基于自训练的开放词汇语义分割图像像素重对齐方法

杨安逸[1,2]　刘齐波[1]　樊亚文[1]　周全[3]
[1]江苏省智能信息处理与通信技术重点实验室，南京邮电大学，中国
[2]广西高校智慧行业软件重点实验室，梧州学院，中国
[3]南通先进海洋研究院，东南大学，中国

摘要

开放词汇语义分割（OVSS）旨在为图像中任意类别（包括未见类别）分配像素级标签，但其面临着一个挑战：如何将视觉-语言模型中的图像级语义迁移到细粒度密集预测任务中。本文介绍了 STSeg，一个将像素级语义对齐与自适应自训练相结合的新颖框架。该框架设计了一个双视觉嵌入聚合模块，通过交叉注意力机制，将来自 CLIP 模型的全局语义特征，与冻结的 SAM 模型所提取的空间精度特征进行融合，生成兼具语义信息与空间细节的高质量像素表征。在此基础上，框架中的自训练模块通过迭代优化，对 SAM 引导生成的伪标签进行细化，使模型能够在训练过程中自主更新知识体系、优化分割性能。此机制不仅使 STSeg 对新类别具备良好的泛化能力，还实现了推理阶段无需额外监督的动态自适应。在多个 OVSS 基准数据集上的广泛评估验证了 STSeg 的有效性。该框架展现出极具竞争力的分割精度与强泛化性，充分证明了其在实际场景中的应用潜力。

关键词：开放词汇语义分割，视觉-语言模型，自训练。

1. 引言

语义分割的目标是为图像中的每个像素分配对应的语义标签。传统方法[1,2]通常预设类别集合为封闭集，这使得模型无法识别训练中未见过的新类别。为解决这一局限性，近年来的相关研究聚焦于开放词汇语义分割[3,4]，其目标是使模型能够对任意语义类别的像素完成分割，包括那些在训练集中不存在的类别。

随着大规模视觉-语言模型（如 CLIP[5]和 ALIGN[6]）的出现，学术界开始广泛探索这类模型在开放词汇语义分割任务中的应用。早期方法通常采用两阶段框架[7,8]：首先由一个独立的分割模型将图像划分为不同的区域[9]，再利用冻结的 CLIP 模型对这些区域进行分类，如图 1（a）所示。然而，这种策略不仅会产生较大的计算开销，还容易增加类别误判的概率。近期的单阶段方法则尝试在一个统一的视觉-语言框架内微调 CLIP，直接实现像素级分割，无需额外搭建分割网络，如图 1（b）所示。例如，SAN[10]提出了用于掩码注意力机制的侧边适配器，MaskCLIP[11]移除了全局池化以实现令牌级分类，CATSeg[12]利用图像和文本特征之间的像素级余弦相似度进行分割，SED[13]则采用分层编码器-解码器来简化聚合过程[14]。传统的微调策略在训练中侧重于已见类别的学习，忽视了对 CLIP 模型本身具备的、针对未见类别的图像-文本对齐能力的保留[15]。同时，大多数方法仅实现了图像级别的语义对齐，未能提供语义分割任务所需的像素级语义对应关系。

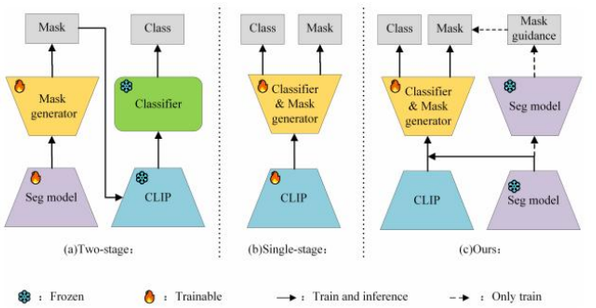

图 1.(a)两阶段模型，由额外的掩码生成网络生成候选掩码；(b)单阶段模型，在统一模型内完成分割与分类；(c)所提的 STSeg 架构，融合预训练 SAM 模块和自训练分支，更高效地建模图像与文本的关联关系。

针对上述问题，本文提出全新的 STSeg 框架，该框架由双视觉嵌入聚合模块和自训练模块组成。在双视觉嵌入聚合模块中，我们引入冻结的 SAM[16]编码器。该模型在提取像素级特征方面的表现优于 CLIP，通过交叉注意力机制融合了来自 CLIP 的全局语义和 SAM 的像素级精细特征，既强化了模型的细粒度特征提取能力，也为图像-文本成本图的生成提供了支撑。此外，我们提出了一种自适应自训练机制，能够从无标签数据中动态地生成并利用伪标签，使模型的泛化能力持续提升。具体来说，框架中引入了一个冻结的 SAM 分支，将模型的初始分割结果作为提示输入 SAM 解码器，由其生成细化后的分割掩码。借助 SAM

通讯作者：周全，quan.zhou@njupt.edu.cn。本研究部分由 NSFC（编号：62476139）和 NSFJS（编号：BK2024023）资助。

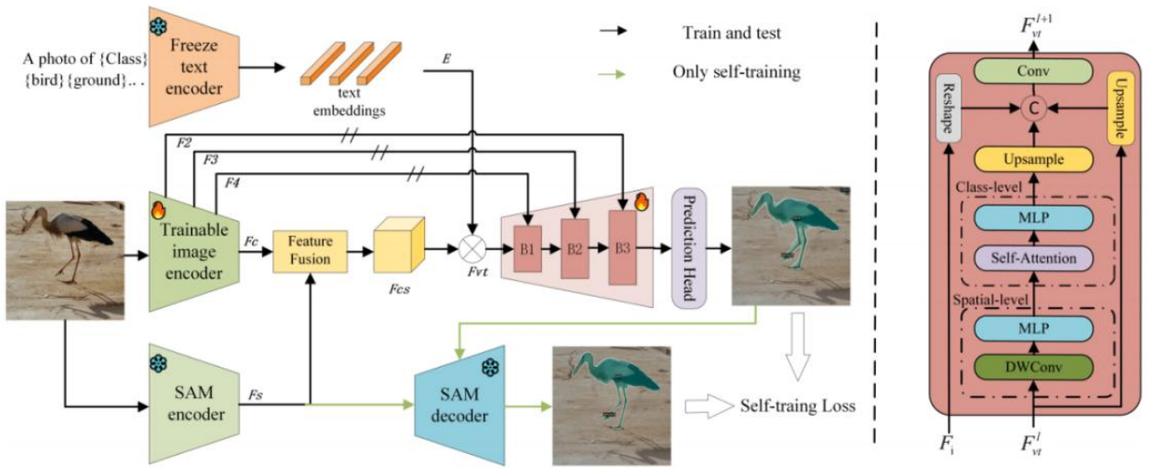

图 2.STSeg 架构整体示意图：首先通过双视觉嵌入聚合模块融合图像特征，与文本特征计算余弦相似度得到初始成本体积；随后由成本感知解码器完成分割；同时引入冻结的 SAM 分支，实现模型的自训练。

提供的监督信息，既能缓解模型对训练类别的过拟合问题，也能进一步提升模型的泛化能力。如图 1（c）所示，我们的方法整合了单阶段和两阶段范式的优势，显著提高了模型对已见和未见类别的分割与识别性能。总之，本文的主要贡献如下：（1）针对 CLIP 在像素级语义表征方面的局限性，我们引入 SAM 模型，并设计了一种可学习的特征融合策略，实现了两种模型视觉特征的有效融合。（2）提出了一种新颖的自训练框架，将开放词汇分割的训练与推理过程统一，通过用模型自主生成的伪标签进行迭代细化，逐步提升模型性能。（3）在多个开放词汇分割基准数据集上开展大量实验，验证了所提方法的有效性。

2. 方法

本节提出了一个适用于开放词汇语义分割的全新框架 STSeg。该框架包含三个核心模块：双视觉嵌入聚合模块（DEA）、成本感知解码器（CAD）和自训练（ST）模块。具体来说，双视觉嵌入聚合模块对两个不同的图像编码器提取的多尺度视觉特征进行聚合，生成更适配语义分割任务的细粒度特征表示。接着，通过计算视觉特征和文本特征之间的相似度，得到低分辨率的成本图，将其作为粗分割预测结果输入至成本感知解码器。此外，本文引入了自训练模块，利用模型自身生成的伪标签对模型进行迭代细化，显著提升模型对未见类别的泛化能力。

2.1 双视觉嵌入聚合模块（DEA）

对于给定输入图像，通过 ConvNeXt-based 的编码器提取图像级特征F_c，通过 SAM 编码器提取像素级特征F_s。先将F_s进行对齐处理，使其空间分辨率和嵌入维度均与F_c保持一致，再通过一系列注意力操作完成两类特征融合。在交叉注意力机制中，查询向量由F_c投影得到，键向量和值向量由F_s投影得到：

$$Q = F_C{*}W_Q,\ K = F_S{*}W_K,\ V = F_S{*}W_V, \quad (1)$$

交叉注意力的输出结果结合残差连接和归一化操作后，可表示为：

$$F' = LN\left(Q + \sigma\left(\frac{QK^{\top}}{\sqrt{d}}\right)V\right), \quad (2)$$

接着，对融合特征F'施加自注意力层，其中查询、键、值向量均由F'生成，得到：

$$F'' = LN\left(F' + \sigma\left(\frac{F'W'_Q(F'W'_K)^{\top}}{\sqrt{d}}\right)F'W'_V\right), \quad (3)$$

最后，将特征输入带有残差连接和层归一化的前馈网络，得到最终的融合特征表征：

$$F_{cv} = LN(F'' + FFN(F'')). \quad (4)$$

其中，LN 表示层归一化，$\sigma(\cdot)$表示 Softmax 激活函数。

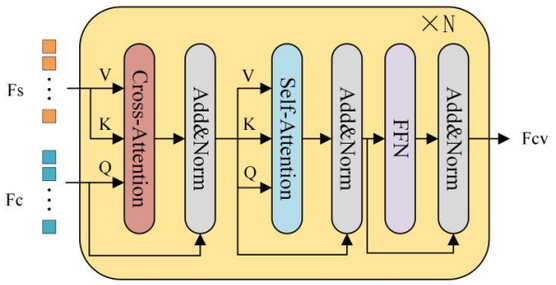

图 3.双视觉嵌入聚合模块工作流程：首先让 SAM 特征与 CLIP 特征进行交叉注意力计算，实现两种表征的有效信息交互；随后对融合特征施加自注意力，进一步提升其全局建模能力和特征一致性。

2.2 成本感知解码器（CAD）

受 SED[17]启发，本文采用成本体积表示的方式来实现视觉特征与文本提示的对齐。对于给定类别集合$\{T_1, ..., T_N\}$，我们构建多组文本模板（例如，“一张T_n的照片”），并将其编码为文本嵌入特征$E \in \mathbb{R}^{N\times P\times D_t}$。对于图像中每个像素点$(x, y)$，计算视觉特征$F_{cv}(x, y)$与文本嵌入特征 $E(n, p)$的余弦相似度，得到成本体积：

$$F_{vt}(x,y,n,p) = \frac{F_{cv}(x,y)\cdot E(n,p)}{\|F_{cv}(x,y)\|\,\|E(n,p)\|}, \quad (5)$$

其中，x 和 y 为像素的空间位置坐标，n 为类别索引，p 为文本模板索引。由此得到的初始成本体积F_{vt}的维度为$H_c \times W_c \times N \times P$。

接着，成本体积F_{vt}由一个成本感知解码器进行细化，该解码器借鉴 SED 的设计，包含三个核心模块，如图 2 右侧所示：用于空间聚合的深度卷积、用于类别聚合的线性自注意力，以及用于细节增强的多尺度特征融合。这一设计将语义对齐信息与局部视觉特征深度融合，为像素级精准分割预测提供支撑。

2.3 自训练模块（ST）

开放词汇语义分割的一个关键挑战在于未见类别缺乏真实的像素级标注，无法进行直接监督训练。为了解决这一问题，我们设计了一种自训练机制，融合主分割分支（成本感知解码器，CAD）和基于 SAM 的辅助分支的互补优势，构建一个闭环学习系统。如图 1 中蓝色路径所示，辅助分支基于冻结的 SAM 解码器构建，在保留 SAM 强大的空间先验能力的同时不引入额外的可训练参数。该分支同时接收由 SAM 编码器提取的图像特征和主分支输出的初始分割掩码。这些初始分割掩码作为提示，引导 SAM 对分割边界进行细化。为确保提示的可靠性，我们采用了一种门控策略，仅将置信度得分高于 0.9 的类别选为有效提示，再输入至 SAM 解码器进行细化。通过这种方式，辅助分支可以将初始预测结果细化为空间精度更高的伪标签，并将其反馈至主分支作为模型训练的替代监督信号。在每轮训练完成后，初始分割掩码和伪标签均会同步进行迭代更新，使监督信号随模型的学习过程动态调整，持续优化。

本框架的训练过程分为两个阶段。第一阶段，模型在具有完整像素级标注的 COCO-Stuff 数据集上进行训练，将标注类别定义为已见类别。此阶段的主要目标是让模型掌握对已见类别的准确分割，为视觉-语言对齐建立坚实的基础。第二阶段，在包含未见类别的目标数据集（如 ADE20K、PASCALContext）上进行自训练，该阶段无真实标签，将辅助分支生成的细化伪标签作为监督信息，从而增强模型对未见类别的识别能力。训练损失函数定义为：

$$\mathcal{L}_{s1} = -\frac{1}{N}\sum_{c=1}^{N}\omega_c y_c^{gt}\log(y_c^p), \begin{cases}\omega_c = \alpha, c \in C\\ \omega_c = 1-\alpha, c \in C'\end{cases} \quad (6)$$

其中，C 和C'分别表示已见类别和未见类别集合。在实际实验中，设置$\alpha = 0.3$，为已见类别分配较小的损失权重，为未见类别分配较大的损失权重，从而强化模型对未见类别的学习。

3. 实验

3.1 数据集与评估指标

数据集：遵循现有开放词汇语义分割方法的通用实验设置，选取大规模 COCO-Stuff 数据集[18]作为训练集，该数据集包含约 118,000 张密集标注的图像，覆盖 171 个不同的语义类别。为验证所提 STSeg 框架的有效性，并与当前领域的最新方法进行对比，我们在 ADE20K、PASCAL VOC、PASCAL-Context 等多个被广泛使用的语义分割基准数据集上开展实验。特别地，A-847 和 A-150 数据集源自 ADE20K 数据集，PC-459 和 PC-59 源自 PASCAL-Context 数据集。

评价指标：遵循 OVSS 的标准协议，本文选取平均交并比（mean Intersectionover Union）作为评价指标，该指标计算所有类别交并比的平均值，能全面反映模型的分割精度。

3.2 实现细节

本文以预训练的视觉-语言模型 CLIP 为骨干网络，采

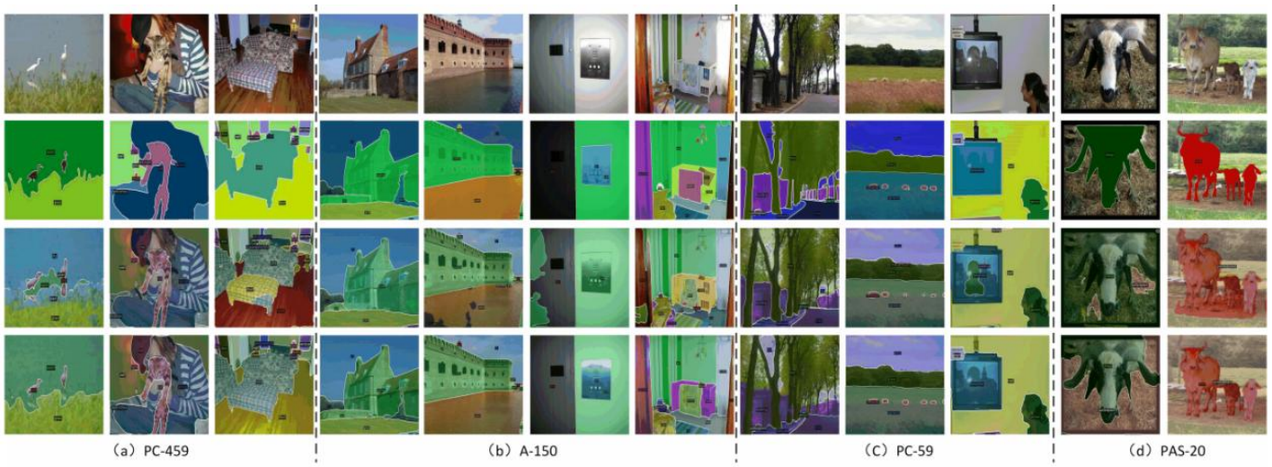

图 4.各基准数据集的分割结果可视化：第一列为原始图像，第二列为真实标注，第三列为 SED 方法的分割结果，第四列为所提方法的分割结果。

表 1.不同开放词汇分割数据集上，所提方法与最先进方法的平均交并比（%）对比。

Method	A-847	PC-459	A-150	PC-59	PAS-20	Avg.
SPNet [3]	-	-	-	24.3	18.3	-
ZS3Net [4]	-	-	-	19.4	38.3	-
LSeg [22]	-	-	-	-	47.4	-
LSeg+ [23]	2.5	5.2	13.0	36.0	-	-
Han et al. [24]	3.5	7.1	18.8	45.2	83.2	31.6
GroupViT [25]	4.3	4.9	10.6	29.5	50.7	20.0
OpenSeg [26]	4.4	7.9	17.5	40.1	-	-
DeOP [27]	7.1	9.4	22.9	48.8	91.7	36.0
OVSeg [28]	7.1	11.0	24.8	53.3	92.6	37.8
SAN [15]	10.1	12.6	27.5	53.8	94.0	39.6
CAT-Seg [12]	12.0	19.0	31.8	57.5	94.6	43.0
SCAN [29]	10.8	13.2	30.8	58.4	**97.0**	42.0
SED [13]	11.4	18.6	31.6	57.3	94.4	42.7
CEL [30]	9.7	12.6	29.9	55.6	91.8	39.9
EBSeg [31]	11.1	17.3	30.0	56.7	94.6	41.9
BBN [32]	**12.8**	19.1	31.9	**59.0**	94.8	43.5
Ours	12.7	**19.7**(+0.6)	**33.2**(+1.3)	58.9	95.7	**43.7**(+0.2)

用分层 ConvNeXt-B 网络作为图像编码器；将文本嵌入特征的维度设为 640，类别模板数量设为 80，训练过程中保持文本编码器冻结；实验基于 4 块 NVIDIA RTX 3090 显卡进行训练，优化器选用 AdamW，初始学习率设为 2×10^{-4}，权重衰减系数设为 1×10^{-4}，训练总迭代次数为 80000 次；训练和推理阶段，均将输入图像裁剪至 768×768 像素。

在模型推理阶段，我们直接舍弃训练阶段的辅助分支，最终的分割输出结果仅由主分支产生，从而保证了计算效率。因此，辅助分支在训练期间引入的额外计算量不会影响模型部署，使得该框架兼具轻量化特性和实际应用价值。尽管移除了辅助分支，但主分支已通过伪标签的迭代细化完成了性能提升，因此模型仍然保留了自训练带来的性能增益。这种训练-推理范式既在训练阶段利用伪标签提升了模型对开放词汇类别的识别能力，又在测试阶段实现了高效、鲁棒的性能表现。

3.3 实验结果

在本项工作中，我们将 STSeg 与几种最先进的开放词汇语义分割（OVSS）方法进行了比较。在五个基准数据集上的结果总结在表 1 中。实验结果表明，我们的方法在 PC-459 数据集和 A-150 数据集上取得了最佳性能。具体来说，在 PC-459 数据集上，STSeg 的性能较 CAT-Seg、SED 和 EBSeg 分别提升 5.8%、11.4%和 14.4%。在 A-150 数据集上，较 CAT-Seg、SED 和 SCAN 分别提升 4.4%、5.06%和 7.79%。在 A-847 数据集和 PC-59 数据集上，STSeg 的性能与近期提出的 BBN 方法相当，二者仅有 0.1 的差距。尽管 STSeg 在 PAS20 上排名第二，但与性能最佳的两阶段方法 SCAN 相比，STSeg 在推理速度上展现出了显著优势，凸显了其在分割精度和效率之间达到了良好的平衡。此外，从所有数据集上的平均性能来看，STSeg 较当前最优方法提升 0.2 个百分点，进一步证实了其综合优越性。

图 4 展示了不同方法在各数据集上分割结果的全面定性对比。为了进一步验证框架的鲁棒性，我们在多个具有挑战性的基准数据集上展示了可视化结果，包括 PC-459、A-150、PC-59 和 PAS-20。在这些数据集上，STSeg 的优势显著。例如，在第九列（羊）和第七列（椅子和毯子）的分割结果中，STSeg 实现了比 SED 等竞争方法更清晰、更准确的分割边界，成功捕捉到了先前方法常常遗漏的精细轮廓。在语义一致性方面，STSeg 同样表现突出。典型案例为第十列的分割结果，电视屏幕上显示的人像被 STSeg 正确地归为属于“电视”类别而非“人”。这表明 STSeg 不仅在像素级定位方面表现优异，还能在语义模糊的复杂场景中实现更可靠的类别分配。此外，自训练模块利用 SAM 增强的伪标签进一步细化了边界预测，并显著提高了模型，对例如“电视”、“标志”等未见类别的泛化能力。上述定量和定性的结果有力地证实了 STSeg 在解

决复杂实际场景下开放词汇语义分割任务的有效性。

3.4 消融实验

为验证 STSeg 中各核心模块的有效性，本文设计了针对性的消融实验，实验结果如表 2 所示。基线模型仅由 CLIP 编码器和成本感知解码器构成，未融入 SAM 相关组件或自训练模块。在此基线模型基础上，我们首先添加双视觉嵌入聚合模块，该模块引入了空间增强的语义特征，使模型在 A-847、A-150 和 PC-59 数据集上的性能分别提升了 5.26%、1.89%和 1.22%，验证了该模块在特征融合与细粒度特征提取上的有效性。单独为基线模型引入自训练模块时，基线模型在 A-847、PC-459 和 A-150 数据集上的性能分别提升 7.02%、4.30%和 2.85%，凸显了自适应伪标签细化的有效性。最后，同时引入两个提出的模块，构成了我们的最终模型。

表 2.STSeg 各模块的性能影响。展示了将不同模块融入基线模型后的实验结果。

DEA	ST	A-847	PC-459	A-150	PC-59	PAS-20
		11.4	18.6	31.6	57.3	94.4
✓		12.0	18.9	32.2	58.0	94.9
	✓	12.2	19.4	32.5	58.2	94.7
✓	✓	12.7	19.7	33.2	58.9	95.7

4. 结论

本文针对开放词汇语义分割任务，提出了一种全新的自训练框架 STSeg。该框架融合 SAM 捕获的丰富空间信息与 CLIP 的强大视觉语义表示能力，有效地整合了先前相互独立的分割和分类模型的优势。这种联合设计使模型能够同时利用细粒度的图像结构特征和高层的语义理解能力完成分割任务。此外，本文提出了一种创新的自训练策略，通过模型自主生成伪标签并进行迭代细化的方式，显著提升了模型对未见过的新类别的泛化能力。STSeg 实现了空间精度与语义丰富性的统一，不仅为开放词汇语义分割研究提供了一种全新的方法，更从全新的视角为该领域在复杂、动态的实际场景中的后续研究提供了思路。

5. 参考文献

[1] J. Jain, J. Li, M. T. Chiu, A. Hassani, N. Orlov, and H. Shi, “Oneformer: One transformer to rule universal image segmentation,” in CVPR, 2023, pp. 2989-2998.

[2] X. Pan, T. Ye, Z. Xia, S. Song, and G. Huang, “Slide-transformer: Hierarchical vision transformer with local self- attention,” in CVPR, 2023, pp. 2082-2091.

[3] Y. Xian, S. Choudhury, Y. He, B. Schiele, and Z. Akata, “Semantic projection network for zero-and few-label semantic segmentation,” in CVPR, 2019, pp. 8256-8265.

[4] M. Bucher, T.-H. Vu, M. Cord, and P. Perez, “Zero-shot semantic segmentation,” in NeurIPS, 2019, pp. 468-479.

[5] 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 ICML, 2021, pp. 8748-8763.

[6] C. Jia, Y. Yang, Y. Xia, Y.- T. Chen, Z. Parekh, H. Pham, Q. Le, Y.- H. Sung, Z. Li, and T. Duerig, “Scaling up visual and vision-language representation learning with noisy text supervision,” in ICML, 2021, pp. 4904-4916.

[7] Z. Ding, J. Wang, and Z. Tu, “Open vocabulary panoptic segmentation with maskclip,” arXiv preprint arXiv:2208.08984, 2022.

[8] F. Liang, B. Wu, X. Dai, K. Li, Y. Zhao, H. Zhang, P. Zhang, P. Vajda, and D. Marculescu, “Open-vocabulary semantic segmentation with mask-adapted clip,” in CVPR, 2023, pp. 7061-7070.

[9] J. Ding, N. Xue, G. Xia, and D. Dai, “Decoupling zero-shot semantic segmentation,” in CVPR, 2021, pp. 11573-11582.

[10] M. Xu, Z. Zhang, F. Wei, H. Han, and X. Bai, “Side adapter network for open-vocabulary semantic segmentation,” in CVPR, 2023, pp. 2945-2954.

[11] C. Zhou, C. C. Loy, and B. Dai, “Extract free dense labels from clip,” in CVPR, 2022, pp. 696-712.

[12] S. Cho, H. Shin, S. Hong, A. Arnab, P. H. Seo, and S. Kim, “Catseg: Cost aggregation for open-vocabulary semantic segmentation,” in CVPR, 2024, pp. 4113-4123.

[13] B. Xie, J. Cao, J. Xie, F. S. Khan, and Y. Pang, “Sed: A simple encoder-decoder for open vocabulary semantic segmentation,” in CVPR, 2024, pp. 3426-3436.

[14] S. Cho, S. Hong, and S. Kim, “Cats++: Boosting cost aggregation with convolutions and transformers,” IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022.

[15] M. Xu, Z. Zhang, F. Wei, H. Hu, and X. Bai, “Side adapter network for open-vocabulary semantic segmentation,” in CVPR, 2023, pp. 2945-2954.

[16] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L.

Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.- Y. Lo, et al., “Segment anything,” arXiv preprint arXiv:2304.02643, 2023.

[17] X. Gu, T.- Y. Lin, W. Kuo, and Y. Cui, “Open-vocabulary object detection via vision and language knowledge distillation,” in ICLR, 2022, pp. 1-20.

[18] H. Caesar, J. Uijlings, and V. Ferrari, “Coco- stuff: Thing and stuff classes in context,” in CVPR, 2018, pp. 1209-1218.

[19] B. Zhou, H. Zhao, X. Puig, T. Xiao, S. Fidler, A. Barriuso, and A. Torralba, “Semantic understanding of scenes through the ade20k dataset,” International Journal of Computer Vision, vol. 127, pp. 302-321, 2019.

[20] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, “The pascal visual object classes (voc) challenge,” International Journal of Computer Vision, vol. 88, pp. 303-338, 2010.

[21] R. Mottaghi, X. Chen, X. Liu, N.- G. Cho, S.- W. Lee, S. Fidler, R. Urtasun, and A. Yuille, “The role of context for object detection and semantic segmentation in the wild,” in CVPR, 2014, pp. 891-898.

[22] B. Li, K. Q. Weinberger, S. Belongie, V. Koltun, and R. Ranftl, “Language-driven semantic segmentation,” in ICLR, 2022, pp. 1-13.

[23] Golnaz Ghiasi, Xiuye Gu, Yin Cui, and Tsung-Yi Lin, “Scaling open-vocabulary image segmentation with image-level labels,” in ECCV, 2022, p. 540-557.

[24] K. Han, Y. Liu, J. H. Liew, H. Ding, J. Liu, Y. Wang, Y. Tang, Y. Yang, J. Feng, Y. Zhao, et al., “Global knowledge calibration for fast open-vocabulary segmentation,” in ICCV, 2023, pp. 797-807.

[25] J. Xu, S. De Mello, S. Liu, W. Byeon, T. Breuel, J. Kautz, and X. Wang, “Groupvit: Semantic segmentation emerges from text supervision,” in CVPR, 2022, pp. 18134-18144.

[26] G. Ghiasi, X. Gu, Y. Cui, and T.- Y. Lin, “Scaling open-vocabulary image segmentation with image-level labels,” in ECCV, 2022, pp. 540-557.

[27] C. Han, Y. Zhong, D. Li, K. Han, and L. Ma, “Zero-shot semantic segmentation with decoupled one- pass network,” in ICCV, 2023, pp. 1086-1096.

[28] F. Liang, B. Wu, X. Dai, K. Li, Y. Zhao, H. Zhang, P. Zhang, P. Vajda, and D. Marculescu, “Open- vocabulary semantic segmentation with mask- adapted clip,” in CVPR, 2023, pp. 7061-7070.

[29] Y. Liu, S. Bai, G. Li, Y. Wang, and Y. Tang, “Open-vocabulary segmentation with semantic-assisted calibration,” in CVPR, 2024, pp. 3491-3500.

[30] S. D. Dao, H. Shi, D. Phung, and J. Cai, “Class enhancement losses with pseudo labels for open-vocabulary semantic segmentation,” IEEE Transactions on Multimedia, vol. 26, pp. 8442-8453, 2024.

[31] X. Shan, D. Wu, G. Zhu, Y. Shao, N. Sang, and C. Gao, “Open-vocabulary semantic segmentation with image embedding balancing,” in CVPR, 2024, pp. 28412-28421.

[32] Y. Pan, R. Sun, Y. Wang, W. Yang, T. Zhang, and Y. Zhang, “Purify then guide: A bi- directional bridge network for open-vocabulary semantic segmentation,” IEEE Transactions on Multimedia, vol. 35, no. 1, pp. 343-356, 2025.