类感知对比学习的弱监督语义分割

白雪飞[1],许文杰[1],王渊辉[1],王文剑[2*]
(1. 山西大学计算机与信息技术学院,山西太原 030006;
2. 山西大学计算智能与中文信息处理教育部重点实验室,山西太原 030006)

摘 要: 图像级弱监督语义分割方法通常采用类激活图定位目标物体,但现有方法生成类激活图时存在目标区域激活不足或背景区域误激活等问题. 文章提出了一种类感知对比学习的弱监督语义分割框架,通过融合文本提示与图像类别信息,提升模型对目标区域的精确定位能力. 首先,文章分析了不同文本提示模板对各类别类激活图的影响,在此基础上,为了获取更具适应性的类别表示,本文构建了一个上下文提示集,并设计上下文提示动态选择策略,根据图像目标区域与文本提示之间的相似性获取最合适的上下文提示. 其次,采用图像-文本对比学习方法,以增强模型在处理图像与文本语义对齐任务中的表现,并设计了对比损失函数监督模型的训练过程. 最后,提出一个类别特定的背景抑制模块,抑制与目标类别紧密相关的背景区域的误激活,从而生成更加完整和紧凑的类激活图,实现更精确的语义分割. 文章在通用数据集PASCAL VOC 2012和MS COCO 2014中对提出的模型进行实验验证,mIoU值分别达到71.9%和43.9%,性能优于所有对比方法,有效提升了弱监督语义分割精度.

关键词: 弱监督语义分割;类激活图;类感知;对比学习;文本提示
基金项目: 国家自然科学基金(No.U21A20513,No.62476157);太行山西省实验室技术攻关专项资助项目(No.THYF-JSZX-24010200)
中图分类号: TP751 **文献标识码:** A **文章编号:** 0372-2112(2025)06-1741-14
电子学报URL:http://www.ejournal.org.cn **DOI:**10.12263/DZXB.20250024
第二十七届中国科协年会学术论文

Class-Aware Contrastive Learning for Weakly Supervised Semantic Segmentation

BAI Xue-fei[1], XU Wen-jie[1], WANG Yuan-hui[1], WANG Wen-jian[2*]
(1. *School of Computer and Information Technology*, *Shanxi University*, *Taiyuan*, *Shanxi* 030006, *China*;
2. *Key Laboratory of Computational Intelligence and Chinese Information Processing of Ministry of Education*, *Shanxi University*, *Taiyuan*, *Shanxi* 030006, *China*)

Abstract: In image-level weakly supervised semantic segmentation (WSSS), class activation map (CAM) are commonly used to localize object regions. However, existing methods often encounter challenges such as under-activation in object regions and erroneous activation in background regions when generating CAM. This paper proposes a class-aware contrastive learning (CA-CL) framework for weakly supervised semantic segmentation, which significantly enhances the model's ability to accurately localize object regions by integrating text prompts and image category information. Firstly, we analyze the influence of different text prompt templates on the class activation maps of various categories, on this basis, to obtain more adaptive class representations, we construct a contextual prompt set and design a dynamic contextual prompt selection strategy. This strategy generates the most appropriate contextual prompts based on the similarity between image object regions and text prompts. Secondly, we adopt an image-text contrastive learning approach to enhance the model's performance in aligning image and text semantics, and we design a contrastive loss function to guide the model training process. Finally, we introduce a class-specific background suppression module to mitigate erroneous activation in background regions closely related to object categories, thereby generating more complete and compact class activation maps and achieving more precise semantic segmentation. Experiments conducted on benchmark datasets PASCAL VOC 2012 and

收稿日期:2025-01-07;修回日期:2025-03-19;责任编辑:覃怀银
*通讯作者:王文剑

MS COCO 2014 demonstrate the effectiveness of the proposed framework, achieving mIoU values of 71.9% and 43.9%, respectively. The results demonstrate superior performance compared to existing methods, significantly improving the accuracy of weakly supervised semantic segmentation.

Key words: weakly supervised semantic segmentation; class activation maps; class-aware; contrastive learning; text prompt

Foundation Item(s): National Natural Science Foundation of China (No.U21A20513, No.62476157); Key Technologies Program of Taihang Laboratory in Shanxi Province (No.THYF-JSZX-24010200)

1 引言

图像语义分割是计算机视觉领域的核心任务之一,其目标是为图像提供像素级别的分类预测,即为每个像素确定所属的语义类别. 全监督语义分割方法依赖于完整的像素级标注,这种数据标注过程需要耗费大量的人力、时间和资金. 为了降低语义标注成本,提高分割效率,研究人员提出了弱监督语义分割(Weakly Supervised Semantic Segmentation, WSSS)方法,即通过有限的标注信息实现高质量的图像语义分割. 常用的弱监督语义标注包括图像级标注[1,2]、边界框标注[3,4]、涂鸦标注[5,6]和点标注[7]等,其中图像级标注由于成本较低,使用方便,受到了研究人员的广泛关注.

基于图像级标注的弱监督语义分割方法通常遵循以下步骤:首先将图像的类别标注作为监督信息,使用分类网络生成类激活图(Class Activation Map, CAM)[8],从而突出和定位图像中的目标物体;其次,通过条件随机场(Conditional Random Field, CRF)[9]或像素间关系网络(Inter-pixel Relations Network, IRNet)[10]等方法对CAM进行细化,生成伪标签;最后,利用生成的伪标签训练语义分割模型[11,12]. 然而,CAM倾向于突出对于分类结果影响最显著的图像区域,容易忽略其他有用的线索,导致生成的伪标签难以覆盖整个目标物体,甚至会导致与目标类别相关的背景区域被误激活(例如,火车图像中的轨道和站台),影响最终的分割结果. 为了解决这一问题,许多传统方法致力于引导网络关注更多的目标区域,如种子区域生长[13]、对抗擦除[14]和共现关系解耦[15]等方法,或者引入自监督学习、注意力机制[16]等,以生成更为完整的CAM. 尽管这些方法在一定程度上提升了分割网络的性能,但仍无法避免CAM目标区域激活不足和背景区域误激活等问题.

近年来,对比语言-图像预训练模型(Contrast Language-Image Pretraining, CLIP)[17]在零样本分类任务中展现出显著的性能. 在弱监督语义分割领域中,其有效性通过CLIP模型的研究[18~22]已被证实,如Xie等人[18]提出一种跨语言图像匹配框架CLIMS,利用包含类别信息的文本提示与图像目标区域之间的对应关系生成CAM. Murugesan等人[19]提出一种提示类学习策略,使用多个与图像类别标签密切相关的同义词来获得更好的文本提示,增强文本与图像的匹配效果. Lin等人[20]通过图像与文本提示匹配,直接从CLIP生成CAM,进一步提高分割性能. 尽管这些方法在一定程度上提升了分割效果,但它们仅采用单一的上下文提示模板,未能充分考虑不同类别在语义和视觉上的多样性,限制了模型对于不同类别特征的捕捉能力.

虽然CLIP模型在弱监督语义分割中有效缓解了传统方法中CAM目标区域激活不足的问题,但是由于目标物体与周围环境的复杂交互及类别间的语义重叠,模型在处理具有相似视觉特征但语义不同的区域时,容易将背景误识别为前景. 因此,如何减少背景区域的误激活也是目前结合CLIP模型的弱监督语义分割方法中的研究重点,一些研究通过构建通用的背景集以抑制所有类别的背景误激活,如CLIMS[18]对所有类别构建一个通用的背景集,使用固定的阈值对背景文本筛选后进行背景的抑制;CLIP-ES[20]中同样使用通用的背景类别集,将前景类别的文本特征与背景文本特征结合,使模型同时考虑前景的激活和背景的抑制. 然而,这种通用的背景集缺乏针对性,无法考虑到每个类别独特的背景特征,可能会导致背景抑制效果不足,或者错误地将一些属于目标的特征识别为背景特征,产生过度抑制,从而影响模型的分割效果.

针对上述问题,本文提出一种类感知对比学习(Class-Aware Contrastive Learning, CA-CL)的弱监督语义分割框架,综合使用文本提示和图像类别2种监督信息,提升模型对目标区域的定位能力,实现较为准确的语义分割,主要贡献如下.

首先,本文分析了不同上下文提示对于生成CAM的影响,在此基础上构建了一个上下文提示集,并提出一种上下文提示动态选择策略(Dynamic Context Prompt Selection, DCPS),旨在选择与图像目标区域最具相关性的上下文提示,将其与图像类别标签相融合,获得更加符合类别特性的文本提示.

其次,为了增强图像特征与文本提示之间的对应关系,采用图像-文本对比学习(Image-Text Contrastive Learning, ITCL)模块,以强化图像目标区域与对应文本之间的相关性,指导模型学习到更精确的特征表示.

最后,针对目标区域的背景误激活问题,引入类别特定的背景抑制模块(Category Specific Background

Suppression,CSBS),通过为每个类别构建与其紧密相关的背景文本,并与不同的上下文组合为背景文本提示,然后利用自适应阈值进行筛选,减少目标物体背景的误激活现象.

在PASCAL VOC 2012数据集[23]和MS COCO 2014数据集[24]上的大量实验结果表明,本文提出的方法优于其他同类方法,验证了其在弱监督语义分割任务中的有效性.

2　相关工作

2.1　现有的弱监督语义分割

现有的弱监督语义分割方法通常遵循3个阶段的学习过程:第1阶段,利用图像的类别标签作为监督信息,通过分类网络生成CAM,用于定位图像中的目标区域;第2阶段,通过CRF或IRNet等后处理方法对CAM进行细化,生成精确的伪标签;第3阶段,利用生成的伪标签训练语义分割模型.其中,如何生成更完整的CAM是WSSS的研究重点,研究人员提出了各种方法来提高CAM质量.Wei等人[14]提出一种对抗擦除的方法来挖掘目标物体的相关区域,通过迭代擦除当前已挖掘的区域,迫使分类网络不断发现新的潜在区域来改善目标物体区域的激活.Wang等人[16]提出自监督等变注意力机制,利用了图像仿射变换前后类激活图的一致性,同时引入像素相关性模块对类激活图进行约束.Zhang等人[25]提出一种互补补丁的方法,将图像划分为互补的两部分,分别进行激活,然后结合两部分的类激活图作为监督信息引导激活区域的扩展.这些方法在一定程度上提升了分割网络的性能,但仍存在目标区域激活不足和背景区域误激活等问题.Lee等人[26]强调了伪标签生成阶段阈值对于其生成的重要性,并提出激活操纵网络来优化阈值,提高模型对目标和背景的区分能力.Chen等人[27]针对共现背景误激活问题,提出了假阳性修正的方法,利用CAM区域中的背景线索作为目标类别的假阳性信息来指导模型训练.文献[28,29]通过结合显著图[30]来生成伪标签,然而,由于显著图缺乏物体类别信息,仍然会产生错误激活问题.此外,一些端到端的WSSS方法[31~33]将CAM生成、伪标签的细化和语义分割网络的训练集成在一个统一的框架内,避免了复杂的分阶段处理,简化了模型训练过程,但分割性能通常不及多阶段方法.

2.2　对比语言-图像预训练

CLIP由图像编码器和文本编码器组成[17],能够对图像和文本进行编码并测量其相似性.该模型通过从互联网上收集的4亿对图像-文本进行了预训练,将图像中更广泛的视觉概念与开放世界场景中相应的文本标签联系起来,成功应用于许多下游任务.在弱监督语义分割任务中,文献[18~22]利用CLIP,通过文本提示来引导模型生成更精细的CAM,如Xie等人[18]引入CLIP,提出一种跨语言图像匹配框架CLIMS,利用文本提示与图像目标区域之间的对应关系来生成CAM,但由于使用了单一的上下文提示模板和通用背景集,限制了CLIP的能力.Murugesan等人[19]在CLIMS的基础上,给定多个与图像类别标签密切相关的同义词,通过计算余弦相似度来生成更好的文本提示,从而进一步增强文本与图像的匹配效果,但仅考虑了修改类别标签对文本提示的影响.Lin等人[20]采用文本驱动策略,直接从CLIP生成CAM,并简化了CAM的细化阶段,但该模型难以处理一些复杂场景.Deng等人[21]提出一种问答式跨语言图像匹配框架QA-CLIMS,使用统一视觉语言理解与生成的预训练模型(Bootstrapping Language-Image Pretraining,BLIP)[34],通过问答的方式获取目标物体的前景和背景信息,然后结合CLIP模型生成CAM,但BLIP的使用会带来较高的计算成本.Jang等人[22]提出密集对齐学习网络DALNet,采用双层对齐策略获取目标全局和局部的特征,从而精确定位目标物体,但其使用了单一的上下文提示模板与固定的背景提示,这在复杂场景中表现不佳.以上工作均使用了单一的上下文提示模板,这限制了模型在处理不同场景时的适应性和泛化能力.

2.3　对比学习

对比学习方法的核心在于拉近相似样本之间的距离,同时推远不相似样本之间的距离,从而促使模型学习到具有高区分度的特征表示[35~37].在弱监督语义分割任务中,对比学习被应用于多个阶段:(1)在特征学习阶段,对比学习被用于预训练网络以提取更具区分度的特征,例如Xie等人[38]通过对比学习来增强前景和背景之间的分离效果,提出了类不可知激活图,改进了CAM的质量;Zhou等人[39]在区域感知表示中使用对比学习来增加相同类别的相似性,同时减少不同类别的相似性,提高区域之间的区分度.(2)在伪标签生成阶段,对比学习用于增强伪标签的质量,确保特征空间中类内距离近而类间距离远,例如Yuan等人[40]将伪标签中高置信度区域的特征作为正样本,同时将潜在的背景区域或其他类别的特征作为负样本,通过对比学习进一步细化伪标签.与这些工作不同,本文利用图像-文本对比学习来增强图像目标区域与文本提示之间的相关性,帮助模型更好地区分目标区域与背景区域.

3　本文方法

为了缓解CAM目标区域激活不足及背景误激活问题,本文提出一种CA-CL的弱监督语义分割框架,该框

架针对不同类别选择合适的文本提示，通过对比学习训练模型，同时进行背景抑制，从而生成更完整的CAM，提升分割性能．如图1所示，CA-CL主要包括5个部分：CAM生成、DCPS、ITCL、CSBS和分割．

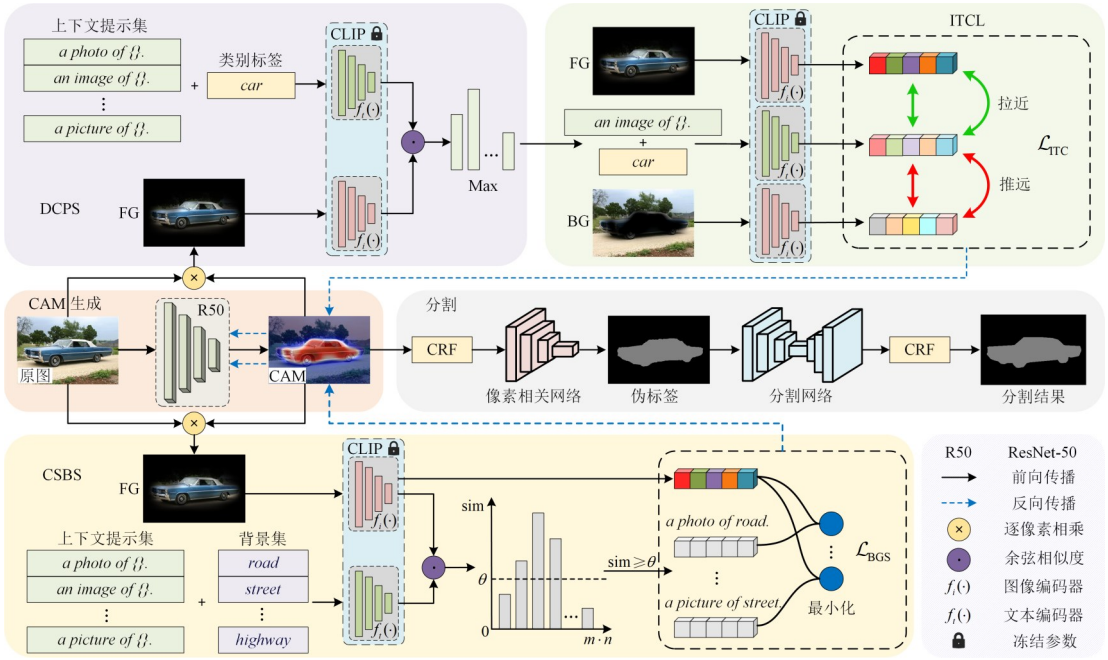


图1 类感知对比学习的弱监督语义分割框架

对于输入图像及其对应的图像级标签进行操作：(1)利用骨干网络ResNet-50[41]生成初始CAM，并与原图计算得到图像的目标前景区域和背景区域．(2)将上下文提示集中的提示与类别标签进行组合构建文本提示，与图像前景区域经CLIP编码后，通过DCPS，自动选择出与前景区域具有最高相关性的文本提示，对不同类别构建合适的文本表示．(3)通过ITCL模块处理图像特征与文本提示之间的对应关系，拉近目标文本提示与前景区域的距离，同时推远其与背景区域的距离，辅助模型学习到更准确的特征表示．(4)在CSBS中，结合上下文提示集和背景集生成背景提示，与前景区域进行相似度计算，找出相似度大于自适应阈值的背景提示来抑制背景区域，减少目标物体背景的误激活现象．通过以上步骤，为每个目标类别生成了高质量的CAM．对得到的CAM采用CRF修正后训练IRNet，进一步扩展激活区域得到精细的伪标签．最后训练分割网络获得预测结果并通过CRF修正得到最终的语义分割结果．

3.1 本文的弱监督语义分割

本文使用预训练的ResNet-50作为生成CAM的骨干网络，具体步骤如下：给定输入图像$\boldsymbol{X}$及其对应的图像级标签$k\in\mathbb{R}^{1\times K}$，首先提取输入图像$\boldsymbol{X}$的特征映射$\boldsymbol{Z}\in\mathbb{R}^{C\times H\times W}$，其中$K$表示类别数量，$C$、$H$和$W$分别表示图像的通道数、高和宽．然后在网络分类层对特征映射和可学习矩阵$\boldsymbol{W}\in\mathbb{R}^{C\times K}$应用Sigmoid函数如下：

$$\boldsymbol{P}_k(h,w)=\sigma\left(\boldsymbol{W}_k^{\mathrm{T}}\boldsymbol{Z}(h,w)\right) \tag{1}$$

其中，$\boldsymbol{P}_k\in\mathbb{R}^{K\times H\times W}$为图像$\boldsymbol{X}$对于类别$k$的初始CAM，将在后续处理流程中细化，$\sigma$为Sigmoid函数，$\boldsymbol{Z}(h,w)$为图像$\boldsymbol{X}$在$(h,w)$上的特征表示．

使用骨干网络生成初始类激活图$\boldsymbol{P}_k$后，将输入图像$\boldsymbol{X}$和$\boldsymbol{P}_k$进行逐像素相乘，从而得到图像的目标前景区域图像FG及背景区域图像BG如下：

$$\begin{cases}\mathbf{FG}=(\boldsymbol{X}\cdot\boldsymbol{P}_k)\\ \mathbf{BG}=\left(\boldsymbol{X}\cdot(1-\boldsymbol{P}_k)\right)\end{cases} \tag{2}$$

3.2 上下文提示动态选择策略

CLIP模型文本编码器的标准输入提示具有如下格式：上下文提示[CTX]加上类别名称[CLS]，并以标点符号“.”结尾[17]．在以往引入CLIP模型的弱监督语义分割方法[18~22]中，采用的都是单一上下文提示语句，这些提示语句并不一致，例如CLIMS[18]中使用了“*a photo of {}.*”，CLIP-ES[20]中使用了“*a clean origami {}.*”．为了探讨不同上下文提示在弱监督语义分割任务中的作用，本文在不同的前景类别上使用多种上下文提示进行实验，重点关注不同上下文提示对于不同类别CAM的影响差异，实验结果如图2和图3所示．

分析图2可以看出，不同的上下文提示语句对于各类别CAM的准确性存在显著差异，例如：对于船类别，文本提示语句“*a picture of {}.*”对比其他上下文能够获得更完整的CAM，而“*an image of {}.*”在该类别上表现

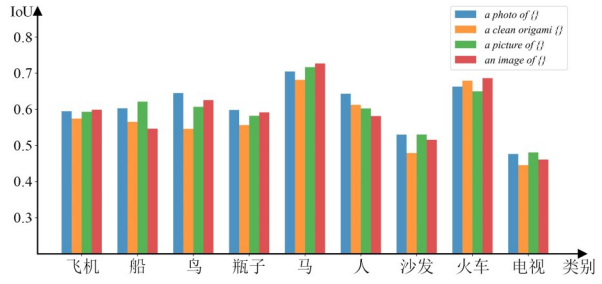


图2　不同上下文提示下各类别CAM的交并比(IoU)

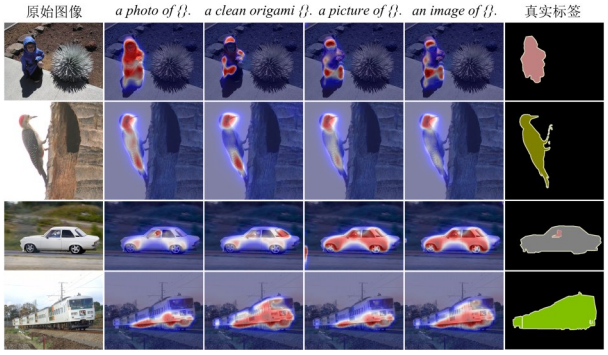


图3　不同上下文提示生成CAM的可视化对比

较差;同样,对于人类别,"*a photo of {}.*"的效果明显优于其他提示语句. 图3展示了不同的上下文提示在各类别上生成的初始CAM的可视化结果,前2行中,"*a photo of {}.*"在人和鸟的类别上都生成了较为精确的CAM;第3行中,"*a picture of {}.*"和"*an image of {}.*"在激活汽车类别时表现更优;最后1行中,"*a clean origami {}.*"和"*an image of {}.*"为火车类别生成了较为完整的CAM. 这些现象表明,不同的上下文提示对于不同类别CAM的激活完整程度具有显著影响. 因此,在弱监督语义分割方法中引入CLIP模型时,上下文提示不应局限于使用单一的提示模板,而是应当根据类别的特性动态选择上下文提示,以充分优化分割效果.

为了使每个类别获得更具适应性的文本提示,从而更好地关联图像特征,本文创建了一个包含m个句子的上下文提示集[CTX-Set],旨在根据类别的特性动态调整文本提示,以生成更为完整的CAM. 基于文献[18~22]的研究成果,本文在构建上下文提示集[CTX-Set]时,首先收集了一些常用的文本模板,如"*a photo of {}.*"和"*a clean origami {}.*",这些模板在WSSS任务中表现出良好的稳定性. 其次,CLIP模型的训练数据来源于互联网上的图像-文本对,这些文本涵盖了日常生活场景中的多种语言描述,为进一步增强与CLIP模型训练语料的适配性,上下文提示集[CTX-Set]中加入了生活中常见的描述图像的上下文,如"*a picture of {}.*""*an image of {}.*"等. 这种多样化的上下文提示能够使模型根据不同类别进行动态调整,不仅增强了模型的泛化能力,还降低了模型对于单一提示模板的依赖性.

在此基础上,本文提出一种DCPS,使用上下文提示集[CTX-Set]中不同上下文与类别标签组合,生成文本提示,根据其与图像前景区域的匹配程度,自动选择出最合适的上下文提示. 具体来讲,使用CLIP中的文本编码器$f_t(\cdot)$和图像编码器$f_i(\cdot)$将文本提示和图像前景区域**FG**映射到公共特征空间. 其中,**FG**经过$f_i(\cdot)$编码后,得到图像表示向量$\boldsymbol{v}_k^{i,f}$;包含m个句子的上下文提示集与类别标签k组合后,生成m个文本提示,通过$f_t(\cdot)$编码后,得到文本表示向量$\boldsymbol{v}_{k,j}^{t,f}$如下:

$$\begin{cases}\boldsymbol{v}_k^{i,f}=f_i(\mathbf{FG})\\ \boldsymbol{v}_{k,j}^{t,f}=f_t(T_{k,j}^f)\end{cases}\tag{3}$$

其中,$T_{k,j}^f$表示类别k的第j个文本提示,$j\in\{1,2,3,\cdots,m\}$. 通过计算二者的余弦相似度寻找与**FG**相关性最高的上下文提示语句,余弦相似度的计算如下:

$$\mathrm{sim}(\boldsymbol{v}_k^{i,f},\boldsymbol{v}_{k,j}^{t,f})=\frac{\boldsymbol{v}_k^{i,f}\cdot\boldsymbol{v}_{k,j}^{t,f}}{\left|\boldsymbol{v}_k^{i,f}\right|\times\left|\boldsymbol{v}_{k,j}^{t,f}\right|}\tag{4}$$

其中,"·"表示向量之间的点积运算,"×"表示2个向量模长的乘积. 式(4)计算得到一个长度为m的相似度向量,该向量中的每个元素表示相应文本表示向量与图像表示向量之间的相似程度. 通过选取该向量中的最大值,可以确定最合适的上下文提示,其对应的文本提示与**FG**具有最高的相似度,表明该文本提示与图像目标区域的关联性最强.

3.3　图像-文本对比学习

为了增强图像目标区域与对应文本提示之间的相关性,提高模型区分前景与背景的能力,从而更准确地激活目标区域,本文采用图像-文本对比学习方法.

通过DCPS获得与类别k相似度最高的文本提示T_k^f后,将该文本提示通过CLIP的文本编码器$f_t(\cdot)$编码后得到表示向量$\boldsymbol{v}_k^{t,f}$,将图像前景区域**FG**和图像背景区域**BG**分别通过CLIP的图像编码器$f_i(\cdot)$编码,得到表示向量$\boldsymbol{v}_k^{i,f}$和$\boldsymbol{v}_k^{i,b}$如下:

$$\begin{cases}\boldsymbol{v}_k^{t,f}=f_t(T_k^f)\\ \boldsymbol{v}_k^{i,f}=f_i(\mathbf{FG})\\ \boldsymbol{v}_k^{i,b}=f_i(\mathbf{BG})\end{cases}\tag{5}$$

对于图像语义分割问题而言,图像中类别k的前景区域**FG**表示向量$\boldsymbol{v}_k^{i,f}$,应该与其前景文本表示向量$\boldsymbol{v}_k^{t,f}$有较高的相似性,而目标背景区域**BG**表示向量$\boldsymbol{v}_k^{i,b}$,应该与文本表示向量$\boldsymbol{v}_k^{t,f}$有较大的差异性. 因此,本文在图像-文本对比学习中,对于文本提示$\boldsymbol{v}_k^{t,f}$,将目标前景区域**FG**表示向量$\boldsymbol{v}_k^{i,f}$定义为正样本,将目标背景区域**BG**表示向量$\boldsymbol{v}_k^{i,b}$定义为负样本. 为了使模型在训练过程中拉近正样本之间的距离,推远负样本之间的距离,本文设计对比损失函数如下:

$$\mathcal{L}_{\mathrm{ITC}}=\frac{1}{N}\sum_{i=1}^{N}\max(0,\ \mathrm{margin}-s_{k,i}^{f,f}+s_{k,i}^{b,f}) \tag{6}$$

$$\begin{cases} s_k^{f,f}=\mathrm{sim}(\boldsymbol{v}_k^{i,f},\boldsymbol{v}_k^{t,f}) \\ s_k^{b,f}=\mathrm{sim}(\boldsymbol{v}_k^{i,b},\boldsymbol{v}_k^{t,f}) \end{cases} \tag{7}$$

其中，N为样本数；margin为正负样本余弦相似度（$s_k^{f,f}$和$s_k^{b,f}$）之间的最小差距，确保模型在训练过程中能够有效区分正负样本. 在ITCL中，通过最小化损失函数$\mathcal{L}_{\mathrm{ITC}}$，模型能够拉近图像前景区域与对应前景文本提示之间的距离，推远背景区域与前景文本提示之间的距离，从而使模型更有效地捕捉目标物体的特征，使得生成的CAM逐渐接近目标物体.

3.4 类别特定背景抑制

尽管DCPS和ITCL能够生成相对完整的CAM，但在一些复杂场景中，目标物体与背景可能还存在一定的相似性，此时仅依靠前景类别的文本提示来定位目标对象，可能会导致背景误激活现象. 因此，本文设计了一个CSBS，综合考虑图像前景与背景之间的特征，旨在有效减少背景的误激活现象，使得模型能够更准确地区分复杂场景中的前景和背景，从而生成更精确的CAM.

为了实现精准的背景抑制，本文为每个类别设计了一个背景集[BGD-Set]，其中包含了n个可能与该类别共同出现的背景单词. 具体而言，对于[CLS]中的每个类别k，将以下查询语句作为ChatGPT[42]的输入："列出n个可能与类别k共同出现的背景的英文单词"，ChatGPT将返回一个包含n个背景单词的列表，这样，每个类别都能获得n个相关的背景元素.

由于每张图像中的背景元素数量不固定，[BGD-Set]中可能存在冗余信息，这些与图像特征无关的背景文本可能会对模型的性能产生影响. 因此，本文提出自适应阈值筛选方法，根据背景文本与图像前景区域之间的相似度，自动筛选出与目标区域相关性较高的背景文本进行背景抑制，从而有效地减少背景区域的误激活.

对于图像$\mathbf{X}$中的类别k，将其前景区域$\mathbf{FG}$经过$f_i(\cdot)$编码后，得到图像前景区域的表示向量$\boldsymbol{v}_k^{i,f}$. 同时，使用包含m个上下文提示的[CTX-Set]与包含n个背景文本的[BGD-Set]组合，生成$m\cdot n$个背景文本提示，通过$f_t(\cdot)$编码后，得到每个背景文本的表示向量$\boldsymbol{v}_{k,l}^{t,b}$如下：

$$\begin{cases} \boldsymbol{v}_k^{i,f}=f_i(\mathbf{FG}) \\ \boldsymbol{v}_{k,l}^{t,b}=f_t(T_{k,l}^{b}) \end{cases} \tag{8}$$

其中，$T_{k,l}^{b}$表示类别k的第l个背景文本提示，$l\in\{1,2,3,\cdots,m\cdot n\}$. 得到背景文本提示后，对其进行自适应阈值筛选. 首先，计算图像前景区域$\boldsymbol{v}_k^{i,f}$与每个背景文本提示$\boldsymbol{v}_{k,l}^{t,b}$之间的余弦相似度，评估每个背景文本与前景区域的相似性：

$$\mathrm{sim}(\boldsymbol{v}_k^{i,f},\boldsymbol{v}_{k,l}^{t,b})=\frac{\boldsymbol{v}_k^{i,f}\cdot\boldsymbol{v}_{k,l}^{t,b}}{\left|\boldsymbol{v}_k^{i,f}\right|\times\left|\boldsymbol{v}_{k,l}^{t,b}\right|} \tag{9}$$

上式计算得到一个长度为m的相似度向量，该向量中每个元素表示相应背景文本提示与图像前景区域的相似程度. 然后，为阈值θ设定一个初始值，根据θ筛选出与前景区域相似度大于该阈值的背景文本提示，这些提示被认为是可能在前景区域中出现的背景：

$$t_l=\begin{cases} 1, & \mathrm{sim}(\boldsymbol{v}_k^{i,f},\boldsymbol{v}_{k,l}^{t,b})\geqslant\theta \\ 0, & \mathrm{sim}(\boldsymbol{v}_k^{i,f},\boldsymbol{v}_{k,l}^{t,b})<\theta \end{cases} \tag{10}$$

其中，$t_l\in\{0,1\}$，用于指示图像前景区域与背景文本的相似度是否超过阈值，从而决定是否将该背景文本作为有效的背景进行抑制. 阈值θ在训练过程中根据背景抑制损失函数$\mathcal{L}_{\mathrm{BGS}}$进行动态更新，使其能够根据图像内容和类别特征自适应调整：

$$\theta'=\theta-\eta\frac{\partial\mathcal{L}_{\mathrm{BGS}}}{\partial\theta} \tag{11}$$

其中，θ为当前阈值；θ'为更新后的阈值；η为学习率；$\frac{\partial\mathcal{L}_{\mathrm{BGS}}}{\partial\theta}$为损失函数$\mathcal{L}_{\mathrm{BGS}}$关于$\theta$的梯度. 利用自适应阈值筛选出的背景文本进行背景抑制，可以去除那些前景区域中误激活的背景区域，有效缓解背景误激活问题. 背景抑制损失函数$\mathcal{L}_{\mathrm{BGS}}$定义为

$$\mathcal{L}_{\mathrm{BGS}}=-\sum_{k=1}^{K}\sum_{l=1}^{m\cdot n}y_k\cdot\log\left(1-\mathrm{sim}(\boldsymbol{v}_k^{i,f},\boldsymbol{v}_{k,l}^{t,b})\cdot t_l\right) \tag{12}$$

其中，y_k为类别k的真实标签，$t_l\in\{0,1\}$. 在模型训练过程中，通过最小化损失函数$\mathcal{L}_{\mathrm{BGS}}$，骨干网络会逐渐减少CAM中与类别$k$相关的背景区域的误激活，从而生成较为准确的CAM，提升语义分割的性能.

3.5 语义分割

经过DCPS、ITCL和CSBS的共同优化后，模型能够为每个类别生成高质量的CAM. 生成CAM后，首先利用CRF消除噪声并精细化边界得到伪标签. 然后使用伪标签训练IRNet，进一步扩展和优化激活区域，从而获得更加准确的伪标签. 最后，使用这些伪标签训练分割网络，并通过CRF对结果进行修正，得到最终的分割结果.

3.6 模型训练目标

为了进一步优化类激活图的质量，本文设计了一个像素级的区域正则化项$\mathcal{L}_{\mathrm{REG}}$，用于约束类激活图的大小，确保模型准确定位目标区域，防止过度激活无关区域：

$$\mathcal{L}_{\mathrm{REG}}=\frac{1}{KHW}\sum_{k=1}^{K}\sum_{h=1}^{H}\sum_{w=1}^{W}\mathbf{P}_k(h,w) \tag{13}$$

因此，在CA-CL训练过程中共使用了3个损失函数，包括$\mathcal{L}_{\mathrm{ITC}}$、$\mathcal{L}_{\mathrm{BGS}}$和$\mathcal{L}_{\mathrm{REG}}$. 模型的总体训练目标为

$$\mathcal{L}=\alpha\cdot\mathcal{L}_{\mathrm{ITC}}+\beta\cdot\mathcal{L}_{\mathrm{BGS}}+\gamma\cdot\mathcal{L}_{\mathrm{REG}} \tag{14}$$

其中,超参数α、β和γ为可调整的损失权重,旨在确保模型在训练过程中能够充分利用文本提示信息,准确激活目标区域,并有效抑制背景区域的误激活,生成更加完整的CAM,从而提升语义分割精度.

4　实验

4.1　数据集和评估指标

本文在PASCAL VOC 2012[23]和MS COCO 2014[24]数据集上评估了所提出的CA-CL框架. 其中,PASCAL VOC 2012包含20个目标类别和1个背景类别,训练集、验证集和测试集分别包含1 464、1 449和1 456张图像. 本文遵循文献[18~22]的设定,采用10 582张图像的增强训练集训练模型. 由于PASCAL VOC 2012数据集不公开测试集真值,本文将测试集结果提交到官方评估服务器评估其性能. MS COCO 2014数据集中含有80个目标类别和1个背景类别,训练集和验证集分别包含82 783和40 504张图像. 本文所有实验均采用平均交并比(mean Intersection over Union, mIoU)作为评价指标.

4.2　实验细节

本文使用ResNet-50作为骨干网络,输入图像被随机重新缩放,并通过随机裁剪到512×512,以及水平翻转来增强,采用SGD作为默认优化器,学习率的调度使用余弦退火策略,批处理大小为16,Epoch设置为20,初始学习率为0.000 25,权重衰减为0.000 1. 损失函数$\mathcal{L}_{\mathrm{ITC}}$中margin的值设为0.5. 模型总体损失函数中超参数α、β和γ分别设置为40、1和0.4. 上下文提示集[CTX-Set]中选取4句上下文提示(m=4). 背景集[BGD-Set]中每个类别包含了5个可能与该类别共同出现的背景单词(n=5). 在MS COCO 2014数据集的实验中,设置保持不变.

生成CAM后,本文采用CRF修正得到伪标签,并使用IRNet进一步细化伪标签. 在语义分割阶段,本文采用基于ResNet-101[41]的DeepLabV2[2]作为分割网络,将细化后的伪标签作为真实标签来训练分割网络DeepLabV2. 在PASCAL VOC 2012数据集上进行分割实验时,使用MS COCO数据集预训练的权重,批处理大小为10,初始学习率设置为0.005,采用SGD优化器,动量和权重衰减分别为0.9和0.000 5,共进行40 000次迭代训练. 在MS COCO 2014数据集上进行分割实验时,使用ImageNet[43]数据集预训练的权重,共进行100 000次迭代训练,其他设置保持不变.

本文实验均基于PyTorch框架实现,在配备40 GB显存的NVIDIA A100 GPU上完成训练.

4.3　实验结果

本文对比了本文方法与其他方法生成的CAM和伪标签质量,图4展示了CAM的可视化结果,表1列出了结果对比.

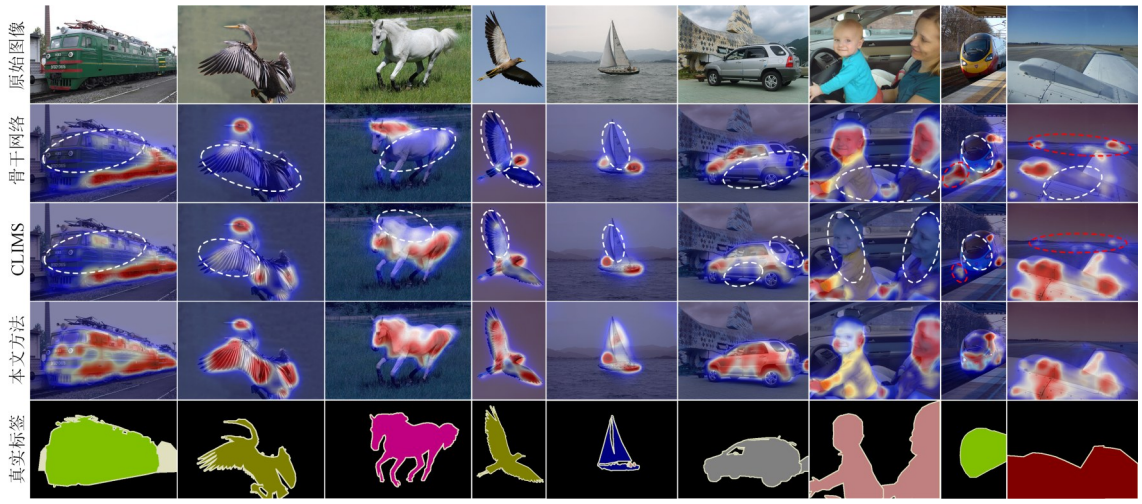


图4　不同方法在PASCAL VOC 2012训练集上生成CAM的可视化结果

图4展示了本文方法、骨干网络以及CLIMS模型生成CAM的可视化对比结果,其中,白色虚线圈表示未能激活的目标区域,红色虚线圈表示与类相关的背景区域的误激活. 实验结果表明,本文方法能够激活更完整的目标区域,同时有效抑制周围背景区域. 具体来说,在图4的前7列中,骨干网络和CLIMS对火车、鸟、马和船等区域激活不足,不能完整地覆盖目标区域,尤其是对于船类别,这2种方法都未能激活图中的船帆区域. 与之相比,本文提出的CA-CL所生成的CAM更加完整紧凑,包含了更多合理的目标区域. 分析图4最后2列可以看出,骨干网络和CLIMS不仅存在目标区域激活不足的情况,还存在与类别相关的背景区域的误激活. 然而,本文所提出的CA-CL方法,通过DCPS为每幅图像生成合适的文本提示,利用ITCL准确激活目标区域,

同时在CSBS的监督下有效地减少背景区域的误激活，生成了更加完整的CAM.

表1列出了本文方法与其他方法在PASCAL VOC 2012训练集上生成CAM和伪标签的实验结果，本文提出方法所生成CAM的mIoU值达到64.4%，明显优于CLIMS[18]和POLE[19]等方法. 此外，本文在实验中对生成伪标签的性能进行了比较，结果如表1最后1列所示，本文方法生成的CAM经过细化后，生成伪标签的mIoU值达到76.0%，相比DALNet[22]方法提高了0.8个百分点，相较于POLE方法提高了1.8个百分点，在所有对比方法中表现最佳，充分验证了本文方法的有效性.

表1 不同方法在PASCAL VOC 2012训练集上的结果对比

方法	骨干网络	mIoU/%	
		CAM	伪标签
MCTformer(CVPR2022)[44]	ViT-B	61.7	69.1
CLIMS(CVPR2022)[18]	ResNet-50	56.6	70.5
W-OoD(CVPR2022)[45]	ResNet-50	59.1	72.1
AMN(CVPR2022)[26]	ResNet-50	62.1	72.2
FPR(ICCV2023)[27]	ResNet-50	63.8	66.4
D2CAM(ICCV2023)[46]	ResNet-50	58.0	71.4
ToCo(CVPR2023)[32]	ViT-B	—	73.6
APC(EAAI2024)[33]	ViT-B	—	74.6
TKP-PCL(SMC2024)[47]	ViT-B	—	74.6
POLE(WACV2024)[19]	ResNet-50	59.0	74.2
DALNet(ECCV2024)[22]	ViT-B	—	75.2
本文方法	ResNet-50	64.4	76.0

为了进一步评估分割结果的质量，本文使用生成的伪标签，在基于ResNet-101的DeepLabV2网络上训练，得到最终的分割模型，并在PASCAL VOC 2012数据集的验证集和测试集上进行验证. 如图5所示，展示了本文方法在验证集上部分分割图的可视化结果，并将其与CLIMS和CLIP-ES[20]的结果比较，结果表明，本文提出的CA-CL能够更准确地分割目标区域，同时有效区分目标物体周围的背景区域. 具体来说，在前5列比较简单的场景中，CLIMS和CLIP-ES对于火车、飞机、人和马等区域中都存在目标区域分割不足的情况，而且还将背景区域错误地识别为前景对象. 与之相比，CA-CL的分割结果中目标区域更加完整，并且背景区域也能够得到较好的分割. 在图5后4列较复杂的场景中，CLIMS和CLIP-ES的分割结果都出现了目标区域分割不准确、前景对象边界不清晰和背景误分割等问题，例如倒数第3列，CLIMS和CLIP-ES方法将桌上的其他物品错误地预测为瓶子，倒数第2列的图像中，人物的腿部区域不能被正确地分割出来，与之相比，CA-CL能够在一些复杂场景下获得更准确的分割图，边界也比较清晰. 综上所述，本文所提出的CA-CL在不同类别及复杂场景中都取得了较好的分割效果，优于所有对比方法.

表2列出了本文方法在PASCAL VOC 2012验证集上的语义分割结果. 其中I表示图像级监督，I+L表示在使用图像级标签外，还额外使用了文本提示. 实验结果表明，本文提出的CA-CL方法在验证集和测试集上的mIoU值分别达到71.8%和71.9%，在验证集上，相比POLE提高了0.3个百分点，比DALNet提高了0.4个百分点；在测试集上，相比CLIMS高出1.9个百分点，比CLIP-ES、POLE和DALNet均高出0.5个百分点，优于所有对比方法，获得了较好的弱监督语义分割精度.

为了进一步验证本文方法对于多目标和小尺寸目标的分割性能，本文在MS COCO 2014数据集上进行了相关实验. 与PASCAL VOC 2012相比，该数据集类别更为丰富，包含大量具有多个目标类别的图像，对模型的性能要求更高. 实验结果如表3所示，CA-CL在MS COCO 2014验证集上的mIoU值达到了43.9%，比同为文本提示的方法QA-CLIMS和DALNet分别高出0.7和

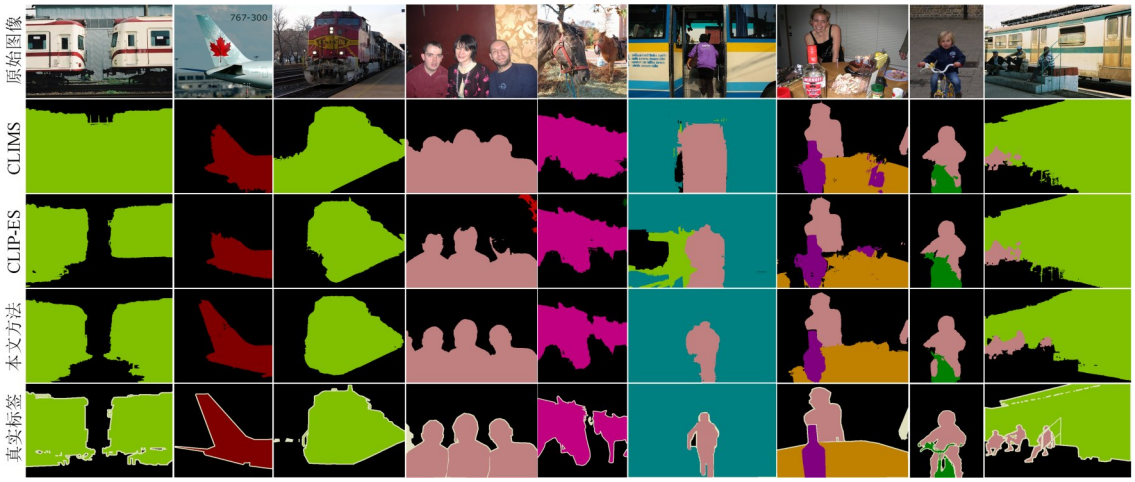


图5 不同方法在PASCAL VOC 2012验证集上的可视化结果

表2　不同方法在PASCAL VOC 2012验证集上的结果对比

监督	方法	mIoU/%	
		验证集	测试集
I	MCTfomer(CVPR2022)[44]	61.7	69.1
I	AMN(CVPR2022)[26]	70.7	70.6
I	VWL(IJCV2022)[48]	70.6	70.7
I	SIPE(CVPR2022)[49]	68.2	69.5
I	FPR(ICCV2023)[27]	70.3	70.1
I	D2CAM(ICCV 2023)[46]	71.2	70.7
I	ToCo(EAAI 2023)[33]	69.8	70.5
I	LPCAM(CVPR2023)[50]	70.1	70.4
I	SMA(WACV2024)[37]	70.9	70.8
I	DSCNet(CVPR2024)[51]	70.3	71.1
I	MCC(WACV2024)[52]	70.3	71.2
I	SFC(AAAI2024)[53]	70.2	71.4
I+L	CLIMS(CVPR2022)[18]	70.4	70.0
I+L	CLIP-ES(CVPR2023)[20]	71.1	71.4
I+L	POLE(WACV2024)[19]	71.5	71.4
I+L	DALNet(ECCV2024)[22]	71.4	71.4
I+L	本文方法	71.8	71.9

1.2个百分点,优于所有对比方法,充分验证了本文所提方法的有效性.

4.4　消融实验

本文对设计的3个模块进行了消融实验,包括DCPS、ITCL和CSBS,每个模块在引导骨干网络生成CAM的过程中发挥不同的作用. 随后,本文探讨了每个损失函数的影响,评估其在训练过程中的优化效果. 此外,本文还分析了上下文提示集[CTX-Set]中上下文提示模板的选取和背景集[BGD-Set]中各类别背景词的数量配置. 所有消融实验均在PASCAL VOC 2012训练集上进行.

表3　不同方法在MS COCO 2014验证集上的结果对比

监督	方法	mIoU/%
I	SIPE(CVPR2022)[49]	40.6
I	MCTformer(CVPR2022)[44]	42.0
I	ToCo(CVPR2023)[32]	42.3
I	LPCAM(CVPR2023)[50]	42.8
I	MCC(WACV2024)[52]	42.3
I+L	QA-CLIMS(ACM2023)[21]	43.2
I+L	DALNet(ECCV2024)[22]	42.7
I+L	本文方法	43.9

4.4.1　上下文提示动态选择策略的影响

本文使用DCPS获取的上下文提示与多种不同的单一上下文提示,包括"*a photo of {}.*""*a clean origami {}.*""*a picture of {}.*""*an image of {}.*"和"*a snapshot of {}.*",分别对CA-CL进行训练,并对比其可视化结果. 如图6所示,不同上下文提示对不同类别的CAM目标区域的激活程度不同,例如第1行使用"*an image of {}.*"作为上下文生成的文本提示对火车类别的激活程度较好,第2行使用"*a photo of {}.*"为人类别生成了更完整的CAM,而本文设计的DCPS能够根据不同类别动态选择最合适的上下文生成文本提示,从而激活更加完整的目标区域,生成较为完整的CAM. 本文还比较了不同单一上下文提示与DCPS生成CAM的mIoU值,结果如表4所示,可以看出,采用DCPS生成CAM的mIoU值达到了64.4%,在性能上显著优于单一的文本提示,能够有效地引导骨干网络生成更完整的CAM.

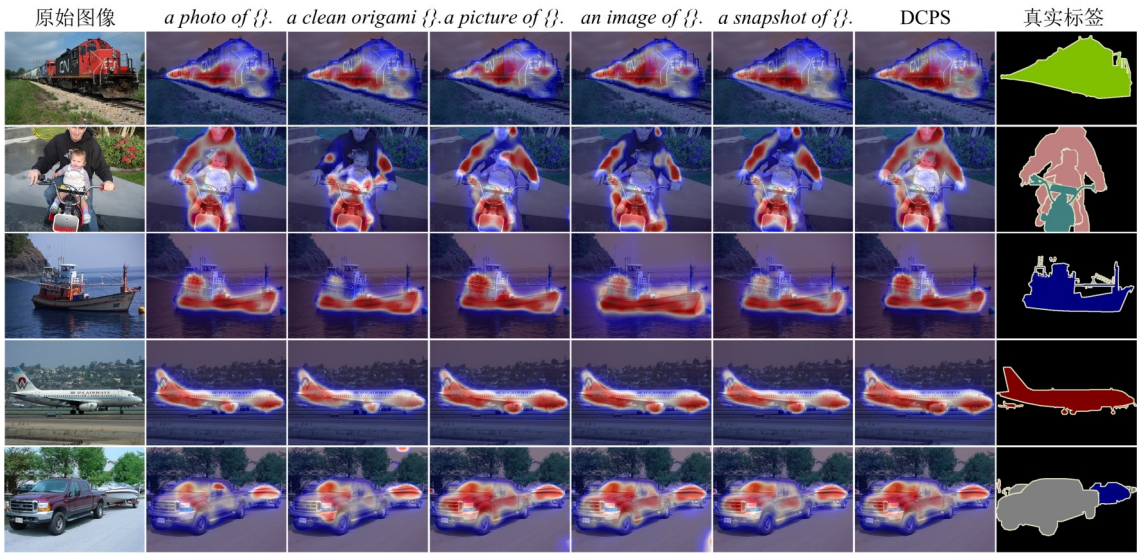


图6　不同上下文提示与DCPS所生成CAM的可视化结果

表4 不同上下文提示与DCPS生成CAM的结果对比

上下文提示	mIoU/%
a photo of {}.	62.8
a clean origami {}.	59.8
an image of {}.	62.7
a picture of {}.	61.9
a snapshot of {}.	60.3
DCPS	64.4

4.4.2 图像-文本对比学习的效果

在CA-CL框架中,利用DCPS生成的文本提示与图像前景区域和背景区域进行对比学习.通过最大化前景区域与其对应文本的相似度,同时最小化背景区域与前景文本的相似度,生成的CAM能够逐渐接近目标物体.本文使用DCPS和ITCL组合方法实验,如图7第2行所示,使用DCPS+ITCL方法生成的CAM能够较为完整地覆盖目标物体,但是部分场景还存在背景区域误激活现象.实验结果如表5所示,使用DCPS和ITCL组合生成CAM的mIoU值达到58.4%,证明了该方法在前景目标区域定位上的有效性.

4.4.3 类别特定背景抑制模块的作用

尽管DCPS+ITCL方法能够确保CAM较为完整地覆盖目标物体,但其没有充分考虑类相关背景的误激活问题,背景区域的误激活会产生不必要的噪声,影响分割精度.为了解决这一问题,本文引入了CSBS,通过构建类别特定的背景集,结合自适应阈值筛选背景文本提示,有效减少了背景区域的误激活,结合背景抑制损失函数,CSBS能够显著提高CAM的完整性,有效抑制背景干扰.加入背景抑制模块生成CAM的可视化结果如图7第3行所示,在DCPS+ITCL的基础上引入CSBS,生成的CAM显著减少了背景区域的误激活,例如,在前4列中,采用DCPS+ITCL方法的船、汽车和火车等目标在激活时通常伴随着背景区域的误激活,而引入CSBS后,能够有效抑制这些背景区域的误激活.图7的后4列进一步展示了CSBS在复杂场景或多目标图像中的背景抑制效果,实验结果表明,即使在复杂场景中,CSBS仍能显著减少背景误激活,帮助生成更精确的CAM.例如,在包含汽车、自行车、火车和人等多目标图像中,CSBS均能有效抑制背景误激活,提升目标区域的定位精度.表5展示了引入CSBS后生成CAM的mIoU值,达到64.4%,性能得到显著提升.综上,CSBS有效地减少了背景区域的误激活,生成了更加准确和完整的CAM.

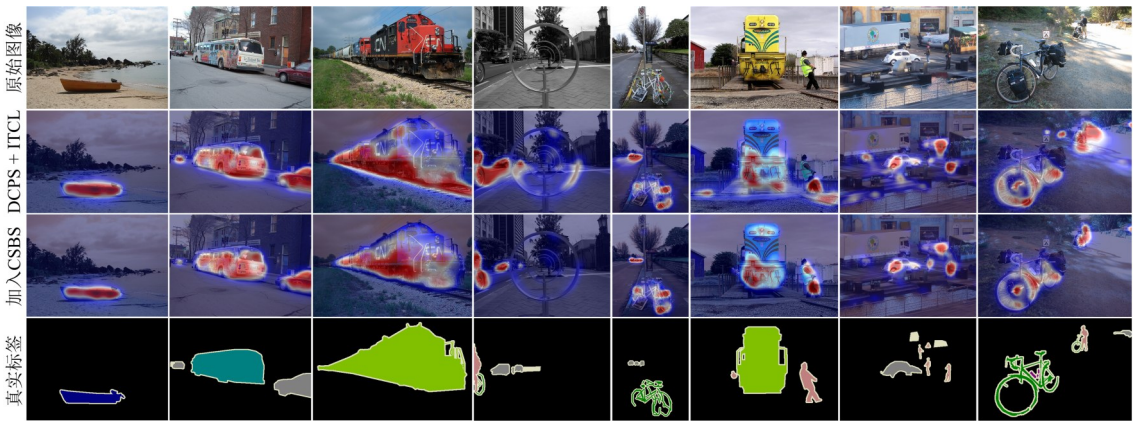


图7 不同方法生成CAM的可视化结果

表5 不同方法生成CAM的结果对比

DCPS+ITCL	加入CSBS	mIoU/%
√	—	58.4
√	√	64.4

4.4.4 损失函数的影响

本文评估了不同损失函数对模型性能的影响,实验结果如表6所示,在仅使用对比学习损失时,生成CAM的mIoU值为57.9%;在结合对比学习损失和区域正则化损失时,mIoU值提升至58.4%;在结合对比学习损失和背景抑制损失时,CAM的质量进一步提高,mIoU值达到62.7%;最终,在结合所有损失函数时,生成的CAM达到最佳效果,mIoU值为64.4%.以上结果表明,每个损失函数都能够有效提升CAM的质量.

表6 不同损失函数组合生成CAM的结果对比

对比学习损失	背景抑制损失	区域正则化损失	mIoU/%
√	—	—	57.9
√	—	√	58.4
√	√	—	62.7
√	√	√	64.4

4.4.5 上下文提示集中模板的选取

本文在CA-CL中对[CTX-Set]的上下文提示模板的选择进行系统分析,研究不同数量上下文提示模板

对模型性能的影响. 首先,本文收集了5个常用的上下文提示模板,包括"*a photo of {}.*""*a clean origami {}.*""*a picture of {}.*""*an image of {}.*"和"*a snapshot of {}.*",基于这些模板设计不同数量的组合,即针对每种数量设置,列举所有的模板组合,按照默认配置进行对比实验,并选取其中性能最佳的结果进行分析,以探讨不同模板的选取对模型性能的影响,实验结果如表7所示.

表7　[CTX-Set]中不同数量上下文组合生成CAM的最佳性能与训练时间

模板数量	1	2	3	4	5
最佳mIoU/%	62.8	63.1	63.7	64.4	62.4
训练时间/h	1.7	2.0	2.4	2.7	3.1

由表7可知,在PASCAL VOC 2012训练集上,随着模板数量的增加,模型性能逐步提升,这说明多样化的上下文信息能够为模型提供更全面的语义特征,激活更加完整的目标区域. 当m=4时,模型性能最佳,此时[CTX-Set]中上下文提示模板分别为"*a photo of {}.*""*an image of {}.*""*a picture of {}.*"和"*a snapshot of {}.*". 然而,当模板数量增加到5个时,模型性能出现下降,这可能是由于模板过多引入了冗余信息,从而影响了性能. 同时,增加模板数量延长了模型的训练时间,并对计算成本需求更大,对于较大的数据集,这种影响会更加显著. 此外,实验中还发现,模型的整体性能与[CTX-Set]中所采用的不同模板组合有关,与各个模板的排列顺序无关,也并不依赖于单个模板的独立性能.

综合考虑整体性能与计算成本,本文最终确定使用4个上下文提示模板进行组合(m=4),具体为"*a photo of {}.*""*an image of {}.*""*a picture of {}.*"和"*a snapshot of {}.*".

4.4.6　背景集中各类别背景词数量配置

本文对背景集[BGD-Set]中各类别的背景词数量n进行了消融实验,以评估其对CA-CL的影响. 具体而言,首先设定n=3,4,5,6,7,使用ChatGPT获取背景词,然后进行对比实验,实验结果如表8所示,当n取值为3或4时,模型的性能表现不佳,这可能是由于背景词过少,导致背景信息不充分,从而限制模型对背景区域的抑制能力;当n=5时,模型的性能最佳,同时计算成本和训练时间相对适中;当n>5时,模型性能出现了略微下降,这可能是由于引入了冗余的背景信息,产生干扰,导致背景过度抑制. 此外,背景词数量的增加提高了计算成本和训练时间,且在类别数量较多的数据集中,这种影响尤为明显. 最终,本文选择n=5作为[BGD-Set]的背景词数量.

表8　[BGD-Set]中不同背景词数量生成CAM的性能与训练时间

背景词数量	3	4	5	6	7
mIoU/%	59.7	61.5	64.4	63.6	63.7
训练时间/h	2.5	2.6	2.7	2.9	3.1

5　结论

本文提出了一种CA-CL的弱监督语义分割框架,利用文本提示引导模型生成高质量的CAM. 该框架引入DCPS、ITCL和CSBS,有效提升了模型定位目标物体的能力,并减少背景区域的误激活,能够生成较为完整的类激活图,提升了弱监督语义分割的性能. 在PASCAL VOC 2012和MS COCO 2014数据集上进行的大量实验验证了本文方法的有效性. 然而,尽管CA-CL在生成完整的CAM和抑制背景误激活方面取得了良好效果,但是也存在对文本提示和背景集的依赖. 未来的研究将聚焦于生成高质量的文本提示和优化背景抑制策略,以进一步提升模型的分割性能.

参考文献

[1] KOLESNIKOV A, LAMPERT C H. Seed, Expand and Constrain: Three Principles for Weakly-Supervised Image Segmentation[M]//Computer Vision-ECCV 2016. Cham: Springer International Publishing, 2016: 695-711.

[2] AHN J, KWAK S. Learning pixel-level semantic affinity with image-level supervision for weakly supervised semantic segmentation[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2018: 4981-4990.

[3] DAI J F, HE K M, SUN J. BoxSup: Exploiting bounding boxes to supervise convolutional networks for semantic segmentation[C]//2015 IEEE International Conference on Computer Vision (ICCV). Piscataway: IEEE, 2015: 1635-1643.

[4] PAPANDREOU G, CHEN L C, MURPHY K P, et al. Weakly-and semi-supervised learning of a deep convolutional network for semantic image segmentation[C]//2015 IEEE International Conference on Computer Vision (ICCV). Piscataway: IEEE, 2015: 1742-1750.

[5] LIN D, DAI J F, JIA J Y, et al. ScribbleSup: Scribble-supervised convolutional networks for semantic segmentation[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2016: 3159-3167.

[6] VERNAZA P, CHANDRAKER M. Learning random-walk label propagation for weakly-supervised semantic segmentation[C]//2017 IEEE Conference on Computer Vi-

sion and Pattern Recognition (CVPR). Piscataway: IEEE, 2017: 2953-2961.

[7] BEARMAN A, RUSSAKOVSKY O, FERRARI V, et al. What's the Point: Semantic Segmentation with Point Supervision[M]//Computer Vision-ECCV 2016. Cham: Springer International Publishing, 2016: 549-565.

[8] ZHOU B, KHOSLA A, LAPEDRIZA A, et al. Learning deep features for discriminative localization[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2016: 2921-2929.

[9] KRÄHENBÜHL P, KOLTUN V. Efficient inference in fully connected CRFs with Gaussian edge potentials[EB/OL]. (2012-10-20)[2025-02-01]. https://arxiv.org/abs/1210.5644v1.

[10] AHN J, CHO S, KWAK S. Weakly supervised learning of instance segmentation with inter-pixel relations[C]// 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2019: 2209-2218.

[11] CHEN L C, PAPANDREOU G, KOKKINOS I, et al.Semantic image segmentation with deep convolutional nets and fully connected CRFs[J]. Computer Science, 2014(4): 357-361.

[12] 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.

[13] HUANG Z L, WANG X G, WANG J S, et al. Weakly-supervised semantic segmentation network with deep seeded region growing[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2018: 7014-7023.

[14] WEI Y C, FENG J S, LIANG X D, et al. Object region mining with adversarial erasing: A simple classification to semantic segmentation approach[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2017: 6488-6496.

[15] ZHANG D, ZHANG H W, TANG J H, et al. Causal intervention for weakly-supervised semantic segmentation[C]// NIPS'20: Proceedings of the 34th International Conference on Neural Information Processing Systems. New York: ACM, 2020: 655-666.

[16] WANG Y D, ZHANG J, KAN M N, et al. Self-supervised equivariant attention mechanism for weakly supervised semantic segmentation[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2020: 12275-12284.

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

[18] XIE J H, HOU X X, YE K, et al. CLIMS: Cross language image matching for weakly supervised semantic segmentation[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2022: 4473-4482.

[19] MURUGESAN B, HUSSAIN R, BHATTACHARYA R, et al. Prompting classes: Exploring the power of prompt class learning in weakly supervised semantic segmentation[C]//2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). Piscataway: IEEE, 2024: 290-301.

[20] LIN Y Q, CHEN M H, WANG W X, et al. CLIP is also an efficient segmenter: A text-driven approach for weakly supervised semantic segmentation[C]//2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2023: 15305-15314.

[21] DENG S H, ZHUO W, XIE J H, et al. QA-CLIMS: Question-answer cross language image matching for weakly supervised semantic segmentation[C]//Proceedings of the 31st ACM International Conference on Multimedia. New York: ACM, 2023: 5572-5583.

[22] JANG S, YUN J, KWON J, et al. DIAL: Dense Image-text Alignment for Weakly Supervised Semantic Segmentation[M]//Computer Vision-ECCV 2024. Cham: Springer Nature Switzerland, 2024: 248-266.

[23] EVERINGHAM M, VAN GOOL L, WILLIAMS C K I, et al. The pascal visual object classes (VOC) challenge[J]. International Journal of Computer Vision, 2010, 88(2): 303-338.

[24] LIN T Y, MAIRE M, BELONGIE S, et al. Microsoft COCO: Common objects in context[M]//Computer Vision-ECCV 2014. Cham: Springer International Publishing, 2014: 740-755.

[25] ZHANG F, GU C C, ZHANG C Y, et al. Complementary patch for weakly supervised semantic segmentation[C]// 2021 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway: IEEE, 2021: 7222-7231.

[26] LEE M, KIM D, SHIM H. Threshold matters in WSSS: Manipulating the activation for the robust and accurate segmentation model against thresholds[C]//2022 IEEE/

CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2022: 4320-4329.

[27] CHEN L Y, LEI C Y, LI R H, et al. FPR: False positive rectification for weakly supervised semantic segmentation[C]//2023 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway: IEEE, 2023: 1108-1118.

[28] JIANG P T, YANG Y Q, HOU Q B, et al. L2G: A simple local-to-global knowledge transfer framework for weakly supervised semantic segmentation[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2022: 16865-16875.

[29] ZHOU T F, ZHANG M J, ZHAO F, et al. Regional semantic contrast and aggregation for weakly supervised semantic segmentation[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2022: 4289-4299.

[30] SIMONYAN K, VEDALDI A, ZISSERMAN A. Deep inside convolutional networks: Visualising image classification models and saliency maps[EB/OL]. (2014-08-19)[2025-02-01]. https://arxiv.org/abs/1312.6034.

[31] RU L X, ZHAN Y B, YU B S, et al. Learning affinity from attention: End-to-end weakly-supervised semantic segmentation with transformers[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2022: 16825-16834.

[32] RU L X, ZHENG H L, ZHAN Y B, et al. Token contrast for weakly-supervised semantic segmentation[C]//2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2023: 3093-3102.

[33] WU W Y, DAI T H, CHEN Z, et al. APC: Adaptive patch contrast for weakly supervised semantic segmentation[EB/OL]. (2024-07-15)[2025-02-01]. https://arxiv.org/pdf/2407.10649.pdf.

[34] LI J, LI D, XIONG C, et al. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation[C]//Proceedings of the International Conference on Machine Learning. Piscataway: IEEE, 2022: 12888-12900.

[35] CHEN T, KORNBLITH S, NOROUZI M, et al. A simple framework for contrastive learning of visual representations[C]//Proceedings of the International Conference on Machine Learning. Piscataway: IEEE, 2020: 1597-1607.

[36] HE K M, FAN H Q, WU Y X, et al. Momentum contrast for unsupervised visual representation learning[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2020: 9729-9738.

[37] KWON J, LEE E, CHO Y, et al. Learning to detour: Shortcut mitigating augmentation for weakly supervised semantic segmentation[C]//2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). Piscataway: IEEE, 2024: 808-817.

[38] XIE J H, XIANG J F, CHEN J L, et al. C2 AM: Contrastive learning of class-agnostic activation map for weakly supervised object localization and semantic segmentation[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2022: 979-988.

[39] ZHOU T F, ZHANG M J, ZHAO F, et al. Regional semantic contrast and aggregation for weakly supervised semantic segmentation[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2022: 4289-4299.

[40] YUAN K H, SCHAEFER G, LAI Y K, et al. A multi-strategy contrastive learning framework for weakly supervised semantic segmentation[J]. Pattern Recognition, 2023, 137: 109298.

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

[42] ROUMELIOTIS K I, TSELIKAS N D. ChatGPT and open-AI models: A preliminary review[J]. Future Internet, 2023, 15(6): 192.

[43] DENG J, DONG W, SOCHER R, et al. ImageNet: A large-scale hierarchical image database[C]//2009 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2009: 248-255.

[44] XU L, OUYANG W L, BENNAMOUN M, et al. Multi-class token transformer for weakly supervised semantic segmentation[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2022: 4300-4309.

[45] LEE J, OH S J, YUN S, et al. Weakly supervised semantic segmentation using out-of-distribution data[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2022: 16876-16885.

[46] WANG C W, XU R T, XU S B, et al. Treating pseudo-labels generation as image matting for weakly supervised

semantic segmentation[C]//2023 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway: IEEE, 2023: 755-765.

[47] WU W Y, DAI T H, HUANG X W, et al. Top-K pooling with patch contrastive learning for weakly-supervised semantic segmentation[C]//2024 IEEE International Conference on Systems, Man, and Cybernetics (SMC). Piscataway: IEEE, 2024: 5270-5275.

[48] RU L X, DU B, ZHAN Y B, et al. Weakly-supervised semantic segmentation with visual words learning and hybrid pooling[J]. International Journal of Computer Vision, 2022, 130(4): 1127-1144.

[49] CHEN Q, YANG L X, LAI J H, et al. Self-supervised image-specific prototype exploration for weakly supervised semantic segmentation[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2022: 4278-4288.

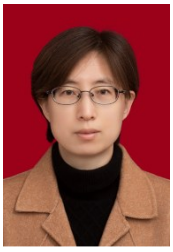
[50] CHEN Z Z, SUN Q R. Extracting class activation maps from non-discriminative features as well[C]//2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2023: 3135-3144.

[51] LAI Q, VONG C M. Weakly-supervised semantic segmentation via dual-stream contrastive learning of cross-image contextual information[EB/OL]. (2024-05-08)[2025-02-01]. https://arxiv.org/pdf/2405.04913.pdf.

[52] WU F W, HE J X, YIN Y F, et al. Masked collaborative contrast for weakly supervised semantic segmentation[C]//2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). Piscataway: IEEE, 2024: 851-860.

[53] ZHAO X Q, TANG F L, WANG X Y, et al. SFC: Shared feature calibration in weakly supervised semantic segmentation[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2024, 38(7): 7525-7533.

作者简介



白雪飞 女,1980年出生于山西省吕梁市. 现为山西大学计算机与信息技术学院副教授、硕士生导师. 主要研究方向为图像处理、机器学习.
E-mail: baixuefei@sxu.edu.cn

王渊辉 男,2000年出生于山西省长治市. 现为山西大学计算机与信息技术学院硕士研究生. 主要研究方向为图像处理、机器学习.
E-mail: 202322407043@email.sxu.edu.cn

许文杰 男,2001年出生于山西省晋中市. 现为山西大学计算机与信息技术学院硕士研究生. 主要研究方向为图像处理、机器学习.
E-mail: 202322409023@email.sxu.edu.cn

王文剑 女,1968年出生于山西省太原市. 现为计算智能与中文信息处理教育部重点实验室副主任、教授、博士生导师. 主要研究方向为机器学习、计算智能、图像处理.
E-mail: wjwang@sxu.edu.cn