**睿尔曼新零售机器人落地方案**

**技术方案（v1.0）**

**目录**

[**睿尔曼新零售机器人落地方案** 1](#_Toc177473860)

[**1. 整体方案技术架构 4**](#_Toc177473862)

[**1.1 项目目标** 4](#_Toc177473863)

[**1.2 产品形态** 4](#_Toc177473864)

[**1.3产品工作形式** 8](#_Toc177473865)

[**2. 国内外研究现状 9**](#_Toc177473866)

[**2.1 根据RGB图像的目标检测网络分类现状（周俊桦）** 9](#_Toc177473867)

[**2.2 6-DoF位姿估计现状（林伟鸿）** 9](#_Toc177473868)

[**2.3 抓取姿态估计现状（施翰煜）** 11](#_Toc177473869)

[**3. 项目关键研究以及拟解决的关键问题 11**](#_Toc177473870)

[**3.1 关键研究内容（林伟鸿）** 11](#_Toc177473871)

[**3.2 拟解决关键问题：** 12](#_Toc177473872)

[3.2.1实时场景下物体目标检测识别技术（周俊桦） 12](#_Toc177473873)

[3.2.2高精度的物体实时位姿估计及位姿跟踪技术（林伟鸿） 13](#_Toc177473874)

[3.2.3常见物体的实时抓取估计技术（施翰煜） 13](#_Toc177473875)

[3.2.4复杂环境中机器人作业运动轨迹规划技术（黄泊凯） 14](#_Toc177473876)

[**4. 拟采取的研究解决方案 14**](#_Toc177473877)

[**4.1 总体技术架构（林伟鸿）** 14](#_Toc177473878)

[4.1.1实时场景下物体目标检测识别技术（周俊桦） 15](#_Toc177473879)

[4.1.2高精度的物体实时位姿估计及位姿跟踪技术（林伟鸿） 17](#_Toc177473880)

[4.1.3常见物体的实时抓取估计技术（施翰煜） 21](#_Toc177473881)

[4.1.4复杂环境中机器人作业运动轨迹规划技术（黄泊凯） 25](#_Toc177473882)

[**5. 预期成果与工作计划（林伟鸿） 25**](#_Toc177473883)

[**5.1 预期成果** 25](#_Toc177473884)

[**5.2 项目研究计划** 26](#_Toc177473885)

# 整体方案技术架构

本项目旨在开发一款高度自动化的智能机器人系统，用于超市环境中的补货和出货任务。该系统将集成先进的感知传感器和机械臂技术，以实现高效、准确的物品处理和库存管理。

* 1. **项目目标**

(1)自动化补货：根据所给指令将补货篮中的物体补货到货架上，通过机器人自带的感知传感器，规划行动路线，精简补货流程形成自动化。

(2)智能出货：根据顾客所需购买产品，自动规划到达货架的路线，通过传感器感知环境并从货架上取得商品并放置于商品篮中。

* 1. **产品形态**

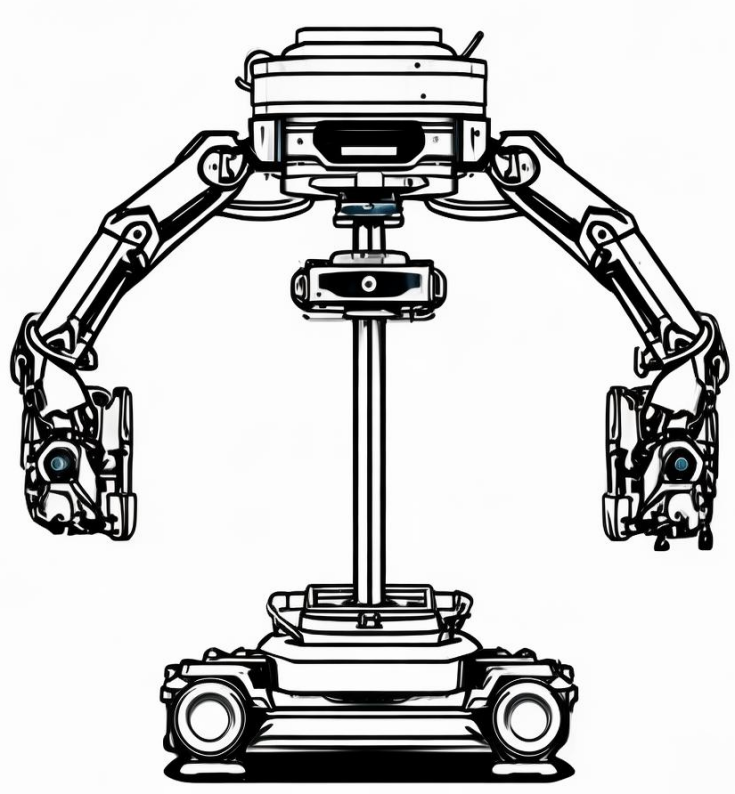
产品由四个重要硬件机械部分组成，分别是：1. 用于操作的7自由度升降双臂系统；2. 用于夹取或者吸取的末端执行工具；3. 用于感知环境的单相机或多相机系统；4. 用于移动的底盘基座以及相关硬件联动控制系统。

图1-1机器人产品形态简笔概念图

其中：

(1)双臂升降系统用于使得6自由度机械臂能够处理不同高度的货架商品，扩大机器人的工作空间，更好的避免双臂系统之间的操作影响。可以选用力纳克的LC3 IC：



图1-2：力纳克DL6

LC3 IC 电动升降柱配有内置控制器，便于集成到工业应用中。LC3 IC 可以 100 mm/sec 的速度推动或拉动 1000 N 的负载。速度可调，不受负载影响。

以及SKF FRE系列升降柱



图1-3：SKF FRE系列升降柱

SKF FRE系列可以 10-15mm/sec 的速度推动或拉动 2000 N 的负载。动态偏载250Nm。

(2)吸取或夹取的末端执行工具，可以采用市面技术上较为成熟的二指夹爪和独立吸盘。二指夹爪可选用Robotiq 2F-85:



图1-4：Robotiq 2F-85

Robotiq 2F-85夹爪最大行程为85mm,夹力为20-235N，抓握荷载5KG，位置分辨精度可达0.4mm。

以及Onrobot RG6：



图1-5：Onrobot RG6

Onrobot RG2夹爪最大行程为160mm，夹力25-120N，抓握荷载6KG.位置分辨率0.1mm。

吸盘可选择Robotiq E Pick吸盘：



图1-6：Robotiq E Pick

吸盘采用电动式，最大负载10KG，真空度80%。

以及范德华力吸盘-VFC:



图1-7：范德华力吸盘-VFC

吸盘采用电动式，最大有效荷载1KG。

(3)机器人视觉系统：机器人需要相机组成的视觉系统来感知周围的环境，在经过手眼标定后，机械臂基座与相机之间的转换关系确定，才能让机器人具有感知周围环境的能力。在头部安装相机可以让机器人获取更大范围的视角，相机可以选用Gemini336L：



图1-7：Gemini336L

Rgb模式支持1280x800@60fps，深度模式支持1280x800@30fps，推荐工作区间0.25-6.0m。

以及realsenseD435i：



图1-8：RealsenseD435i

Rgb模式支持1920x1080@30fps，深度模式支持1280x720@30fps，推荐工作区间0.3-10.0m。

**1.3产品工作形式**

(1)商品管理系统

本项目的商品管理系统是用于管理商店的商品详细信息、库存容量、商品定价等业务流程的软件系统，主要功能包括：

1)商品信息管理：实时管理记录商品类目、所在货架、入库时间、保质期等信息，并能完成添加、修改和删除指令。

2)库存管理：实时监控库存水平，自动计算补货需求，确保库存充足，在发起补货指令后，指引机器人前往补货平台，完成补货。

3)订单管理：处理客户发出的订单指令，指引机器人前往货架处，夹取指定货物于货物篮中。

4)促销等其他管理：创建和管理促销活动、打折等其他操作。

(2)补货

当机器人受到来自商品管理系统的补货指令后，将前往理货台。理货台上放置需要配送到货架上的商品，机器人根据头部相机的视觉系统以及理货台上的重量传感器联合识别当前货架上是否存在货物，当识别到物品分类后，机器人会使用合适的执行工具夹取或吸取商品，并移动到货架放置商品。

理货指令

机器人移动到理货台

视觉系统和重量传感器联合识别

分拣商品

移动并放置商品到对应货架

图1-9：机器人理货流程

(3)取货

当机器人收到取货指令时，根据管理系统内部给定的商品货架坐标移动到指定点位，通过视觉感知夹取商品，放置于货篮中并带回。

移动到指令位置

取货指令

识别商品

分拣商品

图1-10:机器人取货流程

(4)调度

中央调度软件系统部署在服务器或者独立设备单元中，并与各移动平台的嵌入式边缘硬件控制系统连接。

当中央调度系统接受到来自仓储管理软件系统的移动运输指令及坐标时，规划并生成行驶路线，随后下发至离起始坐标最近的闲置作业平台或者是在某一作业平台完成上一指令任务后下发，启动平台开始作业。同时在作业过程中，借助平台搭载的嵌入式硬件控制终端及其上的传感器测算实际坐标并调整路线，来避免移动过程中的机械碰撞。

(5)运输

移动平台边缘硬件控制系统由边缘计算工作站和所搭载的传感器组成。

当移动平台一旦通电启动，则硬件控制系统将一直工作至平台断电关闭。系统工作期间，负责维持平台姿态平衡稳定，检测并防止可能发生的机械碰撞，以保证运输物品不会移位甚至脱落。同时通过与中央调度系统的联动来实现行驶过程中的避障，以保证平台乃至整个作业区域的正常运输运转。

# 国内外研究现状

* 1. **根据RGB图像的目标检测网络分类现状（周俊桦）**

RGB图像的目标检测在各个场景的应用起到重要作用，场景涵盖军事，工业，医学等诸多领域。依托于RGB图像的目标检测技术，可以在多场景下对不同对象进行定位，分类的工作。近年来，基于RGB图像的目标检测分类网络在深度学习的浪潮中取得了飞速发展，一系列具有里程碑意义的模型和算法相继问世，极大地推动了这一领域的技术进步。

自2015年起，Joseph Redmon提出了Yolo（You Only Look Once）的目标检测网络架构，YOLOv1是一种革命性的目标检测算法，它将目标检测任务视为一个单一的回归问题，可以在实时的情况下直接从图像中预测边界框和类别概率，这是一种有锚框的检测方法。在目标检测的领域中，除了Yolo架构的模型，同时还有基于transformer架构的模型，例如由Facebook团队提出的DETR（Detection Transformer），通过自注意力机制实现了端到端的检测，为目标检测领域带来了新的视角。

随后的时间中，CornerNet创新性地通过检测目标的角点来定位目标，这一方法避免了传统锚框设计带来的复杂度。CenterNet在此基础上进一步发展，通过检测目标的中心点来提高检测的准确率。同年，EfficientDet提出了EfficientNet作为主干网络，并通过复合缩放方法优化了网络结构和检测速度，使得实时目标检测成为可能。

YOLOv4结合了多种先进的图像处理技术和网络架构，进一步提升了YOLO系列在目标检测领域的性能。不久，-anchor的Anchor-Free检测方法如FCOS（Fully Convolutional One-Stage Object Detection）也开始受到关注，它们通过消除锚框设计，简化了检测流程。

近期在工业上表现优异的模型是Yolo及其种类繁多的变体，Yolo系模型的检测精度和速度在工业上都达成了较好的平衡。近期在Yolo系的模型当中，出现了很多值得注意的新方法：值得关注的创新是跨阶段部分网络（CSPNet）的应用。CSPNet通过将基础层的特征图分为两部分，一部分直接连接到后面的层，另一部分则经过多个卷积层后再连接，这种设计有效地减少了计算量，并提高了梯度传递的效率，这对于模型的训练和优化至关重要。在数据准备和增强方面，Yolo通过采用多种数据增强技术，如Mosaic和CutMix，YOLOv4能够生成更多样化的训练样本，从而提高模型在复杂场景下的泛化能力。

* 1. **6-DoF位姿估计现状（林伟鸿）**

六自由度（6-DoF）目标姿态估计在各种应用中具有重要意义，这些应用包括增强现实、机器人操作和自动驾驶等。由于光照和遮挡的变化，不同姿态的对象在不同视角下的外观可能会有很大差异，这给从 单一彩色图像中估计6-DoF对象姿态带来了巨大挑战。

基于学习的早期阶段方法通常一次性估计6-DoF姿态。Yu等人提出使用卷积神经网络直接回归对象中心和距离以及旋转分量。近期的工作则提出首先估计观测图像与对象模型之间的2D-3D对应关系，然后使用PnP算法求解对象姿态。这些一次性方法的性能通常受到网络容量的限制，并且容易受到不良条件（如光照变化和遮挡）的影响。PVNet基于关键点的预测，在点云质量良好的情况下在LINEMOD数据集上可以左右99%以上的精度。

随着深度学习的普及，为了避免昂贵的深度传感器，部分算法使用纯粹RGB预测位姿估计的方法，代表的方法有YOLOV5-6D，利用yolo网络的主题框架，引入空间注意力，相比于其他的RNN算法，在推理时常和预测准确度中做到了不错的平衡。 SAM-6D通过两个步骤来实现零样本 6D 物体姿态估计，包括实例分割和姿态估计，给定任意目标物体，SAM-6D 利用两个专用子网络，即实例分割模型（ISM）和姿态估计模型（PEM），来从 RGB-D 场景图像中实现目标；其中，ISM 将 SAM 作为一个优秀的起点，结合精心设计的物体匹配分数来实现对任意物体的实例分割，PEM 通过局部到局部的两阶段点集匹配过程来解决物体姿态问题。表现最出色的方法通常包含姿态细化过程，这一步骤极大提升了性能。FoundationPose 为减少大规模训练的人工工作，利用LLM和Diffusion开发了一种新的合成数据生成Pipeline。为了弥补无模型和基于模型的设置之间的差距，FoundationPose利用以对象为中心的神经场来进行随后的渲染和新视图RGBD渲染。通过两层网络，粗位姿估计和姿态细化，结果表明相比直接预测物体位姿大大提高了准确度，降低了模型参数量。

* 1. **抓取姿态估计现状（施翰煜）**

机器人准确且无碰撞的执行抓取动作的能力在人工智能领域具有广泛的应用潜力，例如装配、拾取和放置以及包装。从单个点云预测抓取点及抓取姿态仍然存在很多困难，一次成功的抓取不仅需要了解物体的几何形状，还需要了解其物理特性，包括质量、形状和摩擦力。基于深度学习的抓取检测算法，它们可以分为三大类，其中最流行的一种基于 RGB-D 图像输入检测可抓取点的方法。Lenz等人提出了一种具有两个网络的级联方法，该方法首先修剪掉不太可能的抓取点，然后使用更大的网络评估剩余的抓取点。Redmon等人提出了一种更快、更准确，以单步方式直接回归抓取姿势的网络。这些方法生成的抓取姿势被约束在 2D 平面中，这限制了抓取姿势的自由度。

随着单目物体 6D 姿态估计的快速发展，一些研究人员通过预测物体的 6D 姿态并将预定义的抓取姿态投影到场景中。Breyer等人提出在杂乱环境中进行抓取的方法，这些方法依赖于单一视点的点云观测，直接在完全观察到的点云上运行，是工作在整体场景级别上的，没有明确推理单个对象实例，将这些方法应用于拾放或目标驱动的抓取任务，通常需要额外的后处理或外部组件。

René Zurbrügg等人提出一种用于以对象为中心的抓取的端到端架构。 该方法使用来自单个任意观察方向的点云数据作为输入，并为场景中每个部分观察的对象生成以实例为中心的表示。 这种表示进一步用于杂乱桌面场景中的对象重建和抓取检测。结果表明相比于单一视点的点云观测，该方法在杂乱环境中的抓取能力更好，并且具有很强的泛化性。

* 1. **路径规划及移动避障（黄泊凯）**

路径规划及行驶避障等此类对作业区域内单台乃至多台机械设备运作的调度规划技术，随着机器人领域行业的兴起也一起发展。当下在多个领域产业路径规划及行驶避障技术都有相当规模的应用，尤其是物流仓储行业。但是随着作业区域范围的扩大，作业机械种类和数量的增多，实现路径规划以及调度和避障控制成为了维持系统正常运转的巨大挑战。

由于机器人行业本身的行业领域特性，路径规划方法随着早期机器人行业的研究被一起提出并被研究。其中，来自斯坦福大学的研究团队开发的Shakey机器人就是早期应用路径规划概念和技术的代表之一。

路径规划在多数场景下指的是，对作业区域进行2D平面建图，在此基础上，预测生成区域内的作业平台最优的移动行驶路线。早期阶段的路径规划方法一般采用计算机领域的图论算法，例如 Dijikstra算法、 A\*算法、 RRT算法等经典图论算法，上述算法均关注在 2D 图中如何生成点对点的移动路径。但是早期阶段的机器人应用场景中，作业区域相对狭小，作业平台数量相对较少，工作任务也比较简单单一。近些年随着机器人行业的进一步发展，还有机器人操作系统（ROS等）和嵌入式系统的成熟，以及传感器的种类的丰富和制造水平的提升，机器人的应用场景领域迅速扩大，应用规模爆炸式增长，传统算法及其优化衍生出的算法逐渐无法满足越来越复杂的应用要求。随着人工智能领域的发展成熟，当下在机器人行业开始将传统路径规划方法与深度学习结合。其中比较典型的一种是基于传统路径规划方法，结合深度强化学习DRL，并借用不同的神经网络结构处理来自不同传感器的各种模态输入，随后通过融合环境特征，就组合形成了SLAM技术。借助硬件控制系统，实现路径规划调度和行驶避障。

# 项目关键研究以及拟解决的关键问题

* 1. **关键研究内容（林伟鸿）**

本项目是基于视觉驱动的机器人超市场景，结合人工智能实现机器人在无人超市中完成取货补货流程。整体研究内容架构如下图所示：

实际作业场景验证

应用验证

* 精确目标检测识别
* 目标实例分割
* 基于RGB-D的位姿估计
* 基于RGB-D的抓取估计
* 空间路径规划

复杂商品目标检测与实例分割技术

实时高精度物体位姿估计技术

物体稳定抓取点估计技术

机器人移动避障以及运动规划技术

4项关键技术

自动取货送货机器人具身智能技术

数据采集困难，复杂商品准确分类和实例分割困难

不同商品的抓取点估计不同，且抓取后需要稳定姿态放置

机器人移动与路径碰撞检测困难

关键难点

场景提炼

图 3-1：项目总体研究框架

* 1. **拟解决关键问题：**

项目关键解决问题主要从实时场景下物体目标检测识别技术、高精度的物体实时位姿估计及位姿跟踪技术、常见物体的实时抓取估计技术、复杂环境中机器人作业运动轨迹规划技术等方面展开，具体如下：

### 3.2.1实时场景下物体目标检测识别技术（周俊桦）

机器人在识别货篮中的商品时，模型需要通过机载摄像机拍下的照片来对货篮中的商品进行识别，定位。在本项目中，必然遇到一些训练上的问题，首先是数据采集，训练一个可用的模型需要大量的训练集，在该项目中我们需要对种类繁多的商品去进行识别，几何倍数增长的数据量给数据集的制作带来了挑战。除此之外还需要解决好商品之间的遮挡问题和细粒度识别问题和商品的自由角度识别。以上的三个问题在业界均是相当具有挑战性的难题。对于遮挡问题，可以先让模型识别表层的商品进行分类。介于该项目的特殊性，机器人会对模型的工作对象进行交互（例如：抓取等动作），这表示随着机器人抓取动作的进行，商品间的遮挡问题难度可能会降低，因为货篮中的商品在随着抓取行为的进行而减少。对于细粒度识别问题 ，需要对模型的结构采取更改，选取合适的头部和颈部以能适应我们的商品定位工作。对于第三个问题，由于商品是随机的，杂乱的堆积在货篮之中，那么商品不一定和模型中的提供的锚框紧密框选（对象和锚框可能存在极大的空隙），这会导致商品的识别性能极大的降低。可以也通过选用合适的模型，采用无锚框识别的网络结构，可以提高模型性能。另外一个值得注意的问题是，在机器人处理的商品中，可能会出现弱纹理的检测对象，弱纹理表面会造成特征提取困难的问题，同时也对物品的中心点定位提出了要求。从多尺度预测的角度上考虑，弱纹理物体可能在特定的预测尺度上仍然难以被检测到。在某些尺度上，弱纹理物体的特征可能仍然不足以被模型识别。

### 3.2.2高精度的物体实时位姿估计及位姿跟踪技术（林伟鸿）

机器人抓取需要3D相机告诉机器人当前的物体位置，并由算法给出物体的抓取位置，目前常用的解决方法一方面是计算物体的6-Dof位姿，并根据抓取物体单独设定抓取点，完成一次单独的抓取动作。但 由于本项目涉及的商品数量多且商品种类未来可能增加，单纯依靠人力设定抓取点不具备很好的泛化性。另一方面是直接根据获取的点云针对单个点云单独生成抓取估计，在限定条件的情况下，算法可以快速给出末端执行器的抓取点位估计，具有良好的泛化性。但是在本项目中，抓取后需要考虑的一大问题是如何将商品稳定放置到指定地点。综合以上两种方法，结合两种网络是一个可预见的解决方案，在预测抓取点的同时，预测物体的当前位姿并进行位姿跟踪。根据项目需要，需要一个兼顾实时性和稳定性的方法，如何在复杂物体中，有效利用有效利用高精度的物体实时位姿估计及位姿跟踪技术对于机器人抓取来说至关重要。这不仅涉及到对物体的精确识别和定位，还需要实时更新物体的位姿信息。考虑到这些问题，采用相对轻量级的网络来保证位姿的估计方法的实时性，融合物体的先验mask、深度等多种信息，构造形状先验网络，可有效克服物体种类多的影响，使物体位姿跟踪更具有鲁棒性。

### 3.2.3常见物体的实时抓取估计技术（施翰煜）

在确定了待抓取物品的种类与形态后，机器人需要对物体的可抓取点进行预估以及预测出一个可行的，无碰撞的抓取姿态。在杂乱的环境中一次成功抓取需要多层次的场景理解：首先，机器人需要分析单个物体的几何属性以找到可行的抓取。 这些抓取需要符合局部对象的几何形状。其次，对于每个提出的抓取，机器人需要推理与场景中其他对象的交互。最后，机器人必须计算无碰撞抓取轨迹，同时考虑目标物体的几何形状。大多数抓取检测算法直接以整体方式预测抓取姿势，这不会捕获环境的可组合性。由于本项目涉及的商品数量较多，商品形态各异，摆放位置也存在多种的可能性，因此以目标为中心实行抓取的这一做法显得格外重要。目前常用的解决方法是由René Zurbrügg等人提出的ICGNet。这是一种在杂乱环境中用于目标驱动抓取的统一架构，允许从单视点云进行以目标为中心的抓取和对象重建。相较于Breyer等人提出的基于单一视点的点云观测，该方法无需依靠外部的分割模块与增加模型的复杂的，并结合了全景分割、重建和抓取检测，捕获到了环境的可组合性，可以更好的适应本项目的复杂抓取环境。

### 3.2.4复杂环境平台作业路径规划（黄泊凯）

中央调度系统在接受来自货物管理系统的移动运输指令后，首先需要解算出起始结束位置的坐标，并基于两点进行端到端的路径规划并下发至离起始位置最近的闲置作业平台或者在某一作业平台完成上一指令任务后下发。

但是在具体的规划过程中，需要考虑两个关键点：一是在作业区域内，必定会同时存在多台作业机械，调度系统需要实现多点对多点的协同调度规划，同时需要避免平台之间行驶路径上的阻碍和冲突碰撞；二是在规划过程中，起始位置处不可能刚好有作业平台且处于闲置状态，因此需要结合全作业区域所有平台位置及其工作状态（是否闲置）来规划生成行驶路径。

因此在规划过程中需要一个任务队列储存任务指令，一个表存储移动平台工作状态，二者需要实现一个顺序匹配，来保证每一个任务都被执行，每一作业平台不会出现闲置浪费资源。同时结合传统路径规划的图论算法，例如D3QN PER算法，并通过深度学习调整收敛相关参数，来实现多对多规划情况下，每条路径以及全局整体尽可能优的调度效率，进而保证平台搭载的机器人高效率作业。

### 3.2.5平台移动避障及姿态控制（黄泊凯）

平台启动后，特别在作业过程中，需要通过边缘硬件控制系统并借助平台所搭载的传感器来实现对平台运动姿态平衡稳定的控制和稳定运行。

其中的关键要点有三：一是平台作业过程中运动姿态的稳定；二是在作业过程中与物体抓取识别系统进行联动，调整平台姿态以辅助机械爪盘更好的操作；三是通过平台搭载的传感器实时检测平台周遭环境（尤其是是行驶路线区域的前方），在出现意外阻碍时，及时更新系统地图信息并通过系统进行告警，随后通过边缘硬件控制系统实施避障。

上述三个关键点的共同之处是，边缘硬件控制系统需要与传感器乃至其他系统模块配合联动。这就对边缘硬件控制系统的数据处理分析能力，即算力资源提出了相当的要求，因此作为边缘硬件控制系统核心的边缘计算工作站其中的芯片需要具备相当的性能和专门针对端侧部署的算法及模型。

# 拟采取的研究解决方案

* 1. **总体技术架构（林伟鸿）**

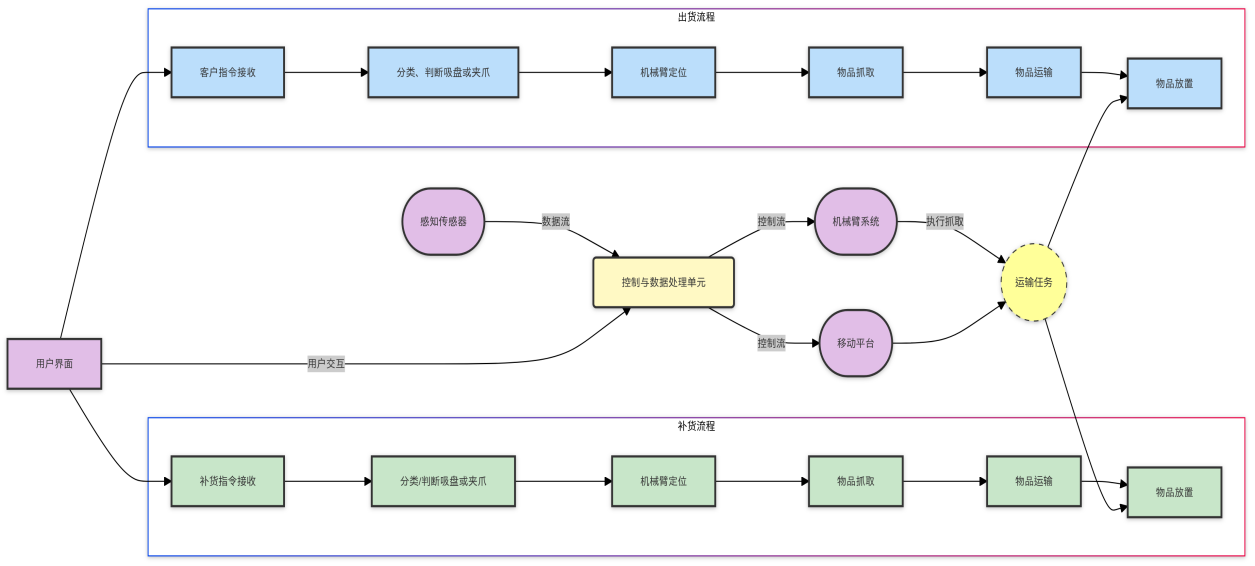


图4-1：整体架构图

### 4.1.1实时场景下物体目标检测识别技术（周俊桦）

拟采用Yolo-World结合Efficient-SAM对分类物体进行更细一步的实例分割，帮助机器人更好地对检测对象进行定位。

图4-2显示了Yolo-World相对目前的SOTA模型Clip的横向对比，在LVIS（LVIS：全称是“Large Vocabulary Instance Segmentation”，是一个用于实例分割的大词汇数据集。它包含了超过1000个类别，每个类别都有大量的实例。LVIS的目的是推动计算机视觉算法在处理开放世界场景（open-world scenarios）时的性能，这里的开放世界场景指的是测试时算法可能会遇到训练时未见过的类别）数据集中，达到相似的mAP时，YOLO-WORLD展现出了更高的FPS，这意味着YOLO-WORLD可以处理实时条件下的任务并且可以达到相当好的效果。

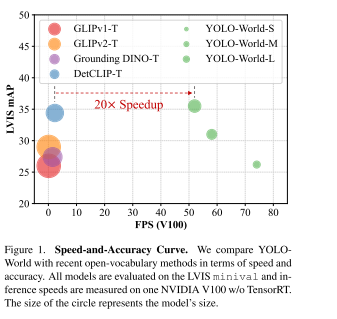


图4-2：YOLO-WORLD mAP评分图

YOLO-WORLD由一个YOLO检测器、一个文本编码器和一个可重新参数化的视觉语言路径聚合网络(repv1 - pan)组成。给定输入文本，YOLO-World中的文本编码器将文本编码为文本嵌入。YOLO检测器中的图像编码器从输入图像中提取多尺度特征。然后利用repv1 - pan通过利用图像特征和文本嵌入之间的跨模态融合来增强文本和图像的表示。

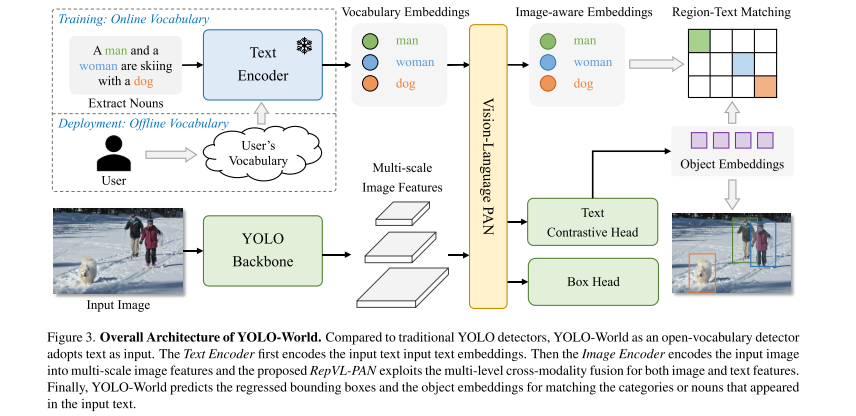


图4-3：YOLO-WORLD架构图

YOLO-World主要是基于YOLOv8开发的，所以它同样包含一个Darknet主干作为图像编码器，一个用于多尺度特征金字塔的路径聚合网络(PAN)，以及一个用于边界盒回归和对象嵌入的头部。

给定文本T，该模型采用CLIP预训练的Transformer文本编码器提取相应的文本嵌入w=TextEncoder(T)∈RC×D其中C为名词个数，D为嵌入维数。当输入文本是标题或引用表达式时，采用n-gram算法提取名词短语，然后将其馈送到文本编码器当中。

该模型也搭建了可重新参数化的视觉语言PAN，结构如下：

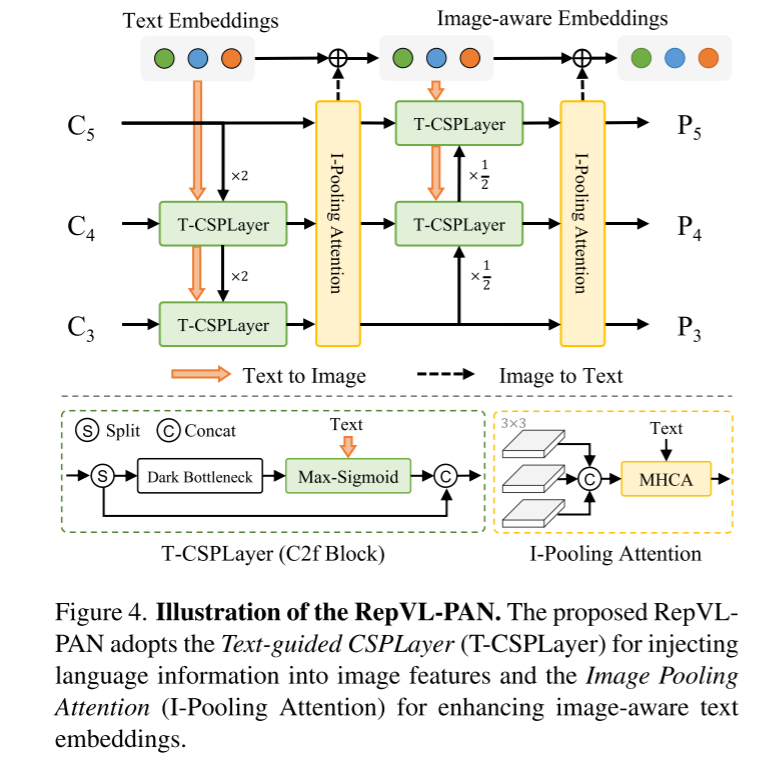


图4-4：RepVL-PAN结构图

图4-4显示了所提出的RepVL-PAN的结构，它遵循PAN中自上而下和自下而上的路径，用多尺度图像特征{C3，C4，C5}建立特征金字塔{P3，P4，P5}。此外，文本引导CSPLayer (T-CSPLayer)和图像池注意(i -池注意)进一步增强图像特征和文本特征之间的交互作用，从而提高开放词汇的视觉语义表示能力。在推理过程中，离线词汇嵌入可以被重新参数化为卷积层或线性层的权重，以供部署。

自顶向下或自底向上融合后使用跨阶段部分层(CSPLayer)。我们对CSPLayer(也称为C2f)进行扩展，将文本引导加入到多尺度图像特征中，形成文本引导CSPLayer。具体来说，给定文本嵌入W和图像特征Xl∈RH×W ×D (l∈{3,4,5})，我们采用最后一个黑暗瓶颈块之后的max-sigmoid关注，通过以下方式将文本特征聚合为图像特征:



其中更新后的Xl’与跨阶段特征拼接为输出，δ表示sigmoid函数。

为了利用图像感知信息增强文本嵌入，该模型使用了图像集中注意力，聚合图像特征来更新文本嵌入。我该模型利用多尺度特征上的最大池化来获得3×3区域，从而产生总共27个补丁令牌X˜∈R27×D。然后通过以下方式更新文本嵌入:



### 4.1.2高精度的物体实时位姿估计及位姿跟踪技术（林伟鸿）

在分类物体后，需要选定合适的末端执行器工具，吸取或夹取商品，同时对商品进行姿态估计以及跟踪，以便感知商品实时姿态以及确定后续放置姿态。

在深度学习时代之前，许多基于手工制作特征的方法，比如SIFT、FPFH、VFH和点对特征（PPF），被设计用于对象姿态估计。然而，当面对复杂场景时，这些方法在准确性和鲁棒性方面存在缺陷。这些传统方法现在已经被利用深度神经网络学习高维特征表示的数据驱动深度学习方法所取代，从而提高了准确性和对复杂环境的鲁棒性。在本项目中，需要考虑到使用轻量级的网络做到兼顾实时性和精确性。故选用FoundationPose作为基础网络，FoundationPose借助于两层网络估计物体位姿，而不是单层网络直接回归物体的位姿，大大减少了网络层参数，Bop是专门用于物体位姿估计排序的榜单，为位姿估计领域做出了极大的贡献，FoundationPose在精度上截至24年3月一直处于榜首位置。

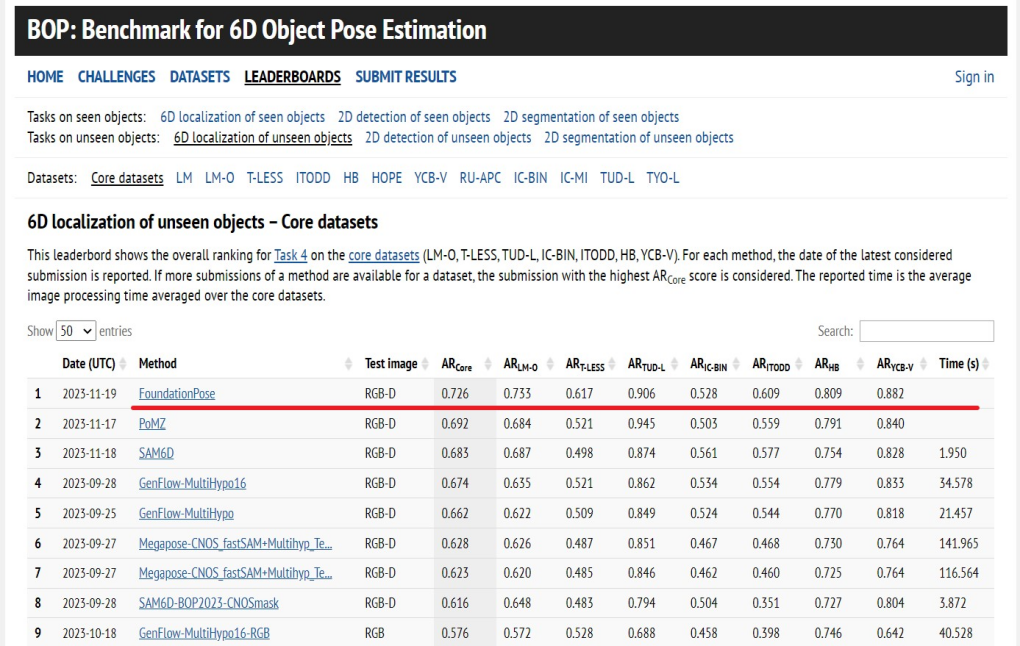


图4-5：Bop排名



图4-6：Foundation与不同算法的对比

为了使模型具有良好的泛化性，使其对未见过的新模型快速进行姿态估计，需要大量的数据集，但真实数据集在现实世界中的采集十分耗费人力与物理。

1. 基于LLM与Diffusion结合的大规模数据集合成

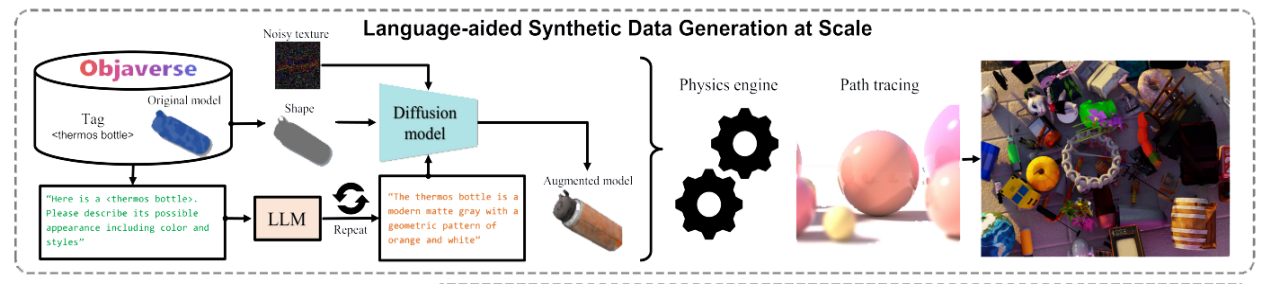


图4-7：数据集生成PipeLine

得益于LLM与Diffusion以及真实物理引擎的快速发展，仿真环境下可以采集到高质量的数据。如上图所示，首先在原有的3d模型上获取没有纹理的基础模型，并通过LLM固定句式的描述生成不同纹理描述，添加一定噪声纹理后交由Diffusion模型生成带有不同纹理的3d模型，通过这种数据增强的方式，可以迅速扩展原有的模型数量；新生成的带有纹理的模型在真实物理引擎的渲染下，可以随机生成在预设的场景环境中，并通过路径追踪取得物体的位姿。经过大量数据集训练后，模型的鲁棒性会进一步提升，并能让模型通过微调适应未来新的商品分类。

有别于直接回归生成的方法，FoundationPose采取两步位姿估计，先假定一个初始位姿，根据位姿 细化网络网络迭代更新这些初始位姿，最后根据位姿排名网络输出一个位姿分数，巧妙的设计绕过直接回归生成方法所带来的非全局最优解的问题，并且算法参数量和稳定性大大增加。

1. 粗位姿估计的生成和位姿微调

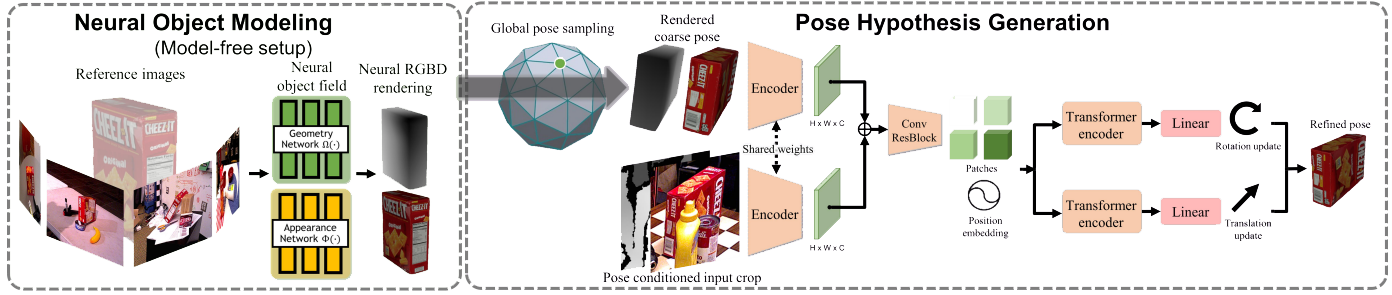


图4-8：物体粗位姿生成以及位姿细化

1. 姿态初始化

给定输入一张rgb图片，物体的二进制mask编码，物体模型，深度图，相机内参。如图3.2.4所示，预先定义Global pose sampling，以从不同视角获取渲染图片并作为初始化旋转变换R。同时根据给定mask和深度，计算场景中物体mask框内的平均深度作为初始化平移变化t。这样的粗姿态获取设计避免了类似其他算法繁杂的位姿初始化过程，预先定义好的视角可以很好适配不同情况且减少网络负担。

1. 位姿调整

姿态初始化后的姿态与实际姿态之间有很大的差距，通过创建一个姿态细化网络，将粗姿态对应的渲染图和来自给定的rgb图片的裁剪作为输入，该网络输出一个姿态更新，包括旋转和平移更新量。如图所示，先使用共享的CNN编码器从两个rgb输入分支提取特征，特征图连接之后并嵌入位置信息用以标记，通过transformer编码器模块处理后，再通过mpl层单独输出旋转更新IMG_256和平移更新IMG_257

IMG_256

IMG_257

网络使用L2监督损失 IMG_258，其中 IMG_259表示权重，IMG_260表示ground truth。

1. 姿态选择

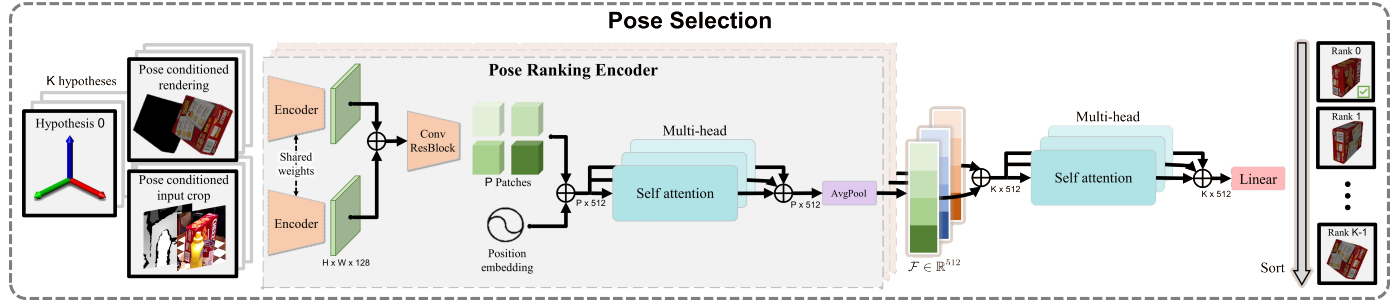


图4-9：位姿选择

完成上述姿态更新后，将获得一个细化的姿态假设列表，使用一个分层的姿态排名网络计算他们的分数，选择具有最高分数的姿态作为最终估计。这样的网络设计通过特定的迭代次数，可以稳定的从粗姿态估计到位姿调整获取一个最佳位姿。相比于单步直接回归的网络算法运行速度大大提升。

与位姿调整的网络相似，输入为与位姿调整后相对应的渲染图和给定rgb图片，使用与细化网络中相同的特征提取主干架构。所提取的特征被连接、位置标记并被转发到多头自注意模块，以便利用图像全局信息进行比较，姿态排序网络经过平均池化，输出描述渲染和给定rgb之间的对齐质量的特征向量，为了充分利用所有的位姿假设信息，再对所有k个姿态假设进行多头子注意力，最终得到有效排名得分，获取最佳位姿。

1. 姿态跟踪

相比于位姿估计，姿态跟踪不需要预先Global pose sampling，只需要计算两帧之间物体的变化，根据上述网络跟新旋转向量和平移向量。

6D位姿估计依赖于高精准的3D模型，但若暂时没有3D模型，也有一些商业或开源的方案，例如Qlone和luma ai大模型，只需要几秒的视频，就能生成逼真的三维模型。

Qlone 是一款基于增强现实 (AR) 技术的 3D 扫描应用程序，它允许用户使用智能手机或平板电脑轻松创建物体的 3D 模型。只需围绕物体移动设备，Qlone 就会自动捕捉物体的图像，并利用 AR 技术构建物体的 3D 模型。同时Qlone 提供了一些基本的模型编辑功能，例如删除不需要的部分、修复模型缺陷等。

luma ai同样可以从一系列照片/视频中自动生成 3D 模型，无需使用专业的 3D 扫描设备。通过照片采集，图像配准，深度估计，表面重建，纹理映射等步骤生成逼真的3D模型。

### 4.1.3常见物体的实时抓取估计技术（施翰煜）

为了让机器人在杂乱的环境中抓取物体，抓取过程中需要考虑物体的几何形状、与其他物体的交互，以及避免碰撞的抓取路径。当前的方法通常基于单视角的点云数据进行操作，忽略了对单个物体实例的处理，增加了抓取的复杂性。ICGNet通过实例级建模来进行抓取预测，它引入了一种端到端的架构，可以从单视角的点云数据中生成每个物体的实例表示，并进行物体重建和抓取预测。该方法的核心是提取实例级的特征嵌入，这些特征可以用于建模接触点、抓取的可行性以及物体的形状。

1. 抓取表示

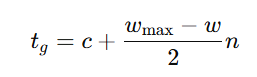
ICGNet提出了基于接触点的抓取表示，考虑了夹爪的接触点、表面法线和抓取方向。其抓取表示的公式如下：

**抓取信息表示：**

其中：

表示接触点（坐标），表示接触表面的法线向量， 表示离散抓取方向的成功概率（可行性评分）为夹爪的开口宽度，范围为：。

**抓取位置的计算：**



其中：

表示夹爪的中心点位置。

**抓取方向的计算：**

图片包含 文本

描述已自动生成

其中：

表示抓取方向的旋转矩阵；表示离散的抓取角度；为绕轴旋转​的旋转矩阵；表示重力方向的向量，为物体表面的法线方向。

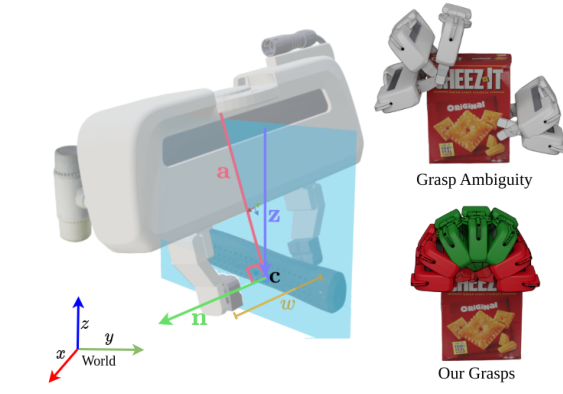


图4-10：抓取的几何表示

在上图中，左图说明了抓取时接触点、法线和抓取方向的定义，右图展示了不同抓取表示中的抓取歧义性，强调了多个抓取方向可能都有效。

1. 模型结构

**输入数据**：该方法使用单个深度摄像头捕捉的点云数据作为输入。

**编码器：**ICGNet首先通过稀疏3D U-Net架构提取体素级的稀疏和密集特征，然后通过自注意力机制对这些特征进行多次迭代精炼，以得到每个物体的实例表示。

**解码器：**使用隐式神经场预测物体占用和抓取可行性。这些预测通过多层感知器（MLP）实现，输入包括世界坐标和实例特征嵌入。

**输出数据：**

ICGNet模型的输出包括以下几个部分，它们共同实现了对场景中物体的抓取和重建：

1)实例分割标签

输出每个点的实例标签，即每个点所属的物体实例。

输出每个点的语义标签，即每个点的物体类别（如瓶子、盒子等）。

2)物体占用预测

输出每个物体的占用值，即预测每个实例的体素是否被占用。这些值用于物体的三维重建，生成每个物体的完整形状。

3)抓取姿势预测

接触点：输出夹爪的接触点。

表面法线：输出接触点处的表面法线，用于确定夹爪的方向。

抓取可行性分数：输出抓取的成功概率，表示在给定接触点和方向下，抓取成功的概率。

夹爪宽度：输出夹爪的开口宽度，用于确定夹爪的大小适应物体。

4)抓取方向预测

输出多个可能的抓取方向，每个方向对应不同的夹爪接触点和表面法线。模型还会预测抓取的旋转矩阵，用于确定夹爪的姿态。

这些输出结合起来，可以生成：

每个物体的三维重建模型；每个物体实例的抓取点、抓取方向、抓取宽度等信息，从而实现机器人在复杂场景中的抓取任务。

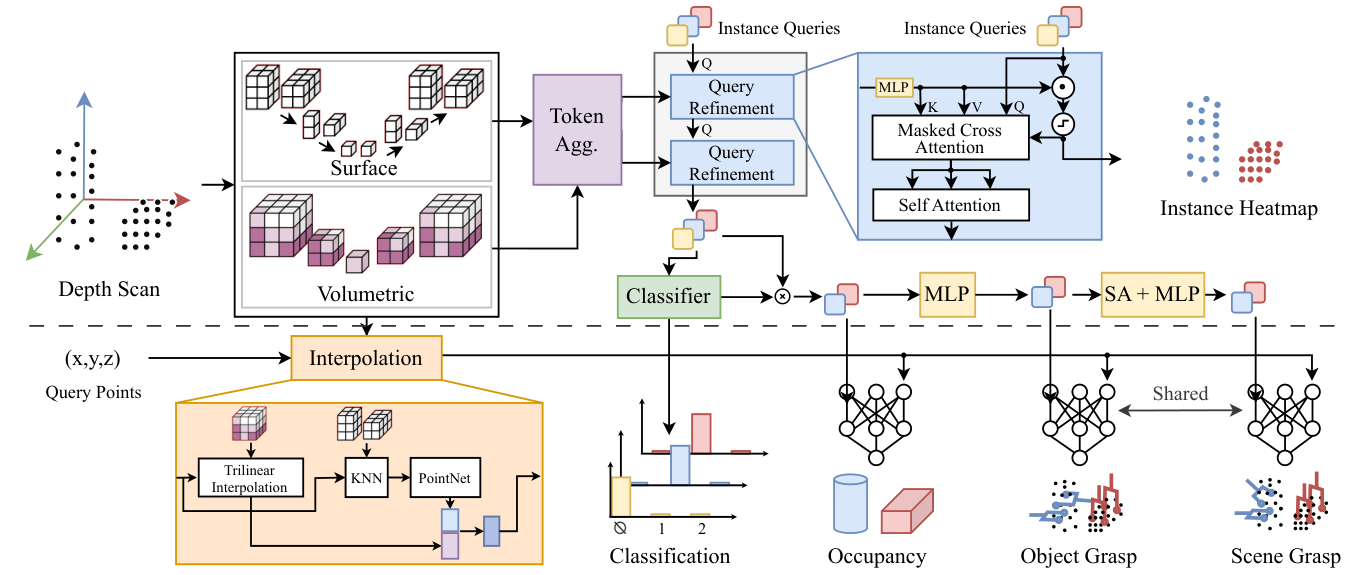
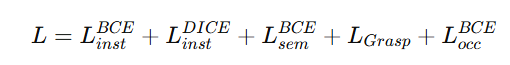


图4-11：模型结构

上图展示了ICGNet模型的整体结构。输入的点云首先经过体素化处理，提取多尺度的特征，之后通过自注意力机制生成实例查询。这些查询随后用于实例重建和抓取预测。

1. 损失函数

模型通过联合训练的方式进行优化，损失函数由多个子任务的损失组成，具体包括：



其中：

和是用于实例分割的损失；是用于语义分割的损失；是抓取预测的损失，包括多类交叉熵损失和夹爪宽度的L2损失；是用于占用预测的二元交叉熵损失。

ICGNet不仅可以用于传统的物体抓取任务，还支持目标驱动的抓取，即针对指定物体进行抓取（如“抓取瓶子”）。此外，它能够避免抓取后物体之间的碰撞。

在实验中，ICGNet成功地将仿真中的学习成果迁移到真实世界的机器人抓取任务上，并取得了优异的表现。为了可以更好的抓取物品，需要对物体进行重建，同时要求模型能够认识到物体的各个组成部分。ICGNet支持目标驱动的抓取的这一特性可以用于抓取瓶盖的任务中。

### 4.1.4复杂环境平台作业路径规划（黄泊凯）

中央调度系统在接受来自货物管理系统的移动运输指令后，需要结合全局所有作业平台的位置以及工作状态进行一个多端对多端的协同路径规划调度。目前在机器人以及上下游相关行业中，多智能体路径搜索(multi-agent path finding, MAPF)方法是此类场景下该问题主流的解决方案之一。而这其中，分布式多智能体路径搜索算法相对于集中式算法，缺乏全局信息且易出现死锁,扩展性较差。因此经过分析对比，采用集中式多智能体路径搜索方法。

另外，根据本项目的应用场景，可以确定的是，在路径规划过程中，作业平台在补货和出货的过程中，分别有着相对固定的起始点和终点，利用这一点将可以极大的降低规划调度难度。

针对本次项目的应用场景，我们首先提出一个多智能体路径搜索模型的Nutcracker-CBS算法框架，其中包含有两个模块：一是目标点分配模块；一是路径构建模块。

图4-12：Nutcracker-CBS 图示

描述已自动生成算法框架

1. 目标点分配模块

传统目标点分配采用匈牙利算法( Hungarian algorithm),通过矩阵变换得到最佳一对一分配,由于无法随 Nutcracker-CBS 主循环增量式计算,当机器人数量的增长时,计算时耗显著增加,因此不适用于 NutcrackerCBS 框架。为此,本章首先构建软约束形式的一对一目标点分配模型,然后基于改进星鸦优化实现增量式求解。

a. 改进星鸦优化算法

Nutcracker-CBS算法中的目标点分配模块主要采用了改进的星鸦优化算法（nutcracker optimization algorithm, NOA）以实现增量式地高效求解软约束构型的目标点分配问题,减少计算开销，降低对边缘处理器的计算资源性能要求。当中的改进主要针对 NOA算法 的觅食与存储,找回与恢复两组策略。公式如下图，

文本, 信件

描述已自动生成

文本, 信件

描述已自动生成

文本, 信件

描述已自动生成

文本, 信件

描述已自动生成

b.模块算法流程

模块由结点,二叉搜索树和搜索森林三个层面构成。结点代表一种多智能体路径规划方案,主要记录了其受到的约束(constraints),各机器人路径( Ps),目标点分配( Ta),各机器人路径代价总和( Sc),以及机器人规划顺序(agentorder)。每棵二叉搜索树代表一种目标点分配方案,由结点分裂扩展而成。算法在分裂结点时,总是继承其根结点的目标点分配方案,并调整约束和路径,从而求解在该目标点分配方案下的无碰撞路径。搜索森林包括多棵搜索树,以此存储不同的目标点分配方案,并对不同目标点分配方案下的所有已发现但未被处理的结点统一排序,实现目标点分配和路径规划的联合寻优。

另外，如图中伪代码流程所示，在算法执行过程中，根据当前结点中的路径长度动态更新邻接矩阵，从而得到各机器人与不同目标点之间距离的准确值，并据此重构目标点分配方案，以保证最终输出结果尽可能优。

文本

低可信度文本

低可信度描述已自动生成图4-13：紧耦合目标点分配伪代码流程

（2）路径构建模块

经典 CBS 算法存在约束构建缺乏系统性,路径构建无法估计避碰代价,不同智能体路径构建互不考虑的问题,在狭窄空间中方案构建效率低下,路径冗长。分别构建回退式约束构建机制,引入避碰路径估计的绕道机制和数据共享底层路径规划机制,提升搜索效率和质量。

一些文字和图案

中度可信度描述已自动生成

图4-14：冲突分类

在路径构建过程中，冲突分类处理有助于 MAPF 算法的分析和改进。我们将路径构建中的不同路径之间的冲突分为关键冲突,半关键冲突和非关键冲突,并以此作为分组依据,分解原 MAPF 问题来分析解决。

1. 路径回退式约束构建机制

关键冲突会产生大量的结点分裂和时耗。传统 CBS 所构建的避碰机制路径冗余,难以执行,原因在于算法总是仅针对当前冲突位置构建约束,而大多数冲突是由于之前某个位置策略不当所导致。因此若提前若干位置产生避让分支路径,一方面避免了在狭窄区域构建绕道策略的难题,另一方面增加避碰分支路径与冲突位置的距离,便于底层算法得到符合车辆运动学约束的路径。因此这里提出一种新型约束构建策略,沿车辆路径回退若干路径点构建约束,提升 CBS 算法避碰策略的搜索效率。

1. 避碰路径估计的绕道机制

半关键冲突虽然存在绕道路线,但传统 CBS 仅通过 碰撞位置处的约束强制下层路径规划避让某一位置,导 致底层算法无法正确理解约束并从根本上构造无碰撞路 径。 如果在下层路径搜索即将到达碰撞位置时估计出由 于避碰带来的代价值并及时调整路径,能够带来更高的 求解效率和更短的路径总和。

1. 数据共享底层路径规划机制

在底层路径搜索中也考虑到其余智能体的影响,可 以显著减轻上层模块的负担,提高效率。 为此,本节在底 层路径构建算法中加入无人车之间的距离惩罚项,在栅 格点选择时降低多机器人冲突的概率,从而尽可能构建 出无碰撞路径

手机屏幕截图

中度可信度描述已自动生成

### 4.1.5平台移动避障及姿态控制（黄泊凯）

边缘硬件控制系统在收到来自中央调度系统下发的路径后，控制硬件执行路径指令。同时整体基于SLAM（Simultaneous Localization and Mapping）技术，在多传感器多模态输入情况下，结合深度学习优化来实现平台作业过程中运动姿态的稳定以及调整平台姿态以辅助机械爪盘更好的操作和出现意外阻碍时的避障操作。



图4-15：SLAM系统整体结构

1. 平台作业过程中运动姿态平衡稳定

该部分主要关注移动过程中和运动速度方向变化时的平台稳定性。在移动过程中保证平台稳定性的关键在于对机械臂姿态的调整以尽可能降低重心；在平台运动过程中速度方向变化时（如转弯，停止，启动等）保证平台稳定性的关键在于对驱动电机的控制并配合传动装置来实现速度和转向的平稳改变，而不至于因速度或者力矩的突变造成的颠簸导致物品移位甚至脱落。

考虑到本项目主要运用场景是轻小物品抓取运输，物品重量相对机械臂和移动平台可以忽略不计，即对移动平台整体重心影响极小，认为运动过程中对机械臂姿态的调整在本次项目初期暂时不需要着重考虑。如果后期应用场景拓展到更广的范围，尤其是抓取运输物品质量较大情况下，再增加对机械臂姿态调整的控制组件算法，借助内置传感器监测结果并经电控组件测算移动平台整体重心，若其偏移值超出预设阈值，则通过PID控制算法来调整机械臂姿态和平衡重心。

实现速度和转向的平稳改变，主要依靠移动平台的机载电控组件以及内置传感器实时检测实现。在平台接近路线转向点或者起终点附近时，通过PID控制算法等来控制编码电机输出马力的渐变以实现平台速度和力矩的平稳渐变。

1. 调整平台姿态联动辅助机械爪盘操作

当到达货架指定位置，机械臂放置商品不可避免要考虑到碰撞检测。在机械臂复杂关节操作下，很难做到像人一般灵活的避开障碍并伸到货架深处拿去商品。

如下图，通过在货架附近微调平台位置，同时与机械臂联动调整姿态，使得机械臂能够以一个有利的位置抓取放置物品。在联动调整过程中，边缘硬件控制系统通过PID算法等，实现调整过程中平台整体姿态的稳定，以防止出现机械碰撞等意外。



图4-16：平台姿态联动调整示意



图4-17：货架示意

c. 行驶避障

基于SLAM技术，同时结合人工输入校正，对作业区域整体进行建图，构建2D格栅地图。在系统作业过程中基于激光雷达、视觉、惯性测量单元IMU等多传感器多模态输入并结合深度学习（例如视觉方面的DeepVO）实时更新移动平台位置，并维护平面地图，尤其是对其中可行驶区域内的障碍物进行标记。

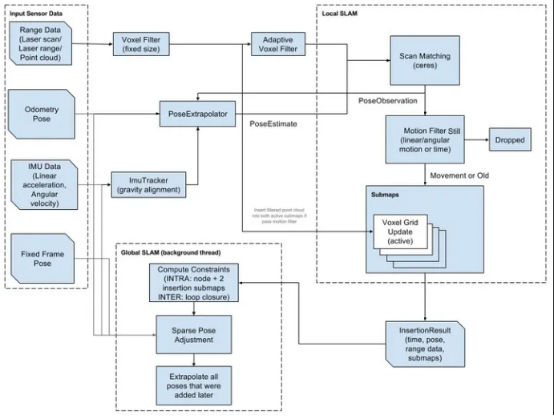


图4-18：SLAM结构及关系如下

在接收到来自控制终端的指令后，该模块以此时移动平台所在位置坐标为起点，指令指定位置坐标为终点。

在移动过程中，为了防止因不可控因素（比如因人为操作失误导致货物掉落在地或者是其他作业平台抓取装置异常货物脱落等）造成的行驶区域出现障碍物导致阻塞。通过移动平台上装载的传感器（激光雷达、单双/目摄像头等）对平台行驶路线前方乃至周围区域进行探测并实时反馈结果。

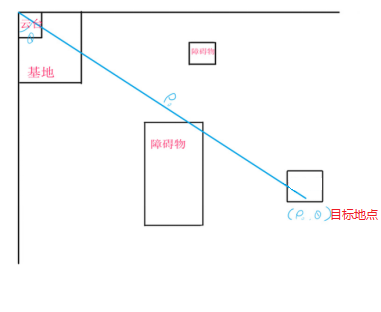


图4-19：避障示意

一旦探测到异常（障碍物），通过SLAM技术，借助平台位置的坐标更新地图信息，标定障碍物坐标并通过系统向操作人员告警。

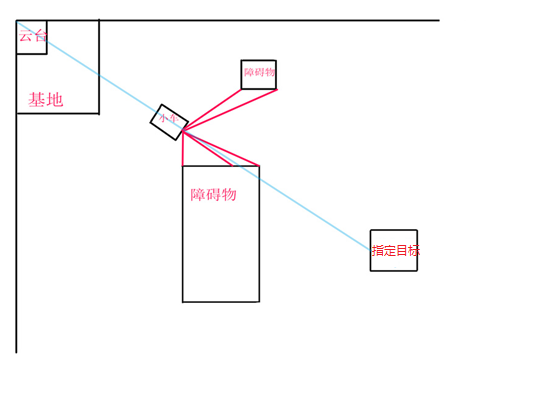


图4-20：避障示意

接着由移动平台机载电控组件触发中断，借助机载传感器并通过轨迹控制，向垂直于原行驶路线前进矢量方向的方向移动，直至前方区域的传感器探测结果为无阻碍。

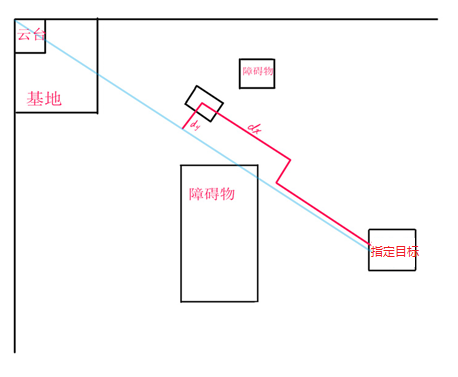


图4-21：避障示意

随后恢复原路线规划方案继续移动。

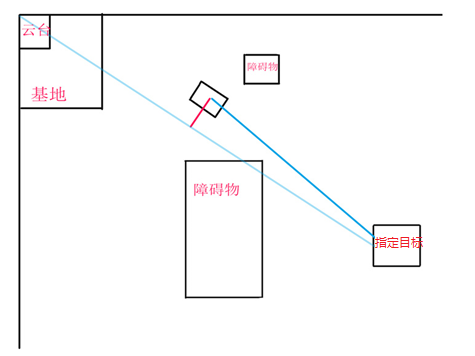


图4-22：避障示意

# 预期成果与工作计划（林伟鸿）

* 1. **预期成果**

开展机器人超市具身智能技术研发，针对超市场景，在大量商品的补货取货流程下，实现鲁棒高精度的商品目标检测与实例分割、6-DoF位姿估计、基于RGB-D商品抓取点估计以及机器人作业实时动态规划控制，预期成果与指标具体如下：

1. 目标检测识别准确率：针对典型商品，在实验室模拟测试场景下达到93%以上；
2. 目标6-DoF位姿检测精度：实验室模拟测试场景下位姿估计的平均距离（ADD）不超过2cm；其中：平均距离（ADD）指计算真值姿势和估计姿势转换的两个3D模型中对应点之间距离的平均值；
3. 位姿识别的图像处理帧率：目标检测识别与位姿估计达到15FPS以上的图像处理速度，支撑机器人作业动作连贯性；
4. 机器人动态规划实时性：相对位姿变化时，机器人运动响应时间小于0.5s，实现对相对位姿变化的实时响应，并需要具备在障碍空间条件下的自主作业规划能力。
5. 知识产权：发明专利申请5项；软件著作权3项。其中：
   1. 在商品目标检测识别方面，完成发明专利申请1项，软件著作权1项；
   2. 在商品位姿估计与抓取点估计方面，完成发明专利申请2项，软件著作权1项；
   3. 在机器人作业实时动态规划控制方面，完成发明专利申请2项，软件著作权1项。
6. 联合人才培养：项目实施期内，联合培养2-3名从事机器人视觉智能、自主规划技术研发的硕士研究生。
   1. **项目研究计划**

本项目研发工作计划需1年时间，具体工作计划如下：

|  |  |
| --- | --- |
| 时间 | 实施阶段 |
| 2024.11~2025.01 | 1. 项目详细技术方案制定； 2. 试验环境搭建； 3. 机器人视觉检测技术研发； 4. 数据集收集制作； 5. 6-DoF位姿估计等算法研发与仿真模拟验证； 6. 实时运动规划算法研发与仿真模拟验证； |
| 2025.02~2025.05 | 1. 机器人视觉检测、位姿估计与实时运动规划等算法的模拟场景应用测试验证与算法集成测试 2. 完成发明专利与软件著作权申请。 |
| 2025.06~2025.10 | 1. 进行电网作业机器人视觉检测与运动规划技术的实际应用验证； 2. 进一步完善算法，完成项目技术成果交付。 3. 完成项目验收。 |