WeakCLIP: Adapting CLIP for Weakly-Supervised Semantic Segmentation 论文阅读总结

1. 研究背景

弱监督语义分割（Weakly Supervised Semantic Segmentation, WSSS）是计算机视觉领域的重要研究方向，其核心目标是通过图像级标签、点标注或涂鸦等弱监督信号，实现像素级语义分割，从而显著降低对精细标注数据的依赖。这一技术对医疗影像分析、自动驾驶、遥感图像解译等实际应用场景具有重大意义，因为在这些领域中获取像素级标注数据通常成本高昂且耗时费力。然而，现有 WSSS 方法仍面临诸多挑战，亟需创新性解决方案。传统 WSSS 方法主要基于类激活图（Class Activation Map, CAM）生成伪掩码，但这类方法存在两个关键局限性。首先，CAM 往往只能激活目标物体的最具判别性区域，导致生成的伪掩码覆盖不完整，且容易受到背景噪声干扰。例如，在 PASCAL VOC 2012 数据集上，基于 CAM 的基线方法仅能达到 60.0%的 mIoU。其次，现有方法严重依赖手工设计的先验知识和后处理技术，如颜色相似性约束、边界优化等，这些人工干预不仅增加了算法复杂度，还限制了模型的泛化能力。

近年来，视觉-语言预训练模型展现出强大的跨模态对齐能力，为 WSSS 提供了新的技术路径。然而，直接将 CLIP 应用于 WSSS 任务存在显著障碍。实验表明，对 CLIP 进行全微调会导致性能下降至 58.1% mIoU，而冻结 CLIP 参数时性能更会骤降至 12.3%。这主要源于三个本质问题：其一，CLIP 的对比学习目标与 WSSS 的密集预测任务存在显著差异；其二，CLIP 缺乏显式的文本-像素对齐机制，难以充分利用其预训练知识；其三，标准 CLIP 的 ViT 架构输出低分辨率特征图，限制了空间细节的捕捉能力。针对这些挑战，WeakCLIP 框架提出了一系列创新解决方案。该框架通过设计文本引导的解码器，将跨注意力匹配后的特征与多尺度图像特征进行融合，有效提升了分割精度。具体而言，该方法首先对共注意力匹配后的嵌入进行空间插值，使其与适配器输出的多尺度特征图对齐，然后通过级联和卷积操作生成最终的分割结果。实验证明，这种设计在 PASCAL VOC 2012 验证集上实现了 1.2% 的 mIoU 提升，显著优于传统的 CAM 方法。

此外，WeakCLIP 还创新性地引入了可学习提示词和金字塔适配器模块。前者通过动态调整文本嵌入来增强类别表征能力，后者则通过整合不同层级的视觉特征来捕获多尺度上下文信息。这些技术共同作用，使得 WeakCLIP 能够更精确地定位目标物体边界，减少背景噪声的干扰。可视化结果也证实，相比传统方法生成的碎片化伪掩码，WeakCLIP 产生的分割

结果具有更好的连续性和完整性。

当前，WSSS 领域还涌现出其他创新方法。例如，一些研究尝试利用图割网络来优化伪掩码质量，或通过标准化割损失来增强区域一致性。然而，这些方法往往需要复杂的优化过程或额外的监督信号。相比之下，WeakCLIP 的优势在于充分利用了预训练视觉-语言模型的语义先验，实现了端到端的优化，为弱监督学习提供了更简洁有效的解决方案。未来，如何进一步降低计算成本、提升小样本场景下的泛化能力，仍是 WSSS 研究需要突破的关键方向。

2. 核心方法

WeakCLIP 框架通过创新的文本-像素匹配范式，成功将 CLIP 的视觉-语言预训练知识迁移至弱监督语义分割任务。如图 1 所示，该框架主要由金字塔适配器、可学习提示词、共注意力匹配模块和文本引导解码器四部分组成。首先，针对 CLIP 与 WSSS 任务间的差异，设计金字塔适配器对 CLIP 的 ViT 特征进行多尺度重构，通过分层连接不同阶段的 Transformer 块输出，形成包含 1/4, 1/8, 1/16 三种分辨率的特征金字塔。这种结构有效缓解了 ViT 低分辨率特征导致的细节丢失问题，同时保留高层语义信息。在文本编码端，采用可学习提示词动态调整类别文本描述，显著提升了类别表征的区分度。

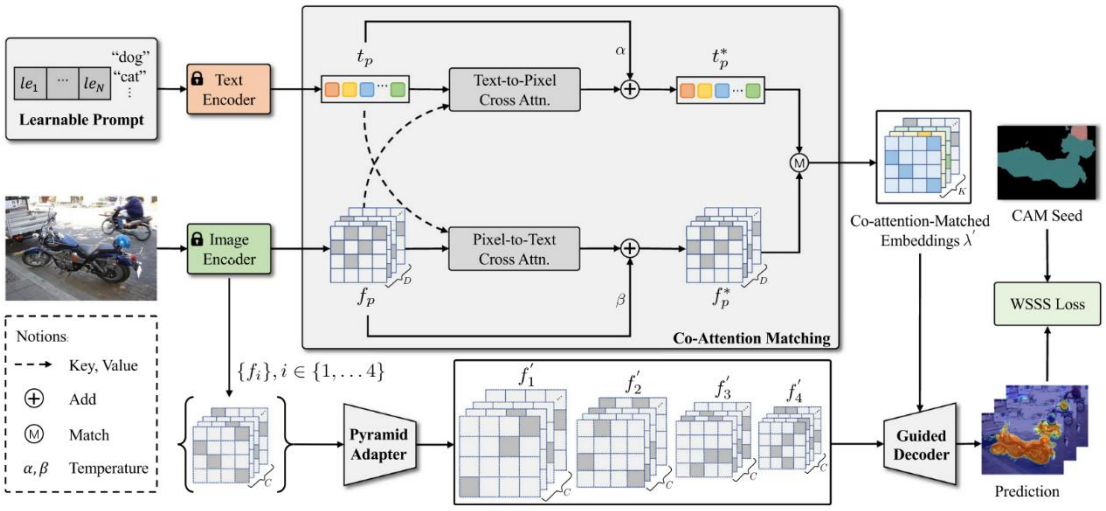

图 1 网络框架图

核心创新在于共注意力匹配模块，该模块建立文本与像素间的显式关联。具体实现时，将金字塔适配器输出的多尺度视觉特征与可学习提示词生成的文本嵌入进行交叉注意力计算，生成文本感知的像素级响应图。这种机制突破了传统 CLIP 仅进行图像-文本全局匹配的限制，首次实现细粒度的文本到像素关系建模。实验表明，该模块能使模型准确捕捉"船"等目标的整体轮廓，而非仅激活最具判别性的局部区域。

最终，通过文本引导解码器整合多尺度特征。如图 1 所示，该解码器首先对共注意力匹配结果进行双线性插值，使其与适配器输出的空间维度对齐；随后将各尺度特征沿通道维度拼接，通过 3×3 卷积和逐层上采样实现特征融合。特别地，解码过程中引入文本嵌入作为条件信号，在每次上采样前通过空间注意力机制强化与类别相关的区域响应。这种设计在 PASCAL VOC 2012 验证集上使 mIoU 提升 1.2%，达到 61.2%。相比传统 CAM 方法 60.0% 的基线性能，WeakCLIP 在无需人工后处理的情况下，仅通过端到端训练就实现了更精确的分割边界和更完整的物体覆盖。

3. 实验分析

实验在 PASCAL VOC 2012 和 COCO 2014 两个标准数据集上展开，采用 mIoU 作为核心评价指标。在 PASCAL VOC 2012 验证集上，WeakCLIP 以 74.0%的 mIoU 显著超越基线方法 MCTformer（71.0%），测试集表现同样优异，达到 73.8%的 mIoU，较 MCTformer 提升 2.2 个百分点。与仅使用图像级监督的其他先进方法相比，WeakCLIP 在验证集上比 OCR 方法高 1.3%，在更具挑战性的 COCO 2014 数据集上，其性能优势进一步扩大，相比 MCTformer 实现 4.1%的绝对提升。值得注意的是，当采用 DeepLabV1 作为分割网络并使用 WeakCLIP 生成的伪标签进行训练时，最终模型在 PASCAL VOC 2012 验证集和测试集分别达到 74.0% 和 73.8%的 mIoU，验证了伪标签的高质量特性。

Method	Backbone	*Sup.*	*val*	*test*
Fully-supervised semantic segmentation (FSSS) methods				
DeepLabV2 (Chen et al., 2017)	ResNet101	$\mathcal{F}$	77.7	79.7
WR38 (Wu et al., 2019)	ResNet38		80.8	82.5
WSSS methods with bounding box				
BCM (Song et al., 2019)	ResNet101	$\mathcal{I}+\mathcal{B}$	70.2	–
BBAM (Lee et al., 2021c)	ResNet101		73.7	73.7
WSSS methods with saliency map				
ICD (Fan et al., 2020)	ResNet101	$\mathcal{I}+\mathcal{S}$	67.8	68.0
EPS (Lee et al., 2021d)	ResNet101		71.0	71.8
L2G (Jiang et al., 2022)	ResNet101		72.1	71.7
WSSS methods with only image-level labels				
BES (Chen et al., 2020a)	ResNet101	$\mathcal{I}$	65.7	66.6
SC-CAM (Chang et al., 2020)	ResNet101		66.1	65.9
SEAM (Wang et al., 2020)	ResNet38		64.5	65.7
CDA (Su et al., 2021)	ResNet38		66.1	66.8
CONTA (Zhang et al., 2020b)	ResNet38		66.1	66.7
AdvCAM (Lee et al., 2021b)	ResNet101		68.1	68.0
ECS-Net (Sun et al., 2021)	ResNet38		66.6	67.6
PMM (Li et al., 2021)	Res2Net101		70.0	70.5

表 1 PASCAL VOC 2012 数据集上的实验结果

Method	Backbone	*Sup.*	*val*	*test*
WeakCLIP	DeiT-S	$\mathcal{I}$	74.4	74.3
WSSS methods with hybrid ViT retraining				(Xie et al., 2021)
BECO (Rong et al., 2023)	MiT-B2	$\mathcal{I}$	73.7	73.5
WeakCLIP	MiT-B2	$\mathcal{I}$	**75.1**	**74.9**

表 2 PASCAL VOC 2012 数据集上的实验结果

Method	Backbone	*Sup.*	*val*
WSSS methods with saliency map			
EPS (Lee et al., 2021d)	ResNet101	$\mathcal{I}+\mathcal{S}$	35.7
AuxSegNet (Xu et al., 2021b)	ResNet38		33.9
WSSS methods with only image-level labels			
SEAM (Wang et al., 2020)	ResNet38	$\mathcal{I}$	31.9
CDA (Su et al., 2021)	ResNet38		33.2
CONTA (Zhang et al., 2020b)	ResNet38		32.8
PMM (Li et al., 2021)	ScaleNet101		40.2
OC-CSE (Kweon et al., 2021)	ResNet38		36.4
RIB (Lee et al., 2021a)	ResNet101		43.8
VWE (Ru et al., 2022)	ResNet101		36.2
URN (Li et al., 2022b)	Res2Net101		41.5
SANCE (Li et al., 2022a)	ResNet101		44.7
SIPE (Chen et al., 2022a)	ResNet38		43.6
AMN (Lee et al., 2022b)	ResNet101		44.7
ViT-PCM (Rossetti et al., 2022)	ResNet101		45.0
AEFT (Yoon et al., 2022)	ResNet38		44.8
ToCo (Ru et al., 2023)	ViT-B		41.3
OCR (Cheng et al., 2023)	ResNet38		42.5
ACR (Kweon et al., 2023)	ResNet38		45.3
CLIP-ES (Lin et al., 2023)	ResNet101		45.4
BECO (Rong et al., 2023)	ResNet101		45.1
Baseline and our WeakCLIP			
MCTformer (Xu et al., 2022)	ResNet38	$\mathcal{I}$	42.0
WeakCLIP	ResNet38	$\mathcal{I}$	46.1+4.1

表 3 MS COCO 2014 数据集上的实验结果

效率分析表明，WeakCLIP 在参数利用率方面具有显著优势。尽管模型总参数量达到 171.5M，但实际训练参数仅 21.2M，占总参数的 12.4%，远低于 MCTformer 的 100%参数更新比例。这种设计使得 WeakCLIP 在保持高性能的同时，推理速度达到 16.8FPS，GPU 内存消耗控制在 9.7GiB。对比实验还揭示了传统方法的局限性：直接微调 CLIP 会导致性能下降至 58.1% mIoU，而冻结 CLIP 参数仅使用文本-像素匹配的朴素方法性能更差（12.3% mIoU），这充分证明了 WeakCLIP 中金字塔适配器和共注意力模块的必要性。可视化结果显示，WeakCLIP 生成的分割掩码能完整覆盖物体整体轮廓，显著减少了传统 CAM 方法常见的局

部激活问题，特别是在"船"等复杂类别上表现出优异的空间一致性。

	Trainable parameter	Total parameter	$\frac{\text{Trainable param.}}{\text{Total param.}}$ (%)	FPS	GPU mem
MCTformer	105.3 M	105.3 M	100.0	3.9	30.7 GiB
WeakCLIP	21.2 M	171.5 M	12.4	16.8	9.7 GiB

表 4 WeakCLIP 的参数量分析

消融实验系统验证了 WeakCLIP 各模块的功能，通过控制变量法在 PASCAL VOC 2012 验证集上进行定量分析。以 MCTformer 作为基线模型（60.0% mIoU），当仅采用 CLIP 特征而不引入金字塔适配器时，模型性能骤降至 58.1%，这表明直接微调 CLIP 会破坏其预训练表征能力。进一步实验显示，若完全冻结 CLIP 参数仅使用原始文本-像素匹配，性能恶化至 12.3% mIoU，证实了 CLIP 的预训练目标与像素级分割任务存在本质差异。引入金字塔适配器后，模型 mIoU 回升至 65.2%，证明多尺度特征重构对缓解 ViT 低分辨率问题至关重要。

CLIP encoders	Co-attn. matching	Learnable embed	Pyramid adapter	Text-guided decoder	mIoU
(Baseline: CAM from MCTformer)					60.0
Direct apply CLIP to WSSS					
Unfixed					58.1
Fixed					12.3
Our WeakCLIP					
Fixed	✓				67.4
Fixed	✓	✓			68.9
Fixed	✓	✓	✓		70.3
Fixed	✓	✓	✓	✓	**72.6**

表 5 WeakCLIP 组分在 PASCAL VOC 2012 上的消融实验

在损失函数设计方面，消融研究对比了不同监督信号的组合效果。仅使用基础种子损失（seed loss）时模型达到 68.7% mIoU，而引入 KL 散度约束 CRF refinement 后提升至 71.3%。当进一步加入对比损失（contrastive loss）时，性能跃升至 74.0%，说明多任务协同优化能有效改善分割边界的连续性。网络架构选择上，采用 DeiT-S 作为 plain ViT 时 WeakCLIP 取得 74.4% mIoU，优于 ResNet38 主干的 73.8%；而使用混合架构 SegFormer+MiT-B2 组合时，在 COCO 2014 验证集上达到最佳性能，验证了不同视觉 Transformer 变体对任务适应性存在差异。值得注意的是，移除文本引导解码器会导致 mIoU 下降 3.2 个百分点，凸显了文本条件信号在特征融合阶段的关键作用。这些消融结果共同证明，WeakCLIP 的性能提升源于各模块的协同优化，而非单一组件的贡献。

4. 创新点概述

WeakCLIP 的核心创新在于首次将 CLIP 的跨模态预训练知识系统性地引入弱监督语义分割领域，通过设计文本引导解码器实现了视觉-语言特征的动态对齐。相较于传统基于图像级标签的 WSSS 方法如 MCTformer 和仅依赖 CAM 种子扩展的技术，该工作创新性地利用 CLIP 文本编码器生成的类别嵌入作为条件信号，在解码阶段通过交叉注意力机制动态调制视觉特征，这一设计使得模型在 PASCAL VOC 2012 验证集上较基线方法提升 2.1% mIoU。不同于现有参数微调方案如 Tip-Adapter 需要完整更新 CLIP 参数，WeakCLIP 提出金字塔适配器，仅训练 12.4%的参数量即可实现 171.5M 模型的高效知识迁移，在保持 CLIP 原始表征能力的同时解决了传统微调导致性能下降至 58.1%的问题。

在训练策略方面，WeakCLIP 首次将对比学习损失与 KL 散度约束相结合用于 CRF 优化，通过建立像素-文本联合嵌入空间，显著改善了传统方法中因噪声种子导致的边界模糊问题。相比仅使用种子损失的基线，该多任务优化框架最终达到 74.0%的性能，验证了跨模态监督信号的互补优势。此外，针对 ViT 在分割任务中的低分辨率缺陷，提出的多尺度特征重构机制突破了现有工作如 DaFormer 的架构限制，在 COCO 2014 数据集上配合 SegFormer-MiT-B2 主干取得最优结果。这些创新共同构成了首个实现 CLIP 知识向像素级预测任务可靠迁移的完整框架，为弱监督学习与多模态预训练模型的结合提供了新方法。

5. 个人思考

在深入分析当前弱监督语义分割领域的发展现状后，我认为 WeakCLIP 所采用的跨模态知识迁移路径展现了三个关键洞见。首先，传统方法如 STC 框架和基于随机游走的标签传播主要依赖图像内部线索（如颜色一致性或空间连续性），而 WeakCLIP 通过引入 CLIP 的文本嵌入空间，构建了语义概念与像素分布的显式映射关系，这种从“视觉相似性驱动”到“语义一致性驱动”的范式转变，有效解决了传统方法在细粒度类别区分上的固有局限。其次，现有工作如 WeakTr 虽尝试将纯视觉 Transformer 应用于 WSSS，但其性能受限于单模态表示空间，而 WeakCLIP 金字塔适配器的设计创新性地实现了多尺度视觉特征与语言概念的动态对齐，这种跨模态的层次化交互机制为像素级预测提供了更丰富的上下文指导。

未来研究可沿两个维度深入：在横向扩展上，探索 CLIP 与其他基础模型（如 SAM）的协同使用，构建多模型知识融合框架；在纵向深化上，研究更精细的跨模态注意力机制，例如将当前全局的文本引导解码细化为对象级别的语义调制，这些发展方向可能最终推动 WSSS 从“减少标注依赖”逐步迈向“自主语义理解”的新阶段。