中图法分类号:TP181;TP37　　**文献标识码**: A　　**文章编号**: 1006-8961(2025)12-3855-15

论文引用格式:Bai X F, Wang Y H, Xu W J, Jiang G X and Wang W J. 2025. Dual-stage guided weakly supervised semantic segmentation with Gaussian correction. Journal of Image and Graphics, 30(12):3855-3869(白雪飞, 王渊辉, 许文杰, 姜高霞, 王文剑. 2025. 融合高斯修正的双阶段指导弱监督语义分割. 中国图象图形学报, 30(12):3855-3869)[DOI:10.11834/jig.250040]

融合高斯修正的双阶段指导弱监督语义分割

白雪飞[1],王渊辉[1],许文杰[1],姜高霞[1],王文剑[2*]

1. 山西大学计算机与信息技术学院,太原　030006; 2. 计算智能与中文信息处理教育部重点实验室(山西大学),太原　030006

摘　要:**目的**　端到端的弱监督语义分割模型因其高效的训练效率备受关注,然而现有研究还存在语义信息提取不充分、生成的伪标签质量较低等不足。针对上述问题,本文提出一种基于知识蒸馏的端到端弱监督语义分割框架,通过双阶段知识交互模块增强学生网络和教师网络之间的知识传递,同时借助高斯修正模块对伪标签进行修正。**方法**　首先,设计双阶段知识交互模块强化教师网络和学生网络的特征学习过程,有效降低训练过程中的噪声干扰。其次,为了生成高质量的伪标签,设计了高斯修正模块,通过拟合类激活图的分布,利用EM(expectation maximization)算法估算每个像素点的噪声概率,并依据与邻域像素的相似度关系修正伪标签,进而提升弱监督语义分割网络的性能。**结果**　本文方法在PASCAL VOC 2012(pattern analysis, statical modeling and computational learning visual object classes 2012)和MS COCO 2014(Microsoft common objects in context 2014)数据集上的mIoU(mean intersection over union)值分别达到74.8%和42.3%,优于其他对比方法。**结论**　通过双阶段知识交互模块以及高斯修正模块,有效降低了图像内部噪声以及潜在的标签噪声对训练过程的影响,并且改善了伪标签生成不完整的问题,与现有方法相比取得了显著的性能提升,在端到端的弱监督语义分割方法中展现出明显的优越性,具有一定的研究价值。

关键词:深度学习;端到端弱监督语义分割;高斯混合模型(GMM);知识蒸馏;类激活图(CAM)

Dual-stage guided weakly supervised semantic segmentation with Gaussian correction

Bai Xuefei[1], Wang Yuanhui[1], Xu Wenjie[1], Jiang Gaoxia[1], Wang Wenjian[2*]

1. School of Computer and Information Technology, Shanxi University, Taiyuan　030006, China; 2. Key Laboratory of Computational Intelligence and Chinese Information Processing (Shanxi University), Ministry of Education, Taiyuan　030006, China

Abstract: **Objective** Weakly supervised semantic segmentation (WSSS) aims to reduce the cost associated with annotating "strong" pixel-level labels by using "weak" labels, such as points, bounding boxes, image-level class labels, and scribbles. Among these, image-level class labels are the most cost-effective and readily available; however, leveraging them for precise segmentation remains a considerable challenge. A widely used WSSS approach based on image-level class labels generally comprises the following steps: 1) training a neural network for image classification using the class labels; 2) using the trained network to generate class activation maps (CAMs), which serve as seed regions for the segmentation

收稿日期:2025-02-17;**修回日期**:2025-05-27;**预印本日期**:2025-06-03

*通信作者:王文剑　wjwang@sxu.edu.cn

基金项目:国家自然科学基金项目(62476157, U21A20513, 62276161, 62576198, 62576201);山西省重点研发计划资助(2022020201010);太行山西省实验室科技攻关专项项目(THYF-JSZX-24010200)

Supported by: National Natural Science Foundation of China(62476157, U21A20513, 62276161, 62576198, 62576201); Key R&D Program of Shanxi Province, China(2022020201010); Key Technologies Program of Taihang Laboratory in Shanxi Province(THYF-JSZX-24010200)

task; and 3) refining these CAMs into pseudo-labels, which are then used as the ground truth to supervise a segmentation network. These steps can be integrated into a single collaborative stage; typically, single-stage frameworks are highly efficient due to their simplified training pipeline. However, the quality of pseudo-labels is crucial to the overall performance of semantic segmentation. High-quality pseudo-labels result in superior segmentation outcomes, whereas noisy or inaccurate pseudo-labels hinder the capability of the model to learn meaningful features. WSSS based on image-level labels faces considerable challenges due to the absence of precise positional and shape-related information, making it difficult to generate accurate segmentation maps. These challenges have led to the development of various approaches, which can be broadly categorized into two types: single-stage methods and multistage methods. Although single-stage methods offer greater efficiency and simplify the overall training process, they often produce less accurate pseudo-labels. This condition is due to the limited refinement of CAMs, resulting in imprecise supervision signals that ultimately degrade segmentation performance. Aiming to alleviate these limitations, a simple yet novel single-stage WSSS framework that incorporates knowledge distillation is introduced to enhance pseudo-label quality without relying on any additional external supervision. The framework enhances the feature learning process within the teacher-student network using a dual-stage knowledge distillation module. This module allows the student network to acquire more dynamic and informative knowledge from the teacher network while preserving key features, thereby enhancing the overall robustness of the student model. Moreover, to further improve segmentation accuracy, a pseudo-label correction module based on a Gaussian mixture model (GMM) is introduced. This module refines the pseudo-labels by modeling the distribution of the CAMs, resulting in highly accurate and reliable supervision signals. The combination of dual-stage knowledge distillation and the Gaussian correction module ensures accurate learning and improved segmentation results, even under weak supervision signals such as image-level labels. Ultimately, the proposed method effectively mitigates the impact of noise during training and enhances the accuracy of the generated pseudo-labels, resulting in superior semantic segmentation outcomes in WSSS tasks. **Method** A novel weakly-supervised semantic segmentation method, aimed at addressing the challenges posed by noisy data points and weak supervision, is proposed. First, a dual-stage knowledge interaction module is introduced to enhance the feature learning process of the teacher and student networks. By enabling highly effective knowledge exchange between the two networks, the proposed approach notably reduces the impact of noise during training, leading to robust feature extraction. Additionally, a Gaussian correction module is proposed to enhance the quality of pseudo-labels. This module refines the pseudo-labels by modeling the distribution of class activation maps. By fitting the distribution more accurately, the module corrects potential errors in the pseudo-labels, ensuring that the model learns from high-quality, refined labels. Therefore, the method boosts the overall performance of weakly-supervised semantic segmentation, making it more robust to noise and improving segmentation accuracy. This method provides a promising solution for weakly-supervised segmentation tasks. **Result** The mIoU values of this method on the PASCAL VOC 2012 and MS COCO 2014 datasets were 74.8% and 42.3%, respectively, surpassing other comparative methods. Specifically, on the PASCAL VOC 2012 dataset, the proposed method achieved a 3.7% improvement over ToCo, an 8.8% enhancement compared to AFA, a 7.5% increase relative to TSCD, and 1.1% compared to BECO. On the MS COCO 2014 dataset, the method improved performance by 2.2% compared to TSCD, 3.4% compared to AFA, and 5.3% compared to AuxSegNet+. Additionally, the mIoU values of different categories are compared on the PASCAL VOC 2012 validation set. The experimental results showed that the method outperformed the competing methods in 16 categories. Notably, for the background class, the method achieved an mIoU of 92.4%, the highest among all methods evaluated. This result indicates that the method effectively leverages the Gaussian correction module to reduce misclassification of background regions, thereby improving segmentation performance. Furthermore, the method achieved notable improvements in categories such as bird, bottle, car, chair, and cow, further demonstrating its effectiveness. **Conclusion** The proposed method effectively mitigates the impact of noise during training and address the issue of incomplete pseudo-label generation through the integration of a dual-stage knowledge distillation module and a Gaussian correction module. This approach achieves remarkable performance improvements compared to existing methods. Overall, the results demonstrate notable advantages in end-to-end weakly supervised semantic segmentation and holds considerable research value.

Key words: deep learning; end-to-end weakly supervised semantic segmentation; Gaussian mixture model (GMM);

knowledge distillation; class activation map (CAM)

0 引 言

语义分割作为计算机视觉领域的一项基础任务,与传统图像分割相比,具有更高的复杂性和应用价值。语义分割不仅需要将图像划分为多个区域,还需要对图像中的每个像素进行分类,以实现对图像内容的精确分析。随着深度学习技术的持续进步,语义分割取得显著发展,并在自动驾驶(Zhou等,2022b)、医学图像处理(Du等,2011)以及遥感图像探测(张文凯 等,2022)等领域得到广泛应用。特别是在全监督语义分割领域,卷积神经网络展现出卓越的预测性能。然而,这些方法往往依赖于精确标注的数据,这种数据标注不仅成本高昂且耗时费力,限制了其大规模应用。为降低分割模型对数据标注的依赖,弱监督语义分割(weakly supervised semantic segmentation,WSSS)成为研究热点,通过利用点标签(Bearman 等,2016)、边界框(Chang 等,2020)、图像级标签(Kolesnikov 和 Lampert,2016)或涂鸦(Lin等,2016)等低成本标注实现语义分割模型的训练。其中,基于图像级标签的方法因其标注成本低,吸引了众多研究者的关注,逐渐成为计算机视觉领域的一个研究热点。

图像级标签仅提供图像中物体类别的存在信息,缺乏具体的位置信息和形状信息,因此,基于图像级标签的弱监督语义分割面临巨大挑战。目前的研究可大致分为两种主要方法:两阶段方法和端到端方法。

两阶段弱监督语义分割方法首先利用图像级标签训练一个分类网络,生成类激活图(class activation map,CAM)(Zhou等,2016),以获取目标物体的粗略位置信息。随后通过后处理技术生成伪标签,并用于语义分割模型的训练。此方法的关键在于如何生成准确完整的CAM以及如何有效处理伪标签中的噪声,这两个因素直接影响后续分割模型的训练效果和准确性。

针对CAM生成不完整的问题,一些研究人员提出擦除的方法,通过迭代擦除最具判别力的区域提高CAM的质量(Wei 等,2017;Singh 和 Lee,2017)。尽管这种方法在一定程度上可以改善CAM的完整性,但其存在一些不足:一方面,迭代过程不仅计算耗时,而且难以确定最优迭代次数;另一方面,网络在后续迭代中可能会错误地激活背景区域,导致CAM质量下降。相比之下,区域生长方法成为另一种有效的类激活图扩展策略。Kolesnikov 和 Lampert(2016)提出一种经典的SEC(seed,expand and constrain)模型,通过种子区域的动态扩展优化CAM质量。然而,SEC模型存在一个固有的局限性,即其对初始种子区域的选择是静态的,一旦设定,这些种子区域在整个分割过程中保持不变,缺乏动态调整的能力。为此,Huang等人(2018)对SEC进行改进,提出DSRG(deep seeded region growing)方法,通过动态更新种子区域调整边界,使得生成的CAM覆盖范围更广且定位更准确。针对伪标签中的噪声问题,Wang等人(2018)提出结合显著性引导的边界细化方法,Yao等人(2021)通过显著图的形状先验优化伪标签质量。尽管这些方法在一定程度上提升了伪标签的质量,但由于显著图无法提供物体类别信息,难以用于网络的训练阶段,对于网络训练的帮助比较有限。

近年来,Transformer因其自注意力机制具备全局特征建模能力,逐渐引入到WSSS任务中(Dosovitskiy等,2021)。例如,Gao等人(2021)提出TS-CAM(token semantic coupled attention map)模型,通过结合语义信息和注意力机制增强弱监督学习下的对象定位能力。Xu 等人(2022)提出MCTformer,利用多个类标记之间的交互关系有效捕获特定类的注意力,从而提升CAM的区分能力。虽然两阶段方法可以实现较优的分割性能,但其分离式训练流程会导致计算成本高以及训练效率低等问题。

端到端的语义分割方法简化了训练流程,无需分阶段处理,在一个统一的框架内完成从图像级标签到像素级分割的过程。尽管这种方法较为高效,但仍面临如何生成高质量伪标签的挑战。为此,Araslanov 和 Roth(2020)提出 1Stage 方法,通过采用像素自适应掩膜细化技术,针对局部一致性、语义保真度和完整性3个关键性能指标进行优化,显著提升了语义分割任务的效率和准确性。Zhang等人(2020)利用条件随机场(conditional random field,CRF)模型捕捉像素之间的关系,在CAM上生成更

加平滑且准确的分割边界，进一步改善了伪标签的质量。此外，Ru等人（2022）根据Transformer模型中的语义相似度对伪标签中的类别区域进行扩散，并提出一种像素自适应细化模块，对初始生成的伪标签进行修正。尽管现有研究在提升伪标签质量方面已取得一定进展，但仍存在诸多局限，其中初始CAM的质量尤为关键。一旦CAM质量欠佳，其所提供的监督信息便相对有限，可能会对后续模型的学习效果产生不利影响。为了提升CAM的质量，Lee等人（2019）提出FickleNet，通过随机选择隐藏单元获得用于图像分类的激活分数，并利用膨胀卷积积累激活区域，从而生成更大且更准确的目标区域。Wu等人（2021）则通过探索多个输入图像之间的语义相关性和差异性，提出一种嵌入式判别注意机制（embedded discriminative attention mechanism，EDAM），生成较高质量的CAM。

为了进一步提高伪标签的质量并突破现有方法的局限，自监督学习（Doersch等，2015）和知识蒸馏（Hinton等，2015）方法引入到弱监督语义分割领域。其中，自监督学习能够通过挖掘图像内部结构特征生成监督信号，降低对人工标注的依赖，并借助图像内在结构生成高质量的伪标签，从而增强模型的分割性能（Wang等，2020；Zhang等，2021）。与此同时，知识蒸馏方法通过将教师模型的知识迁移到学生模型，显著提升了学生模型的特征学习能力与伪标签生成的质量（Gou等，2021）。其中，在线蒸馏将教师和学生模型同步优化，实现了动态调整和知识的持续传递，表现出较大的灵活性与鲁棒性。Xu等人（2023）提出一种自对应蒸馏方法（self correspondence distillation，SCD），融合自监督学习和在线蒸馏的优势，以同一模型的教师—学生网络自身的特征对应关系为蒸馏目标，强化了学生网络的特征学习能力，改善了局部不一致性，进而提升了分割边界的准确性。Zhou等人（2022a）提出的iBOT（image BERT pre-training with online tokenizer）框架，通过指数移动平均（exponential moving average，EMA）平滑模型参数的更新，帮助模型更加稳定地学习特征表示。

尽管EMA可以通过平滑教师网络的参数更新来稳定学生网络的学习过程，但是在只有图像类别标签的弱监督语义分割领域中，教师网络在噪声标签的监督下可能产生不准确的特征表示。这些不准确的特征表示通过EMA传递给学生网络，可能导致学生网络学习到错误的特征表示，从而影响最终的分割性能。

为了解决上述问题，本文提出双阶段知识交互模块，该模块通过不同的训练阶段有效传递知识，减少噪声干扰，提升CAM的质量，并增强模型对目标物体细节部分的捕捉能力。此外，针对端到端弱监督语义分割框架中CAM训练不足而导致伪标签存在噪声的问题，本文提出高斯修正模块，该模块通过调整伪标签中不确定性区域和噪声部分的标签值，有效减少伪标签中的错误信息，为分割网络的训练提供了更加准确的监督信号，从而避免噪声干扰对学习过程产生负面影响，提高网络的分割性能。

综上所述，本文主要贡献如下：1）构建双阶段知识交互模块，促进学生模型更好地学习教师模型的知识，有效减少训练过程中的噪声干扰，从而提升模型整体性能。2）为了缓解伪标签生成不精确的问题，探索了一种基于高斯混合模型的伪标签修正模块。该模块能充分建模类激活图中像素间的关系，进而生成更加准确和可靠的伪标签。3）设计了不确定区域掩膜损失函数，降低了不确定区域对模型训练的不利影响，增强了类激活图定位目标物体区域的能力。4）在常用的PASCAL VOC 2012（pattern analysis，statical modeling and computational learning visual object classes 2012）数据集（Everingham等，2015）和MS COCO 2014（Microsoft common objects in context 2014）数据集（Lin等，2014）上进行实验，实验结果验证了所提出算法的有效性和可行性。

1 本文方法

1.1 整体框架概述

本文提出一种基于知识蒸馏的端到端弱监督语义分割方法，整体框架如图1所示，该方法采用编码器—解码器结构，在图像级分类标签的监督下实现语义分割。具体而言，特征提取阶段采用ViT（vision Transformer）（Dosovitskiy等，2021）编码器，其自注意力机制可以有效捕获全局上下文信息，从而提升分割性能。

为进一步提升模型训练效果，本文基于教师—学生自蒸馏机制设计了双阶段知识交互模块（dual-stage knowledge distillation，DSKD）。该模块主要分

为 3 个阶段：预热网络阶段、扩展网络阶段和分割阶段。

在预热网络阶段，通过对教师网络和学生网络的中间层 Patch token 施加语义一致性约束，帮助学生网络在训练初期建立稳固的特征表示，从而便于后续的知识传递和特征对齐。

在扩展网络阶段，学生模型的参数不再通过梯度反向传播进行更新，而是采用 EMA 机制，将教师模型的参数平滑地迁移到学生模型中。然后，借助 Projector 模块对预热阶段处理后的特征进行空间投影，提高模型对特征空间的理解和对齐能力；之后，根据投影后的 Patch token 计算损失 L_{PUNC}，促进网络在不确定区域间的有效信息交流。

与此同时，在分割阶段，利用预热阶段输出的 Patch token 生成 CAM，并基于该 CAM 得到初始伪标签；接着利用高斯修正模块（Gaussian correction，GC）对伪标签进行修正，提升其质量，使其可以更准确地反映目标物体的边界和形状。最后，这些经过修正的伪标签作为解码器的监督信号，指导模型生成最终的分割结果，确保分割的准确性和可靠性。

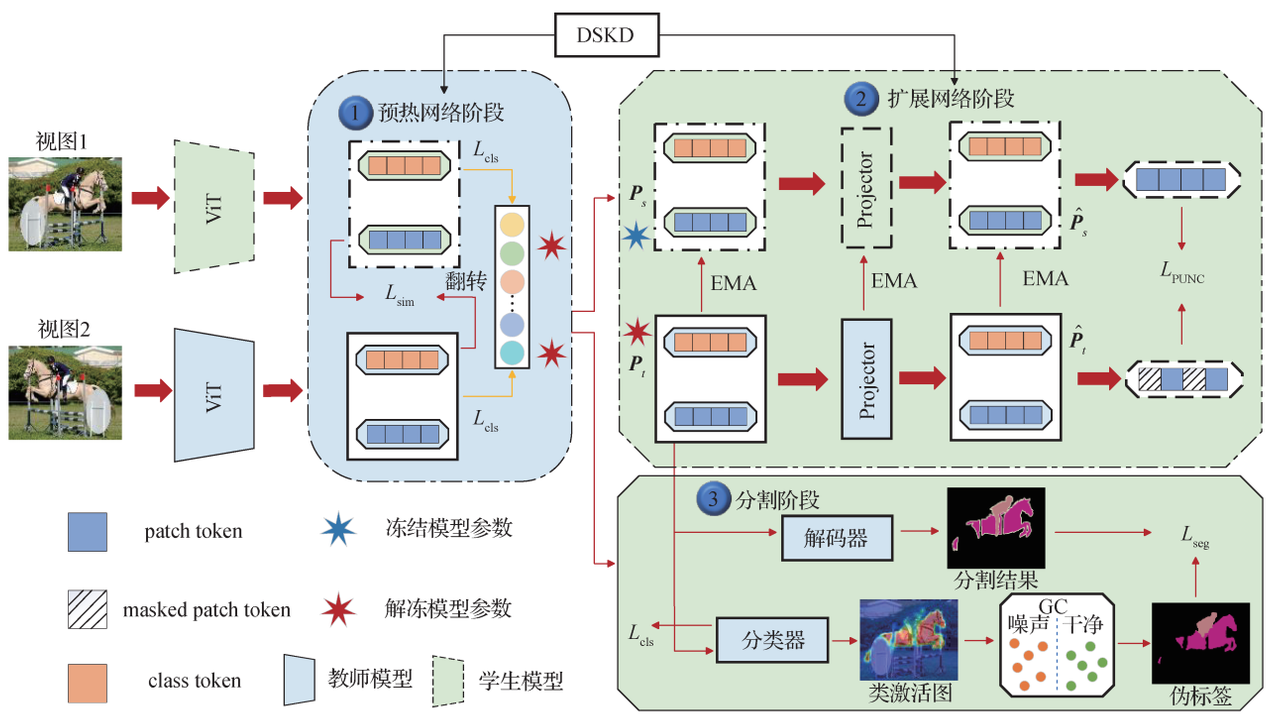


图 1　整体框架

Fig. 1　Overall framework

1.2　特征提取

本文方法采用 ViT 作为特征提取器，利用其强大的全局上下文建模能力获取高质量的图像特征。然而，ViT 在处理语义分割任务时，容易导致 Patch token 过度平滑，丢失关键的细节信息，影响分割的准确性。为了缓解这一问题，本文引入 PTC（patch token contrast）模块（Ru 等，2023），以提升特征提取效果。具体实施过程如下：

1）中间层语义信息提取。首先，从 ViT 的第 m 个中间层提取 patch token，并在该层附加一个辅助分类器，生成辅助 CAM，计算过程为

$$\boldsymbol{M}_c = \sum_i \boldsymbol{\theta}_{c,i} \boldsymbol{P}_{:,i}$$

$$\boldsymbol{CAM}_c(\boldsymbol{P}, \boldsymbol{\theta}) = \frac{ReLU(\boldsymbol{M}_c)}{\max(ReLU(\boldsymbol{M}_c))} \tag{1}$$

$$\boldsymbol{A}_m = \boldsymbol{CAM}(\boldsymbol{P}_m, \boldsymbol{\theta}_m) \tag{2}$$

式中，$\boldsymbol{P}_{:,i}$ 表示 Patch token 中第 i 个通道的所有像素值，$\boldsymbol{\theta}_{c,i}$ 表示类别 c 对应的第 i 个特征通道的权重，$\boldsymbol{P}_m$ 是第 m 个中间层的 patch token，$\boldsymbol{\theta}_m$ 是辅助分类器的权重，$\boldsymbol{A}_m$ 表示生成的辅助 CAM。

然后，使用两个阈值 β_l 和 β_h（$0 < \beta_l < \beta_h < 1$）将 $\boldsymbol{A}_m$ 划分为可靠的前景区域与背景区域，并进一步为每个 patch token 赋予对应的前景或背景伪标签 $\boldsymbol{Y}_m$。

具体为

$$\boldsymbol{Y}_{m,i}=\begin{cases}1 & \boldsymbol{A}_{m,i}>\beta_h\\ 0 & \boldsymbol{A}_{m,i}<\beta_l\end{cases} \tag{3}$$

2）最终patch token监督。在得到$\boldsymbol{Y}_m$之后，将具有相同语义标签的patch token对定义为正对，即$\boldsymbol{Y}_i=\boldsymbol{Y}_j$，而具有不同语义标签的token对定义为负对，即$\boldsymbol{Y}_i\neq\boldsymbol{Y}_j$。之后，利用$\boldsymbol{Y}_m$对最后一层输出的patch token进行监督。最后，PTC模块的损失函数L_{PTC}定义为

$$L_{\text{PTC}}=\frac{1}{N^+}\sum_{\boldsymbol{Y}_i=\boldsymbol{Y}_j}(1-CosSim(\boldsymbol{P}_i,\boldsymbol{P}_j))+\frac{1}{N^-}\sum_{\boldsymbol{Y}_i\neq\boldsymbol{Y}_j}CosSim(\boldsymbol{P}_i,\boldsymbol{P}_j) \tag{4}$$

$$CosSim(\boldsymbol{P}_i,\boldsymbol{P}_j)=\frac{\boldsymbol{P}_i\cdot\boldsymbol{P}_j}{\|\boldsymbol{P}_i\|\|\boldsymbol{P}_j\|} \tag{5}$$

式中，$\boldsymbol{P}_i$与$\boldsymbol{P}_j$分别表示图像中第i个和第j个patch token的特征向量，$CosSim(\boldsymbol{P}_i,\boldsymbol{P}_j)$表示$\boldsymbol{P}_i$与$\boldsymbol{P}_j$之间的余弦相似度，$N^+/N^-$表示正/负对的数量。

通过PTC损失优化最终的patch token，使其在特征空间中保持适当的区分度，从而提升ViT的特征提取能力。

1.3 双阶段知识交互模块

通过引入PTC模块优化特征提取阶段，增强了图像特征的语义表示。基于此，在本节中进一步探讨了如何在弱监督条件下通过知识蒸馏提升模型性能。

具体而言，在知识蒸馏中，教师模型通过梯度信息进行参数更新，而学生网络的参数则通过EMA逐步更新，从而稳定地从教师网络继承知识，提升学习的准确性和稳定性。学生网络的参数更新过程为

$$\boldsymbol{\theta}_t=\beta\cdot\boldsymbol{v}_t+(1-\beta)\cdot\boldsymbol{\theta}_{t-1} \tag{6}$$

式中，$\boldsymbol{\theta}_t$表示在时间步t的学生网络参数，$\boldsymbol{v}_t$表示在时间步t的教师网络参数，β是平滑系数，用于控制教师网络参数对学生网络更新的影响程度。$\boldsymbol{\theta}_{t-1}$表示在时间步$t-1$的学生网络参数。

从式(6)可以看出，尽管EMA能够平滑学生网络的学习过程，帮助学生网络稳定地从教师网络继承知识，但是在弱监督语义分割任务中，图像级标签仅提供图像的类别信息，教师网络可能受到噪声标签影响，提取到不准确的特征表示。这些误导性的特征信息通过EMA传递给学生网络，而EMA的平滑作用主要体现在稳定性上，致使模型在训练阶段对噪声信息的关注度不足，导致CAM定位精度下降，进而影响目标区域的识别与分割。

为了缓解上述问题，本文设计了DSKD模块，包括预热网络阶段、扩展网络阶段和分割阶段。

1.3.1 预热网络阶段

如图1所示，在初始预热阶段，利用在线蒸馏的优势减少噪声干扰，帮助学生网络学习可靠的特征表示。具体而言，教师网络和学生网络均设置为可训练状态，教师网络提取翻转增强的图像特征，学生网络提取原始图像特征，在此基础上，计算两者中间层Patch token的相似度损失，以促进教师网络与学生网络之间的特征一致性，具体为

$$L_{\text{sim}}=-\frac{1}{c}\sum_{c}\frac{Flip(\boldsymbol{P}_t)\cdot\boldsymbol{P}_s}{|Flip(\boldsymbol{P}_t)||\boldsymbol{P}_s|} \tag{7}$$

式中，$\boldsymbol{P}_t$表示教师网络的Patch token，$\boldsymbol{P}_s$表示学生网络的Patch token，c表示通道数，$Flip$表示翻转操作。同时，结合Class token与真实标签的分类损失，进一步提升模型在定位目标物体方面的能力，分类损失函数为

$$L_{\text{cls}}=-\frac{1}{C}\sum_{c=1}^{C}[y_c\log(\hat{y}_c)+(1-y_c)\log(1-\hat{y}_c)] \tag{8}$$

式中，$\hat{y}_c$表示第c类物体的存在概率，图像级标签向量中的第c个分量$y_c\in\{0,1\}$表示第c类是否存在。

1.3.2 扩展网络阶段

在扩展网络阶段冻结学生网络的参数，通过EMA平滑更新学生模型参数，确保知识的稳定传递。

然后通过Projector模块（Caron等，2021）对预热阶段处理后的patch token进行更精细的对齐，得到投影特征$\widehat{\boldsymbol{P}}_s$以及$\widehat{\boldsymbol{P}}_t$。具体而言，Projector模块首先通过3层感知机对输入特征进行非线性变换，然后利用权重归一化的全连接层将特征映射到新的空间，从而实现特征的投影和对齐。由于教师网络和学生网络中的某些patch token可能存在不明确归类为前景或背景的像素，导致损失计算受到干扰。因此，本文设计不确定区域掩膜策略。通过设置阈值l_{th}和h_{th}，选择位于两个阈值之间的CAM区域，从而生成感兴趣区域$\boldsymbol{ROI}$（region of interest），具体为

$$\boldsymbol{ROI}(i,j)=\begin{cases}1 & l_{\text{th}}<A(i,j)<h_{\text{th}}\\ 0 & \text{其他}\end{cases} \tag{9}$$

式中，A表示通过式(1)生成的CAM，$A(i,j)$表示

CAM沿着通道维度的最大值。

随后，将投影特征以及$\boldsymbol{ROI}$送入L_{PUNC}中进行损失计算，以鼓励教师网络与学生网络在不确定区域的特征表达上进行有效的信息交互，减弱了噪声对CAM生成的影响，并提升模型在不确定区域的表现。具体为

$$L_{\mathrm{PUNC}}=\frac{1}{\sum \boldsymbol{ROI}}\sum_{i}\left(-\sigma_s(\widehat{\boldsymbol{P}}_s)\sigma_t\left(\widehat{\boldsymbol{P}}_t\right)\right)\boldsymbol{ROI}_i \tag{10}$$

式中，σ_s为softmax操作，σ_t为log_softmax操作。

1.3.3　分割阶段

在分割阶段，利用预热阶段输出的Patch token生成CAM，并基于该CAM得到初始伪标签，随后，在高斯修正模块GC中对这些伪标签进行修正，产生更加精确的监督信号以指导解码器的训练，从而提升模型在分割任务中的表现。

然而，由于CAM在预热阶段训练不充分，所生成的伪标签中通常存在大量噪声，严重影响了后续训练的监督质量。为缓解这一问题，本文提出一种基于高斯混合模型（Gaussian mixture model，GMM）（Xuan等，2001）的伪标签修正模块。假设像素点的标签置信度呈现双峰分布，其中高置信度的像素点更可能准确地表示真实的标签，因为它们反映了模型对这些像素类别的高确定性。相反，低置信度的像素点则更容易受到噪声影响，导致数据不准确（Jiang等，2024）。基于这一假设，本文使用GMM对CAM输出的像素置信度进行拟合，将像素点划分为高置信度和低置信度两类，从而估算每个像素点的噪声概率，拟合过程为

$$\mathcal{N}(x|\mu,\sigma^2)=\frac{1}{\sqrt{2\pi}\,\sigma}\mathrm{e}^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{11}$$

假设高斯混合模型拟合的两个分布分别为噪声分布和干净分布，那么总的概率密度为

$$p(x)=\sum_{k=1}^{K}\lambda_k\mathcal{N}(x_n|\mu_k,\sigma_k^2) \tag{12}$$

式中，$K=2$，μ_k、σ_k^2表示每一个分布的均值和方差，λ_k表示每一个分布的权重系数。

对于隐变量μ_k、σ_k^2，可以使用EM（expectation maximization）算法确定。具体而言，先计算每个像素点属于干净分布和噪声分布的后验概率，其计算式为

$$\gamma(x_n|\mu_k,\sigma_k^2)=\frac{\lambda_k\mathcal{N}(x_n|\mu_k,\sigma_k^2)}{\sum_{j=2}^{J=2}\lambda_j\mathcal{N}(x_n|\mu_j,\sigma_j^2)} \tag{13}$$

式中，$\mathcal{N}(x_n|\mu_k,\sigma_k^2)$表示第$k$个高斯分布在$x_n$处的概率密度。

在E步，固定参数μ_k、σ_k^2以及权重系数λ_k，并根据式（13）更新$\gamma(x_n|\mu_k,\sigma_k^2)$。

在M步，固定$\gamma(x_n|\mu_k,\sigma_k^2)$值，更新参数$\mu_k$、$\sigma_k^2$以及权重系数$\lambda_k$，其计算式为

$$\mu_k=\frac{\sum_{n=1}^{N}\gamma(x_n|\mu_k,\sigma_k^2)x_n}{\sum_{n=1}^{N}\gamma(x_n|\mu_k,\sigma_k^2)} \tag{14}$$

$$\sigma_k^2=\frac{\sum_{n=1}^{N}\gamma(x_n|\mu_k,\sigma_k^2)(x_n-\mu_k)^2}{\sum_{n=1}^{N}\gamma(x_n|\mu_k,\sigma_k^2)} \tag{15}$$

$$\lambda_k=\frac{1}{N}\sum_{n=1}^{N}\gamma(x_n|\mu_k,\sigma_k^2) \tag{16}$$

式中，N表示像素点的总个数。

迭代E步和M步，直到收敛或达到最大迭代次数。最后估计噪声概率，具体为

$$\widehat{G}_n=\frac{\lambda_1\mathcal{N}(x_n|\mu_1,\sigma_1^2)}{\sum_{t=1}^{2}\lambda_t\mathcal{N}(x_n|\mu_t,\sigma_t^2)} \tag{17}$$

式中，$\widehat{G}_n$代表类激活图中的每个像素点对应的噪声概率，$\mathcal{N}(x_n|\mu_1,\sigma_1^2)$表示$x_n$属于较小标签置信度的噪声分布，$\lambda_1$表示属于较小标签置信度噪声分布的权重系数。

随后，根据噪声概率的分布特性，设定高概率阈值$\theta_h(0<\theta_h<1)$与低概率阈值$\theta_l(0<\theta_l<1)$，从而分别定义高概率噪声点集合$\boldsymbol{R}_h(\widehat{G}_n>\theta_h)$和不确定概率噪声点集合$\boldsymbol{R}_u(\theta_l<\widehat{G}_n<\theta_h)$。然后，对集合$\boldsymbol{R}_h$中的噪声点进行选择性剔除和重新分类，对集合$\boldsymbol{R}_u$中的噪声点进行修正。

经分析可知，$\boldsymbol{R}_h$中的噪声点通常集中在图像的背景区域和目标物体的边界处。对于每个高概率噪声像素点，本文计算其与4个相邻像素的余弦相似度。假设$\boldsymbol{F}_h$表示当前像素点，特征维度为$\widehat{d}$，并将其拉伸为形状为$(1,\widehat{d})$的张量。选取当前像素点的4个邻域像素点对应的特征向量，分别记为$\boldsymbol{F}_{h1}$、$\boldsymbol{F}_{h2}$、$\boldsymbol{F}_{h3}$和$\boldsymbol{F}_{h4}$。接下来，通过式（4）计算$\boldsymbol{F}$与4个邻域像素点的余弦相似度，得到一个长度为4的相似度值向量$\boldsymbol{S}_h$，具体为

$$\boldsymbol{S}_h = \left[CosSim\left(\boldsymbol{F}_h, \boldsymbol{F}_{h1}\right), \cdots, CosSim\left(\boldsymbol{F}_h, \boldsymbol{F}_{h4}\right)\right] \tag{18}$$

如果向量$\boldsymbol{S}_h$中存在小于预设阈值τ的值，说明当前像素与某一邻域像素之间的相似度较低，可能表明当前像素与邻域像素在类别上的不一致。特别地，低相似度值通常表明当前像素与邻域像素的语义标签存在差异（例如，一个属于前景，一个属于背景），这可能意味着当前像素位于物体的边界，并且属于噪声点，表明当前像素可能已被错误分类。此时，这类像素的标签将被重新标定为背景标签0。此外，对于$\boldsymbol{R}_h$中不满足上述条件的其余像素点，由于它们通常位于背景区域，并且对网络训练的影响较小，本文将这些像素的标签值设置为255，以便在后续损失计算中忽略它们。因此，$\boldsymbol{R}_h$中的伪标签修正表示为

$$pseudo_h = \begin{cases} 0 & \exists s \in \boldsymbol{S}_h, s < \tau \\ 255 & \forall s \in \boldsymbol{S}_h, s \geqslant \tau \end{cases} \tag{19}$$

式中，s表示相似度向量$\boldsymbol{S}_h$中的一个分量。

而集合$\boldsymbol{R}_u$中的噪声点通常集中在图像内部以及多个目标物体的连接处。由于前景和背景像素在其特征表示上通常包含不同的语义信息，在特征空间中，二者应当有明显的区分。因此，本文利用余弦相似度判断前景与背景之间的相似性，从而有效地抑制噪声并强化前景—背景的语义分离。具体而言，通过式(4)计算当前像素$\boldsymbol{F}_u$与4个邻域像素点的余弦相似度，得到一个长度为4的相似度值向量$\boldsymbol{S}_u$，具体为

$$\boldsymbol{S}_u = \left[CosSim\left(\boldsymbol{F}_u, \boldsymbol{F}_{u1}\right), \cdots, CosSim\left(\boldsymbol{F}_u, \boldsymbol{F}_{u4}\right)\right] \tag{20}$$

当相似度值向量$\boldsymbol{S}_u$中出现小于预设阈值τ的情况时，表明当前像素点与其邻域像素点中同时存在前景像素和背景像素，导致当前像素点可能位于目标物体的边界处。在这种情况下，当前像素的标签值被重新标定为背景标签0。对于$\boldsymbol{R}_u$中不满足上述条件的其余像素点，根据相似度值最大的邻域像素标签值更新其伪标签。因此，$\boldsymbol{R}_u$中伪标签修正为

$$pseudo_u = \begin{cases} 0 & \exists s \in \boldsymbol{S}_u, s < \tau \\ ps(\max(\boldsymbol{S}_u)) & \forall s \in \boldsymbol{S}_u, s \geqslant \tau \end{cases} \tag{21}$$

式中，$\max(\boldsymbol{S}_u)$表示具有相似度最高的相邻像素，$ps(\max(\boldsymbol{S}_u))$表示与当前像素相似度最高的邻域像素所对应的标签值，s表示相似度向量$\boldsymbol{S}_u$中的一个分量。

在GC模块的修正作用下，伪标签得到了显著优化，可以进一步提升语义分割任务的预测准确性。具体而言，修正后的伪标签作为监督信号，指导解码器生成更加精确的分割结果。为进一步优化模型的训练过程，本文采用常用的交叉熵损失函数作为分割损失，具体为

$$L_{\text{seg}} = -\frac{1}{N}\sum_{n=1}^{N}\sum_{i=1}^{H}\sum_{j=1}^{W}\sum_{c=1}^{C} y_{ijc}\log(\tilde{P}_{ijc}) \tag{22}$$

式中，$\tilde{P}_{ijc}$表示模型预测的第i行第j列像素属于类别c的概率，y_{ijc}表示真实标签中第i行第j列像素的类别，如果是类别c，则为1，否则为0。通过与真实标签对比，修正后的伪标签有效地纠正了预测偏差，特别是在一些难以区分的边界区域和复杂背景区域，缓解了传统方法中常见的误分割现象。

综上所述，本文提出的弱监督语义分割模型中，总的损失函数为

$$L_{\text{all}} = L_{\text{cls}} + \lambda_1 L_{\text{PTC}} + \lambda_2 L_{\text{sim}} + \lambda_3 L_{\text{PUNC}} + \lambda_4 L_{\text{seg}} \tag{23}$$

2 实验结果及分析

2.1 数据集与参数设置

本文实验在数据集PASCAL VOC 2012（Everingham等，2015）和MS COCO 2014（Lin等，2014）上进行。PASCAL VOC 2012数据集有20个目标类别和1个背景类别，训练集图像10 582幅，验证集图像1 449幅，测试集图像1 456幅。MS COCO 2014数据集包含80个目标类别和1个背景类，训练集图像82 081幅，验证集图像40 137幅。在整个训练过程中只使用图像级标签。所有模型采用PyTorch框架实现，在两张24 GB显存的Nvidia 4090上进行训练。

采用预训练ImageNet的ViT-B作为Transformer的编码器。卷积解码器参考了DeepLab-LargeFOV（Chen等，2018）。在本文中使用了轻量级的数据增强：将图像随机缩放裁剪至448×448像素，缩放比例在(0.32，1.0)之间，长宽比在(3/4，4/3)之间。

在PASCAL VOC 2012数据集上的训练总迭代次数设置为20 000次，其中预热阶段训练迭代次数为2 000次。在MS COCO 2014数据集上总迭代次数设置为80 000次，预热阶段训练迭代次数为5 000次。在预热阶段训练中，学生网络和教师网络均采用AdamW进行优化。教师网络的基础学习率设置为6E-5，学生网络的基础学习率设置为6E-6，并通

过余弦调度进行衰减。在预热阶段训练完成之后，学生网络停止梯度更新，通过EMA动量进行更新。教师网络的动量设置为0.996，并在训练过程中逐步升至1.0，学生网络的动量设置为0。Projector模块包括一个3层感知机和一个权重归一化的全连接层（Caron等，2021）。在训练过程中损失函数权重$(\lambda_1, \lambda_2, \lambda_3, \lambda_4)$设置为(0.2, 0.5, 0.1, 0.1)，感兴趣区域的前背景阈值设置为$(l_{\text{th}}, h_{\text{th}}) = (0.2, 0.7)$。

实验中，采用平均交并比（mean intersection over union，mIoU）作为评估指标，通过计算预测结果和真实标签之间的平均交并比衡量不同模型的分割性能，计算式为

$$mIoU = \frac{1}{C}\sum_{c}\frac{\boldsymbol{P}_c \cap \boldsymbol{G}_c}{\boldsymbol{P}_c \cup \boldsymbol{G}_c} \tag{24}$$

式中，$\boldsymbol{P}_c$表示生成的分割区域，通常由伪标签或最终分割结果表示，$\boldsymbol{G}_c$表示真实标注，作为评估依据。

2.2 性能分析及消融实验研究

为了研究不同模块在模型中的有效性，本文在PASCAL VOC 2012数据集上进行消融实验，结果如表1所示。基线表示不使用GC和DSKD。引入DSKD后，模型的分割效果与基线相当，但如图2所示，在引入DSKD后，模型在一些细节部分的分割效果显著优于基线方法。引入GC后，分割精度较引入GC之前提升了1.14%，表明GC模块在伪标签修正方面发挥了重要作用，从而有效提升了分割精度。伪标签可视化结果见图2。

表1 伪标签生成消融实验

Table 1 Ablation experiment of pseudo label generation

模块			mIoU/%
Baseline	DSKD	GC	
√	–	–	68.57
√	√	–	68.53
√	√	√	**69.67**

注：加粗字体表示各列最优结果。“√”表示采用，“–”表示未采用。

如图2所示，第1行表示未使用任何模块生成的伪标签。第2行表示仅使用DSKD模块生成的伪标签。第3行表示引入DSKD模块和GC模块生成的伪标签。第4行表示真实标签。

从第1列的对比图可以看出，红色方框内的目标物体在使用GC和DSKD之后可以清晰地分割出狗和人之间的间隙，并识别出更多的目标区域。具体而言，GC模块通过计算当前像素与周围像素的相似度，并将最相似的像素标签赋予当前像素，从而有效地修正了伪标签。与此同时，DSKD模块在训练初期通过减少噪声干扰，进一步增强了模型对目标区域的识别能力，从而提升了分割精度。然而，狗的腿未能准确分割，这可能是因为GC模块在修正伪标签时，误将腿部区域的标签赋予了狗与人之间相似度最高的区域，从而导致分割错误。

在第2列的结果中，牛的尾巴和腿部在本文方法中被有效分割出来。这是由于该图像仅包含一个类别，伪标签修正过程相对简单，有效避免了误分割情况。在相似度计算中，尾巴区域的特征与牛的身体部分相似，从而确保了该区域被正确地识别为牛的组成部分，确保了分割的准确性。

在第3列中，红色方框内的目标物体相较于基线实现了更为清晰的分割，这得益于DSKD模块在训练初期有效降低了噪声对训练过程的干扰，从而提升了模型对目标区域的识别能力。然而，与真实标签相比，红色方框内物体存在间隙。这可能有两种原因：1）利用高斯混合模型拟合CAM分布时误将红色方框区域识别为高概率噪声点，直接将其标签值设置为背景像素0；2）红色方框区域被识别为不确定噪声点，但是在相似度计算中，由于当前像素与邻近像素的相似度值低于阈值，因此，标签被重新修正为背景像素0。

在第4列中，结合GC模块和DSKD模块后，马的腿得以准确分割，进一步验证了该模块在细节分割中的有效性。

此外，本文还对使用不同模块生成的CAM进行可视化，如图3所示。图3中基线表示不使用GC和DSKD生成的类激活图。+DSKD表示引入了DSKD模块，+GC表示引入了GC模块。从图3可以看出，本文方法可以很好地排除一些噪声的影响，并且可以更完整地激活目标物体区域。

为了进一步探索更适合本文模型的噪声概率阈值参数，进行了多次实验，最终确定$\theta_h = 0.85$以及$\theta_l = 0.25$，实验结果如表2所示。可以看出，当$\theta_h = 0.85$、$\theta_l = 0.25$时，无论是CAM还是分割结果，均达到最优结果。

为确定损失函数中各项的最优权重，本文设计

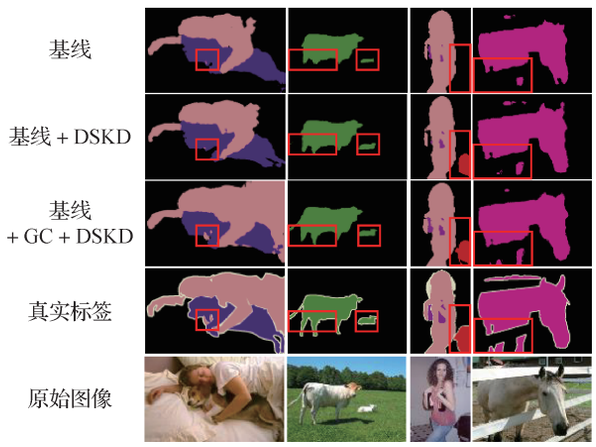
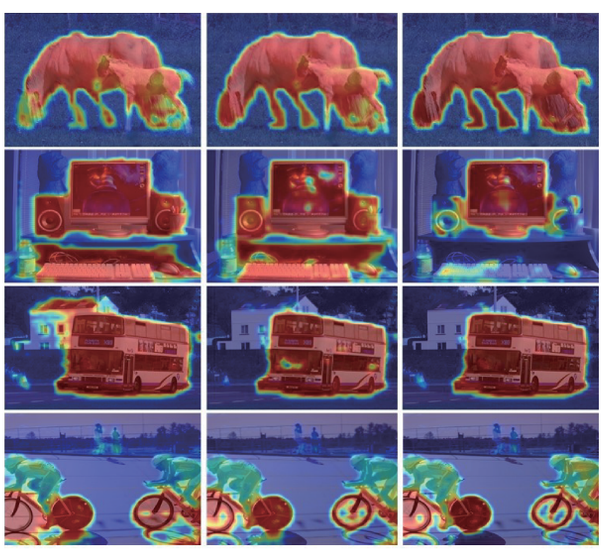


图2　消融实验伪标签可视化

Fig. 2　Visualization of pseudo labels in ablation experiments



(a) 基线　(b) + DSKD　(c) + GC + DSKD

图3　CAM可视化

Fig. 3　CAM visualization

((a) baseline; (b) + DSKD; (c) + GC + DSKD)

表2　不同阈值生成CAM以及分割结果精度对比

Table 2　Comparison of CAM image generation and segmentation accuracy using different thresholds

θ_h	θ_l	mIoU/%	
		分割结果	CAM
0.75	0.35	69.012	70.966
0.85	0.25	**69.670**	**71.429**
0.90	0.20	69.390	70.892

注:加粗字体表示各列最优结果。

并开展了多组对比实验,其中λ_4的取值遵循ToCo(token contrast)(Ru等,2023)以及TSCD(self correspondence distillation)(Xu等,2023)中的设置。如表3所示,最终确定λ_1、λ_2和λ_3取值为0.2、0.5和0.1,此时分割结果达到了最优。

表3　不同损失函数权重生成分割结果精度对比

Table 3　Comparison of segmentation accuracy using different loss function weight settings

λ_1	λ_2	λ_3	mIoU/%
0.1	0.4	0.1	67.19
0.1	0.5	0.1	67.18
0.1	0.6	0.1	68.04
0.2	0.4	0.1	68.82
0.2	0.5	0.1	**69.67**
0.2	0.5	0.2	68.95
0.2	0.6	0.1	68.56

注:加粗字体表示最优结果。

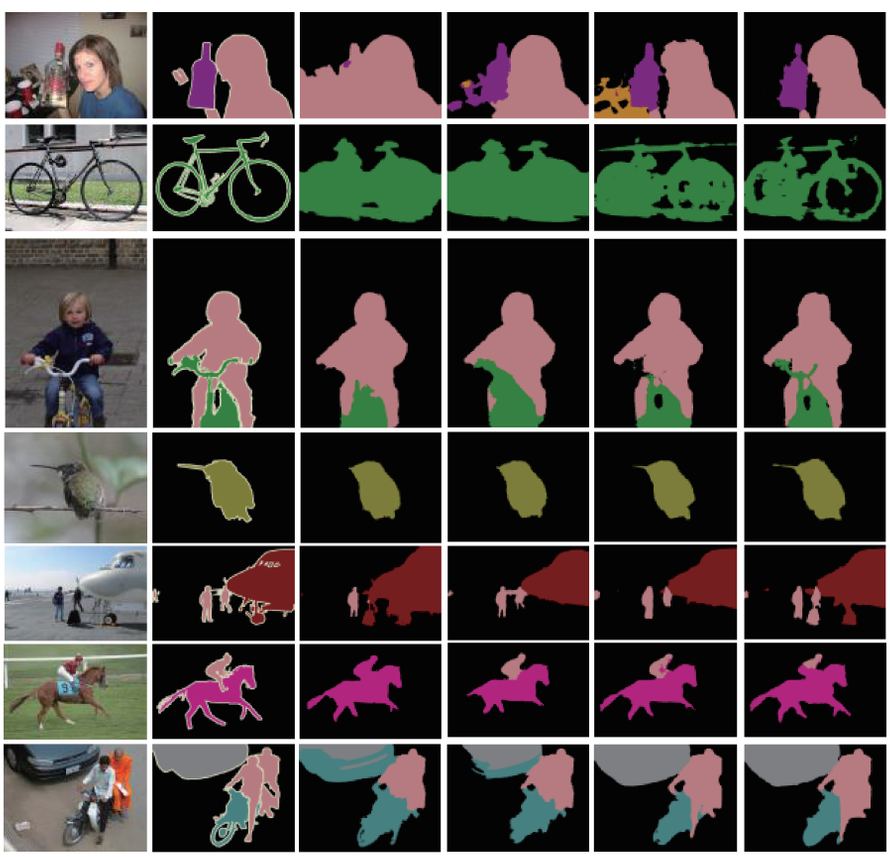
2.3　分割实验结果对比

为了进一步验证本文方法在提升类激活图完整性以及伪标签分割精度方面的效果,在PASCAL VOC 2012训练集和MS COCO 2014数据集上进行语义分割实验,并在PASCAL VOC 2012的测试集与验证集以及MS COCO 2014的验证集上对其产生效果进行评估。

图4展示了本文方法与对比方法AFA(affinity from attention)(Ru等,2022)、TSCD(Xu等,2023)和ToCo(Ru等,2023)在一些验证集图像上的分割结果对比。第1行图像中包含多个间隙的物体,AFA方法未能有效区分这些间隙,且错误地将部分背景区域识别为前景。尽管TSCD以及ToCo方法成功分割出了杯子,但仍未能有效解决背景区域的误分割问题。相比之下,本文方法不仅准确分割出了杯子,还显著减少了背景区域的误分割现象,展现了更强大的分割能力。

在第2行关于自行车的图像中,由于自行车结构存在大量间隙,使得分割任务变得更加复杂。AFA、TSCD和ToCo方法仅能识别出自行车的整体轮廓,未能捕捉到间隙的细节,而本文方法则有效分割出自行车的间隙区域,展现出优越的性能。对于第3、6、7行图像中的复杂目标,物体连接处容易出现误分类。AFA、TSCD和ToCo方法在这些区域的分割效果均不理想。第3行中,3种方法均未能正确分割出手与自行车的连接部位;第6行中,AFA和

TSCD将人与马错误地识别为一个整体，尽管ToCo方法成功分割了两者，但仍存在部分误分类问题。第7行中，AFA和TSCD未能准确区分腿与摩托车，而ToCo虽然成功识别到了腿和两人之间的间隙，但依然有不足之处。相比之下，本文方法可以对这些复杂目标的细节区域进行较为精确的分割。第4行鸟类图像中，AFA、TSCD和ToCo方法均未能准确分割鸟嘴这一细小部位。相比之下，本文方法可以精确识别并分割，显示出对小目标物体分割的优越性。第5行中，本文方法相较于AFA、TSCD和ToCo，在人与飞机的细节分割及小目标分割上表现更优。



(a) 原图　(b) 真实标签　(c) AFA　(d) TSCD　(e) ToCo　(f) 本文

图4　各方法在 PASCAL VOC 2012 验证集的分割结果对比

Fig. 4　Comparison of segmentation results of different methods on PASCAL VOC 2012 val set

((a) original images; (b) ground truth; (c) AFA; (d) TSCD; (e) ToCo; (f) ours)

尽管本文方法在整体上取得了较好的分割性能，但是在某些物体上仍存在误分割现象。例如，在第5行中，标牌被误识为人物，反映出模型在处理颜色相近、视觉相似的物体时仍难以判别。此外，在第6行中，虽然模型能够较好地分割出马的整体区域，但却将人腿误判为马的一部分，导致对人物的分割效果不如TSCD和ToCo。这些问题说明本文方法在处理语义相近或颜色相似区域时的局限性，也为后续工作提供了改进方向。总体而言，本文方法在处理多种复杂场景时均展现出了优秀的分割性能，相较于对比方法具有显著优势。

表4列出了本文方法与RRM(reliable region mining)(Zhang等，2020)、1Stage(Araslanov和Roth，2020)、AuxSegNet+(Xu等，2025)、DuPL(dual student framework with trustworthy progressive learning)(Wu等，2024)、AFA、TSCD、ToCo等端到端的弱监督语义分割方法，以及SEAM(self-supervised equivariant attention mechanism)(Wang等，2020)、SC-CAM(Chang等，2020)、AdvCAM(Lee等，2021a)、ReCAM(Chen等，2022b)、RIB(reducing information bottleneck)(Lee等，2021b)、URN(uncertainty estimation via response scaling for noise mitigation)(Li等，2022)、

BECO（boundary-enhanced co-training）（Rong 等，2023）等两阶段的弱监督语义分割方法在PASCAL VOC 2012数据集上所生成伪标签的精度对比。

表4 各方法在PASCAL VOC 2012数据集上生成伪标签精度对比

Table 4 Comparison of pseudo-label accuracy of different methods on PASCAL VOC 2012 dataset

类别	方法	基干网络	mIoU/%	
			验证集	测试集
两阶段弱监督语义分割方法	SEAM	R38	64.5	65.7
	SC-CAM	R101	66.1	65.9
	AdvCAM	R101	68.1	68.0
	ReCAM	R101	68.5	68.4
	RIB	R101	68.3	68.6
	AuxSegNet+	R38	70.7	70.9
	URN	R101	69.5	69.7
	BECO	MiT-B2	73.7	73.5
端到端弱监督语义分割方法	1Stage	R38	62.7	62.9
	RPM	R38	62.7	64.3
	TSCD	MIT-B1	67.3	67.5
	AFA	MIT-B1	66.0	66.3
	DuPL	ViT-B	73.3	72.8
	ToCo	ViT-B	71.1	72.2
	本文	ViT-B	**74.8**	**73.9**

注：加粗字体表示各列最优结果。

从表4可以看出，本文方法明显优于其他方法。在验证集上，与ToCo、AFA和TSCD相比，本文方法结果提升了3.7%、8.8%和7.5%。在测试集上，与ToCo、AFA和TSCD相比，本文方法结果提升了1.7%、7.6%和6.4%。通过使用DSKD模块，削弱了噪声对训练过程的干扰，特别是对类激活图定位目标物体区域的影响。在此基础上，通过GC模块进一步精炼伪标签的生成过程，从而有助于生成更加准确和完整的分割结果。

为了进一步验证本文算法对于多目标和小尺寸目标的分割性能，本文在类别更丰富且包含大量多目标图像的MS COCO 2014数据集上进行实验。表5给出了本文方法与AuxSegNet+、TSCD、AFA、SEAM、MCTformer、URN、CDA（Su等，2021）、SIPE（Chen等，2022a）等算法在MS COCO 2014验证集上的性能对比。可以看出，本文方法在所有对比的方法中达到了最优，并且性能超过两阶段方法。

表5 各方法在MS COCO 2014数据集上伪标签精度对比

Table 5 Comparison of pseudo-label generation accuracy of different methods on MS COCO 2014 dataset

类别	方法	基干网络	mIoU/%
两阶段弱监督语义分割方法	SEAM	R38	31.9
	MCTformer	R38	42.0
	AuxSegNet+	R38	37.0
	CDA	R38	33.2
	URN	R101	40.7
	SIPE	R101	40.6
端到端弱监督语义分割方法	TSCD	MIT-B1	39.2
	TSCD+CRF	MIT-B1	40.1
	AFA	MIT-B1	38.0
	AFA+CRF	MIT-B1	38.9
	本文	ViT-B	**42.3**

注：加粗字体表示各列最优结果。

表6列出了各分割模型在PASCAL VOC 2012验证集上得到的21个类别的mIoU值对比结果。可以看出，本文方法在17个类别上优于所比较的方法。尤其是在背景类（bkg）上，本文方法取得92.4%的mIoU值，相比其他方法表现最佳。表明本文方法可以有效利用GC模块，减少背景区域的误分割，从而提升分割性能。此外，在bird、bottle、car、chair、cow等类别中，本文方法也取得了显著的效果，进一步验证了其有效性。

3 结 论

本文提出一种融合高斯修正的端到端弱监督语义分割方法，有效提升了伪标签质量和分割性能。本文设计DSKD模块增强了教师网络和学生网络的特征对齐，有效减少了噪声对训练过程的干扰。同时，在自监督条件下，通过高斯混合模型对生成的CAM进行建模，并结合噪声概率及像素间相似度关系进一步优化伪标签的质量。之后，在PASCAL VOC 2012和MS COCO 2014数据集上的大量实验结果验证了本文方法的优越性，与TSCD和ToCo等先进方法相比，本文方法展现出更强的鲁棒性。

表 6　各方法在PASCAL VOC 2012验证集上不同类别的mIoU对比

Table 6　mIoU comparison of different classes and methods on PASCAL VOC 2012 val set

/%

方法	点mIoU	bkg	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow
TSCD	67.3	87.4	70.6	**61.6**	75.2	55.4	62.8	75.1	57.7	77.4	39.4	77.4
AFA	66.0	85.8	71.4	58.8	73.8	57.7	57.8	77.8	66.7	77.7	27.7	79.5
ToCo	71.1	89.9	81.7	35.3	68.2	62.1	76.3	83.7	80.4	87.7	24.6	88.1
本文	**74.8**	**92.4**	**86.5**	47.6	**80.5**	**65.8**	**80.3**	**85.1**	**82.1**	**90.6**	**42.1**	**91.1**

方法	table	dog	horse	mbk	person	plant	sheep	sofa	train	tv
TSCD	**71.7**	66.6	69.1	49.7	76.8	42.9	60.1	42.9	60.5	53.3
AFA	43.6	74.1	68.7	64.2	62.7	51.3	75.6	39.2	59.2	43.9
ToCo	54.8	**87.0**	84.1	76.0	68.1	65.8	85.7	42.6	57.8	**65.6**
本文	65.1	86.9	**88.2**	**77.8**	**82.7**	**66.7**	**88.7**	**51.7**	**61.3**	57.1

注：加粗字体表示各列最优结果。

尽管本文方法在分割细节方面取得了显著改进，但在处理语义相近或视觉相似的场景时，可能会出现一些误分割，尤其在背景区域的判别上表现出一定的局限性。主要原因在于伪标签生成过程在边界区域存在不确定性，且模型对相似语义类别的判别能力仍有限。未来的研究将聚焦于提升模型在语义和视觉相似区域中的区分能力，进一步优化伪标签生成与校正机制，并在保持端到端结构优势的基础上，持续提升模型的分割精度与泛化能力。

参考文献（References）

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: 4253-4262 [DOI: 10.1109/cvpr42600.2020.00431]

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]

Caron M, Touvron H, Misra I, Jegou H, Mairal J, Bojanowski P and Joulin A. 2021. Emerging properties in self-supervised vision transformers//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision. Montreal, Canada: IEEE: 9650-9660 [DOI: 10.1109/iccv48922.2021.00951]

Chang Y T, Wang Q S, Hung W C, Piramuthu R, Tsai Y H and Yang M H. 2020. Weakly supervised semantic segmentation via subcategory exploration//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 8991-9000 [DOI: 10.1109/cvpr42600.2020.00901]

Chen L C, Papandreou G, Kokkinos I, Murphy K and Yuille A L. 2018. DeepLab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. IEEE Transactions on Pattern Analysis and Machine Intelligence, 40(4): 834-848 [DOI: 10.1109/tpami.2017.2699184]

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 Z Z, Wang T, Wu X W, Hua X S, Zhang H W and Sun Q R. 2022b. Class re-activation maps for weakly-supervised semantic segmentation//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE: 969-978 [DOI: 10.1109/cvpr52688.2022.00104]

Doersch C, Gupta A and Efros A A. 2015. Unsupervised visual representation learning by context prediction//Proceedings of 2015 IEEE International Conference on Computer Vision. Santiago, Chile: IEEE: 1422-1430 [DOI: 10.1109/iccv.2015.167]

Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D and Zhai X H. 2021. An image is worth 16 × 16 words: transformers for image recognition at scale [EB/OL]. [2025-02-17]. https://arxiv.org/pdf/2010.11929.pdf

Du Y Z, Arslanturk E, Zhou Z and Belcher C. 2011. Video based noncooperative iris image segmentation. IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics, 41(1): 64-74 [DOI: 10.1109/tsmcb.2010.2045371]

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 W, Wan F, Pan X J, Peng Z L, Tian Q, Han Z J, Zhou B L and Ye Q X. 2021. TS-CAM: token semantic coupled attention map for weakly supervised object localization//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision. Montreal, Canada: IEEE: 2886-2895 [DOI: 10.1109/iccv48922.2021.00288]

Gou J P, Yu B S, Maybank S J and Tao D C. 2021. Knowledge distillation: a survey. International Journal of Computer Vision, 129(6): 1789-1819 [DOI: 10.1007/s11263-021-01453-z]

Hinton G, Vinyals O and Dean J. 2015. Distilling the knowledge in a neural network [EB/OL]. [2025-02-17]. https://arxiv.org/pdf/1503.02531.pdf

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/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 7014-7023 [DOI: 10.1109/cvpr.2018.00733]

Jiang G X, Zhang J, Bai X F, Wang W J and Meng D Y. 2024. Which is more effective in label noise cleaning, correction or filtering?//Proceedings of the 38th AAAI Conference on Artificial Intelligence. Vancouver, Canada: AAAI: 12866-12873 [DOI: 10.1609/aaai.v38i11.29183]

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]

Lee J, Choi J, Mork J C and Yoon S. 2021b. Reducing information bottleneck for weakly supervised semantic segmentation [EB/OL]. [2025-02-17]. https://arxiv.org/pdf/2110.06530.pdf

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. Long Beach, USA: IEEE: 5267-5276 [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: 4071-4080 [DOI: 10.1109/cvpr46437.2021.00406]

Li Y, Duan Y Q, Kuang Z H, Chen Y M, Zhang W and Li X M. 2022. Uncertainty estimation via response scaling for pseudo-mask noise mitigation in weakly-supervised semantic segmentation//Proceedings of the 36th AAAI Conference on Artificial Intelligence. Virtually: AAAI: 1447-1455 [DOI: 10.1609/aaai.v36i2.20034]

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/CVF 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]

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]

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: 16846-16855 [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]

Singh K K and Lee Y J. 2017. Hide-and-Seek: forcing a network to be meticulous for weakly-supervised object and action localization//Proceedings of 2017 IEEE International Conference on Computer Vision. Venice, Italy: IEEE: 3544-3553 [DOI: 10.1109/iccv.2017.381]

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: 7004-7014 [DOI: 10.1109/iccv48922.2021.00692]

Wang X, You S D, Li X and Ma H M. 2018. Weakly-supervised semantic segmentation by iteratively mining common object features//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 1354-1362 [DOI: 10.1109/cvpr.2018.00147]

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: 12275-12284 [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/CVF Conference on Computer Vision and Pattern Recognition. Honolulu, USA: IEEE: 1568-1576 [DOI: 10.1109/cvpr.2017.687]

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: 16765-16774 [DOI: 10.1109/cvpr46437.2021.01649]

Wu Y C, Ye X C, Yang K Q, Li J D and Li X Q. 2024. DuPL: dual student with trustworthy progressive learning for robust weakly supervised semantic segmentation//Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 3534-3543 [DOI: 10.1109/cvpr52733.2024.00339]

Xu L, Bennamoun M, Boussaid F, Ouyang W L, Sohel F and Xu D. 2025. Auxiliary tasks enhanced dual-affinity learning for weakly supervised semantic segmentation. IEEE Transactions on Neural Networks and Learning Systems, 36(3): 5082-5096 [DOI: 10.1109/tnnls.2024.3373566]

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: 4310-4319 [DOI: 10.1109/cvpr52688.2022.00427]

Xu R T, Wang C W, Sun J X, Xu S B, Meng W L and Zhang X P. 2023. Self correspondence distillation for end-to-end weakly-supervised semantic segmentation//Proceedings of the 37th AAAI Conference on Artificial Intelligence. Washington, USA: AAAI: 3045-3053 [DOI: 10.1609/aaai.v37i3.25408]

Xuan G R, Zhang W and Chai P Q. 2001. EM algorithms of Gaussian mixture model and hidden Markov model//Proceedings of 2001 International Conference on Image Processing. Thessaloniki, Greece: IEEE: 145-148 [DOI: 10.1109/icip.2001.958974]

Yao Y Z, Chen T, Xie G S, Zhang C Y, Shen F M, Wu Q, Tang Z M and Zhang J. 2021. Non-salient region object mining for weakly supervised semantic segmentation//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 2623-2632 [DOI: 10.1109/cvpr46437.2021.00265]

Zhang B F, Xiao J M, Wei Y C, Sun M J and Huang K Z. 2020. 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 F, Gu C C, Zhang C Y and Dai Y C. 2021. Complementary patch for weakly supervised semantic segmentation//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision. Montreal, Canada: IEEE: 7242-7251 [DOI: 10.1109/iccv48922.2021.00715]

Zhang W K, Liu W J, Sun X, Xu G L and Fu K. 2022. Multi-source features adaptation fusion network for semantic segmentation in high-resolution remote sensing images. Journal of Image and Graphics, 27(8): 2516-2526 (张文凯, 刘文杰, 孙显, 许光銮, 付琨. 2022. 多源特征自适应融合网络的高分遥感影像语义分割. 中国图象图形学报, 27(8): 2516-2526) [DOI: 10.11834/jig.210054]

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

Zhou J H, Wei C, Wang H Y, Shen W, Xie C H, Yuille A and Kong T. 2022a. iBOT: image BERT pre-training with online tokenizer [EB/OL]. [2025-02-17]. https://arxiv.org/pdf/2111.07832.pdf

Zhou Y Y, Yun X Y, Chai C, Liu Z Y, Fan W X and Luo X. 2022b. Efficient textual explanations for complex road and traffic scenarios based on semantic segmentation [EB/OL]. [2025-02-17]. https://arxiv.org/pdf/2205.14118.pdf

作者简介

白雪飞，女，副教授，主要研究方向为图像处理和机器学习。E-mail：baixuefei@sxu.edu.cn

王文剑，通信作者，女，教授，主要研究方向为机器学习、计算智能和图像处理。E-mail：wjwang@sxu.edu.cn

王渊辉，男，硕士研究生，主要研究方向为图像处理和机器学习。E-mail：wangyuanhui2@sxu.edu.cn

许文杰，男，硕士研究生，主要研究方向为图像处理和机器学习。E-mail：202322409023@email.sxu.edu.cn

姜高霞，男，副教授，主要研究方向为机器学习和数据挖掘。E-mail：jianggaoxia@sxu.edu.cn