OneFormer3D: One Transformer for Unified Point Cloud Segmentation

摘要

在计算机视觉领域，三维点云实例分割是一个很重要的研究主题，在地理信息系统、医学影像、自动驾驶、增强现实和视频监控等领域有着巨大的实际应用价值。本文从点云分割的原理出发，总结了目前主要研究方法及其原理和网络架构，对已发表的主流点云处理和实例分割方法进行分析，并基于 CVPR 2024 论文 *OneFormer3D: One Transformer for Unified Point Cloud Segmentation* 的开源代码，复现并实现了多 GPU 协同训练，然后将此模型应用于开源数据集 UrbanBIS 的实例分割任务，在训练时间和分割精度上实现了比原有方法更好的效果。

关键词： 3D 实例分割；点云；Transformer；Oneformer3D

1 引言

三维点云分割的尺度主要包含三类，即语义分割、实例分割和全景分割。语义分割为每个语义类别输出一个掩码，使得点云中的每个点被分配一个语义标签。实例分割返回一组单个对象的掩码。全景分割为每个前景对象预测一个掩码，并为每个背景点预测一个语义标签。上述三种分割任务实际上都隐含着预测一组掩码这一步骤，但它们通常是完全不同的架构。语义分割方法依赖于 U-Net 网络。实例分割方法将语义分割模型与基于聚类、目标检测或 Transformer Decoder 的聚合方案相结合。全景分割则是结合语义分割和实例分割的结果，将三维点云分为 stuff 类和 thing 类，为每一个 stuff 类的点预测语义标签，为每一个 thing 类的点预测语义和实例标签。

实例分割不是一个孤立的目标检测任务，而是一个从语义分割出发，逐步精细化的自然演变过程。实例分割的思想来源于分类，其本质是在对每个对象进行分类或对输入中的有序列表进行交替命名时，预测整个输入的过程。三维实例分割也与三维语义分割和三维目标检测等任务紧密相连。三维语义分割服务于预测点的语义标签，但它不会分割不相似的实例。相反，三维目标检测预测每个特定对象的边界框，但无法呈现目标对象的详细掩码。因此，三维实例分割可以看作是三维目标检测和语义分割的集成任务。

本文所选择复现的论文 *OneFormer3D: One Transformer for Unified Point Cloud Segmentation* 是第一个多任务统一 3D 分割框架，通过一个单一的网络架构同时处理多个 3D 视觉任务，避免了为每个任务单独设计模型的复杂性。Oneformer3D 使用 SPFormer 作为 baseline，在 Transformer decoder 中并行添加语义查询和实例查询，以统一预测语义和实例分割掩码 [**?**]。

然后，确定了基于 Transformer 的三维实例分割性能不稳定的原因，并通过新的查询选择机制和新的高效匹配策略解决了问题。最后，提出了一个只需训练一次的单一统一模型，即使它们专门针对每个任务进行了调整，也能优于三维语义、三维实例和三维全视角分割方法。

在本文选择复现的工作中，官方开源的代码仅支持单卡训练，无法充分利用 GPU 资源。另一方面，Oneformer3D 的开源代码仅支持 Scannet、Scannet200 [1] 和 S3DIS [2] 三个室内场景数据集。本文将从这两个角度出发，对原工作进行改进。

2 相关工作

2.1 点云处理深度学习方法

点云是无秩序的、非结构化的，这意味着直接应用标准 CNN 是不可行的。2017 年，能够直接在非规则点云上进行处理的 PointNet 提出后，点云语义分割领域基本上都是对 PointNet 的一系列改进。改进的方向主要包含四种：点 MLP 方法、点卷积方法、基于 RNN 的方法和基于图的方法。

2.1.1 点 MLP 方法

PointNet 开创性地将深度学习直接用于三维点云任务 [3]，提出了解决点云刚性变换 (旋转或平移) 不变性的 T-net 和解决点云置换不变性的 max-pooling，设计了能够进行分类和分割任务的网络结构。PointNet++ 在 PointNet 的基础上结合 2D CNN 感受野的思想，解决了其无法获得局部特征，很难对复杂场景进行分析的问题 [4]。在 PointNet 的基础上加入了多层次特征提取结构，其作用是不断的提取局部特征，然后扩大局部范围，最后得到一组全局的特征。PointNeXt 讨论了训练策略改进对于模型性能提升的影响 [5]，同时提出了 InvResMLP 对 PointNet++ 的模型进行缩放，优化其网络结构，在 S3DIS 数据集上取得了更好的分割效果。后来，提出了一种高效且轻量级的网络，称为 RandLA-Net [6]，用于大规模点云分割。这个网络利用随机点取样，在内存和计算方面取得了显著的高效率。进一步提出了一个局部特征聚合模块来捕捉和保留几何特征。

2.1.2 点卷积方法

PointCNN 以类似于 PointNet 中 T-Net 的思路，提出了通过 X 矩阵来将邻居点的 feature 矩阵变得与邻居点的数据无关，这种方法被称作 X-Conv [7]。A-CNN 认为 PointNet 中的多尺度采样过程实际上是有重叠的，同时只进行一次 PointNet 操作不能很好的编码局部的点云结构，因此提出了环形卷积方法，可以在每个本地环形区域内定义任意大小的卷积核，从而有助于获得对 3D 物体更优的几何表示 [8]。PointConv 提出了一种将 2D CNN 扩展到 3D 点云中的方法，并提出 efficient PointConv 结构，降低了 PCNN 卷积的显存占用 [9]。KPConv 定义了一个基于 kernel points 的显式卷积核，通过在中心点为球心的球体内取若干核心点的方式计算中心点特征，将针对图像的可变形卷积扩展到点云领域 [10]。

2.1.3 基于 RNN 的方法

3P-RNN 构建了一个高效的金字塔池化模型来提取 3D 点云的局部信息，再通过一个双向的 RNN 提取空间的点云全局依赖性 [11]。两个 RNN 通过不同的方向扫描 3D 空间提取信息，最终达到良好的 3D 语义分割的效果。RSNet 将无序的点云特征映射为有序的特征向量序列，以便可以应用传统的端到端学习算法，提取局部相关性特征的计算复杂度相对较小 [12]。

2.1.4 基于图的方法

由于点云是离散的，缺乏拓扑关系，通过建立点与点之间的联系，可以增强点云的表达能力，DGCNN 提出了一个新的操作 EdgeConv，在保证置换不变性的同时捕获局部几何信息，同时 EdgeConv 可以被集成，嵌入多个已有的点云处理框架中 [13]。不同于此前的工作，SPG 不是去逐点进行分类，而是将多个点组成的点集看作一个完整的整体，对每个点集再进行分类，并且可以描述相邻物体之间的关系 [14]。SPG 的大小是由场景中简单结构的数量来确定的，而不是点的总数，基于 SPG 这种表达方式，可以在不损失主要精细细节的情况下对大场景的点云运用深度学习进行处理，从而解决大规模点云的分割问题。GACNet 通过建立每个点与周围点的图结构，并通过引入注意力机制计算中心点与每一个邻接点的边缘权重，从而使得网络能在分割的边缘部分取得更好的效果 [15]。

2.2 点云实例分割方法

与语义分割相比，实例分割更具有挑战性，因为它需要更精确和更细粒度的点推理，不仅需要区分具有不同语义的点，还需要分离具有相同语义的实例。总的来说，现有的方法可以分为两类：基于候选区域提案的方法和无候选区域提案的方法。

2.2.1 基于候选区域提案的方法 (Region proposal based methods)

该方法一般分为两个主要步骤，分别是候选区域提案 (边界框提案、对象提案) 和精化任务 (掩码预测)。3D-BoNet 的原理正是遵循这种思路，该方法同时进行三维边界框的回归和场景中所有对象点级掩码的预测，而不需要聚类、特征采样、非极大值抑制或投票等后处理任务 [16]。然而，3D-BoNet 也存在一些限制，例如无法学习不同输入点云的权重，缺乏先进的特征融合组件来同时改善语义和实例分割。GSPN 利用一种名为基于区域点网络 (R-PointNet) 的创新结构，允许可调整的建议增强和实例分割生成 [17]。引入了 Point RoIAlign 层来积累特征，并允许网络处理提案。GSPN 通过加强几何理解，从而排除低客观性的提案，达到更好的分割效果。PanopticFusion 通过将 2D 全景分割扩展到 3D 映射，实现了大规模三维重建、语义标注和实例分割。此方法首先利用 2D 语义和实例分割网络获得像素级的全景标签，然后将这些标签集成到体素地图中 [18]。进一步使用全连接 CRF 实现精确分割，该系统能够实现高质量的语义映射和判别性目标识别。

LIDARSeg 是一个基于室外 LiDAR 点云的三维实例分割网络，用于小物体的分割和定位，该方法使用自注意力模块学习点云鸟瞰视图上的特征表示，根据预测的水平中心和高度限制得到最终的实例标签 [19]。3D-SIS 可用于 RGB-D 扫描的实例语义分割。其主要贡献是能够同时从几何和 RGB 输入中学习，从而允许精确的实例估计 [2]。该网络为所有对象建立

一个边界框回归，然后进行实例 (掩码) 分割。由于该网络是全卷积的，这意味着它可以在单个镜头中运行良好，适用于大规模的 3D 环境。GICN 基于高斯实例中心网络，将在整个场景中传播的实例中心的分布近似为高斯中心热图 [20]。中心实例大小预测、包围盒生成和最终实例掩码由预测的热图得到。

总体而言，基于候选区域提案是一类直观明了的方法，实例分割结果通常具有较好的客观性。然而，这些方法需要进行多级训练，并对冗余提案进行剪枝。因此，它们通常是耗时且计算成本昂贵的。

2.2.2 无候选区域提案的方法 (Region proposal free methods)

无候选区域提案方法没有目标检测模块，这些方法通常将实例分割视为语义分割后的后续聚类步骤。现有的大多数方法都是基于属于同一个实例的点应该具有非常相似的特征这一假设，因此，这些方法的关注点主要集中在判别性特征学习和点分组上。

SGPN 首先为每个点学习一个特征和语义图，然后引入一个相似度矩阵来表示每个成对特征之间的相似度 [21]。为了学习更有判别力的特征，它们使用双铰链损失来相互调整相似度矩阵和语义分割结果。最后，采用启发式非极大值抑制方法将相似点合并为实例。由于构造相似矩阵需要较大的内存消耗，因此该方法的可扩展性受到限制。MASC 是一种新颖的基于简单高效的过程来学习点与点之间相似度，从而将它们组合成实例的网络 [22]。该网络采用稀疏卷积，并提出了一种基于学习到的多尺度亲和力的聚类算法来解决 3D 实例分割问题。但由于聚类算法是顺序执行的，即使在理论上是并行的，网络也是缓慢的。

PointGroup 网络由语义分割分支和偏移预测分支组成 [23]。为了获得更好的分组结果，进一步使用了双集合聚类算法和 ScoreNet，从而将每个点推送到自己的对象质心。然而，该模型需要一个精化部分来缓解影响实例组合的语义不准确的问题。ISBNet 区别于传统的实例分割方法，采用了一种不需要点聚类的方法，将实例表示为内核并通过动态卷积解码实例掩码 [24]。提出了一种名为“实例感知最远点采样”的简单策略来对候选进行采样，并使用受 PointNet++ 启发的本地聚合层对候选特征进行编码。同时提出了一种在动态卷积中预测和利用 3D 轴对齐边界框的方式来提高性能。

综上所述，无候选区域提案是一类节省计算资源的方法。其优点是不需要计算昂贵的区域提案。然而，由于这些方法没有显式地检测对象边界，因此这些方法分组的实例片段的客观性通常较低。

3 本文方法

3.1 本文方法概述

如图 1所示，基线组件以蓝色表示，改进内容用红色突出显示。OneFormer3D 框架继承自 SPFormer [25]（3D 实例分割网络），因为它具有直接的流水线、快速推理以及在训练和推理期间内存占用较小的特点。

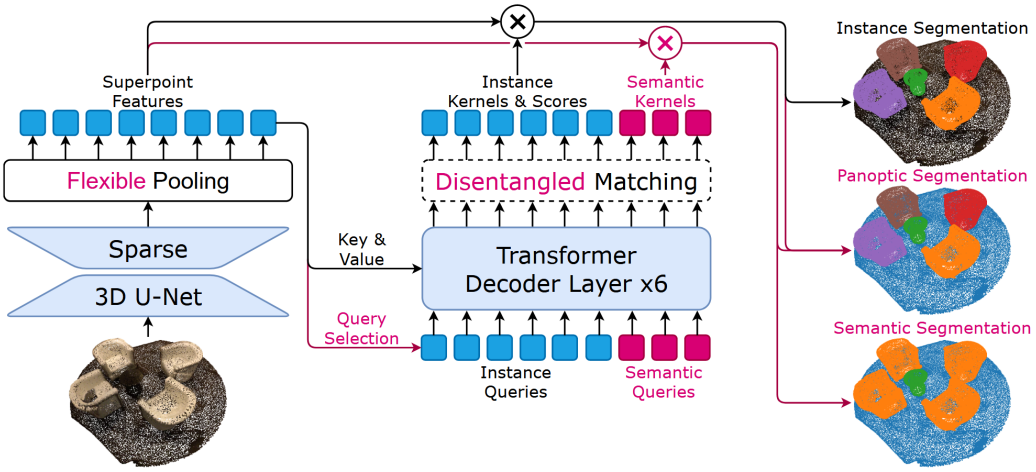

图 1. 方法示意图

本框架使用稀疏 3D U-Net 提取逐点特征，这些特征通过灵活池化传递，通过简单地对超点中的特征进行平均，获得超点特征。超点特征作为 Transformer decoder 的键和值，该解码器还接受可学习的语义和实例查询作为输入。解码器通过交叉注意机制捕获超点信息，并输出一组学习到的核，每个核代表一个单一对象掩码的实例标识（来自实例查询）或语义区域（来自语义查询）。采用分解匹配策略以端到端的方式训练实例核。因此，经过训练的 OneFormer3D 可以无缝地解决语义、实例和全景分割。

3.2 解耦匹配

先前最先进的基于 2D Transformer 的方法和基于 3D Transformer 的方法利用基于匈牙利算法的二分匹配策略。这种常用的方法有一个主要缺点：提案和真实实例之间过多的有意义的匹配使得训练过程持久且不稳定。Oneformer3D 使用了一个简单的技巧，降低了预测实例和真是实例之间匹配的需要。由于实例查询是用超点的特征初始化的，因此该实例查询可以与该超点明确匹配。假设一个超点只能属于一个实例，这给出了超点和地面实况对象之间的对应关系。通过将所有内容放在一起，模型可以在真实对象、超点、实例查询和从该实例查询派生的实例提案之间建立对应关系。最后，通过跳过中间对应关系，可以直接将实例提案与地面实况实例进行匹配。获得的对应关系解开了提案和真实实例的二分图，作者将其称为解耦匹配。

尽管如此，提案的数量仍然超过了真实实例的数量，因此需要过滤掉与真实对象不对应的提案以获得二分匹配。解耦匹配技巧简化了成本函数优化，因为可以将成本矩阵中的最多权重设置为无穷大：

$$\hat{\mathcal{C}}_{ik} = \begin{cases} \mathcal{C}_{ik} & \text{if } i\text{-th superpoint} \in k-\text{th object} \\ +\infty & \text{otherwise} \end{cases} \tag{1}$$

3.3 损失函数定义

将提案与真实实例进行匹配后，最终可以计算实例损失。分类错误会受到交叉熵损失 $\mathcal{L}_{cls}$ 的惩罚。此外，对于提案和真实实例之间的每次匹配，将超点掩码损失计算为二进制交叉熵

$\mathcal{L}_{bce}$ 和 Dice 损失 $\mathcal{L}_{dice}$ 之和。语义损失 $\mathcal{L}_{sem}$ 被定义为二元交叉熵。总损失 $\mathcal{L}$ 的公式为:

$$\mathcal{L} = \beta \cdot \mathcal{L}_{cls} + \mathcal{L}_{bce} + \mathcal{L}_{dice} + L_{sem} \tag{2}$$

其中，β 设定为 0.5。

4 复现细节

4.1 与已有开源代码对比

本文在复现 Oneformer3D 的工作时主要参考了 Oneformer3D 的官方开源代码。在原始代码中，作者基于 mmdetection3d 框架实现了 Oneformer3D 的功能，但仅支持单 GPU 训练，无法进行分布式训练。在研究了代码结构以及 mmdetection3d 的官方文档后，基于 pytorch-lightning 包实现了 Oneformer3D 的分布式训练。

此外，原始的 Oneformer3D 代码仅支持 Scannet、Scannet200 和 S3DIS 三个室内数据集的分割任务，缺少在室外大场景数据集中的实验和应用。UrbanBIS [26] 是用于大规模三维城市理解研究的综合数据集，涵盖六个真实城市场景，总面积 10.78 平方公里，共有 3370 座建筑物。数据集由 113346 个航空摄影测量视图、25 亿点云和 2.8 亿三角面片构建而成。该数据集不仅提供了丰富的城市对象语义注释，每座建筑也都被赋予了实例标签，这是首个拥有高精建筑物级别三维实例标注的真实城市大场景数据集，也是第一个引入了细粒度建筑用途子类别标注的 3D 数据集。

本文参考原代码中的数据预处理方式，将 UrbanBIS 数据集处理成 S3DIS 数据集的形式，首次将 Oneformer3D 运用于室外大场景数据集，并取得了优良的分割效果。

4.2 实验环境搭建

Python 3.7+
CUDA 10.0+
PyTorch 1.8+
mmdetection3d v1.1.0

4.3 数据集转换

S3DIS 数据集一共分为 6 个 Area，每个 Area 内分割成不同的场景，整个场景的点云存在场景的根目录下，按照标签分割的点云存在 Annotation 目录下，以物品种类和编号作为文件名，同时作为语义和实例标签。UrbanBIS 分为五个区域，每个区域划分了训练集和测试集，同一个区域的训练集和测试集加起来是整个区域的点云。实例标签仅标注了建筑，建筑以外的点标记为-100。

数据集预处理方式如下：把 UrbanBIS 的五个区域作为 5 个 S3DIS 中的 Area，分割后的小块作为和 S3DIS 中 “office” 级别的区域，将实例标签还原成实例类别名称保存为文件名。然而 S3DIS 中的语义标签和实例标签都标注了 13 个类别，但 UrbanBIS 里只有建筑标注了实例标签，因此将建筑物按照 building_1、building_2 ... 的方式标注实例标签，其它语义的类别全部标注为一个实例，例如将整个区域内的车辆全部标注为 vehicle_1。

最终，数据经过预处理后得到如下文件结构。

```
项目根目录/
  data/
    urbanbis/
      collect_indoor3d_data.py
      indoor3d_util.py
      instance_mask/
      meta_data/
      points/
      s3dis_data/
      s3dis_infos_Area_1.pkl
      s3dis_infos_Area_2.pkl
      s3dis_infos_Area_3.pkl
      s3dis_infos_Area_4.pkl
      s3dis_infos_Area_5.pkl
      seg_info/
      semantic_mask/
      Stanford3dDataset_v1.2_Aligned_Version
```

如下图所示，S3DIS 形式的数据集中实例的 GT 被单独提取，对应 UrbanBIS 中的建筑实例被单独提取。

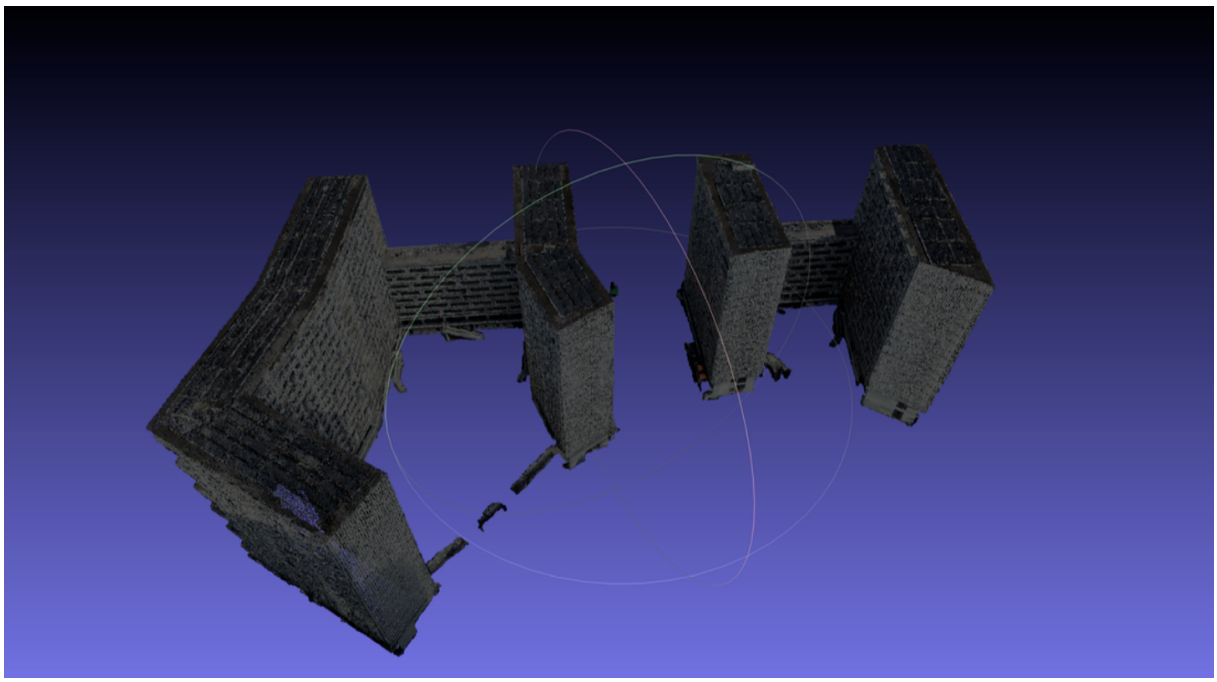

图 2. 数据集中建筑实例 GT

4.4 创新点

在复现 Oneformer3D 的过程中，本文取得了以下创新成果：

实现分布式训练：原始 Oneformer3D 代码仅支持单 GPU 训练，限制了模型的扩展性和训练效率。本文通过深入研究代码结构和 mmdetection3d 的官方文档，采用 pytorch-lightning 框架，成功实现了 Oneformer3D 的分布式训练，提高了模型训练的效率和可扩展性。

扩展至室外大场景数据集：原始代码仅支持 Scannet、Scannet200 和 S3DIS 等室内数据集，缺乏在室外大场景中的应用。本文将 UrbanBIS 数据集处理成与 S3DIS 相似的格式，使 Oneformer3D 首次应用于大规模三维城市理解任务，并在 UrbanBIS 数据集上取得了优异的分割效果，验证了模型在复杂城市环境中的适用性。

5 实验结果分析

本文的实验结果如下图所示，其中 B-Seg 方法为 UrbanBIS 数据集的原生方法，表中的实验结果为 Qingdao 场景的分割结果；Oneformer3D 的实验结果为在 S3DIS 数据集的分割结果；Mine 代表 Oneformer3D 模型在 UrbanBIS 的 Qingdao 场景的分割结果。从表中可以看

表 1. 实验结果

方法	mAP_{25}	mAP_{50}	mAP
B-Seg	0.453	0.550	0.672
Oneformer3D	0.593	0.781	0.864
Mine	0.687	0.799	0.876

出，将 Oneformer3D 的框架扩展到室外大场景数据集，可以取得与原工作相近甚至更好的分割结果，从数据集的角度看，新的分割框架可以大幅提升 UrbanBIS 的分割结果。

分割结果可视化如下图所示。

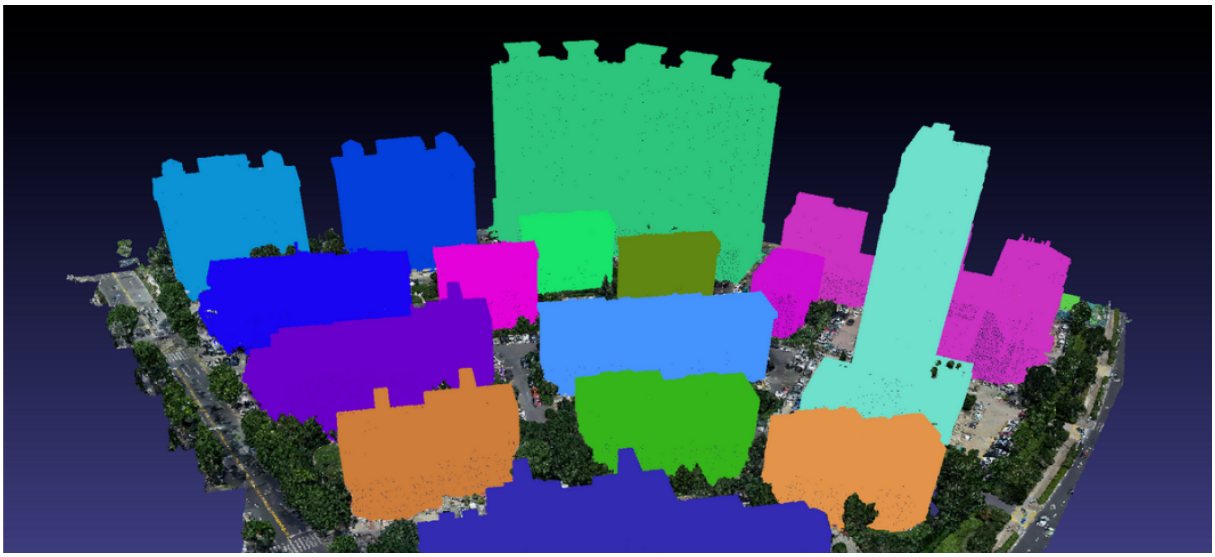

图 3. 分割结果

6 总结与展望

本文复现了 Oneformer3D 的工作，并将其应用于大规模室外三维城市数据集 UrbanBIS 中。通过对原始 Oneformer3D 代码进行改进，成功实现了分布式训练，克服了单 GPU 训练的限制。与此同时，本文通过将 UrbanBIS 数据集处理成类似 S3DIS 数据集的格式，首次在室外大场景数据集上验证了 Oneformer3D 的分割能力，并获得了良好的实验效果。实验结果表明，Oneformer3D 在大规模城市场景分割任务中具有较强的适应性和较为优秀的性能。

尽管本文的实验结果表现出 Oneformer3D 在室外大场景数据集上的有效性，但仍然存在一些可以改进的地方。目前实验仅在 UrbanBIS 数据集上进行，未来可以将 Oneformer3D 扩展到更多室外大场景数据集，以进一步验证模型的鲁棒性和泛化能力。

此外，在观察实验的分割结果后，发现在一些建筑的分割中，出现了丢失某些外立面的情况，如下图所示，这极大阻碍了分割效果的提升。

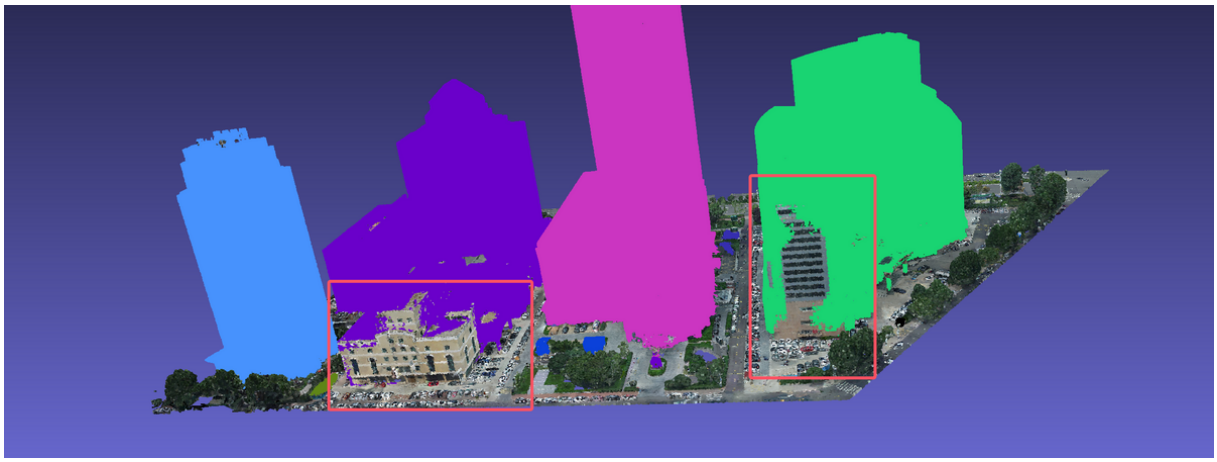

图 4. 外立面丢失可视化

在分析原理和方法后，推测有可能是因为直接将室内场景的点云分割方法运用与室外大场景数据集中，原始的体素划分以及聚类方法不能很好的适应建筑物的分割。未来可以在这个方向上对模型进行进一步改进，研究实例尺度与分割网络参数之间的关系，进一步提升分割模型的鲁棒性和分割效果。

此外，本文在复现 Oneformer3D 的工作时，重点关注了其在点云实例分割任务上的效果，而 Oneformer3D 是一个端到端的统一分割框架，在语义分割和全景分割上仍有其探索空间。未来可以在这个方向上进一步研究 Oneformer3D 在室外大场景的语义分割和全景分割表现，以拓展其使用场景。

参考文献

[1] Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas A. Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2432–2443, 2017.

[2] Ji Hou, Angela Dai, and Matthias Nießner. 3d-sis: 3d semantic instance segmentation of rgb-d scans. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4416–4425, 2018.

[3] C. Qi, Hao Su, Kaichun Mo, and Leonidas J. Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 77–85, 2016.

[4] C. Qi, L. Yi, Hao Su, and Leonidas J. Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. In *Neural Information Processing Systems*, 2017.

[5] Guocheng Qian, Yuchen Li, Houwen Peng, Jinjie Mai, Hasan Hammoud, Mohamed Elhoseiny, and Bernard Ghanem. Pointnext: Revisiting pointnet++ with improved training and scaling strategies. *ArXiv*, abs/2206.04670, 2022.

[6] Qingyong Hu, Bo Yang, Linhai Xie, Stefano Rosa, Yulan Guo, Zhihua Wang, Agathoniki Trigoni, and A. Markham. Randla-net: Efficient semantic segmentation of large-scale point clouds. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11105–11114, 2019.

[7] Yangyan Li, Rui Bu, Mingchao Sun, Wei Wu, Xinhan Di, and Baoquan Chen. Pointcnn: Convolution on x-transformed points. In *Neural Information Processing Systems*, 2018.

[8] Artem Komarichev, Zichun Zhong, and Jing Hua. A-cnn: Annularly convolutional neural networks on point clouds. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7413–7422, 2019.

[9] Wenxuan Wu, Zhongang Qi, and Fuxin Li. Pointconv: Deep convolutional networks on 3d point clouds. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9613–9622, 2018.

[10] Hugues Thomas, C. Qi, Jean-Emmanuel Deschaud, Beatriz Marcotegui, François Goulette, and Leonidas J. Guibas. Kpconv: Flexible and deformable convolution for point clouds. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6410–6419, 2019.

[11] Xiaoqing Ye, Jiamao Li, Hexiao Huang, Liang Du, and Xiaolin Zhang. 3d recurrent neural networks with context fusion for point cloud semantic segmentation. In *European Conference on Computer Vision*, 2018.

[12] Kehua Guo, Changchun Shen, Bin Hu, Min Hu, and Xiaoyan Kui. Rsnet: Relation separation network for few-shot similar class recognition. *IEEE Transactions on Multimedia*, 25:3894–3904, 2023.

[13] Rongting Zhang, Guangyun Zhang, Jihao Yin, Xiuping Jia, and Ajmal S. Mian. Mesh-based dgcnn: Semantic segmentation of textured 3-d urban scenes. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–12, 2023.

[14] Loic Landrieu and Martin Simonovsky. Large-scale point cloud semantic segmentation with superpoint graphs. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4558–4567, 2017.

[15] Appendix for : Graph attention convolution for point cloud semantic segmentation. 2019.

[16] Bo Yang, Jianan Wang, Ronald Clark, Qingyong Hu, Sen Wang, A. Markham, and Agathoniki Trigoni. Learning object bounding boxes for 3d instance segmentation on point clouds. In *Neural Information Processing Systems*, 2019.

[17] L. Yi, Wang Zhao, He Wang, Minhyuk Sung, and Leonidas J. Guibas. Gspn: Generative shape proposal network for 3d instance segmentation in point cloud. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3942–3951, 2018.

[18] Gaku Narita, Takashi Seno, Tomoya Ishikawa, and Yohsuke Kaji. Panopticfusion: Online volumetric semantic mapping at the level of stuff and things. *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 4205–4212, 2019.

[19] Feihu Zhang, Chenye Guan, Jin Fang, Song Bai, Ruigang Yang, Philip H. S. Torr, and Victor Adrian Prisacariu. Instance segmentation of lidar point clouds. *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 9448–9455, 2020.

[20] Shih-Hung Liu, Shan Yu, Shao-Chi Wu, Hwann-Tzong Chen, and Tyng-Luh Liu. Learning gaussian instance segmentation in point clouds. *ArXiv*, abs/2007.09860, 2020.

[21] Weiyue Wang, Ronald Yu, Qiangui Huang, and Ulrich Neumann. Sgpn: Similarity group proposal network for 3d point cloud instance segmentation. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2569–2578, 2017.

[22] Chen Liu and Yasutaka Furukawa. Masc: Multi-scale affinity with sparse convolution for 3d instance segmentation. *ArXiv*, abs/1902.04478, 2019.

[23] Li Jiang, Hengshuang Zhao, Shaoshuai Shi, Shu Liu, Chi-Wing Fu, and Jiaya Jia. Pointgroup: Dual-set point grouping for 3d instance segmentation. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4866–4875, 2020.

[24] T.D. Ngo, Binh-Son Hua, and Khoi Duc Minh Nguyen. Isbnet: a 3d point cloud instance segmentation network with instance-aware sampling and box-aware dynamic convolution.

2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 13550–13559, 2023.

[25] Jiahao Sun, Chunmei Qing, Junpeng Tan, and Xiangmin Xu. Superpoint transformer for 3d scene instance segmentation, 2022.

[26] Guoqing Yang, Fuyou Xue, Qi Zhang, Ke Xie, Chi-Wing Fu, and Hui Huang. Urban-bis: a large-scale benchmark for fine-grained urban building instance segmentation. *ACM SIGGRAPH 2023 Conference Proceedings*, 2023.