E-mail: jig@aircas.ac.cn
Website: www.cjig.cn
Tel: 010-58887035
中国图象图形学报
JOURNAL OF IMAGE AND GRAPHICS

中图法分类号:TP391　　**文献标识码**: A　　**文章编号**: 1006-8961(2024)05-0028-23

论文引用格式:Xiang W K, Zhou Q, Cui J C, Mo Z Y, Wu X F, Ou W H, Wang J D and Liu W Y. 2024. Weakly supervised semantic segmentation based on deep learning. Journal of Image and Graphics, 29(05):0028-0050(项伟康, 周全, 崔景程, 莫智懿, 吴晓富, 欧卫华, 王井东, 刘文予. 2024. 基于深度学习的弱监督语义分割方法综述. 中国图象图形学报, 29(05):0028-0050)[DOI:10.11834/jig.230628]

基于深度学习的弱监督语义分割方法综述

项伟康[1,2],周全[1,2*],崔景程[1],莫智懿[2],吴晓富[1],欧卫华[3],王井东[4],刘文予[5]

1. 南京邮电大学通信与信息工程学院,南京　210000; 2. 梧州学院广西高校智慧行业软件重点实验室,梧州　543002;
3. 贵州师范大学大数据与计算机科学学院,贵阳　550000; 4. 百度在线网络技术(北京)有限公司,北京　100000;
5. 华中科技大学电子信息与通信学院,武汉　430000

摘　要: 语义分割是计算机视觉领域中的基本任务,旨在为每个像素分配语义类别标签,实现对图像的像素级理解。近年来,得益于深度学习的发展,基于深度学习的全监督语义分割方法取得了巨大的进展。然而,这些方法往往需要大量带有像素级标注的训练数据,标注成本巨大,限制了其在诸如自动驾驶、医学图像分析以及工业控制等实际场景中的应用。为了降低数据的标注成本并进一步拓宽语义分割的应用场景,研究者们越来越关注基于深度学习的弱监督语义分割方法,希望通过诸如图像级标注、最小包围盒标注、线标注和点标注等弱标注信息实现图像的像素级分割预测。本文首先对语义分割任务进行了简要介绍,并分析了全监督语义分割所面临的困境,从而引出弱监督语义分割。然后,介绍了相关数据集和评估指标。接着,根据弱标注的类型和受关注程度,从图像级标注、其他弱标注以及大模型辅助这3个方面回顾和讨论了弱监督语义分割的研究进展。其中,第2类弱监督语义分割方法包括基于最小包围盒、线和点标注的弱监督语义分割。最后,本文分析了弱监督语义分割领域存在的问题与挑战,并就其未来可能的研究方向提出建议,旨在进一步推动弱监督语义分割领域研究的发展。

关键词: 语义分割;深度学习;弱监督语义分割(WSSS);图像级标注;最小包围盒标注;线标注;点标注;大模型

Weakly supervised semantic segmentation based on deep learning

Xiang Weikang[1,2], Zhou Quan[1,2*], Cui Jingcheng[1], Mo Zhiyi[2], Wu Xiaofu[1], Ou Weihua[3], Wang Jingdong[4], Liu Wenyu[5]

1. School of Communication and Information Engineering, Nanjing University of Posts and Telecommunications, Nanjing　210000, China;
2. Guangxi Colleges and Universities Key Laboratory of Intelligent Industry Software, Wuzhou University, Wuzhou　543002, China;
3. School of Big Data and Computer Science, Guizhou Normal University, Guiyang　550000, China;
4. Baidu Online Network Technology (Beijing) Co., Ltd, Beijing　100000, China;
5. School of Electronic Information and Communications, Huazhong University of Science and Technology, Wuhan　430000, China

Abstract: Semantic segmentation is an important and fundamental task in the field of computer vision. Its goal is to assign a semantic category label to each pixel in an image, achieving pixel-level understanding. It has wide applications in areas, such as autonomous driving, virtual reality, and medical image analysis. Given the development of deep learning in recent years, remarkable progress has been achieved in fully supervised semantic segmentation, which requires a large amount of

收稿日期:2023-09-12;**修回日期**:2023-12-05;**预印本日期**:2023-12-12

*通信作者:周全　quan.zhou@njupt.edu.cn

基金项目:国家自然科学基金项目(61876093,62262005);广西高校智慧行业软件重点实验室开放研究项目(2023B01)

Supported by: National Natural Science Foundation of China(61876093,62262005); Guangxi Colleges and Universities Key Laboratory of Intelligent Industry Software(2023B01)

training data with pixel-level annotations. However, accurate pixel-level annotations are difficult to provide because it sacrifices substantial time, money, and human-label resources, thus limiting their widespread application in reality. To reduce the cost of annotating data and further expand the application scenarios of semantic segmentation, researchers are paying increasing attention to weakly supervised semantic segmentation (WSSS) based on deep learning. The goal is to develop a semantic segmentation model that utilizes weak annotations information instead of dense pixel-level annotations to predict pixel-level segmentation accurately. Weak annotations mainly include image-level, bounding-box, scribble, and point annotations. The key problem in WSSS lies in how to find a way to utilize the limited annotation information, incorporate appropriate training strategies, and design powerful models to bridge the gap between weak supervision and pixel-level annotations. This study aims to classify and summarize WSSS methods based on deep learning, analyze the challenges and problems encountered by recent methods, and provide insights into future research directions. First, we introduce WSSS as a solution to the limitations of fully supervised semantic segmentation. Second, we introduce the related datasets and evaluation metrics. Third, we review and discuss the research progress of WSSS from three categories: image-level annotations, other weak annotations, and assistance from large-scale models, where the second category includes bounding-box, scribble, and point annotations. Specifically, image-level annotations only provide object categories information contained in the image, without specifying the positions of the target objects. Existing methods always follow a two-stage training process: producing a class activation map (CAM), also known as initial seed regions used to generate high-quality pixel-level pseudo labels; and training a fully supervised semantic segmentation model using the produced pixel-level pseudo labels. According to whether the pixel-level pseudo labels are updated or not during the training process in the second stage, WSSS based on image-level annotations can be further divided into offline and online approaches. For offline approaches, existing research treats two stages independently, where the initial seed regions are optimized to obtain more reliable pixel-level pseudo labels that remain unchanged throughout the second stage. They are often divided into six classes according to different optimization strategies, including the ensemble of CAM, image erasing, co-occurrence relationship decoupling, affinity propagation, additional supervised information, and self-supervised learning. For online approaches, the pixel-level pseudo labels keep updating during the entire training process in the second stage. The production of pixel-level pseudo labels and the semantic segmentation model are jointly optimized. The online counterparts can be trained end to end, making the training process more efficient. Compared with image-level annotations, other weak annotations, including bounding box, scribble, and point, are more powerful supervised signals. Among them, bounding-box annotations not only provide object category labels but also include information of object positions. The regions outside the bounding-box are always considered background, while box regions simultaneously contain foreground and background areas. Therefore, for bounding-box annotations, existing research mainly starts from accurately distinguishing foreground areas from background regions within the bounding-box, thereby producing more accurate pixel-level pseudo labels, used for training following semantic segmentation networks. Scribble and point annotations not only indicate the categories of objects contained in the image but also provide local positional information of the target objects. For scribble annotations, more complete pseudo labels can be produced to supervise semantic segmentation by inferring the category of unlabeled regions from the annotated scribble. For point annotations, the associated semantic information is expanded to the entire image through label propagation, distance metric learning, and loss function optimization. In addition, with the rapid development of large-scale models, this paper further discusses the recent research achievements in using large-scale models to assist WSSS tasks. Large-scale models can leverage their pretrained universal knowledge to understand images and generate accurate pixel-level pseudo labels, thus improving the final segmentation performance. This paper also reports the quantitative segmentation results on PASCAL VOC 2012 dataset to evaluate the performance of different WSSS methods. Finally, four challenges and potential future research directions are provided. First, a certain performance gap remains between weakly supervised and fully supervised methods. To bridge this gap, research should keep on improving the accuracy of pixel-level pseudo labels. Second, when WSSS models are applied to real-world scenarios, they may encounter object categories that have never appeared in the training data. This encounter requires the models to have a certain adaptability to identify and segment unknown objects. Third, existing research mainly focuses on improving the accuracy without considering the model size and inference speed of WSSS networks, posing a major challenge for the deployment of the model in real-world

applications that require real-time estimations and online decisions. Fourth, the scarcity of relevant datasets used to evaluate different WSSS models and algorithms is also a major obstacle, which leads to performance degradation and limits generalization capability. Therefore, large-scale WSSS datasets with high quality, great diversity, and wide variation of image types must be constructed.

Key words: semantic segmentation; deep learning; weakly supervised semantic segmentation (WSSS); image-level annotations; bounding-box annotations; scribble annotations; point annotations; large-scale models

0 引 言

语义分割是计算机视觉领域中极具挑战性的重要任务,其目的是为图像中的每个像素分配一个预定义的语义类别标签。图像语义分割广泛应用于自动驾驶、智能机器人、医学图像分析、遥感图像分析等诸多现实场景,具有广泛的应用前景和重大的现实意义。

传统的图像分割方法主要基于手工设计的特征和经典的机器学习算法。例如,先采用局部二值模式(local binary patterns, LBP)(Ojala等,1994)、方向梯度直方图(histogram of oriented gradients, HOG)(Dalal和Triggs, 2005)和尺度不变特征变换(scale-invariant feature transform, SIFT)(Lowe, 2004)等特征提取方法来提取图像的颜色、纹理、边缘和空间结构等信息;接着使用阈值分割、区域生长、边缘检测、图割算法等进行图像分割。以GrabCut(Rother等,2004)为例,它通过用户交互和迭代优化的方式,由用户提供的初始边界框实现对前景和背景的分割。传统的图像分割算法计算效率高、模型解释性强,但依赖于手工设计的特征和分割算法的选择,其表示能力和鲁棒性相对较差,无法处理复杂场景的图像分割(青晨 等,2020)。

自2015年,图像语义分割的里程碑模型全卷积神经网络(fully convolutional network, FCN)(Long等,2015)被提出后,深度学习(Hinton和Salakhutdinov, 2006)开始在图像分割领域得到广泛应用。FCN使用全卷积层代替传统的全连接层,允许接受任意尺寸的输入图像。它通过多层卷积和池化操作学习到图像的多尺度特征表示,并通过跳跃连接的机制融合多尺度特征,以提高分割结果的准确性。得益于深度学习的蓬勃发展,基于深度学习的语义分割方法大量涌现,在准确性、效率和应用领域的拓展等方面不断取得突破(Minaee等,2022)。

这些基于深度学习的语义分割方法大都采用全监督的训练范式,即通过大量精细的像素级标注图像来训练深度神经网络模型。然而,这种数据标注过程需要耗费大量的人力、时间和资金。而对于某些复杂的场景和某些特殊的专业领域(如医学图像分析、遥感图像理解等),人工标注的准确性也无法完全保证。毫无疑问,这些问题严重制约了图像语义分割的进一步发展。

为了缓解全监督语义分割方法对像素级标注的巨大需求,并进一步提升语义分割实际应用的可扩展性,研究者们开始关注弱监督语义分割(weakly supervised semantic segmentation, WSSS)方法。WSSS是指在训练过程中仅利用相对较少的标注信息,而不是精细的像素级标注信息来训练语义分割模型。Lin等人(2014)的研究表明,对一幅图像进行像素级标注的时间成本约是最小包围盒标注的15倍,图像级标注的60倍。因此,使用弱标注监督信息可以大大降低数据标注过程所需的人力、时间和资金成本。如图1所示,根据所提供监督信息的多寡,弱标注主要可分为:图像级标注(Chen和Sun, 2023; Kweon等, 2023)、最小包围盒标注(Dai等, 2015; Oh等, 2021)、线标注(Lin等, 2016; Xu等, 2021)以及点标注(Bearman等,2016)等。

WSSS的关键问题在于如何充分利用弱标注信息并结合有效的训练策略和模型设计来缩小其与全监督语义分割之间的性能差距(Shen等,2023)。本文旨在对不同类型的WSSS方法进行分类和总结,分析其存在的问题和挑战,为进一步推动WSSS领域的研究和应用提供参考。根据所提供监督信息的多寡,传统WSSS方法主要分为4个类别:基于图像级标注、最小包围盒标注、线标注和点标注的WSSS。此外,由于近年来大模型(Zhao等,2023)的兴起和快速发展,其在WSSS方面的应用也越来越引起众多学者的关注(Chen等,2023a)。因此,本文并没有采用传统的类别划分方法(任冬伟等,2022),而是根

据标注类型和受关注的程度，从图像级标注、其他弱标注以及大模型辅助的 WSSS 这 3 个方面回顾现有 WSSS 方法。其中，第 2 个类别又包括基于最小包围盒标注、线标注和点标注的 WSSS 方法。

本文组织结构如图 2 所示。第 1 节介绍相关数据集和评估指标；第 2 节回顾现有基于图像级标注的 WSSS 方法；第 3 节介绍基于其他弱标注的 WSSS 方法；第 4 节介绍基于大模型的 WSSS 方法；第 5 节总结并讨论现有 WSSS 领域存在的问题和挑战，以及未来可能的研究方向。

(a) 图像级标注　(b) 最小包围盒标准　(c) 线标注　(d) 点标注

图 1　弱监督语义分割标注示意图

Fig. 1　Examples of weak annotations for segmentation

((a)image-level annotation;(b)bounding-box annotation;(c)scribble annotation;(d)point annotation)

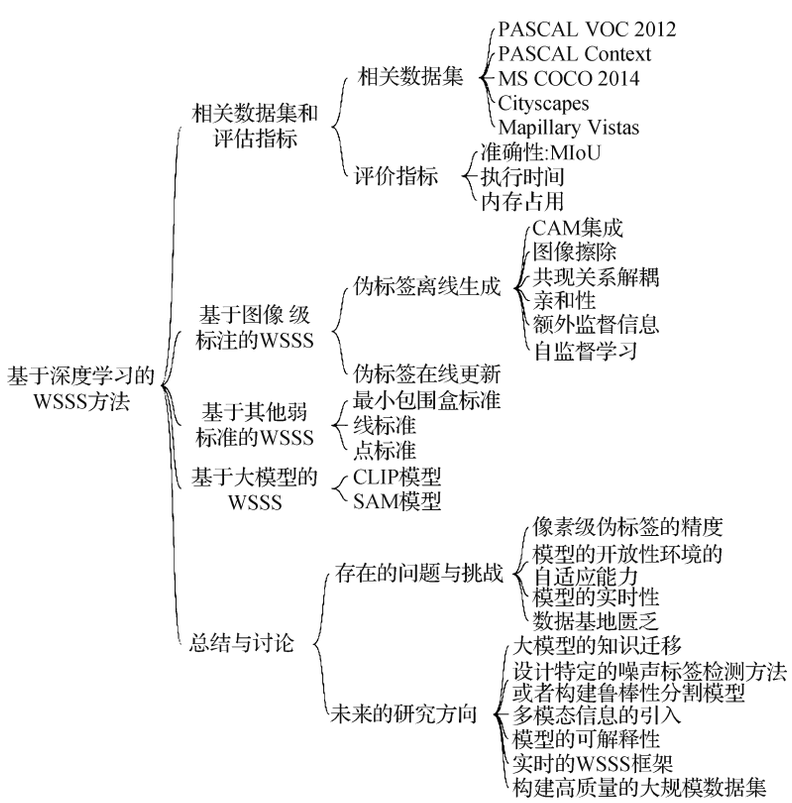

图 2　本文组织结构

Fig. 2　The structure of this paper

1　相关数据集和评估指标

1.1　相关数据集

在语义分割领域，常用的数据集主要包括 PASCAL VOC 2012（pattern analysis, statistical modeling and computational learning visual object classes 2012）（Everingham 等，2015）、PASCAL Context（Mottaghi 等，2014）、MS COCO 2014（Microsoft common objects in context 2014）（Lin 等，2014）、Cityscapes

(Cordts 等, 2016) 和 Mapillary Vistas (Neuhold 等, 2017)等。

PASCAL VOC 2012 数据集是目前最常用的语义分割数据集之一,常作为该领域模型的评估标准。该数据集总共包含 1 464 幅训练图像、1 449 幅验证图像和 1 456 幅测试图像,涉及人、动物、车辆和日常用品等20个常见类别。

PASCAL Context 数据集是在 PASCAL VOC 2010 基础上构建的语义分割数据集,包含 540 个类别。但是由于其中大多数类别数据太过稀疏,研究者们通常选用其中出现频率最高的 59 种类别作为语义标签,而将其余所有的类别全部当成背景。它总共 10 103 幅图像,其中 4 998 幅用于训练,5 105 幅用于验证。

MS COCO 2014 数据集是一个更具挑战性的大规模语义分割数据集。该数据集包含 80 个类别,涉及人、动物、交通工具、电子产品等。其中 83 K 幅图像用于训练,40 K 幅图像用于验证,41 K 幅图像用于测试。

Cityscapes 数据集是一个以城市街景为主的语义分割数据集。该数据集包含 30 个类别,涵盖道路、车辆、行人和交通标志等元素。它包含 5 000 幅具有精细像素级标注的图像,其中 2 975 幅用于训练,500 幅用于验证,1 525 幅用于测试。

Mapillary Vistas 数据集是一个大规模街景图像语义分割数据集,包含 25 K 幅高分辨率街景图像,涉及车辆、行人、标志牌等 65 个语义类别。其中,18 K 幅图像用于训练,2 K 幅图像用于验证、5 K 幅图像用于测试。

1.2 评估指标

一个有效的评估指标对于弱监督语义分割模型性能的衡量是至关重要的。常用的评价指标主要包括平均交并比(mean intersection over union, mIoU)、执行时间和内存占用等。

mIoU 通常用于衡量弱监督语义分割模型的准确性,具体计算为

$$mIoU = \frac{1}{k+1}\sum_{i=0}^{k}\frac{X \cap Y}{X \cup Y} \tag{1}$$

式中,X 表示真实结果,Y 表示预测结果,k 表示总类别个数。本文在后续的性能比较中报告了相关模型的 mIoU 结果。

模型的执行时间也是评判其性能的重要指标之一,但大多数 WSSS 的研究论文没有给出相应的结果,这可能是由于执行时间受硬件配置的影响。如果硬件配置不同,那么执行时间的结果将没有意义。

内存占用是指模型在设备中运行时所需的内存峰值。如果内存占用过大,那么只有最新的硬件设备才能满足需求,这对模型的适用性会产生不利影响。然而,几乎没有任何研究论文提及其模型的内存占用情况。

2 基于图像级标注的弱监督语义分割

图像级标注仅提供图像中所含的物体类别信息,缺少物体在图像中所处的位置信息。相比于像素级标注,图像级标注极大地降低了数据标注的人力、时间和资金成本。然而,由于图像级标注提供的监督信息最少,在这种监督信息下的 WSSS 任务也最具挑战性(任冬伟 等,2022)。该任务的关键在于如何将图像类别信息转化为对应目标的位置信息。

如图 3 所示,现有方法主要遵循两阶段的流程。其中,在第 1 阶段,主要采用 Zhou 等人(2016)提出的类激活图(class activation map, CAM)方法,获得初始种子区域以初步定位目标位置,然后通过对该初始种子区域进行优化得到高质量的像素级伪标签。

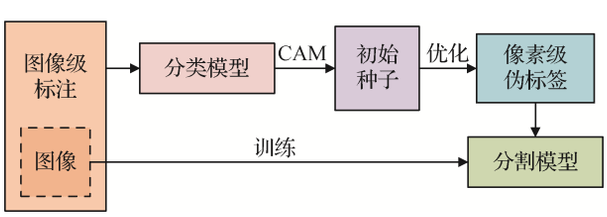

图 3 基于图像级标注的弱监督语义分割流程图(Shen 等, 2023)

Fig. 3 The mainstream pipeline for semantic segmentation with image-level annotations (Shen et al., 2023)

生成 CAM 初始种子区域的具体流程如图 4 所示,图中 $W_1, W_2, \cdots, W_n$ 分别表示分类器关于类别 c 的权重参数。首先从分类模型骨干网的最后一层卷积层中获取特征图,然后使用分类器关于类别 c 的权重参数对该特征图进行加权,获得原图像关于类别 c 的 CAM。该 CAM 反映了图像中每个位置属于类别 c 的概率,通过筛选 CAM 中的高置信度区域,生成初始种子区域,实现图像类别信息到位置信息的转化。在第 2 阶段,使用第 1 阶段生成的像素级伪标签作为

监督信息训练语义分割模型。

根据第2阶段语义分割模型训练过程中像素级伪标签是否需要进行更新，基于图像级标注的WSSS方法又可以分为：伪标签离线生成（Jiang等，2019）和伪标签在线更新（Ru等，2022）两个类别。伪标签离线生成方法认为两阶段模型相互独立，假设在第1阶段就可以得到足够准确的伪标签用于训练第2阶段分割网络。相比之下，伪标签在线更新方法认为第一阶段伪标签生成的质量无法保证，需要迭代更新两阶段模型进行联合优化。因此这类方法认为两阶段模型并不独立，第1阶段伪标签生成和第2阶段分割结果需要彼此交互。接下来，本文将回顾和总结这两类方法的研究进展。

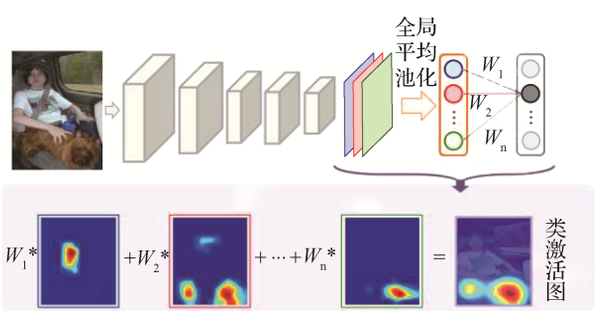

图4　生成CAM初始种子区域的流程图（Zhou等，2016）

Fig. 4　The pipeline for generating CAM initial seed areas (Zhou et al., 2016)

2.1　基于伪标签离线生成的图像级弱监督语义分割

基于伪标签离线生成的图像级WSSS方法试图在第1阶段中生成最优的像素级伪标签，在第2阶段中保持像素级伪标签不变，并利用这些像素级伪标签来训练语义分割模型。由于CAM初始种子区域直接从分类网络生成，一般只关注图像中的辨别性区域，准确性较低，无法直接作为像素级伪标签。因此，伪标签离线生成的图像级WSSS主要研究如何从粗糙的CAM种子区域生成更加准确的像素级伪标签。这些生成策略大致可以分为以下6类：1）基于CAM集成的方法；2）基于图像擦除的方法；3）基于共现关系解耦的方法；4）基于亲和性的方法；5）基于额外监督信息的方法；6）基于自监督学习的方法。

2.1.1　基于CAM集成的方法

由于CAM通常无法覆盖对应目标的全部区域，即存在欠激活现象，因此一种直观的方法是通过不同CAM的集成来获得更完整的目标激活区域。

Wei等人（2018）发现不同空洞率的空洞卷积不仅可以有效地扩大卷积核的感受野，还可以提高分类模型关注非辨别性目标区域的能力。基于这个发现，Wei等人（2018）提出使用不同空洞率的空洞卷积来生成不同的CAM，并对其进行集成以获得更完整的目标激活区域。Lee等人（2019）提出的FickleNet（fickle network）随机隐藏特征图中某些位置的特征表达，探索特征图中不同位置组合对目标区域激活的影响，并集成各种位置组合所生成的CAM，从而识别物体的辨别性区域和非辨别性区域。

与Wei等人（2018）和Lee等人（2019）的方法不同，Jiang等人（2019）发现由分类模型产生的CAM激活区域在模型训练过程中会不断变化。因此他们提出一种OAA（online attention accumulation）策略，在训练过程中为训练图像的每个类别在线维护一个CAM以对之前所有阶段的CAM进行集成。随着训练过程的进行，该CAM会发现更完整的目标激活区域。在此OAA策略的基础上，Jiang等人（2022a）还加入了注意力丢弃层以扩大模型训练过程中CAM激活区域的变化范围，有利于激活图像中更多非辨别性区域。

2.1.2　基于图像擦除的方法

由于分类模型通常只关注图像中最具辨别性的区域，有的学者提出在训练过程中根据当前CAM中的激活区域对图像进行擦除，强制分类模型寻找图像中新的激活区域，最后合并新激活区域和旧激活区域以实现对目标物体的完整覆盖。

Kim等人（2017）提出了两阶段像素级伪标签生成方法：在第1阶段，通过传统CAM方法找到图像中最具辨别性的物体区域；在第2阶段，通过擦除第1阶段所发现的辨别性区域的图像特征从而迫使分类模型发现新目标区域。最后组合两个阶段的CAM以捕获更完整的目标区域。Kweon等人（2021）通过实验得出：在仅擦除图像中某一类物体最具辨别性区域的情况下，该分类器也能发现其他类别物体的非辨别性区域。在此基础上，Kweon等人提出基于特定类别的对抗擦除框架，随机选择单个需要擦除的目标类别并在原图像上进行擦除，激发分类器发现其他物体类别非辨别性区域的潜能。

在图像擦除过程中，CAM可能会扩展到背景区域。为此，一些学者（Wei等，2017；Hou等，2018；Sun等，2021）设计了一系列改进的图像擦除方法。Wei

等人(2017)在引入对抗性擦除的基础上,提出PSL(prohibitive segmentation learning)方法,使用预测的分类置信度对生成的不同CAM进行加权,缓解不可靠CAM所带来的错误干扰。Hou等人(2018)提出了一个简单而有效的自擦除网络,利用常见场景的背景特征具有相似性这一先验知识有目的地抑制CAM向背景区域的扩散,鼓励网络从特定区域而不是整幅图像中发现更完整的目标区域。Sun等人(2021)提出的Ecs-Net(erased CAM supervision network)从原始图像生成的CAM中采样高置信度的像素作为额外监督信息来指导图像擦除后CAM的生成,缓解CAM的错误扩散问题。

2.1.3 基于共现关系解耦的方法

虽然上述两种方案缓解了CAM无法覆盖目标整体区域的问题,但CAM仍存在可能在非目标类别物体上出现错误激活的缺陷,即存在过激活现象。其原因可能是存在所谓的共现类别。例如,“马”和“人”经常出现在同一幅图像中,“地板”和“沙发”也经常出现在同一幅图像中,“马”和“人”、“地板”和“沙发”分别构成2对共现类别。这种共现类别的存在会误导分类模型关注错误的目标区域,造成CAM的过激活。

一些方法(Zhang等,2020b;Su等,2021;Chen等,2022b)尝试通过解耦图像类别间这种共现关系来优化CAM以获得高质量的像素级伪标签。Zhang等人(2020b)构建了一个因果推理框架来实现对图像、上下文信息和图像级标注之间的因果关系分析,并在此基础上提出了CONTA(context adjustment)方法,解耦了图像共现类别间的因果关系,以消除分类模型中存在的共现问题。Su等人(2021)从数据增强的角度出发,提出了一种CDA(context decoupling augmentation)方法,通过复制粘贴的操作来消除前景物体和上下文信息之间的依赖性。具体而言,他们将一幅图像中的前景物体粘贴到不同的图像上,从而使得图像中的前景物体与其余背景部分没有固定的上下文关系,鼓励分类模型更多地关注前景物体本身而不是其共现类别或背景。针对医学图像前景和背景边界模糊以及共现现象严重这两个特点,Chen等人(2022b)提出了C-CAM(causal class activation map)方法,该方法由两个因果链驱动,其中类别因果链使用因果干预迫使分类模型关注前景物体,结构因果链通过形状先验以约束前景物体的边界,二者结合进一步提升了医学图像分割的性能。

2.1.4 基于亲和性的方法

虽然上述方案一定程度上解决了CAM的欠激活和过激活问题,但没有考虑图像元素间的相似度。在图像级WSSS领域中,一般利用亲和性来表示图像像素之间或patch之间的相似度。它可以通过计算颜色、纹理等低级视觉特征或者语义信息等高级视觉特征之间的距离来表达。

在基于低级视觉特征的亲和性应用方面,Kolesnikov和Lampert(2016)提出的边界损失函数,利用颜色和空间位置等低级视觉特征的亲和性来约束像素级伪标签的生成,缓解其边界模糊的问题。Huang等人(2018)提出的DSRG(deep seeded region growing)方法使用种子区域生长策略将种子区域中高置信度像素的语义信息传递给具有相似颜色或纹理特征的相邻像素,以生成更完整的像素级伪标签。

在基于高级语义信息的亲和性应用方面,Ahn和Kwak(2018)和Xu等人(2022)作出了一定的贡献。其中,Ahn和Kwak(2018)设计了AffinityNet(affinity network)用于预测相邻像素的语义亲和性,并通过随机游走策略将语义信息从CAM中的激活区域传到未激活区域,生成了更为准确的像素级伪标签。此外,AffiniNet是以CAM中的高置信度区域作为监督信息进行训练的,并不需要额外的训练数据和标注信息。Xu等人(2022)提出了一种基于Transformer(Vaswani等,2017)的MCTformer(multi-class token transformer)网络,创新地使用多个类别token学习不同类别间的辨别性表示,并借助Transformer的自注意力机制学习patch之间的语义亲和性用于进一步细化特定类别的注意力图,从而显著提高伪标签的准确性。

考虑到单幅图像所蕴含的亲和性信息的局限性,一些方法(Sun等,2020;Wu等,2021;Li等,2021;Zhou等,2022)将亲和性关系扩展到不同图像之间,以便从多个图像中获取更全面和准确的亲和性关系,从而提高生成伪标签的准确性。Sun等人(2020)在分类器中集成了两个注意力模块:一个协同注意力模块用于从给定的一对图像中识别语义相同的区域;另一个对比注意力模块强调图像中特有语义类别的学习。通过这两个注意力模块的互补性完成图像间语义亲和性的准确建模,从而更好地挖掘图像间的语义关系并优化伪标签的生成过程。

Wu等人(2021)提出一个CMA(collaborative multi attention)模块,借助自注意力机制将单幅图像内像素间亲和性建模推广到整个Batch内所有图像。Li等人(2021)将图神经网络(Scarselli等,2009)引入亲和性关系建模过程。其中输入图像用图节点表示,图像间亲和性关系用图的边表示,利用图神经网络和共同注意力机制实现任意数量图像的亲和性关系建模,以估计更可靠的伪标签。同时,为了防止模型过度关注一组图像中的相同语义区域,Li等人进一步提出图丢弃层,促进图像中特有类别的学习。Zhou等人(2022)进一步将图像间的亲和性关系扩展到数据集级别,提出了一种RCA(regional contrast and aggregation)方法。通过记忆存储库存储整个数据集中不同类别的特征表达,并对其进行K-Means聚类(MacQueen,1967)以获得不同类别物体的代表性原型。借助这些代表性原型与输入图像的亲和性关系来增强模型对输入图像的语义理解,从而获得更准确的像素级伪标签。

2.1.5 基于额外监督信息的方法

上述方法只是从图像特征角度着手解决图像级WSSS问题,并没有考虑引入额外的监督信息。为此,许多学者尝试引入物体边界以及显著性图等额外监督信息来缩小弱监督与全监督之间的差距。

1)物体边界信息。Chen等人(2020)和Li等人(2022a)通过引入生成的边界信息来约束CAM的优化。Chen等人(2020)通过生成的CAM获得少量的边界标签,接着使用该边界标签训练BENet(boundary exploration network)以预测更多的物体边界,最后使用这些预测的边界信息为CAM的优化提供约束。Li等人(2022a)提出了SANCE(segmentation assisted noiseless contour exploration)模型,该模型包含两个分支:轮廓分支和分割分支。轮廓分支预测边界信息,分割分支预测物体定位图。二者相辅相成,从而生成更可靠的像素级伪标签。Rong等人(2023)从噪声标签学习的角度出发,提出了边界增强策略,利用生成的边界信息对分割模型的预测结果进行约束,提高了边界区域的预测准确性。

2)显著性图。一些方法(Joon Oh等,2017;Yu等,2019;Lee等,2021c;Chen等,2023c)将显著性图引入WSSS领域,以探索物体分割的边界信息。与直接预测物体边界的方法不同,这类方法通常需要一个额外的显著性检测模型。Joon Oh等人(2017)将图像级WSSS问题分解为两个独立的问题:确定物体位置和确定物体边界。先通过CAM中的高置信度区域定位前景物体和背景区域的初始位置,接着使用训练好的显著性检测模型预测类不可知的物体显著性以确定边界,最后结合位置和边界信息实现类特定的像素级伪标签生成。Yu等人(2019)提出了一个统一的多任务学习框架SSNet(saliency and segmentation network)。通过显著性聚合模块关联图像级WSSS任务和显著性检测任务,实现两者的联合建模和优化。Lee等人(2021c)提出的EPS(explicit pseudo-pixel supervision)方法同样利用了CAM与显著性图之间的互补性,通过显著性图提供的边界信息来约束CAM的扩张。Chen等人(2023c)提出了I2CRC(inter- and intra-class relation constrained)框架,使用显著性引导的类不可知距离模块拉近类内特征的距离,并引入一个类特定的距离模块增强类间特征的差异性,获得更完整的像素级伪标签。

2.1.6 基于自监督学习的方法

自监督学习(Jing和Tian,2021)是一种无需人工标注的训练方法,通过设计特定的网络架构和损失函数,模型能够从未标注的训练数据中学习潜在特征。由于其无需人工标注的特性,自监督学习可以很好地应用于图像级WSSS领域。

孪生网络(siamese networks)(Grill等,2020)从网络架构设计的角度将自监督学习引入到WSSS任务。它通常由两个或多个子网络组成,每个输入样本经过各自的子网络进行编码,通过比较编码后的输出使得网络学习特征之间的相似性和差异性,从而增强特征表示的一致性。Wang等人(2020)提出的SEAM(self-supervised equivariant attention mechanism)方法采用了Siamese网络结构,并构建了跨视图正则化损失函数,对原图像CAM和经过数据增强变换后图像的CAM进行一致性正则化,有效缓解了CAM的过激活和欠激活问题。Jo和Yu(2021)提出了Puzzle-CAM方法,将原图像裁剪成4个子图并对其重组得到重组图像。通过重构正则化损失以最小化重组图像和原图像之间的差异,从而获得更完整的目标激活。Zhang等人(2021b)提出了一种由3个子网络构成的CPN(complementary patch network)方法,采用随机掩码的方式生成两张互补图像,并通过3个正则化项缩小两张互补图像生成的CAM与原图

像CAM之间的差距。Jiang等人(2022b)发现当使用图像的局部区域替换整个输入图像时,分类模型可以关注到物体的更多细节。由此他们提出了L2G(local to global)方法,先利用局部网络从输入图像的多个局部区域提取细节信息,然后利用构建的迁移损失鼓励全局网络在线学习局部网络提取的细节信息,从而产生高质量的CAM,提高分割的性能。Peng等人(2023)提出了一种自适应网络调整策略,通过对分类网络架构进行不同程度的调整,生成图像的不同视图,从而构建一致性正则化以实现自监督学习。

与上述引入Siamese网络结构实现自监督学习的方法不同,一些方法(Du等,2022;Xie等,2022b)从损失函数设计的角度解决图像级WSSS问题。其代表性工作就是采用对比学习算法(He等,2020)。对比学习通过最大化类间特征距离,同时最小化类内特征距离的方式增强特征表达能力,实现前景目标和背景区域的区分,从而生成更准确的像素级伪标签。Du等人(2022)提出了一种像素到原型的对比学习方法,使得图像中每个像素接受来自每个类别可靠原型的监督,并通过数据增强手段引入跨试图一致性正则化,显著提高了CAM和像素级伪标签的质量。Xie等人(2022b)提出的C^2AM(class-agnostic activation map)方法利用无标签图像生成类不可知激活图,接着通过该激活图获得图像的前景特征表示和背景特征表示,最后使用对比学习建模前景和背景之间的关系,为CAM的优化提供指导。

还有一些学者(Chang等,2020;Zhang等,2020c;Lee等,2021a;Chen等,2022a)认为仅仅设计鲁棒的损失函数不足以得到精确的伪标签,进而通过设计一系列正则化项来进一步提升伪标签的精度。Zhang等人(2020c)提出了一种拆分与合并的策略,通过构建的差异正则化项对同幅图像的不同CAM进行约束,有利于物体整体区域的挖掘。Chang等人(2020)引入了一种自监督的子类别探索任务,将一个类别分成几个不同的子类别,使用图像特征的聚类中心为每类前景物体分配各自的子类别标签,促进分类模型对前景目标整体区域的关注。Lee等人(2021a)提出了一种AdvCAM(adversarial class activation map)方法,通过反对抗的方式优化图像的CAM以增加图像对应类别的分类得分,并构建了一种正则化方式来抑制无关区域的错误激活和辨别性区域的过度激活。Chen等人(2022a)提出的SIPE(self-supervised image-specific prototype exploration)方法为每幅图像估计原型并用其对该图像进行重新激活以生成IS-CAM(image-specific class activation map),使用GSC(general-specific consistency)对原CAM和IS-CAM进行一致性正则化,增强模型的自校正能力。

2.2 基于伪标签在线更新的图像级弱监督语义分割

基于伪标签在线更新的图像级WSSS方法采用端到端的训练策略,通过训练一个模型,将像素级伪标签生成过程和语义分割模型训练过程进行联合优化,在更新语义分割模型参数的同时优化像素级伪标签,最终达到二者的联合最优结果。这类方法不需要对像素级伪标签进行离线处理,训练流程简单。

Araslanov和Roth(2020)总结了图像级弱监督语义分割方法设计的3大准则:局部一致性、语义准确性和完整性。局部一致性是指具有相似外观的相邻像素应该共享相同的语义标签;语义准确性是指模型关注的区域应该具有正确的语义标签;完整性指的是模型应当关注物体的整体区域而不是局部。Zhang等人(2020a)提出了一种双分支方法RRM(reliable region mining),其中分类分支用于生成CAM并从中选择可靠区域作为分割分支的监督信息,分割分支用于实现图像的像素级分割。此外,Zhang等人(2020a)还构建了一个新损失函数用于联合局部信息和全局信息。在RRM方法的基础上,Zhang等人(2021c)提出了两种新损失函数:自适应亲和性损失用于建模图像像素之间的关系,增强语义传播的可靠性;标签重分配损失用于识别错误的像素级伪标签,防止网络过拟合该错误标签。

卷积神经网络(convolutional neural network,CNN)采用较小卷积核进行局部卷积运算,因此无法探索图像全局信息;反观Transformer网络,由于其采用的多头自注意力机制(multi-head self-attention,MHSA)能够编码所有token之间的相互关系,因而具有全局感受野的优势(Vaswani等,2017)。因此,Ru等人(2022)提出了AFA(affinity from attention)模块,从Transformer的MHSA中学习全局语义亲和性用于细化像素级伪标签,并设计了一个像素自适应细化模块利用图像低级视觉信息增强像素级伪标签的局部一致性。在此基础上,Ru等人(2023)通过实

验发现虽然 ViT(vision transformer)(Dosovitskiy 等，2021)可以弥补 CNN 局部感受野的缺陷，但同时也会带来过度平滑的问题。基于这一发现，Ru 等人提出了一种 ToCo(token contrast)方法，引入了 PTC(patch token contrast)模块和 CTC(class token contrast)模块。PTC 模块通过网络中间层的语义多样性信息监督最终的 patch token，以缓解 ViT 引起的过渡平滑问题。CTC 模块通过对比整幅图像和局部区域的类别 token 来促进其一致性，加强模型对前景目标非辨别性区域的关注。

Chen 等人(2023b)认为通过优化 CAM 种子区域生成像素级伪标签的方法既复杂又耗时，限制了模型的高效性，因此提出了 MDBA(multi-granularity denoising and bidirectional alignment)方法。通过显著性图直接生成伪标签，并构建一种端到端的多粒度去噪模块用于解决显著性图存在的噪声问题。此外，Chen 等人(2023b)还考虑到显著性图不包含物体的类别信息，无法适用于包含多类物体的复杂图像处理过程。为解决这个问题，他们提出了一种双向对齐机制：在输入端将包含单类物体的简单图像合成为复杂图像用作模型的训练数据；在输出端合成相应简单图像的显著性图，以此作为像素级伪标签来监督复杂图像的分割过程。

2.3 讨论

图像级 WSSS 在仅使用图像级标注的情况下实现了像素级的分割，极大程度地降低了标注成本。然而，由于图像级标注仅包含目标类别信息，缺乏目标位置信息，该任务也兼具很大的挑战性。表 1 展示了一些代表性的图像级 WSSS 方法在 PASCAL VOC 2012(Everingham 等，2015)验证集和测试集上的性能表现。

基于伪标签离线生成的图像级 WSSS 方法先通过图像级标注获得最优的像素级伪标签，然后基于得到的最优像素级伪标签训练语义分割模型，并且在语义分割模型的训练过程中不再更新像素级伪标签。这种方法的模型训练流程较为复杂，无法进行端到端的训练。早期基于 CAM 集成、图像擦除等技术的伪标签离线生成方法主要依赖于图像级标注自身提供的有限监督信息，其性能提升空间有限，例如 Lee 等人(2019)提出的 FickleNet 方法仅获得了 65.3% 的 mIoU。近年研究中通过引入额外监督信息或自监督学习的方式，性能提升显著，例如 Du 等人(2022)的方法取得了 73.6% 的 mIoU。这主要是由于引入额外监督信息的方法提供了新的监督信号，而自监督学习在无需标注的情况下学习数据的内在结构特征。这些方法打破了图像级标注的限制，因此获得了更显著的性能提升。但是，这些方法仍存在一些问题，例如 CAM 种子区域在处理小目标、中间镂空的物体和具有复杂形状的物体时可能非常不准确，给后续生成像素级伪标签的过程带来了很大的挑战。

基于伪标签在线更新的图像级 WSSS 方法不需要事先生成固定的像素级伪标签，而是在整个模型训练过程中进行像素级伪标签和分割结果的联合优化，模型训练流程较为简单。然而，目前这类方法的准确性稍有不足，例如 Ru 等人(2022)提出的 AFA 方法仅获得 66.3% 的 mIoU，与伪标签离线生成方法的最优性能相比仍存在一定的差距。这主要受限于模型训练的难度。未来可以探索如何优化模型结构或提出新的训练策略以降低其训练难度，充分发挥该方法的潜力。此外，值得关注的是，Ru 等人(2023)提出的 ToCo 方法获得了 71.2% 的 mIoU，性能明显领先其他伪标签在线更新方法。这主要归因于其设计的 PTC 和 CTC 模块以及采用 Transformer 骨干网所带来的强大特征表达能力。

3 基于其他弱标注的弱监督语义分割

其他弱标注包括最小包围盒标注、线标注和点标注。相比于图像级标注，这些弱标注提供了较为丰富的监督信息，有助于更有效地预测像素级分割结果。本节将详细介绍并比较这类 WSSS 方法。

3.1 基于最小包围盒标注的弱监督语义分割

最小包围盒标注是一种贴合物体四周边界的最紧致矩形框的标注形式。它不仅提供物体的类别标签，还额外包含物体的数量和粗略位置信息，是一种比图像级标注更强大的监督信号。因此这类方法往往可以取得比图像级标注方法更优异的性能(任冬伟 等，2022)。由于最小包围盒外的图像区域全部被划分为背景，而最小包围盒内同时存在前景和背景区域，所以基于最小包围盒标注的 WSSS 任务的关键在于如何准确地区分最小包围盒内的前景物体和背景区域。如图 5 所示，目前该领域分割方法主要包括两个步骤：1)从最小包围盒中挖掘前景物体

表1 基于图像级标注的弱监督语义分割方法PASCAL VOC 2012上的性能表现

Table 1 Performance comparison of weakly supervised semantic segmentation methods based on image-level annotations on PASCAL VOC 2012 dataset

/%

伪标签处理方式	方法	发表	骨干网	mIoU	
				验证集	测试集
伪标签离线生成	Kolesnikov和Lampert(2016)	ECCV2016	VGG16 / VGG16	50.7	51.7
	Kim等人(2017)	ICCV2017	VGG16 / VGG16	53.1	53.8
	Wei等人(2017)	CVPR2017	VGG16 / VGG16	55.0	55.7
	Joon Oh等人(2017)	CVPR2017	VGG16 / -	55.7	56.7
	Wei等人(2018)	CVPR2018	VGG16 / VGG16	60.4	60.8
	SeeNet(Hou等,2018)	NeurIPS2018	VGG16 / ResNet101	63.1	62.8
	DSRG(Huang等,2018)	CVPR2018	VGG16 / ResNet101	61.4	63.2
	AffinityNet(Ahn和Kwak,2018)	CVPR2018	ResNet38 / ResNet38	61.7	63.7
	FickleNet(Lee等,2019)	CVPR2019	VGG16 / ResNet101	64.9	65.3
	SEAM(Wang等,2020)	CVPR2020	ResNet38 / WideResNet38	64.5	65.7
	Chang等人(2020)	CVPR2020	ResNet38 / ResNet101	66.1	65.9
	Jiang等人(2019)	ICCV2019	VGG16 / ResNet101	65.2	66.4
	Chen等人(2020)	ECCV2020	ResNet50 / ResNet101	65.7	66.6
	CONTA(Zhang等,2020b)	NeurIPS2020	ResNet38 / WideResNet38	66.1	66.7
	Zhang等人(2020b)	ECCV2020	VGG16 / ResNet50	66.6	66.7
	CDA(Su等,2021)	ICCV2021	ResNet38 / WideResNet38	66.1	66.8
	Sun等人(2020)	ECCV2020	ResNet38 / ResNet101	66.2	66.9
	Ecs-Net(Sun等,2021)	ICCV2021	ResNet38 / ResNet38	66.6	67.6
	AdvCAM(Lee等,2021a)	CVPR2021	ResNet50 / ResNet101	68.1	68.0
	Kweon等人(2021)	ICCV2021	ResNet38 / ResNet38	68.4	68.2
	CPN(Zhang等,2021b)	ICCV2021	ResNet38 / ResNet38	67.8	68.5
	Li等人(2021)	AAAI2021	- / ResNet101	68.2	68.5
	I2CRC(Chen等,2023c)	TMM2023	VGG16 / ResNet101	69.3	69.5
	SIPE(Chen等,2022a)	CVPR2022	ResNet50 / ResNet101	68.8	69.7
	Wu等人(2021)	CVPR2021	ResNet38 / ResNet101	70.9	70.6
	MCTformer(Xu等,2022)	CVPR2022	DeiT-S / WideResNet38	71.9	71.6
	L2G(Jiang等,2022b)	CVPR2022	ResNet38 / ResNet101	72.1	71.7
	EPS(Lee等,2021c)	CVPR2021	ResNet38 / ResNet101	71.0	71.8
	RCA(Zhou等,2022)	CVPR2022	ResNet38 / -	72.2	72.8
	SANCE(Li等,2022a)	CVPR2022	ResNet101 / ResNet101	72.0	72.9
	Du等人(2022)	CVPR2022	ResNet38 / ResNet101	72.6	73.6
伪标签在线更新	RRM(Zhang等,2020a)	AAAI2020	ResNet38 / ResNet38	62.6	62.9
	Araslanov和Roth(2020)	CVPR2020	WideResNet38 / WideResNet38	62.7	64.3
	Zhang等人(2021c)	ACMMM21	WideResNet38 / WideResNet38	63.9	64.8
	AFA(Ru等,2022)	CVPR2022	MiT-B1 / MiT-B1	66.0	66.3
	Chen等人(2023b)	TIP2023	ResNet101 / ResNet101	69.5	70.2
	ToCo(Ru等,2023)	CVPR2023	ViT-B / ViT-B	71.1	71.2

注:"/"两边分别表示用于伪标签生成和图像分割的骨干网。"-"表示论文中未提及。

和背景区域的信息，生成像素级伪标签；2）使用该像素级伪标签训练全监督语义分割模型，实现目标图像的分割。

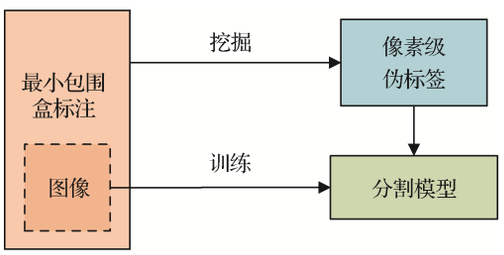

图 5　基于最小包围盒标注的弱监督语义分割流程图（Shen 等，2023）

Fig. 5　The mainstream pipeline for semantic segmentation with bounding-box annotations（Shen et al.，2023）

Dai 等人（2015）最早对该领域进行了探索，提出了 BoxSup（bounding-box supervision）方法。第 1 步采用 MCG（multiscale combinatorial group）（Arbeláez 等，2014）等无监督区域建议方法提取一系列物体候选区域；第 2 步选择其中与最小包围盒重叠区域最大的候选区域作为初始像素级伪标签；第 3 步将其作为监督信号训练一个语义分割模型；第 4 步使用该模型的预测结果更新像素级伪标签。以上第 2—4 步进行重复迭代直到最后收敛。该方法通过像素级伪标签和语义分割模型的交替更新，不断提高像素级伪标签的准确率和语义分割模型的性能。与 Dai 等人（2015）的方法不同，Papandreou 等人（2015）采用 DenseCRF（dense conditional random field）（Krähenbühl 和 Koltun，2011）方法通过最小包围盒生成像素级伪标签，并采用交叉验证方法来调整 DenseCRF 的超参数，尽可能地提高像素级伪标签的准确率。此外，Papandreou 等人（2015）还提出了一种改进的 EM（expectation maximization）（Dempster 等，1977）算法，交替更新像素级伪标签和模型参数。该方法在仅使用少量像素级标注的情况下有效地提升了模型的分割性能。Khoreva 等人（2017）将最小包围盒视为带噪声的像素级标注，将迭代训练视为一种去噪策略。在这种思想下，他们提出了一种简洁高效的 SDI（simple does it）方法，在无需迭代训练的情况下获得了优异的分割性能。具体而言，他们提出了一种改进的 GrabCut 方法，称为 GrabCut+，并将 GrabCut+和 MCG 进行结合，选择二者所得候选区域的交集部分作为像素级伪标签训练语义分割模型，充分利用了 GrabCut+和 MCG 的可靠性。

还有一些方法（Song 等，2019；Kulharia 等，2020）通过计算每个前景物体在最小包围框中的填充率来指导分割过程。Song 等人（2019）设计了一种填充率引导的自适应损失，利用每个类别的平均填充率引导模型关注图像中响应程度较强的区域，减轻错误标记像素对模型的干扰。同时，考虑到来自同一类的两个物体可能由于形状和姿势的变化而具有不同的填充率，Song 等人（2019）通过聚类的方法将每个语义类别分成几个子类来进一步细化填充率，充分发挥填充率对像素级伪标签的指导作用。此外，为了消除无关区域对分割模型预测结果的影响，Song 等人（2019）还提出了 BCM（box-driven class-wise masking）模块，为每个前景物体分别生成一个掩码用于屏蔽无关区域，并提供其形状和位置信息，极大程度地降低了后续图像分割流程的难度。为了缓解 WSSS 中像素级伪标签的噪声问题，Kulharia 等人（2020）通过预测每类物体的注意力图来优化像素级交叉熵损失，并计算每类物体在最小包围框中的填充率来约束注意力图的生成，从而使注意力图更好地聚焦前景区域，减少不正确的梯度传播。

Oh 等人（2021）认为图像中背景区域的特征表达具有一致性，这种一致性可以用来区分最小包围盒内的前景和背景区域，并由此提出了背景感知池化 BAP（background-aware pooling）方法。该方法使用注意力图区分最小包围盒内的前景和背景像素并聚合其中的前景特征，从而产生更准确的像素级伪标签。此外，由于 WSSS 中生成的像素级伪标签不可避免地包含噪声，Oh 等人（2021）还提出了噪声感知损失 NAL（noise-aware loss），减轻分割模型训练过程中错误伪标签的影响，增强了模型的抗噪能力。Lee 等人（2021b）借鉴了图像级 WSSS 中像素级伪标签的生成流程，提出了 BBAM（bounding-box attribution map）方法。首先使用最小包围盒标注训练一个目标检测模型，然后利用该模型从最小包围盒中提取某些特定区域以获得与整个最小包围盒区域相同的目标检测结果，最后基于这些特定区域使用 DenseCRF 生成像素级伪标签用于后续分割模型的训练。

3.2　基于线标注的弱监督语义分割

线标注是指在物体内部画一条连续的线段并标明物体类别作为监督信息。与图像级标注相比，线

标注提供了部分像素的位置信息;与最小包围盒标注相比,线标注缺乏物体的边界信息。由于线标注的稀疏性,所以基于线标注的WSSS关键在于如何将语义信息从稀疏的线段传递到其他未标记的像素(Shen等,2023)。如图6所示,目前该领域分割方法主要包括两个步骤:1)以线标注为初始种子区域,利用像素间亲和性等先验知识将线标注信息传递到其他未标记的区域,生成像素级伪标签;2)基于生成的像素级伪标签训练语义分割模型,实现图像的分割。

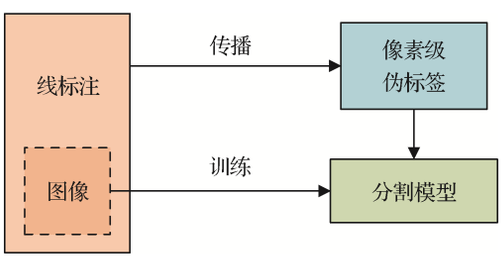

图6 基于线标注的弱监督语义分割流程图(Shen等,2023)

Fig. 6 The mainstream pipeline for semantic segmentation with scribble annotations(Shen et al., 2023)

由于图模型的结构特点和强大建模能力,Lin等人(2016)和Xu等人(2021)利用图模型将语义信息从线标注区域传到其他未标记的区域。为提高计算效率,Lin等人(2016)先将图像划分为若干个超像素,并基于低级视觉特征(如颜色和纹理)构建超像素之间的亲和性关系。接着通过图模型将语义信息从线标注区域传递到其他未标记区域,生成像素级伪标签。最后基于像素级伪标签训练分割模型,在整个训练过程中交替更新图模型和分割模型的参数直到实现最优的分割结果。Xu等人(2021)对Lin等人(2016)的方法做出改进,使用多尺度特征而不是低级视觉特征来构建超像素之间的亲和性关系,增强了语义信息传播的可靠性。

Vernaza和Chandraker(2017)提出的RAWKS(random-walk weakly-supervised segmentation)方法根据随机游走命中概率定义标签传播过程,从而生成像素级伪标签。此外,该方法还进一步集成了一个边界预测器以约束标签传播过程的边界,极大程度地提高了像素级伪标签的准确性。Ke等人(2021)将对比学习引入到线标注的WSSS任务,提出了一种像素到片段的对比学习方法SPML(semi-supervised pixel-wise metric learning)。该方法引入了4种类型的对比关系,即低级图像相似性、语义标注、共现性和特征亲和性。将具有相同语义的像素在特征空间中拉近,具有不同语义的像素在特征空间中分开,从而获得高质量的像素级特征表达和分割结果。

上述方法需要采用额外的模型来生成像素级伪标签。为简化训练流程,一些方法(Tang等,2018a;Tang等,2018b;Zhang等,2021a)通过设计恰当的损失函数直接从线标注预测像素级分割结果,提高了模型的训练效率。受半监督学习思想和浅层图像分割准则的启发,Tang等人(2018a)设计了两种新的损失函数:标注像素的部分交叉熵损失和未标注像素的归一化分割损失。第1个损失用于驱动模型从线标注中学习部分像素的正确语义信息,而第2个损失用于增强相邻像素的一致性,使得模型能够更有效地捕捉图像中相似部分的共性特征。随后,Tang等人(2018b)又提出了一个改进版本,引入了基于CRF(conditional random field)(Lafferty等,2001)的损失函数,避免了额外的CRF优化步骤,进一步简化了分割流程。由于大多数现有的正则化损失主要针对低级视觉特征,无法准确描述复杂情况下像素间的关系。为此,Zhang等人(2021a)提出了DFR(dynamic feature regularized)正则化损失,结合了图像的低级视觉特征和高级语义特征,以增强模型对不同像素间关系的表达能力。此外,Zhang等人(2021a)还设计了一个特征一致性模块,利用分割预测结果中的高置信度区域作为监督信息,以增强像素级特征的类内相似性和类间离散性,进一步提升模型对不同像素间关系的表达能力。

除Tang等人和Zhang等人之外,Wang等人(2019)也探索了如何从线标注直接预测像素级分割结果,并提出了BPG(boundary perception guidance)方法。该方法由边界回归和预测细化两个模块组成:边界回归模块使用类不可知的边界图来引导模型探索不同物体间的边界信息;预测细化模块通过迭代上采样策略生成多尺度特征图,结合高级语义特征和低级视觉特征以逐步改善像素级分割结果。

3.3 基于点标注的弱监督语义分割

点标注是指在图像内对每个物体分别标注一个点或几个点并标明类别信息作为监督信号,是线标注的一种简化形式。与线标注类似,点标注的WSSS任务的关键在于如何将语义信息从稀疏标注的点传

递到其他未标记像素。如图 7 所示，目前该领域分割方法主要包括两个步骤：1）通过标签传播、距离度量学习、损失函数优化等方法将点标注信息传递到其他未标记的区域，生成像素级伪标签；2）基于生成的像素级伪标签训练语义分割模型，实现图像的分割。

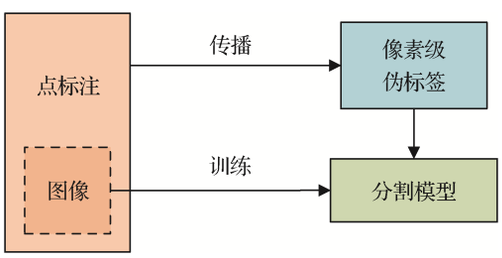

图 7　基于点标注的弱监督语义分割流程图（Shen 等，2023）

Fig. 7　The mainstream pipeline for semantic segmentation with point annotations（Shen et al.，2023）

Bearman 等人（2016）最早将点标注应用于 WSSS 任务。他们利用前景物体上的任意点作为监督信息，并引入物体性先验损失，通过计算图像中每个像素点属于前景物体的概率来区分图像中的前景和背景区域。McEver 和 Manjunath（2020）将图像级 WSSS 领域的 CAM 方法拓展到点标注领域。首先使用点标注的类别信息和位置信息来引导 CAM 的生成，称为 PCAM（point supervised class activation map），然后使用 IRNet（inter-pixel relation network）（Ahn 等，2019）优化 PCAM 以产生像素级伪标签，最后基于生成的像素级伪标签训练一个全监督语义分割模型。

Papadopoulos 等人（2017）和 Maninis 等人（2018）认为前景物体上任意点所包含的监督信息过于薄弱，由此提出了一种极点标注方法。通过标注图像中前景物体最上、最下、最左和最右这 4 个边界点来提供监督信息，以标注成本的少量增加换取了比任意点标注更丰富的监督信息。具体而言，Papadopoulos 等人（2017）将极点标注与 GrabCut 方法相结合，获得了优异的分割性能。Maninis 等人（2018）提出的 DEXTR（deep extreme cut）方法进一步探究了极点标注在语义分割、实例分割、视频分割和交互式分割这 4 个领域的应用，进一步验证了极点标注的有效性。

Qian 等人（2019）提出了一种基于点标注的弱监督场景解析方法，通过距离度量学习（Weinberger 和 Saul，2009）方法最小化同类嵌入向量间距离，并最大化不同类嵌入向量间距离。除此之外，Qian 等人（2019）还采用了跨图像的学习模式，以充分利用图像间的上下文信息，进一步优化嵌入向量的特征表达。Li 等人（2022a）提出了一种基于点标注的弱监督全景分割方法。通过将每个物体实例编码为特定的卷积核模板，并用该模板直接对高分辨率特征进行卷积，实现了前景物体和背景区域的统一预测和分割。

3.4　讨论

与图像级 WSSS 方法相比，基于最小包围盒标注、线标注和点标注的 WSSS 在标注成本和分割准确率之间进行权衡，通过增加标注成本的方式来换取分割性能的提升。表 2 展示了这些 WSSS 方法在 PASCAL VOC 2012 上的性能表现。其中，骨干网表示用于图像分割的特征提取网络。

最小包围盒标注不仅提供了前景物体的类别信息，还提供了粗略的位置信息和边界信息，是一种相对较强的弱标注信号。研究表明，直接将最小包围盒内的所有像素视为前景，并以此训练一个语义分割模型，也能取得一定的分割效果（Papandreou 等，2015）。现有的研究主要关注如何挖掘最小包围盒内的前景和背景线索，从而准确区分最小包围盒内的前景和背景区域。这些研究取得了相当不错的进展，使得基于最小包围盒标注的 WSSS 获得了显著的性能提升，普遍优于基于图像级标注的 WSSS 方法。例如，Oh 等人（2021）提出的方法在 PASCAL VOC 2012 的测试集上获得了 76.1% 的 mIoU。然而，其与全监督语义分割方法相比，仍然存在一定的提升空间。近年来，基于最小包围盒标注的 WSSS 研究进展放缓，主要受限于以下原因：最小包围盒标注提供的信息量是有限的，现有模型对这些信息的利用率已经达到了瓶颈，要想进一步提升分割性能，需要引入海量新的标注数据，但标注成本难以接受。因此，现有最小包围盒标注的数据集难以为模型提供新的信息，这成为了性能提升的瓶颈。如何在有限标注资源下，提升模型的分割性能，是这类 WSSS 任务面临的难点。

线标注不仅指明了图像中所含物体的类别，还提供了目标物体的局部位置信息，但缺乏物体的边界信息。为了充分利用线标注的有限信息，现有的线标注 WSSS 方法主要使用不同的标签传播机制、

表2　基于其他弱标注的弱监督语义分割方法PASCAL VOC 2012上的性能表现

Table 2　Performance comparison of weakly supervised semantic segmentation methods based on other weak annotations on PASCAL VOC 2012 dataset

/%

标注形式	方法	发表	骨干网	mIoU	
				验证集	测试集
最小包围盒标注	Papandreou等人(2015)	ICCV2015	VGG16	60.6	62.2
	BoxSup(Dai等,2015)	ICCV2015	VGG16	62.0	64.6
	SDI(Khoreva等,2017)	CVPR2017	ResNet101	69.4	–
	BCM(Song等,2019)	CVPR2019	ResNet101	70.2	–
	BBAM(Lee等,2021b)	CVPR2021	ResNet101	73.7	73.7
	Oh等人(2021)	CVPR2021	ResNet101	74.6	76.1
	Kulharia等人(2020)	ECCV2020	ResNet101	76.4	–
线标注	RAWKS(Vernaza和Chandraker,2017)	CVPR2017	ResNet101	61.4	–
	Lin等人(2016)	CVPR2016	VGG16	63.1	–
	Tang等人(2018a)	CVPR2018	ResNet101	74.5	–
	Xu等人(2021)	ICCV2021	ResNet101	74.9	–
	Tang等人(2018b)	ECCV2018	ResNet101	75.0	–
	BPG(Wang等,2019)	IJCAI2019	ResNet101	76.0	–
	SPML(Ke等,2021)	arXiv2021	ResNet101	76.1	–
	DFR(Zhang等,2021a)	arXiv2021	Swin-T-Base	82.8	82.9
点标注	Bearman等人(2016)	ECCV2016	VGG16	46.1	–
	Papadopoulos等人(2017)	ICCV2017	VGG16	58.4	–
	DEXTR(Maninis等,2018)	CVPR2018	–	70.0	–
	PCAM(McEver和Manjunath,2020)	arXiv2020	ResNet50	70.5	–
	Li等人(2022a)	TPAMI2022	–	70.9	–

注:“–”表示论文未提及具体网络结构或实验结果。

构建特定的损失函数和引导模块等改进策略,将线标注信息传递到未标注区域,并引导模型完成物体边界信息的提取,从而生成更准确的像素级伪标签以指导图像分割,取得了显著的性能提升。例如,基于CNN骨干网的SPML(Ke等,2021)在PASCAL VOC 2012验证集上获得了76.1%的mIoU,基于Transformer骨干网的DFR(Zhang等,2021a)在测试集上获得了82.9%的mIoU。但是,目前基于线标注的WSSS任务仍面临一些难点,例如线标注的稀疏性对分割边界的干扰仍然存在,模型无法有效分割目标内部区域,尤其是在复杂场景或存在遮挡的情况下。这些问题使得其与全监督语义分割方法的性能差距仍然存在。

点标注指明了图像中某些特定点的类别信息和位置信息,是线标注的一种简化形式。点标注凭借极低的标注成本成为一种高效的标注方式,在针对细长形状物体的标注场景中尤为突出。相比于图像级标注,点标注包含部分像素点的位置信息,但基于点标注的WSSS方法性能略低于图像级标注。例如,PCAM(McEver和Manjunath,2020)在PASCAL VOC 2012验证集上仅获得70.5%的mIoU,低于当前图像级标注WSSS的最优性能。这主要是由于相关点标注数据集的缺乏,导致关于点标注的WSSS研究相对较少,未能充分利用点标注的有效信息。构建点标注的大规模数据集,并设计高效的点标注传播策略来缓解其稀疏性,是基于点标注WSSS任

务的一个潜在研究方向。

4 基于大模型的弱监督语义分割

近年来，大模型（Zhao等，2023）的研究日益火热，受到了研究者的广泛关注。在深度学习领域，大模型通常是指具有数十亿甚至更多参数的神经网络模型。尽管这类模型通常需要消耗大量的计算资源和存储空间进行训练和推理，也依赖包含海量样本的大规模数据集，但通常具有更强的表达能力和泛化能力。通过针对特定任务进行迁移学习（Pan和Yang，2010）和微调，大模型可以有效地助力各种下游任务。因此，一些工作开始尝试如何使用大模型辅助生成更准确的像素级伪标签。本节将回顾总结一些基于CLIP（contrastive language-image pre-training）（Radford等，2021）模型和SAM（segment anything model）（Kirillov等，2023）模型的WSSS方法。

CLIP模型是在大规模图像和文本数据上进行预训练得到的。通过学习图像和文本标签之间的对应关系，CLIP能够更好地利用文本标签中的语义信息，从而改善图像级WSSS的性能（Xie等，2022a）。Xie等人（2022a）提出CLIP跨语言图像匹配框架CLIMS（cross language image matching），旨在提高目标区域激活的完整性并抑制共现类别和背景的错误激活。为此他们设计了3个基于CLIP的损失函数：物体区域和文本标签匹配损失、背景区域和文本标签匹配损失以及共现背景抑制损失。Lin等人（2023）提出了CLIP-ES（contrastive language-image pre-training efficient segmenter）框架，利用CLIP的零样本识别能力来抑制非目标类和背景的错误激活，并改进了CLIP的文本输入形式以更好地适应WSSS任务。此外，Lin等人还提出了置信度引导损失迫使第2阶段的语义分割模型关注高置信度区域，减轻错误像素级伪标签的干扰。Xu等人（2023）提出了统一的Transformer网络架构，来学习目标物体的视觉token和文本token，并结合两种token进一步促进CAM的完整激活。其中，视觉token从图像中捕获语义信息，文本token则利用CLIP模型从标签文本中提取互补信息。

SAM是一个通用的图像分割大模型，在各类图像分割任务中表现出色。一些研究者将SAM引入WSSS任务中，以辅助生成高质量的像素级伪标签。具体而言，Chen等人（2023a）将CAM与SAM相结合，使用CAM的语义信息为SAM的分割结果分配特定的类别标签，从而生成准确的像素级伪标签。Sun等人（2023）将Grounding DINO（grounding distillation with no labels）（Liu等，2023）与SAM方法相结合，克服了SAM模型不支持类别标签作为监督信号的限制。

表3展示了基于大模型的WSSS方法在PASCAL VOC 2012验证集和测试集上的性能表现。得益于庞大参数量和海量训练数据，大模型可以对不同语义内容进行准确的建模和表达。当应用于监督信息不足的WSSS任务时，大模型可以利用其预训练得到的通用语义知识来理解图像内容，生成更准确的像素级伪标签，从而提升模型的分割性能。例如，Lin等（2023）的方法在PASCAL VOC 2012的测试集上取得了73.9%的mIoU，Sun等人（2023）的方法取得了77.1%的mIoU。这些性能表现已经逼近甚至超过了其他研究已久的WSSS方法，这充分证明了大模型的强大特征表达和建模能力。未来的

表3 基于大模型的弱监督语义分割方法PASCAL VOC 2012上的性能表现

Table 3 Performance comparison of weakly supervised semantic segmentation methods based on large-scale models on PASCAL VOC 2012 dataset

/%

大模型类别	方法	发表	骨干网	MIoU	
				验证集	测试集
CLIP	CLIMS（Xie等，2022a）	CVPR2022	ResNet101	70.4	70.0
	Xu等人（2023）	CVPR2023	ResNet38	72.2	72.2
	Lin等人（2023）	CVPR2023	ResNet101	73.8	73.9
SAM	Sun等人（2023）	arXiv2023	ResNet101	77.2	77.1

研究可以继续探索更有效的优化策略，或使用蒸馏技术减小模型尺寸，以进一步提升大模型在WSSS任务上的有效性和适用性。但是，基于大模型的WSSS研究也存在一些挑战：1）过大的模型参数量导致训练和调参变得非常困难，且十分耗费计算资源；2）蒸馏时如何保留模型的关键语义知识是一个难点。

5 总结与讨论

基于深度学习的WSSS是计算机视觉领域的一个重要研究方向，具有巨大的理论研究价值和广泛的应用前景。在现实世界中，像素级注释需要花费大量的人力、物力和时间成本，这一点在针对大规模数据集的标注场景中尤为突出。而WSSS通过不同形式的弱标注信号实现像素级的分割预测，极大程度地缓解了所需的像素级标注成本，为语义分割的实际应用提供了可行的解决方案。近年来涌现了大量基于深度学习的WSSS方法，模型的分割性能取得了显著的提升，然而与全监督方法相比，其仍存在一定的提升空间。此外，这些方法之间可能存在一定的相关性，即采用了相似的策略来缩小弱标注信息和像素级标注之间的差距。接下来，本文将分析WSSS领域存在的问题与挑战，并就其未来可能的研究方向提出建议。

5.1 存在的问题和挑战

基于深度学习的WSSS方法的关键在于缩小弱监督信息和像素级标注之间的差距。其根本策略在于通过给定的弱标注生成像素级伪标签，并基于该伪标签训练语义分割模型。虽然现有的研究通过不同的标签传播策略、引入自监督学习或额外监督信号等方法尽可能地提高像素级伪标签的精度，但其与真实的像素级标签仍存在一定的差距。如何提高生成的像素级伪标签的精度仍然是一个亟待解决的问题。

当WSSS模型应用于现实生活场景中，可能遇到在训练数据中从未出现的物体类别，这就需要模型具备一定的自适应能力来识别和分割未知的目标物体。因此，如何增强模型对开放性环境的自适应能力是一个极大的挑战。

另一方面，现实生活场景通常对实时性有较高的需求。但是现有的研究主要关注如何提高WSSS模型的分割精度，对模型推理速度的关注度不够。如何在保证一定精度的前提下，提高模型的运算速度，这是WSSS领域的另一大挑战。

此外，虽然Gao等人（2023）提出了非监督语义分割数据集，但相关数据集的匮乏仍然是WSSS领域的主要障碍之一。目前研究工作更多地关注算法改进，而较少考虑数据集构建。随着算法性能不断提升，数据集匮乏的问题也日益凸显。目前相关的WSSS方法主要在PASCAL VOC 2012（Everingham等，2015）和MS COCO 2014（Lin等，2014）等少量公开数据集上进行评估。这些公开数据集包含的类别数量有限，对稀有类别的覆盖也不足。这些问题导致现有WSSS模型训练难度的提升和性能的下降，并限制了模型在特定领域的应用。

5.2 未来的研究方向

为了缩小弱监督方法和全监督方法之间的差距，应继续研究如何提高生成的像素级伪标签的精度，探索更有效的像素级伪标签优化方法。基于大模型的WSSS方法是一个可行的研究方向。例如，近几年新出现的CLIP模型（Radford等，2021）和SAM（Kirillov等，2023）模型可以用于辅助像素级伪标签的优化。如第3节所述，已经有部分学者对此做出了初步探索，并取得了可观的成效。但是，如何设计适配于具体WSSS任务的特定网络结构和损失函数，实现大模型预训练知识的有效迁移，仍需要进一步探索。此外，由于像素级伪标签不可避免地存在一些噪声，因此基于像素级伪标签训练语义分割模型可以看做是一种噪声标签学习问题。通过设计特定的噪声标签检测方法来鉴别和过滤错误的像素级伪标签，或者构建鲁棒性分割模型以抵抗像素级伪标签中的噪声干扰也是未来值得研究的方向。

针对WSSS模型在开放环境中的自适应问题，可以通过迁移学习（Pan等，2010）或零样本学习（Xian等，2019）将从已知类别学习到的知识迁移到未知目标类别上。另一种可行的替代方案是引入多模态信息（Baltrušaitis等，2019），例如文本，将分割模型推广到未知的物体类别，增强模型的零样本分割能力。此外，增强模型的可解释性（Guidotti等，2019）也有利于模型对未知目标类别的处理。可解释性可以增强用户对模型行为的理解，从而对未知目标类别的判断进行监督，引导模型的优化。

当前的WSSS方法主要关注如何提高模型的分

割精度，导致模型结构和训练策略较为复杂，模型参数量大，很难进行实时的分割预测，不利于其在现实应用场景的部署。因此，研究者也可以关注如何开发实时的 WSSS 框架，在模型准确性和实时性之间进行权衡，即在尽可能保证模型准确性的情况下，加快模型的推理速度。

针对数据集匮乏的问题，构建高质量的、包含不同类型图像且具有更大挑战性的大规模 WSSS 数据集，是推动该领域发展的潜在研究方向。通过构建相关数据集，可以降低模型过拟合的风险，增强模型对不同类型图像的泛化能力。而针对特定应用场景构建定制化数据集，也将有利于模型更好地适配各种应用场景，推动该领域的发展。

参考文献（References）

Ahn J, Cho S and Kwak S. 2019. Weakly supervised learning of instance segmentation with inter-pixel relations//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 2204-2213 [DOI: 10.1109/cvpr.2019.00231]

Ahn J and Kwak S. 2018. Learning pixel-level semantic affinity with image-level supervision for weakly supervised semantic segmentation//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 4981-4990 [DOI: 10.1109/cvpr.2018.00523]

Araslanov N and Roth S. 2020. Single-stage semantic segmentation from image labels//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 4252-4261 [DOI: 10.1109/cvpr42600.2020.00431]

Arbeláez P, Pont-Tuset J, Barron J, Marques F and Malik J. 2014. Multiscale combinatorial grouping//Proceedings of 2014 IEEE Conference on Computer Vision and Pattern Recognition. Columbus, USA: IEEE: 328-335 [DOI: 10.1109/cvpr.2014.49]

Baltrušaitis T, Ahuja C and Morency L P. 2019. Multimodal machine learning: a survey and taxonomy. IEEE Transactions on Pattern Analysis and Machine Intelligence, 41(2): 423-443 [DOI: 10.1109/TPAMI.2018.2798607]

Bearman A, Russakovsky O, Ferrari V and Li F F. 2016. What's the point: semantic segmentation with point supervision//Proceedings of the 14th European Conference on Computer Vision. Amsterdam, the Netherlands: Springer: 549-565 [DOI: 10.1007/978-3-319-46478-7_34]

Chang Y T, Wang Q S, Hung W C, Piramuthu R, Tsai Y H and Yang M H. 2020. Weakly-supervised semantic segmentation via sub-category exploration//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 8988-8997 [DOI: 10.1109/cvpr42600.2020.00901]

Chen L Y, Wu W W, Fu C C, Han X and Zhang Y T. 2020. Weakly supervised semantic segmentation with boundary exploration//Proceedings of the 16th European Conference on Computer Vision. Glasgow, UK: Springer: 347-362 [DOI: 10.1007/978-3-030-58574-7_21]

Chen Q, Yang L X, Lai J H and Xie X H. 2022a. Self-supervised image-specific prototype exploration for weakly supervised semantic segmentation//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE: 4278-4288 [DOI: 10.1109/cvpr52688.2022.00425]

Chen T, Yao Y Z and Tang J H. 2023b. Multi-granularity denoising and bidirectional alignment for weakly supervised semantic segmentation. IEEE Transactions on Image Processing, 32: 2960-2971 [DOI: 10.1109/TIP.2023.3275913]

Chen T, Yao Y Z, Zhang L, Wang Q, Xie G S and Shen F M. 2023c. Saliency guided inter-and intra-class relation constraints for weakly supervised semantic segmentation. IEEE Transactions on Multimedia, 25: 1727-1737 [DOI: 10.1109/tmm.2022.3157481]

Chen T L, Mai Z D, Li R W and Chao W L. 2023a. Segment anything model (SAM) enhanced pseudo labels for weakly supervised semantic segmentation [EB/OL]. [2023-05-09]. https://arxiv.org/pdf/ 2305.05803.pdf

Chen Z, Tian Z Q, Zhu J H, Li C and Du S Y. 2022b. C-CAM: causal CAM for weakly supervised semantic segmentation on medical image//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE: 11666-11675 [DOI: 10.1109/cvpr52688.2022.01138]

Chen Z Z and Sun Q R. 2023. Extracting class activation maps from non-discriminative features as well//Proceedings of 2023 IEEE/ CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada: IEEE: 3135-3144 [DOI: 10.1109/CVPR52729.2023.00306]

Cordts M, Omran M, Ramos S, Rehfeld T, Enzweiler M, Benenson R, Franke U, Roth S and Schiele B. 2016. The cityscapes dataset for semantic urban scene understanding//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA: IEEE: 3213-3223 [DOI: 10.1109/cvpr.2016.350]

Dai J F, He K M and Sun J. 2015. BoxSup: exploiting bounding boxes to supervise convolutional networks for semantic segmentation//Proceedings of 2015 IEEE International Conference on Computer Vision. Santiago, Chile: IEEE: 1635-1643 [DOI: 10.1109/iccv.2015.191]

Dalal N and Triggs B. 2005. Histograms of oriented gradients for human detection//Proceedings of 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. San Diego, USA: IEEE: 886-893 [DOI: 10.1109/CVPR.2005.177]

Dempster A P, Laird N M and Rubin D B. 1977. Maximum likelihood

from incomplete data via the EM algorithm. Journal of the Royal Statistical Society: Series B (Methodological), 39(1): 1-22 [DOI: 10.1111/j.2517-6161.1977.tb01600.x]

Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X H, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S, Uszkoreit J and Houlsby N. 2021. An image is worth 16x16 words: transformers for image recognition at scale//Proceedings of the 9th International Conference on Learning Representations. [s. l.]: OpenReview.net

Du Y, Fu Z H, Liu Q J and Wang Y H. 2022. Weakly supervised semantic segmentation by pixel-to-prototype contrast//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE: 4310-4319 [DOI: 10.1109/CVPR52688.2022.00428]

Everingham M, Eslami S M A, Van Gool L, Williams C K I, Winn J and Zisserman A. 2015. The Pascal visual object classes challenge: a retrospective. International Journal of Computer Vision, 111(1): 98-136 [DOI: 10.1007/s11263-014-0733-5]

Gao S H, Li Z Y, Yang M H, Cheng M M, Han J W and Torr P. 2023. Large-scale unsupervised semantic segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 45(6): 7457-7476 [DOI: 10.1109/TPAMI.2022.3218275]

Grill J B, Strub F, Altché F, Tallec C, Richemond P H, Buchatskaya E, Doersch C, Pires B A, Guo Z D, Azar M G, Piot B, Kavukcuoglu K, Munos R and Valko M. 2020. Bootstrap your own latent a new approach to self-supervised learning//Proceedings of the 34th International Conference on Neural Information Processing Systems. Vancouver: Curran Associates Inc: 1786

Guidotti R, Monreale A, Ruggieri S, Turini F, Giannotti F and Pedreschi D. 2019. A survey of methods for explaining black box models. ACM Computing Surveys, 51(5): 1-42 [DOI: 10.1145/3236009]

He K M, Fan H Q, Wu Y X, Xie S N and Girshick R. 2020. Momentum contrast for unsupervised visual representation learning//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 9726-9735 [DOI: 10.1109/cvpr42600.2020.00975]

Hinton G E and Salakhutdinov R R. 2006. Reducing the dimensionality of data with neural networks. Science, 313 (5786): 504-507 [DOI: 10.1126/science.1127647]

Hou Q B, Jiang P T, Wei Y C and Chen M M. 2018. Self-erasing network for integral object attention//Proceedings of the 32nd International Conference on Neural Information Processing Systems. Montréal, Canada: Curran Associates Inc: 547-557

Huang Z L, Wang X G, Wang J S, Liu W Y and Wang J D. 2018. Weakly-supervised semantic segmentation network with deep seeded region growing//Proceedings of 2018 IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 7014-7023 [DOI: 10.1109/CVPR.2018.00733]

Jiang P T, Han L H, Hou Q B, Cheng M M and Wei Y C. 2022a. Online attention accumulation for weakly supervised semantic segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 44 (10): 7062-7077 [DOI: 10.1109/tpami.2021.3092573]

Jiang P T, Hou Q B, Cao Y, Cheng M M, Wei Y C and Xiong H K. 2019. Integral object mining via online attention accumulation//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision. Seoul, Korea (South): IEEE: 2070-2079 [DOI: 10.1109/iccv.2019.00216]

Jiang P T, Yang Y Q, Hou Q B and Wei Y C. 2022b. L2G: a simple local-to-global knowledge transfer framework for weakly supervised semantic segmentation//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE: 16865-16875 [DOI: 10.1109/cvpr52688.2022.01638]

Jing L L and Tian Y L. 2021. Self-supervised visual feature learning with deep neural networks: a survey. IEEE Transactions on Pattern Analysis and Machine Intelligence, 43(11): 4037-4058 [DOI: 10.1109/TPAMI.2020.2992393]

Jo S and Yu I J. 2021. Puzzle-CAM: improved localization via matching partial and full features//Proceedings of 2021 IEEE International Conference on Image Processing. Anchorage, USA: IEEE: 639-643 [DOI: 10.1109/icip42928.2021.9506058]

Joon Oh S, Benenson R, Khoreva A, Akata Z, Fritz M and Schiele B. 2017. Exploiting saliency for object segmentation from image level labels//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA: IEEE: 5038-5047 [DOI: 10.1109/cvpr.2017.535]

Ke T W, Hwang J J and Yu S X. 2021. Universal weakly supervised segmentation by pixel-to-segment contrastive learning//Proceedings of the 9th International Conference on Learning Representations. [s. l.]: OpenReview.net

Khoreva A, Benenson R, Hosang J, Hein M and Schiele B. 2017. Simple does it: weakly supervised instance and semantic segmentation//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA: IEEE: 1665-1674 [DOI: 10.1109/cvpr.2017.181]

Kim D, Cho D, Yoo D and Kweon I S. 2017. Two-phase learning for weakly supervised object localization//Proceedings of 2017 IEEE International Conference on Computer Vision. Venice, Italy: IEEE: 3554-3564 [DOI: 10.1109/iccv.2017.382]

Kirillov A, Mintun E, Ravi N, Mao H Z, Rolland C, Gustafson L, Xiao T T, Whitehead S, Berg A C, Lo W Y, Dollár P and Girshick R. 2023. Segment anything [EB/OL]. [2023-04-05]. https://arxiv.org/pdf/2304.02643.pdf

Kolesnikov A and Lampert C H. 2016. Seed, expand and constrain: three principles for weakly-supervised image segmentation//Proceedings of the 14th European Conference on Computer Vision. Amsterdam, the Netherlands: Springer: 695-711 [DOI: 10.1007/978-3-319-46493-0_42]

Krähenbühl P and Koltun V. 2011. Efficient inference in fully connected CRFs with Gaussian edge potentials//Proceedings of the 24th International Conference on Neural Information Processing Systems. Granada, Spain: Curran Associates Inc: 109-117

Kulharia V, Chandra S, Agrawal A, Torr P and Tyagi A. 2020. Box2Seg: Attention weighted loss and discriminative feature learning for weakly supervised segmentation//Proceedings of the 16th European Conference on Computer Vision. Online: Springer: 290-308 [DOI: 10.1007/978-3-030-58583-9_18]

Kweon H, Yoon S H, Kim H, Park D and Yoon K J. 2021. Unlocking the potential of ordinary classifier: class-specific adversarial erasing framework for weakly supervised semantic segmentation//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision. Montreal, Canada: IEEE: 6974-6983 [DOI: 10.1109/iccv48922.2021.00691]

Kweon H, Yoon S H and Yoon K J. 2023. Weakly supervised semantic segmentation via adversarial learning of classifier and reconstructor//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada: IEEE: 11329-11339 [DOI: 10.1109/CVPR52729.2023.01090]

Lafferty J D, McCallum A and Pereira F C N. 2001. Conditional random fields: probabilistic models for segmenting and labeling sequence data//Proceedings of the 18th International Conference on Machine Learning. Williams College, USA: Morgan Kaufmann Publishers Inc

Lee J, Kim E, Lee S, Lee J and Yoon S. 2019. FickleNet: weakly and semi-supervised semantic image segmentation using stochastic inference//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Angeles, USA: IEEE: 5262-5271 [DOI: 10.1109/CVPR.2019.00541]

Lee J, Kim E and Yoon S. 2021a. Anti-adversarially manipulated attributions for weakly and semi-supervised semantic segmentation//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 4070-4078 [DOI: 10.1109/cvpr46437.2021.00406]

Lee J, Yi J, Shin C and Yoon S. 2021b. BBAM: bounding box attribution map for weakly supervised semantic and instance segmentation//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 2643-2651 [DOI: 10.1109/cvpr46437.2021.00267]

Lee S, Lee M, Lee J and Shim H. 2021c. Railroad is not a train: Saliency as pseudo-pixel supervision for weakly supervised semantic segmentation//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 5491-5501 [DOI: 10.1109/cvpr46437.2021.00545]

Li J, Fan J S and Zhang Z X. 2022a. Towards noiseless object contours for weakly supervised semantic segmentation//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE: 16835-16844 [DOI: 10.1109/cvpr52688.2022.01635]

Li X Y, Zhou T F, Li J W, Zhou Y and Zhang Z X. 2021. Group-wise semantic mining for weakly supervised semantic segmentation//Proceedings of the 35th AAAI Conference on Artificial Intelligence. [s. l.]: AAAI: 1984-1992 [DOI: 10.1609/aaai.v35i3.16294]

Li Y W, Zhao H S, Qi X J, Chen Y K, Qi L, Wang L W, Li Z M, Sun J and Jia J Y. 2022b. Fully convolutional networks for panoptic segmentation with point-based supervision. IEEE Transactions on Pattern Analysis and Machine Intelligence, 45(4): 4552-4568 [DOI: 10.1109/tpami.2022.3200416]

Lin D, Dai J F, Jia J Y, He K M and Sun J. 2016. ScribbleSup: scribble-supervised convolutional networks for semantic segmentation//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA: IEEE: 3159-3167 [DOI: 10.1109/cvpr.2016.344]

Lin T Y, Maire M, Belongie S, Hays J, Perona P, Ramanan D, Dollár P and Zitnick C L. 2014. Microsoft COCO: common objects in context//Proceedings of the 13th European Conference on Computer Vision. Zurich, Switzerland: Springer: 740-755 [DOI: 10.1007/978-3-319-10602-1_48]

Lin Y Q, Chen M H, Wang W X, Wu B X, Li K, Lin B B, Liu H F and He X F. 2023. CLIP is also an efficient segmenter: a text-driven approach for weakly supervised semantic segmentation//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada: IEEE: 15305-15314 [DOI: 10.1109/CVPR52729.2023.01469]

Liu S L, Zeng Z Y, Ren T H, Li F, Zhang H, Yang J, Li C Y, Yang J W, Su H, Zhu J and Zhang L. 2023. Grounding DINO: marrying DINO with grounded pre-training for open-set object detection [EB/OL]. [2023-03-20]. https://arxiv.org/pdf/2303.05499.pdf

Long J, Shelhamer E and Darrell T. 2015. Fully convolutional networks for semantic segmentation//Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston, USA: IEEE: 3431-3440 [DOI: 10.1109/CVPR.2015.7298965]

Lowe D G. 2004. Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision, 60(2): 91-110 [DOI: 10.1023/B:VISI.0000029664.99615.94]

MacQueen J. 1967. Some methods for classification and analysis of multivariate observations//Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability. Oakland, USA: University of California Press: 281-297

Maninis K K, Caelles S, Pont-Tuset J and van Gool L. 2018. Deep extreme cut: from extreme points to object segmentation//Proceedings of 2018 IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 616-625 [DOI: 10.1109/cvpr.2018.00071]

McEver R A and Manjunath B S. 2020. PCAMs: weakly supervised semantic segmentation using point supervision [EB/OL]. [2020-07-10]. https://arxiv.org/pdf/2007.05615.pdf

Minaee S, Boykov Y Y, Porikli F, Plaza A J, Kehtarnavaz N and Terzopoulos D. 2022. Image segmentation using deep learning: a survey. IEEE Transactions on Pattern Analysis and Machine Intelligence, 44(7): 3523-3542 [DOI: 10.1109/TPAMI.2021.3059968]

Mottaghi R, Chen X J, Liu X B, Cho N G, Lee S W, Fidler S, Urtasun R and Yuille A. 2014. The role of context for object detection and semantic segmentation in the wild//Proceedings of 2014 IEEE Conference on Computer Vision and Pattern Recognition. Columbus, USA: IEEE: 891-898 [DOI: 10.1109/cvpr.2014.119]

Neuhold G, Ollmann T, Rota Bulo S and Kontschieder P. 2017. The Mapillary vistas dataset for semantic understanding of street scenes//Proceedings of 2017 IEEE International Conference on Computer Vision. Venice, Italy: IEEE: 4990-4999 [DOI: 10.1109/iccv.2017.534]

Oh Y, Kim B and Ham B. 2021. Background-aware pooling and noise-aware loss for weakly-supervised semantic segmentation//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 6909-6918 [DOI: 10.1109/cvpr46437.2021.00684]

Ojala T, Pietikainen M and Harwood D. 1994. Performance evaluation of texture measures with classification based on Kullback discrimination of distributions//Proceedings of the 12th International Conference on Pattern Recognition. Jerusalem, Israel: IEEE: 582-585 [DOI: 10.1109/ICPR.1994.576366]

Pan S J and Yang Q. 2010. A survey on transfer learning. IEEE Transactions on Knowledge and Data Engineering, 22(10): 1345-1359 [DOI: 10.1109/TKDE.2009.191]

Papadopoulos D P, Uijlings J R R, Keller F and Ferrari V. 2017. Extreme clicking for efficient object annotation//Proceedings of 2017 IEEE International Conference on Computer Vision. Venice, Italy: IEEE: 4940-4949 [DOI: 10.1109/iccv.2017.528]

Papandreou G, Chen L C, Murphy K P and Yuille A L. 2015. Weakly- and semi-supervised learning of a deep convolutional network for semantic image segmentation//Proceedings of 2015 IEEE International Conference on Computer Vision. Santiago, Chile: IEEE: 1742-1750 [DOI: 10.1109/iccv.2015.203]

Peng Z L, Wang G C, Xie L X, Jiang D S, Shen W and Tian Q. 2023. USAGE: a unified seed area generation paradigm for weakly supervised semantic segmentation//Proceedings of 2023 IEEE/CVF International Conference on Computer Vision. Paris, France: IEEE [DOI: 10.1109/ICCV51070.2023.00064]

Qian R, Wei Y C, Shi H H, Li J C, Liu J Y and Huang T. 2019. Weakly supervised scene parsing with point-based distance metric learning//Proceedings of the 33rd AAAI Conference on Artificial Intelligence. Honolulu, USA: AAAI: 8843-8850 [DOI: 10.1609/aaai.v33i01.33018843]

Qing C, Yu J, Xiao C B and Duan J. 2020. Deep convolutional neural network for semantic image segmentation. Journal of Image and Graphics, 25(6): 1069-1090 (青晨, 禹晶, 肖创柏, 段娟. 2020. 深度卷积神经网络图像语义分割研究进展. 中国图象图形学报, 25(6): 1069-1090) [DOI: 10.11834/jig.190355]

Radford A, Kim J W, Hallacy C, Ramesh A, Goh G, Agarwal S, Sastry G, Askell A, Mishkin P, Clark J, Krueger G and Sutskever I. 2021. Learning transferable visual models from natural language supervision//Proceedings of the 38th International Conference on Machine Learning. [s.l.]: ACM: 8748-8763

Ren D W, Wang Q L, Wei Y C, Meng D Y and Zuo W M. 2022. Progress in weakly supervised learning for visual understanding. Journal of Image and Graphics, 27(6): 1768-1798 (任冬伟, 王旗龙, 魏云超, 孟德宇, 左旺孟. 2022. 视觉弱监督学习研究进展. 中国图象图形学报, 27(6): 1768-1798) [DOI: 10.11834/jig.220178]

Rong S H, Tu B H, Wang Z L and Li J J. 2023. Boundary-enhanced Co-training for weakly supervised semantic segmentation//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada: IEEE: 19574-19584 [DOI: 10.1109/CVPR52729.2023.01875]

Rother C, Kolmogorov V and Blake A. 2004. "GrabCut": interactive foreground extraction using iterated graph cuts. ACM Transactions on Graphics, 23(3): 309-314 [DOI: 10.1145/1015706.1015720]

Ru L X, Zhan Y B, Yu B S and Du B. 2022. Learning affinity from attention: end-to-end weakly-supervised semantic segmentation with transformers//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE: 16825-16834 [DOI: 10.1109/CVPR52688.2022.01634]

Ru L X, Zheng H L, Zhan Y B and Du B. 2023. Token contrast for weakly-supervised semantic segmentation//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada: IEEE: 3093-3102 [DOI: 10.1109/CVPR52729.2023.00302]

Scarselli F, Gori M, Tsoi A C, Hagenbuchner M and Monfardini G. 2009. The graph neural network model. IEEE Transactions on Neural Networks, 20(1): 61-80 [DOI: 10.1109/TNN.2008.2005605]

Shen W, Peng Z L, Wang X H, Wang H Y, Cen J Z, Jiang D S, Xie L X, Yang X K and Tian Q. 2023. A survey on label-efficient deep image segmentation: Bridging the gap between weak supervision and dense prediction. IEEE Transactions on Pattern Analysis and Machine Intelligence, 45(8): 9284-9305 [DOI: 10.1109/TPAMI.2023.3246102]

Song C F, Huang Y, Ouyang W L and Wang L. 2019. Box-driven class-wise region masking and filling rate guided loss for weakly supervised semantic segmentation//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Angeles, USA: IEEE: 3136-3145 [DOI: 10.1109/cvpr.2019.00325]

Su Y K, Sun R Z, Lin G S and Wu Q Y. 2021. Context decoupling augmentation for weakly supervised semantic segmentation//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision. Montreal, Canada: IEEE: 6984-6994 [DOI: 10.1109/iccv48922.2021.00692]

Sun G L, Wang W G, Dai J F and van Gool L. 2020. Mining cross-image semantics for weakly supervised semantic segmentation//Proceedings of the 16th European Conference on Computer Vision. Glasgow, UK: Springer: 347-365 [DOI: 10.1007/978-3-030-58536-5_21]

Sun K Y, Shi H Q, Zhang Z M and Huang Y M. 2021. ECS-Net: improving weakly supervised semantic segmentation by using connections between class activation maps//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision. Montreal, Canada: IEEE: 7263-7272 [DOI: 10.1109/iccv48922.2021.00719]

Sun W X, Liu Z Y, Zhang Y H, Zhong Y R and Barnes N. 2023. An alternative to WSSS? An empirical study of the segment anything model (SAM) on weakly-supervised semantic segmentation problems [EB/OL]. [2023-06-18]. https://arxiv.org/pdf/2305.01586.pdf

Tang M, Djelouah A, Perazzi F, Boykov Y and Schroers C. 2018a. Normalized cut loss for weakly-supervised CNN segmentation//Proceedings of 2018 IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 1818-1827 [DOI: 10.1109/cvpr.2018.00195]

Tang M, Perazzi F, Djelouah A, Ayed I B, Schroers C and Boykov Y. 2018b. On regularized losses for weakly-supervised CNN segmentation//Proceedings of the 15th European Conference on Computer Vision (ECCV). Munich, Germany: Springer: 524-540 [DOI: 10.1007/978-3-030-01270-0_31]

Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A N, Kaiser Ł and Polosukhin L. 2017. Attention is all you need//Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach, USA: Curran Associates Inc: 6000-6010

Vernaza P and Chandraker M. 2017. Learning random-walk label propagation for weakly-supervised semantic segmentation//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA: IEEE: 2953-2961 [DOI: 10.1109/cvpr.2017.315]

Wang B, Qi G J, Tang S, Zhang T Z, Wei Y C, Li L H and Zhang Y D. 2019. Boundary perception guidance: a scribble-supervised semantic segmentation approach//Proceedings of the 28th IJCAI International Joint Conference on Artificial Intelligence. Macao, China: Morgan Kaufmann: 3663-3669 [DOI: 10.24963/ijcai.2019/508]

Wang Y D, Zhang J, Kan M N, Shan S G and Chen X L. 2020. Self-supervised equivariant attention mechanism for weakly supervised semantic segmentation//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 12272-12281 [DOI: 10.1109/cvpr42600.2020.01229]

Wei Y C, Feng J S, Liang X D, Cheng M M, Zhao Y and Yan S C. 2017. Object region mining with adversarial erasing: a simple classification to semantic segmentation approach//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA: IEEE: 6488-6496 [DOI: 10.1109/cvpr.2017.687]

Wei Y C, Xiao H X, Shi H H, Jie Z Q, Feng J S and Huang T S. 2018. Revisiting dilated convolution: a simple approach for weakly-and semi-supervised semantic segmentation//Proceedings of 2018 IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 7268-7277 [DOI: 10.1109/CVPR.2018.00759]

Weinberger K Q and Saul L K. 2009. Distance metric learning for large margin nearest neighbor classification. The Journal of Machine Learning Research, 10: 207-244

Wu T, Huang J S, Gao G Y, Wei X M, Wei X L, Luo X and Liu C H. 2021. Embedded discriminative attention mechanism for weakly supervised semantic segmentation//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 16760-16769 [DOI: 10.1109/cvpr46437.2021.01649]

Xian Y Q, Lampert C H, Schiele B and Akata Z. 2019. Zero-shot learning-a comprehensive evaluation of the good, the bad and the ugly. IEEE Transactions on Pattern Analysis and Machine Intelligence, 41(9): 2251-2265 [DOI: 10.1109/TPAMI.2018.2857768]

Xie J H, Hou X X, Ye K and Shen L L. 2022a. CLIMS: cross language image matching for weakly supervised semantic segmentation//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE: 4473-4482 [DOI: 10.1109/cvpr52688.2022.00444]

Xie J H, Xiang J F, Chen J L, Hou X X, Zhao X D, Shen L L. 2022b. C2AM: contrastive learning of class-agnostic activation map for weakly supervised object localization and semantic segmentation//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE: 989-998 [DOI: 10.1109/cvpr52688.2022.00106]

Xu J S, Zhou C W, Cui Z, Xu C Y, Huang Y G, Shen P C, Li S X and Yang J. 2021. Scribble-supervised semantic segmentation inference//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision. Montreal, Canada: IEEE: 15334-15343 [DOI: 10.1109/iccv48922.2021.01507]

Xu L, Ouyang W L, Bennamoun M, Boussaid F and Xu D. 2022. Multi-class token transformer for weakly supervised semantic segmentation//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE: 4300-4309 [DOI: 10.1109/cvpr52688.2022.00427]

Xu L, Ouyang W L, Bennamoun M, Boussaid F and Xu D. 2023. Learning multi-modal class-specific tokens for weakly supervised dense object localization//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada: IEEE: 19596-19605 [DOI: 10.1109/CVPR52729.2023.01877]

Yu Z, Zhuge Y Z, Lu H C and Zhang L H. 2019. Joint learning of saliency detection and weakly supervised semantic segmentation//Proceedings of 2019 IEEE/CVF International Conference on Com-

puter Vision. Seoul, Korea (South): IEEE: 7223-7233 [DOI: 10.1109/ICCV.2019.00732]

Zhang B F, Xiao J M, Wei Y C, Sun M J and Huang K Z. 2020a. Reliability does matter: an end-to-end weakly supervised semantic segmentation approach//Proceedings of the 34th AAAI Conference on Artificial Intelligence. New York, USA: AAAI: 12765-12772 [DOI: 10.1609/aaai.v34i07.6971]

Zhang B F, Xiao J M and Zhao Y. 2021a. Dynamic feature regularized loss for weakly supervised semantic segmentation [EB/OL]. [2022-03-05]. https://arxiv.org/pdf/2108.01296.pdf

Zhang D, Zhang H W, Tang J H, Hua X S and Sun Q R. 2020b. Causal intervention for weakly-supervised semantic segmentation//Proceedings of the 34th International Conference on Neural Information Processing Systems. Vancouver, Canada: Curran Associates Inc: #56

Zhang F, Gu C C, Zhang C Y and Dai Y C. 2021b. Complementary patch for weakly supervised semantic segmentation//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision. Montreal, Canada: IEEE: 7222-7231 [DOI: 10.1109/iccv48922.2021.00715]

Zhang T Y, Lin G S, Liu W D, Cai J F and Kot A. 2020c. Splitting vs. merging: mining object regions with discrepancy and intersection loss for weakly supervised semantic segmentation//Proceedings of the 16th European Conference on Computer Vision. Glasgow, UK: Springer: 663-679 [DOI: 10.1007/978-3-030-58542-6_40]

Zhang X R, Peng Z L, Zhu P, Zhang T Y, Li C, Zhou H Y and Jiao L C. 2021c. Adaptive affinity loss and erroneous pseudo-label refinement for weakly supervised semantic segmentation//Proceedings of the 29th ACM International Conference on Multimedia. Chengdu, China: ACM: 5463-5472 [DOI: 10.1145/3474085.3475675]

Zhao W X, Zhou K, Li J Y, Tang T Y, Wang X L, Hou Y P, Min Y Q, Zhang B C, Zhang J J, Dong Z C, Du Y F, Yang C, Chen Y S, Chen Z P, Jiang J H, Ren R Y, Li Y F, Tang X Y, Liu Z K, Liu P Y, Nie J Y and Wen R J. 2023. A survey of large language models [EB/OL]. [2023-06-29]. https://arxiv.org/pdf/2303.18223.pdf

Zhou B L, Khosla A, Lapedriza A, Oliva A and Torralba A. 2016. Learning deep features for discriminative localization//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA: IEEE: 2921-2929 [DOI: 10.1109/CVPR.2016.319]

Zhou T F, Zhang M J, Zhao F and Li J W. 2022. Regional semantic contrast and aggregation for weakly supervised semantic segmentation//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE: 4289-4299 [DOI: 10.1109/cvpr52688.2022.00426]

作者简介

项伟康，男，硕士研究生，主要研究方向为弱监督语义分割。E-mail：1022010310@njupt.edu.cn

周全，通信作者，男，副教授，主要研究方向为计算机视觉、深度学习和模式识别。E-mail：quan.zhou@njupt.edu.cn

莫智懿，男，教授，主要研究方向为机器学习与模式识别、机器视觉和智能视频及监控。E-mail：mofly_0214@qq.com

崔景程，男，本科生，主要研究方向为计算机视觉。E-mail：b22011826@njupt.edu.cn

吴晓富，男，研究员，主要研究方向为机器学习、计算机视觉和统计信号处理。E-mail：xfuwu@njupt.edu.cn

欧卫华，男，教授，主要研究方向为计算机视觉、图神经网络、跨媒体检索与推荐。E-mail：ouweihuahust@gmail.com

王井东，男，计算机视觉首席架构师，主要研究方向为计算机视觉、多媒体以及机器学习。E-mail：welleast@outlook.com

刘文予，男，教授，主要研究方向为人工智能、机器学习和计算机视觉。E-mail：liuwy@hust.edu.cn