DOI:10.19651/j.cnki.emt.2416644

基于激活调制的双分支弱监督语义分割*

王家莉[1,2] 谭 棉[1,2] 冯夫健[1,2]
(1.贵州民族大学数据科学与信息工程学院 贵阳 550025;2.贵州省模式识别与人工智能系统重点实验室 贵阳 550025)

摘 要:图像级标注的语义分割因具有友好的注释和令人满意的性能而被广泛研究。针对类激活图激活区域稀疏、前背景间语义模糊问题,提出基于激活调制的双分支弱监督语义分割网络。该网络以 Resnet50 和 Vision Transformer 作为双分支特征提取网络,并设计激活调制模块嵌入卷积分支,该模块迫使模型激活中间分数的像素,生成紧凑的类激活图,从而缓解类激活图激活区域稀疏的问题。其次,提出基于余弦退火衰减的动态阈值调整策略,该策略在训练过程中自适应的确定背景最高阈值,使更多低置信前景像素参与到分割训练中,生成完整且准确的分割图。在 PASCAL VOC 2012 以及 MS COCO 2014 数据集上验证该网络的有效性。PASCAL VOC 2012 验证集和测试集上的 mIou 值分别为 74.2%和 74.0%,在 MS COCO 2014 验证集上的 mIou 值为 45.9%。实验结果表明,该网络可以解决前背景颜色相似场景下的误分割问题并取得优异的分割性能。

关键词:弱监督学习;语义分割;类激活图;激活调制;动态阈值
中图分类号:TP391.4;TN791 **文献标识码**:A **国家标准学科分类代码**:520.60

Dual branch weakly supervised semantic segmentation based on activation modulation

Wang Jiali[1,2] Tan Mian[1,2] Feng Fujian[1,2]
(1. College of Data Science and Information Engineering, Guizhou Minzu University, Guiyang 550025, China;
2. Key Laboratory of Pattern Recognition and Intelligent Systems of Guizhou Province, Guiyang 550025, China)

Abstract: Semantic segmentation with image-level annotation has been widely studied for its friendly annotation and satisfactory performance. Aiming at the problem of sparse activation regions and semantic ambiguity between foreground and background of class activation maps, a dual-branch weakly supervised semantic segmentation network based on activation modulation is proposed. The network uses Resnet50 and Vision Transformer as a two-branch feature extraction network, and designs an activation modulation module embedded in the convolutional branch, which forces the model to activate the intermediate fraction of pixels to generate a compact class activation map, thus alleviating the problem of sparse activation regions of class activation maps. Second, a dynamic threshold adjustment strategy based on cosine annealing decay is proposed, which adaptively determines the highest background threshold during the training process, so that more low-confidence foreground pixels are involved in the segmentation training, and complete and accurate segmentation maps are generated. The effectiveness of the network is verified on the PASCAL VOC 2012 as well as MS COCO 2014 datasets. mIou values are 74.2% and 74.0% on the PASCAL VOC 2012 validation and test sets, respectively, and 45.9% on the MS COCO 2014 validation set. The experimental results show that the network can solve the mis-segmentation problem and achieve excellent segmentation performance in scenes with similar front background colours.

Keywords: weakly supervised learning; semantic segmentation; class activation map; activation modulation; dynamic threshold

0 引 言

语义分割是计算机视觉中的一项重要任务,旨在为图像中的每一个像素分类,已被广泛应用于自动驾驶、医学成像分析和遥感图像解释[1]等应用场景。然而,语义分割所需的像素级注释是劳动密集以及耗时的,为了缓解全监督

收稿日期:2024-08-10
*基金项目:贵州省科技计划项目(QKHJCZK2022YB195,QKHJCZK2023YB143,QKHPTRCZCKJ2021007,QKHJCZK2022YB197)、贵州省教育厅自然科学研究项目(QJJ2022015,QJJ2023061,QJJ2023012,QJJ2022047,QJJ2024063)、贵州民族大学博士研究启动项目(GZMUZK[2024]QD04)资助

语义分割注释像素级标签所耗费的昂贵成本，研究人员开始研究使用图像级、点、涂鸦以及边界框 4 种相对较弱的标签来完成语义分割任务。其中，图像级注释是最弱的监督，它仅指示所关注的对象是否存在于图像中。图像级标签已广泛存在于像 ImageNet 这样的大规模数据集中，由于其可访问性，越来越多的学者专注于研究基于图像级标签的语义分割。本文工作也属于仅使用图像级标签的范畴。

类激活图（class activation map，CAM）的提出加速了弱监督语义分割的研究进程。CAM 由分类网络生成，通常仅响应最具辨别力的特征。然而，语义分割任务需要获取整个对象的边界信息。因此，分类与分割任务之间存在天然的鸿沟。为了弥补差距，学者们旨在研究能够生成紧凑的、边界平滑的伪标签。例如，Wang 等[2]针对 CAM 部分激活问题，设计了孪生网络 SEAM，该网络通过对输入图像进行仿射变换，设计等变交叉正则化损失迫使两个分支产生激活一致的类激活图。Jiang 等[3]基于局部图像块代替输入图像到分类网络中可以激活更多细节区域这一发现，设计了 L2G 网络，该网络利用全局网络在线学习局部图像块注意力图的补充知识，旨在生成高质量的注意力图，但该方法依赖离线显著性图作引导，存在分割图语义对象缺失的问题。Xu 等[4]以 CAM 作为自蒸馏对象，通过补充语义信息来增强网络特征学习的能力，并设计变异感知细化模块对伪标签进行细化，缓解了现有方法因缺乏全面语义信息提取而导致的伪标签质量低下的问题。Yang 等[5]为了解决弱监督语义分割获得完整语义区域难的问题，对输入图像进行小尺度、中尺度和大尺度变换，将其输入到权重共享的三分支网络，对不同尺度的 CAM 进行融合，使模型从多尺度图像中提取丰富的语义信息。Qin 等[6]认为卷积神经网络部分激活的原因在于模型训练过程中频繁激活特征，致使模型只保留有利于实现分类任务的特征。为此，设计了补偿分支，对两个分支生成的 CAM 执行加权融合。上述方法主要通过设计权重共享的多分支网络架构扩展 CAM 的激活区域，由于相同的底层操作，多个视图生成的 CAM 仍然相似，CAM 部分激活的问题没有得以解决。

Vision Transformer（Vit）作为第一个专门为计算机视觉任务设计的 Transformer 模型，最近在多个视觉任务上取得了性能突破。Vit 凭借多头注意力机制强大的远程上下文建模能力，在弱监督语义分割邻域取得一定的研究成果。例如，Ru 等[7]从多头注意力中学习语义亲和力来细化初始伪标签并结合低级图像外观信息设计像素自适应掩码细化模块。Ru 等[8]为了解决 Vit 中的过平滑问题，分别对 class token 和 patch token 设计对比损失。Kim 等[9]从聚类思想出发，设计聚类注意力模块与多头注意力机制并行嵌入到 Transformer 模块中，该模块对相似的 patch token 执行聚类运算，有助于模型均匀的激活整个对象区域。Xu 等[10]提出 MCTformer 为每一个类别添加可学习的 class token 学习特定类别的注意力图。虽然该方法生成了较优的伪掩码，但是该方法需要修改模型架构。上述方法为了从 CAM 中导出可靠的伪标签监督训练分割网络，需要设置两个背景阈值（β_h 和 β_l）对 CAM 中的像素划分为前景、背景以及不确定区域。为了保证前景的绝对置信，需要设置较高的阈值 β_h，这无疑丢失了部分分数低的置信前景像素。换言之，这部分像素在分割训练中没有得到学习，导致前背景或不同对象间语义模糊问题。

综上所述，尽管基于图像级语义分割任务的研究已取得一定的成果，但相关工作大多侧重研究如何扩展 CAM 的激活区域，从而缩小分类与分割任务之间的差距。上述方法存在不同对象间语义模糊的问题，特别是在前背景图像颜色相似的场景下，这严重削弱了弱监督语义分割的性能。

受协同训练范式的启发，本文提出基于激活调制的双分支弱监督语义分割网络。针对 CAM 部分激活问题，设计以 Resnet50 和 Vit 为骨干的双分支网络，并提出激活调制模块嵌入至卷积分支。该模块分别从通道和空间维度调整特征图的激活分布，迫使模型激活更多中间分数的像素，同时抑制最敏感和最不敏感像素的激活。此外，为了从 CAM 中导出具有更多置信前景像素的伪标签参与到后续的分割训练中，针对前背景语义模糊问题，提出了余弦退火衰减动态阈值调整策略。该策略在训练过程中自适应的确定 β_h，随着迭代次数的增加，背景最高阈值 β_h 以余弦函数的形式逐渐下降，最后稳定在最优阈值。

本文贡献点总结如下：

1）针对 CAM 部分激活问题，设计不同架构的双分支网络，并为卷积分支嵌入激活调制模块。该模块通过调整特征图的激活分布，迫使网络关注具有中间分数的 CAM；

2）针对前背景语义模糊的问题，设计基于余弦退火衰减的动态阈值调整策略。该策略能够让更多置信前景像素参与到后续的分割训练中，缓解了前背景间颜色相似时的误分割问题。

1　基于激活调制的双分支弱监督语义分割网络

1.1　双分支架构

目前，基于多分支的弱监督语义分割方法，大多通过设计权重共享的多分支网络缓解 CAM 部分激活的问题。然而，由于相同的底层操作，模型初始化参数相同，浅层网络学习到的颜色、纹理以及边界信息倾向于同质化，不同视图生成的 CAM 是相似的，CAM 仅响应辨别力特征的问题没有得到解决。为此，受协同训练范式的启发，本文设计了不同网络架构的双分支网络，整体网络架构如图 1 所示。具体地，设计以 Resnet50 和 Vit 为分支的特征提取网络。尽管 Vit 具有强大的远程建模能力，但它存在过平滑问题。遵循 ToCo[8]方法，在第 10 个 Transformer 模块添加辅助分类层，从辅助分类层生成辅助 CAM，并导出伪标签监督训练 Vit 分支最终的 CAM。为了使两个分支生成激活一

致的 CAM,设计了语义一致正则化损失,其表达式如下:

$$\ell_{er} = \lambda_1 \| \boldsymbol{M}^{CNN} - \boldsymbol{M}^{Aux} \|_1 + \lambda_2 \| \boldsymbol{M}^{CNN} - \boldsymbol{M}^{Vit} \|_1 \quad (1)$$

其中,$\boldsymbol{M}^{CNN}$ 表示由卷积分支生成的 CAM,$\boldsymbol{M}^{Aux}$ 表示在 Vit 中辅助分类层生成的 CAM,$\boldsymbol{M}^{Vit}$ 表示 Vit 分支最后层 Transformer 模块的 patch token 生成的 CAM,λ_1 和 λ_2 表示两个超参数,$\|.\|_1$ 表示 L_1 正则化。

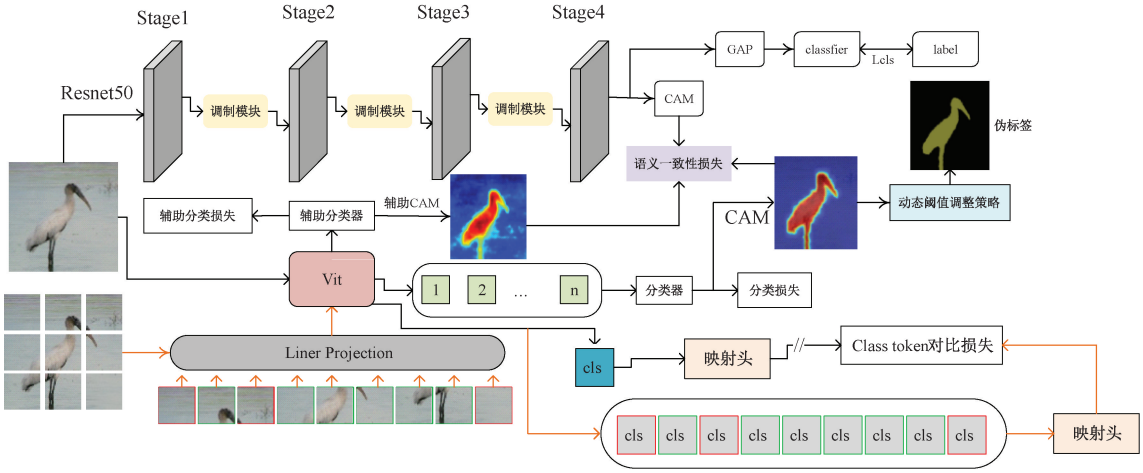


图 1 激活调制双分支网络架构

Fig. 1 Activation modulated two-branch network architecture

1.2 激活调制模块

激活调制模块利用调制函数从通道-空间维度调整特征图的激活分布,使网络能够关注次要的目标区域,激活更多目标对象的响应。调制模块如图 2 所示。

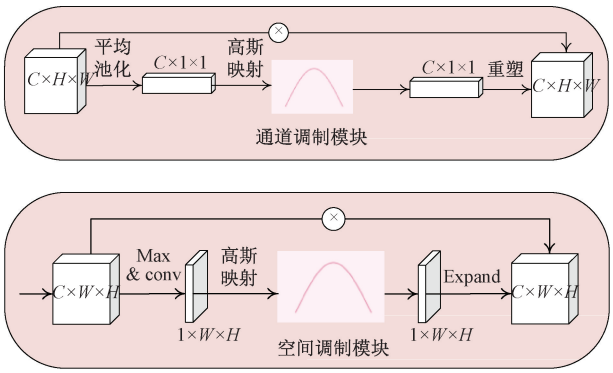


图 2 激活调制模块

Fig. 2 Activating the modulation module

本文以 Resnet50 作为卷积分支,将激活调制模块嵌入至卷积分支的 Stage1,Stage2,Stage3 后调整特征图的激活分布。具体而言,首先,通道激活调制模块通过池化层和卷积层显示建模,将残差块的特征图 $\boldsymbol{F}_i \in \mathbb{R}^{C\times H\times W}(i=1,2,3)$ 执行全局平均池化,获得全局向量 $\boldsymbol{z}^i \in \mathbb{R}^{C\times 1\times 1}$,对 $\boldsymbol{z}^i$ 求均值和方差。其次,将均值和方差输入到高斯调制函数对特征分布进行调整。将调制后的特征分布扩展到特征图 $\boldsymbol{F}^i$ 相同的形状大小 $\boldsymbol{F}^i_{out}$。上述过程可以表述为:

$$\boldsymbol{F}^i_{out} = R(g(conv(avg(\boldsymbol{F}_i)))) \quad (2)$$

其中,$R(.)$ 表示扩展操作,$g(.)$ 表示调制函数,将在下一节详细介绍。$conv(.)$ 表示卷积操作,$avg(.)$ 表示全局平均池化层。最后,将调制后的特征图 $\boldsymbol{F}^i_{out}$ 与调制前的特征图 $\boldsymbol{F}_i$ 执行逐像素相乘生成调制后的特征图。上述过程可以表述为:

$$\boldsymbol{A}_i = \boldsymbol{F}^i_{out} \otimes \boldsymbol{F}_i \quad (3)$$

其中,$\boldsymbol{A}_i$ 表示调制后的特征图,$\otimes$ 表示逐像素相乘。

空间激活调制模块与通道激活调制模块执行级联操作。换言之,通道激活调制模块的输出作为空间激活调制模块的输入。首先,对张量 $\boldsymbol{A}_i \in \mathbb{R}^{C\times H\times W}$ 在通道维度上执行 max 运算和 mean 运算,将获得的张量沿着通道维度进行拼接。其次,将拼接后的张量输入到一个卷积核大小为 7×7 的卷积层,将输出的张量 $\boldsymbol{V} \in \mathbb{R}^{1\times H\times W}$ 输入到高斯调制函数中,调整特征图的激活分布。最后对调制后的特征图和通道激活调制模块的输出进行逐像素相乘融合特征。上述过程可以表述为:

$$\boldsymbol{F}^{s_i}_{out} = R(g(conv^{7\times 7}(cat(\max(\boldsymbol{A}_i)), mean(\boldsymbol{A}_i)))) \otimes \boldsymbol{A}_i \quad (4)$$

其中,$conv^{7\times 7}$ 表示卷积核大小为 7×7 的卷积层,$cat(.)$ 表示沿着通道维度拼接,$\max(.)$ 表示取最大值运算,$mean(.)$ 表示求均值运算。

1.3 调制函数

激活调制模块应用调制函数重新分配特征图的激活分布,使模型关注具有中间激活值的特征,并抑制最敏感和最不敏感像素的激活。其中,最敏感像素对应于 CAM 中最具辨别力的特征,最不敏感像素是背景。

$$V_A = g(V_{A_f}) \quad (5)$$

其中,V_{A_f} 表示特征图的激活值,V_A 表示调制后的特征图激活值,$g(.)$ 表示高斯函数,服从标准正态分布。高斯分布的均值 μ 和方差 σ 分别由 V_{A_f} 计算:

$$\mu = \frac{1}{M}\sum_{i=1}^{M}(V^i_{A_f}) \quad (6)$$

$$\sigma = \sqrt{\frac{1}{M}\sum_{i=1}^{M}(V_{A_f}^{i} - \mu)^2} \tag{7}$$

其中，M 表示特征图中的所有激活值，本文按照 μ 和 σ 的设置来重新映射激活值的分布。

1.4　基于余弦退火衰减的动态阈值调整策略

前背景图像颜色相似场景下的误分割问题归因为语义对象模糊问题，其本质是模型在训练过程中缺乏对该像素的学习所导致的。

基于 Transformer 的弱监督语义分割方法，通过对 β_h 设置较高的阈值，从 CAM 中导出绝对置信的伪标签。然而，这种设置硬阈值的方法无疑过滤掉部分低置信绝对前景像素，被过滤掉的像素无法参与到后续的分割训练中。换言之，这部分像素没有得到充分学习，致使最终的分割结果在前背景间语义模糊，尤其是当图像的前背景颜色相似时出现误分割问题。为此，本文旨在设计一种动态阈值调整方法，在模型训练过程中，自适应的确定 β_h，使更多低置信前景像素参与到后续分割训练中，从而生成密集的分割图。

受余弦退火学习率衰减的启发，提出余弦退火衰减的动态阈值调整策略。该策略在训练过程中遵循余弦函数图像调整最高阈值，模型训练初期设置较高的阈值，有助于模型跳出局部最优解，并在每次迭代训练过程中使用该策略寻找全局最优阈值。余弦退火衰减动态阈值调整策略可以表述为：

$$th(t) = th(0) - \frac{1}{2}(th(0) - th(T))(1 - \cos(\frac{t\pi}{T})) \tag{8}$$

其中，t 表示当前迭代，T 表示训练迭代总数。

1.5　损失函数

图像级标签 $y_c \in R^{C\times 1}$ 是网络唯一可以使用的监督，在卷积分支末端采用全局平均池化层获得分类预测向量 $\boldsymbol{Z}$。同时，在 Vit 的辅助分类层以及网络末端添加全局最大池化层获得分类预测向量。分类损失采用多标签软边缘损失[11]：

$$\ell_{cls}(\boldsymbol{Z}, y_c) = -\frac{1}{C-1}\sum_{c=1}^{C}[y_c \log(\frac{1}{1+e^{-z_c}}) + (1-y_c)\log(\frac{e^{-z_c}}{1+e^{-z_c}})] \tag{9}$$

形式上，双分支网络的分类损失为卷积分支分类损失、Vit 分支辅助分类层分类损失以及 Vit 分支的分类损失的线性组合。

$$\ell_{cls} = \ell_{cls}(\boldsymbol{Z}_{CNN}, y_c) + \ell_{cls}(\boldsymbol{Z}_{Aux}, y_c) + \ell_{cls}(\boldsymbol{Z}_{Vit}, y_c) \tag{10}$$

其中，$\boldsymbol{Z}_{CNN}$ 表示卷积分支的预测分类向量，$\boldsymbol{Z}_{Aux}$ 表示 Vit 辅助分类层的预测分类向量，$\boldsymbol{Z}_{Vit}$ 表示 Vit 分支分类层的预测分类向量。

为了防止 Vit 中的过平滑问题对分割性能的削弱，本文遵循文献[8]的损失函数设计，分别引入 patch token 对比损失以及 class token 对比损失。其定义如下：

$$\ell_{ptc} = \frac{1}{N^+}\sum_{y_i=y_j}(1-\cos sim(\boldsymbol{F}_i, \boldsymbol{F}_i)) + \frac{1}{N^-}\sum_{y_i\neq y_i}\cos sim(\boldsymbol{F}_i, \boldsymbol{F}_j) \tag{11}$$

$$\ell_{ctc} = \frac{1}{N^+}\sum_{q^+}\log\frac{e^{(p^{\mathrm{T}}q^+/\tau)}}{e^{(p^{\mathrm{T}}q^+/\tau)} + \sum_{q^-}e^{(p^{\mathrm{T}}q^-/\tau)+\varepsilon}} \tag{12}$$

其中，$\cos sim(.)$ 计算余弦相似度，N^+ 表示正样本的数量，N^- 表示负样本的数量，τ 表示温度因子，ε 是一个很小的常数，$q^+ \in Q^+$，$q^- \in Q^-$，Q^+ 表示从不确定区域裁剪局部图像的 class token 的集合，Q^- 表示从背景区域裁剪局部图像的 class token 的集合。

对于单阶段弱监督语义分割，本文对模型获得的伪标签采用像素自适应细化模块[7]进行细化。分割损失采用交叉熵损失 ℓ_{seg}。训练整个网络的总损失定义为：

$$\ell_{total} = \ell_{cls} + \ell_{er} + w_1\ell_{ptc} + w_2\ell_{ctc} + w_3\ell_{seg} \tag{13}$$

其中，w_1，w_2，w_3 表示 3 个超参数，用来重新调整损失项的权重。

2　实验结果与分析

2.1　实验设置

1)数据集

本文在两个弱监督语义分割标准数据集上进行实验。PASCAL VOC 2012 数据集由 20 个语义类别及 1 个背景类别组成。其中，训练集、验证集以及测试集分别包含 1 464、1 449 和 1 456 张图像。遵循大多数工作的做法，采用增强训练集 SBD 的 10 582 张图像训练模型。由于 PASCAL VOC 2012 数据集不公开测试集真值，本文同样将测试集结果提交到官方评估服务器评估其性能。MS COCO 2014 数据集由 80 个语义类别及 1 个背景类别组成。遵循文献[12]的做法，没有目标类别的图像从数据集中剔除，剩下的 82 081 张图像作为训练集，40 137 张图像作为验证集。

2)评价指标

为了评估和对比各种方法的性能，采用平均交并比(mean intersection over union，MIOU)作为评价指标，其表达式如下：

$$MIOU = \frac{1}{C+1}\sum_{i=0}^{C}\frac{P_{ii}}{\sum_{i=0}^{C}P_{ij} + \sum_{i=0}^{C}P_{ji} - P_{ii}} \tag{14}$$

其中，C 表示类别数，$C+1$ 表示加上背景类，i 表示真实值，j 表示预测值，P_{ij} 表示把第 i 类像素预测为第 j 类像素的像素个数。

3)实施细节

在训练阶段，Vit 分支采用 AdamW 优化器，预热 1 500 次迭代，预热率和权重衰减率分别设置为 1×10^{-6} 和 0.9，迭代训练学习率为 6×10^{-5}，并采用多项式调度器

进行衰减。卷积分支采用 SGD 优化器,学习率和权重衰减分别设置为 0.001 和 1×10^{-4},动量大小设置为 0.9。对于 PASCAL VOC 2012 数据集的实验,批量大小和迭代次数分别设置为 4 和 20K,采用多裁剪数据增强裁剪出全局图像和局部图像,其图像大小分别为 448×448 和 96×96。背景阈值 $(th(l),th(0),th(T))$ 设置为(0.25,0.7,0.55)。式(1)中的 λ_1 和 λ_2 设置为 0.1,式(12)中的温度因子 τ 遵循 ToCo 方法,设置为 0.5,式(13)中的 w_1,w_2 和 w_3 遵循 ToCo 相同的设置,分别设置为 0.2、0.5 和 0.1。

MS COCO 2014 数据集上的实验,批量大小设置为 8,迭代 80 K 次,背景阈值 $(th(l),th(0),th(T))$ 设置为(0.25,0.65,0.55),其余设置与 PASCAL VOC 2012 相同。

分割解码器由两个卷积核大小为 3×3 的卷积层和一个卷积核大小为 1×1 卷积层组成,使用简单的 MLP 层进行预测。在推理阶段,遵循语义分割[13]常见做法,采用多尺度测试和条件随机场进行后处理。本文采用 Pytorch 深度学习框架,所有实验均部署在两张显存大小为 40 G 的 A100 显卡上。

2.2 消融实验与分析

本节的消融实验旨在研究双分支架构设计、通道调制模块、空间调制模块、动态阈值调整策略以及它们的组合在分割中的有效性。在 PASCAL VOC 2012 数据集的验证集上进行一系列实验,结果如表 1 所示。所有消融实验均没有采用条件随机场进行细化,最优性能均用加粗字体表示。

表 1 PASCAL VOC 2012 数据集验证集上分析各组件有效性

Table 1 Component validity analysed on the validation set of the PASCAL VOC 2012

基线	双分支	动态阈值	通道调制	空间调制	Seg/%
√					67.8
√	√				70.9
√	√	√			70.8
√	√		√		69.6
√	√			√	70.4
√	√		√	√	72.2
√	√	√	√		71.4
√	√	√		√	72.1
√	√	√	√	√	73.3

1)各组件有效性分析

由表 1 可知,仅采用双分支网络架构,相比基线模型具有 3.1% 的性能增益,显著提升模型分割性能。实验结果表明,设计不同架构的双分支网络有利于扩展 CAM 激活区域。当将设计的激活调制模块(通道调制模块+激活调制模块)嵌入到卷积分支时,分割结果达到 72.2%,与基线相比具有 4.4% 的性能增益。实验结果表明,激活调制模块可以促进模型激活具有中间分数的像素,从而迫使模型激活更多目标像素。表 1 中还报告了激活调制模块的子集,仅加入通道调制模块时,分割性能为 69.6%,与基线相比性能增益为 1.8%;仅加入空间调制模块时,可以获得 70.4% 的分割性能,性能增益为 2.6%。激活调制模块的子集实验表明,无论加入通道调制模块还是空间调制模块,都有助于提高基线的分割性能。此外,当仅对双分支网络加入基于余弦退火衰减动态阈值调整策略时,模型具有 70.8% 的分割性能,与基线的性能增益为 3%。实验结果表明,动态阈值调整策略可以使更多置信前景像素参与到分割训练中,缓解前背景语义模糊问题。当对模型采用完整的模块以及策略时,分割性能进一步提升,获得最佳分割性能 73.3%,与基线模型相比具有 5.5% 的性能增益。实验结果表明,本文提出的网络可以生成完整的分割图。

2)双分支架构设计的有效性分析

为了验证本文设计的双分支架构的有效性,采用过度激活率(over-activation rate,OR)作为评价指标。其中,过度激活率定义为:$OR = FP/(TP+FP)$,FP 表示伪标签假阳性像素的数量,TP 表示伪标签真阳性像素的数量。OR 率越高,说明模型激活更多非目标类别的像素,低质量的 CAM 使导出的伪标签性能低下,最终导致分割图质量下滑。基线与双分支架构成比较如图 3 所示。由图 3 可知,基线在一些类别存在较高的过度激活率,从而导致了较低的分割性能。相反,本文设计的双分支架构将基线多个类别的过度激活率降低了 15% 以上,甚至在“cow”、“person”类别的过度激活率仅在 5% 以下。实验结果表明,设计不同架构的双分支架构网络有效抑制了基线模型的过度激活率,使 CAM 激活更多目标像素。

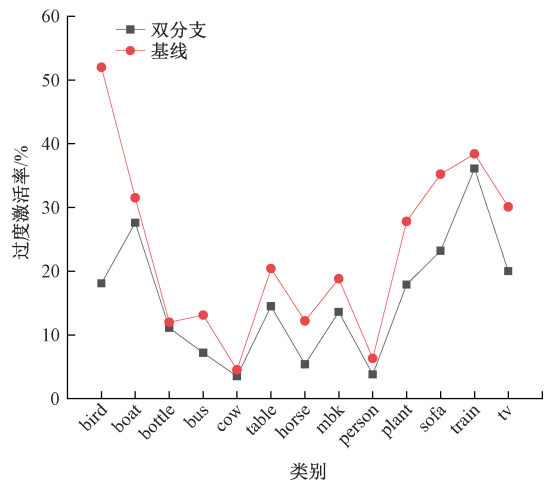


图 3 基线与双分支架构比较

Fig. 3 Baseline vs. two-branch architecture comparison

3)语义一致正则化方法有效性分析

在双分支架构设计中,为了使卷积分支与 Vit 分支生成激活一致的 CAM,本文设计了语义一致正则化损失。表 2 展示了不同正则化方法带来的分割性能,Seg 表示分割精度。采用 L_1 正则化作为语义一致正则化损失函数时,具有 73.3% 的分割性能,分析主要原因是 L_1 正则化的稀疏特性,从而使 CAM 仅激活特定类别的像素,同时抑制背景噪声。当采用 L_2 正则化作为损失函数时,分割性能为 72.6%,实验结果稍逊色 L_1 正则化,分析主要原因是 L_2 正则化的平滑特性所致,不可避免的激活背景像素,从而相对降低了分割性能。为此,本文采用 L_1 正则化作为训练双分支网络架构的语义一致正则化损失。

表 2　不同语义正则化方法分析

Table 2　Analysis of different semantic regularization methods　%

正则化方法	CAM	Aux_cam	Seg
L_1 正则化	73.5	69.0	73.3
L_2 正则化	74.1	69.5	72.6

4)动态阈值调整策略阈值分析

在激活调制双分支弱监督语义分割网络中,$th(t)$ 是一个动态背景阈值,在训练过程中逐渐降低到 $th(T)$。动态阈值调整策略旨在让更多置信前景像素参与到分割头的训练,生成准确且边界平滑的分割图。表 3 展示了不同的 $th(t)$ 阈值设置下的分割结果 CAM 为类激活图,Aux_cam 为辅助类激活图。由表 3 可知,当阈值设置为 0.55 时,模型取得最佳性能 73.3%。当阈值降低到 0.5 时,模型性能开始下降,这是因为设置较低的背景阈值,使得噪声像素也被视为置信前景像素,在分割训练中,降低了分割图的质量。

表 3　余弦衰减策略最高背景阈值设置

Table 3　Maximum background threshold settings for cosine attenuation strategy　%

阈值	CAM	Aux_cam	Seg
0.65	72.8	68.4	71.2
0.60	72.4	67.6	71.4
0.55	73.5	69.0	**73.3**
0.50	73.2	69.4	72.1

5)模型复杂性分析

表 4 展示了基线模型和本文方法模型的计算量(Flops)以及参数量(Params)。本文设计的网络架构在计算量和参数量上分别低于基线模型 4.1 G 和 5.3 M,分割性能却达到了 73.3%。实验结果表明,本文所设计的激活调制双分支网络架构对分割任务的有效性。

表 4　模型复杂性分析

Table 4　Analysis of model complexity

方法	Flops/G	Params/M	Seg/%
基线	71.8	91.6	67.8
本文	67.7	86.3	73.3

6)激活调制模块组合方法分析

为了分析通道调制模块和空间调制模块组合方法的有效性,本文分别对通道-空间调制模型采用串行和并行的组合方法。表 5 展示了不同组合方法的分割性能。当采用并行结构时,获得的分割性能仅有 71.4%。然而,对通道-空间调制模块采用串行组合时,分割性能为 73.3%。与并行结构相比,具有 1.9% 的性能增益。实验结果表明,采用串行的组合方法有利于模型提取丰富的特征信息,使模型更加关注具有中间分数的像素。为此,本文在通道调制模块后级联空间调制模块。

表 5　激活调制模块不同组合方法分析

Table 5　Analysis of different combinations of activation modulation module methods　%

组合方法	CAM	Aux_cam	Seg
并行	72.3	67.1	71.4
串行	**73.5**	**69.0**	**73.3**

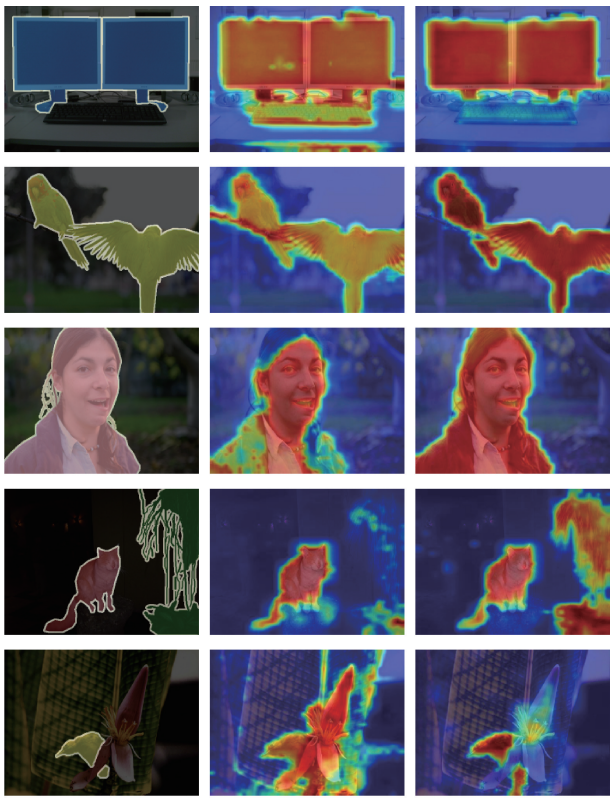
2.3　与先进算法的比较

1) CAM 及伪标签的对比

图 4 所示为可视化本文网络生成的类激活图,与采用相同网络架构的 ToCo 方法相比,本文方法生成的 CAM 更完整且准确,解决了 ToCo 的过度激活问题。值得注意的是,第 3 行图像由于人的“头发”与背景极度相似,导致 ToCo 在激活“person”这个对象时,把“头发”视为背景类对待而出现误分割问题,最终导致分割图分割不完整。相反,本文设计的网络成功的响应了整个对象区域。与 ToCo 相比,完整的激活目标对象并恢复了 ToCo 方法丢失的语义对象,例如第 4 行。此外,表 6 定量比较了现有弱监督语义分割方法生成的伪标签性能(前 3 行为两阶段弱监督语义分割,后 4 行为单阶段弱监督语义分割),Train 表示训练集分割精度;Val 表示验证集分割精度。在单阶段方法中,伪标签直接由 CAM 生成,而两阶段方法需要对初始 CAM 进行细化后产生最终的伪标签。由表 6 可知,本文方法无论在 PASCAL VOC 2012 数据集的验证集还是测试集都获得了最优性能,在验证集和测试集上分别达到 73.5%和 74.9%的伪标签性能。实验结果证明了设计的激活调制模块和动态阈值调整策略可以扩展 CAM 的激活区域并缓解前背景像素语义模糊问题。

2)最终的分割结果对比

表 7 展示了 MS COCO 2014 数据集验证集上的分割



(a) 图像&真值 (a) Images & Ground truth　(b) ToCo方法 (b) The ToCo method　(c) 本文 (c) Ours

图 4　可视化对比类激活图

Fig. 4　Visualisation comparison CAM

表 6　PASCAL VOC 2012 数据集伪标签性能比较

Table 6　Comparison of pseudo-labelling performance on the PASCAL VOC 2012

方法	基线模型	Train/%	Val/%
MCTformer[10] CVPR'22	V1_WR38	69.1	_
RCA[14] CVPR'22	V2_VGG16	71.4	_
ACR[15] CVPR'23	V2_WR38	72.3	_
AFA[7] CVPR'22	MiT_B1	68.7	66.5
ToCo[8] CVPR'23	ViT_B	72.2	70.5
ClusterCAM[9] Access'24	Deit_S	_	71.4
本文	ViT_B	**74.9**	**73.5**

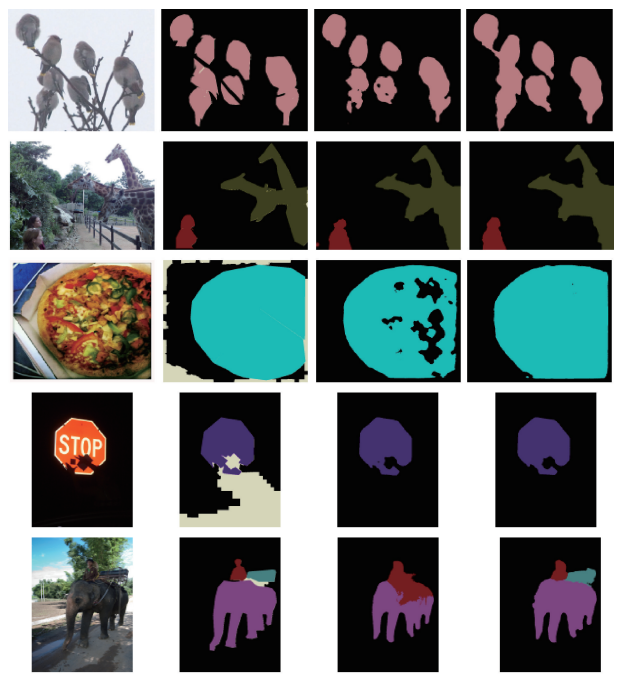
性能(前 10 行为两阶段弱监督语义分割,后 6 行为单阶段弱监督语义分割),MS COCO 2014 场景更复杂,对弱监督语义分割更具有挑战性。由表 7 可知,本文方法在该数据集上获得最优分割性能 45.9%,与两阶段弱监督语义分割方法 LPCAM 相比,具有 0.6% 的性能增益。与采用相同骨干模型 ToCo 相比,具有 3.6% 的性能增益。

图 5 所示为定性比较 ToCo 方法和本文方法在 MS COCO 2014 验证集上的分割结果。实验结果表明,即使在

表 7　MS COCO2014 数据集分割性能比较

Table 7　Comparison of segmentation performance on MS COCO2014

方法	监督方式	基线模型	COCO Val/%
SEAM[2] CVPR'20	I	V1_WR38	31.9
L2G[3] CVPR'22	I+S	V2_R101	44.2
MCTformer[10] CVPR'22	I	V1_WR38	42.0
ACR[15] CVPR'23	I	V2_WR38	45.3
AdvCAM[16] CVPR'21	I	V2_R101	44.4
URN[17] AAAI'22	I	V2_R101	40.7
ESOL[18] NeurIPS'22	I	V2_R101	42.6
BECO[19] CVPR'23	I	V3_R101	45.1
LPCAM[20] CVPR'23	I	V1_WR38	45.5
OCR[21] CVPR'23	I	V1_WR38	42.5
SCD[4] AAAI'23	I	MiT_B1	40.1
AFA[7] CVPR'22	I	MiT_B1	38.9
ToCo[8] CVPR'23	I	ViT_B	42.3
ClusterCAM[9] Access'24	I	Deit_S	41.8
SLRNet[22] IJCV'22	I	WR38	35.0
本文	I	ViT_B	**45.9**



(a) 图像 (a) Images　(b) 真值 (b) Ground truth　(c) ToCo方法 (c) ToCo method　(d) 本文 (d) Ours

图 5　MS COCO 2014 数据集分割结果

Fig. 5　The MS COCO 2014 dataset segmentation results

前背景颜色相似的场景下,本文方法的分割结果仍然具有连续性,如第 3 行。与 ToCo 相比,获得的分割结果更加完整。如第 5 行,恢复了 ToCo 丢失的目标对象。

表 8 展示了在 PASCAL VOC 2012 数据集两阶段弱监督语义分割方法以及单阶段弱监督语义分割方法的分割性能(前 11 行为两阶段弱监督语义分割,后 6 行为单阶段弱监督语义分割),Test 表示测试集分割精度。由表 8 可知,本文方法的分割结果获得先进性能,甚至优于采用显著性图作为额外监督的两阶段方法。与采用相同骨干网络的 ToCo 方法相比,验证集和测试集分别获得 3.1% 和 1.8% 的性能增益。本文方法获得优异的分割性能,得益于设计的双分支网络架构,该架构充分利用卷积和 Transformer 模型的优势,生成完整的分割图。

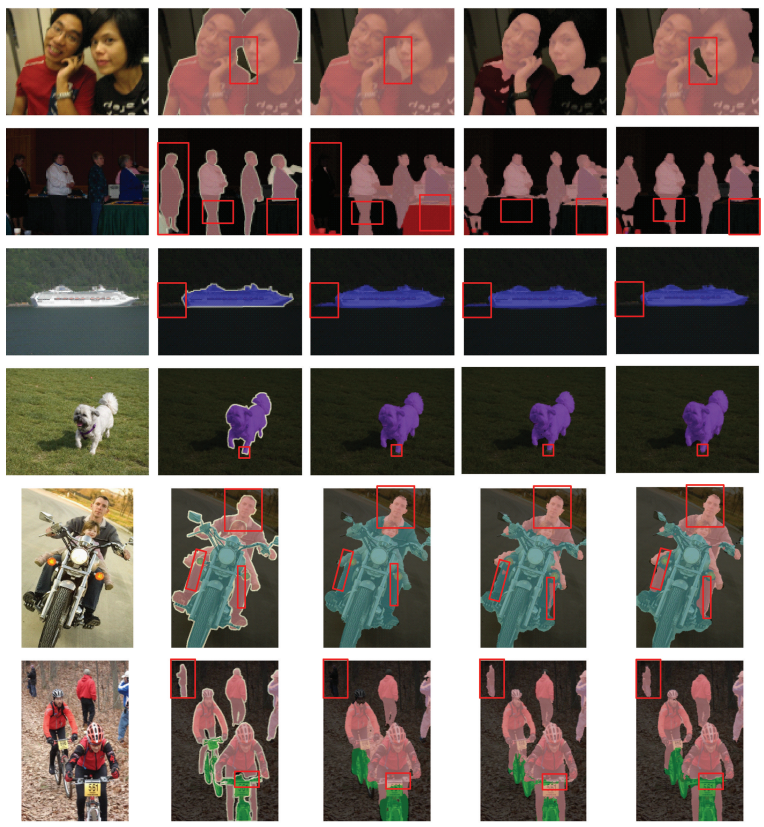
图 6 所示为定性比较 AFA[7]、ToCo 方法以及本文方法的分割结果。由可视化结果可知,AFA 存在过分割问题,例如第 3 行。值得注意的是,由前背景颜色相似导致的误分割问题影响最终的分割性能,例如第 1、2 行。实验结果表明,本文所设计的方法能够解决前背景颜色相似时的误分割问题,生成较完整的分割图。报告测试集详细的分割结果。

本文方法与现有方法在 PASCAL VOC 2012 数据集验证集上进行详细的分割精度对比,结果如表 9 所示。由表 9 中结果可知,在“bottle”、“cat”、“chair”、“cow”、“table”、“dog”等多个类别获得最优分割性能,证明了所提方法的有效性。

表 8　PASCAL VOC 2012 数据集分割性能比较

Table 8　Comparison of segmentation performance on PASCAL VOC 2012

方法	监督方式	基线模型	VOC	
			Val/%	Test/%
SEAM[2] CVPR'20	I	V1_WR38	64.5	65.7
L2G[3] CVPR'22	I+S	V2_R101	72.1	73.0
MCTformer[10] CVPR'22	I	V1_WR38	71.9	71.6
RCA[14] CVPR'22	I+S	V2_R101	72.2	72.8
ACR[15] CVPR'23	I	V2_WR38	71.9	71.9
AdvCAM[16] CVPR'21	I	V2_R101	68.1	68.0
URN[17] AAAI'22	I	V2_R101	69.5	69.7
ESOL[18] NeurIPS'22	I	V2_R101	69.9	69.3
BECO[19] CVPR'23	I	V3_R101	72.1	71.8
LPCAM[20] CVPR'23	I	V1_WR38	72.6	72.4
OCR[21] CVPR'23	I	V1_WR38	72.7	72.0
SCD[4] AAAI'23	I	MiT_B1	67.3	67.5
AFA[7] CVPR'22	I	MiT_B1	66.0	66.3
ToCo[8] CVPR'23	I	ViT_B	71.1	72.2
ClusterCAM[9] Access'24	I	Deit_S	70.3	70.7
SLRNe[22] IJCV'22	I	WR38	67.2	67.6
本文	I	ViT_B	**74.2**	**74.0**



(a) 图像 (a) Images　(b) 真值 (b) Ground truth　(c) AFA方法 (c) The AFA method　(d) ToCo方法 (d) ToCo method　(e) 本文 (e) Ours

图 6　可视化对比 PASCAL VOC 2012 分割结果

Fig. 6　Visualisation comparison of PASCAL VOC 2012 segmentation results

表 9 不同算法在 PASCAL VOC 2012 验证集上的分割精度

Table 9 Segmentation accuracy of different algorithms on the PASCAL VOC 2012 validation set %

方法	bkg	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow
SEAM	88.8	68.5	33.3	**85.7**	40.4	67.3	78.9	76.3	81.9	29.1	75.7
EPS	91.7	**89.4**	40.6	84.7	**67.0**	71.6	**87.8**	**82.7**	87.4	33.6	81.9
文献[23]	90.3	79.3	35.2	83.6	63.6	71.9	80.3	74.9	84.5	30.5	79.8
SEC	82.4	62.9	26.4	61.6	27.6	38.1	66.6	62.7	75.2	22.1	53.5
MCOF	87.0	78.4	29.4	68	44.0	67.3	80.3	74.1	82.2	21.1	70.7
SBCN	90.2	78.5	34.0	83.9	62.7	71.9	77.5	73.8	84.5	29.0	77.0
DSCP	88.9	77.6	31.3	73.2	59.8	71.0	79.2	74.5	80	15.1	73.3
CIAN	88.2	79.5	32.6	75.7	56.8	72.1	85.3	72.9	81.7	27.6	73.3
本文	**92.2**	83.3	**44.9**	84.7	63.5	**78.6**	84.8	80.5	**88.7**	**37.6**	**87.5**
方法	table	dog	horse	mbk	person	plant	sheep	sofa	train	tv	mIou
SEAM	48.1	79.9	73.8	71.4	75.2	48.9	79.8	40.9	68.1	51.6	64.5
EPS	37.3	82.5	82.9	**76.6**	82.8	54.0	79.7	39.1	**85.4**	51.7	71.0
文献[23]	52.8	69.2	76.2	66.3	81.9	47.6	70.6	37.9	63.1	63.3	66.8
SEC	28.3	65.8	57.8	62.3	52.5	32.5	62.6	32.1	45.4	45.3	50.7
MCOF	28.2	73.2	71.5	67.2	53.0	47.7	74.5	32.4	71.0	45.8	60.3
SBCN	51.5	66.7	75.9	65.7	80.9	45.1	68.6	37.8	63.5	61.0	65.7
DSCP	10.2	76.1	72.2	69.1	72.1	39.9	73.9	14.6	70.3	53.1	60.8
CIAN	39.8	76.4	77.0	74.9	66.8	46.6	81.0	29.1	60.4	53.3	64.3
本文	**55.8**	**83.9**	**86.7**	74.7	**83.0**	**65.9**	**89.7**	**59.7**	64.1	**67.8**	**74.2**

3 结 论

针对类激活图激活区域稀疏问题,设计了不同网络架构的双分支特征提取网络,并为卷积分支设计激活调制模块,该模块抑制卷积分支最敏感和最不敏感像素的激活,迫使模型激活具有中间分数的像素,解决了 CAM 部分激活的问题。针对前背景语义模糊问题,特别是在前背景颜色相似的场景下,本文设计基于余弦退火衰减动态阈值调整策略,在模型训练过程中,自适应调整背景最高阈值,使得更多低分数的置信前景像素参与后续的分割训练中,生成完整的分割图。基于上述策略,本文提出基于激活调制的双分支弱监督语义分割网络,在 PASCAL VOC 2012 以及 MS COCO 2014 数据集上验证该方法的有效性和可行性。实验结果表明,本文方法取得了先进分割性能,能够生成准确、完整且紧凑的分割图。

未来工作将致力于研究不依赖于类激活图的伪标签导出方法,类激活从分类网络中产生存在固有的缺陷,不利于语义分割这种密集下游任务。此外,以在线学习的方式重新学习伪标签,并校正语义模糊的像素,有助于提高伪标签的质量,进而提升弱监督语义分割的性能。

参考文献

[1] HOSSAIN M D, CHEN D M. Segmentation for object based image analysis: A review of algorithms and challenges from remote sensing perspective[J]. ISPRS Journal of Photogrammetry and Remote Sensing, 2019, 150: 115-134.

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

[3] JIANG P T, YANG Y Q, HOU Q B, et al. L2g: A simple local-to-global knowledge transfer framework for weakly supervised semantic segmentation[C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022: 16886-16896.

[4] XU R T, WANG CH W, SUN J X, et al. Self correspondence distillation for end-to-end weakly-supervised semantic segmentation[C]. AAAI Conference on Artificial Intelligence, 2023, 37(3): 3045-3053.

[5] YANG G Q, ZHU CH, ZHANG Y. A self-training framework based on multi-scale attention fusion for weakly supervised semantic segmentation[C]. 2023 IEEE International Conference on Multimedia and Expo(ICME), 2023: 876-881.

[6] QIN J, WU J, XIAO X F, et al. Activation modulation and recalibration scheme for weakly supervised semantic segmentation[C]. AAAI Conference on Artificial Intelligence, 2022, 36(2): 2117-2125.

[7] RU L X, ZHAN Y B, YU B SH, et al. Learning affinity from attention: End-to-end weakly-supervised

semantic segmentation with transformers[C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022: 16846-16855.

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

[9] KIM Y W, KIM W. Clustering-guided class activation for weakly supervised semantic segmentation [J]. IEEE Access, 2024, 12: 4871-4880.

[10] XU L, OUYANG W L, BENNAMOUN M, et al. Multi-class token transformer for weakly supervised semantic segmentation[C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022: 4310-4319.

[11] 李利荣,丁江,梅冰,等. 基于像素注意力特征融合的城市街景语义分割算法研究[J]. 电子测量技术, 2023, 46(20):184-190.
LI L R,DING J,MEI B, et al. Semantic segmentation method for urban street scenes based on pixel attention feature fusion [J]. Electronic Measurement Technology, 2023, 46(20):184-190.

[12] 杨大伟,迟津生,毛琳. 基于边界辅助的弱监督语义分割网络[J]. 计算机应用研究, 2024,41(2): 623-628,634.
YANG D W, CHI J SH, MAO L. Weakly supervised semantic segmentation network based on boundary assistance[J]. Application Research of Computers, 2024,41(2):623-628,634.

[13] 黄聪,杨珺,刘毅,等. 基于改进DeeplabV3+的遥感图像分割算法[J]. 电子测量技术, 2022, 45(21): 148-155.
HUANG C, YANG J, LIU Y, et al. Remote sensing image segmentation algorithm based on improved DeeplabV3 + [J]. Electronic Measurement Technology, 2022,45(21):148-155.

[14] ZHOU T F, ZHANG M J, ZHAO F, et al. Regional semantic contrast and aggregation for weakly supervised semantic segmentation [C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022: 4299-4309.

[15] KWEON H, YOON S H, YOON K J. Weakly supervised semantic segmentation via adversarial learning of classifier and reconstructor[C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023: 11329-11339.

[16] LEE J, KIM E, YOON S. Anti-adversarially manipulated attributions for weakly and semi-supervised semantic segmentation [C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021: 4071-4080.

[17] LI Y, DUAN Y Q, KUANG ZH H, et al. Uncertainty estimation via response scaling for pseudo-mask noise mitigation in weakly-supervised semantic segmentation [C]. AAAI Conference on Artificial Intelligence, 2022, 36(2): 1447-1455.

[18] LI J L, JIE Z Q, WANG X, et al. Expansion and shrinkage of localization for weakly-supervised semantic segmentation [J]. Advances in Neural Information Processing Systems, 2022, 35: 16037-16051.

[19] RONG SH H, TU B H, WANG Z L, et al. Boundary-enhanced co-training for weakly supervised semantic segmentation[C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023: 19574-19584.

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

[21] CHENG Z S, QIAO P CH, LI K H, et al. Out-of-candidate rectification for weakly supervised semantic segmentation [C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023: 23673-23684.

[22] PAN J W, ZHU P F, ZHANG K H, et al. Learning self-supervised low-rank network for single-stage weakly and semi-supervised semantic segmentation[J]. International Journal of Computer Vision, 2022, 130 (5): 1181-1195.

[23] 白雪飞,李文静,王文剑. 基于显著性背景引导的弱监督语义分割网络[J]. 模式识别与人工智能, 2021, 34(9): 824-835.
BAI X F, LI W J, WANG W J. Saliency background guided network for weakly-supervised semantic segmentation[J]. Pattern Recognition and Artificial Intelligence, 2021, 34(9):824-835.

作者简介

王家莉,硕士研究生,主要研究方向为弱监督语义分割。
E-mail:1918183420@qq.com

谭棉(通信作者),副教授,主要研究方向为自然图像抠图、微计算。
E-mail:tanmian@gzmu.edu.cn

冯夫健,教授,博士,主要研究方向为智能计算、微计算及其应用。
E-mail: fujian_feng@gzmu.edu.cn