arXiv:2304.01198v2 [cs.CV] 7 Aug 2023

基于解耦单次网络的开放词汇语义分割

Cong Han[1*] Yujie Zhong[1*] Dengjie Li[1] Kai Han[2†] Lin Ma[1†]美团 [2]香港大学
hancong0911@163.com jaszhong@hotmail.com kaihanx@hku.hk

摘要

近年来，开放词汇语义分割问题受到了越来越多的关注，表现最佳的方法均基于双流网络：一条流用于候选掩码生成，另一条流则使用预训练的视觉语言模型进行分割区域分类。然而，现有的双流方法需要将大量（多达一百个）图像裁剪块输入视觉语言模型，效率极低。为解决这一问题，我们提出了一种网络，对每张输入图像只需通过视觉语言模型进行一次前向处理。具体而言，我们首先提出一种新颖的网络适配方法，称为图块分离，用于限制预训练视觉编码器中图块嵌入之间的有害干扰。随后，我们提出分类锚点学习，以促使网络在空间上聚焦于更具判别性的分类特征。大量实验表明，所提出的方法表现出色，超越了当前最先进的方法，同时推理速度快 *4* 到 *7* 倍。代码：*https://github.com/CongHan0808/DeOP.git*

1. 引言

语义分割是计算机视觉中的一项关键任务，旨在将图像像素划分为具有语义意义的区域，并预测其类别标签。以往的语义分割网络主要针对数据集定义的一组固定语义类别。近年来，随着 CLIP [32]和 ALIGN [19]等大规模预训练视觉语言模型（VLM）的兴起，开放词汇（也称零样本）语义分割日益受到关注。

OpenSeg [15] 是较早用于开放词汇语义分割任务的方法之一。该方法采用类别无关掩码表示可能的语义区域，再利用从预训练视觉语言模型中提取的文本嵌入对其进行分类（在 OpenSeg 中使用的是 ALIGN [19] ）。由于只需将图像输入视觉编码器一次即可完成分割，因此效率较高。**然而，单一的**

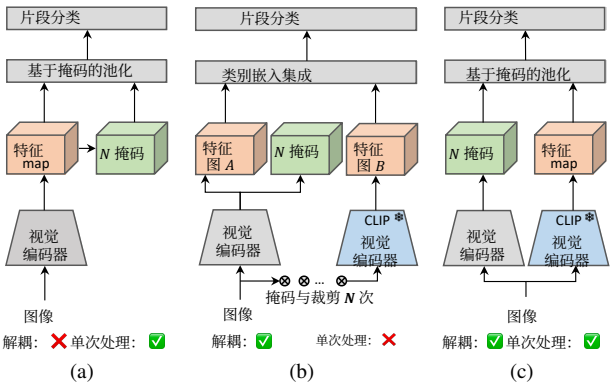


图 1：开放词汇语义分割的三种宏观架构对比：**(a) 耦合**网络，例如 OpenSeg [15]；**(b) 解耦多遍处理**网- 络，例如 SimBaseline [38]；**(c)解耦单次处理** 网- 络（我们的基线）。解耦结构能够保持预训练视觉语言模型的通用性，而单遍机制则带来了较高的计算效率。

在训练过程中更新的共享视觉编码器不可避免地会破坏预训练视觉语言模型原有的视觉语言对齐。因此，为确保其零样本能力，需要使用大量语义概念进行训练。例如，必须使用 Localized Narratives [31]进行训练。我们将 OpenSeg 称为耦合式单次处理方法。另一类方法包括 ZegFormer [10]、SimBaseline [38]和 OVSeg [26], ，它们将两个子任务解耦，并设计了双流架构。在这种架构中，类别无关掩码提议流和掩码分类流使用两个独立的骨干网络，从而保留视觉语言模型视觉编码器的通用性。**然而，这类方法的缺点是计算开销较大，因为候选掩码生成的图像裁剪块需要逐一输入视觉编码器**。我们将这类方法称为解耦式多次处理方法。

在本研究中，我们试图回答一个问题：能否开发一种框架，在保持 VLM 零样本能力的同时实现较高的计算效率？为探索这一可能性，我们首先基于两个关键设计原则——解耦 和单次处理，设计了一种基线架构。如图 1(c) 所示，我们的基线方法

*贡献相同。
†通讯作者。

采用了两个解耦的视觉骨干网络，同时将预训练视觉语言模型（即 本研究中的 CLIP）的视觉编码器保持冻结，以保留其泛化能力。此外，为实现高效率，分类流中的图像只需处理一次。然而，我们通过实验发现，这种基线架构的性能并不理想。我们进一步应用提示学习，使 CLIP 视觉编码器和文本编码器都适应当前的分割任务，但基线模型与此前解耦方法之间的性能差距仍然显著。

经过全面分析，我们发现，分类性能不佳的主要原因在于片段分类，而非掩码的质量或召回率。具体而言，我们确定了导致这一问题的两个主要原因：**(1) 在原始 CLIP 视觉骨干网络中，属于不同分割区域的图块嵌入之间存在过多交互**。 相比之下，在多次处理解耦方法中，不同掩码中的图块嵌入在通过视觉-语言模型时完全没有交互。尽管属于不同掩码的图像块可以为分类提供一些上下文信息，但不同片段嵌入之间过度交互实际上会损害分类性能。**(2) 通过对图块嵌入进行基于掩码的池化得到的最终片段嵌入，并非分类的最优表示**。 这是因为掩码提议网络以类别无关方式训练，掩码中的权重仅表示每个图块嵌入属于特定分割区域的可能性，而与类别无关。因此，基于这些掩码对图块嵌入进行池化并非分类的最优方案。因此，我们认为，解决这些问题有望在保持单次处理分割模型计算效率的同时，提升其分类性能。

为此，我们提出了一种名为解耦单次处理网络（DeOP）的双流框架，并分别通过广义图块分离（GPS）和分类锚点学习（CAL）来缓解上述两个问题。**广义图块分离**可以视为对 CLIP 视觉编码器（例如 ViT-Base）中的图块令牌/嵌入执行的一种“分离”操作。其旨在减少编码器中图块令牌之间的有害干扰，同时保持 CLIP 的嵌入空间。**分类锚点学习**旨在寻找更适合片段分类的图像块，我们将其称为分类锚点。具体做法是在 CLIP 视觉编码器末端附加一个模块，由该模块为每个掩码候选区域生成一张热力图，指示后续分类空间池化时应关注哪些图块嵌入（即锚点）。

我们在公开基准上进行了大量实验，并证明 DeOP 在数据集内和跨数据集评估中始终优于以往方法，同时比其他多遍方法（例如 SimBaseline）效率显著更高，从而验证了我们提出的 GPS 和 CAL 的有效性。

2. 相关工作

视觉-语言预训练。视觉-语言预训练旨在建立图像-文本表征之间的联系，以用于图像-文本检索、视觉问答、文本生成图像和密集预测等任务。一种具有代表性的预训练模型 CLIP [32]，利用包含 4 亿组（图像、文本）配对数据的海量互联网数据集学习图像表示，在零样本图像分类方面表现出优异的性能。为提升预训练视觉-语言模型在下游任务中的性能，研究者提出了文本提示学习，通过向文本提示中添加额外的可学习令牌，使模型适应新任务。研究表明，文本提示学习在少样本和零样本设置下均能在各种任务中实现出色的泛化能力，例如 [47, 21, 14]。

语义分割。语义分割是计算机视觉中的一项基础任务，用于将图像中属于同一类别的部分归为一组。全卷积网络（FCNs） [33] 是基于深度网络的语义分割领域的开创性工作，将该任务定义为逐像素分类任务。随后出现了多种变体，例如 ASPP [5, 6], 、PPM [43], 和 OCNet [39], 。近年来，MaskFormer [9, 8] 提出预测一组二值掩码，每个掩码都对应一个全局类别标签预测结果。

开放词汇语义分割。开放词汇语义分割通过基于像素的方法（例如 ZS3Net [2], CaGNet [16], SPNet [36], Fusioner [29] ）或基于区域的语义分割方法（例如 OpenSeg [15], ZegFormer [10], SimBaseline [38], OVSeg [26] 和 ODISE [37]），为未见物体类别学习逐像素分类器。ZS3Net 使用生成模型，而 CaGNet 融入上下文感知特征生成。SPNet 使用固定的词嵌入投影矩阵。ZegFormer、SimBaseline 以及两项同期工作 OVSeg 和 ODISE 会生成逐片段嵌入，随后进行类别无关分组和逐片段开放词汇分类。本文采用了与它们类似的解耦架构。值得注意的是，该解耦架构可轻松与更先进的网络（例如 SAM 系列 [22, 44, 40]或 Mask DINO [24]）结合，以改进掩码提议生成。

3. 基线方法

问题定义。 我们遵循 ZS3Net [2], 中零样本分割的设置，其中所有类别都定义为

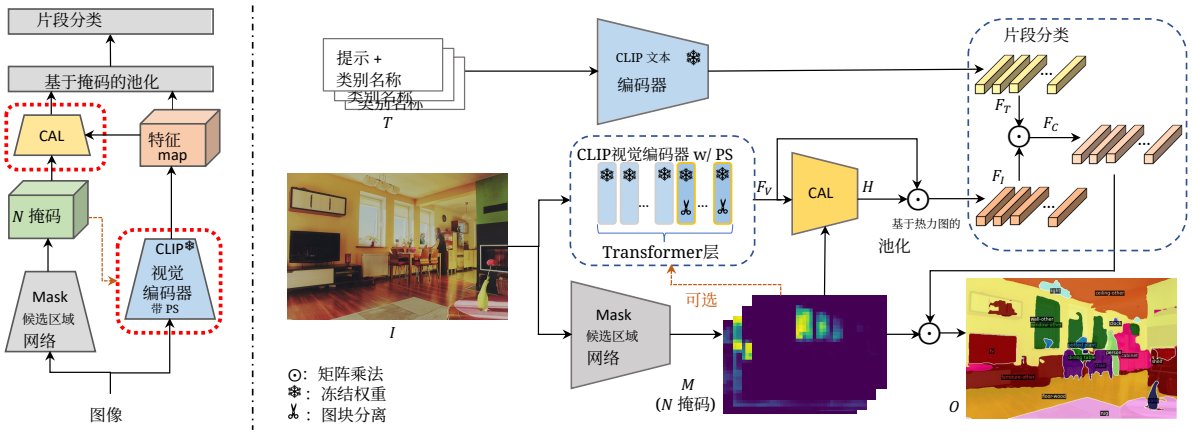


图 2：**左侧**：DeOP 概览。与基线网络相比，额外的图块分离（PS）和分类学习（CAL）模块以红色虚线标出。**右侧**：详细流程。我们将图像输入掩码提议网络以生成 N 候选掩码 M，同时将图像输入 CLIP视觉编码器（含 PS）以获取视觉特征 F_V。将 M 和 F_V 输入 CAL 模块，为 M 生成分类锚点热力图 H。使用 H 进行基于热力图的池化，以获取 F_I。CLIP 文本编码器生成文本嵌入 F_T，并将其用作分类器，以获得每个掩码的分类分数。最后，结合 F_C 和 M 得到区域预测。

$C = C_s \bigcup C_u$ 和 $C_s \bigcap C_u = \varnothing$，其中 C_s 是已见类别集合，C_u 是未见类别集合。在训练阶段，模型需要将图像映射到具有已见类别的不同区域。在推理阶段，模型预测所有已见类别和未见类别的结果。

如前所述，我们旨在基于两个关键原则解决开放词汇语义分割任务：一是保持预训练视觉语言模型对未见类别的泛化能力，二是实现高计算效率。接下来，我们首先提出一个具备这两种所需能力的基线网络（第3.1节），然后介绍一种改进的基线方法（即 基线+），该方法通过在基线上应用提示学习得到（第3.2节）。最后，我们在第3.3节讨论所提出基线方法的局限性。

3.1. 基线网络架构

图 1(c) 展示了基线方法的解耦式单次处理架构。该网络包含两个分支：类别无关掩码提议网络和掩码分类网络。掩码提议网络将像素分组，生成一系列类别无关掩码。借助该提议网络，我们可以将像素级分类转化为区域级分类。掩码分类分支由一个冻结的 CLIP视觉编码器和通过 CLIP文本编码器获得的分类器组成。

具体而言，我们将图像输入掩码提议网络和 CLIP视觉编码器，分别生成 N 候选区域掩码 M 和图像特征图。随后，将N 候选区域掩码应用于 CLIP 特征图，通过基于掩码的池化 [15]获取 N 特征嵌入。另一方面，类别嵌入 F_C （即 分类器）是..... 通过计算视觉特征 F_I 与 CLIP 文本编码器生成的文本特征嵌入 F_T之间的点积获得。该编码器对文本类别进行编码。最后，我们通过矩阵乘法将 N 候选区域掩码与其类别预测结果相结合，得到最终预测结果 O。我们可以将其表述为 $F_C = F_T * F_I$和 $O = F_C * M$。

基于 MaskFormer 的掩码提议网络。我们通过改造 MaskFormer [9]，构建了一个类别无关变体作为掩码提议网络。与逐像素分类分割模型不同，MaskFormer 是一种掩码分类模型，它预测一组掩码，每个掩码都对应一个全局类别标签。为了使 MaskFormer 实现类别无关，我们只需移除类别嵌入组件，同时保留与掩码相关的组件。（有关修改后 MaskFormer 结构的详细说明，请参阅 A。）

基线的训练。在我们的基线方法中，只需训练掩码提议网络，而分类分支无需训练。训练候选区域网络时，仅使用掩码损失，并移除分类损失。此外，我们以离线方式学习文本提示，与 [38]相同。

3.2. 通过提示学习改进基线

值得注意的是，由于为了保留 CLIP 的零样本能力而冻结了 CLIP视觉编码器，提升基线分类性能的一种直接方法是提示学习 [47]。提示学习可以微调 CLIP 的嵌入空间，使其更好地适应当前任务，同时不破坏视觉与语言之间的对齐。在本研究中，我们同时采用视觉提示调优和

文本提示学习来增强我们的基线，由此得到的方法记作baseline+。

视觉提示调优。 视觉提示调优 [20](VPT) 是一种高效且有效的方法，可替代视觉领域大型 Transformer模型的完全微调。我们遵循 VPT 的方法，对预训练的 CLIP视觉编码器模型进行微调。具体而言，我们以逐元素方式将一组可学习提示添加到图像块嵌入序列中。训练后，这些提示可以轻微调整 CLIP 的视觉嵌入空间，使其更适用于分割任务。

在线文本提示学习。 在 SimBaseline 中，文本提示通过分类任务进行离线学习。为了进一步提升分割性能，我们提出以端到端的方式，将它们与整个分割网络一同训练。具体而言，可学习的文本提示首先以离线方式初始化为 [38],，随后与视觉提示一同针对语义分割任务进行训练。

基线训练+。 基线+需要进行端到端训练，以学习 CLIP 编码器的视觉提示和文本提示。由于模型针对分割任务进行训练，我们采用了骰子损失 [25] 和焦点损失 [27], 的组合，这是语义分割中的常见做法。

3.3. 局限性

从实验结果来看，基线+ 方法与以往解耦方法（即 ZegFormer 和 SimBaseline）之间的性能差距仍然显著，如表3所示。由此可推断，差距主要由片段分类造成，因为它们的掩码提议网络相似。本文关注导致分类性能较差的两个问题：（1）原始 CLIP 视觉编码器可能无法感知图块嵌入与语义片段之间的对应关系，而且属于不同分割区域的图块嵌入可能会在编码器中发生过度或无目的的交互。需要注意的是，以往的解耦方法不存在这一问题，因为属于不同分割区域的图块嵌入不会相互交互，这是由于经过掩码处理的图像分别通过视觉编码器。（2）通过对图块嵌入进行基于掩码的池化所得到的最终片段嵌入并不适合分类。掩码提议网络以类别无关方式训练，因此，输出的掩码可能不适用于片段分类的基于掩码的池化。

4. 我们的方法 – DeOP

为缓解第3.3节讨论的两个问题，我们提出一种适配视觉Transformer的方法，称为图像块分离，以减少CLIP视觉编码器中图像块嵌入之间的有害交互；同时提出分类锚点学习（CAL），以便在用于分类的基于掩码的池化中生成更好的表示。

整体流程。 图 2（左）展示了我们提出的框架 DeOP 的概览。该框架遵循基线方法的双流设计。两个新增组件——图块分离和 CAL——以红色虚线标出。所提出的图块分离应用于 CLIP视觉编码器中的一个或多个层，而 CAL 则通过在 CLIP视觉编码器之后添加一个基于查询的热图解码器来实现。其余流程（如图 2 右侧所示），包括掩码提议网络、由文本生成的分类器以及片段分类，均与基线网络相同。在 DeOP 的训练中，采用与基线相同的损失+ 。值得注意的是，在推理时，DeOP 不需要像 [10, 38]中那样对分类流和掩码提议流的分类分数进行启发式集成。

下文将详细介绍第4.1节提出的图块分离方法，以及第4.2节的分类锚点学习。

4.1. 图块分离

通常，由于Transformer层中的自注意力层，基于ViT的CLIP视觉编码器提取的特征图包含大量整幅图像的全局信息。然而，这一特性可能会对语义分割任务造成过度干扰。

分类分支旨在对每个区域进行分类。尽管来自其他区域的一定程度的信息可以提供一些上下文并有利于分类，但过多的信息可能会产生负面影响，分散分类注意力。

为此，我们提出了一种视觉Transformer适配方法，称为**图块分离**（PS），用于限制CLIP视觉编码器中图块嵌入之间的干扰。下文将比较两种类型的图块分离：掩码引导图块分离和广义图块分离。它们可以灵活地应用于基于ViT的视觉编码器中的一个或多个层。

掩码引导图块分离。 掩码引导图块分离（MPS）是一种用于引导 Transformer 层中图块嵌入之间交互的操作。如图 3（左）所示，MPS 使用候选掩码（由掩码提议网络生成）替代自注意力操作中键与查询之间的乘法，使每个图块嵌入主要与属于同一预测分割区域的图块交互，而与属于其他预测分割区域的图块几乎不发生交互。每个图块嵌入的掩码引导注意力由 N 提议掩码导出。具体而言，第 i 个图块嵌入的掩码引导注意力通过

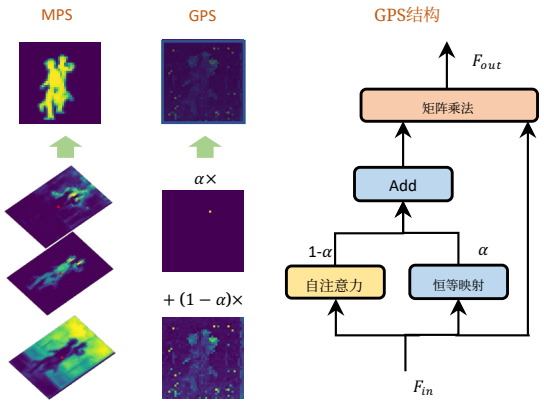


图3：**左**：MPS 与 GPS 在自注意力层中如何获取最终注意力热图（针对特定图块嵌入）的比较。**右**：GPS 的结构，它用于替换视觉Transformer中Transformer层的原始自注意力层。

公式1计算得出：

$$M_i = \sum_{j=1}^{N} M_{j,i} * M_j, \tag{1}$$

其中，M_i 是第 i个图块嵌入的掩码注意力图，M_j 是第 j 个掩码，$M_{j,\ i}$ 是第 i个图块嵌入对应的掩码M_j 的值。

广义图块分离。 不同于利用候选区域掩码作为强先验来引导图块交互的 MPS，本文提出了一种广义图块分离（GPS），且不需要先验（图3中间）。

GPS 将自注意力中的查询-键乘法替换为查询-键乘法与单位矩阵的加权组合。从数学上讲，GPS 可表示为公式2，其中 E 是一个单位矩阵，$Attn$ 是自注意力 [34] 矩阵。在一种极端情况下，当 $\alpha = 0$时，GPS 退化为普通自注意力。在另一种极端情况下，当 $\alpha = 1$时，GPS 等价于完全图块分离（即自注意力层中不发生任何交互）。超参数 α 使这两种极端情况之间能够平滑过渡。

$$F_{out} = (\alpha * E + (1-\alpha) * Attn) \times F_{in}. \tag{2}$$

如图 2 所示，图块分离可以多次应用于不同的Transformer层；但在实践中，我们发现仅将其应用于最后一个Transformer层就是一种有效的选择。

4.1.1 与现有适配方法的比较

嵌入空间得以保留。以往的适应方法，例如 ViT-adapter [7]，通常需要更新骨干网络，这很容易破坏原始嵌入空间。然而，与 VPT [20], 类似，所提出的图块分离能够保留预训练视觉语言模型（例如 CLIP）的原始视觉嵌入空间，从而

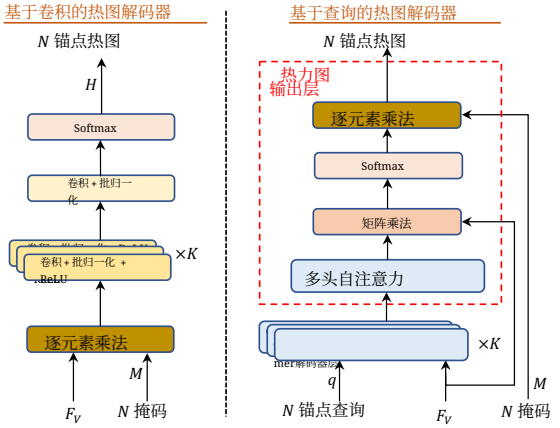


图 4：用于CAL的基于卷积的热图解码器与新型基于查询的热图解码器之间的比较。

维持视觉语言对齐和零样本泛化能力。这一特性符合我们的第一项设计原则。

无需训练。 与现有方法（如ViT-adapter [7] 和VPT [20]）不同，这些方法会向骨干网络引入额外参数，因此需要进一步微调；图块分离不引入额外参数，也无需训练。此外，图块分离可以与先前的适配方法（例如 VPT）协同工作，从而更好地适配当前任务。

4.2. 基于查询的分类锚点学习

为克服第二个局限，即聚焦于更具判别性的图像块进行分类（即 分类锚点），本文提出了分类锚点学习（CAL）。例如，要对人进行分类，网络可能会比身体其他部位更关注头部。本质上，CAL与目标检测器训练中的锚点选择/学习理念相似 [42, 41, 13], ，但其目标是为每个候选掩码输出一张锚点热力图。为此，我们提出了一个基于查询的热图解码器模块。如图 2 所示，热图解码器（图中以CAL标示）连接在CLIP视觉编码器之后。它可以生成注意力热图，指示哪些图像块更适合从分类角度表示这些分割区域。该模块与整个分割网络进行端到端训练。下文将对此进行详细说明。

4.2.1 锚点热图解码器

我们首先介绍所提出的基于查询的热图解码器，然后介绍一种基于卷积的简单替代方案。两者都能够实现CAL，并在第 5.3 节中进行了实验比较。

基于查询的热图解码器。总体而言，解码器接收 N 个可学习的锚点查询（记作 q）、特征

方法	解耦	遍	COCO-Stuff				VOC			
			hIoU	Seen	未见	FPS	hIoU	Seen	未见	FPS
ZS3Net [2]	✗	1	15.0	34.7	9.5	-	28.7	77.3	17.7	-
CaGNet [16]	✗	1	18.2	35.5	12.2	-	39.7	78.4	25.6	-
ZegFormer [10]	✓	$N^{'}$	34.8	36.6	33.2	-	73.3	86.4	63.6	-
SimBaseline [38]	✓	$N^{'}$	37.8	**39.3**	36.3	1.11	77.5	83.5	72.5	2.66
DeOP（我们的方法）	✓	1	**38.2**	38.0	**38.4**	**4.37**	**80.8**	**88.2**	**74.6**	**6.55**

表1：开放词汇设定下与其他方法在 COCO-Stuff 和 Pascal VOC 上的比较。*passes* 表示输入图像通过 CLIP 视觉编码器的次数。请注意，$N^{'}$ 取决于输入图像，因为 ZegFormer 和 SimBaseline 只将得分高于阈值的图像裁剪块输入 CLIP，而 $N^{'} \leq N$。

图（由 CLIP视觉编码器输出，记作 F_V）和N 个候选掩码（记作 M）。输出为 N 个分类锚点热力图 H。基于查询的热图解码器结构如图 4（右）所示。解码器由 K 个标准 Transformer解码器层和一个新颖的热图输出层组成。K 个 Transformer层通过关注输入图像的特征图来更新锚点查询 q。

热图解码器的关键组件是**热图输出层**。与输出为更新后查询的标准 Transformer 解码器层不同，所提出的热图解码器层旨在生成锚点热力图 H。具体而言，锚点查询首先经过一个多头自注意力层，然后与特征图相乘，并经过 Softmax，以获得每个锚点查询对应的注意力图。最后，注意力图与 N 候选掩码相乘，使每个锚点查询都能感知其对应的掩码/分割区域。具体来说，热图输出层可表示为公式3:

$$H = softmax(\frac{q \times F_V}{\sqrt{d}}) * M, \tag{3}$$

其中 $q \in \mathbb{R}^{B\times N\times C}$ 表示 N 锚点查询，$F_V \in \mathbb{R}^{B\times C\times H\times W}$ 表示由 CLIP视觉编码器生成的特征图，d 表示 F_V的维度，$M \in \mathbb{R}^{B\times N\times H\times W}$ 是N 掩码，$*$ 表示逐元素相乘。

基于卷积的热图解码器。 所提出的基于查询的热图解码器并非实现 CAL 的唯一方式。在本文中，我们还将基于查询的热图解码器与卷积对应结构进行比较（如图 4 左侧所示）。这个基于卷积的锚点热图解码器首先将每个候选掩码应用于特征图，生成 N 加权特征图。每个加权特征图都会输入由 K 卷积-BN [18]-ReLU [1] 层构成的堆栈，其中卷积层的卷积核大小为 3，通道数为 512）。最后，额外的卷积层将通道维度映射为 1，再经由 Softmax 层生成 N锚点热图。从数学上讲，基于卷积的锚点热图解码器可表示如下:

$$H = \sigma(conv(CBR_K(\cdots CBR_1(F * M)))), \tag{4}$$

其中，CBR 表示卷积-BN-ReLU层，σ 表示 Softmax 操作。

4.2.2 基于分类锚点热图

这里介绍如何利用分类锚点热图进行分割区域分类。我们只需沿用 [15], 中描述的基于掩码的池化，唯一的区别是，池化时使用学习得到的分类锚点热图，而不是候选掩码。我们将这一过程称为基于热力图的池化。其形式化表示为 $F_I = H \odot F_V$，其中 F_I 是候选掩码/分割区域的分类嵌入。

4.2.3 关于 CAL 的讨论

所提出的 CAL 能够保留 CLIP 原有的视觉对齐能力，因为它只学习一个空间热力图，用于降低或提高 CLIP 输出的某些图块嵌入的权重。此外，尽管 CAL 在已见类别上训练，但它也能很好地迁移到未见类别。这一出色特性已在第 5.3 节中通过定量实验得到验证。

5. 实验

数据集。 我们使用四个广泛应用的图像分割数据集研究 DeOP：COCO-Stuff [3, 28], PASCAL VOC 2012 [12], PASCAL Conext [30] 和 ADE20K [45, 46]。关于数据划分和评估指标，我们直接沿用 [38]。数据集统计信息、数据划分和评估指标的详情见补充材料。

实现细节。 我们基于 Detectron2 [35]开展实验。我们选用 ResNet-101c [4] ，它将 ResNet-101 [17]的首个 7×7 卷积层替换为连续的 3 个 3×3 卷积层，作为 MaskFormer 的骨干网络；并默认将 Transformer 解码器中的查询 N （即候选掩码）数量设为 100。对于 CLIP，我们使用 ViT-B/16 [11] 模型的文本编码器和图像编码器。实验在 8 块 Tesla V100 32G GPU 上进行，批次大小设为 32。PASCAL VOC 2012 和 COCO-Stuff 的训练迭代次数分别为 20k 和 60k。速度测试在 COCO-Stuff 数据集上进行，批次大小为 1，使用 1 块 Tesla V100 32G GPU。

方法	骨干网络	训练数据集	De.	遍数	VOC-20	PC-59	A-150	PC-459	A-847
ZS3Net [2]	R101	Pascal VOC 2012	✗	1	38.3	19.4	-	-	-
LSeg [23]	R101	Pascal VOC 2012	✗	1	47.4	-	-	-	-
OpenSeg [15]	R101	COCO	✗	1	60.0	36.9	15.3	6.5	4.0
OpenSeg [15]	R101	COCO +定位叙述	✗	1	63.8	40.1	17.5	7.9	4.4
SimBase [38]	R101c	COCO-Stuff-156	✓	N'	88.4	47.7	20.5	-	7.0
DeOP（我们的方法）	R101c	COCO-Stuff-156	✓	1	**91.7**	**48.8**	**22.9**	**9.4**	**7.1**

表2：在跨数据集设置下，与其他方法在 Pascal VOC、Pascal Context 和 ADE20K 上的比较。每个数据集后面的数字表示类别数量。 $De.$ 表示解耦网络。 $Loc.\ Narr.$ 表示 Localized Narratives [31]。

方法	COCO-Stuff				VOC			
	pAcc	hIoU	Seen	未见	pAcc	hIoU	Seen	未见
基线	15.6	7.0	6.2	8.0	46.6	30.7	37.5	26.0
基线+	27.6	8.3	8.0	8.7	64.3	33.4	53.2	24.4
基线+搭配 PS	53.9	25.8	25.9	25.4	72.7	47.6	64.0	37.9
基线+使用 CAL	54.2	23.2	28.8	21.0	88.1	70.5	83.3	61.2
DeOP（我们的方法）	**62.2**	**38.2**	**38.0**	**38.4**	**92.5**	**80.8**	**88.2**	**74.6**

表3：消融实验分析 PS（即图块分离）和 CAL（即分类锚点学习）的有效性。

5.1. 开放词汇设定下的结果

我们通过在 COCO-Stuff 数据集和 PASCAL VOC 2012 数据集上训练模型，在开放词汇分割设置下评估我们的方法。训练期间，我们仅使用训练数据集中属于已见类别的图像，同时纳入验证数据集中的所有图像。表 1 展示了与其他方法的比较。

如表1所示，DeOP显著优于以往方法。值得注意的是，我们在 COCO-Stuff 数据集上取得了 38.2 的 hIoU，在 PASCAL VOC 2012 数据集上取得了 80.8 的 hIoU。相较于 ZegFormer 和 SimBaseline，性能提升主要归功于所提出的图块分离方法以及用于分类的学习式锚点热力图。此外，我们认为，这些多遍方法会针对每个候选分割区域彻底去除上下文（即分割区域周围的信息）（这是由于预处理中的掩码操作），这可能会影响性能；而 DeOP 在单遍架构中能够自然地利用上下文信息。

推理速度。 为评估我们方法的效率，我们在表 1 中比较了我们的方法与 SimBaseline 的推理速度。值得注意的是，我们的方法使推理速度（帧/秒）提升约 3 至 4× 倍。这体现了采用单遍机制的优势与效率，这也是我们的第二项设计原则。

5.2. 跨数据集设置下的结果

我们通过在 COCO-Stuff 数据集上训练模型，并在不进行微调的情况下在其他数据集上评估模型，来评测我们的方法在跨数据集设置下的表现。表2显示，得益于解耦设计（即第一项设计原则）和冻结的视觉CLIP编码器，我们的方法在其他数据集上具有良好的泛化能力，并且优于所有以往方法。具体而言，我们的方法在 Pascal VOC 数据集上取得了 91.7 mIoU，在 Pascal Context 数据集上取得了 48.8 mIoU，在 ADE20K数据集上取得了 22.9 mIoU。值得注意的是，随着类别数量从左向右增加，相应的性能指标逐渐下降，这表明提升分类能力至关重要。

提示学习、PS 和 CAL 的有效性。 在表 3 中，我们验证了所提出的各个组件相较于普通基线的有效性。正如基线+所示，提示学习有助于提升像素准确率（pAcc）：在 COCO-Stuff 数据集上，pAcc 从 15.6 提升至 27.6；在 VOC 数据集上，则从 46.6 提升至 64.3。PS 和 CAL 均能提升所有指标。不出所料，DeOP（包含提示学习、PS 和 CAL）在所有指标上均显著优于其他方法，这表明 PS 和 CAL 能够协同工作，并能很好地泛化到未见类别。在补充材料中，我们详细分析了文本提示学习与视觉提示学习的不同组合。

关于PS的分析。 如第4.1节所述，我们比较了两种图块分离类型：MPS和GPS。结果如表4所示。首先可以得出的结论是，这两种PS的表现都优于原始自注意力。总体而言，GPS的表现优于MPS。我们认为，这是因为先验（即 加权提议掩码）过强，可能并非最优。有趣的是，最佳组合是最后一层采用完全图块分离。值得注意的是，我们还尝试将PS应用于视觉Transformer中的更多Transformer层。然而，我们发现PS



图 5：热力图的可视化。从左到右依次为：原始图像、CLIP视觉编码器最后一层中某个图像块 i 的自注意力热图、另一个图像块 j的自注意力热图、包含图像块 i 的分割区域的候选掩码，以及为图像块 i学习得到的分类锚点热力图。三行中的类别分别为：人（已见）、草（**未见**）和飞盘（**未见**）。

方法	α_{-2}	α_{-1}	hIoU	Seen	**未见**
基线	0	0	7.0	6.2	8.0
MPS	-	-	12.5	11.0	14.4
GPS	0	0.5	20.0	17.1	24.0
GPS	0	1.0	**23.8**	**21.5**	**26.6**
GPS	0.5	1.0	16.2	15.2	17.4
GPS	0.5	0.5	13.3	12.1	14.8

表4：关于图块分离的消融研究。 α_{-1} 和 α_{-2}分别表示视觉Transformer中最后一层和倒数第二层的超参数 α 。实验基于我们的基线方法（不进行提示学习）。

如果将其应用于较浅的层（即 第一层），可能会破坏嵌入空间，导致性能不佳，如表 9 所示。

CAL 分析。 我们比较了第 4.2 节中介绍的 CAL 所用的基于查询和基于卷积的热图解码器。对于基于查询的解码器，我们还研究了解码器深度（K）的影响。表 5 的结果表明，基于查询的热图解码器优于基于卷积的解码器，并且性能对热图解码器的深度不敏感。

Type	K	pAcc	hIoU	Seen	**未见**
DeOP（不含 CAL）	0	53.9	25.7	25.9	25.4
基于卷积	3	58.8	31.4	31.4	31.4
基于查询	1	**62.2**	**38.2**	**38.0**	**38.4**
基于查询	3	62.2	37.7	37.9	37.6
基于查询	5	62.2	38.0	38.0	38.0

表 5：分类锚点学习的消融研究。 K 表示热图解码器中的层数。

方法	pAcc	hIoU	Seen	**未见**	FPS
SimBaseline	62.1	38.6	**40.7**	37.0	0.27
DeOP（我们的方法）	**62.6**	**40.0**	40.4	**39.6**	**1.85**

表6：基于ViT-L的CLIP视觉编码器的结果。

更大型 CLIP 视觉编码器上的结果。我们还研究了更强大的 CLIP 视觉编码器对结果的影响。具体而言，我们在 COCO-Stuff 数据集上使用基于 ViT-L/14-336 的 CLIP 视觉编码器训练 SimBaseline 和 DeOP 模型。如表 6 所示，我们的方法在未见类别上始终优于 SimBaseline。此外，DeOP 在推理时实现了更显著的加速（至少 7×）。

5.3. 可视化C

比较A 个性意力掩码。 我们在图 5 中可视化了注意力图、掩码候选和锚点热力图。显然，图像块的自注意力图杂乱无章，无法为分割任务提供有用信息。CAL 生成的注意力图关注图像块周围的局部信息（例如关注树木，以帮助对草进行分类）；CAL 还使模型能够更加关注对分类更具判别性的区域，例如，人的头部比身体更为重要。更多热力图可视化结果和分割结果见补充材料。

6. 结论

在本研究中，我们提出了一种名为 DeOP 的高效开放词汇语义分割框架。与以往的解耦方法不同，DeOP 在视觉语言模型中仅需对输入图像进行一次处理。具体而言，DeOP 配备了我们提出的图块分离机制，用于限制视觉语言模型视觉编码器中图块嵌入之间的冗余干扰；同时采用分类锚点学习，以识别视觉上具有判别性的区域，从而提升分割区域分类效果。我们进行了大量实验来验证该方法，结果发现，DeOP 在同数据集和跨数据集评估中均优于以往方法，同时在推理阶段显著快于其他多遍方法。

致谢 本研究得到中国国家重点研发计划（编号：2022ZD0118700）、香港研究资助局杰出青年学者计划（项目编号：27208022）以及香港大学基础研究种子基金的资助。

参考文献

[1] Abien Fred Agarap. 使用修正线性单元（ReLU）进行深度学习。*ArXiv*，abs/1803.08375，2018年。[2] Max Bucher、Tuan-Hung Vu、Matthieu Cord 和 Patrick P´erez。零样本语义分割。*ArXiv*，abs/1906.00817，2019年。[3] Holger Caesar、Jasper R. R. Uijlings 和 Vittorio Fer- rari。COCO-Stuff：上下文中的事物类和背景类。*2018 IEEE/CVF*计算机视觉与模式识别会议，第1209–1218页，2016年。[4] Liang-Chieh Chen、George Papandreou、Iasonas Kokkinos、Kevin Murphy 和 Alan L Yuille。DeepLab：使用深度卷积网络、空洞卷积和全连接条件随机场进行语义图像分割。*IEEE*模式分析与机器智能汇刊，40(4)：834–848，2017年。[5] Liang-Chieh Chen、George Papandreou、Iasonas Kokkinos、Kevin P. Murphy 和 Alan Loddon Yuille。DeepLab：使用深度卷积网络、空洞卷积和全连接条件随机场进行语义图像分割。*IEEE*模式分析与机器智能汇刊，40 848，2016 834：–。

[6] Liang-Chieh Chen、George Papandreou、Florian Schroff 和 Hartwig Adam。重新思考用于语义图像分割的空洞卷积。*ArXiv*，abs/1706.05587，2017年。[7] Zhe Chen、Yuchen Duan、Wenhai Wang、Junjun He、Tong Lu、Jifeng Dai 和 Y. Qiao。面向密集预测的视觉Transformer适配器。*ArXiv*，abs/2205.08534，2022年。[8] Bowen Cheng、Ishan Misra、Alexander G. Schwing、Alexan- der Kirillov 和 Rohit Girdhar。用于通用图像分割的掩码注意力掩码Transformer。*2022IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR)*，第1280–1289页，2021年。[9] Bowen Cheng、Alexander G. Schwing 和 Alexander Kir- illov。逐像素分类并非语义分割所需的一切。载于*Neural Information Processing Systems*，2021年。[10] Jian Ding、Nan Xue、Guisong Xia 和 Dengxin Dai。解耦零样本语义分割。*2022 IEEE/CVFConference on Computer Vision and Pattern Recognition (CVPR)*，第11573–11582页，2021年。[11] Alexey Dosovitskiy、Lucas Beyer、Alexander Kolesnikov、Dirk Weissenborn、Xiaohua Zhai、Thomas Unterthiner、Mostafa Dehghani、Matthias Minderer、Georg Heigold、Syl- vain Gelly、Jakob Uszkoreit 和 Neil Houlsby。一张图像值16×16个词：用于大规模图像识别的Transformer。*ArXiv*，abs/2010.11929，2020年。[12] M. Everingham、L. Van Gool、C. K. I. Williams、J. Winn 和 A. Zisserman。PASCAL视觉目标类别（VOC）挑战赛。《国际计算机视觉杂志》，88(2):303–338，2010年6月。[13] Chengjian Feng、Yujie Zhong、Yu Gao、Matthew R Scott 和 Weilin Huang。TOOD：任务对齐的一阶段目标检测。载于*2021 IEEE/CVF International Conference on Computer Vision(ICCV)*，第3490–3499页。IEEE计算机学会，2021年。[14] Chengjian Feng、Yujie Zhong、Zequn Jie、Xiangxiang Chu、Haibing Ren、Xiaolin Wei、Weidi Xie 和 Lin Ma。Prompt-det：利用未经筛选的图像实现开放词汇检测。载于欧洲计算机视觉会议，2022年。[15] Golnaz Ghiasi、Xiuye Gu、Yin Cui 和 Tsung-Yi Lin。利用图像级标签扩展开放词汇图像分割。载于欧洲计算机视觉会议，2021年。[16] Zhangxuan Gu、Siyua n Zhou、Li Niu、Zihan Zhao 和 Liqing Zhang。面向零样本语义分割的上下文感知特征生成。第*28*届*ACM*国际多媒体会议论文集，2020年。[17] Kaiming He、X. Zhang、Shaoqing Ren 和 Jian Sun。用于图像识别的深度残差学习。*2016 IEEE Confer-ence on Computer Vision and Pattern Recognition (CVPR)*，第770–778页，2015年。[18] Sergey Ioffe 和 Christian Szegedy。批归一化：通过减少内部协变量偏移加速深度网络训练。*ArXiv*，abs/1502.03167，2015年。[19] Chao Jia、Yinfei Yang、Ye Xia、Yi-Ting Chen、Zarana Parekh、Hieu Pham、Quoc V. Le、Yun-Hsuan Sung、Zhen Li 和 Tom Duerig。利用噪声文本监督扩展视觉与视觉语言表征学习。载于国际机器学习会议，2021年。

[20] Menglin Jia、Luming Tang、Bor-Chun Chen、Claire Cardie、Serge J. Belongie、Bharath Hariharan 和 Ser Nam Lim。视觉提示调优。*ArXiv*，abs/2203.12119，2022年。[21] Chen Ju、Tengda Han、Kunhao Zheng、Ya Zhang 和 Weidi Xie。通过提示视觉语言模型实现高效视频理解。载于欧洲计算机视觉会议，2021年。[22] Alexander Kirillov、Eric Mintun、Nikhila Ravi、Hanzi Mao、Chloe Rolland、Laura Gustafson、Tete Xiao、Spencer White- head、Alexander C. Berg、Wan-Yen Lo、Piotr Doll´ar 和 Ross Girshick。分割一切。*arXiv:2304.02643*，2023年。[23] Boyi Li、Kilian Q. Weinberger、Serge J. Belongie、Vladlen Koltun 和 Ren´e Ranftl。语言驱动的语义分割。*ArXiv*，abs/2201.03546，2022年。[24] Feng Li、Hao Zhang、Huaizhe Xu、Shilong Liu、Lei Zhang、Lionel M Ni 和 Heung-Yeung Shum。Mask dino：面向目标检测与分割的统一Transformer框架。载于*IEEE/CVF*计算机视觉与模式识别会议论文集，第3041–3050页，2023年。[25] Xiaoya Li、Xiaofei Sun、Yuxian Meng、Junjun Liang、Fei Wu 和 Jiwei Li。用于数据不平衡NLP任务的Dice损失。载于语言学计算学会年会，2019年。[26] Feng Liang、Bichen Wu、Xiaoliang Dai、Kunpeng Li、Yinan Zhao、Hang Zhang、Peizhao Zhang、P´eter Vajda 和 Diana Marculescu。利用掩码适配的CLIP实现开放词汇语义分割。*ArXiv*，abs/2210.04150，2022年。[27] Tsung-Yi Lin、Priya Goyal、Ross B. Girshick、Kaiming He 和 Piotr Doll´ar。用于密集目标检测的焦点损失。*IEEE*模式分析与机器智能汇刊，42:318–327，2017年。[28] Tsung-Yi Lin、Michael Maire、Serge J. Belongie、James Hays、Pietro Perona、Deva Ramanan、Piotr Doll´ar 和 C. Lawrence Zitnick。Microsoft COCO：上下文中的常见物体。载于欧洲计算机视觉会议，2014年。[29] Chao Ma、Yu-Hao Yang、Yanfeng Wang、Ya Zhang 和 Weidi Xie。使用冻结的视觉语言模型进行开放词汇语义分割。载于英国机器视觉会议，2022年。[30] Roozbeh Mottaghi、Xianjie Chen、Xiaobai Liu、Nam-Gyu Cho、Seong-Whan Lee、Sanja Fidler、Raquel Urtasun 和 Alan Yuille。上下文在真实场景中目标检测与语义分割中的作用。载于*IEEE*计算机视觉与模式识别会议 *(CVPR)*，2014年。[31] Jordi Pont-Tuset、Jasper Uijlings、Soravit Changpinyo、Radu Soricut 和 Vittorio Ferrari。通过Localized Narratives连接视觉与语言。载于*ECCV*，2020年。[32]Alec Radford、Jong Wook Kim、Chris Hallacy、Aditya Ramesh、Gabriel Goh、Sandhini Agarwal、Girish Sastry、Amanda Askell、Pamela Mishkin、Jack Clark、Gretchen Krueger 和 Ilya Sutskever。从自然语言监督中学习可迁移视觉模型。载于国际机器学习会议，2021年。[33] Evan Shelhamer、Jonathan Long 和 Trevor Darrell。用于语义分割的全卷积网络。*2015IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*，第3431–3440页，2014年。[34] Ashish Vaswani、Noam M. Shazeer、Niki Parmar、Jakob Uszkoreit、Llion Jones、Aidan N. Gomez、Lukasz Kaiser 和 Illia Polosukhin。Attention Is All You Need。*ArXiv*，abs/1706.03762，2017年。[35] Yuxin Wu、Alexander Kirillov、Francisco Massa、Wan-Yen Lo 和 Ross Girshick。Detectron2。https://github.com/facebookresearch/detectron2，2019年。[36] Yongqin Xian、Subhabrata Choudhury、Yang He、Bernt Schiele 和 Zeynep Akata。面向零标签和少标签语义分割的语义投影网络。*2019*年*IEEE/CVF*计算机视觉与模式识别会议 *(CVPR)*，第8248–8257页，2019年。[37] Jiarui Xu、Sifei Liu、Arash Vahdat、Wonmin Byeon、Xi- aolong Wang 和 Shalini De Mello。基于文本到图像扩散模型的开放词汇全景分割。*arXiv*预印本 *arXiv:2303.04803*，2023年。[38] Mengde Xu、Zheng Zhang、Fangyun Wei、Yutong Lin、Yue Cao、Han Hu 和 Xiang Bai。基于预训练视觉语言模型的开放词汇语义分割简单基线。载于欧洲计算机视觉会议，2021年。[39] Yuhui Yuan、Lang Huang、Jianyuan Guo、Chao Zhang、Xilin Chen 和 Jingdong Wang。OCNet：用于语义分割的目标上下文。《国际计算机视觉杂志》，129:2375–2398，2021年。[40] Chaoning Zhang、Dongshen Han、Yu Qiao、Jung Uk Kim、Sung-Ho Bae、Seungkyu Lee 和 Choong Seon Hong。更快地分割一切：面向移动应用的轻量级SAM。*arXiv*预印本 *arXiv:2306.14289*，2023年。[41] Shifeng Zhang、Cheng Chi、Yongqiang Yao、Zhen Lei 和 Stan Z Li。通过自适应训练样本选择弥合基于锚框与无锚框检测之间的差距。载于*IEEE/CVF*计算机视觉与模式识别会议论文集，第9759–9768页，2020年。[42] Xiaosong Zhang、Fang Wan、Chang Liu、Rongrong Ji 和 Qixiang Ye。Freeanchor：学习匹配锚框以进行视觉目标检测。神经信息处理系统进展，32，2019年。[43] Hengshuang Zhao、Jianping Shi、Xiaojuan Qi、Xiaogang Wang 和 Jiaya Jia。金字塔场景解析网络。*2017*年*IEEE*计算机视觉与模式识别会议 *(CVPR)*，第6230–6239页，2016年。[44] Xu Zhao、Wenchao Ding、Yongqi An、Yinglong Du、Tao Yu、Min Li、Ming Tang 和 Jinqiao Wang。快速分割一切，2023年。[45] Bolei Zhou、Hang Zhao、Xavier Puig、Sa nja Fidler、Adela Barriuso 和 Antonio Torralba。通过ADE20K数据集进行场景解析。载于*IEEE*计算机视觉与模式识别会议论文集，第633–641页，2017年。[46] Bolei Zhou、Hang Zhao、Xavier Puig、Tete Xiao、Sanja Fidler、Adela Barriuso 和 Antonio Torralba。通过ADE20K数据集实现对场景的语义理解。《国际计算机视觉杂志》，127:302–321，2019年。[47] Kaiyang Zhou、Jingkang Yang、Chen Change Loy 和 Zi- wei Liu。为视觉语言模型学习提示。《国际计算机视觉杂志》，130:2348，2021年2337：–。

附录

A. 类别无关掩码提议网络

本节将详细介绍掩码提议网络。我们采用改进版的 MaskFormer [9] 作为掩码提议网络，主论文第 3.1节对此有所提及。图6展示了类别无关掩码提议网络的整体框架。该模型架构与 MaskFormer 基本相同，但移除了分类分支。我们使用骨干网络提取图像特征 F_B。图像特征输入像素解码器和 Transformer 解码器，以生成 N 掩码嵌入和逐像素嵌入 F_P。最后，我们通过矩阵乘法组合 N 掩码嵌入和 F_P，得到掩码 M。第 C.2 节评估了提议掩码的质量。

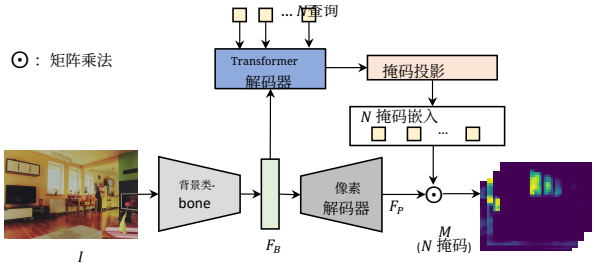


图6：类别无关掩码提议网络的架构。

B. 视觉提示学习

正如正文第 3.2 节所述，我们通过使用提示学习微调预训练模型，提升了基线方法的性能。正文已详细介绍文本提示学习，因此这里我们将进一步阐述所采用的视觉提示学习。视觉提示学习的架构如图7所示。我们比较了两种视觉提示学习形式：前置提示和添加提示。这两种形式的区别在于提示与图像块嵌入的结合方式。 *Prepending prompts* 指在 O 图像块嵌入之前前置 P提示嵌入，从而得到 $P+O$ 嵌入，并与一个类别嵌入结合。*Adding prompts* 则以逐元素方式，将一个提示嵌入添加到每个图像块嵌入中，并且提示嵌入的长度应与图像块嵌入的长度相同。

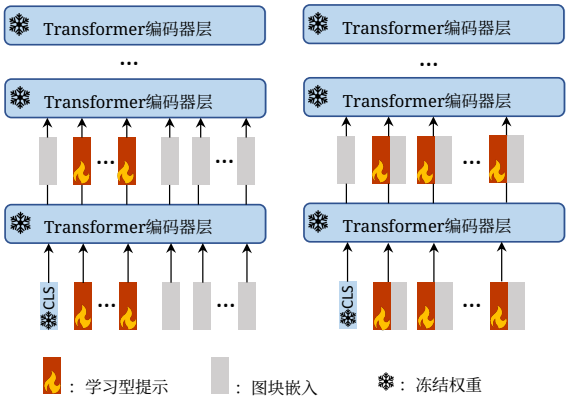


图7：视觉提示调优概述。有两种实现方式：前置提示（**左侧**）和添加提示（**右侧**）。

本节首先介绍数据集和评估指标的详细信息，随后进一步分析我们的方法，包括对不同提示学习方法的消融研究、对 Pascal VOC 数据集上（主要组件）有效性的分析，以及掩码提议网络的性能。

B.1. 数据集与评估指标

数据集。COCO-Stuff是一个大型语义分割数据集，涵盖171个类别，包括80个物体类别和91个背景类。它包含117k张训练图像和5k张验证图像。**PASCAL VOC** 包含来自20个类别的11,185张训练图像和1,449张验证图像。**PAS-CAL Context** 是一组针对 PAS- 的补充标注 CAL VOC 2010。它包含 4,998 张训练图像和 5,005 张验证图像。我们选取了一个包含 59 个常见类别的子集供使用。**ADE20K** 包含超过 20K 张以场景为中心的训练图像和 2k 张验证图像。总共有 150 个语义类别，包括背景类和离散物体。

C. 实验

数据划分。我们选择两种数据划分，分别用于在零样本语义分割（ZS3）设置和跨数据集设置下验证我们的方法。对于 ZS3 设置，我们遵循文献 [2]中的类别划分。具体而言，在 COCO-Stuff 上，我们选择 156 个类别作为已见类别，其余 15 个类别作为未见测试类别。在 Pascal VOC 2012 上，我们选择 15 个类别作为已见类别，其余 5 个类别作为未见测试类别。对于跨数据集设置，我们在 COCO-Stuff 已见类别数据集上训练模型，并在其他数据集上进行验证。

评估指标。 遵循以往工作，我们测量已见类别和未见类别的逐像素分类准确率（pAcc）与平均交并比（mIoU），分别记为 mIoU(S) 和 mIoU(U)。此外，沿用以往工作的做法 [36], ，我们计算已见类别和未见类别的交并比调和平均值（hIoU），其计算方式为 $hIoU = \frac{2*mIoU(S)*mIoU(U)}{mIoU(S)+mIoU(U}$。对于跨数据集验证，我们使用 mIoU 作为评估指标。



图 8：热力图的可视化。四行中涉及的三个类别分别为：绵羊（已见）、大象（已见）、滑板（**未见**）和道路（**未见**）。

Type	端到端	pAcc	hIoU	Seen	**未见**
Text	✗	15.6	7.0	6.2	8.0
Text	✓	25.7	7.2	7.6	6.9
Vision A.	✓	22.6	6.4	6.3	6.5
Vision P.	✓	16.7	6.7	6.4	7.4
Both	✓	**27.6**	**8.3**	**8.0**	**8.7**

表7：提示学习的消融研究。 *Vision P.*和 *Vision A.* 分别表示在视觉提示调优中前置提示和添加提示。 *Both* 指将端到端文本提示学习与添加视觉提示学习相结合。

C.1. 提示学习

提示学习分析。 在表7中，我们评估基于基线方法的不同提示学习方案在分割任务中的有效性。论文正文第3.2节提出通过提示学习改进基线方法。为评估提示学习的影响，我们基于基线方法，使用不同类型的提示开展实验。

方法	All	Seen	**未见**
Recall@30	0.59	0.57	0.71
Recall@50	0.45	0.44	0.56

表 8：COCO-Stuff 数据集上的掩码候选召回率。*Recall*@30 和 *Recall*@50 分别表示交并比为 30% 和 50% 时的召回率。

方法	层	hIoU	Seen	**未见**
基线	-	7.0	6.2	8.0
GPS	[1]	4.6	4.5	4.7
GPS	[1-6]	-	0.0	0.1
GPS	[10-12]	1.9	1.3	3.3

表 9：广义图块分离的消融研究。

行方法。我们可以观察到，所有提示学习方法都能提升模型性能。将端到端文本提示学习与添加视觉提示相结合时，取得了最佳结果。

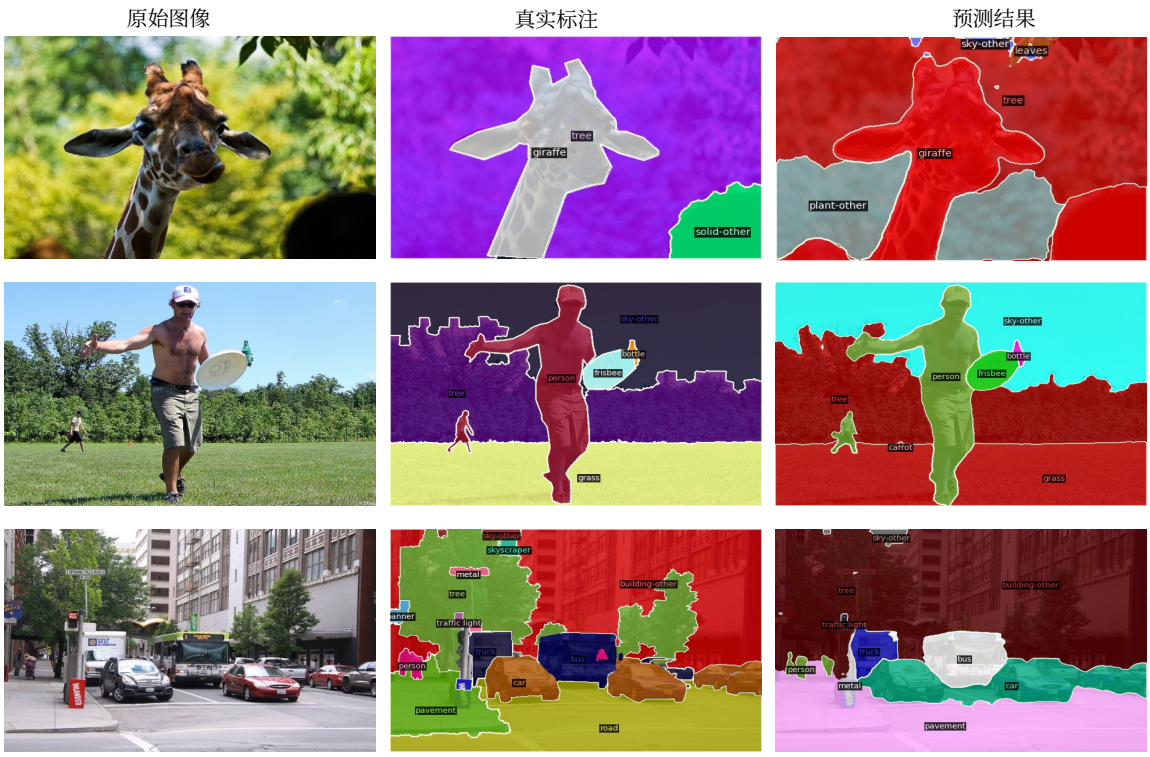


图 9：语义分割的可视化。从左到右依次为：原始图像、真实标注语义分割图和预测结果。

C.2. 类别无关掩码提议网络的泛化能力

为了评估仅使用已见类别图像训练的掩码提议网络的泛化能力，我们在表 8 中报告了该网络在 COCO-Stuff 数据集上的掩码召回率。我们计算了两个指标： $Recall@30$ 和 $Recall@50$。结果表明，掩码提议网络能够为已见类别和未见类别提供类别无关掩码。该网络表现出令人满意的泛化性能。

D. 可视化

热力图可视化。 我们在图 8 中展示了热力图的可视化结果，这些是主论文第 5.3 节的补充案例。Transformer编码器层中的自注意力会从全局范围检索信息，这实际上可能引入更多噪声，因此不利于分割任务。$sheep$ 类、 $elephant$ 类和 $skateboard$ 类属于“物体”类别，而 $road$归为“背景类”。这表明我们的方法在“物体”和“背景类”类别上都能取得良好效果。CAL 模块能够引导模型关注更具辨别性的区域，从而增强其准确分类不同类别的能力。

分割结果。 如图 9 所示，我们展示了 DeOP 的可视化预测结果。我们的方法能够有效分割属于不同类别的区域。即使面对未见类别（长颈鹿、树、飞盘 和 草）的区域，该方法也能取得出色的结果，体现了其在零样本分割任务中的卓越泛化能力和有效性。