开放词汇分割

张瑞程 (22354189)

2025 年 1 月 广东 · 深圳

摘 要: 语义分割是计算机视觉领域的一项重要任务，然而，传统方法由于受限于封闭语义集和高昂的标注成本，难以有效扩展至未见类别。开放词汇语义分割通过引入自然语言描述，实现了对图像中任意类别的像素级分割，因而成为近年来研究的热点。本文深入分析了单阶段与双阶段框架在开放词汇语义分割中的设计与实现，重点探讨了 SED 和 OpenSeg 模型在图像标题数据利用、类别早期拒绝策略及掩码提示微调等方面的创新，以及它们对视觉-语言数据的不同处理方式。最后，本文总结了开放词汇语义分割的前沿研究方向，包括弱标注训练、多任务学习以及生成式扩散模型的潜在应用。通过梳理现有技术与研究趋势，本文对开放词汇语义分割提供了自己的思考和见解。

关键词: 开放词汇分割，视觉语言模型，自然语言监督，语义扩展，多任务学习。

1 引言

语义分割是一个基本的计算机视觉任务之一，旨在解析图像中每个像素的语义类别。传统的语义分割方法假设语义类别是封闭集，在推理过程中难以识别看不见的语义类别，这使其在现实应用中具有极大局限。一方面，类注释的匮乏使得模型难以全面、精准地学习各类别特征；另一方面，封闭集类定义极大地限制了模型对新类别、新概念的拓展能力，使其在面对超出训练集范畴的未知类别时显得力不从心。此外，高昂的标记成本不仅加重了研究负担，还阻碍了技术的快速迭代与广泛应用。举例而言，作为语义分割算法基准测试常用数据集的 COCO [1]，其涵盖的类别仅为 80 个，但现实中的自然场景图像复杂多样，所包含的物体类型往往远超这一数目，这无疑为基于传统方法的深度模型应用增添了重重困难。

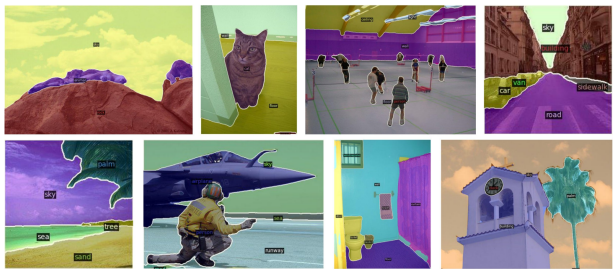

图 1: 开放词汇分割展示。上图中白色的类别标注为训练数据集中真实存在的标签，彩色的类别标注为通过语义理解泛化生成的标签。

之前一些解决方案采用零样本学习 (ZSL) 的方式扩展分割器。这类方法旨在打破传统分割模型的局限，使模型能够将有注释的可见对象类知识拓展至其他未见类别。具体而言，在训练阶段，模型仅依赖可见对象类的注释信息，严格杜绝使用未见类的注释。多数 ZSL 方法借助词嵌入投影技术构建针对未见类的分类器，试图利用预定义的语义关系来推断新类别的归属。然而，ZSL 方法存在着不容忽视的缺陷。由于在学习过程中缺乏对实际未见对象的实例学习，模型只能将这些未知对象在训练期间一概视作背景处理，这就导致在推理阶段，模型仅仅依据预先设定的词嵌入来识别新类，无法深入挖掘未见类的视觉特性以及它们与其他类别之间的内在关联。这导致 ZSL 方法在面对全新类别时，往往难以给出令人满意的分割结果，其应用范围与精度受到了极大的限制。

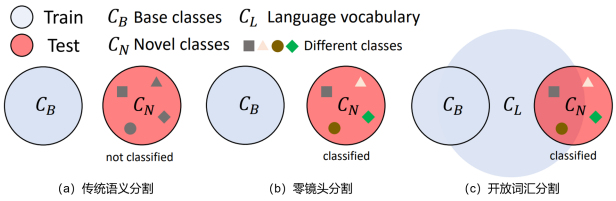

图 2: 传统语义分割、零样本分割和开放词汇分割之间的概念比较。不同的形状代表不同的新类别，颜色表示新对象的掩码和类别预测。(a) 在传统语义分割中，模型只需要识别新类并将它们标记为 “未知”。(b)，在零样本设置中，模型必须将未知类成功分割并分类为特定类别。(c)，在开放词汇设置中，该模型可以借助大型语言模型词汇知识 C_L 对训练中从未见过的对象进行分割和分类。

近年来，开放词汇语义分割作为一项新兴且极具潜力的研究方向崭露头角。该方法创新性地引入自然语言描述，旨在实现对图像中任意类别物体的精准像素级分割，无论这些类别是否在训练集中出现过。与零样本学习有相似之处，开放词汇语义分割模型同样基于基类的学习与新类的推理展开训练，但二者的关键区别在于，开放词汇分割充分利用了与视觉相关的语言词汇数据，如丰富多样的图像标题，将其作为开放词汇设置下的辅助监督信息（如图2）所示。之所以采用语言数据作为辅助弱监督，主要基于以下两方面的考量：

1. 从成本效益角度来看，语言数据具有显著优势。相较于传统的掩码注释，图像标题等语言数据来源广泛，易于收集，能够以较小的代价为模型训练提供丰富的

[1]COCO 数据集是一个大规模的、标注丰富的图像数据集，包含 80 个常见物体类别的边界框、实例分割轮廓、人体关键点以及自然语言描述字幕等标注信息，广泛应用于对象检测、实例分割、关键点检测和图像字幕生成等视觉任务。

信息支撑，有效缓解了标记成本过高带来的压力。

2. 语言数据蕴含着巨大的语义拓展潜力。它能够提供极为宽泛的词汇量，远远超越了预定义的基本类别范畴。在图像标题中，不仅包含封闭集中的物体名称，还会涉及新类别对象、各种属性描述以及动态动作等信息。通过在训练过程中巧妙融合这些字幕信息，模型得以接触到更为丰富多元的语义表达，从而极大地提升了自身的可扩展性与通用性。

随着大规模图像-文本对数据的不断涌现，视觉语言模型（VLM）在各类视觉任务中展现出卓越的性能。VLM 的核心优势在于将图像与语言词汇精确对齐至同一特征空间，有效跨越了视觉与语言之间的数据鸿沟，使得模型能够在两种模态之间灵活转换并协同学习。许多开放词汇分割方法敏锐地捕捉到这一特性，巧妙利用 VLM 中学习到的视觉-文本对齐机制，成功突破封闭集与开放集场景的限制，为实际应用中的语义分割任务提供了强有力的技术支撑。

通过对前沿研究的深入分析，当前开放词汇语义分割的方法主要分为两类：一种是单阶段（one-stage）策略，该方法直接结合视觉编码器与文本编码器对图像-文本对进行联合训练，通过紧密的信息融合，让模型在学习过程中同时捕捉视觉与语义特征，从而提升分割和理解能力；另一种是双阶段（two-stage）策略，该方法首先利用现有技术生成与类别无关的掩码提议，随后通过预训练的视觉语言模型对这些掩码进行分类。单阶段策略注重构建紧密耦合的编码与解码架构，强化视觉与文本信息的深度交互；而双阶段策略则利用已有分割模型的对象提取能力，简化场景理解过程。尽管两种策略的重点有所不同，但它们的共同目标都是实现视觉特征与语义特征的高效对齐。

本文将围绕这两条核心技术路线，对开放词汇语义分割的研究进展进行系统剖析与全面介绍，并在总结当前研究方向的基础上，提出相关思考与展望。

2 One - stage 框架

在开放词汇语义分割领域，基于单阶段框架的方法直接对单一视觉 - 语言模型进行拓展，以此现高效的开放词汇分割任务。具体而言，多项前沿研究尝试对图像编码器的结构进行优化调整，去除其最后一层的池化操作，改为生成高分辨率的像素级特征图，再经过一系列精心设计的解码策略输出最终的分割结果。在众多基于单阶段框架的方法中，我选择了一项极具代表性的作品 SED [1] 进行深入分析。

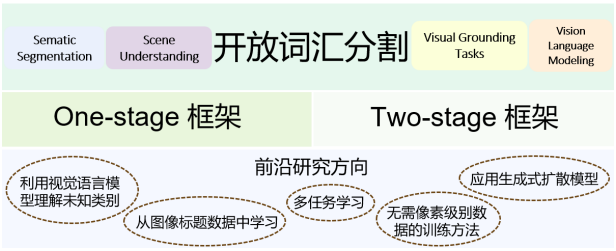

图 3: 调研总览。

2.1 SED

SED (A **S**imple **E**ncoder-**D**ecoder for Open-Vocabulary Semantic Segmentation) 为 one-stage 开放语义分割的典范之作，它首先使用层次图像编码器和文本编码器分别提取视觉特征和文本嵌入，通过计算视觉特征图和文本嵌入之间的余弦相似度获得像素级的 cost map（其可以理解为低分变率的初步分割结果图）。然后使用一个渐进融合解码器在空间和类别维度不断进行特征增强，优化 cost map 的分割精度和类别预测。同时，为了提高推理速度，作者引入了类别早期拒绝方案，在早期解码器层拒置信度低的类别，从而减小计算开销。SED 的创新在于其精心设计地编码器-解码器架构，包括基于层次编码器的 cost map 生成和逐渐融合的 cost map 解码，以及通过类别早期拒绝方案进行的推理加速。

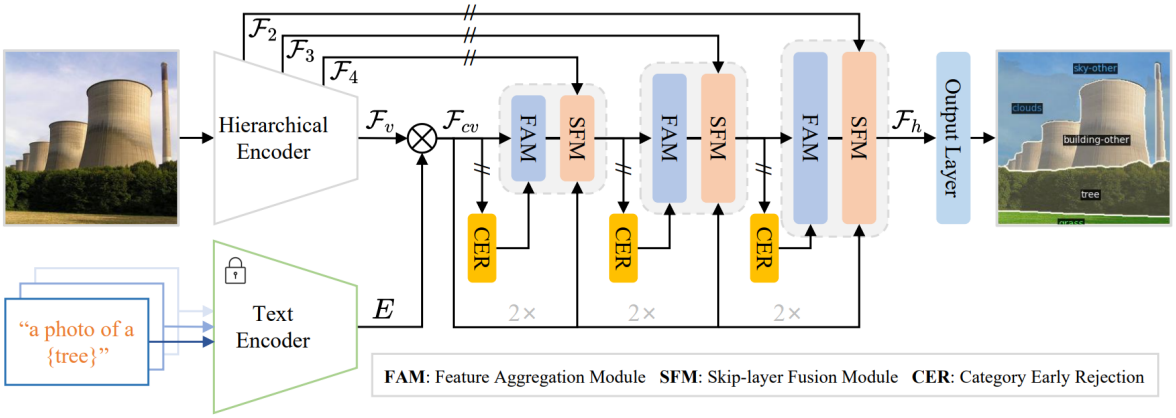

图 4: SED 的结构。SED 首先使用分层视觉编码器（可学习）和文本编码器（冻结）来生成像素级图像-文本 cost map。然后，由一个渐进融合解码器来增强并融合视觉编码器和 cost map 的不同特征，最后输出带有语义的分割掩码。同时，类别早期拒绝 (CER) 在不牺牲性能的情况下加速推理速度。

2.1.1 Cost map 的生成和解码

视觉方面，给定一个输入图像 $I \in \mathbb{R}^{H\times W\times 3}$，SED 首先利用一个分层编码器来提取多尺度特征图，记为 F_2, F_3, F_4, F_5。这些特征图相对于输入尺寸的步幅为 4, 8, 16, 32 像素。为了对齐输出的视觉特征和文本嵌入，SED 在最后一个特征图 F_5 上附加一个 MLP 层，以获得对齐的视觉特征图 $F_v \in \mathbb{R}^{H_v\times W_v\times D_t}$，其中 D_t 等于文本嵌入的特征维度，H_v 是 $H/32$，W_v 是 $W/32$。**语言**方面，给定任意一组类别名称 $\{T_1, ..., T_N\}$，作者使用经典的提示模板策略来生成关于类别名称 T_n 的不同文本描述 $S(n) \in \mathbb{R}^P$，例如 “一张 $\{T_n\}$ 的照片，许多 $\{T_n\}$ 的照片，...”。N 代表类别的总数，P 是每个类别的模板数量。通过将 $S(n)$ 输入到文本编码器，SED 获得文本嵌入，记为 $E = \{E_1, ..., E_N\} \in \mathbb{R}^{N\times P\times D_t}$。

通过计算视觉特征图 F_v 和文本嵌入 E 之间的余弦相似度，SED 获得像素级成本图 F_{cv} 可以表示为

$$F_{cv}(i,j,n,p) = \frac{F_v(i,j)\cdot E(n,p)}{\|F_v(i,j)\|\|E(n,p)\|}$$

，其中 i, j 表示二维空间位置，n 表示文本嵌入的索引，p 表示模板的索引。因此，初始成本图 F_{cv} 的尺寸为 $H_v \times W_v \times N \times P$。初始成本图通过一个卷积层来生成解码器的输入特征图 $F^1_{dec} \in \mathbb{R}^{H_v\times W_v\times N\times D}$，这个向量是后续分割的关键，它为 N 个可能存在的类别[2]分别构建一张特征图，每张特征图的大小为 $H_v \times W_v$，具有 D 个通道。

之后的逐渐融合解码过程则是通过一系列特征增强、混合、跨层链接的方法不断从 cost map 中精细化分割掩码和类别预测，此部分的创新点比较有限，不再详述。

2.1.2 类别早期拒绝策略

由上述分析可知，解码器的输入特征图为 $F^1_{dec} \in \mathbb{R}^{H_v\times W_v\times N\times D}$，$N$ 为全集中所有可能存在的类别，因此，解码过程的计算成本和显存占用与全部语义类别的数量 N 成正比。实际上，大多数图像只包含少数几个语义类别。为了提高推理速度，SED 引入了类别早期拒绝策略（Category Early Rejection，缩写为 CER），在早期解码器层识别存在的类别并拒绝不存在的类别。被拒绝类别对应的特征图将从当前解码器层中移除，后续解码器层仅考虑保留的类别。

CER 在解码器每个层后添加辅助卷积分支分别预测分割图，并通过真实掩码进行监督。为了避免对模型训练产生负面影响，禁止分支的梯度反向传播到解码器。在推理过程中，CER 在分割图上采用 top-k 策略来预测存在的语义类别。具体而言，我们选择每个像素最大响应的前 k 个类别生成一个特征图子集，将其输入到下一个解码器层。

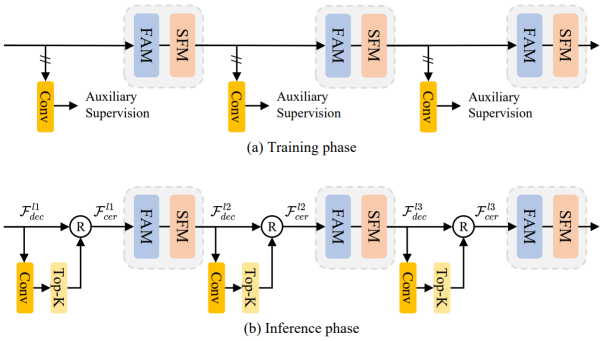

图 5: 类别早期拒绝策略（CER）的结构。(a) 在训练过程中，CER 在每个解码器层之后附加一个辅助卷积来预测由 ground-truth 监督的分割图。(b) 在推理过程中，CER 采用 top-k 策略来预测现有类别并为下一个解码器层拒绝的不存在的类别。

图5展示了 CER 的推理过程。首先从 F^{l1}_{dec} 预测出分割图，并采用 top-k 策略选择 N_{l1} 个类别。然后，CER 移除非选中类别的特征图，并生成新的特征图 $F^{l1}_{cer} \in \mathbb{R}^{H_v\times W_v\times N_{l1}\times D}$。生成的特征图 F^{l1}_{cer} 被馈送到下一个解码器层。这样，大多数不存在的类别在早期层被拒绝，从而提高了解码器的推理速度。

3 Two - stage 框架

受启发于视觉语言模型，例如 CLIP，优越的开放词汇分类能力，一些工作提出使用预先训练的视觉语言模型进行掩码分类的分割方式。此框架通过两阶段方法分解开放分割任务：他们首先生成与类别无关的掩码提议，然后利用预训练的 CLIP 执行开放词汇分类（如图6所示）。他们的成功依赖于两个假设：(1) 该模型可以生成与类别无关的掩码提议（2）预训练的 CLIP 可以将其分类性能转移到掩码图像提议中。

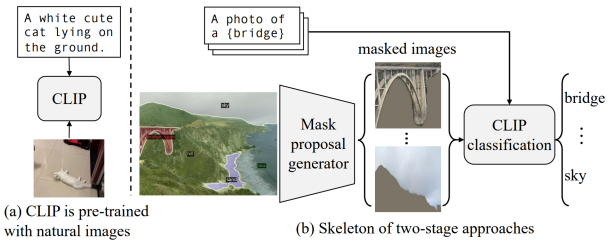

图 6: Two - stage 开放词汇语义分割方法首先生成与类别无关的掩码提议，然后利用预训练的 CLIP 进行开放词汇分类。CLIP 模型的输入为被裁剪的 mask 图像，这些图像与自然图像有很大的域差距。

这种分阶段的方法通过将开放词汇的图像分割任务分解为两个子任务——实例分割和类别预测，实现了任务的简化。这种策略直观且高效，但要达到理想的效果，需要满足以下三个条件：首先，分割器（视觉模块）必须具备高性能；其次，分类器（视觉-语言模块）也需表现出色；最后，两者之间需要有良好的协同作用。随着技术的进步，前两个

[2]这里的 N 为任务定义的全集，包括训练中可获得的类别和推理中不存在的类别。显然，一张实际图片中存在物体种类数一定小于 N，跟据我的理解，在这些不存在类别对应的特征图接近全 0 向量。这是因为在计算余弦相似度时，等效为视觉特征作为“键”去查询文本空间中的“值”，只有当查询匹配时才会得到强度较高的特征向量图。

条件已基本得到满足。然而，如何利用视觉-语言模型来有效预测未知的掩码类别，仍然是一个巨大的挑战。主要难点可以概括为：

1. 两阶段框架中，两个独立的网络分别负责遮罩生成和分类任务，这不仅降低了计算效率，也增加了训练的复杂性。

2. 预训练的视觉-语言模型通常在自然图像上进行训练，这与掩码图像的领域存在显著差异，影响了其在分类任务上的表现。

3. 为了适应特定任务，往往需要对预训练的视觉-语言模型中的视觉-语言特征进行重新调整。如何实现这一过程的高效微调，是一个充满挑战的问题。

Algorithm 1: Two - stage 算法框架

Input: 自然图像: $p(\mathcal{T})$.
Output: 分割掩码图像 $\mathcal{T}_{mc}$.

使用训练好的分割器将输入图像 $\mathcal{T}_)$ 分割为 N 个无类别掩码;
while $N > 0$ **do**
  从 $\mathcal{T}_i$ 中去提取第 n 个掩码 t_n;
  把原图中除 t_n 之外的区域全部设为 0 得到 mask 图像 t_{mn};
  使用训练好的视觉语言模型（如 CLIP）对 t_{mn} 进行类别预测;
  更新 N: $N = N - 1$;
end
将所有具有类别的 mask 图像 t_{mn} 合并为完整分割图像 $\mathcal{T}_{mc}$.

总而言之，这种分阶段策略虽然在理论上简化了问题，但在实践中，如何优化网络间的协作以及如何调整模型以适应特定任务，仍然是研究者们需要克服的关键难题。OpenSeg [2] 作为一个经典的 two-stage 开放词汇框架对解决上述问题提出了有效的方案。

3.1 OpenSeg

OpenSeg 认为开放词汇分割领域所面临的挑战主要在于 CLIP 模型在对掩码区域进行分类时存在一定的局限性。为了克服这一瓶颈，他们建议在一系列掩码图像区域及其对应的文本描述上对 CLIP 模型进行微调。为了收集训练数据，作者们深入挖掘了现有的图像标题数据集，如 COCO Captions。他们利用 CLIP 模型的能力，将掩码图像区域与图像标题中的名词进行精确匹配。与具有固定类（例如 COCO-Stuff）的更精确和手动注释的分割标签相比，作者发现他们的嘈杂但多样化的数据集可以更好地保留 CLIP 的泛化能力。除了微调整个模型外，OpenSeg 还使用了一种“掩码提示调优”的方法来利用掩码图像中的“空白”区域，在最大限度保留 CLIP 泛化性的前提下，通过极小参数的调整成功实现任务适应。

3.1.1 从图像标题中收集掩码-类别对

在开放词汇分割任务中，CLIP 模型面临的一个主要挑战是如何将其在自然图像上训练得到的视觉能力，迁移到对掩码图像进行类别预测的新场景中。普通 CLIP 的视觉训练数据为自然图像，而在开放词汇分割中则需对掩码图像进行类别预测，这种跨域泛化的需求限制了 CLIP 在开放词汇分割中的性能。为了解决这一问题，研究者们提出了一种创新的方法：构建一个专门针对掩码-类别对的数据集，并对现有的 CLIP 模型进行微调。

与传统的分割标签数据集相比，图像标题提供了更为丰富和广泛的信息，它们涵盖了更大的词汇范围。例如，在图7中，图像标题是“There are apple and orange and teapot.”。尽管“apple”和“orange”是 COCO-Stuff 中的有效类，但其他概念是无效的类并且被忽略（如 teapot）。基于这一观察，作者设计了一种自标记策略来提取掩码类别对。具体来说，给定图像，首先使用预训练的 MaskFormer 来提取掩码提议。同时，从相应的图像标题，使用现成的语言解析器提取所有名词，并将它们视为潜在的类。然后，使用 CLIP 将最匹配的掩码提议与每个类配对。通过这样的方式，作者收集了 1.3M 掩码类别对，其中包含 27K 个唯一名词。实验结果表明，这种虽然存在一定噪声但内容丰富多样的掩码类别数据集，在提升模型性能方面，明显优于传统的手动分割标签。

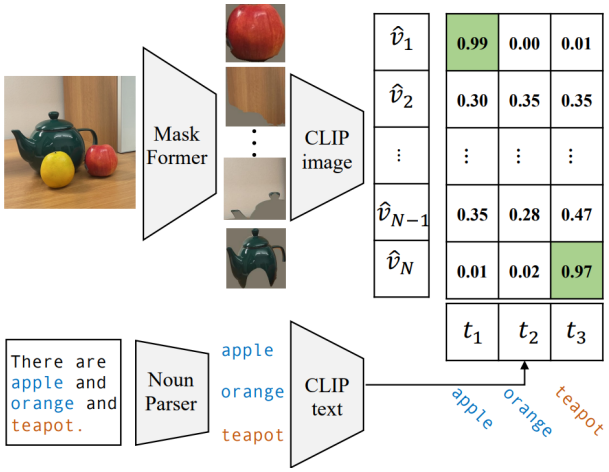

图 7: 对于上图的图像-标题对，只有“苹果”和“橙子”是 COCO 中的有效类别。但如果通过从字幕中提取名词，还可以得到一个新的“茶壶”类别。

3.1.2 掩码提示微调

在收集数据集之后，一个自然的问题是如何有效地微调 CLIP 模型。掩码图像与自然图像之间最显著的区别在于，掩码图像中的背景像素被设置为零，导致出现许多“空白区域”。当将遮罩图像输入到 CLIP 时，图像将被划分为不重

叠的块，然后进行标记化处理，这些空白区域随后会变成零标记。这会带来两个明显的问题：

1. 这些标记不包含有用的信息，造成了无必要的计算和显存开销；
2. 为模型带来 domain shift（因为这样的标记在自然图像中不存在），并导致性能下降。

为了缓解这个问题，作者提出了一种基于视觉提示调整的技术：掩码提示微调。具体来说，当输入到 CLIP 时，掩码图像中的 0 token 被一种可学习的掩码提示向量（learnable prompt）所替代，非掩码位置仍保持为 visual tokens。在微调 CLIP 阶段，只调整 learnable prompt, 而固定 CLIP 的参数。这样的微调方式有多点好处：

1. 是针对分割任务而特定设计的；
2. 训练的参数量减少很多，训练效率得到有效提升；
3. 可以同时多个任务，对于每个任务只需要调整相应的 prompt 即可，而不需要调整 CLIP。更适合多任务的场景；
4. 实验证明，mask prompt tuning 对于性能的提升是有益的，也就是解决了刚刚提到的 domain shift 的问题；
5. 个人认为，这种方式固定 CLIP，调整 learnable prompt 的方式，对于保证 CLIP 的泛化性是友好的。

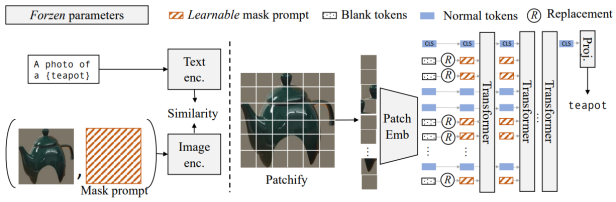

图 8: 掩码提示微调用可学习的掩码提示替换掩码补丁中的 0 token，可以在不改变 CLIP 权重的情况下使其适应掩码图像。

OpenSeg 项目在构建掩码-类别对以及实施掩码提示微调方面所做的工作，核心目标是在保持 CLIP 模型泛化能力的同时，实现对特定任务的适应性。这种方法不仅为两阶段开放词汇分割任务提供了一个高效的解决方案，而且为其他相关任务提供了宝贵的洞见。在考虑将 CLIP 应用于开放词汇（open-vocab）或其他开放领域（open-domain）的任务时，关键在于如何巧妙地调整模型，而不损害其固有的泛化特性。

尤其是掩码提示微调技术，对于希望利用 CLIP-based method 在特定类型的任务上调整之后面向一个 open-domain 的场景的工作，建议在调整过程中尽量减少对模型原始参数的改动。采用提示（prompt）调整的方法，可以作为一种更为温和且有效的策略。这种方法不仅有助于保护 CLIP 的泛化能力，而且能够减小训练成本，实现在保持泛化性的基础上的任务适应性。

4 前沿研究方向

基于 One-stage 和 Two-stage 框架，在开放词汇分割领域的研究工作不断涌现，提出了众多创新解决方案。

4.1 利用视觉语言模型理解未知类别

视觉语言模型（VLM）在大规模图片-文本对上进行训练，因而具备识别包括基础类别和新类别在内的各种类别的天然能力。在开放词汇分割领域，一个直观的策略是利用 VLM 的文本特征来替代传统的闭集分类器，使分割模型能够识别新类别。尽管 VLM 主要通过图像级别的对比学习进行预训练，缺乏处理像素级别任务（如语义分割）的能力，但当前许多研究正致力于调整和微调 VLM，以适应开放词汇分割任务。

4.2 从图像标题数据中学习

除了 VLM 在大规模数据上训练得到的分类性能，图像标题提供了另一种广泛存在且易于获取的数据类型。与预定义的类别不同，标题中描述的物体可能包含新类别，为模型提供了在训练过程中接触新类别弱标注的机会，这也是开放词汇与零样本学习的核心区别。多篇研究提出了不同的方法，以更好地利用图片标题信息，提取新类别，扩展模型在新类别上的识别能力。

4.3 无需像素级别数据的训练方法

在大多数开放词汇的研究中，尽管模型不需要新类别的像素级别标注进行训练，但仍需基础类别的像素级别标注，这增加了数据收集的难度。为了解决这一问题，有研究提出了仅使用图像标题进行训练的方法，仅依赖弱标注的标题，即可训练出能够识别广泛语义空间的检测器/分类器。例如，GroupViT[3] 提出了基于分组机制的分割方法，通过图像-标题的对比损失函数进行训练，摆脱了分割模型对 mask 标注的依赖。

4.4 多任务学习

场景理解任务包括图片/视频理解、目标检测、语义分割等。在开放词汇领域，这些任务之间如何相互促进，以及不同任务的数据集是否能够通过共同训练提升性能，是一个值得探究的问题。一些研究提出了综合性模型，通常采用 Transformer 架构，产生不同任务的输出，并能在多个任务的数据集上进行训练。例如，OpenSeed[4] 构建了一个框架，可以同时在开放词汇检测和分割数据集上训练，发现联合训练两个任务的数据能够提升检测和分割的性能。

4.5 应用生成式扩散模型

近期，生成式扩散模型（Generative Diffusion Models, DM）在图像生成领域取得了显著成就。如何利用 DM 提升开放词汇分割任务的性能，目前主要有两种方案。第一种方案 [5] 是利用 DM 生成逼真且高度差异化的图像，为新类别创建伪图片原型，并在推理时与输入图片进行比对，以确定预测类别。第二种方案 [6] 是提取 DM 中间层的文本-语义联系知识，用于分割模型的分类器，可能达到甚至超越视觉语言模型（VLM）的效果。

4.6 3D 开放词汇场景分割

与图像和视频相比，点云数据的注释成本更高，特别是对于密集预测任务。因此，关于词汇场景理解 3D 开放的研究更加紧迫。目前，3D 开放词汇场景理解的解决方案主要集中在设计投影函数，以此来更充分地利用 2D 视觉语言模型（VLM）的优势。未来，将 2D 模型的知识有效对齐到 3D 模型，开发出全新的解决方案，将成为该领域的重要发展方向。

这些前沿研究方向不仅展示了开放词汇分割领域的多样性和活力，而且为未来的研究提供了丰富的思路和可能性。

5 开放词汇分割面临的问题与挑战

尽管开放词汇语义分割近年来取得了显著进展，但仍然存在诸多问题和挑战有待解决。

- **数据质量与标注成本**：开放词汇语义分割任务需要大规模、高质量的图像-文本对数据，现有数据集中的掩码和类别标注需要大量人工参与，尤其对于复杂场景和新颖类别，标注的主观性和准确性难以保障。虽然社交媒体和网络提供了大量图像和文本，但其中包含的语义信息具有偏差且质量不均。
- **泛化能力的限制**：现有的视觉语言模型（VLM）在处理未见类别时的泛化能力仍显不足，模型在训练过程中未接触的类别可能难以通过简单的词汇投影方式实现有效分割，而且对于复杂语义或动态场景，模型缺乏有效的推理机制。
- **模型复杂度与实时性**：许多先进方法需要复杂的模型架构和大规模的计算资源，在处理大规模数据或高分辨率图像时的效率难以满足实际需求，限制了其在嵌入式设备或实时场景中的应用；
- **缺乏统一评估标准**：目前，开放词汇语义分割领域缺乏统一的评估基准和标准。不同方法在不同数据集上报告的指标难以直接比较，阻碍了技术进步的量化分析。

针对上述挑战，未来研究需从以下方向展开：

- 优化数据采集和标注策略，降低数据成本并提升数据质量；
- 提升模型的跨域泛化能力，通过多模态融合和预训练优化，增强对新类别的理解能力；
- 简化模型架构，提升计算效率，增强其实时性和应用广度；
- 建立统一的评估标准和基准，推动技术进步的公平对比和量化分析。

6 结论

开放词汇语义分割通过结合自然语言描述与视觉信息，为传统语义分割的局限性提供了有效解决方案。单阶段框架通过视觉与语言特征的高效融合，显著提升了模型的实时性与性能；双阶段框架则借助预训练的视觉语言模型，实现了复杂场景下的精准分类。尽管当前方法在模型效率与域适应性方面仍有改进空间，但利用图像标题和生成式扩散模型为进一步提升模型的泛化能力开辟了新路径。此外，多任务学习的探索表明，开放词汇语义分割有望成为未来视觉任务中的重要组成部分，为构建更加通用的人工智能系统提供基础支持。

参考文献

[1] Bin Xie, Jiale Cao, Jin Xie, FahadShahbaz Khan, and Yanwei Pang. Sed: A simple encoder-decoder for open-vocabulary semantic segmentation. Nov 2023.

[2] Feng Liang, Bichen Wu, Xiaoliang Dai, Kunpeng Li, Yinan Zhao, Hang Zhang, Peizhao Zhang, Peter Vajda, and Diana Marculescu. Open-vocabulary semantic segmentation with mask-adapted clip. Oct 2022.

[3] Jiarui Xu, ShaliniDe Mello, Sifei Liu, Wonmin Byeon, Thomas Breuel, Jan Kautz, and Xiaolong Wang. Groupvit: Semantic segmentation emerges from text supervision.

[4] Hao Zhang, Feng Li, Xueyan Zou, Shilong Liu, Chunyuan Li, Jianfeng Gao, Jianwei Yang, and Lei Zhang. A simple framework for open-vocabulary segmentation and detection. Mar 2023.

[5] Jiarui Xu, Sifei Liu, Arash Vahdat, Wonmin Byeon, Xiaolong Wang, ShaliniDe Mello, UcSan Diego, and Nvidia Nvidia. Open-vocabulary panoptic segmentation with text-to-image diffusion models.

[6] Laurynas Karazija, Iro Laina, Andrea Vedaldi, and Christian Rupprecht. Diffusion models for zero-shot open-vocabulary segmentation.