SED: A Simple Encoder-Decoder for Open-Vocabulary Semantic Segmentation

摘要

开放式词汇语义分割致力于将像素从一组开放的类别中区分为不同的语义组。现有的方法大多探索利用预训练的视觉语言模型，其中关键是采用图像级模型进行像素级分割任务。本文提出了一种用于开放词汇语义分割的简单编码器-解码器 SED，它包括一个基于分层编码器的成本图生成和一个具有类别早期拒绝的渐进融合解码器。基于分层编码器的成本图生成采用分层骨干而不是普通变换器来预测像素级图像-文本成本图。与普通变换器相比，分层骨干网更好地捕获了局部空间信息，并且相对于输入大小具有线性计算复杂度。我们的渐进融合解码器采用自上而下的结构，将成本图和不同骨干层的特征图结合起来进行分割。为了加快推理速度，我们在解码器中引入了一种类别早期拒绝方案，该方案在解码器的早期层拒绝了许多不存在的类别，在不降低精度的情况下最多可加速 4.7 倍。在多个开放式词汇语义切分数据集上进行了实验，证明了我们的 SED 方法的有效性。当使用 ConvNeXt-B 时，我们的 SED 方法在 ADE20K 上实现了 31.6% 的 mIoU 评分，在单个 A6000 上每幅图像上有 150 个类别，每个类别为 82 毫秒（ms）。

关键词：开放式词汇；语义分割; 分层编码器；渐进融合解码器；类别早期拒绝；

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 [4, 20, 29] 假设语义类别是闭集，在推理过程中难以识别未见的语义类别。为了决解这个问题，现在已经有工作探索了开放词汇语义分割 [3,28]，旨在分割属于任意语义类别的像素。

最近，视觉语言模型，如 CLIP [22] 和 ALIGN [15]，从数以百万的图像-文本配对数据中学习对齐的图像-文本特征表示。预训练的视觉语言模型在识别开放词汇类别方面表现出优异的泛化能力。这激发了一系列研究工作，探索使用视觉语言模型进行开放词汇语义分割 [10]。最初，研究工作主要采用两阶段框架 [18]，直接适应开放词汇分词的视觉语言模型。具体来说，他们首先生成类别不可知论的掩码提案，然后采用预训练的视觉语言模型将这些提案分类到不同的类别中。然而，这种两阶段框架使用两个独立的网络进行掩码生成和分类，从而阻碍了计算效率。此外，它没有充分利用上下文信息。

与上述两阶段方法不同，基于单阶段框架的方法直接扩展了单视觉语言模型进行开放式分词。几种方法去除了图像编码器最后一层的池化操作，并生成像素级特征图进行分割。例

如，MaskCLIP [32] 删除了 CLIP 图像编码器最后一层的全局池，并使用值嵌入和文本嵌入来直接预测像素级分割图。CAT-Seg [9] 首先生成像素级图像文本成本图，然后通过空间和类聚合对成本图进行细化。虽然与两阶段方法相比，这些方法取得了良好的性能，但我们注意到它们存在以下局限性。首先，MaskCLIP 和 CAT-Seg 都采用普通变压器 ViT [12] 作为骨干，这会受到弱局部空间信息和低分辨率输入大小的影响。为了解决这些问题，CAT-Seg 引入了一个额外的网络来提供空间信息。然而，这会带来额外的计算成本。其次，CAT-Seg 的计算成本随着开放词汇类数量的增加而显著增加。

为了解决上述问题，本工作提出了一种简单而有效的编码器-解码器方法，称为 SED，用于开放词汇语义分割。本工作提出的 SED 包括一个基于分层编码器的成本图生成和一个具有类别早期拒绝的渐进融合解码器。基于分层编码器的成本图生成采用分层骨干而不是普通变换器来预测像素级图像-文本成本图。与普通变换器相比，分层骨干网更好地保留了不同层次的空间信息，并且相对于输入大小具有线性计算复杂度。渐进融合解码器逐渐将来自不同层次骨干网的特征图和成本图结合起来进行分割预测。为了提高推理速度，本工作在解码器中设计了一种类别早期拒绝方案，该方案有效地预测了现有类别，并在解码器的早期层拒绝了不存在的类别。在多个开放式词汇语义切分数据集上进行了综合实验，揭示了所提出的贡献在准确性和效率方面的优点。

2 相关工作

2.1 语义分割

传统的语义分割方法主要包括基于 FCN 的方法和基于 transformer 的方法。最初，研究人员专注于基于 FCN 的方法。Long 等人 [20] 提出了最早的全卷积网络之一，该网络融合了深层和浅层特征以改进分割。随后，许多基于 FCN 的变体被提出。一些方法利用空间金字塔网络 [4] 和编码器-解码器结构 [1] 来提取局部上下文信息。一些方法 [13] 利用注意力模块来提取非局部上下文信息。最近，研究人员专注于开发基于 transformer 的方法。一些方法 [6] 使用 transformer 作为骨干来提取深度特征，而一些方法 [7] 将 transformer 视为分段解码器。

2.2 视觉语言模型

视觉语言模型旨在学习图像表示和文本嵌入之间的联系。最初，研究人员基于预训练的视觉和语言模型开发了视觉语言模型 [5, 25]，并探索了在不同的下游任务中使用图像-文本对对其进行联合微调。相比之下，CLIP [22] 从网站收集大规模的图像-文本配对数据，并通过语言监督从头开始学习视觉特征。学习的大规模数据 CLIP 在不同的零样本任务上具有优异的性能。ALIGN [15] 不是使用经过清理的图像文本配对数据，而是从嘈杂的图像文本数据集中学习视觉语言表示。为了实现这一目标，ALIGN 采用了具有对比损耗的双编码器结构，在下游任务上实现了良好的 zero-shot 性能。最近，Cherti 等人 [8] 对比语言视觉学习进行了深入分析。Schuhmann 等人 [24] 构建了一个十亿图像文本配对数据集，用于训练大规模视觉语言模型。

2.3 开放词汇语义分割

开放式词汇语义分割旨在分割任意类别。最初，研究人员 [2,26,30] 探索了通过学习到的特征映射将视觉特征与预训练的文本嵌入对齐。随着大规模视觉语言模型 CLIP 的成功 [22]，研究人员开始探索使用 CLIP 进行开放词汇语义分割。一些方法 [16,27] 采用两阶段框架，首先预测与类无关的掩码建议，然后将这些建议分为不同的类别。为了提高第二阶段的分类性能，OVSeg [18] 对蒙版图像及其文本注释上的 CLIP 模型进行了微调。Ding 等人 [11] 将掩码令牌与预训练的 CLIP 模型集成在一起，用于掩码细化和分类。ODISE [30] 采用文本到图像扩散模型来生成掩模建议并进行分类。为了提高开放词汇表的性能，ODISE [30] 进一步使用从预训练的 CLIP 裁剪的特征进行掩码分类。

相比之下，一些方法采用单阶段框架。LSeg [17] 在预训练的 CLIP 文本嵌入的指导下学习像素级图像特征。MaskCLIP [32] 去除了自注意力池层以生成像素级特征图，并采用文本嵌入来预测最终的分割图。SAN [28] 沿着冻结的 CLIP 模型引入了一个侧适配器网络，以执行掩码预测和分类。FC-CLIP [31] 采用冻结卷积 CLIP 来预测类无关掩码，并采用掩码池特征进行分类。CAT-Seg [9] 生成像素级成本图，并细化成本图以进行分割预测。我们提出的方法受到 CAT-Seg 的启发，即通过成本图微调图像编码器不会降低其开放词汇能力，但存在显著差异：(1) 我们的 SED 是一个更简单的框架，没有额外的骨干，具有更好的性能和更快的推理速度。(2) 我们的 SED 采用分层图像编码器来生成成本图并执行跳跃层融合，这可以显著提高性能，并且相对于输入大小具有线性计算成本。(3) 在解码器中，我们引入了一种简单的大核操作和渐进融合进行特征聚合，并设计了一种在不牺牲性能的情况下加速的类别早期拒绝策略。

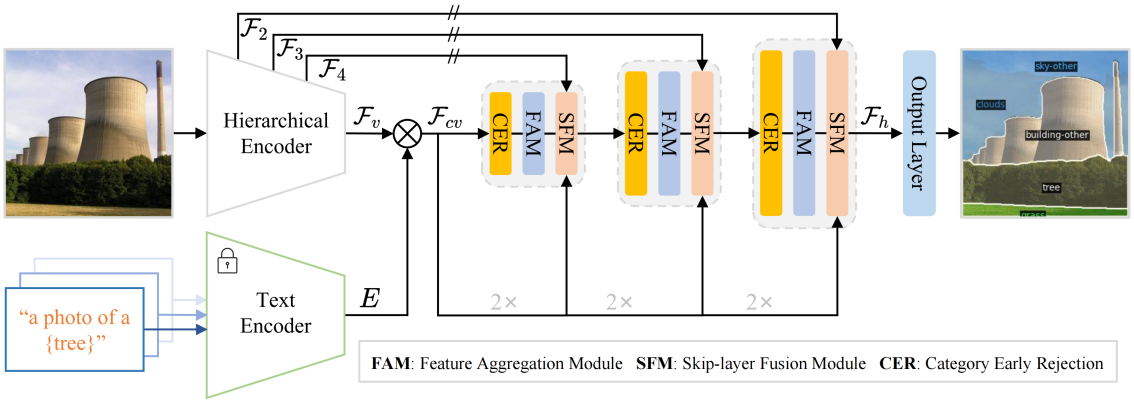

图 1. SED 的总体架构图

3 本文方法

在本节中，我们描述了我们提出的用于开放词汇语义分割的编码器-解码器，称为 SED。图 1显示了我们提出的 SED 的总体架构，它包括两个主要组件：基于分层编码器的成本图生成和具有类别早期拒绝的渐进融合解码器。在我们基于分层编码器的成本图生成中，我们采用分层图像编码器和文本编码器为解码器生成像素级图像文本成本图 Fcv 和分层特征图 F2、F3、F4。我们的渐进融合解码器采用特征聚合模块（FAM）和跳过层融合模块（SFM）来逐步组合像素级成本图 Fcv 和分层特征图 F2、F3、F4，以生成高分辨率特征图 Fh。基于 Fh，

我们采用输出层来预测不同类别的分割图。此外，在解码器中使用类别早期拒绝（CER）策略来早期拒绝不存在的类别，以提高推理速度。

3.1 基于分层编码器的成本图

基于分层编码器的成本图生成（HECG）采用视觉语言模型 CLIP [22] 生成像素级图像-文本成本图。具体来说，我们首先采用分层图像编码器和文本编码器分别提取视觉特征和文本嵌入。然后，我们计算这两个特征之间的像素级成本图。现有的方法，如 MaskCLIP [32] 和 CAT-Seg [9]，采用普通变换器作为图像编码器来生成像素级成本图。如前所述，普通变压器的局部空间信息相对较弱，并且相对于输入大小具有二次复杂性。为了解决这些问题，我们建议使用分层骨干网作为成本图生成的图像编码器。分层编码器可以更好地捕获局部信息，并且相对于输入大小具有线性复杂度。成本图生成描述如下。

给定输入图像 $I \in \mathbb{R}^{H\times W\times 3}$，我们首先利用分层编码器 ConvNeXt [19, 23] 提取多尺度特征图，表示为 $\mathcal{F}_2, \mathcal{F}_3, \mathcal{F}_4, \mathcal{F}_5$。这些特征图相对于输入大小具有 4、8、16、32 像素的步长。为了对齐输出视觉特征和文本嵌入，在最后一个特征图 F5 处附加 MLP 层，以获得对齐的视觉特征图 $\mathcal{F}_v \in \mathbb{R}^{H_v\times W_v\times D_t}$，其中 D_t 等于文本嵌入的特征维度，H_v 为 $H/32$，W_v 为 $W/32$。给定一组任意的类别名称 $\{T_1, .., T_N\}$，我们使用提示模板策略 [9, 14] 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，一张许多 $\{T_N\}$ 的照片，…”。N 表示类别的总数，P 是每个类别的模板数量。通过将 S（n）馈送到文本编码器，我们得到文本嵌入，表示为 $E = \{E_1 .. E_N\} \in \mathbb{R}^{N\times P\times D_t}$。通过计算视觉特征图 $\mathcal{F}_v$ 和文本嵌入 E 之间的余弦相似度 [22]，我们得到像素级成本图 $\mathcal{F}_{cv}$，如下所示

$$\mathcal{F}_{cv}(i,j,n,p) = \frac{\mathcal{F}_v(i,j)\cdot E(n,p)}{||\mathcal{F}_v(i,j)||||E(n,p)||} \tag{1}$$

其中 i，j 表示 2D 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 $\mathcal{F}_{cv}$ 的大小为 $H_v\times W_v\times N\times D$。初始成本图通过卷积层生成解码器的输入特征图 $\mathcal{F}^{l1}_{dec} \in \mathbb{R}^{H_v\times W_v\times N\times D}$。

3.2 渐进融合解码器

语义分割极大地受益于高分辨率特征图。然而，编码器生成的成本图 $\mathcal{F}_{cv}$ 具有相对较低的分辨率和较高的噪声。因此，直接使用成本图进行预测，不利于生成高质量的分割图。为了解决这个问题，我们提出了一种渐进融合解码器（GFD）。GFD 通过将两个模块（包括特征聚合模块（FAM）和滑雪运动员融合模块（SFM））级联到多层中，逐渐生成高分辨率特征图 $\mathcal{F}_h$。FAM 旨在模拟局部区域和不同类别之间的关系，而 SFM 旨在使用分层编码器的浅特征来增强特征图的局部细节。

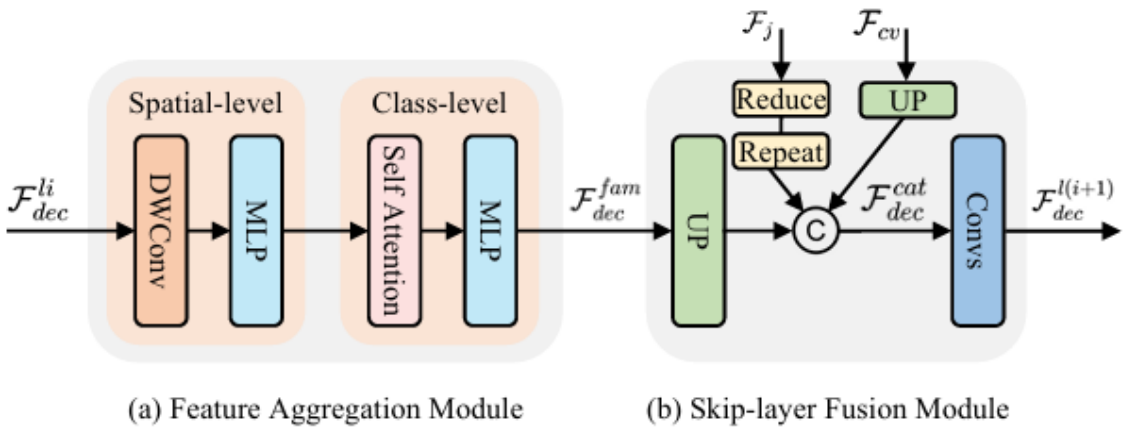

图 2. 渐进融合解码器的结构图

功能聚合模块: 图 2（a）显示了具有空间级和类级融合的特征聚合模块（FAM）的设计。我们首先进行空间级融合来模拟局部区域的关系。先前的工作 [21] 已经证明，大核卷积运算是一种简单但高效的结构，用于捕捉局部信息。受此启发，我们采用大核卷积进行空间级融合。具体而言，输入特征图 F_{dec}^{l1} 经过深度卷积层和 MLP 层。深度卷积层有一个 9×9 的深度卷积和一个层范数运算，MLP 层包含两个线性层和一个 GeLU 层。此外，我们在卷积层和 MLP 层中都使用残差连接。在空间级聚合之后，我们进一步沿类别维度应用中的线性自关注操作来执行类级特征聚合。特征聚合模块（FAM）生成的特征图表示为 $\mathcal{F}_{dec}^{fam}$。

跨层融合模块: 特征映射在空间上比较粗糙，缺乏局部细节信息。相比之下，层次编码器中的浅层特征映射 $\mathcal{F}_2, \mathcal{F}_3, \mathcal{F}_4$ 包含了丰富的细节信息。为了结合这些局部细节进行分割，我们引入了跨层融合模块，将低分辨率特征图 $\mathcal{F}_{dec}^{fam}$ 与高分辨率特征图 $\mathcal{F}_2, \mathcal{F}_3, \mathcal{F}_4$ 逐步结合。如图 2(b) 所示，我们首先使用反卷积操作将低分辨率特征映射 Ff 的分辨率提高 2 倍。然后，我们使用卷积运算将对应的高分辨率特征图 $\mathcal{F}_j$，$j \in 2, 3, 4$ 的通道维数降为 16 倍，并将降维后的特征图重复 N 次，使其与 $\mathcal{F}_{dec}^{fam}$ 具有相同的类别维数。然后，我们将上采样的特征图和重复的特征图拼接在一起。为了融合更多的信息，我们还对初始代价图 $\mathcal{F}_{cv}$ 进行上采样和连接。最后，我们将连接的特征映射到 $\mathcal{F}_{dec}^{cat}$ 通过两个卷积层馈送，以生成输出特征 $\mathcal{F}_{dec}^{cat}$。从 [9] 中可以看出，直接将梯度反向传播到图像编码器会降低开放词汇语义分割的性能。因此，我们直接阻止梯度反向传播从跳过层融合模块到图像编码器。

我们的实验表明，与普通 transformer 相比，分层编码结合跨层融合显著提高了性能。这可能是由于分层编码器能够为分词提供丰富的局部信息，并且停止梯度反向传播避免了对开放词汇分词能力的负面影响。

3.3 类别早期拒绝

渐进融合解码器的计算成本与语义类别的数量成正比。当类别数量非常大时，推理时间会显著增加。事实上，大多数图像只包含几个语义类别。因此，解码器的大部分推理时间都用于计算不存在类别的特征。为了提高推理速度，我们引入了一种类别早期拒绝方案来识别这些现有类别，并在早期解码器层拒绝不存在的类别。从当前解码器层中删除与被拒绝类别相对应的特征图，接下来的解码器层只考虑保留的类别。

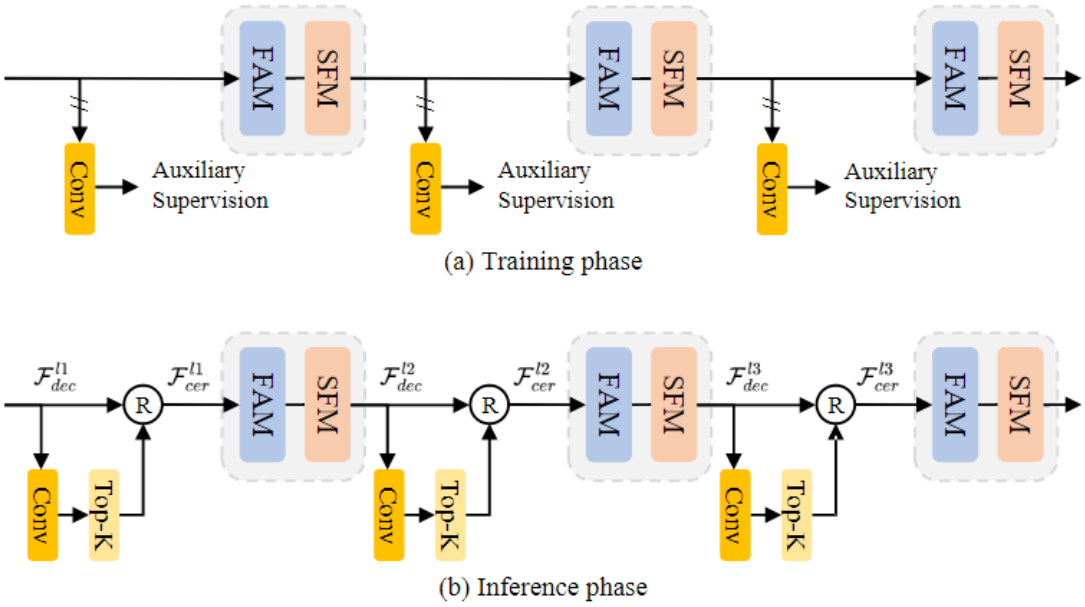

图 3. 类别早期拒绝的结构图

在训练过程中，如图 3 (a) 所示，我们在每一层之后添加辅助卷积分支，分别预测分割图，这些图由 ground-truth 监督。为了避免对模型训练的负面影响，我们停止了它们向解码器的梯度反向传播。在推理过程中，我们在分割图上采用 top-k 策略来预测现有的语义类别。具体来说，我们为每个像素选择响应最大的前 k 个类别，并从所有像素中生成一个类别的联合集，该集合被馈送到下一个解码器层。我们观察到 k=8 可以确保大多数现有类别得到认可。图 3 (b) 显示了推理过程中类别的早期拒绝。我们首先从 $\mathcal{F}_{dec}^{l1}$ 预测分割图，并采用 top-k 策略选择 N_{l1} 类别。然后，我们移除未选定类别的特征图，并生成输出特征图 $\mathcal{F}_{cer}^{l1} \in \mathbb{R}^{H_v \times W_v \times N_{l1} \times D}$。生成的特征图 $\mathcal{F}_{cer}^{l1}$ 被馈送到解码器层。同样，我们为以下层生成类别较少的特征图。因此，大多数不存在的类别在早期层被拒绝，这提高了解码器的推理速度。

4 复现细节

4.1 与已有开源代码对比

复现工作参考了作者在 github(https://github.com/xb534/SED) 上提供的代码。

4.2 数据集

训练集：COCO-Stuff，包含 118k 张密集标注的图像，171 个不同的语义类别。
测试集：

- ADE20K: 有两个不同的测试集：A-150 和 A-847。测试集 A-150 共有 150 个类别，而测试集 A-847 有 847 个类别。
- PASCAL VOC: 包含 20 个不同的对象类别。

- PASCAL-Context: 有两个不同的测试集：PC-59 和 PC-459。测试集 PC-59 有 59 个类别，而测试集 PC-459 有 459 个类别。

4.3 复现细节

采用预训练的视觉语言模型 CLIP 作为基础模型,其中分层骨干 ConvNeXt-B 或 ConvNeXt-L 被用作分层图像编码器。ConvNeXt-B 文本嵌入的特征维度 D_t 为 640，ConvNeXt-L 为 768，类别模板的数量 P 为 80，特征图的通道数 F_{dec}^{l1} 为 128。我们冻结文本编码器，只训练图像编码器和渐进融合解码器。我们在 4 个 NVIDIA 4090 上训练我们的模型，mini-batch 为 4 张图像。优化器 AdamW 采用初始学习率为 2×10−4，权值衰减为 1×10−4。

5 实验结果分析

5.1 复现结果

Method	VLM	Feature backbone	Training dataset	A-847	PC-459	A-150	PC-59	PAS-20
SPNet [46]	-	ResNet-101	PASCAL VOC	-	-	-	24.3	18.3
ZS3Net [2]	-	ResNet-101	PASCAL VOC	-	-	-	19.4	38.3
LSeg [27]	ViT-B/32	ResNet-101	PASCAL VOC-15	-	-	-	-	47.4
LSeg+ [18]	ALIGN	ResNet-101	COCO-Stuff	2.5	5.2	13.0	36.0	-
Han et al. [22]	ViT-B/16	ResNet-101	COCO Panoptic [26]	3.5	7.1	18.8	45.2	83.2
GroupViT [48]	ViT-S/16	-	GCC [40]+YFCC [44]	4.3	4.9	10.6	25.9	50.7
ZegFormer [13]	ViT-B/16	ResNet-101	COCO-Stuff-156	4.9	9.1	16.9	42.8	86.2
ZegFormer [11]	ViT-B/16	ResNet-101	COCO-Stuff	5.6	10.4	18.0	45.5	89.5
SimBaseline [50]	ViT-B/16	ResNet-101	COCO-Stuff	7.0	-	20.5	47.7	88.4
OpenSeg [18]	ALIGN	ResNet-101	COCO Panoptic [26]+LOc. Narr. [36]	4.4	7.9	17.5	40.1	-
DeOP [21]	ViT-B/16	ResNet-101c	COCO-Stuff-156	7.1	9.4	22.9	48.8	91.7
PACL [34]	ViT-B/16	-	GCC [40]+YFCC [44]	-	-	31.4	50.1	72.3
OVSeg [28]	ViT-B/16	ResNet-101c	COCO-Stuff+COCO Caption	7.1	11.0	24.8	53.3	92.6
CAT-Seg [11]	ViT-B/16	ResNet-101	COCO-Stuff	8.4	16.6	27.2	**57.5**	93.7
SAN [51]	ViT-B/16	-	COCO-Stuff	10.1	12.6	27.5	53.8	94.0
SED (Ours)	ConvNeXt-B	-	COCO-Stuff	**11.4**	**18.6**	**31.6**	57.3	**94.4**

图 4. 作者实验结果图

	A-150	A-847	PAS-20	PC-59	PC-459
论文结果	31.6	11.4	94.4	57.3	18.6
复现结果	31.6	11.0	93.5	57.0	18.1

图 5. 复现实验结果图

评价指标为 mIoU，从图 4 和图 5 可以看出，复现性能基本与原论文相当。

5.2 推理时间

	A-150	A-847	PAS-20	PC-59	PC-459
推理时间/s	0. 130	0. 533	0. 075	0. 079	0. 314

图 6. 不同数据集推理时间

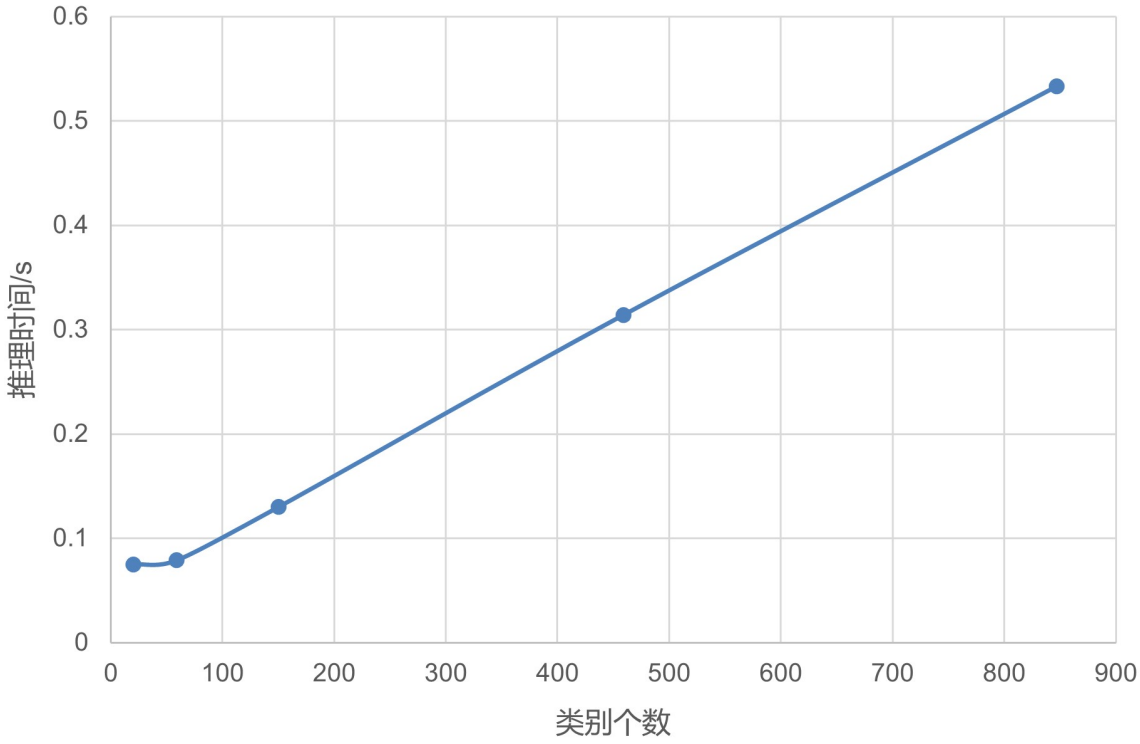

图 7. 推理时间与类别数关系图

图 6 记录的是不同数据集下，每张图片需要的推理时间，根据数据集的类别个数，再绘制了图 7。结合图 6 和图 7，可以看到，推理时间与类别个数基本呈线性关系。

6 总结与展望

本报告基于原论文，介绍了一种称为 SED 的方法，用于开放词汇语义分割。SED 包括基于分层编码器的成本图生成和具有类别早期拒绝的渐进融合解码器。首先使用分层编码器生成像素级图像-文本成本图。基于分层编码器中生成的成本图和不同的特征图，采用渐进融合解码器生成高分辨率特征图进行分割。为了提高速度，在解码器中引入了类别早期拒绝方案来早期拒绝不存在的类别。其在多个数据集上的实验揭示了有效性。本报告还展示了复现细节以及复现结果，从结果可以看出，复现效果基本与原论文相当。而对进一步的研究方向，

一是该方法在类别个数较多的情况下，分割效果依旧很差，仍有很大改进的空间；二是其早期拒绝方案的适用场景可以进一步拓展到其他的任务中去，提高其他任务下的推理速度。

参考文献

[1] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla SegNet. A deep convolutional encoder-decoder architecture for image segmentation. *arXiv preprint arXiv:1511.00561*, 5, 2015.

[2] Yifan Bai, Zeyang Zhao, Yihong Gong, and Xing Wei. Artrackv2: Prompting autoregressive tracker where to look and how to describe. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19048–19057, 2024.

[3] Maxime Bucher, Tuan-Hung Vu, Matthieu Cord, and Patrick Pérez. Zero-shot semantic segmentation. *Advances in Neural Information Processing Systems*, 32, 2019.

[4] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2017.

[5] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In *European conference on computer vision*, pages 104–120. Springer, 2020.

[6] Zhe Chen, Yuchen Duan, Wenhai Wang, Junjun He, Tong Lu, Jifeng Dai, and Yu Qiao. Vision transformer adapter for dense predictions. *arXiv preprint arXiv:2205.08534*, 2022.

[7] Bowen Cheng, Alex Schwing, and Alexander Kirillov. Per-pixel classification is not all you need for semantic segmentation. *Advances in neural information processing systems*, 34:17864–17875, 2021.

[8] Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2818–2829, 2023.

[9] Seokju Cho, Heeseong Shin, Sunghwan Hong, Anurag Arnab, Paul Hongsuck Seo, and Seungryong Kim. Cat-seg: Cost aggregation for open-vocabulary semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4113–4123, 2024.

[10] Jian Ding, Nan Xue, Gui-Song Xia, and Dengxin Dai. Decoupling zero-shot semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11583–11592, 2022.

[11] Zheng Ding, Jieke Wang, and Zhuowen Tu. Open-vocabulary panoptic segmentation maskclip. 2022.

[12] Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[13] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88:303–338, 2010.

[14] Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. Open-vocabulary object detection via vision and language knowledge distillation. *arXiv preprint arXiv:2104.13921*, 2021.

[15] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pages 4904–4916. PMLR, 2021.

[16] Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of naacL-HLT*, volume 1, page 2. Minneapolis, Minnesota, 2019.

[17] Boyi Li, Kilian Q Weinberger, Serge Belongie, Vladlen Koltun, and René Ranftl. Language-driven semantic segmentation. *arXiv preprint arXiv:2201.03546*, 2022.

[18] 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. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7061–7070, 2023.

[19] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11976–11986, 2022.

[20] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015.

[21] Chao Peng, Xiangyu Zhang, Gang Yu, Guiming Luo, and Jian Sun. Large kernel matters–improve semantic segmentation by global convolutional network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4353–4361, 2017.

[22] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[23] Ignacio Rocco, Relja Arandjelovic, and Josef Sivic. Convolutional neural network architecture for geometric matching. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6148–6157, 2017.

[24] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294, 2022.

[25] Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490*, 2019.

[26] Bart Thomee, David A Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. Yfcc100m: The new data in multimedia research. *Communications of the ACM*, 59(2):64–73, 2016.

[27] Dongming Wu, Wencheng Han, Tiancai Wang, Yingfei Liu, Xiangyu Zhang, and Jianbing Shen. Language prompt for autonomous driving. *arXiv preprint arXiv:2309.04379*, 2023.

[28] Yongqin Xian, Subhabrata Choudhury, Yang He, Bernt Schiele, and Zeynep Akata. Semantic projection network for zero-and few-label semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8256–8265, 2019.

[29] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in neural information processing systems*, 34:12077–12090, 2021.

[30] Jiarui Xu, Sifei Liu, Arash Vahdat, Wonmin Byeon, Xiaolong Wang, and Shalini De Mello. Open-vocabulary panoptic segmentation with text-to-image diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2955–2966, 2023.

[31] Qihang Yu, Ju He, Xueqing Deng, Xiaohui Shen, and Liang-Chieh Chen. Convolutions die hard: Open-vocabulary segmentation with single frozen convolutional clip. *Advances in Neural Information Processing Systems*, 36:32215–32234, 2023.

[32] Chong Zhou, Chen Change Loy, and Bo Dai. Extract free dense labels from clip. In *European Conference on Computer Vision*, pages 696–712. Springer, 2022.