软件学报 ISSN 1000-9825, CODEN RUXUEW
[doi: 10.13328/j.cnki.jos.007265] [CSTR: 32375.14.jos.007265]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

基于语义调制的弱监督语义分割*

李军侠[1,2], 苏京峰[1,2], 崔 滢[3], 刘青山[1,2]

[1](南京信息工程大学 计算机学院, 江苏 南京 210094)

[2](江苏省大气环境与装备技术协同创新中心, 江苏 南京 210094)

[3](浙江工业大学 计算机科学与技术学院, 浙江 杭州 310023)

通信作者: 刘青山, E-mail: qsliu@nuist.edu.cn

摘 要: 图像级标注下的弱监督语义分割方法通常采用卷积神经网络 (CNN) 生成类激活图以精确定位目标位置, 其面临的主要挑战在于 CNN 对全局信息感知能力的不足导致前景区域过小的问题. 近年来, 基于 Transformer 的弱监督语义分割方法利用自注意力机制捕捉全局依赖关系, 解决了 CNN 的固有缺陷. 然而, Transformer 生成的初始类激活图会在目标区域周围引入大量背景噪声, 此时直接对初始类激活图进行使用并不能取得令人满意的效果. 通过综合利用 Transformer 生成的类与块间注意力 (class-to-patch attention) 以及区域块间注意力 (patch-to-patch attention) 对初始类激活图进行联合优化, 同时, 由于原始的类与块间注意力存在误差, 对此设计一种语义调制策略, 利用区域块间注意力的语义上下文信息对类与块间注意力进行调制, 修正其误差, 最终得到能够准确覆盖较多目标区域的类激活图. 在此基础上, 构建一种新颖的基于 Transformer 的弱监督语义分割模型. 所提方法在 PASCAL VOC 2012 验证集和测试集上 *mIoU* 值分别达到 72.7% 和 71.9%, MS COCO 2014 验证集上 *mIoU* 为 42.3%, 取得了目前较为先进的弱监督语义分割结果.

关键词: 语义分割; 弱监督学习; 语义上下文; Transformer; 类激活图

中图法分类号: TP391

中文引用格式: 李军侠, 苏京峰, 崔滢, 刘青山. 基于语义调制的弱监督语义分割. 软件学报. http://www.jos.org.cn/1000-9825/7265.htm

英文引用格式: Li JX, Su JF, Cui Y, Liu QS. Semantic-modulation-based Weakly Supervised Semantic Segmentation. Ruan Jian Xue Bao/Journal of Software (in Chinese). http://www.jos.org.cn/1000-9825/7265.htm

Semantic-modulation-based Weakly Supervised Semantic Segmentation

LI Jun-Xia[1,2], SU Jing-Feng[1,2], CUI Ying[3], LIU Qing-Shan[1,2]

[1](School of Computer Science, Nanjing University of Information Science and Technology, Nanjing 210094, China)

[2](Jiangsu Collaborative Innovation Center of Atmospheric Environment and Equipment Technology, Nanjing 210094, China)

[3](College of Computer Science and Technology, Zhejiang University of Technology, Hangzhou 310023, China)

Abstract: Image-level weakly supervised semantic segmentation usually uses convolutional neural networks (CNNs) to generate class activation maps to accurately locate targets. However, CNNs have a limited capacity to perceive global information, which results in excessively narrow foregrounds. Recently, Transformer-based weakly supervised semantic segmentation has utilized self-attention mechanisms to capture global dependencies, addressing the inherent defects of CNNs. Nevertheless, the initial class activation map generated by a Transformer often introduces a lot of background noise around the target area, resulting in unsatisfactory performance if used directly. This study comprehensively utilizes both class-to-patch and patch-to-patch attention generated by a Transformer to optimize the initial class activation map. At the same time, a semantic modulation strategy is designed to correct errors in the class-to-patch

* 基金项目: 国家重点研发计划 (2022YFC2405600); 国家自然科学基金 (62272235, 62102364, U21B2044); 浙江省自然科学基金 (LY22F020016)
收稿时间: 2023-09-08; 修改时间: 2024-01-11; 采用时间: 2024-07-25; jos 在线出版时间: 2025-01-08

attention, using the semantic context information of the patch-to-patch attention. Finally, a class activation map that accurately covers more target areas is obtained. On this basis, a novel model for weakly supervised semantic segmentation based on a Transformer is constructed. The *mIoU* of the proposed method reaches 72.7% and 71.9% on the PASCAL VOC 2012 validation and test sets, respectively, and 42.3% on the MS COCO 2014 validation set, demonstrating that the proposed method achieves improved performance in weakly supervised semantic segmentation.

Key words: semantic segmentation; weakly supervised learning; semantic context; Transformer; class activation map

语义分割是计算机视觉领域一个非常重要且基础的研究方向, 该任务利用计算机的特征表达来模拟人类对图像的识别过程, 为给定图像的每一个像素分配一个语义类别标签. 语义分割在许多领域具有广泛的应用, 如图像识别、自动驾驶、医学图像分析、场景理解和视频分析等, 它可以帮助计算机更好地理解图像中的内容, 从而实现自动化的场景理解和决策. 近年来随着深度学习技术的蓬勃发展[1], 语义分割技术也取得了长足的发展与进步, 其中全监督的语义分割模型被广泛应用并取得了优秀的性能[2]. 然而训练全监督的语义分割模型需要大规模的像素级标注数据, 而像素级标注数据的获取往往难度大且耗时耗力. 为了解决这个问题, 许多工作开始转向研究弱监督语义分割技术. 弱监督语义分割是指只使用弱标注的数据对语义分割模型进行训练, 常用的弱标注包括边界框标注[3]、涂鸦标注[4]、点标注[5]以及图像级标注[6−8]. 其中图像级标注仅需要给出图像存在的具有目标类别信息, 并不需要指出目标类别在图像中的位置, 极大地减少了数据标注的时间和代价. 此外, 大规模基于图像级标注的训练数据可以从在线的多媒体分享网站中快速且方便的获取, 这也极大地缓解了训练数据规模不足的问题. 基于图像级标注的弱监督图像语义分割技术也因此成为计算机视觉领域一大学术研究热点. 本文特别关注使用图像级标注的弱监督语义分割.

如何从图像级标注中推断出高质量且稠密的位置信息, 进而基于推断的伪标注数据构建图像语义分割网络是基于图像级标注弱监督图像语义分割方法面临的关键和难点问题. 类激活图 (CAM)[9]的提出提供了一种只使用图像级标注来获取位置信息的有效方法, 其在分类网络的基础上通过对不同特征映射加权平均得到每个类别对应的鉴别区域. 对于图像级标注的弱监督语义分割, 大多数现有方法通常使用以下流程来解决: 1) 利用图像级标注训练卷积神经网络 (CNN), 生成类激活图以获得种子区域; 2) 对种子区域进行一定约束的扩张以获得伪标签; 3) 使用伪标签作为真实标签来训练全监督语义分割网络. 然而, 卷积神经网络产生的类激活图存在一个问题, 即它倾向于激活一个局部的有辨别力的区域, 而忽略了完整的对象区域, 导致不完全激活问题[6−8]. 最近有研究证明这是由于卷积神经网络的固有特性导致的, 即卷积神经网络中的卷积操作只能捕获小范围的特征依赖性[10], 无法探索全局特征关系, 导致激活对象区域过小, 从而影响生成的伪标签质量, 最终难以得到理想的弱监督语义分割结果.

最近, vision Transformer (ViT)[11]在许多计算机视觉任务中取得了巨大的成功[12], 这主要得益于其本身的自注意力机制, 该机制可以对全局特征关系进行建模, 有效克服卷积神经网络的上述缺点. 因此许多研究人员开始将 ViT 引入弱监督语义分割任务中, 并取得了优异的成果, 例如 TS-CAM[13]、MCTformer[14]等, 通常这些方法会先得到一个粗糙的初始类激活图, 之后直接使用 ViT 生成的原始类与块间注意力进行类激活图的计算. 然而, 本文通过实验发现, ViT 生成的原始类与块间注意力往往存在误差 (如图 1(b) 所示, 对于狗所对应的类与块间注意力来说, 黑色方框所标注的区域块的注意力存在误差), 此时如果直接使用原始类与块间注意力对初始类激活图进行计算往往得不到理想的结果.

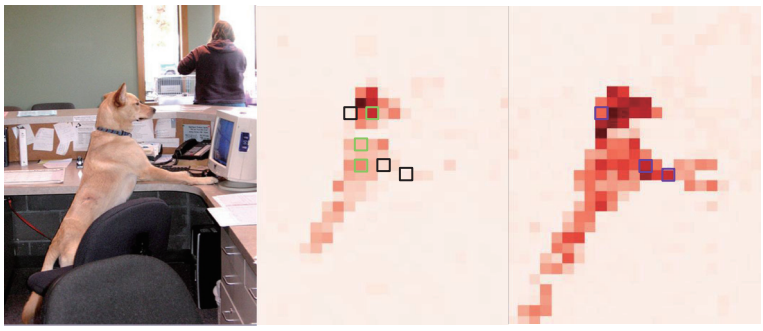
为了解决上述问题, 在本文中构建了一种基于 ViT 的类激活图联合优化框架, 通过综合利用 ViT 生成的类与块间注意力以及区域块间注意力对初始类激活图进行联合优化, 得到能够完整且准确覆盖地目标区域的类激活图. 此外, 在构建的基于 ViT 的类激活图联合优化框架中, 提出了一种语义调制策略, 根据区域块间注意力的语义上下文信息来修正类与块间注意力中存在的误差 (如图 1(c) 紫色方框所标注的区域块所示, 本文方法可以对存在误差的注意力进行有效修正, 显著提升其准确性). 本文主要贡献包括以下 3 点.

1) 构建了一种基于 ViT 的类激活图联合优化框架, 综合利用 ViT 生成的类与块间注意力以及区域块间注意力对初始类激活图进行联合优化, 得到可以较为准确且全面地覆盖前景目标区域的类激活图, 同时有效抑制背景噪声.

2) 针对 ViT 生成的原始类与块间注意力存在误差的问题, 提出了一种语义调制策略, 利用区域块间注意力的

语义上下文信息来有效修正误差, 从而获得更准确的类与块间注意力.

3) 在常用的 PASCAL VOC 2012 和 MS COCO 2014 数据集上进行实验, 所提出的弱监督语义分割方法实现了先进的性能, 验证了所提算法的可行性与有效性.



(a) 原图 (b) 调制前的类与块间注意力可视化图 (c) 调制后的类与块间注意力可视化图

图 1 类与块间注意力可视化图

1 相关工作

1.1 弱监督语义分割

现有的弱监督语义分割方法通常使用卷积神经网络生成类激活图, 进而生成像素级伪标签, 之后以完全监督的方式使用伪标签来训练语义分割网络. 然而, 基于卷积神经网络得到的类激活图只能够激活目标对象中最具辨别力的部分, 无法对语义分割网络的学习提供足够的监督. 为此, 许多工作致力于获得更好的类激活图. 例如, 对抗性擦除[15]利用擦除策略擦除目标对象中最具辨别力的区域, 然后重新训练分类网络, 从而迫使网络激活剩余目标区域. 受分类网络在不同训练阶段所关注的对象区域一直变化这一特性的启发, OAA[16]通过整合不同阶段的激活区域以生成更完整的类激活图. 还有一些工作[17]提出从多个输入图像中捕捉语义相似性和相异性以此来获得更完整的目标区域. SC-CAM[8]利用子类别目标进行对象区域挖掘, 从而发现更多对象区域. SEAM[6]设计了一个探索像素上下文相关性的模块, 并提出利用仿射变换下的一致性约束来得到高精度的种子区域. CPN[18]提出一种互补补丁网络, 通过缩小互补图像生成的类激活图和原始图像生成的类激活图之间的差距来获得具有更多关于目标种子信息的类激活图. 此外, CPN 还提出了一个像素-区域相关模块, 通过利用特征映射和类激活图之间的像素-区域关系来增强上下文信息. L2G[19]提出了一种从局部到全局的知识转移方法, 该方法包括一个局部网络和全局网络. 首先利用局部网络学习丰富的对象细节知识, 然后全局网络在线从局部网络中学习互补的知识, 从而获得更完整的对象关注. AMR[20]提出了一种新颖的激活调制和重校准方法, 该方法利用聚光灯分支和补偿分支对类激活图加权修正, 从而提供重校准的监督信号. 其中补偿分支提出了注意力调节模块, 按照通道-空间的顺序重新学习特征重要性的分布, 这有助于显式地建模通道相关性和空间编码, 以自适应地调节面向分割任务的激活响应; 此外, AMR 还针对双分支引入了一种交叉伪监督机制作为语义相似的正则化机制来相互细化两个分支. 其他的一些工作通过迭代的方法改进了类激活图, 例如, AffinityNet[7]提出学习相邻坐标像素对之间的语义相似性, 并应用随机游走 (RW) 来进一步细化种子区域. IRNet[21]利用邻域像素对之间的关系作为约束, 进一步探索了用于确定像素亲和性的边界激活图. 虽然这些方法最终都获得了更好的类激活图, 但是它们通常都是基于卷积神经网络, 因此该类方法往往受限于局部感受野, 无法探索全局特征关系. 本文通过使用 ViT 结构来避免此问题.

1.2 ViT

2020 年提出的 ViT 首次成功将 Transformer 架构应用于计算机视觉领域, 并在许多视觉任务中取得了出色的表现. TS-CAM[13]首次尝试在弱监督目标定位中使用 ViT, 它首先将类别信息从类 token 重新分配到块 token 中, 并将块 token 进行维度变换重塑为语义感知图, 然后将类与块注意力和语义感知图结合以获取位置线索. AFA[22]从

ViT 的注意力机制中学习可靠的语义亲和力, 并利用学习到的语义亲和力去调节初始伪标签. ToCo[23]设计了一个 patch token contrast 模块, 使用从中间层派生的伪标记关系来监督最终的 patch token, 其可以让语义区域更加对齐, 从而产生更准确的类激活图. 同时, ToCo 还提出了 class token contrast 模块, 从不确定区域中随机地取切片, 并最小化局部与全局图像分别对应的类标记间的特征差异, 促进目标区域的表示一致性. 此外, 卷积神经网络与 ViT 结合的工作也受到了越来越多的关注. TransCAM[24]利用 ViT 分支产生的区域块间注意力来优化卷积神经网络分支生成的类激活图.

与本文方法接近的一个工作是 MCTformer[14], 它在类激活图优化阶段直接使用 ViT 生成的原始类与块间注意力对初始类激活图进行优化. 然而据实验观察, 此时得到的原始类与块间注意力存在误差, 若直接对其进行使用往往得不到理想的结果. 与其不同, 本文方法并不直接使用原始类与块间注意力来对初始类激活图进行优化, 而是设计了一种语义调制策略, 利用区域块间注意力的语义上下文信息对类与块间注意力进行调制, 修正其误差, 使得类与块间注意力的准确性显著提高. 之后, 综合利用调制后的类与块间注意力以及区域块间注意力来联合优化初始类激活图, 得到能够准确覆盖较多前景目标区域的类激活图.

2 本文方法

本文构建了一种基于 ViT 的框架, 用于图像级标注下的弱监督语义分割任务, 该框架总体结构如图 2 所示, 主要由 3 部分组成: 1) 利用 ViT 对输入图像进行特征提取并生成粗糙的初始类激活图; 2) 语义调制策略, 在区域块间注意力中根据语义亲和力选择语义相关块, 之后使用语义相关块的注意力对类与块间注意力进行修正; 3) 综合利用类与块间注意力以及区域块间注意力来联合优化初始类激活图, 得到更加完整且准确地覆盖目标区域的类激活图.

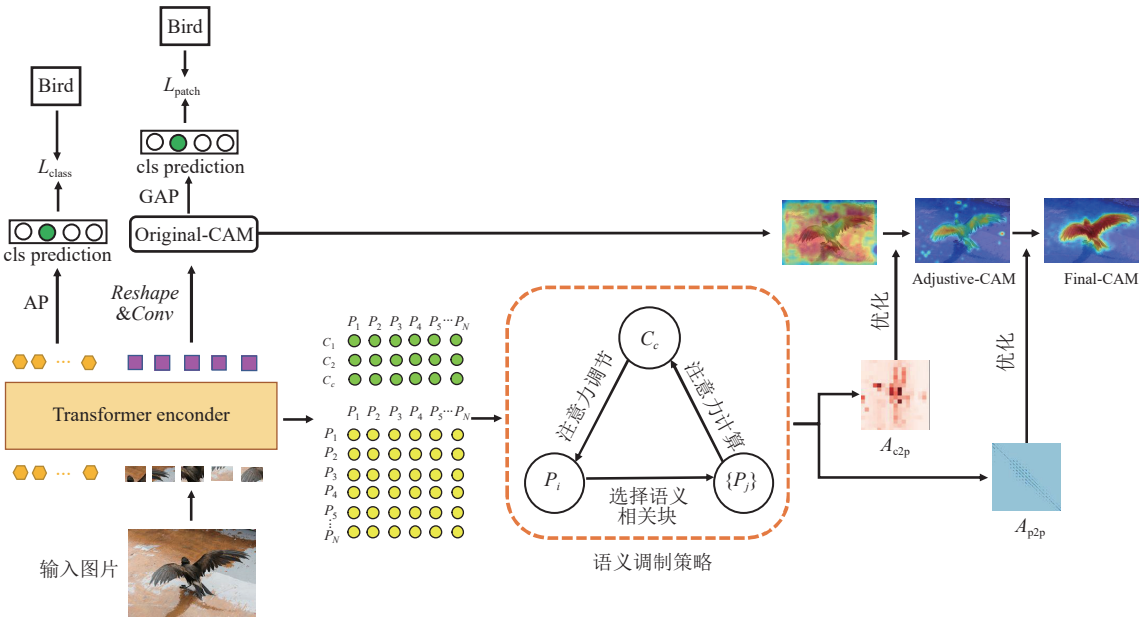


图 2 类激活图联合优化框架

2.1 初始类激活图生成

假设输入的 RGB 图像, 宽度为 W, 高度为 H, 首先将其切分为 $N = w \times h$ 个不重叠的块, 这里 $w=W/P$, $h=H/P$, 其中 P 代表一个区域块的高度和宽度, 然后将切分好的块通过线性映射构造 N 个初始块 token. 此外, 生成 C 个可学习的初始类 token, 其中 C 表示数据集中目标类别的总数, 并将它们与 N 个初始块 token 进行拼接, 之后添加一个可学习的位置编码, 得到输入 token $T_{in} \in R^{(C+N)\times D}$, 其中 D 代表每个 token 嵌入的维度. 将 T_{in} 输入到由 L 个连续

编码层组成的 Transformer 编码器中进行特征提取, 得到最终的输出 token $T_{\text{out}} \in R^{(C+N)\times D}$, 其中每个编码层由一个多头注意力 (MHA) 模块和一个多层感知器 (MLP) 组成. 最终的输出 token T_{out} 可以进一步划分为输出类 token $T_{c\_\text{out}} \in R^{C\times D}$ 和输出块 token $T_{p\_\text{out}} \in R^{N\times D}$. 最后将 T_{p_out} 进行重组 (*Reshape*) 并送到卷积层 (*Conv*) 中, 得到含有 C 个输出通道的初始类激活图 *Original-CAM:*

$$Original\text{-}CAM = Conv(Reshape(T_{p\_\text{out}})) \tag{1}$$

2.2 语义调制策略

从 Transformer 编码器中的多头注意力模块出发可以得到注意力 $A \in R^{M\times(C+N)\times(C+N)}$ (M 代表注意力头的个数), 计算如下:

$$A = \left(\frac{QK^{\mathrm{T}}}{\sqrt{d}}\right) \tag{2}$$

其中, Q, K 分别表示输入经过线性投影得到的 query 矩阵及 key 矩阵, d 表示缩放因子. 取 A 中所有注意力头的平均值得到 $\overline{A} \in R^{(C+N)\times(C+N)}$, 进一步, 通过 $\overline{A}$ 可以分别得到类与块间注意力 $\overline{A}_{c2p} \in R^{C\times N}$ 以及区域块间注意力 $\overline{A}_{p2p} \in R^{N\times N}$, 其中 $\overline{A}_{c2p} = \overline{A}[1:C, C+1:C+N]$, $\overline{A}_{p2p} = \overline{A}[C+1:C+N, C+1:C+N]$. 对得到的原始类与块间注意力进行可视化, 发现部分注意力存在误差, 如图 1(b) 黑色方框所标注的区域块的注意力. 同时发现, 绿色方框标注的区域块的注意力相对准确, 而且黑色方框标注的区域块和绿色方框标注的区域块在原图中的特征非常相近. 于是思考利用绿色方框标注的区域块的注意力来修正黑色方框标注的区域块的注意力.

因此本文设计了一种语义调制策略, 如图 3 所示. 具体步骤如下: 针对类 c 与块 i 之间的注意力, 首先从区域块间注意力中得到区域块 i 与其他每一个区域块间的语义亲和力, 接下来按照语义亲和力从大到小的顺序对所有的区域块进行排序, 之后把排在前 30% 的区域块选择出来, 并把选择出来的区域块称为区域块 i 的语义相关块, 根据前面的分析可知, 语义相关块就是与区域块 i 特征相似的一些区域块. 然后, 从类 c 所对应的类与块间注意力中将类 c 与语义相关块的注意力取出并进行计算, 得到类 c 与块 i 之间的注意力调节因子 $r(c,i) \in R^{1\times1}$, 计算如下:

$$r(c,i) = \frac{1}{S}\sum \overline{A}_{c2p}(c,j) \tag{3}$$

其中, $c \in \{1,2,\ldots,C\}$ 表示目标类别, i, j 表示块, $i \in \{1,2,\ldots,N\}$, $j \in U$, U 表示块 i 的语义相关块集合, S 表示 U 中语义相关块的数量. $\overline{A}_{c2p}(c,j)$ 表示类 c 与块 j 之间的注意力. 接下来使用注意力调节因子 $r(c,i)$ 对类 c 与块 i 之间的注意力进行调节:

$$\overline{A}_{c2p}(c,i) = \overline{A}_{c2p}(c,i) + \alpha \times r(c,i) \tag{4}$$

其中, α 代表调制系数. 之后, 对语义调制后的类与块间注意力进行可视化, 如图 1(c) 所示, 可以发现与调制前类与块间注意力对比, 调制后类与块间注意力准确性显著提高.

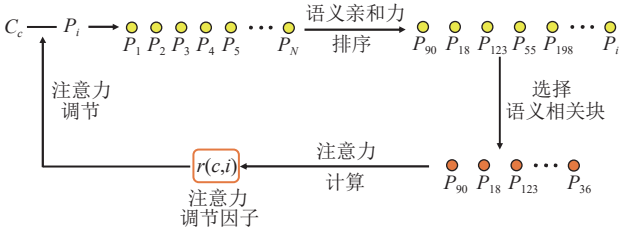


图 3 语义调制策略

2.3 类激活图优化

如图 4 中第 2 列和第 3 列所示, 可以发现 CNN 生成的初始类激活图往往只会覆盖图像中最具辨别力的区域. 最近已经有工作指出这个问题是由 CNN 的卷积操作引起的, 因为每次卷积核和图像进行卷积操作的时候, 卷积核所覆盖的区域只是图像的一小块, 故进行卷积操作的时候, 只能捕获小范围的特征依赖性, 无法捕获全局特征依赖

关系, 导致目标区域不完整. 然而, 由于 ViT 中自注意力机制的存在, ViT 能够捕获输入数据的全局特征依赖关系, 也就是说, ViT 能关注到输入数据中所有位置的信息, 因此当把图像分解为一系列的区域块, 并把这些区域块作为输入数据输入到 ViT 中后, ViT 中的自注意力机制可以获取每个区域块与其他区域块之间的关系. 因此 ViT 生成的初始类激活图可以覆盖较多的目标区域, 但是此时得到的类激活图会在目标区域周围引入大量背景噪声, 导致背景部分过度激活, 如图 4 中第 3 列所示. 在本文中, 通过综合利用 ViT 生成的类与块间注意力以及区域间注意力对初始类激活图进行联合优化, 得到更高质量的类激活图. 首先将初始类激活图 *Original-CAM* 与类与块间注意力逐元素相乘, 得到初步优化后的调制类激活图 *Modulated-CAM*:

$$Modulated\text{-}CAM = Original\text{-}CAM \bullet \overline{A}_{c2p} \tag{5}$$

其中, $\bullet$ 表示哈达玛积. 然后再利用区域块间注意力进一步优化调制类激活图 *Modulated-CAM*, 得到最终类激活图 *Final-CAM*:

$$Final\text{-}CAM = Modulated\text{-}CAM * \overline{A}_{p2p} \tag{6}$$

其中, $*$ 表示矩阵乘法. 经过类与块间注意力以及区域块间注意力优化后的类激活图能够覆盖较多的目标区域, 如图 4 中第 4 列所示. 接下来, 使用随机游走生成高质量的伪标签来训练语义分割网络.

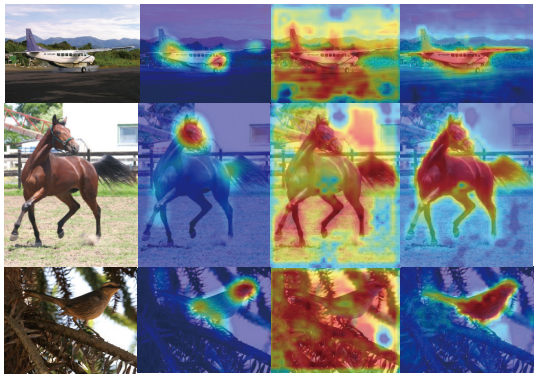


图 4　不同方法 CAM 可视化结果比较

2.4 损失函数

首先对输出类 token 中的每个类 token 采用平均池化 (AP) 操作得到预测类分数 $\hat{y}_{\text{cls}}$, 然后与真实标签做多标签交叉熵损失计算, 得到分类损失 L_{cls}:

$$\hat{y}_{\text{cls}}(c) = \frac{1}{D}\sum_{i=1}^{D} T_{c\_\text{out}}(c,i) \tag{7}$$

$$L_{\text{cls}} = \frac{1}{C}\sum_{k=1}^{C} y(k)\log\sigma(\hat{y}_{\text{cls}}(k)) + (1-y(k))\log(1-\sigma(\hat{y}_{\text{cls}}(k))) \tag{8}$$

其中, $c \in \{1,2,\ldots,C\}$ 表示类别, $T_{c\_\text{out}}(c,i)$ 表示 T_{c_out} 对应于第 c 个类 token 中第 i 个位置处的值, $y(k)$ 表示第 k 个类别的真实标签, $\hat{y}_{\text{cls}}(k)$ 表示第 k 个类别的预测类分数, σ 表示 Sigmoid 激活函数. 此外, 对初始类激活图 *Original-CAM* 进行全局平均池化 (GAP) 操作得到预测类分数 $\hat{y}_{\text{pat}}$, 与真实标签做多标签交叉熵损失计算, 得到分类损失 L_{pat}:

$$\hat{y}_{\text{pat}}(c) = \frac{1}{w \times h}\sum_{i=1}^{w}\sum_{j=1}^{h} Original\text{-}CAM(c,i,j) \tag{9}$$

$$L_{\text{pat}} = \frac{1}{C}\sum_{k=1}^{C} y(k)\log\sigma(\hat{y}_{\text{pat}}(k)) + (1-y(k))\log(1-\sigma(\hat{y}_{\text{pat}}(k))) \tag{10}$$

其中, w, h 表示 *Original-CAM* 的宽和高, $Original\text{-}CAM(c,i,j)$ 表示 *Original-CAM* 中第 c 个类别对应 (i,j) 位置处

的值, $\hat{y}_{\text{pat}}(k)$ 表示第 k 个类别的预测类分数. 最后, 总损失 L_{total} 定义为:

$$L_{\text{total}} = L_{\text{cls}} + L_{\text{pat}} \tag{11}$$

3 实验与分析

3.1 实验细节

本文使用在 ImageNet 上预训练的 DeiT-S[25]作为主干网络. 在训练过程中使用标准的数据增强方法, 包括随机缩放、随机水平翻转、颜色抖动和随机裁剪. 对训练图像进行归一化处理, 并将其大小调整为 256×256, 然后裁剪成 224×224 的大小作为网络模型输入. 使用 Adam 优化器来优化网络模型, 同时使用批量大小 64 对模型进行了 60 个周期的训练. 损失函数采用多标签交叉熵损失, 初始学习率设置为 5E–4. 对于语义分割, 遵循前人的工作[16]使用基于 ResNet38[26]的 DeepLabv1. 在进行推理时, 使用多尺度测试以及 CRF 进行后处理, 其中 CRF 的超参数设置如文献 [14] 所建议的. 本文所提模型是基于 PyTorch 深度学习框架实现的, 在 Ubuntu 环境下使用两张 NVIDIA GTX 2080 Ti 显卡进行训练.

3.2 数据集与评价指标

在 PASCAL VOC 2012[27]和 MS COCO 2014[28]两个数据集上进行实验, 验证所提框架的可行性和有效性. PASCAL VOC 2012 数据集有 21 个类别, 包括 20 个目标类和 1 个背景类. 该数据集分为 3 部分: 训练集 (包括 1 464 幅图像)、验证集 (包括 1 449 幅图像) 和测试集 (包括 1 456 幅图像). 同时遵循前人工作[18], 使用包含 10 582 幅图像的扩充训练集进行训练. MS COCO 2014 数据集有 81 个类别, 包括 80 个目标类和 1 个背景类, 训练集和验证集分别包含 82 081 副和 40 137 幅图像.

本文使用平均交并比 (mean intersection over union, *mIoU*) 作为评价标准来衡量所提方法在 PASCAL VOC 2012 和 MS COCO 2014 数据集上的语义分割性能. *mIoU* 定义为预测分割结果与真实分割结果的交集区域与并集区域之间的比值, 其度量的是预测分割结果与真实分割结果之间的相似性. 计算公式如下:

$$mIoU = \frac{1}{k+1}\sum_{i=0}^{k}\frac{p_{ii}}{\sum_{j=0}^{k}p_{ij}+\sum_{j=0}^{k}(p_{ji}-p_{ii})} \tag{12}$$

其中, k 表示数据集中的目标类别总数, i 表示真实值, j 表示预测值, p_{ij} 表示将真实值为 i, 预测为类别 j 的像素数量. *mIoU* 的取值范围是 0–1 之间, 数值越高表示预测的分割结果与真实分割结果的重叠程度越好, 即预测的分割结果的准确性越高. 此外, 从官方的 PASCAL VOC 在线评测服务器上获得 PASCAL VOC 2012 测试集上的语义分割结果.

3.3 模型复杂性

本文在表 1 展示了所提方法与 MCTformer 以及两篇 2023 年具有代表性的 SOTA 方法的比较结果, 该比较基于计算复杂度、参数数量、推断速度、运行时间和内存占用. 从表 1 可以看出, 本文方法与 MCTformer 相比, 在各种比较参数差距不大的情况下, 取得了更优异的效果; 而与 CLIP-ES[29]以及 LPCAM[30]两种最新 SOTA 方法相比, 在仅使用它们总运行时间约 1/4 的情况下, 就达到了和它们相似的结果, 充分说明了本文所提方法的优越性.

表 1 模型复杂性比较

方法	会议	MACs (G)	Params (M)	FPS	Time (h)	Memory usage (MB)
MCTformer[14]	CVPR 2022	4.7	21.7	15.5	7.7	2 507
CLIP-ES[29]	CVPR 2023	17.6	149.6	1.78	35.5	1 325
LPCAM[30]	CVPR 2023	55.9	70.4	3.16	35.9	6 805
本文方法	—	4.7	21.7	9.36	7.9	2 575

3.4 消融实验

3.4.1 语义调制策略的影响

为了进一步分析本文提出的语义调制策略带来的影响, 本节给出了 PASCAL VOC 2012 训练集上的类激活图

结果以及验证集上的分割结果, 如表 2 所示. 在这里, 使用 MCTformer[14]作为基准模型, 该模型没有使用语义调制策略, 它对应的类激活图 *mIoU* 和分割结果 *mIoU* 分别为 61.7% 和 71.9%. 本文所提框架使用了语义调制策略, 对应的类激活图 *mIoU* 和分割 *mIoU* 分别为 63.6% 和 72.7%, 分别比基准模型提升了 1.9% 和 0.8%. 该结果充分证明了语义调制策略的有效性.

表 2　语义调制策略的影响 (%)

方法	类激活图*mIoU*	分割*mIoU*
基准	61.7	71.9
本文方法	**63.6**	**72.7**

在语义调制策略中, 当对类与某个区域块之间的注意力进行调节时, 首先会将与当前区域块语义相关的区域块选择出来, 然后对语义相关块的注意力进行计算以获得注意力调节因子, 之后再进行注意力调节. 由于选择不同数量的语义相关块所带来的影响是不同的. 本节设置了 5 个不同的语义相关块数量占比, 分别为总块数的 10%、20%、30%、40%、50%, 并给出了 PASCAL VOC 2012 训练集上对应的类激活图结果, 如图 5 所示. 发现当选择的语义相关块数量比为 30% 时, 获得了最好的 *mIoU* 结果. 因此, 在本文中将选择的语义相关块数量占比设置为 30%, 此时语义相关块对类与块之间注意力误差的调节效果最好, 对应的类激活图质量最高.

在本文中, 使用公式 (4) 对类与块间注意力进行调节, 其中 α 为调制系数, 不同的调制系数所带来的影响是不同的, 下面对调制系数 α 进行分析, 并在表 3 中给出了 PASCAL VOC 2012 训练集上的类激活图对比结果. 从表 3 可以看出, 当调制系数设置为 1.2 时对应的 *mIoU* 值最高, 达到了 63.6%. 因此, 在本文所有实验中, 调制系数设置为 1.2, 此时, 类与块间注意力经过调节后存在的误差最少, 对应的类激活图质量最高.

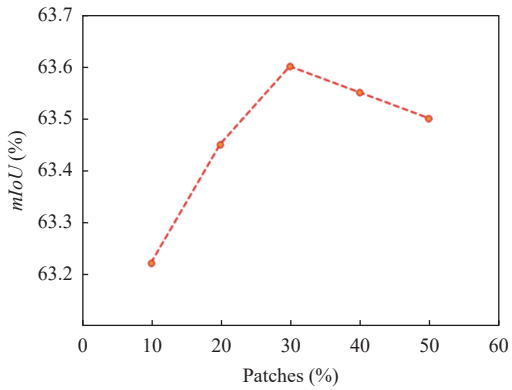


图 5　语义相关块数量占比对结果的影响

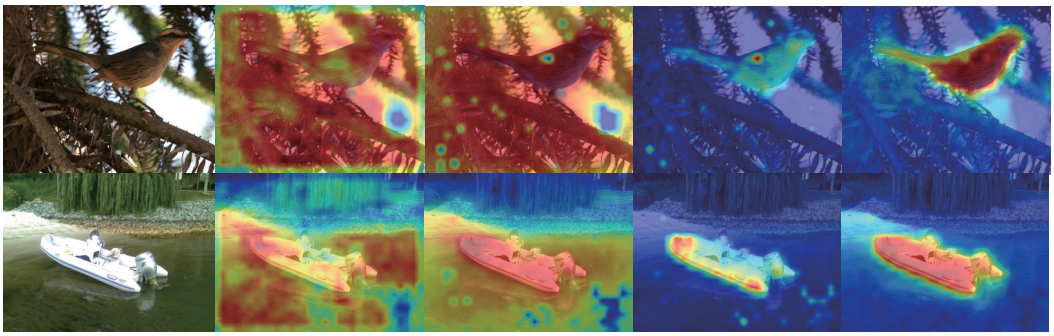
表 3　调制系数对结果的影响

α	类激活图*mIoU* (%)
0.6	63.45
0.8	63.51
1.0	63.55
1.2	**63.60**
1.4	63.54
1.6	63.52
1.8	63.50

3.4.2　注意力联合优化的影响

本文所提框架使用 ViT 对输入图像进行特征提取并生成初始类激活图, 然而此时得到的初始类激活图会在激活的目标区域周围引入了大量背景噪声, 导致背景部分过度激活, 如图 6(b) 所示. 针对这个问题, 广泛使用的一个解决方法是利用 ViT 自身产生的注意力来对过度激活的类激活图进行额外优化. 在图中展示了使用不同类型注意力对初始类激活图进行优化的可视化结果, 可以发现, 单独使用区域块间注意力进行优化的效果并不好, 如图 6(c) 所示, 许多背景部分同样被过度激活, 通过分析可知, 在使用区域块间注意力优化初始类激活图的过程中所涉及的主要操作是矩阵乘法, 而初始类激活图中许多激活值是错误的, 因此初始类激活图与区域块间注意力进行矩阵乘法计算后得到的激活值同样是错误的, 最终导致效果不好; 单独使用类与块间注意力进行优化的效果比使用区域块间注意力好, 如图 6(d) 所示, 背景噪声大大减少, 而且目标区域也能被较为准确地激活, 然而此时存在两个问题: 第一, 有些目标区域没有被正确激活; 第二, 有些被激活的目标区域激活值较小. 这时考虑使用类与块间注意力和区域块间注意力的结合来优化初始类激活图. 因为经过类与块间注意力优化得到的类激活图中存在的错误激活值

较少, 此时在与区域块间注意力进行矩阵乘法计算可以得到较为准确的激活值. 通过实验发现, 综合利用类与块间注意力和区域块间注意力来优化初始类激活图可以得到比单独使用类与块间注意力或区域块间注意力更好的效果, 其可视化结果如图 6(e) 所示.



(a) 原图 (b) 初始类激活图 (c) 区域块间注意力 (d) 类与块间注意力 (e) 区域块间注意力+类与块间注意力

图 6 不同类型注意力优化生成的激活图结果

同时对应的类激活图 *mIoU* 结果如表 4 所示, 侧面印证了使用类与块注意力和区域块间注意力的结合能够更好地优化初始类激活图.

因此本文使用类与块间注意力以及区域块间注意力对初始类激活图进行联合优化, 分别得到调节类激活图以及最终类激活图. 下面对不同阶段得到的类激活图进行分析, 并在表 5 中给出了 PASCAL VOC 2012 训练集上的类激活图对比结果. 从表 5 可以看出, Original-CAM 的 *mIoU* 值为 46.1%, 接下来使用类与块间注意力优化得到 Modulated-CAM, 此时的 *mIoU* 值提高到 59.3%, 然后使用区域块间注意力对 Modulated-CAM 进一步优化得到 Final-CAM, 其 *mIoU* 值可以达到 63.6%.

表 4 PASCAL VOC 2012 训练集上类激活图结果

方法	类激活图 *mIoU* (%)
初始类激活图	46.1
初始类激活图+区域块间注意力	54.2
初始类激活图+类与块间注意力	59.3
初始类激活图+类与块间注意力+区域块间注意力	**63.6**

表 5 注意力联合优化的影响

方法	类激活图*mIoU* (%)
Original-CAM	46.1
Modulated-CAM	59.3
Final-CAM	**63.6**

3.5 与先进算法的比较

3.5.1 PASCAL VOC 2012

类激活图与伪标签对比: 表 6 给出了不同方法对应类激活图和伪标签的比较结果 (粗体标记最佳结果). 如表中第 2 列所示, 本文方法在 PASCAL VOC 2012 数据集训练集上取得了最优的类激活图结果, 其 *mIoU* 值达到了 63.6%. 相比于其他一些最新方法, 例如 AMR[20]、AFA[22]和 CLIMS[31], 本文类激活图的 *mIoU* 值分别高出 6.8%、11.0% 和 7.0%. 受启发于前人工作[6,8], 采用 PSA[7]对类激活图进行后处理, 以得到高质量的像素级伪标签. 本文方法经过 PSA 后处理得到的伪标签在 PASCAL VOC 2012 训练集上达到了 70.5% 的 *mIoU*, 比 AFA[22]和 MCTformer[14]分别高出 1.8% 和 1.4%, 如表中第 3 列所示. 该实验结果很好地证明了本文方法在生成高质量类激活图和高精度伪标签方面的有效性.

分割结果对比: 本节将本文方法与最新的基于图像级标注的弱监督语义分割模型进行分割性能比较, 本文方法使用 ImageNet 上预先训练的 ResNet38 对伪标签进行全监督训练. 表 7 给出了在 PASCAL VOC 2012 数据集验证集和测试集上分割对比结果. 其中, I 表示图像级标签, S 表示显著性图, 粗体标记最佳结果. 如表 7 所示, 本文方法在 PASCAL VOC 2012 验证集和测试集上 *mIoU* 指标分别达到了 72.7% 和 71.9%, 在没有使用额外监督的情况下

(显著图), 其结果与最新 SOTA 方法所得到的结果接近. 此外, 相比较于使用额外显著性监督的最新 SOTA 方法 Mat-Label[35], 本文方法在验证集和测试集上所得到的结果虽然与其具有差距, 但是仅相差 0.6% 和 2.1%. 进一步印证了本文方法的优越性.

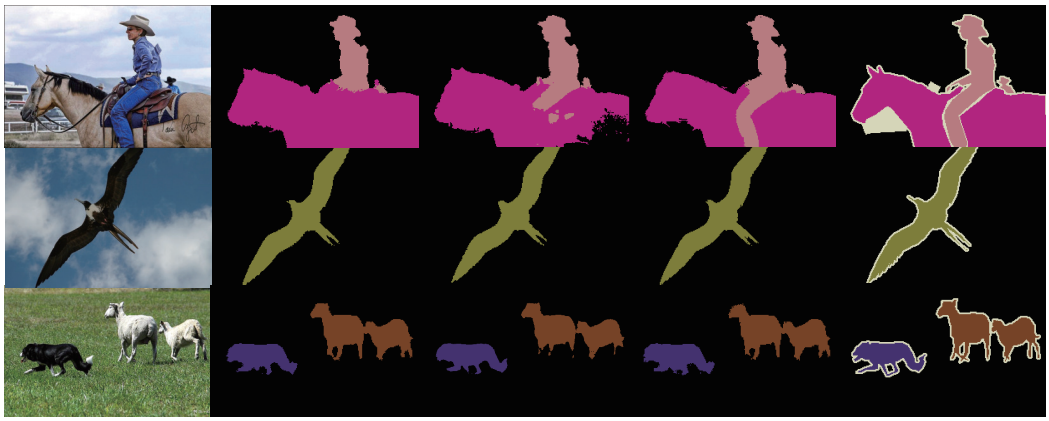
表 6　本文方法与其他方法在 PASCAL VOC 2012 训练集上类激活图和伪标签比较结果 (%)

方法	会议	类激活图*mIoU*	伪标签*mIoU*	方法	会议	类激活图*mIoU*	伪标签*mIoU*
SC-CAM[8]	CVPR 2020	50.9	63.4	AMR[20]	AAAI 2022	56.8	69.7
SEAM[6]	CVPR 2020	55.4	63.6	CLIMS[31]	CVPR 2022	56.6	**70.5**
EDAM[32]	CVPR 2021	52.8	68.1	AFA[22]	CVPR 2022	52.6	68.7
AdvCAM[33]	CVPR 2021	55.6	69.9	MCTformer[14]	CVPR 2022	61.7	69.1
CPN[18]	ICCV 2021	57.4	67.8	本文方法	—	**63.6**	**70.5**
PMM[34]	ICCV 2021	58.2	61.5				

表 7　本文方法与其他方法在 PASCAL VOC 2012 验证集和测试集上分割结果比较 (%)

方法	会议	监督信息	验证集 *mIoU*	测试集 *mIoU*	方法	会议	监督信息	验证集 *mIoU*	测试集 *mIoU*
EDAM[32]	CVPR 2021	I+S	70.9	70.9	AMN[42]	CVPR 2022	I	69.5	69.6
EPS[36]	CVPR 2021	I+S	71.0	71.8	W-OoD[43]	CVPR 2022	I	70.7	70.1
InferCAM[37]	WACV 2022	I+S	70.8	71.8	SIPE[44]	CVPR 2022	I	68.8	69.7
RCA[38]	CVPR 2022	I+S	72.2	72.8	ViT-PCM[45]	ECCV 2022	I	70.3	70.9
L2G[19]	CVPR 2022	I+S	72.1	71.7	AEFT[46]	ECCV 2022	I	70.9	71.7
Mat-Label[35]	ICCV 2023	I+S	**73.3**	**74.0**	ToCo[23]	CVPR 2023	I	69.8	70.5
SEAM[6]	CVPR 2020	I	64.5	65.7	ACR[47]	CVPR 2023	I	71.9	71.9
CDA[39]	ICCV 2021	I	66.1	66.8	BECO[48]	CVPR 2023	I	72.1	71.8
URN[40]	AAAI 2022	I	69.5	69.7	CLIP-ES[29]	CVPR 2023	I	71.1	71.4
MCTformer[14]	CVPR 2022	I	71.9	71.6	LPCAM[30]	CVPR 2023	I	72.6	**72.4**
ReCAM[41]	CVPR 2022	I	68.5	68.4	OCR[49]	CVPR 2023	I	**72.7**	72.0
CLIMS[31]	CVPR 2022	I	69.3	68.7	本文方法	—	I	**72.7**	71.9

同时, 后文图 7 给出了本文方法在验证集上的分割结果示例, 可以看出, 与 CLIP-ES 相比, 本文方法在分割物体的细节处理方面表现更好, 如图 7 第 3 行中羊的腿部, 与 MCTformer 相比, 本文方法在分割物体的完整性方面表现更好, 如图 7 第 1 行中马的身体.



(a) 原图　(b) MCTformerr　(c) CLIP-ES　(d) Ours　(e) 真实标签

图 7　PASCAL VOC 2012 验证集上的分割结果

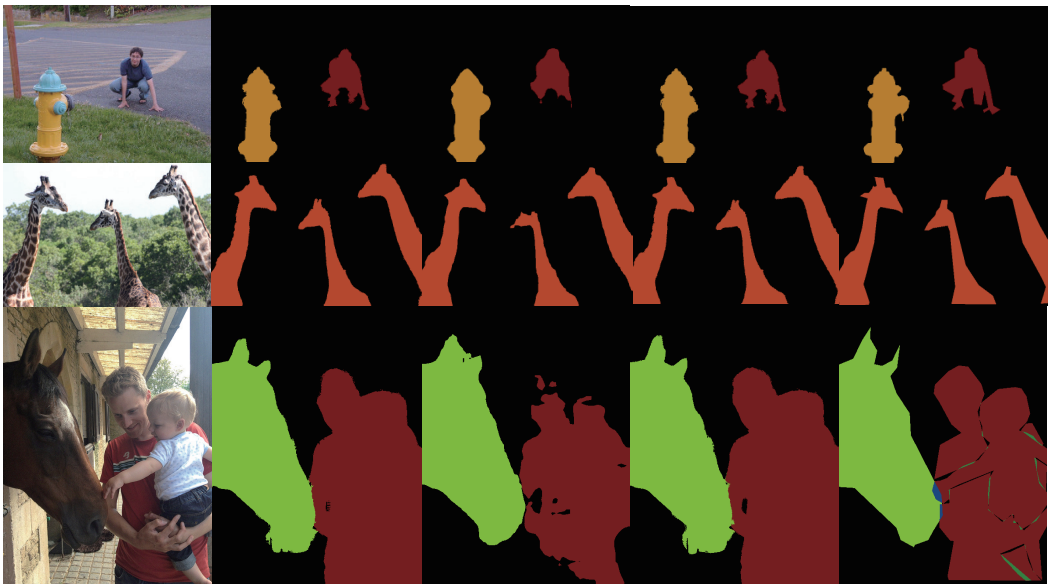
3.5.2 MS COCO 2014

表 8 给出了本文方法在 MS COCO 2014 验证集上与其他方法的分割性能比较结果. 其中, I 表示图像级标签, S 表示显著性图, 粗体标记最佳结果. 本文方法获得了 42.3% 的 *mIoU*, 相比于其他的一些最新方法, 例如 LPCAM [30] 和 OCR[49], 所提方法得到了与他们类似的结果.

表 8 本文方法与其他方法在 MS COCO 2014 验证集上分割结果比较 (%)

方法	会议	监督信息	验证集*mIoU*	方法	会议	监督信息	验证集*mIoU*
EDAM[32]	CVPR 2012	I+S	—	AMN[42]	CVPR 2022	I	44.7
EPS[36]	CVPR 2021	I+S	35.7	W-OoD[43]	CVPR 2022	I	—
InferCAM[37]	WACV 2022	I+S	—	SIPE[44]	CVPR 2022	I	43.6
RCA[38]	CVPR 2022	I+S	36.8	ViT-PCM[45]	ECCV 2022	I	45.0
L2G[19]	CVPR 2022	I+S	44.2	AEFT[46]	ECCV 2022	I	44.8
Mat-Label[35]	ICCV 2023	I+S	**45.6**	ToCo[23]	CVPR 2023	I	41.3
SEAM[6]	CVPR 2020	I	31.9	ACR[47]	CVPR 2023	I	45.3
CDA[39]	ICCV 2021	I	33.2	BECO[48]	CVPR 2023	I	45.1
URN[40]	AAAI 2022	I	40.7	CLIP-ES[29]	CVPR 2023	I	**45.4**
MCTformer[14]	CVPR 2022	I	42.0	LPCAM[30]	CVPR 2023	I	42.8
ReCAM[41]	CVPR 2022	I	—	OCR[49]	CVPR 2023	I	42.5
CLIMS[31]	CVPR 2022	I	—	本文方法	—	I	42.3

同时, 后文图 8 给出了本文方法在验证集上的分割结果示例, 从图 8 可以很清楚地看到, 与 CLIP-ES[29]相比, 本文方法在分割物体的细节处理方面表现更好, 如图 8 第 1 行中人的手部, 第 2 行中长颈鹿的头部, 与 MCTformer[14] 相比, 本文方法在分割物体的完整性方面表现更好, 如图 8 第 3 行中人的头部.



(a) 原图 (b) MCTformer (c) CLIP-ES (d) Ours (e) 真实标签

图 8 MS COCO 2014 验证集上的分割结果

3.6 局限性与不足

本文所提方法虽然取得了目前较为先进的性能, 但是在一些具有挑战性的场景中仍然存在局限性与不足. 例

如不能够很好地对透明物体进行分割, 如图 9 第 1 行所示, 本文方法没有准确地区分中间的瓶子. 此外, 本文方法在对于小物体的分割上也存在不足, 如图 9 第 2 行所示, 本文方法没有成功识别出图中的猫.

图 9 PASCAL VOC 2012 验证集上的分割结果

4 结 论

为了解决初始类激活图中前景区域过小、背景噪声过多的问题, 本文构建了一种基于 ViT 的类激活图联合优化框架. 首先, 针对 ViT 生成的原始类与块间注意力中存在的误差, 设计了一种语义调制策略, 利用区域块间注意力的语义上下文信息对其进行修正, 提高其准确性; 之后, 综合利用修正后的类与块间注意力以及区域块间注意力对初始类激活图进行联合优化. 最终得到的类激活图在准确覆盖目标区域的同时较好地抑制了背景噪声. 一系列的对比实验充分证明了本文所提方法的优越性及其有效性.

References:

[1] Bai C, Huang L, Chen JN, Pan X, Chen SY. Optimization of deep convolutional neural network for large scale image classification. Ruan Jian Xue Bao/Journal of Software, 2018, 29(4): 1029–1038 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/5404.htm [doi: 10.13328/j.cnki.jos.005404]

[2] Tian X, Wang L, Ding Q. Review of image semantic segmentation based on deep learning. Ruan Jian Xue Bao/Journal of Software, 2019, 30(2): 440–468 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/5659.htm [doi: 10.13328/j.cnki.jos.005659]

[3] Khoreva A, Benenson R, Hosang J, Hein M, Schiele B. Simple does it: Weakly supervised instance and semantic segmentation. In: Proc. of the 2017 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017. 876–885. [doi: 10.1109/CVPR.2017.181]

[4] Lin D, Dai JF, Jia JY, He KM, Sun J. ScribbleSup: Scribble-supervised convolutional networks for semantic segmentation. In: Proc. of the 2016 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016. 3159–3167. [doi: 10.1109/CVPR.2016.344]

[5] Bearman A, Russakovsky O, Ferrari V, Fei-Fei L. What's the point: Semantic segmentation with point supervision. In: Proc. of the 14th European Conf. on Computer Vision. Amsterdam: Springer, 2016. 549–565. [doi: 10.1007/978-3-319-46478-7_34]

[6] Wang YD, Zhang J, Kan MN, Shan SG, Chen XL. Self-supervised equivariant attention mechanism for weakly supervised semantic segmentation. In: Proc. of the 2020 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020. 12275–12284. [doi: 10.1109/CVPR42600.2020.01229]

[7] Ahn J, Kwak S. Learning pixel-level semantic affinity with image-level supervision for weakly supervised semantic segmentation. In: Proc. of the 2018 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 4981–4990. [doi: 10.1109/CVPR.2018.00523]

[8] Chang YT, Wang QS, Hung WC, Piramuthu R, Tsai YH, Yang MH. Weakly-supervised semantic segmentation via sub-category exploration. In: Proc. of the 2020 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020. 8988–8997. [doi: 10.1109/CVPR42600.2020.00901]

[9] Zhou BL, Khosla A, Lapedriza A, Oliva A, Torralba A. Learning deep features for discriminative localization. In: Proc. of the 2016 IEEE Conf. on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016. 2921–2929. [doi: 10.1109/CVPR.2016.319]

[10] Chen ZW, Wang CA, Wang YB, Jiang GN, Shen YH, Tai Y, Wang CJ, Zhang W, Cao LJ. LCTR: On awakening the local continuity of Transformer for weakly supervised object localization. In: Proc. of the 36th AAAI Conf. on Artificial Intelligence. AAAI Press, 2020. 710–718. [doi: 10.1609/aaai.v36i1.19918]

[11] Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai XH, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S, Uszkoreit J, Houlsby N. An image is worth 16x16 words: Transformers for image recognition at scale. In: Proc. of the 9th Int'l Conf. on Learning Representations. OpenReview.net, 2021.

[12] Shi ZN, Chen HP, Zhang D, Shen XJ. Pre-training-driven multimodal boundary-aware vision Transformer. Ruan Jian Xue Bao/Journal of Software, 2023, 34(5): 2051–2067 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/6768.htm [doi: 10.13328/j.cnki.jos.006768]

[13] Gao W, Wan F, Pan XJ, Peng ZL, Tian Q, Han ZJ, Zhou BL, Ye QX. TS-CAM: Token semantic coupled attention map for weakly supervised object localization. In: Proc. of the 2021 IEEE/CVF Int'l Conf. on Computer Vision. Montreal: IEEE, 2021. 2866–2875. [doi: 10.1109/ICCV48922.2021.00288]

[14] Xu L, Ouyang WL, Bennamoun M, Boussaid F, Xu D. Multi-class token Transformer for weakly supervised semantic segmentation. In: Proc. of the 2022 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. New Orleans: IEEE, 2022. 4300–4309. [doi: 10.1109/CVPR52688.2022.00427]

[15] Wei YC, Feng JS, Liang XD, Cheng MM, Zhao Y, Yan SC. Object region mining with adversarial erasing: A simple classification to semantic segmentation approach. In: Proc. of the 2017 IEEE Conf. on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017. 6488–6496. [doi: 10.1109/CVPR.2017.687]

[16] Jiang PT, Hou QB, Cao Y, Cheng MM, Wei YC, Xiong HK. Integral object mining via online attention accumulation. In: Proc. of the 2019 IEEE/CVF Int'l Conf. on Computer Vision. Seoul: IEEE, 2019. 2070–2079. [doi: 10.1109/ICCV.2019.00216]

[17] Sun GL, Wang WG, Dai JF, van Gool L. Mining cross-image semantics for weakly supervised semantic segmentation. In: Proc. of the 16th European Conf. on Computer Vision. Glasgow: Springer, 2020. 347–365. [doi: 10.1007/978-3-030-58536-5_21]

[18] Zhang F, Gu CC, Zhang CY, Dai YC. Complementary patch for weakly supervised semantic segmentation. In: Proc. of the 2021 IEEE/CVF Int'l Conf. on Computer Vision. Montreal: IEEE, 2021. 7222–7231. [doi: 10.1109/ICCV48922.2021.00715]

[19] Jiang PT, Yang YQ, Hou QB, Wei YC. L2G: A simple local-to-global knowledge transfer framework for weakly supervised semantic segmentation. In: Proc. of the 2022 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. New Orleans: IEEE, 2022. 16865–16875. [doi: 10.1109/CVPR52688.2022.01638]

[20] Qin J, Wu J, Xiao XF, Li LJ, Wang XG. Activation modulation and recalibration scheme for weakly supervised semantic segmentation. In: Proc. of the 36th AAAI Conf. on Artificial Intelligence. AAAI Press, 2022. 2117–2125. [doi: 10.1609/aaai.v36i2.20108]

[21] Ahn J, Cho S, Kwak S. Weakly supervised learning of instance segmentation with inter-pixel relations. In: Proc. of the 2019 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 2209–2218. [doi: 10.1109/CVPR.2019.00231]

[22] Ru LX, Zhan YB, Yu BS, Du B. Learning affinity from attention: End-to-end weakly-supervised semantic segmentation with Transformers. In: Proc. of the 2022 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. New Orleans: IEEE, 2022. 16825–16834. [doi: 10.1109/CVPR52688.2022.01634]

[23] Ru LX, Zheng HL, Zhan YB, Du B. Token contrast for weakly-supervised semantic segmentation. In: Proc. of the 2023 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Vancouver: IEEE, 2023. 3093–3102. [doi: 10.1109/CVPR52729.2023.00302]

[24] Li RW, Mai ZD, Zhang ZB, Jang J, Sanner S. TransCAM: Transformer attention-based CAM refinement for weakly supervised semantic segmentation. Journal of Visual Communication and Image Representation, 2023, 92: 103800. [doi: 10.1016/j.jvcir.2023.103800]

[25] Touvron H, Cord M, Douze M, Massa F, Sablayrolles A, Jégou H. Training data-efficient image Transformers & distillation through attention. In: Proc. of the 38th Int'l Conf. on Machine Learning. PMLR, 2021. 10347–10357.

[26] Wu ZF, Shen CH, van den Hengel A. Wider or deeper: Revisiting the ResNet model for visual recognition. Pattern Recognition, 2019, 90: 119–133. [doi: 10.1016/j.patcog.2019.01.006]

[27] Everingham M, Van Gool L, Williams CKI, Winn J, Zisserman A. The PASCAL visual object classes (VOC) challenge. Int'l Journal of Computer Vision, 2010, 88(2): 303–338. [doi: 10.1007/s11263-009-0275-4]

[28] Lin TY, Maire M, Belongie S, Hays J, Perona P, Ramanan D, Dollár P, Zitnick CL. Microsoft COCO: Common objects in context. In: Proc. of the 13th European Conf. on Computer Vision. Zurich: Springer, 2014. 740–755. [doi: 10.1007/978-3-319-10602-1_48]

[29] Lin YQ, Chen MH, Wang WX, Wu BX, Li K, Lin BB, Liu HF, He XF. CLIP is also an efficient segmenter: A text-driven approach for weakly supervised semantic segmentation. In: Proc. of the 2023 IEEE/CVF Conf. on Computer Vision and Pattern Recognition.

Vancouver: IEEE, 2023. 15305–15314. [doi: 10.1109/CVPR52729.2023.01469]

[30] Chen ZZ, Sun QR. Extracting class activation maps from non-discriminative features as well. In: Proc. of the 2023 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Vancouver: IEEE, 2023. 3135–3144. [doi: 10.1109/CVPR52729.2023.00306]

[31] Xie JH, Hou XX, Ye K, Shen LL. CLIMS: Cross language image matching for weakly supervised semantic segmentation. In: Proc. of the 2022 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. New Orleans: IEEE, 2022. 4473–4482. [doi: 10.1109/CVPR52688.2022.00444]

[32] Wu T, Huang JS, Gao GY, Wei XM, Wei XL, Luo X, Liu CH. Embedded discriminative attention mechanism for weakly supervised semantic segmentation. In: Proc. of the 2021 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Nashville: IEEE, 2021. 16765–16774. [doi: 10.1109/CVPR46437.2021.01649]

[33] Lee J, Kim E, Yoon S. Anti-adversarially manipulated attributions for weakly and semi-supervised semantic segmentation. In: Proc. of the 2021 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Nashville: IEEE, 2021. 4071–4080. [doi: 10.1109/CVPR46437.2021.00406]

[34] Li Y, Kuang ZH, Liu LY, Chen YM, Zhang W. Pseudo-mask matters in weakly-supervised semantic segmentation. In: Proc. of the 2021 IEEE/CVF Int'l Conf. on Computer Vision. Montreal: IEEE, 2021. 6964–6973. [doi: 10.1109/ICCV48922.2021.00688]

[35] Jo S, Yu IJ, Kim K. MARS: Model-agnostic biased object removal without additional supervision for weakly-supervised semantic segmentation. In: Proc. of the 2023 IEEE/CVF Int'l Conf. on Computer Vision. Paris: IEEE, 2023. 614–623. [doi: 10.1109/ICCV51070.2023.00063]

[36] Lee S, Lee M, Lee J, Shim H. Railroad is not a train: Saliency as pseudo-pixel supervision for weakly supervised semantic segmentation. In: Proc. of the 2021 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Nashville: IEEE, 2021. 5495–5505. [doi: 10.1109/CVPR46437.2021.00545]

[37] Sun WX, Zhang J, Barnes N. Inferring the class conditional response map for weakly supervised semantic segmentation. In: Proc. of the 2022 IEEE/CVF Winter Conf. on Applications of Computer Vision. Waikoloa: IEEE, 2022. 2653–2662. [doi: 10.1109/WACV51458.2022.00271]

[38] Zhou TF, Zhang MJ, Zhao F, Li JW. Regional semantic contrast and aggregation for weakly supervised semantic segmentation. In: Proc. of the 2022 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. New Orleans: IEEE, 2022. 4299–4309. [doi: 10.1109/CVPR52688.2022.00426]

[39] Su YK, Sun RZ, Lin GS, Wu QY. Context decoupling augmentation for weakly supervised semantic segmentation. In: Proc. of the 2021 IEEE/CVF Int'l Conf. on Computer Vision. Montreal: IEEE, 2021. 7004–7014. [doi: 10.1109/ICCV48922.2021.00692]

[40] Li Y, Duan YQ, Kuang ZH, Chen YM, Zhang W, Li XM. Uncertainty estimation via response scaling for pseudo-mask noise mitigation in weakly-supervised semantic segmentation. In: Proc. of the 36th AAAI Conf. on Artificial Intelligence. AAAI Press, 2022. 1447–1455. [doi: 10.1609/aaai.v36i2.20034]

[41] Chen ZZ, Wang T, Wu XW, Hua XS, Zhang HW, Sun QR. Class re-activation maps for weakly-supervised semantic segmentation. In: Proc. of the 2022 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. New Orleans: IEEE, 2022. 959–968. [doi: 10.1109/CVPR52688.2022.00104]

[42] Lee M, Kim D, Shim H. Threshold matters in WSSS: Manipulating the activation for the robust and accurate segmentation model against thresholds. In: Proc. of the 2022 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. New Orleans: IEEE, 2022. 4330–4339. [doi: 10.1109/CVPR52688.2022.00429]

[43] Lee J, Oh SJ, Yun S, Choe J, Kim E, Yoon S. Weakly supervised semantic segmentation using out-of-distribution data. In: Proc. of the 2022 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. New Orleans: IEEE, 2022. 16897–16906. [doi: 10.1109/CVPR52688.2022.01639]

[44] Chen Q, Yang LX, Lai JH, Xie XH. Self-supervised image-specific prototype exploration for weakly supervised semantic segmentation. In: Proc. of the 2022 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. New Orleans: IEEE, 2022. 4288–4298. [doi: 10.1109/CVPR52688.2022.00425]

[45] Rossetti S, Zappia D, Sanzari M, Schaerf M, Pirri F. Max pooling with vision Transformers reconciles class and shape in weakly supervised semantic segmentation. In: Proc. of the 17th European Conf. on Computer Vision. Tel Aviv: Springer, 2022. 446–463. [doi: 10.1007/978-3-031-20056-4_26]

[46] Yoon SH, Kweon H, Cho J, Kim S, Yoon KJ. Adversarial erasing framework via triplet with gated pyramid pooling layer for weakly supervised semantic segmentation. In: Proc. of the 17th European Conf. on Computer Vision. Tel Aviv: Springer, 2022. 326–344. [doi: 10.1007/978-3-031-19818-2_19]

[47] Kweon H, Yoon SH, Yoon KJ. Weakly supervised semantic segmentation via adversarial learning of classifier and reconstructor. In: Proc. of the 2023 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Vancouver: IEEE, 2023. 11329–11339. [doi: 10.1109/CVPR52729.2023.01090]

[48] Rong SH, Tu BH, Wang ZL, Li JJ. Boundary-enhanced co-training for weakly supervised semantic segmentation. In: Proc. of the 2023

IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Vancouver: IEEE, 2023. 19574–19584. [doi: 10.1109/CVPR52729.2023.01875]

[49] Cheng ZS, Qiao PC, Li KH, Li SH, Wei PX, Ji XY, Yuan L, Liu C, Chen J. Out-of-candidate rectification for weakly supervised semantic segmentation. In: Proc. of the 2023 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Vancouver: IEEE, 2023. 23673–23684. [doi: 10.1109/CVPR52729.2023.02267]

附中文参考文献:

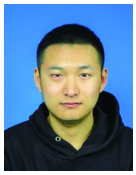
[1] 白琮, 黄玲, 陈佳楠, 潘翔, 陈胜勇. 面向大规模图像分类的深度卷积神经网络优化. 软件学报, 2018, 29(4): 1029–1038. http://www.jos.org.cn/1000-9825/5404.htm [doi: 10.13328/j.cnki.jos.005404]

[2] 田萱, 王亮, 丁琪. 基于深度学习的图像语义分割方法综述. 软件学报, 2019, 30(2): 440–468. http://www.jos.org.cn/1000-9825/5659.htm [doi: 10.13328/j.cnki.jos.005659]

[12] 石泽男, 陈海鹏, 张冬, 申铉京. 预训练驱动的多模态边界感知视觉 Transformer. 软件学报, 2023, 34(5): 2051–2067. http://www.jos.org.cn/1000-9825/6768.htm [doi: 10.13328/j.cnki.jos.006768]

李军侠(1985－), 女, 博士, 副教授, CCF 专业会员, 主要研究领域为计算机视觉, 人工智能.

崔滢(1986－), 女, 博士, 副教授, CCF 专业会员, 主要研究领域为计算机视觉, 人工智能.



苏京峰(1997－), 男, 硕士生, CCF 学生会员, 主要研究领域为计算机视觉.

刘青山(1975－), 男, 博士, 教授, 主要研究领域为图像与视频理解, 机器学习, AI+气象.