6. 实施细节

6.1. 与类别无关的 3D 分割器

我们采用 ISBNet [50]的架构作为我们的
由于其公开发布的实现,该网络是一个与类别无关的 3D 提案网络。该网络处理

彩色点云 $P \in R^{N\times 6}$ 并输出一系列
K 个二元 3D 实例掩码 $M \in \{0, 1\}^{额\times N}$ 。其核心是
3D UNet 主干[16],利用 3D
稀疏卷积[15],处理输入以产生特征图 F^{3D}
点云。随后,
基于采样策略的实例编码器重新细化这些特征以生成特定于实例的内核,并且

边界框参数。最后阶段涉及一个框感知动态卷积,它使用这些实例

核和掩码特征,通过相应的框预测进行增强,计算每个的二进制掩码

实例。

在推理过程中,我们利用交集而不是并集
(IoU)预测分数来滤除质量较低的掩模,
阈值为 0.2。这个分数对于
对象类别 在训练期间,IoU 预测头是
根据预测结果之间的 IoU 值进行训练
掩码及其对应的地面真相,由二分匹配算法确定。接下来,我们

使用超点[39,55]来完善我们的
实际点云结构与提案的匹配。此步骤
确保我们的分割与点云的空间组织一致。最后,我们丢弃任何

分数少于 50 分的小提案。

6.2. 开放词汇 2D 分割器

在本研究中,我们采用了四个二维开放词汇实例
分割器:Grounded-SAM2 [2],DETIC [88], SEEM [90],
和 ODISE [73]。以下是这些
使用分段器:
(a)对于接地 SAM,我们利用 Swin-B 接地
DINO 解码器[42]已在各种
数据集包括 COCO [44]、 O365 [61]、 GoldG [37, 53],
OpenImage [38]、 ODinW-35 [41]和RefCOCO [33]。
该模型用于生成边界框
给定的文本提示,框和文本阈值均已设置
到 0.4。随后,这些生成的边界框
通过 ViT-L 分段任意模型[34]
生成实例掩码。为了处理每个文本查询标题,我们将其分成多个块,每个块包含
10 个类,
适应 77 个 token 解码器的限制。最后,我们应用 IoU 的非最大抑制

阈值0.5来获得最终的边界框。
(b)对于 DETIC,我们遵循[47] 的方法,使用在 ImageNet-21K 数据集
[8]上预训练的 Swin-B 模型(该数据集有 21K 个类别)
作为文本查询。我们将置信度阈值设置为 0.5。
(c)对于 SEEM,我们采用 Focal-T 视觉解码器,
在 RefCOCO 和 LVIS [19] 上进行训练,并使用 logit
得分阈值为 0.4。与 Grounded-SAM 类似,SEEM
遵循查询处理和后处理程序。
(d)对于 ODISE,我们使用预训练标签 COCO 版本。该模型由稳定扩散补充

[57]在 LAION [59]数据集的一个子集上进行预训练,
并使用 Mask2Former [6]作为掩码生成器。我们将置信度阈值设置为 0.5。

6.3. S3DIS 和副本数据集

(a)对于缺少原始网格数据的 S3DIS 数据集,
我们应用超点方法
Transformer [56]直接从
3D 点云数据。对于具有大量点(例如 1M 个点)的场景,我们对点云进行子采样

提高 4 倍,实现高效处理。
(b)对于 Replica 数据集,我们采用基于 Felzenszwalb 和 Huttenlocher 的
高效基于图的图像分割方法[14]的网格分割工具 3来创建超点。[64] 提供了语义
和实例分割的依据。

6.4. 3D 对象提案形成过程

采用分层合并顺序和

凝聚式合并策略如图 1 所示。合并得到的 3D 点云区域

跨各个帧{r1, r2, . . . , rT }的过程,
算法将这些独立碎片的区域合并起来
(见图6)递归地将其转化为格式正确的,结果为
高质量增强 3D 提案。

6.5. 点云-图像投影

建立三维点云之间的对应关系
并且对于 RGB-D 序列 V 的每一帧,我们采用
针孔相机投影原理。给定一个 3D 点
云 $P = \{pi\}^{否}_{我=1} \in R^{N\times 6}$,对于特定帧 t,
我们考虑其深度图像 $Dt \in R^{高\times 宽}$,内在矩阵
$Kt \in R^{3\times 3}$ 和外部矩阵 $[R|c]_{吗} \in R^{3\times 4}$,在哪里
R 是一个 3D 旋转矩阵,c 是一个 3D 平移向量。
旋转和平移的复合矩阵转换
从全局框架(点云)到
相机在时间 t 的帧。我们计算投影矩阵
将 3D 点映射到 2D 图像坐标如下:

$$\Pi t = Kt \cdot [R|c]_{吗} \quad (4)$$

然后三维点的二维投影 $pi = [x^{(3d)}_{吗}, 和^{(3d)}_{吗}, 和^{(3d)}_{吗}] \in P$ 由以下公式给出:

2https://github.com/IDEA-Research/Grounded-任何事物的细分

3 https://github。 com/ScanNet/ScanNet/树/主控/分段器

算法1 3D对象提议形成

输入： T 个每帧合并的点云区域{rt} t=1。
输出:增强的 3D 提议集 r。

```
1:函数HIERARCHICAL TRAVERSE -(s:开始,e:结束)
2:      如果s = e那么
3:          否则返回rs                    ▷ 在{rt}中查找 t=1
4:
5:          m ← (s + e)/2
6:          rleft ←层次遍历(s, m)
7:          rright ←层次遍历(m+1， e)
8:          r ← (右左∪右)
9:          Cr ← COST MATRIX(r) ▷ 遵循等式 (1)
    在主论文中
10:         r ←聚集聚类(r, Cr)
11:         返回r
12:     如果结束
13:结束函数
14: r←分层遍历（1， T）
```

$$\text{和}_{\text{页}}^{(2d)} \cdot \frac{+_{\text{页}}^{(2d)}}{\text{和}_{\text{页}}^{(2d)}} = \Pi t \cdot \frac{+_{\text{页}}^{(3d)}}{\text{和}_{\text{页}}^{(3d)}} \qquad (5)$$

其中 z $_{\text{页}}^{(2d)}$ 是投影深度值,x $_{\text{页}}^{(2d)}$ ，和$_{\text{页}}^{(2d)}$ 是
2D 像素坐标。接下来,我们丢弃任何
投影超出图像边界,定义为
+$_{\text{页}}^{(2d)}$ ∈/ [0, W − 1] 或 y$_{\text{页}}^{(2d)}$ ∈/ [0, H − 1]。为了解决该视点内的
遮挡问题,我们进一步过滤掉
它们的预测深度和实际深度之间的差异
深度图像中相应像素处记录的深度超过某个深度阈值τdepth：

$$|z_{\text{页}}^{(2d)} - Dt[\ y_{\text{页}}^{(2d)}\ ,\ x_{\text{页}}^{(2d)}\ ]| > \tau\text{深度} \qquad (6)$$

7. 附加分析

报告了深度阈值τdepth的消融研究
表14中。总体而言， τdepth = 0.1 可获得最佳性能
曼斯。

RGB-D 子采样因子的消融研究
图像显示在表15中。默认情况下,我们对
图像数量增加 10 倍。将子采样因子增加到 20 或 40 会稍微降低性能

达到 AP 成绩的 17.1 分。减少图片数量
太多会产生更糟糕的结果。我们还报告了总运行时间
（以小时为单位)对最后一列的 Scan-Net200 整个验证集进行推理。

τdepth	AP	APhead	APcom	APtail
0.2	17.4	17.7	15.6	19.3
0.1	18.2	0.05	18.9	16.5	19.2
18.7	0.025		17.7	16.4	22.8
17.7	0.01	16.7	17.6	17.6	18.6
			16.3	13.8	21.2

表 14. 消融对深度阈值τdepth 的影响。

使用 3D	子因子	AP	APhead	APcom	APtail	时间(h)
✓	10（默认）	23.7	27.8	21.2	21.8	20 + 2.3
	10（默认）	18.2	18.9	16.5	19.2	20 + 2
	20	17.9	17.9	16.5	19.6	10 + 1
	40	17.4	17.3	16.7	18.5	5 + 0.5
	80	16.5	16.7		15.4	17.1
	160	13.2	12.4		12.4	15.2
	320	9.0	8.6		8.0	10.7

表15.RGB-D图像的子采样因子研究。

在 ScanNet200 [58]和 Scan-Net++ [80]上进行类无关评估我们进一步
检查了 Open3DIS 在 ScanNet200 和 Scan-Net++ 数据集上生成的掩
码提议的质量。在 ScanNet200 中,采用 3D 主干 ISBNet,Open3DIS
(2D + 3D) 在生成高质量 3D 提议方面表现出优于现有方法的性能,如表
16所示。在 ScanNet++ 中，

与以前的方法不同,我们仅使用 100 个子采样
每个 3D 场景有 2D RGB-D 帧（为了提高计算效率）。仅使用 2D 数据的
结果显示出良好的前景
结果,如表17所示。

评估 2D 中类别无关掩码的质量
上下文中,我们利用 2D-G-3DIP 生成的所有掩码
模块无需任何后处理,通常会产生
尽管以牺牲准确率为代价,但召回率仍然很高。
3D 掩膜,我们从 ISBNet 中选择了前 100 个掩膜
他们的信心得分。随后,为了评估
开放词汇能力,与类别无关的掩码经过
通过选择前 k 个（其中 k 的范围大约在 300 到 600 之间)具有最高
CLIP 分数的掩码进行后处理。最终置信度分数设置为 1.0（Open-
Mask3D）。

8.定性结果

8.1 从单幅图像构建 3D 提案

为了获得高质量的 3D 增强提案，
是保证2D掩模版有效抬高的关键
从单一图像到 3D 场景。广泛的重叠
2D 掩模通常覆盖多个物体,并且由于相机不完美导致点与像素配对的敏
感性
校准是导致先前仅依赖于点的方法性能不佳的主要因素

方法	AP AP50 AP25 AR AR50 AR25
超点 DBSCAN	5.0 12.7 38.9
[13] 14.4 27.5	1.6 5.5 32.1
OVIR-3D [47] （Detic）	38.8
蒙版聚类[75] (CropFormer) 17.4 33.3 46.7	
ISBNet [50] （3D）	40.2 50.0 54.6 66.8 80.4 87.4
我们的（接地 SAM）	29.7 45.2 56.8 49.0 70.0 34.6 43.1 48.5 66.2 83.2
我们的（3D + 地面 SAM）	81.6 31.5 45.3 51.1 61.2 87.1 41.5 51.6 56.3 91.4
我们的（SAM）	74.8 90.9 97.5
我们的（3D+SAM）	97.8

表 16.ScanNet200 上的类别无关评估[58]（2024 年 3 月 19 日更新）。

方法	AP AP50 AP25 AR AR50 AR25	笔记
ISBNet [50] （3D）	6.2 10.1 16.2 10.9 16.9	25.2 预训练的 Scannet200
SAM3D [78] 14.2 29.4	7.2	
SAM 制导图切割[18] 12.9 25.3 43.6		
段3D [26] 12.0 22.7 37.8		
SAI3D [82] （SAM)17.1 31.1 49.5		
我们的（SAM） 18.5 33.5 44.3 35.6 63.7我们的（SAM） 20.7 38.6 47.1 40.8 75.7	82.7	每个场景 100 帧
	91.8	每个场景的所有帧

表 17.ScanNet++ 上的类无关评估[80]（2024 年 3 月 19 日更新）。

几何交并比（IoU）。在图7 中，SAM3D
[78]掩模版分散在很宽的区域，而 OVIR-3D
[47]掩模版噪声大且碎片化。Open3DIS，
然而，通过考虑超点并使用平均的 3D 深度特征合并它们来解决这些问题。

我们的方法在 3D 和 2D 中实现了一致性，从而
单个 2D 图像上相应掩模的 3D 点云区域明显更清晰。

8.2 在 2D-G-3DIP 中使用超点的原因

我们选择使用 3D Superpoints 作为我们创新的 2D-G-3DIP 模块的代表。

3D Superpoints 的灵感来自于它们卓越的能力
精确地概括物体的形状和边界
在 3D 场景中。本质上，当我们检查一个物体时
在 3D 环境中，我们发现 3D 超点的子集可以准确、完整地覆盖该对象的

形状，如图8 所示。
尽管深度可能会带来一些缺陷
传感器，以前的方法[47,78]通常依赖于
基于点云 - 图像投影技术生成
逐点 3D 实例掩码。然而，这种方法通常会产生一组稀疏的 3D 提案，并且一些点
可能会
变得模糊，导致掩模版不完整（见图10）。
相比之下，我们的 Open3DIS 采用了独特的方法。我们
为点组（特别是 3D 超点）分配权重，并利用 3D 深度特征和 ge-
度量交并比 (IoU) 计算。这
独特的组合使我们能够产生超点
3D 实例蒙版更加详细，
比以前的方法更精确。这些
即使在存在遮挡的情况下，蒙版也能提供 3D 场景中物体实例的更细粒度表示

和不完美之处。

8.3. ScanNet200 上的更多定性结果，副本和 S3DIS

ScanNet200。我们展示了应用于广泛的 Scannet200 数据集的 Open3DIS 的
可视化。在图9 中，我们
显示经过 Open3DIS 处理过的场景
以及它们对应的实例地面实况（实例 GT）。尽管 Scan-net200 数据集的规模
相当大，但需要注意的是，地面实况

注释可能会忽略某些相对较小的物体
在场景中。这些省略的对象代表
黑点表示没有被标记的实例。Open3DIS 利用 2D 和 3D 分割器来

生成全面的 3D 实例蒙版，确保
甚至非常小的物体也会被覆盖。尽管我们
继续使用 Scannet200 数据集进行评估，主要是因为它包含了广泛的对象类别，
我们预计 Open3DIS 将展示

应用于细粒度时性能显著优越
3D 实例分割数据集。

与其他方法相比,如图10所示
仔细观察,Open3DIS 擅长制作更精细的 3D
掩模可以有效地覆盖具有复杂和模糊几何结构的物体。另一方面,OVIR-3D 依
赖于 2D 分割器并直接扩展 2D 掩模

通过基于点的交并法实现 3D 场景
（IoU)匹配。这种方法会导致次优的掩码
质量,尽管它有能力发现稀有对象类别。
相比之下,OpenMask3D 采用 3D 实例分割器
并使用 CLIP 模型评估每个 3D 实例。
虽然这种方法在某些情况下可能会带来好处，
它损害了开放词汇 3D 实例分割 (Open-Vocabulary 3DIS) 的通用性。特别是，

OpenMask3D 可能难以识别稀有物体类别
在训练期间扩大类别数量时。
正文中的表 3 说明了这些
差异。OpenMask3D 在 Scannet20 上训练时，
平均准确率 (AP) 得分为 12.6,而
Open3DIS 以令人印象深刻的 19.0 AP 分数超越了最先进的方法。这一显著的
表现
差距凸显了Open3DIS在处理多样化和具有挑战性的3D实例分割任务方面的
优势。

复制品。我们的方法在
副本数据集如图11a 所示。

S3DIS。我们的方法在
S3DIS 数据集在图11b 中进行了可视化。

8.4. 开放词汇场景探索

我们在 ARKitScenes [3]上展示了 Open3DIS 卓越的开放词汇场景探索能力

（图12a)和 ScanNet200 [58] （图12b)数据集，
因包含大量场景而闻名
多样而稀有的物体。具体来说,我们展示了
系统根据各种属性（如材质、颜色、可供性和使用情况)查询实例对象的能力。我
们有意排除与类别无关的 3D Seg-menter 组件,从而将我们的方法推向

接近零样本实例分割方法。值得注意的是,在具有挑战性的场景中,例如识别物
体
比如一张便条纸、一张马的图片或者一瓶橄榄油
石油,Open3DIS 优于其他方法[47, 51, 64, 78]
其中一些方法很难检测到
这些物体,更不用说准确定位它们了。请参阅
现场演示的补充视频。

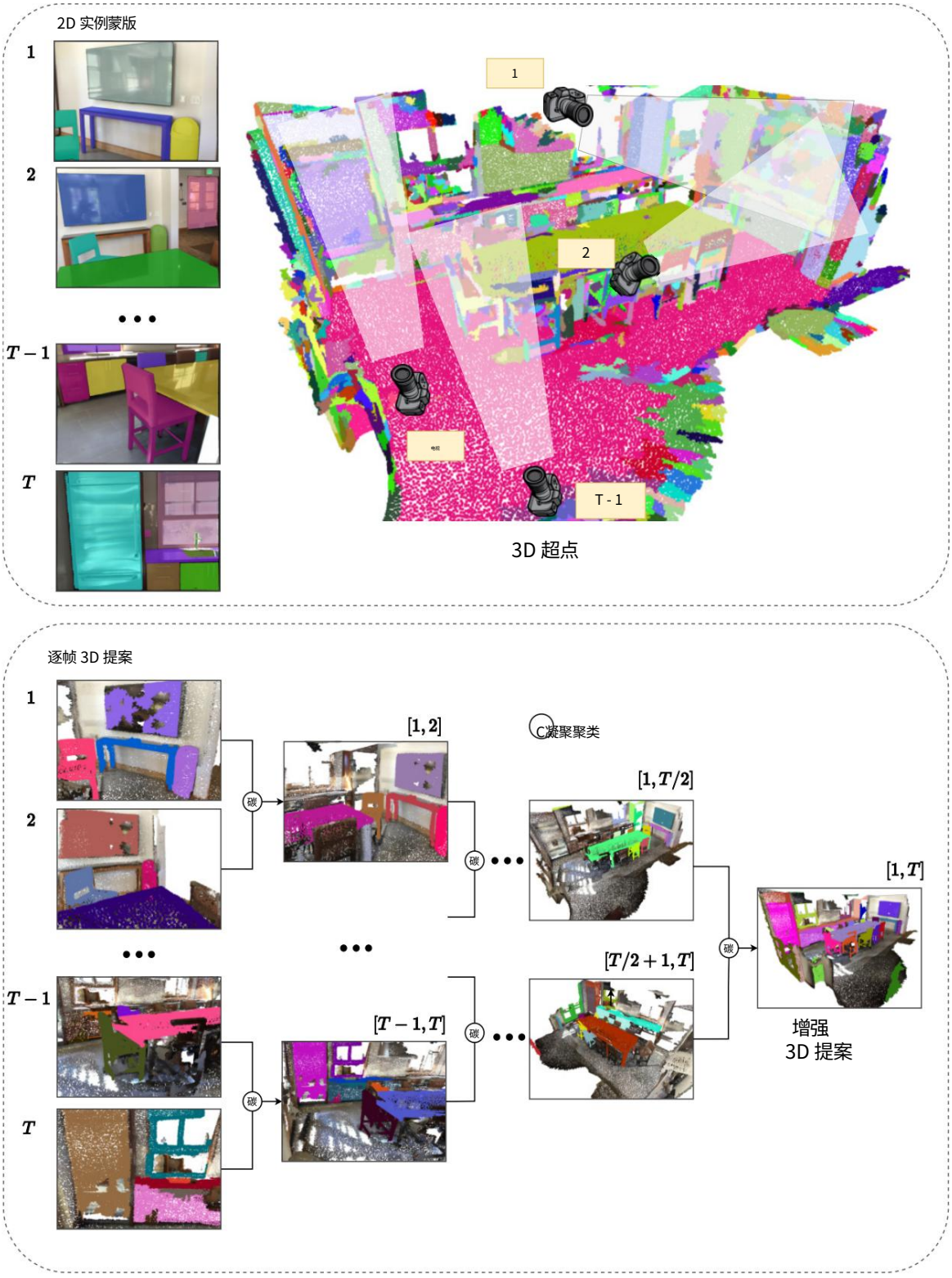

图 6.（顶部）2D-G-3DIP 模块利用 2D 每帧实例掩码，通过利用 3D 超点来生成每帧 3D 提案。（底部）我们提出的分层合并。这些提案被视为点云区域，并在多个视图中进行分层合并过程，从而产生最终的增强 3D 提案（最好以彩色显示）。