分类号：TP391.41　　学校代码：10363
密　级：公开　　学　　号：2220120120

安徽工程大学
Anhui Polytechnic University

硕士学位论文

题目　基于机器视觉的工业机器人无序抓取关键技术研究

论文作者　解晓民
指导教师　王海(教授)
学科（专业）　机械工程
研究方向　机器视觉

论文提交日期：2025 年 5 月 20 日

分类号：TP391.41　　单位代码：10363

密　级：公开　　学　　号：2220120120

基于机器视觉的工业机器人无序抓取关键技术研究

RESEARCH ON KEY TECHNOLOGY OF DISORDERED GRASPING OF INDUSTRIAL ROBOTS BASED ON MACHINE VISION

学生姓名：解晓民

指导教师：王海(教授)

专　　业：机械工程

研究方向：机器视觉

论文答辩日期：2025 年 5 月 12 日

安徽工程大学学位论文原创性声明

本人郑重声明：我恪守学术道德，崇尚严谨学风。所呈交的学位论文，是本人在导师的指导下，独立进行研究工作所取得的成果。除文中已明确注明和引用的内容外，本论文不包含任何其他个人或集体已经发表或撰写过的作品及成果的内容。论文为本人亲自撰写，我对所写的内容负责，并完全意识到本声明的法律后果由本人承担。

学位论文作者签名：解晓民

日期： 2025 年 5 月 12 日

安徽工程大学学位论文版权使用授权书

学位论文作者完全了解学校有关保留、使用学位论文的规定，同意学校保留并向国家有关部门或机构送交论文的复印件和电子版，允许论文被查阅或借阅。本人授权安徽工程大学可以将本学位论文的全部或部分内容编入有关数据库进行检索，可以采用影印、缩印或扫描等复制手段保存和汇编本学位论文。

保密 □，在 __ 年解密后适用本版权书。

本学位论文属于

不保密 ☑。

学位论文作者签名：解晓民 指导教师签名：王海

日期：2025 年 5 月 12 日　　日期：2025 年 5 月 12 日

基于机器视觉的工业机器人无序抓取关键技术研究

摘要

随着智能制造的发展，工业机器人无序抓取技术成为提升自动化水平的关键挑战。目前智能制造的各种应用，如工业图像分析和处理、工业目标检测、缺陷检测、语义分割等，都在使用机器视觉，并且已经提供了相当有前景的结果。对于传统识别抓取方法难以适应复杂多变的作业场景，而基于机器视觉的无序抓取技术通过融合目标检测、姿态估计与运动规划算法，显著提升了机器人的环境感知与操作能力。因此，本研究围绕工业机器人无序抓取中的关键技术展开，结合深度学习和多模态数据融合方法，提出了一系列创新算法与实验验证方案。本研究中的主要工作有：

(1) 多模态数据集构建：通过融合 ScanNet 公共数据集与自采集的 RGB-D 数据，构建了包含 2416 张图像的多模态数据集，涵盖 10 类常见工业物体。针对彩色图与点云数据，分别采用 RoboFlow 与 CloudCompare 工具完成精细化标注，并通过高斯滤波、自适应参数变换等预处理方法提升数据质量。

(2) 单模态目标检测算法改进：基于 YOLOv8 框架，提出 AC-YOLO 网络，显著提升了工业机器人无序抓取任务的目标检测性能。网络在 Neck 模块中引入 GSConv 特征融合与 ESlim-Neck 结构，并设计三维空间注意力机制，使模型在复杂背景中表现出更强的鲁棒性与检测精度。此外，通过对模型激活函数与损失函数的优化，实现了检测速度与精度的双提升。实验表明，改进后的模型在 mAP50 指标上达到 83.6%，参数量降低至原始模型

的 54.5%，显著提升了检测精度与效率，体现了 AC-YOLO 网络识别算法在无序抓取任务上的优秀表现。

(3) 基于多模态特征的三维分割与识别算法设计：构建了 RGB-D 数据对齐的点云生成流程，并设计了基于 PointNet 与 E-ResNet 模块的三维分割网络。提出 Point RGB-D 网络，融合 RGB 图像与点云特征，结合改进的 E-ResNet 模块与稠密特征融合策略，通过多模态特征的融合和细化，本研究实现了对目标对象的高精度三维识别。在 Dice 系数和 mIoU 指标上分别达到 88.3%和 79.1%，优于主流点云分割算法：PointNet++、VoteNet 等，同时模型参数量仅为 75.3MB。实验验证表明，该方法在分割精度和效率方面均具有显著优势，为后续机器人在三维空间中的识别和操作能力提供了有力支持。

(4) 机器人智能分拣实验：详细阐述了相机内参标定、三维坐标定位以及机械臂运动学分析的相关方法。结合 RGBD Fusion 网络和 Refinement 模块的设计，本研究实现了目标物体的高精度姿态识别。通过 ROS 仿真平台搭建 UR3 机械臂控制环境，实验结果显示，系统对 10 类目标的分拣成功率平均达 86.8%，单次抓取识别耗时低于 31ms，所提出的方法在复杂环境下的目标分拣任务中具有较高的识别速度和精确性。

经检测，本研究提出的方法为工业机器人无序抓取任务提供了完整的解决方案，如网络结构搭建、实验结果对比等。在复杂场景下的目标检测、三维分割与机器人运动规划方面展现出显著优势，为智能制造中的自动化分拣系统开发提供了技术基础。

关键词：工业机器人，无序抓取，机器视觉，多模态融合，位姿估计，ROS 仿真

RESEARCH ON KEY TECHNOLOGY OF DISORDERED GRASPING OF INDUSTRIAL ROBOTS BASED ON MACHINE VISION

ABSTRACT

With the development of smart manufacturing, industrial robot disordered grasping technology has become a key challenge to enhance automation. Various applications of intelligent manufacturing, such as industrial image analysis and processing, industrial target detection, defect detection, semantic segmentation, etc., are currently using machine vision and have provided quite promising results. For traditional recognition grasping methods are difficult to adapt to complex and changing operational scenarios, while machine vision-based disordered grasping technology significantly improves the robot's environment perception and manipulation capabilities by fusing target detection, attitude estimation and motion planning algorithms. Therefore, this study focuses on the key technologies in disordered grasping for industrial robots, combining deep learning and multimodal data fusion methods, and proposing a series of innovative algorithms and experimental validation schemes.The main works in this study are:

(1) Multimodal dataset construction: By fusing the ScanNet public dataset with self-collected RGB-D data, a multimodal dataset containing 2416 images is constructed, covering 10 categories of common industrial objects. For the color map and point cloud data, RoboFlow and CloudCompare tools are used to complete the fine annotation, and Gaussian filtering, adaptive parameter transformation and other preprocessing methods are used to improve the data

quality.

(2) Improvement of unimodal target detection algorithm: Based on the YOLOv8 framework, the AC-YOLO network is proposed, which significantly improves the performance of target detection in the disordered grasping task of industrial robots. The network introduces GSConv feature fusion with ESlim-Neck structure in the Neck module, and designs a 3D spatial attention mechanism, so that the model exhibits stronger robustness and detection accuracy in complex backgrounds. In addition, by optimizing the activation function and loss function of the model, the double improvement of detection speed and accuracy is achieved. Experiments show that the improved model achieves 83.6% in the mAP50 index, and the number of parameters is reduced to 54.5% of the original model, which significantly improves the detection accuracy and efficiency, reflecting the excellent performance of AC-YOLO network recognition algorithm in the disordered grasping task.

(3) Design of 3D segmentation and recognition algorithm based on multimodal features: the point cloud generation process for RGB-D data alignment is constructed, and a 3D segmentation network based on PointNet and E-ResNet modules is designed. The Point Rgb-D network is proposed to fuse RGB image and point cloud features, combined with the improved E-ResNet module and dense feature fusion strategy, through the fusion and refinement of multimodal features, this study realizes high-precision three-dimensional recognition of target objects. It achieves 88.3% and 79.1% in Dice coefficient and mIoU indexes respectively, which is better than the mainstream point cloud segmentation algorithms: PointNet++, VoteNet, etc., and at the same time, the number of model references is only 75.3MB. experimental validation shows that the method has significant advantages in both segmentation accuracy and efficiency, which provides the subsequent robot's ability of recognizing and manipulating in the three-dimensional space with It provides strong support for

the recognition and operation ability of the robot in 3D space.

(4) Robot Intelligent Sorting Experiment: The relevant methods of camera internal parameter calibration, 3D coordinate localization and robotic arm kinematics analysis are elaborated in detail. Combined with the design of RGBD Fusion network and Refinement module, this study realizes high-precision attitude recognition of target objects. The UR3 robotic arm control environment is built through the ROS simulation platform, and the experimental results show that the system's sorting success rate for 10 types of targets reaches 86.8% on average, and the single grasping recognition consumes less than 31ms, and the proposed method has high recognition speed and accuracy in the target sorting task in complex environments.

After testing, the method proposed in this study provides a complete solution for industrial robots' disordered grasping task, such as network structure construction and comparison of experimental results. It shows significant advantages in target detection, 3D segmentation and robot motion planning in complex scenes, and provides a technical basis for the development of automated sorting systems in intelligent manufacturing.

Xie XiaoMin (Mechanical Engineering)

Supervised by Prof.Wang Hai

KEY WORDS:industrial robot, disordered grasping, machine vision, multimodal fusion, position estimation, ROS simulation

目录

插图清单

表格清单

第1章 绪论

1.1 研究背景

根据人口普查数据显示，我国正面临日益严峻的人口老龄化问题，适龄劳动人口已出现负增长，人口红利逐渐消退[1]。这一变化对我国制造业的发展，特别是劳动密集型中小企业，带来了深远影响[2]。在当前形势下，实现制造业的可持续增长已刻不容缓，而劳动密集制造业的转型升级正朝着智能制造方向迈进。将机器人技术与人工智能深度整合，能够在一定程度上缓解人口老龄化对制造业的冲击，为行业注入新动力。

为促进智能制造和机器人产业的高质量发展，我国制定了《中华人民共和国国民经济和社会发展第十四个五年规划和 2035 年远景目标纲要》（以下简称《纲要》）[3]。《纲要》明确提出，要以智能制造为主要方向，重点攻克智能感知等关键技术，深入实施智能制造工程，加快制造业的数字化、网络化和智能化转型。另一方面，《纲要》把机器人定位为现代产业核心装备和新兴技术的重要载体，凸显其在新一代信息技术深度融合环境中的关键作用，以推动产业向智能化升级迈进。文件还强调，必须加大新型机器人产品的研发和推广力度，为特定领域及细分场景提供定制化、专业化解决方案，从而打造新的产业竞争优势。

在政策支持下，我国传统的劳动密集制造业正逐步演变为以工业机器人主导的生产体系。然而，目前大多数工业机器人仍采用示教编程或离线编程，按照预设轨迹执行任务[4]。这种方式依赖工人熟练度，费时费力，不够智能，难以适应复杂多变的作业场景，尤其在抓取任务中，机器人对物体的预设位置高度依赖。当零件位置发生变化时，抓取往往失败，显著影响了工业机器人的实际应用效率。

为应对上述挑战，智能制造技术结合机器视觉的智能感知能力，为工业机器人提供了一种全新解决方案[5]。通过机器视觉，机器人能够实时感知和识别环境中的物体，无需依赖示教或离线编程，从而显著提升抓取精度与灵活性。这一视觉引导的抓取技术，可增强机器人在复杂环境中的适应性，并扩大其应用场景。然而，传统工业机器人在抓取精度、对物体位置和形状的敏感性以及对未知环境的适应能力等方面仍存在不足，制约了其推广应用。

针对这些限制，基于机器视觉的无序抓取技术应运而生[6]。通过视觉传感器（如RGB-D 相机）获取环境信息，结合图像处理、目标检测、姿态估计和抓取检测等算法，机器人能够实现对不同形状和随机放置物体的精准抓取[7]。这项技术不仅提升了机器人的智能化水平和适应能力，还拓宽了其应用范围，进一步增强了工业机器人的经济效益和社会价值[8]。

1.2 国内外研究现状

1.2.1 无序抓取技术概括

无序抓取技术旨在赋予机器人在未知物体形状、位置及姿态下，从复杂环境中自动抓取物体的能力[9]。这一技术不仅具有重要的实际应用价值，同时也涉及计算机视觉、机器人学和人工智能等多学科交叉领域的研究挑战[10]。作为一种非接触、无损的感知手段，基于机器视觉的无序抓取技术可以替代人工完成分拣等高强度任务。其核心在于通过视觉传感器获取环境信息，并结合图像处理、目标检测、姿态估计、抓取点检测以及运动规划等算法，为机器人提供精准的操作指令和实时反馈控制。这项技术的引入不仅提升了机器人的智能化水平和环境适应能力，也显著拓展了其应用场景。

图 1-1 基于机器视觉的分拣平台

图 1-1 展示了一个基于机器视觉的分拣平台。该平台能够有效降低对人力的依赖，提高检测效率与自动化水平。相较于人工视觉，机器视觉在识别精度、一致性和应对复杂多变环境的能力上表现更为出色，为分拣任务提供了更优解决方案。

当前，工业机器人无序抓取中的研究主要集中在以下几个方面：1.物体图像分割。2.物体的检测和识别。3.抓取姿态识别检测。所采用的方法主要包括以下几类：

1.基于机器学习的抓取方法：此类方法利用卷积神经网络(CNN)直接从图像中预测抓取区域或抓取点。机器学习方法能够处理多种形状和尺寸的物体，具备较强的泛化

能力。然而，这种方法依赖于大量的训练数据和高计算资源，训练过程耗时较长[11]。本次主要研究探索此方法。

2.基于强化学习的抓取方法：通过深度强化学习(DRL)，机器人可以在试错过程中学习最优抓取策略[12]。强化学习方法具有良好的适应性，能够应对动态环境和复杂任务，但其训练过程需要较长时间，并且对探索机制的效率要求较高。

3.基于分析法的抓取方法：此类方法利用物体的几何和动力学特性，构建稳定的抓取方案。基于分析的方法能够提供理论上的抓取稳定性保障和力反馈，但其效果依赖于精确的物体模型和高质量的传感器信息。

综上所述，基于机器视觉的无序抓取技术是一项具有广泛应用前景的先进技术，随着计算能力的提升和算法的不断优化，基于机器视觉的无序抓取技术有望在智能制造[14]、服务机器人[15]、医疗辅助[16]等更多领域得到广泛应用[17]。

1.2.2 语义分割研究现状

作为计算机视觉的关键研究方向之一，语义分割技术在智能化应用中扮演着重要角色。该技术已广泛应用于多个前沿领域，涵盖自动驾驶系统的环境感知、计算摄影的智能处理、人机交互的视觉理解，以及图像检索系统的核心算法构建等场景。值得关注的是，全球科技企业及创新机构均在此领域投入大量研发资源，既包括谷歌、百度等跨国科技巨头，也涌现出商汤科技(SenseTime)、旷视科技(Megvii)等专注于计算机视觉的创新企业。尤其在智能机器人与无人驾驶系统领域，该技术已发展成为构建环境感知与决策系统的核心技术支撑。

在机器人无序抓取任务中图像分割方法种类较多，主要分为两类，即传统分割方法以及基于机器学习的分割方法。传统分割方法包括基于阈值、区域、边缘检测等信息完成无序抓取任务图像的分割。Xu 等[18]通过引入各向异性滤波预处理机制，有效增强了原始图像的边缘特征并抑制噪声干扰，进而构建阈值自适应调节模型控制曲线扩散系数，显著提升了图像分割效率。Cigla 等[19]创新性地将超像素过分割结果作为图论节点，通过优化网络拓扑结构实现了算法复杂度的有效降低。Yu 等学者[20]基于数学形态学理论框架，开发了多尺度结构元素融合的边缘检测算法，在分割精度与计算速度间取得良好平衡。值得注意的是，尽管这些传统方法在特定场景中取得进展，其仍普遍面临模型泛化能力不足、特征表达能力有限等技术瓶颈，尤其在复杂背景下的边缘信息保留与细微特征分割精度方面存在显著局限。而卷积神经网络凭借其强大的特征

抽象能力和感受野扩展特性，在全局特征表征方面展现出显著优势，逐渐成为语义分割任务的主流技术路线。

Ciresan 团队[21]于 2012 年首次将卷积神经网络应用于像素级语义分割任务，通过滑动窗口采样策略将局部图像块输入模型进行逐像素分类预测，为深度学习方法在该领域的应用奠定基础。随着技术迭代，全卷积网络(Fully Convolutional Networks, FCN)框架在 2015 年取得突破性进展，其核心创新在于采用 VGG16 等经典架构作为骨干网络，通过全卷积化改造实现端到端训练，并利用反卷积操作实现特征图分辨率恢复，最终通过 Softmax 层输出空间维度保持的分割结果[22]。这种架构设计不仅突破了传统方法对输入尺寸的限制，更在特征复用机制上取得重要进展。尽管 FCN 框架具有里程碑意义，其仍存在参数量大、计算成本高等局限性，特别是在特征图上采样过程中易产生细节信息丢失问题[23]。为此，Ronneberger 团队[24] 同年提出具有编码器-解码器对称拓扑的 U-Net 架构，通过跨层特征拼接机制实现多尺度特征融合，显著提升了边缘细节的保留能力。该模型的创新设计启发后续研究相继开发出 Attention U-Net[25] 与 U-Net++[26] 等改进架构，前者通过引入通道注意力机制优化特征选择过程，后者采用密集跳跃连接增强特征传递效率，这些演进方向共同推动了分割精度与模型效率的协同提升。

1.2.3 目标检测研究现状

目标识别与定位技术作为机器视觉的核心挑战之一，其核心任务在于精确识别色彩图像中的特定目标并完成空间标注。该技术已深度渗透至智能驾驶[27]、安防监控[28]、生物特征识别[29]等关键应用场景，持续推动着产业智能化进程。

传统方法主要依赖滑动窗口遍历机制结合人工特征工程，典型代表包括 Viola-Jones 检测框架[30] 与 HOG 特征检测系统[31]。Viola-Jones 框架通过积分图加速计算与级联分类器设计，在人脸检测领域取得早期突破；HOG 检测器则基于梯度方向直方图构建特征描述子，配合支持向量机实现物体识别。然而，这类方法普遍面临多尺度适应性问题，HOG 检测器需通过图像金字塔重构应对目标尺寸变化，导致计算复杂度显著上升[32]。

随着深度学习技术的突破性进展，目标检测领域逐步形成以卷积神经网络为主导的技术体系[33]，其核心优势在于自动特征学习能力与端到端优化特性[34]。当前主流方法可分为两阶段检测与单阶段检测两大技术路线：

(1)两阶段检测

在两阶段检测架构演进方面，Girshick 团队[35] 提出的 R-CNN 框架开创性地将区域建议与特征提取相结合，通过选择性搜索生成候选区域并利用卷积网络提取特征。后续改进方案中，SPP-Net[36] 引入空间金字塔池化实现多尺度特征融合，Fast R-CNN[37] 通过感兴趣区域池化层统一特征维度，而 Faster R-CNN 则创新性地集成区域建议网络(RPN)，实现了建议生成与目标检测的联合优化。这类方法虽具有较高检测精度，但存在计算冗余与实时性不足等固有缺陷[38] 。

(2)单阶段检测

为提升检测效率，单阶段检测架构应运而生。Redmon 团队[39] 开发的 YOLO 系列模型将检测任务转化为回归问题，通过全局图像分析实现实时检测，后续迭代版本通过多尺度预测与损失函数优化持续提升性能。SSD 检测框架[40] 结合多层级特征图与预设锚框设计，有效增强了对不同尺度目标的适应能力。Lin 团队[41] 提出的 RetinaNet 创新性地采用特征金字塔网络(FPN)与焦点损失函数(Focal Loss)，显著改善了类别不平衡问题。值得关注的是，Law 团队[42] 提出的 CornerNet 突破传统锚框范式，通过角点关键点检测与特征池化策略实现目标定位，为单阶段检测开辟了新的技术路径。后续的 YOLOv8[43] 、YOLOv10[44] 系列算法都借鉴了无锚框模式，采用了各种优化策略以及训练技巧，进一步提高了目标检测的精度。

1.2.4 姿态识别研究现状

机器人抓取姿态估计的早期研究主要集中于基于分析的方法，通过构建力封闭性、稳定性及任务兼容性等数学模型确保抓取可靠性[45] 。其中，力封闭条件与任务约束被视为实现成功抓取的核心要素，如 Li 等[46] 通过几何推导建立三指抓手的力封闭判定准则，同时给出二维/三维空间平衡抓取的充要条件。随着任务导向需求的增强，Haschke R 等[47] 提出基于抓取质量度量的优化框架，通过约束任务方向的最大操作空间与接触力分布来提升姿态适应性。这类方法虽能通过物理约束获得理论最优解，但其计算复杂度随约束条件呈指数级增长[48] ，且依赖精确的力学模型与环境先验[49] ，难以适应动态非结构化场景的实时需求。

为突破传统方法的局限性，研究者开始探索基于特征学习的路径[50] ，通过传感器数据提取环境特征并建立与抓取位姿的映射关系[51] 。相比分析式方法，这类机器学习策略展现出更强的环境适应能力，特别是在未知物体姿态估计方面表现突出。随着机器学习和深度学习在物体检测等领域的突破性进展，基于神经网络的抓取姿态估计方

法迅速发展[52]，主要形成两种技术范式：

(1)基于抓取鲁棒函数

在基于抓取鲁棒函数的方法中，Dex-Net 系列研究具有里程碑意义。Dex-Net1.0[53]通过对称采样生成候选姿态，结合力封闭概率预测实现最优抓取选择，其概率模型综合考量了机械手参数、物体位姿及摩擦系数等不确定性因素。Dex-Net2.0[54] 创新性地构建多模态数据集，并开发 GQ-CNN 模型实现深度图像到抓取姿态的端到端预测。研究[55] 进一步将该框架拓展至仓储拣选场景，构建时间离散的 POMDP 决策模型。Dex-Net4.0[56] 针对双机械手系统进行优化，通过差异化训练策略实现异构抓手的协同控制。

(2)基于结构化抓取表达

基于结构化抓取表达的方法则聚焦于姿态参数的显式建模。早期，Saxena A 等[57]采用多视点融合策略生成三维抓取点云，而 Le QV 等[58] 通过接触点对分析提升抓取稳定性，但这类方法缺乏角度与开合距离参数。对此，Jiang 团队[59] 提出七自由度旋转矩形框表达法，结合两级搜索框架实现高效姿态优化。Lenz I 团队[60] 改进研究，将其简化为五维参数体系，并设计级联卷积网络进行参数精炼，该范式后被广泛采用。后续，Wang Z 团队[61] 通过分割-检测协同架构实现物体分离与姿态估计的联合优化，Asif U 等[62] 则开发了层次化森林模型分析姿态-类别联合概率分布。尽管两级架构精度优势明显，其计算复杂度制约了实时性[63]，故马倩倩等[64] 转向单阶段解决方案，进行基于 DenseNet 的特征融合模型、并行 RGB-D 特征提取架构以及像素级姿态生成网络，在保持精度的同时显著提升推理速度。

1.3 主要研究内容

本文以机器学习模型为基础实现分割，识别网络。首先基于公共数据集 ScanNet 进行扩充，完善数据库。然后对图像进行了数据预处理。首先通过基于单模态识别算法 AC-YOLOV8 实现对物体的目标检测。另一方面，通过改进的 SAM 网络实现对物体色彩图与深度图的掩码分割，然后分割得到的掩码输入到基于多模态特征融合算法 PRD 中完成对目标的三维分割识别，最后通过 ROS 系统完成模拟机器人智能分拣实验。

章节安排如下：

第 1 章：首先介绍了基于工业机器人无序抓取的可研究性以及实用性，同时介绍了国内外对于无序抓取的研究方法，详细介绍了卷积神经网络在工业无序抓取任务上的

成果，并介绍了国内外对其的研究现状。

第2章：介绍了工业机器人无序抓取所需数据集的构建过程，包括彩色图像和点云数据的采集、预处理及标注工作。数据集的处理包括图像的基础变换、机器学习自适应参数优化以及点云滤波等技术。本研究使用了高斯滤波等方法对点云数据进行质量提升，同时构建了标注精确的彩色图与点云结合的数据集。最终将数据集分为 10 个类别，共 2416 张图像。最终完成了标准化的多模态数据集，为后续的算法开发与验证奠定了基础。

第 3 章：本章研究了单模态目标检测算法的改进，提出了基于 YOLOv8 的 AC-YOLO 网络，显著提升了工业机器人无序抓取任务的目标检测性能。在 Neck 模块中引入 GSConv 和 ESlim-Neck 结构，并结合三维空间注意力机制 TDSA，使模型在复杂背景中表现出更强的鲁棒性与检测精度。此外，通过对模型激活函数与损失函数的优化，实现了检测速度与检测准确度的提升。

第 4 章：本章探讨了基于多模态特征的三维分割与识别算法，构建了 RGB-D 数据对齐的点云生成流程，并设计了基于 PointNet 与 E-ResNet 模块的三维分割网络。通过多模态特征的融合和细化，本研究实现了对目标对象的高精度三维识别。实验验证表明，该方法在分割精度和效率方面均具有显著优势，为后续机器人在三维空间中的识别和操作能力提供了有力支持。

第5章：本章围绕机器人智能分拣实验展开研究，详细阐述了相机内参标定、三维坐标定位以及机械臂运动学分析的相关方法。结合 RGBD Fusion 网络和 Refinement 模块的设计，本研究实现了目标物体的高精度姿态识别，并通过 ROS 平台验证了算法的可行性和稳定性。实验结果表明，所提出的方法在无序抓取环境下的目标分拣任务中具有较高的精确性和一定的抗干扰能力。

第6章：总结本论文所做工作，指出工作中存在的不足，以及展望未来，指出工业机器人无序抓取可以发展的方向。

第2章 工业机器人无序抓取数据集构建

2.1 数据来源及流程

2.1.1 数据集采集

当前使用的大多数相机都是普通的 RGB 彩色相机，属于二维相机类型。这些相机拍摄的图像将物体信息压缩为平面，无法提供物体的三维空间数据，因此只能依赖人眼主观推测物体的尺寸和距离，既缺乏数据支持，又缺乏准确性。相比之下，深度相机能够捕获物体的三维空间信息。随着技术的发展，深度相机已经经历了多次迭代，并且衍生出多种不同的原理，如双目立体、时间飞行法、结构光视觉原理。接下来，将依次介绍这三种深度相机的工作原理。

(1) 双目立体深度相机。这类相机通过视觉差实现三维感知。其基本配置包括两个摄像头，分别位于左右两侧，同时拍摄图像。由于相机的空间位置存在差异，两幅图像之间会产生视角差异，通过计算这种视差，便可以推算出目标物体的三维信息。图 2-1 展示了双目立体深度相机的原理示意图。

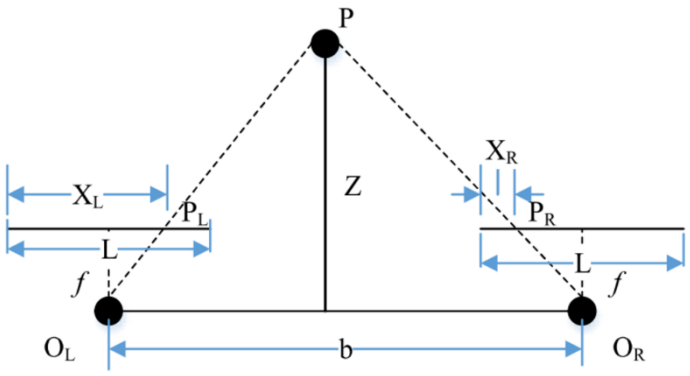


图 2-1 双目立体深度相机的原理示意图

左右两个深度摄像头 O_L 和 O_R 的视觉中心的连线的距离为基线 b，观测点 P 在深度摄像头 O_L 和 O_R 的成像点分别为 P_L 和 P_R。因为光线沿着理想直线进行传播，故观测点点 P 就是两个深度摄像的视觉中心点与成像点连线的交点。而 X_L 和 X_R 分别表示视觉中心 O_L 和 O_R 点到左成像面的距离。由几何关系可知，观测点 P_L 和 P_R 之间的距离为：

$$P_L P_R = b - \left(x_L - \frac{L}{2} \right) - \left(\frac{L}{2} - x_R \right) = b - \left(x_L - x_R \right) \tag{2-1}$$

根据相似三角形理论可以得出：

$$\frac{b-\left(x_L-x_R\right)}{Z-f}=\frac{b}{Z} \tag{2-2}$$

则可以得到点 P 到投影中心平面的距离 Z：

$$\mathrm{Z}=\frac{b\times f}{x_L-x_R} \tag{2-3}$$

(2)结构光深度相机采用主动式编码光投影技术。其工作原理是通过相机上的近红外激光发射器将光束投射到目标物体上，接着使用红外摄像头捕捉物体表面的结构光图像。捕获到的图像信息会传输给芯片进行计算。由于物体是三维的，投射到物体表面的光线会发生形变，利用形变的程度，通过三角测量原理可以计算出物体的深度数据，并据此重建三维结构。因此，结构光技术的核心在于将光进行结构化，通过对比光在目标表面不同区域的形变情况，恢复目标的三维信息。图 2-2 为结构光相机示意图。

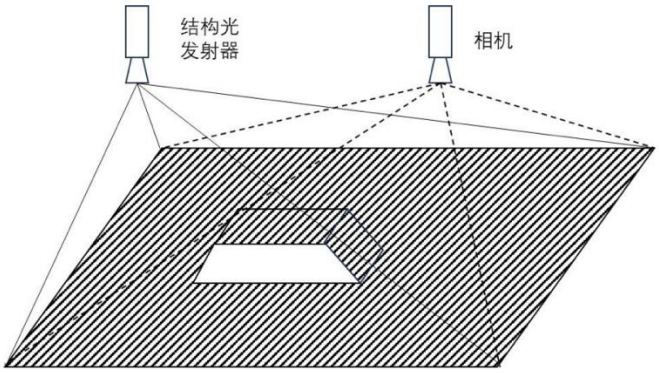


图 2-2 结构光相机示意图

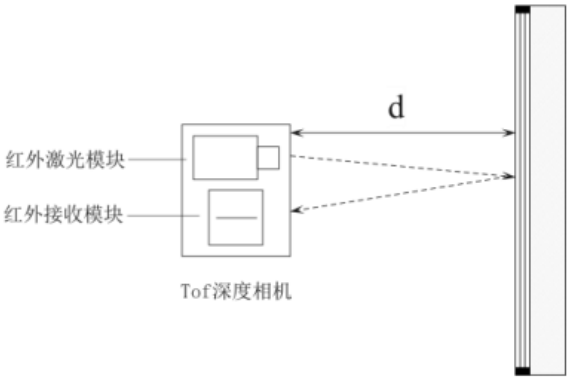


图 2-3 TOF 相机的工作原理图

(3)飞行时间(Time of Flight, TOF)深度传感技术基于光波渡越时间的精确计量实现三维重建。其核心原理通过发射端产生经调制的近红外光波，经目标物体表面反射后，

由接收端的高精度光电传感阵列捕获回波信号。通过解析发射与接收信号间的相位偏移量，结合光速常量建立严格的时空映射关系，可计算获得高精度的深度信息。TOF深度相机能够以非常高的速度获取深度信息，并且在低光环境下也能稳定工作。TOF相机的工作原理如图 2-3 所示。其中的距离值按式(2-4)计算。

$$d = \frac{1}{2} \times c \times \Delta t = \frac{1}{2} \times c \times t_p \times \frac{q_{n+1}}{q_n + q_{n+1}} \tag{2-4}$$

其中，Δt 为飞行间隔；c 为光速；Δt_p 为单次光脉冲的间隔；q_n 是前一次快门获得的电荷量；q_{n+1} 为当前快门收集到的电荷量。

为系统评估不同三维成像原理的深度传感设备性能特征，本文对其进行综述并将其总结为了表格的形式，如表 2-1 所示：

表 2-1 不同三维成像原理的特征

技术维度	基于双目原理	基于结构光原理	基于 TOF 原理
测距方式	被动式立体视觉	主动式编码光投影	主动式光子计时
工作原理	基于视差原理来计算目标点的位置偏差	通过投射与接收到编码图案的偏差	根据发送与接收光信号的的时间差
测量精度	毫米级(0.1-5m)	亚毫米级(<1m)	厘米级(0.5-5m)
测量范围	0.3-5m	0.2-3m	0.5-10m
影响因素	受光线和目标表征影响	易受反光影响	不受光线和目标表征影响
分辨率	低	较高	高
帧率	中	高	低
技术优势	硬件集成简易，设备成本优势显著，室内外普适性强	弱光环境鲁棒性优异，高精度表面重建能力突出	远距离探测性能卓越，动态场景适应性优良
技术瓶颈	立体匹配计算负载较高，无纹理目标失效概率>60%	近距测量盲区(<0.2m)，镜面材质重建误差>15%	光子散射导致远距精度衰减(>5m 误差率上升 40%)
应用场景	室内 SLAM 定位，农业机械视觉导航	工业机器人视觉引导，精密装配，生物特征识别（3D 人脸支付）	自动驾驶障碍物感知（50m 范围），虚拟现实空间映射

TOF 适用于低光环境，且高精度、高帧率的应用场景，符合本次的要求，故采用。采集主要通过如图 2-4 所示数据采集平台实现这一平台包含了 realsenseL515 相机、计算

平台包含的图像采集模块以及图像处理模块等关键组成部分，用以实现图像的捕获、处理和分析，以便进行自动化的分割和识别。具体而言，镜头的作用是将光聚焦并传输到相机的感光元件上；相机则将光学图像转换为电子信号，随后由图像采集模块转换为数字信号，供计算机处理；图像处理模块最终对这些数字图像进行分析，完成特征提取、模式识别等步骤。

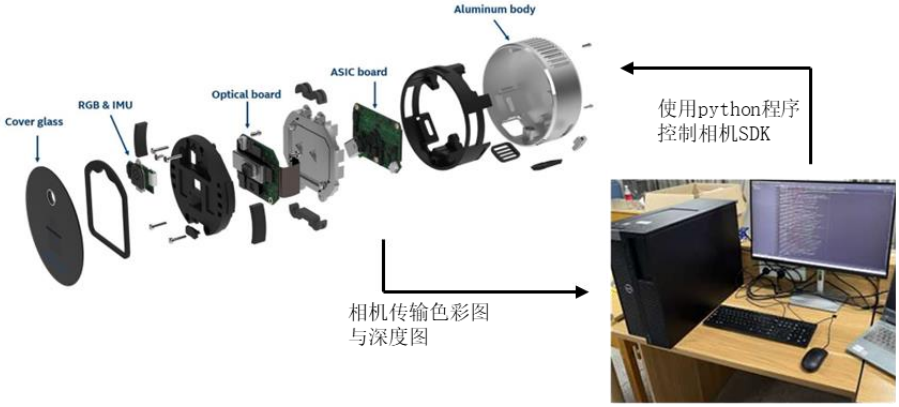


图 2-4 数据采集平台

Intel RealSense L515 是英特尔公司推出的一款采用固态激光雷达扫描技术，在维持紧凑型工业设计与低功耗特性的同时，实现亚毫米级测量精度与扩展量程的深度相机。它结合了深度感知技术与激光扫描原理，旨在为各种应用场景提供精准的三维深度数据，尤其适用于机器人视觉、增强现实、3D 建模以及物体识别等领域。与传统的 RGB-D 相机不同，RealSense L515 采用了固态激光扫描技术，能够实现更高的精度和更远的测量距离，同时保持较小的体积和低功耗。RealSense L515 的硬件架构主要由激光发射器、接收传感器以及 RGB 摄像头等组件组成。激光发射器发射激光脉冲，经过物体反射后，接收传感器测量光的飞行时间，利用光的传播时间差来计算距离，从而生成精准的深度图。。此外，RGB 摄像头提供与深度数据同步的高质量彩色图像，用于生成精确的彩色三维图像。RealSense L515 在室内应用中表现出色，尤其适合对精度要求较高的 3D 重建、室内导航和环境扫描等任务。它可以在短至 0.25 米、长至 9 米的范围内获取高精度的深度数据，适用于多种行业，如医疗健康、工业自动化和虚拟现实等。RealSense L515 的主要性能参数如表 2-2 所示：

表 2-2 L515 主要性能参数

参数	数值	单位
外观尺寸	102 × 25 × 25	mm
工作范围	0.25 - 9	m

RGB 摄像头分辨率	1280 × 720/640×640	像素
RGB 摄像头帧率	30	帧/s
激光扫描分辨率	1024 × 768/640×640	像素
激光扫描帧率	30	帧/s
主板接口	USB 3.1 Gen 1	-
工作环境光强度	0 -100,000	lux
操作温度范围	0 - 40	℃

2.1.2 数据集介绍

在 RGB-D 数据集的研究中，数据集的质量和数量直接影响模型的训练效果，尤其是在多模态学习和深度学习模型的训练中，数据的多样性和覆盖度对模型的泛化能力起着至关重要的作用。RGB-D 数据集包含了同时获得色彩图像和深度图数据的信息，这使得其在计算机视觉、机器人感知、物体识别等领域具有广泛应用。以下是一些当前流行的 RGB-D 数据集，它们广泛用于推动深度学习和机器视觉模型的研究与发展。

NYU Depth V2 数据集[65] 该数据集由纽约大学发布，包含 1449 张室内 RGB 图像和深度图。每张图像都配有详细的像素级语义标签，涵盖了多种室内场景，包括厨房、客厅、办公室等。NYU Depth V2 数据集是一个广泛使用的 RGB-D 数据集，尤其在语义分割、物体检测、3D 重建等任务中具有重要应用。它为深度学习模型提供了丰富的训练数据，尤其是对模型在复杂环境中的鲁棒性具有重要意义。ScanNet 数据集[66] 是一个种类丰富的 RGB-D 数据集，包含超过 1600 个不同场景的扫描图像，覆盖了各种室内环境。数据集提供了 RGB 图像、深度图、3D 点云以及与之相关的语义标签信息。ScanNet 数据集适用于多种任务，如 3D 场景理解、物体检测、姿态估计等。它特别适用于机器人与自动驾驶技术中的环境感知与决策支持。SUN RGB-D 数据集[67] 是由斯坦福大学发布的一个大规模 RGB-D 数据集，包含超过 9000 张室内图像，图像来自 77 个不同场景类别。数据集中的每张图像都有 RGB 图像、深度图以及 3D 标签。该数据集在物体检测、场景解析、深度图生成等任务中被广泛应用。SUN RGB-D 数据集特别适用于研究多模态数据的融合与分析，支持多种基于深度学习的智能视觉应用。Microsoft Kinect for Windows 数据集[68] 作为 Kinect 设备的官方数据集，Microsoft Kinect for Windows 数据集提供了大量包含 RGB 和深度数据的图像，数据集包括不同的动作捕捉和人机交互场景。该数据集的一个关键特点是具有丰富的动态数据，尤其适用于人

体姿态估计、手势识别等任务。由于 Kinect 设备的普及，这一数据集被广泛应用于机器人学、虚拟现实和增强现实领域的研究。

这些RGB-D数据集的建立为基于深度学习的机器人感知、3D重建、环境理解等研究提供了丰富的训练资源。基于本课题目标应用对象，为了提供数据集的泛化性，拟结合 ScanNets 数据集和自采样图片，构建数据集。通过多源数据融合方法，可以提升数据样本的多样性与丰富度，而且有效降低了模型过拟合风险，从而增强了训练模型的泛化性能与适用性。数据集的种类数量分布如图 2-5 所示。

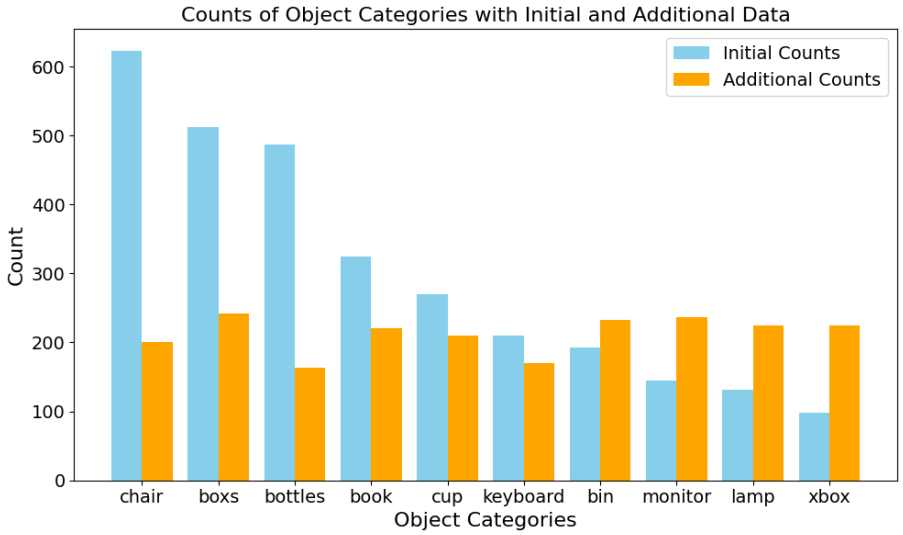


图 2-5 数据集种类数量分布

其数据集是一项目标检测任务数据，数据集收集了常见的物件，椅子、桌子。数据集共计 2416 张图像。其中，具体类别数量见表 2-3：

表 2-3 数据集种类数量

种类	椅子	箱体	瓶子	书本	杯子	键盘	垃圾桶	屏幕	灯	xbox
数量	824	754	651	545	480	380	425	381	356	322

ScanNet 数据集中的图像分辨率为 640×480，即标准的 VGA 分辨率。经 realsense 相机采集的每张图像的分辨率也为 640×480 像素。数据集中的部分图像如图 2-6 所示。

图 2-6 数据集部分影像图片，包含彩色图与其对应的深度图

综上所述，本研究构建的组合数据集为机器人无序抓取任务提供了一个全面且可靠的训练数据库。该数据库不仅涵盖了丰富的物体图像，而且在种类和数量分布上均经过精心设计，完全满足本课题的研究需求。鉴于本课题采用基于机器学习的方法进行预测，数据集的预处理和标签系统的重构成为确保研究质量的关键环节。具体而言，本研究将对数据集中的图像数据进行标准化预处理，并对标签系统进行精细化重构，以确保数据的一致性和可解释性。数据处理流程将按照图 2-7 所示的实验流程执行，以保障数据处理的准确性和实验结果的可靠性。

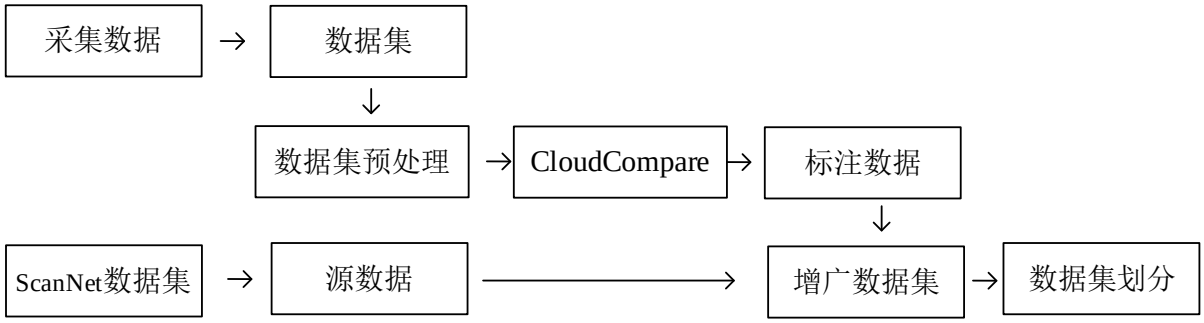


图 2-7 标签制定流程

2.2 色彩图数据集预处理

为了提高无序抓取任务的识别精度，图像预处理是必不可少的步骤。图像预处理主要是对采集到的图像进行一系列处理，以去除无关信息、恢复重要特征、提升目标信息的可检测性，从而为后续的图像分析和抓取任务提供更有价值的数据。其主要作用包括：

1. 修正图像误差：通过校正采集过程中的系统误差与随机误差，处理图像的几何变换问题，如平移、旋转、缩放和翻转等。

2. 图像尺寸调整：将图像调整为适合后续 YOLO 网络模型输入的尺寸要求，例如进行灰度化、尺度变换等操作，以便与模型输入的格式兼容。

3. 图像质量增强：增强图像的对比度和细节，使物体特征更加突出，从而有助于后续抓取模型的识别。

4. 噪声抑制：使用如平均滤波、中值滤波等技术减少图像中的噪声和其他小波动，降低不必要的干扰信息。

5. 边缘与特征提取：对图像进行边缘检测和特征点提取，为无序抓取任务中的物体定位和抓取策略提供数据支持。这些预处理操作有助于提高无序抓取过程中物体识

别和定位的准确性，从而优化整个抓取流程。

2.2.1 基本变换

线性变换是一种简单的图像增强方法，通过对图像的像素值进行线性映射，能够改变图像的亮度和对比度。通过这种映射，较高的像素值会被压缩，较低的像素值则会得到增强，从而可以得到反转的图像效果。该变换在一些特定的图像处理任务中，如图像反色或图像补偿，具有重要应用。在该变换中，原始图像的像素值 通过以下公式：

$$s = 255 - \gamma \tag{2-5}$$

对数变换是一种常用的图像增强技术，其核心原理是通过非线性映射调整图像的灰度分布。具体而言，该变换将图像中低灰度值区域的像素值进行扩展，同时压缩高灰度值区域的像素值，从而有效增强图像低灰度部分的细节信息。此外，对数变换能够显著压缩具有较大像素值动态范围的图像，使其灰度分布更加集中，从而突出图像中的关键细节特征。与之相反，反对数变换则侧重于增强图像高灰度部分的细节信息。对数变换的数学表达式如下：

$$s(x,y) = c\log(1+\gamma(x,y)) \tag{2-6}$$

式中：$s(x,y)$ 表示变换后的像素值；$\gamma(x,y)$ 表示输入的像素值；c 是调节对比度的权重。

伽玛变换是一种非线性变换方法，通常用于图像对比度增强，特别是在图像的低亮度区域。它通过将图像的像素值应用指数函数来放大差异，从而增强图像的细节和对比度中的较暗区域变得更加明亮，从而揭示出图像的细节。由于函数的特性，较小的像素值会经过显著的放大，因此特别适合增强低亮度部分的细节。伽玛变换如下式所示：

$$s = c\gamma^{\varphi} \tag{2-7}$$

根据伽玛系数φ的不同取值，伽玛变换主要可分为以下两种情况：当φ>1 时，伽玛变换适用于处理漂白（过亮）的图像。此时，变换会对图像的灰度级进行压缩，降低高灰度区域的亮度，从而恢复图像的细节信息，并改善整体对比度当φ<1 时，伽玛变换适用于处理过暗的图像。此时，变换会增强低灰度区域的亮度，扩展图像的有效像素区间，从而显著提升图像的对比度，使得原本难以分辨的细节更加清晰可见。线

性变换、对数变换、伽马变换的曲线如图 2-8 所示。

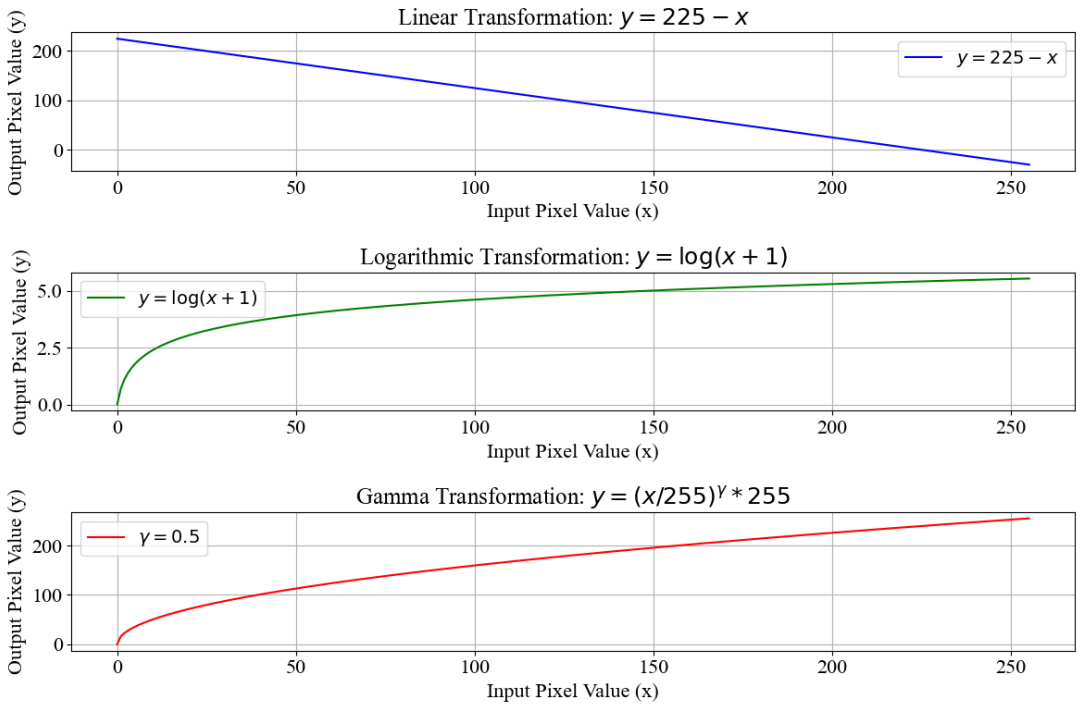


图 2-8 基本变换函数图

2.2.2 基于机器学习的自适应参数变换

三种方式各有优缺，为了更好的处理效果，考虑以下组合方式，结合亮度反转、对数变换和伽玛变换，三者进行加权平均也就是进行：

$$y = \frac{(\omega_1(225-x) + \omega_2 \log(1+x) + \omega_3 x^{\varphi})}{\omega_1 + \omega_2 + \omega_3} \tag{2-8}$$

在机器学习领域，函数的目标是寻找最优的 $\boldsymbol{\omega}$ 值，通过损失函数不断迭代更新，在模型训练过程中，本文采用图像质量和处理时间作为联合目标函数，同时也为了少过拟合的发生采用拉普拉斯平滑，如式(2-9)所示。

$$\begin{aligned} \mathcal{L} &= \alpha \cdot \mathcal{L}_{\text{quality}} + \beta \cdot \mathcal{L}_{\text{efficiency}} + \lambda \cdot \mathcal{L}_{\text{laplacian}} \\ &= \alpha \cdot (1 - \text{SSIM}) + \beta \cdot \text{runtime} + \lambda \cdot \Delta I(i, j) \end{aligned} \tag{2-9}$$

式中：α, β, λ 决定各指标在优化过程中的权重，指标分别是图像相似度评价，处理运行时间和正则化。本次取值分别为 0.5、0.4、0.1。结构相似性 (SSIM, Structural Similarity Index)如下式。

$$\text{SSIM}(x, y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)} \tag{2-10}$$

式中的符号含义见表 2-4。

表 2-4 结构相似性符号含义

计算公式	含义
$\mu_x=\frac{1}{N}\sum_{i=1}^{N}x_i,\mu_y=\frac{1}{N}\sum_{i=1}^{N}y_i$	图像 x和 y的局部均值
$\sigma_x^2=\frac{1}{N-1}\sum_{i=1}^{N}\left(x_i-\mu_x\right)^2,\sigma_y^2=\frac{1}{N-1}\sum_{i=1}^{N}\left(y_i-\mu_y\right)^2$	图像 x和 y的局部方差
$\sigma_{xy}=\frac{1}{N-1}\sum_{i=1}^{N}\left(x_i-\mu_x\right)\left(y_i-\mu_y\right)$	图像的协方差
C_1,C_2 为较小常数，这里取 0.01 与 0.05	避免分母为零

而 $\Delta I(i,j)$ 是图像像素的拉普拉斯算子，定义为：

$$\Delta I(i,j)=4I(i,j)-I(i-1,j)-I(i+1,j)-I(i,j-1)-I(i,j+1) \tag{2-11}$$

式中：$I(i,j)$ 表示 (i,j) 点的像素值，之后加一减一操作是以像素块为单位。式子鼓励图像在局部区域内保持一致性，减小噪声和突变。

最后，进行迭代，损失率曲线如图 2-9。

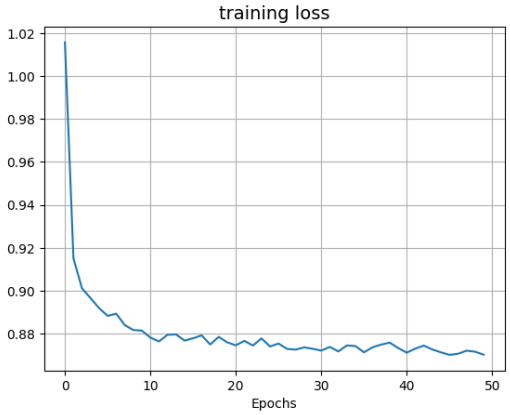


图 2-9 训练损失曲线图

实验进行了 50 次迭代，可以看出基本在 20 个周期后趋于稳定，最后迭代出的参数见表 2-5：

表 2-5 迭代得到的参数

参数	数值
ω_1	0.8693
ω_2	1.1618
ω_3	0.9318

这个组合方式的目标是在不同亮度级别下突出缺陷的特征，使得在图像中更容易

检测和识别。具体操作为：开始时，对图像进行亮度反转，以突出缺陷的轮廓和边缘。接着，应用对数变换来以提高整体对比度，确保缺陷与背景的区分度更大。最后，使用伽玛变换对图像进行微调，增强图像的低亮度区域，使得缺陷的细节在这些区域更加清晰可见。

2.3 点云数据集预处理

2.3.1 点云滤波介绍

点云滤波是点云数据处理流程中的关键预处理步骤，其主要目标是提升点云数据的质量，并为后续的点云分析、建模与理解任务提供更加可靠的数据基础。点云滤波的作用可以概括为以下几个方面：

1. 去除噪声，点云数据通常会受到相机的系统误差（数学原理方面等）、环境干扰（光线明暗交界等）或测量不精确等因素的影响，导致数据中存在噪声。滤波可以有效去除这些噪声点，使得点云数据更加干净和精确，减少噪声对后续实验的进一步处理，如配准、分割等的影响。

2. 降采样与数据压缩，点云数据通常非常庞大，是一般色彩图像的十倍以上。过于密集的点云会增加计算量，影响后续网络的效率。通过滤波可以进行降采样，减少点云中不必要的冗余，达到数据压缩的目的，同时保持重要的几何信息和结构特征。

3. 平滑处理，点云数据由于噪声、传感器精度等因素呈现出不规则的几何形状。滤波可以平滑点云数据，减少突变或尖锐的噪声，保持原始结构的连续性和光滑性，提高点云的可视化效果和后续的分析精度，尤其是在进行后续网络进行特征提取时。

4. 边界点去除，在本次实验采集过程中，可以明显看到很多离群点，边界点或孤立点会明显影响点云的质量，尤其是在面对无序抓取问题时的物体检测、分割等任务中。滤波可以去除这些边界点和孤立点，使得点云数据更加精确和连贯。

5. 保留关键特征，点云数据在滤波时进行降噪和简化数据，但也应该确保保留点云中的关键几何特征，如表面形状、边缘或角点等。

6. 优化点云表示，在如三维重建、模型简化等的场景下，点云的表示非常重要。通过滤波，可以优化点云的数据结构，使得数据更加适应特定的应用场景，如生成低多边形网格或进行快速渲染。

2.3.2 基本滤波方法

体素网格滤波(Voxel Grid Filtering, VGF)：是一种基于空间划分的点云降采样方法。其核心思想是将点云数据划分为多个规则的三维体素网格，并在每个网格中保留一个代表点，通常为网格内所有点的质心或中心点，从而实现点云的降采样。常用于减少点云的密度并加速计算。

统计离群点去除(Statistical Outlier Removal, SOR)：根据每个点的邻域信息，删除远离其他点的离群点，用于去除噪声。其流程如图 2-10 所示。

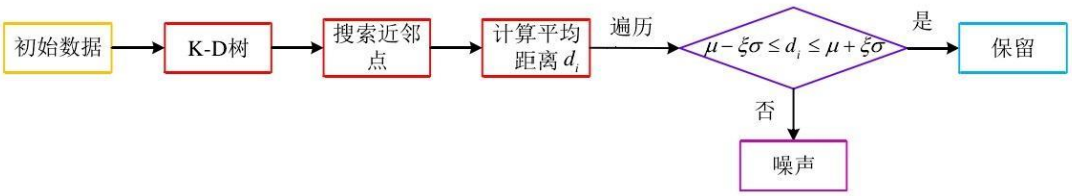


图 2-10 SOR 方法流程图

高斯滤波(Gaussian Filtering, GF)：一种基于高斯核函数的点云平滑方法，主要用于去除点云中的噪声。其基本原理是通过对每个点及其邻域内的点进行加权平均，利用高斯函数赋予不同点以不同的权重，从而实现对点云的平滑处理。其数学表达式如下式：

$$F(x,y,z)=\frac{\sum_{i=1}^{N} p_i G\left(\left\|p_i-p_o\right\|\right)}{\sum_{i=1}^{N} G\left(\left\|p_i-p_o\right\|\right)} \tag{2-12}$$

其中，p_o 为当前点的坐标值； p_i 为邻域内第 i 个点的坐标值；N 为邻域内的点数；$F(x,y,z)$为滤波后点的坐标值；G 为高斯核，如下式：

$$G=e^{-\frac{d^2}{2\sigma^2}} \tag{2-13}$$

移动最小二乘法(Moving Least Squares, MLS)是一种基于局部最小二乘优化的平滑方法，广泛应用于点云数据的平滑和曲面重建。它的核心思想是通过局部最小二乘拟合方法对每个点进行平滑，从而保留数据的整体结构，并消除噪声或不规则性。

半径邻域滤波(Radius Outlier Removal, ROR)是一种用于点云数据处理的去噪方法，主要通过检查点的邻域来识别和移除离群点。它的基本原理是：如果一个点的邻邻域内点数少于某个阈值，则认为这个点是噪声或离群点，将其移除。

2.3.3 点云质量评价

本研究设计了一种基于离散度分析的点云质量评估模型，通过统计点云数据集与最优拟合平面间的空间分布特性，量化评估数据精度与噪声强度。技术实施方案包含以下环节：

1. 选取目标样本。首先，在需要检测的表面采样符合评价要求区域。通过数据裁剪方法，从原始点云数据中提取出用于质量评价的样本数据。样本数据的选取应确保其能够代表被测物体的表面特征，并避免包含明显的异常点或噪声点。

2. 最优平面建模。基于筛选样本采用最小二乘准则构建理想平面模型，其数学表达式可表示为

$$z = ax + by + c \tag{2-14}$$

通过建立正规方程组的矩阵表达式：

$$Av = l \tag{2-15}$$

求解得到平面参数向量的最小二乘解：

$$v = \left(A^T A\right)^{-1} A^T l \tag{2-16}$$

其中，对应值如下：

$$A = \begin{bmatrix} x_1 & y_1 & 1 \\ \cdots & \cdots & \cdots \\ x_n & y_n & 1 \end{bmatrix}; v = \begin{bmatrix} a \\ b \\ c \end{bmatrix}; l = \begin{bmatrix} z_1 \\ \cdots \\ z_n \end{bmatrix} \tag{2-17}$$

其中，$\left(x_1, y_1, z_1\right), \cdots, \left(x_n, y_n, z_n\right)$表征输入样本数据集的空间坐标矩阵。

3. 依据几何学距离公式对每个采样点进行平面偏离量计算。基于正态分布假设，点云与理想平面间的距离分布应满足正态分布，故应有对应的平均值 μ 及方差 σ。

4. 粗差剔除。根据正态分布的定理，一般认为出现在 3σ 之外的为误差点，故删除(μ-3σ，μ+3σ)区间以外的点云数据。

5. 方差表示的是，观察数据偏离中心趋势的离散程度。故把方差 σ 作为点云数据质量评定标准。

对统计离群点移除滤波(SOR)，体素网格滤波(VGF)，移动最小二乘法(MLS)，半径邻域滤波(ROR)和高斯滤波(GF)。上述方法的点云数据质量见表 2-6，可以看出，高斯滤波在均值和方差上表现良好，故后续采用高斯滤波。

表 2-6 各滤波方法的质量评价

	均值 μ/m	方差 σ/m^2
SOR	-0.00217	0.04183
VGF	0.00746	0.05734
MLS	-0.01673	0.00479
ROR	-0.00417	0.00235
GF	0.00074	0.00993

下面对高斯滤波后的效果进行展示，如图 2-11。可以看出，很多在采集过程中的噪声问题，离群点问题都得到了一定程度的解决，完成了下采样，同时也保持了原本的一些几何特征。

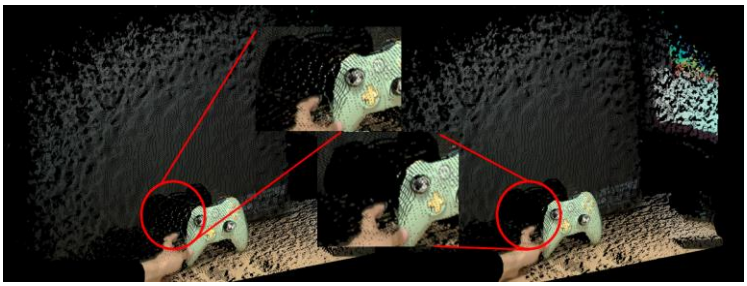


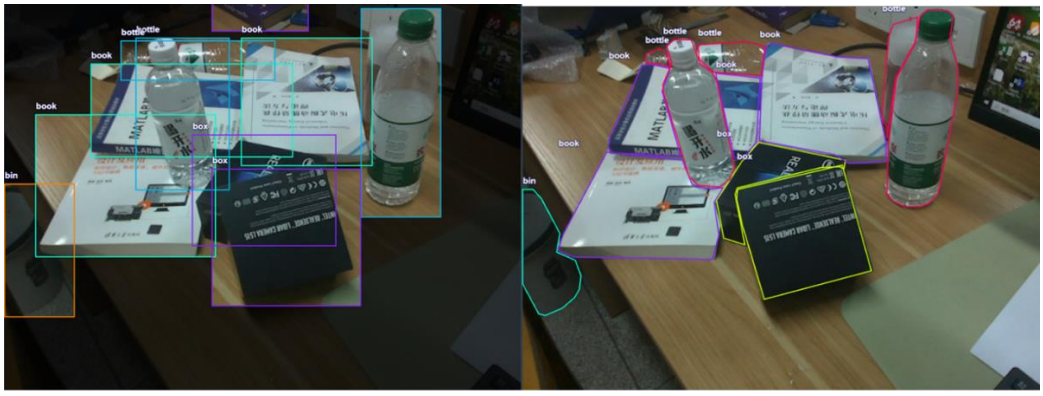
图 2-11 滤波效果展示图

2.4 整体数据集建立

经过数据预处理后共有 2416 张 png 格式的彩色图，2416 张 npy 格式的深度图和两者对齐后的 2416 个 pcd 格式的点云文件。然后利用标注工具进行标注。当前常见的标注工具有：Roboflow、Makesense、Image Labeler 与 CVAT (Computer Vision Annotation Tool)等。本次采用 Roboflow 标注工具。对于点云数据的标注工具也有很多，如 CloudCompare，LabelCloud，Pointly 等。本文选取的是 CloudCompare 标注工具。

2.4.1 彩色图数据集标注

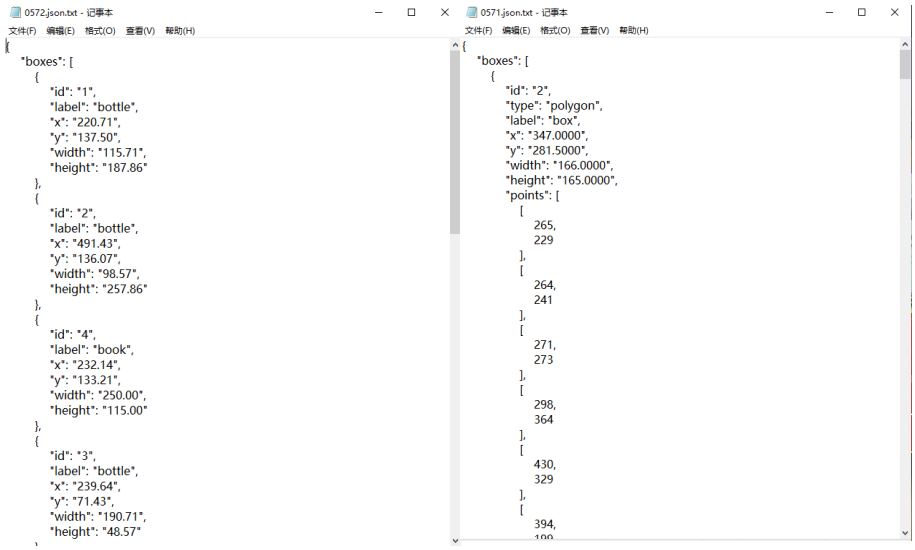
现对数据预处理后的共计 2416 张 png 格式的彩色图片进行数据集标注，分为目标检测和语义分割两部分。标签展示如图 2-12。



（a）分类数据标注　　（b）分割数据标注

图 2-12 标签展示图

每张图片含有有多个物体类别。标注的结果以 json 文件格式保存，分类文件信息中包括目标序号，标签，标注框位置、标注框宽高；分割文件信息中包括，序号，分割框类型，标签名称，标注框中心点位置和宽高和分割转折点的位置。文件内容见图 2-13 所示。



0572.json.txt - 记事本

```
{
  "boxes": [
    {
      "id": "1",
      "label": "bottle",
      "x": "220.71",
      "y": "137.50",
      "width": "115.71",
      "height": "187.86"
    },
    {
      "id": "2",
      "label": "bottle",
      "x": "491.43",
      "y": "136.07",
      "width": "98.57",
      "height": "257.86"
    },
    {
      "id": "4",
      "label": "book",
      "x": "232.14",
      "y": "133.21",
      "width": "250.00",
      "height": "115.00"
    },
    {
      "id": "3",
      "label": "bottle",
      "x": "239.64",
      "y": "71.43",
      "width": "190.71",
      "height": "48.57"
```

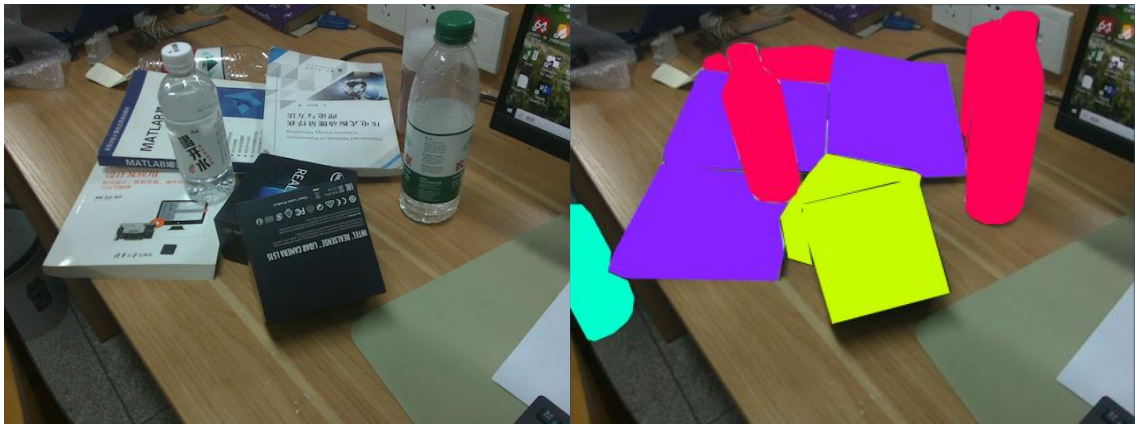
0571.json.txt - 记事本

```
{
  "boxes": [
    {
      "id": "2",
      "type": "polygon",
      "label": "box",
      "x": "347.0000",
      "y": "281.5000",
      "width": "166.0000",
      "height": "165.0000",
      "points": [
        [
          265,
          229
        ],
        [
          264,
          241
        ],
        [
          271,
          273
        ],
        [
          298,
          364
        ],
        [
          430,
          329
        ],
        [
          394,
```

（a）分类标签json文件　　（b）分割标签json文件

图 2-13 标注文件内容图

由于 JSON 文件无法直接用于网络训练，因此需要通过 Python 程序提取其中的文本数据。对于特征分类标签，提取的关键信息包括物体类别、标注框中心点位置以及标注框的宽度和高度，这些信息被保存为 TXT 格式文件，以便于后续模型读取和处理。对于分割标签，提取的关键信息则保存为 PNG 图像格式，其中每个像素值对应物体的类别标签，从而为图像分割任务提供准确的标注数据。这一数据处理如图 2-14 所示：



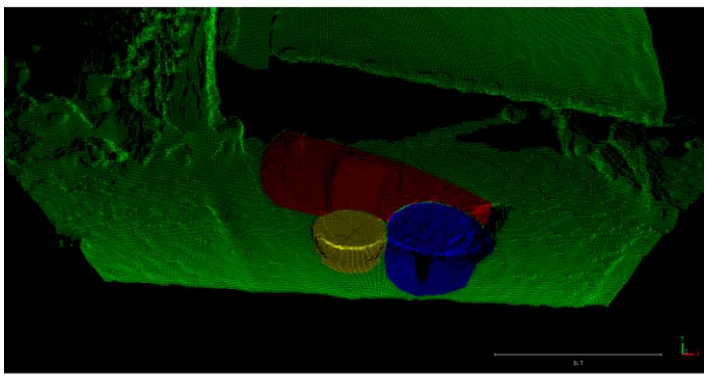
（a）原始图像　　（b）分割标签图像

图 2-14 分割图像展示

2.4.2 点云数据集标注

现对经过深度图与彩色图对齐后 pcd 格式点云，共计 2416 张，在 cloudcompare 中进行数据集标注。标签展示如图 2-15。

(a) 标注单类别展示



(b) 整体标注展示

图 2-15 点云标签可视化展示

每张图片含有有多个物体类别。标注的结果以 txt 文件格式保存。将标注好的标签数据导出为 ASCII Cloud 数据，以文本格式展示，如图 2-16。每行都代表一个点云中的

点，每列都含有对应点的相应信息。从左到右依次为，X 坐标，表示点在三维空间中的 X 轴位置。Y 坐标，表示点在三维空间中的 Y 轴位置。Z 坐标，表示点在三维空间中的 Z 轴位置。红色通道的颜色值，其取值范围为 0-255（RGB 颜色空间中的红色分量）。绿色通道的颜色值。蓝色通道的颜色值。后续表示属于不同类别的值。

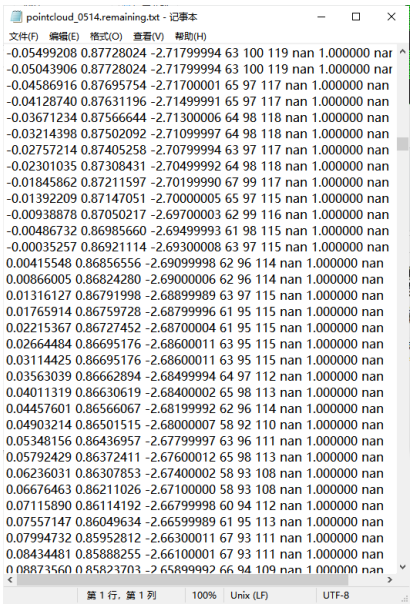


图 2-16 点云 txt 文件展示

2.5 本章小结

本章介绍了工业机器人无序抓取所需数据集的构建过程，包括彩色图像和点云数据的采集、预处理及标注工作。数据集的处理包括图像的基础变换、机器学习自适应参数优化以及点云滤波等技术。本研究使用了高斯滤波等方法对点云数据进行质量提升，同时构建了标注精确的彩色图与点云结合的数据集。在建立数据集时，对 ScanNet 数据集进行扩充，将数据集分为 10 个类别，共 2416 张图片。最终完成了标准化的多模态数据集，为后续的算法开发与验证奠定了基础。

第3章 基于单模态识别算法研究

3.1 引言

在现代机器人技术领域，自动化抓取任务作为一种关键应用，广泛应用于仓储物流、制造业以及智能服务等多个行业。机器人抓取任务的核心挑战之一就是在无序环境中快速、准确地识别和抓取目标物体，这对于提高生产效率和减少人工干预具有重要意义。然而，物体在无序环境中的形态和位置变化多端，给目标识别带来了巨大难度，尤其是当物体表面特征不明显或者背景复杂时，传统的视觉识别方法往往表现不佳。近年来，基于深度学习的目标检测算法在机器视觉领域取得了显著进展，尤其是YOLO系列网络在实时性和精度上表现出色。YOLOv8作为YOLO系列中的较新版本，其在小物体检测、特征提取以及速度优化方面具有显著优势，已经在多个领域获得了应用。对于机器人无序抓取任务，YOLOv8 通过高效的卷积神经网络结构，能够快速准确地识别多种物体，成为该领域的研究热点。

然而，尽管该系列算法在无序抓取任务中表现出色，但仍面临一些挑战。首先，对于小物体或者遮挡严重的物体，YOLOv8 模型会出现漏检和误检，特别是在本次的无序抓取的复杂场景中。其次，现有的单模态图像识别方法对于物体的深度信息和空间位置的感知能力较弱，这在一些高精度要求的任务中仍有局限性。尽管通过引入多尺度特征融合和注意力机制有所改善，但如何进一步提升检测精度和处理复杂场景的能力，依然是当前研究的难点。

因此，本文提出了一种新的网络架构 AC-YOLO(Anarchic Crawling-YOLO)。该架构以经典 YOLOv8 为主框架，在 Neck 模块中引入 GSConv 模块，部分语义信息的丢失，并结合了 FedFusion 模块设计了 ESlim-Neck 结构。该结构具有更好的稳定性和兼顾更多的数据信息。针对无序抓取的场景，在检测头部分设计了三维空间注意机制(Three-Dimensional Space Attention，TDSA)，来克服 SE 注意机制没有对通道和空间位置进行区分的问题。最后，本文还给出了实验结果的可视化和量化评价方案，并和多个网络进行对比，以体现网络应用场景的适用性。整体网络流程图如图 3-1。

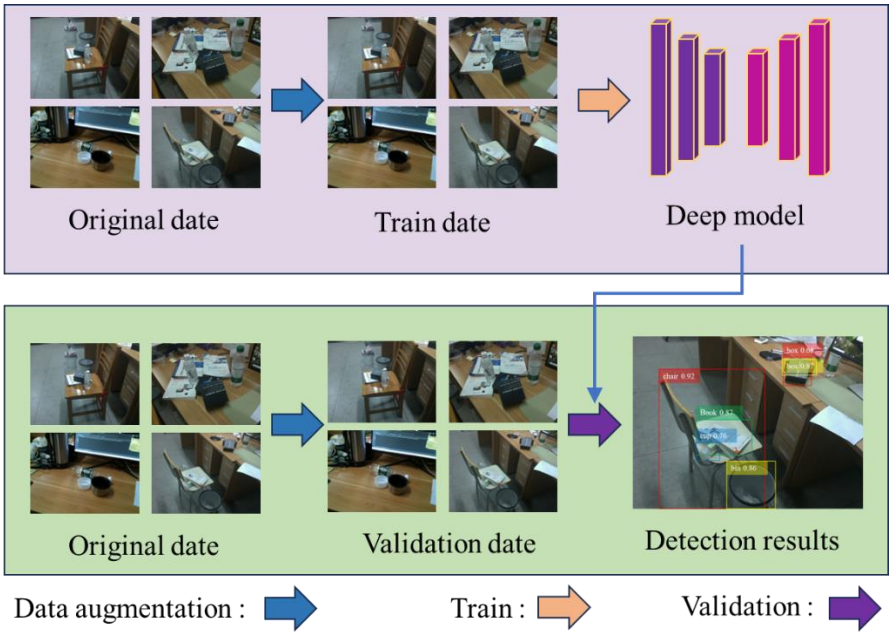


图 3-1 AC-YOLOv8 网络整体流程图

3.2 基于 YOLOv8 的识别算法

经典 YOLOv8 网络架构分析。YOLOv8 是一种高效的单阶段目标检测模型，其架构主要由 Backbone、Neck 和 Head 三部分组成。以下对各模块的功能与设计进行详细分析：

1. 输入预处理。输入图像首先被调整为固定尺寸(640×640×3)，以满足网络的输入要求。这种标准化操作确保了模型对不同分辨率图像的处理一致性。

2. 主干网络(Backbone)。主干网络采用 CSPDarknet 结构，其核心特点包括：(1)多尺度特征提取：通过不同尺寸的卷积核捕获图像的低级边缘、纹理及高级语义信息。(2) 跳跃连接：利用跨层连接增强特征传递效率，避免深层网络中的梯度消失问题。

3. 颈部网络(Neck)。颈部网络由特征金字塔网络(FPN)与路径聚合网络(PAN)构成，用于多尺度特征融合：FPN：通过自上而下的路径传递高层语义信息，增强对小目标的检测能力。PAN：通过自下而上的路径传递低层细节信息，提升对大目标的定位精度。两者的结合使模型能够同时捕获不同尺度的目标特征，显著提升检测性能。

4. 检测头(Head)。检测头在三个不同分辨率的特征图上工作，分别负责大、中、小尺寸目标的检测。其输出包括：边界框参数：中心点坐标(x,y)、宽度与高度(w,h)及置信度评分。类别概率：每个边界框属于不同类别的概率分布。为优化检测结果，模型采用非极大值抑制(NMS)算法去除冗余框：根据置信度评分筛选候选框。

计算框之间的交并比(IoU)，去除重叠率高且置信度低的框。保留最优检测结果，输出目标边界框、类别标签及置信度评分。网络结构图如图 3-2：

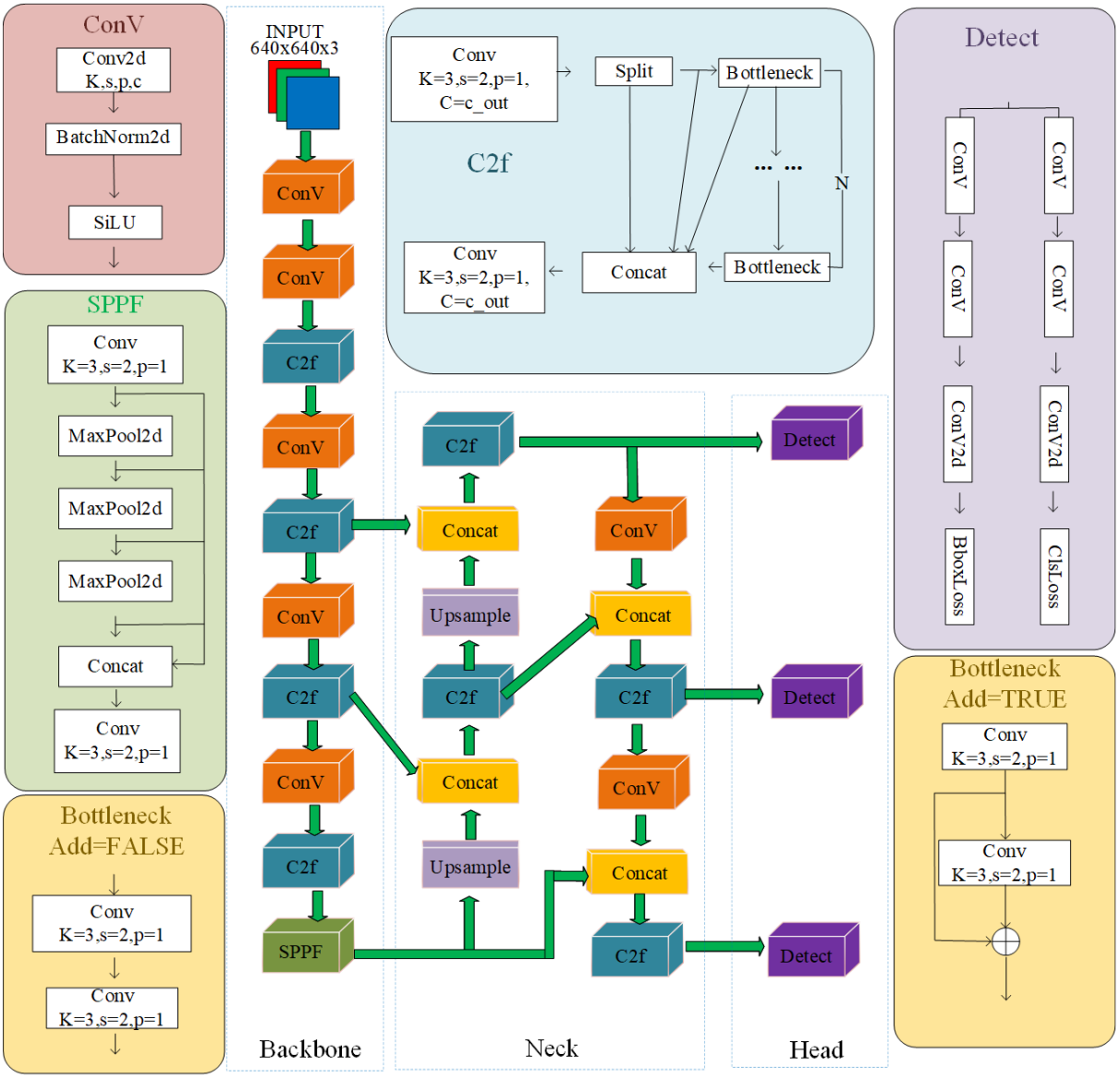


图 3-2 YOLOv8 网络结构图

在经典的 YOLOv8 网络中，输入图像在 Neck 模块中经过卷积和上采样处理后，空间信息逐渐传递到通道中，这一过程在编码特征图维度和通道时会导致少量语义信息的丢失。为了解决这一问题，同时使得 YOLO 网络更适用于机器人无序抓取目标检测算法的场景，AC-YOLO 网络在特征提取、特征融合以及网络轻量化设计等方面进行了多项改进。具体而言，经典 YOLO 网络中 Neck 层中的 FPN+PAN 结构改进为 GSConv 与 FF-Slim-Neck 组合的特征融合网络。这种改进增强了图像关键信息的学习，也提高了特征融合的效率，使得网络能够更好地处理多尺度目标检测任务。

进一步而言，本文在 Neck 模块中引入了 GSConv(Group Shuffle Convolution)结构。模块旨在增强网络对全局信息的捕捉能力，能够在减少计算量的同时保留更多的空间信息，从而有效缓解语义信息丢失的问题。

然后为了进一步平衡检测精度、检测速度和计算成本之间的关系，本文在 GSConv 中引入了 FF-Slim-Neck 结构。FF-Slim-Neck 通过优化特征融合路径和减少冗余计算，能够在保持较高检测精度的同时显著提升检测速度，尤其在小物体检测任务中表现出色。

最后，在 Head 层中引入了 TDSA(Three-dimensional Space Attention)注意力机制。TDSA 机制利用三维权重在空间和时间维度上同时操作，以捕捉三维数据中的局部特征，进而生成注意力权重图。该权重图指示了输入数据中每个位置的重要性，帮助网络来增强特征。

基于上述改进，本文提出了一种基于 YOLOv8 的算法框架，称为 AC-YOLO，其整体结构如图 3-3 所示。AC-YOLO 网络通过结合 GSConv、FF-Slim-Neck 和三维空间注意机制 TDSA，在保持高效计算的同时显著提升了检测精度和速度，尤其适用于无序场景下的工业机器人抓取检测任务。

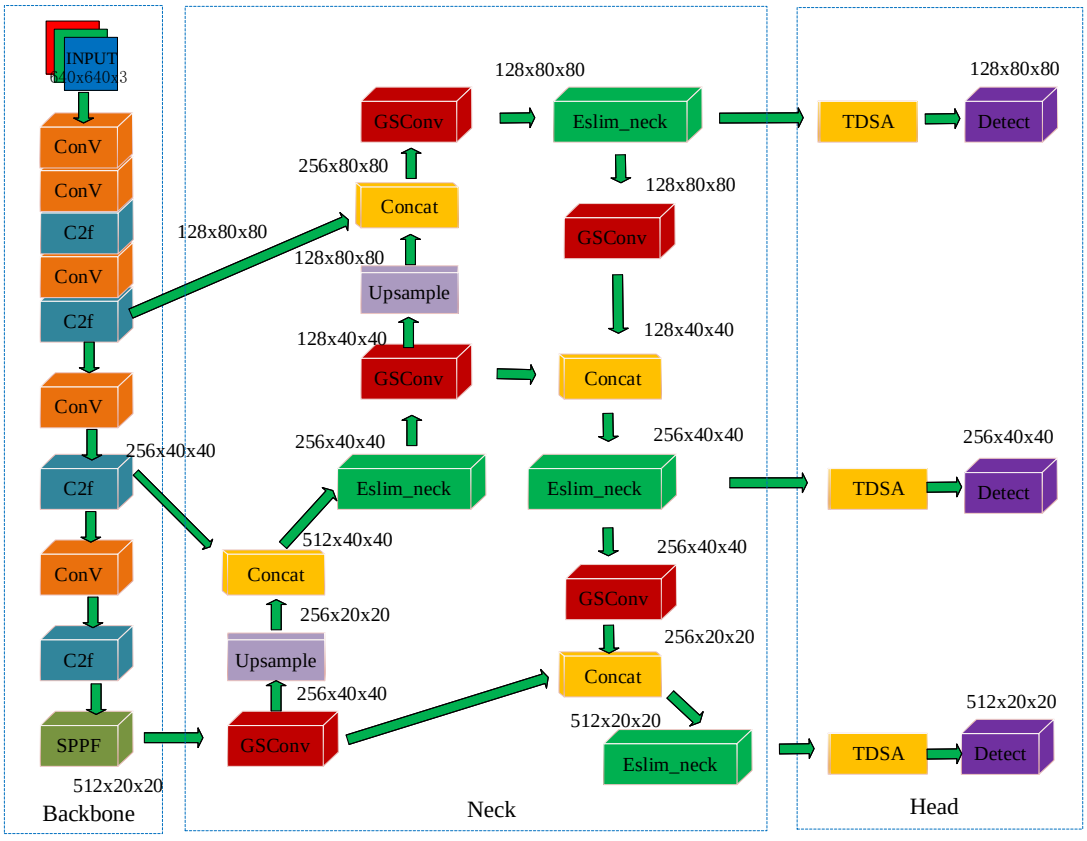


图 3-3 AC-YOLOv8 网络结构图

3.2.1 特征融合模块

本文在 Neck 模块中引入了 GSConv 结构，其结构旨在增强网络对全局信息的捕捉能力，同时也通过分组卷积和通道混洗操作，来在模型复杂度不增加的情况下保留更多的空间信息，从而有效缓解语义信息丢失的问题。进一步而言，该结构有以下特点：

1.传统卷积核的感受野有限，难以捕捉远距离特征，GSConv 通过扩大感受野，使每个输出位置都能获取输入图像的全局信息。

2.在空间维度上执行卷积，利用大尺寸卷积核和空洞卷积来覆盖更大区域，从而捕捉全局特征。

3.GSConv 的卷积核参数在空间上共享，减少了参数量，提升了计算效率。其结构如图 3-4 所示

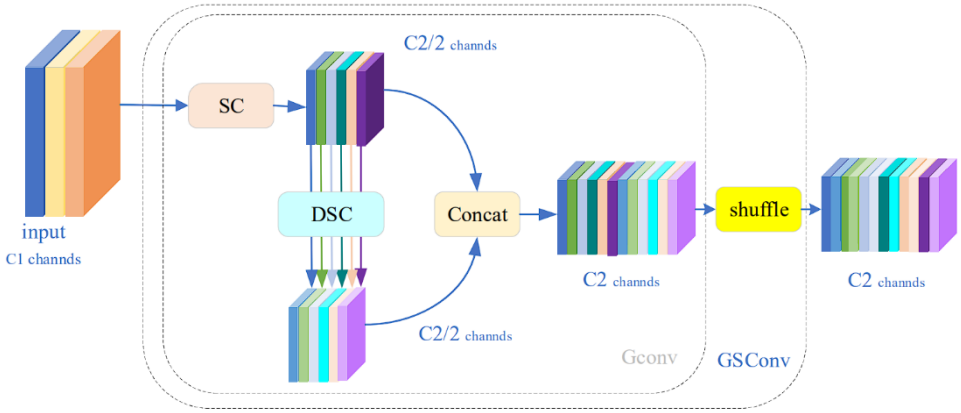


图 3-4 GSConv 模块结构图

GSConv 能够有效保留每个通道的关联性，从而特征的信息量。具体而言，对于给定的输入特征图，大小为$C_1 \times H \times W$，首先进行卷积操作。这一模块主要包含两种卷积操作：标准卷积(Standard Convolution, SC)和深度可分离卷积(Depthwise Separable Convolution, DSC)。

1. 使用 SC 对特征图进行分组卷积操作，此时输出 $C_2/2 \times H \times W$ 的特征图，完成对局部特征的提取。

2. 借鉴残差连接的思想，对输入特征进行 DSC 操作，使得输出特征更加注重空间上的联系，也提高了运算速度。另一方面，并将输入与输出进行 concat 操作，在通道维度上进行拼接，得到大小为 $C_2 \times H \times W$ 的融合特征图。使得输出特征有了 SC 的局部特征提取能力和 DSC 的高效特征表示能力。

3. 最后，将前一步的输出放入 Shuffle 模块，也就是对融合后的特征图进行通道重

排列操作。Shuffle 操作通过打乱通道顺序，来让不同通道之间可以相互学习，此时输出大小为 $C_2\times H\times W$ 的特征图。

颈部模块采用了 ESlim-Neck 结构，该结构的设计学习了 DenseNet、VoVNet 和 CSPNet 等网络的优点，结合了广义学习方法，旨在提升特征融合的效率并降低计算复杂度。具体而言，首先对输入的 $C_1\times H\times W$ 特征进行卷积操作，需要注意的是 ESlim-Neck 只对通道数进行改变，并不会改变特征的高和宽，经卷积操作后通道数减半，此时输出为 $C_1/2$。随后，进入 GSConv 模块输出通道数为 $C_2/4$ 的特征图。再将其输入到 FCConv 模块进行特征融合，此时特征图的通道数依然为 $C_2/2$。FCConv 模块如图 3-5 所示，先对输入的特征 X 提取局部特征 $E_l(x)$和全局特征 $E_g(x)$，将其放到逐通道卷积中得到融合特征。

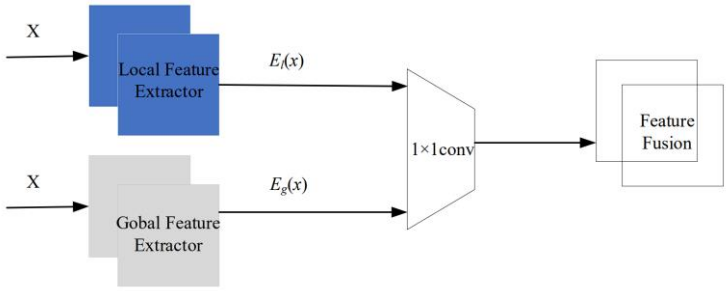


图 3-5 FCConv 模块结构图

整个过程 FCConv 模块数学表达式如式子：

$$F_{\text{conv}}\left(E_l(x),E_g(x)\right)=W_{\text{conv}}\left(E_l(x)\,\|\,E_g(x)\right) \tag{3-1}$$

最后，通过残差连接，将 FCConv 模块和 Conv 模块的输出进行连接，完成这一模块的输出，此时的输出特征图为 $C_2\times H\times W$。ESlim Neck 整体结构如图 3-6：

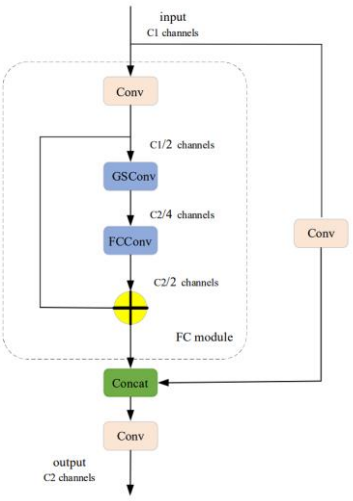


图 3-6 E-Slim-Neck 结构图

3.2.2 注意力学习模块

卷积神经网络核心在于卷积层的设计，该层通过局部感受野机制提取输入数据的空间特征。具体而言，卷积层由多个可学习的滤波器构成，每个滤波器在输入图像上滑动并进行逐点乘加运算，生成对应的特征映射。这些映射能够逐层捕获从低级边缘、纹理到高级语义结构的层次化特征，从而实现对图像内容的有效表征。通过堆叠多个卷积层，CNN 能够逐步提取从低级到高级的语义特征，从而实现对复杂视觉任务的高效处理。卷积核对应其特征图，而特征图在进一步处理后得到输出结果。这一过程的数学表达如下：令 i 表示对应层的函数，如式所示：

$$Y_i = F_i(X_i) \tag{3-2}$$

式中，Y_i 是输出特征；F_i 是卷积运算；X_i 是输入数据，X_i 是一个形状为 (H_i, W_i, C_i) 的张量，而 H_i 、W_i 和 C_i 表示数据的高、宽和通道数。进一步而言，整体网络可以由式(3-2)组成，如式所示：

$$N = \bigoplus_{i=1..s} F^{L_i}_i(X_{(H_i, W_i, C_i)}) \tag{3-3}$$

最终得到的 N 即为设计的神经网络。

本研究采用的注意力机制基于“挤压-激励模块”(Squeeze-and-Excitation Module, SE)，其核心思想是通过全局最大池化提取全局上下文信息，进而动态调整各通道的权重分布。然而，SE 模块在实际应用中存在以下不足：

1. 感知能力受限：缺乏对通道间空间布局的建模能力，难以适应机器人无序抓取任务中复杂的目标分布。

2. 计算开销较大：权重计算过程涉及大量矩阵运算，导致网络复杂度增加，限制了模型从神经元中提取深层特征的能力。

为了克服这些挑战，本文引入了三维空间注意机制(Three-Dimensional Space Attention, TDSA)。TDSA 机制通过结合空间维度和通道维度的注意力权重，能够更精确地捕捉目标物体在三维空间中的分布和重要性差异，其核心思想是区分特征图上的不同空间位置，并单独处理所有通道，为每个神经元分配唯一的权重。其主要进行以下步骤：

1. TDSA 注意力机制使用三维卷积核在空间和时间维度上同时操作，以捕捉数据中的局部特征。

2. 将捕捉到的特征生成注意力权重图，该权重图指示了输入数据中每个位置的重要性。特别的，权重图通过 sigmiod 函数进行归一化，确保权重总和为 1。

3. 特征加权：将生成的注意力权重与输入特征图进行逐元素相乘，从而增强重要特征并抑制不重要的特征。

4. 残差连接：为了保持原始信息，将加权后的特征与原始特征进行残差连接，形成最终输出。TDSA 结构如图 3-7 所示

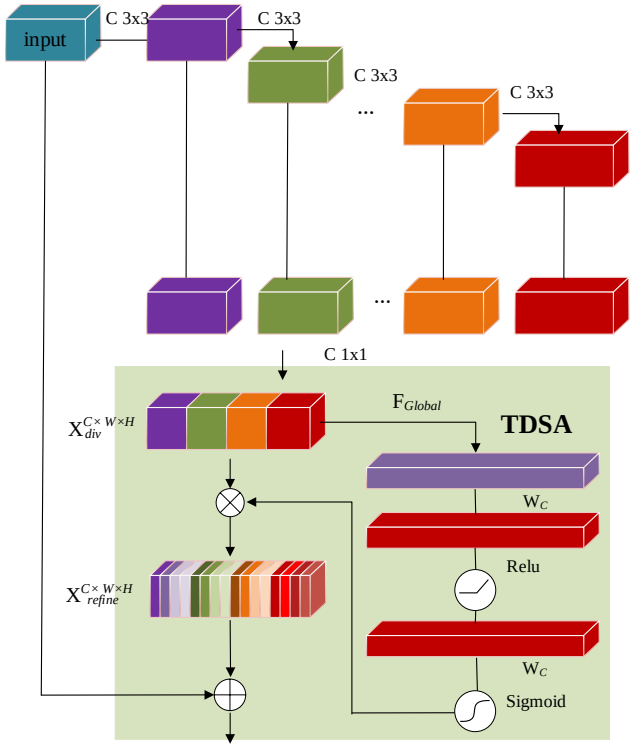


图 3-7 TDSA 结构图。图中 C_{3x3}，C_{1x1} 表示 3×3 和 1×1 的卷积核，F_{Global} 表示全局池化，W_C 是全连接层，TDSA 是三维通道注意力， 表示元素逐个相乘， 表示元素逐个相加

具体而言，给定输入特征图的尺寸为 W×H×C，字符含义同上文。输入进入多个 3×3 的标准卷积，得到多个输出结果，将结果进行在通道维度上的拼接，图中使用不同的颜色标识，拼接后的结果作为 TDSA 模块的输入。TDSA 首先通过全局池化操作捕获特征图的空间上下文信息，然后利用卷积和多层感知机(MLP)计算每个空间位置和通道的注意力权重。这些权重能够动态调整特征图中每个神经元的重要性，从而增强关键特征的表示能力。在这之后分别使用了 Rule 和 Sigmiod 激活函数。此外，TDSA 将现有的注意力模块整合到每个网络块中，进一步提升了检测头层的输出质量。通过这种设计，TDSA 不仅能够显著提高模型对目标物体的定位精度，还能增强模型在复杂场景下的鲁棒性，同时保持较低的计算复杂度。

3.2.3 激活函数

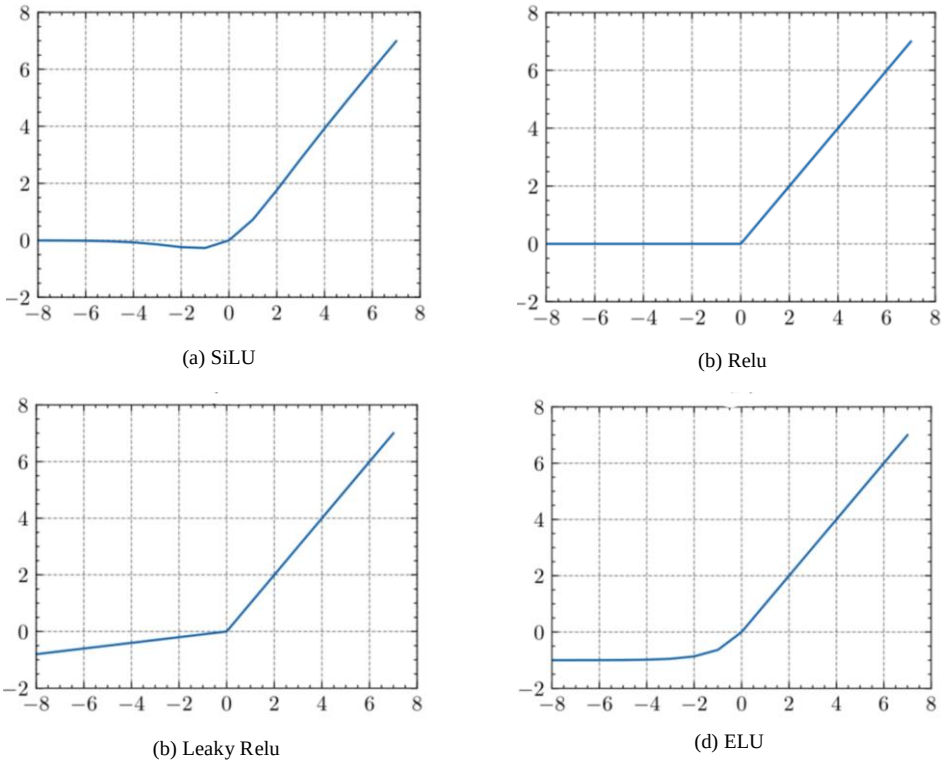
为防止在反向传播中出现梯度消失，本文选用在函数特性上具有连续可微性质的SiLU 激活函数，其函数图像如图 3-8(a)所示。相较于图 3-8(b)的 Leaky ReLU、图 3-8(c)的 ReLU 以及图 3-8(d)的 ELU。与其他激活函数相比，SiLU 具有具备无上界有下界、平滑、非单调的特性，并具备自稳定能力。因其导数在权重上存在零点，可以起到隐式正则化的作用，能够抑制大权重的学习，从而避免模型过拟合。各激活函数表达式如下式所示：

$$f(x)=\frac{x}{1-e^{-x}} \tag{3-4}$$

$$f(x)=\max(0,x) \tag{3-5}$$

$$f(x)=\begin{cases} x & ,x>0 \\ ax & ,x\le 0 \end{cases} \tag{3-6}$$

$$f(x)=\begin{cases} x & ,x>0 \\ \alpha(e^{x}-1) & ,x\le 0 \end{cases} \tag{3-7}$$



(a) SiLU　(b) Relu

(b) Leaky Relu　(d) ELU

图 3-8 激活函数函数图像

3.2.4 损失函数

AC-YOLO 的总损失函数是由多个子损失函数组成的，主要包括定位损失(Localization Loss)、置信度损失(Confidence Loss)、分类损失(Classification Loss)以及对于无序抓取识别任务的改进项损失 Focal Loss。每个损失项的细节如下：

1. 定位损失，用于评价模型预测的边界框与真实边界框之间的差距。AC-YOLO 采用的定位损失结合了 IoU (Intersection over Union)，采用了 GIoU (Generalized Intersection over Union)损失，以及边界框中心的预测误差。其核心目标是确保网络能够准确地定位物体边界。GIoU 定义及其损失如下：

$$\mathrm{GIoU}(A,B)=\frac{|A\cap B|}{|A\cup B|}-\frac{|C-(A\cup B)|}{|C|} \tag{3-8}$$

$$\mathcal{L}_{GIoU}=1-\mathrm{GIoU}(A,B) \tag{3-9}$$

其中，A 和 B 分别是预测框和真实框；$|A\cap B|$和$|A\cup B|$分别是预测框和真实框的交集区域面积和并集区域面积；C 表示包含 A和 B的最小闭合矩形区域面积，也就是最小外接矩形(bounding box)；$|C-(A\cup B)|$是并集外的区域面积，用来衡量框的分离度。

2. 置信度损失，AC-YOLO 网络中置信度损失用于评估模型对目标是否存在的信心。这包括两部分：一是对于有物体的区域（正样本）进行置信度评分的误差，二是对于无物体的区域（负样本）进行的误差。采用做法是采用二进制交叉熵损失 (Binary Cross-Entropy Loss) 来度量每个区域是否包含物体。损失形式如式：

$$\mathcal{L}_{conf}=\sum_{i}\left[1_{obj}\cdot\mathrm{BCE}(S,\hat{S})+1_{noobj}\cdot\mathrm{BCE}(0,\hat{S})\right] \tag{3-10}$$

其中，1_{obj}表示物体存在时的指示函数；1_{noobj}表示非目标存在时的指示函数；S 是预测置信度；$\hat{S}$ 是真实置信度。BCE 的数学公式见下式：

$$\mathrm{BCE}(y,\hat{y})=-\left(y\cdot\log(\hat{y})+(1-y)\cdot\log(1-\hat{y})\right) \tag{3-11}$$

3. 分类损失，用于度量网络对每个目标类别的预测错误。AC-YOLO 网络使用交叉熵损失 (Cross-Entropy Loss) 来计算类别的误差，目的是使得网络能将目标分类为正确的类别。损失形式如式：

$$\mathcal{L}_{cls}=-\sum_{i}y_i\cdot\log(\hat{y}_i) \tag{3-12}$$

4. Focal Loss，来应对类别不平衡等问题，AC-YOLO 网络使用 Focal Loss 来减少简单背景区域对训练的影响，提高难例的关注度。公式如下：

$$\mathcal{L}_{focal} = -\alpha_t(1-p_t)^{\gamma}\log(p_t) \tag{3-13}$$

其中，p_t是模型预测为目标的概率；α_t和γ是调整权重；α_t是用于调整目标与非目标的权重，也就是在目标与非目标两者数量差距较大时使得模型加大较少类别样本的权重；γ用于难分类样本的情况，使模型关注那些被错误划分且难以学习的样本，减少易学习样本的影响。具体而言，当p_t趋近 1，说明该样本是易学习的样本，此时调整权重$(1-p_t)^{\gamma}$是趋向 0，此时损失的权重较小。当p_t趋近于 0，即负样本被分到正样本，且该样本为前景的概率特别小，此时$(1-p_t)^{\gamma}$调制因子趋向于 1，此时损失增大。不同的γ取值所对应的 Focal Loss 如图 3-9。

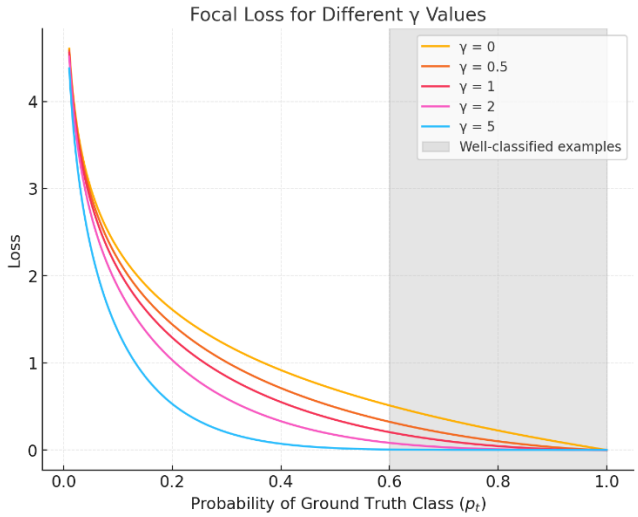


图 3-9 不同的γ取值所对应的 Focal Loss 曲线图

3.2.5 网络训练和参数配置

能否为机器学习和深度学习模型选择正确的超参数，影响着模型中能否提取出特征数据。故在本次 AC-YOLO 网络中，超参数的设计显得尤为重要。例如，学习率过小，会极大降低收敛速度，增加训练时间；学习率过大，可能导致参数在最优解两侧来回振荡。表 3-1 展示了本次超参数的具体配置。

表 3-1 超参数配置

超参数	具体值
预热迭代次数	10
预热动量	0.925
预热偏置学习率	0.02
迭代周期	470

初始学习率	0.125
最终学习率	0.003
动量	0.863
权重衰减	0.00008
批量大小	32
输入图尺寸	640×480

这些超参数通过共同作用，优化 AC-YOLO 网络的训练过程。热身参数，如热身周期、热身动量确保训练初期稳定，学习率参数，如初始学习率和最终学习率控制收敛速度和稳定性，而正则化参数，如权重衰减帮助提升模型的泛化能力。这些设置平衡了 AC-YOLO 的高效性和精度需求。

3.3 实验结果与展示

3.3.1 评价标准及实验平台

在图像预处理部分，采用图像相似度评价和处理运行时间来进行量化评价，具体细节参考第二章相关内容，在此不再复述。

为评估目标检测算法的性能，本研究选取像素精度(PA)、平均像素精度(mPA)、平均交并比(mIoU)及模型参数量作为核心指标，前三个指标计算需基于混淆矩阵，也就是假设总类别数为 k+1，定义 p_{ij} 为实际类别 i 被预测为类别 j 的像素数，其中 p_{ii} 代表真阳性， p_{ij} 为假阳性， p_{ji} 为假阴性。

像素精度表征正确分类像素占比，其计算式为：

$$PA = \frac{\sum_{i=0}^{k} p_{ii}}{\sum_{i=0}^{k}\sum_{j=0}^{k} p_{ij}} \tag{3-14}$$

平均像素精度通过逐类计算正确分类像素比例后取均值获得：

$$mPA = \frac{1}{(k+1)}\sum_{(i=0)}^{k}\frac{p_{ii}}{\sum_{j=0}^{k} p_{ij}} \tag{3-15}$$

mAP50 指标衡量是当 IoU 阈值为 0.5 时，模型的平均精度。具体来说，mAP50 计算的是所有类别的 AP 的平均值，其中 AP 是在 IoU 阈值为 0.5 时计算的。mAP50-95 指标衡量是模型在 IoU 阈值从 0.5 到 0.95 范围内的平均精度。

平均交并比反映各类别预测区域与真实区域的重合程度：

$$MIoU = \frac{1}{(k+1)} \sum_{(i=0)}^{1} \frac{p_{ii}}{\sum_{j=0}^{k} p_{ij} + \sum_{j=0}^{k} p_{ji} - p_{ii}} \tag{3-16}$$

式中分母项通过双重求和消除重复计数的 p_{ii}。式中，$\sum_{j=0}^{k} p_{mj}$ 包含真实类别 m 的像素总数；$\sum_{j=0}^{k} p_{jm}$ 则为预测为类别 m 的像素总数。

实验环境参数详见表 3-2 所示硬件配置。

表 3-2 实验平台参数表

实验设备	型号
设备	Dell Precision 7960 塔式
平台	Ubuntu 20.04
CPU	Intel(R)至强® W5-3423
GPU	A100 Tensor Core GPU
System memory	128GB
SSD	Dell M.2 Class 40 2280 - 2TB

3.3.2 基于机器学习的预处理结果展示

本节提出一种基于机器学习的自适应参数变换图像预处理算法，相较于伽马变换、对数变换等传统线性变换方法，该方案通过动态优化参数显著提升了工业场景下机器人无序抓取数据集的适应能力。

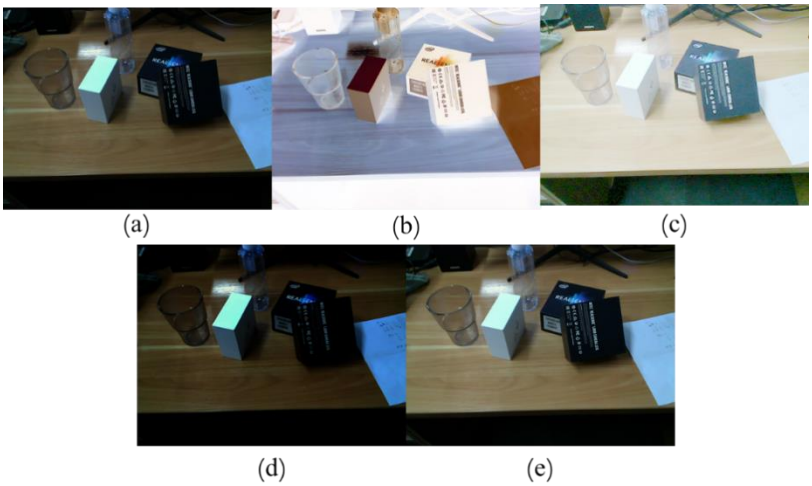


图 3-10 不同数据增强的测试结果

(a)原图；(b)线性变换；(c)对数变换；(d)伽马变换；(e)复合变换

如图 3-10 所示，通过随机采样数据集图像进行预处理对比实验可见，本算法在图

像增强效果及目标特征可辨识度方面展现出更优的泛化性能。特别地，其参数自适应性有效克服了传统方法需人工预设参数的局限性，经处理后的图像在目标检测任务中呈现出更高的轮廓清晰度与纹理区分度，为后续识别模型提供了更高质量的输入数据。

实验在本次数据集中随机选取了 50 张图像，进而比较了伽马变换、对数变换、线性变换和自适应参数变换在图像相似度评价和处理运行时间上的量化性能。如表 3-3 所示，通过比较图像相似度评价和处理运行时间上这两方面，可以可以发现，实验方法虽然在运行时间方面较长，但差距较小，且 SSIM 评分最高，可以得出自适应参数变换可以有效提高工业机器人无序抓取数据集的图像预处理表现。

表 3-3 不同预处理效果对比

指标	线性变换	对数变换	伽玛变换	Ours
SSIM	0.218	0.475	0.587	0.872
Runtime/s	0.056	0.160	0.143	0.173

3.3.3 AC-YOLO 目标检测实验结果展示

在本节中，通过消融实验分析了在 Neck 模块中引入 GSConv 结构以及结合 FedFusion 设计 ESlim-Neck 解码器的性能表现。从表 3-4 可以看出，在工业机器人无序抓取任务中，包含 GSConv 块的网络架构与包含 ESlim-Neck 的网络网络架构在相同的实验条件下，模型参数体量分别为原始 YOLO 的 71.80%和 54.56%。尽管模型复杂度显著降低，其识别精度仍高于原始 YOLO。

表 3-4 GSConv 与 ESlim-Neck 的消融实验

项目	YOLOv8	GSConv	ESlim-Neck
pa	0.764	0.803	0.794
mpa	0.751	0.794	0.781
miou	0.776	0.812	0.827
params	71.3MB	51.2MB	38.9MB

训练过程如图 3-11 所示。实验结果显示，定位损失、分类损失和置信度损失均较快收敛，模型在测试集上的准确率达到 0.817，召回率为 0.791，mAP50 为 0.836，mAP50-95 为 0.387，表明模型性能优越。进一步分析表明，GSConv 结构的引入增强了特征表达能力，而 ESlim-Neck 模块优化了特征融合效率，从而显著提高了模型在无序抓取任务中的适应性和鲁棒性。

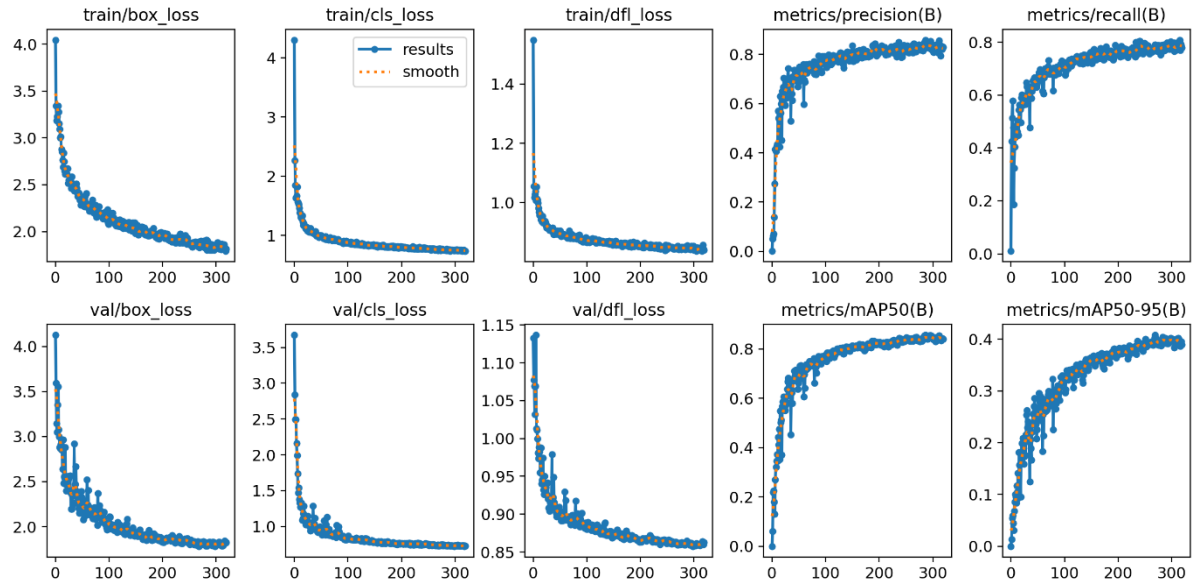
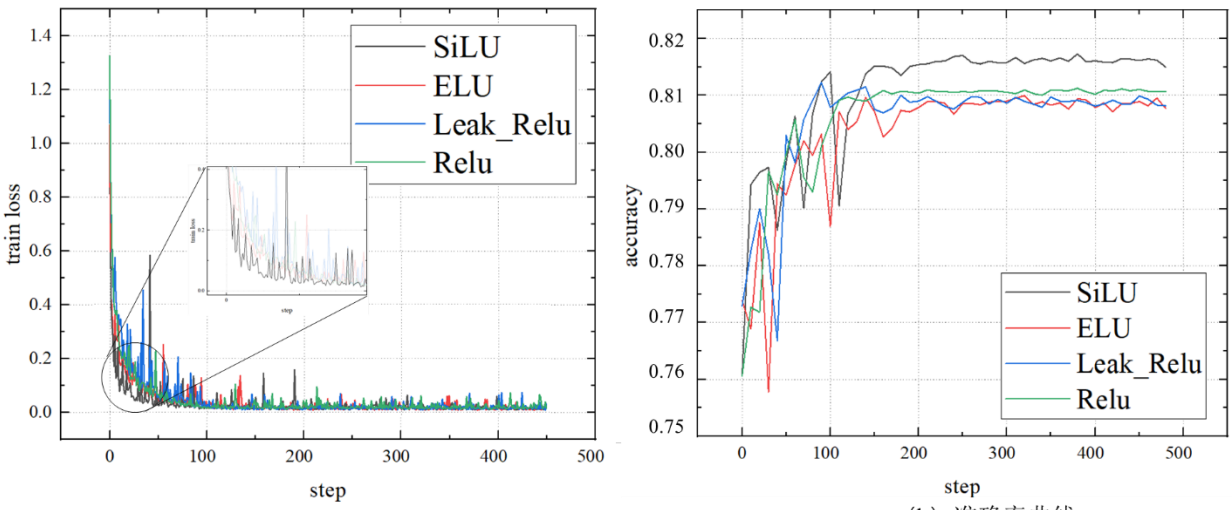


图 3-11 AC-YOLO 网络训练过程图。包含训练集和验证集的 box 损失、分类损失和置信度损失曲线图，以及 AC-YOLO 模型的准确率、召回率、mAP50 和 mAP50-95。

为评估激活函数对 AC-YOLO 网络检测性能的影响，本研究对比分析了 SiLU、ELU、Leaky_ReLU 和 ReLU 四种典型激活函数的训练收敛特性。如图 3-13(a)(b)所示，通过可视化训练损失曲线与验证集精度曲线可见，在工业机器人无序抓取场景中，采用 SiLU 激活函数的网络展现出更优的特性，其训练损失值较快收敛同时目标检测精度高于其他。这种性能提升主要源于 SiLU 函数连续可微的 S 型结构，相较于传统分段线性激活函数，能更有效地传递梯度信息并缓解神经元饱和现象。

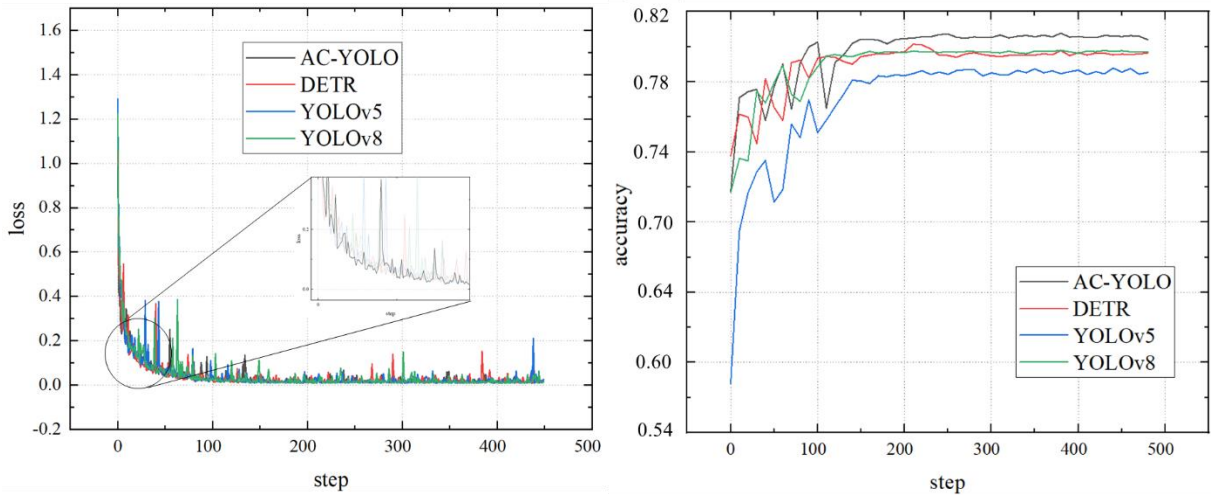


(a) 训练损失曲线　　(b) 准确率曲线

图 3-12 不同激活函数的训练过程图

3.3.4 不同单模态算法的识别结果

该研究比较了当下流行的目标检测网络在工业机器人无序抓取数据集上的结果，以验证 AC-YOLO 的性能。如图 3-13(a) 和图 3-13(b) 所示，可以看出，与其他网络相比，AC-YOLO 网络可以在学习过程中更快地趋于稳定，同时在准确率方面的表现也更高。



(a) 损失曲线　　(b) 准确率曲线

图 3-13 目前流行模型的损失和训练过程

后续实验中，为了全面评估改进的 AC-YOLO 模型在无序抓取任务中的性能，我们选取了不同的目标检测网络进行对比实验，包括 DEtection TRansformer[69] (DETR)、原始 YOLOv8[43] 以及我们提出的改进模型。为了直观展示不同模型的检测效果，我们在图 3-14 中给出了它们在相同测试图像上的检测结果。

具体而言，第一列显示的是原始输入图像，第二列至第四列分别展示了 DETR、YOLOv8 以及我们改进模型的检测结果。从图中可以观察到，各个模型均能识别出场景中的主要目标（如书本、杯子、椅子等），但检测精度和目标边界框的定位存在差异。例如，DETR 在某些目标的置信度较低，且可能存在漏检的情况；YOLOv8 相较于 DETR 在检测速度和准确率上有所提升，但仍然存在部分目标未能准确识别或定位偏差的情况。而我们提出的 AC-YOLO 模型在多个目标类别的检测上均表现出更高的置信度，目标边界框更加准确，且能够检测到更完整的目标信息，这表明我们的改进方法在检测任务上取得了更好的效果。综上，图 3-14 直观地展示了不同目标检测网络在相同场景下的检测结果对比，进一步验证了我们方法的有效性。

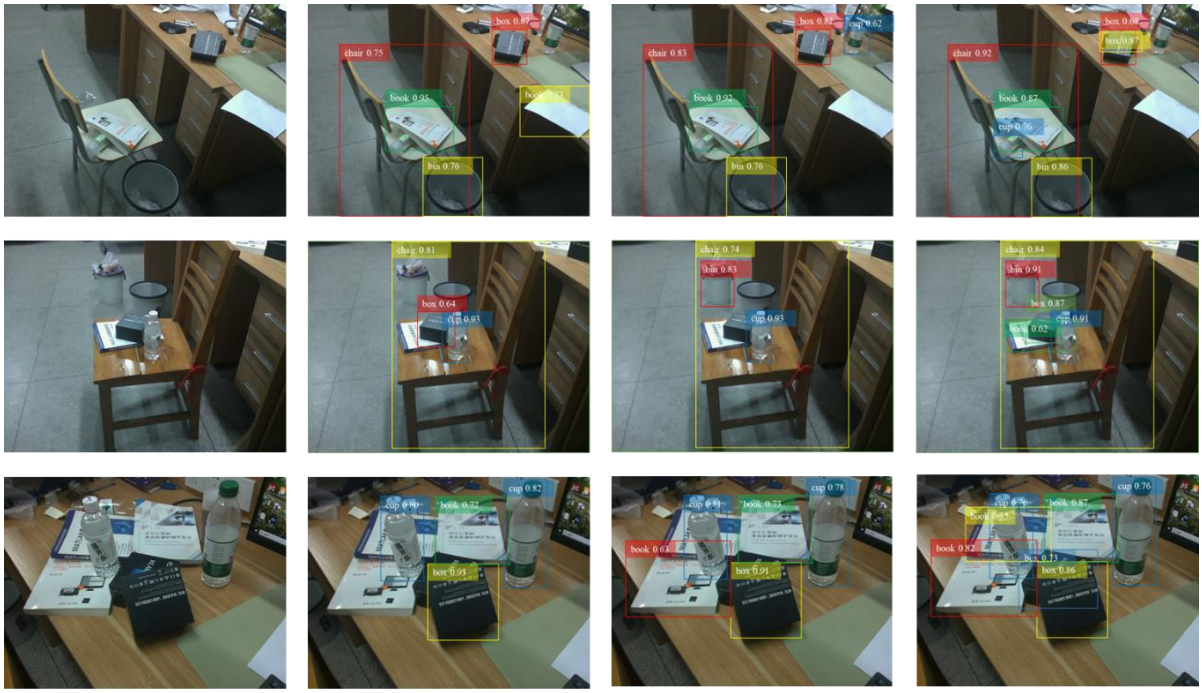


图 3-14 不同网络的检测结果

表 3-5 中的结果量化评价了 AC-YOLO 网络的性能。在同样的工业机器人无序抓取识别任务中，在 PA、mPA 和 mIOU 这三个方面的量化得分可以分析出，AC-YOLO 相较于 YOLOv5、YOLOv8 和 DETR 均展现出系统性优势。具体而言，在像素准确度方面，AC-YOLO 以 81.7%的得分显著优于基准模型，表明其在像素级分类任务中具有更强的判别能力；在平均像素准确度指标上，AC-YOLO 达到 80.9%，较次优模型 YOLOv8 提升 2.7 个百分点，反映其在不同类别间特征提取的均衡性显著增强，有效缓解了传统模型在多目标场景下的类别偏差问题；尤其值得注意的是，AC-YOLO 在 mIOU 指标上以 92.1%的突破性表现超越对比模型，表明网络通过融合注意力机制与多尺度特征金字塔结构，显著提升了目标边界的定位精度与区域重叠率。实验数据验证了 AC-YOLO 网络架构改进的有效性，其性能优势源于特征融合卷积层优化、跨层特征交互机制的引入以及解码阶段的空间注意力重加权策略，这些技术创新共同增强了模型对复杂工业场景中形变、遮挡目标的鲁棒性表征能力。

表 3-5 当前流行网络的实验结果对比

项目	YOLOv5	YOLOv8	DETR	AC-YOLO
pa	0.783	0.796	0.794	0.817
mpa	0.775	0.782	0.763	0.809
miou	0.853	0.875	0.884	0.921

3.4 本章小结

本章研究了单模态目标检测算法的改进，提出了基于YOLOv8的AC-YOLO网络，显著提升了工业机器人无序抓取任务的目标检测性能。在 Neck 模块中引入 GSConv 和 ESlim-Neck 结构，并结合三维空间注意力机制 TDSA，使模型在复杂背景中表现出更强的鲁棒性与检测精度。此外，通过对模型激活函数与损失函数的优化，实现了检测速度与精度的双提升。实验结果表明，AC-YOLO 在 mAP 和模型大小等指标上均优于现有基准模型，体现了 AC-YOLO 网络识别算法在无序抓取任务上的优秀表现。

第4章 基于多模态特征的三维分割识别算法研究

4.1 引言

本章提出的无序抓取目标识别与分割旨在解决物体分割与提取任务中的目标与背景划分、小目标分割、目标遮挡识别等问题。主要流程是是对色彩数据和深度图进行分割，生成掩码图，作为提示，输入到分割识别网络中，另一方面，通过对色彩数据和深度图进行对齐操作，生成点云，对其进行滤波等预处理操作，输入到网络中，完成对目标的三维分割识别。图 4-1 展示了本次实验的工作流程。

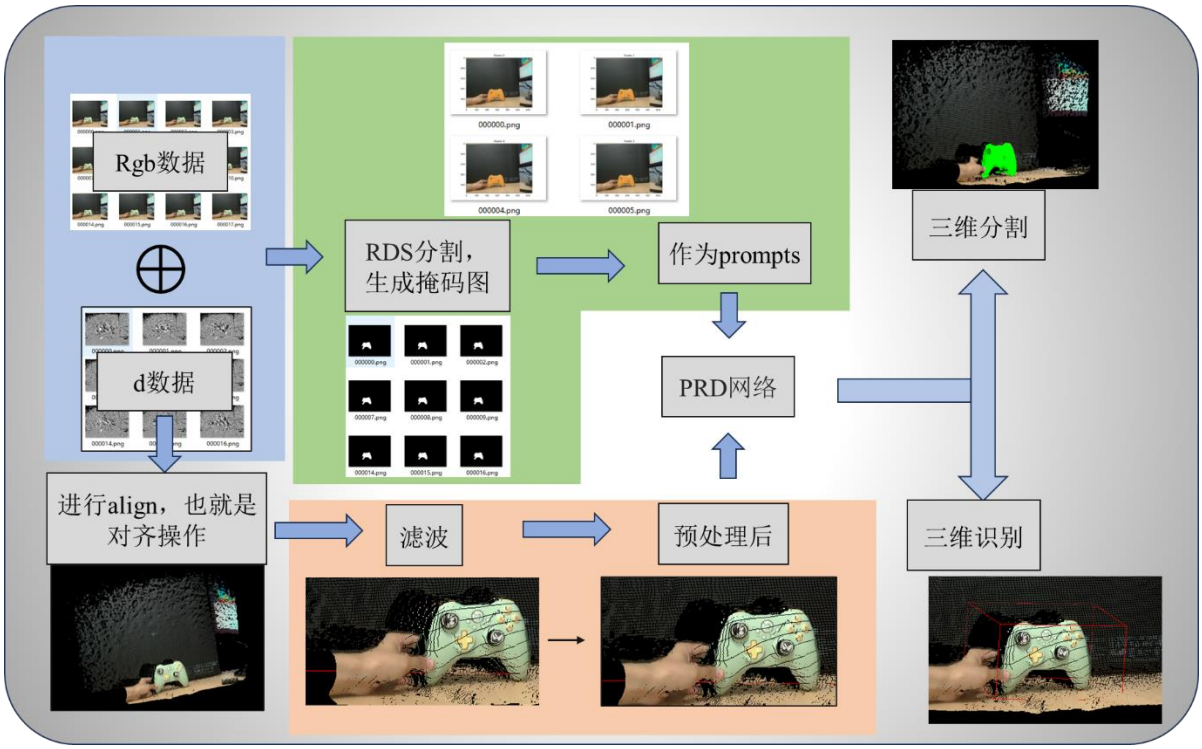


图 4-1 三维分割识别工作流程图

4.2 RGBD 数据三维对齐生成点云

4.2.1 小孔成像原理

小孔成像模型描述了空间点通过摄像头投影到图像平面上的过程。该模型通常通过一个简单的透视投影关系来表达，假设光线从三维空间的点穿过摄像头的投影中心，最后在图像平面上形成投影。在此过程中，存在以下三组坐标：三维空间坐标系：$P_{ir}=[X,Y,Z]^T$，描述点在深度摄像头坐标系下的三维空间位置。图像平面坐标系：

$p_{ir}=[x,y,1]^T$，描述点在图像平面上的投影位置。深度值：Z 表示点到摄像头的深度，单位通常为米。

小孔成像模型通过内参矩阵 H_{ir} 将三维点 P_{ir} 投影到图像平面上：

$$p_{ir}=H_{ir}P_{ir} \tag{4-1}$$

根据小孔成像原理，三维点的投影过程如下：

1.将三维点 P_{ir} 转化为齐次坐标 $[X,Y,Z,1]^T$。

2.使用内参矩阵 H_{ir} 对点进行透视投影。

3.得到图像平面上的齐次坐标 $p_{ir}=[x,y,1]^T$，其中：$x=\frac{(f_x\cdot X)}{Z}+c_x$ ，$y=\frac{(f_y\cdot Y)}{Z}+c_y$ "

反过来，如果知道图像上的投影点 $p_{ir}=[x,y,1]^T$ 和深度值 Z，可以通过内参矩阵的逆矩阵 H_{ir}^{-1} 恢复三维点的空间坐标 P_{ir}：

$$P_{ir}=H_{ir}^{-1}p_{ir} \tag{4-2}$$

此时的计算为：

1.从 p_{ir} 中分离出像素坐标 x,y 和深度值 Z。

2.通过矩阵求逆关系恢复 X,Y,Z：

$$\begin{aligned} X&=\frac{(x-c_x)\cdot Z}{f_x}\\ Y&=\frac{(y-c_y)\cdot Z}{f_y}\\ Z&=Z \end{aligned} \tag{4-3}$$

4.2.2 深度图与彩色图对齐

由于数据集包含色彩图像和深度图像两部分，后续需要借助点云形式来处理数据，故使用点云对齐操作。针对包含可见光与深度信息的异构数据集，本研究采用三维空间对齐技术实现数据融合，具体流程如下：

1. 深度信息采集。通过深度传感器获取目标场景的深度映射图，记录每个像素对应的空间距离信息；2. 色彩信息提取。同步采集 RGB 三通道光学影像，获取与深度图时空同步的表面纹理特征；3. 彩色点云生成。建立深度-色彩映射关系的关键操作包含：1）投影坐标构建：基于针孔相机模型，将深度图像素坐标(*x*,*y*)与其深度观测值 z 组合

为投影向量 $p=(x, y, z)$；2）空间坐标反演：通过 RealSense L515 相机标定参数中的内参矩阵 H，如式(4-4)所示，乘以 p 得到对应的空间点坐标：

$$H_{rgb}=\begin{bmatrix} f_{x\_rgb} & 0 & c_{x\_rgb} \\ 0 & f_{y\_rgb} & c_{y\_rgb} \\ 0 & 0 & 1 \end{bmatrix} \tag{4-4}$$

式中，对应的本次 realsenseL515 实验平台的数据见表 4-1。

表 4-1 相机内参矩阵值

内参	值
f_{x_rgb}	596.9665
c_{x_rgb}	326.0782
f_{y_rgb}	597.5341
c_{y_rgb}	239.8642

3）由于 P 是该点在 realsense 坐标系下的坐标，因此需要将其转换到 RGB 摄像头的坐标系下，具体的，就是乘以一个旋转矩阵 R，再加上一个平移向量 T，得到 P_{rgb}；4）用内参矩阵 H 乘以 P_{rgb}，得到 p_{rgb}，p_{rgb} 也是一个三维向量，其 x 和 y 坐标即为该点在 RGB 图像中的像素坐标，读取该像素点的色彩值；5）对深度图像中的每一个像素都做上述操作，得到配准后的深度图。经上述操作后得到的对齐效果如图 4-2。

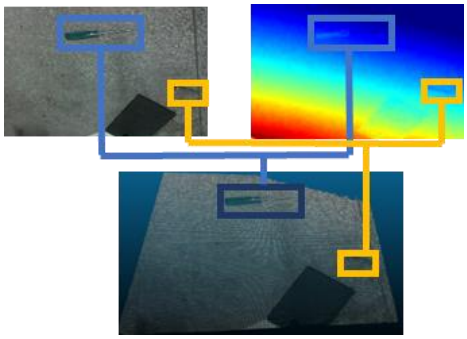


图 4-2 对齐效果示意图

4.3 三维分割网络算法设计

4.3.1 网络整体结构

首先介绍分割网络，图像分割在视觉理解中具有至关重要的作用，广泛应用于自然图像与工业图像等多个领域。近年来，视觉基础模型(Vision Foundation Models，VFMs)在多个任务上取得了优异的性能，尤其是 Segment Anything Model(SAM1)及其后继版本 SAM2 展示了强大的分割能力。本次重点在于如何改进 SAM 模型，使其在本次

无序抓取的图像分割任务时更加灵活、适应性更强。

RDS(RGB-D Segment)整体框架包含编码器(Encoder)、DAE 块(Downsampling And Extraction of multi-scale features)、适配器(Adapters)、解码器(Decoder)以及分割输出部分。整个网络采用了 U 形结构，输入图像经过编码器逐步下采样并提取特征，随后通过解码器逐步上采样生成分割结果。编码器部分使用 Hiera 骨干网络来代替 SAM 的视觉 Transformer(ViT)作为编码器，通过引入适配器模块(Adapter)，实现了高效的参数微调，使得参数更适应本次指定的分割任务。之后通过 DAE 块减少通道数量并增强特征表示。DAE 块在有助于在保持分割信息的同时，减小计算复杂度。在编码器中提取的多尺度特征经 DAE 块处理后，被输入到解码器中。解码器采用了经典的 U-Net 设计，逐步对特征进行上采样。网络整体结构如图 4-3：

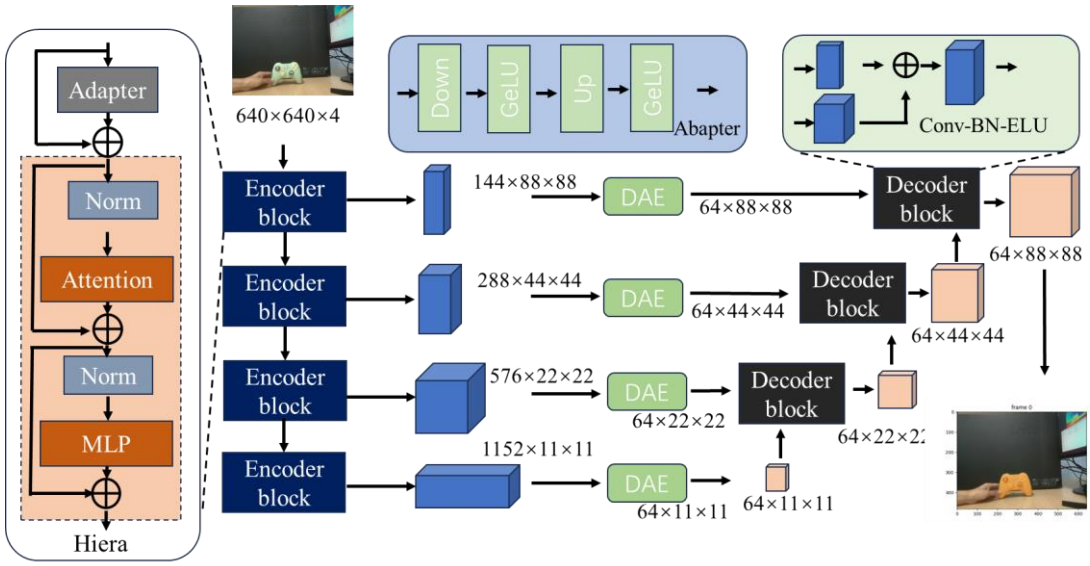


图 4-3 RDS 网络结构图

4.3.2 网络细节模块设计

具体而言，编码器部分，Hiera 具有层次化结构，更适合捕捉多尺度的特征，适合之后的 U 型网络结构，RDS 网络的编码器采用了 Hiera 骨干网络。与 SAM 预训练的 ViT 编码器不同，Hiera 具有层次化结构，更适合捕捉多尺度的特征，这对 U 形网络设计尤为关键。给定输入图像$I \in R^{3\times H\times W}$，其中$H$和$W$分别代表图像的高度和宽度，Hiera 输出四个层次化的特征$X_i \in R^{C_i\times H/2^{i+1}\times W/2^{i+1}}$（$I \in 1,2,3,4$），其中$C_i$是每个层次中的通道数量。具体地，对于本次网络，通道数C_i分别为 144, 288, 576 和 1152。

适配器部分，由于 Hiera 的参数很大（Hiera 为 214M），因此执行完全微调不总是

可行的。因此，需要冻结 Hiera 的参数，并在 Hiera 的每个多尺度块之前插入适配器，以实现参数高效的微调。在框架中的每个适配器都由一个用于下采样的线性层、一个 GeLU 激活函数和另一个用于上采样的线性层和一个最终的 GeLU 激活组成。

DAE 块部分，DAE 块的作用是减少通道数量,将其减少到 64 并增强特征表示。这种处理有助于减少计算量，同时保留关键的分割信息。这一部分灵感来自于 Inception，具有不同内核的多分支卷积层和尾随的膨胀池化和卷积层。进一步来说，首先，在每个分支中采用了瓶颈结构，由 1×1 卷积层组成，之后的分支分别为 1×1，3×3 和 5×5 卷积层，各分支下一步都采用 3×3 卷积层，但步长分别为 1，3 和 5，然后各分支由一个 1×1 卷积层进行汇总，并借鉴 Resnet 残差连接的思想，和 input 进行连接。最后，经过 ELU 函数激活。DAE 块的设计见图 4-4 中提出的网络架构。

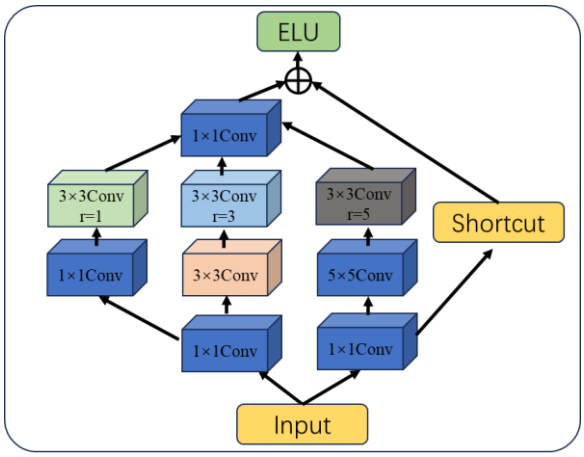


图 4-4 DAE 模块整体架构图

解码器部分，解码器每一层中包含两个“Double Conv”模块，即两层卷积+批量归一化+ReLU 激活函数的组合，并将特征逐步还原到输入图像的尺寸，其中'Conv' 表示 3×3 卷积层，'BN' 表示批量归一化。每个解码器模块的输出特征通过 1×1 卷积分割头，产生分割结果 S_i（$i \in 1,2,3$），具体地，对于本次 RDS 网络对应分辨率为：88×88，44×44，22×22，分这些分割结果 S_i 与真实标签(GT，Ground Truth)进行迭代。

4.3.3 损失函数

二值交叉熵(Binary Cross Entropy, BCE)损失函数是二值图像分割中的常用损失函数之一。它用于衡量模型预测的二值输出与实际标签之间的差异。在二值图像分割任务中，目标是将每个像素分类为两类之一：前景或背景。通常，我们的模型输出是一个与输入图像大小相同的概率图（概率矩阵），表示每个像素属于前景（通常标记为 1）

的概率。BCE 损失函数通过比较模型输出的概率值和真实的标签值来衡量模型的表现。其数学表达式如下:

$$\mathrm{BCE}=-\frac{1}{N}\sum_{i=1}^{N}\left[y_i\cdot\log\left(\hat{y}_i\right)+\left(1-y_i\right)\cdot\log\left(1-\hat{y}_i\right)\right] \tag{4-5}$$

其中，N是图像中的像素总数；y_i是第i个像素的真实标签（0 或 1）；$\hat{y}_i$是模型预测的第i个像素属于前景的概率（0 到 1 之间的值）。

但对于传统的 BCE Loss，存在：1.只是简单的将每个像素求 BCE 再平均，忽视了目标对象的结构；2.对于小目标而言，整张图像的 loss 会被背景类所主导，导致难以对前景进行学习；3.对象的边缘位置像素非常容易分类错误，不应该与其他位置像素一样给予相似的权重。而解决方案自然是对不同位置的像素进行加权。具体来说，权重最高的地方应该是地方的边缘位置，离边缘越远则越低。针对本次三维分割网络的 weighted BCE Loss 表达式如下：

$$L_{wbce}^{s}=-\frac{\sum_{i=1}^{H}\sum_{j=1}^{W}\left(1+\gamma\alpha_{ij}\right)\sum_{l=0}^{1}\mathbf{1}\left(g_{ij}^{s}=l\right)\log\Pr\left(p_{ij}^{s}=l|\ \Psi\right)}{\sum_{i=1}^{H}\sum_{j=1}^{W}\gamma\alpha_{ij}} \tag{4-6}$$

其中，$\mathbf{1}(\bullet)$是指示函数；γ 是超参数；符号 $l\in\{0,1\}$ 表示两种标签；p_{ij}^{s}和 g_{ij}^{s} 是图像中像素位置(i, j)的预测和真实值；Ψ 表示模型的所有参数；而$\Pr\left(p_{ij}^{s}=l|\ \Psi\right)$表示预先确定的概率；$\alpha_{ij}$ 指的就是(i, j)位置像素的权重。如果不进行加权的话，则相当于 α_{ij} 恒为 1。α_{ij} 按下式计算：

$$\alpha_{ij}^{s}=\left|\frac{\sum_{m,n\in A_{ij}}g_{mn}^{s}}{\sum_{m,n\in A_{ij}}1}-g_{ij}^{s}\right| \tag{4-7}$$

其中，g_{ij}^{s}表示(i, j)位置的真值(1 或 0，对应前景或背景)；A_{ij}表示(i, j)位置周围的像素；m，n 表示位置(m, n)。为了更好理解，取特殊值来讨论。假设g_{mn}^{s} 均为 0，g_{ij}^{s}为 1，相当于当前像素为前景，而周围像素均为背景，是一个小目标，应该给予高权重。类似的，如果g_{mn}^{s}均为 0，g_{ij}^{s}也为 0，表明当前位置与周围位置均为背景，此时对应着一个低权重。

另一方面，Intersection over Union (IoU) Loss 是二值图像分割任务中的一种损失函

数，用于评估模型分割结果与真实标签之间的重叠程度。它基于 IoU（也称为 Jaccard 系数），直接优化预测与真实标签之间的区域相似性。IoU 损失函数在前景区域较少或类别不平衡的情况下有较好的性能，因为它对前景区域的重叠程度具有敏感性。在二值图像分割中，预测结果通常是一个概率图，表示每个像素属于前景（标记为 1）的概率。IoU 度量则用于评估预测结果的前景区域与真实标签的前景区域之间的相似度。IoU 的计算公式为：

$$IoU = |A \cup B| \ / \ |A \cap B| \tag{4-8}$$

其中，｜A∩B｜表示预测前景与真实前景的交集像素数（即两个集合中共有的前景像素数）；｜A∪B｜表示预测前景和真实前景的并集像素数（即两者中至少有一个是前景的像素数）。而 IoU 的 loss 为

$$L_{IoU} = 1 - \frac{\sum_{i=1}^{N} p_i g_i}{\sum_{i=1}^{N} \left(p_i + g_i - p_i g_i \right)} \tag{4-9}$$

接下来是类似的思想，为了更加注重边缘位置，分割小目标等，故采用 weighted IoU Loss。数学表述式如下

$$L_{wiou}^{s} = 1 - \frac{\sum_{i=1}^{H}\sum_{j=1}^{W} \left(gt_{ij}^{s} \times p_{ij}^{s} \right) \times \left(1 + \gamma \alpha_{ij}^{s} \right)}{\sum_{i=1}^{H}\sum_{j=1}^{W} \left(gt_{ij}^{s} + p_{ij}^{s} - gt_{ij}^{s} \times p_{ij}^{s} \right) \times \left(1 + \gamma \alpha_{ij}^{s} \right)} \tag{4-10}$$

式中，各个符号代表与式(4-6)类似，在此不再赘述。公式的思想依旧为离边缘近的像素对 IOU 计算的贡献更大。

最终的损失函数为：

$$L_{\text{RDSloss}} = L_{wbce}^{s} + L_{wiou}^{s} \tag{4-11}$$

4.4 基于 PointNet 网络的三维识别算法

4.4.1 网络整体结构

所提出的 PRD(Point Rgb-D)网络旨在预测给定场景的单视图 RGB-D 图像，预测在视锥体内 3D 场景中每个类别标签 $C = \{c_i\}_{i=1}^{N+1}$。这里，N 是对象类别的数量，c_{N+1} 表示背景空位置，而 c_i 表示不同类别的对象占据的位置。

目标识别网络首先使用对色彩图和深度图进行编码，本次采用改进的 resnet 网络作

为编码器，然后将 RGB 特征 f_{rgb}，深度特征 f_d 和过程中的 f_{fuse} 传递至下一模块，其中 f_{fuse} 如下式

$$f_{fuse}=\sum_{i=1}^{4}\{f_{rgb}^{2^{i+1}}+f_{d}^{2^{i+1}}\} \tag{4-12}$$

其中，i 代表不同分辨率的特征层；2^{i+1} 分别表示 1/4, 1/8, 1/16, 和 1/32 特征层，之后是 FGF(Feature-guided Fusion)模块，前述三种特征通过残差块连接。具体来说，f_{fuse} 使用线性插值来让较小分辨率的融合特征进行上采样，以匹配 RGB 特征的分辨率。随后，上采样的特征与 RGB 特征连接，并将生成的特征作为残差块 $ResBlock_{rgb}$ 的输入。然后，$ResBlock_{rgb}$ 生成相应的融合特征 d_{rgb} 和置信度图 c_{rgb}。同样的，$ResBlock_d$ 迭代生成相应的融合特征 d_d 和置信度图 c_d。这个过程可以用式表示。

$$\begin{aligned}\left(c_{rgb},d_{rgb}\right)&=\mathrm{ResBlock}_{rgb}\left(cat\left(f_{fuse},f_{rgb}\right)\right)\\\left(c_{d},d_{d}\right)&=\mathrm{ResBlock}_{d}\left(cat\left(f_{fuse},f_{d}\right)\right)\end{aligned} \tag{4-13}$$

另一方面，RGB 数据和 D 数据进行对齐融合后组成点云数据，作为输入。输入层以单帧点云数据集（表征为 n×3 的二维张量，n 为点云数量，3 对应空间坐标维度）为基础架构。如图 4-5 所示，系统首先通过与 T-Net 网络学习的变换矩阵相乘实现空间对齐，该机制通过建立可学习的仿射变换空间，有效增强模型对刚性变换的鲁棒性。经多层感知机逐层提取点云特征后，再次通过特征域 T-Net 进行特征空间校准，形成双重几何不变性保障体系。在特征的各个维度上执行最大池化操作，得到最终的全局特征。最后，将全局特征通过 MLP 进行分类预测，并与网络输出的分类分数进行加权平均。

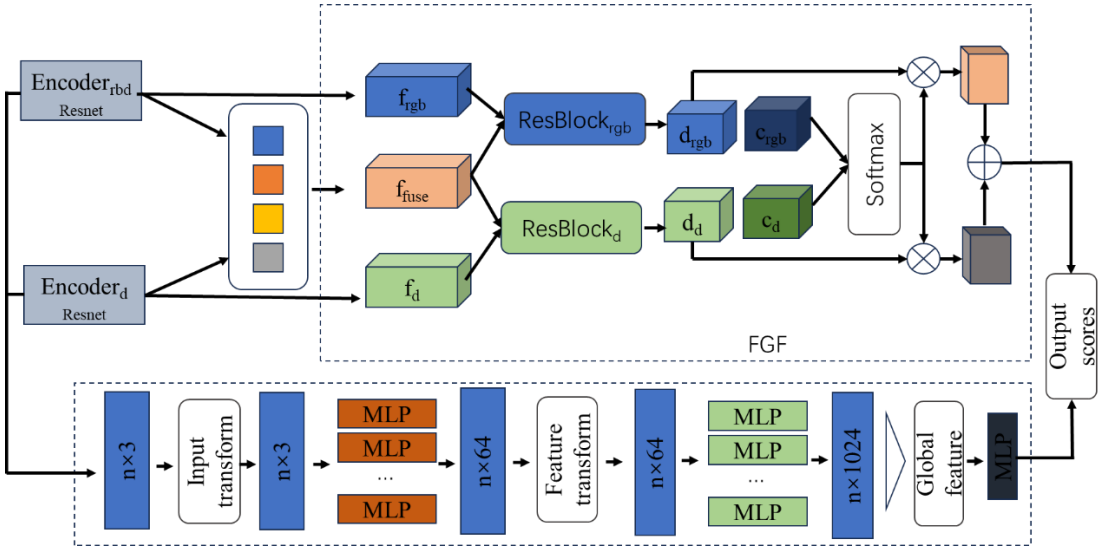


图 4-5 PRD 网络结构图

4.4.2 E-Resnet 模块设计

E-ResNet 模块经输入后有五个部分最后给出输出。首先是 Stage 0 进行输入数据的预处理，后 4 个 Stage 由多个参数不同的 Bottleneck（以下简称 BTNK）组成。Stage 1 至 4 分别包含 3、4、6、3 个 BTNK。整体架构如图 4-6 所示：

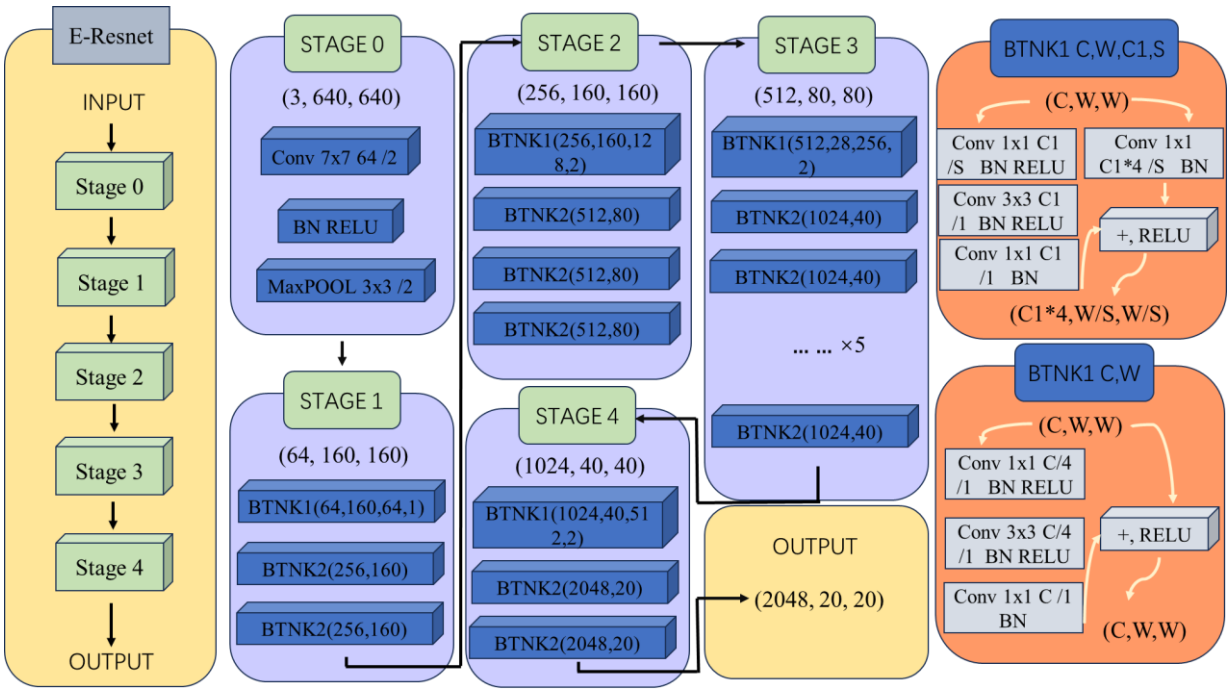


图 4-6 E-ResNet 整体网络架构图

具体而言，Stage 0：(3,640,640)指输入通道数、高和宽，即(C,H,W)。本次的输入高度和宽度相等，故用(C,W,W)简化表示。该 stage 采用经典卷积-归一化-激活-池化流程。卷积层中的参数依次为卷积核大小，数量和步长。卷积层的输出尺寸按式子计算：

$$output\_size = \frac{input\_size + 2 \times padding - kernel\_size}{stride} + 1 \tag{4-14}$$

BN 层也就是批量归一化层。其目的是预处理使我们的 Batch 的特征满足正态分布规律，这样能够加速网络的收敛。首先会统计每个通道数目所有点的像素值，求得均值和方差，然后在每个通道上分别用该点的像素值减均值除方差得到该点的像素值。也就是，给定输入最小批次量 $\mathcal{B} = \{x_{1...m}\}$，求最小批次均值，方差，然后进行标准化，完成所给输出 $\{y_i = BN_{\gamma,\beta}(x_i)\}$。过程如下式。

$$\begin{aligned}\mu_{\mathcal{B}} &\leftarrow \frac{1}{m}\sum_{i=1}^{m} x_i \\ \sigma_{\mathcal{B}}^2 &\leftarrow \frac{1}{m}\sum_{i=1}^{m}\left(x_i-\mu_{\mathcal{B}}\right)^2 \\ \hat{x}_i &\leftarrow \frac{x_i-\mu_{\mathcal{B}}}{\sqrt{\sigma_{\mathcal{B}}^2+\epsilon}} \\ y_i &\leftarrow \gamma\hat{x}_i+\beta \equiv BN_{\gamma,\beta}\left(x_i\right)\end{aligned} \tag{4-15}$$

输入 x_i 经过 BN 层后，还需要用通过 γ 调整数据的方差，通过 β 调整数据的均值。而 γ 和 β 在反向传播中学习。本次 E-Resnet 还使用了偏置，如下式子：

$$\begin{aligned}y_i &= \frac{x_i-u}{\sqrt{\sigma^2+\varepsilon}} \\ y_i^b &= \frac{x_i^b-u^b}{\sqrt{\sigma^2\left(x^b\right)+\varepsilon}} \\ x_i^b-u^b &= x_i+b-(u+b)=x_i-u \\ \sigma^2\left(x^b\right) &= \frac{1}{m}\sum_{i=1}^{m}\left(x_i^b-u^b\right)^2 \\ &= \frac{1}{m}\sum_{i=1}^{m}\left(x_i+b-(u+b)\right)^2 \\ &= \frac{1}{m}\sum_{i=1}^{m}\left(x_i-u\right)^2 \\ &= \sigma^2(x)\end{aligned} \tag{4-16}$$

RELU 指 ReLU 激活函数。该 stage 中第 2 层为 MAXPOOL，即最大池化层。

之后，Stage 1 至 Stage 4 采用改进型残差结构，核心模块为两类 BK 单元，其数学形式化定义如下：

Bottleneck 类型I(BTNK1)：设输入张量 x∈ℝ^{C×W×W}，经残差路径 F(x)=W_2(σ(W_1(x)))和恒等路径 G(x)=W_3(x)处理后，输出为：

$$y=\sigma(F(x)+G(x)) \tag{4-17}$$

其中 W_1∈ℝ^{C1×C×1×1}、W_2∈ℝ^{C2×C1×3×3}、W_3∈ℝ^{C_{out}×C×1×1}为卷积权重；σ表示 ReLU 激活函数。该结构通过可学习投影矩阵 W_3 实现维度匹配，满足当 C_{out}≠C 时残差连接可行性。

Bottleneck 类型Ⅱ(BTNK2)：当输入输出通道数一致时，简化为：

$$y=\sigma(F(x)+x) \tag{4-18}$$

此时 G(x)=x 保持恒等映射，避免引入额外参数。各 stage 参数配置规律如表 4-2 所

示：

表 4-2 stage 参数配置细节

Stage	S 值	通道关系(C1/C)	下采样特性
1	1	C1=C=64	空间维度保持(W=160)
2-4	2	C1=0.5C	空间维度缩减(W←W/2)

特别地，Stage 1 的 BTNK1 模块采用等通道设计(C1=C)，而后续 stage 通过降维策略(C1=0.5C)降低计算复杂度。下采样操作通过设置卷积步长 S=2 实现，使特征图在 Stage 2-4 中逐级压缩。这种分层特征提取机制在保证信息完整性的同时，有效提升了模型对多尺度目标的表征能力。

4.4.3 损失函数

损失函数包括两个部分，包含分类损失和 3D 包围框估计。分类损失包括多类交叉熵损失分类损失，以及特征变换矩阵的损失函数。3D 包围框估计是基于交并比损失函数所改进。多类交叉熵损失是用于多分类问题的一种。通过衡量模型预测与真实标签之间的误差，采用多类交叉熵损失，具体公式如下

$$\mathcal{L}_{cla}=\frac{1}{2}\sum_{i=1}^{2}\mathrm{CE}(\hat{s}^{i},s^{i})=-\frac{1}{N}\sum_{i}\sum_{c=1}^{N}y_{ic}\log(p_{ic}) \tag{4-19}$$

式中，$\hat{s}^{i}$ 表示模型预测标签；s^{i} 表示物体标注类别；N 表示类别数量；y_{ic} 是符号函数（取值 0 或 1），如果样本 i 的真实类别等于 c 取 1，否则取 0；p_{ic} 表示观测样本 i 属于类别 c 的预测概率。

另，因为特征变换矩阵的维度较大，为了能够保证网络的训练效果，需要添加正则项。正则项有着用于约束特征变换矩阵的性质。损失函数的数学表达式如下：

$$\mathcal{L}_{reg}=\| TT^{T}-I\|_{F}^{2} \tag{4-20}$$

其中，T是 T-Net 输入的变换矩阵；I是单位矩阵（与 T维度相同）；$\|\bullet\|_{F}^{2}$ 表示 Frobenius 范数的平方。这个正则项用于约束 T接近正交，即 $TT^{T}\approx I$，从而避免不稳定变换。

交并比是目标检测任务中用于评估预测框与真实框重合度的重要指标，其数学定义为：

$$IoU = \frac{\left|T_{3\mathrm{Dbox}} \cap P_{3\mathrm{Dbox}}\right|}{\left|T_{3\mathrm{Dbox}} \cup P_{3\mathrm{Dbox}}\right|} \tag{4-21}$$

然而，传统 IoU 存在以下缺陷：1.梯度消失问题：当预测框与真实框无交集时，IoU 值为 0，导致梯度无法回传，模型参数无法更新。2.区分能力不足：对于 IoU 相同的预测框，无法进一步区分其方向、形状及尺度差异。为解决上述问题，本研究提出一种改进的 IoU 损失函数，使其综合考量了预测框和真实框间的重叠量及其相似性，其公式如下：

$$\mathcal{L}_{\mathrm{box}} = 1 + \lambda\Theta + \frac{\vartheta\left(C^{\mathrm{p}}, C^{t}\right)}{l} - IOU \tag{4-22}$$

式中，λ 为权重系数，用于平衡损失项的贡献；Θ 为两框的形状相似性度量，通过最小凸包区域的最长距离归一化计算，两者的计算公式如式(4-23)所示；C^{p} 和 C^{t} 分别为预测框和标签框的几何中心位置；ϑ 为预测点和标签点直接的欧氏距离；l 为所有最小凸包区域的最长距离。

$$\lambda = \frac{\Theta}{\Theta + 1 - I_{\mathrm{OU}}}$$

$$\Theta = \frac{4}{\pi^2}\left(\arctan\frac{w^{\mathrm{p}}}{l^{\mathrm{p}}} - \arctan\frac{w^{\mathrm{t}}}{l^{\mathrm{t}}}\right)^2 + \left(\arctan\frac{l^{\mathrm{p}}}{h^{\mathrm{p}}} - \arctan\frac{l^{\mathrm{t}}}{h^{\mathrm{t}}}\right)^2 + \left(\arctan\frac{h^{\mathrm{p}}}{w^{\mathrm{p}}} - \arctan\frac{h^{\mathrm{t}}}{w^{\mathrm{t}}}\right)^2 \tag{4-23}$$

式中，l^{p}，w^{p}，h^{p} 分别为预测框的长、宽、高；l^{t}，w^{t}，h^{t} 对应的真实框的长、宽、高。

综上所述，PRD 网络的整体损失函数数学表达式如下：

$$\mathcal{L} = \mathcal{L}_{\mathrm{cls}} + \mathcal{L}_{\mathrm{reg}} + \mathcal{L}_{\mathrm{box}} \tag{4-24}$$

4.5 实验结果与分析

4.5.1 评价标准及实验平台

点云分割任务需处理无序、非结构化的三维数据，其稀疏性和几何复杂性使得传统图像分割的评估指标难以直接适用。尤其在物体边界模糊或类别分布不均衡的场景中，指标需同时满足对分割区域重叠度的敏感性以及对噪声的鲁棒性。而 Dice 相似性系数是一种广泛用于评估分割任务重叠度的指标，其定义为预测区域与真实标签的交集点数的两倍与两者点集总和的比值，其数学表达式如下：

$$dice = \frac{2\sum_{x\in\Omega} p_l(x)g_l(x)}{\sum_{x\in\Omega} p_l^2(x) + \sum_{x\in\Omega} g_l^2(x)} \tag{4-25}$$

式中，$p_l(x)$为像素属于背景的概率；$g_l(x)$为像素属于前景的概率。Dice 与交并比相比，Dice 对假阴性(False Negative)的惩罚更缓和，这一特性在点云分割中尤为重要——点云的非均匀采样和稀疏性可能导致预测点与真实点的微小位置偏移，而 Dice 通过归一化处理降低了此类误差对评估结果的敏感性。

另一方面，在点云目标检测领域，系统的评价指标需综合几何匹配度、分类置信度及参数量大小等多维度因素。为评估目标检测算法的性能，本研究选取的评价指标如下：

1. 交并比

衡量预测边界框与真实框的空间重合度，定义为两框交集体积与并集体积之比：

$$\mathrm{IoU} = \frac{\mathcal{V}_{\text{预测}} \cap \mathcal{V}_{\text{真实}}}{\mathcal{V}_{\text{预测}} \cup \mathcal{V}_{\text{真实}}} \tag{4-26}$$

其中，$\mathcal{V}$表示三维边界框体积。当 IoU≥τ（阈值，本次取 0.5）时，判定为真正例(TP)。

2. 混淆矩阵衍生指标

精确率(Precision)：正例预测中真实正例占比。召回率(Recall)：真实正例中被正确检测的比例 。F1-Score：调和平均平衡精确率与召回率。这些如下式

$$\begin{aligned} P &= \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}} \\ R &= \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} \\ F_1 &= 2\cdot\frac{P\cdot R}{P+R} \end{aligned} \tag{4-27}$$

3. 平均精度(Average Precision, AP)

通过积分精确率-召回率曲线下面积计算，反映多阈值下的综合性能。采用 PASCAL VOC 插值法：

$$\mathrm{AP} = \frac{1}{11}\sum_{r\in\{0,0.1,\ldots,1\}} P_{\text{interp}}(r) \tag{4-28}$$

其中，$P_{interp}(r) = max_{r'\geq r}P(r')$为插值后的精确率。

4. 均值平均精度(mean Average Precision, mAP)

多类别检测任务中，各类别 AP 的算术平均：

$$\mathrm{mAP}=\frac{1}{N_{\mathrm{cls}}}\sum_{c=1}^{N_{\mathrm{cls}}}\mathrm{AP}_c \tag{4-29}$$

N_{cls} 为类别总数。

所有实验均在统一平台上进行，如表 4-3 所示：

表 4-3 实验平台参数表

实验设备	型号
设备	Dell Precision 7875 塔式
平台	Unbuntu 20.04
CPU	Intel(R)至强® W5-3423
GPU	NVIDIA GeForce RTX A2000
System memory	256GB
SSD	Samsung 990 Pro 2TB

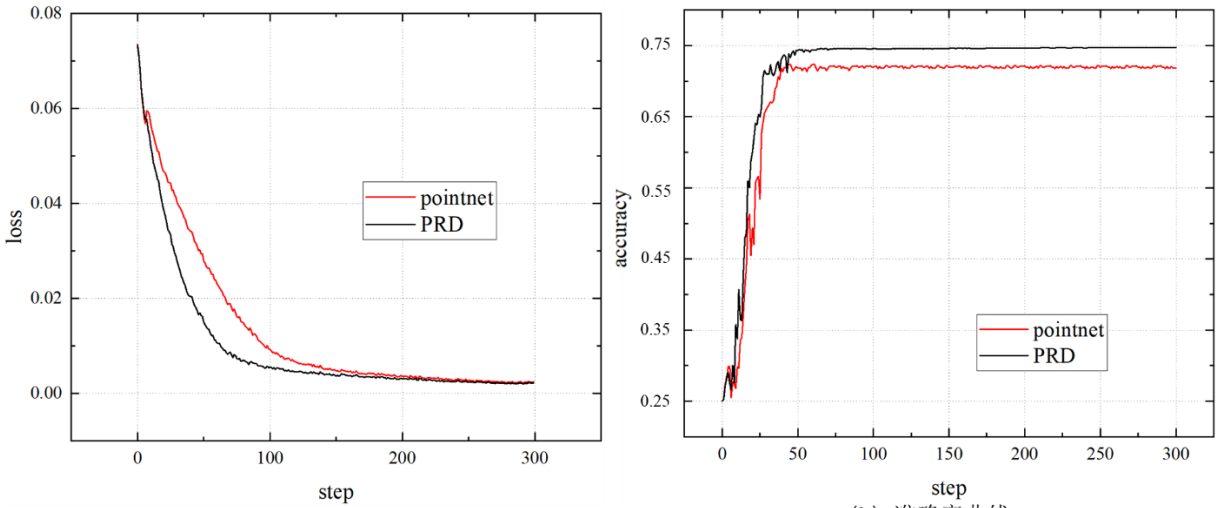
4.5.2 实验结果展示

为验证本文方法在工业机器人无序抓取任务上的性能，使用第二章的混合数据集进行三维识别实验，并与目前流行的方法进行对比。数据集共计 2416 张图片，其中，1933 张图片用作训练，483 个用作测试。在实验过程中，使用 Adam 算法进行优化，初始学率为 0.02，批量大小 batch_size 为 32，迭代次数 epoch 为 300。网络训练参数如表 4-4：

表 4-4 网络训练参数表

超参数	具体值
预热迭代次数	10
预热动量	0.7
预热偏置学习率	0.2
迭代周期	300
初始学习率	0.02
最终学习率	0.001
动量	0.85
权重衰减	0.005
批量大小	32
输入图尺寸	640×480

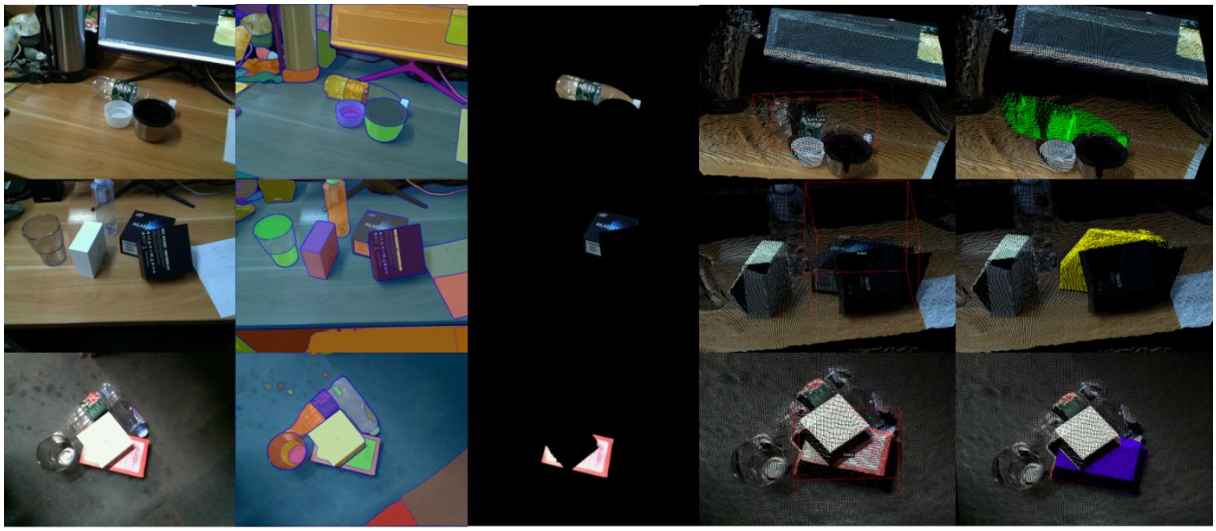
训练过程如图 4-7(a)和图 4-7(b)所示。可以看出，相比基准网络 pointnet，实验网络 PRD 在训练 loss 曲线上，收敛速度更快，在迭代周期达到 100 次左右时，趋于稳定而基准网络需要迭代周期达到 150 次左右才能趋于稳定。在准确率方面，两个网络在训练初期都出现了震荡的情况，但 PRD 网络震荡幅度较小。在后期，准确率趋于稳定后，PRD 网路的准确率达到了 75%左右，高于基准网络 pointnet。综合两点，实验网络在损失率迭代，准确率表现方面均优于基准网络，可以表明其良好性能。



(a) 损失曲线　　(b) 准确率曲线

图 4-7 训练过程图

在实验过程中，我们为了更直观地展示检测和分割的结果，对输入图像进行了多种形式的可视化处理，并在图 4-8 中展示了不同阶段的结果。从左至右依次为：(a)原始图像、(b)整体分割结果、(c)目标物分割结果、(d)3D 目标识别框、(e)3D 点云分割。



(a) 原图　(b) 整体分割图　(c) 目标物分割图　(d) 3D识别框　(e) 3D点云分割

图 4-8 实验结果可视化

其中，整体分割图(b)显示了场景中所有可见区域的语义分割结果，不同类别的物体被赋予了不同的颜色，便于观察整体的检测情况。目标物分割图(c)进一步突出了感兴趣的目标，使得对关键物体的识别更加清晰。3D 目标识别框(d)展示了在三维空间中对目标物体的检测情况，其中每个检测框对应一个具体的目标，便于后续的定位和抓取任务。而 3D 点云分割(e)则对三维点云数据进行了精细化处理，通过颜色区分不同目标，实现了更精确的三维感知。

从可视化结果可以看出，我们的方法能够较为准确地完成从二维图像到三维点云的目标检测与分割，分割边界清晰，目标识别精度较高，符合无序抓取等任务的需求。这进一步验证了我们方法在复杂场景下的有效性和鲁棒性。

为了更量化考量网络的具体性能，现将网络与目前流行的网络，如，pointnet[70] 、pointnet++[71] 、3D-SIS[72] 和 VoteNet[73] 进行对比，结果如表 4-5。可以分析得到，PRD 模型比原始 pointnet 模型在各项数值方面均有所提高，Dice 提升了约 5.7 个百分点,准确率上升了约 2.8 个百分点，只在参数大小方面略大于原始网络，在有限代价内显著提高了检测性能是值得的。在与其他主流网络的比较中，PRD 模型表现出较为优越的性能，仅次于 3D-SIS 网络。然而，值得注意的是，3D-SIS 的参数量高达 136.7M，而 PRD 模型仅为其一半左右，这意味着在保证较高检测精度的同时，PRD 模型具备更快的运行速度和更高的计算效率，因而更适合工业机器人无序抓取等对实时性有较高要求的应用场景。

表 4-5 当下主流的三维识别网络实验结果对比

项目	pointnet	pointnet++	3D-SIS	VoteNet	ours
Dice	0.826	0.863	0.887	0.864	0.883
pa	0.723	0.761	0.753	0.731	0.751
mpa	0.718	0.734	0.751	0.717	0.748
miou	0.704	0.774	0.786	0.761	0.791
params	69.1MB	83.5MB	136.7MB	73.6MB	75.3MB

4.6 本章小结

本章探讨了基于多模态特征的三维分割与识别算法，构建了 RGB-D 数据对齐的点云生成流程，并设计了基于 PointNet 与 E-ResNet 模块的三维分割网络。通过多模态特

征的融合和细化，本研究实现了对目标对象的高精度三维识别。实验验证表明，该方法在分割精度和效率方面均具有显著优势，为后续机器人在三维空间中的识别和操作能力提供了有力支持。

第5章 机器人智能分拣实验

为确保机械臂在作业过程中有效规避障碍并准确抓取目标物体，需获取目标物体与障碍物在机械臂坐标系中的精确位姿信息。实现该目标需满足以下条件：机械臂能够精准定位并按规划路径运动，这要求能求解出机械臂各连杆与关节间的运动学关系，完成相机内外参数标定并确定已知物体、相机及机械臂间的空间位姿参数。故本章首先构建实验机械臂的运动学模型及连杆坐标系，继而开展正向与逆向运动学求解，随后阐述相机内参标定、外参标定以及手眼标定方法，从而为后续研究建立完整的运动学模型与视觉感知基础。

5.1 相机内参标定

相机标定本质为相机参数估计过程，旨在建立三维空间点与二维图像投影间的几何映射关系。标定精度直接影响系统对目标物体的位姿解算精度及抓取操作准确性。标定过程中会涉及到数个坐标系，它们之间的变换关系如图 5-1：

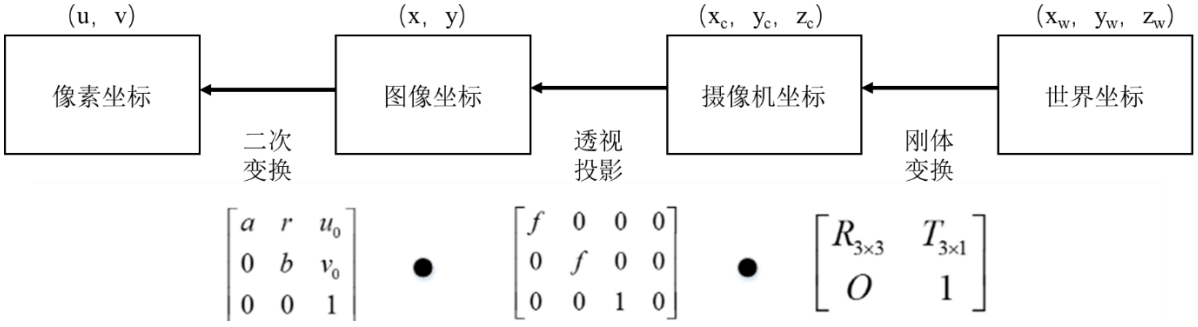


图 5-1 各个坐标系之间的变换关系

涉及到四个坐标系：世界坐标系($Ow - x_w y_w z_w$)，作为基准三维直角坐标系，可定义；相机坐标系($Oc - x_c y_c z_c$)，以光学中心为原点，坐标系的对应方向为 x 轴与 y 轴，而 z 轴为光学方向的延伸；图像坐标系($O' - x'y'$)，以成像平面中心为基准，坐标轴与传感器物理边界对齐；像素坐标系($O - xy$)，则以图像左上角为原点，通过离散化量化图像坐标系形成。各坐标系间的更直观的关联如图 5-2 所示：

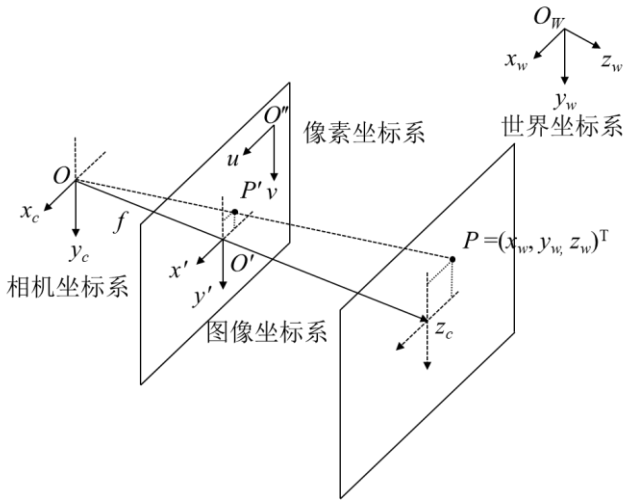


图 5-2 各个坐标系之间的直观展示图

三维物体至二维图像的几何投影本质上是上述坐标系的链式变换过程。实际应用中，受镜头光学特性影响，图像坐标系转换会产生非线性几何畸变，其具体畸变类型及表现形式如图 5-3 所示：

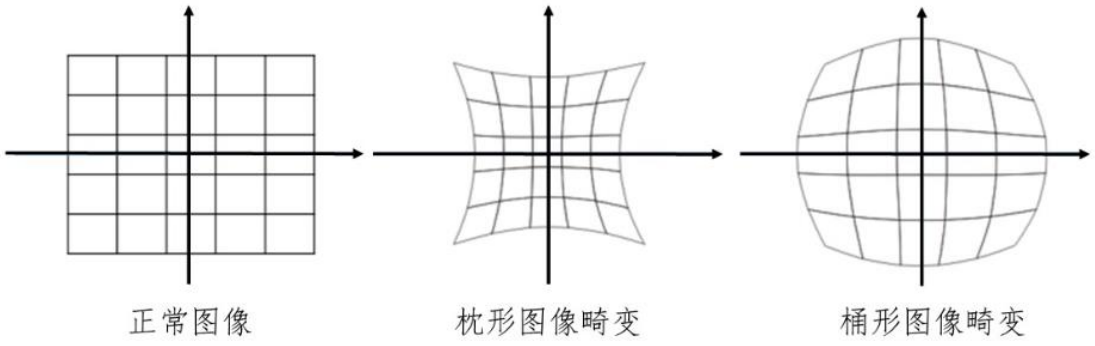


图 5-3 不同畸变类型的表现形式图

为减小图像畸变的影响，需对其进行矫正。整个过程的数学表述可以用下式展示：

$$\begin{aligned}\begin{bmatrix}X_w\\Y_w\\Z_w\\1\end{bmatrix}&=\begin{bmatrix}R&0\\T&1\end{bmatrix}\begin{bmatrix}\frac{Z_c}{f}&0&0\\0&\frac{Z_c}{f}&0\\0&0&Z_c\\0&0&1\end{bmatrix}\begin{bmatrix}dx&0&-u_0dx\\0&dy&-v_0dy\\0&0&1\end{bmatrix}\begin{bmatrix}u_1\\v\\1\end{bmatrix}\\&=\begin{bmatrix}R&0\\T&1\end{bmatrix}\begin{bmatrix}\frac{Z_cdx}{f}&0&\frac{-Z_cu_0dx}{f}\\0&\frac{Z_cdy}{f}&\frac{-Z_cv_0dy}{f}\\0&0&Z_c\\0&0&1\end{bmatrix}\begin{bmatrix}u\\v\\1\end{bmatrix}\end{aligned}\tag{5-1}$$

式中，$\begin{bmatrix} R & 0 \\ T & 1 \end{bmatrix}$为深度相机的外参；$\begin{bmatrix} \frac{Z_c dx}{f} & 0 & \frac{-Z_c u_0 dx}{f} \\ 0 & \frac{Z_c dy}{f} & \frac{-Z_c v_0 dy}{f} \\ 0 & 0 & Z_c \\ 0 & 0 & 1 \end{bmatrix}$为深度相机的内参，包括 x、y 轴的焦距、主点、畸变系数等；Zc为到相机中心的距离。

本次采用 realsense 相机内参标定，流程如下。1.等待标定启动。2.启动成功，使标定板在摄像头的蓝色框范围内移动。3.改变摄像头的视角，使图像上的标定板扫过蓝色框，直至所有的蓝色消失。4.通过不停移动图片让软件获取 15 张在不同位姿下的图片（深度）。5.再不停移动图片获取 15 张 RGB 图 6.标定完成。具体操作如图 5-4

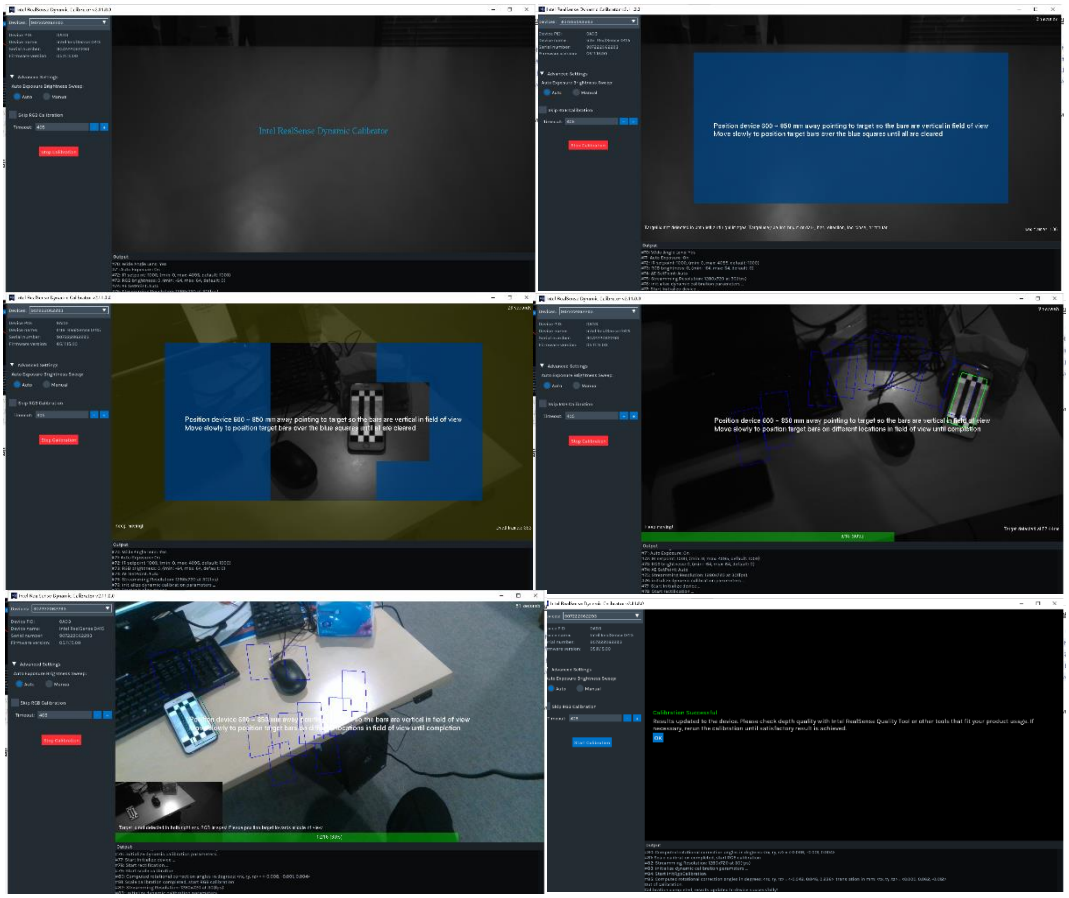


图 5-4 RealsenseL515 相机内参标定过程图

在 camera_calibration_4Hz-report-cam.pdf 中显示三个相机重投影误差 reprojection error，图 5-5 是本次标定的结果。图中 X 轴(Image Index)这一轴表示标定过程中采集的图像编号，也就是说，每个数据点对应一张标定图像。Y 轴（重投影误差）代表计算得到的重投影误差值，单位通常为像素，其中 Error X 和 Error Y 分别代表水平和垂直方

向的误差。这个值表示在该图像中，通过相机模型计算得到的投影点与实际检测到的角点之间的欧氏距离。按下式计算：

$$e=\sqrt{(u-\hat{u})^2+(v-\hat{v})^2} \tag{5-2}$$

其中，(u,v)为实际检测到的图像坐标；$(\hat{u},\hat{v})$为根据内外参数计算得到的投影坐标。误差越小，说明相机内参和外参的拟合程度越好，标定精度越高。通过观察各图像的误差分布，可以看出图像的误差保持在较低范围，也就是 1pix 范围内。

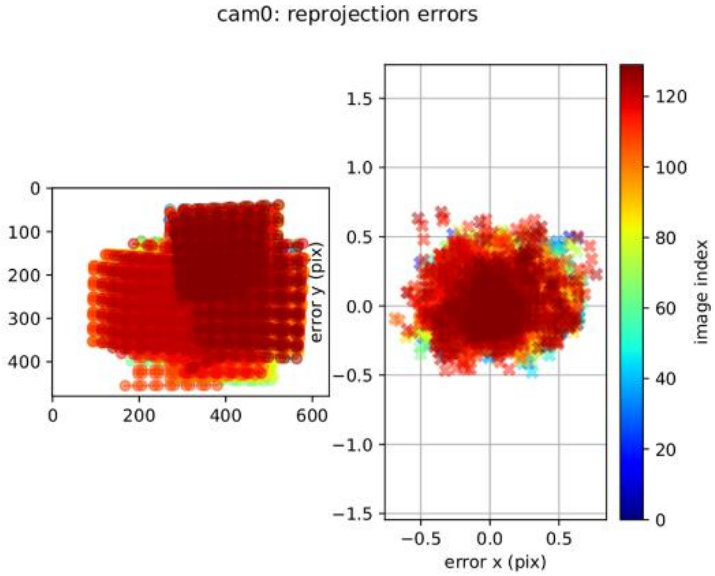


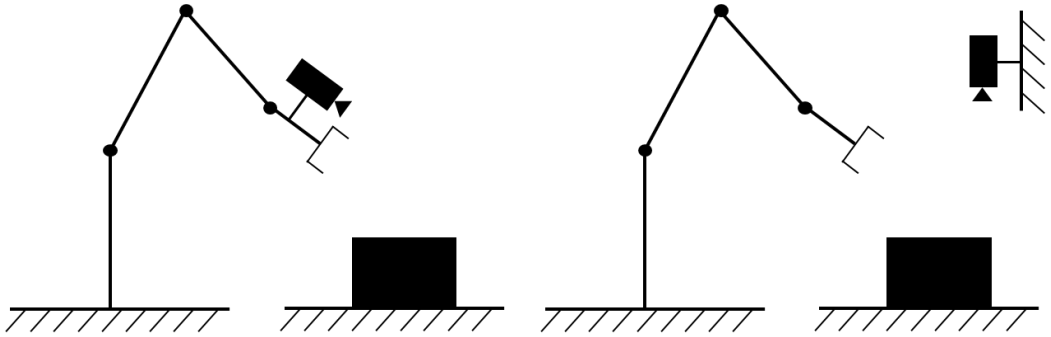
图 5-5 标定误差结果图

5.2 三维坐标定位

为实现目标物体与机器人系统间的精准空间交互控制，需通过手眼标定建立机器人末端执行器与视觉传感器间的精确位姿转换关系。本节重点阐述视觉引导机器人系统的标定方法与实现过程，其核心在于求解相机坐标系与机器人末端工具坐标系间的刚性变换矩阵，该参数矩阵的准确性直接决定视觉伺服控制系统的空间配准精度。

5.2.1 系统手眼标定

手眼标定根据传感器安装方式可分为两种典型配置模式：当相机固定于机器人末端执行器时(Eye-in-Hand)，需通过复合运动解算相机与工具坐标系的相对位姿；当相机独立安装于工作空间时(Eye-to-Hand)，则需构建相机坐标系与机器人基坐标系的空间映射关系。两种配置模式的如图 5-6 所示



（a）眼在手上　　（b）眼在手外

图 5-6 两种配置类型

考虑到生产实际的需要，本次无序抓取的应用场景，采用的是当相机独立安装于工作空间时的标定方式，即眼在手外的类型，进行手眼标定。

5.2.2 转换坐标关系

为实现多坐标系间的精确空间配准，本研究涉及四个核心坐标系：基坐标系(B, base)、末端执行器坐标系(G, gripper)、相机坐标系(C, cam)和标定板坐标系(T, target)。如图 5-7 所示：

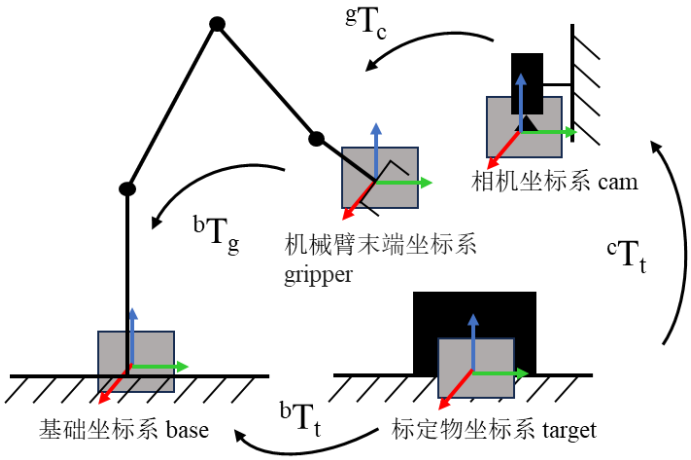


图 5-7 坐标转换关系图

基于齐次坐标变换的传递性原理，在相机移动至不同位置对标定板进行拍照的过程中，各坐标系间的位姿变换矩阵满足以下链式乘法法则：

$$ {}^{b}T_{t} = {}^{b}T_{h}\,{}^{g}T_{c}\,{}^{c}T_{t} \tag{5-3} $$

本研究提出基于空间位姿守恒原理的相机-执行器联合标定方案，其技术实施流程如下：

第一步，将高精度棋盘格标定板固定于机器人工作空间内，确保其与基坐标系的

空间变换关系恒定。标定板如图 5-8。

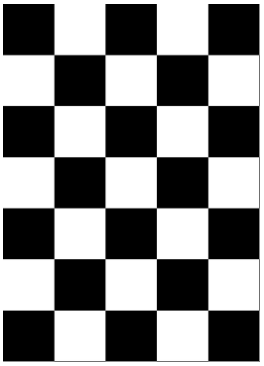


图 5-8 标定板图

第二步，操控机械臂携带视觉传感器完成 N 次(N≥10)多视角观测：每次移动后确保标定板完整成像于视场。记录各采样点执行器终端位姿，同步存储包含标定板的数字图像集。在标定过程中，标定板与基础坐标系的位姿关系不变，即 T 保持不变。同理，在标定过程中相机与机械臂末端的位姿关系 T 保持不变。故可对式(5-3)进行变换：

$$\begin{aligned} {}^{b}T_{(1)t} &= {}^{c}T_{(1)t}\,{}^{g}T_{c}\,{}^{b}T_{(1)g} \\ {}^{b}T_{(2)t} &= {}^{c}T_{(2)t}\,{}^{g}T_{c}\,{}^{b}T_{(2)g} \end{aligned} \tag{5-4}$$

由于 ${}^{b}T_{(1)t} = {}^{b}T_{(2)t}$，故可得到下式：

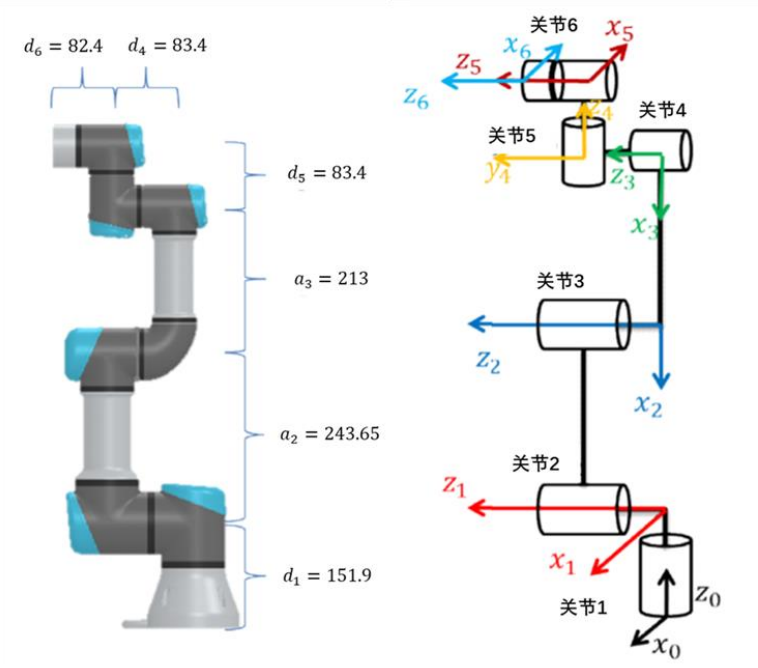
$$\begin{aligned} {}^{c}T_{(1)t}\,{}^{g}T_{c}\,{}^{b}T_{(1)g} &= {}^{c}T_{(2)t}\,{}^{g}T_{c}\,{}^{b}T_{(2)g} \\ {}^{c}T_{(2)t}^{-1}\,{}^{c}T_{(1)t}\,{}^{g}T_{c} &= {}^{g}T_{c}\,{}^{b}T_{(2)g}\,{}^{b}T_{(1)g}^{-1} \end{aligned} \tag{5-5}$$

即求解：$AX=XB$，A：目标坐标系到相机坐标系的转换。B：机械臂末端坐标系到基础坐标系的转换。方程建立后，采集多组数据，共同求解出 T，再将 T 带回式(5-3)即可获得每组数据对应的 T。至此，UR3 机械臂的眼在手外的标定的工作完成。

5.3 机械臂运动学分析

5.3.1 机械臂运动学建模

本研究选用 Universal Robots 公司研发的 UR3 六自由度协作机械臂作为实验平台。该型机械臂凭借其紧凑型结构设计（总臂长 500mm，工作半径 500mm）与高重复定位精度(±0.1mm)，特别适用于精密装配场景与自动化教学实验。其模块化关节结构可实现末端执行器在三维空间内的全向运动，各旋转关节均具备 360° 连续转动能力，最大负载参数为 3kg。UR3 机械臂尺寸数据和坐标系建立如图 5-9：



(a)机械臂尺寸　　(b)机械臂坐标系

图 5-9 UR3 机械臂图

基于 Denavit-Hartenberg(D-H)参数法的运动学建模，该建模方法通过定义连杆坐标系间四个基本几何参数（连杆转角 α、连杆长度 a、连杆偏距 d、关节角 θ），系统化构建机械臂运动学链模型。表 5-1 展示 UR3 机械臂的 D-H 参数，其中基础参数值源自厂商技术文档，关节变量 θ_1-θ_6表征各旋转关节的实时运动状态。

表 5-1 UR3 的 D-H 参数

关节	θ_i	$d_i(m)$	$a_i(m)$	$\alpha_i(rad)$
1	θ_1	0.15190	0	$\pi/2$
2	θ_2	0	-0.24365	0
3	θ_3	0	-0.21300	0
4	θ_4	0.08340	0	$\pi/2$
5	θ_5	0.08340	0	$-\pi/2$
6	θ_6	0.08240	0	0

通过上述 D-H 参数建模，可完整描述机械臂各连杆坐标系间的齐次变换关系，为后续正/逆运动学分析奠定模型基础。

5.3.2 基于直线插补点的机械臂运动规划

机械臂轨迹规划方法依据控制域可分为两类：关节空间规划与笛卡尔空间规划。前者通过构造关节角位移的时间函数，在运动学约束下直接生成各关节运动轨迹；后

者则通过构建末端执行器在笛卡尔空间中的连续位姿序列，经逆运动学求解间接驱动关节运动。本研究采用笛卡尔空间规划方法，其优势在于能直接约束末端执行器的空间运动轨迹，保障作业过程的可控性。

在基坐标系统中，给定运动轨迹起点和期望点的笛卡尔坐标分别为 $A_{Start}(X_S,Y_S,Z_S)$ 和 $A_{Finish}(X_F,Y_F,Z_F)$，通过在三个坐标轴之间插入点，引导机械臂执行点对点轨迹运动，示意图如图 5-10 所示：

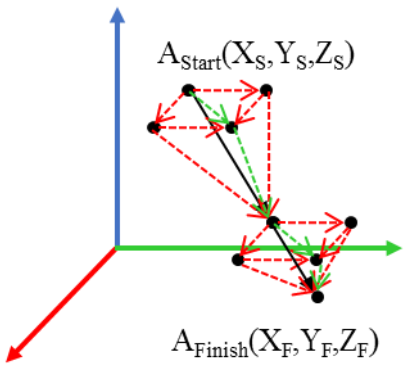


图 5-10 笛卡尔坐标系图

首先，计算运动轨迹中起点和终点之间的距离。计算距离的公式如下：

$$d=\sqrt{(X_s-X_f)^2+(Y_s-Y_f)^2+(Z_s-Z_f)^2} \tag{5-6}$$

插补点的增量和实时坐标表示为：

$$\begin{cases} X=X_s+\sigma(X_f-X_s) \\ Y=Y_s+\sigma(Y_f-Y_s) \\ Z=Z_s+\sigma(Z_f-Z_s) \end{cases} \tag{5-7}$$

式中，σ 是固定的时间单位，即标准化的时间操作符，按下式得到：

$$s(t)=\begin{cases} \dfrac{1}{2}at^2, & (0\le t\le T_b) \\ \dfrac{1}{2}aT_b^2+aT_b(t-T_b), & (T_b<t\le 1-T_b) \\ \dfrac{1}{2}aT_b^2+aT_b(t-T_b)-\dfrac{1}{2}a(t+T_b-1)^2, & (1-T_b<t\le 1) \end{cases} \tag{5-8}$$

式中，a 是抛物线段的加速度；T_b 是抛物线段的时间，代表标准化时间操作符的累积的总时间；σ=0 代表起点，σ=1 代表期望点。由于该算法控制机械臂在 X, Y 和 Z 轴上直线移动，因此三个坐标轴之间的运动轨迹满足线性关系：

$$\begin{cases} Z = \delta_1 X \\ Z = \delta_2 Y \\ Y = \delta_3 X \end{cases} \quad , \quad \begin{cases} \delta_1 = \dfrac{Z_f - Z_s}{X_f - X_s} \\ \delta_2 = \dfrac{Z_f - Z_s}{Y_f - Y_s} \\ \delta_3 = \dfrac{Y_f - Y_s}{X_f - X_s} \end{cases} \tag{5-9}$$

通过式可知，对机械臂进行逆运动学求解，通过指定起点、目标点和插值点位置进行计算，即可将坐标点转换成关节角度变量，并通过控制机械臂的关节角度变量，进而规划机械臂的运动路径，控制其完成抓取任务。

5.3.3 基于解析法的机械臂逆运动学计算

针对多关节系统的运动学反解问题，现有技术路线主要包括几何法、闭式解析法及迭代逼近法三类。本研究通过多维技术评估确定最优解算方案：1. 几何法：基于空间几何关系建立解算模型，虽具有物理意义明确、实现简单的特点，但在处理高自由度系统，如六轴串联机械臂，时存在多解筛选困难及建模复杂度指数增长的问题；2. 闭式解析法：通过代数方程推导获得精确解析解，计算效率显著占优，时间复杂度O(1)，但受限于机械臂构型需满足 Pieper 准则等特定约束条件；3. 数值迭代法：采用梯度下降等优化算法逐步逼近解，虽具备强构型适应性，但存在时间成本过高，平均迭代次数>50 次、雅可比矩阵病态等固有限制。考虑到本研究需要进行实时的解算，因此采用闭式解析法。

定义相邻连杆坐标系$i-1$至i的齐次坐标变换关系：

$$A_i = Rot_z(\theta_i)Trans_z(d_i)Rot_x(\alpha_i)Trans_x(a_i) \tag{5-10}$$

将旋转变换矩阵与平移算子展开可得完整变换矩阵：

$$A_i(\theta_i) = \begin{bmatrix} c_{\theta_i} & -s_{\theta_i}c_{\alpha_i} & s_{\theta_i}s_{\alpha_i} & a_i c_{\theta_i} \\ s_{\theta_i} & c_{\theta_i}c_{\alpha_i} & -c_{\theta_i}s_{\alpha_i} & a_i s_{\theta_i} \\ 0 & s_{\alpha_i} & c_{\alpha_i} & d_i \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{5-11}$$

其中，s 代表 sin；c代表cos，下同；α_i 为扭角参数。基于 D-H 参数表进行级联运算，得到相邻连杆的齐次变换。矩阵表达式依次为：

$$A_1(\theta_1)=\begin{bmatrix} c_1 & 0 & s_1 & 0 \\ s_1 & 0 & -c_1 & 0 \\ 0 & 1 & 0 & d_1 \\ 0 & 0 & 0 & 1 \end{bmatrix} A_2(\theta_2)=\begin{bmatrix} c_2 & -s_2 & 0 & a_2c_2 \\ s_2 & c_2 & 0 & a_2s_2 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} A_3(\theta_3)=\begin{bmatrix} c_3 & s_3 & 0 & a_3c_3 \\ s_3 & -c_3 & 0 & a_3s_3 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

$$A_4(\theta_4)=\begin{bmatrix} c_4 & 0 & -s_4 & 0 \\ s_4 & 0 & c_4 & 0 \\ 0 & -1 & 0 & d_4 \\ 0 & 0 & 0 & 1 \end{bmatrix} A_5(\theta_5)=\begin{bmatrix} c_5 & 0 & s_5 & 0 \\ s_5 & 0 & -c_5 & 0 \\ 0 & 1 & 0 & d_5 \\ 0 & 0 & 0 & 1 \end{bmatrix} A_6(\theta_6)=\begin{bmatrix} c_6 & -s_6 & 0 & 0 \\ s_6 & c_6 & 0 & 0 \\ 0 & 0 & 1 & d_6 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

(5-12)

将上述各个矩阵相乘，便可以得到机械臂的正运动学方程，即：

$$T=A_1A_2A_3A_4A_5A_6=\begin{bmatrix} n_x & o_x & a_x & p_x \\ n_y & o_y & a_y & p_y \\ n_z & o_z & a_z & p_z \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{5-13}$$

式中的简写为下式，展开如下：

$$\begin{aligned}
n_x &= s_1s_5c_6+c_1\left[\cos(\theta_4-\theta_3-\theta_2)c_5c_6-\sin(\theta_4-\theta_3-\theta_2)c_6\right] \\
n_y &= s_1\cos(\theta_4-\theta_3-\theta_2)c_5c_6-c_1c_6s_5-s_1\sin(\theta_4-\theta_3-\theta_2)s_6 \\
n_z &= -c_5c_6\sin(\theta_4-\theta_3-\theta_2)-\cos(\theta_4-\theta_3-\theta_2)s_6 \\
o_x &= s_1s_5s_6-c_1\left[c_6\sin(\theta_4-\theta_3-\theta_2)+\cos(\theta_4-\theta_3-\theta_2)c_5s_6\right] \\
o_y &= -s_1\sin(\theta_4-\theta_3-\theta_2)c_5+\left[-s_1\cos(\theta_4-\theta_3-\theta_2)c_5+c_1s_5s_5\right]s_6 \\
o_z &= -\cos(\theta_4-\theta_3-\theta_2)c_6+c_5\sin(\theta_4-\theta_3-\theta_2)s_6 \\
a_x &= -c_5s_1-c_1\cos(\theta_4-\theta_3-\theta_2)s_5 \\
a_y &= c_1c_5-\cos(\theta_4-\theta_3-\theta_2)s_1s_5 \\
a_z &= -\sin(\theta_4-\theta_3-\theta_2)s_5 \\
p_x &= -(d_4+d_6c_6)s_1+c_1\left[a_2c_2+a_3c_{23}-d_5\sin(\theta_4-\theta_3-\theta_2)+d_6\cos(\theta_4-\theta_3-\theta_2)s_5\right] \\
p_y &= (d_4+d_6c_6)c_1+s_1\left[a_2c_2+a_3c_{23}-d_5\sin(\theta_4-\theta_3-\theta_2)+d_6\cos(\theta_4-\theta_3-\theta_2)s_5\right] \\
p_z &= d_1-d_5\cos(\theta_4-\theta_3-\theta_2)+a_2s_2+a_3c_{23}
\end{aligned} \tag{5-14}$$

首先对矩阵A_1进行反变换，然后左乘：

$$A_1^{-1}T = A_2A_3A_4A_5A_6$$

$$\begin{bmatrix} c_1 & s_1 & 0 & 0 \\ 0 & 0 & 1 & -d_1 \\ s_1 & -c_1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \times \begin{bmatrix} n_x & o_x & a_x & p_x \\ n_y & o_y & a_y & p_y \\ n_z & o_z & a_z & p_z \\ 0 & 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} n_xc_1+n_ys_1 & o_xc_1+o_ys_1 & a_xc_1+a_ys_1 & p_xc_1+p_ys_1 \\ n_z & o_z & a_z & -d_1+p_z \\ -n_yc_1+n_xs_1 & -o_yc_1+o_xs_1 & -a_yc_1+a_xs_1 & -p_yc_1+p_xs_1 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

$$= \begin{bmatrix} [1,1] & [1,2] & c_6s_5 & 0 \\ [2,1] & [2,2] & -s_5s_6 & 0 \\ 0 & [3,2] & -\sin(\theta_4-\theta_3-\theta_2)s_5 & -c_5 \\ [4,1] & [4,2] & -(d_4+d_6c_5) & 1 \end{bmatrix}$$

(5-15)

式中矩阵对应位置的值展开如下：

$$[1,1] = \cos(\theta_4-\theta_3-\theta_2)c_5c_6 - \sin(\theta_4-\theta_3-\theta_2)s_6$$
$$[1,2] = -c_5c_6\sin(\theta_4-\theta_3-\theta_2) - \cos(\theta_4-\theta_3-\theta_2)s_6$$
$$[2,1] = -\sin(\theta_4-\theta_3-\theta_2)c_6 - \cos(\theta_4-\theta_3-\theta_2)c_5s_6$$
$$[2,2] = -\cos(\theta_4-\theta_3-\theta_2)c_6 + c_5\sin(\theta_4-\theta_3-\theta_2)s_6$$
$$[3,2] = -\cos(\theta_4-\theta_3-\theta_2)s_5$$
$$[4,1] = a_2c_2 + a_3c_{23} - d_5\sin(\theta_4-\theta_3-\theta_2) + d_6\cos(\theta_4-\theta_3-\theta_2)s_5$$
$$[4,2] = a_2c_2 + a_3c_{23} - d_5\sin(\theta_4-\theta_3-\theta_2) + d_6\cos(\theta_4-\theta_3-\theta_2)s_5$$

按照式(5-15)展开，由于等号左右的矩阵对应位置的元素相等，故矩阵的第三行第三列和第三行第四列等式展开如下：

$$\begin{aligned} -c_5 &= -a_yc_1 + a_xs_1 \\ -\left(d_4 + d_6c_5\right) &= -p_yc_1 + p_xs_1 \end{aligned} \tag{5-16}$$

求解得θ_1、θ_5。同理，由矩阵的第三行第一列和第三行第二列元素可知：

$$\frac{-s_5s_6}{c_6s_5} = \frac{-o_yc_1+o_xs_1}{-n_yc_1+n_xs_1} \tag{5-17}$$

解得θ_6。同理，由矩阵的第二行第三列元素可知：

$$-\sin(\theta_4-\theta_3-\theta_2)s_5 = a_z \tag{5-18}$$

同时，对矩阵A_2、A_3和A_4依次进行反变换：

$$A_4^{-1}A_3^{-1}A_2^{-1}A_1^{-1}T = A_5A_6$$

$$= \begin{bmatrix} [1,1] & [1,2] & [1,3] & [1,4] \\ -n_yc_1 + n_xs_1 & -o_yc_1 + o_xs_1 & -a_yc_1 + a_xs_1 & -p_yc_1 + p_xs_1 + d_4 \\ [3,1] & [3,2] & [3,3] & [3,4] \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{5-19}$$

$$= \begin{bmatrix} c_5c_6 & -c_5s_6 & s_5 & s_5d_6 \\ c_6s_5 & -s_6s_5 & -c_5 & -c_5d_6 \\ s_6 & c_6 & 0 & d_5 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

式中矩阵对应位置的值展开如下：

$$[1,1] = (n_xc_1 + n_ys_1)\cos(\theta_4 - \theta_3 - \theta_2) - n_z\sin(\theta_4 - \theta_3 - \theta_2)$$
$$[3,1] - (n_xc_1 + n_ys_1)\sin(\theta_4 - \theta_3 - \theta_2) - n_z\cos(\theta_4 - \theta_3 - \theta_2)$$
$$[1,2] = (o_xc_1 + o_ys_1)\cos(\theta_4 - \theta_3 - \theta_2) - o_z\sin(\theta_4 - \theta_3 - \theta_2)$$
$$[3,2] = -(o_xc_1 + n_ys_1)\sin(\theta_4 - \theta_3 - \theta_2) - o_z\cos(\theta_4 - \theta_3 - \theta_2)$$
$$[1,3] = (a_xc_1 + a_ys_1)\cos(\theta_4 - \theta_3 - \theta_2) - a_z\sin(\theta_4 - \theta_3 - \theta_2)$$
$$[3,3] = -(a_xc_1 + a_ys_1)\sin(\theta_4 - \theta_3 - \theta_2) - a_z\cos(\theta_4 - \theta_3 - \theta_2)$$
$$[1,4] = (p_xc_1 + p_ys_1)\cos(q_4 - q_3 - q_2) - (p_z - d_1)\sin(q_4 - q_3 - q_2) - a_2c_3c_4 - a_2s_3s_4 - a_3c_4$$
$$[3,4] = -(p_xc_1 + p_ys_1)\sin(q_4 - q_3 - q_2) - (p_z - d_1)\cos(q_4 - q_3 - q_2) + a_2c_3s_4 + a_3s_4 - a_2s_3c_4$$

由矩阵的第一行第四列和第二行第四列元素可知：

$$\begin{gathered} s_5d_6 = (p_xc_1 + p_ys_1)\cos(q_4 - q_3 - q_2) \\ -(p_z - d_1)\sin(q_4 - q_3 - q_2) - a_2c_3c_4 - a_2s_3s_4 - a_3c_4 \\ d_5 = -(p_xc_1 + p_ys_1)\sin(q_4 - q_3 - q_2) \\ -(p_z - d_1)\cos(q_4 - q_3 - q_2) + a_2c_3s_4 + a_3s_4 - a_2s_3c_4 \end{gathered} \tag{5-20}$$

将式(5-18)和式(5-20)联立，解得θ_2、θ_3 和θ_4。至此，本次 UR3 机械臂的运动学分析求解完成。

5.4 RGBD Fusion 物体姿态识别网络

针对现有单模态位姿估计方法在跨模态特征融合方面的局限性，本研究提出基于 RGB-D 6D 位姿估计模型。该架构通过独立设计的 RGB 特征提取分支与深度特征编码分支，分别捕获纹理语义信息与空间几何特征，充分利用 RGB 和深度图这两个互补的数据源。并使用密集融合网络来提取各像素的特征，进而估计姿态。

现有的方法很难同时满足严重遮挡下准确的姿态估计和实时识别的要求。当时的

PoseCNN 方法和 MCN 方法，这些方法需要详细的事后改进步骤来充分利用 3D 信息，例如前者中的高度定制的迭代最近点过程和 MCN 中的多视图假设验证方案。这些细化步骤不能与最终目标一起优化，不能满足实时性要求；而可以满足实时性要求的方法如 PointFusion 和 Frustrum PointNet 方法，也就是直接从点云数据中估计 6D 位姿，这些方法在遮挡环境下性能是不足的.

5.4.1 网络整体架构

整体架构分为两个阶段。第一阶段，语义解析网络，采用改进型 PoseCNN 架构对 RGB 图像实施实例分割，生成目标物体的精确空间掩膜，再通过掩膜确定其在深度图中的位置，并把这部分裁剪下来。此时获得的 image crop 和 masked point cloud 都当做网络下一阶段的输入。下一阶段就是根据输入的掩膜进行姿态估算，其分为四个部分。1.把色彩掩膜的逐像素信息放到全卷积网络中，对应到 color embeddings(H * W * d_rgb)。2.使用 PointNet 的变体处理经过 mask 获得目标点云的数据，映射成一个带有几何信息的几何编码。3.由于遮挡和分割错误，来自前一步骤的特征集可能包含其他对象或背景部分上的像素的特征，所以不能进行简单地堆叠。这里提出了一个逐像素密集融合方法，把 color embeddings 和几何编码进行像素级别的融合，然后经过基于无监督的置信度评分网络进行姿态估算。4.对初步估算的姿态结果进行提炼优化。整体网络结构图如图 5-11：

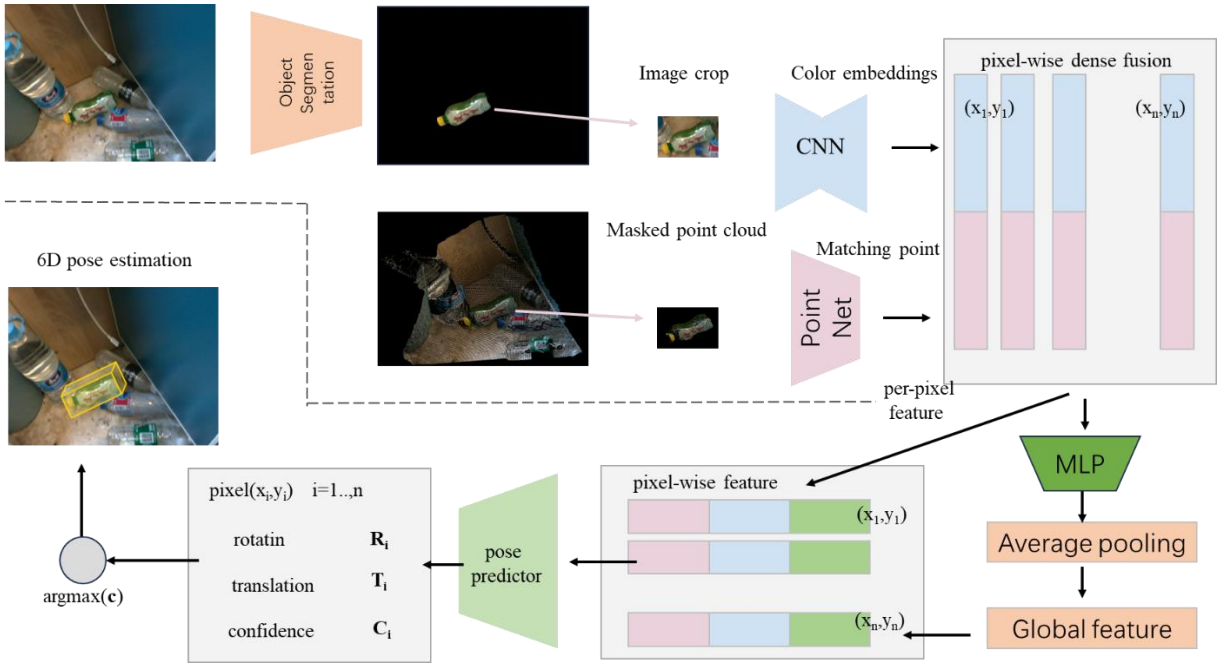


图 5-11 RGBD Fusion 网络整体架构图

网络中的语义分割模块使用了编码器-解码器结构，得到通道数为“类别数+1”的输出。因将背景也视为一种。RGBD Fusion网络基于像素级语义分割的视觉处理方法中，编码方案采用 one-hot 形式生成多通道分割输出。，是二元的 mask。而将 mask 对应的像素点映射到点云上的点有两种方法：

(1) 若是有序点云，则点云点的 index 可通过 RGB 像素点在图像按行 flatten 的单行像素点中的index来得到：点云点index = 图片cols * 像素点row_index + 像素点col_index 然后根据点云点的 index，可以索引到点云点的 x,y,z 值。

(2) 若是无序点云，则点云点的 x,y,z 值，可以通过相机内参来获得：点云点的坐标值为相机坐标系中的坐标值，而 RGB 像素点的坐标值为图像坐标系中的坐标值。两者之间的转换可由相机的内参矩阵来完成。

本文所述的RGBD Fusion网络架构通过以下技术路径实现特征整合：1.稠密特征融合机制：针对每个像素点构建色彩-深度联合表征，通过将 RGB 图像特征向量与对应点云深度特征沿特征维度进行融合拼接，形成像素级融合特征；2.分层特征处理流程：初级特征整合阶段将融合特征输入多层感知机(MLP)进行非线性变换采用平均池化算子提取全局上下文特征，该对称函数的置换不变性克服了点云数据的无序性挑战。通过通道维度拼接将全局特征广播至各局部融合特征，构建具有空间感知能力的增强特征表示；3.自监督姿态预测系统：每个增强特征向量对应生成一个候选姿态参数，网络同步输出各姿态假设的置信度评估值，构建端到端的自监督选择机制。该架构通过像素级稠密融合保持空间细节，结合全局上下文引导的特征增强策略，最终输出经置信度加权的一组姿态估计结果。RGBD Fusion使用一种自监督的方式选择其中最优的预测姿态，即对于每个预测姿态都对应输出一个置信度分数作为判断依据。网络模型对于姿态的预测和置信度分数的计算的学习，是通过对损失函数的设计来在训练优化中实现的。

使用的损失函数是最小化 3D 模型上的采样点在真实姿态下的坐标与在预测姿态下的坐标之间的欧氏距离的均值。数学表达式如下：

$$L_i^p = \frac{1}{M}\sum_j \left\| (Rx_j + t) - (\hat{R}_i x_j + \hat{t}_i) \right\| \tag{5-21}$$

其中，M 为在 3D 模型上的随机采样点的点数。对于由不同的融合特征预测出的不同结果，都有一个如上的损失函数。

但上述的损失函数，使用的对象为形状非对称的物体。若对象为形状对称的物体，

则存在有多个不同的 pose 可作为解，满足该最优化问题。对于比如球型的物体，甚至存在有无限个 pose 可作为该最优化问题的解。因此，对于形状对称的物体，上述的损失函数（或者说用于学习的目标函数），将变得模棱两可。这不利于网络的训练。于是，在RGBDFusion网络中，对于形状对称的物体，使用了另一个损失函数，数学表达式如下：

$$L_i^p = \frac{1}{M}\sum_j \min_{0<k<M} \left\| (Rx_j + t) - (\hat{R}_i x_k + \hat{t}_i) \right\| \tag{5-22}$$

式子就是对于每个在预测姿态下的3D模型采样点，都去寻找它在真实姿态下的3D模型采样点中的最近点，并计算这对采样点之间的距离，构建基于最小化逐点距离均值的约束条件，通过 L1 范数度量预测姿态与真实位姿间的点云配准误差，该优化目标通过迭代式地缩小预测与真值间的几何偏差，最终收敛于满足严格点云配准要求的唯一解。此时，这一约束体现在强制模型满足精确的逐点对应关系，消除因物体几何对称性导致的次优解

该损失函数在灵感来自于迭代最近点的思想，但算法步骤相反。具体地，在逐次的优化迭代过程中，不断调整，使得各对最近点之间的距离都在逐渐减小。慢慢地从可能的错误的一对多或多对一的各对最近点对，解耦并逼近正确的一对一的各对最近点对，因为其距离之和的最小值在理论上可以取到 0，而其他情况基本上都不可能取到 0，除非数据的取值和分布的巧合性。

5.4.2 Refinement 模块

Refinement 模块也就是迭代式位姿优化算法。初始位姿估计模块输出的位姿参数 T_{init} 存在亚像素级误差，需通过迭代精化机制进行优化。如图 5-12 所示，该算法通过构建位姿残差学习框架实现渐进式位姿修正，其数学表达为：

$$T_{refined} = \prod_{k=1}^{K} \Delta T_k \cdot T_{init} \tag{5-23}$$

式中，$\Delta Tk=[\Delta Rk|\ \Delta tk]$ 表示第 k 次迭代的位姿修正量；K 为最大迭代次数。具体实施流程如下：

1. 位姿初始化：将初始位姿 T_{init} 作用于物体点云 P 生成对齐点云 $P' = T_{init} \circ P$ 。

2. 多模态特征融合：通过 PointNet 网络提取 P' 的几何特征 F_{geo} ，与原始 RGB 特征

F_{rgb} 进行稠密特征融合。

3. 将所有的融合特征利用 MLP 提取出全局特征，并把该全局特征送进 pose residual estimator 中，计算出 R 和 t 的残差。

4. 将上一步得到的 R 和 t 的残差应用于作为本次迭代输入的点云上，得到的微调的点云即为本次迭代的输出。

5. 重复迭代直到达到条件。迭代终止条件：当 R 和 t 的残差小于特定值时终止迭代（本次 Rϵ=0.01rad, tϵ=0.005m）。

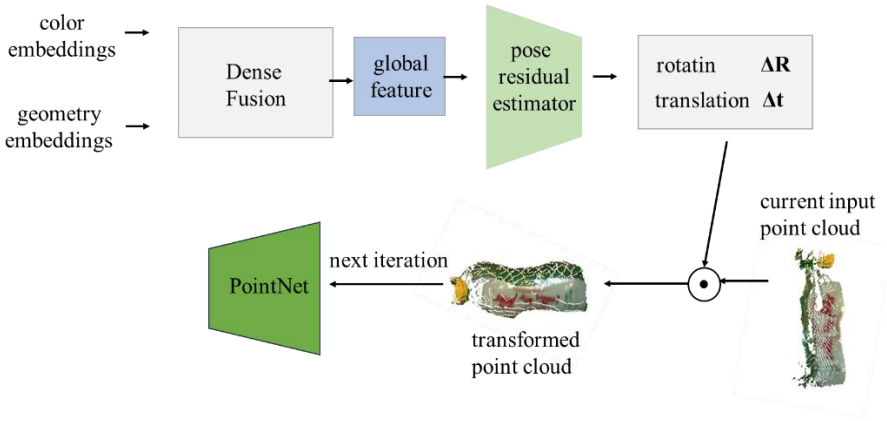


图 5-12 Refinement 模块细节

需要注意的是利用 DenseFusion 的主体模块预测出的 pose 为以上的迭代过程中物体的初始化 pose，此后利用 refine 模块预测出的 pose residual 是叠加在其上的微调量。

5.4.3 实验结果

位姿估计精度的评估采用三项核心指标：2D 重投影误差、5cm5°精度阈值以及算法实时性。其中，2D 重投影误差通过计算目标物体关键点在图像中的投影位置与实际检测位置的欧氏距离均值进行量化，公式如下：

$$e_{REP} = \left\| P_i - CH\mu \right\|_2 \tag{5-24}$$

其中，Pi是目标平面像素坐标；C 是摄像机内参矩阵，是成像几何投影关系建模核心参数；H 是刚体变换矩阵，三维空间位姿的齐次坐标表征；μ 是像素最大似然混合分布均值。若误差值低于 4 像素，则判定为有效估计。2D 重投影误差准确率(Rep(%))是此误差阈值下，满足重投影误差小于该阈值的点的比例。

5cm5°精度阈值用于综合评价平移与旋转误差，需同时满足平移误差 e_{TE} <5 cm 和旋转误差 e_{RE} <5°。5cm5°精度准确率是此误差阈值下，满足要求样本的比例。平移和

旋转误差按下式计算：

$$e_{TE} = \left\| t - t' \right\|_2$$
$$e_{RE} = \arccos\left[\frac{\left(\mathrm{tr}(RR'^{-1}) - 1 \right)}{2} \right] \tag{5-25}$$

其中R是平移矩阵；R'是旋转矩阵。

为验证 RGBD Fusion 网络的性能优势，本研究对比了 GDR-Net[74]（RGB 单模态）、CloudPose[75]（点云单模态）、G2L-Net[76]（RGBD 多模态）以及本文方法在相同数据集下的表现，结果如表 5-2 所示：

表 5-2 姿态识别任务结果对比

方法	GDR-Net	CloudPose	G2L-Net	Ours
Rep(%)	76.67	79.29	80.20	83.14
5cm5°(%)	61.73	63.47	64.88	64.64
Time(fps)	25	13	23	31

通过结果比对可以发现，RGBD Fusion位姿估计网络重投影精度与实时性上均优于 GDR-Net、CloudPose 和 G2L-Net 位姿估计网络，且其处理速度达到 31 帧/秒，满足实时操作需求，验证了算法改进的有效性。

在量化指标评价对比后，为了更直观的评判姿态识别网络的性能，下面对姿态识别的结果进行可视化如图 5-13。可以看出可视化结果满足姿态识别要求，但同时揭示当前局限：在存在遮挡的环境下，存在部分物体未能正确识别的问题，这为后续研究指明改进方向。

图 5-13 姿态识别可视化结果

5.5 ROS 仿真实验

5.5.1 MoveIt 运动规划插件与 Rviz 仿真插件

为实现机器人运动规划与状态监控的一体化验证，本研究基于 ROS 框架集成 MoveIt 运动规划模块与 Rviz 三维可视化工具。MoveIt 作为核心控制器，其功能架构如主要包含以下功能链：

1.模型解析：通过 ROS 参数服务器加载 UR3 机械臂的 URDF 模型，解析各关节运动学参数与连杆约束条件。

2.轨迹规划：采用笛卡尔空间插补算法，将用户输入的末端执行器目标位姿(/move_group/goal)转化为关节角度序列(/joint_trajectory)。

3.碰撞检测：基于OMPL库实时计算运动路径的可行性，避免机械臂与场景障碍物发生干涉。

Rviz 则通过订阅多模态数据话题(/camera/rgb/image_raw、/point_cloud 及/tf)，动态渲染机器人位姿、环境点云及目标检测结果。两者的协同工作为虚实融合的系统调试提供了高保真仿真环境。

5.5.2 运动学仿真模型建立

精确的机器人运动学模型是仿真实验的基础。本研究基于 UR3 机械臂的 SolidWorks 装配体，通过 sw2urdf 插件将其转换为 URDF 格式。转换过程如图 5-14 所示。

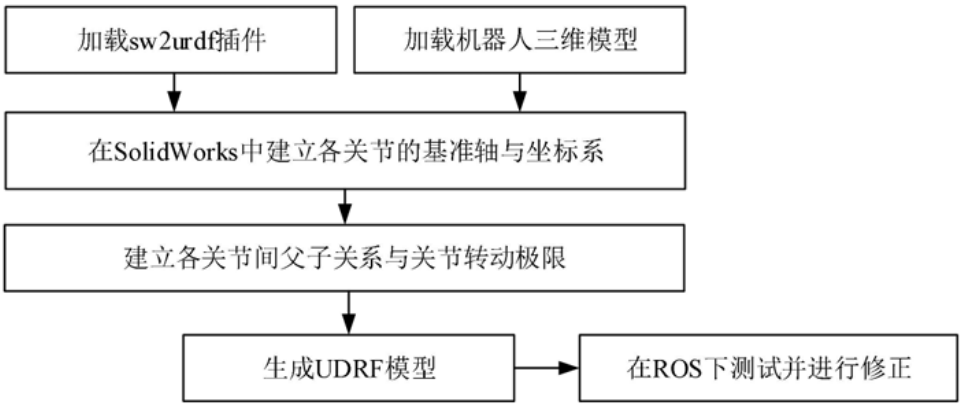


图 5-14 转换为 URDF 格式流程图

转换过程需定义以下关键参数：1.连杆属性：包括质量、惯性矩、几何尺寸、连杆长度 a、偏置距离 d 等。2.关节约束：设置旋转关节的转动范围及运动学类型。3.坐标

系绑定：依据 D-H 参数法建立各机器人连杆 link 与机器人关节 joint 坐标系如图 5-15，确保正运动学计算的准确性。

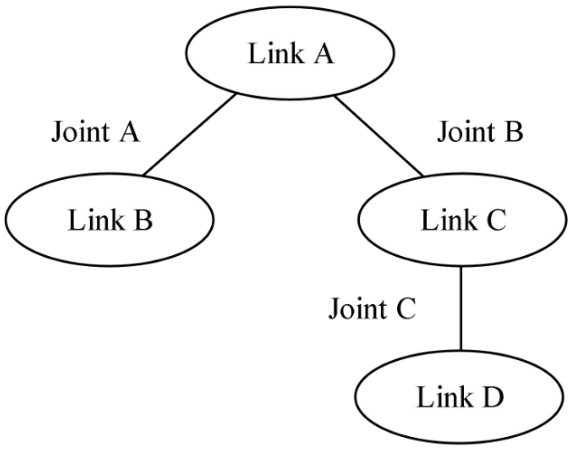
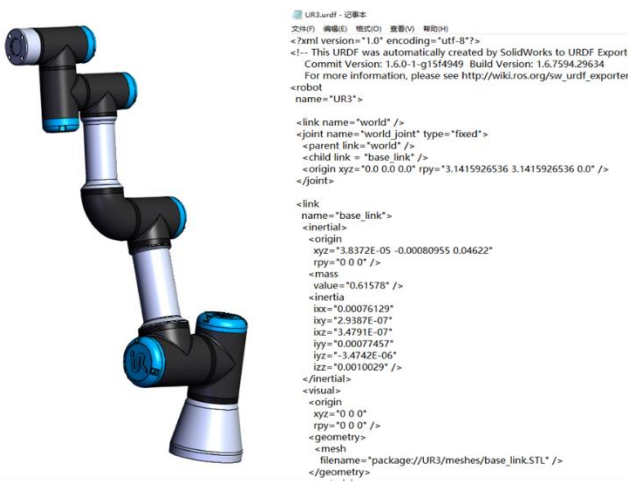


图 5-15 机械臂连杆与关节关系图

生成的 URDF 文件经 ROS 的 check_urdf 工具验证后，通过 MoveIt 的 Setup_Assistant 配置运动规划组（这里采用 arm_group 与 gripper_group），并设置基坐标系(base_link)与末端工具坐标系的转换关系。最终生成包含碰撞矩阵、虚拟关节定义及控制器接口的配置文件包，如图 5-16 所示。。



(a)URDF模型　　(b)URDF文件

图 5-16 生成 UR3 的 URDF 模型与对应的文件

5.5.3 实验流程

本实验进行的 ROS 通讯节点如图 5-17 所示。具体而言，首先订阅节点/UR3_move_group 来获取 UR3 关节转动角度数据，然后通过仿真控制器中的/move_group 节点发布的消息获取控制 UR3 运动所需数据。相机 camera 发布相机获取图像的节点/realsense2_camera，用于对目标进行三维目标检测的节点/prdnet_ros 和进行

6D 姿态估计的节点/rgbd_fusion_ros，用于发布眼在手外手眼标定坐标转换关系的节点/eye_to_hand，发布 UR3 机械臂关节和位置姿态的节点/UR3_joint_state 与/UR3_robot_state。最后，发布用于控制 UR3 机器臂使其按照/move_group 给出的规划路径的运动控制节点/UR3/ros_control 与/UR3/traj_controller_spawn 完成整体话题发布。

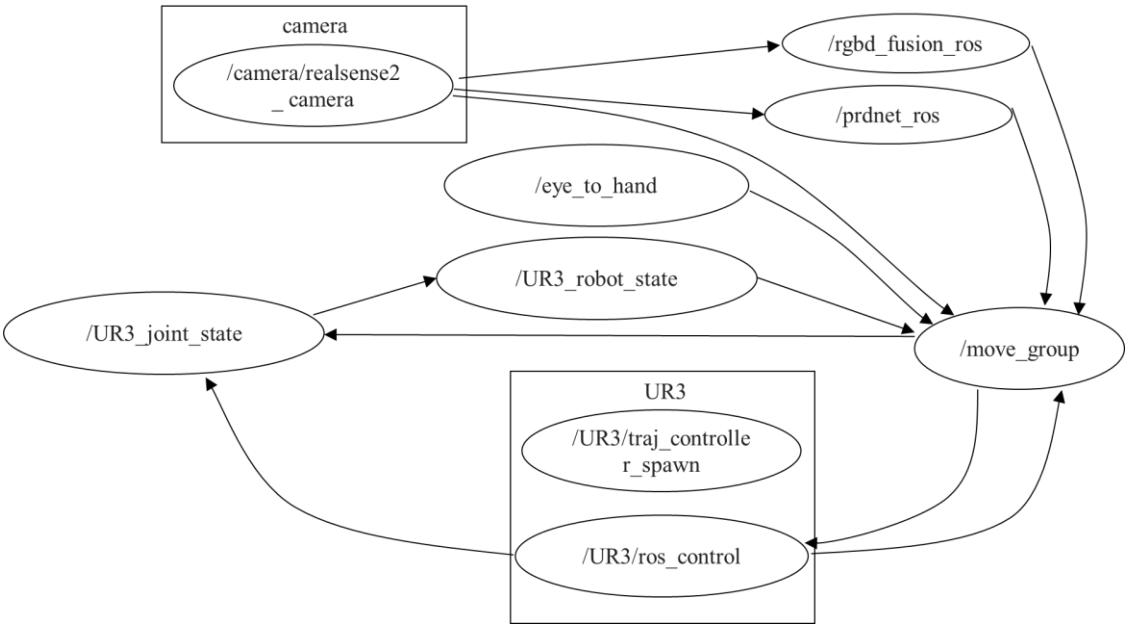


图 5-17 ROS 通讯节点图

为验证系统在复杂场景下的鲁棒性，实验设计涵盖多目标识别、动态轨迹规划及抓取成功率统计三个核心环节，其工作流如下：

1.环境初始化：在 Rviz 中加载包含 10 类目标物体及随机干扰物的仿真场景。

2.多模态感知：目标检测：通过/prdnet_ros 节点调用第四章的 PRDNet 算法，实时分割 RGB 图像中的目标区域。位姿估计：/rgbd_fusion_ros 节点融合 RGB 与深度特征，输出 6D 位姿参数（旋转矩阵 R 与平移向量 t）。

3.坐标转换与轨迹生成：利用手眼标定矩阵 T_C^B 将相机坐标系下的目标位姿转换至基坐标系。MoveIt 基于目标位姿序列生成无碰撞的关节运动轨迹，并通过/traj_controller 发布角度指令。

4.抓取执行与复位：UR3 机械臂按规划路径执行抓取动作后，返回初始位姿并触发新一轮检测-规划循环，直至场景清空。

5.5.4 仿真结果

首先在 RVIZ 中引入 UR3 机器人模型待检测场景，如图 5-18 所示，RVIZ 界面的左

边是各个窗口的状态栏，主界面是 realsense 采集的点云图像和 TF 坐标系机械臂模型，下方是采集的对应色彩图。

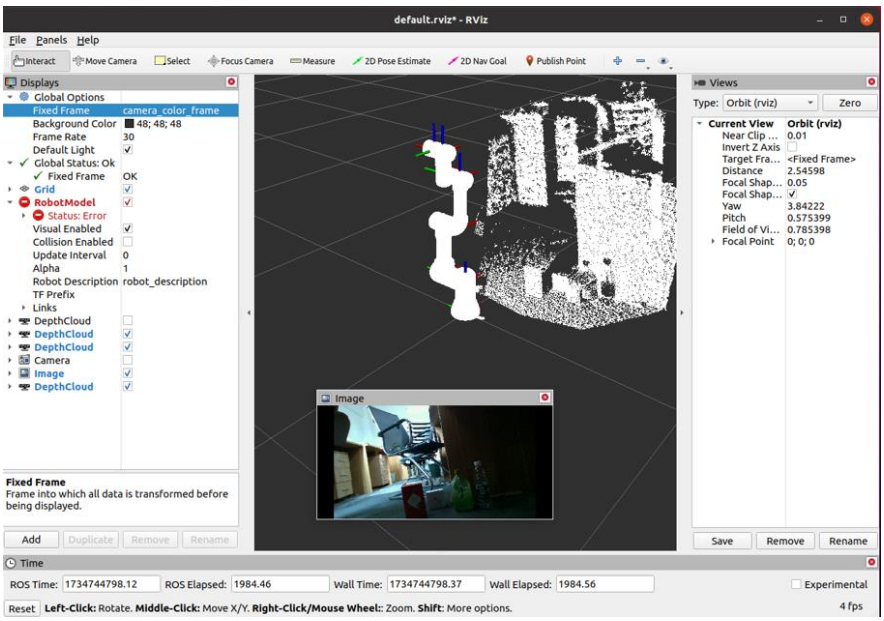
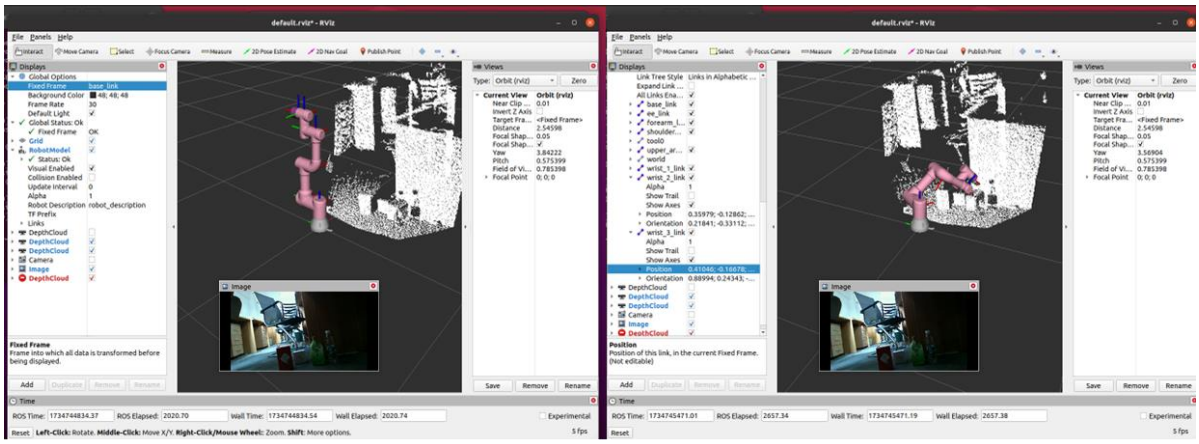


图 5-18 初始状态图

经过前一节所述的工作流，完成实验，如图 5-19



(a)开始实验　　(b)到达目标点

图 5-19 实验过程图

为避免实验的偶然性，故进行 500 次分拣实验（不同类别各 50 次），抓取成功率指标如表 5-3。

表 5-3 抓取实验结果

种类	椅子	箱体	瓶子	书本	杯子	键盘	垃圾桶	屏幕	灯	xbox
正确	46	42	48	41	42	46	43	42	42	43
总计	50	50	50	50	50	50	50	50	50	50
正确率	92%	84%	96%	82%	84%	93%	86%	84%	84%	86%

5.6 本章小结

本章围绕机器人智能分拣实验展开研究，详细阐述了相机内参标定、三维坐标定位以及机械臂运动学分析的相关方法。结合 RGBD Fusion 网络和 Refinement 模块的设计，本研究实现了目标物体的高精度姿态识别，并通过 ROS 平台验证了算法的可行性和稳定性。实验结果表明，所提出的方法在复杂环境下的目标分拣任务中具有较高的抗干扰性，同时较好地平衡了精确性与推理速度的取舍。

总结与展望

总结

本文针对工业机器人无序抓取任务，结合机器视觉和多模态数据融合方法，提出了创新的算法与实验验证方案，显著提升了目标检测、三维分割与识别精度，并为自动化分拣系统提供了技术支持。面对人口时代，有限的人力资源是远远不够的，因此需要智能制造技术来辅助工业生产。而本文研究的模型能够帮助工厂实现生产图像预处理，目标检测识别，生产对象的三维分割识别，降本增效，对后续智能制造系统的开发尽了绵薄之力。

本研究主要工作包括以下四个方面：

(1) 多模态数据集构建：通过融合 ScanNet 公共数据集与自采集的 RGB-D 数据，构建了包含 2416 张图像的多模态数据集，涵盖 10 类常见工业物体。采用高斯滤波、自适应参数变换等预处理方法提升数据质量，并通过 RoboFlow 与 CloudCompare 工具完成精细化标注，为后续算法训练与验证奠定了基础。

(2) 单模态目标检测算法改进：基于 YOLOv8 框架提出 AC-YOLO 网络，在 Neck 模块中引入 GSConv 特征融合与 ESlim-Neck 结构，并设计三维空间注意力机制(TDSA)。实验表明，改进后的模型在 mAP50 指标上达到 83.6%，参数量降低至原始模型的 54.5%，显著提升了检测精度与效率。

(3) 多模态三维分割与识别算法设计：构建 RGB-D 数据对齐的点云生成流程，提出融合 RGB 图像与点云特征的 Point Rgb-D 网络，结合改进的 E-ResNet 模块与稠密特征融合策略，在 Dice 系数和 mIoU 指标上分别达到 88.3%和 79.1%，优于主流点云分割算法，如 PointNet++、VoteNet 等。

(4) 机器人智能分拣实验验证：通过 ROS 仿真平台搭建 UR3 机械臂控制环境，结合 RGBD Fusion 网络与 Refinement 模块实现目标物体的高精度姿态识别。实验结果显示，系统对 10 类目标的分拣成功率平均达 86.8%，单次抓取识别耗时低于 31ms，验证了算法在工业生产实际应用的可行性。

展望

本文提出的方法在工业机器人无序抓取任务中展现了显著优势，但仍存在以下不

足，未来可从以下方向进一步优化：

(1) 数据集的多样性与泛化性：当前数据集主要基于实验室环境构建，未来可引入更多复杂工业场景，如动态光照、遮挡严重环境的数据，并探索跨模态数据增强技术，进一步提升模型的泛化能力。

(2) 算法的实时性与轻量化：尽管 AC-YOLO 与 Point Rgb-D 网络在效率上已取得改进，但在高动态场景下仍需更快的推理速度。未来可通过网络剪枝、量化或边缘计算优化模型部署，满足工业场景的实时性需求。

(3) 多机器人协同与动态规划：当前实验基于单机械臂分拣任务，未来可扩展至多机器人协同抓取，并结合强化学习实现动态环境下的自适应运动规划。

(4) 系统集成与实际验证：目前算法验证主要在仿真环境中完成，后续需在真实工业场景中搭建硬件平台，结合前端交互系统实现完整的“感知-决策-执行”闭环，推动技术落地应用。

参考文献

[1] 梁婧,王静.我国制造业结构变迁、发展趋势与政策思考[J].国际金融,2023,(07):57-65.

[2] 李磊,何艳辉.机器人应用与个体就业——基于人口普查数据的研究[J].财贸研究,2024,35(11):31-42.

[3] 《国家及各地区国民经济和社会发展第十四个五年规划和 2035 年远景目标纲要》[J].中国信息界,2022,(05):110.

[4] 侯智,张艾.工业机器人产业发展现状与人才需求分析[J].机器人产业,2023,(04):95-100.

[5] 王洪亮.智能机器人技术在智能制造中的应用[J].电子技术,2023,52(09):382-383.

[6] 刘亚欣,王斯瑶,姚玉峰,等.机器人抓取检测技术的研究现状[J].控制与决策,2020,35(12):2817-2828.

[7] 罗华,尚俊云,李瑞峰,等.一种面向工业机器人的高精度三维无序抓取系统[J].导航与控制,2023,22(04):106-116.

[8] 曲建利.机器人生产的价值属性及社会价值——基于马克思主义劳动价值论[J].学理论,2020,(07):29-30.

[9] 王连庆,钱莉.基于视觉引导的工业机器人无序抓取系统设计[J].制造业自动化,2022,44(03):86-89+196.

[10] 吴昉,王伟,陈海永.机器人视觉智能识别与无序抓取实验系统设计[J].机器人技术与应用,2022,(05):38-45.

[11] 邱益,张志浩,梁杰.基于深度学习的无序件抓取实验系统开发[J].实验室研究与探索,2022,41(02):84-88+99.

[12] 董豪,杨静,李少波,等.基于深度强化学习的机器人运动控制研究进展[J].控制与决策,2022,37(02):278-292.

[13] 何浩源,尚伟伟,张飞,等.基于深度神经网络的多指灵巧手抓取手势优化[J].机器人,2023,45(01):38-47.

[14] 罗华,尚俊云,李瑞峰,等.一种面向工业机器人的高精度三维无序抓取系统[J].导航与控制,2023,22(04):106-116.

[15] Wang B, Tao F, Fang X, et al. Smart manufacturing and intelligent manufacturing: A comparative review[J]. Engineering, 2021, 7(6): 738-757.

[16] Lee I. Service robots: a systematic literature review[J]. Electronics, 2021, 10(21): 2658.

[17] Bulla C, Parushetti C, Teli A, et al. A review of AI based medical assistant chatbot[J]. Res Appl Web Dev Des, 2020, 3(2): 1-14.

[18] Xu A, Wang L, Feng Setl al. Threshold-based level set method of image segmentation[C]. 2010 Third International Conference on Intelligent Networks and Intelligent Systems, 2010: 703-706.

[19] Cigla C, Alatan A A. Region-based image segmentation via graph cuts[C]. 2008 15th IEEE International Conference on Image Processing, 2008: 2272-2275.

[20] Yu-Qian Z, Wei-Hua G, Zhen-Cheng Cetl al. Medical images edge detection based on mathematical morphology[C]. 2005 IEEE engineering in medicine and biology 27th annual conference, 2006: 6492-6495.

[21] Ciregan, Dan, Ueli Meier, and Jürgen Schmidhuber. Multi-column deep neural networks for image classification[C]. 2012 IEEE Conference on Computer Vision and Pattern Recognition. 2012: 3642-3649.

[22] Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation[C]. Proceedings of the IEEE conference on computer vision and pattern recognition. 2015: 3431-3440.

[23] Tao X, Paoletti M E, Han L, et al. A new deep convolutional network for effective hyperspectral unmixing[J]. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2022, 15: 6999-7012.

[24] Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation[C].Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18. 2015: 234-241.

[25] Guo C, Szemenyei M, Yi Y, et al. Sa-unet: Spatial attention u-net for retinal vessel segmentation[C].2020 25th international conference on pattern recognition (ICPR). 2021: 1236-1242.

[26] Li B, Wu F, Liu S, et al. CA‐Unet++: An improved structure for medical CT scanning based on the Unet++ Architecture[J]. International Journal of Intelligent Systems, 2022, 37(11): 8814-8832.

[27] 邹军,张世义,李军.基于深度学习的目标检测算法综述[J].传感器世界,2023,29(08):9-15.

[28] 蔡嘉磊,茅智慧,李君,等.基于深度学习的目标检测算法与应用综述[J].网络安全技术与应用,2023,(11):41-45.

[29] 郑太雄,江明哲,冯明驰.基于视觉的采摘机器人目标识别与定位方法研究综述[J].仪器仪表学报,2021,42(09):28-51.

[30] Prasanna G V S, Pavani K, Singh M K. Spliced images detection by using Viola-Jones algorithms method[J]. Materials Today: Proceedings, 2022, 51: 924-927.

[31] Surasak T, Takahiro I, Cheng C, et al. Histogram of oriented gradients for human detection in video[C].2018 5th International conference on business and industrial research (ICBIR). IEEE, 2018: 172-176.

[32] Lim S H, Erichson N B, Hodgkinson L, et al. Noisy recurrent neural networks[J]. Advances in Neural Information Processing Systems, 2021, 34: 5124-5137.

[33] Subakan C, Ravanelli M, Cornell S, et al. Attention is all you need in speech separation[C].ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2021: 21-25.

[34] Krizhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutional neural networks[J]. Advances in neural information processing systems, 2012, 25.

[35] Girshick R, Donahue J, Darrell T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation[C].Proceedings of the IEEE conference on computer vision and pattern recognition. 2014: 580-587.

[36] He K, Zhang X, Ren S, et al. Spatial pyramid pooling in deep convolutional networks for visual recognition[J]. IEEE transactions on pattern analysis and machine intelligence, 2015, 37(9): 1904-1916.

[37] Girshick R. Fast r-cnn[C].Proceedings of the IEEE international conference on computer vision. 2015: 1440-1448.

[38] Ren S, He K, Girshick R, et al. Faster r-cnn: Towards real-time object detection with region proposal networks[J]. Advances in neural information processing systems, 2015, 28.

[39] Redmon J, Divvala S, Girshick R, et al. You only look once: Unified, real-time objectdetection[C].Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 779-788.

[40] Liu W, Anguelov D, Erhan D, et al. Ssd: Single shot multibox detector[C].Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14. 2016: 21-37.

[41] Lin T Y, Goyal P, Girshick R, et al. Focal loss for dense object detection[C].Proceedings of the IEEE international conference on computer vision. 2017: 2980-2988.

[42] Law H, Deng J. Cornernet: Detecting objects as paired keypoints[C].Proceedings of the European conference on computer vision (ECCV). 2018: 734-750.

[43] Lou H, Duan X, Guo J, et al. DC-YOLOv8: small-size object detection algorithm based on camera sensor[J]. Electronics, 2023, 12(10): 2323.

[44] Mao M, Lee A, Hong M. Efficient Fabric Classification and Object Detection Using YOLOv10[J]. Electronics (2079-9292), 2024, 13(19).

[45] Sahbani A, El-Khoury S, Bidaud P. An overview of 3D object grasp synthesis algorithms[J].Robotics and Autonomous Systems, 2012, 60(3): 326-36.

[46] Li JW, Liu H, Cai HG. On computing three-finger force-closure grasps of 2-D and 3-D objects[J]. IEEE Transactions on Robotics and Automation, 2003, 19(1): 155 -61.

[47] Haschke R, Steil JJ, Steuwer I, et al. Task-oriented quality measures for dextrous grasping[C]. International Symposium on Computational Intelligence in Robotics and Automation (ICRA),Barcelona, Spain, 2005: 689-694.

[48] Bicchi A, Kumar V. Robotic grasping and contact: A review[C]. IEEE Internationa Conference on Robotics and Automation (ICRA), San Francisco, United States, 2000: 348 -353.

[49] Bohg J, Morales A, Asfour T, et al. Data-driven grasp synthesis—a survey[J]. IEEE Transactions on Robotics, 2013, 30(2): 289-309.

[50] Turk MA, Pentland AP. Face recognition using eigenfaces[C]. Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR), Lahaina, United States, 1991: 586-587.

[51] Lowe DG. Object recognition from local scale-invariant features[C]. Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR), Kerkyra, Greece, 1999:1150-1157.

[52] Shao L, Ferreira F, Jorda M, et al. Unigrasp: Learning a unified model to grasp with multifingered robotic hands[J]. IEEE Robotics and Automation Letters, 2020, 5(2): 2286 -93.

[53] Mahler J, Pokorny FT, Hou B,et al. Dex-net 1.0: A cloud-based network of 3d objects for robust grasp planning using a multi-armed bandit model with correlated rewards[C]. IEEE international conference on robotics and automation (ICRA), Stockholm, Sweden, 2016:1957-1964.

[54] 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[J]. arXiv preprint arXiv: 1703.09312, 2017.

[55] Mahler J, Goldberg K. Learning deep policies for robot bin picking by simulating robust grasping sequences[C]. Conference on robot learning (CoRL), Auckland, New Zealand, 2017:515-524.

[56] Mahler J, Matl M, Satish V, et al. Learning ambidextrous robot grasping policies[J]. Science Robotics, 2019, 4(26): 59-67.

[57] Saxena A, Driemeyer J, Ng AY. Robotic grasping of novel objects using vision[J].The International Journal of Robotics Research, 2008, 27(2): 157 -73.

[58] Le QV, Kamm D, Kara AF, et al. Learning to grasp objects with multiple contact points[C].IEEE International Conference on Robotics and Automation (ICRA), Stockholm, Sweden,2010: 5062-5069.

[59] Jiang, Yun, Stephen Moseson, and Ashutosh Saxena. Efficient grasping from rgbd images:Learning using a new rectangle representation[C]. IEEE International conference on robotics and automation (ICRA), Shanghai, China, 2011: 3304-3311.

[60] Lenz I, Lee H, Saxena A. Deep learning for detecting robotic grasps[J]. The International Journal of Robotics Research, 2015, (4-5): 705-724.

[61] Wang Z, Li Z, Wang B, et al. Robot grasp detection using multimodal deep convolutional neural networks[J]. Advances in Mechanical Engineering, 2016, 8(9): 2304 -2311.

[62] Asif U, Bennamoun M, Sohel FA. RGB-D object recognition and grasp detection using hierarchical cascaded forests[J]. IEEE Transactions on Robotics, 2017, 33(3): 547-564.

[63] 夏晶,钱堃,马旭东,等.基于级联卷积神经网络的机器人平面抓取位姿快速检测[J].机器人, 2018, 40(06): 794-802.

[64] 马倩倩,李晓娟,施智平.轻量级卷积神经网络的机器人抓取检测研究[J].计算机工程与应用,2020,56(10):141-148.

[65] Ma F, Karaman S. Sparse-to-dense: Depth prediction from sparse depth samples and a single image[C]. 2018 IEEE international conference on robotics and automation (ICRA). IEEE, 2018: 4796-4803.

[66] Dai A, Chang A X, Savva M, et al. Scannet: Richly-annotated 3d reconstructions of indoor scenes[C].Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 5828-5839.

[67] Song S, Lichtenberg S P, Xiao J. Sun rgb-d: A rgb-d scene understanding benchmark suite[C].Proceedings of the IEEE conference on computer vision and pattern recognition. 2015: 567-576.

[68] Dolatabadi E, Taati B, Mihailidis A. Concurrent validity of the Microsoft Kinect for Windows v2 for measuring spatiotemporal gait parameters[J]. Medical engineering & physics, 2016, 38(9): 952-958.

[69] Li Y, Miao N, Ma L, et al. Transformer for object detection: Review and benchmark[J]. Engineering Applications of Artificial Intelligence, 2023, 126: 107021.

[70] Qi C R, Su H, Mo K, et al. Pointnet: Deep learning on point sets for 3d classification and segmentation[C].Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 652-660.

[71] Qian G, Li Y, Peng H, et al. Pointnext: Revisiting pointnet++ with improved training and scaling strategies[J]. Advances in neural information processing systems, 2022, 35: 23192-23204.

[72] Hou J, Dai A, Nießner M. 3d-sis: 3d semantic instance segmentation of rgb-d scans[C].Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019: 4421-4430.

[73] Bülow G. Detecting deformable linear objects in 3D point clouds using Deep Hough Voting[J]. 2024.

[74] Feng W, Zhang J, Zhou Y, et al. GDR-Net: A geometric detail recovering network for 3D scanned objects[J]. IEEE Transactions on Visualization and Computer Graphics, 2021, 28(12): 3959-3973.

[75] Gao G, Lauri M, Wang Y, et al. 6d object pose regression via supervised learning on point clouds[C].2020 IEEE International Conference on Robotics and Automation (ICRA). IEEE, 2020: 3643-3649.

[76] Chen W, Jia X, Chang H J, et al. G2l-net: Global to local network for real-time 6d pose estimation with embedding vector features[C].Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020: 4233-4242.

攻读硕士学位期间发表学术论文及参研项目

（一）参与的科研项目

[1] 基于 3D 视觉的复杂工件缺陷检测机理和实验研究. 重点研发与成果转化项目，课题编号：2023yf092

（二）发表的学术论文

[1] An Enhanced YOLOv5-Based Algorithm for Metal Surface Defect Detection[J]. Applied Sciences, 2023, 13(20): 11473. (JCR Q4，作者三作)

[2] 基于改进 MobileNet 网络的多类别螺母识别研究[J].安徽工程大学学报,2025.(已录用，待发表，作者一作)

[3] Research on 3D segmentation and recognition of point cloud data based on PointNet network [J].IEEE Transactions on Pattern Analysis and Machine Intelligence,2025(初审，作者一作)

致谢

行文至此，落笔为终。回首三年的硕士求学生涯，既有探索未知时的迷茫与艰辛，也有突破瓶颈后的喜悦与成长。这段旅程的每一步都离不开师长、同门、亲友的指导与支持，在此谨以最诚挚的谢意献给所有给予我帮助的人。

饮水思源，师恩难忘。衷心感谢我的导师王海教授！从论文选题到框架设计，从实验验证到成果凝练，王老师始终以渊博的学识和严谨的治学态度为我指明方向。他开阔的学术视野、敏锐的科研洞察力以及因材施教的教学理念，让我深刻领悟到学术研究的真谛。生活中，王老师谦和的为人与豁达的处世态度，更让我学会了如何平衡科研与生活。师者如光，虽微致远，能成为王老师的学生，是我求学路上莫大的幸运。

同舟共济，携手前行。感谢赵亚玲、李帅、张子豪等师兄师姐在科研入门阶段的倾囊相授，你们无私分享的文献资源、实验技巧与项目经验，为我奠定了扎实的研究基础；感谢解勇正、袁义凯、卢良玉等同窗挚友，无数个并肩攻关的日夜、思维碰撞的讨论，让科研之路充满温度；感谢金傲龙、屈润昌、李保、朱文博、陈佳佳等师弟师妹在数据采集、代码调试中的鼎力相助，你们的热情与活力为实验室注入了蓬勃生机。团队协作的力量让我深知，科研从不是一个人的孤军奋战。

春晖寸草，亲情永驻。感谢父母和兄长数十年如一日的默默付出，你们始终以最朴素的方式支持我的每一次选择，用包容与理解为我筑起最温暖的港湾；感谢挚友们的鼓励与陪伴，低谷时的关怀与成功时的喝彩，皆成为我砥砺前行的动力。

以梦为马，未来可期。谨以此文献给所有关心我的人，愿我们保持热爱，奔赴下一场山海！