使用语言分割任何 3D 对象

李承俊 [1], Yuyang Zhao [2], 2、Gim Hee Lee

[1] 韩国大学 新加坡国立
[2] 大学https://cvrp-sole.github.io

(a)“我可以洗手吗？” (b)“玩游戏的设备。” (c)“我想看看外面。”

图 1:使用各种语言指令查询 SOLE 时的定性结果。 SOLE 具有高度的通用性,可以使用各种语言指令来分割相应的实例,包括但不限于 (a) 视觉问题、(b) 属性描述和 (c) 功能描述。

摘要。在本文中,我们研究了具有自由形式语言指令的开放词汇 3D 实例分割 (OV-3DIS)。早期的研究仅依赖带注释的基本类别进行训练,因此对未见过的新类别的泛化能力有限。最近的研究通过生成与类别无关的掩码或将广义掩码从 2D 投影到 3D 来缓解对新类别的不良泛化能力,但忽略了语义或几何信息,导致性能不佳。相反,直接从 3D 点云生成可泛化但与语义相关的掩码将产生更好的结果。在本文中,我们引入了使用语言分割任何 3D 对象 (SOLE),这是一个语义和几何感知的视觉语言学习框架,通过直接从 3D 点云生成与语义相关的掩码具有很强的泛化能力。具体来说,我们提出了一个多模态融合网络,将多模态语义纳入主干和解码器中。此外,为了将 3D 分割模型与各种语言指令对齐并提高掩码质量,我们引入了三种类型的多模态关联作为监督。我们的 SOLE 在 ScanNetv2、ScanNet200 和 Replica 基准上的表现远远优于以前的方法,尽管训练中没有类注释,但结果甚至接近全监督方法。此外,大量的定性结果证明了我们的 SOLE 对语言指令的通用性。

关键词:开放集·3D实例分割·多模态

平等贡献。

1 简介

3D实例分割旨在检测、分割和识别
三维场景中的物体实例分割是三维场景理解的关键任务之一。有效且可推广的三维实例
分割在三维场景理解中具有巨大的潜力。
现实世界的应用，包括但不限于自动驾驶、增强现实 (AR) 和虚拟现实 (VR)。由于其重要性，3D

实例分割在计算机视觉领域取得了显著的成功
社区[20, 53, 60]。以前的3D实例分割模型主要关注封闭集设置，其中训练和测试阶段共享

相同的类别。然而，形状各异的新颖和看不见的类别，
语义含义在实际应用中是不可避免的。无法将语义信息分割成
这种情况大大缩小了应用范围。

鉴于闭集设置的强烈局限性，开集三维实例
分割（OS-3DIS），旨在检测和分割看不见的类别
根据说明在社区中引入和调查。大多数
一些研究[10, 26, 44]利用类别名称或描述作为分割指令，这也被称为开放词汇 3D 实例分割
(OV-3DIS)。早期的方法[10, 11, 66]将类别拆分为每个

将数据集分为基础集和新集。训练中只有基础类别可用
阶段，但模型有望在推理过程中细分出新的类别。
由于训练过程中缺乏新的类别，这些方法很容易过度拟合
基础类别，因此在新类别上表现不佳。此外，当它们

在不同分布的数据上进行评估。在这方面，最近
论文[26, 40, 58, 64]借助 2D
基础模型[45,50,76]。具体来说， [26,58]学习与类别无关的 3D 掩模
从掩码注释中获取点云，然后将点云投影到二维图像，从基础模型中获得类标签。 [40,64]使用
二维图像预测二维实例
开放词汇实例分割模型[76]并将它们融合以获得 3D
预测。然而，类无关掩码和 2D 投影掩码忽略了
掩码生成中的语义和几何信息分别导致
达到次优性能。相比之下，我们直接预测语义相关的
来自 3D 点云的掩模，产生更好、更通用的 3D 掩模。

在本文中，我们提出 SOLE：使用语言分割任何 3D 对象 _ _ _
规避上述 OV-3DIS 问题。实现可推广的
开放集三维实例分割，我们的 SOLE 需要两个主要属性：直接从三维点云生成和分类三维掩模，以及
响应
到自由形式的语言指令。3D分割网络需要
与语言指令保持一致，直接对实例进行分段和分类
从点云中提取。为此，我们构建了一个多模态融合网络，其中包含两个
主要技术：1)将从预训练的多模态二维语义分割模型[14]获得的逐点 CLIP 特征合并到主干网络中，

增强模型的通用性；2)引入跨模态解码器
整合语言领域特征信息，促进多模态知识的有效融合。此外，我们提高了泛化能力。

使用新颖的视觉语言学习框架,在各种场景和语言指令中实现 3D 分割,使用三种类型的多模态关联训练 3D 分割网络:1)掩模-视觉关联,2)掩模-标题关联和 3)掩模-实体关联。这些关联改善了语言指令对齐,并通过更丰富的语义信息增强了 3D 掩模预测。

我们的视觉语言学习框架 (SOLE) 配备了多模态融合网络和三种多模态关联,在 ScanNetv2 [8]、ScanNet200 [52]和 Replica [56]基准上的表现远远优于以前的成果。此外,SOLE 可以响应自由形式的查询,包括但不限于问题、属性描述和功能描述（图1和图4）。总之,我们的贡献如下:

– 我们为 OV-3DIS 提出了一个视觉语言学习框架 SOLE。为 SOLE 设计了一个多模态融合网络,该网络可以直接从具有多模态信息的 3D 点云中预测与语义相关的掩码,从而产生高质量且可通用的片段。

– 我们提出了三种类型的多模态关联,以改善 3D 分割模型与语言之间的对齐。这些关联提高了掩码质量和对语言指令的响应能力。

– SOLE 在 ScanNetv2、Scannet200 和 Replica 基准上取得了最佳结果,结果甚至接近全监督的对应结果。此外,大量定性结果表明 SOLE 可以回答各种语言问题和指令。

2 相关工作

闭集 3D 实例分割。3D 实例分割旨在检测、分割和识别 3D 场景中的对象实例。先前的研究[4、12、19、20、23、30、37、43、53、57、60、61、65、68、72]主要考虑闭集设置,其中训练和测试类别相同。这些方法在特征提取和解码过程上有所不同。随着Transformer模型的发展,掩码预测成为一种比传统框检测解码方法更高效、更有效的方法。Mask3D [53]从场景中采样固定数量的点作为查询,然后直接使用注意机制预测最终的掩码,从而获得更好的结果。然而,无论采用何种解码方法,闭集方法都缺乏处理看不见的类别的能力,从而阻碍了它们在现实世界中的应用。

开放词汇 2D 分割。由于最近大规模视觉语言模型的成功[1、5、15、28、50、69、70] ,开放词汇或零样本2D分割取得了显著成就[6、9、14、16、18、32、33、35、41、51、62、63、71、75] 。共同的关键思想是利用 2D 多模态基础模型[28、50]将图像级嵌入转移到像素级下游任务。LSeg [33]、 OpenSeg [14]和OVSeg [35]将像素级或掩码级视觉特征与基础模型中的文本特征对齐

用于开放词汇语义分割。其他工作如 X-Decoder [78]，
FreeSeg [49]和 SEEM [79]提出了更统一的开放词汇框架
分割，包括实例分割、全景分割和引用分割。

开放词汇 3D 场景理解。取得了显著的成功
开放词汇二维分割（OV-2DS)的进展激发了
开放词汇 3D 分割。然而，OV-2DS 中的技术无法
由于缺乏三维多模态基础模型，无法直接将二维图像和三维点云迁移到三维领域。因此，研究人员提出将二维图像和三维点云对齐，
云，从而将二维基础模型提升到三维。对于开放词汇三维语义分割， [2,10,17,25,27,46,54,55]构建与任务无关的
逐点
从二维基础模型[50]中提取特征表示，然后使用这些特征来查询三维场景中的开放词汇概念。这些工作重点

单纯将语义信息从二维转移到三维，限制了实例级识别任务的应用。在这方面，开放词汇三维

引入实例分割（OV-3DIS） [11, 26, 40, 44, 58, 64]来检测
并在 3D 场景中分割出各种类别的实例。PLA [10]及其变体[11,66]将训练类别分为基础类别和新类别，并训练

仅带有基类注释的模型。OpenMask3D [58]和 OpenIns3D [26]
从掩码注释中学习与类别无关的 3D 掩码，然后使用相应的 2D 图像从基础模型中获取类别标签。最近，

研究人员还研究了无需训练即可将二维实例分割模型[76]的二维预测直接提升到三维[40, 64]。先前的研究极大地

促使 OV-3DIS 得到改进。但由于语义泛化能力较差、掩模质量不高，结果还不尽如人意

预测。考虑到以前工作的局限性，我们显著提高了
OV-3DIS 通过设计一个具有多模式的视觉语言学习框架
网络和各种多模式关联。

3 方法

目的。开放词汇 3D 实例分割（OV-3DIS)的目标
使用自由格式语言指令定义如下：给定一个 3D 点云
P ∈ RM×C对应的二维图像 I 和实例级三维蒙版 m，
我们的目标是训练一个没有真实类别的 3D 实例分割网络
注释。在推理过程中，给定一个文本提示 q，训练后的 3D 实例
分割网络必须检测并分割相应的实例。

掩码预测基线。我们在基于 Transformer 的框架上构建了我们的框架
3D 实例分割模型 Mask3D [53]，将实例分割任务视为掩码预测范式。具体来说，Transformer

使用带有掩码查询的解码器来分割实例。给定从场景中选择的Nq 个查询，使用交叉注意来聚合来自

点云到实例查询。经过几个解码器层之后， Nq个查询变为
Nq 个掩码及其对应的语义预测。在训练过程中，匈牙利
采用匹配[31]来匹配和训练模型与真实标签

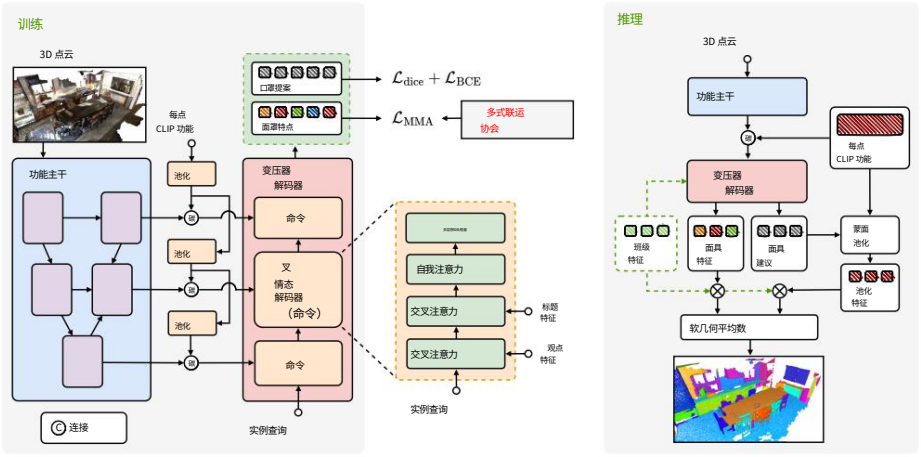

图 2:SOLE 的整体框架。SOLE 建立在具有多模态自适应性的基于 Transformer 的实例分割模型上。对于模型架构,骨干特征与每点 CLIP 特征集成,然后输入跨模态解码器 (CMD)。CMD 将逐点特征和文本特征聚合到实例查询中,最终分割实例,这些实例由多模态关联监督。在推理过程中,预测的掩码特征与每点 CLIP 特征相结合,增强了开放词汇的性能。

和掩码。在推理阶段,将具有正确语义分类结果的Nq 个掩码作为最终输出。我们的 SOLE 利用基于 Transformer 架构的掩码预测范式,其中模型仅使用没有真实标签的掩码进行训练,以实现可推广的 OV-3DIS。

概述。 SOLE 的整体架构如图 2 所示。为了实现具有自由形式语言指令的开放词汇实例分割,我们使用多模态信息改进了基于 Transformer 的实例分割模型:主干中的逐点 CLIP 特征（第3.1 节)和解码器中的文本信息（第3.2 节）。此外,为了在没有真实类别标签的情况下实现更好的泛化能力,我们在目标实例上构建了三种类型的多模态关联:掩码-视觉关联、掩码-标题关联和掩码-实体关联来训练 SOLE。借助多模态框架和关联,我们的 SOLE 可以根据各种语言提示有效地分割实例。

3.1 主干特征集成

使用预训练模型[29, 73, 74]初始化主干是提高下游任务性能的有效方法,尤其是在下游数据不丰富的情况下。对于 3D 开放集设置,由于 3D 数据有限,利用 2D 基础模型至关重要。因此,我们遵循[46],根据相机姿势将 2D 图像的预训练视觉特征投影到 3D 点云。为了保持细粒度和可推广的特征,我们使用 2D 基础模型来构建 3D 点云。

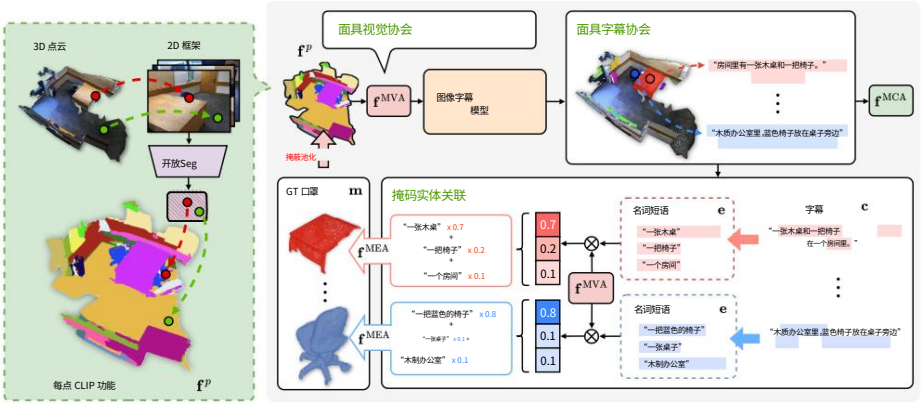

图 3:三种类型的多模态关联实例。对于每个真实实例掩码,我们首先汇集每个点的 CLIP 特征以获得掩码-视觉关联 f MVA。随后,f 标题和相应的文本特征 f 关联。最后,从掩码标题中提取名词短语,并通过多模态注意聚合它们的嵌入以获得掩码-实体关联 f MEA。这三种多模态关联 被输入到 CLIP 空间字幕模型中,为每个掩码生成字幕,称为 Mask-Caption 用于监督 SOLE,以获得使用自由格式语言指令分割 3D 对象 的能力。

为了实现这些目标,我们利用 OpenSeg [14]作为 2D 主干。这些特征包含 CLIP [50]特征空间中的视觉信息,与文本信息对齐。

由于图像级对比训练导致 CLIP 特征空间主要关注语义信息,因此仅利用投影特征无法在实例分割上取得最佳性能。为此,我们用 3D 实例分割主干训练 a ∈ RM×D并将其特征 f 与投影的 2D CLIP 特征 f p ∈ RM×C相结合:

(1)

其中 M 表示点的数量,D 和 C 表示 3D 实例分割主干的特征维数和投影的 2D 特征,
分别。注意从 3D 主干中提取不同分辨率的特征,并分别与 2D CLIP 特征合并。如图 2 所示,采用与 3D 主干相同的池化策略来
CLIP 特征,对齐分辨率。最后,合并逐点特征
具有多种分辨率的图像被输入到跨模态解码器中。

3.2 跨模态解码器(CMD)

投影 2D CLIP 特征提供了可泛化的视觉信息,但语言信息未明确集成,限制了对语言指令的响应能力。为了解决这个问题,我们引入了跨模态

解码器（CMD)将文本信息纳入我们框架的解码过程中。具体来说,每个 CMD 模块包含三个注意层。

实例查询首先从 CLIP 组合的主干特征 f b中提取视觉信息。然后,CLIP 文本特征被投影到第二个注意层的键和值上,并结合文本领域知识。在训练过程中,CLIP 文本特征是从每个目标掩码的标题特征中获得的， fMCA ∈ R Nc×C （详情见第3.3节）,而在推理过程中,它可以是查询实例的描述或其他形式的语言指令,如视觉问题或功能属性。最后,将自注意力应用于实例查询以进一步改进表示。通过将来自 CLIP 的多模态知识与多级 CMD 作为解码器融合,SOLE 可以对各种语言指令做出高质量的响应。

3.3 视觉语言学习

我们进行视觉语言学习,以使我们的 SOLE 能够实现可推广的 OV-3DIS。为了有效地响应各种语言指令,我们利用来自目标掩码注释的多模态信息来监督分割网络。具体来说,提出了三种层次粒度的监督类型:1) 掩码-视觉关联,2) 掩码-标题关联和 3) 掩码-实体关联。

掩码-视觉关联 (MVA)。利用二维图像与三维点云之间的对应关系,我们可以通过对Nm个目标实例掩码 m = [m1, m2, . . . , mNm]内的每个点 CLIP 特征取平均值来获得实例级 CLIP 视觉特征MVA f 。实例级 CLIP 视觉特征可以作为监督,间接将三维分割模型与 CLIP 文本空间对齐。此外,作为三维点云与语言之间的中间表示,掩码-视觉关联也是以下两个细粒度关联的基础。∈R Nm×C

掩码-字幕关联 (MCA)。尽管位于 CLIP 特征空间中,但掩码-视觉关联并不是一种准确且精确的语言监督。

相反,直接用语言指令监督模型会产生更好的结果。由于 CLIP [50] 具有很强的泛化能力,社区中广泛研究了从 CLIP 空间生成文本[34,42,59]。由于MVA掩模视觉关联中的实例级 CLIP 视觉特征 f 位于 CLIP 视觉空间中,我们可以将它们输入到 CLIP 空间字幕生成模型（DeCap [34])以获得掩模字幕 c = [c1, c2, ..., cNm]。然后将掩模字幕输入到 CLIP 文本模型中以提取掩模字幕关联 f MCA。此关联表示实例掩模的语言信息,在训练期间用于 CMD 融合文本信息。

掩码实体关联（MEA）。虽然掩码标题关联可以为语义和几何提供详细的语言描述,但对于特定类别来说,它可能会产生歧义。如图3的示例所示。

桌子的掩码标题是“房间里的一张木桌和一把椅子”。这样的标题可能会导致模型混淆椅子和桌子,或者将两个实例误解为一个。因此,引入更细粒度的视觉语言关联对于更好的语义学习非常重要。

由于对象通常是标题中的名词,我们可以提取名词的实体级描述并将其与实例进行匹配。

具体来说,如图 3 所示,我们首先提取每个掩码字幕ci的所有名词短语ei ,并从 CLIP 文本编码器 T 中获取每个名词短语的文本特征,如下所示:

(2)

其中 E(·) 表示提取名词短语的 NLP 工具， Ni表示从掩码标题ci中获得的名词数量。实体可以以硬匹配或和 表示
软匹配的方式与掩码匹配。直观地,最相似的实体可以看作是掩码标签。然而,这种硬匹配存在两个主要问题。首先,生成的标题和相似度结果可能不准确,导致错误的监督。其次,虽然实体是正确的,但硬匹配忽略了上下文中的几何信息,从而削弱了对语言指令的响应能力。为此,我们提出了一种软匹配,通过多模态注意来获取掩码-实体关联。

具体来说,根据掩码特征和实体特征之间的注意力图Ac,e获得第 i 个掩码 f 的聚合实体特征： 事物
ʁ

(3)

中 f MVA,其 表示第 i 个掩码的掩码视觉关联特征,f ek是第 i 个掩码标题中第 k 个实体的 CLIP 文本特征。利用聚合实体特征,可以将 3D 掩码与特定实例类别对齐。
ʁ

3.4 训练和推理

训练。三种类型的多模态关联是学习可推广的 3D 实例分割模型的有效监督。我们遵循掩码预测范式来训练分割模型,该模型通过匈牙利匹配将真实实例与预测的掩码进行匹配[31]。具体而言,第 i 个预测掩码与第 j 个真实实例之间的匹配成本计算如下:

(4)

其中 p(·, ·) 表示预测实例与真实实例之间的 softmax 概率。在匹配掩码和真实实例后,

模型使用 mask 和语义损失的组合进行训练。具体来说，所有三种类型的关联都用于对模型进行语义监督。对于每个关联，我们遵循[77]使用 focal loss [36]和 dice loss的组合，这可以确保每个类的分割结果是独立生成的。第 j 个真实值的语义多模态关联损失Lj

好的

掩码是：

(5)

其中 a ∈ {MVA, MCA, MEA} 表示三种类型的关联，y = sigmoid(fm表示匹配的二进制标签。预测与关联 a 之间的 p a 杰 是失是 mask 损失和语义损失的组合： σ(j) σ(j) 。整体训练损) 是语义概率 · f j

(6)

其中 m σ(j)表示与第 j 个目标掩码匹配的预测掩码。

推理。在推理过程中，我们将 CLIP 的视觉特征与预测的掩码特征相结合，以实现更好的泛化能力。具体来说，在获得 3D 掩码后，将每个点的 CLIP 特征在掩码内进行池化。

然后将池化的 CLIP 特征和 mask 特征输入到分类器中，得到各自的分类概率 p(f m) 和 p(f p)，最终概率由它们之间的软几何平均得出：

(7)

其中τ是增加置信度的指数，本文将其设置为0.667。对于基准评估，我们使用所有类别名称的CLIP文本特征作为分类器。对于响应其他语言指令，我们使用相应语言指令的CLIP文本特征作为二元分类器。

4 实验

4.1 实验设置

数据集。我们在流行的场景理解数据集 ScanNetv2 [8]、ScanNet200 [52]和 Replica [56]上对 SOLE 进行了评估，包括闭集和开集 3D 实例分割任务。ScanNetv2 [8]是一个流行的室内点云数据集，有 18 个实例类，其中“其他家具”类由于其模糊性而被忽略。ScanNet200 [52]是 Scan-Netv2 的细粒度注释版本，包含 200 个类的头部（66 个类别）、公共（68 个类别)和尾部（66 个类别)子集。对于 ScanNetv2 和 ScanNet200，我们评估了闭集设置和分层开放集设置。Replica [56]是一个高质量的合成数据集，带有 48 个实例类别注释。按照[58]的做法，我们在 Replica 中的八个场景中对开放集实例分割进行了评估，包括 {office0、office1、office2、office3、office4、room0、room1 和 room2}。

表 1: 闭集 3D 实例分割设置的比较
ScanNetv2 [8]。 SOLE 与类分割方法、掩码训练方法进行了比较
以及全监督对应模型（上限）。 SOLE 的表现优于所有 OV-DIS
方法并取得与全监督模型相媲美的结果。

方法	B/N AP AP50 AP25体素大小				
解放军[10]	10/7	-	21.9	-	2厘米
地区 PLC [66]	10/7	-	32.3	-	2厘米
低维斯3D [11]	10/7	-	31.2	-	2厘米
OpenIns3D [26]	-/7	-	27.9	42.6	2厘米
体素大小为 4cm 的 SOLE	-/7	31.6	58.5	72.5	4厘米
SOLE 无文本支持	-/7	41.1	57.1	65.9	2厘米
SOLE（我们的）	-/7	52.3 72.4 81.7			2厘米
解放军[10]	8/9	-	25.1	-	2厘米
地区 PLC [66]	8/9	-	32.2	-	2厘米
低维斯3D [11]	8/9	-	38.1	-	2厘米
OpenIns3D [26]	-/9	-	19.5	27.9	2厘米
体素大小为 4cm 的 SOLE	-/9	31.9	57.5	73.6	4厘米
SOLE 无文本支持	-/9	42.9	59.6	70.7	2厘米
SOLE（我们的）	-/9	50.4 68.3 75.2			2厘米
OpenIns3D [26] -/17 SOLE w 4cm体素大		-	28.7	38.9	2厘米
小 -/17 SOLE w/o文本支持 -/17 SOLE (我们		30.8	52.5	70.9	4厘米
的) -/17 44.4 62.2 71.4		35.0	50.2	60.2	2厘米
					2厘米
Mask3D [53] （完全支持） 17/-		55.2	73.7	83.5	2厘米

实施细节。继 Mask3D [53] 之后，我们采用 Minkowski-UNet [7]作为主干。特征主干提取 5 个尺度的点特征，
而 4 层 Transformer 解码器则迭代细化实例查询。我们的
模型使用 AdamW [39]优化器训练了 600 个 epoch 。学习率
设置为 $1 \times 10-4$ ，周期性衰减。在训练中，我们设置$\lambda MMA = 20.0$， $\lambda dice = 2.0$且$\lambda BCE = 5.0$作为损失权重。

基线。我们主要将 SOLE 与现有的两大流派进行比较
OV-3DIS：类分割方法[10, 11, 66]和掩码训练方法[26, 58]。
类别分割方法[10, 11, 66]将训练类别分为基础类别和新类别
类别。所有掩码注释和基本类别标签均用于训练
模型。与这些方法相比，我们只在
掩码注释，并在拆分的新类别上与它们进行比较。掩码训练方法[26, 58]使用掩码注释训练与类别无关的掩码生成器，并使用 2D 基础模型获得语义预测。设置

的口罩训练方法与我们的方法类似，我们直接与它们进行比较
所有类别。