Open3DIS:基于 2D 掩码引导的开放词汇 3D 实例分割

Phuc Nguyen1* Tuan Duc Ngo1,4* Evangelos Kalogerakis4
Chong Gan2,4 Anh Tran1 Cuong Pham1,3 Khoi Nguyen1

1VinAI 研究 2MIT-IBM Watson AI Lab 3上海邮电技术学院 4马萨诸塞大学阿默斯特分校

{v.phucnda,v.anhtt152,v.khoindm}@vinai.io {tdngo,kalo}@cs.umass.edu

ganchuang@csail.mit.edu cuongpv@ptit.edu.vn

https://open3dis.github.io/

图 1.左图:虽然领先的开放词汇 3D 实例分割方法（如 OpenMask3D [64]和 OVIR-3D [47])通常处理小的或模糊的实例,特别是那些来自不常见类别的实例,Open3DIS 在分割此类情况方面表现出色。它在 ScanNet200 [58]上,其平均精度比现有方法高出约 1.5 倍。右图： Open3DIS 聚合了来自基于点云的实例分割器和基于 2D 图像的网络。我们的方法结合了新颖的组件（红色和黄色框)执行跨多帧的 2D 蒙版到点云的聚合和映射,以及 3D 感知特征提取有效地将对象提议与文本查询进行比较。

抽象的

我们推出了 Open3DIS,这是一种新颖的解决方案,旨在解决 3D 场景中的开放词汇实例分割问题。3D 环境中的对象呈现出不同的形状、尺度和颜色,从而精确实例级识别是一项具有挑战性的任务。开放词汇场景理解方面的最新进展通过采用与类别无关的 3D 实例提议网络进行对象定位并学习每个 3D 掩模的可查询特征,在该领域取得了重大进展。

虽然这些方法可以生成高质量的实例提议,但它们很难识别小规模和几何模糊的对象。我们方法的关键思想是一个新模块,用于聚合跨框架并将它们映射到几何连贯的点云区域作为高质量对象提案解决上述限制。然后将它们与 3D 类别无关的实例提案涵盖范围广泛现实世界中的物体。为了验证我们的方法,我们在三个著名数据集上进行了实验,包括 ScanNet200、S3DIS 和 Replica,结果显示,在分割具有不同特征的对象方面具有显著的性能提升类别优于最先进的方法。

1. 简介

本文讨论了开放词汇 3D 点云实例分割 (OV-3DIS) 这一具有挑战性的问题。给定一个由点云表示的 3D 场景,我们寻求获得任意类的一组二进制实例掩码感兴趣的,这可能在训练阶段不存在。这个问题的出现是为了克服

: 平等贡献

传统的全监督三维实例分割
（3DIS)方法[21, 22, 50, 60, 63, 66, 81, 84]，
受封闭集框架的约束 将识别限制在预先确定的一组对象类中

通过训练数据集。这项任务在机器人和 VR 系统中有着广泛的应用。此功能可以
使机器人或代理能够识别和定位物体
使用文本描述的 3D 环境中的任何类型的
详细说明名称、外观、功能等。

目前，针对 OV-3DIS 的研究很少
[10, 11, 47, 64]。最近， [64]建议使用
预先训练的3DIS模型实例提议网络来捕捉3D点云场景的几何结构和

生成高质量的实例掩码。然而，这种方法在识别稀有物体方面面临挑战，因为

它们在 3D 点云场景中不完整出现，并且
预训练的 3D 模型对于此类不常见类别的检测能力有限。另一种方法是

利用现成的二维开放词汇理解
模型[47, 78]很容易捕捉到新的类别。尽管如此，将这些 2D 提议从图像转换为 3D 点
云场景是一项具有挑战性的任务。这是因为
事实上，2D 提案只捕捉了
3D 对象，还可能包括不相关的区域，例如
背景。这两种方法总结如下
图1.

在本文中，我们介绍了一种 OV-3DIS 方法 Open3DIS，它扩展了预定义概念集之外的理解能力。给定一个 RGB-D 图像序列和相应的 3D 重建点云

场景中，Open3DIS 解决了现有方法的局限性。它通过使用 3D 实例网络和 2D-guide-3D 实例提议模块来补充 3D 实例提议的两个来源，以实现足够的 3D 对象二进制实例掩码。该模块（我们的主要贡献)从点中提取几何相干区域

在多帧二维预测掩模的指导下，将图像传输到云中，并将它们聚合成更高质量的三维图像
提案。后来，逐点特征提取聚合
以多尺度方式对每个实例进行 CLIP 特征
跨多个视图，构建实例感知点
用于开放词汇实例分割的云特征。

为了评估 Open3DIS 的开放词汇能力，
我们在 ScanNet200 [58]、 S3DIS
[1] 和 Replica [62]数据集。Open3DIS 在 OV-3DIS 中取得了最佳结果，远远超过了之前的研究成果。尤其是与

在大规模数据集ScanNet200上领先的方法。

总而言之，我们工作的贡献如下：

1. 我们推出了“2D 引导的 3D 提案模块”
   通过使用聚合的二维实例对内聚点云区域进行聚类来创建精确的三维提案
   来自多视角 RGB-D 图像的蒙版。
2. 我们引入了一种新颖的逐点特征提取
   开放词汇 3D 对象提议的方法。
3. Open3DIS 在 Scan-Net200、S3DIS 和 Replica 数据集上取得了最佳效果，其性能与全监督方法相当。

2.相关工作

开放词汇二维场景理解方法旨在
在测试中识别基类和新类
基础类在训练过程中可见，而新类
类不是。根据识别任务的类型，我们
可以将它们归类为开放词汇对象检测
（OVOD） [32、46、52、67、79、83、87] 、开放词汇语义分割
（OVSS） [9、40、42、70、72、90] ，以及
开放词汇实例分割（OVIS） [20,29,65 ，
69, 85, 86]. 处理小说
类是利用预先训练的视觉文本嵌入
模型，例如 CLIP [54]或 ALIGN [30]，作为基础类别和新类别共存的联合文本-图像嵌入，
为了将模型的功能转移到基类上
到新的类别。然而，这些方法不能简单地
扩展到3D点云，因为3D点云无序且密度不平衡，并且外观和形状的变化比2D图像大得多。

全监督 3D 实例分割 (F-3DIS)
旨在将 3D 点云分割成训练类实例。F-3DIS 的方法可分为

主要分为三类：基于框的[25, 76, 81]、基于聚类的
[5, 12, 31, 66, 68]，以及基于动态卷积的[21, 22,
45, 50, 60, 63, 71]技术。基于框的方法检测
并分割每个 3D 提案内的前景区域
框来获取实例掩码。基于集群的方法采用
将预测的对象质心分组到聚类中或
构建树或图结构，然后将其分解为子树或子图[28, 43]。对于第三个

Mask3D [60]和 ISBNet [50]等提出了一种动态卷积方法，其内核代表不同的对象实例，并与逐点特征进行卷积

导出实例掩码。在本文中，我们使用 ISBNet 作为
3D 网络，但需要进行必要的调整才能输出 3D
与类别无关的提案。

开放词汇 3D 语义分割 (OV-3DSS)
和对象检测（OV-3DOD)以开放词汇的方式实现对 3D 场景的语义理解，

包括可供性、材料、活动和属性
在看不见的环境中。此功能突出显示
在最近的工作[17, 24, 51]中，OV-3DSS 和[4, 48, 89]中
OV-3DOD。尽管如此，这些方法不能精确地
使用 3D 实例蒙版定位和区分 3D 对象，
因此无法完整描述 3D 物体的形状。

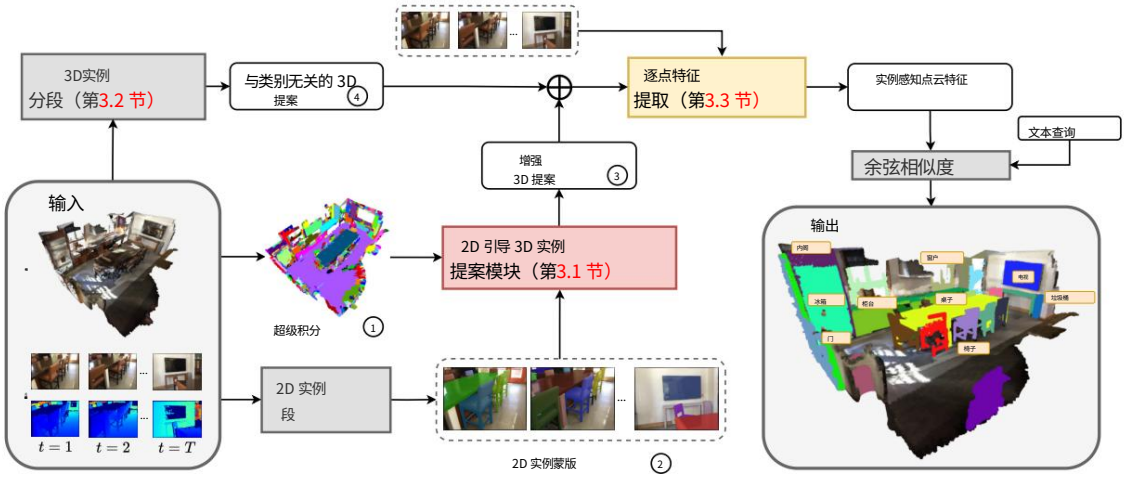

图 2. Open3DIS 概览。预先训练的类无关 3D 实例分割器会提出初始 3D 对象，而 2D 实例分割器为视频帧生成掩码。我们的 2D-Guided-3D 实例提议模块（第3.1 节）结合了超点和 2D 实例掩码来增强 3D 提案，并将其与初始 3D 提案集成。最后，逐点特征提取模块（第 3.3 节）将实例感知点云 CLIP 特征与文本嵌入关联起来以生成最终实例掩码。

开放词汇 3D 实例分割（OV-3DIS）涉及对可见和不可见类别的划分（在将 3D 点云训练成实例。方法 OV-3DIS 可分为 3 组：基于开放词汇语义分割、基于文本描述和 3D 提案对比学习、以及 2D 开放词汇

供电方法。第一组包括 OpenScene [51]和 Clip3D [23]利用以下聚类技术 DBScan 对 OV-3DSS 结果进行扫描，生成 3D 实例提议。然而，它们的质量取决于聚类准确性并且可能导致不可靠的结果。在

另一方面，第二组由 PLA [11]、 Re-gionPLC [77]和 Lowis3D [10]组成，专注于训练 3D 实例提议网络以及预测提议与其对应文本标题之间的对比开放词汇。然而，当类别数量增加时，这些方法难以处理，并且可能会降低其区分不同对象类别的能力。对于

最后一组OpenMask3D [64]使用预先训练的 3DIS 模型生成与类别无关的 3D 提案，随后根据其 CLIP 得分进行分类 2D 掩模投影。类似地，OpenIns3D [27]采用预先训练的 3DIS 模型，并通过以下方式解决问题其 Mask-Snap-Lookup 模块利用合成场景跨多个尺度的图像。然而，挑战也随之而来用于识别小型或具有独特几何结构的不常见物体类别。相反，OVIR-3D [47]、 SAM3D [78]、 SAM-Pro3D [74]、 MaskClustering [75]和 SAI3D [82]利用

预训练的二维开放词汇模型生成二维实例掩码，然后将其反向投影到相关的三维点云上。然而，

带有对象的 2D 分割蒙版会导致前景对象中包含背景点，从而导致 3D 提案的质量欠佳。尽管如此，该团队相对于其他团队的优势在于他们利用 2D 预训练模型处理大规模数据集，例如

CLIP [54]或 SAM [35]可以扩展到数百类如 ScanNet200 [58]。在最后一组之后，Open3DIS 通过以下方式生成高质量的 3D 实例建议：将 3DIS 网络的 3D 蒙版与提案相结合通过在 2D 实例掩码的指导下对几何相干区域（超点）进行分组来生成。这

补充了来自 3D 网络。我们的方法擅长捕捉稀有物体在保留三维几何结构的同时，实现 OV-3DIS 领域最先进的性能。

3. 方法

我们的方法处理 3D 点云和 RGB-D 序列，产生一组三维二进制掩码，表示场景中的对象实例。我们假设已知相机每一帧的参数。我们的架构如下图所示

图2. 与先前的研究[11, 64, 77]类似，我们采用了 3DIS 网络模块直接提取对象建议来自 3D 点云。该模块利用 3D 卷积和注意力机制，捕获空间和

用于稳健的 3D 对象实例检测的结构关系。尽管它有诸多优势，但稀疏点云、采样伪影和噪声可能会导致遗漏物体，尤其是对于小物体，例如图1 中的纸巾盒。

我们的方法集成了一个新颖的 2D-Guided-3D 实例提议模块，利用在大型图像数据集上训练的二维实例分割网络来更好地捕捉

单个图像中较小的物体。然而，由此产生的 2D 蒙版可能只能捕捉实际 3D 对象实例的部分由于遮挡（图2-2 ）。为了解决这个问题，我们提出

召回 Recallhead Recallcom Recalltail			
仅限 3D 61.63 仅限 2D 68.61	81.92 76.66	53.68 74.73	12.06 34.68
2D 和 3D 73.29	87.48	74.16	34.31

表 1.2D、3D 或组合提案的召回率（%）。

一种通过以下方式构建 3D 对象实例建议的策略
分层聚合和合并点云区域
来自同一物体的反向投影 2D 掩模。为了增强鲁棒性和几何同质性，我们使用

“超点” [14] 。这产生了
完整的对象实例，补充提取的
3DIS 网络。表1对 Scan-net200 数据集[58]的详细分析显示，

整合二维时，召回率，尤其是稀有类别
和 3D 提案。
为了实现开放词汇分类，我们还采用了逐点特征提取模块来构建跨 3D 点云的密集特征图。在

接下来的章节中，我们将更详细地解释我们的模块，
从 2D 引导 3D 实例建议模块开始
这是我们的主要贡献。

3.1. 2D引导3D实例建议模块

该模块以 3D 点云 P = {pn}作为输入
其中 N 是点的数量，并且pi ∈ R
3D 坐标和 RGB 颜色。此外，它还接收
RGB-D 视频序列 V = {(It, Dt, Πt)}
每个帧 t 包含 RGB 图像It、深度图Dt，以及
相机矩阵Πt （即用于将 3D 点投影到图像上的内在矩阵和外在矩阵的乘积

平面）。输出包括K1 个二进制实例掩码
以K1 × N二进制矩阵M1表示（图2-3 ）。

超点。在预处理步骤中，我们利用
[14]的方法将点分组为几何同质区域，称为超点（图2-1 ）。这产生了

一组 U 超点{qu}
是点的二进制掩码。超点增强了处理
提高我们管道后期阶段的效率，并做出贡献
到格式良好的候选对象实例。

每帧超点合并。对于所有输入帧，
我们利用预先训练的二维实例分割器，采用
Grounding-DINO [46]和 SAM [36]。网络输出一组 2D 掩模（图 2-2 ）。对于每个 2D 掩模，
索引 m（在所有帧中唯一），我们计算 IoU
ou,m与每个超点qu在投影所有点时
使用已知相机将qu投射到掩模 m 的图像平面上
矩阵，不包括相机视野之外的点，
并确定包含投影点的图像像素。
超点被认为与
如果 IoU 高于阈值ou,m > τiou，则为 2D 掩码。
然而，2D 掩模可能包括背景区域或

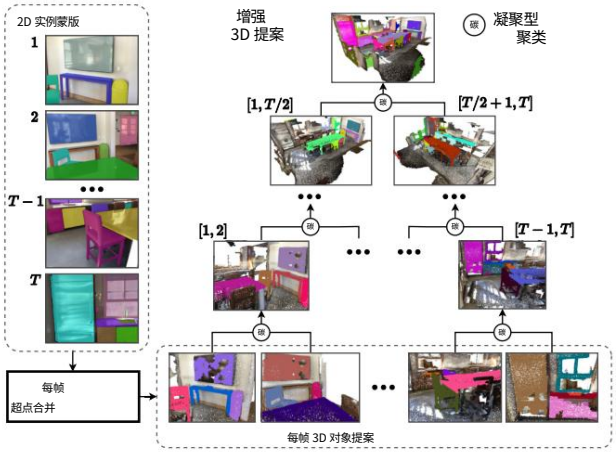

图 3. 2D 引导 3D 实例提案模块。我们使用每帧超点合并生成初始 3D 提案，

然后对 RGB-D 序列进行分层遍历
使用凝聚聚类来合并帧之间的区域集。

附近物体的部分，因此仅凭 IoU 不足以确定属于 3D 提案的超点。为了解决

为此，我们利用 3D 提议网络的 3D 主干 - N×D3D
工作[50,60]提取每个点的特征F
并测量这些超点之间的特征相似度
其特征由平均点特征决定 - 1×D3D 3D
图尔斯 ∈R 。对于每个 2D 实例掩码m2D
我们启动一个点云区域ri，其中超点与掩码具有最大的 IoU。我们扩展这个区域

通过与满足
重叠条件（τiou）并且具有最高的余弦相似度s
= maxu′εri cos(f ,f)与那些
已经处于高于阈值ri的区域（s > τsim）
（我们将在结果中讨论所有阈值的影响
部分）。增长持续到没有其他重叠或
找到相邻的超点。我们的超点合并
与单独使用点或其他合并策略（见表7）相比，该方法可以生成更完善的点

每帧对应于 2D 掩模的云区域。

3D 对象建议形成。要创建 3D 对象建议，一种选择是利用点云区域

从个体合并过程中获得
帧。然而，这会导致提案碎片化，仅捕获对象实例的部分，因为这些区域对应于
单个视图的 2D 蒙版（图2-2 ）。

解决这个问题，我们合并来自不同点云的区域
自下而上地构建框架，创造更完整、
连贯的 3D 对象蒙版。凝聚聚类将帧对中的区域集组合起来，直到没有兼容的

对保留。结果集包括合并区域和独立区域，可与其他区域集匹配

来自后续帧。在以下段落中，我们
讨论此过程中的三个关键设计选择：（a）
区域对之间的匹配得分，（b）匹配概率

区域集之间的连接,以及（c)帧的顺序或
用于匹配和合并的区域集。

匹配得分。对于一对点云区域 (ri , rj)，
我们根据（a)特征相似度和（b)重叠度定义匹配分数。它们基于特征的相似度

s 通过重新之间的余弦相似度来衡量
gions 的特征向量 f 3D ，或 我, = 余弦（f 3D ,f 3D ），哪个
依次计算为它们的点特征的平均值。虽然这衡量了区域是否属于

相同物体形状的样本可能会产生具有相同几何形状的重复样本的高相似性。为了解决这个问题，
我们还考虑了重叠程度,表示为
欠债 i,j =两个区域ri之间的IoU(ri , rj) ，RJ，
预计在重叠区域会很高
同一实例。如果两个区域
基于特征的相似度和 IoU 分数满足 > τsim
和 o i,j > τiou （与每帧检测中使用的阈值相同）
每点合并）。我们的方法结合了匹配
基于点云深度特征和几何的分数
结构,从而产生更连贯和明确的观点
云区域与其他策略的比较（见表7）。

凝聚聚类过程。合并区域集
{里} 我=1 和{rj} j=1 从不同的框架到统一的
设置{rl} 大号 l=1, 其中 L ⩽ I + J,我们采用凝聚法
聚类[49]。我们首先将它们连接成一个“活跃集” {rl}
我+我 l=1 。我们计算每个条目ci ,j
大小为 (I + J) × (I + J) 的二元成本矩阵 C 为：

$$c_{i,j} = o_{我,吉} \tau iou \odot s > \tau sim \quad , \qquad (1)$$

其中,(·)为指示函数,⊙为AND运算符。
凝聚聚类过程迭代合并
根据成本矩阵 C 在“活动集”内划分区域,并持续更新该矩阵,直到没有进一步的
合并是可能的 由 C 中不存在任何正元素来表示。

合并顺序。我们探索了两种合并策略:一种是顺序合并,即在连续帧之间合并区域集,然后将结果集进一步与

下一帧,以及层次顺序,其中包括
在单独的过程中合并非连续帧之间的区域集。分层方法形成二叉树，

每个级别合并来自连续对的集合
上一级别（见图3）。实验部分提供了详细信息和性能分析。

3.2. 3D实例分割网络

网络设计。该网络直接处理3D点
云来生成 3D 对象实例掩码。我们采用已建立的 3D 实例分割网络,如

Mask3D [60]和 ISBNet [50]作为我们的骨干。对于每个
对象候选,根据采样点计算出的核

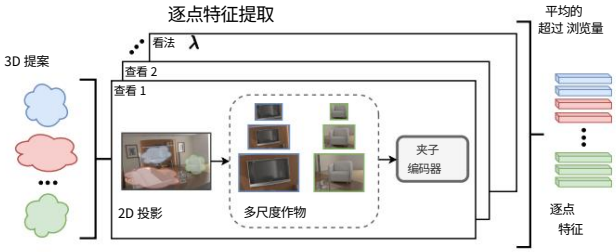

图 4.逐点特征提取。每个 3D 提案都经过投影到顶部 λ 视图上并进行多尺度裁剪[64]，
提取 CLIP 特征。然后将得到的提议特征在各个视图之间进行平均,并累积到点云特征中。

及其邻居与逐点特征进行卷积,以
预测二进制掩码。在我们的开放词汇场景中，
我们排除了语义标签头,只关注
二进制实例掩码头。输出由K2×N二进制矩阵M2中的K2个二进制掩码组成（见图2-4 ）。

合并对象实例建议。我们只需添加
将集合M2的提议添加到M1中,形成最终的 K 集合
大小为 K×N 的提案 M。注意,我们在这里应用 NMS 来删除大小为

重叠IoU阈值τdup。

3.3. 逐点特征提取

在流程的最后阶段,我们从组合提案集中为每个 3D 对象提案计算一个特征向量。每个提案的特征向量可用于各种

基于实例的任务,例如与文本提示进行比较
在 CLIP 空间中[54]。与之前的开放词汇
实例分割方法[64],它使用 top-λ
帧/视图方法,我们采用了一种更“3D 感知”的池化策略。该策略在

点云,考虑每个点的频率
在每个视图中都能看到（见图4）。我们的理由是
在顶部λ视图中更频繁出现的点应该对提案的特征向量做出更多贡献。

设 f 夹子 l,k DCLIP∈R 是 2D CLIP 图像特征
第 λ 个视图中的第 k 个实例， vλ ∈ {0, 1} 否 是可见性
视图 λ 和m3D 钾 ∈ {0, 1} 否 是第 k 个提案
M 中的二进制掩码。我们获得逐点 CLIP 特征
F 夹子 ∈R n×DCLIP 作为：

$$F^{夹子} = 内华达州_{钾} \; (nl * f^{夹子}_{l,k}) * m3D_{我} \quad , \qquad (2)$$

其中 是元素乘法（如果
必要)并且 NV(x) 是 x 的 L2 归一化向量。
文本查询 ρ 与 3D 掩码之间的最终得分
三维钾 是 CLIP 文本之间的平均余弦相似度
嵌入eρ和掩码内的所有点,具体为：

$$s^{夹子}_{k,p} = \frac{1}{|m3D_{钾}|_n} 余弦（F^{夹子} * m3D_{K}, eρ), \qquad (3)$$

其中|m3D$_{掩}$ | 是第 k 个掩码中的点数。

4.实验

4.1. 实验装置

数据集我们主要在具有挑战性的数据集 ScanNet200 [58] 上进行实验,该
数据集包含 1,201 个训练场景和 312 个验证场景,包含 198 个对象类别。

该数据集非常适合评估具有长尾分布的真实世界开放词汇场景。此外,我们
还对 Replica [62] （48 个类别)进行了实验

和 S3DIS [2] （13 个类别)与先前的方法[10, 11] 进行比较。Replica 有 8
个评估场景,而 S3DIS
包括 6 个区域的 271 个场景,其中第 5 区用于
评估。我们遵循分类方法
[11]用于 S3DIS。值得注意的是,我们省略了 Scan-NetV2 [7]上的
实验,因为它与 ScanNet200 相比相对容易
和相同的输入点云。

评估指标。我们使用 IoU 阈值为 50% 和 25% 的标准 AP 指标进行评估。此
外,我们
计算 IoU 阈值从 50% 到 95% 的 mAP

5% 增量。对于 ScanNet200,我们报告类别组特定的APhead、 APcom和APtail。

实施细节。要处理 ScanNet200 和
S3DIS 扫描效率高,我们对 RGB-D 进行了下采样
帧数增加 10 倍。我们的方法利用
Grounded-SAM 框架1。我们使用数据集类
名称作为生成 2D 实例掩码的文本提示,
随后使用τdup = 0.5 的 NMS 来处理重叠实例。我们生成超点的实现来自
[39, 55]。在逐点特征提取中,

每个提案都会被投影到所有视点,我们选择投影数量最多的前 λ=5 个视图

点。对于 CLIP,我们使用 ViT-L/14 [54]。我们遵循
OpenMask3D[64]通过将置信度得分设置为 1.0
每个 3D 提案。

4.2 与先前研究的比较

设置 1:ScanNet200。定量评估
ScanNet200 数据集总结在表2 中。根据[64],我们利用与类别无关的 3D 提议网络

在 ScanNet200 训练集上进行训练,然后在验证集上测试 OV-3DIS。采用我
们的 2D-Guided-3D
实例提案模块,Open3DIS 达到 18.2 和
在 AP 和APtail 中达到 19.2。我们的表现优于 OVIR-3D [47]和
OpenMask3D [64]在 AP 上分别以 +5.2 和 +2.8 的优势领先,
并在APtail指标中超越所有其他方法,甚至超越全监督方法。这强调了我们
的 2D-Guided-3D 实例提议模块的有效性,

1https://github.com/IDEA-Research/Grounded-
任何事物的细分

这对于制作精确的 3D 实例蒙版非常有效
独立于任何 3D 模型。结合 ISBNet 的类无关 3D 提案,可以提高我们的性能

在 AP、 AP50和AP25中分别为 23.7、29.4 和 32.8 与之前的方法相比,
AP 提高了 1.5 倍。
令人印象深刻的是,我们的方法与完全监督技术相媲美,分别达到约 96% 和
88%
ISBNet 和 Mask3D 的 AP 分数,并且在
APcom和APtail。这一表现凸显了合并 2D 和 3D 方案的优势,并展示了

我们的模型在分割稀有物体方面的熟练程度。

为了评估我们的方法的普遍性,我们进行了一项额外的实验,其中类不可
知论
3D 提案网络被训练好的网络取代
仅基于 ScanNet20 数据集。然后,我们将
ScanNet200 实例类别分为两组:基础
组,由 51 个语义相似的类组成
ScanNet20 类别,以及其余的新组
类。我们在表3中报告了APnovel、 APbase和 AP。我们的
与 PLA [11]、 OpenMask3D [64]相比,我们提出的 Open3DIS 取得了优
异的性能,且差距很大
在新类和基类中都如此。值得注意的是,PLA [11]经过训练
对比学习技巧,在
数百个小说类别。

设置 2:Replica。我们进一步在 Replica 数据集上评估了我们的方法的零样
本能力,结果如下:
详见表4。考虑到几个 Replica 类别与 ScanNet200 类别具有语义相似性,

为了保持真正的零样本场景,我们省略了此数据集的类无关 3D 提议网络（仅
使用来自 2D 的提议）。在此约束下,我们的方法

仍然优于 OpenMask3D [64]和 OVIR-3D [47]
AP 的利润率分别为 +5.0 和 +7.0。

设置 3:S3DIS。与 PLA [11]的设置一致,我们
在基类上训练了一个完全监督的 3DIS 模型
S3DIS 数据集,然后在两者上测试模型
基础类和新类。结果如表5 所示,
我们以 APB 形式报告绩效 $_{50}$ 和
$_{接入点网络}$，分别代表基础类别和新类别的AP50。Open3DIS的表现明显优于

APN 中的现有方法 $_{50}$，实现两倍以上
分数。这一出色的表现凸显了我们的方法在处理看不见的类别方面的有效性,

在2D基础模型的支持下。

我们用任意文本查询得出的定性结果。我们
可视化文本驱动的 3D 实例的定性结果
图5 中的分割。我们的模型成功地分割了
根据不同类型的输入文本提示生成实例,涉及标签中不存在的对象类别,

对象的功能、对象的分支和其他属性。

方法	设置	3D 提案 AP	AP50	AP25	APhead	APcom	APtail
国际标准书库[50]	全力支持	24.5	32.7		37.6	38.6	20.5	12.5
面具3D [60]	全力支持	26.9	36.2		41.4	39.8	21.7	17.9
OpenScene [51] + DBScan [13] †	开放词汇	没有任何	2.8	7.8	18.6	2.7	3.1	2.6
OpenScene [51] + Mask3D [60]	开放词汇	Mask3D [60]	11.7	15.2	无 14.2	17.8	13.4	11.6	9.9
SAM3D † [78]	开放词汇		6.1			21.3	7.0	6.2	4.6
OBIR-3D † [47]	开放词汇	没有任何	13.0	24.9	8.8	32.3	14.4	12.7	11.7
OpenIns3D [27]	开放词汇	面具3D [60]	10.3			14.4	16.0	6.5	4.2
OpenMask3D [64]	开放词汇	掩模3D [60]	15.4	19.9		23.1	17.1	14.1	14.9
我们的（仅限 2D）	开放词汇	没有任何	18.2	26.1		31.4	18.9	16.5	19.2
我们的（仅限 3D）	开放词汇	伊斯兰学校网络[50]	18.6	23.1	伊斯兰	27.3	24.7	16.9	13.3
我们的（2D 和 3D）	开放词汇	学校网络[50]	23.7	29.4		32.8	27.8	21.2	21.8

表 2. ScanNet200 上的 OV-3DIS 结果。 †方法在 ScanNet200 上进行了调整和评估。我们提出的方法实现了最高 AP，在所有指标上均优于以前的方法。最佳结果以粗体显示，而第二好结果则以下划线显示。

图 5. 我们的方法在开放词汇实例分割上的定性结果。我们使用任意文本查询实例掩码涉及 ScanNet200 标签中不存在的对象类别的提示。对于每个场景，我们展示具有与查询嵌入具有最高相似度得分。这些可视化强调了模型的开放词汇能力，因为它成功识别和分割在 3D 提议网络训练阶段从未遇到过的对象。

方法	预训练	APnovel	APbase	AP
打开Mask3D	ScanNet200	15.0	16.2	15.4
我们的	ScanNet200	22.6	26.7	23.7
PLA（基础 15）	ScanNet20	0.3	10.8	3.2
解放军 (20 基数)	ScanNet20	0.3	15.8	4.5
OpenScene + Mask3D	ScanNet20	7.6	11.1	8.5
打开Mask3D	ScanNet20	11.9	14.3	12.6
我们的	ScanNet20	16.5	25.8	19.0

表 3. OV-3DIS 在ScanNet200数据集上的结果，使用在 ScanNet20 上训练的类无关 3D 提议网络。

方法	3D提案	AP	AP50	AP25
OpenScene + Mask3D	掩码3D	10.9	15.6	17.3
OpenMask3D	Mask3D	13.1	18.4	24.2
奥布莱恩-3D †	没有任何	11.1	20.5	27.5
我们的（仅限 2D）	没有任何	18.1	26.7	30.5
我们的（仅限 3D）	信息服务网	14.9	18.8	23.6
我们的（2D 和 3D）	信息服务网	18.5	24.5	28.2

表 4. OV-3DIS 在Replica数据集上的结果。†我们采用源[47]的代码到这个数据集。

方法	B8/N4 亚太$_{50}$	B8/N4 接入点网络$_{50}$	B6/N6 亚太$_{50}$	B6/N6 接入点网络$_{50}$
LSeg-3D [11]	58.3	0.3	41.1	0.5
解放军[11]	59.0	8.6	46.9	9.8
低维斯3D [10]	58.7	13.8	51.8	15.8
我们的	60.8	26.3	50.0	29.0

表 5. OV-3DIS 对 S3DIS 的APB结果 $_{50}$ 和APN $_{50}$.

环境	AP	APhead	APcom	APtail
A1:OpenScene（蒸馏）	3.3	5.5	2.4	1.7
A2:OpenScene（融合）	17.5	21.5	17.1	13.3
A3:OpenScene（集成）5.6 B:Mask-wise 功能	6.4	4.8	5.7	
	22.2	25.9	19.3	21.4
C: 逐点特征	23.7	27.8	21.2	21.8

表 6. 按掩码提取和按点提取的比较
使用 Open3DIS 实例提议集进行分类的特征。

使用超点过滤条件	AP	APhead	APcom	APtail
✓ 深度特征	18.2	18.9	16.5	19.2
✓ 无欧几	15.9	16.5	14.3	17.0
✓ 里得分布	16.0	16.4	14.1	17.6
没有任何	12.0	12.6	11.2	12.2

表 7. 2D-G-3DIP 不同配置的消融。

合并策略	合并命令	AP	APhead	APcom	APtail
匈牙利	顺序		13.2	13.9	11.3
匈牙利	层次结构		16.1	16.1	13.3
凝聚	顺序		16.9	17.8	16.1
凝聚	层次		18.2	18.9	16.5

表 8. 不同合并配置的消融。

3D 段。	AP	APhead	APcom	APtail
Mask3D [60]	23.7	26.4	ISBNet [50]	22.5
	23.7	27.8		21.2

表 9. 不同 3D 分割器上的消融。

4.3. 消融研究

为了验证我们方法的设计选择，一系列消融
研究在ScanNet200的验证集上进行。

开放词汇的不同类型特征研究
分类见表6。前三行
（设置 A1-A3），我们使用 OpenScene [51]提取的逐点特征图对我们的
3D 提案。其中，融合方法直接
将 2D 图像中的 CLIP 特征投影到 3D 点上
云，AP 中结果最高，为 17.5。在设置 B 中，
我们采用类似于[64]的策略，为每个提取特征
通过将 3D 提案投影到顶部 λ 视图上进行遮罩，
其 AP 为 22.2。我们的逐点特征提取（设置 C）超越了这些，实现了最佳 AP
得分为23.7，证实了我们的设计选择。

2D引导3D实例建议模块研究
见表7。我们提出的方法（第 1 行）利用超点将 3D 点合并到区域中并过滤异常值
基于特征空间的余弦相似度，实现 AP
18.2。禁用此过滤会显著降低 AP 2.3。
相比之下，一种更基本的方法（第 3 行）依靠欧几里得距离来消除异常超点，得出
AP 为 16.0，表明欧几里得算法的有效性较低
距离用于噪声过滤。我们的基线（最后一行）仅基于 2D 掩码对 3D 点进行分组，显著降低了 AP 至 12.0，强调了超点的必要性

合并以有效地创建 3D 提案。
我们研究不同的合并配置，包括
合并策略和合并顺序见表15。具体来说，我们首先建立两个

区域集，然后匹配对被合并到新的细化区域，而不匹配的区域则保持不变。使用
与所提出的凝聚聚类相比，匈牙利匹配产生的结果较差，AP 下降了约 2.0。
采用顺序合并顺序会导致 AP 性能略有下降，约为 1.0。最佳结果是

当聚集聚类与
层次化合并顺序。

2D 段。	AP	APhead	APcom	APtail
似乎[90]	21.5	26.5	19.6	18.0
奥德赛[73]	21.6	26.0	19.5	19.1
导弹防御系统[88]	22.2	26.8	20.0	19.2
	23.7	27.8	21.2	21.8

表 10. 不同 2D 分割器上的消融。

τiou	0.3	0.5	0.7	0.9	0.95
亚太地区	17.7	17.8	18.0	18.2	16.9
AP50	25.4	25.8	25.9	26.1	24.1

表 11. τiou 上的消融。

τ模拟	0.5	0.7	0.8	0.9	0.95
亚太地区	14.2	14.6	17.2	18.2	16.2
AP50	21.0	21.8	25.1	26.1	23.8

表 12. τsim 上的消融。

查看精选	前 1 名	前 5 名	前 10 名	前 20 名	全部
美联社	21.2	23.7	22.6	22.5	22.5
AP50	27.3	29.4	28.7	29.0	29.1

表 13.顶部 λ 视图上的烧蚀选择。

分割器的消融研究。表9和表 10显示了我们对各种类别无关的 3D 分割器和开放词汇 2D 分割器的比较分析。

研究结果表明，利用 ISBNet [50]或
Mask3D [60]达到了类似的性能水平，AP 为 23.7。结合来自

SEEM [90]、 Detic [88]或 ODISE [73]导致 AP 略微下降 1.4，我们将其归因于不够精细
这些模型产生的输出。

对可见性阈值和相似性阈值的不同值进行消融研究。我们报告了性能

我们的版本仅使用来自 2D-G-3DIP 的提议
表11和表 12中可见性阈值和相似度阈值的不同值。

关于不同价值观的研究见
表13.仅依赖最高的观点
预计点数导致 AP 分数降低至 21.2。
相反，将浏览次数提高到 10 次或更多也会
产生更糟糕的结果，可能是由于劣质的存在，
遮挡的 2D 掩模。λ=5 报告最佳性能。

5.讨论

我们提出了一种 3D 场景中的开放词汇实例分割方法，该方法聚合了来自
以几何连贯的方式基于点云的实例分割器和基于二维图像的网络。

局限性。我们的类无关 3D 提案和 2D 引导 3D 实例提案模块目前独立运行，它们的输出组合起来获得

最终的 3D 提案集。更好的整合策略，
这些模块相互增强，
协同时尚，将是一个有趣的未来方向。