



华南理工大学  
South China University of Technology

# 专业学位硕士学位论文

## 语言视觉激光多模态 融合的机器人导航方法

作者姓名 杨礼铭

学位类别 电子信息硕士(计算机技术)

指导教师 毕盛 副教授  
覃争鸣 高级工程师

所在学院 计算机科学与工程学院

论文提交日期 2025年4月17日



# **Language vision laser multi-modal fusion robot navigation method**

A Dissertation Submitted for the Degree of Master

**Candidate: Liming Yang**

**Supervisor: Associate Prof.Sheng Bi**

South China University of Technology

Guangzhou, China



分类号：TP3

学校代号：10561

学 号：202221043765

## 华南理工大学硕士学位论文

# 语言视觉激光多模态融合的机器人导航方法

作者姓名：杨礼铭

指导教师姓名、职称：毕盛 副教授

申请学位级别：工学硕士

学科专业名称：计算机技术

研究方向：机器人室内导航

论文提交日期：2025年5月30日

论文答辩日期：2025年5月30日

学位授予单位：华南理工大学

学位授予日期： 年 月 日

答辩委员会成员：

主席：张平

委员：徐红云 黄艳 袁华 潘丹



# 华南理工大学

## 学位论文原创性声明

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，本论文不包含任何其他个人或集体已经发表或撰写的成果作品。对本文的研究做出重要贡献的个人和集体，均已在文中以明确方式标明。本人完全意识到本声明的法律后果由本人承担。

作者签名： 日期： 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留、使用学位论文的规定，即：研究生在校攻读学位期间论文工作的知识产权单位属华南理工大学。学校有权保存并向国家有关部门或机构送交论文的复印件和电子版，允许学位论文被查阅（除在保密期内的保密论文外）；学校可以公布学位论文的全部或部分内容，可以允许采用影印、缩印或其它复制手段保存、汇编学位论文。本人电子文档的内容和纸质论文的内容相一致。

本学位论文属于：

保密（校保密委员会审定为涉密学位论文时间：\_\_年\_\_月\_\_日），于\_\_年\_\_月\_\_日解密后适用本授权书。

不保密，同意在校园网上发布，供校内师生和与学校有共享协议的单位浏览；同意将本人学位论文编入有关数据库进行检索，传播学位论文的全部或部分内容。

(请在以上相应方框内打“√”)

作者签名： 日期：

指导教师签名： 日期：

作者联系电话： 电子邮箱：

联系地址(含邮编)：广东省广州市天河区华南理工大学（五山校区）3号楼



## 摘要

随着机器人需求市场的不断扩大，机器人逐渐从实验室跻身到酒店服务、工厂物流、家政服务、医疗看护、教育娱乐等各行各业之中，为推动生产力的持续发展做出了巨大贡献。在众多的应用场景之中移动机器人主要以室内环境作为其主要的工作场景，以自主导航作为其完成其他复杂任务的基础功能。目前大部分广泛应用于室内服务的移动机器人都采用点云激光进行建图与实时定位来实现自主导航，然而这种方法无法利用图像丰富的特征信息进行导航，容易在平坦、多重复场景这类特征不明显的环境中定位失败，已成为亟待解决的难题。针对在移动机器人室内导航过程中，单一使用视觉语言导航算法无法充分利用语义中的方位和环境中的感知信息、无法导航至目标半米内的问题，本文提出了一种语言视觉激光多模态融合的机器人导航方法。首先，在全局路径规划中，标记地图中的导航点，保留其位姿、图像、点云图和各点之间的拓扑信息，通过多模态融合网络得到各导航点与目标的匹配权值，结合 dijkstra 算法和方位优化算法，规划出全局路径导航点序列。然后在局部路径规划中，通过特征提取、特征融合和运动模块在局部未知环境中探索目标，将多线激光与单目相机进行联合标定，进一步通过目标检测、点云聚类和坐标变换方法得到目标具体位姿，发布导航任务以完成局部路径的规划。最后，通过仿真实验和真实环境实验，验证所提出的导航方法的有效性和可行性。本文的主要贡献如下：

- (1) 本文提出了一种全局路径规划导航方法。与前人的工作相比，针对静态目标导航任务所提出的全局路径规划导航方法基于单目相机、激光雷达等多种传感器和基于多模态特征融合神经网络，增强系统对当前环境和导航过程中的认知和感知能力，再通过方位优化算法筛除噪声导航点，提高导航点选择的正确率的同时提高后续规划的计算响应速度，最后通过导航点规划算法加权融合多种策略进一步提高导航的准确率和导航效率。
- (2) 本文提出了一种未知环境的目标物体探索方法。与前人的工作相比，针对动态目标导航任务所提出的局部目标物体探索方法基于多特征提取和融合的方法，在同一嵌入空间内利用注意力机制融合视觉特征和文本特征，有效的构建了视觉表示和目标物体所在导航方向的关联，使系统能够通过探索找到在变化的环境中的目标物体。
- (3) 本文设计了一套单目相机和多线激光融合的图像点云融合方法，联合视觉观察

的认知信息和多线点云的感知信息让移动机器人能够有效地在仿真环境和真实环境中依据自然语言指令完成目标导航任务，在仿真环境和真实机器人上部署并完成一系列可行性与性能测试，实验结果表明该方法具有一定的有效性和优越性。

**关键词：**移动机器人；自主导航系统；多传感器融合；路径规划

## Abstract

With the continuous expansion of the robot demand market, robots gradually from the laboratory to hotel services, factory logistics, domestic service, medical care, education and entertainment and other industries, to promote the sustainable development of productivity has made great contributions. In many application scenarios, the mobile robot mainly takes indoor environment as its main working scene, and autonomous navigation as its basic function to complete other complex tasks. At present, most of the mobile robots widely used in indoor services use point cloud laser for mapping and real-time positioning to achieve autonomous navigation. However, this method cannot make use of the rich feature information of images for navigation, and it is easy to fail to locate in the environment with unclear features such as flat and multiple repeated scenes, which has become an urgent problem to be solved. In order to solve the problem that the single visual language navigation algorithm can not make full use of the semantic orientation and the perceptual information in the environment, and can not navigate to the target within half a meter in the indoor navigation process of mobile robots, this paper proposes a multi-modal fusion robot navigation method based on language vision laser. First of all, in global path planning, navigation points in the map are marked, their pose, image, point cloud image and topological information between points are retained, and the matching weights of each navigation point and the target are obtained through multi-modal fusion network. Combining dijkstra algorithm and orientation optimization algorithm, the global path navigation point sequence is planned. Then, in local path planning, the target is explored in the local unknown environment through feature extraction, feature fusion and motion module, the multi-line laser and monocular camera are jointly calibrated, and the specific pose of the target is obtained through target detection, point cloud clustering and coordinate transformation methods, and the navigation task is released to complete the local path planning. Finally, the effectiveness and feasibility of the proposed navigation method are verified by simulation experiments and real environment experiments. The main contributions of this paper are as follows:

- (1) This paper presents a global path planning navigation method. Compared with previous work, the proposed global path planning navigation method for static target navigation tasks is based on a variety of sensors such as monocular camera and Lidar and

a multi-modal feature fusion neural network to enhance the system's cognition and perception of the current environment and navigation process, and then filters out the noise navigation points through the orientation optimization algorithm. The accuracy rate of navigation point selection is improved, and the computational response speed of subsequent planning is improved. Finally, the navigation accuracy and efficiency are further improved through the weighted integration of navigation point planning algorithms.

- (2) This paper proposes an exploration method for target objects in an unknown environment. Compared with previous works, the local target object exploration method proposed for the dynamic target navigation task is based on the method of multi-feature extraction and fusion. Within the same embedding space, the attention mechanism is utilized to fuse visual features and text features, effectively constructing the association between the visual representation and the navigation direction where the target object is located. Enable the system to find the target objects in the changing environment through exploration.
- (3) In this paper, a set of image point cloud fusion method based on monocular camera and multi-line laser fusion is designed. By combining cognitive information of visual observation and perception information of multi-line point cloud, the mobile robot can effectively complete target navigation tasks according to natural language instructions in simulation environment and real environment, and a series of feasibility and performance tests are deployed and completed on simulation environment and real robot. The experimental results show that this method has certain effectiveness and superiority.

**Keywords:** Mobile Robot; Autonomous Navigation System; Multi-Sensor Fusion; Path planning

# 目 录

<b>摘 要</b> .....	I
<b>Abstract</b> .....	III
<b>插图目录</b> .....	VIII
<b>表格目录</b> .....	X
<b>第一章 绪论</b> .....	1
1.1 研究背景和意义 .....	1
1.2 国内外研究现状 .....	3
1.2.1 基于 SLAM 技术的导航方法 .....	3
1.2.2 视觉语言导航方法 .....	5
1.2.3 具身智能 .....	7
1.3 研究内容 .....	7
1.4 组织架构 .....	8
<b>第二章 相关工作</b> .....	10
2.1 ROS 系统的通信与导航 .....	10
2.2 联合标定 .....	11
2.3 目标检测网络 .....	13
2.4 多模态特征融合网络 .....	16
2.5 点云聚类 .....	18
2.6 PID 控制器 .....	20
2.7 本章小结 .....	21
<b>第三章 全局路径规划导航方法</b> .....	22
3.1 导航任务定义 .....	22
3.2 导航框架设计 .....	23
3.3 指令语义提取模块 .....	26
3.4 多模态融合网络模块 .....	28
3.5 方位优化算法 .....	32
3.6 导航点规划算法 .....	35
3.7 本章小结 .....	37

<b>第四章 局部路径规划导航方法</b>	38
4.1 引言	38
4.2 导航框架设计	39
4.3 特征提取模块	41
4.3.1 提取局部特征	42
4.3.2 提取全局特征	43
4.3.3 提取目标特征	44
4.4 特征融合模块	44
4.4.1 特征融合编码器	46
4.4.2 特征融合解码器	46
4.4.3 LSTM 网络	48
4.5 运动模块	48
4.6 图像点云融合模块	51
4.7 本章小结	54
<b>第五章 实验设计与结果分析</b>	55
5.1 仿真环境实验设计与评估	55
5.1.1 数据集	55
5.1.2 实验参数	56
5.1.3 实验设备	57
5.1.4 评价指标	58
5.1.5 对比实验	59
5.1.6 消融实验	63
5.2 真实环境实验设计与评估	66
5.2.1 实验设备	66
5.2.2 实验环境	68
5.2.3 导航实验	70
5.3 本章小结	75
<b>结 论</b>	77
<b>参考文献</b>	80
<b>攻读硕士学位期间取得的研究成果</b>	87

致 谢 ..... 88

## 插图目录

图 1-1 多功能服务机器人 . . . . .	2
图 2-1 ROS 系统组件图 . . . . .	10
图 2-2 ROS 导航框架图 . . . . .	11
图 2-3 激光雷达和单目相机联合标定 . . . . .	12
图 2-4 YOLOV10 网络 . . . . .	15
图 2-5 Transformer 网络结构 . . . . .	17
图 2-6 障碍物点云随着距离增加变稀疏 . . . . .	20
图 2-7 PID 算法控制电机转速 . . . . .	20
图 3-1 目标物体导航过程 . . . . .	23
图 3-2 之前工作存在的问题 . . . . .	24
图 3-3 全局路径规划导航框架 . . . . .	25
图 3-4 导航图示例 . . . . .	26
图 3-5 R2R 指令集中的方位指示分布 . . . . .	27
图 3-6 指令语义提取模块 . . . . .	28
图 3-7 CLIPD 多模态融合网络框架 . . . . .	30
图 3-8 Vision Transfomer 结构 . . . . .	31
图 3-9 CLIDP 实现的核心伪代码 . . . . .	32
图 3-10 地图与实时位姿坐标变换 . . . . .	33
图 4-1 局部路径规划导航框架 . . . . .	40
图 4-2 未知环境下的导航探索框架 . . . . .	41
图 4-3 局部特征提取流程 . . . . .	42
图 4-4 全局特征提取流程 . . . . .	43
图 4-5 目标特征提取流程 . . . . .	44
图 4-6 未知环境导航示例图 . . . . .	45
图 4-7 encoder 结构图 . . . . .	47
图 4-8 decoder 结构图 . . . . .	47
图 4-9 LSTM 结构图 . . . . .	48
图 4-10 机器人的运动控制方法 . . . . .	49

图 4-11 里程计计算	50
图 4-12 预选框选择策略	53
图 4-13 根据平面成像与点云数据求目标位姿	53
图 5-1 仿真机器人模型	57
图 5-2 仿真环境与可视化点云	58
图 5-3 LM-Nav 和 LVL-Nav 路径对比图	62
图 5-4 导航指标变化图	64
图 5-5 消融实验结果	67
图 5-6 镰神 C16 激光雷达	68
图 5-7 灵遨移动机器人	68
图 5-8 嵌入式开发板	69
图 5-9 真实环境	71
图 5-10 图像点云融合可视化	72
图 5-11 真实环境成功导航示例	73
图 5-12 真实环境失败导航示例	74

## 表格目录

表 3-1 不同的语言模型测试结果 . . . . .	27
表 3-2 模型进行图像文本匹配的正确率 . . . . .	29
表 4-1 特征融合 Transfomer 网络主要参数 . . . . .	46
表 5-1 Gazebo 与其他仿真环境对比 . . . . .	55
表 5-2 机器人运动参数 . . . . .	56
表 5-3 实验平台配置参数 . . . . .	57
表 5-4 仿真环境实验定量结果对比 . . . . .	61
表 5-5 消融实验结果 . . . . .	65
表 5-6 硬件系统设计 . . . . .	69
表 5-7 硬件设备参数 . . . . .	70
表 5-8 真实环境实验定量结果对比 . . . . .	75

# 第一章 绪论

第十四届全国人民代表大会第三次会议通过的政府工作报告进一步强调，要加速推进人工智能与实体经济的深度融合，实施“人工智能+”战略进而构建具有全球竞争力的数字产业集群，通过数字化转型推动经济高质量的发展，从而达到提升人民生活质量<sup>[1]</sup>的目的。在这一背景之下机器人导航的研究与应用迎来了新的发展机遇。随着人工智能技术的不断进步和应用场景的深化，机器人目标物体导航<sup>[2]</sup>作为机器人与人工智能交叉领域的关键技术持续受到学术界和产业界的高度关注。多模态融合的导航技术近年来在自动驾驶、智能家居、服务机器人等领域取得了显著进展<sup>[3]</sup>，尤其是在复杂和动态环境中的应用潜力得到了广泛认可。尽管技术持续不断的突破，已知环境下的机器人导航仍然面临诸多挑战。本文聚焦于已知室内环境中的目标物体导航问题<sup>[4]</sup>，这一场景对传统视觉导航方法提出了更高的要求。传统方法通常依赖于预先构建的环境地图来实现机器人的定位与路径规划，但在复杂多样且动态变化的室内环境中，无法十分有效地完成导航至目标半米内的任务<sup>[5]</sup>。同时，基于语言视觉的导航方法仅依赖于视觉图像获取环境信息，无法充分利用环境中的感知信息高效地完成导航任务<sup>[6]</sup>。

本文针对室内已知环境下的多模态融合导航方法的研究及实现这一主题提出了一种语言视觉激光多模态融合的机器人导航方法，并将该方法在真实的移动机器人上进行部署，搭建成了一套已知环境下的机器人视觉导航系统。本章的主要内容包括：室内已知环境下语言视觉激光多模态融合导航方法的研究背景和意义、国内外研究现状、研究内容和论文组织架构。

## 1.1 研究背景和意义

随着我国经济的快速增长和平价经济时代的到来，物流运输、仓库存储和酒店等各行各业都在寻找能够实现缩减人工成本的同时提高工作效率和精确性的新兴技术手段。而移动机器人凭借其安全性高、灵活性好、效率突出等优势逐渐成为这些行业中的重要工具<sup>[7]</sup>，例如用于物流仓储的自动化搬运机器人、农业领域的智能喷洒设备以及医疗环境中的自主消毒装置等，如图1-1。即使应用领域不断地扩展和变化，自主导航能力仍然是移动机器人实现位置变更和环境交互的基础，是其完成特定多样化任务需求的核心技术支撑。随着这一技术的不断发展，机器人研究社区衍生出许多针对不同的场景下不同的室内已知环境导航方法来实现这种自主导航技术。这种技术主要分为激光雷达导航<sup>[8]</sup>和视觉导航<sup>[9]</sup>两种方法。



图 1-1 多功能服务机器人

早期的自主导航技术主要依赖于激光雷达来实现，其完整的导航系统整合了多个领域的算法。目前大多数的机器人研究都基于机器人操作系统 ROS<sup>[10]</sup> 进行开发，其导航框架通常包括以下步骤：利用实时定位与建图（Simultaneous Localization and Mapping, SLAM）<sup>[11, 12]</sup> 技术，激光雷达通过旋转式激光发射装置获取平面障碍物的精确距离信息，使移动机器人能够在未知环境中实时构建环境地图并同步确定自身所处的空间位姿，接着通过如 A\* 算法<sup>[13]</sup> 这类全局路径规划算法在所构建的栅格地图中生成全局路径，为了让移动机器人在变化的环境中避免碰撞的同时尽可能沿着生成的全局路径行驶，大多数机器人都采用时间弹性带（Timed Elastic Band, TEB）算法<sup>[14]</sup> 作为其局部路径规划算法进行导航，最终抵达目标点。这种预先构建精确栅格地图的传统的导航方法在静态环境中表现良好，但在动态环境中却显得力不从心。此外，这类方法难以处理视觉和语义等富含特征的信息，极大地限制了其应用范围。

人工智能技术的快速更新迭代使得机器能够通过深度学习实现对特定目标的识别与理解，这类技术赋予机器从海量数据中自主学习的能力而无需依赖人工干预。它不仅能够处理可以存储在关系数据库中的结构化数据，还能高效地分析和学习例如文本、图像和音频等这类非结构化数据，得益于深度学习这些方法的强大的学习能力，使其已在众多任务中展现出卓越的性能。人工智能算法在推荐搜索、语音识别与翻译、语义分割、目标检测等领域等典型的应用场景任务中的准确性和效率甚至超越了人类水平。总的来说这种基于深度学习模型的视觉导航方法在动态复杂多变的环境中表现良好，但这种方法无法充分利用环境中的感知信息可靠且高效地完成导航任务，其在静态环境中的

导航准确率和导航效率不如传统依赖于激光的导航方法。

上述的这些问题都制约了室内服务机器人的发展，我们急需找到一种能够动态调整以适应静态环境和复杂多变环境的导航方法。激光雷达导航方法具有更为完整的开发社区，它能利用激光雷达的感知信息预先构建精确栅格地图，并在实时导航的过程中可靠地避障并导航至目标点。在这个过程中仍存在着部分的难题。首先，需要读取并提取自然语言指令中的关键信息。其次，需要设计一种多模态融合网络，提取环境中获取的感知特征、认知特征、指令语义特征和地图特征以获得潜在的目标点的概率。最后，还需要利用指令中可能存在的方位信息设计一种方位优化算法，以提高导航的成功率。同时，由于激光雷达导航方法依赖于地图导航点的设置，无法适应于多变的环境，无法导航至目标半米内。因此还需要设计一种视觉导航方法。首先，需要设计适用于视觉导航的深度学习网络结构，以提高导航的成功率和效率。其次，必须提高模型对未知环境的适应性以提供一定的探索能力。最后，还需解决如何将训练好的模型部署到计算受限的真实机器人上的问题。综上所提出的所有问题，本文致力于语言视觉激光多模态融合的导航方法，旨在提出一套有效的室内目标物体导航方法并应用于真实机器人，使其具备类似人类的全局性和灵活性。这将有助于解决机器人在复杂多变环境下的导航需求，推动机器人产业的发展。

## 1.2 国内外研究现状

针对不同的传感器和应用场景，目标物体导航可以分为基于 SLAM 的导航和视觉语言导航。基于 SLAM 的导航是指利用激光雷达或摄像头预先在环境中创建好栅格地图，在地图中标记各导航点并利用简单的拓扑信息和目标信息实现导航。它通常应用于环境变动不大的场景，如仓储物流中的自动引导机器人、工业自动化中的生产线物料运输机器人、酒店迎宾送餐机器人等。视觉语言导航是指利用摄像头和其他传感器并借助计算机视觉技术，在环境中实现自主导航、自主避障的过程。这种导航常用于车辆导航、行人导航或地图服务中。此外，随着实体机器人在各行各业的需求不断扩大，强调智能体的感知、决策与行动都必须通过自身与环境的动态交互来实现的具身智能逐渐从理论转向了实践。

### 1.2.1 基于 SLAM 技术的导航方法

根据感知设备的不同 SLAM 技术可以划分为基于激光雷达和基于视觉传感器的两大技术路线。利用激光雷达进行环境感知与建图的技术体系的发展历程较视觉方案更为

长远。在卡尔曼滤波理论的发展过程中研究者们先后提出了扩展型和无损型两种改进算法，这两种算法在当时迅速成为 SLAM 领域的主流解决方案。其中扩展型卡尔曼滤波在实际应用中存在一定局限性，它不仅需要预先确定运动轨迹，并且在系统模型和噪声统计特性不明确的情况下还可能出现算法不收敛这种致命问题。

Murphy<sup>[15]</sup> 等人在 Rao-Blackwellised 粒子滤波中引入状态变量边缘化技术，有效缩减状态空间维度并提升采样效率。然而这种方案存在计算资源消耗过大的问题，在复杂特征的场景下会对 SLAM 算法的实时性造成负面影响。为了解决上述问题，Grisetti<sup>[16, 17]</sup> 等人进一步优化了 RBPF 框架并提出了 Gmapping 算法。该算法通过融合机器人运动学模型与局部观测数据，从而显著降低了位姿预测的不确定性。此外，其改进的重采样策略有效缓解了粒子退化的现象，使该方法成为激光 SLAM 领域兼具鲁棒性和实用性的代表性解决方案。后来，Cartographer<sup>[18]</sup> 和 LSD-SLAM<sup>[19]</sup> 这类基于图优化的 SLAM 解决方案采用前后端分离的架构设计，其中，前端模块负责处理机器人位姿的实时更新，后端模块则专注于全局位姿与环境特征的联合优化，二者协同工作构成了完整的 SLAM 系统。这种架构虽然在复杂场景下具有较高的建图精度，但在计算效率方面存在一定局限。Zhang<sup>[20]</sup> 等人将 SLAM 这一复杂问题划分成两种算法，一种以高频但低保真度执行里程计，另一种以低频但高保真度用于点云的精细匹配，提出了针对激光点云进行特征提取的激光 SLAM 系统。在轻量级的 SLAM 算法之中，Shan<sup>[21]</sup> 等人开发的 LeGO-LOAM 算法针对地面场景进行了专门优化。他们的方法通过识别点云数据中的平面和边缘特征求解计算连续扫描间的六维位姿变换，并最终实现了适用于计算受限的嵌入式平台的高效定位方案。LO-Net<sup>[22]</sup> 框架创新性地结合了激光雷达与深度学习技术，通过采用端到端训练模式，使移动机器人的位姿预测精度得到显著提升。孙海<sup>[23]</sup> 等人为了解决仓储搬运机器人的导航需求，采用基于激光的同步定位与地图构建技术来通过环境中的线性特征提取从而实现了室内地图的构建，该方案不仅解决了仓储环境中的特征识别难题，还推动了激光 SLAM 技术在实际工程中的应用发展。

基于视觉传感器的 SLAM 方法主要通过处理图像数据来实现机器人的位置感知，尽管它的研究历史与激光 SLAM 方法更短，但随着深度学习技术的不断发展也同样取得了显著的技术进展。该技术在状态估计方面主要形成了两种方法论：一种是通过提取图像特征点进行计算，另一种则直接利用图像像素信息进行处理。基于特征点的运动估计方法的核心思想是通过提取图像中的特征点并建立匹配关系，进而推算出机器人的位移和姿态变化。Davision<sup>[24, 25]</sup> 等人提出了一种 MonoSLAM 系统，通过一种主动地映射

方法、测量方法以及单目特征初始化和特征方向估计方法实现了一种能够进行实时定位的纯视觉 SLAM 系统，它是视觉 SLAM 方法的起源因此具有里程碑的意义。PTAM<sup>[26]</sup> 算法在视觉 SLAM 领域实现了重要突破，他们首次采用前端分离的架构设计，并创新性地引入了关键帧机制实现了地图构建与位姿跟踪的并行处理。但该算法在后端优化效率方面存在明显不足，这一局限促使后续研究者在他们成果的基础之上将非线性优化理论作为视觉 SLAM 的重点研究方向。此后 ORB-SLAM<sup>[27]</sup> 系统进一步拓展了视觉 SLAM 的应用范围，可支持单目、双目及深度相机等多种传感器模式，同时还通过引入闭环检测机制显著提升了定位精度。直接法视觉 SLAM 通过原始图像数据直接计算相机运动。LSD-SLAM<sup>[28]</sup> 是直接法的经典实现方案，它采用直接处理图像信息的策略而无需特征提取即可完成机器人的位姿估计，通过使用视频流中可用的数百张图像与整个密集模型对齐来精确跟踪相机的运动，在快速运动下具有卓越的跟踪性能。Foster<sup>[29]</sup> 等人提出了一种半直接单目视觉里程计 (Semi-direct Monocular Visual Odometry, SVO) 算法，SVO 消除了对昂贵的特征提取和强大的运动估计匹配技术的需求，直接作用于像素强度，使其定位精度和鲁棒性大大提高。

## 1.2.2 视觉语言导航方法

Wen<sup>[30]</sup> 等人设计了一种基于思维树的视觉语言网络 (Visual Language model with a Tree-of-Thought Net, VLTNet)，在多路径选择和必要时的回溯时创新性地使用 ToT 推理框架，能够以更高的准确性做出全局决策，在涉及复杂自然语言作为目标指令的场景中表现出色。Unl<sup>[31]</sup> 等人提出了一种由用于初始检测的 GLIP 视觉语言模型和用于验证的 InstructionBLIP 模型组成的双组件框架，解决传统方法对于标记数据的依赖局限，提高模型目标导航过程中的语义理解，从而增强机器人在不熟悉环境中的自主性。Alvare<sup>[32]</sup> 等人针对视觉语义导航问题提出了一种新的具身代理解决方案，在基于 ROS 的系统上设计了一种可以兼容视觉语义导航模型 (Visual Semantic Navigation Model, VSN-model) 的新型导航框架，以便任何的 VSN 模型都可以轻松部署在任何兼容 ROS 的机器人中，并在真实的环境中进行测试，以提高具身代理的性能和效率。但这种方法对于不同的 VSN 解决方法存在着明显的性能差异。Yuan<sup>[33]</sup> 等人提出了一种可以在目标导航任务中进行可靠前沿选择的方法，引入了一种由多元化专家前沿分析 (DEFA) 和共识决策 (CDM) 组成二点多专家决策框架，解决基于基础模型的系统中常见的荒谬或不相关的推理，在 RoboTHOR 和 HM3D 数据集上都展现了最先进的性能，擅长导航到未经训练的对象或目标。Jones<sup>[34]</sup> 等人通过利用自然语言作为通用的跨模态基础，在大数据集不

易获得的异构传感器模态上微调视觉运动通用策略，将多模态对比损失与基于感觉的语言生成损失相结合来编码高级语义，提出了能够适用于大型视觉语言动作模型的 FuSe 导航方法。

Long<sup>[35]</sup> 等人引入了一种新颖的视觉语言导航 (Visual Language Navigation, VLN) 学习框架，使用具有不同能力的大型模型作为领域专家构建了导航代理 DiscussNav，在每一步行动之前会通过涵盖指令理解、环境感知和完成度估计在内的专家讨论，用于纠正无意的错误和筛选不一致的运动决策以有效地促进目标导航。Zhou<sup>[36]</sup> 等人提出了一种纯粹基于大语言模型 (Large Language Models, LLMs) 的指令跟踪导航代理 NavGPT，可以显示地执行包括将指令分解为子目表、整合与导航任务相关的常识知识、从观察到的场景中识别目标、跟踪导航进度以及通过计划调整来适应异常情况这种导航的高级规划。尽管 NavGPT 在 R2R 任务测试过程中的性能仍然低于经过训练的模型，但可以通过调整 LLMs 的多模态输入和显示推理来优化基于学习的模型。后来他们通过拟合 VLN 专业模型和基于 LLM 的导航范式之间的差距，同时保持 LLM 在生成语言和导航推理方面的解释能力，从而整合了一种优化后的导航代理 NavGPT-2<sup>[37]</sup> 来实现有效的动作预测和导航推理。Liu<sup>[38]</sup> 等人提出了一种由对话系统 DRAGON 提供支持的引导机器人，能够将环境与自然语言联系起来。通过理解用户的命令，将用户的自由格式语言与环境相结合，并通过口语为用户提供语义信息，引导用户在地图上找到所需的地标。

2024 年，Yokoyama<sup>[39]</sup> 等人借助人类推理的方式设计了一种视觉语言边界地图 (Vision Language Frontier Maps, VLFM) 来帮助代理在新环境中导航到新环境中看不见的语义目标。该方法根据观测深度图构建栅格占据地图以确定潜在的导航路线，在通过观测图像和预先训练的视觉语言模型生成基于语言的价值图，并据此来确定最优导航路线。该方法在按照路径长度加权完成对象目标导航任务指标中取得了最先进的结果。An<sup>[40]</sup> 等人提出了一种基于拓扑规划导航框架 (Evolving Topological Planning Navigation, ETPNav)，它利用基于 transformer 的跨模态规划器根据拓扑图和指令生成导航计划，再通过试错启发式避障控制器执行该计划。Guo<sup>[41]</sup> 等人借助 ROS 框架提出了一种有效的目标导航策略 EONS，首先对常见的室内物体进行语义关联分析，利用 Mask R-CNN 和残差连接网络建立物体语义关联模型，通过 ROB-SLAM 系统方法构建了一个高可用性的环境图，在移动机器人导航的过程中寻找可达的最优路径以完成导航。Shah<sup>[42, 43]</sup> 等人提出了一种由三个大型预训练模型协同合作构成的 LM-Nav 导航方法，包括将自然语言指令转化为目标序列大语言模型、将目标序列与环境地图中的节点进行特征关联的视

觉语言模型 (VLM) 和负责构建环境拓扑图并利用凸优化技术规划从起点到终点最优路径的视觉导航模型 (VNM)。

### 1.2.3 具身智能

Gupta<sup>[44]</sup> 等人通过构建了一种深度进化强化学习框架模拟实验验证了复杂环境能够促进形态智能进化，证明进化可加速适应性行为的遗传，为具身智能的进化机制与物理形态优化提供了理论支持，推动自适应机器人研发。Cao<sup>[45]</sup> 等人提出了一种高效的强化学习方法 (Causal Action Empowerment, CAE) 用于具身主体，通过识别并利用状态、动作和奖励之间的因果关系来提取可控的状态变量，为高影响行为的优先级重新加权动作，从而提高具身智能体对因果意识行为的执行效率。Fan<sup>[46]</sup> 等人提出了一种具有自主设计、决策和任务执行的具身智能框架，通过视觉语义控制和实时反馈循环增强系统的鲁棒性，验证了 GPT-4 等大语言模型在工业机器人中的潜力。Wen<sup>[47]</sup> 等人通过两阶段训练机制解决数据利用率低的问题，再根据不同的机器人轨迹数据来开发统一的动力学感知模型以增强具身机器人动作的可行性。Lin<sup>[48]</sup> 等人为具身智能系统设计了一种全新的三维感知算法 BIP3D，通过预先训练的 2D 视觉基础模型和空间增强器模块来分别增强语义理解和空间理解，使系统能够实现多视图、多模态特征融合和三维感知功能。目前的具身智能研究大都通过整合视觉、语言、动作等多模态的数据来赋予机器人强大的环境感知和任务理解能力，逐渐从模块化设计转向端到端、从仿真环境到仿真与真实世界协同训练。但该领域依然存在触觉传感器精度低能耗高、硬件成本高、非结构化真实物理交互数据处理难、多模态异构数据难以融合等问题亟待突破。

## 1.3 研究内容

本文主要研究语言视觉激光多模态融合的机器人导航方法的研究及实现，研究方向是多模态融合的目标物体导航方法，并设计一个导航系统，在移动机器人上部署全局和局部路径规划算法，并在仿真环境和真实环境中进行目标物体导航实验。

本文的目标物体导航是在环境中根据多模态的数据到达预期的目标物体。现有的工作通常通过建图来标记环境中存在的目标位置，或是通过训练深度强化学习模型作为代理实时预测动作，以到达指定目标。但上述的方法无法通过视觉信息进行自主探索，忽略了激光雷达所获取的感知信息对于导航的约束和指导，从而导致系统导航成功率低、导航效率低和无法导航到目标半米内的问题。针对以上问题，本文提出一种多模态融合的目标物体导航方法，具体来说，该方法将导航任务拆分成全局路径规划和局部路

径规划两个部分。在全局路径规划中，标记地图中的导航点，保留其位姿、图像、点云图和各点之间的拓扑信息，通过多模态融合网络得到各导航点与目标的匹配权值，结合dijkstra 算法和方位优化算法规划出全局路径导航点序列。然后，在局部路径规划中，将多线激光与单目相机进行联合标定，结合目标检测、点云聚类和坐标变换方法得到目标具体位姿，发布导航任务以完成局部路径的规划，实现导航到目标半米内的闭环任务。Gazebo 数据集上的实验表明，该方法在测试环境中优于最先进的方法，实验结果证明了该方法的有效性和效率。多模态融合的导航系统具有环境感知、目标认知、路径规划和自主导航四个方面的功能，可以完成真实环境下目标物体导航任务。

总的来说，本文的主要贡献如下：

- (1) 本文提出了一种全局路径规划导航方法。与前人的工作相比，针对静态目标导航任务所提出的全局路径规划导航方法基于单目相机、激光雷达等多种传感器和基于多模态特征融合神经网络，增强系统对当前环境和导航过程中的认知和感知能力，再通过方位优化算法筛除噪声导航点，提高导航点选择的正确率的同时提高后续规划的计算响应速度，最后通过导航点规划算法加权融合多种策略进一步提高导航的准确率和导航效率。
- (2) 本文提出了一种未知环境的目标物体探索方法。与前人的工作相比，针对动态目标导航任务所提出的局部目标物体探索方法基于多特征提取和融合的方法，在同一嵌入空间内利用注意力机制融合视觉特征和文本特征，有效的构建了视觉表示和目标物体所在导航方向的关联，使系统能够通过探索找到在变化的环境中的目标物体。
- (3) 本文设计了一套单目相机和多线激光融合的图像点云融合方法，联合视觉观察的认知信息和多线点云的感知信息让移动机器人能够有效地在仿真环境和真实环境中依据自然语言指令完成目标导航任务，在仿真环境和真实机器人上部署并完成一系列可行性与性能测试，实验结果表明该方法具有一定的有效性和优越性。

## 1.4 组织架构

本文的全部内容共有五章，各个章节的内容概括如下：

第一章：绪论。首先简单描述本章的内容；然后描述已知环境下语言视觉激光多模态融合的机器人导航方法的研究背景和意义，即国家的战略支持机器人视觉导航等人工

智能的研发应用，传统的建图定点导航方法和视觉导航方法不适应家居服务机器人的应用场景，基于多模态融合的导航方法应运而生；接着介绍多线激光导航和视觉-语言导航两个应用场景的国内外研究现状；再简单概括本文的研究内容；最后简要概括本文各个章节的内容。

**第二章：相关工作。**主要描述了本文提出的语言视觉激光多模态融合的导航方法中使用的一些理论方法，包括 ROS 系统及传感器间通讯方法、联合标定、目标检测网络、多模态特征融合网络、点云聚类。

**第三章：全局路径规划导航方法。**主要描述了本文所提出的已知环境下的多模态融合导航方法的全局路径规划部分。首先介绍了目标导航的问题定义和全局路径规划方法的整体模型。最后分别介绍了指令提取模块、多模态融合网络模块、方位优化模块和导航点规划算法模块。

**第四章：局部路径规划导航方法。**主要描述了本文所提出的已知环境下的多模态融合导航方法的局部路径规划部分，包括特征提取模块、特征融合模块和运动模块。

**第五章：实验结果与分析。**主要描述我们所提出的结合全局路径规划方法、局部路径规划方法的 LVL-Nav 方法在仿真环境和真实环境下的实验评估结果，包括在仿真环境中不同方法的对比实验来验证本文所提出方法的有效性、针对提出方法的各关键组件进行消融实验分析各个组件的作用和在真实环境中进行目标物体导航的实验。

**总结与展望：**首先对本文所完成的工作及其实现效果进行概述，并针对系统设计中的不足之处提出了未来研究的改进方向与可能的研究思路。

## 第二章 相关工作

本文主要研究语言视觉激光多模态融合的机器人导航方法的研究及实现，研究方向是多模态融合的目标物体导航方法。本章则主要介绍这个研究内容所设计到的技术，其中包括 ROS 系统及传感器间的通讯方法、联合标定、目标检测网络、多模态特征融合网络、点云聚类算法和 PID 控制器。

### 2.1 ROS 系统的通信与导航

ROS(Robot Operating System) 是一个运行于计算机操作系统上层开放源代码的机器人元操作系统，它提供了一整套灵活、模块化的通用框架，帮助工程师、研究人员和教育工作者等开发人员快速构建复杂的机器人系统。在经过开源机器人基金会 (Open Source Robotics Foundation, OSRF) 多年的精心维护和推广后，其社区活跃度、功能包丰富度、高扩展和高开发便利性上都表现十分出色，ROS 已经成为机器人领域最流行的开发框架之一。

ROS 系统主要由节点管理者 (ROS Master)、话题 (Topic)、发布者 (Publisher) 和订阅者 (Subscriber) 这四个部门共同组成，如图2-1所示。其中，节点是 ROS 中的基本计算单

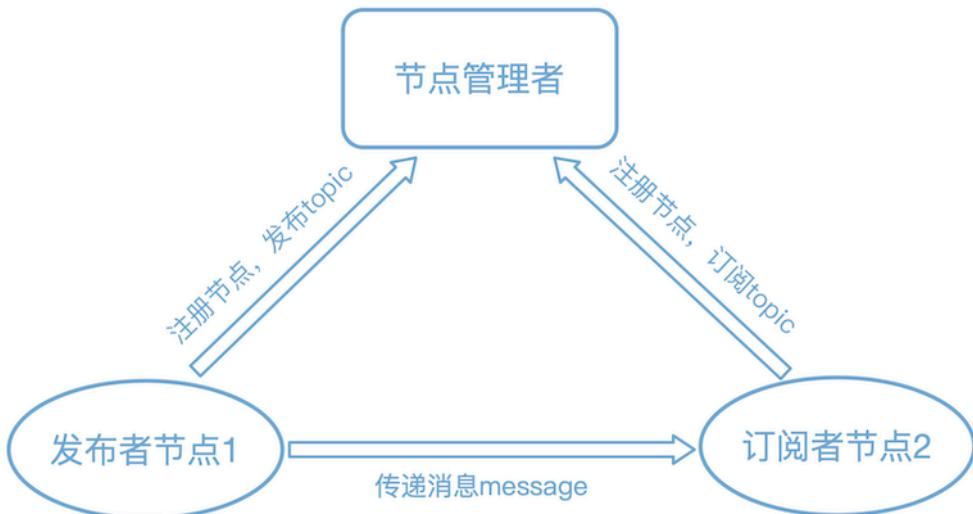


图 2-1 ROS 系统组件图

元，负责执行如传感器数据采集、运动控制或算法处理等特定的任务，而一个完整的机器人系统通常由多个节点组成。话题是 ROS 中实现数据交换的核心机制。节点可以通过节点管理者发布消息到话题，或者通过订阅话题来接受消息，当订阅者想要获取某个特定发布者所发布的消息，只需要保证发布和订阅双方都监听同一个话题即可实现进程

间的通讯。

在用户侧，ROS 提供了一套机器人导航框架如图2-2所示。该框架接收其它组件提供的地图、传感器信息、定位坐标关联、里程计信息，根据用户指定的导航目标点发布机器人执行的速度指令以运动到目标点。具体来说，ROS 使用 SLAM 算法，通过获取传感器数据(如激光雷达或摄像头数据)和机器人运动信息来实时构建环境地图，然后，在创建的二维栅格地图的基础上，通过自适应蒙特卡洛(Adaptive Monte Carlo Localization, AMCL)等定位算法和传感器数据来实时估计机器人的位姿，最后，提供了 move\_base 作为导航框架主体，由局部、全局代价地图、全局规划期、局部规划期、失败恢复行为等组件来执行导航任务。

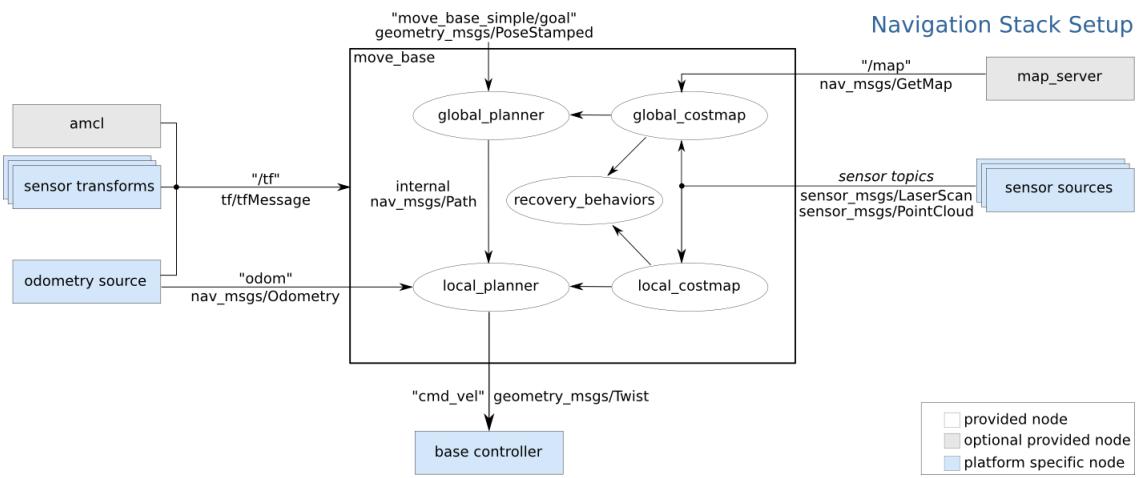


图 2-2 ROS 导航框架图

除了上述的核心功能以外，ROS 还提供了丰富的工具和生态系统，如可以实时显示机器人传感器数据、地图、路径等信息的可视化工具 RViz、允许开发者在虚拟环境中测试机器人算法的高保真物理仿真环境 Gazebo、用于记录机器人运行时的传感器数据工具 ROS Bag，来帮助开发者更高效地开发和调试机器人系统。本文采用 ROS 系统的通信机制、导航方法和仿真环境来实现目标物体导航。

## 2.2 联合标定

在移动机器人上搭载的激光雷达和单目相机处于不同的坐标系之中，导致它们分别获取的点云数据和视觉图像信息无法直接一一对应起来，因此需要将这些传感器进行联合标定，以便于将点云数据映射到视觉图像中进行多模态融合。本文使用张正友标定法<sup>[49]</sup>对激光雷达和单目相机进行联合标定，具体的过程如下图2-3所示。图中存在以单目相机的成像中心为原点的相机坐标系  $\text{Point}_C = [X_C, Y_C, Z_C]^T$ 、以激光雷达为原点的

激光雷达坐标系  $\text{Point}_L = [X_L, Y_L, Z_L]^T$  和以黑白棋盘格标定板的左下角为原点的世界坐标系  $\text{Point}_W = [X_W, Y_W, Z_W]^T$  三个不同的坐标系。联合标定就是要求解出一个用于

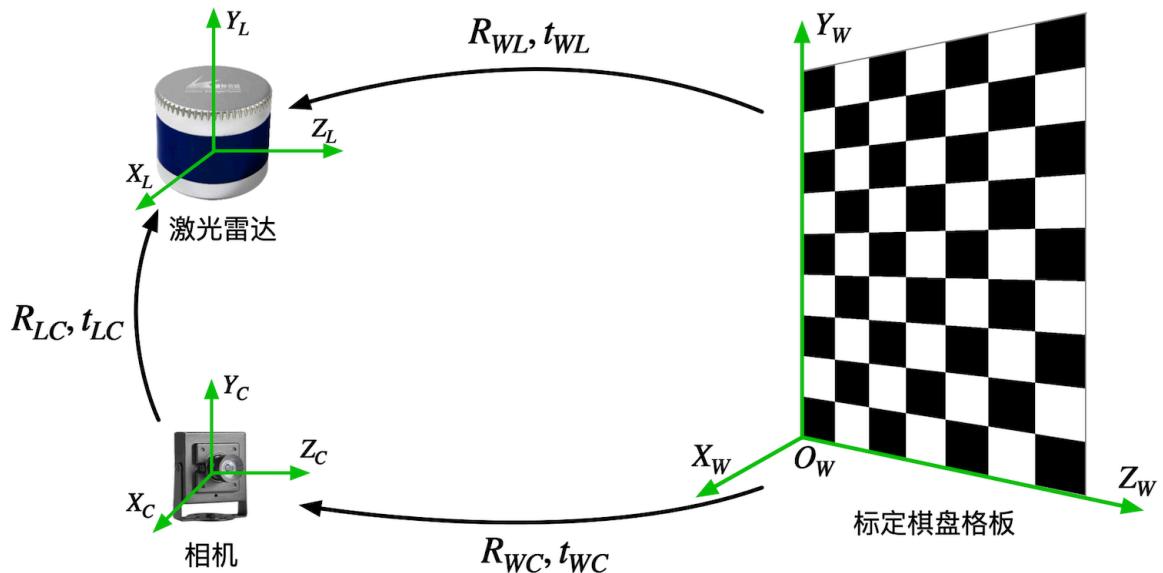


图 2-3 激光雷达和单目相机联合标定

翻转和平移坐标系的旋转矩阵和平移矩阵，即图中的  $R_{LC}$  和  $t_{LC}$ ，使得激光雷达的点云能够映射到图像之中，将雷达感知到的物体与相机认知识别到的物体对应起来。为此需要借助棋盘格标定板来求解它们直接的变换关系矩阵。

首先，相机坐标系中的三维坐标点  $\text{Point}_C$  与像素平面上的二维坐标点  $P_{uv}$  之间的转换关系可以表示为：

$$ZP_{uv} = Z \begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = K \begin{bmatrix} x_C \\ y_C \\ z_C \end{bmatrix} = K \cdot \text{point}_C \quad (2-1)$$

然后，在激光雷达坐标系下空间的任意一点都可以通过旋转矩阵  $R_{LC}$  和平移矩阵  $t_{LC}$  在另一坐标系中进行表示，它们之间的转换关系可以表示为：

$$\text{Point}_C = \begin{bmatrix} x_C \\ y_C \\ z_C \end{bmatrix} = R_{LC} \cdot \begin{bmatrix} x_L \\ y_L \\ z_L \end{bmatrix} + t_{LC} = R_{LC} \cdot \text{Point}_L + t_{LC} \quad (2-2)$$

将上式左项移动到右侧，变形可得：

$$\delta(R_{LC}, t_{LC}, \text{point}_C, \text{point}_L) = R_{LC} \cdot \text{point}_C + t_{LC} - \text{point}_L \quad (2-3)$$

因此，借助棋盘格标定板可以在激光雷达和单目相机两个坐标系间获取多组互相匹配的特征点，进而使得上式中的  $\delta(R_{LC}, t_{LC}, \text{point}_C, \text{point}_L)$  取得最小，即可得到旋转矩阵  $R_{LC}$  和平移矩阵  $t_{LC}$ ，即：

$$(R_{LC}, t_{LC}) = \arg \min_{R_{LC}, t_{LC}} \frac{1}{2} \sum \|\delta(R_{LC}, t_{LC}, \text{point}_C, \text{point}_L)\|^2. \quad (2-4)$$

上述非线性最小二乘方程可以使用 Levenberg-Marquadt 算法<sup>[50]</sup> 将其化成线性方程进行求解。获得的旋转平移矩阵将用于后续第四章的局部路径部分。

## 2.3 目标检测网络

在智能体进行自主导航的过程中，其搭载的算法模型通过视觉处理器对实时采集的视觉图像进行解析，从而获取包括各类物体的分类属性和空间坐标在内的特征参数。这些关键信息用于精准定位感兴趣的目标，并辅助智能体矫正可能出错的动作决策以逼近通向终点的最优路径。

目标检测的任务是从视觉图像之中认知并准确框出特定的目标物体，属于计算机视觉领域中需要解决的基础问题。与图像分类任务不同，目标检测不仅需要识别出图像中存在的物体类别，还需要精确地用锚框 (Anchor) 的形式定位出每个物体的位置信息及其对应的置信度分数。在早期还未出现基于深度学习的检测算法之时，开发者们大多都使用机器学习分类器<sup>[51]</sup> 或是依赖于人工设计的特征提取方法来实现初级的目标检测。但这类传统方法存在特征设计复杂、泛化能力有限等问题，难以应对复杂场景下的目标检测任务。

目标检测技术的发展与深度学习方法的更新迭代紧密相关，在卷积神经网络 (Convolutional Neural Network, CNN)<sup>[52]</sup> 主导的图像特征解析方法于公开数据集的目标检测竞赛中大放异彩后，这类技术方法迅速渗透到包括目标检测领域的各类人工智能算法之中并实现了革命性的突破，也逐渐形成了基于端到端思想的单阶段 (One-stage) 算法和基于区域生成的两阶段 (Two-stage) 算法这两大目标检测技术路线。

前者将目标检测任务视为一个统一的回归问题，一次性直接在图像上进行目标分类和锚框预测而不需要额外的区域生成步骤。因此这类方法凭借这种端到端的结构更适用于实时目标检测的场景，能够有效满足工业级边缘设备对毫秒级响应时间的严苛需求，但在跟踪小物体或密集目标时效果较差。经典的 One-stage 目标检测模型包括：

- (1) YOLO<sup>[53]</sup>: 早期 YOLO 系列的算法采用网格划分策略，它将输入图像分割成等

尺寸的单元，让每个单元同时预测若干预设锚框的空间坐标、类别概率及置信度分数。在生成密集预测结果后，系统再通过非极大值抑制技术筛选最优边界框以去除冗余检测。这类方法的显著特点是检测效率高。得益于其单阶段检测架构，它无需区域建议等预处理步骤即可完成目标识别。然而，YOLO 过于依赖非极大值抑制进行后处理以得到正确的检测框，这阻碍了其的端到端部署，并且对推理延迟产生了不利影响。除此之外，YOLO 的各种组件设计缺乏全面深入的检查，导致明显的计算冗余，限制了模型的能力。与两阶段的目标检测算法相比，在面对小尺度目标或背景复杂的图像时 YOLO 的检测精度存在一定局限。

- (2) SSD<sup>[54]</sup>: SSD 通过预训练深度卷积主干网络在多个层级的特征空间中生成不同分辨率的特征映射，以此实现不同尺度大小物体的定位，但这种方法所使用的先验框参数十分依赖于人工预设，无法在训练过程中实现自适应调整，使得这类方法训练得出得模型泛化能力受限于参数调优策略，并且对小尺寸的目标识别效果仍然较差，存在特征提取不充分的情况。

Two-stage 算法将整个目标检测任务解耦为目标定位与目标识别两个阶段，首先通过区域生成网络 (RPN) 在特征提取层输出潜在得目标区域，接着在各候选目标区域进行类别判定和边界框回归以完成目标检测任务。Two-stage 算法一般在正确率指标上会表现得更好，但其在检测时可能需要占用更多得资源、消耗更多的时间。代表性的 Two-stage 目标检测模型包括：

- (1) RCNN 系列：包括 RCNN<sup>[55]</sup>、fastRCNN<sup>[56]</sup>、FasterRCNN<sup>[57]</sup> 等。这类算法的核心包括以下三点：1) 通过区域生成网络 (Region Proposal Network, RPN) 在主干网络中的特征空间生成目标候选框；2) 针对每个候选区域都会通过一个预训练的卷积神经网络进行特征提取，抑或是对每个候选区域使用 RoI<sup>[58]</sup> 池化进行固定大小的缩放；3) 最后利用交叉熵损失函数与平滑 L1 损失共同优化分类置信度与边界框坐标参数。但这类方法存在检测实时性不足的问题。
- (2) 基于 Transformer：DETR<sup>[59]</sup> 是 Transformer 目标检测算法的开篇之作。它通过引入 Transformer 架构将目标检测过程视为一个由图形到集合的预测问题，消除了如锚框生成和非极大值抑制等后处理过程，通过二分匹配和一个转换器、编码器、解码器的结构和端到端的方式来进行目标的预测和类别的区分。但这类方法的训练时间长，模型的收敛速度慢，且在小物体的检测时性能较差。

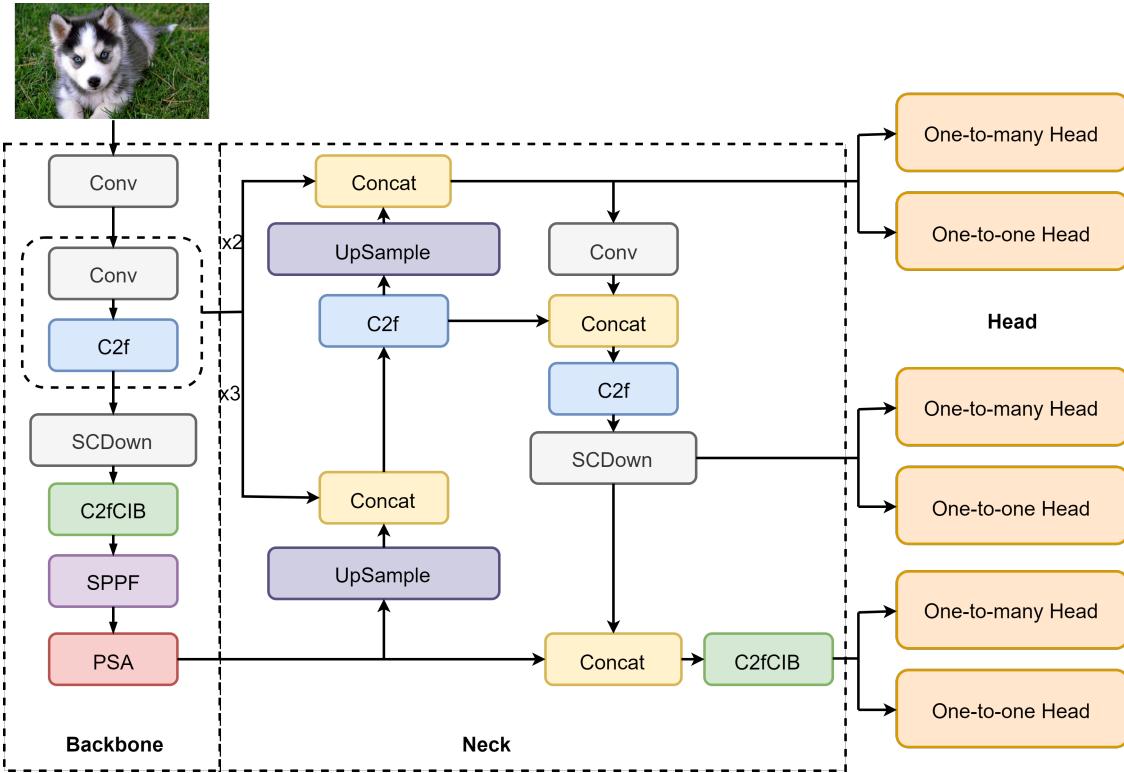


图 2-4 YOLOV10 网络

为了满足整个导航过程中目标检查所需的实时性和可靠性，本文使用的目标检测网络是 YOLOV10 网络<sup>[60]</sup>，如图2-4所示。YOLOV10 与 YOLOV8<sup>[61]</sup> 相比在整体的网络结构上基本保持一致，网络分为骨干网络 (Backbone)、颈部 (Neck)、头部 (Head) 三个部分。首先将图片输入骨干网络中提取图像的全局和局部特征，在颈部通过 Upsampling、Concat 和注意力机制卷积网络增强特征的表达能力，实现有效的特征提取和融合，并在头部将颈部提取的特征映射到最终的输出空间，生成网络的最终预测结果。然而，相较于后者，YOLOV10 为了实现更加轻量化的端到端部署而做出了几点重要的优化：

- (1) 轻量化分类头 (Lightweight Classification Head): 在 YOLOV8 网络结构中 Head 部分的分类头的参数量和计算量比回归头更大，但后者对检测精度的影响更大，因此减少分类头的卷积参数量以达到轻量化模型的目的。
- (2) 空间-通道分离下采样 (Spatial-Channel decoupled downsampling, SCDown): YOLOV8 使用一个标准卷积时实现空间下采样和通道变换，SCDown 将这两种操作进行解耦，先通过逐点卷积调节通道维度，然后通过深度卷积进行空间下采样，保证降低计算成本的同时最大限度保留信息。
- (3) 精度驱动的模型设计: 在小模型规模的深层阶段使用大核卷积 (Large-kernel Conv) 来扩大感受野，增强模型能力；针对计算开销过大的自注意力机制设计

了一种高效的部分自注意力 (Partial self-attention, PSA)，对分辨率最低的特征的一半进行计算，将对于全局的学习能力以较小的计算成本融入到网络中。通过这些方法可以在不显著增加计算成本的情况下提升模型的性能。

- (4) 基于秩的块设计：YOLOV10 使用了一种分层特征优化的紧凑反转模块 (CIB) 结构，它通过低复杂度的深度卷积方法来实现局部空间的特征融合，同时结合轻量化的点卷积操作完成跨通道信息交互，这一设计有效缓解了传统检测模型中由同构模块堆叠所引发的参数冗余问题。

## 2.4 多模态特征融合网络

多模态特征融合是指通过整合多种来源或不同形式的数据信息来训练一种能够提升系统认知能力或执行效果的技术，这种技术旨在充分发挥各模态间的优势互补特性从而提升系统在模式辨识、类别划分和内容生成等任务中的表现能力。

在机器人执行导航任务的探索阶段将采用特征融合框架对异构感知信息进行融合，通过语义编码器提取的文本描述特征、目标物体识别算法提取的语义局部信息、深度卷积模型提取的环境特征来得到可以指导机器人进行局部导航的动作决策。现阶段的多模态融合方法按照融合的阶段的不同可以分为以下三种类型：

- (1) 特征级融合：这种融合方法是在神经网络的核心处理模块之前，通常是在数据输入环节，将多种模态的特征进行整合。比如，在数据输入阶段就将视觉信息和文本信息进行融合。
- (2) 模型级融合：这种融合方式选择在神经网络的中间层级进行多模态信息的整合。具体做法可以是将各模态经过独立学习后的特征表示进行合并，然后再进行后续的网络处理。
- (3) 决策级融合：该策略在完成各模态的独立处理后在决策或输出层进行融合。在每个模态的数据都经过单独处理后，通过将各模态的输出结果进行融合以形成最终决策。这种方法的优势在于其灵活性高且能兼容经过预训练的单模态模型。

本文将采用第三种融合方法，即使用 Transformer 网络进行后期的决策级融合，Transformer 网络的操作流程如图2-5所示。Transformer 网络的核心创新在于采用了多头自注意力机制<sup>[62]</sup>。自注意力的核心原理是通过计算序列中各元素间的关联度来生成相应的注意力权重，再基于这些权重对序列元素进行加权整合从而达到自注意的目的。多头机制则通过并行使用多个注意力单元使每个单元能够学习到不同的权重分布，从而让模

型能够在多个特征子空间中对输入序列进行多样化表征。这种设计方法使 Transformer 能够突破传统模型的局部窗口限制来实现对序列全局信息的有效捕捉。多个注意力单元可以分别聚焦于不同类型的信息特征，这样可以显著增强网络在处理多模态数据时的特征融合能力。具体来说：

$$\begin{aligned} \text{Attention}(Q, K, V) &= \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \\ \text{head}_i &= \text{Attention}\left(QW_i^Q, KW_i^K, VW_i^V\right) \\ \text{MultiHead}(Q, K, V) &= \text{Concat}(\text{head}_1, \dots, \text{head}_n)W^O \end{aligned} \quad (2-5)$$

其中，查询 ( $Q$ )、键 ( $K$ ) 和值 ( $V$ ) 由输入序列通过三个线性变换获得的矩阵， $W_i^Q$ 、 $W_i^K$  和  $W_i^V$  分别是  $Q$ 、 $K$  和  $V$  的权重矩阵， $d_k$  表示特征维度， $\text{head}_i$  表示第  $i$  个注意力头， $W^O$  表示连接多个注意力头输出的权重矩阵  $\text{Attention}$  表示子注意力机制， $\text{MultiHead}$  是多头自注意力机制。

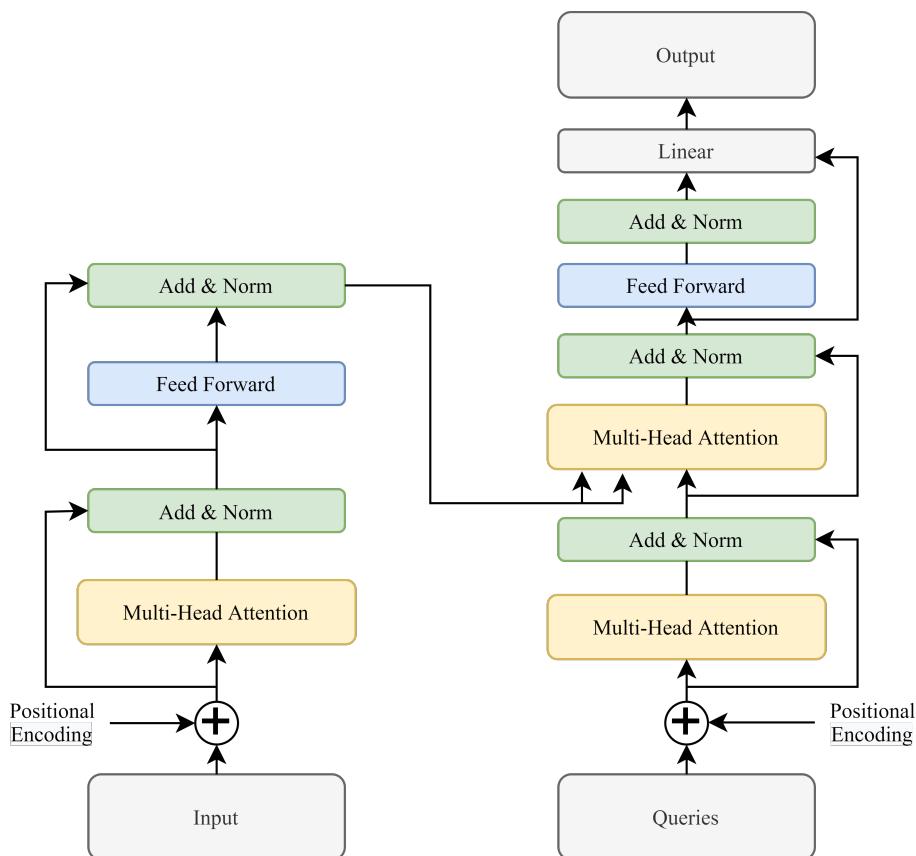


图 2-5 Transformer 网络结构

## 2.5 点云聚类

三维激光雷达获取的点云具有精度高、空间坐标信息准确的特点，因此在移动机器人的导航过程中常用于实时定位与环境感知，但点云数据是一种非结构化的离散数据，在环境感知的过程中需要通过高效的聚类算法来区分从属于不同目标的点云。目前主流的聚类方法一般会使用基于局部表面曲率、法向量方向这类几何属性与密度梯度、高程变化这类空间分布特征进行特征编码，便于将点云映射至可聚类特征域，欧式聚类就是这类算法的典型例子，它通过 KD-Tree 加速邻域搜索实现  $O(n * \log n)$  级时间复杂度的快速点云分割，但也需要根据雷达点云近密远稀的特性对距离阈值进行动态调整。针对不同的应用场景还存在密度聚类、超体聚类等算法，它们在耗时和聚类准确率方面各有优势。在目标物体导航的过程中要求系统需要更强的实时性，因此本文采用欧式聚类<sup>[63]</sup>方法对三维点云进行预处理。

欧式聚类法在点云密集的情况下需要进行额外的优化以保证其实时性，这里采用点云栅格化和 kd 树对算法进行加速优化<sup>[64]</sup>。栅格化方法首先将扫描区域划分为若干网格单元，将三维点云投影至二维平面，保持 z 轴数值不变的同时，将每个网格内点的 x、y 坐标统一为该网格中心点坐标。随后进行去重处理，对于 x、y 坐标相同且 z 值相近的点只保留一个代表性点，通过栅格化方法对点云数据进行处理能够显著减少计算复杂度。除此之外，本文利用 PCL 库中的 KdTree->setInputCloud() 函数将栅格化后的点云构建为 k 维二叉树结构，借助 kd 树结构优化近邻搜索过程，进一步提升欧式聚类算法的效率，本文改进的欧式聚类算法的伪代码见 Algorithm1。

距离阈值是用来区分不同簇点云的重要参数，当设定的距离阈值  $d_{th}$  小于一个点集  $P_m = \{p_m \in P\}$  与另一个点集  $P_n = \{p_n \in P\}$  之间的最小距离，则可以判定这两个点集为两簇不同的点集。因此点云簇为两簇不同的点云的条件可以表示为

$$\min \|p_m - p_n\| \geq d_{th} \quad (2-6)$$

近密远稀是激光雷达这类设备的共同特点，表现为空间中点云的密度会随着其与原点的距离的减小而逐渐增加、增加而逐渐减小，如图2-6所示，随着车辆与激光雷达之间的距离增大，其反射生成的点云分布密度显著降低。传统的欧式聚类算法依赖于固定的距离阈值来划分障碍物，这种方法难以应对点云密度随距离变化的情况。本文通过调整距离阈值参数设计了一种自适应的欧式聚类方法来解决空间点云发散而导致聚类效果

---

**Algorithm 1** 可变距离阈值的欧式聚类算法

---

**Input:** 激光点云  $P$ **Output:** 点云簇集合  $C$ 

```

1: 对激光点云  $P$  进行栅格化和 kd 树预处理;
2: 创建点云簇列表  $C$ 
3: 创建当前点云簇  $c$ 
4: for  $p_i \in P$  do
5:   que.push( $p_i$ )                                 $\triangleright$  遍历点云  $P$  中的每个点，并将当前点  $p_i$  加入队列
6:   while !que.empty() do
7:     ThresholdGet( $p_i$ )  $\rightarrow d_{th}$                  $\triangleright$  计算当前点对应的距离阈值
8:     KdtreeSearch( $P, p_i, d_{th}$ )  $\rightarrow P_i^k$            $\triangleright$  使用 kd 树寻找  $p_i$  的邻近点
9:     for  $p_j \in P_i^k$  do
10:    que.push( $p_j$ )                                $\triangleright$  遍历  $p_i$  的临近点，并将其加入队列中
11:   end for
12:    $c = c \cup p_i$                              $\triangleright$  将找到的这簇点云加入结果集中
13:   que.pop()
14: end while  $c = \emptyset$ 
15: end for

```

---

不好的问题，如(2-7)所示，距离阈值参数  $d_{th}$  会随着点云与原点距离的增大而逐渐增大。

$$d_{th} = \begin{cases} 5cm & 0 < Range \leq 1.5m \\ 10cm & 1.5m < Range \leq 3.0m \\ 15cm & 3.0 < Range \leq 5m \\ 20cm & 5m < Range \end{cases} \quad (2-7)$$

本文使用的十六线激光雷达在不同距离的聚类测试中表现出明显的波动，当雷达与目标之间的距离超过 5 米后 z 轴方向的点云密度衰减十分严重，聚类的可靠性明显下降，因此我们在自适应方法中添加了终止机制，即  $d_{th}$  在 5m 后的范围不再迭代更新。

点云聚类算法的结果用于后文第四章所介绍的点云映射中，将点云感知信息与图像认知信息的结果相融合，以获得目标检测算法认知到的目标物体的精确位置信息。

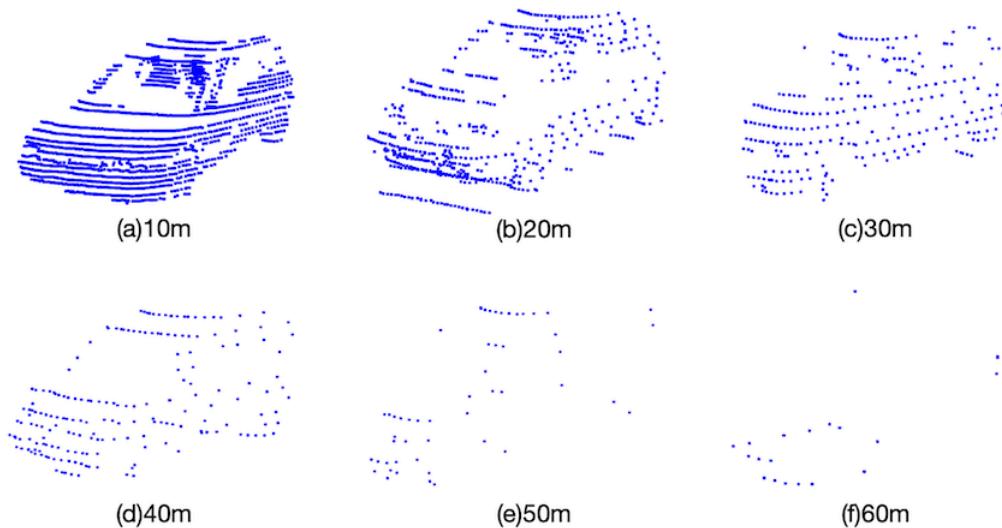


图 2-6 障碍物点云随着距离增加变稀疏

## 2.6 PID 控制器

PID 控制器是一种广泛使用的反馈控制系统，能够根据目标值与当前值之间的偏差来调整控制变量，从而使系统稳定地达到设定目标。具体而言，PID 控制器通过调节比例、积分和微分来控制机器人的运动，确保其到达预定的位置或保持稳定的姿态，如图2-7，它简单易用，适用于大多数线性控制问题，具有很好的实时性和鲁棒性。

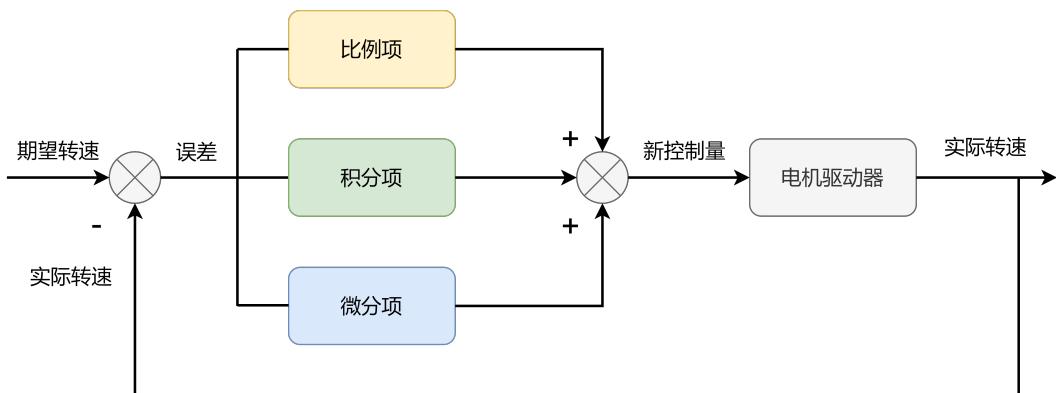


图 2-7 PID 算法控制电机转速

比例项是 PID 控制器中最基础的组成部分，它直接根据期望值与实际值之间的偏差来进行调整，如式2-8，其中  $P$  表示比例控制输出， $K_p$  表示比例增益，即在误差变化时控制器输出的变化幅度， $e(t)$  表示当前时刻的误差，当误差越大时，比例项的输出也

就越大，以帮助系统更快速的接近目标值。

$$P = K_p \cdot e(t) \quad (2-8)$$

积分项是 PID 控制器中的第二个组成部分。积分项的目的是消除系统在长时间运行中可能出现的静态误差，确保系统达到并保持在期望的目标位置，如式2-9，其中  $I$  表示积分控制输出， $K_i$  表示积分增益，能够决定积分项对控制输出的影响程度， $e(t)$  表示误差。当系统的输出长期无法达到设定的目标时，积分项就会逐渐地累积来弥补系统长期积累的偏差，一直到该误差消除。

$$I = K_i \cdot \int_0^t e(t) dt \quad (2-9)$$

当积分增益过大时会导致系统响应过度，引发系统的超调和震荡，这时候就需要通过微分项来预测误差的变化趋势以减少系统发生超调和震荡的可能。与前面介绍的两者不同，微分项主要关注的是误差变化的速度而非当前的误差值，如式2-10，其中  $D$  表示微分控制输出， $K_d$  表示决定了微分项对控制输出的影响程度的微分增益， $\frac{d}{dt}e(t)$  表示用来预测未来误差变化趋势的误差变化速率。通过微分项的预测误差和提前调整，系统能够避免震荡和过调现象的发生，让系统更加平稳的响应。

$$D = K_d \cdot \frac{d}{dt}e(t) \quad (2-10)$$

通常情况下积分项、比例项和微分项会一起使用，即通过比例项消除实时误差，利用积分项来消除长期累计的误差，而微分项则通过对误差变化率的敏感度调整来预测误差的变化趋势，减少系统发生超调和震荡的概率，三者相辅相成一同构成一个完整的 PID 控制器，使其能够更精准、更平稳地调节系统响应。

PID 控制器将用于后文第四章所介绍的运动模块中，以执行特征提取、融合模块所输出的导航动作。

## 2.7 本章小结

本章主要内容是目标物体导航过程中涉及到的一些常用的算法原理和关键技术，主要介绍了 ROS 系统及传感器间的通讯方法、联合标定、目标检测网络、多模态特征融合网络、点云聚类算法和 PID 控制器。

### 第三章 全局路径规划导航方法

为实现移动机器人既能根据自然语言指令中的目标进行导航，又能完成导航至目标半米内的任务，本文提出了一种语言视觉激光多模态融合的机器人导航 (Language Vision Lidar Navigation, LVL-Nav) 方法。为了能够完成上述任务，我们提出的 LVL-Nav 将导航任务拆分为已知环境中下的全局路径规划导航和未知环境下的局部路径规划两个部分。本章将详细介绍 LVL-Nav 方法中的全局路径规划导航方法。主要有目标物体导航任务定义，然后描述了所提出的全局路径规划导航方法的框架设计，最后分别描述了框架中的指令语义提取模块、多模态融合网络模块、方位优化算法和导航点规划算法四个部分。其中指令语义提取模块将非结构化的语言指令通过大语言模型映射到具体的目标空间和方位空间，前者作为多模态融合网络模块中的目标特征与图像编码器和深度图编码器所提取环境的图像特征、深度特征拼接得到的环境特征进行跨模态融合，帮助智能体理解多种模态之间的关系；后者则通过方位优化算法辅助筛选掉冗余的导航点，最后再通过导航点规划在最大化目标与导航点的匹配程度的同时最大化导航成功率。

#### 3.1 导航任务定义

多模态融合的目标物体导航任务要求代理从室内环境中的随机起始位置根据自然语言指令导航到用户依次指定的物体类别，如打印机、餐桌、储物柜等，并且代理被允许使用以第一视角的视觉观察 (通常为 RGB 图像) 和激光雷达，其问题可以定义如下：给定自然语言指令  $I$  和智能体在环境中的初始位姿  $S$ ，通过解析指令要求，在建好的栅格地图中的导航点集中依次推理选择与指令目标序列  $\{l_i\}$  相匹配的导航点序列  $\{v_i\}$ ，并最终在目标位置的半米内停下以完成导航任务。

一次完整的目标物体导航过程如图3-1所示，代理在  $t=0$  的初始时刻被初始化在房间右上角餐桌附近的  $S$  点，同时能获得单目相机和激光雷达第一视角的 RGB 图像和点云数据，以及一条导航指令 “Go all the way to the refrigerator, then turn around to the sofa, straight forward past the bed, and stop at the edge of the chair.”。然后代理通过全局路径规划生成任务执行序列，随后在构建的拓扑路径网络中进行节点间的移动，需要在餐桌附近依次导航到冰箱、沙发、寝室的床和床边的座椅，在经过所有子目标之后在最终目标的位置附近停下。其中，蓝色箭头依次所指的目标是智能体所推理出的所有目标节点。

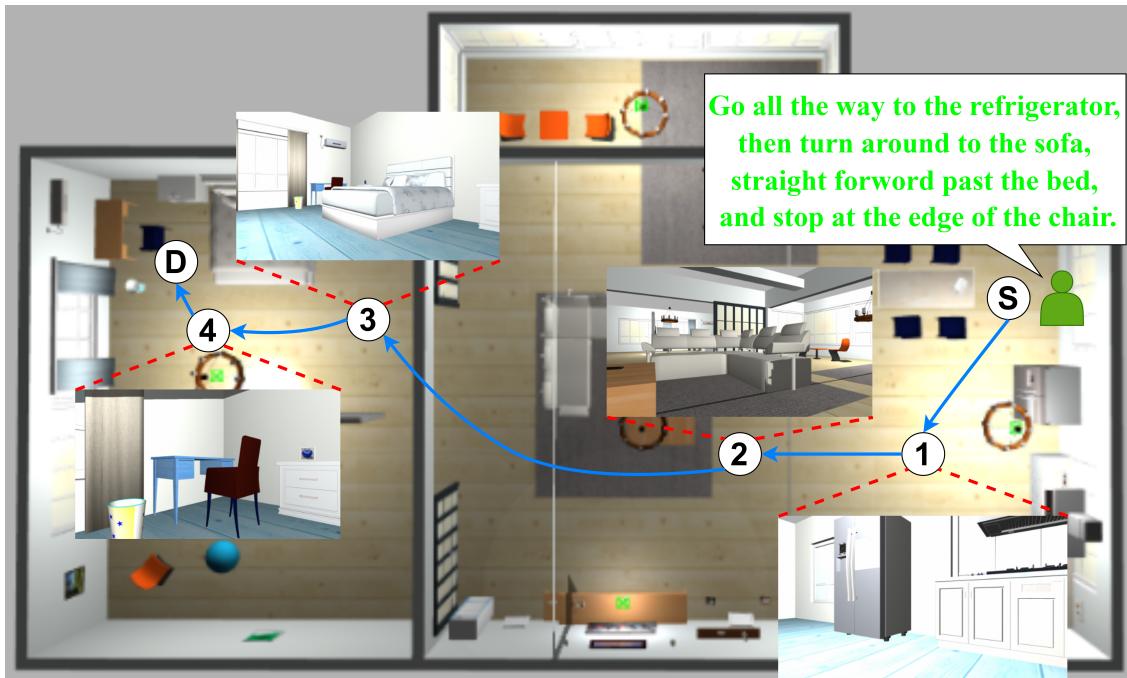


图 3-1 目标物体导航过程

## 3.2 导航框架设计

目标物体导航是机器人导航中的一个重要子任务和研究方向，其主要目的是通过代理对环境的感知和认知信息依次到达指令给定的目标物体。现有的目标物体导航方法中，Ke<sup>[65]</sup> 等人提出了一种 FAST 模型，使用贪婪搜索算法将目标对象描述的编码与动作空间的编码相结合，以导航到下一个目标点。Hong<sup>[66]</sup> 等人提出了一种新的语言和视觉实体关系图，通过在语言元素与视觉实体之间传递信息，实现了不同模态之间的高效协同，并将其用于捕捉并模拟文本与视觉之间以及视觉实体之间的相互关系。Wang<sup>[67]</sup> 等人针对代理无法实时捕捉环境布局和进行长期规划这一限制，提出了一种结构化场景记忆 (Structured Scene Memory, SSM) 导航框架，它可以自适应地捕获和表征环境中的视觉和几何特征以支持当前决策，并模拟迭代算法进行远程推理以实现高效的全局规划，他们的算法在 R2R 和 R4R 数据集上的实验结果取得了最优秀的性能。Pashevich 等人针对代理无法准确地处理长序列的子任务和理解复杂的人类指令，提出了一种多模态转换器 (Episodic Transformer, E.T.)，利用合成指令作为中间表示，将理解环境的视觉特征与复杂多样的自然语言指令解耦。Shah<sup>[42]</sup> 等人提出的视觉语言导航方法使机器人能够根据目标图像和环境拓扑图进行导航。Huang<sup>[68]</sup> 等人创新性地提出了一种基于空间映射表示的方法 VLMaps，它直接将预训练的视觉语言特征与物理世界的 3D 重建融合在一起，通过大语言模型可以将自然语言命令转换为一系列可以定位在地图中的开放词汇导航

目标，并且允许在多个机器人之间共享以动态生成新的障碍地图。王滔<sup>[69]</sup>等人通过场景语义分割和三维重建的方法构建语义地图，再利用语言模型理解并执行指令。

这些目标物体导航方法如图3-2所示，他们大都强调在环境中提取物体语义、位置、物体关系等丰富的视觉特征表示，并将其与栅格地图中的导航点相关联，以此来告诉代理应该导航到哪个导航点，但这种做法过于关注环境中存在什么物体，缺失了已知环境的拓扑信息，没有利用语义中的方位信息进行导航点的筛选，且视觉语言模型在光线条件变化大的室内的鲁棒性较低，从而导致了导航成功率和导航效率低的问题。

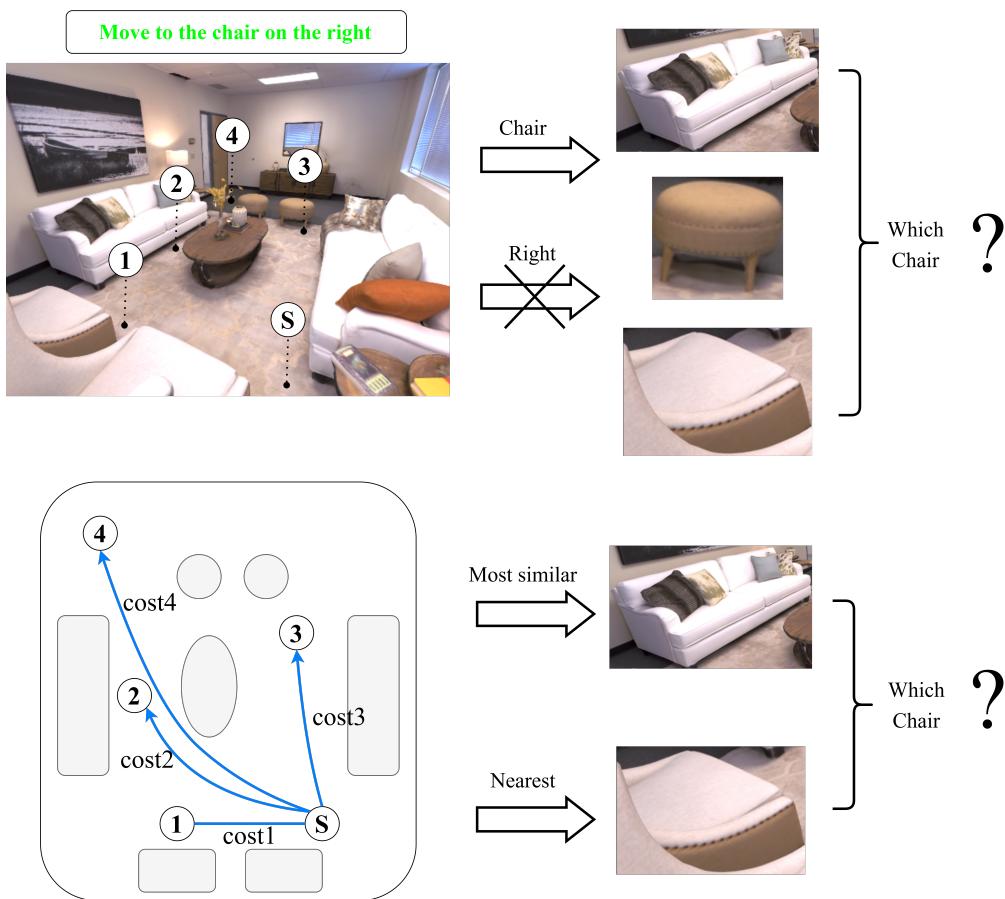


图 3-2 之前工作存在的问题

因此，如何在进行图像观察中的物体语义、位置、物体关系等目标物体视觉表示和栅格地图中所标记导航点的匹配过程中，利用环境中不同导航节点和机器人当前位姿所构成的拓扑信息来辅助代理进行更精确地匹配，优化传统的视觉语言模型以克服其在曝光或欠曝环境下模型性能差的问题，再通过从语言指令中提取的方位信息进行方位优化以筛选掉不在目标方向上的冗余导航点，从而提高指导代理依次导航到目标点的导航成功率和导航效率。

在全局路径规划中我们的导航框架如图3-3所示。全局路径规划主要由四个关键模

块算法构成。首先，输入的导航指令在经过大语言模型处理之后得到目标序列和动作序列，前者与环境中存在的多个可能的导航点图像一同作为多模态融合网络的输入得到导航点匹配特征，后者在经过方位优化算法后筛选掉冗余的导航点，再基于栅格地图和初始位姿计算获得环境的拓扑图以补足环境拓扑特征，最后通过导航点规划算法综合考虑多种导航策略以获得最终的导航点序列，依次发布导航点以完成导航任务。

我们的方法考虑了之前工作中不足的地方并加以改进。具体来说，视觉图像导航方法的效果在很大程度上取决于多模态融合网络的性能，多模态融合网络通过引入深度信息，优化传统视觉语言模型，提高模型在光线条件变化大的室内的鲁棒性，此外，本文针对室内环境存在多种相同目标的情况，提出了一种方位优化方法，结合指令中提供的方位信息与移动机器人的实时位姿，筛选不符合方位条件的冗余导航点，进一步提高全局路径规划中导航点与目标匹配的正确率，最后通过环构建的环境拓扑图结合上述两个算法结果进行导航点规划算法，最终获得导航点序列以执行全局路径导航。

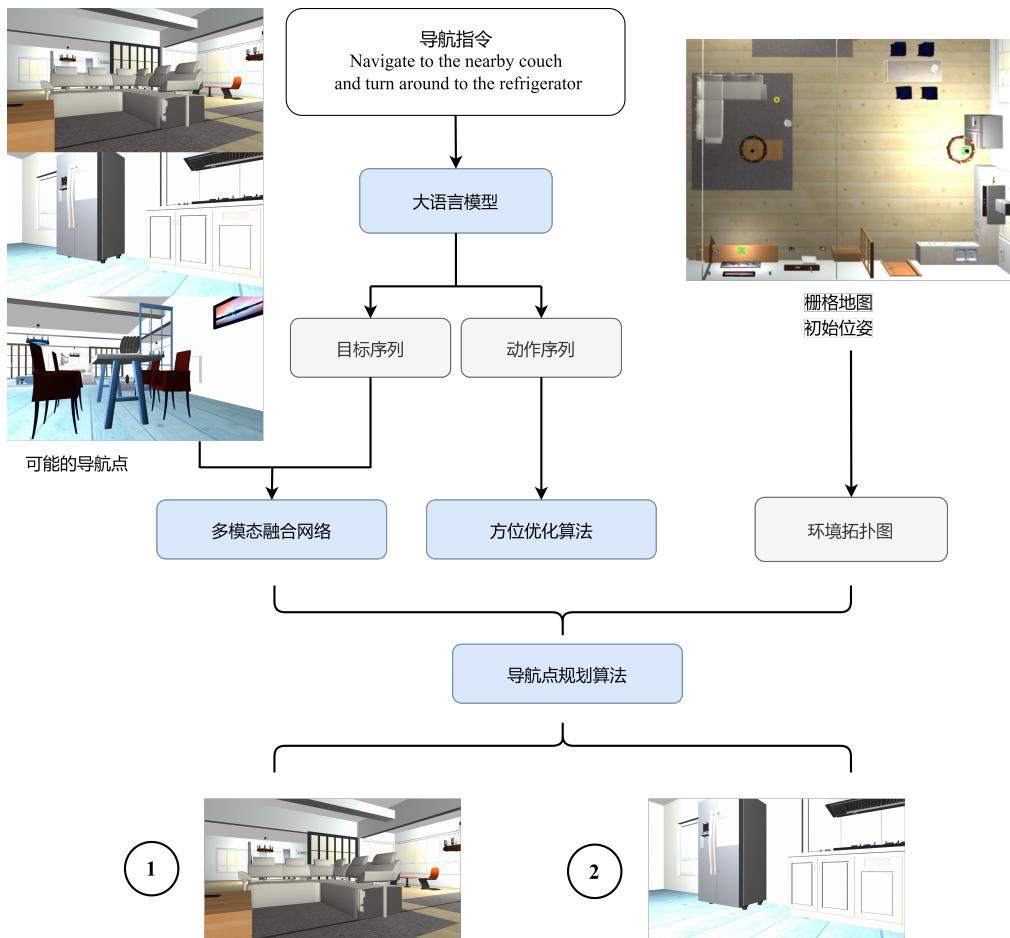


图 3-3 全局路径规划导航框架

### 3.3 指令语义提取模块

房间到房间（Room-to-Room, R2R）是一个用于验证视觉语言导航（VLN）模型的基准数据集<sup>[70]</sup>。该数据集的环境信息通过拓扑导航网络进行表示，如图3-4所示，其中每个顶点代表移动机器人在环境中的可达点，同时该点还会对应一个在该位置获取的360度全景视觉观测，连接于顶点之间的边则代表可达性关系。目标物体导航任务要求智能体根据输入的自然语言指令生成正确的目标点导航序列。举例来说在给出指令“Turn left and exit the room. Keep walking along the hall past the kitchen area. Wait by the doorway to the dining table area.”后，智能体根据对指令的理解从起始点开始执行导航任务，它需要根据推理的结果在导航图的节点间进行移动并最终到达目标位置。

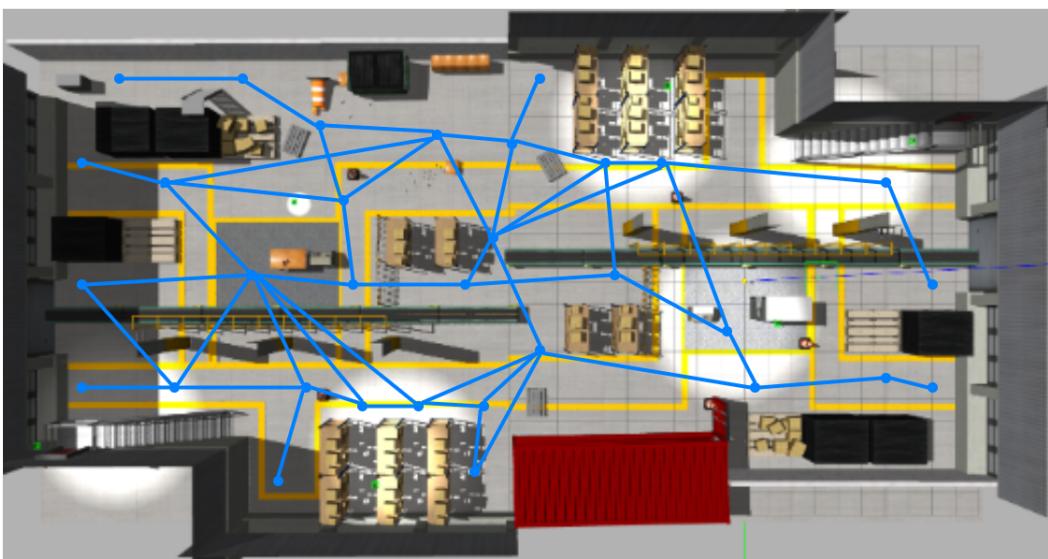


图 3-4 导航图示例

视觉语言导航技术使人类能够使用自然语言与机器人进行交互，使机器人能够基于对环境的感知和对语言指令的理解来执行相应的导航任务。但目前大多数关于视觉语言导航的研究仍然面临诸多挑战，其中最突出的问题之一是这类方法难以适应人类语言的复杂性和多样性，因为许多现有方法往往忽视或无法全面提取指令中所蕴含的例如目标位置和方位描述这类关键信息，这在很大程度上限制了该技术在实际智能制造环境中的广泛应用。通过统计 R2R 数据集中存在的部分指令的得到单词分布图3-5。就分布结果表明我们可以通过利用大语言模型来更准确地解析指令中的目标和方位信息，基于提取出的方位信息可以使智能体更有效地过滤掉复杂室内环境中由于方向不同而产生的大量错误目标，从而显著提升导航任务的执行精度和鲁棒性。这一能力对于提高视觉语言导航系统的实际应用价值至关重要。



图 3-5 R2R 指令集中的方位指示分布

本文提出并采用了一种专门设计的任务提示模板来解析指令中的目标和方位信息，该模板能够高效且准确地引导语言模型将各种不同形式的导航指令解析并转换为导航任务执行所需的具体目标和方位序列。通过这种方法我们能够为多模态融合网络以及方位优化算法提供更加清晰且具体的任务语义信息，从而提升导航系统的理解能力和执行效率。在实验的过程中我们通过输入任意形式的导航指令，选用了 Noun Chunks、GPT3.5、GPT4.0 和深度求索 (DeepSeek) 四种不同的语言模型进行自然语言导航指令解析测试，测试结果如表3-1所示。实验结果表明在我们设计的导航指令提示任务中，

表 3-1 不同的语言模型测试结果

Noun Chunks	GPT3.5	GPT4.0	DeepSeek
0.80	0.88	0.96	0.96

Noun Chunks 和 GPT3.5 对复杂句式的处理能力较差，在面对简写、俚语等语言变体导航指令时输出的目标序列和方位序列存在错误、乱序的情况，这样的解析结果会影响系统整体导航的成功率。除此之外，虽然 GPT4.0 在我们设计的任务中取得了较好的结果，

但是移动机器人在接收指令后的很长一段时间都在等待模型响应，其缓慢的响应速度会降低系统整体的导航效率。而 DeepSeek 则通过垂直领域预训练、知识增强和验证机制的技术路线，在保证它本身能准确地完成特定的语言推理任务的同时还能快速高效地进行响应，因此本文选用 DeepSeek<sup>[71]</sup> 作为核心语言模型。

为了提高语言模型提取结果的准确性，我们通过详细描述任务内容同时提供正确的目标与方位提取示例来引导模型学习合理的输出模式，使其生成的结果尽可能符合预期要求。具体来说在输入一条随意表达的导航指令后，该模型会自动分析并提取句中所包含的目标与方位信息，并将其转换为标准化的数据格式以方便后续导航任务的执行。模型的提取示例结果如图3-6所示，在我们的实验测试中该模型对不同风格的导航指令均表现出较高的适应能力和准确性，这进一步证明了该方法在导航任务中的有效性。

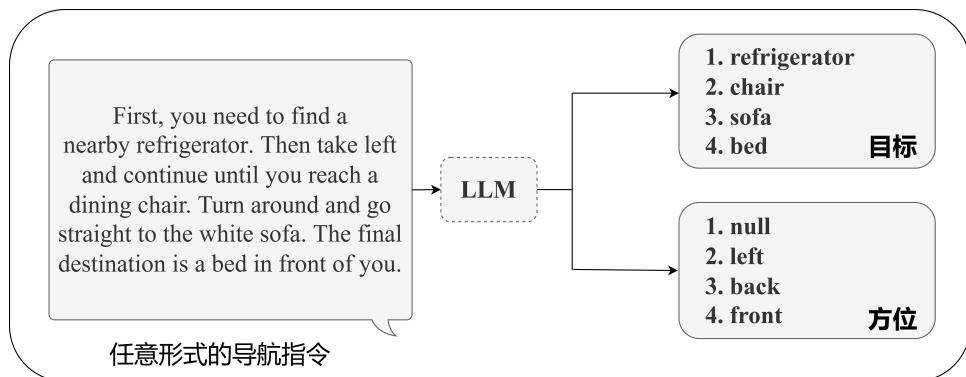


图 3-6 指令语义提取模块

### 3.4 多模态融合网络模块

CLIP (Contrastive Language-Image Pre-training) 多模态融合网络<sup>[72]</sup> 是一种基于对比学习的多模态融合模型，它通过将图像和文本映射到同一个嵌入空间中使匹配的图像-文本对的特征向量尽可能接近，不匹配的图像-文本对的特征向量尽可能远离。这种学习策略使得 CLIP 在没有特定任务训练数据的情况下依然能够有效地执行分类和搜索任务，使得该模型展现出极高的泛化能力和适应性。因此 CLIP 依靠其具有的极大灵活性和通用性在众多计算机视觉和自然语言处理任务中展现出优越的性能，同时也在多模态建模领域得到了广泛关注。

在利用视觉语言进行导航的任务中，一个核心挑战是如何高效且准确地将语义目标与环境图像进行匹配，并在此基础上规划出一条符合指令描述的全局导航点序列，以确保导航智能体能够按照预期完成路径规划。InteriorNet<sup>[73]</sup> 是由帝国理工学院和 KooLab

联合创建的大规模、多传感器、照片级真实感室内场景数据集，该数据集包含约 100 万个来源于真实的生产制造的家具 CAD 模型和 2200 万个室内布局。为了研究 CLIP 模型在不同光照环境下的表现，我们随机选取了该数据集中的若干图像并通过计算直方图将其划分为 4 个曝光程度不同的子数据集。这 4 个子数据集分别对应图像过度欠爆、轻微欠爆、正常曝光、明显过曝的情况。随后使用 CLIP 模型在这些数据集上进行了图像-文本匹配测试以此评估模型在不同光照条件下匹配的正确率。实验结果如表3-2所示，这表明在光照条件良好的室内环境下 CLIP 能够稳定且可靠地完成图像与文本的匹配任务，但当图像出现曝光过度或欠曝的情况时，CLIP 的图像-文本匹配正确率却明显下降并表现出较大的不稳定性。这一实验表明尽管 CLIP 在标准光照条件下能够提供较为理想的匹配效果，但其在室内的复杂光照环境下仍然存在一定的局限性。

表 3-2 模型进行图像文本匹配的正确率

色阶范围	[0,60]	[61,120]	[121,180]	[181,255]
CLIP	32.43%	<b>99.12%</b>	89.38%	37.72%
CLIDP	<b>44.95%</b>	98.23%	<b>92.04%</b>	<b>42.28%</b>

为了解决现有视觉语言导航系统在不同光照条件下匹配能力不稳定的问题，本文针对多模态融合网络的特征提取能力进行了改进，引入了一种深度图像编码器以提取环境中的深度信息从而扩展了模型所能捕捉的特征维度。这种改进能够确保模型在曝光过度或欠曝的复杂光照条件下，仍然可以依赖深度特征信息进行准确的图像-文本匹配，而不仅依赖于颜色或纹理等受光照影响较大的视觉特征。基于这一思路，我们提出了一种新的多模态融合框架 CLIDP (Contrastive Language-Image-Depth Pre-training)，其整体架构如图3-7所示。

CLIDP 多模态融合网络在预训练模型 ViT-B-32<sup>[72]</sup> 的基础上，通过引入的深度信息进行了进一步的优化和重训练，其意在解决的关键任务是在给定的 ((图像, 深度图), 文本) 数据对中，预测其中存在的所有可能配对的概率。这一过程通过训练一个图像编码器和一个深度图编码器来分别提取环境的 RGB 图像和深度特征信息，并将这两个部分的特征拼接，以形成完整的环境特征表示。

首先，图像编码器使用与 CLIP 类似的 ViT(Vision Transformer) 结构，如图3-8所示，图像切分与嵌入 (Patch Embedding) 操作会将输入的图像划分为固定大小的非重叠小

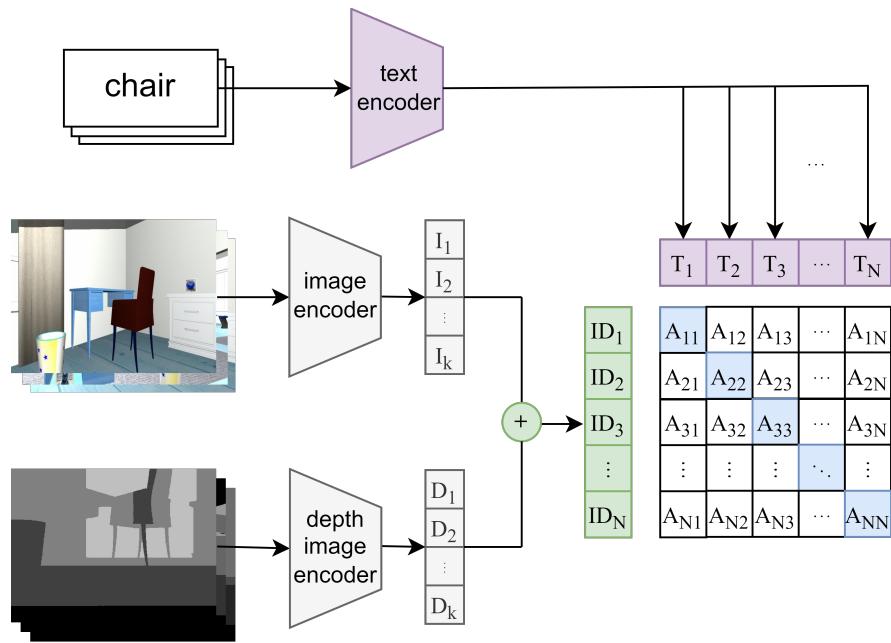


图 3-7 CLIPD 多模态融合网络框架

块，并将小块通过线性变换投影到高维特征空间形成 Patch 嵌入向量，通过将大尺度的图像分解成更小的部分可以获得更细粒度的信息，帮助模型更好地理解图像内容，接着位置编码 (Position Embedding) 则会给每个独立的 Patch 叠加一个可学习的正余弦位置编码使模型能够区分不同位置的 Patch，提高空间信息建模能力，然后在编码器 (Transfomer Encoder) 处通过由多层自注意力机制、前馈神经网络和残差连接等组成的 Transfomer 结构对每个独立的 Patch 进行特征提取以捕获全局依赖关系，最后，在分类头 (Classification Head) 通过 MLP 进行分类，输出最终预测结果。

深度图编码器则使用基于深度残差网络 (Deep Residual network) 的 ResNeet 架构，通过引入跳跃连接和注意力机制，从而在解决了深层网络中梯度消失的问题的同时，提高模型对深度特征的识别和利用能力。此外，模型还联合训练一个文本编码器，以从文本指令中提取语义特征，并通过最大化数据集中互相匹配的环境特征与文本特征的余弦相似度(3-1)，确保模型能够准确地学习环境特征和语言文本特征之间的对应关系。除此之外，在训练过程中我们采用反向传播算法不断优化多模态融合网络的参数，使模型在多种环境条件下都能可靠地进行匹配。其核心训练流程的伪代码如图3-9所示。

我们在相同的数据集上对 CLIDP 网络的图像-文本匹配准确率进行了测试，实验结果如表3-2所示。通过对分析可以看出 CLIDP 在光照条件复杂的室内环境中尤其是在曝光过度或欠曝光的情况下表现出了显著的优势，与原始 CLIP 模型相比其匹配准确率依然保持较高水平。这表明 CLIDP 通过融合深度信息使多模态网络能够更有效地感知

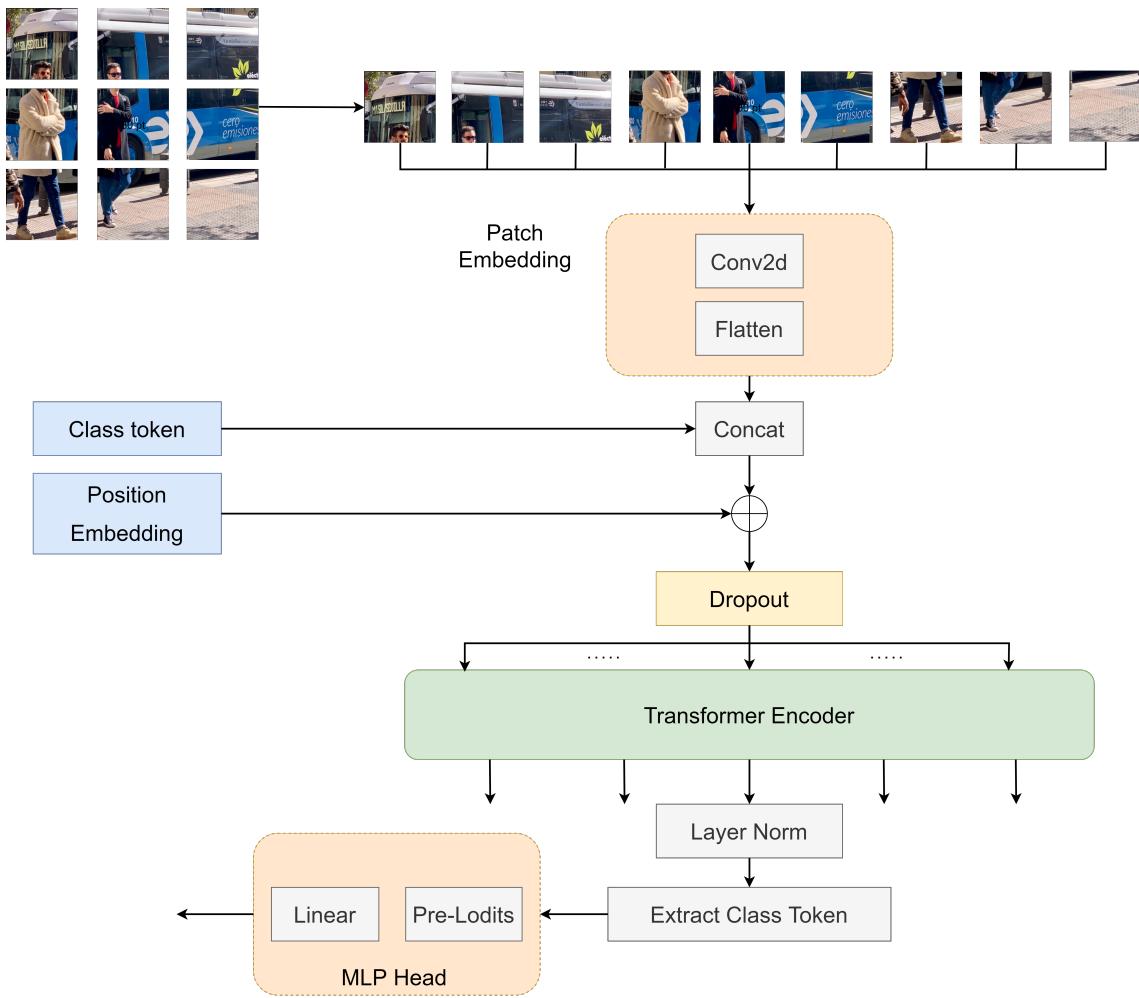


图 3-8 Vision Transfomer 结构

和理解场景结构，从而增强了模型对光照变化的鲁棒性。这一改进能够提升 CLIDP 在复杂室内环境下的图像-文本匹配精度，使其在现实应用中具有更强的适应能力。多模态融合网络 CLIDP 为后续的导航点规划算法提供可靠的目标-图像匹配信息。

$$\cos(\theta) = \frac{A \times B}{\|A\| \|B\|} = \frac{\sum_{i=1}^n A_i B_i}{\sqrt{\sum_{i=1}^n A_i^2} \sqrt{\sum_{i=1}^n B_i^2}} \quad (3-1)$$

```

# image_encoder          - Vision transformer
# depth_image_encoder   - Vision transformer
# text_encoder           - Text transformer
# I[n, h, w, c]          - 预处理图像集
# D[n, h, w, c]          - 预处理深度图集
# T[n, l]                - 预处理文本集
#W_i[d_i, d_e]           - 图像学习权重
#W_d[d_d, d_e]           - 深度图学习权重
#W_t[d_t, 2 * d_e]       - 文本学习权重
# 提取每个模态的特征
I_f = image_encoder(I)      # [n, d_i]
D_f = depth_image_encoder(D) # [n, d_d]
T_f = text_encoder(D)        # [n, d_t]
# 多模态嵌入
I_e = l2_normalize(np.dot(I_f, W_i), axis = 1) # [n, d_e]
D_e = l2_normalize(np.dot(D_f, W_d), axis = 1) # [n, d_e]
T_e = l2_normalize(np.dot(T_f, W_t), axis = 1) # [n, 2*d_e]
# 环境特征
ID_e = np.concatenate((I_e, D_e, axis = 1))
# 余弦相似度
logits = np.dot(ID_e, T_e.T)
# 构造损失函数
labels = np.arange(n)
loss_id = cross_entropy_loss(logits, labels, axis = 0)
loss_t = cross_entropy_loss(logits, labels, axis = 1)
loss = (loss_id + loss_t) / 2

```

图 3-9 CLIDP 实现的核心伪代码

### 3.5 方位优化算法

传统的视觉语言导航方法主要依赖于从图像中提取的目标特征和环境特征进行推理导航，但在包含多个相同物体的复杂室内环境之中，图像中的目标特征和环境特征会交织在一起，并且会忽略目标与其他物体之间的空间位置关系，导致视觉语言导航的准确率下降。此外，第一人称视觉观察中的所有物体视觉信息都会被同时处理，这要求代理需要在环境中所有可能的物体中进行判断，可能会浪费大量的时间去处理与目标关联度较小的物体，而这种情况在目标物体不显眼、环境物体过多的环境中尤为突出，这增加了后续模型计算推理的复杂性，同时也降低了导航的效率，并且可能导致计算效率低下。

在室内环境进行视觉语言导航的实验中，当环境的多个不同位置都包含相似或相同的物品时，智能体难以仅仅依赖多模态融合网络和 Dijkstra 算法根据自然语言指令准确执行导航任务。为了解决这一问题本文引入了一种方位优化算法，它通过分析自然语言指令中包含的方位语义来有效地筛选出冗余的导航点进而改善导航路径的准确性和

效率。具体而言，通过语言模型可以从各种形式的自然语言指令中提取出 null、front、back、left 和 right 这五种常见的方位指代。这些方位指代帮助代理在进行导航规划时不需要考虑环境中所有存在的所有导航点，而是专注于与目标物体相关的区域和物体从而更高效地筛选出有用的信息，同时也能达到减少不必要的计算的目的，避免了在视觉上和空间上不相关的物体对目标识别的干扰。除此之外，这种方位优化方法还能辅助代理确定目标位置与当前机器人位姿之间的相对方位，让代理在环境复杂、物体众多的情况下也能专注于目标物体所在的特定区域而不会受到无关物体的干扰，剔除那些不在指定方位上的冗余导航点从而使代理可以更快、更准确地到达目标，这也增强了导航系统的整体性能。方位优化过程能够大幅提升全局路径规划中生成导航点序列的准确度，使得导航路线更加贴合指令的要求。在第五章中我们通过仿真实验和消融实验验证了该方位优化算法在室内环境下进行视觉语言导航的有效性。

在 ROS(Robot Operating System) 系统中，每个导航点的坐标都是基于地图坐标系 (map) 来定义的。为了实现精确的路径规划和导航，必须通过坐标变换将基于地图坐标系下的导航点坐标转换为机器人当前位姿坐标系下的坐标，如图3-10所示。这一数学转换可以通过式(3-2)进行。

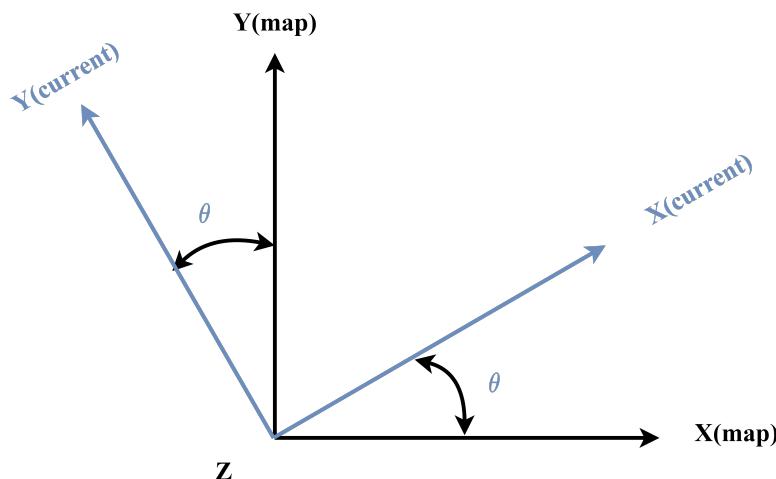


图 3-10 地图与实时位姿坐标变换

$$P_B' = \begin{pmatrix} \cos \theta & \sin \theta \\ -\sin \theta & \cos \theta \end{pmatrix} (P_B - P_A) \quad (3-2)$$

其中  $P_A$  表示机器人在地图坐标系中的位置， $P_B$  表示某个导航点在地图坐标系中的位置， $P_B'$  表示导航点在机器人坐标系下的相对位置，而  $\theta$  则表示机器人在地图坐标系中的偏航角，它代表了机器人当前位置相对于地图坐标系的旋转角度。通过计算这一角

度，可以将环境中的所有导航点坐标转换为机器人坐标系中的坐标，从而确保全局路径规划中的导航点坐标是准确的。

然而，偏航角  $\theta$  本身无法直接测量或获得，它必须通过旋转矩阵进行求取。在 ROS 的 Navigation 框架中，可以使用 tf 库中的 TransformListener 来监听机器人在地图坐标系下的位姿信息。位姿信息是通过四元数  $q = (x, y, z, w)$  来表示的，其中  $(x, y, z)$  表示机器人的位置，而  $w$  则表示旋转参数。通过监听对应话题获取这一四元组后，便可以将四元组转换成式3-3表示的旋转矩阵  $R$

$$R = \begin{bmatrix} 1 - 2y^2 - 2z^2 & 2(xy - zw) & 2(xz + yw) \\ 2(xy + zw) & 1 - 2x^2 - 2z^2 & 2(yz - xw) \\ 2(xz - yw) & 2(yz + xw) & 1 - 2x^2 - 2y^2 \end{bmatrix} \quad (3-3)$$

除此之外，机器人当前的位姿的欧拉角是在地图坐标系下，分别通过绕 X 轴旋转  $\gamma$  角、绕 Y 轴旋转  $\beta$  角以及绕 Z 轴旋转  $\alpha$  角获得的。这些角度的组合构成了机器人在地图坐标系下的完整姿态。式3-4给出了旋转矩阵的计算方式，通过这个旋转矩阵可以有效地表示机器人在地图坐标系中的朝向。

$$R_{XYZ}(\gamma, \beta, \alpha) = \begin{bmatrix} c\alpha \cdot c\beta & c\alpha \cdot s\beta \cdot s\gamma - s\alpha \cdot c\gamma & c\alpha \cdot s\beta \cdot c\gamma + s\alpha \cdot s\gamma \\ s\alpha \cdot c\beta & s\alpha \cdot s\beta \cdot s\gamma + c\alpha \cdot c\gamma & s\alpha \cdot s\beta \cdot c\gamma - c\alpha \cdot s\gamma \\ -s\beta & c\beta \cdot s\gamma & c\beta \cdot c\gamma \end{bmatrix} \quad (3-4)$$

式中： $c\alpha$  表示  $\cos \alpha$ ， $s\alpha$  表示  $\sin \alpha$

通过联立式3-3和3-4可以通过四元组来表示出欧拉角  $\alpha$ 、 $\beta$  和  $\gamma$  如式3-5

$$\begin{aligned} \beta &= \text{Atan2}\left(-r_{31}, \sqrt{r_{11}^2 + r_{21}^2}\right) \\ \alpha &= \text{Atan2}\left(\frac{r_{21}}{c\beta}, \frac{r_{11}}{c\beta}\right) \\ \gamma &= \text{Atan2}\left(\frac{r_{32}}{c\beta}, \frac{r_{33}}{c\beta}\right) \end{aligned} \quad (3-5)$$

其中  $r_{ij}$  表示旋转矩阵中第 i 行第 j 列元素， $\text{Atan2}(x, y)$  表示  $x/y$  的反正切值。

式3-5计算得到的角度  $\alpha$ ，即为机器人当前位姿在地图坐标系中的偏航角  $\theta$ ，将其代回式3-2进行坐标转换，可以计算出各个地图坐标系线下的导航点在机器人坐标系下的精确坐标。通过与语言模型解析的方位指令进行对比，系统能够筛选出哪些导航点位于指令所指定的方位范围内，从而剔除那些不符合指令要求的冗余导航点，为机器人提供精确的路径点序列。这一优化过程显著提升了机器人在执行视觉语言导航任务时的准确

性和效率。

### 3.6 导航点规划算法

在拥有环境拓扑图的导航点选择任务中通常会使用 Dijkstra 算法进行决策，根据最短距离的贪心思路帮助移动机器人进行导航点的选择，以期更快地找到目标。然而这种方法无法将导航目标与环境物体和物体之间的位置关系结合起来，且这种导航方法大多使用于静态的导航环境之中，对于可能会因为物体的摆放而改变可导航路径的复杂环境会使得原本的拓扑图失效，降低代理导航正确率和导航效率。基于这一思路，我们给全局路径规划任务进行建模，设计了一种全新的导航点规划算法。

全局路径规划可以描述为在遵循自然语言指令中给出的目标的情况下最大化导航成功的概率。具体来说，室内导航环境可以由一系列的房间  $E = e_1, e_2, \dots, e_k$  共同组成，环境中存在由一系列导航点  $V = v_1, v_2, \dots, v_n$  组成的拓扑图  $G$ ，图中的节点代表了环境中的枢纽，边的权值则代表一个节点导航到另一个节点的路径代价。一次导航任务  $\tau \in D$  由导航环境、拓扑图、初始位姿  $S$  和导航指令  $I$  确定，因此我们将每次导航任务表示成  $\tau = (E, G, S, I)$ 。在先前的小节中我们通过指令语义提取模块对导航指令  $I$  进行解析，提取出导航指令中的目标序列  $\bar{l} = l_1, l_2, \dots, l_n$ ，导航点规划算法就是要找到一个导航点序列  $\bar{v} = v_1, v_2, \dots, v_n$ ，使得该序列能够最大化目标与导航点之间的匹配程度，并尽可能提高导航成功率，如式3-6。

$$P(v_i|l_i) P(c_{\bar{v}} = 1|\bar{v}) = \max_{1 \leq t_i \leq \dots \leq t_n \leq n} \prod_{1 \leq i \leq n} p(v_{t_i}|l_i) \cdot \prod_{1 \leq j \leq n} \mu^{Dis(v_j, v_{j+1})} \quad (3-6)$$

其中  $P(c_{\bar{v}} = 1|\bar{v})$  表示根据导航点序列  $\bar{v}$  能够完成导航的概率， $P(v_i|l_i)$  表示导航节点  $v_i$  与给定的一个目标  $l_i$  相匹配的概率，该概率通过多模态融合网络进行预测输出， $\mu \in (0, 1)$ ， $\mu^{Dis(v_j, v_{j+1})}$  代表从导航节点  $v_j$  导航到  $v_{j+1}$  的距离代价，即随着拓扑图中所存在的任意两个导航点  $v_j$  与  $v_{j+1}$  距离的增大，他们成为自然语言指令所指示的导航任务中相邻导航点的概率越小。

为了求得式3-6一个最优的解  $\bar{v}$ ，我们对其进行求导，得到一个对于序列  $\bar{t} = t_1, t_2, \dots, t_n$  单调递增的函数：

$$R(\bar{v}, \bar{t}) = \sum_{i=1}^n CLIDP(v_{t_i}, l_i) + \log(\mu) \cdot \sum_{j=1}^{T-1} Dis(v_j, v_{j+1}) \quad (3-7)$$

当且仅当  $(\bar{v}, \bar{t})$  最大化式3-7时，可以同时使得序列  $\bar{v}$  最大化式3-6

为了求得最大化式3-6的解  $(\bar{v}, \bar{t})$ , 可以通过构造一个辅助函数, 再利用动态规划的方法得到  $R$  的全局最优解。综上所述, 对于  $\forall i \in \{0, 1, \dots, n\}$  和  $\forall v \in V$ , 定义一个辅助函数  $Q(i, v)$ , 表示以  $v$  结尾的所有导航点序列与目标点序列  $(l_1, l_2, \dots, l_n)$  匹配的最大值:

$$Q(i, v) = \max_{\substack{\bar{v}=(v_1, v_2, \dots, v_j), v_j=v \\ \bar{t}=(t_1, t_2, \dots, t_i)}} R(\bar{v}, \bar{t}) \quad (3-8)$$

我们通过算法2描述了结合环境拓扑图、多模态融合网络的导航点规划算法的具体过程。从起始位姿开始, 通过动态规划的方式, 最大化目标匹配度和最小化路径代价, 以获得全局路径规划输出的导航点序列。

---

**Algorithm 2** 导航点规划算法

---

**Input:** 目标序列  $\bar{l} = l_1, l_2, \dots, l_n$ , 拓扑图  $G$ , 起始位姿  $S$  和指令  $I$

**Output:** 导航点序列  $\text{Path} = [\bar{v} = v_1, v_2, \dots, v_n]$

- 1: 对  $\forall i \in 1, 2, \dots, n$  和  $\forall v \in V$  初始化  $Q(i, v) = -\infty$
  - 2: 初始化  $Q(0, v)$
  - 3: **for**  $i \in 1, 2, \dots, n$  **do**
  - 4:   对  $\forall v \in V$ , 通过递推公式求解  $Q(i, v)$
  - 5:   Path 中添加导航点  $(\text{argmax}(Q(i, *)))$
  - 6:    $S = \text{argmax}(Q(i, *))$
  - 7: **end for**
- 

在上述的导航点规划算法中, 首先输入通过语言模型提取的目标序列和环境拓扑图, 设置导航起始点  $S$ , 初始化  $Q(0, v)$  为导航起始点  $S$  到各个导航点  $v$  的最短路径长度, 当  $i \geq 1$  时, 我们通过动态规划递推公式计算每一时刻的  $Q(i, v)$  值:

$$Q(i, v) = \max \left( Q(i-1, v) + \text{CLIDP}(v, l_i), \max_{\omega \in \text{neighbors}(v)} \lambda \cdot Q(i, \omega) + \mu \cdot \text{Dis}(v, \omega) \right) \quad (3-9)$$

其中,  $\text{neighbors}(v)$  表示拓扑图中节点  $v$  的邻接点,  $\lambda \in (0, 1)$  表示缩放参数, 目的是减少先前的导航点选择对后续决策的影响。在每一次的递推过程中, 我们采用贪心策略选取能够使得  $Q(i, v)$  取得最大值、且经过方位优化筛选后的节点作为当前目标所对应的导航点, 并将其设为下一递推过程的起点, 直到遍历完目标序列, 得到全局路径规划的导航点序列  $\bar{v} = v_1, v_2, \dots, v_n$ 。

### 3.7 本章小结

本章主要内容是目标物体导航过程中涉及到的一些常用的网络在导航过程中的作用以及这些网络的算法原理和操作流程，主要介绍了指令语义提取模块、多模态融合网络模块、方位优化算法和导航点规划算法。

## 第四章 局部路径规划导航方法

本章将详细介绍 LVL-Nav 方法中的局部路径规划导航方法，包含描述全局路径规划方法不足之处的引言，针对该不足所提出的局部路径规划方法的导航框架设计，最后分别描述了框架中的特征提取模块、特征融合模块、运动模块和图像点云融合模块四个主体部分。

### 4.1 引言

目标物体导航方法旨在使移动机器人能根据自然语言指令中的多种目标进行导航，并要求机器人能够导航至目标位置的半米范围内。使机器人能够理解和执行指令是完成这一任务的关键之一，但基于导航点的全局路径规划导航方法侧重于将移动机器人引导到环境中预设的导航点，在许多应用场景中实际选取的导航点可能只是路径上的中转位置，仅仅代表该点距离目标较近而不代表机器人已经到达最终的实际目标旁。出于这类方法的限制，我们所提出的全局路径规划方法在目标没有明确的预设导航点的情况下无法十分有效地应对需要精确导航到目标物体的任务。除此之外，在未知或动态变化的环境中机器人需要能够根据感知信息实时地调整自己的路径并同时发现和定位到目标物体。传统的导航方法并不具备让机器人在局部环境中进行自主探索的能力，无法完成机器人在未知环境中的探索任务。因此要在更复杂的应用场景中实现精准的导航，移动机器人不仅需要依赖于导航点的引导，还需要具备自主探索、识别目标并作出相应决策的能力。

在现有的视觉语言导航方法中，2020 年，Krantz<sup>[74]</sup> 等人开发了序列到序列的基准模型和能够融合多种特征的多模态注意模型，用以完成连续环境中的视觉语言导航。Du<sup>[75, 76]</sup> 等人提出了基于视觉神经网络的目标物体导航方法，使用 Transformer 表征环境特征并输入强化学习网络学习导航策略，降低了模型训练时的试错成本。Wang<sup>[77]</sup> 等人为了解决不同机器人之间难以传递导航技能而提出了一种基于元学习的视觉感知泛化策略，在视觉感知模型中分别使用与模型无关的元学习算法和基于指标的元学习算法，以便在可见和不可见的环境中更好地泛化，使智能体能够快速适应新的相机配置。Fang<sup>[78]</sup> 等人提出了一种基于强化学习的目标物体导航方法，提高了智能体的泛化能力。Fukushima<sup>[79]</sup> 等人提出基于对象记忆 Transfomer 的目标物体导航方法，使用对象场景存储器存储长期场景和对象序列，实现室内环境高效导航。朱威<sup>[80]</sup> 等人提出了一种结合优势结构和最小化目标 Q 值的深度强化学习导航算法，加快多目标连续导航训练过程

中的收敛速度。这类方法基本都采用从视觉观察中提取物体语义、物体位置和相对位置等特征信息，将其通过编码的嵌入层网络以构建丰富的视觉表示，以此来告诉模型周围环境的特征。除此之外模型还会关注当前视觉观察中与目标物体相关联的区域的方向信息，这使得代理能够朝着正确的方向进行探索导航。

根据上述的问题和已有的解决思路，本章提出了一种局部路径规划方法。我们的目标是根据指令中目标物体的名称，通过特征融合、特征提取网络模型输出的离散动作在局部环境中进行自主探索，同时利用视觉图像信息识别出目标物体，在视觉观察中定位目标物体后再由图像点云融合算法计算获得目标的精确位姿，转换坐标系后发布导航任务完成局部环境目标物体导航。该方法由特征提取模块、特征融合模块、运动模块和图像点云融合模块构成。

## 4.2 导航框架设计

未知环境下视觉导航系统需要实现未知环境探索、动作执行和图像点云融合导航三个方面的功能，因此我们设计了由未知环境探索节点、运动节点、图像点云融合节点构成的局部路径规划导航系统，如图所示4-1。

其中，未知环境探索节点由多模态特征提取器与跨模态融合器一同组成，该节点基于 Transformer 的并行注意力架构对视觉观察的场景特征向量与目标对象语义描述特征进行融合，进而生成用以描述导航方向指引向量以及表征空间拓扑结构的环境特征。接着，在 ROS 机器人导航框架之下通过动作执行节点封装可能执行的离散环境探索动作，以期找到导航目标，当发现目标时则进入图像点云融合节点，通过点云聚类算法、目标检测算法和 IoU 度量算法获得精准的目标物体的位置信息，最后发布导航任务完成目标物体导航。

具体来说，特征提取模块采用预训练参数初始化的残差卷积网络来解析场景的宏观语义表征，同时通过基于注意力机制的检测模型捕获包含局部对象的语义分布及其几何编码信息在内的细粒度区域感知特征，并通过参数化嵌入映射层来生成具有可解释性的目标特征向量。接着，在特征融合模块中对全局特征、局部特征和目标特征进行融合，将获得的方向特征、环境特征和上一时刻的动作特征通过 LSTM 网络进行编码生成导航动作和动作评分，通过运动模块中封装好的离散动作函数进行执行，以此来指导代理在局部未知环境中进行探索。接着，在视觉和多线激光雷达进行联合标定并确定发现目标后，通过 YOLOV10 目标检测算法认知环境中存在的目标的同时，再通过点云聚类感

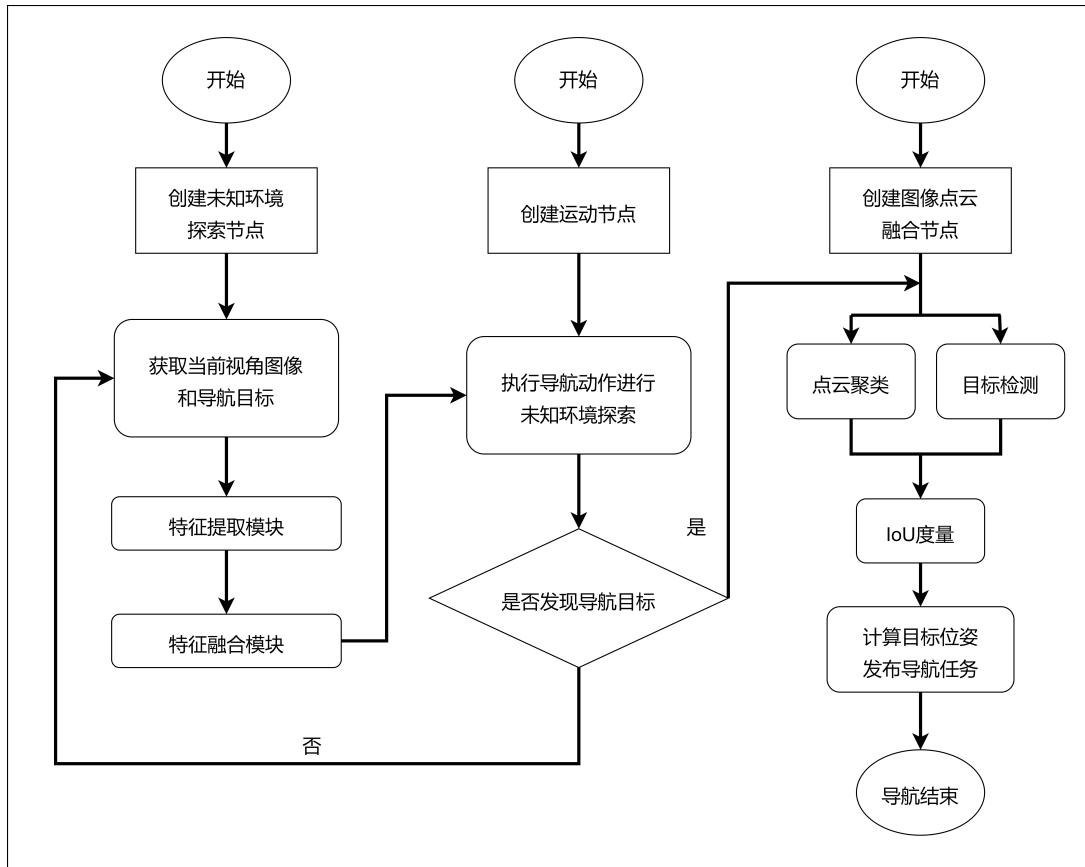


图 4-1 局部路径规划导航框架

知获得目标与移动机器人之间的相对位置信息，通过图像点云融合模块获得目标位姿，再经过坐标变换得到目标在 map 坐标系下的坐标。最后再运动模块中利用 Navigation 导航框架发布导航任务完成导航到目标旁以完成局部导航，进一步完成移动机器人执行导航的闭环任务。

未知环境下的导航探索主要由特征提取模块和特征融合模块构成，如图4-2，其中特征提取模块从移动机器人上的单目相机获取第一人称环境视觉图像，并将其送入不同的特征提取网络中，分别提取表征各个区域物体语义、位置信息的  $100 \times 256$  局部特征、表征导航环境中机器人所处位置的位姿信息的  $49 \times 256$  全局特征和表征探索导航目标的  $1 \times 256$  目标特征。接着再通过主要由 encoder 特征强化、decoder 特征融合和 LSTM 这三个结构组成的特征融合模块生成表征离散动作空间的动作概率分布函数。具体来说特征融合模块将特征提取获得的三个不同特征进行强化融合，并通过 LSTM 网络进行导航序列建模以生成  $1 \times 4$  的动作特征，最后由运动模块进行探索任务。

上述操作过程可以用公式表示为

$$X_{\text{local}}, X_{\text{global}}, X_{\text{target}} = \text{Extract}(X_{\text{image}}, X_{\text{text}}) \quad (4-1)$$

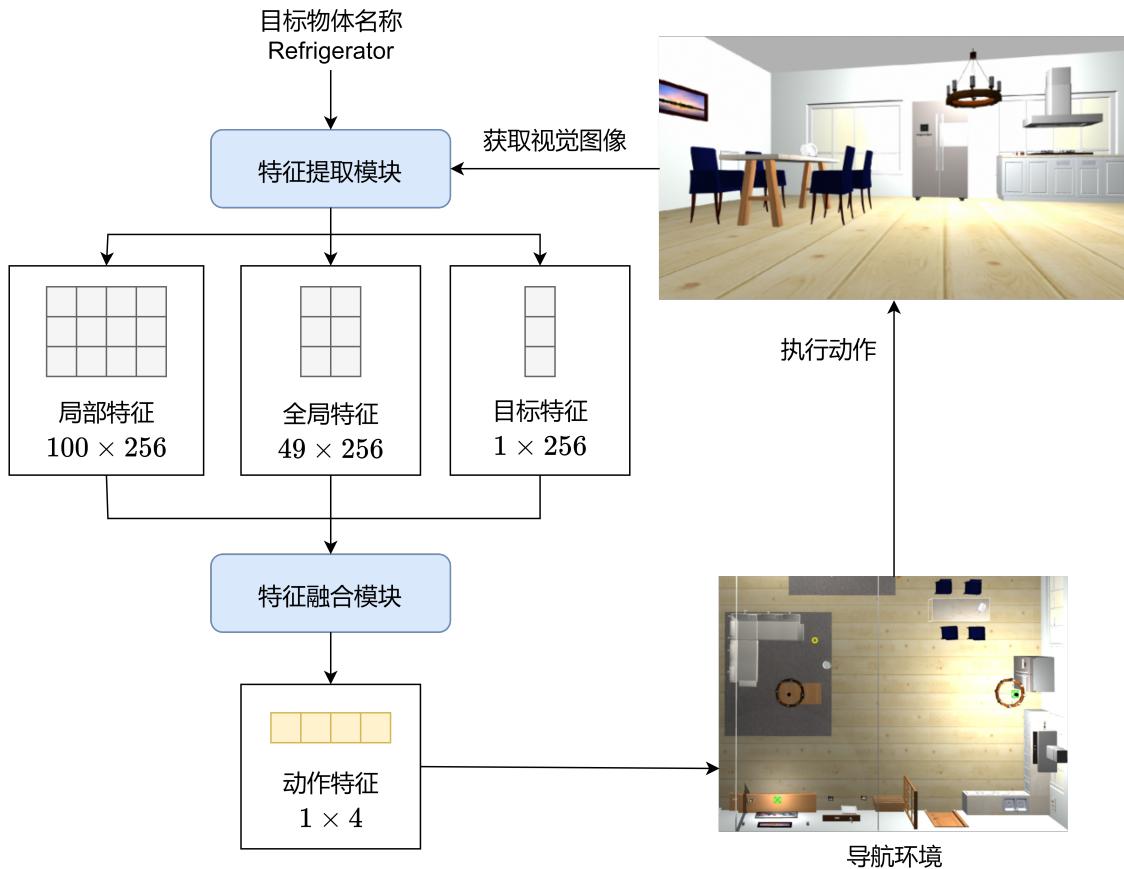


图 4-2 未知环境下的导航探索框架

$$X_{\text{action}} = \text{Fusion}(X_{\text{local}}, X_{\text{global}}, X_{\text{target}}) \quad (4-2)$$

其中  $\text{Extract}()$ ,  $\text{Fusion}()$  分别表示特征提取模块、特征融合模块,  $X_{\text{local}}$ ,  $X_{\text{global}}$ ,  $X_{\text{target}}$ ,  $X_{\text{image}}$ ,  $X_{\text{text}}$ ,  $X_{\text{action}}$  分别表示特征提取模块中的局部特征、全局特征、目标特征、第一人称视觉图像、输入目标和特征融合模块的动作特征。

### 4.3 特征提取模块

在未知环境的探索导航过程中使用多层次的特征提取策略能够充分利用环境信息,使机器人在无先验知识的情况下更高效、可靠地探索并接近目标。在未知环境下的探索导航过程中,特征提取模块主要由三个分别提取局部特征、全局特征和目标特征的不同网络组成,可以用公式表示为

$$X_{\text{local}} = \text{Concat}(\text{ReLU}(\text{Linear}(\text{DETR}(X_{\text{image}}))), X_{\text{corelation}}) \quad (4-3)$$

$$X_{\text{global}} = \text{Flatten}(\text{Conv}(\text{ResNet}(X_{\text{image}})) + X_{\text{position}}) \quad (4-4)$$

$$X_{\text{target}} = \text{Embedding}(X_{\text{text}}) \quad (4-5)$$

经过特征提取模块处理后，提取出的用于描述局部场景内物体语义属性及方位信息的区域特征、描述整体环境状态的全局特征以及反映待识别目标语义特性的对象特征均被标准化为 256 维向量并集成于同一嵌入空间之中。这一对齐操作有效促进了多模态特征的协同交互，为后续特征融合环节构建视觉观测数据与目标实体间的深度语义关联奠定了结构基础，能够实现异构特征的高效融合。

#### 4.3.1 提取局部特征

局部特征的提取流程如图4-3所示。在一开始我们使用预训练的 DETR 网络来提取第一人称视觉图像，用于表征当前视觉观察的局部信息、帮助代理认知视觉观察图像中物体代表的语义信息和位置信息，得到  $100 \times 256$  的局部特征，其中我们的 DETR 网络使用 ResNet-50 作为主干网络，采用六层编码器、六层解码器、八个独立注意力头和维度为 2048 的前馈神经网络搭建而成。然后将该特征经过全连接层 Linear 和 ReLU 激活函数以得到  $100 \times 249$  的局部特征，再将表征推理结果的  $100 \times 7$  关联特征与局部特征进行连接，得到包含环境所存在的物体语义信息、位姿信息、可靠性在内的  $100 \times 249$  的局部特征。

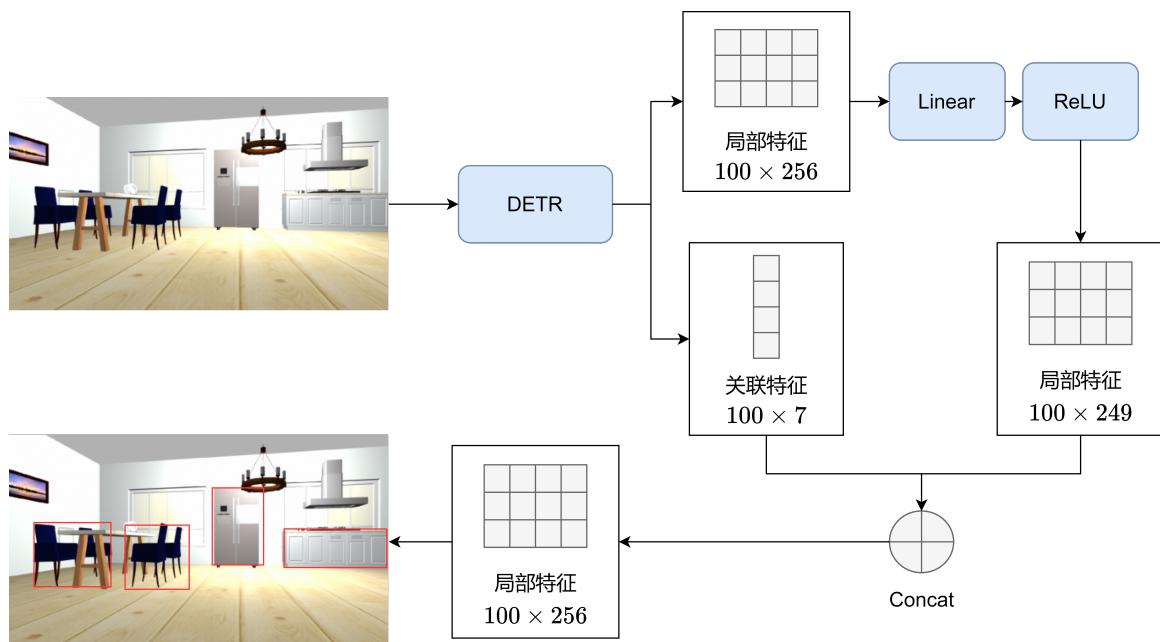


图 4-3 局部特征提取流程

### 4.3.2 提取全局特征

全局特征的提取流程如图4-4所示。我们构建了基于 ResNet18 架构的特征提取网络用于解析机器人本体视角采集的 RGB 图像数据。该编码器通过深度卷积网络提取场景的空间拓扑表征来为智能体提供空间定位的环境认知信息。首先使用斯坦福大学开源的 ImageNet 数据集对我们搭建的 ResNet18 网络进行预训练，将第一人称观察视觉图像输入到该预训练的网络中再经过 Conv 卷积层和 ReLU 激活函数得到  $7 \times 7 \times 512$  的全局特征，然后将输出的环境表征张量与三角函数位置嵌入生成的  $7 \times 7 \times 256$  维位置编码矩阵进行 Add 操作和 Flatten 操作得到维度为  $49 \times 256$  的融合特征向量以完成多模态感知信息的空间对齐。其中正余弦位置编码是一种固定、无训练参数的编码方法，它可以表示为式4-6、4-7，其中 pos 表示位置索引， $i$  表示特征维度索引， $d_{\text{model}}$  表示嵌入维度即模型隐藏层维度，分母  $10000^{2i/d_{\text{model}}}$  表示用于确保不同维度的编码值在不同的频率范围内变化的缩放因子。

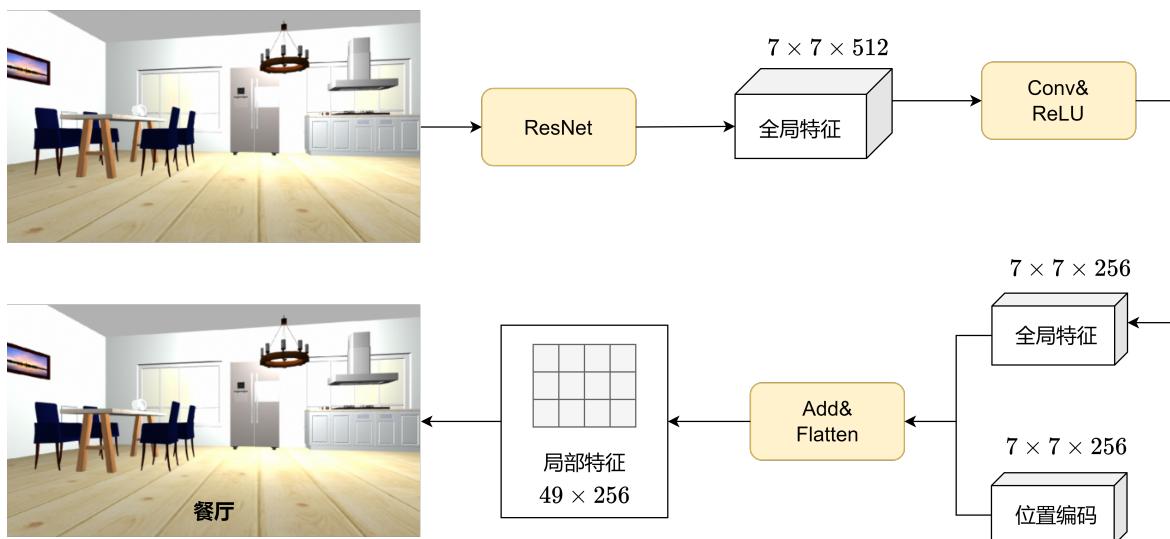


图 4-4 全局特征提取流程

$$PE_{(pos,2i)} = \sin \left( \frac{pos}{10000^{2i/d_{\text{model}}}} \right) \quad (4-6)$$

$$PE_{(pos,2i+1)} = \cos \left( \frac{pos}{10000^{2i/d_{\text{model}}}} \right) \quad (4-7)$$

### 4.3.3 提取目标特征

目标特征提取流程如图4-5所示。我们在特征构建阶段定义词表容量  $V = 32$  与嵌入空间维度  $D=256$  作为核心超参数并以此初始化嵌入矩阵  $M$ ，接着通过输入目标实体在词表中的离散索引  $I$  来执行嵌入层的索引查询操作，从映射关系矩阵  $M$  中抽取对应的目标语义向量进而完成符号空间到连续向量空间的转换。在训练过程中，该目标特征会不断被优化，使语义相近的信息在嵌入空间中更加接近，从而增强特征融合模块在构建文本语义与视觉语义关联方面的能力。利用嵌入层网络对目标物体的单词进行编码可以生成用于指代目标物体语义的目标特征。在后续的特征融合过程中，可以借助注意力机制筛选出视觉观察中与目标物体相关的物体特征，从而精准确定导航的方向以快速在未知环境中找到目标。



图 4-5 目标特征提取流程

## 4.4 特征融合模块

特征融合模块主要由特征融合编码器、特征融合解码器和 LSTM 网络构成。其中特征融合编码器主要由具有自注意力机制的 Transformer encoder 组成，它将特征提取模块获得的局部特征通过自注意力机制编码强化，获得能够表征表示物体与一和位置信息的强化局部特征，并将其作为键值对输入到特征融合解码器之中。特征融合解码器同样由具有多头自注意力机制的 Transfomer decoder 组成，它将表征目标物体语义信息和机器人所处环境位置信息的目标特征和全局特征进行强化编码，然后将作为键值对的局部特征与他们连结而成查询一同输入到网络之中，通过 decoder 进行融合得到环境特征和方向特征，最后，经过 LSTM 网络输出代理在未知环境下进行探索的离散导航动作。

我们通过观察人类在环境中寻找目标物体的过程发现，当目标物体出现在视觉观察范围内时，我们会直接按照最近的路线去靠近目标，而当目标物体未出现在观察范围内时，我们则会前往环境中存在的所有物体中与目标关联性最强的物体附近进行寻找。我们按照这样的思路去搭建特征融合网络以指导代理能够更精准的在局部未知环境中进行导航，如图4-6所示，环境中存在硬纸盒、手提电脑和花盆三个物体。当需要导航到视觉观察中所发现的目标物体花盆时，代理就会根据标目标物体出现在视觉观察中的位

置进行决策，执行向右前方进行移动的动作。当需要导航到视觉观察中未出现的目标物体鼠标时，代理则会先判断视觉观察中所出现的所有物体与目标物体的关联性，然后导航到手提电脑这一最可能会找到鼠标的目标旁。

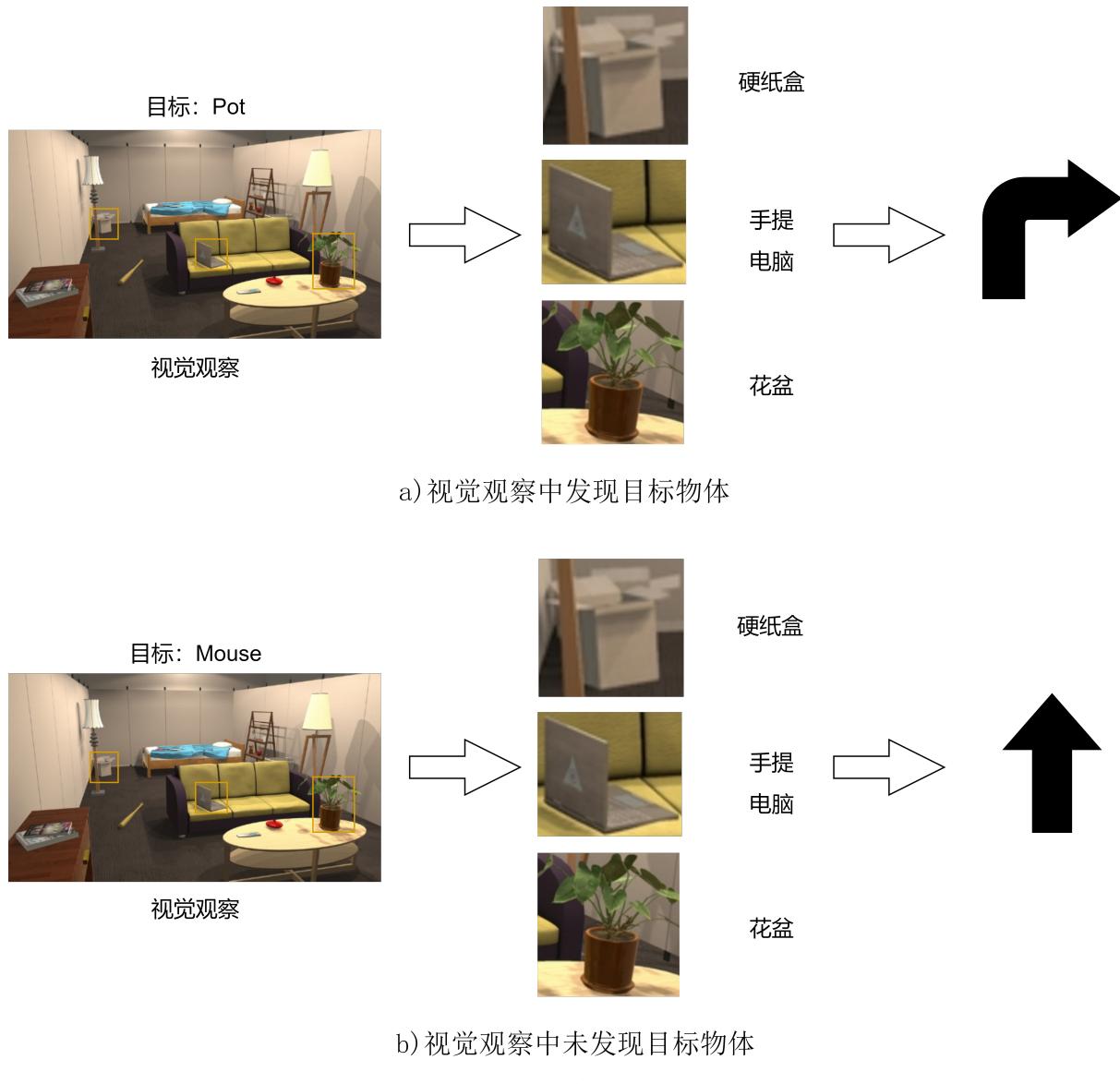


图 4-6 未知环境导航示例图

因此，为了使移动机器人能够正确定位其当前所在位置，需要从视觉观察中提取环境特征，同时还需要提取局部特征以感知周围环境中物体的语义信息和位置信息。此外，代理应建立目标特征与全局特征、局部特征之间的紧密关联，从而有效引导方向特征的生成。同时，还需融合当前的全局特征与局部特征以获取环境的整体表征，并构建方向特征与环境特征相对应的序列，为后续模块提供历史经验信息，帮助 LSTM 网络正确地输出导航动作以指导代理在局部未知环境中进行精准高效的导航。特征融合模块的操作流程可以用如下公式表示，其中 *Encoder* 表示特征融合编码器，*Decoder* 表示特征

融合解码器。

$$X_{\text{encoder}} = \text{Encoder}(X_{\text{local}}) \quad (4-8)$$

$$X_{\text{direction}}, X_{\text{environment}} = \text{Decoder}(\text{Concat}(X_{\text{global}}, X_{\text{target}}, X_{\text{parameter}}), X_{\text{encoder}}) \quad (4-9)$$

$$X_{\text{action}}^n = \text{Linear}(\text{LSTM}(X_{\text{direction}}, X_{\text{environment}}, X_{\text{action}}^{n-1})) \quad (4-10)$$

#### 4.4.1 特征融合编码器

特征融合模块中的编码、解码器基于 Transfomer 进行搭建，它的参数如表4-1所示，即由输入维度为 256、层数为 2、自注意力头为 8、前馈神经网络模型的维度为 512 共同组成。encoder 结构如图4-7所示，我们将  $100 \times 256$  维的局部特征作为输入进入到 encoder 中的自注意力层中，经过第一次的残差连接、层归一化和前馈神经网络后经过第二次的残差连接和层归一化，输出得到强化后的  $100 \times 256$  维局部特征。在 encoder 强化表征环境中局部物体的语义信息和未知信息的局部特征的过程之中，分为 100 个维度的 256 个序列的每一个序列都通过网络的自注意力机制与其他序列进行交互，以学习环境中物体之间的关联信息，帮助在当前视觉观察环境中未能发现目标的代理朝着与目标物体最相关联的物体方向前进，进而发现目标物体。

表 4-1 特征融合 Transfomer 网络主要参数

n_head	num_encoder_layers	num_decoder_layers	dim_feedforward
8	2	2	512

#### 4.4.2 特征融合解码器

特征融合的解码器 decoder 结构如图4-8所示。我们将特征提取模块输出表征代理所处环境的  $49 \times 256$  维全局特征、表征当前导航目标的  $1 \times 256$  维目标特征和可学习的  $1 \times 256$  维目标特征进行连结，得到  $51 \times 256$  维的特征输入到 decoder 之中，在经过第一个多头自注意力层、残差连接、层归一化之后，将 Transfomer encoder 输出的  $100 \times 256$  维强化后的局部特征作为键值对、 $51 \times 256$  维的特征作为查询一同输入到第二个多头自注意力层之中，经过残差连接归一化层、前馈神经网络和最后一次的残差连接和层归一化后，得到  $51 \times 256$  维的特征，使用该特征中与  $1 \times 256$  维的目标特征和  $1 \times 256$  维的可学习的参数特征相对应位置的向量，作为表征环境中与目标相关联物体的信息，以帮

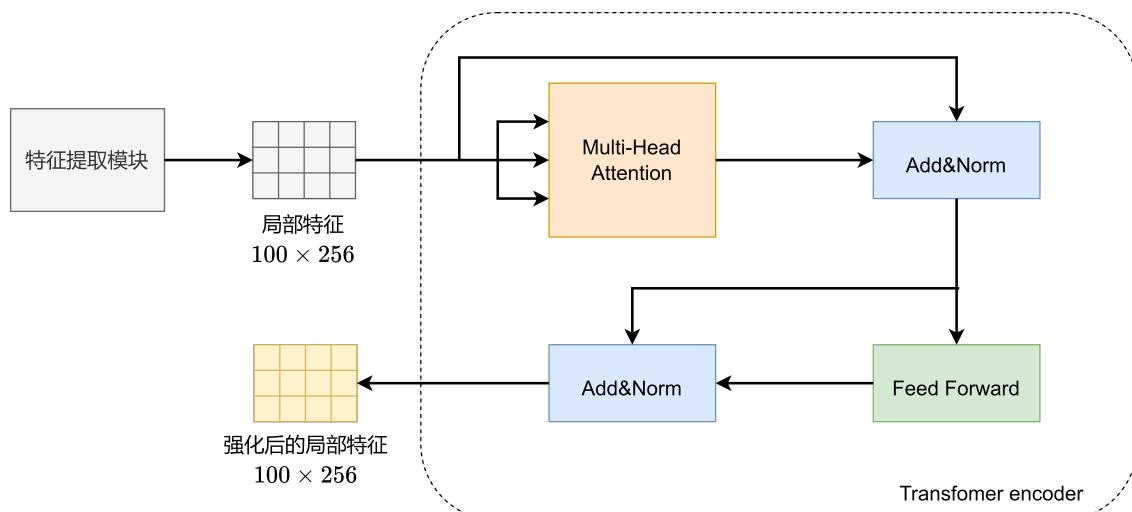


图 4-7 encoder 结构图

助代理确定导航方向的方向特征，还有表征全局特征和局部特征共同联合的环境特征。

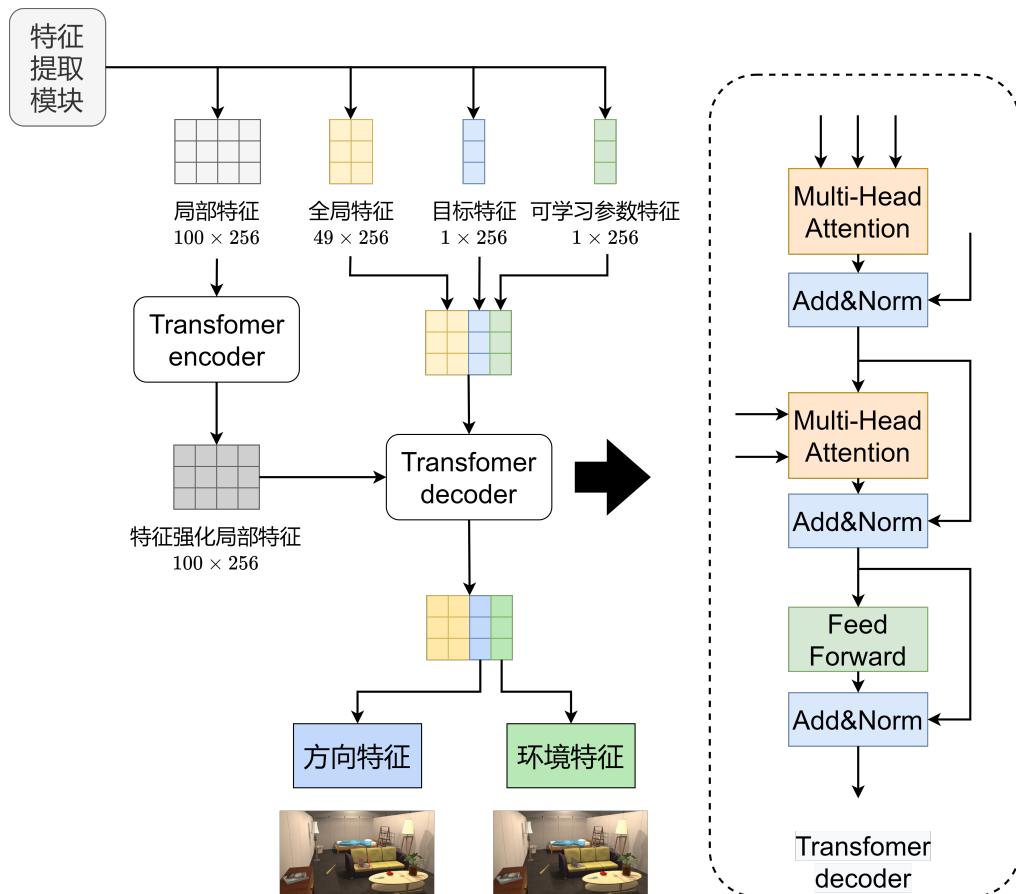


图 4-8 decoder 结构图

#### 4.4.3 LSTM 网络

LSTM 是一种能够有效地捕捉序列数据中的长期依赖关系的特殊循环神经网络，传统的神经网络或其他类型的 RNN 可能会遗忘过去的特征信息，而 LSTM 通过引入遗忘门、输入门和输出门从而有效地解决了这一问题，这种循环结构使得它在基于历史信息进行决策或是当任务需要依赖较长时间之前的状态时的场景中表现出色。而在机器人执行导航这种需要处理时序信息和长期依赖的任务的过程中，需要代理根据上一时刻的导航动作、方向特征和环境特征提取有用的历史信息，使用 LSTM 结构的网络根据以往的经验做出更精准的决策，从而达到显著提高导航系统的决策能力的目的。

通过 LSTM 网络输出局部未知环境下的导航动作流程如图4-9所示。我们将特征融合解码器输出的  $1 \times 256$  维方向特征、 $1 \times 256$  维环境特征和表示未知环境下代理所执行的  $1 \times 4$  维导航动作进行拼接，得到  $1 \times 516$  维特征向量作为双层 LSTM 网络的输入，最终输出  $1 \times 4$  维的导航动作交由运动模块所封装的导航动作执行，并且该当前时刻的导航动作在经过一个线性层后作为下一时刻的动作特征输入到 LSTM 网络中。

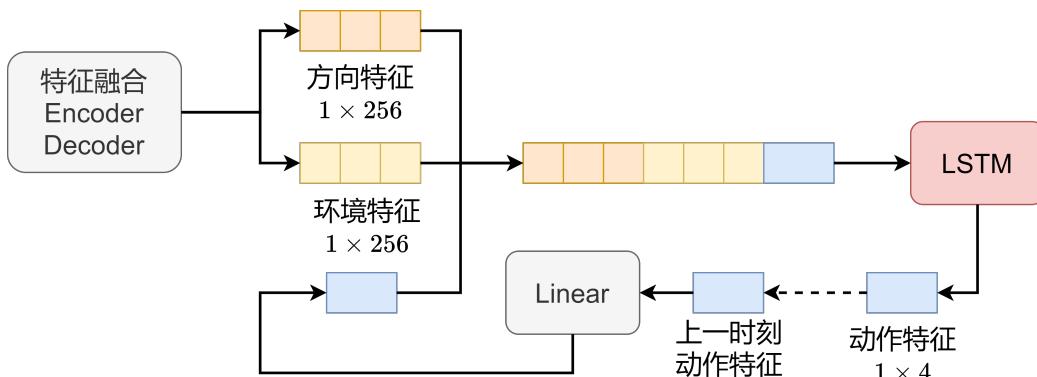


图 4-9 LSTM 结构图

#### 4.5 运动模块

在运动模块中，我们通过 ROS 的分布式通信框架来订阅目标位姿信息，依托差分驱动模型来实现机器人在环境中的移动探索。

本系统研发的移动机器人基于差速转向原理实现底盘驱动。在机械结构设计上，我们通过独立控制两侧驱动电机的转速差异来完成运动控制。当两侧驱动轮保持相同转速时机器人将保持直线行驶；当两侧转速产生差值时，根据差速传动特性机器人会以自身轴线为中心进行转向运动。基于这种原理，我们可以通过调整双轮速度参数实现不同曲率的弧线运动与精确转向定位。在制动控制方面，驱动系统可通过切断电机供电使设备

实现急停功能。

具体如图4-10所示，底盘控制系统的工作流程包含指令下发与动态修正两个关键环节。主控制器首先将预设转速参数传输至电机驱动模块，但由于机械传动间隙、地面摩擦系数变化等客观因素，机器人车轮的实际转速值与理论设定值往往存在偏差。因此系统设计了一种闭环控制架构，驱动模块不仅需要输出 PWM 控制信号，还通过正交编码器实时采集电机转动参数。在 10ms 的采样周期内，控制模块会持续统计编码器脉冲数量并将其转换为移动距离参数进行反馈。主控单元运用运动学方程对这些原始数据进行坐标变换与轨迹拟合，最终生成毫米级精度的实时位姿信息从而保障移动平台的精准定位能力。

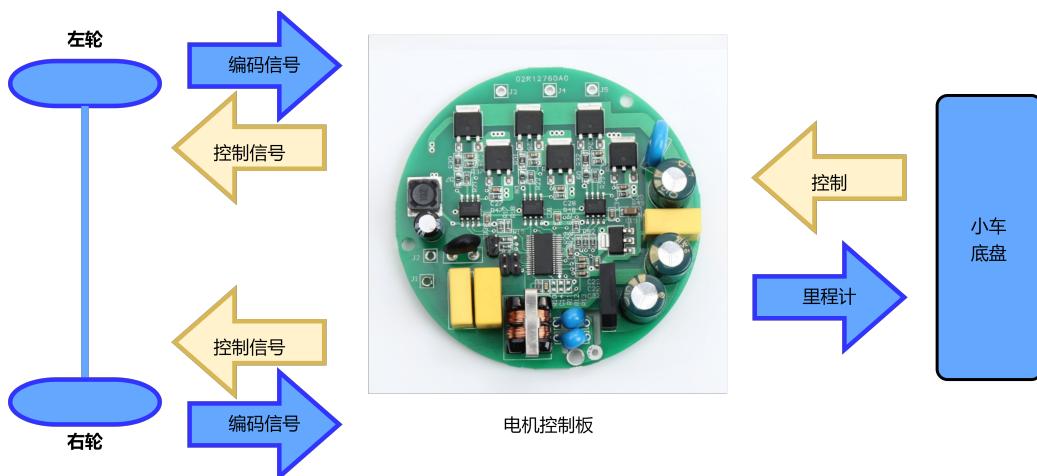


图 4-10 机器人的运动控制方法

在控制移动机器人移动的过程中通常会受到地面摩擦阻力的影响，出现左右轮的实际转速和想要控制的速度不同的情况，导致移动机器人无法按照计划的轨迹进行移动，这时候就需要使用第二章所介绍的 PID 控制算法来调整小车的运动，使其能够精准的按照目标轨迹进行移动。

在移动机器人根据多特征融合网络的输出进行局部范围内的探索以期找到目标物体时，我们需要计算出每一时刻的里程计数据，并将其与订阅获取的实时里程计数据进行比较，从而计算出当前已经移动的距离，达到让机器人准确地进行移动的目的。具体来说里程计包括位姿（位置和方向角）和运动速度（线速度与角速度）两个重要的信息，它的计算过程如图4-11所示。在整个机器人运动的过程当中，假设用  $p_i$  来表示在时刻  $i$  的位姿，图中的  $p_1, p_2, \dots, p_n$  则代表机器人的整个运动过程中的轨迹。在很短的时间内，机器人从  $p_1$  运动到  $p_2$ ，那么我们就可以根据机器人当前的位姿  $p_1$  及其左右轮的速度  $V_{left}$  和  $V_{right}$ ，通过微积分的方法来推算机器人下一个时间步  $p_2$  处的线速度、角速度及

其位姿，从而计算获得完整的里程计数据。

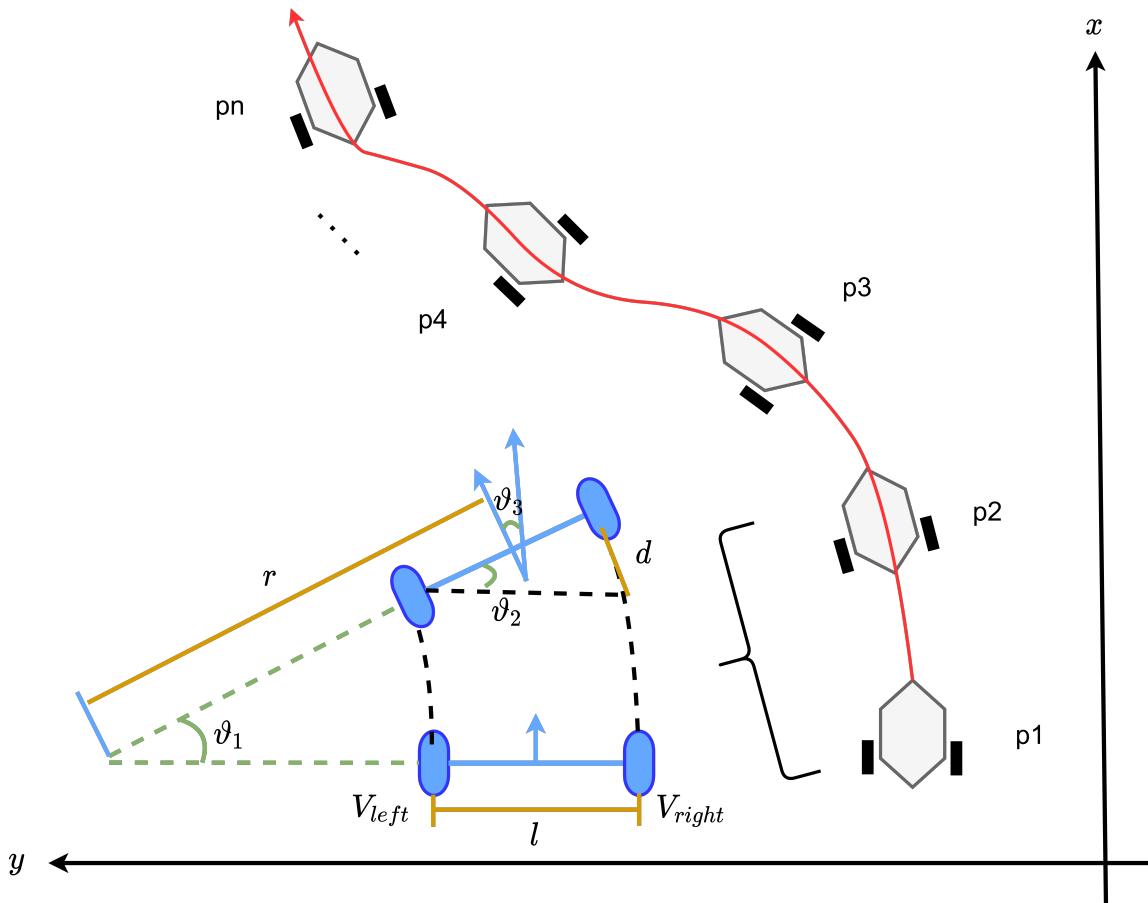


图 4-11 里程计计算

在机器人移动的三维空间中存在多个坐标系，我们构建了存在于导航环境中相对静止的全局坐标系作为机器人本体运动的参考坐标系，在任意的时刻  $t$ ，机器人的位姿可以表示为  $P_t = (x_t, y_t, \theta_t)$ ，其中  $(x_t, y_t)$  表示机器人的位姿在世界坐标系  $x$ 、 $y$  平面上的投影坐标， $\theta_t$  表示朝向角。导航系统则基于差分轮式运动学模型进行驱动，该模型将平台的运动学状态解耦为绕瞬态旋转中心 (Instantaneous Center of Rotation, ICR) 的刚体转动与平移分量的合成运动，其运动可以近似视为围绕某一瞬时旋转中心，以半径  $r$  沿圆弧轨迹运动。已知左右轮的速度  $V_{left}$ 、 $V_{right}$  及两轮间距  $l$ ，可以通过式(4-11)、(4-12)和(4-13)来计算机器人在  $t$  时刻的线速度  $v$ 、角速度  $\omega$  以及旋转半径  $r$

$$v = \frac{v_{left} + v_{right}}{2} \quad (4-11)$$

$$\omega = \frac{v_{right} - v_{left}}{l} \text{ (rad/s)} \quad (4-12)$$

$$r = \frac{v}{\omega} = \frac{l}{2} \cdot \frac{v_{right} + v_{left}}{v_{right} - v_{left}} \quad (4-13)$$

在实际计算中为了推算出机器人在当前时刻的位姿，我们基于速度积分法进行计算。设在  $t$  时刻移动机器人的位姿为  $P_t = (x_t, y_t, \theta_t)$ ，在微笑的时间序间隔  $\Delta t$  内进行的运动过程可以等效为恒定速度的匀速运动，因此可以通过如下递推式(4-14)、(4-15)和(4-16)计算出结果。

$$\theta_{t+1} = \theta_t + \omega * \Delta t \quad (4-14)$$

$$x_{t+1} = x_t + r * (\sin \theta_{t+1} - \sin \theta_t) \quad (4-15)$$

$$y_{t+1} = y_t - r * (\cos \theta_{t+1} - \cos \theta_t) \quad (4-16)$$

该方法能够高效、准确地计算机器人每个时刻的位姿，并依次累积得到完整的运动轨迹，为导航和路径规划提供精确的里程计数据支持。在具体的实现中我们使用面向对象的编程思想进行封装，将其用于探索未知环境。

## 4.6 图像点云融合模块

未知环境的目标物体导航任务要求代理要具备识别出环境中存在着的目标物体的功能。多线激光雷达具备高精度的深度测量能力，它能够提供目标的距离感知信息和三维结构，但点云数据会随着距离的变大而逐渐变得稀疏，在追踪远距离目标时效果有限。单目相机则具有高分辨率的视觉信息，能够捕捉目标的纹理、颜色和边缘特征等认知信息，但它缺乏直接的深度感知能力，同时也容易受到光照条件等环境因素的影响。相比于使用单一传感器进行目标检测，使用多线激光和单目相机的联合检测方法在各种应用场景中展现出显著的优势，能够有效降低误检和漏检率。首先图像点云融合方法能够显著提升代理对远距离目标的检测能力，即使在单目相机难以判别目标的远距离区域依旧可以通过多线激光雷达来获取准确的深度信息。此外目前的视觉言语导航方法大多仅依赖于网络模型输出的动作进行决策，使用导航至目标 3 米内即算完成导航任务这一笼统指标并不利于后续利用机械臂执行的下游任务的执行，这就需要多线激光提供的准确距离信息来辅助代理进行导航以完成导航至目标半米内的任务。

局部路径规划方法中的图像点云融合模块使用了第二章所介绍的 YOLOV10 目标检测网络、优化后的欧式点云聚类算法和多线激光单目相机联合标定方法。多线激光获

取的点云在经过点云聚类之后通过外参矩阵重投影到目标检测二维图像上，并且使用 IoU 重叠度 (Intersection over Union) 来判断目标检测框中的目标物体与点云聚类结果是否匹配，当点云聚类结果与目标检测框不匹配时就会被舍弃，当结果匹配时就会被纳入结果集，并将出所选定的区域框中所有点云的均值作为该目标的距离，用以发布最终导航目标点完成完整的导航。

IoU 具有非负性、不可分辨的同一性、对称性和尺度不变这类属性，这使得通过这个方法来判断两个任意形状 A 和 B 之间的相似性与它们的空间尺度无关，所以它被广泛用作计算机视觉中许多任务的评估指标，如像素级图像分割、2D 和 3D 对象检测等。因此，本文使用 IoU 度量法来衡量有限的点云聚类算法结果和目标检测算法结果之间的相似性。具体来说针对两个有限样本集合  $A$  和  $B$ ，他们之间的 IoU 被定义为他们的交集除以他们之间的并集，如式4-17。

$$\text{IoU}(A, B) = \frac{A \cap B}{A \cup B} = \frac{A \cap B}{|A| + |B| - A \cap B} \quad (4-17)$$

根据点云聚类和目标检测的结果可以获得两个不同的预选框，将它们通过预选框匹配算法计算他们之间的 IoU 度量，如算法3。当预选框之间的 IoU 低于 0.5 时，根据非极大值抑制准则将两者判定为独立目标；当 IoU 处于区间  $0.4 <= \text{IoU} <= 0.8$  时，根据置信度加权策略选取两个候选框的重叠区域作为最终检测框；对于 IoU 大于 0.8 的强相关预选框则通过最小外接矩形算法生成融合后的最终检测框，如图4-12。

---

### Algorithm 3 IoU 度量算法

---

$$A_1(x_1, y_1), B_1(x_2, y_1), C_1(x_2, y_2), D_1(x_1, y_2)$$

**Input:** 检测框的四个角坐标:  $A_1(x'_1, y'_1), B_1(x'_2, y'_1), C_1(x'_2, y'_2), D_1(x'_1, y'_2)$

$$x_1 \leq x_2, y_2 \leq y_1, x'_1 \leq x'_2, y'_2 \leq y'_1$$

**Output:** 两个预选框之间的 IoU 结果

- 1: The area of  $B_c$ :  $\text{Area}_c = (x_2 - x_1) \times (y_1 - y_2)$
- 2: The area of  $B_d$ :  $\text{Area}_d = (x'_2 - x'_1) \times (y'_1 - y'_2)$
- 3: The area of overlap:

$$\text{Area}_{\text{overlap}} = (\max(x_2 - x'_2) - \min(x_1, x'_1)) \times (\max(y_1 - y'_1) - \min(y_2, y'_2))$$

- 4:  $\text{IoU} = \text{Area}_{\text{overlap}} / (\text{Area}_c + \text{Area}_d - \text{Area}_{\text{overlap}})$
- 

在经过预选框选择策略之后获得的最终边界框将用于测定目标在机器人坐标系下

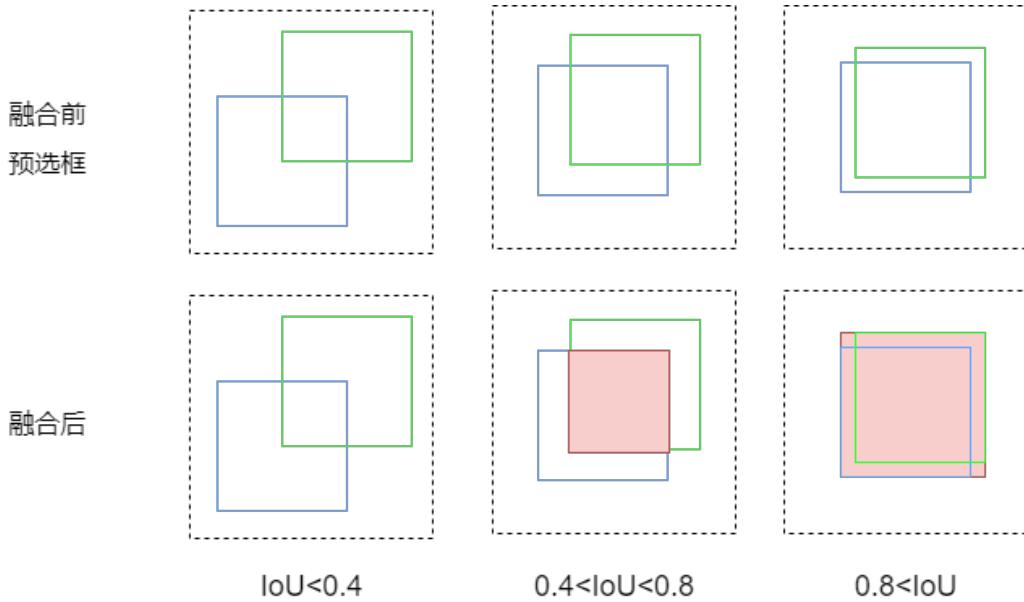


图 4-12 预选框选择策略

的位姿，如图4-13。通过最终边界框的中点  $(u_{mid}, v_{mid})$  和单目相机的水平视域、垂直视域，可以分别计算出目标物体与机器人  $x$  轴正方向的水平夹角  $\phi$ 、目标物体与机器人  $y$  轴正方向的水平夹角  $\varphi$ 。然后将投影在最终边界框中的点云距离进行均值相加以获得机器人与目标之间的距离  $D$ ，根据  $\phi$ 、 $\varphi$  和  $D$  可以计算出机器人坐标系下目标的位姿。最

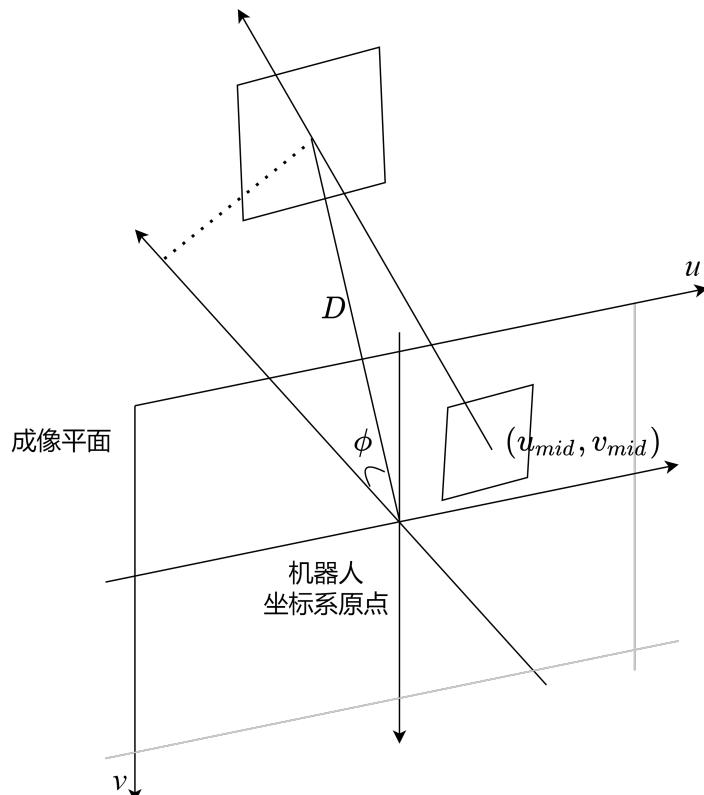


图 4-13 根据平面成像与点云数据求目标位姿

后，通过 ROS 框架中 tf 话题所提供的数据和四元组转欧拉角方法即可将机器人坐标系下的目标位姿转换成地图坐标系下的目标位姿，发布导航即可完成完整的目标物体导航任务。需要注意的是，在实验的过程中我们发现当目标物处于桌子上等不可达的位置时，移动机器人将根据 AMCL 规划出的路径移动到环境的边界处，这通常与环境中离目标最近的导航点不一致，因此我们在发布最后的导航点时会先检查该点是否可达，当其不可达时图像点云融合模块将输出环境中与最终目标导航点相距最近的可达点作为替代。具体的仿真环境、现实环境和消融实验实验结果见第五章。

## 4.7 本章小结

本章提出了一种局部路径规划方法，该方法由特征提取模块、特征融合模块、运动模块和图像点云融合模块构成。首先根据指令中目标物体的名称，通过特征融合、特征提取网络模型输出的离散动作，依靠运动节点在局部环境中进行自主探索，同时利用视觉图像信息识别出目标物体，在视觉观察中定位目标物体后再由图像点云融合算法计算获得目标的精确位姿，转换坐标系后发布导航任务完成局部环境目标物体导航。

## 第五章 实验设计与结果分析

本章将详细介绍我们所提出的结合全局路径规划方法和局部路径规划方法的 LVL-Nav 方法在仿真环境和真实环境下的实验评估结果，包括在仿真环境中不同方法的对比实验来验证本文所提出方法的有效性、针对提出方法的各关键组件进行消融实验分析各个组件的作用和在真实环境中进行目标物体导航的实验。

### 5.1 仿真环境实验设计与评估

#### 5.1.1 数据集

在仿真实验中，我们使用 Gazebo 中的离线数据集来进行导航实验以测试我们所提出方法的有效性。Gazebo 是一个广泛用于机器人导航控制研究的强大开源仿真环境，它提供了一个高质量的物理引擎来模拟机器人在复杂的环境中的运动、感知、认知和交互，如表 5-1 所示，其中可导航指的是导航代理可以在仿真环境中进行自主移动以探索不同的空间环境；可交互指的是可以手动调整环境中物体的摆放位置或是添加目标物体；物体状态表示环境中的物体会随着时间轴而发生变化，比如钟表会不停地走；动态照明是指随环境中的光源可以动态进行调整，且环境中的物体会随着光源位置的变化而产生不同的阴影；3D 资源库指的该仿真环境提供可以添加在环境中不同物体的社区资源；真实对应指的是所使用的仿真世界不是随机生成而是与真实环境一一对应还原的。相较于其他仿真环境，Gazebo 能够提供更全面丰富的可视化系统和导航功能。

表 5-1 Gazebo 与其他仿真环境对比

仿真环境	可导航	可交互	物体状态	动态照明	3D 资源库	真实对应
Gazebo	✓	✓	✓	✓	✓	✓
iGibson	✓	✓				
Matterport3D	✓					
Minos	✓					
Habitat	✓					

在实验的过程中我们使用了类民住房、类酒店、类工厂 3 种具有不同尺度不同特征的场景，每种类型的场景中都有 5 个不同样式的环境，环境中共存在有 1000 多个物体，且每个环境都有 120 多条自然语言导航指令用于导航测试。自然语言导航指令有单目标、三目标和五目标三种不同的形式，每个环境之下的每种指令都有 40 多条用于测

试导航方法的性能。本文所设计的 LVL-Nav 方法可以选择 32 个不同的物体作为导航目标，包括手提电脑、鼠标、花盆、台灯、沙发等室内常见家具。

在全局路径规划过程中，我们通过搭载在移动机器人上的十六线激光雷达在各仿真环境中构建栅格地图，根据所建的栅格地图设置通路中的导航点，并通过 ROS 导航节点进行导航，记录导航时间构建用邻接矩阵的形式表示的拓扑图。此外，我们获取环境中已经设置的所有导航点的四个方向的位姿信息和对应的视觉观察图像作为全局路径规划导航的依据，用于多模态融合网络模块匹配导航点和目标和执行后续的导航任务。

在局部路径规划过程中，需要获取仿真环境中的特征信息来训练特征提取、特征融合网络模型。具体而言，我们将每个环境都进行网格化，每个网格代表移动机器人能在环境中进行移动的离散动作步，我们将每个到导航点所获得视觉观察送入到预训练的 ResNet18 网络中提取特征，融合将该导航点的坐标和该点的特征作为键值对进行存储，除此之外，我们还将环境中所有的所有目标物体的位姿、名称和可以看到该目标物体的导航点作为键值对进行存储。最终获得由导航点、目标物体和最佳动作共同构成的预训练数据集。在训练时，模型首先获得的目标物体名称和当前导航点位姿，然后它会根据导航点所对应的特征输出一个导航动作并和最佳动作计算交叉熵。在不同环境的测试过程中，我们随机给定不同的初始位姿和导航指令。

### 5.1.2 实验参数

在仿真环境中，我们所搭建的移动机器人的运动参数如下表 5-2 所示。

表 5-2 机器人运动参数

最大线速度	最大角速度	最大线加速度	最大角加速度	线速度分辨率	角速度分辨率
0.6m/s	25°/s	0.2m/s <sup>2</sup>	30°/s <sup>2</sup>	0.01m/s	1°/s

全局路径规划中的多模态融合网络通过在 Interiornet 数据集筛选后的 20w 个数据集进行训练，使用 Adam 优化器以  $10^{-4}$  的学习率更新 CLIDP 网络，我们使用 20 个异步代理训练 5000 次 episode，每次的 batch\_size 为 128，且每 100 次保存一次模型数据和精度，然后将得到的模型在划分的验证集里测试精度，最终选取精度最高的模型作为后续导航所使用的模型。局部路径规划中的特征提取、特征融合模块通过预训练的模型权重进行初始化，使用 Adam 优化器以  $10^{-5}$  的学习率更新模型，我们使用 20 个异步代理训练 300 万次 episode，每 10 万次保存一次模型数据和精度，然后将得到的 30 个模型在验证集里测试精度，最终选取精度最高的模型作为后续导航所使用的模型。

### 5.1.3 实验设备

所有的仿真实验是在笔记本电脑上进行的，模型的训练则是在服务器上进行的，它们的主要配置参数如表所示5-3。我们使用双系统的方式在笔记本电脑上安装 Ubuntu20.04 并搭建 ROS 所依赖的相关环境，使用 VSCode 作为代码编辑、环境配置管理和远程连接工具。在服务器上则使用 docker 容器进行模型训练环境的配置和依赖的安装。

表 5-3 实验平台配置参数

软硬件	服务器配置	笔记本电脑配置
GPU	NVIDIA GeForce RTX 2080Ti × 3	NVIDIA GeForce RTX 3050Ti
CPU	Intel Xeon E5-2620 v4 (32)@3	AMD Ryzen 9 5900HS
内存	128GB	16GB
存储	4T	1T
操作系统	Ubuntu18.04	Ubuntu18.04
开发环境	Python3.8+Pytorch1.9	Python3.8+Pytorch1.9

仿真差速驱动机器人的建模如图5-1所示，机器人的底盘由半径为 10 厘米的圆柱状

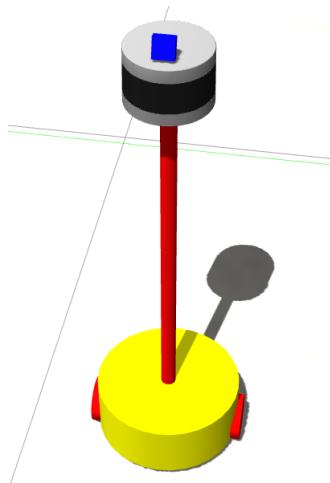


图 5-1 仿真机器人模型

机身、两个宽度为 1.5 厘米的驱动轮和两个球状万向轮搭建而成，底盘上搭载了一个水平扫描范围为 360 度、垂直扫描范围 30 度的十六线激光雷达和单目相机，移动机器人的机身长度为 50 厘米，宽度为 23 厘米。本文在三维动态模拟器 Gazebo 中进行仿真实验，仿真环境和环境中的点云可视化示例如图5-2所示，为了保证后续不同方法能够进

行公平的比较，我们统一使用多线激光 SLAM 算法扫描各个环境进行全局地图的创建，并且在不同方法的实验过程中均不改变移动机器人上所有设备、ROS 导航算法的参数。

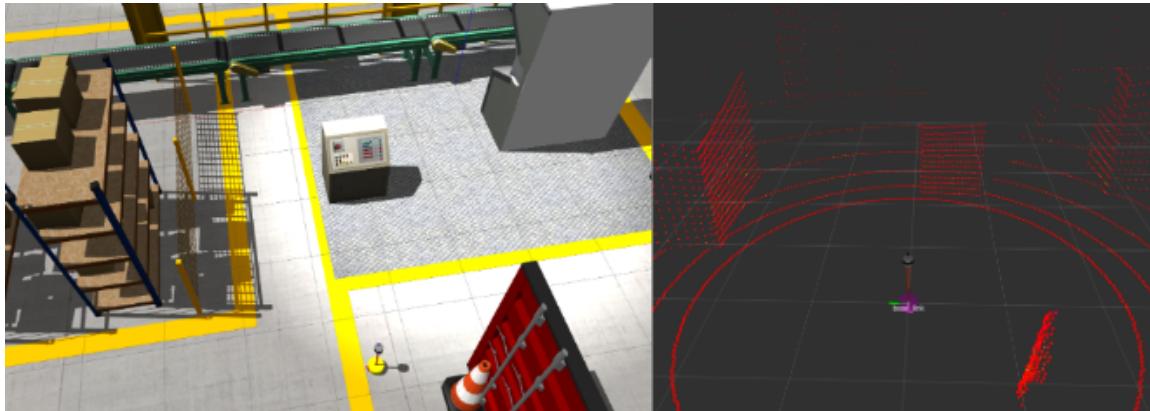


图 5-2 仿真环境与可视化点云

#### 5.1.4 评价指标

本文通过导航成功率 (SR)、导航路径匹配度 (PMD) 和导航效率 (SPL) 三个方面来评估导航方法的综合性能，SR 衡量导航的有效性，MD 衡量导航符合指令要求的程度，SPL 衡量导航效率，其具体表达公式如下。

- (1) **成功率 (Success Rate, SR)**。成功率指移动机器人根据自然语言指令进行导航的过程中，成功完成导航所占的比例，取值范围从 0 到 1。

$$SR = \frac{1}{N} \sum_{n=1}^N S_n \quad (5-1)$$

其中  $N$  表示导航总次数， $S_n$  表示执行第  $n$  次导航的指标，导航成功时值为 1，导航失败时值为 0。

- (2) **路径匹配度 (Path Matching Degree, PMD)** 路径匹配度指移动机器人根据自然语言指令进行导航的过程中，依次经过的目标与指令要求的目标相匹配所占的比例，取值范围从 0 到 1。

$$PMD = \frac{1}{N \cdot M} \sum_{n=1}^N \left( \sum_{m=1}^M a_{nm} \right) \quad (5-2)$$

其中  $M$  表示导航指令中存在的目标个数， $a_{nm}$  表示导航子目标匹配指标，导航点匹配时值为 1，导航点不匹配时值为 0。

- (3) **成功路径长度加权比率 (Success weighted by Path Length, SPL)** 成功路径长度加权比率指移动机器人根据自然语言指令进行导航的过程中，根据实际执行路

径和最优路径的关系从 0 到 1 取值，当实际执行路径与导航最优路径越接近则取值越高，代表导航的效率越高。

$$SPL = \frac{1}{N} \sum_{n=1}^N \left( S_n \cdot \left( \frac{T_n}{t_n} \right) \right) \quad (5-3)$$

其中  $T_n$  表示完成第 n 次导航时所需的最短时间， $t_n$  表示完成第 n 次导航时的实际耗时。

## 5.1.5 对比实验

### 5.1.5.1 对比方法

- (1) Random Policy：随机策略模型的代理在随机的时间步中，根据平均的概率分布去选择动作空间中的动作进行执行，或是直接在环境中停止以结束导航任务。
- (2) Baseline：基准导航策略直接将局部路径规划中的特征提取得到的局部、全局和目标特征直接送入 LSTM 网络进行离散动作推理。用于对比验证本文所提出方法中特征融合模块的作用。
- (3) CLIP<sup>[72]</sup>：CLIP 基于对比学习的方法提出了一种多模态融合的网络结构，通过有效地将图像和文本信息结合在一起提供更强的跨模态理解能力，克服了传统模型依赖单一数据源的缺陷，使移动机器人能够理解环境中的目标物体和语义信息以指导导航任务的执行。
- (4) EONS<sup>[41]</sup>：EONS 借助 ROS 框架提出了一种有效的目标导航策略。首先对常见的室内物体进行语义关联分析，利用 Mask R-CNN 和残差连接网络建立物体语义关联模型，通过 ROB-SLAM 系统方法构建了一个高可用性的环境图，在移动机器人导航的过程中寻找可达的最优路径以完成导航。
- (5) ViNG<sup>[42]</sup>：与传统的基于地图或几何推理的导航方法不同，ViNG 提出了一种创新的视觉语言导航方法，通过学习如何从视觉目标图像中推断导航策略和先前观察到的数据构建环境的拓扑图，通过导航点建议 (Waypoint Proposal)、拓扑图剪枝 (Graph Pruning) 和负挖掘 (Negative Mining) 多种核心策略，使机器人能够根据目标图像和环境拓扑图在现实世界环境中实现高效率的自主导航。
- (6) LM-Nav<sup>[43]</sup>：LM-Nav 通过三个大型预训练模型协同合作，完成了移动机器人根据自然语言指令进行导航的全部流程，包括负责将用户输入的自然语言指令转化为具有顺序的目标序列大语言模型 GPT3、负责将目标序列与环境地图中的

节点进行特征关联以建立具有先后导航顺序的导航点视觉语言模型 (VLM) 和负责构建环境拓扑图并利用凸优化技术规划从起点到终点最优路径的视觉导航模型 (VNM)。

(7) LVL-Nav(Ours): 提出了一种多模态融合的导航方法，将导航过程拆分成全局和局部路径规划两个部分。通过引入深度信息优化传统的视觉语言模型，提出一种方位优化算法筛选不符合方位条件的冗余导航点，结合导航点规划算法进一步提高全局路径规划中导航点的正确率，再通过特征提取、融合网络在局部未知环境中进行探索，最后由图像点云融合模块完成目标物体导航任务。

### 5.1.5.2 对比结果

在和近三年流行方法的对比试验如表5-8所示。本文所提出的方法的导航指标在五目标的导航测试指令中的成功率 SR(77.4%)、路径匹配度 PDM(77.4%) 和成功路径长度加权比率 (62.1%) 三个指标中的性能表现都是最好的。在与五目标指令的基准模型的对比中，本文所提出方法的成功率 SR(+22.9%)、路径匹配度 PDM(+23.5%) 和成功路径长度加权比率 SPL(+22.6%) 三个指标都大大超过了基准模型。在五目标指令实验中与当前表现最好的 LM-Nav 模型相比，本文的方法在成功率 SR(+1.9%)、路径匹配度 PDM(+3.5%) 和成功路径长度加权比率 (+8.7%) 三个指标上也都有所提升，这是因为我们在全局路径规划中通过引入深度信息优化多模态融合网络，提高模型在光线条件变化大的室内的鲁棒性，提出了一种方位优化法，结合指令中提供的方位信息与移动机器人的实时位姿，筛选不符合方位条件的冗余导航点，再通过导航点规划算法进一步提高全局路径规划所生成导航点的正确率，然后在局部路径规划中我们设计了一种特征提取、融合网络能够有效地在局部未知环境中进行探索，设计并实现了一种单目相机和多线激光传感器融合的局部导航方法，以完成导航至目标半米内的闭环任务，这表明了我们所提出方法的有效性。

相比于我们所提出的方法，CLIP 将图像和文本嵌入同一空间中来判断导航目标，但在动态变化的室内环境中容易发生物体移动、光线变化等情况，或是环境中的多个位置都存在相同的目标物体时，该模型缺乏足够的动态适应和区分不同方向正确目标的能力，无法有效地将导航点与目标做出正确的匹配从而导致导航失败，这导致该方法在导航成功率方面的指标表现不佳。

EONS 借助构建的语义关联模型在环境中通过直接寻找目标物体或是与目标最相关

表 5-4 仿真环境实验定量结果对比

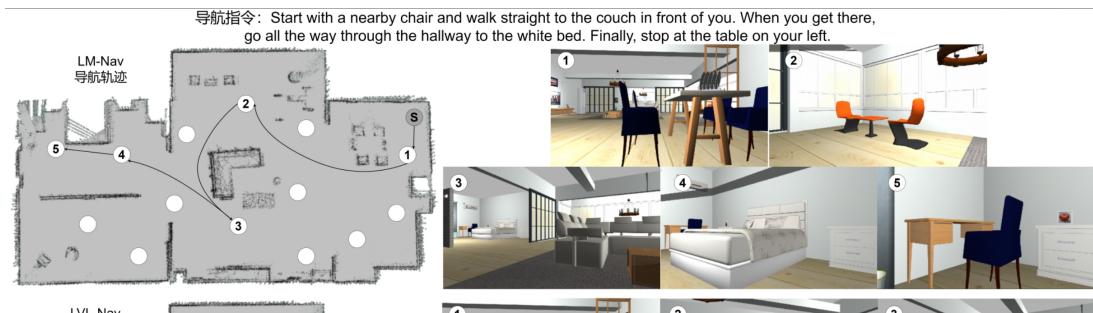
Method	I=1			I=3			I=5		
	SR	PMD	SPL	SR	PMD	SPL	SR	PMD	SPL
Random	6	6	3.9	0	0	0	0	0	0
CLIP	66.2	66.2	65.3	42.2	43.7	64.9	32.7	36.5	60.6
Baseline	73.1	73.1	48.3	67.4	69.5	45.8	51.8	53.9	39.5
EONS	78.5	78.5	31.7	75.0	76.8	29.4	64.1	65.5	26.2
ViNG	82.6	82.6	63.3	79.3	78.9	62.2	70.7	71.1	59.0
LM-Nav	85.3	85.3	57.6	79.4	77.1	54.9	72.8	73.9	53.4
<b>Ours</b>	<b>88.4</b>	<b>88.4</b>	<b>66.7</b>	<b>82.3</b>	<b>82.9</b>	<b>65.2</b>	<b>74.7</b>	<b>77.4</b>	<b>62.1</b>

联的物体从而引导代理执行导航任务。但这种方法依赖于训练过程中环境物体之间的固定关联信息，在静态的环境中表现较为出色，但对于存在多种移动物体的动态环境中，这种目标导航策略无法及时更新目标位置和状态，从而引发导航的错误。除此之外，当 EONS 在导航过程中未能第一时间观察到目标时，会根据关联模型寻找关联性最强的物体进行探索，这种探索策略占用了大部分的导航时间，导致该方法在导航效率方面的指标表现不佳。

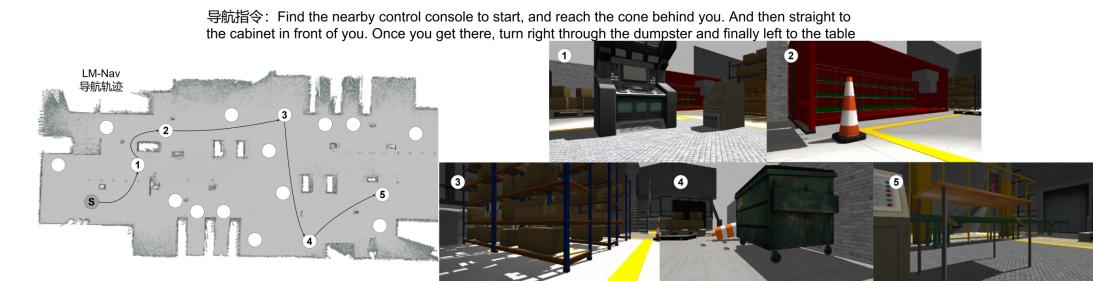
ViNG 通过结合视觉目标学习和构建环境拓扑图实现了基于视觉的目标导航，因此在导航效率指标上取得了不错的成绩，但是它过度依赖于视觉目标的识别的定位，在很有可能出现遮挡或曝光、欠曝的复杂的室内环境中容易导致目标识别失败，模型无法正确的指导代理进行有效的探索从而导致导航任务的失败。这导致该方法在导航成功率方面的指标表现不佳。

LM-Nav 通过大语言模型有效的解析导航任务，利用大量导航数据训练视觉语言模型和视觉导航模型，在多种复杂环境中指导代理生成正确的导航动作，这种联合多种大模型在环境中执行导航任务的方式具有一定的有效性，但这种导航方法在执行任务的过程中依赖于模型的推理速度，等待各种大模型响应的时间占据了部分的导航时间，因此它的导航效率并不高。

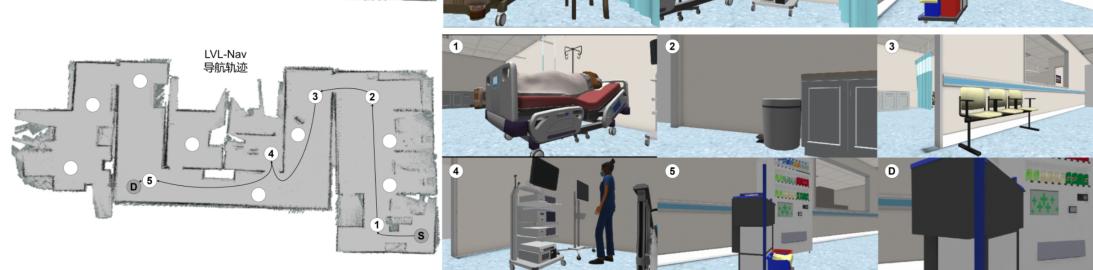
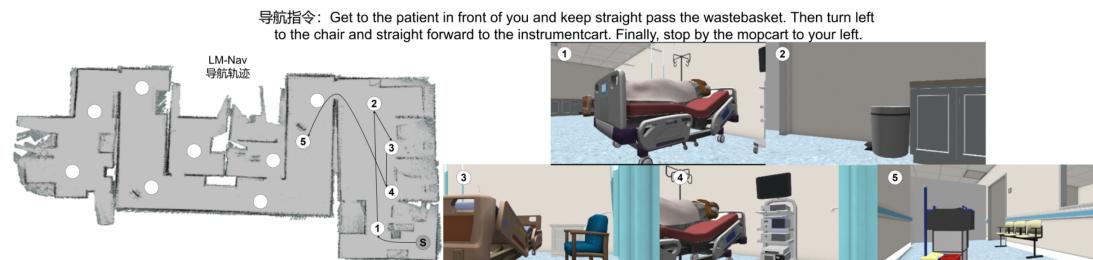
LM-Nav 在实验中取得了次优的结果，该方法与我们所提出的 LVL-Nav 导航方法在根据不同指令执行导航任务时的轨迹结果如图 5-3 所示，其中 S 表示随机初始化的移动机器人起点，标号 1, 2, ..., 5 分别表示算法根据指令所执行经过的子目标导航点，D 表



a) 类家庭仿真环境下的导航



b) 类工厂仿真环境下的导航



c) 类医院仿真环境下的导航

图 5-3 LM-Nav 和 LVL-Nav 路径对比图

示本文提出方法中局部路径规划获得的导航终点。具体来说，在类家庭仿真环境5-3a中 LM-Nav 方法无法正确匹配到曝光状态下的沙发，引导移动机器人移动至错误的导航点，而本文提出的 LVL-Nav 则通过引入深度信息提高模型在曝光或欠曝的室内环境中的鲁棒性，在全局路径规划中正确地匹配到了曝光状态下的沙发；在5-3b、5-3c中 LM-Nav 方法无法正确避免环境中所存在的其他相似物体影响导航的规划，当指令要求移动到路障或椅子旁时，将移动机器人引导至错误的导航点，而本文所提出的方位优化法能够结合指令中提供的方位信息与移动机器人的实时位姿，筛选不符合方位条件的冗余导航点，在全局路径规划中正确地匹配指令所指方向的目标。图5-3和表5-8统计的结果表明，本文所提出的方法不依赖于导航点的初始化，能够在动态变化复杂环境中指导代理正确地匹配导航点，在局部路径规划部分高效执行探索动作并可靠地完成导航至目标半米内的闭环任务。

在对比实验的过程中，我们把本文全局路径规划中的多模态融合方法和 LM-Nav 方法在训练过程中每 200 个 epoch 保存一次的模型在验证集上测试精度，如图5-4所示。将两种方法在测试集上的成功率 SR 和路径匹配度 PMD 进行对比，可以发现本文提出的方法在经过足够的训练后的表现优于 LM-Nav。本方法的模型中添加了深度图特征提取架构，学习的速度会更慢，我们的方法在第 1800 个 epoch 时的 SR 只有 59.8%，而 LM-Nav 方法则达到了 68.9%(+9.1%)。此外，我们的方法在训练的过程中收敛更慢，在第 2400 个 epoch 时 LM-Nav 方法基本收敛到精度最高值，而我们的方法则在 3400 个 epoch 时候才基本收敛到精度最高值。另一方面，我们的方法在 3400 次 epoch 后得到的成功率最高值 74.7%(+1.9%) 比 LM-Nav 方法的成功率最高值 72.8% 在测试集上的表现会更好；我们的方法在 1800 次 epoch 后得到的路径匹配度比 LM-Nav 方法的路径匹配度在测试集上表现得更好，并在第 3400 次 epoch 后得到路径匹配度最高值 77.4%(+3.5%) 比 LM-Nav 方法得路径匹配度最高值 73.9% 匹配度更高，这得益于我们提出的多模态融合网络框架能够结合深度特征信息归纳表征视觉观察和目标物体文本特征，能够在复杂光照条件下正确通过利用深度信息进行图像-文本匹配，从而提升导航的成功率。

### 5.1.6 消融实验

为了逐步验证本文所提出的多模态融合网络、导航点规划、方位优化和局部路径规划算法的有效性，我们对上述各关键模块进行仿真消融实验。我们在仿真环境下进行的消融实验中依次添加多模态融合模块、导航点规划算法、方位优化算法、局部路径规划算法来测试各个模块对于指导代理进行高效且正确地导航的有效性，根据他们分别执行

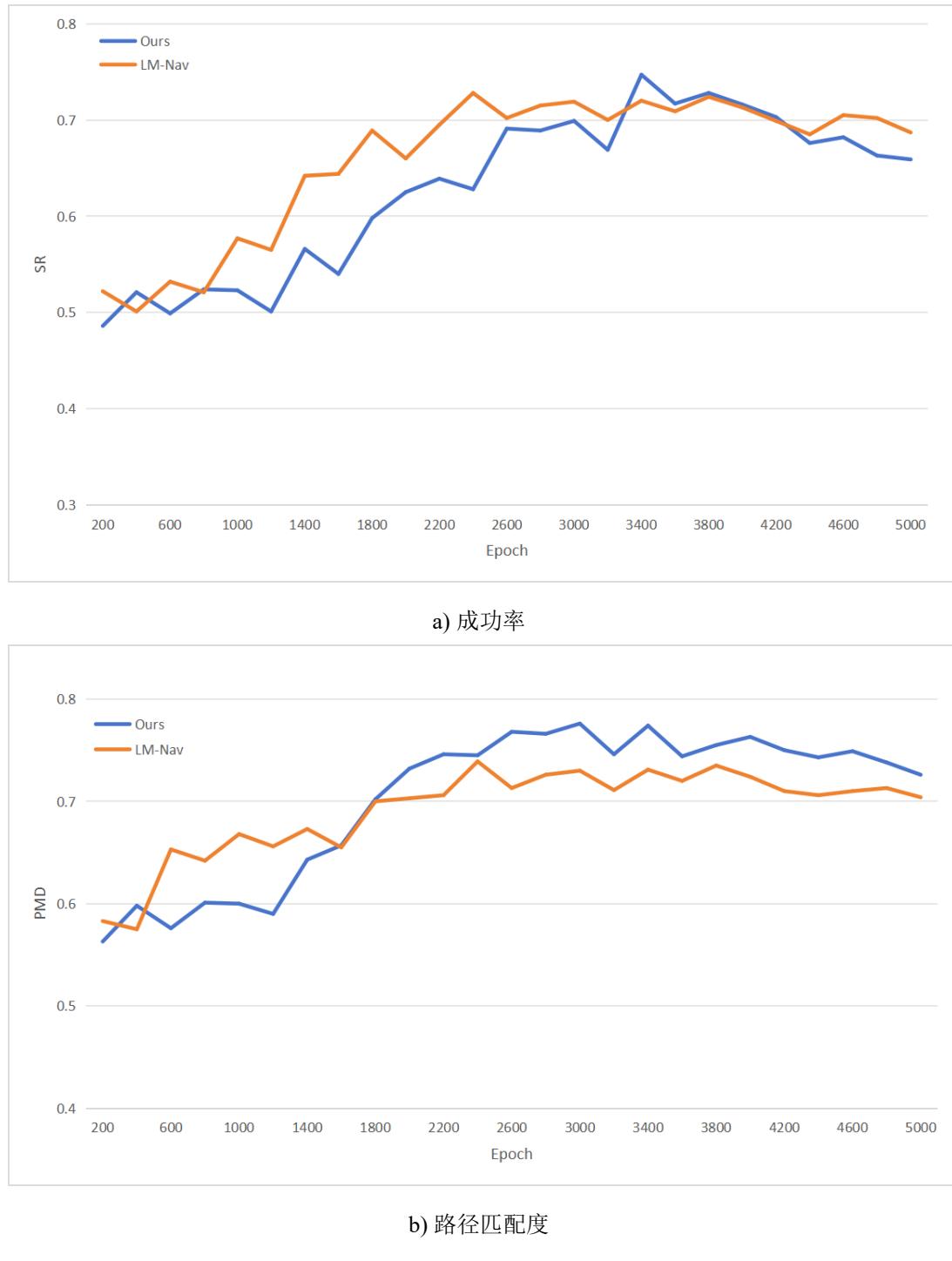


图 5-4 导航指标变化图

基于自然语言指令后记录的评价指标结果如表5-5所示，我们分析了各模块的作用。

(1) 多模态融合模块：我们直接使用多模态融合网络进行指令目标与导航点之间的匹配，并将匹配的结果直接发布至/move\_base\_simple 话题之中交由 ROS 导航节点执行导航。实验结果表明在单目标指令中成功率 SR(-22.2%) 成功路径长度加权比率 (-1.4%)，在五目标指令中成功率 SR(-25.1%)、路径匹配度 PMD(-25.5%)

表 5-5 消融实验结果

Method		+ 多模态融合	+ 导航点规划	+ 方位优化	+ 局部路径规划
I=1	SR	66.2	78.3	83.4	<b>88.4</b>
	SPL	65.3	54.0	62.7	<b>66.7</b>
I=5	SR	49.6	60.9	68.1	<b>74.7</b>
	PMD	51.9	61.7	69.4	<b>77.4</b>
	SPL	59.3	46.6	57.2	<b>62.1</b>

和成功路径长度加权比率 (-2.8%)，由于缺失了导航点规划算法所引入的加权最短距离算法，忽略了自然语言指令导航所具有循序渐进的特性，使该导航方法只能锁定环境中与指令目标最相像目标从而导致导航的失败。

- (2) 导航点规划算法：我们在全局路径规划中的多模态融合模块基础之上引入了导航点规划算法，将算法匹配的导航点序列交由 ROS 节点执行导航。实验结果表明单目标指令中成功率 SR(-10.1%) 成功路径长度加权比率 (-12.7%)，在五目标指令中成功率 SR(-13.8%)、路径匹配度 PDM(-15.7%) 和成功路径长度加权比率 (-15.5%)，导航点规划算法能依赖于指令目标与环境匹配的相似度和拓扑图中的距离进行规划，相比单纯的多模态融合算法在导航成功率上提升了很多，但当环境中的不同位置存在多个相同目标的环境时，该导航方法无法根据指令信息筛选掉错误的导航点从而导致导航的失败，且该规划算法计算开销大，导航效率很低。
- (3) 方位优化算法：我们将完整的全局路径规划算法直接进行实验，将算法匹配的导航点序列交由 ROS 节点执行导航。实验结果表明单目标指令中成功率 SR(-5.0%) 成功路径长度加权比率 (-4.0%)，在五目标指令中成功率 SR(-6.6%)、路径匹配度 PDM(-8.0%) 和成功路径长度加权比率 (-4.9%)，首先方位优化算法能够有效的剔除不再指令要求方位上的相似目标，该方法在具有多个相同目标的环境中进行导航尤为出色，并且方位优化算法能提前筛选大部分的待规划导航点以减少导航点规划算法的计算开销，大幅地提高了导航效率。但这种方法受限于环境中导航点的初始化设置，当导航点中未能出现目标或是导航点未能设置在目标半米之内的时候会导致导航的失败。
- (4) 局部路径规划：添加了由特征提取、融合和图像点云融合模块的局部路径规划

方法让代理能够在导航点中未能出现目标的情况下进行局部范围内的自主探索，并在探索的过程中锁定目标，然后通过图像点云融合模块将激光雷达感知和视觉观察认知信息相结合，以获得目标的最终位姿，在进行坐标变换之后发布导航即可完成完整的目标物体导航任务。

消融实验的结果表明多模态融合模块、导航点规划算法、方位优化算法、局部路径规划算法均能有效地提高移动机器人的导航成功率、导航匹配度和导航效率，且随着指令中目标数量的增多，本文提出的 LVL-Nav 方法具有更高的导航匹配度和导航成功率。

在消融实验过程中取出一组五目标指令下进行导航的结果如图5-5所示。图中5-5a表示仅使用多模态融合网络根据五目标指令所执行的路径和导航到各子目标所捕获的图像，机器人无法利用环境中的距离和方位信息选取导航点、无法区分不同位置的相同目标，错误地选取了1、5两个相同导航点，且无法导航至目标半米内。5-5b为仅使用多模态融合和导航点规划算法根据相同的五目标指令所执行的导航，机器人无法利用指令中的方位信息，错误地选择左前方最近的椅子作为第一个导航点。5-5c为仅使用多模态融合、导航点规划和方位优化算法根据相同指令所执行的导航，移动机器人能够正确地通过多模态融合网络进行匹配，再利用导航点规划和方位优化进行筛选，得到了正确的导航点序列，但受限于导航点位姿的初始化，无法完成导航至目标半米内的任务。5-5d为添加了局部路径规划方法所执行的正确导航，全局路径规划得到了正确的导航点序列，局部路径规划完成导航至目标半米内的任务。

对比仿真消融实验结果表明，本文所提出的多模态融合网络、导航点规划、方位优化和局部路径规划算法均能十分有效地帮助移动机器人理解指令目标、利用环境信息，使移动机器人能够根据指令高效且正确地执行导航。

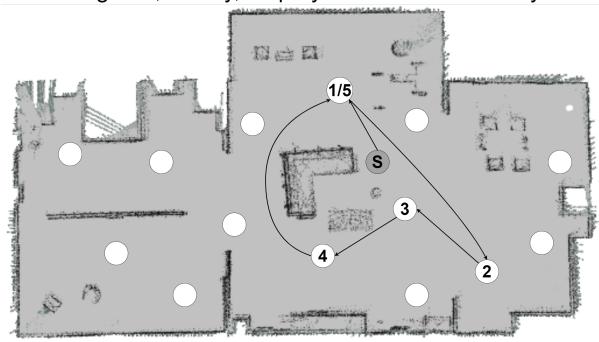
## 5.2 真实环境实验设计与评估

由于仿真环境实验相对理想，移动机器人在导航的过程中不受过多传感器测量误差与其他外界因素干扰，且无法完全模拟真实机器人的计算性能，故需要真实环境实验进一步验证本文提出方法的有效性。

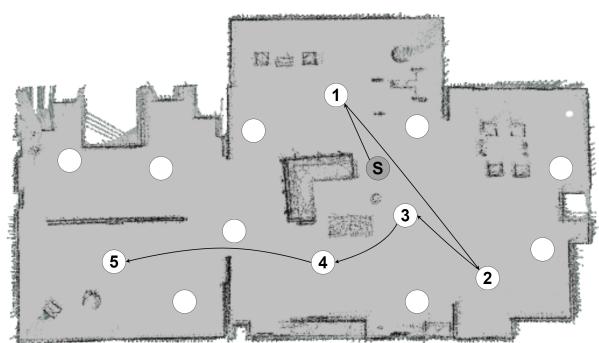
### 5.2.1 实验设备

为进一步验证所提出算法在实际应用中的可行性，选用搭载了镭神 C16 激光雷达和杰锐微通 HF899 单目相机的灵邀移动机器人进行真实环境实验，其中激光雷达和用于进行导航实验的移动机器人如图5-6、5-7所示。

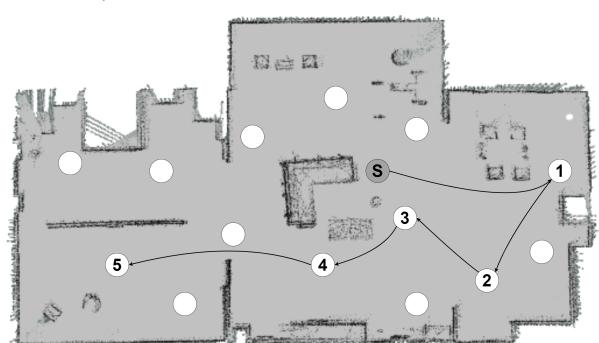
Start with the nearest chair in your right, get to the door on your left. Then turn right to the sofa and go through a foldingdoor, Finally, stop by the chair in front of you.



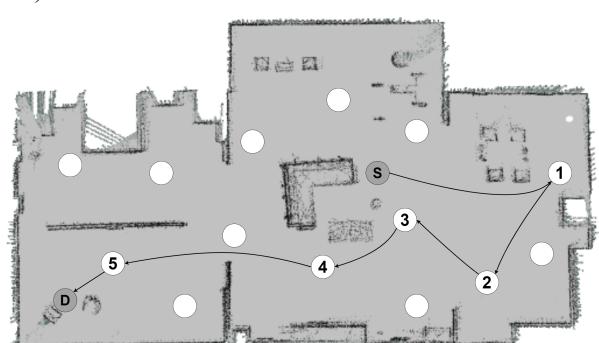
a) 多模态融合导航路径



b) 多模态融合 + 导航点规划导航路径

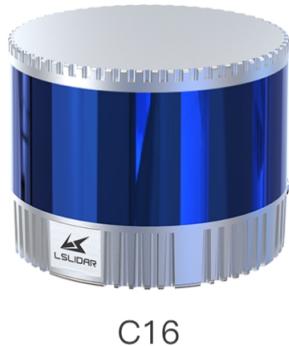


c) 多模态融合 + 导航点规划 + 方位优化导航路径



d) LVL-Nav 方法导航路径

图 5-5 消融实验结果



C16

图 5-6 镍神 C16 激光雷达

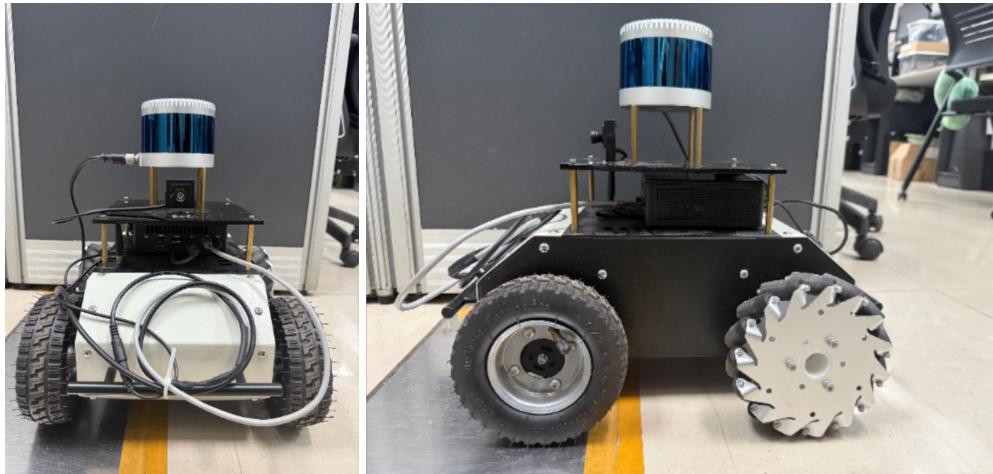


图 5-7 灵邀移动机器人

如表5-6所示，我们将全局路径规划、ROS 导航控制系统部署到 UP Xtreme i11 嵌入式开发板上；将局部路径规划中的特征提取、融合网络模型部署在 Jetson TX 嵌入式开发板上；灵邀机器人搭载码盘系统，通过 IMU 等装置实验机器人的移动；镍神 C16 激光雷达负责全局路径规划中的建图、导航及局部路径规划中的图像点云融合；杰锐微通 HF899 单目相机负责局部路径规划中提供第一视觉视觉观察用于特征提取、融合和图像点云融合；其中 UP Xtreme i11 和 Jetson TX 嵌入式开发板如图5-8所示，设备的参数如表5-7所示。

### 5.2.2 实验环境

本文的真实实验场景搭建在学院楼内的实验室，总面积约有 45 平方米。实验场景包含成排的桌子、一个饮水机、一个打印机、一个冰箱、两个书架、几张椅子，布局设计成室内办公桌的样式，真实实验环境及其栅格地图如图5-9所示。我们从 32 个室内常见物体中随机选择多个连续的指令目标在真实环境中进行实验，包括冰箱、打印机这种不存在复数的目标，也包括椅子、笔记本电脑等这种存在复数的目标。真实环境中的实

表 5-6 硬件系统设计

设备	功能
UP Xtreme i11	部署全局路径规划、驱动机器人
Jetson TX	部署导航模型进行探索推理
灵邀机器人	移动机器人
镭神 C16 激光雷达	建图、导航及图像点云融合
杰锐微通 HF899 单目相机	获取视觉观察



a) UP Xtreme i11

b) Jetson TX

图 5-8 嵌入式开发板

验直接使用仿真环境中由全局路径规划和局部路径规划共同构成的 LVL-Nav 方法进行测试，不在现实测试环境中采集数据来重训练模型，同时也不改变各算法参数，以此来验证所提出的方法在真实环境下的迁移效果，并找出方法的优势和局限。

真实实验的运行环境为内存 16GB 的 64 位 Ubuntu18.04 操作系统，平台为 ROS Melodic。通过局域网完成网络连接配置后，使用笔记本电脑执行远程操作，控制机器人完成建图、记录导航点与图像并根据指令执行导航。在实验的设计上，本文采用随机抽取并随机排序的方法在上述的目标物体中进行选择，在实验环境中也随机布局这些被选中的物体。当移动机器人的局部路径规划在执行完成图像点云融合算法并输出“导航结束”时作为移动机器人完成一次导航的信号，若在导航的过程中正确地根据指令经过所要求的所有目标，并在最后的目标半米内停止则表示本次真实环境下的导航成功，除

表 5-7 硬件设备参数

软硬件	UP Xtreme i11	Jetson TX2
GPU	Intel(R) Iris(R) Xe Graphics	256-core NVIDIA Pascal architecture GPU
CPU	11th Gen intel r core(tm) i7-1165G7	Dual-core NVIDIA Denver 2 64-bit CPU Quad-core Arm Cortex-A57 MPCore processor
AI 性能	1.0TFLOPS	1.33TFLOPS
内存	2GB	8GB
存储	24GB	32GB
功率	65w	7.5w-15w

此之外，在整个导航的过程中若发生过碰撞或导航超时则都判定为导航失败。

### 5.2.3 导航实验

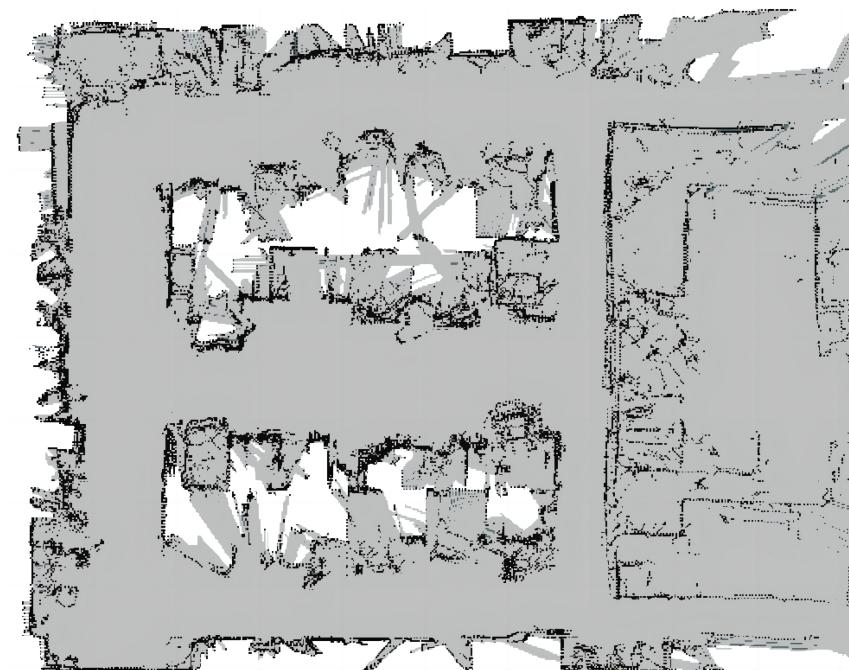
实验中先将单目相机和激光雷达通过张正友标定法进行联合标定，得到的外参数据如式5-4所示，我们发现仅依靠由多模态融合、导航点规划和方位优化组成的全局路径规划已经能较为可靠地依照自然语言指令进行导航点匹配，但仍然存在指令目标移动导致目标不在导航点旁的情况，从而导致仅依靠全局路径规划导航任务的失败。局部路径规划可以在局部环境中通过将视觉观察进行特征融合、特征提取进而实现探索，当目标出现在视觉观察中时可以在图像点云融合模块中结合目标检测、点云聚类和坐标变换方法将空间点云数据映射到视觉图像之中，映射结果如图5-10所示，包括来自于图像目标检测结果的物体分类信息和来自于十六线激光雷达通过聚类后物体相对于激光源的距离信息，如若指令需要导航到目标“person”的附近，代理则会在局部探索到目标后融合的结果将点云平均距离作为目标与移动机器人之间的实际距离，再根据平面成像原理依次计算出目标在机器人坐标系和 map 坐标系下的具体位姿。最终通过发布导航任务以完成局部路径的规划，实现导航到目标半米内的闭环任务。

$$R = \begin{bmatrix} -0.0419 & 0.0487 & 0.9979 \\ -0.9985 & -0.0379 & -0.0401 \\ 0.0359 & -0.9981 & 0.0502 \end{bmatrix}, t = \begin{bmatrix} 0.1549 \\ -0.0253 \\ 0.0952 \end{bmatrix} \quad (5-4)$$

为了验证本文所提出方法在复杂多变真实环境中依照指令进行导航的能力，我们给定只在方位和目标上存在细微差距的单目标、三目标和五目标导航指令进行实验，其中



a) 真实实验环境

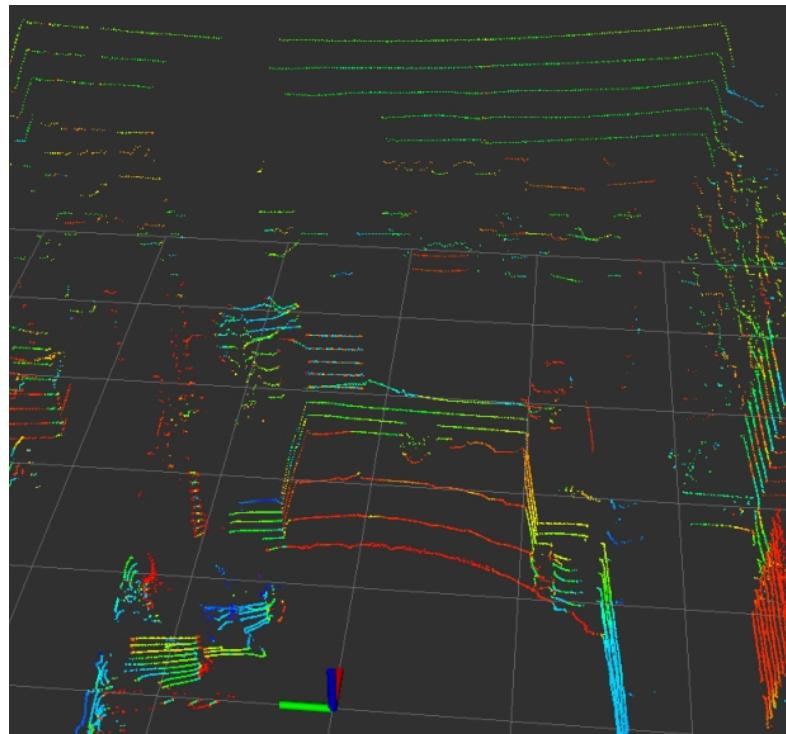


b) 真实实验环境栅格地图

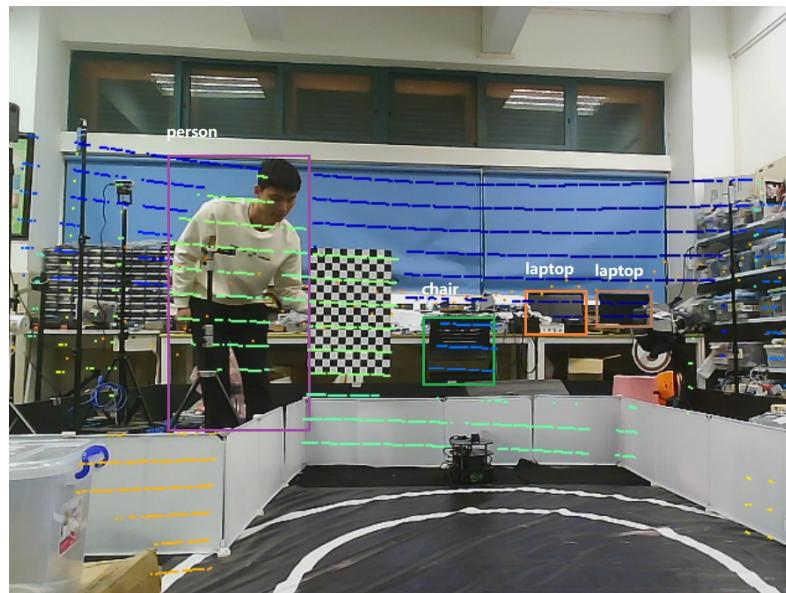
图 5-9 真实环境

五目标的实验示例如图5-11和5-12所示。

前者是一次真实环境成功导航示例，学院楼实验室的门口出发，根据全局路径规划获得的目标导航点之后，在 ROS 中发布目标位姿依次进行目标物体导航，在直行经过储物柜之后移动机器人根据匹配的目标点继续导航到前方的冰箱旁，然后右转依次经过书架和风扇这两个导航点，接着得益于全局路径规划中的方位优化和多模态融合网络，



a) 激光点云可视化



b) 融合效果图

图 5-10 图像点云融合可视化

在最后目标椅子的导航点选择中代理不会错误地选择在环境中离自己最近的风扇旁的其他椅子，而是前往指令要求的前方最近的椅子，从而正确匹配离工作台椅子旁最近的导航点，最后根据局部路径规划中的特征提取融合和运动模块在视觉观察中进行探索，检测到最终目标后通过图像点云融合模块获得环境中椅子的真实位姿，发布位姿后执行以顺利完成导航任务。

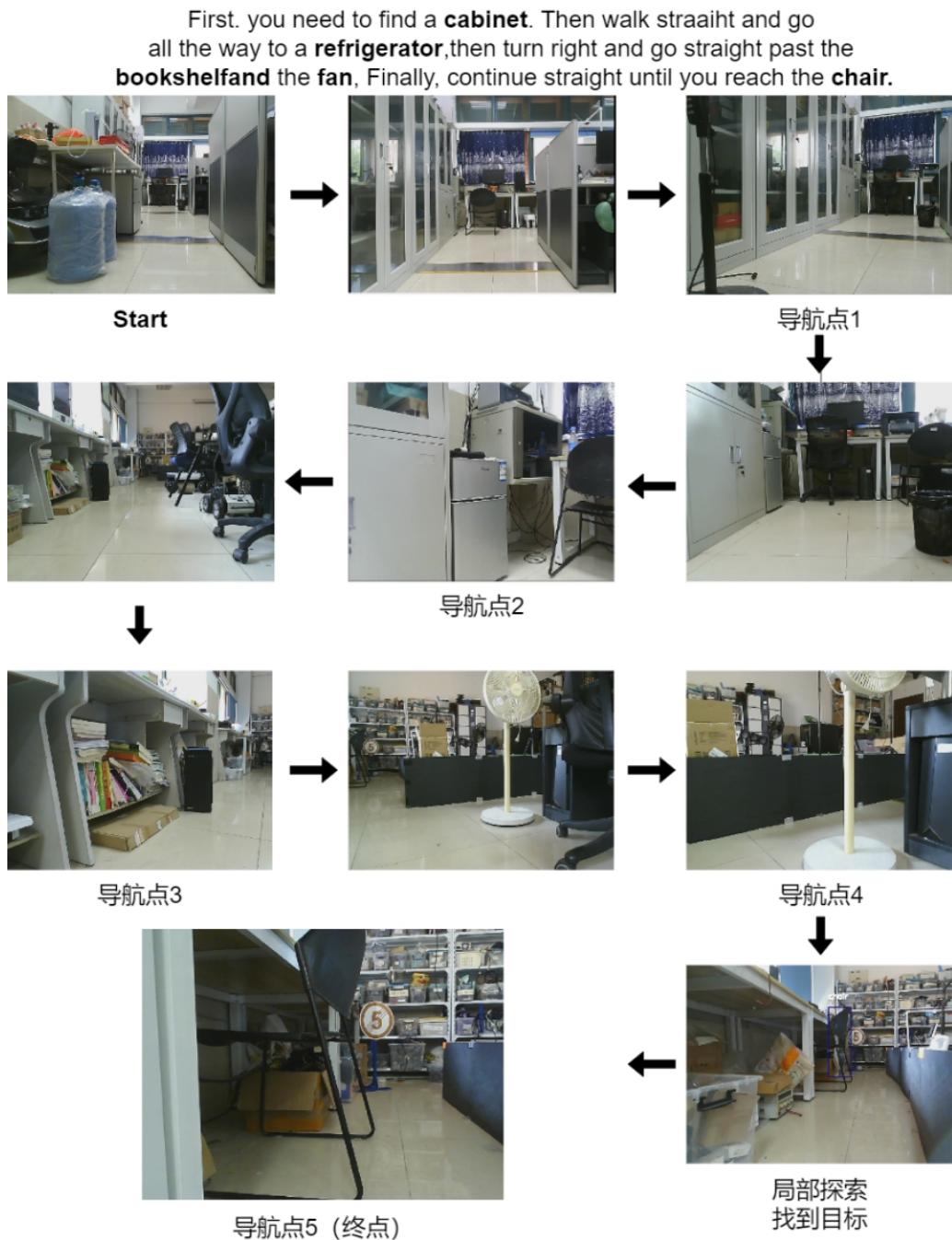


图 5-11 真实环境成功导航示例

后者是一次真实环境失败导航示例，我们特意选用只具有细微差别但完全不同的导航指令来测试代理理解指令并正确执行导航的能力，所以同样地在学院楼实验室的门口出发，根据全局路径规划获得的目标导航点之后，在 ROS 中发布目标位姿依次进行目标物体导航，在直行经过储物柜之后根据方位指示前往右方，在经过有很多椅子较为狭窄的环境后导航到双肩背包旁，接着左转经过摆放着水杯的桌角导航到同一个风扇旁，但经过我们的实验发现，移动机器人在 50cm 这种十分狭窄的环境中无法跨过风扇脚驶出狭窄环境，最终在通道里调整姿态并因为导航超时导致导航失败，这是由于在真实的

First, you need to find a **cabinet**. Then turn right and go all the way to a black **backpack**, then turn left and go straight past the **water cup** and the **fan**. Finally, after taking a right turn next to a **chair**.

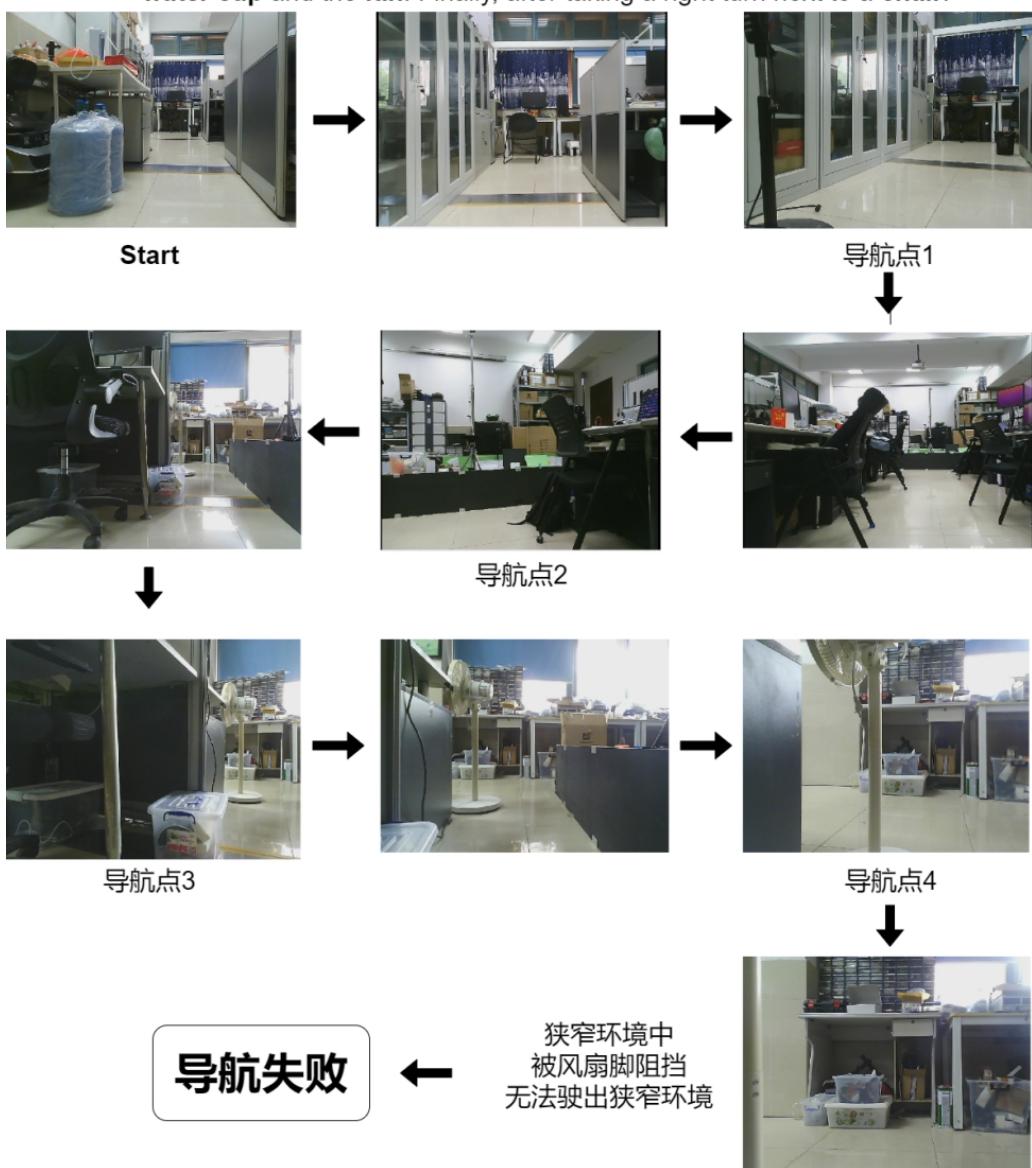


图 5-12 真实环境失败导航示例

室内环境存在大量不变移动变化的障碍物，这导致移动机器人无法十分有效地在狭窄的环境中进行可靠地导航。

本文在真实实验场景中针对不同的自然语言导航指令分别进行实验，目标物体从“mouse”“laptop”等室内常见物体中进行随机选择，目标物体摆放在真实场景中的随机可见位置，和近三年流行方法的实验定量结果对比如表5-8所示。本文所提出的方法在真实环境五目标的导航过程中成功率 SR(60.1%)、路径匹配度 PDM(64.5%) 和成功路径长度加权比率 SPL(52.7%) 三个指标中的性能表现都是最好的，这验证了我们所提出方法在真实环境中进行导航的有效性。真实环境下的实验结果表明本文提出的导航方法以

表 5-8 真实环境实验定量结果对比

Method	I=1			I=3			I=5		
	SR	PMD	SPL	SR	PMD	SPL	SR	PMD	SPL
EONS	67.0	67.0	38.9	51.8	53.5	23.7	47.4	49.2	21.5
ViNG	69.4	69.4	53.7	65.2	66.3	49.4	55.2	56.8	48.5
LM-Nav	74.2	74.2	48.3	66.7	69.1	44.6	58.8	60.9	43.3
<b>Ours</b>	<b>76.9</b>	<b>76.9</b>	<b>56.0</b>	<b>68.3</b>	<b>71.4</b>	<b>54.9</b>	<b>60.1</b>	<b>64.5</b>	<b>52.7</b>

可靠且高效地根据任意形式的自然语言指令和环境拓扑信息，在光照条件变化大的复杂室内环境中融合感知和认知信息正确规划出全局路径，具备区分导航指令中的细微方向、目标差别的能力，同时局部路径规划允许移动机器人不依赖于环境中导航点位置的设置，使其能够正确的导航到目标半米内。

另一方面，我们在测试的过程中也发发现了方法中的不足之处。由于真实环境和仿真环境的差异问题导致了局部路径规划中的特征提取、融合模型性能的下降，当目标未能出现在移动机器人的视觉观察之中时，局部探索方法在存在遮挡物体的环境中存在失败的可能从而导致导航的失败；并且由于真实的室内环境摆放大量的障碍物，这导致移动机器人无法十分有效地在狭窄的环境中进行可靠地导航，在与椅角或桌角发生碰撞后无法继续导航从而导致导航的失败。

### 5.3 本章小结

本章主要介绍了本文所提出的语言视觉激光多模态融合的机器人导航方法在仿真、真实环境进行目标物体导航的实验，我们通过和近三年表现良好方法的对比试验验证了本文所提出方法的有效性，即通过引入深度信息优化多模态融合网络，提高模型在光线条件变化大的室内的鲁棒性；提出了一种方位优化法筛选不符合方位条件的冗余导航点；通过导航点规划算法进一步提高全局路径规划所生成导航点的正确率；设计了一种特征提取、融合网络能够有效地在局部未知环境中进行探索，设计并实现了一种单目相机和多线激光传感器融合的局部导航方法以完成导航至目标半米内的闭环任务；通过消融实验分析并验证了所提出方法中各个模块的有效性。

实验结果表明，我们所提出的方法在仿真、真实环境中的导航表现基本达到了预期，但受限于真实、仿真环境之间的差异以及真实狭窄环境对局部探索导航和基于目标

点导航两方面的影响，移动机器人在真实环境中也存在导航失败的问题亟待解决。

## 结 论

本文主要研究语言视觉激光多模态融合的机器人导航方法的研究及实现，研究方向是多模态融合的目标物体导航方法，并设计一个导航系统，在移动机器人上部署全局和局部路径规划算法，并在仿真环境和真实环境中进行目标物体导航实验。

本文的目标物体导航是在环境中根据多模态的数据到达预期的目标物体。现有的工作通常通过建图来标记环境中存在的目标位置，或是通过训练深度强化学习模型作为代理实时预测动作，以到达指定目标。但上述的方法无法通过视觉信息进行自主探索，忽略了激光雷达所获取的感知信息对于导航的约束和指导，从而导致系统导航成功率低、导航效率低和无法导航到目标半米内的问题。针对以上问题，提出一种多模态融合的目标物体导航方法，具体来说，该方法将导航任务拆分成全局路径规划和局部路径规划两个部分。在全局路径规划中，标记地图中的导航点，保留其位姿、图像、点云图和各点之间的拓扑信息，通过多模态融合网络得到各导航点与目标的匹配权值，结合 dijkstra 算法和方位优化算法规划出全局路径导航点序列。然后，在局部路径规划中，将多线激光与单目相机进行联合标定，结合目标检测、点云聚类和坐标变换方法得到目标具体位姿，发布导航任务以完成局部路径的规划，实现导航到目标半米内的闭环任务。Gazebo 数据集上的实验表明，该方法在测试环境中优于最先进的方法，实验结果证明了该方法的有效性和效率。多模态融合的导航系统具有环境感知、目标认知、路径规划和自主导航四个方面的功能，可以完成真实环境下目标物体导航任务。

总的来说，本文的主要贡献如下：

- (1) 本文提出了一种全局路径规划导航方法。与前人的工作相比，针对静态目标导航任务所提出的全局路径规划导航方法基于单目相机、激光雷达等多种传感器和基于多模态特征融合神经网络，增强系统对当前环境和导航过程中的认知和感知能力，再通过方位优化算法筛除噪声导航点，提高导航点选择的正确率的同时提高后续规划的计算响应速度，最后通过导航点规划算法加权融合多种策略进一步提高导航的准确率和导航效率。实验结果表明该方法具有一定的有效性和优越性。
- (2) 本文提出了一种未知环境的目标物体探索方法。与前人的工作相比，针对动态目标导航任务所提出的局部目标物体探索方法基于多特征提取和融合的方法，在同一嵌入空间内利用注意力机制融合视觉特征和文本特征，有效的构建了视

觉表示和目标物体所在导航方向的关联，使系统能够通过探索找到在变化的环境中的目标物体。

- (3) 本文设计了一套单目相机和多线激光融合的图像点云融合方法，联合视觉观察的认知信息和多线点云的感知信息让移动机器人能够有效地在仿真环境和真实环境中依据自然语言指令完成目标导航任务，在移动机器人平台对所提出方法进行了测试加以验证。

本文专注于研究一种适用于室内服务机器人的目标物体导航方法，并致力于构建一个可以部署在真实移动机器人上的语言视觉激光多模态融合的导航系统，使移动机器人既能根据自然语言指令进行全局路径规划，完成导航至目标半米内的任务。多模态融合导航系统具有环境认知与感知、路径规划和自主导航探索这三个主要方面的功能，可以完成室内真实环境下的多目标物体导航任务，该方法在多目标物体连续导航的仿真测试环境中优于最先进的方法，实验结果证明了该方法的有效性和效率。对于未来的展望，本文有以下三个方向可以进行考虑：

- (1) 在真实环境的导航实验之中，我们采用激光雷达结合低功耗计算平台的导航架构在常规场景下表现稳定，但在复杂狭窄环境等受限空间内导航系统对误差源的容忍性下降，容易出现轨迹规划失效、运动控制失稳等问题。可以根据地图中障碍物与全局路径的几何关系标记狭窄环境并生成合适通行位姿对，在机器人出入被标记的狭窄环境时自动切换相应导航策略。通过这种全局成本地图膨胀化以规划更安全的全局路径的方法，机器人根据合适通行位姿分段规划全局路径以自适应环境
- (2) 在真实环境的导航实验之中，由于真实环境和仿真环境的差异问题导致了局部路径规划中的特征提取、融合模型性能的下降，当目标未能出现在移动机器人的视觉观察之中时，局部探索方法在存在遮挡物体的环境中存在失败的可能从而导致导航的失败。可以通过在局部路径规划的特征提取模块加入视觉观察的深度图，帮助模型更精准地定位物体的位置信息、理解环境的拓扑结构和空间关系，进一步提高模型对于空间中的物体位置的感知，从而更高效、可靠地执行局部探索任务。
- (3) 目标物体导航的环境中存在众多不同尺寸的物体，而多尺度特征可以增强模型的鲁棒性使其更具有抗干扰性，因此可以考虑添加多尺度特征可以更好地区分不同尺度下的物体。不同尺度的特征可以在不同的情况下提供更好的表征，让

模型能够更好地应对光照变化、遮挡、旋转等因素的影响。另一方面，添加多尺度特征会增加计算复杂度，可以通过金字塔结构或多尺度卷积等方法共享计算从而提高效率。这种方法能够提供更丰富的语义信息、适应不同尺度的物体、增强模型的鲁棒性，并且能够更好地结合局部和全局信息，从而提高图像处理任务的性能和效率。

## 参考文献

- [1] 中华人民共和国. 政府工作报告[EB/OL]. 2025. [https://www.gov.cn/yaowen/liebiao/202503/content\\_7010586.htm](https://www.gov.cn/yaowen/liebiao/202503/content_7010586.htm).
- [2] Zhang T, Hu X, Xiao J, et al. A survey of visual navigation: From geometry to embodied AI[J]. Engineering Applications of Artificial Intelligence, 2022, 114: 105036.
- [3] Majumdar A, Aggarwal G, Devnani B, et al. Zson: Zero-shot object-goal navigation using multimodal goal embeddings[J]. Advances in Neural Information Processing Systems, 2022, 35: 32340-32352.
- [4] Sun J, Wu J, Ji Z, et al. A survey of object goal navigation[J]. IEEE Transactions on Automation Science and Engineering, 2024.
- [5] Mavrogiannis C, Baldini F, Wang A, et al. Core challenges of social robot navigation: A survey[J]. ACM Transactions on Human-Robot Interaction, 2023, 12(3): 1-39.
- [6] Li S E. Reinforcement learning for sequential decision and optimal control[J], 2023.
- [7] Reddy K, Gharde P, Tayade H, et al. Advancements in robotic surgery: a comprehensive overview of current utilizations and upcoming frontiers[J]. Cureus, 2023, 15(12).
- [8] Zhang J, Singh S, et al. LOAM: Lidar odometry and mapping in real-time.[C]//Robotics: Science and systems: vol. 2: 9. 2014: 1-9.
- [9] Kazerouni I A, Fitzgerald L, Dooly G, et al. A survey of state-of-the-art on visual SLAM[J]. Expert Systems with Applications, 2022, 205: 117734.
- [10] Stanford Artificial Intelligence Laboratory et al. Robotic Operating System[EB/OL]. (2018-05-23). <https://www.ros.org/>.
- [11] Dellaert F, Fox D, Burgard W, et al. Monte carlo localization for mobile robots[C]// Proceedings 1999 IEEE international conference on robotics and automation (Cat. No. 99CH36288C): vol. 2. 1999: 1322-1328.
- [12] Thrun S, Fox D, Burgard W, et al. Robust Monte Carlo localization for mobile robots[J]. Artificial intelligence, 2001, 128(1-2): 99-141.
- [13] Hart P E, Nilsson N J, Raphael B. A formal basis for the heuristic determination of minimum cost paths[J]. IEEE transactions on Systems Science and Cybernetics, 1968, 4(2): 100-107.

- [14] Rösmann C, Feiten W, Wösch T, et al. Efficient trajectory optimization using a sparse model[C]//2013 European Conference on Mobile Robots. 2013: 138-143.
- [15] Murphy K, Russell S. Rao-Blackwellised particle filtering for dynamic Bayesian networks[G]//Sequential Monte Carlo methods in practice. Springer, 2001: 499-515.
- [16] Grisetti G, Stachniss C, Burgard W. Improving grid-based slam with rao-blackwellized particle filters by adaptive proposals and selective resampling[C]//Proceedings of the 2005 IEEE international conference on robotics and automation. 2005: 2432-2437.
- [17] Grisetti G, Stachniss C, Burgard W. Improved techniques for grid mapping with rao-blackwellized particle filters[J]. IEEE transactions on Robotics, 2007, 23(1): 34-46.
- [18] Hess W, Kohler D, Rapp H, et al. Real-time loop closure in 2D LIDAR SLAM[C]//2016 IEEE international conference on robotics and automation (ICRA). 2016: 1271-1278.
- [19] Engel J, Schöps T, Cremers D. LSD-SLAM: Large-scale direct monocular SLAM[C]//European conference on computer vision. 2014: 834-849.
- [20] Zhang J, Singh S. Low-drift and real-time lidar odometry and mapping[J]. Autonomous robots, 2017, 41: 401-416.
- [21] Shan T, Englot B. Lego-loam: Lightweight and ground-optimized lidar odometry and mapping on variable terrain[C]//2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). 2018: 4758-4765.
- [22] Li Q, Chen S, Wang C, et al. Lo-net: Deep real-time lidar odometry[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019: 8473-8482.
- [23] 孙海, 任翠平, 卢军, 等. 激光测距在仓储搬运机器人运动中的应用[J]. 电子技术与软件工程, 2017(1): 103-104.
- [24] Davison A J, Reid I D, Molton N D, et al. MonoSLAM: Real-time single camera SLAM[J]. IEEE transactions on pattern analysis and machine intelligence, 2007, 29(6): 1052-1067.
- [25] Davison A J. Real-Time Localisation and Mapping with a Single Camera[J]. 情報処理学会研究報告 = IPSJ SIG technical reports, 2003, 2003(2): 107-114.

- [26] Klein G, Murray D. Parallel tracking and mapping for small AR workspaces[C]//2007 6th IEEE and ACM international symposium on mixed and augmented reality. 2007: 225-234.
- [27] Mur-Artal R, Montiel J M M, Tardos J D. ORB-SLAM: a versatile and accurate monocular SLAM system[J]. IEEE transactions on robotics, 2015, 31(5): 1147-1163.
- [28] Newcombe R A, Lovegrove S J, Davison A J. DTAM: Dense tracking and mapping in real-time[C]//2011 international conference on computer vision. 2011: 2320-2327.
- [29] Forster C, Pizzoli M, Scaramuzza D. SVO: Fast semi-direct monocular visual odometry[C]//2014 IEEE international conference on robotics and automation (ICRA). 2014: 15-22.
- [30] Wen C, Huang Y, Huang H, et al. Zero-shot object navigation with vision-language models reasoning[C]//International Conference on Pattern Recognition. 2025: 389-404.
- [31] Unlu H U, Yuan S, Wen C, et al. Reliable semantic understanding for real world zero-shot object goal navigation[C]//International Conference on Pattern Recognition. 2025: 135-150.
- [32] Gutiérrez-Álvarez C, Ríos-Navarro P, Flor-Rodríguez-Rabadán R, et al. Visual semantic navigation with real robots[J]. Applied Intelligence, 2025, 55(2): 206.
- [33] Yuan S, Unlu H U, Huang H, et al. Exploring the reliability of foundation model-based frontier selection in zero-shot object goal navigation[C]//International Conference on Pattern Recognition. 2025: 119-134.
- [34] Jones J, Mees O, Sferrazza C, et al. Beyond sight: Finetuning generalist robot policies with heterogeneous sensors via language grounding[J]. ArXiv preprint arXiv:2501.04693, 2025.
- [35] Long Y, Li X, Cai W, et al. Discuss before moving: Visual language navigation via multi-expert discussions[C]//2024 IEEE International Conference on Robotics and Automation (ICRA). 2024: 17380-17387.
- [36] Zhou G, Hong Y, Wu Q. Navgpt: Explicit reasoning in vision-and-language navigation with large language models[C]//Proceedings of the AAAI Conference on Artificial Intelligence: vol. 38: 7. 2024: 7641-7649.

- [37] Zhou G, Hong Y, Wang Z, et al. Navgpt-2: Unleashing navigational reasoning capability for large vision-language models[C]//European Conference on Computer Vision. 2024: 260-278.
- [38] Liu S, Hasan A, Hong K, et al. DRAGON: A dialogue-based robot for assistive navigation with visual language grounding[J]. IEEE Robotics and Automation Letters, 2024.
- [39] Yokoyama N, Ha S, Batra D, et al. Vlfm: Vision-language frontier maps for zero-shot semantic navigation[C]//2024 IEEE International Conference on Robotics and Automation (ICRA). 2024: 42-48.
- [40] An D, Wang H, Wang W, et al. Etpnav: Evolving topological planning for vision-language navigation in continuous environments[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2024.
- [41] Guo Y, Xie Y, Chen Y, et al. An Efficient Object Navigation Strategy for Mobile Robots Based on Semantic Information[J/OL]. Electronics, 2022, 11(7). <https://www.mdpi.com/2079-9292/11/7/1136>. DOI: 10.3390/electronics11071136.
- [42] Shah D, Eysenbach B, Kahn G, et al. Ving: Learning open-world navigation with visual goals[C]//2021 IEEE International Conference on Robotics and Automation (ICRA). 2021: 13215-13222.
- [43] Shah D, Osiński B, Levine S, et al. Lm-nav: Robotic navigation with large pre-trained models of language, vision, and action[C]//Conference on robot learning. 2023: 492-504.
- [44] Gupta A, Savarese S, Ganguli S, et al. Embodied intelligence via learning and evolution[J]. Nature communications, 2021, 12(1): 5721.
- [45] Cao H, Feng F, Huo J, et al. Causal action empowerment for efficient reinforcement learning in embodied agents[J]. Science China Information Sciences, 2025, 68(5): 5-19.
- [46] Fan H, Liu X, Fuh J Y H, et al. Embodied intelligence in manufacturing: leveraging large language models for autonomous industrial robotics[J]. Journal of Intelligent Manufacturing, 2025, 36(2): 1141-1157.
- [47] Wen Y, Lin J, Zhu Y, et al. Vidman: Exploiting implicit dynamics from video diffusion model for effective robot manipulation[J]. Advances in Neural Information Processing Systems, 2024, 37: 41051-41075.

- [48] Lin X, Lin T, Huang L, et al. BIP3D: Bridging 2D Images and 3D Perception for Embodied Intelligence[J]. ArXiv preprint arXiv:2411.14869, 2024.
- [49] Zhang Z. A flexible new technique for camera calibration[J]. IEEE Transactions on pattern analysis and machine intelligence, 2002, 22(11): 1330-1334.
- [50] Moré J J. The Levenberg-Marquardt algorithm: implementation and theory in numerical analysis[J]. Lecture notes in mathematics, 1977, 630: 105-116.
- [51] Viola P, Jones M. Rapid object detection using a boosted cascade of simple features[C]// Proceedings of the 2001 IEEE computer society conference on computer vision and pattern recognition. CVPR 2001: vol. 1. 2001: I-I.
- [52] Taye M M. Theoretical understanding of convolutional neural network: Concepts, architectures, applications, future directions[J]. Computation, 2023, 11(3): 52.
- [53] 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.
- [54] Liu W, Anguelov D, Erhan D, et al. Ssd: Single shot multibox detector[C]//Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14. 2016: 21-37.
- [55] 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.
- [56] Girshick R. Fast r-cnn[C]//Proceedings of the IEEE international conference on computer vision. 2015: 1440-1448.
- [57] Ren S, He K, Girshick R, et al. Faster R-CNN: Towards real-time object detection with region proposal networks[J]. IEEE transactions on pattern analysis and machine intelligence, 2016, 39(6): 1137-1149.
- [58] Doukas C, Maglogiannis I. Region of interest coding techniques for medical image compression[J]. IEEE Engineering in medicine and Biology Magazine, 2007, 26(5): 29-35.
- [59] Carion N, Massa F, Synnaeve G, et al. End-to-end object detection with transformers[C]//European conference on computer vision. 2020: 213-229.

- [60] Wang A, Chen H, Liu L, et al. Yolov10: Real-time end-to-end object detection[J]. Advances in Neural Information Processing Systems, 2025, 37: 107984-108011.
- [61] Varghese R, Sambath M. Yolov8: A novel object detection algorithm with enhanced performance and robustness[C]//2024 International Conference on Advances in Data Engineering and Intelligent Computing Systems (ADICS). 2024: 1-6.
- [62] Niu Z, Zhong G, Yu H. A review on the attention mechanism of deep learning[J]. Neurocomputing, 2021, 452: 48-62.
- [63] Liu H, Song R, Zhang X, et al. Point cloud segmentation based on Euclidean clustering and multi-plane extraction in rugged field[J]. Measurement Science and Technology, 2021, 32(9): 095106.
- [64] Guo Z, Liu H, Shi H, et al. KD-tree-based euclidean clustering for tomographic SAR point cloud extraction and segmentation[J]. IEEE Geoscience and Remote Sensing Letters, 2023, 20: 1-5.
- [65] Ke L, Li X, Bisk Y, et al. Tactical rewind: Self-correction via backtracking in vision-and-language navigation[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019: 6741-6749.
- [66] Hong Y, Rodriguez C, Qi Y, et al. Language and visual entity relationship graph for agent navigation[J]. Advances in Neural Information Processing Systems, 2020, 33: 7685-7696.
- [67] Wang H, Wang W, Liang W, et al. Structured scene memory for vision-language navigation[C]//Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition. 2021: 8455-8464.
- [68] Huang C, Mees O, Zeng A, et al. Visual language maps for robot navigation[C]//2023 IEEE International Conference on Robotics and Automation (ICRA). 2023: 10608-10615.
- [69] 王湉, 范峻铭, 郑湃. 基于大语言模型的人机交互移动检测机器人导航方法[J]. 计算机集成制造系统, 2024, 30(5): 1587-1594.
- [70] Anderson P, Wu Q, Teney D, et al. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 3674-3683.

- [71] Guo D, Yang D, Zhang H, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning[J]. ArXiv preprint arXiv:2501.12948, 2025.
- [72] Radford A, Kim J W, Hallacy C, et al. Learning transferable visual models from natural language supervision[C]//International conference on machine learning. 2021: 8748-8763.
- [73] Li W, Saeedi S, McCormac J, et al. Interiornet: Mega-scale multi-sensor photo-realistic indoor scenes dataset[J]. ArXiv preprint arXiv:1809.00716, 2018.
- [74] Krantz J, Wijmans E, Majumdar A, et al. Beyond the nav-graph: Vision-and-language navigation in continuous environments[C]//Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVIII 16. 2020: 104-120.
- [75] Du H, Yu X, Zheng L. Learning object relation graph and tentative policy for visual navigation[C]//Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VII 16. 2020: 19-34.
- [76] Du H, Yu X, Zheng L. Vtnet: Visual transformer network for object goal navigation[J]. ArXiv preprint arXiv:2105.09447, 2021.
- [77] Wang T, Wu Z, Wang D. Visual perception generalization for vision-and-language navigation via meta-learning[J]. IEEE Transactions on Neural Networks and Learning Systems, 2021, 34(8): 5193-5199.
- [78] Fang Q, Xu X, Wang X, et al. Target-driven visual navigation in indoor scenes using reinforcement learning and imitation learning[J]. CAAI Transactions on Intelligence Technology, 2022, 7(2): 167-176.
- [79] Fukushima R, Ota K, Kanezaki A, et al. Object memory transformer for object goal navigation[C]//2022 International conference on robotics and automation (ICRA). 2022: 11288-11294.
- [80] 朱威, 洪力栋, 施海东, 等. 结合优势结构和最小目标 Q 值的深度强化学习导航算法.[J]. 控制理论与应用, 2024, 41(4): 716-728.

## 攻读硕士学位期间取得的研究成果

一、已发表（包括已接受待发表）的论文，以及已投稿、或已成文打算投稿、或拟成文投稿的论文情况（只填写与学位论文内容相关的一部分）：

序号	作者（全体作者，按顺序排列）	题目	发表或投稿刊物名称、级别	发表的卷期、年月、页码	与学位论文哪一部分（章、节）相关	被索引收录情况
1	毕盛, 杨礼铭, 董敏, 沈煜	语言视觉激光多模态融合的机器人导航方法	小型微型计算机系统, 中文核心	已录用定稿	第三章部分内容	CNKI
2	董敏, 谭皓禹, 杨礼铭, 沈煜, 陈章韶, 毕盛	面向独居老人的智能家居监护系统	嵌入式技术与智能系统	2024 年 1(1)	第四章部分内容	Hans

二、与学位内容相关的其它成果（包括专利、著作、获奖项目等）（只填写与学位论文内容相关的一部分）：

序号	作者（全体作者，按顺序排列）	题目	成果类型	状态	受理/登记时间
1	谭皓禹; 杨礼铭; 沈煜	面向独居老人的智能家居监护系统	中国研究生电子设计竞赛全国总决赛一等奖	已获取	2023-08-16
2	毕盛; 杨礼铭; 董敏	一种视觉与多线激光融合的移动机器人导航系统	发明专利, 专利号: CN202410023917.7	已公开	2024-04-12

## 致 谢

首先我要感谢父母二十余年的养育之恩，你们一直是我前行路上最坚实的后盾，尊重我的选择，用无私的爱与辛勤的付出让我心无旁骛地行走在自己所选择的道路之上。惟愿未来我能成为你们的骄傲，亦如你们一直是我的港湾。

衷心感谢毕盛老师给予的指导和支持。从论文选题的反复推敲到框架搭建的精准指导，从实验设计的严谨把关到逐字逐句的审阅批注，您始终以渊博的学识、创新的思维和精益求精的治学态度为我指明方向。您对科研的热忱与对学生的包容，将永远激励我在未来的征途上笃行不怠。

感谢 B3351 实验室的师兄师姐为我的开发工作提供了硬件平台以及其它技术支持，让我能够按照个人兴趣，充分展示我的能力。

感谢实验室的谭皓禹、沈煜等同门师兄弟给予的帮助，为我的工作提供了硬件平台和其他技术帮助，也难忘无数个深夜并肩实验的时光，你们不厌其烦地与我探讨技术细节，慷慨分享珍贵数据；难忘比赛之前一起排查、解决问题；更难忘失落时的暖心鼓励与成功时的击掌相庆。这段同舟共济的岁月，因你们的陪伴而熠熠生辉。

同时本文的工作离不开前人发布的大量论文和论文开源代码，在此，我对所无私的科研人员和所有活跃在开源社区第一线的开发者们致以崇高的敬意。

三年时光如白驹过隙，愿此去经年，我们都能在各自的征途上发光发热，在多年以后再次相见时也能举杯畅谈。