

SS-EPA：基于补丁语义亲和力增强的单阶段弱监督语义分割方法

姜景杰¹⁾ 周肖桐¹⁾ 杨欢¹⁾ 张轩豪¹⁾ 杨钧茗¹⁾

¹⁾(南京信息工程大学, 计算机学院、网络空间安全学院, 南京, 中国, 210044)

摘要 图像级弱监督语义分割 (Image Level Weakly Supervised Semantic Segmentation, Image Level WSSS) 通常利用类激活图 (Class Activation Map, CAM) 来生成伪标签 (Pseudo Label)。先前WSSS方法无论使用CNN还是Transformer框架，多数都采用多阶段方法，需要分阶段地训练模型和采取不同的训练策略，多个阶段间的复杂交互较为繁琐。且先前的方法通常直接通过ViT中的语义亲和力优化CAM，对计算资源要求较高，还可能会给CAM带来错误和误导。本文提出了一种名为SS-EPA (Single Stage WSSS with Enhanced Patch Affinity) 的单阶段WSSS方法，集成了端到端式的多头自注意力CAM优化方法，利用ViT中的补丁语义亲和力 (Patch Affinity) 信息，对从补丁令牌生成的初始CAM执行优化。为了进一步解决语义亲和力中噪声、错误和注意力图过于庞大的问题，本文提出了头平均注意力融合增强模块 (Head Average Attention Fusion, HAAF)。HAAF通过对ViT中不同头的注意力权重执行平均操作来聚合语义信息，去除头重复关注、包含无效信息的冗余问题，提取更加精简有效的信息，极大减少计算资源占用。在Pascal VOC 2012数据集上的实验表明，本文所提方法可以显著优化生成的CAM和伪标签，语义分割性能在验证集和测试集上分别达到了72.4%和73.3%。相较先前的单阶段方法，SS-EPA误分类的概率更小，且有更加完整和准确的对象边界，充分验证了本文方法的有效性。

关键词 计算机视觉；语义分割；弱监督学习；Transformer；语义亲和力

SS-EPA: A Single-Stage Weakly Supervised Semantic Segmentation Method Based on Enhanced Patch Affinity

Jingjie Jiang¹⁾ Xiaotong Zhou²⁾ Huan Yang³⁾ Xuanhao Zhang³⁾ Junming Yang³⁾

¹⁾(School of Computer Science, Nanjing University of Information Science&Technology, Nanjing, 210044, China)

Abstract Image Level Weakly Supervised Semantic Segmentation (Image Level WSSS) often uses Class Activation Maps (CAM) to generate pseudo labels. Previous WSSS methods, whether based on CNN or Transformer frameworks, typically follow multi-stage approaches, requiring separate training phases and complex interactions. These methods also optimize CAM using semantic affinity in ViT, demanding high computational resources and introducing potential errors. This paper proposes a single-stage WSSS method called SS-EPA (Single Stage WSSS with Enhanced Patch Affinity), which integrates an end-to-end CAM optimization process using patch affinity information from ViT. To address issues of noise and large attention maps, a Head Average Attention Fusion (HAAF) module is introduced, which averages attention weights from different heads, reducing redundant and irrelevant information, and lowering computational costs. Experiments on the Pascal VOC 2012 dataset show that SS-EPA significantly improves CAM and pseudo labels, achieving segmentation accuracy of 72.4% on the validation set and 73.3% on the test set. Compared to previous single-stage methods, SS-EPA reduces misclassification and produces more complete object boundaries, demonstrating its effectiveness.

Keywords Computer Vision, Semantic Segmentation, Weakly Supervised Learning, Transformer, Semantic Affinity

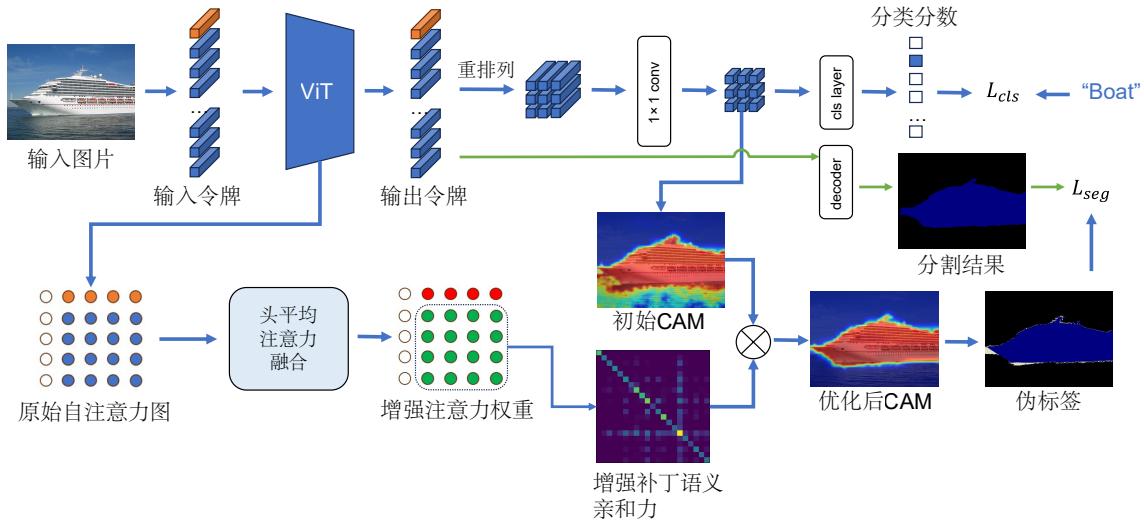


图 1 SS-EPA 整体框架。SS-EPA 首先将输入图像分成多个补丁，每个补丁线性转换成补丁令牌，并连接一个类别令牌。ViT输出令牌一方面经过重拍列和 1×1 卷积生成初始CAM，并由分类层输出分类分数。另一方面经过分割decoder生成语义分割结果。然后从ViT中提取多头自注意力图，并通过HAAF模块进行增强，获取增强补丁语义亲和力。最后通过增强补丁语义亲和力对初始CAM进行优化，并生成伪标签用于监督分割任务。

1 引言

深度学习推动图像分割显著进步，但依赖精确标注样本，成本高昂。弱监督语义分割（WSSS）技术因此出现，WSSS仅需粗略标注的样本，大幅降低了样本获取的难度。弱监督标注分为边框级标注[1, 2]、涂鸦级标注[3]和图像级标注[4, 5]。其中最难利用的是图像级标注，因为它通常只提供图片的分类标签，包含极少可利用的语义信息，也是多数研究人员最热衷于研究的方法之一。本工作只使用图像级弱标注。

先前的WSSS方法通常采用多阶段方法，即先训练一个分类网络来生成类激活图Class Activation Map (CAM) [6]，获取类别在图像中的位置信息。然后通过扩展优化CAM生成伪标签，最后利用伪标签全监督地训练语义分割模型。虽然多阶段方法通常可以获得更精准的CAM和更优秀的分割性能，但通常需要分阶段地训练模型和不同的训练策略，耗费大量的计算资源和时间来进行训练和优化，这限制了其在大规模数据集或实时应用中的实用性。通过CNN生成的CAM，存在只激活最显著区域的缺点，原因是CNN感受

野有限，对全局信息捕获不完善。Vision Transformer (ViT) [7]在其它视觉任务中的巨大成功，引起了WSSS领域研究人员的广泛关注[4, 5, 8]。ViT是一种基于Transformer架构的视觉模型，它将输入图片划分为小的补丁（patch），并利用Transformer来建模这些补丁之间的关系。且ViT采用无卷积架构，避免了卷积带来的先验约束，如局部性和平移不变性，使ViT具有更好的可扩展性。基于ViT的WSSS方法首先通过补丁令牌生成粗略的初始CAM，再利用多头自注意力中包含的语义亲和力信息对初始CAM进行优化[5]。然而Transformer中不同深度层的多头自注意力可能关注不同部分，如浅层更关注局部结构、纹理颜色等，深层能捕获更广泛和抽象的视觉语义信息，直接将其与CAM相乘可能会产生错误与误导。且ViT注意力图十分庞大，直接提取所有注意力权重会占据大量计算资源。

本文提出一种单阶段WSSS方法SS-EPA (Single Stage WSSS with Enhanced Patch Affinity)，和一种头平均注意力融合增强模块 (Head Average Attention Fusion, HAAF)。针对先前CAM优化多数集中在多阶段方法上的问题，本文提出一

种单阶段WSSS方法SS-EPA，集成了端到端式多头自注意力CAM优化方法。SS-EPA从ViT的自注意力中提取补丁语义亲和力信息，并用于优化初始CAM。本文将该端到端的优化方法集成到单阶段WSSS方法中，不会影响其完整性和一致性。针对注意力图较为庞大，且不同深度的注意力特性各不相同的问题，本文提出一种头平均注意力融合增强模块HAAF。HAAF对来自不同层多头自注意力中的语义亲和力进行融合增强，并用于优化从Transformer生成的CAM。HAAF通过对多头自注意力中的各头权重进行平均，聚合不同语义信息，减少不同头重复关注相似区域的冗余信息。随后，通过全局平均池化聚合每个注意力图的全局特征，并将其输入多层感知机，提取更复杂的特征关系。最终获得融合后的增强注意力图，充分考虑了不同层次注意力的重要性。HAAF可以去除头重复关注、包含无效信息的冗余问题，显著降低计算资源消耗并提升效率。该方法还能减少每个头对噪声或异常的敏感度，提高模型鲁棒性。

为了验证本文提出的SS-EPA和HAAF模块的有效性，本文在Pascal VOC 2012数据集上评估了SS-EPA的CAM、伪标签和分割结果的性能表现，并与基线方法ToCo[4]相比较。对比实验、消融实验以及各种可视化结果表明，本文所提方法可以显著优化生成的CAM，伪标签和分割模型误分类的概率更小，且有更加完整和准确的对象边界。在VOC验证集上与基线相比，伪标签和分割性能分别提升了2.2%和1.3%的mIoU分数，充分验证了本文方法的有效性。

总的来说，本文主要贡献包括以下三个方面：

- 提出了一种名为SS-EPA的单阶段WSSS方法，集成了端到端式多头自注意力CAM优化方法。在不影响单阶段方法的完整性和一致性的前提下，集成了利用补丁语义亲和力信息优化初始CAM的方法，使CAM更加精细和准确，从而生成更加优质的伪标签用于训练分割模型。

- 提出一种头平均注意力融合增强模块(HAAF)，来解决语义亲和力信息包含噪声与错误，以及注意力图较为庞大的问题。通过对注意力的不同头的权重做平均，HAAF可去除冗余信息并提高模型鲁棒性，利用多层感知机的交互能力，HAAF可以充分考虑来自不同层注意力的重要性，对包含语义亲和力的自注意力完成简化和增强。

- 在Pascal VOC 2012数据集上的实验表明，本文方法可以显著优化生成的CAM，最终的分割模型性能相比以前的单阶段方法有了实质性的改进，且实现了与一些多阶段方法相当的性能。

2 相关技术

2.1 图像级弱监督语义分割

图像级WSSS是所有WSSS方法中挑战性最大的一种。该方法仅依赖于图像的分类标签。与其它形式的弱标注，如涂鸦标注和边框标注相比，图像级标注所提供的信息量更为有限，因此在实现像素级精细分割的任务上难度更高。利用图像级标注的WSSS方法通常生成CAM来获取图像分类时的关注区域，从而捕获特定于类别的定位信息[9]。但CAM激活的区域通常只会覆盖最明显最具判别力的对象区域，而忽略其它非判别性的区域。一些工作专注于生成高质量CAM，如对抗性擦除方法[10]，通过擦除最具判别力的对象区域来迫使模型关注其它非判别性的区域，可以一定程度上优化缓解CAM激活不全的问题。还有工作利用子类别探索[11]、自监督注意力机制[12]和多图像语义信息[13]来获得更精准的CAM。最近也有工作[14, 15]利用了CLIP[16]模型，利用CLIP对图像和文本强大的上下文理解能力抑制背景像素的激活，更专注于前景区域。然而这些方法大多集中于多阶段WSSS方法，且需要分阶段地训练模型和不同的训练策略，多个阶段间的复杂交互较为繁琐。本文提出的单阶段WSSS方法SS-EPA，集成了端到端式多头自注意力CAM优化方法，减少了流程复杂

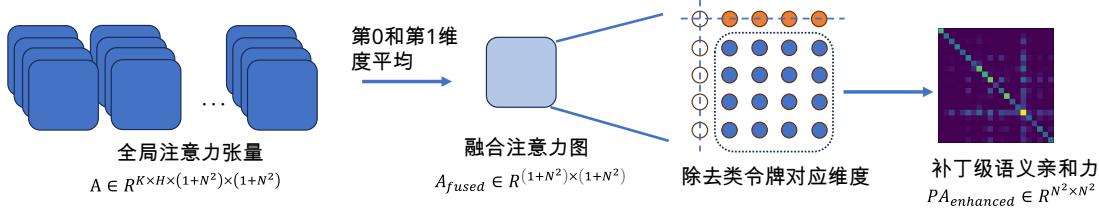


图 2 补丁语义亲和力获取流程图

性。

2.2 弱监督语义分割中的ViT

先前的WSSS方法大多建立在CNN网络之上，存在只激活最显著区域的缺点。ViT[7]凭借其强大的全局上下文建模能力，在WSSS任务中取得成功。TS-CAM[17]提出借助ViT和多头自注意力的特性生成CAM，来充分利用ViT的长距离建模能力。MCTformer[5]强调了ViT中类令牌的重要性，通过嵌入多个类令牌并强制它们学习不同类的激活图，并利用特定类别的注意力图优化CAM。AFA[8]通过额外的模块学习多头自注意力中的语义亲和力，改善了CAM的覆盖区域，缓解了CAM难以捕捉完整的目标区域的问题。ToCo[4]通过利用ViT中间层的伪标记关系来监督最终的补丁标记，从而解决ViT的过度平滑问题。然而，先前的方法通常利用ViT中的语义亲和力优化CAM，对计算资源要求较高，且直接利用可能会给CAM带来错误和误导。且ViT中多头注意力的设计目的是为了捕捉不同依赖关系，但实践中一些注意力头往往关注相似的区域或信息，导致不同头之间存在相似性，产生冗余。本文提出的头平均注意力融合增强模块（HAAF）通过多头平均化去除上述冗余信息，淡化可能捕捉到的噪声或者无效的注意力模式，减少单个头对特定噪声的敏感度，提高模型的鲁棒性。

3 方法

3.1 概述

本节首先介绍了SS-EPA的整体框架，SS-EPA是一种改进的单阶段WSSS方法，集成了端到端式多头自注意力CAM优化方法。SS-EPA首先通过ViT对输入图片分类，并生成初始CAM。然后结合补丁语义亲和力优化CAM，并生成伪标签用于监督分割任务，实现图像分类和语义分割的联合

学习。如图1所示，SS-EPA是单阶段WSSS方法，集成了端到端式多头自注意力CAM优化方法，相比传统多阶段方法简化了流程。本文提出了头平均注意力融合增强模块（HAAF），来去除不同头重复关注相似区域的冗余信息，并提高模型鲁棒性。解决了多头自注意力图较为庞大，且直接利用语义亲和力会带来噪声与错误的问题[5]。

3.2 SS-EPA框架

SS-EPA首先将输入图片拆分为 $N \times N$ 个补丁，并通过线性转换为补丁令牌序列 $T_{patch} \in R^{N^2 \times D}$ ，其中D是嵌入维度。生成一个维度同样为D的类令牌 $T_{cls} \in R^{1 \times D}$ ，将类令牌与补丁令牌链接，并添加位置编码构成ViT编码器的输入令牌序列 $T_{input} \in R^{(1+N^2) \times D}$ 。ViT backbone具有K个Transformer编码层，每个编码层包含一个多头自注意力和一个多层次感知机，以及分别用于两个子层前的层归一化。ViT编码器接收输入令牌序列 $T_{input}^i, i = (1, 2, \dots, K)$ ，并输出令牌序列 $T_{out}^i \in R^{(1+N^2) \times D}, i = (1, 2, \dots, K)$ 。最后一层Transformer编码层的输出令牌序列 $T_{out}^K \in R^{(1+N^2) \times D}$ ，去除类令牌对应维度并重排列可得补丁令牌序列 $T_{out_patch} \in R^{N \times N \times D}$ ，并执行 1×1 卷积操作将令牌维度变为物体类别数量，公式如下：

$$CAM = conv_{1 \times 1}(T_{out_patch}) \quad (1)$$

其中， $conv_{1 \times 1}$ 的输入通道为D，输出通道为物体类别数C，卷积核大小为 1×1 。通过上式可获得来自补丁令牌的初始类激活图 $CAM \in R^{N \times N \times C}$ 。参照[8]的方法，通过一个全局最大池化层（Global Max Pooling）

来聚合补丁令牌 $T_{\text{out_patch}}$ 信息，然后通过全连接层来计算分类分数 cls_score ，分类损失函数使用多标签软边距损失（Multi Label Soft Margin Loss）作为损失函数 L_{cls} ，公式如下：

$$L_{\text{cls}}(x, y) = -\frac{1}{C} \sum_{i=1}^C [y \log(\sigma(x)) + (1 - y) \log(1 - \sigma(x))] \quad (2)$$

其中 x, y 分别是模型预测分数与真值标签， $\sigma(x)$ 表示Sigmoid函数的输出，即：

$$\sigma(x) = \frac{1}{1 + e^{-x}} \quad (3)$$

ViT backbone使用的是一标准的Transformer多头自注意力，首先将输入令牌归一化，并通过全连接层将其转换为一个查询 $Q \in R^{(1+N^2) \times D}$ 和一组键值 $K \in R^{(1+N^2) \times D}, V \in R^{(1+N^2) \times D}$ ，注意力计算采用[18]中的缩放点积注意力（Scaled Dot-Product Attention），计算公式如下：

$$\text{Attn}(Q, K, V) = \left(\text{Softmax} \frac{QK^T}{\sqrt{D}} \right) V \quad (4)$$

从中可以提取多头自注意力图 $A_{\text{map}} = QK^T$ ，其中 $A_{\text{map}}^i \in R^{H \times (1+N^2) \times (1+N^2)}, i = 1, \dots, K$ ， H 为多头自注意力头的个数。此操作不会带来任何额外的计算资源消耗，因为多头自注意力权重是Transformer在计算时产生的副产物。然后在第0个维度上进行concatenate操作将 K 层注意力图串联起来，获得全局注意力张量 $A \in R^{K \times H \times (1+N^2) \times (1+N^2)}$ ，该注意力张量十分庞大，在3.3节中将讨论如何减小计算资源占用。

全局注意力张量 A 中蕴含了补丁语义亲和力信息，将注意力图 A 在第0和第1个维度上进行平均来聚合来自不同层和不同头的注意力信息，得到 $A_{\text{fused}} \in R^{(1+N^2) \times (1+N^2)}$ ，除去其中类令牌对应的维度，如图2所示，剩下的注意力权重可作为补丁级语义亲和力PatchAffinity $\in R^{N^2 \times N^2}$ 。由于从补丁令牌生成的初始类激活图CAM存在大量噪声与错误，所以需要补丁级语义亲和力对其进行优化，优化公式如下：

$$\text{CAM}_{\text{refined}} = \text{PatchAffinity} \times \text{CAM} \quad (5)$$

通过上式可获得通过原始补丁语义亲和力优化后的类激活图 $\text{CAM}_{\text{refined}} \in R^{N \times N \times C}$ ，相比初始CAM对目标的覆盖性更好，错误激活区域更少，且可以激活更多目标区域。

3.3 头平均注意力融合增强模块（HAAF）

鉴于Transformer中不同深度的层的多头自注意力可能关注不同部分，如浅层更关注局部结构、纹理颜色等，深层能捕获更广泛和抽象的视觉语义信息，所以不能简单地将来自不同层的多头自注意力平均来聚合语义信息。且一个标准ViT backbone (vit_base_patch16_224) 的多头自注意力图十分庞大 (batchsize为2时，显存占用超过12GB)，对计算资源要求较高。本文提出头平均注意力融合增强模块（HAAF）来解决上述问题，如图3所示。

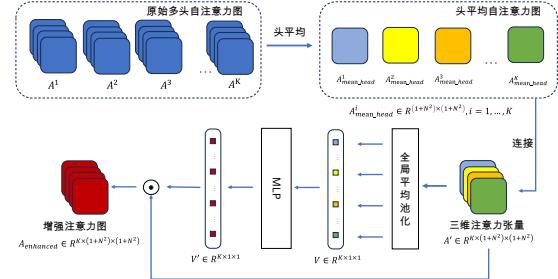


图3 HAAF结构图

对于3.2节中从backbone中提取的多头自注意力图 $A_{\text{map}}^i \in R^{H \times (1+N^2) \times (1+N^2)}, i = 1, \dots, K$ ，HAAF首先采用头平均操作去除维度 H ，有助于去除冗余信息并减少 H 倍的显存占用，得到 $A_{\text{mean_head}}^i \in R^{(1+N^2) \times (1+N^2)}, i = 1, \dots, K$ 。然后在第0个维度上进行concatenate操作将 K 层注意力图串联起来，获得全局注意力张量 $A' \in R^{(K \times (1+N^2) \times (1+N^2))}$ 。全局平均池化通过平滑特征表示和增强泛化能力，相较于全局最大池化，在减少噪声和防止过拟合方面更具优势。所以本文通过全局平均池化聚合 K 层注意力图的全局特征，得到聚合后的长度为 K 的特征向量 $V \in R^{K \times 1 \times 1}$ ，并将特征向量

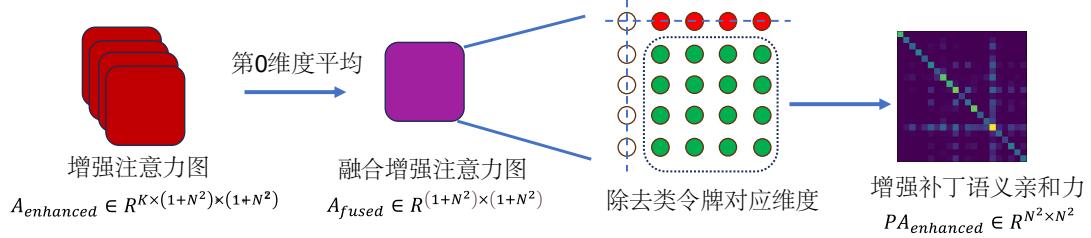


图 4 增强补丁语义亲和力获取流程图

输入多层感知机中相互作用，提取更复杂的特征相互关系，多层感知机输出相同形状的特征向量 $V' \in R^{(K \times 1 \times 1)}$ 。获得多层感知机输出的特征向量 V' 后，将全局注意力张量 A' 与特征向量 V' 结合，公式如下：

$$A'_{enhanced} = A' \odot V' \quad (6)$$

其中 \odot 表示逐元素相乘符号。通过上式可获得充分考虑了不同层注意力重要性的增强注意力图 $A'_{enhanced} \in R^{K \times (1+N^2) \times (1+N^2)}$ 。经过头平均后的注意力图更加稳定，不易受到单个注意力头学习偏差的影响。

图4展示了不同层增强注意力图的融合过程，对增强注意力图 $A'_{enhanced}$ 在第0个维度 K 上进行平均操作，可得到融合增强注意力图 $A'_{fused} \in R^{(1+N^2) \times (1+N^2)}$ 。除去其中类令牌对应的维度，剩下的增强注意力权重可作为增强后的补丁级语义亲和力 $PA_{enhanced} \in R^{N^2 \times N^2}$ ，如图4所示。通过HAAF增强后的补丁语义亲和力 $PA_{enhanced}$ 相比增强前减少了噪声与错误，并且充分考虑了不同层注意力的重要性。通过特征向量 V' 对每层注意力进行加权。最后利用 $PA_{enhanced}$ 对CAM优化，过程与3.2节介绍的CAM优化过程类似，优化公式如下：

$$CAM'_{refined} = PA_{enhanced} \times CAM \quad (7)$$

其中 CAM 是来自补丁令牌的初始类激活图 $CAM \in R^{N \times N \times C}$ ，通过上式可获得通过增强补丁语义亲和力优化后的类激活图 $CAM'_{refined} \in R^{(N \times N \times C)}$ 。 $CAM'_{refined}$ 相比直接利用语义亲和力优化的 $CAM_{refined}$ ，拥有更全面的激活区域和更精细的对象边界，且有更高的鲁棒性。

3.4 模型训练与损失函数

如图1所示，使用多标签软边缘损失作为分类损失 L_{cls} ，使用交叉熵损失作为分割损失 L_{seg} 。参照基线方法ToCo[3]，本文使用了辅助分类损失 L_{m_cls} ，以及令牌对比损失 L_{ptc} 和 L_{ctc} 。此外，为了进一步提高性能，还按照先前的方法[8, 19–21]，采用了正则化损失 L_{reg} ，所以SS-EPA的损失最终定义如下：

$$\begin{aligned} L = & L_{cls} + \lambda_1 L_{seg} + \lambda_2 L_{m_cls} \\ & + \lambda_3 L_{ptc} + \lambda_4 L_{ctc} + \lambda_5 L_{reg} \end{aligned} \quad (8)$$

其中，超参数 $\lambda_i, i = 1, 2, \dots, 5$ 用于平衡不同损失的权重。

4 实验

4.1 实验设置

4.1.1 数据集

本文在Pascal VOC 2012[22]数据集上评估所提方法。Pascal VOC 2012包含20个前景类别和1个背景类别。它有三个子集：训练集（train）、验证集（val）和测试集（test），分别包含1464、1449 和1456张图片。按照先前方法[4, 5, 8, 12]的常用做法，本文进一步利用SBD数据集将VOC训练集图片数量扩充至10582。在训练过程中，本文严格只使用图像级分类标签用于监督模型训练。

4.1.2 评估指标

与先前方法一样，本文使用平均交并比（mean Intersection over Union，mIoU），作为CAM质量、伪标签质量以及语义分割模型性能的评估指标。本文方法在Pascal VOC测试集上的评估结果由官方在线评估服务器给

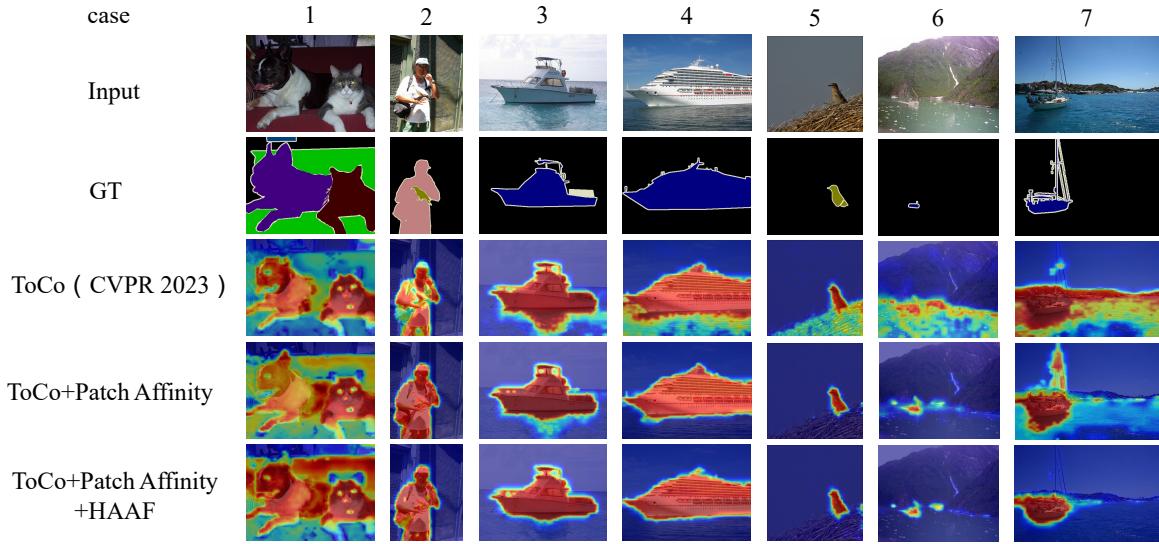


图 5 生成CAM可视化结果，从上到下依次为输入图片（input），真值标签（GT），ToCo生成CAM，SS-EPA生成CAM（不使用HAAF），SS-EPA生成CAM（使用HAAF）。红框部分为显著提升区域。

出。

4.1.3 实现细节

本文利用在ImageNet数据集[23]上预训练的ViT-B (vit_base_patch16_224) [7]作为backbone，它有12层Transformer编码层，12个注意力头，嵌入维度为768。卷积解码器使用LargeFOV [24]，它由两个膨胀系数为5的 3×3 卷积和一个 1×1 卷积预测层构成。

输入图片被随机裁剪为 448×448 的大小。模型共训练20000个迭代，batch-size设置为4，模型优化器采用AdamW[25]，学习率在前1500个迭代中逐渐提升到 6×10^{-5} ，并在后续根据多项式调度器衰减。公式8中的权重因子 $\lambda_i, i = 1, 2, \dots, 5$ 在前2000个迭代分别设置为 $(0, 1.0, 0.2, 0.5, 0)$ ，2000个迭代后分别设置为 $(0.1, 1.0, 0.2, 0.5, 0.05)$ 。

4.2 实验结果

4.2.1 CAM与伪标签

图5呈现了SS-EPA生成CAM的可视化结果，并与基线模型ToCo相比较。可以看出，SS-EPA在不使用HAAF的情况下，生成的CAM比ToCo更少，说明利用原始补丁语义亲和力优化CAM，可以显著减少初始CAM中的噪声，纠正错误激活的背景区域，使CAM更加精准和细

表 1 伪 标 签 生成 定 量 评 估 (MS: Multi Scale, CRF: dense CRF) (单位 mIoU%)

Method	Backbone	train	val
Multi-Stage WSSS Methods			
ViT-PCM[26]	ViT-B	67.7	66.0
MCTformer[5]	Deit-S	69.1	-
LPCAM[27]	Deit-S	-	70.8
SFC[28]	ResNet101	73.7	-
POLE[29]	ResNet50	74.2	-
Single-Stage WSSS Methods			
RRM[21]	ResNet38	-	65.4
1Stage[30]	ResNet38	66.9	65.3
SLRNet[31]	ResNet38	67.1	66.2
AFA[8]	MiT-B1	68.7	66.5
MCC[32]	Deit-B	75.1	72.2
ToCo[4]	ViT-B	74.5	72.2
ToCo+MS+CRF[4]	ViT-B	77.3	74.6
SS-EPA	ViT-B	77.1	74.2

化。在使用HAAF后，SS-EPA可能够发现一些未被初始CAM激活的前景区域（如图5中case 1与case 2），并生成错误更少的CAM（如图5中case 3至case 7），这说明了本文提出的HAAF可以进一步优化补丁语义亲和力，从而生成更优质的CAM。

表1呈现了利用CAM生成伪标签的定量评估结果，在VOC训练集和验证集上进行评估，并与一些先进的WSSS方法进行比较。结果表明，本文提出的SS-EPA比现有的单阶段WSSS方法更好，且达到了与一些多阶段WSSS方法相当的性能。与基线方法ToCo相比较，无论是否使用MS (Multi Scale)

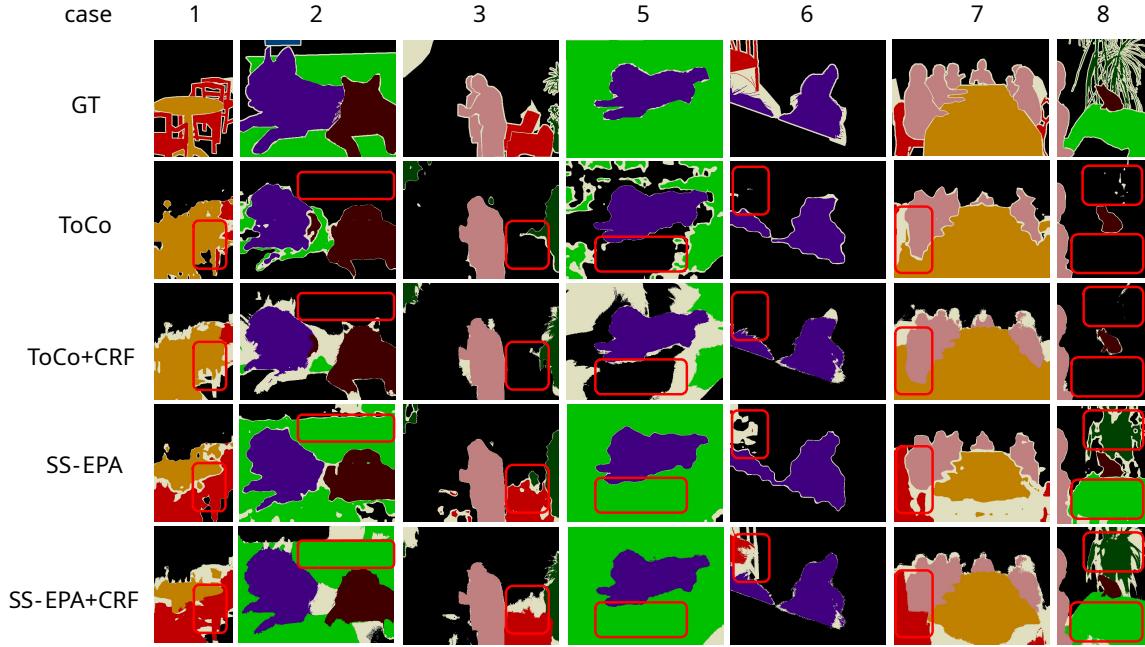


图 6 生成伪标签的可视化结果，从上到下依次为分割真值标签（GT），ToCo 分割结果（不加CRF），ToCo 分割结果（加CRF），SS-EPA 分割结果（不加CRF），SS-EPA 分割结果（加CRF）。红框部分为显著提升区域。

和CRF（DenseCRF），SS-EPA 的性能都要优于ToCo。

都高于ToCo（如图6中case 1 与 case 7），且能识别到一些ToCo 无法识别到的目标（如图6中case 2 至 case 6）。

4.2.2 分割结果

表2中呈现了在Pascal VOC 2012 数据集上的定量语义分割结果，比较了本文提出的SS-EPA 与其它的WSSS 方法在mIoU 分数上的表现。SS-EPA 用ImageNet 预训练的ViT-B（vit_base_patch16_224）作为 backbone，在验证集和测集上分别达到了72.4% 和73.3% 的mIoU 分数，比基线方法ToCo 提升了1.3% 和1.1% 的mIoU 分数。结果表明，SS-EPA 的性能优于现有的利用图像级标签的单阶段WSSS 方法。此外，SS-EPA 与许多多阶段WSSS 方法的性能相当，证明了本文所提方法的有效性。

图7展示了SS-EPA、ToCo和真实标签的分割结果。可视化结果表明，本文提出的SS-EPA成功分割了图像中的多个对象，并且与ToCo相比，SS-EPA分类的准确度更高（如图7中Val的case 1、3、4，Test的case 5、7、8），能正确发现一些ToCo中误分类为背景的前景目标（如图7中Val的case

图6是SS-EPA 生成伪标签的可视化结果，包括使用DenseCRF[40] 后处理前后的结果。可视化结果表明，无论是否使用CRF，SS-EPA 生成的伪标签在准确度上

表2 分割结果定量评估（单位mIoU%）

Method	Backbone	train	val
<i>Multi-Stage WSSS Methods</i>			
ReCAM[33]	ResNet101	68.5	68.4
ViT-PCM[26]	ResNet101	70.3	70.9
CLIMS[14]	ResNet101	70.4	70.0
AMN[34]	ResNet101	70.7	70.6
EDAM[35]	ResNet101	70.9	70.6
SFC[28]	ResNet101	71.2	72.5
POLE[29]	ResNet50	71.5	71.4
MCTformer[5]	Deit-S	71.9	71.6
L2G[36]	ResNet101	72.1	71.7
BECO[37]	ResNet101	72.1	71.8
RCA[38]	ResNet38	72.2	72.8
LPCM[27]	Deit-S	72.6	72.4
OCR[39]	ResNet38	72.7	72.0
<i>Single-Stage WSSS Methods</i>			
RRM[30]	ResNet38	62.6	62.9
1Stage[30]	ResNet38	62.7	64.3
AFA[8]	MiT-B1	66.0	66.3
SLRNet[31]	ResNet38	67.2	67.6
MCC[32]	Deit-B	70.3	71.2
ToCo[4]	ViT-B	71.1	72.2
SSEPA	ViT-B	72.4	73.2

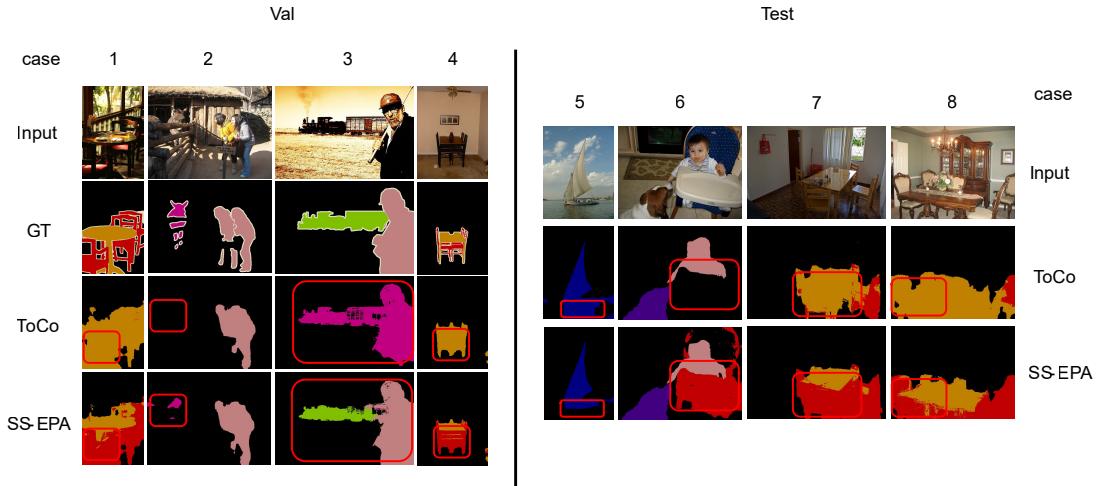


图 7 分割结果的可视化结果，左半部分为VOC验证集（Val），从上到下依次为输入图片（input），分割真值标签（GT），ToCo分割结果，SS-EPA分割结果。右半部分为VOC测试集（Test），从上到下依次为输入图片（input），ToCo分割结果，SS-EPA分割结果。红框部分为显著提升区域。

2, Test的case 6），且整体对象边界都更加完整和准确。

4.3 消融实验

4.3.1 补丁语义亲和力分析

表3呈现了关于伪标签和分割结果的消融实验定量评估。结果表明，使用不加HAAF增强的补丁语义亲和力后，SS-EPA可以生成更加优质的伪标签并提高分割性能。其中伪标签提升了2.8%，分割性能在验证集和测试集上分别提升了0.6%和1.1%，分割的准确率更高。从图5可看出尽管在未使用HAAF的情况下存在噪声与错误，补丁语义亲和力仍有效优化了初始CAM。

表3 分割结果定量评估（单位mIoU%）

Method	Pseudo label(train)	Seg(val)	Seg(test)
ToCo	77.3	71.1	72.21
SS-EPA(w/o Patch Affinity)	76.2	71.3	71.62
SS-EPA(with Patch Affinity)	79.0	71.9	72.73
SS-EPA(with Patch Affinity HAAF)	79.5	72.4	73.34

4.3.2 HAAF分析

如3.3节中所说，补丁语义亲和力存在噪声与错误，直接使用补丁语义亲和力并不合适。本文提出的HAAF模块显著减少了语义亲和力中的噪声和错误，并减少计算资源占用。从表3中可以看出，HAAF进一步提升了

伪标签和分割结果的mIoU分数，其中伪标签提升了0.5%，分割性能在验证集和测试集上分别提升了0.5%和0.6%。

表4 SS-EPA计算资源占用实验结果评估（单位GB）

Method	Backbone	Batchsize 1	Batchsize 2
SS-EPA(w/o Patch Affinity)	ViT-B	6.6	10.4
SS-EPA(with Patch Affinity)	ViT-B	13.2	23.6
SS-EPA(with Patch Affinity HAAF)	ViT-B	8.3	12.6

表4展示了SS-EPA的计算资源占用评估，分别评估了batchsize 1 和batchsize 2 的实验结果。结果表明，在不使用HAAF的情况下，整个SS-EPA需要占据较高的计算资源，batchsize 为1时需要13.2GB显存，batchsize 为2时则需要23.6GB。而HAAF可以将计算资源占用降低到8.3GB和12.6GB，显著减少了对计算资源的需求，提升了计算效率。

4.3.3 Backbone分析

表5展示了不同backbone下的SS-EPA和基线方法ToCo的实验结果评估。结果表明，SS-EPA在使用不同backbone的情况下比ToCo更好，在VOC验证集和测试集上的分割性能都更加优秀。其中表现最好的backbone是vit-base-patch16-224。与使用更高分辨率的vit-base-patch16-384相比，低分辨率的vit-base-patch16-224具有更好的

表 5 SS-EPA 不同 Backbone 实验结果评估 (单位 mIoU%)

Method	Backbone	Depth	Img_size	Seg(val)	Seg(test)
ToCo	vit-small-patch16-224	8	224 × 224	55.0	52.7
SS-EPA	vit-small-patch16-224	8	224 × 224	57.6	50.2
ToCo	vit-base-patch16-384	12	384 × 384	71.1	71.8
SS-EPA	vit-base-patch16-384	12	384 × 384	71.7	71.9
ToCo	vit-base-patch16-224	12	224 × 224	71.1	72.2
SS-EPA	vit-base-patch16-224	12	224 × 224	72.4	73.3

泛化能力，不太容易过拟合到训练数据中的特定细节。而 vit-small-patch16-224 只有 8 层 Transformer 块，参数量和计算量都相对较少，导致其在捕捉图像中的复杂特征和细节时能力有限。

5 结论

本文工作主要有以下两点：第一是提出了一种名为 SS-EPA 的单阶段 WSSS 方法，集成了端到端式多头自注意力 CAM 优化方法；第二是提出一种头平均注意力融合增强模块 (HAAF)，来进一步优化语义亲和力中的噪声和错误。具体而言，本文首先提出了 SS-EPA 这个单阶段 WSSS 方法，将端到端式多头自注意力 CAM 优化方法，在不影响单阶段方法的完整性和一致性的前提下，集成到单阶段 WSSS 框架中。鉴于语义亲和力信息包含噪声与错误，以及注意力图较为庞大，本文提出了头平均注意力融合增强模块 (Head Average Attention Fusion, HAAF)。通过对注意力的不同头的权重做平均，HAAF 可去除冗余信息并提高模型鲁棒性。利用多层感知机的交互能力，HAAF 可以充分考虑来自不同层注意力的重要性，对包含语义亲和力的自注意力完成简化和增强。实验结果表明，SS-EPA 可以显著优于其它单阶段 WSSS 方法，并达到与一些多阶段 WSSS 方法相当的性能。SS-EPA 端到端式的设计，减少了中间步骤的计算和存储要求，对计算资源受限的环境更友好。本文方法虽然取得了更优秀的分割性能，但在计算开销和局部特征学习上仍有提升空间。后续研究将骨

干网络 ViT 更换成更加强大的 Transformer 变体如 EfficientFormer [41] 或 Swin Transformer [42]，通过引入高效注意力机制来进一步减少参数量和计算量，或通过滑动窗口的局部注意力来更好地捕捉局部信息。

参 考 文 献

- [1] J. Dai, K. He, and J. Sun, “Boxsup: Exploiting bounding boxes to supervise convolutional networks for semantic segmentation,” in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1635–1643.
- [2] B. Zhang, J. Xiao, J. Jiao, Y. Wei, and Y. Zhao, “Affinity attention graph neural network for weakly supervised semantic segmentation,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 11, pp. 8082–8096, 2021.
- [3] D. Lin, J. Dai, J. Jia, K. He, and J. Sun, “Scribblesup: Scribble-supervised convolutional networks for semantic segmentation,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 3159–3167.
- [4] L. Ru, H. Zheng, Y. Zhan, and B. Du, “Token contrast for weakly-supervised semantic segmentation,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 3093–3102.

- [5] L. Xu, W. Ouyang, M. Bennamoun, F. Boussaid, and D. Xu, “Multi-class token transformer for weakly supervised semantic segmentation,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 4310–4319.
- [6] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, “Learning deep features for discriminative localization,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2921–2929.
- [7] A. Dosovitskiy, “An image is worth 16x16 words: Transformers for image recognition at scale,” *arXiv preprint arXiv:2010.11929*, 2020.
- [8] L. Ru, Y. Zhan, B. Yu, and B. Du, “Learning affinity from attention: End-to-end weakly-supervised semantic segmentation with transformers,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 16 846–16 855.
- [9] J. Ahn and S. Kwak, “Learning pixel-level semantic affinity with image-level supervision for weakly supervised semantic segmentation,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 4981–4990.
- [10] Y. Wei, J. Feng, X. Liang, M.-M. Cheng, Y. Zhao, and S. Yan, “Object region mining with adversarial erasing: A simple classification to semantic segmentation approach,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1568–1576.
- [11] Y.-T. Chang, Q. Wang, W.-C. Hung, R. Piramuthu, Y.-H. Tsai, and M.-H. Yang, “Weakly-supervised semantic segmentation via sub-category exploration,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 8991–9000.
- [12] Y. Wang, J. Zhang, M. Kan, S. Shan, and X. Chen, “Self-supervised equivariant attention mechanism for weakly supervised semantic segmentation,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 12 275–12 284.
- [13] X. Li, T. Zhou, J. Li, Y. Zhou, and Z. Zhang, “Group-wise semantic mining for weakly supervised semantic segmentation,” in *Proceedings of the AAAI conference on artificial intelligence*, vol. 35, no. 3, 2021, pp. 1984–1992.
- [14] J. Xie, X. Hou, K. Ye, and L. Shen, “Clims: Cross language image matching for weakly supervised semantic segmentation,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 4483–4492.
- [15] Y. Lin, M. Chen, W. Wang, B. Wu, K. Li, B. Lin, H. Liu, and X. He, “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*, 2023, pp. 15 305–15 314.
- [16] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, “Learning transferable visual models from natural language supervision,” in *International conference on machine learning*. PMLR, 2021, pp. 8748–8763.

- [17] W. Gao, F. Wan, X. Pan, Z. Peng, Q. Tian, Z. Han, B. Zhou, and Q. Ye, “Ts-cam: Token semantic coupled attention map for weakly supervised object localization,” in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 2886–2895.
- [18] A. Vaswani, “Attention is all you need,” *Advances in Neural Information Processing Systems*, 2017.
- [19] M. Tang, F. Perazzi, A. Djelouah, I. Ben Ayed, C. Schroers, and Y. Boykov, “On regularized losses for weakly-supervised cnn segmentation,” in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 507–522.
- [20] B. Zhang, J. Xiao, and Y. Zhao, “Dynamic feature regularized loss for weakly supervised semantic segmentation,” *arXiv preprint arXiv:2108.01296*, 2021.
- [21] B. Zhang, J. Xiao, Y. Wei, M. Sun, and K. Huang, “Reliability does matter: An end-to-end weakly supervised semantic segmentation approach,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 07, 2020, pp. 12 765–12 772.
- [22] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, “The pascal visual object classes (voc) challenge,” *International journal of computer vision*, vol. 88, pp. 303–338, 2010.
- [23] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, “Imagenet: A large-scale hierarchical image database,” in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.
- [24] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, “Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 4, pp. 834–848, 2017.
- [25] I. Loshchilov, “Decoupled weight decay regularization,” *arXiv preprint arXiv:1711.05101*, 2017.
- [26] S. Rossetti, D. Zappia, M. Sanzari, M. Schaerf, and F. Pirri, “Max pooling with vision transformers reconciles class and shape in weakly supervised semantic segmentation,” in *European conference on computer vision*. Springer, 2022, pp. 446–463.
- [27] Z. Chen and Q. Sun, “Extracting class activation maps from non-discriminative features as well,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 3135–3144.
- [28] X. Zhao, F. Tang, X. Wang, and J. Xiao, “Sfc: Shared feature calibration in weakly supervised semantic segmentation,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 7, 2024, pp. 7525–7533.
- [29] B. Murugesan, R. Hussain, R. Bhattacharya, I. Ben Ayed, and J. Dolz, “Prompting classes: exploring the power of prompt class learning in weakly supervised semantic segmentation,” in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2024, pp. 291–302.
- [30] N. Araslanov and S. Roth, “Single-stage semantic segmentation from image labels,” in *Proceedings of the IEEE/CVF*

- conference on computer vision and pattern recognition*, 2020, pp. 4253–4262.
- [31] J. Pan, P. Zhu, K. Zhang, B. Cao, Y. Wang, D. Zhang, J. Han, and Q. Hu, “Learning self-supervised low-rank network for single-stage weakly and semi-supervised semantic segmentation,” *International Journal of Computer Vision*, vol. 130, no. 5, pp. 1181–1195, 2022.
- [32] F. Wu, J. He, Y. Yin, Y. Hao, G. Huang, and L. Cheng, “Masked collaborative contrast for weakly supervised semantic segmentation,” in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2024, pp. 862–871.
- [33] Z. Chen, T. Wang, X. Wu, X.-S. Hua, H. Zhang, and Q. Sun, “Class reactivation maps for weakly-supervised semantic segmentation,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 969–978.
- [34] M. Lee, D. Kim, and H. Shim, “Threshold matters in wsss: Manipulating the activation for the robust and accurate segmentation model against thresholds,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 4330–4339.
- [35] T. Wu, J. Huang, G. Gao, X. Wei, X. Wei, X. Luo, and C. H. Liu, “Embedded discriminative attention mechanism for weakly supervised semantic segmentation,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 16 765–16 774.
- [36] P.-T. Jiang, Y. Yang, Q. Hou, and Y. Wei, “L2g: A simple local-to-global knowledge transfer framework for weakly supervised semantic segmentation,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 16 886–16 896.
- [37] S. Rong, B. Tu, Z. Wang, and J. Li, “Boundary-enhanced co-training for weakly supervised semantic segmentation,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 19 574–19 584.
- [38] T. Zhou, M. Zhang, F. Zhao, and J. Li, “Regional semantic contrast and aggregation for weakly supervised semantic segmentation,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 4299–4309.
- [39] Z. Cheng, P. Qiao, K. Li, S. Li, P. Wei, X. Ji, L. Yuan, C. Liu, and J. Chen, “Out-of-candidate rectification for weakly supervised semantic segmentation,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 23 673–23 684.
- [40] L.-C. Chen, “Semantic image segmentation with deep convolutional nets and fully connected crfs,” *arXiv preprint arXiv:1412.7062*, 2014.
- [41] Y. Li, G. Yuan, Y. Wen, J. Hu, G. Evangelidis, S. Tulyakov, Y. Wang, and J. Ren, “Efficientformer: Vision transformers at mobilenet speed,” *Advances in Neural Information Processing Systems*, vol. 35, pp. 12 934–12 949, 2022.
- [42] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, “Swin transformer: Hierarchical vision transformer using shifted windows,” in

Proceedings of the IEEE/CVF international conference on computer vision, 2021, pp. 10 012–10 022.