分类号：TP242.6　　学校代码：10363
密　级：公开　　学　　号：2220320127

安徽工程大学
Anhui Polytechnic University

硕士学位论文

题目　基于词袋模型的视觉SLAM回环检测改进算法研究

论 文 作 者　徐奥
校 内 导 师　孙新柱　副教授
校 外 导 师　于晓东　高级工程师
专业学位类别　电子信息
研究方向（领域）　人工智能与系统集成

论文提交日期：2025 年 5 月 15 日

分类号：TP242.6　　单位代码：10363
密　级：公 开　　学　　号：2220320127

题　目　基于词袋模型的视觉 SLAM 回环检测改进算法研究

题　目　Research on Improved Algorithm for Visual SLAM Loop Closure Detection Based on Bag-of-Words Model

学生姓名： 徐奥

校内导师： 孙新柱（副教授）

校外导师： 于晓东（高级工程师）

专业学位类别： 电子信息

研究方向（领域）： 人工智能与系统集成

论文答辩日期：2025 年 05 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：

日期：2025年 5月 14日

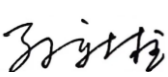
安徽工程大学学位论文版权使用授权书

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：  　　指导教师签名： 

日期：2025年 5月 14日　　日期：2025年 5月 14日

基于词袋模型的视觉 SLAM 回环检测改进算法研究

摘　要

同步定位与地图构建（Simultaneous Localization and Mapping, SLAM）指的是移动机器人通过搭载的传感器实时获取环境信息，在未知环境中实现自身定位和环境地图的构建。根据传感器的不同，主要分为激光 SLAM 和视觉 SLAM 两类；相较于激光 SLAM，视觉 SLAM 因其低成本、传感器融合能力强以及广泛的场景适用性，逐渐成为 SLAM 研究领域的主流方向。在视觉 SLAM 中，回环检测是实现高精度定位与地图优化的关键技术，通过识别相机是否重新进入先前访问过的区域，建立当前帧与历史帧之间的约束关系，修正由累计误差所导致的定位漂移。然而，随着场景和环境的多样性增加，基于传统词袋模型的回环检测方法，缺乏对环境变化或新场景出现时视觉词典的动态更新能力；同时，忽略了环境中物体间的空间结构关系，导致回环检测召回率低和 SLAM 系统定位误差较大。因此，如何增强词袋模型的动态适应能力，并提升回环检测的召回率成为本文亟待解决的问题。

针对传统视觉 SLAM 算法在回环检测环节中词袋模型算法生成的视觉单词对场景代表不强，以及无法根据环境变化进行视觉词典动态更新的问题，采用基于 ConvNeXt 孪生网络的动态视觉词典更新算法。首先，选用 ConvNeXt 作为骨干网络，构建具有共享权重的孪生网络架构，通过动态特征融合模块整合双路网络不同视角下提取的特

征；其次，通过主成分分析算法对特征进行降维处理，并使用分层密度聚类算法去除数据中的噪声，构建视觉词典；最后，在离线视觉词典的基础上，结合层级树检索机制和节点更新策略，使得词典能够针对环境变化进行动态更新。

针对传统视觉 SLAM 算法在回环检测过程中忽略空间物体结构关系的问题，采用基于结构相似性评估的回环检测算法。首先，在回环检测过程中，通过结构相似性评估算法计算图像的全局结构相似性评分，以捕捉图像之间的整体结构信息；其次，利用动态权重机制，融合图像的局部特征评分与全局结构相似性评分，从而获得更为准确的回环候选帧；最后，设计了回环候选帧剔除机制，通过历史回环结果剔除错误的回环候选帧，以降低回环误判概率。

为验证本文算法的性能，在公开数据集 TUM、KITTI 和 EUROC 中分别对基于 ConvNeXt 孪生网络的动态视觉词典更新算法和基于结构相似性评估的回环检测算法进行实验，并与主流的 ORB-SLAM3、OV2SLAM 算法进行对比分析，验证本文所提算法在视觉 SLAM 回环检测的有效性和优势。同时，在自主搭建的硬件平台上进行真实场景实验，验证算法的可行性，为视觉 SLAM 回环检测算法的应用研究提供参考。

关键词：同时定位与地图构建，词袋模型，动态更新，结构相似性评估，回环检测

RESEARCH ON IMPROVED ALGORITHM FOR VISUAL SLAM LOOP CLOSURE DETECTION BASED ON BAG-OF-WORDS MODEL

ABSTRACT

Simultaneous Localization and Mapping (SLAM) refers to the capability of mobile robots to achieve self-localization and environmental map construction in unknown environments through real-time sensor data acquisition. Based on sensor types, SLAM implementations are primarily categorized into LiDAR-based SLAM and vision-based SLAM. Compared with LiDAR-SLAM, visual SLAM has emerged as the predominant research direction due to its cost-effectiveness, superior sensor fusion potential, and broad scenario adaptability. In VSLAM systems, loop closure detection serves as a critical technology for achieving high-precision localization and map optimization. It establishes constraint relationships between current frames and historical frames by identifying whether the camera re-enters previously visited areas, thereby correcting localization drift caused by accumulated errors. However, traditional bag-of-words model-based loop detection methods exhibit limitations in dynamic visual vocabulary updating when confronted with environmental variations or novel scenarios. Simultaneously, their neglect of spatial structural relationships between environmental objects leads to low recall

rates in loop detection and substantial localization errors in SLAM systems. Therefore, enhancing the dynamic adaptability of BoW models and improving loop detection recall rates constitute the primary challenges addressed in this study.

To address the limitations of traditional BoW-based visual SLAM algorithms specifically, the weak scene representation of generated visual words and the inability to adapt to environmental changes—this paper proposes a dynamic visual vocabulary update algorithm based on a ConvNeXt Siamese network. First, ConvNeXt is adopted as the backbone to construct a Siamese architecture with shared weights, and a dynamic feature fusion module is employed to integrate features extracted from different viewpoints. Then, Principal Component Analysis is applied for dimensionality reduction, and a hierarchical density-based clustering algorithm is used to remove noise and construct the visual vocabulary. Finally, based on an offline-initialized vocabulary, a hierarchical tree search mechanism and a node update strategy are introduced to enable dynamic updates of the vocabulary in response to environmental changes.

To address the issue of traditional visual SLAM loop closure algorithms neglecting spatial structural relationships, this paper introduces an improved loop closure detection algorithm based on structural similarity assessment. First, a global structural similarity score is computed for candidate images to capture the overall structural information between

images. Then, a dynamic weighting mechanism is applied to fuse local feature scores with global structural similarity scores, yielding more accurate loop closure candidates. Finally, a candidate frame rejection mechanism is designed, which utilizes historical loop closure results to eliminate incorrect candidates, thereby reducing false-positive loop closures.

To evaluate the performance of the proposed algorithms, experiments were conducted on the publicly available TUM, KITTI, and EUROC datasets. The experiments assessed both the dynamic visual vocabulary updating algorithm based on the ConvNeXt Siamese network and the loop closure detection algorithm based on structural similarity evaluation. A comparative analysis was performed against state-of-the-art ORB-SLAM3 and OV2SLAM algorithms to validate the effectiveness and advantages of the proposed methods in visual SLAM loop closure detection. Additionally, real-world experiments were conducted on a self-developed hardware platform to verify the feasibility of the proposed algorithms, providing valuable insights for the application research of visual SLAM loop closure detection algorithms.

Xu Ao(Electronic Information)

Supervised by Associate Professor Sun Xinzhu

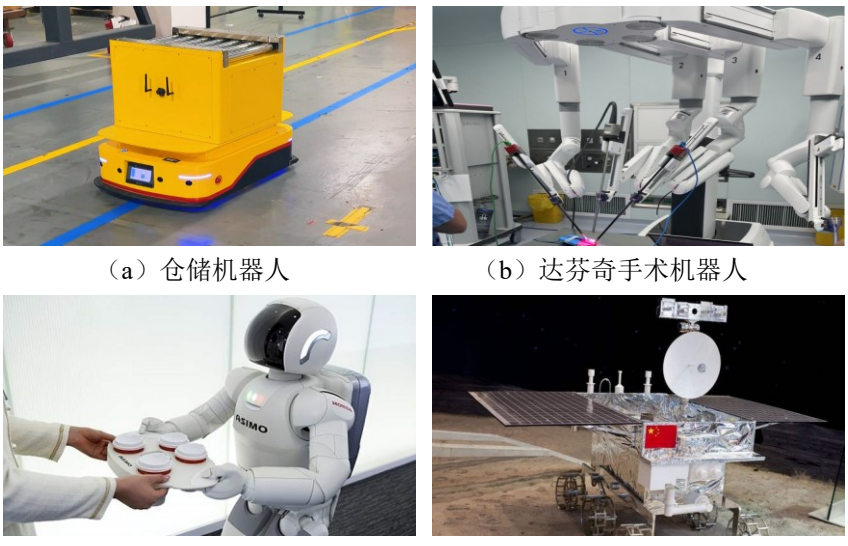
KEY WORDS: Simultaneous Localisation and Mapping, Bag-of-Words Model, Dynamic Updating, Structural Similarity Index Measurement, Loop Closure detection

目 录

第 1 章 绪论

1.1 本文研究背景及意义

机器人技术作为 21 世纪最具影响力的关键技术之一，不仅彻底重塑了传统工业生产模式，而且还在深刻改变着人类社会的组织形式和生活方式。早期的机器人主要用于完成高重复性和高风险的操作任务。随着人工智能和机器学习技术的突破性进展，机器人技术实现了从自动化向智能化的范式跃迁，其影响力已延伸至众多关键领域，成为推动产业变革的核心驱动力[1][2]。在工业制造领域，智能机器人不仅承担传统装配、搬运等任务，还通过深度学习实现质量检测与故障诊断；在医疗领域，以达芬奇手术机器人为代表的智能化手术系统，已在全球范围内广泛应用于临床实践，累计完成数百万例手术，显著提升治疗效果；在社会服务领域，教育机器人和家庭服务机器人正通过个性化教学方案和全方位服务支持，逐渐改善人们的生活质量。在航天领域，机器人技术被广泛应用于执行人类无法完成的高风险任务，如外太空探索、卫星维修和深空探测等。机器人技术发展水平不仅代表了国家高新技术发展先进程度，同时也反应出国家在科技领域的综合实力。而同步定位与地图构建（Simultaneous Localization and Mapping, SLAM）作为移动机器人实现自主定位和建图的核心技术，已成为当前机器人领域研究的热点方向[3][4]。机器人应用示例，如图 1-1 所示。



（a）仓储机器人　（b）达芬奇手术机器人

（c）家庭服务机器人　（d）月球探测器

图 1-1 机器人应用示例

SLAM 技术使得移动机器人在没有先验环境信息的情况下，依托传感器实时获取观测数据，并基于概率推理和非线性优化理论，联合推演机器人在全局坐标系中的六自由度位姿，以及环境的三维拓扑结构，从而实现机器人在未知环境中的定位与地图构建[5][6][7]。根据所使用传感器的不同，现有 SLAM 系统主要分为两大类型：激光 SLAM 和视觉 SLAM。激光 SLAM 通过激光雷达发射激光束并接收反射信号，利用飞行时间原理生成精确的三维点云数据，进而通过点云配准算法实现位姿估计[8][9]。激光 SLAM 的优点在于高精度的空间感知能力，尤其在结构化环境中，能够生成高密度的三维点云数据并提供精确的地图。然而，激光 SLAM 也存在一些局限性，如激光雷达的高成本、对环境光照的依赖、以及点云稀疏性导致的建图精度问题。视觉 SLAM 通过相机采集的图像序列进行特征提取与匹配，并结合运动估计与非线性优化来估计相机的位姿和构建地图[10][11]。与激光 SLAM 相比，视觉 SLAM 借助于低成本的消费级 CMOS 相机、多模态数据融合的潜力和较强的稠密环境感知能力，逐渐在服务机器人、增强现实等领域获得广泛应用，因此成为 SLAM 领域的重要研究方向。视觉 SLAM 系统运行效果图，如图 1-2 所示。

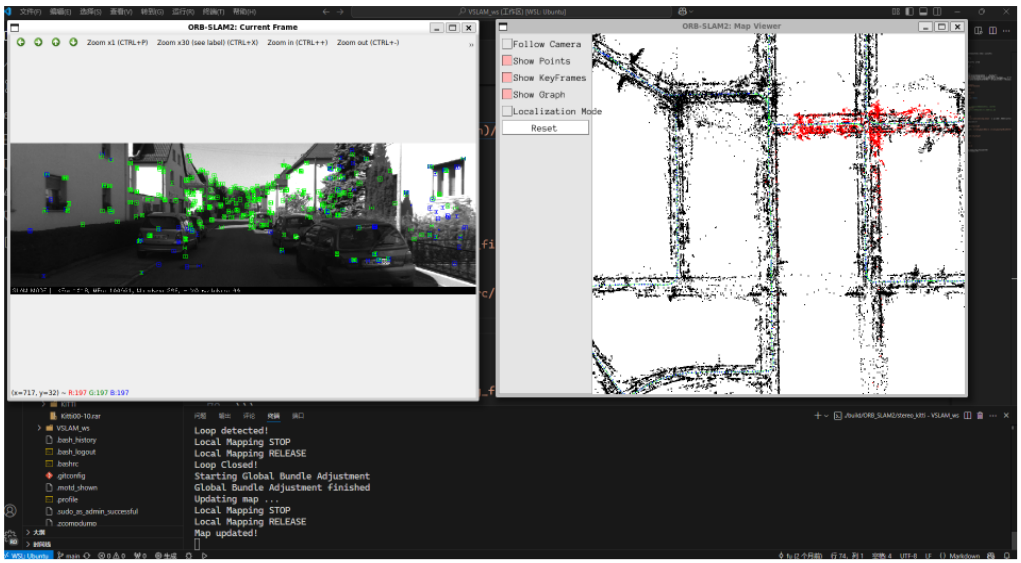


图 1-2 视觉 SLAM 系统运行效果图

视觉 SLAM 在长时间运行过程中，由于累计误差的存在，容易导致定位漂移，从而影响全局一致的地图构建。为了进行误差修正，回环检测模块被引入 SLAM 系统[12]。目前主流回环检测方法是基于图像间的匹配完成的，该方法通过查询当前图像与历史图像的相似关系来实现回环检测，因此回环检测转变为相似

度计算问题。词袋模型算法通过查询视觉词典，用视觉单词向量表示图像，图像间的相似度判断变成了向量间的距离计算[13]。在视觉 SLAM 中，回环检测模块采用词袋模型算法构建视觉词典，并通过查询视觉词典，将当前帧与历史关键帧数据库进行相似性度量，从而判断机器人是否进入过历史场景。当系统识别到重访历史场景时，基于空间几何约束，建立当前帧与历史帧之间的约束关系，并将其转化为图优化问题[14]。此时，时空关联机制将局部误差均匀分布至全局位姿图，从而消除轨迹误差，显著提升 SLAM 系统的定位精度和地图全局一致性。回环检测技术已深度集成于 ORB-SLAM3[15]、Dyna-SLAM[16]等开源框架，并为无人机自主导航、智能仓储机器人等应用提供了鲁棒的时空感知基础。

随着 SLAM 技术在动态开放环境和大规模复杂场景中的应用，基于传统词袋模型的视觉 SLAM 回环检测方法表现出一定缺陷：（1）现有方法采用量化局部特征描述子，并建立倒排索引结构，离线构建视觉词典作为场景检索的基准框架；但在长期运行的场景中，环境光照变化、动态物体干扰以及季节更替等因素会引起场景语义变化，静态词典无法及时更新，导致视觉单词与真实场景的语义关联性逐渐减弱。（2）当前主流方法仅依赖局部特征点间的视觉相似性度量，忽视了场景中物体间的空间拓扑关系。而纯粹基于外观匹配的范式在面对重复纹理区域时极易发生误匹配，从而导致回环检测的召回率降低，进而影响 SLAM 系统的定位精度。因此，如何提升回环检测在复杂环境中的适应性和精度，成为亟待解决的关键问题[17][18]。

综上所述，在复杂场景中引入视觉词典的动态更新机制，并融合相似性度量中的空间结构关系，对视觉 SLAM 系统的回环检测与系统定位精度具有显著作用，且为后续的后端优化与建图提供了坚实的基础。本文以提升视觉 SLAM 系统定位精度和回环检测算法的召回率为目标，针对传统回环检测算法中词袋模型在场景表征能力和空间结构特征缺失方面的不足，以及视觉词典面对场景变化无法更新的问题，提出了一种基于 ConvNeXt 孪生网络的动态视觉词典更新算法和基于结构相似性评估的回环检测算法。重点研究了增强词典模型中的视觉单词表征能力，设计了视觉词典的动态更新机制；并采用融合物体空间结构信息相似度度量方法，以及设计了回环候选帧剔除机制的回环检测算法，有效提高了回环检测的精度和 SLAM 系统鲁棒性。

1.2 国内外研究现状

1.2.1 视觉 SLAM 研究现状

视觉 SLAM 是机器人在未知环境中，仅依靠相机实时感知环境、同步确定自身位置并构建环境地图的关键技术。核心在于通过图像处理和计算机视觉算法，从连续的图形序列中提取特征信息，进而实现高精度的定位与建图。视觉 SLAM 的理论基础可追溯至 20 世纪 80 年代，Smith 等人在 1986 年的研究中首次将概率论方法引入机器人定位与建图领域，奠定了视觉 SLAM 的早期理论框架。然而，受制于当时计算能力和传感器技术的限制，相关研究主要集中在实验室环境中，未能实现广泛应用。直到 2006 年，Durrant-Whyte 与 Bailey 正式提出“SLAM”术语并构建了完整的概率图优化理论框架，推动了视觉 SLAM 研究进入系统化阶段[19]。根据帧与帧之间的关联方式不同，将视觉 SLAM 系统分为直接法和特征点法。

直接法视觉 SLAM 根据最小化连续图像间的光度误差来估算相机运动和环境结构，用于构建稠密或半稠密地图。2011 年，Newcombe 等[20]提出的 DTAM 算法最早的基于直接法的视觉 SLAM 方法之一，通过密集图像跟踪来实现机器人在未知环境中的定位和地图构建。DTAM 无需特征提取即可在纹理稀缺的环境中进行定位与建图，适用于低纹理或动态场景；但其计算资源消耗和对环境光照变化的依赖，导致在实时性和鲁棒性方面存在一定局限。2014 年，Engel 等[21]提出 LSD-SLAM 算法，该算法通过直接法优化图像之间的像素光度误差，从而实现准确的深度估计和地图构建。LSD-SLAM 通过直接使用图像的像素强度信息进行优化，能在低纹理的环境中保持较高的稳定性和精度，适合大规模场景的实时定位和建图。但是由于直接法依赖图像亮度信息，导致对光照变化较为敏感。2017 年，Engel 等[22]提出的 DSO 算法，是一种基于直接法和稀疏优化的视觉里程计算法，旨在通过优化图像中的稀疏像素点来估计相机的位姿变化。虽然在低纹理环境中表现优越，且在动态环境中具备较强的鲁棒性；但是由于没有回环检测机制，导致误差会逐渐积累，影响 SLAM 系统精度。2020 年，Zubizarreta 等[23]提出 DSM 算法，DSM 结合了 DSO 的稀疏直接法特性与全局优化策略，在保持高精度和高效率的同时，显著提升了地图的一致性与可扩展性，适用于大规模

场景下的建图任务。由于依赖光度一致性假设，DSM 对光照变化较为敏感，且全局优化计算开销较大，限制了其实时性和在资源受限平台上的应用。2023 年，贾嫣晗等[24]提出一种双目直接法视觉 SLAM，该算法利用双目相机的基线信息实现深度估计的物理尺度恢复，同时通过双目光度一致性约束优化位姿，在保持直接法计算效率的前提下提升定位精度。

特征点法视觉 SLAM 是通过提取图像中的显著特征点（如角点、边缘或局部图案）来实现机器人的定位和建图的技术。与直接法 SLAM 不同，特征点法依赖于图像的局部特征来建立场景的表示，并使用这些局部特征进行相机的位姿估计与地图的构建，而且在光照变化条件下具有更强的稳定性，因此在视觉 SLAM 算法的中主要使用特征点法。2007 年，Davision 等[25]提出了 MonoSLAM 算法，该算法使用单目相机获取场景图像帧，并通过特征点法提取特征点，位姿优化阶段采用了卡尔曼滤波器方法。然而，由于滤波器容易受到环境噪声的影响，会导致 SLAM 系统稳定性下降。2015 年，Mur-Artal 等[26]提出了 ORB-SLAM 算法，包含跟踪、局部建图和回环检测三个主要线程，在特征点提取过程中使用了 ORB 特征，这种方法具有较高的鲁棒性，有效提升了帧与帧之间特征点的跟踪精度。2017 年，Mur-Artal 等[27]基于 ORB-SLAM 提出了 ORB-SLAM2 系统，该系统对不同类型的传感器（如单目、双目和 RGB-D 相机）进行了优化，并在检测到回环后及时对帧间约束关系进行优化，进一步提升了系统的稳定性。ORB-SLAM2 系统框架，如图 1-3 所示。

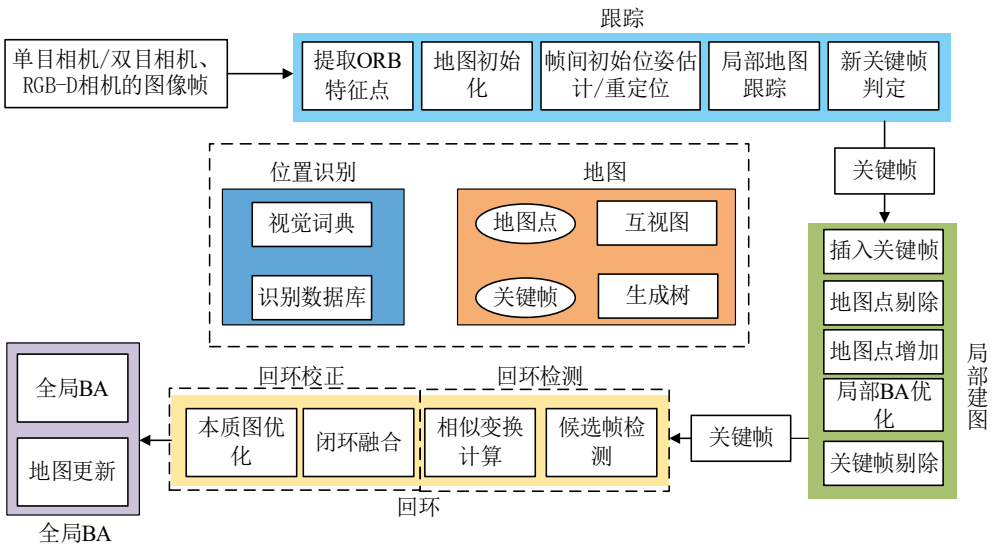


图 1-3 ORB-SLAM2 系统框架图

2018 年，张涛等[28]提出了一种改进的双目视觉里程计算法，采用改进的随机采样一致性算法（RANdom SAmple Consensus，RANSAC）算法选择优质特征点，有效降低了特征匹配错误率并提高了匹配效率。同时，提出了一种双向重投影的位姿估计算法，减少了视觉里程计的累积误差，提高了位姿估计精度，但该方法在动态物体干扰场景下的鲁棒性不足。2020 年，Carlos 等[15]提出了 ORB-SLAM3 算法，加入了多地图复用机制，在室内外环境中，相较于 ORB-SLAM2，ORB-SLAM3 展现出了更高的稳定性，但其在弱纹理环境中仍存在特征跟踪失效问题。2023 年，张洪等[29]在 ORB-SLAM2 的基础上，引入了旋转和平移量策略作为是否创建关键帧的依据，这一创新有效降低了系统的误差，提高了定位精度和鲁棒性；但该策略对运动量阈值的设定较为敏感，在剧烈运动场景中易产生关键帧冗余，且未解决动态物体导致的特征误匹配问题。

1.2.2 回环检测研究现状

在视觉 SLAM 领域，当机器人在某个位置的估计发生偏差时，误差会随时间逐渐累积，从而导致总体误差的增大。这种累积误差会对机器人的地图构建和定位精度造成显著影响。为了解决这个问题，回环检测被引入视觉 SLAM 中。回环检测的主要功能是判断机器人当前所处的场景是否与之前经历过的场景相同，建立地图回环约束，从而调整地图构建和定位策略[30]。当系统确认发生回环时，系统会比对新获取的图像与历史图像，通过重定位来修正位置误差，并更新地图，最终有效提升定位精度和系统稳定性。因此，实现高效的回环检测对于提升视觉 SLAM 系统的整体性能至关重要。回环检测校正示意，如图 1-4 所示。

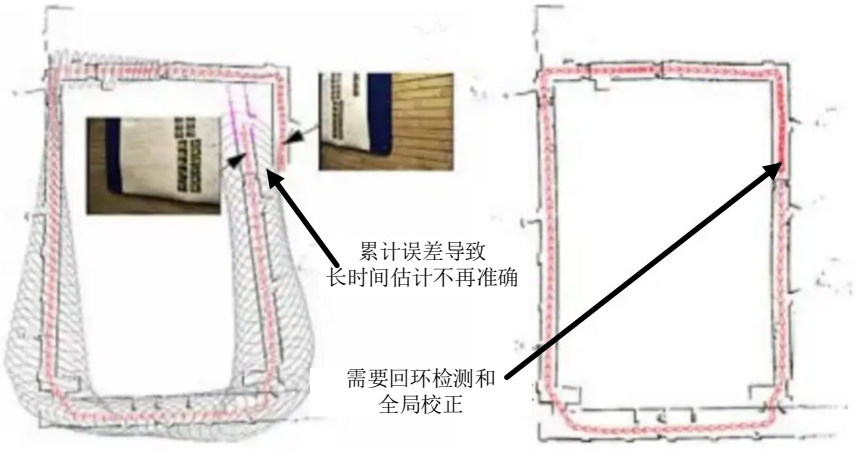


图 1-4 回环检测校正示意图

视觉 SLAM 的回环检测早期的研究主要集中在传统特征匹配方法上，如 SIFT 和 SURF 等。这些方法能够在一定程度上实现回环检测，但在大规模环境和动态场景下的表现有限。随着研究的深入，回环检测逐渐开始借助局部地图信息来提高鲁棒性。2007 年，Clemente 和 Davison[31]提出了一种“分段式”地图概念，大幅提升了回环检测在动态和复杂环境中的鲁棒性，然而，由于空间特征需要构建稀疏地图，最终的回环检测结果往往因为信息量不足而不够精确[32]。2008 年，Williams 等[33]提出了重定位系统，利用三点位姿采样法和一致性随机采样法对位姿进行计算，从而提高了回环检测的实时性。尽管如此，该方法需要大量存储空间来保存地图信息，对内存的要求较高，存在一定的局限性。2012 年，Galvez-Lopez 等[34]提出了 DBoW2（Declarative Bag of Words Library 2），通过树状结构分层构建视觉词典，使用逆序索引加速计算过程。该方法利用高效的词袋模型和逆序索引技术，大幅提升了回环检测的速度和准确性。随着深度学习技术的不断发展，视觉 SLAM 的回环检测也取得了一些突破。2017 年，Xiang Gao 等[35]提出了一种基于深度神经网络的回环检测方法，通过神经网络从图像中学习特征，使得回环检测在复杂环境中变得更加鲁棒。然而，该方法的计算资源消耗较大，且泛化能力有限，难以适应未见过的场景或数据，这限制了在实际应用中的效果和可靠性。除了基于视觉信息的回环检测方法，结合视觉与惯性信息的融合方法也备受关注。2019 年，Tong Qin[16]等人提出了基于视觉与惯性传感器数据融合的导航定位系统——VINS-Fusion。通过同时使用相机和惯性测量单元的数据，VINS-Fusion 能够有效地利用视觉信息来补偿惯性传感器的漂移。然而，该方法存在传感器误差叠加问题，尤其在复杂环境或长时间运行中，误差会逐渐累积，影响系统的稳定性和可靠性。除了基于视觉信息的回环检测方法，基于语义信息的描述方法也越来越受到关注。2021 年，Yuan 等[36]提出了 SVG-Loop 方法，该方法将视觉信息与语义信息相结合，在公共数据集 KITTI 上表现出较好的回环检测性能。然而，由于标注成本较高，许多用于特征提取的模型参数通常是基于通用数据集直接训练得到的，导致算法在特定场景中的泛化能力较弱[37]。

目前，主流的回环检测方法主要是基于传统词袋模型算法。词袋模型算法通过提取大量图像特征，并使用聚类分析算法对特征点描述子进行分类，进而通过词频-逆文档频率（Term Frequency-Inverse Document Frequency，TF-IDF）加权机

制赋予权重，将每一类特征视为“视觉单词”，构建视觉词典；将图像之间的相似度判断则转变为向量之间的距离计算，从而通过查询当前图像与历史图像的相似性来判断是否发生回环。该方法在较大规模场景的视觉 SLAM 中，能有效解决回环检测的问题。2003 年，Sivic 等[38]引入了词袋模型，通过降噪处理的图像特征构建特征字典，并通过字典中的特征单词来描述环境，从而判断是否发生回环。然而，这种方法在处理大规模或动态环境时受到性能限制，特别是在特征提取和匹配阶段。2005 年，Newman 及其团队[39]提出了一种基于特征存储库的算法，及时保存图像帧的特征信息，随后将当前帧与历史关键帧进行匹配，判断是否发生回环。虽然这一方法能较好地应对实时计算，但随着存储库中图像数量的增加，匹配速度和内存管理的问题逐渐凸显。2008 年，Angeli 等[40]在词袋模型的基础上，增加了图像形状和颜色信息的表达，同时构建了增量式字典，增强了特征信息的表征能力。通过结合贝叶斯概率模型和图像分类模型来判断回环，提升了准确性。然而，增量式字典构建仍然面临内存管理和实时性的问题，尤其是在大规模地图和长时间运行时。2013 年，Labbe 等[41]提出了一种基于图像颜色和形状特征的增量式字典构建方法，满足了机器人实时计算的需求。然而，由于增量式字典构建需要大量内存管理，使得该方法在大规模应用中的实时性存在挑战。2018 年，林俊钦等[42]通过利用空间信息筛选候选图像，减少了搜索范围，并通过比较当前帧与候选帧的相似性确认回环的存在。这一方法有效提高了回环检测的速度，但在动态环境中，空间信息筛选会受到环境变化的影响，从而影响检测精度。2021 年，Garcia-Fidalgo 等[43]提出了 OV2SLAM 算法，通过引入多层树状结构到二进制视觉词典中，使得视觉词典可以高效地检索、插入和删除新的视觉单词，从而提供了更高的灵活性和性能。虽然该方法能显著提升回环检测的效率，但在复杂环境下，性能仍然受限。

1.3 研究内容与章节安排

本文为实现复杂场景下移动机器人高精度定位和建图，提出一种基于词袋模型的视觉 SLAM 回环检测改进算法，该算法包含视觉词典构建和回环检测两个环节。在视觉词典环节，为了解决传统视觉 SLAM 算法在回环检测中的视觉单词代表性不足和无法根据环境变化动态更新词典的问题，提出基于 ConvNeXt 孪

生网络的动态视觉词典更新算法。在回环检测环节，为了解决传统视觉 SLAM 算法在回环检测中的忽略空间结构和错误候选帧的问题，提出基于结构相似性评估的回环检测算法。论文总体技术路线如图 1-5 所示。

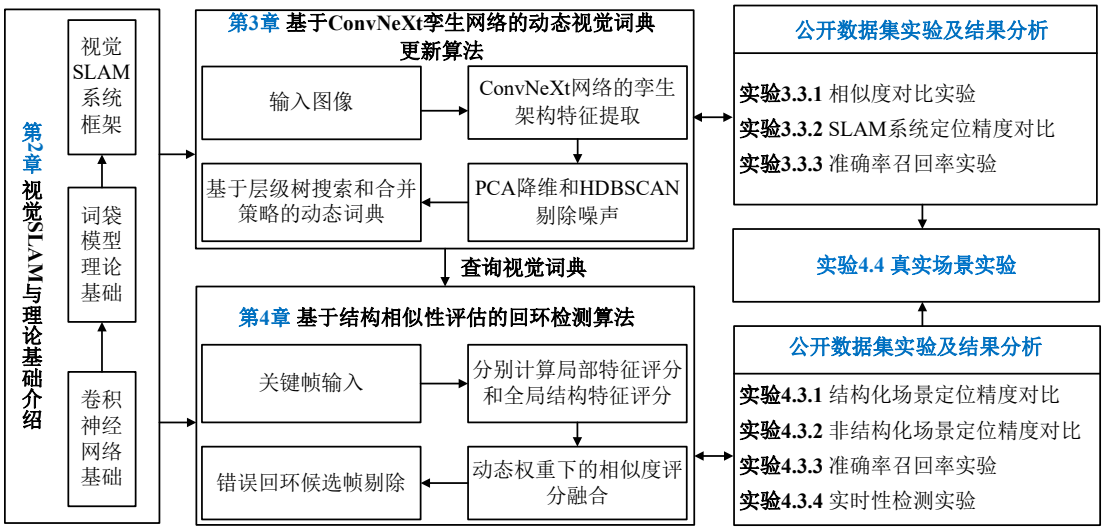


图 1-5 论文总体技术路线图

各章节主要内容如下：

第 1 章，绪论。首先介绍 SLAM 技术背景和应用案例，分析总结了当前视觉 SLAM 研究现状。然后重点介绍了回环检测研究现状，并对本文研究内容进行论述，最后对本文内容进行详细介绍。

第 2 章，视觉 SLAM 与词袋模型、卷积神经网络基础理论。首先分别对视觉 SLAM 框架中的视觉里程计、后端优化、回环检测、建图等模块进行理论介绍和数学建模。然后介绍词袋模型算法工作原理与基础理论，最后介绍卷积神经网络的发展过程和工作原理。为基于词袋模型的视觉 SLAM 回环检测改进算法研究打下理论基础。

第 3 章，基于 ConvNeXt 孪生网络的动态视觉词典更新算法。为了解决传统视觉 SLAM 算法在回环检测中的视觉单词代表性不足和无法根据环境变化动态更新词典的问题，提出基于 ConvNeXt 孪生网络的动态视觉词典更新算法。首先，选用 ConvNeXt 构建共享权重的孪生网络提取特征；其次，采用动态特征融合模块（Dynamic Feature-fusion Module，DFM）整合双路网络不同视角下的特征，确保信息的有效融合；然后，通过主成分分析（Principal Components Analysis，PCA）进行降维处理，减小计算复杂度，同时使用分层密度聚类（Hierarchical Density-Based Spatial Clustering，HDBSCAN）去除噪声，并构建离线视觉词典；最后，

在离线视觉词典的基础上，结合层级树的高效检索机制，利用合并策略和节点递归重建实现视觉词典的动态更新。在公开数据集中对本章算法进行验证。

第 4 章，基于结构相似性评估的回环检测算法。为了解决传统视觉 SLAM 算法在回环检测中的忽略空间结构的问题，提出基于结构相似性评估的回环检测算法。首先，采用结构相似性评估算法计算图像的全局结构相似性评分，捕捉图像之间的整体结构信息，以提高回环检测的准确性；其次，通过动态权重机制，将图像的局部特征评分与全局结构相似性评分融合，优化回环候选帧的选择，提高候选帧的准确性和相关性；最后，设计回环候选帧剔除机制，通过历史回环结果对错误的回环候选帧进行剔除，降低回环误判的概率，进一步提升回环检测的鲁棒性。在公开数据集中对本章算法进行验证，同时在真实场景中对本文整体算法进行验证。

第 5 章，全文总结与展望。对本文的研究内容进行详细的工作总结，并分析本文所提算法优缺点，以及对下一步工作计划的展望。

第 2 章 视觉 SLAM 与词袋模型、卷积神经网络基础理论

本章节主要介绍视觉 SLAM 系统框架、词袋模型基础理论和卷积神经网络基础理论。首先，分别对视觉 SLAM 系统的视觉里程计、后端优化、回环检测和建图的相关理论进行介绍；然后，对词典模型基础理论进行详细叙述；最后，介绍卷积神经网络基础理论。为后续基于词袋模型的视觉 SLAM 回环检测改进算法研究垫定理论基础。

2.1 视觉 SLAM 系统框架

视觉 SLAM 系统由传感器信息获取、视觉里程计、后端优化、回环检测和建图模块组成[44]，系统框架如图 2-1 所示。传感器信息获取模块负责获取相机数据并进行预处理，确保后续模块接收到高质量数据。视觉里程计作为前端的核心模块，通过特征点匹配建立帧间约束关系，估计帧间位姿，并同步构建局部地图。后端优化模块将视觉里程计估计的位姿与回环检测提供的场景重定位信息相融合，通过全局优化抑制误差累积。回环检测模块识别是否到达已访问过的场景，建立当前帧与历史帧之间的位姿约束，为后端优化提供回环约束。建图模块根据优化后的轨迹生成环境地图。视觉 SLAM 系统通过“前端实时追踪—后端全局优化—回环误差修正”的协同机制，实现高精度定位与全局一致性地图构建。本节主要介绍视觉里程计和回环检测，同时对后端优化与建图模块进行简要阐述。

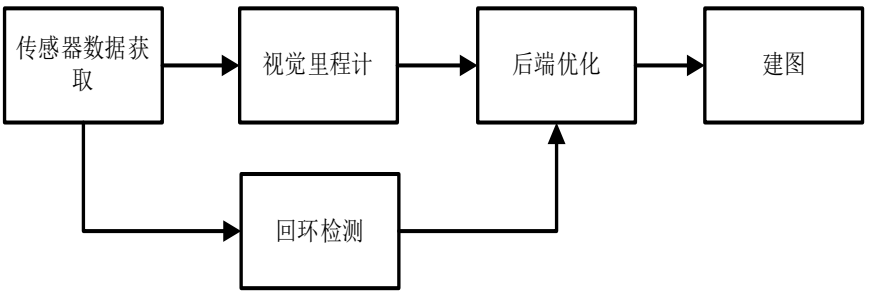


图 2-1 视觉 SLAM 系统框架图

2.1.1 视觉里程计

视觉里程计的核心任务是基于图像序列估计机器人在空间中的位姿变化，关

键步骤是提取并匹配特征点，从而建立帧间约束关系，最终求解相机的位姿。在视觉 SLAM 系统中，特征点法因其对环境光照变化的鲁棒性较强，且计算效率较高，而被广泛采用。

视觉 SLAM 系统中常用的特征点包括 SIFT、SURF 和 ORB 等。ORB 特征由于计算效率高、对旋转和光照变化具有一定的鲁棒性，在视觉 SLAM 中得到了广泛应用。ORB 特征由 FAST 角点检测器和 BRIEF 描述子两部分组成[45]。其中，FAST 角点检测器用于从图像中提取特征点，通过对像素邻域的亮度变化进行快速判别，以高效地检测关键点。FAST 角点检测如图 2-2 所示。

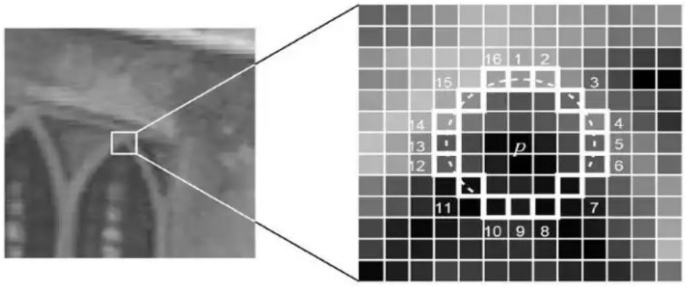


图 2-2 FAST 角点检测

BRIEF 描述子通过在特征点周围定义固定大小的邻域，从中随机选取 n 组像素对 (I_p, I_q)，若 I_p 的灰度值小于 I_q，则对应二进制位设为 1，否则设为 0，最终串联形成长度为 256 的描述子。BRIEF 描述子可视化图，如图 2-3 所示。

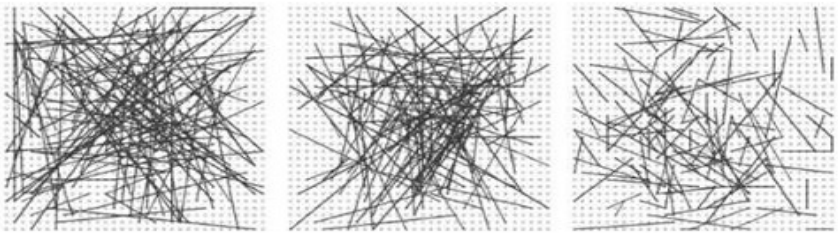


图 2-3 描述子可视化图

原始 BRIEF 对旋转变化较为敏感，因此 ORB 通过灰度质心法计算特征点主方向，并对采样点对进行旋转，以增强其旋转不变性。灰度质心法利用角点邻域内像素灰度分布的质心偏移量确定角点的方向。

在完成特征提取与匹配后，SLAM 系统基于特征点匹配估计相机的位姿，以实现运动跟踪和地图构建。位姿求解依赖于几何约束与优化模型，根据不同传感器类型和场景需求划分为以下核心方法：

1、2D-2D 对极几何法：基于两幅图像中的匹配特征点构建对极约束，通过

基础矩阵或单应矩阵求解相对位姿[46]。对极几何法需借助 RANSAC 剔除误匹配点，并对本质矩阵进行分解以获取候选位姿，但由于单目相机无法直接测量深度，求解结果存在尺度不确定性。

2、3D-2D PnP（Perspective-n-Point）法：已知一组 3D 空间点及其在图像中的 2D 投影，采用最小化重投影误差的方法直接求解相机位姿[47]。常见的 PnP 算法包括 DLT、P3P 以及 EPnP。PnP 法在双目、RGB-D 系统中可直接获取尺度信息，但依赖精确的 3D 点云数据。

3、3D-3D ICP（Iterative Closest Point）法：ICP 法针对点云数据，采用 SVD 分解计算旋转矩阵和平移向量[48]，通过迭代最小化对应点之间的欧式距离求解位姿。ICP 法适用于 LiDAR/深度相机，但在部分遮挡场景易陷入局部最优解。

2.1.2 后端优化

视觉 SLAM 在长时间运行过程中，由于误差不可避免，导致误差逐渐积累，出现定位漂移，对构建全局一致的地图产生重大影响。后端优化模块则通过全局信息校正误差，以确保高精度位姿估计和地图的长期一致性。常见的后端优化方法包括非线性优化和图优化，光束法平差（Bundle Adjustment，BA）[49]和位姿图优化（Pose Graph Optimization，PGO）为核心。

BA 采用非线性最小二乘框架，通过联合优化相机位姿和三维地图点坐标，使多视角间的重投影误差最小化。BA 充分利用稀疏矩阵特性，并采用高效求解器（如 Ceres、g2o）进行优化，使其在高精度三维重建和 SLAM 系统中得到广泛应用。由于 BA 计算复杂度较高，直接对所有关键帧执行全局优化的计算开销较大，因此在实时 SLAM 系统中，通常采用滑动窗口法，仅对最近的若干帧进行优化，以在精度与计算开销之间取得平衡。

PGO 则简化了优化模型，仅将相机位姿作为图优化的顶点，并利用相邻帧的相对位姿约束或回环检测结果构建优化问题。PGO 采用高斯-牛顿法或列文伯格-马夸尔特算法进行优化，计算复杂度相对较低，适用于大规模场景的全局位姿调整。由于 PGO 仅优化位姿，不涉及 3D 地图点，计算负担远小于 BA，因此常用于视觉 SLAM 系统的回环检测后端优化，如 VINS-Mono、ORB-SLAM3 等算法均采用该方法进行全局轨迹优化。

2.1.3 回环检测

回环检测是视觉 SLAM 系统中的核心模块之一，主要任务是识别机器人是否经过已知的重复场景，通过有效的回环检测，系统能够消除由于累积误差所导致的定位漂移，从而确保整个轨迹和地图的全局一致性。回环检测的关键在于评估当前帧与历史帧之间的相似度，进而判断是否存在回环现象。一旦检测到回环，系统便能利用回环约束进行优化，纠正先前的轨迹误差，增强系统的定位精度和鲁棒性。回环检测不仅有助于提升地图的精度，还能有效提升长时间运行 SLAM 系统的稳定性和可靠性。

目前词袋模型是回环检测中的常用方法。词袋模型方法的基本思想是将图像中的特征点映射到视觉词典中，通过查询视觉词典来比较当前帧与历史帧的相似度，从而判断是否发生回环。因此，词袋模型能够在大规模场景中高效地完成回环检测，进而支持视觉 SLAM 系统在复杂环境中的稳定运行。回环检测的过程通常包括以下几个主要步骤（基于词袋模型的回环检测框图，如图 2-4 所示）：

1、回环候选帧筛选。回环检测首先通过图像特征提取与匹配，结合词袋模型方法，筛选出存在回环的候选图像。词袋模型将图像中的局部特征点映射到视觉词典中，将局部特征描述子转化为离散的视觉单词。然后，通过查询视觉词典计算当前帧与历史帧的相似度，利用 TF-IDF 加权余弦相似度来判断是否大于相似度阈值，以此来衡量当前帧与历史帧的相似性，从而筛选出与当前帧相似的历史帧作为回环候选帧。

2、几何一致性校验。几何一致性校验依赖于基础矩阵或单应矩阵，验证当前图像与候选图像之间的匹配关系是否具备真实的几何一致性。几何一致验证方法是基于 RANSAC 算法，通过对部分匹配点进行随机采样与模型估计，并通过迭代优化来去除误匹配，确保最终结果的精度和鲁棒性。

3、回环候选帧确认。当某一历史帧通过几何一致性验证与当前帧建立起稳定的匹配关系后，系统便确认该回环匹配的有效性，并进一步估算当前帧的位姿相对于历史帧的变化。此时，回环检测的成功不仅提供了回环场景的信息，还为系统后端优化提供了有力的约束。

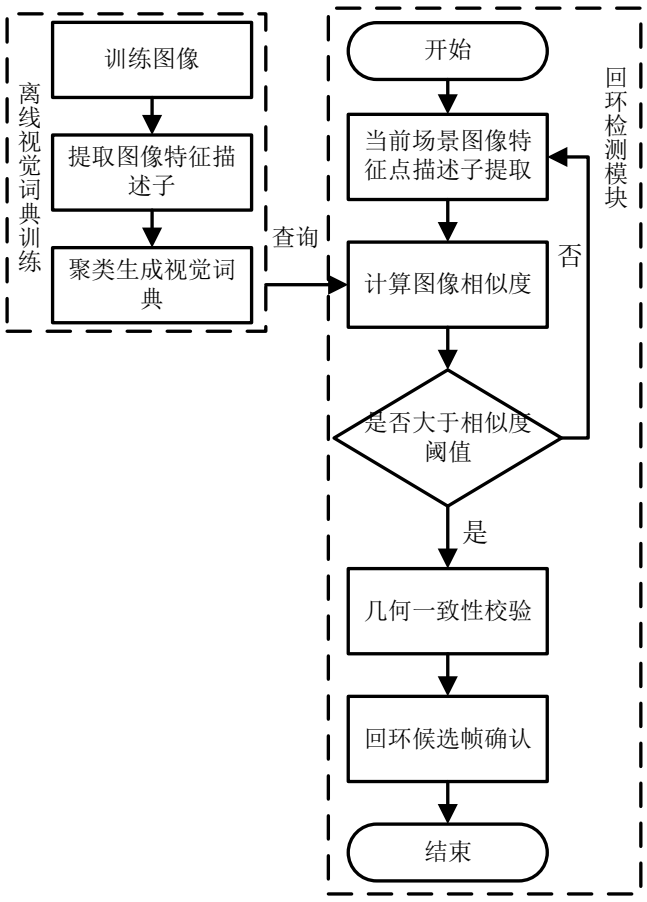


图 2-4 基于词袋模型的回环检测框图

通过上述步骤，回环检测不仅能够识别重复场景，还能有效消除由累积误差引起的定位漂移，从而显著提升整个系统的全局一致性与地图构建精度。这对于长期运行的 SLAM 系统至关重要，尤其在大规模场景和复杂环境中，回环检测对提高系统的鲁棒性和精度具有不可或缺的作用。

2.1.4 建图

视觉 SLAM 的建图模块根据机器人运动轨迹生成环境地图，确保定位精度和地图的全局一致性。建图分为稀疏建图和稠密建图两种方式。稀疏建图通过提取环境中的关键特征点构建地图，适用于大规模场景，具有较高的计算效率，但在环境细节的表现上有所不足。稠密建图则利用深度信息或三维重建生成更加精细的环境模型，能够捕捉更多的细节，但计算复杂度较高，适合需要高精度建模的应用。建图模块不仅需要生成环境地图，还要保持地图的局部和全局一致性。局部地图用于优化当前位姿，而全局地图则随着机器人运动不断更新，确保长期定位精度。

2.2 词袋模型基础理论

词袋模型最初应用于自然语言处理领域，广泛用于文本分类与信息检索任务。该模型将一段文本表示为词频向量，忽略单词之间的顺序关系，通过统计词汇出现频率来衡量文档间的相似性。由于其表示简洁、计算高效的特点，词袋模型逐渐被引入到计算机视觉领域，特别是在视觉 SLAM 中的回环检测与场景识别等任务中发挥了重要作用。

视觉词袋模型借鉴了自然语言处理中的“文档-单词”结构：将图像视为文档，局部特征点类比为单词，进而构建视觉词典。与文本处理不同的是，该模型忽略了特征点在图像中的空间关系与几何结构，仅关注视觉单词的出现频率。这种无序统计的方法提高了特征匹配的效率，使视觉 SLAM 系统在大规模或复杂环境中能够高效执行回环检测与重定位任务。词袋模型可视化，如图 2-5 所示。

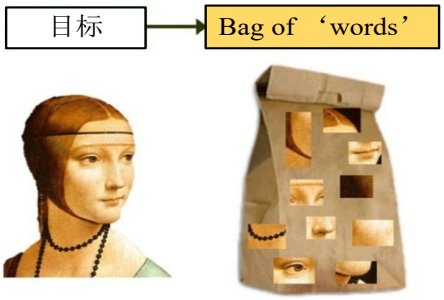


图 2-5 词袋模型可视化

在视觉 SLAM 中，词袋模型通过聚合图像的局部特征点构建“视觉词典”，从而实现对图像内容的紧凑表达与高效匹配。首先，利用局部特征提取算法（如 ORB、SIFT、BRIEF 等）从大规模图像数据集中提取描述子；其次，采用 K-Means 聚类算法对所有描述子进行聚类，生成k个聚类中心，每个聚类中心即视为一个“视觉单词”；最终，这些聚类中心构成视觉词典，图像由其包含的视觉单词频率直方图进行表示，实现图像的量化编码。为了进一步增强词袋模型的判别能力，视觉词袋模型引入 TF-IDF 加权机制。TF-IDF 通过衡量视觉单词在整个数据集中出现的频率来调整权重，有效抑制了高频但低区分度的通用特征（如天空、墙面纹理）对匹配精度的干扰。同时，TF-IDF 能够增强在特定场景中具有代表性和区分度的视觉单词的影响力，从而提升回环检测的准确性。词汇树如图 2-6 所示。

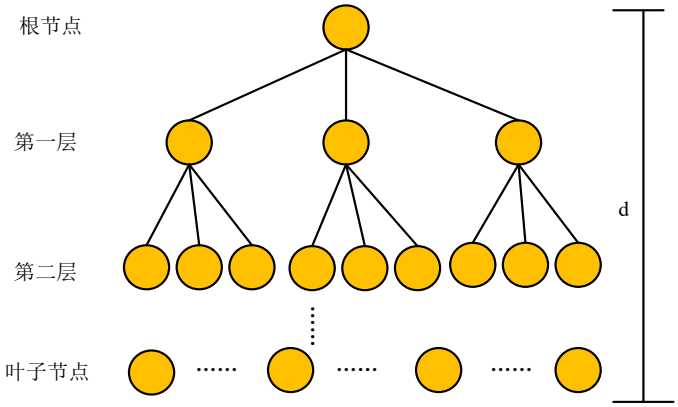


图 2-6 词汇树结构示意图

词袋模型通过查询预先构建的视觉词典，将图像中的局部特征描述子映射到最近的视觉单词，从而将原始高维特征转换为离散的视觉单词表示。对于任意输入图像，所有局部特征被量化为一个视觉单词分布向量，用以紧凑表达图像的内容信息。对于任意输入图像 A，图像特征表示为视觉单词向量，如式（2-1）所示。

$$v_A = \{(\omega_1,\eta_1),(\omega_2,\eta_2)\ldots(\omega_N,\eta_N)\} \tag{2-1}$$

式中 η_N 是 TF-IDF 权重，ω_N 是视觉单词，v_A 是视觉单词 A 的向量化视觉词典是固定的，因此只需要一个向量即可描述整幅图像。通过该向量化表示，图像间的相似性评估可简化为向量间的相似度计算。比较当前帧 v_A 与历史帧 v_B 之间的相似度，如式（2-2）所示。

$$s\left(v_A - v_B\right) = 2\sum_{i=1}^{N}|v_{Ai}| + |v_{Bi}| - |v_{Ai} - v_{Bi}| \tag{2-2}$$

式中 v_{Ai} 和 v_{Bi} 图像 A 和 B 在第 i 个视觉单词上的频率或权重。

词袋模型的核心在于构建并使用一个视觉词典，该词典将图像中的局部特征描述子映射为离散的视觉单词。根据词典的更新方式，词袋模型可分为静态词典和动态词典两种类型。

静态词典是在系统初始化时构建并固定的，通过对大量训练图像进行特征提取和聚类得到。静态词典在回环检测过程中为每一帧图像提供一个固定的视觉单词表示，利用这些词汇的频率或权重向量进行匹配和相似度计算。由于词典是固定的，静态词典具有较高的计算效率，能够快速响应查询和匹配。然而，静态词典的缺点也十分明显，在长时间运行的系统中，无法有效应对场景变化和环境的动态特性，导致回环检测的失败或系统性能下降。

与静态词典不同，动态词典可以在 SLAM 系统运行过程中进行更新。动态词典通过引入新视觉单词来适应环境的变化，并能根据新的图像数据逐步扩充或调整词典内容。相比之下，动态词典通过不断更新词典内容，能够更好地应对长期运行中的环境变化，提升回环检测的准确性和 SLAM 系统的鲁棒性。同时，动态词典能够通过合并、删除等操作优化词汇表，确保词典的紧凑性和区分度。因此，针对传统静态词典在视觉单词代表性不足以及无法根据环境变化动态更新词典的问题，本文在第三章提出了一种基于 ConvNeXt 孪生网络的动态视觉词典模型改进算法。该算法不仅可以动态更新词典，还能提升视觉单词的区分能力，进一步提高回环检测的精度和系统的适应性。

2.3 卷积神经网络基础理论

卷积神经网络（Convolutional Neural Network，CNN）是一种专为处理视觉数据设计的深度学习模型，核心工作原理包括局部感受野、权重共享和多层特征提取。相比于传统的全连接神经网络（Fully Connected Neural Network，FCNN），CNN 通过局部连接、权重共享和层级特征提取，有效减少参数数量，提高计算效率，并具备较强的平移不变性，广泛应用于计算机视觉任务，如图像分类、目标检测、和自动驾驶感知等领域。CNN 主要由卷积层、池化层和全连接层组成，并结合激活函数和归一化等技术，提取图像特征并进行分类。CNN 网络结构示意，如图 2-7 所示。

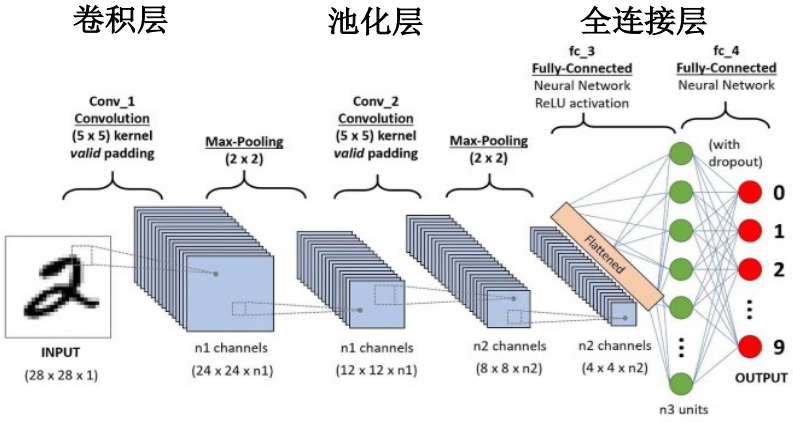


图 2-7 CNN 网络结构示意图

卷积操作使用一组小的滤波器（或卷积核）滑动遍历图像，计算加权求和后生成特征图，从而捕捉边缘、纹理等局部模式。卷积计算公式，如式（2-3）所示。

$$Y(i,j)=\sum_{m}\sum_{n}X(i+m,j+n)\bullet W(m,n)+b \tag{2-3}$$

式中 $X(i+n,j+n)$ 表示输入的像素值，$W(m,n)$ 为卷积核权重，b 为偏置项。卷积层具有局部连接、权重共享等特点，能高效提取边缘、纹理、形状等特征，并减少参数计算量。权重共享机制使同一卷积核在整个图像中复用，大幅减少参数数量，提高计算效率，同时增强模型对平移变化的泛化能力。

池化层是 CNN 中的一种下采样操作，主要用于降低特征图的尺寸、减少计算量，并增强模型的平移不变性。池化层通过对局部区域的像素进行统计运算，常见方法包括最大池化和平均池化。最大池化选取池化窗口内的最大值，保留最显著特征，计算公式，如式（2-4）所示。

$$Y(i,j)=\max(i+m,j+n) \tag{2-4}$$

式中 $X(i+n,j+n)$ 表示输入的像素值。平均池化计算窗口内所有像素的均值，实现平滑特征图的效果，如式（2-5）所示。

$$Y(i,j)=\frac{1}{k^2}\sum_{m=0}^{k-1}\sum_{n=0}^{k-1}X(i+m,j+n) \tag{2-5}$$

式中 k 为池化窗口大小。池化层的关键参数包括池化窗口大小、步长和填充。

全连接层是神经网络中的关键组成部分，位于模型的后端阶段，负责将前面卷积层或池化层提取的局部特征整合为全局特征，并最终输出预测结果。相比于卷积层的局部感受野特性，全连接层中的每个神经元都与上一层的所有神经元相连，因此具有强大的特征融合和信息整合能力。FC 层的计算本质上是一个矩阵乘法，如式（2-6）所示。

$$Y=f(WX+\mathrm{b}) \tag{2-6}$$

式中 W 是权重矩阵，X 是输入特征，b 为偏置项，$f(\bullet)$ 为非线性激活函数，Y 为输出结果。全连接层通过矩阵运算实现特征整合，并通过激活函数引入非线性能力，使得神经网络能够学习复杂的特征表示。

CNN 的多层结构使其能够从低级到高级逐步提取图像特征。而随着计算机视觉领域的不断发展，新的架构和方法开始挑战 CNN 的主导地位。MobileNet 和 EfficientNet 等轻量级 CNN 优化了参数效率，使神经网络能在移动设备上高效运行。与此同时，ViT（Vision Transformer）等基于自注意力机制的模型展现出超越

CNN 的潜力，在某些计算机视觉任务中取得更优表现。尽管如此，CNN 凭借其在特征提取、计算效率和迁移学习方面的优势，仍然在许多视觉任务中占据主导地位，并不断演化，如 ConvNeXt 进一步融合了 Transformer 的设计思想，为 CNN 的发展开辟了新的方向。

2.4 本章小结

本章节主要介绍视觉 SLAM 系统框架、词袋模型的基础理论以及卷积神经网络的基本原理。首先，系统性地阐述视觉 SLAM 的视觉里程计、后端优化、回环检测和建图模块；随后，深入探讨词袋模型的基本原理，介绍其在视觉 SLAM 中的应用及其优缺点；最后，概述卷积神经网络的基础理论，解析其特征提取能力及在计算机视觉任务中的优势，并引出 ConvNeXt 网络。通过这些理论的介绍，为后续基于词袋模型的视觉 SLAM 回环检测改进算法研究奠定坚实基础。

第 3 章 基于 ConvNeXt 孪生网络的动态视觉词典更新算法

针对传统视觉 SLAM 算法在回环检测中的视觉单词代表性不足和无法根据环境变化动态更新词典的问题，本章提出一种基于 ConvNeXt 孪生网络的动态视觉词典更新算法。首先，构建具有参数共享机制的双流 ConvNeXt 孪生网络架构，通过动态特征融合模块聚合多视角特征信息，在保留局部几何细节的基础上，增强模型对复杂场景的表征鲁棒性；其次，通过主成分分析（PCA）对高维特征描述子进行降维，并结合分层密度聚类算法（HDBSCAN）实现噪声鲁棒性特征筛选，构建初始离线视觉词典；最后，设计基于层级树结构的动态更新机制，在离线视觉词典的基础上，判断描述子是否匹配到对应的视觉单词，并通过判断匹配次数，结合层级树的递归重建机制和合并更新策略，实现了视觉词典的动态更新和优化。本章提出的基于 ConvNeXt 孪生网络的动态视觉词典更新算法框架，如图 3-1 所示。

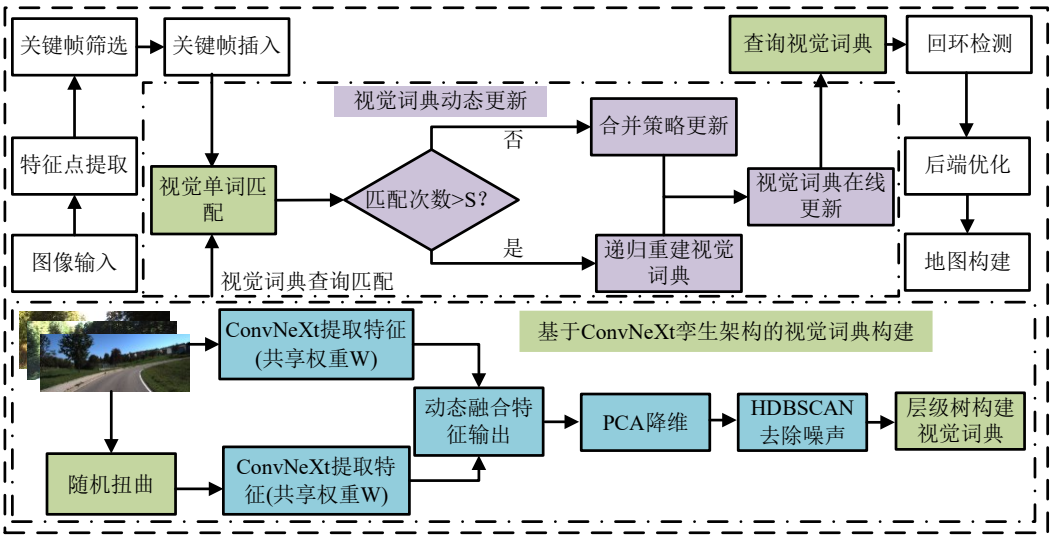


图 3-1 基于 ConvNeXt 孪生网络的动态视觉词典更新算法框架

3.1 基于 ConvNeXt 孪生网络的特征提取

3.1.1 孪生网络架构

孪生网络是一种基于对比学习的神经网络节构，主要用于解决相似度度量与匹配问题。该网络结构由两个结构完全相同且共享权重的子网络组成，每个子网

络分别接收来自不同时刻或不同视角的图像输入，并通过各自的特征提取模块生成相应的特征表示。由于共享权重机制，孪生网络确保了特征提取过程的一致性，有助于提升特征的表征能力，从而在复杂环境下仍能保持稳定的匹配性能。

不同神经网络在特征提取能力与计算效率方面差异显著，因此选择合适的骨干网络对孪生结构设计至关重要。经典的深度卷积神经网络模型，如 VGG19、AlexNet、GoogleNet、和 ResNet，均可作为候选骨干网络，且各具结构特色。为评估不同 CNN 模型在回环检测任务中的表现，本文基于公开数据集 KITTI 的 02 序列开展了精度测试。02 序列采集自室外公共道路场景，包含多个回环位置，适合用于评估神经网络在回环检测中的匹配能力。TOP-1 Accuracy 是回环检测中常用的评估指标，用于衡量模型能否准确识别真实回环帧，反映其特征表征与图像匹配性能。不同网络 TOP-1 Accuracy 对比，如图 3-2 所示，为本章中骨干网络的选择提供参考依据。

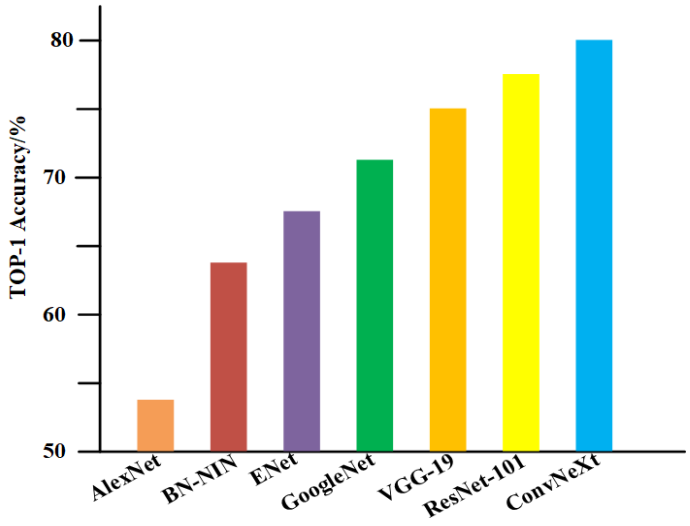


图 3-2 网络 TOP-1 Accuracy 对比图

从图 3-2 可以看出，ResNet-101 和 VGG-19 相较于早期的浅层网络模型，在回环检测任务中表现出更高的准确率，这主要得益于在网络架构设计上的持续优化。随着网络深度的增加（如 ResNet-101、VGG-19），模型在特征提取方面具备更强的表达能力，能够捕捉更丰富且具有区分性的图像特征，从而提高了回环检测的准确性。相比之下，ConvNeXt 在保持卷积神经网络优势的基础上，融合了 Transformer 的设计理念，并对传统卷积模块进行了结构性重构，使其在特征提取效率和多尺度信息建模能力上进一步增强，从而使 ConvNeXt 在 KITTI 02 序列中的 TOP-1 Accuracy 达到了 80%，展现出良好的判别能力。因此，最终选

用ConvNeXt作为孪生网络结构中的主干网络，以兼顾特征表达能力与计算效率。

本章的孪生神经网络架构接收一组配对数据作为输入，两个样本分别通过结构相同且共享参数的 ConvNeXt 网络进行处理。借助共享网络参数的特征映射机制，输入数据被映射至同一潜在空间，生成对应的特征向量。随后，通过动态融合特征模块（DFM）对这些特征进行进一步构建，以获得图像的最终特征表示。本章孪生网络架构，如图 3-3 所示。

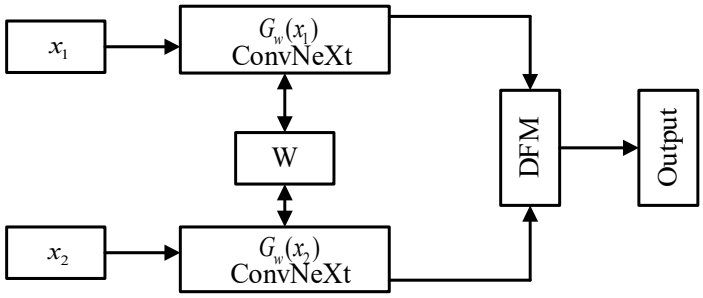


图 3-3 本章孪生网络架构图

3.1.2 主干网络模块

ConvNeXt 采用分层多阶段架构，包括四个主要阶段，每个阶段通过堆叠多个 ConvNeXt Block 逐步提取高层语义特征，同时伴随空间分辨率的递减和通道数的递增。在具体设计上，ConvNeXt Block 引入倒置瓶颈结构，并通过 7×7 深度卷积实现全局感受野，避免了 Transformer 自注意力机制带来的$o(n^2)$计算复杂度。此外，网络采用 1×1 卷积和层归一化（Layer Norm）来调整通道数，以优化特征表示能力。为了增强非线性表达能力，ConvNeXt 采用 GeLU 激活函数替代 ReLU，使梯度流更加平滑，并通过残差连接融合原始特征与处理后的特征，有效防止信息丢失，提升网络的训练稳定性和表达能力。

ConvNeXt 网络通过层级化的结构逐步提取图像特征，并进行深度表征学习。首先，输入图像经过一个 4×4 的卷积层进行初步编码，该过程类似于 ViT 和 Swin Transformer 的初始特征编码，通过步长为 4 的卷积操作将图像映射到高维特征空间，并降低计算量。随后，特征图进入四个级联的特征提取阶段，每个阶段由多个 ConvNeXt Block 组成，逐步提取不同层级的特征信息，提升网络表征能力。ConvNeX 网络架构图，如图 3-4 所示。

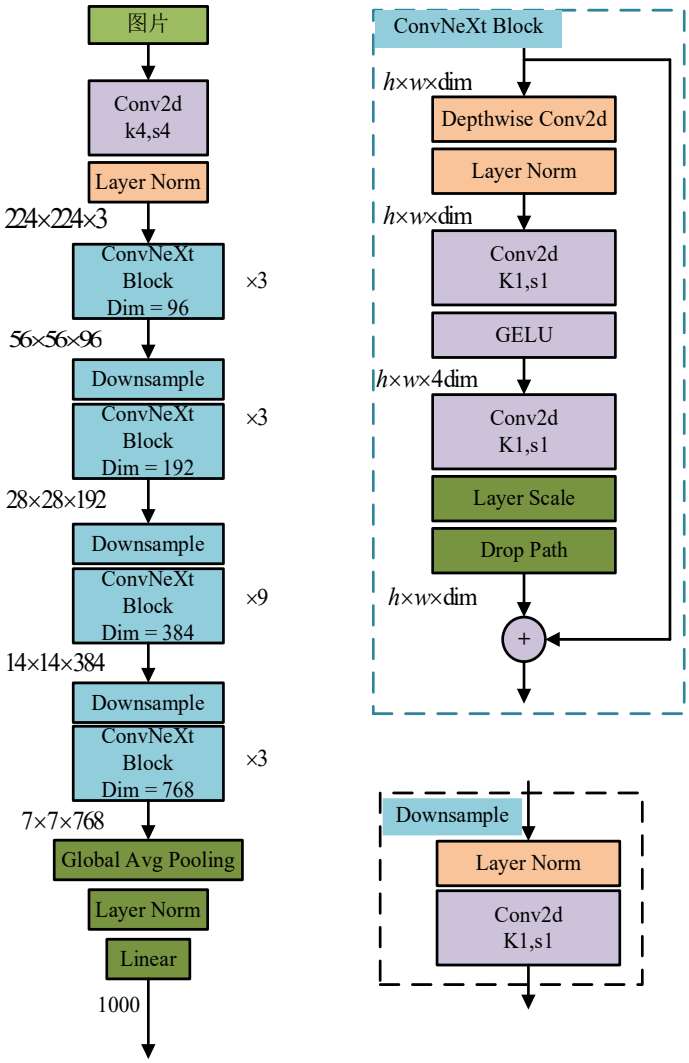


图 3-4 主干网络模块图

ConvNeXt Block 将输入特征经过 7×7 深度可分离卷积进行特征提取，卷积后的特征经过层归一化，消除特征分布的变化，以提高训练的稳定性。随后，网络通过 1×1 卷积调整通道数，并采用 GELU 激活函数。GELU 引入高斯分布的概率特性，使小于 0 的输入值缓慢衰减而非直接截断，从而使梯度流更加平滑，有助于模型的收敛与泛化能力，如式（3-1）所示。

$$GELU(X) = X \bullet \Phi(X) \tag{3-1}$$

式中 $\Phi(x)$ 为标准正态分布的累积分布函数，X 表示输入特征值。最后，ConvNeXt Block 采用残差连接将原始输入特征与变换后特征逐元素相加，如式（3-2）所示。

$$Y = X + X'' \tag{3-2}$$

随着阶段的深入，特征图的空间分辨率逐步降低，而通道数逐步增加，以提取更高级别的语义特征。在第 l 阶段，特征提取表示，如式（3-3）所示。

$$F_l = B_l(F_{l-1}) \tag{3-3}$$

式中 B_l 为该阶段的 ConvNeXt Block 处理过程，F_{l-1} 为上一阶段的特征图。ConvNeXt Block 中，一阶段主要提取边缘、纹理等低级特征；二阶段通过步长为 2 的 2×2 卷积进行降采样，并提取局部结构信息；三阶段进一步降采样，增强对复杂物体的表征能力；四阶段负责全局特征整合，为后续任务提供高级语义信息。

经过四个阶段的特征提取后，最终的特征图输入到全局平均池化层，将空间维度进一步压缩成单个特征向量，如式（3-4）所示。

$$z_c = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} F_{i,j,c} \tag{3-4}$$

式中 $F \in \mathbb{R}^{H \times W \times C}$ 为输入特征图，H 和 W 分别是特征图的高度和宽度，C 是通道数。

3.1.3 动态特征融合模块

动态特征融合模块（Dynamic Feature-fusion Module，DFM）旨在提升特征表示的能力。为了缓解不同特征图之间的差异性，DFM 采用空间注意力机制，动态调整特征图中不同区域的重要性。空间注意力机制通过增强特征密集区域的像素权重，使得网络可以更关注那些信息丰富、特征突出的区域。具体来说，输入特征图首先经过全局平均池化和全局最大池化，分别提取全局信息并突出显著特征。其次，池化后的特征图在通道维度上进行拼接，以整合不同尺度的空间信息。然后，经过 3×3 和 1×1 卷积层以及激活函数处理，生成权重分配特征图，用于动态调整各像素点的关注度，最终通过逐元素相乘操作将输入特征图与权重分配特征图相结合。空间注意力机制，如图 3-5 所示。

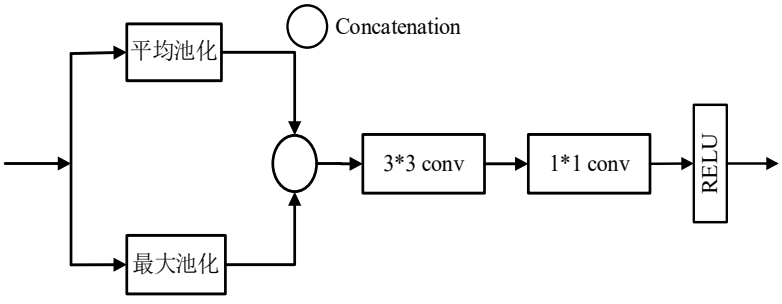


图 3-5 空间注意力机制

在完成空间注意力处理后，DFM 模块采用拼接和加法两种操作方式来融合来自不同模态的特征。拼接和加法的多层次、多方式融合策略使 DFM 能够更有效地整合和强化特征信息，进而提升模型的表示能力。首先，拼接操作通过在通道维度上将特征图拼接，扩展了特征空间的维度，从而保留了更多的细节信息。随后，特征图通过加法操作进一步融合，在保持特征紧凑性的同时，增强同类特征的表达能力。最后，为了将特征输出为二进制形式，DFM 应用 Sigmoid 激活函数将特征值映射到[0,1]区间，并通过阈值化操作将其转换为二进制特征。这一过程进一步确保了特征的二值化输出，符合特定的任务需求。DFM 计算公式，如式（3-5）所示。

$$DFM(F_{ri},F_{li})=\sigma\{C[C(Sa(F_{ri}),Sa(F_{li})),A(Sa(F_{ri})\oplus Sa(F_{li}))]\} \tag{3-5}$$

式中，F_{ri} 和 F_{li} 分别表示图层 i 中的 ConvNeXt 特征；$\sigma(\bullet)$ 为 sigmoid 激活函数；$C(\bullet)$ 表示不同特征拼接后；经过 3×3 卷积；A 为加法后的 3×3 卷积；$Sa(\bullet)$ 表示空间注意力机制模块。动态特征融合模块如图 3-6 所示。

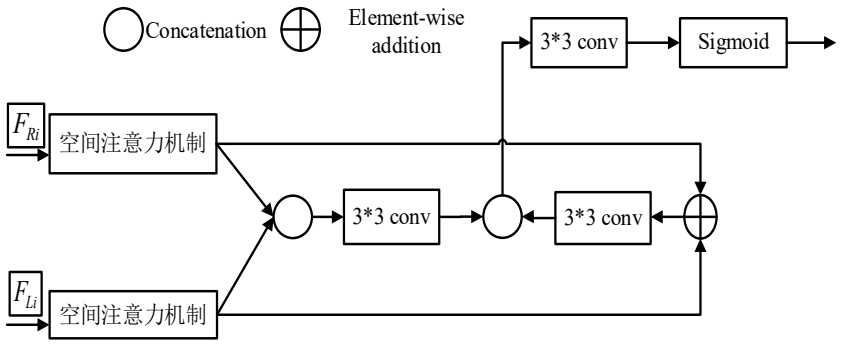


图 3-6 动态特征融合模块

3.2 动态词典构建与更新

传统的视觉 SLAM 系统通常依赖词袋模型来实现回环检测，通过构建的静态视觉词典，对图像描述子进行索引和匹配，从而判断系统是否回到之前的地点。然而，视觉词典构建完成便无法根据环境变化进行动态更新，回环检测算法无法有效识别新的视觉特征，从而产生定位误差。因此，提出动态更新视觉词典机制，通过 PCA 进行降维处理，并采用 HDBSCAN 方法构建离线视觉词典；在离线视觉词典的基础上，结合层级树检索机制和合并更新策略，实现词典动态更新。

3.2.1 PCA 降维

由孪生网络架构提取的特征具有高维度特性，高维特征会增加计算复杂度，并引入冗余信息，从而影响算法的运行效率和匹配精度。因此，本章算法设计的数据降维模块，通过 PCA 法将高维的图像特征描述子映射到低维空间，最大程度上保留描述子所代表的信息。相比较于其他降维算法，PCA 法将高维数据映射到低维空间时，能够捕捉到原始数据中最重要的信息，减少了信息的损失；同时，提取到的特征更加独立和具有代表性。算法计算步骤如下：

1、数据均值中心化。对于提取到的所有 ORB 特征描述子向量按行组成一个数据矩阵 $X_{m\times n}$，其中每一行代表一个特征向量。为确保主成分分析能够更好地捕捉数据的方差，对数据矩阵进行均值中心化操作，即减去每个特征的均值，使得每个特征的均值为 0。因为 PCA 对于初始变量的方差非常敏感，使用均值中心化可以将数据转换为可比较的尺度，从而使得特征之间具有可比性。均值中心化计算公式，如式(3-6)所示。

$$\overline{x_i} = x_i - \mu \tag{3-6}$$

式中 $\overline{x_i}$ 是均值中心化后的数据，x_i 是原始数据，μ 是特征的均值。

2、计算均值中心化后的数据矩阵 $\overline{X}_{m\times n}$ 的协方差矩阵。计算协方差矩阵是 PCA 的基础，有助于找到数据中的主要变异性方向，以便实现数据降维、去除冗余信息和提高数据的可解释性。协方差矩阵计算公式，如式(3-7)所示：

$$C_{n\times n} = \frac{1}{m}\bar{X}^{T}_{m\times n}\bar{X}_{m\times n} \tag{3-7}$$

式中 $C_{n\times n}$ 为协方差矩阵，m 为样本数，$\overline{X}_{m\times n}$ 为均值中心化后的特征矩阵。

3、对协方差矩阵 $C_{n\times n}$ 进行特征值分解，得到 n 个特征值 $\lambda_1,\lambda_2,\ldots,\lambda_n$ 和对应的特征向量 $T_1,T_2,\cdots\cdots,T_n$。求解出的 n 个特征值从大到小排列，并将其标准化，使得每个特征向量的模为 1。数据从 n 维投影至 k 维($k<n$)，取前 k 个特征值对应的特征向量组成特征矩阵 $V_{n\times k}$，如式（3-8）所示。

$$V_{n\times k}=\left[T_1,T_2,\cdots,T_K\right] \tag{3-8}$$

4、将原始数据投影到新的特征空间，得到降维后的数据。投影公式，如式（3-9）所示。

$$Y_{m\times k} = X_{m\times n} \times V_{n\times k} \tag{3-9}$$

式中 $Y_{m\times k}$ 为降维后的数据矩阵，$X_{m\times n}$ 为所有特征构成的矩阵，$V_{n\times k}$ 为特征矩阵。

3.2.2 HDBSCAN 聚类

HDBSCAN 算法是由 Campello 等提出的一种基于密度聚类与层次聚类相结合的算法，该算法根据互达距离的度量方法对不同族群构建树形层次结构，以实现不同密度族群的聚类；与其他聚类算法相比，HDBSCAN 算法具有更好的鲁棒性和泛化性能，且能够有效地处理噪声和离群点，识别为单独的噪声类别。对于真实世界的数据非常有用，因为数据中往往会存在一些异常值或者不属于任何聚类的点。算法计算步骤如下：

1、空间变换，即用互达距离表示两个数据点之间的距离。数据点与第 k 个最近邻样本点的距离称为核心距离，表示为 $core_k(y)$，如式（3-10）所示。

$$core_k(y) = d(y, N^k(y)) \tag{3-10}$$

式中 y 是原始数据点，$N^k(y)$ 是第 k 个点，d 代表欧几里得距离。数据点与数据点之间的距离称为互达距离，表示为 $d_{mreach-k}(a,b)$，如式（3-11）所示。

$$d_{mreach-k}(a,b) = \max\{core_k(a), core_k(b), d(a,b)\} \tag{3-11}$$

式中 $core_k(a)$ 是 a 的核心距离，$core_k(b)$ 是 b 的核心距离，$d(a,b)$ 是 a 和 b 之间的欧几里得距离。

2、构建最小生成树。分层密度聚类算法通过 Prim 算法，将数据看作一个加权图，其中数据点为顶点，任意两点之间的边的权重等于这些点之间的相互可达距离，根据权重阈值构建最小生成树。

3、构建层次聚类结构（聚类树）。根据互达距离对最小生成树的边按照增序排序，然后循环访问，通过并查集来确定每条边所关联的两个簇，将每个边归类一个新的簇。

4、压缩聚类树。确定最小簇大小后，对聚类树进行压缩拆分，拆分出来的簇大小与最小簇大小进行比较，剔除数量小于最小簇大小的一方，最终得到一个拥有少量节点的聚类树。

5、提取簇。将压缩聚类树的每个叶节点都选定为某个簇，自下而上遍历整

棵聚类树，根据其稳定性对簇进行提取。λ 和稳定性计算公式，如式（3-12）所示。

$$\lambda = \frac{1}{d(a,b)} \tag{3-12}$$

式中 λ 为距离的倒数，$d(a,b)$ 为 a 和 b 点之间的欧几里得距离。

$$S_{duster} = \sum_{p \in C_{duster}} (\lambda_p - \lambda_{death}) \tag{3-13}$$

式（3-13）中 λ_p 是聚类簇从父簇分离出去时的 λ 值；λ_{death} 是聚类簇分裂为 2 个子簇时的 λ 值；$C_{cluster}$ 为聚类簇集合。

通过选择每个簇中到其他簇成员的平均距离最小的点作为代表，提取簇中心，增强视觉单词的表征能力。提取簇中心计算公式，如式(3-14)所示：

$$AverageDis\tan ce(a) = \frac{1}{n-1} \sum_{a \in C, a \neq b} d(a,b) \tag{3-14}$$

式中对于簇 C 中的点 a，计算平均距离到其他簇成员的距离 $d(a,b)$ 的总和，其中 b 是簇 C 中的其他成员，n 为簇 C 中点的数量。

在构建视觉词典的词汇树过程中，系统自根节点开始逐层向下扩展。首先，从输入的图像描述子中随机选取k个样本作为初始聚类中心；随后，计算其余描述子到各聚类中心的欧几里得距离，并将每个描述子分配给距离最近的中心。接着，对每个新形成的聚类子集重复上述聚类操作，直至某一子集中包含的描述子数量小于设定的预设值 M 。层级树示例，如图 3-7 所示。

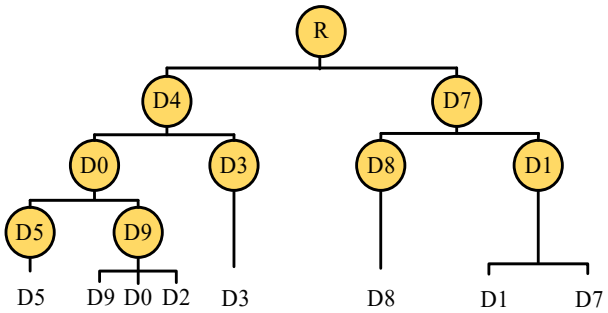


图 3-7 层级树示意图（K=2,M=3）

视觉词典构建词汇树会为每个视觉单词赋予频率-逆文档频率(TF-IDF)的权重占比。通过权重占比来实现对特征点的向量化以及实现较好的检索效果。权重公式，如式(3-15)所示。

$$\eta_i = \frac{n_{id}}{n_d} \lg \frac{N}{N_i} \tag{3-15}$$

式中 n_{id} 是视觉单词 i 在图像 d 出现的频次，n_d 是图像 a 视觉单词总数，N 是图像序列总数，N_i 是所有图像中 i 出现的频次。视觉单词在图像 a 中出现次数越多，则 *TF* 越高；视觉单词 i 在其他图像上出现次数少，则 *IDF* 越高，这使得视觉单词具有很好区分度，权重也就越高。视觉单词向量化，如式(3-16)所示。

$$A = \{(\omega_1, \eta_1), (\omega_2, \eta_2) \ldots (\omega_N, \eta_N)\} \overset{def}{=} V_A \tag{3-16}$$

式中 η_N 是 TF-IDF 权重，ω_N 是视觉单词，ω_N 是视觉单词 A 的向量化。

3.2.3 动态词典更新

设计动态词典更新策略，通过查询当前描述子是否匹配得到对应的描述子，根据在视觉词典中匹配的次数，与设定的查询次数阈值 S 进行比较来进行动态词典更新，以适应新的环境和图像特征。

如果查询匹配次数未超过阈值 S，则表示当前叶节点存储的视觉单词与新描述子存在相似但并未完全匹配。此时，系统认为新描述子与该叶节点中的已有描述子存在较小的匹配误差，在这种情况下，合并策略被触发。合并策略基于描述子之间的相似度，通过“位与”运算（AND）来识别和合并相似的描述子。“位与”运算确保了描述子的原始特征信息不被改变，同时也保证了词典的有效性和更新的精确性，如式（3-17）所示。

$$W_i^{t+1} = W_i^t \wedge A_q \tag{3-17}$$

式中 W_i^t 为 t 时刻匹配的单词，则在 $t+1$ 时刻进行单词更新。

若查询匹配次数超过阈值 S 时，系统认为新描述子与该叶节点中的已有描述子存在较大的误差，当前视觉单词在词典中尚未包含。此时，系统将通过递归重建机制来扩展词典。系统会重新构建当前叶节点，将新描述子添加到原有描述子集合中，并重新组织节点结构，以适应新的描述子。具体过程如下：

1、选择需要重建的叶节点。假设当前叶节点为 L_i，其中包含原始的描述子集合 $D_i = \{d_i\}$。如果该节点无法匹配当前单词 W_i^t，则进入重建过程。

2、扩展叶节点描述子集合。根据阈值判断和描述子匹配结果，系统将新的

描述子d_t添加到该叶节点的描述子集合$D_i^{'}$中：

3、重建叶节点结构。为了重新组织当前叶节点L_i的描述子集合$D_i^{'}$，需要重新分配这些描述子，并更新聚类中心。选择k个新的聚类中心，并随机选取初始聚类中心$C_1,C_2,\ldots\ldots,C_k^{'}$。对每个描述子$D_i \subset D_i^{'}$，计算其到所有聚类中心的欧几里距离，并将其分配给最近的聚类中心，如式（3-18）所示。

$$C_i = \arg\min_k \| D_i - C_k \| \tag{3-18}$$

式中是描述子D_i到聚类中心C_k的欧几里距离。然后根据新的描述子分配，重新计算聚类中心，聚类中心是该簇内所有描述子的均值，如式（3-19）所示。

$$C_k^{'} = \frac{1}{|C_k|}\sum_{D_i \in C_K} D_i \tag{3-19}$$

式中$|C_k|$是簇k中的描述子数量。

4、重新构建叶节点。完成聚类后，当前叶节点L_i会包含一个或多个新的聚类簇。如果聚类产生了多个簇，当前叶节点则被分割成多个子节点。当某个子节点中包含的描述子数量预设值$M(M > S)$时，停止递归操作。

当层级树中叶子节点的总数超过设定的阈值N时，算法将触发合并操作，通过评估叶子节点之间的相似度，将具有高度相似性的节点合并为一个新的节点，从而去除冗余并增强树的表达能力，提升树结构的区分能力。

3.3 公开数据集实验及结果分析

为验证本章算法优点和性能，设计以下实验：相似度评分实验、算法定位精度实验、准确率和召回率实验。本文实验所用平台软硬件配置：CPU 为 Inter i5-12400F（六核心处理器，十二线程，主频率 2.5GHz），内存 16GB，GPU 为 RTX3060，显存为 12GB，操作系统为 Ubuntu 18.04 LTS。

3.3.1 相似度对比实验

为验证本章算法在词袋模型中的改进性能，首先采用 LaMAR 数据集对视觉词典进行离线训练。随后，从 KITTI 数据集的 09 序列中选取部分图像，分别与 DBoW3 和 iBoW-LCD 算法进行词典构建与相似度评分的对比实验。DBoW3 算法是 DBoW2 的改进版本，通过多方面的改进和优化，提供了更强大和高效的图

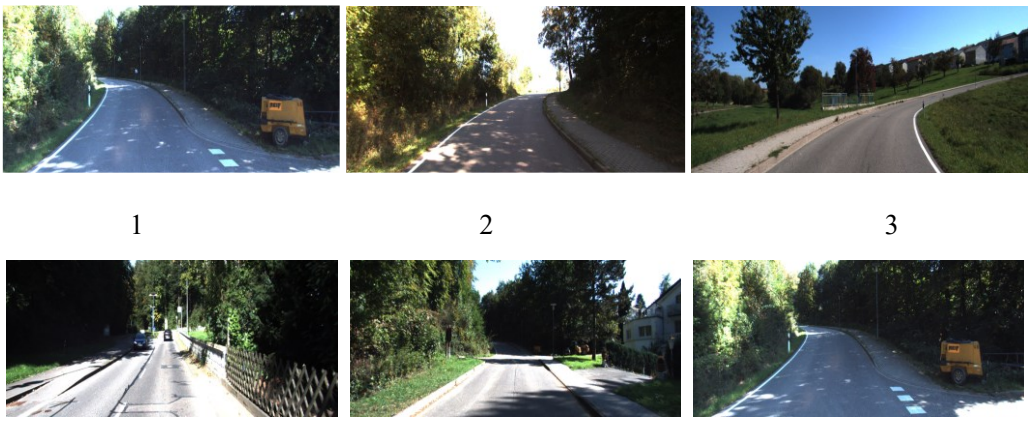
像描述和检索能力。iBoW-LCD 则采用在线增量式词典构建策略，在系统运行过程中动态生成并更新视觉单词，完全不依赖离线训练。LaMAR 数据集为混合数据集，具有室内外街道场景 35880 张图片，与 KITTI 数据集的城市道路驾驶场景不同，LaMAR 的户外部分主要为街道行人视角场景，包括人行道、建筑外立面和城市开放空间。

相同数据集下，特征点提取数为 1000，生成 K=10，L=6 规模的词典，DBoW3 算法、本章算法降维前和本章算法降维后构建词典单词数分别为 994834、923749 和 595639，词典构建规模与查询时间对比，如表 3-1 所示。对比结果表明，在相同特征点提取数量条件下，本章算法生成更紧凑的词典，降维后词汇量减少了 35.5%，仅为传统词袋模型的 59.9%。同时，在进行相似度查询时，本章算法的查询时间更短，显著提升了回环检测的计算效率。该性能优势得益于算法中引入的 PCA 降维和 HDBSCAN 聚类方法，通过提取描述子簇中心，有效压缩了词典规模，同时提升了图像信息的表达能力与检索效率。综上所述，本章算法在保持词袋模型结构与功能的前提下，以更小的词汇量实现更高效的回环检测查询性能。

表 3-1 词典构建规模与查询时间对比

	词典规模	查询时间
DBoW3 算法	994834	152ms
Ours 算法（PCA 降维前）	923749	143ms
Ours 算法（PCA 降维后）	595639	120ms

从 KITTI 数据集的 09 序列中选取部分图像进行相似度评分的对比实验。KITTI 数据集 09 序列一共包含 1590 帧灰度图像，提供了真实轨迹信息。其中，图像 1（0018 帧）与图像 6（1586 帧）形成回环，图像 2（0147 帧）、图像 3（0220 帧）、图像 4（1447 帧）和图像 5（1542 帧）均为干扰项，如图 3-8 所示。



1　　2　　3

4　　5　　6

图 3-8 相似度实验图片

采用 DBoW3 算法、iBoW-LCD 算法以及本章所提算法进行对比，评分结果分别如表 3-2、表 3-3 和表 3-4 所示，其中每组图像的最高相似度评分以加粗方式标注。从整体结果来看，三种算法在评分数值上存在较大差异。即使在 DBoW3 和 iBoW-LCD 中出现评分为 0 的情况，本章算法仍能输出有效的相似度评分，显示出更强的鲁棒性。本章算法之所以表现更优，主要得益于采用 ConvNeXt 孪生网络提取图像特征，并通过动态融合模块增强特征的判别能力与表征能力，从而使生成的视觉单词更具代表性，显著减小图像间的 L1 范数距离，提升相似度评分。例如，在图 1 与图像 6 的回环匹配中，DBoW3、iBoW-LCD 及本章算法的相似度评分分别为 0.01611、0.01772 和 0.02571，且均为对应图像的最高得分。实验结果表明，本章算法不仅构建了更紧凑的视觉词典，还显著增强了视觉单词的表征能力，有效提升了图像匹配和检索的准确性。

表 3-2 DBoW3 实验结果

图	1	2	3	4	5	6
1	1	**0.01217**	0.00558	0.00370	0.00187	**0.01611**
2	0.01217	1	0.00196	0	0	0.00334
3	0.00558	0.00196	1	0.00152	0.00524	0.00593
4	0.00370	0	0.00152	1	0.00754	0
5	0.00187	0	0.00524	**0.00754**	1	0.00785
6	**0.01611**	0.00334	**0.00593**	0	**0.00785**	1

表 3-3 iBoW-LCD 算法实验结果

图	1	2	3	4	5	6
1	1	**0.01339**	**0.00614**	0.00407	0.00206	**0.01772**
2	0.01339	1	0.00216	0	0	0.00367
3	0.00614	0.00216	1	0.00167	0.00576	0.00652
4	0.00407	0	0.00167	1	0.00829	0
5	0.00206	0	0.00576	**0.00829**	1	0.00864
6	**0.01772**	0.00367	0.00652	0	**0.00864**	1

表 3-4 Ours 实验结果

图	1	2	3	4	5	6
1	1	**0.01596**	**0.01950**	0.00791	0.01216	**0.02571**
2	0.01596	1	0.01573	0.00799	0.01380	0.01533
3	0.01950	0.01573	1	0.00768	0.01378	0.01160
4	0.00791	0.00799	0.00768	1	0.01217	0.00593
5	0.01216	0.01380	0.01378	**0.01217**	1	0.01402
6	**0.02571**	0.01533	0.01160	0.00593	**0.01402**	1

3.3.2 SLAM 系统定位精度对比实验

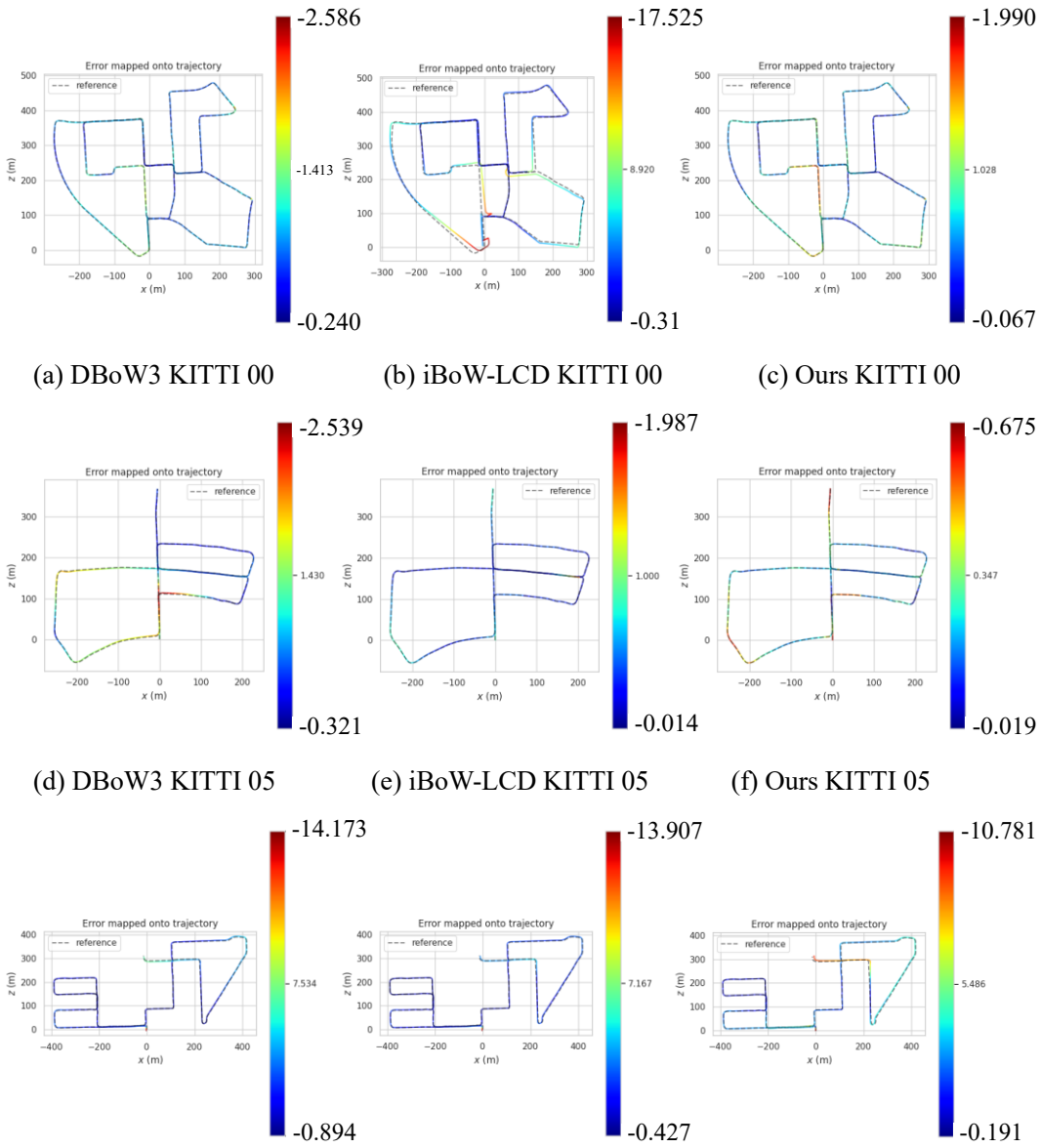
为了验证本章所提算法中动态词典更新的效果，保持其他参数配置不变的情况下，分别将与 DBoW3、iBoW-LCD 算法进行 SLAM 系统定位实验。本章算法与 DBoW3 算法的视觉词典为 3.3.1 节中 LAMAR 数据集训练的；本章算法和 DBoW3 算法与 ORB-SLAM3 系统结合，iBoW-LCD 算法与 OV2SLAM 进行结合，观察视觉词典更新对系统定位精度的影响。本节仅选取 KITTI 数据集中部分带有回环序列进行测试。KITTI 数据集由德国卡尔斯鲁厄理工学院和丰田美国技术研究院联合创办，是目前国际上最大的自动驾驶场景下的计算机视觉算法评测数据集。KITTI 数据集部分图像，如图 3-9 所示。

图 3-9 KITTI 数据集部分图像

本节使用 evo 测评工具评测本章算法与 DBoW3、iBoW-LCD 算法。评价指标采用绝对轨迹误差(absolute trajectory error，ATE)和相对位姿误差(relative pose error，RPE)中的均方根误差(root mean squared error，RMSE)来衡量定位精度。ATE 反映算法评估轨迹和真实轨迹的全局一致性和精度，而 RPE 反映了轨迹漂移程度；均方根误差(RMSE)数值越低代表系统鲁棒性越高。

KITTI 数据集轨迹对比，如图 3-10 所示；其中，黑色虚线表示真实轨迹，红色表示最大轨迹误差，蓝色表示最小轨迹误差。DBoW3 在 00、05、08 序列最大绝对轨迹误差分别为 2.586m、2.539m、14.173m；iBoW-LCD 在 00、05、08 序列最大绝对轨迹误差分别为 17.525m、1.987m、13.907m；本章算法在 00、05、08 序列最大绝对轨迹误差分别为 1.990m、0.675m、10.781m。整体来看，本章算法在所有序列中均实现了更低的定位误差，在 00 和 05 序列中显著优于对比算法。其中，iBoW-LCD 在 00 序列中由于词典在线更新过程中未能及时识别关键回环帧，导致累积误差放大；而 DBoW3 虽在多数序列中表现稳定，但固定词典结构在场景变化剧烈（如 08 序列）时适应性较差，误差也随之增加。相比之下，本

章算法通过引入基于孪生网络的结构特征提取模块与动态融合机制，显著增强了视觉单词的表达能力；同时，通过动态更新视觉词典，算法能够避免传统静态词典在大规模或新场景中容易出现的回环误判的问题。实验结果验证了本章算法在复杂场景下减少了 SLAM 系统定位精度误差，具有更高的准确性和鲁棒性。



(a) DBoW3 KITTI 00　(b) iBoW-LCD KITTI 00　(c) Ours KITTI 00

(d) DBoW3 KITTI 05　(e) iBoW-LCD KITTI 05　(f) Ours KITTI 05

(g) DBoW3 KITTI 08　(h) iBoW-LCD KITTI 08　(i) Ours KITTI 08

图 3-10 KITTI 数据集部分轨迹对比图

本章算法使用改进后的词袋模型训练视觉词典，训练数据集是独立于实验场景的 LaMAR 数据集。本章算法与 DBoW3 算法、iBoW-LCD 算法在 KITTI 数据集部分序列的定位精度对比结果，如表 3-5 所示。相较于 DBoW 和 iBoW-LCD，本章算法在 ATE 和 RPE 方面均有所降低，表明在回环检测和轨迹优化方面具有

一定优势。在 00 序列和 08 序列中，由于场景中存在较多回环信息，各算法均能有效地检测回环并进行轨迹修正，因此 ATE 误差相差不大，整体误差收敛较为接近。在 02 序列和 09 序列中，由于为公路场景，环境结构高度重复，导致不同帧之间的特征匹配容易发生混淆。受此影响，无论是本章算法，还是 DBoW3 和 iBoW-LCD，系统误差均相对较大。在 05 序列和 06 序列中，由于回环次数较少，词袋模型对场景的全局约束能力较弱，因此 ATE 和 RPE 误差的下降幅度更为明显。总体来看，本章算法相比较于 DBoW3 算法、iBoW-LCD 算法平均 ATE 降低 12.59％、9.34%，平均 RPE 降低 27.25％、19.83%。实验结果表明，本章算法通过动态更新词典的能力，能够更好地适应环境变化，提高回环检测的有效性，从而优化 SLAM 轨迹估计的精度。

表 3-5 SLAM 系统定位精度对比

KITTI 数据集	DBoW3 (RMSE/m)		iBoW-LCD (RMSE/m)		Ours (RMSE/m)	
	ATE	RPE	ATE	RPE	ATE	RPE
KITTI 00	3.808	0.027	3.671	0.027	3.217	0.024
KITTI 02	9.398	0.027	9.454	0.027	8.132	0.027
KITTI 05	1.992	0.044	1.795	0.035	1.461	0.025
KITTI 06	2.67	0.188	2.369	0.168	2.204	0.117
KITTI 07	1.687	0.017	1.497	0.015	1.347	0.015
KITTI 08	3.821	0.039	3.486	0.035	3.322	0.032
KITTI 09	11.425	0.036	11.283	0.036	10.738	0.035

3.3.3 准确率召回率实验

为了进一步评估算法整体性能，本章采用 KITTI 数据集部分带有回环的序列进行测试。准确率是算法提取的所有回环中确实是真实回环的概率，而召回率是指在所有真实回环中正确检测出来的概率。本章算法与 DBoW3 算法、iBoW-LCD 算法召回率对比，如表 3-6 所示。表 3-6 数据准确率为 100%时算法达到的最大召回率，即检测到的回环中全部为真实回环的情况下，算法能够找到的真实回环比例。实验结果表明，本章算法的回环检测召回率相较 DBoW3 提高 12.25%，相较 iBoW-LCD 提高 21.06%。在 KITTI 00 序列中，得益于本章提出的动态特征融合模块，增强了图像的局部细节感知与空间结构表达能力，使得回环检测更加

精准；相比之下，iBoW-LCD 虽具有一定的增量词典构建能力，但在视觉单词表征能力仍存在不足，导致在该序列中的召回率明显低于本章算法。在 KITTI 09 序列中，iBoW-LCD 在动态场景中的鲁棒性有限，易受到误匹配影响；而 DBoW3 由于使用静态词典，更难适应环境变化，召回率均低于本章算法。在 KITTI 02 和 07 序列中，本章算法整体性能仍优于 iBoW-LCD 和 DBoW3。综上所述，本章算法在多数典型场景中均表现出较高的鲁棒性与识别能力，验证了所提方法在实际 SLAM 回环检测应用中的优势。

表 3-6 本章算法与 DBoW2 算法准确率对比

数据集	召回率		
	DBoW3	iBoW-LCD	Ours
KITTI00	0.734	0.691	0.843
KITTI02	0.098	0.071	0.108
KITTI05	0.560	0.449	0.572
KITTI06	0.256	0.268	0.277
KITTI07	0.168	0.231	0.236
KITTI08	0.865	0.656	0.806
KITTI09	0.038	0.155	0.210

3.4 本章小结

本章围绕回环检测中视觉词袋模型存在的表征能力不足和缺乏更新机制的问题，提出一种基于 ConvNeXt 孪生网络的动态视觉词典更新算法。为增强视觉单词的表征能力，通过引入 ConvNeXt 作为共享主干的孪生网络，并融合动态特征融合模块，有效提升了对图像细节和结构的感知能力，从而增强了视觉单词的判别表达。然后，算法引入 PCA 降维与 HDBSCAN 聚类，压缩特征维度并剔除噪声数据，构建结构紧凑、鲁棒性强的初始视觉词典。最后，针对视觉词典缺乏更新机制的问题，在离线视觉词典的基础上，结合层级树检索与节点合并策略，实现视觉词典的动态更新。实验结果表明：本章算法使视觉 SLAM 定位精度 ATE RMSE 相较于 DBoW3、iBoW-LCD 平均降低 12.59％、9.34%，回环检测召回率相较于 DBoW3、iBoW-LCD 平均提升 3.98%、19.11%。验证了本章算法在视觉 SLAM 回环检测和定位精度的显著优势。

第 4 章 基于结构相似性评估的回环检测算法

针对传统视觉 SLAM 回环检测算法仅关注图像中的局部特征，从而导致回环检测召回率低以及系统定位误差较大的问题，提出一种基于结构相似性评估的视觉 SLAM 回环检测改进算法。首先，由关键帧公共单词筛选得到待检测关键帧，分别计算待检测关键帧对应共视关键帧组的局部特征评分和结构相似性评分；其次，通过动态权重获得融合后的相似度评分，并比较共视关键帧组的相似度总得分（S 是共视关键帧组相似度总得分，$S_{\min}$ 是由实验得到的阈值），只取得分超过阈值的组中分数最高的关键帧作为回环候选帧。最后，根据相对误差判断结果（S_1 和 S_2 分别是相邻帧的相似度评分），决定是否对回环候选帧进行候选帧剔除，再根据历史回环结果对错误候选帧进行剔除，并通过连续性检测获得回环结果。基于结构相似性评估的回环检测算法框架，如图 4-1 所示。

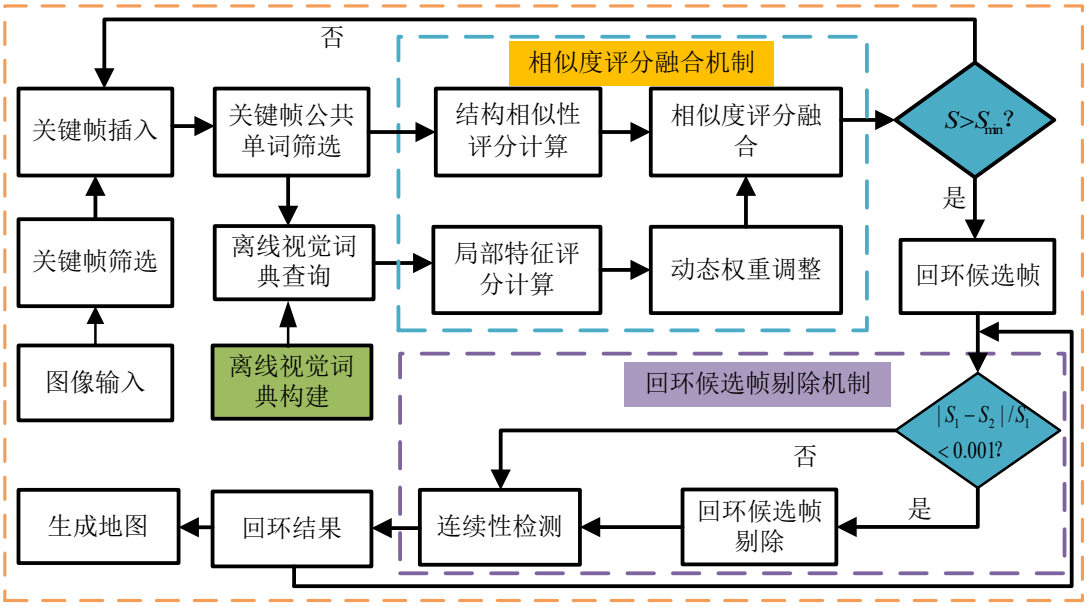


图 4-1 基于结构相似性评估的回环检测算法框架

4.1 相似度评分融合

基于词袋模型的回环检测方法在视觉 SLAM 中广泛应用，主要依赖图像的局部特征来进行图像匹配，例如常用的 ORB 描述子。该方法在捕捉图像中的关键点和特征信息方面具有较高的效率和精度，因此可以有效地进行图像匹配和回

环检测。然而由于局部特征匹配主要关注特征点的局部信息，忽略图像的整体结构特征，这使得回环检测在某些情况下无法充分捕捉图像的全局变化，从而影响检测结果的准确性，导致回环误判或漏检的情况。为了解决这一问题，提出一种相似度评分融合算法，通过融合图像的全局结构信息与局部特征信息，克服仅依赖局部特征匹配的不足。该算法引入结构相似性评估算法，旨在通过综合考虑图像的整体结构和局部特征，进一步提高回环检测的精度和鲁棒性。

4.1.1 局部特征评分

基于词袋模型的回环检测方法，通过对图像局部特征的评分来实现图像匹配，从而有效地进行回环检测。图像中的每个局部特征描述子都会与词典中预先定义的视觉单词进行匹配，通过计算描述子与词典中每个视觉单词之间的距离来确定最匹配的视觉单词。通过这种方式，图像中的每个局部特征描述子都会映射到词典中最相似的视觉单词上，最终实现图像之间的相似度计算。局部特征 ORB 特征点提取，如图 4-2 所示。

图 4-2 局部特征 ORB 特征点提取图

局部特征评分的计算使得图像的局部特征描述子被转化为一个视觉单词序列。每个图像的局部特征通过与词典中视觉单词的匹配关系，得到一个低维的向量表示。当多个图像的局部特征与词典中的视觉单词匹配时，可以通过统计匹配的频率和分布，计算图像之间的相似度。词袋模型通过考虑每个视觉单词在不同图像中的出现频率（TF）以及其在所有图像中的逆文档频率（IDF），为每个视觉单词赋予一个权重。这一权重能够反映视觉单词的代表性和区分度，进而在回环检测中提高匹配的准确性和稳定性。视觉单词量化公式，如式（4-1）所示。

$$A=\{(\omega_1,\eta_1),(\omega_2,\eta_2)\ldots(\omega_N,\eta_N)\}\overset{def}{=}V_A \tag{4-1}$$

式中，ω_n是视觉单词，η_N是权重，A代表图像，V_A是图像的视觉单词向量化表示。查询到图像的视觉单词向量后，通过视觉词典可以用向量V_A描述图像A。为了比较两幅图像的局部特征评分，采用相似度计算方法来衡量两个视觉单词向量的相似程度。计算公式，如式（4-2）所示。

$$s\left(v_A - v_B\right) = 2\sum_{i=1}^{N} |v_{Ai}| + |v_{Bi}| - |v_{Ai} - v_{Bi}| \tag{4-2}$$

式中，v_A和v_B分别是图像A和图像B的向量化表示，v_{Ai}和v_{Bi}图像A和B在第i个视觉单词上的频率或权重。

4.1.2 全局结构相似性评分

在视觉 SLAM 回环检测中，传统词袋模型方法主要依赖基于词袋模型的特征匹配进行相似性计算，容易受到视觉词典量化误差的影响，局部特征点在处理纹理相似但结构不同的场景时存在误匹配的问题。因此，引入结构相似性评估算法，可以更有效地衡量图像间的相似性结构；相似性评估算法基于图像的亮度、对比度和结构信息，能够更符合人眼感知，对光照变化和噪声具有更强的鲁棒性；结构相似性评估能够更精确地度量两幅图像在结构上的相似性，而不仅仅依赖于局部特征之间的差异。为进一步提升结构相似性评估算法在回环检测中的性能，本章提出一种基于 Sigmoid 变换、高斯模糊和空间权重的改进算法，使算法的对图像结构和特征的表达能力得到了显著增强，并且更好地处理图像中的全局结构和细节。

Sigmoid变换作为一种常见的非线性变换方法，通过对图像的亮度和对比度进行调节，能够增强图像中的细节，特别是在低对比度或光照不均的情况下。这种变换可以使得图像中的低对比度区域得到更好的突出，从而提高结构相似性评估的精度。在结构相似性评估算法中引入Sigmoid变换后，图像的亮度和对比度不均匀的区域能够得到更为均衡的感知，增强算法对于不同图像区域之间细微差异的敏感度。设有图像X和Y，像素值为$x(i,j)$、$y(i,j)$，图像大小为$M \times N$。Sigmoid变换公式，如式（4-3）所示。

$$S\left[x(i,j)\right] = \frac{1}{1+e^{-x(i,j)}} \tag{4-3}$$

式中 $x(i,j)$ 为原像素值，$S\left[x(i,j)\right]$ 为变换后的值。

高斯模糊能够对图像进行平滑处理，消除一些细节噪声和小尺度的局部差异。通过引入高斯模糊，结构相似性评估算法对图像中由于噪声或微小扰动引起的局部特征变化的敏感性得到抑制，从而减少这些噪声干扰对回环检测结果的影响。此外，高斯模糊还能够使得图像的全局结构更加平滑，有助于评估图像的宏观相似度，从而提升回环检测在复杂环境下的稳定性和准确性。对图像Sigmoid变换后进行高斯模糊，从而求得像素值亮度均值、方差和协方差。高斯模糊函数，如式（4-4）所示。

$$G(i,j)=\frac{1}{2\pi\sigma^2}\exp(-\frac{i^2+j^2}{2\sigma^2}) \tag{4-4}$$

式中 $G(i,j)$ 是高斯核在图像中坐标 (i,j) 的值，σ 为高斯分布的标准差。图像进行高斯模糊的计算过程是将图像中每个像素与高斯核进行卷积，如式（4-5）所示。

$$\bar{x}(i,j)=\sum_{i=1}^{M}\sum_{j=1}^{N}S[x(i,j)]\bullet G(i,j) \tag{4-5}$$

传统的结构相似性评估算法在计算相似性时，所有区域的权重是均等的，这导致算法在图像的某些重要区域（如关键物体或特征区域）对比度较低时，无法充分反映这些区域的结构信息。为此，本章通过空间权重，分析图像中每个区域的重要性，将较高的权重赋予那些关键结构区域，而对不重要区域给予较低的权重。这种权重分配机制使得算法更加关注图像中的关键结构信息，从而提高回环检测算法对不同区域的敏感性。将亮度、对比度和结构相似性的加权乘积得到最终的结构相似性评估值。空间权重计算公式，如式（4-6）所示。

$$\omega(i,j)=\exp(-\tfrac{i^2+y^2}{2\sigma^2}) \tag{4-6}$$

式中 σ 为空间权重的标准差。最终的结构相似性评估值，如式（4-7）所示。

$$S_s=\frac{1}{N}\left\{\sum_{i=1}^{N}\left(\frac{2\bar{\mu}_X\bar{\mu}_Y+C_1}{\bar{\mu}_X^2+\bar{\mu}_Y^2+C_1}\cdot\frac{2\bar{\sigma}_{XY}+C_2}{\bar{\sigma}_X^2+\bar{\sigma}_Y^2+C_2}\cdot\omega_i\right)\right\} \tag{4-7}$$

式中，N 是整张图片划分窗口数量，$\bar{\mu}_X$ 和 $\bar{\mu}_Y$ 分别是图像 X 和 Y 在窗口内的局部均值，$\bar{\sigma}_X^2$ 和 $\bar{\sigma}_Y^2$ 分别是图像 X 和 Y 在窗口内的局部方差，$\bar{\sigma}_{XY}$ 是图像 X 和 Y 在窗口内的局部协方差，C_1 和 C_2 是稳定性常数，是位于 i 窗口处的空间权重。

4.1.3 动态权重

由于不同场景下的图像会表现出不同特点，使用相同的权重导致算法对图像相似度评估不准确。因此采用动态权重调整策略，依据局部特征评分变化趋势调整权重参数。如果评分变化趋势表明图像局部特征占比较大，则通过增加局部特征的权重以更准确地捕捉图像的细节信息；相反，则通过增加全局结构的权重以更好地捕捉图像的结构信息。具体步骤如下：1）设置初始权重α_t为0.5。2）设置大小为10帧的图像数据收集窗口，记录每幅图像的局部特征评分，并计算窗口内图像局部特征评分的平均值(β_t)。3）根据窗口间平均值大小判断评分变化趋势，如果β_t大于β_{t-1}，增加权重；反之，减小权重；当β_t与β_{t-1}之间的误差小于0.001时，则保持权重不变。计算公式，如式（4-8）所示。

$$\alpha_t=\begin{cases}\alpha_{t-1}+\delta, & 当\ \beta_t>\beta_{t-1}\\ a_{t-1}-\delta, & 当\ \beta_t<\beta_{t-1}\\ a_{t-1}, & 当\ |\beta_t-\beta_{t-1}|<0.001\end{cases} \tag{4-8}$$

式中，α_t是更新后的权重值，α_{t-1}是上一次调整后的权重值，β_{t-1}是上一时刻窗口的平均值，δ是调整步长，由实验取值为0.025。

为了更准确评估图像相似度，减少回环误判，将调整后的权重用于评分融合，以实现全局图像结构与局部图像特征的相似度评分互补。融合相似度评分计算公式，如（4-9）所示。

$$S=\alpha_t\cdot V_s+(1-\alpha_t)\cdot S_s \tag{4-9}$$

式中S是融合后的图像相似度评分，α_t是权重值，V_s是图像的局部特征评分，S_s是图像的结构相似性评分。

4.2 回环候选帧剔除

回环检测需要将当前图像与所有历史图像进行逐帧比较，在面对大规模场景时，随着场景图像帧数的增加，相似度评分存在过拟合，而导致候选帧出现误判断，进而产生错误的回环，严重影响整个地图的准确性；因此，设计回环候选帧剔除机制来应对这一问题。大规模场景发生动态变化的几率增加，但一些稳定的环境特征仍会持续存在，保留在历史回环结果中。所以，可以利用历史回环结果，

根据场景的变化情况来调整阈值，适应不同环境下的场景变化，从而有效的剔除错误候选帧。

回环候选帧剔除机制通过计算两幅图像相似度评分的相对误差来判断图像间是否过拟合。设定相对误差阈值为 0.01，即当两幅图像相似度评分的相对误差不超过 1%时，存在错误候选帧，反之则不存在。如果不存在错误候选帧，则跳过回环候选帧剔除。相对误差计算，如式（4-10）所示：

$$|(S_1 - S_2)/S_1| < 0.01 \tag{4-10}$$

式中，S_1 和 S_2 分别是相邻的回环候选帧相似度评分，误差阈值 0.01 由实验得到。

由于哈希差异度能够准确反映图像间细微的差异，因此采用哈希差异度剔除错误候选帧。算法使用拉普拉斯金字塔对图像进行多尺度分解，分别计算每个尺度的哈希值，然后对每个尺度的哈希值进行加权计算，从而得到多尺度的哈希值 H_{Dhash}。哈希值计算，如式（4-11）所示。

$$H_{Dhash} = \frac{1}{n}\left[\sum_{i=1}^{n} hash(I_{scale_i}) \times \omega_i\right] \tag{4-11}$$

式中，I_{scale_i} 是第 i 层，$hash()$ 是差值哈希计算函数，计算相邻像素灰度值差异，ω_i 是权重。哈希差异度计算公式，如公式（4-12）所示。

$$H_D = 1 - d(H_X, H_Y)/p \tag{4-12}$$

式中，$d(H_X, H_Y)$ 是图像间哈希值的汉明距离，p 是哈希值的位数，H_D 是当前回环候选帧的哈希差异度。具体流程如下：首先，根据相对误差判断结果决定是否执行候选帧剔除操作；当误差小于设定阈值时，判定为存在错误回环，执行错误候选帧剔除；否则，直接进入连续性检测阶段。随后，分别计算当前回环候选帧与历史回环结果之间的哈希差异度，并根据历史回环次数判断是否已发生过回环：若未发生回环，则设定固定阈值为 0.6；若已发生回环，则设定动态阈值，该阈值为历史回环结果中最大哈希差异度的 0.8 倍。最后，根据该阈值对候选帧进行筛选，剔除误匹配帧，保留更为准确的候选帧用于后续回环确认。回环候选帧剔除方法，如算法 4-1 所示。

算法 4-1 回环候选帧剔除

输入：回环候选帧集合，历史回环结果
输出：筛选后的回环候选帧集合
参数：阈值 threshold，哈希差异度
1：**IF** $|(S_1 - S_2)/S_1| < 0.01$：
2：　**FOR** i in 回环候选帧集[ ]:
3：　使用公式 8 计算回环候选帧和历史回环结果哈希差异度
4：　**IF** 还未发生回环
5：　　初始阈值设置为 0.6
6：　　**IF** 回环候选帧哈希差异度> threshold:
7：　　剔除错误候选帧
8：　　**Return** 筛选后的回环候选帧集合
9：　　**End IF**
10：　**ELSE**:
11：　　比较历史回环结果哈希差异度
12：　　动态阈值设置为历史回环结果中哈希差异度最大值的 0.8 倍
13：　　**IF** 回环候选帧哈希差异度> threshold:
14：　　剔除错误候选帧
15：　　**Return** 筛选后的回环候选帧集合
16：　　**End IF**
17：　**End IF**
18：　**End FOR**
19：**ELSE:**
20：　直接进行连续性检测
21：**End IF**

4.3 公开数据集实验及结果分析

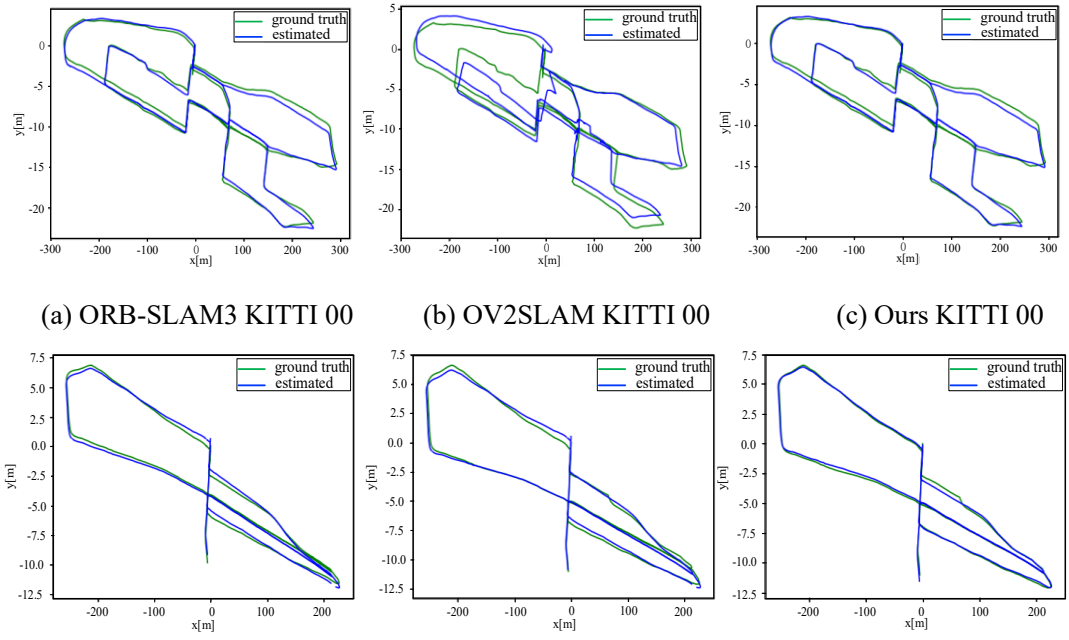
本章实验所用平台软硬件配置与 3.3 节公共数据集实验一致。为验证本章算法有效性，分别设计以下实验：词袋模型相似度评分实验，结构化场景定位结果与分析，非结构化场景定位结果与分析，准确率召回率实验，实时性实验。

4.3.1 结构化场景定位结果与分析

结构化场景通常包含明确的对象、组织结构和相对可预测的视觉模式。为了进一步验证本章所提算法整体性能，选取公共数据集 KITTI 中 00 序列~09 序列作为测试数据集。KITTI 数据集由德国卡尔斯鲁厄理工学院和丰田美国技术研究院联合创办，包含市区、乡村和高速公路等场景采集的真实图像数据，以上场景符合结构化场景特征。

本章使用 evo 测评工具评测算法性能，比较 ORB-SLAM3、OV2SLAM 与本章算法在结构化场景下 SLAM 系统定位误差大小，评价指标采用 3.3.2 节中介绍的绝对轨迹误差（ATE)和相对位姿误差(RPE)中的均方根误差(RMSE)来衡量定位精度。ORB-SLAM3 是基于传统 ORB 特征的视觉 SLAM 算法，使用 DBoW3 离线训练字典，通过匹配当前帧和历史关键帧之间的特征点，识别相似场景进行回环检测。OV2SLAM 是基于 LK 光流的视觉 SLAM 算法，回环检测与 ORB-SLAM3 基本一致，不同的是使用 iBoW-LCD 在线构建字典。

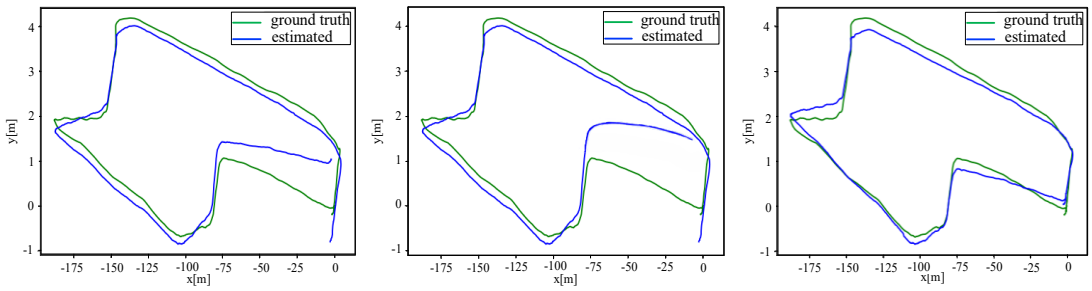
KITTI 数据集部分序列轨迹对比，如图 4-3 所示；其中，绿色线条是真实轨迹；蓝色线条是评估轨迹，即算法轨迹。图 4-3(a)-(c)中，00 序列中存在多处回环，本章算法与 ORB 算法估计误差基本一致，部分轨迹 ORB-SLAM3 误差更大；OV2SLAM 算法由于在线构建词典，视觉单词辨识度较弱，导致系统轨迹误差较大。图 4-3(d)-(f)中，由于 ORB-SLAM3 和 OV2SLAM 仅比较图像局部特征评分，导致轨迹在回环处误差较大。图 4-3(g)-(i)中，ORB-SLAM3 和 OV2SLAM 出现了无法回环的情况，而本章算法实现了回环。整体来看，ORB-SLAM3 和 OV2SLAM 在回环检测中受限于局部特征和词典构建的不足，导致误差较大和回环失败。相比之下，本章算法通过全局结构与局部特征融合及回环候选帧剔除机制，提高了回环检测的精度与鲁棒性，确保了算法轨迹与真实轨迹的高吻合度。



(a) ORB-SLAM3 KITTI 00　(b) OV2SLAM KITTI 00　(c) Ours KITTI 00

(d) ORB-SLAM3 KITTI 0　(e) OV2SLAM KITTI 05　(f) Ours KITTI 05

图 4-3 KITTI 数据集部分序列轨迹对比图



(g) ORB-SLAM3 KITTI 07　　(h) OV2SLAM KITTI 07　　(i) Ours KITTI 07

图 4-3 KITTI 数据集部分序列轨迹对比图（续）

KITTI 数据集 SLAM 系统定位精度统计对比，如表 4-1 所示。ORB-SLAM3 在 00、03、04、05、06、07 和 08 序列系统定位精度较好；其中 01、02 序列为室外公路场景，由于仅依靠局部特征评分，导致系统定位精度较差；09 序列是室外场景的回环，ORB-SLAM3 未检测到回环导致系统误差较大。OV2SLAM 在线构建视觉词典，由于视觉单词不够准确，导致在 00、02、08、09 序列系统定位误差较大；在 03、04、05、06 和 07 序列存在回环较少的简单场景中，系统定位精度较好。本章算法与其他算法相比，其中 00、05、06、07、08 序列为城市道路场景，包含较多建筑物，建筑物与道路能够形成清晰的结构化场景，SLAM 系统定位精度误差最小；01、02、03、04、09 序列为室外公路场景，虽然建筑物较少，存在较多树木，但是存在较为清晰的结构场景，与其他算法相比 SLAM 系统精度较小。由表 4-1 可知，本章算法与 ORB-SLAM3、OV2SLAM 相比，ATE RMSE 分别平均减少 19.12%、28.10%，RPE RMSE 分别平均减少 13.57%、31.64%。实验结果表明，本章算法能够减少 SLAM 系统定位精度误差，具有更高的准确性和鲁棒性。

表 4-1 KITTI 数据集 SLAM 系统定位精度对比

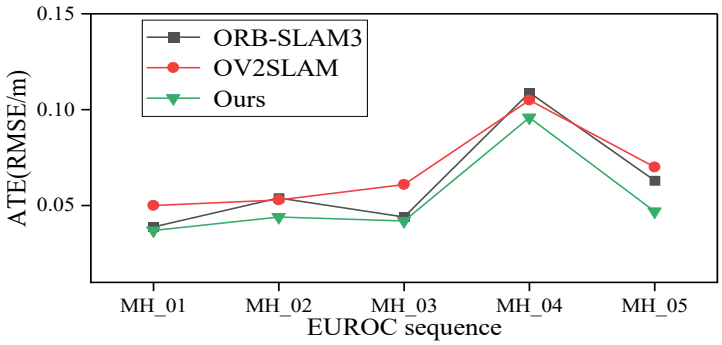
数据集	ORB-SLAM3 (RMSE/m)		OV2SLAM (RMSE/m)		Ours (RMSE/m)	
	ATE	RPE	ATE	RPE	ATE	RPE
KITTI 00	1.309	0.028	2.061	0.044	1.253	0.027
KITTI 01	13.647	0.050	14.761	0.053	12.853	0.049
KITTI 02	6.294	0.028	6.771	0.027	5.757	0.026
KITTI 03	0.658	0.016	1.387	0.016	0.532	0.016
KITTI 04	2.258	0.037	1.283	0.017	0.171	0.018
KITTI 05	0.871	0.017	2.323	0.031	0.815	0.015
KITTI 06	0.888	0.020	2.165	0.026	0.708	0.017
KITTI 07	1.606	0.016	2.166	0.016	0.543	0.014
KITTI 08	3.811	0.039	3.620	0.085	3.498	0.038
KITTI 09	6.127	0.029	5.611	0.039	4.176	0.022

4.3.2 非结构化场景定位结果与分析

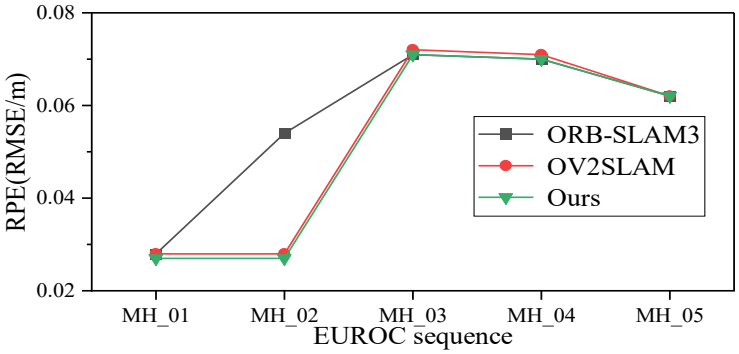
与结构化场景相对，非结构化场景中的图像内容更加复杂多样，为验证本章所提算法在非结构场景中的性能，选取 EUROC 数据集中的 MH01~MH05 序列作为测试数据集。EUROC 数据集是一个用于评估视觉惯性导航（VINS）算法的性能的数据集，是由瑞士苏黎世联邦理工学院制作的室内无人机双目视觉惯性数据集，包含室内和工厂的场景，其场景内包含较多复杂而随机的元素，且缺乏明显的层次结构，符合非结构化场景特征；EUROC 数据集部分图像，如图 4-4 所示。本章使用 evo 测评工具评测算法性能，评价指标采用 3.3.2 节中介绍的绝对轨迹误差(ATE)和相对位姿误差(RPE)中的均方根误差(RMSE)。

图 4-4 EUROC 数据集部分图像

在 EUROC 数据集上的绝对轨迹误差（ATE）与相对位姿误差（RPE）对比结果，如图 4-5 所示。由于 ORB-SLAM3 与 OV2SLAM 均仅依赖图像局部特征评分，难以充分利用图像的全局信息，导致 ATE RMSE 普遍偏高；尽管 OV2SLAM 与本章算法在 RPE RMSE 方面差距相对较小，但在整体定位精度上仍存在显著差异。本章提出的算法在特征匹配阶段引入了动态权重调整策略，根据图像内容动态调整结构相似性评分与局部特征评分之间的权重比例，从而在保持图像整体结构约束的同时，充分发挥局部特征在精细配准与姿态估计中的作用。在非结构化场景下尤为有效，能够增强系统对结构缺失、纹理重复等复杂环境的鲁棒性。实验结果显示，在 EUROC 数据集 MH01-MH05 序列中，本章算法在 ATE 和 RPE 两个误差指标上均取得最优表现，显著优于对比方法。与 ORB-SLAM3 和 OV2SLAM 相比，本章算法的 ATE RMSE 平均降低 13.92%、21.53%，RPE RMSE 平均降低 9.82%、1.53%。验证了本章算法不仅在典型结构化场景中具备高精度定位能力，在非结构化环境下亦能有效抑制累积误差，保证系统稳定性与精度。



（a）ATE RMSE



（b）RPE RMSE

图 4-5 EUROC 数据集 ATE 和 RPE 对比

4.3.3 准确率召回率实验

为了进一步评估本章所提算法整体性能，使用测试数据集 KITTI 数据集和 EUROC 数据集进行准确率召回率实验。在回环检测任务中，准确率是算法提取的所有回环中确实是真实回环的概率，而召回率是指在所有真实回环中正确检测出来的概率。虽然大多数系统将准确率作为优化指标，但是错误的回环检测会对视觉 SLAM 的稳定性造成不可挽回的损失。因此，本章更关注准确率为 100%时，算法达到的最大召回率；测试数据集回环检测召回率对比，如表 4-2 所示。由表 4-2 可知，在结构化场景 KITTI 数据集中，00 序列、06 序列、07 序列和 09 序列存在明显的结构化场景，召回率提升较大；02 序列存在较多相似场景（公路两旁都是树），导致召回率较低。在非结构化场景 EUROC 数据集中，只有 MH03-MH05 序列存在回环，数据集中工厂室内场景范围较小，存在回环次数较少，召回率都较高，本章算法与 ORB-SLAM3、OV2SLAM 相比召回率依旧有所提升。实验结果表明，本章算法与 ORB-SLAM3、OV2SLAM 相比回环检测召回率有一定提升，召回率分别平均提升 18.49%、26.73%。

表 4-2 测试数据集回环检测召回率对比

数据集	召回率		
	ORB-SLAM3	OV2SLAM	Ours
KITTI00	0.942	0.512	0.960
KITTI02	0.091	0.068	0.108
KITTI05	0.546	0.553	0.572
KITTI06	0.250	0.115	0.507
KITTI07	0.189	0.368	0.622
KITTI08	0.886	0.803	0.906
KITTI09	0.065	0.201	0.310
MH 03	0.903	0.893	0.909
MH 04	0.898	0.890	0.908
MH 05	0.892	0.891	0.907

4.3.4 实时性实验

为了评估本章所提回环检测算法的实时性，在 KITTI 数据集中的 00~09 序列上进行实时性对比测试。表 4-3 为回环检测算法耗时的具体数据，图 4-6 展示了回环检测的耗时与图像数量之间的关系。实验结果表明，随着图像数量的增加，回环检测的耗时显著上升，因为当前图像都需要与之前所有的图像进行相似度对比，检测是否发生回环。由图 4-6 可见，本章算法在实时性上表现出了显著优势，随着图像数量的增加，回环检测的时间增长较为缓慢，实时性得到了显著提升。这种性能提升主要得益于回环候选帧剔除机制能够有效去除错误候选帧，减少不必要的计算，从而缩短回环验证的时间。因此，本章算法不仅在回环检测精度上表现出色，在实时性方面也取得了优异的平衡，能够满足实时处理的需求。

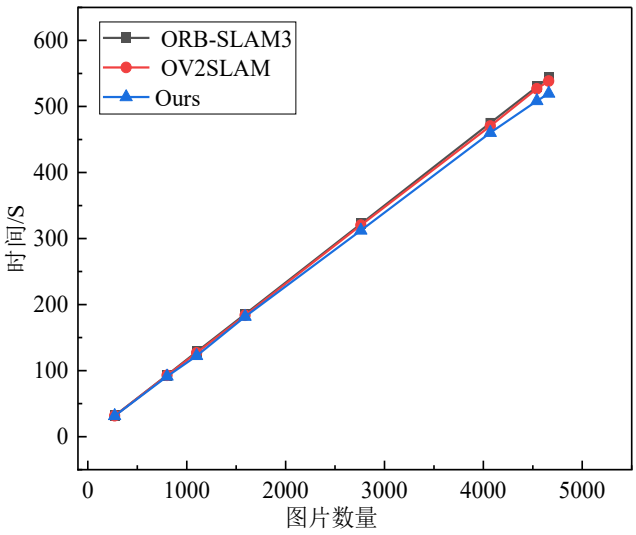


图 4-6 算法耗时与图像数量关系

表 4- 3 实时性对比

数据集标号	图片数量	耗时/s		
		ORB-SLAM3	OV2SLAM	Ours
04	271	31.63	31.59	31.41
03	801	93.29	92.87	91.18
01,06,07	1100	128.39	126.83	122.37
09	1590	185.58	183.72	181.78
05	2761	322.25	320.19	312.36
08	4070	475.03	470.31	460.17
00	4541	529.86	526.73	508.47
02	4661	544.01	538.52	519.72

4.4 真实场景实验

4.4.1 硬件实验平台

为了验证本文算法在真实场景中的可行性，使用移动机器人 Husky 上面的 RealSense D435i 采集真实场景进行实验，如图 4-7 所示。实验平台硬件配置为：CPU 是 I7-10875H 处理器，GPU 是 GTX1080 显卡，运行内存为 32G，操作系统为 Ubuntu 18.04，主要参数如表 4-4 所示。通过该实验平台采集真实场景数据对本文算法进行真实场景实验。

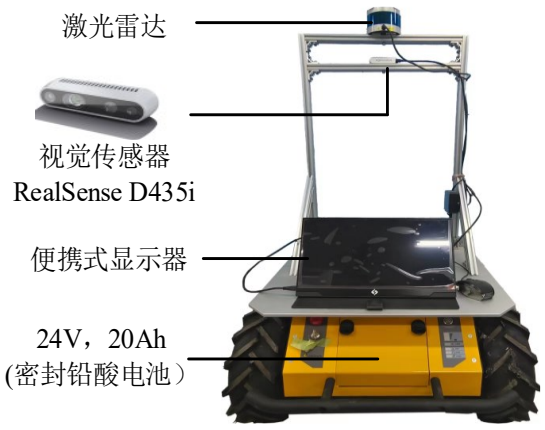


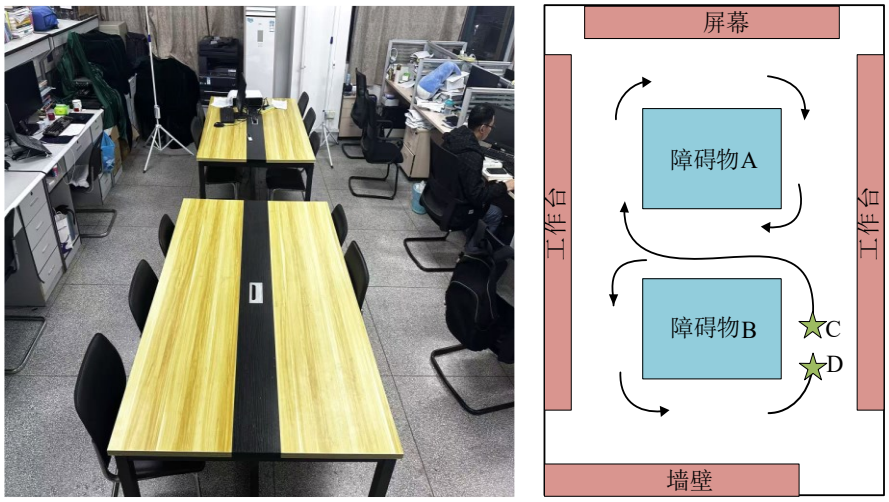
图 4-7 移动机器人和相机传感器

表 4- 4 移动机器人主要参数

变量名称	参数	数值
运行速度	v	$1m/s$
相机采样频率	h	$30fps$
方向变化范围	θ	$0\sim\pi$
尺寸	$L\times H\times W$	$990mm\times670mm\times390mm$

4.4.2 室内真实场景实验

为验证本文提出基于词袋模型的视觉 SLAM 回环检测改进算法可行性，采用室内结构化真实场景进行实验。移动机器人 Husky 通过 RealSense D435i 深度相机采集数据集，构建具有回环的室内真实场景。在实验中机器人围绕障碍物进行 8 字回环。C 点为起始点，激动机器人按照箭头指示的路径运行，最终到 D 点形成回环，真实场景图片和场景布局图如图 4-8 所示。使用本章算法与 ORB-SLAM3、OV2SLAM 算法进行分析对比，实验中，本文算法视觉词典为独立于室内场景的数据集训练的，ORB-SLAM3 算法为传统词袋模型算法进行回环检测，OV2SLAM 算法为在线词袋算法进行回环检测。



(a)真实场景图片　　(b)真实场景布局图

图 4-8 室内真实场景与布局图

真实场景不同算法轨迹对比图，如图 4-9 所示。ORB-SLAM3 算法在回到 C 点过程中由于场景特征稀疏导致出现了假回环，对算法轨迹修正错误，与真实轨迹出现较大误差。OV2SLAM 则是由于场景太小，无法提取到太多视觉单词，导致无法形成回环。本文算法在词袋模型中通过更新视觉词典和回环检测过程加入结构相似性评估，使得系统能够识别正确回环，并修正轨迹，构建全局一致的地图。本文算法相比于 ORB-SLAM3、OV2SLAM 的 ATE RMSE 降低了 41.37%、85.09%，RPE RMSE 降低了 47.67%、64.51%，如表 4-5 所示。

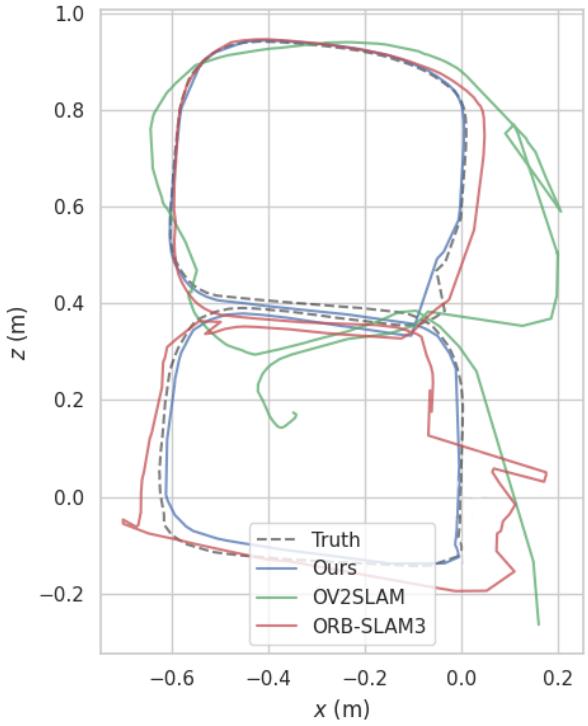


图 4-9 室内真实场景不同算法轨迹对比图

表 4-5 室内真实场景轨迹误差精度对比（单位：m）

算法	绝对轨迹误差(ATE)	相对位姿误差（RPE）
ORB-SLAM3	0.029	0.021
OV2SLAM	0.114	0.031
本文算法	0.017	0.011

4.4.3 室外真实场景实验

为验证本文提出基于词袋模型的视觉 SLAM 回环检测改进算法可行性，采用室外真实场景实验。移动机器人 Husky 通过 RealSense D435i 深度相机采集数据集，构建具有回环的室外真实场景。该场景为室外学校公共道路场景，主要包括圆形花坛与方形绿化带。圆形花坛为中存在较多车辆以及有行人经过，导致场景较为复杂；方形绿化带则存在较多结构化场景，对回环检测存在影响。真实场景平面图，如图 4-10 所示。采用 A-B-C-D-A 的红色路径对算法进行实验，实验对比算法采用与 4.4.2 节中的相同算法，其中本文算法视觉词典为独立于室外场景的数据集训练的。

图 4-10 室外真实场景平面图

不同算法轨迹对比，如图 4-11 所示。ORB-SLAM3 算法由于场景只计算局部特征，导致在花坛处出现轨迹错误，且由于场景中相似场景较多，导致出现了回环误判，对算法轨迹修正错误，在回环处 A 点与真实轨迹出现较大误差。OV2SLAM 则是提取到太多相似视觉单词，无法形成回环，导致轨迹错误。本文算法在词袋模型中通过回环检测过程加入结构相似性评估以及回环候选帧剔除，使得系统能够识别相似场景，检测出正确回环，并修正轨迹，构建全局一致的地图。本文算法相比于 ORB-SLAM3、OV2SLAM 的 ATE RMSE 降低了 31.40%、41.55%，RPE RMSE 降低了 33.33%、43.94%，如表 4-6 所示。

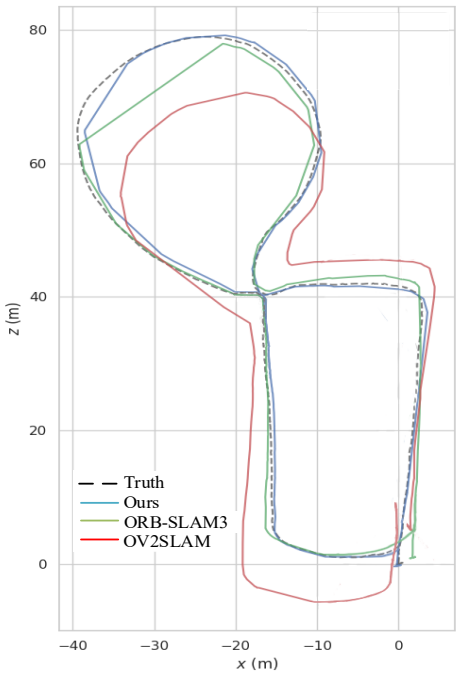


图 4-11 室外真实场景不同算法轨迹对比图

表 4-6 室外真实场景轨迹误差精度对比（单位：m）

算法	绝对轨迹误差(ATE)	相对位姿误差（RPE）
ORB-SLAM3	0.121	0.111
OV2SLAM	0.142	0.132
本文算法	0.083	0.074

4.5 本章小结

本章针对传统视觉 SLAM 回环检测算法仅关注局部特征而忽视图像整体结构的问题，提出了一种基于结构相似性的视觉 SLAM 回环检测改进算法。该算法在回环检测过程中引入了结构相似性评估算法，通过计算图像的全局结构相似性评分，并结合动态权重机制与局部特征评分相融合，从而显著提高了回环候选帧匹配的准确性。本章算法充分考虑了空间物体的结构关系，有效弥补了传统算法忽视全局信息的不足；此外，本章还设计了基于历史回环结果的候选帧剔除机制，能够有效筛除错误的回环候选帧，减少回环误判，从而进一步优化回环检测性能，提高系统的鲁棒性与准确性。实验环节通过公共数据集在结构化与非结构化场景下的验证本章算法；本章算法相比 ORB-SLAM3 和 OV2SLAM，ATE RMSE 分别降低 19.12%和 28.10%，回环检测的召回率分别提高 18.49%和 26.73%，结果表明本章算法在回环检测的优势。同时搭建硬件平台对本文整体算法进行真实场景实验，验证了本文算法可以为视觉 SLAM 系统带来更高的定位精度和更准确的回环检测效果。

第 5 章 全文总结展望

5.1 全文总结

本文从基于词袋模型的视觉 SLAM 回环检测研究现状出发，以实现复杂场景下移动机器人高精度定位和建图为目标，提出一种基于词袋模型的视觉 SLAM 回环检测改进算法。围绕基于词袋模型的视觉 SLAM 回环检测中视觉单词表征能力弱、词典更新滞后及缺乏结构信息利用这一关键问题，提出基于 ConvNeXt 孪生网络的动态视觉词典更新算法和基于结构相似性评估的回环检测算法。通过引入 ConvNeXt 孪生网络提升视觉单词表征能力，并构建动态词典更新机制增强词典对环境变化的适应性；进一步融合图像局部特征与全局结构信息，构建更具判别力的相似性评估方法，并引入回环候选帧剔除模块，有效剔除错误匹配并筛选出高置信度的回环候选帧，最终建立完整的回环检测系统架构。在公开数据集与自主搭建的移动机器人平台上进行的实验验证，结果表明本文算法在回环检测准确率与系统定位精度方面均优于 ORB-SLAM3 与 OV2SLAM 等主流算法，具备良好的鲁棒性与应用前景。

论文主要研究内容和创新点如下：

（1）针对传统视觉 SLAM 算法在回环检测中存在的视觉单词代表性不足及词典无法随环境变化动态更新的问题，本文提出了一种基于 ConvNeXt 孪生网络的动态视觉词典更新算法。首先，采用 ConvNeXt 构建共享权重的孪生网络；其次，设计动态特征融合模块，整合双路网络不同视角下提取的特征，确保信息的充分融合与表达能力的增强。随后，利用主成分分析（PCA）进行降维，以降低计算复杂度，同时借助分层密度聚类（HDBSCAN）去除噪声，构建更加稳健的离线视觉词典。最后，在离线视觉词典的基础上，结合层级树的高效检索机制，通过合并策略与节点递归重建实现视觉词典的动态更新，提升回环检测的灵活性与准确性。本文算法在公开数据集 KITTI 及真实场景中进行了验证，实验结果表明本文算法与 DBoW3、iBoW-LCD 相比 ATE 的 RMSE 平均降低 12.59％、9.34%，回环检测召回率平均提升 3.98%，iBoW-LCD 提升 19.11%。验证了该算法在提高回环检测召回率和鲁棒性方面具有显著优势。

（2）针对传统视觉 SLAM 算法在回环检测过程中忽略场景空间结构信息的问题，提出了一种基于结构相似性评估的回环检测算法。首先，利用结构相似性评估计算图像的全局结构相似性评分，从整体层面捕捉图像间的结构信息，从而提升回环检测的准确性。其次，引入动态权重融合机制，结合图像的局部特征评分与全局结构相似性评分，优化回环候选帧的筛选过程，提高候选帧的匹配精度与相关性。最后，设计回环候选帧剔除机制，通过历史回环结果对潜在误匹配帧进行筛选与剔除，降低回环误判概率，进一步增强回环检测的鲁棒性。本文算法在公开数据集 KITTI、EUROC 和真实场景中进行了实验验证，实验结果表明与 ORB-SLAM3 算法和 OV2SLAM 算法进行对比分析，绝对轨迹误差分别减少 19.12%、28.10%，回环检测召回率分别提升 18.49%、26.73%。验证了该算法能够有效提高回环检测的准确性与稳定性。

5.2 展望

本文围绕“基于词袋模型的视觉 SLAM 回环检测改进算法”进行研究，并通过公开数据集和硬件平台对本文算法进行验证。实验表明本文算法在视觉 SLAM 回环检测中具有较高的鲁棒性，系统建图更准确。然而，仍有一些问题需要更进一步研究与探索：

（1）动态场景下定位问题。虽然本文提出的算法通过动态特征融合模块和结构相似性评估考虑了全局和局部信息的结合，但对于快速变化或具有较强动态性的环境，视觉单词被动态物体影响，导致词典匹配不准确，进而影响回环检测的效果。因此，如何提高算法对动态环境的适应能力，仍是未来改进的重要方向。

（2）稠密场景地图构建。基于本文的研究，下一步将进一步改进地图构建过程，利用准确的回环检测信息来提高稠密场景地图的精度。通过整合回环检测结果，可以有效减少地图中的误差积累，确保稠密地图在大规模环境中的准确性和鲁棒性。

参考文献

[1] 张继贤,刘飞.视觉 SLAM 环境感知技术现状与智能化测绘应用展望[J].测绘学报,2023,52(10):1617-1630.

[2] 阴贺生,裴硕,徐磊,等.多机器人视觉同时定位与建图技术研究综述[J].机械工程学报,2022,58(11):11-36.

[3] 刘鑫,王忠,秦明星.多机器人协同 SLAM 技术研究进展[J].计算机工程,2022,48(5):1-10.

[4] 王耀南,江一鸣,姜娇,等.机器人感知与控制关键技术及其智能制造应用[J].自动化学报,2023,49(3):494-513.

[5] Fischer T, Milford M. Event-based visual place recognition with ensembles of temporal windows[J]. IEEE Robotics and Automation Letters, 2020, 5(4): 6924-6931.

[6] Molloy T L, Fischer T, Milford M, et al. Intelligent reference curation for visual place recognition via Bayesian selective fusion[J]. IEEE Robotics and Automation Letters, 2020, 6(2): 588-595.

[7] Cao Y, Beltrame G. VIR-SLAM: Visual, inertial, and ranging SLAM for single and multi-robot systems[J]. Autonomous Robots, 2021, 45(6): 905-917.

[8] Zhu J, Li H, Zhang T. Camera, LiDAR, and IMU based multi-sensor fusion SLAM: A survey[J]. Tsinghua Science and Technology, 2023, 29(2): 415-429.

[9] 刘铭哲,徐光辉,唐堂,等.激光雷达 SLAM 算法综述[J].计算机工程与应用,2024,60(1):1-14.

[10] 孙新柱,龚光强,陈孟元等.室内场景下基于曼哈顿约束的多重特征视觉 SLAM 方法[J].中国惯性技术学报,2023,31(9):890-899.

[11] Zhuang L, Zhong X, Xu L, et al. Visual SLAM for unmanned aerial vehicles: Localization and perception[J]. Sensors, 2024, 24(10): 2980-2304.

[12] Tsintotas K A, Bampis L, Gasteratos A. The revisiting problem in simultaneous localization and mapping: a survey on visual loop closure detection[J]. IEEE Transactions on Intelligent Transportation Systems, 2022, 23(11): 19929-19953.

[13] 陈孟元,李鹏飞,符乙,等.低光条件下结合 FFE-Net 网络的视觉 SLAM 算法[J].

中国惯性技术学报,2025,33(1):36-45.

[14] Samadzadeh A, Nickabadi A. Srvio: Super robust visual inertial odometry for dynamic environments and challenging loop-closure conditions[J]. IEEE Transactions on Robotics, 2023, 39(4): 2878-2891.

[15] Campos C, Elvira R, Rodríguez J J G, et al. Orb-slam3: An accurate open-source library for visual, visual–inertial, and multimap slam[J]. IEEE Transactions on Robotics, 2021, 37(6): 1874-1890.

[16] Bescos B, Fácil J M, Civera J, et al. DynaSLAM: Tracking, mapping, and inpainting in dynamic scenes[J]. IEEE robotics and automation letters, 2018, 3(4): 4076-4083.

[17] Yu J, Shen S. SemanticLoop: Loop closure with 3D semantic graph matching[J]. IEEE Robotics and Automation Letters, 2022, 8(2): 568-575.

[18] Ma J, Ye X, Zhou H, et al. Loop-closure detection using local relative orientation matching[J]. IEEE Transactions on Intelligent Transportation Systems, 2021, 23(7): 7896-7909.

[19] Durrant-Whyte H, Bailey T. Simultaneous localization and mapping: part I[J]. IEEE Robotics and Automation Magazine, 2006, 13(2): 99-110.

[20] Newcombe R A, Lovegrove S J, Davison A J. DTAM: Dense tracking and mapping in real-time[C].2011 international conference on computer vision (ICCV 2011). Barcelona Spain, IEEE, 2011: 2320-2327.

[21] Engel J, Schöps T, Cremers D. LSD-SLAM: Large-scale direct monocular SLAM[C]. European conference on computer vision (ICCV 2011). Barcelona, Spain, IEEE, 2014: 834-849.

[22] Engel J, Koltun V, Cremers D. Direct sparse odometry[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 40(3): 611-625.

[23] Zubizarreta J, Aguinaga I, Montiel J M M. Direct sparse mapping[J]. IEEE Transactions on Robotics, 2020, 36(4): 1363-1370.

[24] 贾嫣晗,邹风山,徐方,等.完全在线的双目直接法视觉 SLAM 算法[J].控制与决策,2023,38(11):3093-3102.

[25] 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.

[26] 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.

[27] 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.

[28] 张涛,陈浩,闫捷等.基于双向重投影的双目视觉里程计[J].中国惯性技术学报, 2018, 26(6):752-759.

[29] 张洪,于源卓,邱晓天.基于改进关键帧选择的 ORB-SLAM2 算法[J].北京航空航天大学学报,2023,49(1):45-52.

[30] 李博洋,刘思健,崔明月,等.基于最小回环检测的多车协同 SLAM 框架[J].电子学报,2021,49(11):2241-2250.

[31] CLEMENTEL， DAVISONA． Mapping large loops with a single hand-held camera[C]. Robotics Science and Systems (RSS 2007). Bei Jing,China, IEEE, 2007, 2(2) :297-304．

[32] 安平,王国平,余佳东,等.一种高效准确的视觉 SLAM 闭环检测算法[J].北京航空航天大学学报,2021,47(1):24-30.

[33] Williams B, Cummins M, Neira J, et al. An image-to-map loop closing method for monocular SLAM[C] 2008 IEEE/RSJ International Conference on Intelligent Robots and Systems(IROS 2008). Nice, France, IEEE, 2008: 2053-2059．

[34] Gálvez-López D, Tardos J D. Bags of binary words for fast place recognition in image sequences[J]. IEEE Transactions on Robotics, 2012, 28(5): 1188-1197.

[35] Gao X, Zhang T. Unsupervised learning to detect loops using deep neural networks for visual SLAM system[J]. Autonomous Robots, 2017, 41: 1-18.

[36] Yuan Z, Xu K, Zhou X, et al. SVG-Loop: Semantic–visual–geometric information-based loop closure detection[J]. Remote Sensing, 2021, 13(17): 3520-3544.

[37] 李康宇,王西峰,徐斌,等.非结构化环境下基于外观的闭环检测研究综述[J].机器人,2023,45(2):238-256.

[38] Sivic, Zisserman. Video Google: A text retrieval approach to object matching in videos[C]Proceedings Ninth IEEE International Conference on Computer Vision.

(ICCV 2003) Nice, France, IEEE, 2003: 1470-1477 .

[39] Newman P, Ho K. SLAM-loop closing with visually salient features[C]. Proceedings of the 2005 IEEE International Conference on Robotics and Automation (ICRA 2005), Barcelona, Spain, IEEE, 2005: 635-642.

[40] Angeli A, Filliat D, Doncieux S, et al. Fast and incremental method for loop closure detection using bags of visual words[J]. IEEE Transactions on Robotics, 2008, 24(5): 1027-1037.

[41] Labbe M, Michaud F. Memory Management for Real-Time Appearance-Based Loop Closure Detection[C]. IEEE International Conference on Intelligent Robots and Systems (IROS 2011), San Francisco, IEEE, 2011,1271-1276.

[42] 林俊钦,韩宝玲,罗庆生,等.基于 NDT 匹配和改进回环检测的 SLAM 研究[J]. 光学技术,2018,44(2):152-157.

[43] Ferrera M, Eudes A, Moras J, et al. OV2SLAM: A fully online and versatile visual SLAM for real-time applications[J]. IEEE Robotics and Automation Letters, 2021, 6(2): 1399-1406.

[44] 王朋,郝伟龙,倪翠,等.视觉 SLAM 方法综述[J].北京航空航天大学学报,2024,50(2):359-367.

[45] Rublee E, Rabaud V, Konolige K, et al. ORB: An efficient alternative to SIFT or SURF[C] 2011 International Conference on Computer Vision (ICCV 2011). Barcelona, Spain, IEEE, 2011: 2564-2571.

[46] Pollefeys M, Nistér D, Frahm J M, et al. Detailed real-time urban 3d reconstruction from video[J]. International Journal of Computer Vision, 2008, 78: 143-167.

[47] Engel J, Sturm J, Cremers D. Semi-dense visual odometry for a monocular camera[C]. Proceedings of the IEEE International Conference on Computer Vision (ICCV 2013). Sydney, Australia, IEEE, 2013: 1449-1456.

[48] Kuçak R A, Erol S, Erol B. The strip adjustment of mobile LiDAR point clouds using iterative closest point (ICP) algorithm[J]. Arabian Journal of Geosciences, 2022, 15(11): 1017.

[49] Liu Z, Zhang F. Balm: Bundle adjustment for lidar mapping[J]. IEEE Robotics and Automation Letters, 2021, 6(2): 3184-3191.

攻读学位期间发表的学术论文和科研成果

一、发表学术论文

[1] 徐奥,孙新柱,陈孟元.基于结构相似性评估的视觉 SLAM 回环检测算法[J/OL].重庆工商大学学报(自然科学版),1-11[2025-03-19].http://kns.cnki.net/kcms/detail/50.1155.N.20241203.1143.002.html.

[2] Yang S, Xu A, Li P, et al. Visual SLAM Algorithm Based on YOLOv5 in Dynamic Scenario[C]. 2023 China Automation Congress (CAC). IEEE, 2023: 2640-2645.

[3] Li P, Xu A, Cheng H, et al. Optimization algorithm for closed-loop detection based on similarity fusion mechanism[C]. 2023 38th Youth Academic Annual Conference of Chinese Association of Automation (YAC). IEEE, 2023: 1340-1344.

二、参与的科技项目

[1] 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）

[2] 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）

[3] 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）

致谢

行文至此，落笔为终。三年的研究生生涯即将画上句号，回首求学之路，虽不乏艰辛，但更多的是收获与成长。在此，谨向所有给予我支持、指引与温暖的师长、亲友致以最诚挚的谢意。

首先，衷心感谢我的导师孙新柱老师。感谢孙老师在学术道路上的悉心指导与严格要求。从论文选题、框架设计到研究推进，孙老师始终以渊博的学识、严谨的治学态度和耐心的指导为我指明方向。孙老师对学术的热情、对细节的追求以及对学生的包容，让我深刻理解了科研精神的真谛。同时，也衷心感谢实验室的陈孟元老师，在课程学习、课题讨论和论文评审中给予的宝贵建议。陈老师对理论深度的极致追求与充满智慧的批判视角，使我在方法论层面实现了质的突破。两位老师的学术品格，恰似双星辉映，指引着我在求真之路上稳步前行。此外，也要衷心感谢我的校外导师于晓东，在专业实践期间对于我的指导与关心，不仅让我加深了对专业知识的理解，也帮助我提升了独立思考和解决问题的能力。

其次，感谢已经从实验室毕业的师兄们。他们不仅在专业知识上给予了我极为宝贵的指导，更以自己的实际行动树立了学习的榜样。同时，感谢同门李鹏飞、符乙、许瑞珩、杨苏朋、丁帅和我的室友于卓越、李淼、杜鹏，我们在一起生活学习三年，在硕士研究生期间相互帮助。回想起无数个共同奋斗的日日夜夜，一起携手攻克了一个又一个难关。这段旅程充满了挑战，但正因为有你们的智慧分享和温暖陪伴，让这一切都变得格外有意义和难忘，我相信这是我人生中的一段不可忘却的宝贵经历。

然后，感谢我的父母。二十余载求学路，是你们毫无保留的支持与包容，让我心无旁骛地追逐学术理想。每当我因压力疲惫时，你们永远是最温暖的港湾。特别感谢我的爱人吴丛，感谢你长久以来的陪伴，是你让我有了坚持下去的动力，也感谢你的鼓励与包容。父母的默默付出、爱人的包容鼓励，始终是我最坚实的后盾。寥寥数语，难表深情，唯愿以未来的努力回报这份沉甸甸的爱。

最后，感谢在论文评审与答辩中付出辛劳的各位专家。求学时光虽已落幕，但探索真理的步履永不停歇。愿以此文为起点，在未来的道路上继续坚守初心，不负所期。

尚德教学　唯实惟新