技术架构

## **1. 技术方案**

方案介绍

1. 项目概述：

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：
2. 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。
3. 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。

核心内容

针对整体方案，通过整个流程模块进行分解，大致可以分为一下几个模块：

1. 多自由度双臂机械臂系统：具有6自由度，末端执行器为二指夹爪和吸盘，配有至少一个相机系统用于感知环境，并且可以承重日常生活中的常见物体以及商品组合后提交给顾客的购物篮。
2. 机器人移动平台：安装有激光雷达的移动平台，用于机械臂移动过程中的路径规划
3. 控制与数据处理平台：具有较高的数据处理能力的平台，用于整体机器人系统的运算推理。

## **技术架构**

思维导图



## **核心流程**

### 根据指令进行物体分类与实例分割：（周俊桦）

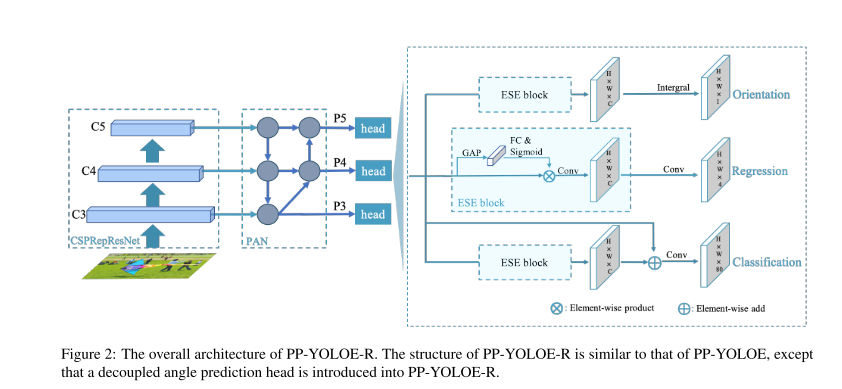
（1）简要介绍

该流程所要实现的任务是能够对货篮中的商品进行识别，分类，从而帮助机械臂实现商品的定位，使机械臂完成特定商品的选取工作。通过该模型，机械臂可以在取货前对货篮拍摄一张图片，在经过处理后，模型能为机械臂提供机械臂当前位置的表面商品的相对位置。

选取PP-YOLOE-R进行物体的分类工作，PP-YOLO提出了种基于PP-YOLOE高效无锚框旋转目标检测器PP-YOLOE-R。在本次项目中，机械臂需要处理的场景大多是自由角度的商品，即商品相对于摄像头画幅边框的角度可能并不是平行或者垂直的。这时选用传统的YOLO模型可能会导致商品识别性能下降的问题。

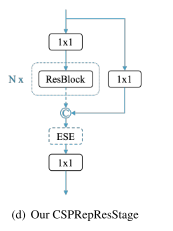
（2）PP-YOLORE-R工作细节

1）模型架构：PP-YOLOE模型的主干（BackBone）是CSPRepResNet，其颈部（Neck）是路径聚合网络（PAN），头部（Head）是高效任务对齐头部。在YOLO网络中BackBone负责对输入图像进行特征的提取，Neck负责对特征图进行多尺度特征融合，并把这些特征传递给预测层，Head进行最终的回归预测。



2）模型工作流程

1.当一张照片进入模型时，会先经过CSPRepRESNet进行特征图的生成。这个网络是一个由三个卷积层和四个后续阶段组成的主干，由RepResBlock堆叠而成。RepResBlock结构如图所示。在每个阶段，使用跨阶段部分连接来避免大量3 × 3卷积层带来的大量PA参数和计算负担。



这能有效地减轻机械臂的计算压力，使其处理任务时能更加从容。

1. 位于颈部的PANet会对在主干生成的特征图进行特征融合，PANet引入了自底向上的路径，使得底层信息更容易传递到高层顶部。最终将特征融合图传递到头部。
2. 在目标检测中，分类和定位之间的任务冲突是一个众所周知的问题。在头部中，模型会对接受的特征融合图进行回归预测。PP-YOLOE-R采用了角度预测头，这是由一个通道注意层和一个卷积层组成的，非常轻巧。该头部采用DFL预测角度。不同于ln-norm学习狄拉克δ分布，DFL旨在学习角度的一般分布。具体来说，以偶区间ω离散角度，得到以积分形式表示的预测θ，可以表示为:

descript

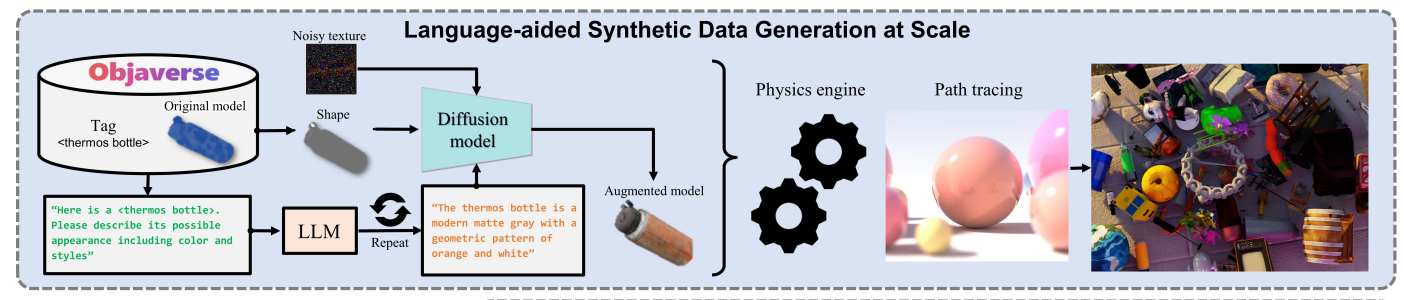
其中pi表示角度在每个区间内下降的概率。在就角度预测头中，ESE block会对特征融合图进行池化（GAP）和在特征函数层的处理，最终实现定位，回归，分类的工作

### 物体6D位姿估计与位姿跟踪：（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。

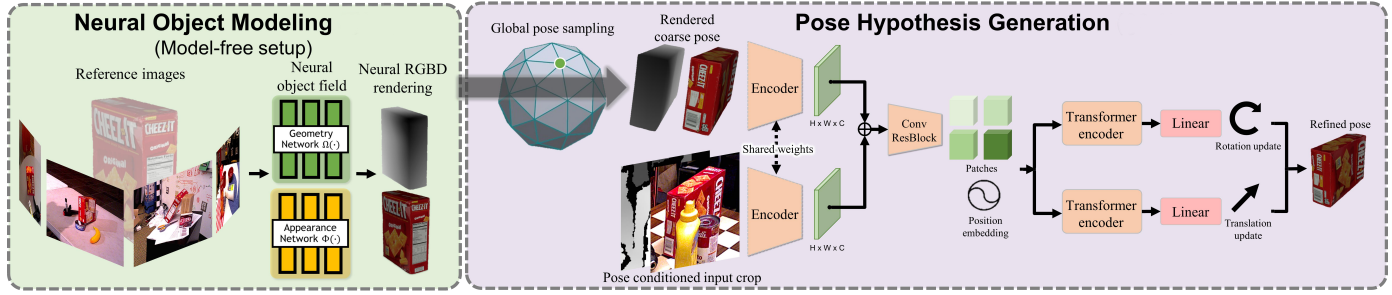
在深度学习时代之前，许多基于手工制作特征的方法，比如SIFT、FPFH、VFH和点对特征（PPF），被设计用于对象姿态估计。然而，当面对复杂场景时，这些方法在准确性和鲁棒性方面存在缺陷。这些传统方法现在已经被利用深度神经网络学习高维特征表示的数据驱动深度学习方法所取代，从而提高了准确性和对复杂环境的鲁棒性。

1. 基于LLM与Diffusion结合的大规模数据集合成



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，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。

1. 粗位姿估计的生成和位姿微调



* 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。如上图所示，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。

* 1. 位姿调整

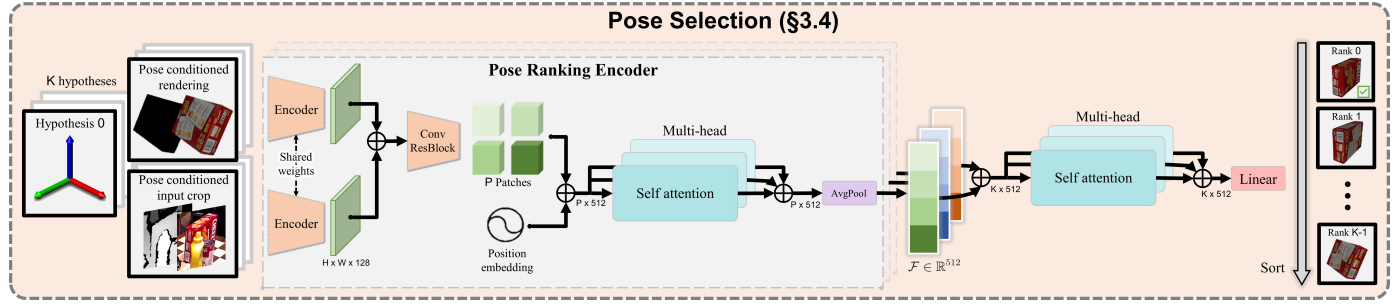
姿态初始化后的姿态与实际姿态之间有很大的差距，通过创建一个姿态细化网络，将粗姿态对应的渲染图和来自给定的rgb图片的裁剪作为输入，该网络输出一个姿态更新，包括旋转和平移更新量。如图所示，先使用共享的CNN编码器从两个rgb输入分支提取特征，特征图连接之后并嵌入位置信息用以标记，通过transformer编码器模块处理后，再通过mpl层单独输出旋转更新descript和平移更新descript

descript

descript

网络使用L2监督损失 descript，其中 descript表示权重，descript表示ground truth。

1. 姿态选择



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。

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，以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。

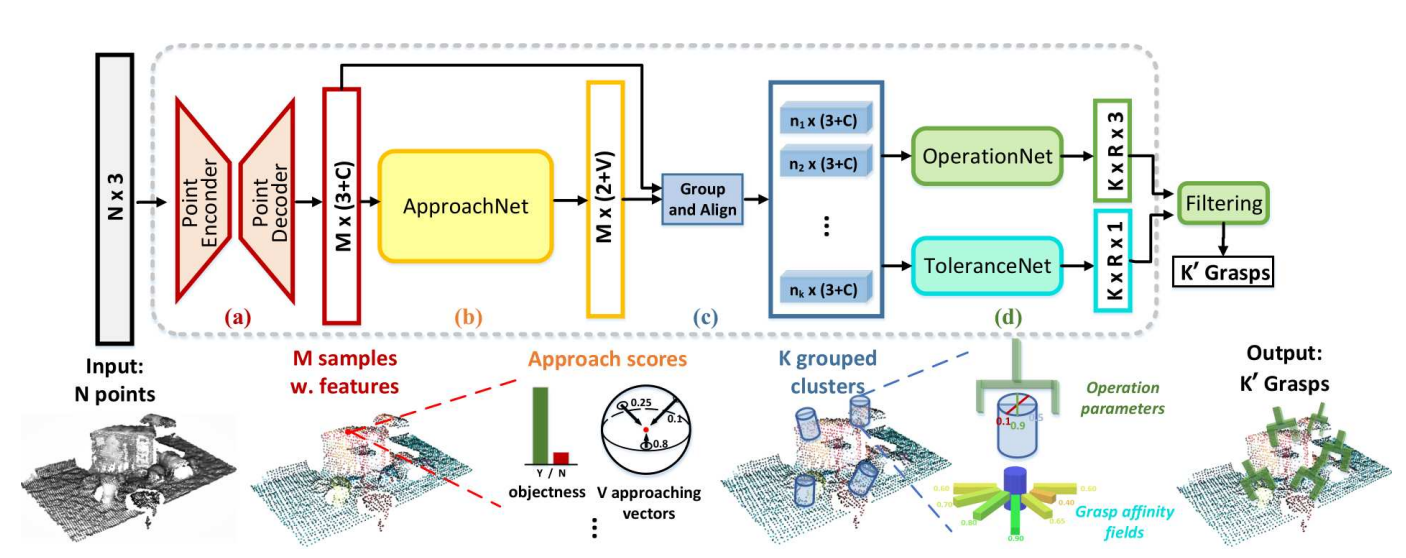
luma ai同样可以从一系列照片/视频中自动生成 3D 模型，无需使用专业的 3D 扫描设备。通过照片采集，图像配准，深度估计，表面重建，纹理映射等步骤生成逼真的3D模型。

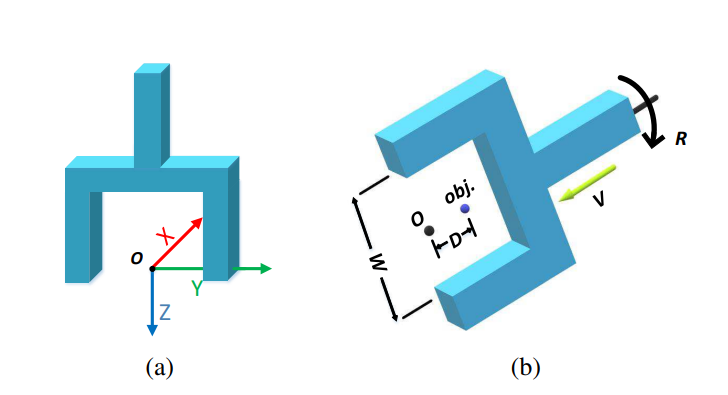
### 机械臂二指抓取点估计与吸盘吸取点位估计：（施翰煜）

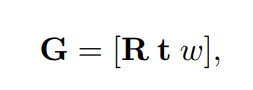
机械臂二指夹爪和吸盘点位估计是两种不同的网络，前者通常根据物体的几何形状、点云信息，计算每个点云的抓取力闭合可能性的方向，并考虑物理学重心等约束条件，根据夹爪最大量程生成抓取估计；而吸盘点估计，通常选择物体重心于稳定面的垂直方向作为吸取点，吸盘头为柔性可伸缩，所以针对不同物体设置不同的伸缩值，仅依赖6D位姿估计，从上往下吸取，近似退化成平面吸取即可。

Graspnet是一个开源稳定的二指夹爪网络，基于这个网络有一些添加不同约束条件的新的网络，这里做简单介绍：

GraspNet：一种基于端到端3D的抓取姿势检测网络。使用PointNet++骨干网络捕获点云几何特征做为输入，接着通过点编码器-解码器提取特征并进行采样，接着使用Approach Network预测接近向量，获得接近向量后，通过Operation Network进一步预测平面内旋转、接近距离、夹具宽度和抓取置信度，此时已经可以完成抓取的操作了，为了以提高抓取姿势预测的鲁棒性，接着使用Tolerance Network预测每次抓取的扰动容忍度。



1. 抓取姿态表示。在上图a的坐标系中定义二指夹模型。上图b中的obj表示待抓取的物品，GraspNet需要预测接近向量V，物品到二指夹的距离D和二指夹宽度W。将该抓取姿态用一个向量表示：



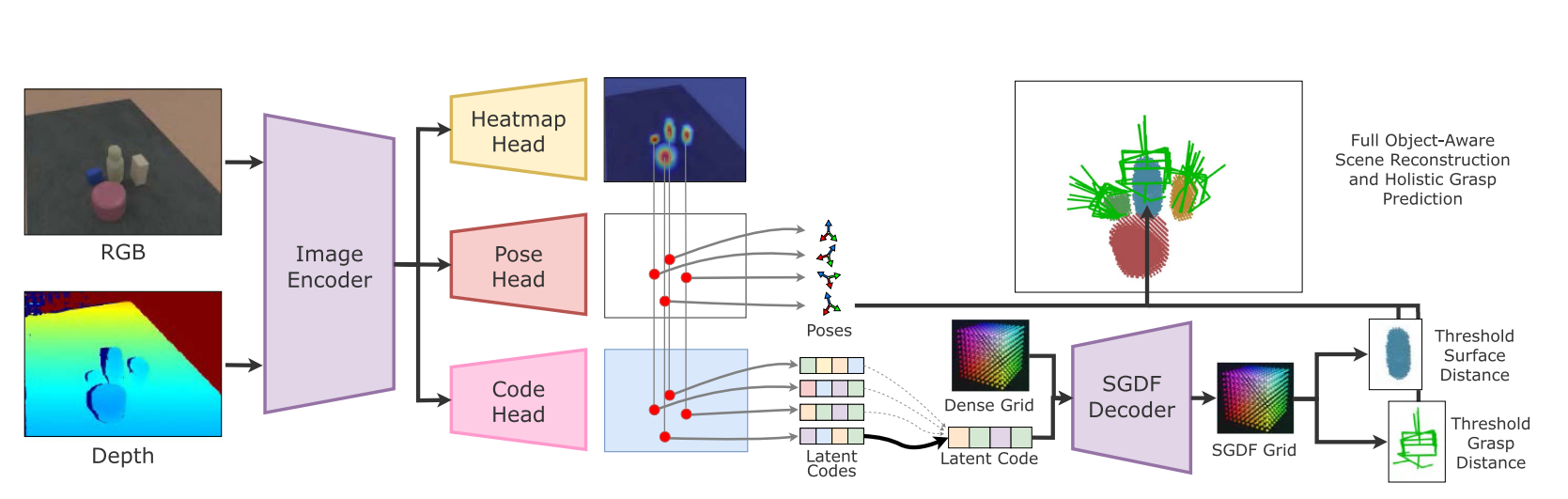
其中R表示二指夹的方向，w是二指夹的宽度，t是二指夹的中心。

2.Approach Network。由于某些方向由于遮挡而不适合抓取，因此要估计可行的接近方向和抓取点。Approach Network以PointNet++作为基础网络，该网络以大小为 N × 3 的原始点云作为输入，输出一组点的特征，并预测每个点是否可抓取的置信度以及可能的接近向量。网络的输出头会将可行的接近向量分类为 V 个预定义的视点。 同时，对于每个点，方法网络输出两个值来预测其可抓取或不可抓取的置信度。接着，对这些抓取点进行采样，得到若干个虚拟接近向量。

3.Operation Network。在Approach Network中获取到可抓取点获得接近向量后，需要进一步预测平面内旋转、接近距离、夹具宽度和抓取置信度。为了更好的学习，所有采样点都转换为一个新的坐标，旋转网络将对齐的点云作为输入，并预测每个物品的分类分数和归一化残差，以及相应的抓取宽度和置信度。

4.Tolerance Network。经过前面的步骤，端到端的网络已经可以预测准确的抓取姿势。 由于可行的抓取姿势是无限的，而人类倾向于选择能够容忍更大误差的抓取姿势。Tolerance Network正是用于模拟这种行为，用于预测抓取的鲁棒性，即抓取姿态的容忍度。给定一个真实的抓握姿势，Tolerance Network会在球体空间中搜索其邻居，以查看抓握仍然稳健且抓握分数 s > 0.5 的最远距离。

CenterGrasp：该方法通过RGB-D图像编码器预测物体姿态，并通过解码器重建物体的3D形状和抓取位置，能够在复杂场景中实现准确的抓取。GraspNet 使用基于点云的神经网络，而CenterGrasp 采用基于RGB-D输入的整体形状重建进行6-DoF抓取估计，CenterGrasp 集成了对象感知，能够区分不同物体并基于形状和潜在嵌入进行抓取预测，这是GraspNet 所没有的。CenterGrasp在推理时，图像编码器检测每个对象的中心点以及各自的 6D 位姿和潜在形状向量，再由建议形状和抓握距离函数 (SGDF) 解码器重建 3D 形状和有效抓取。



1.图像编码器。使用 ResNet50 拓扑生成 RGB 和深度图像的低分辨率特征表示（上图Depth），连接并将其输入ResNet18FPN网络中，最后使用生成的特征作为Heatmap Head、Pose Head和Code Head的输入。

2.Heatmap Head预测每个像素的物体概率，Pose Head将给定像素处物体的 6D 姿势预测为向量，Code Head预测该像素处物体的形状和可能的抓取。

3.形状和抓取距离函数 (SGDF) 解码器。将 3D 坐标和潜在Code映射到形状距离和抓取距离。 形状距离是一个标量值，表示样本坐标与对象表面上最近点之间的有符号距离。 抓取距离是距样本坐标最近的有效抓取的完整 6D 姿态，以与对象框架对齐并以样本坐标本身为中心的参考框架表示。

ICGNet：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。 图像被转换为 点云并输入到模型中。 该模型联合重建每个物体的完整 3D 形状，并预测不会与其他物体和环境发生碰撞的稳定抓取。

1.Encoder。给定世界框架中的点云，Encoder利用用于表面特征的稀疏 3D-UNet 架构和用于体积特征的密集 3DUNet 架构以多种分辨率提取稀疏和密集特征，通过Token Aggregation得到丰富的稀疏特征。运用多个实例查询细化模块将体素化场景分解为单独的实例。此外，依靠分类头为每个潜在表示分配一个类，并使用无对象类过滤掉不匹配的查询。

2.Decoder。给定现三维坐标，使用一系列具有残差连接的 MLP预测每个实例的占用率和抓取的可靠性。通过插值模块合并表面和体积特征，丰富了坐标的信息。

3.Query Refinement。给定一组 K 个实例，应用一系列屏蔽的交叉注意力和自注意力来提取以实例为中心的信息，并给出从U-Net主干提取的场景级特征。 采用 Mask3D并添加MLP层来进一步处理提取的场景特征。 使用基于体素位置的傅里叶位置编码，并使用最远点采样来对初始实例查询位置进行采样。

4.插值模块。由于场景信息有限，直接将输入坐标或相应的位置编码传递给占用网络会产生次优的重建和掌握预测。因此采用小型 PointNet 进行特征聚合，该特征聚合依赖于 KNN 特征和到最近邻居的距离。密集和稀疏特征都通过 MLP 连接并馈送，以提取每个坐标的丰富嵌入。

### 机械臂导航与运动避障：（黄泊凯）

最后根据当前6D位姿跟踪与预设放置位姿规划机械臂到指定位置并松开末端夹持器完成一次物品运输

## **详细设计**

## **方案执行**

整体架构图

