分类号：TP242.6　　单位代码：10363
密　级：公开　　学　　号：2210920115

题　　目 基于零样本学习的未知物体抓取关键技术研究

题　　目 RESEARCH ON KEY TECHNIQUES OF GRASPING UNSEEN OBJECTS BASED ON ZERO-SHOT LEARNING

学　生　姓 名：张靖

校　内　教　师：曹雏清

校　外　教　师：刘志恒

专业 学位 类别：计算机技术

研究方向(领域)：人工智能及应用

论文答辩日期：2024.5.24

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：张靖

日期：2024年5月23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：张靖　　指导教师签名：李新涛

日期：2024年 5月 23日　　日期：2024年 5月 23日

基于零样本学习的未知物体抓取关键技术研究

摘 要

随着近些年机器人行业的快速发展，越来越多的学者关注到了机器人抓取领域，尤其是对于未知物体的检测及抓取，成为了机器人智能抓取领域中的热门话题。在未知物体的抓取问题中，检测未知物体的位置以及姿态都非常重要。目前，面对单一已知物体执行抓取作业的研究已经有了很大的突破且在实际中得到了应用，但在真实环境中，物体模型的未知使得未知物体的抓取仍有极大的挑战性。

因此，本文基于零样本学习理论研究机器人检测未知物体的方法，并且为了高效地完成机器人抓取任务，还对未知物体的位姿估计进行了研究，主要研究内容包括：

第一，基于零样本学习模型对未知物体进行检测，将视觉特征和语义特征联系起来，将图像和文本映射到同一表示空间，并通过对比不同图像和文本对之间的相似性和差异性进行训练，从而学习到具有良好泛化能力的特征表示方法用于提高零样本学习模型对未知物体的检测能力；第二，通过局部特征匹配获取足够多的目标图像对的匹配点对，读取深度信息后将其作为 ICP 的初始点云进行配准，以得到目标的 6D 姿态；第三，结合上述两者设计了一种用于未知物体的抓取网络并进行实验验证。实验证明，该抓取网络在实际场景中抓取系统的平均检测准确率为 90%，综合抓取成功率为 79%。

关键词：零样本学习，未知物体，机器人抓取，姿态估计，点云配准

RESEARCH ON KEY TECHNIQUES OF GRASPING UNSEEN OBJECTS BASED ON ZERO-SHOT LEARNING

ABSTRACT

With the rapid development of the robot industry in recent years, more and more scholars have paid attention to the field of robot grasping, especially the detection and grasping of unknown objects, which has become a hot topic in the field of robot intelligent grasping. In the problem of unknown object grasping, detecting the position and pose of unknown objects is very important. At present, the research on grasping a single known object has made a great breakthrough and has been applied in practice. However, in real environments, the unknown object model still makes the grasping of unknown objects a great challenge.

Therefore, this paper studies the method of robot detection of unknown objects based on zero-shot learning theory, and in order to complete the robot grasping task efficiently, the pose estimation of unknown objects is also studied. The main research contents include:

Firstly, unknown objects are detected based on zero-shot learning model, visual features and semantic features are linked, image and text are mapped to the same representation space, and the similarity and difference between different image and text pairs are compared for training. The feature representation with good generalization ability is learned to improve the detection ability of the zero-shot learning model for unknown objects. Secondly, enough matching point pairs of the target image pair were obtained by local feature matching, and the depth information was read as the initial point cloud of ICP for registration to obtain the 6D pose of the target. Thirdly, a grasping network for

unknown objects is designed and verified by experiments. Experiments show that the average detection accuracy of the grasping system in the actual scene is 90%, and the comprehensive grasping success rate is 79%.

KEY WORDS: zero-shot learning, unknown object, robot grasping, pose estimation, point cloud registration

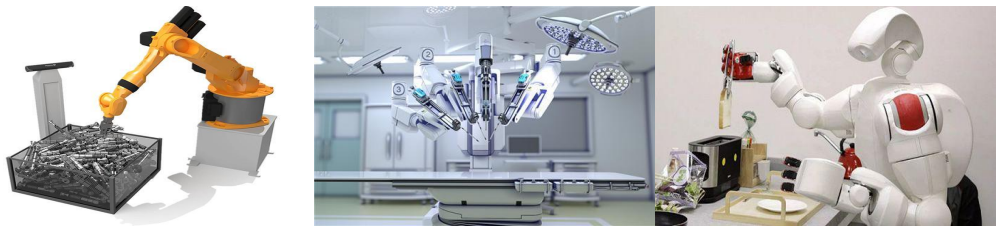
目 录

第 1 章 绪论

1.1 研究背景和意义

机器人是众多先进学科复杂技术相辅相成的自动化设备，在医疗、工业制造、运输、人工智能等领域都有广泛应用。机器人自主抓取是新一代机器人的发展方向。现有的智能机器人很大程度上可以在人类的帮助下进行相当复杂的抓取工作，如工业制造、家庭服务、医疗服务等。尽管智能机器人的研究已经应用广泛，但是其智能化水平依然有待提高，很多情况下，还无法进行智能交互。很多学者都在研究如何令机器人在抓取任务中也有人类的思考能力，例如，刀通常应该用握在刀柄而不是刀片，一杯加热茶的杯子最好使夹爪抓在把手上，而不是杯子的边缘。经过许多学者对机器人自主抓取的研究，机器人能够非常灵活地应用于各种环境，为机器人提供了执行抓取任务奠定了扎实的基础。

无论是在工业制造[1]、家庭服务[2-3]、医疗[4]等行业，具备自主抓取能力的机器人都展现出了广泛的应用价值。图 1-1 中是不同种类的机器人被分别部署于不同场景执行着各自抓取任务的状态。如图 1-1(a)所示，机器人末端安装吸盘作为夹持器，用于工业场景抓取随意摆放的零部件，能够代替人类进行一些简单的抓取摆放工作，使工厂生产效率得到有效提升。如图 1-1(b)所示，是应用在手术室的医疗机械臂，可以用于协助复杂的手术进行，医疗机械臂相比于人类的手臂稳定性更高、精度更高，在提升手术成功率上有着重要作用。如图 1-1(c)所示，该机器人可以完成家庭生活中很多简单的、重复的工作，不仅显著提升了人类生活的幸福感，更能够协助生活不便的老年人和残障人士完成一些原本无法胜任的工作，从而极大地增强了他们生活的便捷性。



(a) (b) (c)

图 1-1 机器人应用展示

Figure 1-1 Robot application demonstration

现有的机器人技术在绝大部分任务的执行过程中表现出较高的完成度，然而，在存在遮挡的环境中或者面对未知目标的情况下，机器人在任务执行方面的成功率仍有待加强。因此，在机器人领域中，机器人的智能抓取工作有非常大的发展空间。随着机器人在各个环境下的应用增加，抓取任务也随之多样化，使得未知物体的识别和抓取也成为了目前研究的热点。

在深度学习技术还没有推广的时候，抓取未知物体首先通过模板匹配或是人工示教得到该物体对应的抓取信息，再通过这些信息对物体抓取姿态进行规划。模板匹配指的是将大量的物体模板提前录入，在抓取时对待抓取物体与模板进行匹配，人工示教指的是使物体的类别不发生变化，对机器人进行抓取抓类别物体的抓取示教。两者及其耗费人力、时间，且无法在物体较丰富的环境下进行抓取任务。后来使用机器学习方法，首先通过建立不同的分类器对未知物体进行分类，再对该类别物体进行抓取规划，因此泛化性低，不适合用于未知物体类别较多的场景。相较于耗费大量人力资源的机器学习方法，卷积神经网络能够已有的数据集中独立地发现数据的深层特征，并且保持高效性。目前使用较广的方法是使用深度卷积神经网络算法对未知物体进行识别和检测。然而，在研究机器人对未知物体进行抓取的过程中，仍旧存在许多困难和挑战需要克服，比如，环境的复杂性使得机器人很难快速准确地识别目标的位置和形状；其次，未知物体的多样性使得机器人需要具备良好的适应性和泛化性。

了解了机器人抓取现状，为了提高机器人抓取未知物体的准确性与实时性，本文基于零样本学习的方法来研究该问题，并使用结合局部特征匹配和点云配准的姿态估计网络来提高未知物体姿态估计的泛化性，从而提高机器人自主抓取能力，推动了机器人领域的发展。

1.2 国内外研究现状和问题

1.2.1 基于视觉的物体识别研究现状

在机器人抓取物体过程中，物体识别是实现抓取任务的关键步骤之一。使用的算法要根据颜色、尺寸、周围物体等特征来检测图像中是否包含目标物体，若有，则需要将该物体所在的区域精确标记出来。由于机器视觉技术的蓬勃发展，物体识别方法已经从传统的物体识别方法慢慢变成使用神经网络的物体识别方法，融合了机器学习、计算机视觉等相关技术，在机器人研究领域中依然是一个

重要的研究方向。

本节将系统全面地阐述三种物体识别方法的研究现状，包括基于模板信息、浅层特征和零样本学习的方法。并针对后续研究的重心，重点阐述基于零样本学习的物体识别方法。

(1)基于模板信息的物体识别

基于模板信息的物体识别方法是物体检测及分类领域中常见的方法之一，也常用于在机器人抓取中。针对某一物体，若已存在对应模板，可通过归一化互相关函数(Normalized Cross Correlation,NCC)作为评估指标，其定义如式 1-1。通过像素级精度深入剖析目标图像与模板图像间的相关性，以实现对物体类别的识别。

$$\mathrm{F}(f,t)=\frac{1}{n-1}\frac{(f(x,y)-\hat{f})\cdot(t(x,y)-\hat{t})}{\sigma_f\sigma_t} \tag{1-1}$$

公式中，设定 f 为目标图像，模板图像则用 t 表示。NCC 函数则用 $\mathrm{F}(f,t)$ 来计算这两者之间的相似性；$f(x,y)$、$t(x,y)$ 分别代表了 f 和 t 在特定坐标 (x,y) 处的像素值；而 $\hat{f}$、$\hat{t}$ 则分别是 f 和 t 整体的像素均值，反映了图像的亮度平均水平。此外，σ_f、σ_t 分别表示 f 和 t 的像素标准差，用于衡量图像像素值的离散程度，n 表示图像中的像素数量。

然而，摄像机放置位置与角度的差异会导致物体在图像中呈现出不同的尺度，进而引发尺度偏差问题。这种偏差会在很大程度上影响物体识别的准确性。为了解决此类问题，学者们进行了许多研究[5-6]。Kato 等[7]提出一种基于黑白模板的改进方法来减少投影变换引起的识别误差。模板包含黑色边框和多样化的填充图案。通过图像阈值分割识别黑色边框和边缘线，并通过解析图像坐标与模板信息计算出投影变换关系。利用变换后的模板进行物体匹配和识别。

采用基于模板匹配的物体识别方法，可以充分利用其简洁性和高效的计算速度优势，然而，它只适用于对固定环境及物体的识别，并需要借助物体模板信息。因此，在实际应用中，这种方法的适用范围存在一定的限制和局限性。

(2)基于浅层特征的物体识别方法

用特征表达方法对待识别的物体的视觉特征进行对应表示，是物体识别领域中最耗费资源的难题之一。经典的人工设计特征通常依据物体的形态、纹理、大小、色泽等特性进行精心构建，接下来将对构建方法进行详尽的阐述。

为了详尽描述物体的形状轮廓特征，设计了傅里叶描述子(Fourier Descriptor)[8]，该方法将物体的边缘视为一条闭合曲线，并将曲线上每个点的坐标视为以曲线长度为周期的函数。通过对该函数进行傅里叶级数展开，得到一组函数系数，能有效刻画物体边缘的形状特征。在计算机视觉领域，尺度不变特征变换(Scale-invariant feature transform, SIFT)[9]同样用途广泛，通过检测极值特征点，并对这些特征点的方向、尺度等属性详尽描述，从而生成局部特征描述子，该方法能够有效地应对图像的旋转、亮度变化等问题，因此在机器视觉中应用广泛，是一种具有里程碑意义的特征[12-14]。加速鲁邦特征(Speeded Up Robust Features, SURF)[10]是 SIFT 的改进特征，增加了积分图像以及 harris 角点信息的描述[15]，在人脸识别相关工作中应用广泛。梯度方向直方图(Histograms of Oriented Gradients, HOG)[11]描述子统计缘方向分布直方图与图像局部区域的梯度方向，以描述图像特征，该方法可以精准捕捉目标物体边缘的形状特性，因此广泛用于行人检测中[16-17]。

这类物体识别方法具有显著优势，即采用经过人工设计的特征，从而易于理解和优化。然而，这类方法也面临一些挑战。首先，所用特征需要根据指定的任务进行设计，因此设计成本较高；其次，这些方法难以捕捉到图像的全局特征，且所用特征表达能力有限，无法有效描述视觉信息中的抽象特征。综上，这些算法通用性低，且泛化能力也较弱。

(3)基于零样本学习的物体识别方法

目前，深度学习技术备受关注，其在多个任务中取得了令人瞩目的研究成果。然而，这些模型通常需要大量训练样本才能实现良好的性能，并且训练出的分类器仅能对已知物体进行分类。因此，基于零样本学习的未知物体识别方法应运而生，旨在解决未知物体的识别问题。

简单来讲，假设模型已经能够识别马，老虎和熊猫了，现在需要该模型也识别斑马，那么只需要让模型知道一个拥有马的形状、老虎一样的条纹、熊猫一样

的颜色的对象是斑马，那么模型在输入斑马图片的时候就知道这是斑马了，实现网络如图 1-2 所示。

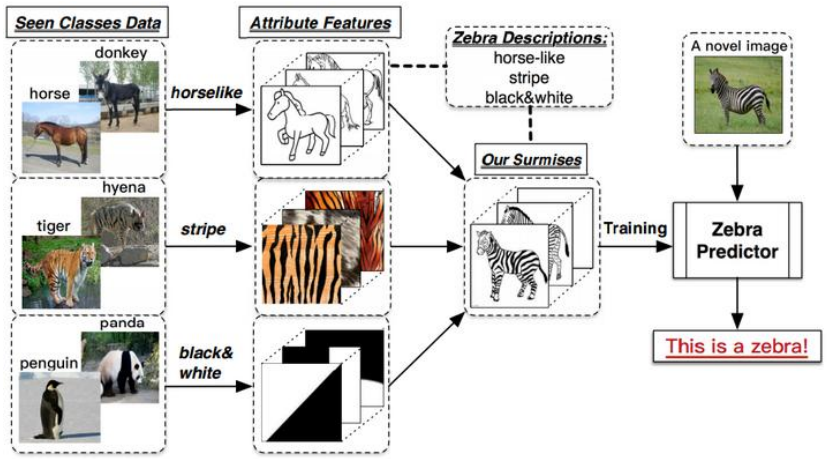


图 1-2 零样本学习介绍

Figure 1-2 Introduction of zero-shot learning

训练集和测试集表示如图 1-3，训练集数据 x_{tr} 及其标签 y_{tr}，包含了模型需要学习的类别，测试集数据 x_{te} 及其标签 y_{te}，包含了模型需要辨识的类别；训练集类别的描述 A_{tr}，以及测试集类别的描述 A_{te}；在分类任务中采用语义向量 $a_i \in A$ 来表示每种类别 $y_i \in Y$，每个属性都通过一个维度来表示，例如“有条纹”、“尖嘴”、“光滑”等，假设一个类别拥有某些属性，则在相应的维度上，将其设置为非零值。

在零样本学习中，利用 x_{tr} 和 y_{tr} 来训练模型，而模型能够识别 x_{te}，因此模型需要了解所有类别的 A_{tr} 和 A_{te}。

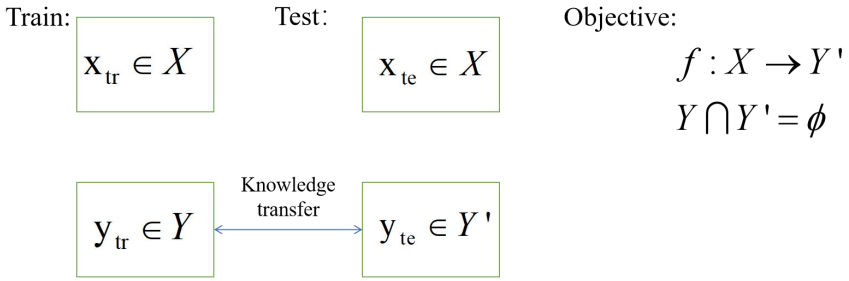


图 1-3 训练集与测试集的表示

Figure 1-3 Representation of training set and test set

基于零样本学习的物体识别方法为未知物体的识别打开了思路，本文也将使

用相关理论进行未知物体的抓取研究。

1.2.2 机器人抓取研究现状

在机器人应用系统中，抓取操作是机器人作业任务的关键之一，涵盖了物体分拣、工具操作等多样化的作业需求。但是，机器人在抓取领域一直面临着一个具有挑战性的问题，即抓取检测。这个问题需要综合考虑多个因素，包括机械臂运动轨迹、物体是否遮挡、环境是否结构化，抓取位置等。因此，人们也将这个问题称为抓取综合问题(Grasp Synthesis problem)[18]。处理抓取综合问题的方法可以概括为两种方式[19]，基于解析分析法的物体抓取识别和基于数据驱动法的物体抓取识别。

(1)基于解析分析法

基于解析分析法通常可以转化为稳定性、灵巧度等限制下的目标优化问题，目标是指机械臂夹爪受力闭合的抓取。

为了实现机械臂夹爪的力闭合，一般会简化夹爪模型与物体模型[20-21]。Bicchi等[22]深入探讨了静态不确定的抓取过程中接触力与接触模型之间的内在联系，结果表明并非所有接触力都需要进行详尽计算，而是仅需关注其中一部分关键力，因此可以通过优化算法，有效减少末端执行器的控制自由度，提高抓取效率和精度。Prattichizzo 等[23]和 Rosales 等[24]通过在末端控制器和物体相碰的位置引入了弹簧机制，将机械手抓取物体的配置问题转化成一个参数优化的问题。针对测量物体位置产生的误差，Zhang 等[25]提出了粒子滤波算法，现被广泛用于估计物体动态物理模型的参数。然而，该方法当前只能用于追踪二维物体，且还处于理论阶段，不能应用于现实任务。

在实际的抓取任务中，机器人系统误差和随机误差导致对末端控制器与待抓取物体之间的相对位置无法精准估算。因此，Rodriguez 等[26]针对三指机械手进行牢笼配置，即使末端控制器存在无法忽视的位置偏差，仍能稳定且准确地抓取物体。Seo 等[27]分成两个步骤执行物体抓取的操作，第一步发现接触位置之间的牢笼配置并控制双指执行抓取，第二部通过额外的接触对物体进行约束，以达到稳定抓取的目的。

(2)基于数据驱动方法

本文研究的是基于数据驱动方法下机器人的抓取识别，接下来重点阐述该方

法的研究近况。

基于数据驱动方法是一种研究机器人成功抓取与目标物体之间的抽象映射关系的方法。自从 Graspit[33]方法的推广，物体抓取方法的研究正逐渐聚焦于利用数据信息和先验知识策略的方向。

基于数据驱动方法可以将抓取对象分为已知物体和未知物体。在抓取已知物体时，经常建立物体在三维空间中的模型，随后将所获取的信息同已有的模型相匹配，以实现精确的抓取操作。然而，对于未知模型物体的抓取，则需要采用不同的策略。首先，需要处理该物体的特征，以提取其独特的属性。接着，将这些特征与已知物体比较，找出可能的匹配项。最后，通过设置损失函数优化模型，确保识别的准确度，从而成功抓取。

已知物体的抓取是根据已有的先验知识对新数据进行抓取估计的一种技术手段。因此，该类抓取问题可以总结为物体抓取位姿的估计问题。Miller 等[28]假设在已知物体三维模型的基础上，专注于识别并抓取那些形状简单、规则化且差异显著的物体，如球体和柱体等。通过采用该方法，能够有效减少抓取候选项的数量，从而显著提升识别效果。Borst 等[29]为了生成一组可行的抓取方案，首先利用表面特征信息进行提取，随后通过启发式算法对这些方案进行深入的可行性评估。研究者 Hsiao 等[30]基于快速扫描随机树的方法，旨在识别物体表面最适合抓取的区域。该方法利用球形、柱形等外形特征对物体进行描述，进而根据这些描述出的抓取特性来估计物体的抓取位姿。尽管这些算法在仿真环境下取得了显著成效，但在实际的机器人抓取任务中，其性能仍有待进一步提升。

许多研究者开始研究人类抓取物体的方法，并从中获得启发，取得了无数的研究成果。Ciocarlie 等[31]通过深入研究人类神经网络的科学知识，发现人类对手部的控制精度远不及人手自身的精细调节能力，这表明人类在控制手部时实际上采用了欠驱动的方式。基于这一发现，作者将欠驱动理论直接应用于机械手的研究中，旨在通过减少机械手的配置空间来高效找到期望的抓取位置。同时，还采用了特征值抓取方法，以进一步确定机械手的稳定抓取配置，从而提升抓取的准确性和效率。Ekvall 等[32]收集了大量数据，这些数据记录了人们戴着抓取指定物体时手部的运动情况。通过对这些海量数据的深入学习，能够更好地理解人类抓取的习惯，从而为机械手的抓取策略设计提供有力的依据。Kroemer 等[34]采用了

强化学习方法，依赖于一个底层控制器来精准地执行抓取操作。在抓取过程中实时将抓取效果反馈给控制器，以便计算相关函数并优化抓取策略。

然而，上述方法都存在着建立物体 3D 模型的需求以及应用领域的限制。首先，建立物体的 3D 模型非常复杂，需要大量的时间进行采样和模拟，其次，已知模型物体的抓取方法在实际使用中限制很大，不适合应用于各种非结构对象的环境。

抓取未知物体是指抓取模型未知且不属于训练集的对象。与抓取已知物体不同，机器人需要将未知物体与已知物体进行对比分析确定最适合抓取的区域，从而提高抓取成功率。

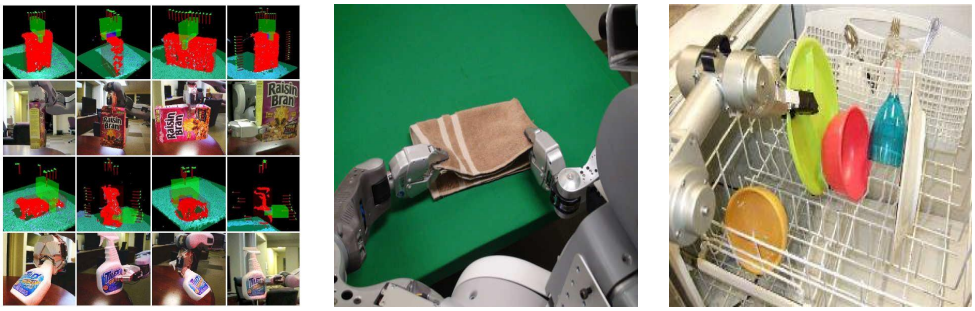
在机器人抓取过程中，通常所选取的抓取区域会展现出相似的视觉特征，如形状、纹理等局部特征。为精确描述待抓取物体，需要明确阐述其与类似物体之间的距离度量，并将其作为该物体的特征描述，以指导抓取操作。El-Khoury 等[35]为了模拟人类抓取动作的规律，采用了以物体的原型作为待抓取物体的定义的方法。该方法通过利用简单的几个几何模型来替代三维物体的各个部分，从而实现了对物体形状和结构的准确描述，再利用机器学习检测物体适合抓取的部位。Boularias 等[36]用马尔科夫随机树(MRF)定义物体模型，随机树节点定义为物体的 3D 点云，这些节点信息包含该点云是否可以抓取的判断，实验结果显示，该方法可以高效区分出 3D 物体上最适合抓取的地方。

也有研究者通过物体原始形状的信息进行抓取识别。Morales 等[37]依赖视觉反馈确定具体抓取对象，并在平面进行抓取时，依据物体的形状信息来选定适宜的抓取姿态。同时使用最近邻方法优化抓取效果，通过抓取经验的学习来得到最优的抓取姿态。Dunes 等[38]采用多角度采样的方式对物体外形进行建模，生成二次曲面，并根据其短轴和质心信息配置抓取信息并推断机械臂夹爪抓取方向。

Rao 等[39]根据 2D 颜色信息和 3D 距离信息是机器人得到物体的大小和外形后通过支持向量机训练抓取。Bohg 等[40]建议了基于对称性条件的物体形状重建方法，首先对待抓取对象原始的 3D 点云模型进行恢复，用来得到该物体完整 3D 模型，此时可以通过抓取已知物体的方法对其进行抓取质量计算。

在这之外，还有研究者获得最佳抓取未知物体方式是通过建立抓取方式和物体特征之间的对应关系。Hsiao 等[41]使用物体外形和经过分割后的点云情况进行

启发式抓取生成假设，并使用手指数量、手间手掌之间距离等特征权重表选择出得分最高的抓取方法，如图 1-4(a)所示。Maitin 等[42]通过检测深度图像边界的技术，同时提取物体边缘的位置和角点的信息可以完成可以抓取纸袋等可变形物体，如图 1-4(b)所示。Saxena 等[43]提出了利用线性回归算法对物体可抓取部位的图像信息进行学习的抓取部位检测方法，在真实环境下进行实验验证了算法的可行性，并获得了优异的抓取效果，如图 1-4(c)所示。



(a)3D 点云信息抓取　　(b) 可变形物体抓取　　(c)2D 视觉信息抓取

图 1-4 基于视觉信息的未知模型物体抓取识别

Figure 1-4 Grasping and recognition of unseen model objects based on visual information

由于深度学习技术在人工智能领域的快速推广，许多人研究人员开始在机器人抓取中使用深度学习方法[44-46]。Redmon 等[47]通过输入包含物体的一张图像并使用卷积神经网络方法(Convolutional Neural Network, CNN)检测待抓取对象的抓取区域。Finn 等[48]通过 CNN 模型对物体的抓取数据进行采集获得其被抓取过程中的视觉特征点并构建一套抓取策略，来完成对未知物体的抓取估计。Pinto[49]和 Levine[45]通过自监督学习方法的运用，引导机器人学习并收集待抓取对象的抓取数据。该方法实验时展现了不错的抓取效果，但是依赖于庞大的样本集来训练物体模型，这不仅存在过拟合的风险，同时也导致模型计算成本高昂，降低抓取效率。

(3) 研究现状总结

上述机器人抓取研究现状表明，解析分析方法利用机器人夹爪实现力闭合，从而稳定、高效地抓取物体。为了确保机器人抓取过程的平衡性和稳定性，必须使用刚体模型假设或线性摩擦约束等不切实际的假设，导致这些方法的仿真结果都很好，但没有实际应用的指导意义，抓取效果远达不到满足机器人实时抓取要

求的总体水平。

基于数据驱动的方法更侧重于识别物体并进行描述，例如描述抽象属性、测量相似性等等。该方法一般通过大量数据建立相应的算法模型，在机械手和物体之间获得数据连接，然后创建抓取方法来识别和抓取物体。然而，这种方法缺乏一个严格的数学模型，无法预测捕获不符合预期的对象的策略。

在机器人抓取领域，基于数据驱动的物体抓取方法为应用机器人抓取任务奠定了坚实的基础，同时抓取效果优异。然而，在实际应用中，机器人不可避免地会遇到非结构化环境以及数据收集困难等挑战。因此，当前该领域面临的核心问题是如何高效、准确地识别物体，并精准地获取其抓取姿态，以确保机器人能够顺利完成抓取任务。

1.3 本文的主要研究内容

本文为了解决未知物体的抓取问题，并从零样本学习的角度进行深入分析。为此，我们选择了零样本学习方法来检测定位未知物体，并通过局部特征匹配与点云配准结合的方法估计物体 6D 姿态，从而执行抓取工作。本文的主要研究内容如下：

(1)根据零样本学习理论，我们成功训练了一个可训练的未知物体检测网络。该网络具备在无需任何训练样本的情况下，准确定位并识别未知物体的能力。这一创新为机器人抓取领域解决了物体识别困难的问题。

(2)现有的物体六自由度姿态估计方法侧重于处理已训练过的对象，针对未知物体纹理细节较弱或在有遮挡、光照复杂的非结构化环境中的六自由度姿态估计仍是一个具有挑战性的问题，提出了一个物体六自由度姿态估计的网络。首先通过局部特征匹配获取图像对中足够多的匹配点对，其次，通过传感器读取匹配点对的深度信息得到相应点云数据并将其作为后续点云精准配准的初始点云进行配准，最终得到源点云相对于目标点云的旋转矩阵和平移向量，即未知物体在机器人坐标系下的六自由度姿态。

(3)本文提出了一种端到端的未知物体抓取方法，该方法基于未知物体检测网络和抓取建议生成网络。为了验证该方法在机器人实际抓取任务过程中的可行性，我们搭建了一个机器人抓取实验平台，并进行了针对未知物体的抓取测试实验。

1.4 本文的章节安排

本文基于零样本学习对未知物体的定位和识别进行了研究，并设计了抓取建议网络用于有效、安全地执行抓取任务。

章节安排如下：

第一章绪论，介绍本文的研究的背景和意义，简单说明了本文的主要工作内容，再介绍近年来，学者们对未知物体识别和抓取的研究现状，最后总结本文主要贡献。

第二章基于零样本学习的未知物体检测，本文介绍了零样本学习的基本概念，并探讨了在未知物体和视觉识别中容易被忽视的背景问题以及定位和识别未知物体所需解决的难题。通过联合建模视觉特征与语义域之间的相互作用，我们能够更好地理解和应对这些挑战。

第三章未知物体 6D 姿态估计，本章首先通过局部特征匹配获取图像对中的匹配点，读取匹配点的深度信息得到点云并将其作为 ICP 的初始点云进行配准，最终得到源点云相对于目标点云相的旋转矩阵和平移向量，该网络可以用于实时估计任何看不见的物体的 6D 姿态，无需任何再训练且泛化性高。

第四章抓取研究与实验分析，搭建机器人抓取平台，设计机器人对未知物体的抓取实验。

第五章总结与展望，总结了本文的主要工作的内容与贡献，并对后续工作的展望进行简单阐述。

第 2 章 基于零样本学习的未知物体检测

2.1 引 言

零样本学习(Zero Shot Learning,ZSL)是一种先进的机器学习方法，其主要目标是解决在训练阶段未出现的类别的分类问题。在传统的监督学习方法中，模型需要通过大量的标记训练样本来学习如何对新的物品进行分类和抓取。然而，这种方法在面对新的、未知的物体时可能会遇到困难，因为机器人可能从未见过这些物品，也就无法从训练数据中学习到如何抓取它们。

在这种情况下，零样本学习方法就显得尤为重要。它的基本思想是利用已知类别的数据来训练一个视觉-语义交互模型，然后通过这个模型将知识泛化到未知类别的数据上。这样，即使机器人从未见过需要抓取的物品，也能够识别并成功抓取这个物品。零样本学习的实现主要依赖于两个关键技术：属性表示和语义映射。属性表示是将物体的特征转化为一种可以被计算机理解和处理的形式，例如将物体的颜色、形状、纹理等特征转化为数值向量。语义映射则是建立已知类别和未知类别之间的语义关系，例如通过类比、推理等方式。

尽管零样本学习在理论上具有很大的潜力，但在实际应用中还面临着许多挑战。首先，零样本学习假设的场景是一个非常简单的情况，即只有一个主导类别出现在一个图像中；其次，是对一个完整的场景进行未知物体识别，而在实践中，属性和语义描述通常只与单个对象相关，并非整个场景；第三，零样本学习为基本任务中未知类物体提供了答案，如分类和检索，但它不能缩放到高级任务，如场景解释和上下文建模，这需要对场景中所有突出对象进行基本推理。第四，全局属性更容易受到背景变化、视点、外观和尺度变化以及实际挑战的影响，如遮挡和杂波。因此，对于存在属于多个对象类别的不同竞争属性的复杂场景，图像级零样本学习的应用效果差强人意。因此，本章主要研究在图像中可能受到语义类描述噪声影响的情况下，通过将视觉图像特征与语义标签信息同时联系起来，预测每个实例在给定图像中的精确位置并输出类别。

2.2 零样本学习相关介绍

零样本学习研究过程中，认为已知类别和未知类别之间拥有一个共享的映射关系，如式 2-1。该函数衡量视觉特征$\theta(x)$和语义特征$\phi(y)$之间的相容性，W是需要学习的映射矩阵。

$$\mathrm{F}(x,y:W)=\theta(x)^{T}W\phi(y) \tag{2-1}$$

目前零样本学习的方法一般分成两种：基于相容性的方法和基于生成对抗网络的方法。在本文的研究范围内，主要使用基于相容性的零样本学习方法，因此着重介绍该方法的相关理论背景，以便这部分讨论的重点更加清晰明确。基于相容性的零样本学习方法在实际应用中性能优越该方法核心在于准确捕捉物体在语义空间或是视觉空间中投影之间的相关性。

在模型的训练过程中，首先利用已知类别的视觉特征和属性语义信息，建立并学习它们之间的映射关系。之后，在测试过程中，使用这些已构建的映射关系计算未知物体的视觉特征与属性语义之间的相容性，最终选择相容性得分最高的未知物体类别作为预测结果。这种视觉-语义映射有效地连接了视觉空间与语义空间。

如图 2-1 所示，在左侧，我们利用卷积神经网络（CNN）提取已知和未知类别图像的视觉特征$\theta(x)$。而在右侧，我们则定义了物体的属性$\phi(y)$。位于两者之间的共享嵌入空间，该空间的投影方式可以是从视觉空间映射至语义空间、从语义空间映射至视觉空间，或将两者共同投影至隐空间，通过这样的方式，可以在零样本学习的框架下实现更为准确和高效的物体分类。表达式$\theta(x)^{T}W\phi(y)$旨在量化在共享空间中两者的相容性，该值越大，该类别的属性语义与对应图像中的物体匹配度越高。

在训练过程中，模型通过连续迭代计算得到映射矩阵W。随后，在测试过程中，首先从未知类别的图像中提取视觉特征，并利用先前计算出的映射矩阵W与属性语义$\phi(y)$进行相乘操作。接着，我们根据计算得到的相容性得分来选取得分最高的类别。相容性得分最高的类别该表该未知物体投影得到的$\theta(x)^{T}W$与类别 y 的视觉特征$\phi(y)$具有最高的相似性。基于这一逻辑，最终将类别 y 的标签分

配给该未知类别。通过这样的处理，能够实现零样本学习对未知类别的物体进行准确分类。

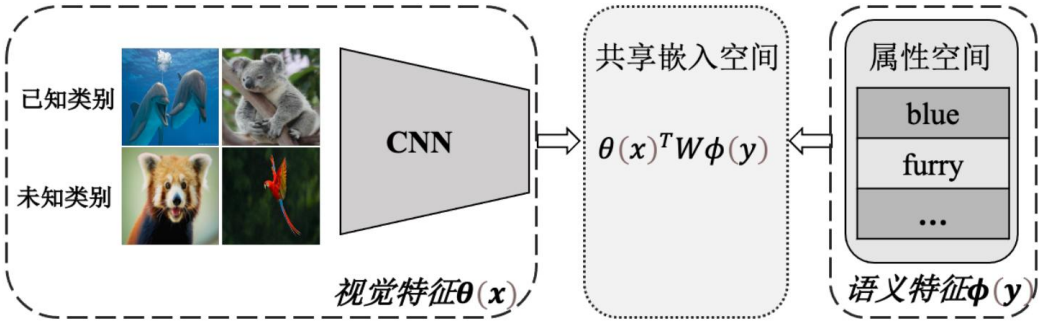


图 2-1 视觉特征与语义特征的映射

Figure 2-1 Mapping between visual features and semantic features

2.2.1 视觉空间投影到语义空间

为了学习映射矩阵 W，一般将视觉空间映射至语义空间，通过优化过程使视觉特征与语义空间中的表示之间的差异最小化。

零样本学习核心目标在于对未知类别准确分类，其关键在于各类别与其语义特征的映射关系的一致性。因此，在训练过程中，若 W 能更精准地刻画类别与其语义特征之间的内在联系，则模型在测试过程中将展现出更强的未知类别识别能力。为实现这一目标，需确保视觉特征 $\theta(x)$ 与语义特征 $\phi(y)$ 具备足够的区分性，从而保证映射关系的精确性。

Akata 等[50]提出了基于属性的标签嵌入模型(ALE)，因为作者认为各个属性之间具有一定的相关性，因此在模型学习的过程中应当充分考虑全部属性的影响。ALE 为了高效学习视觉特征与语义特征之间的相关关系，将两者之间的双线性关系称为相容性。通过设定一系列限制条件，确保同类图片与语义属性之间的相容性高于不同类图片与语义属性之间的相容性。这种策略有助于模型更准确地捕捉和描述图像与语义之间的内在联系，进而提升零样本学习的性能。

引入了转换矩阵计算空间相容性，该矩阵被称为视觉特征和语义特征之间的映射矩阵 W。空间相容性得分 F 代表图像特征与语义特征的相似性，其数值越高，标志着两个特征之间的匹配性越高，其可能是同类别的概率也越大。在训练过程中，该模型采取学习算法以优化 W，从而使得对于给定一张图片，其正确标签排名最优化，脱颖而出。在预测阶段，采用式 2-2 所示的预测函数，结果选

择具有最高相容性分数的图像类别，成为最终的预测结果。

$$f(x:w)=\arg\max_{y\in Y}F(x,y:W) \tag{2-2}$$

根据各个属性之间拥有一定的相关性这一想法，Akata 等人[51]认为仅考虑对象的属性可能无法完全反映语义空间中的对象类别信息，因此，在提取人工定义的属性的基础上，结合无监督自然语言处理模型[53]提取的词向量和 WordNet 模型[52]提供的相似关系，提出了一种结构联合嵌入(SJE)方法。SJE 综合考虑了通过三种不同方式提取语义特征并计算与图像特征之间的相容性。在此过程中，引入了包含三个不同映射矩阵的相容性函数，这些矩阵各自代表了其相关性，具体计算如式 2-3 所示。

$$\mathrm{F}(x,y:\{W\}_{1\ldots k})=\sum_k\alpha_k\theta(x)^TW_k\phi_k(y)\ \text{s.t.}\sum_k\alpha_k=1 \tag{2-3}$$

α_k 代表不同嵌入方式的权重，其权重和为 1。结合输出嵌入，可以表示所有的信息。损失函数为如式 2-4 所示。

$$\mathrm{L_{SJE}}=\frac{1}{N}\sum_{n=1}^{N}\max_{y\in Y}\{0,1(x_n,y_n,y)\} \tag{2-4}$$

在零样本学习问题中，语义特征在连接已知类别与未知类别方面起到了关键作用，因此逐渐引起了研究人员的广泛关注。Guo 等[54]发现零样本学习中，属性的信息量及其对模型的贡献存在差异，认为不同属性价值不同，应采取相应的处理策略。基于此，提出了根据每个属性所携带的信息量以及其对分类的贡献度进行筛选的方法，然而，该方法忽视了属性之间存在的语义联系[55]，因此在去除某些属性的同时也会影响到属性间的相关关系。

研究者们开始研究属性是否具有区分能力，同时保持各属性之间的相关性。Liu 等[56]认为不同的属性可以根据其建模的贡献度赋予不同的权重来解决该问题，因此提出了通过融合不同层级的注意力向量来更好地学习视觉特征和语义特征之间的关系的方法。

Zhu 等[57]发现在学习语义与视觉的映射时，视觉特征本身的区分能力一样不可忽视。针对此问题，研究人员提出了一种通过语义引导能够自动发觉物体最具区分性的部位的多注意力定位模型，从而提高网络在细粒度条件下对物体的识别能力。但在提取多个目标区域时存在一定程度的冗余性，单一区域提取的视觉特征与语义特征之间的相关性也存在一定的弱化现象。这些问题可能在一定程度上

影响了模型的性能，因此需要在后续研究中加以改进和优化。

Li 等[58]认为视觉和语义的鉴别性都一样重要，因此提出了一种基于鉴别性的模型。首先提取原始图像与精细图像的视觉特征，再将这两种不同尺度的特征映射到语义空间或隐空间之中。在这一过程中，精细图像在维持物体整体性的同时，有效地剔除了一部分的环境信息，从而实现了视觉特征与语义特征之间差异的平衡。

2.2.2 语义空间投影到视觉空间

当视觉特征的维度远超过语义特征的维度时，ZSL 模型通常使用深度神经网络从图像中提取其高维视觉特征，如通过 ResNet[59]提取出的特征有 2048 个维度，而语义特征一般是几十到几百个维度。当视觉特征不经其他处理就直接投影至语义空间时，投影特征间的距离将会变窄，导致某个特征会成为许多个未知类别特征的近邻点，造成枢纽点问题[60]。

在零样本学习的测试阶段，一般采取最近邻算法，但必须充分考虑、处理枢纽点问题，以确保模型能够准确地进行预测。Shigeto 等人[60]证实了采用岭回归确实会导致枢纽点问题的放大，因此提出将语义特征有效地投影至视觉空间来缓解该问题。Romera-Paredes 等[61]提出了极简零样本学习模型(ESZSL)，它在类别语义和视觉分类器学习嵌入函数时建立了一种约束关系，属性识别可以最小化多个分类器的误差来进行隐式学习，同时考虑到分类任务中每个属性的重要性。然而，许多现实生活中没有有序空间结构的数据导致目前对卷积神经网络的研究被限制。这些图结构中的节点连接方式各异，在处理图数据时，需综合考虑节点的属性信息与结构信息，以确保对图结构的全面而准确的理解。因此，需要使用图卷积神经网络[62]自动地同时学习图的特征信息和结构信息。

在训练阶段，每个节点会转变自身的特征信息，并将这些信息传递给相邻的节点。随后，每个节点会收集来自相邻节点的特征信息，并对这些汇总的信息进行非线性变换，以增强模型的性能。简言之，通过提取和转换图中每个节点的特征信息，融合节点的局部信息，并通过非线性变换使各节点都可以被相邻节点影响。

Wang 等[63]在零样本学习中引入图卷积网络，提出了图卷积零样本学习(GCNZ)，样本中类别的语义代表了图中的节点，在训练过程中，使用已知类别

的语义学习 GCN 参数矩阵；在测试过程中，利用学习到的参数矩阵获取未知类别的标签。该网络的每次更新都会整合相邻节点的信息，因此，节点不仅蕴含丰富的语义信息，还承载着结构信息。在 GCN 中，随着层数的增加，参与操作的信息量也相应增多，这使得模型的泛化能力逐渐增强，感知域也随之拓宽。然而，不容忽视的是，在图结构的传播过程中，相距较远的节点所共享的知识可能会受到拉普拉斯平滑(laplacian smoothing)[64]的影响，导致其被逐渐稀释，影响最终结果。为了避免相关知识被稀释，应当尽量 GCN 的层数。然而，这种做法可能会降低节点间信息的混合程度，进而影响最终结果的准确性。Michael 等人[65]提出 DGP 模型，旨在解决层数较少时可能出现的过度平滑问题。通过促进信息在这些节点之间的传播，从而优化了整体的信息流动效率。然而，DGP 模型可能会扰乱节点之间原有的结构，导致原有的祖先节点以及后代节点都有可能转变为某节点的邻近节点，进而与它们共享相同的权值。为了解决此问题，DGP 引入了加权方案来衡量每个节点的贡献，以获得良好的结果。

2.2.3 投影到隐空间

除上述两种投影方式外，将视觉和语义特征一起投影到隐空间同样是非常优异的方式。

Elyor Kodirov 等人[66]在编码阶段投影视觉特征至一个隐空间，随后，在解码阶段将这些隐特征重新投影至视觉空间，尽可能恢复原始视觉特征，确保信息的完整性和准确性。Jiang 等[67]提出了同时将语义特征与视觉特征投影至隐空间，既保持隐特征的鉴别性，又约束其保持语义信息，在此过程中，确保语义信息间的关系得以保留。近年来人们使用端对端的模型来建立和训练模型，将视觉特征分别投影至隐空间及语义空间，以得到更多信息。这种双重映射策略进一步完善了所学映射矩阵[56-57]。

2.3 零样本目标检测

在 2.1 节中，我们指出了零样本学习在未知物体识别方面存在的一些问题。为了解决这些问题，很多研究者[50-51]对传统的零样本学习方法进行了改进，使其不仅能够对未知物体进行分类，还能够在没有可视化示例的情况下识别和定位新对象类的每个实例，以满足目标检测的需求，这种物体识别方法被称为零样本目标检测，本节将具体介绍该方法，并在此基础上进行改进，最后通过实验验证并实现检测及定位未知物体的功能，为后续抓取的研究提供了足够的先验知识。

2.3.1 整体框架介绍

本文使用的框架包含两个主要模块，第一个模块对象级特征获取模块，第二个模块为物体检测模块，将视觉特征与语义嵌入相结合，以实现未知物体的检测，整体框架与流程如图 2-2 所示，接下来将对此进行详细阐述。

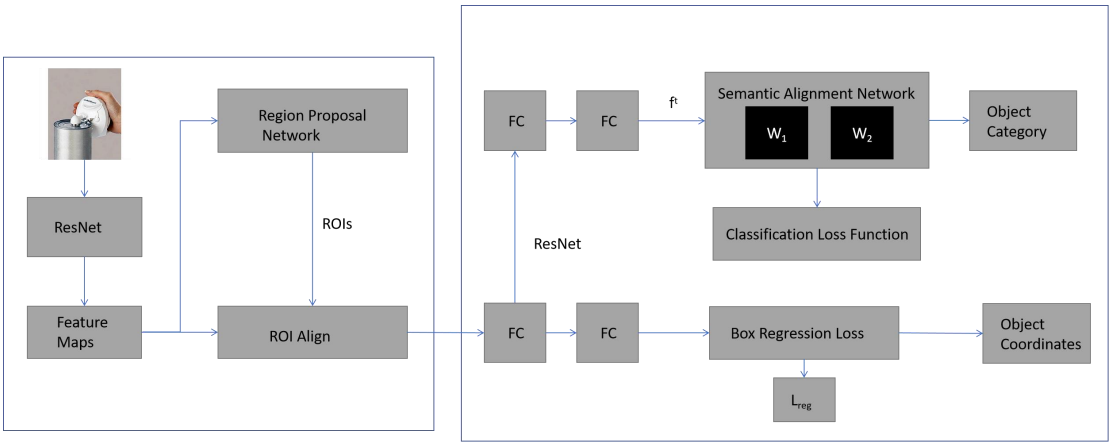


图 2-2 零样本目标检测框架

Figure 2-2 Zero-shot object detection framework

首先，先对零样本目标检测中使用到的符号进行说明。定义 $S=\{1,...,S\}$ 表示已知类，它们的示例在训练阶段使用，S 表示它们的总数；还有另一组未知类 $U=\{S+1,...,S+U\}$，其实例仅在测试阶段可用；另外用 $C=S\bigcup U$ 表示所有对象类的集合，这样 $C=S\bigcup U$ 就表示标签空间的基数。

再将相似的对象类分成一组来定义一组元类。这些元类 $M=\{z_m:m\in[1,M]\}$ 表示，其中 M 表示元类的总数，$z_m=\{k\in C\ \text{s.t.g(k)=m}\}$。这里，g(k) 是一个映射函数，它将每个类 k 映射到它对应的元类 $z_{g(k)}$。元类是互斥的，即 $\bigcap_{m+1}^{M}z_m=\phi$ 且

$\bigcup_{m+1}^{M} z_m = C$。

所有训练图像的集合用 χ^{S} 表示，其中包含所有已知对象类的实例。所有包含未知对象类样本的测试图像的集合用 χ^{U} 表示。每个测试图像 $x \in \chi^{U}$ 包含至少一个未知类的实例。

通过 word2vec 对每个类定义一个 d 维词向量 v_c。对每一个标签都定义为 y_i，因为目标检测任务还涉及识别背景类，另外再引入扩展标签集：$S' = S \bigcup y_{bg}$，$C' = C \bigcup y_{bg}$，$M' = M \bigcup y_{bg}$，其中 $y_{bg} = \{C+1\}$ 是表示背景标签的单例集。

零样本目标检测的实现可以描述为给映射空间一张图像 $\chi = \chi^{S} \bigcup \chi^{U}$，并输出 C'。零样本目标检测的目的是学习一个映射函数使正则化经验风险 $(\hat{\Re})$ 最小。映射函数公式可以表示为 $\arg\min_{\Theta} \hat{\Re}(f(x;\Theta)) + \Omega(\Theta)$。

式中，训练过程中的 $x \in \chi^{S}$，Θ 为参数集，$\Omega(\Theta)$ 为学习权值的正则化。映射函数具有形式如公式 2-5 所示。

$$f(x;\Theta) = \arg\min_{y \in C} \max_{b \in \mathrm{B}(x)} \mathrm{F}(x, y, b; \Theta) \tag{2-5}$$

其中，$\mathrm{F}(\cdot)$ 是一个兼容性函数，$\mathrm{B}(x)$ 是一个给定的图像 x 中所有边界框建议的集合。

2.3.2 对象级特征获取模块

针对输入图像 x，采用特征提取网络 ResNet 对其进行特征提取，产生相应的特征图，并将特征提取网络所生成的特征图纳入区域建议网络(Region Proposal Network,RPN)[68]和 ROI Align 的输入体系中。RPN 旨在识别特征图中的候选目标框(Anchor)，并对其进行坐标的精细调整。随后，基于调整后的特征图与候选目标框，我们利用 ROI Align 层精准响应图像中的特定感兴趣区域，从而提升目标检测的准确性与效率。每个感兴趣的区域对象，其结果对象级特征值 f 都得以充分展现。

RPN 通过滑动窗口的方式生成目标框，目标框长宽比一般为 1：1，1：2 和 2：1，面积比一般为 128*128，256*256，521*521，图 2-3 为生成效果。

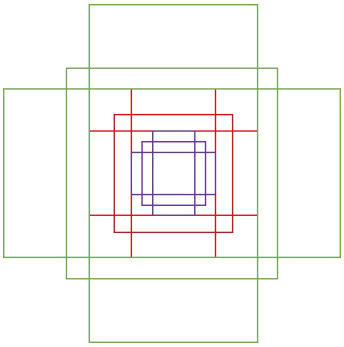


图 2-3 Anchor 生成示意图

Figure 2-3 Anchor generation diagram

在 RPN 网络中输入大小为30∗40 的特征图，会生成30∗40∗90＝10800个初始框。对这些初始框进行判断，确认是否含有目标并进行回归与筛选。

RPN 模块既可以独立使用，也可以作为 Faster R-CNN 的一部分进行使用，但后者生成候选区域时数量会相对较多。因此，为了提升模型的性能，我们引入了非极大抑制(Non-Maximum Suppression,NMS)筛选掉多余的候选框。在本文中，NMS 阈值被设为 0.7。若经过筛选后，剩下的数量仍然超过预设的值，则仅保留预设值内的候选框。这样的策略有助于在保证候选框质量的同时，控制候选框的数量，从而优化模型的训练效果。

ResNet 网络引入了残差结构，为了避免训练过程中出现梯度消失或爆炸的问题，在网络中使用 skip Connection 为下一层网络输出多尺度特征信息，以支持更深层次的网络设计。图 2-4 为 ResNet34 的网络结构。

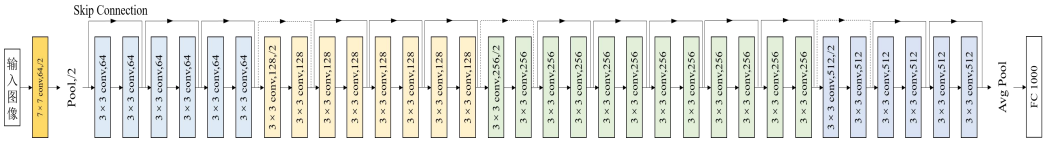
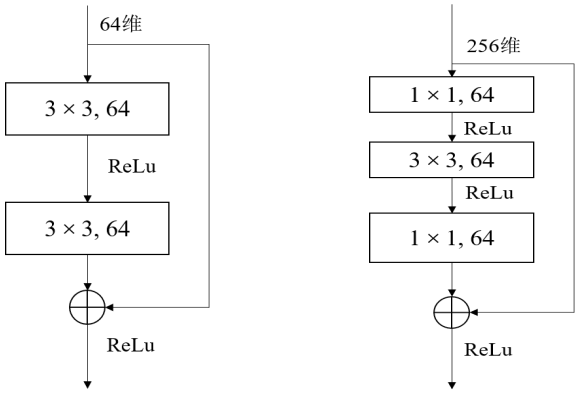


图 2-4 ResNet34 结构

Figure 2-4 Network structure of ResNet34

残差网络两种结构如图 2-5 所示，2-5(b)使用了1×1 和3×3 两种规格的卷积核，既可以减少网络运行过程中无意义参数的产生，也可以利用残差网络来搭建更深的网络，因此使用较多。



(a) (b)

图 2-5 残差网络

Figure 2-5 Residual network

解决梯度消失的原理如下式 2-8 和 2-9 所示：

$$H(x^{l+1}) = H(x^{l}) + x \tag{2-8}$$

$$H'(x^{l+1}) = H'(x^{l}) + x \tag{2-9}$$

式中，$H(x^{l})$ 代表第 l 步的输出，$H(x^{l+1})$ 则代表第 $l+1$ 步的输出。网络梯度发现下降时，对其进行求导，并将导数设置成常量，从而有效预防梯度消失现象的发生。

另外，本文对 ROI Pooling 进行了优化处理。在计算目标感兴趣区域时，由于 Region Proposals 过程所得到的 x, y, w, h 等参数往往不是整数，因此需要进行取整操作以便后续计算，但这种取整过程会导致误差产生。例如，输入 400×400 大小的图片以及一个 300×300 大小且位置已知的目标框，网络在经历五次下采样后，会获得一个缩小至原图 $\frac{1}{32}$ 的特征图，目标框尺寸也会相应变化。

然而，当进行 ROI Pooling 操作时，由于目标框在特征图上的对应尺寸(9.37)并非整数，无法直接找到对应的区域。为解决这一问题，我们通常会对该数值进行取整处理，在特征图上寻找最接近的整数区域进行池化。但这样做会产生一定的误差，如图 2-6 所示，这种取整操作可能导致目标区域在特征图上的定位不准确。

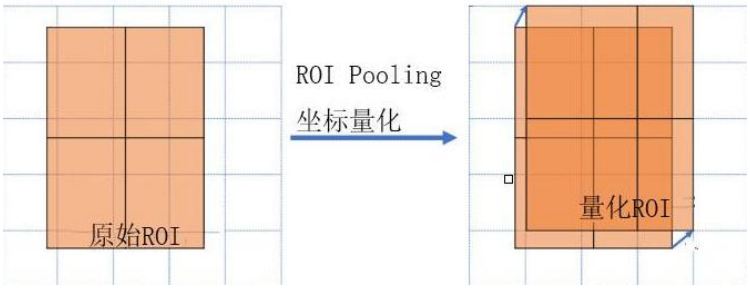


图 2-6 误差的产生

Figure 2-6 The generation of errors

根据图 2-6 所示，通过 ROI Pooling 后得到的目标框相较于原始目标框存在了偏移，引入了不小的误差。在后续步骤中，目标框还需经历上采样处理，这一过程中会进一步累积误差。经计算，误差超过 11 个像素。后续在将目标框输入 FC 层中时还会进行同样的操作，这一过程中同样会出现与第一步类似的误差，导致误差累积。

在两次取整操作中，每次操作都会导致分割框的位置发生偏移，并且该偏移量会随着迭代次数的增加呈现指数级增长，从而对最终的检测结果产生影响。因此，我们不再使用 ROI Pooling 来提取感兴趣区域的特征，而是采用 ROI Align。不同于 ROI Pooling，ROI Align 避免了对坐标取整的操作，而是通过插值的方法获取浮点数坐标，使小数部分得以保留。

ROI Align 保留了感兴趣区域在分割池化后得到的浮点数坐标。在此过程中，感兴趣区域被 ROI Align 划分为一定数量的单元，具体大小由 FC 层决定。如图 2-7 所示，感兴趣区域被划分为四个单元。随后，在各单元中选取采样点并利用插值法精确计算这些点的坐标。最后，最大池化这些采样点。

通过 ROI Align 方法，我们成功解决了网络在进行回归时分割框位置出现偏移的问题，并使小目标物体的检测精度得到了显著提高。

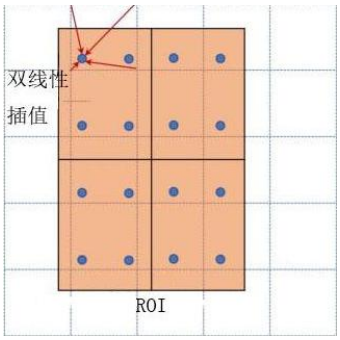


图 2-7 改进之处

Figure 2-7 Improvements

2.3.3 物体检测模块

第二个模块专注于物体检测，其结构包含两个分支，即分类分支和回归分支。这两个分支均接收对象级特性 f 作为输入。

顶部分支作为分类分支，经过训练，能够预测每个候选框所属的对象类别。这可以分配一个类 $c \in C'$，它可以是已知类、未知类或背景类别。该分支捕捉已知类与未知类之间的语义关系，使得其能够更准确地识别不同类别的对象，为后续的物体检测任务提供有力的支持。

该分支的核心网络为语义对齐网络(Semantic Alignment Network，SAN)，该网络的核心是一个可调的全连接层(Fully Connected layer,FC)。通过调整参数 $W_1 \in \mathbb{R}^{d \times d}$，该层能够将输入的视觉特征向量投影至一个 d 维的语义空间。随后，这些特征映射被进一步投影至固定的语义嵌入 $W_2 \in \mathbb{R}^{d \times (C+1)}$ 上，以实现与文本信息的对齐。这里的语义嵌入采用无监督方式通过文本挖掘获得，具体在本文中我们使用 Word2vec 嵌入方法。给定顶部分支中 SAN 的特征表示 (f^t) 作为输入，整个操作过程可以通过公式 2-6 进行表达。

$$o = (W_1 W_2)^T f^t \tag{2-6}$$

其中，o 表示输出的预测分数，W_2 是通过将所有类别的语义向量堆叠而成的，其中也包含了背景类，在处理背景类时，我们采用平均词向量作为其嵌入表示，该平均词向量的计算方式如式 2-7 所示。

$$v_b = \frac{1}{C} \sum_{c=1}^{C} v_c \tag{2-7}$$

采用平均词向量作为背景类的嵌入方法，其优势在于能够包容部分对象(IoU<0.5)，并且通过平均嵌入的方式能够更为精确地模拟背景类所涵盖的语义。另外，对于确保词向量间关系的一致性至关重要，否则可能导致检测性能不理想。W_1 定义的投影可以调整，这是与固定的语义嵌入 W_2 不同的地方，这也有助于自动地更新语义嵌入的权重，使其与视觉特征相匹配，同时 W_1 和 W_2 没有应用非线性激活函数可以避免计算量过大的问题。

底部分支为回归分支，主要负责对边界框进行回归操作。通过添加相应的偏

移量，该分支能够精确匹配预测边框与实际位置，为后续的物体 6D 姿态估计奠定坚实基础。

2.3.4 网络训练

本文训练过程分为两步，第一步训练只使用可见类 χ^S 的训练集训练网络 Faster-RCNN[69]，并学习模型相关参数，Faster-RCNN 网络结构如图 2-8 所示。

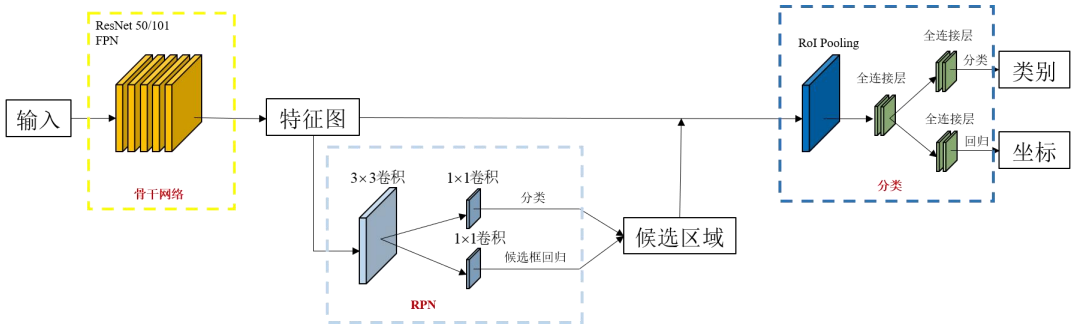


图 2-8 Faster R-CNN 网络结构

Figure 2-8 Faster R-CNN network structure

第二步训练使用可见类图像 χ^S 再次训练，但对 Faster-RCNN 模型进行了修改，将 Faster-RCNN 最后一层的分类分支用语义对齐网络和一个损失函数替换掉。权值 W_1 随机初始化，W_2 固定为对象类的语义向量，在训练过程中不更新。

$y_i \in S'$ 对于每个建议都有相应的真实标签，积极的建议属于可见类 S，而消极的建议只包含背景。在本文中，设置 RPN stride= 16，因此，对于每一幅给定的图像，RPN 都会输出 16 个积极和 16 个消极的建议。现在，当对象建议通过 ROI 池和随后的密集层传递时，将为每个 ROI 计算一个特征表示 f_i。该特征被转发到分类分支和回归分支。总体损失是这两个分支中各自损失的总和，即分类损失和边界框回归损失，损失定义为公式 2-10。

$$L(o_i, b_i, y_i, b_i^*) = \arg\min_{\Theta} \frac{1}{T} \sum_i (L_{cls}(o_i, y_i) + L_{reg}(b_i, b_i^*) \tag{2-10}$$

其中 Θ 为网络参数，o_i 为分类分支输出，$T = N \times R$ 为 N 幅图像训练集中的 ROI 总数。b_i 和 b_i^* 分别是预测的和真实的边界框的参数化坐标，y_i 表示第 i 个对象建议的真实标签。

分类损失 L_{cls} 如公式 2-11 所示。

$$L_{cls}(o_i, y_i) = \lambda L_{mm}(o_i, y_i) + (1-\lambda) L_{mc}(o_i, g(y_i)) \tag{2-11}$$

其中，超参数λ控制了两个损失之间的权衡。我们通过执行传统上看到的目标检测任务来修复它。我们为此使用了 ILSVRC 检测数据集的验证集。

损失同时涉及已知类别和未知类别，包括最大边际损失L_{mm}和元类聚类损失L_{mc}如公式，分别为公式 2-12 和 2-13。

$$L_{mm}(o_i, y_i) = \frac{1}{|C \setminus y_i|} \sum_{c \in C \setminus y_i} \log(1 + \exp(o_c - o_{y_i})) \tag{2-12}$$

$$L_{mc}(o_i, g(y_i)) = \frac{1}{|M'' || Z|} \sum_{c \in M''} \log(1 + \exp(o_c - o_j)) \tag{2-13}$$

其中，集合$M'' = \{M \setminus z_{g(y_i)}\}$和$Z = \{z_{g(y_i)}\}$，$o_k$表示类$k \in S$的预测响应。$L_{mm}$试图将真类的预测响应从其他类中分离出来。相比之下，L_{mc}将属于不同元类的类进一步分开，并试图将每个元类的对象聚类在一起。L_{mc}损失利用元类定义将类似的类聚类在一起，通过将未知类与类似已知的类相关联，这有助于识别未知类的可视化实例。因此，元类的定义对零样本目标检测中特别有用，其中语义关系非常有助于理解未知类。

在图 2-9 中，使用不同的损失函数进行训练过程中，可视化了修改后的嵌入空间。左边为仅使用最大边际损失L_{mm}进行训练得到的效果，而右边结合使用了L_{mm}和聚类损失L_{mc}，可以明显看出在右边语义相似的类更紧密地联系在一起。

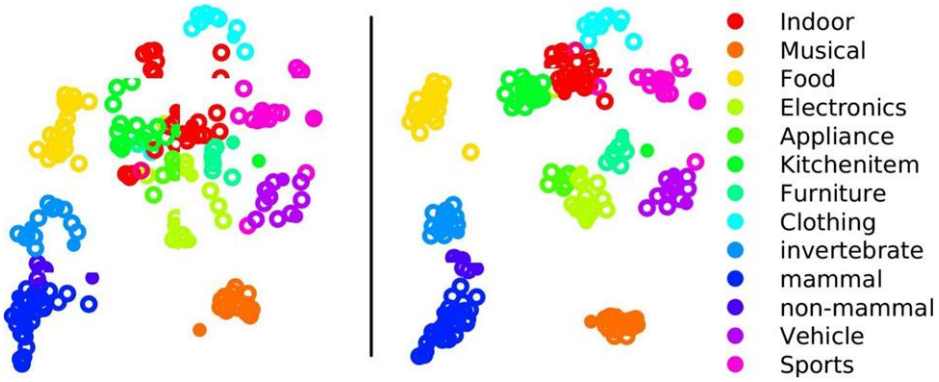


图 2-9 聚类损失对学习嵌入的影响

Figure 2-9 The effect of clustering loss on learning embeddings

元类的设置对学习嵌入的影响是非常大的，比如与动物相关的元类在语义空

间上表现出显著的接近性，而服饰类物体与音乐器具的元类则呈现出较远的距离。这种清晰的语义分离对于提升零样本检测性能具有积极的影响。另外，基于元类的聚类损失并不会对细粒度检测造成负面影响，与总体损失 L_{cls} 的聚类部分 L_{mc} 相比，超参数 λ 的设定更加注重最大裕度损失 L_{mm}，从而确保了检测精度的保持与提升。

回归损失微调了每个已知类 ROI 的边界框。对于每个 f_i，我们得到 $4\times S$ 值，它们代表每个对象实例的边界框的 4 个参数化坐标。基于参数化的真实坐标计算了回归损失。在训练过程中，由于没有视觉示例，没有对背景和未知类别进行边界框预测，但未知类的视觉特征与类似的已知类的视觉特征相关，可以通过一个达到最大响应的密切相关的已知类对象的边界框建议来近似一个未知对象的边界框。使用公式 2-14 对分类分支的每个输出预测值进行归一化。

$$\hat{o}_c = \frac{o_c}{\| v_c \|_2 \| f^t \|_2} \tag{2-14}$$

式 2-14 基本计算了修改后的词向量与图像特征之间的余弦相似度。这种归一化方法将预测值映射到 0 到 1 的范围内。如果最大响应 $\hat{o}_c$ 中的 $c \in C$ 归属于 y_{bg}，则对象建议将其归类为背景；如果一个对象最大预测响应 $\hat{o}_u$，$u \in U$ 且高于阈值 α 则作为未知类，用公式 2-15 可以表达该计算方法。

$$y_u = \arg\max_{u \in U} \hat{o}_u \ \text{s.t.} \hat{o}_u > \alpha \tag{2-15}$$

另一个检测分支找到 b_i，它是可见类的边界框的参数化坐标集。选择了一个预测响应最大的类的边界框，将其作为未知类归属元类的分类依据。对于标记任务，只需使用映射函数 $g(\cdot)$ 就可以为任何看不见的标签分配一个元类。

在一些特殊情况下，甚至可能无法获取未知类类别信息，此时，可以用已知类和背景类字向量来取代已知类、未知类、背景类字向量来作为固定嵌入 $W_2 \in \mathbb{R}^{d\times(S+1)}$ 来训练整个框架，只产生最大边际损失 L'_{mm}，定义如公式 2-16。

$$L'_{mm}(o_i, y_i) = \frac{1}{|S' \setminus y_i|} \sum_{c \in S' \setminus y_i} \log(1 + \exp(o_c - o_{y_i})) \tag{2-16}$$

假设，对于一个对象提议，向量$\mathbf{o}\in\mathbb{R}^{S+1}$只包含已知类和背景类，如果背景类获得最高的分数，我们忽略该建议；对于其他情况，我们将向量$\mathbf{o}$按降序排序计算索引列表$\mathbf{I}$和排序列表$\hat{\mathbf{o}}$，具体如公式 2-17。

$$\hat{\mathbf{o}},\mathbf{I}=sort(\mathbf{o})\ \text{s.t.},o_j=\hat{o}_{lj} \tag{2-17}$$

然后，将$\hat{\mathbf{o}}$的前 K 个分数值(s.t.,$K\le S$)及其对应的词向量使用公式 2-18 合并。

$$\mathbf{e}_i=\sum_{k+1}^{K}\hat{\mathbf{o}}_k v_{li} \tag{2-18}$$

$\mathbf{e}_i$是一个对象建议的语义空间投影，它是由前 K 个已知类类概率加权的单词向量的组合。最后的预测是通过找到$\mathbf{e}_i$和所有未知词向量之间的最大余弦相似度，如公式 2-19。

$$y_u=\arg\max_{u\in U}\cos(\mathbf{e}_i,\mathbf{v}_u) \tag{2-19}$$

2.4 实验与分析

2.4.1 数据集获取

本实验所使用的数据集为 ILSVRC-2017 检测数据集，该数据集包含 200 个对象类别。在训练方面，数据集包括 456567 张图像和 478807 个围绕对象实例的边界框注释。验证数据集包含 20121 张图像，完全注释了 200 个对象类别，其中包括 55502 个对象实例。随机选择了 23 个未知类，其余的 177 个类别被认为是已知类。因为零样本目标检测要求训练集中不包含任何已知类的实例，所得到的测试集有 19008 张图像和 19931 个边界框。在服务器中进行模型的训练，训练的硬件和软件的环境配置如表 2-1 所示。

表 2-1 环境配置

Table 2-1 Environment configuration

模块	环境配置
操作系统	Ubuntu18.04
CPU	Intel Xeon 4210R 2.4GHz
内存	64G
显卡	NVIDIA GeForce A40
CUDA Care	CUDA 11.1/cudnn8.0.4
深度学习框架	Tensorflow2.4

ILSVRC-2017 检测数据集虽然自带层次结构，但并不是树状结构，本文为了方便训练与测试，选择了 M=14，将数据集中 200 类别分为 14 个元类(包括 People)，分别为 lndoor，Musical，Food，Electronics，Appliance，Kitchenitem，Furniture，Clothing，Invertebrate，Mammal，Non-mammal，Vehicle，Sports，People。对于 ILSVRC-2017 检测数据集，每个类的实例数遵循长尾分布，具体如图 2-10 所示。

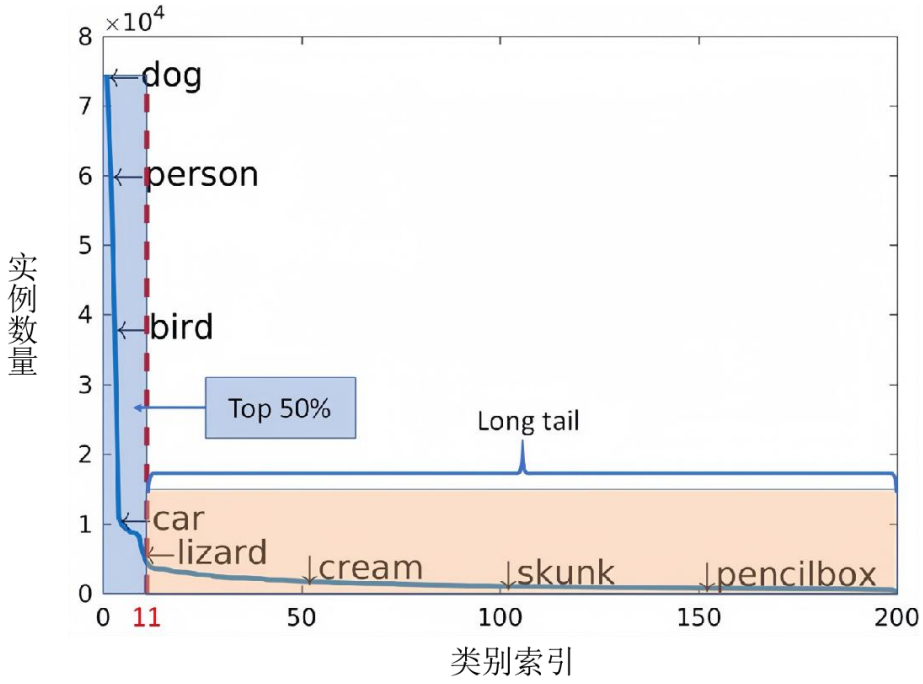


图 2-10 ILSVRC- 2017 数据集长尾分布

Figure 2-10 Long-tail distribution of ILSVRC-2017 dataset

可以看出，200 个类别中只有 11 个高度频繁的类，覆盖了分布的前 50%。这种分布方式会对网络的检测效果影响较大，因此，在训练的第二步中，使用了分布频率较低的类别数据，使每个未知类别和其相似的已知类之间取得平衡。在第一阶段的训练过程中使用了 1000 万个 mini-batches 中准备 280 万个 mini-batches 用于第二阶段的训练，并保证每个未知类从同一元类的其他的已知类中获取至少 10000 个实例，但对于某些未知类，其所属元类中所有类别的实例也不到 10000，但也有未知类物体所属的元类，其实例远超于 10000 个，因此，需要从这些实例丰富的元类中随机挑选 10000 个数据来平衡训练集，280 万个 mini-batches 中剩下的实例则作为背景。

2.4.2 评估指标

本文使用的评估指标为单个未知类的平均精度(AP)和所有未知类的平均精度(mAP)。因为本文中有元类的概念，所以在评估过程中，我们还使用零样本元类检测(ZSMD)精度作为评价的一部分。零样本元类检测是指给定一个测试图像 $x \in \chi^{U}$，目标是定位一个未知类 $u \in U$ 的每一个实例并将其分类为一个元类 $m \in M$。

2.4.3 实验与结果分析

首先，使用 2.3.4 节中预先训练的权值初始化共享层，训练 Fast-RCNN 层的 RPN 和检测网络，将较短的图像尺寸调整为 600 像素，RPN stride=16，三个目标框比例 128，256 和 512 像素，三个高宽比 1：1，1：2 和 2：1，非极大抑制(NMS)=0.7。使用的优化器是 Adam，学习速率为 10^{-5}，β_1=0.9 是用在第一步没有任何数据增强的情况下，β_2=0.999，用于第二步通过重复对象建议进行数据增强后训练网络。

第一阶段训练过程中在没有任何属性文本的情况下使用所有已知类数据训练了一个原始的 Faster-RCNN，在该网络分类层中可以得到一个向量 $\mathbf{o} \in \mathbb{R}^{S+1}$，也可以得到 Faster-RCNN 的检测精度，并选择 ZSOD[70]。模型一起进行数据对比，在损失 L_{cls}、L'_{mm} 下进行实验验证，结果如表 2-2。

表 2-2 各网络的 mAP 值

Table 2-2 MAP values for each network

任务	ZSD					ZSMD				
方法	Faster-RCNN	ZSOD		OURS		Faster-RCNN	ZSOD		OURS	
损失函数	无	L_{cls}	L'_{mm}	L_{cls}	L'_{mm}	无	L_{cls}	L'_{mm}	L_{cls}	L_{cls}
mAP	12.7	15.0	11.4	17.1	16.4	12.9	14.8	11.6	16.8	16.8

可以看出，在 ZSD 任务中，相较于 Faster-RCNN，在训练过程中使用物体属性的 ZSOD 模型和我们的网络都表现较好，因为语义的训练可以增加对未知物体的识别精度，L_{cls} 获得了比 L'_{mm} 更高的 mAP，因为 L_{cls} 在训练过程中同时考虑了未知类的语义和元类的语义；在 ZSD、ZSMD 两个任务中，使用元类的方法比 Faster-RCNN 与 ZSOD 模型都要优异，mAP 值分别是 16.4 与 16.8，因为它

的损失函数L'_{mm}利用了来自元类定义的高级语义关系，这在另两个方法中是不存在的。

在表 2-3 中，我们记录了网络检测每个未知类别的 AP 值，可以看出，与已知类视觉相似的未知类获得了更好的检测效果，在不存在相似的已知类的未知类物体检测中，使用损失函数L'_{mm}的方法表现得更好，分析原因是因为同一元类内的各类别的视觉差异，导致元类注释不能提供足够的监督，使得训练过程中，无法将未知类与已知类联系起来，导致检测效果一般。

表 2-3 实验结果

Table 2-3 Experimental results

网络	Faster-RCNN	ZSOD		OURS	
损失函数	无	L_{cls}	L'_{mm}	L_{cls}	L'_{mm}
p.box	0.0	0.0	5.6	0.0	**7.9**
syringe	3.9	8.0	1.0	**9.2**	1.0
harmonica	**0.5**	0.2	0.1	0.2	0.1
maraca	0.0	**0.2**	0.0	**0.2**	0.1
burrito	36.3	39.2	27.8	**41.2**	28.8
pineapple	**2.7**	2.3	1.7	2.3	2.0
bowtie	1.8	1.9	1.5	**2.3**	1.6
s.trunk	1.7	3.2	1.6	**3.3**	1.7
d.washer	12.2	11.7	7.2	**15.0**	8.1
canopener	2.7	4.8	2.2	**5.3**	2.3
p.rack	**7.0**	0.0	0.0	0.0	0.1
bench	1.0	0.0	4.1	0.0	**4.3**
e.fan	0.6	7.1	5.3	**8.2**	5.6
iPod	22.0	23.3	26.7	**27.8**	27.6
scorpion	19.0	25.7	65.6	27.7	**67.1**
snail	1.9	5.0	4.0	**5.2**	4.0
hamster	40.9	50.5	47.3	**54.9**	48.6
tiger	75.3	75.3	71.5	**79.2**	72.7
ray	0.3	0.0	21.5	0.0	**21.6**
train	28.4	44.8	51.1	45.6	**52.3**
unicycle	**17.9**	7.8	3.7	8.0	3.8
golfball	12.0	28.9	26.2	**30.0**	28.6
h.bar	4.0	4.5	1.2	**4.6**	1.2

具体的检测效果如图 2-11 所示。预测分数阈值 0.3 远低于传统对象检测中使用的值(一般大于 0.5)。因为基于零样本学习的未知物体检测难度高于已知物体，

整个训练过程中没有使用任何未知类的实例，预测的置信度不能像已知类物体那样强。此外，基于零样本学习的未知物体检测方法需要将视觉特征与有噪声的语义词向量相对应，这也降低了对检测效果的整体效果。

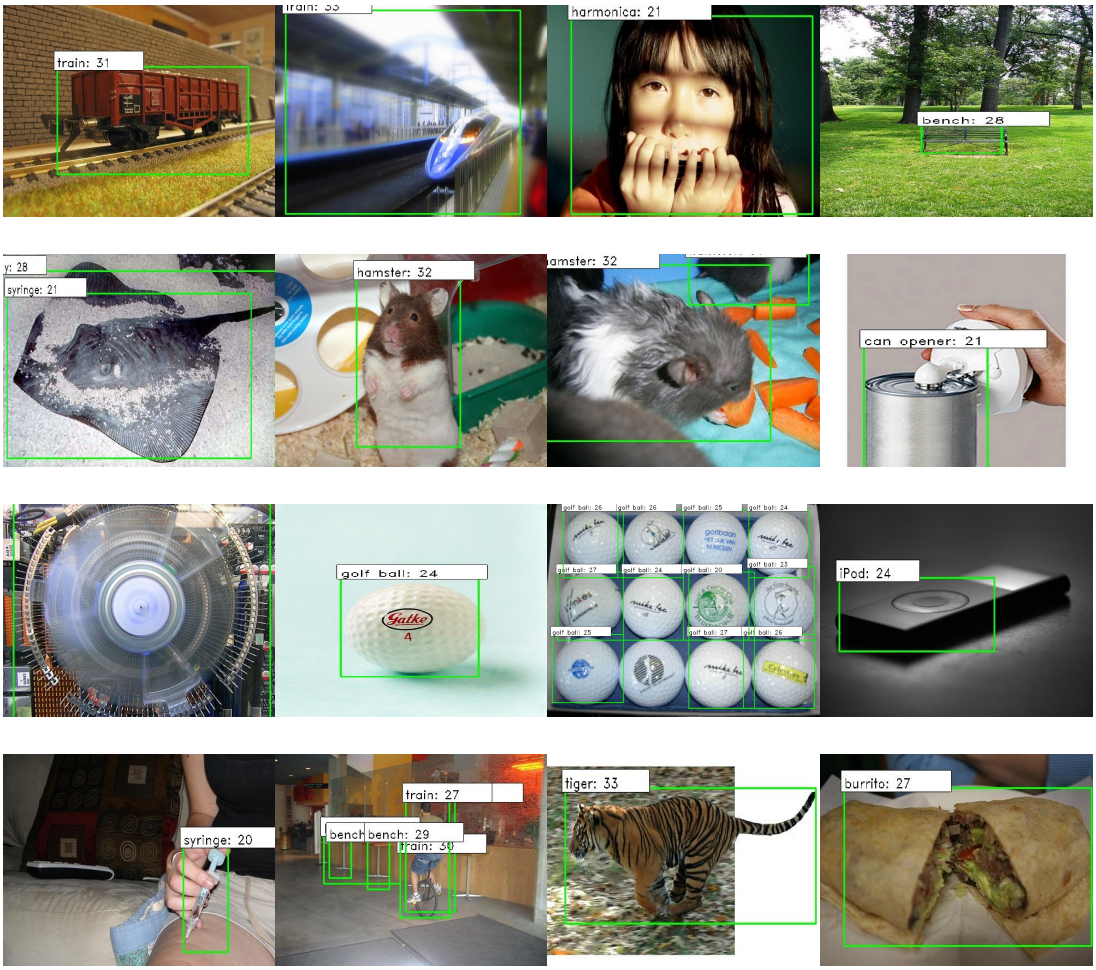


图 2-11 实验结果

Figure 2-11 Experimental results

另外，我们使用了不同的实验设置来比较我们的方法。第一，不使用使用任何预训练模型，训练了最大边际损失为 L_{mm}，训练是进行了相同次数的迭代，即 1000 万个 mini-batches，每个 mini-batches 中有一个图像，得到了 mAP = 5.7 。mAP 低的一个主要原因可能是网络在这些迭代中没有完全收敛。因此，选择了预先训练的主干初始化，这可以加快网络的收敛速度；第二，在从主干训练前排除了所有重叠的未知类后，得到了 mAP = 12.7 。这表明，主干中重叠的存在可以导致在检测设置中得到更高的结果；第三，在这个实验中，使用 softmax 来代替本文的最大边际损失 L_{mm}，得到 mAP = 13.8，而本文中使用最大边际损失 L_{mm} 时

mAP = 15，可以看出，本文使用的最大边际损失更适合于 ZSD，因为它可以以更好的方式对齐视觉特征和语义。相比之下，softmax 损失试图对齐视觉特征及其真实语义，但不能最大限度地将未知类从已知类中分离出来，具体实验结果如表 2-4 所示。

表 2-4 消融实验结果

Table 2-4 Results of ablation experiments

方法	不使用预训练模型	排除所有重叠的未知类	Softmax 损失	OURS
mAP	5.7	12.7	13.8	15

综上所述，基于零样本学习的未知物体检测网络具备在没有任何实例的情况下准确定位未知物体的能力，为后续的机器人抓取未知物体并进行分类提供了研究基础。

2.5 本章小结

在本章中，主要介绍了零样本学习相关知识和基于零样本学习改进的零样本检测相关算法。首先展开了零样本学习与其映射方法的综述，简单讲述了零样本学习识别未知物体的过程；其次本章还介绍基于零样本学习在没有可视化示例的情况下识别和定位未知类的每个实例的检测方法，尤其是对未知物体定位的实现为第 3 章对物体进行 6D 姿态估计研究打下了坚实的基础。

第 3 章 未知物体 6D 姿态估计

3.1 引 言

在现代工业和日常生活中，机器人抓取已成为一项基本任务。物体六自由度(Six Degrees Of Freedom,6D)姿态表示物体坐标系和相机坐标系之间的几何转换，由三维旋转和三维平移组成，6D 姿态估计是机器人抓取与操作的关键步骤之一，本章在第 2 章检测定位到未知物体的基础上对该物体进行多角度图像获取，通过局部特征匹配率先得到足够多的匹配点作为 ICP 精准配准的初始点云，再进行 ICP 精准配准得到目标物体的 6D 姿态。

3.2 网络框架介绍

我们的目标是训练一个不需要任何人工注释、快速且高精度的网络，使其能够直接得到物体的 6D 姿态。为了实现这一点，我们使用 RGB-D 传感器从多个方位拍摄获取对象的基准图片，将这些图片作为基准与机器人坐标系下的图片进行 LoFTR 局部特征匹配[71]获取足够多的特征点并得到像素级的匹配点集，通过 RGB-D 的深度信息得到相应的 3D 点云，作为 ICP[72]配准算法的初始点云，再通过 ICP 将其优化配准得到相对于机器人的 3D 点云模型坐标，具体流程如图 3-1 所示。

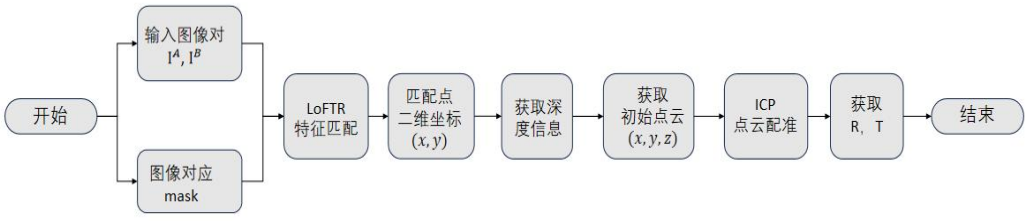


图 3-1 物体位姿获取流程

Figure 3-1 Process For obtaining the pose

3.2.1 LoFTR 局部特征匹配

首先，我们使用 LoFTR 局部特征匹配获得较好的匹配点集，具体实现过程见图 3-2。

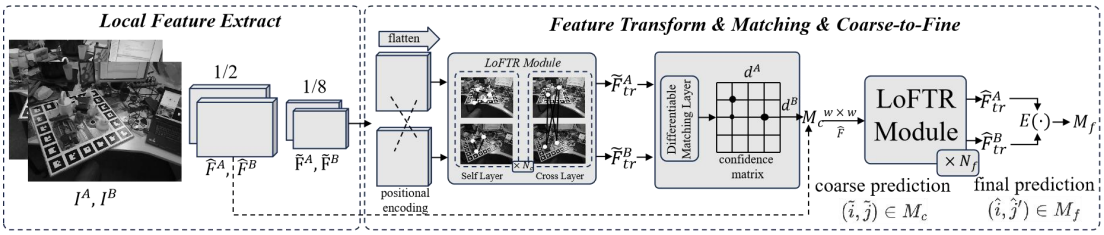


图 3-2 LoFTR 局部特征匹配

Figure 3-2 Matching LoFTR local features

LoFTR 局部特征匹配方法使用了 Transformer[73]模块中的自注意层和互注意层来处理卷积网络中分离的密集局部特征。

Transformer 是一种基于注意力机制的序列模型，如图 3-3 所示。与传统的循环神经网络(RNN)和卷积神经网络(CNN)不同，Transformer 仅使用自注意力机制(self-attention)来处理输入序列和输出序列，因此可以并行计算，极大地提高了计算效率。下面是 Transformer 的详细解释。

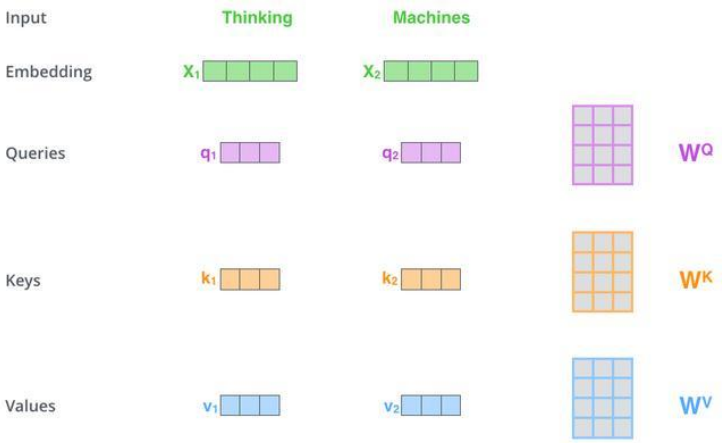


图 3-3　Transformer 模型

Figure 3-3 Transformer model

自注意力机制是 Transformer 的核心部分，它允许模型在处理序列时，将输入序列中的每个元素与其他元素进行比较，以便在不同上下文中正确地处理每个元素。自注意力机制中有三个重要的输入矩阵：查询矩阵 Q(query)、键矩阵 K(key)和值矩阵 V(value)。这三个矩阵都是由输入序列经过不同的线性变换得到的，如图 3-4 所示。

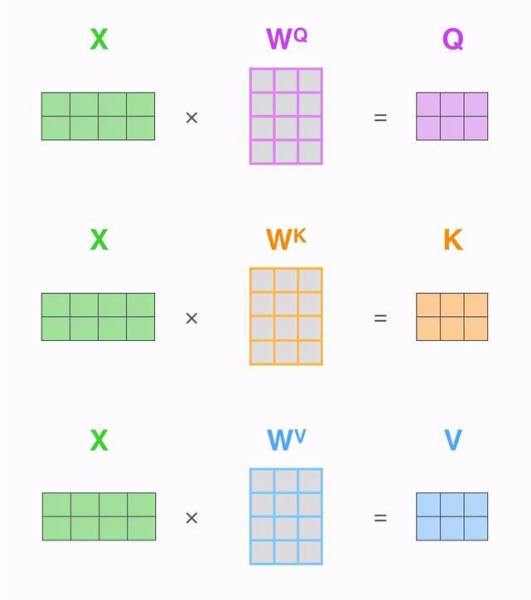


图 3-4 输入矩阵

Figure 3-4 Input matrix

然后，查询矩阵 Q 与键矩阵 K 的乘积经过一个 softmax 函数，得到一个与输入序列长度相同的概率分布，该分布表示每个元素对于查询矩阵 Q 的重要性。最后，将这个概率分布乘以值矩阵 V 得到自注意力向量，表示将每个元素的值加权平均后的结果，如 3-5 所示。

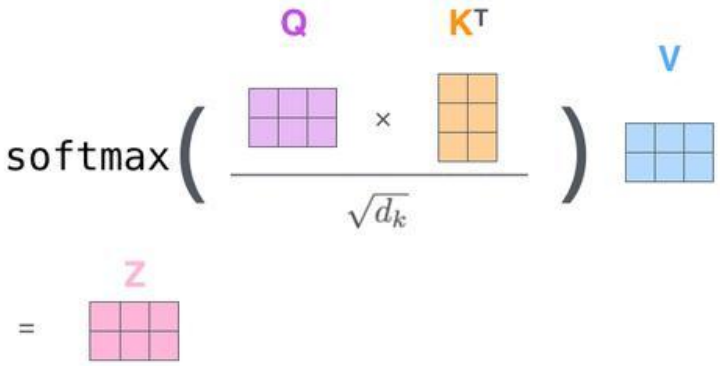


图 3-5 概率分布计算

Figure 3-5 Probability distribution calculation

CrossAttention 实际上是指编码器和解码器之间的交叉注意力层。在这一层中，解码器会对编码器的输出进行注意力调整，以获得与当前解码位置相关的编码器信息，如图 3-6 所示。

CrossAttention 的计算过程：

(1)编码器输入(通常是来自编码器的输出)：它们通常被表示为 enc_inputs，

大小为(batch_size, seq_len_enc, hidden_dim)。

(2)解码器的输入(已生成的部分序列)：它们通常被表示为 dec_inputs，大小为(batch_size, seq_len_dec, hidden_dim)。

(3)解码器的每个位置会生成一个查询向量(query)，用来在编码器的所有位置进行注意力权重计算。

(4)编码器的所有位置会生成一组键向量(keys)和值向量(values)。

(5)使用查询向量(query)和键向量(keys)进行点积操作，并通过 softmax 函数获得注意力权重。

(6)注意力权重与值向量相乘，并对结果进行求和，得到编码器调整的输出。

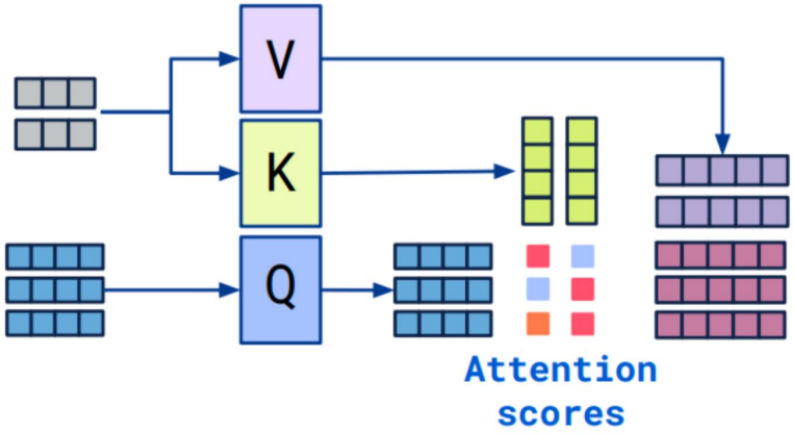


图 3-6 交叉注意力层

Figure 3-6 Cross attention

Transformer 的在 LoFTR 局部特征匹配中应用大大提高了匹配效率，接下来将具体介绍 LoFTR 局部特征匹配过程。

首先，通过从低特征分辨率(仅为 1/8 图像维度)的密集匹配中，选择具有高匹配置信度，并利用基于相关的方法将它们细化到高分辨率亚像素级别，转换成具有反映上下文和位置信息的特征。此外，多次运用自注意力层和互注意力层来学习匹配优先级，使其在低纹理、运动模糊或图像模式重复的区域，能够生成数量充足、质量优良的匹配结果。

在 Linemod[79]数据集上的具体匹配效果如图 3-7 所示，Linemod 数据集将在实验阶段介绍，接下来将具体阐述分析该模块网络。

图 3-7　LoFTR 局部特征匹配效果

Figure 3-7 The effect of LoFTR

使用 FPN 标准卷积结构从图像对 I^A、I^B 中提取原始图像维度 1/8 的粗粒度特征图 $\tilde{F}^A$ 和 $\tilde{F}^B$ 以及原始图像维度 1/2 的细粒度特征图 $\hat{F}^A$ 和 $\hat{F}^B$。为了提取与位置和上下文相关的局部特征，将 $\hat{F}^A$ 和 $\hat{F}^B$ 展平为一维向量并添加位置编码，然后输入 LoFTR 模块中的 Self-attention 层、Cross-attention 层进行处理，并交错重复 Nc 次输出易于匹配的特征 $\tilde{F}_{tr}^A$ 和 $\tilde{F}_{tr}^B$，计算两者之间的得分矩阵 $\mathbf{S}$，如公式 3-1 所示。

$$\mathbf{S}(i,j) = < \tilde{F}_{tr}^A(i), \tilde{F}_{tr}^B(j) > \tag{3-1}$$

其中 $<\cdot,\cdot>$ 表示内积。

使用 SuperGlue 中的最优传输层(OT)进行匹配，同时在 $\mathbf{S}$ 的两个维度上应用 softmax 得到一个置信矩阵 $\mathbf{P_c}$，如公式 3-2 所示。

$$\mathbf{P_c}(i,j) = \text{soft}\max(\mathbf{S}(i,\cdot))_j \cdot \text{soft}\max(\mathbf{S}(\cdot,j))_i \tag{3-2}$$

在公式 3-2 的基础上可以选择置信度高于 θ_c 的匹配，并进一步执行最近邻(MNN)准则，得到粗粒度匹配 M_c，如公式 3-3 所示。

$$M_c = \{(\tilde{i},\tilde{j}) \mid \forall(\tilde{i},\tilde{j}) \in \text{MNN}(\mathbf{P_c}), \mathbf{P_c}(\tilde{i},\tilde{j}) \geq \theta_c\} \tag{3-3}$$

对于每一个粗粒度匹配 $(\tilde{i},\tilde{j})\in M_c$，首先在细粒度特征映射 $\hat{F}^A$ 和 $\hat{F}^B$ 上确定 $(\hat{i},\hat{j})$ 的位置，然后裁剪出大小为 $w\times w$ 的局部窗口，将其输入一个较小的 LoFTR 模块(N_f层)，生成以 $\hat{i}$ 和 $\hat{j}$ 为中心的局部特征图 $\hat{F}^A_{tr}(i)$ 和 $\hat{F}^B_{tr}(j)$。

然后将 $\hat{F}^A_{tr}(i)$ 的中心向量与 $\hat{F}^B_{tr}(j)$ 中的所有向量计算相关，并生成一个 heatmap，通过计算概率分布的期望，可以得到在在 I^B 上具有亚像素精度的最终位置 $\hat{j}$。最终产生细粒度的匹配项 M_f。LoFTR 的匹配结果如图 3-8 所示。

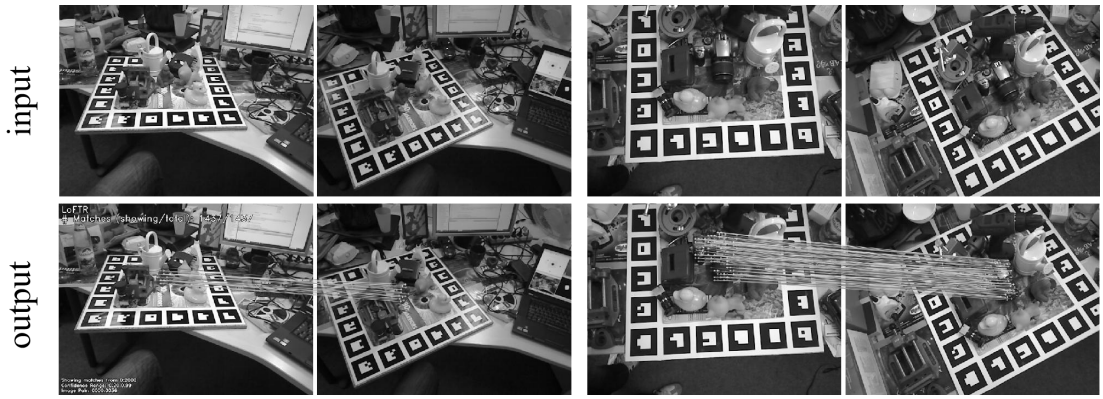


图 3-8 LoFTR 局部特征匹配结果

Figure 3-8 The result of LoFTR matching

3.2.2 ICP 精准配准

在得到最终匹配预测之后再利用传感器的固有矩阵，将匹配点转换为包含三维几何信息的点云 $A=\{a_1,a_2,...,a_n\}$ 与 $B=\{b_1,b_2,...,b_n\}$，将其作为配准的初始点云，然后引入精细配准 ICP 算法，实现目标的精细配准，获得更准确的目标姿态信息，具体流程如图 3-9 所示。

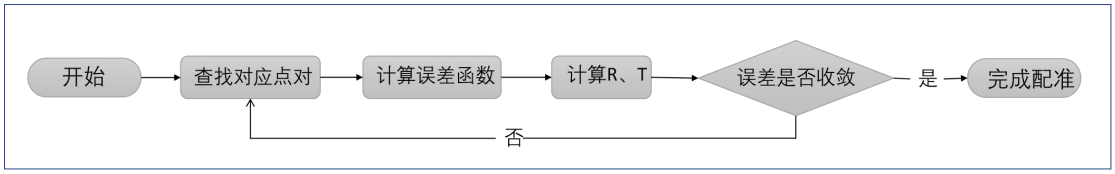


图 3-9 ICP 精细配准流程

Figure 3-9 ICP fine registration process

ICP 算法的核心是迭代计算位姿变换矩阵，使源点集 A 与目标点集 B 之间的误差函数最小，从而得到最优解。点云 A 中的点 a_i 找到与点云 B 的欧氏距离最短

的点b_i，将a_i和b_i作为相应的点对，计算变换矩阵，然后计算配准误差$\mathrm{E}(\mathbf{R},\mathbf{T})$，如公式 3-4 所示。本文的目标则是旋转矩阵$\mathbf{R}$和平移向量$\mathbf{T}$使得$\mathrm{E}(\mathbf{R},\mathbf{T})$最小。

$$\mathrm{E}(\mathbf{R},\mathbf{T})=\frac{1}{\mathrm{N}}\sum_{i=1}^{n}\|b_i-(\mathbf{R}a_i+\mathbf{T})\|^2 \tag{3-4}$$

式中，n 为点云数量，接下来将阐述具体计算过程。

首先，根据点云坐标信息，计算对应的重心，如公式 3-5。

$$\begin{cases}\mu_a=\dfrac{1}{\mathrm{N}}\sum_{i=1}^{n}a_i\\ \mu_b=\dfrac{1}{\mathrm{N}}\sum_{i=1}^{n}b_i\end{cases} \tag{3-5}$$

计算旋转矩阵和平移向量及其误差函数$\mathrm{E}(\mathbf{R},\mathbf{T})$，使得其值最小。

预先设定阈值ε和最大迭代次数N_{max}，将上一步的变换矩阵作用于点云A，得到新的待配准点云A'，计算A'和B的距离误差 d，见公式 3-6，若$\mathrm{d}<\varepsilon$或者当前迭代次数大于N_{max}，则结束算法，否则将A'设置为新的要进行配准的点云，重复上述步骤，直到满足算法结束的条件。

$$\mathrm{d}=\frac{1}{\mathrm{N}}\sum_{i=1}^{\mathrm{N}}\|A'-B\|^2 \tag{3-6}$$

针对点云配准和特征提取网络的特点，本文采用的损失函数包括 2 个部分：两点云之间倒角距离、预测和真实变换矩阵的误差、源点云和目标点云的局部几何结构特征差等属性。本文采用两个点云之间的倒角距离L_{AB}损失来刻画两个点云之间的空间距离，如公式 3-7 所示。

$$\mathrm{L}_{AB}=\frac{1}{\mathrm{N}}\sum_{a_i\in A}\min_{b_i\in B}\|A-B\|_2^2+\frac{1}{\mathrm{N}}\sum_{b_i\in B}\min_{a_i\in A}\|B-A\|_2^2 \tag{3-7}$$

式中，a_i是点云A中的一个点，b_i是点云B中的一个点，N 是点云数量。在两个点云提取特征后经过平均池化得到点云平均特征，计算两个平均特征的欧几里得距离L_{D}，如公式 3-8 所示。

$$\mathrm{L}_{\mathrm{D}}=\|f(A)-f(B)\|_2 \tag{3-8}$$

式中，$f(\cdot)$为平均池化得到的点云平均特征。

所以，最终的损失函数L_{tatal}如公式 3-9 所示。

$$\mathrm{L}_{tatal}=\mathrm{L}_{AB}+\mathrm{L}_{D} \tag{3-9}$$

3.3 实验与分析

3.3.1 数据集与评价指标

网络的训练在服务器(操作系统:Ubuntu 18.04 LTS，CPU:Intel Xeon 4210 R 2.4GHz，GPU:NVIDIA GeForce A40)上进行。

本文采用公共数据集 Linemod 对网络进行验证。Linemod 数据集已被广泛应用于深度学习目标姿态估计领域。该数据集由单目标和多目标两个子数据集组成，每个子数据集包含 13 个子集，每个子集中提供了 1300 张与该目标图像相关联的虚拟 3D 控制点文件以及目标 3D 模型。针对每个控制点文件，6D 姿态和实例语义分割掩码均已被注释标记。该数据集是用于评估 6D 姿态估计任务的基准数据集之一，许多方法都采用该数据集进行评估[83-84]。

此外，为了验证在有遮挡环境下的物体 6D 姿态估计效果，本文还引入了 Occlusion Linemod 数据集。这一数据集是在 Linemod 数据集的基础上，通过为每个场景添加详细的注释而创建的，旨在模拟不同程度的遮挡情况。每个图像中的遮挡程度都各不相同，从而提供了丰富的遮挡场景样本，有助于全面评估我们的算法在有遮挡条件下的性能表现。

六维物体姿态估计中常用的四个标准指标，其中一个度量标准是估计姿态的平均距离(ADD)，它主要测量地面真实姿态和估计姿态之间的平均距离，在 Linemod 数据集中用于评价非对称物体。ADD 的计算公式如公式 3-10 所示。当平均距离小于误差值时，认为估计的 6D 姿态是正确的。本实验将误差阈值设置为对象模型最大直径的 10%。

$$\mathrm{ADD}=\frac{1}{\mathrm{N}}\sum_{x\in M}\|(\mathbf{R}x+\mathbf{T})-(\hat{\mathbf{R}}x+\hat{\mathbf{T}})\| \tag{3-10}$$

式中，$[\mathbf{R},\mathbf{T}]$ 表示地面真实姿态，$[\hat{\mathbf{R}},\hat{\mathbf{T}}]$ 表示估计的姿态。M 表示对象坐标系中的对象三维模型，x 是属于模型 M 的点，N 是点云数量。由于对称物体旋转角度的不确定性，我们采用平均最近点距离(ADD-S)作为对称物体的评价准则。ADD-S 计算三维地面真实模型上最近点与预测模型之间的平均距离，具体定义如公式 3-11。

$$ADD-S=\frac{1}{\mathrm{N}}\sum_{x1\in M}\min_{x2\in M}\|(\mathbf{R}x_1+\mathbf{T})-(\hat{\mathbf{R}}x_2+\hat{\mathbf{T}})\| \tag{3-11}$$

3.3.2 实验结果与分析

为验证本文姿态估计的效果，将本文算法与目前物体 6D 位姿估计方法中较先进且常用的模型进行了比较。使用的模型分别为 BB8[74]模型、GDR-NET[75]模型及 PVNet[76]模型，比较结果如表 3-1 所示。

表 3-1　Linemod 上不同算法的 ADD(S)

Table 3-1 ADD(S) of different algorithms on Linemod

方法	BB8 模型	GDR-NET 模型	PVNet 模型	Ours 模型
Ape	40.4	38.2	**43.6**	43.1
Benchvise	89.1	84.1	**99.9**	89.3
Camera	84.2	82.3	**86.8**	86.1
Can	92.6	90.7	**95.5**	94.2
Cat	46.3	50.0	**79.3**	68.2
Drill.	68.2	68.1	**96.4**	74.5
Duck	32.1	34.2	**52.5**	47.2
Egg.	96.4	93.9	**99.2**	97.1
Glue	91.2	91.1	**95.7**	93.3
Hole	79.5	77.8	**82.0**	80.4
Iron	90.8	88.2	**98.9**	91.0
Lamp	83.7	80.3	**99.3**	87.3
Phone	84.6	73.0	**92.4**	87.6
Mean	75.3	73.2	**86.3**	80.0

其中，因为 PVNet 网络充分使用了 Linemod 数据集中的真实数据进行训练，所以该模型的效果远好于 BB8 模型、GDR-NET 模型以及本文的模型，但本文的模型在 Ape、Camera、Can、Glue 等物体类别中都获得了与 PVNet 模型相近 ADD(S)值，本文的模型相较于 BB8 模型，ADD(S)提高了 4.7%，相较于 GDR-NET 模型，ADD(S)提高了 6.8%，但在 Cat、Duck 等弱纹理物体上，本文的模型效果非常优异，比 BB8 模型分别提高了 21.9%、15.1%，比 GDR-NET 分别提高了 18.2%、13%，说明本文的模型在物体纹理及几何细节较弱时也表现优异。分析原因是本文中所采用的 LoFTR 局部特征匹配方法在设计上汲取了 Transformer 的优势，并借助自注意层和互注意层获得了两幅图像的特征描述符，从而达到了更为准确全面的效果。此种方法带来的全局接受域特性有效提升了网络性能，让网络在纹理较弱、细节少的区域同样可以实现密集匹配。Linemod 数据集结果的可视化如图 3-10 所示，根据旋转矩阵 **R** 和平移向量 **T** 移动模型中的点，并将其投影到 RGB 图像上。

图 3-10　Linemod 数据集可视化

Figure 3-10 Linemod data set visualization

因为 RGB-D 相机在黑暗或强光条件下捕捉图像，只影响 RGB 图像，而深度图像没有影响，为了验证本文的网络在光照变化下的鲁棒性，本文通过改变同一场景中 RGB 图像的亮度值进行实验。

在光照变化条件下的 6D 姿态估计可视化结果如图 3-11 所示，经过比较可以发现，该网络在应对光照变化条件下的姿态估计难题时，同样表现出了良好的性能。

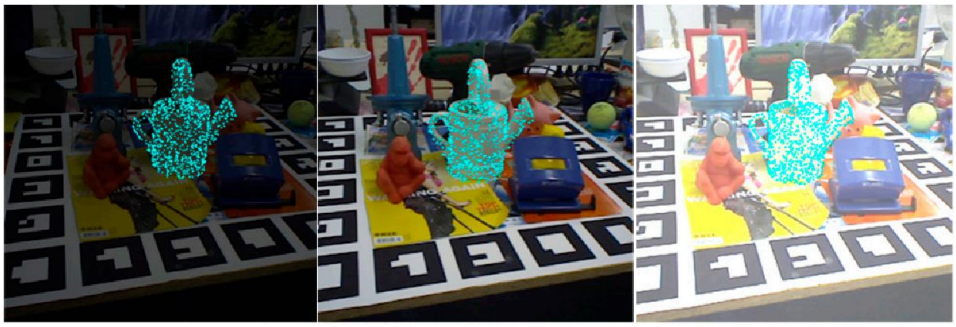


图 3-11 光照变化下姿态估计结果

Figure 3-11 Results of pose estimation under changes in lighting conditions

表 3-2 显示了本文网络在 Occlusion Linemod 数据集的实验结果，图 3-12 为 Occlusion Linemod 数据集结果的可视化。本文网络估计的 6D 姿态结果表明在 Linemod 数据集上 ADD(S)值为 80.0%，在 Occlusion Linemod 数据集上 ADD(S)值为 78.0%，ADD(S)指标有轻微下降，经过分析是因为注意力机制的网络生成的注意力图得分高的像素点由于物体存在遮挡，使得对应关系存在偏差。虽然因为 PVNet 数据集进行了真实样本训练表现最优越，但本文模型的 ADD(S)值比

BB8 模型高了 3.6%，比 GDR-NET 模型高了 5.2%。证明本文的模型在有严重的遮挡和背景杂波环境下对物体 6D 姿态估计的性能相较于其他模型表现更优异。

表 3-2　Occlusion Linemod 上不同算法的 ADD(S)

Table 3-2 ADD(S) of different Occlusion Linemod algorithms

方法	BB8 模型	GDR-NET 模型	PVNet 模型	Ours 模型
Ape	39.4	38.0	**42.3**	42.1
Benchvise	87.3	83.8	**96.7**	88.9
Camera	83.2	81.6	**85.4**	85.1
Can	91.9	90.4	**94.5**	93.9
Cat	45.3	49.5	**78.1**	67.9
Drill.	67.1	67.8	**94.9**	73.8
Duck	31.0	33.7	**51.7**	47.1
Egg.	95.7	93.5	**98.3**	96.5
Glue	90.6	91.0	**94.2**	93.0
Hole	78.5	76.9	**81.3**	80.1
Iron	89.7	88.1	**97.6**	90.4
Lamp	83.4	79.6	**98.4**	86.9
Phone	84.2	72.4	**91.0**	87.3
Mean	74.4	72.8	**85.0**	78.0

图 3-12　Occlusion Linemod 数据集可视化

Figure 3-12 Occlusion Linemod data set visualization

另外，本文还对网络在噪声下的物体 6D 姿态估计精度做了验证实验。众所周知，当 RGB-D 相机在真实环境中捕捉图像时，噪声对深度图像的影响非常大。深度图像中每个像素的数据是从物体到相机的距离，以 mm 为单位。本文通过在 Linemod 数据集的深度图像的每个像素中添加随机噪声，随机噪声数范围为 (−25，25)，以此来验证本文的 6D 姿态估计网络对噪声的鲁棒性。表 3-3 显示了

该网络在随机噪声增加条件下的性能，可以看出，该网络在噪声情况下也有较好的精度。

表 3-3　随机噪声下网络精度

Table 3-3 Network accuracy under random noise

噪音范围/mm	0	5	10	15	20	25
精度	99.4%	99.4%	99.3%	99.2%	99.1%	99.1%

3.4 本章小结

本章面对复杂非结构化环境下面对未知物体位姿估计困难的问题，研究设计了融合局部特征匹配与点云配准的姿态估计网络以用于未知物体的姿态估计。首先，利用 LoFTR 局部特征匹配算法获取足够多的特征点并得到像素级的匹配点集，通过 RGB-D 的深度信息得到相应的 3D 点云，并将其作为 ICP 配准算法的初始点云优化配准效果。实验表明，该方法可以有效避免 ICP 算法过度依赖初始点云的问题，加快了迭代效率，提高了配准的精度与速度。由于点云配准是 6D 姿态估计的重要步骤，解决了该问题可以有效提升未知对象 6D 姿态估计的效率。

第 4 章 抓取系统搭建与实验

4.1 引 言

针对本文提出的基于零样本学习的未知物体抓取方法，我们搭建了机器人抓取系统，并进行了相关实验以验证其可行性。考虑到抓取任务的复杂性和灵活性，我们选择了一款双臂协作型机器人 Baxter 机器人进行实验验证。同时，我们还根据第二章和第三章提出的未知物体检测网络和抓取候选生成网络进行了实验验证。

4.2 机器人抓取的实现

4.2.1 机器人抓取实验平台介绍

本文的机器人抓取实验平台如图 4-1 所示，由上位机、Kinect 相机、Baxter 机器人组成。为了收集抓取场景中的多样化数据，我们将利用 Kinect 摄像头进行采集。随后，Baxter 机器人将负责接收来自计算机的数据，这些数据涵盖了目标处理、抓取位置参数的优化选择等关键信息。基于这些数据，机器人将操纵其机械臂精准地执行对目标的抓取操作。

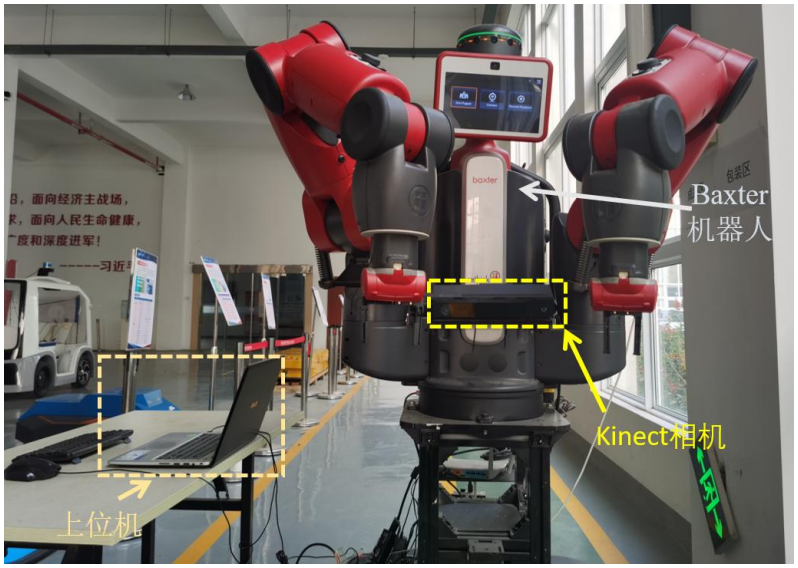


图 4-1 机器人抓取实验平台

Figure 4-1 Robot grasping experimental platform

本文进行实验验证时，所采用的上位机系统与前文离线训练、实验所使用的

计算机配置相同。表 4-1 是该计算机平台的具体配置信息。

表 4-1 计算机配置

Table 4-1 Computer configuration

模块	环境配置
操作系统	Ubuntu18.04
CPU	Intel Xeon 4210R 2.4GHz
内存	64G
显卡	NVIDIA GeForce A40
CUDA Care	CUDA 11.1/cudnn8.0.4
深度学习框架	Tensorflow2.4

Baxter 是一个双臂协作机器人，具体信息如表 4-2 所示。

表 4-2 Baxter 部件信息

Table 4-2 Baxter part information

部件	详细介绍
单臂自由度	7
单臂展开距离	104 cm
定位精度	+/- 5 mm
最大末端负载	2.2 KG
手指长度	62 mm
手指宽度	15 mm
开合宽度	3-11 cm
抓取力度	10 N

在机器人的每个手指末端都特别配备了橡胶衬套，强化了两者间的摩擦力，从而显著提升了抓取效果。在机器人的左手腕部装配了 Kinect 摄像头，并巧妙地利用其右手执行抓取动作，可以准确获取待抓取物体的三维图像信息，这样的布局使得机器人在执行抓取任务时，能够更精确地识别并锁定目标物体。为进一步优化抓取操作的便利性，将机器人的抓取任务设定在与其底座处于相同高度的桌面上进行，这样的设计不仅提高了机器人抓取任务的便捷性，也有效提升了其执行效率。由机器人操作系统(ROS)处理机器人与上位机之间的通信，并使用无碰撞逆运动学求解器 IK[77]求解逆运动学。

为了确保机器人在执行抓取任务时能够准确识别并定位目标物体，需要将其位姿坐标从现实世界中的坐标系转换成机器人坐标系下是非常重要的，接下来将详细阐述对 Kinect 相机进行位姿标定的相关计算方法。

相机标定目的是将相机坐标系下的目标物体数据转换至机器人坐标系下。在机器人视觉系统中，深度相机的安装方式主要分为眼在手与眼在外两种模式。本

文选择了眼在手模式，即将深度相机安装于机器人的末端关节处，这种安装方式能够确保相机与机器人末端执行器的紧密配合，从而提高视觉信息的实时性和准确性。Kinect 深度相机安装如图 4-2 所示。

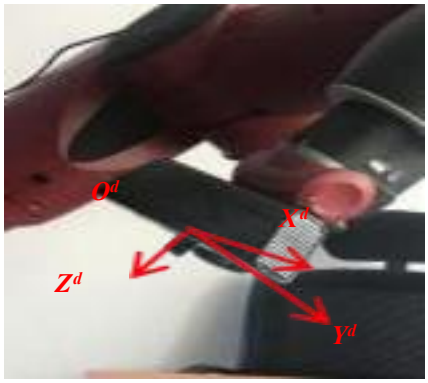


图 4-2 眼在手

Figure 4-2 Eye in hand

在本文的机器人抓取系统有三个坐标系，分别为机器人基坐标系 $O_rX_rY_rZ_r$，相机坐标系 $O_dX_dY_dZ_d$，夹持器坐标系 $O_gX_gY_gZ_g$，如图 4-3 所示，各个坐标系方向具体定义如表 4-3 所示。

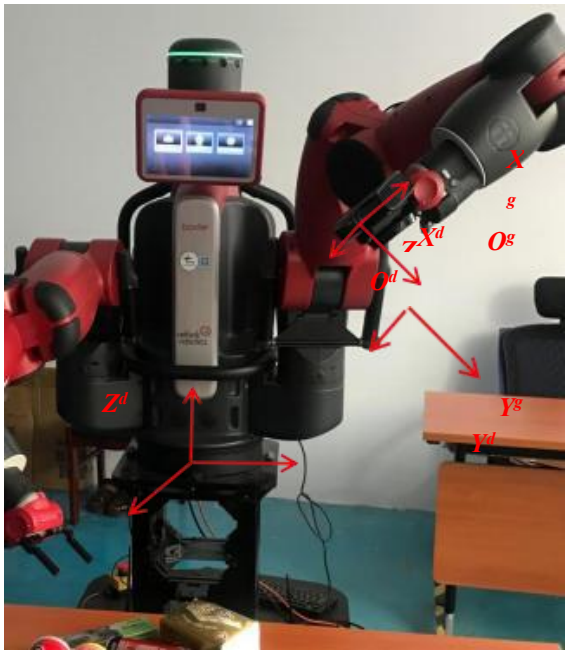


图 4-3 机器人坐标系

Figure 4-3 Robot coordinate system

表 4-3 坐标系表示方式及定义

Table 4-3 Coordinate system representation and definition

坐标系	各点的表示	方向备注
$O_rX_rY_rZ_r$	F^r	X_r 正方向：机器人前方 Y_r 正方向：机器人左方 Z_r 正方向：机器人上方 原点 O_r：基座平面中心
$O_dX_dY_dZ_d$	F^d	X_d 正方向：与成像平面坐标系 X 轴平行 Y_d 正方向：与成像平面坐标系 Y 轴平行 Z_d 正方向：相机朝向的方向 原点 O_d：相机光心位置
$O_gX_gY_gZ_g$	F^g	X_g 按照右手笛卡尔坐标系确定 Y_g 平行于夹持器开合方向且指向机器人的左侧 Z_g 正方向：指向远离夹持器的方向 原点 O_g：夹持器夹爪末端中间位置

其中相机坐标系中提到的成像平面坐标系指的是 Kinect 深度相机成像过程中的坐标系，具体如图 4-4 所示。

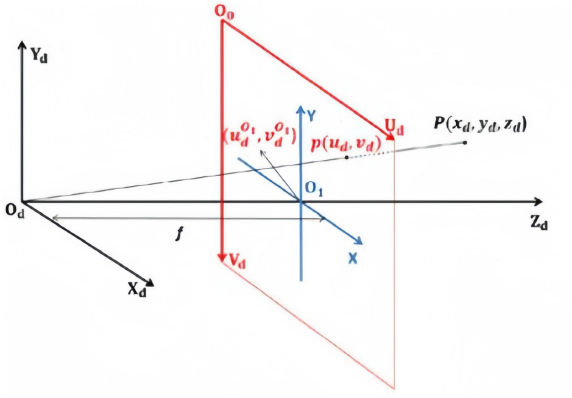


图 4-4 Kinect 深度相机成像过程

Figure 4-4 Kinect depth camera imaging process

Baxter 利用开发函数可以获取其夹爪末端位姿的信息，并同时获取 Kinect 相机的点云数据。

经过以上步骤的处理，我们可以得到丰富的点对信息，这些点对精准地反映了机器人坐标系 $F^r=\{q_1,q_2,...,q_n\}$ 和相机坐标系 $F^d=\{p_1,p_2,...,p_n\}$ 之间的相对位置关系。基于这些点对数据，我们可以进一步计算得出坐标系间的旋转矩阵 **R** 和平移向量 **T**，具体的计算过程如式(4-1)所示。

$$(\mathbf{R},\mathbf{T})=\arg_{(\mathbf{R},\mathbf{T})}\min\sum_{i=1}^{n}\|(\mathbf{R}p_i+\mathbf{T})-q_i\|^2 \tag{4-1}$$

式中，n 为点对数，可以得到 F^d 在 F^g 下坐标系下的位姿矩阵 ${}^{r}_{g}\mathrm{T}=(\mathbf{R},\mathbf{T})$，再使用正运动学获得末端夹爪在 F^r 坐标系中的位姿矩阵，得到相机在 F^r 坐标系下的位姿矩阵 ${}^{r}_{d}\mathrm{T}=(\mathbf{R},\mathbf{T})$，如式(4-2)所示。

$${}^{r}_{d}\mathrm{T}={}^{r}_{g}\mathrm{T}\,{}^{g}_{d}\mathrm{T} \tag{4-2}$$

根据上述步骤，得到了目标物体在机器人坐标系下的点云数据，这一转换过程确保了数据的准确性和一致性，为后续的研究和应用提供了可靠的基础。

4.2.2 抓取流程介绍

本文的实验以 Ubuntu 18.04 操作系统为平台，依托机器人控制系统 ROS 实现各模块间的通信机制。ROS 系统是机器人领域应用极广的开源模块化技术，其设计初衷是为机器人开发者提供一个高效且便捷的平台。本文实验根据第二章与第三章的内容展开为图像获取、物体识别、位姿估计以及机械臂抓取规划四个模块，这些模块共同构成了实验的核心架构。

实验借助 ROS 系统的有效运用，成功实现了各模块间的协同作业，确保了实验数据的精准度和实时性。此外，ROS 系统不仅提供了多样化的工具和沟通机制，还展现出强大的跨平台性能，兼容多种编程语言，有效实现了不同模块间的信息交互。图 4-5 详细展示了各模块间的通信关系，进一步验证了 ROS 系统在实验中的优越性和实用性。

在 ROS 通信系统中，功能模块作为关键的通信单元，被称为节点，发挥着至关重要的作用。每个节点都具备独一无二的标识符，可以在控制系统明确其位置。

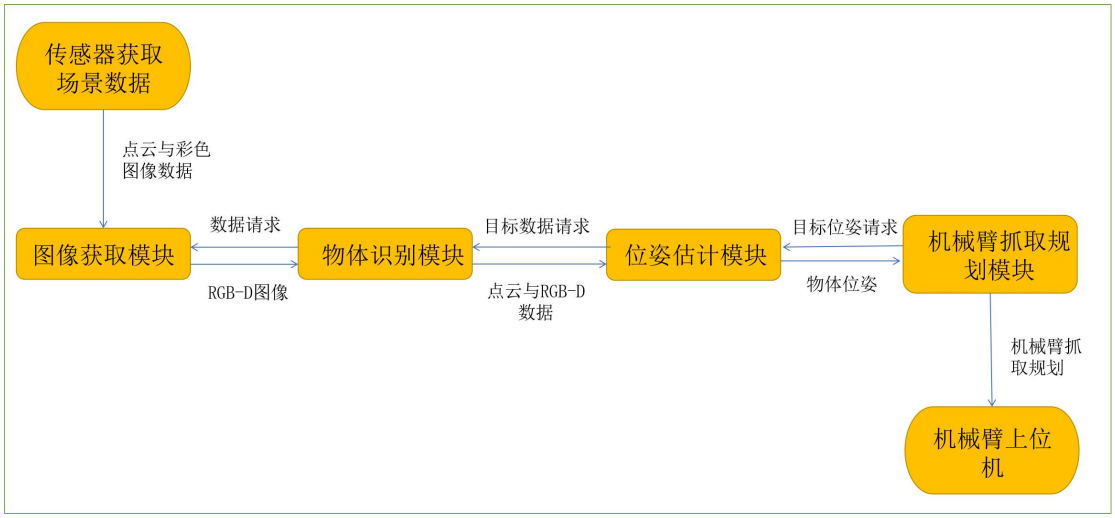


图 4-5 机器人抓取流程

Figure 4-5 Robot grasp process

该系统提供了三种差异化的通信模式：主题发布、服务调用以及动作执行。各节点或是服务端或是客户端，用一样的名称及类型连接。一旦客户端发起服务调用，将执行相应的函数，并将处理结果实时反馈给客户端。另外，ROS 还拥有模拟环境以及可视化工具 Rviz，大大方便了实验过程中的调试及其他功能。

本文主要通过 ROS 系统，对目标物体进行图像获取直至抓取规划的完整操作流程。接下来，将对该流程中展开详尽的阐述。

(1)图像获取模块

本文首先利用 Kinect 传感器获取目标物体的彩色及深度图像。其中，基于零样本学习的未知物体识别和位姿估计中的局部特征匹配均依赖于彩色图像，而位姿估计中的 ICP 精准配准则使用深度图像。

本模块作为服务端，借助 ROS 话题的通信机制，将 Kinect 采集的数据传送至 ROS 系统内部。紧接着，本模块订阅相关话题，根据预设参数对获取的原始数据展开预处理操作，包括了点云滤波、点云下采样等环节。将预处理后的图像数据发送给后续模块，以供进一步的分析和处理。图 4-6 详尽地展示了该模块输入与输出关系上的具体流程。

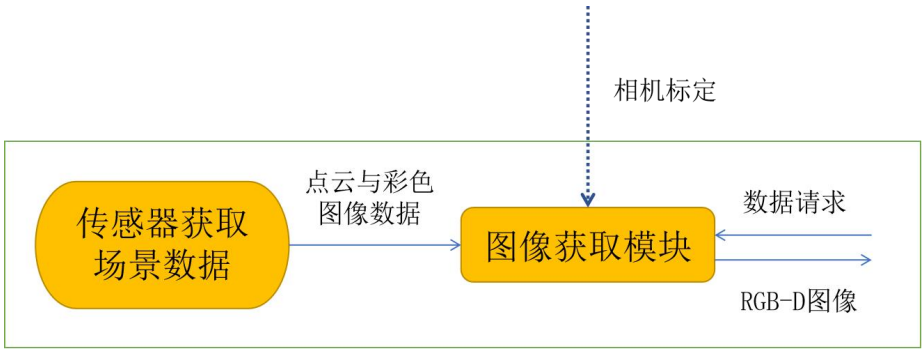


图 4-6 图像获取模块的输入输出关系

Figure 4-6 Input/output relationship of the image acquisition module

(2)物体识别模块

该模块基于改进的零样本检测网络对物体进行定位识别，主要任务在于根据 Kinect 采集的 RGB 图像进行目标物体的识别任务。首先以客户端的身份从前一个模块获取数据，并对接收到的 RGB 数据实施处理，通过我们的算法，可检测出未知物体并输出其精确坐标，进一步将这些物体坐标传输至后一个模块，致力于获取更多目标物体数据，为后续位姿估计工作提供支持。数据具体的传输过程如图 4-7 中所示。

图 4-7 物体识别模块数据传输

Figure 4-7 Data Transmission on the object identification module

(3)位姿估计模块

物体的位姿估计模块通过物体识别模块获取检测到的物体坐标并控制机械臂获取足够多的图像数据，再通过获得的图像对进行局部特征匹配并将其匹配结果用于 ICP 精准匹配，最终获得目标物体相对于机械臂的位姿数据，最后，在获取位姿数据后以服务端的身份将数据传至下一模块，该模块数据的输入输出如图 4-8 所示。

图 4-8 位姿估计模块数据传输

Figure 4-8 Data transmission of the pose estimation module

在本模块中，为了获取尽可能多的各角度目标信息，需对配备 Kinect 深度相机的机器人手臂进行精确的运动控制。在本文实验中，我们设定机器人手臂末端的运动轨迹始终围绕一个固定半径为 80 cm 的球形区域，并确保其沿着半球中心进行 100°的旋转，以确保我们能够从多个视角捕获到全面的三维数据。

首先将 Kinect 深度相机的视角精确地对准桌面工作区的中心，并维持与桌面呈 50°的夹角。机器人手臂的具体运动方向见图 4-9，箭头为手臂的移动路径。通过控制机器人手臂沿着这一方向移动，我们能够引导深度相机采集多帧深度图像数据，可以保证深度相机捕捉到工作区内物体的所有信息。为了防止机械臂在抓取物体过程中出现碰撞情况，需要使用无碰撞轨迹运动[78]方法规划机械臂的运动路径。

图 4-9 Kinect 深度相机运动路径

Figure 4-9 Kinect depth camera motion path

(4)机械臂抓取规划模块

根据图 4-10 的展示，该模块的工作流程如下：首先，该模块作为客户端从

上一模块中获取物体位姿，基于这些信息，本模块进一步生成目标物体的抓取位姿以及机械臂运动轨迹，来实现对抓取任务的精确控制。

在标定实验的基础上，获得目标物体相对于机械臂的位姿信息，随后，利用抓取位姿以及机械臂运动规划的算法，生成机械臂关节的轨迹，并将相关参数发送至机械臂控制系统。最终，控制系统根据生成的轨迹指令，指导机械臂抓取目标物体。

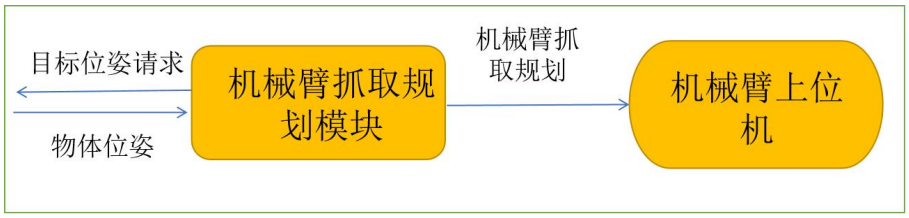


图 4-10 机械臂抓取规划模块

Figure 4-10 The robot arm grabs the planning module

机器人的轨迹规划指的是在获取目标物体的抓取姿态之后，控制末端夹爪从当前位置移至目标位置的过程。由于机器人抓取任务无需约束机器人夹爪的轨迹，而是当前位置至目标位置的移动，因此本文主要在机器人的构形空间中进行轨迹规划[80-81]。

本文使用 MoveIt!软件包进行机械臂轨迹规划工作。该软件包含有大量的轨迹规划算法，并能够通过人机通信接口实时监控机器人的工作状态。Move_group 节点能够与机器人系统的任何独立组件交互，是 Movelt!的核心组成部分。使用 Movelt!进行轨迹规划非常容易操作，创建一个机器人通用的 URDF 文件即可。在轨迹规划的过程中，规划器会生成机器人从当前状态至目标状态且不会发生接触的路径，并将该路径发送到控制器中，使机器人能够根据设定的轨迹移动。在物体抓取实验中，我们使用规划器 RRT-Connect[82]完成了一批由当前姿态至目标姿态的路径规划。

4.3 仿真实验

为了验证本文网络的有效性并规避实际抓取中可能发生的问题对算法精确性的潜在干扰，本文使用 Gazebo 虚拟实验平台进行了仿真实验。

Gazebo 作为一款开源且备受信赖的机器人仿真工具，在机器人行为模拟领域具有显著优势。它主要用于构建一个虚拟环境，用以模拟机器人及其他物体在其中的互动行为。通过 Gazebo 工具，可以有效地测试开发并且调试机器人各类

控制算法及一些任务的规划，从而提升机器人技术的实用性和可靠性。Gazebo 通过运用其物理引擎，能够精准地模拟如碰撞检测、摩擦力和重力等物理现象，从而显著提升了机器人模拟运动的真实性。另外，Gazebo 强大的功能还可以模拟出多个机器人在相同的环境中交互、协作等行为，从而验证了整个系统的协同能力。

本文的仿真实验包含以下四个模块：首先是配置模块，负责仿真环境的搭建与参数的设定，同时加载实验需要的组件；其次是视觉算法模块，该模块在获取目标物体视觉信息后依次使用零样本检测、位姿估计算法从而得到其抓取参数；紧接着是机器人轨迹规划模块，在获取抓取参数后规划机器人的运动轨迹以实现精确抓取；最后是控制模块，该模块负责将规划好的轨迹转化为具体的控制指令，驱动机器人执行抓取任务。这四个模块共同构成了本文的仿真实验框架。

在 Gazebo 仿真环境中，首先安装 Baxter 机器人仿真包，安装成功后可以在 Gazebo 的 3D 视图区看到 Baxter 机器人，如图 4-11 所示。

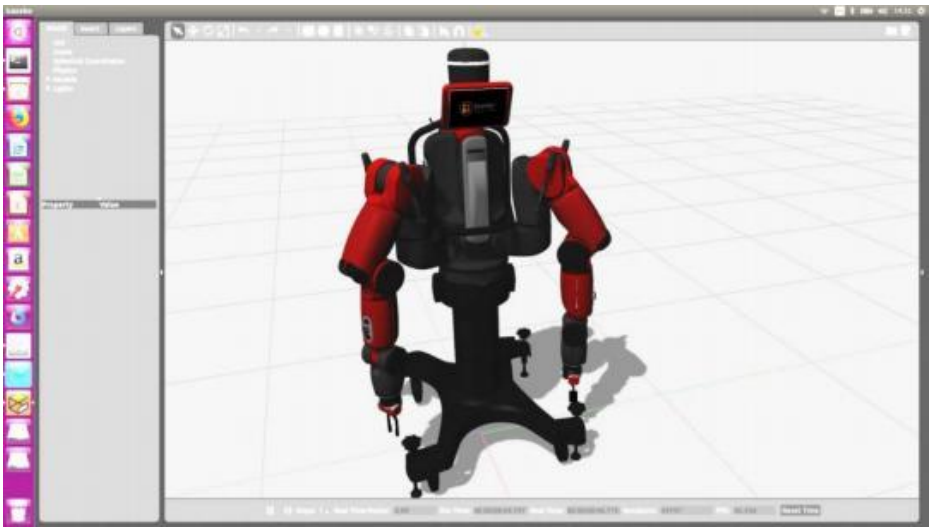


图 4-11　gazebo 视图中的 Baxter 机器人

Figure 4-11 Gazebo's Baxter robot

仿真实验软件框架如图 4-12 所示，自上而下分别为显示层、硬件层以及算法层。首先使用配置模块生成仿真实验所需的硬件模型，以确保实验的准确性和有效性，通过 Kinect 深度相机可以获得待抓取物体的场景数据，再通过算法层进行检测并估计其 6D 位姿，使用 Movelt!获取抓取参数，最后控制机械臂抓取物体。

图 4-12 仿真实验软件框架

Figure 4-12 Software framework of Gazebo

为评估算法性能，本文实验的抓取对象用了几个不同的物体，随机摆放在实验台上。使用 move_group 以及 Rviz 窗口，算法层将检测到目标物体并估计其 6D 位姿，MoveIt!会依据算法层输出的结果精准完成轨迹规划，同时生成机器人抓取位姿，这些结果将在 Rviz 中实时显示。

算法层会通过 topic 的形式将输出的结果发送到 MoveIt!进行轨迹规划的节点中，当此节点获得对应信息后，它会调用对应的接口来创建两个规划组，分别生成夹爪夹取物体的位姿以及机械臂运动的轨迹，来完成抓取任务。

本次仿真实验的运行环境为配备 Intel Xeon 4210R 2.4GHz 处理器、64GB 内存以及 Ubuntu 18.04 操作系统的计算机。

为了全面评估机器人抓取性能，每次都随机设置目标物体的摆放位置，并进行五轮抓取实验，总计尝试了 500 次的抓取。在每次抓取过程中，我们都详细记录了目标物体的检测准确率、位姿估计准确率以及最终的抓取成功率。

其中，评价检测准确率，本文主要关注机器人是否能够成功检测到场景中待抓取的目标；而位姿估计准确率的评价则侧重于检测到的物体 6D 位姿是否在预设的误差范围内。经过详细的实验数据分析，我们得出了如表 4-4 所示的实验结果。

表 4-4 仿真实验结果

Table 4-4 Results of simulation experiments

轮次	抓取次数	检测准确率	位姿估计准确率	抓取成功率
1	100	92%	90%	85%
2	100	91%	91%	84%
3	100	91%	90%	86%
4	100	91%	90%	83%
5	100	90%	91%	84%
合计	500	91.0%	90.4%	84.4%

仿真实验的结果已详细列于上表中，具体数据如下：目标物体的检测准确率达到了 91.0%，位姿估计的准确率也高达 90.4%，而最终的抓取成功率则为 84.4%。这些数据充分展示了机器人在抓取任务中的性能表现。

4.4 抓取实验

为了进一步验证本课题所提出的网络的有效性，本节使用 4.2.1 中所述机器人抓取平台进行抓取实验。在实验中，从第二章基于零样本学习的未知物体识别中划分的未知类物体中选取了 5 种物体作为被抓取目标，分别为塑料瓶、小药瓶、订书机盒、小黄人玩偶、钳子。对每类物体进行 20 次抓取，在每次抓取实验中，我们采取了单独且随机的放置方式，确保每个物体都被独立地放置在桌面上。

分别统计了对 5 种物体抓取实验的检测准确率和抓取成功率。其中，检测准确率的评价标准在于机器人是否能够精确识别场景中待抓取的目标，并成功输出其位置和姿态；而抓取成功率则是通过计算机器人夹爪成功抓取并举起物体至 20 厘米高度，且物体在此过程中未掉落的比例来统计的。

图 4-13 展示了成功抓取该 5 类物体的结果。

图 4-13　抓取示例

Figure 4-12 Grasping example

详细的实验结果如表 4-5 所示。

表 4-5 抓取实验结果

Table 4-5 Results of the grasping experiment

物体	塑料瓶	小药瓶	订书机盒	小黄人玩偶	钳子	合计
抓取次数	20	20	20	20	20	100
检测准确率	90%	91%	91%	89%	89%	90.0%
抓取成功率	80%	80%	85%	75%	75%	79%

在表 4-3 所示的抓取实验结果中，由于被抓取对象塑料瓶、小药瓶、订书机盒的形状是相对规则的，所以抓取成功率均在 80%及以上，而在抓取小黄人玩偶、钳子时会出现较多次不成功的抓取，使得抓取成功率均只有 75%。本文在仿真环境中进行的抓取实验取得了较好的结果，其中平均检测准确率达到了 91%，机器人抓取成功率也高达 84.4%；而在实际实验环境中，我们同样对抓取性能进行了测试。实验结果显示，检测准确率为 90%，而抓取成功率则为 79%。这些数据充分证明了本文所提出方法的有效性和实用性。

在抓取实验过程中，该系统的成功率会受到多种因素的影响。经过深入分析，我们发现误差产生的可能原因主要包括以下几个方面：

在实验过程中，本文的网络仍旧存在不少会影响成功率的因素。分析了成功率降低的可能原因，主要概括成以下 5 个方面：

(1)相机的安装位置、角度等因素可能导致传感器测量偏差，环境噪声的干扰也会对深度相机获取的距离信息产生一定的误差；

(2)机器人进行末端夹爪与基座之间的位姿转换时，不可避免地会存在一定的手眼标定误差；

(3)从第三章的实验结果中可以明显观察到，6D 位姿估计过程中存在一定的误差，这种误差会很大程度上影响抓取实验的准确率，因此需要进一步分析和优化位姿估计算法，以提高机器人抓取的准确性和稳定性；

(4)机器人轨迹规划的成功率会受到众多因素的共同影响，首先，机器人的负载情况、其运动学特性和摩擦力等内在因素，会直接影响运动轨迹的稳定性和准确性；其次，外部运行环境、目标物体形状以及位置变化、传感器精度及响应速度等外在因素，同样会对轨迹规划产生不小的影响；

(5)机器人在执行抓取任务时，其执行机构的误差是一个不可忽视的因素。

这种误差主要源自四个方面：第一，源于制造和装配过程中的不精确性而产生的机械结构误差；第二，传感器精度不足或控制算法不精确导致的运动控制误差；第三，机器人所处环境的温度、湿度、振动等变化产生的环境干扰误差；最后，不同的负载条件也会对机器人的运动性能产生不同程度的影响。

因此，在设计和控制机器人时，需要充分考虑这些误差因素，以提高机器人执行抓取任务的准确性和稳定性。

通过前述的抓取实验，可以得出以下结论：对于形状规则的未知物体，本文所构建的网络展现了卓越的鲁棒性；而当面对形状不规则的未知物体时，尽管抓取效果相较于规则物体稍有不足，但失败率保持在较低水平。同时，对于非规则物体的抓取判断，该网络仍能保持较高的正确率。这一结果表明，本文方法在抓取不同形状的未知物体时均具备较好的泛化性和稳定性。

4.5 本章小结

本章对所提出的未知物体抓取网络进行了实验验证和分析。首先介绍了实验环境和实验平台，然后介绍了抓取流程，随后，在 Gazebo 环境中对所提的抓取网络进行了有效性验证。其中，本文抓取网络的检测准确率高达 91.0%，位姿估计准确率也高达 90.4%，抓取成功率为 84.4%；最终，我们在实际环境中选择了五组物体进行抓取实验，并详细统计了抓取成功率。实验结果显示，在实际场景中，我们的抓取网络检测准确率达到了 90%，抓取成功率则为 79%。这些数据充分证明了本文提出的抓取网络在实际应用中的泛化性和鲁棒性。

第 5 章 总结与展望

5.1 工作总结

本文对机器人检测并抓取未知物体进行了研究。针对现有的机器人抓取未知物体网络模型泛化性能力低和训练时间过长，以及复杂非结构化环境下位姿估计鲁棒性差的问题，本文基于零样本学习与位姿估计相关模型及理论进行优化设计，进行了基于零样本学习的未知物体抓取研究。主要研究工作如下：

(1)参考现有的资料，介绍基于零样本学习及其映射方式，介绍在此基础上改进的零样本检测模型，对其进行研究并改进，使用 ROI Align 替换 ROI pooling 解决了网络在回归分割框位置的偏移问题，并提高了小目标的检测精度，也为后续的抓取任务提供了便利。

(2)面对复杂非结构化环境下面对未知物体位姿估计困难的问题，研究设计了融合局部特征匹配与点云配准的姿态估计网络以用于未知物体的姿态估计。首先，利用 LoFTR 局部特征匹配算法获取足够多的特征点并得到像素级的匹配点集，通过 RGB-D 的深度信息得到相应的 3D 点云，并将其作为 ICP 配准算法的初始点云优化配准效果。实验表明，该方法可以有效避免 ICP 算法过度依赖初始点云的问题，加快了迭代效率，提高了配准的速度及精度。

(3)最后，搭建了包含网络和硬件在内的机器人抓取实验平台，并开展了相关的抓取实验以验证本文所提出网络的有效性。在实验开始之前，我们首先进行了精确的机器人手眼标定，随后利用 ROS 系统中的 MoveIt!软件进行了轨迹规划。接着，我们在 Gazebo 仿真环境中展开了五轮抓取仿真实验，实现了高达 91.0%的检测准确率，位姿估计的准确性也达到了 90.4%，同时抓取成功率稳定在 84.4%。最终，在实际环境下，我们展开了五组不同物体的抓取实验，并对抓取成功率进行了统计。实验结果显示，在实际场景中，抓取系统的平均检测准确率达到了 90%，而综合抓取成功率则为 79%。综合对比仿真实验与实际实验的测试结果，本文所提出的基于零样本学习的未知物体抓取方法展现出了显著的性能，证明了其在实际机器人抓取任务中的泛化性与鲁棒性。

5.2 进一步工作和展望

本文针对机器人抓取未知物体困难的问题，通过零样本检测方法和未知物体位姿估计网络的研究之后，提出了一些有效的解决方法。但是本文所提出的方法仍旧存在可以优化的地方，需要在以后的工作中进行研究分析从而做出更进一步，主要包括以下几个方面：

（1）针对本文所提出的零样本检测网络的改进工作，是通过将视觉特征映射到语义空间的方式实现的，在后续的研究工作中，我们将尝试使用基于生成对抗网络的方式，对比两者的优劣程度以获取更好的检测识别方式；

(2)针对本文提出的结合局部特征匹配和点云配准的姿态估计网络，尽管可以有效地对场景中的物体进行位姿估计，但只研究了针对单一物体的位姿估计状况，在有多个物体的情况下估计效果不够理想，后期将会继续研究网络对多个物体甚至是不同物体的位姿估计能力，以实现相关的任务；

(3)本文所研究的抓取目的是成功抓取未知物体，但在生活中，机器人的抓取作业会根据特定的任务抓取指定的部位并进行相关动作，而我们的网络无法完成这一点。因此后期将会继续研究算法在面向任务的抓取中的可行性，更安全更智能地进行抓取。

参考文献

[1] 李金莉.探析工业机器人机器视觉系统在工件分拣中的应用[J].内燃机与配件, 2019, (17): 2.

[2] 徐方, 张希伟, 杜振军.我国家庭服务机器人产业发展现状调研报告[J].机器人技术与应用, 2009, (2): 6.

[3] 袁政华, 夏雪.浅析家庭型服务机器人的发展前景[J].信息记录材料, 2019, 20(7): 2.

[4] 曹心言,李润源,陈嘉郡,等.张拉整体机器人构型与运动控制研究现状及进展[J].机器人外科学杂志(中英文),2024,5(02):130-137.

[5] Barnea D I , Silverman H F .A Class of Algorithms for Fast Digital Image Registration[J].IEEE Transactions on Computers, 2009, C-21(2):179-186.

[6] Zitova B , Flusser J .Image registration methods: a survey[J].Image and vision computing, 2003,21(11): 977-1000.

[7] Kato H , Billinghurst M .Marker tracking and hmd calibration for a video-based augmented reality conferencing system[C]. Augmented Reality, 1999. (IWAR'99) Proceedings. 2nd IEEE and ACM International Workshop on. 1999: 85-94.

[8] Lowe D G .Distinctive image features from scale-invariant keypoints[J]. International journal of computer vision, 2004,60(2): 91-110.

[9] Bay H , Tuytelaars T , Van Gool L .Surf: Speeded up robust features[C]. European conference on computer vision. 2006: 404-417.

[10] Dalal N , Triggs B .Histograms of oriented gradients for human detection[C]. Computer Vision and Pattern Recognition, 2005.CVPR 2005.IEEE Computer SocietyConference on. 2005,1: 886-893.

[11] Dalal N .Histograms of oriented gradients for human detection[J].Proc of Cvpr, 2005.

[12] Ng P C , Henikoff S .SIFT: predicting amino acid changes that affect protein

function[J].Nucleic Acids Research, 2003, 31(13):3812-3814.

[13] Kumar P , HIenikoff S , Ng P C .Predicting the effects of coding nonsynonymous variants on protein function using the SIFT algorithm[J].Nature protocols, 2009,4(7): 1073-1081.

[14] Milborrow S , Nicolls F .Active shape models with SIFT descriptors and MARS[C].Computer Vision Theory and Applications (VISAPP),2014 International Conference on. 2014,2:380-387.

[15] Harris C, Stephens M. A combined corner and edge detector.[C].Alvey vision conference. 1988,15: 10-5244.

[16] Wang X , Han T X , Yan S .An HOG-LBP human detector with partial occlusion handling[C].Computer Vision, 2009 IEEE 12th International Conference on. 2009:32-39.

[17] Pang Y , Yuan Y , Li X, et al. Efficient HOG human detection[J]. Signal Processing,2011,91(4): 773-781.

[18] Sahbani A , El-Khoury S , Bidaud P . An overview of 3D object grasp synthesis algorithms[J]. Robotics and Autonomous Systems,2012,60(3): 326-336.

[19] Bohg J , Morales A , Asfour T , et al. Data-driven grasp synthesis-a survey[J]. IEEE Transactions on Robotics, 2014,30(2): 289-309.

[20] Murray R M , Li Z , Sastry S S , et al. A mathematical introduction to robotic manipulation[M].CRC press,1994: 1-18.

[21] Siciliano B , Khatib O .Springer handbook of robotics[M].Springer, 2016:274-295.

[22] Bicchi A , Kumar V .Robotic grasping and contact: A review[C]. Robotics and Automation, 2000.IEEE International Conference on. 2000,1: 348-353.

[23] Prattichizzo D , Malvezzi M , Gabiccini M , et al. On the manipulability ellipsoids of underactuated robotic hands with compliance[J]. Robotics and Autonomous Systems, 2012, 60(3): 337-346.

[24] Rosales C , Suárez R , Gabiccini M , et al. On the synthesis of feasible and prehensile robotic grasps[C]. Robotics and Automation (ICRA), 2012 IEEE

International Conference on. 2012: 550-556.

[25] Zhang L , Trinkle J C .The application of particle filtering to grasping acquisition with visual occlusion and tactile sensing[C]. Robotics and Automation (ICRA), 2012 IEEE International Conference on. 2012: 3805-3812.

[26] Rodriguez A , Mason M T , Ferry S .From caging to grasping[J]. The International Journal of Robotics Research, 2012, 31(7): 886-900.

[27] Seo J , Kim S , Kumar V . Planar, bimanual, whole-arm grasping[C]. Robotics and Automation (ICRA), 2012 IEEE International Conference on. 2012: 3271-3277.

[28] Miller A T , Knoop S , Christcnsen H I, et al. Automatic grasp planning using shape primitives[C].Robotics and Automation,2003. Proceedings.ICRA'03. IEEEInternational Conference on.2003,2: 1824-1829.

[29] Borst C , Fischer M , Hirzinger G .Grasping the dice by dicing the grasp[C]. Intelligent Robots and Systems, 2003.(IROS 2003). Proceedings. 2003 IEEE/RSJ Inter-national Conference on. 2003,4: 3692-3697.

[30] Hsiao K , Kaelbling L P , Lozano-Perez T .Robust grasping under object pose uncertainty[J]. Autonomous Robots, 2011,31(2): 253-268.

[31] Ciocarlie M T , AllenPK .Hand posture subspaces for dexterous robotic grasping[J].The International Journal of Robotics Research,2009,28(7): 851-867.

[32] Ekvall S , Kragic D .Learning and evaluation of the approach vector for automatic grasp generation and planning[C].Robotics and Automation, 2007IEEE Interna-tional Conference on. 2007: 4715-4720.

[33] Miller A T , Allen P K .GraspIt!: A versatile simulator for robotic grasping[J].IEEE Robotics & Automation Magazine, 2005, 11(4):110-122.

[34] Krocmer O , Detry R , Piater J , et al.Combining active learning and reactive control for robot grasping[J]. Robotics and Autonomous Systems, 2010,58(9): 1105-1116.

[35] EI-Khoury S , Sahbani A .Handling objects by their handles[C]. IEEE/RSJ International Conference on Intelligent Robots and Systems. 2008:1689-1695.

[36] Boularias A , Kroemer O , Peters J . Learning robot grasping from 3-d images with markov random fields[C].Intelligent Robots and Systems (IROS), 2011IEEE/RSJInternational Conference on. 2011: 1548-1553.

[37] Morales A , Sanz P J , Pobil A P , et al. Vision-based three-finger grasp synthesis constrained by hand geometry[J]. Robotics and Autonomous Systems, 2006,54(6):496-512.

[38] Dune C , Marchand E , Collowet C , et al. Active rough shape estimation of unknown objects[C].Intelligent Robots and Systems, 2008.IROS 2008. IEEE/RSJInternational Confcrencc on. 2008: 3622-3627.

[39] Rao D , Le Q V , Phoka T , et al. Grasping novel objects with depth segmentation[C].Intelligent Robots and Systems (IROS), 2010 IEEE/RSJ International Conferenceon. 2010: 2578-2585.

[40] Bohg J , Johnson-Roberson M , Leon B , et al. Mind the gap-robotic grasping under incomplete observation[C]. Robotics and Automation (ICRA),2011 IEEE International Conference on. 2011:686-693.

[41] Hsiao K , Chitta S , Ciocarlie M , et al.Contact-reactive grasping of objects with partial shape information[C]. Intelligent Robots and Systems (IROS),2010 IEEE/RSJInternational Conference on. 2010: 1228-1235.

[42] Maitin-Shepard J , Cusumano-Towner M , Lei J , et al. Cloth grasp point detection based on multiple-view geometric cues with application to robotic towel folding[C].Robotics and Automation (ICRA),2010 IEEE International Conference on. 2010:2308-2315.

[43] Saxena A , Driemeyer J , Ng A Y .Robotic grasping of novel objects using vision[J].The International Journal of Robotics Research, 2008,27(2): 157-173.

[44] Lenz I , Lee H , Saxena A . Deep learning for detecting robotic grasps[J]. The International Journal of Robotics Research, 2015,34(4-5): 705-724.

[45] Levine S , Pastor P , Krizhevsky A , et al. Learning hand-eye coordination for robotic grasping with large-scale data collection[C]. International Symposium on Experi-mental Robotics. 2016: 173-184.

[46] Wang Z , Li Z , Wang B , et al. Robot grasp detection using multimodal deep convolutional neural networks[J]. Advances in Mechanical Engincering , 2016,8(9).

[47] Redmon J , Angelova A . Real-time grasp detection using convolutional neural networks[C].Robotics and Automation (ICRA),2015 IEEE International Conferenceon. 2015: 1316-1322.

[48] Finn C , Tan X Y , Duan Y , et al. Deep spatial autoencoders for visuomotor learning[C].Roboties and Automation (ICRA),2016 IEEEInternational Conference on.2016:512-519.

[49] Pinto L , Gupta A . Supersizing self-supervision: Learning to grasp from 50k tries and 700 robot hours[C].Robotics and Automation (ICRA),2016 IEEE International Conference on. 2016: 3406-3413.

[50] Akata Z , Perronnin F , Harchaoui Z , et al. Label-embedding for attributebased classification [C]. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2013: 819-826.

[51] Akata Z , Reed S , Walter D , et al. Evaluation of output embeddings for finegrained image classification [C]. In Proceedings of the IEEE conference on computer vision and pattern recognition, 2015: 2927-2936.

[52] Fellbaum C . WordNet [J]. The encyclopedia of applied linguistics, 2012.

[53] Pennington J , Socher R , Manning C D .Glove: Global vectors for word representation [C]. In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), 2014: 1532-1543.

[54] Guo Y , Ding G , Han J , et al. Zero-shot learning with attribute selection [C].In Thirty-Second AAAI Conference on Artificial Intelligence, 2018.

[55] Jayaraman D , Grauman K . Zero-shot recognition with unreliable attributes [C]. In Advances in neural information processing systems, 2014:3464-3472.

[56] Liu Y , Guo J , Cai D , et al. Attribute attention for semantic disambiguation in zero-shot learning [C]. In Proceedings of the IEEE International Conference on Computer Vision, 2019: 6698-6707.

[57] Zhu Y , Xie J , Tang Z , et al. Semantic-guided multi-attention localization for zero-shot learning [C]. In Advances in Neural Information Processing Systems, 2019: 14943-14953.

[58] Li Y , Zhang J , Zhang J , et al. Discriminative learning of latent features for zero-shot recognition [C]. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018: 7463-7471.

[59] Szegedy C , Ioffe S , Vanhoucke V , et al. Inception-v4, inception-resnet and the impact of residual connections on learning[C]// Proceedings of the AAAI conference on artificial intelligence. 2017 31.

[60] Shigeto Y , Suzuki I , Hara K , et al. Ridge regression, hubness, and zeroshot learning [C]. In Joint European Conference on Machine Learning and Knowledge Discovery in Databases, 2015: 135-151.

[61] Romera-Paredes B , Torr P . An embarrassingly simple approach to zero-shot learning[C]. In International Conference on Machine Learning, 2015: 2152-2161.

[62] Kipf T N , Welling M . Semi-supervised classification with graph convolutional networks [J]. arXiv preprint arXiv:1609.02907, 2016.

[63] Wang X , Ye Y , Gupta A . Zero-shot recognition via semantic embeddings and knowledge graphs[C]. In Proceedings of the IEEE conference on computer vision and pattern recognition, 2018: 6857-6866.

[64] Lopes M T , Gioppo L L , Higushi T T , et al. Automatic bird species identification for large number of species[C]. In 2011 IEEE International Symposium on Multimedia, 2011: 117-122.

[65] Kampffmeyer M , Chen Y , Liang X , et al. Rethinking knowledge graph propagation for zero-shot learning [C]. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019: 11487-11496.

[66] Kodirov E , Xiang T , Gong S . Semantic autoencoder for zero-shot learning [C]. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017: 3174-3183.

[67] Jiang H , Wang R , Shan S , et al. Learning discriminative latent attributes for

zero-shot classification [C]. In Proceedings of the IEEE International Conference on Computer Vision, 2017: 4223-4232.

[68] 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.

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

[70] Mao Q , Wang C , Yu S , et al. Zero-Shot Object Detection With Attributes-Based Category Similaritys[J].IEEE Transactions on Circuits and Systems II: Express Briefs,67(5):921-925.

[71] Sun J , Shen Z , Wang Y , et al.LoFTR: Detector -free local feature matching with transformers[C]//Computer Vision and Pattern Recognition.IEEE,2021.

[72] Besl P J , Mckay N D .A Method for registration of 3-D shapes[J]. Proceedings of SPIE-The International Society for Optical Engineering, 1992,14(3):239-256.

[73] Vaswani A , Shazeer N , Parmar N , et al.Attention Is All You Need[J].arXiv, 2017.DOI:10.48550/arXiv.1706.03762.

[74] Rad M , Lepetit V .BB8: A scalable, accurate, robust to partial occlusion method for predicting the 3D poses of challenging objects without using depth[C]//2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, IEEE,2017:3848-3856.

[75] Wang G , Manhardt F , Tombari F , et al. GDR-NET: geometry-guided direct regression network for monocular 6D object pose estimation[C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR). Nashville, TN, USA, 2021: 16606-16616.

[76] Peng S , Liu Y , Huang Q , et al. PVNet: pixel-wise voting network for 6DoF pose estimation[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, CA, USA, IEEE, 2019: 4556-4565.

[77] Starke S , Hendrich N , Krupke D , et al. Evolutionary multi-objective inverse kinematics on highly articulated and humanoid robots[C]//2017 IEEE/RSJ

International Conference on Intelligent Robots and Systems (IROS). IEEE, 2017:6959-6966.

[78] Schulman J , Ho J , Lee A X , et al. Finding Locally Optimal, Collision-Free Trajectories with Sequential Convex Optimization[C]//Robotics: science and systems. 2013, 9(1): 1-10.

[79] Hinterstoisser S , Lepetit V , Ilic S , et al.Model Based Training, Detection and Pose Estimation of Texture-Less 3D Objects in Heavily Cluttered Scenes[C] //Asian Conference on Computer Vision.Springer, Berlin, Heidelberg, 2012.

[80] 王佳,胡旭晓,吴跃成.基于 UR5 机械臂的轨迹跟踪控制算法研究[J].软件工程,2022,25(06):45-48+17.

[81] 朱萌,孟焯,张豪等.基于 ROS 的 6 自由度机械臂运动轨迹规划[J].组合机床与自动化加工技术,2021(04):1-3+9.

[82] Kuffner J J , LaValle S M .RRT-connect: An efficient approach to single-query path planning[C]V/Proceedings 2000 ICRA. Millennium Conference. IEEE InternationalConference on Robotics and Automation.Symposia Proceedings (Cat. No. 00CH37065).IEEE,2000,2: 995-1001.

[83] He Y , Sun W , Hhuang H , et al.Sun, Pvn3d: A deep point-wise 3d keypoints voting network for 6dof pose estimation[C]//Computer Vision and Pattern Recognition.IEEE, 2020.

[84] Wang C , Xu D , Zhu Y , et al.Densefusion: 6d object pose estimation by iterative dense fusion[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).IEEE,2020.

附 录

攻读学位期间发表的学术论文目录

[1] 张靖,于伶.基于零样本学习的检测网络[J].计算机应用文摘，2024，第 40 卷 17 期.

[2] 张靖,于伶.结合局部特征匹配和点云配准的姿态估计网络[J/OL].重庆工商大学学报(自然科学版):1-9.

[3] 张靖,曹雏清.一种基于零样本学习的机器人抓取方法及系统.（第一作者，实质审查中，公开号：CN117681204A）

攻读硕士学位期间参加的竞赛及科研情况

[1] 第六届全国机器人专利创新创业大赛.二等奖.参赛作品：《非均匀环境下的多选择 CFAR 检测方法研究》。

[2] 铝顶装配无人工厂项目设计与研发：研究机器人端拾器的设计，以适合各个工位物料的抓取；设计零件的机器人抓取定位结构以及组装定位精度。

[3] 压缩机机壳三次元自动生产线的设计与研发：设计取料机械手臂，由 1 个伺服电机驱动从上原点位置下降，到下方的双工位站后，对工件进行印油，然后用吸盘将工件拾起，回到上原点等待三次元手臂手臂接料。

致 谢

岁月如梭，时光飞逝，三年的研究生生活即将画上圆满的句号。回首过去，我深感收获颇丰。在此，我要衷心感谢导师们的悉心指导，同学们的相互扶持，以及家人始终如一的关爱与支持。正因为有你们的陪伴，我的学习之路更加温馨而充实，这段时光也成为了我人生中难以忘怀的美好回忆。

首先，我要感谢导师曹雏清老师和刘志恒老师对我的指导和支持，在学术和人生方面给予了深刻的启发和帮助。论文的每个环节都得到了详细的指导和建议，对选题、方案设计、撰写修改等方面提供了宝贵的指导。在此向两位老师致以深深的感谢和崇高的敬意！

其次，对芜湖哈特机器人产业技术研究院有限公司的领导和同事们表示衷心的感谢。学习和工作过程中给予了悉心的指导和无私的帮助，传授了有效的学习方法，使获益良多。支持与帮助为毕业论文成功奠定了坚实基础。

然后，感谢同学和室友们的关心与帮助。正是你们的陪伴与支持，让我能够顺利走过研究生生涯的每一个阶段。在此，我向你们表示由衷的感谢。

最后，感谢百忙之中评阅本文，并提出宝贵意见和出席论文答辩的各位专家教授，不妥之处，敬请指正!