计算机科学
COMPUTER SCIENCE

基于GraspNet的多阶段无序混装抓取方法

于灵鑫，陈艺博，曲浩君，厉广伟，李金屏

引用本文

于灵鑫，陈艺博，曲浩君，厉广伟，李金屏. 基于GraspNet的多阶段无序混装抓取方法[J]. 计算机科学, 2026, 53(4): 318-325.

YU Lingxin, CHEN Yibo, QU Haojun, LI Guangwei, LI Jinping. Multi-stage Grasping Method for Unordered Mixed Objects Grasping Based on GraspNet [J]. Computer Science, 2026, 53(4): 318-325.

相似文章推荐（请使用火狐或 IE 浏览器查看文章）

Similar articles recommended (Please use Firefox or IE to view the article)

基于深度学习的刚体位姿估计方法综述

Survey of Rigid Object Pose Estimation Algorithms Based on Deep Learning

计算机科学, 2023, 50(2): 178-189. https://doi.org/10.11896/jsjkx.211200164

基于大型场景下的多相机标定方法

Multi-camera Calibration Method Based on Large-scale Scene

计算机科学, 2022, 49(11A): 211200054-6. https://doi.org/10.11896/jsjkx.211200054

基于半直接方法的序列影像直线特征跟踪匹配算法

Line Tracking and Matching Algorithm Based on Semi-direct Method in Image Sequence

计算机科学, 2019, 46(6A): 270-273.

多通道用户界面中的目标选择技术

计算机科学, 2000, 27(1): 48-50.

基于相似关系的社会集合论

Social Set Theory Based on Similarity Relations

计算机科学, 2013, 40(5): 54-57.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.250600124

基于 GraspNet 的多阶段无序混装抓取方法

于灵鑫　陈艺博　曲浩君　厉广伟　李金屏
济南大学信息科学与工程学院　济南 250022
山东省网络环境智能计算技术重点实验室(济南大学)　济南 250022
山东省"十三五"高校信息处理与认知计算重点实验室(济南大学)　济南 250022
(1379993737@qq.com)

摘　要　用于工业分拣领域的机械装置通常是针对特定应用场景和特定产品而设计的,面对多种物品无序堆叠的场景,其普适性和智能性往往较差。当前基于 3D 结构光相机的点云匹配抓取技术虽在一定程度上提升了柔性生产能力,但受限于硬件成本高昂,以及特征描述能力有限、计算复杂度高、对遮挡敏感等固有缺陷,难以满足无序混装抓取需求。近年来以 GraspNet 为代表的深度学习抓取技术发展迅速,通过双目相机实现位姿估计,但仍存在目标选择策略欠优、位姿评分机制具有局限性、位姿定位偏差大等问题。针对上述挑战,提出一种改进型三阶段抓取算法。第一阶段,针对目标选择策略欠佳的问题,通过融合 YOLOv10 目标检测与 SAM 分割模型,结合优化的目标选择算法,即选择无遮挡、距离近的目标,有效解决了堆叠遮挡场景下的目标选择策略不佳难题。第二阶段,对 GraspNet 位姿估计框架进行改进,即通过引入基于点云表面法向量的位姿筛选机制,重构更加合理的评分机制,进而获取高精度抓取位姿。第三阶段,设计位姿微调策略,即采用"悬停对齐-垂直抓取"的分层控制架构,最大程度消除执行过程中的累积误差,有效解决位姿定位偏差大、实际抓取不准确问题。实验结果表明,该方法显著提升了复杂场景下的抓取效率、操作可靠性和跨场景泛化能力,同时由于使用双目相机取代了 3D 结构光相机,还显著降低了系统成本,为工业自动化提供了高性价比的解决方案。

关键词:无序混装抓取;位姿估计;目标选择;姿态优化;双目相机
中图分类号　TP242

Multi-stage Grasping Method for Unordered Mixed Objects Grasping Based on GraspNet

YU Lingxin,CHEN Yibo,QU Haojun,LI Guangwei and LI Jinping
School of Information Science and Engineering,University of Jinan,Jinan 250022,China
Shandong Provincial Key Laboratory of Network Based Intelligent Computing(University of Jinan),Jinan 250022,China
Shandong College and University Key Laboratory of Information Processing and Cognitive Computing in 13th Five-year(University of Jinan),Jinan 250022,China

Abstract　Mechanical devices used in industrial sorting are typically designed for specific application scenarios and products,often exhibiting poor versatility and intelligence when faced with unordered mixed object grasping. Current point cloud matching grasping technologies based on 3D structured light cameras have improved flexible production capabilities to a certain extent. However,they are constrained by high hardware costs,limited feature description capabilities,high computational complexity,and sensitivity to occlusions,making it difficult to meet the demands of unordered mixed object grasping. In recent years,deep learning-based grasping technologies,represented by GraspNet,have developed rapidly,achieving pose estimation through binocular cameras. Nevertheless,these methods still suffer from suboptimal target selection strategies,limitations in pose scoring mechanisms,and significant pose localization errors. To address these challenges,this study proposes an improved three-stage grasping algorithm. In the first stage,the YOLOv10 object detection model is fused with the SAM segmentation model,combined with an optimized target selection algorithm that prioritizes unobstructed and closer targets,effectively solving the problem of poor target selection strategies in stacked and occluded scenarios. In the second stage,the GraspNet pose estimation framework is enhanced

到稿日期:2025-06-19　返修日期:2025-08-19
基金项目:山东省科技型中小企业创新能力提升工程(2022TSGC1047);中央引导地方科技发展项目(YDZX2024078);济南大学 2023 年学科交叉会聚建设项目(XKJC-202310)
This work was supported by the Shandong Provincial Project of Innovation Ability Enhancement Engineering for Technology Oriented Small and Medium-sized Enterprises(2022TSGC1047),Central Guidance Funding Projects for Local Scientific and Technological Development of Shandong Province(YDZX2024078) and University of Jinan Disciplinary Cross-Convergence Construction Project 2023 (XKJC-202310).
通信作者:李金屏(ise_lijp@ujn.edu.cn)

by introducing a pose filtering mechanism based on point cloud surface normals and reconstructing the scoring mechanism to obtain high-precision grasping poses. In the third stage, a pose fine-tuning strategy is designed using a hierarchical control architecture of "hover alignment-vertical grasping" to effectively eliminate cumulative errors during execution, ultimately addressing the issue of inaccurate real-world grasping. Experimental results demonstrate that this method significantly improves grasping efficiency, operational reliability, and cross-scenario generalization capabilities in complex environments. Moreover, by replacing 3D structured light cameras with binocular cameras, the system cost is significantly reduced, providing a cost-effective solution for industrial automation.

Keywords Unordered mixed objects grasping, Pose estimation, Target selection, Pose optimization, Binocular camera

1 引言

在工业自动化进程中，机器人技术的飞速发展显著提升了生产效率和操作灵活性[1]。然而，在实际应用场景中，无序混装环境（见随机堆叠的工业零件图 1(a)、待分拣的快递包裹图 1(b)）对传统固定编程方式提出了严峻挑战。这类场景具有物品种类多样、空间分布随机且存在大量遮挡的特点，要求机器人具备高度的自适应能力和智能决策能力，以高效完成抓取任务。

(a)待分拣的工业活塞　　(b)待分拣快递

图 1　无序混装场景示意图

Fig. 1　Diagram of unordered mixed objects

抓取检测作为机器人抓取操作的核心技术，其性能直接影响机械臂操作的可靠性。其中，目标物品的位姿估计精度或最优抓取位姿的获取是关键挑战之一。视觉传感技术凭借其在环境感知与位姿解算中的独特优势，长期以来受到计算机视觉与机器人学研究者的广泛关注[2]。然而，在无序混装场景中，物品之间存在大量的遮挡且动态变化频繁[3]，现有方法的精度和鲁棒性面临巨大的挑战。

早期研究主要依赖高精度结构光传感器与多视图几何方法，但结构光相机价格昂贵。基于匹配点的 3D 目标位姿估计是广泛应用的方法之一，基于 2D-3D 的位姿估计方法是通过离线阶段标注 3D 模型的关键点，与在线阶段检测到的 2D 图像中的关键点匹配，结合 PnP 算法[4]计算物品的 6 自由度(6-DoF)位姿。常用的 2D 特征描述子包括 SIFT[5]，ORB[6]，HOG[7]等。然而这类方法依赖纹理特征，易受光照条件影响[8]，尤其在处理缺乏纹理的工业零件时表现有限。基于 3D-3D 的方法则在在线阶段直接匹配点云特征，通过 ICP[9]算法得到精确的位姿。常用的 3D 特征描述子包括 SHOT[10]，FPFH[11]，Spin-Image[12]等。然而，这类方法对点云质量和计算资源需求高，且在噪声、遮挡及动态场景下的鲁棒性仍存在局限。

近年来，随着深度学习理论的突破性进展和大规模标注数据集（如 Cornell[13]，Jacquard[14]，YCB-Video[15]，GraspNet-1Billion[16]）的构建，基于数据驱动的位姿估计范式取得了显著优势[17]，端到端的神经网络架构（如 PointNet＋＋[18]，VoxelNet[19]，PVNet[20]和 GraspNet[16]）能够直接从原始点云数据中学习具有判别性的特征，为机器人抓取任务提供了更为可靠的技术路径。

GraspNet 作为基于深度学习位姿估计的代表性工作，其抓取检测结果在单一应用场景中展现了良好的泛化性能，这一表现很大程度上得益于 GraspNet-1Billion 数据集的超大规模特性。然而，在无序混装场景的应用中，观察到以下若干问题亟待解决。

1）GraspNet 目标选择策略欠优。主要原因在于过度依赖位姿评分机制，但在无序混装场景下，完整 RGB-D 信息更为关键。大量堆叠遮挡情况下，不同物品的抓取优先级应与其 RGB-D 信息完整性密切相关，信息越完整越适合作为抓取目标。然而当前策略未能充分体现这一特性，限制了其合理性和有效性。

2）GraspNet 的位姿评分机制在无序混装场景中表现出一定的局限性。主要体现在其未能有效考量周围物品对抓取位姿的约束与干扰，增加了夹爪在执行抓取任务时的碰撞风险。

3）机械臂实际执行的位姿与模型输出存在一定的偏差。传感器误差和标定误差等因素，导致最终的抓取位姿与预期存在偏差，影响了最终的抓取成功率。

针对上述问题，本文提出了一种基于 GraspNet 的多阶段无序混装抓取方法。本文的主要贡献如下：

1）优化目标选择算法：提出融合 YOLOv10 目标检测与 SAM 分割模型，结合优化的目标选择算法，通过优先选择无遮挡且距离较近的目标，有效解决了堆叠遮挡场景中目标选择策略不佳的问题。

2）重构抓取姿态选择机制：对 GraspNet 位姿估计框架进行改进，通过点云特征解耦与抓取亲和力场预测的协同优化，设计基于点云表面法向量的姿态筛选机制，重构位姿评分策略，从而获取高精度抓取位姿。

3）设计位姿微调策略：采用“悬停对齐-垂直抓取”的分层控制架构，有效消除执行过程中的累积误差，最终解决实际抓取不准确问题。

2 预备知识

2.1 YOLO 和 SAM 原理

目标检测和分割是机器人视觉感知中的两项关键技术，可以将感兴趣目标从复杂背景中分离出来。YOLO[21]（You Only Look Once）是一种高效的目标检测算法，其核心思想是通过单次前向传播，同时预测目标的边界框和类别。相比传统的两阶段检测方法（如 Faster R-CNN[22]），YOLO 具有更快的推理速度以及更成熟的工业相关产品，适合实时性要求较高的场景。本文采用的 YOLOv10[23] 在前代版本的基础上进一步提升了检测精度和鲁棒性。

为了进一步提升目标提取的精度，本文引入了 SAM[24]（Segment Anything Model）。SAM 是一种基于提示的通用分割模型，能够根据用户输入（如点、框或文本提示）生成高质量的像素级分割掩码（Mask）。相比传统分割方法，SAM 具有更强的泛化能力，能够在未见过的复杂场景中实现零样本分割。

本文利用 YOLOv10 检测到的目标边界框作为 SAM 的提示输入，从而实现对目标物品的精确分割。这种结合方式充分发挥了 YOLO 的高效检测能力和 SAM 的高精度分割能力，为后续的目标选择和抓取任务提供了可靠支持。

2.2 GraspNet

GraspNet 位姿估计框架[16] 流程如图 2 所示。其网络架构采用 PointNet++ 主干网络提取点云特征，首先，通过 Approach Network 基于 300 个预定义视角预测可行接近方向；其次，Operation Network 结合分类与回归策略，精确解算夹爪平面内旋转（12 角度分箱）与宽度参数；最后，Tolerance Network 通过最大容忍偏移距离预测增加强抓取抗干扰能力。Tolerance Network 通过预测最大容忍偏移距离，增强抓取对环境扰动的抗干扰能力，提升实际抓取位姿预测的精度。

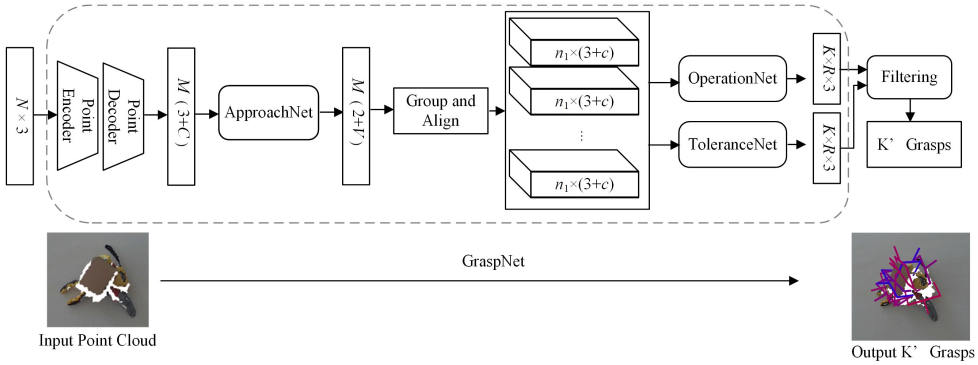


图 2 GraspNet 框架

Fig. 2 Framework of GraspNet

3 基于 GraspNet 的多阶段无序混装抓取方法

为解决无序混装场景中存在的目标选择策略欠优、位姿评分机制具有局限性、位姿定位偏差大等问题，本文提出了一种基于 GraspNet 的多阶段无序混装抓取方法。算法流程如图 3 所示，接下来将从最优目标选择、抓取姿态优选、姿态微调 3 个方面介绍本文算法。

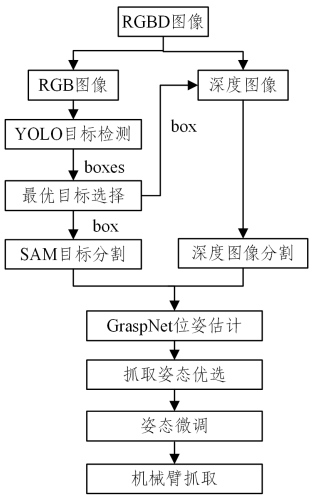


图 3 基于 GraspNet 的多阶段无序混装抓取方法流程图

Fig. 3 Flowchart of multi-stage unordered mixed grasping method based on GraspNet

3.1 最优目标选择算法

为选择无遮挡、距离近的目标，有效解决堆叠遮挡场景下的目标选择策略不佳难题，本文改进了最优目标选择算法，结合最顶部和最离群两种规则，先选择离群目标再选择最顶部目标，以减小抓取时夹爪碰撞其他目标的风险。

假设检测到的边界框列表为 $B=\{b_1, b_2, \cdots, b_n\}$，其中 n 为 YOLOv10 目标检测到的边界框的总个数，每个边界框 b_i 由其坐标 $(x_{\min,i}, y_{\min,i}, x_{\max,i}, y_{\max,i})$ 表示；深度图 D 是一个二维数组，每个像素 $D_{x,y}$ 值表示该位置到相机的深度（单位：毫米）。

3.1.1 选择最离群的目标

最离群的目标是指独立性较强的目标，通常与物品堆完全分离。这类目标往往不受遮挡，其几何特征清晰可见，便于机械臂抓取。同时，抓取最离群的目标不会对其他工件造成干扰，减少了碰撞其他目标的风险。

算法的核心是通过计算每个边界框与其他边界框的最小距离，找出距离其他边界框最远的边界框。具体的步骤如下：

1）对于每个边界框 b_i，计算其中心点 (x_i, y_i)：$x_i=\frac{x_{\min,i}+x_{\max,i}}{2}$，$y_i=\frac{y_{\min,i}+y_{\max,i}}{2}$

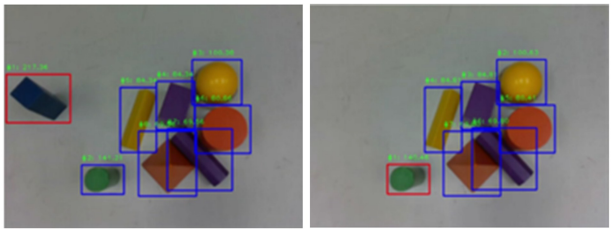
2）对于每个边界框 b_i，计算其与其他边界框的最小距离 d_i：

$$d_i=\min_{j\neq i}\sqrt{(x_i-x_j)^2+(y_i-y_j)^2}$$

其中，j 遍历所有的其他边界框。

3）选择具有最大最小距离的边界框 b_{outlier}：$b_{\text{outlier}} = \arg\max_i d_i$

利用上述步骤，对 RGB-D 图像进行目标检测，并选择最离群的目标。如图 4 所示，将不同形状、大小的积木块随机摆放在工作台上，进行最离群目标选择后依次将最离群目标移走，选择的顺序如图 4(a)、图 4(b)所示，红色框被认为是最离群的目标。



(a)第一顺序 (b)第二顺序

图 4 离群目标优选测试图(电子版为彩图)

Fig. 4 Diagram of outlier target priority selection test

3.1.2 选择最顶部的目标

最顶部的目标位于堆叠结构的上层，受其他目标遮挡的程度最小，其 RGB 图像和深度信息最为完整，最有利于位姿估计。此外，抓取最顶部的目标对下方工件的影响也最小，能够有效避免抓取操作导致的碰撞。具体步骤如下：

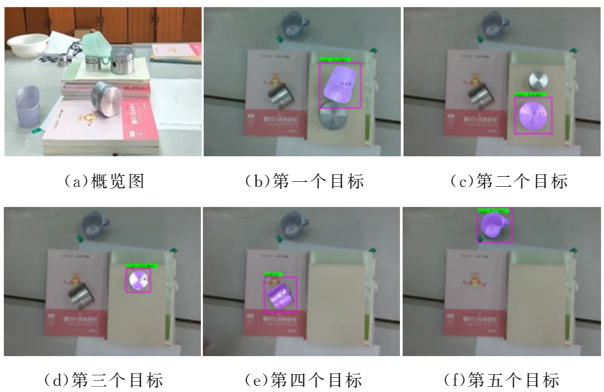
1）初始化最小深度 $d_{\min} = \infty$，最顶端边界框 $b_{\text{top}} = \text{None}$。

2）对于每个边界框 b_i，提取深度图对应的区域 $D_i = \{D_{x,y} \mid x_{\min,i} \leqslant x \leqslant x_{\max,i}, y_{\min,i} \leqslant y \leqslant y_{\max,i}\}$。

3）过滤掉无效的深度值(0 或超出量程的值)得到 D_i'，如果 D_i' 为空，则跳过该边界框；否则，计算该边界框的最小深度值 $d_i = \min(D_i')$。

4）如果 $d_i < d_{\min}$，则 $d_{\min} = d_i$，$b_{\text{top}} = b_i$。

利用上述步骤，对 RGB-D 图像进行目标检测，并选择最顶部的目标。如图 5 侧视图所示，将不同大小的活塞、杯托等多个目标错落有致地摆放在工作台上，进行最顶部目标选择(红色框被认为是最顶部目标)后依次将最顶部目标移走，选择的顺序如图 5(b)－图 5(f)所示。



(a)概览图 (b)第一个目标 (c)第二个目标

(d)第三个目标 (e)第四个目标 (f)第五个目标

图 5 顶部目标优选测试图(电子版为彩图)

Fig. 5 Diagram of top target priority selection test

3.2 抓取姿态优选机制

GraspNet 模型输出大量候选抓取位姿后，采用非极大值抑制(Non-Maximum Suppression，NMS)结合 Tolerance Network 的力闭合分析(Force-closure Analysis)，通过计算抓取姿态在不同方向上可承受的最大位移扰动，对候选位姿进行排序并选择得分最高的抓取方案。该方法在非堆叠物品场景中表现良好，但在无序混装场景中，高容忍度的抓取位姿可能覆盖超出目标物品的空间范围，导致夹爪在执行过程中碰撞非目标物品。

为提升无序混装抓取姿态的鲁棒性与适配性，本文通过点云特征解耦与抓取亲和力场预测的协同优化，设计了一种基于点云表面法向量和接近向量(见图 6)夹角结合抓取点几何中心约束的姿态筛选规则。具体步骤如下：

1)物品表面法向量估计。对于每个候选抓取点，通过点云局部几何特征估计其表面法向量 $\boldsymbol{n}$。该法向量通过随机采样一致性算法(RANSAC[25])拟合局部平面获得，并经过方向校正以确保校正法向量 $\boldsymbol{n}_{\text{corrected}}$ 与全局坐标系的桌面垂直方向一致：

$$\boldsymbol{n}_{\text{corrected}} = \begin{cases} \boldsymbol{n}, & \text{if } \boldsymbol{n} \cdot (0,0,1) < 0 \\ -\boldsymbol{n}, & \text{otherwise} \end{cases} \tag{1}$$

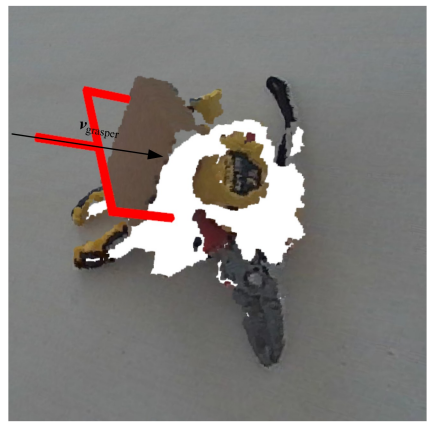


图 6 接近向量示意图

Fig. 6 Diagram of approach vector

2)计算接近向量 $\boldsymbol{v}_{\text{grasper}}$ 与法向量 $\boldsymbol{n}_{\text{corrected}}$ 的对齐程度，定义为两向量夹角的余弦值：

$$Score_{\text{align}} = \frac{\boldsymbol{v}_{\text{grasper}} \cdot \boldsymbol{n}_{\text{corrected}}}{\| \boldsymbol{v}_{\text{grasper}} \| \cdot \| \boldsymbol{n}_{\text{corrected}} \|} \tag{2}$$

对齐评分越高，表明抓取姿态越垂直于物品表面。这一评分反映了抓取姿态与物品表面的匹配程度，是抓取稳定性的重要指标。

3)几何中心约束。为了确保抓取点靠近物品的几何中心，计算模型输出的候选抓取点位置 p_{grasp} 与目标物品点云几何中心 $\boldsymbol{c}$ 的距离偏差：

$$d = \| \boldsymbol{p}_{\text{grasp}} - \boldsymbol{c} \|_2 \tag{3}$$

其中，$\boldsymbol{c}$ 为目标物品点云所有点的平均位置：

$$\boldsymbol{c} = \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{p}_i, \boldsymbol{p}_i \in \text{目标点云} \tag{4}$$

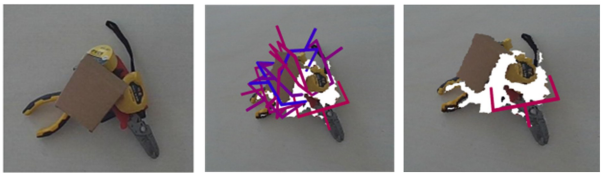
4)结合对齐评分与距离偏差，最终筛选得分定义为：

$$Final\ Score = \lambda \cdot Score_{\text{align}} + (1-\lambda) \cdot e^{\left(-\frac{d^2}{2\sigma^2}\right)} \tag{5}$$

其中，λ 为平衡系数，建议取 0.7；σ 为距离衰减系数(通常设为目标物品直径的 10%)。这一约束确保抓取点靠近物品的几何中心，从而提高抓取的稳定性。

通过上述规则，优先选择最垂直于物品表面且靠近几何

中心的抓取点，这一综合评分机制综合考虑了抓取姿态的局部几何特征和全局几何约束，显著提升了抓取姿态的精准性与稳定性，更符合无序混装抓取任务。图 7(b)所示为 GraspNet 的输出的候选姿态和最优抓取姿态，图 7(c)为本文改进评分机制得到的候选姿态和最优抓取姿态。



(a)RGB 图 (b)GraspNet 输出候选和最优

(c)本文输出候选和最优

图 7 姿态筛选机制测试图

Fig. 7 Diagram of pose filtering mechanism test

3.3 姿态微调策略

针对机械臂抓取任务中的标定和误差传递导致的位姿偏差问题，提出一种基于图像反馈的两阶段位姿调整策略。该方法通过“悬停对齐-垂直抓取”的分层控制架构，在保持系统简洁性的同时显著提升抓取精度。具体实现流程如下：

1)初始位姿规划：利用目标检测模型(如 YOLOv10)输出的抓取位姿参数，将机械臂末端执行器(夹爪)移动至目标物品正上方的预设安全高度(Z 方向偏移量 $\Delta z=150$ mm)。该高度基于目标物品高度分布统计特性确定，确保夹爪在垂直方向与目标物品无干涉。

2)视觉对齐校正：在悬停状态下，通过手眼标定相机同步获取夹爪与目标物品的图像。利用上一步得到的目标检测框计算目标图像中心(u_t，v_t)，预先定义夹爪中心坐标(u_g，v_g)，计算图像平面偏差 $\Delta u=u_t-u_g$，$\Delta v=v_t-v_g$。

3)平面位姿补偿：通过相机内参矩阵将像素偏差转换为基坐标系下的平移补偿量：

$$\Delta x=\frac{\Delta u\cdot d\cdot k_x}{f_x}$$
$$\Delta y=\frac{\Delta v\cdot d\cdot k_y}{f_y} \tag{6}$$

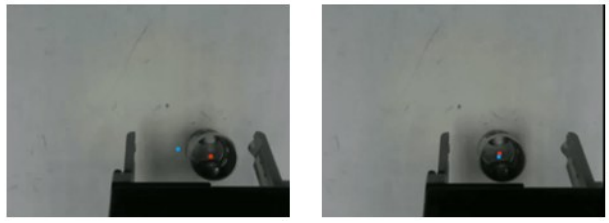
其中，d 为当前悬停高度；k_x，k_y 为像素当量系数；f_x，f_y 为相机焦距。驱动机械臂进行水平位移补偿，实现夹爪与目标的空间对齐。

4)安全下降与动态调整：完成对齐后，机械臂以匀速垂直下降。在下降过程中持续采集图像数据，当检测到夹爪与目标物品的相对像素偏差超过阈值($\Delta u^2+\Delta v^2>5$ 像素^2)时，触发动态补偿机制，通过比例控制律实时调整末端位姿。

$$\Delta x'=k_p\Delta u$$
$$\Delta y'=k_p\Delta v \tag{7}$$

其中，k_p 为控制系数。通过实时调整，确保夹爪在接近目标时能够精确对齐。

步骤 1)－步骤 3)为悬停对齐控制阶段，步骤 4)为垂直抓取控制阶段。标定误差和传递误差导致的姿态偏差如图 8(a)所示，经本文方法改进后如图 8(b)所示，消除了上述偏差。



(a)姿态偏差图 (b)姿态修正图

注：红色点为物品中心，蓝色点为抓取点位置。

图 8 姿态修正测试图(电子版为彩图)

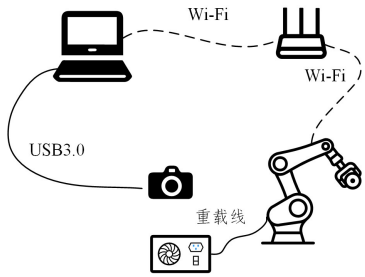
Fig. 8 Diagram of pose correction test

4 实验结果与分析

4.1 实验硬件平台

搭建的机械臂无序混装抓取系统硬件平台如图 9(a)所示，主要由睿尔曼机械臂、RealSense D435i 相机、二指夹爪、上位机、路由器组成，其规格参数如表 1 所列，其连接关系为无序混装抓取系统拓扑，如图 9(b)所示。

(a)硬件平台



(b)系统拓扑

图 9 机械臂无序混装抓取系统硬件平台

Fig. 9 Robotic arm unordered mixed grasping system hardware platform

表 1 实验平台硬件配置

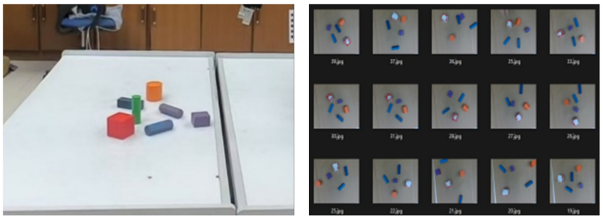
Table 1 Experimental platform hardware configuration

硬件	规格参数	数量	安装位置
机械臂	RML63-6F，6DOF	1	固定在工作台
相机	D435i，640 * 480@30 fps	1	“眼在手上”
夹爪	EG2-5E 最大开度 85 mm	1	机械臂末端
上位机	i7-12700H，RTX3070Ti	1	
路由器	CR6606，Wi-Fi6	1	

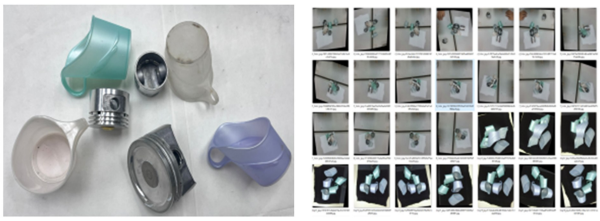
4.2 实验细节和实验数据

为了验证本文方法的泛化能力，实验中选用了两类不同形状和材质的物体作为研究对象。第一类为规则几何形状的

物体，包括长方体、正方体和圆柱体，如图 10(a)和图 10(b)所示，材质为合成木质材料；第二类为形状不规则且对光照敏感的物体，包括活塞和杯托，如图 10(c)和图 10(d)所示。所有实验中，堆叠的最大高度统一设置为 15 cm。



(a)实验工件示意图一　(b)部分数据集示例一



(c)实验工件示意图二　(d)部分数据集示例二

图 10　数据集示意图

Fig. 10　Diagram of datasets

本文在无序混装场景中构建了两种数据集(规则物体和不规则物体)，每种数据集都包含 400 张图像，所有图像通过“眼在手上”的 D435i 相机统一采用 640×480 分辨率在不同距离、角度下拍摄，涵盖上述物品多种姿态，并使用 LabelImg 进行精确边界框标注，使用 NVIDIA RTX 3070 Ti GPU 与 PyTorch 框架训练，初始学习率为 0.01，采用 SGD 优化器，批量大小为 20，经过 150 轮训练，得到 YOLOv10 目标检测模型。

对于 RANSAC(随机抽样一致性算法)，设置两个实验参数。一个是迭代次数，本文设置为 100。其原因在于，在无序混装场景中，由于遮挡、点云稀疏性等问题，可能无法获得完整的表面点云数据。因此，100 次迭代能够平衡计算效率与稳定性，在大多数情况下能提供可靠的法向量估计结果。另一个参数是采样半径 r，本文通过目标检测框的 x 方向或 y 方向的变化辅助动态调整采样半径，如图 11 所示。其思路如下：

1)设边界框左下角为(x_1,y_1)，右上角为(x_2,y_2)，则其 x 方向长度为：

$$d_{\text{x\_pixel}}=|x_2-x_1| \tag{8}$$

2)根据相机成像原理，将像素尺度转换为实际空间中的物理尺度。假设边界框中心点对应的深度值为 D(单位：mm)，相机内参为 f_x，则可估算出边界框的实际长度为：

$$d_{x\_\text{real}}=\frac{d_{\text{pixel}}\cdot D}{f_x} \tag{9}$$

3)同理可得 d_{y_real}。

4)取 x 方向和 y 方向中实际距离较小的值作为参数调整的依据：

$$d_{\text{real}}=\min(d_{x\_\text{real}},d_{y\_\text{real}}) \tag{10}$$

5)根据实际 d_{real} 长度设置 RANSAC 半径 r，通过引入比例系数 $\alpha(\alpha=\frac{1}{4})$，使其在保留局部几何特征(如边缘、角点)的同时，有效抑制了噪声和异常点的影响。采样半径与目标尺寸成正比：

$$r=\alpha d_{\text{real}} \tag{11}$$

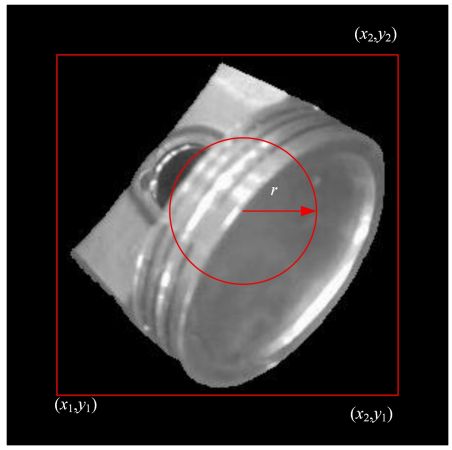


图 11　采样半径示意图

Fig. 11　Schematic diagram of sampling radius

4.3　实验评价指标

本研究为验证无序混装抓取系统的实际应用性能表现，采用抓取成功率 g_{succeed} 作为核心量化指标。

$$g_{\text{succeed}}=\frac{g_s}{g_s+g_f}\times 100\% \tag{12}$$

其中，g_s 代表成功抓取的次数，g_f 代表抓取失败的次数。在抓取实验中，如果成功夹住物品并将其放到指定区域，且物品没有掉落，则为抓取成功，反之则为抓取失败。

4.4　对比实验

为验证本文提出的基于 GraspNet 的多阶段无序混装抓取方法的有效性及其对不同形状物体的泛化能力，本文在真实物理环境中分别开展了针对规则物体、不规则物体以及混合物体(规则与不规则物体混合)的抓取实验，并在相同实验条件下与原始 GraspNet 方法进行对比测试。每轮实验包含 50 次独立抓取操作，每次抓取前均对物体进行随机摆放，以模拟实际生产中的无序混装场景。每类物体的实验共进行 5 轮，累计完成 250 次抓取；3 类物体共计完成 15 轮、总计 750 次抓取实验，全面评估了所提方法在复杂抓取场景下的稳定性与适应性。

实验结果如表 2 所列，结果表明，本文提出的多阶段无序混装抓取方法在各类抓取任务中均显著优于原始 GraspNet 方法。在总计 750 次抓取实验中，GraspNet 的成功率为 67.1%，而本文方法达到 85.7%，提升了 18.6 个百分点，表现出更强的整体性能。在不同物体类型的任务中，两种方法表现出一致的趋势：在规则物体抓取任务中，GraspNet 的成功率为 69.2%，本文方法提升至 92.4%，提升了 23.2 个百分点；在不规则物体抓取任务中，部分物体存在表面反光、深度信息缺失或几何结构复杂等问题，导致点云数据质量下降，GraspNet 的表现相较于抓取规则物体时有所下降，但本文方法仍实现了 82.4%的抓取成功率，表现出明显优势，验证了其在面对复杂形状、光照敏感及堆叠遮挡等挑战时依然具备

较强的适应能力与抓取稳定性；在规则与不规则物体混合抓取任务中，GraspNet 的成功率为 66.4%，而本文方法仍能达到 82.4%，进一步说明本文方法具备良好的泛化能力与环境适应性。

表 2 无序混装抓取对比实验结果

Table 2 Experimental comparison results of unordered mixed grasping

实验轮次	抓取次数	Graspnet 成功次数	Graspnet 成功率/%	本文成功次数	本文成功率/%
1	50	35	70.0	47	94.0
2	50	33	66.0	46	92.0
3	50	37	74.0	48	96.0
4	50	33	66.0	44	88.0
5	50	35	70.0	46	92.0
总计 1	**250**	**173**	**69.2**	**231**	**92.4**
6	50	30	60.0	40	80.0
7	50	31	62.0	41	82.0
8	50	38	76.0	41	82.0
9	50	32	64.0	39	78.0
10	50	33	66.0	45	90.0
总计 2	**250**	**164**	**65.6**	**206**	**82.4**
11	50	31	62.0	37	74.0
12	50	36	72.0	44	88.0
13	50	30	60.0	44	88.0
14	50	32	64.0	40	80.0
15	50	37	74.0	41	82.0
总计 3	**250**	**166**	**66.4**	**206**	**82.4**
总计	**750**	**503**	**67.1**	**643**	**85.7**

注：1－5 为规则物体抓取结果；6－10 为不规则物体抓取结果；11－15 为混合物体抓取结果。

综上所述，本文方法不仅在整体抓取性能上明显优于原始 GraspNet 方法，而且在面对视觉感知误差、目标遮挡和复杂堆叠等现实工业场景中的典型问题时，展现出更高的鲁棒性和实用性，具有较强的工程应用潜力。

4.5 消融实验

为深入评估本文所提出的多阶段抓取策略中各模块的功能有效性，验证优化目标选择算法、抓取姿态优选机制以及位姿微调策略在复杂场景下的实际作用，本文在无序混装抓取环境下开展了系统性的消融实验。实验基于原始 GraspNet 方法，并在规则物体的堆叠场景中逐步引入不同组合的改进模块。每组实验包含 50 次独立抓取循环，共进行 5 轮测试，结果取平均值，以提升实验数据的稳定性与可信度。

消融实验结果如表 3 所列，实验表明，本文方法在无序混装场景下的机械臂抓取任务中具有显著优势。GraspNet 单独使用时成功率为 69.2%，整体性能有限。仅引入优化的目标选择算法后，成功率提升至 79.2%，这表明本文的目标选择算法有效解决了堆叠遮挡场景中目标选择策略不佳的问题，能够有效筛选出更易抓取的目标，为后续的姿态优选和微调提供了良好的基础。仅引入优化的姿态优选机制的成功率为 85.2%，说明本文的抓取姿态的选择机制充分考虑了无序混装场景中的复杂环境问题，能有效规避周围物品对抓取位姿的干扰，显著提高了抓取操作的成功率。仅启用姿态微调策略的成功率为 76.4%，效果有限，说明双目相机的设备误差对抓取的影响较小，在进行姿态微调后可以满足抓取的需要。两两结合的抓取成功率均表现出协同作用。最终，在全部 3 个模块的共同作用下，整体抓取成功率达到了 92.4%，相较于基线提升了 23.2%，充分验证了本文多阶段抓取策略的有效性。

表 3 无序混装抓取消融实验结果

Table 3 Ablation experiment results of unordered mixed grasping

模块				抓取成功率/%
GraspNet	1	2	3	
√				69.2
√	√			79.2
√		√		85.2
√			√	76.4
√	√	√		87.6
√	√		√	81.6
√		√	√	88.0
√	√	√	√	92.4

注：1 为最优目标选择模块；2 为抓取姿态优选模块；3 为姿态微调模块。

从模块影响程度来看，抓取姿态优选模块对性能提升贡献最大，其次是最优目标选择模块，而姿态微调模块虽提升幅度较小，但对末端执行阶段的稳定性具有不可替代的作用。三者协同工作，显著提升了机械臂在无序混装抓取任务中的抓取成功率和鲁棒性。

结束语 本文针对工业无序混装抓取系统成本高昂、特征描述能力有限、计算复杂度高、对遮挡敏感等问题，提出了一种基于双目视觉的改进型多阶段抓取算法。通过融合目标优选、优化姿态优选机制与位姿微调构建多阶段无序混装抓取方法，有效解决了 GraspNet 方法在目标选择策略欠优、位姿评分机制具有局限性、位姿定位偏差大等方面的问题。实验结果表明，相较于 GraspNet 方法，本文算法在无序混装场景下实现了位姿估计准确性、抓取成功率的显著提升，消融实验进一步验证了各模块设计的有效性。该方法在硬件成本降低的同时，显著增强了在无序混装工业场景中抓取系统的鲁棒性，为柔性制造提供了高性价比的解决方案。未来研究将聚焦于多层料框、高密度堆叠等更复杂场景的扩展应用，并探索轻量化模型部署以提升实时性，最终实现工业自动化场景的全面降本增效。

参 考 文 献

[1] GUO H K. Application of Artificial Intelligence Technology in Mechanical Automation [J]. Electronic Technology, 2024, 53(10): 218-219.

[2] ZHAO Y, HUANG Q. Application of Intelligent Sensors in Industrial Automation [J]. Smart China, 2025(1): 126-128.

[3] YAN J X. Research on Robotic Sorting Technology for Stacked Parts Based on Deep Learning [D]. Hangzhou: Zhejiang University, 2024.

[4] ZHANG H J, XIONG Z, LAO D B, et al. Monocular vision measurement system based on EPNP algorithm[J]. Infrared and Laser Engineering, 2019, 48(5): 0517005.

[5] LOWE D G. Distinctive image features from scale invariant keypoints [J]. International Journal of Computer Vision, 2004, 60(2): 91-110.

[6] RUBLEE E, RABAUD V, KONOLIGE K, et al. ORB: an efficient alternative to SIFT or SURF[C]// 2011 International Con-

ference on Computer Vision. New York: IEEE, 2011: 2564-2571.

[7] DALAL N, TRIGGS B. Histograms of oriented gradients for human detection[C]// 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. New York: IEEE, 2005: 886-893.

[8] ZHANG Q P, CAO Y. Research on three-dimensional reconstruction algorithm of weak textured objects in indoor scenes [J]. Laser & Optoelectronics Progress, 2021, 58(8): 0810017.

[9] BESL P J, MCKAY N D. Method for registration of 3-D shapes [C]// Proceedings of SPIE. 1992: 586-606.

[10] TOMBARI F, SALTI S, DI STEFANO L. Unique signatures of histograms for local surface description[C]// Computer Vision—ECCV 2010. Heidelberg: Springer, 2010: 356-369.

[11] RUSU R B, BLODOW N, BEETZ M. Fast point feature histograms(FPFH) for 3D registration[C]// 2009 IEEE International Conference on Robotics and Automation. New York: IEEE, 2009: 3212-3217.

[12] JOHNSON A E. Spin-images: a representation for 3-D surface matching: CMU-RI-TR-97-47[R]. Pittsburgh: Carnegie Mellon University, 1997.

[13] JIANG Y, MOSESON S, SAXENA A. Efficient grasping from rgbd images: Learning using a new rectangle representation [C]// 2011 IEEE International Conference on Robotics and Automation. IEEE, 2011: 3304-3311.

[14] DEPIERRE A, DELLANDRÉA E, CHEN L. Jacquard: A large scale dataset for robotic grasp detection. [C]// RSJ International Conference on Intelligent Robots and Systems(IROS). IEEE, 2018: 3511-3516.

[15] XIANG Y, SCHMIDT T, NARAYANAN V, et al. Posecnn: A convolutional neural network for 6d object pose estimation in cluttered scenes[J]. arXiv: 1711.00199, 2017.

[16] FANG H S, WANG C, GOU M, et al. Graspnet-1billion: A large-scalebenchmark for general object grasping[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 11444-11453.

[17] KLEEBERGER K, BORMANN R, KRAUS W, et al. A survey on learning-based robotic grasping[J]. Current Robotics Reports, 2020, 1: 239-249.

[18] QI C R, YI L, SU H, et al. Pointnet++: Deep hierarchical feature learning on point sets in a metric space[C]// Advances in Neural Information Processing Systems. 2017.

[19] ZHOU Y, TUZEL O. Voxelnet: End-to-end learning for point cloud based 3d object detection[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 4490-4499.

[20] PENG S, LIU Y, HUANG Q, et al. Pvnet: Pixel-wise voting network for 6dof pose estimation[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019: 4561-4570.

[21] JIANG P, ERGU D, LIU F, et al. A Review of Yolo algorithm developments[J]. Procedia Computer Science, 2022, 199: 1066-1073.

[22] GIRSHICK R. Fast R-CNN[C]// Proceedings of the IEEE International Conference on Computer Vision. 2015: 1440-1448.

[23] WANG A, CHEN H, LIU L, et al. Yolov10: Real-time end-to-end object detection[J]. Advances in Neural Information Processing Systems, 2024, 37: 107984-108011.

[24] KIRILLOV A, MINTUN E, RAVI N, et al. Segment anything [C]// Proceedings of the IEEE/CVF International Conference on Computer Vision. 2023: 4015-4026.

[25] 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.

YU Lingxin, born in 2000, postgraduate. Her main research interests include pattern recognition, computer vision and robot control.

LI Jinping, born in 1968, Ph.D, professor, is a member of CCF(No. 06393S). His main research interests include artificial intelligence, pattern recognition, computer vision, digital image processing and optimization algorithms.

(责任编辑:喻藜)