

基于点云的类别级物体姿态估计

栗仁武^{1,2}, 张凌霄^{1,2}, 高林^{1,2}, 李淳芃^{1,2}, 蒋浩^{1,2}

(1. 中国科学院计算技术研究所, 北京 100190;

2. 中国科学院大学, 北京 100049)

摘要: 针对类别级的物体姿态估计问题, 提出一种仅将深度相机扫描的点云作为输入, 在仅知道目标物体点云类别的情况下, 准确估计目标物体三维位姿的方法。该方法不需要依赖大量的带标签的人工标注数据集, 仅使用虚拟仿真技术模拟生产的数据, 即可在真实数据集上取得较高的精度。该方法首先对输入点云进行背景噪声过滤, 之后通过中心预测模块对点云做标准归一化, 再使用基于对应类别模板点云变形的的方法预测其标准坐标系坐标, 最后通过最小二乘法获得目标物体的三维位姿。实验结果表明, 该方法在真实数据上具有更好的泛化性能和更高的精度。

关键词: 姿态估计; 点云处理; 深度学习; 模式识别

中图分类号: TP391

文献标志码: A

doi: 10.11959/j.issn.2096-6652.202227

Category-level object pose estimation from depth point cloud

LI Renwu^{1,2}, ZHANG Lingxiao^{1,2}, GAO Lin^{1,2}, LI Chunpeng^{1,2}, JIANG Hao^{1,2}

1. Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China

2. University of Chinese Academy of Sciences, Beijing 100049, China

Abstract: Aiming at the problem of category-level object pose estimation, a method was proposed to accurately estimate the pose of the target object by only taking the point cloud scanned by the depth camera as the input, with knowing the category of input point cloud only. The method did not rely on a huge amount of labeled dataset, but used virtual data produced by simulation instead, which achieved better accuracy on real-world dataset. This method first filtered the background noise of the input point cloud. Then standardized the point cloud through the well-designed center prediction module. After that, the normalized object coordinate space would be estimated through a shape template deformation module. Finally, the pose would be obtained from least squares. Experiments on real-world dataset demonstrates that the method achieve higher accuracy and better generalization ability.

Key words: pose estimation, point cloud processing, deep learning, pattern recognition

0 引言

机械臂为了抓取目标, 首先需要对场景进行理解, 识别出感兴趣目标, 之后求解感兴趣目标相对于机器人相机坐标系的姿态, 然后在感兴趣目标表

面选取合适的抓取点, 传送给机器人的决策系统进行抓取。机械臂需要通过视觉传感器(如 RGB 相机或者深度相机, 以及相应的感知算法)感知周围环境。相较于传统的 RGB 相机, 深度相机能够感知场景中物体的深度信息, 可以帮助机器人更准确

收稿日期: 2022-02-16; 修回日期: 2022-04-01

通信作者: 李淳芃, cpli@ict.ac.cn

基金项目: 科技创新 2030—“新一代人工智能”重大项目 (No.2018AAA0103002); 中科院科技服务网络计划 (No.KFJ-STIS-QYZD-129)

Foundation Items: The National Key Research and Development Program of China (No.2018AAA0103002), Science and Technology Service Network Initiative, CAS (No.KFJ-STIS-QYZD-129)

地感知现实世界。随着 Kinect、RealSense 等深度相机的推出, 能够得到深度信息的深度相机得到了巨大发展和普及, 并被广泛作为机械臂的“眼睛”感知现实世界。如何通过深度相机感知并理解工作场景, 检测或者分割出目标物体、对目标姿态进行准确的估计、选取合适的抓取点成为机械臂研究的核心问题。

近年来, 基于深度学习的目标检测和分割取得了令人瞩目的成果, 相关技术也比较成熟, 而且已经成功在工业界大量部署应用。目标姿态估计一般指 6 自由度的姿态估计, 即目标物体相对于当前相机坐标系的刚性变换: $(R, t) \in \text{SE}(3)$, 其中 $R \in \text{SO}(3)$, $t \in \mathbb{R}^3$ 。三维旋转空间具有搜索空间大、高度非线性、优化困难的特点, 而且姿态估计问题对精度要求更高。目前的 6 自由度姿态估计方法大多集中在实例级别上, 传统的基于图像模板或者点云特征的匹配方法存在效率低下、对遮挡敏感、难以处理对称物体的问题。本文针对类别别的姿态估计问题, 在只知道目标物体类别的情况下, 通过深度相机扫描出的点云, 估计出目标物体的姿态。

1 背景与相关工作

物体姿态估计在机器人抓取和增强现实领域都起着至关重要的作用。6 自由度姿态估计的目的是得到物体在相机坐标系下的旋转矩阵和平移向量。根据实际运行阶段是否存在目标物体对应的三维模型, 可以将 6 自由度姿态估计划分为实例级别姿态估计和类别别姿态估计。一般来说, 对于每个三维目标物体, 实例级别姿态估计需要有相应的三维模型来做匹配, 而类别别姿态估计则对此没有要求, 只需要知道目标物体的类别。

1.1 实例级别姿态估计

(1) 基于图像模板匹配的方法

LINEMOD^[1]作为模板匹配方法的代表, 以颜色-深度 (RGB-D) 信息为输入, 能够在复杂背景下做到三维物体的实时检测和定位, 在弱纹理甚至无纹理的情况下, 依然能获取鲁棒的姿态估计, 更重要的是, 该方法不需要冗长的训练时间。该方法主要包含 3 个步骤: 模板采集、模板训练和模板匹配。首先, 模板采集阶段, 通过深度相机从多个视角、多个距离、多个方向对目标物体的三维模型进行拍摄; 其次, 模板训练阶段, 使用第一步获取的不同视角的 RGB-D 数据建立多视角、多模态的局

部特征模板; 最后, 模板匹配阶段, 对于实际场景下的目标物体, 通过图像特征提取、点云特征提取和对应的特征匹配方法, 获取图像特征点与三维模型的一一对应关系, 便可通过迭代最近点 (iterated closest point, ICP) 或者透视点 (perspective-n-point, PnP) 方法进行 6 自由度的位姿求解。基于该方法, Hinterstoisser S 等人^[2-3]做了进一步的改进和发展。

(2) 基于点对特征匹配的方法

Drost B 等人^[4]采用全局建模、局部匹配的思想首次提出了点对特征 (point pair feature, PPF), 简化了在三维场景中匹配三维模型的过程。以带法向信息的点云为输入, 对于点云中的任意两个点, 以两点距离和两点的法向量之间的几何关系形成的四维特征为 PPF。在离线建模阶段, 建立哈希表存储三维模型的所有 PPF, 并将其作为模型的整体描述。在在线测试阶段, 将全局坐标系作为场景点云和模型点云坐标系变换的桥梁, 降低了刚体变换的自由度, 使得位姿计算更加简单, 同时借助广义霍夫变换的思想, 将位姿空间参数化, 对参数空间进行投票, 最后获得可靠的位姿估计。该方法对轻微的形变和遮挡鲁棒性好, 抗噪声干扰能力强, 而且可以做多目标的匹配。但是由于该方法采用 PPF 对模型进行描述, 而对称物体对称点的特征相同, 因此该方法对对称物体存在歧义; 另外, PPF 提取和位姿投票过程中的高计算量也带来了效率低下的问题。Choi C 等人^[5]针对 PPF 做了进一步改进, 将两点各自的颜色信息引入 PPF, 将原先的四维特征扩充到十维, 即颜色点对特征 (color point pair feature, CPPF), 该特征描述子在彩色点云数据上取得了比原始 PPF 更好的结果。Choi C 等人^[6]继续对 PPF 做改进, 提出使用不同种类的点对关系, 比如使用边缘点为点对构建 PPF 特征, 该方法大大减少了匹配点的数量, 在离线建模阶段和在线测试阶段都大大提升了效率, 并且在包含大量平面的点云数据中取得了良好的效果。Vidal J 等人^[7]提出对 PPF 方法的进一步继承和改进, 该方法将基于 PPF 的系列方法总结为 6 个步骤: 预处理、PPF 提取、匹配、候选位姿估计、聚类和后处理, 并且对每个步骤做了针对性的改进, 在多个数据集上取得了优异的位姿估计结果。最近, Guo J W 等人^[8]针对复杂场景下的点云数据, 基于 PPF 提出一种新的基于物体中心与 PPF 的相对几何关系的物体中心投票机制, 该机制能够处理多目标和同一目标多实例的场

景, 并且鲁棒性和效率都更高。

(3) 基于深度学习的方法

近年来, 借助神经网络强大的表达能力, 深度学习方法在实例级别的 6 自由度位姿估计问题上已经超越了传统的匹配方法。Xiang Y 等人^[9]以 RGB 图片为输入, 不需要事先准备好对应的三维模型, 而是使用 3 个分支(语义分割、中心投票和旋转预测)直接输出 6 自由度位姿估计的参数。Li Y 等人^[10]使用 Xiang Y 等人^[9]提出的方法先得到目标物体的初始位姿, 计算当前位姿渲染得到的图像与真实位姿渲染得到的图像之间的损失并优化, 得到一个相对位姿量, 并且不断迭代优化位姿。Xu D F 等人^[11]提出将 RGB 特征与点云特征进行融合, 取得了良好的效果。之后 Wang C 等人^[12]提出一种逐点(像素点和三维点)的特征融合方式, 取得了更加准确的位姿估计结果。He Y S 等人^[13-14]另辟蹊径, 首先通过 RGB 特征与点云的融合特征预测目标物体的有序关键点, 之后建立预测关键点与事先标注的三维模型关键点之间的对应关系, 通过最小二乘拟合^[15]或 PnP^[16]方法估计目标物体的位姿, 并使用 ICP 方法^[17]进行最终位姿的修正。Labbé Y 等人^[18]将多视角的 RGB-D 数据作为输入, 对相机位姿和目标物体姿态联合优化求解。另外一些方法^[19-24]也在 RGB-D 数据上取得了良好的位姿估计效果。

1.2 类别别姿态估计

Sahin C 等人^[25]首次聚焦类别别姿态估计问题, 提出一种以深度图为输入, 适应纹理、光照和姿态变化等造成的形状差异的模型。作者在训练阶段提取三维计算机辅助设计(computer aided design, CAD)模型的骨架, 根据骨架将 CAD 模型分块, 创建随机森林, 将骨架结构作为形状不变特征, 并用于表示类别内实例的局部特征, 在测试阶段, 将新的目标物体分块, 在随机森林中进行比对。Wang H 等人^[26]提出一种被同一类别内部所有实例共享的规范表示归一化对象坐标空间(normalized object coordinate space, NOCS), 该方法学习输入图像中每一个像素到 NOCS 中三维坐标的映射, 从而能够通过 Umeyama 算法^[27]和随机采样一致(random sample consensus, RANSAC)方法^[28]得到从 NOCS 到原始三维空间的相似变换, 继而得到目标物体的姿态估计。Chen D S 等人^[29]针对不同实例间的形状差异, 提出规范形状空间(canonical shape space, CASS), 将同类别内所有实例统一起来。该方法以

单张 RGB-D 图像为输入, 训练 CASS 下的点云变分自编码器(variational autoencoder, VAE)^[30], 并将其作为隐式表示, VAE 将任何视角下的 RGB-D 图像映射为与视角无关的隐式表示, 通过和原始输入的特征进行对比, 计算相应的目标物体姿态。

Tian M 等人^[31]提出一种基于学习模板变形场的方法——形状先验变形(shape prior deformation, SPD)方法。该方法的核心思想是求解 NOCS 下类别模板点云 M 到观测点云 P 的变形场 D 和点到点的映射关系 A , 得到 P 在 NOCS 的点云 $Q=A(M+D)$, 之后便可借助最小二乘等方法间接求得 NOCS 点云 Q 到相机坐标系的观测点云 P 之间的相似变换, 即点云 P 的位姿。具体来说, 该方法首先以已知类别为先验, 在离线数据集上得到该类别在 NOCS 的平均点云, 并将其作为该类别的模板点云; 之后通过共享参数的神经网络分别提取观测物体与模板物体在 RGB 模态和点云模态上的特征, 并进行深度特征融合, 模板点云在观测点云特征的指导下学习该模板点云到观测点云在 NOCS 中的非刚性变形场与对应关系, 进而计算出相似变换, 得到观测物体的位姿估计。

2 类别别点云姿态估计方法

本文方法以深度相机扫描到的目标点云为输入, 预测其相对当前相机坐标系的位姿。本文方法主要包括 3 个模块: 点云去噪模块、中心预测模块以及基于类别模板的变形场预测的姿态估计模块。首先, 点云去噪模块针对输入点云中不可避免的背景噪声设计数据生成策略, 通过虚拟仿真数据模拟现实世界中结构光深度相机采集到的数据, 进而设计了一个二分类的点云分割网络, 借此识别出点云中的背景噪声并过滤; 中心预测模块以过滤后的点云为数据, 预测其中心点, 将其平移至其中心并归一化, 以便后续的姿态估计模块更容易学习其潜在表征; 最后本文采用基于类别模板的变形场预测的姿态估计模块, 以去噪且归一化后的点云为输入, 预测对应类别的模板点云到输入点云的变形场, 通过最小二乘法求解得到其姿态。

2.1 点云去噪

目标点云一般是由基于二维图像的分割方法(比如 MaskRCNN)从深度图中分割出前景, 再投影到三维坐标空间得到的, 目标周边的背景点容易被误识别为前景点, 从而对目标姿态估计产生影

响。为了抑制背景点被误识别带来的影响，作为首个直接在点云数据上应用深度学习的方法，PointNet 使用均值池化或者最大池化的方法来满足与点云顺序无关的要求，其网络结构简单，网络参数少，并在点云分类和点云分割任务中取得了优异的性能。Charles R Q 等人^[32]提出的 PointNet 使用若干共享参数的多层感知机 (shared MLP)，以点云中每个点的三维坐标为输入，将点云中的每个点分别映射到高维特征空间，得到在高维特征空间的表征，并使用该表征来预测每个点的标签。本文将点云去噪问题定义为前景点和背景点之间的二分类问题，借助 PointNet 模型将深度图投影得到的点云分割为两个部分，分别为前景点和背景点，前景点是想要保留的部分，背景点作为对后续步骤的干扰项应舍去。

2.1.1 数据生成

深度学习模型依赖于海量的训练数据，因此需要事先获取足够的训练数据。人工标注现实世界的点云的类别标签费时费力，因此本文尝试从虚拟仿真数据中模拟生成带有背景点的目标点云，以供深度学习模型进行训练。本文随机在背景中生成点状噪声、条状噪声、块状噪声，并且对前景进行随机的剔除与裁剪来模拟结构光深度相机因光照纹理等因素造成的孔洞。

具体来说，记当前虚拟仿真样本深度图为 D ，其前景掩码为 M ，与 M 形状相同的全 1 矩阵为 I ，与 M 形状相同的全 0 矩阵为 O 。虚拟数据生成步骤如下。

步骤 1：在 I 上随机选取两点 A 和 B ，连接 AB 两点，并填充为 0，得到 I' 。生成新的前景掩码 $M' = I' \odot M$ 。并重复步骤 1 若干次，来模拟真实深度图中的孔洞。

步骤 2：在 O 上随机选择若干点置为 1，得到

O' 。生成新的前景掩码 $M' = O' \oplus M$ 。以此模拟点状噪声。

步骤 3：在 O 上随机选择若干点置为 1，并且使用膨胀操作将其扩张，得到 O' 。生成新的前景掩码 $M' = O' \oplus M$ 。以此模拟块状噪声。

步骤 4：在 O 上随机选择两点 A 和 B ，连接 AB 两点，并填充为 1，得到 O' 。生成新的前景掩码 $M' = O' \oplus M$ 。重复步骤 4 若干次，以此模拟线状噪声。

上述步骤中的 $\odot$ 表示逐点乘运算， $\oplus$ 表示逐点加运算。前景点云 F 和背景点云 B 可由下列计算式定义：

$$F = \text{Proj}(D \odot M) \quad (1)$$

$$B = \text{Proj}(D \odot (M' - M)) \quad (2)$$

其中，Proj 表示将深度值投影到三维空间形成点云。完整点云 P 由前景点云 F 和背景点云 B 合并构成，赋予前景点云 F 中每个点的标签为 1，背景点云 B 中的每个点的标签为 0。

2.1.2 分类网络

如图 1 上半部分所示，输入点云 $P \in \mathbb{R}^{n \times 3}$ ，分类网络模型以多层感知机 (multilayer perceptron, MLP) 为编码器，将点云映射到高维特征空间，得到其逐点特征，其维度为 $n \times 512$ ，然后该逐点特征被送入由 MLP 组成的解码器，通过 Sigmoid 函数得到每个点属于前景点的概率：

$$l = \text{MLP}_1(\text{MLP}_2(P)) \quad (3)$$

$$\rho = \sigma(l) \quad (4)$$

其中， ρ 为解码器输出的未归一化的概率， σ 为 Sigmoid 函数。得到每个点属于前景点的概率之后便可以借此过滤背景噪声，以 0.5 为阈值，概率值大于 0.5 的点保留，小于或等于 0.5 的点舍去，

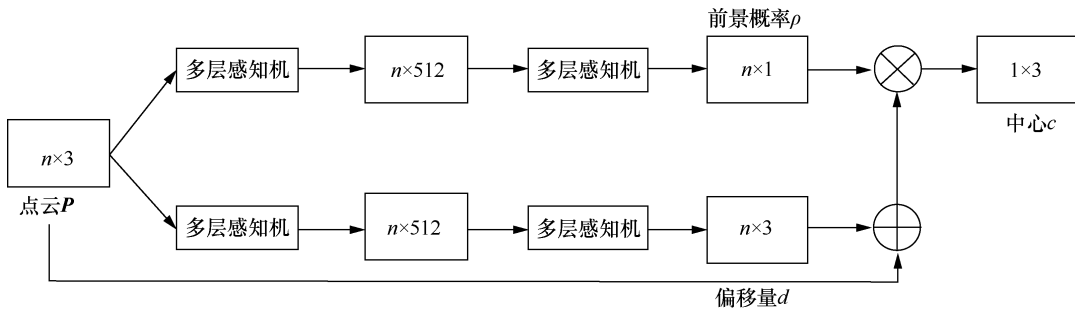


图 1 点云去噪与中心预测

并且从保留的点中随机选取点代替舍去的点，从而组成过滤后的新点云 P ，并保持点云的大小不变。

2.2 中心预测

由于相机视角的丰富性，不同视角下采集到的目标点云尺度和距离相差很大。为了便于深度学习模型更好地学习到输入点云的潜在表征，对输入进行归一化是必要的。

如图1下半部分所示，对于输入点云 $P \in \mathbb{R}^{n \times 3}$ ，同样以 MLP 为编码器，将点云映射到高维特征空间，得到其逐点特征，并送入共享 MLP，回归出每个点 $p_i \in P$ 到其中心点的偏移量 d_i ：

$$d = \text{MLP}^1(\text{MLP}^2(P)) \quad (5)$$

赋予属于前景点概率更高的点更多的权重，概率更低的点更少的权重。如图1所示，借助点云去噪模块中得到的前景点概率 ρ 对 d 进行加权求和，进而求得 P 的中心位置 c ：

$$c = \sum_{i=1}^N \rho_i (p_i - d_i) \quad (6)$$

一旦获取到点云 P 的中心，便可以对 P 进行归一化，得到归一化之后的点云 P' 。具体来说，对于 P' 中的每一点 P'_i ，先将其平移到点云的中心，再使用包围球半径进行尺度的归一化，即：

$$p'_i = p_i - c \quad (7)$$

$$s = 2 \times \max(|p'_{i1}|, |p'_{i2}|, |p'_{i3}|) \quad (8)$$

$$p'_i = p'_i / s \quad (9)$$

其中， p'_{i1} 、 p'_{i2} 、 p'_{i3} 分别为点 p'_i 的 x 坐标、 y 坐标和

z 坐标， s 为点云 P 最小包围球的半径。

2.3 基于类别模板的变形场预测的姿态估计

受 Tian M 等人^[31]提出的 SPD 方法的启发，本文使用基于类别模板变形场预测的方式来估计姿态。与 SPD 方法不同的是，首先，SPD 方法依赖于 RGB 特征与点云特征做逐点的特征融合，而在实际的应用环境中（尤其是工业场景），RGB 图像很难获取，并且虚拟仿真数据中的 RGB 图像的真实感和纹理丰富度与现实世界中实际采集到的图像差异巨大，同时，基于同一类别（比如杯子）的物体几何形状类似但纹理信息变化多的观察，本节提出的姿态估计模块仅将点云作为输入；其次，SPD 方法对于从真实环境中采集到的带噪声和孔洞的点云来说鲁棒性不够好，没有对此做针对性的处理，如前文所述，本文对点云进行去噪，提高了输入点云的质量，并且基于中心预测模块做了中心化和尺度归一化，使得网络更容易学习到数据的共性特征。

具体来说，本节介绍的基于类别模板的变形场预测网络框架如图2所示，输入为点云 $P' \in \mathbb{R}^{n \times 3}$ 和对应类别的模板点云 $M \in \mathbb{R}^{m \times 3}$ ，经过网络结构相同但是参数不同的 PointNet 编码器，分别得到 P' 和 M 的局部特征，每个点的特征维度为 512，之后分别对其做平均池化，得到 P' 和 M 的全局特征。

P' 在 M 的指导下，具体地说， P' 的全局特征和局部特征与 M 的全局特征被拼接在一起，经过多层感知机和 SoftMax 层输出预测的对应矩阵 $A \in \mathbb{R}^{n \times m}$ 。类似地， M 在 P' 的指导下，使用相似的特征融合方式，预测 M 到 P' 的变形场 $D \in \mathbb{R}^{m \times 3}$ 。模板点云 M 在变形场 D 的作用下得到

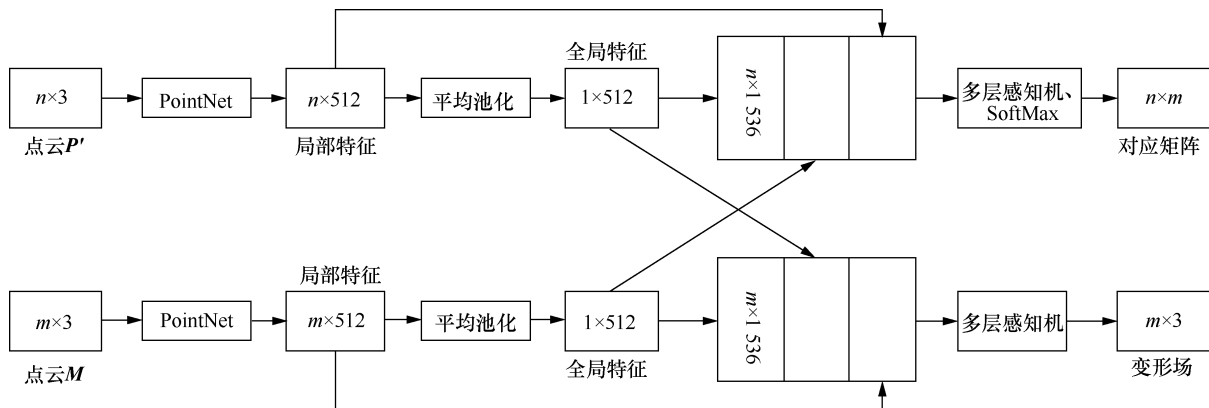


图2 基于类别模板的变形场预测网络框架

变形之后的 NOCS 点云坐标 $\mathbf{Q}$ 为：

$$\mathbf{Q} = \mathbf{A}(\mathbf{M} + \mathbf{D}) \in \mathbb{R}^{n \times 3} \quad (10)$$

得到 NOCS 点云 $\mathbf{Q}$ 和与其一一对应的观测点云 $\mathbf{P}'$ 后，可以使用基于最小二乘的 Umeyama 算法计算出观测点云相对于当前相机坐标系的相对位姿。

2.4 损失函数

(1) 点云去噪损失函数

将点云去噪问题定义为前景点和背景点的分类问题，通过式 (1) 和式 (2) 得到了虚拟仿真模拟数据的前景点和背景点的标签 $\hat{\rho}$ ，前景点标签为 1，背景点标签为 0。由于前景点和背景点存在类别不平衡问题，一般而言，背景点占比很小，因为在该分类问题中使用 Focal Loss 优化函数，具体地说，定义优化函数为：

$$L_{\text{denoise}} = -\alpha(\hat{\rho} - \rho)^\gamma \log(\rho) - (1 - \alpha)\rho^\gamma \log(1 - \rho) \quad (11)$$

其中 $\alpha = 0.25$, $\gamma = 2$ 。

(2) 中心预测损失函数

在中心预测模块中，预测点云 $\mathbf{P}$ 中每个点 p_i 到其中心的位移 d_i ，很容易根据点云 $\mathbf{P}$ 的真实中心算出其真实位移 $\hat{d}_i$ ，并将其作为监督信号。以 $L1$ 函数为损失函数：

$$L_{\text{center}} = \frac{1}{n} \sum_{i=1}^n \|d_i - \hat{d}_i\| \quad (12)$$

(3) 变形场预测损失函数

假设输入点云 $\mathbf{P}$ 对应 NOCS 的点云坐标 $\mathbf{P}_{\text{nocs}}$ 已知，则使用倒角距离 (chamfer distanc) 衡量 $\mathbf{Q}$ 与 $\mathbf{P}_{\text{nocs}}$ 之间的距离：

$$L_{\text{recon}} = d_{\text{CD}}(\mathbf{Q}, \mathbf{P}_{\text{nocs}}) \quad (13)$$

其中， d_{CD} 为倒角距离，定义如下：

$$d_{\text{CD}}(X, Y) = 0.5 \times (d(X, Y) + d(Y, X)) \quad (14)$$

其中， $d(X, Y)$ 表示点云 X 和点云 Y 的距离，定义如下：

$$d(X, Y) = \sum_{x \in X} \min_{y \in Y} \|x - y\|_2^2 \quad (15)$$

(4) 对应损失函数

通过对应矩阵 $\mathbf{A}$ 得到从模板点云到输入点云之间的一一映射，因此，可以通过对 $\mathbf{Q}$ 和其对应的真实值 $\mathbf{Q}_{\text{gt}}$ 进行约束，间接优化对应矩阵 $\mathbf{A}$ ， L_{cor} 为映射关系的误差损失：

$$L_{\text{cor}} = L1_{\text{smooth}}(\mathbf{Q}, \mathbf{Q}_{\text{gt}}) \quad (16)$$

(5) 正则损失函数

$\mathbf{A}$ 代表了从模板点云到输入点云之间的映射关系，因此，需要对 $\mathbf{A}$ 施加正则约束 L_{entropy} ，使其每一行尽量服从峰值分布。 $\mathbf{D}$ 表示模板点云的变形场，点云的变形一般是局部的，需要限制变形程度，因此对 $\mathbf{D}$ 施加正则约束 L_{def} 。具体定义如下：

$$L_{\text{entropy}} = \frac{1}{n} \sum_i \sum_j -A_{i,j} \log A_{i,j} \quad (17)$$

$$L_{\text{def}} = \frac{1}{m} \sum_{d_i \in D} \|d_i\|_2 \quad (18)$$

(6) 总损失函数

综上所述，总的损失函数 L_{total} 为各部分损失函数的和：

$$L_{\text{total}} = \lambda_1 L_{\text{denoise}} + \lambda_2 L_{\text{center}} + \lambda_3 L_{\text{recon}} + \lambda_4 L_{\text{cor}} + \lambda_5 L_{\text{entropy}} + \lambda_6 L_{\text{def}} \quad (19)$$

其中， $\lambda_1 = \lambda_2 = 1.0$, $\lambda_3 = 5.0$, $\lambda_4 = 1.0$, $\lambda_5 = 1 \times 10^{-4}$ 。

3 实验

3.1 实验数据

本文基于 NOCS 提供的虚拟仿真数据集——CAMERA25 数据集进行训练，在真实场景中采集的 REAL275 数据集上进行测试，与 NOCS 方法、SPD 方法进行对比分析。NOCS 数据集包含 6 个类别的物体：瓶子 (bottle)、碗 (bowl)、罐子 (can)、相机 (camera)、笔记本电脑 (laptop) 和杯子 (mug)。其中 CAMERA25 数据集是将人工合成的三维网格模型渲染到真实场景中生成的数据，一共有 300 K 的合成图片，其中 25 K 被用作验证集，训练集中包含 1 085 个物体，测试集中包含 184 个物体。REAL275 数据集为从真实场景中采集到的图片，其中从 7 个场景中采集到的 4 300 张图片被用来训练，从 6 个场景中采集到的 2 750 张图片被用来验证。

3.2 评价指标

使用不同阈值下的交并比 (intersection over union, IoU) 评价目标物体的三维检测精度， $3D_{50}$ 表示预测物体的有向包围框和真实值的交并比在 50% 以上的比例， $3D_{75}$ 表示预测物体的有向包围框和真实值的交并比在 75% 以上的比例，该指标越高越好。使用不同旋转、平移阈值下的平均精度来评价目标物体的姿态估计精度。记 $n^\circ m$ cm 为旋转误

差不超过 n° 、平移误差不超过 m cm 的阈值。设置 4 种不同的阈值，分别是 $5^\circ 2$ cm、 $5^\circ 5$ cm、 $10^\circ 2$ cm 和 $10^\circ 5$ cm，每个阈值下的指标同样是越高越好。对于轴对称物体，比如部分瓶子、碗和罐子，不计入相对于轴的旋转损失。对于带柄的杯子，视柄的可视性将其归入对称物体或者非对称物体。对于柄不可见的杯子，将其归入非对称物体；对于柄可见的杯子，将其归入对称物体。

3.3 实现细节

在搭载 NVIDIA1080Ti 显卡的 Ubuntu 系统电脑上完成所有实验。使用 PyTorch 框架进行深度学习模型的训练与测试，设置批处理参数为 16，初始学习率为 0.0001，学习率衰减策略为分段衰减，每 10 个迭代轮次 (epoch) 衰减一次学习率，衰减因子为 0.1，最大迭代 epoch 为 100。

3.4 实验结果

为了验证本文提出的方法的有效性，与 NOCS 方法、SPD 方法进行了对比，在 CAMERA25 数据集上进行训练，在 REAL275 测试集上进行测试，并采用第 3.2 节中描述的评价指标进行评估，结果见表 1。从表 1 中可以看出，本文提出的方法在交并比和不同阈值下的姿态估计准确度均有明显提升，表明本文方法在真实数据上有更好的泛化性能，显示了本文方法的优越性。同时，为了更加直观地展示实验结果，还提供了 4 组姿态估计的可视化结果，如图 3 所示，其中每组结果左侧为本文方法的结果，右侧为 SPD 方法的结果，蓝色框表示预测有向包围框，绿色框表示真实值有向包围框。深色坐标系表示预测姿态坐标系，浅色坐标系表示真实值姿态坐标系。可以看出，相比于 SPD 方法，本

文提出的方法预测的有向包围框更加紧致（比如第一组结果中的笔记本电脑和第三组结果中的笔记本电脑的有向包围框），预测的姿态也更加准确（比如第二组结果中左下角的杯子和第四组结果中的杯子和笔记本）。

表 1 本文方法与 NOCS 方法、SPD 方法的姿态估计精度

方法	3D ₅₀	3D ₇₅	5°2 cm	5°5 cm	10°2 cm	10°5 cm
NOCS	78.0%	30.1%	7.2%	10.0%	13.8%	25.2%
SPD	70%	37.9%	11.2%	12.9%	32.2%	41.3%
本文方法	61.2%	42.6%	23.3%	29.7%	41.8%	58.1%

4 结束语

本文提出了一种基于深度图点云的类别别物体姿态估计方法，仅使用虚拟仿真数据集模拟结构光深度相机，并用此进行深度学习模型训练。该方法在没有使用真实标注数据集参与训练的情况下，在真实数据集上表现出了优秀的姿态估计精度。在未来的工作中，将进一步提升姿态估计的精度。

参考文献:

- [1] HINTERSTOISSER S, HOLZER S, CAGNIART C, et al. Multimodal templates for real-time detection of textureless objects in heavily cluttered scenes[C]//Proceedings of 2011 International Conference on Computer Vision. [S.l.:s.n.], 2011.
- [2] HINTERSTOISSER S, CAGNIART C, ILIC S, et al. Gradient response maps for real-time detection of textureless objects[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2012, 34(5): 876-888.
- [3] HINTERSTOISSER S, LEPETIT V, ILIC S, et al. Model based train-

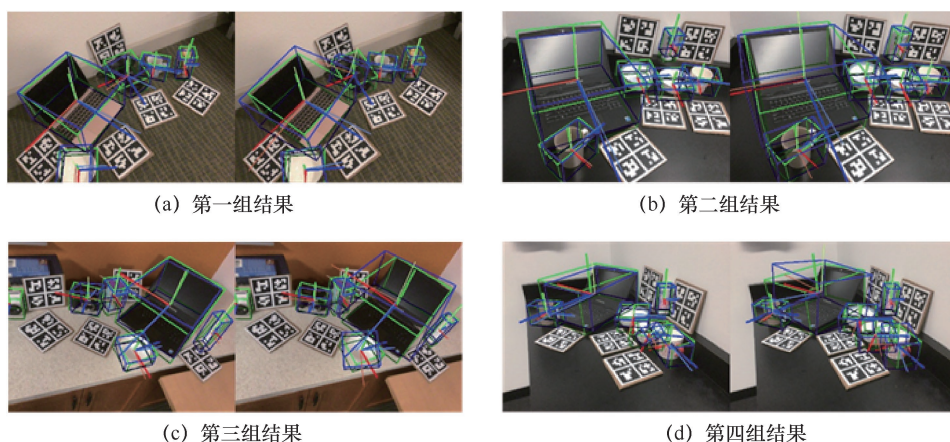


图 3 本文方法与 SPD 方法可视化结果

- ing, detection and pose estimation of texture-less 3D objects in heavily cluttered scenes[C]//Proceedings of the Asian Conference on Computer Vision. [S.l.:s.n.], 2013.
- [4] DROST B, ULRICH M, NAVAB N, et al. Model globally, match locally: efficient and robust 3D object recognition[C]//Proceedings of 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2010: 998-1005.
 - [5] CHOI C, CHRISTENSEN H I. 3D pose estimation of daily objects using an RGB-D camera[C]//Proceedings of 2012 IEEE/RSJ International Conference on Intelligent Robots and Systems. Piscataway: IEEE Press, 2012: 3342-3349.
 - [6] CHOI C, TREVOR A J B, CHRISTENSEN H I. RGB-D edge detection and edge-based registration[C]//Proceedings of 2013 IEEE/RSJ International Conference on Intelligent Robots and Systems. Piscataway: IEEE Press, 2013: 1568-1575.
 - [7] VIDAL J, LIN C Y, LLADÓ X, et al. A method for 6D pose estimation of free-form rigid objects using point pair features on range data[J]. *Sensors*, 2018, 18(8): 2678.
 - [8] GUO J W, XING X J, QUAN W Z, et al. Efficient center voting for object detection and 6D pose estimation in 3D point cloud[J]. *IEEE Transactions on Image Processing: a Publication of the IEEE Signal Processing Society*, 2021, 30: 5072-5084.
 - [9] XIANG Y, SCHMIDT T, NARAYANAN V, et al. PoseCNN: a convolutional neural network for 6D object pose estimation in cluttered scenes[C]//Proceedings of Robotics: Science and Systems XIV. Robotics: Science and Systems Foundation, 2018.
 - [10] LI Y, WANG G, JI X Y, et al. DeepIM: deep iterative matching for 6D pose estimation[J]. *International Journal of Computer Vision*, 2020, 128(3): 657-678.
 - [11] XU D F, ANGUELOV D, JAIN A. PointFusion: deep sensor fusion for 3D bounding box estimation[C]//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2018: 244-253.
 - [12] WANG C, XU D F, ZHU Y K, et al. DenseFusion: 6D object pose estimation by iterative dense fusion[C]//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2019: 3338-3347.
 - [13] HE Y S, SUN W, HUANG H B, et al. PVN3D: a deep point-wise 3D keypoints voting network for 6DoF pose estimation[C]//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2020: 11629-11638.
 - [14] HE Y S, HUANG H B, FAN H Q, et al. FFB6D: a full flow bidirectional fusion network for 6D pose estimation[C]//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2021: 3002-3012.
 - [15] ARUN K S, HUANG T S, BLOSTEIN S D. Least-squares fitting of two 3-D point sets[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1987, 9(5): 698-700.
 - [16] LEPETIT V, MORENO-NOGUER F, FUA P. EPnP: an accurate O(n) solution to the PnP problem[J]. *International Journal of Computer Vision*, 2008, 81(2): 155-166.
 - [17] BESL P J, MCKAY N D. Method for registration of 3-D shapes[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1992, 14(2): 239-256.
 - [18] LABBÉ Y, CARPENTIER J, AUBRY M, et al. CosyPose: consistent multi-view multi-object 6D pose estimation[J]. *arXiv preprint*, 2020, arXiv:2008.08465.
 - [19] PARK K, PATTEN T, VINCZE M. Pix2Pose: pixel-wise coordinate regression of objects for 6D pose estimation[C]//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE Press, 2019: 7667-7676.
 - [20] LI Z G, WANG G, JI X Y. CDPN: coordinates-based disentangled pose network for real-time RGB-based 6-DoF object pose estimation[C]//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE Press, 2019: 7677-7686.
 - [21] KÖNIG R, DROST B. A hybrid approach for 6DoF pose estimation[J]. *arXiv preprint*, 2020, arXiv:2011.05669.
 - [22] HODAN T, BARÁTH D, MATAS J. EPOS: estimating 6D pose of objects with symmetries[C]//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2020: 11700-11709.
 - [23] HAGELSKJÆR F, BUCH A G. PointVoteNet: accurate object detection and 6 DOF pose estimation in point clouds[C]//Proceedings of 2020 IEEE International Conference on Image Processing. Piscataway: IEEE Press, 2020: 2641-2645.
 - [24] SUNDERMEYER M, DURNER M, PUANG E Y, et al. Multi-path learning for object pose estimation across domains[C]//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2020: 13913-13922.
 - [25] SAHIN C, KIM T K. Category-level 6D object pose recovery in depth images[C]//Proceedings of the European Conference on Computer Vision. Cham: Springer, 2019.
 - [26] WANG H, SRIDHAR S, HUANG J W, et al. Normalized object coordinate space for category-level 6D object pose and size estimation[C]//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2019: 2637-2646.
 - [27] UMEYAMA S. Least-squares estimation of transformation parameters between two point patterns[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1991, 13(4): 376-380.
 - [28] FISCHLER M A, BOLLES R C. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography[J]. *Communications of the ACM*, 1981, 24(6): 381-395.
 - [29] CHEN D S, LI J, WANG Z, et al. Learning canonical shape space for category-level 6D object pose and size estimation[C]//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2020: 11970-11979.
 - [30] AN J, CHO S. Variational autoencoder based anomaly detection using reconstruction probability[J]. *Special Lecture on IE*, 2015, 2(1): 1-18.

- [31] TIAN M, ANG M H, LEE G H. Shape prior deformation for categorical 6D object pose and size estimation[C]//Proceedings of the European Conference on Computer Vision. Cham: Springer, 2020: 530-546.
- [32] CHARLES R Q, HAO S, MO K C, et al. PointNet: deep learning on point sets for 3D classification and segmentation[C]//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2017: 77-85.



高林（1987- ），男，博士，中国科学院计算技术研究所副研究员，主要研究方向为计算机图形学和几何处理。

[作者简介]



栗仁武（1998- ），男，中国科学院计算技术研究所、中国科学院大学硕士生，主要研究方向为点云处理和几何深度学习。



李淳芑（1980- ），男，博士，中国科学院计算技术研究所副研究员，主要研究方向为虚拟现实、人机交互和计算机图形学。



张凌霄（1995- ），男，中国科学院计算技术研究所助理工程师，主要研究方向为计算机图形学和几何处理。



蒋浩（1982- ），男，博士，中国科学院计算技术研究所副研究员，主要研究方向为群体行为模拟。