计算机科学
COMPUTER SCIENCE

IBSNet:用于估计单视角扫描点云交互平分面的神经隐式场

袁右文, 金朔, 赵玺

引用本文

袁右文, 金朔, 赵玺. IBSNet:用于估计单视角扫描点云交互平分面的神经隐式场[J]. 计算机科学, 2025, 52(8): 195-203.

YUAN Youwen, JIN Shuo, ZHAO Xi. IBSNet:A Neural Implicit Field for IBS Prediction in Single-view Scanned Point Cloud [J]. Computer Science, 2025, 52(8): 195-203.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.240900086

IBSNet:用于估计单视角扫描点云交互平分面的神经隐式场

袁右文　金　朔　赵　玺
西安交通大学计算机科学与技术学院　西安 710049
(2193412689@stu.xjtu.edu.cn)

摘　要　三维物体之间的空间关系分析对于多物体场景的理解及合成具有重要意义。传统的三维空间关系分析方法计算物体之间的交互平分面(Interaction Bisector Surface,IBS)并进一步提取其特征。然而,当输入为单视角扫描点云时,由于数据完整性的缺失,使用传统方法往往难以计算出准确的交互平分面,从而极大地影响了下游任务(如场景分类、分析、合成等)。针对此问题,提出一种面向单视角扫描点云的交互平分面估计方法,使用神经网络框架 IBSNet 估计双物体的差分无符号距离场,然后基于这种隐式距离场的表示提取交互平分面。在 ICON 数据集上对该方法与其他方法(几何方法、IMNet、Grasping Field)进行了对比实验,并测试了各个方法在面对不同残缺程度和噪声程度的单视角扫描点云时的鲁棒性。实验结果表明,该方法对于残缺的单视角扫描点云有一定的鲁棒性,可以有效地估计出形状之间的交互平分面。

关键词:空间关系分析;交互平分面;单视角扫描点云;神经隐式场;无符号距离场
中图分类号　TP391

IBSNet:A Neural Implicit Field for IBS Prediction in Single-view Scanned Point Cloud

YUAN Youwen,JIN Shuo and ZHAO Xi
College of Computer Science and Technology,Xi'an Jiaotong University,Xi'an 710049,China

Abstract　The analysis of spatial relationships between 3D objects is of great significance for scene understanding and interaction. For example,by analyzing the spatial relationship between the robot and the object,the robot can be guided to grasp the object more accurately. By learning the spatial relationship between objects in the real scene,it can guide the generation of virtual scenes that look more natural or better meet the needs of interaction. However,because the single-view scanned point clouds gotten by RGB-D cameras or LiDAR usually have many artifacts and noise,existing methods for analyzing the spatial relationships of objects are often difficult to make accurate predictions when faced with single-view scanned point clouds,which makes these methods impractical for practical applications. For handling the spatial relationship analysis of single-view scanned point clouds,this paper uses the interaction bisector surface(IBS) to express spatial relationships,and proposes a differential unsigned distance field of dual-object to implicitly represent IBS. Inspired by the implicit function learning methods widely used in recent years,this paper designs a neural implicit field to fit the differential unsigned distance field. This neural implicit field takes the single-view scanned point clouds of two objects as input and returns the different unsigned distance field of the two objects. This network uses two multi-layer self-attention point cloud encoders to extract the features of the two input point clouds and combines these features after that. Then these features are inputted into a dual-object unsigned distance decoder to get the unsigned distance value of the query points. Comparative experiments of this method with other methods(Geometry Method,IMNet and Grasping Field) are conducted on the ICON dataset. It simulates single-view scans of each scene from 26 different viewpoints to get the single-view scanned point clouds and split the whole dataset into training set and test set based on a single scene. The robustness of each method is also tested when facing single-view scanning point clouds with different degrees of incompleteness and noise. Experimental results show that the proposed neural implicit field is very robust to the input single-view scanned point clouds with different degrees of incompleteness,and can efficiently predict IBS with accurate shapes.

Keywords　Spatial relationship analysis,Interaction bisector surface,Single-view scanned point cloud,Neural implicit field,Unsigned distance field

到稿日期:2024-09-13　返修日期:2025-01-24
基金项目:国家重点研发计划(2022YFB3303202);国家自然科学基金(62072366,U23A20312)
This work was supported by the National Key Research and Development Program of China(2022YFB3303202) and National Natural Science Foundation of China(62072366,U23A20312).
通信作者:赵玺(xi.zhao@xjtu.edu.cn)

1 引言

在一个三维场景中，不同物体之间通常存在多种空间关系。这些相对底层的空间关系与场景的高层次语义有着潜在且重要的联系。因此，表示、分析、理解三维物体之间的空间关系成为了研究的重要方向，对于依赖场景分析、场景合成的应用（如 VR、AR、机器人等领域）具有重要意义。例如，通过分析机械臂和物体之间的空间关系，可以指导机械臂更加准确地抓握物体[1-2]；通过对真实场景中物体之间空间关系的学习，可以指导生成看起来更加自然或更加符合交互需求的虚拟场景。

在现有的相关研究中，研究者采用了不同的空间关系表示方式。其中，交互平分面（Interaction Bisector Surface，IBS）[3]作为一种直观、高效的空间关系表示，已经被广泛应用于场景分析和物体识别等任务中。交互平分面通过显式的方式表示空间关系，它被定义为空间中到两个物体距离相等的点的集合。交互平分面在几何上具有确切的定义，它是维诺图的子集，因此具有良好的几何特征。

传统的计算方法需要先计算诺维图，再从中抽取出交互平分面。这种方法需要基于网格或点云表示的物体的完整几何形状进行计算。这意味着传统算法对于模型的残缺和噪声非常敏感。如果希望得到高质量的交互平分面，则必须提供高质量、完整的网格数据。然而，实际应用中通常使用的深度相机、激光雷达等设备只能得到点云数据，且往往因为物体间或物体自身的遮挡而存在残缺和噪声，这导致传统的交互平分面的计算方法难以投入实际应用中。因此，从更常见的单视角扫描点云数据中估计物体之间的交互平分面十分重要。

随着机器学习和神经网络技术的发展，很多工作采用了隐式函数学习方法（Implicit Function Learning，IFL）来表示物体表面[4-6]或者进行表面重建[7-8]。这类方法通过训练一个神经隐式场来得到空间中连续的点到物体表面的距离。例如，DeepSDF[4]用神经网络拟合一个连续的有符号距离函数来得到空间中点到物体表面的有符号距离。NDF[9]采用神经网络拟合无符号距离场，从而表示出非封闭的表面和开放流形等更复杂的拓扑结构，进一步拓展了隐式函数学习的适用范围。与传统的网格或体素的方法相比，神经隐式场可以更加容易地表示拓扑结构复杂的表面。同时，由于其可以表示连续的表面，因此 IFL 方法可以输出任意精细程度的表面。

受隐式函数学习方法的启发，本文提出了一个能够估计单视角扫描点云交互平分面的网络，称为 IBSNet。通过在训练过程中充分挖掘输入点云的局部和全局特征，以及物体点云之间的语义关系，该网络在面对存在残缺和噪声的输入点云时，估计出的交互平分面更加接近于完整形状计算的结果。具体地，本文提出差分无符号距离场的概念。通过使用差分无符号距离场表示交互平分面，本文方法可以很方便地将已有的点云特征提取方法应用到交互平分面的估计中。

本文的主要贡献总结如下：

1）提出了一种利用差分无符号距离场表示双物体的交互平分面的方法。根据空间中点在差分无符号距离场中的值，可以精确地判断该点是否在交互平分面上，从而获取理想的交互特征表示。

2）提出了一种端到端的、基于神经隐式场的单视角扫描点云物体交互平分面估计网络（IBSNet）。该网络可用于查询空间中点在双物体的差分无符号距离场中的值。相较于先前的方法，本文方法可以充分挖掘单视角扫描点云的潜在信息，并且对输入的不同残缺程度的点云具有很好的鲁棒性。

3）在 ICON[10]使用的双物体空间关系的数据集上对本文方法与传统的交互平分面计算方法进行了对比，并测试了各个方法在面对不同残缺程度的单视角扫描点云时的鲁棒性。实验证明了本文方法对于残缺点云数据能够估计出形状更加准确的交互平分面。

2 相关工作

2.1 点云特征提取

点云的特征提取对于分析场景和物体之间的关系具有重要意义。3D ShapeNets[11]和 VoxNet[12]开创性地将以往应用在图像领域的 2DCNN 网络拓展为 3DCNN 并应用在体素上，为三维形状的特征表示和分类提供了一个基于深度学习的新范式。OctNet[13]为了解决以往工作中体素分辨率偏低的问题，采用了不平衡的八叉树对空间进行层次划分，成功利用输入数据的稀疏性提升了网络处理高分辨率体素的能力。此外，OctNet 还利用八叉树层次化的特点，在不同尺度上捕捉更加丰富的三维信息，进一步提升了网络的处理能力。同时，使用 OctNet 在 3D 对象分类、3D 物体方向估计和 3D 物体语义分割任务中进行了验证，均取得了极佳的效果。然而，基于体素的方法天然带有一定局限性，包括计算代价大、噪声敏感、体素化导致的信息损失等。

PointNet[14]是第一个直接处理点云数据的端到端深度学习模型，它无需进行数据格式转换，直接利用神经网络处理点云数据。它的设计思路充分考虑了点云数据具有置换不变性和旋转不变性的特点，可以处理不同姿态和排列的点云数据。PointNet 的优势在于它能够直接处理点云数据，无需对原始数据进行额外的预处理或转换。同时，它具有旋转不变性和置换不变性，可以处理不同姿态和排列的点云数据。此外，它还能够处理可变数量的点，适用于不同大小和密度的点云输入。然而，PointNet 也存在一定局限性。首先，它忽略了点云中的局部结构信息，无法处理具有细粒度特征的点云数据。其次，由于其采用了 MLP 网络，因此对于大规模的点云数据可能面临计算和内存开销较大的问题。基于自注意力机制的网络已在自然语言处理和图像分析领域取得了巨大成功，Point Transformer[15]和 Point Cloud Transformer[16]将注意力机制用于点云特征提取。Point Transformer 根据点云的特点改造了经典的自注意力机制，其只在点的邻域内采用注意力计算，计算复杂度较低。Point Cloud Transformer 还将局部信息引入编码过程中，进一步提升了网络性能。Transformer 的注意力机制在提取点云特征方面具有很好的潜力，其同样也被应用于点云的目标检测[17]和点云配准[18]中。

本文在神经隐式场的编码器中使用了自注意力机制，以提升编码器对稀疏点云的特征聚合能力。

2.2 空间关系表示

物体之间的空间关系描述了物体之间的相对位置关系。之前的工作中，曾经出现过各种空间关系表示方法，这些方法大体可以分为以下两类。

一类将多个交互实体的特征进行组合，以隐式地获取空间关系表示。例如，Karunratanakul 等[19]基于手与物体的空间关系建立了抓握场（Grasping Field），并利用这种抓握场实

现了估计给定物体抓握姿势和从 RGB 图像重建手和物体两个任务。Zhao 等[20]构建了一个基于空间关系的双分支条件点云补全网络，使用特征组合的方式获取双物体的空间关系特征，并利用空间关系引导点云补全。

相较于特征组合的方式隐含地表达空间关系，另一类方法尝试以更加直观、可解释性更强的方式表达空间关系。例如，Zhao 等[3]提出了交互平分面(IBS)。IBS 通过多个物体之间特定的几何结构编码了多个实体的形状信息和位置关系，是一种直观、高效的空间关系表示方法。She 等[2]使用 IBS 作为手和物体之间的空间关系，并从交互平分面中提取局部和全局特征来捕捉更丰富的交互信息。Huang 等[21]进一步拓展了 IBS 表示方式并提出了交互面(Interaction Interface，ITF)。Xuan 等[22]将 IBS 用于人与场景交互动作生成。在生成动作时，该方法计算人与物体的 IBS 作为约束，从而获得更好的生成结果。Hu 等[10]注意到一个物体的功能应该取决于它和其他物体的关系，而不是孤立的物体本身，并设计了一种称为交互上下文的用于编码物体上下文信息的描述子(Interaction Context，ICON)。ICON 使用 IBS 和 IR 两种空间关系表示方法分别计算交互特征，并通过对象检索实验验证了其有效性。Zhao 等[23]将空间关系引入场景合成工作中，并设计了空间覆盖特征(Space Coverage Feature，SCF)，可以在频域对物体周围的开放空间进行编码。结合 IBS 与 SCF，该方法使得现有的自动场景合成方法具有处理复杂结构物体的能力。Huang 等[24]提出了神经交互场(NIF)和神经交互模板(NIT)，可用于描述如何操作物体。其用神经网络表达的逐点 SCF 特征，来编码给定物体周围的开放空间。然而，使用特定几何结构来表达空间关系也存在一定局限。由于这类方法的基本思路是设计某些对物体几何形状、物体间空间关系敏感的几何结构，并通过该几何结构的形状、拓扑特征表示物体间的空间关系，因此其对输入数据的质量有很高的要求。当输入数据存在残缺、噪声等问题时，计算出的几何结构可能会具有错误的拓扑结构和形状。这类方法对数据质量的敏感性导致其对残缺、噪声数据的鲁棒性不足。

另一类工作将物体之间的空间关系抽象成场景图表示(3D Scene Graph)。该类方法基于 3D 目标检测方法，预测场景中每个物体的 3D 包围盒，然后将每个包围盒作为场景图的一个节点。连接一对相邻节点的边包含了两个节点的相对位置关系和两个物体之间的语义关系。Wald 等[25]首先使用点云分割网络提取每个物体的空间和语义特征，得到每个物体的 3D 包围盒；然后根据提取的特征预测一个全连接的特征图；最后使用图卷积网络(Graph Convolutional Network，GCN)得到场景图。考虑到场景图关于物体之间空间关系的表示往往比较粗糙，只包括上、下、前、后等大致方位，Liu 等[26]进一步细化方位描述。然而，场景图表示方法将物体抽象为包围盒，因此其很难表示物体之间局部的空间关系。

本文提出的基于神经隐式场的交互平分面重建方法可以在面对不同残缺程度的物体点云时，保证输出高质量的交互平分面，且局部细节保留得比较完整。

2.3 神经隐式场

隐式场是一种通过函数描述空间中的某种性质的方式，常用的隐式表示方式有符号距离场(Signed Distance Field，SDF)和无符号距离场(Unsigned Distance Field，UDF)。由于神经网络具有强大的函数拟合能力，采用神经网络表达隐式场成为研究热点。DeepSDF[4]是深度隐式重建的代表，其使用神经隐式场拟合 3D 物体的连续有符号距离函数(SDF)。该方法能够用 3D 空间点及其 SDF 值训练神经隐式场，实现了离散数据到连续 SDF 值的映射。但是 DeepSDF 用一个全局特征向量表示物体模型的几何特征，导致网络在局部细节上表现不佳。DeepLS[27]将整个模型划分为多个部分，每个局部均由一个特征向量表示。这种做法为每个局部分配了一个特征，从而提升了神经网络表现模型细节的能力。

深度隐式重建可以从有限的离散数据中学习表达物体模型的连续的神经隐式场，但其对于新的输入需要重新训练。深度隐式形状学习通过编码器将任意形状的输入点云编码为一个表示其形状信息的特征向量，然后再经过隐式场解码器将原始输入点云转换为隐式场。IMNet[28]是深度隐式形状学习的代表，它设计了一个隐式场解码器，可以从输入点云的特征向量中解码查询点的 SDF 值。通过该过程，对于任意一个输入点云，IMNet 都可以重建对应的连续神经隐式场。

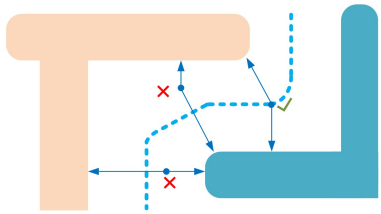
本文在利用神经隐式场表示交互平分面时，采用了深度隐式形状学习的方法，以便网络能够输出任意一对物体的交互平分面。

3 本文方法

本文提出了一种用于估计双物体的差分无符号距离场的神经隐式场，并通过在差分无符号距离场的零值面上随机采样足够的点来拟合交互平分面。本章首先介绍了差分无符号距离场的定义，然后介绍了本文提出的 IBSNet 网络结构、多层自注意力点云编码器、特征增强隐式场解码器，以及训练使用的损失函数，最后介绍了通过迭代策略在差分无符号距离场的零值面上进行采样并重建交互平分面点云的过程。

3.1 差分无符号距离场

无符号距离场 UDF 表示空间中点到物体表面的最小距离，而交互平分面上的点到两个物体表面的距离是相同的。根据交互平分面的这个特点，可以结合两个物体的无符号距离场表示交互平分面，如图 1 所示。



注：黄色部分为物体 A，蓝色部分为物体 B，中间的蓝色虚线为 A 与 B 之间的 IBS。

图 1 差分无符号距离场(电子版为彩图)

Fig. 1 Differential unsigned distance field

具体来说，将两个物体的无符号距离场分别记为 U_1 和 U_2，则一个差分无符号距离场 U_{diff} 可以表示为：

$$U_{\text{diff}}=|U_1-U_2| \tag{1}$$

其零值面代表了物体之间的交互平分面，故物体之间的交互平分面可以表示为所有在差分无符号距离场中值为 0 的点的集合：

$$IBS=\{p\in R^3\,|\,U_{\text{diff}}(p)=0\} \tag{2}$$

其中，p 表示空间中的一个点。

基于该思路，本文利用神经隐式场获得输入的一对点云

的无符号距离场，并从中提取差分无符号距离场的零值面。

3.2 网络结构

IBSNet 的网络结构如图 2 所示，其输入是两个物体的单视角扫描点云和一个查询点，输出是该查询点在差分无符号距离场中的值 。根据该值可以判断该点是否在交互平分面上，所有位于交互平分面上的点可以用来重建物体之间的交互平分面。对于一对输入的点云，如果将每个物体的点云单独输入神经隐式场中并获得对应的无符号距离场，同样可以计算出这对物体的差分无符号距离场的零值面。然而这种方式忽略了两个物体之间的语义关系和几何关系。本文提出的 IBSNet 将两个物体的点云同时作为神经隐式场的输入，通过将两个物体的特征进行组合来构建物体间的空间关系。

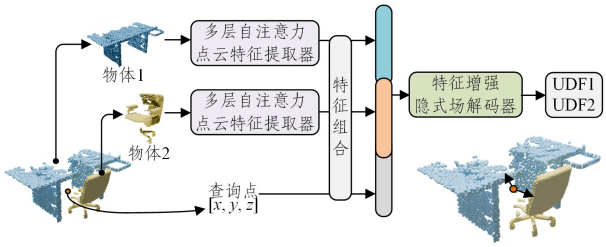


图 2 IBSNet 的网络结构

Fig. 2 Framework of IBSNet

IBSNet 采用了 Encoder-Decoder 结构，该结构主要由两个多层自注意力点云编码器和一个特征增强隐式场解码器组成。Encoder 提取两个输入单视角扫描点云的特征，随后采用特征组合的方式得到表示空间关系的特征向量，之后将查询点的坐标与空间关系特征向量进行拼接并输入解码器中，最后解码器得到查询点在差分无符号距离场中的值。如果得到的值小于阈值，则可以认为查询点在交互平分面上。

3.3 多层自注意力点云编码器

针对输入的单视角扫描点云存在点密度不均匀的问题，本文在点云编码器中使用了自注意力机制进行骨架点特征提取。Point Transformer[15] 在 PointNet 的基础上，在每一层骨架点进行邻域特征聚合时，使用了基于邻域自注意力的特征聚合方法。相较于 PointNet，这种基于邻域自注意力的方法能够调整邻域内每个点对骨架点特征的贡献权重。本文使用的多层自注意力点云编码器的结构如图 3 所示。首先，对于输入编码器的单视角扫描点云 $S\in R^{N\times 3}$，使用两个一维卷积得到逐点特征。然后进行 3 轮骨架特征提取与骨架自注意力运算，每轮骨架点个数为该轮输入点云的 1/4，特征长度变为原来的两倍。经过骨架特征提取后，此时骨架点有 $N/64$ 个，每个点的特征向量长度为 $8\times d$。最后进行最大池化，得到全局特征向量。

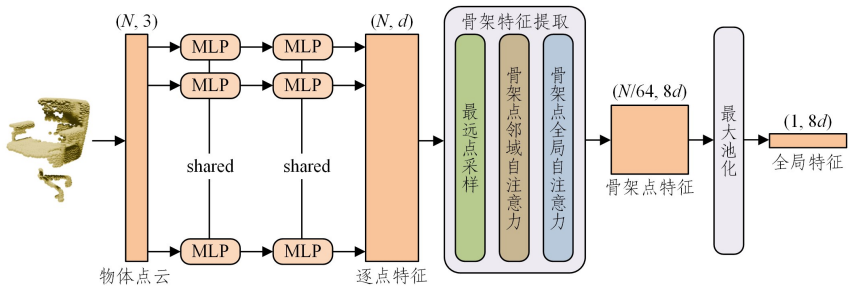


图 3 多层自注意力点云编码器

Fig. 3 Multi-layer self-attention point cloud encoder

3.4 特征增强隐式场解码器

受 DeepSDF[4] 和 IMNet[28] 的启发，IBSNet 的解码器同样使用了一个带跳层连接的 MLP，如图 4 所示。具体而言，解码器有 6 个线性层，除第一层外，每一层的输出拼接上输入解码器的全局特征向量作为下一层的输入。这样做的好处是，在解码过程中可以不断地强化输入物体的形状信息和查询点的位置信息，避免网络层数过深导致信息丢失。

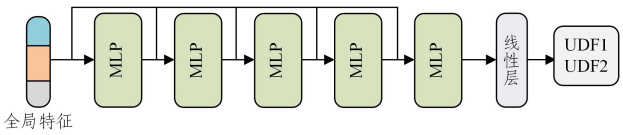


图 4 双物体无符号距离解码器

Fig. 4 Dual object unsigned distance decoder

3.5 损失函数

在网络的训练阶段，对于随机采样的查询点，将网络估计的 U_{diff} 值与真实的 $\hat{U}_{\text{diff}}$ 值进行比较优化。本文使用 L2 范数损失函数进行监督：

$$Loss=\frac{1}{2k}\sum_{i=1}^{k}(U_{\text{diff}}(x_i)-\hat{U}_{\text{diff}}(x_i))^2 \tag{3}$$

其中，k 为采样点的数量，x_i 为空间中的采样点的坐标。

3.6 迭代生长法重建交互平分面

构造交互平分面的方法分为两步：首先筛选出足够多的位于交互平分面上的采样点，然后根据需要采用重建算法将采样点转换为网格表示。其中，最核心的一步是筛选交互平分面上的采样点。

本文采用了一种渐进式的方法来进行点的筛选，该方法的核心思想是在交互平分面附近空间进行搜索，而不是在整个空间进行穷举，以达到加速的目的。具体地，首先在空间中随机选取一些初始点，然后从初始点中筛选靠近交互平分面的点，再继续从该点出发去搜寻其他交互平分面上的点。具体算法伪代码如算法 1 所示。

算法 1 交互平分面重建算法

输入：两个物体的点云 S_1 和 S_2

输出：交互平分面点云 P

1. 设定目标点数 n、初始阈值 t 和最小阈值，并在空间中随机选取一些点得到初始的 P
2. while $t<t_{min}$，或 P 中的点数量小于 n do
3. 对 P 中的每个点 p 使用 IBSNet 得到 $U_{diff}(p)$. 如果 $U_{diff}(p)\geqslant t$，则将 p 移出 P
4. 在 P 中每个点 p 半径为 t 的范围内随机取 10 个点，并加入 P 中
5. 令 t 为原来的一半

6. end while

7. 使用最远点采样对集合 P 下采样，得到有 n 个点的交互平分面点云

4 实验

4.1 实验数据

本实验使用的数据集是 ICON 中给出的双物体空间关系数据集，该数据集包含 9 个交互类别，分别为：桌与椅、包与挂钩、花与花瓶、衣撑与衣架、衣撑与衣服、篮子与物体、平板车与物体、书架与物体、帽子与衣架。每个类别包含数十条三维网格表示的数据。对完整网格数据，本文从半径为 1 的球面上均匀地取 26 个视角，对每个视角进行了模拟单视角扫描。具体扫描的方法与点云补全算法（如 PCN[29]）思路一致，即首先模拟拍摄深度图，然后进行反投影。为了对模拟扫描算法进行加速，本文在每次采样光线时仅在“种子”射线附近采样，有效减小了搜索空间。具体来说，首先将每对物体的 Mesh 归一化到以原点为球心的单位球内。然后，对于每个视角，在视线内均匀采样 r 个点，并在每个点的位置附近随机发射 m 条光线。随后，计算射线与两个 Mesh 的相交关系，若其与某个 Mesh 相交，则第一个交点处形成对应点云中的一个点，并记该条光线为“种子”光线。在每轮迭代中，不断在种子光线邻域内增加采样的射线密度，直到采样得到的点数量达到预定的标准。最后，对得到的点云使用最远点采样，得到固定点数为 n 的单视角扫描点云。迭代光线投射法获得的数据如图 5 所示。

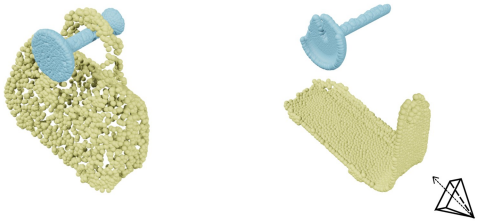


图 5 完整点云与本文方法采集的单视角扫描点云

Fig. 5 Complete point cloud and the single-view scanned point cloud collected by the proposed method

为了获取训练网络需要的查询点及对应的真实的值，对于每个样本，在空间中随机选取 50 000 个点，分别计算每个点到两个物体的 Mesh 表面的最近距离，得到 $\hat{U}_1$ 和 $\hat{U}_2$，进而得到：

$$\hat{U}_{\text{diff}} = |\hat{U}_1 - \hat{U}_2| \tag{4}$$

本文将数据随机划分为训练集和测试集，其中训练集占 70%，测试集占 30%。数据划分的粒度为场景，即同一场景所得到的扫描点云仅出现在训练集或测试集中。

4.2 实现细节

本文实验在 Ubuntu 22.04.1 LTS 操作系统上进行，使用的显卡型号为 NVIDIA GeForce RTX 3090。输入单视角扫描点云的点数为 2048，经过两个共享 MLP 层后得到两个物体的逐点特征，维度均为 2048×32。每轮提取上一轮 1/4 的点作为骨架点，并通过 MLP 将特征维度扩张为原来的 2 倍，然后进行 3 轮自注意力运算，此时剩余骨架点点数为 32，特征维度为 256。将该骨架点特征执行平均池化，并通过两个维度为（256，256）的 MLP 层后得到全局特征。将两个物体各自的 256 维特征及查询点坐标进行拼接，形成 515 维的解码器输入。解码器为多层 MLP，第一层维度为（515，512），中间四层维度为（515＋512，512），最后的线性层维度为（512，2）。本文在训练集上共训练 50 代，耗时 5.5 h。在使用迭代生长法重建交互平分面时，本文取初始阈值 T 为 0.005，取最小阈值 $T_{\min}$ 使得 $\log_2 \frac{T}{T_{\min}} = 100$，目标点数为 15 000。

4.3 点云与 Mesh 形状相似度评价指标

本文得到的交互平分面是点云形式的，为了评估不同方法重建交互平分面点云 S 的精确度，取交互平分面真值 Mesh 的顶点集合 M，并计算单向倒角距离评估 S 与 M 的相似度：

$$CD_t(S \rightarrow M) = \frac{1}{\|S\|} \sum_{x \in S} \min_{t \in M} \|x - t\|_2^2 \tag{5}$$

其中，x 和 t 分别是 S 和 M 中的点。从式（5）可以看出，单向倒角距离越小，结果与真值的相似度越大，结果越精确。

4.4 结果与评估

4.4.1 交互平分面结果

图 6 展示了使用本文方法得到的交互平分面。

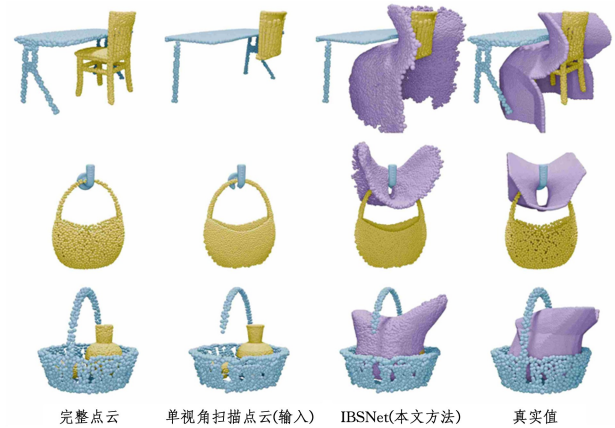


图 6 交互平分面结果

Fig. 6 Result of reconstructed IBS

从图中可以看出，对于残缺程度不大的物体，使用 IBSNet 可以得到形状十分接近真值的交互平分面。对于残缺程度较大的物体，该方法可以有效地结合两输入物体的信息，并对各自的物体类别做出合理估计。此外，从图中第二个场景可以看出，通过使用自注意力机制，IBSNet 对于具有较大数据尺度差异的场景也可以得到合理的结果。不论两物体尺度较为接近或差异较大，本文方法均可以估计出形状合理的交互平分面。

4.4.2 与其他方法的重建效果对比

为了评估本文算法的有效性，将 IBSNet 与 Grasping Field[19] 和 IMNet[28] 进行比较。

Grasping Field 是一个手部姿态生成网络，可以给出抓握给定物体的最佳手部姿态。本文利用其估计物体之间无符号距离场的能力来重建交互平分面。IMNet 是一个隐式形状学习网络，它无需重复训练，即可得到任意输入点云的神经隐式场。本文采用两个完全相同的 IMNet 网络实现双物体的形状估计，并计算出交互平分面。

本文在可视化结果对比中还加入了使用传统方法计算交

互平分面的“几何方法[3]”，如图 7 所示。同时，展示了各方法对于同一场景在不同视角下获得的单视角扫描点云的结果，如图 8 所示。从图 7 可以看出，虽然传统方法在输入点云残缺程度较小时（如第一个场景）求出的交互平分面具有几乎完美的拓扑，但是当输入点云残缺程度较大时，传统方法求出的交互平分面的拓扑结构受到了明显的影响，甚至出现了错误的拓扑结构，而本文方法仍然能够通过估计物体类别和空间关系等信息得到拓扑正确的交互平分面。此外，本文方法求出的交互平分面相比其他深度学习方法明显更加平滑，且能够更大程度地保留细节特征。从图 8 也可以看出，IMNet 和 Grasping Field 方法重建的交互平分面存在更多穿插现象，而本文方法对于同一场景不同视角下的单视角扫描点云能够稳定地重建出质量更高的交互平分面。表 1 列出了 IBSNet，Grasping Field（GF）和 IMNet 重建交互平分面的逐类别单向倒角距离。可以看出，IBSNet 在 7 个类别上优于对比方法，且均值有所提升。其中，IBSNet 得到的交互平分面精度明显优于 IMNet，这是因为 IMNet 方法将两个点云独立输入网络中，而未考虑二者之间的语义关系，证明了本文采用特征组合方式的有效性。本文还展示了各方法重建每个场景的交互平分面的时间开销，如表 2 所列。

表 1 本文方法与其他方法的逐类别单向倒角距离（↓）

Table 1 Comparison of the proposed method and other methods in class-by-class one-way chamfering distance（↓）

方法	类别 1	类别 2	类别 3	类别 4	类别 5	类别 6	类别 7	类别 8	类别 9	平均
GF	23.98×10^{-3}	$\mathbf{16.24\times10^{-3}}$	31.96×10^{-3}	16.24×10^{-3}	$\mathbf{19.59\times10^{-3}}$	29.06×10^{-3}	17.98×10^{-3}	27.89×10^{-3}	9.17×10^{-3}	21.05×10^{-3}
IMNet	22.92×10^{-3}	18.13×10^{-3}	30.24×10^{-3}	14.59×10^{-3}	24.33×10^{-3}	26.35×10^{-3}	21.29×10^{-3}	28.92×10^{-3}	11.30×10^{-3}	21.46×10^{-3}
IBSNet（Ours）	$\mathbf{19.98\times10^{-3}}$	17.43×10^{-3}	$\mathbf{28.04\times10^{-3}}$	$\mathbf{14.57\times10^{-3}}$	21.14×10^{-3}	$\mathbf{28.99\times10^{-3}}$	$\mathbf{17.29\times10^{-3}}$	$\mathbf{27.32\times10^{-3}}$	$\mathbf{8.52\times10^{-3}}$	$\mathbf{20.27\times10^{-3}}$

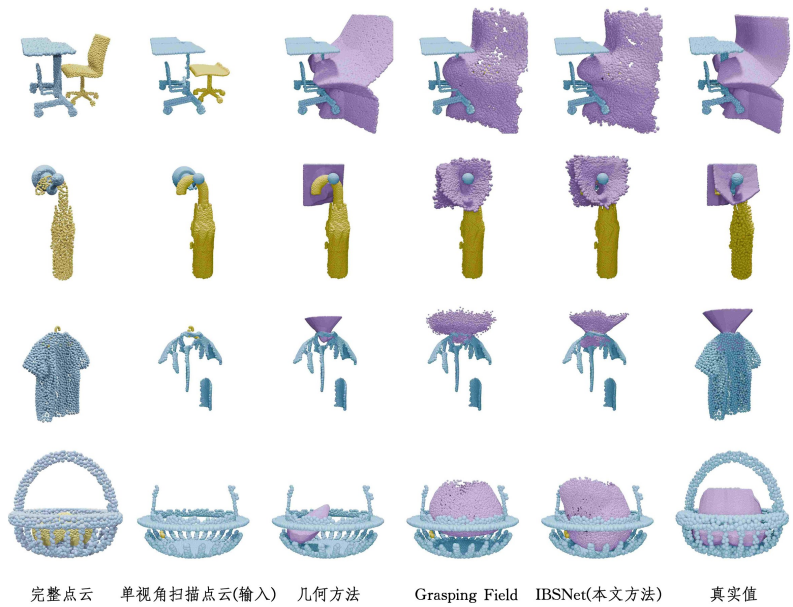
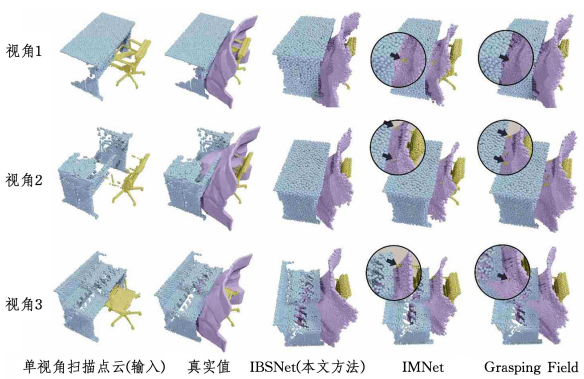


图 7 不同方法的交互平分面重建结果对比

Fig. 7 Comparison of IBS reconstruction results using different methods



注：箭头指向的部分存在穿插。

图 8 不同方法对于同一场景不同视角的交互平分面重建结果对比

Fig. 8 Comparison of IBS reconstruction results of different methods for the same scene from different perspectives

表 2 各方法重建交互平分面耗时

Table 2 Time consuming of each method to reconstruct IBS

方法	t/scene/s
IBSNet（Ours）	2.3
IMNet	1.4
GF	2.5
几何方法	**0.3**

4.4.3 不同方法对输入点云残缺的鲁棒性

为了验证不同方法对于残缺的输入点云的鲁棒性，本小节统计了不同残缺程度下各种方法的重建精度，并绘制了输入点云残缺程度-交互平分面精度折线图。本文通过计算完整模型点云到单视角扫描点云的单向倒角距离来衡量点云的残缺程度，记为 $CD_{c\to p}$。随着点云的残缺程度增大，完整模型与其形状相似度下降，$CD_{c\to p}$ 的值也随之增大。从图 9 可以看出，通过完整模型点云到残缺输入点云的单向倒角距离来衡量点云的残缺程度是合理的。

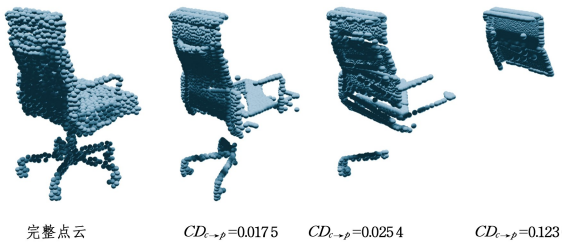


图 9 同一场景不同视角下得到的点云及对应的 $CD_{c\to p}$ 值

Fig. 9 Point clouds scanned under different views of the same scene and the corresponding $CD_{c\to p}$

本文将 $CD_{c\to p}$ 值划分为多个区间，表示多个残缺等级，等级越高表示点云越残缺。然后，对比不同方法在各残缺等级下的平均单向倒角距离，如图 10 所示。从图中可以看出，几何方法虽然在点云残缺程度小时重建交互平分面的效果最优，但是受点云残缺程度影响很大。Grasping Field 和 IBSNet 在残缺程度较大时均优于几何方法。得益于多层自注意力点云编码器的运用，IBSNet 在不同残缺程度下有着更稳定的效果。

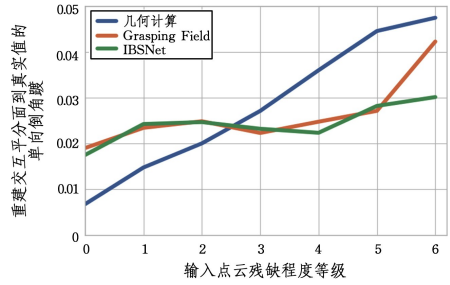


图 10 不同方法在各残缺等级下的平均单向倒角

Fig. 10 Average one-way chamfer of different methods at different defect levels

图 11 展示了不同方法对不同残缺程度的单视角扫描点云的重建效果。可以看出，本文方法在面对不同残缺程度的点云时都可以保持相对不错的重建质量，局部的细节特征得到了很好保留，同时不存在穿插现象，而其他两个方法重建的交互平分面均存在穿插现象，且丢失了细节。

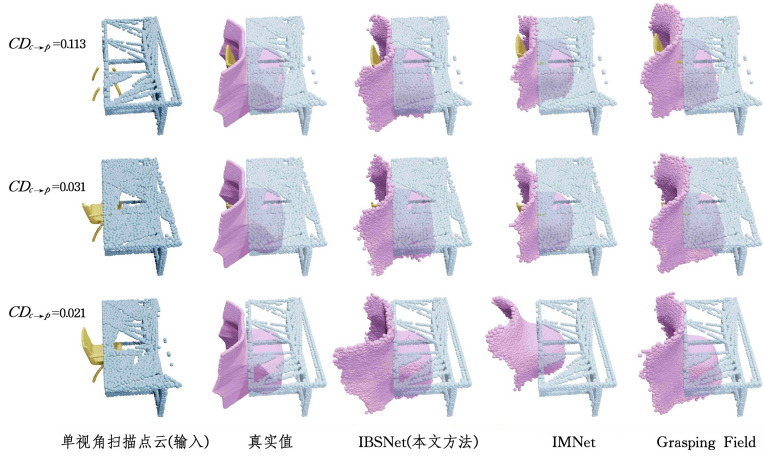


图 11 不同残缺程度下的重建交互平分面效果对比

Fig. 11 Comparison of IBS reconstruction results for the single view scanned point cloud with different degrees of incompleteness

4.4.4 不同方法对输入点云噪声的鲁棒性

考虑到实际应用中采用不同设备扫描得到的单视角扫描点云存在不同程度的噪声，本文进一步对比了各方法对于不同噪声程度的单视角扫描点云的鲁棒性。在单视角扫描点云的基础上叠加高斯噪声得到带噪声的单视角扫描点云：

$$p_{\text{noised}}=p+Gaussian(\sigma,s) \tag{6}$$

其中，p 表示空间中的一个点，p_{noised} 表示叠加噪声后得到的点，σ 和 s 分别是高斯分布的均值和方差。为简单起见，$\sigma=0$。本文取 s 为 0，0.01，0.02，0.05，0.07，0.1 和 0.2，得到 7 种噪声程度的点云。与 4.4.3 节相同，这里也采用完整点云到带噪声点云的单向倒角距离来衡量噪声程度。各方法在不同噪声程度下的平均单向倒角距离如图 12 所示。

从图 12 中可以看出，相比于传统的几何计算，基于学习的 3 种方法对于噪声具有明显更好的鲁棒性。在实验中，当噪声程度较大时，几何计算的方法有很大概率无法重建出交互平分面。相比 IMNet 和 Grasping Field，IBSNet 在单视角扫描点云的噪声程度较大时重建出来的交互平分面质量最好。

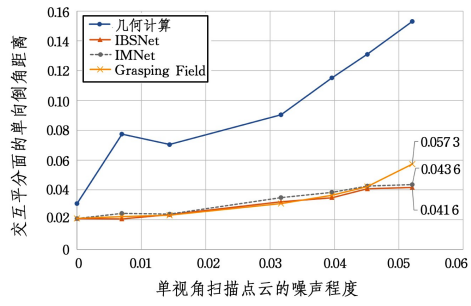


图 12 不同方法在不同噪声程度下的平均单向倒角距离

Fig. 12 Average one-way chamfer distance of different methods under different noise levels

4.4.5 消融实验

本文分别将 IBSNet 的编码器和解码器替换为 Grasping Field 方法的网络结构：1）GF，编码器和解码器都使用 Grasping Field 网络；2）GF＋IBSNet-D，使用 Grasping Field 替换 IBSNet 的编码器；3）IBSNet-E＋GF，使用 Grasping Field 替换 IBSNet 的解码器；4）IBSNet，完整的 IBSNet 网络。本文仍然使用重建交互平分面到原始 Mesh 的单向倒角距离评估作

为重建质量的评价指标，表 3 列出了不同方法的效果。从表中可以看出，在 Grasping Field 网络结构的基础上，将编码器或解码器替换为本章设计的编码器或解码器均对交互平分面重建结果有积极影响，这说明提出的基于自注意力机制的点云特征编码器和特征增强隐式场解码器均是有效的，且性能超过了基线方法。

表 3 消融实验中各种方法的逐类别单向倒角距离(↓)

Table 3 Class-by-class one-way chamfer distance of various methods in ablation experiments(↓)

方法	类别 1	类别 2	类别 3	类别 4	类别 5	类别 6	类别 7	类别 8	类别 9	平均
GF	23.98×10^{-3}	$\mathbf{16.24}\times10^{-3}$	31.96×10^{-3}	16.24×10^{-3}	19.59×10^{-3}	29.06×10^{-3}	17.98×10^{-3}	27.89×10^{-3}	9.17×10^{-3}	21.05×10^{-3}
GF+IBSNet-D	21.44×10^{-3}	17.22×10^{-3}	30.03×10^{-3}	17.02×10^{-3}	$\mathbf{18.92}\times10^{-3}$	29.08×10^{-3}	17.44×10^{-3}	28.99×10^{-3}	9.02×10^{-3}	20.98×10^{-3}
IBSNet-E+GF	20.26×10^{-3}	17.67×10^{-3}	$\mathbf{27.43}\times10^{-3}$	15.58×10^{-3}	19.22×10^{-3}	$\mathbf{27.32}\times10^{-3}$	17.52×10^{-3}	28.55×10^{-3}	$\mathbf{8.15}\times10^{-3}$	20.46×10^{-3}
IBSNet	$\mathbf{19.98}\times10^{-3}$	17.43×10^{-3}	28.04×10^{-3}	$\mathbf{14.57}\times10^{-3}$	21.14×10^{-3}	28.99×10^{-3}	$\mathbf{17.29}\times10^{-3}$	$\mathbf{27.32}\times10^{-3}$	8.52×10^{-3}	$\mathbf{20.27}\times10^{-3}$

4.4.6 交互平分面重建算法的效率分析

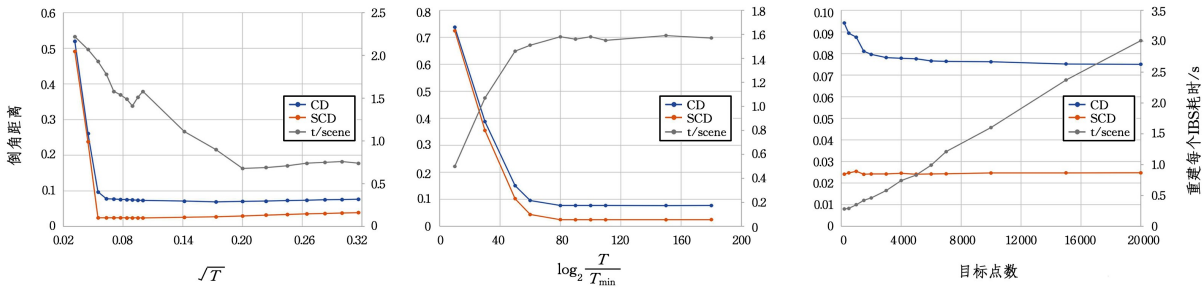
本文进一步探究了不同超参数的设置对于迭代生长交互平分面重建方法的效果和时间开销的影响。为了更加细致地对比不同情况下重建出的交互平分面的质量，本文还采用了完整的倒角距离作为衡量指标：

$$CD(S,V)=\frac{1}{\|S\|}\sum_{x\in S}\min_{t\in V}\|x-t\|_2^2+\frac{1}{\|V\|}\sum_{x\in V}\min_{t\in S}\|x-t\|_2^2 \tag{6}$$

其中，S 表示交互平分面点云，V 表示交互平分面真值 Mesh 的顶点集合，x 和 t 分别是 S 和 V 中的点。

各超参数对算法的影响如图 13 所示，其中 CD，SCD 和 t/scene 分别表示完整的倒角距离、单向倒角距离，以及重建每个场景的交互平分面消耗时间。结合图 13，可以取 $T=0.07$，$\log_2\frac{T}{T_{\min}}=80$，目标点数为 8000。

此时，本文方法重建的交互平分面平均单向倒角距离为 20.6×10^{-3}，重建每个交互平分面平均耗时 0.84 s。对比表 2 可知，该组超参数设置可以在保持重建的交互平分面质量无明显下降的情况下极大减少时间开销。从图 13 也可以看出，本文提出的迭代生长法具有很好的拓展性。根据重建质量和重建速度的优先级，可以选择合适的一组超参数以适应实际需要。



(a)初始阈值 T (b)最小阈值 $T_{\min}$ (c)目标点数

图 13 不同超参数对迭代生长法重建交互平分面质量和时间开销的影响

Fig. 13 Effects of different hyperparameters on the quality and time cost of reconstructing IBS using iterative growing method

结束语 为了解决单视角扫描点云的交互平分面估计问题，本文提出了一种基于神经隐式场估计交互平分面的方法。具体地，设计了一个神经隐式场(IBSNet)，用于查询空间中的点在双物体的差分无符号距离场中的值。通过与传统方法的对比，证明了本文提出的 IBSNet 可以估计出形状和细节准确、表面平滑的交互平分面。各方法对不同残缺程度的点云估计交互平分面的对比实验，证明了本文方法对单视角扫描点云具有很好的鲁棒性。最后进行了消融实验，证明了本文提出的多层自注意力点云编码器和特征增强隐式场解码器的有效性。

参 考 文 献

[1] HUANG Z Y, XU J Z, DAI S S, et. al. NIFT: Neural interaction field and template for object manipulation[C]// 2023 IEEE International Conference on Robotics and Automation (ICRA). IEEE, 2023: 1875-1881.

[2] SHE Q J, HU R Z, XU J Z, et. al. Learning high-DOF reaching-and-grasping via dynamic representation of gripper-object interaction[J]. ACM Transactions on Graphics, 2022, 41(4): 1-14.

[3] ZHAO X, WANG H, KOMURA T, et. al. Indexing 3D Scenes Using the Interaction Bisector Surface[J]. ACM Transactions on Graphics, 2014, 33(3): 1-14.

[4] PARK J J, FLORENCE P, STRAUB J, et al. DeepSDF: Learning continuous signed distance functions for shape representation[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019: 165-174.

[5] CHEN Z Q, ZHANG H. Learning implicit fields for generative shape modeling[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019: 5939-5948.

[6] SITZMANN V, ZOLLHÖFER M, WETZSTEIN G, et. al. Scene representation networks: Continuous 3d-structure-aware neural scene representations[J]. arXiv: 1906.01618, 2019.

[7] WANG P, LIU L J, LIU Y, et. al. NeuS: Learning Neural Im-

plicit Surfaces by Volume Rendering for Multi-view Reconstruction[C]// NIPS 2021. 2021：27171-27183.

[8] FU Q C，XU Q S，ONG Y S，et al. Geo-Neus：Geometry-consistent neural implicit surfaces learning for multi-view reconstruction[J]. Advances in Neural Information Processing Systems，2022，35：3403-3416.

[9] CHIBANE J，MIR A，PONS-MOLL G. Neural unsigned distance fields for implicit function learning[C]// Proceedings of the 34th International Conference on Neural Information Processing Systems. 2020：21638-21652.

[10] HU R，ZHU C，VAN KAICK O，et al. Interaction context（ICON）：towards a geometric functionality descriptor[J]. ACM Transactions on Graphics，2015，34(4)：1-12.

[11] WU Z，SONG S，KHOSLA A，et al. 3D ShapeNets：A Deep Representation for Volumetric Shapes[C]// 2015 IEEE Conference on Computer Vision and Pattern Recognition（CVPR）. 2015：1912-1920.

[12] MATURANA D，SCHERER S. VoxNet：A 3D Convolutional Neural Network for real-time object recognition[C]// 2015 IEEE/RSJ International Conference on Intelligent Robots and Systems（IROS）. 2015：922-928.

[13] RIEGLER G，ULUSOY A O，GEIGER A. OctNet：Learning Deep 3D Representations at High Resolutions[C]// 2017 IEEE Conference on Computer Vision and Pattern Recognition（CVPR）. 2017：3577-3586.

[14] CHARLES R Q，SU H，KAICHUN M，et al. PointNet：Deep Learning on Point Sets for 3D Classificationand Segmentation[C]// 2017 IEEE Conference on Computer Vision and Pattern Recognition（CVPR）. 2017：652-660.

[15] ZHAO H，JIANG L，JIA J，et al. Point Transformer.[C]// 2021 IEEE/CVF International Conferenceon Computer Vision（ICCV）. 2021：16259-16268.

[16] GUO M H，CAI J X，LIU Z N，et al. PCT：Point cloud transformer[J]. Computational Visual Media，2021(7)：187-199.

[17] LIU M Y，YANG Q M，HU G H，et al. 3D point cloud object detection algorithm based on Transformer[J]. Journal of Northwestern Polytechnical University，2023，41(6)：1190-1197.

[18] LIU X H，BAI Z Y，XU Z，et al. Multi-guided Point Cloud Registration Network Combined with Attention Mechanism[J]. Computer Science，2024，51(2)：142-150.

[19] KARUNRATANAKUL K，YANG J，ZHANG Y，et al. Grasping Field：Learning Implicit Representations forHuman Grasps[C]// 2020 International Conference on 3D Vision(3DV). 2020：333-344.

[20] ZHAO X，ZHANG B，WU J，et al. Relationship-Based Point Cloud Completion[J]. IEEE Transactions on Visualization and Computer Graphics，2022，28(12)：4940-4950.

[21] HUANG Z Y，DAI S S，XU K，et. al. DINA：Deformable INteraction Analogy[J]. Graphical Models，2024，133：101217.

[22] XUAN H B，LI X Z，ZHANG J S，et al. Narrator：Towards Natural Control of Human-Scene Interaction Generation via Relationship Reasoning[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision. 2023：22268-22278.

[23] ZHAO X，HU R，GUERRERO P，et al. Relationship templates for creating scene variations[J]. ACM Transactions on Graphics，2016，35(6)：1-13.

[24] HUANG Z，XU J，DAI S，et al. NIFT：Neural Interaction Field and Template for Object Manipulation[C]// 2023 IEEE International Conference on Robotics and Automation（ICRA）. 2022：1875-1881.

[25] WALD J，DHAMO H，NAVAB N，et al. Learning 3d semantic scene graphs from 3D indoor reconstructions[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020：3961-3970.

[26] LIU Y Y，LONG C J，ZHANG Z X，et al. Explore Contextual Information for 3D Scene Graph Generation[J]. IEEE Transactions on Visualization and Computer Graphics，2023，29(12)：5556-5568.

[27] CHABRA R，LENSSEN J E，ILG E，et al. Deep Local Shapes：Learning Local SDF Priors for Detailed 3D Reconstruction[C]// Computer Vision—ECCV 2020，Lecture Notes in Computer Science. 2020：608-625.

[28] CHEN Z，ZHANG H. Learning Implicit Fields for Generative Shape Modeling[C]// 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition（CVPR）. 2019：5939-5948.

[29] YUAN W，KHOT T，HELD D，et al. PCN：Point Completion Network[C]// 2018 International Conference on 3D Vision（3DV）. 2018：728-744.

YUAN Youwen，born in 2001，postgraduate. His main research interests include 3D point cloud processing and analysis，and 3D interaction relationship analysis.

ZHAO Xi，born in 1985，Ph. D，professor，Ph. D supervisor，is a member of CCF（No. 86701M）. Her main research interests include 3D data analysis and processing and synthesis.

（责任编辑：何杨）