表 1:不同方法在 ScanNet200 数据集上的类无关 3D 实例分割结果。
参照[34],我们与传统聚类方法和 VFM 辅助的 3D 场景
感知方法。Speed 的单位是每帧毫秒,其中 VFM 和其他部分的速度为
另行报告。

方法	类型	视觉调频	AP AP50 AP25		速度
HDBSCAN [19]	离线 1.6 Nunes 等人 [22]	–	5.5	32.1	–
	离线 2.3 Felzenszwalb 等 [7] 离线 5.0 UnScene3D		7.3	30.5	–
[25]		–	12.7	38.9	–
	离线 15.9 DINO [2]		32.2	58.5	–
SAMPro3D [31]	离线 SAM [12] 18.0 SAI3D [34]		32.8	56.1	–
	离线语义SAM [15] 30.8 SAM3D [33]		50.5	70.6	–
	在线的	独自的	20.2 35.7 37.9		55.5 1369+1518
我们是我	在线的	独自的	58.8 35.9 56.3		75.0 1369+80
们是-E	在线的	快速SAM [39]			74.0 20+80

表 2:不同方法从 ScanNet200 到 SceneNN 和 3RScan 的数据集迁移结果。我们
直接在其他数据集上评估表1中的模型,以展示其泛化能力。

方法	类型	ScanNet200→SceneNN ScanNet200→3RScan AP AP50 AP25 AP AP50 AP25			
SAMPro3D 离线 12.6 25.8 SAI3D 离线 18.6 34.7		53.2	3.9	8.0	21.0
		65.7	5.4	11.8	27.4
SAM3D	在线 15.1 30.0 在线26.6 46.2在	51.8	6.2 13.0 10.3		33.9
我们是	线 23.4 43.0	63.1	23.6 10.2 22.4		50.7
ESAM-E		60.0			48.5

4.1 基准和实施细节

我们在四个数据集上评估了我们的方法:ScanNet [6]、ScanNet200 [24]、 SceneNN [10]和
3RScan [30]。ScanNet包含 1513 个扫描场景,其中我们使用 1201 个序列进行
训练,其余 312 用于测试。ScanNet200 提供了更细粒度的注释
ScanNet 的场景,包含 200 多个类别。SceneNN 包含 50 个高质量
扫描场景并添加实例和语义标签。根据[32],我们选择了 12 个干净的序列
测试。3RScan 是一个更具挑战性的室内数据集,其中 RGB-D 序列是通过
快速移动的摄像机。我们选择其测试集进行测试,其中包含 46 个场景。每个数据集
提供已摆姿势的 RGB-D 序列和带有标签的重建点云。

基准:我们在四个基准上比较了不同的方法。首先,我们与表 1 中的 VFM 辅助 3D 实例分割方法进行了比较。我们
在 ScanNet200 上训练了不同的方法
训练集（如果需要)并以类无关的方式在 ScanNet200 验证集上对其进行评估。
对于离线方法,每个场景的输入是重建的点云和 RGB-D 视频,其中
对重建的点云进行预测。对于在线方法,输入是流式
RGB-D 视频,我们将St上的最终预测结果映射到重建的点云上
使用最近邻插值进行比较。

由于零样本方法（如 SAM3D)不需要训练。为了公平地将它们与可学习的
方法,我们进一步在 SceneNN 和 3RScan 上评估表 1 中的模型,无需微调。这
表2所示的基准验证了不同方法的泛化能力。

我们还与表 3 中的在线 3D 实例分割方法进行了比较。
工作[16; 32],我们在 ScanNet 训练集上训练不同的方法,并在 ScanNet 上进行评估
验证集和SceneNN。

最后,我们评估了 198 个类别的开放词汇 3D 实例分割能力
表 4 中的 ScanNet200。由于 ESAM 输出与类无关的 3D 掩码,因此有两种方法可以
将其扩展到开放词汇 3D 分割。首先是将与类别无关的 3D 掩码输入到
开放词汇掩码分类模型,如OpenMask3D [29],该模型在

表 3:不同方法在 ScanNet 和 SceneNN 数据集上的 3D 实例分割结果。

方法	类型	扫描网			场景神经网络		
		AP	AP50	AP25	AP	AP50	AP25
TD3D [13]	离线	71.3	78.8	86.7	–	–	–
Oneformer3D [14]	离线	78.1	84.6	91.3	–	–	–
INS 卷积 [16]	在线 - TD3D-MA	57.4	–	–	–	–	–
[32]	在线	39.0	60.5	71.3	26.0	42.8	59.2
ESAM-E	在线	38.4	57.7	72.9	27.3	42.1	56.4
ESAM-E+FF [26]	在线	40.8	58.9	75.7	30.2	48.6	63.6

SAI3D [34]。第二种是采用开放词汇二维分割模型来获取类别
每个 2D 掩模的标签。由于 3D 掩模和 2D 掩模之间存在一一对应关系
在 ESAM 中,我们可以相应地获取每个 3D 掩模的类别标签。这里我们遵循 SAI3D
采用第一种方法,并与之进行比较。

实施细节:按照[32],我们进行训练
ESAM 分为两个阶段。首先,我们训练单视图
ScanNet(200)-25k 上的感知模型
具有单独 RGB-D 的 ScanNet(200) 子集
帧,没有基于内存的适配器和损失
三个辅助任务。接下来我们微调
RGB-D 序列上的单视图感知模型
带有适配器和全损失。为了减少内存占用,我们随机抽取 8 个相邻
每次迭代时,每个场景的 RGB-D 帧。在超参数方面,我们设置$\phi = 0.5$,
$\varepsilon = 1.75$,$\tau = 0.02$,$\alpha = 0.5$,$\beta = 0.5$。

表 4:ScanNet200 数据集上的开放词汇 3D 实例分割结果。

方法	AP	AP50	AP25
SAI3D	9.6	14.7	19.0
我们是	13.7	19.2	23.9

4.2 与最新技术的比较

我们将我们的方法与性能最佳的 VFM 辅助 3D 实例分割方法进行了比较
以及如上所述的在线 3D 实例分割方法。我们提供了三个版本的
ESAM,即ESAM、ESAM-E和ESAM-E+FF。 ESAM 采用 SAM 作为 VFM,而 ESAM-E
采用 FastSAM [39]实现实时推理。ESAM-E+FF 不仅采用了
FastSAM,还将 FastSAM 主干提取的图像特征融合到点云中,如下所示
[26]. 我们还包括一些可视化结果以进行定性评估。

根据表 1,在类别无关的 3D 实例分割任务中(即 3D“分割任何事物”
任务”) ,我们的 ESAM 与以前的方法相比建立了新的最佳水平,甚至包括
离线方法。请注意,在线方法感知 3D 场景更具挑战性
与离线替代方案相比,离线方法直接处理完整的重建 3D
场景,而在线方法处理部分和嘈杂的帧。尽管准确率很高,但 ESAM
也比以前的方法快得多。由于高效的
架构设计和快速合并策略,而像 SAM3D 这样采用手工制作的
合并策略每帧需要超过 1 秒的时间。当用更快的替代方案替换 SAM 时
FastSAM、ESAM-E 可以实现实时在线 3D 实例分割,速度约为 10 FPS,而
准确率还是比以前的方法高很多。

在泛化能力方面,ESAM 也表现出色。如表所示
2、当直接迁移到其他数据集时,ESAM 与
零样本方法。我们注意到 SAI3D 在 3RScan 数据集上的表现甚至比 SAM3D 更差,这是
因为它高度依赖干净的重建 3D 网格和精确对齐的 RGB 帧。
在 3RScan 中,相机移动速度很快,因此 RGB 图像和相机姿势很模糊。

我们在 ScanNet200 上对上述方法的预测进行了可视化,如图 4 所示。ESAM
可以预测准确且细粒度的 3D 实例分割掩模,同时能够处理
实时流式传输 RGB-D 视频。我们还提供了在线可视化,以进一步展示
图5中显示了ESAM的实用性。更多细节可以在我们的视频演示中查看。

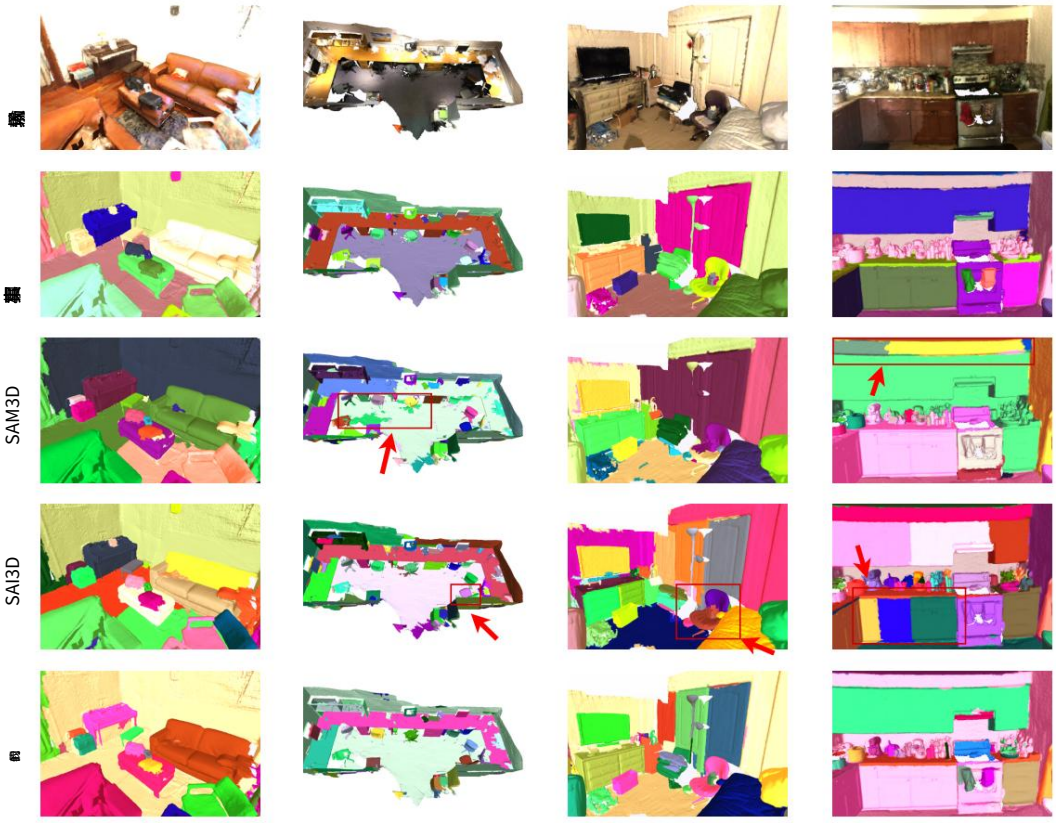


图4:不同3D实例分割方法在ScanNet200数据集上的可视化结果。
如红框所示,SAM3D 预测了噪声掩模,而 SAI3D 则倾向于过度分割
实例分成多个部分。

如表 3 和表 4 所示,与
以前的在线 3D 实例分割方法和开放词汇 3D 实例分割
方法。

4.3 ESAM分析

数据高效学习。我们通过仅使用 20% 或 50% 的训练集来减少训练
样本,并报告
表 5 中ESAM 在 Scan-Net200 上的类无关性能。结果
表明,
即使 ESAM 性能下降,
有限的训练数据。这是因为 2D VFM 已经提供了良好的初
始化,因此学习
ESAM 的一部分很容易收敛。

表 5:经过训练的 ESAM 的表现
包含部分训练集。

比例	AP	AP50	AP25
100%	37.9	58.8	75.0
50%	37.0	58.4	75.4
20%	34.4	55.8	74.2

推理时间分解。我们分解了不包括 VFM 的 ESAM 推理时间
在表 6 中。时间感知主干由稀疏卷积 U-Net 和几个
基于内存的适配器。合并过程包括相似度计算、二分匹配
和掩码/表示更新。由于高效的设计,解码器和合并操作
ESAM 仅占用一小部分推理时间。

消融研究。我们首先对 ESAM 进行消融研究,以验证
提出的方法。对于架构设计,我们在 ScanNet200-25k 上进行实验并报告
表 7 中列出了类无关 AP 和除 SAM 之外的每帧的平均推理延迟（毫秒）。
可以看出,几何感知池化将性能提升了 1.4%,同时
计算开销。请注意,单个视图上的预测误差将累积在

表 6:不包括 VFM 的 ESAM 推理时间（毫秒)的分解。

骨干		解码器	合并			全部的
3D-Unet	适配器		相似度匹配	更新		
41.0	28.0	5.0	0.7	0.3	5.0	80

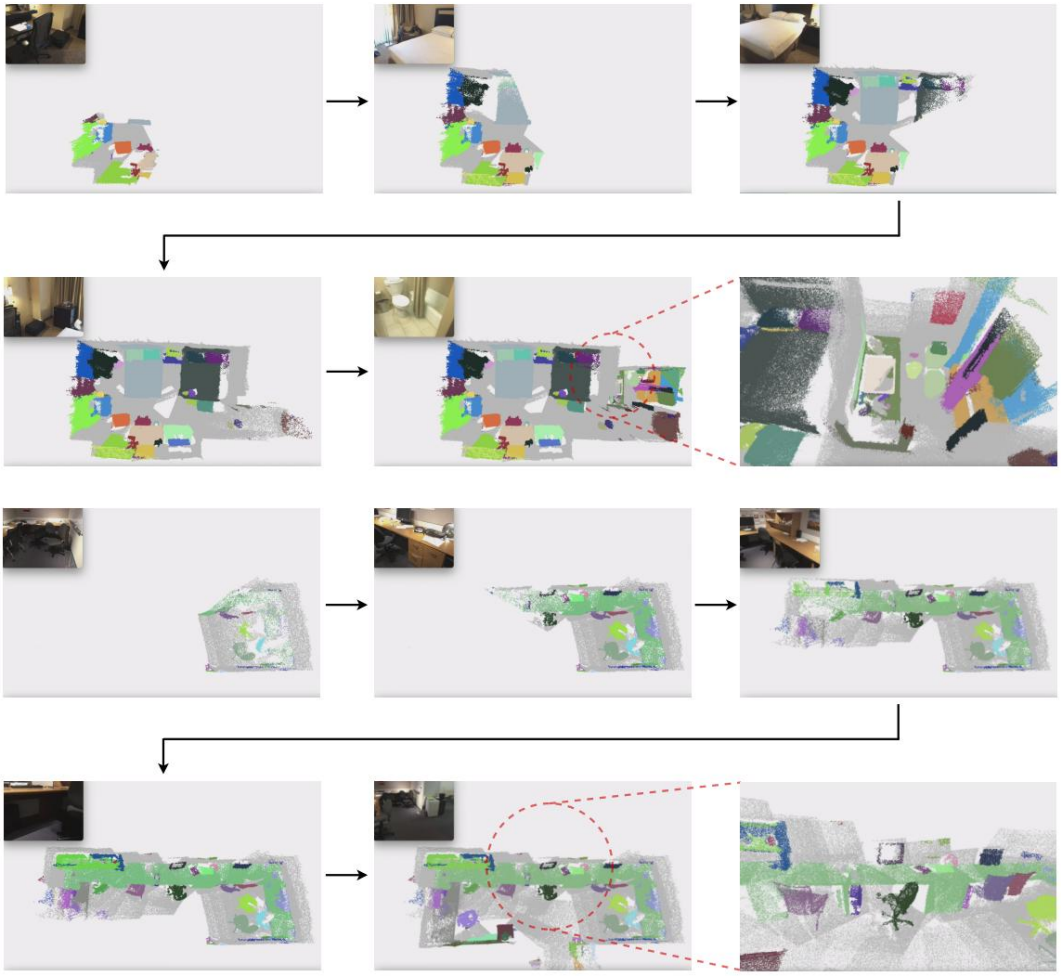


图 5:ScanNet200 数据集上的 ESAM 在线可视化。请参阅我们的视频演示
项目页面了解更多详细信息。

整个场景,因此 ScanNet200-25k 上的高 AP 对最终性能贡献很大。我们可以
还观察到,ESAM 中的双层设计与
耗时的F = FP策略,而与完全超点F = FS策略相比,延迟仅略有增加。对于合并策略,我们在 ScanNet200
上比较了不同的设计
报告了 AP,如表 8 所示。结果表明,每个辅助任务对于质量
掩模合并。我们注意到几何相似性对掩模合并的影响最为显著。
最终表现。这是因为大多数掩码对可以根据距离排除。

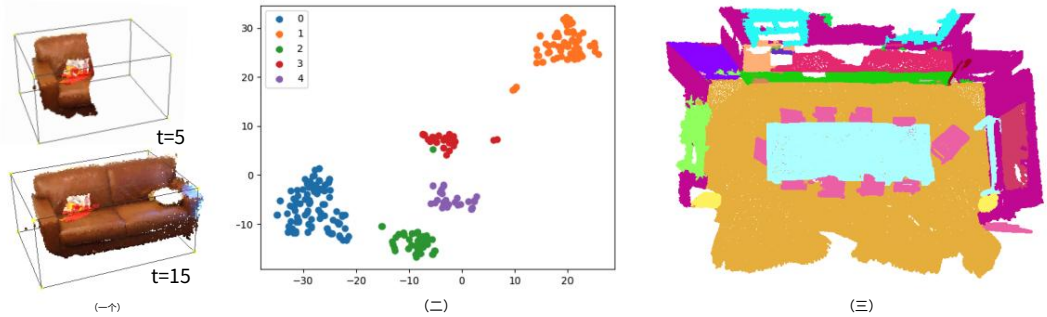
辅助任务的可视化。我们还将辅助任务的预测可视化为
全面理解 ESAM。从图 6 (a) 可以看出,该模型
仅通过部分观察就能预测物体的整体几何形状。t-SNE 可视化
图 6 (b) 验证了模型成功学习了判别性查询表征
对象匹配。最后,图 6 (c) 中的语义分割结果表明我们的 ESAM 可以
学习令人满意的语义表示并可扩展到3D语义分割任务。

表7:架构设计的效果。

方法	AP	延迟
用平均池化替换 G	56.3	仅设置
F = FS	45.6	仅设置
F = FP	58.0	最终模型
	57.7	

表 8:合并策略的效果。

方法	美联社
删除方框表示	28.7
删除对比表示	31.6
删除语义表示	34.8
最终模型	37.9



(一个) (二) (三)

图 6:我们合并策略的辅助任务的可视化。(a)3D 框预测几何相似性。我们将不同时刻的物体边界框可视化。(b) t-SNE 对对比相似性实例特定表示的可视化。不同颜色表示不同的实例,不同的点表示不同帧处的实例特征。(c) 基于查询的语义分割,用于实现语义相似性。

5 结束语

在这项工作中,我们提出了 ESAM,这是一个利用视觉基础模型的高效框架用于在线、实时、细粒度、通用和开放词汇的 3D 实例分割。我们提出利用几何感知池化将 VFM 生成的二维掩码提升到三维查询,这是然后使用双路径查询解码器来细化查询并生成准确的 3D 实例掩码。然后,根据查询-掩码对应关系,我们设计了三个辅助任务来表示每个 3D 掩码在三个判别向量中,这使得能够通过矩阵运算快速合并掩模。广泛的在四个数据集上的实验结果表明,ESAM 取得了领先的性能,在线和实时推理以及强大的泛化能力。ESAM 在开放词汇和数据高效环境中也显示出巨大的潜力。我们相信 ESAM 为如何有效地利用 2D VFM 实现具体感知。

潜在的局限性。尽管性能令人满意,但仍然存在一些局限性ESAM。首先,ESAM是否实时取决于所采用的VFM。目前我们采用SAM和 FastSAM,其中只有 FastSAM 可以实现实时推理。然而,我们认为在不久的将来,将会有更高效、性能更好、功能更多的2D VFM,并且 ESAM 可以随着 2D VFM 的改进而得到进一步的改进。其次,3D U-Net而基于内存的特征提取适配器相对较重,这占了大部分ESAM 3D 部分的推理时间。如果我们能够使主干更加高效,这留待以后的工作。

参考

[1]乔什·阿希姆 (Josh Achiam)、史蒂文·阿德勒 (Steven Adler)、桑迪尼·阿加瓦尔 (Sandhini Agarwal)、拉马·艾哈迈德 (Lama Ahmad)、伊尔格·阿卡亚 (Ilge Akkaya)、弗洛伦西亚·莱奥尼·阿莱曼 (Florencia Leoni Aleman)、迪奥戈·阿尔梅达、扬科·阿尔滕施密特、萨姆·奥尔特曼、Shyamal Anadkat 等。 Gpt-4 技术报告。 arXiv 预印本 arXiv:2303.08774,2023 年。

[2] Mathilde Caron、Hugo Touvron、Ishan Misra、Hervé Jégou、Julien Mairal、Piotr Bojanowski 和 Armand Joulin。自监督视觉转换器的新兴特性。在 ICCV,第 9650–9660 页,2021 年。

[3] Devendra Singh Chaplot、Dhiraj Prakashchand Gandhi、Abhinav Gupta 和 Russ R Salakhutdinov。目的使用面向目标的语义探索进行目标导航。NeurIPS,33:4247–4258,2020 年。