

多尺度物体杂乱堆叠场景下的机器人抓取位姿检测

郭万金^{1,2} 王晨阳¹ 叶洲闻¹ 徐明坤¹ 党金虎¹ 梁培栋³ 张磊¹

(1. 长安大学智能制造系统研究所, 西安 710064; 2. 长三角哈特机器人产业技术研究院, 芜湖 241007;
3. 福建(泉州)先进制造技术研究院, 泉州 362000)

摘要: 针对多尺度物体杂乱堆叠场景因物体边界模糊、相互遮挡,易导致机器人稳定抓取位姿检测困难的问题,提出一种多尺度特征融合的机器人稳定抓取位姿检测方法。首先,设计一种结合 KD-tree 加速欧氏聚类与高斯采样的两阶段点云预处理方法,提取稳定抓取目标点并抑制环境噪声。其次,通过多尺度圆柱采样与位置编码联合提取抓取中心的局部几何特征,并引入多头自注意力机制提取全局上下文特征,提升抓取稳定性与抓取检测精度。最后,在 GraspNet-1Billion 数据集上进行抓取位姿准确度评估实验,并在 UR10 机器人上进行真实场景下的抓取实验;实验结果表明,所提方法在 3 类别数据集中的平均精度 (AP) 分别为 73.46%、65.64% 和 26.29%,随机杂乱场景下机器人的平均抓取成功率为 85.71%,三层堆叠实验场景下机器人的抓取成功率为 81.40%,验证了所提方法的有效性。

关键词: 机器人抓取; 抓取位姿检测; 多尺度特征融合; 注意力机制

中图分类号: TP242.6 文献标识码: A 文章编号: 1000-1298(2026)15-0407-11 **OSID:** 

Robotic Grasp Pose Detection in Cluttered Scenes with
Multi-scale Object Stacking

Guo Wanjin^{1,2} Wang Chenyang¹ Ye Zhouwen¹ Xu Mingkun¹ Dang Jinhu¹ Liang Peidong³ Zhang Lei¹

(1. Institute of Smart Manufacturing Systems, Chang'an University, Xi'an 710064, China

2. Yangtze River Delta HIT Robot Technology Research Institute, Wuhu 241007, China

3. Fujian (Quanzhou) Institute of Advanced Manufacturing Technology, Quanzhou 362000, China)

Abstract: Robotic grasp pose detection in cluttered scenes with multi-scale objects remains challenging due to severe occlusion and ambiguous object boundaries, which often leads to unstable grasp predictions. To address this problem, a robotic stable grasp pose detection method was proposed based on multi-scale feature fusion. Firstly, a two-stage point cloud preprocessing strategy combining KD-tree-accelerated Euclidean clustering and Gaussian sampling was designed to extract stable grasp target points while effectively suppressing environmental noise. This preprocessing step improved the reliability of candidate grasp points in cluttered and stacked scenes. Secondly, local geometric features around grasp centers were extracted through multi-scale cylindrical sampling combined with positional encoding, enabling the representation of grasp-relevant geometry under different spatial scales. On this basis, a multi-head self-attention mechanism was introduced to capture global contextual features among grasp candidates, which further enhanced grasp stability and improved grasp pose detection accuracy under occlusion. Finally, extensive experiments were conducted on the GraspNet-1Billion dataset to evaluate grasp pose accuracy, along with real-world grasping experiments on a UR10 robotic platform. Experimental results showed that the proposed method achieved average precision (AP) values of 73.46%, 65.64%, and 26.29% on the Seen, Similar, and Novel subsets, respectively. In real-world

收稿日期: 2026-02-26 修回日期: 2026-03-16

基金项目: 国家自然科学基金项目(52275005)、陕西省自然科学基金基础研究计划项目(2025JC-QYXQ-027)、安徽博士后科研项目(2026B1324)、中央高校基本科研业务费专项资金项目(300102255202、300102253201)、中国博士后科学基金项目(2024M760002)和安徽省机器视觉检测与感知重点实验室开放基金项目(KLMVI-2025-HIT-06)

作者简介: 郭万金(1983—),男,副教授,博士,主要从事机器人与智能控制研究,E-mail: guowanjin@chd.edu.cn

通信作者: 张磊(1985—),男,副教授,博士,主要从事多智能体系统与柔性智能生产线设计研究,E-mail: zhlei0202@chd.edu.cn

experiments, the robot attained an average grasp success rate of 85.71% in random cluttered scenes and 81.40% in three-layer stacked scenarios. These results demonstrated the effectiveness of the proposed method for stable grasp pose detection in cluttered scenes with multi-scale objects.

Key words: robot grasping; grasping pose detection; multi-scale feature fusion; attention mechanism

0 引言

通用目标抓取位姿检测是机器人在现实环境中实现自动化作业与人机交互的基础能力。然而,在物流分拣、果蔬采摘以及家庭服务等实际应用场景中,物体通常以多尺度、杂乱堆叠的形式分布于有限空间内,不同尺度物体的堆叠边界与接触区域受堆叠角度和遮挡关系影响显著,呈现出明显的边界模糊性,导致机器人不易检测到稳定抓取位姿。如何在多尺度物体杂乱堆叠场景下针对被遮挡物体实现高精度、高鲁棒性及高稳定性的抓取检测,已成为机器人领域的关键研究热点^[1-2]。

现有抓取检测研究可分为基于二维图像及RGB-D信息的抓取位姿检测^[3-5]与基于6-DoF姿态估计的抓取位姿检测^[6-9],后者依托三维空间位姿进行建模,更适配杂乱堆叠场景下的抓取任务需求^[10-14]。

针对杂乱堆叠环境中物体尺度差异显著与点云特征缺失的问题,Gui等^[15]通过点云序列化注意力机制强化特征提取主干网络,显著提高抓取检测精度,Yu等^[6,16]分别从类别级位姿估计与无形状先验类别级位姿估计角度提升未知物体抓取的泛化能力,Li等^[17]通过语义与碰撞联合学习增强抓取安全性,Yu等^[18]提出的MTGrasp通过多尺度建模改善不同尺寸物体的抓取适应性。Niu等^[19]针对低照度环境引入视觉增强策略,提升弱光条件下的抓取鲁棒性,Tian等^[20]通过对向接触点表示与双解码结构,实现多目标场景下的高效抓取检测。Lu等^[21]通过混合物理评价指标引入力闭合、重力与碰撞约束,Hoang等^[22]通过上下文感知建模提高了遮挡条件下无碰撞抓取生成的可靠性。针对严重遮挡与堆叠场景,文献[23-24]分别提出NG-Net与CGNet,通过上下文感知与物理度量增强复杂场景下抓取检测的鲁棒性。

尽管现有研究在提升抓取检测性能方面取得了显著进展,在实际应用中,机器人往往需要面对多尺度物体在有限空间内杂乱堆叠所形成的复杂场景,在此类场景中,物体因随机姿态与严重相互遮挡导致边界模糊,难以实现准确的空间分离与稳定抓取区域判定。一方面,由于局部点云信息缺失,现有方法难以充分建模物体间的遮挡关系,极易导致抓取

点选取不稳定及位姿分布离散;另一方面,背景杂乱与环境噪声进一步削弱了特征表达的鲁棒性,使得抓取检测精度与抓取稳定性难以兼顾。

针对多尺度物体杂乱堆叠场景中因物体边界模糊、相互遮挡严重,导致机器人稳定抓取位姿困难的问题,本文提出一种多尺度特征融合的机器人稳定抓取位姿检测方法。首先,设计一种结合KD-tree加速欧氏聚类与高斯采样的两阶段点云预处理方法,通过对场景点云进行聚类分割与概率筛选,提取接近物体质心且包含稳定几何信息的抓取目标点,抑制环境噪声干扰。其次,通过多尺度圆柱采样与位置编码联合提取抓取中心的局部几何特征,并引入多头自注意力机制提取全局上下文特征,实现多尺度特征的加权融合,增强模型对遮挡关系与结构差异的表征能力,从而提升抓取稳定性并提高抓取检测精度。最后,在GraspNet-1Billion数据集上进行抓取位姿准确度评估实验,并在UR10机器人平台上开展真实场景抓取实验。对比其他抓取检测方法,验证所提多尺度特征融合机器人稳定抓取位姿检测方法的有效性。

1 抓取位姿检测方法总体结构

为提升多尺度物体杂乱堆叠场景下机器人抓取位姿检测稳定性,本文提出一种多尺度特征融合的机器人稳定抓取位姿检测方法,所提方法的总体结构如图1所示。

首先,利用RGB图像、深度图像及物体掩码构建场景点云,将其输入融合KD-tree加速欧氏聚类与高斯采样的两阶段点云预处理模块,对原始点云进行聚类分割与抓取目标点采样;随后采用稀疏卷积网络对点云进行初步几何特征提取,再由PointNet编解码器提取高维语义特征,特征经过多层感知机(Multilayer perceptron, MLP)处理后,通过双阈值筛选提取关键抓取目标点,得到候选点云集合。

针对多尺度物体杂乱堆叠场景中局部结构尺度差异显著导致抓取中心特征缺失的问题,利用多尺度圆柱采样模块与采样点编码器提取候选点云局部特征,并引入位置编码以增强局部特征的空间表达能力,获得表征抓取中心几何结构的多尺度局部特征向量。同时,为弥补局部特征在严重遮挡条件下

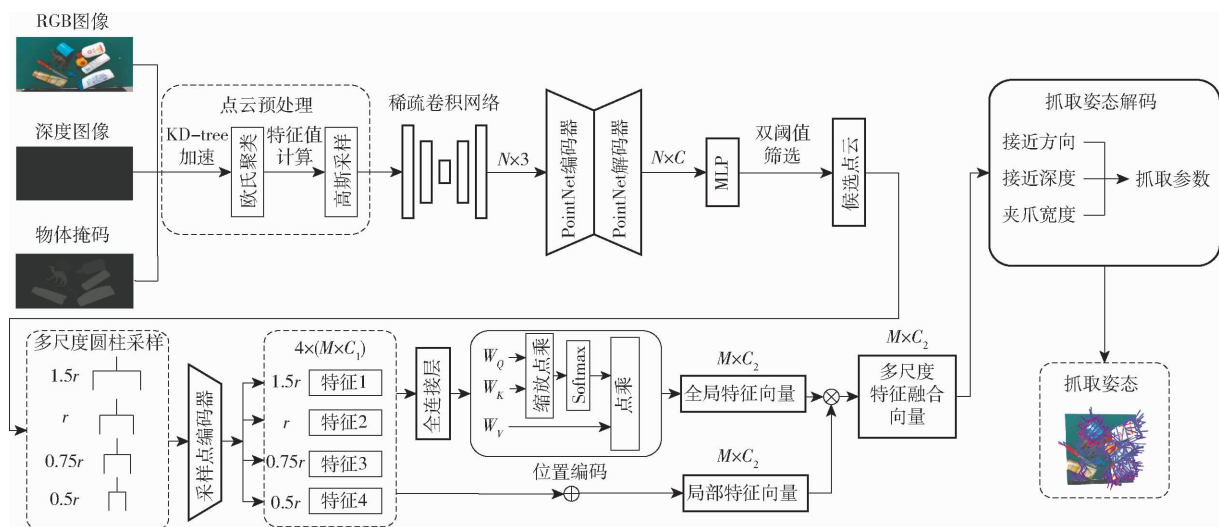


图 1 抓取位姿检测方法总体结构图

Fig. 1 Overall framework of proposed grasp pose detection method

对整体结构关系建模能力不足的问题,通过多头自注意力机制对候选点云进行全局上下文建模,提取全局特征向量,并与局部多尺度特征进行加权融合,形成增强的多尺度特征融合表示。最后,将融合后的特征输入抓取姿态解码模块,对抓取接近方向、抓取深度及夹爪宽度等参数进行并行回归,输出最终的机器人抓取参数及其对应的可视化抓取姿态。

2 多尺度特征融合的机器人稳定抓取位姿检测方法

在本文中,机器人的所有抓取操作与位姿计算均以相机坐标系为全局统一基准,机器人抓取位姿的几何关系示意图如图 2 所示。图 2 中坐标系 $O-xyz$ 为相机坐标系, $O_g-X_gY_gZ_g$ 为夹爪坐标系,在夹爪坐标系中, Z_g 轴指向夹爪的正前方,表示机器人的抓取接近方向; X_g 轴平行于夹爪的张合平面,表示夹爪手指开闭的滑动方向; Y_g 轴垂直于 X_g 轴和 Z_g 轴构成的平面,满足右手坐标系定则。

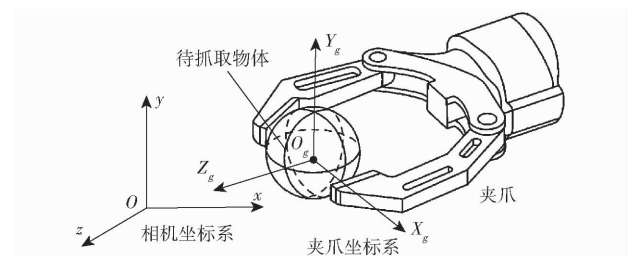


图 2 机器人抓取位姿几何关系示意图

Fig. 2 Geometric relationship of robot grasp pose

机器人的抓取位姿 G 描述了将夹爪坐标系转换至相机坐标系的刚体变换关系^[14],包含夹爪在相机坐标系下的三维空间位置、姿态以及夹爪开口宽度,其定义为

$$G = [R, t, \omega] \quad (1)$$

式中, $R \in SO(3)$ 表示夹爪坐标系相对于相机坐标系的 3×3 正交旋转矩阵,用于描述夹爪在三维空间中的姿态以及抓取接近方向, $t \in \mathbf{R}^3$ 表示从相机坐标系原点 O 指向夹爪抓取中心 O_g 的空间位置向量,用于描述抓取点在三维空间中的位置, $\omega \in \mathbf{R}^+$ 表示夹爪的自适应开口宽度,用于确定夹爪在抓取时的张开距离。

2.1 两阶段点云预处理

现有抓取位姿检测方法常采用均匀或随机采样来降低点云计算复杂度。然而,在多尺度物体杂乱堆叠场景中,此类采样方法难以有效抑制背景干扰,且在目标点选择上易受噪声与遮挡影响,从而导致抓取目标点稳定性不足。针对上述问题,本文设计一种融合 KD-tree 加速欧氏聚类与高斯采样的两阶段点云预处理方法。首先,通过聚类将场景点云划分为相对独立的对象区域,减少不同物体之间的相互干扰;随后基于高斯概率模型对各对象区域进行采样,优先保留具有较高表面完整性与几何稳定性的抓取候选区域,从而在保证计算效率的同时有效抑制环境噪声。不同于侧重均匀空间覆盖的最远点采样 (Farthest point sampling, FPS) 方法^[25],高斯概率密度采样通过对点云分布进行概率建模,引导采样过程聚焦于几何分布稳定且表面连续的区域,从而更有利于保留与抓取稳定性密切相关的关键抓取目标点。

为将多尺度物体杂乱堆叠场景中的物体进行初步分离并剔除背景噪声,第 1 阶段采用欧氏聚类方法对原始点云进行处理。考虑到传统欧氏聚类在大规模点云上的近邻搜索耗时较高,本文引入 KD-tree 构建空间索引结构,以加速邻域查询过程。经处理后,场景点云被划分为若干独立的点云簇集合:

$C = \{C_1, C_2, \dots, C_m\}$, 对于每一个点云簇 C_i , 记其中包含的点为 $p_{i,j}$ (表示第 i 个簇中的第 j 个点), 为刻画该点云簇的空间分布特性, 分别计算其几何中心 μ_i 和协方差矩阵 Σ_i, μ_i 和 Σ_i 的计算式分别为

$$\mu_i = \frac{1}{k_i} \sum_{j=1}^{k_i} p_{i,j} \quad (2)$$

$$\Sigma_i = \frac{1}{k_i - 1} \sum_{j=1}^{k_i} (p_{i,j} - \mu_i)(p_{i,j} - \mu_i)^T \quad (3)$$

式中, k_i 为第 i 个点云簇中点的总数量; Σ_i 反映了物体簇在三维空间中的散布情况, 为后续构建高斯分布模型提供了参数基础。

在获得各独立点云簇的统计特征后, 第 2 阶段基于多元高斯分布模型对抓取候选区域进行采样。利用式(2)、(3)得到的邻域几何中心 μ_i 与协方差矩阵 Σ_i , 构建描述该簇空间分布的高斯概率密度函数。多元高斯分布的归一化常数 η_i 为

$$\eta_i = \frac{1}{(2\pi)^{3/2} |\Sigma_i|^{1/2}} \quad (4)$$

式中, $(2\pi)^{3/2}$ 是三维坐标系的常数, $|\Sigma_i|$ 表示协方差矩阵的行列式。利用该常数, 对于点云簇 C_i 内点 $p_{i,j}$ 的高斯概率密度函数 $f(p_{i,j})$ 表示为

$$f(p_{i,j}) = \eta_i \exp \left(-\frac{1}{2} (p_{i,j} - \mu_i)^T \Sigma_i^{-1} (p_{i,j} - \mu_i) \right) \quad (5)$$

概率密度较高的区域通常对应于点云分布密集、表面连续且远离噪声的区域, 往往位于物体的主体或稳定接触表面而非几何突变明显的边缘或角点区域。高斯采样可以在采样过程中避开不稳定的边缘抓取点, 从而提高后续抓取姿态检测的稳定性与可执行性。

在多尺度物体杂乱堆叠场景中, 为了进一步增强检测的鲁棒性, 本文引入簇均衡采样策略, 通过为每个独立点云簇分配相等的采样点数量, 避免了大尺度物体主导采样结果的问题; 同时, 依据高斯分布特性确定质心优先级, 引导采样集中于高置信度区域, 并通过协方差矩阵保留各方向上的几何可变性。

在计算效率方面, 通过引入 KD-tree 加速结构, 整个预处理过程的时间复杂度主要由近邻搜索主导, 约为 $O(N \log N)$, 相较于基于全局遍历的核密度估计 (Kernel density estimation, KDE) 方法 (时间复杂度为 $O(N^2)$) 具有更高的计算效率, 能够满足机器人抓取任务实时性的需求。

2.2 融入注意力机制的多尺度特征融合

在多尺度物体杂乱堆叠场景中局部点云信息易受遮挡与噪声影响, 可靠的抓取位姿检测不仅需要评估抓取中心在局部几何上的可行性与碰撞安全

性, 还需结合全局上下文信息, 避免因物体间遮挡与空间耦合而引发的抓取冲突。从任务角度来看, 优质的抓取中心通常应在多种抓取配置下均具备较高的可行性, 例如不同的接近方向、抓取深度及平面内旋转角度, 从而在存在感知误差与执行扰动的情況下仍具备较高的成功率与容错性。基于上述考虑, 本文在多尺度局部几何建模基础上, 引入全局上下文建模能力, 从大规模点云中筛选稳定且鲁棒的候选抓取中心。

2.2.1 抓取中心筛选

首先, 对场景中每个点 p_i 在不同视点 j 、旋转角度 k 、接近深度 l 的抓取配置进行评估。定义 E_q 为配置的有效性 (若配置有效则取 1, 否则取 0), S_q^p 为该配置的抓取质量评分, q 表示由抓取中心点 p_i 、视点索引 j 、夹爪绕接近方向的旋转角度 k 以及接近深度 l 共同构成的一组抓取姿态参数组合。 σ_q 为该配置的无冲突概率, 该点处有效抓取配置的总数量 N_i 计算式为

$$N_i = \sum_{j=1}^V \sum_{k=1}^A \sum_{l=1}^D E_q \quad (6)$$

式中 V ——候选视点的离散采样数量

A ——夹爪绕接近方向的旋转角离散采样数量

D ——沿接近方向的抓取深度离散采样数量

该指标用于衡量该点作为抓取中心时, 在不同姿态参数扰动下的抓取可行空间大小。若 $N_i > 0$, 则该点的平均抓取质量 $\bar{g}_i$ 和平均无碰撞率 $\bar{\sigma}_i$ 定义为

$$\bar{g}_i = \frac{1}{N_i} \sum_{j=1}^V \sum_{k=1}^A \sum_{l=1}^D E_q S_q^p \quad (7)$$

$$\bar{\sigma}_i = \frac{1}{N_i} \sum_{j=1}^V \sum_{k=1}^A \sum_{l=1}^D E_q \sigma_q \quad (8)$$

其中平均抓取质量 $\bar{g}_i$ 反映了该点在多种抓取配置下的整体抓取质量水平; 平均无碰撞率 $\bar{\sigma}_i$ 则反映了其在多尺度物体杂乱堆叠场景中规避碰撞干扰的能力。

在多尺度物体相互遮挡的场景中, 仅依据抓取质量或无碰撞率中的单一指标, 易导致抓取中心选择偏向局部最优而忽略全局安全性。为此, 本文构建一种兼顾抓取稳定性与环境安全性的双阈值筛选机制。首先, 将平均无碰撞率作为该点的基础评分 s_i^l , 即 $s_i^l = \bar{\sigma}_i$, 并对其进行归一化处理得到

$$\hat{s}_i^l = \frac{s_i^l - \min s_i^l}{\max s_i^l - \min s_i^l} \quad (9)$$

随后, 引入加权系数 α , 结合平均抓取质量与归一化后的基础评分计算综合高阈值评分 s_i^h , 即

$$s_i^h = \alpha \bar{g}_i + (1 - \alpha) \hat{s}_i^l \quad (10)$$

该双阈值评分机制使网络在抓取中心筛选过程

中能够隐式建模物体之间的空间关系,倾向于选择远离遮挡干扰且在多种抓取配置下均表现稳定的点作为候选抓取中心。最终,根据低阈值与高阈值约束,剔除低置信度区域,得到用于后续多尺度特征建模的候选抓取中心集合。

2.2.2 融合多尺度圆柱采样的局部特征提取

抓取位姿检测的鲁棒性在很大程度上依赖于对抓取中心局部几何特征的有效建模。然而,在多尺度物体杂乱堆叠的复杂场景中,由于物体之间相互遮挡严重、边界与接触区域呈现模糊性,局部点云往往存在信息缺失与分布不均的问题。当采样尺度与目标物体实际尺寸或稳定接触区域不匹配时,不同采样结果之间的几何特征差异显著,缺乏多尺度结构约束的局部特征表示容易使检测的抓取姿态偏离物体的稳定受力区域,从而影响抓取位姿检测在杂乱堆叠场景中的稳定性。

为有效反映多尺度物体杂乱堆叠场景中抓取中心区域在不同空间范围内的几何特性,本文采用融合多尺度圆柱采样与位置编码的局部特征建模方式,首先,利用一组与夹爪半径成比例的同心圆柱体(比例因子 $\tau = 0.50, 0.75, 1.00, 1.50$)对局部点云进行多尺度采样。其次,对不同尺度下获得的采样点云分别通过采样点编码器进行特征编码,提取包含不同层级几何语义的局部特征。

针对候选抓取中心的局部特征,为增强对物体间空间关系的感知能力,利用相对位置编码 $\delta(\cdot)$ 建模抓取中心与其邻域点之间的空间关系,从而在存在遮挡与边界模糊的情况下,提升局部特征对稳定抓取区域的判别能力。每个尺度 τ_k 的局部特征 f_{i,τ_k}^{con} 表达式为

$$f_{i,\tau_k}^{con} = \max_{c_j \in N(c_i, r)} \gamma(f_j + \delta(\|c_i - c_j\|_2)) \quad (11)$$

式中 $N(c_i, r)$ ——以 c_i 为中心、半径为 r 的局部邻域内的抓取中心集合

f_j ——邻域内抓取中心 c_j 的特征向量

$\delta(\cdot)$ ——位置编码函数

$\gamma(\cdot)$ ——特征映射函数

max——逐元素最大值池化操作,用于提取显著的邻域特征

c_i ——当前待编码的候选抓取中心坐标

c_j ——局部邻域 $N(c_i, r)$ 内第 j 个领域抓取中心坐标

最终,将各尺度下获得的局部特征向量进行拼接,并通过线性映射进行融合,得到统一的局部特征向量

$$f_i^{con} = \text{MLP}([f_{i,0.50}^{con}, f_{i,0.75}^{con}, f_{i,1.00}^{con}, f_{i,1.50}^{con}]) \quad (12)$$

通过在不同尺度感受野下联合建模局部几何结构,增强了对不同尺寸物体特征的表达能力,并解决抓取位姿不稳定的问题。

2.2.3 融入注意力机制的全局特征提取

在多尺度圆柱采样获得的局部特征能够反映抓取中心区域几何细节的基础上,仍需进一步建模候选抓取中心之间的空间依赖关系。在多尺度物体杂乱堆叠场景中,物体之间往往存在显著的遮挡与空间耦合,仅依赖局部邻域信息难以判断某一抓取配置是否会与周围物体产生潜在冲突。因此,在局部特征之上引入注意力机制,以对抓取位姿进行全局一致性约束。

在对局部特征进行最大池化操作以保留具有判别力的几何信息并抑制冗余噪声后,引入多头自注意力机制对候选抓取中心之间的空间拓扑关系进行建模,以捕获能够反映物体间遮挡与空间约束的全局特征。为在多尺度物体杂乱堆叠场景下并行建模候选抓取中心之间的多样化空间依赖关系,在筛选后的候选点集上采用多头自注意力机制。

首先,将每个候选抓取中心的局部特征向量 f_i^{con} 分别映射到 h 个不同的特征子空间中,以并行刻画不同抓取中心之间在遮挡与空间分布层面的潜在关联模式。对于第 m 个注意力头 ($m = 1, 2, \dots, h$),其计算过程为

$$head_m = \text{Softmax}\left(\frac{(f^{con} W_m^Q)(f^{con} W_m^K)^T}{\sqrt{d_k}}\right)(f^{con} W_m^V) \quad (13)$$

式中, W_m^Q 、 W_m^K 、 W_m^V 分别为第 m 个头的查询(Query)、键(Key)和值(Value)的线性投影矩阵; d_k 为键向量维度, $\sqrt{d_k}$ 为缩放因子; $\text{Softmax}(\cdot)$ 用于归一化生成注意力权重。随后,将所有注意力头的输出进行拼接(Concatenate),通过线性投影层融合信息,同时引入残差连接以保留原始局部几何信息,最终得到融合局部与全局上下文约束的全局特征向量 f_i^{attn}

$$f_i^{attn} = \text{Concat}(head_1, head_2, \dots, head_h) W^O + f_i^{con} \quad (14)$$

式中 W^O ——输出层的线性投影矩阵

通过并行建模多个注意力头,网络能够同时关注局部抓取接触区域的几何细节以及不同物体之间的全局空间关联信息,从而在遮挡严重的堆叠场景中对抓取位姿检测施加全局一致性约束,降低潜在抓取冲突的发生概率。

2.2.4 多尺度特征融合

经过局部多尺度建模与全局上下文建模后,

特征中同时包含局部几何细节与跨物体空间约束信息。为避免不同层级特征直接叠加带来的信息冗余,并聚焦于稳定抓取最相关的判别特征,采用门控融合机制对特征进行进一步整合,由于在全局特征提取阶段已通过残差连接将局部几何信息与全局上下文信息进行初步融合,输出特征 f_i^{attn} 同时包含局部与全局两类约束。为了进一步提炼特征并抑制冗余信息,将融合后的特征矩阵 $F \in \mathbf{R}^{M \times C_2}$ 输入多层感知机进行通道变换与非线性映射,得到最终的多尺度特征融合表示,为后续抓取姿态解码提供支持。

2.3 抓取姿态解码

在获得融合局部几何与全局空间约束的多尺度特征表示后,需要将高维特征映射为可直接执行的抓取参数。考虑到抓取位姿各参数在物理意义与学习难度上的差异,采用分层回归策略对抓取姿态进行解码,以降低多参数联合回归在复杂堆叠场景中的学习难度。

夹爪位姿子解码器采用正弦-余弦编码(Sine-cosine encoding)对抓取坐标系的接近方向向量进行回归,以有效避免角度周期性带来的不连续问题,提升位姿检测在多尺度物体杂乱堆叠场景中的稳定性。同时,结合抓取中心与其局部点云表面之间的几何拓扑关系,对抓取深度偏移量进行估计,以避免夹爪在遮挡环境中过度接近或发生碰撞。夹爪宽度子解码器通过MLP检测自适应抓取宽度 ω ,并通过Sigmoid激活函数将检测结果约束在夹爪的物理开口范围内,从而确保在复杂堆叠条件下抓取执行的物理可行性与安全性。

2.4 损失函数

训练采用多任务损失,联合平衡聚类精度、抓取质量回归、抓取宽度检测、遮挡分类和可见性估计。总体损失函数是所有子损失的加权总和,即

$$L = \lambda_1 L_c + \lambda_2 L_g + \lambda_3 L_\omega + \lambda_4 L_o + \lambda_5 L_v \quad (15)$$

式中,聚类损失 L_c 和遮挡分类损失 L_o 采用二元交叉熵损失函数;抓取质量损失 L_g 和夹持器宽度损失

L_ω 采用Smooth-L1损失函数;能见度估计损失 L_v 采用均方误差损失函数; λ_1 、 λ_2 、 λ_3 、 λ_4 和 λ_5 分别为对应损失函数的权重。

3 数据集实验

为验证所提方法在多尺度物体杂乱堆叠场景下的有效性与鲁棒性,在大规模公开数据集GraspNet-1Billion上开展抓取位姿准确度评估实验,并将本文方法与当前主流的抓取检测算法进行准确度对比。

3.1 实验参数设置

实验基于GraspNet-1Billion数据集集中的RealSense子集进行,GraspNet-1Billion数据集是目前针对杂乱堆叠场景下6-DoF抓取姿态检测的主流大规模基准数据集之一。为全面评估模型的泛化能力,测试集被划分为已知物体(Seen)、相似物体(Similar)和新物体(Novel)3个子测试集。实验评估遵循GSNet的标准指标,主要指标包括不同摩擦条件下的平均精度(Average precision, AP),其中摩擦因数 μ 的取值范围为[0.2,1.0],分别统计AP、AP_{0.8}和AP_{0.4}。在具体实现细节上,网络输入采用本文提出的两阶段点云预处理方法,提取出15000个高斯采样点作为输入。随后,经过双阈值筛选保留1024个候选抓取中心,并从离散化的300个接近方向、12个平面内旋转角度以及4个接近深度中解码出最终的抓取参数。模型基于PyTorch框架搭建,并在NVIDIA RTX 3080GPU上进行训练。优化器采用Adam,训练总周期(Epoch)为12,批次大小(Batch size)为2,初始学习率为0.001,且每个Epoch的衰减率设为0.95,对于损失函数,将 λ_1 、 λ_2 、 λ_3 、 λ_4 和 λ_5 分别设为10、15、10、10和100。

3.2 实验结果与分析

为了验证本文提出的方法在多尺度物体杂乱堆叠场景下的抓取检测性能,本文在GraspNet-1Billion数据集的RealSense子集上进行对比实验,结果如表1所示。

表1 GraspNet-1Billion (RealSense)数据集对比实验结果

Tab. 1 GraspNet – 1Billion (RealSense) dataset comparison experiment results									%
方法	Seen			Similar			Novel		
	AP	AP _{0.8}	AP _{0.4}	AP	AP _{0.8}	AP _{0.4}	AP	AP _{0.8}	AP _{0.4}
PointNetGPD ^[9]	25.96	33.01	15.37	22.68	29.15	20.76	9.23	9.89	2.74
GraspNet ^[25]	27.56	33.43	16.95	26.11	34.18	14.23	10.55	11.25	3.98
Parallel Graph Network ^[11]	50.12	59.28	42.34	43.90	53.17	35.58	16.94	20.36	8.37
GSNet ^[14]	65.70	76.25	61.08	53.75	65.04	45.97	23.98	29.93	14.05
MTGrasp ^[18]	70.57	81.43	65.89	61.29	73.29	53.54	27.79	34.57	15.79
本文方法	73.46	83.08	67.96	65.64	78.60	57.94	26.29	33.98	14.16

实验结果表明,本文方法在 Seen、Similar 及 Novel 3 个测试子集上的各项指标均优于 PointNetGPD、GraspNet 等早期算法,验证了所提抓取位姿检测方法在多尺度物体杂乱堆叠的大规模点云场景中的有效性与鲁棒性。

在 Similar(相似物体)测试集中,本文方法表现出较强的特征提取能力,AP、 $AP_{0.8}$ 和 $AP_{0.4}$ 分别达到 65.64%、78.60% 和 57.94%,相较于文献[18]中的 MTGrasp 方法分别提升了 4.35、5.31、4.40 个百分点,同时显著优于 GSNet 的 53.75%、65.04% 和 45.97%。这是由于 Similar 测试集中的物体在整体类别与几何轮廓上比较相似,但在局部尺寸比例、曲率分布以及潜在抓取接触区域结构等细节层面存在差异。针对该类局部差异,本文方法通过多尺度局部特征建模并结合空间上下文信息,有效缓解了遮挡与尺度变化对抓取中心判别的干扰,能够更稳定地聚焦于抓取目标点,从而提高了 Similar 子集上的抓取检测精度。

此外,在 Seen(已知物体)子集中,本文方法的 AP 达到 73.46%,分别高于 MTGrasp(70.57%)和 GSNet(65.70%),表明在已知几何结构条件下,本文方法能够保证抓取位姿检测的精度与稳定性;在 Novel(未知物体)子集中,本文方法整体性能与 GSNet、MTGrasp 等主流方法处于同一水平,表明本文方法在面对未见几何结构时具备一定的泛化能力。

然而,对比表 1 数据发现,本文方法在 Novel 数据集上的 AP 为 26.29%, $AP_{0.8}$ 为 33.98%, $AP_{0.4}$ 为 14.16%,在抓取评价指标下略低于 MTGrasp。该现象主要与本文采用的高斯概率密度采样方法有关,高斯概率密度采样方法在多尺度物体杂乱堆叠场景中能够有效抑制背景噪声并聚焦物体主体区域,从而提升相似与已知物体场景下抓取检测的稳定性;但其概率分布特性更倾向于保留几何中心区域点云,在一定程度上弱化了对物体边缘、角点及曲率突变区域的表达。由于这些区域在 6-DoF 抓取任务中往往对应关键的对向接触特征,当面对几何结构差异较大的未知物体时,该特性可能影响候选抓取覆盖范围,从而对 Novel 子集上的检测性能产生一定影响。

对模型在 GraspNet-1Billion 数据集上的抓取位姿估计结果进行了可视化,如图 3 所示。图 3a 为原始 RGB 图像,图 3b 为抓取位姿可视化结果,抓取位姿由抓取中心点和抓取姿态共同构成,抓取中心点确定抓取的空间位置,抓取姿态确定夹爪空间朝向与旋转方式。图 3c 为检测到的抓取中心点,抓取中心分布在对象曲面上,并集中在稳定、无碰撞的区域。

图 3d 为检测到的抓取姿态,抓取姿态定义为夹爪在三维空间中的朝向与旋转方式,表示夹爪接近目标物体的方向向量(接近方向)以及夹爪手指开闭所在的基准平面,图 3d 突出显示本文方法通过捕获全局空间上下文得到的适应复杂物体表面的夹爪姿态。

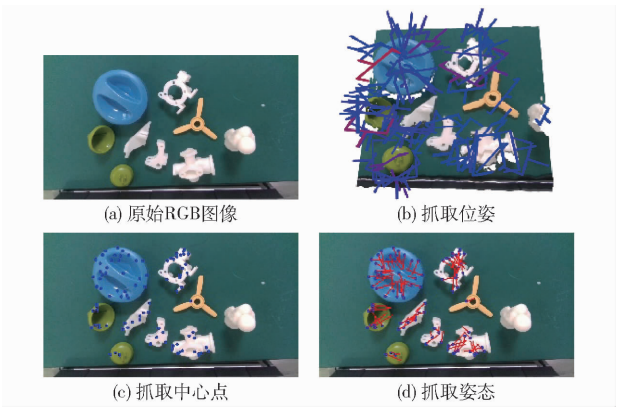


图 3 抓取位姿可视化

Fig. 3 Grasping pose visualization

3.3 消融实验

为验证所提出多尺度特征融合抓取位姿检测方法中高斯采样(Gaussian sampling, GS)、多尺度采样(Multi-scale sampling, MSS)、多头自注意力机制(Multi-head self-attention, MHSA)在多尺度物体杂乱堆叠场景下的作用,在 GraspNet-1Billion 数据集上开展消融实验,实验结果如表 2 所示。基线模型不包含高斯采样、多尺度采样及多头自注意力机制,其余实验在此基础上逐步引入组成方法,以分析各方法对抓取检测性能的影响。

表 2 消融实验结果

Tab. 2 Ablation results

%

GS	MSS	MHSA	AP		
			Seen	Similar	Novel
			67.12	53.75	24.31
✓			70.37	63.03	24.49
✓	✓		70.64	63.56	24.35
✓	✓	✓	73.46	65.64	26.29

注:“✓”表示包含此方法。

实验结果表明,相较于基线方法,引入高斯采样后,模型在 Seen、Similar 及 Novel 3 个子集上的性能均有所提升。特别是在 Similar(相似物体)测试集中,AP 由 53.75% 提升至 63.03%,表明高斯采样在多尺度物体杂乱堆叠场景中能够有效抑制背景噪声与冗余点云,使模型更加聚焦于物体主体区域,从而为后续特征建模提供更稳定的抓取候选点分布;引入多尺度圆柱采样后,模型在 Seen 与 Similar 子集上的性能继续提升,说明多尺度局部特征有助于刻画不同尺寸物体在堆叠与遮挡条件下的局部几何差异;引入多头自注意力机制后,模型在 3 个子集上的性能均达到最

取循环,随后等待下一次抓取指令,依次对场景中剩余目标物体进行抓取操作。

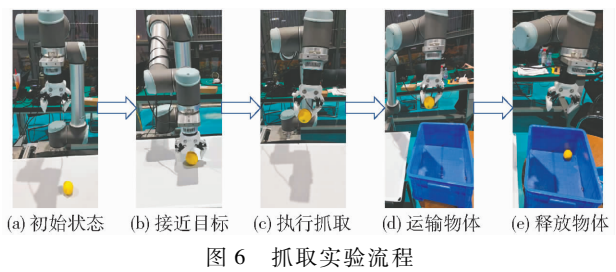


图 6 抓取实验流程

为客观衡量多尺度物体杂乱堆叠场景下的抓取执行效果,本文采用检测成功率 (Success rate, SR) 与抓取成功率 (Completion rate, CR) 作为核心评价指标。其中,成功检测目标并生成满足几何约束与夹爪开口限制、且通过碰撞检测的抓取位姿,记为一次检测成功;完成完整抓取执行流程则记为一次抓取成功。

4.2.2 随机杂乱场景抓取实验

实验中针对每个随机摆放的杂乱场景进行 5 次测试,考虑到抓取位姿检测结果的随机性,每个物体首次抓取失败后,允许再尝试 2 次。本文方法与文献[14]中 GSNet 方法在真实随机场景下的抓取实验对比结果如表 3 所示。

表 3 真实随机杂乱场景抓取实验结果对比
Tab.3 Comparison of grasping experimental results in real random cluttered scenes

场景	物体数量	GSNet			本文方法		
		尝试次数	SR/%	CR/%	尝试次数	SR/%	CR/%
场景 1	5	31	87.10	80.64	29	93.10	86.21
场景 2	5	33	84.85	75.76	30	86.67	83.33
场景 3	6	39	82.05	76.92	35	91.43	85.71
场景 4	6	41	80.49	73.17	36	91.67	83.33
场景 5	7	49	77.55	71.43	39	94.87	89.74
场景 6	7	47	80.85	74.47	41	90.24	85.37
平均值	6	40	81.67	75.00	35	91.43	85.71

表 3 中,尝试次数表示机器人在该场景下的总抓取尝试次数。实验结果显示,本文方法在不同复杂度的抓取测试场景中均保持了较高的稳定性,平均抓取成功率达到 85.71%,在相似物品场景(如场景 3)中,本文方法抓取成功率达到 85.71%;同时,在堆叠场景(如场景 5)中,本文方法抓取成功率达到 89.74%,验证了所提多尺度特征融合方法在真实随机杂乱场景下具备良好的抓取稳定性与抓取鲁棒性,能够有效满足多尺度物体随机杂乱场景下对稳定抓取的应用需求。

为直观评估不同方法在随机杂乱场景中的表现,将本文方法与 GraspNet、GSNet 方法在真实随机杂乱场景下的抓取检测结果进行对比,如图 7 所示。

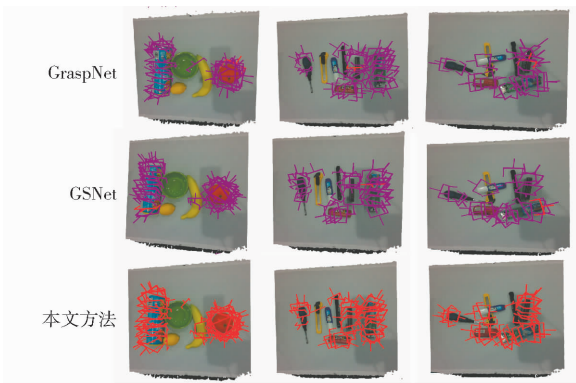


图 7 抓取姿态对比

Fig.7 Grasp posture comparison

GraspNet 受限于稀疏点云表示与有限的上下文推理能力,在面对细小物体或纹理相近的物体(如紧挨着的文具)时,往往出现漏检现象,且抓取点的覆盖率较低。GSNet 虽然在此基础上有所改进,但其生成的抓取姿态较为分散且杂乱,由于缺乏有效的碰撞约束,其检测结果中包含较多潜在发生碰撞或接触不稳定的抓取位姿,且极易在形状相似的相邻物体间产生重叠与混淆。本文提出的多尺度特征融合机器人稳定抓取位姿检测方法能够聚焦于对象的特定几何区域,在物体中心附近生成了稳定抓取位姿,展现出了较好的鲁棒性与抓取稳定性。

4.2.3 堆叠场景抓取实验

为进一步验证本文所提方法在多尺度物体紧密堆叠、遮挡关系显著增强的杂乱堆叠场景中的抓取性能,本节构建了两个具有不同结构复杂度的三层堆叠实验场景,并分别开展抓取实验,堆叠场景 7 主要用于定性展示本文方法在中等遮挡条件下的抓取执行过程,堆叠场景 8 用于进一步通过定量实验对抓取性能进行系统评估。实验场景如图 8 所示。

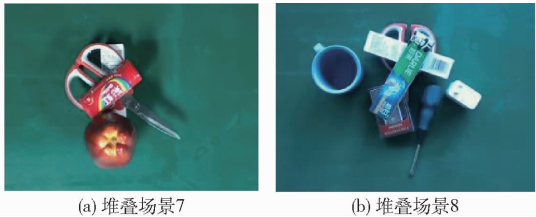


图 8 多尺度物体堆叠实验场景

Fig.8 Multi-scale object stacking experimental scenes

其中,堆叠场景 7 包含 4 个物体,堆叠结构为最上面 2 个物体、中间 1 个物体、最下面 1 个物体,整体堆叠密度较低、物体遮挡程度较轻,用来作为基础验证场景;场景 8 包含 7 个物体,堆叠结构为最上面 1 个物体、中间 2 个物体、最下面 4 个物体,堆叠密度更高、物体间遮挡更严重、空间布局更复杂。

图 9 展示了在堆叠场景 7 中本文方法的抓取执行过程。图 9 中第 1 行给出了物体被逐个抓取后场

景中剩余物体的 RGB 图像,用以直观反映抓取过程中场景结构的动态变化;第 2 行为本文所提方法检测到的各阶段最优抓取位姿;第 3 行为夹爪运动至对应最优抓取位姿时的实时状态照片;第 4 行为抓取过程中机器人的正面位姿照片。图 9a~9d 表示机器人对堆叠场景中物体的抓取顺序。

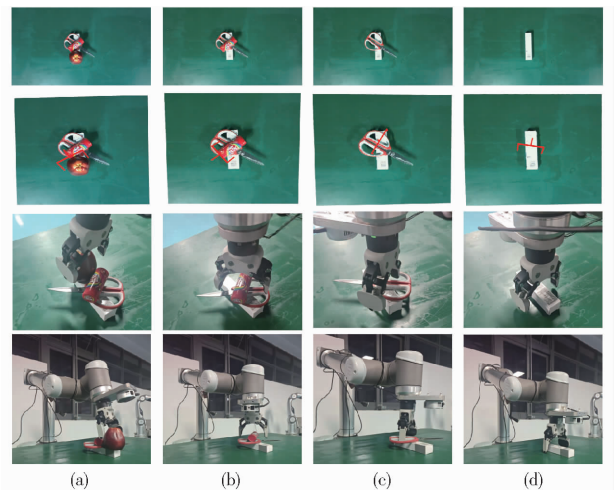


图 9 堆叠场景 7 中的抓取执行过程示例

Fig.9 Example of grasp execution process in stacking scene 7

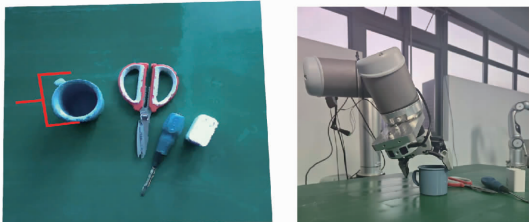
为验证本文方法在多尺度物体杂乱堆叠复杂场景中的效果,在堆叠场景 8 中进行了 5 次重复抓取测试,并详细记录对比了各方法的尝试次数、检测成功率(SR)和抓取成功率(CR)。实验结果显示,GraspNet 方法在 57 次尝试中,SR 和 CR 分别为 78.95% 和 52.64%;GSNet 方法在 51 次尝试中,SR 和 CR 分别为 86.28% 和 64.71%;本文方法在 43 次尝试中,SR 达 95.35%,CR 达 81.40%。实验结果表明,在堆叠场景 8 的三层杂乱堆叠复杂条件下,本文方法不仅能够以更少的尝试次数准确检测出可行抓取位姿,还具备较好的抓取稳定性与任务执行鲁棒性。

为深入探究不同抓取方法在堆叠场景中的性能差异,明确各方法抓取失败的核心原因,对堆叠场景 8 的实验结果进行了细化统计。根据失败发生的阶段与表现形式,将抓取失败原因划分为物体识别失败、抓取位姿不合理以及抓取过程中物体掉落 3 类。统计结果显示,GraspNet 方法在上述 3 类情况中的失败次数分别为 12 次、6 次和 9 次;GSNet 方法分别为 7 次、7 次和 4 次;本文方法对应的失败次数分别降至 2 次、5 次和 1 次。

在抓取位姿检测环节,本文方法的识别失败次数仅为 2 次,这表明,本文方法在多尺度物体杂乱堆叠状态下能够有效缓解物体遮挡、边界重叠导致的物体边界不清晰、特征提取困难等关键问题,能够较好地识别出堆叠场景中的目标物体。抓取稳定性方

面,抓取过程中物体掉落次数仅为 1 次,进一步分析发现,本文方法检测的抓取中心点集中在物体躯干区域,即物体质心附近。物体质心附近的抓取中心点能够有效提升夹爪与物体的接触稳定性,减少抓取过程中物体的晃动、滑移,进而降低物体掉落概率,保障抓取操作的可靠执行。

结合堆叠场景的空间约束特性进一步分析可知,位姿不合理主要体现在部分检测抓取位姿超出机械臂的可达工作空间范围,该问题主要源于感知阶段未引入机械臂运动学可达性约束,其典型示例如图 10 所示。后续可通过将可达性约束引入抓取位姿筛选或联合规划模块进一步优化。



(a) 抓取位姿示例

(b) 机器人位姿示例

图 10 堆叠场景 8 中的抓取失败示例

Fig.10 Grasp failure cases in stacking scene 8

5 结论

(1) 针对多尺度物体杂乱堆叠场景中由遮挡严重、边界模糊及尺度差异引起的抓取位姿检测困难问题,提出了一种多尺度特征融合的机器人稳定抓取位姿检测方法。通过在点云预处理阶段引入基于 KD-tree 加速的欧氏聚类与高斯概率采样相结合的两阶段方法,在遮挡条件下筛选出具有稳定几何结构的抓取候选点,降低环境噪声与冗余点对后续抓取检测的干扰。

(2) 通过多尺度圆柱采样与位置编码对抓取中心邻域的局部几何特征进行刻画,并进一步引入多头自注意力机制,建模抓取中心之间的全局空间关系,实现了局部几何细节与全局上下文信息的协同融合,解决由于多物体遮挡与尺度变化导致抓取中心判别困难的问题,提升抓取位姿检测的准确性与抓取稳定性。

(3) 在 GraspNet-1Billion 数据集与 UR10 机器人抓取实验平台上的实验结果验证了所提方法的有效性。在 GraspNet-1Billion 数据集上,本文方法在 Seen、Similar、Novel 3 个子集上的平均精度(AP)分别达到 73.46%、65.64% 和 26.29%;在 UR10 机器人真实抓取实验中,随机杂乱场景下的平均抓取成功率达到 85.71%,三层堆叠实验场景下机器人的抓取成功率为 81.40%,验证了本文方法在多尺度物体杂乱堆叠场景中的有效性。

参 考 文 献

- [1] 黄赛珂, 周风余, 朱振威, 等. 基于推抓技能协同的家庭复杂场景下的物品抓取方法[J]. 机器人, 2026, 48(1): 1 – 10, 37.
- Huang Saikē, Zhou Fengyu, Zhu Zhenwei, et al. An object grasping method in complex household scenes based on pushing and grasping skills collaboration[J]. Robot, 2026, 48(1): 1 – 10, 37. (in Chinese)
- [2] 欧阳安, 黄慧星, 张安琪, 等. 果蔬采摘机器人研究进展与展望[J]. 农业机械学报, 2026, 57(5): 19 – 38, 49.
- Ouyang An, Huang Huixing, Zhang Anqi, et al. Research progress and prospects of fruit and vegetable harvesting robots[J]. Transactions of the Chinese Society for Agricultural Machinery, 2026, 57(5): 19 – 38, 49. (in Chinese)
- [3] Kumra S, Joshi S, Sahin F. Antipodal robotic grasping using generative residual convolutional neural network[C]//2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE, 2020: 9626 – 9633.
- [4] Tian H, Song K, Li S, et al. Lightweight pixel-wise generative robot grasping detection based on RGB – D dense fusion[J]. IEEE Transactions on Instrumentation and Measurement, 2022, 71: 1 – 12.
- [5] Cao H, Chen G, Li Z, et al. Efficient grasp detection network with Gaussian-based grasp representation for robotic manipulation[J]. IEEE/ASME Transactions on Mechatronics, 2022, 28(3): 1384 – 1394.
- [6] Yu S, Zhai D H, Xia Y. Robotic grasp detection based on category-level object pose estimation with self-supervised learning[J]. IEEE/ASME Transactions on Mechatronics, 2023, 29(1): 625 – 635.
- [7] Yu S, Zhai D H, Xia Y, et al. Robotic grasp detection with 6-D pose estimation based on graph convolution and refinement[J]. IEEE Transactions on Systems, Man, and Cybernetics: Systems, 2024, 54(6): 3783 – 3795.
- [8] 马程宏, 谭励, 曾伟, 等. 基于视触觉感知的番茄尺寸和姿态解析方法[J]. 农业机械学报, 2024, 55(1): 223 – 232.
- Ma Zenghong, Tan Li, Zeng Wei, et al. Analytical method and evaluation of tomato size and posture based on visual and tactile perception[J]. Transactions of the Chinese Society for Agricultural Machinery, 2024, 55(1): 223 – 232. (in Chinese)
- [9] Liang H, Ma X, Li S, et al. PointNetGPD: detecting grasp configurations from point sets[C]//2019 IEEE International Conference on Robotics and Automation (ICRA). IEEE, 2019: 3629 – 3635.
- [10] He R, Chen H, Fu M. Interest point selection and feature extraction in six-DoF grasp pose detection[J]. IEEE/ASME Transactions on Mechatronics, 2024, 29(6): 4723 – 4733.
- [11] Zhuang C, Wang H, Niu W, et al. A parallel graph network for generating 7-DoF model-free grasps in unstructured scenes using point cloud[J]. Robotics and Computer-Integrated Manufacturing, 2025, 92: 102879.
- [12] Tang W, Tang K, Zi B, et al. High precision 6-DoF grasp detection in cluttered scenes based on network optimization and pose propagation[J]. IEEE Robotics and Automation Letters, 2024, 9(5): 4407 – 4414.
- [13] 黄玲涛, 王琳轩, 孔紫静, 等. 面向无模型物体的机器人视觉-力觉协同交互方法[J]. 农业机械学报, 2025, 56(10): 770 – 779.
- Huang Lingtao, Wang Linxuan, Kong Zijing, et al. Vision-force collaborative interaction method for model-free objects in robotic systems[J]. Transactions of the Chinese Society for Agricultural Machinery, 2025, 56(10): 770 – 779. (in Chinese)
- [14] Wang C, Fang H S, Gou M, et al. Graspness discovery in clutters for fast and accurate grasp detection[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). 2021: 15964 – 15973.
- [15] Gui H, Pang S, He X, et al. High-performance grasp pose detection via point cloud serialization attention[J]. Pattern Recognition, 2026, 171: 112099.
- [16] Yu S, Yin J, Zhai D H, et al. Fast and accurate category-level object pose estimation without shape priors for robotic grasp detection[J]. IEEE Transactions on Automation Science and Engineering, 2025, 22: 17696 – 17709.
- [17] Li Y, Kong T, Chu R, et al. Simultaneous semantic and collision learning for 6-DoF grasp pose estimation[C]//2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE, 2021: 3571 – 3578.
- [18] Yu S, Zhai D H, Xia Y. MTGrasp: multiscale 6-DoF robotic grasp detection[J]. IEEE/ASME Transactions on Mechatronics, 2024, 30(1): 156 – 167.
- [19] Niu M, Lu Z, Chen L, et al. Vergnet: visual enhancement guided robotic grasp detection under low-light condition[J]. IEEE Robotics and Automation Letters, 2023, 8(12): 8541 – 8548.
- [20] Tian H, Song K, Xu J, et al. Antipodal-points-aware dual-decoding network for robotic visual grasp detection oriented to multi-object clutter scenes[J]. Expert Systems with Applications, 2023, 230: 120545.
- [21] Lu Y, Deng B, Wang Z, et al. Hybrid physical metric for 6-DoF grasp pose detection[C]//2022 IEEE International Conference on Robotics and Automation (ICRA). IEEE, 2022: 8238 – 8244.
- [22] Hoang D C, Stork J A, Stoyanov T. Context-aware grasp generation in cluttered scenes[C]//2022 IEEE International Conference on Robotics and Automation (ICRA). IEEE, 2022: 1492 – 1498.
- [23] Shi M, Hou J, Li Z, et al. NG – Net: no-grasp annotation grasp detection network for stacked scenes[J]. Journal of Intelligent Manufacturing, 2025, 36(2): 1477 – 1490.
- [24] Yu S, Zhai D H, Xia Y. CGNet: robotic grasp detection in heavily cluttered scenes[J]. IEEE/ASME Transactions on Mechatronics, 2022, 28(2): 884 – 894.
- [25] 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 (CVPR), 2020: 11444 – 11453.