

无抓取标注的堆叠物体抓取位姿检测方法

石敏¹⁾, 侯京召¹⁾, 朱登明^{2)*}, 李兆歆^{3,4)}, 庞家繁¹⁾, 郭诗盛¹⁾

¹⁾ (华北电力大学控制与计算机工程学院 北京 102206)

²⁾ (中国科学院计算技术研究所前瞻研究实验室 北京 100190)

³⁾ (中国农业科学院农业信息研究所农业信息技术研究室 北京 100081)

⁴⁾ (农业农村部农业大数据重点实验室 北京 100081)

(mdzhu@ict.ac.cn)

摘 要: 针对现有的抓取检测网络需要大量标注数据进行训练且难以适应新物体抓取检测的问题, 提出一种适用于堆叠场景下的物体 6DoF 抓取位姿检测方法. 该方法由模型可抓取位姿模板库、待抓物体选择网络和抓取映射 3 部分组成. 首先基于几何外观特征与力封闭原理, 在物体模型上生成一组满足力封闭的 6DoF 抓取位姿, 构建可抓取位姿模板库; 然后分割原始场景点云, 基于可见性、遮挡性和置信度等确定待抓物体; 接着根据待抓物体在场景中的位姿, 将模板库中的抓取位姿映射到该物体上; 最后选择与待抓物体质心距离最近且无碰撞 6DoF 抓取位姿, 实现稳定抓取. 实验在 6 种类型物体堆叠场景中进行, 结果表明, 该方法取得 96.2% 的平均抓取成功率; 相较于 PointNetGPD, 其抓取成功率提升 5~20 个百分点.

关键词: 6DoF 抓取检测; 无抓取标注; 堆叠场景; 机械臂抓取; 位姿检测

中图法分类号: TP391.41 **DOI:** 10.3724/SP.J.1089.2023-00691

No-Label Stacked Object Grasp Pose Detection Algorithm

Shi Min¹⁾, Hou Jingzhao¹⁾, Zhu Dengming^{2)*}, Li Zhaoxin^{3,4)}, Pang Jiafan¹⁾, Guo Shisheng¹⁾

¹⁾ (School of Control and Computer Engineering, North China Electric Power University, Beijing 102206)

²⁾ (Foresight Research Laboratory, Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190)

³⁾ (Agricultural Information Technology Research Laboratory, Chinese Academy of Agricultural Sciences, Beijing 100081)

⁴⁾ (Key Laboratory of Agricultural Big Data, Ministry of Agriculture and Rural Affairs, Beijing 100081)

Abstract: Aiming at the problem that existing grasping detection networks require a large amount of labeled data for training and are difficult to adapt to new object grasping detection, a 6DoF grasping position detection method is proposed for stacked scenes. The method consists of a library of model grasping pose templates, an object selection network, and a grasping mapping. Based on the geometric appearance features and the force closure principle, a set of 6DoF grasping poses satisfying the force closure are generated on the object model to construct a grasping pose template library; then, the original field attraction cloud is segmented, and the objects to be grasped are identified based on the visibility, occlusion, and confidence; then, based on the pose of the objects to be grasped in the scene, the grasping poses in the template library are mapped onto the objects; finally, the nearest and non-colliding 6DoF grasping poses are selected and mapped with the objects to be grasped; and the nearest center-of-mass and non-colliding 6DoF grasping pose is se-

收稿日期: 2023-10-21; 修回日期: 2024-11-26. 基金项目: 国家重点研发计划(2020YFB1710400); 国家自然科学基金(61972379); 山东省新旧动能转换重大产业攻关项目(2021-55); 2023 年苏州市前沿技术研究项目“基于数字孪生技术的元宇宙数字影棚内容创作平台研发”. 石敏(1975—), 女, 博士, 副教授, 硕士生导师, CCF 会员, 主要研究方向为虚拟现实; 侯京召(1999—), 男, 硕士研究生, 主要研究方向为抓取检测; 朱登明(1973—), 男, 博士, 副研究员, 硕士生导师, CCF 会员, 论文通信作者, 主要研究方向为虚拟现实、计算机图形学; 李兆歆(1983—), 男, 博士, 助理研究员, 主要研究方向为三维计算机视觉; 庞家繁(2001—), 男, 硕士研究生, 主要研究方向为抓取检测; 郭诗盛(2001—), 男, 硕士研究生, 主要研究方向为抓取检测.

lected. Finally, the closest collision-free 6DoF grasping pose to the center of mass of the object to be grasped is selected to achieve stable grasping. The experiments are conducted in six types of object stacking scenes, and the results show that the method achieves an average success rate of 96.2%, which is 5–20 percentage points higher than that of PointNetGPD.

Key words: 6DoF grasping detection; no grasping annotation; stacked scene; robotic arm grasping; position detection

随着智能机械臂广泛应用于物体拣取、装配焊接、包装分拣等工业生产过程,其中的关键技术“抓取检测”^[1]已经成为计算机视觉领域的研究热点.在工业生产中,抓取场景中的物体多呈堆叠且无序排列,从该场景中选择出合适的待抓物体并生成无碰撞的抓取位姿,是一项具有挑战性的任务.

早期研究主要是基于二维平面的抓取检测^[2-6],通过数学分析、深度学习等方法评估抓取接触点或者抓取矩形的质量,得到最优的抓取位姿.虽然此类方法的模型简单、检测速度快,但存在二维平面的抓取方向固定,缺乏灵活性的问题.

针对二维平面抓取检测存在的问题,学者提出6自由度(6 degree of freedom, 6DoF)抓取检测.由于6DoF抓取检测方向任意,具有更强的灵活性和可扩展性,因此成为目前主流的抓取检测方法.

6DoF抓取检测大致分为2类:(1)基于分类的抓取检测方法^[7-9].首先对大量候选抓取样本进行采样,然后利用各种方法评估抓取质量,即根据一些评价指标对候选抓取位姿进行分类,选择质量最高的抓取位姿;(2)基于回归的抓取检测方法^[10-14].通过含有大量抓取标注的数据集训练网络模型,利用网络模型直接预测抓取位姿.前者虽然适用于未知物体的抓取,但并不能很好地应用于堆叠场景(包含多个无序、堆叠物体的场景)的抓取检测;后者虽然在堆叠场景上实现了较高的抓取成功率,但它需要大量含有抓取标注的数据集训练网络.

本文提出了一种无需抓取标注的6DoF抓取检测方法,通过抓取位姿模板库构建、待抓物体选择以及抓取位姿映射,有效地预测堆叠场景中的物体6DoF抓取位姿.首先,扩展了IPA-Bin picking数据集,增加了遮挡标注,构建了无抓取标注的抓取检测数据集G;然后,提出了一种模型可抓取位姿模板库构建方法,可以基于外观几何特征和力封闭原理自动生成一组均匀覆盖模型的6DoF抓取位姿;接着,针对堆叠场景提出了一种6DoF抓取检测方法,通过在网络中增加遮挡分支和可见性分支,并结合质心约束和碰撞检测,使得所预测的

抓取位姿具有更高的稳定性.

1 相关工作

1.1 二维平面抓取检测

通常,二维平面抓取检测通过数学分析、深度学习等方法评估抓取接触点或者抓取矩形的质量,选取最优的抓取位姿.DexNet 2.0^[2]首先从深度图中采样出多个候选抓取位姿,然后选用卷积神经网络(convolutional neural network, CNN)评估抓取位姿质量,选择最高质量的抓取位姿;Morrison等^[6]提出GGCNN,能够针对输入的深度图直接预测抓取质量、抓取宽度和抓取角度;Lenz等^[4]提出包含2个网络模块的抓取检测方法,第1个网络模块生成抓取矩形集合作为候选抓取,第2个网络模块基于RGBD图像提取的特征对这些候选抓取进行排序,选择排名最高的矩形作为最终抓取对象.二维平面抓取检测的网络相对简单,抓取方向是自上而下的,缺乏灵活性.

1.2 6DoF抓取检测

(1)基于分类的抓取检测.ten Pas等^[7]首先在点云表面生成大量6DoF候选抓取样本,然后基于抓取点的对拓概念二分类候选抓取样本,在抓取实验中取得了93%的抓取成功率;他们还提出GPD^[8],在感兴趣区域生成候选抓取后,将这些候选抓取编码到多通道图像中,并使用CNN评估每个候选抓取的分数.Liang等^[9]提出了PointNetGPD,通过对夹爪内部点云进行几何分析来评估抓取质量.基于分类的抓取检测方法适用于未知物体的抓取,具有良好的泛化性,但该方法无法很好地适用于堆叠场景.

(2)基于回归的抓取检测.Fang等^[10]提出GraspNet-1Billion数据集和一个抓取检测网络,该网络以编码-解码的方式直接回归出抓取位姿.Qin等^[11]提出以PointNet++^[15]为主干网络的抓取检测方法,首先回归出可能的抓取位姿,然后评估每个抓取位姿.Zhao等^[13]提出一个端到端的抓取检测网

络 REGNet, 包括得分网络(score network, SN)、抓取区域网络(grasp region network, GRN)和细化网络(refinement network, RN) 3 个模块, 分别用于选择高抓取置信度的点、基于 SN 模块所选的点生成抓取, 以及优化抓取. Ni 等^[14]提出的 PointNet++ Grasping 是一种端到端的方法, 可以直接预测所有抓取的位姿、类别和分数. Kleeberger 等^[12]也提出一种端到端的方法 PQ-Net, 可以直接回归出每个抓取位姿及其对应的抓取成功率, 选择抓取成功率最高的位姿作为输出. 基于回归的抓取检测方法需要大量已标注的抓取位姿数据进行网络训练, 如 YCB^[16], Cornell Grasping dataset^[17] 或者 GraspNet-1Billion 数据集; 其中, GraspNet-1Billion 数据集包含超过 10 亿个抓取位姿标注, 标注数据的时间成本很高.

2 本文方法

本文提出一种无抓取标注的 6DoF 抓取检测方法, 其抓取检测流程如图 1 所示. 给定相机捕

捉到的单视角点云 c , 本文方法的目标是生成 6DoF 抓取位姿. 类似于 PointNetGPD, 本文将 6DoF 抓取位姿定义为 $g = (c, r)$, 其中, $c = (x, y, z) \in \mathbb{R}^3$ 和 $r = (r, r, r) \in \mathbb{R}^3$ 分别表示抓取位置和抓取方向.

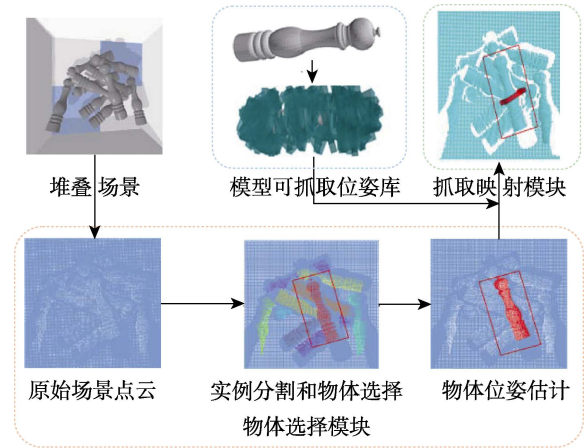


图 1 本文抓取检测流程

本文方法包括模型可抓取位姿模板库构建、物体选择和抓取映射 3 部分, 如图 2 所示.

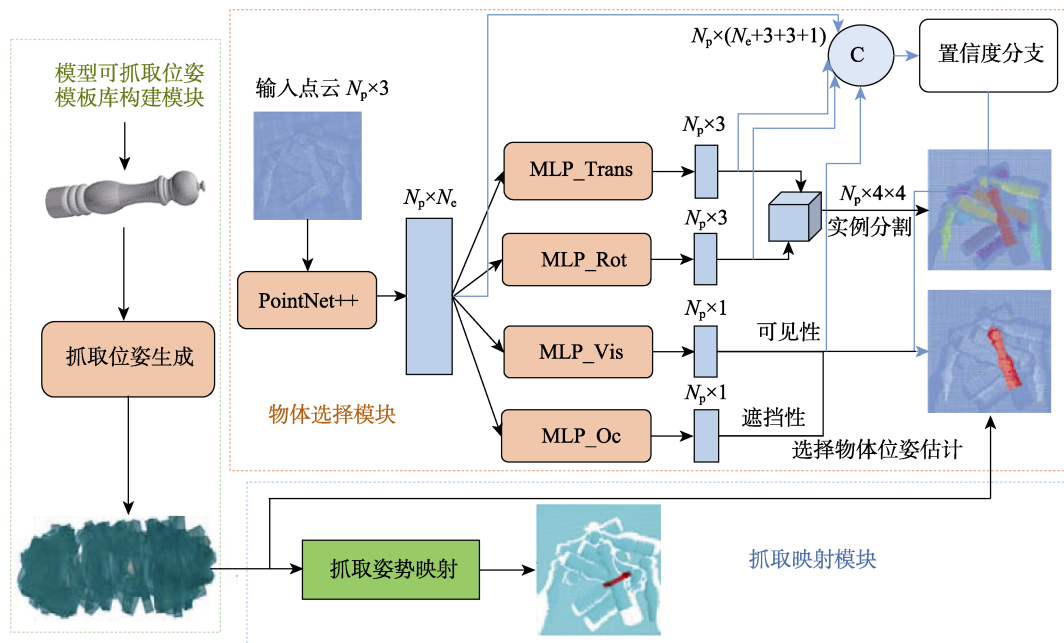


图 2 本文方法结构

2.1 可抓取位姿模板库构建

模型位姿模板库包含一组均匀覆盖在物体表面的抓取位姿, 其稳定将影响到最终堆叠场景中抓取的稳定性. 对于二指夹抓, 稳定的抓取位姿需要满足力封闭原理. 力封闭即二指夹爪在夹住待抓物体时, 夹爪 2 根手指会对物体产生 2 个闭合

力. 这 2 个闭合力形成 2 个摩擦锥. 如图 3 所示, 当两者合力(图中连线)处于 2 个摩擦锥内时, 则满足力封闭; 否则, 不满足力封闭.

基于此, 本文对物体表面采样得到的抓取位姿进行力封闭判断, 最终选择一组满足力封闭的 6DoF 抓取位姿, 该组抓取位姿均匀地覆盖模型表面.

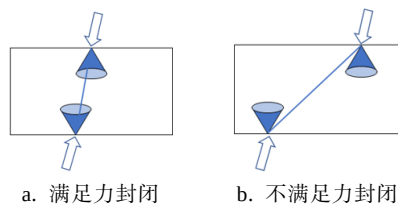


图3 力封闭示意图

除了满足力封闭, 为了验证抓取位姿的碰撞性, 还要对二指夹爪进行建模. 如图4所示, 2个平行的长方体用于表示夹爪的2根手指, 二指夹爪模型 m 可以由一组参数表示 $m = \{m_l, m_w, m_t, m_d\}$, 其中, m_l 和 m_w 表示手指的长度和宽度, m_t 表示手指的厚度, m_d 表示夹爪完全打开时手指之间的距离.

模型可抓取位姿模板库构建模块以物体模型 M 和夹爪模型 m 作为输入, 输出是一组 6DoF 抓取位姿集合 G' . 首先在模型表面采样一个接触点 c_1 , 并根据模型表面在该点的切向量计算抓取闭合方向 D_{close} , 计算得到模型表面另一个接触点 c_2 ; 然后根据 c_1 和 c_2 计算抓取方向 $R = R_0 \times R_\theta$,

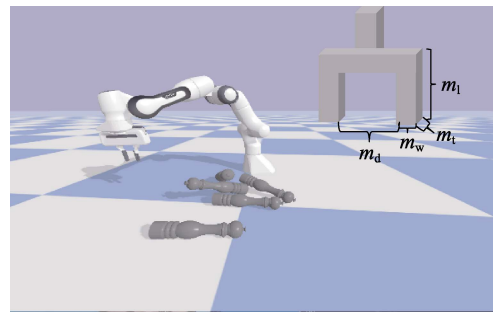


图4 建立夹爪模型

并判断 c_1 和 c_2 是否力封闭, 在 c_1 和 c_2 满足力封闭的情况下, 让夹爪沿着抓取闭合方向 D_{close} 闭合, 与模型 M 的 2 个接触点分别为 $P_{contact1}$ 和 $P_{contact2}$; 然后在它们满足力封闭情况下, 计算抓取位姿 g' , 其抓取位置为 P_{center} , 抓取方向矩阵为 R , 若抓取中心 g' 与模板库 G' 中每个抓取位姿中心的距离大于阈值 d , 则将该抓取位姿 g' 加入到模板库中. 最后迭代以上步骤, 直到 G' 中的抓取数量达到 N 个. 该方法的具体步骤如图5所示.

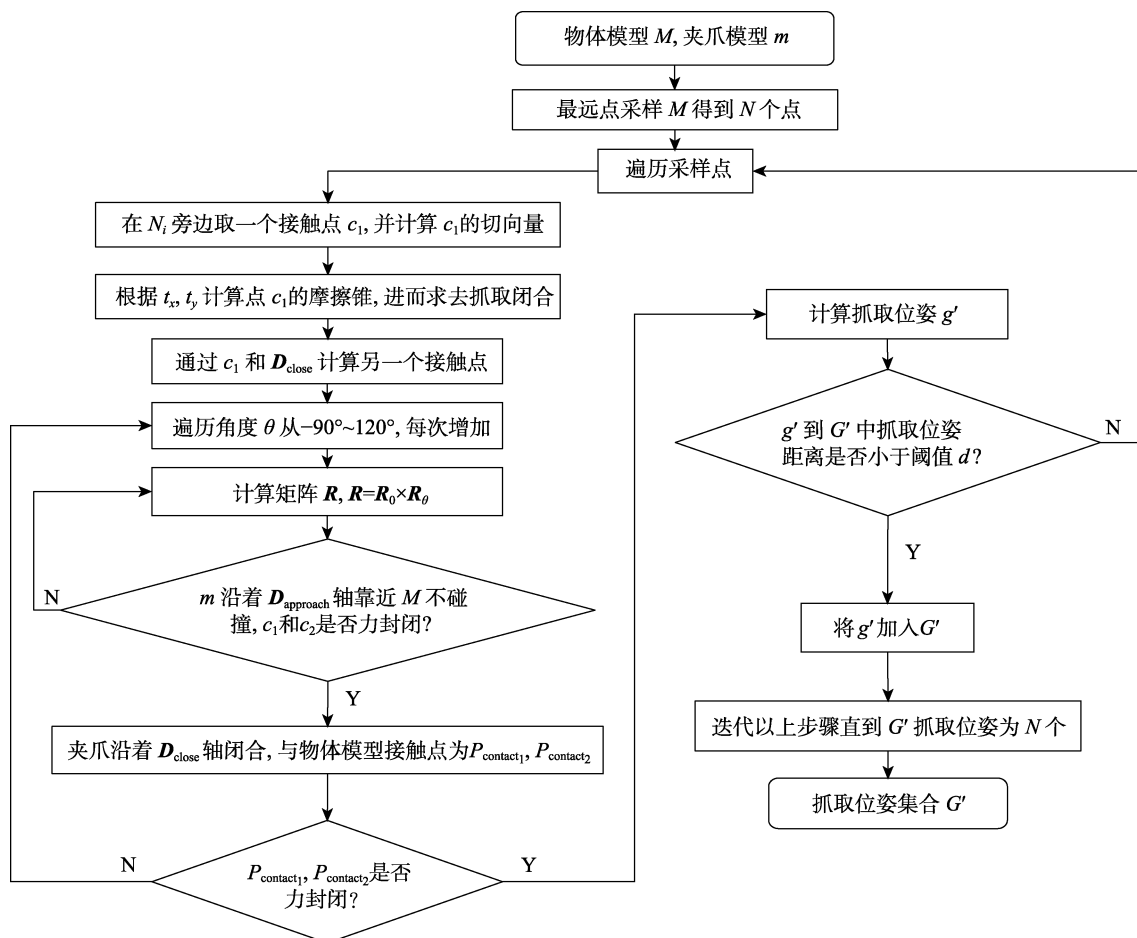


图5 模型可抓取位姿模板库构建流程

图 5 中, $\mathbf{R}_0 = [\mathbf{D}_{\text{approach}}, \mathbf{D}_{\text{close}}, \mathbf{V}_{\text{axis}}]$. 其中,
 $\mathbf{D}_{\text{approach}_x} = \mathbf{D}_{\text{close}_y}, \mathbf{D}_{\text{approach}_y} = -\mathbf{D}_{\text{close}_x}, \mathbf{D}_{\text{approach}_z} = 0,$
 $\mathbf{V}_{\text{axis}} = \mathbf{D}_{\text{approach}} \times \mathbf{D}_{\text{close}},$

$$\mathbf{R}_\theta = \begin{bmatrix} \cos(\theta) & 0 & \sin(\theta) \\ 0 & 1 & 0 \\ -\sin(\theta) & 0 & \cos(\theta) \end{bmatrix},$$

$$\mathbf{P}_{\text{center}} = (\mathbf{P}_{\text{contact}_1} + \mathbf{P}_{\text{contact}_2}) / 2 - \mathbf{R}_0 \times \mathbf{m}_1.$$

2.2 物体选择

分析多物体堆叠场景可知, 场景中的物体相互堆叠; 上层物体在相机视野里往往暴露的部分较多, 可见性更高; 底部物体由于被上层物体遮挡, 在相机视野里往往只暴露出部分, 因此可见性较低. 可见性是指场景点云中属于该物体的点数占该物体完全暴露在相机视野所拥有有点数的比例; 其取值为[0,1], 可见性越高, 值越大.

对于无序堆叠场景, 如果选择的待抓物体处于堆的上方, 可见性高且无遮挡, 则该物体更易于抓取, 且抓取时不与场景中剩余物体发生碰撞, 抓取更加稳定. 因此, 本文需要对场景各物体进行实例分割, 并预测其可见性、遮挡性等信息, 以选择合适的待抓物体. 如图 2 所示, 抓取物体选择包含实例分割、可见性预测、遮挡性预测和置信度预测任务. 该模块的输入是一个大小为 $N_p \times 3$ 的堆叠场景点云, 本文使用 PointNet++ 作为特征提取网络. PointNet++ 已广泛应用于点云分割、位姿估计等任务, 它在提取全局特征的同时可保证局部特征的提取. 通过 PointNet++ 提取得到 $N_p \times N_e$ 维特征, 将该特征共享到 5 个多层感知机(multi-layer perceptron, MLP)分支, 预测 N_p 个点的平移、旋转、可见性、遮挡性和置信度, 表示为 MLP_Trans, MLP_Rot, MLP_Vis 和 MLP_Oc. 网络总损失 L 由每个分支损失的加权和确定, 包括实例分割分支 L_{is} 、可见性预测分支 L_v 、遮挡预测分支 L_o 和置信度预测分支 L_{con} , 即

$$L = L_{\text{is}} + \alpha L_v + \beta L_o + \lambda L_{\text{con}}.$$

其中, α, β, λ 表示比例常数.

(1) L_v . 在堆叠场景中, 位于底部的物体曝光在相机视野里的部分比较少, 在生成的场景点云中表现为属于该物体的点数目比较少. 可见性低的物体不利于抓取, 而且由于点数较少, 包含的特征也较少, 不利于网络的学习过程. 因此, 本文在网络中添加 L_v 来降低底部物体的影响, 增加物体实例分割准确率. 通过 MLP_Vis 共享 $N_p \times N_e$

维的特征, 逐点回归可见性. L_v 的计算公式为

$$L_v = \sum \|V_{\text{vis}_i}^{\text{pred}} - V_{\text{vis}_i}^{\text{gt}}\|^2.$$

其中, $V_{\text{vis}_i}^{\text{pred}}$ 表示第 i 个点可见性的预测值, $V_{\text{vis}_i}^{\text{gt}}$ 表示第 i 个点可见性真实值.

(2) L_{is} . 将 PointNet++ 提取到的 $N_p \times N_e$ 特征分别输入到 MLP_Trans 和 MLP_Rot. 其中, MLP_Trans 用于回归每个点所属对象的中心位置, MLP_Rot 用于回归每个点所属对象的旋转, 该旋转以欧拉角为单位. 组合 MLP_Trans 的输出 $T_{\text{MLP}}^{\text{pred}}(N_p \times 3)$ 和 MLP_Rot 的输出 $R_{\text{MLP}}^{\text{pred}}(N_p \times 3)$ 为 $N_p \times 4 \times 4$ 的旋转平移变换 $\mathbf{T}$, 即

$$\mathbf{T} = T_{\text{MLP}}^{\text{pred}}(N_p \times 3) + R_{\text{MLP}}^{\text{pred}}(N_p \times 3).$$

为了更好地进行损失计算, 将 $\mathbf{T}$ 映射到欧几里得空间中, 表示为 $R(\mathbf{T})$, 则可以通过欧几里得距离衡量 2 个变换之间的差异. L_{is} 的计算为

$$L_{\text{is}} = \sum v_i^{\text{gt}} \times D_{\text{distance}}(R(\mathbf{T}_{\text{pred}}), R(\mathbf{T}_{\text{gt}})).$$

为了降低物体可见性低、特征少对网络学习过程的负面影响, 选择可见性作为该损失权重, D_{distance} 的计算公式为

$$D_{\text{distance}}(R(\mathbf{T}_1), R(\mathbf{T}_2)) = \min_{t_1 \in R(\mathbf{T}_1), t_2 \in R(\mathbf{T}_2)} \|t_1 - t_2\|.$$

受到 PPR-Net++^[18] 的聚类启发, 本文认为, 属于同一个物体的点回归预测的 $T_{\text{MLP}}(N_p \times 3)$ 相近, 对其进行聚类并求均值, 可以得到多个不同的点簇, 每个点簇对应一个实例分割后的物体.

(3) L_o . 堆叠场景中, 可见性高的物体并不一定比低的物体更利于抓取, 如图 6 所示, 物体 1 的可见性比物体 2 低, 但由于其没有遮挡, 物体在抓取时不会与场景中其他物体发生碰撞, 抓取更稳定. 在满足不遮挡的前提下, 本文优先选择可见性高的物体作为待抓物体. 通过 MLP_Oc 共享 $N_p \times N_e$ 维的特征, 逐点回归遮挡性. L_o 的计算公式为

$$L_o = -\frac{1}{n} \sum [o_i^{\text{gt}} \ln o_i^{\text{pred}} + (1 - o_i^{\text{gt}}) \ln (1 - o_i^{\text{pred}})].$$

(4) L_{con} . 采用聚类的方式估计待抓物体位姿. 由经验可知, 参与聚类的点的平移和旋转误差越小, 聚类的结果越准确, 位姿估计也会更准确, 因此, 本文引入置信度机制. 如图 7 所示, 置信度分支由一个 MLP 和 Softmax 组成, 置信度分支的输入为 $N_p \times (N_e + 3 + 3 + 1)$. 其中, N_e 表示 PointNet++ 提取到的特征维度; 3, 3, 1 分别表示 $T_{\text{MLP}}^{\text{pred}}$, $R_{\text{MLP}}^{\text{pred}}$ 和

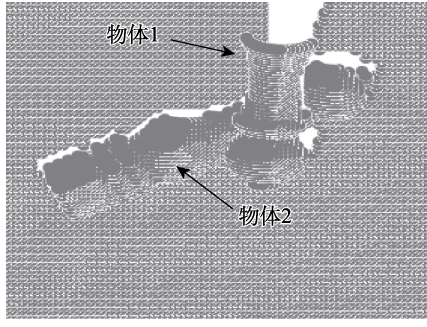


图6 堆叠物体示意图

$V_{\text{vis}}^{\text{pred}}$ 的维度. 输出是 $N_p \times 2$ 的二分类向量 $\mathbf{p}$, 每个点的 $p_i = [p_{i_cls0}, p_{i_cls1}]$, $p_{i_cls0} \in [0,1]$, $p_{i_cls1} \in [0,1]$, p_{i_cls1} 表示正样本概率. 令 $C_{\text{con}_i}^{\text{pred}} = p_{i_cls1} \in [0,1]$, 则

$$L_{\text{con}} = -\frac{1}{n} \sum \left[x_i \ln C_{\text{con}_i}^{\text{pred}} + (1-x_i) \ln (1-C_{\text{con}_i}^{\text{pred}}) \right].$$

其中, 各个点标注 $x_i = t_i \& r_i$. t_i, r_i 取值分别为

$$\begin{cases} t_i = \begin{cases} 1, e_{ti} < d_t \\ 0, e_{ti} \geq d_t \end{cases} \\ r_i = \begin{cases} 1, e_{ri} < d_r \\ 0, e_{ri} \geq d_r \end{cases} \end{cases}.$$

其中, e_{ti} , e_{ri} , d_t 和 d_r 分别表示平移误差、旋转误差、平移误差阈值和旋转误差阈值.

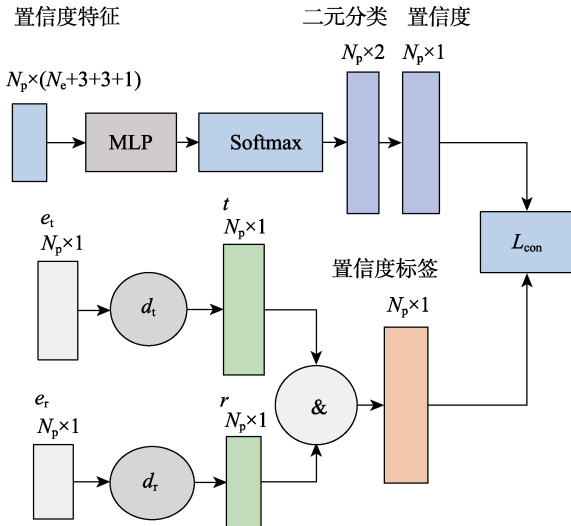


图7 置信度分支结构

根据 L_{con} 可知, $C_{\text{con}_i}^{\text{pred}}$ 用来学习 x_i , 点的置信度越高, 该点的平移预测误差和旋转预测误差就越小, 位姿估计的精度也会越高.

物体选择. 设 O_i 表示实例分割后的第 i 个物体, O_{0_i} 表示第 i 个物体的遮挡性, O_{v_i} 表示第 i 个

物体所属点可见性的均值, O_{t_i} 表示第 i 个物体所属点 T_{trans} 的 Z 方向值的均值. 本文设计一种物体抓取优先度策略为

$$O_{\text{score}_i} = a \times O_{0_i} + b \times O_{v_i} + c \times \exp(-O_{t_i}).$$

其中, a, b, c 表示比例常数.

对 O_{score} 从大到小排序, 并作为抓取顺序, 在满足能生成无碰撞 6DoF 抓取位姿的条件下, 选择 O_{score} 最大的物体作为待抓物体. 由置信度分支可知, 预测的平移误差和旋转误差越小, 置信度越高, 为了提高待抓物体位姿估计的精准度, 选择待抓物体置信度最高的 150 个点的旋转均值 $\bar{\mathbf{R}}_{\text{rot}}$, 平移均值 $\bar{\mathbf{T}}_{\text{select}}$ 作为待抓物体的旋转平移矩阵 $\mathbf{R}_{\text{select}}$, 即

$$\mathbf{R}_{\text{select}} = \begin{bmatrix} \bar{\mathbf{R}}_{\text{rot}} & \bar{\mathbf{T}}_{\text{select}} \\ 0 & 1 \end{bmatrix}.$$

2.3 抓取映射

由于 6DoF 抓取检测得到的是场景中的抓取位姿, 因此需要将模型上的抓取位姿变换到场景待抓物体上. 为了保证抓取的稳定性, 在选择抓取位姿时需要考虑质心约束和碰撞约束. 质心约束指抓取位姿应尽可能靠近质心, 抓取位姿距离物体的质心越近, 那么在物体受到外力(如重力)作用时产生的力矩就越小, 而较小的力矩意味着物体更易保持平衡, 抓取更稳定. 碰撞约束指夹爪在该抓取位姿不与场景中物体发生碰撞, 当与场景发生碰撞时会扰乱场景, 导致待抓物体位姿偏移, 抓取失败. 为此, 在设计抓取映射方法时, 考虑质心约束和碰撞约束. 首先将模型库中的抓取位姿逐一变换到场景中待抓物体上, 然后筛选出无碰撞且抓取趋近方向与 Z 轴夹角小于 30° 的抓取位姿, 最后选择筛选位姿中与质心距离最近的抓取位姿作为输出. 其中, S_C 表示场景点云映射方法步骤如下.

输入. 待抓物体位姿 $\mathbf{R}_{\text{select}}$, 模型抓取位姿 G' .

输出. 6DoF 抓取位姿 G_{out} .

$G_{\text{grasp}} = \emptyset$.

for $i = 1$ to $\text{len}(G')$ do

$G_{\text{grasp}'} = G'_i \times \mathbf{R}_{\text{select}}$;

if $O(G_{\text{grasp}'} \cap S_C \neq \emptyset)$ and $G_{\text{grasp}'}$ 的趋近方向与 Z 轴的夹角小于 30°

$G_{\text{grasp}} = G_{\text{grasp}'} \cup G_{\text{grasp}}$.

end for

遍历 G_{grasp} , 选择其中与待抓物体质心距离最近的抓取位姿 G_{grasp} 作为输出, 即指抓取位姿中心到质心的距离最小.

3 实验与结果分析

3.1 数据集构建

IPA-Binpicking 数据集^[19]是实例分割、位姿估计公开数据集. 因为无遮挡物体抓取更稳定, 所以在选择待抓物体时需要保证其无遮挡, 而 IPA-Binpicking 数据集不包含遮挡标注, 无法用来预测遮挡信息. 为此, 本文在该数据集的基础上构造了数据集, 用于对网络进行训练, 其中包含 Pepper, GearShaft, Tless29, Bunny, Tless20 和 Gear 这 6 种物体, 每个物体有 1200 个堆叠场景, 训练集和测试集划分为 4:1.

为了降低对计算机硬件的需求, 生成点云时采用均匀降采样的方法, 从深度图每 4 个像素点中取一个像素点, 然后还原成点云. 以 GearShaft 的场景深度图为例, 其深度图尺寸为 512×512 像素, 采用均匀降采样方法后, 得到的点云包含的点数为 $(512/4) \times (512/4) = 16384$. 每个场景的标注信息如下: (1) p_c . 通过 IPA-Binpicking 数据集场景深度图均匀降采样后生成的三维点云; (2) T_{trans} . IPA-Binpicking 数据集场景中物体对应的三维平移; (3) R_{rot} . IPA-Binpicking 数据集场景中物体对应的三维旋转; (4) V_{vis} . IPA-Binpicking 数据集场景中物体的可见性度量. (5) O_c . 场景中物体的遮挡性, O_c 的计算公式为

$$O_c = \begin{cases} 0, & \text{if 物体被遮挡} \\ 1, & \text{else} \end{cases}.$$

3.2 实验场景构建

本文使用 Pybullet 仿真平台模拟工业应用中的一种常见设置: 将机械臂 Panda 固定在地面上, 抓取器设置为二指夹爪, 摄像机固定在场景正上方. 选择 6 种物体, 通过将物体从固定高度任意方向自由落向地面, 创建无序堆叠的场景, 并获取场景点云, 如图 8 所示. 实验中, 所有物体都是密度均匀的刚体, 它们的质量设置为 1 kg, 物体表面的摩擦系数为 0.6, 夹爪最大夹持力为 30 N. 由于 PointNetGPD 适用于场景包含少量物体的抓取, 为了验证其有效性, 为每个物体搭建 2 种实验场景:

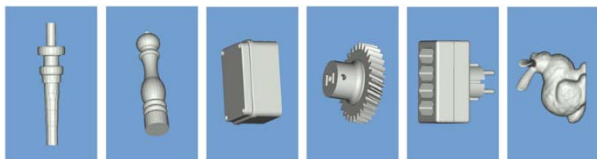


图 8 实验所用 6 种物体

(1) 包含 1~3 个物体, 物体数目较少, 遮挡情况较为轻微; (2) 包含 4~10 个物体, 底部物体遮挡严重.

为每种物体创建了 300 个场景, 总共创建 1800 个场景, 其中, 1200 个用于抓取实验, 600 个用于评估时间效率.

第 1 种场景实验中, 分别对 6 种物体进行 100 次抓取, 计算抓取成功率; 第 2 种场景实验中, 在 6 种物体上再次执行 100 次抓取, 计算抓取成功率. 任意方向自由下落保证每次抓取实验场景都是随机的、不同的, 6 种物体上 2 次抓取的实验场景如图 9 所示.

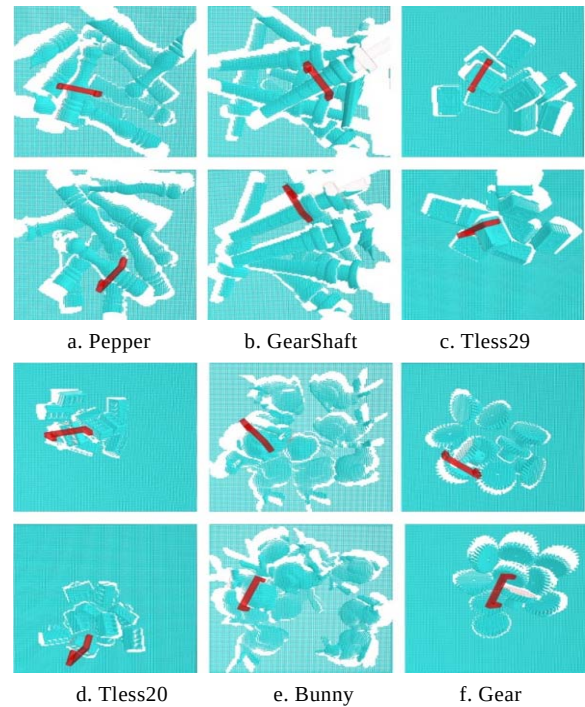


图 9 6 种类型物体 2 次抓取的实验场景

为了评估本文方法和 PointNetGPD 方法^[9]的时间效率, 分别对 6 种物体执行 100 次抓取, 其中, 第 1 种场景 50 次, 第 2 种场景 50 次, 并通过计算 100 次抓取检测的平均时间评估时间效率.

3.3 实验参数设置

本文实验中, 特征提取维度 $N_e = 128$, $\alpha = 70$, $\beta = 20$, $\lambda = 20$, $\gamma = 10$, $a = 0.6$, $b = 0.2$, $c = 0.2$. 平移误差阈值 $d_t = 0.005$, 旋转误差阈值 $d_r = 0.005$, 夹爪模型 $(m_l, m_w, m_d, m_t) = (0.05, 0.01, 0.10, 0.0018)$. 本文算法使用 Pytorch 实现, 物体选择模块使用 Adam optimizer 训练, batch size 大小为 4、初始学习率为 0.001, 学习率每 80 步衰减 0.7.

3.4 评价指标

采用与 PointNetGPD 方法一致的抓取成功率

g_{succed} 作为本文的评价指标, 计算公式为

$$g_{\text{succed}} = \frac{g_s}{g_s + g_f}$$

其中, g_s 表示抓取成功的次数, g_f 表示抓取失败的次数. 进行一次抓取实验时, 夹爪成功地抓起物体并移动到场景上方, 且物体不掉落, 记为抓取成功; 否则, 记为抓取失败.

3.5 实验

3.5.1 对比实验

在第1种场景中, 将本文方法与 PointNetGPD 方法^[9]的抓取成功率进行实验对比, 结果如表1所示. 可以看出: (1) 2种方法都有很高的抓取成功率. (2) 因为 PointNetGPD^[9]在选择抓取位姿时只考虑抓取位姿得分, 该得分与抓取位姿内部点云有关, 与抓取位姿到质心的距离无关, 距离质心远的抓取位姿可能比距离质心近的抓取位姿得分更高, 所以该方法会生成位于物体末端的抓取位姿; 而该位姿远离物体的质心, 抓取位姿距离物体的质心越远, 则在物体受到外力(如重力)作用时产生的力矩就越大, 不易保持平衡, 在夹爪移动过程物体会脱落, 导致抓取失败, 降低抓取成功率. (3) 本文方法选择抓取位姿时考虑质心距离约束, 选择离质心最近的无碰撞抓取位姿作为输出, 解决了 PointNetGPD^[9]中的问题. (4) 与 PointNetGPD^[9]相比, 本文方法的抓取成功率高 5~14 个百分点.

表1 在第1种场景中2种方法的抓取成功率对比 %

方法	Pepper	GearShaft	Tless29	Bunny	Tless20	Gear
Point-NetGPD ^[9]	94	90	88	84	83	85
本文	99	99	98	98	93	98

注. 粗体表示最优值.

在第2种场景中, 将本文方法与 PointNetGPD 方法^[9]的抓取成功率进行实验, 结果如表2所示. 可以看出: (1) PointNetGPD^[9]的抓取成功率均低于本文方法, 有2个原因: 一是该方法有时生成的抓取位姿位于物体的末端, 远离物体的质心, 抓取不稳定, 导致抓取失败; 二是该方法不能很好地应用于堆叠场景, 由于该场景中的物体往往遮挡严重, 它在生成抓取位姿时只考虑夹爪内部点云的质量, 未考虑待抓取的物体是否被其他物体遮挡或被场景中其他物体压住, 以及抓取时会与场景中的其他物体碰撞, 因此导致抓取失败. (2) 本文方法不仅可以有效地处理堆叠场景, 而且可以保证待抓取物体不被遮挡, 与 PointNetGPD^[9]相比, 本文方

表2 在第2种场景中2种方法的抓取成功率对比 %

方法	Pepper	GearShaft	Tless29	Bunny	Tless20	Gear
Point-NetGPD ^[9]	84	82	74	77	75	80
本文	99	98	94	97	91	98

注. 粗体表示最优值.

法的抓取成功率提高 15~20 个百分点.

综合分析表1和表2可以看出, 6种物体抓取成功率从高到低的排名顺序为 Pepper, GearShaft, Gear, Bunny, Tless29 和 Tless20. 对6种抓取对象的失败案例分析如下: (1) 对于 Pepper, GearShaft, Gear 和 Bunny, 抓取失败主要受对象选择模块的遮挡分支预测精度的影响, 偶尔会将被抓取对象的一端轻微遮挡视为未遮挡, 导致抓取失败. (2) 除了遮挡分支预测精度的影响外, Tless29 和 Tless20 还受到位姿估计预测精度的影响, 与夹爪尺寸相比, 由于这2个抓取对象的尺寸较小, 因此对位姿估计精度度有更高的要求, 当位姿估计精度误差较大时, 模型上的抓取位姿映射到场景中不再满足力封闭条件, 抓取不稳定, 导致抓取失败; 尽管如此, 本文方法依然在没有抓取位姿标注的 Tless29 堆叠场景上实现了 94% 的抓取成功率、在 Tless20 的堆叠场景上实现了 91% 的抓取成功率.

对本文方法与 PointNetGPD 方法^[9]的时间效率进行实验, 结果如表3所示. 可以看出, PointNetGPD^[9]的时间效率远低于本文方法, 该方法需要花费大量时间进行抓取位姿采样评估以生成最终的抓取位姿; 而本文方法直接回归出待抓物体, 然后通过模型可抓取位姿映射得到最终的抓取位姿, 无需抓取位姿采样评估, 节省了大量的时间.

表3 2种方法抓取检测时间对比 s

方法	Pepper	GearShaft	Tless29	Bunny	Tless20	Gear
Point-NetGPD ^[9]	3.381	3.553	4.415	3.634	4.862	4.363
本文	1.095	1.110	1.385	0.996	1.308	1.556

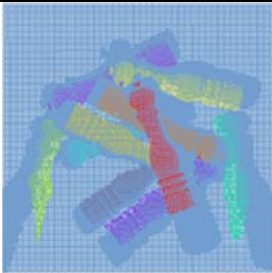
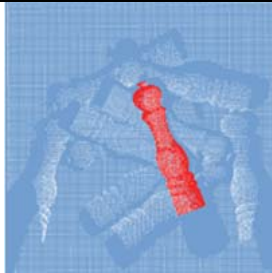
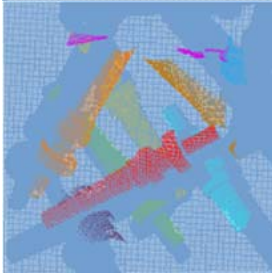
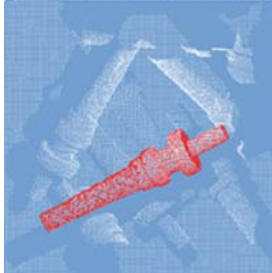
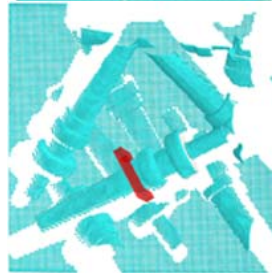
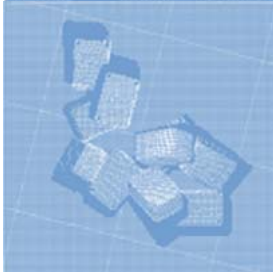
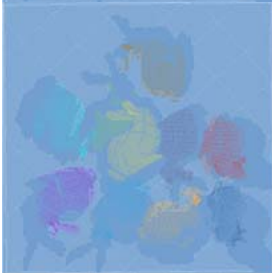
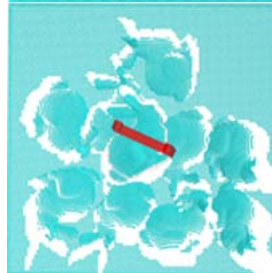
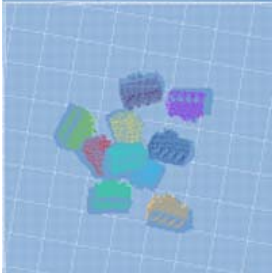
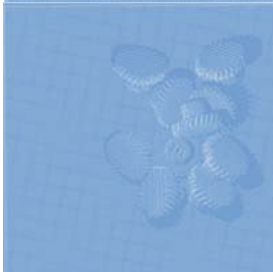
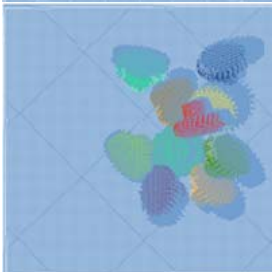
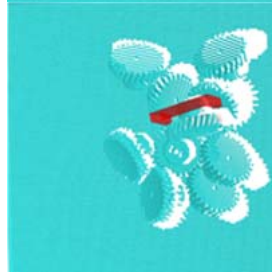
注. 粗体表示最优值.

表4所示为本文方法在第2种场景下的定性结果. 其中, 属于同一对象的点用相同的颜色表示, 红色物体表示物体选择模块选择出的待抓取物体.

3.5.2 消融实验

通过消融实验, 验证本文方法中各分支的效果. 以本文不加入各个分支的模型为 Baseline, 在2种场景下, 通过在 Baseline 上逐渐添加可见分支和遮挡分支进行消融实验. 其中, “v”表示可见分

表 4 在第 1 种场景中 6 种物体可视化结果

物体类别	场景点云	实例分割	待抓物体位姿估计	抓取位姿
Pepper				
GearShaft				
Tless29				
Bunny				
Tless20				
Gear				

支,“o”表示遮挡分支。
在第 1 种场景下的消融实验结果如表 5 所示。

可以看出,与 Baseline 相比,添加可见分支后,最大抓取成功率提高 3 个百分点;添加遮挡分支后,最大

抓取成功率提高 6 个百分点;添加可见分支和遮挡分支后(即本文方法),抓取成功率平均提高 4.8 个百分点,两者的提升效果可以叠加,证明了两者的有效性.由于第 1 种场景包含较少的物体,因此很少出现上层物体被下层物体遮挡的情况,Baseline 选择最上方物体作为待抓物体满足无遮挡、可见性高、位于场景上方,可以获得较高的抓取成功率.

在第 2 种场景下的消融实验结果如表 6 所示.可以看出,与 Baseline 相比,添加可见分支后,最大抓取成功率提高 6 个百分点;添加遮挡分支后,最大抓取成功率提高 8 个百分点;添加可见分支和遮挡分支后,抓取成功率平均提高 8.8 个百分点,两者的效果可以叠加.与第 1 种场景相比,第 2 种

场景包含的物体更多,顶部的物体可能被底部的物体遮挡,此时 Baseline 简单选择最上方物体作为待抓物体可能不满足无遮挡条件,夹爪夹起物体时可能会因为待抓物体一端被遮挡掉落,降低抓取成功率.与 Baseline 相比,添加遮挡分支后,可以保证待抓物体不被遮挡,抓取稳定,提高抓取成功率;同样,添加可见分支后,可以保证所选待抓物体在相机感知范围内具有更高的可见性,意味着待抓物体在场景点云中占据更多的点,网络学习到的特征更丰富,有利于待抓物体位姿估计,位姿估计结果越准确则映射到场景中的抓取位姿更稳定,提高了抓取成功率.同时,加入可见分支和遮挡分支可以保证抓取更稳定.

表 5 第 1 种场景下的消融实验结果对比

%

方法	Pepper	GearShaft	Tless29	Bunny	Tless20	Gear
Baseline	96	94	94	91	88	93
Baseline+v	97	97	96	93	89	95
Baseline+o	98	97	97	97	92	96
本文(Baseline+o+v)	99	99	98	98	93	98

注.粗体表示最优值

表 6 第 2 种场景下的消融实验结果对比

%

方法	Pepper	GearShaft	Tless29	Bunny	Tless20	Gear
Baseline	88	86	89	87	86	88
Baseline+v	94	92	90	92	88	91
Baseline+o	96	94	91	95	89	96
本文(Baseline+o+v)	99	98	94	97	91	98

注.粗体表示最优值.

3.5.3 讨论

本文方法中,模型可抓取位姿模板库构建方法基于外观几何特征和力封闭原理,可以在不同类型的物体模型上自动生成一组均匀覆盖模型的 6DoF 抓取位姿,适用于不同外观几何特征的物体抓取检测;在选择待抓物体时,能够保证被选择物体无遮挡、可见性高和位于场景上方,因此即使场景中物体数目增多,与遮挡物体相比,这些物体更容易被稳定抓取.但是,本文使用的夹爪为二指刚性夹爪,不适用于抓取草莓、西红柿等易受损的对象,而多指软夹爪适用于抓取此类物体.另外,本文方法虽然不需要在训练集中包含抓取位姿标注,但是为提高抓取稳定性,需要增加遮挡标注等.

4 结语

本文提出一种无需抓取标注 6DoF 抓取检测方

法,以原始场景点云作为输入,首先通过物体选择网络回归出得分最高的待抓物体位姿;然后将模型可抓取位姿模板库的 6DoF 抓取位姿映射到场景中;最后在满足无碰撞的条件下,选择离待抓物体质心最近抓取位姿作为输出.本文使用的数据集 G 是基于工业数据集 IPA-Binpicking 扩展得到的,在堆叠场景抓取实验中,与 PointNetGPD 相比,本文方法的抓取成功率提高 5~20 个百分点.

未来,将研究一种适用于多种夹爪类型的抓取检测方法,使得本文方法可以拓展到农业采摘任务中;另外,结合大模型,探索对未知物体的稳定、准确的抓取检测.

参考文献(References):

- [1] Du G G, Wang K, Lian S G, *et al.* Vision-based robotic grasping from object localization, object pose estimation to grasp esti-

- mation for parallel grippers: a review[J]. *Artificial Intelligence Review*, 2021, 54(3): 1677-1734
- [2] Mahler J, Liang J, Niyaz S, *et al.* Dex-Net 2.0: deep learning to plan robust grasps with synthetic point clouds and analytic grasp metrics[OL]. [2023-10-21]. <https://arxiv.org/abs/1703.09312>
 - [3] Wang S F, Jiang X, Zhao J, *et al.* Efficient fully convolution neural network for generating pixel wise robotic grasps with high resolution images[C] // *Proceedings of the IEEE International Conference on Robotics and Biomimetics*. Los Alamitos: IEEE Computer Society Press, 2019: 474-480
 - [4] Lenz I, Lee H, Saxena A. Deep learning for detecting robotic grasps[J]. *The International Journal of Robotics Research*, 2015, 34(4-5): 705-724
 - [5] Depierre A, Dellandréa E, Chen L. Optimizing correlated graspability score and grasp regression for better grasp prediction[OL]. [2023-10-21]. <https://arxiv.org/abs/2002.00872v1>
 - [6] Morrison D, Corke P, Leitner J. Closing the loop for robotic grasping: A real-time, generative grasp synthesis approach [OL]. [2023-10-21]. <https://arxiv.org/abs/1804.05172>
 - [7] ten Pas A, Platt R. Using geometry to detect grasp poses in 3D point clouds[M] // *Bicchi A, Burgard W. Robotics Research*. Cham: Springer, 2017: 307-324
 - [8] ten Pas A, Gualtieri M, Saenko K, *et al.* Grasp pose detection in point clouds[J]. *The International Journal of Robotics Research*, 2017, 36(13-14): 1455-1473
 - [9] Liang H Z, Ma X J, Li S, *et al.* PointNetGPD: detecting grasp configurations from point sets[C] // *Proceedings of the International Conference on Robotics and Automation*. Los Alamitos: IEEE Computer Society Press, 2019: 3629-3635
 - [10] Fang H S, Wang C X, Gou M H, *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*. Los Alamitos: IEEE Computer Society Press, 2020: 11441-11450
 - [11] Qin Y Z, Chen R, Zhu H, *et al.* S4G: amodal single-view single-shot SE(3) grasp detection in cluttered scenes[C] // *Proceedings of the Conference on Robot Learning*. New York: PMLR, 2020: 53-65
 - [12] Kleeberger K, Völkl M, Moosmann M, *et al.* Transferring experience from simulation to the real world for precise pick-and-place tasks in highly cluttered scenes[C] // *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*. Los Alamitos: IEEE Computer Society Press, 2020: 9681-9688
 - [13] Zhao B L, Zhang H B, Lan X G, *et al.* RegNet: REgion-based grasp network for single-shot grasp detection in point clouds[OL]. [2023-10-21]. <https://arxiv.org/pdf/2002.12647v1>
 - [14] Ni P Y, Zhang W G, Zhu X X, *et al.* PointNet++ grasping: learning an end-to-end spatial grasp generation algorithm from sparse point clouds[C] // *Proceedings of the IEEE International Conference on Robotics and Automation*. Los Alamitos: IEEE Computer Society Press, 2020: 3619-3625
 - [15] Qi C R, Yi L, Su H, *et al.* PointNet++: deep hierarchical feature learning on point sets in a metric space[C] // *Proceedings of the 31st International Conference on Neural Information Processing Systems*. Red Hook: Curran Associates Inc, 2017: 5105-5114
 - [16] Xiang Y, Schmidt T, Narayanan V, *et al.* PoseCNN: a convolutional neural network for 6D object pose estimation in cluttered scenes[C] // *Proceedings of the 14th Robotics: Science and Systems*. San Francisco: RSS, 2018
 - [17] Jiang Y, Moseson S, Saxena A. Efficient grasping from RGBD images: learning using a new rectangle representation[C] // *Proceedings of the IEEE International Conference on Robotics and Automation*. Los Alamitos: IEEE Computer Society Press, 2011: 3304-3311
 - [18] Zeng L, Lv W J, Dong Z K, *et al.* PPR-Net++: accurate 6-D pose estimation in stacked scenarios[J]. *IEEE Transactions on Automation Science and Engineering*, 2022, 19(4): 3139-3151
 - [19] Kleeberger K, Landgraf C, Huber M F. Large-scale 6D object pose estimation dataset for industrial bin-picking[C] // *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*. Los Alamitos: IEEE Computer Society Press, 2019: 2573-2578