一种视触觉融合驱动的抓取目标形状重建方法

张怡阳 虞承舜 李彤 帅宇昕 宋荆洲 陈钢
（北京邮电大学智能工程与自动化学院，100876，北京）

摘要：针对非结构化家庭服务场景下，机器人抓取目标形状重建面临的抓取目标特性多样化和视线遮挡等挑战，提出一种视触觉融合驱动的抓取目标形状重建方法。首先，设计块特征转化网络，将单视角输入视觉点云与触觉阵列转化为块特征代理。其次，利用Transformer架构实现视触觉跨模态特征融合与补全，设计几何感知模块预测模拟局部几何关系。最后，结合形状查询器生成基于动态条件的查询嵌入，设计解码器折叠重建模块，补全基于块特征预测的点云，输出完整目标形状点云。实验结果表明：所提出方法对机器人执行抓取任务的抓取成功率有显著提升，形状重建性能相较于现有最佳基线模型提升了12.5%，在陌生物体上的平均准确度高达80.3%，同时触觉模态的融入使模型的倒角距离下降了1.532。所提方法通过跨模态数据互补可以实现更完整、鲁棒性更强的抓取目标点云形状重建，且在处理复杂多变目标时具有较高的鲁棒性与泛化能力，为非结构化家庭服务场景下，机器人抓取目标的形状重建提供了一种解决方案。

关键词：形状重建；视觉点云；触觉阵列；视触觉融合
中图分类号：TP242.6 **文献标志码：**AA
文章编号：0253－987X(XXXX)XX－0001－17

A Method of Reconstructing the Shape of Grasping Target Driven by Visual-Tactile Fusion

ZHANG Yiyhang,YU Chengshun,LI Tong,SHUAI Yuxin,SONG Jingzhou,CHEN Gang
(School of Intelligent Engineering and Automation, Beijing University of Posts and Telecommunications, Beijing 100876, China)

Abstract: To address the challenges of diverse target characteristics and visual occlusion in grasping target shape reconstruction for robots in unstructured home service scenarios, this paper proposes a shape reconstruction method for grasping targets driven by visual-tactile fusion. First, a block feature transformation network is designed to transform the single-view input point cloud and the tactile array into patch feature proxies. Second, the Transformer architecture is utilized to achieve visual-tactile cross-modal feature fusion and completion, and a geometric perception module is designed to predict and model local geometric relationships. Finally, combined with a shape query module to generate dynamic condition-based query embeddings, a decoder folding reconstruction module is designed to complete the point cloud based on patch feature predictions, thereby outputting the complete target

收稿日期：2026-03-24。

shape point cloud. The experimental results demonstrate that the proposed method significantly improves the grasp success rate of robotic grasping tasks. Compared with the current best baseline model, the proposed approach achieves a 12.5% improvement in shape reconstruction performance and attains an average accuracy of up to 80.3% on unseen objects. Furthermore, the integration of the tactile modality reduces the Chamfer distance by 1.532. Through cross-modal data complementarity, the proposed method achieves more complete and robust point cloud shape reconstruction for grasping targets, demonstrating strong robustness and generalization capabilities when handling complex and variable objects, thereby providing a solution for shape reconstruction of grasping targets for robots in unstructured home service scenarios.

Keywords: shape reconstruction; visual point cloud; tactile array; visual-tactile fusion

随着具身智能的发展，机器人正逐渐走进日常生活。家庭场景具有高度非结构化的特点，存在机器人抓取目标物体的几何形状多样、表面光学属性复杂等多样化特性的挑战。在这种场景下，获取目标物体完整的3D几何形状，是机器人实现灵巧抓取的前提[1]。点云能直接、精确地提供环境的三维几何信息，使机器人能够理解物体形状、位置并进行安全的物理交互。相较于RGB-D图像，视觉点云直接从三维空间采样，可以提供与机器人运动规划无缝衔接的几何数据，且对颜色变化不敏感，可靠性更高。由于点云属于一种不规则的非结构化数据[2]，可采用深度学习方法通过训练3D卷积神经网络，完成形状重建[3]。Pistilli等[4]提出了基于深度图卷积的形状估计网络。然而，基于深度学习的点云补全方法由于缺乏精确的局部形状约束，并不能很好地捕捉局部区域的几何细节和结构[5]。

在非结构化的家庭服务灵巧抓取任务中，视觉感知面临不可逾越的物理瓶颈：机器人交互时产生严重遮挡，且透明、高反光或无纹理物体导致深度相机失效[6]。触觉感知[7]可以填补视觉信息的空白，提供不受光照与材质影响的局部几何真值。触觉感知的实现依赖于触觉传感器，现已能够捕获机器人抓取过程中的高分辨率力学信息[8]。Bauza等[9]基于Gelslim传感器[10]，高精度地重建了物体的形状。Suresh等[11]将触觉信息作为高斯势添加到既有高斯过程图中，实现了精确的物体形状和表面重建；Ottenhaus等[12]利用隐式高斯过程对物体局部表面进行了高精度重建。因此，基于触觉可以实现精确的局部形状重建，能够对视觉信息进行补充。

视触空间特征融合的目标形状重建模型融合了双模态信息互补优势，解决了单一视觉模态的局限性。Zhang等[13]提出以自适应动态加权的方式实现视觉和触觉的信息融合。Smith等[14]将视触觉信息融合进图表，在网格表示中预测物体形状。Watkins-Valls等[15]采集物体背面的触觉信息，预测物体几何模型。Ottenhaus等[16]采用基于深度学习的不确定性驱动的探索策略进行形状重建。

而已知深度学习启发式算法的准确性和可靠性较差，并且融合视触觉信息的形状重建难以实现高分辨率的点云重建，因此如何基于视触觉特征融合实现高分辨率目标点云形状重建至关重要。

针对非结构化家庭服务场景下，机器人抓取目标的形状重建问题，本文提出一种基于视触觉信息融合的形状重建方法。首先，利用深度相机获取单视角场景点云，通过预处理得到高质量输入点云，同时采集目标接触区域触觉阵列。然后，建立融合视触觉空间特征的目标形状重建模型，将视觉点云与触觉阵列映射为块特征代理，设计了几何感知模块，显式建模点云的局部几何结构，提出了动态形状查询器，能够根据视觉编码器的输出自适应生成缺失部分的查询嵌入，并结合触觉特征实现跨模态条件解码，实现视触觉空间特征的融合以及目标点云的补全。最后，实现抓取目标的形状重建工作。该方法通过跨模态数据互补，摆脱了环境变化对重建稳定性的影响，可以低成本实现家庭服务场景下的全面且高鲁棒的形状重建，且对形状重建和机器人抓取领域具有积极意义。同时，该重建方法在搭载不同视觉与触觉传感器的机器人平台上也具有较好的通用性。

1 目标完整形状重建模型

本方法提出了一种基于集合到集合的视触觉点云形状重建模型，该模型将视觉点云和触觉阵列转换为一系列块特征，并利用Transformer架构的序列到序列生成能力和跨模态特征融合能力来实现视觉点云与触觉阵列的特征配准、融合和补全，最终实现抓取目标点云的形状重建任务。

1.1 目标物体形状重建模型架构

基于视触觉融合的物体形状重建网络的整体框架如图1所示。首先，对复杂场景点云进行预处理，获取视觉输入点云。在视觉输入端，对输入的视觉点云进行下采样以获得中心点，然后使用轻量级动态图卷积神经网络（DGCNN）来提取中心点周围的局部特征。将空间位置嵌入添加到视觉局部特征后，使用Transformer编码器架构来预测缺失部分的块特征。在解码阶段，查询器增添多接触区域的稀疏触觉特征，用于解码器的代理预测。在形状重建阶段，使用折叠重建网络以从粗到细的方式补全点云。

模型利用Transformer架构的序列到序列生成能力和跨模态特征融合能力来实现视触觉点云的特征融合和补全，最终实现抓取目标的形状重建任务。首先，将视觉点云和触觉阵列转换为块特征，用于表示点云中的局部区域。然后，类比语言翻译，将点云补全建模为一组到一组的翻译任务。其中，转换器将部分点云的块特征作为输入，并生成缺失部分的块特征。具体来说，给定代表部分点云的块特征集合$\Gamma=\{\mathbf{F}_1,\mathbf{F}_2,...,\mathbf{F}_N\}$，$N$为输入点云块特征数

量，将点云补全过程建模为集到集的转换问题，表达式为

$$H=A_{\mathrm{D}}(Q,A_{\mathrm{E}}(\Gamma)) \tag{1}$$

式中：A_{E}和A_{D}分别为编码器和解码器模块；$A_{\mathrm{E}}(\Gamma)$为编码器的输出特征；Q为解码器的动态查询特征；$H=\{\mathbf{H}_1,\mathbf{H}_2,...,\mathbf{H}_M\}$为缺失点云的预测块特征集合，$M$为缺失点云的预测块特征数量。

编码器-解码器架构分别由编码器和解码器中的多头自注意力层组成。编码器中的自注意力层首先使用长距离和短距离信息更新代理特征。然后，前馈网络使用多层感知机（MLP）架构进一步更新代理功能。解码器利用自注意力和交叉注意力机制来学习结构知识。自注意力层使用全局信息增强局部要素，而交叉注意力层则关注编码器的查询和输出之间的关系。为了预测缺失部分的块特征，使用动态查询嵌入，这使得解码器更加灵活，可以针对不同类型的对象及其缺失信息进行调整。

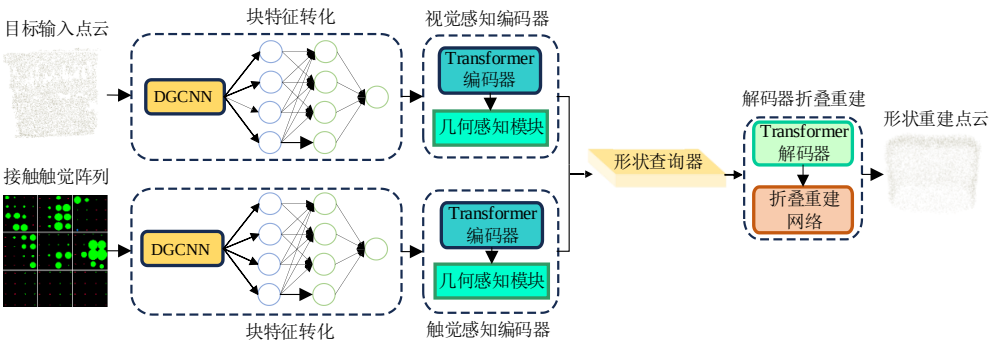


图1 基于视触觉融合的物体形状重建网络框架

Fig. 1 Network framework for object shape reconstruction based on visual-tactile fusion

1.2 视觉点云预处理

深度相机直接获取的点云较为复杂，为获取高质量输入点云，需要对场景点云进行预处理。

利用随机样本一致（RANSAC）算法，通过随机取点拟合出桌面平面，采用基于坐标轴阈值的空间框架方法从桌面点云中分离出目标物体。提取出物体点云后，采用统计滤波方法，对点云进行去噪，同时保留点云的有效部分。然后通过点云平滑化减少点云表面的粗糙度和不规则性。

1.3 输入端块特征转化

本方法改善了点云补全网络输入端的馈送方式，将获得的单视角视觉点云和触觉阵列表示为1组块特征。视觉块特征表示视觉不完整点云的各显著局部区域，触觉块特征表征各接触区域对应

的触觉阵列。

首先，进行最远点采样，以在部分点云中定位固定数量的中心点。然后使用具有分层下采样的轻量级DGCNN从输入点云中提取点中心的块特征，特征$\mathcal{F}_i$是一个特征向量，用于捕获q_i周围的局部结构，可以计算为

$$\mathcal{F}_i=\mathcal{F}_i'+\boldsymbol{\varphi}(q_i),i=1,2,...,N \tag{2}$$

式中：$\mathcal{F}_i'$为使用DGCNN模型提取的点q_i的领域特征向量；$\boldsymbol{\varphi}$为通过MLP捕获点q_i对应块特征的全局位置信息；q_i为第i个输入点云块特征中心点。

1.4 几何感知模块

为了方便Transformer更好地利用点云的三维几何结构的感应偏置，本方法设计了一个即插即用的几何感知模块来模拟几何关系，结构如图2所示。图中，T_Q为查询坐标；T_{k}为键特征；T_{m}为局部几何特征。

该模块使用K邻近算法（KNN）模型来捕获点云中的几何关系。首先，给定查询坐标，根据键坐标查询最近的键特征。然后，通过线性层的特征聚合来学习局部几何结构，并进行最大池化操作。最后，将几何特征和语义特征连接起来并映射到原始尺寸以形成输出查询特征Q。

1.5 形状查询器

查询特征Q用作待预测代理的初始状态。为了确保查询正确反映完整形状的框架，提出了一个查询生成器模块来生成基于编码器输出的动态条件的查询嵌入，结构如图3所示。图中，Q_{v}为视觉特征查询；Q_{t}为触觉特征查询；$\mathcal{F}_i^{Q_{\mathrm{v}}}$视觉查询嵌入特征；$\mathcal{F}_i^{Q_{\mathrm{t}}}$为触觉查询嵌入特征；$C$为块特征中心点坐标。

首先，在最大池化操作之后使用线性投影来总结。然后，直接使用线性投影层生成全局特征，这些特征可以重塑为缺失点云中心点坐标。最后，将编码器的全局特征和坐标连接起来，并使用MLP生成查询嵌入特征$\mathcal{F}_i^{Q}$，表达式为

$$\mathcal{F}_i^{Q}=\mathrm{MLP}(\mathrm{concat}(f_g,c_i)),g=1,2,...,M,i=1,2,...,M \tag{3}$$

式中：concat（·）为合并查询；MLP（·）为通过MLP进行特征映射；f_g为第g个全局特征向量；c_i为第i个缺失点云块特征中心点坐标。

合并触觉查询特征$\mathcal{F}_i^{Q_{\mathrm{t}}}$表达式为

$$\mathcal{F}_i^{Q}=\mathrm{concat}(\mathcal{F}_i^{Q_{\mathrm{v}}},\mathcal{F}_i^{Q_{\mathrm{t}}}),i=1,2,...,M \tag{4}$$

式中：$\mathcal{F}_i^{Q}$为全局视触觉特征。

在编码阶段，视觉模态和触觉模态各自引入了位置嵌入，为了节省计算成本和降低过拟合风险，直接将依据特征排序混合获取的全局视触觉特征送入解码器。

1.6 解码器折叠重建模块

本方法提出了一种多尺度的点云生成框架，以全分辨率恢复缺失的点云。为了减少冗余计算，再次使用查询生成器生成的坐标作为缺失点云的局部中心。然后利用折叠重建网络$f(\mathbf{H}_i)$来恢复以预测代理为中心的详细局部形状，

表达式为

$$O_i=f(\mathbf{H}_i)+c_i,\ i=1,2,...,M \tag{5}$$

式中：O_i是以c_i为中心的相邻点的集合。本方法可预测点云的缺失部分，并将它们与输入点云连接起来，以获得完整的对象。

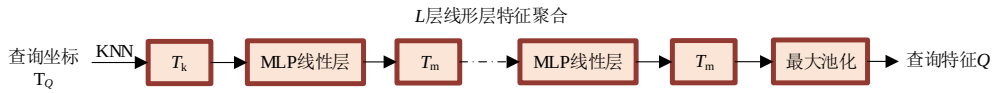


图2 几何感知模块

Fig. 2 Module of geometric perception

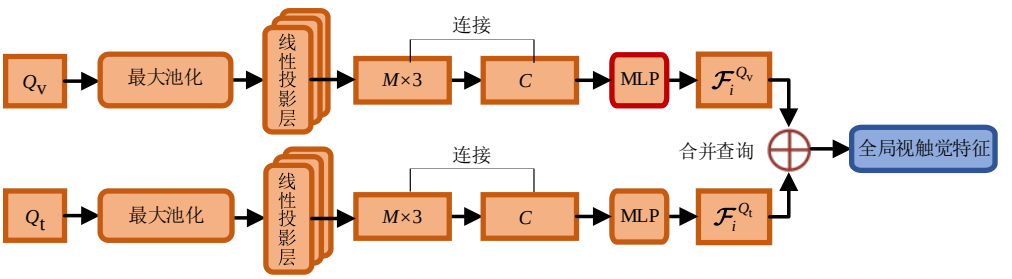


图3 形状查询生成器模块

Fig. 3 Shape query generator module

1.7 损失函数设计

点云补全的损失函数提供输出质量的定量测量。本方法引入倒角距离（CD）度量来设计损失函数，其为与点排列方式无关的指标，对点云的无序性具有不变性，适合用来衡量点云之间的相似性，倒角距离表达式为

$$d_{\mathrm{CD}}(S,W)=\frac{1}{n}\sum_{s\in S}\min_{w\in W}\|s-w\|+\frac{1}{m}\sum_{w\in W}\min_{s\in S}\|s-w\| \tag{6}$$

式中：d_{CD}为2个点云之间的倒角距离；S与W为2个待比较的点云集合；n为点云S包含点的数量；m为点云W包含点的数量；$\min\|\cdot\|$为对于一个点云中的某个点，另一个点云中的点距其最近的距离；s为点云S中的任意一个点；w为点云W中的任意一个点。

具体而言，将预测的局部中心和输入点云的中心连接起来形成整个抓取对象的局部中心；此外，直接使用高分辨率点云X来监督稀疏点云Y以激励相似分布。损失函数有如下具体表示。

（1）L_0为针对局部中心点的局部约束损失项。设Y为预测的局部中心点云，包含n_Y个点；X为高分辨率真实点云，包含n_X个点，L_0表达式为

$$L_0=\frac{1}{n_Y}\sum_{y\in Y}\min_{x\in X}\left\|y-x\right\|+\frac{1}{n_X}\sum_{x\in X}\min_{y\in Y}\left\|x-y\right\| \tag{7}$$

式中：y为点云Y中的任意一个点；x为点云X中的任意一个点；第一项表示从预测点云到真实点云的单向距离；第二项表示从真实点云到预测点云的单向距离。

（2）L_1为针对完整点云的全局约束损失项。设Z为补全的完整点云，包含n_Z个点；L_1表达式为

$$L_1=\frac{1}{n_Z}\sum_{z\in Z}\min_{x\in X}\left\|z-x\right\|+\frac{1}{n_X}\sum_{x\in X}\min_{z\in Z}\left\|x-z\right\| \tag{8}$$

式中：z为点云Z中的任意一个点；第一项表示从补全点云到真实点云的单向距离；第二项表示从真实点云到补全点云的单向距离。

（3）L为损失函数。表达式为

$$L=L_0+L_1 \tag{9}$$

最终损失函数是L_0和L_1之和，其意义在于L_0监督局部中心点，使其分布与真实点云相似；L_1监督补全点云，使其完整度和质量更高；通过两者的联合优化，既保证局部中心点的准确性，也提升了点云补全的整体性能。

2 数据集部署与训练策略

本方法基于ShapeNet数据集[17]在搭建的机器人抓取目标形状重建实验平台上进行训练和测试，然后再迁移到VTG数据集[18]。

（1）实验平台搭建。机器人抓取目标形状重建实验平台由UR10六自由度机械臂、ZED 2i深度相机、Robotiq三指手以及uSkin触觉传感器组成，平台结构如图4所示。

（2）ShapeNet数据集部署。原始ShapeNet数据集中包含55个类别物体的模型，首先将原始的ShapeNet数据集分为2个部分：ShapeNet-34数据集包含34个已知的类别物体，ShapeNet-21数据集包含21个相似的类别物体。在已知的类别中，随机从每个类别中选取100个物体，构建一个包含3 400个物体的测试集，其余的物体用于训练集，最终得到4 6765个物体模型用于训练。本方法还构建了另一个包含2 305个来自21个类似类别物体的测试集。

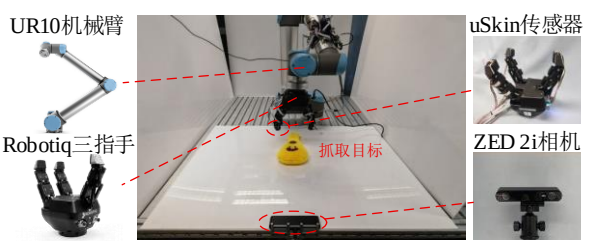


图4 机器人抓取目标形状重建实验平台

Fig. 4 Experimental platform for reconstructing the shape of robot grasping targets

（3）VTG测试数据集部署。为更好地模拟家庭服务场景中抓取物体的形状重建，选择VTG数据集作为陌生类别物体测试集，其包含18个物体及其对应的接触触觉数据，以及3个视角的点云。每个物体选取10组接触触觉数据以及对应的10个上视角点云、10个左视角点云、10个右视角点云，利用上述数据以及扫描得到的物体完整形状点云构建一个包含180组触觉数据、540个单视角点云、18个物体完整形状点云的测试集。部分VTG数据集对象示例如图5所示。

（a） 罐头 （b） 网球 （c） 橡胶棒 （d） 棒球

图5 VTG数据集部分物体示例

Fig. 5 Examples of objects in the VTG dataset

（4）双输入数据获取。对于单一模态数据集ShapeNet，在一个样本中，视觉点云和触觉阵列可以通过如下方法获取。以VTG数据集为例，从上、左、右3个视角，使用ZED 2i采集物体的点云，从而获取3个视角的目标场景点云，经过预处理后得到作为输入的单视角点云；通过安装的uSkin传感器，获取包含432个数据点的抓取目标物体触觉阵列信息，并进行触觉数据可视化，从而获取触觉阵列。

在配备有2个GPU（NVIDIA GeForce GTX 3090Ti）、1个CPU（Intel i9-10900X 3.7 GHz）和64 GB内存的平台上进行模型训练。为了优化，我们使用学习率为5×10^{-6}、权重衰减为1×10^{-4}的Adam优化器。批量大小设置为8，使用CD损失函数对模型进行200个迭代周期的训练。

3 实验结果与讨论

在具体实验中，本文从对象中采样8 192个点作为重建目标物体点云真实值。

3.1 评价指标

考虑到指定了目标物体点云点数，因此将精确率作为准确度指标。对于重建点云U和目标点云V给定匹配阈值$\delta>0$，根据点$u_i\in U$与$v_j\in V$的几何距离$\|u_i-v_j\|$是否小于该阈值来判断是否可以认为两点匹配成功。具体地，精确率P定义为预测点云U中正确匹配到目标点云V的点的比例，表达式为

$$P=\frac{|\{u_i\in U:\min\limits_{v_j\in V}\|u_i-v_j\|\le\delta\}|}{n_U} \tag{10}$$

式中：δ匹配阈值；$\min\limits_{v_j\in V}\|u_i-v_j\|$表示点$u_i\in U$到目标点云$V$中最近点的距

离，若其不大于δ，则认为u_i被正确匹配；n_U为重建点云U的点数。δ=4时模型性能评价指标的区分度最高。

召回率或覆盖率R定义为目标点云V中被重建点云U覆盖的点的比例，表达式为

$$R=\frac{|\{v_j\in V:\min_{u_i\in U}\ /\!/\ v_j-u_i\ /\!/\leq\delta\}|}{n_V} \tag{11}$$

式中：n_V为目标点云V的点数。

则F1分数F_{F1}表达式为

$$F_{\mathrm{F1}}=2\frac{PR}{P+R} \tag{12}$$

F1分数是精确率和召回率的调和平均值，综合考虑了点云预测的完整性和准确性。

除此之外，我们还将损失函数设计中的CD作为重建点云和真实点云几何结构匹配度的指标，CD越小代表点云重建结果越接近真实表面。

3.2 有效性实验

为了验证所提方法的有效性，我们选取了7个在日常家庭服务场景中常见、且未参与模型训练的真实物体（苹果、香蕉、易拉罐、棒球、橡胶棒、T型积木、带孔法兰）在搭建的机器人抓取目标形状重建实验平台上进行闭环验证。图6展示了在评价指标区分度最高的δ为4时，模型的重建推理效果。

3.3 真实抓取任务验证评估

为验证本文方法对下游抓取任务的实际影响，本文选取了VTG数据集中，家庭服务场景常见的5种物体（网球、棒球、牛奶、易拉罐、把手），设计了两种对比执行策略进行GraspNet[19]抓取测试。

（1）基线策略，仅依赖单视角点云进行GraspNet抓取执行。

（2）本文策略，机器人进行预抓取，采集视觉与触觉信息，使用本文模型在线重建目标形状点云，再进行GraspNet抓取执行。

针对2种对比执行策略，每种物体均进行50次随机位姿的抓取尝试。实验结果如表1所示，在残缺视觉下，由于物体背部几何信息丢失，极易发生空抓或边缘滑脱，平均抓取成功率仅为72.4%。而采用本文重建网络补全形状后，平均抓取成功率提升了11.2%，抓取成功率高达83.6%。这证明了本文所提方法能有效地提升机器人抓取任务成功率。

3.4 一般性泛化实验

本节进行一般性泛化实验，研究现有方法（PoinTr[20]、SnowflakeNet[21]、LAKe-Net[22]、SeedFormer[23]、AdaPoinTr[24]、CompleteDT[25]、Dapointr[26]、FACNet[27]）和本文方法在面向多类别对象时的表现性能。本文将测试集分为3个难度，分别为简单、中等、困难。在目标点云重建实验中，难度指补全任务的困难程度，难度较高的任务通常意味着输入点云中缺失的信息较多。本文将简单、中等、困难难度下

的输入点云点数设定为输入视觉点云数量的100%、75%、50%。此外，本文在VTG测试集上的测试中，额外引入了模型单次前向推理时间的评估，对模型推理速度进行量化分析。

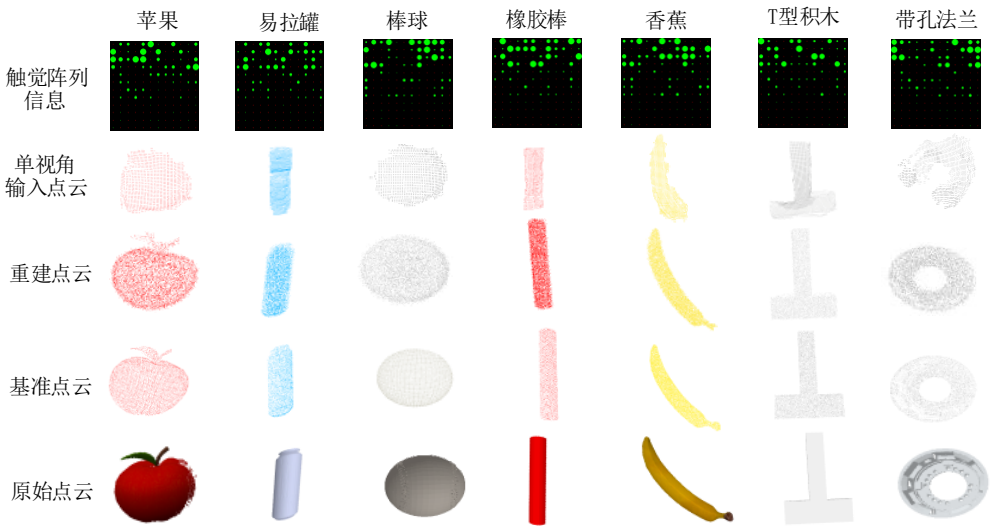


图6 不同目标物体的重建效果

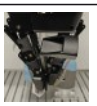
Fig. 6 Reconstruction results for different objects

一般性泛化实验结果见图7~图9，结果展示了δ=4时，模型在不同类别物体上的平均推理的测试效果，其所给结果为预测样本指标的均值。实验中的基线模型均引入了触觉管道，以输入单视角点云与触觉阵列图像。在已知类别物体上，本文提出的形状重建网络模型的重建性能相较于最优基线模型Dapointr提升了12.5%；在相似类别物体上，与AdaPointr相比，本文方法在3种困难程度下，CD指标分别提高了1.163、1.412和2.271，且本方法的平均CD和平均准确度指标显著地优于所有基线模型；在陌生类别物体上，模型的平均准确度高达80.3%。在3种跨数据集场景下，本模型均表现出更低的CD值和更高的准确度，这验证了其对不同形状物体更好的泛化能力。并且随着任务难度的增加，模型的CD值增幅显著小于其他基线模型，这表明模型在处理复杂形状时更稳定。详细实验数据见表2、表3与表4。其中D_{CDs}、D_{CDm}、D_{CDh}分别表示3种难度下模型的CD值，D_{CDavg}表示平均CD值。

本文模型在实现最优重建精度的同时，单次前向推理耗时仅约为37.27 ms。相比于轻量模型FACNet，本模型由于折叠重建模块从粗到细的推理过程，推理时间增加，但相较于其他基线模型，本文方法凭借轻量级DGCNN，显著减少了单次前向推理时间，成功实现了计算效率与重建精度的有效平衡。

表1 抓取任务实验结果

Table 1 Experimental results on the grasping task

抓取目标		抓取成功率/%	
		基线策略	本文策略
网球		70.0	86.0
棒球		72.0	84.0
牛奶		88.0	90.0
易拉罐		74.0	82.0
把手		58.0	76.0

3.5 消融实验

为了全面检验本文方法核心模块设计与训练策略的有效性，分别针对网络架构的关键组件和损失函数项展开了详细的消融研究。所有消融模型均在VTG测试集上进行评估，以保证对比基准的一致性。

(1)模型架构消融。为验证基于视触觉融合的物体形状重建网络框架中各模块的贡献，本文设计了逐步递增关键组件的消融方案，结果如表5所示。

由表可知，模型A是用于单视角视觉点云补全的普通模型，它使用标准Transformer架构。模型B是由在编码器和解码器之间添加查询生成器构成。由该模型可以看到，查询生成器有效引导了缺失特征的预测方向，将倒角距离减少了0.503。模型C为引入DGCNN，该模型从输入单视角视觉点云中提取特征时，能够使CD显著减少到5.320。模型D为将几何模块添加到所有Transformer模块，其性能得到进一步提高，这证明了模块学习的几何结构的有效性。最后，模型E为引入触觉通道，得到融合视触觉信息的完整模型，该模型的触觉模态提供了高精度的局部几何真值约束，使CD进一步减少了1.532，这也证明了触觉模态对于目标重建的有效性。

(2)损失函数消融。为了验证联合损失函数的必要性，本文在保持模型E不变的前提下，对损失函数项进行了消融验证，测试结果见表6。

当网络仅采用局部约束L_0监督时，由于缺乏对最终高分辨率点云生成的全局约束，局部点云较为发散，导致CD值较高，且整体准确度偏低。相反，当仅采用全局约束L_1时，由于网络在生成稀疏局部中心时缺乏直接监督，导致解码器代理预测过程容易陷入局部最优，虽然最终指标优于前者，但重建细节仍存在结构性缺陷。实验结果表明，联合损失函数在各项指标上均取得了最优表现，证明了该联合损失函数在提升模型重建精度上的有效性。

3.6 跨硬件平台泛化性实验

为了验证模型不依赖于特定的视觉相机型号或触觉传感器数量，本节在机

器人抓取目标形状重建实验平台的基础上另外设计了2种不同的硬件实验平台：平台V1将三指uSkin传感器减少至单指uSkin传感器；平台V2将ZED 2i更换为Intel Realsense D435i。对于平台V1、V2，直接使用基准平台训练完成的模型权重在VTG测试集上进行前向推理实验，实验结果见表7。

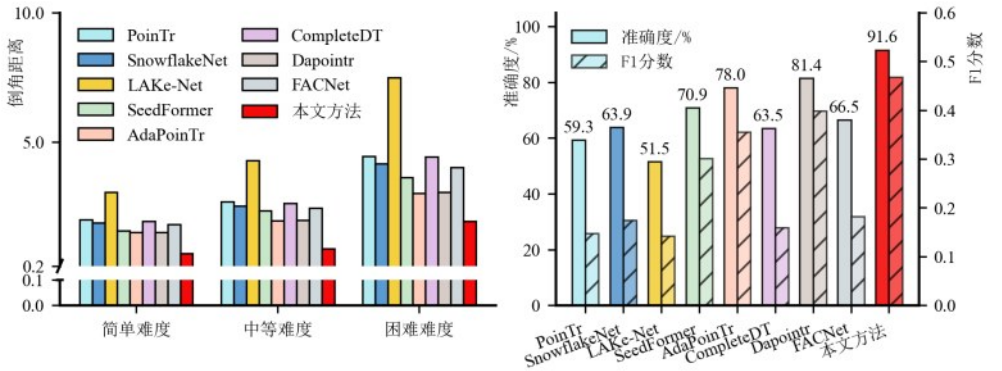


图7 ShapeNet-34测试集上模型性能对比结果

Fig. 7 Comparison of models performance on the ShapeNet-34 test dataset

（a）不同难度等级下倒角距离 （b）准确度与F1分数对比

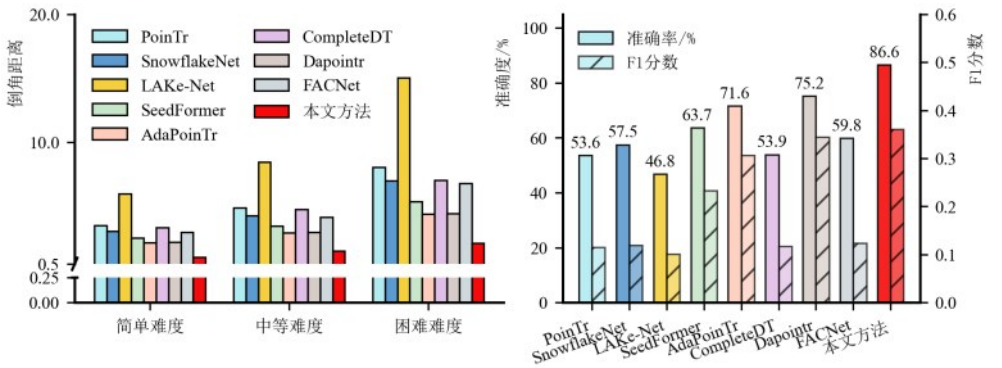


图8 ShapeNet-21测试集上模型性能对比结果

Fig. 8 Comparison of models performance on the ShapeNet-21 test dataset

（a）不同难度等级下倒角距离 （b）准确度与F1分数对比

在实验平台V1上，模型的CD值上升至3.421，准确度下降至76.2%，性能下降约4.1%，即使触觉传感器数量大幅减少，模型仍能保持较高重建精度，验证了模型对触觉传感器部分缺失的鲁棒性；在实验平台V2上，模型的CD值为3.286，准确度为77.5%，性能下降仅3.5%，说明模型对不同型号深度相机采集的点云具有良好的适应能力。

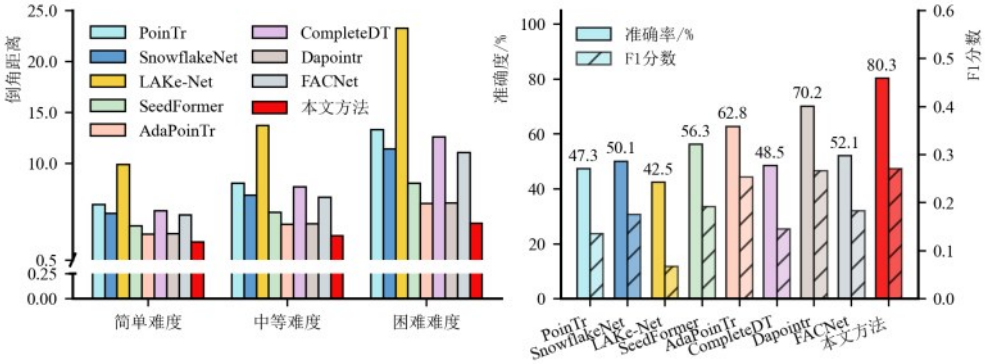


图9 VTG测试集上模型性能对比结果

Fig. 9 Comparison of models' performance on the VTG test dataset

（a）不同难度等级下倒角距离 （b）准确度与F1分数对比

表2 ShapeNet-34测试集实验结果

Table 2 Experimental results on the ShapeNet-34 test dataset

模型	D_{CDs}	D_{CDm}	D_{CDh}	D_{CDavg}	准确度/%	F1分数
PoinTr	1.994	2.682	4.441	3.039	59.3	0.147
SnowflakeNet	1.862	2.514	4.148	2.841	63.9	0.175
LAKe-Net	3.060	4.278	7.486	4.941	51.5	0.142
SeedFormer	1.562	2.343	3.632	2.512	70.9	0.301
AdaPoinTr	1.498	1.962	3.022	2.161	78.0	0.355
CompleteDT	1.929	2.636	4.422	2.996	63.5	0.159
Dapointr	1.513	1.982	3.052	2.182	81.4	0.398
FACNet	1.806	2.439	4.024	2.756	66.5	0.182
本文方法	0.696	0.877	1.298	0.957	91.6	0.468

表3 ShapeNet-21测试集实验结果

Table 3 Experimental results on the ShapeNet-21 test dataset

模型	D_{CDs}	D_{CDm}	D_{CDh}	D_{CDavg}	准确度/%	F1分数
PoinTr	3.509	4.898	8.062	5.490	53.6	0.115
SnowflakeNet	3.072	4.270	7.003	4.782	57.5	0.119
LAKe-Net	5.980	8.451	15.042	9.824	46.8	0.101
SeedFormer	2.521	3.454	5.381	3.785	63.7	0.233
AdaPoinTr	2.175	2.936	4.404	3.172	71.6	0.307
CompleteDT	3.355	4.772	7.040	5.056	53.9	0.117
Dapointr	2.197	2.965	4.448	3.203	75.2	0.344
FACNet	2.980	4.142	6.793	4.638	59.8	0.124
本文方法	1.012	1.524	2.133	1.556	86.6	0.361

4 结 论

表4 VTG测试集实验结果

Table 4 Experimental results on the VTG test dataset

模型	D_{CDs}	D_{CDm}	D_{CDh}	D_{CDavg}	准确度/%	F1分数	单次前向推理时间/ms
PoinTr	5.930	8.065	13.307	9.101	47.3	0.136	30.86
SnowflakeNet	5.063	6.886	11.413	7.787	50.1	0.176	660.23
LAKe-Net	9.883	13.688	23.256	15.609	42.5	0.068	782.59
SeedFormer	3.862	5.175	8.052	5.696	56.3	0.192	3411.21
AdaPoinTr	3.067	4.018	6.067	4.384	62.8	0.253	38.46
CompleteDT	5.329	7.709	12.584	8.541	48.5	0.145	52.91
Dapointr	3.096	4.054	6.128	4.426	70.2	0.267	41.87
FACNet	4.911	6.679	11.071	7.554	52.1	0.183	11.62
本文方法	2.297	2.885	4.128	3.103	80.3	0.271	37.27

表5 模型架构消融实验结果

Table 5 Results of the model architecture ablation experiment

模型	形状查询器	DGCNN	几何感知模块	触觉	倒角距离	准确度/%
A	否	否	否	否	6.821	56.2
B	是	否	否	否	6.318	58.7
C	是	是	否	否	5.320	60.9
D	是	是	是	否	4.635	72.7
E	是	是	是	是	3.103	80.3

表6 损失函数消融实验结果

Table 6 Results of the loss function ablation experiment

损失项	L_0	L_1	倒角距离	准确度/%
仅局部中心点损失项	是	否	4.207	71.9
仅全局点云损失项	否	是	3.876	74.6
联合损失函数	是	是	3.103	80.3

针对非结构化家庭服务场景下，机器人抓取目标的点云形状重建问题，提出一种基于视触觉信息融合的高精度形状重建方法。该方法提出块特征表示法，将视觉点云和触觉阵列转换成块特征，通过Transformer架构和折叠重建网络实现目标形状从粗到细的补全，有效地平衡了计算效率与重建质量，可以有效提升抓取任务的成功率。

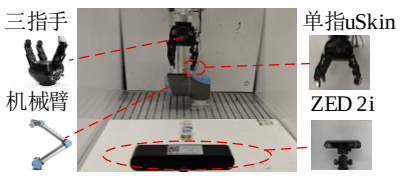
该方法在不同测试集上的平均CD和平均准确度都显著优于其他基线模型，在陌生类别物体上的平均准确度达

到了 80.3%，单次前向推理时间仅为 37.27 ms。

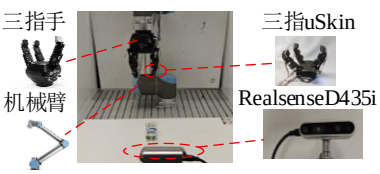
表 7　跨硬件平台泛化性实验结果

Table 7　Results of the cross-hardware platform generalization experiment

实验平台	倒角距离	准确度/%
基准平台	3.103	80.3
平台 V1	3.421	76.2
平台 V2	3.286	77.5



（a）　实验平台 V1



（b）　实验平台 V2

图 10　跨硬件泛化性实验平台

Fig. 10　Cross-hardware generalization experiment platforms

消融实验表明，几何感知模块与动态查询生成器的结合，使模型能够显式建模点云局部几何关系，使 CD 指标进一步下降；触觉模态的加入则直接提升了模型的重建精度，验证了模型跨模态特征融合架构的优越性；联合损失函数通过多尺度联合监督机制有效提升了模型重建精度。

本文方法可为家庭服务场景中机器人抓取目标的形状重建提供技术支撑，能够降低人工建模成本，对遮挡、阴影等实际工况具有良好的适用性，研究成果可定位为经过实验验证的应用基础研究成果，具备进一步开展工程应用验证和实际场景推广的潜力。

通过在公开数据集及真实搭建的机器人实验平台上的验证表明，本文提出的目标形状重建方法具有很高的准确性和泛化性，同时在搭载不同传感器的机器人平台上也有较好的通用性，但基于视触觉信息的形状重建算法只关注目标物体的几何特征，在处理不同类别物体时的重建结果缺乏语义一致性。因此，进一步研究语义增强的形状重建方法，构建包含视觉-触觉-语义信息的多模态融合目标形状重建网络是后续研究的重点。

参考文献

[1] LIU L, Yang W, Fei B. GeeNet: robust and fast point cloud completion for ground elevation estimation towards autonomous vehicles[J]. Frontiers of Information Technology & Electronic Engineering, 2024, 25 (7): 938-950.

[2] 武湘怡,叶海良,曹飞龙.基于平滑-锐化图卷积的点云补全方法[J/OL].计算机应用,1-12[2026-03-20].https://link.cnki.net/urlid/51.1307.TP.20250924.1815.006.
WU Xiangyi, YE Hailiang, CAO Feilong. Point cloud completion method based onsmooth-sharpen graph convolution[J/OL]. Journal of Computer Applications, 1-12[2026-03-20]. https://link. cnki. net/urlid/51.1307.TP.20250924.1815.006.

[3] 陆春媚, 杨志景. 多级精细化反卷积点云

补全网络[J]. Journal of Computer Engineering & Applications, 2023, 59(17).
LU Chunmei, YANG Zhijing. Multistage Refinement of Deconvolution Point Cloud Complementation Network[J]. Journal of Computer Engineering & Applications, 2023, 59(17).

[4] PISTILLI F, Fracastoro G, Valsesia D, et al. Learning robust graph-convolutional representations for point cloud denoising [J]. IEEE Journal of Selected Topics in Signal Processing, 2020, 15(2): 402-414.

[5] 石键瀚.基于融合触觉和视觉的点云补全理论研究[D].南京邮电大学,2024. DOI:10.27251/d.cnki.gnjdc.2024.000420.

[6] 陈仁祥, 邱天然, 杨黎霞, 等. 基于空间信息聚合的遮挡目标抓取位姿检测[J]. Optics and Precision Engineering, 2024, 32 (18): 2792-2802.
CHEN Renxiang, QIU Tianran, YANG Lixia. Occludedtarget graspingdetection method based on spatial information aggregation[J]. Optics and Precision Engineering, 2024, 32(18): 2792-2802.

[7] SHENG Q, Zhou Z, Li J, et al. A Comprehensive Review of Humanoid Robots[J]. SmartBot, 2025, 1(1): e12008.

[8] Li T, Yan Y, Yu C, et al. A comprehensive review of robot intelligent grasping based on tactile perception[J]. Robotics and Computer-Integrated Manufacturing, 2024, 90: 102792.

[9] BAUZA M, Canal O, Rodriguez A. Tactile maping and localization from high-resolution tactile imprints[C]//2019 International Conference on Robotics and Automation (ICRA). IEEE, 2019: 3811-3817.

[10] DONLON E, DONG S Y, LIU M, et al. GelSlim: a high-resolution, compact, robust, and calibrated tactile-sensing finger [C]//Proceedings of 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems. Piscataway: IEEE Press, 2018: 1927-1934.

[11] SURESH S, Si Z, Mangelson J G, et al. ShapeMap 3-D: Efficient shape map** through dense touch and vision[C]//2022 International Conference on Robotics and Automation (ICRA). IEEE, 2022: 7073-7080..

[12] OTTENHAUS S, Miller M, Schiebener D, et al. Local implicit surface estimation for haptic exploration[C]//2016 IEEE-RAS 16th International Conference on Humanoid Robots (Humanoids). IEEE, 2016: 850-856.

[13] ZHANG P, Bai L, Shan D, et al. Visual–tactile fusion object classification method based on adaptive feature weighting[J]. International Journal of Advanced Robotic Systems, 2023, 20(4): 17298806231191947.

[14] SMITH E, Meger D, Pineda L, et al. Active 3D shape reconstruction from vision and touch[J]. Advances in Neural Information Processing Systems, 2021, 34: 16064-16078.

[15] WATKINS-VALLS D, Varley J, Allen P. Multi-modal geometric learning for grasping and manipulation[C]//2019 International conference on robotics and automation (ICRA). IEEE, 2019: 7339-7345.

[16] OTTENHAUS S, Renninghoff D, Grimm R, et al. Visuo-haptic grasping of unknown objects based on gaussian process implicit surfaces and deep learning[C]// 2019 IEEE-RAS 19th International Conference on Humanoid Robots (Human-

oids). IEEE, 2019: 402-409.

[17] CHANG A X, Funkhouser T, Guibas L, et al. Shapenet: An information-rich 3d model repository[J]. arXiv preprint arXiv: 1512.03012, 2015.

[18] LI T, Yan Y, Yu C, et al. VTG: A Visual-Tactile Dataset for Three-Finger Grasp[J]. IEEE Robotics and Automation Letters, 2024.

[19] Fang H S, Wang C, Gou M, et al. Graspnet-1billion: A large-scale benchmark for general object grasping[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020: 11444-11453.

[20] YU X, Rao Y, Wang Z, et al. Pointr: Diverse point cloud completion with geometry-aware transformers[C]//Proceedings of the IEEE/CVF international conference on computer vision. 2021: 12498-12507.

[21] XIANG P, Wen X, Liu Y S, et al. Snowflakenet: Pointcloud completion by snowflake point deconvolution with skip-transformer[C]//Proceedings of the IEEE/CVF international conference on computer vision. 2021: 5499-5509.

[22] TANG J, Gong Z, Yi R, et al. Lake-net: Topology-aware point cloud completion by localizing aligned keypoints[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022: 1726-1735.

[23] ZHOU H, Cao Y, Chu W, et al. Seedformer: Patch seeds based point cloud completion with upsample transformer [C]//European conference on computer vision. Cham: Springer Nature Switzerland, 2022: 416-432.

[24] YU X, Rao Y, Wang Z, et al. AdaPoinTr: Diverse point cloud completion with adaptive geometry-aware transformers [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45(12): 14114-14130.

[25] Li J, Guo S, Wang L, et al. CompleteDT: Point cloud completion with information-perception transformers[J]. Neurocomputing, 2024, 592: 127790.

[26] Li Y, Zhou Q, Gong J, et al. Dapointr: Domain adaptive point transformer for point cloud completion[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2025, 39(5): 5066-5074.

[27] Yu X, Li J, Wong C C, et al. FACNet: Feature alignment fast point cloud completion network[J]. Computational Visual Media, 2025, 11(1): 141-157.