

中图分类号: TP242 文献标识码: A 文章编号: 1006-8961(2024)01-0218-13

论文引用格式: Wan J H, Liu X P, Chen L L, Ao S, Zhang P and Guo Y L. 2024. Geometric attribute-guided 3D semantic instance reconstruction. Journal of Image and Graphics, 29(01):0218-0230(万骏辉, 刘心溥, 陈莉丽, 敖晟, 张鹏, 郭裕兰. 2024. 几何属性引导的三维语义实例重建. 中国图象图形学报, 29(01):0218-0230)[DOI:10.11834/jig.230106]

几何属性引导的三维语义实例重建

万骏辉¹, 刘心溥², 陈莉丽³, 敖晟¹, 张鹏^{1*}, 郭裕兰²

1. 中山大学电子与通信工程学院, 深圳 518107; 2. 国防科技大学电子科学学院, 长沙 410005;
3. 军事科学院国防科技创新研究院人工智能研究中心, 北京 100071

摘要: 目的 语义实例重建是机器人理解现实世界的一个重要问题。虽然近年来取得了许多进展,但重建性能易受遮挡和噪声的影响。特别地,现有方法忽视了物体的先验几何属性,同时忽视了物体的关键细节信息,导致重建的网格模型粗糙,精度较低。针对这种问题,提出了一种几何属性引导的语义实例重建算法。方法 首先,通过目标检测器获取检测框参数,并对每个目标实例进行检测框盒采样,从而获得场景中对应的残缺局部点云。然后,通过编码器端的特征嵌入层和Transformer层提取物体丰富且关键的细节几何信息,以获取对应的局部特征,同时利用物体的先验语义信息来帮助算法更快地逼近目标形状。最后,本文设计了一种特征转换器以对齐物体全局特征,并将其与前述局部特征融合送入形状生成模块,进行物体网格重建。结果 在真实数据集ScanNet v2上,本文算法与现有最新方法进行了全面的性能比较,实验结果证明了本文算法的有效性。与性能排名第2的RfD-Net相比,本算法的实例重建指标提升了8%。此外,本文开展了详尽的消融实验以验证算法中各个模块的有效性。结论 本文所提出的几何属性引导的语义实例重建算法,更好地利用了物体的几何属性信息,使得重建结果更为精细、准确。

关键词: 场景重建; 三维点云; 语义实例重建; 网格生成; 目标检测

Geometric attribute-guided 3D semantic instance reconstruction

Wan Junhui¹, Liu Xinpu², Chen Lili³, Ao Sheng¹, Zhang Peng^{1*}, Guo Yulan²

1. School of Electronics and Communication Engineering, Sun Yat-sen University, Shenzhen 518107, China;
2. College of Electronic Science and Technology, National University of Defense Technology, Changsha 410005, China;
3. Artificial Intelligence Research Center, National Innovation Institute of Defense Technology, Academy of Military Sciences, Beijing 100071, China

Abstract: Objective The objective of 3D vision is to capture the geometric and optical features of the real world from multiple perspectives and convert this information into digital form, enabling computers to understand and process it. 3D vision is an important aspect of computer graphics. Nonetheless, sensors can only provide partial observations of the world due to viewpoint occlusion, sparse sensing, and measurement noise, resulting in a partial and incomplete representation of a scene. Semantic instance reconstruction is proposed to solve this problem. It converts 2D/3D data obtained from multiple

收稿日期: 2023-03-20; 修回日期: 2023-06-05; 预印本日期: 2023-06-12

* 通信作者: 张鹏 zhangpeng5@mail.sysu.edu.cn

基金项目: 国家自然科学基金项目(U20A20185, 61972435); 广东省基础与应用基础研究基金项目(2022B1515020103); 广东省科技计划项目(2019B121203006); 深圳市科技计划资助项目(KQTD20190929172704911)

Supported by: National Natural Science Foundation of China (U20A20185, 61972435); Guangdong Basic and Applied Basic Research Foundation (2022B1515020103); Science and Technology Planning Project of Guangdong Science and Technology Department under Grant Guangdong Key Laboratory of Advanced IntelliSense Technology (2019B121203006)

sensors into a semantic representation of the scene, including modeling each object instance in the scene. Machine learning and computer vision techniques are applied to achieve high-precision reconstruction results, and point cloud-based methods have demonstrated prominent advantages. However, existing methods disregard prior geometric and semantic information of objects, and the subsequent simple max-pooling operation loses key structural information of objects, resulting in poor instance reconstruction performance. **Method** In this study, a geometric attribute-guided semantic instance reconstruction network (GANet), which consists of a 3D object detector, a spatial Transformer, and a mesh generator, is proposed. We design the spatial Transformer to utilize the geometric and semantic information of instances. After obtaining the 3D bounding box information of instances in the scene, box sampling is used to obtain the real local point cloud of each target instance in the scene on the basis of the instance scale information, and then semantic information is embedded for foreground point segmentation. Compared with ball sampling, box sampling reduces noise and obtains more effective information. Then, the encoder's feature embedding and Transformer layers extract rich and crucial detailed geometric information of objects from coarse to fine to obtain the corresponding local features. The feature embedding layer also utilizes the prior semantic information of objects to help the algorithm quickly approximate the target shape. The attention module in the Transformer integrates the correlation information between points. The algorithm also uses the object's global features provided by the detector. Considering the inconsistency between the scene space and the canonical space, a designed feature space Transformer is used to align the object's global features. Finally, the fused features are sent to the mesh generator for mesh reconstruction. The loss function of GANet consists of two parts: detection and shape losses. Detection loss is the weighted sum of the instance confidence, semantic classification, and bounding box estimation losses. Shape loss consists of three parts: Kullback-Leibler divergence between the predicted and standard normal distributions, foreground point segmentation loss, and occupancy point estimation loss. Occupancy point estimation loss is the cross-entropy between the predicted occupancy value of the spatial candidate points and the real occupancy value. **Result** The experiment was compared with the latest methods on the ScanNet v2 datasets. The algorithm utilized computer aided design (CAD) models provided by Scan2CAD, which included 8 categories, as ground truth for training. The mean average precision of semantic instance reconstruction increased by 8% compared with the second-ranked method, i. e., RfD-Net. The average precision of bathtub, trash bin, sofa, chair, and cabinet is better than that from RfD-Net. In accordance with the visualization results, GANet can reconstruct object models that are more in line with the scene. Ablation experiments were also conducted on the corresponding dataset. The performance of the complete network was better than the other four networks, which included a GANet that replaced ball sampling with box sampling, replaced the Transformer with PointNet, and removed the semantic embedding of point cloud features and feature transformation. The experimental results indicate that box sampling obtains more effective local point cloud information, the Transformer-based point cloud encoder enables the network to use more critical local structural information of the foreground point cloud during reconstruction, and semantic embedding provides prior information for instance reconstruction. Feature space transformation aligns the global prior information of an object, further improving the reconstruction effect. **Conclusion** In this study, we proposed a geometric attribute-guided network. This network considers the complexity of scene objects and can better utilize the geometric and attribute information of objects. The experiment results show that our network outperforms several state-of-the-art approaches. Current 3D-based semantic instance reconstruction algorithms have achieved good results, but acquiring and annotating 3D data are still relatively expensive. Future research can focus on how to use 2D data to assist in semantic instance reconstruction.

Key words: scene reconstruction; three-dimensional point cloud; semantic instance reconstruction; mesh generation; object detection

0 引言

语义实例重建是一项综合性任务,它的目的是

从部分观测中估计场景中实例物体的语义类别和姿态,并恢复这些物体的完整形状。作为机器人与现实世界的三维环境交互的关键步骤(龙霄潇等, 2021),已广泛应用于许多视觉应用中,如虚拟现实、

视觉导航和室内设计。然而,由于室内空间结构的复杂性、高遮挡和多变性,语义实例重建仍然是一个开放的挑战。

在过去的几年里,关于语义实例重建问题有许多解决方案。一种经典的方法是基于二维图像的场景实例重建。Gkioxari 等人(2019)提出的 Mesh R-CNN (mesh region based convolutional neural networks)先预测二维图像中的实例物体,之后再预测物体的三维形状。然而这种方法缺乏准确的深度信息,会导致物体的位置定位出现偏差。因此,之后的研究工作更侧重于对三维数据的处理。与图像相比,点云的几何信息有助于精准定位物体(Guo 等, 2020)。然而,点云具有稀疏性和不规则性等特点,在点云上直接使用基于网格的卷积神经网络(convolutional neural network, CNN)进行场景实例重建存在一定的困难。因此一种方法是用体素网格或截断的符号距离函数(truncated signed distance function, TSDF)来表示三维场景,并使用三维卷积神经网络(three dimensional convolutional neural networks, 3D-CNN)预测体素标签,从而在实例层面上重建场景。Hou 等人(2020)提出的 RevealNet 首先对稠密体素网格进行实例分割,然后用 3D-CNN 进行实例重建。由于内存成本随体素分辨率的增加而不断增长,因此这种方法的计算成本巨大。此外,该方法可恢复的实例分辨率十分有限,无法重建出场景中的高保真物体。另一种方法是通过检索最相似的 CAD (computer aided design)模型来替换真实场景中的残缺物体。Avetisyan 等人(2019)提出的 Scan2CAD 首先基于 CAD 模型和原始扫描对场景内所有实例进行姿态预测,然后根据估计的姿态对每个模型进行对齐匹配。然而,这些方法在真实扫描和 CAD 模型之间存在几何保真度方面的差异。此外,形状匹配的效率也与 CAD 数据集的规模有关。

Nie 等人(2021)提出的 RfD-Net 在原始点云上同时完成了目标检测和语义物体重建任务。其首先在场景点云里定位物体,然后采样物体的局部点云以提取局部细节特征,并融合物体的全局先验特征进行网格重建。该方法可以识别和重建高分辨率网格物体,取得了突出的效果。然而该方法未能充分考虑到场景中物体的几何属性,导致无法重建出高保真的物体。首先,场景中的物体具有不同的种类和几何尺寸,这就导致了它们具有较大的形状差异。

其次,物体的摆放位置也十分复杂,例如部分书架放置在桌子上方,许多椅子放置在桌子下方,桌子和椅子之间互相嵌入,但 RfD-Net 采用统一尺度球体采样的方法,并没有有效地利用物体的先验几何尺度信息,从而采样到大量属于其他物体的噪声点,丢失了大量有效的物体前景点,导致重建质量效果一般。此外,其在提取局部细节特征时对全部前景点采用最大池化操作,这会导致网络难以有效地提取出物体关键点的几何信息。实际上,在基于点云的语义实例重建的算法流程中,场景中的物体具有丰富的先验几何和语义信息。因此,三维目标检测器提取出来的检测框可作为实例的先验几何尺度信息,以引导进行更精确的局部点云采样。Nie 等人(2020)在基于二维图像进行实例重建的方法 Total3D (total 3D understanding)中,融合了先验语义信息,引导网络生成更高质量的物体模型。Liu 等人(2022)提出的 AGFA-Net (adaptive global feature augmentation network for point cloud completion)以物体点云关键的几何坐标信息作为引导,利用 Transformer 结构自适应地增强物体的细节和全局特征,有效地提升了物体点云补全的性能。Fu 等人(2023)提出的 VAPCNet (viewpoint-aware 3D point cloud completion network)学习部分点云的抽象表征,以区分空间中的各种视角,然后利用准确的视角信息和点云的局部几何特征来预测缺失的点,进一步提升了物体点云补全的性能。

针对这些问题和思路,本文提出一种几何属性引导的语义实例重建算法 (geometric attribute-guided network, GANet)。本文算法可以从场景的稀疏物体点云中重建出高质量的物体网格,其主要创新点如下:1)为了有效利用物体的几何细节和语义属性信息,针对场景物体的特点,重新设计了物体检测框的采样方案。此外,为了利用关键点的几何信息引导细节特征的融合,设计了一种基于 Transformer 的三维点云特征提取方法。该方法捕获了物体中更多的高质量前景点,从而增强了特征的表征性,有效地提升了物体重建的性能。2)针对场景物体重建存在特征空间不一致的问题,提出了一个特征转换模块将全局特征和局部几何特征进行对齐。在特征融合时,保证了全局特征和局部特征对物体重建的贡献度一致,提升物体重建的性能。3)为了证明所提算法的鲁棒性和有效性,在真实点云场景

数据集ScanNet v2(Dai等,2017a)上对物体进行了重建。实验结果表明,本文算法的实例重建平均精度均值(mean average precision, mAP)比RfD-Net方法高8%。

1 相关工作

1.1 三维形状补全

三维形状补全任务的目的是从一个物体的部分扫描中恢复其完整的形状。目前的形状补全方法可以分为基于点、基于体素、基于符号距离函数(signed distance function, SDF)和基于网格的方法。这些方法大多遵循一个共同的算法流程。具体来说,首先对不完全扫描的点云用卷积神经网络进行编码,然后预测不可见部分、孔洞和缺陷表面。Yuan等人(2018)提出的PCN(point completion network)网络是一种开创性的点云形状补全方法。然而,该方法通常只能补全有限个数的点,无法表征高分辨率的物体。Dai等人(2017b)以稠密体素网格作为输入,构建三维编码器—解码器的网络结构以进行物体形状补全,得到了具有高分辨率的三维物体网格。Mescheder等人(2019)提出了OccNet(occupancy network)用于恢复物体在任意分辨率下的连续形状。它将物体的真实点云编码为隐式特征,然后预测重建物体的候选空间采样点的占用情况,最后采用移动立方体算法(marching cubes algorithm)(Lorensen和Cline,1987)进行网格生成。三维形状补全可以生成物体的完整三维模型,本文算法利用其完成三维实例重建中的重建任务。具体来说,即使用OccNet的解码器进行空间点的占用预测,并最终生成网格模型进行实例重建。

1.2 场景补全

场景补全不是关注单个物体,而是从一个输入场景的2D/3D观测中恢复可见和不可见场景(包括物体)的完整几何形状。前期工作曾使用过多种复杂的插值方法来完成场景补全任务。目前的主流方法均采用深度学习策略来提升场景补全的效果。Roldão等人(2022)根据输入的数据格式,将当前的方法分为基于二维表示的方法、基于三维网格的方法和基于三维点云的方法。基于二维表示的方法以深度图或范围图像作为输入,并基于二维卷积神经网络构建整个场景补全的框架。Han等人(2019)输

入单一深度图像,联合三维场景体积重建和三维深度预测,完成场景补全任务。基于三维网格的方法主要是利用三维占用网格或TSDF网格来编码输入数据。这种表示方式的优点是可以方便地转化为二进制网格,便于压缩。最后基于三维卷积神经网络框架,在可见区域和遮挡区域预测所有体素网格的语义和实例标签。SPSG(self-supervised photometric scene generation)由Dai等人(2021)提出,其首先通过体积融合将RGB-D(RGB-depth)扫描和相机姿态用TSDF表示,然后利用二维视图做引导,以自监督学习的方式估计遮挡区域的场景几何和颜色,从而将输入的RGB-D图像重建为高质量彩色三维模型。由于点云的稀疏性和不规则性,基于原始三维点云的场景补全方法较少。Zhong和Zeng(2020)提出了一种语义场景补全方法SPCNet(semantic point completion network)。该方法采用编码—解码器结构,其中一个观测点编码器用于提取观测点的特征,另一个遮挡点解码器将前述特征映射到遮挡点上。在此基础上,SPCNet进一步引入图像点特征,并结合纹理信息以提升语义场景补全的性能。

场景补全是指推断出场景(包括背景和物体)的三维语义的同时,补全其三维形状的任务。本文的研究目标是场景实例形状重建,即从输入数据中恢复出场景中的物体实例及其三维形状(不包括背景)。这两个任务都是基于三维点云或RGB-D图像的,其相关性在于它们都需要对输入数据进行深度理解,以便更好地推断出场景的三维语义和物体实例几何结果。

1.3 语义实例重建

除了对单一物体进行形状补全和对整个场景进行恢复之外,还有一种语义实例重建任务,这也是本文关注的重点。该任务首先需要场景内的实例进行定位,随后对被定位的实例进行重建。现有一些方法用CAD模型代替场景里的实例以实现语义实例重建。Scan2CAD(Avetisyan等,2019)通过设计的3D-CNN网络学习真实物体和合成物体之间对应的热力图,最后基于热力图通过位姿优化算法对齐相应的CAD模型完成重建。另一类方法则通过相应的网格重建网络生成实例结果。Hou等人(2020)提出的RevealNet将三维扫描离散到TSDF网格中,利用3D-CNN预测语义实例的占用网格情况。Nie等人(2020)提出的Total3D在二维RGB图像上进行物

体检测和网格重建。RfD-Net(Nie等,2021)采用直接处理点云的方式,它将候选方案编码到潜在空间中,并使用隐函数预测候选空间采样点的占用情况以进行网格重建。Tang等人(2022)提出的DIMR(disentangled instance mesh reconstruction)使用了更多的标注信息,先对场景点云进行实例和语义分割,然后利用一个预训练的网络自编码器来完成网格重建。

2 本文方法

2.1 网络框架

本文提出的基于几何属性引导的语义实例重建网络GANet的整体框架如图1所示。网络以真实场景点云作为输入,输出场景内物体的网格模型。具

体而言,1)首先使用点云三维目标检测器获取实例目标的特征 F_p ,生成场景内的物体检测框,从而定位候选对象。2)随后通过目标筛选策略来选择高质量的三维检测框,并利用检测框的几何信息引导前景点采样,分组和对齐检测框内的局部点云。物体局部点云的几何和语义属性信息被输入到去噪器中,进行前景点分割。3)进一步地,设计了一个基于Transformer的特征提取器以提取点云特征。该特征提取器在物体关键点的几何属性引导下学习了实例点云局部特征 F_o ,并采用设计的特征转换器对齐物体的全局先验特征 F_g ,两者串联融合得到 F_m 。4)最后,使用网格生成器学习规范空间中的占用函数来表示实例的形状,并根据占用点的占用情况,利用移动立方体算法生成网格,同时根据物体几何参数将重建的网格放置在场景中。

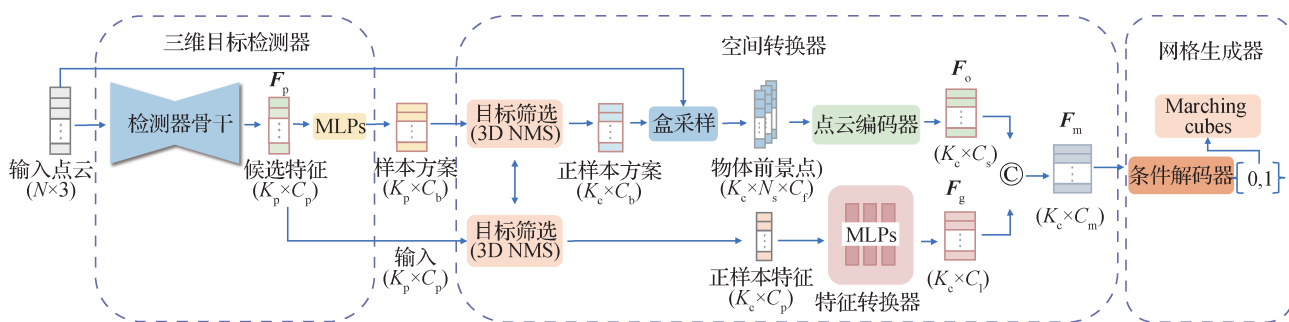


图1 几何属性引导的语义实例重建算法(GANet)框架图

Fig. 1 The framework of geometric attribute guided network for semantic instance reconstruction (GANet)

2.2 几何属性引导的空间转换

2.2.1 目标筛选

为了保证检测的准确率,目标检测器通常在初始检测阶段会产生冗余的候选目标。与图像不同,三维点云具有很强的稀疏性,很多空间实际上是没有物体实例的。如果考虑所有的候选目标,将会涉及大量的负样本。在本文算法的流程中,考虑到被检测到的候选目标需要送入形状生成模块进行物体重建,因此大量负样本会影响形状生成器的学习效果。此外,考虑全部候选目标的方案也会带来计算量过大的问题,降低了方法的效率。因此在训练阶段,本文算法会尽可能地从所有候选目标中保留更高质量的物体。具体来说,GANet会选择置信度得分较高的 K_c 个候选目标。在推理阶段,由于候选对象的数量不受限制,因此本文算法使用三维非极大值抑制(non-maximum suppression, NMS)来消除冗余的

负样本。经过目标筛选后, K_c 个正样本被保留下来。

2.2.2 几何属性引导的目标点采样

本文算法在获得 K_c 个正样本后,通过在场景点云中采样这些样本的局部点云,以获取更精细的物体信息。现有方法通常采用球体采样的方案获得局部点云。具体来说,以候选检测框的中心为球体中心,设置一个半径 $r = 1$ 的球,通过随机采样得到球内的 N_s 个候选点。实际上,这种采样策略在语义实例重建任务里存在一定的问题。因为场景里的物体种类和大小不一,对所有物体进行固定半径的球体采样并没有考虑到物体的几何属性多样性。对于过小的物体来说,会采集多余的噪声点;对于过大的物体来说,会丢失物体的关键特征点。为解决该问题,本文算法采用了盒采样的方案,利用检测框在场景内采样目标候选点,并通过几何属性引导获取更多属于物体的局部点。

盒采样与球体采样的对比如图2所示。若检测框内的点数大于 N_s , 则对框内的点进行最远点采样 (farthest point sampling, FPS); 若点数小于 N_s , 则复制里面的点进行补充。进行盒采样后, 每个实例获取到了更多的有效物体点, 同时也减少了其他物体的噪声点。之后, 本文算法将不同物体中的点对齐到一个规范统一的坐标系, 对输入点进行归一化 (包括去中心和旋转的操作)。规范空间处理消除了形状重建中不同物体空间平移和旋转的偏差, 降低了形状学习的难度。

由于从真实场景中获得的三维点往往是有噪声

的, 因此进一步利用一个去噪器来提取前景点。去噪器由 PointNet (Qi 等, 2017) 组成, 通过对每个点进行特征提取, 进而判断是否属于前景点。其中物体采样点的三维坐标 (归一化处理后) 和每个点相对于场景地面之间的高度差作为网络输入。为提高特征鉴别力, 本文算法还将物体的类别先验信息添加到输入数据中。具体来说, GANet 对检测到的物体对象类别进行一位有效编码 (one-hot), 再与前述几何特征进行串联。随后, GANet 将特征值为 0 的点视为背景点, 使用基于 Transformer 的点云编码器从预测的前景点中提取局部几何特征 $F_o \in \mathbf{R}^c$ 。

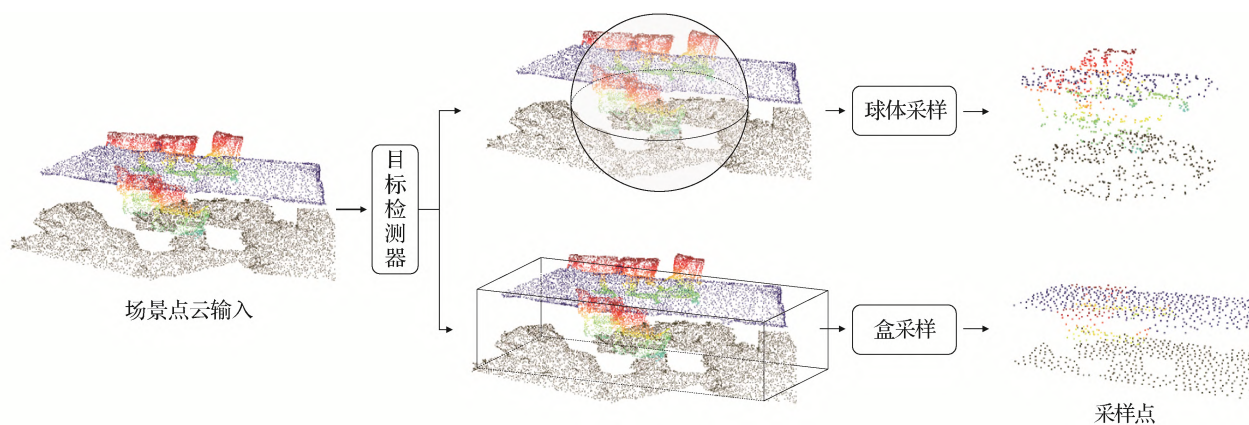


图2 盒采样与球体采样的效果对比图

Fig. 2 The comparative visualization of box sampling and ball sampling effects

2.2.3 特征转换器

目标检测器得到的物体特征实际上是包含了物体类别和大小的信息, 该特征可作为物体形状生成的先验信息。由于此特征是从整个场景原始点云中提取而来, 而从检测框内采样的点云经过了归一化处理, 且物体的局部点都处于规范空间中, 因此, 规范空间中的局部点云特征和真实空间中的物体先验特征是特征不对齐的, 这在一定程度上会干扰物体的形状生成。这里, 本文设计了一个简单的多层感知机将全局物体先验特征转换到规范空间下。考虑到全局物体先验特征和局部物体几何特征是权重相当的, 因此同时利用两个多层感知器 (multi-layer perceptron, MLP) 将全局物体先验特征从 128 维升至 512 维, 增加特征的代表性, 得到转换后的物体特征, 具体为

$$F_g = MLP_3 \left\{ MLP_2 \left\{ MLP_1 (F_p) \right\} \right\} \quad (1)$$

式中, F_p 代表目标检测器得到的物体特征, F_g 代表

转换后的物体特征。

2.2.4 物体点云特征编码器

获取到物体前景点后, 本文算法使用 AGFA-Net (Liu 等, 2022) 中基于 Transformer 的多尺度点云特征编码器来获取物体的局部细节特征。图3为点云特征编码器的结构图。特征编码器的目标是生成一个特征向量 F_o , 该特征项包含了局部点云的局部几何细节特征和全局形状特征, 用于后续的解码过程。为了增强物体特征的表达效果, 本文算法同样利用了物体的类别先验信息。具体来说, 被保留下来的物体前景点, 其几何特征与前述的语义 one-hot 向量进行串联, 得到语义几何特征矩阵 F_{ge} 。从实验结果可知, 类别编码提供了形状先验, 有助于提升物体重建的效果。其中, 特征嵌入层由两个 MLP 串联组成, 用以将输入数据映射到特征空间中, 从而更好地拟合深度网络。将 F_{ge} 输入到特征嵌入层中, 生成一个尺寸为 $N_s \times 32$ 的特征矩阵 F_{in} 。然后本文算法对 F_{ge} 进行 FPS, 利用其对应的空间索引从 F_{in} 里取出尺

寸为 $(N_s/2) \times 32$ 的关键点特征矩阵 $F_{in,0}$ 。将 F_{in} 和 $F_{in,0}$ 分别转换为 K, V (此处 K 和 V 相同) 和 Q , 送入多头注意力层 (Vaswani 等, 2017)。其中, 注意力层得到的特征矩阵与 Q 构成残差连接。为了增强编码器的表征能力, 本文算法利用残差连接的多层感知机以进一步更新特征。更新后的特征与 $F_{in,0}$ 进行串联, 送入自注意力层 (Vaswani 等, 2017), 得到第 1 阶段的特征向量 $F_{in,1}$ 。之后本文算法按照这个操作由

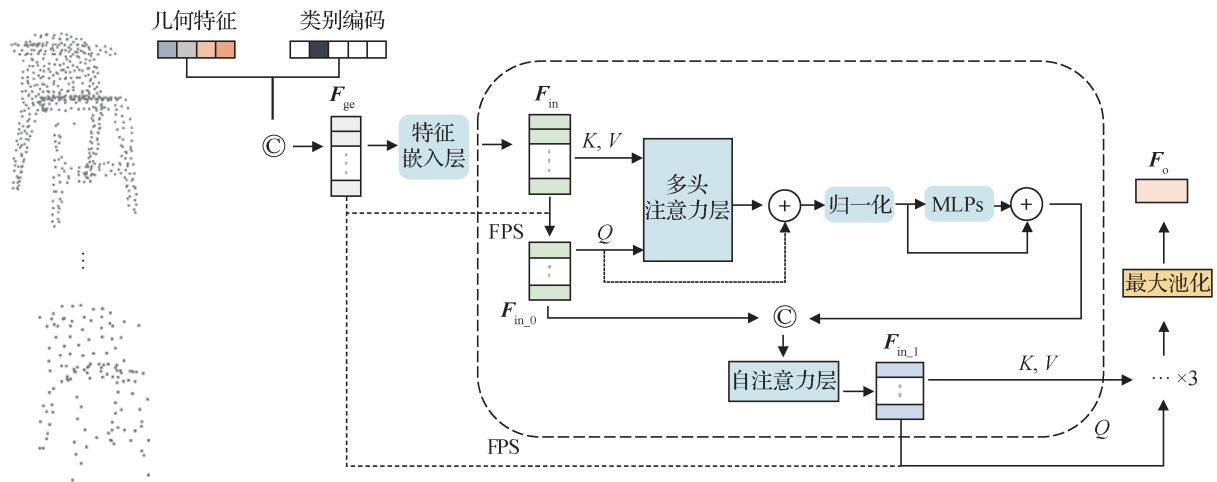


图3 基于Transformer的点云特征编码器结构图

Fig. 3 The framework of Transformer-based point cloud feature encoder

2.3 目标形状生成

本文算法参考 OccNet (Mescheder 等, 2019), 将目标的三维形状表示成一个占用函数, 形状生成的过程如下: 首先初始化一个立方体, 在立方体内统一采样空间点 $p \in \mathbf{R}^3$, 占用函数根据输入的物体特征 F_m 预测这些空间点的二元占用值 $o \in \{0, 1\}$ (0 表示物体表面外, 1 表示物体表面内)。具体来说是采用条件批处理归一化层对输入进行处理, 使解码器能根据物体的特征来回归每个候选空间点的占用值。另外, 本文算法参考 RfD-Net 将形状解码器构建为一个概率生成模型。利用 OccNet 设计的编码器, 从输入点 p 、占用值 o 和物体特征 F_m , 预测均值和标准差 (μ, α) 来近似标准正态分布。对此分布进行采样, 产生一个向量 $z \in \mathbf{R}^l$ 。用单层 MLP 将 p 和 z 处理成等维, 然后进行求和, 并用 5 个条件块来预测空间点最终的占用值 o 。在推理阶段, 本文算法将向量 z 设置为零, 并使用移动立方体算法生成物体网格模型。

本文与 RfD-Net 的区别在于: 1) 采样策略。本文采用一种几何属性引导的目标点盒采样而非固定

粗到细迭代 3 次, 不断提取并融合物体的关键点特征。最后一层经过最大池化后得到一个可表征实例的特征向量 F_o 。

最后将物体全局先验特征和物体局部细节特征进行串联, 送入网格生成器, 即

$$F_m = F_g \odot F_l \quad (2)$$

式中, \odot 代表串联操作, F_g 代表物体全局先验特征, F_l 代表物体局部细节特征, F_m 代表融合后的物体特征。

尺度的球体采样, 从而能提取到更多有效的前景点。2) 特征编码。RfD-Net 对所有点使用简单的最大池化操作, 丢失了许多几何关键信息。本文算法采用一种基于 Transformer 的点云特征编码器, 可由粗到细地提取局部细节特征。3) 特征融合。为了更好地融合局部细节特征和全局先验特征, 本文设计了一种特征转换器以进行特征对齐, 而 RfD-Net 并未考虑该问题。

2.4 损失函数

本文算法的损失函数主要由检测损失和形状重建损失两部分组成。检测损失用于监督检测器所生成的候选物体及其对应相关的检测框参数; 形状重建损失用于监督实例重建的训练过程。

2.4.1 检测损失

参考 VoteNet, 三维目标检测的损失函数包括置信度损失、边界框 (中心 c 、目标大小 s 、航向角 θ) 估计损失和语义分类 (语义标签 s_i) 损失。具体为

$$L_{det} = L_c + L_s + L_\theta + \lambda_{con} L_{con} + \lambda_l L_l + L_v \quad (3)$$

式中, 交叉熵损失用于定义置信度损失 L_{con} , Smooth-L1 损失用于定义中心损失 L_c 。本文算法将尺度损

失 L_v 和方向损失 L_o 定义为分类 (交叉熵) 损失和回归 (Smooth-L1) 损失的混合形式。同时, 交叉熵损失还用于定义语义分类损失 L_l 。此外, 本文算法利用了一个额外的投票损失 L_v 。对上述损失进行加权并相加, 得到最终的检测损失 L_{det} 。

2.4.2 形状重建损失

首先, 对于每个候选目标检测框中的点, 本文算法采用交叉熵来监督前景点分割损失 L_{seg} 。形状生成器在训练中会学习一个近似标准正态分布的均值和标准差 (μ, α) , 本文算法从此分布中采样一个潜在向量 z 以预测空间候选点的占用值。最终的形状重建损失为

$$L_{shape} = \frac{1}{N_d} \sum_{i=1}^{N_d} \left[\sum_{j=1}^K L_{ce}(\hat{o}_{i,j}, o_{i,j}) + KL(\hat{p}(z_i) \| p(z_i)) \right] + \lambda_{seq} L_{seq} \quad (4)$$

式中, KL 代表 KL 散度 (Kullback-Leibler divergence), L_{ce} 代表交叉熵损失函数。 $\hat{o}_{i,j}$ 和 $o_{i,j}$ 分别代表第 i 个正样本物体中第 j 个空间候选点的预测占用值和真实占用值。 $\hat{p}(z_i)$ 代表的是隐式向量 z_i 的预测分布, $p(z_i)$ 是与之对应的标准正态分布。

最终的损失函数由检测损失和形状重建损失加权求和组成。即

$$L = L_{det} + \lambda L_{shape} \quad (5)$$

3 实验与分析

本节首先介绍数据集的细节、损失函数、评估指标和实验细节。然后将本文算法 GANet 与几种最先进的方法进行比较。最后通过消融实验来证明 GANet 所提出的各个模块的有效性。

3.1 数据集

本文使用 ScanNet v2 (Dai 等, 2017a)、ShapeNet (Chang 等, 2015) 和 Scan2CAD (Avetisyan 等, 2019) 3 个数据集开展实验测试。ScanNet v2 数据集由 1 513 个真实场景扫描数据组成, 通常用于三维语义、实例分割和目标检测等任务的性能评估。本文算法采用了与 RfD-Net 方法相同的点云预处理操作 (在训练和测试数据集中随机采样 80 K 个点), 并参考 RfD-Net 和 RevealNet 进行训练集/测试集划分。由于 Scan2CAD 已将 ShapeNet 中的物体 CAD 模型与 ScanNet v2 场景中的实例进行对齐, 因此本文算法

将 Scan2CAD 中的物体边界框和对齐的物体 CAD 模型作为目标检测和形状生成的监督信息。

3.2 评估指标

本文从三维目标检测和语义实例重建两个方面进行评估。本文参考 RfD-Net, 使用三维交并比 (intersection over union, IoU) 阈值为 0.5 的平均精度均值 (mAP) 来评估三维目标检测和语义实例重建的性能。其中, 语义实例重建指标衡量的是预测的物体网格在三维场景中与对应完整物体网格之间的重合度。为了进行公平对比, 本文采用与 RfD-Net 一样的设置, 对物体进行体素化处理。

3.3 实验细节

本文采用 PyTorch 深度学习框架, 使用两张 NVIDIA GeForce RTX3090 显卡来进行训练。网络的训练分为 3 个阶段: 目标检测预训练阶段、形状重建阶段和端到端联合训练。训练过程中的 batchsize 设置为 8。预训练阶段采用的初始学习率为 1×10^{-3} , 形状重建阶段采用的初始学习率为 1×10^{-4} , 端到端训练采用的初始学习率为 5×10^{-5} , 每 80 个周期减小到原来的 0.5 倍。每个阶段的 epoch 都设置为 240。

式 (3) 的检测损失中, λ_{con} 和 λ_l 分别设置为 0.5 和 0.1, 式 (4) 中的 λ_{seq} 设置为 100, 目标检测预训练阶段中式 (5) 的 λ 设置为 0, 形状重建阶段, λ 设置为 1, 联合训练阶段, λ 设置为 0.005。

3.4 实验结果分析

3.4.1 定量比较

1) 语义实例重建方面。本文将 GANet 方法与现有的方法 DIMR、RfD-Net 和 RevealNet 进行了比较。与 RevealNet 类似, 本文还比较了实例分割算法 3D-SIS (3D semantic instance segmentation) (Hou 等, 2019) 和形状完成算法 Shape Completion (Dai 等, 2017b) 的组合以及扫描完成算法 ScanComplete (Dai 等, 2018) 和 3D-SIS 的组合。

本文算法与其他算法在 ScanNet v2 和 Scan2CAD 数据集上的语义实例重建实验结果的对比如表 1 所示。其中, “*” 代表该方法提供的模型权重文件重新在同一平台上进行推理。由表 1 可得, 在重建指标 mAP@0.5 上, 本文算法获得了比其他方法更好的性能, 在大多数类别上的重建效果都达到了最佳。DIMR 采用实例分割—重建的框架。由于场景点云残缺严重, 导致其预测的实例位置和尺寸无法与真值精

确匹配,从而影响了最终性能。具体而言,GANet方法对浴缸、垃圾桶、沙发和橱柜的语义实例重建精度均值优于整体排名第2的RfD-Net方法,主要是因为GANet方法采用先验几何信息引导的盒采样方案,相

比固定尺度球体采样方案,获取到了物体更多的有效局部信息。在局部点的几何属性引导下,网络可提取出物体的关键特征,而不是全局的粗略特征,从而使重建结果的关键结构更符合物体本身。

表1 ScanNet v2和Scan2CAD数据集上的语义实例重建实验结果的对比

Table 1 Comparisons of semantic instance reconstruction results on the ScanNet v2 and Scan2CAD datasets

方法	display	bathtub	trashbin	sofa	chair	table	cabinet	bookshelf	mAP
3D-SIS(Hou等,2019)+Shape Completion(Dai等,2017b)	2.27	1.14	1.68	14.86	9.93	<u>3.90</u>	7.11	3.03	5.49
ScanComplete(Dai等,2018)+3D-SIS(Hou等,2019)	1.65	4.55	11.25	9.09	9.09	0.64	0.18	5.45	5.24
RevealNet(Hou等,2020)	13.16	13.64	18.19	24.79	15.87	11.28	8.60	10.60	14.52
RfD-Net*(Nie等,2021)	<u>23.13</u>	<u>15.88</u>	17.92	<u>41.61</u>	10.12	0.94	<u>26.35</u>	9.09	<u>18.13</u>
DIMR(Tang等,2022)	23.36	12.29	21.68	35.01	3.51	0.65	19.37	2.16	14.76
本文	21.43	21.84	<u>20.42</u>	44.65	<u>10.16</u>	1.13	27.90	<u>9.09</u>	19.58

注:加粗字体表示各列最优结果,下划线字体表示各列次优结果。

2)三维目标检测方面。本文算法与现有算法,如3D-SIS、MLCVNet(Xie等,2020)、RevealNet和RfD-Net进行了比较,实验结果如表2所示。其中,w/o和w/分别表示未用和已用生成网格结果来优化检测结果。

表2 ScanNet v2数据集上的三维目标检测结果对比

Table 2 Comparison of 3D object detection results on ScanNet v2

算法	输入	mAP/%
3D-SIS(Hou等,2019)	Geo+Image	25.70
MLCVNet(Xie等,2020)	Geo	33.40
RevealNet(Hou等,2020)	Geo	29.29
RfD-Net*(Nie等,2021)(w/o opt)	Geo	38.60
RfD-Net*(Nie等,2021)(w/ opt)	Geo	39.08
本文(w/o opt)	Geo	38.80
本文(w/ opt)	Geo	39.20

注:加粗字体表示各列最优结果。

从表2可以看出,3D-SIS和RevealNet的结果较差,主要是因为它们都对场景进行了网格化处理,丢失了物体的几何细节信息。考虑到该问题,MLCVNet方法省略网格化步骤,直接对场景点云进行处理,因此获得了较好的检测性能。由于MLCVNet方法仅对目标检测器进行训练而忽视了形状生成器,导致mAP仍不太理想,仅有33.40%。相比之下,RfD-Net方法

融合了目标检测器和形状生成器,并对两者进行联合训练,取得了更高的mAP,即39.08%。但RfD-Net方法在融合过程中引入了过多的噪声信息,且并未执行特征对齐。本文方法GANet通过引入有效的盒采样策略和特征对齐模块很好地解决了上述问题,因此取得了最好的三维目标检测结果。

3.4.2 定性比较

在ScanNet v2测试集上的定性结果如图4所示。在图4第1行可以看到:RfD-Net方法重建出来的桌子和椅子质量不佳,它们相互之间几乎完全重叠。主要是因为场景里的桌子和椅子本身就紧密地靠在一起,球体采样的方案会提取到大量属于其他物体的噪声点,影响重建的效果。本文GANet方法采用检测框盒采样的方案能获取到更多属于目标实例的前景点,减少噪声点的影响。此外,可以观察到DIMR方法所预测的桌子位置和尺寸与真实值有所偏差,主要是因为其采用的点云实例分割技术难以准确地预测残缺物体的尺寸和中心。特别地,若桌子并未与其他物体靠在一起(图4第2、3行)时,RfD-Net方法难以重建出精细的桌子网格模型,而DIMR方法仅能恢复桌面且错误地预测了朝向。相比之下,本文方法能捕获到物体的几何关键信息,更准确地判断候选点的占用情况,恢复出更符合原场景的网格模型。

3.4.3 消融实验

本文算法的核心设计主要有检测框盒采样、基

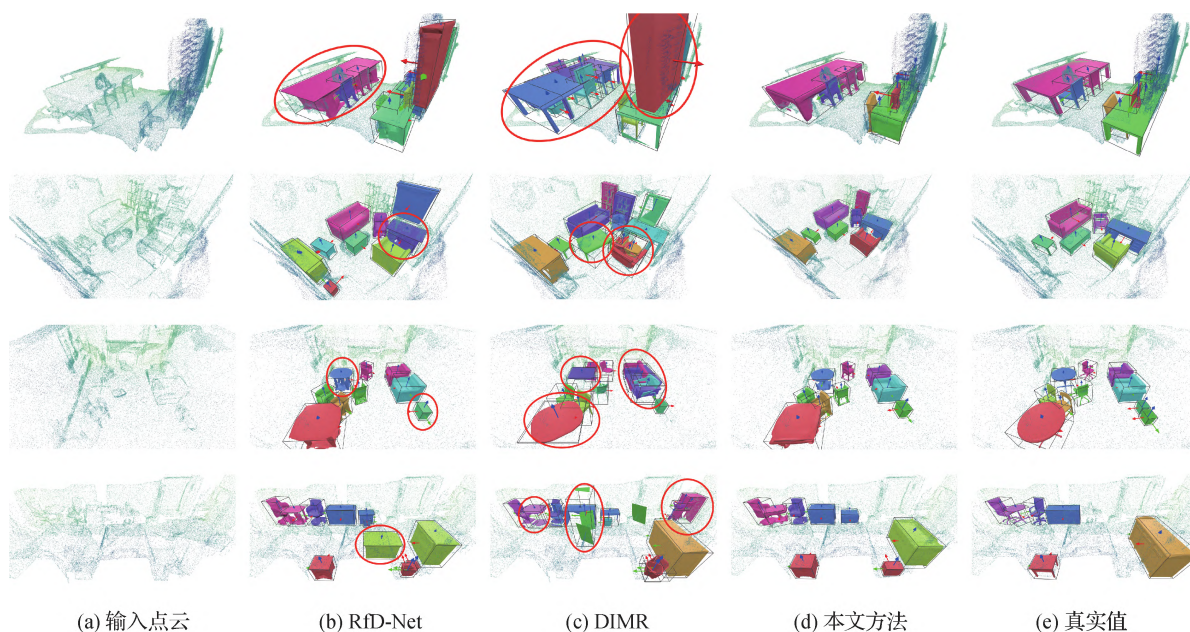


图4 本文算法在ScanNet v2数据集上部分场景进行语义实例重建的定性结果

Fig. 4 Qualitative results of semantic instance reconstruction for some scenarios on ScanNet v2
(a) input point clouds; (b) RfD-Net; (c) DIMR; (d) ours; (e) ground truth)

于Transformer的物体点云特征编码器、语义先验的嵌入和特征空间转换模块。为了验证各个设计的有效性,本文在真实场景数据集上进行了5组实验。包括完整的GANet模型([E]);将GANet中的检测框采样替换为球体采样([A]);将GANet中的点云特征编码器替换为PointNet网络([B]);点云编码器和去噪器只有几何信息输入,没有语义先验信息([C]);去掉特征转换模块([D])。

表3给出了消融实验的结果。其中,w/o代表缺失相应模块。从结果中可以观测到,检测框采样的方案将实例重建结果的精度从18.60%提升到19.58%。这是因为检测框采样获取到了更多有效的物体前视点,同时降低了其他物体噪声信息对于重建目标物体的负面影响。对比实验[B]和实验

表3 消融实验对比结果

Table 3 Comparison results of ablations studies

实验编号	网络模型	mAP/%	
		目标检测	实例重建
[A]	GANet w/o box sample	38.30	18.60
[B]	GANet w/o Transformer	38.62	17.73
[C]	GANet w/o category	38.52	18.88
[D]	GANet w/o transformation	38.29	19.15
[E]	GANet	39.20	19.58

注:加粗字体表示各列最优结果。

[E],如果本文算法将基于Transformer的点云特征编码器换成PointNet网络,结果将降低为17.73%。主要是因为原方案对所有局部点的细节信息只进行简单的最大池化,丢失了物体关键点信息。与完整模型相比,实验[C]的性能下降为18.88%,证明了物体语义先验信息一定程度上能提升物体重建的性能。实验[D]直接将检测器提取的特征与物体局部几何特征串联起来,而本文算法首先通过特征转换器对方案特征进行对齐,然后再进行串联。从表3中可以看出,特征转换器提高了物体重建的性能。

3.4.4 场景点云对比实验

本文进一步测试GANet算法在不同稀疏性场景点云中的效果。本文算法将场景输入点云的数目固定为20 K、40 K和80 K,分别进行训练和测试。本文算法在不同稀疏性场景上的结果对比和实例重建效果图分别如表4和图5所示。

表4 不同规模输入点的比较

Table 4 Comparison with different scales of input points

输入点云数量	mAP/%	
	目标检测	实例重建
20 K	35.99	15.08
40 K	39.24	17.36
80 K	39.20	19.58

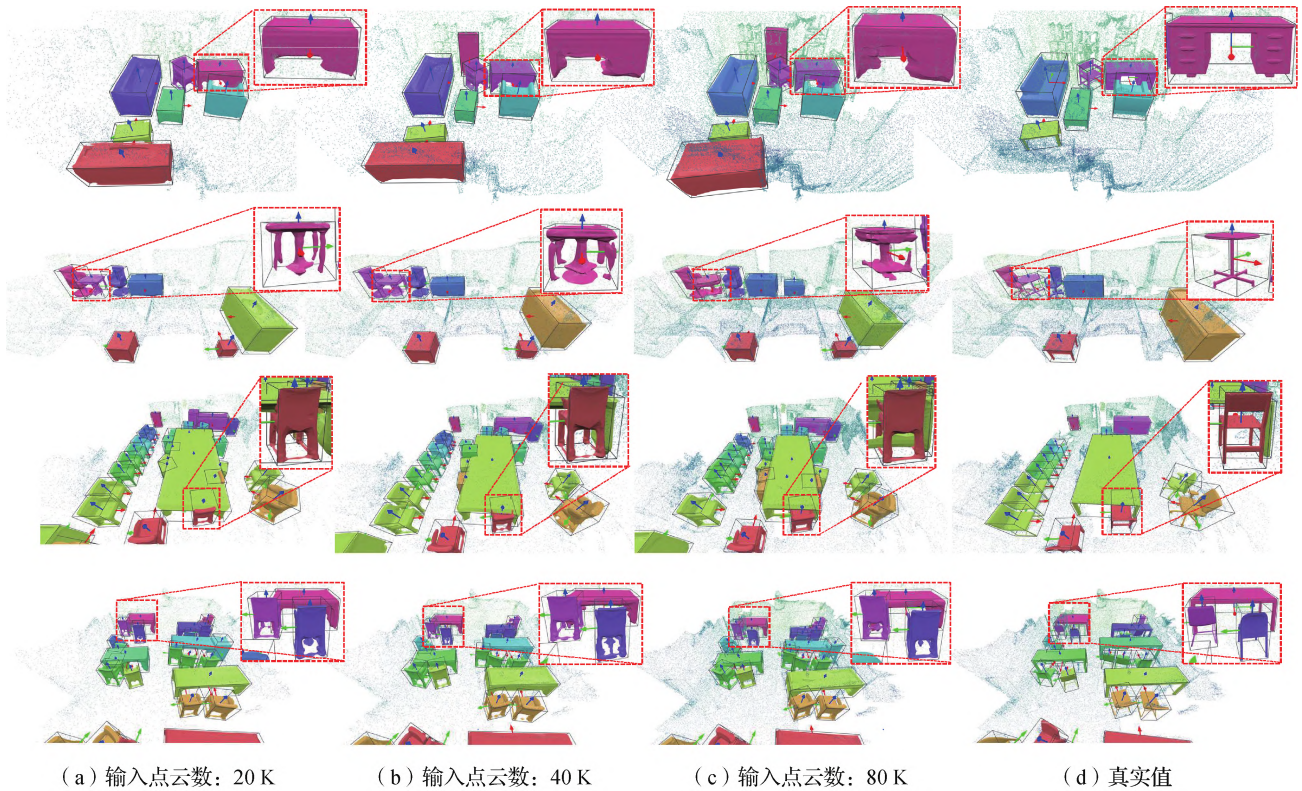


图5 不同点云稀疏度下的实例重建结果对比

Fig. 5 Comparison of the instance reconstruction results with different point sparsity
(a) 20 K input points; (b) 40 K input points; (c) 80 K input points; (d) ground truth

表4和图5中的结果表明,即使场景点云非常稀疏,仅有20 K个点,本文算法依然能得到令人满意的实例重建结果。

4 结 论

本文提出了一个新的语义实例重建算法GANet,用于点云的场景物体重建。GANet有效地利用物体的几何尺度信息,设计了检测框盒采样策略,获取到了更多有效的物体前景点。进一步地,为了有效利用物体的前景点信息,本文算法采用了基于Transformer的点云特征编码器,以物体的关键几何属性信息为指导,提取物体的关键局部细节特征。同时,物体的语义属性信息的嵌入进一步提升了重建的性能。与RfD-Net算法相比,本文算法在ScanNet v2数据集上取得了8%的实例重建性能提升。但值得注意的是,本文算法依然是基于先理解后重建的框架来完成语义实例重建任务,理解的结果会较大地影响后续的重建结果。因此,如何设计一种更为完善的算法及损失函数,使得理解与重建任务

之间可以互相促进,将是未来三维语义实例重建的主要研究方向之一。

参考文献(References)

- Avetisyan A, Dahnert M, Dai A, Savva M, Chang A X and Nießner M. 2019. Scan2CAD: learning CAD model alignment in RGB-D scans//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 2609-2618 [DOI: 10.1109/CVPR.2019.00272]
- Chang A X, Funkhouser T, Guibas L, Hanrahan P, Huang Q X, Li Z M, Savarese S, Savva M, Song S R, Su H, Xiao J X, Yi L and Yu F. 2015. ShapeNet: an information-rich 3D model repository [EB/OL]. [2023-03-20]. <https://arxiv.org/pdf/1512.03012.pdf>
- Dai A, Chang A X, Savva M, Halber M, Funkhouser T and Nießner M. 2017a. ScanNet: richly-annotated 3D reconstructions of indoor scenes//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA: IEEE: 5828-5839 [DOI: 10.1109/CVPR.2017.261]
- Dai A, Qi C R and Nießner M. 2017b. Shape completion using 3D-encoder-predictor CNNs and shape synthesis//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA: IEEE: 5868-5877 [DOI: 10.1109/CVPR.

- 2017.693]
- Dai A, Ritchie D, Bokeloh M, Reed S, Sturm J and Nießner M. 2018. ScanComplete: large-scale scene completion and semantic segmentation for 3D scans//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 4578-4587 [DOI: 10.1109/CVPR.2018.00481]
- Dai A, Siddiqui Y, Thies J, Valentin J and Nießner M. 2021. SPSSG: self-supervised photometric scene generation from RGB-D scans//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 1747-1756 [DOI: 10.1109/CVPR46437.2021.00179]
- Fu Z H, Wang L G, Xu L, Wang Z Y, Laga H, Guo Y L, Boussaid Farid and Bennamoun M. 2023. VAPCNet: viewpoint-aware 3D point cloud completion//Proceedings of 2023 IEEE/CVF International Conference on Computer Vision. Paris, France: IEEE: 12108-12118
- Gkioxari G, Johnson J and Malik J. 2019. Mesh R-CNN//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision. Seoul, Korea (South): IEEE: 9785-9795 [DOI: 10.1109/ICCV.2019.00988]
- Guo Y L, Wang H Y, Hu Q Y, Liu H, Liu L and Bennamoun M. 2020. Deep learning for 3D point clouds: a survey. IEEE Transactions on Pattern Analysis and Machine Intelligence, 43 (12): 4338-4364 [DOI: 10.1109/TPAMI.2020.3005434]
- Han X G, Zhang Z X, Du D, Yang M D, Yu J M, Pan P, Yang X, Liu L G, Xiong Z X and Cui S G. 2019. Deep reinforcement learning of volume-guided progressive view inpainting for 3D point scene completion from a single depth image//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 234-243 [DOI: 10.1109/CVPR.2019.00032]
- Hou J, Dai A and Nießner M. 2019. 3D-SIS: 3D semantic instance segmentation of RGB-D scans//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 4416-4425 [DOI: 10.1109/CVPR.2019.00455]
- Hou J, Dai A and Nießner M. 2020. RevealNet: seeing behind objects in RGB-D scans//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 2098-2107 [DOI: 10.1109/CVPR42600.2020.00217]
- Liu X P, Ma Y X, Xu K, Wan J W and Guo Y L. 2022. AGFA-Net: adaptive global feature augmentation network for point cloud completion. IEEE Geoscience and Remote Sensing Letters, 19: #7004505 [DOI: 10.1109/LGRS.2022.3198799]
- Long X X, Cheng X J, Zhu H, Zhang P J, Liu H M, Li J, Zheng L T, Hu Q Y, Liu H, Cao X, Yang R G, Wu Y H, Zhang G F, Liu Y B, Xu K, Guo Y L and Chen B Q. 2021. Recent progress in 3D vision. Journal of Image and Graphics, 26(6): 1389-1428 (龙霄潇, 程新景, 朱昊, 张朋举, 刘浩敏, 李俊, 郑林涛, 胡庆拥, 刘浩, 曹汛, 杨睿刚, 吴毅红, 章国锋, 刘焯斌, 徐凯, 郭裕兰, 陈宝权. 2021. 三维视觉前沿进展. 中国图象图形学报, 26(6): 1389-1428 [DOI: 10.11834/jig.210043]
- Lorensen W E and Cline H E. 1987. Marching cubes: a high resolution 3D surface construction algorithm. ACM SIGGRAPH Computer Graphics, 21(4): 163-169 [DOI: 10.1145/37402.37422]
- Mescheder L, Oechsle M, Niemeyer M, Nowozin S and Geiger A. 2019. Occupancy networks: learning 3D reconstruction in function space//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 4455-4465 [DOI: 10.1109/CVPR.2019.00459]
- Nie Y Y, Han X G, Guo S H, Zheng Y J, Chang J and Zhang J J. 2020. Total3DUnderstanding: joint layout, object pose and mesh reconstruction for indoor scenes from a single image//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 52-61 [DOI: 10.1109/CVPR42600.2020.00013]
- Nie Y Y, Hou J, Han X G and Nießner M. 2021. RiD-Net: point scene understanding by semantic instance reconstruction//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 4608-4618 [DOI: 10.1109/CVPR46437.2021.00458]
- Qi C R, Su H, Mo K C and Guibas L J. 2017. PointNet: deep learning on point sets for 3D classification and segmentation//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA: IEEE: 77-85 [DOI: 10.1109/CVPR.2017.16]
- Roldão L, De Charette R and Verroust-Blondet A. 2022. 3D semantic scene completion: a survey. International Journal of Computer Vision, 130 (8): 1978-2005 [DOI: 10.1007/s11263-021-01504-5]
- Tang J X, Chen X K, Wang J B and Zeng G. 2022. Point scene understanding via disentangled instance mesh reconstruction//Proceedings of the 17th European Conference on Computer Vision. Tel Aviv, Israel: Springer: 684-701 [DOI: 10.1007/978-3-031-19824-3_40]
- Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A N, Kaiser Ł and Polosukhin L. 2017. Attention is all you need//Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach, USA: Curran Associates Inc.: 6000-6010
- Xie Q, Lai Y K, Wu J, Wang Z T, Zhang Y M, Xu K and Wang J. 2020. MLCVNet: multi-level context VoteNet for 3D object detection//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 10447-10456 [DOI: 10.1109/CVPR42600.2020.01046]
- Yuan W T, Khot T, Held D, Mertz C and Hebert M. 2018. PCN: point completion network//Proceedings of 2018 International Conference on 3D Vision (3DV). Verona, Italy: IEEE: 728-737 [DOI: 10.1109/3DV.2018.00088]

Zhong M and Zeng G. 2020. Semantic point completion network for 3D semantic scene completion//Proceedings of the 24th European Conference on Artificial Intelligence. Santiago de Compostela, Spain: IOS Press: 2824-2831

作者简介

万骏辉,男,硕士研究生,主要研究方向为三维目标检测与场景实例重建。E-mail: wanjh8@mail3.sysu.edu.cn

张鹏,通信作者,男,副教授,主要研究方向为计算机视觉与

机器学习。E-mail: zhangpeng5@mail.sysu.edu.cn

刘心溥,男,博士研究生,主要研究方向为三维目标检测与点云形状补全。E-mail: liuxinpu@nudt.edu.cn

陈莉丽,女,助理研究员,主要研究方向为计算机视觉与机器学习。E-mail: chenlili8209@nudt.edu.cn

敖晟,男,博士研究生,主要研究方向为三维特征学习、匹配和点云配准。E-mail: aosh@mail2.sysu.edu.cn

郭裕兰,男,副教授,主要研究方向为三维视觉与机器学习。E-mail: yulan.guo@nudt.edu.cn