OpenIns3D:捕捉和查找 3D 开放词汇实例分割

Zhening Huang [1] Xiaoyang Wu [2] Xi Chen [2]
Hengshuang Zhao [2] Lei Zhu [3,4] 约翰·拉森比 [1]

1剑桥大学2香港大学
3香港科技大学
4香港科技大学（广州）

https://zheninghuang.github.io/OpenIns3D/

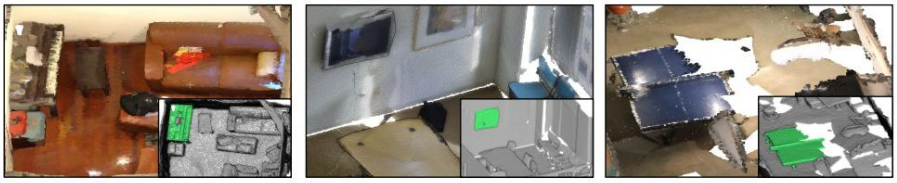


图 1:使用 OpenIns3D 进行复杂查询 3D 实例分割。

摘要。在本文中,我们引入了 OpenIns3D,一种新的仅支持 3D 输入的
3D 开放词汇场景理解框架。OpenIns3D
框架采用了“Mask-Snap-Lookup”方案。“Mask”模块
在 3D 点云中学习与类别无关的掩码提议，“Snap”模块生成多个尺度的合成场景级图像,并利
用 2D 视觉语言模型提取有趣的物体,并且

“查找”模块搜索“Snap”的结果,为建议的掩码分配类别名称。这种方法虽然简单,但实现了

在广泛的 3D 开放词汇中表现出色
在室内和室外数据集上,OpenIns3D 可轻松实现不同 2D 检测器之间的切换,而无需重新训练。
当与强大的 2D 开放世界模型集成时，

它在场景理解任务中取得了优异的成绩。此外，
当与 LLM 驱动的 2D 模型结合时,OpenIns3D 表现出
令人印象深刻的理解和处理高度复杂文本的能力
需要复杂推理和现实世界知识的查询。

关键词:开放词汇理解·3D场景理解·视觉语言模型

1 简介

3D 场景理解在自动驾驶、机器人传感、AR/VR 和制造业等各个领域发挥着关键作用。

通讯作者。

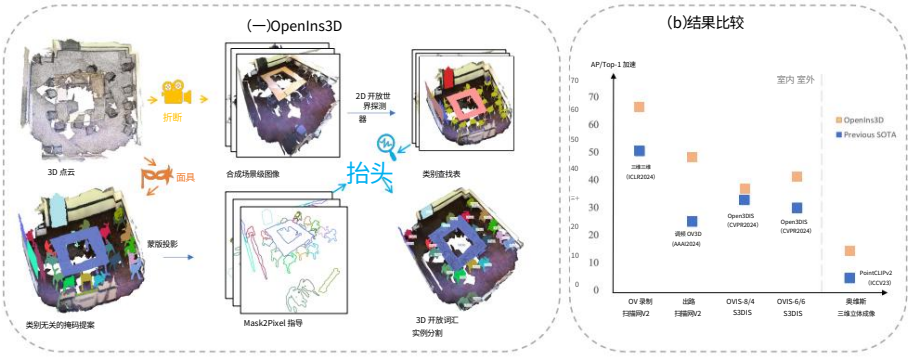


图 2:OpenIns3D 和定量结果的高级说明。(a)
OpenIns3D 遵循“Mask-Snap-Lookup”步骤来理解开放词汇场景。(b)在室内和室外数据集上都取得了一系列 SOTA 结果。OV-Rec:开放词汇对象识别。OVOD:开放词汇对象检测。OVIS:开放词汇实例分割。PointCLIPV2 [50]; Uni3D [46];Open3DIS [26];FM-OV3D [42]

虽然 3D 闭集理解的发展相对成熟,但开放词汇环境下的场景理解仍处于起步阶段。闭集理解只能处理一组预定义的概念和场景,但在面对不熟悉的概念或语言用法的变化时无法提供有效的响应。这一限制影响了其在动态和不断变化的环境中的表现。

得益于互联网规模的图像文本数据集,二维图像开放词汇理解取得了重大进展 [4,7,12,18,28,39,44,45]。

然而,与可以轻松从互联网上收集的 2D 数据不同,构建大规模 3D 文本数据集是一项挑战。因此,实现 3D 开放词汇理解最可行的方法是利用 2D 图像来连接语言和 3D 数据。在这个方向上,已经出现了一些值得注意的研究,例如 OpenScene [27]、PLA 系列 [10, 11, 40] 和 CLIP2Scene [6]。这些研究利用对齐良好的 2D 图像和 3D 点云进行特征提取或使用 2D 字幕模型来构建 3D 文本对。然而,这些方法的一个先决条件是拥有对齐良好的 2D 图像和 3D 点云。这意味着摆出的姿势的 2D 图像、相关的深度图和相机模型需要可作为网络的输入。在现实生活中,有许多情况是无法获得 2D 图像或对齐 2D 和 3D 数据所需的信息。例如,为了节省存储空间,使用 LiDAR 生成的点云通常不附带 2D 图像(示例数据集包括 [14, 30, 37])。在从不同传感器的多个扫描配准中获得点云或从 3D 模拟/CAD 模型转换而来的情况下 [13, 25],通常不存在 2D 图像。

我们相信,开发一个不依赖于对齐良好的二维图像的三维开放词汇框架是有意义的,因为这将简化部署前

要求并增强其在广泛场景中的适用性。为此
最后,我们介绍 OpenIns3D,一个旨在有效执行 3D 的框架
开放词汇场景理解任务,无需依赖 2D 对齐图像。总体而言,OpenIns3D 包含三个核心步骤:Mask、Snap 和 Lookup。
图 2a 给出了 OpenIns3D 的整体说明。

Mask:给定一个 3D 点云,OpenIns3D 的第一部分学习与类无关的
使用掩码提议模块 (MPM) 来生成掩码提议。这个过程经过训练
没有任何分类标签。为了控制掩码的质量,MPM 提出了一个可学习的掩码评分模块来预测每个掩码的质量
输出并实现一系列掩码过滤技术以丢弃无效的、
低质量掩码。MPM 输出场景中与类别无关的掩码列表。

Snap:生成多个经过校准的合成场景级图像,并
优化相机姿态和固有参数。这些图像专门
设计为覆盖部分或全部相关口罩,旨在尽量减少
需要多次渲染。而不是单独预测
每个掩码提议 [23, 43, 50],场景级图像被输入到 2D 开放词汇模型中,以同时理解所有有趣的对象

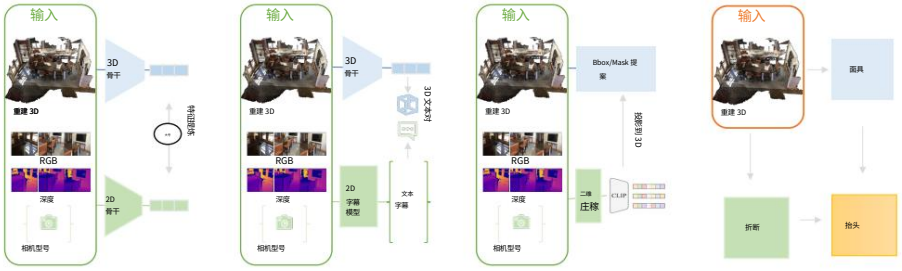
场景中存在的。然后构建一个类查找表 (CLT) 来存储
所有检测到的物体类别以及它们各自的像素位置。

查找:为了精确确定每个图像中 mask 提案的位置,
构建了 Mask2Pixel 地图。这些地图投射了所有 3D 掩码建议
到 Snap 中使用的相机参数相同的 2D 图像上。在 Lookup
阶段,OpenIns3D 在 Mask2Pixel 图的帮助下搜索 CLT
精确地为 3D 掩模提案分配类别名称。将来自多个视图的结果组合起来以建立初始掩模分类结果。对于

其余的掩码,在本地范围内执行类似的查找过程,以
便于分类。最后,通过移除
全局和本地查找后都缺少类别分配的掩码。

OpenIns3D 在广泛的比较中表现出色
与其他方法相比,如图 2b 所示。这种简单灵活的方法在各种 3D 开放词汇任务上取得了一系列最佳 (SOTA) 结果。具体来说,在室内 S3DIS [1] 数据集和室外 STPLS3D 数据集的开放词汇实例分割上取得了最佳结果

[5] 数据集。OpenIns3D 甚至比同时期的 OpenMask3D [36] 表现更好
大量使用二维图像,例如,在具有挑战性的
副本 [34] 数据集。我们还发现“Snap and Lookup”是一个强大的开放词汇对象识别引擎,它实现了 SOTA 对象
在 ScanNet [9] 上的识别任务中,表现优于之前的 SOTA,10 亿
参数 3D 基础模型 [46]。最后,在转换 mask 建议时
转化为 3D 边界框,OpenIns3D 还在以下方面取得了最佳成果:
在 ScanNet 上进行开放词汇对象检测 (OVOD),优于之前的
图像相关方法 [42]。OpenIns3D 的设计还允许在无需重新训练的情况下更改 2D 检测器。这为模型提供了
能够与最新的二维开放词汇模型一起发展。此外,当二维检测器与大型局域网相结合时,



(a)3D特征蒸馏 (b)构建3D文本对 (c)CLIP 和投影 (d)Mask-Snap-Lookup（我们的）

图 3:四类开放词汇 3D 场景理解模型。a)3D 特征蒸馏框架,其中 2D 图像用作桥梁

将语言相关的特征提炼到 3D 世界中,典型作品包括 OpenScene [27]
和 Clip2Scene [6]。b) 构建 3D 文本对,其中使用 2D 字幕模型
构建 3D 文本对进行特征学习,典型方法包括 PLA 系列 [10, 11, 40] c) CLIP 和投影,从 2D 图像中裁剪出对象

然后经过 CLIP 处理,将结果直接投影到三维空间中,包括 OpenMask3D [36]、OV-3DET [23]、CLIP2 [41] 和 Open3DIS [26]。d) OpenIns3D

OpenIns3D 能够实现复杂的查询理解能力。当与 LISA [20]（一个基于 LLM 的推理分割模型)集成时，
OpenIns3D 表现出强大的能力来理解高度复杂的

语言查询并以 3D 形式执行推理分割,如图所示
图 1.总之,我们的贡献是：
- OpenIns3D 采用独特的管道,无需
  对齐良好的图像。这种方法在
  一系列基准,并具有理解高度复杂
  输入查询。
- 建议的“捕捉和查找”组合可以充当强大的 3D 对象识别引擎,尤其是对于从场景级扫描中提取的嘈
  杂 3D 对象。通过实现具有较大余量的最先进结果,证明了此功能。

2 相关工作

3D 开放词汇理解。与图像相比,3D 开放词汇理解的进展相对较慢。在

3D 物体分类任务,类似 PointCLIP [43]、PointCLIPV2 [50] 的方法，
CLIP2Point [16] 将 3D 点云投影到深度图中,并将它们链接起来
2D 模型进行分类。然而,这些方法在
场景级理解,其中点经常重叠且不完整。对于
场景级理解,大多数工作主要集中于利用对齐良好的 2D 姿势图像、深度图和点云 [6, 10, 11, 27, 31,
40, 41]。
一个值得注意的例子是 OpenScene [27],它能拍摄摆姿势的 2D 图像、深度
地图和 3D 数据作为输入,进行特征蒸馏,将 2D 语言对齐特征从图像转移到 3D 点云。同样，

Clip2Scene [6] 通过使用六个摄像头捕获的相应图像校准 LiDAR 点云来构建密集的像素点对。然而,由于这些方法仅关注语义级别的理解,因此很难实现实例级别的理解。相比之下,PLA [10] 及其后续工作 RegionPLC [40] 和 Lowis3D [11] 利用 2D 字幕模型构建 3D 文本对以学习特征。然而,PLA 系列工作依靠二分类头将输入对象分类为基本类别或新类别,并且该二分类头对不同基本新分割的可迁移性非常有限,这对灵活的应用构成了挑战。一项当前的工作 OpenMask3D [36] 利用对齐良好的 2D 图像来学习掩模提案的特征,在开放词汇实例分割中取得了令人印象深刻的结果。Open3DIS [26] 遵循相同的程序,使用图像增强了掩模提案和掩模分类,从而获得了更好的性能。这些方法的一个常见问题是它们依赖于输入中对齐良好的 3D 和 2D 对,而这在实际应用中可能并不总是可用的。我们在图 3 中总结了这些方法之间的主要区别。简化输入要求可提高方法的灵活性和兼容性。在这项工作中,我们探索如何在不依赖 2D 图像的情况下进行 3D 开放词汇理解。

从 3D 生成图像。过去,基于投影的方法已被广泛探索用于 3D 理解,并且已被证明有利于获得互补特征。例如,MVCNN [35] 将 3D 对象投影到不同的视图以帮助特征学习,而 LAR [3] 引入对象中心投影方法以从各个角度生成 3D 对象的图像,从而协助视觉接地任务。此外,虚拟视图融合 [19] 采用原始相机姿势但扩大了视野,从而增强了 2D 特征传输。然而,这些方法面临着诸如最佳视图选择、对象遮挡、投影期间信息丢失和渲染时间长等挑战。在开放词汇设置的背景下,投影图像的质量对模型性能起着至关重要的作用。在我们的工作中,我们评估了不同的投影方法及其与 2D 开放词汇模型的兼容性,以确定实现良好结果且实施效率高的最佳解决方案。

3 OpenIns3D

3.1 基线和挑战

基线。我们采用最近的 3D 实例分割主干 Mask3D [33] 来生成掩码提议,从而构建了一个朴素基线。为了使掩码提议模块 (MPM) 适合开放词汇设置,我们删除了 Mask3D 中使用分类标签的所有组件。后来,PointCLIP [50] 被用于掩码理解。这种朴素方法虽然仅满足 3D 输入的要求,但渲染时间长且性能不令人满意（见表 7）。这个基线模型存在几个问题。

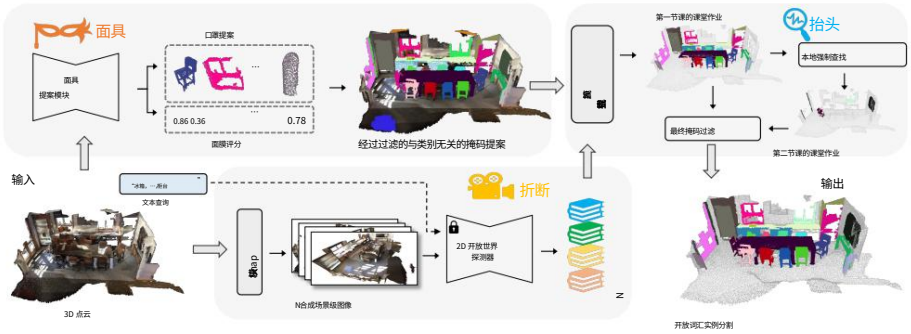


图 4:OpenIns3D 的通用流水线 OpenIns3D 首先使用 MPM 处理点云以生成 3D 掩码提案和掩码分数。然后，Snap 模块（详见图 5)渲染 N 个合成场景级图像,这些图像随后与输入的文本查询一起传递到 2D 开放世界模型中。来自 2D 模型的检测结果存储在类查找表 (CLT) 中。最后,将掩码提案和 CLT 都输入到查找模块中,其中在全局级别执行 Mask2Pixel 引导查找（详见图 6），然后在局部级别执行局部强制查找以解锁掩码提案的语义。最后的掩码过滤会细化掩码提案并得到最终结果。

挑战 1:过多的掩码提案。在 Mask3D [33] 中,掩码提案通过掩码分类逻辑进行过滤,在类别无关的设置中被删除。因此,需要一种有效的掩码过滤方案。

挑战 2:3D 实例质量低。从场景扫描中提取的 3D 实例通常是破碎的、不完整的和稀疏的。因此,通过简单投影生成的图像不易被 2D VL 模型理解。

挑战 3:缺乏背景信息。人类通过场景理解识别不完美的 3D 点云对象。然而,孤立的实例点云投影缺乏这种背景。一种解决方案是将蒙版和背景都投影到图像上,但这会引入分散注意力的元素,可能会混淆 CLIP 等分类级模型。

挑战 4:投影图像与自然图像之间的领域差距。渲染图像通常与训练中使用的自然图像有很大不同,这对 2D 视觉语言模型的理解提出了挑战。

3.2 总体框架

本节我们针对上述四个问题,介绍OpenIns3D的设计。OpenIns3D的流程如图4所示。

口罩:与类别无关的口罩提案

在 Mask Proposal Module (MPM) 中,我们引入了两种简单的设计,旨在过滤基线模型中生成的低质量 mask。

Mask 评分。受 [7, 17, 18] 的启发,我们将 Mask Module 生成的实例查询输入到浅层 MLP 模块中,以预测质量,即

mask 的 IoU。预测的 IoU (IoUm)由 ground truth 监督
训练阶段的IoU (IoUgt)值,计算如下:
二分匹配中的预测掩码及其匹配的真实掩码。
对于不匹配的预测掩码（IoUu），我们标记真实 IoU 值
为零。损失使用 L2 计算。为了避免 IoU 预测过低，
引入超参数γ来降低不匹配损失的权重
因此,MPM 的总损失函数为:

（1）

蒙版过滤。为了提高蒙版质量,应用了三个过滤器。首先,我们
保留模型预测 IoU 分数高于阈值 β 的掩码,确保只保留高质量的掩码。其次,从

SAM [18],我们通过比较两个二进制掩码来关注稳定掩码,这两个二进制掩码分别来自
使用不同的阈值对相同的底层软掩模进行处理。具体来说,我们
引入偏移值 α,并选择掩码,其中
阈值掩膜（一个为 −α,另一个为 +α)超过 80%。最后，
场景中的小物体经常会导致无效的提议,因此我们过滤掉了 mask
点数低于 N_{\text {min}} 的提案 通过采用这些技术,掩模提案的数量
将会减少,从而产生更清洁、
为后续的面具理解任务提供更高质量的面具（挑战1）。

Snap:合成场景级视图生成

从点渲染图像可能是一项耗时的任务,尤其是当
渲染任务数量很高。我们提出了一种多尺度合成场景级
图像方案。
相机姿势选择。Snap 模块以三个角度捕捉场景级图像
尺度:全局、角落和广角,如图 5 所示。对于全局级图像,相机位于场景上方,直接指向场景中心。对于角落图像,相机
位于中心上方，

指向角落,而对于广角图像,相机的位置
在 3×3 网格交互点处,指向最远的角落。使用相机位置坐标 P_{\text {cam}} 和向上
,目标坐标 P_{\text {target}} ，
场景 U 轴,可以使用 Lookat 函数来确定姿势
矩阵姿势。更详细的数学公式如下
补充材料。
相机固有校准。一旦建立了相机外部矩阵,就可以通过修改

相机的固有参数。目标是确保整个场景或
特定部分包含在所捕获的图像中。为了实现这一点
目标,我们初始化任意相机本征矩阵,然后调整焦距
通过缩放来调整图像中投影的区域,从而实现图像长度（fx和fy)和主点坐标（cx和cy)的缩放。缩放是
通过重新调整图像中的投影区域来实现的
坐标空间。例如:如果预定义区域的投影点

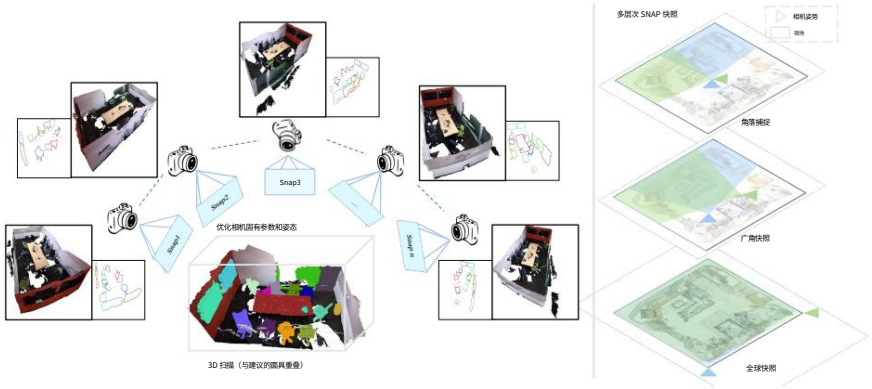


图 5:Snap 和 Mask2Pixel 地图。进行多尺度快照以渲染具有不同细节级别的图像,以便理解场景,包括广角快照、广角快照和全局快照。相机位于场景顶部并指向中心或角落,并使用校准的内在矩阵确定视野。使用定义的相机模型,构建 Mask2Pixel 地图以存储 2D 图像中每个 3D 掩模的位置（使用相同的颜色表示 2D-3D 对应关系)以指导类别名称的搜索。

位于 x 域 [-1000, -192] 范围内的图像坐标中,我们校准的固有参数将此范围转换为 x 中的 [0, 1000]。重要的是,我们保留了 x 和 y 坐标之间的纵横比,以保持最终图像的比例而不发生任何失真。此过程可确保充分利用每个捕获的图像并涵盖场景内的所有感兴趣区域（挑战 2 和 3）。

类别查找表。获得 N 张合成场景级图像后,我们将其输入 2D 开放词汇检测器。通过针对感兴趣的类别提供的文本查询,可以获得合成图像中检测到的对象列表。

随后,有关检测到的物体的信息（包括其位置和类别)将存储在指定的类别查找表 (CLT) 中。稍后将检索此表以将类别类别分配给 3D 掩码提案（挑战 4）。

Lookup:通过搜索进行口罩分类

我们进行多级搜索,为从 Mask 步骤生成的 mask 提案分配类别标签。

Mask2Pixel 引导查找。我们引入了 Mask2Pixel 引导查找 (MGL) 来在 CLT 中进行搜索。该概念涉及使用与生成 2D 图像相同的相机外部和内部矩阵将每个 3D 掩模提案投影到 2D 平面上,如图 5 所示。了解图像中每个掩模的精确像素位置后,我们可以通过 CLT 进行精确搜索,以确定每个掩模最可能的类别。MPM 的开发通过整合深度信息来考虑遮挡。为了完成匹配,我们遵循三步方法:1. 基于

在掩模到二维平面的投影上,我们选择最匹配的类
根据 IoU 值对类别进行分类;2. 如果最佳匹配对象的 IoU 值
在 2D 平面上低于 20%,则匹配被忽略。3. 我们汇总结果
从多个角度制定最终预测,计算概率
使用其标准化的平均 IoU 值对分数进行评估,如图 6 所示。
本地强制查找。虽然
Mask2Pixel 引导查找分配
用类别来掩盖提案,
有些面具可能不对应
CLT 中的对象。为了解决这个问题,
我们引入了本地强制查找
(LEL)方法。我们从 2D 场景级中裁剪出
剩余的蒙版
使用放大边界框的图像
并使用 2D 检测器对其进行处理以促进检
测。选择
为了获得最佳视图,我们引入了一种遮挡报告方法来
评估每个蒙版的遮挡情况

投影,然后选择前 K 个
LEL 视图。有关 Occlu-sion 报告的更多详细信息,请
参阅补充材料。

最终的蒙版细化。
以前的查找方法,大量的
获得掩码建议的比例
类别预测。所有掩码
在
MGL 和 LEL 阶段已被淘汰。

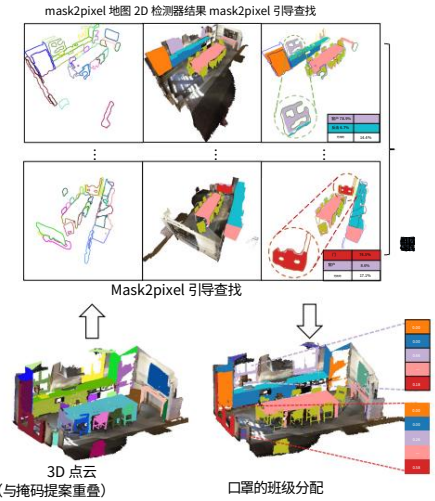


图 6:Mask2Pixel 引导查找图。2D 检测结果与投影掩码之间的 IoU 是

为 3D 蒙版分配类名的指导。
多幅图像结果被组合在一起。

4 实验

数据集和评估方案。我们在五个数据集上测试了 OpenIns3D,其中包括四个室内数据集,即 S3DIS [2]、ScanNetv2 [9]、ScanNet200 [31],
Repica [34] 和一个室外数据集 STPLS3D [5]。其中,S3DIS、Scan-Netv2 和 ScanNet200 是从 RGB-D 图像生成的室内点云数据集,Replica 是一个照片级逼真的 3D 室内场景重建数据集,而

STPLS3D 是一个航空摄影测量构建的户外数据集。我们专门使用这些数据集中的彩色 3D 数据,没有使用任何
2D 图像、姿势或深度图。按照先前研究 [10] 的设置,我们
排除了 ScanNetv2 中的“其他家具”类别和 S3DIS 中的“杂物”类别
由于含义模糊,Replica 和 ScanNet200 的评估方法如下:
OpenMask3D [36] 的设置。对于 STPLS3D,我们合并了低、中、
和高植被类别合并为一个“植被”类别,并保留其余所有类别。
实施细节。对于 S3DIS、ScanNetv2、Scannet200 和 STPLS
数据集,MPM 模块在训练时不使用任何类别标签,并且
λ 设置为 0.1,以减少零 IoU 的权重。Replica 的 mask 提案

表 1:ScanNetv2 上的零样本对象分类。OpenIns3D 的 Snap 和 Lookup 方法用于掩码分类,超越了所有以前的方法,包括最新的语言对齐的大规模 3D 基础模型 [46]。

方法	平均 床 车厢 椅子 沙发 桌子 门 风 背景图片 中心 书桌 冰箱 浴室 淋浴间 盥洗池
点CLIP [50]	6.3 0.0 0.0 0.0 0.0 0.0 0.7 0.0 0.0 91.8 0.0 0.0 0.0 0.0 15.0 0.0 0.0 0.0 0.0 0.0
PointCLIP V2 [43]	11.0 0.0 0.0 23.8 0.0 0.0 0.0 0.0 0.0 7.8 0.0 90.7 0.0 0.0 0.0 0.0 0.0 64.4 0.0 0.0
CLIP2Point [16]	24.9 20.8 0.0 85.1 43.3 26.5 69.9 0.0 20.9 1.7 31.7 27.0 0.0 1.6 46.5 0.0 22.4 25.6
带有 TP 的 PointCLIP。	26.1 0.0 55.7 72.8 5.0 5.1 1.7 0.0 77.2 0.0 0.0 0.0 51.7 0.3 0.0 0.0 40.3 85.3 49.2
带 TP 的 CLIP2Point。	35.2 11.8 3.0 45.1 27.6 10.5 61.5 2.6 71.9 0.3 33.6 29.9 4.7 11.5 72.2 92.4 86.1 34.0 38.5 32.6 67.2 69.3 42.3 18.3 19.1 4 .0 62.6 1.4 12.7 52.8 40.1 9.1 59.7 41.0 71.0 45.5
片段2 [41]	
Uni3D [46]	45.8 58.5 3.7 78.8 83.7 54.9 31.3 39.4 70.1 35.1 1.9 27.3 94.2 13.8 38.7 10.7 88.1 47.6
OpenIns3D	60.8 85.2 27.4 87.6 77.3 46.9 54.8 64.2 71.4 9.9 80.8 82.7 71.6 61.4 38.7 0.0 87.9 85.7

在 ScanNet200 上进行训练,随后在 OpenMask3D [36] 上进行训练。Snap 模块捕获大小为 1000 \times 1000 的图像,包括 16 个全局快照、4 个角快照和 4 个广角快照。对于 S3DIS,场景的顶部 0.5 米被移除,因为房间是封闭的。对于 STPLIS3D,我们遵循 Mask3D 将大型室外场景分割成 50m \times 50m 的块,并将相机提升至 10m。补充材料中介绍了更多实施细节。

开放词汇点云识别。我们首先在相同的设置下（即仅使用 3D 输入)评估所有现有 3D 开放词汇模型中 OpenIns3D 的性能。这些模型最常用于识别任务的测试,包括 PointCLIP(v1&v2) [50] [43]、Clip2Point [16]、CLIP2 [41]。我们还与大规模 3D 基础模型进行了比较,例如 Uni3D [46],它有 10 亿个参数,并使用大规模 3D 形状和图像-文本对进行训练。我们遵循他们的评估方案,并报告了 ScanNetv2 上实例分类的 Top-1 准确率。结果列于表 1 中。

开放词汇实例分割。我们采用了各种比较方案以与现有方法保持一致。对于 3D 开放词汇实例分割,我们与 PLA [10] 及其后续工作 RegionPLC [40] 和 Lowis3D [11] 进行了比较。为了进行公平的比较,我们遵循了它们的类别划分,并比较了我们在新类别上的结果,如表 2 所示。对于 STPLS3D,我们将 OpenIns3D 与分类模块为 PointCLIP 和 PointCLIPV2 [50] 的基线模型进行了比较（表 5）。我们还探索了 OpenIns3D 在更具挑战性的数据集上的性能,数据集包含更多的类别。具体来说,我们在 Replica（表 4)以及 ScanNet200（表 6)上比较了 OpenIns3D 与 OpenMask3D、OpenScene 的性能。在 OpenMask3D 之后,我们使用 Mask3D 为 OpenScene 生成掩码提案以供评估。

开放词汇对象检测。由于在进行这项工作时,关于 3D 开放词汇实例分割的研究有限,我们还选择了 3D 开放世界对象检测领域的一些最新方法进行更全面的评估。评估是通过将生成的掩码转换为轴对齐的边界框来进行的,结果如表 3 所示。

表 2:S3DIS 和
ScanNetv2。我们比较了在新类别上定义的零样本性能
在 PLA 系列工作中。在 S3DIS 数据集上取得了显著的改进，
在 ScanNetv2（B/N:Base/Novel)上观察到了具有竞争力的结果。

奥维斯	S3DIS			扫描网v2			
方法	B/N	AP50	AP25	B/N	AP50	AP25	需要 2D
解放军 [10]	8/4	08.6	-	10/7	21.9	-	✓
地区PLC [40]	8/4	-	-	10/7	32.3	-	✓
低维斯3D [11]	8/4	13.8	-	10/7	31.2	-	✓
Open3DIS [26]	8/4	26.3	-	10/7	-	-	✓
Mask3d+PointClip [50]	–/4	05.4	10.3	–/7	04.5	07.8	
开放Ins3D	–/4	37.0	39.3	–/7	27.9	42.6	
改进	–/4	(+10.7)	(+29.0)	–/7	-	(+34.8)	
解放军 [10]	6/6	09.8	-	8/9	25.1	-	✓
地区PLC [40]	6/6	-	-	8/9	32.2	-	✓
低维斯3D [11]	6/6	15.8	-	8/9	38.1	-	✓
Open3DIS [26]	6/6	29.0	-	8/9	-	-	✓
Mask3d+PointClip [50]	–/6	08.5	10.6	–/9	05.6	06.7	
开放Ins3D	–/6	33.0	38.9	–/9	19.5	27.9	
改进	–/6	(+4.0)	(+28.3)	–/9	-	(+21.2)	
Mask3d+PointClip [50]	–/12	08.6	09.3	–/17 –/	04.5	14.4	
OpenIns3D改进	–/12	28.3	29.5	17	28.7	38.9	
	–/12	(+19.7)	(+20.2)	–/17	(+24.2)	(+24.5)	

表 3： ScanNet 中针对未见类别的开放词汇对象检测(AP25) 。

方法	平均	厕所	床	椅子	沙发	梳妆台	桌子	柜子	书架	枕头	水槽
OV-PointCLIP [43]	3.1	6.6	2.3	6.3	3.9	0.7	7.2	1.0	1.4	0.2	2.8
OV-Image2Point [38]	0.6	0.4	0.0	0.2	0.8	0.8	0.2	1.7	4.2	4.6	1.2
Detic-ModelNet [47]	0.0	0.5	1.8	0.5	12.7	49.0	7.3	18.6	2.8	14. 3	2.4
Detic-ImageNet [47]	23.1	14.7	21.5	55.0	38.0	19.2	41.9	23.8	3.5	0.4	
OV-3DETIC [23]					2.6						
L3Det [49]											
FM-OV3D [42]											
开放Ins3D	43.7	79.5	70.5	76.9	15.8	0.0	53.1	40.1			
改进	(+19.1)	(+23.2)	(+31.7)	(+57.7)	–						

方法	0.7	2.1	0.6	0.8
OV-PointCLIP [43]				
OV-Image2Point [38]	1.0	0.9	0.0	0.1
Detic-ModelNet [47]	8.1	1.2	0.0	0.7
Detic-ImageNet [47]		0.3	0.0 0.7	
OV-3DETIC [23]		4.5	3.9 21.1	
L3Det [49]		17.3	23.4 27.9	
FM-OV3D [42]		6.0	17.4 8.8	
开放Ins3D		41.2	7.1 53.1	
改进	(+30)	(+25.4)	(+23.9)	(+25.2)

5 结果与讨论

5.1 与 SOTA 的比较

首先,OpenIns3D 在开放词汇方面表现出色
点云识别,超越了所有以前的方法,包括大规模三维基础模型 15%。这证明了零样本的有效性

“Snap” 和 “lookup” 方案。随着识别能力的增强,3D 开放词汇对象检测在 ScanNet 数据集中的表现
也取得了很大的进步。例如,对于 3D
分割,与 PLA 系列 [10, 11, 40] 和最新作品 Open3DIS [26] 相比,OpenIns3D 不需要对齐图像作为输入，
（仍然需要点的 RGB 信息）,在 S3DIS 上取得更高的结果
数据集,在 4 个新类别划分和 6 个新类别划分中均如此。在
SPTLS3D、OpenIns3D 的表现比基线模型 PointCLIPV2 高出 9.3%

表 4:室内复制品中的 OVIS
数据集。

模型	2D AP AP50 AP25
OpenScene [27]（2D融合） ✓10.9 15.6 17.3	
OpenScene [27] (2D/3D 编码) ✓ 8.2 10.4 13.3	
OpenMask3D [36] ✓ 13.1 18.4 24.2	
OpenScene [27]（3D蒸馏）	8.2 10.5 12.6
开放Ins3D	13.6 18.0 19.7
改进	(+5.4) (+7.5) (+7.1)

表 5:户外 OVIS
STPLS3D 数据集

模型	AP AP50 AP25
点CLIP [50]	02.0 02.6 04.0
PointCLIPv2 [43]	02.1 03.1 05.2
开放Ins3D	11.4 14.2 17.2
改进	(+9.3) (+11.9) (+12.0)

表 6:ScanNet200 验证集上的 3D 实例分割结果。
与无二维输入模型相比,OpenIns3D 表现出了强大的性能。
然而,在处理常见小物体时,出现了明显的局限性
和尾部课程。

模型	使用 2D APhead APcommon APtail AP AP50 AP25						
OpenScene（2D融合)[27]	✓	13.4	11.6	9.9	11.7	15.2	17.8
OpenScene (2D/3D 编码) [27] ✓		11.0	3.2	1.1	5.3	6.7	8.1
打开Mask3D	✓	17.1	14.1	14.9	15.4	19.9	23.1
OpenScene（3D蒸馏）		10.6	2.6	0.7	4.8	6.2	7.2
开放Ins3D		16.0	6.5	4.2	8.8 10.3 14.4		
改进		(+5.4)	(+3.9)	(+3.6) (+4.0) (+4.1) (+7.2)			

在 AP 中。在 Replica 数据集上,OpenIns3D 甚至优于 OpenMask3D，
它依赖于对齐良好的图像来理解掩码。在 Scan-Net200 的情况下,OpenIns3D 与所有其他 3D 相比获得了最高的性能
输入基线。然而,我们观察到尾部的表现有所下降
以及 ScanNet200 中的常见类别。性能下降可能是
归因于低质量和不清晰的较小物体的重建
ScanNet200 场景。这是 OpenIns3D 的一个明显限制,但对于更好的重建,例如在 Replica 中,以及对于 ScanNet200 中的 Head 类，
OpenIns3D 仍然表现出不错的性能。

综上所述,如果仅使用 3D 数据作为输入,OpenIns3D 在所有现有方法中表现出色,并且优于许多现有的

最先进的方法需要二维图像。它也显示出一定的局限性
在 3D 场景中无法很好重建的小物体上。

5.2 消融研究

掩模质量消融。在对 ScanNetv2 进行评估后,我们评估了
使用平均精度 (AP) 得分来评估与类别无关的掩码质量。我们
将所有类别视为通用类别,因为预测与类别无关。评估是在 ScanNetv2 验证集上进行的。表 8 展示了

掩模评分和掩模过滤设计的有效性。

表 7:渲染和推理时间消融。在典型
具有 50 个掩码的 ScanNet 场景。OpenIns3D 需要更少的渲染和推理时间。

渲染	2D 主干图像数量 图像尺寸Trender Tinfer 需要图片		(宽 × 高) (秒/场景)	(秒/场景)	(秒/场景)	总AP25 (%)	
点式	250	夹子	1282	5.2	15.3	20.5	9.3
拉尔	250	夹子	1282	14.3	18.7	33.0	10.5
蒙版渲染	250	夹子	1282	42.6	19.5	62.1	7.3
OpenIns3D（我们的）	8	G-恐龙 10002		2.3	6.2	8.5	29.8
OpenIns3D（我们的）	8	奥迪斯	10002	2.3	8.2	10.5	35.1

表 8:MPM 消融。MS:掩模评分。
MF:掩码过滤。

方法	AP50	AP25
Mask3d-监督 CA-Mask3d CA-	74.7	80.9
Mask3d + MS 50.2	47.5	49.2
(+02.7) 53.3 (+04.1)		
CA-Mask3d + MF 61.6 (+14.1) 71.0 (+21.8)		
CA-Mask3d + MS + MF 64.6 (+17.0) 73.4 (+24.2)		

表 9:Abla- 的浏览次数
LEL:本地强制查找

IDX4	8 16 LEL AP50 AP25			
我✓				18.3 27.1
二		✓		22.7 35.1
三			✓	24.8 37.5
四			✓✓	28.7 38.9

多视图消融。我们还研究了使用不同数字
视图数（表 9）。增加查找模块中使用的视图数量
可带来更好的结果。此外,强制查找功能提供了最终
改善结果。

投影和 2D 骨干消融。我们进行了全面的
研究各种渲染方法及其与 2D 主干的相互作用,以确定合适的方法。我们报告了渲染时间和推断

每种方法的性能（表 7）,更多详细信息请参见
补充材料。关键的观察是场景级渲染
和理解方法在速度上表现出色,同时也表现出强大的
性能。从 Grounding Dino [22] 切换到最新的 ODISE [39] 也
带来了性能的提升,表明 OpenIns3D 框架可以轻松受益于 2D 开放世界探测器的快速发展。

跨域分析。评估MPM的泛化能力
我们在不同的领域训练并测试了 OpenIns3D
数据集,如表 10 所示。与基线相比,跨域模型在两个数据集上也表现出令人印象深刻的性能。
值得注意的是,在 ScanNetv2 的 17 个类别中,有 11 个类别在 S3DIS 中不存在。

OpenIns3D 在 S3DIS 上训练,在这些模型中仍然取得了不错的表现
看不见的课程。

自由流动语言能力。捕捉和查找方案外包了
将理解任务隐藏到 2D 视觉语言模型中。因此,当与由 LISA [20] 等大型语言模型 (LLM) 驱动的 2D 模型集成时,
OpenIns3D 可以执行基于推理的分割任务（如图 7 所示）。例如,当给出查询“在脑电波期间用铅笔写下想法”时,

风暴”,OpenIns3D 准确地分割了白板,而对于“家具
提供与朋友一起休闲娱乐的体验”,它能够精准识别和细分
台球桌。

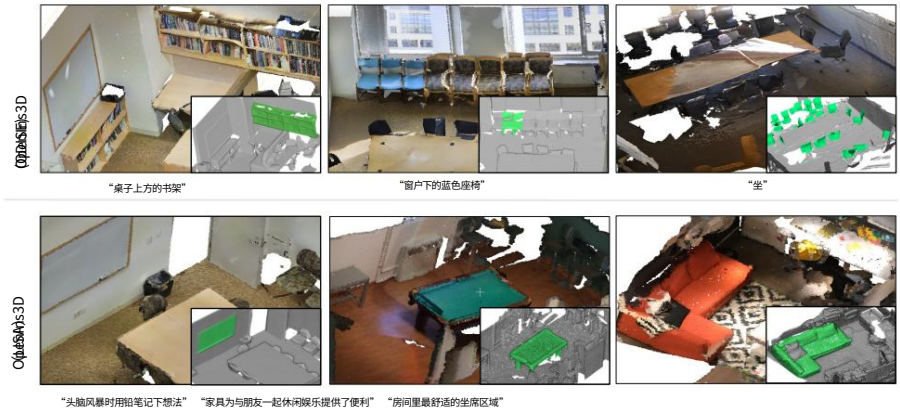


图 7:OpenIns3D 的定性结果。OpenIns3D (ODISE) 演示
管理多功能词汇的能力。OpenIns3D（LISA)可以进行3D推理分割。

表 10:跨域消融。我们在两个不同的数据集上训练并测试了 OpenIns3D,以检查其跨域能力。虽然 S3DIS 和 ScanNetV2 具有
非重叠类别,OpenIns3D 表现出良好的泛化能力。

测试	模型	训练数据	AP50	AP25
ScanNetv2 [9]	Mask3D-Pointclip [50]	扫描网v2	04.5	14.4
	开放Ins3D	扫描网v2	28.7	38.9
	开放Ins3D	S3DIS	21.5	33.6
S3DIS [1]	Mask3D-Pointclip [50]	S3DIS	03.5	06.8
	开放Ins3D	S3DIS	28.3	29.5
	开放Ins3D	扫描网v2	14.2	19.8

6 结论

实现 3D 开放词汇场景理解是一项具有挑战性的任务,主要是因为缺乏大量的 3D 文本数据。目前,这方面的大多数工作
该领域专注于使用 2D 图像来弥合 3D 和语言之间的差距。
然而,这不仅需要 2D 和 3D 之间的良好对齐,而且
由于在改变时需要付出巨大努力进行再训练,因此发展缓慢
2D 主干。OpenIns3D 引入了一个新的管道,即 Mask-Snap-Lookup,
这项任务的 Mask 模块在 3D 域中生成真实的 mask,
Snap 在 2D 域中渲染场景级图像,而 Lookup 模块
将结果从 2D 精确链接到 3D。此流程不需要图像输入,
实现了更好的性能,并且可以与 2D 模型无缝演进而无需训练。我们希望我们的工作能为研究人员提供新的视角
致力于开放世界 3D 场景理解。

致谢

本研究得到了英国剑桥大学格顿学院研究生研究奖、英国国家公路技术学院（由 EPSRC 博士培训中心赞助）、香港特别行政区创新科技署推出的 InnoHK 基金、中国国家自然科学基金（编号 62201484）、香港大学启动基金和香港大学基础研究种子基金的支持。我们非常感谢杨云汉在探索渲染技术方面的帮助，以及王成耀分享他对 Mask3D 的实现。

附录

简而言之，OpenIns3D 是一种用于 3D 开放世界场景理解的新管道，它仅使用 3D 彩色点云作为输入，使其更易于在各种场景中部署。我们还提供了 OpenIns3D 在 ScanNetv2 [9]、S3DIS [1] 和 STPLS3D [5] 数据集上实例分割和对象检测的表现的详细类别结果，以供将来的工作进行比较。在开发过程中，我们测试了各种渲染方法并记录了它们的不同性能，这证明了场景级渲染如何从其他渲染方法中脱颖而出。还介绍了 OpenIns3D 更详细的实现细节和方法。章节结构如下：

有关方法论的更多详细信息

与类别无关的掩码提议模块。我们修改了 Mask3D [33] 中需要分类标签的模块，使其成为与类别无关的设置。这包括 1. 删除匈牙利匹配中的语义概率成分，2. 消除语义分类损失，3. 丢弃基于分类 logits 的排名，以及 4. 摆脱基于分类 logits 的过滤。相反，我们添加了掩码评分和掩码过滤模块来获取高质量的掩码提议。

局部强制查找。在这里，我们详细解释了我们提出的遮挡报告模块，该模块可以有效评估所有合成图像中遮挡的情况。具体来说，执行以下四个步骤：

– 步骤 1. 点计数数组:我们通过构建 3D
维度为 W \times H \times (M+1) 的数组,其中 M, 表示
掩码,+1 表示背景点。此数组将表示为
PC,即点数,因为它是用来存储点数的
3D 蒙版投影到图像中的每个像素上。例如,如果像素
在坐标 i,j 处,3D 掩模 k 中的两个点占据了
投影,PT_{i,j,k} 将被分配值 2。
– 步骤 2. 最关键点识别:利用投影过程中生成的深度图,我们构建一个名为 FP 的二维数组
尺寸为 W \times H,用于识别每个 ,
像素并指示原始掩码编号。例如,如果像素 i,j 的
最前面的点是从 Mask k 投影的,我们表示 FP_{i,j} = k – 步骤 3. 遮挡率计
算:评估遮挡率 (OR)
对于特定图像中的掩码 k,我们计算以下公式:

其中 T 表示掩模 k 中的点的总数。
– 步骤 4. 所有图像报告:最后,我们对所有图像重复步骤 1-3,以
获得所有图像中每个遮挡率的总体报告,
形成最终的遮挡报告。

使用遮挡报告选择最佳视图后,合成场景级图像
裁剪以聚焦于特定的掩码提案,然后通过 2D 重新处理
检测器。结果也是在 Mask2Pixel 的帮助下搜索的,在这种情况下
二元掩模映射形成掩模的最终分类预测,
如图8所示。

B 实施细节

零样本物体识别

零样本物体识别结果采用“Snap and
Lookup”模块将类别名称分配给地面真相面具。Snap
模块拍摄 24 幅图像,Lookup 模块使用 ODISE [39] 提取
潜在的有趣对象,并为面具分配标签。对于
未分配标签或分配了错误的标签,top-1 分类
被标记为假阴性。
本次比较中的大多数其他方法都使用以对象为中心的渲染方法
将深度图或点云投影到图像中以进行分类。
Uni3D [46] 也经过了大量图像和文本对的预训练,
作为 3D 形状。OpenIns3D 的场景级图像渲染,具有增强的效果,
被证明在物体识别任务中更为有效。

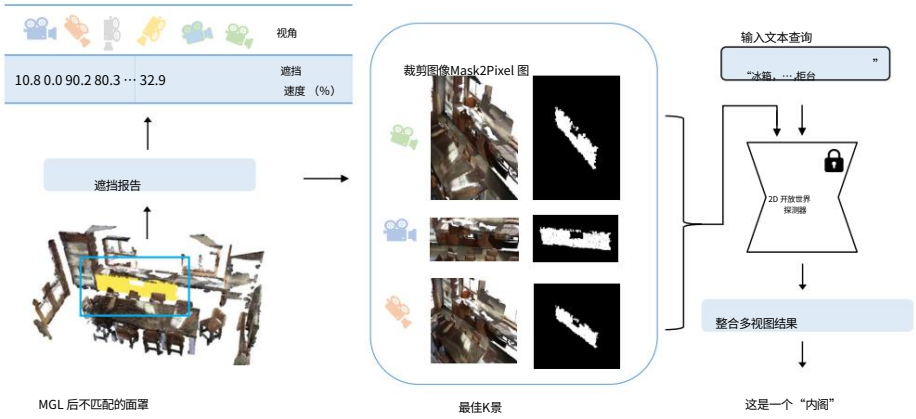


图 8:局部强制查找说明。第一阶段的剩余蒙版首先通过遮挡报告模块来选择最佳 K 个视图。所选图像在由 2D 检测器处理之前会被裁剪，以产生分类结果。

开放词汇实例分割

掩码。掩码提议模块基于轻量级版本的 Mask3D [33] 构建，具有三个解码器层。对于掩码质量评分模块，我们将 λ 设置为 0.1，以降低零 IOU 掩码的权重。掩码提议模块使用 ADAM 优化器进行训练，学习率为 0.0003，并应用单周期调度程序。对于 ScanNetv2、S3DIS 和 STPLS3D 实验，掩码提议模块在所有类别的类无关掩码上进行训练，以学习提议掩码。对于 Replica 和 ScanNet200，我们遵循 OpenMask3D [36] 并使用 ScanNet200 预训练权重进行掩码提议。

快照。我们捕获了 24 张场景图像，包括 16 张全局图像、4 张角落图像和 4 张广角图像。我们使用 PyTorch3D 光栅化渲染器来渲染图像。对于所有数据集，我们捕获了尺寸为 1000 x 1000 的图像，以在速度和性能之间取得很好的平衡。此外，为了避免天花板造成的遮挡效应，我们丢弃了 S3DIS 和 ScanNet 数据集中顶部 0.3m 的点。因此，S3DIS 数据集中的天花板类别在 AP 结果中被完全丢弃并分配为 0。对于 STPLS3D，摄像机位置位于场景顶部上方 5m 处以获得更好的视野。

查找。在查找阶段，我们仅当结果已在至少两个视图中得到验证时才为每个掩码分配分类标签。在局部强制查找的情况下，我们使用两倍于目标掩码大小的边界框裁剪图像。裁剪后的图像随后被输入到 2D 检测器中以优化结果。

Mask2Pixel 图（在本例中为二值图）用于精确搜索检测结果，如图 8 所示。

表 11:ScanNet200 验证的 3D 实例分割结果
设置。带有 RGBD 图像的 OpenIns3D 表现出了颇具竞争力的性能。

模型	使用 2D	APhead	APcommon	APtail	AP	AP50	AP25
OpenScene（2D融合)[27]	✓	13.4	11.6	9.9	11.7	15.2	17.8
OpenScene (2D/3D 合成) [27]	✓	11.0	3.2	1.1	5.3	6.7	8.1
OpenMask3D	✓	17.1	14.1	14.9	15.4	19.9	23.1
OpenIns3D与 rgbd	✓	19.2	14.2	14.2	15.9	20.6	23.3
OpenScene（3D蒸馏）		10.6	2.6	0.7	4.8	6.2	7.2
开放Ins3D		16.0	6.5	4.2	8.8	10.3	14.4

表 12:副本上的 3D 实例分割结果。OpenIns3D 具有
RGBD 图像表现出颇具竞争力的性能。

模型	2D	AP	AP50	AP25
OpenScene [27]（2D 融合）	✓	10.9	15.6	17.3
OpenScene [27] (2D/3D 加密)	✓	8.2	10.4	13.3
OpenMask3D [36]	✓	13.1	18.4	24.2
Open3DIS [26]	✓	18.5	24.5	28.2
OpenIns3D与 rgbd	✓	21.1	26.2	30.6
OpenScene [27]（3D蒸馏）		8.2	10.5	12.6
开放Ins3D		13.6	18.0	19.7

开放词汇对象检测

开放词汇对象检测是通过将掩码提议转换为轴对齐的边界框来进行的。我们遵循相同的 Snap 和 Lookup
开放词汇实例分割设置中的实现细节
边界框理解。

C 用 RGB-D 替换 Snap

虽然 OpenIns3D 主要利用 3D 输入框架和所提出的
“Mask-Snap-Lookup”管道,我们还在以下场景中评估其性能：
提供 2D 图像,以便与其他方法进行比较。
对框架所做的修改是将 Snap 模块替换为
可用的 RGB-D 图像。为了进行评估,我们选择了两个广泛的数据集：
ScanNet200 [31] 和 Replica [34]。结果列于表 11 和
12. OpenIns3D 搭配 RGB-D 在 Replica 数据集上表现出色,同时
在 ScanNet200 数据集上具有竞争力。两种方法都使用 Yoloworld [8] 作为
2D 探测器。（2024 年 8 月 1 日更新）

D 详细结果

开放词汇实例分割

表 13 和 14 提供了 OpenIns3D 在 S3DIS 和 ScanNetv2 数据集上的每个类别的结果。PLA 中的新（未见过）类别以蓝色突出显示。
表 15 表示 STPLS3D 数据集的每个类别的结果，与 PointCLIP 和 PointCLIPV2 进行了比较。

在 S3DIS 中，OpenIns3D 在这些新颖的类别中始终取得较高的成绩。
我们将此归因于 S3DIS 中 3D 点数据的高质量，这为 2D Snap 图像中的物体检测提供了有利条件。然而，对于柱子和梁等类别，OpenIns3D 很难产生理想的结果。上限结果标记为 0，如实现细节部分所述。

在ScanNetv2中，点云数据质量不是很高。因此，
Snap 输出质量有限，导致性能略低。
在 STPLS3D 中，OpenIns3D 的表现远胜于 PointCLIP 和 PointCLIPV2。这是意料之中的，正如零样本分类任务中已经证明的那样。然而，在自行车、摩托车、标志和灯杆等非常小的物体上，其性能并不那么强劲。这是因为 Snap 模块将相机放置在高角度，导致这些较小物体可用的像素数量有限。

开放词汇对象检测。

我们还在表 16 中展示了所有类别的物体检测结果，其次是 OV-3DETIC 类别。OpenIns3D 在大多数类别中都表现出色。

跨域分析。

表 17 展示了跨域 OpenIns3D 模型的各类别结果，该模型在 S3DIS 上训练并在 ScanNetv2 上测试。进行此表比较是为了展示 mask 模块的泛化能力。

E 图像生成的其他尝试

图 9 和表 18 展示了我们在得出合成场景级图像提供最佳解决方案的结论之前探索过的替代 2D 图像渲染方法。我们在此记录了该过程以供将来参考。

尝试 I、II：受到 LAR [3] 成功的启发，我们将相机放置在物体周围，并投影点云以为每个蒙版生成多视图图像。然而，这些方法生成的图像超出了 CLIP 模型的识别能力，尤其是对于分割或重建效果不佳的蒙版，导致结果不令人满意。

图 9:从 3D 生成 2D 图像的尝试可视化。I:LAR 点投影;II:LAR 点背景投影;III:网格渲染;IV 场景中的网格渲染;V:网格背景渲染;VI:从原始 2D 图像裁剪;VII:从网格进行场景级渲染;VIII:从点进行场景级渲染。性能可在表 18 中找到。

表 13:S3DIS AP50 上的 3D 开放词汇实例分割的每类结果。新类别上的表现以蓝色标记。

方法划分														
解放军 [10]	B8/N4 89.5 100.0 50.8 00.0 35.3 36.2 60.5 00.1 84.6 01.9 00.8 59.4 B6/N6 89.5 60.2 17.9 00.0 41.5 10.2 02.1 00.6 86.2 45.1 00.1 02.2 OpenIns3D –/N12 00.0 84.4 29.0 00.0 00.0 62.6 25.2 25.5 52.0 60.0 00.0 00.0													

表 14:ScanNet AP50 上的 3D 开放词汇实例分割的每类结果。新类别上的表现以蓝色标记。

方法划分																	
解放军 [10]	B13/N4 50.5 77.0 82.9 43.4 75.4 49.0 46.0 43.7 46.5 33.7 23.2 54.1 49.6 56.0 97.8 47.5 85.8 B10/N7 53.7 62.7 11.2 70.5 27.2 47.7 45.7 30.0 01.5 39.9 40.8 50.6 68.6 84.6 92.9 24.6 00.0 B8/N9 45.1 77.4 82.2 84.2 74.2 48.9 51.0 30.0 00.5 02.1 16.8 44.9 28.3 35.1 94.3 16.6 00.0 OpenIns3D –/N17 24.3 52.5 75.7 61.6 40.6 39.7 45.5 54.8 0.5 33.5 16.7 48.1 18.5 4.3 50.1 16.8 7.6																

尝试 III、IV、V:我们将注意力转移到场景网格上,使用光栅化渲染方法而不是简单的点投影。虽然这些方法带来了一些改进,但当提出的遮罩不完美时,它们仍然被证明具有很大的局限性。质量不理想的遮罩占了遮罩的很大一部分,这使得这是一个不充分的解决方案。此外,每个遮罩的渲染都需要大量时间,这使得部署不切实际。

尝试 VI:然后,我们尝试使用原始图像并裁剪图像中的蒙版进行评估(VI)。我们认为这将提供最佳质量的图像,使它们最有可能通过 2D 模型进行识别。

我们使用遮挡报告方法从所有帧中选择前 K 个视图,并使用放大的边界框裁剪掉遮罩像素。这种方法取得了显著的效果,这主要归功于 2D 图像的高质量。然而,我们最终放弃了这种方法,因为担心它在一般场景中的适用性。

第七、第八次尝试:将重点转移到场景级渲染,我们的模型开始产生高质量的结果。通过从远处观察所有损坏的实例并结合大量上下文信息,物体变得清晰可辨。因此,场景级图像与用于训练 2D 视觉语言模型的图像的领域差距很小。

F 局限性和未来工作

OpenIns3D 存在一些局限性,需要在未来的研究中进一步探讨。

表 15:3D 开放词汇实例分割的每类结果
在 STPLS3D AP50 上。所有模型均以零样本方式进行测试。

方法	意思是	建筑	地	植被	车	人	杆	垃圾桶	灯	篱笆	路	建筑
点CLIP [50]	2.7	15.3	0.4	10.2	06.6	00.0	00.0	00.0	00.0	00.0	00.0	00.0
PointCLIPV2 [43] 3.2		20.3	0.2	12.3	5.8	00.0	00.0	00.0	00.0	00.0	00.0	00.0
开放Ins3D	14.1	40.4	01.2	54.2	24.2	30.0	05.5	02.1	03.0	00.0	00.0	08.3

表 16:3D 开放词汇对象检测的详细结果。我们展示了所有
ScanNet20 上的结果,其次是 OV-3DETIC [23]。

方法	意思是	墙	床	椅	沙发	桌子	柜台	架子	水槽	厕所	浴缸	冰箱	书桌	柜子	门	窗户	垃圾桶	灯	包
3D 检测 [48]	15.0	53.3	24.9	15.8	31.4	11.5	09.1	02.1	09.4	17.0	29.2	27.5	20.0	13.7	00.0	00.0	00.0	17.7	04.8	03.0	09.5
CLIP-3D [29]	12.7	44.8	23.8	17.5	12.6	04.9	13.2	01.9	04.0	11.4	17.6	32.2	14.9	11.4	02.4	00.5	14.5	08.6	07.5	05.1	04.7
OV-3DETIC [23]	18.8	57.3	42.3	27.1	31.5	08.2	14.2	03.0	05.6	23.0	31.6	56.3	11.0	19.7	00.8	00.3	09.6	10.5	03.8	02.1	02.7
开放Ins3D	37.1	79.5	70.5	76.9	15.8	00.0	53.1	40.1	41.2	07.1	53.1	14.3	32.1	29.1	04.8	55.6	40.4	41.1	02.6	48.0	06.2

表 17:ScanNetv2 上 OVOD 的 OpenIns3D 跨域分析
AP25. OpenIns3D 在跨域数据集上取得了有竞争力的结果,甚至在
类别 在训练数据集中不可用,以蓝色突出显示。相比
与 OVOD 上的其他 SOTA 模型相比,跨域 OpenIns3D 仍然具有竞争力的性能。MPM-SC:在 ScanNetv2 上训练的 MPM;MPM-S3:在 S3DIS 上训练的 MPM。

方法	盒子	墙	床	椅	沙发	桌子	门	窗户	架子	图片	柜台	书桌	浴帘	浴缸	马桶	水槽	冰箱
OpenIns3D MPM-SC	17.1	57.5	74.5	59.2	36.9	29.3	47.5	26.4	0.0	31.1	32.2	55.4	39.1	0.0	57.4	42.1	6.6
OpenIns3D MPM-S3	16.1	43.5	45.7	41.8	28.6	17.7	18.3	31.9	1.2	1.0	29.3	23.1	20.1	8.0	63.6	16.4	1.7
下型号																	
PointCLIP P-3DE	6.0	4.8	45.2	4.8	2.2	PointCLIPV2 P-3DE	19.3	21.0	61.9	15.6	7.4	8	13.2	4.6	-	-	1.0	4.0	-	-	-	-	13.4	6.5
OV-3DET P-2DE	3.0	42.3	27.1	31.5	14.2	9.6									-	-	12.4	21.4	-	-	-	-	14.5	16.8
															-	5.6	-	0.3	19.7	10.5	11.0	-	57.3	31.6	56.3

依赖地面实况实例掩码:类似于 SAM [18]、OpenIns3D
仍然依赖于地面真相面具监督。虽然它确实被证明具有
能够概括以前从未见过的面具,提供
大量与类无关的掩码可能会有所帮助。像 Un-Scene3D [32]、Segment3D [15] 这样的方法可以
作为掩码的替代方法
提案,将其与 Snap 和开放词汇查找模块链接起来
理解。这需要进一步调查。

– 语义分割性能有限:OpenIns3D 严重依赖于
过滤来改进掩码提议,直接丢弃质量低的掩码。虽然这种方法有利于实例分割,因为它减少了错误

正面实例,它可能会限制其在语义分割中的表现。

表 18:Snap 和 Lookup 模块的演进。对应的图像
可视化效果如图 9 所示。场景级渲染不仅需要更少的图像
而且与其他预掩码渲染级别相比也取得了出色的效果。

*:VI 的图像大小不固定,因为它取决于原件上的遮罩区域的大小
图像。

索引	方法	工作强度	所需图像	使用 2D 图像	大小	2D 主干	AP50	AP25
I	LAR-每个掩模的点投影		250		1282	夹子 夹	5.3	8.6
II	每个掩模的 LAR-point-bg-projection		250		1282	子 夹子	6.3	10.5
III	每个蒙版的网格渲染		250		1282	6.7 7.3	6.8	7.2
IV	每个掩码的 IV 网格场景渲染				1282			
V	每个遮罩的 V mesh-bg-rendering				1282	剪辑 4.3 5.3		
VI	crop-original2d 每个遮罩			✓		剪辑 24.3 29.6		
VII	每个场景的场景网格渲染				10002	奥德赛 28.7 38.9		
VIII	场景点渲染	每个场景	250250250	8	8	10002 奥德赛 21.5 33.6		

表 19:与 OpenScene 和其他框架的语义比较
分割。我们的框架优先考虑掩码质量,并牺牲整体语义
分割结果。

语义段。	米卢						麦考林					
方法	书架	书桌	沙发	厕所	平均值		书架	书桌	沙发	厕所	平均值	
3DGenZ [24]	6.3	3.3	13.1	8.1	7.7		13.4	5.9	5.9	26.3	12.9	
MSeg 投票 [21]	47.8	40.3	56.5	68.8	53.3		50.1	67.7	67.7	81.0	66.6	
OpenScene-LSeg [27]	67.1	46.4	60.2	77.5	62.8		85.5	69.5	69.5	90.0	78.6	
OpenScene-OpenSeg [27]	64.1	27.4	49.6	63.7	51.2		73.7	73.4	73.4	95.3	79.0	
开放Ins3D	54.8	16.7	61.6	50.6	45.9		59.0	32.3	76.7	79.8	61.9	

我们还计算了OpenIns3D在
OpenScene [27] 报告了四个类别,如表 19 所示。我们的
与 OpenScene 相比,该方法在语义方面仍然存在差距
分割。

– 小物体性能:OpenIns3D 的性能最终取决于
与点云本身的质量密切相关。非常
较小或由稀疏点云组成,在
渲染图像,因为它们要么占据图像像素的一小部分
或者过于零散而无法被二维模型检测到。

G 可视化

其他数据集上的零样本性能。OpenIns3D 能够部署
在任何彩色的 3D 扫描上,无论是否有 2D 图像对应物。
为了证明这一点,我们提供了一些在基于激光雷达的数据集（如 ArkitScene Lidar)上部署
OpenIns3D 的演示,这些数据集的 2D 图像不可用,并且
Mattport3D,具有不同风格的室内房间。结果显示
在图 11 和图 10 中,我们展示了两个掩码提案（在
ScanNet 200) 以及最终的检测结果。

口罩建议。图 12 和 13 给出了口罩的定性评估
提议模块。学习到的掩码提议与
地面实况掩码，通常会捕获额外的未标记掩码。这证明了我们的无类标签学习方案在生成高质量类无关掩码提案
方面的有效性。此外，通过应用 Mask

评分和掩码过滤技术，我们能够连接碎片化或
脆弱的掩模，从而显著提高掩模质量。这些进步为理解 Snap 和 Lookup 奠定了坚实的基础

方案。

Snap 可视化。图 14、15 和 16 展示了
Snap 模块，我们主要展示全局 snap。使用建议的姿势和
Snap 模块能够从点云生成质量不错的图像，无论数据集是室内还是

户外的。

查找结果可视化。查找模块有效地将二维结果链接到
3D。在这里，我们展示了从所有三个数据集中得出的结果的可视化
（图 17、18、19）。