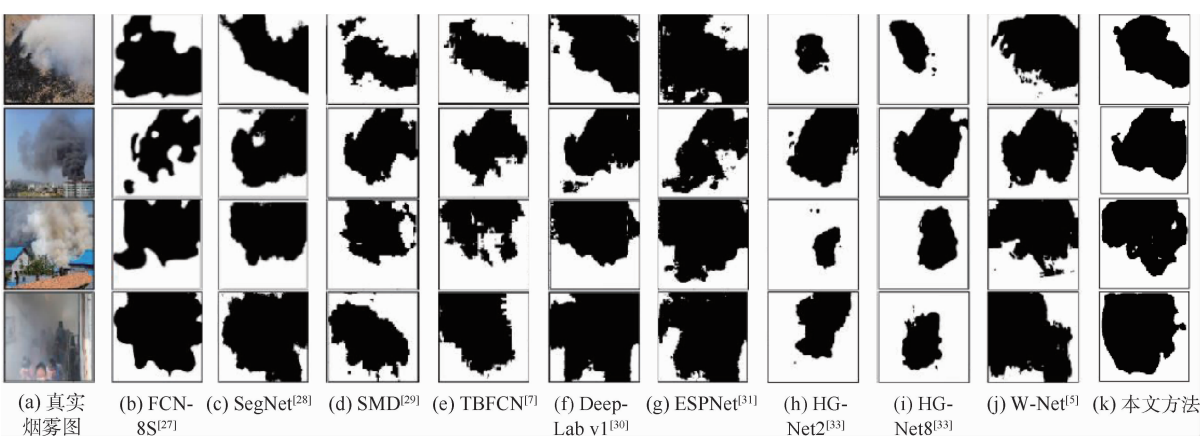



图 6　真实图像分割结果

Fig. 6　Segmented results for real images

通过肉眼观察比较结果,实验再次印证了本文模型能够有效减少烟雾分割中偏离主要烟雾区域的不合常理噪点,第 2、3 行图像中,本文模型的优势尤为明显。

### 3.3 剥离实验

为了验证模型每一个部分的重要性,分别得到本文方法的 3 种变体。图 7(a)为只保留双向定位空间注意力模块,记为 ResNet + BDA。图 7(b)显示了只保留多尺度通道注意力融合模块的变体网络结构,记为 ResNet + MSCA。图 7(c)为 2 个模块均保留且串行连接,ResNet 提取特征后先经过双向定位空间注意力模块,再经过多尺度通道注意力融合模块,记为 ResNet + MSCA 串联 BDA。本文方法实际上是 2 个模块均保留且并行连接,ResNet 提取特征后同时经过双向定位空间注意力模块和多尺通道注意力融合模块,并通过逐点相加方式连接,如图 3 所示。考虑到避免网络的大量参数消耗,本文没有将连接方式由逐点相加方式修改为通道维度连接再降维的对比模型。表 2 所示的实验结果表明,2 个注意力模块都有效改善了分割效果,并联方式连接 2 个模块的效果好于串行方式依次通过 2 个模块。

为了进一步验证注意力模块的有效性,本文对注意力加权后的特征图进行了可视化,如图 8 所示。可以看出,烟雾存在的区域已经能够被赋予更高的注意力权重,经过模型后续的卷积和逐步上采样,局部信息会被进一步精细化,从而进一步优化分割结果。

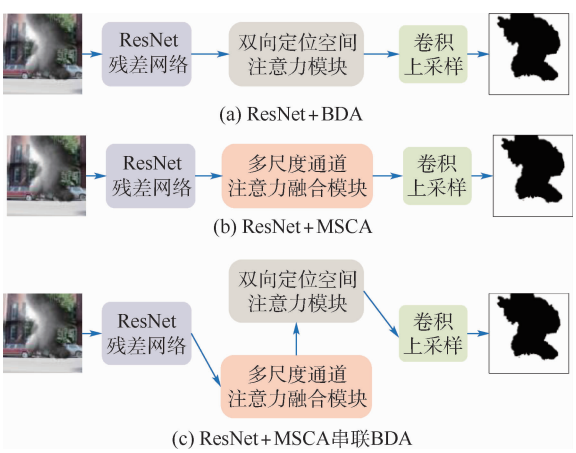



图 7 本文方法的变体

Fig. 7 Variants of the proposed method

**表 2 剥离实验效果**

**Table 2 Ablation experimental results**

| 网络结构变体 | mIoU/% | | |
|---|---|---|---|
| | DS01 | DS02 | DS03 |
| ResNet + BDA | 71.61 | 72.45 | 72.89 |
| ResNet + MSCA | 70.12 | 71.79 | 72.11 |
| ResNet + MSCA 串联 BDA | 72.49 | 73.26 | 73.98 |
| ResNet + MSCA 并联 BDA (本文方法) | 73.13 | 73.81 | 74.25 |

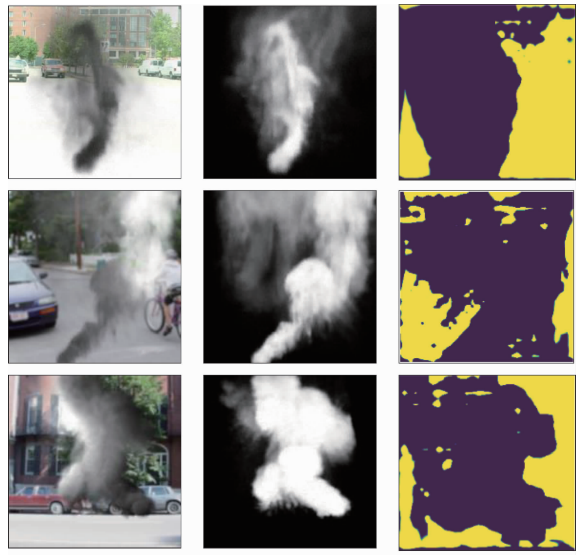



图 8 注意力机制加权后的特征图

Fig. 8 Weighted feature maps by attention mechanism

### 3.4 场景实验

在真实烟雾场景中,时常会出现更具挑战性的情形,如雾和云也具有与烟雾相似的特性,可能会造成混淆。在低质量图像及复杂背景下,烟雾分割的难度也会随之加大。为验证本文方法的适应性,设计了在天空背景及复杂场景下的烟雾分割实验,实验结果如图 9 所示。

第 1 组图像对比了蓝色天空下对于黑色烟(见图 9(a))、白色浓烟(见图 9(b))、白色淡烟(见图 9(c))的适应性效果,实验表明,本文模型能够在无云天气良好情况下分割出不同颜色及浓度的烟雾。其中,黑色烟雾效果好于白色烟雾,浓烟分割效果好于淡烟,这是由于浓而深色的烟雾与背景图片具有较高的对比度。

第 2 组图像探究了云和雾对分割效果的影响,分别在淡云天气(见图 9(d))、有雾天气(见图 9(e))、浓云天气(见图 9(f))对烟雾分割效果进行对比,实验表明,雾和云会在一定程度上影响分割效果,但本文模型依旧能够整体捕捉到云、雾与烟的差异性,有效从云、雾背景中分割出烟雾。

第 3 组图像实验了本文模型在复杂背景中的分割情况,分别在 2 个烟雾源(见图 9(g))、3 个烟雾源(见图 9(h))、多个烟雾源(见图 9(i))的图像中进行烟雾分割,实验表明,本文模型在复杂场景中也具有较好的鲁棒性。


图 9　真实场景可视化实验结果

Fig. 9　Visualized experimental results of real scenes

## 4　结　论

针对烟雾区域连续成片状的特点及边缘像素点不易分割的问题,本文从注意力机制角度出发,研究烟雾分割网络的性能提升方法。考虑到烟雾检测对于实时性的要求,为避免大尺寸矩阵运算带来的过大内存消耗和过高的计算复杂度,本文分别设计了双向定位空间注意力模块和多尺度通道注意力融合模块,并结合 ResNet50 特征提取模块,提出了全局烟雾注意网络。在虚拟烟雾数据集和真实烟雾数据集上分别进行了实验。实验结果表明,本文提出的 GSA-Net 网络能够在避免传统注意力机制由于大矩阵运算产生的过大内存负担和过高计算复杂度的同时,实现全局注意力信息建模。此外,在通道维度能够自适应地对不同尺度的特征图进行通道加权和共享,更全面地捕获通道维度的依赖关系,从而辅助优化分割效果。剥离实验结果也表明,本文提出的 2 个模块对烟雾分割性能都有很好的提升作用。为验证本文模型的推广性能,还在更具挑战性的真实场景进行测试,取得了很好的实验结果。

在未来工作中,将探索更加高效的长距离依赖特征提取网络模型,提升算法精度,同时降低计算复杂度。

## 参考文献（References）

[1] 夏雪,袁非牛,章琳,等. 从传统到深度:视觉烟雾识别、检测与分割[J]. 中国图象图形学报, 2019, 24(10):1627-1647.
XIA X, YUAN F N, ZHANG L, et al. From traditional methods to deep ones: Review of visual smoke recognition, detection, and segmentation[J]. Journal of Image and Graphics, 2019, 24(10):1627-1647(in Chinese).

[2] 金博. 森林防火:全国森林火灾分月统计(2017)[M]//国家林业和草原局. 中国林业年鉴(2018). 北京:中国林业出版社,2018:138.
JIN B. Forest fire prevention forest fire by months(2017)[M]// State Forestry and Grassland Administration. China forestry yearbook(2018). Beijing: China Forestry Publishing House, 2018:138(in Chinese).

[3] YUAN F N, ZHANG L, WAN B Y, et al. Convolutional neural networks based on multi-scale additive merging layers for visual smoke recognition[J]. Machine Vision and Applications, 2019, 30(2):345-358.

[4] YUAN F N, ZHANG L, XIA X, et al. Deep smoke segmentation[J]. Neurocomputing, 2019, 357:248-260.

[5] YUAN F N, ZHANG L, XIA X, et al. A wave shaped deep neural network for smoke density estimation[J]. IEEE Transactions on Image Processing, 2020, 29:2301-2313.

[6] CHEN L C, PAPANDREOU G, SCHROFF F, et al. Rethinking atrous convolution for semantic image segmentation[EB/OL]. (2017-12-05)[2021-09-01]. https://arxiv.org/abs/1706.05587.

[7] ZHANG Z, ZHANG C, SHEN W, et al. Multi-oriented text detection with fully convolutional networks[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2016:4159-4167.

[8] RONNEBERGER O, FISCHER P, BROX T. U-Net: Convolutional networks for biomedical image segmentation[C]// International Conference on Medical Image Computing and Computer-Assisted Intervention. Berlin: Springer, 2015:234-241.

[9] MNIH V, HEESS N, GRAVES A, et al. Recurrent models of visual attention[C]// Proceedings of the 27th International Conference on Neural Information Processing Systems. New York: ACM, 2014:2204-2212.

[10] 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.

[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] WOO S, PARK J, LEE J Y, et al. CBAM: Convolutional block attention module[C]// Proceedings of the European Conference on Computer Vision (ECCV). Berlin: Springer, 2018:3-19.

[13] PARK J, WOO S, LEE J Y, et al. BAM: Bottleneck attention module[EB/OL]. (2018-07-18)[2021-09-01]. https://arx-

iv.org/abs/1807.06514v2.

[14] 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.

[15] YUAN Y, HUANG L, GUO J, et al. OCNet: Object context network for scene parsing[EB/OL]. (2021-03-15)[2021-09-01]. https://arxiv.org/abs/1809.00916v4.

[16] 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.

[17] HOU Q, ZHOU D, FENG J. Coordinate attention for efficient mobile network design[EB/OL]. (2021-03-04)[2021-09-01]. https://arxiv.org/abs/2103.02907v1.

[18] 张娜,王慧琴,胡燕.粗糙集与区域生长的烟雾图像分割算法研究[J].计算机科学与探索,2017,11(8):1296-1304.
ZHANG N, WANG H Q, HU Y. Smoke image segmentation algorithm based on rough set and region growing[J]. Journal of Frontiers of Computer Science and Technology, 2017, 11(8): 1296-1304(in Chinese).

[19] LIN G H, ZHANG Y M, ZHANG Q X, et al. Smoke detection in video sequences based on dynamic texture using volume local binary patterns[J]. KSII Transactions on Internet and Information Systems, 2017, 11(11): 5522-5536.

[20] FILONENKO A, HERNÁNDEZ D C, JO K H. Fast smoke detection for video surveillance using CUDA[J]. IEEE Transactions on Industrial Informatics, 2018, 14(2): 725-733.

[21] TAO C Y, ZHANG J, WANG P. Smoke detection based on deep convolutional neural networks[C]// Proceedings of 2016 International Conference on Industrial Informatics—Computing Technology, Intelligent Technology, Industrial Information Integration. Piscataway: IEEE Press, 2016: 150-153.

[22] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. ImageNet classification with deep convolutional neural networks[C]// Proceedings of the 25th International Conference on Neural Information Processing Systems. New York: ACM, 2012: 1097-1105.

[23] YIN Z J, WAN B Y, YUAN F N, et al. A deep normalization and convolutional neural network for image smoke detection[J]. IEEE Access, 2017, 5: 18429-18438.

[24] YUAN F, ZHANG L, XIA X, et al. A gated recurrent network with dual classification assistance for smoke semantic segmentation[J]. IEEE Transactions on Image Processing, 2021, 30: 4409-4422.

[25] 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.

[26] HE K, ZHANG X, REN S, et al. Spatial pyramid pooling in deep convolutional networks for visual recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015, 37(9): 1904-1916.

[27] LONG J, SHELHAMER E, DARRELL T. Fully convolutional networks for semantic segmentation[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2015: 3431-3440.

[28] BADRINARAYANAN V, KENDALL A, CIPOLLA R. SegNet: A deep convolutional encoder-decoder architecture for image segmentation[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(12): 2481-2495.

[29] WANG W, SHEN J, SHAO L. Video salient object detection via fully convolutional networks[J]. IEEE Transactions on Image Processing, 2018, 27(1): 38-49.

[30] CHEN L C, PAPANDREOU G, KOKKINOS I, et al. DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018, 40(4): 834-848.

[31] MEHTA S, RASTEGARI M, CASPI A, et al. ESPNet: Efficient spatial pyramid of dilated convolutions for semantic segmentation[C]// Proceedings of the European Conference on Computer Vision (ECCV). Berlin: Springer, 2018: 552-568.

[32] ISLAM M A, NAHA S, ROCHAN M, et al. Label refinement network for coarse-to-fine semantic segmentation[EB/OL]. (2017-03-01)[2021-09-01]. https://arxiv.org/abs/1703.00551.

[33] NEWELL A, YANG K, DENG J. Stacked hourglass networks for human pose estimation[C]// Proceedings of the European Conference on Computer Vision(ECCV). Berlin: Springer, 2016: 483-499.

# Improved spatial and channel information based global smoke attention network

DONG Zeshu[1], YUAN Feiniu[1,*], XIA Xue[2]

(1. College of Information, Mechanical and Electrical Engineering, Shanghai Normal University, Shanghai 201418, China;
2. School of Information Management, Jiangxi University of Finance and Economics, Nanchang 330032, China)

**Abstract**: Smoke has the characteristics of semi-transparency, irregularity and blurry boundaries, leading to the challenging task of image smoke segmentation. To solve these problems, we propose an attention modeling method to extract the correlation of long-distance information. The attention method can capture the long-distance dependency of pixels and continuity of regions, so as to reduce the misclassification of discontinuous smoke regions. To avoid large memory consumption of large matrix multiplication and high computational complexity, we modify both spatial and channel attention structures to design a bi-direction attention (BDA) and a multi-scale channel attention (MSCA), which are used to compensate for lost spatial information by global pooling in attention methods. In addition, we propose a global smoke attention network, which combines residual networks with attention models to reduce memory consumption and computational complexity without sacrificing global correlation information. Experimental results show that the proposed network achieves the mean intersection over union of 73.13%, 73.81% and 74.25% on the three virtual smoke test datasets of DS01, DS02 and DS03, respectively, and it outperforms most of the existing state-of-the-art methods.

**Keywords**: smoke segmentation; bi-directional localization; spatial attention; multi-scale fusion; channel attention

**Received**: 2021-09-14; **Accepted**: 2021-10-01; **Published online**: 2021-10-28 15:31

**URL**: kns.cnki.net/kcms/detail/11.2625.V.20211027.2052.002.html

**Foundation items**: National Natural Science Foundation of China (61862029,62062038); Project of Education Department of Jiangxi Province (GJJ201117)

* **Corresponding author**. E-mail: yfn@ustc.edu

http://bhxb.buaa.edu.cn jbuaa@buaa.edu.cn
**DOI**: 10.13700/j.bh.1001-5965.2021.0524

# 基于图对比的上下位关系检测

张雅丽[1], 方全[2,*], 王允鑫[1], 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**: 上下位关系是自然语言处理(NLP)下游任务的基础,因此上下位关系检测是自然语言处理领域备受关注的问题。针对现有词嵌入方法采用随机初始化词向量,不能很好地捕获上下位关系不对称和可传递的特性,且现有模型没有充分利用预测向量与真实投影之间关系的局限性,提出了一种基于图对比学习的上下位关系检测(HyperCL)方法。引入图对比学习进行数据增强,基于最大化局部和全局表示的互信息,学习具有鲁棒性的词特征表示。所提方法学习了将下位词的词向量投影到上位词和非上位词,同时能够更好地区分嵌入空间中的上位词和非上位词,从而提高了检测精度。在2个基准数据集上的实验结果表明,所提模型比现有方法在准确率上提升了0.03以上。

**关 键 词**: 自然语言处理(NLP); 上下位关系检测; 图对比学习; 数据增强; 词嵌入
**中图分类号**: TP18
**文献标志码**: A **文章编号**: 1001-5965(2022)08-1480-07

上下位关系(hypernymy)在自然语言中是一种基础的语义关系,用来表示概念之间的层次隶属关系。层次关系在知识表示和推理方面起着至关重要的作用。因此,上下位关系检测是自然语言处理(natural language processing, NLP)研究中的重要基石,可用以提升个性化推荐[1]、词汇蕴含[2]和网页查询理解[3]等下游任务的准确性。研究人员在自由文本语料库上进行了大量的上下位关系检测工作,归属为2类方法:模式匹配法和分布式表示法[4]。

模式匹配法最早追溯到Hearst-patterns[5],该模式使得大型语料库成为显式模板is-a词对的良好资源。通过模板提取模式词对$(x,y)$,如"$y$ such as $x$"和"$x$ and other $y$"[6],其中,$x$为上位词,$y$为下位词。模式匹配法的精度很高,但提取的词对比较稀疏。分布式表示法则是遵循或受分布包含假设(distributional inclusion hypothesis, DIH)[7]的启发,认为下位词的上下文集合应该包含在上位词的上下文集合中,利用术语的分布式向量来预测上下位关系。例如,文献[8-9]使用无监督上下位关系检测方法预测2个术语间是否存在上下位关系;文献[10]将DIH映射到无监督嵌入模型中,从而获得包含上下位关系信息的潜在特征向量;文献[11-12]采用监督算法预测上下位关系,把术语对用嵌入向量表示,放入支持向量机(support vector machine, SVM)或逻辑回归分类器中进行预测。但分布式表示法不能区分更细粒度的词汇关系,可能存在"词汇记忆"问题。除此之外,有的方法把模式匹配法和分布式表示法结合来提高性能[13]。近年来,文献[10,14-15]提出了学习词嵌入来捕获上下位关系特性的方法,能够建模术语在映射空间中将词向量映射到其上位词词向

**收稿日期**: 2021-09-06; **录用日期**: 2021-09-17; **网络出版时间**: 2021-11-01 14:53
**网络出版地址**: kns.cnki.net/kcms/detail/11.2625.V.20211101.1030.003.html
**基金项目**: 国家自然科学基金(62072456,62036012); 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**: *张雅丽, 方全, 王允鑫, 等. 基于图对比的上下位关系检测[J]. 北京航空航天大学学报, 2022, 48(8): 1480-1486. ZHANG Y L, FANG Q, WANG Y X, et al. Hypernymy detection based on graph contrast[J]. Journal of Beijing University of Aeronautics and Astronautics, 2022, 48(8): 1480-1486 (in Chinese).*

量,这使得该方法在保持分布式表示法优点的同时克服了“词汇记忆”的问题[16]。线性投影模型[17]的目的在于:学习线性投影矩阵,使下位词词向量映射到其对应上位词的词向量的误差最小。还有方法联合学习上下位和非上下位关系的表示,训练一个二分类模型判断词对是否存在上下位关系。

但是上述方法都有一定的局限性:①现有词嵌入方法都是用神经网络模型随机初始化词向量,不能很好地表示不同关系词语的特征,导致映射的网络会有重叠的可能,也就是说,一个词既属于上位词,也属于非上位词(下位词、同义词等);②虽然学习了上下位关系和非上下位关系的投影网络,但是没有同时考虑预测向量与真实上位词和非上位词向量的距离问题,可能会使预测结果不够精确。

针对以上局限性,本文提出了一种基于图对比学习的上下位关系检测(HyperCL)模型。图对比学习方法近年来在视觉图形等领域发挥着重要作用,也出现了多种关于图对比学习的方法。例如,最大化局部和全局表示一致性的方法[18];将最大化互信息思想引入异构图中[19];关注优化节点嵌入和图嵌入之间的相似度[20]。本文引入图对比学习方法,通过在不同增强子图下最大化特征一致性来学习节点的特征表示。

本文模型主要由 2 个部分组成:①在学习特征方面,为了进一步提高词语特征的表达能力,设计了一个图对比学习网络层,能够通过互信息最大化学到节点的全局表示;②在关系检测方面,提出上下位关系检测层,建模上位词和非上位词映射网络,并同时利用 2 种类别关系的映射来提高检测精度。

总的来说,本文主要工作包括:

1) 提出使用图对比学习全局节点特征的方法,学到的节点是以目标节点为中心的整个子图的所有节点表示,即利用数据或任务特定的增强,注入期望的特征不变性。

2) 提出了一种新的损失函数,在将下位词词向量映射到真实上位词和非上位词的同时,还考虑预测上位词词向量和 2 个真实向量在映射空间的距离问题,减小上位词和非上位词映射重叠的可能,更好地捕获上下位关系,使得模型预测的结果更加精确。

3) 在 2 个公共的基准数据集的实验结果表明,与当前最好的方法相比,模型在检测上下位关系的精确度上提升了 0.03 以上。

# 1　方　法

## 1.1　问题描述

上下位关系检测任务的目的是:给定一个未知关系词对($x$,$y$),判断 $x$ 和 $y$ 是否有上下位关系。具体来说,给定下位词的词向量 $\boldsymbol{y}$,经过网络层学习得到上位词的预测词向量 $\boldsymbol{P}$ 和非上位词的预测词向量 $\boldsymbol{N}$,通过距离函数判断 $\boldsymbol{P}$ 与 $\boldsymbol{y}$ 是否存在上下位关系。

## 1.2　整体框架

图 1 给出了本文模型的架构。通过图对比学习,本文模型可以学习到互信息最大化的节点特征,并且学习到与术语相似的非邻居节点的特征。本文模型主要由 2 个部分组成:①图对比模块;②上下位关系检测模块。首先,图对比模块经过数据增强得到术语的全局特征表示 $\boldsymbol{x}$。然后,将 $\boldsymbol{x}$ 送入上下位关系检测模块,先经过隐藏层得到 $\boldsymbol{x}'$,将 $\boldsymbol{x}$ 与 $\boldsymbol{x}'$级联,经过线性层,得到预测上位词向量 $\boldsymbol{P}$ 和非上位词向量 $\boldsymbol{N}$。最后,用 $s$ 分数判定未知词对上下位关系。

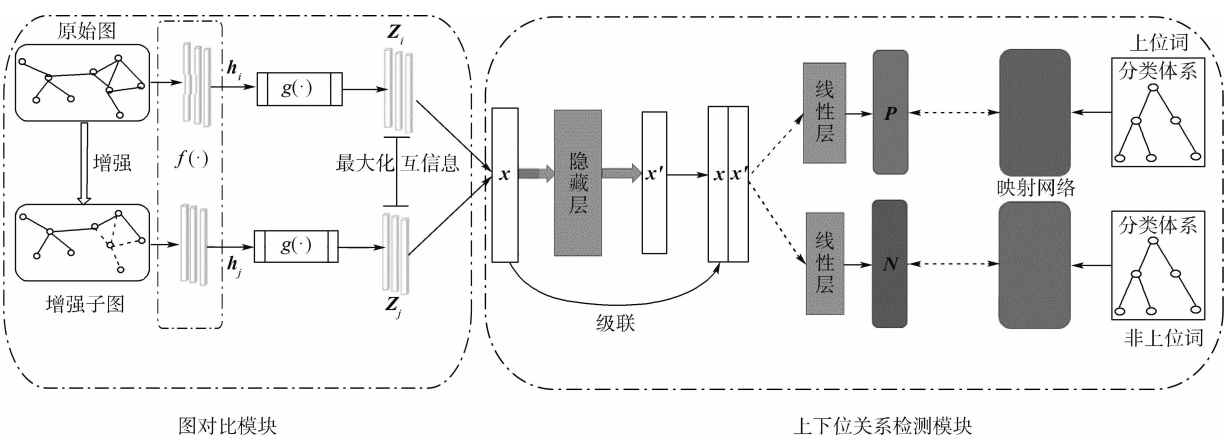



图 1　图对比的上下位关系检测模型

Fig. 1　Hypernymy detection based on graph contrastive model

### 1.3 图对比模块

图对比模块的目标是通过数据增强学习到不受扰动的节点表示。在进行数据增强前，先将 Glove 模型随机初始化的词向量利用术语间的关系构成图 $G$。$\boldsymbol{X}=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_n\}$ 为 Glove 模型初始化的词向量，$n$ 为图中的节点数，$\boldsymbol{x}_i\in\mathbf{R}^F$ 表示节点 $i$ 的特征。以邻接矩阵 $\boldsymbol{A}\in\mathbf{R}^{n\times n}$ 的形式表示节点之间真实关系信息。$\boldsymbol{A}$ 可能是由任意实数（或任意边特征）组成，但在实验中都假设图是未加权的，也就是如果 $i\rightarrow j$ 有上下位关系，那么 $A_{ij}=1$，否则 $A_{ij}=0$。

在假设图 $G$ 的语义会被保留在其局部结构的前提下，用随机游走构成子图的数据增强方式得到 2 个子图 $G_i$ 和 $G_j$，在尝试多种构图方法后（移除节点、移除边、随机游走构成子图等），发现用随机游走构成子图方法得到的词向量最好。因此，本文选择随机游走构成子图方法。

学习一个图卷积编码器，编码器表示为

$$f:\mathbf{R}^{n\times F}\times\mathbf{R}^{n\times n}\rightarrow\mathbf{R}^{n\times F'} \tag{1}$$

$$f(\boldsymbol{X},\boldsymbol{A})=\sigma(\hat{\boldsymbol{D}}^{-\frac{1}{2}}\hat{\boldsymbol{A}}\hat{\boldsymbol{D}}^{-\frac{1}{2}}\boldsymbol{X}\boldsymbol{\Theta}) \tag{2}$$

式中：$\hat{\boldsymbol{A}}=\boldsymbol{A}+\boldsymbol{I}_N$ 为带有自循环的邻接矩阵；$\hat{\boldsymbol{D}}$ 为其对应的单位矩阵；$\sigma$ 为非线性激活 ReLU 函数；$\boldsymbol{\Theta}\in\mathbf{R}^{F\times F'}$ 为应用于每个节点的可学习线性变换；$f(\boldsymbol{X},\boldsymbol{A})=\boldsymbol{H}=\{\boldsymbol{h}_1,\boldsymbol{h}_2,\cdots,\boldsymbol{h}_n\}$ 为每个节点 $i$ 的表示 $\boldsymbol{h}_i\in\mathbf{R}^F$。

图卷积编码器是一类灵活的节点嵌入结构，通过在局部节点邻域上重复聚合来生成节点 $\boldsymbol{h}_i$，该节点为聚合了以节点 $i$ 为中心的子图表示，而不仅仅是节点本身。2 个增强子图 $G_i$ 和 $G_j$ 用同一个编码器学习节点表示。

为了得到增强子图的全局表示向量 $\boldsymbol{Z}_i$ 和 $\boldsymbol{Z}_j$，用非线性转换层 $g(\cdot)$ 将得到的局部表示聚合成全局的图表示。

用一个分类器来最大化局部互信息，分类器表示为

$$D:\mathbf{R}^F\times\mathbf{R}^F\rightarrow\mathbf{R} \tag{3}$$

$D(\boldsymbol{h}_i,s)$ 表示分配给该节点-全局表示的概率分数（包含在原始图 $G$ 中的节点分数会更高）。

在一个小批量中随机采样 $n$ 个样本点，每个增强后的样本点都是正样本，其余 $(2n-1)$ 个增强数据为负样本。损失函数定义如下：

$$\zeta=-\lg\frac{\exp(\mathrm{sim}(\boldsymbol{Z}_i,\boldsymbol{Z}_j)/\tau)}{\sum_{k=1,k\neq i}^{2n}\exp(\mathrm{sim}(\boldsymbol{Z}_i,\boldsymbol{Z}_j)/\tau)} \tag{4}$$

式中：$\tau$ 为温度参数；$\mathrm{sim}(\boldsymbol{Z}_i,\boldsymbol{Z}_j)$ 为这 2 个向量之间的余弦相似度。上述损失函数本质上最大化 $\boldsymbol{Z}_i$ 和 $\boldsymbol{Z}_j$ 之间互信息的下限。

### 1.4 上下位关系检测模块

上下位关系检测模块学习下位词映射到其真实上位词和非上位词，并且根据预测上位词词向量与 2 个真实向量的距离判断预测词对是否为上位词。

模块的一端输入是真实上位词和非上位词的映射网络，由两部分训练集构成，用 $C^{(+)}$ 和 $C^{(-)}$ 表示。其中，$C^{(+)}$ 是从分类知识体系中直接获取的，而 $C^{(-)}$ 来自于 2 个部分：①正确上下位关系的相反方向上下位关系词对，如 $C^{(-)}=\{(y,x)\mid(x,y)\in C^{(+)}\}$；②用反方向上下位关系词对随机匹配术语对和共下位词词对。

该模块的另一端输入是用图对比学习得到的词嵌入 $\boldsymbol{x}$。将 $\boldsymbol{x}$ 经过隐藏层，得到 $H(\boldsymbol{x},\theta_C^{(s)})$，$\theta_C^{(s)}$ 为共享映射参数，可以提高分布式语义中多任务学习的效率[20]。将输入 $\boldsymbol{x}$ 和 $H(\boldsymbol{x},\theta_C^{(s)})$ 做通道连接。

$$\mathrm{cat}(\boldsymbol{x},\boldsymbol{x}')=\boldsymbol{x}\otimes H(\boldsymbol{x},\theta_C^{(s)}) \tag{5}$$

式中：cat 为 $\boldsymbol{x}$ 和 $H(\boldsymbol{x},\theta_C^{(s)})$ 的级联向量。级联的目的是：如果共享参数的部分不起作用时，也可以拥有原词向量的效果。

将 $\mathrm{cat}(\boldsymbol{x},\boldsymbol{x}')$ 分别输入 2 个不同线性层，$\theta_T^{(-)}$ 和 $\theta_C^{(-)}$ 为映射参数的集合。得到预测上位词词向量 $\boldsymbol{P}$：$P(\mathrm{cat}(\boldsymbol{x},\boldsymbol{x}'),\theta_C^{(+)})$ 和非上位词词向量 $\boldsymbol{N}$：$N(\mathrm{cat}(\boldsymbol{x},\boldsymbol{x}'),\theta_C^{(-)})$，再与真实下位词向量 $\boldsymbol{y}$ 对比。损失函数如下：

$$\mathrm{loss}=\begin{cases}\tanh(\|\boldsymbol{P}-\boldsymbol{y}\|_2^2-\|\boldsymbol{N}-\boldsymbol{y}\|_2^2)+1 \\ \qquad\qquad y\in\text{hypernyms} \\ \tanh(\|\boldsymbol{N}-\boldsymbol{y}\|_2^2-\|\boldsymbol{P}-\boldsymbol{y}\|_2^2)+1 \\ \qquad\qquad y\in\text{Non-hypernyms}\end{cases} \tag{6}$$

因为一个术语 $\boldsymbol{x}$ 可能是上位词也可能是下位词，所以映射网络会有重叠的可能，对检测结果会有一定影响，该损失函数可以使上位词预测词向量与真实上位词词向量距离很近，与真实非上位词词向量距离很远，从而使检测结果更加精确。

在模型训练后，用 $s(x,y)\in(-1,1)$ 分数[4]判断词对 $(x,y)$ 中 $x$ 和 $y$ 是否具有上下位关系：

$$s(x,y)=\tanh(\|\boldsymbol{N}-\boldsymbol{y}\|_2^2-\|\boldsymbol{P}-\boldsymbol{y}\|_2^2) \tag{7}$$

式中：tanh 为激活函数。若 $s>0$，则判断 $y$ 是 $x$ 的下位词；否则，判断 $y$ 不是 $x$ 的下位词。

## 2　实　验

### 2.1　数据准备

实验使用 BLESS 和 WBLESS 二个基准数据集来验证本文模型的有效性。这 2 个数据集主要用于上下位关系检测的二分类任务。本文只使用 BLESS[10] 数据集的名词-名词子集,共 6 433 个词对,训练集和测试集比例为 6:4。WBLESS[12] 包含 1 668个词对,由 2 种类型的关系组成:上下位关系和其他(包括反方向上下位关系,同义词关系和全称-别名关系等),训练集、验证集和测试集比例为 6:1:3。

### 2.2　比较方法

本文选取了 7 种先进的模型进行比较:

1) Santus[8]:一种基于熵的分布语义模型。

2) Weeds[12]:基于对特征向量对训练线性 SVM 的分布法。

3) Kiela[9]:基于概念的相关图,引入概念的视觉属性用于词汇蕴涵检测。

4) Nguyen[10]:一种神经网络模型,学习用于上下位检测和方向性的层次嵌入。

5) Roller[6]:对比了模板匹配法和分布表示学习法在多个上下位任务上的性能。

6) Wang[21]:基于词嵌入映射网络区分上下位关系和其他语义关系。

7) Wang-MWP[22]:一种多 Wahba 映射模型,用于区分基于词嵌入的上下位和非上下位关系,MWP 模型主要用于监督关系分类任务,但通过添加小的修改也可以以一种无监督的方式预测上下位关系。

### 2.3　评估方法与参数设置

本文使用平均准确率(average precision,AP)作为评估指标,该指标能够对检测结果做出较为精确的估计,其也用在其他模型中[9-10,12]。初始化词向量训练,本文选择许多先进方法[10,23]使用的 Glove 模型。在 2 个英语基准数据集上用相同维度的词向量作为嵌入,评估所有模型的实验效果。

图对比学习模块中的初始词嵌入向量的维度 $d$ 设置为 100,隐藏层的维度为 521,输出的词向量维度为 100;在上下位关系检测模块,使用 200 维的全连接层作为隐藏层,tanh 函数作为激活函数,用 Adam 优化算法学习模型参数,epoch 为 1 000,学习率为 0.000 1,小批量设置为 64。

## 3　实验结果分析

### 3.1　上下位关系检测结果分析

从表 1 可以看出,在 BLESS 数据集上,与 Roller[6] 模型相比,本文模型的 AP 提高了 0.03。原因在于:本文模型利用了节点及节点间关系的全局特征表示,而不是随机初始化节点表示。

在 WBLESS 数据集上,本文方法比 Wang[21] 方法提高了 0.08,比 Nguyen[10] 模型提高了 0.09,与监督模型 Wang-MWP[22] 模型相比,本文方法也分别高出了 0.02 和 0.04。主要因为:本文模型在嵌入空间中能更好地区分上位词和非上位词,避免了映射网络重叠的可能,进一步提高了检测精度。如表 1 所示,本文方法的结果最好。

**表 1　不同方法的 AP 值比较**

**Table 1　Comparison of AP with different methods**

| 方法 | AP | |
|---|---|---|
| | BLESS | WBLESS |
| Santus[8] | 0.87 | |
| Weeds[12] | | 0.75 |
| Kiela[9] | 0.88 | 0.75 |
| Nguyen[10] | 0.92 | 0.87 |
| Roller[6] | 0.96 | 0.87 |
| Wang[21] | 0.96 | 0.88 |
| Wang-MWP[22] | 0.97 | 0.92 |
| HyperCL | 0.99 | 0.96 |

### 3.2　神经网络结构分析

为了验证神经网络对本文模型的影响,进一步分析了神经网络结构对模型性能的影响。以 BLESS 数据集为例,在 BLESS 数据集上实验了隐藏层和隐藏单元数量改变对预测精度的影响。为了减少模型训练的随机性,训练模型 10 次并取其平均值,结果如图 2 和图 3 所示。

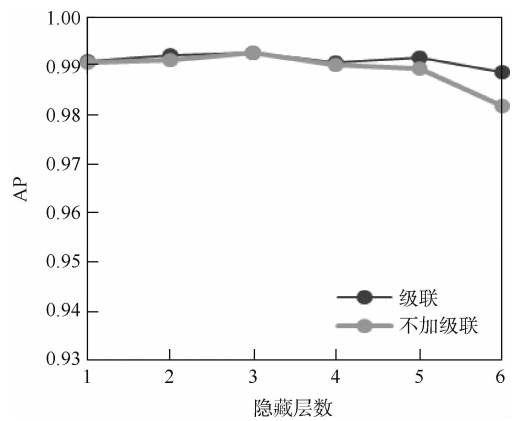



图 2　改变隐藏层数时 AP 的性能变化

Fig. 2　Performance changes in AP with hidden layer number changing

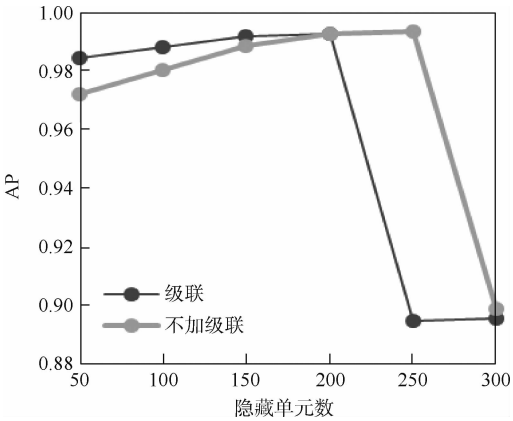



图 3 改变隐藏单元数时 AP 的性能变化

Fig. 3 Performance changes in AP with hidden unit number changing

图 2 的结果表明,隐藏层数对本文模型影响不大,加级联的模型比不加级联的模型效果较好。从图 3 中可以看到,当隐藏单元数小于 200 时,加级联的模型效果比不加级联模型的效果要好;在隐藏单元数大于 200 时,加级联的模型性能迅速下降,而不加级联的模型在单元数大于 250 时才下降,性能下降可能是过拟合造成的。总的来说,相对隐藏单元数来说,隐藏层数对本文模型的影响不大,也从侧面说明本文模型也具有相对较高且稳定的性能。

### 3.3 消融实验

为了验证本文模型中图对比学习模块的有效性,在 2 个数据集上做了消融实验,epoch 为 1 000,实验结果保留 3 位小数,结果如表 2 所示。

**表 2 不同词嵌入下的检测性能比较**

**Table 2 Comparison of detection performance under different word embeddings**

| 输入 | AP | |
|---|---|---|
| | BLESS | WBLESS |
| HyperCL-no | 0.987 | 0.922 |
| HyperCL | 0.992 | 0.961 |

本文对比了 2 种不同词嵌入训练方法:随机初始化词向量和图对比学习得到的词向量,HyperCL-no 表示随机初始化词向量,HyperCL 表示用图对比学习训练词向量。实验结果表明,加入图对比学习后,在 BLESS 数据集上 AP 提高了 0.005,WBLESS 数据集上提高了 0.039,证明了使用图对比学习训练词向量能够进一步提高检测精度。

为了验证本文中提出的损失函数对于模型性能的影响,在 2 个数据集上进行了消融实验。将效果最好的 Wang[21] 模型作为 Baseline,epoch 为 1 000,实验结果保留 3 位小数。结果如表 3 所示。

**表 3 不同损失函数下的检测性能比较**

**Table 3 Comparison of detection performance with different loss functions**

| 输入 | AP | |
|---|---|---|
| | BLESS | WBLESS |
| Baseline-no | 0.964 | 0.887 |
| HyperCL-no | 0.987 | 0.922 |
| Baseline | 0.945 | 0.870 |
| HyperCL | 0.992 | 0.961 |

表 3 中:Baseline-no 表示使用随机初始化词向量,Baseline 表示使用图对比学习的词向量。从表 3 的实验结果可以看出,在使用随机初始化词向量,用 Baseline 的损失函数时,本文模型在 2 个数据集上的 AP 比 Baseline 分别提高了 0.023 和 0.035;使用图对比学习学到的词向量,用损失函数时,跟 Baseline 相比,精度分别提高了 0.047 和 0.091,由此证明了损失函数能够减小映射网络重叠的可能性,进一步提高检测精度。但是,也发现 Baseline 用随机初始化词向量比用图对比学习向量的精度要高,猜测可能是因为图对比学习到的表示更能表现每个节点的特征,增加了在映射网络中重叠的可能,导致精度下降,而本文提出的新的损失函数能更好地避免该问题。

## 4 结 论

1)本文模型引入图对比学习,通过在不同增强子图下最大化特征一致性来学习节点的特征表示,能够更好地捕获上下位关系的特性。

2)提出能够充分利用映射网络中上位词和非上位词关系的损失函数,使得预测结果有更进一步的提升。

3)在 2 个基准数据集上的实验表明,所提模型可以更有效地进行上下位关系检测任务。

本文只研究了上下位关系的二分类,关注了在英语数据集的模型性能。在将来,会将工作继续扩展到多分类、多语言的语义关系检测,并在其他多分类数据集上对其进行评估。

### 参考文献(References)

[1] ZHANG Y C, AHMED A, JOSIFOVSKI V, et al. Taxonomy discovery for personalized recommendation[C]// Proceedings of the 7th ACM International Recommendation on Web Search and Data Mining. New York: ACM, 2014: 243-252.

[2] VULIC I, GERZ D, KIELA D, et al. HyperLex: A large-scale evaluation of graded lexical entailment[J]. Computational Linguistics, 2017, 43(4): 781-835.

[3] WANG Z Y, ZHAO K, WANG H X, et al. Query understanding

through knowledge-based conceptualization[C]// Proceedings of the International Joint Conferences on Artificial Intelligence. Palo Alto:AAAI,2015:3264-3270.

[4] WANG C Y,YAN J C,ZHOU A,et al. Transductive non-linear learning for Chinese hypernym prediction[C]// Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics,2017:1394-1404.

[5] HEARST M. Automatic acquisition of hyponyms from large text corpora[C]// Proceedings of the 14th International Conference on Computational Linguistics,1992:539-545.

[6] ROLLER S,KIELA D,NICKEL M. Hearst patterns revisited: Automatic hypernym detection from large text corpora[C]// Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics,2018:358-363.

[7] GEFFET M,DAGAN I. The distributional inclusion hypotheses and lexical entailment[C]// Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics,2015:107-114.

[8] SANTUS E,LENCI A,LU Q,et al. Chasing hypernyms in vector spaces with entropy[C]// Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics,2014:38-42.

[9] KIELA D,RIMELL L,VULIC I,et al. Exploiting image generality for lexical entailment detection[C]// Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing,2015:119-124.

[10] NGUYEN K A,KOPER M,WALDE S S I,et al. Hierarchical embeddings for hypernymy detection and directionality[C]// Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing,2017:233-243.

[11] ROLLER S,ERK K,BOLEDA G. Inclusive yet selective: Supervised distributional hypernymy detection[C]// Proceedings of the 25th International Conference on Computational Linguistics,2014:1025-1036.

[12] WEEDS J,CLARKE D,REFFIFIN J,et al. Learning to distinguish hypernyms and co-hyponyms[C]// Proceedings of the 25th International Conference on Computational Linguistics,2014:2249-2259.

[13] SHWARTZ V,GOLDBERG Y,DAGAN I. Improving hypernymy detection with an integrated path-based and distributional method[C]// Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics,2016:2389-2398.

[14] YU Z,WANG H,LIN X,et al. Learning term embeddings for hypernymy identification[C]// Proceedings of the International Joint Conferences on Artificial Intelligence. Palo Alto:AAAI,2015:1390-1397.

[15] LUU A T,TAY Y,HUI S C,et al. Learning term embeddings for taxonomic relation identification using dynamic weighting neural network[C]// Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing,2016:403-413.

[16] 汪诚愚,何晓丰,宫学庆,等. 面向上下位关系预测的词嵌入投影模型[J]. 计算机学报,2020,43(5):869-883.
WANG C Y,HE X F,GONG X Q,et al. Word embedding projection models for hypernymy[J]. Journal of Computers,2020,43(5):869-883(in Chinese).

[17] FU R J,GUO J,QIN B,et al. Learning semantic hierarchies via word embeddings[C]// Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics,2014:1199-1209.

[18] VELICKOVIC P,FEDUS W,HAMILTON W L,et al. Deep graph infomax[C]// ICLR 2019,2019:1-17.

[19] REN Y X,LIU B,HUANG C,et al. Heterogeneous deep graph infomax[EB/OL]. (2020-11-13)[2021-09-01]. https://arxiv.org/abs/1911.08538.

[20] SUN F Y,HOFFMANN J,VERMA V,et al. InfoGraph: Unsupervised and semi-supervised graph-level representation learning via mutual information maximization[EB/OL]. (2020-01-17)[2021-09-01]. https://arxiv.org/abs/1908.01000.

[21] WANG C Y,HE X F,ZHOU A Y. Improving hypernymy prediction via taxonomy enhanced adversarial learning[C]// Proceedings of the 33rd AAAI Conference on Artificial Intelligence. Palo Alto:AAAI,2019:7128-7135.

[22] WANG C Y,FAN Y,HE X F,et al. A family of fuzzy orthogonal projection models for monolingual and cross-lingual hypernymy prediction[C]// The World Wide Web Conference. New York:ACM,2019:1965-1976.

[23] PHAM N,LAZARIDOU A,BARONI M. A multitask objective to inject lexical contrast into distributional semantics[C]// Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing,2015:21-26.

# Hypernymy detection based on graph contrast

ZHANG Yali[1], FANG Quan[2,*], WANG Yunxin[1], Hu Jun[2], QIAN Shengsheng[2], XU Changsheng[2]

(1. Henan Institute of Advanced Technology, Zhengzhou University, Zhengzhou 450000, China;
2. National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China)

**Abstract**: Hypernymy is the foundation of many downstream tasks in natural language processing (NLP), so hypernymy detection has received considerable attention in the field of NLP. Adopting random initialization word vectors, existing word embedding methods cannot well capture the asymmetry and transferability of hypernymy, or make full use of the relationship between the prediction vector and the real projection. To address these problems, a novel method is proposed for detecting hypernymy based on graph contrastive learning (HyperCL). Firstly, HyperCL is introduced for data enhancement, and robust word feature representations are learned based on maximizing mutual information between local and global representations. Secondly, the proposed method learns how to project the hyponym vector to its hypernym and non-hypernym, and better distinguish the hypernym and non-hypernym in the embedded space, thus improving the detection accuracy. Experimental results on two benchmark datasets show that the proposed model increases the accuracy by more than 0.03, compared with the existing methods.

**Keywords**: natural language processing (NLP); hypernymy detection; graph contrastive learning; data augmentation; word embedding

**Received**: 2021-09-06; **Accepted**: 2021-09-17; **Published online**: 2021-11-01 14:53

**URL**: kns.cnki.net/kcms/detail/11.2625.V.20211101.1030.003.html

**Foundation items**: National Natural Science Foundation of China (62072456,62036012); 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.1001-5965.2021.0525

# 基于立体图像的多路径特征金字塔网络3D目标检测

苏凯祺, 阎维青*, 徐金东
(烟台大学 计算机与控制工程学院, 烟台 264005)

**摘 要**: 3D目标检测是计算机视觉和自动驾驶中一项重要的场景理解任务。当前基于立体图像的3D目标检测方法大多没有充分考虑多个目标之间的尺度存在较大差异,从而尺度小的物体容易被忽略,导致检测精度低。针对这一问题,提出了一种基于立体图像的多路径特征金字塔网络(MpFPN)3D目标检测方法。MpFPN对特征金字塔网络进行了扩展,增加了自底向上的路径、由上至下的路径及输入特征图到输出特征图之间的连接,为联合区域提议网络提供了更高语义信息和更细粒度空间信息的多尺度特征信息。实验结果表明:在3D目标检测KITTI数据集上,无论在场景简单、中等、复杂情况下,所提方法获得的结果都优于比较方法的结果。

**关 键 词**: 3D目标检测; 特征金字塔网络(FPN); 立体图像; 多尺度; 深度学习
**中图分类号**: TP391
**文献标志码**: A **文章编号**: 1001-5965(2022)08-1487-08

3D目标检测是最重要的场景理解任务之一,在自动驾驶和增强现实等领域有着广泛的应用。根据传感器的类型,可以将3D目标检测分为基于点云的方法[1-4]、基于单目图像的方法[5-8]和基于立体图像的方法[9-12]。基于点云(如来自LiDAR的点云)的3D目标检测可以实现最佳性能,但LiDAR传感器却是最昂贵的。此外,LiDAR的感知范围相对较短,数据分辨率相对稀疏。单目相机是最便宜、最方便安装的,但是仅使用单个图像的3D检测缺乏可靠的深度信息。与LiDAR相比,双目相机并不昂贵,并且可以为远处的小物体提供更密集的信息,甚至还可以通过自定义的基线设置感知更长的距离,而与单目相机相比,双目相机可以提供准确的深度信息。因此,基于立体图像的3D目标检测方法成为当前研究的热点。基于立体图像的3D目标检测将立体图像对作为输入,将物体的定向3D边界框作为输出。由于立体视觉的深度误差随距离呈二次方增加,如果仅依赖于带标注的3D边界框,则在训练阶段中没有使用深度图的3D目标检测是一项艰巨的任务。在没有深度监督的情况下,基于立体的3D目标检测仍是一项挑战任务。

目标检测的任务是找到图像中所有感兴趣的区域并确定其位置和类别。但是,2D目标检测只能确定目标的像素坐标,缺少目标大小和深度信息,因此在现实场景下的应用具有局限性。而3D目标检测引入了第三维度,能够确定更详细的目标尺寸和位置信息,在自动驾驶和虚拟现实等方面具有更好的应用前景。同时,第三维度的引入也带来了很多挑战,如多种传感器的选择、精确的

**收稿日期**: 2021-09-06; **录用日期**: 2021-09-17; **网络出版时间**: 2021-10-18 09:56
**网络出版地址**: kns.cnki.net/kcms/detail/11.2625.V.20211015.1719.002.html
**基金项目**: 国家自然科学基金(61801414,62072391,62066013); 山东省自然科学基金(ZR2019MF060); 山东省高等学校科研计划重点项目(J18KZ016)
* **通信作者**. E-mail: wqyan@tju.edu.cn

**引用格式**: *苏凯祺, 阎维青, 徐金东. 基于立体图像的多路径特征金字塔网络3D目标检测[J]. 北京航空航天大学学报, 2022, 48(8): 1487-1494. SU K Q, YAN W Q, XU J D. 3D object detection based on multi-path feature pyramid network for stereo images [J]. Journal of Beijing University of Aeronautics and Astronautics, 2022, 48(8): 1487-1494 (in Chinese).*

深度信息获取、不同场景下的目标检测等。

在目标检测中,不同图像中目标的尺度或同一图像中的多个目标之间的尺度存在较大差异,这种尺度差异问题是基于图像的目标检测方法面临的最核心的挑战之一。基于深度学习的目标检测方法,随着卷积神经网络(convolutional neural network, CNN)的不断加深,提取图像语义特征信息的能力不断加强,浅层的空间信息却相对丢失,导致深层特征图无法提供细粒度的空间信息来对目标进行精准定位。针对这一问题,学者提出了很多 2D 目标检测不同尺度特征的网络结构。例如,NAS-FPN[13]利用神经架构搜索来自动设计特征网络拓扑;M2Det[14]提出了 U 型模块来聚合多尺度特征;EfficientDet[15]提出了双向跨尺度连接和加权特征融合的 BiFPN;文献[16]提出了跨尺度特征聚合网络;文献[17]采用了多层次多通道方法生成多尺度候选框。基于立体图像的 3D 目标检测方法大多没有充分考虑尺度问题,因此,本文提出了多路径特征金字塔网络(multi-path feature pyramid network, MpFPN)用于 3D 目标检测,该网络结构在检测尺度大且细节丰富的目标时,以增强语义信息作为分类依据,在检测尺度小且偏差容忍小的目标时,考虑了细粒度的空间信息对目标进行精准定位。本文方法在训练期间不依赖 LiDAR 数据,仅将带有相应标注的 3D 边界框的 RGB 图像作为训练数据。首先,将立体图像对分别进行特征提取,MpFPN 为联合区域提议网络(region proposal network, RPN)提供了更高语义信息和更细粒度空间信息的多尺度特征信息。然后,将左右图像特征输入到联合 RPN 以产生联合左右感兴趣区域(region of interest, RoI)对。最后,利用立体回归信息和预测关键点对 3D 边界框进行预测。与现有方法进行比较,实验结果表明,本文方法在 3D 目标检测 KITTI 数据集上取得了更好的检测结果。

# 1 相关工作

## 1.1 基于点云的 3D 目标检测

由于点云可以提供准确的 3D 信息,大多数最新的 3D 目标检测方法都将其作为检测 3D 目标的信息输入。F-ConvNet[1]用 2D 图像生成 2D 框建议,并进一步用于裁剪对象级点云以生成 3D 边界框。PointRCNN[2]将框架调整为两阶段的网络,并且仅使用原始点云作为输入来直接生成 3D 目标提议。VoteNet[3]通过使用点特征采样和分组的投票聚类来检测目标。PV-RCNN[4]采用 3D 卷积网络从体素化点云中学习 3D 特征,将提取的 3D 特征进一步转化为鸟瞰特征图。

## 1.2 基于单目图像的 3D 目标检测

由于单目相机比 LiDAR 或立体相机便宜且安装灵活,使用单目相机进行 3D 目标检测自然成为工业界和学术界的要求。M3D-RPN[5]在观察到的 2D 投影和未观察到的深度尺寸中通过几何推理定位单目图像中的 3D 目标。MonoPSR[6]扩展了最先进的 2D 目标检测器,通过 2D 检测将方向和比例估计转移到 3D 空间中来回归目标 3D 边界框的方向及尺寸。MonoPair[7]通过考虑成对样本的关系来改进遮挡物体的建模。RAR-Net[8]直接从图像估计 3D 检测,无需预测中间 3D 场景表示。

## 1.3 基于立体图像的 3D 目标检测

尽管基于点云的 3D 目标检测已经达到了很高的精度,然而 LiDAR 传感器的昂贵价格却让很多人无法使用智能驾驶系统。随着深度学习的发展,基于立体视觉的 3D 目标检测精度得到了提高。3DOP[9]使用立体图像对估计 3D 点云特征,通过贪婪算法估计每个 3D 候选提议,提出了一个以 3D 目标提议为输入的 3D 目标检测网络,以预测准确的 3D 边界框。Stereo R-CNN[10]使用从粗到精的 3D 边界框估计方法,利用立体 RPN 预测的信息计算粗略的 3D 目标边界框,通过使用基于区域的光度对齐恢复准确的 3D 边界框。Pseudo-LiDAR[11]为基于立体图像的 3D 目标检测引入了两步方法,先将从立体图像估计的视差图转换为伪 LiDAR 点的点云,再利用现有基于点云的模型进行 3D 目标检测,并在单目和立体输入上都达到了最先进的性能。iDispNet[12]仅预测感兴趣区域物体上像素的视差,并将其转化为实例点云,输入到现有 3D 检测器进行 3D 边界框回归。

# 2 本文方法

本文方法流程如图 1 所示。本文方法由 3 个阶段组合而成:首先,输入一对左右图像进行特征提取;然后,将这些特征输入到联合 RPN 以获取相应的左右提案对;最后,利用立体回归信息和预测关键点对 3D 边界框进行预测。在特征提取方面,本文所提的 MpFPN 网络对 FPN 进行了扩展,增加了一条自底向上的路径来强化金字塔的空间信息,以及一条由上至下的路径来增强金字塔的语义信息,并在同一尺度的原始输入特征图和最终输出特征图之间增加一条连接来减少信息丢失。因此,本文提出的 MpFPN 网络为联合 RPN 提

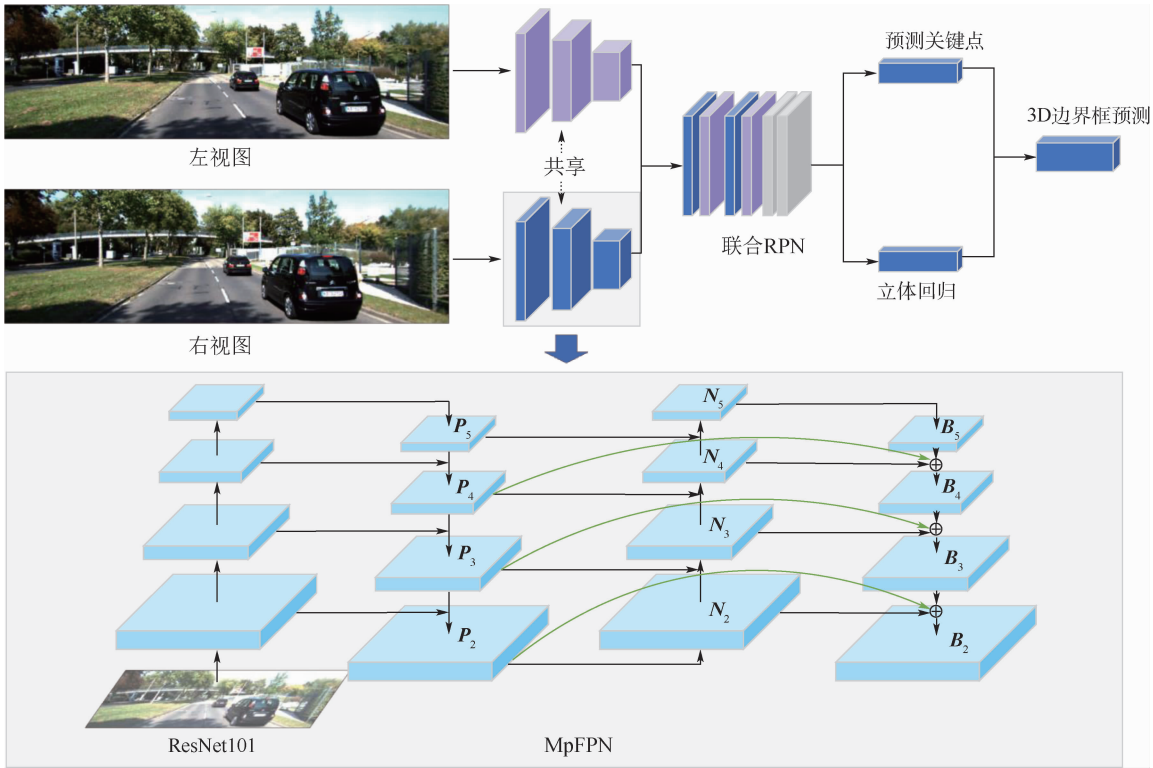



图 1　本文网络的整体框架

Fig. 1　Overall framework of proposed network

供了更高语义信息和更细粒度空间信息的多尺度特征信息。

### 2.1　MpFPN

多尺度特征融合的目的是聚合不同分辨率的特征。给出一个多尺度特征列 $\boldsymbol{P}^{\text{in}}=(\boldsymbol{P}_1^{\text{in}},\boldsymbol{P}_2^{\text{in}},\boldsymbol{P}_3^{\text{in}},\cdots,\boldsymbol{P}_n^{\text{in}})$,其中 $\boldsymbol{P}_i^{\text{in}}$ 表示第 $i$ 级别特征,目标是找到能够有效聚合不同特征并输出新特征列表的数据转换 $f$: $\boldsymbol{P}^{\text{out}}=f(\boldsymbol{P}^{\text{in}})$。传统的由上至下 FPN[18] 采用 2 ~ 5 级别特征 $\boldsymbol{P}^{\text{in}}=(\boldsymbol{P}_2^{\text{in}},\boldsymbol{P}_3^{\text{in}},\boldsymbol{P}_4^{\text{in}},\boldsymbol{P}_5^{\text{in}})$,其中,$\boldsymbol{P}_i^{\text{in}}$ 表示分辨率为输入图像的 $1/2^i$ 的特征级别。例如,如果输入图像的分辨率为 $320\times320$,则 $\boldsymbol{P}_2^{\text{in}}$ 表示分辨率为 $80\times80$ 的特征级别 $2(320/2^2=80)$,而 $\boldsymbol{P}_5^{\text{in}}$ 表示分辨率为 $10\times10$ 的特征级别 5。传统的 FPN 以从上至下的方式聚合多尺度特征:

$$\begin{cases}\boldsymbol{P}_5^{\text{out}}=\text{Conv}(\boldsymbol{P}_5^{\text{in}})\\ \boldsymbol{P}_4^{\text{out}}=\text{Conv}(\boldsymbol{P}_4^{\text{in}}+\text{Resize}(\boldsymbol{P}_5^{\text{out}}))\\ \boldsymbol{P}_3^{\text{out}}=\text{Conv}(\boldsymbol{P}_3^{\text{in}}+\text{Resize}(\boldsymbol{P}_4^{\text{out}}))\\ \boldsymbol{P}_2^{\text{out}}=\text{Conv}(\boldsymbol{P}_2^{\text{in}}+\text{Resize}(\boldsymbol{P}_3^{\text{out}}))\end{cases}\tag{1}$$

式中:Resize 用于分辨率匹配的上采样或下采样操作;Conv 用于特征处理的卷积运算。

PANet[19] 增加了自底向上的路径来解决 FPN 受到单向信息流限制的问题,且 PANet 比 FPN 和 NAS-FPN 的目标检测精度更高。因此,本文 MpFPN 网络对 FPN 进行的扩展是先增加了一条自底向上的路径。自底向上的路径使得顶层特征获得了更多的空间信息,强化了金字塔的空间信息,不仅缩短了从底层到顶层的信息流,也通过传播低级模式的强响应增强了整个特征层次结构的定位能力。原本 FPN 的顶层特征是经过卷积神经网络卷积得到的,从底层到顶层甚至要经过一百多层的卷积,导致顶层特征拥有了高语义却丢失了细粒度的空间信息。而增加的自底向上的路径却仅需不到 10 层的卷积,便可以使原本的底层细粒度空间信息传递最终的顶层特征,使得顶层特征拥有了更细粒度的空间信息。更细粒度的空间信息可以对尺度小且偏差容忍小的目标进行更精确的定位。本文遵循 FPN 中的设置,定义生成具有相同空间大小特征图的层处于同一网络阶段。每个特征级别对应一个阶段。以 ResNet[20] 为基本结构,用 $\{\boldsymbol{P}_2,\boldsymbol{P}_3,\boldsymbol{P}_4,\boldsymbol{P}_5\}$ 表示 FPN 生成的由上至下的特征级别。相对应的,自底向上的扩展路径从最低级别 $\boldsymbol{P}_2$ 开始并逐渐接近 $\boldsymbol{P}_5$。从 $\boldsymbol{P}_2$ 到 $\boldsymbol{P}_5$,空间大小使用系数 2 逐渐进行下采样。使用 $\{\boldsymbol{N}_2,\boldsymbol{N}_3,\boldsymbol{N}_4,\boldsymbol{N}_5\}$ 来表示与 $\{\boldsymbol{P}_2,\boldsymbol{P}_3,\boldsymbol{P}_4,\boldsymbol{P}_5\}$ 对应的新生成的自底向上的特征映射(注:$\boldsymbol{N}_2$ 就是 $\boldsymbol{P}_2$,没有任何处理)。

在此基础之上,本文方法又增加了一条由上至下的路径。由上至下的路径,通过对来自更高金字塔级别的空间特征图进行上采样来产生更高

分辨率的特征,向下传播高级语义信息到整个金字塔特征层次,增强了金字塔的语义信息。FPN 融合的特征信息来自于不同的深度,不同深度特征之间表征能力不同,存在着显著的语义鸿沟。而增加的由上至下路径,通过再次的不同深度之间特征融合来缓和特征之间的语义差距,进一步加深了网络层数来获得更高语义信息。更高语义特征信息可以对尺度大且细节丰富的目标进行更精确的分类。MpFPN 使用 $\{\boldsymbol{B}_2,\boldsymbol{B}_3,\boldsymbol{B}_4,\boldsymbol{B}_5\}$ 来表示与 $\{\boldsymbol{N}_2,\boldsymbol{N}_3,\boldsymbol{N}_4,\boldsymbol{N}_5\}$ 对应的特征映射(同样,$\boldsymbol{B}_5$ 就是 $\boldsymbol{N}_5$,没有任何处理)。由上至下的路径通过对来自更高金字塔级别的空间上更粗糙但语义上更强大的特征图进行上采样,来产生更高分辨率的特征。这些特征通过横向连接从自底向上的路径中得到增强。每个横向连接合并来自自底向上路径和自顶向下路径的相同空间大小的特征图。自底向上的特征图具有较低级别的语义,但能激活更准确的定位,因为其子采样次数更少。

本文在多路径的金字塔网络中增加了输入特征图和输出特征图之间的连接,在不增加网络存储成本的情况下融合了更多的特征信息,从而减少了因层数增加导致的信息丢失问题。在扩展 FPN 的过程中,增加了自底向上和由上至下的路径,使得网络层数增加,信息传递路径变长,该过程导致原始信息丢失。因此,在多路径的金字塔网络中增加输入特征图和输出特征图之间的连接,因为输入特征图是直接输入到输出特征图来进行相加操作,所以并不会增加网络存储成本,同时融合了更多的特征信息,从而减少了因层数增加导致的信息丢失。例如,将输入特征图 $\boldsymbol{P}_4$ 与输出特征图 $\boldsymbol{B}_4$ 进行相加操作。

如图 1 所示,本文方法增加的自底向上路径都通过横向连接获取更高语义信息的特征图 $\boldsymbol{B}_i$ 和更高分辨率的特征图 $\boldsymbol{N}_{i-1}$,并生成中间特征图,通过连接将原始输入特征图 $\boldsymbol{P}_{i-1}$ 输入到中间特征图来生成新的特征图 $\boldsymbol{B}_{i-1}$。首先,对每个特征图 $\boldsymbol{B}_i$ 进行 2 倍的空间分辨率上采样(使用最近邻上采样)以增大空间大小。其次,通过横向连接将特征图 $\boldsymbol{N}_{i-1}$ 的每个元素和上采样图相加。然后,通过连接将特征图 $\boldsymbol{P}_{i-1}$ 与融合的特征图逐元素相加。最后,融合的特征图由一个 $3\times3$ 的卷积层处理,以生成用于后续子网的 $\boldsymbol{B}_{i-1}$。这是一个迭代过程,并在逼近 $\boldsymbol{N}_2$ 后终止。在新增路径中,本文将特征图的通道数始终固定为 256。注意,特征图 $\boldsymbol{N}_{i-1}$ 和特征图 $\boldsymbol{P}_{i-1}$ 与上采样图进行融合的先后顺序对结果没有影响,因为最终的特征图依旧是 3 个特征图的融合。

### 2.2 联合 RPN

RPN 是一种基于滑动窗口的前景检测器。特征提取后,利用一个 $3\times3$ 卷积层来减少通道,再利用 2 个全连接层对目标进行分类,并对每个输入位置的边界框偏移量进行回归。左右特征图在每个尺度上进行连接,进行感兴趣区域提取。左右真实框的联合区域被指定为物体分类的目标。如果锚框与联合真实框之一的交叉比(intersection-over-union, IoU)高于 0.7,则为其分配正标签,如果其与任何联合框的 IoU 低于 0.3,则为其分配负标签。于是,正锚框往往包含左右 2 个对象区域。当计算偏移量时,先计算正锚框相对于包含在目标联合真实框中的左右真实框的偏移量;再分别为左右回归分配偏移量,使得左右图像的目标提议从同一组锚框生成,确保左右感兴趣区域之间的正确对应关系。RPN 回归器有 6 个回归项:$[\Delta u,\Delta w,\Delta u^*,\Delta w^*,\Delta v,\Delta h]$,其中,$u$、$v$ 分别为图像上 2D 框的水平、垂直坐标,$w$、$h$ 分别为框的宽、高,且上标“$*$”为右图上的相对应的项,左右框的 $\Delta v$ 和 $\Delta h$ 一致。最终在适当的金字塔特征级别上分别对左特征图和右特征图应用 RoI Align[21] 方法。根据 2 个分支的不同要求,将 RoI 特征分别输入关键点预测分支和立体回归分支,但只有左 RoI 特征输入到关键点预测分支。立体回归分支被用来预测目标类别、立体边界框、尺寸和视角。

### 2.3 3D 边界框预测

在立体回归分支中,先使用 2 个连续的全连接层提取语义特征,再使用 4 个分支分别预测目标类别、立体边界框、尺寸和视角。本文使用 Mask R-CNN[21] 方法对关键点进行预测。先将 $14\times14$ 的 RoI 特征图输入到 6 个连续 $3\times3$ 卷积层,再用一个 $2\times2$ 反卷积层上采样输出尺度到 $28\times28$,最终预测关键点。本文利用预测的关键点和回归的立体边界框信息来预测 3D 边界框。3D 边界框可以用 $(x,y,z,\theta)$ 来表示,其中,$(x,y,z)$ 表示 3D 边界框的中心位置,$\theta$ 表示水平方向。给定左右 2D 框、关键点和回归的维度,则可以通过最小化 2D 框和关键点的重投影误差来求解 3D 边界框。本文从立体边界框和关键点中提取了 7 个测量值:$z=\{u_1,v_t,u_r,v_b,u_1',v_r',u_p\}$,代表了 2D 左方框的左、上、右、下边缘,2D 右方框的左、右边界,以及关键点的 $u$ 坐标。给定关键点就可以推断出 3D 框和 2D 框之间的对应关系,投影变换对应的约束计算如下:

$$\begin{cases} v_{\mathrm{t}} = \left(y - \dfrac{h}{2}\right) \Big/ \left(z - \dfrac{w}{2}\sin\theta - \dfrac{l}{2}\cos\theta\right) \\ u_{\mathrm{l}} = \left(x - \dfrac{w}{2}\cos\theta - \dfrac{l}{2}\sin\theta\right) \Big/ \\ \qquad \left(z + \dfrac{w}{2}\sin\theta - \dfrac{l}{2}\cos\theta\right) \\ u_{\mathrm{p}} = \left(x + \dfrac{w}{2}\cos\theta - \dfrac{l}{2}\sin\theta\right) \Big/ \\ \qquad \left(z - \dfrac{w}{2}\sin\theta - \dfrac{l}{2}\cos\theta\right) \\ \vdots \\ u'_{\mathrm{r}} = \left(x - b + \dfrac{w}{2}\cos\theta + \dfrac{l}{2}\sin\theta\right) \Big/ \\ \qquad \left(z - \dfrac{w}{2}\sin\theta + \dfrac{l}{2}\cos\theta\right) \end{cases} \tag{2}$$

式中:$b$ 为立体相机的基线长度;$w$、$h$、$l$ 分别为 3D 边界框的宽、高、长。

### 2.4　实现细节

本文采用 ResNet101 作为骨干网络,在 ImageNet 分类数据集[22]上进行预训练,并去掉了最后的全连接层和池化层来适用 MpFPN。本文翻转训练集中的图像,交换左右图像,同时分别镜像视点角度和关键点,以进行数据增强。由于考虑到 2D 空间到 3D 空间之间的投影约束,本文采用的是文献[10]中定义的多任务损失。在训练期间,每个迷你批次包含 1 个立体对和 512 个采样感兴趣区域。使用 SGD 训练网络,其权重衰减为 0.000 5,动量为 0.9。学习率初始化设置为 0.001,每 5 个周期降低 0.1。

## 3　实　验

### 3.1　数 据 集

本文方法在经典的 3D 目标检测 KITTI 数据集[23]上进行了评估,该数据集有7 481个用于训练的图像,7 518 个用于测试的图像。KITTI 官方网站限制了提交给服务器以评估测试集的访问权限。因此,本文遵循与文献[10]相同的训练和验证划分,将 7 481 个训练图像划分为3 712个训练图像和 3 769 个验证图像。

### 3.2　评价指标

为了全面评估本文方法的性能,报告了 IoU 阈值分别为 0.5 和 0.7 的汽车类别的 3D 平均精度($AP_{3D}$)和鸟瞰视角平均精度($AP_{bev}$)。根据 2D 图像中物体的遮挡/截断和物体的大小,KITTI 数据具有 3 个级别的难易度设置,分别为简单、中等(主要指标)和困难。KITTI 数据集基准中的 AP 计算由原始 Pascal VOC 基准提议的11 个召回位置改进成 40 个召回位置。

### 3.3　3D 目标检测评估

为了进行比较,本文在表 1 中总结了从单目到双目的 3D 目标检测方法的主要结果。在简单、中等和困难情况下,本文方法在所有 IoU 阈值上均优于基于单目图像的目标检测方法。这是因为:单目图像在深度估计中固有存在尺度模糊,且深度估计的误差随距离的增加而增大,使基于单目图像的目标检测方法深度估计误差较大。而本文方法则是基于立体图像进行深度估计,相对而言误差较小,导致精度优于基于单目图像的目标检测方法。与基于立体图像的目标检测方法相比,本文方法取得了更好的检测结果。3DOP[9]对图像特征仅做了单一处理,没有考虑到空间的立体特性,而本文方法是将左右立体图像对进行联合来共同提取 RoI,既保证了左右感兴趣区域之间的正确对应关系,也充分考虑到了空间立体性。Stereo R-CNN[10]采用 ResNet-101 + FPN 的结构进行特征提取,FPN 是一种在目标检测中通用的网络,其设计核心简单,但没有充分考虑多尺度问题。本文 MpFPN 网络对 FPN 进行了扩展,增加了自底向上的路径、由上至下的路径,并在多路径的金字塔网络中增加了输入特征图到输出特征图

**表 1　KITTI 验证集上汽车类别的 $AP_{bev}$ 和 $AP_{3D}$**

**Table 1　$AP_{bev}$/$AP_{3D}$ of car category on KITTI validation set**　%

| 方法 | 输入 | IoU = 0.5 | | | IoU = 0.7 | | |
|---|---|---|---|---|---|---|---|
| | | 简单 | 中等 | 困难 | 简单 | 中等 | 困难 |
| MonoGRNet[24] | M | 54.21/50.51 | 39.69/36.97 | 33.06/30.82 | 24.97/13.88 | 19.44/10.19 | 16.30/7.62 |
| M3D-RPN[5] | M | 55.37/48.96 | 42.49/39.57 | 35.29/33.01 | 25.94/20.27 | 21.18/17.06 | 17.90/15.21 |
| AM3D[25] | M | 72.64/68.86 | 51.82/49.19 | 44.21/42.24 | 43.75/32.23 | 28.39/21.09 | 23.87/17.26 |
| 3DOP[9] | S | 55.04/46.04 | 41.25/34.63 | 34.55/30.09 | 12.63/6.55 | 9.49/5.07 | 7.59/4.10 |
| TLNet[26] | S | 62.46/59.51 | 45.99/43.71 | 41.92/37.99 | 29.22/18.15 | 21.88/14.26 | 18.83/13.72 |
| Stereo R-CNN[10] | S | 87.13/85.84 | 74.11/66.28 | 58.93/57.24 | 68.50/54.11 | 48.30/36.69 | 41.47/31.07 |
| 本文方法 | S | 87.62/86.49 | 75.04/72.62 | 59.31/58.04 | 69.44/55.26 | 49.36/37.94 | 42.11/32.38 |

注:S 表示双目图像对作为输入,M 表示单目图像作为输入。“/”前数据为 $AP_{bev}$,“/”后数据为 $AP_{3D}$。

之间的连接,充分考虑了多目标尺度差异问题。MpFPN 为联合 RPN 提供了更高语义信息和更细粒度空间信息的多尺度特征信息,有效提高了 3D 目标检测的精度。具体来说,在 IoU = 0.7 的困难情况下,本文方法的结果比 Stereo R-CNN 提高了 1.31%;在中等情况下可以看到类似的观察结果,性能提高了 1.25%。

如表 2 所示,将本文方法与遵循两阶段网络的基于立体图像的 Pseudo-LiDAR[11] 进行了比较:①通过 PSMNet[27] 进行深度图估计;②通过 AVOD[28] 进行 3D 边界框回归。PL + FP 方法将生成的深度图作用于 3D 目标检测模型,重点考虑了深度图信息,然而在检测模型中仅仅利用了传统的 3D 检测模型,没有对模型进行进一步改进,也没有充分考虑目标前后遮挡及小目标问题,从而出现目标漏检情况。而本文方法分析了传统检测网络的不足,设计了多尺度 3D 目标检测网络,为目标检测提供了更高语义信息和更细粒度空间信息的多尺度特征信息,即使物体被遮挡或小尺度目标也可以很好地检测。因此,通过表 2 可以看出,本文方法精度缩小了差距,甚至在物体检测难度为中等、困难情况下,本文方法的精度高于 Pseudo-LiDAR 方法。

### 3.4 消融实验

表 3 比较了有无自底向上和由上至下路径及有无输入特征图到输出特征图之间路径的实验结果。其中,有自底向上和由上至下路径在表 3 中用 Path 表示,有输入特征图和输出特征图之间路径在表 3 中用 Conn 表示。如表 3 所示,逐渐增加路径到模型中,并测试每一个组合模型的平均精度,可以看到模型精度是逐渐提高的。结果表明,本文方法通过获得更丰富的多尺度信息来提高 3D 目标检测精度。因此,本文方法通过将 2 种方式结合在一起来增强特征信息,进而提高模型的检测精度。

**表 2 本文方法与 Pseudo-LiDAR[11] 方法在 KITTI 验证集上汽车类别的 $AP_{bev}$ 和 $AP_{3D}$**

**Table 2 $AP_{bev}$ and $AP_{3D}$ of car category on KITTI validation set between the proposed method and Pseudo-LiDAR[11] method** %

| 方法 | $AP_{bev}$(IoU = 0.7) | | | $AP_{3D}$(IoU = 0.7) | | |
|---|---|---|---|---|---|---|
| | 简单 | 中等 | 困难 | 简单 | 中等 | 困难 |
| 本文方法 | 69.44 | 49.36 | 42.11 | 55.26 | 37.94 | 32.38 |
| PL + FP[11] | 69.7 | 48.1 | 41.8 | 54.9 | 36.4 | 31.1 |

**表 3 在 KITTI 数据集上对于 MpFPN 方法的消融实验**

**Table 3 Ablation experiment of MpFPN approach on KITTI dataset** %

| Path | Conn | $AP_{bev}$(IoU = 0.7) | | | $AP_{3D}$(IoU = 0.7) | | |
|---|---|---|---|---|---|---|---|
| | | 简单 | 中等 | 困难 | 简单 | 中等 | 困难 |
| × | × | 65.92 | 46.11 | 40 | 52.25 | 34.69 | 30.27 |
| √ | × | 68.01 | 48.15 | 41.21 | 54.78 | 36.88 | 31.42 |
| √ | √ | 69.44 | 49.36 | 42.11 | 55.26 | 37.94 | 32.38 |

### 3.5 检测结果示例

图 2 展示了 KITTI 验证集中几个场景的定性检测结果。可以观察到,在常见的街道场景中,本文方法可以准确检测场景中的物体,并且检测到的 3D 框在前视图像和点云上都很好地对齐。特别是当物体离相机很远而导致目标尺度小,本文方法仍然能够获得准确的检测结果。

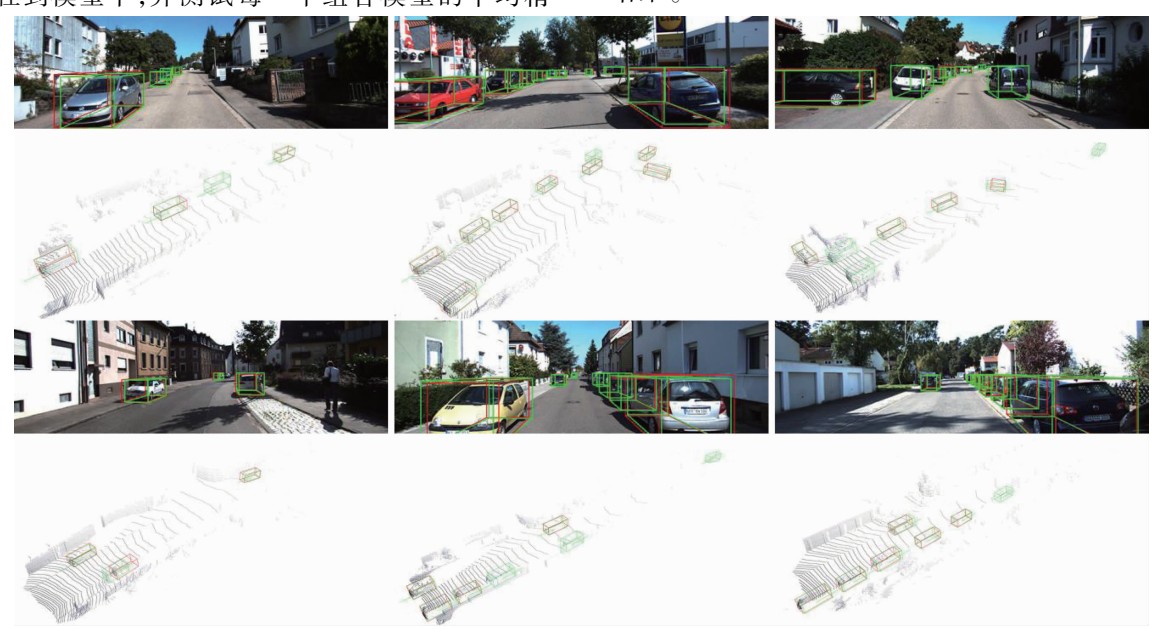


图 2 KITTI 验证集上的 3D 目标检测结果

Fig. 2 3D object detection results on KITTI validation set

## 4 结　论

1）针对 3D 目标检测任务，本文提出了一种基于立体图像的多路径特征金字塔网络 3D 目标检测方法。

2）设计了多路径特征金字塔网络（MpFPN）用于 3D 目标检测的多尺度特征提取。MpFPN 对特征金字塔网络进行了扩展，增加了自底向上的路径、由上至下的路径及输入特征图到输出特征图之间的连接，为联合 RPN 提供了更高语义信息和更细粒度空间信息的多尺度特征信息，充分考虑了多目标尺度差异问题，并通过实验验证了 MpFPN 的有效性。

3）实验结果表明，与现有方法进行比较，本文方法在 3D 目标检测 KITTI 数据集上取得了更好的检测结果。在没有深度数据监督的情况下，本文方法在 3D 目标检测任务上优于现有方法的性能。与一些具有深度数据监督的方法相比，本文方法也获得了可比的性能。

### 参考文献（References）


[1] WANG Z, JIA K. Frustum ConvNet: Sliding frustums to aggregate local point-wise features for a modal 3D object detection [C] // Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). Piscataway: IEEE Press, 2019: 1742-1749.

[2] SHI S, WANG X, LI H. PointRCNN: 3D object proposal generate-on and detection from point cloud [C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2019: 770-779.

[3] QI C R, LITANY O, HE K, et al. Deep Hough voting for 3D object detection in point clouds [C] // Proceedings of the IEEE International Conference on Computer Vision. Piscataway: IEEE Press, 2019: 9277-9286.

[4] SHI S, GUO C, JIANG L, et al. PV-RCNN: Point-voxel feature set abstraction for 3D object detection [C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2020: 10529-10538.

[5] BRAZIL G, LIU X. M3D-RPN: Monocular 3D region proposal network for object detection [C] // Proceedings of the IEEE International Conference on Computer Vision. Piscataway: IEEE Press, 2019: 9287-9296.

[6] KU J, PON A D, WASLANDER S L. Monocular 3D object detection leveraging accurate proposals and shape reconstruction [C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2019: 11867-11876.

[7] CHEN Y, TAI L, SUN K, et al. MonoPair: Monocular 3D object detection using pairwise spatial relationships [C] // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2020: 12093-12102.

[8] LIU L, WU C, LU J, et al. Reinforced axial refinement network for monocular 3D object detection [C] // European Conference on Computer Vision. Berlin: Springer, 2020: 540-556.

[9] CHEN X, KUNDU K, ZHU Y, et al. 3D object proposals using-stereo imagery for accurate object class detection [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 40(5): 1259-1272.

[10] LI P, CHEN X, SHEN S. Stereo R-CNN based 3D object detection for autonomous driving [C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2019: 7644-7652.

[11] WANG Y, CHAO W L, GARG D, et al. Pseudo-LiDAR from visual depth estimation: Bridging the gap in 3D object detection for autonomous driving [C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2019: 8445-8453.

[12] SUN J, CHEN L, XIE Y, et al. Disp R-CNN: Stereo 3D object detection via shape prior guided instance disparity estimation [C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2020: 10548-10557.

[13] GHIASI G, LIN T Y, LE Q V. NAS-FPN: Learning scalable feature pyramid architecture for object detection [C] // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2019: 7036-7045.

[14] ZHAO Q, SHENG T, WANG Y, et al. M2Det: A single-shot object detector based on multi-level feature pyramid network [C] // Proceedings of the AAAI Conference on Artificial Intelligence. Palo Alto: AAAI, 2019: 9259-9266.

[15] TAN M, PANG R, LE Q V. EfficientDet: Scalable and efficient object detection [C] // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2020: 10781-10790.

[16] 曹帅，张晓伟，马健伟. 基于跨尺度特征聚合网络的多尺度行人检测 [J]. 北京航空航天大学学报，2020，46(9)：1786-1796.
CAO S, ZHANG X W, MA J W. Transscale feature aggregation network for multiscale pedestrian detection [J]. Journal of Beijing University of Aeronautics and Astronautics, 2020, 46(9): 1786-1796 (in Chinese).

[17] 李晓光，付陈平，李晓莉，等. 面向多尺度目标检测的改进 Faster R-CNN 算法 [J]. 计算机辅助设计与图形学学报，2019，31(7)：1095-1101.
LI X G, FU C P, LI X L, et al. Improved faster R-CNN for multi-scale object detection [J]. Journal of Computer-Aided Design & Computer Graphics, 2019, 31(7): 1095-1101 (in Chinese).

[18] 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.

[19] LIU S, QI L, QIN H, et al. Path aggregation network for instance segmentation [C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE

Press,2018:8759-8768.

[20] 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.

[21] HE K,GKIOXARI G,DOLLÁR P,et al. Mask R-CNN[C]// Proceedings of the IEEE International Conference on Computer Vision. Piscataway:IEEE Press,2017:2961-2969.

[22] DENG J,DONG W,SOCHER R,et al. ImageNet:A large-scale hierarchical image database[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway:IEEE Press,2009:248-255.

[23] GEIGER A,LENZ P,URTASUN R. Are we ready for autonomous driving? The KITTI vision benchmark suite[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway:IEEE Press,2012:3354-3361.

[24] QIN Z,WANG J,LU Y. MonoGRNet:A geometric reasoning network for monocular 3D object localization[C]// Proceedings of the AAAI Conference on Artificial Intelligence. Palo Alto: AAAI,2019:8851-8858.

[25] MA X,WANG Z,LI H,et al. Accurate monocular 3D object detection via color-embedded 3D reconstruction for autonomous driving[C]// Proceedings of the IEEE International Conference on Computer Vision. Piscataway:IEEE Press,2019:6851-6860.

[26] QIN Z,WANG J,LU Y. Triangulation learning network:From monocular to stereo 3D object detection[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway:IEEE Press,2019:7615-7623.

[27] CHANG J R,CHEN Y S. Pyramid stereo matching network[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway:IEEE Press,2018:5410-5418.

[28] KU J,MOZIFIAN M,LEE J,et al. Joint 3D proposal generation and object detection from view aggregation[C]// Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). Piscataway:IEEE Press,2018:1-8.

# 3D object detection based on multi-path feature pyramid network for stereo images

SU Kaiqi, YAN Weiqing*, XU Jindong

(School of Computer and Control Engineering, Yantai University, Yantai 264005, China)

**Abstract**: 3D object detection is an important scene understanding task in computer vision and autonomous driving. However, most of these methods do not fully consider the large differences in scales between multiple objects. Thus, objects with a small scale are easily ignored, resulting in low detection accuracy. To address this problem, this paper proposes a 3D object detection method based on multi-path feature pyramid network (MpFPN) for stereo images. MpFPN extends feature pyramid network, adding a bottom-up path, top-down path, and connections between input and output features. It provides multi-scale feature information with higher semantic information and finer-grained spatial information for union region proposal network. Experimental results show that the proposed method achieves better results than comparative methods in easy, moderate and hard scenarios on the 3D object detection dataset KITTI.

**Keywords**: 3D object detection; feature pyramid network (FPN); stereo image; multi-scale; deep learning

**Received**: 2021-09-06; **Accepted**: 2021-09-17; **Published online**: 2021-10-18 09:56

**URL**: kns.cnki.net/kcms/detail/11.2625.V.20211015.1719.002.html

**Foundation items**: National Natural Science Foundation of China (61801414,62072391,62066013); Shandong Provincial Natural Science Foundation (ZR2019MF060); Shandong Province Higher Educational Science and Technology Key Program (J18KZ016)

* **Corresponding author**. E-mail: wqyan@tju.edu.cn

http://bhxb.buaa.edu.cn jbuaa@buaa.edu.cn
**DOI**: 10.13700/j.bh.1001-5965.2021.0535

# 基于时空注意力机制的新冠肺炎疫情预测模型

鲍昕[1], 谭智一[1,*], 鲍秉坤[1], 徐常胜[2]

(1. 南京邮电大学 通信与信息工程学院, 南京 210003;
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 100080)

**摘 要**: 新冠肺炎疫情持续蔓延给人类社会带来深远影响,准确预测各地区的病毒传播趋势对防控疫情而言至关重要。现有研究主要基于传统的时序预测模型和传染病模型,鲜有考虑疫情地区关联复杂和时序依赖性强的特点,限制了其疫情预测的性能。为此,针对新冠肺炎疫情的预测任务,提出了一种时空注意力驱动的自编码器框架。通过引入空间注意力机制捕捉病毒感染序列间的动态空间关联性,利用时间注意力机制挖掘病毒感染序列中复杂的时序依赖性,以此实现对不同地区的新冠肺炎病毒传播趋势的准确预测。在模型的编码器端,融合空间注意力机制的长短期记忆(LSTM)网络,关联目标地区与其他地区的病毒感染序列,提取该区域近期新冠肺炎疫情的时序特征。在模型的解码器端,将时间注意力机制引入基于 LSTM 网络的解码器中,通过捕捉病毒感染序列的时序依赖性推测未来的新冠肺炎疫情趋势变化。在多个公开的新冠肺炎疫情数据集上对所提模型进行验证,实验结果表明:所提模型的预测性能超越了 LSTM 等模型;在公开的欧洲部分国家新冠肺炎疫情数据集上,预测误差指标 RMSE 和 MAE 分别降低了 22.3% 和 25.0%,在中国部分省级单位新冠肺炎疫情数据集上,RMSE 和 MAE 分别降低了 10.1% 和 10.4%。

**关 键 词**: 新冠肺炎疫情预测; 注意力网络; 时空序列预测; 长短期记忆(LSTM)网络; 自编码器

**中图分类号**: TP183

**文献标志码**: A **文章编号**: 1001-5965(2022)08-1495-10

新冠肺炎疫情给全人类生命安全和全球公共卫生安全带来了巨大的威胁和挑战,对世界经济造成了严重冲击和损害。针对性的防疫政策和措施是遏制新冠肺炎疫情蔓延的关键,而疫情的传播趋势信息是制定防疫政策和措施的重要基石。因此,基于历史疫情数据预测未来不同地区的疫情趋势对疫情传播防控而言具有重要的意义。

与已知的全球性传染病(如严重急性呼吸综合征 SARS 和甲型 H1N1 流感)不同,新冠肺炎病毒在传播过程中具有如下特点:①地区间疫情关联复杂。新冠肺炎病毒在潜伏期内具有很强的传染性[1],容易随人流大范围传播,造成不同区域间的疫情存在高度的关联性[2]。②时序依赖性强。新冠肺炎病毒的潜伏期一般为 5 ~ 6 天,也可能长达 14 天[3],而且无症状感染者的存在,使得新冠肺炎病毒的宿主难以被及时发现,导致当前的感染人数与过去不同时段的感染情况存在着复杂的时序依赖性。上述特点使得新冠肺炎疫情预

**收稿日期**: 2021-09-07; **录用日期**: 2021-09-17; **网络出版时间**: 2021-11-02 14:13
**网络出版地址**: kns.cnki.net/kcms/detail/11.2625.V.20211101.1605.010.html
**基金项目**: 国家重点研发计划(2020AAA0106200); 国家自然科学基金(6193000388,61872424); 江苏省自然科学基金(BK20200037, BK20210595)

* **通信作者**. E-mail: tzy@njupt.edu.cn

**引用格式**: *鲍昕, 谭智一, 鲍秉坤, 等. 基于时空注意力机制的新冠肺炎疫情预测模型[J]. 北京航空航天大学学报, 2022, 48(8): 1495-1504. BAO X, TAN Z Y, BAO B K, et al. Prediction model of COVID-19 based on spatiotemporal attention mechanism[J]. Journal of Beijing University of Aeronautics and Astronautics, 2022, 48(8): 1495-1504 (in Chinese).*

测成为一项极具挑战性的任务。

研究人员开展了大量新冠肺炎疫情预测方面的研究工作,主要包括基于传统时序预测模型和基于传染病模型 2 类方法[4-5]。基于传统时序预测模型的方法大多通过拟合疫情数据与特定的时序预测模型来分析疫情的传播规律和发展趋势。然而,这类方法只关注和建模单一地区的疫情数据,忽视了由于人口流动造成的不同地区的疫情在空间上存在的复杂关联,因而限制了其最终的预测性能[6]。基于传染病模型的预测方法主要根据疫情的传播特点对经典的传染病模型进行扩展,如增加潜伏期、无症状感染者等[7]。然而,这类方法通常基于恒定的传染强度和特定的传播函数,并不能有效地建模疫情数据在时序上的动态依赖性。本文针对新冠肺炎疫情的传播特点,提出了一种时空注意力驱动的自编码器框架,引入空间注意力机制和时间注意力机制来建模病毒感染序列中的空间关联性和时序依赖性,实现了对不同地区疫情传播趋势的准确预测。具体而言,在模型的编码器端,先引入空间注意力机制,学习目标地区与其相关地区的病毒感染序列之间的关联,再基于该关联信息引入相关地区的疫情时序数据,并将其与目标地区的疫情时序数据融合,经由长短期记忆(long short-term memory,LSTM)网络抽取用于预测目标地区疫情的强化时序特征;在模型的解码器端,通过融合时间注意力机制与 LSTM 网络,有效捕捉疫情时序中的动态时序依赖性,更加准确预测未来一段时间内的疫情动态。在开放的新冠肺炎疫情数据集上进行的大量评估实验表明,本文模型优于 LSTM 等模型。本文的主要创新点如下:

1)利用空间注意力机制捕获不同地区疫情之间的动态关系,建模了相关地区疫情对预测目标地区疫情传播趋势的影响。

2)引入时间注意力机制捕捉不同时间间隔内的疫情动态间的时序依赖关系,更好地估计未来一段时间内的疫情动态。

3)提出一种基于自编码器结构的时空序列预测模型,分别在编码器端和解码器端融合了基于空间注意力机制的多地区疫情关联建模和基于时间注意力机制的疫情时序依赖性挖掘,实现了端到端的新冠肺炎疫情预测。

## 1 相关工作

### 1.1 新冠肺炎疫情传播趋势预测

现有的新冠肺炎疫情传播趋势预测工作大致可分为 2 类方法:①基于传统的时序预测模型。这类方法将传统的时序预测模型,包括差分整合移动平均自回归(autoregressive integrated moving average,ARIMA)模型[8]、LSTM 网络[5]和先知模型[9]等,应用到各种疫情预测任务中。例如,Nguyen 等[10]提出利用 LSTM 网络预测台湾中部地区呼吸系统疾病的病例数。这类方法借助 LSTM 和 ARIMA 等模型良好的泛化能力,可以有效挖掘疫情序列数据中复杂的时序关系。因此,基于传统的时序预测模型及这类模型的变种也被广泛应用于新冠肺炎疫情的预测任务中。例如,Rauf 等[11]提出了一种基于门控循环单元(gated recurrent unit,GRU)的国家级新冠肺炎疫情传播趋势预测框架,综合考虑了时变的疫情信息与当地人口、年龄结构等外部因素。但是这类方法大多只关注和建模单一地区疫情数据中的时序关系,忽视了不同地区疫情数据之间的空间关联,在面对地区间疫情关联复杂的新冠肺炎疫情预测任务时,其性能仍有待提升。②基于传染病模型的方法。传染病数学建模是了解传染病的传播规律、预测传染病的传播趋势和分析防疫策略效果的重要方式之一。例如,Wang 等[12]提出利用 SEIR 模型预测严重急性呼吸综合征(SARS)的传播趋势,成功预测其拐点,并评估了防疫措施的有效性。在新冠肺炎疫情开始在全球范围内蔓延时,研究人员最初利用经典的传染病模型(如 SIR 和 SEIR 等)来分析和预测疾病的传播趋势,并利用分析结果指导预防和控制措施[13-14]。随着对新冠肺炎疫情的深入了解,部分相关工作根据疫情传播的新特点,如潜伏期、无症状感染者等,在经典的传染病模型的基础上进行了一定的扩展[15]。例如,Liu 和 Fong[7]在经典的 SEIR 模型的基础上,考虑了无症状感染者、死亡患者和转阴复阳等人群,提出了融合这些因素的 SEAIRD 模型,相比于传统的 SEIR 模型,该模型对疫情数据的拟合程度更好。然而现有的基于传染病模型的方法都是根据当前时段的感染情况估算感染率,认为在任意时刻,感染者对暴露人群具有恒定的感染率[16]。然而,新冠肺炎病毒具有不稳定的潜伏期及无症状感染者的存在,使得病毒宿主被发现和隔离的概率不尽相同,导致疫情在时序上存在复杂的关联。因此,基于恒定感染率的传染病模型方法难以有效建模这一时序关联,致使其缺乏对疫情动态的刻画能力,具有一定的局限性。

### 1.2 时空序列预测

准确预测新冠肺炎疫情的传播需要综合考虑

疫情序列数据中的时序关联信息和序列间的空间关联信息。因此,新冠肺炎疫情预测可被视为一个时空序列预测任务。

时空序列是时间序列在空间维度上的扩展,可以视为多个时间序列的集合[17]。时空序列预测任务中,不仅需要考虑序列内的时序依赖性,还要考虑序列间的空间关联性。传统的时空序列预测方法将时空序列看作是多个独立时序来处理,缺乏对序列数据之间的关联关系进行建模,因此,难以挖掘出时空序列数据中复杂的时空模式。近年来兴起的基于深度学习的方法在时空序列预测问题上取得了更好的预测效果[18]。时空序列预测方法中,建模序列间空间关联的方式包括以下 2 种:①从序列数据特征中挖掘序列数据间潜在的空间关联性。这种方式通常有 2 种具体的实现途径:第1 种途径是基于注意力机制挖掘序列间的空间关联,如 Liang 等[19]提出了一种多级注意力机制来捕捉不同空气质量传感器读数序列间的关联性;第 2 种途径是将时空序列视为一系列有时间顺序的图片,用卷积神经网络(convolutional neural network,CNN)提取数据中的空间关联,如 Yao 等[20]提出了一种多视角时空神经网络模型 DMVST-Net 来解决出租车需求预测问题,该模型分别利用 CNN 和 LSTM 来提取不同区域出租车需求序列中的空间关联性和时序依赖性。②引入与序列数据对应的实体之间的关系等外部信息,通过实体关系嵌入构建序列间的空间关联性。例如,Wang 等[21]在预测交通道路流量的问题上引入了不同道路的位置和连接关系来建模车流量序列之间的空间关联性,在交通流量预测问题上取得了不错的预测效果。然而,这些时空序列预测模型所建模的序列间关联性大多是静态的,而在新冠肺炎疫情数据中,地区疫情序列间的关联是随时间变化的,这种静态关联建模的方式并不适用于提取地区疫情间的动态关联。此外,与大气数据和交通流量等具有明显规律性或周期性的数据不同,新冠肺炎疫情序列数据在后验分布上不存在明显的周期性和规律性,其时序依赖性更为复杂。

## 2　新冠肺炎疫情时空序列预测模型

### 2.1　问题描述

本文中的新冠肺炎疫情预测任务是:基于任意给定地区的近期疫情数据,预测该地区在未来一段时间内的每日新增确诊人数序列。假设共有 $N$ 个地区,每个地区的实时疫情数据包括每日新增确诊、累计确诊、新增死亡及累计死亡病例数等 $d$ 维特征。用 $\boldsymbol{X}^i=\{\boldsymbol{x}_1^i,\boldsymbol{x}_2^i,\cdots,\boldsymbol{x}_T^i\}\in\mathbf{R}^{T\times d}$ 表示地区 $i$ 最近 $T$ 个时间段的疫情数据,$\boldsymbol{x}_t^i\in\mathbf{R}^d$ 表示地区 $i$ 在第 $t$ 个时间段的疫情数据。由于任意地区的疫情都可能与其他地区的疫情存在关联,模型的输入数据中也包括过去 $T$ 时间段内所有地区的疫情数据,即 $\boldsymbol{X}=\{\boldsymbol{X}^1,\boldsymbol{X}^2,\cdots,\boldsymbol{X}^N\}$。基于上述符号设定,将本文的预测目标,即目标地区在未来 $\tau$ 个时间段内的新增确诊人数表示为 $\hat{\boldsymbol{Y}}=\{\hat{\boldsymbol{y}}^1,\hat{\boldsymbol{y}}^2,\cdots,\hat{\boldsymbol{y}}^N\}\in\mathbf{R}^{N\times\tau}$。

### 2.2　整体框架

本文提出的时空注意力驱动的新冠肺炎疫情传播趋势预测方法的框架如图 1 所示。模型的整体架构采用了自编码器的结构,编解码器中各包含一个 LSTM 网络,分别用于提取疫情序列数据的时序特征和预测的未来疫情趋势。具体来说,本文模型主要包括 2 个关键部分:①编码器端的空间注意力机制。编码器端的 LSTM 网络融合了空间注意力机制,用以从不同地区的疫情数据中学习地区间的疫情关联。如图 1 所示,从疫情时空序列数据中动态地提取地区间疫情的关联性而非静态的关系嵌入。②解码器端的时间注意力机制。在解码器端的 LSTM 网络中融入了时间注意力机制,使其能够基于与当前输出最相关的时序位置上的序列信息生成预测结果,从而提升模型对疫情数据的刻画能力。

### 2.3　地区间新冠肺炎疫情的关联性建模

图 2 为欧洲部分国家每日新增确诊人数曲线。可以明显看出,不同国家的疫情曲线间存在复杂的关联性。以 2020 年 4 月至 11 月期间为例,由于人们防疫意识薄弱,地区间的大范围人口流动造成新冠肺炎疫情在相关地区内迅速蔓延。也就是说,对于任意地区,该区域当前的疫情不仅与该地区过去的本土病例有关,还可能与来自其他地区的输入病例有关。因此,在预测特定地区新冠肺炎疫情时,需要考虑相关地区的疫情对目标地区的影响。鉴于不同地区间疫情的相互影响是随时间变化的,本文引入了一种新的空间注意力机制来捕捉不同地区之间疫情的动态关联性。空间注意力单元的模块结构如图 1 中右边虚线框所示。假设地区 $i$ 为疫情预测的目标地区,地区 $k$ 为地区 $i$ 的相关地区,则 $t$ 时刻 2 个地区疫情之间的相关性权重 $e_t^k$ 可由如下公式计算:

$$e_t^k=\boldsymbol{v}_s^{\mathrm{T}}\tanh(\boldsymbol{w}_s[\boldsymbol{h}_{t-1}^i,\boldsymbol{s}_{t-1}^i]+\boldsymbol{w}_s'\boldsymbol{X}^k\boldsymbol{u}_s+\boldsymbol{b}_s)\quad(1)$$

式中:$\boldsymbol{v}_s,\boldsymbol{u}_s,\boldsymbol{b}_s\in\mathbf{R}^T$,$\boldsymbol{w}_s\in\mathbf{R}^{T\times 2m}$ 和 $\boldsymbol{w}_s'\in\mathbf{R}^{T\times T}$ 为模

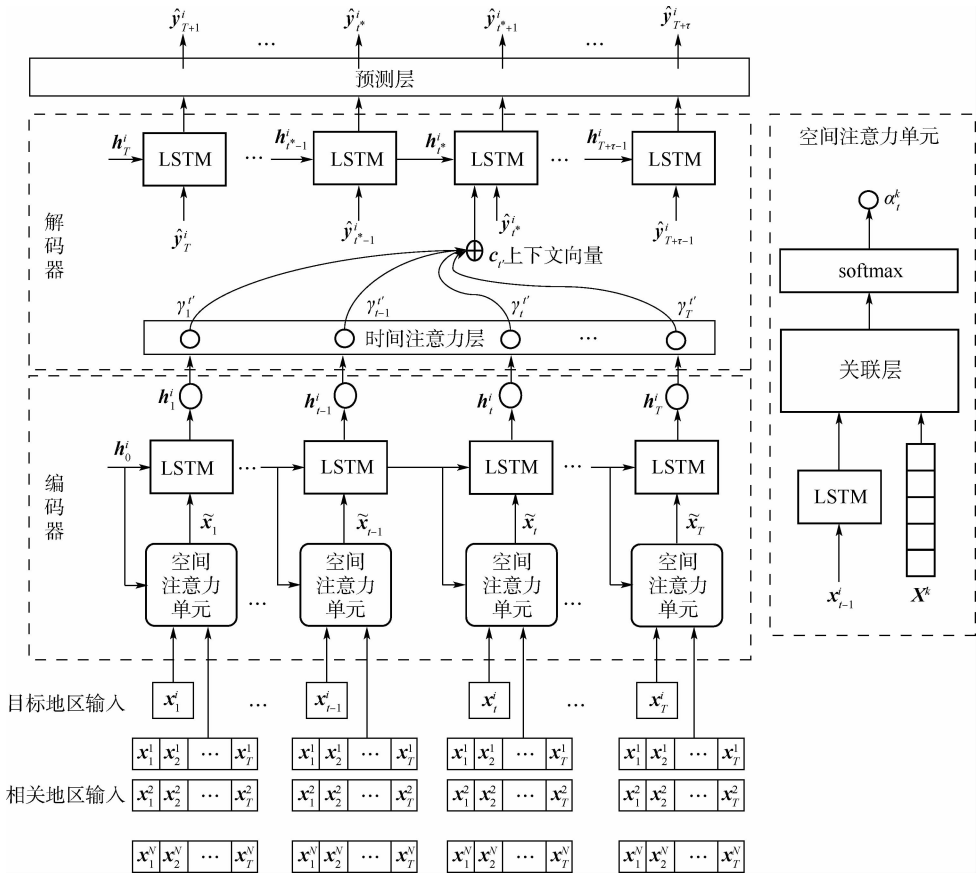



图 1 基于时空注意力机制的新冠肺炎疫情预测框架

Fig. 1 COVID-19 prediction framework based on spatial temporal attention mechanism

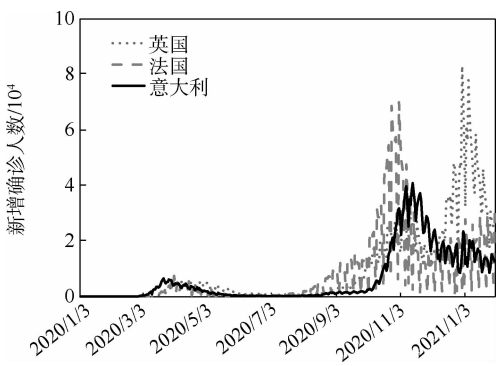



图 2 欧洲部分国家新增确诊人数曲线

Fig. 2 Curves of new cases in some European countries

型中可学习的参数。

在式(1)中,目标地区 $i$ 对地区 $k$ 的疫情的相关性权重基于地区 $k$ 的疫情数据 $\boldsymbol{X}^k$ 和本地近期疫情数据计算得到。而本地的近期疫情数据则利用编码器中上一时间步的隐藏状态 $\boldsymbol{h}_{t-1}^i$ 和细胞状态 $\boldsymbol{s}_{t-1}^i$ 来表示。

式(2)中用 softmax 函数对式(1)计算得到的相关性权重进行归一化,获得最终的地区 $i$ 对地区 $k$ 的注意力权重。

$$\alpha_t^k = \frac{\exp(e_t^k)}{\sum_{i=1}^{N}\exp(e_t^i)} \tag{2}$$

利用式(2)中的注意力权重值来计算空间注意力加权后的输出向量:

$$\tilde{\boldsymbol{x}}_t = (\alpha_t^1\boldsymbol{x}_t^1, \alpha_t^2\boldsymbol{x}_t^2, \cdots, \alpha_t^N\boldsymbol{x}_t^N) \tag{3}$$

### 2.4 新冠肺炎疫情的时序依赖性建模

由于新冠肺炎疫情数据中当前的确诊人数与过去不同时段的情况存在着复杂的时序关联,本文在解码器端融入一个时间注意力机制来提取疫情序列数据中的时序依赖性。该时间注意力机制可以基于当前预测输出位置,对目标序列中不同时序位置上的信息赋予特定的相关性权重,以生成预测输出序列。具体来说,用式(4)来计算预测时间步 $t'$ 的输出与编码器中不同时间间隔的隐藏状态输入的相关性权重 $u_{t'}^o$。

$$u_{t'}^o = \boldsymbol{v}_{\mathrm{d}}^{\mathrm{T}}\tanh(\boldsymbol{w}_{\mathrm{d}}[\boldsymbol{h}_{t'-1}^i, \boldsymbol{s}_{t'-1}^i] + \boldsymbol{w}'_{\mathrm{d}}\boldsymbol{h}_o^i + \boldsymbol{b}_{\mathrm{d}}) \tag{4}$$

式中:$o \in \{1,2,\cdots,T\}$;$\boldsymbol{h}_o^i$ 表示编码器 $o$ 时刻对应输出的隐藏状态;$t' \in [T, T+\tau]$;$\boldsymbol{h}_{t'-1}^i$ 和 $\boldsymbol{s}_{t'-1}^i$ 分别表示解码器中 $t'-1$ 时刻的隐藏状态和细胞状态;$\boldsymbol{v}_{\mathrm{d}}, \boldsymbol{b}_{\mathrm{d}} \in \mathbf{R}^m$,$\boldsymbol{w}_{\mathrm{d}} \in \mathbf{R}^{m\times 2n}$ 和 $\boldsymbol{w}'_{\mathrm{d}} \in \mathbf{R}^{m\times m}$ 为通过训练学到的模型参数。

利用 softmax 函数对 $u_{t'}^o$ 进行归一化处理,得到 $t'$ 时刻的输出对 $o$ 时刻隐藏状态的注意力权重 $\gamma_{t'}^o$,计算式如下:

$$\gamma_{t'}^{o} = \frac{\exp(u_{t'}^{o})}{\sum_{j=1}^{T}\exp(u_{t'}^{j})} \tag{5}$$

利用注意力权重对相应的编码器隐藏状态输出进行加权求和,得到上下文向量 $\boldsymbol{c}_{t'}$,计算公式如下:

$$\boldsymbol{c}_{t'} = \sum_{o=1}^{T}\gamma_{t'}^{o}\boldsymbol{h}_{o}^{i} \tag{6}$$

### 2.5　基于自编码器的新冠肺炎疫情预测框架

本文模型采用自编码器结构,在编码器中,将空间注意力网络的输出向量 $\tilde{\boldsymbol{x}}_t$ 作为编码器新的输入,来更新 $t$ 时刻的隐藏状态。

$$\boldsymbol{h}_{t}^{i} = f_{\mathrm{e}}(\boldsymbol{h}_{t-1}^{i}, \tilde{\boldsymbol{x}}_{t}) \tag{7}$$

式中:$\boldsymbol{h}_{t-1}^{i}$表示上一时刻的隐藏状态;$f_{\mathrm{e}}$ 表示编码器中的 LSTM 单元结构。

在解码器端,利用计算得到时间注意加权的上下文向量 $\boldsymbol{c}_{t'-1}$后,结合解码器 $t'-1$ 时刻的输出 $\hat{\boldsymbol{y}}_{t'-1}^{i}$,更新 $t'$时刻的隐藏状态。

$$\boldsymbol{h}_{t'}^{i} = f_{\mathrm{d}}(\boldsymbol{h}_{t'-1}^{i}, [\hat{\boldsymbol{y}}_{t'-1}^{i}; \boldsymbol{c}_{t'}]) \tag{8}$$

式中:$f_{\mathrm{d}}$ 表示解码器中的 LSTM 单元结构。

结合上下文向量 $\boldsymbol{c}_{t'}$和解码器中更新后的隐藏状态输出 $\boldsymbol{h}_{t'}^{i}$,并送入全连接神经网络的预测输出层,得到最终的预测结果 $\hat{\boldsymbol{y}}_{t'}^{i}$。

$$\hat{\boldsymbol{y}}_{t'}^{i} = \boldsymbol{v}_{\mathrm{o}}^{\mathrm{T}}(\boldsymbol{w}_{\mathrm{m}}[\boldsymbol{c}_{t'}; \boldsymbol{h}_{t'-1}^{i}] + \boldsymbol{b}_{\mathrm{m}}) + b_{\mathrm{o}} \tag{9}$$

式中:$\boldsymbol{w}_{\mathrm{m}} \in \mathbf{R}^{n\times(m+n)}$,$\boldsymbol{v}_{\mathrm{o}}, \boldsymbol{b}_{\mathrm{m}} \in \mathbf{R}^{n}$,$b_{\mathrm{o}} \in \mathbf{R}$。

本文使用均方误差(MSE)作为损失函数,并使用反向传播算法来训练模型参数。损失函数的定义如下:

$$L(\theta) = \frac{1}{N}\sum_{i=1}^{N}\|\boldsymbol{y}^{i} - \hat{\boldsymbol{y}}^{i}\|_{2}^{2} \tag{10}$$

## 3　实　验

### 3.1　数 据 集

为了验证本文模型的性能,在世界卫生组织(WHO)发布的新冠肺炎疫情数据集上进行实验。在实验中,按地域将数据集划分为欧洲部分国家新冠肺炎疫情数据集和中国部分省级单位新冠肺炎疫情数据集。

欧洲部分国家新冠肺炎疫情数据集中包括欧洲国家从 2020 年 1 月 24 日到 2021 年 3 月 22 日期间每日新增确诊人数、累计确诊人数、新增死亡人数及累计死亡人数的序列数据。

中国部分省级单位新冠肺炎疫情数据集以省级单位(含直辖市、特别行政区)划分,记录了省级单位从 2020 年 1 月 22 日到 2020 年 10 月 15 日期间每日实时确诊数据,包括累计确诊人数、康复者人数和死亡人数等。中国防疫措施得当,2020 年 4 月开始,大多数省份的新增确诊人数几乎为零,因此,将中国部分省级单位新冠肺炎疫情数据集上的预测目标设置为每日的累计确诊人数。表 1 中列举了 2 个数据集的详细信息及训练集和测试集的数据时间范围划分情况。

**表 1　数据集描述**

**Table 1　Dataset description**

| 数据集 | 预测目标 | 时间范围 | 训练集时间划分 | 测试集时间划分 |
|---|---|---|---|---|
| 欧洲部分国家新冠肺炎疫情数据集 | 新增确诊人数 | 2020/1/24—2021/3/22 | 2020/1/24—2020/12/24 | 2020/12/25—2021/3/22 |
| 中国部分省级单位新冠肺炎疫情数据集 | 累计确诊人数 | 2020/1/22—2020/10/15 | 2020/1/22—2020/8/22 | 2020/8/23—2020/10/15 |

### 3.2　评价标准和实验设置

为了评估预测性能,本文采用均方根误差(RMSE)、平均绝对误差(MAE)及平均绝对百分比误差(MAPE)作为模型的评价指标,其是时空序列预测任务中普遍使用的评价指标,RMSE 和 MAE 描述了预测值和真实值的实际差异大小,而 MAPE 描述了预测误差相对于真实值的百分比。度量指标值越低,预测值与真实值越接近,模型的预测性能也就越好。

RMSE 的计算公式如下:

$$\mathrm{RMSE} = \sqrt{\frac{1}{T}\sum_{t=1}^{T}(\boldsymbol{y}_{t}^{i} - \hat{\boldsymbol{y}}_{t}^{i})^{2}} \tag{11}$$

MAE 的计算公式如下:

$$\mathrm{MAE} = \frac{1}{T}\sum_{t=1}^{T}|\boldsymbol{y}_{t}^{i} - \hat{\boldsymbol{y}}_{t}^{i}| \tag{12}$$

MAPE 的计算公式如下:

$$\mathrm{MAPE} = 100\,\frac{1}{T}\sum_{t=1}^{T}\frac{|\boldsymbol{y}_{t}^{i} - \hat{\boldsymbol{y}}_{t}^{i}|}{\boldsymbol{y}_{t}^{i}} \tag{13}$$

本文实验主要分为模型性能验证和消融实验 2 个部分。在模型性能验证的对比实验中,将所有预测模型的输入序列长度设置为 $T=7$,预测输出序列长度设置为 $\tau=7$。在相同输入输出序列长度下,比较本文模型与其他模型在疫情预测上的性能差异。在消融实验中,对比了不同输入序列长度和预测输出序列长度下,本文模型(STA-EP)与去掉空间注意力机制和去掉时间注意力机制的 2 个变体模型的性能差异。从 $T=\{7,14,21,28,35,42,49\}$ 选取时间窗口,从 $\tau=\{7,8,9,10,11,12,13\}$ 选择输出序列长度,进行实验。在训练阶段,初始学习率设置为 0.001,模型的 drop-

out设置为0.3。为了优化模型的性能,在模型编码器和解码器中采用stacked LSTM,层数设置为2,并采用网格搜索从{32,64,128,256}确定隐藏层数。实验中,用80%的数据作为训练集,20%的数据作为验证集来优化模型的参数设置。

### 3.3 性能验证

本节实验对比了几种当前新冠肺炎疫情预测研究中比较有代表性的时序预测方法,包括LSTM、GRU和Seq2Seq模型,以及通过静态关系嵌入建模时空序列数据的空间关联性的预测模型T-GCN。

1) LSTM[8]。一种基于门控机制的循环神经网络,能够在一定程度上缓解普通RNN中存在的长期时序依赖性问题。

2) GRU[11]。LSTM网络的一种变体,通过简化时序单元结构提升其计算效率。

3) Seq2Seq[22]。使用一个循环神经网络将输入序列编码为特征表示,并通过另一个循环神经网络迭代地进行预测。本节实验中,该模型的循环神经网络均采用LSTM网络。

4) T-GCN[23]。利用图卷积网络(GCN)学习交通网络图中道路的连接关系,建模交通时空序列数据的空间关联性,利用GRU获取数据的时间依赖性。本节实验中,通过嵌入地区间的相邻关系图建模地区疫情的空间关联性。

表2和表3分别为本文模型和对比方法在2个数据集上的结果比较(表中仅展示了部分数据结果)。由表2可知,以平均RMSE和MAE作为标准,本文模型的性能比欧洲部分国家新冠肺炎疫情数据集中最先进的baseline方法Seq2Seq模型表现更加优异,平均RMSE和MAE分别降低了22.3%和25.0%。由表3可知,本文模型的误差指标RMSE和MAE比中国部分省级单位新冠肺炎疫情数据集上最先进的baseline方法T-GCN模型分别下降了10.1%和10.4%。此外,从预测结果来看,本文模型比对比方法在MAPE指标上也有显著提高。从模型性能比较的实验结果可以看出,本文方法RMSE、MAE及MAPE指标都要优于对比方法,验证了本文方法在新冠肺炎疫情预测问题上的有效性。

造成上述性能差异的主要原因在于:对比方法中代表性的时序预测方法只考虑了单个地区的疫情序列数据,忽视了地区疫情序列数据间的关联性,致使其不能有效考虑和估计地区间的疫情影响,从而带来更大的预测误差。值得注意的是,T-GCN模型通过嵌入地区间的相邻关系图,建模地区疫情间的空间关联性。在中国部分省级单位新冠肺炎疫情数据集上,相对于时序预测方法,取得了一定的成效;然而在欧洲部分国家新冠肺炎

**表2 不同模型在欧洲部分国家新冠肺炎疫情数据集上的预测性能比较(部分数据)**

**Table 2 Prediction performance comparison among different methods in COVID-19 epidemic dataset of some European countries (partial data)**

| 国家 | LSTM | | | GRU | | | Seq2Seq | | |
|---|---|---|---|---|---|---|---|---|---|
| | RMSE | MAE | MAPE | RMSE | MAE | MAPE | RMSE | MAE | MAPE |
| 阿尔巴尼亚 | 703.92 | 691.41 | 0.86 | 701.24 | 688.64 | 0.86 | 662.86 | 650.38 | 0.81 |
| 丹麦 | 938.59 | 903.71 | 0.86 | 929.71 | 894.01 | 0.87 | 901.43 | 860.9 | 0.81 |
| 罗马尼亚 | 3 188.53 | 3 075.03 | 0.97 | 3 184.12 | 3 070.45 | 0.97 | 3 157.08 | 3 042.02 | 0.95 |
| 奥地利 | 1 786.8 | 1 736.26 | 0.94 | 1 781.35 | 1 729.26 | 0.94 | 1 750.72 | 1 696.74 | 0.91 |
| 希腊 | 1 036.59 | 980.39 | 0.88 | 1 032.73 | 975.95 | 0.88 | 1 002.07 | 943.65 | 0.83 |
| 德国 | 13 128.95 | 12 358.32 | 0.99 | 13 125.13 | 12 354.14 | 0.99 | 13 102.8 | 12 325.16 | 0.98 |
| 英国 | 26 716.02 | 26 339.62 | 1.68 | 26 711.37 | 26 335.18 | 1.62 | 26 705.26 | 26 295.7 | 0.99 |
| 法国 | 20 001.07 | 3 911.05 | 0.99 | 19 997.16 | 18 644.59 | 0.99 | 19 951.54 | 18 591.85 | 0.98 |
| 平均 | 4 112.48 | 3 911.05 | 1.68 | 4 105.65 | 3 903.92 | 1.62 | 4 075.76 | 3 868.85 | 1.98 |

| 国家 | T-GCN | | | STAEP(本文模型) | | |
|---|---|---|---|---|---|---|
| | RMSE | MAE | MAPE | RMSE | MAE | MAPE |
| 阿尔巴尼亚 | 798.227 3 | 764.024 9 | 0.96 | 565.97 | 552.767 | 0.65 |
| 丹麦 | 1 013.16 | 819.04 | 0.96 | 754.97 | 679.68 | 0.43 |
| 罗马尼亚 | 3 271.24 | 3 088.82 | 0.98 | 1 563.26 | 1 398.67 | 0.49 |
| 奥地利 | 1 868.26 | 1 769.33 | 0.97 | 1 064.07 | 997.13 | 0.51 |
| 希腊 | 1 252.66 | 1 063.47 | 0.96 | 740.09 | 675.71 | 0.44 |
| 德国 | 13 385.03 | 11 823.22 | 0.99 | 10 243.3 | 9 217.24 | 0.62 |
| 英国 | 27 062.34 | 25 433.32 | 0.99 | 23 676.71 | 23 183.08 | 0.75 |
| 法国 | 20 541.78 | 19 158.85 | 0.99 | 17 127.35 | 15 549.89 | 0.72 |
| 平均 | 4 383.86 | 3 922.16 | 1.12 | 3 166.93 | 2 902.36 | 0.92 |

表 3　不同模型在中国部分省级单位新冠肺炎疫情数据集上的预测性能比较(部分数据)

Table 3　Prediction performance comparison among different methods in COVID-19 epidemic dataset of some Chinese provinces (partial data)

| 省(直辖市、特别行政区) | LSTM | | | GRU | | | Seq2Seq | | |
|---|---|---|---|---|---|---|---|---|---|
| | RMSE | MAE | MAPE | RMSE | MAE | MAPE | RMSE | MAE | MAPE |
| 安徽省 | 910.56 | 911.04 | 0.91 | 909.37 | 909.06 | 0.91 | 829.33 | 829.72 | 0.83 |
| 湖南省 | 938.96 | 938.78 | 0.92 | 937.39 | 937.31 | 0.92 | 857.86 | 858.03 | 0.84 |
| 河南省 | 1 198.18 | 1 198.53 | 0.93 | 1 196.62 | 1 196.84 | 0.93 | 1 117.44 | 1 116.79 | 0.87 |
| 湖北省 | 68 058.96 | 68 059.28 | 0.99 | 68 057.51 | 68 057.81 | 0.99 | 67 977.7 | 67 976.45 | 0.99 |
| 江西省 | 854.42 | 855.05 | 0.91 | 853.39 | 853.39 | 0.91 | 773.9 | 773.36 | 0.82 |
| 重庆 | 504.077 | 503.97 | 0.86 | 502.49 | 502.63 | 0.86 | 422.71 | 423.7 | 0.83 |
| 香港 | 4 886.68 | 4 886.25 | 0.98 | 4 885.07 | 4 884.77 | 0.98 | 4 805.51 | 4 804.7 | 0.96 |
| 江苏省 | 585.54 | 585.47 | 0.87 | 584.08 | 584.21 | 0.87 | 504.39 | 504.22 | 0.75 |
| 平均 | 2 661.7 | 2 661.39 | 0.78 | 2 660.65 | 2 660.38 | 0.92 | 2 592.01 | 2 591.38 | 0.59 |

| 省(直辖市、特别行政区) | T-GCN | | | STAEP(本文模型) | | |
|---|---|---|---|---|---|---|
| | RMSE | MAE | MAPE | RMSE | MAE | MAPE |
| 安徽省 | 953.89 | 942.18 | 0.95 | 106.16 | 86.22 | 0.16 |
| 湖南省 | 981.89 | 969.9 | 0.95 | 122.41 | 99.24 | 0.16 |
| 河南省 | 1 241.01 | 1 226.61 | 0.96 | 212.95 | 186.52 | 0.1 |
| 湖北省 | 68 101.24 | 68 086.82 | 0.99 | 54 061.26 | 54 059.26 | 0.8 |
| 江西省 | 897.24 | 882.82 | 0.95 | 96.81 | 77.65 | 0.15 |
| 重庆 | 547.06 | 535.36 | 0.92 | 72.21 | 58.87 | 0.14 |
| 香港 | 4 981.62 | 4 970.01 | 0.99 | 7 666.55 | 7 536.1 | 1.51 |
| 江苏省 | 628.56 | 616.86 | 0.93 | 79.7 | 64.47 | 0.23 |
| 平均 | 2 231.44 | 2 222.01 | 1.82 | 2 006.32 | 1 990.52 | 0.51 |

疫情数据集上,平均 RMSE 和平均 MAE 却比简单的时序预测方法大,这是因为相比于中国以相邻地区关联为主的省级单位间疫情关联,欧洲各国由于一体化进程带来各国间频繁且不受地理距离限制的人口流动,其国家间的疫情关联显得更加复杂。此外,在疫情的不同发展时期,地区间采取防疫措施的类型和力度差异等多方面因素导致地区间的疫情关联是动态变化的。而本文模型在自编码器的结构基础上引入空间注意力机制,通过动态捕捉不同地区疫情的关联性,建模了其他地区疫情对目标地区疫情的影响。此外,通过在解码器端融合时间注意力机制,更加有效地挖掘疫情序列的时序依赖性,进一步提高了模型的预测性能。

### 3.4　消融实验

为了进一步验证本文提出的新冠肺炎疫情时空序列预测模型中每个部分的有效性,利用完整的模型与 2 个变体模型在欧洲部分国家新冠肺炎疫情数据集上进行了消融实验。STAEP_NS 表示本文模型去掉编码器中的空间注意力机制的模型,STAEP_NT 表示 STAEP 去掉解码器中的时间注意力机制的模型。此外,通过改变输入和输出序列的长度,进一步对比不同关键性参数下模型性能的差异。

#### 3.4.1　输入序列长度变化下的消融实验

本节实验中,将预测输出序列长度固定为 7 天,不断调整编码器端的输入序列长度。实验结果如图 3 所示。可以看出,随着编码器端输入序列长度的增加,最初预测误差会有一个逐渐减小的过程,这是因为更长的输入序列中包含了更

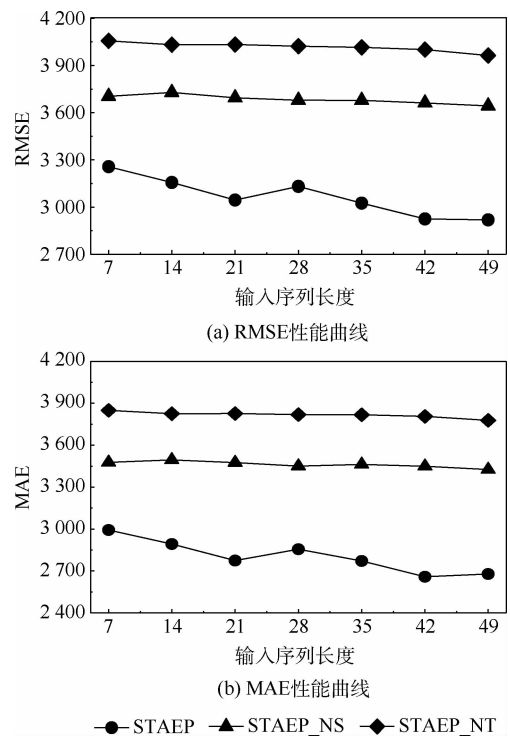



图 3　输入序列长度对模型性能的影响曲线

Fig. 3　Influence curves of input sequence length on model performance

多的历史疫情信息,这些信息可以使模型从中学到更丰富的数据特征,因此能够更加有效地预测疫情的动态变化。但是当输入序列过长时,由于LSTM 网络难以建立时间跨度过大的历史依赖关系,模型性能提升速度呈现明显减缓的趋势。另外,在不同输入序列长度下,本文模型比 2 个变体模型的表现更加优异。STAEP_NT 变体由于去掉了解码器中的时间注意力机制,在处理长序列输入时难以关注到对当前预测输出比较关键的时序位置信息,导致预测误差明显增加。在STAEP_NS 变体中,去掉了编码器中的空间注意力机制,其性能也远逊于本文模型,说明通过空间注意力机制捕捉到的地区间疫情关联有助于提升对目标地区疫情动态预测的效果。

3.4.2 输出序列长度变化下的消融实验

本节实验中,将编码器输入序列长度固定为 7 天,逐步增加解码器端输出序列的长度。实验结果如图 4 所示。3 种方法在欧洲部分国家新冠肺炎疫情数据集上的表现具有相似的趋势。逐步增加预测输出序列长度时,模型的预测误差会逐渐增大。可以看到,模型性能下降趋势是比较平缓的,说明本文模型对输入数据的利用是充分有效的。而从消融实验的对比结果来看,图 4 和图 3表现出相似的趋势。基于时空注意力机制的方法相较于2 种变体模型具有更好的预测性能,尤其是相对于 STAEP_NT 变体模型而言,二者的性能差距更为明显。图 4 的结果进一步验证了本文提出的时空注意力机制对模型性能的贡献。

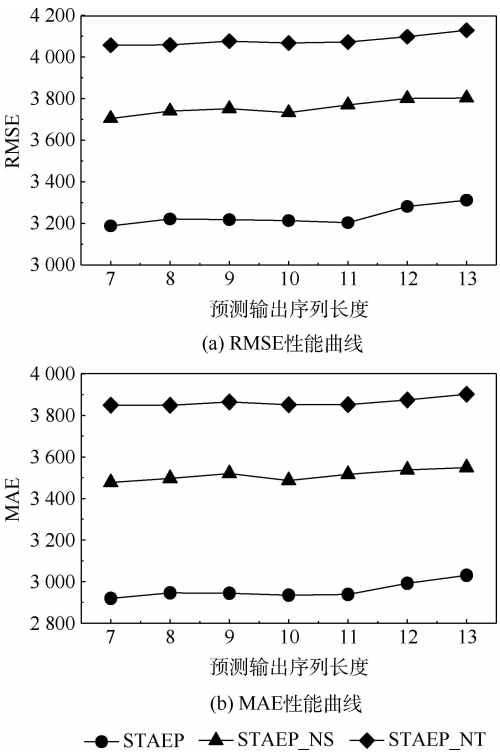



图 4 输出序列长度对模型性能的影响曲线

Fig. 4 Influence curves of output sequence length on model performance

### 3.5 案例分析

为了进一步验证本文模型在新冠肺炎疫情预测任务上的有效性,分析了模型在不同情况下预测不同地区疫情传播趋势的表现。欧洲新冠肺炎疫情主要集中在英国、法国、意大利等发达国家。部分人口基数大的国家确诊人数呈爆发式增长,同时,大量外来输入和输出病例导致国家之间疫情相互关联。通过引入时空注意力机制,从目标地区和其他相关地区疫情序列中挖掘潜在的疫情关联,比关注单一地区和静态关系嵌入的建模方式取得了更好的预测效果。在中国部分省级单位新冠肺炎疫情数据集中,湖北省确诊基数庞大,疫情发展初期恰处于春运期间,新冠肺炎病毒随着人口流动大范围迅速传播。与湖北省相邻的省份(如安徽省、河南省等)最先受到影响,这也解释了利用地区相邻关系建模疫情的空间关联性比只关注单个地区的疫情建模方式取得更好的预测效果。随着国家防疫政策的发布,中国新冠肺炎疫情得到了有效控制,逐渐趋向于平稳。此时,不同地区间疫情的关联主要体现在疫情走势的时序相似性,而不是位置上的相邻关系。因此,通过注意力机制建模地区间疫情的动态关联性比简单地嵌入地区间的相邻关系的方式更有效。

## 4 结 论

1)本文提出了一种基于时空注意力机制和自编码器网络结构的新冠肺炎疫情传播趋势预测模型,实现对不同地区未来疫情的动态发展趋势的准确预测。

2)设计的时空注意力机制挖掘了时空序列数据中动态复杂时序依赖性和空间关联性,不仅考虑了单个地区疫情序列数据的复杂时序依赖性,同时也考虑了相关地区过去的疫情对预测目标地区疫情趋势的影响。

3)在公开的新冠肺炎疫情数据集上进行了实验,验证了本文模型的有效性。在欧洲部分国家新冠肺炎疫情数据集上,预测误差指标 RMSE 和 MAE 分别降低了 22.3% 和 25.0%,在中国部分省级单位新冠肺炎疫情数据集上,RMSE 和 MAE 分别降低了 10.1% 和 10.4%。

本文工作还有进一步提升的空间,可行的提升途径包括考虑相关政府发布的防疫措施,地区人口数量、结构分布及医疗资源情况等对新冠肺炎疫情传播的影响。

## 参考文献（References）


[1] LEI S,JIANG F,SU W,et al. Clinical characteristics and outcomes of patients undergoing surgeries during the incubation period of COVID-19 infection[J]. EClinicalMedicine,2020,21:100331.

[2] MASTAKOURI A A,SCHÖLKOPF B. Causal analysis of Covid-19 spread in Germany[EB/OL]. (2020-08-03)[2021-09-01]. https://arxiv.org/abs/2007.11896v1.

[3] XIAO C,ZHOU J,HUANG J,et al. C-Watcher:A framework for early detection of high-risk neighborhoods ahead of COVID-19 outbreak[EB/OL]. (2021-03-01)[2021-09-01]. https://arxiv.org/abs/2012.12169v2.

[4] YANG Z,ZENG Z,WANG K,et al. Modified SEIR and AI prediction of the epidemics trend of COVID-19 in China under public health interventions[J]. Journal of Thoracic Disease,2020,12(3):165-174.

[5] ACHTERBERG M A,PRASSE B,MA L,et al. Comparing the accuracy of several network-based COVID-19 prediction algorithms[J]. International Journal of Forecasting,2022,38(2):489-504.

[6] GUPTA R,PANDEY G,CHAUDHARY P,et al. Machine learning models for government to predict COVID-19 outbreak[J]. Digital Government:Research and Practice,2020,1(4):26.

[7] LIU X X,FONG S. Towards a realistic model for simulating spread of infectious COVID-19 disease[C]// Proceedings of the 2020 the 4th International Conference on Big Data and Internet of Things. New York:ACM,2020:96-101.

[8] KUMAR M,GUPTA S,KUMAR K,et al. Spreading of COVID-19 in India,Italy,Japan,Spain,UK,US:A prediction using ARIMA and LSTM model[J]. Digital Government:Research and Practice,2020,1(4):24.

[9] WANG P,ZHENG X,LI J,et al. Prediction of epidemic trends in COVID-19 with logistic model and machine learning technics[J]. Chaos Solitons & Fractals,2020,139:110058.

[10] NGUYEN K L P,CHEN H W,YANG C T,et al. Implementation of a respiratory disease forecasting model using LSTM for central Taiwan[M]// KIM K J,KIM H Y. Information science and applications. Berlin:Springer,2020:441-450.

[11] RAUF H T,LALI M I U,KHAN M A,et al. Time series forecasting of COVID-19 transmission in Asia pacific countries using deep neural networks[J/OL]. Personal and Ubiquitous Computing,2021 (2021-01-10)[2021-09-01]. https://doi.org/10.1007/s00779-020-01494-0.

[12] WANG J,MCMICHAEL A J,MENG B,et al. Spatial dynamics of an epidemic of severe acute respiratory syndrome in an urban area[J]. Bulletin of the World Health Organization,2006,84(12):965-968.

[13] BISWAS K,KHALEQUE A,SEN P. Covid-19 spread:Reproduction of data and prediction using a SIR model on Euclidean network[EB/OL]. (2020-03-16)[2021-09-01]. https://arxiv.org/abs/2003.07063.

[14] GHAMIZI S,RWEMALIKA R,CORDY M,et al. Data-driven simulation and optimization for Covid-19 exit strategies[C]// Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York:ACM,2020:3434-3442.

[15] WAN H,CUI J A,YANG G J. Risk estimation and prediction of the transmission of coronavirus disease-2019 (COVID-19) in the mainland of China excluding Hubei province[J]. Infectious Diseases of Poverty,2020,9:116.

[16] ZHENG N,DU S,WANG J,et al. Predicting COVID-19 in China using hybrid AI model[J]. IEEE Transactions on Cybernetics,2020,50(7):2891-2904.

[17] 黎维,陶蔚,周星宇,等. 时空序列预测方法综述[J]. 计算机应用研究,2020,37(10):2881-2888.
LI W,TAO W,ZHOU X Y,et al. Survey of spatio-temporal sequence prediction methods[J]. Application Research of Computers,2020,37(10):2881-2888(in Chinese).

[18] SHI X,CHEN Z,WANG H,et al. Convolutional LSTM network:A machine learning approach for precipitation nowcasting[M]. Cambridge:MIT Press,2015.

[19] LIANG Y,KE S,ZHANG J,et al. GeoMAN:Multi-level attention networks for geo-sensory time series prediction[C]// Proceedings of the 27th International Joint Conference on Artificial Intelligence. New York:ACM,2018:3428-3434.

[20] YAO H,WU F,KE J,et al. Deep multi-view spatial-temporal network for taxi demand prediction[C]// Proceedings of the AAAI Conference on Artificial Intelligence,2018:2588-2595.

[21] WANG X,MA Y,WANG Y,et al. Traffic flow prediction via spatial temporal graph neural network[C]// Proceedings of the Web Conference 2020. New York:ACM,2020:1082-1092.

[22] SUTSKEVER I,VINYALS O,LE Q V. Sequence to sequence learning with neural networks[C]// Advances in Neural Information Processing Systems,2014:3104-3112.

[23] ZHAO L,SONG Y,ZHANG C,et al. T-GCN:A temporal graph convolutional network for traffic prediction[J]. IEEE Transactions on Intelligent Transportation Systems,2019,21(9):3848-3858.

# Prediction model of COVID-19 based on spatiotemporal attention mechanism

BAO Xin[1], TAN Zhiyi[1,*], BAO Bingkun[1], XU Changsheng[2]

(1. School of Communications and Information Engineering, Nanjing University of Posts and Telecommunications, Nanjing 210003, China;
2. National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100080, China)

**Abstract**: The continuous spread of the COVID-19 has brought profound impacts on human society. For the prevention and control of virus spreading, it is critical to predict the future trend of epidemic situation. Existing studies on COVID-19 spread prediction, based on classic SEIR models or naive time-series prediction models, are rarely considering the characteristics of complex regional correlation and strong time series dependence in the process of epidemic spread, which limits the performance of epidemic prediction. To this end, we propose a COVID-19 prediction model based on auto-encoder and spatiotemporal attention mechanism. The proposed model estimates the trend of COVID-19 by capturing the dynamic spatiotemporal dependence between the epidemic situation sequences of different regions. In particular, a spatial attention mechanism is implemented in the encoder section for every given region to capture the dynamic correlation between the epidemic situation time-series of the region and those of the related regions. Based on the leant correlation, an long short-term memory (LSTM) network is then applied to extract the epidemic sequential features for the given region by combining the recent epidemic situations of the region and the related regions. On the other hand, to better predict the dynamic of the future epidemic situation, temporal attention is introduced into an LSTM network-based decoder to capture the temporal dependence of the epidemic situation sequence. We evaluate the proposed model on several open datasets of COVID-19, and experimental results show that the proposed model outperforms the state-of-the-art models. The metrics of RMSE and MAE of the proposed model on the COVID-19 epidemic dataset of some European countries decreased 22.3% and 25.0%. The metrics of RMSE and MAE of the proposed model on the COVID-19 epidemic dataset of some Chinese provinces decreased 10.1% and 10.4%.

**Keywords**: prediction of COVID-19; attention network; spatiotemporal sequence prediction; long short-term memory (LSTM) network; auto-encoder

**Received**: 2021-09-07; **Accepted**: 2021-09-17; **Published online**: 2021-11-02 14:13

**URL**: kns.cnki.net/kcms/detail/11.2625.V.20211101.1605.010.html

**Foundation items**: National Key R & D Program of China (2020AAA0106200); National Natural Science Foundation of China (6193000388, 61872424); Natural Science Foundation of Jiangsu Province (BK20200037, BK20210595)

* **Corresponding author**. E-mail: tzy@njupt.edu.cn

http://bhxb.buaa.edu.cn jbuaa@buaa.edu.cn
**DOI**: 10.13700/j.bh.1001-5965.2021.0521

# 基于三维 Saab 变换的高光谱图像压缩方法

徐艾明, 黄宇星, 沈秋*
(南京大学 电子科学与工程学院, 南京 210023)

**摘 要**: 高光谱图像中存储了丰富的光谱信息,具有极大的应用价值,但现有大部分高光谱图像压缩方法难以同时兼顾图像中的空间冗余与谱间冗余,导致压缩性能受到局限。针对该问题,提出了一种基于三维修正偏置的子空间(Saab)变换的高光谱图像压缩方法。采用三维 Saab 变换对高光谱图像的分块进行空间光谱信息融合的降维操作,同时去除谱间冗余和局部空间冗余;利用高效率视频编码(HEVC)中的帧内编码模块进一步去除空间冗余和统计冗余;实现低失真、高比率的高光谱图像压缩。在多个高光谱图像数据集上的实验结果表明,所提方法在同码率下重建图像的信噪比(SNR)比采用主成分分析(PCA)降维的方法至少提高 0.62 dB,在高码率的情况下性能优于张量分解的压缩方法。同时,验证了不同降维方法对分类任务的性能影响,结果表明,所提方法更好地保留了图像中的重要特征,在低码率的情况下仍可以保持较高的分类精度。

**关 键 词**: 修正偏置的子空间(Saab)变换; 空间光谱信息融合; 高效率视频编码(HEVC); 高光谱图像; 图像压缩
**中图分类号**: V221+.3; TB553
**文献标志码**: A **文章编号**: 1001-5965(2022)08-1505-10

近年来,随着高光谱成像技术的迅速发展,高光谱图像已被广泛应用到气象遥感、地质检测、智能监控、医疗诊断等许多领域中[1-4]。不同于普通 RGB 图像,高光谱图像能够覆盖物体上百个光谱波段的信息[5-6],可以帮助研究者在不同的光谱波段观察物体。然而,在当前有限的传输信道带宽下,具有较高的空间分辨率和光谱分辨率的高光谱图像很难进行存储和传输。庞大数据量与有限带宽之间的矛盾严重限制了高光谱图像在计算机视觉、遥感等领域的应用。因此,高光谱图像的压缩已经成为国内外学者关注研究的重点。

高光谱图像的数据巨大是因为数据存在着极大的空间冗余和谱间冗余[7]。高光谱图像压缩的目的就是在一定的失真内,利用一系列降维压缩方法去除冗余。大部分高光谱图像压缩方法都是来源于二维灰度图像压缩,将高光谱图像的各波段视作独立的图像进行处理,忽略了谱间冗余。因此,传统的图像压缩方法在应用于高光谱图像压缩时,往往存在着不能很好处理谱间冗余的缺点。针对该问题,研究者们提出了许多压缩方法,大致可分为以下 7 类:基于变换的方法[8-10]、基于预测的方法[11-12]、基于矢量量化的方法[13-15]、基于压缩感知的方法[16]、基于张量分解的方法[6,17-18]、基于稀疏表示的方法[19]及基于深度学习的方法[20]。由于变换压缩方法有着高压缩性能和灵活的码率控制机制等诸多优点[21],目前大部分压缩方法是基于变换的。其中,主成分分析(principle component analysis,PCA)法和小波变换是最广泛使用的相关方法。许多方法将 PCA 用于减少光谱维数,并与传统图像编码算法相结合,用

**收稿日期**: 2021-09-06; **录用日期**: 2021-10-01; **网络出版时间**: 2021-10-29 14:28
**网络出版地址**: kns.cnki.net/kcms/detail/11.2625.V.20211028.2041.005.html
**基金项目**: 国家自然科学基金 (U1936202,62071216)
* **通信作者**. E-mail: shenqiu@nju.edu.cn

**引用格式**: *徐艾明, 黄宇星, 沈秋. 基于三维 Saab 变换的高光谱图像压缩方法[J]. 北京航空航天大学学报, 2022, 48(8): 1505-1514. XU A M, HUANG Y X, SHEN Q. Hyperspectral image compression method based on 3D Saab transform[J]. Journal of Beijing University of Aeronautics and Astronautics, 2022, 48(8): 1505-1514 (in Chinese).*

于高光谱图像压缩,取得了不错的压缩效果[9,22]。

要想进一步提升变换压缩方法的性能,使用新的变换函数常常是最为有效的。Kuo 等[23]提出的修正偏置的子空间(subspace approximation with adjusted bias,Saab)变换在 PCA 的基础上添加了直流锚向量与偏置向量,通过分析输入数据的统计分布来确定对输出空间的合理变换,在 RGB 图像分类任务中取得了优异的表现。虽然 Saab 变换最初是为了分类任务而提出的,但其在压缩任务上也存在着极大的潜力。Saab 变换中增加的直流锚向量能够有效保留数据的低频信息,以及更多包含图像轮廓和重要细节的信息。而在 PCA 变换矩阵基础上选取的锚向量矩阵可以集中图像绝大部分的能量与信息,构成最佳线性近似子空间。因此,Saab 变换在二维图像[24]和 RGB 视频[25]上都取得了不错的压缩效果。可视化实验[24]也验证了 Saab 变换的锚向量具有很强的特征表达能力。因此,Saab 变换在高光谱图像压缩上的应用具有很大的研究价值。

为了同时兼顾空间冗余和谱间冗余,本文提出了一种基于三维 Saab 变换的高光谱图像压缩方法。首先,对高光谱图像进行空间光谱信息的融合,将图像转化为向量集合。然后,利用 Saab 变换对向量集合进行降维,通过直流锚向量保留数据的低频信息,同时通过控制交流锚向量数量降低光谱维度,联合去除局部空间冗余与谱间冗余。最后,将降维后的向量集合展开为矩阵集合,并采用高效率视频编码[26](high efficiency video coding,HEVC)帧内编码对各维数据和锚向量矩阵进行无损压缩,以进一步去除空间冗余和统计冗余。图像的重建算法与压缩算法具有对称性,因此图像的重建就是压缩的反向操作。

## 1 高光谱图像压缩方法

本节详细介绍基于三维 Saab 变换的高光谱图像压缩方法,整体流程如图 1 所示,主要分为图像压缩模块和图像重建模块。2 个主要模块具有对称性,因此图像重建模块就是图像压缩模块的反操作。

图像压缩模块由三维 Saab 变换和 HEVC 无损压缩 2 部分组成。三维 Saab 变换先对高光谱图像进行空间光谱信息的融合,将图像转化为向量集合,再对向量集进行降维,联合去除局部空间冗余与谱间冗余。用 HEVC 帧内编码对降维后的数据进行无损压缩,进一步去除空间冗余和统计冗余。

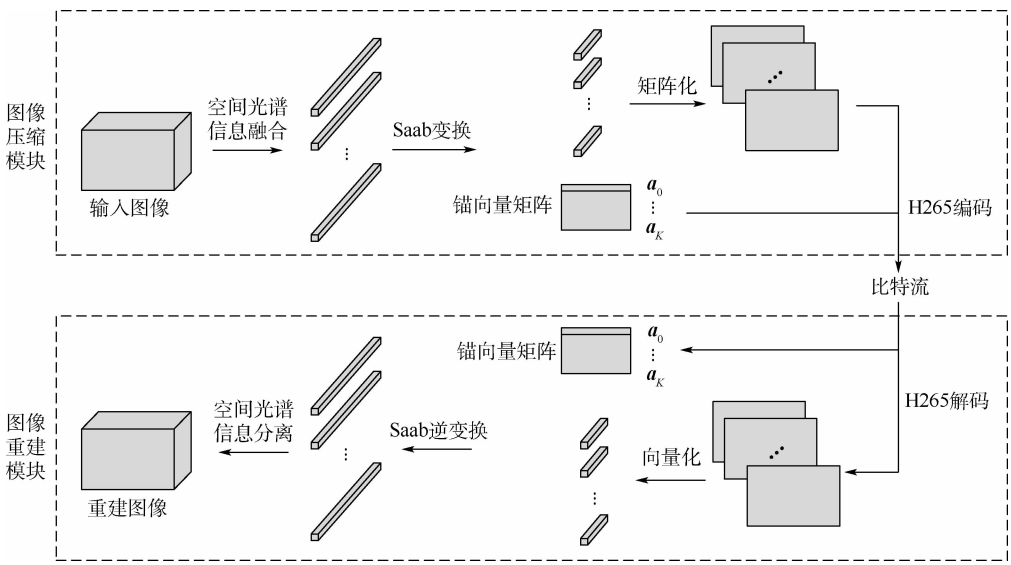



图 1 基于三维 Saab 变换的高光谱图像压缩方法流程

Fig. 1 Flowchart of hyperspectral image compression method based on 3D Saab transform

### 1.1 问题定义

基于变换的压缩方法通常是将高光谱图像从空间域变换到频域,变换后的低频变换系数包含图像轮廓和重要细节的信息,而图像的能量也集中到了低频变换系数上。这样仅保留低频变换系数就可以保存图像的绝大部分信息,实现很高的压缩比。

令数据样本 $\boldsymbol{x}_i=(x_{i1},\cdots,x_{in},\cdots,x_{iN})^{\mathrm{T}}\in\mathbf{R}^N$,输入为 $M$ 个 $N$ 维样本,则输入为矩阵 $\boldsymbol{X}=(\boldsymbol{x}_0,\boldsymbol{x}_1,\cdots,\boldsymbol{x}_{M-1})\in\mathbf{R}^{N\times M}$;变换后的样本 $\boldsymbol{y}_i=(y_{i1},\cdots,y_{ik},\cdots,y_{iK})^{\mathrm{T}}\in\mathbf{R}^K$,输出为 $M$ 个 $K$ 维样本,则输出为矩阵 $\boldsymbol{Y}=(\boldsymbol{y}_0,\boldsymbol{y}_1,\cdots,\boldsymbol{y}_{M-1})\in\mathbf{R}^{K\times M}$。变换过程

可表示为

$$y_{ik} = \sum_{n=1}^{N} a_{kn} x_{in} = \boldsymbol{a}_k^{\mathrm{T}} \boldsymbol{x}_i \tag{1a}$$

或

$$\begin{cases} \boldsymbol{y}_i = \boldsymbol{A}^{\mathrm{T}} \boldsymbol{x}_i \\ \boldsymbol{Y} = \boldsymbol{A}^{\mathrm{T}} \boldsymbol{X} \end{cases} \tag{1b}$$

式中:变换矩阵 $\boldsymbol{A} = (\boldsymbol{a}_0, \boldsymbol{a}_1, \cdots, \boldsymbol{a}_{K-1}) \in \mathbf{R}^{N \times K}$;$\boldsymbol{a}_k = (a_{k1}, \cdots, a_{kn}, \cdots, a_{kN})^{\mathrm{T}} \in \mathbf{R}^N$ 为 $\boldsymbol{A}$ 中的变换向量。

从而输出到输入的反变换可以表示为

$$x_{in} = \sum_{k=1}^{K} u_{nk} y_{ik} = \boldsymbol{u}_n^{\mathrm{T}} \boldsymbol{y}_i \tag{2a}$$

或

$$\begin{cases} \boldsymbol{x}_i = \boldsymbol{U}^{\mathrm{T}} \boldsymbol{y}_i \\ \boldsymbol{X} = \boldsymbol{U}^{\mathrm{T}} \boldsymbol{Y} \end{cases} \tag{2b}$$

式中:反变换矩阵 $\boldsymbol{U} = (\boldsymbol{u}_0, \boldsymbol{u}_1, \cdots, \boldsymbol{u}_{N-1}) \in \mathbf{R}^{K \times N}$;$\boldsymbol{u}_n = (u_{n1}, \cdots, u_{nk}, \cdots, u_{nK})^{\mathrm{T}} \in \mathbf{R}^K$ 为 $\boldsymbol{U}$ 中的变换向量。$\boldsymbol{U}$ 为 $\boldsymbol{A}$ 的逆矩阵,即 $\boldsymbol{U} = \boldsymbol{A}^{-1}$。

一般来说,变换矩阵 $\boldsymbol{A}$ 是通过数据集合 $D$ 计算得来的。给定一个变换矩阵的集合 $A$,则最优变换矩阵 $\boldsymbol{A}^*$ 的优化目标如下:

$$\boldsymbol{A}^* = \arg\max_{\boldsymbol{A} \in \mathcal{A}} M(\boldsymbol{Y}) \tag{3}$$

式中:$M(\boldsymbol{Y})$ 表示变换后输出 $\boldsymbol{Y}$ 的某项属性。能够令 $M(\boldsymbol{Y})$ 的值最大的变换矩阵就是所求的最优变换。在图像压缩中,$M(\boldsymbol{Y})$ 常常被定义为与压缩效率密切相关的属性,如能量压缩特性或去相关性能力。

PCA 的优化目标是最大化去相关性,而且 PCA 同样具有出色的能量压缩特性,其变换矩阵的计算过程如下:

1) 已知输入 $\boldsymbol{X} = (\boldsymbol{x}_0, \boldsymbol{x}_1, \cdots, \boldsymbol{x}_{M-1}) \in \mathbf{R}^{N \times M}$,即 $M$ 个 $N$ 维样本。对所有样本进行中心化,即去除样本之间的均值。中心化后的样本为

$$\overline{\boldsymbol{X}} = (\boldsymbol{x}_0 - \bar{\boldsymbol{x}}, \boldsymbol{x}_1 - \bar{\boldsymbol{x}}, \cdots, \boldsymbol{x}_{M-1} - \bar{\boldsymbol{x}}) \tag{4}$$

式中:样本之间的均值 $\bar{\boldsymbol{x}}$ 为

$$\bar{\boldsymbol{x}} = \frac{1}{M} \sum_{i=0}^{M-1} \boldsymbol{x}_i \tag{5}$$

2) 计算中心化后样本的协方差矩阵。

$$\boldsymbol{C} = \overline{\boldsymbol{X}}\, \overline{\boldsymbol{X}}^{\mathrm{T}} \in \mathbf{R}^{N \times N} \tag{6}$$

3) 对协方差矩阵 $\boldsymbol{C}$ 进行特征值分解,求出其特征值及对应的特征向量。按照特征值的大小,从大到小对特征向量进行排序,并选取前 $K$ 个特征值对应的特征向量作为变换向量 $\boldsymbol{a}_k \in \mathbf{R}^N (k = 0, 1, \cdots, K-1)$。这样能够使样本投影到输出空间的方差最大,平方误差最小,即满足最大可分性与最近重构性。变换矩阵 $\boldsymbol{A}$ 就是由这 $K$ 个特征向量构成,输出的维数 $K$ 根据压缩的需要提前指定。

假定需要压缩的高光谱图像的大小为 $H \times W \times C$,其中,$H$、$W$、$C$ 分别为高光谱图像的高度、宽度、通道数。很明显,如果用 PCA 对图像进行降维压缩,则输入必须是二维数据。因此,需要先将高光谱图像转化为二维数据的形式。常用的处理方法有以下 2 种:

1) 将高光谱图像的各波段视作独立的图像进行处理,即对 $C$ 张大小为 $H \times W$ 的二维灰度图分别进行 PCA 变换。

2) 将高光谱图像的各像素点视作独立的样本进行处理,即将高光谱图像转化为 $H \times W$ 个维度为 $C$ 的向量,再对该向量组成的矩阵进行 PCA 变换。

很明显,第 1 种方法忽略了谱间冗余,而第 2 种方法则忽略了空间冗余,两者都无法同时兼顾空间冗余和谱间冗余。

### 1.2　三维 Saab 变换

#### 1.2.1　空间光谱融合

传统的图像压缩方法往往存在着不能同时处理谱间冗余和空间冗余的缺点。针对这一问题,本文选择对高光谱图像进行空间光谱信息的融合,将图像转化为向量集合。具体过程如图 2 所示。假定输入的高光谱图像大小为 $H \times W \times C$,首先,按照空间位置将其分为 $M$ 个非重叠的图像子块,子块的大小为 $B \times B \times C$。然后,将每个子块按照通道分为 $C$ 个 $B \times B$ 的矩阵,并将矩阵展开为向量。最后,按照通道顺序依次将各向量连接,得到维度为 $D$ 的向量。这样执行下来就能将高光谱图像转化为由 $M$ 个 $D$ 维向量组成的集合。其中,$M = (H \times W)/(B \times B)$,$D = B \times B \times C$。

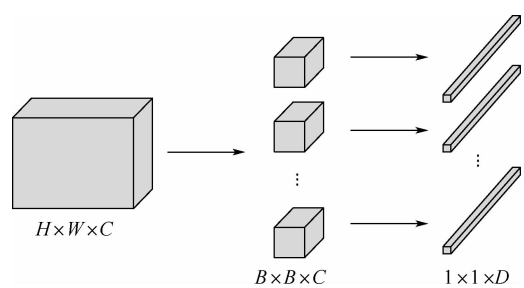



图 2　空间光谱信息融合

Fig. 2　Fusion of spatial and spectral information

#### 1.2.2　Saab 变换

Saab 变换[23]是一种基于子空间映射的变换方法,具有较强的可解释性和鲁棒性。其通过分析输入数据的统计分布来确定对输出空间的合理变换,在 PCA 的基础上添加了直流锚向量与偏置向量,进一步增强了变换的能量压缩特性和去相

关性能力。Saab 变换已经被用于手写数字识别、目标分类和 RGB 视频压缩等任务中,本文采用 Saab 变换对高光谱数据进行降维。

由 1.2.1 节可知,输入为 $M$ 个 $D$ 维样本,即输入为矩阵 $\boldsymbol{X}=(\boldsymbol{x}_0,\boldsymbol{x}_1,\cdots,\boldsymbol{x}_{M-1})\in\mathbf{R}^{D\times M}$。而本文需要将样本从 $D$ 维降至 $K$ 维,则变换后的输出为 $M$ 个 $K$ 维样本,即输出为矩阵 $\boldsymbol{Y}=(\boldsymbol{y}_0,\boldsymbol{y}_1,\cdots,\boldsymbol{y}_{M-1})\in\mathbf{R}^{K\times M}$。输入中的某个样本可以表示为向量 $\boldsymbol{x}_i=(x_{i1},\cdots,x_{in},\cdots,x_{iD})^{\mathrm{T}}\in\mathbf{R}^D$,输出中的某个样本可以表示为向量 $\boldsymbol{y}_i=(y_{i1},\cdots,y_{ik},\cdots,y_{iK})^{\mathrm{T}}\in\mathbf{R}^K$,则 Saab 变换可表示为

$$y_{ik}=\sum_{n=1}^{D}a_{kn}x_{in}+b_k=\boldsymbol{a}_k^{\mathrm{T}}\boldsymbol{x}_i+b_k \tag{7a}$$

或

$$\begin{cases}\boldsymbol{y}_i=\boldsymbol{A}_{\mathrm{Saab}}^{\mathrm{T}}\boldsymbol{x}_i+\boldsymbol{b}\\ \boldsymbol{Y}=\boldsymbol{A}_{\mathrm{Saab}}^{\mathrm{T}}\boldsymbol{X}+\boldsymbol{B}\end{cases} \tag{7b}$$

式中:变换矩阵 $\boldsymbol{A}_{\mathrm{Saab}}=(\boldsymbol{a}_0,\boldsymbol{a}_1,\cdots,\boldsymbol{a}_{K-1})\in\mathbf{R}^{D\times K}$,也称作锚向量矩阵;$\boldsymbol{a}_k=(a_{k1},\cdots,a_{kn},\cdots,a_{kD})^{\mathrm{T}}\in\mathbf{R}^D$ 为 $\boldsymbol{A}_{\mathrm{Saab}}$ 中的变换向量,也称作锚向量;$\boldsymbol{b}=(b_0,b_1,\cdots,b_{K-1})^{\mathrm{T}}\in\mathbf{R}^K$ 为偏置向量;偏置矩阵 $\boldsymbol{B}=(\boldsymbol{b},\cdots,\boldsymbol{b})\in\mathbf{R}^{K\times M}$。

在 Saab 变换中,锚向量矩阵 $\boldsymbol{A}_{\mathrm{Saab}}$ 是由直流锚向量 $\boldsymbol{a}_0$ 和交流锚向量 $\boldsymbol{a}_k(k=1,2,\cdots,K-1)$ 构成的。其中,直流锚向量为

$$\boldsymbol{a}_0=\frac{1}{\sqrt{D}}(1,1,\cdots,1)^{\mathrm{T}}\in\mathbf{R}^D \tag{8}$$

而交流锚向量 $\boldsymbol{a}_k(k=1,2,\cdots,K-1)$ 的计算过程如下:

1)对 $M$ 个样本进行中心化,即去除样本之间的均值。令 $\boldsymbol{x}_i'=\boldsymbol{x}_i-\bar{\boldsymbol{x}}$,则中心化后的样本为

$$\overline{\boldsymbol{X}}=(\boldsymbol{x}_0-\bar{\boldsymbol{x}},\boldsymbol{x}_1-\bar{\boldsymbol{x}},\cdots,\boldsymbol{x}_{M-1}-\bar{\boldsymbol{x}})=(\boldsymbol{x}_1',\boldsymbol{x}_2',\cdots,\boldsymbol{x}_{M-1}')$$

式中:样本之间的均值为 $\bar{\boldsymbol{x}}=\frac{1}{M}\sum_{i=0}^{M-1}\boldsymbol{x}_i$。

2)去除样本中的直流分量。去除直流分量后的样本为

$$\boldsymbol{z}_i=\boldsymbol{x}_i'-(\boldsymbol{a}_0^{\mathrm{T}}\boldsymbol{x}_i'+b_0)\boldsymbol{1} \tag{9}$$

式中:$\boldsymbol{1}$ 为单位恒向量;去除直流分量后的样本矩阵为 $\boldsymbol{Z}=(\boldsymbol{z}_0,\boldsymbol{z}_1,\cdots,\boldsymbol{z}_{M-1})$。

3)计算样本的协方差矩阵。

$$\boldsymbol{C}=\boldsymbol{Z}\boldsymbol{Z}^{\mathrm{T}}\in\mathbf{R}^{D\times D} \tag{10}$$

4)对协方差矩阵 $\boldsymbol{C}$ 进行特征值分解,求出其特征值及对应的特征向量。按照特征值的大小,从大到小对特征向量进行排序,并选取前 $K-1$ 个特征值对应的特征向量作为交流锚向量 $\boldsymbol{a}_k\in\mathbf{R}^D(k=1,2,\cdots,K-1)$。这样能够使样本投影到输出空间的方差最大,平方误差最小,即满足最大可分性与最近重构性。

因此,直流锚向量与这 $K-1$ 个交流锚向量构成了输入数据的最佳线性近似子空间,而该子空间也正是 Saab 变换的输出空间。此外,Saab 变换可以通过直流锚向量保留数据的低频信息,以及更多包含图像轮廓和重要细节的信息,有效提高重建质量。

偏置值 $b_k$ 的选取规则为:所有偏置项相等,并保证每个输出都是非负值,即要求满足:

$$y_{ik}=\sum_{n=1}^{D}a_{kn}x_{in}+b_k=\boldsymbol{a}_k^{\mathrm{T}}\boldsymbol{x}_i+b_k\geqslant 0 \tag{11}$$

$$b_k=b\sqrt{K} \tag{12}$$

值得注意的是,输出的维度 $K$ 是根据压缩的需要提前指定的。由于本文在 Saab 变换前对高光谱图像进行了空间光谱信息的融合,输入的维度 $D=B\times B\times C$ 是高光谱图像通道数 $C$ 的 $B\times B$ 倍。因此,如果需要高光谱图像降维后的通道数为 $K'$,则 Saab 变换中指定的维度 $K=B\times B\times K'$。

综上所述,本文中 Saab 变换的过程如下所示。

**输入**:$M$ 个 $D$ 维样本 $\boldsymbol{X}=(\boldsymbol{x}_0,\boldsymbol{x}_1,\cdots,\boldsymbol{x}_{M-1})\in\mathbf{R}^{D\times M}$。

**输出**:$M$ 个 $K$ 维样本 $\boldsymbol{Y}=(\boldsymbol{y}_0,\boldsymbol{y}_1,\cdots,\boldsymbol{y}_{M-1})\in\mathbf{R}^{K\times M}$。

1)计算直流锚向量。

$$\boldsymbol{a}_0\leftarrow\frac{1}{\sqrt{D}}(1,1,\cdots,1)^{\mathrm{T}}\in\mathbf{R}^D$$

2)对所有样本进行中心化。

$$\boldsymbol{x}_i'\leftarrow\boldsymbol{x}_i-\frac{1}{M}\sum_{i=0}^{M-1}\boldsymbol{x}_i$$

3)去除样本中的直流分量。

$$\boldsymbol{z}_i\leftarrow\boldsymbol{x}_i'-(\boldsymbol{a}_0^{\mathrm{T}}\boldsymbol{x}_i'+b_0)\boldsymbol{1}$$

4)计算样本的协方差矩阵:$\boldsymbol{C}\leftarrow\boldsymbol{Z}\boldsymbol{Z}^{\mathrm{T}}$。

5)对协方差矩阵 $\boldsymbol{C}$ 做特征值分解。

6)取最大的 $K-1$ 个特征值对应的特征向量作为交流锚向量 $\boldsymbol{a}_k\in\mathbf{R}^D(k=1,2,\cdots,K-1)$。

7)计算锚向量矩阵:$\boldsymbol{A}_{\mathrm{Saab}}\leftarrow(\boldsymbol{a}_0,\boldsymbol{a}_1,\cdots,\boldsymbol{a}_{K-1})$。

8)选取合适的偏置:$b_k\leftarrow b\sqrt{K}$,$\boldsymbol{b}\leftarrow(b_k,\cdots,b_k)^{\mathrm{T}}\in\mathbf{R}^K$,$\boldsymbol{B}\leftarrow(\boldsymbol{b},\cdots,\boldsymbol{b})\in\mathbf{R}^{K\times M}$。

9)对中心化的样本进行变换。

$$\boldsymbol{Y}\leftarrow\boldsymbol{A}_{\mathrm{Saab}}^{\mathrm{T}}\boldsymbol{X}+\boldsymbol{B}$$

### 1.3 HEVC 无损压缩

HEVC[26] 是 H.264/AVC 以后的新一代视频图像编码技术。为了进一步去除数据中存在的空

间冗余和统计冗余,本文采用 HEVC 帧内编码对 Saab 变换降维后的数据进行无损压缩。与H.264 类似,HEVC 帧内编码采用了基于块的多方向帧内预测方式,但其亮度预测方向从 H.264 的 8 种增加到了 33 种,加上平面和直流模式共计 35 种帧内预测模式,更加细致灵活[27],有效提升了对方向性结构的预测效果。此外,HEVC 帧内编码还定义了自适应的平滑滤波器,对参考像素进行适当的预滤波处理,有效减少了噪声的影响,并提高了编码的精度与效率。

已知降维后得到了 $M$ 个 $K$ 维样本,即 $\boldsymbol{Y}=(\boldsymbol{y}_0,\boldsymbol{y}_1,\cdots,\boldsymbol{y}_{M-1})\in\mathbf{R}^{K\times M}$。具体操作过程为:首先,将数据的空间光谱信息分离,即将每个 $1\times1\times K$ 维向量 $\boldsymbol{y}_i\in\mathbf{R}^K$ 转化为大小为 $B\times B\times K'$的数据块,并将这 $M$ 个数据块合并为 $H\times W\times K'$的图像,其中 $K'=K/(B\times B)$;然后,将图像分为 $K'$个大小为 $H\times W$ 的灰度图,并分别进行 HEVC 无损压缩。锚向量矩阵 $\boldsymbol{A}_{\text{Saab}}$也采用 HEVC 无损压缩。

### 1.4　图像重建模块

如图 1 所示,图像重建模块与图像压缩模块具有对称性,因此图像重建模块就是图像压缩模块的反操作。具体过程为:首先,将码流解码,得到 $H\times W\times K'$的图像和锚向量矩阵 $\boldsymbol{A}_{\text{Saab}}$;然后,对图像进行空间光谱信息融合得到 $M$ 个 $K$ 维向量构成的向量组 $\boldsymbol{Y}$,并对其进行 Saab 反变换,得到 $M$ 个 $D$ 维向量构成的向量集合 $\hat{\boldsymbol{X}}$;最后,对该向量集进行空间光谱信息分离,重建出 $H\times W\times C$ 的高光谱图像。

空间光谱信息融合和空间光谱信息分离的具体过程在 1.2.1 节和 1.3 节中已分别介绍过,此处不再赘述。Saab 反变换可表示为

$$\hat{\boldsymbol{X}}=\boldsymbol{U}_{\text{Saab}}^{\text{T}}(\boldsymbol{Y}-\boldsymbol{B}) \tag{13}$$

式中:反变换矩阵 $\boldsymbol{U}_{\text{Saab}}=(\boldsymbol{u}_0,\boldsymbol{u}_1,\cdots,\boldsymbol{u}_{D-1})\in\mathbf{R}^{K\times D}$;$\boldsymbol{u}_n=(u_{n1},\cdots,u_{nk},\cdots,u_{nK})^{\text{T}}\in\mathbf{R}^K$ 为 $\boldsymbol{U}_{\text{Saab}}$中的变换向量。

由于 Saab 变换是正交变换,变换矩阵 $\boldsymbol{A}_{\text{Saab}}$的各行之间互相正交,$\boldsymbol{U}_{\text{Saab}}=\boldsymbol{A}_{\text{Saab}}^{-1}=\boldsymbol{A}_{\text{Saab}}^{\text{T}}$。如果保留所有的特征向量作为交流锚向量,就可以完成数据 100% 的重建,但在降维过程中只保留了前 $K-1$ 个特征向量,因此只能完成数据的近似重建。

## 2　实验结果及分析

本节在多个高光谱图像数据集上设计了充足的实验,对比本文方法和现有压缩方法以验证方法的有效性。具体来说,本文在 Salinas 数据集上对比了多种不同压缩方法的压缩性能,在多个数据集上对比了降维到不同维度时本文方法与 PCA 的压缩性能,在 Salinas 数据集上对比了本文方法与 PCA 的重建图像分类性能。

### 2.1　数据集

本文采用了 3 个标准的高光谱图像数据集作为实验数据,分别为 Salinas 数据集[28]、PaviaU 数据集[28]、Botswana 数据集[28]。

Salinas 数据集是由 AVIRIS 传感器获取的美国 Salinas 山谷地带的图像场景,该数据空间包含 512×217 个像素,224 个光谱波段。本文移除 20 个水吸收波段,将剩余 204 个波段作为实验数据。图 3 展示了 Salinas 数据集的三通道伪彩色图。PaviaU 数据集是由 ROSIS 传感器获取的意大利 Pavia 大学的图像场景,该数据空间包含 610×340 个像素,115 个光谱波段。本文移除 12 个水吸收波段,将剩余 103 个波段作为实验数据。图 4 展示了 PaviaU 数据集的三通道伪彩色图。Botswana 数据集是由 Hyperion 传感器获取的南非 Botswana 三角洲区域的图像场景,该数据空间包含 1 476×256 个像素,242 个光谱波段。本文移除了水吸收和噪声波段,将剩余 145 个波段作为实验数据。图 5 展示了 Botswana 数据集的三通道伪彩色图。

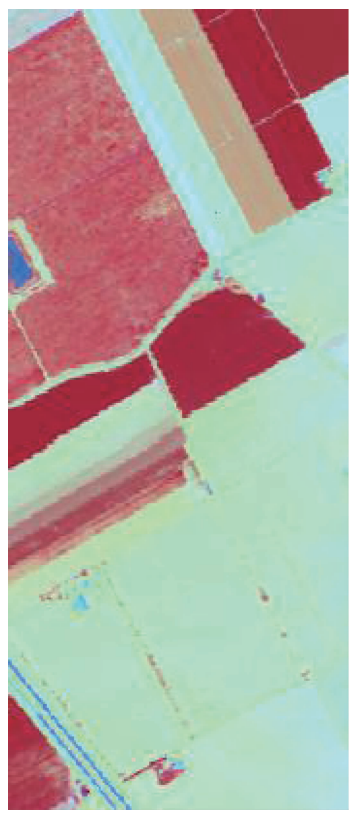


图 3　Salinas 数据集伪彩色图

Fig. 3　Pseudo RGB image of Salinas dataset

图 4 PaviaU 数据集伪彩色图

Fig. 4 Pseudo RGB image of PaviaU dataset

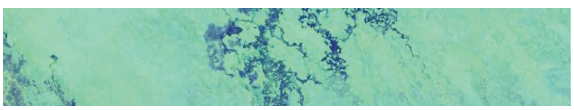


图 5 Botswana 数据集伪彩色图

Fig. 5 Pseudo RGB image of Botswana dataset

## 2.2 不同方法的压缩性能比较

为了验证本文方法能够有效提升高光谱图像的压缩性能,在 Salinas 数据集上对比了多种不同的压缩方法,分别为 CBTD[17]、PARAFAC[6]、TUCKER-ALS[18]这 3 种张量分解方法及 PCA 和 HEVC。其中,CBTD 是基于张量分解的高光谱图像压缩方法,其与 TUCKER-ALS 都是对高光谱图像进行 Tucker 分解,将图像分解为 1 个核张量和 3 个不同维度的因子矩阵,而这些因子矩阵通常被认为是不同维度上的主成分;PARAFAC 是利用 CP(CANDECAMP/PARAFAC)分解将高光谱图像分解为秩一张量之和;PCA 是将高光谱图像的各像素点视作独立数据处理的 PCA 变换,其余设置与本文方法一致;HEVC 是将高光谱图像当作视频序列进行编码,即将高光谱图像的各波段视为视频序列的不同帧,并对其进行 HEVC 视频编码。

本文提出的基于三维 Saab 变换的高光谱图像压缩方法中,空间光谱信息融合时的分块大小 $B$ 设置为 4,通过改变三维 Saab 变换降维的维度 $K$ 来控制码率与重建质量。为适应分块的大小,高光谱图像的空间尺寸应为 4 的倍数。因此,本文移除了数据集中空间边缘部分的少量像素。具体来说,本文取 Salinas 数据集空间的 512 × 216 个像素、PaviaU 数据集空间的 608 × 340 个像素、Botswana 数据集的全部像素作为实验数据。本文的 HEVC 帧内编码和 HEVC 视频编码使用 ffmpeg 开源代码,调用 ffmpeg 中命令对降维后的灰度图像和锚向量矩阵分别进行压缩。由于 HEVC 视频编码最大只支持 12 bit 的像素,编码时需要将 16 位的图像分为前 8 位和后 8 位的 2 张图像,并对前 8 位的图像采用较小的量化参数,对后 8 位的图像采用较大的量化参数,以达到最优的压缩效果。2.3 节和 2.4 节中的实验同样采用上述实验设置。

本文采用信噪比(signal to noise rate,SNR)作为压缩性能的评价指标。SNR 的值越高,表示压缩性能越好。假定原图 $\boldsymbol{X} \in \mathbf{R}^{H \times W \times C}$,重建图像为 $\hat{\boldsymbol{X}} \in \mathbf{R}^{H \times W \times C}$,则 SNR 的计算公式为

$$\mathrm{SNR} = 10\lg\left(\frac{\|\boldsymbol{X}\|^2}{\|\boldsymbol{X} - \hat{\boldsymbol{X}}\|^2}\right) \tag{14}$$

对于高光谱图像,比特率(bit rate,BR)采用的单位是 bpppb(bit per pixel per band,比特数每像素每波段)。

不同压缩方法在 Salinas 数据集上的压缩性能如图 6 所示。很明显,本文方法在同码率下重建图像的 SNR 值比采用 PCA 降维的方法至少提高了 0.62 dB,有力证明了三维 Saab 变换可以有效提高高光谱图像的重建质量。此外,本文方法在同码率下取得了高出 HEVC 较多的 SNR 值,充分证明了本文方法中三维 Saab 变换对整体压缩效果具有主要贡献。当然,HEVC 表现不佳的原因主要在于:高光谱图像不同波段之间并没有空间位置的变化,只存在光谱信息的变化,而 HEVC 的帧间预测主要针对于空间信息的改变,并不能很好地处理光谱信息的变化。

在比特率较低时,本文方法的 SNR 值优于除 CBTD 方法以外的其他对比方法;在比特率较高时,

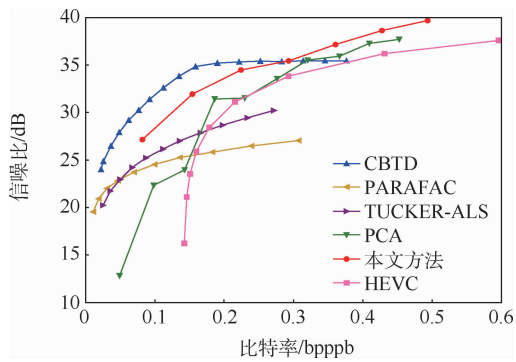



图 6 不同方法在 Salinas 数据集上的压缩性能

Fig. 6 Compression performance of different methods on Salinas dataset

本文方法的 SNR 值优于所有对比方法。CBTD 方法认为高光谱图像的谱间相关性很强，只需在光谱维度上保留较小的因子矩阵，便可以很好地还原光谱信息，从而减小比特率。因此，在重建精度较低时，CBTD 方法能够取得比三维 Saab 变换更低的比特率；但随着比特率的增大，三维 Saab 变换能够更好地保留光谱信息，而 CBTD 方法过于忽视谱间信息，因此三维 Saab 变换在高比特率时能够取得更好的压缩性能。

### 2.3　本文方法与 PCA 各维度压缩性能比较

三维 Saab 变换对高光谱图像进行了空间光谱信息的融合，同时在 PCA 的基础上添加了直流锚向量与偏置向量，通过分析输入数据的统计分布来确定对输出空间的合理变换。为了更加深入地探究三维 Saab 变换相比于 PCA 的优越性，本文在 Salinas 数据集、PaviaU 数据集、Botswana 数据集上对比了降维到不同维度时本文方法与 PCA 的压缩性能。

实验结果如图 7 所示。在 3 个数据集上，降维到较低维度的情况下，本文方法的 SNR 值远高于 PCA。这有力证明了三维 Saab 变换中的空间光谱信息融合操作能够同时兼顾空间冗余和谱间冗余，有效保留了空间信息和光谱信息。此外，这也表明三维 Saab 变换通过直流锚向量能够保留数据的低频信息，以及更多包含图像轮廓和重要细节的信息，有利于提高重建质量。而在降维到相同维度的情况下，本文方法的 SNR 值也都高于 PCA。这进一步证明了直流锚向量的重要作用，也表明在 PCA 变换矩阵基础上选取的锚向量矩阵可以集中图像绝大部分的能量与信息，构成最佳线性近似子空间，有助于提升压缩性能。

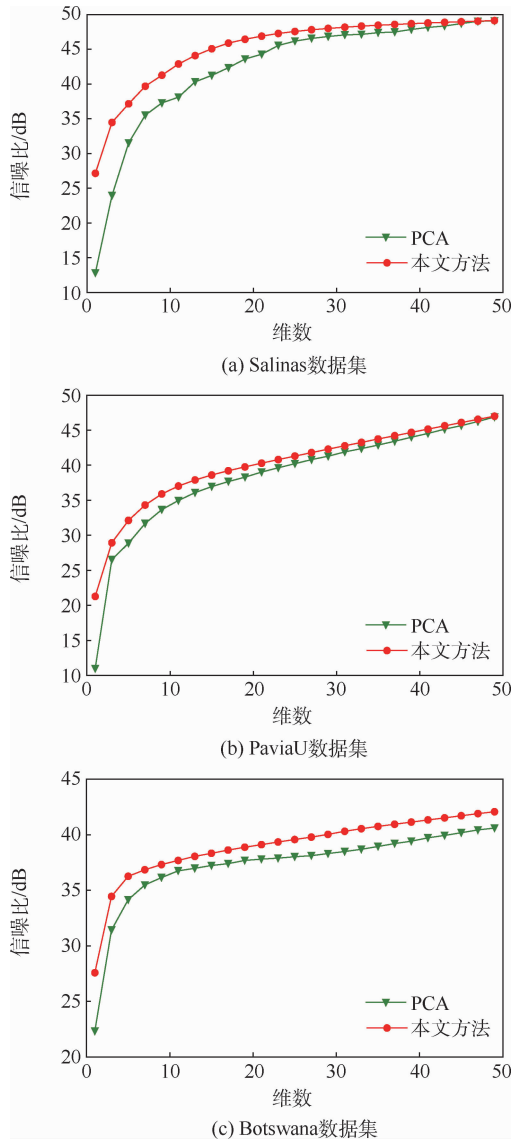



图 7　本文方法和 PCA 在不同数据集上的压缩性能

Fig. 7　Compression performance of the proposed method and PCA on different datasets

### 2.4　重建图像的分类性能

高光谱图像中存储了丰富的光谱信息，在分类、分割、目标检测等诸多视觉任务中都有着重要应用。然而，许多高光谱图像压缩方法往往都注重于在压缩信号中保留更多有利于重建的信息，而非保留更多有利于其他视觉任务的重要特征信息，如分类任务中的判别特征信息等。例如，PCA 虽然可以取得不错的压缩性能，但不一定能捕捉到图像中所有的判别信息。这是因为对于分类任务很重要的判别特征不一定拥有很高的信号能量。因此，重建图像在视觉任务中的表现可以一定程度上反映出压缩方法的信息保存能力，且具有重要的参考价值。

许多视觉任务的核心目标是实现更好的像素级分类。为了更加深入地探究三维 Saab 变换相比 PCA 的优越性，本文利用 SGD（基于随机梯度下降的线性支持向量机）和 3D-CNN[29] 这 2 种高光谱图像分类方法，比较了本文方法与 PCA 在 Salinas 数据集重建图像上的分类性能。本文随机划分了数据集的 30% 作为训练集，70% 作为测试集。2 种分类器都是在无损的训练集上训练到收敛，再在重建图像上测试分类性能。分类指标采用总体分类精度。

实验结果如图 8 所示。在 Salinas 数据集上，降维压缩到相同维度的情况下，与 PCA 相比，三维 Saab 变换的重建图像拥有更高的分类准确率。这表明三维 Saab 变换中的空间光谱信息融合操作能够有效保留空间特征和光谱特征，有利于分类等视觉任务。值得注意的是，三维 Saab 变换的重建图像在低维的情况下仍然可以保持较高的分类精度，有力证明了空间光谱信息融合和直流锚

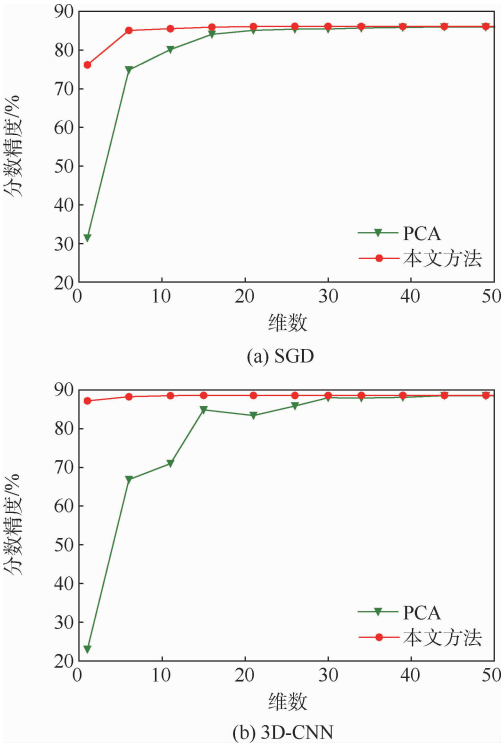



图 8 在 Salinas 数据集重建图像上的分类精度

Fig. 8 Classification accuracy of reconstructed image on Salinas dataset

向量的特征表达能力与信息保留能力。此外,在 3D-CNN 分类器上,PCA 的重建图像在 21 维时有小幅的分类精度下降。这一定程度上证明了 PCA 并不能捕捉到图像中所有的判别信息,其保留的图像信息并不全部有助于分类任务。

## 3 结 论

1) 针对高光谱图像存在大量冗余信息,且 PCA 等传统方法无法兼顾空间冗余与谱间冗余的问题,本文提出一种基于三维 Saab 变换的高光谱图像压缩方法。

2) 采用三维 Saab 变换对高光谱图像的分块进行空间光谱信息融合的降维操作,联合去除局部空间冗余与谱间冗余,使用 HEVC 帧内编码模块对降维后的数据进行无损压缩,进一步去除空间冗余和统计冗余。此外,三维 Saab 变换的直流锚向量可以有效保留数据的低频信息,与交流锚向量一起构成最佳线性近似子空间,有助于提升重建质量。

3) 实验结果表明,本文方法能够有效处理高光谱图像中的各种冗余信息,保留包含图像轮廓和重要细节的特征信息,有效提升了压缩性能。此外,本文方法能够更好地保留图像中的重要特征,在低码率的情况下仍然可以保持较高的分类精度。

## 参考文献 (References)

[1] 马晨光,曹汛,季向阳,等. 高分辨率光谱视频采集研究[J]. 电子学报,2015,43(4):783-790.
MA C G,CAO X, JI X Y,et al. Research on high resolution hyperspectral capture technique[J]. Acta Electronica Sinica, 2015,43(4):783-790(in Chinese).

[2] LEITNER R, BIASIO M D, ARNOLD T, et al. Multi-spectral video endoscopy system for the detection of cancerous tissue [J]. Pattern Recognition Letters,2013,34(1):85-93.

[3] CHO W,JANG J,KOSCHAN A,et al. Hyperspectral face recognition using improved inter-channel alignment based on qualitative prediction models[J]. Optics Express,2016,24(24): 27637-27662.

[4] SANTARA A,MANI K,HATWAR P,et al. BASS Net: Band-adaptive spectral-spatial feature learning neural network for hyperspectral image classification[J]. IEEE Transactions on Geoscience and Remote Sensing,2017,55(9):5293-5301.

[5] LANDGREBE D. Hyperspectral image data analysis[J]. IEEE Signal Processing Magazine,2002,19(1):17-28.

[6] FANG L Y,HE N J,LIN H. CP tensor-based compression of hyperspectral images[J]. Journal of the Opitical Society of America A:Optics,Image Science,and Vision,2017,34(2): 252-258.

[7] 王成. 高光谱图像压缩的方法研究[D]. 南京:南京理工大学,2014:5-6.
WANG C. Researches of hyperspectral image compression methods[D]. Nanjing:Nanjing University of Science and Technology,2014:5-6(in Chinese).

[8] ZHANG J, LIU G Z. A novel lossless compression for hyperspectral image by context-based adaptive classified arithmetic coding in wavelet domain[J]. IEEE Geoscience and Remote Sensing Letters,2007,4(3):461-465.

[9] DU Q, FOWLER J E. Hyperspectral image compression using JPEG2000 and principal component analysis[J]. IEEE Geoscience and Remote Sensing Letters,2007,4(2):201-205.

[10] WANG X H,TAO J Z,SHEN Y T,et al. Distributed source coding of hyperspectral images based on three-dimensional wavelet[J]. Journal of the Indian Society of Remote Sensing, 2018,46(4):667-673.

[11] SHINDE T S,TIWARI A K,LIN W Y. Low-complexity adaptive switched prediction-based lossless compression of time-lapse hyperspectral image data[C]//2019 IEEE Global Conference on Signal and Information Processing (GlobalSIP). Piscataway: IEEE Press,2019:1-5.

[12] CANG S, WANG A. Research on hyperspectral image reconstruction based on GISMT compressed sensing and interspectral prediction[J]. International Journal of Optics,2020, 2020(12): 1-11.

[13] RYAN M J,ARNOLD J F. The lossless compression of AVIRIS images by vector quantization[J]. IEEE Transations on Geoscience and Remote Sensing,1997,35(3):546-550.

[14] RYAN M J,PICKERING M R. An improved M-NVQ algorithm for the compression of hyperspectral data[C]//IEEE 2000 In-

ternational Geoscience and Remote Sensing Symposium. Piscataway:IEEE Press,2000,2:600-602.

[15] MOTTA G,RIZZO F,STORER J A. Partitioned vector quantization:Application to lossless compression of hyperspectral images[C]//2003 International Conference on Multimedia and Expo. Piscataway:IEEE Press,2003:111-553.

[16] 宋娟. 基于分布式信源编码的多光谱图像/视频压缩技术研究[D]. 西安:西安电子科技大学,2012.
SONG J. Researches on comression of multispectral images/video based on distributed source coding[D]. Xi'an:Xidian University,2012(in Chinese).

[17] LI R,PAN Z B,WANG Y,et al. The correlation-based tucker decomposition for hyperspectral image compression[J]. Neurocomputing,2021,419:357-370.

[18] ZHANG L F,ZHANG L P,TAO D C,et al. Compression of hyperspectral remote sensing images by tensor approach[J]. Neurocomputing,2015,147:358-363.

[19] FU W,LI S T,FANG L Y,et al. Adaptive spectral-spatial compression of hyperspectral image with sparse representation[J]. IEEE Transactions on Geoscience and Remote Sensing,2017,55(2):671-682.

[20] DUA Y M,SINGH R S,PARWANI K,et al. Convolution neural network based lossy compression of hyperspectral images[J]. Signal Processing:Image Communication,2021,95:116-255.

[21] DUA Y M,KUMAR V,SINGH R S. Comprehensive review of hyperspectral image compression algorithms[J]. Optical Engineering,2020,59(9):090902.

[22] DU Q,LY N,FOWLER J E. An operational approach to PCA + JPEG2000 compression of hyperspectral imagery[J]. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing,2013,7(6):2237-2245.

[23] KUO C C J,ZHANG M,LI S Y,et al. Interpretable convolutional neural networks via feedforward design[J]. Journal of Visual Communication and Image Representation,2019,60:346-359.

[24] LI N,ZHANG Y F,ZHANG Y,et al. On energy compaction of 2D Saab image transforms[C]//2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC). Piscataway:IEEE Press,2019:466-475.

[25] LI N,ZHANG Y,KUO C C J. Explainable machine learning based transform coding for high efficiency intra prediction[EB/OL]. (2020-11-21)[2021-09-01]. https://arxiv.org/abs/2012.11152.

[26] SZEV S,BUDAGAVI M,SULLIVAN G J. High efficiency video coding (HEVC):Algorithms and architectures[M]. Berlin:Springer,2014:91-112.

[27] LAINEMA J,BOSSEN F,HAN W J,et al. Intra coding of the HEVC standard[J]. IEEE Transactions on Circuits and Systems for Video Technology,2012,22(12):1792-1801.

[28] GRANA M,VEGANZONS M A,AYERDI B. Hyperspectral remote sensing scenes[EB/OL]. [2021-09-01]. http://www.ehu.eus/ccwintco/index.php/Hyperspectral_Remote_Sensing_Scenes.

[29] CHEN Y S,JIANG H L,LI C Y,et al. Deep feature extraction and classification of hyperspectral images based on convolutional neural networks[J]. IEEE Transactions on Geoscience and Remote Sensing,2016,54(10):6232-6251.

# Hyperspectral image compression method based on 3D Saab transform

XU Aiming, HUANG Yuxing, SHEN Qiu*

(School of Electronic Science and Engineering, Nanjing University, Nanjing 210023, China)

**Abstract**: Hyperspectral images contain rich and valuable spectral information, which brings great challenges to storage and transmission. However, most current hyperspectral image compression methods cannot consider spatial and spectral redundancy simultaneously, resulting in limited compression performance. We present a hyperspectral image compression method based on 3D subspace approximation with adjusted bias (Saab) transform. 3D Saab transform is firstly applied to hyperspectral image blocks, which performs spatial-spectral fusion and dimensionality reduction on blocks to remove spectral redundancy and local spatial redundancy simultaneously. Then, we use intra mode of high efficiency video coding (HEVC) to further remove spatial and statistical redundancy. Experimental results demonstrate that the proposed method can improve the signal-to-noise ratio (SNR) by at least 0.62 dB as compared with principle component analysis (PCA) based algorithm. At a high bit rate, the proposed method outperforms the state-of-art tensor decomposition compression method. We also evaluate the impact of different dimensionality reduction methods on classification, which demonstrates that the proposed method can better retain important features, with improved classification accuracy at a low bit rate.

**Keywords**: subspace approximation with adjusted bias (Saab) transform; fusion of spatial spectral information; high efficiency video coding (HEVC); hyperspectral image; image compression

**Received**: 2021-09-06; **Accepted**: 2021-10-01; **Published online**: 2021-10-29 14:28

**URL**: kns.cnki.net/kcms/detail/11.2625.V.20211028.2041.005.html

**Foundation items**: National Natural Science Foundation of China (U1936202, 62071216)

* **Corresponding author**. E-mail: shenqiu@nju.edu.cn

http://bhxb.buaa.edu.cn jbuaa@buaa.edu.cn
**DOI**: 10.13700/j.bh.1001-5965.2021.0527

# 真实场景水下语义分割方法及数据集

马志伟[1], 李豪杰[1], 樊鑫[1], 罗钟铉[1], 李建军[2], 王智慧[1,*]
(1. 大连理工大学 国际信息与软件学院, 大连 116621; 2. 杭州电子科技大学 计算机学院, 杭州 310018)

**摘 要**: 随着水下生物抓取技术的不断发展,高精度的水下物体识别与分割成为了挑战。已有的水下目标检测技术仅能给出物体的大体位置,无法提供物体轮廓等更加细致的信息,严重影响了抓取效率。为了解决这一问题,标注并建立了真实场景水下语义分割数据集DUT-USEG,该数据集包含6 617张图像,其中1 487张具有语义分割和实例分割标注,剩余5 130张图像具有目标检测框标注。基于该数据集,提出了一个关注边界的半监督水下语义分割网络(US-Net),该网络通过设计伪标签生成器和边界检测子网络,实现了对水下物体与背景之间边界的精细学习,提升了边界区域的分割效果。实验表明:所提方法在DUT-USEG数据集的海参、海胆和海星3个类别上相较于对比方法提升了6.7%,达到了目前最好的分割精度。

**关 键 词**: 水下生物抓取; 语义分割; 半监督学习; 弱监督学习; 边界检测
**中图分类号**: TP37; TP242.6
**文献标志码**: A **文章编号**: 1001-5965(2022)08-1515-10

近年来,随着水下机器人目标抓捕需求的不断上升,诸多研究开始关注水下机器视觉领域。目前的水下视觉研究主要包括水下图像增强[1-3]和水下目标检测[4-6]。水下图像增强的目的是提高水下图像的质量,水下目标检测的目的是为水下机器人的抓捕提供物体的识别与定位。然而,目标检测只能够为物体提供一个矩形包围框,无法给出物体轮廓等更加细致的信息,尤其是当水下环境和物体本身难以区分时,即使有目标检测的结果,仍然很难完成水下目标的精准抓取,而计算机视觉中的语义分割任务则能够很好地完成物体与背景之间的区分。因此,本文专注于水下机器视觉的一个新研究问题,即水下图像语义分割。

为了进行该项研究,本文提出了真实场景水下语义分割数据集DUT-USEG,该数据集包括6 617张水下图像,包含了海参、海胆、扇贝和海星4个类别,其中1 487张图像具有本文手工添加的语义分割标注和实例分割标注,剩余的5 130张图像具有目标检测框标注(这些目标检测标注是由Liu等[7]完成的)。DUT-USEG数据集已经发布在https://github.com/baxiyi/DUT-USEG。

基于该数据集,本文进一步进行了水下图像语义分割的研究,提出了一个包括伪标签生成器和边界检测子网络的半监督水下语义分割网络(underwater segmentation network, US-Net)。针对水下图像中物体与水下环境难以区分的问题,本文设计了一个边界检测网络,该网络通过融合多个不同尺度的特征图来完成类边界的学习。此外,为了有效利用数据集中无语义分割标注的数据,本文设计了一个伪标签生成器,通过已有的分类标注和框标注分别得到类激活图和框注意力图,将二者融合后通过阈值筛选得到语义分割的伪标签,利用生成的伪标签和已有的标注数据共同监督边界检测网络的学习,伪标签的生成一定程度上解决了监督信息稀少的问题。实验表明,本文方法在DUT-USEG数据集的精度优于已有

**收稿日期**: 2021-09-06; **录用日期**: 2021-10-01; **网络出版时间**: 2022-01-28 11:58
**网络出版地址**: kns.cnki.net/kcms/detail/11.2625.V.20220127.1107.001.html
**基金项目**: 国家自然科学基金(61976038,61932020,61772108,U1908210)
* **通信作者**. E-mail: zhwang@dlut.edu.cn

**引用格式**: *马志伟, 李豪杰, 樊鑫, 等. 真实场景水下语义分割方法及数据集[J]. 北京航空航天大学学报, 2022, 48(8): 1515-1524. MA Z W, LI H J, FAN X, et al. A real scene underwater semantic segmentation method and related dataset[J]. Journal of Beijing University of Aeronautics and Astronautics, 2022, 48(8): 1515-1524 (in Chinese).*

方法。

本文的主要工作包括：

1）提出了一个真实场景水下语义分割数据集 DUT-USEG，本文这是第一个开源的真实场景水下图像语义分割数据集。

2）针对水下语义分割问题，提出了一个半监督的语义分割网络（US-Net），该方法通过设计一个伪标签生成器和边界检测子网络以解决水下物体与背景之间边界难以区分的问题。

3）实验表明，本文提出的 US-Net 网络在 DUT-USEG 数据集上达到了目前最好的语义分割结果。

## 1 相关工作

### 1.1 水下目标检测

水下目标的抓捕需要先对水下生物进行定位，因此，目前有很多工作对水下目标检测进行研究。Li 等[5]使用 Fast RCNN 网络进行鱼类的检测和识别。Villon 等[6]将深度学习方法与定向网格直方图（HOG）和支持向量机（support vector machine，SVM）方法在珊瑚礁鱼类检测中进行了比较，证明了深度学习方法在水下目标检测中的优越性。Chen 等[4]提出了一个利用多个高级特征图以改进小物体目标检测的网络，并设计了样本加权损失函数来监督网络的学习。文献[7-8]针对水下目标检测数据缺少的问题，提出了相关的数据集。与本文最相关的是 Liu 等[7]提出的数据集，本文数据集是在该数据集的基础上制作的。

与以上水下目标检测工作不同，本文专注于水下图像的语义分割，认为语义分割得到的像素级分割结果对于完成更精确的水下目标抓取能够提供帮助。

### 1.2 基于框的弱监督语义分割

基于全卷积神经网络（fully convolutional network，FCN）[9]的语义分割方法已经发展成熟，但是由于语义分割的标注成本较高，目前有很多工作开始研究弱监督语义分割，与本文较为相关的方法是基于框进行弱监督语义分割。He 等[10]提出了一种以框为监督，在自动生成区域建议和训练卷积网络之间进行迭代的方式来逐步改善分割掩码的方法。Khoreva 等[11]指出，通过 GrabCut[12]或 MCG[13]等传统方法先得到伪标签，经过单次训练就可以得到很好的效果。Song[14]和 Lee[15]等通过不同的方式对伪标签质量进行改善，从而达到了更好的分割结果。Zhang 等[16]使用高斯注意力图和引导梯度反向传播图作为额外输入，为准确分割提供先验线索。

与以上方法不同，本文方法主要关注于水下图像的语义分割。由于水下图像环境的复杂性及水下能见度较低等问题，很多水下物体与周围环境之间的区分度很低，导致物体的边界很难与背景区分开。同时，由于水下物体之间经常出现彼此挨得很近的情况，物体与物体之间的边界也很难分割，图 1 展示了这 2 个问题的样例图像。图 1中，第 1 行的 2 张图像展示了物体与背景之间的边界难以区分的问题；第 2 行 2 张图像展示了物体之间边界难以区分的问题。因此，将现有方法直接应用于水下图像的语义分割会出现很多问题。GrabCut 等传统算法容易将框内的整个区域都当做前景区域，而基于深度学习的方法则经常分割不出目标物体，或只分割出目标物体的一小部分。为了解决之前方法存在的问题，本文设计了一个边界检测子网络来检测类之间的边界，从而达到更好区分前景与背景的目的。

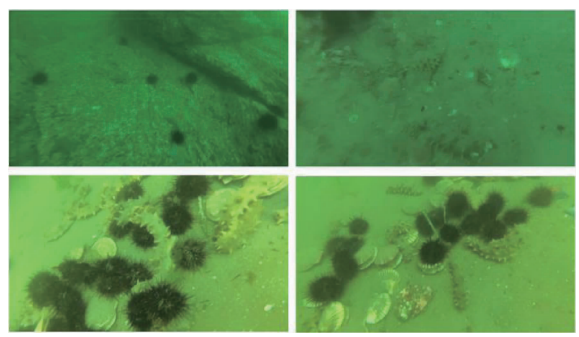


图 1 边界难以分割的样例图像

Fig. 1 Sample image with indivisible border

## 2 数据集

### 2.1 数据集收集与标注

由于目前还没有开源的水下图像语义分割数据集，为了完成水下语义分割的研究，本文以之前的目标检测水下数据集为基础构建了一个水下图像分割数据集，并对其中的 1 487 张图像进行了语义分割和实例分割的标注，剩余的 5 130 张图像保留了原数据集的目标检测标注。本文的标注是通过 LabelMe 标注软件完成的，具体做法是：使用多边形类型的标注方式将物体的周围使用点构成封闭多边形，再将多边形内部区域生成对应物体的掩码，流程如图 2 所示。由于本文是以实例为单位进行标注的，在完成语义分割标注的同时，也完成了实例分割的标注。

本文的标注包括海参、海胆、扇贝和海星 4 个类别，使用的是 COCO 数据集[17]的标注格式，为

了方便语义分割任务的研究,本文像 PASCAL VOC 2012[18]一样提供了语义分割的掩码图像。图 3 展示了语义分割标注的一些样例。可以看到,本文数据集包含了多种不同的水下场景。

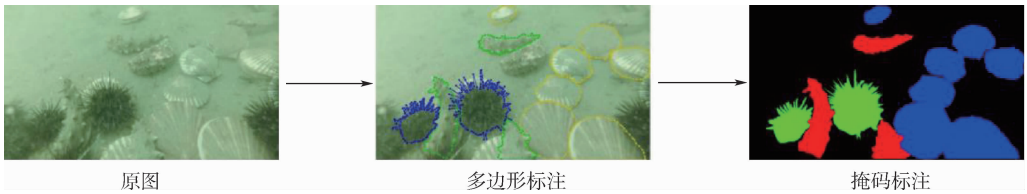



图 2　标注过程

Fig. 2　Process of labeling

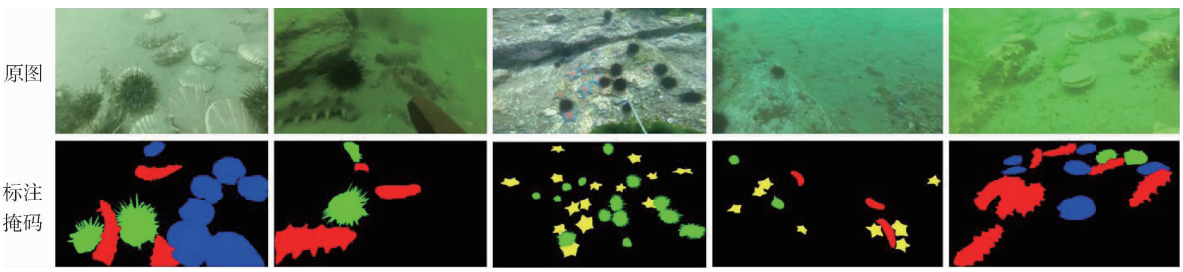



图 3　语义分割的标注样例

Fig. 3　Examples of annotations for semantic segmentation

## 2.2　数据集的相关统计

由于在水下机器人实际抓捕过程中,使用的摄像头型号往往并不固定,采集到图像的分辨率也会呈现较大差异,如在 URPC 2018 的比赛数据集中包含了 4 种不同分辨率的图像。因此,为使本文数据集能够更真实地模拟水下抓捕的实际情况,本文也在数据集中包含了多种不同分辨率的图像,其中最低分辨率为 586 × 480,最高分辨率达到了 3 840 × 2 160。图 4 展示了本文数据集中不同分辨率图像的数量统计情况。可以看到,1 000以上高分辨率的图像占到了图像总数的一半左右,由于在训练过程中进行下采样造成的细节信息损失问题,对高分辨率图像进行分割也是语义分割问题中的一个难题[19]。

本文数据集中包含了 4 类水下生物,分别为海参、海胆、扇贝和海星,图 5 展示了 4 个类别对应实例所占的数量。可以看到,本文数据集存在一定的类别不均衡问题,海胆的实例数目很多,而扇贝的实例数目较少,这是由海洋复杂的生态环境决定的。由于海洋生物本身在一定区域内就存在不同类别数目不均衡的情况,当为了获取如海参、扇贝这类数量较少的生物图像时,不可避免地会将海胆等数量较多的生物也拍摄其中,图 6 展示了这种情况的一些样例图像。

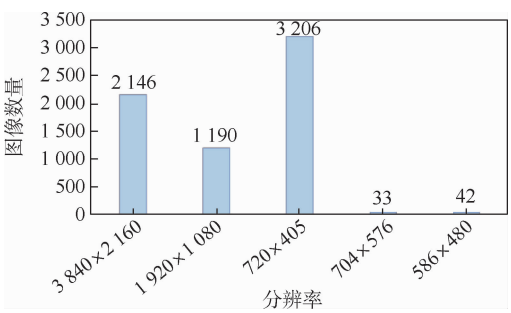



图 4　本文数据集中不同分辨率图像数量统计

Fig. 4　Statistics on number of images with different resolutions in the proposed dataset

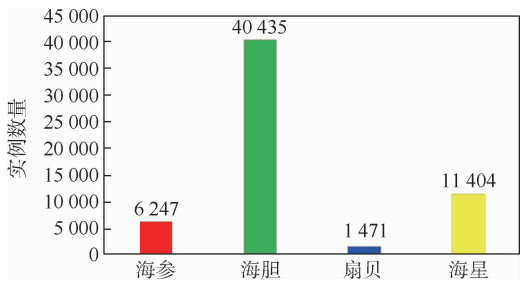



图 5　本文数据集中不同类别对应的实例数量统计

Fig. 5　Statistics on number of instances of different categories in the proposed dataset

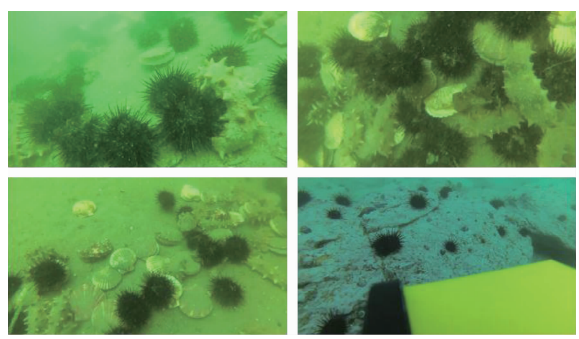


图 6　海洋生物不同类别数目不均衡问题的样例图像

Fig. 6　Sampleimages showing the unbalanced numbers of different categories marine organisms

## 3 方 法

本文提出的 US-Net 整体网络框架如图 7 所示,该网络包含一个伪标签生成器和一个边界检测网络,将分别在 3.1 节和 3.2 节进行详细介绍。

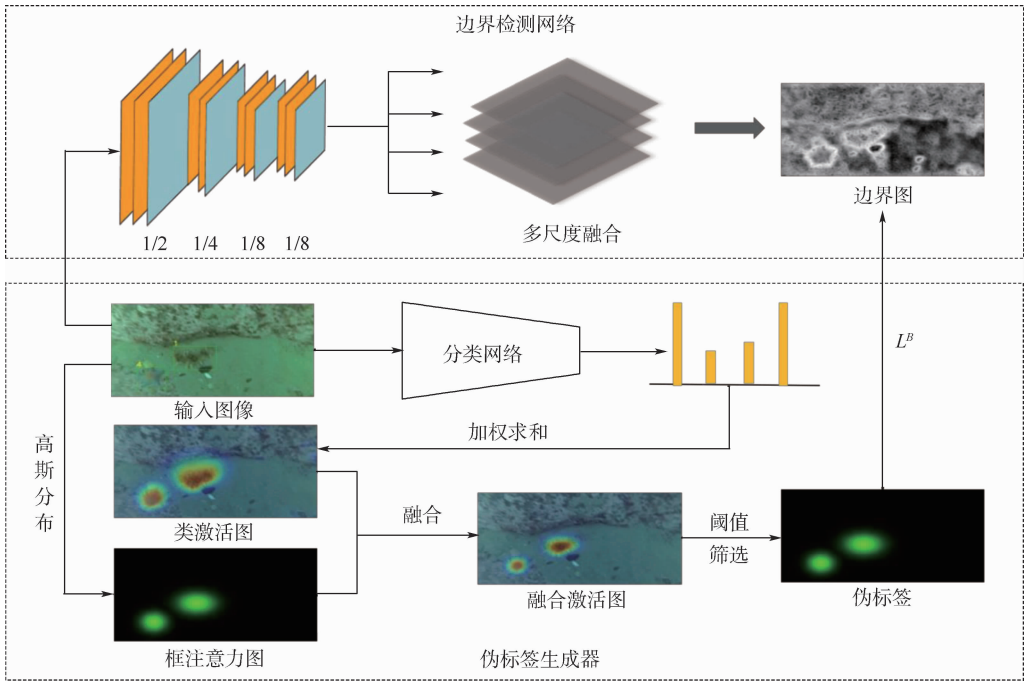



图 7 US-Net 网络框架

Fig. 7 Framework of US-Net

### 3.1 伪标签生成器

由于在训练过程中仅有部分数据为语义分割标注数据,对于没有语义分割标注的数据,通过分类和目标检测框的标注信息为其生成伪标签。伪标签通过融合类激活图和框注意力图生成伪标签,并通过 $L^B$ 损失函数来监督边界网络的学习,如图 7 上部分所示。

#### 3.1.1 类激活图

为了结合分类网络的信息,本文使用一种类激活图的方法[20]。该方法先将分类网络最后一层卷积输出的特征图进行全局平均池化,再通过一个全连接层得到各个类别对原特征图的权重,使用该权重对原特征图进行加权求和从而得到原特征图中的各个类的判别性区域。

假设分类网络的最后一层卷积得到的特征图大小为 $H \times W \times K$,令 $f_k(x,y)$ 表示在特征图第 $k$ 层的 $(x,y)$ 位置的激活值。将该特征图进行全局平均池化后得到 $F_k = \frac{1}{H \times W}\sum_{(x,y)} f_k(x,y)$。假设在全连接层中 $F_k$ 对应的权重为 $w_k^c$,$M_c$ 表示在类别 $c$ 上的类激活图,则

$$M_c(x,y) = \sum_k w_k^c f_k(x,y) \tag{1}$$

该方法得到的效果如图 7 中的类激活图所示。可以看到,类别的高激活度区域在一定程度上反映了该类别物体的形状和位置信息,这对于生成可靠的语义分割伪标签是很有帮助的。

#### 3.1.2 基于框的高斯注意力

对于框标注信息的利用,本文借鉴了 Zhang 等[16]的做法,使用一个框中心为均值,框的长和宽为标准差的二维高斯分布作为高斯注意力图来表示框包含的位置信息。

具体来说,对于一个框 $B^i$:$[x^i, y^i, w^i, h^i]$,其中,$(x^i, y^i)$ 为框中心的坐标,$w^i$、$h^i$ 分别为框的宽、高,那么框 $B^i$ 的高斯注意力图可以表示为

$$G(x,y) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} e^{-\frac{1}{2(1-\rho^2)}F(x,y)} \tag{2}$$

$$F(x,y) = \frac{(x-\mu_1)^2}{\sigma_1^2} - 2\rho\frac{(x-\mu_1)(y-\mu_2)}{2\sigma_1\sigma_2} + \frac{(y-\mu_2)^2}{\sigma_2^2} \tag{3}$$

式中:$\mu_1 = x^i$,$\mu_2 = y^i$;$\sigma_1 = w^i$,$\sigma_2 = h^i$;$\rho$ 为相关系数。

#### 3.1.3 伪标签生成

为了得到更加可靠的伪标签,先对类激活图和框注意力图进行融合,在融合过程中,考虑到分类网络的性能上限问题,会有一些图像没有被正确分类。对于分类错误的情况,直接使用该图像

的真实类别对应的框生成高斯注意力图,用来替代在该类别上的激活图。对于正确的分类,仍然需要考虑 2 种情况:①类别对应的框内在类激活区域(使用区域内的激活值均大于某个阈值作为判断),这种情况下直接使用类激活图和框注意力图的点对点乘积作为结果,可以使类激活图中超出物体本身的激活区域(在框的边缘及超出框的部分)得到有效抑制;②当框内不存在类激活区域时,仍然直接使用框注意力图作为结果。

具体来说,对于一张图像 $I$ 及其对应的一个框 $B^j$:$[x^j,y^j,w^j,h^j]$,$(x^j,y^j)$ 为框左上角的坐标,$w^j$、$h^j$ 分别为框的宽、高。假设 $B^j$ 对应的真实类别为 $c_g$,$I$ 在分类网络中的预测类别集合为 $C$,则融合结果 $S_{c_g}$ 的计算式为

$$S^j_{c_g}(x,y) = \begin{cases} G^j(x,y) & c_g \notin C \\ G^j(x,y) & c_g \in C \text{ and } \mathrm{cond}_1 \\ G^j(x,y)\cdot M_c(x,y) & c_g \in C \text{ and } \mathrm{cond}_2 \end{cases} \tag{4}$$

$$\mathrm{cond}_1 = \{\forall x,y, M_c(x,y) < \varepsilon\}_1 \tag{5}$$

$$\mathrm{cond}_2 = \{\exists x,y, M_c(x,y) \geqslant \varepsilon\}_1 \tag{6}$$

式中:$x \in (x^j, x^j + w^j)$,$y \in (y^j, y^j + h^j)$;$\{\mathrm{cond}_i\}_1$ 表示一个布尔值,满足 $\{\}$ 中的条件则为 true,否则为 false;$G^j$ 和 $M_c$ 分别为式(2)和式(1)中得到的结果;$\varepsilon$ 为设定的阈值。

在完成融合后,根据前背景阈值对式(4)中生成的融合激活图 $S$ 进行筛选,从而得到伪标签。依据实际经验,本文使用 0.3 作为前景阈值,0.05 作为背景阈值。也就是说,当某点类的激活值大于 0.3 时,认为该点属于这个类,如果有多个点均大于 0.3,则取激活值最大的类;如果某点各个类激活值均小于 0.05,则认为其为背景;忽略不满足以上 2 个条件的其他像素,认为这些像素是不可靠的,无法作为监督训练的标签。

### 3.2　边界检测网络

为了解决水下图像的边界分割问题,本文受文献[21]的启发设计了一个边界检测网络,网络的结构如图 7 下部分所示,该网络结构通过融合不同尺度的信息来更好地学习边界图。

对于有语义分割掩码标注的数据,直接使用其标注作为标签。对于只有框标注的数据,利用伪标签生成器得到其伪标签。根据获得的掩码标签中像素间类别的关系,将像素划分为同类像素对和不同类像素对 2 个集合。为了减小不必要的计算,只对距离较近的像素进行类关系的判断,使用一个距离阈值进行限制。

具体来说,对于 2 个像素点 $p_i$ 和 $p_j$:

$$P = \{(i,j) \mid \|x_i - x_j\| < \gamma, \forall i \neq j\} \tag{7}$$

$$P^+ = \{(i,j) \mid L(p_i) = L(p_j), (i,j) \in P\} \tag{8}$$

$$P^- = \{(i,j) \mid L(p_i) \neq L(p_j), (i,j) \in P\} \tag{9}$$

式中:$\gamma$ 为 2 个像素之间的最大欧氏距离;$L$ 为之前得到的伪标签。这样得到的 $P^+$ 即为属于同类的像素对(包括同属于前景类或同属于背景类),$P^-$ 为不属于同一类的像素对。

有了以上集合划分后,根据这些像素对来设计边界检测器网络的损失函数。如果 2 个像素属于同类像素,那么它们之间的连线应该均为该类别的像素,即它们之间不应该存在边界。反之,应该存在边界。因此,使用式(10)来衡量 2 个像素 $p_i$ 和 $p_j$ 之间的语义相关性 $a_{ij}$:

$$a_{ij} = 1 - \max B(x_k) \qquad k \in l(i,j) \tag{10}$$

式中:$l(i,j)$ 表示 $p_i$ 和 $p_j$ 连线上的像素点;$B(x_k)$ 为 $x_k$ 点属于边界的概率值,也就是说,$B$ 为最终要得到的边界概率图。当 $p_i$ 和 $p_j$ 为同类像素,希望 $a_{ij}$ 接近 1,反之,希望 $a_{ij}$ 接近 0。根据这个关系,本文将最终的损失函数设计成一个二元的交叉熵损失函数:

$$L^B = -\left(\sum_{(i,j)\in P^+}\frac{\ln a_{ij}}{|P^+|} + \sum_{(i,j)\in P^-}\frac{\ln(1-a_{ij})}{|P^-|}\right) \tag{11}$$

得到边界概率图后,使用文献[21]中随机游走的方式逐步迭代来完善式(4)中得到的融合激活图,使得类的激活区域向边界进行扩张,从而得到分割掩码,再经过条件随机场[22]后处理得到最终的分割结果。

## 4　实验结果

为了验证本文方法在真实水下场景 DUT-USEG 数据集上的有效性,将有语义分割标注的数据一部分划分为测试集,并将另一部分和只有框标注的数据作为训练集。具体来说,该训练集包含 5 863 张图像,其中 733 张为有语义分割标注的图像,约占训练集总数的 1/8,测试集包含 754 张图像。

### 4.1　实现细节

#### 4.1.1　超参数设置

在融合类激活图和框注意力图时,需要设定阈值来判断框内是否有激活区域,即式(5)和式(6)中的 $\varepsilon$,本文实验中取 $\varepsilon = 0.3$。

在生成伪标签时,需要设置前景阈值和背景阈值,在本文实验中分别将前景阈值和背景阈值

设置为 0.3 和 0.05。

在实现边界检测网络的损失函数时,需要设置像素点对之间的最大距离,即式(7)中的 $\gamma$,本文实验中将 $\gamma$ 设置为 10。

#### 4.1.2 训练细节

本文的所有训练过程是在一块 Tesla V100-SXM2-32GB 显卡上完成的。对于获取类激活图的分类网络,将初始学习率设置为 0.01,并使用学习率衰减的策略[23]训练了 30 个 epoch;对于边界检测网络,将初始学习率设置为 0.1,同样使用学习率衰减策略训练 30 个 epoch,训练中使用的优化方法是随机梯度下降法。本文的分类网络和边界检测网络均使用的是在 COCO 数据集[17]上预训练的 ResNet101[24]网络结构。

#### 4.1.3 其他细节

在生成类激活图时,为了得到更好的效果,本文将原图进行了不同尺度的缩放得到多张类激活图,并进行加和平均。

在对类激活图和框注意力图进行融合的过程中,由于类激活图是经过分类网络下采样后的结果,是原图尺寸的 1/4,而高斯注意力图则与原图尺寸相同。因此,先对类激活图使用双线性插值的方式进行 4 倍上采样,再对类激活图与框注意力图进行点对点乘积。

在测试过程中,观察到由于测试数据并不在分类网络的训练数据中,在测试数据上得到的类激活图效果很差,因此测试过程中舍弃类激活图,直接使用框的高斯注意力图作为激活图并使用边界图进行完善。

### 4.2 对比实验

将本文 US-Net 与同类方法(测试阶段具有框的先验知识)进行对比,对比方法包括 GrabCut[12]和 GGANet[16]。

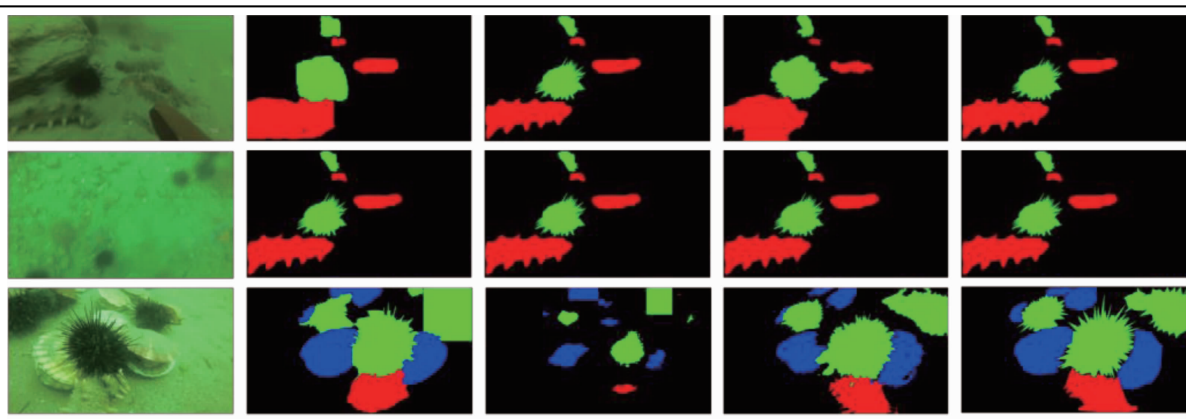
由于 GrabCut 在无法分割出前景时会将整个框内的所有像素作为前景,为了对比的公平性,将 GGANet 和本文方法的预测结果也做了同样的处理。由于 GGANet 的网络在本文数据集上训练后无法收敛,对比实验中使用的是原文中在 Pascal VOC 2012 上训练得到的用于分割未见类的模型。对比实验结果如表 1 所示。在 4 个类别上分别计算了交并比 IOU,并将 4 个类别和背景类别一起计算平均值,得到了平均交并比 mIOU,mIOU也是语义分割中最常见的评价指标。此外,还将 3 个方法在频率加权交并比 FWIOU 和像素精确度 PA 两个指标上进行了比较。

可以看到,本文方法在海参、海胆和海星 3 个类别上交并比都达到了最好,在平均交并比、频率加权交并比和像素精确度 3 个指标也均达到了最好;但在扇贝类别上的 IOU 低于 GrabCut,可能原因在于:本文数据集中扇贝的数量较少(见图 5),而 GrabCut 是一种传统方法,性能不会受到训练数据量的影响。图 8 展示了本文方法和其他 2 种方法的可视化结果。可以看到,由于边界检测器的作用,本文方法在边界区域有更好的分割效果。

**表 1 对比实验结果**

**Table 1 Results of contrast experiments**

| 方法 | IOU/% | | | | mIOU/% | FWIOU/% | PA/% |
|---|---|---|---|---|---|---|---|
| | 海参 | 海胆 | 扇贝 | 海星 | | | |
| GrabCut | 52.5 | 63.8 | 77.8 | 58.2 | 70.1 | 97.6 | 98.6 |
| GGANet | 32.1 | 42.9 | 52.8 | 34.5 | 52.1 | 96.6 | 98.0 |
| US-Net(本文方法) | 54.9 | 67.7 | 70.4 | 63.6 | 71.1 | 98.1 | 98.9 |



(a) 输入图像 (b) GrabCut (c) GGANet (d) 本文方法 (e) 标注

图 8 对比实验可视化结果

Fig. 8 Visualization results of contrast experiments

不同于 GrabCut 经常将框内大部分区域分割为前景,本文的分割结果更加接近物体的实际形状,这使得本文方法对水下生物的抓取具有更强的指导意义。

本文将 3 种方法的运行速度进行了比较,由于 GrabCut 是一种传统方法,没有参数,也不需要网络的前向传播过程,相比 GrabCut,本文方法的速度并无优势。但是,对比同为深度学习方法的 GGANet,本文方法在运行速度上具有较大优势,实验结果如表 2 所示。以测试集中每张图像的预测时间作为衡量指标。虽然两者的网络结构均基于 ResNet101,在参数量上没有明显差别,但由于本文方法在预测边界图和通过边界图迭代获取分割图的过程中是对整张图进行的,而 GGANet 则是先对一张图像中每个框内的物体单独进行分割,再进行融合。因此,在运行时间上,本文方法更具优势。

此外,本文实验是基于目标检测标注框进行的,测试阶段同样也可以在已有目标检测网络结果的基础上进行,从而使得本文方法在实际水下目标抓取中具有更灵活的应用价值。

**表 2　运行时间对比实验结果**

**Table 2　Results of contrast experiment for running time**

| 方法 | 运行时间/(s·张$^{-1}$) |
|---|---|
| GGANet | 38.9 |
| US-Net(本文方法) | 47.6 |

## 4.3　消融实验

### 4.3.1　边界检测网络消融实验

为了验证边界检测网络对实验结果的影响,进行了边界检测器网络的消融实验,实验结果如表 3 所示。可以看到,在测试阶段,使用边界检测网络对实验结果有明显提升。此外,从图 9 的边界图可视化效果可以看出,边界检测网络确实具有检测出水下生物边界的能力。

**表 3　边界检测网络消融实验结果**

**Table 3　Results of ablation experiment for boundary detection network**

| 方法 | mIOU/% |
|---|---|
| 只用高斯注意力图 | 68.8 |
| 高斯注意力图 + 边界检测网络(本文方法) | 71.1 |

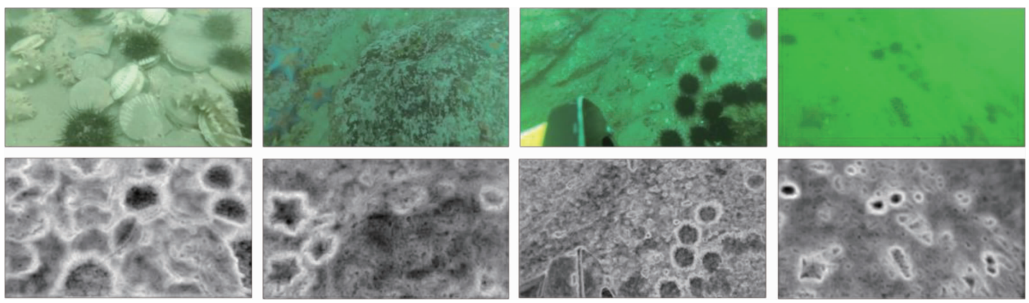


图 9　边界图可视化结果

Fig. 9　Visualization results of boundary maps

### 4.3.2　伪标签生成器消融实验

为了验证伪标签生成器对实验结果的提升效果及融合类激活图和高斯注意力图的必要性,对伪标签生成器进行了消融实验,实验结果如表 4 所示。可以看到,相比于不使用伪标签生成器(即只用部分标注数据做半监督),只使用框高斯注意力图作为伪标签训练边界检测网络对实验结果的提升有限,而融合了类激活图后,对结果提升明显。

**表 4　伪标签生成器消融实验结果**

**Table 4　Results of ablation experiment for pseudo label generator**

| 方法 | mIOU/% |
|---|---|
| 不使用伪标签生成器 | 67.6 |
| 高斯注意力图作为伪标签 | 68.6 |
| 高斯注意力图 + 类激活图作为伪标签(本文方法) | 71.1 |

### 4.3.3　超参数消融实验

在生成伪标签的前景阈值和边界损失函数中像素对的最大距离(即式(7)中的 $\gamma$)2 个超参数上,本文进行了消融实验,实验结果如表 5 和表 6 所示。

从表 5 的实验结果可以看到,当前景阈值过高或过低时,都会影响伪标签的生成质量,从而造成性能的下降,当阈值选取为 0.3 时,实验效果最佳。

表 6 中的实验结果表明,当像素对最大距离 $\gamma$ 增加时,由于考虑了更大范围像素之间的类别关系,实验结果会有所提升。观察到当 $\gamma$ 从 5 提升到 10 后,实验结果有明显提升,但当 $\gamma$ 从 10 增加到 15 后,对实验结果的提升并不明显,而实际训练时间和显存占用则明显增加,因此从精度和效率进行综合考虑,最终将 $\gamma$ 设置为 10。

表5 伪标签前景阈值消融实验结果

Table 5 Results of ablation experiment for foreground threshold of pseudo labels

| 前景阈值 | 0.2 | 0.3 | 0.4 |
|---|---|---|---|
| mIOU % | 70.5 | 71.1 | 69.7 |

表6 像素对最大距离消融实验结果

Table 6 Results of ablation experiment for maximum distance of pixel pairs

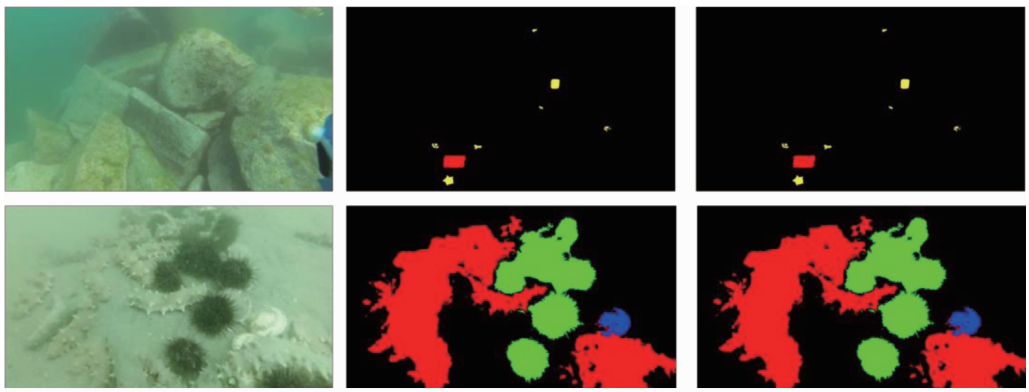
| 最大距离 $\gamma$ | 5 | 10 | 15 |
|---|---|---|---|
| mIOU/% | 70.2 | 71.1 | 71.3 |

### 4.4 局限性分析

由于在本文数据集中存在大量的高分辨率图像(见图4),在训练和测试过程中需要对这些图像进行下采样,以减小计算量,这就导致一些小物体在下采样过程中损失了较多信息,本文方法在部分高分辨率图像的小物体上表现较差,如图10第1行图像所示。本文准备在之后的工作中尝试更好的处理高分辨率图像。

此外,由于本文的边界检测网络检测的只是类别之间的边界,导致一些同类但相距很近的物体之间的边界难以被区分,如图10第2行图像所示。未来会尝试设计一个检测实例之间边界的网络来解决这一问题。

运行速度较慢也是本文方法在实际水下抓捕应用场景中存在的一个问题,由于在测试阶段需要先通过网络生成边界图,再通过随机游走的方式进行迭代才可获得最终语义分割结果,整个网络运行时间较长,未来会尝试设计一个一阶段的网络来直接得到语义分割结果。



(a) 输入图像 (b) 本文结果 (c) 标注

图10 失败案例

Fig. 10 Failure examples

## 5 结 论

1)为了推进水下图像语义分割方向的研究,本文提出了一个水下语义分割数据集DUT-USEG,该数据集包含6 617张水下图像,其中1 487张图像具有语义分割和实例分割标注。

2)本文提出了一个半监督语义分割网络(US-Net),主要包含伪标签生成器和边界检测子网络,用来解决水下图像中物体与背景之间边界难以区分的问题。实验表明,该方法在目前的同类方法中达到了最高精度。

未来会在水下图像分割领域做更深入的研究,将力争解决目前方法中存在的高分辨率小物体分割不准确、同类物体边界分割不准确及运行速度较慢的问题。

### 参考文献(References)

[1] LI C, GUO C, REN W, et al. An underwater image enhancement benchmark dataset and beyond[J]. IEEE Transactions on Image Processing, 2019, 29: 4376-4389.

[2] LI C, ANWAR S, PORIKLI F. Underwater scene prior inspired deep underwater image and video enhancement[J]. Pattern Recognition, 2020, 98: 107038.

[3] GUO Y, LI H, ZHUANG P. Underwater image enhancement using a multiscale dense generative adversarial network[J]. IEEE Journal of Oceanic Engineering, 2019, 45(3): 862-870.

[4] CHEN L, LIU Z, TONG L, et al. Underwater object detection using invert multi-class adaboost with deep learning[C]//2020 International Joint Conference on Neural Networks (IJCNN). Piscataway: IEEE Press, 2020: 1-8.

[5] LI X, SHANG M, QIN H, et al. Fast accurate fish detection and recognition of underwater images with fast R-CNN[C]//OCEANS 2015-MTS/IEEE Washington. Piscataway: IEEE Press, 2015: 1-5.

[6] VILLON S, CHAUMONT M, SUBSOL G, et al. Coral reef fish detection and recognition in underwater videos by supervised machine learning: Comparison between deep learning and HOG + SVM methods[C]// International Conference on Advanced Concepts for Intelligent Vision Systems. Berlin: Springer, 2016:

160-171.

[7] LIU C W,LI H J,WANG S C,et al. A dataset and benchmark of underwater object detection for robot picking[EB/OL]. (2021-06-10)[2021-09-01]. https://arxiv.org/abs/2106.05681.

[8] JIAN M,QI Q,DONG J,et al. The OUC-vision large-scale underwater image database[C]//2017 IEEE International Conference on Multimedia and Exposition (ICME). Piscataway:IEEE Press,2017:1297-1302.

[9] LONG J,SHELHAMER E,DARRELL T. Fully convolutional networks for semantic segmentation[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway:IEEE Press,2015:3431-3440.

[10] DAI J,HE K,SUN J. BoxSup:Exploiting bounding boxes to supervise convolutional networks for semantic segmentation[C]// Proceedings of the IEEE International Conference on Computer Vision. Piscataway:IEEE Press,2015:1635-1643.

[11] KHOREVA A,BENENSON R,HOSANG J,et al. Simple does it:Weakly supervised instance and semantic segmentation[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway:IEEE Press,2017:876-885.

[12] ROTHER C,KOLMOGOROV V,BLAKE A. "GrabCut" interactive foreground extraction using iterated graph cuts[J]. ACM Transactions on Graphics,2004,23(3):309-314.

[13] PONT-TUSET J,ARBELAEZ P,BARRON J T,et al. Multiscale combinatorial grouping for image segmentation and object proposal generation[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,2016,39(1):128-140.

[14] SONG C,HUANG Y,OUYANG W,et al. Box-driven class-wise region masking and filling rate guided loss for weakly supervised semantic segmentation[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway:IEEE Press,2019:3136-3145.

[15] LEE J,YI J,SHIN C,et al. BBAM:Bounding box attribution map for weakly supervised semantic and instance segmentation[EB/OL]. (2021-03-16)[2021-09-01]. https://arxiv.org/abs/2103.08907.

[16] ZHANG P,WANG Z,MA X,et al. Learning to segment unseen category objects using gradient gaussian attention[C]//2019 IEEE International Conference on Multimedia and Exposition (ICME). Piscataway:IEEE Press,2019:1636-1641.

[17] 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.

[18] EVERINGHAM M,VAN GOOL L,WILLIAMS C K I,et al. The pascal visual object classes (VOC) challenge[J]. International Journal of Computer,2015,111:98-136.

[19] CHENG H K,CHUNG J,TAI Y W,et al. CascadePSP:Toward class-agnostic and very high-resolution segmentation via global and local refinement[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway:IEEE Press,2020:8890-8899.

[20] ZHOU B,KHOSLA A,LAPEDRIZA A,et al. Learning deep features for discriminative localization[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway:IEEE Press,2016:2921-2929.

[21] AHN J,CHO S,KWAK S. Weakly supervised learning of instance segmentation with inter-pixel relations[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway:IEEE Press,2019:2209-2218.

[22] KRÄHENBÜHL P,KOLTUN V. Efficient inference in fully connected CRFS with gaussian edge potentials[J]. Advances in Neural Information Processing Systems,2011,24:109-117.

[23] LIU W,RABINOVICH A,BERG A C. ParseNet:Looking wider to see better[EB/OL]. (2015-11-19)[2021-09-01]. https://arxiv.org/abs/1506.04579.

[24] 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.

# A real scene underwater semantic segmentation method and related dataset

MA Zhiwei[1], LI Haojie[1], FAN Xin[1], LUO Zhongxuan[1], LI Jianjun[2], WANG Zhihui[1,*]

(1. International School of Information Science & Engineering, Dalian University of Technology, Dalian 116621, China;
2. School of Computer and Software, Hangzhou Dianzi University, Hangzhou 310018, China)

**Abstract**: Underwater object recognition and segmentation with high accuracy have become a challenge with the development of underwater object grabbing technology. The existing underwater object detection technology can only give the general position of an object, unable to give more detailed information such as the outline of the object, which seriously affects the grabbing efficiency. To address this problem, we label and establish underwater semantic segmentation dataset of a real scene (DUT-USEG). The DUT-USEG dataset includes 6 617 images, 1 487 of which have semantic segmentation and instance segmentation annotations, and the remaining 5 130 images have object detection box annotations. Based on this dataset, we propose a semi-supervised underwater semantic segmentation network (US-Net) focusing on the boundaries. By designing a pseudo label generator and a boundary detection subnetwork, this network realizes the fine learning of boundaries between underwater objects and background, and improves the segmentation effect of boundary areas. Experiments show that the proposed method improves by 6.7% in three categories of holothurian, echinus, and starfish in DUT-USEG dataset, and achieves state-of-the-art results.

**Keywords**: underwater object grabbing; semantic segmentation; semi-supervised learning; weakly supervised learning; boundary detection

**Received**: 2021-09-06; **Accepted**: 2021-10-01; **Published online**: 2022-01-28 11:58

**URL**: kns.cnki.net/kcms/detail/11.2625.V.20220127.1107.001.html

**Foundation items**: National Natural Science Foundation of China (61976038, 61932020, 61772108, U1908210)

* **Corresponding author**. E-mail: zhwang@dlut.edu.cn

http://bhxb.buaa.edu.cn jbuaa@buaa.edu.cn
**DOI**: 10.13700/j.bh.1001-5965.2021.0597

# 外观动作自适应目标跟踪方法

熊珺瑶, 王蓉*, 孙义博
(中国人民公安大学 信息与网络安全学院, 北京 100038)

**摘 要**: 为降低目标运动时产生的外观形变对目标跟踪的影响,在 DaSiamese-RPN 基础上进行改进,提出了一种外观动作自适应的目标跟踪方法。在孪生网络的子网络中引入外观动作自适应更新模块,融合目标的时空信息和动作特征;利用 2 种欧氏距离分别度量真实图和预测图之间的全局和局部差异,并对二者加权融合构建损失函数,加强预测目标特征图与真实目标特征图之间全局和局部信息的关联性。在 VOT2016、VOT2018、VOT2019 和 OTB100 数据集上进行测试,实验结果表明:在 VOT2016 和 VOT2018 数据集上,预测平均重叠率分别提高 4.5% 和 6.1%;在 VOT2019 数据集上,准确度提高 0.4%,预测平均重叠率降低 1%;在 OTB100 数据集上,跟踪成功率提高 0.3%,精确度提高 0.2%。

**关 键 词**: 目标跟踪; 外观动作自适应; 孪生网络; 特征融合; 外观形变
**中图分类号**: TP183
**文献标志码**: A **文章编号**: 1001-5965(2022)08-1525-09

当目标在跟踪过程中出现复杂运动时,跟踪目标外观和大小发生变化,会造成目标的丢失。

近年来,为了适应目标复杂动作时产生的外观变化,降低此变化给跟踪带来的干扰,基于孪生网络的跟踪器不断改进。DaSiamese-RPN[1]跟踪器通过引入干扰感知模块,增加训练时的负样本,使模型有效地捕捉更多上下文信息以适应目标外观变化,从而提高跟踪性能;SiamRPN ++[2]、SiamMask[3]和 SiamDW[4]通过将 ResNet[5]、ResNeXt[6]和 MobileNet[7]等深层神经网络引入基于孪生网络的视频跟踪器中,获得更深层的局部性特征,从而处理跟踪过程中目标运动时引起的纵横比变化;Siam R-CNN[8]通过引入重检测器和动态编程算法(the dynamic programming algorithm,TDPA),重新检测并跟踪前期视频帧中的所有目标,得到运动轨迹,选出当前时间步长中与真实标签最接近的运动目标,从而找回丢失目标;基于 UpdateNet[9]的更新策略应用于基于孪生网络的跟踪器中,通过引入卷积神经网络(convolutional neural network,CNN)实现跟踪过程中模板特征的在线更新,使得模板信息更丰富,从而增强跟踪器性能。虽然现有的基于孪生网络的目标跟踪方法已表现出较高的准确性和鲁棒性,但目标的外观和动作的变化通常很大,前期跟踪模板信息没能适应目标变化进行更新,可能会导致跟踪器的早期故障。大多数跟踪方法使用简单的线性算法或单一的卷积层对前期跟踪模板进行更新,这种策略获取的特征较少,同时在实际应用场景下无法适应目标运动产生的变化速率。

本文在 DaSiamese-RPN 的基础上提出了一种外观动作自适应目标跟踪方法。在特征提取的子网络中引入外观动作自适应更新模块,先利用模块中的 ACTION-Net[10],以多路径激励的方式融合目标的时空信息及动作特征,使网络有效地识别目标外观和动作的变化;然后,通过 2 种欧氏距离分别对全局和局部的差异进行度量,并对二者加权融

**收稿日期**: 2021-10-09; **录用日期**: 2021-10-29; **网络出版时间**: 2021-11-16 09:52
**网络出版地址**: kns.cnki.net/kcms/detail/11.2625.V.20211115.1929.001.html
**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**: *熊珺瑶, 王蓉, 孙义博. 外观动作自适应目标跟踪方法[J]. 北京航空航天大学学报, 2022, 48(8): 1525-1533.*
*XIONG J Y, WANG R, SUN Y B. Appearance and action adaptive target tracking method[J]. Journal of Beijing University of Aeronautics and Astronautics, 2022, 48(8): 1525-1533 (in Chinese).*

合构建损失函数,以此加强预测目标特征图与真实目标特征图之间全局和局部信息的关联性。

## 1 DaSiamese-RPN 方法

DaSiamese-RPN 跟踪方法沿用 Siamese-RPN 的主干网络[11-12],将目标跟踪的问题化为学习目标外观模板特征和搜索区域之间特征表示的互相关性的问题,为解决视频中出现多个相似目标而对跟踪目标产生干扰的问题,采用数据增强的方式对数据集进行扩充,增加了训练时目标的种类,同时在训练时利用相同种类但不同目标的负图片对模型进行训练,使得网络可以对同类进行区分,从而提升跟踪的准确度。DaSiamese-RPN 跟踪方法整体框架如图 1 所示。

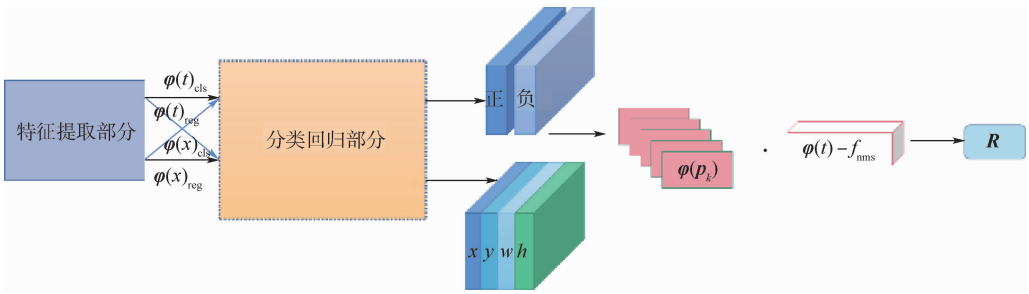



图 1 DaSiamese-RPN 整体框架

Fig. 1 Overall framework of DaSiamese-RPN

DaSiamese-RPN 跟踪框架由 2 个部分组成,即特征提取部分(见图 2(a))和分类回归部分(见图 2(b))。特征提取部分采用孪生网络对模板帧和当前帧目标进行特征提取,孪生网络利用 AlexNet 作为特征提取主干网络,生成特征分别记为 $\varphi(t)$ 和 $\varphi(x)$。分类回归部分利用区域建议网络(region proposal network,RPN),将特征提取部分生成的特征图划分为分类和回归分支,即 $\varphi(t)_{cls}$、$\varphi(t)_{reg}$ 和 $\varphi(x)_{cls}$、$\varphi(x)_{reg}$,为了在跟踪过程中更好地利用上下文视频帧信息,增强跟踪的鲁棒性,在 RPN 网络中利用非最大值抑制(NMS)选出一定量的干扰样本与当前帧搜索区域进行响应,特征之间的响应方式如式(1)所示,利用干扰样本与当前帧搜索区域进行响应并计算响应分数,过程如式(2)所示。

$$f(\boldsymbol{d}_j,\boldsymbol{p}_k)=\varphi(\boldsymbol{d}_j)\cdot\varphi(\boldsymbol{p}_k) \tag{1}$$

$$f_{nms}=\frac{\hat{a}\sum_{j=1}^{n}a_j f(\boldsymbol{d}_j,\boldsymbol{p}_k)}{\sum_{j=1}^{n}a_i} \tag{2}$$

式中:$n$ 为选取的干扰样本的数量;$\boldsymbol{d}_j$ 为非极大值抑制选出的高于某阈值的干扰物;$\boldsymbol{p}_k$ 为 RPN 网络生成的区域建议框,实验中共设置 5 个区域建议框,即 $k=5$;$f(\boldsymbol{d}_j,\boldsymbol{p}_k)$ 表示干扰物与区域建议框之间的响应函数;$\hat{a}$ 为权重因子,设为 0.5;$a_j$ 为每个干扰物的权重,统一设为 1;“·”表示向量之间逐元素相乘。

最终的响应分数以模板特征图与干扰物特征和检测区域建议框之间互相关分数的差值表示,选取响应分数最高的区域建议框作为跟踪结果 $\boldsymbol{R}$,如下:

$$\boldsymbol{R}=\arg\max(\varphi(t)-f_{nms})\cdot\varphi(\boldsymbol{p}_k) \tag{3}$$

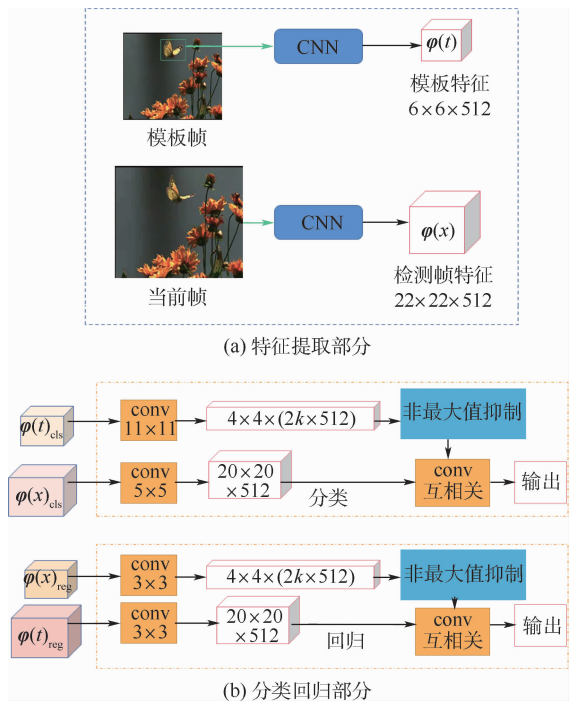



图 2 DaSiamese-RPN 跟踪框架

Fig. 2 Tracking framework of DaSiamese-RPN

## 2 改进方法

为适应目标运动时产生的外观和大小变化对目标跟踪的干扰,简单地对跟踪前期收集的目标模板特征进行线性更新不能满足运动带来的复杂变化。本文利用外观动作自适应跟踪方法对跟踪过程中前期收集的目标特征模板进行更新。外观动作自适应跟踪方法以 DaSiamese-RPN 网络为基

础,在孪生网络的模板帧特征提取子网络引入外观动作自适应更新模块,并利用 2 种欧氏距离分别度量全局和局部信息的差异构建损失函数以获取更丰富的目标模板特征。外观动作自适应目标跟踪框架如图 3 所示。

外观动作自适应更新模块以初始帧所标记的目标真实标签特征 $\boldsymbol{A}^{\mathrm{GT}}$、跟踪前期所有目标轨迹的累积特征模板 $\tilde{\boldsymbol{A}}_{i-1}$、当前帧目标的预测特征模板 $\boldsymbol{A}_i$ 作为输入,其中,$i$ 表示当前帧的序号。目的是通过卷积神经网络自适应地对累积特征模板进行更新。外观动作自适应网络通过学习函数 $\sigma$ 来实现,本质是通过整合当前帧目标的视觉信息更新之前的累积特征模板。为获得高度可靠的信息,在进行更新时还考虑了初始真实标签模板的信息。网络实现过程如下:

$$\tilde{\boldsymbol{A}}_i = \sigma(\boldsymbol{A}^{\mathrm{GT}}, \tilde{\boldsymbol{A}}_{i-1}, \boldsymbol{A}_i) \tag{4}$$

外观动作自适应更新模块由 2 层卷积层和 Action-Net 组成,具体是一个 $1\times1\times3C\times96$ 的卷积层,通过 ReLU 函数激活后,输入 $1\times1\times C\times96$ 的卷积层,将 $C$ 设为 512,将卷积输出的结果送入 Action-Net。外观动作自适应更新模块框架如图 4 所示。

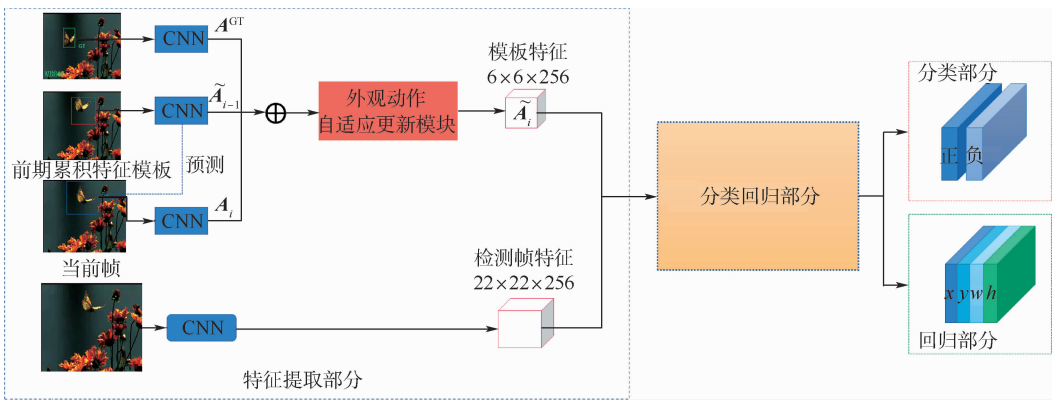



图 3　外观动作自适应目标跟踪框架

Fig. 3　Framework of appearance and action adaptive target tracking

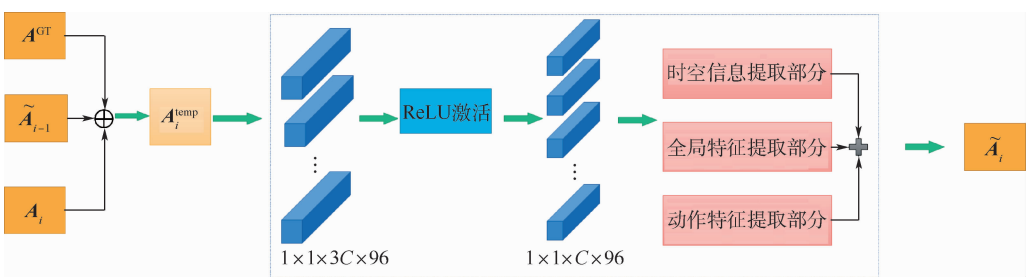



图 4　外观动作自适应更新模块框架

Fig. 4　Framework of appearance and action adaptive module

## 2.1　Action-Net

Action-Net 通过一个轻量级网络获取视频图像中目标的时空信息和运动特征,由 3 个模块组成,分别为时空激励(spatio-temporal excitation, STE)、通道激励(channel excitation, CE)和动作激励(motion excitation, ME)。Action-Net 框架如图 5 所示。

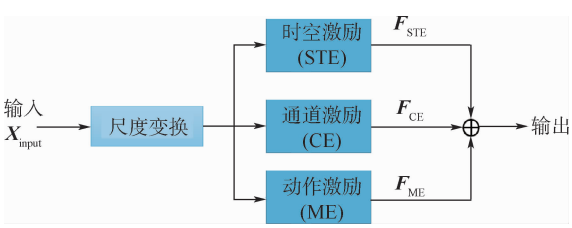



图 5　Action-Net 框架

Fig. 5　Framework of Action-Net

时空激励结构如图 6(a)所示,用于获取所有通道上的时空信息,生成时空信息掩码 $\boldsymbol{M}_{\mathrm{STE}} \in \mathbf{R}^{N\times T\times1\times H\times W}$,并与特征进行结合,给视频帧特征赋予时空信息。具体实现方法是:输入一个 5 维向量 $\boldsymbol{X}_{\mathrm{input}} \in \mathbf{R}^{N\times T\times C\times H\times W}$,$N$ 为批次量,$T$ 为将完整视频划分为短视频的段数,$C$ 为通道数,$H$ 为输入视频帧的高度,$W$ 为输入视频帧的宽度。将输入向量在通道维度上取平均并进行尺度变化,得到 $\tilde{\boldsymbol{X}} \in \mathbf{R}^{N\times1\times T\times H\times W}$,将 $\tilde{\boldsymbol{X}}$ 输入一个 $3\times3\times3$ 的卷积中,并再一次进行尺度变化,输出为 $\boldsymbol{X}^{*}_{\mathrm{STE}} \in \mathbf{R}^{N\times T\times1\times H\times W}$,通过 Sigmoid 函数激活,得到时空信息掩码。时空激励的输出表示为输入向量与进行时空信息激励之后的视频帧特征的加和,如式(5)所示,最终输出向量表示方法如式(6)所示。

$$\boldsymbol{M}_{\mathrm{STE}} = \delta(\boldsymbol{X}^{*}_{\mathrm{STE}}) \tag{5}$$

$$\boldsymbol{F}_{\mathrm{STE}} = \boldsymbol{X}_{\mathrm{input}} + \boldsymbol{M}_{\mathrm{STE}} \odot \boldsymbol{X}_{\mathrm{input}} \tag{6}$$

通道激励结构如图 6(b)所示,用于获取视频帧上的全局信息,生成全局信息掩码 $\boldsymbol{M}_{\mathrm{CE}} \in \mathbf{R}^{N\times C\times T\times 1\times 1}$,并与输入特征进行结合,给视频帧特征赋予全局信息。具体实现方式是:给定输入的 5 维向量,通过空间平均池化获取特征图的全局信息,如下:

$$\boldsymbol{X}_{\mathrm{avg}} = \frac{1}{H\times W}\sum_{h=1}^{H}\sum_{w=1}^{W}\boldsymbol{X}_{\mathrm{input}}\{:,:,:,h,w\}$$

$$\boldsymbol{X}_{\mathrm{avg}} \in \mathbf{R}^{N\times T\times C\times 1\times 1} \tag{7}$$

将 $\boldsymbol{X}_{\mathrm{avg}}$ 以通道缩减率 $r=16$ 进行通道压缩,送入全连接层,进行尺度变换后送入核为 3 的 1×1卷积中,再送入全连接层,输出为 $\boldsymbol{X}^{*}_{\mathrm{CE}} \in \mathbf{R}^{N\times T\times C\times 1\times 1}$,利用 Sigmoid 激活函数激活得到全局信息掩码。通道激励的输出表示为输入向量与进行全局信息激励之后的视频帧特征向量的加和,如式(8)所示,最终输出向量如式(9)所示。

$$\boldsymbol{M}_{\mathrm{CE}} = \delta(\boldsymbol{X}^{*}_{\mathrm{CE}}) \tag{8}$$

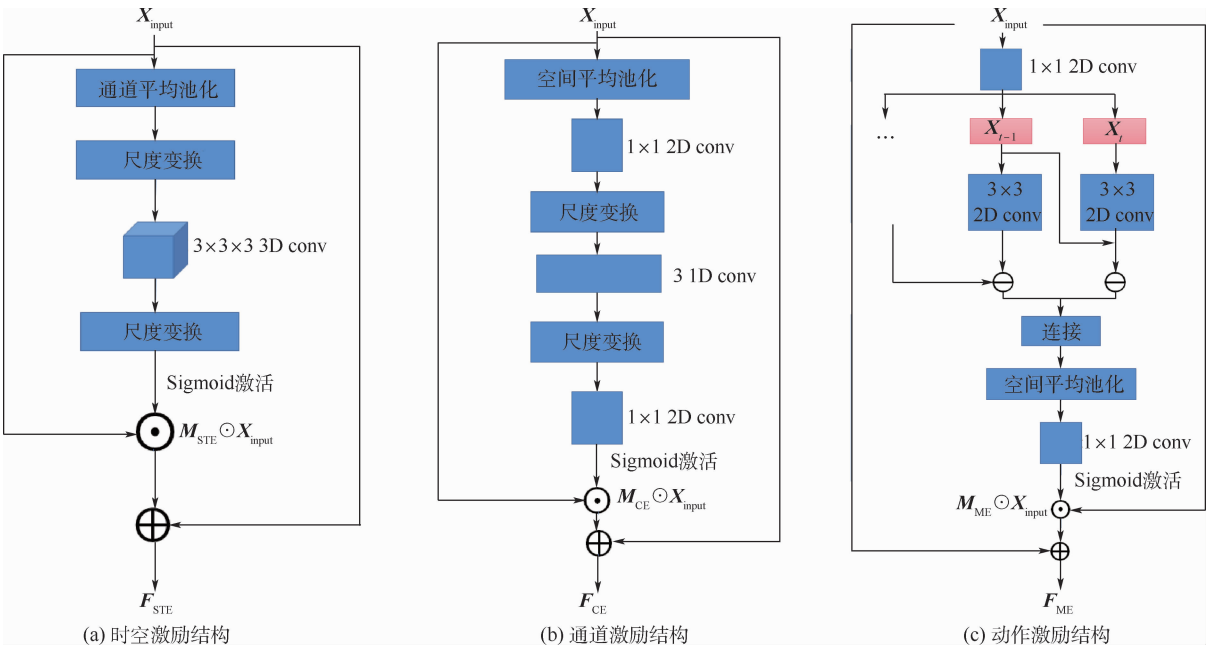
$$\boldsymbol{F}_{\mathrm{CE}} = \boldsymbol{X}_{\mathrm{input}} + \boldsymbol{M}_{\mathrm{CE}} \odot \boldsymbol{X}_{\mathrm{input}} \tag{9}$$

动作激励结构如图 6(c)所示,用于获取视频帧中的运动特征,利用在相邻帧之间的差值进行建模,思路与前 2 个模块类似。具体实现方法是对输入的视频段中每一帧图像进行处理。将输入的 5 维向量送入一个 2 维的 1×1 卷积,每一个视频帧的特征向量为 $\boldsymbol{X}_T = [\boldsymbol{X}_1, \boldsymbol{X}_2, \cdots, \boldsymbol{X}_t]$。运动特征建模的方式是:将上一帧的特征与当前帧送入 3×3的卷积层后得到的特征相减,将整段视频的特征连接,并将最后一帧的运动特征用 0 进行填充,得到整个视频的全局运动特征 $\boldsymbol{X}^{*}_{\mathrm{ME}} \in \mathbf{R}^{N\times T\times \frac{C}{r}\times H\times W}$,输出送入全连接层后,通过 Sigmoid 函数激活得到动作特征掩码 $\boldsymbol{M}_{\mathrm{ME}}$。动作激励的输出表示为输入向量与进行动作特征激励之后的输出向量的加和,如式(10)所示,最终的输出如式(11)所示。

$$\boldsymbol{M}_{\mathrm{ME}} = \delta(\boldsymbol{X}^{*}_{\mathrm{ME}}) \tag{10}$$

$$\boldsymbol{F}_{\mathrm{ME}} = \boldsymbol{X}_{\mathrm{input}} + \boldsymbol{M}_{\mathrm{ME}} \odot \boldsymbol{X}_{\mathrm{input}} \tag{11}$$

最终 Action-Net 的输出是将 3 个模块的输出进行逐元素相加。




(a) 时空激励结构 (b) 通道激励结构 (c) 动作激励结构

图 6 Action-Net 各部分结构

Fig. 6 Action-Net structure of each part

## 2.2 损失函数

损失函数用于判定真实值与模型预测值之间的差异,通过最小化损失函数优化网络模型。大多数跟踪模型仅仅利用预测图特征与真实标签特征的欧氏距离构建损失函数,只考虑到了像素误差,忽略了预测和真实标签之间的全局性差异。本文利用 2 种欧氏度量分别度量真实标签和预测模板之间的全局和局部差异,并将两者加权联合构建损失函数。损失函数构建框架如图 7 所示。

利用欧氏距离建模真实标签与预测特征标签之间的像素级差异,构建局部信息损失,并利用 $L_2$ 正则进行标准化,局部信息损失函数定义为

$$L_{\mathrm{P}} = \left\|\sigma(\boldsymbol{A}^{\mathrm{GT}}, \boldsymbol{A}_{i-1}, \tilde{\boldsymbol{A}}_{i})\right\|_2 \tag{12}$$

仅仅利用局部信息差异训练更新模板网络忽略了预测图和真实标签之间全局性的差异。针对以上问题,引入全局池化获取真实标签与预测图特征的全局信息,采用欧氏距离度量,并利用 $L_1$ 正则标准化。全局信息损失函数定义为

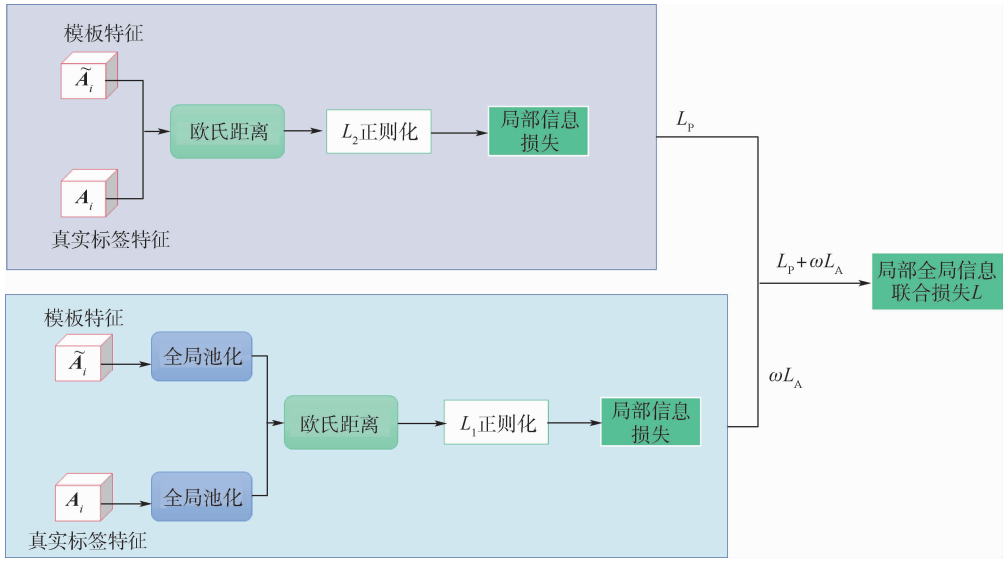



图 7　全局局部信息联合损失框架

Fig. 7　Framework of global and local information combination loss function

$$L_A = \| \mathrm{Avgpool}(\sigma(\sigma(\boldsymbol{A}_0^{GT}, \boldsymbol{A}_{i-1}, \tilde{\boldsymbol{A}}_i) - \boldsymbol{A}_{i+1}^{GT})) - \mathrm{Avgpool}(\boldsymbol{A}_{i+1}^{GT}) \| \tag{13}$$

本文联合局部和全局信息损失构建损失函数,为增强局部信息和全局信息之间的关联性及研究局部损失和全局损失对模型优化的影响,采用加权联合的方式,如下:

$$L = L_P + \omega L_A \tag{14}$$

式中:$\omega$ 为全局信息系数,实验中将 $\omega$ 设为 1 000、500、100、0。

## 3　仿真实验与结果分析

### 3.1　数 据 集

实验使用大规模单目标跟踪数据集 LaSOT[13] 训练模型。LaSOT 由 70 个目标类别、1 400 个视频序列组成,共计 352 万帧,每个类别包含 20 个序列。除此之外,LaSOT 数据集提供了 1 000 帧以上的较长序列,满足进行较长时间的跟踪。训练使用 LaSOT 数据集的一个子集,包含 20 个随机选择的目标及序列,共计 45 578 帧。

训练的模型分别在 VOT2016[14-15]、VOT2018[16]、VOT2019[17] 及 OTB100[18] 数据集上进行测试。

VOT2018 数据集由 60 个序列组成,共 21 356 帧,VOT2016 与 VOT2018 数据集有 10 个不同的序列,VOT2019 数据集与 VOT2018 数据集相比,被替换了 20% 的序列。对于 VOT 系列数据集,采用预测平均重叠率(EAO)、准确度(Accuracy)及鲁棒性(Robustness)对不同跟踪器进行比较。

OTB100 数据集由 100 个序列组成,平均长度为 590 帧,采用跟踪成功率和精确度对不同跟踪器的性能进行比较。

### 3.2　仿真实验环境

本文实验在 NVIDIA GTX3090Ti GPU 上使用 Pytorch 框架进行,以 DaSiamese-RPN 作为基础跟踪器,主干网络使用 AlexNet 提取特征,进行 2 次实验。实验 1 直接应用 DaSiamese-RPN 进行跟踪;实验 2 引入外观动作自适应更新模块,分 2 个阶段训练,每个阶段训练 50 轮,共训练 100 轮,以批次量为 64 进行训练。第 1 阶段学习速率在每个周期从 $10^{-5}$ ~ $10^{-6}$ 以对数方式降低,第 2 阶段在第 1 阶段最佳模型的基础上进行训练,学习速率在每个周期从 $10^{-7}$ ~ $10^{-8}$ 以对数方式降低,利用随机梯度下降(SGD)对模型进行优化,动量设置为 0.9,权重衰减参数设置为 0.000 5,为研究全局信息和局部信息对模型性能的影响,分别将全局信息系数设为 1 000、500、100、0 进行实验。

### 3.3　实验与结果评估

本文实验在 VOT2016、VOT2018、VOT2019 及 OTB100 数据集上进行测试,分别对原始 DaSiamese-RPN 跟踪模型、引入外观动作自适应更新模块的目标跟踪模型进行性能评估。

在 VOT2016 数据集上进行测试的结果如表 1 所示。引入外观动作自适应更新模块后,相比 DaSiamese-RPN 跟踪器,跟踪性能有所提升。全局信息系数设置为 0 时,准确度提升了 1.7%,鲁棒性降低了 0.6%,预测平均重叠率提升了 1.4%;全局信息系数设置为 1 000 时,准确度提升了 0.3%,鲁棒性降低了 2.4%,预测平均重叠率提

升了4.4%;全局信息系数设置为500时,准确度下降了0.1%,鲁棒性降低了2.4%,预测平均重叠率提升了3.1%;全局信息系数设置为100时,效果最好,准确度提升了0.4%,鲁棒性降低了3.4%,预测平均重叠率提升了4.5%。

在VOT2018数据集上进行测试的结果如表2所示。引入外观动作自适应更新模块后,相比DaSiamese-RPN跟踪器,跟踪性能有所提升。全局信息系数设置为0时,准确度提升了1.5%,鲁棒性降低了4.2%,预测平均重叠率提升了2.6%;全局信息系数设置为1 000时,准确度提升了1.6%,鲁棒性降低了5.1%,预测平均重叠率提升了4.6%;全局信息系数设置为500时,效果最好,准确度上升了1.6%,鲁棒性降低了7.9%,预测平均重叠率提升了6.1%;全局信息系数设置为100时,准确度提升了1.6%,鲁棒性降低了5.1%,预测平均重叠率提升了3.9%。

在VOT2019数据集上进行测试的结果如表3所示。引入外观动作自适应更新模块后,相比DaSiamese-RPN跟踪器,跟踪准确度有所提升,其他性能有所下降。全局信息系数设置为0时,准确度提升了0.1%,鲁棒性上升了2%,预测平均重叠率下降了0.5%;全局信息系数设置为1 000时,准确度提升了0.3%,鲁棒性上升了2.5%,预测平均重叠率下降了0.4%;全局信息系数设置为500时,准确度上升了0.4%,鲁棒性上升了2.5%,预测平均重叠率下降了1%;全局信息系数设置为100时,准确度提升了0.3%,鲁棒性上升了3%,预测平均重叠率下降了1.2%。

**表1 VOT2016数据集测试结果**

**Table 1 Results of testing on VOT2016 dataset**

| 模型 | 准确度/% | 鲁棒性/% | 预测平均重叠率/% |
|---|---|---|---|
| DaSiamese-RPN | 61 | 22 | 41.1 |
| 本文($\omega=0$) | 62.7 | 21.4 | 42.5 |
| 本文($\omega=1\ 000$) | 61.3 | 19.6 | 45.5 |
| 本文($\omega=500$) | 60.9 | 19.6 | 44.2 |
| 本文($\omega=100$) | 61.4 | 18.6 | 45.6 |

**表2 VOT2018数据集测试结果**

**Table 2 Results of testing on VOT2018 dataset**

| 模型 | 准确度/% | 鲁棒性/% | 预测平均重叠率/% |
|---|---|---|---|
| DaSiamese-RPN | 56.9 | 33.7 | 32.6 |
| 本文($\omega=0$) | 58.4 | 29.5 | 35.2 |
| 本文($\omega=1\ 000$) | 58.5 | 28.6 | 37.2 |
| 本文($\omega=500$) | 58.5 | 25.8 | 38.7 |
| 本文($\omega=100$) | 58.5 | 28.6 | 36.5 |

在OTB100数据集上进行测试的结果如表4所示。引入外观动作自适应更新模块之后,相比DaSiamese-RPN跟踪器,性能有较小提升。全局信息系数设为0时,跟踪成功率提升了0.3%,精确度提升了0.2%;全局信息系数设为1 000时,跟踪成功率降低了0.1%。

将输出结果进行可视化,结果如图8和图9所示。

由图8可以看出,外观动作自适应跟踪方法在遮挡场景下有利于目标找回,图8(c)发现目标在跟踪过程中丢失,但是在图8(d)中,外观动作自适应跟踪方法将目标找回,DaSiamese-RPN跟踪方法中目标持续丢失。由图9可以看出,外观自适应跟踪方法在目标动作复杂变化场景下表现良好,图9(c)可以看出,在目标动作发生复杂变化时,外观动作自适应跟踪方法在调整全局局部信息联合损失的权重下有较好的表现,但是改进前方法使目标丢失。

**表3 VOT2019数据集测试结果**

**Table 3 Results of testing on VOT2019 dataset**

| 模型 | 准确度/% | 鲁棒性/% | 预测平均重叠率/% |
|---|---|---|---|
| DaSiamese-RPN | 58.2 | 52.7 | 27.2 |
| 本文($\omega=0$) | 58.3 | 54.7 | 26.7 |
| 本文($\omega=1\ 000$) | 58.5 | 55.2 | 26.8 |
| 本文($\omega=500$) | 58.6 | 55.2 | 26.2 |
| 本文($\omega=100$) | 58.5 | 55.7 | 26 |

**表4 OTB100数据集测试结果**

**Table 4 Results of testing on OTB100 dataset**

| 模型 | 跟踪成功率/% | 精确度/% |
|---|---|---|
| DaSiamese-RPN | 64.6 | 85.9 |
| 本文($\omega=0$) | 64.9 | 86.1 |
| 本文($\omega=1\ 000$) | 64.5 | 85.5 |
| 本文($\omega=500$) | 64.6 | 85.8 |
| 本文($\omega=100$) | 64.8 | 86 |

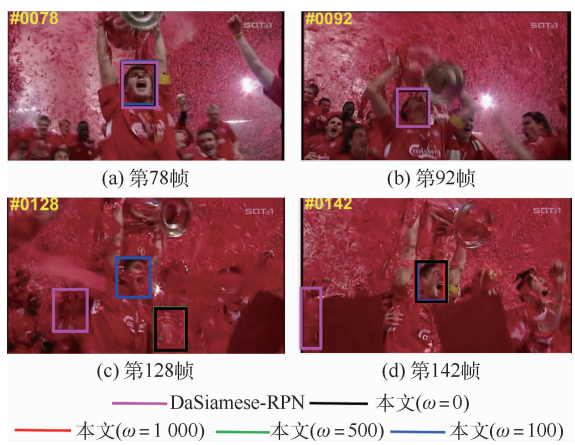
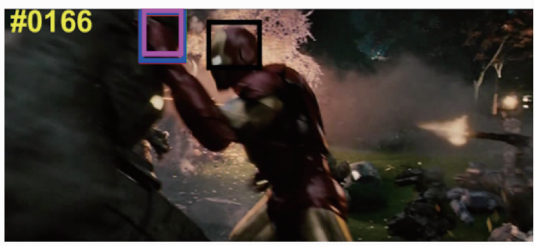



图8 遮挡场景下可视化

Fig. 8 Visualization of occlusion scene


(a) 第89帧


(b) 第118帧




(c) 第166帧

DaSiamese-RPN　本文($\omega$=0)　本文($\omega$=1 000)　本文($\omega$=500)　本文($\omega$=100)

图 9　动作复杂变化下可视化

Fig. 9　Visualization under complex changes of movement

### 3.4　与其他方法的对比

将外观动作自适应目标跟踪方法与前沿跟踪方法在 VOT2016、VOT2018 及 OTB100 数据集进行了效果对比,如表 5 ~ 表 7 所示。

**表 5　不同方法在 VOT2016 数据集上的对比**

**Table 5　Comparison with different methods on VOT2016 dataset**

| 方法 | 预测平均重叠率/% | 准确度/% | 鲁棒性/% |
|---|---|---|---|
| MemTrack[19] | 27.3 | 53.3 | 144.1 |
| SiamFC[11] | 23.5 | 52.9 | 190.8 |
| SiamRPN[12] | 26.2 | 53.8 | 42.4 |
| SiamRPNpp[2] | 39.3 | 61.8 | 23.8 |
| 本文 | 45.6 | 61.4 | 18.6 |

**表 6　不同方法在 VOT2018 数据集上的对比**

**Table 6　Comparison with different methods on VOT2018 dataset**

| 方法 | 预测平均重叠率/% | 准确度/% | 鲁棒性/% |
|---|---|---|---|
| DRT[20] | 35.5 | 51.8 | 20.1 |
| RCO[16] | 37.6 | 50.7 | 15.5 |
| UPDT[21] | 37.9 | 53.6 | 18.4 |
| SiamRPN[12] | 38.4 | 50.5 | 14 |
| MFT[16] | 38.6 | 50.3 | 15.9 |
| LADCF[22] | 38.9 | 50.3 | 15.9 |
| SiamRPNpp[2] | 35.2 | 57.6 | 27 |
| 本文 | 38.7 | 58.5 | 25.8 |

**表 7　不同方法在 OTB100 数据集上的对比**

**Table 7　Comparison with different methods on OTB100 dataset**

| 方法 | 跟踪成功率/% | 精确度/% |
|---|---|---|
| SiamFC[11] | 58.9 | 79.4 |
| GradNet[23] | 63.9 | 86.1 |
| C-RPN[24] | 63.9 | 85.2 |
| SiamRPN[12] | 63.7 | 85.1 |
| SiamRPNpp[2] | 64.8 | 85.3 |
| FENG[25] | 61 | 73 |
| SNLT[26] | 67 | 80 |
| 本文 | 64.9 | 86.1 |

比较不同方法与外观动作自适应跟踪方法的性能可以看出,本文方法表现出更好的性能,充分解决了跟踪目标复杂动作带来的干扰问题,具有一定的有效性。

## 4　结　论

为解决目标复杂运动时产生的外观形变对目标跟踪的影响,本文在 DaSiamese-RPN 基础上进行改进,提出了一种外观动作自适应的目标跟踪方法。

1) 在孪生网络的特征提取子网络中引入外观动作自适应更新模块,采用多激励通道融合目标的时空信息和动作特征,降低目标动作变化带来的干扰。

2) 利用 2 种欧氏距离分别度量真实图和预测图之间全局和局部性差异,并将两者加权联合来构建损失函数,加强预测目标特征图与真实目标特征图之间全局和局部信息的关联性。

3) 在 VOT2016、VOT2018 和 OTB100 数据集上进行测试,结果均优于原基线方法,表明本文方法能够更好地适应目标的复杂动作变化,跟踪的准确度和鲁棒性都有所提升。

下一步的研究将尝试将外观动作自适应更新模块与其他跟踪方法结合,在提高跟踪速度和性能方面进行研究,着重解决遮挡等外部环境干扰带来的问题。

### 参考文献 (References)

[1] 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 (ECCV). Berlin: Springer, 2018: 101-117.

[2] LI B, WU W, WANG Q, et al. Evolution of Siamese visual tracking with very deep networks[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2019: 16-20.

[3] WANG Q, ZHANG L, BERTINETTO L, et al. Fast online object tracking and segmentation: A unifying approach[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2019: 1328-1338.

[4] ZHANG Z P, PENG H W. Deeper and wider Siamese networks for real-time visual tracking[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2019: 4591-4600.

[5] 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.

[6] XIE S N, GIRSHICK R, DOLLAR P, et al. Aggregated residual transformations for deep neural networks[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2017: 1492-1500.

[7] HOWARD A G, ZHU M, CHEN B, et al. MobileNets: Efficient convolutional neural networks for mobile vision applications [EB/OL]. (2017-04-17) [2021-10-01]. https://arxiv.org/abs/1704.04861.

[8] VOIGTLAENDER P, LUITEN J, TORR P H S, et al. Siam R-CNN: Visual tracking by re-detection[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2020: 6578-6588.

[9] ZHANG L, GONZALEZ-GARCIA A, WEIJER J, et al. Learning the model update for Siamese trackers[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2019: 4010-4019.

[10] WANG Z, SHE Q, SMOLIC A. ACTION-Net: Multipath excitation for action recognition[EB/OL]. (2021-03-11) [2021-10-01]. https://arxiv.org/abs/2103.07372.

[11] BERTINETTO L, VALMADRE J, HENRIQUES J F, et al. Fully-convolutional Siamese networks for object tracking[C]// Proceedings of the European Conference on Computer Vision (ECCV). Berlin: Springer, 2016: 850-865.

[12] TAO R, GAVVES E, SMEULDERS A W M. Siamese instance search for tracking[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2016: 1420-1429.

[13] FAN H, LIN L, YANG F, et al. LaSOT: A high-quality benchmark for large-scale single object tracking[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2019: 5374-5383.

[14] KRISTAN M, MATAS J, LEONARDIS A, et al. A novel performance evaluation methodology for single-target trackers[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2016, 38(11): 2137-2155.

[15] KRISTAN M, LEONARDIS A, MATAS J, et al. The visual object tracking VOT2016 challenge results[C]// Proceedings of the European Conference on Computer Vision (ECCV). Berlin: Springer, 2016: 2-5.

[16] KRISTAN M, LEONARDIS A, MATAS J, et al. The sixth visual object tracking VOT2018 challenge results[C]// Proceedings of the European Conference on Computer Vision (ECCV). Berlin: Springer, 2018: 3-53.

[17] KRISTAN M, MATAS J, LEONARDIS A, et al. The seventh visual object tracking VOT2019 challenge results[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2019.

[18] WU Y, LIM J, YANG M H. Object tracking benchmark[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015, 37(9): 1834-1848.

[19] YANG T, CHAN A B. Learning dynamic memory networks for object tracking[C]// Proceedings of the European Conference on Computer Vision (ECCV). Berlin: Springer, 2018: 152-167.

[20] SUN C, WANG D, LU H, et al. Correlation tracking via joint discrimination and reliability learning[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2018: 489-497.

[21] BHAT G, JOHNANDER J, DANELLJAN M, et al. Unveiling the power of deep tracking[C]// Proceedings of the European Conference on Computer Vision (ECCV). Berlin: Springer, 2018: 483-498.

[22] XU T Y, FENG Z H, WU X J, et al. Learning adaptive discriminative correlation filters via temporal consistency preserving spatial feature selection for robust visual object tracking[J]. IEEE Transactions on Image Processing, 2019, 28(11): 5596-5609.

[23] LI P, CHEN B, OUYANG W, et al. GradNet: Gradient-guided network for visual object tracking[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2019: 6162-6171.

[24] FAN H, LING H. Siamese cascaded region proposal networks for real-time visual tracking[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2019: 7952-7961.

[25] FENG Q, ABLAVSKY V, BAI Q X, et al. Real-time visual object tracking with natural language description[C]// 2020 IEEE Winter Conference on Applications of Computer Vision (WACV). Piscataway: IEEE Press, 2020: 700-709.

[26] FENG Q, ABLAVSKY V, BAI Q, et al. Siamese natural language tracker: Tracking by natural language descriptions with Siamese trackers[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2021: 5851-5860.

# Appearance and action adaptive target tracking method

XIONG Junyao, WANG Rong*, SUN Yibo

(School of Information Technology and Network Security, People's Public Security University of China, Beijing 100038, China)

**Abstract**: On the basis of DaSiamese-RPN, a target tracking approach of appearance and action adaptation is proposed to limit the effect of appearance deformation on target tracking when the target is moving. First of all, the appearance and action adaptive module is introduced in the subnet of the Siamese network, which integrates the object's spatial information and action feature. Secondly, the global and local divergence between the actual and predicted feature maps are measured by using two Euclidean distances, and the loss function is constructed by weighting the fusion of the two, so as to strengthen the correlation between the global and local information. Finally, tests were conducted on the VOT2016, VOT2018, VOT2019, and OTB100 datasets. The experimental results showed that the expected average overlap was improved by 4.5% and 6.1% in the VOT2016 and VOT2018 datasets respectively. On the VOT2019 dataset, accuracy increased by 0.4% and expected average overlap decreased by 1%; The tracking success rate was improved by 0.3% and accuracy increased by 0.2% when evaluated on the OTB100 dataset.

**Keywords**: target tracking; appearance and action adaptive; Siamese network; feature integration; appearance deformation

**Received**: 2021-10-09; **Accepted**: 2021-10-29; **Published online**: 2021-11-16 09:52

**URL**: kns.cnki.net/kcms/detail/11.2625.V.20211115.1929.001.html

**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.1001-5965.2021.0600

# 基于多标签协同学习的跨域行人重识别

李慧, 张晓伟*, 赵新鹏, 路昕雨
(青岛大学 计算机科学技术学院, 青岛 266071)

**摘 要**: 跨域是行人重识别的重要应用场景,但是源域与目标域行人图像在光照条件、拍摄视角、成像背景与风格等方面的表观特征差异性是导致行人重识别模型泛化能力下降的关键因素。针对该问题,提出了基于多标签协同学习的跨域行人重识别方法。利用语义解析模型构造了基于语义对齐的多标签数据表示,以引导构建更关注行人前景区域的局部特征,达到语义对齐的目的,减少背景对跨域重识别的影响。基于行人图像全局特征和语义对齐后的行人局部特征,利用协同学习平均模型生成行人重识别模型的多标签表示,减少跨域场景下噪声硬标签的干扰。利用协同学习网络框架联合多标签的语义对齐模型,提高行人重识别模型的识别能力。实验结果表明:在 Market-1501→DukeMTMC-reID、DukeMTMC-reID→Market-1501、Market-1501→MSMT17、DukeMTMC-reID→MSMT17 跨域行人重识别数据集上,与 NRMT 方法相比,平均精度均值分别提高了 8.3%、8.9%、7.6%、7.9%,多标签协同学习方法具有显著的优越性。

**关 键 词**: 跨域行人重识别; 语义对齐; 全局特征; 多标签表示; 协同学习
**中图分类号**: TP37; TP277
**文献标志码**: A **文章编号**: 1001-5965(2022)08-1534-09

在智能视频监控中,行人是智能监控视频分析中的主体,而跨域行人重识别是公共安全领域智能化的重要挑战技术之一。跨域行人重识别旨在弥补单一、固定摄像头的视觉局限,识别跨摄像机设备下的同一标识的行人图像,被广泛应用于智能视频监控、智能安防等领域,具有重要的研究意义和应用价值。

跨域行人重识别的目的是预测来自不同相机的 2 幅图像是否属于同一个人。单域行人重识别已经取得巨大成功,而源域数据集和目标域数据集之间存在的数据偏差和场景差异,导致跨域行人重识别方法将源域训练出的行人重识别模型用于目标域会产生显著的性能下降[1-2],尤其是在光照变化、背景杂乱、人体姿势变化和摄像机角度变化的影响下,同一 ID 标识的行人在不同视图之间的特征辨识度明显降低。为此,跨域行人重识别技术从基于深度卷积神经网络的全局特征开始向更细粒度的局部特征方向发展。根据对行人局部区域划分方式的不同,基于行人语义部件的重识别模型可以分为物理硬划分和语义软划分 2 种。物理硬划分不需要部件标签,较为简单,如 Sun 等[3]使用 PCB 模型将行人图像均匀划分为 $p$ 个水平带来提取行人局部特征。虽然不需要额外的部件监督信息,但是基于人体区域的硬划分过于粗糙,对于行人姿态多变、空间分布差异突出及遮挡严重等情形下跨域行人重识别性能下降的问题没有解决。

基于语义软划分的行人重识别模型对上述问题取得了一定的效果,如通过姿态估计的不同阶段提取不同语义层次的特征,并通过合并不同语

**收稿日期**: 2021-10-10; **录用日期**: 2021-10-29; **网络出版时间**: 2021-11-12 08:58
**网络出版地址**: kns.cnki.net/kcms/detail/11.2625.V.20211110.1924.004.html
**基金项目**: 国家自然科学基金(61902204); 山东省自然科学基金(ZR2019BF028)
* **通信作者**. E-mail: by1306114@buaa.edu.cn

**引用格式**: *李慧, 张晓伟, 赵新鹏, 等. 基于多标签协同学习的跨域行人重识别[J]. 北京航空航天大学学报, 2022, 48(8): 1534-1542. LI H, ZHANG X W, ZHAO X P, et al. Multi-label cooperative learning for cross domain person re-identification[J]. Journal of Beijing University of Aeronautics and Astronautics, 2022, 48(8): 1534-1542 (in Chinese).*

义层次的特征,来对齐不同图像中的人体部件区域,增强局部细节信息的表示能力[4]。Kalayeh 等[5]提出集成人体解析模型到行人重识别模型中,通过解析人体部件模型获取各部件概率图,从而得到更精细的部件特征。但是,基于语义软划分的行人重识别模型过于依赖人体语义的划分结果,从而制约着跨域行人重识别方法的精度。

近年来,基于无监督域自适应的风格迁移[1]、属性识别[6]和目标域标签估计[7-8]的方法已被证明有助于缓解跨域行人重识别的域迁移问题。例如,Zhong 等[9]从风格迁移的方向入手,利用相机不变性来缓解跨域行人重识别的跨域偏移问题。在属性识别方面,Lin 等[10]将属性识别与行人重识别的全局特征结合起来,构造一个全新的损失函数,来提高行人重识别模型的性能。在基于聚类生成的目标域伪标签方面,Fan 等[2]提出自步学习的方法,利用无监督聚类获得伪标签来训练网络,通过学习获得可靠的目标域标签估计。

虽然上述基于无监督的跨域行人重识别方法在很大程度上提升了行人重识别模型的性能,但是仍然远远落后于有监督的单域行人重识别方法,主要原因是:无监督学习易受背景噪声的影响,且缺乏标签的监督信息。为生成可靠的目标域标签,Wang 和 Zhang[11]将无监督行人重识别转化为多标签分类任务,利用基于记忆的非参数分类器,将多标签分类和单标签分类联合在一个框架中,逐步寻找真实可靠的标签,但是这种多标签分类过于依赖分类器得分,缺乏动态适应性。为此,本文提出了一种多标签协同学习(multi-label collaborative learning,MCL)模型,通过 2 个无监督网络的平均模型进行交替监督,并且利用语义对齐模块(semantic alignment module,SAM)生成的多标签数据,减小背景噪声对跨域行人重识别任务的影响,提高跨域行人重识别模型的泛化能力。

综上所述,本文创新如下:

1)针对跨域行人重识别的背景噪声干扰和行人图像硬划分带来的语义不对齐问题,提出一种基于行人语义解析的语义对齐模块 SAM,其通过人体解析模型对齐行人语义特征,减小背景噪声对于识别任务的影响,也提高了模型的泛化能力。

2)为解决跨域行人重识别中目标域硬伪标签存在的噪声干扰问题,对全局特征和语义对齐后的行人前景特征分别进行软监督生成多软标签表示,减少噪声的影响。

3)为提升跨域行人重识别在未标记的目标域内的性能,本文提出了一个多标签协同学习模型 MCL,其基于协同学习框架利用语义对齐模块 SAM 联合训练多个标签来优化行人重识别模型,提高了跨域行人重识别方法的性能。

4)在 Market-1501→DukeMTMC-reID、DukeMTMC-reID→Market-1501、Market-1501→MSMT17、DukeMTMC-reID→MSMT17 跨域行人重识别数据集上的实验结果,与当前先进的跨域行人重识别方法 NRMT 相比,本文方法的平均精度均值(mAP)分别提高了 8.3%、8.9%、7.6%、7.9%。

## 1　相关工作

### 1.1　行人语义对齐

由于行人姿态变化和空间失配问题,导致提取局部特征时往往出现行人语义特征不对齐的问题,一些研究试图通过硬划分方法来提取语义特征。例如,Zhang 等[12]使用硬分割的方式提取行人局部特征,为了解决硬划分带来的局部特征空间距离不对齐的问题,采用动态规划法寻求最短路径,达到对齐局部特征的目的,并采用相互学习的框架来增强模型的效果,但是这种硬划分对于姿态变化不敏感,误差较大。由此,许多软划分方法被提出对齐行人局部特征。例如,Zheng 等[13]针对行人重识别中的行人错位问题提出了 PIE 框架,融合了 PoseBox 和置信度来纠正由摄像机视角、人员运动和探测器错误引起的姿势变化,并实现了基于部件对齐的行人匹配,但框架过于复杂,且姿态估计模型和行人重识别数据集之间的标注存在很大的偏差。Zhu 等[14]提出由身份引导的人体语义解析方法来实现部件对齐的目的,并设计了基于特征图的级联聚类方法生成人体部位的伪标签,减小人体解析模型带来的误差。Guo 等[15]针对背景中的行人附属物提出了使用自我注意力机制来关注非人体部分,结合人体解析提取人体部件特征获得更好的行人特征表示。以上方法都是针对有标签的单域行人重识别,并且语义分析的精度严重影响了重识别的性能,对于跨域行人重识别的语义对齐问题还缺乏深入的研究。

### 1.2　多标签学习

近年来,无监督行人重识别由于其潜在地解决了跨域行人重识别模型的可扩展性问题而引起了越来越多的关注和研究,但是如何训练一个有

效的伪标签是极具挑战性的问题。HCT[16]利用层次聚类的方法获得了硬伪标签监督模型的训练,但是标签存在很大的噪声,且过于单一。互均值教学[17]为无监督域自适应行人重识别设计了一个对称框架,包含硬伪标签和软标签,虽然很大程度上抑制了聚类噪声的影响,但是只从全局特征考虑,忽略了局部特征。Yu 等[18]提出了学习一个软多标签(即一个实值标签似然向量而不是单一的伪标签)来监督无监督模型,虽然考虑到无监督图像与多个图像的相似度,从而得到软标签,但是图像外观存在很大的相似性,往往带来较大的误差。现有的标签学习往往只考虑单一的全局标签,忽略了行人重识别图像的复杂性。

## 2 多标签协同学习

### 2.1 多标签协同学习模型架构

本文提出的多标签协同学习模型 MCL 以 MMT[17]为基础框架,如图 1 所示。利用协同学习的思想,训练 2 个具有不同初始化的相同网络 Net1 和 Net2 来协同学习,同时利用 2 个网络的平均模型 Mean Net 生成的实时软标签来监督网络的训练,减少噪声的干扰。为了同时训练这 2 个网络,将相同的图像进行批处理后,对 2 个网络同时进行训练,但是分别对图像进行随机擦除[19]、裁剪和翻转。

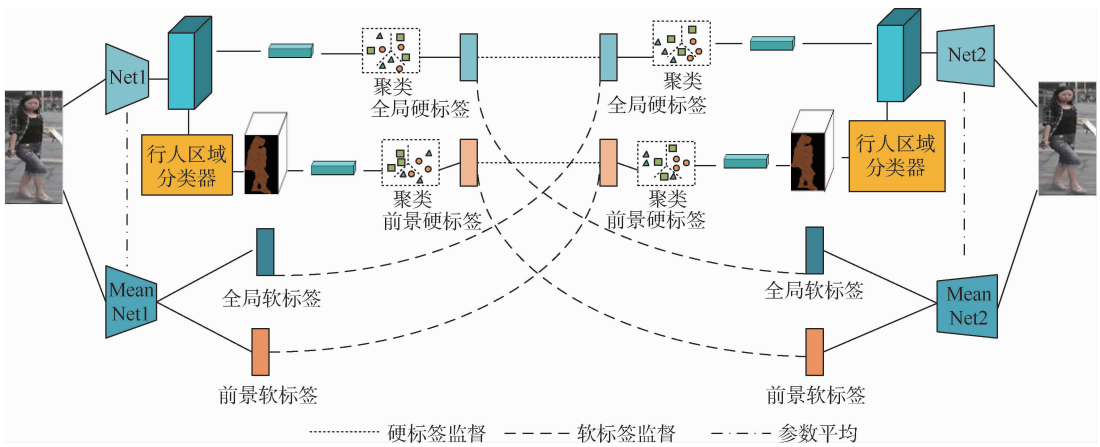



图 1 多标签协同学习跨域行人重识别框架

Fig. 1 Framework of cross-domain person re-identification on multi-label collaborative learning

由于 MMT[17]只通过全局特征来监督模型的训练,容易受到背景噪声的干扰,而且只通过全局软标签进行监督,标签过于单一。因此,本文提出了多标签协同框架 MCL,不仅通过网络提取全局特征,还利用语义对齐模块 SAM 通过解析模型提取前景特征进行池化后分别聚类,获得全局和前景特征的硬伪标签,减少了背景噪声的干扰。同时,利用2 个平均模型分别对全局特征和前景特征进行分类预测,得到全局和前景的多个软标签数据来监督模型的训练。

本文通过交叉熵损失和三元组损失来优化多标签协同学习模型 MCL,其中,语义对齐模块 SAM 通过提取前景特征来避免背景的干扰,同时通过全局和前景的多个软标签数据来减少硬标签噪声的影响,优化跨域行人重识别方法。

### 2.2 行人语义对齐模块

通过主干网络 ResNet50[20]提取全局特征图,定义主干映射函数为 $f_\theta$,对于行人重识别数据集 $x_i$,全局特征映射 $\boldsymbol{M}_g$ 如下:

$$\boldsymbol{M}_g^{c\times h\times w} = f_\theta(x_i) \tag{1}$$

式中:$\theta$ 为主干参数;$c$ 为通道数;$h$、$w$ 为分辨率。

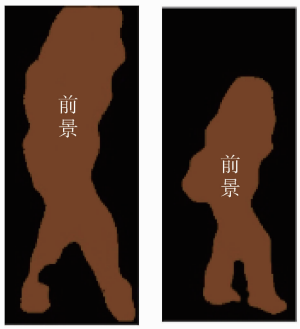
由于直接硬划分会导致语义特征不对齐问题,如图 2(a)所示,SSG[21]采用将全局特征均匀水平划分为两部分的方法,分别提取上半身、下半身两部分的局部特征,这种硬划分方法在行人图像不对齐时会出现严重的特征错位问题;本文采用对行人姿态变化具有更强适应能力的语义解析方法提取像素级行人特征,如图 2(b)所示,通过提取行人区域特征,对齐图像像素级的行人区域,同时也剔除了背景的干扰,使模型更关注行人区域。

利用语义解析网络 PSPNet[22]训练行人区域分类器,通过分类器将行人重识别数据集中的行人图像划分前景和背景区域进行训练学习,对每个像素点$(x,y)$ 使用线性分类器 softmax 来区分行人区域,表达式为

$$P(x,y) = \mathrm{softmax}(\boldsymbol{W}^{\mathrm{T}}\boldsymbol{M}_g(x,y)) = \frac{\exp(\boldsymbol{W}^{\mathrm{T}}\boldsymbol{M}_g(x,y))}{\sum_{i=1}^{P}\exp(\boldsymbol{W}_i^{\mathrm{T}}\boldsymbol{M}_g(x,y))} \tag{2}$$

式中:$P(x,y)$为属于行人区域的预测概率;$\boldsymbol{W}^{\mathrm{T}}$ 为


(a) 行人硬划分




(b) 行人语义对齐

图 2　硬划分与语义对齐对比效果

Fig. 2　Comparison effect of hard division and semantic alignment

行人区域分类器的可训练权重矩阵。

对于行人重识别数据集 $x_i$,通过行人区域分类器获得行人数据集的前景概率图,用 $P_k(x,y)$ 表示属于前景部分 $k$ 的像素 $(x,y)$ 的预测概率,利用全局特征图提取前景部分的特征图,如下:

$$\boldsymbol{M}_{\mathrm{f}} = P_k \circ \boldsymbol{M}_{\mathrm{g}} \tag{3}$$

式中:"∘"指点乘运算;$\boldsymbol{M}_{\mathrm{f}}$ 为前景特征图。

通过前景特征表示引导模型关注行人区域,减少背景的干扰,达到语义对齐的目的。

### 2.3　基于全局特征和语义前景特征的多标签表示

通过无监督聚类产生的硬标签监督模型的训练如图 3 所示。分别对全局特征 $\boldsymbol{M}_{\mathrm{g}}$ 和前景特征 $\boldsymbol{M}_{\mathrm{f}}$ 进行全局平均池化和前景平均池化,通过无监督聚类的方式获得全局硬标签 $D_{\mathrm{g}}^{\mathrm{h}} = \{(x_i,\tilde{y}_{\mathrm{g}}^{\mathrm{h}}) \mid_{i=1}^{N_{\mathrm{t}}}\}$ 和前景硬标签 $D_{\mathrm{f}}^{\mathrm{h}} = \{(x_i,\tilde{y}_{\mathrm{f}}^{\mathrm{h}}) \mid_{i=1}^{N_{\mathrm{t}}}\}$ 来监督模型的训练。其中,$(x_i,\tilde{y}_{\mathrm{g}}^{\mathrm{h}})$ 和 $(x_i,\tilde{y}_{\mathrm{f}}^{\mathrm{h}})$ 分别表示第 $i$ 个图像通过无监督聚类产生的全局硬标签和前景硬标签,$N_{\mathrm{t}}$ 为目标域图像的个数。

基于聚类产生的硬标签会不可避免地存在聚类错误的噪声数据,因此通过平均模型 Mean Net 生成的软标签来减小硬标签的噪声影响。如图 4 所示,通过平均模型 Mean Net 对目标域图像的全局特征 $\boldsymbol{M}_{\mathrm{g}}$ 进行分类预测,得到全局特征的软标签数据 $D_{\mathrm{g}}^{\mathrm{s}} = \{(x_i,\tilde{y}_{\mathrm{g}}^{\mathrm{s}}) \mid_{i=1}^{N_{\mathrm{t}}}\}$,$(x_i,\tilde{y}_{\mathrm{g}}^{\mathrm{s}})$ 表示第 $i$ 个图像通过平均模型分类预测的全局软标签。

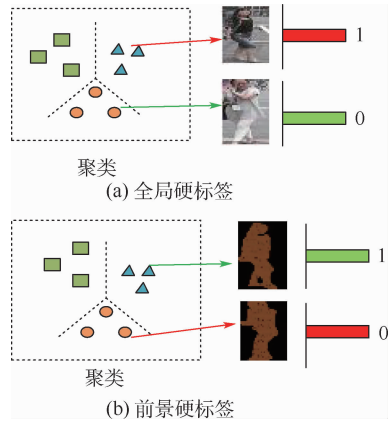



图 3　全局硬标签与前景硬标签示意图

Fig. 3　Schematic diagram of global hard label and foreground hard label

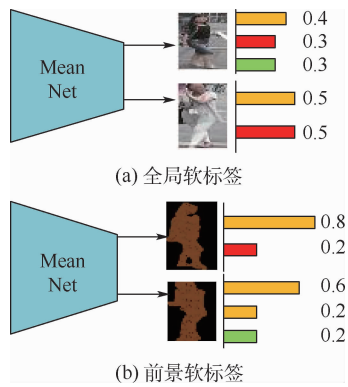



图 4　全局软标签与前景软标签示意图

Fig. 4　Schematic diagram of global soft label and foreground soft label

为了减少背景噪声的影响,利用语义对齐模块 SAM 训练的行人区域分类器提取行人图像前景概率图,通过式(3)提取行人前景特征 $\boldsymbol{M}_{\mathrm{f}}$,利用平均模型 Mean Net 对目标域图像的前景特征 $\boldsymbol{M}_{\mathrm{f}}$ 进行分类预测,获得前景特征的软标签数据 $D_{\mathrm{f}}^{\mathrm{s}} = \{(x_i,\tilde{y}_{\mathrm{f}}^{\mathrm{s}}) \mid_{i=1}^{N_{\mathrm{t}}}\}$,$(x_i,\tilde{y}_{\mathrm{f}}^{\mathrm{s}})$ 表示第 $i$ 个图像通过平均模型分类预测的前景软标签。

分别联合全局多标签数据 $D_{\mathrm{g}} = \{x_i,(\tilde{y}_{\mathrm{g}}^{\mathrm{h}},\tilde{y}_{\mathrm{g}}^{\mathrm{s}}); 1 \leqslant i \leqslant N_{\mathrm{t}}\}$ 和前景多标签数据 $D_{\mathrm{f}} = \{x_i,(\tilde{y}_{\mathrm{f}}^{\mathrm{h}},\tilde{y}_{\mathrm{f}}^{\mathrm{s}}); 1 \leqslant i \leqslant N_{\mathrm{t}}\}$ 来监督行人重识别模型的训练,减小了噪声数据的影响和背景干扰,提高了模型的识别能力。多标签数据可表示为

$$D = \{x_i,(\tilde{y}_{\mathrm{g}}^{\mathrm{h}},\tilde{y}_{\mathrm{g}}^{\mathrm{s}},\tilde{y}_{\mathrm{f}}^{\mathrm{h}},\tilde{y}_{\mathrm{f}}^{\mathrm{s}}); 1 \leqslant i \leqslant N_{\mathrm{t}}\} \tag{4}$$

### 2.4　损失函数

利用交叉熵损失和三元组损失训练网络,通过联合训练全局特征和前景特征的多标签表示优化多标签协同学习模型 MCL。

目标域图像用 $x_i^{\mathrm{t}}$ 来表示,用特征变换函数 $F(\cdot \mid \theta_1)$ 和 $F(\cdot \mid \theta_2)$ 表示 2 个协同网络,伪标签用 $C^{\mathrm{t}}(F(x_i^{\mathrm{t}} \mid \theta)$ 表示。首先,通过无监督聚类获得 2 个协同网络的全局硬标签(global hard la-

bel,GHL)和前景硬标签(foreground hard label,FHL);然后,利用平均模型 Mean Net 对全局和前景特征进行分类预测,获得全局软标签(global soft label,GSL)和前景软标签(foreground soft label,FSL);最后,联合全局和前景特征多标签数据监督模型的训练。

交叉熵损失 $L_{\mathrm{id}}^{\mathrm{t}}(\theta)$ 和三元组损失 $L_{\mathrm{tri}}^{\mathrm{t}}(\theta)$ 表达式如下:

$$L_{\mathrm{id}}^{\mathrm{t}}(\theta) = \frac{1}{N_{\mathrm{t}}}\sum_{i=1}^{N_{\mathrm{t}}} L_{\mathrm{ce}}(C^{\mathrm{t}}(F(x_i^{\mathrm{t}}\mid\theta)),\tilde{y}_i^{\mathrm{t}}) \tag{5}$$

$$L_{\mathrm{tri}}^{\mathrm{t}}(\theta) = \frac{1}{N_{\mathrm{t}}}\sum_{i=1}^{N_{\mathrm{t}}} \max(0,\|F(x_i^{\mathrm{t}}\mid\theta) - F(x_{i,\mathrm{p}}^{\mathrm{t}})\| + m - \|F(x_i^{\mathrm{t}}\mid\theta) - F(x_{i,\mathrm{n}}^{\mathrm{t}}\mid\theta)\|) \tag{6}$$

式中:$L_{\mathrm{ce}}$ 为交叉熵损失函数;$\tilde{y}_i^{\mathrm{t}}$ 为对应 $x_i^{\mathrm{t}}$ 的伪标签;$\|\cdot\|$ 表示 $L_2$ 范数距离;$x_{i,\mathrm{p}}^{\mathrm{t}}$、$x_{i,\mathrm{n}}^{\mathrm{t}}$ 分别为每个小批次中最难正、负样本;$m$ 为三元组损失距离边缘,用于控制正负样本的距离。通过实验发现,当 $m=0.5$ 时,训练效果最好。

通过式(5)和式(6)计算全局多标签数据 $D_{\mathrm{g}}$ 的损失函数如下:

$$\begin{aligned}L_{\mathrm{g}}(\theta_1,\theta_2) = {} & (1-\lambda_1)(L_{\mathrm{id\_h}}^{\mathrm{g}}(\theta_1)+L_{\mathrm{id\_h}}^{\mathrm{g}}(\theta_2)) + \\ & \lambda_1(L_{\mathrm{sid\_s}}^{\mathrm{g}}(\theta_1\mid\theta_2)+L_{\mathrm{sid\_s}}^{\mathrm{g}}(\theta_2\mid\theta_1)) + \\ & (1-\lambda_2)(L_{\mathrm{tri\_h}}^{\mathrm{g}}(\theta_1)+L_{\mathrm{tri\_h}}^{\mathrm{g}}(\theta_2)) + \\ & \lambda_2(L_{\mathrm{stri\_s}}^{\mathrm{g}}(\theta_1\mid\theta_2)+L_{\mathrm{stri\_s}}^{\mathrm{g}}(\theta_2\mid\theta_1))\end{aligned} \tag{7}$$

式中:$\lambda_1$、$\lambda_2$ 为权重参数;$L_{\mathrm{id\_h}}^{\mathrm{g}}$ 和 $L_{\mathrm{tri\_h}}^{\mathrm{g}}$ 分别为全局硬标签交叉熵损失和三元组损失;$L_{\mathrm{sid\_s}}^{\mathrm{g}}$ 和 $L_{\mathrm{stri\_s}}^{\mathrm{g}}$ 分别为全局软标签交叉熵损失和三元组损失。

前景多标签数据 $D_{\mathrm{f}}$ 损失函数为

$$\begin{aligned}L_{\mathrm{f}}(\theta_1,\theta_2) = {} & (1-\lambda_3)(L_{\mathrm{id\_h}}^{\mathrm{f}}(\theta_1)+L_{\mathrm{id\_h}}^{\mathrm{f}}(\theta_2)) + \\ & \lambda_3(L_{\mathrm{sid\_s}}^{\mathrm{f}}(\theta_1\mid\theta_2)+L_{\mathrm{sid\_s}}^{\mathrm{f}}(\theta_2\mid\theta_1)) + \\ & (1-\lambda_4)(L_{\mathrm{tri\_h}}^{\mathrm{f}}(\theta_1)+L_{\mathrm{tri\_h}}^{\mathrm{f}}(\theta_2)) + \\ & \lambda_4(L_{\mathrm{stri\_s}}^{\mathrm{f}}(\theta_1\mid\theta_2)+L_{\mathrm{stri\_s}}^{\mathrm{f}}(\theta_2\mid\theta_1))\end{aligned} \tag{8}$$

式中:$\lambda_3$、$\lambda_4$ 为权重参数;$L_{\mathrm{id\_h}}^{\mathrm{f}}$ 和 $L_{\mathrm{tri\_h}}^{\mathrm{f}}$ 分别为前景硬标签交叉熵损失和三元组损失;$L_{\mathrm{sid\_s}}^{\mathrm{f}}$ 和 $L_{\mathrm{stri\_s}}^{\mathrm{f}}$ 分别为前景软标签交叉熵损失和三元组损失。

联合训练多标签数据 $D$,则本文总损失函数为

$$L = \lambda_5 L_{\mathrm{g}}(\theta_1,\theta_2) + (1-\lambda_5)L_{\mathrm{f}}(\theta_1,\theta_2) \tag{9}$$

式中:$L_{\mathrm{g}}$ 和 $L_{\mathrm{f}}$ 分别为全局和前景多标签损失;$\lambda_5$ 为权重参数。

# 3 实验结果分析

## 3.1 数据集及实验设置

本节在国际公开的行人重识别基准数据集 Market-1501[23]、DukeMTMC-reID[24] 和 MSMT17[25] 上评估了本文方法。在 ImageNet[26] 上训练的 ResNet50[20] 作为骨干网,输入图像的大小调整为 256×128。语义解析模型使用 LIP 数据集来训练,再通过图像大小为 256×128 的行人重识别数据集对语义解析网络进行微调。

以预先训练的权重 $\theta_1$ 和 $\theta_2$ 为基础,优化 $L$,损失权重 $\lambda_1=0.5$,$\lambda_2=0.8$,$\lambda_3=0.5$,$\lambda_4=0.8$,$\lambda_5=0.6$。学习速率固定在 0.000 35,共训练 40 个epoch。聚类算法采用 K-Means 聚类算法,对 Market-1501、DukeMTMC-reID、MSMT17 数据集分别设置 500、700、900 个伪标签类别的 $M_{\mathrm{t}}$。采用 Adam 优化器对网络进行优化,权值衰减为 0.000 5。在测试时,采用欧氏距离度量查询图像与图库图像的相似度,且没有使用重排序[27] 等方法。同时使用 mAP 与 Rank-$k$ 精度来评估本文方法,R-1、R-5、R-10 分别表示第 1 张、前 5 张、前 10 张结果的命中率。

## 3.2 消融实验

为了验证各模块的有效性,本文在 Market-1501 和 DukeMTMC-reID 数据集上开展了验证实验。

### 3.2.1 语义对齐模块有效性验证

为了验证语义对齐模块 SAM 的有效性,在基于聚类的硬标签模型上进行了实验。将未使用语义对齐的全局特征的模型作为基线,记为全局特征模型 GFM,硬划分模型记为 HPM,加入语义对齐模块 SAM 的软划分模型记为 SPM。如表 1 所示,在 DukeMTMC-reID→Market-1501 的跨域行人重识别实验中,软划分模型 SPM 较全局特征模型 GFM,mAP 提高了 17.2%,R-1 提高了 12.1%,较硬划分模型 HPM,mAP 提高了 2.2%,R-1 提高了 1.7%。实验证明,加入语义对齐模块 SAM 的软划分模型具有显著的优越性。同样的实验结论,在 Market-1501→DukeMTMC-reID 数据集跨域行人重识别实验中得到了验证。

**表 1 在 DukeMTMC-reID 和 Market-1501 数据集上的语义对齐模块有效性消融实验**

**Table 1 Ablation study for semantic alignment module validity on DukeMTMC-reID dataset and Market-1501 dataset**

| 方法 | Market-1501→DukeMTMC-reID | | | | DukeMTMC-reID→Market-1501 | | | |
|---|---|---|---|---|---|---|---|---|
| | R-1/% | R-5/% | R-10/% | mAP/% | R-1/% | R-5/% | R-10/% | mAP/% |
| GFM | 68.4 | 80.1 | 83.5 | 49.0 | 75.8 | 89.5 | 93.2 | 53.7 |
| HPM | 76.0 | 85.8 | **89.3** | 60.3 | 86.2 | 94.6 | 96.5 | 68.7 |
| SPM | **76.7** | **85.9** | 89.0 | **60.9** | **87.9** | **95.2** | **96.8** | **70.9** |

注:黑体数据为每列最优值。

#### 3.2.2　联合全局和语义前景特征的多标签有效性验证

为了验证联合全局和语义前景特征的多标签有效性,在 Market-1501、DukeMTMC-reID 数据集上进行了实验。将只用全局特征多标签的模型记为 GSL,将只用前景特征多标签的模型记为 FSL,与本文联合全局和前景特征的多标签协同学习模型 MCL (GSL + FSL)进行了对比。从表 2 可以看出,在 DukeMTMC-reID→Market-1501 数据集跨域行人重识别实验中,联合全局和前景特征的多标签协同学习模型 MCL (GSL + FSL)较全局特征多标签的模型 GSL,mAP 提高了 9.4%,R-1 提高了 5.5%,较前景特征多标签的模型 FSL,mAP 提高了 2.8%,R-1 提高了 1.5%。同样的实验结果,在 Market-1501→DukeMTMC-reID 数据集跨域行人重识别实验中得到了验证。实验结果证明,联合全局和前景特征的多标签协同学习模型 MCL 取得了最优的性能。

**表 2　在 DukeMTMC-reID 和 Market-1501 数据集上的多标签有效性消融实验**

**Table 2　Ablation study of multi labels validity on DukeMTMC-reID dataset and Market-1501 dataset**

| 方法 | Market-1501→DukeMTMC-reID | | | | DukeMTMC-reID→Market-1501 | | | |
|---|---|---|---|---|---|---|---|---|
| | R-1/% | R-5/% | R-10/% | mAP/% | R-1/% | R-5/% | R-10/% | mAP/% |
| GSL | 78.0 | 88.8 | 92.5 | 65.1 | 87.7 | 94.9 | 96.9 | 71.2 |
| FSL | 82.4 | **91.1** | **93.4** | 69.0 | 91.7 | 96.5 | 97.7 | 77.8 |
| MCL(GSL + FSL) | **82.5** | **91.1** | 93.2 | **70.5** | **93.2** | **97.1** | **98.1** | **80.6** |

注:黑体数据为每列最优值。

### 3.3　对比实验

将本文方法在 Market-1501、DukeMTMC-reID 和 MSMT17 数据集上开展跨域行人重识别方法的性能对比验证。

首先,与现有的多标签学习方法 MPLP + MMCL[11] 进行对比,与其利用基于记忆的非参数分类器来训练标签分类不同,本文采用协同学习的平均模型来监督标签的训练,监督方式灵活,具有显著优势,实验结果证明了这一点。如表 3所示,在 Market-1501→DukeMTMC-reID 跨数据集的行人重识别实验中,本文方法相较于现有的多标签学习方法 MPLP + MMCL[11],mAP 提高19.1%,R-1 提高 10.1%,在其他数据集上,本文方法同样具有显著的性能提升。

其次,本文方法优于比较方法,具有显著优势。在 Market-1501→DukeMTMC-reID 跨数据集的行人重识别实验中,与 ECN[28]、D-MMD[29]、AD-Cluster[30]、SSG[21]、DG-Net + +[31]、JVTC + [32]、MPLP + MMCL[11]、NRMT[33]、MEB[34] 等方法进行比较,如表 3 所示。本文方法相较于利用多个网络相互学习的跨域行人重识别方法 MEB[34],mAP 提高了 4.4%,R-1 提高了 2.9%。与同样利用 2 个网络的自训练聚类的跨域行人重识别方法 NRMT[33] 相比,mAP 提高了 8.3%,R-1 提高了 4.7%。如表 4 所示,在 DukeMTMC-reID→Market-1501 跨数据集的行人重识别实验中,本文方法 mAP 和 R-1 分别达到了 80.6% 和 93.2%,与跨域行人重识别方法 MEB[34] 相比,mAP 提高了 4.6%,R-1 提高了 3.3%。同时,比跨域行人重识别方法 NRMT[33] 的 mAP 和 R-1 分别提高了 8.9% 和 5.4%。

如表 5 所示,在 Market-1501→MSMT17 跨数据集的行人重识别实验中,本文方法比利用图像

**表 3　在 DukeMTMC-reID 数据集上不同方法的比较**

**Table 3　Comparison with different methods on DukeMTMC-reID dataset**

| 方法 | Market-1501→DukeMTMC-reID | | | |
|---|---|---|---|---|
| | R-1/% | R-5/% | R-10/% | mAP/% |
| ECN[28] | 63.3 | 75.8 | 80.4 | 40.4 |
| D-MMD[29] | 63.5 | 78.8 | 83.9 | 46.0 |
| AD-Cluster[30] | 72.6 | 82.5 | 85.5 | 54.1 |
| SSG[21] | 76.0 | 85.8 | 89.3 | 60.3 |
| DG-Net + +[31] | 78.9 | 87.8 | 90.4 | 63.8 |
| JVTC + [32] | 80.4 | 89.9 | 92.2 | 66.5 |
| MPLP + MMCL[11] | 72.4 | 82.9 | 85.0 | 51.4 |
| NRMT[33] | 77.8 | 86.9 | 89.5 | 62.2 |
| MEB[34] | 79.6 | 88.3 | 92.2 | 66.1 |
| MCL(本文) | **82.5** | **91.1** | **93.2** | **70.5** |

注:黑体数据为每列最优值。

**表 4　在 Market-1501 数据集上不同方法的实验比较**

**Table 4　Comparison with different methods on Market-1501 dataset**

| 方法 | DukeMTMC-reID→Market-1501 | | | |
|---|---|---|---|---|
| | R-1/% | R-5/% | R-10/% | mAP/% |
| ECN[28] | 75.1 | 87.6 | 91.6 | 43.0 |
| D-MMD[29] | 70.6 | 87.0 | 91.5 | 48.8 |
| AD-Cluster[30] | 86.7 | 94.4 | 96.5 | 68.3 |
| SSG[21] | 86.2 | 94.6 | 96.5 | 68.7 |
| DG-Net + +[31] | 82.1 | 90.2 | 92.7 | 61.7 |
| JVTC + [32] | 86.8 | 95.2 | 97.1 | 67.2 |
| MPLP + MMCL[11] | 84.4 | 92.8 | 95.0 | 60.4 |
| NRMT[33] | 87.8 | 94.6 | 96.5 | 71.7 |
| MEB[34] | 89.9 | 96.0 | 97.5 | 76.0 |
| MCL(本文) | **93.2** | **97.1** | **98.1** | **80.6** |

注:黑体数据为每列最优值。

表5 在MSMT17数据集上不同方法的实验比较

Table 5 Comparison with different methods on MSMT17 dataset

| 方法 | Market-1501→MSMT17 | | | | DukeMTMC-reID→MSMT17 | | | |
|---|---|---|---|---|---|---|---|---|
| | R-1/% | R-5/% | R-10/% | mAP/% | R-1/% | R-5/% | R-10/% | mAP/% |
| ECN[28] | 25.3 | 36.3 | 42.1 | 8.5 | 30.2 | 41.5 | 46.8 | 10.2 |
| D-MMD[29] | 29.1 | 46.3 | 54.1 | 13.5 | 34.4 | 51.1 | 58.5 | 15.3 |
| MPLP + MMCL[11] | 40.8 | 51.8 | 56.7 | 15.1 | 43.6 | 54.3 | 58.9 | 16.2 |
| NRMT[33] | 43.7 | 56.5 | 62.2 | 19.8 | 45.2 | 57.8 | 63.3 | 20.6 |
| DG-Net + +[31] | 48.4 | 60.9 | 66.1 | 22.1 | 48.8 | 60.9 | 65.9 | 22.1 |
| MCL(本文) | **57.3** | **68.5** | **73.3** | **27.4** | **58.5** | **70.0** | **74.5** | **28.5** |

注:黑体数据为每列最优值。

生成器来达到域适应的跨域行人重识别方法DG-Net + +[31]的mAP和R-1分别提高了5.3%和8.9%,比跨域行人重识别方法NRMT[33]的mAP和R-1分别提高了7.6%和13.6%。在DukeMTMC-reID→MSMT17跨数据集的实验中,本文方法比跨域行人重识别方法NRMT[33]、DG-Net + +[31]的mAP分别提高了7.9%和6.4%,证明了该方法的有效性。综上所述,本文提出的多标签协同学习跨域行人重识别方法有效提高了跨域行人重识别的性能。

## 4 结 论

本文提出的多标签协同学习的跨域行人重识别模型,在Market-1501、DukeMTMC-reID及MSMT17数据集上都优于对比方法,具有明显的性能优势。其以协同学习平均模型为基础,联合语义对齐的前景特征和全局特征共同训练多标签表示,提高了跨域行人重识别方法的性能,减少了背景噪声的干扰,使模型更关注人体部分,提高了模型的泛化能力,以便其应用于其他场景的数据。

本文相比单域的行人重识别方法,仍有不足之处,问题在于伪标签的质量相比于真实的标签仍有一些差距,下一步的主要工作将放在提高伪标签的质量上,通过对标签的选择上进行深入挖掘,来筛选更有价值更为准确的标签。

## 参考文献(References)


[1] DENG W J,ZHENG L,YE Q X,et al. Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person re-identification[C]// Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press,2018:994-1003.

[2] FAN H,ZHENG L,YANG Y. Unsupervised person re-identification: Clustering and fine-tuning[J]. ACM Transactions on Multimedia Computing Communications,2018,14(4):1-18.

[3] SUN Y F,ZHENG L,YANG Y,et al. Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline)[C]// Proceedings of European Conference on Computer Vision. Berlin:Springer,2018:480-496.

[4] 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 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press,2017:1077-1108.

[5] KALAYEH M M,BASARAN E,GOKMEN M,et al. Human semantic parsing for person re-identification[C]// Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. Piscataway:IEEE Press,2018:1062-1071.

[6] WANG J Y,ZHU X T,GONG S G,et al. Transferable joint attribute-identity deep learning for unsupervised person re-identification[C]// Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. Piscataway:IEEE Press,2018:2275-2284.

[7] LV J M,CHEN W H,LI Q,et al. Unsupervised cross-dataset person re-identification by transfer learning of spatial-temporal patterns[C]// Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. Piscataway:IEEE Press,2018:7948-7956.

[8] LI M X,ZHU X T,GONG S. Unsupervised person re-identification by deep learning tracklet association[C]// Proceedings of European Conference on Computer Vision. Berlin:Springer,2018:737-753.

[9] ZHONG Z,LIANG Z,ZHENG Z D,et al. Camera style adaptation for person re-identification[C]// Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. Piscataway:IEEE Press,2018:5157-5516.

[10] LIN Y T,ZHENG L,ZHENG Z D,et al. Improving person re-identification by attribute and identity learning[J]. Pattern Recognition,2019,95:151-161.

[11] WANG D K,ZHANG S L. Unsupervised person re-identification via multi-label classification[C]// Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press,2020:19874684.

[12] ZHANG X,LUO H,FAN X,et al. AlignedReID:Surpassing human-level performance in person re-identification[EB/OL]. (2018-01-31)[2021-10-01]. https://arxiv.org/abs/1711.08184.

[13] ZHENG L,HUANG Y,LU H,et al. Pose-invariant embedding for deep person re-identification[J]. IEEE Transactions on Image Processing,2019,28(9):4500-4509.

[14] ZHU K,GUO H,LIU Z,et al. Identity-guided human semantic

parsing for person re-identification[C]// Proceedings of European Conference on Computer Vision. Berlin: Springer, 2020: 346-363.

[15] GUO J Y, YUAN Y H, HUANG L, et al. Beyond human parts: Dual part-aligned representations for person re-identification [C]// Proceedings of IEEE International Conference on Computer Vision. Piscataway: IEEE Press, 2019: 19398436.

[16] ZENG K W, NIAN M N, WANG Y H, et al. Hierarchical clustering with hard-batch triplet loss for person re-identification[C]// Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2020: 13657-13665.

[17] GE Y, CHEN D, LI H. Mutual mean-teaching: Pseudo label refinery for unsupervised domain adaptation on person re-identification[C]// Proceedings of the International Conference on Learning Representations. Piscataway: IEEE Press, 2020.

[18] YU H X, ZHENG W S, WU A, et al. Unsupervised person re-identification by soft multilabel learning[C]// Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2019: 2148-2157.

[19] ZHONG Z, ZHENG L, KANG G L, et al. Random erasing data augmentation[EB/OL]. (2017-11-16)[2021-10-01]. https://arxiv.org/abs/1708.04896v1.

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

[21] FU Y, WEI Y C, WANG G S, et al. Self-similarity grouping: A simple unsupervised cross domain adaptation approach for person re-identification[C]// Proceedings of IEEE International Conference on Computer Vision. Piscataway: IEEE Press, 2019: 6112-6121.

[22] ZHAO H, SHI J, QI X, et al. Pyramid scene parsing network [C]// Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2016: 17355193.

[23] ZHENG L, SHEN L Y, LU T, et al. Scalable person re-identification: A benchmark[C]// Proceedings of IEEE International Conference on Computer Vision. Piscataway: IEEE Press, 2015: 1116-1124.

[24] ZHENG Z D, ZHENG L, YANG Y. Unlabeled samples generated by GAN improve the person re-identification baseline in vitro[C]// Proceedings of IEEE International Conference on Computer Vision. Piscataway: IEEE Press, 2017: 17453019.

[25] WEI L H, ZHANG S L, GAO W, et al. Person transfer GAN to bridge domain gap for person reidentification[C]// Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2018: 18347650.

[26] DENG J, DONG W, SOCHER R, et al. ImageNet: A large-scale hierarchical image database[C]// Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2009: 248-255.

[27] ZHONG Z, LIANG Z. Re-ranking person re-identification with k-reciprocal encoding[C]// Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2017: 1318-1327.

[28] ZHONG Z, ZHENG L, LUO Z, et al. Invariance matters: Exemplar memory for domain adaptive person re-identification[C]// Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2019: 598-607.

[29] MEKHAZNI D, BHUIYAN A, EKLADIOUS G, et al. Unsupervised domain adaptation in the dissimilarity space for person re-identification[C]// Proceedings of European Conference on Computer Vision. Berlin: Springer, 2020: 159-174.

[30] ZHAI Y P, LU S J, YE Q X, et al. AD-Cluster: Augmented discriminative clustering for domain adaptive person re-identification[C]// Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2020: 9021-9030.

[31] ZOU Y, YANG X D, YU Z D, et al. Joint disentangling and adaptation for cross-domain person re-identification[C]// Proceedings of European Conference on Computer Vision. Berlin: Springer, 2020: 87-104.

[32] LI J, ZHANG S L. Joint visual and temporal consistency for unsupervised domain adaptive person re-identification[C]// Proceedings of European Conference on Computer Vision. Berlin: Springer, 2020: 483-499.

[33] ZHAO F, LIAO S C, XIE G S, et al. Unsupervised domain adaptation with noise resistible mutual-training for person re-identification[C]// Proceedings of European Conference on Computer Vision. Berlin: Springer, 2020: 526-544.

[34] ZHAI Y P, YE Q X, LU S J, et al. Multiple expert brainstorming for domain adaptive person re-identification[C]// Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2020: 594-611.

# Multi-label cooperative learning for cross domain person re-identification

LI Hui, ZHANG Xiaowei*, ZHAO Xinpeng, LU Xinyu

(College of Computer Science and Technology, Qingdao University, Qingdao 266071, China)

**Abstract**: Cross-domain was an important application scenario in person re-identification, but the apparent difference of person image in illumination condition, shooting angle, imaging background and style between the source domain and target domain was the most important factor that leads to the decline of the generalization ability of person re-identification model. A cross-domain person re-identification method was proposed based on multi-label cooperative learning to solve the problem. Firstly, the semantic parsing model was used to construct the multi-label data based on semantic alignment, which was able to guide us to construct global features that pay more attention to the person area, achieve the purpose of semantic alignment, and reduce the background influence on cross-domain person re-identification. Furthermore, the collaborative learning average model was used to generate a multi-label representation of the person re-identification model based on global and local features after semantic alignment, reducing the interference of noisy hard labels in the cross-domain scenario. Finally, the semantic alignment model of multi-label based on a collaborative learning network framework was combined to improve the identification ability of re-identification model. The experiment results show that on the Market-1501 → DukeMTMC-reID, DukeMTMC-reID → Market-1501, Market-1501 → MSMT17, DukeMTMC-reID→MSMT17 cross-domain person re-identification data set, compared with the current state-of-the-artscross-domain person re-identification method NRMT, the mean average precision of this method is increased by 8.3%, 8.9%, 7.6% and 7.9%, respectively. Multi-label cooperative learning method has obvious advantages.

**Keywords**: cross-domain person re-identification; semantic alignment; global feature; multi-label representation; collaborative learning

**Received**: 2021-10-10; **Accepted**: 2021-10-29; **Published online**: 2021-11-12 08:58

**URL**: kns.cnki.net/kcms/detail/11.2625.V.20211110.1924.004.html

**Foundation items**: National Natural Science Foundation of China (61902204); Shandong Provincial Natural Science Foundation (ZR2019BF028)

* **Corresponding author**. E-mail: by1306114@buaa.edu.cn

http://bhxb.buaa.edu.cn jbuaa@buaa.edu.cn
**DOI**: 10.13700/j.bh.1001-5965.2022.0131

# 基于球场重建的球员运动数据分析

吉晓琪[1], 宋子恺[1], 于俊清[1,2,*]

(1. 华中科技大学 计算机科学与技术学院, 武汉 430074; 2. 华中科技大学 网络与计算中心, 武汉 430074)

**摘 要**: 足球比赛中球员运动数据分析对增加观众的观看体验和辅助教练进行球员评估有着重要意义。球员运动数据分析的难点在于如何定位球员在球场上的坐标,即如何确定足球视频中单帧画面出现的缺损球场与标准二维球场之间的映射关系。针对如何在足球比赛中克服相机的高速移动和视角剧烈变化,设计并提出了利用球场重建与球员跟踪来进行球员运动数据分析的方法。球场重建方面,将足球视频中的球场分组为左中右3部分,每组通过球场分割、球场直线检测、球场直线分组、球场中圈点集合识别和球场关键点匹配来实现缺损球场到标准球场的映射;球员跟踪采用核相关滤波(KCF)跟踪算法,得到了球员运动数据统计的可视化结果。结合球场重建和球员跟踪算法定位球员的标准坐标,统计球员的一系列运动数据并进行可视化分析。提出的球员运动数据分析方法能够准确而快速地统计出球员的运动数据,包括球员坐标、运动轨迹、奔跑速度、活动范围和球员间距。球场重建方面采用图像交并进行评估,交并比达到87%,相比于传统的基于字典查询的方法(交并比为83.3%)准确度提升了3.7%。实验结果表明:所提出的球场重建方法能够更准确地表示球场映射关系,为球员运动数据分析统计提供更好的支持。

**关 键 词**: 足球比赛; 球员数据分析; 球场重建; 球场分割; 球员跟踪; 球员运动检测
**中图分类号**: TP391
**文献标志码**: A **文章编号**: 1001-5965(2022)08-1543-10

随着足球运动的流行,足球比赛视频内容分析和球员运动数据分析吸引了大量的研究人员。足球比赛视频内容分析拥有广泛的应用场景,如辅助裁判判罚的验证、球队球员战术分析、精彩度片段的自动识别、视频注释和摘要提取、基于内容的视频压缩、比赛的自动总结、定制广告插入、图形化球员数据展示、球员和球队的数据统计评估等。

足球比赛视频内容分析研究成果丰硕,但是足球比赛视频球员运动数据分析领域却没有完善的方法。足球比赛视频中的球员运动数据分析面临着以下挑战:①足球比赛视频受环境影响较大,不利于比赛视频帧的图像处理;②获取球员准确定位的复杂度较高;③球场重建方面的研究内容较少;④基于深度学习的球场重建方法存在数据集少、训练耗时高和准确度不足的问题。

针对足球比赛视频中的远镜头视频,本文提出并设计了一种足球比赛中的球员运动数据统计与分析方法。利用基于主色的球场分割算法剔除掉非球场区域,采用霍夫变换[1]和最小二乘法检测球场直线,使用概率决策树实现球场直线分组,并利用区域增长算法识别球场中圈,基于球场直线交点和球场中圈关键点计算单应矩阵实现球场重建,基于球场重建方法和核相关滤波(kernelized

**收稿日期**: 2022-03-09; **录用日期**: 2022-03-25; **网络出版时间**: 2022-03-31 09:19
**网络出版地址**: kns.cnki.net/kcms/detail/11.2625.V.20220329.1820.001.html
**基金项目**: 国家重点研发计划 (2020YFB1805601)
* **通信作者**. E-mail: yjqing@hust.edu.cn

**引用格式**: *吉晓琪, 宋子恺, 于俊清. 基于球场重建的球员运动数据分析[J]. 北京航空航天大学学报, 2022, 48(8): 1543-1552.*
*JI X Q, SONG Z K, YU J Q. Player movement data analysis on soccer field reconstruction [J]. Journal of Beijing University of Aeronautics and Astronautics, 2022, 48(8): 1543-1552 (in Chinese).*

correlation filter, KCF)[2]球员跟踪算法计算出球员的运动数据,并能够进行直观的可视化展示和统计分析。

本文的主要创新点包括球场分割、直线分组及球场重建。在球场分割方面,利用基于主色的球场分割算法,采用 HSV 色彩模式确定阈值,利用腐蚀、膨胀、平滑等形态学操作对结果进行优化,得到球场分割结果。利用霍夫变换和最小二乘法检测直线。在球场直线分组方面,将检测到的球场直线分为 5 组,利用直线的倾斜度及直线之间的关系构建概率决策树实现分组。对于中场区域,基于区域增长算法逐像素识别出中圈的关键点信息,利用中圈关键点弥补直线交点不足的问题。在球场重建方面,对于球场左半场和右半场区域,在直线分组的基础上利用直线间距确定直线及交点的匹配关系。对于球场中场区域,利用中圈点集合进行椭圆拟合得到关键点匹配关系,基于关键点匹配信息计算单应矩阵实现球场映射。

基于球场重建和球员跟踪算法,可以对球员运动数据进行统计与分析。利用单应矩阵和球员跟踪结果得到球员在标准球场上的坐标,并统计球员的运动数据,包括运动轨迹、奔跑速度、球员间距、球员分布关系和球员热力图。

## 1 相关工作

球员运动数据分析方法中的难点在于表示足球比赛视频帧与标准二维球场之间的映射关系,而球员重建映射方面的研究主要分为基于字典查询和基于深度学习 2 种。

基于字典查询的球场映射方面,文献[3]将球场映射注册问题表述为球场轮廓图与单应矩阵生成字典的最邻近搜索,通过关注球场边缘图及球场线,而不是复杂的语义分段,简化了球场表述,提出了使用条件生成对抗网络(conditional generative adversarial nets, CGAN)[4],直接从 RGB 图像生成边缘图像,该方法利用了球场线标记的边缘数据,为了从图像中提取边缘数据,比较使用了 3 种不同的方法,即定向梯度直方图(histogram of oriented gradient, HOG)[5]特征、倒角匹配和卷积神经网络(convolutional neural networks, CNN),在合成的字典中尽可能详细地涵盖了不同摄像机的角度和位置,并将问题简化为求取每一帧边缘图的匹配问题,使用马尔可夫随机场(Markov random field, MRF)来进行辅助优化,对于测试数据,通过训练好的 GAN 模型,生成图像的边缘,并以字典查询的方式完成图像的映射。文献[6]在分层网络中采用了图像翻译网络,对运动场进行细分,将非球场区域与球场区域进行分割,得到仅包含球场区域的图像,对球场线进行提取,并对数据库提取边缘图像,在进行相似度计算时,采用了优化后的 Lucas-Kanade 算法[7],但是与文献[8]提出的方法存在相似的问题,这 2 种方法的瓶颈在于数据库的必要性阻碍了可伸缩性。

基于深度学习的球场映射方面,文献[9]提出了一种训练网络,可以通过自我监督直接回归求取 2 个图像之间的单应性,并提出一种新的 VGG 格式的网络,展示了如何使用 4 个点进行参数化训练来获得良好的深度估计问题,提出了单应性估计的另一种表达形式。文献[10]提出了在无监督的情况下训练深层网络,用来学习估计相对应的单应性,将深度学习的优势与基于特征的优势结合在一起,依赖特征进行单应性的估计,不同之处是其学习的是特征而不是如何定义特征。这些方法与传统方法相比较,改进了推理过程,却没有明显改善。文献[11]提出了一种两阶段深度神经网络(deep neural networks, DNN)图像配准框架,第 1 阶段称为初始化注册网络,进行图像注册的粗略估算,通过单应变换进行参数化,第 2 阶段称为注册误差网络,利用梯度优化对初始化注册网络进行反馈优化,但是其鲁棒性存在较大的提升空间。

文献[8]将图像通过深层网络执行语义分割,再使用分支定界和几何先验的 MRF 对映射进行估计,将问题表述为 MRF 中的分支定界推断,其中能量函数根据语义信息(如球场表明、球场线及球场上的圆)定义,实现了球场映射。

## 2 球员运动数据分析方法

球员运动数据分析方法的核心在于利用关键点单应矩阵的球场重建方法。整体方法流程如图 1所示。通过对足球比赛视频进行分析处理,利用单应矩阵进行球场重建,同时结合球员跟踪

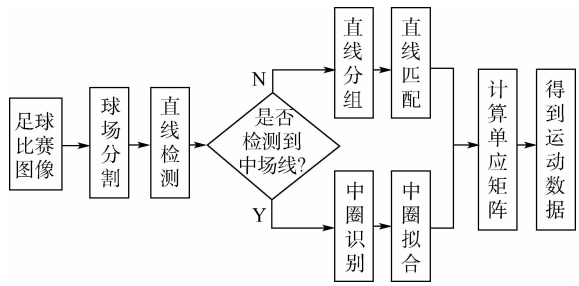



图 1 球员运动数据分析方法整体流程

Fig. 1 Flow chart of player motion data analysis method

坐标计算球员在标准球场上的坐标,基于球员在球场上的坐标计算球员的运动数据,并对这些数据进行统计分析及可视化展示。

### 2.1　球场区域检测

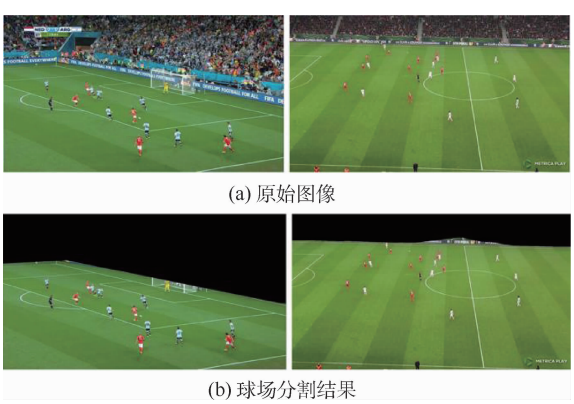
球场区域中最明显的特征是球场直线和球场中圈 2 部分,对球场上的直线和中圈进行准确检测是实现视频帧与标准二维球场映射的关键。本节介绍了球场分割和球场直线检测,并对球场直线进行分组以简化后续求取关键点匹配关系的复杂度,利用球场中圈关键点弥补中场区域直线交点不足的问题。

#### 2.1.1　球场分割和直线检测

对于输入的足球比赛视频帧,采用基于主色的球场分割算法将非球场区域剔除,减少对球场处理时非球场区域的干扰。

球场分割算法步骤为:①彩色空间变换。将输入图像转换到 HSV 彩色空间中。②阈值分割。将球场主色所在的颜色范围确定为阈值,利用基于阈值的分割算法确定球场区域,得到初步分割结果。③形态学处理。通过形态学腐蚀和膨胀操作来平滑输出的图像,以及填充图像中的孔洞。球场分割结果如图 2 所示。

采用传统的图像处理方法对球场直线进行检测。在整个方法中,需要对图像进行预处理,包括将 RGB 图像转换为灰度图像,通过高斯模糊对边缘进行平滑处理,将处理后的图像输入 Canny 边缘特征[12]得到图像的边缘特征。为减少非球场区域的干扰,利用球场分割算法将非球场区域剔除,选择感兴趣的 ROI 球场区域,只对球场区域进行霍夫变换,找到球场上的直线。但是,霍夫变换检测到的球场直线可能有很多,而且不是连续的,因此需要对许多直线进行过滤和筛选。根据直线的倾斜角及直线之间的垂直距离,可以将霍夫变换检测到的直线中属于同一条球场直线的线进行分组。对分组后的直线进行拟合,拟合得到连续的球场直线。



(a) 原始图像

(b) 球场分割结果

图 2　半场区域与中场区域的球场分割结果

Fig. 2　Segmentation results of half field area and midfield area of soccer field

#### 2.1.2　球场直线分组

根据国际足球协会理事会的游戏规则,足球比赛场地由中场线分为 2 部分,每部分分别存在水平线和垂直线。考虑到这些特性,将球场直线分为 5 组(见图 3(a)),分别如下:

1) $G_1 = \{g_{1i}\}_{i=1}^{6}$。位于球场左半区的水平线段,按照从上到下的顺序排序。

2) $G_2 = \{g_{2i}\}_{i=1}^{3}$。位于球场右半区的垂直线段,按照从右到左的顺序排序。

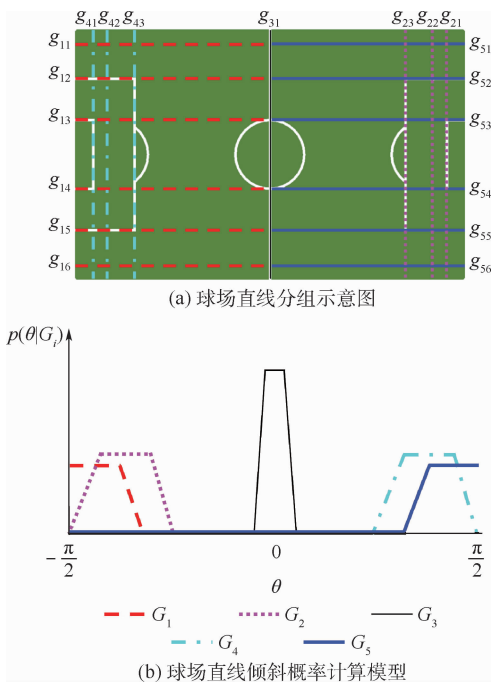
3) $G_3 = \{g_{31}\}$。球场中场线。

4) $G_4 = \{g_{4i}\}_{i=1}^{3}$。位于球场左半区的垂直线段,按照从左到右的顺序排序。

5) $G_5 = \{g_{5i}\}_{i=1}^{6}$。位于球场右半区的水平线段,按照从上到下的顺序排序。

考虑到检测到的直线可能存在噪声,将不属于上述 5 组的直线归类为第 6 组 $G_6$。

将检测到的直线按照倾斜角进行分类。定义倾斜模型为 $T = \{p(\theta \mid G_j)\}_{j=1}^{6}$,其中,$p(\theta \mid G_j)$ 为第 $j$ 组直线对应的概率计算公式。概率计算模型如图 3(b)所示,横坐标表示线段的倾斜角度 $\theta$,范围为 $-\pi/2 \sim \pi/2$,不同形状表示不同的线段分组,纵坐标表示倾斜角度对应的概率。其中,前5 个概率计算公式根据倾斜角度计算,如下:




(a) 球场直线分组示意图

(b) 球场直线倾斜概率计算模型

图 3　球场直线分组

Fig. 3　Straight-line grouping in soccer field

$$p(\theta \mid G_j) \propto \begin{cases} \exp\left[-\dfrac{(\theta-\theta_{j,1})^2}{2\sigma_j^2}\right] & \theta < \theta_{j,1} \\ 1 & \theta \in [\theta_{j,1},\theta_{j,2}] \\ \exp\left[-\dfrac{(\theta-\theta_{j,2})^2}{2\sigma_j^2}\right] & \theta > \theta_{j,2} \end{cases} \tag{1}$$

式中：$\theta_{j,1}$ 和 $\theta_{j,2}$ 分别为第 $j$ 组直线组中对应的最小倾斜角和最大倾斜角；$\sigma_j$ 为第 $j$ 组的标准差。

对于第 6 组倾斜模型，其概率计算定义如下：

$$p(\theta \mid G_6) = \frac{1}{2\pi} \qquad \theta \in [-\pi,\pi) \tag{2}$$

检测到的直线的具体分组由图 4 所示的概率决策树进行分类。概率决策树的构建规则如下：假设检测到的直线有 $n$ 条，则该概率决策树包含 $n$ 层，对于所有非叶子节点，其子节点都是 6 个，第 $i$ 层的节点总数为 $6^i$。对于第 $i$ 层中每个节点，在所检测到的直线中，拥有倾斜角为 $\theta_i$ 的直线，每个节点的概率计算如下：

$$\Pr(G_j \mid \theta_i) = \frac{\Pr(G_j)p(\theta_i \mid G_j)}{\sum_{j=1}^{6}\Pr(G_j)p(\theta_i \mid G_j)} \qquad j \in [1,6] \tag{3}$$

式中：$\Pr(G_j)$ 为该条直线被分配给第 $j$ 组的先验概率，对于每个节点，其先验概率的获取可通过如下规则得到：

1）$G_1$ 和 $G_5$ 最多包含 6 条直线。

2）$G_2$ 和 $G_4$ 最多包含 3 条直线。

3）$G_3$ 仅包含 1 条直线。

4）$G_1$ 中的直线和 $G_5$ 中的直线不会出现在同一张图像中。

5）$G_2$ 中的直线和 $G_4$ 中的直线不会出现在同一张图像中。

6）属于同一组的直线在球场区域范围内不会存在交点。

利用规则 1）~ 规则 3），可从根节点到叶子节点的分析过程中对概率决策树进行剪枝操作，减少计算量，具体操作为：在概率决策树的每层概率计算中，如果某组的直线数目到达上限，则直接设置其先验概率为零。对于规则 4）和规则 5），当确定某条直线的分组后，可以认为剩余的直线不属于其对立组，在计算概率决策树的过程中设置对立组节点的先验概率为零。规则 6）可以通过分析某条直线和已经计算结束的直线之间是否存在交点来进行概率决策树的剪枝操作，将不可能属于的节点的先验概率置为零。

在考虑剪枝操作后，假设某节点的 6 个子节点中所有剩余的可能节点数 $N_g \leqslant 6$，则每个子节点的先验概率为 $1/N_g$。将所有路径上从根节点到叶子节点的概率相乘，得到总概率最高的路径叶子节点便是该条直线的最终分类。

### 2.1.3 中圈点集合识别

由于球场直线交点不满足求取 2 张图像之间单应矩阵的输入集合点数量，需要对中圈识别以获取足够的关键点。

设计如图 5(a) 所示的结构元素 $e_{RG}$，利用该结构元素可以识别获取中圈椭圆上小的不连续点集合，并且该结构元素也能够很好地适应中圈椭圆的曲率。该结构元素由一个主点（最大圆形）构成，从该点延伸出 5 条点线，其中 3 条沿主轴方向，为小圆、方块和三角形的点线，另 2 条分别沿主轴顺时针和逆时针旋转 45°的点线，为叉形和菱形的点线。具体识别过程为：将结构元素 $e_{RG}$ 初始化在最初检测到的椭圆交点像素处（见图 5(b)），在每次迭代中，计算结构元素 $e_{RG}$ 中 5 条点线上白色像素点的数量，下一次迭代中，将 $e_{RG}$ 移动至白色像素点数量最多的点线的开头位

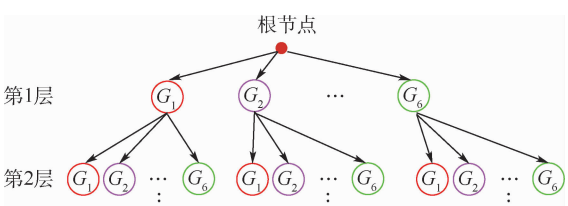



图 4 概率决策树模型

Fig. 4 Probabilistic decision tree model

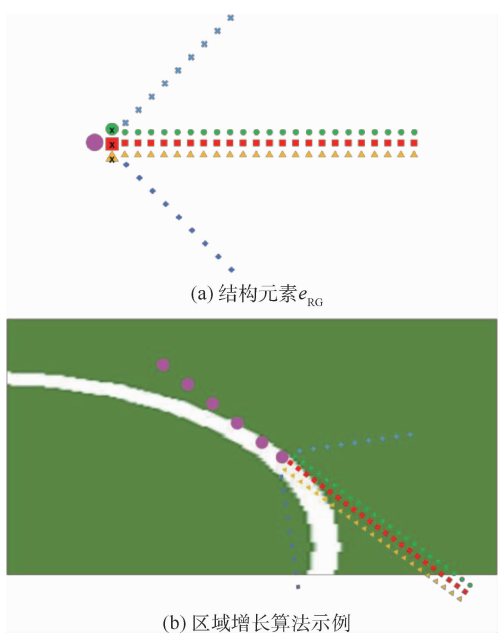



图 5 球场中圈点集合识别

Fig. 5 Recognition of center circle point set in soccer field

置,并将该点添加到中圈椭圆点集合中。对于叉形或菱形的点线,如果计算的白色像素数量最多,还会将 $e_{RG}$ 进行角度旋转,叉形线则逆时针旋转 45°,菱形线则顺时针旋转 45°,通过这种方式,使结构元素 $e_{RG}$ 较好地适应椭圆的曲率。当要添加的像素点已经存在于椭圆点集合当中或者椭圆点集合达到一定数量时,区域增长算法结束。至于中圈不完整的情形,通过正反 2 次来执行算法获得点集合。

## 2.2　球场重建

完成球场直线检测分组和中圈点集合识别后,利用直线和中圈确定球场上的关键点位置,得到关键点的匹配信息以计算足球比赛视频帧与标准二维球场之间的单应矩阵,利用单应矩阵实现足球比赛视频帧中球场区域与标准二维球场之间的映射,完成球场重建工作。

将球场分为左半场、中场、右半场 3 部分,左半场和右半场区域,球场直线足够多,利用球场直线可以得到足够的关键点信息,通过直线之间的水平垂直关系、同组直线之间的距离关系,确定直线的匹配关系,从而确定直线之间交点的关系。球场中场区域,由于球场直线有限,利用球场中圈来计算中圈上的关键点,得到球场中圈上的关键点匹配关系。球场关键点的定义如图 6 所示。

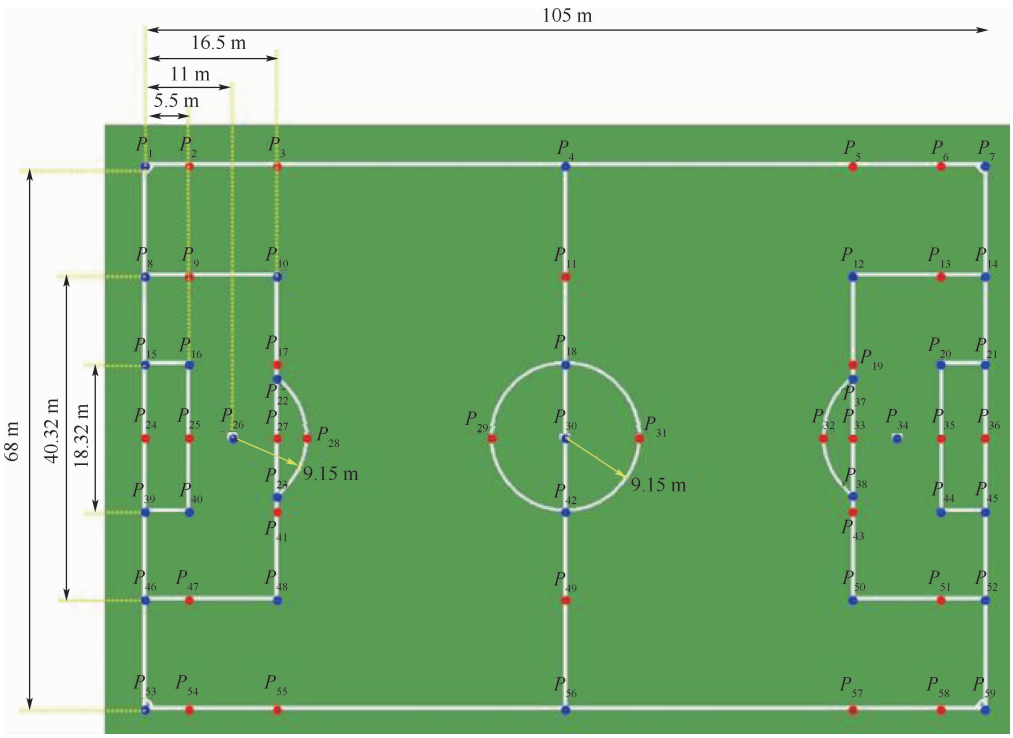
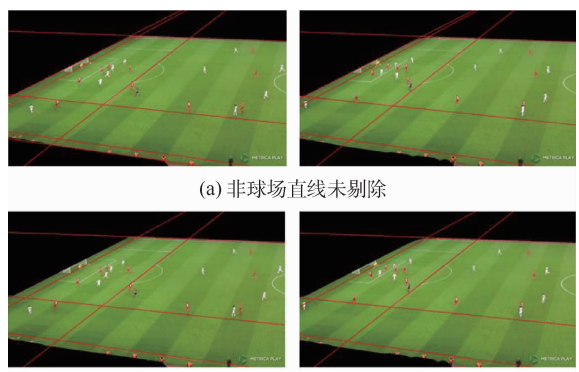



图 6　球场关键点

Fig. 6　Key points in soccer field

球场直线的关键点是垂直球场线和水平球场线之间的交点,为了确定球场线交点与二维球场上关键点的匹配对应关系,需要确定视频帧中检测到的直线与二维标准球场上的球场线之间的对应关系。观察发现,对于检测到的球场线,其特征除了分组用到的倾斜度之外,还有相邻球场线之间的距离关系。利用球场线之间的距离信息来进一步实现球场线之间的匹配。

由于直线检测是在 Canny 边缘检测的图像基础上进行的,在球场分割之后的图像上将球场区域和非球场区域之间的分割处检测为直线。因此,如何识别出最外层的直线是否是球场上的直线是首要任务。通过计算最外层球场直线到非球场区域的距离,利用距离信息来进行判断,可以剔除掉检测到的非球场直线,图 7 显示了最初检测到的球场直线和剔除之后的球场直线的对比实验结果。



(a) 非球场直线未剔除

(b) 非球场直线已剔除

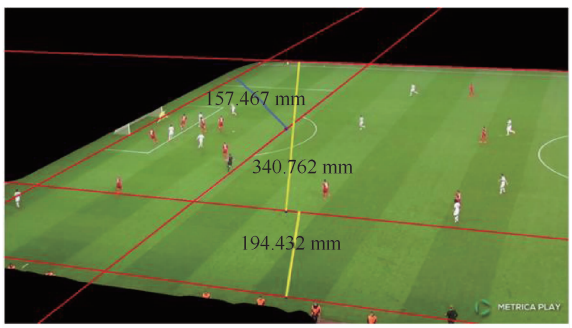
图 7　非球场直线剔除实验结果

Fig. 7　Off-soccer field straight line culling

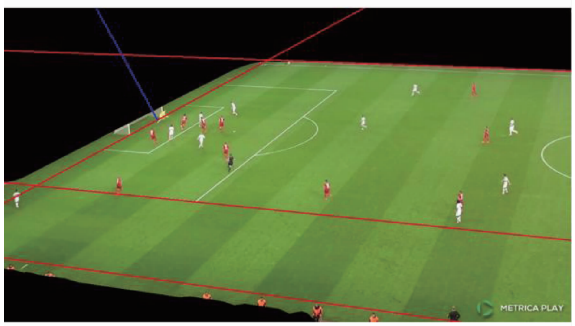
通过对球场直线的筛选,球场区域检测到的直线都属于球场线。可以分析,由于左半场和右半场的范围内直线最多,直线的匹配问题较复杂。球场的两半边,其水平方向的直线最多能检测到 6 条,其垂直方向的直线最多能检测到 3 条。那么对于垂直方向的直线,如果检测到 3 条,就可以直接按照其顺序来一一确定匹配关系。如果检测到 2 条直线,就需要计算这 2 条直线之间的距离,通过距离信息进行判断。同样,为了简化这一过程,通过计算最边缘的直线到非球场区域的垂直距离来确定最外层直线的匹配关系,在最外层直线确定的情况下,剩下的 1 条直线通过距离信息自然能够确定。同理,对于水平方向的直线,也利用这种方法,先确定水平方向最外层的 2 条直线,对于剩下的中间直线,利用距离信息进行确定。球场直线匹配的计算示意图如图 8 所示。在检测到球场直线且直线匹配成功之后,只需要计算求出水平方向直线和垂直方向直线的交点,便可以得到直线关键点,用来进行球场重建。

在得到球场中圈点集合之后,通过最小二乘拟合计算得到中圈椭圆的 5 个参数,包括中圈中心坐标 $C=(x_0,y_0)$、长半轴 $a$、短半轴 $b$ 及椭圆的曲率 $\alpha$。中圈关键点映射的示意图如图 9 所示。

基于球场直线交点匹配关系和球场中圈关键点匹配关系,可以计算单应矩阵 $\boldsymbol{H}$,对于单应矩阵 $\boldsymbol{H}$,其包含 8 个未知量,至少需要 4 对匹配的对应点,且这 4 对匹配的对应点中任意 3 点不共线,可




(a) 最边缘直线到非球场距离




(b) 中间直线间距

图 8 直线匹配距离计算

Fig. 8 Distance calculation of matching straight line

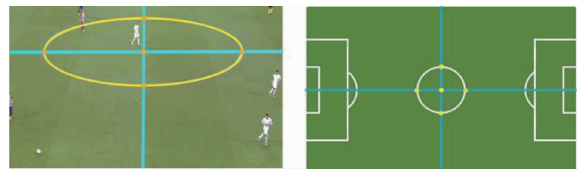


图 9 中圈关键点映射

Fig. 9 Key point mapping in center circle

以求出 2 张图像之间对应的单应矩阵 $\boldsymbol{H}$,实现球场重建。

## 3 球场重建实验

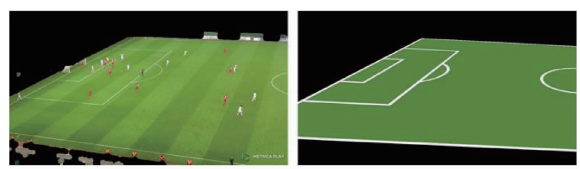
本节实验使用 MATLAB 和 C++编写,采用 OpenCV 视觉库处理图像,球场数据从 Soccer Dataset 数据集[13]获取。在经过处理得到足球比赛视频帧与标准二维球场之间的单应矩阵之后,便可以实现二者之间的映射。选择 793 帧包含球场左中右 3 部分的足球比赛视频帧来进行测试。

利用分割后的球场视频帧同利用单应矩阵进行变换后的标准球场之间的交并比作为评价指标,如下:

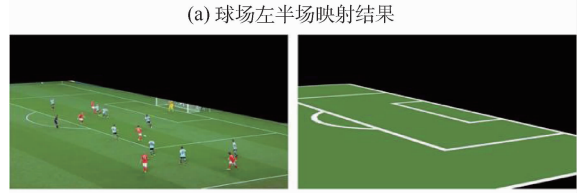
$$\mathrm{IOU}=\frac{C\cap G}{C\cup G} \tag{4}$$

式中:$C$ 表示视频帧中分割后的球场;$G$ 为利用单应矩阵进行变换后的标准球场。

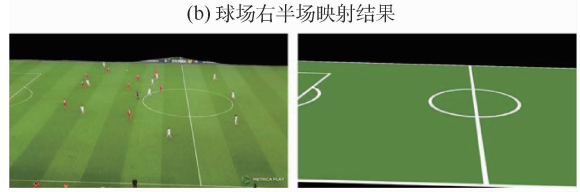
对于球场左半场和右半场 2 部分区域,通过检测直线,实现直线匹配,求取直线的交点,并完成关键点的匹配,通过匹配得到的关键点求出单应矩阵。对于球场中间部分,通过识别中圈的关键点,求得单应矩阵完成球场的映射。图 10 展示了球场映射结果。



(a) 球场左半场映射结果



(b) 球场右半场映射结果

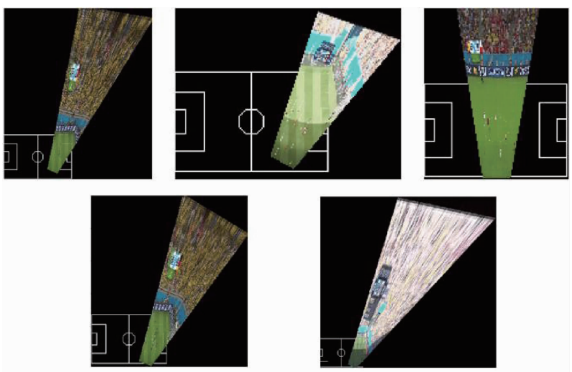


(c) 球场中场映射结果

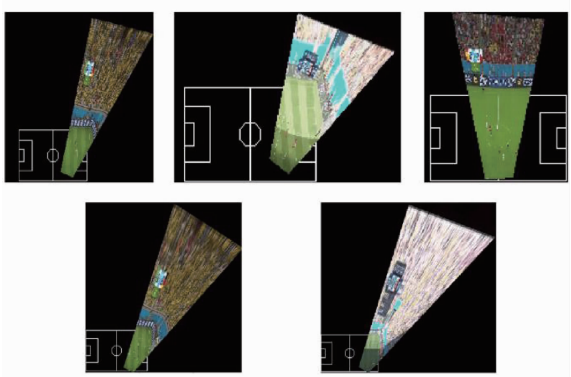
图 10 球场映射结果

Fig. 10 Soccer field mapping results

利用上述计算得到的单应矩阵进行球场映射,计算交并比,通过统计所有测试样例的映射结果,其交并比达到了 87%,能够较好地完成映射,实现后续球员数据获取和统计,对比实验结果如图 11 所示。图 11(a)为采用文献[3]提出的基于字典查询的方法计算得到的实验结果,计算得到的交并比为 83.3%,图 11(b)为本文方法计算得到的实验结果,交并比达到了 87%,相比于其他方法的实验结果在准确度方面有所提升。



(a) 基于字典查询的球场映射结果



(b) 本文方法球场映射结果

图 11　球场映射实验对比

Fig. 11　Experimental comparison of soccer field mapping

## 4　球员运动数据分析

### 4.1　球员跟踪

经过足球比赛画面中的部分可见球场到完整的标准二维球场之间的映射之后,需要利用跟踪算法确定目标球员对象在视频序列中的位置,本文选取 KCF 目标跟踪算法。

KCF 算法[2]是基于相关滤波跟踪算法[14-15]的改进算法,其主要特性包括:使用目标周围区域的循环矩阵采集正负样本,采用脊回归训练目标检测器,并利用循环矩阵在傅里叶空间可对角化的性质,将矩阵运算转换为向量的点乘,大大减少了计算量,提高了运行速度,满足实时性的要求。另外,该算法将线性空间的脊回归通过核函数映射到非线性空间,在非线性空间中通过求解一个对偶问题和一些常见的约束,利用循环矩对角化来简化计算复杂度。

由于 KCF 算法的上述特性,其速度快,效果优异,在足球场景的球员跟踪中表现较好,选取 KCF 算法跟踪目标球员。KCF 算法使用循环位移使得其具有边界效应问题,导致在球员较多时易出现跟踪飘移的情况;另外,该算法的搜索区域固定,在快速运动中容易超出搜索范围,导致球员运动数据的获取存在一定的偏差。

### 4.2　球员数据计算

基于单应矩阵和球员跟踪坐标能够得到球员在标准球场上的坐标位置,计算如下:

$$\begin{bmatrix} x_2 \\ y_2 \\ z_2 \end{bmatrix} = \boldsymbol{H} \begin{bmatrix} x_1 \\ y_1 \\ z_1 \end{bmatrix} \tag{5}$$

式中:$(x_1, y_1)$ 为比赛视频帧中的像素点位置,$z_1 = 1$;$(x_2, y_2, z_2)$ 为映射后在标准球场上的坐标,转换为二维形式为 $(x_2/z_2, y_2/z_2)$。

基于式(5)可以进一步求取球员的其他运动数据:

1)球员运动距离。得到球员 $t_i$ 时刻在标准二维球场上的坐标 $(x_{t_i}, y_{t_i})$ 后,可以计算球员从 $t_{i-1}$ 时刻到 $t_i$ 时刻的运动距离 $d_i$:

$$d_i = \sqrt{(x_{t_i} - x_{t_{i-1}})^2 + (y_{t_i} - y_{t_{i-1}})^2} \tag{6}$$

2)球员运动总路程。球员运动的总路程 $d_{\text{total}}$ 计算式为

$$d_{\text{total}} = \sum d_i \tag{7}$$

3)球员运动间距。利用多目标检测和跟踪算法,可以同时得到 2 个球员 $t_i$ 时刻在二维球场上的坐标位置 $(x_{t_{i1}}, y_{t_{i1}})$、$(x_{t_{i2}}, y_{t_{i2}})$,通过式(8)计算出 2 个球员之间的位置:

$$d_i = \sqrt{(x_{t_{i2}} - x_{t_{i1}})^2 + (y_{t_{i2}} - y_{t_{i2}})^2} \tag{8}$$

4)球员运动速度。计算得到球员从 $t_{i-1}$ 时刻到 $t_i$ 时刻的运动路程 $d_i$ 后,利用式(9)计算得到球员从 $t_{i-1}$ 时刻的坐标位置 $(x_{t_{i-1}}, y_{t_{i-1}})$ 运动到 $t_i$ 时刻的坐标位置 $(x_{t_i}, y_{t_i})$ 的平均运动速度 $v_i$。

$$v_i = d_i/(t_i - t_{i-1}) \tag{9}$$

5)球员活动范围。球员活动范围的计算利用式(10)的椭圆方程来表示:

$$(x - x_0)^2/a^2 + (y - y_0)^2/b^2 = 1 \tag{10}$$

式中:$(x_0, y_0)$ 为椭圆的圆心坐标位置;$a$ 为椭圆的长半轴;$b$ 为椭圆的短半轴。

计算得到球员在 $t_i$ 时刻的坐标位置 $(x_{t_i}, y_{t_i})$ 后,对横纵坐标 $x_{t_i}$ 和 $y_{t_i}$ 进行计算,得到其平均值

$x_{ave}$、$y_{ave}$和标准差 $x_{sta}$、$y_{sta}$。将($x_{ave}$,$y_{ave}$)定义为椭圆的中心,$x_{sta}$和 $y_{sta}$分别定义为椭圆的长半轴和短半轴,从而确定球员的活动范围。椭圆的面积 $S=\pi\times a\times b$ 表示球员的活动范围参数化形式,单位为 $m^2$,如图 12 所示。

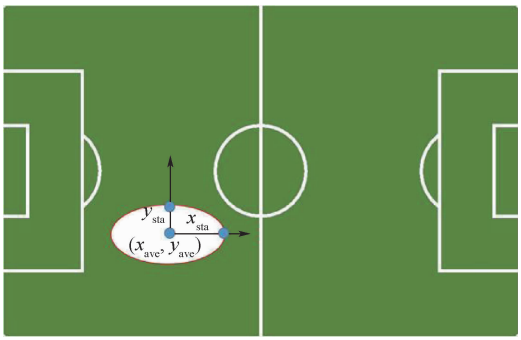



图 12 球员活动范围参数计算

Fig. 12 Calculation of player motion range parameters

### 4.3 球员运动数据统计

基于球员运动距离可以得到球员跑动距离-时间统计图,如图 13 所示,可以看出球员在足球比赛中跑动距离及其变化。

基于球员运动速度可以做球员运动类型的统计,根据运动速度的快慢将运动离散化为冲刺、快跑、慢跑、走、站立 5 种形式,离散方式如表 1 所示。统计分析发现,在慢跑和快跑两方面,前卫球员远远优于后卫和前锋球员。在走和冲刺两方面,后卫和前锋球员要高于前卫球员。图 14 统计了测试视频中 2 名球员不同运动类型的统计信息。足球比赛实际是两方球队的球员在教练的战术安排下,在球场上相互博弈的过程,如果一方占据了进攻的优势,则另一方的后卫球员就需要积极地跑动来进行防守,因此,单从球员不同速度下的跑动距离来判定球员的运动能力不够全面客观,还需要考虑到攻防双方的进攻情况、后卫球员的助攻次数、前卫的跑动范围等因素。

球队的运动距离反映了球员的运动能力,但不能反映出球员在比赛时的活动范围。图 15 展示了球员的跑动示意图及球员的轨迹热力图,可以直观了解到球员在球场上的站位频次情况。

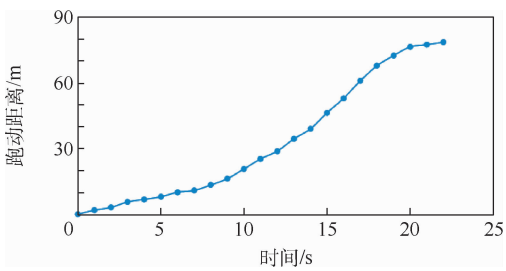



图 13 球员跑动距离-时间统计

Fig. 13 Player running distance-time statistics

表 1 速度离散化说明

Table 1 Description of velocity discretization

| 运动类型分类 | 速度间隔 |
|---|---|
| 站立 | 小于 0.2 m/s |
| 走 | 0.2 ~ 2.1 m/s |
| 慢跑 | 2.1 ~ 3.8 m/s |
| 快跑 | 3.8 ~ 6.1 m/s |
| 冲刺 | 大于 6.1 m/s |

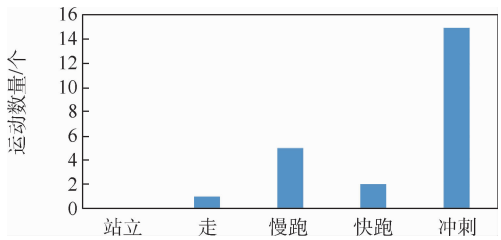



图 14 球员进球时运动类型统计

Fig. 14 Statistics of player's movement types when scoring a goal

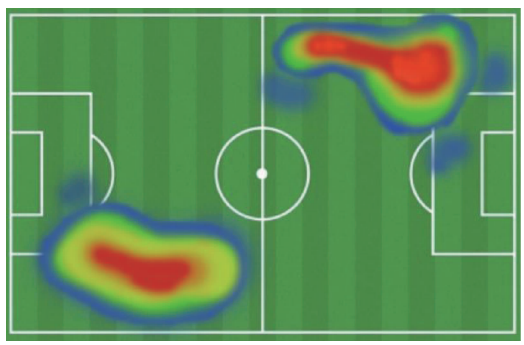


图 15 球员位置热力图

Fig. 15 Player position heat map

## 5 结 论

1)本文方法在二维球场重建方面的交并比能够达到 87%,优于其他对比的二维球场重建方法;在球员运动数据统计方面,可以计算并统计出全场球员的位置、速度、路程、间距、运动类型等数据,并通过热力图可视化展示。实验结果表明,本文方法能够较好地实现球场映射及球员运动数据计算。

2)通过融合球场区域检测、球场重建和球员跟踪的方法,提出并设计了一种球员运动数据统计分析的方法。球场区域检测方面,采用了基于主色的球场分割和基于霍夫变换的直线检测,对球场直线进行分组,对球场中圈进行点集合检测。球场重建方面,利用球场直线及直线交点匹配计算球场左右半场的单应矩阵,利用中圈关键点匹配计算中场单应矩阵,实现球场映射。球员数据统计方面,利用单应矩阵和球员跟踪坐标计算球员的位置、速度、路程、间距、运动类型、热力图等

运动数据并展示。

3）通过改进与集成多个视觉图像算法,实现了对足球比赛视频的运动数据统计与分析的完整流程。获得的足球球员运动数据能够直观地改善观众的观看体验,辅助教练进球球队战术分析和战术安排,有广泛的应用场景。

由于球场直线检测的准确性、球场中圈的识别方法及本文利用的 KCF 跟踪算法的局限性,使得算法的复杂度较高,耗时较长,未来的工作将聚焦在算法优化,提高直线检测准确度,加快球场中圈识别速度,改善球员跟踪算法。

## 参考文献（References）


[1] BALLARD D H. Generalizing the Hough transform to detect arbitrary shapes[J]. Pattern Recognition,1981,13(2):111-122.

[2] 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.

[3] SHARMA R A,BHAT B,GANDHI V. Automated top view registration of broadcast football videos[C]// Proceedings of IEEE Winter Conference on Applications of Computer Vision. Piscataway:IEEE Press,2018:305-313.

[4] MIRZA M, OSINDERO S. Conditional generative adversarial nets[EB/OL]. (2014-11-06)[2022-03-01]. https://arxiv.org/abs/1411.1784.

[5] DALAL N,TRIGGS B. Histograms of oriented gradients for human detection[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press,2005,1:886-893.

[6] CHEN J, LITTLE J J. Sports camera calibration via synthetic data[C]// Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. Piscataway: IEEE Press,2019:2497-2504.

[7] LUCAS B. An iterative image registration technique with an application to stereo vision (DARPA)[C]// Proceedings of DARPA Image Understanding Workshop,1981:121-130.

[8] HOMAYOUNFAR N,FIDLER S,URTASUN R. Sports field localization via deep structured models[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway:IEEE Press,2017:4012-4020.

[9] DETONE D,MALISIEWICZ T,RABINOVICH A. Deep image homography estimation[EB/OL]. (2016-01-13)[2022-03-01]. https://arxiv.org/abs/1606.03798.

[10] NGUYEN T,CHEN S W,SHIVAKUMAR S S,et al. Unsupervised deep homography:A fast and robust homography estimation model[J]. IEEE Robotics and Automation Letters,2018,3(3):2346-2353.

[11] JIANG W,HIGUERA J,ANGLES B,et al. Optimizing through learned errors for accurate sports field registration[C]// Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. Piscataway:IEEE Press,2020:201-210.

[12] CANNY J. Collision detection for moving polyhedra[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,1986,8(2):200-209.

[13] FENG N,SONG Z,YU J,et al. SSET:A dataset for shot segmentation,event detection,player tracking in soccer videos[J]. Multimedia Tools and Applications,2020,79(39):28971-28992.

[14] PUWEIN J,ZIEGLER R,VOGEL J,et al. Robust multi-view camera calibration for wide-baseline camera networks[C]// 2011 IEEE Workshop on Applications of Computer Vision (WACV). Piscataway:IEEE Press,2011:321-328.

[15] BOLME D S,BEVERIDGE J R,DRAPER B A,et al. Visual object tracking using adaptive correlation filters[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway:IEEE Press,2010:2544-2550.

# Player movement data analysis on soccer field reconstruction

JI Xiaoqi[1], SONG Zikai[1], YU Junqing[1,2,*]

(1. School of Computer Science and Technology, Huazhong University of Science and Technology, Wuhan 430074, China;
2. Center of Network and Computation, Huazhong University of Science and Technology, Wuhan 430074, China)

**Abstract**: Objective In soccer matches, player data analysis is crucial to improve the viewing experience for viewers and to aid coaches in performance evaluation. The difficulty of player data analysis is how to locate the coordinates of players on the soccer field, i.e., how to determine the mapping relationship between the defective field appearing in a frame of soccer video and the standard two-dimensional field. Aiming at how to deal with the high-speed movement of the camera and the sharp change of the angle of view in the soccer match, we designed and proposed a method of player motion analysis using field reconstruction and player tracking. For field reconstruction, the field in the soccer video is grouped into three parts: left, center, and right. Each group is mapped from the defective field to the standard field by soccer field segmentation, straight line detection, straight-line grouping, center circle point set identification, and key point matching; the kernelized correlation filter (KCF) tracking algorithm is used for player tracking. Then, using a combination of field reconstruction and player tracking approaches, we determine the standard coordinates of players and generate a set of player motion data and visualization results. The player data analysis method proposed in this paper can accurately and effectively count the player data, including player coordinates, motion trajectory, running speed, activity range, and player spacing. In terms of field reconstruction, image intersection is used for evaluation, and the intersection ratio of our algorithm reaches 87%, which improves 3.7% compared to the traditional dictionary-based reconstruction method (83.3% intersection ratio). The results of the experiments suggest that our field reconstruction method can more precisely depict the field mapping connection and can give greater assistance for the statistical analysis of player data. In this paper, we design and propose a complete algorithm for player data analysis based on soccer field reconstruction and obtain visualization results of player statistics. The soccer field reconstruction method combining the knowledge of soccer has improved in accuracy and efficiency. The player data analysis in this paper can provide data support for soccer fans and practitioners, the field reconstruction method lays a solid foundation for further research in the field of player analysis.

**Keywords**: soccer match; player data analysis; soccer field reconstruction; soccer field segmentation; player tracking; player motion detection

**Received**: 2022-03-09; **Accepted**: 2022-03-25; **Published online**: 2022-03-31 09:19
**URL**: kns.cnki.net/kcms/detail/11.2625.V.20220329.1820.001.html
**Foundation item**: National Key R & D Program of China (2020YFB1805601)
* **Corresponding author**. E-mail: yjqing@hust.edu.cn

# 《北京航空航天大学学报》征稿简则

《北京航空航天大学学报》是北京航空航天大学主办的以航空航天科学技术为特色的综合性自然科学学术期刊(月刊)。本刊以反映航空航天领域研究成果与动态、促进学术交流、培养科技人才和推动科技成果向社会生产力转化为办刊宗旨。本刊为中国自然科学技术核心期刊,并被 Ei Compendex 等国内外权威文献检索系统收录。本刊向国内外公开发行,为进一步提高办刊质量和所刊出文章的学术水平,特制定本简则。

## 1 论文作者及内容

**1.1** 本刊面向海内外所有学者。

**1.2** 主要刊载与航空航天科学技术有关的材料科学及工程、飞行器设计与制造、宇航科学与工程、信息与电子技术、控制技术和自动化工程、流体力学和动力工程、计算机科学及应用技术、可靠性工程与失效分析等领域的研究文章。航空航天科学技术民用方面以及具有航空航天工程背景的应用数学、应用物理、应用力学和工程管理等方面的文章也在本刊优先考虑之列。

## 2 来稿要求

**2.1** 论文应具有创新性、科学性、学术性和可读性。

**2.2** 论文为原创作品,尚未公开发表过,并且不涉及泄密问题。若发生侵权或泄密问题,一切责任由作者承担。

**2.3** 主题明确,数据可靠,图表清晰,逻辑严谨,文字精练,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,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,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,引言,正文,参考文献。首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,作者有权改投它刊。在此之前,请勿一稿多投,否则一切后果自负。

**3.3** 来稿一经刊登,即赠送单行本。

**3.4** 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 405,407,409 房间

**电 话**:(010)82315594,82338922,82314839,82315426

**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

**Beijing Hangkong Hangtian Daxue Xuebao**

（原《北京航空学院学报》）

（月刊　1956年创刊）

第48卷 第8期　2022年8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**　100083
**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) 82315594 82338922 82314839
**电子信箱**　jbuaa@buaa.edu.cn

# JOURNAL OF BEIJING UNIVERSITY OF AERONAUTICS AND ASTRONAUTICS (JBUAA)

(Monthly, Started in 1956)

Vol.48 No.8 August 2022

**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 100083, 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) 82315594 82338922 82314839
**E-mail** jbuaa@buaa.edu.cn
http://bhxb.buaa.edu.cn

中国标准连续出版物号：ISSN 1001-5965 / CN 11-2625/V　　国内定价：50.00元／期

ISSN 1001-5965