基于伪标签去噪和SAM优化的大规模无监督语义分割

杨维静,徐 瑞,顾浩文,陈 涛,舒祥波,姚亚洲*
(南京理工大学计算机与工程学院,江苏南京 210094)

摘 要: 语义分割技术能够对复杂、多元的场景实现细粒度理解,是促进无人系统高效、智能工作的关键技术之一.大规模无监督语义分割旨在从大规模未标记图像中学习语义分割能力.然而,现有方法由于自学习伪标签存在类别混淆和形状表示欠佳的问题,导致最终分割精度较低.为此,本文提出一种伪标签去噪和SAM优化(Pseudo-label Denoising and SAM Optimization,PDSO)方法以解决大规模无监督语义分割问题.本文设计了一种基于去噪的特征微调模块,在基于小损失准则从大规模数据集中筛选出具有干净图像级伪标签的潜在样本后,利用这些干净样本对预训练的主干网络进行微调,使网络获得更稳健的类别表示.为了进一步减少伪标签中的类别噪声,设计了一种基于聚类的样本去噪模块,根据类别占比和样本与聚类中心之间的距离来去除干扰聚类任务的噪声样本,从而提升聚类性能.本文还设计了一种SAM提示优化模块,根据聚类距离识别出图像中的活跃类别,以过滤噪声目标,并将点和框作为SAM的目标提示信息,生成预期的目标掩膜以细化伪标签中目标的边缘.实验结果表明,在大规模语义分割数据集ImageNet-S_{50}、ImageNet-S_{300}和ImageNet-S_{919}的测试集上,本文方法在平均交并比指标上分别达到了45.0%、26.6%和14.5%,显著提高了分割目标的类别准确率和边缘精度.

关键词: 大规模无监督语义分割;图像级去噪;分割一切模型;伪标签;聚类
基金项目: 国家自然科学基金(No.62302217);装备发展部信息系统共用技术预研项目(No.31511030202)
中图分类号: TP751 **文献标识码:** A **文章编号:** 0372-2112(2025)03-0716-12
电子学报URL:http://www.ejournal.org.cn **DOI:**10.12263/DZXB.20240357

Pseudo-label Denoising and SAM Optimization for Large-scale Unsupervised Semantic Segmentation

YANG Wei-jing, XU Rui, GU Hao-wen, CHEN Tao, SHU Xiang-bo, YAO Ya-zhou*
(*School of Computer Science and Engineering*, *Nanjing University of Science and Technology*, *Nanjing*, *Jiangsu* 210094, *China*)

Abstract: Semantic segmentation technology enables fine-grained understanding of complex and diverse scenes and is one of the key technologies to promote efficient and intelligent work of unmanned systems. Large-scale unsupervised semantic segmentation aims to learn semantic segmentation capabilities from a large number of unlabeled images. However, the existing approaches suffer heavily from their noisy self-learned pseudo-labels with poor category and shape representations, leading to low final segmentation accuracy. In this paper, we propose a Pseudo-label Denoising and SAM Optimization (PDSO) approach for large-scale unsupervised semantic segmentation to alleviate the problem mentioned above. Specifically, we first propose a denoising-based feature fine-tuning module, which fine-tunes the pre-trained backbone network with clean image-level pseudo-label samples selected from a large dataset based on a small loss criterion, enabling the network to obtain more robust category representations. To further reduce category noise in pseudo-labels, we propose a clustering-based sample denoising module to discard noisy samples that interfere with clustering based on the category proportion and the distances between samples and cluster centers, thereby enhancing clustering performance. Moreover, we propose a SAM prompt optimization module, which identifies active categories in the image based on clustering distance to filter out noisy targets and uses points and boxes as SAM's target prompt information to generate expected target masks and refine the edges of targets in pseudo-labels. Our proposed PDSO reaches the mIoU of 45.0%, 26.6%, and 14.5% on the test

收稿日期:2024-04-22;修回日期:2024-09-21;责任编辑:朱梅玉
*通讯作者:姚亚洲

set of ImageNet-S_{50}, ImageNet-S_{300}, and ImageNet-S_{919} datasets, respectively, which significantly improves the category accuracy and edge accuracy of the segmented targets.

Key words: large-scale unsupervised semantic segmentation; image-level denoising; segment anything model; pseudo-label; clustering

Foundation Item(s): National Natural Science Foundation of China (No.62302217); Information Systems Common Technology Pre-research Project of Equipment Development Department (No.31511030202)

1 引言

随着大数据和人工智能等前沿技术的飞速发展,从陆地到天空,从物理系统到信息系统,各种类型的无人系统[1~6]大量涌现,在军事、民用等领域得到了广泛应用,对于国家安全、经济发展和民生改善等方面具有重要意义. 例如,自动驾驶[7]系统通过对车辆进行实时、连续控制,有效减少了驾驶过程中的人为失误,提升了行驶效率和安全性. 智能水肥系统根据水土环境变化,实施精准灌溉与施肥,提高了农作物的品质,推动了精准农业的发展,实现农业现代化. 无人系统在现实世界中工作时,首先对环境进行精确感知与分析,如自动驾驶系统需要准确判定周围车辆和行人的位置以及行进轨迹,才能做出恰当的决策并行动. 在计算机视觉技术中,语义分割是无人系统获取环境信息的关键环节. 语义分割旨在将相应的语义类别准确分配给图像的每个像素[8],从而实现对场景的细粒度理解. 传统的语义分割方法依赖大量标注图像指导训练,然而现实场景中的图像往往缺乏标签,且收集和标注现实场景图像需要耗费昂贵的时间和人力成本. 鉴于此,通过自主学习解决分割问题的无监督学习成为了研究的热点[9].

先前一些无监督学习方法[10,11]使用聚类算法为未标记数据生成伪标签,再利用伪标签去训练分割模型. 例如,DeepCluster[11]方法先使用k-means[12]算法对特征进行迭代分组,随后利用生成的伪标签有监督地更新网络权重. 然而聚类伪标签中固有的噪声显著阻碍了这些无监督学习方法的性能. 为了解决这一问题,研究者们进行了许多探索,例如MMT[13]方法基于同步平均教学框架,利用辅助网络生成的更为鲁棒的软标签对目标网络生成的伪标签进行在线优化. 但此类方法中,辅助网络的训练会大幅增加计算成本. DenseCL[14]采用对比学习方法,通过最大化图像与其增强对之间的特征相似性,同时最小化负对之间的相似性来学习表示. 尽管此类自监督学习方法在无监督语义分割问题上取得了不错的进展,但这些方法在训练时不能同时兼顾分割目标的类别和形状表示,并且侧重于小型数据集和简单类别的分割,例如,PASCAL VOC[15]数据集仅涵盖20种类别,包含约2 000张图像,导致这些无监督语义分割方法无法高效地推广到复杂多样的大规模真实世界场景中.

最近,Gao等人[16]提出大规模无监督语义分割任务,致力于在具有大量类别数和大规模数据量的场景下实现自主学习分割能力,并提出PASS[16]方法,基于大规模未标记数据的自学习表示为像素分配类别标签. PASS方法将大规模无监督语义分割任务分解为3个主要步骤:(1)无监督的表示学习;(2)应用基于像素注意力的聚类方法生成伪标签;(3)分割微调. 相较于常规的无监督语义分割方法,PASS方法能够从复杂的大规模数据集中捕获像素级语义信息,在解决大规模无监督语义分割问题时展现出了极大的潜力,但仍存在2个不足:(1)在大规模数据集中,由于不可避免地存在大量易混淆的图像,自监督训练往往受到相似模式和复杂背景的干扰,导致模型难以准确定位并识别出目标,从而造成伪标签中的目标被赋予错误的类别标签,例如,图1第一行的狐狸被错误地识别为兔子;(2)自监督训练的主干网络会引起嘈杂的类别表示,使聚类中心发生偏移,进而阻碍像素级标签的准确分配,并导致较为粗糙的目标边缘分割,如图1第二行所示,帆船的分割伪标签包含大片的背景区域. 深度卷积神经网络具有很强的记忆能力,低质量的伪标签会严重影响分割模型的学习效果,限制分割网络性能的提升.

考虑到常规的无监督语义分割方法局限于解决小规模数据集和简单类别的分割问题[16],本文以PASS方法为基线,使得本文方法可以应用于复杂且多样化的真实场景分割. 针对PASS方法自学习得到的伪标签存在类别混淆和形状表示欠佳的问题,本文致力于利用去噪技术和SAM(Segment Anything Model)[17]模型提高伪标签中目标类别的准确率和边缘的分割精度. 为应对类别噪声问题,设计了一种基于去噪的特征微调(Denoising-based Feature Fine-tuning, DFF)模块,该模块利用由聚类生成的图像级伪标签训练分类网络,并基于小损失准则[18~20]筛选出具有干净伪标签的潜在样本,进而对预训练的主干网络进行微调,以增强其类别表示能力. 同时,鉴于易混淆的图像会引起聚类中心偏移,导致聚类性能下降,设计了一种基于聚类的样本去噪(Clustering-based Sample Denoising, CSD)模块,该模块依据类别占比和聚类距离剔除干扰聚类任务的样本,并对剩余的样本进行重新聚类,以获得更准确的聚

类中心,进一步提升聚类效果.

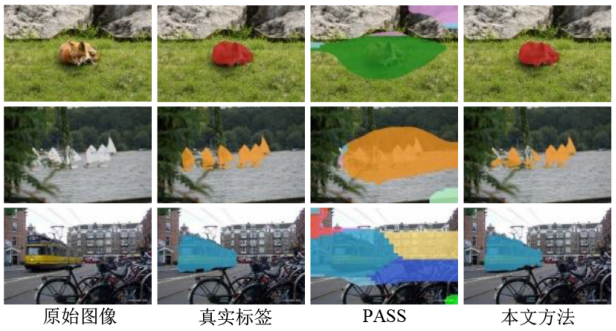


图1 传统方法与本文方法伪标签可视化对比

为了缓解伪标签中目标边缘粗糙的问题,本文设计了一种SAM提示优化(SAM Prompt Optimization,SPO)模块,该模块利用基于点和边界框提示生成的SAM掩膜,细化伪标签的目标边缘.聚类生成的伪标签可以为SAM提示优化模块提供目标的粗略定位框,但当伪标签中存在大量像素级噪声时,噪声目标可能会被错误地定位和分割.因此,SAM提示优化模块首先根据样本与聚类中心之间的距离识别出活跃类别,避免噪声目标被分割.此外,由于伪标签只能提供指示目标大致区域的边界框,SAM提示优化模块进一步利用具有最大像素注意力值的位置作为SAM生成预期目标掩膜的点提示,并为所得掩膜分配相应类别,从而形成最终用于分割网络训练的伪标签.在ImageNet-S_{50}、ImageNet-S_{300}和ImageNet-S_{919}[16]数据集上的大量实验表明,本文所提方法能够有效提升大规模无监督语义分割的效果.本文的贡献主要包含以下几个方面:

(1)提出了一种基于去噪的特征微调模块,利用小损失准则筛选出潜在的干净样本,进而对预训练的主干网络进行微调,有效提升了主干网络的类别表示能力.

(2)提出了一种基于聚类的样本去噪模块,从大规模数据集中去除干扰聚类任务的样本并重新进行聚类,达到修正聚类中心以生成高质量伪标签的目标.

(3)提出了一种SAM提示优化模块,根据聚类距离识别出图像中的活跃类别,有效抑制噪声对象的分割,并利用基于点和边界框提示获取的SAM掩膜,对活跃类别的目标边缘进行优化,有效缓解了伪标签中目标形状表示欠佳的问题.

2 相关基础

2.1 无监督语义分割

语义分割是最基本的计算机视觉任务之一,旨在为图像中的所有像素分配准确的语义标签.相较于分类和目标检测,语义分割能实现更加细粒度的视觉理解,在自动驾驶、环境监测、卫星成像等领域应用广泛.尽管基于深度学习的语义分割模型在经过大量标注数据的有监督训练后获得了卓越的分割效果,但数据的像素级注释所需的高昂时间和人力成本严重限制了其进一步发展.鉴于此,为了降低标注成本,许多研究尝试以无监督的方式解决语义分割问题.早期一些方法试图以语义一致性作为监督信号,在像素级别上学习语义对应关系.例如,IIC[21]通过最大化每对图像的类分配之间的互信息学习聚类.PiCIE[22]则将几何一致性作为一种归纳偏差,促进目标类别的学习.但这些方法过度依赖数据增强,难以在缺乏先验知识时学习语义一致性.近期,无监督语义分割的研究进展受益于利用自监督学习的先验作为监督信号,例如,InfoSeg[23]利用最大化局部像素特征和从自监督学习模型中获得的高层次类特征之间的互信息分割图像.由于从DINO[24]学习到的语义表示对目标外观变化更具鲁棒性,TransFGU[25]利用从DINO中发现的高级语义概念生成像素级的伪标签.STEGO[26]利用DINO提取图像特征,并通过知识蒸馏学习特征之间的对应关系.然而,这些方法在训练时不能同时兼顾分割目标的类别和形状,针对这一问题,本文方法旨在通过优化伪标签中目标的类别和形状来提升模型的分割性能.

2.2 大规模无监督语义分割

常规无监督语义分割方法的研究是在训练数据量和类别有限的情况下进行,因此,这些方法更适用于处理小型数据集和少量类别的分割任务,这种局限性使得它们难以应用于复杂且多样化的真实场景[16],故大规模无监督语义分割亟待研究.作为常规无监督语义分割任务的扩展,大规模无监督语义分割与常规无监督语义分割的不同在于其具有大规模的数据量和多样化的类别.大规模无监督语义分割任务的目标是在没有任何注释数据的情况下,自动将大规模数据分割成具有语义意义的区域.Gao等人[16]通过收集并注释ImageNet[27]数据集中的图像,构建了一个大规模的ImageNet-S数据集,并给出了一个具有明确目标和全面评估协议的大规模无监督语义分割基准,用于评估大规模无监督语义分割方法的性能.为了解决大规模无监督语义分割问题,他们提出PASS方法,结合增强的表示学习策略和像素注意力方案生成伪标签,以进一步训练分割网络.与常规无监督语义分割方法相比,PASS在大规模数据集的分割任务中表现出了显著的性能改进,但其自学习伪标签中包含大量的类别噪声以及目标对象轮廓不清晰,导致分割网络的分割性能降低.因此,本文致力于将去噪技术和SAM模型引入大规模无监督语义分割模型中,促进伪标签质量的提升.

2.3 伪标签去噪

大规模数据的自监督训练不可避免地会产生大量噪声伪标签,而深度神经网络很容易对噪声标签过拟合,因此,大规模数据的自监督伪标签会显著降低分割网络的性能. 为了减轻噪声标签的负面影响,Goldberger[28]和Patrini[29]等人提出了不同的噪声转移矩阵. 然而,在现实世界的数据集中,先验假设不再成立,这些转移矩阵方法也便失效[30]. 在缺乏噪声转移矩阵时,许多研究处理噪声标签的方向是设计更具抗噪性的损失函数. Xu等人[31]提出了基于行列式的互信息损失,可以应用于任何现有的分类神经网络,而不需要考虑噪声模式. Zhou等人[32]提出了一种限制模型输出的新策略,使任何损失对噪声标签都是鲁棒的. 然而,这种策略只有当类别数少且简单时才能很好地执行. 相比之下,基于小损失准则筛选干净样本进行噪声标签处理的样本选择策略在高噪声数据上取得了更好的性能. MentorNet[33]预先训练一个额外的教师网络,以监督学生网络的训练. 在训练期间,教师网络为学生网络提供可能被正确标记的干净样本. Co-teaching[18]也使用2个网络,但每个网络都选择一定数量的小损失样本,并将其反馈给并行模型进行进一步训练. 本文在采用小损失准则的同时,结合聚类距离从大规模数据筛选干净样本,从而提高样本选择的效率和准确性,促进伪标签去噪.

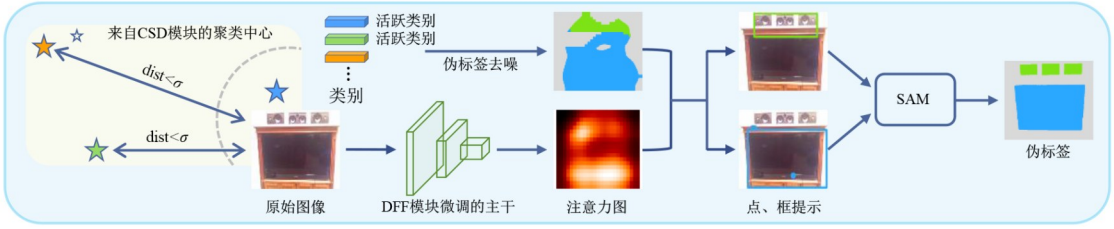
3 方法介绍

3.1 总体框架

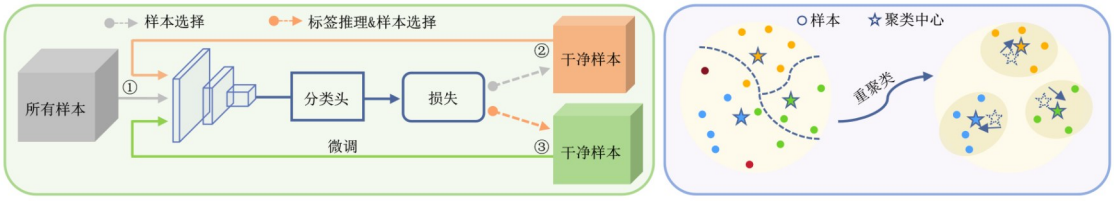
本文提出一种伪标签去噪和SAM优化方法来解决大规模无监督语义分割问题,总体框架如图2所示. 首先,大规模数据集中存在大量易混淆的图像,导致预训练主干学习到过多的错误信息. 本文提出一种基于去噪的特征微调模块,基于小损失准则挑选出潜在的干净样本对无监督的预训练主干进行微调,经微调后的主干网络具备更强的类别表示能力,进而有效降低伪标签中的类别噪声. 此外,为了获得更加准确的聚类中心,促进像素级标签的正确分配,提出了一种基于聚类的样本去噪模块,该模块根据类别占比和聚类距离去除大规模数据集中的噪声样本,并对剩余的样本重新聚类,有效缓解了噪声样本对聚类任务的负面影响,确保生成更高质量的聚类伪标签. 最后,针对分割目标边缘粗糙的问题,本文引入强大的SAM模型,提出一种SAM提示优化模块,该模块首先依据聚类距离识别出图像中的活跃类别,避免噪声对象被定位并分割,再利用点和边界框作为活跃类别的分割目标位置提示,由SAM生成相应类别目标的对象掩膜,实现细化伪标签中目标边缘的目的,从而提高模型对目标边缘的分割性能.

3.2 基于去噪的特征微调模块

大规模数据的自监督训练往往受到图像中相似模式和复杂背景的影响,导致网络不能准确定位并识别出目标对象. 针对方法PASS直接使用自监督学习的特征进行聚类,从而造成伪标签中包含大量的图像级和像素级噪声的问题,本文设计了一种基于去噪的特征微调模块,以减轻噪声信息对特征表示的干扰,提高聚类性能. 考虑到模型优先记忆具有干净标签的训练样本的信息,首先,用聚类生成的伪标签训练分类头,并基于小损失准则选取出干净样本. 然后,利用获取的相对干净的样本再次训练分类网络,并推理出整个大规模数据的图像级伪标签,相较于由方法PASS聚类生成的标签,由分类网络推理获得的图像级伪标签包含的噪声更少. 最后,利用从整个数据集中挑选出的充足数量的干净样本微调预训练的主干网络,使网络获得更加稳健的类别表示.



(a) SAM提示优化模块



(b) 基于去噪的特征微调模块　　(c) 基于聚类的样本去噪模块

图2　伪标签去噪和SAM优化方法结构示意图

给定训练数据集 $D=\left\{\left(\boldsymbol{x}_i, y_i^*\right)\middle|1\leqslant i\leqslant N\right\}$，其中 $\left(\boldsymbol{x}_i,y_i^*\right)$表示样本$\boldsymbol{x}_i$和该样本由聚类生成的图像级标签 y_i^*组成的数据对，N为图片数量．如图2(b)所示，该模块运用到3个分类网络，它们具有相同的主干和分类头并且都加载预训练的主干权重进行初始化．具体而言，该模块首先用数据集D训练分类网络f，在初始的训练周期，当f已经学习了充足的干净样本信息时，挑选出一个样本集 $S=\left\{\left(\boldsymbol{x}_i,\tilde{y}_i\right)\middle|1\leqslant i\leqslant\lceil\alpha N\rceil\right\}$，其中样本$\boldsymbol{x}_i$对应的标签$\tilde{y}_i$有极高可能性是正确标签，$\alpha$控制着$S$中的样本数量．按照小损失准则获得样本集$S$的过程可表示为

$$S=\operatorname{argmin}_{d':|d'|=\lceil\alpha N\rceil}\mathcal{L}\left(f,d'\right) \tag{1}$$

其中，$\mathcal{L}$是交叉熵损失函数．

此后，用数据集S训练分类网络g，训练后对大规模数据进行推理得到数据集 $D'=\left\{\left(\boldsymbol{x}_i,y_i'\right)\middle|\ y_i'=g\left(\boldsymbol{x}_i\right),1\leqslant i\leqslant N\right\}$，其中由分类网络$g$推理得到的标签$y_i'$相比于由聚类生成的图像级标签$y_i^*$更加可靠．同样的，再次基于分类损失从数据集$D'$中挑选出小损失样本，组成一个干净样本集 $S'=\left\{\left(\boldsymbol{x}_i,\tilde{y}_i'\right)\middle|1\leqslant i\leqslant\lceil\beta N\rceil\right\}$，过程如下：

$$S'=\operatorname{argmin}_{d':|d'|=\lceil\beta N\rceil}\mathcal{L}\left(g,d'\right) \tag{2}$$

其中，$\alpha<\beta<1$，S'中的样本数量由β控制．随后，用新的样本集S'训练分类网络h实现微调无监督预训练主干的目的．

3.3 基于聚类的样本去噪模块

尽管本文提出的基于去噪的特征微调模块可以提高主干网络的类别表示能力，但大规模数据集中通常包含大量不利于聚类任务的易混淆图像，导致图像级伪标签的类别分配不均衡[34]．由于深度神经网络很容易对噪声标签过拟合，聚类伪标签中的类别噪声会严重降低模型的分割性能，尤其是对于包含大量错误聚类标签的头部类别的分割．因此，本文设计了一种基于聚类的样本去噪模块进一步提高聚类性能．

该模块基于样本标签的类别占比和样本与聚类中心之间的距离，识别出可能导致聚类中心向错误方向偏移的噪声样本．$\boldsymbol{h}_c$表示类别c的聚类中心，n_c表示分配给类别c的样本数量，$1\leqslant c\leqslant C$，其中C为语义类别的数量．首先，将目标特征$\boldsymbol{h}_i$与$\boldsymbol{h}_c$之间的距离定义如下：

$$d_{ic}=\left\|\boldsymbol{h}_c-\boldsymbol{h}_i\right\|_2 \tag{3}$$

其中，$\|\cdot\|_2$表示一个向量的L_2范数．紧接着，按升序对$\left\{d_{ic},c\in 1,2,\cdots,C\right\}$进行排列，$c_i^*$和$c_i'$表示离目标特征$\boldsymbol{h}_i$最近和第二近的类别，对应的最短和第二短的距离分别为$d_{ic_i^*}$和$d_{ic_i'}$．如果分配给类别c_i^*的样本数量明显大于平均每个类别所占的样本数N/C，则定义类别c_i^*为头部类别，属于头部类别的目标样本被视为潜在的噪声样本．潜在噪声样本的判别过程如下：

$$\text{head}_i=\begin{cases}1, & n_{c_i^*}>\gamma\dfrac{N}{C}\\ 0, & \text{其他}\end{cases} \tag{4}$$

其中，$\gamma>1$，是一个预定义的比例系数．由于可靠的样本通常只靠近一个聚类中心而远离其他聚类中心，因此，将远离所有聚类中心或靠近多个聚类中心的样本视为不可靠的样本，如公式(5)所示：

$$\text{dist}_i=\begin{cases}1, & d_{ic_i^*}>\nabla d\ \text{或}\ d_{ic_i'}-d_{ic_i^*}<\nabla m\\ 0, & \text{其他}\end{cases} \tag{5}$$

最终按如下标准确定噪声样本：

$$\text{noise}_i=\begin{cases}1, & \text{head}_i=1\ \text{且}\ \text{dist}_i=1\\ 0, & \text{其他}\end{cases} \tag{6}$$

将噪声样本去除后，用去噪后的大规模数据集代替原始数据集重新进行聚类．同时，将从基于去噪的特征微调模块获得的微调权重加载到主干中，以获得更有利于聚类任务的类别表示．使用基于聚类的样本去噪模块校正聚类中心后，聚类性能显著提高．

3.4 SAM提示优化模块

尽管基于去噪的特征微调和基于聚类的样本去噪模块显著提高了模型的类别表示能力和聚类性能，但由于模型的形状表示能力较差，聚类生成的伪标签无法捕捉到目标边缘．因此，本文引入了一种强大且通用的视觉分割模型SAM，该模型能够根据用户提示生成指定目标的高精度分割掩膜，对此本文设计了一种SAM提示优化模块以获取目标对象的形状线索．由聚类生成的伪标签可以获得丰富的语义信息和目标对象的粗略定位，但像素级噪声可能导致噪声目标被定位并被分割出来，因此该模块首先根据伪标签中各类别聚类中心与样本的距离判定出图像中的活跃类别，如下所示：

$$a_{ic}=\begin{cases}1, & c=c_i^*\ \text{或}\ d_{ic}<\sigma\\ 0, & \text{其他}\end{cases} \tag{7}$$

其中，a_{ic}表示样本x_i中类别c的活跃状态．只有当类别c是主类别c_i^*或者d_{ic}小于预定义阈值σ时，才将其视为样本x_i的活跃类别．对于$a_{ic}=0$的噪声类别，从伪标签中剔除它们对应的像素，并将它们视为背景．对于$a_{ic}=1$的活跃类别，将伪标签中活跃类别的外接矩形框作为目标定位信息，输入SAM以生成目标掩膜．

获得准确的目标定位是高效利用SAM的关键，然而，伪标签只能提供较为粗糙的目标定位框．因此，该模块进一步利用具有最高像素注意力值的点提示为SAM提供额外的目标位置线索．$\boldsymbol{h}_i\in\mathbb{R}^{L\times H\times W}$代表样本

$\boldsymbol{x}_i$的像素注意力特征,其中H、W和L分别为特征的高度、宽度和通道维度. 首先利用沿着通道维度的平均池化操作来获得空间注意力$\boldsymbol{A}_i$,过程如下:

$$\boldsymbol{A}_i=\frac{1}{L}\sum_{l=1}^{L}\boldsymbol{h}_{i(l)} \tag{8}$$

然后,选择像素注意力值最大的位置P作为目标定位点:

$$P=\mathrm{argmax}_{(h,w)}\boldsymbol{A}_{i(h,w)} \tag{9}$$

如图2(*c*)所示,伪标签中活跃类别的所有外接矩形框和目标定位点都被用作输入SAM的目标位置提示,以生成期望的目标掩膜. 随后将对应的语义类别分配给各目标掩膜,形成用于分割网络训练的最终伪标签.

4　实验结果与分析

4.1　实验设置

4.1.1　数据集和评价指标

遵循基线方法PASS[16],本文使用数据集ImageNet-S_{50}、ImageNet-S_{300}和ImageNet-S_{919}对所提出的方法进行评估. 数据集ImageNet-S_{50}的目标分割类别为50类,其中训练集包含64 431张图像,验证集包含752张图像,测试集包含1 682张图像. 数据集ImageNet-S_{300}的目标分割类别为300类,其中用于训练、验证和测试的图片数量分别为384 862、4 097和9 088. 数据集ImageNet-S_{919}有919个目标分割类别,用于训练、验证和测试的图片数量分别为1 183 322、12 419和27 423. 本文采用平均交并比(mean Intersection over Union, mIoU)[35,36]、图像级准确率(Image-level Accuracy, Img-Acc)和F度量值(F-measure, F_β)作为大规模无监督语义分割任务的性能评价指标.

4.1.2　实验细则

本实验采用方法PASS作为基线,并遵循其设置,即基于网络ResNet-18训练ImageNet-S_{50}数据集,基于网络ResNet-50训练ImageNet-S_{300}和ImageNet-S_{919}数据集. 本文将基线方法PASS在表示学习步骤中训练的迭代次数为400轮的网络ResNet-18和ResNet-50作为预训练的主干. 对于数据集ImageNet-S_{50}、ImageNet-S_{300}和ImageNet-S_{919},在基于去噪的特征微调模块中,分类网络f训练的迭代次数为15轮,批处理大小设置为64和128,分类网络g和h训练的迭代次数为50轮,批处理大小设置为32和128. 样本选择时,比例因子(α,β)设置为(3/5, 2/3). 样本去噪时,比例因子γ设置为1.4,阈值$(\nabla m,\nabla d)$设置为(0.5, 0.2). 像素级去噪时,阈值σ设置为0.9. 下游分割网络的设置和训练也遵循基线方法PASS,以一个1×1卷积层作为分割头并使用交叉熵损失监督训练,训练的迭代次数为20轮,批处理大小设置为256,以便与基线方法进行公平的实验比较. 本文在实验中采用具有动量的随机梯度下降法作为优化器,其中动量大小设置为0.9,权重衰减设置为1×10^{-6},初始学习率为0.6,并随着余弦学习率策略逐渐衰减到6×10^{-6}.

4.2　与现有技术比较

表1、表2和表3分别展示了在ImageNet-S_{50}、ImageNet-S_{300}和ImageNet-S_{919}数据集上本文方法和其他方法的实验结果. 本文除了将所提出的方法与基线方法PASS进行重点对比外,还在表1中给出了典型的无监督语义分割方法(例如MDC[11,18]、PiCIE[18]和MaskCon[37])在ImageNet-S_{50}数据集上的性能,并与本文方法进行对比. 可以看出,本文方法在面对规模大且复杂的数据时能得到更好的分割结果. 具体来说,在ImageNet-S_{50}的验证集和测试集上,本文方法的mIoU分别达到了46.3%和45.0%. 相较于基线方法PASS,本文方法在mIoU、Img-Acc和F_β指标上分别提高13%、8%和18%,证明了本文方法的有效性. 从表1可以看出,即使基线方法PASS使用额外的显著图信息时,本文方法仍然能在验证集和测试集上高出3.0%和2.7%. 对于ImageNet-S_{300}数据集,本文方法可以在验证集上将基线方法的mIoU提高到27.0%,在测试集上提高到26.6%. 在ImageNet-S_{919}数据集上,本文方法的mIoU在验证集上达到15.6%,在测试集上达到了14.5%. 在ImageNet-S_{300}和ImageNet-S_{919}数据集上的实验结果进一步证明了本文方法对具有大规模数据量和类别的数据集能够实现较好的分割性能.

图3展示了本文方法和基线方法在ImageNet-S_{50}数据集上的一些分割结果示例. 可以直观地看出,相较于基线方法PASS,对于图像中的前景目标类别,本文方法实现了更加准确的类别预测,被预测错误的像素也更少,例如,第一列中方法PASS错误将花瓶识别为包,而本文方法能够纠正这一类别混淆错误. 另外,得益于SAM模型对伪标签中目标形状的优化,本文方法使得分割目标的边缘更加清晰,例如第三列中方法PASS只能识别出图像中存在飞机且粗略刻画出飞机所在位置,而本文方法使得模型对飞机轮廓的刻画更加精细.

4.3　消融实验

为验证本文方法中各个模块对于大规模无监督语义分割任务的有效性,本文在ImageNet-S_{50}数据集上设置了一系列的消融实验,定量结果如表4所示. 在基线方法上加入基于去噪的特征微调模块增强网络的类别表示能力后,验证集上mIoU从31.9%提升到33.2%. 再加入基于聚类的样本去噪模块,将干扰聚类的噪声样本从大规模数据集中移除,重新聚类后,又带来了1.7%的性能增益. 加入SAM提示优化模块对伪标签中目标

表1 不同方法在ImageNet-S_{50}数据集上的对比结果

方法	$ImageNet_{1k}$预训练	显著图	mIoU/%		Img-Acc/%		F_β/%	
			验证集	测试集	验证集	测试集	验证集	测试集
MDC[11,18]	√	—	14.6	14.3	44.8	40.8	33.2	32.6
PiCIE[18]	√	—	17.8	17.6	45.0	44.0	32.1	31.6
MaskCon[37]	—	√	24.6	24.2	47.9	47.6	65.7	66.2
$PASS_p$[16]+RC[38]	—	√	42.6	42.1	58.8	61.8	62.1	61.3
$PASS_p$+Sal	—	√	43.3	42.3	64.6	65.2	70.0	69.9
MDC[11,18]	—	—	4.0	3.6	14.9	13.4	31.6	31.3
PiCIE[18]	—	—	5.0	4.5	15.8	14.0	14.6	32.2
$PASS_s$[16]	—	—	29.2	29.3	66.2	65.5	49.0	49.0
$PASS_p$	—	—	32.4	32.0	62.9	64.1	48.7	47.9
本文方法	—	—	46.3	45.0	71.6	72.6	66.3	66.1
ΔGain	—	—	↑13.9	↑13.0	↑8.7	↑8.5	↑17.6	↑18.2

表2 不同方法在ImageNet-S_{300}数据集上的对比结果

方法	mIoU/%		Img-Acc/%		F_β/%	
	验证集	测试集	验证集	测试集	验证集	测试集
ImageNet-S_{300}						
$PASS_s$	16.6	16.0	34.7	32.8	34.4	34.3
$PASS_p$	18.0	18.1	43.9	42.6	47.6	47.5
本文方法	27.0	26.6	56.1	54.2	62.5	62.8
ΔGain	↑9.0	↑8.5	↑12.2	↑11.6	↑14.9	↑15.3

表3 不同方法在ImageNet-S_{919}数据集上的对比结果

方法	mIoU/%		Img-Acc/%		F_β/%	
	验证集	测试集	验证集	测试集	验证集	测试集
$PASS_s$	7.3	6.6	19.9	18.0	34.8	34.6
$PASS_p$	11.5	11.0	24.0	22.3	37.1	36.9
本文方法	15.6	14.5	28.5	27.7	43.7	43.8
ΔGain	↑4.1	↑3.5	↑4.5	↑5.4	↑6.6	↑6.9

的边缘进行细化后，本文方法最终分别在验证集和测试集上达到46.3%和45.0% mIoU.

为了直观地展示本文方法在聚类性能上的提升，本文在ImageNet-S_{50}训练集上任选了4个类别，并使用t-SNE[39]方法对基线方法得到的图像特征以及增加了基于去噪的特征微调和基于聚类的样本去噪模块的方法得到的图像特征进行了可视化，如图4所示，其中不同颜色代表不同类别的样本. 观察可见，相较于基线方法，模型增加了基于去噪的特征微调和基于聚类的样本去噪模块后，相同类别的样本之间更加紧凑且不同类别的簇群之间更加分散，分类错误的样本数量明显减少. 图5通过选择聚类准确率(Accuracy，ACC)和标准化互信息(Normalized Mutual Information，NMI)作为评价指标，定量展示了基于去噪的特征微调和基于聚类的样本去噪模块对于提高聚类性能的有效性. 可见，通过基于去噪的特征微调模块对无监督的预训练主干进行微调后，聚类准确率从68%提高到71%. 利用基于

表4 ImageNet-S_{50}数据集消融实验结果

基线方法	基于去噪的特征微调	基于聚类的样本去噪	SAM提示优化	mIoU/%	
				验证集	测试集
√	—	—	—	31.9	32.0
√	√	—	—	33.2	33.0
√	√	√	—	34.9	35.3
√	—	—	√	40.5	39.1
√	√	√	√	46.3	45.0

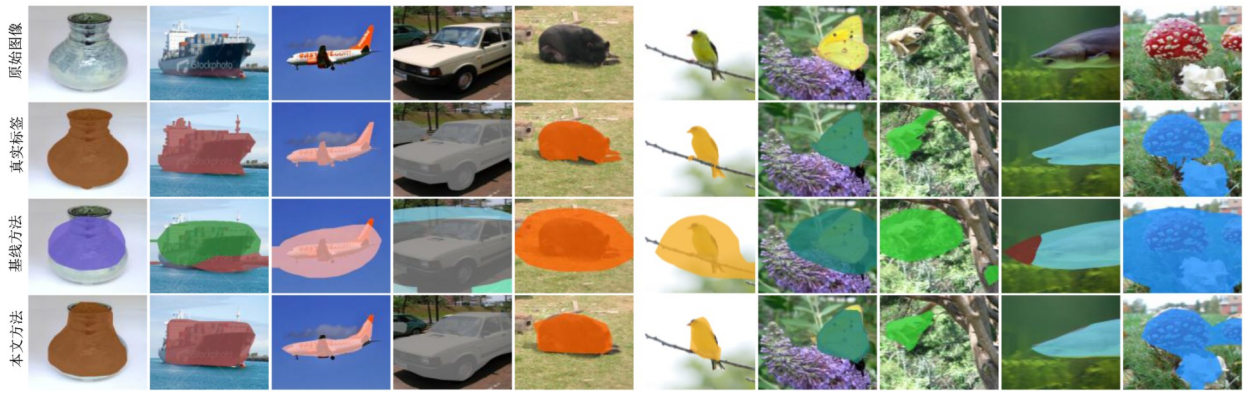


图3 本文方法与基线方法的分割结果可视化对比

聚类的样本去噪模块进一步提高聚类性能,聚类准确率又获得了3%的额外增益. 图6展示了本文所设计的每个模块对伪标签质量的提升,由于没有训练集的像素级真值标签,因此在验证集上报告结果. 可以看出,使用基于去噪的特征微调和基于聚类的样本去噪模块,可以将伪标签的分割性能从30.8% mIoU提高到34.2% mIoU. 通过引入SAM提示优化模块,可以进一步将分割性能提高到43.0%.

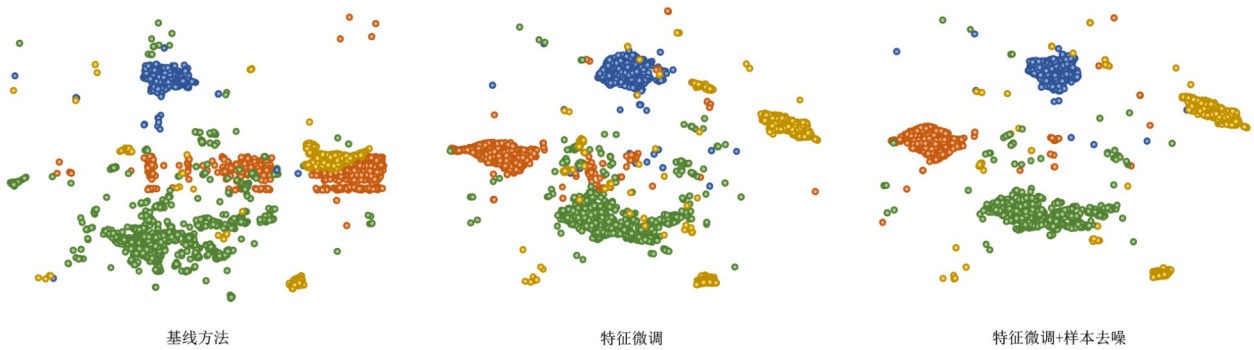


图4　与基线方法在ImageNet-S_{50}训练集上的聚类性能定性对比

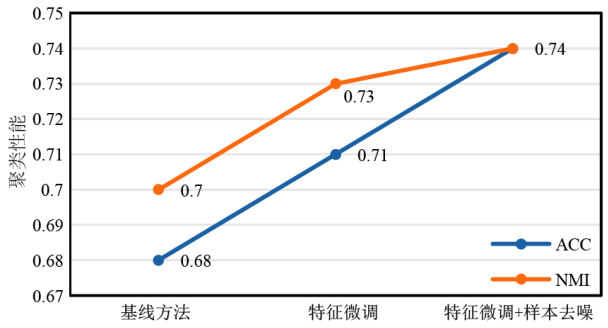


图5　与基线方法在ImageNet-S_{50}训练集上的聚类性能定量对比

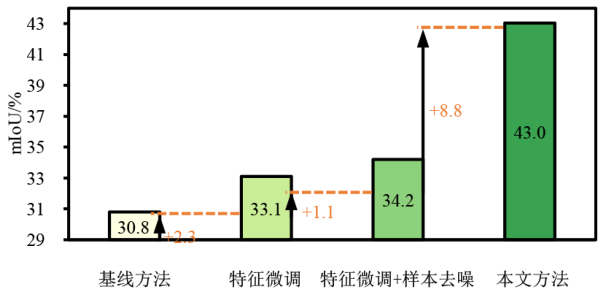


图6　与基线方法在ImageNet-S_{50}验证集上的伪标签质量对比

为验证本文提出的SAM提示优化模块对伪标签形状优化的优越性,表5展示了本文优化模块与其他先进的形状优化策略的性能对比,选择平均交并比和F度量值作为评价指标. 由于没有训练集的像素级真值标签,因此在验证集上报告结果,可以看出,相较于方法CRF[40]和SEPL[41],利用本文提出的SAM提示优化模块对伪标签进行优化能得到更好的分割结果. 图7为CRF、SEPL和SAM提示优化模块对伪标签进行优化的可视化结果,可以看出方法CRF仅能对目标掩膜边界处进行小范围的平滑处理. 方法SEPL对伪标签进行优化后,仍包含了部分背景噪声. 与方法CRF和SEPL相比,SAM提示优化模块对伪标签进行优化后,不仅可以很好地增强目标的边缘特征,还有效过滤了噪声目标.

表5　在ImageNet-S_{50}验证集上不同方法对伪标签形状优化的定量对比

方法	mIoU/%	F_β/%
基线方法	30.8	51.2
基线方法+CRF	31.6	52.1
基线方法+SEPL	34.7	55.4
基线方法+SAM提示优化	37.3	59.8

4.4　参数分析

γ是控制潜在噪声样本范围的参数,影响着基于聚类的样本去噪模块的性能. 在实验中,将γ分别设置为1.2、1.3、1.4、1.5和1.6,再对去除噪声样本后的数据集ImageNet-S_{50}进行聚类,得到的结果如图8所示. 可以看出,当比例因子γ设置为1.4时,评价指标ACC和NMI均达到了最高,聚类任务取得了最优结果. 阈值σ控制着图像中活跃类别的数量,图9展示了阈值σ在数据集ImageNet-S_{50}上对最终分割结果的影响. 可见,阈值σ过小,会使目标对象被过滤,而过大的阈值σ会使噪声目标被引入,导致最终分割精度较低. 阈值σ取值为0.8、0.9和1时,模型的分割精度较为稳定.在本文实验中,设置阈值σ为0.9,此时模型获得了最高的分割精度.

4.5　其他数据实验

为了检验本文方法的通用性,本文还在规模较小的数据集Pascal VOC 2012[15]上进行了实验验证. Pascal VOC 2012是常规无监督语义分割领域最常用的数据集之一,它包含21个目标类别,其中20个前景目标类别,1个背景类别. 经过SBD[42]数据集增强后,训练、验证和测试集分别包含10 582、1 449和1 456张图像.

实验中以ResNet-18作为主干网络，所有参数设置均与ImageNet-S_{50}数据集相同. 表6为本文方法和其他先进方法在数据集Pascal VOC 2012上分割结果的比较，本文方法实现了42.3%的分割精度，比使用DINO作为预训练主干的DINOSAUR[43]方法高出5.1%，进一步证明了本文方法的优越性.

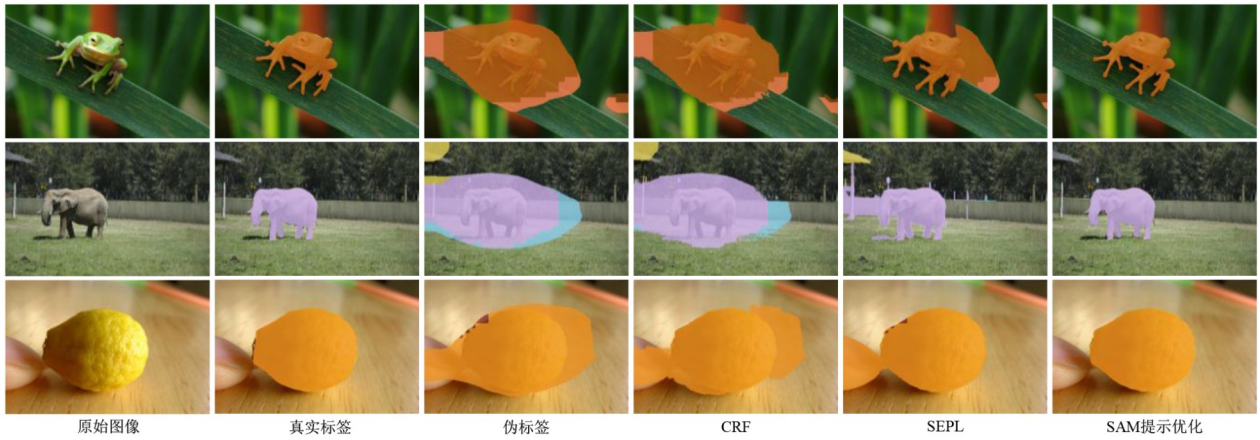


图7 不同方法对伪标签形状优化的可视化对比

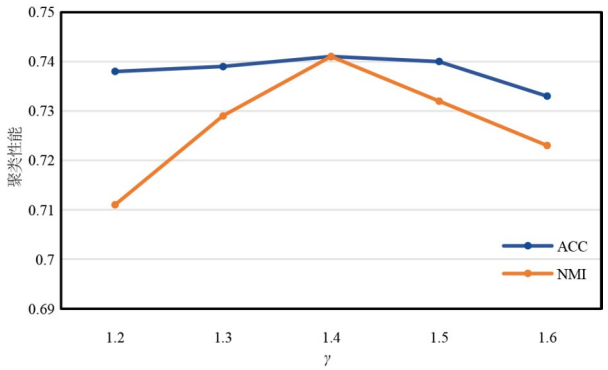


图8 在ImageNet-S_{50}数据集上γ不同取值对聚类性能的影响

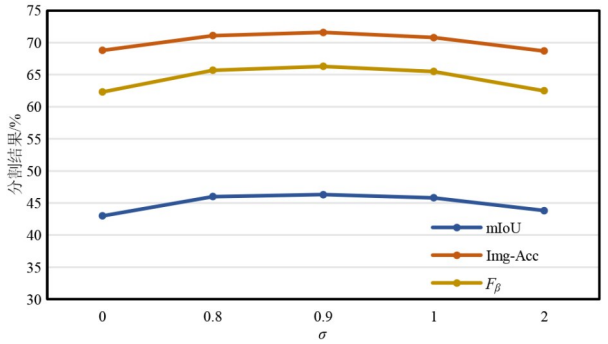


图9 在ImageNet-S_{50}数据集上σ不同取值对分割结果的影响

5 结论

针对大规模无监督语义分割任务中伪标签存在大量类别噪声和目标边缘不准确的问题，本文基于去噪技术和SAM模型对伪标签进行优化. 首先，本文提出了一种基于去噪的特征微调模块，该模块利用潜在的干净样本对无监督的预训练主干进行微调，使主干网络获得更稳健的类别表示. 此外，本文还引入了一种基于聚类的样本去噪模块，依据类别占比和聚类距离，从大规模数据中去除干扰聚类任务的样本并重新进行聚类任务. 最后，本文提出了一种SAM提示优化模块，用于进一步优化伪标签中目标的形状. 该模块识别活跃类别以防止定位到噪声目标，并且其将像素注意力值最大的位置点作为输入SAM的目标位置提示之一，有效弥补了边界框提示的局限性，使得定位信息更加充分. 实验结果表明，与之前先进的大规模无监督语义分割方法相比，本文提出的方法在大规模无监督语义分割任务中展现出了更出色的分割性能.

表6 不同方法在Pascal VOC 2012数据集上的对比结果

方法	mIoU/%
IIC[21]	9.8
SegSort[44]	11.7
MaskCon[37]	35.0
DenseCL[14]	35.1
TransFGU[25]	37.2
DINOSAUR[43]	37.2
DeepSpectral[45]	37.2
Leopart[46]	41.7
HSG[47]	41.9
本文方法	42.3

参考文献

[1] LI T J, LIU J, ZHANG W, et al. UAV-human: A large benchmark for human behavior understanding with unmanned aerial vehicles[C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Pis-

cataway: IEEE, 2021: 16261-16270.

[2] RODRÍGUEZ A C, D´ARONCO S, SCHINDLER K, et al. Mapping oil palm density at country scale: An active learning approach[J]. Remote Sensing of Environment, 2021, 261: 112479.

[3] KELLENBERGER B, MARCOS D, LOBRY S, et al. Half a percent of labels is enough: Efficient animal detection in UAV imagery using deep CNNs and active learning[J]. IEEE Transactions on Geoscience and Remote Sensing, 2019, 57(12): 9524-9533.

[4] LENCZNER G, CHAN-HON-TONG A, LE SAUX B, et al. DIAL: Deep interactive and active learning for semantic segmentation in remote sensing[J]. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2022, 15: 3376-3389.

[5] CHENG Y W, XU H, LIU Y M. Robust small object detection on the water surface through fusion of camera and millimeter wave radar[C]//2021 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway: IEEE, 2021: 15243-15252.

[6] LIN J Y, DIEKMANN P, FRAMING C E, et al. Maritime environment perception based on deep learning[J]. IEEE Transactions on Intelligent Transportation Systems, 2022, 23(9): 15487-15497.

[7] LI J L, DAI H, HAN H, et al. MSeg3D: Multi-modal 3D semantic segmentation for autonomous driving[C]//2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2023: 21694-21704.

[8] HE H Y, CAI J F, PAN Z Z, et al. Dynamic focus-aware positional queries for semantic segmentation[C]//2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2023: 11299-11308.

[9] SEONG H S, MOON W, LEE S, et al. Leveraging hidden positives for unsupervised semantic segmentation[C]//2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2023: 19540-19549.

[10] 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]//2019 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway: IEEE, 2019: 6111-6120.

[11] CARON M, BOJANOWSKI P, JOULIN A, et al. Deep clustering for unsupervised learning of visual features[M]// Lecture Notes in Computer Science. Cham: Springer International Publishing, 2018: 139-156.

[12] LLOYD S. Least squares quantization in PCM[J]. IEEE Transactions on Information Theory, 1982, 28(2): 129-137.

[13] GE Y, CHEN D, LI H. Mutual mean-teaching: Pseudo label refinery for unsupervised domain adaptation on person re-identification[C]//International Conference on Learning Representations. Piscataway: IEEE, 2020: 1-15.

[14] WANG X L, ZHANG R F, SHEN C H, et al. Dense contrastive learning for self-supervised visual pre-training[C]// 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2021: 3023-3032.

[15] EVERINGHAM M, ESLAMI S M ALI, VAN GOOL L, et al. The pascal visual object classes challenge: A retrospective[J]. International Journal of Computer Vision, 2015, 111(1): 98-136.

[16] GAO S H, LI Z Y, YANG M H, et al. Large-scale unsupervised semantic segmentation[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45(6): 7457-7476.

[17] KIRILLOV A, MINTUN E, RAVI N, et al. Segment anything[C]//2023 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway: IEEE, 2023: 3992-4003.

[18] HAN B, YAO Q M, YU X R, et al. Co-teaching: Robust training of deep neural networks with extremely noisy labels[EB/OL]. (2018-04-18) [2024-04-22]. https://arxiv.org/abs/1804.06872v3.

[19] KIM Y, KIM J M, AKATA Z, et al. Large loss matters in weakly supervised multi-label classification[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2022: 14156-14165.

[20] SONG H, KIM M, LEE J G. Selfie: Refurbishing unclean samples for robust deep learning[C]//International Conference on Machine Learning. Lille: PMLR, 2019: 5907-5915.

[21] JI X, VEDALDI A, HENRIQUES J. Invariant information clustering for unsupervised image classification and segmentation[C]//2019 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway: IEEE, 2019: 9865-9874.

[22] HYUN CHO J, MALL U, BALA K, et al. PiCIE: Unsupervised semantic segmentation using invariance and equivariance in clustering[C]//2021 IEEE/CVF Conference

on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2021: 16794-16804.

[23] HARB R, KNÖBELREITER P. InfoSeg: Unsupervised semantic image segmentation with mutual information maximization[M]//Lecture Notes in Computer Science. Cham: Springer International Publishing, 2021: 18-32.

[24] CARON M, TOUVRON H, MISRA I, et al. Emerging properties in self-supervised vision transformers[C]// 2021 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway: IEEE, 2021: 9650-9660.

[25] YIN Z Y, WANG P C, WANG F, et al. TransFGU: A top-down approach to fine-grained unsupervised semantic segmentation[M]//Lecture Notes in Computer Science. Cham: Springer Nature Switzerland, 2022: 73-89.

[26] HAMILTON M, ZHANG Z, HARIHARAN B, et al. Unsupervised semantic segmentation by distilling feature correspondences[C]//International Conference on Learning Representations. Piscataway: IEEE, 2022: 1-26.

[27] 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.

[28] GOLDBERGER J, BEN-REUVEN E. Training deep neural-networks using a noise adaptation layer[C]//International Conference on Learning Representations. Piscataway: IEEE, 2022: 1-9.

[29] PATRINI G, ROZZA A, MENON A K, et al. Making deep neural networks robust to label noise: A loss correction approach[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2017: 1944-1952.

[30] CHEN W K, ZHU C, LI M T. Sample prior guided robust model learning to suppress noisy labels[M]//Lecture Notes in Computer Science. Cham: Springer Nature Switzerland, 2023: 3-19.

[31] XU Y L, CAO P, KONG Y Q, et al. L_DMI: An information-theoretic noise-robust loss function[EB/OL]. (2019-09-08)[2024-04-22]. https://arxiv.org/abs/1909.03388v2.

[32] ZHOU X, LIU X M, WANG C Y, et al. Learning with noisy labels via sparse regularization[C]//2021 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway: IEEE, 2021: 72-81.

[33] JIANG L, ZHOU Z, LEUNG T, et al. Mentornet: Learning data-driven curriculum for very deep neural networks on corrupted labels[C]//International Conference on Machine Learning. New York: ACM, 2018: 2304-2313.

[34] YANG L R, MENG F M, LI H L, et al. Learning with noisy class labels for instance segmentation[M]//Lecture Notes in Computer Science. Cham: Springer International Publishing, 2020: 38-53.

[35] 梁新宇, 林洗坤, 权冀川, 等. 基于深度学习的图像实例分割技术研究进展[J]. 电子学报, 2020, 48(12): 2476-2486.
LIANG X Y, LIN X K, QUAN J C, et al. Research on the progress of image instance segmentation based on deep learning[J]. Acta Electronica Sinica, 2020, 48(12): 2476-2486. (in Chinese)

[36] 蔡超丽, 李纯纯, 黄琳, 等. ED-NAS: 基于神经网络架构搜索的陶瓷晶粒SEM图像分割方法[J]. 电子学报, 2022, 50(2): 461-469.
CAI C L, LI C C, HUANG L, et al. ED-NAS: Ceramic grain segmentation based on neural architecture search using SEM images[J]. Acta Electronica Sinica, 2022, 50(2): 461-469. (in Chinese)

[37] VAN GANSBEKE W, VANDENHENDE S, GEORGOULIS S, et al. Unsupervised semantic segmentation by contrasting object mask proposals[C]//2021 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway: IEEE, 2021: 10052-10062.

[38] CHENG M M, MITRA N J, HUANG X L, et al. Global contrast based salient region detection[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015, 37(3): 569-582.

[39] VAN DER MAATEN L, HINTON G. Visualizing data using t-SNE[J]. Journal of Machine Learning Research, 2008, 9(11): 2579-2605.

[40] KRÄHENBÜHL P, KOLTUN V. Efficient inference in fully connected crfs with Gaussian edge potentials[C]// NIPS'11: Proceedings of the 25th International Conference on Neural Information Processing Systems. New York: ACM, 2011: 109-117.

[41] CHEN T, MAI Z, LI R, et al. Segment anything model (SAM) enhanced pseudo labels for weakly supervised semantic segmentation[C]//Advances in Neural Information Processing Systems. New York: ACM, 2023: 1-14.

[42] HARIHARAN B, ARBELAEZ P, BOURDEV L, et al. Semantic contours from inverse detectors[C]//2011 International Conference on Computer Vision. Piscataway: IEEE, 2011: 991-998.

[43] SEITZER M, HORN M, ZADAIANCHUK A, et al. Bridging the gap to real-world object-centric learning[C]// International Conference on Learning Representations. Piscataway: IEEE, 2023: 1-43.

[44] HWANG J J, YU S, SHI J B, et al. SegSort: Segmentation by discriminative sorting of segments[C]//2019 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway: IEEE, 2019: 7334-7344.
[45] MELAS-KYRIAZI L, RUPPRECHT C, LAINA I, et al. Deep spectral methods: A surprisingly strong baseline for unsupervised semantic segmentation and localization[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2022: 8364-8375.
[46] ZIEGLER A, ASANO Y M. Self-supervised learning of object parts for semantic segmentation[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2022: 14502-14511.
[47] KE T W, HWANG J J, GUO Y H, et al. Unsupervised hierarchical semantic segmentation with multiview cosegmentation and clustering transformers[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2022: 2571-2581.

作者简介

杨维静　女,2000年出生,江苏泰州人. 南京理工大学计算机与工程学院硕士研究生. 主要研究方向为无监督语义分割.
E-mail: yangweijing@njust.edu.cn

陈　涛　男,1993年出生,江苏苏州人. 南京理工大学计算机与工程学院博士后. 主要研究方向为计算机视觉、语义分割、弱监督学习.
E-mail: taochen@njust.edu.cn

徐　瑞　男,1999年出生,江苏邳州人. 南京理工大学计算机与工程学院硕士研究生. 主要研究方向为BEV感知、遥感图像目标检测.
E-mail: 122106222759@njust.edu.cn

舒祥波　男,1986年出生,湖北孝感人. 南京理工大学计算机与工程学院博士生导师. 主要研究方向为计算机视觉、深度学习、模式识别、人工智能、机器学习、大数据.
E-mail: shuxb@njust.edu.cn

顾浩文　男,1998年出生,江苏南通人. 南京理工大学计算机与工程学院博士研究生. 主要研究方向为视频目标分割.
E-mail: guhaowen@njust.edu.cn

姚亚洲　男,1987年出生,江苏连云港人. 南京理工大学计算机与工程学院博士生导师. 主要研究方向为计算机视觉、多媒体技术、机器学习.
E-mail: yazhou.yao@njust.edu.cn