

华中科技大学

本科毕业设计[论文]

结合深度学习的机器人 面向任务抓取检测方法研究

院 系 机械科学与工程学院

专业班级 机器人 2002 班

姓 名 郭炜星

学 号 U202011175

指导老师 王书亭 教授

2024 年 5 月 15 日

学位论文原创性声明

（黑体小2号加粗居中）

本人郑重声明：所呈交的论文是本人在导师的指导下独立进行研究所取得的研究成果。除了文中特别加以标注引用的内容外，本论文不包括任何其他个人或集体已经发表或撰写的成果作品。本人完全意识到本声明的法律后果由本人承担。

作者签名：

年 月 日

学位论文版权使用授权书

本学位论文作者完全了解学校有关保障、使用学位论文的规定，同意学校保留并向有关学位论文管理部门或机构递交论文的复印件和电子版，允许论文被查阅和借阅。本人授权省级优秀学士论文评选机构将本学位论文的全部或部分内容编入有关数据进行检索，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。

本学位论文属于 1、保密口，在解密后适用本授权书。

2、不保密口。

（请在以上相应方框内打“√”）

作者签名：

年 月 日

导师签名：

年 月 日

摘要

对比传统的机械臂抓取方法，基于深度学习的机器人抓取方法具有能适应复杂环境，解除了对抓取场景和抓取物体的限制，可进行端对端的抓取，可自主对网络进行调整从而能在大方向相同的条件下适配多种不同的具体任务即具有任务导向是面向任务的方法等多种优点。

由于基于深度学习的机器人抓取方法上述列举出的优点还有基于某些特定领域的特定优点，该方法目前在零件装配、物流运输、生活服务、物联网等领域得到了越来越多的关注与发展。

但目前在应用于复杂情景时，目标检测和抓取识别算法的融合以及融合算法的准确率（AP）、速度（FPS）的进一步提高还是为解决的难题。

基于上述的问题，本文对基于深度学习的机器人抓取识别算法的研究主要落实在下面部分的具体研究内容上：

在机械臂目标自动识别与抓取的研究中，本文搭建了一个实验平台，旨在实现机械臂对目标物体的自动识别和抓取。为了进行目标分类检测，本文采用了深度学习框架，并使用改进的 YOLOv5 算法进行目标分类检测。该算法在目标检测方面具有较高的准确性和实时性。在物体抓取末态位姿检测方面，本文使用基于 GR-ConvNet 的算法。该算法通过分析目标物体的形状和姿态信息，确定机械臂在抓取过程中的最佳末态位姿。

最终，本文的成功地将目标分类检测和抓取末态位姿检测的代码进行融合，实现了一种综合的深度学习算法。并且在物理样机上进行了实验测试，并得到了令人满意的结果。实验结果表明，结合深度学习算法的面向任务抓取检测框架能够快速准确地进行目标物体的分类识别，并且能够分析出抓取位姿和抓取角度。通过这种融合的方法，机械臂能够实现面向任务的自动识别和抓取目标物体的功能。这意味着机械臂能够根据任务需求自主地辨别不同的目标物体，并且能够准确地抓取它们的位置和角度，从而完成特定的任务。

关键词：抓取检测，深度学习算法，目标识别，位姿预测，深度学习算法改进

Abstract

Compared with the traditional robotic arm grasping method, the robot grasping method based on deep learning has the advantages of being able to adapt to complex environments, lifting the restrictions on grasping scenes and grasping objects, end-to-end grasping, and autonomously adjusting the network, so as to adapt to a variety of different specific tasks under the same conditions in the general direction, that is, it has many advantages such as task-oriented and task-oriented methods.

Due to the advantages listed above and the specific advantages of some specific fields based on the robot grasping method based on deep learning, this method has received more and more attention and development in the fields of parts assembly, logistics and transportation, and Internet of Things. But for now when applied to complex scenarios, the fusion of object detection and grasping recognition algorithms, as well as the further improvement of the accuracy (AP) and speed (FPS) of the fusion algorithm, are still difficult problems to solve.

Based on the above problems, the research on the robot grasping and recognition algorithm based on deep learning in this paper is mainly implemented in the following specific research contents:

In the study of automatic target recognition and grabbing by robotic arms, this paper builds an experimental platform to achieve automatic recognition and grabbing of target objects by robotic arms. In order to perform target classification detection, this article adopts a deep learning framework and uses the improved YOLOv5 algorithm for target classification detection. This algorithm has high accuracy and real-time performance in target detection. In terms of object grasping final pose detection, this article uses an algorithm based on GR-ConvNet. This algorithm determines the optimal final posture of the robotic arm during the grasping process by analyzing the shape and attitude information of the target object.

Finally, this paper successfully fuses the code of object classification detection and terminal pose detection to achieve a comprehensive deep learning algorithm. And

the experimental test was carried out on the physical prototype, and satisfactory results were obtained. Experimental results show that the task-oriented grasping detection framework combined with deep learning algorithm can quickly and accurately classify and identify target objects, and can analyze the grasping posture and grasping angle. Through this fusion approach, the robotic arm is able to realize the task-oriented function of automatically identifying and grasping target objects. This means that the robotic arm is able to autonomously identify different target objects according to the needs of the task and can accurately grasp their position and angle to complete a specific task.

Key words: grasping detection, deep learning algorithms, target recognition, grasping pose prediction, deep learning algorithm improvement

目 录

摘要	I
Abstract.....	II
1 引言	1
1.1 研究背景、课题来源及意义	1
1.2 国内外研究现状及发展趋势	4
1.2.1 目标检测研究现状	4
1.2.2 抓取位姿估计研究现状	5
1.3 本文拟解决的关键问题	9
1.4 国内外市场情况分析及本文项目成果竞争力分析	9
1.4.1 市场情况分析	9
1.4.2 行业竞争对手分析	10
1.4.3 产品技术竞争力分析	10
1.4.4 产品性能竞争力分析	11
1.5 章节规划	11
2 基于深度学习的机械臂抓取理论基础介绍	13
2.1 引言	13
2.2 面向任务的抓取模型搭建	13
2.2.1 相机成像模型	13
2.2.2 手眼模型	17
2.2.3 机械臂运动学模型	20
2.3 卷积神经网络	26
2.3.1 卷积层	27
2.3.2 池化层	28
2.3.3 激活函数	29
2.3.4 全连接层	32
2.4 本章小结	32
3 基于多注意力机制的改进 YOLOv5 模型	34
3.1 引言	34

3.2 基于深度学习的目标检测算法.....	34
3.2.1 YOLOv5 与历代 YOLO 的比较与优势	34
3.3 YOLOv5 介绍.....	37
3.3.1 输入端.....	37
3.3.2 主干	39
3.3.3 Neck.....	41
3.3.4 输出端.....	42
3.4 改进的 YOLOv5 目标检测算法	43
3.4.1 多注意力模块集成分层特征映射主干	43
3.4.2 基于宽高差异值的快速收敛损失函数	46
3.4.3 自适应特征优化的端到端注意力模块	48
3.5 仿真实验验证	50
3.5.1 常用数据集	50
3.5.2 实验抓取数据集制作	50
3.5.3 改良前后算法在官方数据集上的预测结果对比及结果展示.....	51
3.6 本章小结	54
4 基于改进 YOLOv5 与 GR-ConvNet 的机器人面向任务的抓取检测框架.....	55
4.1 引言	55
4.2 GR-ConvNet	56
4.2.1 模型架构	57
4.3 改进的 YOLOv5 算法与 GR-ConvNet 面向任务抓取框架融合	59
4.4 仿真实验验证	60
4.4.1 抓取识别算法常用数据集	60
4.4.2 实验抓取数据集制作	64
4.4.3 模型训练结果	65
4.5 本章小结	67
5 实验设计和结果分析.....	68
5.1 引言	68
5.2 面向任务的抓取检测平台搭建.....	68

5.3 实验验证与结果分析.....	70
5.4 本章小结	76
6 总结与展望	76
6.1 全文总结	76
6.2 后续展望	77

1 引言

1.1 研究背景、课题来源及意义

近年来，机器人作为辅助人类的工作或代替人类的工作的一种手段，已经融入了人类的生活当中，活跃于社会生活的各个领域。比如在工业生产领域中，工厂面临着一系列问题，如人力成本的上升、工作人员管理的复杂性以及安全保障的要求，所以工厂制造业的发展趋势是向着成手工与机械结合的生产方式发展的，主要体现为人工操纵自动化机床从而减少人与高危零件高危环节的接触。

由于现代社会对生产力的要求不断提高，传统的工厂面临着人力成本、工作人员管理和安全保障等问题。在这种情况下，机械臂可以作为一种高效、可靠的工具，能够代替人手完成一些有毒有害气体粉尘、爆炸和触电风险较高的工作。

在智能制造领域^[1]，相继有不少企业开始应用工业机器人。其中有两种典型的工业机器人，一种是焊接机器人，另外一种是搬运机器人。除此之外，在钢铁制造领域中有应用一些特殊功能的工业机器人。

而在日常生活中，机器人也逐步步入了人们的生活的方方面面，其中智能家居机器人是如今为最火热的话题。智能家居中的扫地机器人则是家具机器人中最先落地实施的一种，这种机器人可以让您不用劳心劳力地每日清扫地板的灰尘。这种便民机器人能够让人的生活更加舒心惬意。机器人技术引入住宅环境也能为人们提供更安全舒适的生活，比如当家里长期有老人和儿童居住时，家政机器人能起到额外的安全保障和照顾的作用，比如帮忙拿取存放在高处的瓶瓶罐罐等物品并在这个流程中起到保障主人安全。

而在机器人的各种作业情景中，机器人自主抓取技术无疑是机器人作业的重要核心和基本流程中的一环，同时它也是机器人与环境进行交互最基本和最重要的功能之一。所以，机器人抓取规划目前已经成为了全世界学者和工业界的研究热点。机器人的自主抓取功能已经有着广泛的应用前景，在工业生产中自主抓取技术的进步加速了机器人技术的发展并应用于装配、包装、分拣等各行各业多种有具体任务的应用中。比如，工业机器人需要在生产线上对目标零件进行抓取并且将其放在指定的位置；农用机器人需要利用抓取技术来采摘蔬

菜水果或其他农作物；医疗机器人需要通过精确抓取技术来对医疗器材或目标组织和器官进行抓取来完成样本组织取样或者完成外内科手术等操作；服务机器人需要通过抓取方法来对服务需要的工具或服务对象需要的物品来帮助人们完成各种任务；物流机器人需要通过抓取技术来进行包装、分拣等高效地处理货物的方式。这些都是面向具体任务时需要机器人抓取方法的具体会出现的例子。

然而，机器人在面向具体任务时能否准确抓取仍然是一项具有挑战性和复杂性的工作。而成功抓取的关键是抓取位姿规划，而随着视觉传感器水平的提高和视觉算法研究的不断深入，目前基于视觉的抓取方法已经取代早期的人工监督学习方法，成为机器人感知周围环境的重要途径。

早期研究者从传感器数据（主要是图像信息）中人工提取出特征表达，再利用传统机器学习的方法从监督数据中学习人工特征与抓取位姿间的映射关系^[2]。

这样的方法虽然能够在未知物体上进行抓取经验的迁移，但由于人工设计的特征受限人类的认知而具有局限性，难以进行一些更加复杂的表示，因此通常没法对其他选定的更复杂的任务有效。

借鉴深度学习在计算机视觉等领域的成功，利用深度神经网络^[2]来取代传统机器学习方法进行本文所关注的面向任务的抓取具有较大的研究意义。

所以为了应对在任意环境下对所选定的特定对象的抓取，本文认为结合深度学习目标检测算法的机器人抓取位姿估计^{[4][5][6]}对于面向任务的抓取方法研究会有深远的发展。因此本文认为利用深度学习来优化机器人的抓取检测的流程并且将目标检测算法和抓取识别算法合二为一，最后基于深度学习对机器人检测抓取点和姿态的算法进行训练**错误!未找到引用源。**是解决目前机器人在任意环境下对特定抓取目标及对该目标的抓取位姿的识别误差大，识别速度慢等问题的一些合理的办法。

目前市面上的各类抓取算法都无法避开的是：这些算法都需要人为的对输入的要抓取的物体进行标定或分类，才能避免抓取目标不明确从而导致的抓取失败的问题。所以市面上的大部分抓取检测算法都面临着数据集处理工作量大，处理过程繁琐的问题。为了应对这个问题，实现对特定目标的抓取位姿估计算

法的编写，本文采用了将目标检测算法和抓取位姿估计算法结合的方法，即用基于深度学习的目标检测方法解决了对抓取目标的定位分类问题，又避免了人为处理大规模的数据集导致的错误疏漏问题，进一步提高了抓取的成功率。

因此，本课题决定采用基于目标检测的面向任务的抓取识别算法来解决在任意环境下^[8]以任务目标作为主要驱动的面向任务的抓取位姿规划难题。

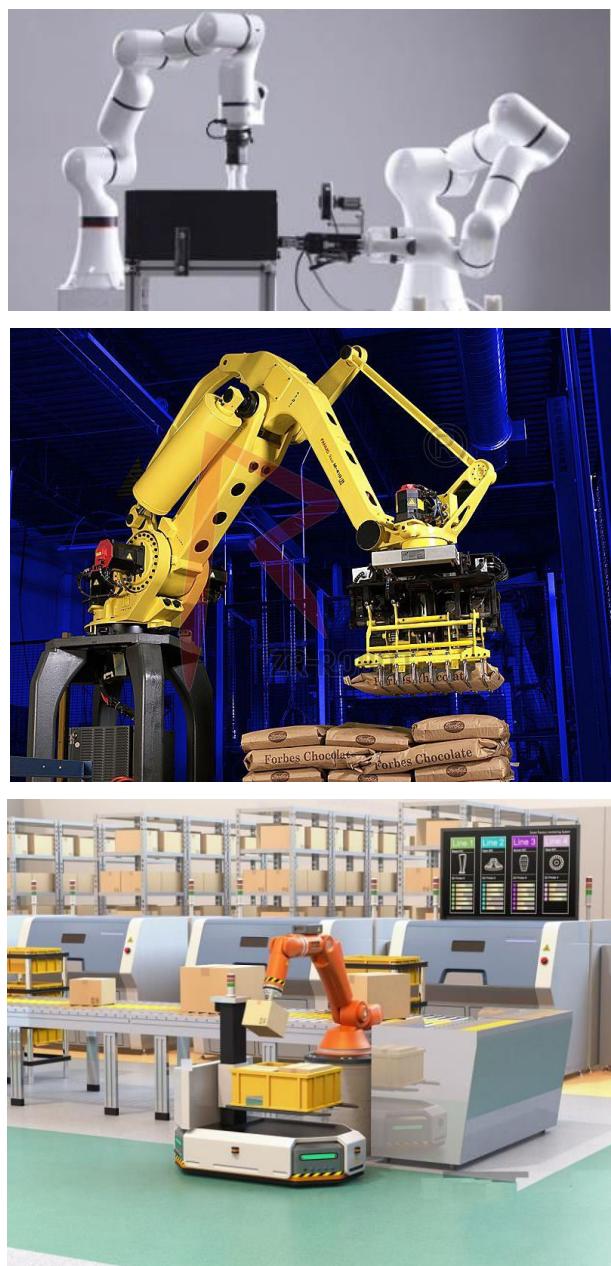


图 1-1 机器人抓取应用场景示例

1.2 国内外研究现状及发展趋势

1.2.1 目标检测研究现状

目标检测是计算机视觉和数字图像处理领域的热门方向，广泛应用于多个领域。它的主要任务是在图像或视频中识别或进一步定位出感兴趣的目标物体。

通过利用计算机视觉技术进行目标检测，可以减少人力资源的消耗，提高工作效率。相比传统的手动标注和识别方法，自动化的目标检测算法能够更快地、准确地完成目标的识别和定位，具有很高的实用价值。

目标检测在智能监控系统中扮演着核心的角色。它可以帮助实现身份识别、实例分割等任务，从而提升监控系统的智能化水平。通过目标检测，监控系统可以自动识别和跟踪感兴趣的目标，实现对目标的实时监控和分析，为安全管理、犯罪预防等提供有力支持。

目前，基于深度学习的目标检测算法主要可以分为两种思路：双阶段目标检测算法和单阶段目标检测算法。双阶段算法精度较高，但其训练速度较慢，代表性的算法包括 R-CNN 错误!未找到引用源。、SPP-net^[10]、Fast R-CNN^[12]等。而单阶段算法则是一种端对端的算法，可以同时进行物体位置估计和类别分类，代表性的算法有 YOLO^[4]、YOLOv2^[14]、YOLOv3^[15]和 SSD^[16]等。

目标检测领域的发展趋势涉及多个方向。其中包括轻量目标检测算法的研究，旨在提高算法的运行效率和性能化；小目标检测的挑战与解决，以应对在图像中尺寸较小的目标检测问题；视频检测的发展，关注于在视频流中实时进行目标检测和跟踪；以及弱监督检测的研究，旨在通过利用较少的监督信息实现准确的目标检测。

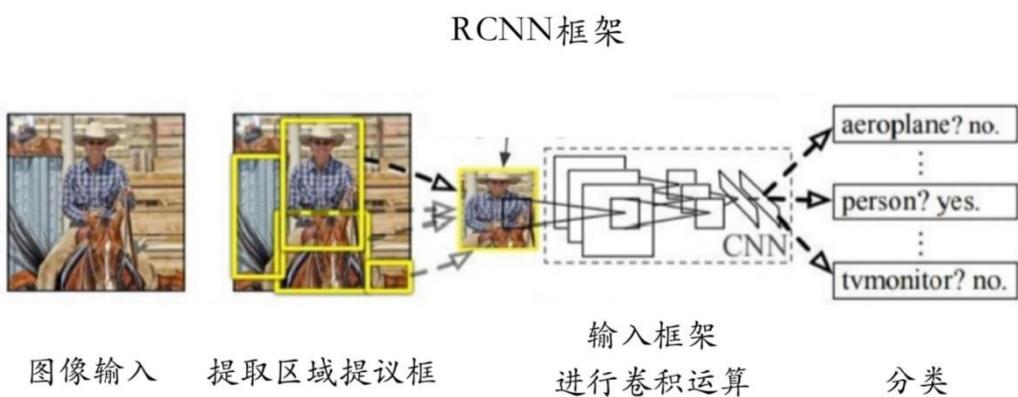


图 1-2 R-CNN 程序框图

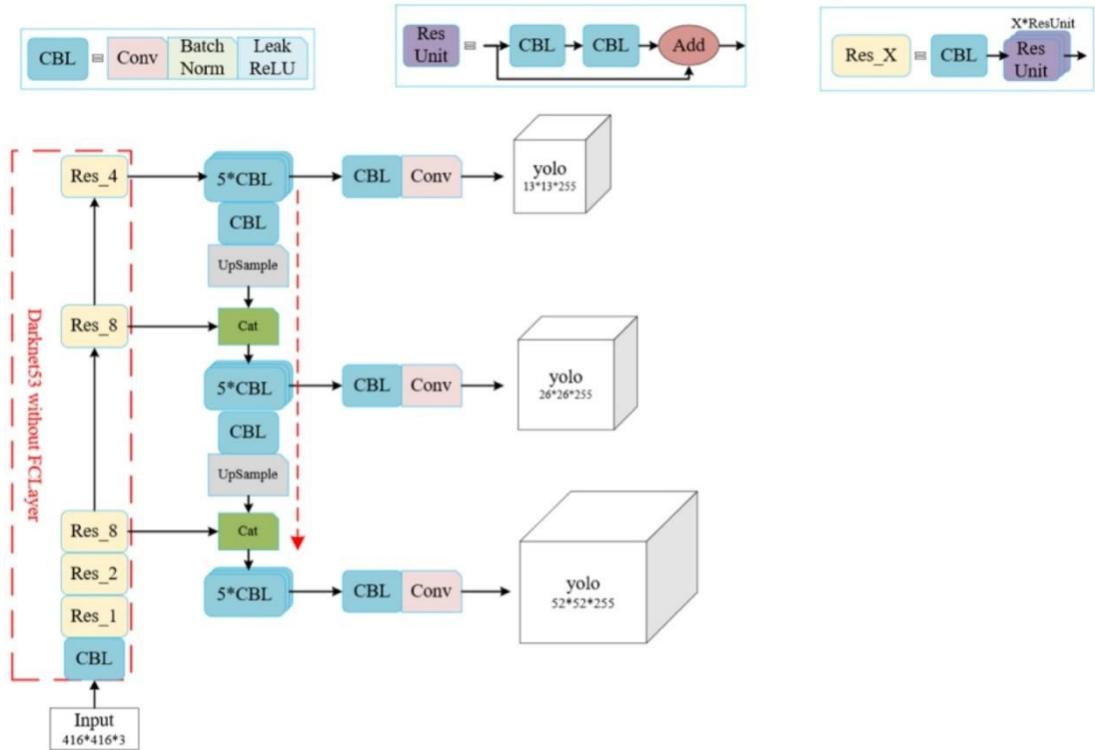


图 1-3 YOLOv3 程序框图

1.2.2 抓取位姿估计研究现状

目前机器人抓取位姿估计技术主要分为两大类错误!未找到引用源。^[17]:一类是传统的分析法，这种传统的分析法是基于物体的几何形状、物理模型、运动学和力学分析来进行抓取位姿的检测^[17]，这种方法也被称为硬编码方法^[18];另一类是基于经验的抓取方法，这种方法通过观察和学习人类的抓取动作，并将这些通过学习得到的经验应用到机器人抓取任务中来。这种方法避免了复杂的物理力学和数学模型的计算，使得抓取过程更加简化搞笑。而反之分析法在确定抓取时需要满足运动学或动力学上的公式以进行抓取位姿的计算，这种方法要求对物体的几何特征和环境的物理属性有深入的了解，并需要进行复杂的数学计算。所以基于分析的抓取方法在理论上更加严谨，但计算复杂度高且对环境要求较高。图 1-4 为分析法在进行抓取检测时所进行的流程和策略^[19]。首先，考虑到环境的影响和机械臂与周边物体或障碍的模型来进行抓取检测。其中，抓取检测是指在机器人抓取任务中找到一组满足面向特定任务的抓取需求的抓取位姿，这些位姿包括抓取角度、抓取宽度和抓取成功率等因素。抓取检测的

目标是确保机器人能够稳定地抓取目标物体，并提高抓取成功的概率。最后，将抓取位姿信息通过手眼模型，相机模型的转换映射传递给机器人执行抓取具体流程。在实际抓取应用时不一定会包含图 1-4 中介绍的所有元素和部分。图 1-5 就分析法所涉及的运动学、动力学、几何结构关系进行了详细的展开说明。在阅读文献的过程中发现分析法的应用中存在计算困难的问题，并且很少考虑面向任务的具体约束。通常情况下，分析法只在遇到问题时才进行解决，而缺乏对任务约束的主动考虑；分析法还有一个主要缺点，它要求目标对象的参数是已知的，这限制了它在非结构化环境中的广泛应用。在现实世界中，很多物体的参数是未知的或者难以准确获取，因此分析法的适用范围受到了限制。为了避免上述缺点，经验法被引入到了抓取检测中，经验法^[18]能够提供机器人更高的环境感知和适应能力，减少对手动建模的依赖。通过观察和学习人类的抓取动作，机器人可以从经验中获得抓取技巧，提高抓取成功率。

另一种方法为纯经验法，这种方法主要依赖于先前已知的成功抓取经验，且该方法通过参量标准对候选抓取位姿进行排序，以选择最适合的抓取方式。该方法参考了学习和分类方法的技术，避免了具体情况具体分析的复杂性，这种方法也被称为比较法^[21]和以知识为基础的方法^[22]。经验法的分类方式多种多样，按照算法将其分为学习式和启发式，如图 1-6 所示。学习式经验法通过人类示范、标记示例或无初始参量的反复试验来学习抓取技巧。机器人通过模仿和学习人类的抓取行为，逐渐提高自身的抓取能力。启发式经验法则依赖于各种启发式方法，将感知到的物体结构数据与候选抓取姿态进行比对。启发式方法可以基于物体的形状、纹理、重心等特征，结合机器人的感知能力，选择最佳的抓取姿态。其中，本课题将采用的抓取策略就属于经验法中的学习式方法。

使用学习的方法检测抓取又可分为两大类，如图 1-7 所示：

相比之下，单阶段算法是一种端到端的抓取方法，它基于视觉运动控制策略。这种方法可以直接从 RGB 图像作为输入，并输出具体的抓取位姿。单阶段算法通过深度学习和视觉感知技术，实现从图像到抓取位姿的端到端映射，使机器人能够直接根据图像进行抓取操作。

是一种需要独立的抓取规划控制系统的抓取算法，被称为双阶段算法。这种算法通过抓取检测生成适合的抓取位姿，然后使用规划控制系统生成相应

的轨迹，从而完成抓取过程。

在双阶段算法中，抓取检测负责从输入数据中提取有关目标物体的信息，并生成一组合适的抓取位姿。这些位姿包括抓取角度、抓取位置等。然后，规划控制系统利用这些位姿信息进行路径规划和控制，使机器人能够准确地执行抓取动作。

二是一种端到端，并基于视觉运动控制策略的抓取方法即单阶段算法，这种方法可以直接从 RGB 图像作为输入，并输出具体的抓取位姿。单阶段算法通过深度学习和视觉感知技术，实现从图像到抓取位姿的端到端映射，使机器人能够直接根据图像进行抓取操作。

ImageNet 的成功和快速计算技术的进步为基于深度学习的学习法抓取方法的发展提供了重要的推动力。ImageNet 是一个包含数百万张图像的大规模数据库，它为深度学习模型的训练提供了丰富的数据资源。同时，随着计算技术的不断发展，深度学习模型的训练和推理速度也大大提高，使得基于深度学习的学习法抓取方法成为可能^{[27][29]}。

RGB-D 传感器的实用性和经济性也在推动使用深度学习技术从图像数据中直接学习物体特征的技术。RGB-D 传感器能够提供丰富的颜色和深度信息，使得机器人可以更准确地感知和理解环境中的物体。同时，随着传感器技术的发展，RGB-D 传感器的成本也逐渐降低，使得更多的研究者和开发者能够使用这种传感器进行基于深度学习的学习法抓取方法的研究和应用。

Pinto 等人^[31]使用了类似于 AlexNet 的架构，他们的 CNN 在设计上借鉴了 AlexNet，并通过增加数据大小的方法来提高网络的泛化能力，使其能够更好地适应新的数据。Varley 等人^[32]提出了一种有趣的方法，通过形状补全来掌握规划，其中使用 3D CNN 在他们自己的数据集上从不同角度捕获的物体的 3D 原型上训练网络。Guo 等错误!未找到引用源。使用触觉数据和视觉数据训练混合深度架构。

Mahler 等^[33]提出了一种抓取质量卷积神经网络(GQ-CNN)，该网络通过在 Dex-Net 2.0 抓取规划器数据集上训练的合成点云数据预测抓取。

Levine 等人^[34]讨论了使用深度学习框架将单眼图像用于机器人抓取的手眼协调。他们使用 CNN 进行抓取成功预测，并进一步使用连续伺服来连续伺服机

械手以纠正错误。Antanas 等人^[34]研发了概率逻辑框架，这种框架可以基于语义对象分类部分来提高机器人的抓取能力。这个框架结合了高级推理和低级把握。高级推理包括对象的可视性、类别和基于任务的信息，而低级推理使用视觉形状特征。

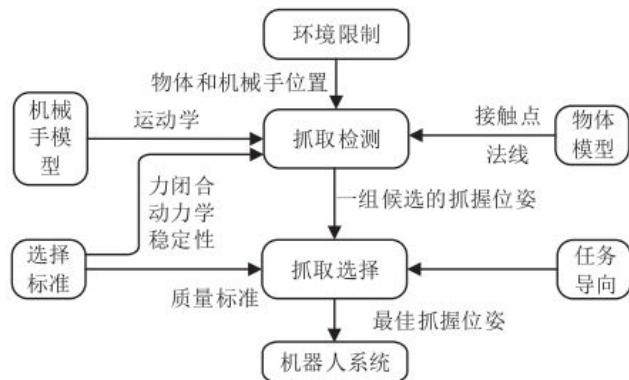


图 1-4 使用分析法抓取检测的策略

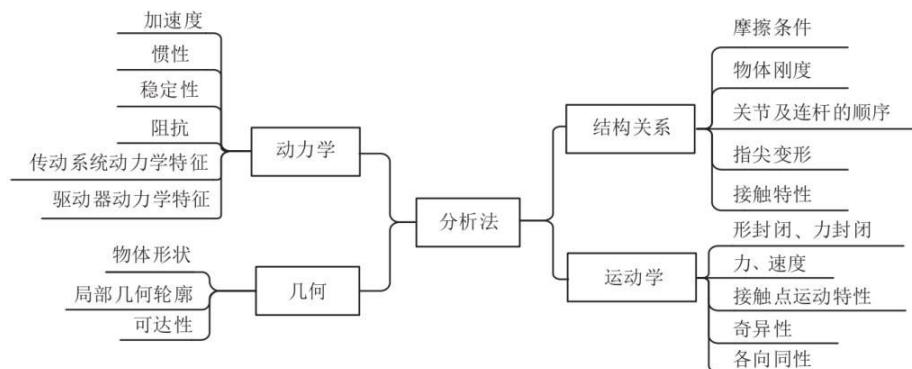


图 1-5 分析法研究所涉及内容

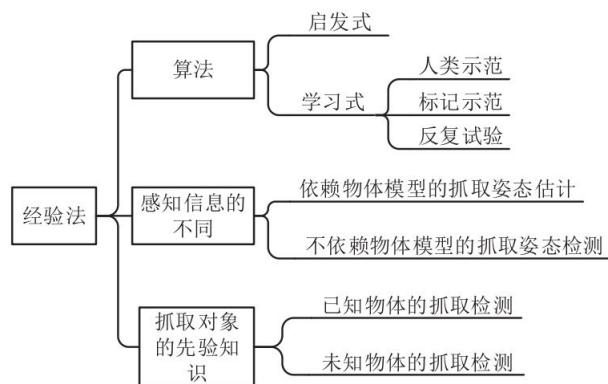


图 1-6 经验法的分类方式

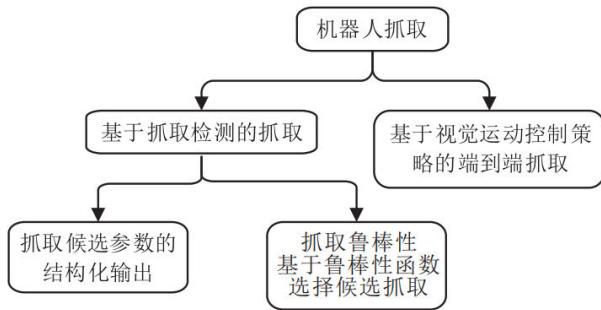


图 1-7 基于学习的方法进行抓取的分类

1.3 本文拟解决的关键问题

本文以机器人自主抓取作业情景为背景，主要研究面向任务的机械臂抓取末态位姿估计方法。考虑到应用场景和研究对象的复杂性，参考国内外的研究方法，现有算法在估算正确率、估算速度还需要提升，且现有算法大多没有针对具体任务对象。为此，本文将重点研究位姿估计所需模型的建立、YOLOv5 目标检测算法的改进^[23]、基于深度学习的抓取检测方法融合与应用^[23]~~错误！未找到引用源。~~^[25]。

本课题的目标是解决上述机器人抓取检测中留存的问题：

- (1) 首先是分别对抓取位姿识别中会出现的模型如机器人的相机成像模型、手眼模型、运动学模型进行分析。
- (2) 其次结合深度学习算法，且应用的算法应对于自制数据集的准确率即 mAP 值达到 50% 以上以及在支持 GPU 的设备上图像处理速度即 FPS 值达到 140 以上。
- (3) 实现面向任务的机器人抓取检测算法，即完成机器人抓取检测的一整套从目标检测到目标图像处理再到抓取识别算法最后到位姿输出的流程，且最终输出结果经过模型坐标系变换后输出的是抓取目标的最终位姿。换言之就是通过将目标检测和抓取识别算法的结合实现机器人从物体识别到实现抓取的综合算法。

1.4 国内外市场情况分析及本文项目成果竞争力分析

1.4.1 市场情况分析

机器人抓取位姿检测算法市场正处于快速增长阶段，主要受益于自动化、

智能制造和物流行业的发展。随着工业 4.0 的推进，越来越多的企业正在寻求高效、精准的机器人抓取方案以提高生产效率和降低运营成本。

1.4.2 行业竞争对手分析

在机器人抓取位姿检测领域，主要的竞争对手包括：

ABB: 提供先进的机器人抓取解决方案，具有强大的工业背景和稳定的市场份额。

Fanuc: 以其高精度和高速度的机器人系统著称，广泛应用于汽车和电子制造业。

KUKA: 专注于智能自动化，提供模块化和灵活的机器人抓取系统。

Universal Robots: 以协作机器人（Cobot）见长，适用于中小企业和多种应用场景。

Cognex: 以机器视觉起家，提供高精度的图像识别和抓取定位系统。

1.4.3 产品技术竞争力分析

本产品在技术上的竞争力体现在以下几个方面：

多注意力模块集成分层特征映射主干：

通过替换的新主干 swin transformer 的多注意力模块的集成，可以更好地捕捉图像中的关键特征，提高算法的鲁棒性和精确度。相比传统的单一注意力机制，本产品的多注意力模块可以显著提升特征提取的效果，适应多种复杂场景尤其是小目标检测的任务。

基于宽高差异值的快速收敛损失函数：

本文所采用的损失函数 EIOU 的设计可以加速模型的收敛，减少训练时间，同时提高定位的准确性。在市场上的其他算法中，大多使用传统的损失函数，本产品在收敛速度和精确度上具有显著优势，能够更快地迭代和部署。

自适应特征优化的端到端注意力模块：

通过自适应特征优化的端到端注意力模块 CBAM，框架可以根据具体任务动态调整特征提取和处理策略，提高整体性能。这一特性使得本产品在处理不同类型和复杂度的任务时具有更高的灵活性和适应性，优于那些固定特征提取流程的算法。

1.4.4 产品性能竞争力分析

首先，本文的产品拥有高精度和高鲁棒性。多层次特征映射和自适应优化使得本产品在不同环境下都能保持高精度的抓取定位，减少误差率。

其次，本文的模型具有很高的损失收敛速度。基于宽高差异值的快速收敛损失函数显著缩短了训练时间，能够更快地部署到生产环境中。

然后，本模型还具有极高的适应性和灵活性。端到端注意力模块的自适应特征优化使得本产品能够应对各种复杂应用场景，从而在多样化的市场需求中具有较强的竞争力。

最后，本产品实现了目标检测和抓取位姿检测算法的融合，实现了面向任务面向具体对象的机器人抓取识别检测。所以，本产品在面向具体任务的情景中，非常实用，且算法的技术和性能上的竞争力均十分突出。通过集成多注意力模块、快速收敛的损失函数以及自适应特征优化，本产品可以在抓取位姿检测领域提供更高的精确度、更快的训练速度和更强的环境适应性。这些优势将使本产品在市场上占据有利地位，应对来自其他竞争对手的挑战。

1.5 章节规划

针对面向任务的机械臂抓取位姿估计方法研究，以提高抓取位姿估计输出稳定性，成功率，输出结果速度为目标，本文重点研究了位姿估计所需的模型建立、目标检测算法的基础知识、目标检测算法的编写与改进、抓取检测算法的框架以及两种算法的融合使用，最终实现了结合深度学习的机械臂抓取末态位姿估计方法，各个章节的具体内容如下：

第一章以多个现实角度为机械臂抓取应用背景介绍了论文的研究目的，并且通过剖析现有的算法中存在的疏漏从而强调了本文所采用的算法的先进性和意义，强调了本文将要解决的关键的问题与相应所需技术。最后还对本文的工作进行了总结。

第二章具体系统的建立了机器人抓取位姿估计所需的各模型以及系统的介绍了后文所需要的深度学习算法的相关知识。关于相关模型建立部分，首先对相机成像模型进行分析，相机成像模型的建立和标定是机械臂位姿估计算法的重中之重；之后建立了机械臂手眼模型，将机械臂和相机之间的手眼关系提炼到数学模型中，方便相机读取的位置姿态信息转换到机械臂坐标系；最后建立

了机械臂运动学模型，方便了后续的位姿估计和运动规划流程研究。关于卷积神经网络相关必要知识介绍的部分，该章节主要介绍了卷积层、池化层、激活函数等后续搭建卷积神经网络所需要的必要知识。

第三章中首先介绍了 YOLOv5 原网络的基础结构并列举了其中将在后续内容中被改进的主干网络的缺点与不足，引出来后续的改动部分并对本文主要改动的三个部分进行了具体的阐述，分别为多注意力模块内嵌的分层特征映射主干替换，基于宽高差异值的快速收敛损失函数改进和自适应特征优化的端到端注意力模块的嵌入。最后进行了仿真实验验证，首先提出了自制数据集的方法与实验过程并介绍了仿真实验中各参数的具体设置，并将改动后设置好参数的算法在同一数据集中的结果（AP、FPS）进行对比从而体现改动的优势与合理性。

第四章研究了抓取检测位姿估计算法（GR-ConvNet）的网络架构和其比起现有其他网络架构的优势性，并且引出了本文的融合架构即基于改进的 YOLOv5 与 GR-ConvNet 的机器人面向任务的抓取检测框架，并具体阐述了本文是如何将目标检测算法与抓取识别算法进行融合操作。最终进行了仿真实验验证，首先提出了基于 GR-ConvNet 的仿 Cornell 数据集格式的自制数据集制作流程，并展示了在该数据集下改进的模型达到的实现面向具体任务的抓取位姿估计的效果且与其他文章中采用标准数据集的算法结果进行了对比。

第五章进行了具体的实验，介绍了实验流程与实验结论。首先介绍了本次实验所使用的硬件平台的型号并且展示了相关图片，接着展示了具体实验流程的图片及数据，最后展示了实验结果与具体与其他相似网络结果的解决进行了比较。

第六章总结了全文内容回顾了研究的流程，分析了本文的研究内容的不足并展望了未来可能进行的研究内容。

2 基于深度学习的机械臂抓取理论基础介绍

2.1 引言

本文所提到的改良的抓取位姿估计方法主要包括视觉引导的目标检测和抓取识别算法的框架构建两部分。在研究相关的算法部分之前，首先要对在进行视觉工作时会涉及到的基础模型进行认识和建立，其中包括了相机成像模型、机械臂和相机的手眼模型、还有机械臂的运动学模型；接着对两部分算法部分涉及到的卷积神经网络基础理论、涉及到的可用神经网络类型进行认知和评估工作；然后对任务导向的面向特定目标的机械臂抓取任务进行分析，确定系统的主要功能；最后基于构建的基础模型，根据面向任务的特点设计能满足本文研究目标的机械臂抓取位姿估计系统总体框架。

2.2 面向任务的抓取模型搭建

机械臂在视觉引导下进行六自由度抓取规划^[50]的过程中涉及到了三个基础模型，分别是获取抓取目标信息的相机成像模型^[51]、负责相机坐标系下位姿向机械臂基座标系下位姿转换的手眼模型^[52]、以及映射关节空间和笛卡尔空间位姿的机械臂运动学模型^[53]。本小节将根据实验设备完成这三个基础模型的构建。

2.2.1 相机成像模型

本文的抓取位姿规划是基于相机视觉的结果进行规划，所以相机的输入是必不可少的部分，只有通过相机的图像输入才能进入下一阶段的目标识别分类，因此我们首先需要对相机成像模型进行建模分析。本文使用的视觉传感器是海康工业相机 MV-CU050-60GM，基于小孔成像的原理，将三维世界坐标系中的点映射到相机成像平面内，其中涉及到了四个坐标系的依次转换，包括了世界坐标系、相机坐标系、图像坐标系和像素坐标系，如下图 2-1 所示，摄像机标定坐标系的变换包括了刚体变换（旋转、平移）和透视变换以及线性变换。



图 2-1 坐标系变换关系

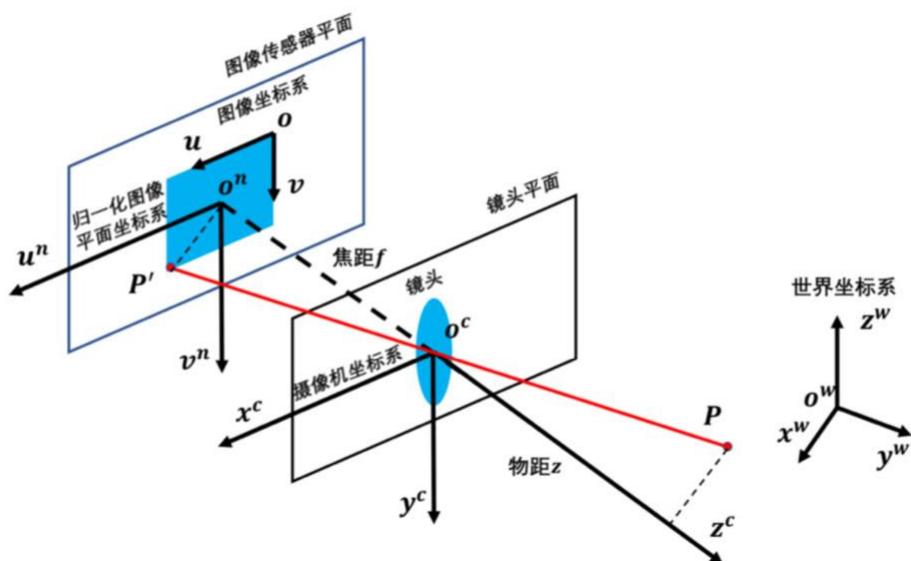


图 2-2 小孔成像原理与各坐标系示意图

本小节将对正投影的计算即描述 3D 点如何进行投影到 2D 像素坐标系进行分析。

首先，从世界坐标系的点通过刚体变换能将其表示为摄像机坐标系的点。

世界坐标系是现实世界中用于描述相机和待测物体实际空间位置的三维直角坐标系 $[x^w \ y^w \ z^w]$ 。它提供了一个参考框架，使我们可以准确地描述物体在空间中的位置和方向。

相机坐标系也是一种三维直角坐标系 $[x^c \ y^c \ z^c]$ ，其原点通常位于相机

镜头的光心位置。在相机坐标系中， x 轴和 y 轴与镜头平面的两边平行， z 轴与镜头的光轴重合且与镜头平面垂直。这个坐标系的选择有助于描述相机内部的光学特性和图像捕获过程。

通过刚体变换，我们可以将点从世界坐标系转变到相机坐标系。刚体变换是一种数学变换，它对物体在空间中的位置和朝向进行变换，而不改变物体的形状。其中，刚体变换包括旋转和平移两个主要操作。旋转操作改变物体的方向和朝向，而平移操作改变物体在空间中的位置。这两个操作通过旋转矩阵 R 和平移矩阵 t 来表示。旋转矩阵 R 描述了物体围绕某个中心点进行的旋转操作。它是一个 3×3 的正交矩阵，其中的元素表示了物体在三个坐标轴上的旋转角度。旋转矩阵 R 保持向量的长度和夹角不变，因此可以保持物体的形状不变。平移矩阵 t 描述了物体在空间中的平移操作。它是一个 3×1 的向量，表示物体在三个坐标轴上的平移距离。平移操作只改变物体的位置，而不改变物体的朝向和形状。具体变换公式如下式 (2-1)

$$\begin{bmatrix} x_c \\ y_c \\ z_c \end{bmatrix} = R \begin{bmatrix} x_w \\ y_w \\ z_w \end{bmatrix} + t \quad (2-1)$$

$$R = \begin{bmatrix} R_{11} & R_{12} & R_{13} \\ R_{21} & R_{22} & R_{23} \\ R_{31} & R_{32} & R_{33} \end{bmatrix}$$

$$t = \begin{bmatrix} t_1 \\ t_2 \\ t_3 \end{bmatrix}$$

图像坐标系也被称为平面坐标系。反映的是图像中物体的实际物理尺寸，单位是毫米（米），原点一般位于图像的中心。坐标原点为成像平面的中点即相机光轴与成像平面的交点。从摄像机坐标系变换到图像坐标系要通过透视变换的方法，以及归一化方法，从而得到归一化的图像坐标。如下式 (2-2)，首先通过相似三角形原理得出相机坐标系下的点和平面坐标系的关系，使用齐次变换表示如下式 (2-3) 所示

$$\begin{cases} x = \frac{f}{z_c} x_c \\ y = \frac{f}{z_c} y_c \end{cases} \quad (2-2)$$

$$\begin{bmatrix} x \\ y \\ 1 \end{bmatrix} = \frac{1}{z_c} \begin{bmatrix} f & 0 & 0 & 0 \\ 0 & f & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} x_c \\ y_c \\ z_c \\ 1 \end{bmatrix} \quad (2-3)$$

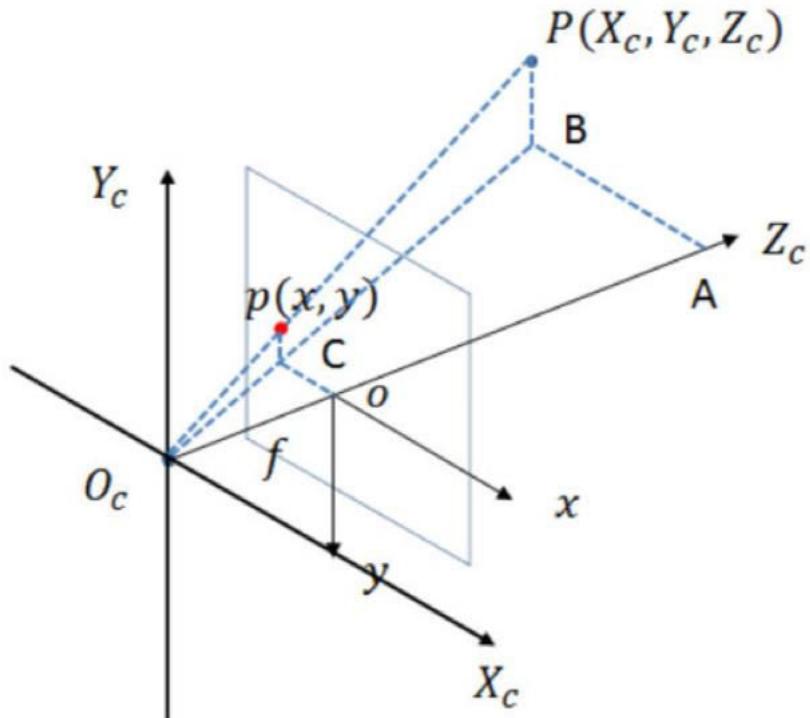


图 2-3 图像坐标系和世界坐标系变换示意图

而图像坐标系的圆心在像素坐标系的表示方式为 $(u_0 \quad v_0)$, 相机采集的图像信息经过一系列变换就最终会以像素阵列形式储存在像素坐标系内, 其坐标原点一般而言都在左上角, 单位是像素 (px)。所以二者的转换不包含旋转, 只是坐标原点和单位不一致, 所以需要通过线性变换将表示在图像坐标系下的点变换到像素坐标系中来, 如下式 (2-4), 在齐次坐标下的表示为

$$\begin{cases} u = \frac{x}{dx} + u_0 \\ v = \frac{y}{dy} + v_0 \end{cases} \quad (2-4)$$

$$\begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = \begin{bmatrix} \frac{1}{dx} & 0 & u_0 \\ 0 & \frac{1}{dy} & v_0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} \quad (2-5)$$

综上所述，通过上面四个坐标系的变换就可以得到点从世界坐标系到像素坐标系的变换表示式 (2-6)

$$z_c \begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = \begin{bmatrix} \frac{1}{dx} & 0 & u_0 \\ 0 & \frac{1}{dy} & v_0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} f & 0 & 0 & 0 \\ 0 & f & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} R & t \\ 0 & 1 \end{bmatrix} \begin{bmatrix} x_w \\ y_w \\ z_w \\ 1 \end{bmatrix} \quad (2-6)$$

所以引入相机内参矩阵 K，外参矩阵 M，图像的深度信息即三维坐标系下的点到光心的距离 d，该深度信息可由深度相机的深度图像得到。

其中，

$$K = \begin{bmatrix} \frac{1}{dx} & 0 & u_0 \\ 0 & \frac{1}{dy} & v_0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} f & 0 & 0 & 0 \\ 0 & f & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} = \begin{bmatrix} f & 0 & u_0 & 0 \\ 0 & \frac{1}{dy} & v_0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} = \begin{bmatrix} f_x & 0 & u_0 & 0 \\ 0 & f_y & v_0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \quad (2-7)$$

$$M = \begin{bmatrix} R & t \\ 0 & 1 \end{bmatrix} \quad (2-8)$$

$$d = z_c \quad (2-9)$$

综上，相机成像的模型可以表示为式 (2-9):

$$d \begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = KM \begin{bmatrix} x_w \\ y_w \\ z_w \\ 1 \end{bmatrix} \quad (2-9)$$

2.2.2 手眼模型

从上述的相机模型的建立以及标定中我们能得到目标在摄像机坐标系下的坐标变换关系，也就是能将目标根据相机坐标系变换到需要的像素坐标系上。但对于本文提到的机械臂抓取位姿估计问题来说，我们要想达到机械臂上的抓取，还需要手眼模型的建立，从而得到相机与工具机械臂之间的变换关系。

手眼模型的建立方式主要由摄像机安装的方式进行分类，大致可以分为两类。一类被称为眼在手上，即在这类手眼关系中，摄像机是固定在机械臂的末端的，随着机械臂的运动相机也随着进行移动，相机相对于机械臂的位置不变。

这种方法的优势在于摄像机能对要抓取的目标物体进行多角度的观察，相机观察过程中不易受到遮挡，灵活多变。该类手眼关系下，模型的建立更多关注于相机坐标系和机械臂末端坐标系的转换关系。另一类则被称为眼在手外，在这种手眼关系中，相机的观察位置、观察角度是固定不变的，相机相对于机械臂的基座坐标系固定不变，这类安装方式中，相机的视场可能会在机械臂移动的过程中被其遮挡。该类手眼关系下，模型的建立应该关注相机坐标系和机械臂基座坐标系之间的关系，两种手眼关系的示意图如下图 2-4 所示。

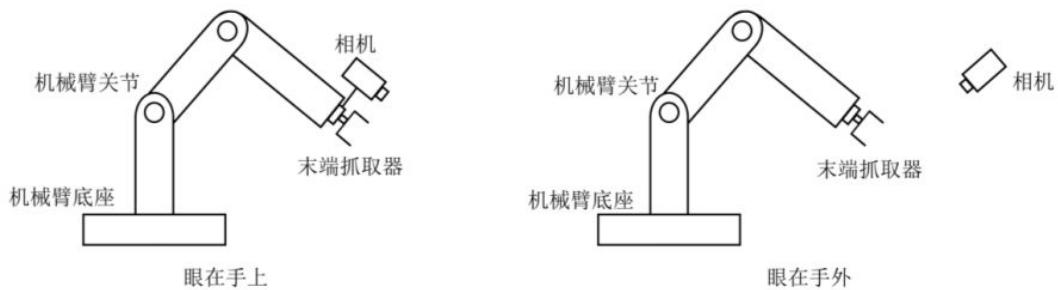


图 2-4 两类手眼关系示意图

本文使用的机械臂为华数工业机器人 HSR-C0605，且本文使用的是眼在手上的安装方式进行抓取作业。

眼在手系统中，进行手眼标定的过程涉及到机械臂基座坐标系、机械臂末端坐标系、相机坐标系和标定板坐标系之间的关系，如图 2-5 所示。为了进行手眼标定，首先需要将标定板放置在相机的视角范围内。其中，标定所用到的标定板通常带有一些已知的特征点或标记，以便在图像中进行识别和定位。在标定过程中，需要移动机械臂到不同的位置，以确保标定板一直在相机的视野内。这样可以获取到不同位置下的标定板图像。同时，在标定过程中需要记录示教器内机械臂末端的坐标以及相机拍摄到的标定板图像。这些数据将用于进行手眼标定算法的计算。手眼标定的目标是确定不同坐标系之间的转换关系，即机械臂基座坐标系、机械臂末端坐标系、相机坐标系和标定板坐标系之间的转换矩阵。这样，当机械臂末端的坐标确定时，就可以通过转换矩阵计算出相机坐标系中的目标物体坐标。在一个手眼系统中，可以得到如下的转换矩阵间的关系：

$$T_{base}^{target} = T_{base}^{tool} T_{tool}^{cam} T_{cam}^{target} \quad (2-10)$$

其中，

T_{base}^{target} 表示目标坐标系其相对于机械臂基坐标系的转换矩阵，

T_{base}^{tool} 表示机械臂末端坐标系相对于机械臂基座标系的变换矩阵，可以通过机械臂示教器读取， T_{tool}^{cam} 表示相机坐标系相对于机械臂末端坐标系的转换矩阵，即所求的手眼矩阵， T_{cam}^{target} 表示目标坐标系相对于相机坐标系的转换矩阵，即相机的外参，这个矩阵可以通过相关图像处理算法得到。

由于在移动过程中，标定板坐标系与机械臂基座标系的相对位置即 T_{base}^{target} 是固定不变的，因此可以构建如下等式 2-11：

$$T_{base^{(1)}}^{tool} T_{tool}^{cam} T_{cam^{(1)}}^{target} = T_{base^{(2)}}^{tool} T_{tool}^{cam} T_{cam^{(2)}}^{target} \quad (2-11)$$

由于基坐标系的刚体变换和相机坐标系的刚体变换即 $T_{base^{(2)}}^{base^{(1)}} T_{cam^{(2)}}^{cam^{(1)}}$ 可知，所以我们可以将上述式子变换为：

$$T_{base^{(2)}}^{tool^{-1}} T_{base^{(1)}}^{tool} T_{tool}^{cam} = T_{tool}^{cam} T_{cam^{(2)}}^{target} T_{cam^{(1)}}^{target^{-1}} \quad (2-12)$$

其中， T_{tool}^{cam} 为未知量且 $T_{base^{(2)}}^{tool^{-1}} T_{base^{(1)}}^{tool} = T_{base^{(2)}}^{base^{(1)}}$ ， $T_{cam^{(2)}}^{target} T_{cam^{(1)}}^{target^{-1}} = T_{cam^{(2)}}^{cam^{(1)}}$ ，都为已知量，通过求解方程可以得到相机坐标系相对于机械臂末端坐标系的转换矩阵，即我们需要的手眼关系。

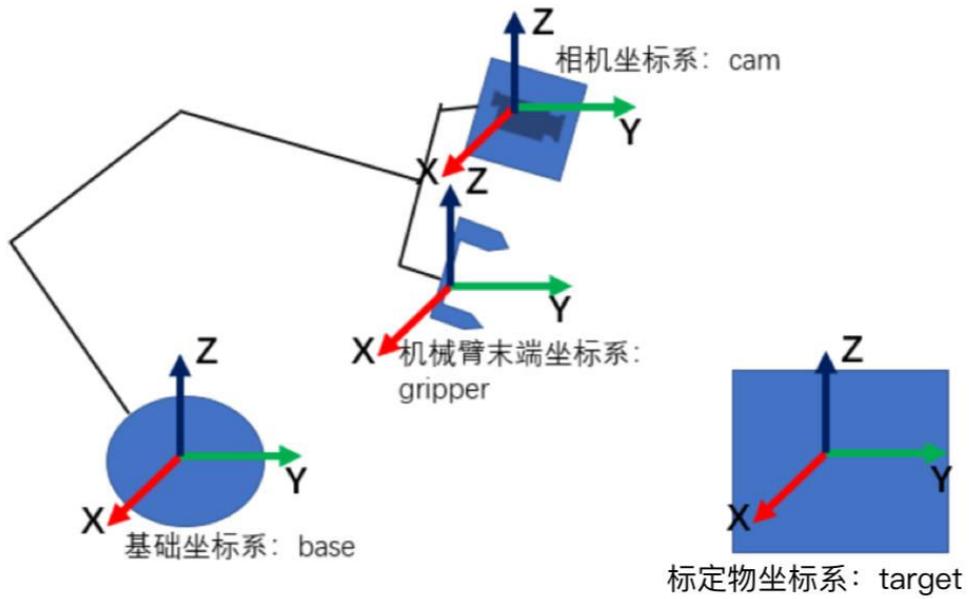


图 2-5 眼在手上坐标系分布示意图

2.2.3 机械臂运动学模型

运动学模型是机械臂进行碰撞检测和路径规划的前提。该模型描述了机械臂关节角度和末端位姿之间的关系，是实现机械臂运动控制和运动规划的基础。

机械臂运动学模型包括正运动学和逆运动学正逆两个方向。其中，正运动学解决的问题是，给定机械臂各个关节的角度，求解机械臂末端的位姿，即确定机械臂末端在笛卡尔空间中的位置和姿态；逆运动学解决的问题是，已知机械臂末端的位姿，求解对应的关节角度。通过逆运动学，可以根据末端所需的位姿，计算出机械臂各个关节所需的角度，从而实现目标位置和姿态的控制。

运动学模型主要研究机械臂的关节角度与机械臂末端在笛卡尔三维空间中的位置姿态关系。它考虑了机械臂的结构和约束条件，通过建立数学模型和运动学方程，描述了关节角度和末端位姿之间的映射关系。这样，就能够根据给定的关节角度计算出末端的位姿，或者根据给定的末端位姿计算出关节角度。本节将对机械臂正运动学和逆运动学分别进行分析：

(1) 机械臂正运动学建模 Denavit 和 Hartenberg 于 1955 年提出了 D-H 模型 (Denavit–Hartenberg parameters)，该模型通过在机械臂各个连杆上建立坐标系来描述了机器人连杆与关节之间关系。如图 2-6，本小节为每个连杆建立了坐

标系，如图所示，连杆*i*的坐标系为 $\{i\}$ ，该坐标系的 z_i 轴与*i*即关节中心线重合； x_i 轴与连接两个连杆的关节中心线的公垂线重合，方向由关节*i*指向关节*i+1*； y_i 轴由右手法则决定；原点为 x_i 轴和 z_i 轴的交点。由图 2-6 可得连杆*i-1*和*i*的坐标系和之间的关系。

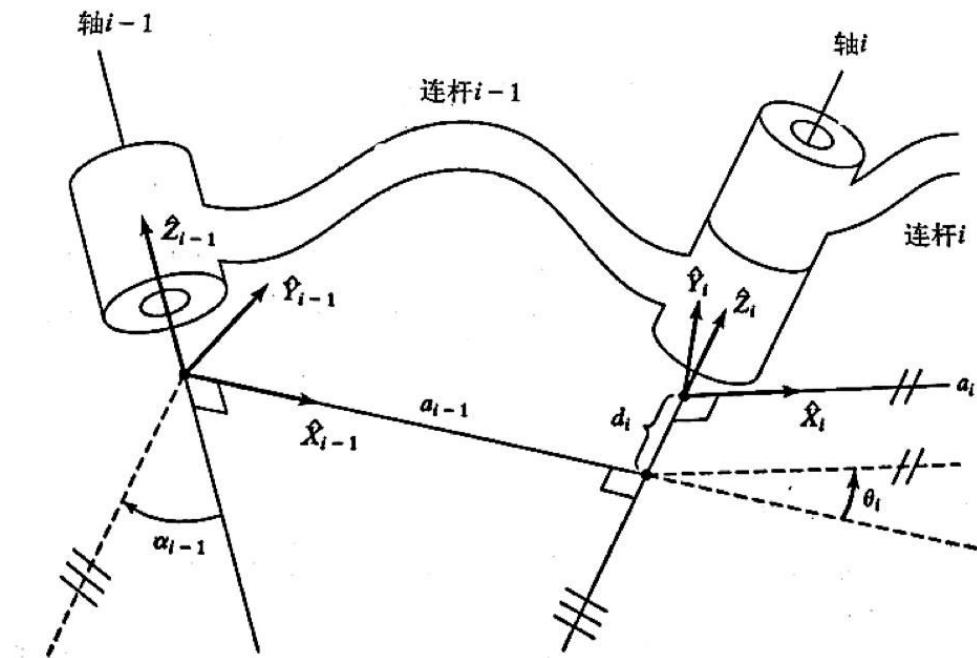


图 2-6 机械臂和连杆之间的关系

其中各个参数的具体含义如下：

a_i : 从 z_i 到 z_{i+1} 沿着 x_i 测量的距离；

α_i : 从 z_i 到 z_{i+1} 绕着 x_i 旋转的角度；

d_i : 从 x_{i-1} 到 x_i 沿着 z_i 测量的距离；

θ_i : 从 x_{i-1} 到 x_i 绕着 z_i 旋转的角度；

通常选择 $a_i \geq 0$ ，因为它代表连杆长度，而 α_i , d_i , θ_i 可正可负。通过上述四个参数，坐标系 $\{i\}$ 相对于坐标系 $\{i-1\}$ 的变换矩阵可以通过四次坐标系刚体变换得到式 (2-13)：

$${}^{i-1}T_i = \text{Rot}(x, \alpha_{i-1}) \text{Trans}(x, a_{i-1}) \text{Rot}(z, \theta_i) \text{Trans}(z, d_i)$$

$$\begin{aligned}
&= \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & c\alpha_{i-1} & -s\alpha_{i-1} & 0 \\ 0 & s\alpha_{i-1} & c\alpha_{i-1} & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & a_{i-1} \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} c\theta_i & -s\theta_i & 0 & 0 \\ s\theta_i & c\theta_i & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & d_i \\ 0 & 0 & 0 & 1 \end{bmatrix} \quad (2-13) \\
&= \begin{bmatrix} c\theta_i & -s\theta_i & 0 & a_{i-1} \\ s\theta_i c\alpha_{i-1} & c\theta_i c\alpha_{i-1} & -s\alpha_{i-1} & -s\alpha_{i-1} d_i \\ s\theta_i s\alpha_{i-1} & c\theta_i s\alpha_{i-1} & c\alpha_{i-1} & c\alpha_{i-1} d_i \\ 0 & 0 & 0 & 1 \end{bmatrix}
\end{aligned}$$

其中，s 是三角函数 sin 的缩写，c 是三角函数里 cos 的缩写，下式同。

本文使用的 D-H 参数表如下述：

表 2-1 机械臂 D-H 参数表

连杆	a_{i-1} (mm)	d_i (mm)	α_{i-1} (deg)	θ_i (deg)
1	0	0	0	$\theta_1(-240,240)$
2	0	0	-90	$\theta_2(-110,130)$
3	a_2	0	0	$\theta_3(0,162)$
4	a_3	d_4	-90	$\theta_4(-200,200)$
5	0	0	90	$\theta_5(-120,120)$
6	0	0	-90	$\theta_6(-360,360)$

其中， $a_2=435$, $a_3=50$, $d_4 = 470$ 。

我们可以通过上式 2-13，以及上述的 D-H 参数表来求任意的相邻连杆坐标系之间的变换，通过正向运动学推导得到机械臂末端坐标系相对于机械臂基坐标系的位姿。

将表中 D-H 参数代入式 (2-13) 中，然后正向推导对接下来的关节进行相邻关节矩阵的依次相乘，可以得到机械臂末端坐标系在基座标系中的位置和朝向如下式：

$${}^0T = {}^1T_1 {}^2T_2 {}^3T_3 {}^4T_4 {}^5T_5 {}^6T = \begin{bmatrix} n_x & o_x & a_x & p_x \\ n_y & o_y & a_y & p_y \\ n_z & o_z & a_z & p_z \\ 0 & 0 & 0 & 1 \end{bmatrix} \quad (2-14)$$

其中，

$${}^1T = \begin{bmatrix} c\theta_1 & -s\theta_1 & 0 & 0 \\ s\theta_1 & c\theta_1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}, \quad {}^2T = \begin{bmatrix} c\theta_2 & -s\theta_2 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ -s\theta_2 & -c\theta_2 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

$$\begin{aligned} {}^2_3 T &= \begin{bmatrix} c\theta_3 & -s\theta_3 & 0 & a_2 \\ s\theta_3 & c\theta_3 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}, \quad {}^3_4 T = \begin{bmatrix} c\theta_4 & -s\theta_4 & 0 & a_3 \\ 0 & 0 & 1 & d_4 \\ -s\theta_4 & -c\theta_4 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \\ {}^4_5 T &= \begin{bmatrix} c\theta_5 & -s\theta_5 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ s\theta_5 & c\theta_5 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}, \quad {}^5_6 T = \begin{bmatrix} c\theta_6 & -s\theta_6 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ -s\theta_6 & -c\theta_6 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \end{aligned}$$

进一步进行推导可得：

$$\begin{cases} n_x = s_1(c_4s_6 - s_4c_5) - c_1(c_{23}(s_4s_6 + c_4c_5) - s_{23}s_5c_6) \\ n_y = c_1(s_4c_5 - c_4s_6) - s_1(c_{23}(s_4s_6 + c_4c_5) + s_{23}s_5c_6) \\ n_z = s_{23}(c_4c_5 + s_4c_6) - c_{23}s_5c_6 \\ o_x = s_1(c_4c_6 - s_4c_5) - c_1(c_{23}(s_4c_6 + c_4c_5) + s_{23}s_5s_6) \\ o_y = c_1(s_4c_5 - c_4c_6) - s_1(c_{23}(s_4c_6 + c_4c_5) - s_{23}s_5c_6) \\ o_z = s_{23}(c_4c_5 + s_4c_6) + c_{23}s_5c_6 \\ a_x = -c_1(c_{23}c_4s_5 + s_{23}c_5) - s_1s_4s_5 \\ a_y = -s_1(c_{23}c_4s_5 + s_{23}c_5) + c_1s_4s_5 \\ a_z = c_4s_5s_{23} - c_5c_{23} \end{cases} \quad (2-15)$$

$$\begin{cases} p_x = c_1(c_2a_2 - s_{23}d_4) + c_1c_{23}a_3 \\ p_y = s_1(c_2a_2 - s_{23}d_4) + s_1c_{23}a_3 \\ p_z = -s_2a_2 - a_3s_{23} - d_4c_{23} \end{cases}$$

式子中， s_1 为 $\sin\theta_1$ 的缩写， c_1 为 $\cos\theta_1$ 的缩写， s_{23} 为 $\sin(\theta_2 + \theta_3)$ 的缩写， c_{23} 为 $\cos(\theta_2 + \theta_3)$ 的缩写。由式子(2-15)和式子(2-16)连立即可获得六轴抓取机械臂的正运动学模型

(2) 机械臂逆运动学建模

机械臂逆运动学解也叫做机械臂的运动学反解，是机器臂抓取工作时运动规划的基础，在本文中的抓取位姿估计后，机械臂需要通过逆运动学模型从而规划如何到达位姿估计函数输出的末态位置和位姿，所以建立机械臂逆运动学反解模型是很重要的。逆运动学主要是通过机械臂末端在笛卡尔坐标系下的位姿确定关节变量，正运动学的解是唯一确定的，但是由于多自由度机械臂的关节通常是冗余的，因此逆运动学的解一般不是唯一的，一个位姿可以反解出多组解。求解逆运动学问题时，通常使用代数法进行求解，首先已知机械臂末端位姿，即：

$${}^0_6T = {}^0_1T(\theta_1) {}^1_2T(\theta_2) {}^2_3T(\theta_3) {}^3_4T(\theta_4) {}^4_5T(\theta_5) {}^5_6T(\theta_6) = \begin{bmatrix} n_x & o_x & a_x & p_x \\ n_y & o_y & a_y & p_y \\ n_z & o_z & a_z & p_z \\ 0 & 0 & 0 & 1 \end{bmatrix} \quad (2-16)$$

将 ${}^0_1T(\theta_1)$ 的逆变换 ${}^0_1T(\theta_1)^{-1}$ 左乘方程 (2-16) 可以得到以下式子:

$${}^0_1T(\theta_1)^{-1} {}^0_6T = {}^1_2T(\theta_2) {}^2_3T(\theta_3) {}^3_4T(\theta_4) {}^4_5T(\theta_5) {}^5_6T(\theta_6) \quad (2-17)$$

逆矩阵 ${}^0_1T(\theta_1)^{-1}$ 可以通过求逆得出:

$${}^0_1T(\theta_1)^{-1} = \begin{bmatrix} c_1 & s_1 & 0 & 0 \\ -s_1 & c_1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \quad (2-18)$$

将式子 (2-18) 代入式 (2-17) 可得:

$$\begin{bmatrix} c_1 & s_1 & 0 & 0 \\ -s_1 & c_1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} n_x & o_x & a_x & p_x \\ n_y & o_y & a_y & p_y \\ n_z & o_z & a_z & p_z \\ 0 & 0 & 0 & 1 \end{bmatrix} = {}^1_6T \quad (2-19)$$

1_6T 可以通过求解式子 (2-13) 得出, 接着令方程 (2-19) 两侧 (2, 4) 位置的元素对应相等得出下式:

$$-\sin\theta_1 p_x + \cos\theta_1 p_y = 0 \quad (2-20)$$

利用三角代换的:

$$p_x = \rho \cos\phi, p_y = \rho \sin\phi \quad (2-21)$$

上式中: $\rho = \sqrt{p_x^2 + p_y^2}; \phi = \text{Atan2}(p_y, p_x) = \text{Aarctan}\left(\frac{p_y}{p_x}\right)$ 将式 (2-21) 代入式 (2-20), 就可以得到 θ_1 的解:

$$\theta_1 = \phi = \text{Atan2}(p_y, p_x) \quad (2-22)$$

令方程 (2-19) 两侧 (1, 4) 和 (3, 4) 位置的元素对应相等可以得到 θ_3 的值:

$$\theta_3 = \text{Atan2}(a_3, d_4) - \text{Atan2}\left(k, \pm\sqrt{a_3^2 + d_4^2 - k^2}\right) \quad (2-23)$$

其中,

$$k = \frac{p_x^2 + p_y^2 + p_z^2 - a_2^2 - a_3^2 - d_4^2}{2a_2} \quad (2-24)$$

由 (2-23) 和 (2-24) 可以得出, θ_3 具有两个解。

接下来继续求解其余角, 已知 θ_1 和 θ_3 , 使用逆矩阵 ${}^0_3T^{-1}$ 左乘方程 (2-16) 可

得出下列等式：

$${}^0_3T^{-1} {}^0_6T = {}^3_4T(\theta_4) {}^4_5T(\theta_5) {}^5_6T(\theta_6) \quad (2-25)$$

等同于：

$$\begin{bmatrix} c_1c_{23} & s_1c_{23} & -s_{23} & -a_2c_3 \\ -c_1s_{23} & -s_1s_{23} & -c_{23} & a_2s_3 \\ -s_1 & c_1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} n_x & o_x & a_x & p_x \\ n_y & o_y & a_y & p_y \\ n_z & o_z & a_z & p_z \\ 0 & 0 & 0 & 1 \end{bmatrix} = {}^3_6T \quad (2-26)$$

3_6T 可通过求解(2-13)得出，令方程(2-26)两侧(1, 4)和(2, 4)位置的元素对应相等可以得出：

$$\begin{aligned} \theta_{23} = \text{Atan2} [& (-a_3 - a_2c_3)p_z + (c_1p_x + s_1p_y)(a_2s_3 - d_4), (-d_4 + a_2s_3)p_z \\ & + (c_1p_x + s_1p_y)(a_2c_3 + a_3)] \end{aligned} \quad (2-27)$$

且，

$$\theta_{23} = \theta_2 + \theta_3 \quad (2-28)$$

且 θ_3 已知对应两个解，代入上式可得：

$$\theta_2 = \theta_{23} - \theta_3 \quad (2-29)$$

所以 θ_2 也对应两个解。

下面继续令(2-26)两侧(1, 3)和(3, 3)位置的元素对应相等可得 θ_4 ，若 $s_5 \neq 0$ 则可得出：

$$\theta_4 = \text{Atan2}(-a_xs_1 + a_yc_1, -a_xc_1c_{23} - a_ys_1c_{23} + a_zs_{23}) \quad (2-30)$$

而当 $s_5 = 0$ 时，关节4的中心线会与关节6的中心线重合，处于机械臂的奇异位置， θ_4 的值可以任意选择。得出 θ_4 后，可以进一步求解出 θ_5 ， θ_6 的值。

进一步使用逆矩阵 ${}^0_4T^{-1}$ 左乘方程(2-16)可以得出下列等式：

$${}^0_4T^{-1} {}^0_6T = {}^4_5T(\theta_5) {}^5_6T(\theta_6) \quad (2-31)$$

其中， ${}^0_4T^{-1}$ 与 4_6T 的求解方法都与上述方法相同，令(2-26)两侧(1, 3)和(3, 3)位置的元素对影响的可以得出：

$$\theta_5 = \text{Atan2}(ss_5, cc_5) \quad (2-32)$$

其中，

$$ss_5 = -a_x(c_1c_{23}c_4 + s_1s_4) - a_y(s_1c_{23}c_{54} - c_1s_4) + a_zs_{23}c_4 \quad (2-33)$$

$$cc_5 = -a_x c_1 s_{23} - a_y s_1 s_{23} - a_z c_{23} \quad (2-34)$$

最后，还是通过逆矩阵 ${}^0T^{-1}{}^0T$ 左乘方程（2-16）得出：

$${}^0T^{-1}{}^0T = {}^5T(\theta_6) \quad (2-35)$$

令方程（2-35）两侧（3, 1）和（1, 1）位置的元素对应相等可以解出：

$$\theta_6 = Atan2(ss_6, cc_6)$$

综上所述，本小节求得了本文所使用的机械臂逆运动学模型，从而得知了各个关节的角度信息，从而完善了机械臂位姿估计的流程。

2.3 卷积神经网络

卷积神经网络（CNN）是一种被广泛应用于图像分割、分类和检测等任务的神经网络结构。它具有处理图像数据的能力，并在许多计算机视觉任务中取得了显著的成果。

CNN 的主要结构包括输入层、卷积层、池化层、激活函数和全连接层等组件，其结构内容如图 2-7 所示。大致来说，输入层的作用是接收图像的像素矩阵或多维矩阵数据，并将其传递给下一层进行处理。卷积层是 CNN 的核心组件，通过应用卷积操作提取图像中的特征。池化层用于减小特征图的尺寸并保留重要的特征信息。激活函数引入非线性变换，增加网络的表达能力。全连接层将特征映射转换为最终的输出结果，例如该构架通常用来输出图像的分类标签或边界框的坐标。每一层在 CNN 中都有特定的功能，通过一层层的处理和特征提取，CNN 能够学习到输入图像中的高级抽象特征，并用于解决各种视觉任务。

CNN 通过逐层处理高效地提取图像中的特征信息，在各种图像处理任务中取得了显著的成果。它能够自动学习并捕捉图像中的局部和全局特征，具有较强的表达能力和泛化能力。由于 CNN 的结构特点和学习能力，它适用于图像分类、物体检测、语义分割等多个领域，并在这些任务中取得了重要的突破。

总之，卷积神经网络是一种强大的图像处理工具，通过逐层处理和特征提取，能够高效地处理图像数据并解决各种图像处理任务。它的应用广泛，并对计算机视觉领域的发展做出了重要贡献。

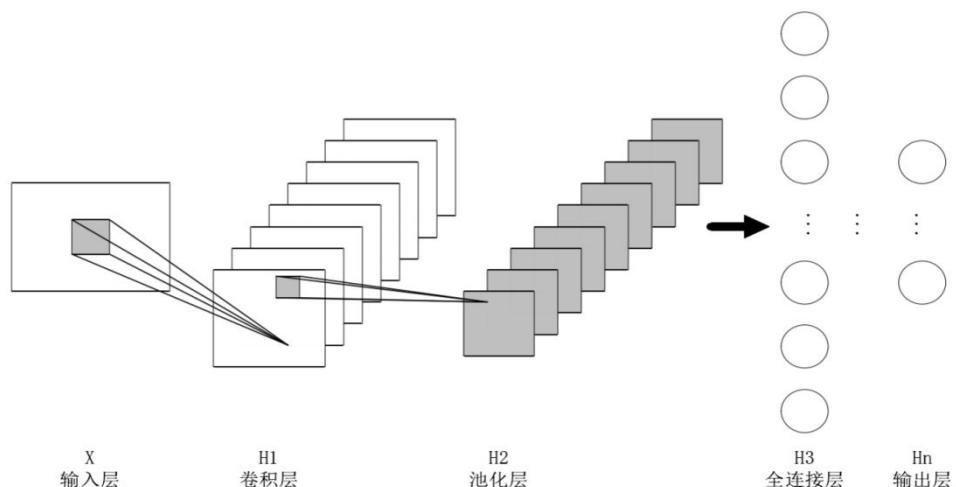


图 2-7 卷积神经网络模型

2.3.1 卷积层

卷积层是卷积神经网络（CNN）中最为重要的组成部分之一。它通过卷积操作对输入图像进行处理，从而学习图像的特征信息，并提取出关键的特征，同时保留了整体的信息。

在卷积层中，卷积核这个部件起到了核心的作用。卷积核实质上是一个小的滤波器，它通过与输入图像进行逐元素的乘积和求和操作，来获取局部的特征信息。卷积操作可以理解为对图像进行滑动窗口的操作，通过不断调整卷积核的位置，提取出图像的不同特征。

卷积核内部包含一组可学习的参数，这些参数在训练过程中通过反向传播算法进行调整和优化。通过不断调整卷积核的参数，卷积层可以学习到不同的特征过滤器，从而对输入图像中的不同特征进行感知和提取。

常见的卷积核尺寸包括 3×3 和 5×5 ，不同的尺寸选择会影响卷积层对图像特征的感知范围。较小的卷积核尺寸可以捕捉到更细节的特征，而较大的卷积核尺寸可以覆盖更广阔的感受野，获取更宏观的特征信息。选择适当的卷积核尺寸对于任务的成功和性能至关重要。

通过逐层堆叠多个卷积层，网络可以逐渐提取出更加抽象和高级的特征表示。这种分层的特征提取过程使得卷积神经网络能够有效地学习到图像的语义信息，并在各种图像处理任务中取得优秀的性能。

图 2-8 为卷积运算的过程,以尺寸为 4×4 的输入矩阵为例,矩阵中的数值分别

为 X_1, X_2, \dots, X_9 。卷积核的尺寸为 3×3 , 对应的权重为 W_1, W_2, \dots, W_9 。卷积的步长被设定设为 1, 通过对 3×3 的输入矩阵使用卷积核进行运算, 得到 y , 可以用下面的公式(2-36)表示:

图 2-8 展示了一个卷积运算的过程, 以一个尺寸为 4×4 的输入矩阵为例。这个过程使用一个 3×3 的卷积核对输入矩阵进行运算, 得到输出结果。

输入矩阵的大小为 4×4 , 其中的数值表示为 X_1, X_2, \dots, X_9 。卷积核的尺寸也为 3×3 , 并且对应的权重为 W_1, W_2, \dots, W_9 。卷积运算的目标是通过对输入矩阵和卷积核进行运算, 得到输出矩阵 y 。

在这个卷积运算中, 步长被设定为 1, 即通过在输入矩阵上以步长为 1 的间隔滑动卷积核, 对应的权重与输入矩阵中的元素进行逐元素相乘, 并将结果相加, 得到输出矩阵 y 。

公式(2-36)可以用来表示这个卷积运算的过程:

$$y = X_1 \times W_1 + X_2 \times W_2 + \dots + X_9 \times W_9 = \sum_{i=1}^9 X_i \times W_i \quad (2-36)$$

每次卷积操作完成后, 保持卷积核不变, 改变 3×3 的输入矩阵的位置, 方向可以变换, 输出矩阵的大小会相应地减小。按照公式(2-36)的步骤重复四次, 就可以得到一个大小为 2×2 的输出矩阵。

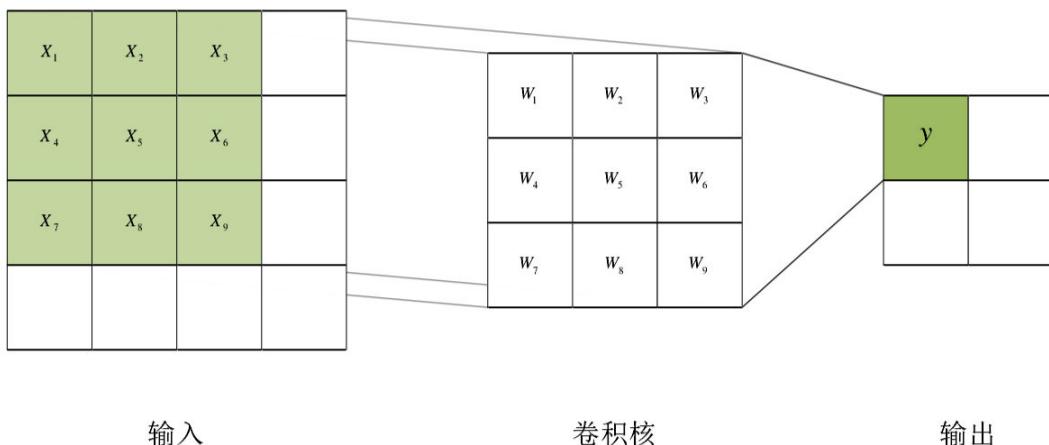


图 2-8 卷积过程示例

2.3.2 池化层

池化层是卷积神经网络中的一个重要组件, 通常位于卷积层之后, 也被称为下采样层。它在图像处理和特征提取过程中起到了关键的作用。

池化层的主要作用是降低特征层的空间分辨率, 并提取具有空间不变性的

特征。通过将局部区域合并成单个值的方式，池化层能够减少特征图的尺寸，同时保留重要的特征信息。

在池化操作中，最常见的方法是最大池化和平均池化，它们在卷积神经网络中起到重要的作用。最大池化是其中一种常用的池化操作，它从局部区域中选择最大值作为输出特征。在最大池化过程中，输入特征图被分割成不重叠的局部区域，然后该模块会在每个区域中选择最大值作为代表性特征。这种方式可以有效地保留图像中的主要特征，并减少特征图的尺寸。另一种常用的池化操作是平均池化，它计算局部区域内像素值的平均值作为输出特征。平均池化将输入特征图分割成不重叠的局部区域，然后计算每个区域中像素值的平均值作为池化后的特征值。平均池化有助于平滑特征图，并减少对细节的敏感性。

上述这两个操作都有助于减少网络的参数数量，降低计算负担，并提高模型的运算效率。

池化层在卷积神经网络中的另一个重要的作用是它的使用减少了过拟合问题。这是通过降低特征层的维度来实现的，从而减少需要拟合的参数数量，降低模型的复杂度。具体而言，池化操作通过将局部区域合并为单个值的方式，减少了特征图的尺寸和维度。这样可以减轻模型对输入数据的细节敏感性，从而提高模型的泛化能力。通过降低特征层的维度，池化层有效地减少了模型需要学习的参数数量，降低了过拟合的风险。

最大池化和平均池化的具体示例过程如图 2-9 所示。

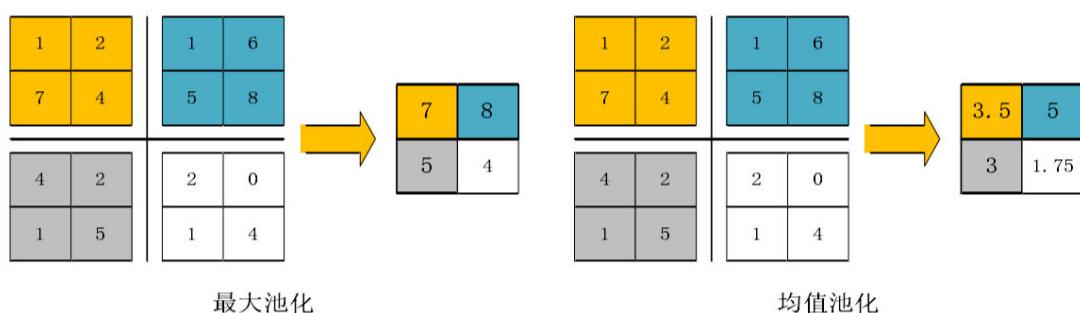


图 2-9 池化过程

2.3.3 激活函数

激活函数(Activation functions)在深度学习网络中扮演着重要的角色，其主要作用是引入非线性因素，使网络能够学习更为复杂的结构^[37]。

在卷积层中加入激活函数的作用是将线性组合的结果转化为非线性的输出。这是因为卷积操作本身是一种线性运算，如果没有激活函数的引入，网络的表达能力将受限于线性变换，无法捕捉到复杂的特征和模式。

常用的激活函数主要有以下几种：

(1) Sigmoid 函数

Sigmoid 函数是一种常用的激活函数，它的取值范围在 0 到 1 之间，导数范围在 0 到 0.25 之间。Sigmoid 函数具有对称性和易于求导的特点，在许多场景中得到广泛应用。它可以将输入值映射到一个概率分布，常用于二分类问题或将输出限制在一定范围内的任务。

然而，在神经网络中使用 Sigmoid 函数可能会引发梯度爆炸或梯度消失的问题。这是因为 Sigmoid 函数在接近饱和区域时，导数接近于 0，导致梯度逐渐消失。当权重初始化在 $[0,1]$ 范围内且网络层数较深时，梯度逐渐趋近于 0，导致梯度消失现象，影响反向传播和训练过程。另一方面，如果权重初始化为随机值且范围为 $(0, +\infty)$ ，可能会出现梯度爆炸现象。在反向传播过程中，梯度值会不断放大，导致网络参数的更新失控，训练过程无法正常进行。

Sigmoid 函数可以表示为：

$$f(x) = \frac{1}{1 + e^{-x}} \quad (2-37)$$

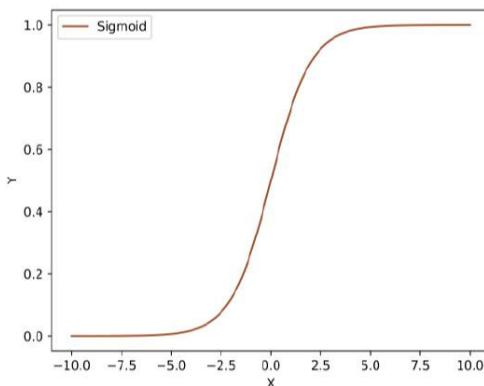


图 2-10 Sigmoid 函数图

(2) Tanh 函数

图 2-11 显示了 Tanh 函数及其导数。

与 Sigmoid 函数相比，Tanh 函数解决了 Sigmoid 函数的原点对称问题，将 0 作为输出的均值。Tanh 函数的取值范围在 -1 到 1 之间，具有 S 形曲线的特点。

它在输入为负时输出负数，在输入为正时输出正数。相比于 Sigmoid 函数，Tanh 函数的输出范围更广，更接近于正态分布。

Tanh 函数在某些情况下可以比 Sigmoid 函数更好地表达非线性关系，有效地减少了迭代次数。它具有较大的梯度，使得网络的学习过程更快速。

然而，尽管 Tanh 函数在一定程度上解决了梯度消失问题，但仍未能完全解决。当网络层数较深时，梯度仍然会逐渐趋近于 0，导致梯度消失现象的出现。这会影响反向传播和网络的训练效果。

Tanh 函数的表达式如下式 (2-38) 所示：

$$f(x) = \tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}} \quad (2-38)$$

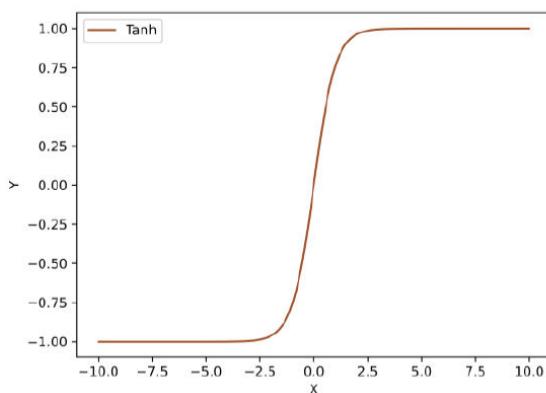


图 2-11 Tanh 函数图

(3) ReLU 函数

ReLU 函数是目前深度神经网络中最常用的激活函数之一。它的表达式如式 (2-39) 所示。

ReLU 函数的输出特性可以通过图 2-11 来观察和理解。当输入 x 大于 0 时，ReLU 函数的输出等于输入 x ；当输入 x 小于等于 0 时，ReLU 函数的输出为 0。因此，ReLU 函数在正数部分保持线性增长，能够更好地捕捉到数据的非线性关系。

此外，ReLU 函数的导数特性也值得注意。当输入 x 大于 0 时，ReLU 函数的导数为 1；当输入 x 小于等于 0 时，ReLU 函数的导数为 0。这意味着在正数部分，ReLU 函数不会引入梯度消失问题，梯度保持不变；而在负数部分，ReLU 函数的梯度为 0，可能导致部分神经元无法更新。

$$f(x) = \max(0, x) \quad (2-39)$$

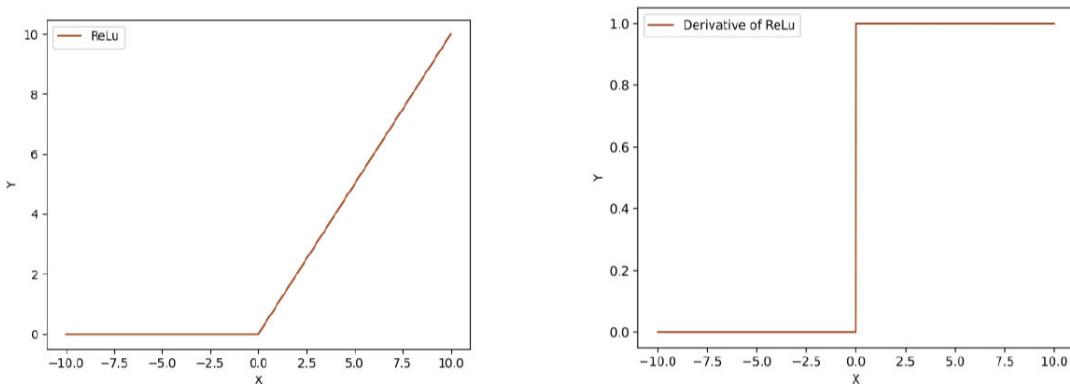


图 2-12 ReLU 函数及其导函数图

2.3.4 全连接层

全连接层是卷积神经网络中位于顶部的一层，其主要作用是充当分类器，将卷积层和池化层提取的特征整合和映射，输出与任务相关的预测结果。在全连接层中，每个神经元都与前一层的所有神经元相连接，这种连接方式赋予了全连接层强大的拟合能力，使其能够学习到输入特征之间更复杂的关联关系。

然而，全连接层的参数量巨大，容易导致过拟合问题。为了解决这个问题，常常会使用正则化方法或者对全连接层的复杂度进行控制，如添加正则化项或使用 Dropout 等技术。这些方法有助于减少全连接层的参数数量，提高模型的泛化能力。

在全连接层的最后一层，通常采用 softmax 激活函数，将网络的输出转化为类别的概率分布。这样做可以使网络直接输出各个类别的预测概率，便于进行多类别分类任务。

在某些特定任务中，如物体检测和分割，可以采用全连接层的变种，例如全局平均池化层。全局平均池化层将整个特征图的空间维度降低为 1，从而显著减少了网络的参数量和计算复杂度，同时保留了关键特征信息。

2.4 本章小结

本章首先对基于深度学习的机械臂抓取理论基础进行了解释，并且提出了理论基础所需要介绍的内容。然后对视觉工作中会涉及到的相机成像模型、机械臂和相机坐标系转换所需的手眼模型还有机械臂的运动学模型进行了详细的解释和建模。随后对后续两部分算法测试改进和融合时所会涉及到的内容进行

了讲解，主要包括深度学习神经网络的卷积层、池化层、激活函数、全连接层等结构。

3 基于多注意力机制的改进 YOLOv5 模型

3.1 引言

目标检测算法的选取与改进在基于任务目标的抓取检测算法研究中显得尤为重要，这主要体现在目标检测算法能提前将目标从堆叠或散乱排放的物品中识别出来，并且排除掉其他的非抓取对象的无用信息，既能够加速抓取检测算法的检验流程，又能够为抓取位姿估计算法排除掉无用的信息，增加了其成功输出结果的成功率。

本章首先介绍了 YOLO 目标检测函数的发展历代版本，引出了 YOLOv5 这种单阶段算法的优势以及最后使用这个算法作为改进算法的框架的理由，接着对 YOLOv5 的各个部分进行了详细的介绍，以及对各个部分的改进方向尤其是主干以及损失函数等方向进行了梳理。具体的改进方向是优化了目标检测算法的主干部分，并对其损失函数进行了改良，并增加了注意力机制来促使模型更好的定位目标。接着本文介绍了检测 YOLOv5 改进情况的数据集，以及使用该数据集下的改进算法和未改进算法的结果对比，并陈述了对将要在实验流程中进行抓取的数据集的制作及针对实验具体对象的目标检测算法的训练过程。

3.2 基于深度学习的目标检测算法

3.2.1 YOLOv5 与历代 YOLO 的比较与优势

YOLO (You Only Look Once)^[13]是一种经典的单阶段目标检测算法，图 3-1 为 YOLO 的结构图，它有许多版本。这些版本包括 YOLOv1、Fast YOLO、YOLOv2、YOLOv3 等等，每个版本都在不同方面进行了改进和优化。

YOLOv1 是最早的 YOLO 版本，它通过将图像划分为网格，并使用边界框和置信度来定位和分类目标物体。然后，通过非最大值抑制来生成最终的检测结果。尽管 YOLOv1 在准确性方面表现出色，但其检测速度相对较慢。

为了提高检测速度，研究人员引入了 Fast YOLO，Fast YOLO 是基于 YOLOv1 的一种轻量化处理算法。该网络采用了一些改进措施，使得算法在保持较高准确性的同时，能够更快地进行目标检测。

接下来问世的 YOLOv2^[14]是对 YOLOv1 的改进版本，该模型引入了

Darknet-19 特征提取网络和 anchor boxes 的概念^[55]。其中，Darknet-19 是一种具有 19 个卷积层的网络结构，用于提取特征，而 anchor boxes 则被用于更好地处理不同比例和形状的目标物体。

进一步改进的 YOLOv3^[15]在快速、准确和泛化等方面进行了优化。它引入了 Darknet-53 作为主干特征提取网络，该网络由 53 个卷积层组成。YOLOv3 还使用了残差连接这种跨层连接技术来提高特征的传递效果，该跨层连接技术使层与层间的信息能够更好的传递和利用，避免了梯度消失问题。YOLOv3 还增加了更多的锚框，即一种预定义的边界框来提高检测性能，这种边界框能够让网络更方便的检测目标的位置和尺度。

YOLOv4 是对 YOLOv3 的改进版本，主要改进了残差块的设计。它引入了大残差块、SPP（Spatial Pyramid Pooling）和 PANNnet 网络，并采用了新的激活函数 Mish，新的损失函数 CIOU。此外，YOLOv4 还应用了 Mosaic 数据增强、学习率余弦退火衰减和标签平滑等操作，进一步提升了检测性能和鲁棒性。

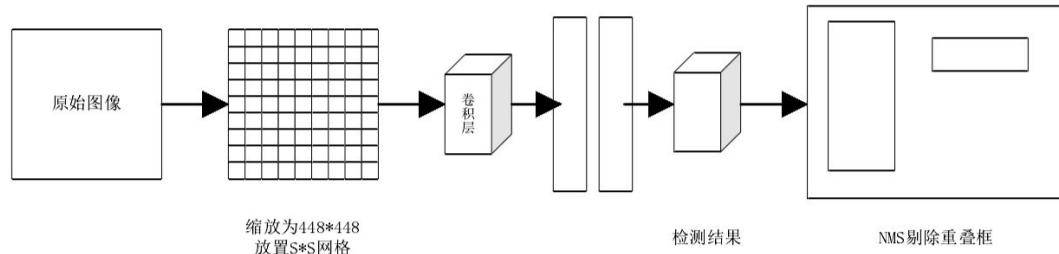


图 3-1 YOLO 架构

YOLOv4 之后新一代问世的版本就被称为 YOLOv5，它对比于 YOLOv4 的提升有许多，其中一个很明显的就是数据增强方面。

在 YOLOv5 算法框架里，数据加载器在数据增强方面起着关键作用，这个模块通过传递和增强每个训练批次的数据来提高模型的训练效果。数据加载器对每个训练批次的数据进行三种主要的数据增强操作：首先是缩放操作，该模块会将输入图像的尺寸调整到适合模型的大小，以便进行有效的训练。其次是色彩空间调整，该模块会通过对图像的亮度、对比度、饱和度等进行调整，增强模型对不同光照条件下的目标检测的鲁棒性。最后是马赛克增强，通过在图像中随机插入马赛克块，会有效解决目标检测任务中的小目标问题错误!未找到

引用源。, 提高对小目标的检测精度和鲁棒性。其中, 马赛克数据增强是 YOLOv5 相对于 YOLOv4 的一个比较重要的新增功能。这个功能使得模型在训练过程中能够学习到目标物体的局部特征和上下文信息, 从而提高了 YOLOv5 网络对小目标的检测效果。

此外, YOLOv5 还引入了自适应锚定框的功能, YOLOv5 的自适应锚定框实际上是对 YOLOv4 的锚定框进行改进的结果。自适应锚定框能够根据数据集中目标的尺寸分布情况, 自动调整锚定框的大小和比例, 从而提高了模型对不同大小目标的识别能力和检测精度。

其次, YOLOv5 是基于 PyTorch 框架构建的目标检测模型, 相比较 Darknet 框架而言, YOLOv5 在环境搭建方面更加友好, 更容易用于训练自定义数据集, 并且相对于 Darknet 框架更易于投入生产环境中使用。

而且, 相较于同类型的竞品模型, YOLOv5 在网络结构、激活函数和优化函数的选择等方面有显著的提升。

在网络结构的改进上, YOLOv5 的网络结构设计可以说是简洁明了, 并且它拥有多种内置版本作为训练的选择, 包括 YOLOv5 s、YOLOv5 m、YOLOv5 l 和 YOLOv5 x 四种模型。这些模型在网络结构上是相同的, 区别在于模型的深度和卷积核个数的不同。通过调整这些参数, 实际操作人员可以根据任务需求来控制模型的复杂度和准确性, 使得 YOLOv5 的使用更加灵活和方便。

图像在层层卷积的过程中, 激活函数的选择正确与否对于深度学习网络的训练能否成功是至关重要的。YOLOv5 的作者使用了 Leaky ReLU^[40]和 Sigmoid 错误!未找到引用源。激活函数。

在 YOLOv5 中, 中间及隐藏层使用了 Leaky ReLU 激活函数, 通过使用 Leaky ReLU 激活函数, YOLOv5 能够更好地处理非线性关系, 提高特征的表达能力和模型的学习能力。最后的检测层中该模型网络则是使用了 Sigmoid 激活函数。而 YOLO v4 选择的激活函数则是 Mish^[42]激活函数。与 ReLU 和 Sigmoid 相比, Mish 激活函数在运算中占用的内存更高, 这个缺点会使得模型的运行更加的缓慢。

YOLO V5 的作者为我们提供了两个优化函数 Adam^[44]和 SGD, 并都预设了与之匹配的训练超参数。默认为 SGD。

其中，本文的面向任务方法设计需要训练较小的自定义数据集，对于这种类型的数据集，Adam 优化函数在模型训练中是更合适的选择，而 YOLOv4 则只有 SGD 优化函数这一选项，不利于本文任务的展开。

最后，YOLO V5s 高达 140FPS 的对象识别速度可以说是一骑绝尘于任何其他的 YOLO 目标检测算法系列中的算法，这个可以算是实时检测的对象识别速度完全能够符合本文的面向任务目标抓取检测任务中选取目标检测算数最看重的实时性这一特点。

3.3 YOLOv5 介绍

3.3.1 输入端

YOLOv5 的输入处理主要由三个部分组成，包括 Mosaic 数据增强、自适应计算锚框和自适应缩放图像，由于 YOLOv5 本身对图像的预处理和数据增强已经达到了本文所需要的程度，所以本章中没有对 YOLOv5 的输入端进行改进，所以本小节是对 YOLOv5 输入端主要功能的介绍。

(1) Mosaic 数据增强

Mosaic 数据增强技术是目标检测算法领域中一种较为强大的数据增强方法，它通过将拼接原始数据集图像，并同时调整这些图像中的标签信息以匹配后续训练，这种特征拼接的方法有助于提升模型的性能。这种方法能够使模型在后续的训练过程中获得更多的场景信息，从而提高对多样性目标的检测能力。

具体来说，Mosaic 数据增强首先会在输入模型的数据集中随机选择四张不同的图像，然后将它们按照设计好的比例拼接成一张新的大尺寸图像。这种拼接的方式能够模拟真实场景中的多个目标物体同时存在的情况。同时，为了保证物体位置信息的准确性，相应的标签信息也会在此进程中进行相应的变换。

在进行拼接时，该模块会通过对图像和标签进行平移、缩放和裁剪等操作，使得拼接后的大图像与单独的原始图像具有相似的尺度和位置分布。这样做的目的是让模型能够学习到各种不同的目标物体排列和相互作用的情况，提高对复杂场景的理解和检测能力。

在目标检测任务中，采用 Mosaic 数据增强技术能够提供更丰富和复杂的图像特征信息。通过采用 Mosaic 数据增强，模型能够观察到多个目标物体之间的

关系。这种关系包括目标物体的相互影响、位置关系、尺度变化等。通过学习这些关系，模型可以更好地理解目标物体之间的上下文信息，提高对目标的检测和识别能力。它对于提升模型对多样性目标的检测能力和泛化能力非常有帮助。由于 Mosaic 数据增强可以合成包含多个目标物体的图像，模型在训练过程中能够接触到更多不同类别、形状和尺度的目标。这有助于模型学习更加全面的特征表示，并提升对不同目标的检测准确性和泛化能力。

(2) 自适应锚框计算

自适应锚框计算技术是 YOLOv5s 中的一项创新，旨在提高目标检测模型对各种尺度和长宽比目标的识别能力。传统的目标检测算法通常使用手工设定的锚框，但这种设定可能导致模型在处理不同尺度和长宽比目标时的性能下降。

自适应锚框计算技术的引入使得模型在训练过程中也能够动态学习和调整锚框的参数，以适应不同目标的尺度和长宽比。相比于固定的手工设定锚框，这种自适应技术能够更好地适应目标的多样性。

上述的自适应锚框的实现方法其实就是在该技术中引入了可学习模块，并在损失函数中增加了与锚框参数相关的惩罚项。这样做的目的是促使模型学习合适的锚框参数，使其能够更好地捕获不同目标的特征。

通过动态调整锚框的位置、尺度和长宽比，模型能够更好地适应不同目标的特征。自适应锚框计算技术提供了一种灵活的方式，使得目标检测模型能够更准确地定位和识别各种尺度和长宽比的目标。在训练过程中，模型通过学习适应性锚框参数，能够捕捉到目标物体的多样性和变化性，从而提高模型的泛化能力。

自适应锚框计算技术的引入使得目标检测模型在面对各种尺度和长宽比目标时更加灵活和准确。相比传统的手工设定锚框的方法，自适应锚框计算技术能够更好地适应不同数据集和场景的需求，提高模型的性能。这项创新为目标检测算法的发展带来了新的可能性，为更好地应对真实世界中多样性目标的检测任务提供了有效的解决方案。

(3) 自适应缩放图片

YOLOv5 引入了自适应缩放技术和优化的输入图像数据集预处理方式，这是该目标检测算法在预处理部分的关键创新之一。传统的目标检测算法在调整

图片数据库的行为中，通常将不同尺寸的图片调整为统一的标准尺寸，但这种做法造成的结果大多是模型计算和推理速度的下降。

相比传统方法，YOLOv5 采用了最小化黑边的策略，根据原始图像的长宽比进行处理，动态调整网络结构以适应不同的分辨率。这一创新使得 YOLOv5 能够更好地处理各种尺寸的物体，提高检测准确性和鲁棒性。

传统方法中，将图像调整为统一尺寸可能导致信息的丢失或形变，从而影响目标检测的准确性。而 YOLOv5 通过自适应缩放技术，更细致地适应输入图片的大小和长宽比，使得模型能够更好地捕捉目标物体的特征。这种方法不仅提高了检测性能，还能够减少计算资源的浪费。

在实际应用中，YOLOv5 的在输入端的灵活性使得它能够适应不同尺寸的输入图片数据库，并保持高效的推理速度。这意味着在处理实际场景中的目标检测任务时，YOLOv5 能够更好地应对不同尺寸物体的检测需求，同时也能保持高性能和高效率。

3.3.2 主干

由图 3-6，即 YOLOv5 的流程图可以看出，CSPdarknet 版本的 backbone 主干部分主要由 Conv 模块、C3 模块和 SPP 结构构成。

Conv 模块的结构示意图如下图 3-2 所示，这个模块是卷积神经网络中的一个最基础的模块，通常有着多种输入以及输出。当特征图片经过输入端的处理后进入 Conv 模块，该特征图片会先经过代表了卷积操作的 Conv2d 卷积层，该卷积层是一个特征提取器，它由多个卷积核组成，其卷积核的个数由输入特征信息矩阵的通道数决定。每个卷积核的尺寸大小、计算步长、填充大小等参数都会决定该卷积核所在的卷积层最终的输出大小及其特征提取能力和感受野大小。

接下来特征图片会经过 BN 归一化层，该层主要用于归一化规范神经网络中的特征值分布。该层的输入通常为一个 batch 中的特征图像，接着归一卷积层会对输入图像的每个通道的图像特征进行均值方差计算，从而实现归一化。最后这些特征会通过仿射变换再进行还原输出，接着通过 SILU' 激活函数实现了对输入特征的转化和输出。

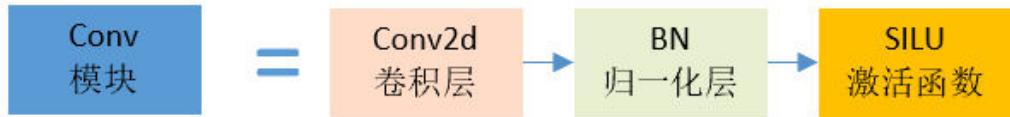


图 3-2 Conv 模块示意图

C3 模块的主要结构特征也如下图 3-3 所示，C3 模块的跨阶段融合策略使得模型可以同时利用两层特征，从而更好的提取细节信息。

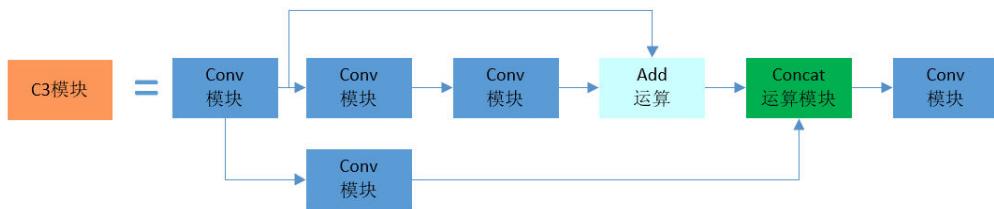


图 3-3 C3 模块示意图

SPP 模块即金字塔池化模块，它利用了最大池化层操作来整合多种感受野输出的特征信息。SPP 的结构网络如下如图 3-4 所示。

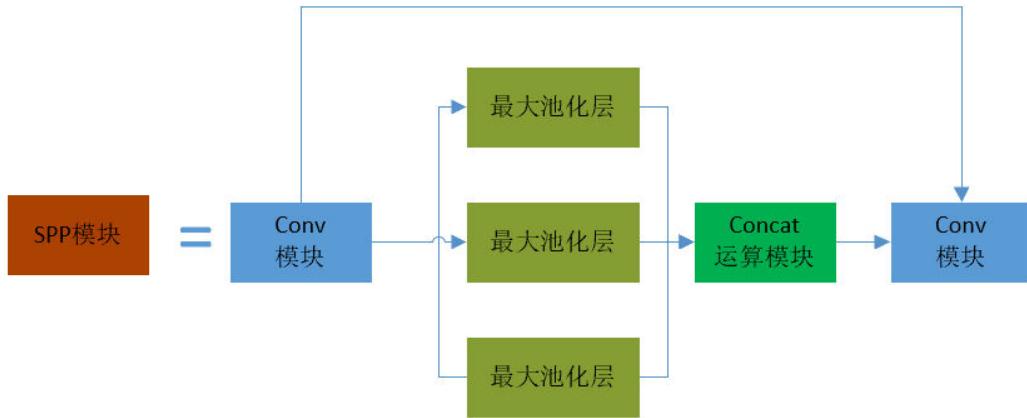


图 3-4 SPP 模块示意图

其中，YOLOv5 的 CSPdarknet 主干网络虽然效果非常强大，但是还具有以下几个缺点：

(1) CSPdarknet 对于一些轻量级的网络来说，参数量还是比较大的。这会导致模型的存储需求较高，在推理计算时需要更多的计算资源。

(2) CSPdarknet 对于小目标检测问题的适用性不高。YOLOv5 原主干网络对小尺寸目标的感知能力相对弱，我们认为这一点会对于面对任务的抓取有着一定的缺失。

(3) 对于密集目标的处理能力较差。由于网络的感受野和特征提取能力限制，该主干网络对密集目标的定位和识别分割的能力不强。

根据上述缺点，本章决定对 YOLOv5 的主干网络进行改动，将其改为 Swin-transformer 网络，该网络较为轻量且对小目标的检测非常强大，从而对于密集目标的处理能力较好。

3.3.3 Neck

Neck 网络是深度学习结构中的一个关键组件，位于骨干网络和输出端之间。其主要作用是对特征进行融合和进一步处理，以提升目标检测模型的性能。Neck 网络在整个深度学习结构中扮演着连接骨干网络和检测头部分的重要角色。它接收来自骨干网络的多层特征，并通过一系列操作对这些特征进行融合和处理，以产生更具表征能力的特征表示。通过 Neck 网络，不同尺度和层次的特征能够有效地整合和利用。这种多层次的特征融合能够捕捉目标物体在不同尺度上的信息，使得模型能够更好地适应不同大小和复杂度的目标物体。

在 YOLOv5 中，Neck 部分在整体中的模型可由下图 3-6 看出，该模型的 Neck 模块将 PAN 与 FPN 相结合的方式也如图 3-5 所示，其中每经过一次 Conv 就相当于进行了一次下采样。它使用的 PAN 网络和 FPN 网络是两种常用的多尺度特征融合方法，它们的融合方式在 YOLOv5 目标检测框架中起到重要作用。

PAN 网络通过多尺度特征融合，能够提升模型对小目标的检测能力。传统的目标检测算法在处理小目标时容易出现信息丢失或分辨率不足的问题，而 PAN 网络通过将不同尺度的特征进行融合，能够更好地捕捉小目标的细节特征，提高检测准确性。在 3-5 中，PAN 网络实质上就是在 FPN 网络中再添加一个自下而上的金字塔模型从而来将底层的定位特征信息递到上层。

FPN 网络引入了不同尺度的特征金字塔结构，使得模型能够更好地适应不同尺寸的目标物体。FPN 网络通过自顶向下和自底向上的特征传播，将来自不同层级的特征进行融合，从而在不同尺度上建立起特征金字塔，提供多尺度的语义信息，增强了模型对不同尺寸目标的感知能力。

拥有 PAN+FPN 结构的 YOLOv5 网络既掌握了输入图片的语义信息又掌握了图片的定位信息，这增强了模型对不同尺寸目标的位置和语义感知能力。

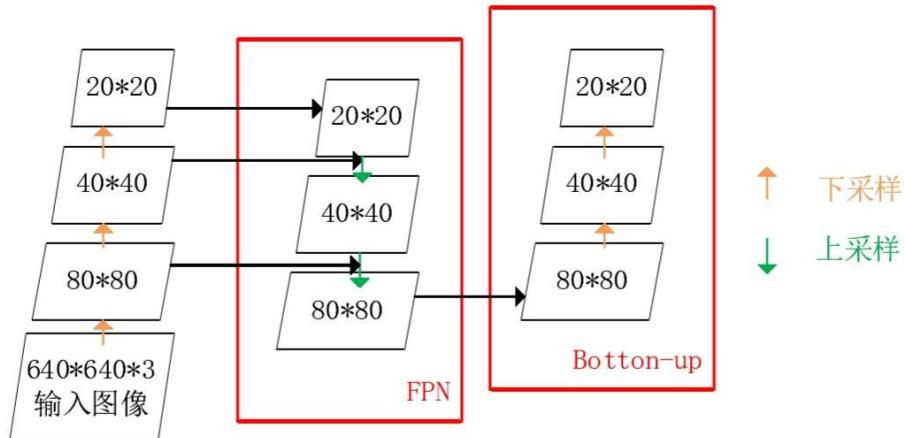


图 3-5YOLOv5 neck 中 FPN+PAN 结构示意图

3.3.4 输出端

YOLOv5 的输出端任务是对经过 Backbone 和 Neck 处理的特征进行分类和预测，以实现目标检测的功能。

输出端采用卷积层和激活函数来完成分类和预测任务。其中，卷积层用于提取特征，激活函数则引入非线性变换，使得模型能够学习更复杂的特征表示。

YOLOv5 使用 GIOU (Generalized Intersection over Union) 损失函数来衡量目标的位置精度。相比传统的 IoU (Intersection over Union) 损失函数，GIOU 损失函数考虑了目标框的尺寸、位置和形状信息，能够更准确地衡量目标检测结果与真实标签之间的差异。

输出端生成不同尺寸的特征图，以适应不同尺寸的目标物体。这种多尺度的特征图可用于检测不同大小的目标，并提供更全局和细节的信息，增强了模型的感知能力和检测准确性。

为了提高输出的精确性和干净度，YOLOv5 使用非极大值抑制 (NMS) 算法来消除高度重叠的多余预测框。NMS 算法会通过筛选出具有最高置信度的框，并抑制与其高度重叠的其他框，从而得到最终的输出结果。

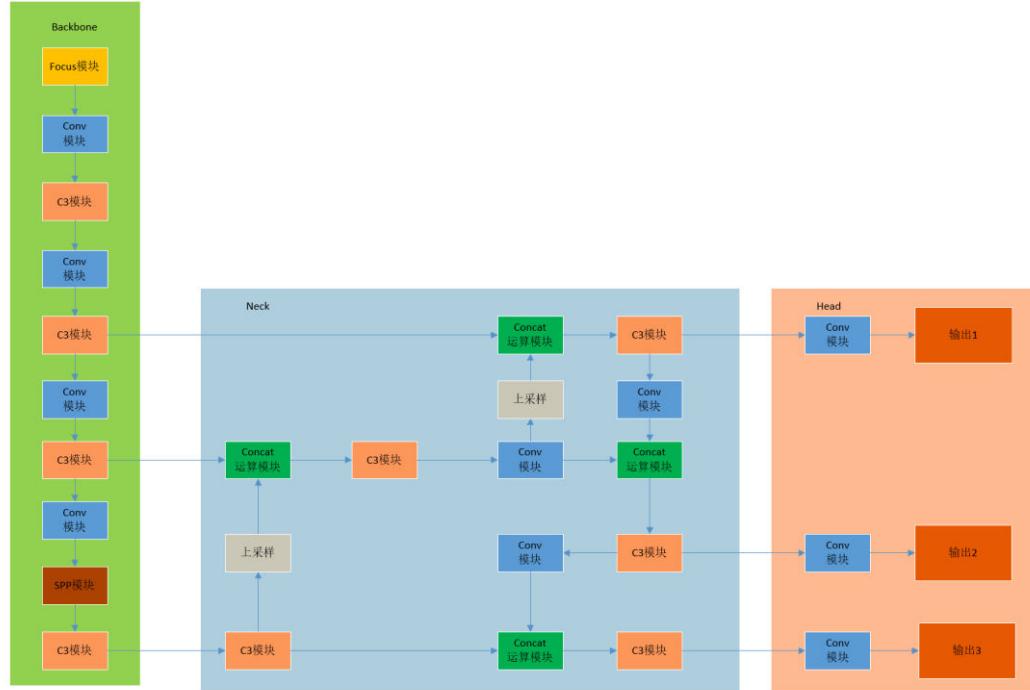


图 3-6 YOLOv5 原框架整体架构

3.4 改进的 YOLOv5 目标检测算法

3.4.1 多注意力模块集成分层特征映射主干

本小结对 YOLOv5 的骨干网络进行了全面的更改，将 YOLOv5 原先使用的 CSPdarknet 网络更改成了性能更强的 swin_transformer^[45]网络架构，以期在本文所用到的更小的自制数据集下达到更好的 mAP 值和更高的 FPS 值。

Swin-transformer 所构建的特征图有层次性，随着下采样次数的提高，该模型会通过合并更深层次的图像（下图灰色示意框）逐渐减小模型的宽和高来建立分层特征映射，并且由于仅在每个局部窗口（下图红色示意框）内计算自关注，可以大大降低运算量节省运算空间，并且还具有输入图像大小的线性计算复杂性，三个下采样概念图如下图 3-1 所示。

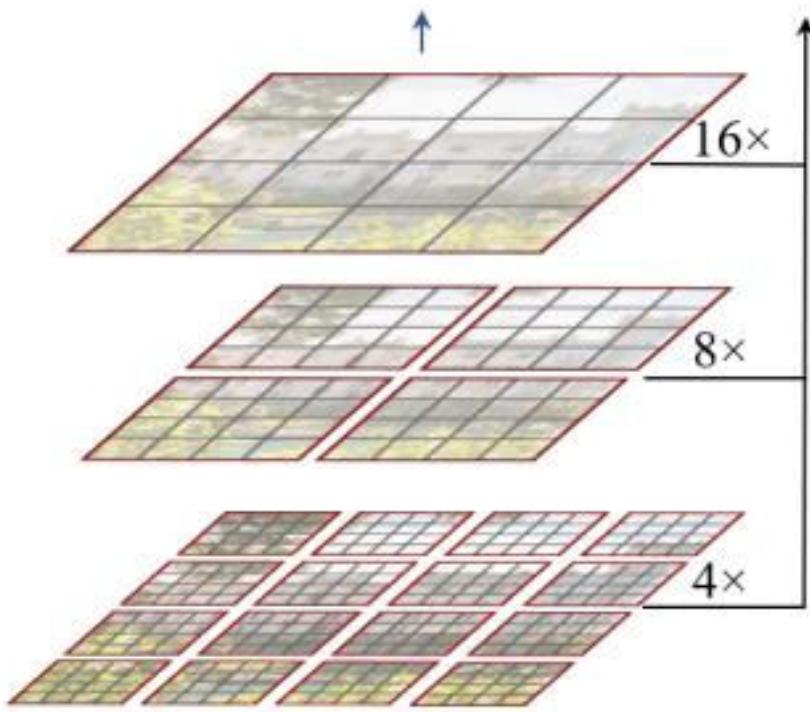


图 3-7 swin-transformer 网络结构图概念图

Swin-transformer 的整体网络具体实际架构如下图 3-2 所示，该模型的具体流程可以清晰地从下图中看出，首先对深度学习网络输入一个宽度为 W ，高度为 H 的三通道 RGB 图像，该图像会进入通道扩张模块，该模块的作用是将图像通过线性嵌入层的卷积操作，按照 4×4 的大小进行分割，分割过后每个小部分都是 $\frac{H}{4} \times \frac{W}{4}$ 的三通道块，接着对这些小块进行深度方向上的拼接叠加，最终图像在经过通道扩张模块后会成为 $\frac{H}{4} \times \frac{W}{4} \times 48$ 的矩阵。然后该模型对扩张后的矩阵进行线性扩张以满足深度网络训练需求，所以上面的矩阵在经过线性通道扩张操作后会变为 $\frac{H}{4} \times \frac{W}{4} \times C$ 的图像矩阵，其中的 C 的数值变化原因是不同类型的任务需求不同，Swin-transformer 模型的所需的 channel 也会随之变化，channel 数的变化最终就会提现到 C 值的变化上。接着，被编码过的包含图像特征的一个个小的局部图像矩形就会进入到 Swin-transformer 模块里，该模块的大致结构如下图 3-3 所示，其中模型内的 LN 为线性常态化模块，W-MSA 为窗口多头注意力机制模块，SW-MSA 为可变窗口多头注意力机制模块，MLP 为全连接层，具体结构如下图 3-4 所示，且其中涉及到的 GELU^[49]激活函数被表示在下式 3-1 中。Swin-transformer 模块还可以进行多次堆叠，多次堆叠可以提高非线性及模

型的特征表征能力，本小节的实现中对该模块进行了两次堆叠。

$$GELU = x \times \phi(x), x \sim N(0,1) \quad (3-1)$$

而经过了第一阶段的卷积操作过后图像矩阵仍会维持 $\frac{H}{4} \times \frac{W}{4} \times C$ 的大小并进入下一个阶段，接下来的几个阶段的格式相同，都是先经过一个下采样模块再经过一个 swin transformer 模块，而每通过一个模块，图像矩阵都会增加一倍的通道数并且缩小一倍的宽度和高度参数，这种下采样方式会导致 swin transformer 这种网络架构形成分层式的特征图，能够更有效地处理不同分辨率的图片的特征，提高了模型的在各种不同图片上的适用性。

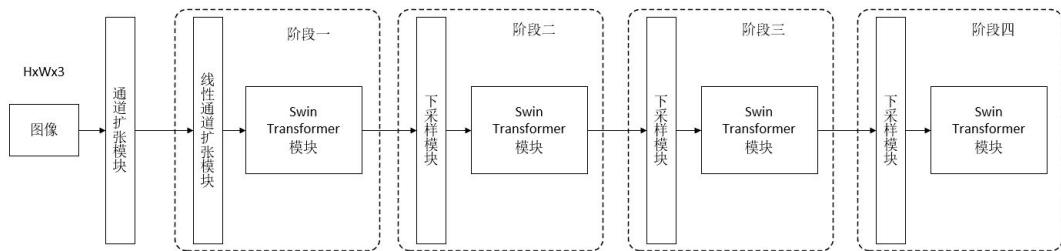


图 3-8swin-transformer 网络结构实际图

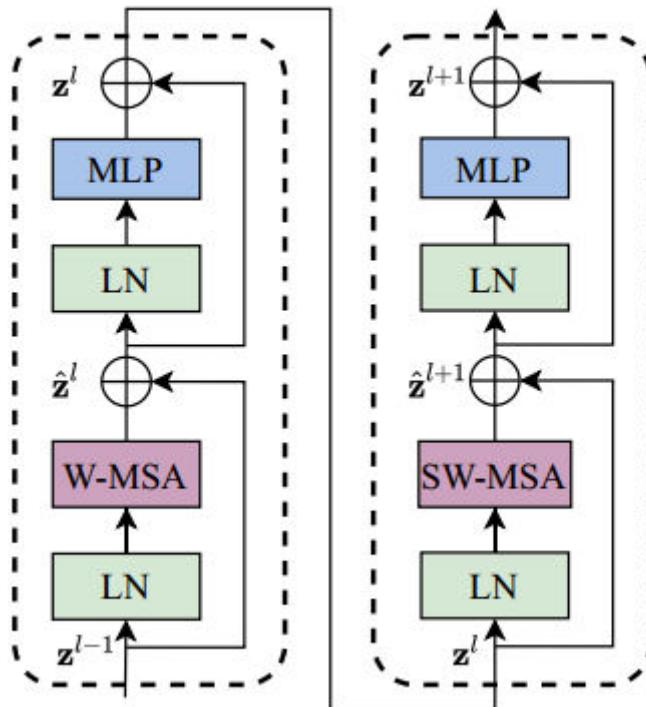


图 3-9 Swin transformer 模块示意图

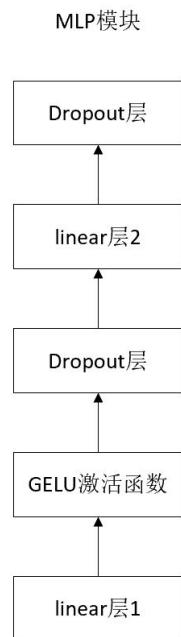


图 3-10 MLP 结构

Swin transformer 主干网络的应用相对于 CSPdarknet 而言，优点主要突出于以下几个方面。首先，Swin transformer 的层次化特征映射使得其比 CSPdarknet 更泛化，有着更好的适用性；其次，该模型的 transformer 特性引入了局部自注意力计算，这种跨窗口信息流动方式使得计算更加高效，也增加了感受野；再者，swin transformer 在本小节的实现方式中进行了多次堆叠，拥有了比 CSPdarknet 更强大的特征表达能力。并且在目标检测上，使用 Swin transformer 能够达到更高的 AP 值，更符合本文需要。

在本小节中，我们在与 cspdarknet 即原骨干网络同根目录中，以 pytorch 为框架将 swintransformer 的整体网络结构进行了实现，并在 YOLOv5 目标检测框架上进行了调用并最终成功通过 Cornell 模型训练了 swin transformer 网络架构中的参数，验证了该架构替换 YOLOv5 主干网络的实用性。

3.4.2 基于宽高差异值的快速收敛损失函数

YOLOv5 在官方版本中使用的损失函数是 GIOU，GIOU^[47]比 IOU^[46]算法优秀的点在于它在其基础上引入了最小外接框，这个举措解决了检测框和真实

框 (A、B) 没有重叠时损失率 loss 等于 0 的问题。但是当检测框和真实框出现包含现象的时候 GIOU 会退化成 IOU。其中，IOU 算法时目标检测中最常见最基础的损失函数，但是有检测框真实框不相交不能反映框间距离等一系列问题。所以本小节决定引入损失函数 EIOU 来解决上述出现的所有问题。

如下式 (3-1)、(3-2) 为 IOU、GIOU 的表达式，

$$IOU = \frac{A \cap B}{A \cup B} \quad (3-1)$$

$$GIOU = IOU - \frac{\text{最小封闭框面积} - A \cup B}{\text{最小封闭框面积}} \quad (3-2)$$

其中，最小封闭框面积就是 A、B 并集的最小闭包面积。

从式子 (3-2) 可以看出，GIOU 具有当检测框和真实框出现包含现象的时候 GIOU 会退化成 IOU 的缺点。且在水平垂直这两个方向上没有给出惩罚机制，进而导致这两个方向的收敛速度较慢。

本小节在考虑到上述 GIOU 的缺点后，又进一步考虑到收敛较慢问题可以通过直接回归 A、B 两框的中心点间欧氏几何距离促进损失函数加速收敛来解决，并且计算两框的宽高的差异值来保证收敛的精确度。

如下式(3-3)为 EIOU 的表达式，

$$EIOU = IOU - \left[\frac{\rho^2(b, b^{gt})}{(w_c)^2 + (h_c)^2} + \frac{\rho^2(w, w^{gt})}{(w_c)^2} + \frac{\rho^2(h, h^{gt})}{(h_c)^2} \right] \quad (3-3)$$

其中， $\rho(b, b^{gt})$ 为两框中心点间的距离， $\rho(w, w^{gt})$ 为两框宽度间的距离， $\rho(h, h^{gt})$ 为两框高度之间的距离。 w_c 和 h_c 分别是最小闭包区域的宽度和高度。

如下图 (3-1) 为具体的函数实现过程。

```
#-----#
# 计算两个框中心点的距离的平方
#-----#
rho2      = ((b2_xmin + b2_xmax - b1_xmin - b1_xmax)**2+(b2_ymin + b2_ymax - b1_ymin - b1_ymax))/4
#-----#
# 计算预测的宽高分别与最小外接框宽高的差值 优点加速了收敛提高了回归精度
#-----#
rho_w2    = ((b2_xmax - b2_xmin) - (b1_xmax - b1_xmin)) ** 2
rho_h2    = ((b2_ymax - b2_ymin) - (b1_ymax - b1_ymin)) ** 2
cw2       = cw ** 2
ch2       = ch ** 2
eiou      = iou - (rho2 / cw2 + rho_w2 / cw2 + rho_h2 / ch2)
# giou     = iou - (enclose_area - union_area) / enclose_area
```

图 3-11 损失函数 EIOU 实现过程

3.4.3 自适应特征优化的端到端注意力模块

注意力机制^[48]源于对人类视觉的研究，在日常生活中，人类能够通过选择性地关注信息的一部分并忽略其他可见信息来加速对信息的比较重要的特征进行很大程度上的吸收消化。这一概念被引入到卷积神经网络中，以提高网络对重要信息的关注和特征的表达能力。

本小节对 YOLOv5 网络架构的改进中所使用的注意力模块为 CBAM (Convolutional Block Attention Module) 注意力机制，该注意力机制是一种模块化的注意力模型，可用于卷积神经网络，并能进行端到端的训练。CBAM 模块包括通道注意力和空间注意力模块，本小节在改进的算法中使用该模块的主要目的是提高对输入图像特征的表达和关注于输入图像的重要特征信息。

通道注意力模块通过全局池化操作计算通道注意力，并使用多层感知机进行处理，生成包含通道注意力机制的特征图。这个模块的作用是对不同通道的特征进行加权，以使网络更加关注重要的通道信息。且该注意力机制能够根据特征图中每个通道的重要性，为每个通道分配不同的权重。这样，网络可以更有针对性地选择和利用特征图中对任务更有贡献的通道，提高模型的性能和泛化能力。

空间注意力模块将通道注意力机制的特征图与输入特征图进行元素级乘法操作，生成经过空间注意力调整后的输入特征。这个模块的作用是根据通道注意力的权重，对输入特征图的不同位置进行加权，以突出重要的空间信息。该模块能够在特征图中根据通道注意力的权重，对不同空间位置的特征进行加权。通过这种方式，网络可以更加关注重要的空间信息，提高对目标的定位和识别能力。

且 CBAM 模块是一种轻量级的注意力机制，适用于任意区域，并可以提供对重要内容的关注和特征优化。它的引入可以改善 YOLOv5 卷积神经网络的表达能力和注意力机制，从而提高模型在视觉任务中的性能。

如下式 3-4，通道注意力机制可以表示为

$$M_c(F) = \text{sigmoid} \left(\text{MLP}(\text{AvgPool}(F)) + (\text{MLPMaxPool}(F)) \right) \quad (3-4)$$

其中，F 为输入特征图像。

空间注意力模块以通道注意力模块的输出特征图作为输入特征图。这两个

模块一起构成 CBAM (Convolutional Block Attention Module) 注意力机制的核心组成部分。其运算流程图如图 (3-2) 的小图 (2)

在空间注意力模块中，首先进行基于通道的全局最大池化和全局平均池化操作，对通道维度进行压缩。这两个操作分别提取出输入特征图中每个通道的最大值和平均值。

接着，该模块会将全局最大池化和全局平均池化的结果进行 concat 操作，并将这两个特征图在通道维度上进行拼接。这样得到的特征图具有两倍的通道数。

经过卷积降维操作和 sigmoid 函数的处理后，该模块会将拼接后的特征图转换为空间注意力特征图。其中，卷积降维操作可以减少参数量，而 sigmoid 函数可以将特征图的每个像素值映射到[0, 1]的范围进行标准化操作。

最后，空间注意力特征图与输入特征图进行元素级乘法操作，将每个位置的像素值与对应位置的空间注意力权重相乘，生成最终的特征图。经过上述对图像的处理后，空间注意力机制能够对输入特征图的不同位置进行加权，从而突出重要的空间信息。

通过两次加权处理的结果就是最终生成 CBAM 注意力特征图像，如图 (3-3)。

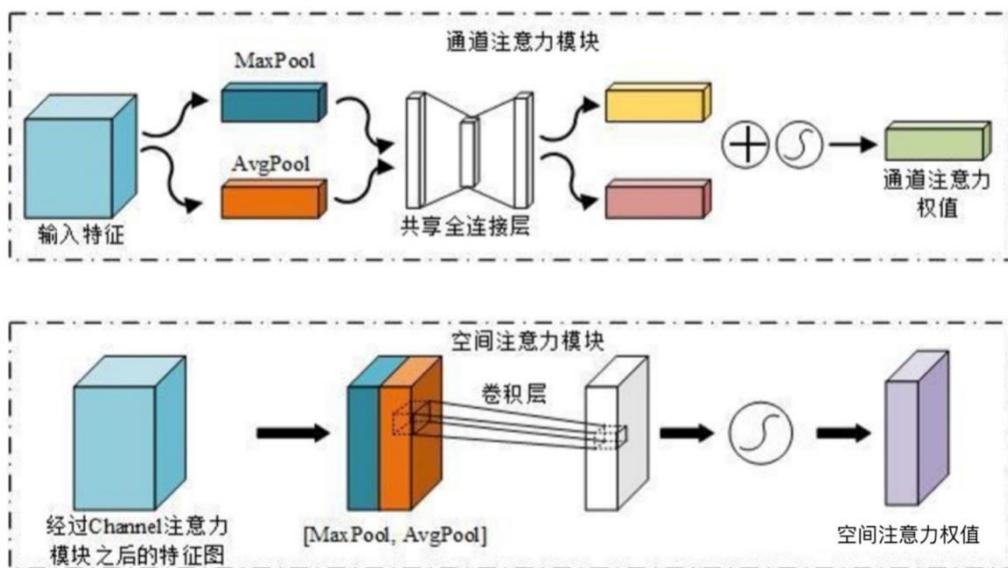


图 3-12 (1) 通道注意力模块 (2) 空间注意力模块

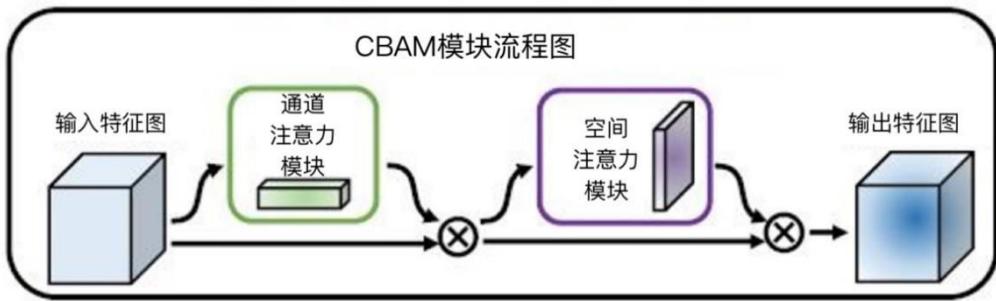


图 3-13 CBAM 注意力机制示意图

3.5 仿真实验验证

3.5.1 常用数据集

常用数据集中，COCO 和 VOC 是在目标检测领域中比较容易提及到的训练数据集。

COCO 数据集是一个大型、丰富的物体检测、分割和字幕数据集，旨在实现场景理解的目标。该数据集包含 91 类目标，拥有 328,000 张影像和 2,500,000 个标签，是目前为止在语义分割领域中最大的数据集之一。COCO 数据集的广泛覆盖和多样性使其成为评估物体检测和分割算法性能的重要基准。

PASCAL VOC 是一个官方数据集，用于测试改良后的 YOLOv5 算法，并验证算法在同一数据集上的准确率和速度改进。VOC 数据集包含 4 个大目标集，包括 20 个目标类型，拥有至少 10 万张图片。该数据集在数据分类和目标检测领域得到广泛应用，是评估算法性能的常用基准之一。

本节使用了 VOC 数据集进行训练和预测对比是评估改良后的 YOLOv5 算法性能。通过在 VOC 数据集上进行训练和测试，可以评估算法在不同目标类型、尺寸和场景下的检测准确性和速度表现。本节使用 VOC 数据集对于改进前的模型和改进以后的模型进行了各类别的 AP 值的对比来体现改进的优点。

3.5.2 实验抓取数据集制作

本小节介绍了实验所需要抓取的物体制作的自用数据集。首先我们使用 1:1 的图像对需要抓取的物体进行各自的拍照和合并拍照，合并拍照的目的主要

是为了训练 YOLOv5 代码对多物体的识别能力。接着我们使用 labelimg 这个 python 工具来对拍下来的物体进行识别框的手动标定，如下图 3-13 为具体图片标定的过程，并且对图片的标定完成后将标定好的框转换成 xml 格式的文件以便于后续的算法的调用。然后将图片的格式按照 VOC 数据集的格式进行整理。最后将模型中的参数进行一系列的更改过后对模型进行了训练。



图 3-13 具体图片标定过程

3.5.3 改良前后算法在官方数据集上的预测结果对比及结果展示

改良前的 YOLOv5 在 VOC 数据集上的 AP 结果和预测速度值已经是一个过关的目标检测函数了，mAP 数值达到了 80 而 FPS 值达到了 51。

在经过了同样的改良后的 YOLOv5 在 VOC 数据集上的的 AP 结果同样也是非常惊人，可以从表 3-1 看出，在物体识别上本章节改进的 YOLOv5 目标检测算法大部分都超出了 YOLOv5 的原框架的识别结果。

该改良算法的平均正确率在大数据集中已经屈指可数，完全可以使用到实用领域范围，而 FPS 也达到 53，53 的 FPS 值意味着每秒可以处理 53 帧的数据，这也代表着该改良算法完全具有实时性，可以胜任本文的位姿估计算法的前置处理函数。

本小节还利用改良后的网络结构对自制数据集进行了一系列训练，训练后对多张输入特征图片进行了识别处理测试，其中一个结果如下图 3-16。从该图也可以看出，该改良深度网络框架在自己制作的数据集的目标识别任务执行上也有着良好的发挥。

表 3-1 改良前后算法在 VOC 数据集上的结果对比

类别	改良前(%)	改良后(%)
aeroplane	91.97	89.49
bicycle	89.49	90.72
bird	82.12	85.23
boat	74.66	77.10
bottle	78.23	85.11
bus	87.56	85.09
car	91.73	89.56
cat	87.01	88.18
chair	66.97	70.53
cow	87.63	88.38
diningtable	74.49	76.35
dog	83.99	84.52
horse	90.72	89.87
motorbike	89.12	90.45
person	88.81	89.15
pottedplant	56.60	60.27
sheep	88.32	89.30
sofa	72.41	76.02
train	86.86	86.93
tvmonitor	81.11	83.94

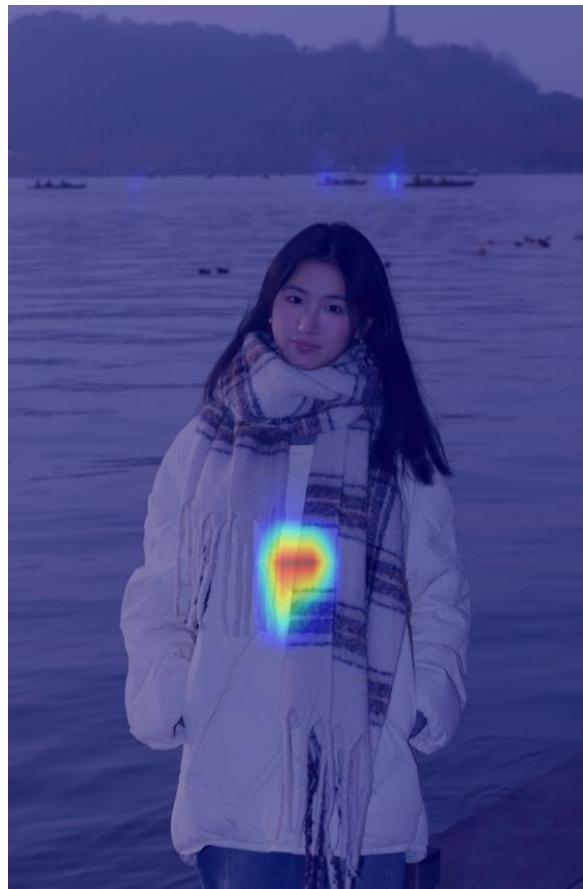


图 3-14 数据集图像预测输出热度图

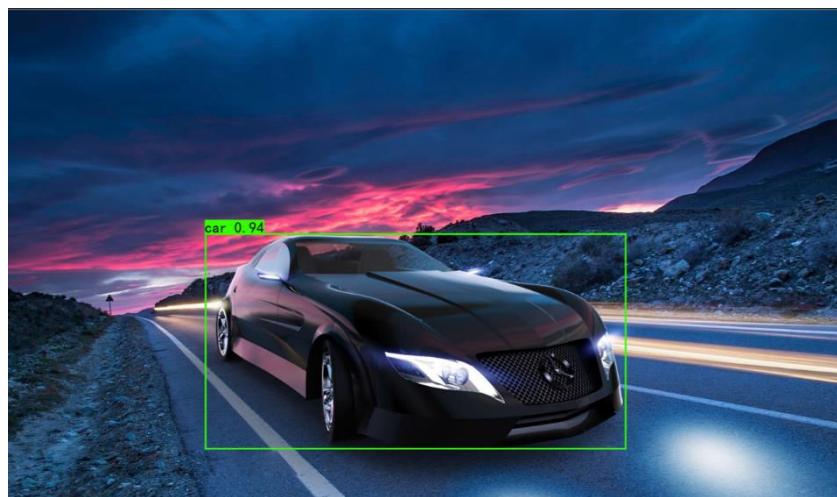


图 3-15 官方数据集图片预测框图

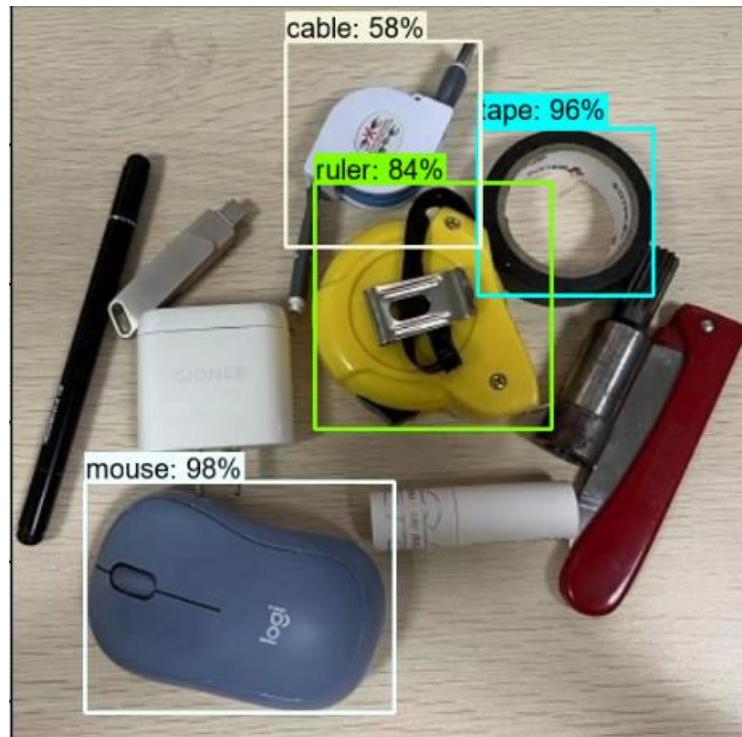


图 3-16 自制数据集图像预测框图

3.6 本章小结

本章首先对本论文的目标检测算法和抓取位姿识别网络融合的网络架构中目标检测算法选择 YOLOv5 的原因进行了剖析，具体的方法为引出了 YOLO 算法的各代版本并进行对比最终交代了 YOLOv5 适用于面对任务的抓取识别方法研究的几个主要原因。之后解释了没有进行改进前的 YOLOv5 算法的大致框架并引出了之后对 YOLOv5 的改进。然后对 YOLOv5 的改进部分进行了详细的解释，本章主要引入了多注意力模块的分层映射主干、基于宽高差异值的快速收敛损失函数以及自适应特征优化的端到端注意力模块这三个优化点。随后进行了仿真实验验证，其目的是验证改进后的 YOLOv5 算法的实用性，以及对比改进前后的代码在数据集上的表现，其中还插入了 YOLOv5 训练和评估常用的数据集 COCO 和 VOC 数据集的介绍。最终可以得出改进后的 YOLOv5 算法有着实用稳定等优点，证明了 YOLOv5 算法改进的成功。

4 基于改进 YOLOv5 与 GR-ConvNet 的机器人面向任务的抓取检测框架

4.1 引言

通过上一章详细介绍了如何对目标检测算法进行改进以及对改进后的目标检测算法进行在传统数据集上进行能力测试，该章节验证了改进后的目标检测算法的先进性和可行度。在对目标检测算法的改进基础上，进一步地，我们需要找到适合的抓取位姿识别代码与之前改进过后的目标检测算法一起进行算法融合操作，将两个算法之间的关系找出并将其紧密联系。

算法融合操作具有许多优点如下：

提高了目标识别和抓取的精度，同时使 GR-ConvNet 在处理小目标问题时更加精确。将目标检测和抓取位姿识别算法结合的做法等同于使得 GR-ConvNet 在运行之前，图像就经过了非常优秀的预处理，大大降低了 GR-ConvNet 在生成残差卷积网络时的复杂度从而提高了模型训练速度，更符合本文中面向任务的模型架构需要。而且 YOLOv5 检测大目标的能力同样出色且具有实时性，这一优点也能最终体现在抓取位姿输出的速度提升上。

保证了在复杂多信息环境下的抓取位姿识别成功率。YOLOv5 负责目标检测的主要优点在于该深度网络可以正确快速的将相机所拍摄的图像中各物体进行分类，并正确实时地识别需要抓取的物体，大大减轻了 GR-ConvNet 在识别物体上的负担。由于将两个算法融合等于将两个算法的不同的特征提取能力和不同的感受野进行结合，所以本小节的举措既能够让抓取位姿输出变得更加全面和稳定，更能够提升输出的质量和速度。该举措还进一步增加了整体模型的适用性，当出现例如一个场景有多个不同性质的抓取对象时，融合了改进的 YOLO 的抓取框架网络能更好的处理抓取目标的多样性。

使得整体模型鲁棒性提升。融合了改进版的 YOLO 深度网络的 GR-ConvNet 框架可以提高整体系统的鲁棒性，使得该模型能够应对多样化的环境及抓取目标的挑战，这一点也与本文所追求的面向具体任务的抓取位姿模型非常契合。例如当出现 GR-ConvNet 不好处理的小目标问题并且网络性能发生了性能下降时，YOLO 改进深度网络能够提供更利于其处理的结果，从而减轻系

统的单点故障，这保证了系统的整体稳定性。

本章先对选定的抓取位姿识别算法进行了网络结果的详细介绍和功能效果展示，接下来就进一步的提出了算法融合的方法和步骤，并且实现了两个算法的融合操作，为进一步进行在硬件平台上的实验操作打好理论基础并准备好了实现代码。

4.2 GR-ConvNet

GR-ConvNet (Generative Residual Convolutional Neural Network)^[25]是一个服务于抓取任务的用于图像识别并输出抓取位姿的卷积神经网络，它主要用于执行机器人对未训练过的物体进行抓取的任务，其研发的目的来用来解决输入为场景的 3 通道 `rgb` 图像或者单通道深度图像，输出为目标物体的可抓取位姿估计以及对物体的抓取预测结果的情景，如下图 4-1 所示为该模型在进行抓取位姿估计时的主要流程和输入输出实例示意，其中，该模型输出的位姿估计包括抓取质量即抓取估计成功率、抓取所需宽度以及抓取所需角度，角度的范围为 $(-\frac{\pi}{2}, \frac{\pi}{2})$ 。这个模型同时也可以称为生成残差卷积神经网络模型，主要是因为该网络的特点就是引入了五个残差层来确保抓取位姿输出的精确性。该模型可以实时速度(0~ 20ms)从 n 通道输入生成具有鲁棒性的对跖抓取位姿预测结果。

该模型在标准数据集和不同的混合对象集上评估了所提出的模型架构，其在 Cornell 和 Jacquard 抓取数据集上分别达到了 97.7% 和 94.6% 的准确率。

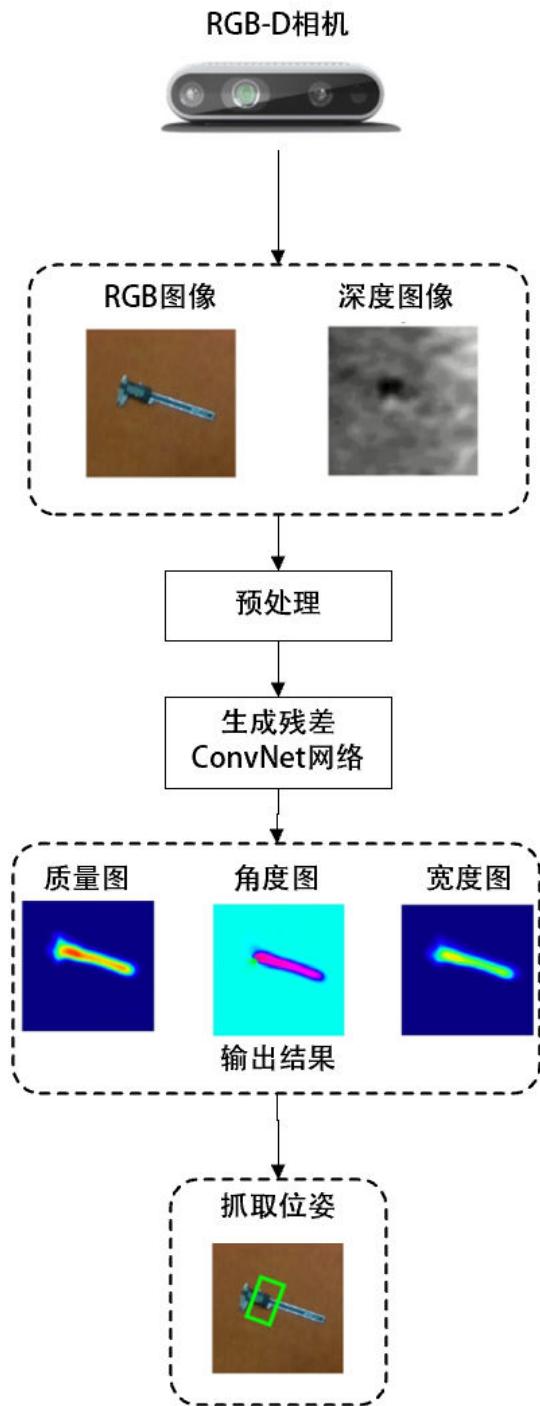


图 4-1 GR-ConvNet 估计抓取位姿时的运行流程

4.2.1 模型架构

图 4-1 展示了提出的 GR-ConvNet 模型，该模型是一种生成架构，采用 3 通道输入 RGB 图像或者单通道深度图像，并以三幅图像的输出形式进行结果输出，

其分别为抓取角度、抓取宽度以及抓取成功率，从而实现逐像素抓取。

如下图 4-2 GR-ConvNet 网络结构所示，图像经过 3 个卷积层，这三个卷积层模块是该深度学习网络中的基础组成部分，一个模块主要由一个 Conv 卷积网络加上一个 Batch 标准化层最后通过一个激活函数组成。然后图像再经过 5 个残差层和卷积转置层，其中卷积转置层与前面的卷积层的区别就在于转置层模块中的 Conv2D 卷积网络是前序网络的转置，最终两个前后的三卷积层对称结构使得图像变回原来的大小并且生成 4 个图像。这些输出图像的内容分别是抓拍质量成功率、抓取所需角度组成 $\cos 2\theta$ 和 $\sin 2\theta$ 以及末端执行器抓取任务时所需的宽度。由于本文使用的反点抓取方法即在物体两边相对的极点进行抓取的方法，在 $\pm \frac{\pi}{2}$ 周围是均匀的，所以本节以两个元素 $\cos 2\theta$ 和 $\sin 2\theta$ 的形式提取角度，这两个元素输出不同的值，这些值组合在一起就可以形成最终所需要的抓取角度。

在该模型中，卷积层用于从输入图像中提取特征，并将其输出传递到残差层。残差层的数量对于抓取检测的输出结果有影响，随着残差层数量的增加，输出特征的数量和质量也会增加，但是一般来说，使用 5 个残差层是最佳选择，因为超出 5 层之后输出特征的质量会因为梯度消失和维度误差的问题而导致下降。

使用残差层的主要目的是使网络能够更好地学习恒等函数。通过引入残差连接，网络可以更轻松地学习输入和输出之间的映射，从而提高网络的性能。此外，卷积转置运算也用于上采样操作，以保留空间特征，从而提高检测的准确性。

该模型的另一个优点是其相对较少的参数数量，这使其适用于计算资源有限且需要快速运算的场景。较少的参数数量意味着模型的存储空间要求较低，并且在推理和训练过程中需要的计算资源也较少。

由于该模型具有轻量化的特性，因此非常适用于需要快速运算的场景，特别是高频率的闭环控制任务。其较快的速度可以满足实时性要求，使得模型可以在实时控制系统中进行高效的运行和响应。

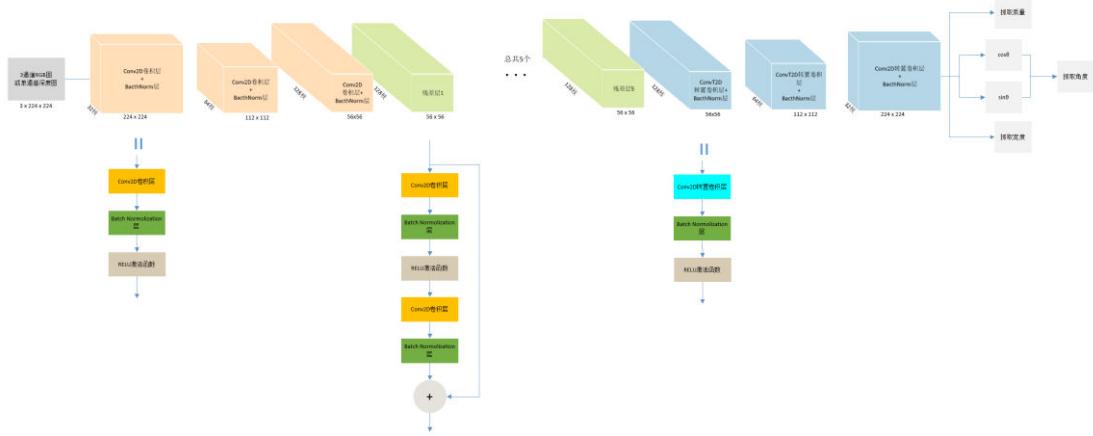


图 4-2 GR-ConvNet 网络结构

4.3 改进的 YOLOv5 算法与 GR-ConvNet 面向任务抓取框架融合

GR-Conv 可以为每个像素生成一个抓取帧，然后选择质量最高的像素作为最终抓取帧。

因此，它能够抓住任何有规则或不规则形状的物体。但在实际环境中，无法根据各种需求选择预定义对象。例如，当一些感兴趣的物体与一组中的其他物体混合在一起，并且它们都在相机的同一视场中时，该算法就会遇到没法解决的错误。

而 YOLOv5 可以在复杂的背景下识别和定位感兴趣的对象，这是 GR-ConvNet 算法所缺乏的。因此，我们提出了一种替代方法，将 GR-ConvNet 与 YOLO 结合起来，形成两步级联结构，称为 YOLO-GRCNN。

YOLO 生成的边界框用于剔除图像中其他无用的区域，并生成 GR-ConvNet 感兴趣的区域。然后将感兴趣区域的深度图像输入到 GR-ConvNet 中，识别出目标最合适抓取位置。YOLO-GRCNN 为流水线结构，如图所示，YOLOv5 负责目标识别和定位，而 GR-ConvNet 负责生成最佳目标抓取姿态。为了实现抓取过程，提出的 YOLO-GRCNN 检测结构主要由以下步骤组成。

首先在抓取之前，对所有需要抓取的目标进行标定。之后使用 YOLOv5 用于检测摄像机提供的 RGB 图像中的物体。接着对由 YOLOv5 分类出来的物体通过 PS 平台进行裁剪。然后通过 rolabelimg 框架对于这些训练过的对象指定多组边界框用于抓取，并对每个边界框的索引进行编码。经过上一步的处理后，深度图像被切割为了只包含感兴趣的对象的图像，并且它与 GR-ConvNet 的输

入图像具有相同的大小。最后使用训练好的 GR-ConvNet，生成质量分数最高的抓取框，并且进一步变换抓取框架，得到机器人坐标系下的抓取方案。

需要注意的是，在最后一个步骤中生成的抓取盒坐标平面为二维像素坐标平面，需要转换为相机坐标系。由于相机是“眼在手上”操作，因此需要根据机械臂的当前位姿计算相机坐标系与机器人本体坐标系之间的变换。最后，推导出机器人基坐标系下物体的最优抓取姿态。

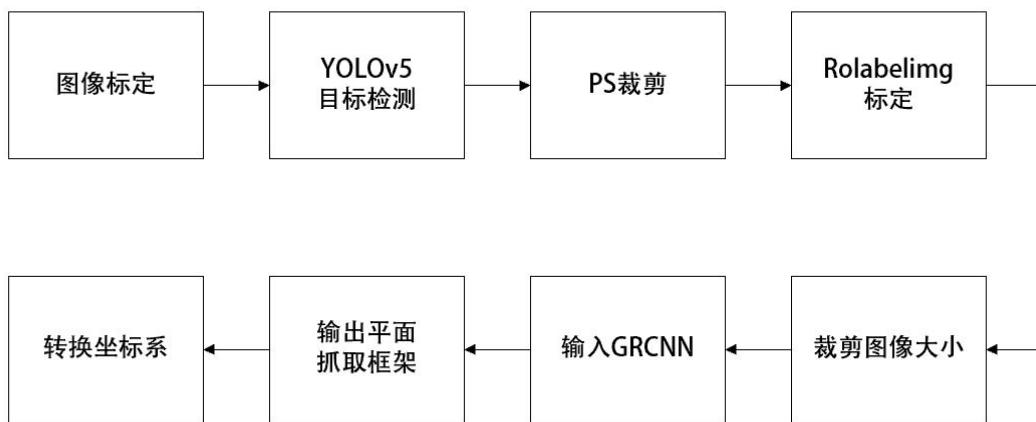


图 4-3 YOLO-GRConvNet 框架

4.4 仿真实验验证

4.4.1 抓取识别算法常用数据集

(1) Cornell 数据集

Cornell 数据集 (Cornell Grasping Dataset) 是一个常用的用于机器人抓取任务的数据集，旨在推动机器人视觉和控制的研究。该数据集由康奈尔大学的机器人实验室 (Cornell Robotic Perception Lab) 创建。

Cornell 数据集包含了一个真实场景中的抓取任务数据库，其中包括了约 1000 个物体实例的 RGB-D 图像、深度图像和抓取姿态的标注信息。这些物体实例包括各种形状和尺寸的物体，如杯子、餐盘、玩具等。数据集提供了多个视角和不同的光照条件下的图像，以增加数据的多样性和挑战性。

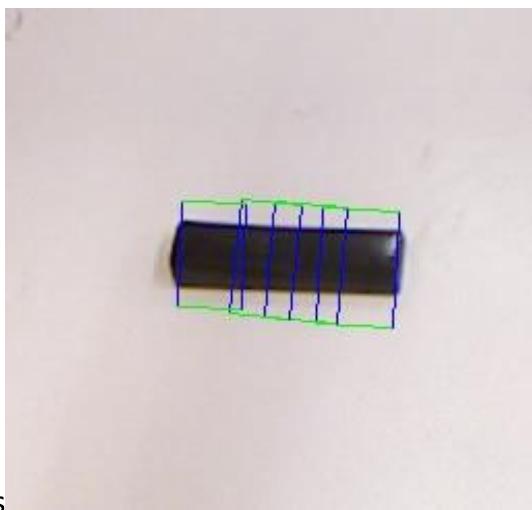
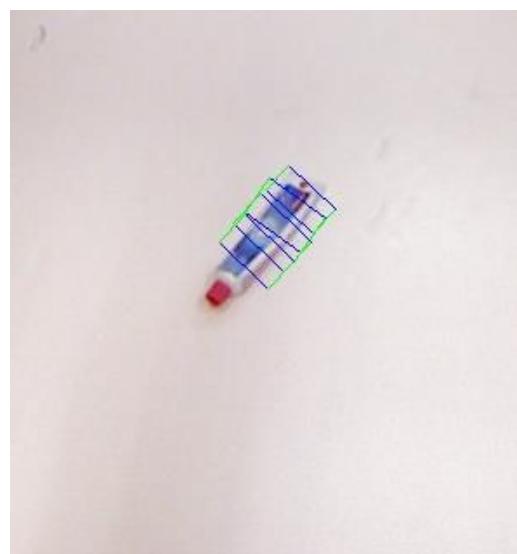
每个物体实例都标注了多个抓取姿态，包括夹爪的位置、朝向和开合状态

等信息。这些标注信息提供了用于训练和评估机器人抓取算法的有效数据。

Cornell 数据集的使用对于机器人抓取研究具有重要意义。它为研究人员提供了一个标准化的数据集，使他们能够在统一的基准上评估不同的抓取算法和系统。研究员可以利用该数据集进行机器人抓取算法的训练、验证和性能评估，进一步改进和优化机器人的抓取能力。

Cornell 数据集还在其中提供了用于抓取算法评估的测量指标，如抓取成功率、物体姿态估计准确性等。这些指标可用于比较不同算法的性能，并促进机器人抓取技术的发展和创新。

Cornell 数据集因此是一个广泛使用的机器人抓取任务数据集，提供了真实场景中的物体实例图像、深度图像和抓取姿态的标注信息，为机器人抓取算法的研究和评估提供了有价值的资源。



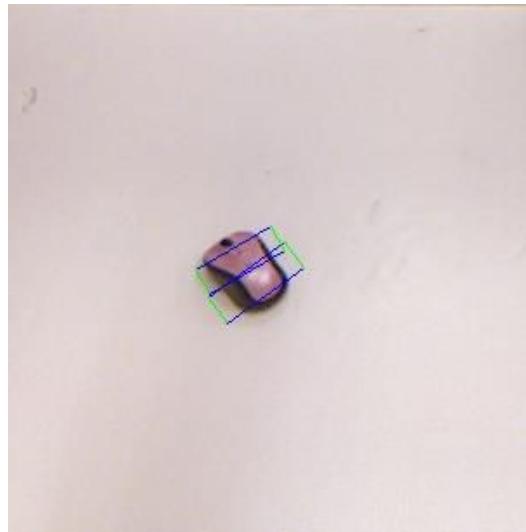


图 4-4 cornell 数据集标定实例

(2) Jacquard 数据集

Jacquard 数据库 (Jacquard Dataset) 是一个用于机器人视觉和控制研究的数据集，由麻省理工学院 (MIT) 的 Jacquard 团队创建。该数据集是为了推动机器人对复杂物体的感知和操作能力而设计的。

Jacquard 数据库也包含了许多 RGB-D 图像和相关的注释信息。这些图像是在真实环境中拍摄的，包括各种不同的物体和场景。数据集中的物体类别涵盖了日常生活中常见的物体，如杯子、书籍、玩具等。

每个物体实例都具有精确的 3D 模型，包括几何形状和表面纹理。此外，Jacquard 数据库还提供了每个物体实例的手动标注，如物体的姿态、抓取点和抓取方向。这些注释信息为机器人抓取算法的开发和评估提供了基准数据。

Jacquard 数据库的独特之处在于它提供了丰富的变化和挑战性。数据集包含了物体的不同视角、光照条件和背景干扰，以模拟真实世界中的复杂环境。这使得研究人员可以测试和改进机器人对物体的感知和操作能力，以应对实际应用中的各种情况。

此外，Jacquard 数据库还提供了用于评估机器人抓取算法性能的测量指标，如抓取成功率、位姿估计误差等。这些指标使研究人员能够定量评估不同算法的效果，并促进机器人感知和操作技术的发展。

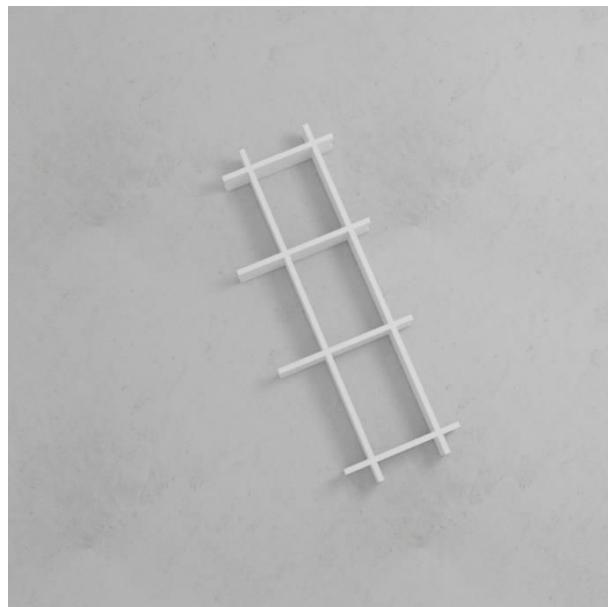


图 4-5 抓取图像实物图

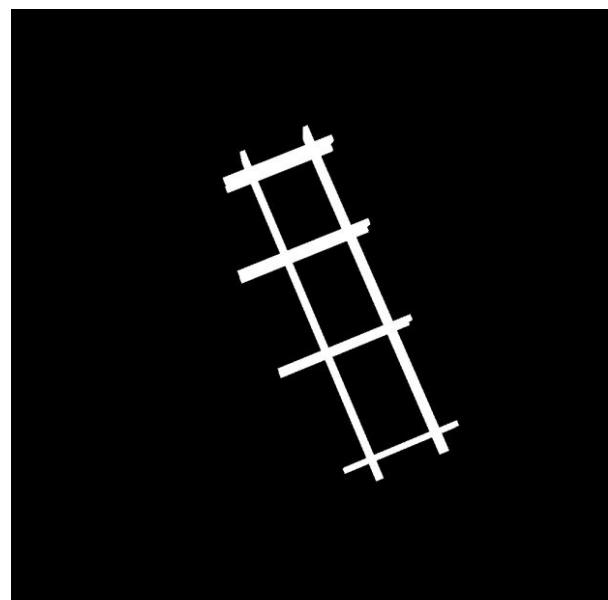


图 4-6 抓取图像掩码图示例



图 4-7 抓取图像深度图示例

4.4.2 实验抓取数据集制作

本小节介绍了实验所需要抓取的物体基于 GR-ConvNet 需求和基于 YOLO-GRConvNet 框架制作的实验自用数据集流程。

对于 GR-ConvNet 的自用数据集而言，首先我们同样使用 1: 1 的图像对需要抓取的物体进行各自的拍照和合并拍照，合并拍照的目的主要是为了训练 GR-ConvNet 网络框架对多物体的目标检测与抓取位姿识别能力。接着我们使用 rolabelimg 这个 python 工具来对拍下来的物体进行抓取框的手动标定，如下图 4-8 为具体图片标定的过程，并且对图片的标定完成后将标定好的框转换成 xml 格式的文件以及将 xml 格式通过 python 工具转为 txt 文件以便于后续的算法的调用。然后将图片的格式按照 cornell 数据集的格式进行整理。最后将模型中的参数进行一系列的更改过后对模型进行了训练。

对于 YOLO-GRConvNet 网络框架所需要的数据集则是直接上一章中使用过的改进 YOLOv5 训练所使用的同一数据集即可。

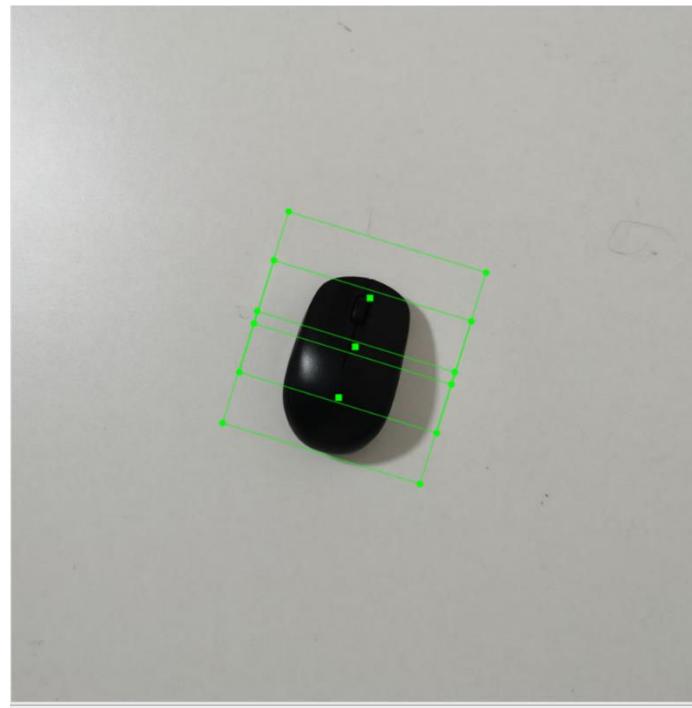


图 4-8 图像抓取框标定流程示意图

4.4.3 模型训练结果

如下表为 GR-ConvNet 在 cornell 标准抓取数据集上的训练结果。下图 4-9 为 GR-ConvNet 对 cornell 官方数据集抓取检测预测结果的实例，下图 4-10 为 GR-ConvNet 对本章自制数据集抓取检测预测结果的实例

表 4-1 GR-ConvNet 训练结果表

参量	训练结果
最优成功率	96.62
最终成功率	92.13
IOU	93.26

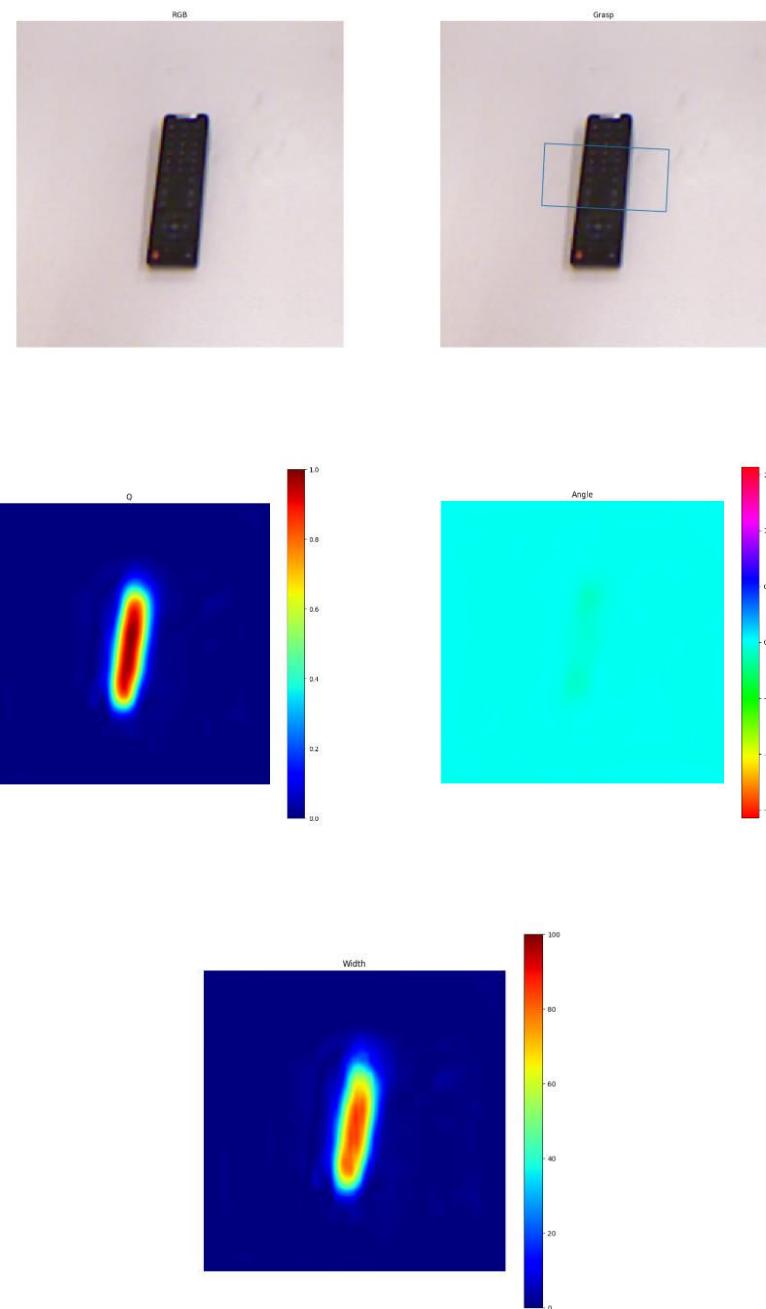


图 4-9 GR-ConvNet 对 cornell 官方数据集物体的抓取结果预测示意图 (1) 抓取原 rgb 图
(2) 抓取选取框图 (3) 抓取成功率 Q (4) 抓取角度 Angle (5) 抓取宽度 Width

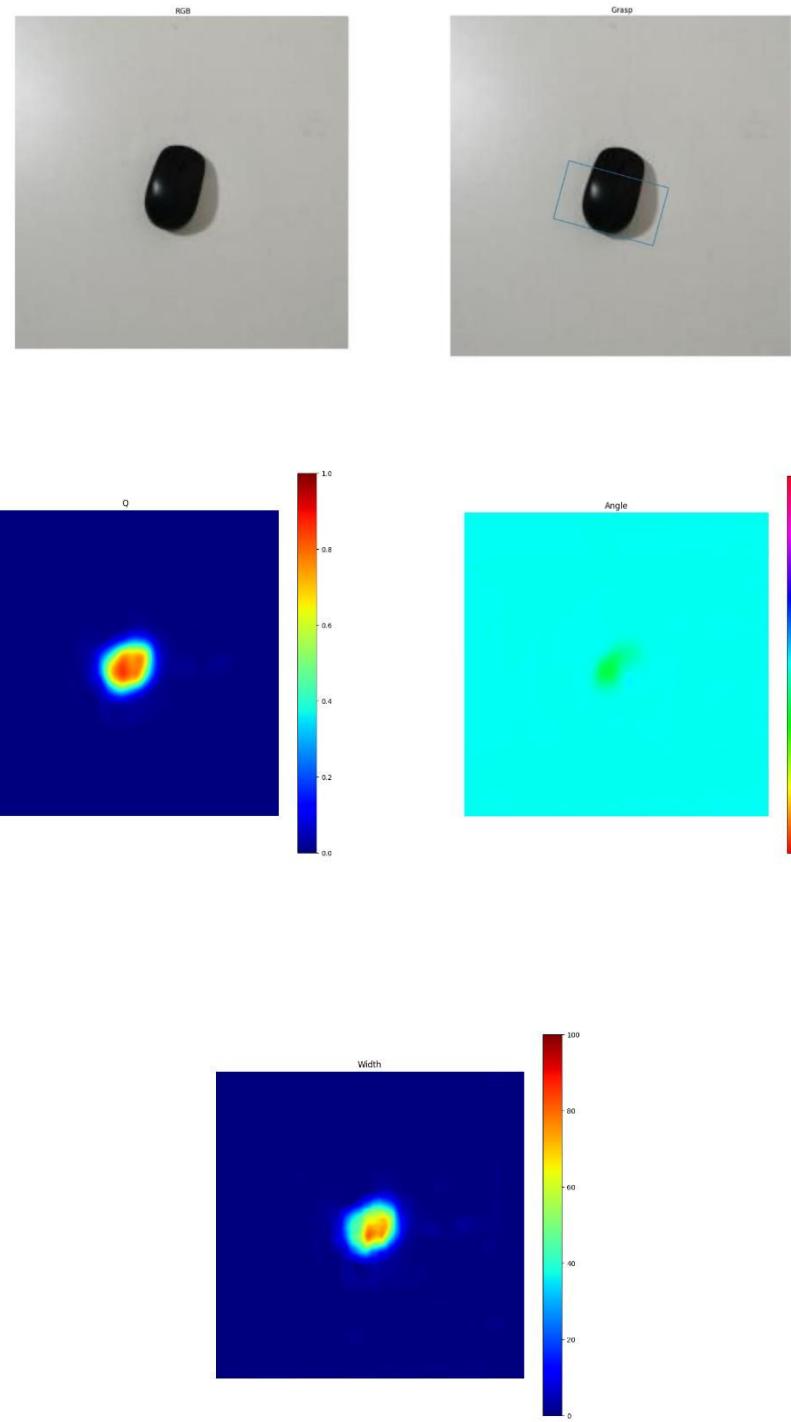


图 4-10 GR-ConvNet 对本章自制数据集物体的抓取结果预测示意图
(1) 抓取原 rgb 图
(2) 抓取选取框图 (3) 抓取成功率 Q (4) 抓取角度 Angle (5) 抓取宽度 Width

4.5 本章小结

本章首先对上述引入的 GR-ConvNet 进行了详细的功能介绍并剖析了它的具体网络架构与该模型的具体运行流程。之后分析了为什么 YOLO 与 GR-

ConvNet 进行融合符合面向任务的抓取检测方法研究的需要，并且详细的讲解了 YOLOv5 框架与 GR-ConvNet 网络架构的融合过程，还提出了总体框架运行流程。随后在进行仿真实验验证之前，本章对 GR-ConvNet 训练和预测时常用的两大数据集 cornell 和 jacquard 进行了一定的介绍，接着便展示了融合网络框架和 GR-ConvNet 网络各自的实验抓取数据集的制作流程，最后展示了网络模型架构在 cornell 官方数据集上的模型训练成果以及该模型架构分别对官方数据集的物体和对自制数据集的物体的输出图片结果。

5 实验设计和结果分析

5.1 引言

为了验证本文所提的面向具体任务的抓取检测位姿识别方法研究技术，本章首先搭建了实验所用到的硬件平台并引入了软件平台，且在实验平台的基础上进行了软件算法系统的移植与整体 YOLO-GRConvNet 框架的应用，对其中的抓取位姿识别模块进行了详细的介绍。最后在真实场景下进行实际抓取任务，结果表明本文所提出的技术能够较好地实现面向任务的抓取检测位姿输出方法。

5.2 面向任务的抓取检测平台搭建

根据抓取任务的需求搭建硬件平台如图 5-1 所示，实验所使用的设备包括华数工业机械臂 HSR-C0605、机械夹爪、海康工业相机 MV-CU050-60GM、气泵桌及目标物体组成的工作区域和计算机等。其中深度相机固定在机械臂末端。



图 5-1 硬件平台

华数工业机械臂型号为 HSR-C0605，如图 5-2（1）所示，具有六个关节，

每个关节都可以进行位置控制和速度控制，可以实现高精度定位，其关键参数如表 5-1 所示。

表 5-1 华数机械臂关键参数

参数名称	参数值	参数名称	参数值
本体重量	26kg	有效负载	5kg
最大工作半径	1005mm	重复定位精度	±0.02mm
自由度	6	关节最大速度	240° /s

机械夹爪使用了一种以气泵为驱动方式的二指夹爪，如图 5-2（1）所示，能够方便的进行紧密的夹取其关键参数如表 5-2 所示

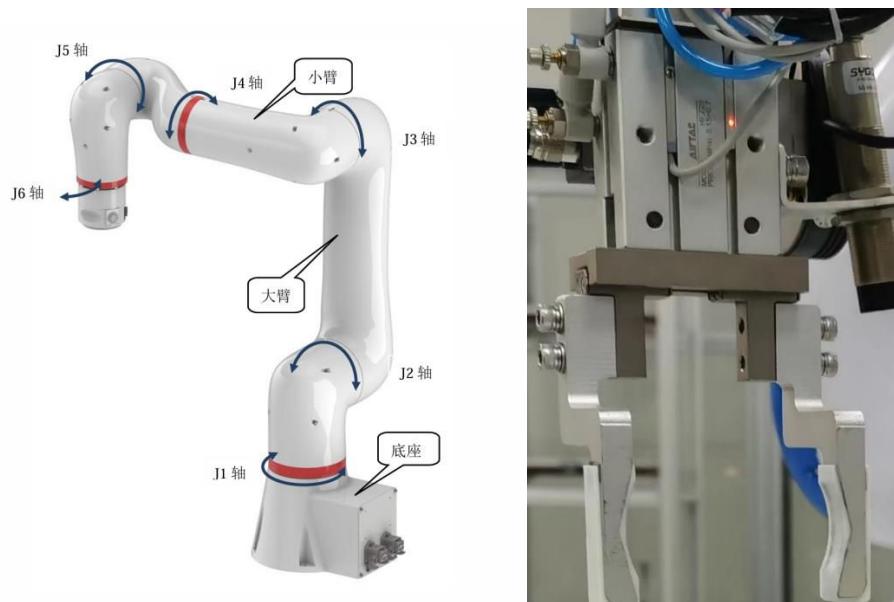
表 5-2 机械夹爪参数

参数名称	参数值	参数名称	参数值
重复精度	±0.01	抓取范围	0-85mm
型号	复动型	工作压强	0.2-0.7Mpa

MV-CU050-60GM 海康工业相机是一款 RGB 相机，用于在抓取检测实验中获取场景的 RGB 信息，该工业相机被广泛应用于机械臂抓取领域，如图 5-2（2）所示。本文使用 MV-CU050-60GM 获取抓取领域内物体图像的 RGB 信息，并将其输入 labelimg 进行滚图像标定。其关键参数如表 5-3 所示。

表 5-3 MV-CU050-60GM 工业相机关键参数

参数名称	参数值	参数名称	参数值
外观尺寸	29×29×30mm	图像帧率	48.2fps
图像分辨率	2592×1944px	像元尺寸	2.2 × 2.2 μm
传感器类型	CMOS	动态范围	65.4dB



(1) HSR-C0605 机械臂和气泵夹爪



(2) MV-CU050-60GM RGB 相机

图 5-2 硬件设备

软件平台使用的时实验室平台 forallbot，该平台由 C++语言为框架，在 Windows10 操作系统平台进行开发完成。

5.3 实验验证与结果分析

本小节为具体的面向任务抓取检测方法应用的实验流程。首先根据第二章式 2-9 构建的相机成像模型和手眼模型对相机进行了标定和手眼标定。其中，相机标定的目的是得到相机内参，采用了张正友标定法，具体通过在不同位置的抛射标定板进行计算，手眼标定也是通过第二章推到的式 2-12，在相机标定的过程中通过相机在世界坐标系的标定同时进行标定。

结束了实验前的标定过程后，本节便进入到正式的实验环节。为了验证本

文抓取位姿估计方法是否具有实用性，本节采用了多目标场景下的抓取环境来对算法的稳定性实用性进行实验考校。由于抓取器抓取质量和抓取宽度的限制，本节选取了质量小于 500g，基于抓取宽度在 10-80mm 的三个物体作为实验对象，如图 5-3 所示。机械臂首先移动到初始坐标进行实验平台上的物体初态位置的拍照与数据采集，接着将黑白的特征图片输送到上位机中，进入到 YOLO-GRConvNet 框架作为输入图像，YOLO-GRConvNet 融合框架的操作流程可以从第四章得知。并且其图片操作流程如图 5-4 所示，YOLO-GRConvNet 在经过改进的 YOLOv5 框架进行目标识别后会输出目标检测预测结果图如图 5-4（1）所示，在经过工具对图像进行裁剪的流程后，图片将会直接输入到整体框架中的抓取位子识别模块内，并进行抓取位姿的输出，具体模型最终输出图像如图 5-5 所示。

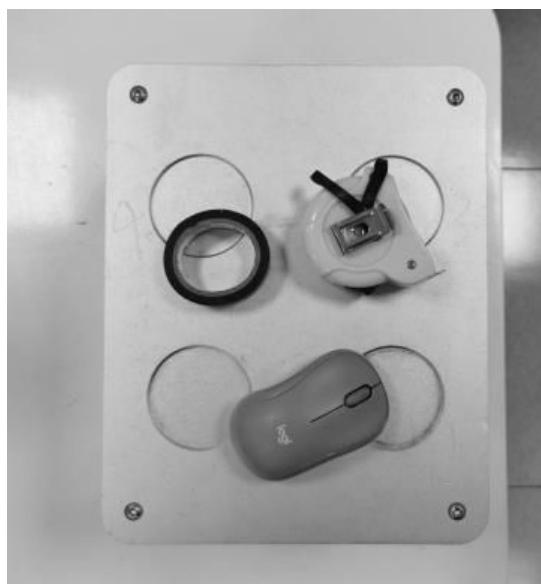
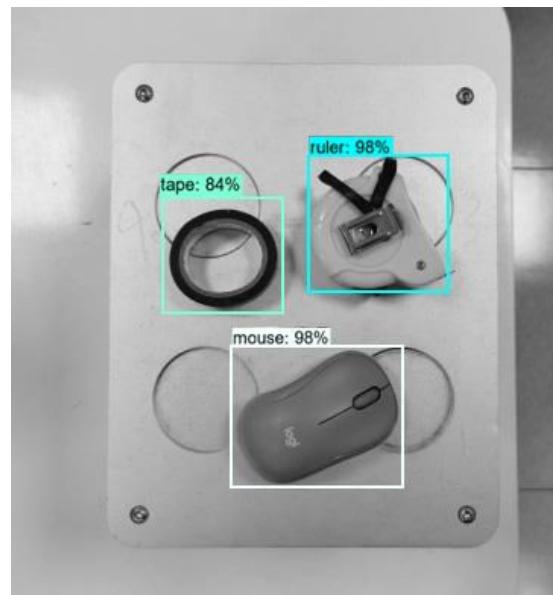


图 5-3 待抓取物体的初态图片示意图



(1) 改进的 YOLOv5 目标检测算法输出图片



(2) 对目标物体进行裁剪后的输出图片

图 5-4 框架运行流程中的中间图片输出和图片处理示

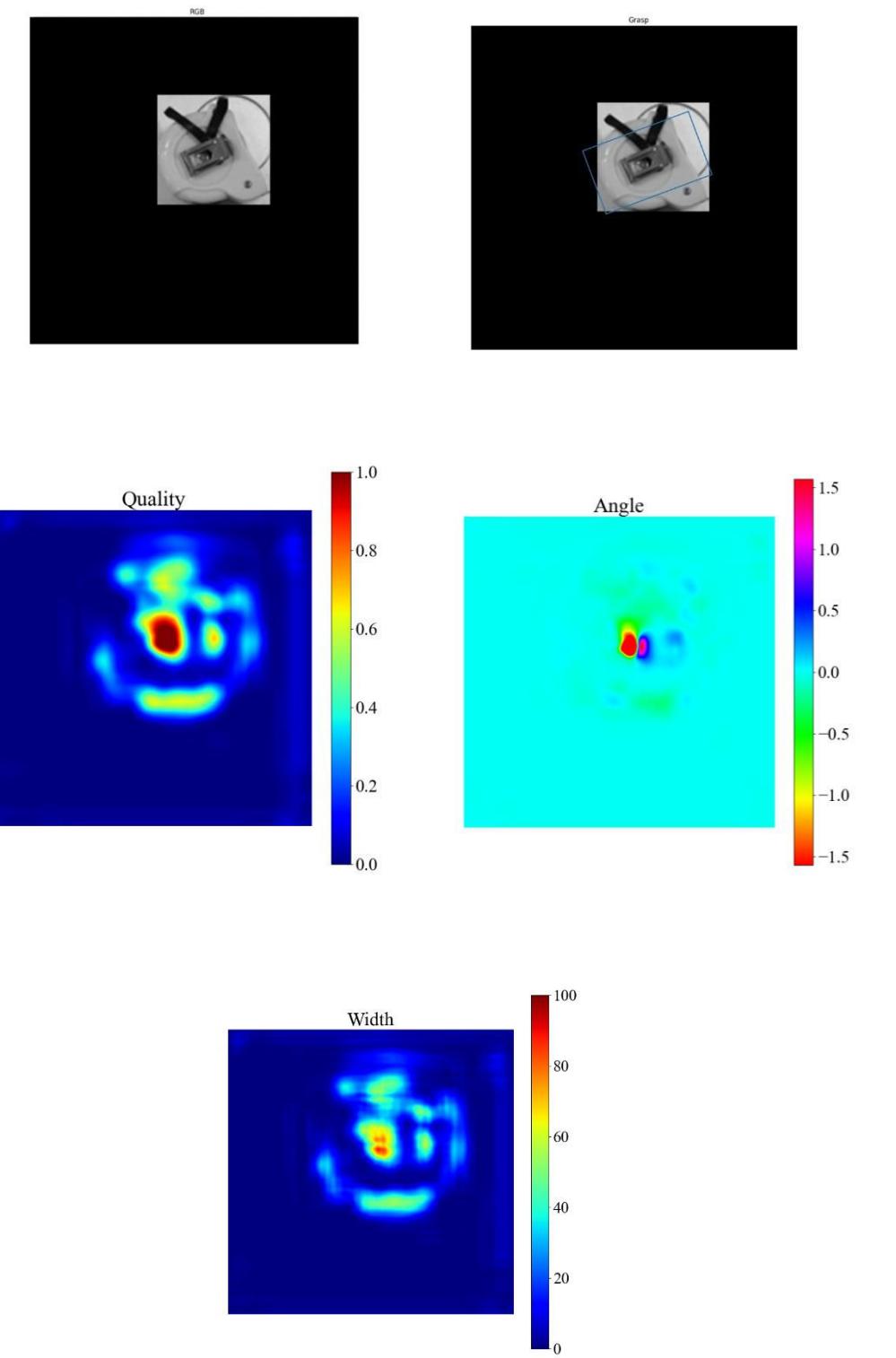


图 5-5 YOLO-GRConvNet 对 cornell 官方数据集物体的抓取结果预测示意图

- (1) 抓取原rgb图 (2) 抓取选取框图 (3) 抓取成功率 Q (4) 抓取角度 Angle (5) 抓取宽度 Width

具体实验抓取过程如下图 5-6 所示，并且成功率达到了 8/10，这说明该方法有实用性与可行性，两步阶联框架的训练参数如下表 5-4、5-5 所示，与代表性抓取框架的对比^[54]如下表 5-6 所示。

表 5-4 目标检测算法训练参数

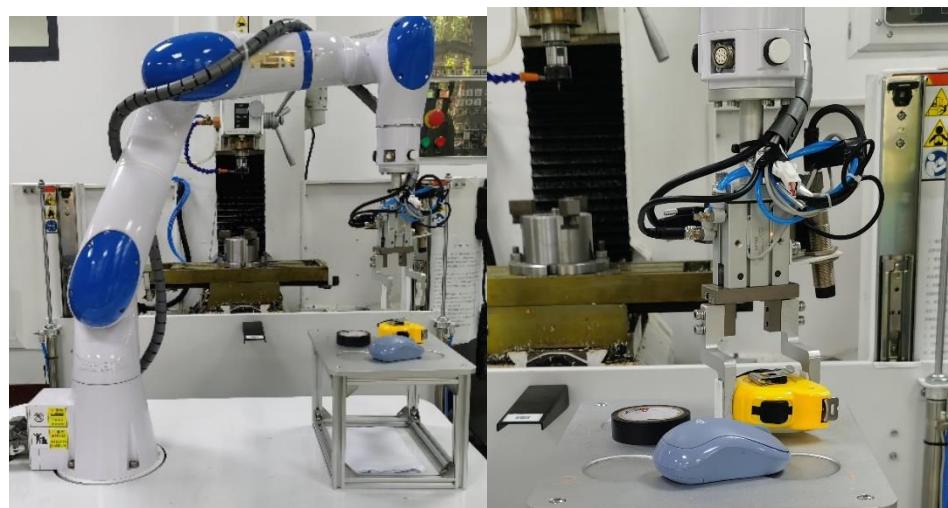
参数	值
主干网络	Swin transformer
训练集数	240
验证集数	80
优化器	Adam
学习率	[1e-3,2e-5]
冻结训练回合数	50
冻结训练 batch_size	12
非冻结训练 batch_size	8
最大训练回合数	100

表 5-5 抓取识别算法训练参数

参数	值
主干网络	GR-ConvNet
优化器	Adam
输入尺寸	224
batch_size	8
epoch	30
最大训练回合数	3000

表 5-6 抓取结果对比表

框架	多类别场景 抓取成功率(%)
	Mask-RCNN+PCA
6D GraspNet ^[56]	62
No classification ^[57]	53
YOLO-GRCConvNet	80



(1) 开始对物体进行抓取



(2) 将物体抓起过程



(3) 将物体抓取到指定位置

图 5-6 抓取测试实验流程实际图

5.4 本章小结

本章为了验证所提出的 YOLO-GRConvNet 框架的有效性，首先搭建了面向任务的抓取检测实验平台，然后在实验平台中引入了软件网络架构，内嵌入机器人抓取检测程序中并且最后基于搭建的实验平台与实验室的软件平台面向具体抓取对象的实现了机器人抓取位姿识别任务和抓取任务。

6 总结与展望

6.1 全文总结

本文对面向任务的机械臂抓取位姿估计方法展开了研究，达成了提高抓取位姿估计输出稳定性，成功率，输出结果速度等目标，本文研究了目标检测算法的基础知识、位姿估计所需的模型建立并重点研究了目标检测算法的改进、抓取检测算法的框架以及两种算法的融合使用，最终实现了结合深度学习的机械臂抓取末端位姿估计方法，本文具体研究工作内容总结如下：

(1) 以多个现实角度分析了机械臂抓取应用背景，并且通过剖析现有的算法中存在的疏漏，并因此提出了本文所采用的具体网络框架，并说明了所采用的算法的先进性和意义，解释了本文拟解决的关键技术问题。

(2) 具体系统的建立了机器人抓取位姿估计所需的各模型以及系统的介绍了所需要的深度学习算法的相关知识。对相机成像模型、机械臂手眼模型和机械臂运动学模型进行模型建立与具体推导，方便了实验中的位姿估计和运动规划流程研究。介绍了关于卷积神经网络相关必要知识，主要研究了卷积层、池化层、激活函数等搭建卷积神经网络所需要的必要知识。

(3) 了解了 YOLOv5 原网络的基础结构并列举了其中将在后续内容中被改进的原网络的缺点与不足，引出了本文最重要的目标检测算法的改动部分并对本文主要改动的三个部分进行了具体的阐述，分别为多注意力模块内嵌的分层特征映射主干替换，基于宽高差异值的快速收敛损失函数改进和自适应特征优化的端到端注意力模块的嵌入。最后对改进后的目标检测网络进行了仿真实验验证，展示了自制数据集的方法与实验过程并将改动后设置好参数的算法在同一数据集中的结果（AP、FPS）进行对比从而体现改动的优势与合理性。

(4) 研究了抓取检测位姿估计算法（GR-ConvNet）的网络架构和其比起现有其他网络架构的优势性，并且研究了改进的 YOLOv5 与 GR-ConvNet 的机器人面向任务的抓取检测框架的融合。最终进行了仿真实验验证，提出了基于 GR-ConvNet 的仿 Cornell 数据集格式的自制数据集制作流程，并展示了在该数据集下改进的模型达到的实现面向具体任务的抓取位姿估计的效果且与其他文章中采用标准数据集的算法结果进行了对比。

(5) 搭建了面向具体任务的抓取检测实验所使用的硬件平台和软件平台并且展示了相关实验具体流程，还展示了具体框架输出的图片及数据，最后展示了实验结果及与其他网络的相同环境的成功率对比体现了该网络框架的合理性与实用性。

6.2 后续展望

本文针对面向任务的机械臂抓取位姿识别方法进行了讨论和研究，基于实验室现有的环境提出了面向具体任务的抓取位姿规划方法，设计了 YOLO-GRConvNet 的抓取位姿检测框架，并搭建了实验平台进行实验验证。但由于研究时间、实验室硬件条件的限制以及抓取规划问题的复杂性，本文方法仍存在许多不足之处，可以在未来的研究工作中进一步完善：

(1) 本文的 YOLOv5 改进方法比较原方有所提升，但可以通过进一步的

研究将不同网络的优势放大并且继续提升该目标识别模块的性能，且改进的途中发现新网络的网络参数比原先的主干网络大了不少，这也导致了新网路的训练时间比起原网络架构要更占资源，更耗费时间，这一点也可以通过进一步的调控网络参数构成进行优化。

(2) 本文所研究的抓取检测方法只有对抓取过程的终态位姿进行预测及结果输出，并没有对全局信息的动态路径规划研究，轨迹研究也是抓取位姿检测研究进行验证的重要因素。在后续的研究中，可以通过增加这一模块来研究机械臂的高正确率自主抓取任务。

参考文献：

- [1] 周宇权.工业机器人技术在智能制造领域中的应用研究[J].产业技术创新, 2022(002):004.
- [2] 刘亚欣, 王斯瑶, 姚玉峰, 等. 机器人抓取检测技术的研究现状[J]. 控制与决策, 2020, 35(12): 2817-2828.
- [3] 张晓寒. 基于机器人视觉识别的抓取控制研究[D].青岛理工大学,2023.
- [4] 刘瑞昊. 基于深度学习的机器人目标检测与抓取系统研究[D].江南大学,2023.
- [5] 王斌. 基于深度图像和深度学习的机器人抓取检测算法研究[D].浙江大学,2019.
- [6] 张亚辉. 基于 Faster R-CNN 目标检测的机器人抓取系统研究[D].中国科学院大学(中国科学院深圳先进技术研究院),2019.
- [7] Li Z, Xu B, Wu D, et al. A YOLO-GGCNN based grasping framework for mobile robots in unknown environments[J]. Expert Systems with Applications, 2023, 225: 119993.
- [8] Liu D, Tao X, Yuan L, et al. Robotic objects detection and gras** in clutter based on cascaded deep convolutional neural network[J]. IEEE Transactions on Instrumentation and Measurement, 2021, 71: 1-10.
- [9] Girshick R, Donahue J, Darrell T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation[C]. Proceedings of the IEEE conference on computer vision and pattern recognition, 2014: 580-587.
- [10] Kulecki B, Młodzikowski K, Staszak R, et al. Practical aspects of detection and grasping objects by a mobile manipulating robot[J]. Industrial Robot: the international journal of robotics research and application, 2021, 48(5): 688-699.
- [11] He K, Zhang X, Ren S, et al. Spatial pyramid pooling in deep convolutional networks for visual recognition[J]. IEEE transactions on pattern analysis and machine intelligence, 2015, 37(9): 1904-1916.
- [12] 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

- [13]Redmon J, Divvala S, Girshick R, et al. You only look once: Unified, real-time object detection[C]. Proceedings of the IEEE conference on computer vision and pattern recognition, 2016: 779-788.
- [14]Redmon J, Farhadi A. YOLO9000: better, faster, stronger[C]. Proceedings of the IEEE conference on computer vision and pattern recognition, 2017: 7263-7271.
- [15]Redmon J, Farhadi A. Yolov3: An incremental improvement[J]. arXiv preprint arXiv:1804.02767, 2018.
- [16] Liu W, Anguelov D, Erhan D, et al. SSD: Single shot multibox detector[C]. European conference on computer vision, 2016: 21-37.
- [17]Bohg J, Morales A, Asfour T, et al. Data-driven graspsynthesis— A survey[J]. IEEE Transactions on Robotics, 2014, 30(2): 289-309.
- [18]Caldera S, Rassau A, Chai D. Review of deep learning methods in robotic grasp detection[J]. Multimodal Technologies and Interaction, 2018, 2(3): 57.
- [19]Sahbani A, El-Khoury S, Bidaud P. An overview of3D object grasp synthesis algorithms[J]. Robotics and Autonomous Systems, 2012, 60(3): 326-336.
- [20]Konidaris G, Kuindersma S, Grupen R, et al. Robotlearning from demonstration by constructing skill trees[J]. The International Journal of Robotics Research, 2012, 31(3): 360-375.
- [21]Kamon I, Flash T, Edelman S. Learning to grasp usingvisual information[C]. Proceedings of IEEE International Conference on Robotics and Automation. Rehovot: Weizmann Science Press of Israel, 1996: 2470-2476.
- [22]Shimoga K. Robot grasp synthesis algorithms: Asurvey[J]. The International Journal of RoboticsResearch, 1996, 15(3): 230-266.
- [23]常飞, 王奔, 张小旭, 等. 基于改进 YOLOv5 的小目标检测方法研究[J]. Smart Rail Transit, 2024, 61(2).
- [24]Feng S, Qian H, Wang H, et al. Real-time object detection method based on YOLOv5 and efficient mobile network[J]. Journal of Real-Time Image Processing, 2024, 21(2): 56.
- [25]Fang Y, Liao B, Wang X, et al. You only look at one sequence: Rethinking

- transformer in vision through object detection[J]. Advances in Neural Information Processing Systems, 2021, 34: 26183-26197.
- [26] Sulabh Kumra, Shirin Joshi, Ferat Sahin,"Antipodal Robotic Grasping using Generative Residual Convolutional Neural Network", IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) 2020,arXiv:1909.04810
- [27] Depierre A, Dellandréa E, Chen L. Jacquard: A large scale dataset for robotic grasp detection[C]//2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE, 2018: 3511-3516.
- [28] P. Schmidt, N. Vahrenkamp, M. Wachter, and T. Asfour, “Grasping of unknown objects using deep convolutional neural networks based on depth images,” in 2018 IEEE International Conference on Robotics and Automation (ICRA). IEEE, 2018, pp. 6831–6838.
- [29] A. Zeng, S. Song, K.-T. Yu, E. Donlon, F. R. Hogan, M. Bauza, D. Ma, O. Taylor, M. Liu, E. Romo, et al., “Robotic pick-and-place of novel objects in clutter with multi-affordance grasping and cross-domain imagematching,” in 2018 IEEE International Conference on Robotics and Automation (ICRA). IEEE, 2018, pp. 1–8.
- [30] L. Pinto and A. Gupta, “Supersizing self-supervision: Learning to grasp from 50k tries and 700 robot hours,” in 2016 IEEE international conference on robotics and automation (ICRA). IEEE, 2016, pp. 3406–3413.
- [31] J. Varley, C. DeChant, A. Richardson, J. Ruales, and P. Allen, “Shape completion enabled robotic grasping,” in 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE, 2017, pp. 2442–2447.
- [32] D. Guo, F. Sun, H. Liu, T. Kong, B. Fang, and N. Xi, “A hybrid deep architecture for robotic grasp detection,” in 2017 IEEE International Conference on Robotics and Automation (ICRA). IEEE, 2017, pp. 1609–1614.
- [33] Mahler J, Liang J, Niyaz S, et al. Dex-net 2.0: Deep learning to plan robust grasps with synthetic point clouds and analytic grasp metrics[J]. arxiv preprint arxiv:1703.09312, 2017.

- [34] Levine S, Pastor P, Krizhevsky A, et al. Learning hand-eye coordination for robotic grasping with deep learning and large-scale data collection[J]. The International journal of robotics research, 2018, 37(4-5): 421-436.
- [35] Antanas L, Moreno P, Neumann M, et al. Semantic and geometric reasoning for robotic gras**: a probabilistic logic approach[J]. Autonomous Robots, 2019, 43: 1393-1418.
- [36] 伍锡如,雪刚刚.基于图像聚类的交通标志 CNN 快速识别算法[J].智能系统学报,2019,14(04):670-678.
- [37] 骆训浩.卷积神经网络中非线性激活函数的研究与应用[D].大连理工大学,2018.
- [38] Yi X, Song Y, Tang X. Weak Supervised Surface Defect Detection Method Basedon Selective Search and CAM[C]//2019 Chinese Automation Congress (CAC).IEEE, 2019: 4386-4391.
- [39] Li B, He Y. An improved ResNet based on the adjustable shortcut connections[J].IEEE Access, 2018, 6: 1 8967-1 8974.
- [40] 刘思诚, 邓皓, 等. 基于 YOLOv5 改进的小目标检测算法[J]. Ordnance Industry Automation, 2022, 12(41): 12.
- [41] Xu, et al. "Reluplex made more practical: Leaky ReLU." 2020 IEEE Symposium on Computers and communications (ISCC)[J]. IEEE, 2020.
- [42] Han J, Moraga C. The influence of the sigmoid function parameters on the speed of backpropagation learning[C]//International workshop on artificial neural networks. Berlin, Heidelberg: Springer Berlin Heidelberg, 1995: 195-201.
- [43] Misra D. Mish: A self regularized non-monotonic activation function[J]. arxiv preprint arxiv:1908.08681, 2019.
- [44] 杨观赐, 杨静, 李少波, 胡建军等. 基于 Dropout 与 ADAM 优化器的改进 CNN 算法[J]. 华中科技大学学报: 自然科学版, 2018, 46(7): 122-127.
- [45] Liu Z, Lin Y, Cao Y, et al. Swin transformer: Hierarchical vision transformer using shifted windows.Proceedings of the IEEE/CVF international conference on computer vision[J]. 2021: 10012-10022.

- [46]Zheng, Zhaohui, et al. "Distance-IoU loss: Faster and better learning for bounding box regression." Proceedings of the AAAI conference on artificial intelligence[J]. Vol. 34. No. 07. 2020.
- [47]Rezatofighi H, Tsoi N, Gwak J Y, et al. Generalized intersection over union: A metric and a loss for bounding box regression. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition[J]. 2019: 658-666.
- [48]祁宣豪, 智敏. 图像处理中注意力机制综述[J]. Journal of Frontiers of Computer Science & Technology, 2024, 18(2).
- [49]李光明, 张倩, 金瑾, 等. 基于 CNN 和 Transformer 的轻量级超分辨率重建网络研究[J]. Computer Science and Application, 2023, 13: 93.
- [50]孙亮, 马江, 阮晓钢. 六自由度机械臂轨迹规划与仿真研究[D]. , 2010.
- [51]卢荣胜, 史艳琼, 胡海兵. 机器人视觉三维成像技术综述[J]. Laser & Optoelectronics Progress, 2020, 57(4): 040001.
- [52]肖帅, 王韬, 张树华, 等. 一种基于手眼相机的空间机械臂在轨标定方法[J]. 空间控制技术与应用, 2022, 48(3): 72-77.
- [53]冷舒, 吴克, 居鹤华. 机械臂运动学建模及解算方法综述[J]. 宇航学报, 2019, 40(11): 1262-1273.
- [54]葛俊彦, 史金龙, 周志强等. 基于三维检测网络的机器人抓取方法[J]. 仪器仪表学报, 2023(8):146-153.
- [55]吴善宝. 基于头盔检测和车牌识别的电动车安全监控系统[D].浙江科技大学, 2024.DOI:10.27840.
- [56]Fang H S, Wang C, Gou M, et al. GraspNet: a large-scale clustered and densely annotated dataset for object grasping[J]. arxiv:1912.13470.
- [57]Ten Pas A, Gualtieri M, Saenko K, et al. Grasp pose detection in point clouds[J]. The International Journal of Robotics Research, 2017, 36(13-14): 1455-1473.