Cascade-CLIP：用于零样本语义分割的级联视觉语言嵌入对齐

李运恒 [1]　李钟毓 [1]　曾泉胜 [1]　侯淇彬 [1,2*]　程明明 [1,2]
[1] 南开大学，计算机学院　[2] 南开大学，深圳福田研究院

2025 年 2 月 17 日

Abstract

预训练视觉-语言模型，例如 CLIP，已成功应用于零样本语义分割。现有的基于 CLIP 的方法主要利用最后一层的视觉特征与文本嵌入对齐，而忽略了包含丰富目标细节的中间层特征。然而，我们发现直接聚合 CLIP 多层视觉特征会削弱模型对新类别的零样本分割能力。由于不同层的视觉特征之间的差异使得直接聚合后特征难以与文本嵌入良好对齐。为解决这个问题，我们通过引入一系列独立的解码器，以级联的方式将多级视觉特征与文本嵌入对齐，形成了一种新颖但简单的框架—Cascade-CLIP。Cascade-CLIP 具有灵活性，可以轻松应用于现有的零样本语义分割方法。实验结果表明，Cascade-CLIP 在分割基准测试上，如 COCO-Stuff、Pascal-VOC 和 Pascal-Context，取得了卓越的零样本性能。代码可通过https://github.com/HVision-NKU/Cascade-CLIP获取。

1. 引言

语义分割作为计算机视觉的基本课题之一，在预测图像中每个像素的类别方面效果显著 (Chen et al., 2021; Huang et al., 2021; Xie et al., 2021; Cheng et al., 2022a)。然而，在封闭集标注图像上训练的语义分割模型 (Zhao et al., 2017; Zeng et al., 2022) 仅能分割预定义的类别。这促使一些研究者研究零样本语义分割模型 (Bucher et al., 2019; Gu et al., 2020; Han et al., 2023b)，这些模型能够分割在训练图像中根本不存在的类别，并且越来越受到关注。

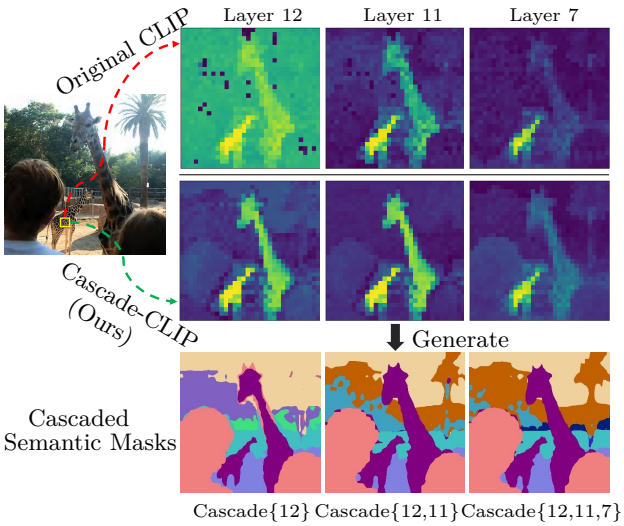
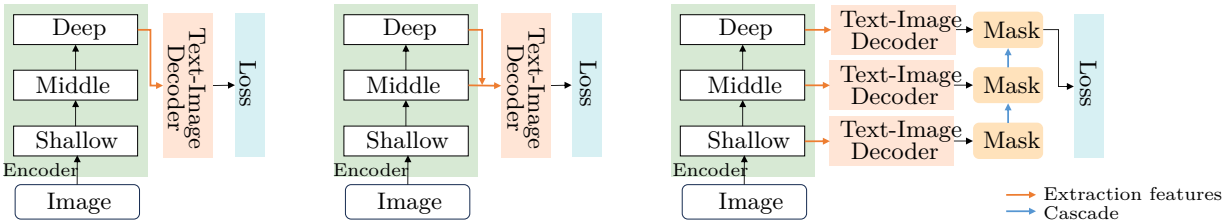


Figure 1: Cascade-CLIP 的动机示意图。余弦相似度图（上方）表明，与最后一层特征（第 12 层）相比，CLIP (Radford et al., 2021) 中间层的视觉特征能够捕捉到更丰富的局部物体细节。

最近，得益于在图像级别上的令人印象深刻的零样本能力，以 CLIP (Radford et al., 2021) 为代表的大规模视觉-语言预训练模型已经被考虑用于零样本语义分割。然而，直接将 CLIP 应用于零样本语义分割任务是无效的，因为它需要密集的像素/区域级预测。两阶段方法 (Xu et al., 2022; Ding et al., 2022) 通过训练的 Proposal 生成器生成区域 Proposal，并将裁剪的掩码区域输入 CLIP 进行零样本分类，从而解决了上述问题。尽管这种范式很好地保留了 CLIP 的图像级零样本能力，但其计算成本较高。一阶段方法 (Zhou et al., 2023; Xu et al., 2024) 通过匹配文本嵌入与从 CLIP 视觉编码器最后一层提取的像素级特征来生成像素级分割，在效率和效果之间实现了良好的平衡。然而，这些方法在分割物体细节，特别是语义对象的边界方面表现较弱。借鉴闭集分割方法的经验 (Zheng et al., 2021;

*侯淇彬是通讯作者



(a) ZegCLIP (Zhou et al., 2023) (b) Multi-layer Features Fusion (c) Cascade-CLIP (Ours)

Figure 2: 三种基于 CLIP 的零样本分割方法。(a) ZegCLIP 依赖于**最后一层**的视觉特征，未考虑中间层的信息。(b) 受 SegFormer (Xie et al., 2021) 启发，我们融合了**中间层**和**最后一层**的特征以增强特征表示，但这种融合破坏了文本与视觉特征之间的相关性。(c) 为了解决这一问题，我们的 Cascade-CLIP **分离**图像编码器，并分别为深层特征和中间层特征**对齐**独立的文本-图像解码器，最后**级联**分割结果。

Table 1: 与不同方法在 COCO-Stuff 164K 和 PASCAL VOC 2012 数据集上的比较[1]。简单地融合多层特征会导致性能下降。

	COCO-Stuff 164K			PASCAL VOC 2012		
Methods	mIoU^S	mIoU^U	hIoU	mIoU^S	mIoU^U	hIoU
ZegCLIP (Baseline)	40.1	39.5	39.8	90.5	78.3	84.0
Multi-layer Fusion	37.7	39.0	38.4	91.2	75.0	82.3
Cascade-CLIP	41.1	43.4	42.2	92.7	83.1	87.7

Hou et al., 2020; Xie et al., 2021; Guo et al., 2022)，捕捉丰富局部细节的有效解决方案是从编码器中聚合多层特征，以改善粗略的分割结果。对于 CLIP 模型，我们观察到从中间层提取的视觉特征包含了丰富的对象细节，如图 1 所示。然而，直接融合多层特征会导致不理想的结果。正如表1所示，将中间层与最后一层特征直接融合 (图 2(b)) 反而会降低相对于基准模型 (图 2(a)) 的性能。基线模型的成功在于有效利用 CLIP 预训练的最后一层视觉特征与文本嵌入之间的关联。然而，多层特征的融合会由于中间层与最后一层特征间的显著差异，破坏原有的视觉-语言相关性，从而削弱 CLIP 在未见类别上的零样本分割能力。此外，特征融合后，不同特征间的差异也破坏了预训练的视觉表示，进一步增加了在微调过程中将视觉特征与文本嵌入对齐的难度。

在本文中，我们改进了视觉和文本嵌入对齐的方法，提出 Cascade-CLIP，一个多层级框架，可以更好地利用 CLIP 的多样化视觉特征，增强其对新类别的迁移能力。具体而言，Cascade-CLIP 将视觉编码器分为多个阶段，确保每个阶段内的特征变化较小。每个阶段配备独立的文本-图像解码器，使用不同的文本嵌入更好地对齐多层级视觉特征，建立更好的视觉-语言相关性。通过这种方式，我们可以整合视觉编码器中的互补多层次语义掩码，以增强分割结果，以增强分割结果，如图 1（底行）所示。通过利用多层级特征，我们首次证明了 Cascade-CLIP 可以显著提高 CLIP 在零样本语义分割任务中的图像到像素的适应性。此外，Cascade-CLIP 具有很好的灵活性，可以无缝应用于现有的最先进方法（如 ZegCLIP (Zhou et al., 2023) 和 SPT-SEG (Xu et al., 2024)），以提升它们在三个常用零样本分割基准上的性能。特别是，由于级联视觉-语言对齐的优势，我们的方法在未见类别上的表现尤为突出，反映出强大的适应能力。本文的贡献总结如下：

- 本文揭示了 CLIP 中间层提取的视觉特征包含有关目标的丰富局部信息。然而，简单融合多级视觉特征会削弱 CLIP 的零样本能力。
- 本文提出了 Cascade-CLIP，一个灵活的多级视觉-语言嵌入对齐框架，能够有效地利用 CLIP 的多级视觉特征，以提高新类别的迁移性。
- 广泛的实验证明了我们的 Cascade-CLIP 在三个广泛使用的基准上的零样本语义分割的有效性。

2. 相关工作

2.1. 预训练视觉语言模型

大规模的视觉-语言模型 (Jia et al., 2021; Kim et al., 2021; Radford et al., 2021) 使用海量的图像-文本对进行预训练，在图像和文本嵌入对齐方面取得了显著进展，并实现了强大的零样本和少样本泛化能力。例

[1] mIoU^S 和 mIoU^U 分别表示已见类和未见类的平均交并比。hIoU 表示已见类和未见类之间的调和平均交并比。

如，最受欢迎的视觉-语言模型之一 CLIP (Radford et al., 2021)，通过对比学习使用 4 亿个图像-文本对进行训练。凭借其强大的零样本识别能力和简洁的结构，CLIP 已被广泛应用于各种下游任务中，包括零样本视觉识别 (Khattak et al., 2023)、密集预测 (Rao et al., 2022)、目标检测 (Gu et al., 2021) 和视觉指称表达 (Wang et al., 2022) 等。本文探讨了如何有效地将 CLIP 强大的泛化能力从图像级别转移到像素级分类任务中。

2.2. 零样本语义分割

零样本语义分割任务执行像素级分类，包括训练期间未见过的类别。先前的工作如 SPNet (Xian et al., 2019)、ZS3 (Bucher et al., 2019)、CaGNet (Gu et al., 2020)、SIGN (Cheng et al., 2021b)、JoEm (Baek et al., 2021) 和 STRICT (Pastore et al., 2021) 专注于学习视觉空间和语义空间之间的映射，以提高从已见类别到未见类别的语义泛化能力。最近的方法大多采用具有强大零样本分类能力的大规模视觉-语言模型（如 CLIP (Radford et al., 2021) 和 ALIGN (Jia et al., 2021)）进行零样本语义分割。一些免训练的方法，例如 ReCo (Shin et al., 2022) 和 CaR (Sun et al., 2023)，直接使用 CLIP 执行零样本语义分割。其他方法，如 MaskCLIP+ (Zhou et al., 2022a)，则应用 CLIP 为未见类别生成伪标注，以训练现有的分割模型，但对未见类别名称的依赖限制了其应用范围。为了解决这一问题，一些工作如 ZegFormer (Ding et al., 2022)、Zsseg (Xu et al., 2022)、FreeSeg (Qin et al., 2023) 和 DeOP (Han et al., 2023a)，将零样本语义分割任务解耦为类别无关的掩码生成过程以及利用 CLIP 进行掩码类别分类的过程。这些方法虽然保留了 CLIP 在图像层面的零样本能力，但由于引入了 Proposal 生成器，计算成本不可避免地增加。与使用重量级 Proposal 生成器不同，ZegCLIP (Zhou et al., 2023) 引入了轻量级解码器，将文本嵌入与从 CLIP 提取的视觉嵌入进行匹配。类似地，SPT-SEG (Xu et al., 2024) 通过整合光谱信息增强了 CLIP 的语义理解能力。尽管上述方法已经成功地将 CLIP 的图像分类能力转化为像素级分割，但仍有很大的改进空间。与之前的工作不同，我们从新的角度出发，探讨了视觉编码器中间层特征在零样本语义分割中的作用。

3. 本文方法

零样本语义分割任务 (Bucher et al., 2019; Zhou et al., 2023) 旨在在一个仅包含已见类别像素标注的数据集上训练后，能够对已见类别 $\mathcal{C}$ 和未见类别 $\hat{\mathcal{C}}$ 进行分割。通常情况下，$\mathcal{C} \cap \hat{\mathcal{C}} = \varnothing$，并且在训练时未见类别 $\hat{\mathcal{C}}$ 的标签是不可用的。关键问题是在对已见类别进行训练时，保留识别未见类别的能力。

3.1. 回顾 ZegCLIP

由于高效性和良好的性能，最近的零样本语义分割方法 (Zhou et al., 2023; Xu et al., 2024) 大多基于单阶段方案。这里，我们以 ZegCLIP 工作 (Zhou et al., 2023) 作为基线进行回顾。

如图 2(a) 所示，ZegCLIP (Zhou et al., 2023) 首先提取了 CLIP 的文本嵌入 C 类作为 $\mathrm{T} = [\mathrm{t}_1, \mathrm{t}_2, \ldots, \mathrm{t}_C] \in \mathbb{R}^{C \times d}$ 和 CLIP 图像的视觉特征，包括 [CLS] 标记 $\mathrm{g} \in \mathbb{R}^{1 \times d}$ 和块标记 $\mathrm{H} \in \mathbb{R}^{N \times d}$，其中 d 是 CLIP 模型的特征维度，N 是块标记的数量。C 是类别的数量，在训练期间 $C = |\mathcal{C}|$，在推理期间 $C = |\mathcal{C} \cup \hat{\mathcal{C}}|$。为了避免过拟合，ZegCLIP (Zhou et al., 2023) 使用了关系描述符 $\widehat{\mathrm{T}} = \mathrm{concat}(\mathrm{T} \odot \mathrm{g}, \mathrm{T}) \in \mathbb{R}^{C \times 2d}$ 代替 T，其中 $\odot$ 和 concat 分别是哈德曼积和连接操作。然后，可以通过在文本-图像解码器中计算文本嵌入 $\widehat{\mathrm{T}}$ 和视觉特征 H 之间的相似性来生成语义掩码 $\mathrm{M} \in \mathbb{R}^{C \times N}$ 。整个过程可以表示为：

$$\mathrm{M} = \mathrm{Softmax}(\mathcal{D}(\phi_q(\widehat{\mathrm{T}}), \phi_k(\mathrm{H}))), \tag{1}$$

其中 $\mathcal{D}(\cdot)$ 表示文本-图像解码器，如图 3右侧所示，ϕ_q 和 ϕ_k 是两个线性投影，用于对齐 $\widehat{\mathrm{T}}$ 和 H 的特征维度。由于视觉特征仅从视觉编码器的最后一层提取，因此以前的方法通常无法很好地识别语义目标的边界。这是因为深层特征携带如图 Fig. 1所示的高级语义全局特征，但与中间层相比，低级局部细节较少，这将是本文重点关注的内容。

3.2. 研究动机

多级特征通常被用于闭集分割模型中 (Zheng et al., 2021; Xie et al., 2021) 以锐化目标分割细节。我们在第 1节中的分析也揭示了 CLIP (Radford et al., 2021) 中层特征能够捕捉丰富的局部目标细节。这促使我们

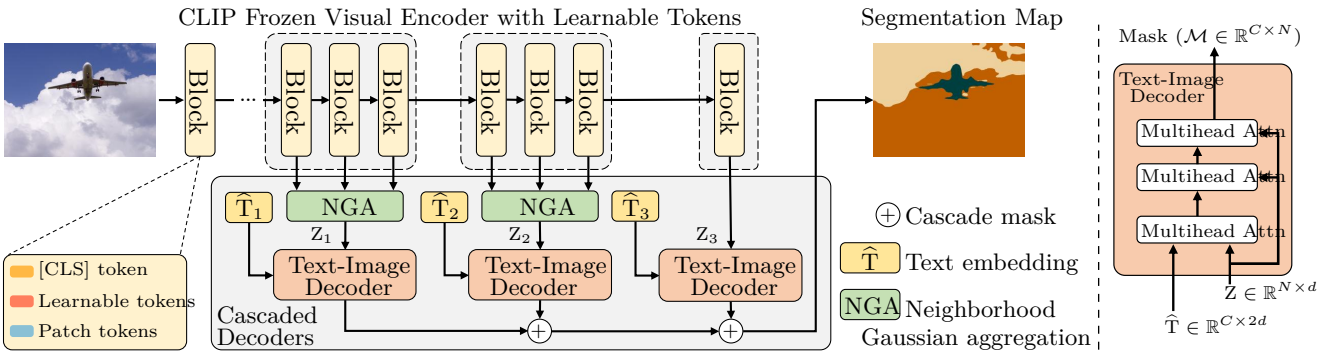
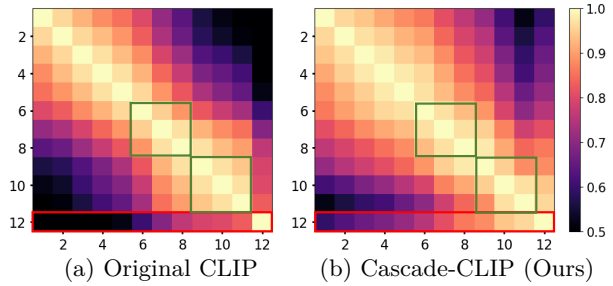


Figure 3: Cascade-CLIP 架构。CLIP 视觉编码器被分为多个阶段。然后，我们使用 NGA 模块聚合每个阶段内块的特征，并为聚合的视觉特征和非共享的文本嵌入分配独立的文本-图像解码器。在文本-图像解码器中（图的右侧部分），受 (Zhang et al., 2022) 启发，分割掩码可以通过多头注意力（Attn）层的缩放点积注意力计算得到。最后，我们结合由不同级联解码器生成的多层次语义掩码，以增强分割预测。（详情请参见第 3.3节。）



(a) Original CLIP (b) Cascade-CLIP (Ours)

Figure 4: 中心核对齐热力图 (Kornblith et al., 2019) 展示了（a）原始 CLIP 模型与（b）Cascade-CLIP（本方法）各网络层间的特征相似性。最后一行（红框区域）展示了最后一层特征与其他网络层特征的相似性。绿框区域则呈现了相邻网络层之间的特征相似性。

研究如何有效地利用这些独特的特征来增强 CLIP 对新类别的可迁移性，这是先前工作中被忽略的。然而，如图 2(b) 中简单地聚合多级视觉特征会降低分割性能。为了分析性能下降的原因，我们尝试可视化 CLIP 视觉特征的中心核对齐图 (Kornblith et al., 2019)，如图 4(a) 所示，该图计算了不同层次之间的相似性。我们观察到浅层和深层特征之间存在显著差异，且差异随着网络深度增加而增大。这表明，由于层次间存在显著差异，直接将多级中间特征整合到最后一个特征中可能会破坏预训练 CLIP 中的视觉-语言嵌入对齐，从而削弱 CLIP 的零样本能力。基于上述分析，我们重点研究如何有效地利用具有丰富局部细节的中间特征来改善零样本分割。为了应对这一挑战，我们提出了两种策略，即级联视觉-语言嵌入对齐和邻域高斯聚合，以更好地将多级视觉特征与文本嵌入对齐。这些策略旨在减少不同层次之间的特征差异，使得中层视觉特征能够与文本嵌入良好对齐并补充深层特征，提高零样本分割的能力。

3.3. 级联对齐框架

Cascade-CLIP 框架概览如 Fig. 3 所示。基本做法是将 CLIP 的视觉编码器分割成多个阶段以提取多级视觉特征，每个阶段的特点略有变化。为了在微调过程中更好地建立视觉与语言的关联，考虑到各阶段特征之间的差异，我们为视觉编码器的每个阶段分配了一个独立的文本-图像解码器。该解码器与 Sec. 3.1 中提到的类似。最后，通过级联来自不同阶段的互补分割掩码生成细化结果。

具体来说，设 H_l 表示第 l 个 Transformer block 的 patch tokens。对于 ViT-B，block 的数量应为 12。首先，我们将 CLIP 的视觉编码器分为 S 个阶段，每个阶段包含一组 Transformer blocks。在每个阶段，例如第 s 阶段，为了更好地利用不同 Transformer blocks 的多级特征，我们引入了一个邻域高斯聚合 (NGA) 模块来聚合这些特征，生成聚合特征 Z_s。我们将在后面详细描述 NGA 模块。然后，对于来自第 s 阶段的输出 Z_s，我们关联了一个通过线性投影从 $\hat{\mathrm{T}}$ 获得的对应文本嵌入 $\hat{\mathrm{T}}_s$。随后，将 Z_s 和 $\hat{\mathrm{T}}_s$ 输入独立的文本-图像解码器以生成语义掩码。最后，我们通过组合多个阶段生成的所有语义掩码代替 Eq. 1 中的单个语义掩码，如下所示：

$$\mathrm{M}_S = \mathrm{Softmax}\left(\sum_{s=1}^{S} \mathcal{D}_s(\widehat{\mathrm{T}}_s, \mathrm{Z}_s)\right), \quad (2)$$

其中 $\mathcal{D}_s(\cdot)$ 表示第 s 个文本-图像解码器。这里，我们使用逐元素求和操作，可以将其视为多个级联解码器输出的集成。

如 Fig. 3 所示，视觉-语言对齐过程可以以级联的方式多次应用于不同的块。在实际操作中，我们不会将文本-图像编码器附加到浅层 Transformer 块，因为浅层特征包含较少的语义信息。我们在消融实验将展示如何划分视觉编码器以利用多级视觉特征。

级联掩码的损失函数给定来自第 s 阶段的文本-图像解码器 $\mathcal{D}_s(\cdot)$，令 $\mathcal{M}_s = \mathcal{D}_s(\widehat{\mathrm{T}}_s, \mathrm{Z}_s)$ 为预测的分割掩码。$\mathcal{M} = \sum_{s=1}^{S} \mathcal{M}_s$ 是多级级联掩码。目标损失函数 $\mathcal{L}^{\text{pixel}}$ 定义为：

$$\mathcal{L}^{\text{pixel}} = \alpha\mathcal{L}^{\text{dice}}(\mathrm{Y}, \mathcal{M}) + \beta\mathcal{L}^{\text{focal}}(\mathrm{Y}, \mathcal{M}), \quad (3)$$

其中 $\mathcal{L}^{\text{dice}}$ 和 $\mathcal{L}^{\text{focal}}$ 分别是使用 Sigmoid 作为激活函数的 dice 损失 dice loss (Milletari et al., 2016) 和 focal loss (Lin et al., 2017)。Y 是真实值。$\{\alpha, \beta\}$ 是两个权重，默认值分别为 $\{1, 100\}$。

为了更好地将中间视觉特征与文本嵌入对齐，我们采用视觉提示调整 (Zhou et al., 2022b; Ding et al., 2022)，通过在冻结编码器的每个块上的视觉特征中引入可学习的标记。在视觉提示调整过程中，级联对齐方式使梯度可以直接反向传播到视觉编码器的中间层。这可以促进中间层特征与文本嵌入的对齐，极大地增强了不同层之间的相似性。我们在图 4(b) 中说明了这一点，这与图 4(a) 有明显的区别。

邻域高斯聚合 为了更好地利用每个 Transformer 块的特征潜力，我们提出了邻域高斯聚合 (NGA) 模块，以在每个阶段内融合多级特征。基于第 3.2节的分析和图 4(b) 的说明，我们观察到随着距离的增加，跨层的特征相似性逐渐下降。因此，我们提出在特征融合时根据它们相对邻域距离为不同的块分配独特的高斯权重。此外，这些权重关于训练数据是可训练的，这使得可以从每个编码器阶段的各个块中获取自适应权重信息。考虑到由 d 个 Transformer 块组成的第 s 阶段编码器，高斯权重 $W_{s,l}$ 和聚合特征 Z_s 可以计算为：

$$\begin{aligned} W_{s,l} &= \exp\left(-\frac{1}{2}\frac{(d-l+1)^2}{\sigma^2}\right), \quad l \in [1, d], \\ \mathrm{Z}_s &= \sum_{l=1}^{d} \mathrm{H}_l \cdot W_{s,l}, \end{aligned} \quad (4)$$

其中高斯函数的方差参数 σ 默认设置为 1. l 对应于 Transformer 块的索引。增加 σ 会使 Transformer 块之间的权重均衡化，而减少 σ 则会导致依赖于单一块特征（如附录 C 中所示）。通过设置方差参数 σ，NGA 模块可以为邻近块分配较高的权重，为远端块分配较低的权重，从而更有效、灵活地整合不同深度层特征。

4. 实验

4.1. 数据集和评价指标

为了评估我们提出方法的有效性，我们在三个广泛使用的基准数据集上进行了大量实验，包括 COCO-Stuff ((Caesar et al., 2018))、Pascal-VOC ((Everingham et al., 2015)) 和 Pascal-Context ((Mottaghi et al., 2014))。已知类别和未知类别的划分遵循之前工作 ((Zhou et al., 2023)) 的常见设置，同时报告了已知类别和未知类别的平均交并比 (mIoU) 和调和平均交并比 (hmIoU)。更多有关数据集统计、数据划分和评价指标的细节可以在附录中找到。

4.2. 实施细节

我们在开源工具箱 MMSegmentation ((Contributors, 2020)) 上实施提出的方法，并使用配备 4 块 NVIDIA RTX 3090 GPU 的机器进行所有实验。采用包含 12 个 Transformer 块的 VIT-B/16 ((Dosovitskiy et al., 2020)) 作为 CLIP ((Radford et al., 2021)) 的图像编码器。每个 GPU 上的批处理大小设置为 4，输入图像分辨率为 512×512。优化器采用 AdamW ((Loshchilov & Hutter, 2019))，并使用 MMSeg 工具箱中的默认训练计划。为了公平比较，在每个数据集上使用与 ZegCLIP ((Zhou et al., 2023)) 相同的训练迭代次数。

4.3. 与最先进方法的对比

为了证明我们的 Cascade-CLIP 的有效性，将评估结果与之前最先进的方法进行了比较，包括双编码器方法（例如，ZegFormer ((Ding et al., 2022))，Zsseg ((Xu

Table 2: 与当前最先进的**归纳式**和**转导式**零样本分割方法在 COCO-Stuff 164K 和 PASCAL VOC 2012 数据集上的对比。R 表示 ResNet (He et al., 2016)。ST 表示使用生成的未见类别的伪标签重新训练模型。我们的 Cascade-CLIP 方法使用一个三阶段级联解码器整合了从图像编码器第 6 层到第 12 层提取的特征：6-8, 9-11, 12。

Methods	Backbone	Segmentor[2]	COCO-Stuff 164K (171)			PASCAL VOC 2012 (20)		
			$\text{mIoU}^S\uparrow$	$\text{mIoU}^U\uparrow$	hIoU↑	$\text{mIoU}^S\uparrow$	$\text{mIoU}^U\uparrow$	hIoU↑
Inductive: training images do not contain any unseen objects.								
ZegFormer (Ding et al., 2022)	R101&CLIP-B	MaskFormer	36.6	33.2	34.8	86.4	63.6	73.3
Zsseg (Xu et al., 2022)	R101&CLIP-B	MaskFormer	39.3	36.3	37.8	83.5	72.5	77.5
DeOP (Han et al., 2023a)	R101&CLIP-B	MaskFormer	38.0	38.4	38.2	88.2	74.6	80.8
Zsseg+MAFT (Jiao et al., 2023)	R101&CLIP-B	MaskFormer	40.6	40.1	40.3	88.4	66.2	75.7
SPNet-C (Xian et al., 2019)	R101	W2V&FT	35.2	8.7	14.0	78.0	15.6	26.1
ZS3Net (Bucher et al., 2019)	R101	W2V	34.7	9.5	15.0	77.3	17.7	28.7
CaGNet (Gu et al., 2020)	R101	W2V&FT	33.5	12.2	18.2	78.4	26.6	39.7
SIGN (Cheng et al., 2021b)	R101	W2V&FT	32.3	15.5	20.9	75.4	28.9	41.7
JoEm (Baek et al., 2021)	R101	W2V	-	-	-	77.7	32.5	45.9
ZegCLIP (Zhou et al., 2023)	CLIP-B	SegViT	40.2	41.4	40.8	91.9	77.8	84.3
Cascade-CLIP (Ours)	CLIP-B	SegViT	41.1	43.4	42.2	92.7	83.1	87.7
Transductive: training images employ the names of unseen classes.								
Zsseg+ST (Xu et al., 2022)	R101&CLIP-B	MaskFormer	39.6	43.6	41.5	79.2	78.1	79.3
FreeSeg (Qin et al., 2023)	R101&CLIP-B	Mask2Former	42.4	42.2	42.3	91.9	78.6	84.7
FreeSeg+MAFT (Jiao et al., 2023)	R101&CLIP-B	Mask2Former	44.1	55.2	49.0	90.0	86.3	88.1
SPNet-C+ST (Xian et al., 2019)	R101	W2V&FT	34.6	26.9	30.3	77.8	25.8	38.8
ZS5Net (Bucher et al., 2019)	R101	W2V	34.9	10.6	16.2	78.0	21.2	33.3
CaGNet+ST (Gu et al., 2020)	R101	W2V&FT	35.6	13.4	19.5	78.6	30.3	43.7
MaskCLIP+ (Zhou et al., 2022a)	R101	DeepLabv2	38.1	54.7	45.0	88.8	86.1	87.4
MVP-SEG+ (Guo et al., 2023)	R101	DeepLabv2	38.3	55.8	39.9	44.9	67.5	54.0
ZegCLIP+ST (Zhou et al., 2023)	CLIP-B	SegViT	40.7	59.9	48.5	92.3	89.9	91.1
TagCLIP+ST (Li et al., 2023)	CLIP-B	SegViT	40.4	60.0	48.3	94.3	92.7	93.5
Cascade-CLIP+ST (Ours)	CLIP-B	SegViT	41.7	62.5	50.0	93.3	93.4	93.4

et al., 2022) 和 DeOP ((Han et al., 2023a))) 和单编码器方法（例如，ZegCLIP ((Zhou et al., 2023)))。

归纳式设置下的比较。如表 2所示，在归纳设置中，Cascade-CLIP 显著提高了性能，该设置下未提供未见类别的特征和标注。值得注意的是，在提升看到类的结果的同时，我们的方法也提高了未见类的性能。例如，在 COCO 和 Pascal VOC 上，对于未见类的 mIoU，Cascade-CLIP 将最先进性能分别提升了 2.0% 和 5.3%，这证明了其在零样本分割中的强大泛化能力。

[2]Segmentors 的描述源于 (Zhu & Chen, 2023). MaskFormer 和 MasksFormer 在 (Cheng et al., 2021a) 和 (Cheng et al., 2022b) 中提出; DeepLabv2 在 (Chen et al., 2018) 中提出; W2V 在 (Mikolov et al., 2013) 中提出; SegViT 在 (Zhang et al., 2022) 中提出。

转换式设置下的比较。我们进一步评估了 Cascade-CLIP 在转导式设置下的可迁移性，该设置下模型通过为未见像素生成伪标签并利用已见像素的真实标签进行重新训练。表2显示，在转换自训练后，我们的模型显著提高了未见类的性能，同时在看到类上始终保持了优秀的性能。为了进一步验证 Cascade-CLIP 的有效性，我们在 PASCAL Context 数据集上与其他方法进行了比较。如表 3所示，Cascade-CLIP 在未见类的 mIoU 方面始终优于其他方法。上述结果清楚地证明了我们提出方法的有效性。关于我们方法的效果和普遍性的其他实验结果，请参见第 4.5 节。

定性结果。图5展示了基线和 Cascade-CLIP 在看到类和未见类上的分割结果。Cascade-CLIP 在看到类和未见类上都显示出令人印象深刻的分割能力，并且可以清晰地区分相似的未见类。例如，我们的方法可以更

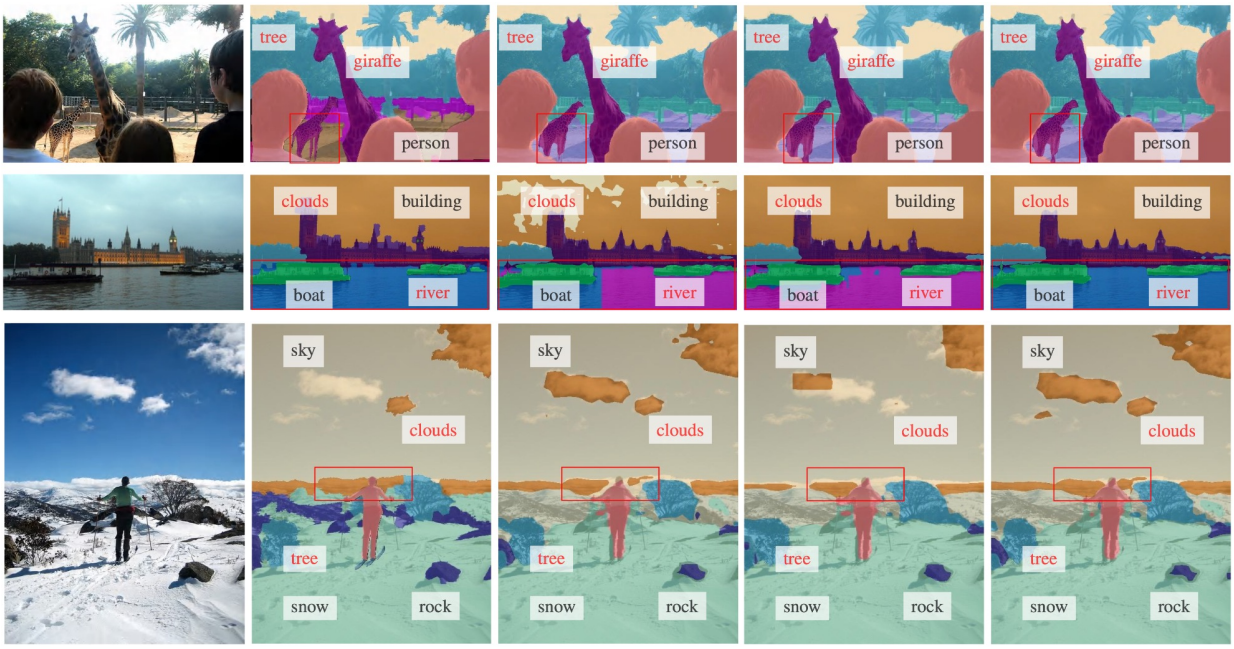


Figure 5: COCO-Stuff 164K 数据集上的定性转导式结果。黑色和红色标签分别表示已见类别和未见类别。

Table 3: 与当前最先进的零样本分割方法在 PASCAL Context 数据集上的比较。

Methods	PASCAL Context (59)		
	$\text{mIoU}^S\uparrow$	$\text{mIoU}^U\uparrow$	hIoU↑
Inductive			
SPNet-C (Xian et al., 2019)	27.1	9.8	14.4
ZS3Net (Bucher et al., 2019)	20.8	12.7	15.8
CSRL (Li et al., 2020)	29.4	14.6	19.5
JoEm (Baek et al., 2021)	33.0	14.9	20.5
ZegCLIP (Zhou et al., 2023)	53.8	45.5	49.3
Cascade-CLIP (Ours)	55.9	47.2	51.2
Transductive			
ZS5Net (Bucher et al., 2019)	27.0	20.7	23.4
ZegCLIP+ST (Zhou et al., 2023)	54.5	41.4	47.1
Cascade-CLIP+ST (Ours)	56.4	55.0	55.7

好地区分‘长颈鹿’区域和‘树’区域（图5(1)），‘船’区域和‘河’区域（图5(2)），以及‘云’区域和‘天空’区域（图5(3)）。更多的定性结果见附录。

4.4. 消融实验

逐组件消融研究。为了理解 Cascade-CLIP 中每个组件的作用，包括级联解码器和 NGA 模块，我们从基线 ZegCLIP 开始，该基线使用 CLIP 最后一层的视觉特征，然后逐步加入每个提出的模块。如表4所示，从编码器的不同块中捕捉到独特且互补的信息，将未见类别的 mIoU 分数提升了 3.1%（第 2 个结果）。在此基础上，引入 NGA 模块以聚合每个分割编码器中物体的丰富局部信息，进一步提升了未见类别的 mIoU 分数（第 3 个结果）。

Table 4: Cascade-CLIP 组件的消融实验。

Cascaded decoders	NGA	$\text{mIoU}^S\uparrow$	$\text{mIoU}^U\uparrow$	hIoU↑
✗	✗	40.1	39.5	39.8
✓	✗	40.8	42.6	41.7
✓	✓	41.1	43.4	42.2

Table 5: 不同块分割方式的效果。{ } 中的元素表示用于融合编码器特征的块编号。当块数超过 1 时，使用 NGA 进行融合。

Block splitting manner	#Decoders	$\text{mIoU}^S\uparrow$	$\text{mIoU}^U\uparrow$	hIoU↑
{4-6}, {7-9}, {10-12}	3	40.4	42.7	41.5
{6-8}, {9-10}, {11-12}	3	40.7	42.4	41.5
{6-8}, {9-11}, {12}	3	41.1	43.4	42.2

所提出块分割方式的效果。级联解码器架构在 Cascade-CLIP 中至关重要，因为它能够保持视觉与语言的关联。我们在表5中的分析表明，将最后一个块独立为一个阶段（第 3 个结果）比其他分割策略的组合（第 1 个和第 2 个结果）更有效。这是因为 CLIP 图像编码器最

Table 6: Cascade-CLIP 视觉编码器中级联解码器数量及对应块数。

Block splitting manner	Number	mIoU^S↑	mIoU^U↑	hIoU↑
Number of cascaded decoders				
{12} (Baseline)	1	40.1	39.5	39.8
{9-11}, {12}	2	40.2	40.2	40.2
{6-8}, {9-11}, {12}	3	41.1	43.4	42.2
{3-5}, {6-8}, {9-11}, {12}	4	40.8	42.5	41.7
Number of blocks				
{8-9}, {10-11}, {12}	2	40.7	41.3	41.0
{6-8}, {9-11}, {12}	3	41.1	43.4	42.2
{4-7}, {8-11}, {12}	4	41.1	41.5	41.3

Table 7: Cascade-CLIP 与其他解码器方法的比较。直接使用末层特征或简单融合多层特（与 Cascade-CLIP 相同的块数）并通过解码器对齐，性能会下降。

Description	Param. (M)	mIoU^S↑	mIoU^U↑	hIoU↑
Last-layer	40.5	40.3	40.6	40.5
Multi-layer fusion	40.5	39.9	35.7	37.7
Cascade-CLIP	40.5	41.1	43.4	42.2

Table 8: 不同聚合方法的比较。† 表示权重可训练。

Description	mIoU^S↑	mIoU^U↑	hIoU↑
Concat	39.9	41.0	40.5
Self-attention	37.7	39.0	38.4
Sum	39.3	42.0	40.6
NGA	40.9	43.0	41.9
Sum†	39.9	41.2	40.5
NGA†	41.1	43.4	42.2

后一层特征与文本嵌入的关联最强，将其匹配到单独的解码器减少了这种相关性的破坏。

级联解码器的数量和每个阶段中的块数。为了展示跨不同层信息融合的重要性，我们在表 6中展示了 Cascade-CLIP 在不同级联解码器数量及每个编码器阶段块数下的性能。可以看到，将级联解码器的数量从 1 增加到 3 逐渐提高了分割性能。这表明与仅使用最后一层特征的先前工作相比，来自各层的特征具有互补性。我们每个阶段的 Transformer 块默认值为 3。将块数减少到 2 会导致性能下降，因为忽略了中层特征。通过级联三个解码器（包括对最后一个块的额外解码器）实现了最佳性能。注意，我们没有使用用起始层特征，因为它们编码的特征语义较少。

Table 9: 独立/共享文本嵌入的效果。

Description	mIoU^S↑	mIoU^U↑	hIoU↑
Shared	40.5	42.4	41.4
Independent	41.1	43.4	42.2

Table 10: 将 Cascade-CLIP 扩展到现有方法以提升零样本分割结果。

Methods	COCO-Stuff 164K		VOC 2012	
	mIoU^S↑	mIoU^U↑	mIoU^S↑	mIoU^U↑
Inductive				
Frozen CLIP	32.3	32.5	85.9	59.5
Frozen CLIP+Ours	36.3	35.3	89.0	69.7
ZegCLIP	40.2	41.4	91.9	77.8
ZegCLIP+Ours	41.1	43.4	92.7	83.1
SPT-SEG	38.0	40.7	92.0	85.0
SPT-SEG+Ours	40.2	43.6	92.1	86.1
Transductive				
ZegCLIP+ST	40.7	59.9	92.3	89.9
ZegCLIP+Ours+ST	41.7	62.5	93.3	93.4
SPT-SEG	40.4	57.5	93.6	92.2
SPT-SEG+Ours+ST	41.7	62.1	93.8	94.4
Fully Supervised				
ZegCLIP	40.7	63.2	92.4	90.9
ZegCLIP+Ours	41.5	64.0	93.7	94.6

为了证明我们的设计在利用 CLIP 的多层次特征方面的有效性，我们还展示了特征余弦相似性图和定性分割结果。在图 6顶部，展示了在训练过程中未包含的未见类别的 Patch 相似性。我们观察到 Cascade-CLIP 的中层包含了关于局部物体的详细信息，包括边界。此外，如图 6底部所示，通过利用这些独特特征，Cascade-CLIP 与仅使用最后一块特征相比，改善了已知和未知类别的分割性能。

Cascade-CLIP **与带有多个解码器的方法对比**。为了证明整合来自不同级联解码器生成的多样化语义掩码的有效性，而不是引入可能影响性能的额外参数，我们基于最后一层特征或多层特征融合构建了一个多解码器模型。如表 7所示，Cascade-CLIP 在参数量相等的条件下，优于最后一层和多层特征方法。这表明仅依赖最后一层特征无法产生互补和增强的分割结果。此外，直接融合特征会导致零样本能力的下降，即使使用多个解码器也无法改善。

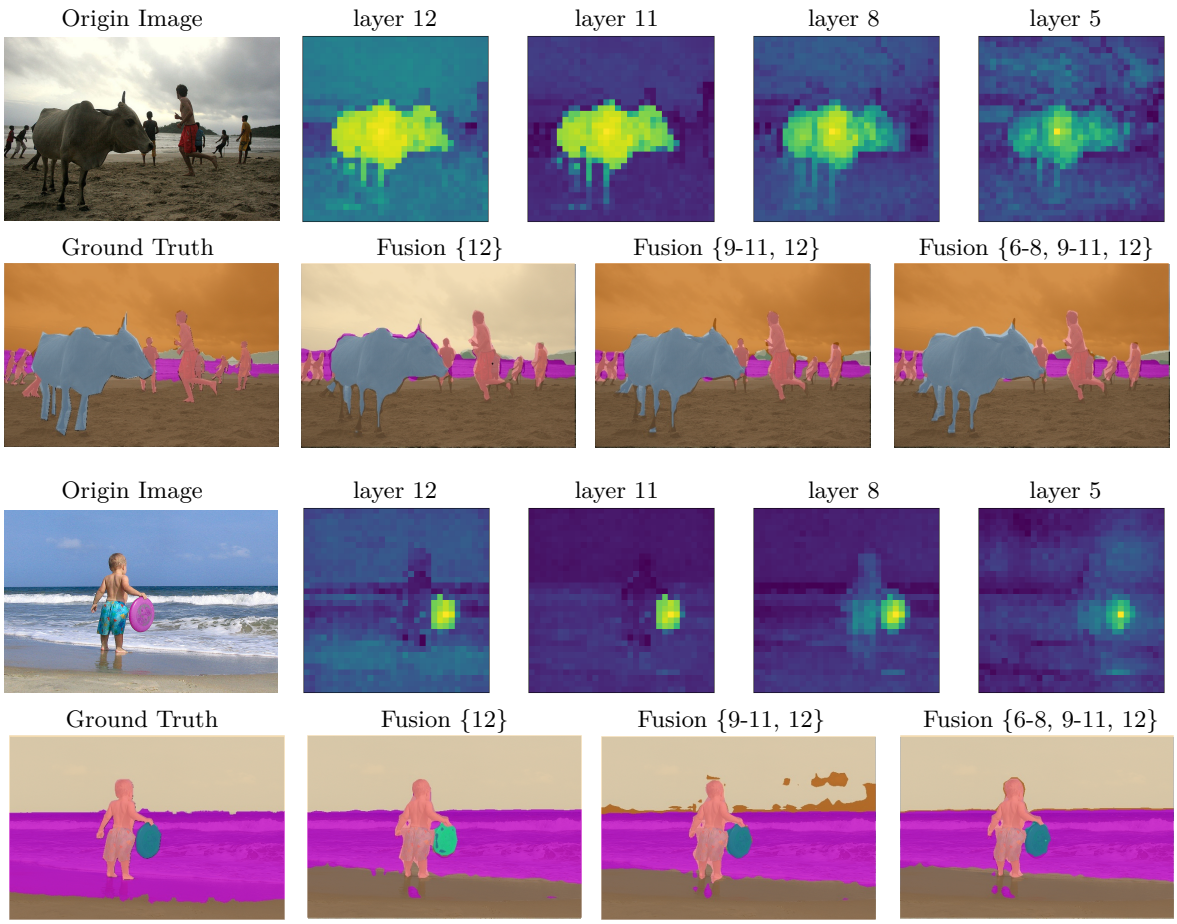


Figure 6: 特征余弦相似度图及定性分割结果的可视化。我们通过对 Cascade-CLIP 中深层和浅层视觉标记的余弦相似度计算，可视化**未见**类别的特征对应关系。归纳式分割结果通过级联不同解码器生成。

NGA **与其他聚合方法的对比**。在每个分割编码阶段，不同层次特征之间的差异会在聚合各层后破坏特征空间。为了克服这个问题，我们提出了邻域高斯聚合（NGA），通过考虑块之间的距离来减少原始特征空间中的中断。如表 8所示，我们的 NGA 优于常见的聚合策略（例如，求和、连接和自注意力）。通过可学习的权重，我们的 NGA 进一步提升了性能。这表明，在进行多层次特征融合时，为远距离特征分配较小权重的我们的 NGA，在改进零样本分割方面比其他特征聚合方法更有优势。

独立/共享文本嵌入的效果。由于来自不同分割编码阶段的特征表现出显著差异，将不同的文本嵌入与每个阶段的特征对齐是至关重要的。这在表 9 的结果中得到验证，我们的带有独立文本嵌入的 Cascade-CLIP 比带有共享文本嵌入的获得了更高的 mIoU 分数。

4.5. 将 Cascade-CLIP 扩展到其他方法

我们的方法是一个用于提高零样本分割能力的通用框架。具体来说，我们可以将级联-CLIP 无缝集成到现有的流行零样本语义分割方法中，例如，Frozen CLIP ((Radford et al., 2021))，ZegCLIP ((Zhou et al., 2023)) 和 SPT-SEG ((Xu et al., 2024))。如表如 10所示，我们的方法可以显著提高这些方法的性能，证明了所提出方法的泛化能力。

5. 结论

本文侧重于利用具有丰富局部细节但与深度特征存在显著差异的 CLIP 中间特征，以增强零样本语义分割。通过引入级联掩码机制，我们提出了 Cascade-CLIP 框架，该框架旨在以级联的方式有效地对齐多级视觉特

征与文本嵌入,，从而增强 CLIP 从图像到像素级别的适应性。实验证明了所提出方法的有效性。

影响声明

我们的工作探索了如何利用预训练 CLIP 的多层次特征来增强模型的零样本能力，这已被证明在下游任务中有效。由于我们的方法并非针对特定应用设计，因此不直接涉及社会问题。

致谢

本研究得到了国家自然科学基金（编号：62225604、62276145）、中央高校基本科研业务费专项资金（南开大学，编号：070-63223049）、中国科协青年人才托举工程（编号：YESS20210377）的资助。计算资源由南开大学超级计算中心（NKSC）提供支持。

参考文献

Baek, D., Oh, Y., and Ham, B. Exploiting a joint embedding space for generalized zero-shot semantic segmentation. In ICCV, pp. 9536–9545, 2021.

Bucher, M., Vu, T., Cord, M., and Pérez, P. Zero-shot semantic segmentation. In NeurIPS, pp. 466–477, 2019.

Caesar, H., Uijlings, J., and Ferrari, V. Coco-stuff: Thing and stuff classes in context. In CVPR, pp. 1209–1218, 2018.

Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F., and Adam, H. Encoder-decoder with atrous separable convolution for semantic image segmentation. In ECCV, pp. 801–818, 2018.

Chen, X., Yuan, Y., Zeng, G., and Wang, J. Semi-supervised semantic segmentation with cross pseudo supervision. In CVPR, pp. 2613–2622, 2021.

Cheng, B., Schwing, A., and Kirillov, A. Per-pixel classification is not all you need for semantic segmentation. In NeurIPS, pp. 17864–17875, 2021a.

Cheng, B., Misra, I., Schwing, A. G., Kirillov, A., and Girdhar, R. Masked-attention mask transformer for universal image segmentation. In CVPR, pp. 1290–1299, 2022a.

Cheng, B., Misra, I., Schwing, A. G., Kirillov, A., and Girdhar, R. Masked-attention mask transformer for universal image segmentation. In CVPR, pp. 1290–1299, 2022b.

Cheng, J., Nandi, S., Natarajan, P., and Abd-Almageed, W. Sign: Spatial-information incorporated generative network for generalized zero-shot semantic segmentation. In ICCV, pp. 9556–9566, 2021b.

Contributors, M. MMSegmentation: Openmmlab semantic segmentation toolbox and benchmark. https://github.com/open-mmlab/mmsegmentation, 2020.

Ding, J., Xue, N., Xia, G.-S., and Dai, D. Decoupling zero-shot semantic segmentation. In CVPR, pp. 11583–11592, 2022.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929, 2020.

Everingham, M., Eslami, S. A., Van Gool, L., Williams, C. K., Winn, J., and Zisserman, A. The pascal visual object classes challenge: A retrospective. IJCV, 111:98–136, 2015.

Gu, X., Lin, T.-Y., Kuo, W., and Cui, Y. Open-vocabulary object detection via vision and language knowledge distillation. In ICLR, 2021.

Gu, Z., Zhou, S., Niu, L., Zhao, Z., and Zhang, L. Context-aware feature generation for zero-shot semantic segmentation. In ACM Multimedia, pp. 1921–1929, 2020.

Guo, J., Wang, Q., Gao, Y., Jiang, X., Tang, X., Hu, Y., and Zhang, B. Mvp-seg: Multi-view prompt

learning for open-vocabulary semantic segmentation. arXiv preprint arXiv:2304.06957, 2023.

Guo, M.-H., Lu, C.-Z., Hou, Q., Liu, Z., Cheng, M.-M., and Hu, S.-M. Segnext: Rethinking convolutional attention design for semantic segmentation. In NeurIPS, pp. 1140–1156, 2022.

Han, C., Zhong, Y., Li, D., Han, K., and Ma, L. Open-vocabulary semantic segmentation with decoupled one-pass network. In ICCV, pp. 1086–1096, 2023a.

Han, K., Liu, Y., Liew, J. H., Ding, H., Liu, J., Wang, Y., Tang, Y., Yang, Y., Feng, J., Zhao, Y., and Wei, Y. Global knowledge calibration for fast open-vocabulary segmentation. In ICCV, pp. 797–807, 2023b.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In CVPR, pp. 770–778, 2016.

Hou, Q., Zhang, L., Cheng, M.-M., and Feng, J. Strip pooling: Rethinking spatial pooling for scene parsing. In CVPR, pp. 4003–4012, 2020.

Huang, Z., Wei, Y., Wang, X., Liu, W., Huang, T. S., and Shi, H. Alignseg: Feature-aligned segmentation networks. IEEE TPAMI, 44(1):550–557, 2021.

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

Jiao, S., Wei, Y., Wang, Y., Zhao, Y., and Shi, H. Learning mask-aware clip representations for zero-shot segmentation. In NeurIPS, pp. 35631–35653, 2023.

Khattak, M. U., Wasim, S. T., Naseer, M., Khan, S., Yang, M.-H., and Khan, F. S. Self-regulating prompts: Foundational model adaptation without forgetting. In ICCV, pp. 15190–15200, 2023.

Kim, W., Son, B., and Kim, I. Vilt: Vision-and-language transformer without convolution or region supervision. In ICML, pp. 5583–5594, 2021.

Kornblith, S., Norouzi, M., Lee, H., and Hinton, G. Similarity of neural network representations revisited. In ICML, pp. 3519–3529, 2019.

Li, J., Chen, P., Qian, S., and Jia, J. Tagclip: Improving discrimination ability of open-vocabulary semantic segmentation. arXiv preprint arXiv:2304.07547, 2023.

Li, P., Wei, Y., and Yang, Y. Consistent structural relation learning for zero-shot segmentation. In NeurIPS, pp. 10317–10327, 2020.

Lin, T.-Y., Goyal, P., Girshick, R., He, K., and Dollár, P. Focal loss for dense object detection. In ICCV, pp. 2980–2988, 2017.

Loshchilov, I. and Hutter, F. Decoupled weight decay regularization. In ICLR, 2019.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. Distributed representations of words and phrases and their compositionality. In NeurIPS, 2013.

Milletari, F., Navab, N., and Ahmadi, S.-A. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In 3DV, pp. 565–571, 2016.

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

Pastore, G., Cermelli, F., Xian, Y., Mancini, M., Akata, Z., and Caputo, B. A closer look at self-training for zero-label semantic segmentation. In CVPR, pp. 2693–2702, 2021.

Qin, J., Wu, J., Yan, P., Li, M., Yuxi, R., Xiao, X., Wang, Y., Wang, R., Wen, S., Pan, X., and Wang, X. Freeseg: Unified, universal and open-vocabulary

image segmentation. In CVPR, pp. 19446–19455, 2023.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., and Sutskever, I. Learning transferable visual models from natural language supervision. In ICML, pp. 8748–8763, 2021.

Rao, Y., Zhao, W., Chen, G., Tang, Y., Zhu, Z., Huang, G., Zhou, J., and Lu, J. Denseclip: Language-guided dense prediction with context-aware prompting. In CVPR, pp. 18082–18091, 2022.

Shin, G., Xie, W., and Albanie, S. Reco: Retrieve and co-segment for zero-shot transfer. In NeurIPS, pp. 33754–33767, 2022.

Sun, S., Li, R., Torr, P., Gu, X., and Li, S. Clip as rnn: Segment countless visual concepts without training endeavor. arXiv preprint arXiv:2312.07661, 2023.

Wang, Z., Lu, Y., Li, Q., Tao, X., Guo, Y., Gong, M., and Liu, T. Cris: Clip-driven referring image segmentation. In CVPR, pp. 11686–11695, 2022.

Xian, Y., Choudhury, S., He, Y., Schiele, B., and Akata, Z. Semantic projection network for zero- and few-label semantic segmentation. In CVPR, pp. 8256–8265, 2019.

Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J. M., and Luo, P. Segformer: Simple and efficient design for semantic segmentation with transformers. In NeurIPS, pp. 12077–12090, 2021.

Xu, M., Zhang, Z., Wei, F., Lin, Y., Cao, Y., Hu, H., and Bai, X. A simple baseline for zero-shot semantic segmentation with pre-trained vision-language model. In ECCV, pp. 736–753, 2022.

Xu, W., Xu, R., Wang, C., Xu, S., Guo, L., Zhang, M., and Zhang, X. Spectral prompt tuning: Unveiling unseen classes for zero-shot semantic segmentation. In AAAI, pp. 6369–6377, 2024.

Zeng, Y., Zhang, X., and Li, H. Multi-grained vision language pre-training: Aligning texts with visual concepts. In ICML, pp. 25994–26009, 2022.

Zhang, B., Tian, Z., Tang, Q., Chu, X., Wei, X., Shen, C., et al. Segvit: Semantic segmentation with plain vision transformers. In NeurIPS, pp. 4971–4982, 2022.

Zhao, H., Shi, J., Qi, X., Wang, X., and Jia, J. Pyramid scene parsing network. In CVPR, 2017.

Zheng, S., Lu, J., Zhao, H., Zhu, X., Luo, Z., Wang, Y., Fu, Y., Feng, J., Xiang, T., Torr, P. H., et al. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In CVPR, pp. 6881–6890, 2021.

Zhou, C., Loy, C. C., and Dai, B. Extract free dense labels from clip. In ECCV, pp. 696–712, 2022a.

Zhou, K., Yang, J., Loy, C. C., and Liu, Z. Learning to prompt for vision-language models. IJCV, 130 (9):2337–2348, 2022b.

Zhou, Z., Lei, Y., Zhang, B., Liu, L., and Liu, Y. Zegclip: Towards adapting clip for zero-shot semantic segmentation. In CVPR, pp. 11175–11185, 2023.

Zhu, C. and Chen, L. A survey on open-vocabulary detection and segmentation: Past, present, and future. arXiv preprint arXiv:2307.09220, 2023.