

分类号: TP242.6

密 级: 公开

学校代码: 10363

学 号: 2220340103



安徽工程大学

Anhui Polytechnic University

硕士学位论文

题 目

动态场景下基于 OneFormer 网络的

SLAM 算法

论 文 作 者

杨苏朋

指 导 教 师

陈孟元（教授）

学 科（专业）

控制科学与工程

研 究 方 向

检测技术与自动化装置

论文提交日期: 2025 年 5 月 15 日

分类号：TP242.6
密 级：公 开

单位代码：10363
学 号：2220340103

题 目 动态场景下基于 OneFormer 网络的 SLAM 算法

题 目 Research on Visual SLAM Algorithm Based on
Target Detection and Segmentation in Dynamic
Scenes

学生姓名：杨苏朋

导 师：陈孟元（教授）

专业学位类别：控制科学与工程

研究方向（领域）：检测技术与自动化装置

论文答辩日期：2025 年 5 月 12 日

安徽工程大学学位论文原创性声明

本人郑重声明：我恪守学术道德，崇尚严谨学风。所呈交的学位论文，是本人在导师的指导下，独立进行研究工作所取得的成果。除文中已明确注明和引用的内容外，本论文不包含任何其他个人或集体已经发表或撰写过的作品及成果的内容。论文为本人亲自撰写，我对所写的内容负责，并完全意识到本声明的法律结果由本人承担。

学位论文作者签名：杨苏朋

日期：2025 年 5 月 15 日

安徽工程大学学位论文版权使用授权书

学位论文作者完全了解学校有关保留、使用学位论文的规定，同意学校保留并向国家有关部门或机构送交论文的复印件和电子版，允许论文被查阅或借阅。本人授权安徽工程大学可以将本学位论文的全部或部分内容编入有关数据库进行检索，可以采用影印、缩印或扫描等复制手段保存和汇编本学位论文。

保密 ☐，在 _____ 年解密后适用本版权书。

本学位论文属于

不保密 ☒

学位论文作者签名：杨苏朋

指导教师签名：陈吉云

日期：2025 年 5 月 15 日

日期：2025 年 5 月 15 日

动态场景下基于 OneFormer 网络的 SLAM 算法

摘 要

随着智能机器人在工业、服务等领域的广泛应用，基于视觉的同步定位与建图技术（Simultaneous Localization and Mapping, SLAM）已成为机器人自主感知的核心模块，受到学者们的持续关注。传统视觉 SLAM 算法通常基于静态环境假设，在面对动态物体干扰时，系统容易产生错误的特征匹配与位姿估计，进而导致定位漂移与地图失真。目前主流的方法为直接剔除动态物体，但针对动态特征占比较大与某些动态物体包含有利信息而不能剔除的场景并不适用。本文为适用所有动态场景，提出了一种基于 OneFormer 通用分割网络的动态 SLAM 跟踪算法，进而提升系统的定位与建图精度。

针对传统视觉 SLAM 在动态场景下动态目标分割精度不足的问题，设计了一种基于改进 OneFormer 网络的动态物体分割算法。该算法以 OneFormer 通用图像分割网络为基础框架，在低层级信号特征增强模块，利用轻量级卷积与特征信号重组强化浅层特征表达，提升了对弱语义特征的判别能力，使网络更聚焦于动态特征提取。在双域协同注意力增强模块，利用通道-空间双域注意力机制自适应增强高维特征动态区域的网络分割权重，提高了网络对动态区域的识别精度。

针对在动态物体占比较大与动态物体包含有利信息的不能剔除场景，传统 SLAM 定位与建图不准确问题，在改进 OneFormer 分割网络的基础上，设计了一种基于动态物体分割与映射的 SLAM 实时跟踪算法。该算法结合网络分割获取到的相关信息，使用相机位姿与物体运动偏转角度估计，实现潜在动态物体判断；并将识别的动态物体映射回网络分割掩膜图上形成分层结构，通过对动态物体的精准显

示与追踪，保留了动态物体上的有效信息，降低了因信息缺失而导致的系统误差，提高了系统位姿估计准确性。

为验证本文所提算法有效性，在公开 TUM 数据集中对算法进行测试，实验数据显示，改进后分割网络像素准确率与平均交并比分别达到 88.6%与 60.4%，相较于改进前实现了 4.4%与 5.1%的绝对提升。所提跟踪算法在绝对轨迹误差均方根值指标上较 DP-SLAM、DynaSLAM 分别降低 17.70%与 19.63%，展现出良好的位姿估计精度与建图性能。此外，在搭建的移动机器人硬件平台上进行室内外真实场景测试，结果表明本文算法具有较强的适应性，为动态场景下的视觉 SLAM 算法提供了重要参考依据。

关键词：同时定位与地图构建，动态场景，深度学习，运动判断，图像映射

SLAM Algorithm Based on OneFormer Network in Dynamic Scenes

ABSTRACT

With the wide application of intelligent robots in industrial and service fields, vision-based Simultaneous Localisation and Mapping (SLAM) has become a core module of robot autonomous perception and has received continuous attention from scholars. Traditional vision SLAM algorithms are usually based on static environment assumptions, and when facing dynamic object interference, the system is prone to erroneous feature matching and position estimation, which leads to localisation drift and map distortion. Currently, the mainstream approach is to directly reject dynamic objects, but it is not suitable for scenes with a large proportion of dynamic features and certain dynamic objects that contain favourable information and cannot be rejected. In this paper, we propose a dynamic SLAM tracking algorithm based on OneFormer universal segmentation network for all dynamic scenarios, which improves the system's positioning and mapping accuracy.

Aiming at the problem of insufficient dynamic target segmentation accuracy of traditional visual SLAM in dynamic scenes, a dynamic object segmentation algorithm based on improved OneFormer network is designed. The algorithm takes OneFormer general image segmentation network as the basic framework. In the low-level signal feature enhancement module, lightweight convolution and feature signal recombination are used to strengthen the shallow feature expression, which improves the discriminative ability of weak semantic features and makes the network more focused on dynamic feature extraction. In the dual-

domain collaborative attention enhancement module, the channel-space dual-domain attention mechanism is used to adaptively enhance the network segmentation weights of high-dimensional feature dynamic regions, which improves the network's recognition accuracy of dynamic regions.

Aiming at the problem of inaccuracy of traditional SLAM positioning and mapping in non-rejectable scenarios where dynamic objects account for a large proportion and dynamic objects contain favourable information, a real-time tracking algorithm for SLAM based on dynamic object segmentation and mapping is designed based on the improvement of the OneFormer segmentation network. The algorithm combines the relevant information obtained by network segmentation, uses camera position and object motion deflection angle estimation to achieve potential dynamic object judgement; and maps the identified dynamic objects back to the network segmentation mask map to form a layered structure, which retains the effective information on the dynamic objects through the accurate display and tracking of the dynamic objects, reduces the systematic error due to the lack of information, and improves the system position estimation Accuracy.

In order to verify the effectiveness of the algorithm proposed in this paper, the algorithm is tested in the public TUM dataset, and the experimental data show that the pixel accuracy and the average intersection and merger ratio of the improved segmentation network reach 88.6% and 60.4%, respectively, which achieve an absolute improvement of 4.4% and 5.1% compared with the pre-improvement. The proposed tracking algorithm reduces 17.70% and 19.63% in absolute trajectory error rms value metrics compared to DP-SLAM and DynaSLAM, respectively, demonstrating good bit-position estimation accuracy and map building performance. In addition, the results of indoor and outdoor real scene tests

on the constructed mobile robot hardware platform show that the algorithm in this paper has strong adaptability, which provides an important reference for the vision SLAM algorithm in dynamic scenes.

Yang Supeng (Electrical engineering)

Supervised by Professor Chen Mengyuan

KEY WORDS: Simultaneous Localization and Map Building, Dynamic Scenes, Deep Learning, Motion Determination, Image Mapping.

目 录

第 1 章 绪论.....	1
1.1 研究背景及意义	1
1.2 国内外研究现状	3
1.2.1 传统视觉 SLAM 研究现状.....	4
1.2.2 动态场景下视觉 SLAM 研究现状.....	5
1.3 本文的组织架构	7
第 2 章 视觉 SLAM 与 OneFormer 网络基础.....	10
2.1 视觉 SLAM 系统框架.....	10
2.1.1 视觉里程计	11
2.1.2 后端优化.....	13
2.1.3 回环检测与建图	14
2.2 动态 SLAM 问题.....	16
2.3 OneFormer 通用分割网络.....	18
2.4 本章小结	21
第 3 章 基于改进 OneFormer 网络的动态物体分割算法	22
3.1 低层级特征信号增强模块 LSSEM	22
3.1.1 轻量级卷积模块 GSConv	22
3.1.2 特征信号重组模块.....	23
3.2 双域协同注意力增强模块 DEAM	25
3.2.1 空间注意力机制 SAM	26
3.2.2 通道注意力机制 CAM.....	26
3.2.3 双域注意力协同工作机制.....	27
3.3 实验结果与分析	27
3.3.1 改进模块提升效果验证实验	28
3.3.2 动态场景下分割性能实验	30
3.3.3 不同算法对比实验.....	31
3.3.4 真实场景下分割效果实验.....	33
3.4 本章小结	34
第 4 章 基于动态物体分割与映射的 SLAM 实时跟踪算法	35

4.1 潜在动态物体判断	35
4.2 动态物体映射与跟踪	37
4.2.1 动态物体映射	37
4.2.2 逐帧映射形成跟踪	37
4.3 实验结果与分析	38
4.3.1 场景中潜在动态物体识别与判断实验	38
4.3.2 动态物体判别及反向映射实验	39
4.3.3 SLAM 系统中算法运行评估实验	40
4.3.4 真实场景中算法运行评估实验	42
4.4 本章小结	46
第 5 章 全文总结展望	47
5.1 全文总结	47
5.2 展望	48
参考文献.....	49
攻读学位期间发表的学术论文与科研成果.....	54
致 谢.....	55

第 1 章 绪论

1.1 研究背景及意义

随着计算机视觉、智能算法与神经网络等前沿技术的发展，新一代智能技术正深刻改变着人类社会的运行方式。机器人技术虽兴起时间较短，却已展现出重要的工业价值，一方面有效替代了人工完成高危复杂作业，另一方面大幅提升了全球工业生产的智能化水平。当前，机器人技术的创新程度已成为衡量国家科技实力与工业现代化水平的重要标志，受到各国政府的高度关注与重点扶持。我国在《“十四五”机器人产业发展规划》中特别强调，要通过提升机器人产业核心竞争力，为推进建设制造强国、保障民生需求提供关键技术保障。移动机器人作为智能机器人的重要分支，凭借其自主移动与环境交互能力，在无人机、智能家居、物流运输等多个领域中发挥重要作用^[1]，具体如图 1-1 所示。



图 1-1 机器人技术应用实例

移动机器人要实现智能化运行，必须解决三个核心问题：环境定位、目标导航与路径决策^[2]。其中，环境定位问题需要通过精确的位置感知来解决，而后两

者则属于运动规划范畴。在现有技术中，SLAM 技术是解决机器人定位问题的有效方案，通过传感器对数据进行采集与分析，在未知环境中实现实时位姿估计与环境建图，为后续的导航与人机交互奠定了基础^[3]。根据传感器配置的不同，SLAM 系统主要分为激光与视觉两大类。激光 SLAM 系统^[4]通常使用激光雷达、编码器与惯性测量单元 IMU 等传感器，具有测量精度高、稳定性好等优势。然而，该系统在复杂场景中性能显著下降，且高昂的硬件成本限制了其应用范围。相比之下，视觉 SLAM 系统基于单目、双目或 RGB-D 相机获取环境图像信息，不仅能捕捉丰富的纹理特征，还具有成本低、体积小等优势。特别是在面对动态场景时，视觉 SLAM 系统展现出更强的环境适应性。随着移动机器人应用场景的不断拓展，对 SLAM 系统的鲁棒性要求日益增高，视觉 SLAM 技术凭借其主动感知能力，正逐渐成为移动机器人自主导航领域的研究热点^[5]。该技术不仅能满足开放环境下的定位需求，还为机器人实现智能决策提供了更丰富的环境信息。

当前主流的视觉 SLAM 系统普遍建立在静态环境假设的基础上，这类系统在缺乏动态干扰的静态场景中能够保持良好的定位与建图性能。然而，当应用于实际动态场景如人流密集的公共场所或交通繁忙的城市道路时，移动物体对视觉观测造成的干扰会严重影响系统的位姿估计。这种干扰不仅会导致建图出现几何畸变与特征拖影等问题，更会破坏地图的空间一致性，使得最终构建的环境地图失去实用价值，严重制约了视觉 SLAM 技术在实际应用中的可靠性。

近年来，深度学习技术的引入为视觉 SLAM^{[6][7]}系统带来了新的研究方向，特别是通过语义分割网络识别动态物体并结合运动一致性分析的方法，有效降低了动态干扰对系统性能的影响。然而，在真实世界的复杂动态场景如城市交叉路口等环境中，现有系统仍面临多重挑战，这些场景中普遍存在的运动物体严重影响了图像特征的提取，同时，动态物体导致的错误匹配还会引发闭环检测失效与误差累积问题。更复杂的是物体间的相互遮挡会导致外观信息缺失，增加语义分割的难度，干扰系统对物体运动状态的准确判断。这些因素共同作用，使得系统在复杂动态环境下的图像分割精度、位姿估计准确性与地图构建质量出现明显下降。为了提升视觉 SLAM 系统在实际应用中的可靠性，研究出一种可以在复杂动态环境下进行良好特征提取、运动估计与动态物体处理的 SLAM 技术具有重

要价值。

本文以 ORB-SLAM2 框架为基础,从动态场景下视觉 SLAM 研究现状出发,以实现复杂动态场景下移动机器人准确性定位与高精度建图为目标,深入分析复杂动态场景移动机器人面临的挑战,通过动态物体分割网络优化以及动态物体映射与跟踪线程共同构建了一种面向复杂动态场景的视觉 SLAM 算法。该算法有效改善了移动机器人在动态场景下因动态物体干扰与信息丢失导致的跟踪失效、轨迹漂移与建图失真等问题,提高了视觉 SLAM 系统的定位与建图能力。

1.2 国内外研究现状

自 SLAM 理论框架于 1986 年问世以来^[8],其技术体系历经近四十年的演进,已经实现了从实验室验证到工业级系统集成的模式转换,应用场景从早期的实验室环境延伸至自动驾驶、工业巡检、航天探测等复杂场景,逐步成为智能系统实现自主环境感知的核心技术。在技术演进历程中,SLAM 经历了两次重大模式转换,第一次是从基于滤波的递推估计向基于优化的批量处理方法转型,第二次则是从纯几何驱动向多模态融合的智能感知升级。

早期 SLAM 系统主要采用卡尔曼滤波、粒子滤波等递推估计算法^{[9][10]},这类方法通过建立环境信息与传感器观测的概率关联模型,在计算资源受限的条件下展现出良好的实时性。然而,其理论根基仅建立在系统当前状态与上一时刻状态相关的假设之上^[11],这种时序关联的局部性约束导致算法无法建立跨时段的历史帧数据关联,从而导致闭环检测机制缺失与位姿估计误差随时间推移产生不可逆的累积漂移,严重制约系统在长期运行中的可靠性。这些问题暴露出传统滤波方法在全局优化方面中的不足,促使研究者寻求新的方法突破。

随着计算机视觉技术的发展,基于视觉传感器的 SLAM 系统逐渐成为主流研究方向。相较于传统滤波,视觉 SLAM 系统采用分层处理^{[12][13]}的思想,将复杂的定位与建图问题分解为多个功能模块协同完成。传感器数据模块通过预处理实现图像去畸变与特征增强,为后续处理提供高质量输入。视觉里程计模块基于多视图几何原理,建立帧间对应关系并计算相机位姿变化,其匹配精度直接影响局部轨迹的准确性。闭环检测模块通过建立数据帧之间的关联,为系统提供关键的全局约束条件。后端优化模块则通过概率图模型整合各类观测信息,在贝叶斯

框架下求解最优位姿估计，有效抑制误差累积，使得系统在保证实时性的同时，能够实现厘米级的高精度定位。

尽管现有框架已取得显著进展，视觉 SLAM 在实际动态环境中仍面临重大技术挑战，特别是在包含行人、车辆等移动物体的开放场景中，动态干扰导致的特征误匹配问题会引发系统性的位姿估计误差。这种误差不仅会造成局部轨迹失真与地图结构畸变，还会破坏闭环检测的有效性，导致整个 SLAM 系统的性能显著下降。

1.2.1 传统视觉 SLAM 研究现状

视觉 SLAM 技术作为机器人环境感知的关键技术，主要围绕环境建模精度提升、实时计算性能优化与场景表征能力完善三个主要研究方向，显著改善了系统的特征匹配、位姿估计与闭环检测等功能，促进了在自动驾驶、工业自动化等领域的实际应用。然而，现有系统仍面临两个主要技术瓶颈，即处理动态复杂环境时的稳定性不足以及计算效率与精度不平衡的问题。要推动该技术的进一步发展，需要在多传感器融合、动态场景建模与算法优化等方面进行持续创新。

2007 年，Andrew J. Davison 团队^[14]提出的 MonoSLAM 系统，开创了基于扩展卡尔曼滤波的视觉 SLAM 技术，其核心贡献在于建立了基于概率图模型的状态估计框架，通过特征点提取构建观测方程，采用卡尔曼滤波对相机位姿与稀疏三维点云进行联合概率估计，首次实现更新频率为 30Hz 的实时单目定位，该系统的突破性在于摆脱了对惯性测量单元 IMU 等辅助传感器的依赖，但其仍受限于单线程架构的设计缺陷，计算复杂度随地图规模呈指数增长，在存在动态干扰的场景中的特征跟踪失败率较高。同年，George Klein 与 David Murray 等人^[15]提出的 PTAM（Parallel Tracking and Mapping）系统，标志着 SLAM 架构设计的模式转变，该系统首次实现跟踪与建图线程的解耦并行，前端线程采用改进的 Shi-Tomasi 角点检测实现亚像素级特征跟踪，后端线程则开创性地引入光束法平差（Bundle Adjustment, BA）进行位姿与地图的全局优化。这种双线程架构使 BA 优化效率提升 47%，同时将轨迹误差降低至 MonoSLAM 的 31%，但该系统在复杂场景下的特征跟踪鲁棒性不足且未实现闭环检测模块搭建，仍需要在后续研究中不断进行改进。

2014 年, Jakob Engel 团队^[16]提出的 LSD-SLAM 系统, 突破特征点法的理论框架而开创基于直接法的稠密建图新方法, 其创新性在于构建光度误差模型, 通过相机运动参数化建立相邻帧间的灰度不变性假设, 采用李代数 $SE(3)$ 表示相机位姿, 利用高斯-牛顿法求解误差函数的最小二乘解。相较于特征点法, 直接法避免了特征提取带来的信息损失, 将尺度漂移误差降低 23%, 但该方法对环境变化敏感, 且在 GPU 加速下仍难以实现实时稠密重建。Christian Forster 团队^[17]同期提出的 SVO (Semi-direct Visual Odometry) 系统使用分层优化架构, 前端采用 FAST 角点检测构建稀疏特征地图, 后端基于直接法对选定像素区域进行光流跟踪, 这种混合策略在保证精度的同时将计算负载降低至传统方法的 40%。Mathieu Labbé 团队^[18]研发的 RTAB-Map 系统在多传感器融合领域也取得突破, 通过构建基于词袋模型的多源特征关联框架, 实现激光雷达点云与 RGB-D 数据的对齐。其创新的记忆管理机制采用工作记忆、短期记忆与长期记忆三级结构将回环检测召回率提升至 95.3%, 显著优于单一传感器系统。

2015-2017 年, Raul Mur-Artal 团队^{[19][20]}推出的 ORB-SLAM 系列确立了特征点法的工程化标准, 创新性的构建跟踪-建图-回环检测多线程并行模型, 通过负载均衡策略使 CPU 利用率提升 65%, 采用改进的 ORB 特征实现 512 位二进制描述子, 使特征匹配速度达到 0.8ms/帧, 设计基于协方差矩阵的关键帧选择机制, 将 BA 优化变量规模压缩至传统方法的 58%。在总体性能测试中, 其绝对轨迹误差的均方根误差 RMSE 达到 1.2cm, 至今仍是特征点法的性能标准, 该系统的最终开源也成为了工业级实地部署的基础框架。

2017 年, Tong Qin 团队^[21]提出的 VINS-Mono 系统实现视觉-惯性紧耦合的理论突破。其创新在于构建 IMU 模型, 将相邻关键帧间的惯性测量值积分转换为相对运动约束, 避免重复积分导致的计算冗余, 通过设计基于滑动窗口的非线性优化器联合优化视觉重投影误差与 IMU 测量残差, 将高速运动场景下的定位误差的均方根误差 RMSE 降低至 0.25m, 较纯视觉方法在系统稳定性上提升 76%, 标志着多模态传感器融合从松耦合向紧耦合的根本转变。

1.2.2 动态场景下视觉 SLAM 研究现状

视觉 SLAM 长期建立在静态环境假设的理论体系之上, 当面对存在行人、

车辆等动态干扰因素的现实应用场景时，其稳定性便会受到严重影响。动态物体引入的时变特征会破坏其数据关联的一致性，导致建图过程产生偏差并引发定位漂移问题，使得系统在实际部署中的鲁棒性与精度下降^[22]。因此，如何实现动态场景下视觉 SLAM 系统定位与建图的双重稳定性成为制约 SLAM 技术发展的关键难题^[23]。

为突破静态环境假设的局限性，学者们围绕动态场景的 SLAM 方法展开深入研究，逐步形成三大技术路线：基于纯几何约束的动态抑制方法、基于深度学习的语义方法以及融合几何与深度学习的混合架构。各类方法在动态特征处理机制上呈现出显著的模式差异^[24]。

在几何方法研究维度，研究者致力于通过物理运动约束筛选静态特征。2017 年，Li S 等人^[25]另辟蹊径，提出深度边缘点关联策略，利用深度不连续区域的几何稳定性构建静态特征子集并通过加权优化机制削弱动态点对位姿估计的影响，但其在动态物体较多的场景中的适应性仍有待提升。2021 年，艾青林团队^[26]创新性地提出特征点动态状态分级理论，通过构建包括极线约束、重投影误差阈值等在内的多层级几何验证模型，实现对动态特征的渐进式剔除，但该方法的局限在于对场景几何结构的完整性依赖较高。

随着深度学习技术的突破，基于语义先验的动态 SLAM 方法展现出强大潜力，2021 年，Liu Y 团队^[27]在 ORB-SLAM3 架构中嵌入并行语义推理线程，通过实时语义分割网络识别动态目标，并在后端优化中施加语义约束，有效降低了动态干扰的影响。然而，此类方法在复杂场景下面临语义误检导致的特征错误去除问题。针对此缺陷，同年，Soares 等人^[28]提出的 Crowd SLAM 系统引入目标检测网络生成潜在动态区域识别框，直接剔除框内所有特征点，但该策略忽略了动态目标周边可能存在的静态特征，如地面标识、建筑边缘等，易在复杂场景中出现跟踪失败的问题，尤其在多动态环境中更为显著。为协调动态环境下的动静态特征，2022 年，Fan 等人^[29]提出 Blitz-SLAM，应用单阶段 BlitzNet^[30]语义分割网络判别静态背景以及动态物体，并结合图像深度信息利用几何约束得到静态背景中的潜在动态特征点，但其对动态点的判定效果仍有较大提升空间。因此，2024 年，Li 等人^[31]提出 DDN-SLAM，通过结合语义检测框架和高斯混合模型，将背景的深度概率进行划分，如果边界框内的动态特征点比例超过设定阈值，则将其

归类为动态物体进行剔除，并应用神经隐式表征^[32]在动态场景中实时构建密集 SLAM。

鉴于单一技术路线的局限性，融合几何与深度学习优势的混合方法成为研究热点。该类方法通常采用“语义预筛+几何精修”的两阶段处理流程，通过深度学习获取语义级动态区域分割后利用几何约束进行特征运动状态判别。2019 年，Xiao L 团队^[33]提出的 Dynamic-SLAM 系统便集成了 SSD 目标检测网络与选择性跟踪算法，通过动态目标检测缩小几何验证范围，但其特征点运动判别模块对极线方向运动的敏感性不足。为此，2022 年，Wu W 团队^[34]设计 Darknet19-YOLOv3 轻量化检测网络与 Depth-RANSAC 联合架构，通过深度信息增强 RANSAC 采样策略的动态点筛选效率，在保持实时性的同时提升位姿估计精度。

几何与深度学习融合的方法虽在动态区域中取得一定突破，但在动态-静态特征交叠区域存在误判风险。为进一步提升动态特征判别精度，研究者们开始探索多模态信息融合机制。2018 年，YUC 团队^[35]开发的 DS-SLAM 系统创新性地融合 SegNet 语义分割网络与运动一致性检验策略，通过分析特征点运动矢量分布差异识别动态区域，但该方法在动态目标沿极线方向运动时，会因投影几何歧义性导致判别失效。同年，Bescos 团队^[36]提出的 DynaSLAM 框架引入 Mask R-CNN 实例分割网络实现像素级动态区域提取并构建多视角几何验证模型，通过分析特征点帧间视差角变化判断其动态属性，该方案虽在多数场景下表现优异，但其在快速运动场景中仍存在误判。

1.3 本文的组织架构

本文从动态场景下视觉 SLAM 研究现状出发，以实现复杂动态场景下移动机器人准确性定位与高精度建图为目标，深入分析复杂动态场景移动机器人面临的挑战，通过动态物体分割网络优化以及动态物体映射与跟踪线程共同构建了一种面向复杂动态场景的视觉 SLAM 算法。该算法有效改善了移动机器人在动态场景下因动态物体干扰与信息丢失导致的跟踪失效、轨迹漂移与建图失真等问题，提高了视觉 SLAM 系统的定位与建图能力。论文总体技术路线如图 1-2 所示。

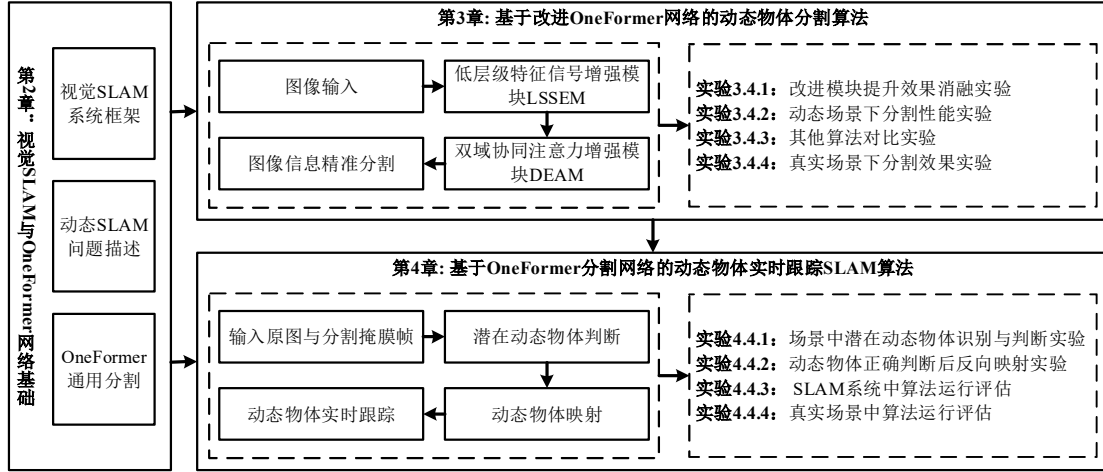


图 1-2 论文总体技术路线

各章内容安排如下：

第 1 章：绪论。本章介绍了动态场景下视觉 SLAM 的研究背景与意义，指出传统 SLAM 在动态环境中因动态物体干扰而导致的定位漂移问题。通过梳理国内外研究现状，对比传统视觉 SLAM 框架在动态场景中的优化方案，引出基于深度学习分割网络的视觉 SLAM 算法。最后对本文的研究内容进行整体概述，并简要介绍了后续章节的组织架构。

第 2 章：视觉 SLAM 与 OneFormer 网络基础。本章首先介绍了视觉 SLAM 系统视觉里程计、后端优化、回环检测与建图等模块的基础理论及工作原理。然后对视觉 SLAM 算法中的动态场景进行描述并分析动态物体对 SLAM 系统产生的影响。最后详细探究 OneFormer 通用分割网络的系统架构，为动态场景下的精确分割与改进奠定理论基础。

第 3 章：基于改进 OneFormer 网络的动态物体分割算法。针对传统视觉 SLAM 在动态场景下动态目标分割精度不足的问题，设计了一种基于改进 OneFormer 网络的动态物体分割算法。该算法以 OneFormer 通用图像分割网络为基础框架，在低层级信号特征增强模块，利用轻量级卷积与特征信号重组强化浅层特征表达，提升了对弱语义特征的判别能力，使网络更聚焦于动态特征提取。在双域协同注意力增强模块，利用通道-空间双域注意力机制自适应增强高维特征动态区域的网络分割权重，提高了网络对动态区域的识别精度。

第 4 章：基于动态物体分割与映射的 SLAM 实时跟踪算法。针对 SLAM 在动态物体占比较大与动态物体包含有利信息不能剔除的场景中定位与建图不准确等问题，在改进 OneFormer 分割网络的基础上设计了一种基于动态物体分割

与映射的 SLAM 实时跟踪算法。算法结合网络分割获取到的相关信息，在潜在动态物体判断阶段，使用相机位姿与物体运动偏转角度估计物体运动状态进而区分动态物体。在动态物体映射与跟踪阶段，将识别得到的动态物体映射回网络分割掩膜图上形成分层结构，通过对动态物体的精准显示与追踪，保留了动态物体上的有效信息，降低了因信息缺失而导致的系统误差，提高了系统位姿估计准确性。

第 5 章：全文总结展望。对全文研究工作进行归纳总结，并在此基础上指出本文算法还存在的不全面之处，对未来的工作进行展望。

第2章 视觉 SLAM 与 OneFormer 网络基础

本章首先介绍视觉 SLAM 系统中视觉里程计、后端优化、回环检测与建图等模块的基础理论及工作原理；然后对视觉 SLAM 算法中的动态场景进行具体描述，并深入分析动态物体对 SLAM 系统性能的影响；最后详细介绍 OneFormer 通用分割网络的系统架构，为动态场景下的精确分割与算法改进提供理论基础。

2.1 视觉SLAM系统框架

视觉 SLAM 作为计算机视觉领域的前沿技术，通过实时环境感知在智能化系统中发挥着重要作用。该技术框架不仅能实现未知环境中的自主定位与导航，还能同步构建高精度的三维场景模型，这些核心功能使其在无人驾驶路径规划、工业动态监测、服务机器人自主导航等典型应用场景中展现出显著的技术优势。特别是在复杂动态环境下的鲁棒性优化与多传感器融合技术方面的突破，进一步拓展了该技术在工程实践中的应用前景。自 SLAM 技术问世以来，经过二十余年的持续创新与发展，目前该技术已趋于成熟，形成了相对完善的理论框架，其相关算法也被广泛应用于各类实际场景中。当前的 SLAM 算法工作流程主要分为四大部分，具体如图 2-1 所示。

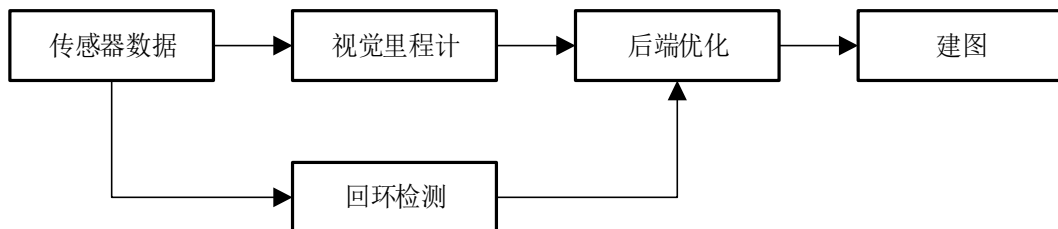


图 2-1 视觉 SLAM 工作流程

视觉里程计作为系统的前端感知模块，承担着实时运动估计的核心任务。该模块通过提取并分析连续图像帧之间的运动特征，运用特征点检测与匹配、光流估计等计算机视觉算法，精确解算相机在三维空间中的相对位姿变化。这一过程不仅为系统提供了高精度的初始位姿估计，同时也为后端优化模块构建了可靠的数据基础。

后端优化模块在系统中发挥着消除累积误差的关键作用。针对视觉里程计在长时间运行中产生的误差累积问题，该模块采用图优化或非线性最小二乘等优化方法，对机器人的运动轨迹与环境地图进行全局优化。通过最小化重投影误差与

位姿图误差，有效校正系统漂移，从而显著提升定位精度与地图的全局一致性。

回环检测模块^[37]作为系统长期稳定运行的核心保障，专门应对机器人在复杂环境中重复访问同一区域时的定位挑战。该模块基于特征描述子匹配和词袋模型等智能识别算法，能够精准识别场景相似性并检测回环事件，其作用不仅体现在能及时修正累积轨迹误差，更能通过闭环约束显著提升系统的全局一致性，从而确保大规模环境下定位结果的可靠性。

建图模块^[38]作为系统的核心输出单元，实现了从感知数据到环境表征的智能化转换。该模块采用多模态地图表达方式，包括二维栅格地图、三维点云地图及语义地图等，通过优化的地图更新与存储策略，为机器人路径规划、动态避障等高层决策提供精确的环境模型。其模块化设计不仅保证了系统的实时性能，更通过良好的可扩展性使视觉 SLAM 系统能够适应工业巡检、服务机器人等多样化应用场景。

四个模块的协同工作构成了一个完整的视觉 SLAM 系统，实现了 SLAM 的核心功能并为机器人自主导航与智能决策提供了坚实的技术支撑。

2.1.1 视觉里程计

视觉里程计作为 SLAM 系统的前端核心模块，在实时运动估计任务中发挥着关键作用。该模块通过高效处理视觉传感器采集的图像序列，基于李群李代数理论构建严格的运动学模型，实现了位姿变换参数的高精度解算。在算法实现层面，视觉里程计主要包含三个核心技术环节，首先通过特征点匹配或光流法建立帧间对应关系，随后利用运动估计模型求解相机位姿变化，最后结合惯性测量参数进行多传感器融合优化。这些环节为 SLAM 系统提供了可靠的位姿估计与运动约束条件，为后续建图与优化奠定了坚实基础。从技术实现方式来看，视觉里程计主要分为基于特征点匹配的间接法和基于光度一致性的直接法两大类。其中，基于 RGB-D 相机的视觉里程计系统因其能直接获取深度信息而具有显著优势：特征点法通过提取并匹配 ORB、SIFT 等特征点实现运动估计，具有较好的鲁棒性；直接法则通过最小化光度误差来求解相机运动，更适合纹理丰富的场景。目前，这些技术方案已在自动驾驶、无人机导航、增强现实等多个领域得到广泛应用。

特征点法中，系统首先通过特征检测算法如 FAST^[39]、ORB^[40]、SIFT^[41]、

SURF^[42]等在图像中定位关键点，随后提取这些特征点的描述子，并通过匹配相邻帧间的特征点对应关系来推算相机位姿变化，这一过程通常包括特征检测、特征描述、特征匹配与位姿估计四个关键步骤。特征点法最显著的优势在于其对光照条件和观测视角变化的强鲁棒性，这使其在静态环境中表现优异，然而当面对动态场景中的运动物体干扰时，特征提取质量会出现明显退化，同时，在特征稀疏的静态场景中，其性能也会显著降低。这种场景依赖性使得特征点法在实际应用中需要根据环境特点进行针对性优化。区别于特征点法，直接法采用了一种基于像素强度梯度的新方法，核心思想是构建光度误差模型，将相机运动参数估计转化为图像序列间的全局光度一致性优化问题。通过分析像素亮度信息的特性，直接法建立了基于原始图像数据的运动约束，避免了传统特征提取过程中的信息损失。在具体实现上，直接法通过最小化相邻帧间的像素灰度误差来实现相机位姿估计，这一特性使其具有两个显著优势：首先，由于不依赖特征点提取，其在处理低纹理场景时表现优异；其次，对运动模糊具有较强的鲁棒性。然而，这种方法也存在一定局限，其对光照条件和视角变化的适应性相对较弱。当前主流的直接法实现方案主要采用光度误差最小化和光度一致性约束两种优化策略，这些方法通过充分利用图像中的光度信息，为运动估计提供了新的解决思路，特别适用于特征点法难以处理的特定场景^[43]。

随着深度学习技术的快速发展，语义视觉里程计的出现为传统运动估计方法带来革命性的变革。该方法的核心突破在于将语义信息引入视觉里程计框架，通过深度神经网络实现场景语义分割，将物体类别信息与传统特征点法或直接法融合，显著提升了运动估计的精度，赋予了系统对动态物体识别和环境语义理解的认知能力。在技术实现层面，语义视觉里程计首先利用语义分割网络对场景中的静态结构和动态物体进行区分，然后将语义信息作为先验知识优化特征点匹配与光度误差计算过程，最终输出具有语义标注的运动估计结果，实现性能提升。这种方法有效解决了传统视觉里程计在动态环境中性能下降的问题，同时为后续的路径规划与决策等深层次任务提供了更丰富的环境语义信息。当前，语义视觉里程计在机器人、自动驾驶等实际应用场景中展现出巨大潜力，正推动视觉 SLAM 技术从单纯的几何重建向具有环境认知能力的智能化系统演进，为机器人在复杂现实环境中的自主导航与智能交互开辟了新的可能性^[44]。

2.1.2 后端优化

在 SLAM 系统的架构设计中, 前端与后端承担着截然不同却又相辅相成的核心任务。前端作为系统的实时感知模块, 主要承担传感器数据处理和初始状态估计的任务。其通过特征提取、数据关联等技术从原始传感器数据中解析出环境特征信息, 为系统提供实时的位姿估计结果, 强调计算效率和实时性, 是后续优化模块处理的基础。后端优化模块承担着全局状态估计与误差校正的功能, 其核心任务是通过概率图模型对前端输出的含噪声观测数据进行联合优化, 进而精确求解机器人六自由度位姿序列与环境地标的三维空间坐标。从功能分工角度分析, 前后端模块形成了良好的互补关系, 前端主要负责视觉特征提取与运动参数解算等几何层面的即时处理, 而后端则基于最大后验估计理论框架, 运用滑动窗口优化或因子图优化等数值计算方法, 实现系统状态变量的全局最优估计。这种分工协作的架构设计, 既保证了系统的实时响应能力, 又确保了定位与建图的全局一致性。

基于上述架构, 卡尔曼滤波算法作为状态估计领域的奠基性估计算法, 蕴含了后端优化的核心思想, 其通过建立在线性高斯系统假设下的递推预测-校正机制, 基于最小均方误差准则实现了状态向量的最优估计。然而, 视觉 SLAM 系统固有的非线性观测模型与运动学方程, 使得传统卡尔曼滤波在工程应用中存在模型失配的局限性。因此, 扩展卡尔曼滤波 (Extended Kalman Filter, EKF) 应运而生, 它通过对非线性系统进行局部线性化近似, 成功地将卡尔曼滤波的思想推广至非线性场景。在 SLAM 技术发展的初期, 扩展卡尔曼滤波因其理论完备性与实现便利性而备受关注, 然而, 随着视觉 SLAM 问题的复杂度不断提升, 扩展卡尔曼滤波逐渐暴露出计算负担过重、对非线性问题处理精度有限等问题, 促使研究者们探索更为高效的后端优化方法^{[45][46]}。

随着技术发展, 现代图优化方法展现出独特优势, 其作为视觉 SLAM 后端优化的重要技术手段, 核心思想是将复杂的系统状态估计问题转化为直观的图结构表示。该方法以顶点表示位姿参数与三维地标坐标等状态变量, 边表示传感器观测方程与运动学约束, 从而构建完整的拓扑模型。基于该模型, 系统可以建立非线性最小二乘问题并利用稀疏矩阵技术进行高效求解, 实现全局状态的最大后验概率估计。这种表示方式不仅保留了问题的本质特性, 还充分发挥了稀疏矩阵

的计算优势，使其成为当前视觉 SLAM 后端处理的主流解决方案。

特别值得关注的是，因子图模型优化方法在传统优化基础上进一步创新，相比传统后端优化方法展现出独特的优势，其能够系统性地融合多源约束信息，并通过增量式优化机制维持解的全局一致性。然而，当环境规模扩大时，特征点数量的超线性增长会导致解算矩阵维度急剧膨胀，显著增加计算资源消耗。为解决这一问题，位姿图优化方法采用状态变量降维策略，通过保留关键帧位姿节点并利用图结构的稀疏特性，有效控制了计算复杂度，实现了相机位姿的全局一致性优化^[47]。

合理选择后端优化方法不仅能够实现对相机位姿与地图点的精确估计，还能显著提升整个 SLAM 系统的性能与稳定性。随着 SLAM 技术的不断发展，后端优化方法也在持续更新，从最初的滤波方法到如今的位姿图优化方法，每一次突破都推动了 SLAM 系统向更高精度、更强鲁棒性的方向迈进。未来，随着计算能力的提升与优化理论的创新，后端优化技术有望在更复杂的场景中发挥更大的作用，为 SLAM 系统的实际应用提供更强大的支持。

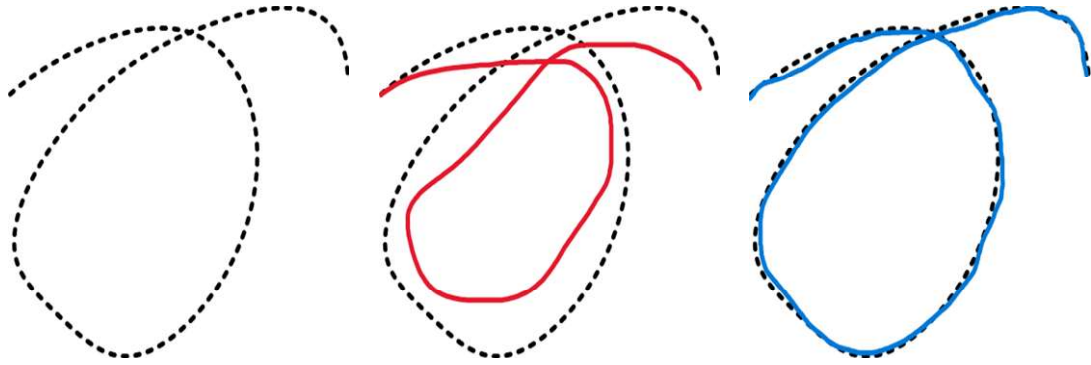
2.1.3 回环检测与建图

在 SLAM 系统中，回环检测^[48]技术扮演着至关重要的角色，其核心功能是解决位姿估计随时间累积的漂移问题。这种漂移现象是 SLAM 系统面临的固有挑战，主要源于传感器测量噪声、运动模型误差以及环境特征提取的不确定性等多重因素的累积效应。随着系统运行时间的逐渐增加，这些误差会不断累积，最终导致估计位姿与实际位姿之间产生显著偏差，严重影响所构建地图的全局一致性。

为有效抑制累积误差的指数级增长，回环检测技术基于场景相似性度量机制，通过识别历史观测数据在空间上的关联性为系统提供全局约束。该技术模块采用视觉词袋模型或深度学习特征匹配方法，建立当前观测与地图先验间的拓扑对应关系，有效校正了里程计漂移导致的系统偏差。

以移动机器人的运动轨迹为例，如图 2-2 所示，当系统仅依赖相邻帧间相对位姿估计时，即使经过局部优化处理，位姿协方差矩阵仍会随时间呈现发散特性。这种误差传播效应导致初始轨迹随运行时间延长逐渐漂移，最终形成如图 2-2(b) 所示的发散轨迹。此时，回环检测机制通过精密算法识别机器人返回原点或熟悉

位置,为轨迹施加新的约束条件,这一约束的引入显著提升了全局位姿估计精度,有效消除累积误差,最终获得如图 2-2 (c) 所示的优化轨迹^[49]。



(a)真实轨迹; (b)真实轨迹于优化后的轨迹; (c)真实轨迹与回环后轨迹

图 2-2 SLAM 轨迹图

回环检测显著提升了位姿估计精度和环境模型的一致性。其通过建立跨时段状态变量间的非线性约束关系,有效抑制了误差传播引发的协方差矩阵发散问题。具体实现方法为通过构建位姿约束残差,将闭环检测问题转化为最小二乘优化问题,从而保证了运动轨迹在 $SE(3)$ 空间中的微分几何特性。这种基于全局拓扑校正的优化方法使系统能够在长时间运行中保持亚厘米级定位精度和三维地图的拓扑准确性。这一技术的成功应用不仅为 SLAM 系统在复杂环境下的持续稳定运行提供了可靠保障,同时也为机器人自主导航和环境探索等实际应用奠定了重要技术基础^[50]。

建图是 SLAM 系统的核心目标^[51],通过多传感器数据融合实现对机器人所处环境的建模。目前主流的地图类型包括稀疏地图、稠密地图和八叉树地图三种典型形式。

稀疏地图的核心思想是基于特征提取算法从环境中选择具有明显区分度的关键点,这些特征点不仅包含精确的几何位置信息,还携带丰富的描述子特征,通过建立特征点之间的关联关系,构建轻量级的语义地图。从系统实现角度分析,稀疏地图具有三重显著优势:首先,其数据存储量通常仅为稠密地图的百分之一到百分之五,极大降低了计算资源消耗;其次,简化的图结构表示使得位姿估计问题可以在低维空间实现高效求解;最后,基于特征点的表示方式天然适合闭环检测和数据关联任务。这些特性使得稀疏地图成为计算资源受限平台中建图的首选解决方案。稠密地图则采用完全不同的构建思路,通过深度传感器或立体视觉系统获取的环境三维点云数据,以规则网格或不规则点集的形式完整记录场景的

几何细节。这种记录方式虽然需要消耗大量存储空间和计算资源，但能够提供毫米级精度的环境分辨率，完整保留场景的几何特征和纹理信息，在需要高精度环境建模的应用场景，如工业检测和虚拟现实等领域中，具有不可替代的重要实用价值。八叉树地图采用创新的层级化空间表示方法，通过三维空间的递推细分机制，在存储效率和表示精度之间实现了智能平衡。该表示方法的核心在于将环境空间划分为多级立方体结构，每个节点精确记录对应三维空间的占用概率和几何属性。这种独特的结构设计赋予了八叉树地图诸多优势：其基于八叉树的数据组织方式确保了高效的空間查询能力，自适应分辨率机制能够根据场景复杂度自动调节细分精度，概率化的占用表示则又有效克服了传感器噪声带来的不确定性影响。在实际工程应用中，八叉树地图凭借其智能化的存储策略，既能满足大规模环境建模的精度要求，又能将内存占用控制在合理范围内。值得注意的是，随着深度学习技术的发展，融合语义信息的八叉树地图正在为机器人系统提供更加丰富和深入的环境理解能力。这三种地图类型并非相互排斥，而是可以根据应用需求进行灵活组合。例如在自动驾驶系统中，可以同时维护稀疏特征地图用于实时定位，稠密地图用于障碍物检测，八叉树地图用于路径规划。这种多层级的地图表示体系，有力地推动了同步定位与地图构建技术向更智能、更高效的方向持续发展^[52]。

2.2 动态SLAM问题

ORB-SLAM3 作为视觉 SLAM 领域的代表性算法，在静态环境下的优异性能已获得广泛验证。然而，当应用于存在动态物体的现实场景时，该算法的性能会出现明显退化，这一现象已成为当前 SLAM 研究领域亟待解决的关键问题。要深入探究动态物体对系统性能的影响机理，需要从 ORB-SLAM3 的核心算法架构展开系统性分析。

SLAM 系统的核心技术流程可以分为前端视觉里程计和后端优化两大部分。前端主要负责特征点的提取与匹配，通过 ORB 特征检测器从连续图像帧中提取具有尺度不变性和旋转不变性的特征点，并基于描述子相似性建立帧间匹配关系。这些匹配特征点构成的几何约束被用于初步估计相机运动。后端优化则采用基于图优化的框架，其中最关键的就是光束法平差（Bundle Adjustment, BA）模块。BA 通过建立重投影误差最小化模型同时对相机位姿和三维地图点位置进行优化

调整。这一优化过程的数学模型如式(2-1)所示：

$$(T^*, p^*) = \arg \min_{(T, p)} \sum_{i=1}^m \sum_{j=1}^n \rho \left(\left\| z_{i,j} - \pi(T_i p_j) \right\|_{\Sigma_{i,j}}^2 \right) \quad (2-1)$$

其中， $T_i \in SE(3)$ 表示第 i 个关键帧的相机位姿，属于特殊欧式群空间； $p_j \in \mathbb{R}^3$ 是第 j 个地图点的三维坐标；观测函数 $\pi: \mathbb{R}^3 \rightarrow \mathbb{R}^2$ 完成从三维空间到二维图像平面的投影变换； $z_{i,j}$ 是实际观测到的特征点坐标； $\Sigma_{i,j}$ 表示该观测对应的协方差矩阵，用于加权不同观测的可靠性；而 $\rho(\cdot)$ 是鲁棒核函数，用于降低异常值的影响。这个非线性最小二乘问题的求解通常采用 Levenberg-Marquardt 算法进行迭代优化。

在理想的静态环境假设下，这一优化框架能够进行良好的位姿估计。然而，当场景中存在动态物体时，其多个关键环节都会受到严重影响。首先在特征提取阶段，动态物体表面与静态特征类似部分也会产生大量稳定的 ORB 特征点，这些特征点会被系统无差别地采集而扰乱系统。在特征匹配过程中，由于动态物体自身的运动特性，其表面特征点在连续帧中会表现出与静态特征完全不同的运动轨迹。例如，一个以速度 v 移动的行人，其表面特征点在图像序列中的运动不仅包含相机运动的视觉误差，还会叠加物体自身运动产生的额外位移，这种复合运动导致的直接后果就是 BA 优化过程中出现严重的异常值干扰。考虑一个典型的动态特征点 p_d ，其在相邻关键帧 k 和 $k+1$ 之间的真实运动可以表示为 $p_d^{k+1} = R_d p_d^k + t_d$ ，其中 R_d 和 t_d 表示动态物体自身的刚体运动，然而系统在 BA 优化时仍然假设这是一个静态点，即期望满足 $p_d^{k+1} = T_{k+1}^{-1} T_k p_d^k$ ，这种错误的静态假设必然导致重投影误差 $\|z_{i,j} - \pi(T_i p_j)\|$ 异常增大。虽然鲁棒核函数 $\rho(\cdot)$ 可以在一定程度上降低这些异常值的影响，但当动态特征点占比较大时，优化结果仍会出现显著偏差，从而对系统稳定性造成影响。

动态物体带来的另一个严重问题是观测遮挡。相机的工作原理本质上是对场景不断采样的过程，而动态物体的运动会不断改变这种采样关系。当动态物体经过相机视野时，会造成两种典型的遮挡情况：一种是永久性遮挡，即静态场景部分被完全遮挡；另一种是动态遮挡，即遮挡区域随时间变化，这两种情况都会显著破坏 SLAM 系统的观测完整性。特别是在基于几何的深度测量过程中，遮挡

会导致匹配特征点对应不同的三维空间点,从而导致计算得到的深度值 z 严重偏离真实值 z_{real} 。这种误差会通过 BA 优化过程在位姿图中持续传播,最终导致系统误差的不断累积放大。

从系统设计的理论层面分析,动态物体的存在实质上破坏了 SLAM 算法所依赖的静态环境基本假设。传统 SLAM 的概率图模型通常将观测噪声建模为零均值高斯分布,而动态物体引入的误差则表现为具有显著偏差的系统性误差。这种本质差异导致基于最大似然估计的光束法平差优化过程会产生根本性的估计偏差。实验数据显示,在典型办公场景中,当存在中等动态干扰时,未经改进的 ORB-SLAM3 系统的绝对轨迹误差(Absolute Trajectory Error, ATE)会从静态环境下的 2cm 显著增加到 15cm 以上,而在高动态的人群密集场景中,轨迹误差进一步增大,最高可达到 50cm^[53]。

针对动态环境下 SLAM 系统面临的挑战,研究者提出了多种创新性解决方案。从技术路线上看,主要包含三类典型方法:基于 RANSAC 的运动估计算法通过统计一致性检验有效区分静态与动态特征点;基于深度学习的语义分割技术直接识别场景中的潜在运动物体;以及融合光流估计与传统 SLAM 框架的方法实时监测和补偿运动干扰提升系统鲁棒性。虽然这些方法显著提升了动态环境中的定位与建图精度,但如何实现算法精度与计算效率的最佳权衡,使 SLAM 系统在动态场景中兼具高可靠性和实用性,仍然是当前研究需要突破的关键技术瓶颈。

2.3 OneFormer通用分割网络

图像分割技术的目标是将像素划分为不同的区域,根据应用需求可分为三种主要任务类型^[54]。语义分割的核心目标是根据语义类别对像素进行分类标注,着重于类别区分而不关注个体实例;实例分割在语义分割基础上进一步细化,需要精确区分同一语义类别中的不同实例,并保持各实例的完整轮廓;全景分割则整合了前两者的特点,既要区分背景区域 *stuff* 和前景物体 *thing* 进行区分,又需准确分割前景中的具体实例,实现对场景的全面解析。这三种分割任务在计算机视觉领域各有侧重,共同构成了完整的图像分割技术体系。

面对这些多样化的分割任务需求,传统解决方案通常需要独立训练多个专用模型,导致系统开发与部署复杂度显著提升。OneFormer 创新性地提出了一种通

用的多任务分割框架，通过引入任务条件化架构设计，使单个模型能够自适应地处理语义分割、实例分割和全景分割三大任务。效果验证实验结果表明，该框架不仅在各项分割任务中均取得较好性能，更通过统一的模型架构大幅降低了系统复杂度。相较于传统多模型方案，OneFormer 可减少约三分之二的计算资源消耗，同时提供高度集成的训练与推理流程，显著提升了图像分割系统的实用性与灵活性。

OneFormer 的核心贡献在于其统一的任务条件查询机制，与传统方法不同，该框架的对象查询是基于任务输入进行的，而不是直接输入到 Transformer 解码器中，这一方法确保了模型的预测能够根据当前任务动态调整，使其在多任务学习中表现卓越，这种创新对于构建一个真正通用的多任务图像分割模型至关重要，使得 OneFormer 能够在语义、实例与全景分割任务上均达到最优性能。OneFormer 的设计灵感来源于 MaskFormer，后者提出了一种通用架构，并在单一模型结构下训练不同的分割任务，实现了性能领先。然而，MaskFormer 的一个主要局限是它无法在同一个模型中同时处理多个任务，仍然需要对每个分割任务分别训练独立的模型，OneFormer 克服了这一限制，它仅凭一个模型与一次训练过程，即可完成多个分割任务，并在所有任务中实现最佳性能。这一突破性进展使得多任务图像分割的实现更加高效与低成本。

OneFormer 采用了任务引导的联合训练策略，在训练过程中，它采用统一的方式对任务的地面真值标签进行抽样，并从全景标注数据中提取所有的对比标签，通过任务引导，该模型使用文本输入来进行任务条件化，这些输入采用“任务是 XX”这一格式，其中“XX”可以是语义、实例或全景分割。由于不同任务的样本地面真值依赖于任务的类别，因此 OneFormer 又引入了对比损失，以确保对象查询能够与从地面真值标签中派生出的输入文本列表对齐，从而提升分割精度。

除了在所有图像分割任务上达到最优性能外，OneFormer 还采用了扩展领域保持方法，进一步优化了基于标准骨干网络的模型表现。另一个显著特性是其任务动态性，即在推理阶段可以动态调整任务设置，若任务值设定为语义分割、实例分割或全景分割，最终的输出会根据相应的后处理步骤进行调整，这一特性使得 OneFormer 适用于多种场景，在适用范围上远超传统方法。

OneFormer 独特的任务引导查询机制与单一联合训练策略，使其能够在语义、实例与全景分割任务中同时实现最优表现。此外，相比于单独训练多个模型的方

案，OneFormer 减少了约三分之二的训练时间、存储需求与推理成本，从而大幅提升了图像分割的通用性与可获取性。作为一种创新性的多任务通用图像分割框架，OneFormer 为计算机视觉领域提供了一种更加高效、经济且易于部署的解决方案。OneFormer 总体框架架构如图 2-3 所示。

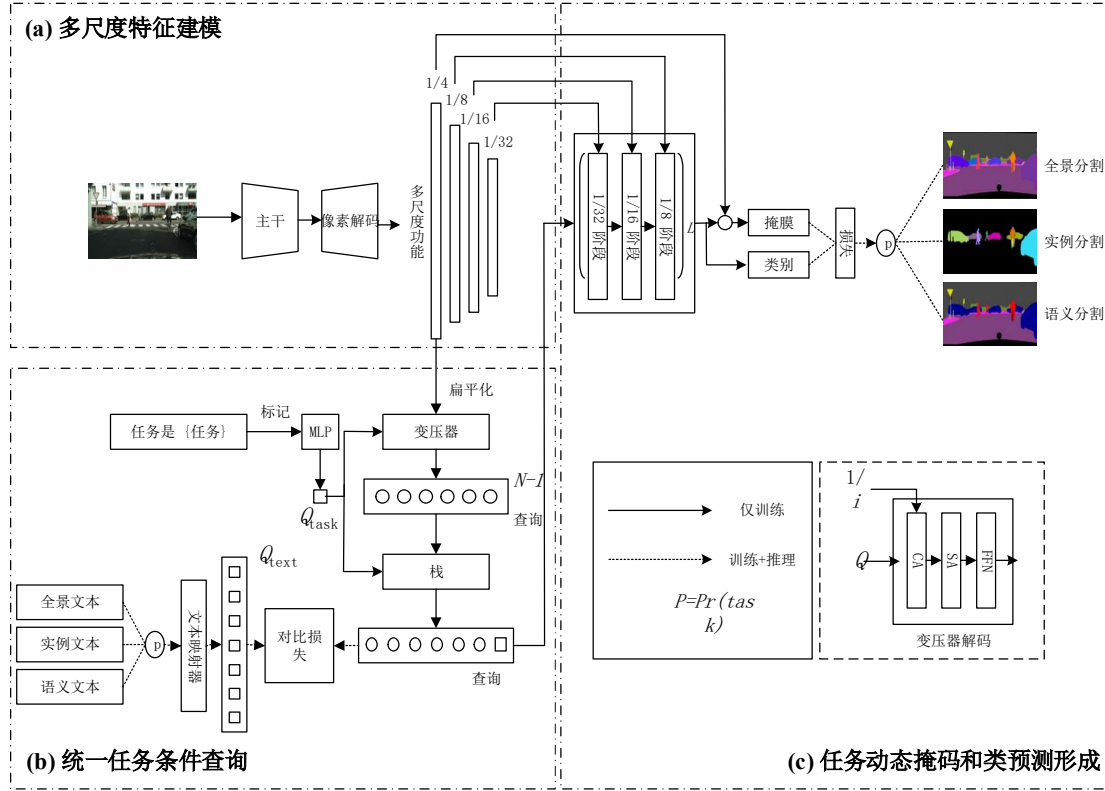


图 2-3 OneFormer 框架架构

其中，图 2-3(a)表示多尺度特征建模阶段，主干提取输入图像的多尺度特征后使用像素解码器。图 2-3(b)表示统一任务条件查询阶段，OneFormer 网络在任务令牌 Q_{task} 的指导下，针对变压器内部制定统一的 $N-1$ 个任务条件对象查询集与扁平的 $1/4$ 尺度特征，将 Q_{task} 与来自转换器的 $N-1$ 个查询连接起来，在训练过程中统一以 $p=1/3$ 对任务进行采样，并使用文本映射器生成相应的文本查询 Q_{text} 。为了计算查询-文本对比损失来学习任务间的区别，可以在推理过程中删除文本映射器，从而使模型参数更有效。图 2-3(c)表示任务动态掩码与类预测形成阶段，OneFormer 网络使用多级变压解码器来获得任务动态类与掩码预测。

网络中的文本映射器结构如图 2-4 所示，OneFormer 使用 6 层转换文本编码器对输入文本列表 Q_{pad} 进行标记与编码，将一组 N_{ctx} 嵌入到编码中，从而获得 N 个文本查询 Q_{text} 。

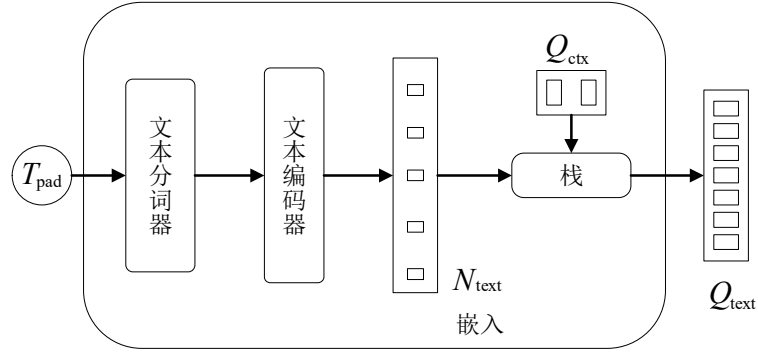


图 2-4 文本映射器

2.4 本章小结

本章围绕视觉 SLAM 技术与 OneFormer 通用分割网络展开系统性介绍。首先，通过对视觉 SLAM 系统的视觉里程计、后端优化、回环检测与建图四大核心模块进行介绍，详细描述了各模块的功能定位及其协同工作机制。其次，基于视觉 SLAM 数学模型分析了动态物体对系统状态估计的影响，重点讨论运动目标位姿估计误差累积及建图不准确等关键问题。最后，系统介绍了改进所使用的 OneFormer 通用分割网络的创新架构及其在多任务分割中的性能优势，论证其在 SLAM 系统中提供高质量语义信息的可行性，为构建基于语义约束的 SLAM 系统奠定了理论基础。

第 3 章 基于改进 OneFormer 网络的动态物体分割算法

本章针对传统视觉 SLAM 在动态场景中动态目标分割精度不足的问题，设计了一种基于改进 OneFormer 网络的动态物体分割算法。该算法以 OneFormer 通用图像分割网络为基本架构，在低层级信号特征增强模块，通过引入轻量级卷积与特征重组机制强化浅层特征表达能力，有效提升了对弱语义特征的判别性能，使网络更加专注于动态特征的提取。在双域协同注意力增强模块，利用通道-空间双域注意力增强机制提升动态区域的特征响应权重，显著提高了网络对动态目标的识别精度。分割网络总体结构图如图 3-1 所示。

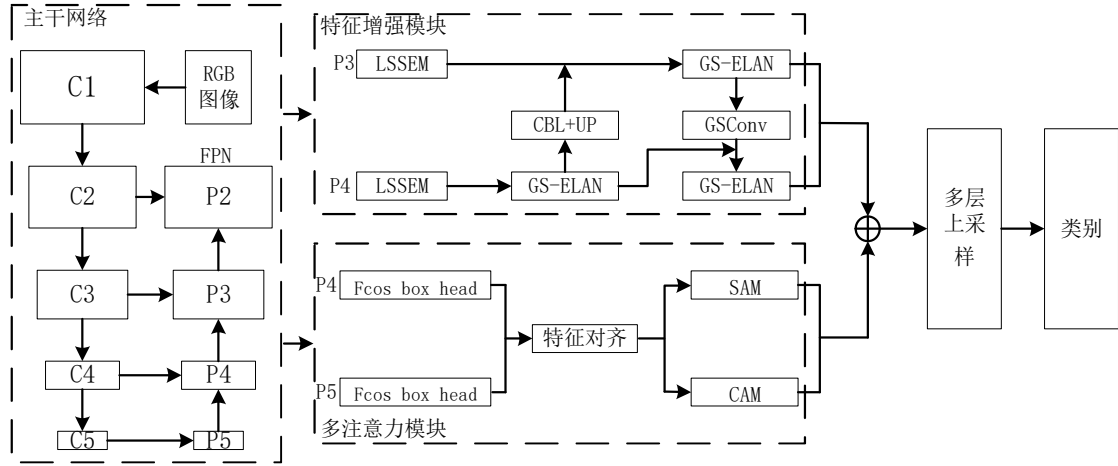


图 3-1 OneFormer 分割网络改进模块示意图

3.1 低层级特征信号增强模块LSSEM

3.1.1 轻量级卷积模块 GSConv

标准卷积（Standard Convolution，SConv）通过全通道点积运算实现特征提取，其输出特征维度由卷积核数量决定。然而，这种密集的卷积操作会产生大量冗余特征，不仅显著增加模型参数量，还会在线性变换过程导致明显的通道信息衰减问题。为此，本章采用轻量级卷积模块 GSConv 替代标准卷积，该模块通过优化卷积运算方式，在保持特征提取能力的同时显著降低了网络计算参数量，GSConv 具体结构框架示意图如图 3-2 所示。

针对标准卷积操作中存在的通道信息衰减问题，GSConv 通过引入深度可分离卷积与通道混洗的协同优化机制实现特征表达能力的增强。具体而言，在架构

设计方面，GSConv 采用了 5×5 大尺寸深度卷积核来替代传统的小尺寸卷积核，在保持参数效率基本不变的前提下，显著扩展了卷积操作的局部感受野，使得网络对小尺度目标的特征捕获能力得到明显提升。在特征融合机制方面，GSConv 引入了通道混洗操作，该操作通过在通道维度上进行有规律的置换重组，实现了不同卷积核间的特征信息交互。这种协同优化机制既能打破传统卷积操作中通道间的信息壁，促使来自不同分支的特征图进行充分的信息交流，又能通过控制混洗的粒度，使分割网络在计算复杂度与特征融合效果之间取得良好平衡。此外，通道混洗操作的引入还进一步优化了图像特征的信息流路径，使得网络在保持轻量化的同时获得了更丰富的特征表示能力。

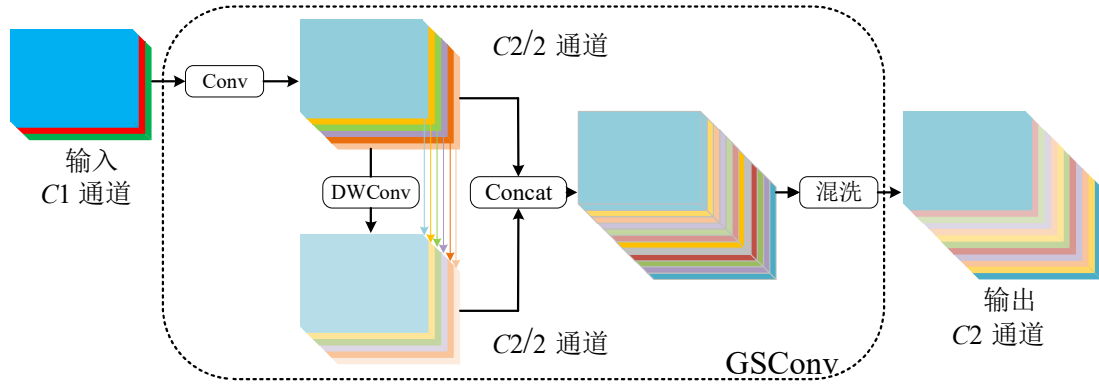


图 3-2 GSConv 结构示意图

GSConv 针对嵌入式 GPU 架构的计算特性也进行了优化设计。在内存访问方面，通过通道压缩策略，将中间特征图的通道维度缩减 50%，使得 DRAM 的访问频次降低约 40%，有效缓解了嵌入式平台常见的内存带宽瓶颈问题。在存储器结构层次方面，通过调整特征图的存储布局，使得中间结果能够更好地利用共享内存与动态核融合技术，从而将多个连续操作合并为单个 CUDA 核函数，显著降低了内核启动开销。

3.1.2 特征信号重组模块

在深度学习分割领域，精准的物体检测分割依赖于早期低层级特征图的空间细节信息。然而由于网络浅层卷积核的堆叠深度不足，这些特征图所携带的语义信息较少，受限于局部感受野的约束，导致模型在复杂动态场景中易出现物体漏检、边缘模糊及纹理混淆等问题。特别是在自动驾驶与 SLAM 定位建图等任务中，特征表达环节目标尺度的剧烈变化格外凸显了传统层级结构的局限性。

本章设计了一种低层级特征信号增强模块（Low-level Semantic Signal

Enhancement Module, LSSEM), 用于提升网络对低层级特征的表达能力并优化感受野范围。模块基于金字塔网络, 结合多尺度池化策略构建多分支架构, 从而拓展网络的横向感知能力, 增强对弱语义特征的提取效果, 提高分割任务中的特征表达能力。模块的具体结构如图 3-3 所示。

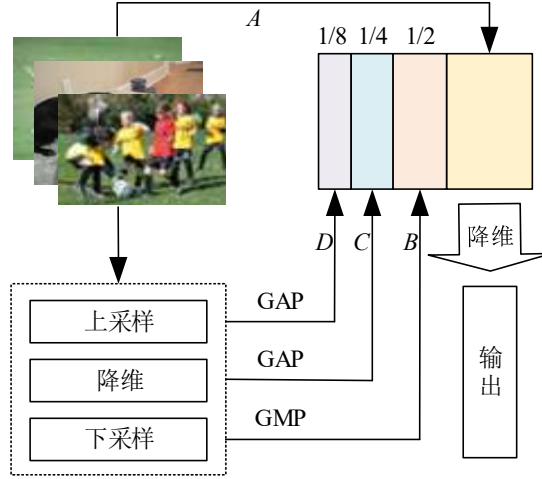


图 3-3 Oneformer 低层级特征信号增强模块 LSSEM

其中, 支路 A 为原尺度特征保留环节, 直接对输入特征图 $P \in \mathbb{R}^{W_{in} \times H_{in} \times C_{in}}$ 进行 1×1 卷积, 以调整通道数为目标输出特征维度 C_{out} , 同时保留原始分辨率的空间细节信息; 支路 B 为粗粒度特征提取环节, 通过全局最大池化(Global Max Pooling, GMP) 对特征图进行下采样至 $1/2$ 尺度, 随后通过 1×1 卷积压缩通道至 $C_{out}/16$ 并借助双线性插值恢复至初始分辨率, 增强对显著目标区域的响应能力; 支路 C 与 D 为细粒度上下文建模环节, 通过全局平均池化(Global Average Pooling, GAP) 分别在 $1/4$ 与 $1/8$ 下采样率下提取多粒度感受野特征, 然后采用 1×1 卷积核进行压缩比为 $1:16$ 的通道维度降维, 进而通过双线性插值填充上采样恢复图像至原始输入分辨率, 最终通过通道拼接实现跨尺度特征融合, 通过 1×1 卷积完成特征通道的维度对齐与信息整合统一, 形成具有空间-通道双重注意力机制的特征表达体系。LSSEM 的计算过程如式(3-1)、(3-2)与(3-3)所示:

$$F_{GMP}^{1/2^n}(P) = F_{GMP}^{1/2^n} \begin{bmatrix} W_p \\ H_p \end{bmatrix} = \begin{bmatrix} W_n \\ H_n \end{bmatrix}, (n = 1, 2, 3) \quad (3-1)$$

$$nT_i = J \begin{bmatrix} W_i \\ H_i \end{bmatrix} = \begin{bmatrix} \frac{W_i + 2p - k}{s} + 1 \\ \frac{H_i + 2p - k}{s} + 1 \end{bmatrix}, (i = 1, 2, 3; w = h = 1) \quad (3-2)$$

$$Q = J \left\{ S_{\text{ff}} \left[y_{\text{up}}(nT_1), y_{\text{up}}(nT_2), y_{\text{up}}(nT_3), P \right] \right\} \quad (3-3)$$

其中, W 与 H 表示输入特征图 P 相关参数; $F_{GMP}^{1/2}$ 表示全局最大池化、 $F_{GAP}^{1/4}$ 与 $F_{GAP}^{1/8}$ 表示全局平均池化, 上角标表示池化的特征图缩放倍率; J 表示卷积核维度为 $w \times h$, 步长为 h , 填充为 p 的常规卷积, 本文选取卷积核为 1×1 , 即 $w=h=1$; $y_{\text{up}}(nT_N)$ 表示双线性插值填充上采样过程; nT_1, nT_2, nT_3 表示 3 条尺度不同的池化特征金字塔的输出; Q 表示低层级特征信号增强模块 LSSEM 得到的新特征图。

通过 LSSEM 模块的处理, 在保持分辨率不变的前提下, 主干网络采集到的低层级输出特征图显著增强, 提升了网络对复杂场景中目标特征的捕捉与解析能力, 优化了网络整体的特征提取性能。

3.2 双域协同注意力增强模块 DEAM

针对 OneFormer 网络在特征融合效率上的不足, 本章设计了一种双域协同注意力增强模块 (Dual-domain Attention Enhancement Module, DAEM)。该模块通过构建级联式的空间注意力机制 (Spatial Attention Mechanism, SAM) 与通道注意力机制 (Channel Attention Mechanism, CAM), 实现了对特征图在空间域与通道域的动态优化。模块的整体架构如图 3-4 所示。

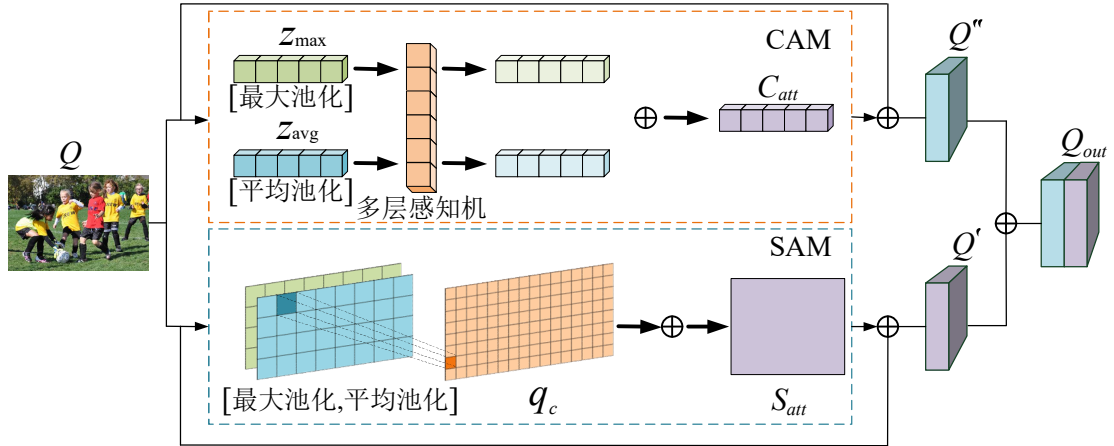


图 3-4 双域协同注意力增强模块 DEAM

DEAM 结合多尺度特征融合与跨维度信息交互, 能有效增强模型对动态目标的特征表达能力, 使网络能够在不同层级特征空间内建立更精细的注意力分布, 从而优化网络的特征提取与表达能力, 提升在复杂动态场景中的网络分割精度。

3.2.1 空间注意力机制 SAM

空间注意力机制的核心目标是通过像素级权重动态分配，强化特征图中具有显著语义信息的空间区域。具体而言，给定输入特征图 $Q \in \mathbb{R}^{C \times H \times W}$ ，采用全局平均池化(Global Average Pooling, GAP)与全局最大池化(Global Max Pooling, GMP)对特征图进行双模态特征提取，生成两组通道增强的特征描述符，具体如式(3-4)与(3-5)所示：

$$F_{GAP}(Q) = \frac{1}{H \times W} \sum_{i=1}^H \sum_{j=1}^W Q_c(i, j) \quad (3-4)$$

$$F_{GMP}(Q) = \max_{(i,j) \in \Omega} Q_c(i, j) \quad (3-5)$$

其中， Ω 表示特征图空间维度， $c \in \{1, 2, \dots, C\}$ 表示通道索引。通过通道维度拼接上述特征，构建多尺度融合特征 $q_c \in \mathbb{R}^{2C \times 1 \times 1}$ ，其计算过程如式(3-6)所示：

$$q_c = \mathcal{C}_{concat} [F_{GAP}(Q), F_{GMP}(Q)] \quad (3-6)$$

采用包含双层卷积的权重生成网络对 q_c 进行空间维度映射，生成空间注意力图 $S_{att} \in \mathbb{R}^{1 \times H \times W}$ ，具体如式(3-7)所示：

$$S_{att} = \sigma(F_{conv}^{3 \times 3}(q_c)) \quad (3-7)$$

其中， $\sigma(\cdot)$ 为 Sigmoid 函数， $F_{conv}^{3 \times 3}$ 表示 3×3 卷积操作。最终，通过将空间注意力图与原始特征图进行自适应加权，输出优化后的特征图 $Q' = Q \otimes S_{att}$ 。

3.2.2 通道注意力机制 CAM

通道注意力机制 CAM 的核心目标是通过动态校准通道权重，解决多尺度特征融合中的通道冗余问题。对于输入特征图 $Q \in \mathbb{R}^{C \times H \times W}$ ，通过全局平均池化 GAP 与全局最大池化 GMP 操作提取通道级统计特征，具体如式(3-8)与(3-9)所示：

$$z_{avg} = \frac{1}{H \times W} \sum_{i=1}^H \sum_{j=1}^W Q(i, j) \quad (3-8)$$

$$z_{max} = \max_{(i,j) \in \Omega} Q(i, j) \quad (3-9)$$

其中， $z_{avg} \in \mathbb{R}^C$ 表示通过全局平均池化 GAP 对空间维度进行平均操作后生成的全局平均响应向量，反映了各通道的整体特征响应； $z_{max} \in \mathbb{R}^C$ 表示通过全局最大池化 GMP 提取每个通道的空间最大值后生成一个反映局部显著特征的统计向量。为了提升通道之间的非线性交互能力，本节设计了一个共享权重的多层

感知机 MLP 网络结构。该网络将 z_{avg} 与 z_{max} 分别作为输入，通过两个全连接层 FC 与 ReLU 激活函数对特征进行变换与优化，从而进一步增强特征表达与模型性能。具体如式(3-10)与(3-11)所示：

$$w_{avg} = W_2 \cdot ReLU(W_1 \cdot z_{avg}) \quad (3-10)$$

$$w_{max} = W_2 \cdot ReLU(W_1 \cdot z_{max}) \quad (3-11)$$

其中， $W_1 \in \mathbb{R}^{C/r \times C}$ 与 $W_2 \in \mathbb{R}^{C \times C/r}$ 分别表示权重矩阵； r 为压缩比，通常设为 16。中间层采用瓶颈结构以减少计算复杂度，ReLU 函数在该过程中引入非线性特性提升模型的表达能力。将两条路径的输出元素 w_{avg} 与 w_{max} 相加并通过 Sigmoid 函数计算得到通道注意力权重 $C_{att} \in \mathbb{R}^{C \times 1}$ ，具体如式(3-12)所示：

$$C_{att} = \sigma(w_{avg} + w_{max}) \quad (3-12)$$

其中， $\sigma(\cdot)$ 将权重映射到[0,1]区间，用于表示各通道的重要性。将 C_{att} 与输入特征图在通道维度上进行加权操作，最终得到优化后的特征图 $Q'' = Q \odot C_{att}$ 。

3.2.3 双域注意力协同工作机制

SAM 与 CAM 模块采用并行-级联混合架构，此机制通过建立空间-通道耦合关系，能有效改善传统注意力机制中维度信息相对孤立的问题。首先在特征预处理环节，将输入特征图分别输入 SAM 与 CAM 分支，完成空间域与通道域的独立权重计算，然后将空间注意力图 S_{att} 与通道注意力图 C_{att} 进行矩阵点乘，生成联合注意力图 $A_{joint} = S_{att} \times C_{att}$ 进行特征融合，最后利用 A_{joint} 对原始特征进行联合加权，输出优化特征 $Q_{out} = Q \otimes A_{joint}$ ，完成特征重构。

3.3 实验结果与分析

本章仿真实验平台配置为：Ubuntu18.04 版本 Linux 操作系统，AMD Ryzen 5 5500 CPU，主频 3.6GHZ，内存 16GB，RX6750GRE GPU 显卡，显存 24GB。

为验证改进后分割算法的有效性，选择在慕尼黑工业大学 CV 组发布的 TUM 标准数据集中进行测试。该数据集利用 Kinect v2 设备在室内结构化场景中采集，包含时序对齐的 640×480 像素 RGB 影像与 16 位深度信息，同时集成采样率为 100Hz 的 IMU 模块，深度传感器经过严格标定，相对测距误差不超过 1%。作为计算机视觉领域的权威测试集，其多模态数据特性与严谨的测量标准使其成为视

觉 SLAM 算法性能评估的首选基准, 目前已在相关领域积累超过万次研究引用, 成为视觉 SLAM 领域的主流评测数据集。本文选取了 TUM 数据集中动态特征最明显的图像序列 fr3_walking_xyz。

3.3.1 改进模块提升效果验证实验

评价网络分割效果的指标主要有:

(1) 交并比(Intersection Over Union, IOU): 作为网络分割任务的核心评价指标, 其数学定义为: 针对特定语义类别, 预测分割区域与真实区域在像素空间中的交集与并集面积的归一化比值。该指标通过计算预测掩膜与真值之间的重叠度, 量化模型输出的空间定位精度, 数值越趋近于 1, 预测结果与真实标注的空间一致性越高。全局平均交并比 (mean IOU, mIOU) 则通过计算所有语义类别 IOU 的算术平均值, 综合评估模型的全类别分割性能。根据计算机视觉领域经验准则, 当 mIOU 超过 0.5 阈值时, 可判定模型具备基本的分割效能, 不同应用场景可能设置差异化评价标准。第 i 类交并比值 $IOU^{[i]}$ 计算公式如式(3-13)所示:

$$IOU^{[i]} = \frac{p_i^i}{\sum_{j=0}^{20} p_j^i + \sum_{j=0}^{20} p_j^i - p_i^i} \quad (3-13)$$

(2) 像素准确率(Pixel Accuracy, PA): 一种全类别综合指标, 作为语义分割的基础性评价指标, 其数学定义为: 模型在测试集上正确归类的累计像素数量与测试集像素总数量比值, 在混淆矩阵的理论框架下, 该指标等价于主对角线元素的算术和与矩阵所有元素之和的商值, 该指标能直观反映整体分类精度。PA 计算公式如式(3-14)所示:

$$PA = \frac{\sum_{i=0}^{20} p_i^i}{\sum_{i=0}^{20} \sum_{j=0}^{20} p_j^i} \quad (3-14)$$

为评估本文提出的改进对算法性能的影响, 本节基于控制变量法在 TUM 数据集的 fr3_walking_xyz 动态数据集上设计了系统性消融实验, 将本节算法与改进前原始 OneFormer 进行 PA、mIOU 与精确率曲线对比, 结果如图 3-5、3-6 与 3-7 所示。

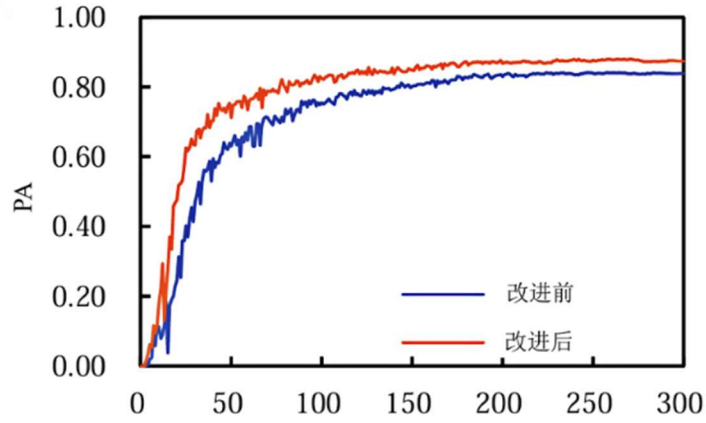


图 3-5 改进前后 PA 值对比

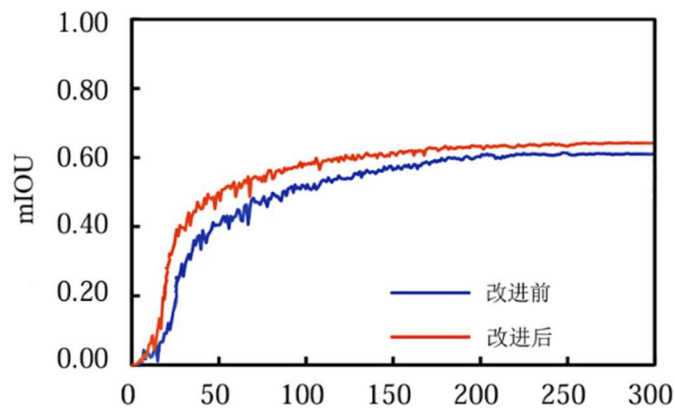


图 3-6 改进前后 mIOU 值

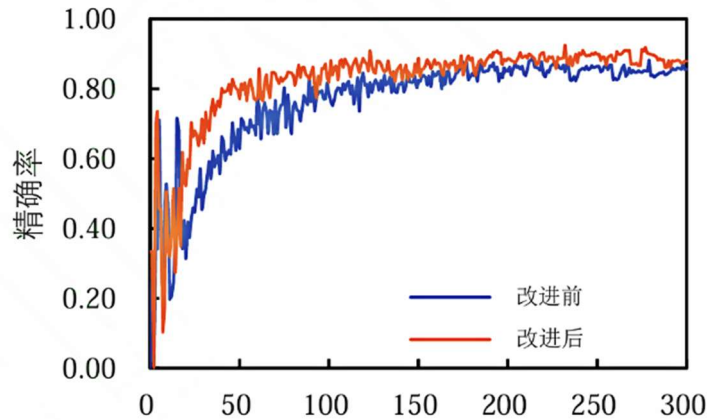


图 3-7 改进前后精确率对比

为系统验证算法改进的有效性,本节在保持相同训练策略的条件下,以改进前原始 OneFormer 网络作为基线模型,通过渐进式引入 GSConv 轻量卷积模块、LSSEM 低层级信号增强模块和 DEAM 双域注意力模块进行对比实验分析。其结果如表 3-1 所示。

表 3-1 消融实验

GSConv	LSSEM	DEAM	PA/%	mIOU/%
			84.9	57.5
√			86.8	58.9
	√		85.4	57.8
		√	86.4	58.1
√	√	√	88.6	60.4

如表 3-1 第一行所示, 原始 OneFormer 在 fr3_walking_xyz 数据集上的像素准确率 PA 为 84.9%, 平均交并比 mIOU 为 57.5%。引入 GSConv 模块后数据如表 3-1 第二行所示, 该模块可以有效提高网络的特征融合能力, 减少网络参数量, 引入后像素准确率 PA 与平均交并比 mIOU 分别提升至 86.8%与 58.9%, 相对于原始 OneFormer 在 PA 指标上提升 2.2%、mIOU 指标上提升 2.4%。添加 LSSEM 模块后数据如表 3-1 第三行所示, 该模块可以有效增强网络对前景信息的特征提取能力, 引入后像素准确率 PA 与平均交并比 mIOU 分别提升至 85.4%与 57.8%, 相对于原始 OneFormer 在 PA 指标上提升 0.6%、mIOU 指标上提升 0.5%。添加 DEAM 模块后数据如表 3-1 第四行所示, 该模块能增强网络对动态物体的分割注意力, 使其对动态物体的表征能力有较大提升, 引入后像素准确率 PA 与平均交并比 mIOU 分别提升至 86.4%与 58.1%, 相对于原始 OneFormer 在 PA 指标上提升 4.4%、mIOU 指标上提升 5.1%。三个改进模块协同工作时数据如表 3-1 第五行所示, 融合后像素准确率 PA 与平均交并比 mIOU 最终提升至 88.6%与 60.4%, 相对于原始 OneFormer 在 PA 指标上提升 4.4%、mIOU 指标上提升 5.1%。综上所述, 本节算法融合了各个模块的优点, 分割精度取得大幅提升。

3.3.2 动态场景下分割性能实验

本节选取 TUM 公开数据集中的动态数据集 fr3_walking_xyz 作为实验基准, 系统验证所提算法在不同动态强度场景下的分割性能提升效果。该数据集以 1280×960 的高清分辨率采集了包含复杂动态干扰的典型室内办公场景, 为算法评估提供了理想的测试环境。在实验设计方面, 本节重点选取数据集中两个沿电脑桌做周期性移动的人体目标作为主要动态干扰源, 通过精心截取动态目标在图

像平面占比呈现梯度递减趋势的五个典型帧（帧号为#727、#126、#577、#182、#318），构建了从高动态占比到低动态占比的完整测试序列。这种渐进式的实验方案能够全面、系统地评估算法在不同动态干扰强度下的分割鲁棒性。改进前后算法在构建的典型动态场景下的分割效果对比情况如图 3-8 所示，其中高动态占比帧主要验证算法对显著运动目标的处理能力，而低动态占比帧则重点考察算法对微弱动态干扰的敏感性。

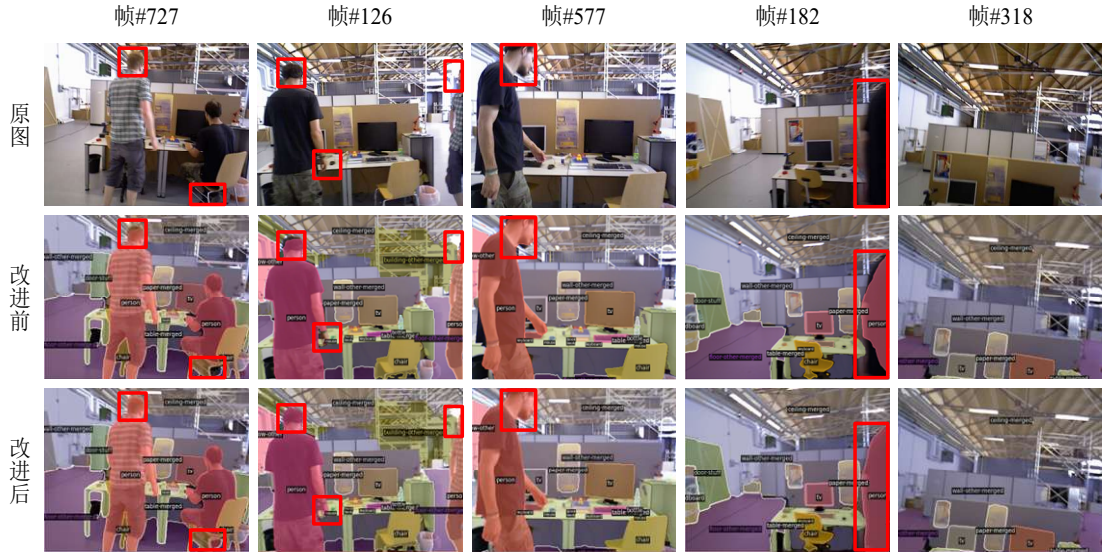


图 3-8 改进前后动态场景下分割效果图

图 3-8 第一行展示了数据集原始 RGB 图像帧，清晰呈现了动态人体与静态办公场景的空间位置关系。图 3-8 第二行展示了改进前原始 OneFormer 分割网络的分割结果，可以看出其在动态目标边缘与运动模糊处出现明显分割缺失；图 3-8 第三行展示了本章通过引入低层级特征信号增强模块 LSSEM 与双域协同注意力增强模块 DEAM 而改进的增强型 OneFormer 分割网络的分割效果，可看出所有的动态人体均完整分割，显著提升了 OneFormer 在动态目标分割上的完整性，特别是在人体四肢等细节区域上。图中红色方框着重标定了算法改进前后的关键差异区域。

3.3.3 不同算法对比实验

本节基于 TUM 公开数据集中的动态数据集 fr3_walking_xyz 开展算法对比实验，实验选取 YOLACT、Mask R-CNN、改进前原始 OneFormer 及改进后 OneFormer 四种分割算法进行多场景分割效果对比实验，该数据集充分集合了动态场景的两个核心问题：高速运动导致的目标区域模糊与多目标交错运动的动态

遮挡。实验具体效果通过图 3-9 中的四组典型场景直观呈现，图中红色方框特别标注了四种算法在这两个典型问题融合区域的响应情况。

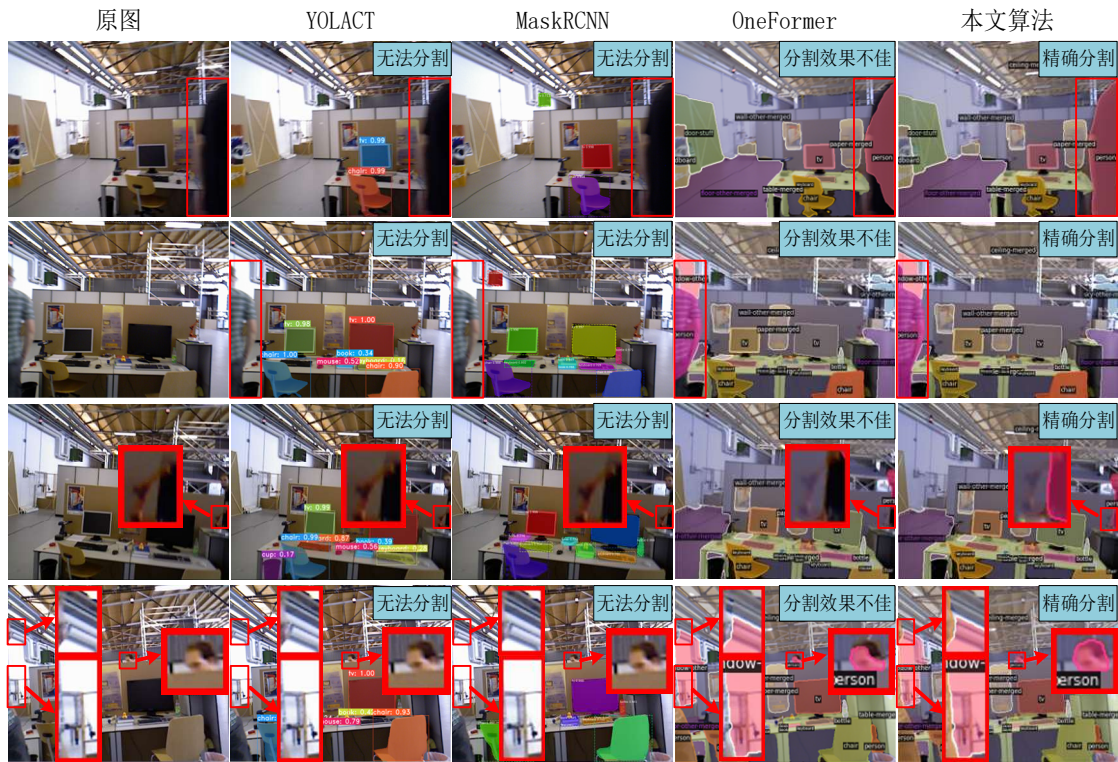


图 3-9 TUM 数据集下 4 种算法分割效果对比图

如图 3-9 第二列与第三列所示，YOLACT 与 Mask RCNN 在动态场景下存在性能局限，两者对场景中的物体均不能识别分割完全，尤其是在当处理图像中出现运动模糊的行人和图像中出现行人与周围环境之间有较大遮挡的情况时，YOLACT 与 Mask RCNN 更是出现了直接无法识别分割的问题。这主要源于两者在网络架构上的缺陷，缺乏专门的目标特征增强机制与有效的注意力增强机制，在特征表征学习中未能建立有效认知模型，因而在分割时不能保证物体边界的完整性，造成在分割过程的中断与分割质量的下降。原始 OneFormer 算法分割效果图如图 3-9 第四列所示，虽然对场景中物体分割精度有了很大提升，针对静态物体可以良好分割，但在动态物体分割方面仍存在明显不足，实验图像显示其在动态物体边缘特征不明显处出现分割缺失。本章提出的改进 OneFormer 分割网络通过双路径架构实现了突破性优化：在上层设计了一种特征增强模块对低层级信息进行增强，同时，融入了一个多注意力模块用于增强算法对主体信息的捕捉能力，削弱其他非必要信息影响，增强分割的效率与准确性，分割效果如图 3-9 第五列所示，场景中所有静态与动态物体均完整分割。

为进一步验证实验效果，实验得出的 PA 值、mIOU 值定量评估数据对比情况如表 3-2 所示。

表 3-2 4 种算法 AP 值对比（单位：%）

Algorithm	PA/%	mIOU/%
Yolact ^[55]	74.7	50.6
MaskRCNN ^[56]	77.1	52.2
Oneformer	84.9	57.5
本文算法	88.6	60.4

数据表明，YOLACT 与 Mask RCNN 在像素分割准确率 PA 与交并比 mIOU 上分别仅有 74.7%、77.1%与 50.6%、52.2%。改进前原始 OneFormer 像素分割准确率 PA 与交并比 mIOU 为 84.9%与 57.5%，改进后 OneFormer 像素分割准确率 PA 与交并比 mIOU 提升至 88.6%与 60.4%，与 Yolact 相比像素分割准确率 PA 取得了 18.6%的提升，交并比 mIOU 取得了 19.4%的提升，与 MaskRCNN 相比像素分割准确率 PA 取得了 14.9%的提升，交并比 mIOU 取得了 15.7%的提升；与改进前原始 OneFormer 相比像素分割准确率 PA 与交并比 mIOU 分别提升 4.4%与 5.0%。

3.3.4 真实场景下分割效果实验

为全面验证本文算法的实际应用性能，在室内外真实动态场景中分别开展系统性测试。室内外真实场景下分割效果如图 3-10 所示。图 3-10(a)与(b)呈现的是以实验室为背景的室内动态测试场景，该场景不仅包含两组以 0.5-1.8m/s 不同速度随机移动的实验人员，还专门设置了两种典型的动态交互对象：被实验人员拿起并随身携带而产生不规则运动的双肩背包与被实验人员推动而产生连续位移的旋转办公椅。值得注意的是，场景中同时存在快速挥动手臂、突然改变行进方向等非规律运动状态。图 3-10(c)与(d)则展示了更具挑战性的大学校园主干道室外动态场景，该场景不仅包含以 0.5-2.5m/s 不同速度行走或奔跑的行人群体，还涉及电动车、轿车等多种移动载具的复杂交互，测试场景中特别设计了行人横穿马路、多辆电动车并排行驶、轿车紧急制动等高动态复杂工况。实验结果表明，本章提出的算法在这些真实场景的各种挑战性条件下均能保持优异的分割性能，准确识别并分割各类动态目标。

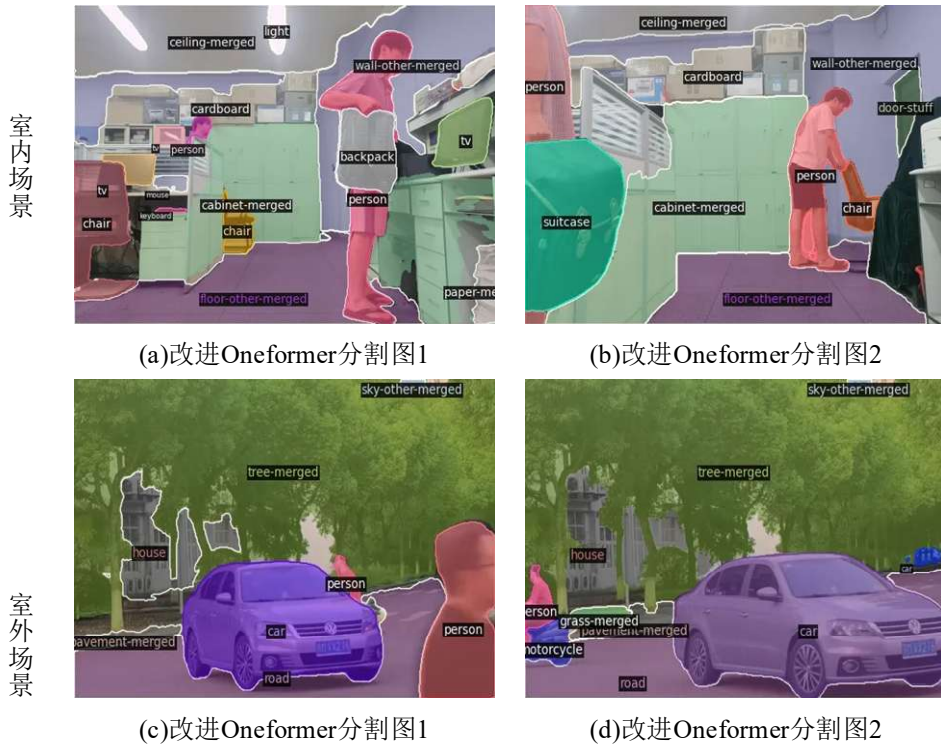


图 3-10 室内外真实场景下分割效果

3.4 本章小结

本章针对传统视觉 SLAM 前端在动态目标影响下对场景中物体分割精度不足的问题，设计了一种基于改进 OneFormer 网络的动态物体分割算法。该算法以 OneFormer 通用图像分割网络为基础框架，在低层级信号特征增强模块，利用轻量级卷积与特征信号重组强化浅层特征表达，提升了对弱语义特征的判别能力，使网络更聚焦于动态特征提取。在双域协同注意力增强模块，利用通道-空间双域注意力机制自适应增强高维特征动态区域的网络分割权重，提高了网络对动态区域的识别精度。在 TUM 数据集下进行改进效果验证实验，结果表明，所提算法的像素准确率 PA 与平均交并比 mIOU 分别达到 88.6%与 60.4%，相对于改进前实现 4.4%与 5.1%的绝对提升，显著提高了动态物体的识别精度，证明了本章算法的有效性。

第 4 章 基于动态物体分割与映射的 SLAM 实时跟踪算法

本章针对 SLAM 系统在动态物体占比较大或动态物体包含有效信息无法剔除的场景中存在定位与建图精度不准确的问题，在改进 OneFormer 分割网络的基础上，设计了一种融合动态物体分割与映射的 SLAM 实时跟踪算法。该算法首先基于网络分割结果，在潜在动态物体判定阶段，通过相机位姿估计结合物体运动偏转角度分析来准确判别物体运动状态。随后在动态物体映射与跟踪阶段，将识别出的动态物体反向映射至分割掩膜图形成分层结构，建立了动态物体的精确跟踪机制，保留了动态物体上的有效信息，从而显著降低了因场景信息缺失导致的系统误差，有效提升了系统的位姿估计精度。算法框架图如图 4-1 所示。

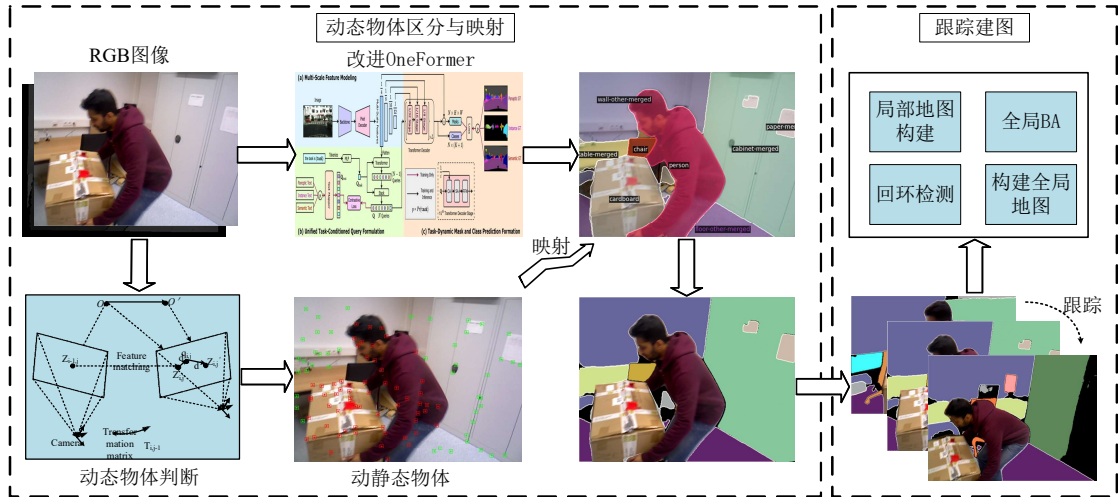


图 4-1 系统总体框架图

4.1 潜在动态物体判断

尽管初步的潜在动态物体分割阶段可以筛选出场景中可能存在的动态元素，但其缺少一个有效的验证机制来确认这些物体是否真的发生了位置变化。为解决这一问题，本节采用了一种结合相机姿态估计与物体运动状态判断的动态识别方法，通过分析相机姿态的变化以及物体在连续帧之间的相对位移，准确评估潜在动态物体的实时运动状态，从而提高动态物体识别的精度。

已知空间中的静态点 p 、相机标定参数和静态点特征深度，当前帧 M_c 是关键帧 M_{c-10} 之后的第 10 帧，可利用参数值和深度信息将图像特征点 p 从关键帧 M_{c-10}

关联到当前帧 M_c ，如式(4-1)所示：

$$\delta_c = \pi(E_{(R,t)}, K, I_{c-10}) \quad (4-1)$$

其中， δ_c 为点 p 投影到当前帧 M_c 中 2D 像素坐标， π 为投影函数， $E_{(R,t)}$ 为相机姿态的相对变换矩阵， R 和 t 分别表示从世界坐标系到相机坐标系的旋转和平移， K 为相机内参数， I_{c-10} 为点 p 在帧 M_{c-10} 上的 3D 投影点。

取 $e_{(R,t)} \in SE(3)$ 为相机姿态的相对变换向量， $SE(3)$ 为特殊欧氏群，定义投影误差函数如式(4-2)所示：

$$E_{(R,t)} \in SE(3) = \left\{ E_{(R,t)} = \begin{pmatrix} R & t \\ 0 & I \end{pmatrix} \mid R \in SO(3) \right\} = \exp(e_{(R,t)}) \quad (4-2)$$

其中， τ_c 为关键帧 M_{c-10} 中 2D 像素坐标 δ_{c-10} 投影到当前帧坐标，联合最小二乘法和高斯-牛顿法优化求解可得到重投影误差项最优解，最小化重投影误差 e_p 如式(4-3)所示：

$$e_p = \tau_c - \pi(E_{(R,t)}, K, I_{c-10}) \quad (4-3)$$

通过整合点特征的误差函数得到一个总的代价函数如式(4-4)所示：

$$\{R, t\} = \operatorname{argmin} \sum_i^A \varepsilon \left(e_p^T(e_{(R,t)}) \mathbf{A}_p e_p(e_{(R,t)}) \right) \quad (4-4)$$

其中， A 为 i 个残差运算所需 3D 点投影至 2D 点数量， ε 为惩罚因子， $\sum_i^A$ 为求和运算， $\mathbf{A}_p$ 为重投影误差的协方差矩阵。

同理，在当前图像帧 M_c 中选定空间点 P 的投影点，利用相对位姿变换矩阵将其坐标系转换至关键帧 M_{c-10} ，由此得到对应的映射点 I'_{c-10} 。随后，沿着视线矢量对基准点 I_{c-10} 与当前点对比并记录视角偏差。为准确评判这一过程，本文在序列帧间建模动态物体的运动，监测潜在动态特征点的视角偏差，步骤如式(4-5)和(4-6)所示：

$$\varphi_c = \omega(\overrightarrow{km_c}, \overrightarrow{\ddot{m}_c}) \quad (4-5)$$

$$\varphi_{c-10} = \omega(\overrightarrow{k\ddot{m}_{c-10}}, \overrightarrow{km_{c-10}}) \quad (4-6)$$

其中， φ_c 和 φ_{c-10} 分别用于表征那些在当前帧 M_c 与参考帧 M_{c-10} 内已成功匹配的特征点的视角偏差程度， ω 为夹角因子系数， m_c 与 $\ddot{m}_c$ 分别标识了特征点及其投影点在世界坐标系统中的实际位置。

最终，将式(4-5)与(4-6)得出的数值求均值同预先定义的阈值 θ 相对照。一旦

超出阈值 θ 的限定，即可判定相应的特征点为动态，相反地，当平均值未达到或恰好等于阈值时，该特征点则被视为静态。

4.2 动态物体映射与跟踪

为了实现动态图像的分层显示结构，对动态物体进行跟踪，本节采用了一种创新的策略，利用平面分割及掩膜构建实现一种层次化的图像表达方式。

4.2.1 动态物体映射

通过对输入 RGB 图像的分析，识别出图像内的不同物体并对其进行分类，系统地提取集合中每个类别图像的颜色值与深度特征纹理等一系列信息，进而输出一组分类标签集合 U ，具体如式(4-7)所示：

$$U = \{U_1^{\text{color}}, U_2^{\text{depth}}, \dots, U_n^{\text{color}}, U_n^{\text{depth}}, \dots\} \quad (4-7)$$

其中， n 表示图像中辨识出的第 n 类物体； U_n^{color} 、 U_n^{depth} 分别表示第 n 类物体的颜色值信息和深度特征纹理信息。

OneFormer 网络能够精准地标记出 RGB 图像中物体类别，并为系统识别到的每一类图像元素在原始 RGB 图像上创建一个特定的掩码 X_n^{mask} 。然后在此基础上利用相机位姿估计与物体运动估计对图像中的动静态掩码进行判断，最终通过精细的映射操作，将动态区域的掩码投射到平面分割图的指定区域。此时，RGB 图像中属于动态部分的所有颜色值信息 U_n^{color} 借助掩码的作用被准确无误地转移到了平面分割图的相应位置，得到图像中的动态物体在掩膜图像上分层映射的结构，其数学表达式如式(4-8)所示：

$$R_i = \left\{ \left\{ X_i^{\text{D-mask}} \mid U_i^{\text{D-color}} \right\} \rightarrow M_i^{\text{D-color}} \right\} \quad (4-8)$$

其中， i 表示区分出的动态区域个数， $X_i^{\text{D-mask}}$ 表示每个动态区域掩码部分， $U_i^{\text{D-color}}$ 表示其对应 RGB 图像区域颜色值信息， $M_i^{\text{D-color}}$ 表示动态掩码区域颜色值信息， R_i 表示最终映射过后的优化区域。

4.2.2 逐帧映射形成跟踪

对于输入的原始图片中连续帧 $N-1$ 、 N 、 $N+1$ ，其包含的动态信息被精确捕

捉，然后单独映射至其对应的掩膜帧上形成动态物体与掩膜的特殊双层呈现效果。输入帧 $N-1$ 、 N 、 $N+1$ 一张张连续切换，其上的动态物体不断运动位置发生变化，映射至掩膜帧上的位置也会快速随之逐帧变动，整个过程连贯后便形成跟踪效果。示意图如图 4-2 所示。

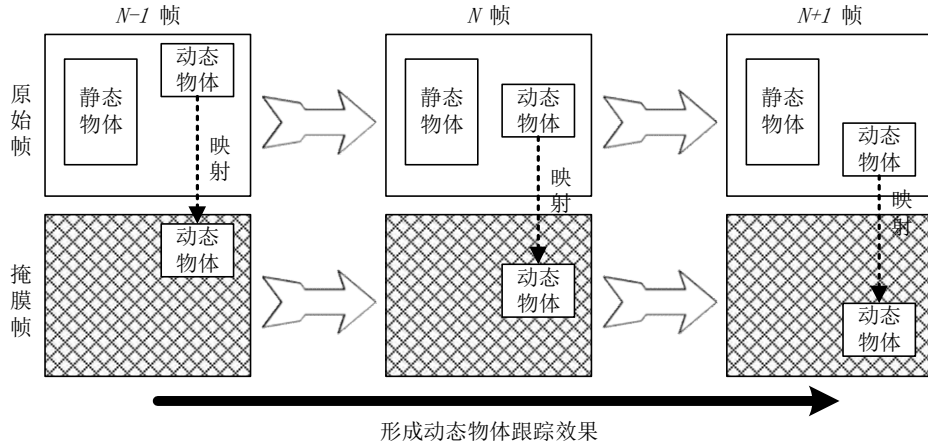


图 4-2 动态物体逐帧映射形成跟踪效果示意图

4.3 实验结果与分析

4.3.1 场景中潜在动态物体识别与判断实验

为系统验证本文提出的动静态特征点区分算法在复杂动态场景下的鲁棒性，基于 TUM 数据集设计了对比实验方案。具体而言，选取了包含多人往复运动的四个典型数据集 fr3_w_xyz、fr3_w_halfs、fr3_w_xyz、fr3_w_static，这些序列在动态目标运动轨迹复杂度方面具有明显差异。在实验设计中，采用采样策略从每个序列中随机抽取具有运动状态代表性的关键帧进行分析。为建立科学对比基准，实验同步比对了经典 SLAM 框架、DynaSLAM 的动态检测模块与本章算法，算法在两位实验者沿非线性轨迹绕桌高速移动的典型动态场景中动态点区分效果图如图 4-3 所示。

如图 4-3(e)-(h)与(i)-(l)所示，DynaSLAM 采用的基于几何约束的检测方法在高速运动区域出现显著漏检现象，其判定结果呈现明显的运动模糊效应。反观本文算法，不仅完整捕获了人体四肢末端的快速运动特征并标记为红色特征点，同时精确保持了桌面与背景等静态特征的稳定识别并标记为绿色特征点。

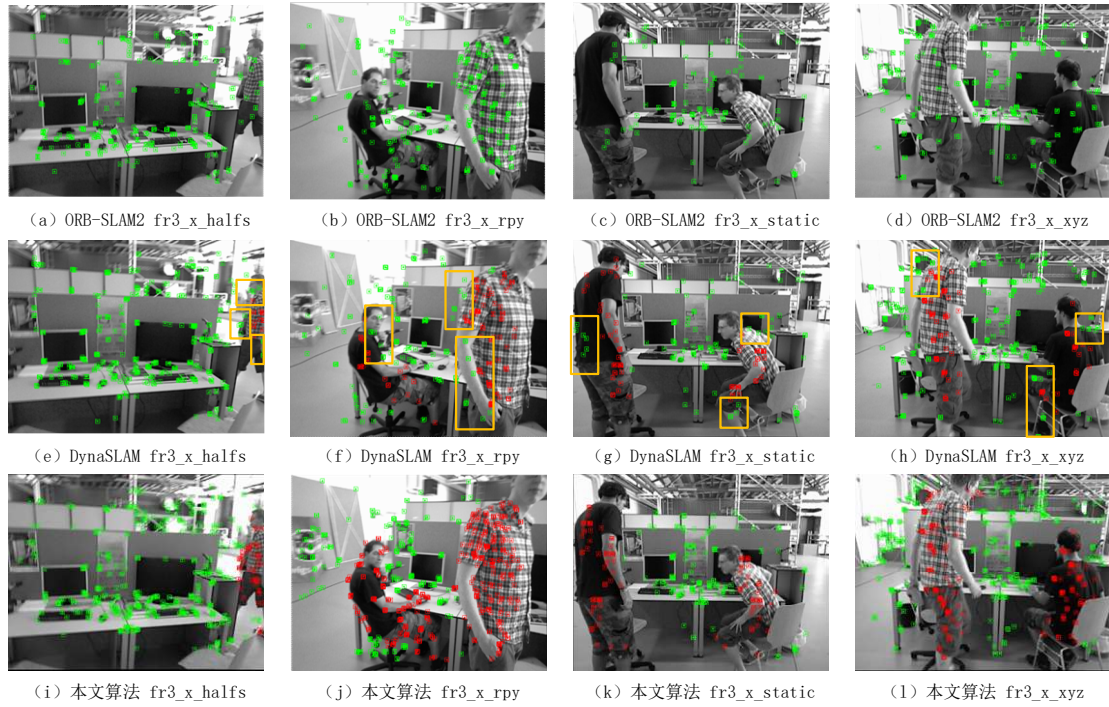


图 4-3 动态物体判断实验效果图

4.3.2 动态物体判别及反向映射实验

为系统验证将动态区域映射回平面分割图上以形成场景中逐帧特异化呈现的跟踪效果算法的有效性，本节构建了分级实验验证体系，选取 TUM 数据集 fr_walking_xyz 序列进行动态物体映射处理能力测试，该序列包含行人运动导致动态/静态区域交替变化的典型场景，能有效检验算法的环境适应性，实验结果如图 4-4 与 4-5 所示。



图 4-4 动态物体映射效果图



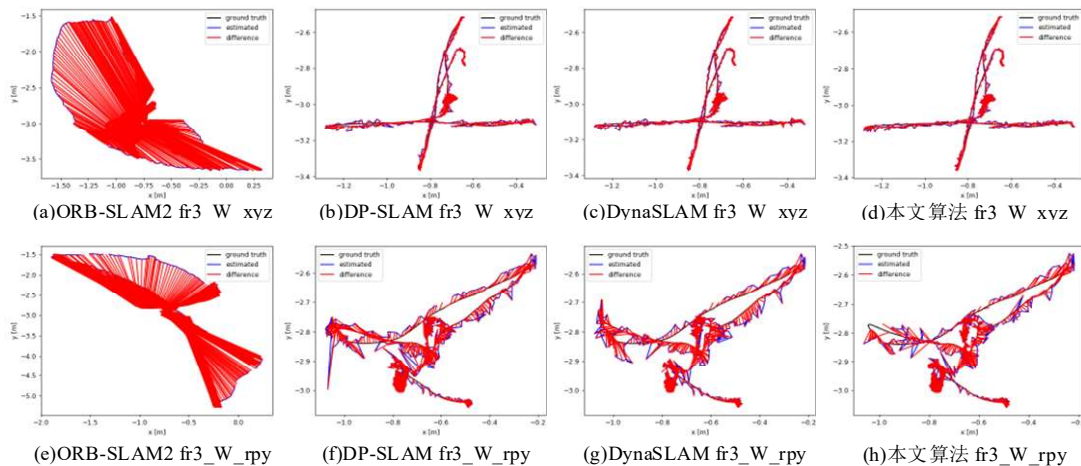
图 4-5 动态物体映射形成跟踪过程效果图

图 4-4 实验选取了动态物体占比呈梯度递减趋势的 5 帧数据帧以验证算法的性能。实验结果表明，在动态区域占比从高到低逐渐变化的过程中，映射算法始终能够保持稳定的跟踪精度，充分证明了该算法在面对动态目标变化时的鲁棒性。图 4-5 展示了本文算法结合改进后的 **OneFormer** 网络，对场景中动态物体进行逐帧追踪与映射的具体实验过程及效果，进一步验证了算法的有效性与可靠性。

上述实验结果证明，本章提出的动态区域映射算法在动态场景下表现出良好性能，能有效实现不同动态占比环境中的精确区域映射与稳定 SLAM 跟踪机制。

4.3.3 SLAM 系统中算法运行评估实验

为了验证本文算法在动态场景下的有效性，采用 TUM 数据集下四个动态序列 **fr3_W_xyz**、**fr3_W_rpy**、**fr3_W_static** 与 **fr3_W_halfs** 对 **ORB-SLAM2**、**DP-SLAM**、**DynaSLAM** 以及本文算法在动态场景中运行效果进行对比与评估。算法轨迹对比结果图如图 4-6 所示。



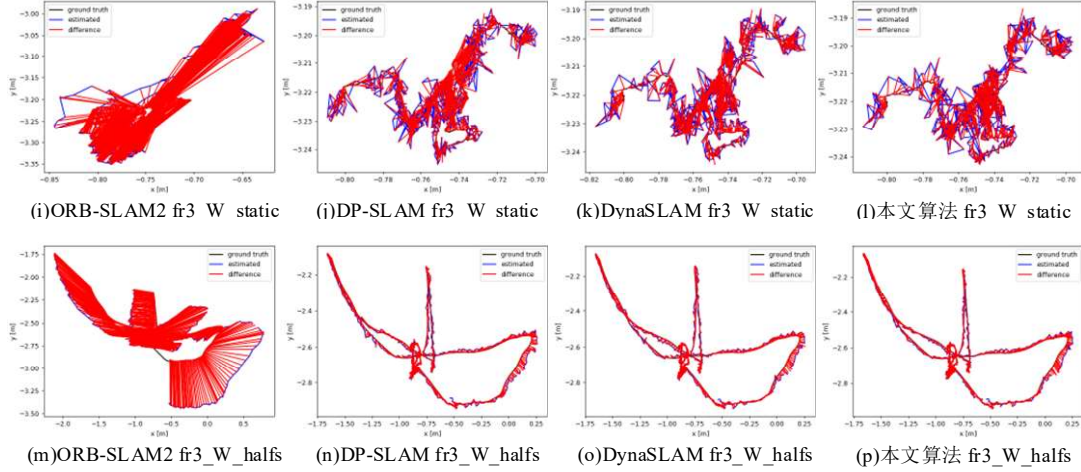


图 4-6 TUM 数据集下 4 种不同算法轨迹图

场景中存在的动态物体会严重影响相机位姿的求解，从而产生较大的轨迹漂移，DP-SLAM 方法通过基于移动概率传播的模型检测动态点，并结合几何约束与语义分割信息进行动态点跟踪。虽然理论上这种方法能有效监测动态点的变化，但实际应用中效果较差。相比之下，DynaSLAM 方法通过明确区分动态与静态特征来估计相机位姿，并在优化过程中施加约束，但这可能因忽视动态特征而导致系统丢失关键信息，从而影响整体的精度。本文利用改进的 **Oneformer** 网络来准确区分动态与静态特征，保留动态特征进行映射，静态特征进行相机位姿估计，动态特征用于物体跟踪，使得动态物体可以从背景中分离并在连续帧间进行动态跟踪，提升了 SLAM 系统的整体性能，有效减轻了动态物体对位姿估计的影响，并为实现更加准确的轨迹定位与地图构建提供了强有力的支持。

为了进一步分析算法的性能，更直观地展示算法的性能表现，采用 SLAM 算法两个核心评估指标：相对位姿误差（Relative Pose Error, RPE）与绝对轨迹误差（Absolute Trajectory Error, ATE）。其中，RPE 用于评估里程计在短距离内的精度，而 ATE 则用于衡量算法在整个轨迹上的全局一致性与准确性。实验结果的具体数值，包括均方根误差（Root Mean Square Error, RMSE）与标准差（Standard Deviation, SD），实验数据分别如表 4-1 与表 4-2 所示。

在 TUM 四个序列上具体分析，本文算法相对于 ORB-SLAM2、DP-SLAM 以及 DynaSLAM 相对位姿误差 RPE 的 RMSE 值与 SD 值分别平均减少 80.88%、25.32%、9.69%与 80.86%、31.40%、14.93%；绝对轨迹误差 ATE 的 RMSE 值与 SD 值分别平均减少 97.58%、17.70%、19.63%与 97.98%、24.56%、18.01%。实验结果证明，在动态环境下本文算法生成的位姿估计与绝对轨迹误差值更小，定

位更加准确。因此可以看出，针对场景中存在动态物体影响相机位姿的求解的问题，本文算法表现良好。

表 4-1 TUM 数据集下 4 种不同算法相对位姿误差对比结果（单位：m）

Seq	ORB-SLAM2		DP-SLAM ^[57]		DynaSLAM ^[58]		本文算法	
	RMSE	SD	RMSE	SD	RMSE	SD	RMSE	SD
fr_w_xyz	0.7533	0.4691	0.1090	0.0644	0.0782	0.0457	0.0516	0.0319
fr_W_rpy	0.9861	0.5222	0.2893	0.1184	0.2413	0.1156	0.2311	0.1122
fr_W_static	0.4104	0.2684	0.0670	0.0497	0.0566	0.0313	0.0441	0.0211
fr_W_half	0.2872	0.0733	0.1588	0.1394	0.1397	0.1072	0.1391	0.0899
average	0.6093	0.3333	0.1560	0.0930	0.1290	0.0750	0.1165	0.0638

表 4-2 TUM 数据集下 4 种不同算法绝对轨迹误差对比结果（单位：m）

Seq	ORB-SLAM2		DP-SLAM ^[57]		DynaSLAM ^[58]		本文算法	
	RMSE	SD	RMSE	SD	RMSE	SD	RMSE	SD
fr3_W_xyz	0.7963	0.4411	0.0147	0.0073	0.0175	0.0081	0.0146	0.0073
fr3_W_rpy	0.7552	0.4103	0.0356	0.0218	0.0368	0.0186	0.0244	0.0123
fr3_W_static	0.5892	0.4305	0.0079	0.0037	0.0069	0.0031	0.0058	0.0029
fr3_W_half	0.7041	0.4188	0.0254	0.0129	0.0245	0.0123	0.0238	0.0119
average	0.7112	0.4252	0.0209	0.0114	0.0214	0.0105	0.0172	0.0086

4.3.4 真实场景中算法运行评估实验

本章在 Husky 机器人上进行实验，Husky 机器人平台参数如表 4-3 所示。本章的实验是通过视觉传感器来采集外界数据，本章采用 Intel RealSenseD435i 深度摄像机作为视觉传感器，D435i 深度摄像机如图 4-7 所示。实验中需要处理动态物体以及语义分割等相关方面的数据，故使用分段主从机模式来进行测试实验，其中主机为个人电脑，从机为加在 Husky 机器人平台上的外接可充电显示器，外接可充电显示器配置参数如表 4-4 所示，改装的 Husky 机器人如图 4-8 所示。

表 4-3 Husky 机器人平台参数

参数名称	参数信息
重量	50 千克（110 磅）
最大有效负载	75 千克（165 磅）
全地形有效负载	20 千克（44 磅）
最大速度	1 米/秒 （3.3 英尺/秒）
工作时间	正常运行 3 小时，待机 8 小时

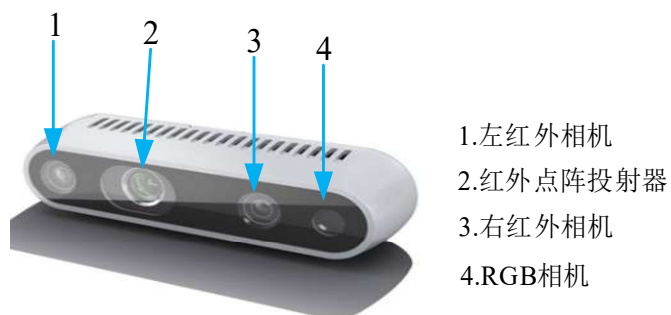


图 4-7 D435i 深度摄像机

表 4-4 可充电显示器配置

硬件名称	参数信息
CPU	Intel Core i5-6300HQ
GPU	NVIDIA GeForce GTX 960
RAM	32G
操作系统	Ubuntu 18.04

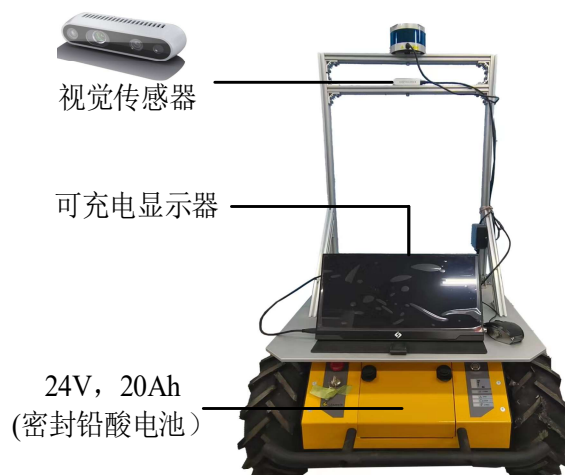


图 4-8 Husky 机器人

室内外真实场景布局如图 4-9(a)(b)所示。室内场景中，机器人从起点 A 沿着 $A \rightarrow B \rightarrow C \rightarrow D \rightarrow E \rightarrow B \rightarrow C \rightarrow F \rightarrow A$ 的路线绕工作台 2 与工作台 3 进行“8”字运动，行人 1 在工作台 1 处拿起背包沿 ab 作往返运动，行人 2 在工作台 2 附近推动椅子沿 cd 作往返运动；室外场景中，机器人沿 $A \rightarrow B \rightarrow C \rightarrow D \rightarrow E$ 的路线在在校园内运动，中间陆续随机有行人与车辆经过。

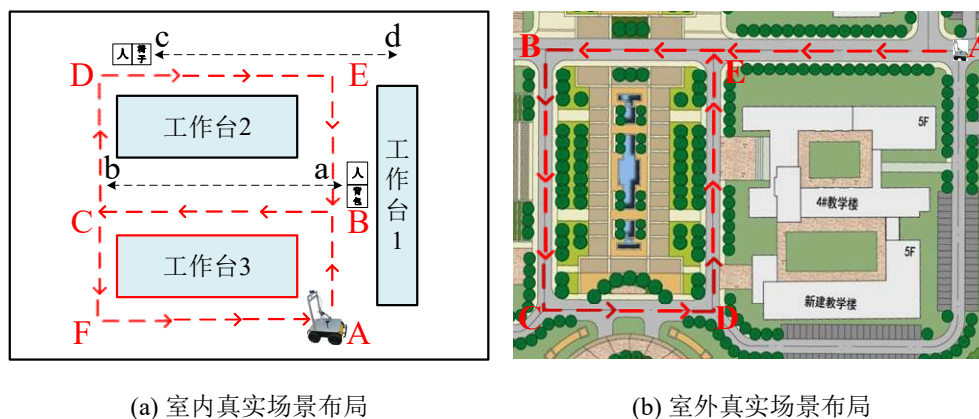


图 4-9 实验平台真实场景

在室内外的真实场景中，分别选择在运行过程中出现动态遮挡与动态模糊的两帧动态帧，使用本章所提算法进行动静态判断，判断效果分别如图 4-10 与图 4-11 中(a)(b)所示，其中绿色为静态特征点，红色为动态特征点。此外，本章动态物体映射算法在室内外真实数据场景中的映射与追踪效果如图 4-10 与图 4-11 的(c)~(f)所示。图中的(c)、(d)、(e)、(f)展示了真实场景中动态物体逐渐变化的 4 帧图像，实验结果表明，算法在动态物体的映射与追踪方面表现出色，能够稳定地捕捉并跟踪目标物体的位置变化，验证了算法在复杂动态场景下的准确性。

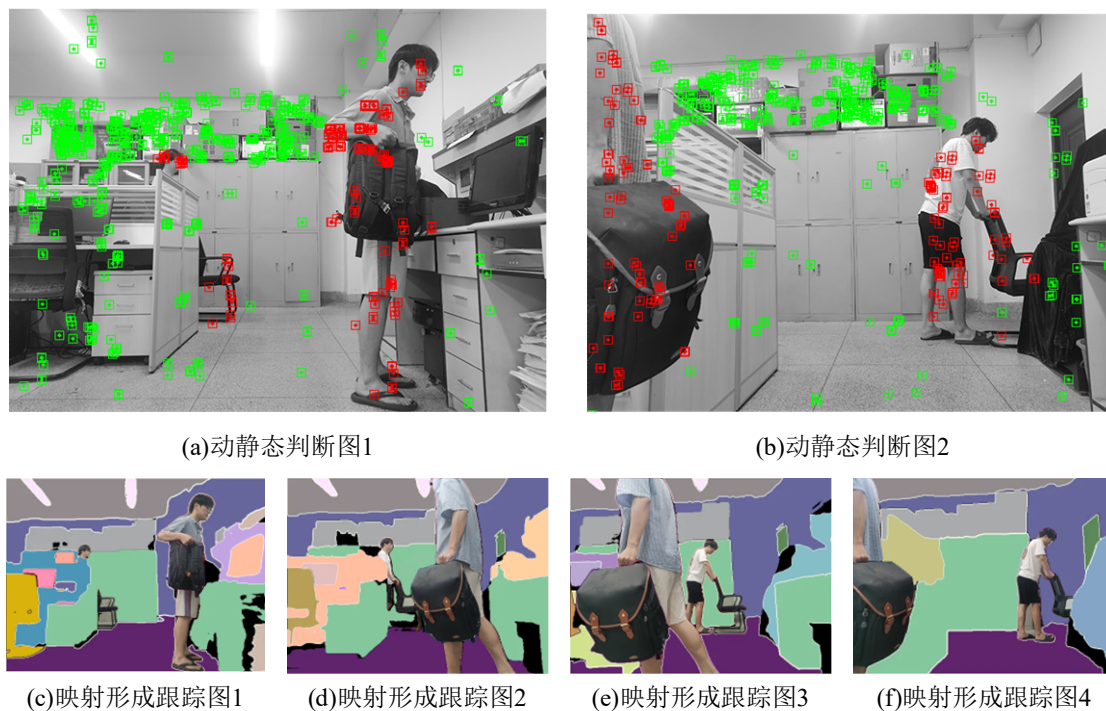


图 4-10 室内真实场景实验效果图



图 4-11 室外真实场景实验效果图

图 4-12 展示了本章 SLAM 算法与 ORB-SLAM2 算法在室内外真实场景中的定位轨迹对比。在复杂动态场景下，本章算法生成的轨迹更接近真实路径，表现出显著的优势与更高的精度。

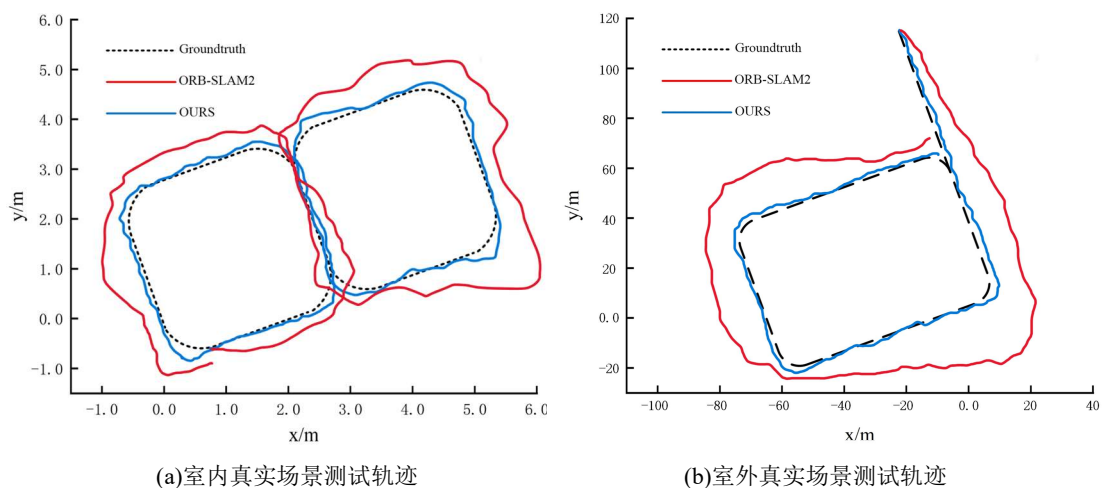


图 4-12 真实场景下算法轨迹图对比

综上所述，本章算法通过引入改进后的 OneFormer 分割网络，对动态物体进行分割、区分、映射与跟踪，在动态环境中展现出较强的鲁棒性与稳定性，有效提升了定位与跟踪的准确性。

4.4 本章小结

本章针对 SLAM 在动态物体占比较大与动态物体包含有利信息不能剔除的场景中定位与建图不准确等问题，在改进 OneFormer 分割网络的基础上设计了一种基于动态物体分割与映射的 SLAM 实时跟踪算法。算法结合网络分割获取到的相关信息，在潜在动态物体判断阶段，使用相机位姿与物体运动偏转角度估计物体运动状态进而区分动态物体。在动态物体映射与跟踪阶段，将识别得到的动态物体映射回网络分割掩膜图上形成分层结构，通过对动态物体的精准显示与追踪，保留了动态物体上的有效信息，降低了因信息缺失而导致的系统误差，提高了系统位姿估计准确性。在公开数据集中对算法进行测试，实验数据显示，所提算法的绝对轨迹误差均方根值较 DP-SLAM、DynaSLAM 分别降低 17.70%与 19.63%，展现出良好的位姿估计精度与建图性能。

第 5 章 全文总结展望

5.1 全文总结

本文从动态场景下视觉 SLAM 研究现状出发, 以实现复杂动态场景下移动机器人准确性定位与高精度建图为目标, 深入分析复杂动态场景移动机器人面临的挑战, 通过动态物体分割网络优化以及动态物体映射与跟踪线程共同构建了一种面向复杂动态场景的视觉 SLAM 算法。该算法有效改善了移动机器人在动态场景下因动态物体干扰与信息丢失导致的跟踪失效、轨迹漂移与建图失真等问题, 提高了视觉 SLAM 系统的定位与建图能力。

论文主要研究内容与创新点概括如下:

(1) 针对传统视觉 SLAM 在动态场景下动态目标分割精度不足的问题, 设计了一种基于改进 OneFormer 网络的动态物体分割算法。该算法以 OneFormer 通用图像分割网络为基础框架, 在低层级信号特征增强模块中, 采用轻量级深度可分离卷积替代标准卷积, 结合特征通道重组, 有效增强了网络对浅层特征的表达能力。在双域协同注意力增强模块中, 构建了通道-空间双域注意力机制, 通过并行处理特征图的通道与空间维度信息, 实现了对动态区域特征的自适应增强。实验结果表明, 所提算法的像素准确率与平均交并比较改进前分别提升了 4.4% 与 5.1%, 显著提高了对动态物体的分割精度。

(2) 针对 SLAM 在动态物体占比较大与动态物体包含有利信息不能剔除的场景中定位与建图不准确等问题, 设计了一种基于动态物体分割与映射的 SLAM 实时跟踪算法。该算法利用改进的分割网络获取物体信息, 在潜在动态物体判断阶段, 通过相机位姿与物体运动偏转角度联合估计运动状态, 实现动态物体的准确区分。在动态物体映射与跟踪阶段, 通过动态物体反向映射, 构建分层掩膜结构, 保留其有效信息, 实现对动态物体的精准追踪。实验结果表明, 所提算法绝对轨迹误差均方根较 DP-SLAM 与 DynaSLAM 分别降低 17.70% 与 19.63%。在搭建的移动机器人硬件平台上进行的室内外真实场景测试表明, 本算法能有效适应复杂动态环境, 为解决动态场景下的视觉 SLAM 问题提供了可靠的参考依据。

5.2 展望

本文通过公开数据集与搭建硬件平台在真实场景中进行验证,证明了所提出算法的可行性,并为视觉 SLAM 算法的进一步发展提供了新的思路。然而,在研究过程中也发现了一些亟待解决的问题:

(1) 相似结构化场景中的定位问题。在相似结构化场景中运行时本文算法难以进行大量有效特征信息的提取,帧与帧之间的位置关系不清晰,这样易造成定位失败。因此,下一步需要在本文构建的算法基础上提高相似结构化场景下定位能力。

(2) 深度学习框架下语义 SLAM 的运行效率。本文所提算法采用语义分割网络对图像数据进行处理,资源占用较大,视觉 SLAM 算法在环境中响应时间较长。因此,如何提高深度学习框架下语义 SLAM 的运行效率是下一步必要解决的问题。

针对以上挑战,未来的研究将致力于进一步优化算法性能与计算效率,以实现更高效、更准确的视觉 SLAM 系统,尤其是在复杂环境与实时应用场景中。

参考文献

- [1] 林韩熙,向丹,欧阳剑,等.移动机器人路径规划算法的研究综述[J].计算机工程与应用,2021,57(18):38-48.
- [2] Cadena C, Carlone L, Carrillo H, et al. Past, present, and future of simultaneous localization and mapping: Toward the robust-perception age[J]. IEEE Transactions on Robotics, 2016,32(6):1309-1332.
- [3] Khah M H ,Nokhodberiz S N ,Montazeri A . Simultaneous localization and mapping in a multi-robot system in a dynamic environment with unknown initial correspondence[J]. Frontiers in Robotics and AI, 2024,10(29):1672-1679.
- [4] Shan T, Englot B. Lego-loam: Lightweight and ground-optimized lidar odometry and mapping on variable terrain[C].// 2018 IEEE International Conference on Intelligent Robots and Systems,2018:4758-4765.
- [5] Henein M, Zhang J, Mahony R, et al. Dynamic SLAM: The need for speed[C].// 2020 IEEE International Conference on Robotics and Automation (ICRA). IEEE, 2020:2123-2129.
- [6] 袁国帅,齐咏生,刘利强,等.一种基于因子图消元优化的激光雷达视觉惯性融合 SLAM 方法[J].电子学报,2023,51(11):3042-3052.
- [7] 汤琴琴,王立喜,侍经纬,等.复杂环境下无人机视觉惯性里程计设计[J].中国惯性技术学报,2025,33(2):140-146.
- [8] Smith C, Cheeseman P. On the representation and estimation of spatial uncertainty[J]. The International Journal of Robotics Research,2017,5(4):56-68.
- [9] 李天成,范红旗,孙树栋.粒子滤波理论、方法及其在多目标跟踪中的应用[J].自动化学报, 2015, 41(12): 1981-2002.
- [10] Li Y, Brasch N, Wang Y, et al. Structure-slam: Low-drift monocular slam in indoor environments[J]. IEEE Robotics and Automation Letters, 2020,5(4):6583-6590.
- [11] 赵功洋,闫利.移动机器人 SLAM 位姿估计的改进四元数无迹卡尔曼滤波[J].测绘学报,2022,51(2):212-223.
- [12] 王旒军,陈家斌,余欢等.RGB-D SLAM 综述[J].导航定位与授时,2017,4(6):9-18.
- [13] Pu H, Luo J, Wang G, et al. Visual SLAM integration with semantic segmentation

- and deep learning: A review[J]. IEEE Sensors Journal,2023,23(19): 2119-2138.
- [14]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, 2017, 29(6): 1052-1067.
- [15]Klein G, Murray D. Parallel tracking and mapping for small AR workspaces[C].// 2017 6th IEEE and ACM International Symposium on Mixed and Augmented Reality. IEEE, 2017: 225-234.
- [16]Engel J,Schöps T,Cremers D.LSD-SLAM: Large-scale direct monocular SLAM[C].// European Conference on Computer Vision. Cham: Springer International Publishing, 2014: 834-849.
- [17]Forster C,Pizzoli M,Scaramuzza D.SVO: Fast semi-direct monocular visual odometry[C].// 2014 IEEE International Conference on Robotics and Automation. IEEE, 2014: 15-22.
- [18]Labbe M, Michaud F. Memory management for real-time appearance-based loop closure detection[C].//2011 IEEE International Conference on Robotics and Systems. IEEE, 2011: 1271-1276.
- [19]Barath D, Cavalli L, Pollefeys M. Learning to find good models in Ransac[C].// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2022: 1744-1753.
- [20]Mur-Artal R, Tardós J D. Orb-slam2: An open-source slam system for monocular, stereo, and rgb-d cameras[J]. IEEE Transactions on Robotics, 2017,33(5):1255-1262.
- [21]Qin T, Li P, Shen S. Vins-mono: A robust and versatile monocular visual-inertial state estimator[J]. IEEE Transactions on Robotics, 2018, 34(4):1004-1020.
- [22]Li Q, Ibrahim H, Chin K, et al. FADM-SLAM: a fast and accurate dynamic intelligent motion SLAM for autonomous robot exploration involving movable objects[J]. Robotic Intelligence and Automation, 2023, 43(3):254-266.
- [23]Xing Z, Zhu X, Dong D. DE-SLAM: SLAM for highly dynamic environment[J]. Journal of Field Robotics, 2022,39(5):528-542.
- [24]Sun Y, Liu M, Meng M. Motion removal for reliable RGB-D SLAM in dynamic

- p>environments[J]. Robotics and Autonomous Systems, 2018,18(21):115-128.
- [25]Li S, Le D. RGB-D SLAM in dynamic environments using static point weighting[J]. IEEE Robotics and Automation Letters, 2017, 2(4): 2263-2270.
- [26]艾青林,刘刚江,徐巧宁.动态环境下基于改进几何与运动约束的机器人 RGBD SLAM 算法[J]. 机器人, 2021,43(2):167-176.
- [27]Liu Y, Miura J. RDS-SLAM: Real-time dynamic SLAM using semantic segmentation methods[J]. IEEE Transactions on Robotics, 2021, 9(18): 2772-2785.
- [28]Soares J, Gattass M, Meggiolaro M. Crowd-SLAM: visual SLAM towards crowded environments using object detection[J]. Journal of Intelligent & Robotic Systems, 2021, 72(2):50-56.
- [29]Fan Y, Zhang Q, Tang Y, et al. Blitz-SLAM: A semantic SLAM in dynamic environments[J]. Pattern Recognition, 2022, 121: 108225.
- [30]Dvornik N, Shmelkov K, Mairal J, et al. Blitznet: A real-time deep network for scene understanding[A]. Proceedings of the IEEE International Conference on Computer Vision[C]. Venice: IEEE, 2017: 4154-4162.
- [31]Li M, He J, Jiang G, et al. DDN-slam: Real-time dense dynamic neural implicit slam with joint semantic encoding[J]. IEEE Robotics and Automation Letters, 2024,5(10):4300-4307.
- [32]Mildenhall B, Srinivasan P P, Tancik M, et al. Nerf: Representing scenes as neural radiance fields for view synthesis[J]. Communications of the ACM, 2021, 65(1): 99-106.
- [33]Xiao L, Wang J, Qiu X, et al. Dynamic-SLAM: Semantic monocular visual localization and mapping based on deep learning in dynamic environment[J]. Robotics and Autonomous Systems, 2019, 17(10):1-16.
- [34] Wu W, Guo L, Gao H, et al. YOLO-SLAM: A semantic SLAM system towards dynamic environment with geometric constraint[J]. Neural Computing and Applications, 2022, 34(22): 2011-2026.
- [35] Yu C, Liu Z, Liu X, et al. DS-SLAM: A semantic visual SLAM towards dynamic environments[C].// 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Madrid, Spain: IEEE, 2018: 1168-1174.

- [36] Bescos B, Fácil J, Civera J, et al. DynaSLAM: Tracking, mapping, and inpainting in dynamic scenes[J]. IEEE Robotics and Automation Letters, 2018, 3(4): 4076-4083.
- [37] 王柯赛,姚锡凡,黄宇,等. 动态环境下的视觉 SLAM 研究评述[J]. 机器人, 2021, 43(6): 715-732.
- [38] 高兴波,史旭华,葛群峰,等.面向动态物体场景的视觉 SLAM 综述[J]. 机器人, 2021, 43(6): 733-750.
- [39] Bansal M, Kumar M. 2D object recognition: a comparative analysis of SIFT, SURF and ORB feature descriptors[J]. Multimedia Tools and Applications, 2021, 80(12): 1839-1857.
- [40] Rublee E, Rabaud V, Konolige K, et al. ORB: an efficient alternative to SIFT or SURF[C].// IEEE International Conference on Computer Vision. Barcelona: IEEE Press, 2011: 2564-2571.
- [41] Lowe G. Object recognition from local scale-invariant features[C].// IEEE International Conference on Computer Vision. Kerkyra: IEEE Press, 2019:1150-1157.
- [42] Bay H, Tuytelaars T, Gool L J V. SURF: speeded up robust features[C].// European Conference on Computer Vision. Graz: Springer Press, 2016: 404-417.
- [43] Zhou L, Wang S, Kaess M. DPLVO: Direct point-line monocular visual odometry[J]. IEEE Robotics and Automation Letters, 2021, 6(4): 2113-2120.
- [44] Singh G, Wu M, Lam S K. Fusing semantics and motion state detection for robust visual SLAM[C].// Proceedings of the IEEE Winter Conference on Applications of Computer Vision. 2020: 2764-2773.
- [45] Li S, Ni P. Square-root unscented kalman filter based simultaneous localization and mapping[C].// 2010 IEEE International Conference on Information and Automation, 2010: 2381-2388.
- [46] Sim R, Elinas P, Little J J. A study of the rao-blackwellised particle filter for efficient and accurate vision-based SLAM[J]. International Journal of Computer Vision, 2017, 44(3): 303-318.
- [47] Chen S, Zhou B, Jiang C, et al. A lidar/visual slam backend with loop closure

- p>detection and graph optimization[J]. Remote Sensing, 2021, 13(14): 2720-2729.
- [48]Zhang G, Zhang T, Zhang C. Accurate real-time SLAM based on two-step registration and multimodal loop detection[J]. Measurement Science and Technology, 2022, 34(2): 22-29.
- [49]陈孟元,丁陵梅,张玉坤.基于改进关键帧选取策略的快速 PL-SLAM 算法[J]. 电子学报, 2022,50(3):608-618.
- [50]Macario Barros A, Michel M, Moline Y, et al. A comprehensive survey of visual slam algorithms[J]. Robotics, 2022, 11(1): 24-31.
- [51]Xue W, Ying R, Gong Z, et al. SLAM based topological mapping and navigation[C].// 2020 IEEE/ION Position, Location and Navigation Symposium (PLANS). IEEE, 2020: 1336-1341.
- [52]方宝富,张旭,王浩.联合深度图聚类与目标检测的像素级分割算法[J].模式识别与人工智能,2022,35(2):130-140.
- [53]陈孟元,韩朋朋,刘金辉,等.动态遮挡场景下基于改进 Transformer 实例分割的 VSLAM 算法[J]. 电子学报, 2023, 51(7): 1812-1825.
- [54]J. Jain, J. Li, M. Chiu, A. Hassani, N. Orlov and H. Shi, OneFormer: One Transformer to Rule Universal Image Segmentation[C].// 2023 IEEE Conference on Computer Vision and Pattern Recognition , Vancouver, BC, Canada, 2023: 2989-2998.
- [55]苏铁明,李鹏博,徐志祥,等.基于改进 YOLACT 的堆垛图像快速分割方法研究 [J].计算机测量与控制,2023,31(12):210-215.
- [56]李士杰,田秀霞.基于 Mask-RCNN 的图像篡改检测模型 [J].计算机仿真,2024,41(7):227-232.
- [57]何武,路巍,汪瑶,等.基于 Player 的室内服务机器人的地图构建与定位系统[J]. 仪表技术,2019,11(5):56-58.
- [58]房立金,王科棋.一种结合深度学习的运动重检测视觉 SLAM 算法[J].计算机工程,2022,48(5):18-26.

攻读学位期间发表的学术论文与科研成果

一、发表学术论文

- 【1】 陈孟元,杨苏朋,许瑞珩,等.改进 OneFormer 分割网络的动态物体实时跟踪 SLAM 算法[J].中国惯性技术学报,2025,33(03):257-266.
- 【2】 S. Yang, A. Xu, P. Li, M. Chen, P. Du and K. Shao, "Visual SLAM Algorithm Based on YOLOv5 in Dynamic Scenario," *2023 China Automation Congress (CAC)*, Chongqing, China, 2023, pp. 2640-2645

二、参与的科技项目

- 【3】 安徽省重点研究与开发计划项目（高新领域），202304a05020073，面向低纹理场景的增强感知视觉移动机器人关键技术与示范应用，2023/08-2026/06，100.0 万元。（皖科高秘〔2023〕250 号）
- 【4】 安徽省高校杰出青年科研项目，2022AH020065，基于视觉-惯性的移动机器人复杂环境仿生 SLAM 算法研究，2022/09-2025/08，100.0 万元。（皖教秘科〔2022〕241 号）
- 【5】 安徽省高校协同创新项目，GXXT-2021-050，复杂环境下移动机器人位姿自适应优化算法研究及应用，2021/01-2023/12，40.0 万元。（皖教秘科〔2021〕58 号）

三、科技竞赛

- 【6】 “游刃有余-低纹理场景下基于增强感知视觉机器人开发”——第十八届“挑战杯”全国大学生课外学术科技作品竞赛——国家级三等奖
- 【7】 “增强感知视觉机器人-为中国智造插上翅膀（EPV-ROBOT）”——第九届安徽省“互联网+”大学生创新创业大赛总决赛高校主赛道——银奖
- 【8】 “游刃有余-低纹理场景下基于增强感知视觉机器人开发”——第十届“挑战杯·华安证券”课外学术科技作品竞赛安徽省一等奖

致 谢

当光标停留在致谢页的这一刻，突然意识到三年的研究生生涯已临近终点。2022 年秋日初见时的忐忑犹在昨日，2025 年盛夏的告别却已近在眼前。安徽工程大学的每个角落——从晨光中的图书馆到深夜依然明亮的实验室，都记录着这段不可复制的成长时光。

衷心感谢我的导师陈老师。从选题方向的确立、论文框架的构建，到最终成稿的反复打磨，您严谨的学术态度和耐心的指导让我受益匪浅。同时感谢实验室的同门伙伴，每次的组会研讨、做实验时的相互支持，都让这段学术之旅充满温度。

感谢我的室友和同学们。三年间我们共同经历的无数个学习到深夜的日子，在宿舍讨论课题的周末，压力大时彼此的鼓励，这些点滴构成了研究生生活最温暖的底色。特别感谢你们在我实验遇到瓶颈时的陪伴，让我在异乡求学时始终感受到集体的力量。

最深沉的谢意献给我的父母和姐姐。二十多年来，你们始终尊重我的每个选择，在我攻读硕士期间给予无条件支持。每次视频时你们藏不住的牵挂，寒暑假回家时桌上永远温热的那碗汤，都是支撑我坚持至今的精神支柱。你们教会我：真正的成长，是带着爱意继续前行。

此外，我要感谢一个特殊的人，虽然现在的我们已经因为对未来生活的不同追求而各自奔赴不同的人生轨迹，但那些相互依偎取暖的时光会永远在记忆里熠熠生辉。感谢你让我看见女性世界的辽阔与困境，教会我爱的能力，也让我明白了自己在感情中的局限。愿我们在各自的远方，都能遇见更好的自己。我会在以后每个阳光明媚的日子里，当阳光穿过树叶的间隙时想起，此时这世上一定会有个女孩正对着这金灿灿的世界肆意微笑。

最后，想对这三年里始终没有放弃的自己说声感谢。那些为修改论文辗转难眠的夜晚，在实验室反复验证数据的坚持，学术汇报前反复练习的执着，终将成为人生路上最坚实的基石。未来，我将带着在安徽工程大学收获的知识与勇气，继续书写新的篇章。