

基于自监督辅助 CLIP 与 Transformer 自注意力的 提示引导式弱监督语义分割方法

魏龙生^{1,2} 李唐强^{1,2} 徐 朗^{1,2} 罗大鹏³

摘 要 针对传统图像级标注的弱监督语义分割方法仍依赖目标数据集的图像级标注,进而难以获取精细空间监督信号的问题,提出一种无需目标域标注的自监督对比语言-图像预训练提示引导方法. 主要解决三个问题:一是传统弱监督缺乏像素级空间约束;二是初始伪标签受限于分类特征的传导偏差;三是预训练特征与分割任务存在语义失配. 对此,提出三点解决方案:1) 设计自监督提示学习机制,替代传统固定的文本模板,利用跨模态注意力动态生成包含图像上下文的可学习提示向量;2) 提出基于损失响应梯度的动态裁剪策略,利用深度神经网络的“早期记忆效应”过滤高梯度的噪声区域,优化伪标签边界精度;3) 构建基于 Transformer 自注意力的类感知亲和图,增强分割结果的空间一致性. 在 PASCAL VOC 2012 和 MS COCO 2014 数据集上的实验结果表明,所提方法在未见目标域标注的设定下, mIoU 指标优于现有的无人工标注弱监督语义分割方法,且在边界保持与背景抑制方面具有显著优势.

关键词 弱监督语义分割; 语言-图像预训练; 自监督学习; 提示学习; 梯度裁剪

引用格式 魏龙生, 李唐强, 徐朗, 罗大鹏. 基于自监督辅助 CLIP 与 Transformer 自注意力的提示引导式弱监督语义分割方法. 自动化学报, 2026, 52(9): 2032–2045

DOI 10.16383/j.aas.c250762 **CSTR** 32138.14.j.aas.c250762

Prompt-guided Weakly Supervised Semantic Segmentation Method Based on Self-supervised Auxiliary CLIP and Transformer Self-attention

WEI Long-Sheng^{1,2} LI Tang-Qiang^{1,2} XU Lang^{1,2} LUO Da-Peng³

Abstract To address the problem that traditional weakly supervised semantic segmentation methods relying on image-level annotations of target datasets still fail to obtain fine-grained spatial supervision signals, this paper proposes a self-supervised contrastive language-image pre-training prompt-guided method that requires no target domain annotations. This paper mainly tackles three key issues: First, the lack of pixel-level spatial constraints in traditional supervision methods; second, the limitation of initial pseudo-labels due to the propagation bias of classification features; third, the semantic mismatch between pre-trained features and segmentation tasks. In response to these problems, three corresponding solutions are proposed. 1) A self-supervised prompt learning mechanism is designed to replace traditional fixed text templates. By using cross-modal attention to dynamically generate learnable prompt vectors containing image context. 2) A dynamic clipping strategy based on loss response gradients is put forward. It filters out high-gradient noise regions by utilizing the “early memory effect” of deep neural networks, thus optimizing the boundary accuracy of pseudo-labels. 3) A class-aware affinity graph based on Transformer self-attention is constructed to enhance the spatial consistency of segmentation results. Experimental results on PASCAL VOC 2012 and MS COCO 2014 datasets show that, under the setting without target domain annotations, the proposed method outperforms existing label-free weakly supervised semantic segmentation methods in terms of mIoU metric, and has significant advantages in boundary preservation and background suppression.

Keywords weakly supervised semantic segmentation; contrastive language-image pre-training; self-supervised learning; prompt learning; gradient clipping

Citation Wei Long-Sheng, Li Tang-Qiang, Xu Lang, Luo Da-Peng. Prompt-guided weakly supervised semantic segmentation method based on self-supervised auxiliary CLIP and Transformer self-attention. *Acta Automatica Sinica*, 2026, 52(9): 2032–2045

收稿日期 2025-12-31 录用日期 2026-03-04

Manuscript received December 31, 2025; accepted March 4, 2026

国家自然科学基金 (62573393), 空间智能控制技术全国重点实验室开放基金课题 (HTKJ2025KL502006) 资助

Supported by National Natural Science Foundation of China (62573393) and State Key Laboratory of Space Intelligent Control (HTKJ2025KL502006)

本文责任编辑 樊彬

Recommended by Associate Editor FAN Bin

1. 中国地质大学 (武汉) 人工智能与自动化学院 武汉 430074
2. 复杂系统先进控制与智能化湖北省重点实验室 武汉 430074
3. 中国地质大学 (武汉) 机械与电子信息学院 武汉 430074

1. School of Artificial Intelligence and Automation, China University of Geosciences, Wuhan 430074 2. Hubei Key Laboratory of Advanced Control and Intelligent Automation for Complex Systems, Wuhan 430074 3. School of Mechanical Engineering and Electronic Information, China University of Geosciences, Wuhan 430074

在图像理解任务中, 语义分割旨在为每个像素分配语义标签, 实现细粒度感知^[1-2]. 尽管全监督方法性能优异, 但其对像素级标注的依赖导致成本高昂, 限制了实际应用^[3]. 因此, 弱监督语义分割 (weakly supervised semantic segmentation, WSSS) 因其可扩展性受到广泛关注^[4]. 其中, 基于图像级标签的方法因标注成本极低而成为主流, 它仅需标注图像中存在的类别, 而无需提供位置信息^[5]. 然而, 这种监督形式存在根本性局限: 图像级标签缺乏对象的空间位置与轮廓信息, 导致模型难以生成完整且准确的对象区域^[6-7]. 为解决该类问题, 传统方法通常采用复杂的多阶段流程, 首先生成粗糙的类激活图 (class activation map, CAM), 再通过亲和传播等方法进行细化以得到伪标签, 最终训练分割模型^[8-9]. 这一范式的流程较为繁琐、计算成本高, 且生成的伪标签常出现区域不完整和背景混淆等问题.

近年来, 对比语言-图像预训练 (contrastive language-image pre-training, CLIP) 模型的兴起为 WSSS 提供了新思路^[10]. CLIP 通过在海量图文对上进行对比学习, 构建了强大的跨模态语义对齐能力^[9]. 在 WSSS 任务中, 图像与其对应的类别标签天然构成一组图文配对数据. 这一固有属性使得 CLIP 模型所蕴含的跨模态语义先验得以充分利用, 进而为伪标签生成过程提供有效的语义约束, 显著提升伪标签的类别一致性与整体质量^[11]. 基于 CLIP 的方法致力于从其潜在空间中自动挖掘类别语义与上下文关系^[12], 通过优化 CAM 生成、引入类感知注意力传播以及设计置信度感知的损失机制^[13], 显著提升了伪标签的完整性并抑制了背景噪声. 然而, 现有主流方法大多仍需依赖目标数据集的图像级标注来指导模型区分特定类别, 未能充分挖掘无标注数据中的自监督结构信息, 限制了其在零人工标注场景下的应用潜力.

针对目标域人工标注依赖及预训练特征域漂移问题, 本文提出一种基于自监督适应 CLIP 的提示引导式 WSSS 方法. 主要创新如下: 1) 引入自监督适应机制, 通过无标注目标数据上微调 CLIP 视觉编码器, 在保留通用表征的同时适配特定任务分布, 有效缓解了直接迁移导致的语义失配; 2) 设计动态上下文提示学习模块, 从图像特征中推理上下文语义以生成可学习提示, 摆脱对手工模板及图像级标签的依赖, 实现了精准的无监督语义激活; 3) 构建基于 Transformer 自注意力的类感知亲和图与梯度统计噪声抑制机制, 利用深度网络的早期记忆效应过滤噪声梯度, 显著提升了伪标签的空间一致性与边界精度^[14]. PASCAL VOC 2012 与 MS COCO

2014 数据集上的实验表明, 该方法在无需目标域人工标注设定下的平均交并比 (mIoU) 优于现有同类方法, 且在边界保持与背景抑制方面优势显著^[15].

1 相关工作

WSSS 通过低成本监督信息完成像素级语义分割任务, 在标注成本高昂的领域展现出显著的应用价值. 然而, 传统 WSSS 方法普遍面临两大核心局限: 一是语义覆盖范围不完整, 难以兼顾图像中所有类别目标; 二是非显著目标定位精度不足, 无法精准捕获视觉占比低或特征不突出的目标区域. CLIP 模型所具备的强大跨模态语义对齐能力, 为突破传统 WSSS 的技术瓶颈提供了全新范式, 现已成为该领域的核心支撑工具. 当前基于 CLIP 的 WSSS 研究主要沿三大方向逐步推进: 其一, 聚焦 CLIP 模型本身的优化, 重点提升其区域感知能力与跨模态 (视觉-语言) 对齐的适配性; 其二, 融合自监督学习策略, 利用大规模无标注数据为弱监督场景学习高判别力特征表示; 其三, 直接基于 CLIP 构建端到端 WSSS 框架. 尽管上述研究取得了阶段性进展, 但现有方案的核心环节仍未脱离对人工定义类别标签或固定文本提示的依赖, 导致模型的视觉语义推理潜力未能得到充分释放. 本文立足上述研究脉络, 以突破人工监督依赖为核心目标, 探索自监督训练与动态提示生成的深度融合机制.

1.1 CLIP 图像文本模型

当前基于 CLIP 的弱监督语义分割研究, 主要致力于解决其在像素级解析中区域感知不足与图文语义粒度不匹配的问题, 以增强对局部语义的捕获能力和对非显著区域的定位精度. 其中, Long-CLIP^[16] 针对 CLIP 文本编码长度受限的问题, 提出保持语义一致性的扩展机制, 在支持更长文本描述输入的同时, 保留了原始语义空间结构, 为引入多属性或复杂场景描述以构建细粒度伪标签提供了文本维度的支持. Alpha-CLIP^[17] 通过引入表示用户指定关注区域的 Alpha 通道, 并利用大规模 RGBA-文本对进行训练, 在不改变图像原始内容的基础上实现区域引导, 显著提升了模型的区域感知能力, 可作为伪标签增强器以改善难分区域的分割效果. TF-CLIP^[18] 提出一种去除文本依赖的视觉 CLIP 架构, 借助 CLIP-Memory 模块与时间记忆扩散机制, 强化了跨帧一致性语义建模能力, 为 CLIP 在缺少文本标注的弱监督场景下适应低资源环境提供了有效策略. WeakCLIP^[19] 结合可学习提示结构、金字塔多尺度适配器与跨模态协同注意力机制, 通过引导文本特征对局部图像区域产生精准响应, 优

化了 CLIP 在语义粒度、空间分辨率及视觉引导方面的局限,提升了伪标签的语义完整性与边界保真度. DeCLIP^[20] 则探索了 CLIP 中间层特征在空间解码中的潜力,通过引入轻量级卷积解码器对高语义层表示进行空间还原,实现了伪造区域的精确定位,其冻结特征与解码结构的设计为 CLIP 在像素级任务中的适配提供了借鉴.

综上所述,现有方法从增强区域感知与提升模态对齐一致性两个方向推进^[21],而本文在此基础上,进一步探讨如何构建高质量的 CLIP 引导伪标签机制,结合区域注意、提示调控与空间引导,以实现更精细且语义一致的弱监督语义分割结果.

1.2 自监督学习

当前自监督学习方法主要致力于解决编码器在无监督条件下结构感知弱、语义对齐能力不足的核心问题,旨在为下游密集预测任务(如语义分割)学习更具判别力与泛化性的特征表示.具体而言,Lim 等^[22]提出一种融合图神经网络的自监督表征学习策略,利用图结构增强局部邻域建模能力,并通过结构保持约束提升模型在复杂区域中的语义一致性,为视觉编码器的结构建模提供了借鉴. Zhang 等^[23]提出一个面向时间序列的统一自监督建模框架,定义“任务-语义一致性”准则以衡量预训练任务对下游任务迁移的影响,其强调的语义结构保持性为视觉-语言模型的预训练机制提供了理论支撑. Chen 等^[24]通过引入低分辨率扰动与多尺度特征对齐机制,引导模型在无监督训练中关注局部细节与边缘,从而强化语义边界感知能力,为分割任务中的边界建模提供了有效支持. Gui 等^[25]则系统地归纳了多模态自监督学习中的五类典型预训练任务,并强调任务语义的一致性与结构保持性是提升模型迁移性能的关键,为构建通用编码器提供了框架性指导.

综上所述,当前自监督学习方法在结构感知、语义对齐与特征迁移等方面取得了显著进展,为 CLIP 等多模态模型在低标注场景下的训练提供了重要基础. 本文基于上述研究成果,结合多尺度重建、结构保持与语义一致性机制,探索以自监督学习方式训练 CLIP 图像编码器,从而在无需大规模图文对监督的前提下提升其特征判别力,为后续弱监督语义分割中的高质量伪标签生成提供更优的视觉表征支撑.

1.3 基于图像级标签的弱监督语义分割

近两年,基于 CLIP 的弱监督语义分割方法虽取得进展,但其主流方案仍普遍依赖人工构造的类

别标签或提示,未能充分释放视觉模型自身的结构感知与语义推理潜力,导致提示泛化性弱、空间定位粗糙等问题. 现有研究主要沿以下方向展开.

- 将 CLIP 作为分割主干以简化流程的方法,如 FrozenCLIP^[26] 在冻结视觉编码器基础上构建解码器进行伪标签精炼, Prompt-CLIP 类方法通过手工提示与 GradCAM 机制提取定位图,在效率与性能间取得平衡,但其提示表达能力仍依赖于固定词汇集.

- 直接采用冻结 CLIP 生成伪标签的方法,如 WeCLIP^[27] 引入 SAM 进行掩码级优化, CLIMS++^[28] 通过图文匹配监督与上下文发现模块提升目标区域完整性,虽改善了初始定位效果,但仍受限于预训练 CLIP 的语义对齐粒度.

- 利用生成模型扩展数据与伪标签质量的工作,如 SeCo^[29] 通过多粒度语义对比提高伪标签判别性, DuPL^[30] 则借助双学生框架缓解确认偏差, GPCD^[31] 结合 GPT 与扩散模型生成多样训练数据,此类方法提升了模型鲁棒性,但整体流程复杂且仍以人工定义为前提.

- 面向提示语义增强的研究,如 Yoon 等^[32] 通过类标签注入机制增强一致性并抑制背景, Wu 等^[33] 利用 GPT 生成类别聚类信息以缓解语义歧义,这些方法虽强化了类别引导,却未突破对文本监督的依赖.

综上,当前基于 CLIP 的 WSSS 方法在效率与性能上虽有提升,但其核心仍依赖于外部提供的类别标签或固定提示,未能实现视觉模型自身语义表征能力的充分挖掘与泛化. 为此,本文提出对 CLIP 进行自监督视觉训练,使其无需任何人工标注即可学习泛化性更强的视觉语义表示,进而动态生成高质量语义提示,以此引导伪标签生成与优化,最终构建一个无需外部类别监督、简洁而高效的弱监督语义分割框架.

2 网络模型

现有基于 CLIP 的弱监督语义分割方法大多依赖预设的静态文本模板(如“a photo of [CLASS]”)进行跨模态对比,这种粗糙的提示形式往往忽略目标与背景间的语义共现关系,极易导致共现背景(如“rail”)被错误激活,严重影响伪标签的定位质量与边界准确性. 为缓解这一问题,一些方法尝试人为构造背景类提示(如“a photo of rail”)以显式抑制干扰区域,但这种方式高度依赖人工先验与语义知识,缺乏可扩展性与泛化能力.

针对上述局限,本文提出一种仅需输入图片、

无需任何图像级标签的自监督 CLIP 训练框架, 实现从提示生成到伪标签优化的完整引导. 具体而言 (如图 1 所示), 本文首先通过自监督机制增强 CLIP 的视觉表征与上下文建模能力, 进而动态生成与图像内容适配的前景及背景语义提示, 彻底摆脱对固定模板与人工先验的依赖; 在此基础上, 设计基于损失响应的梯度裁剪机制, 在训练中施加结构约束, 以保留高置信语义区域、抑制共现背景干扰, 从而显式提升边界建模质量; 同时, 该框架进一步引入语义感知的提示细化策略, 增强模型在类间混淆与边缘模糊区域的判别力. 本文实现了从静态模板依赖到动态上下文感知的范式转变, 在降低人工依赖的同时, 显著提升了伪标签的定位准确性与边界一致性.

2.1 CLIP 自适应文本训练

CLIP 具有视觉骨干网络和文本编码器. 为充分利用其多尺度视觉特征, 如图 2 所示, 引入可学习单元 $P_v, \dots, P_m$, 其中 P_i ($i = v, \dots, m$) 是视觉空间中的投影器, 用于提取来自原图与增强图中

的视觉特征以获得令牌嵌入; MoCo V3 是自监督对比学习框架. CLIP 目标是通过训练使共享的嵌入空间具有可泛化性、判别性, 并对图像变化具有不变性, 从而能自动构造背景提示词.

2.1.1 自适应文本训练

在本方法中, 通过解耦图像的内容与风格成分实现从视觉特征中生成提示. 该过程的具体流程定义如下: 使用 OpenAI 发布的官方预训练 CLIP ViT-B/16 模型权重初始化视觉编码器与文本编码器. 训练数据来源于分割数据集 (PASCAL VOC 2012 和 MS COCO 2014) 的无标注训练图片 $D_{train} = \{x_i\}_{i=1}^N$, N 为批量大小. 需要强调的是, 在此阶段本文仅使用图像数据, 丢弃所有人工提供的图像级类别标签和像素级掩码, 完全遵循无监督设定.

首先, 为提取具有泛化性的视觉特征, 本文使用来自不同层的多尺度内容特征, 通过拼接得到如下表示:

$$F_{content} = \text{Concat}(F_l^{(1)}, F_l^{(2)}, \dots, F_l^{(L)}) \quad (1)$$

其中, $F_{content}$ 表示内容特征; $\text{Concat}(\cdot)$ 表示拼接

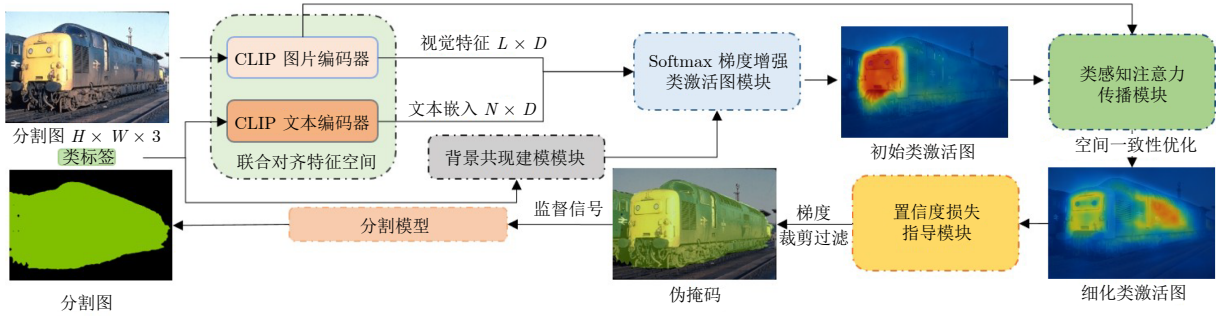


图 1 模型框架

Fig.1 Model backbone

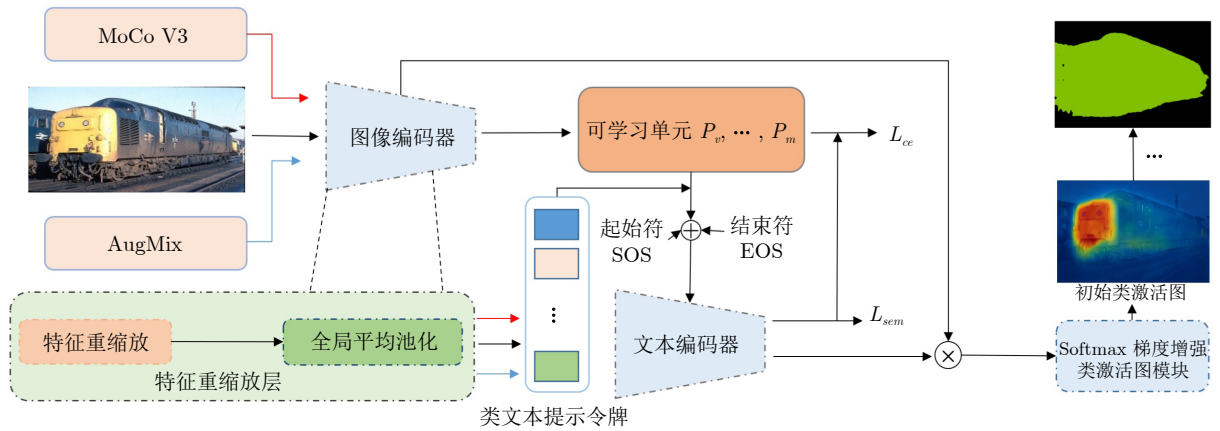


图 2 CLIP 模型框架

Fig.2 CLIP model backbone

函数; $F_l^{(k)}$ 表示第 k 层的内容特征; L 为特征层数. 该多尺度表示的关键在于包含不同层次的低、中、高级特征, 因此相比于仅使用高级语义特征, 其迁移能力更强. 类似地, 使用实例级特征统计表示风格特征, 即通道维度上的均值与标准差, 这些特征来源于第 L 层, 定义为:

$$F_{style} = [\mu(F_L^{(L)}), \sigma(F_L^{(L)})] \quad (2)$$

其中, $\mu(\cdot)$ 和 $\sigma(\cdot)$ 分别表示通道均值与标准差计算; F_{style} 为风格向量.

为将视觉特征映射到提示空间, 图像 x 的初始提示令牌 T_{init} 通过线性投影计算:

$$T_{init} = W_{proj}(F_{content} \oplus F_{style}) + b_{proj} \quad (3)$$

其中, $\oplus$ 表示特征拼接操作; $W_{proj} \in \mathbf{R}^{D \times d}$ 和 b_{proj} 分别为线性投影层的可学习权重矩阵与偏置参数, D 为拼接后的特征总维度, d 为 CLIP 文本编码器的输入维度.

接着, T_{init} 通过特征重缩放层 (FRG) 处理后, 输入到包含 1 个编码器与 M 个解码器的投影器 P_t 中. 通过 P_t 学习 M 个离散的上下文相关提示令牌. 最终, 对于类别 y , 其输入文本编码器的完整提示序列 T_y 定义为:

$$T_y = [v_1, v_2, \dots, v_M, E_y] \quad (4)$$

其中, $v_m = \text{Decoder}_m(P_t)$ 表示由第 m 个解码器生成的第 m 个可学习上下文向量 (context vector); E_y 是类别 y 名称对应的固定词嵌入 (word embedding); $[\cdot]$ 表示序列拼接. 这种设计使得提示不仅包含类别语义, 还融合了图像特定的视觉上下文信息.

为实现自监督适应, 引入伪标签对齐损失 L_{ce} 和语义一致性损失 L_{sem} . 利用冻结的预训练 CLIP 生成高置信度伪标签 $\hat{y}$ 作为监督信号, L_{ce} 定义为:

$$L_{ce} = -\frac{1}{N} \sum_{i=1}^N \sum_{c=1}^C \mathbb{I}(\hat{y}_i = c) \log p(c|x_i) \quad (5)$$

其中, $\mathbb{I}(\cdot)$ 为指示函数; C 为类别数; $p(\cdot|x_i)$ 为后验概率. 同时, 为增强鲁棒性, 对图像施加 AugMix 增强得到 x_{aug} , 并最小化其与原图提示嵌入的距离:

$$L_{sem} = \frac{1}{N} \sum_{i=1}^N \|T(x_i) - T(x_{aug})\|_2^2 \quad (6)$$

其中, $T(\cdot)$ 表示图像的提示嵌入. 最终的总训练目标为 $L_{total} = L_{ce} + \lambda L_{sem}$, 其中 λ 为平衡系数.

2.1.2 图像增强与损失函数

本方法利用对比学习将视觉与文本特征映射到共享的嵌入空间中. 为计算输入 x 属于类别 y 的后

验概率, 采用以下定义 (τ 表示温度超参数):

$$p(y|x) = \frac{\exp\left(\frac{\text{Sim}(P_v(F_x), T_y)}{\tau}\right)}{\sum_{y'=1}^C \exp\left(\frac{\text{Sim}(P_v(F_x), T_{y'})}{\tau}\right)} \quad (7)$$

其中, $\text{Sim}(\cdot, \cdot)$ 表示余弦相似度计算; F_x 为图像 x 的视觉特征; $P_v(F_x)$ 视为视觉嵌入 (该空间促进视觉不变性). 随后, 在每个输入图像的预测概率与对应类别标签之间计算交叉熵损失:

$$L_{ce} = -\frac{1}{N} \sum_{i=1}^N \sum_{y=1}^C \mathbb{I}(y_i = y) \log p(y|x_i) \quad (8)$$

其中, y_i 为图像 x_i 的真实类别标签. 为扩展从不同图像增强方式中获得一致的提示嵌入, 对每个图像采用 AugMix 增强——通过几何变换 (旋转、翻转) 及组合多种数据增强 (亮度变化、模糊), 并使用随机权重混合来增强结果, 实现语义不变下的数据增强. 引入蒸馏损失实现语义一致性, 该损失基于 L_2 范数距离定义: 将原始图像 x 的提示嵌入视为教师模型, 增强图像 x_{aug} 的提示嵌入视为学生模型, 损失函数如下:

$$L_{sem} = \frac{1}{N} \sum_{i=1}^N \|T_{x_i} - T_{x_{i,aug}}\|_2^2 \quad (9)$$

其中, T_{x_i} 和 $T_{x_{i,aug}}$ 分别为原始图像与增强图像的提示嵌入. 最终训练模型的总损失函数为:

$$L_{total} = L_{ce} + \lambda L_{sem} \quad (10)$$

其中, λ 为平衡系数.

2.2 CLIP 细化类激活图

为有效剔除掩码中与目标类别共现的背景区域, 引入提示引导的语义精炼模块. 该模块利用已学习的背景提示引导掩码生成器在生成精细化掩码时排除背景语义信息, 同时保留原始类别匹配目标, 确保对前景类别的对齐. 掩码 M_{init} (由优化 GradCAM 生成) 和 CLIP 生成的提示 T_y 作为融合模块的输入. 首先, M_{init} 沿通道方向被均匀划分为四个掩码特征图 M_1, M_2, M_3, M_4 ; 然后, 对每个分支的特征图应用不同膨胀率 $d = 1, 2, 3, 4$ 的 3×3 膨胀卷积, 以学习不同尺度的特征:

$$M'_i = \text{DilatedConv}(M_i, d = i) \quad i = 1, 2, 3, 4 \quad (11)$$

其中, $\text{DilatedConv}(\cdot)$ 表示膨胀卷积. 不同分支通过逐像素级加法操作相互关联得到多尺度特征:

$$M_{multi-scale} = M'_1 + M'_2 + M'_3 + M'_4 \quad (12)$$

为捕获不同尺度的注意力信息, 对输入特征图

应用全局平均池化 (GAP)、卷积和 Sigmoid 激活, 得到通道级自适应注意力图:

$$A = \text{Sigmoid}(\text{Conv}(\text{GAP}(M_{\text{multi-scale}}))) \quad (13)$$

其中, $\text{GAP}(\cdot)$ 为全局平均池化操作; $\text{Conv}(\cdot)$ 为 1×1 卷积层. 利用注意力图动态融合每个分支的特征图, 并通过跳跃连接防止信息丢失:

$$M_{\text{refine}} = \text{Concat}(A \odot M'_1, A \odot M'_2, A \odot M'_3, A \odot M'_4) + M_{\text{init}} \quad (14)$$

其中, $\text{Concat}(\cdot)$ 表示沿通道维度拼接; $\odot$ 表示逐元素相乘; M_{refine} 为生成的精细化掩码. 引入精炼损失 L_{refine} , 与之前的匹配损失 L_{match} (图文匹配损失) 共同构成总损失函数:

$$L_{\text{cam}} = L_{\text{match}} + \gamma L_{\text{refine}} \quad (15)$$

其中, $L_{\text{refine}} = \text{CE}(M_{\text{refine}}, M_{\text{pseudo}})$, M_{pseudo} 为初始伪标签, $\text{CE}(\cdot, \cdot)$ 为像素级交叉熵损失; γ 为平衡系数. 通过这种方式, 可利用固定背景提示抑制前景掩码中潜在的类相关背景误激活区域; 同时借助图像级监督联合语义匹配与背景抑制目标, 提升分割区域与类别语义的对齐效果, 生成紧凑、完整的目标掩码. 需注意的是, CLIP 模型和提示向量仅在训练阶段用于引导掩码生成器的学习, 推理阶段仅保留掩码生成器, 以提升整体效率与实用性.

2.3 梯度裁剪再细化类激活图

如图 3 所示, 首先将类别标签 Q 与图像块标签 T 一起输入 Transformer 编码器层, 获得相应输出 T'_{cls} 和 T'_{patch} ; 然后使用 L_2 归一化的类令牌范数和补丁令牌范数生成预测序列 Z ; 对 Z 进行 Layer-Norm (LN) 并上采样至图像原始分辨率:

$$Z_{\text{upsample}} = \text{Upsample}(\text{LN}(Z), \text{尺寸为 } H \times W) \quad (16)$$

其中, $\text{Upsample}(\cdot)$ 表示上采样; Z_{upsample} 表示上采样得到的结果; $H \times W$ 为输入图像的原始分辨率.

在弱监督语义分割任务中, 初始伪标签不可避免地包含噪声, 典型表现为背景区域的误激活 (如将“水面”误判为“船”). 根据深度神经网络的记忆效应, 模型在训练初期倾向于优先拟合具有清晰语义模式的简单样本, 此时梯度主要由有效特征驱动; 而在训练后期, 模型为了拟合伪标签中的噪声 (如错误标注), 必须大幅调整参数, 从而在反向传播中产生异常高梯度. 本文将像素级的梯度幅值作为衡量标签噪声置信度的指标, 高梯度往往对应伪标签中的误激活区域, 而低梯度则对应高置信度的正确预测. 因此, 梯度裁剪解码器的核心目标是构建一个动态滤波器, 通过抑制高梯度回传, 阻断模型

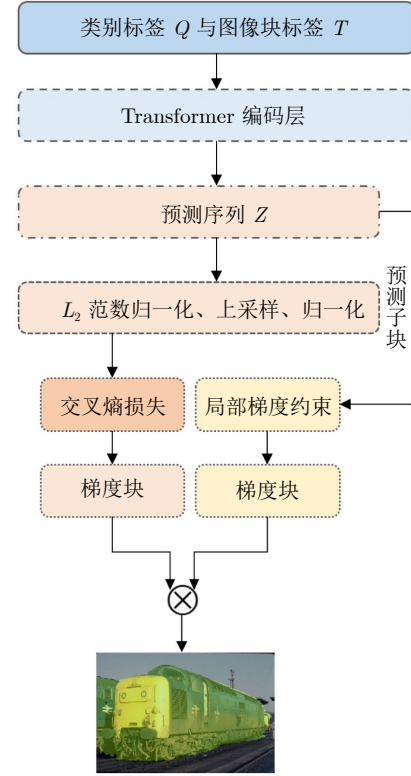


图 3 图像梯度裁剪器

Fig.3 Image gradient clipper

对伪标签噪声的强保留, 同时保留有效特征的学习能力.

这需要平衡两个互补的约束因素: 反映图像整体噪声水平的全局约束与反映区域纹理复杂度的局部约束. 为计算这两种约束, 首先将预测结果 Z_{upsample} 划分为 $K \times K$ 个不重叠子块 (子块尺寸为 $s \times s$, $K \times s = H$). 结合生成的 CAM M_{refine} 及预测子块 Z , 计算梯度子块; 每个梯度子块内的平均梯度 $\bar{G}_{k,l}$ 即为局部约束:

$$\bar{G}_{k,l} = \frac{1}{s^2} \sum_{i=1}^s \sum_{j=1}^s \nabla_{Z_{k,l,i,j}} \text{CE}(Z_{k,l,i,j}, M_{\text{refine},k,l,i,j}) \quad (17)$$

其中, k, l 表示子块坐标. 相应地, 所有子块梯度的均值构成了全局约束:

$$\bar{G}_{\text{global}} = \frac{1}{K^2} \sum_{k=1}^K \sum_{l=1}^K \bar{G}_{k,l} \quad (18)$$

为平衡这两者, 本文采用松弛约束策略, 选择 $\bar{G}_{k,l}$ 和 $\bar{G}_{\text{global}}$ 中的最大值作为最终的动态裁剪阈值 $\bar{G}_{\text{thr}}$:

$$\bar{G}_{\text{thr}}(k, l) = \max(\bar{G}_{k,l}, \bar{G}_{\text{global}}) \quad (19)$$

该阈值设定的理论依据在于: 图像中的高频细

节 (如物体轮廓和纹理边界) 天然具有较大的局部梯度. 若仅以较低的全局平均梯度 $\bar{G}_{\text{global}}$ 作为硬性阈值, 这些关键的结构信息极易被错误地作为噪声裁剪掉. 通过引入局部梯度 $\bar{G}_{k,l}$ 并取两者的最大值, 实际上是为高复杂度区域提供了更高的梯度容忍度, 从而在有效抑制背景突变噪声的同时, 最大程度地保护了目标的边界结构信息.

此外, 为保证训练稳定性, 必须引入梯度裁剪启动阈值 ϵ . 在训练初期, 模型尚未收敛, 此时的高梯度反映的是模型对未知特征的探索而非噪声拟合, 过早裁剪会导致欠拟合. 本文基于对训练过程中平均梯度的统计观察, 将 ϵ 设置为 0.10. 该数值对应于损失函数曲线趋于平缓的收敛拐点. 仅当前 batch 的全局平均梯度 $\bar{G}_{\text{batch}} < \epsilon$ (即模型进入精细调整阶段, 大梯度更可能源于对噪声的强行拟合) 时, 才激活梯度裁剪机制:

$$G_{\text{clip}}(k, l, i, j) = \begin{cases} G(k, l, i, j), & G(k, l, i, j) \leq \bar{G}_{\text{thr}}(k, l) \\ 0, & \text{其他} \end{cases} \quad (20)$$

其中, $G(k, l, i, j)$ 为原始梯度; G_{clip} 为稀疏化后的梯度. 最后, 反向传播掩码梯度的平均值, 通过这种动态选择机制, 迫使网络优先从梯度较小、置信度较高的区域学习, 逐步修正伪标签.

3 实验

3.1 数据集和评价指标

在对 CLIP 模型类泛化与跨数据集迁移能力实验中, 本文主要针对 ImageNet^[34]、Caltech 101^[35]、OxfordPets^[36]、StanfordCars^[37]、Flowers 102^[38]、Food-101^[39]、FGVCAircraft^[40]、SUN 397^[41]、DTD^[42] 9 个图像识别数据集进行分析; 同时在域泛化实验中, 以 ImageNet 作为源数据集, 使用 ImageNet V2^[43]、ImageNet-Sketch^[44] (50 000 张图像)、ImageNet-A^[45] 和 ImageNet-R^[46] 作为目标域. CLIP 核心架构由两层瓶颈网络 (Linear-ReLU-Linear) 和单层带批归一化的 MLP 组成, 视觉和文本编码器均采用 CLIP 预训练的 Transformer 主干 (ViT-B/16). 训练阶段设置: 50 个 epoch, 学习率调度 (初始学习率为 1×10^{-4} , 每 10 个 epoch 衰减 0.5), 每类采样 16 个镜头, 批量大小为 4, 文本提示初始化为 “a photo of [CLASS]”, 上下文长度设为 4; 评价指标为三次运行的平均 top-1 准确率. 分割模型上, 在 PASCAL VOC 2012^[47] 和 MS COCO 2014^[48] 数

据集进行系统评估. PASCAL VOC 2012, 包含训练集 (1 464 张)、验证集 (1 449 张)、测试集 (1 456 张), 涵盖 20 个前景类别和 1 个背景类别; 引入扩展集 (10 582 张) 及 SBD^[49] 数据增强训练多样性. MS COCO 2014, 包含训练图像 (82 081 张)、验证图像 (40 137 张), 涵盖 80 个对象类别和 1 个背景类别, 目标类别更丰富, 覆盖日常场景. 两个数据集均采用 mIoU 作为核心评价指标.

3.2 实验结果

3.2.1 CLIP 表征能力验证

虽然本文研究重点在于弱监督语义分割, 但是为了验证本文提出的自监督适应机制是否在优化分割性能的同时保持了 CLIP 预训练权重的通用泛化能力, 本文评估了模型在图像分类任务上的表现. 同数据集类别泛化、跨数据集泛化 (11 个数据集) 以及域泛化 (5 个变体) 的实验结果汇总于表 1 中.

表 1 CLIP 泛化能力综合评估 (%)
Table 1 Comprehensive evaluation of CLIP generalization capabilities (%)

方法	同数据集泛化		跨数据集泛化	域泛化
	基础类	新类		
CLIP ^[50]	69.34	74.22	66.89	59.08
SLIP ^[51]	69.78	74.26	67.22	59.38
CoOp ^[52]	82.57	63.25	66.35	61.74
CoCoOp ^[53]	80.31	71.39	67.92	62.16
MaPLe ^[54]	82.07	75.16	68.24	62.36
StyLip ^[55]	82.97	75.86	69.64	63.02
本文方法	83.16	77.14	69.89	63.83

如表 1 所示, 在同数据集泛化任务中, 本文在基础类和新类上均取得了最优结果. 更重要的是, 在衡量通用性的跨数据集泛化和域泛化任务中, 本文方法分别取得了 69.89% 和 63.83% 的平均准确率, 均优于 CoOp、MaPLe 等基线方法. 这有力地证明了本文的适应机制不仅成功适配了分割任务, 还通过自监督一致性约束有效抵抗了特征漂移, 保留并增强了 CLIP 的通用视觉表征能力.

3.2.2 分割模型验证

为全面验证本文提出的提示引导弱监督语义分割框架在实际任务中的有效性, 本文对比评估了最终训练出的分割模型在 PASCAL VOC 2012 与 MS COCO 2014 两个数据集上的性能, 平衡系数 γ 设置为 0.3.

在 PASCAL VOC 2012 数据集上, 如表 2 所示, 模型所生成的 CAM 及对应伪真实掩码评估得

表 2 PASCAL VOC 2012 训练集上的 CAM 及对应伪真实掩码评估 (%)

Table 2 CAM and corresponding pseudo-ground truth mask evaluation on the PASCAL VOC 2012 training set (%)

方法	CAM	掩码
MCTformer ^[56]	61.7	69.1
CLIMS ^[57]	56.6	70.5
CLIP-ES ^[58]	58.6	75.0
FPR ^[59]	63.8	66.4
WeakCLIP ^[19]	—	74.8
POLE ^[60]	59.0	74.2
DuPL ^[30]	—	73.0
MCC ^[61]	57.7	71.8
HMCTANet ^[62]	64.5	75.8
本文方法	64.9	78.1

分并不弱于其他基于图像级标签标注的模型. 定量评估结果如表 3 所示, 模型在验证集和测试集上分别取得了 71.3% 和 71.7% 的平均交并比. 在与不依赖图像级标签的纯图像弱监督方法的对比中 (表 3 下半部分), 本文方法取得了显著的领先优势. 相较于 MaskCLIP、MaskCLIP+^[63] 以及 PixelCLIP^[1], 验证集上的性能提升超过 4 个百分点. 这直接验证了基于自监督训练增强 CLIP 视觉表征并以此动态生成语义提示的路径, 能够比现有无需标签的方案更有效地驱动高质量分割. 尽管不使用图像级标签, 但本文方法的性能却达到了与主流依赖图像级标签的弱监督方法 (表 3 上半部分) 相当甚至更优的水平. 本文的结果超越了 SEAM、AdvCAM、OC-CSE 和 CLIMS^[57] 等基于 ResNet 或 ViT 的经典工作, 并且与同样采用 Transformer 架构的 ToCo 和采用 ViT-PCM 的基线相比, 本文方法在验证集上分别取得了 1.5 和 1.0 个百分点的稳定提升, 这凸显了本文框架在复杂目标语义建模与上下文感知上的优越性.

如图 4 所示, 本文展示了若干 PASCAL VOC 2012 验证集中具有代表性的图像分割可视化结果, 涵盖人、车、动物、背景物体等多种场景. 在与不依赖图像级标签的纯图像弱监督方法 (如 MaskCLIP、PixelCLIP 等) 的对比中, 本文模型展现出更强的语义区分能力与边界保持能力. 无论是在多人聚集、骑行场景中, 还是在复杂遮挡背景下的动物、交通工具等样例中, 均能准确识别并分割出多个类别的目标区域, 且目标之间边界清晰、结构完整, 尤其在非显著类 (如猫、酒瓶等) 场景中, 其他方法常出现漏检或粘连, 而本文的模型仍可较为准确地保留目

表 3 PASCAL VOC 2012 数据集 mIoU 实验 (%)

Table 3 mIoU experiment on the PASCAL VOC 2012 dataset (%)

方法	框架	验证集	测试集
SEAM ^[64]	ResNet38	64.5	65.7
AdvCAM ^[65]	ResNet101	68.1	68.0
OC-CSE ^[66]	ResNet38	68.4	68.2
VWE ^[67]	ResNet101	70.6	70.7
CLIMS ^[57]	ResNet101	70.4	70.0
SIPE ^[68]	ResNet101	68.8	69.7
W-OoD ^[69]	ResNet38	70.7	70.1
AMN ^[70]	ResNet101	69.5	69.6
ViT-PCM ^[71]	ResNet101	70.3	70.9
文献 [32] 方法	ResNet38	70.9	71.7
ToCo ^[72]	ViT-B	69.8	70.5
D2CAM ^[73]	ResNet101	71.2	70.7
FPR ^[59]	ResNet101	70.3	70.1
SFC ^[5]	ResNet38	70.2	71.4
CLIP-ES ^[58]	ResNet38	68.9	70.0
MaskCLIP ^[63]	ViT-PCM	67.3	69.5
MaskCLIP+ ^[63]	ViT-PCM	65.1	70.0
PixelCLIP ^[1]	ConvNeXt-B	67.2	—
文献 [74] 方法	ResNet101	71.1	70.9
文献 [75] 方法	ResNet50	70.7	71.5
本文方法	ViT-B	71.3	71.7

标轮廓, 体现出提示引导机制对弱激活区域的补全能力. 与依赖图像级标签的主流弱监督方法 (如 SEAM、AdvCAM、ToCo 等) 相比, 本文的模型在抑制背景误激活、提升语义一致性方面也表现出较强的鲁棒性. 结合背景建模与梯度裁剪机制, 本文方法在结构细节、上下文关联与空间一致性方面均优于对比方法, 充分验证了本文框架在复杂语义分割任务中的有效性与实用性.

如表 4 所示, 在更具挑战性的 MS COCO 2014 验证集上, 本文方法获得 45.1% 的 mIoU, 显著优于 ViT-PCM (45.0%) 与 ToCo (41.3%), 进一步验证了本文方法对目标、细粒度边界和非显著类的强鲁棒性. 此外, 在推理阶段, 本文方法不再依赖 CLIP 模型本身, 而是将其学习的提示信息与伪标签结构注入到最终分割网络中, 有效降低了部署成本并提升了推理效率.

3.3 计算开销与推理效率分析

为全面评估所提方法的实际部署潜力, 本文不仅关注分割精度, 还详细量化了模型在训练与推理阶段的时间与空间复杂度. 所有实验均在单张 NVIDIA

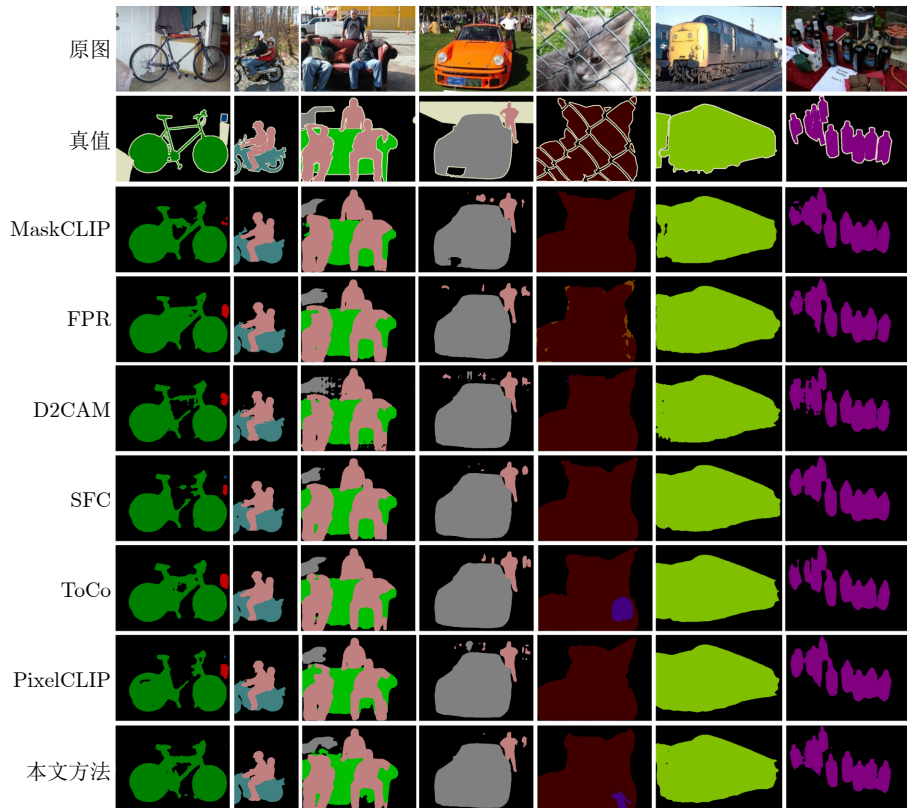


图 4 不同方法在 PASCAL VOC 2012 验证集上的分割可视化对比

Fig. 4 Segmentation visualization comparison of different methods on the PASCAL VOC 2012 validation set

RTX 3090 (24 GB) GPU 上进行, 输入图像分辨率统一调整为 512×512 .

虽然本文方法引入了自监督 CLIP 适应阶段以及随后的分割网络训练, 导致总训练时间相比于免训练方法 MaskCLIP 有所增加, 但以计算换人工的策略是值得的. 如表 5 所示, 本文的自监督适应阶段显存占用为 21.8 GB, 在 3090 显存上限 (24 GB) 的可控范围内总训练时长约为 13.0 h, 相比复杂的 CLIMS 方法 (18.5 h) 仍具有效率优势.

现有的无标签方法 (如 MaskCLIP^[63]、CLIP-ES^[58]) 在推理时仍需依赖庞大的 CLIP 视觉编码器进行密集的视觉注意力计算, 推理速度仅为 12.4 fps 和 8.5 fps, 难以满足实时需求. 相比之下, 本文方法采用了参数解耦策略: 自监督 CLIP 仅作为训练指导者, 在推理阶段完全被移除, 仅保留轻量级的分割网络. 因此, 本文方法的推理速度高达 47.5 fps, 相比 MaskCLIP 提升约 2.8 倍, 模型参数量减少约 67.8%, 完美实现了高精度与实时性的平衡.

3.4 消融实验

为深入剖析本文所提框架中各关键模块的独立贡献, 并验证动态提示机制相对于传统静态模板的

优越性, 本文在 PASCAL VOC 2012 验证集上进行系统的平行对照消融实验. 具体设置如下 (各实验设置对应的 mIoU 结果如表 6 所示):

1) 基线模型 (Baseline): 仅使用 ViT-B/16 分类网络生成初始 CAM, 作为纯视觉基准;

2) 提示策略对比: 使用固定手工模板 (如“a photo of [CLASS]”) 引入 CLIP 先验以及使用本文提出的自监督适应机制, 根据图像内容动态生成上下文提示进行实验对比;

3) 后处理模块验证: 在上述两种提示策略的基础上, 分别叠加 CAM 细化模块引入多尺度卷积与通道注意力和梯度裁剪机制来过滤噪声梯度, 以验证模块的通用性与最终性能上限.

从表 6 的实验结果可以得出以下结论: 首先, 在提示策略层面, 相比于 Baseline (64.9%), 引入静态提示将 mIoU 提升至 66.2%, 验证了 CLIP 跨模态知识的有效性; 而本文的动态提示进一步将其提升至 67.2%, 直接证明了可学习的上下文提示比固定模板更能精准激活目标区域. 其次, 在完整架构层面, 本文构建了基于静态提示的完整变体, 即“静态提示+CAM 细化+梯度裁剪”. 结果显示, 虽然该组合也将性能提升至 70.2%, 证明了 CAM 细化模

表 4 MS COCO 2014 数据集 mIoU 实验 (%)
Table 4 mIoU experiment on the MS COCO 2014 dataset (%)

方法	框架	验证集
SEAM ^[64]	ResNet38	31.9
AdvCAM ^[65]	ResNet101	33.9
OC-CSE ^[66]	ResNet38	36.4
VWE ^[67]	ResNet101	36.2
CLIMS ^[67]	ResNet101	—
SIPE ^[68]	ResNet101	43.6
W-OoD ^[69]	ResNet38	41.5
AMN ^[70]	ResNet101	44.7
ViT-PCM ^[71]	ResNet101	45.0
文献 [32] 方法	ResNet38	44.8
ToCo ^[72]	ViT-B	41.3
D2CAM ^[73]	ResNet101	44.0
FPR ^[59]	ResNet101	45.6
SFC ^[5]	ResNet38	45.9
CLIP-ES ^[58]	ResNet38	44.6
MaskCLIP ^[63]	ViT-PCM	—
MaskCLIP+ ^[63]	ViT-PCM	—
PixelCLIP ^[1]	ConvNeXt-B	—
文献 [74] 方法	ResNet101	41.5
文献 [75] 方法	ResNet50	41.3
本文方法	ViT-B	45.1

表 5 计算复杂度与推理效率对比

Table 5 Comparison of computational complexity and inference efficiency

方法	骨干网络	训练阶段		推理阶段	
		显存 (GB)	时间 (h)	参数量 (M)	速度 (fps)
MaskCLIP ^[63]	ViT-PCM/16	—	—	149.6	12.4
CLIP-ES ^[58]	ConvNeXt-B	16.5	15.0	149.6	8.5
CLIMS ^[67]	ConvNeXt-B	22.1	18.5	58.2	28.6
本文方法	ViT-B/16	21.8	13.0	48.2	47.5

块与梯度裁剪机制的普适性,但仍低于本文最终方案的 71.3%。这一 1.1% 的显著差距有力地证明: 动态提示不仅本身性能更优,而且能为后续的 CAM 细化模块与梯度裁剪机制提供更高质量的初始伪标签,从而确立了动态提示加结构化约束这一范式的整体优势。

为验证本文所提框架对核心超参数的鲁棒性,并在 PASCAL VOC 2012 验证集上确定最优参数配置,本文对自监督损失权重 λ 、CAM 细化平衡系数 γ 以及梯度裁剪启动阈值 ϵ 进行详细的敏感性分析实验,结果如图 5 所示。

● 自监督损失权重 λ 的影响。如图 5(a) 所示,

表 6 不同模块对 PASCAL VOC 2012 验证集性能的影响
Table 6 Impact of different modules on the performance of the PASCAL VOC 2012 validation set

模型设置	提示策略	CAM 细化	梯度裁剪	mIoU (%)
Baseline	—	×	×	64.9
Baseline	静态提示	×	×	66.2
Baseline (细化+裁剪)	静态提示	√	√	70.2
本文方法	动态提示	×	×	67.2
本文方法	动态提示	√	×	69.8
本文方法	动态提示	√	√	71.3

随着 λ 从 0.1 增加至 0.5, 模型性能呈现稳步上升趋势, 这表明适度的自监督约束能够有效防止 CLIP 视觉特征在适应过程中发生灾难性遗忘问题, 保持了其泛化能力。然而, 当 λ 超过 0.5 后, 过强的正则化约束开始抑制主要分类任务的学习, 导致 mIoU 轻微下降。因此, 本文选取 $\lambda = 0.5$ 作为最佳平衡点。

● CAM 细化平衡系数 γ 的影响。如图 5(b) 所示, γ 控制了语义细化模块对初始类激活图的修正力度。实验数据显示, 当 $\gamma = 0.3$ 时模型达到性能峰值 (71.3%)。较小的 γ 无法充分利用上下文注意力来修正粗糙的边界, 而过大的 γ (如 0.5) 则可能引入过度的平滑先验, 导致细节特征丢失。

● 梯度裁剪启动阈值 ϵ 的影响。如图 5(c) 所示, 模型性能对 ϵ 的选择较为敏感。该阈值决定了模型“拒绝”高梯度噪声的严格程度, 当 ϵ 设定为 0.10 时, 能够最有效地平衡噪声过滤与有效特征保留。当 $\epsilon < 0.10$ 时, 过于激进的裁剪策略导致部分包含丰富纹理信息的有效梯度被错误丢弃, 造成欠拟合 (mIoU 降至 68.4%); 反之, 当 $\epsilon > 0.15$ 时, 由于阈值过高, 大量背景噪声梯度得以回传, 削弱了该模块的抗噪能力。

综上所述, 实验结果表明本文方法在参数 $\lambda = 0.5$, $\gamma = 0.3$, $\epsilon = 0.10$ 处表现出最佳的收敛性与鲁棒性, 验证了各模块参数设计的合理性。

4 结束语

本文针对弱监督语义分割中伪标签质量受限、空间一致性弱及背景干扰显著等问题, 区别于传统依赖图像级标签的方法, 提出一种仅需输入图像的自监督 CLIP 框架。该框架通过自监督训练增强 CLIP 的视觉表征能力, 并整合自动背景建模、类感知注意力传播及置信度感知监督, 显著提升了伪标签的语义完整性与空间一致性, 增强了分割网络对复杂场景下目标边界和背景噪声的区分能力。在 PASCAL VOC 2012 与 MS COCO 2014 数据集上的实验表明, 本文方法在 mIoU 上优于现有主流多阶

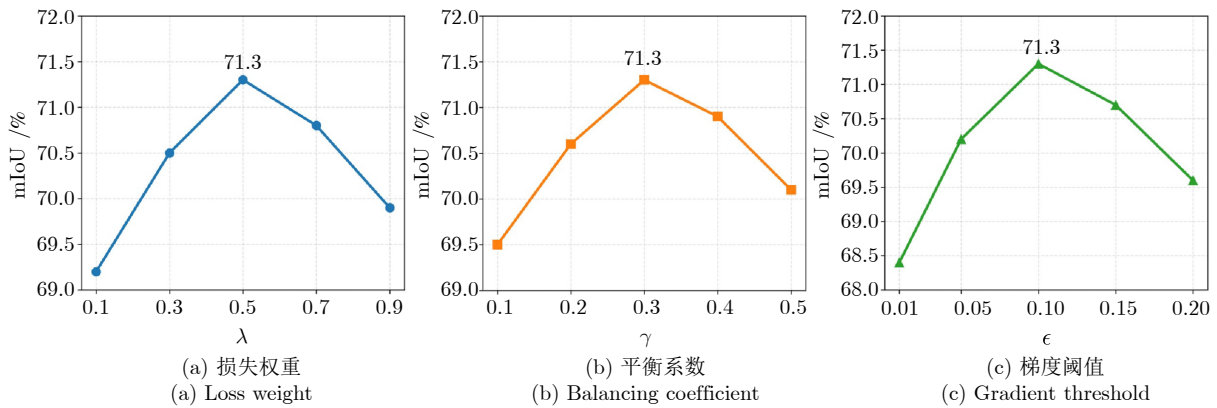


图 5 超参数敏感性分析

Fig. 5 Hyperparameter sensitivity analysis

段方法, 验证了其在效率、准确性与可扩展性上的综合优势。

尽管本文方法在目标轮廓保留上表现优异 (如图 4 所示), 但在处理极小目标或极高重叠场景时, 区分能力仍有待进一步提升。这主要受限于预训练 CLIP 视觉编码器的固定分辨率以及类激活图在密集场景下的响应弥散问题, 导致细微特征难以被精确捕捉。

未来的工作将致力于解决上述问题, 具体方向包括: 1) 细粒度与密集场景优化: 针对极小目标与高重叠场景, 探索引入多尺度高分辨率适配器或实例级对比学习机制, 以增强模型在密集遮挡下的特征判别力与像素级解耦能力; 2) 开放与长尾类别分割: 探索在开放世界及长尾分布下的弱监督分割, 以应对实际场景中类别不平衡与未知类别出现的挑战; 3) 大模型自适应与提示学习: 进一步结合大模型微调与动态提示学习机制, 持续优化语义对齐与上下文建模能力, 推动该方法在遥感、医疗等大规模真实场景中的实用化落地。

参考文献

- Guo Zi-Lin, Wu Dong-Yue, Gao Chang-Xin, Sang Nong. Image-specific classifiers for weakly supervised semantic segmentation. *Acta Automatica Sinica*, 2025, **51**(6): 1191–1204 (郭子麟, 吴东岳, 高常鑫, 桑农. 基于图像特定分类器的弱监督语义分割. *自动化学报*, 2025, **51**(6): 1191–1204)
- Yang Da-Wei, Chi Jin-Sheng, Mao Lin. Weakly supervised semantic segmentation network with multiple boundary references. *Journal of Computer-aided Design & Computer Graphics*, 2025, **37**(1): 120–128 (杨大伟, 迟津生, 毛琳. 多重边界参考的弱监督语义分割网络. *计算机辅助设计与图形学学报*, 2025, **37**(1): 120–128)
- Zhou Kai, Yu Lian-Zhi. Self-activated learning method for weakly supervised semantic segmentation integrating attention mechanism. *Electronic Science & Technology*, 2025, **38**(4): 80–86 (周凯, 于莲芝. 融合注意力机制的弱监督语义分割自激活方法. *电子科技*, 2025, **38**(4): 80–86)
- Feng Xing-Jie, Nan Bo-Gong. Semi-supervised semantic segmentation method based on adversarial learning and consistent regularization. *Computer Applications and Software*, 2025, **42**(1): 182–188 (冯兴杰, 南博公. 基于对抗学习和一致性正则的半监督语义分割方法. *计算机应用与软件*, 2025, **42**(1): 182–188)
- Zhao X Q, Tang F L, Wang X Y, Xiao J M. SFC: Shared feature calibration in weakly supervised semantic segmentation. In: *Proceedings of the 38th AAAI Conference on Artificial Intelligence*. Vancouver, Canada: AAAI Press, 2024. 7525–7533
- Na J, Ha J W, Chang H J, Han D, Hwang W. Switching temporary teachers for semi-supervised semantic segmentation. In: *Proceedings of the 37th International Conference on Neural Information Processing Systems*. New Orleans, USA: Curran Associates Inc., 2024. Article No. 1754
- Zhou T F, Wang W G. Prototype-based semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024, **46**(10): 6858–6872
- Tan C C, Tao R S, Liu H, Gu G H, Wu B Y, Zhao Y, et al. C2P-CLIP: Injecting category common prompt in CLIP to enhance generalization in deepfake detection. In: *Proceedings of the 39th AAAI Conference on Artificial Intelligence*. Philadelphia, USA: AAAI Press, 2025. 7184–7192
- Hou Yu-Shuang, Kong Yu-Qiu, Zhang Li-He, Yin Bao-Cai. Weakly-supervised referring image segmentation based on entity information. *Journal of Computer-aided Design & Computer Graphics*, 2026, **38**(3): 553–562 (侯雨爽, 孔雨秋, 张立和, 尹宝才. 基于实体信息的弱监督指称图像分割方法. *计算机辅助设计与图形学学报*, 2026, **38**(3): 553–562)
- Yang K C, Gu T C, An X, Jiang H Q, Dai X Z, Feng Z Y, et al. CLIP-CID: Efficient CLIP distillation via cluster-instance discrimination. In: *Proceedings of the 39th AAAI Conference on Artificial Intelligence*. Philadelphia, USA: AAAI Press, 2025. 21974–21982
- Yuan B, Zhao D P. A survey on continual semantic segmentation: Theory, challenge, method and application. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024, **46**(12): 10891–10910
- Shi M, Lin S W, Yi Q M, Weng J, Luo A W, Zhou Y C. Lightweight context-aware network using partial-channel transformation for real-time semantic segmentation. *IEEE Transactions on Intelligent Transportation Systems*, 2024, **25**(7): 7401–7416
- Shan J H, Jiang W, Huang Y, Yuan D D, Liu Y H. Unmanned aerial vehicle (UAV)-based pavement image stitching without occlusion, crack semantic segmentation, and quantification. *IEEE Transactions on Intelligent Transportation Systems*, 2024, **25**(11): 17038–17053
- Zhou T F, Wang W G. Cross-image pixel contrasting for se-

- mantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024, **46**(8): 5398–5412
- 15 Xie B, Cao J L, Xie J, Khan F S, Pang Y W. SED: A simple encoder-decoder for open-vocabulary semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, USA: IEEE, 2024. 3426–3436
 - 16 Zhang B C, Zhang P, Dong X Y, Zang Y H, Wang J Q. Long-CLIP: Unlocking the long-text capability of CLIP. In: Proceedings of the 18th European Conference on Computer Vision (ECCV). Milan, Italy: Springer, 2024. 310–325
 - 17 Sun Z Y, Fang Y, Wu T, Zhang P, Zang Y H, Kong S, et al. Alpha-CLIP: A CLIP model focusing on wherever you want. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, USA: IEEE, 2024. 13019–13029
 - 18 Yu C Y, Liu X H, Wang Y Q, Zhang P P, Lu H C. TF-CLIP: Learning text-free CLIP for video-based person re-identification. In: Proceedings of the 38th AAAI Conference on Artificial Intelligence. Vancouver, Canada: AAAI Press, 2024. 6764–6772
 - 19 Zhu L H, Wang X G, Feng J P, Cheng T H, Li Y Y, Jiang B, et al. WeakCLIP: Adapting CLIP for weakly-supervised semantic segmentation. *International Journal of Computer Vision*, 2025, **133**(3): 1085–1105
 - 20 Smeu S, Oneata E, Oneata D. DeCLIP: Decoding CLIP representations for deepfake localization. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). Tucson, USA: IEEE, 2025. 149–159
 - 21 Singha M, Jha A, Banerjee B. GOPro: Generate and optimize prompts in CLIP using self-supervised learning. arXiv preprint arXiv: 2308.11605, 2023.
 - 22 Lim J Y, Lim K M, Lee C P, Tan Y X. SSL-ProtoNet: Self-supervised learning prototypical networks for few-shot learning. *Expert Systems With Applications*, 2024, **238**: Article No. 122173.
 - 23 Zhang K X, Wen Q S, Zhang C L, Cai R Y, Jin M, Liu Y, et al. Self-supervised learning for time series analysis: Taxonomy, progress, and prospects. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024, **46**(10): 6775–6794
 - 24 Chen H Y, Li W B, Gu J J, Ren J J, Sun H Z, Zou X Y, et al. Low-res leads the way: Improving generalization for super-resolution by self-supervised learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, USA: IEEE, 2024. 25857–25867
 - 25 Gui J, Chen T, Zhang J, Cao Q, Sun Z N, Luo H, et al. A survey on self-supervised learning: Algorithms, applications, and future trends. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024, **46**(12): 9052–9071
 - 26 Zhang B F, Yu S Y, Wei Y C, Zhao Y, Xiao J M. Frozen CLIP: A strong backbone for weakly supervised semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, USA: IEEE, 2024. 3796–3806
 - 27 Yang X B, Gong X J. Foundation model assisted weakly supervised semantic segmentation. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). Waikoloa, USA: IEEE, 2024. 512–521
 - 28 Xie J H, Deng S H, Hou X X, Luo Z C, Shen L L, Huang Y W, et al. CLIMS++: Cross language image matching with automatic context discovery for weakly supervised semantic segmentation. *International Journal of Computer Vision*, 2025, **133**(8): 5569–5588
 - 29 Yang Z W, Fu K X, Duan M H, Qu L H, Wang S, Song Z J. Separate and conquer: Decoupling co-occurrence via decomposition and representation for weakly supervised semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, USA: IEEE, 2024. 3606–3615
 - 30 Wu Y C, Ye X C, Yang K Q, Li J D, Li X Q. DuPL: Dual student with trustworthy progressive learning for robust weakly supervised semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, USA: IEEE, 2024. 3534–3543
 - 31 Wu W Y, Dai T H, Chen Z H, Huang X W, Ma F, Xiao J M. Generative prompt controlled diffusion for weakly supervised semantic segmentation. *Neurocomputing*, 2025, **638**: Article No. 130103
 - 32 Yoon S H, Kwon H, Kim H, Yoon K J. Class tokens infusion for weakly supervised semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, USA: IEEE, 2024. 3595–3605
 - 33 Wu W Y, Qiu X L, Song S Q, Chen Z H, Huang X W, Ma F, et al. Prompt categories cluster for weakly supervised semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). Nashville, USA: IEEE, 2025. 3189–3198
 - 34 Krizhevsky A, Sutskever I, Hinton G E. ImageNet classification with deep convolutional neural networks. In: Proceedings of the 26th International Conference on Neural Information Processing Systems. Lake Tahoe, USA: Curran Associates Inc., 2012. 1097–1105
 - 35 Li F F, Fergus R, Perona P. Learning generative visual models from few training examples: An incremental Bayesian approach tested on 101 object categories. In: Proceedings of the Conference on Computer Vision and Pattern Recognition Workshop. Washington, USA: IEEE, 2004. Article No. 178
 - 36 Parkhi O M, Vedaldi A, Zisserman A, Jawahar C V. Cats and dogs. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Providence, USA: IEEE, 2012. 3498–3505
 - 37 Krause J, Stark M, Deng J, Li F F. 3D object representations for fine-grained categorization. In: Proceedings of the IEEE International Conference on Computer Vision Workshops. Sydney, Australia: IEEE, 2013. 554–561
 - 38 Nilsback M E, Zisserman A. Automated flower classification over a large number of classes. In: Proceedings of the 6th Indian Conference on Computer Vision, Graphics & Image Processing. Bhubaneswar, India: IEEE, 2008. 722–729
 - 39 Bossard L, Guillaumin M, van Gool L. Food-101—Mining discriminative components with random forests. In: Proceedings of the 13th European Conference on Computer Vision (ECCV). Zurich, Switzerland: Springer, 2014. 446–461
 - 40 Maji S, Rahtu E, Kannala J, Blaschko M, Vedaldi A. Fine-grained visual classification of aircraft. arXiv preprint arXiv: 1306.5151, 2013.
 - 41 Xiao J X, Hays J, Ehinger K A, Oliva A, Torralba A. SUN database: Large-scale scene recognition from abbey to zoo. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. San Francisco, USA: IEEE, 2010. 3485–3492.
 - 42 Cimpoi M, Maji S, Kokkinos I, Mohamed S, Vedaldi A. Describing textures in the wild. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Columbus, USA: IEEE, 2014. 3606–3613
 - 43 Recht B, Roelofs R, Schmidt L, Shankar V. Do ImageNet classifiers generalize to ImageNet? In: Proceedings of the 36th International Conference on Machine Learning (ICML). Long Beach, USA: PMLR, 2019. 5389–5400
 - 44 Wang H H, Ge S W, Xing E P, Lipton Z C. Learning robust global representations by penalizing local predictive power. In: Proceedings of the 33rd International Conference on Neural Information Processing Systems. Vancouver, Canada: Curran As-

- sociates Inc., 2019. Article No. 943
- 45 Hendrycks D, Zhao K, Basart S, Steinhardt J, Song D. Natural adversarial examples. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville, USA: IEEE, 2021. 15257–15266
 - 46 Hendrycks D, Basart S, Mu N, Kadavath S, Wang F, Dorundo E, et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). Montreal, Canada: IEEE, 2021. 8320–8329
 - 47 Everingham M, van Gool L, Williams C K I, Winn J, Zisserman A. The PASCAL visual object classes (VOC) challenge. *International Journal of Computer Vision*, 2010, **88**(2): 303–338
 - 48 Lin T Y, Maire M, Belongie S, Hays J, Perona P, Ramanan D, et al. Microsoft COCO: Common objects in context. In: Proceedings of the 13th European Conference on Computer Vision (ECCV). Zurich, Switzerland: Springer, 2014. 740–755
 - 49 Jo S, Yu I J, Kim K. RecurSeed and EdgePredictMix: Single-stage learning is sufficient for weakly-supervised semantic segmentation. arXiv preprint arXiv: 2204.06754, 2022.
 - 50 Radford A, Kim J W, Hallacy C, Ramesh A, Goh G, Agarwal S, et al. Learning transferable visual models from natural language supervision. In: Proceedings of the 38th International Conference on Machine Learning (ICML). Virtual Event: PMLR, 2021. 8748–8763
 - 51 Mu N, Kirillov A, Wagner D, Xie S N. SLIP: Self-supervision meets language-image pre-training. In: Proceedings of the 17th European Conference on Computer Vision (ECCV). Tel Aviv, Israel: Springer, 2022. 529–544
 - 52 Zhou K Y, Yang J K, Loy C C, Liu Z W. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 2022, **130**(9): 2337–2348
 - 53 Zhou K Y, Yang J K, Loy C C, Liu Z W. Conditional prompt learning for vision-language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans, USA: IEEE, 2022. 16795–16804
 - 54 Khattak M U, Rasheed H, Maaz M, Khan S, Khan F S. MaPLE: Multi-modal prompt learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Vancouver, Canada: IEEE, 2023. 19113–19122
 - 55 Bose S, Jha A, Fini E, Singha M, Ricci E, Banerjee B. StyLIP: Multi-scale style-conditioned prompt learning for CLIP-based domain generalization. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). Waikoloa, USA: IEEE, 2024. 5530–5540
 - 56 Xu L, Ouyang W L, Bennamoun M, Boussaid F, Xu D. Multi-class token Transformer for weakly supervised semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans, USA: IEEE, 2022. 4300–4309
 - 57 Xie J H, Hou X X, Ye K, Shen L L. CLIMS: Cross language image matching for weakly supervised semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans, USA: IEEE, 2022. 4473–4482
 - 58 Lin Y Q, Chen M H, Wang W X, Wu B X, Li K, Lin B B, et al. CLIP is also an efficient segmenter: A text-driven approach for weakly supervised semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Vancouver, Canada: IEEE, 2023. 15305–15314
 - 59 Cheng B W, Misra I, Schwing A G, Kirillov A, Girdhar R. Masked-attention mask Transformer for universal image segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans, USA: IEEE, 2022. 1280–1289
 - 60 Gao X R, Liu W F. CLIP as auxiliary supervisor for weakly supervised semantic segmentation. In: Proceedings of the International Conference on New Trends in Computational Intelligence (NTCI). Qingdao, China: IEEE, 2024. 21–26
 - 61 Wu F W, He J X, Yin Y F, Hao Y B, Huang G, Cheng L C. Masked collaborative contrast for weakly supervised semantic segmentation. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). Waikoloa, USA: IEEE, 2024. 851–860
 - 62 Luo H L, Zeng Z. Enhancing weakly supervised semantic segmentation through multi-class token attention learning. *The Journal of Supercomputing*, 2025, **81**(1): Article No. 74
 - 63 Shin H, Kim C, Hong S, Cho S, Arnab A, Seo P H, et al. Towards open-vocabulary semantic segmentation without semantic labels. In: Proceedings of the 38th International Conference on Neural Information Processing Systems. Vancouver, Canada: Curran Associates Inc., 2024. Article No. 290
 - 64 Shao Q Q, Yu L J, Guo Y T, Xie X C, Zou J P, Li L. Weakly supervised semantic segmentation of PolSAR image based on improved SEAM. *Journal of Physics: Conference Series*, 2023, **2456**(1): Article No. 012003
 - 65 Duan R J, Ma X J, Wang Y S, Bailey J, Qin A K, Yang Y. Adversarial camouflage: Hiding physical-world attacks with natural styles. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, USA: IEEE, 2020. 997–1005
 - 66 Kweon H, Yoon S H, Kim H, Park D, Yoon K J. Unlocking the potential of ordinary classifier: Class-specific adversarial erasing framework for weakly supervised semantic segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). Montreal, Canada: IEEE, 2021. 6974–6983
 - 67 Ru L X, Du B, Wu C. Learning visual words for weakly-supervised semantic segmentation. In: Proceedings of the 30th International Joint Conference on Artificial Intelligence (IJCAI). Montreal, Canada: ijcai.org, 2021. 982–988
 - 68 Chen Q, Yang L X, Lai J H, Xie X H. Self-supervised image-specific prototype exploration for weakly supervised semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans, USA: IEEE, 2022. 4278–4288
 - 69 Lee J, Oh S J, Yun S, Choe J, Kim E, Yoon S R. Weakly supervised semantic segmentation using out-of-distribution data. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans, USA: IEEE, 2022. 16876–16885
 - 70 Li Y Q, Tuncel E, Chen J. Optimal tracking over an additive white noise feedback channel. In: Proceedings of the 7th Asian Control Conference. Hong Kong, China: IEEE, 2009. 501–506
 - 71 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: Proceedings of the 17th European Conference on Computer Vision (ECCV). Tel Aviv, Israel: Springer, 2022. 446–463
 - 72 Ru L X, Zheng H L, Zhan Y B, Du B. Token contrast for weakly-supervised semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Vancouver, Canada: IEEE, 2023. 3093–3102
 - 73 Wang C W, Xu R T, Xu S B, Meng W L, Zhang X P. Treating pseudo-labels generation as image matting for weakly supervised semantic segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). Paris, France: IEEE, 2023. 755–765
 - 74 Xu M D, Zhang Z, Wei F Y, Lin Y T, Cao Y, Hu H, et al. A simple baseline for open-vocabulary semantic segmentation with pre-trained vision-language model. In: Proceedings of the 17th European Conference on Computer Vision (ECCV). Tel Aviv, Israel: Springer, 2022. 736–753

- 75 Ding J, Xue N, Xia G S, Dai D X. Decoupling zero-shot semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans, USA: IEEE, 2022. 11573–11582



魏龙生 中国地质大学(武汉)人工智能与自动化学院副教授. 主要研究方向为计算机视觉, 模式识别. 本文通信作者.

E-mail: weilongsheng@cug.edu.cn

(WEI Long-Sheng) Associate professor at the School of Artificial Intelligence and Automation, China University of Geosciences. His research interests include computer vision and pattern recognition. Corresponding author of this paper.)



李唐强 中国地质大学(武汉)人工智能与自动化学院硕士研究生. 主要研究方向为深度学习, 图像分割.

E-mail: litq@cug.edu.cn

(LI Tang-Qiang) Master student at the School of Artificial Intelligence and Automation, China University

of Geosciences. His research interests include deep learning and image segmentation.)



徐朗 中国地质大学(武汉)人工智能与自动化学院硕士研究生. 主要研究方向为语义图像分割.

E-mail: cyberlang@163.com

(XU Lang) Master student at the School of Artificial Intelligence and Automation, China University of Geosciences. His main research interest is semantic image segmentation.)



罗大鹏 中国地质大学(武汉)机械与电子信息学院副教授. 主要研究方向为图像/视频处理, 计算机视觉和机器学习.

E-mail: luodapeng@cug.edu.cn

(LUO Da-Peng) Associate professor at the School of Mechanical Engineering and Electronic Information, China University of Geosciences. His research interests include image/video processing, computer vision, and machine learning.)