弱监督语义分割

WeakCLIP: Adapting CLIP for Weakly-Supervised Semantic Segmentation

Lianghui Zhu · Xinggang Wang · Jiapei Feng · Tianheng Cheng ·Yingyue Li · Bo Jiang ·Dingwen Zhang · Junwei Han

基本信息

出版日期：2025

期刊：International Journal of Computer Vision

影响因子：11.6

WeakCLIP: Adapting CLIP for Weakly-Supervised Semantic Segmentation

Lianghui Zhu[1] · Xinggang Wang[1] · Jiapei Feng[1] · Tianheng Cheng[1] · Yingyue Li[1] · Bo Jiang[1] · Dingwen Zhang[2] · Junwei Han[2]

Abstract
Contrastive language and image pre-training (CLIP) achieves great success in various computer vision tasks and also presents an opportune avenue for enhancing weakly-supervised image understanding with its large-scale pre-trained knowledge. As an effective way to reduce the reliance on pixel-level human-annotated labels, weakly-supervised semantic segmentation (WSSS) aims to refine the class activation map (CAM) and produce high-quality pseudo masks. Weakly-supervised semantic segmentation (WSSS) aims to refine the class activation map (CAM) as pseudo masks, but heavily relies on inductive biases like hand-crafted priors and digital image processing methods. For the vision-language pre-trained model, i.e. CLIP, we propose a novel text-to-pixel matching paradigm for WSSS. However, directly applying CLIP to WSSS is challenging due to three critical problems: (1) the task gap between contrastive pre-training and WSSS CAM refinement, (2) lacking text-to-pixel modeling to fully utilize the pre-trained knowledge, and (3) the insufficient details owing to the $\frac{1}{16}$ down-sampling resolution of ViT. Thus, we propose WeakCLIP to address the problems and leverage the pre-trained knowledge from CLIP to WSSS. Specifically, we first address the task gap by proposing a pyramid adapter and learnable prompts to extract WSSS-specific representation. We then design a co-attention matching module to model text-to-pixel relationships. Finally, the pyramid adapter and text-guided decoder are introduced to gather multi-level information and integrate it with text guidance hierarchically. WeakCLIP provides an effective and parameter-efficient way to transfer CLIP knowledge to refine CAM. Extensive experiments demonstrate that WeakCLIP achieves the state-of-the-art WSSS performance on standard benchmarks, i.e., 74.0% mIoU on the *val* set of PASCAL VOC 2012 and 46.1% mIoU on the *val* set of COCO 2014. The source code and model checkpoints are released at https://github.com/hustvl/WeakCLIP.

Keywords Semantic segmentation · Weakly-supervised Learning · CAM refinement · CLIP

Communicated by Gunhee Kim.

✉ Xinggang Wang
  xgwang@hust.edu.cn

  Lianghui Zhu
  lhzh@hust.edu.cn

  Jiapei Feng
  fjp@hust.edu.cn

  Tianheng Cheng
  thch@hust.edu.cn

  Yingyue Li

  Junwei Han

1 Introduction

Weakly-supervised semantic segmentation (WSSS) is an important task to reduce the pixel-level annotation burden which leverages weak supervisions such as image-level classification labels (Ahn & Kwak, 2018; Kolesnikov & Lampert, 2016; Pathak et al., 2015; Pinheiro & Collobert, 2015; Wang et al., 2020; Wei et al., 2017, 2018; Huang et al., 2018; Jiang et al., 2019; Lee et al., 2019a), points (Bearman et al., 2016),

研究背景

本文聚焦于**弱监督语义分割**（Weakly-Supervised Semantic Segmentation, WSSS）任务，旨在解决当前方法在处理**类激活图**（Class Activation Map, CAM）种子时面临的问题

- ➤ 传统的语义分割需要人工标注**像素级**的数据集，费时费力，限制了语义分割的实际应用
- ➤ 现有的CAM优化方法依赖于**先验知识**和**数字图像处理算法**，这些方法存在偏差，导致最后的模型缺乏鲁棒性
- ➤ **CLIP**在自然语言处理上取得了巨大的成功，同时也能高效的迁移到计算机视觉的任务，为WSSS带来了新的机遇

[1]Radford A, Kim J W, Hallacy C, et al. Learning transferable visual models from natural language supervision[C]//International conference on machine learning. PmLR, 2021: 8748-8763.

研究现状

➢ 为了减轻手动标注像素级数据的负担，出现了多种基于不同弱监督信息（**边界框，涂鸦，点和图像级**）的算法。其中基于图像级标签的WSSS最具有挑战性，常使用CAM定位目标，但原始的CAM存在噪声数据，容易出错

➢ **大规模预训练模型**在各领域广泛应用，如CLIP通过对比学习在大量**图像-文本**对上预训练，展现出强大的知识迁移能力。已有研究尝试将CLIP应用于WSSS
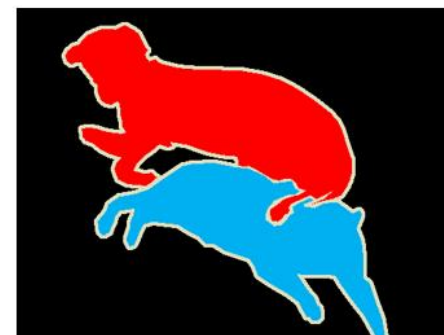
| 图像级标注 | 物体点标注 | 物体框标注 | 物体划线标注 | 像素级标注 |

研究方法

- **Learnable Prompt：** 用可学习嵌入增强类别文本输入，经文本编码器处理和投影得到文本嵌入，满足WSSS像素级理解需求

- **Pyramid Adapter：** 独立于CLIP图像编码器，对不同分辨率特征图处理，获取多尺度特征，避免破坏预训练知识

- **Co-attention Matching：** 用两个交叉注意力模块建模文本与像素双向关系，更新文本和图像嵌入，得到匹配嵌入

- **Text-Guided Decoder：** 将匹配嵌入调整到对应尺寸，与适配器输出特征拼接解码，实现分层融合，提升分割预测效果
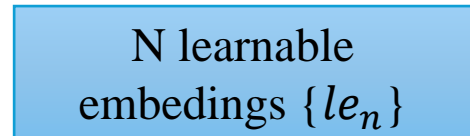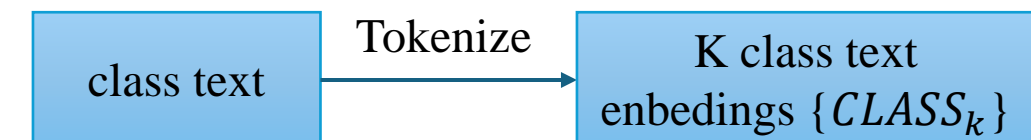
- **WSSS loss：** 平衡种子损失和边界损失

t_p: projected text embedings
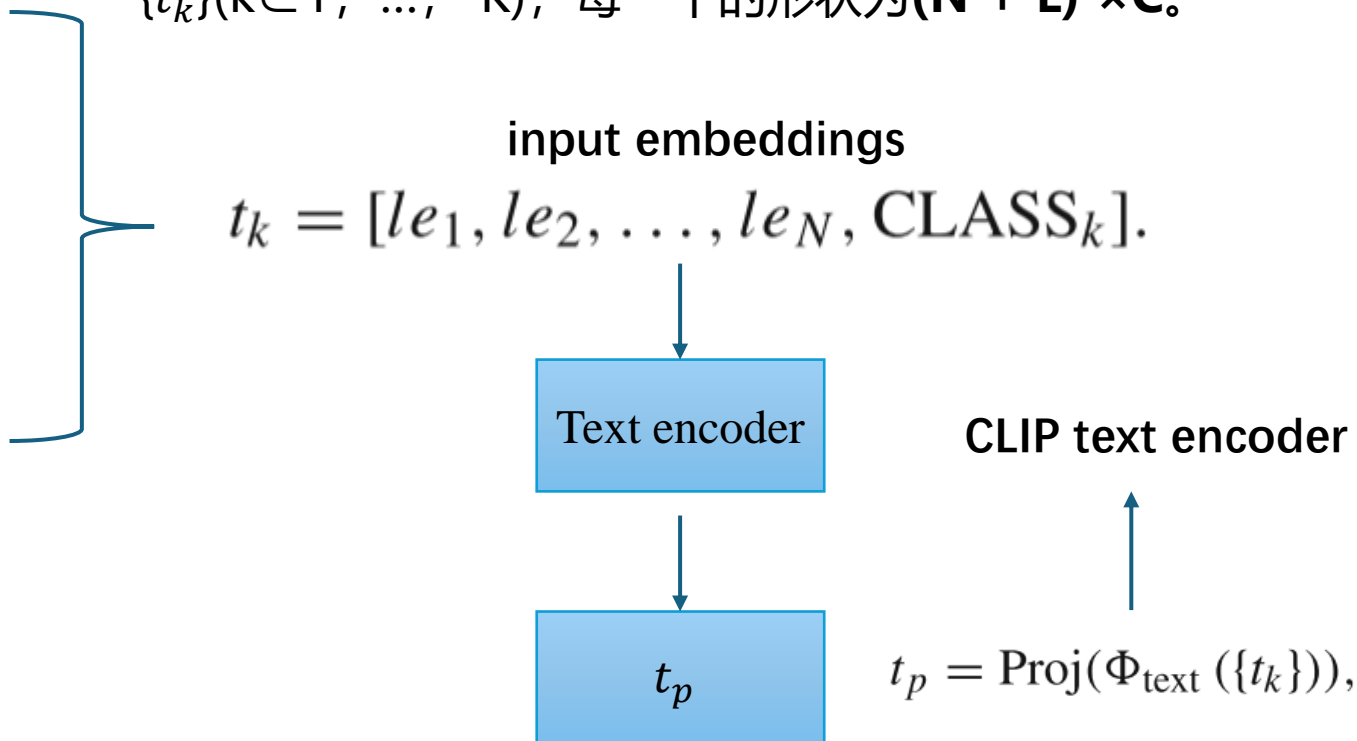
f_p: projected images embedings
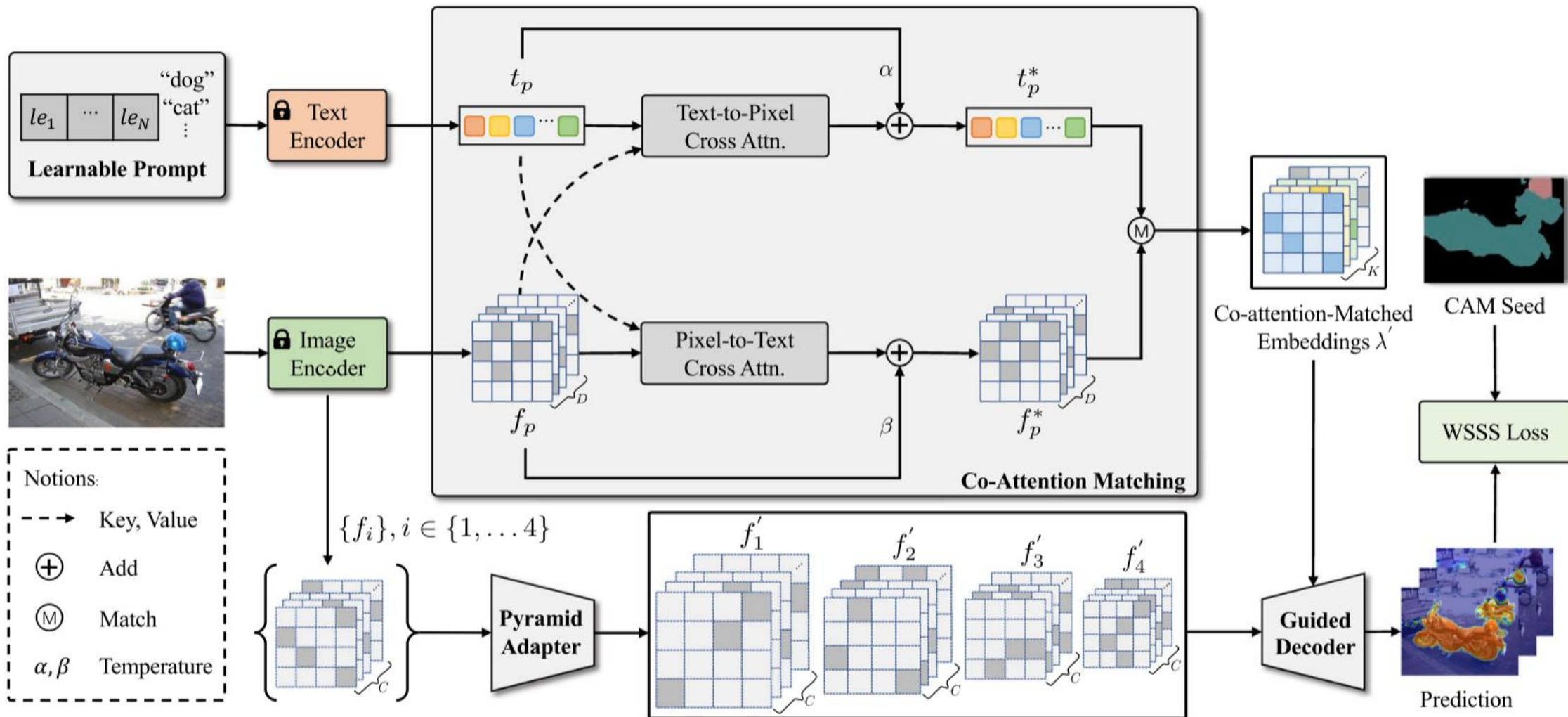
Learnable Prompt

我们首先对类文本进行标记并嵌入到K个类文本嵌入$\{CLASS_k\}$ (k∈1，...，K)，每个的形状为L × C，L是类文本标记的上下文长度，C是transformer宽度。

接下来，我们随机初始化N个可学习嵌入{len}(N∈1，...，N)作为可学习的提示词，每个提示的形状为1 × C。随后，将可学习嵌入拼接到类文本嵌入前面，作为输入嵌入$\{t_k\}$(k∈1，...，K)，每一个的形状为**(N + L) ×C**。

```
class text  --Tokenize-->  K class text
                           enbedings {CLASS_k}

N learnable
embedings {le_n}

learnable prompt
```

input embeddings

$$t_k = [le_1, le_2, \ldots, le_N, \text{CLASS}_k].$$

Text encoder

CLIP text encoder

t_p

$$t_p = \text{Proj}(\Phi_{\text{text}}(\{t_k\})),$$

Pyramid Adapter
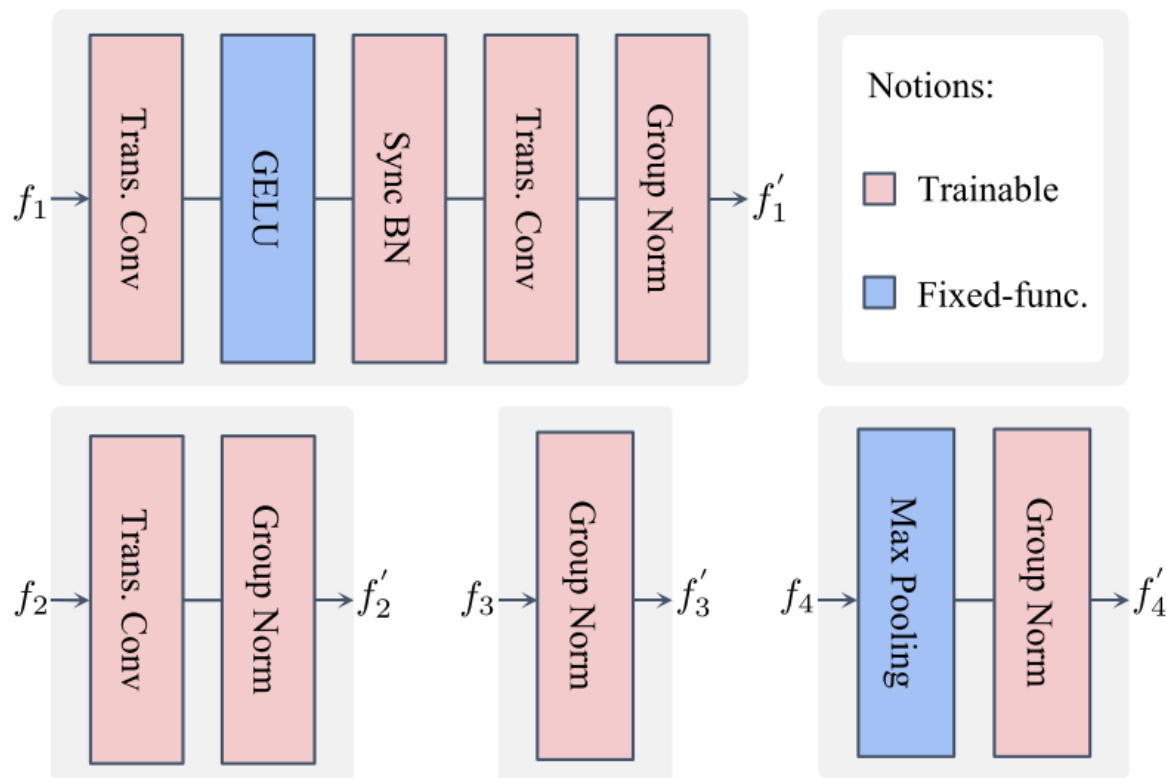
金字塔适配器首先使用来自于**transformer层**的 feature maps$\{f_i\}$(i∈{1，…，4})作为输入。随后，利用转置卷积对f_1和f_2进行上采样。然后，使用最大池化将f_4下采样。最后，金字塔适配器生成一组特征$\{f_i'\}$(i∈{1，…，4})，每一个f_i'拥有不同的分辨率 ($\frac{1}{4}$, $\frac{1}{8}$, $\frac{1}{16}$, $\frac{1}{32}$)，有效地结合了**低级特征和高级特征**

Co-attention Matching

交叉注意模块输出的**文本和图像**通过残差连接与原始t_p和f_p融合，α和β用于平衡交叉注意力输出的影响

learnable temperatures

learnable temperatures

$$t_p^* = \mathrm{CrossAttn}(t_p, f_p) * \alpha + t_p, \qquad f_p^* = \mathrm{CrossAttn}(f_p, t_p) * \beta + f_p,$$

updated text embeddings

updated image embeddings

最后，我们在t_p^*和f_p^*之间进行**文本到图像的匹配**，以获得协同注意力匹配嵌入λ'

$$\lambda' = \mathrm{Match}(t_p^*, f_p^*),$$

λ'的形状：$K \times H \times W$

Text-Guided Decoder

C.B.R. 表示 "Conv-BN-ReLU" 层
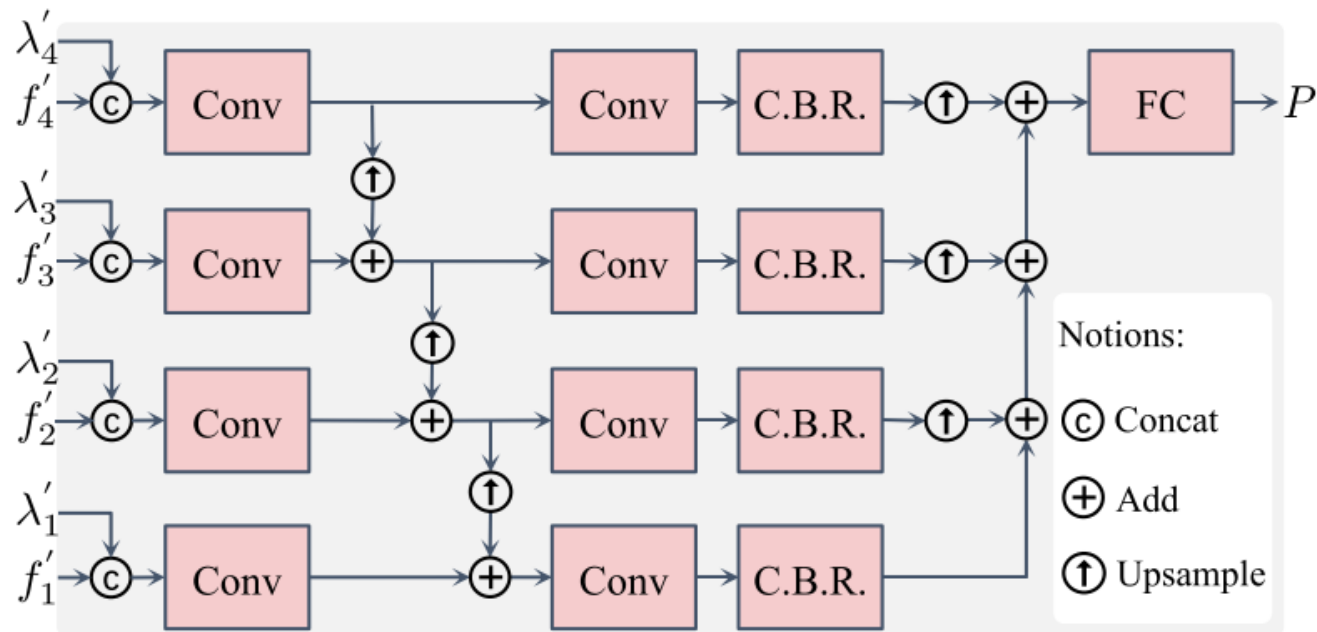
我们首先将协同注意匹配的嵌入λ'调整到与金字塔适配器输出特征$\{f_i'\}$相对应的大小。$\{\lambda_i'\}$表示不同空间大小的协同注意匹配嵌入的集合。然后将$\{\lambda_i'\}$和$\{f_i'\}$拼接在一起，对其进行解码，得到**分割预测P**

$$P = \text{Decoder}_{g}\left(\{f_i'\}, \{\lambda_i'\}\right), i \in \{1, \ldots 4\}.$$

金字塔适配器输出特征和协同注意力匹配嵌入的分层融合产生更具有**鲁棒性**的的分割预测

$$S = \arg\max\left(P\right).$$

在推理过程中，我们对P进行argmax运算，得到最终的**分割结果S**

WSSS loss

WSSS loss = seeding loss + boundary loss

$$L_{\text{seed}} = -\frac{1}{\sum_{k \in K} |M_k|} \sum_{k \in K} \sum_{u \in M_k} \log P_{u,k}$$

M_k：分类为k类别的像素集

$P_{u,k}$：k类别在u位置的分割预测

CRF processed result

$$L_{\text{boundary}} = \frac{1}{n} \sum_{u=1}^{n} \sum_{k \in K} \Psi_{u,k}(I, P) \cdot \log \frac{\Psi_{u,k}(I, P)}{P_{u,k}}.$$

I：原始图像

P：分割预测

n：位置总数

[1]Krähenbühl P, Koltun V. Efficient inference in fully connected crfs with gaussian edge potentials[J]. Advances in neural information processing systems, 2011, 24.

生成伪掩膜

我们使用训练好的WeakCLIP网络生成高质量的**伪掩膜，最后，我们利用生成的****伪掩膜****来训练全监督分割网络，实现****弱监督语义分割**

实验过程

数据集：PASCAL VOC 2012、MS COCO 2014

评估指标：mean intersection over union (mIoU)

Baseline：使用MCTformer产生的CAM作为基准，并训练WeakCLIP模型，产生**伪掩膜**

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
OC-CSE (Kweon et al., 2021)	ResNet38		68.4	68.2
ReCAM (Chen et al., 2022b)	ResNet101		68.5	68.4
CPN (Zhang et al., 2021a)	ResNet38		67.8	68.5
RIB (Lee et al., 2021a)	ResNet101		68.3	68.6
AMR (Qin et al., 2022)	ResNet101		68.8	69.1
VWE (Ru et al., 2022)	ResNet101		70.6	70.7
URN (Li et al., 2022b)	Res2Net101		71.2	71.5
CLIMS (Xie et al., 2022)	ResNet101		70.4	70.0
SANCE (Li et al., 2022a)	ResNet101		70.9	72.2
SIPE (Chen et al., 2022a)	ResNet101		68.8	69.7
W-OoD (Lee et al., 2022a)	ResNet101		70.7	70.1
AMN (Lee et al., 2022b)	ResNet101		69.5	69.6
ViT-PCM (Rossetti et al., 2022)	ResNet101		70.3	70.9
AEFT (Yoon et al., 2022)	ResNet38		70.9	71.7

Baseline and our WeakCLIP

MCTformer (Xu et al., 2022)	ResNet38	$\mathcal{I}$	71.9	71.6
WeakCLIP	ResNet38	$\mathcal{I}$	74.0	73.8

full supervision (F), image-level labels (I), saliency maps (S), and bounding box labels (B)

WeakCLIP在验证集和测试集上的mIoU分别达到**74.0%**和**73.8%**，优于其他仅使用图像级监督的方法，以及部分使用额外显著图监督或边界框监督的方法。

MS COCO 2014

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
WSSS methods with plain ViT retraining			(Strudel et al., 2021)
WeakTr (Rong et al., 2023)	DeiT-S	$\mathcal{I}$	46.9
WeakCLIP	DeiT-S	$\mathcal{I}$	47.3
WSSS methods with hybrid ViT retraining			(Xie et al., 2021)
WeakCLIP	MiT-B2	$\mathcal{I}$	**47.4**

WeakCLIP在验证集上的mIoU达到**46.1%**，比基准MCTformer提高了4.1%，优于其他仅使用图像级监督的方法。使用SegFormer和MiT-B2骨干进行再训练，WeakCLIP在COCO 2014验证集上取得最佳性能。

Number of learnable embeddings	2	4	8	16
mIoU	72.2	72.3	**72.6**	72.3

We mark the best WSSS results in bold

当可learnable embedings的数量设置为8时，性能最好，而其他个数设置的效果略差。这些发现表明， learnable embedings的数量确实会影响结果，选择适当数量的learnable embedings可以增强弱监督语义分割场景下的文本表示。

消融实验

Initial value of temperatures	1	1e−1	1e−2	1e−3
mIoU	72.5	**72.6**	72.3	72.2

我们还进行了消融研究，以探讨可学温度初始值对共同注意匹配的影响，使用初始值为0.1的可学习温度可以获得最佳性能。

消融实验

Table 7 Comparison of per-class segmentation results in terms of IoUs on the PASCAL VOC 2012 *val* set

Method	bkg	Plane	Bike	Bird	Boat	Bottle	Bus	Car	Cat	Chair	Cow
MCTformer[CVPR22] (Xu et al., 2022)	91.9	78.3	39.5	89.9	55.9	76.7	81.8	**79.0**	90.7	32.6	**87.1**
WeakCLIP (ours)	**92.7**	**87.1**	**40.6**	**89.9**	**63.0**	**78.3**	**86.8**	77.9	**90.7**	**33.0**	84.8

Method	Table	Dog	Horse	mbk	Person	Plant	Sheep	Sofa	Train	Tv	mIoU
MCTformer[CVPR22] (Xu et al., 2022)	**57.2**	87.0	**84.6**	77.4	79.2	55.1	**89.2**	47.2	70.4	58.8	71.9
WeakCLIP (ours)	48.4	**88.2**	83.8	**78.4**	**81.4**	**64.9**	87.8	**53.6**	**76.4**	**66.6**	**74.0**

We mark the best WSSS results in bold

Table 8 Comparison of per-class segmentation results in terms of IoUs on the PASCAL VOC 2012 *test* set

Method	bkg	Plane	Bike	Bird	Boat	Bottle	Bus	Car	Cat	Chair	Cow
MCTformer[CVPR22] (Xu et al., 2022)	92.3	84.4	37.2	82.8	**60.0**	**72.8**	78.0	79.0	89.4	31.7	**84.5**
WeakCLIP (ours)	**92.9**	**88.4**	**40.6**	**88.3**	57.6	71.8	**82.6**	**80.0**	**89.9**	**33.1**	82.6

Method	Table	Dog	Horse	mbk	Person	Plant	Sheep	Sofa	Train	Tv	mIoU
MCTformer[CVPR22] (Xu et al., 2022)	**59.1**	85.3	**83.8**	79.2	**81.0**	53.9	85.3	60.5	65.7	57.7	71.6
WeakCLIP (ours)	57.6	**86.0**	82.9	**83.9**	79.6	**66.6**	**86.8**	**60.9**	**72.8**	**64.2**	**73.8**

We mark the best WSSS results in bold

结果表明，WeakCLIP在大多数分类中都取得了优异的性能，证明了其在弱监督语义分割(WSSS)领域的有效性。

消融实验

	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

Furthermore, WeakCLIP presents 4.3× faster training FPS than baseline and saves 68.4% GPU memory

MCTformer具有**105.3M**参数，相比之下，WeakCLIP有**171.5 M**参数，但只训练**21.2 M**参数。WeakCLIP只有12.4%的可训练参数，训练FPS速度提高了4.3倍，节省了68.4%的GPU内存。

可视化结果

我们利用WeakCLIP网络产生的伪掩膜，利用全监督网络训练的分割模型取得了较好的成绩，达到了SOTA的效果

总结
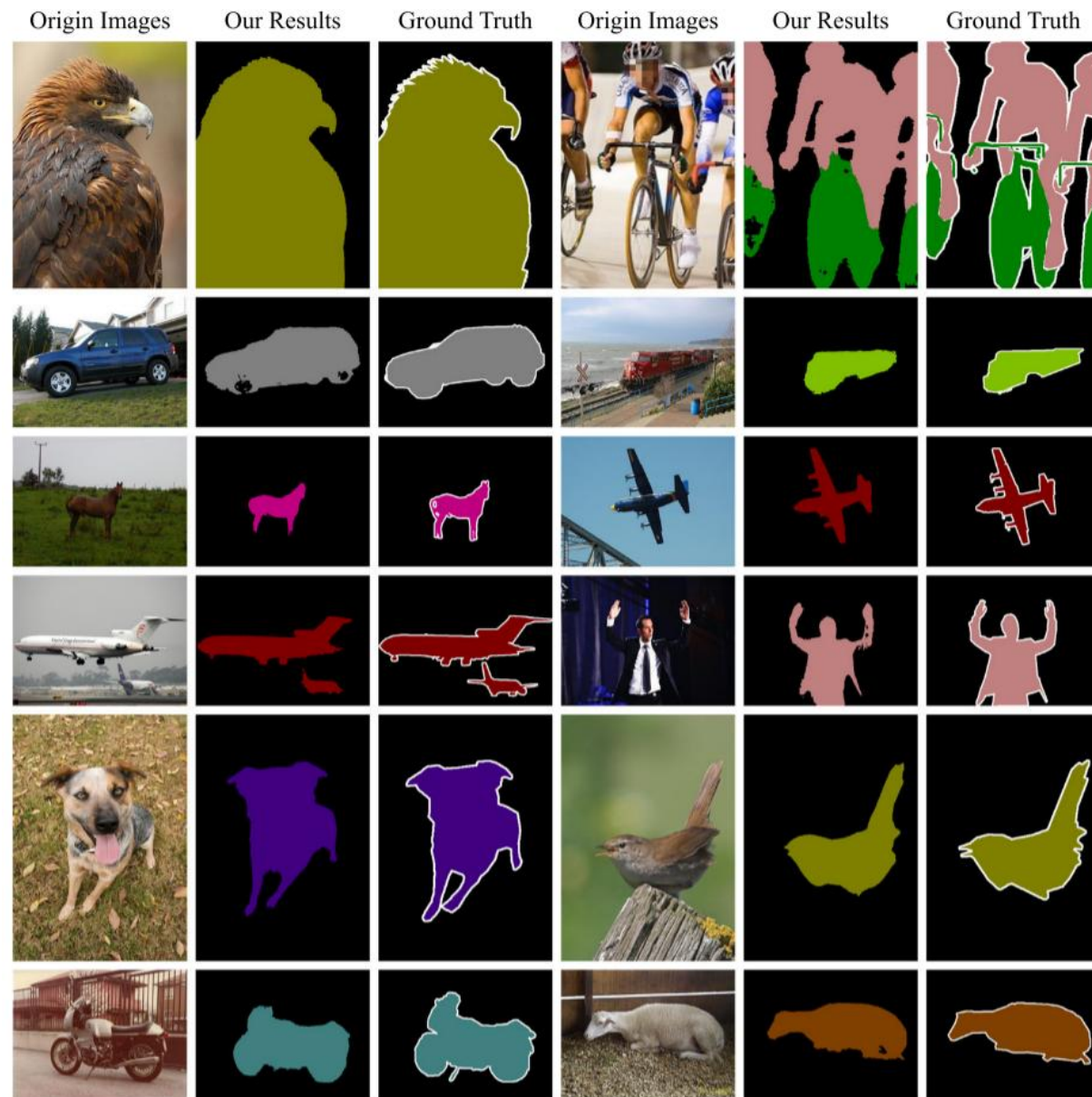
本文提出了名为**WeakCLIP**的新方案，旨在利用预训练CLIP模型的知识来增强弱监督语义分割网络的CAM优化过程，有效解决了将CLIP集成到WSSS中存在的问题。在广泛使用的PASCAL VOC 2012和COCO 2014数据集上的实验结果表明，与以往的WSSS方法相比，WeakCLIP取得了明显的提升。未来，作者计划探索更先进的大规模CLIP，以提升WSSS的像素级理解能力。

Thank you