视觉 SLAM 运动分割技术综述

冯嘉琪 [1] 杨恺伦 [2] 林家丞 [1, 3] 杨观赐 [1, 3]

摘 要 作为移动机器人与自动驾驶领域的关键基础技术, 视觉同时定位与地图构建 (V-SLAM) 在动态环境中面临严峻挑战. 由动态物体引起的特征匹配错误常常导致定位偏差、地图失真以及系统鲁棒性受损. 运动分割技术是提高 V-SLAM 性能的重要手段, 但在复杂动态场景中准确区分静态和动态元素仍极具挑战性. 本文系统梳理 V-SLAM 运动分割研究进展, 根据对环境的潜在假设, 将现有方法分为三个主要研究范式, 并给出各范式的技术原理、代表性策略的核心优势、本质局限及适用边界. 最后展望未来的研究方向.

关键词 视觉 SLAM; 动态环境; 运动分割; 运动理解; 多传感器融合; 移动机器人

引用格式 冯嘉琪, 杨恺伦, 林家丞, 杨观赐. 视觉 SLAM 运动分割技术综述. 自动化学报, 2026, **52**(4): 1−27

DOI 10.16383/j.aas.c250365 **CSTR** 32138.14.j.aas.c250365

A Review of Motion Segmentation Techniques for Visual SLAM

FENG Jia-Qi[1] YANG Kai-Lun[2] LIN Jia-Cheng[1, 3] YANG Guan-Ci[1, 3]

Abstract As a critical foundational technology in the fields of mobile robots and autonomous driving, visual simultaneous localization and mapping (V-SLAM) faces severe challenges in dynamic environments. Feature mismatches induced by dynamic objects frequently lead to localization drift, map distortion, and degradation of system robustness. Motion segmentation technology is an important means of enhancing V-SLAM performance, but accurate discrimination between static and dynamic elements in complex dynamic scenarios remains highly challenging. This paper systematically reviews the research progress on motion segmentation for V-SLAM. Taxonomically categorizing existing methods into three primary research paradigms based on underlying environmental assumptions, we present the technical principles of each paradigm, along with the core strengths, inherent limitations, and applicability boundaries of representative strategies. Finally, future research directions are prospected.

Keywords visual SLAM; dynamic environments; motion segmentation; motion understanding; multi-sensor fusion; mobile robots

Citation Feng Jia-Qi, Yang Kai-Lun, Lin Jia-Cheng, Yang Guan-Ci. A review of motion segmentation techniques for visual SLAM. *Acta Automatica Sinica*, 2026, **52**(4): 1−27

同时定位与地图构建 (simultaneous localization and mapping, SLAM)[1] 技术是移动机器人、自动驾驶和增强现实领域开展自主导航和沉浸式交互体验的技术基础. 在机器人领域, 从地面移动机器人室内巡检、室外测绘到滑移转向机器人爬坡越障以及腿足机器人行走在复杂地形上, SLAM 都是其开展自主定位与环境感知的必备技术[2]. 其中, 视觉同时定位与地图构建 (visual simultaneous localization and mapping, V-SLAM)[3−4] 凭借相机成本优势与丰富的视觉信息获取能力, 能够为服务机器人、巡检机器人等智能装备的自主感知与决策提供核心技术支撑, 有效满足其在复杂场景下的自主运行需求. 但在动态场景下由于相机的前方往往存在多个动态目标, 因此会导致 V-SLAM 系统误跟或者丢失特征点的情况出现, 进而导致定位系统的误差增加, 并进一步产生错误的导航决策结果. 为使位姿估计仍能保持准确, 需对动态与静态特征加以区分, 实现对动态特征的剔除, 这个过程被定义为运动分割[5]. 随着计算机视觉、人工智能与机器人学技术发展, 运动分割由原先单一基于几何特征匹配开始向融合深度学习语义理解转变[6−8], 这使得其能够

收稿日期 2025-08-01 录用日期 2025-12-09
Manuscript received August 1, 2025; accepted December 9, 2025
国家自然科学基金 (62373116, 62473139), 贵州省科技计划项目 (黔科合支撑 [2023] 一般 118), 湖南省重点研发计划 (2025QK3019) 资助
Supported by National Natural Science Foundation of China (62373116, 62473139), Guizhou Provincial Science and Technology Project (QKHZC [2023] 118), and Key R&D Program of Hunan Province (2025QK3019)
本文责任编委 黄华
Recommended by Associate Editor HUANG Hua
1. 贵州大学现代制造技术教育部重点实验室 贵阳 550025 2. 湖南大学人工智能与机器人学院 长沙 410012 3. 贵州大学贵州省电子信息与智能应用国际科技合作基地 贵阳 550025
1. Key Laboratory of Advanced Manufacturing Technology of Ministry of Education, Guizhou University, Guiyang 550025 2. School of Artificial Intelligence and Robotics, Hunan University, Changsha 410012 3. Guizhou Province International Science & Technology Cooperation Base of Electronic Information and Intelligent Applications, Guizhou University, Guiyang 550025

更好地服务于机器人复杂任务场景 (如人机协同、动态避障), 从而提高 V-SLAM 定位的稳定性.

现有的大部分视觉同时定位与建图方法通常建立在静态环境假设之上. 然而, 在真实场景中普遍存在大量的动态物体, 导致该假设在实际应用中往往不再成立. 对于地面机器人, 其运动受地表结构约束, 如果地面上出现动态物体 (比如车辆、行人等), 有可能会打破运动模型的一致性; 同样地, 对于滑移转向机器人, 其运动学参数易受地形与载荷影响, 动态物体的出现也可能会导致运动学参数估算结果出现偏差. 因此, 提升系统在动态环境中的适应性, 关键在于构建对场景运动信息的"感知–理解–处理"闭环.

当前研究多聚焦于运动分割技术, 仅通过识别和分割动态区域, 难以支撑构建具有智能决策能力和强适应性的 V-SLAM 系统. 动态物体的运动状态 (如瞬时速度、加速度)、运动模式 (如刚性运动、非刚性变形、匀速或加减速)、动态交互 (如物体间的相互遮挡与运动关联) 以及多模态运动信息 (如光流场、场景流、几何运动约束) 中, 蕴含着丰富的环境动态特征. 这种让机器人或系统能够像人类一样"看明白"动态场景中的运动行为或变化, 通过感知、分析和解释物体或场景中的运动信息 (如运动位置、运动方式、运动属性、运动关联), 从而推断出运动的目的、意图或背后规律的过程即为运动理解. 简言之, 运动理解就是让机器人或系统不仅能够看到想看的内容 (感知), 还能像人类一样"一叶落而知秋", 看懂运动行为背后蕴含的意义 (认知). 运动理解不仅涵盖运动分割以识别哪里在动, 更延伸至多维度的深度解析: 1) 运动状态解析, 即通过量化动态物体的速度、方向、轨迹等物理特征, 明确其运动方式; 2) 运动模式识别, 即区分物体是刚体运动 (如车辆)、非刚体运动 (如行人、动物) 还是周期性运动等特定模式, 理解其运动性质; 3) 动态交互建模, 即分析动态物体间的相互作用 (如避让、跟随) 及与静态环境的交互 (如遮挡关系), 理解其运动关系; 4) 多模态信息融合, 即综合视觉 (光流、场景流)、几何约束 (对极几何、重投影误差) 及多传感器 (如惯性测量单元 (inertial measurement unit, IMU) 角速度、雷达径向速度) 提供的运动线索, 形成对场景动态性的统一鲁棒表征, 将运动与高层语义关联, 把握其运动本质. 运动理解是动态环境下 V-SLAM 系统具备良好决策能力的关键, 也是对抗动态干扰的认知基础. 运动分割技术作为运动理解的核心基础, 是解决动态环境下 V-SLAM 定位失效与系统性能下降问题的直接且关键手段, 尤其是在复杂动态场景中, 其核心挑战在于如何实时、准确地实现静态与动态物体的有效分割与区分.

本文系统梳理视觉同时定位与建图的运动分割技术, 根据预设的环境条件将现有方法分为静态假设、语义信息和多传感器融合三大类体系, 分析现有方法的优缺点, 总结当前的方法对于运动物体分割效果较差的原因, 并重点论述如何实现语义理解以及深度学习模型的优化、多传感器的深度融合, 从而为突破视觉 SLAM 技术瓶颈、推动其在实际场景中的广泛应用提供理论支撑与技术参考.

1 视觉 SLAM 的运动分割研究现状

从前端像素间匹配过程看, 早期视觉 SLAM 方法主要有两类: 特征点法 (基于特征的间接法) 以及直接法 (基于像素灰度值的直接法)[9]. 其中, 特征点法是基于特征点最小化投影误差优化相机运动, 并通过检测及跟踪图像中有效特征点实现定位及建图的方法, 其代表为 ORB-SLAM (oriented fast and rotated brief SLAM) 系列[10−12]. 此方法利用了后端全局优化技术提高了地图构建的一致性, 整体上算法结构稳定、鲁棒性强. 直接法则是在一对连续图像帧之间基于最小化像素灰度差值来估计机器人的位姿, 典型代表是 LSD-SLAM[13] 与 DSO[14] 等方法, 此类方法适用于环境纹理明显的场景且场景变化缓慢情况. 受大规模场景、运动目标较多等因素限制, 上述两类方法因存在累积误差与特征匹配缺陷, 难以满足大规模场景的应用需求. 尤其在动态环境中, 相机位姿累积误差的不断叠加与特征匹配失效, 会直接导致位姿估计误差增大、地图构建异常等问题. 另外, 动态目标运动轨迹的偏移会进一步造成目标点的位置偏差, 最终影响整体建图效果.

幸运的是, 深度学习为视觉 SLAM 带来了新的机遇[15−18]. 基于卷积神经网络 (convolutional neural networks, CNN) 的语义分割以及目标检测算法能够获得当前场景中物体类别的信息和所在的位置信息, 从而给运动分割提供了语义先验信息; 将语义信息与传统的几何约束相结合, 可以构建出语义 SLAM 系统, 有利于从原始的几何地图向包含语义标签的智能地图跨越[19−22]. 但是, 由于深度学习模型计算量较大, 且缺乏针对带宽的系统约束, 在运算速度上有较大局限性, 特别是在应用场景中较常见的嵌入式设备上这一缺点则更为突出. 另外由于现实世界场景的变化性较强, 不同类型的传感器以及其相应的系统装置又各有差异, 系统在泛化到多场景、多设备的能力方面也始终没有完全令人满意的表现.

与此同时, 多传感器融合也是对视觉 SLAM 的一种补充优化方法[2, 23–25], 它把激光雷达 (light detection and ranging, LiDAR)、IMU 和相机的视觉信息综合利用起来, 能够提高复杂环境下系统定位及建图的精度和鲁棒性. 激光雷达提供的高精度距离信息可弥补视觉测深的不足; IMU 则能以高频次信息支撑系统位姿估计, 即使无法检测到视觉特征点, 依旧可依靠 IMU 高频率获取信息来保持位姿估计的连续性. 但多传感器融合还存在数据同步、数据校准以及融合算法设计等问题, 需进一步研究[26].

随着可微分渲染与神经隐式表征技术的兴起, 以神经辐射场 (neural radiance fields, NeRF) 和 3D 高斯泼溅 (3D Gaussian splatting, 3DGS) 为代表的新方法为 V-SLAM 的运动分割带来了新的技术突破[27–29]. 这类方法通过可微分的场景表征, 能够隐式地学习静态环境与动态物体的分布, 从而实现更为精细的运动分割与背景修复. NeRF[30] 作为一种对 3D 场景的隐式神经表示方法, 通过多层感知器 (multilayer perceptron, MLP) 网络隐式表征场景辐射场, 利用可微分渲染最小化渲染误差, 间接实现动态区域识别. 但也有自身缺陷, 如训练效率低、动态物体破坏多视图一致性导致渲染残差增大、难以进行高分辨率实时渲染, 因此更加适合复杂光照与弱纹理场景的动态分割. 3DGS[31] 以显式 3D 高斯点云为场景表征单元, 通过可微分渲染实现实时高精度重建, 同时利用高斯分布的运动一致性与多视图约束区分静态、动态元素, 兼顾渲染速度与几何精度.

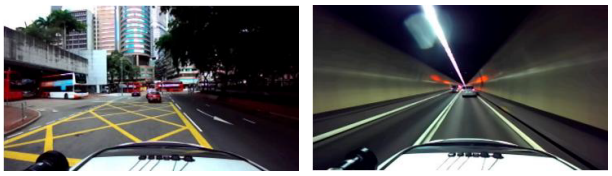
此外, 近年来还出现了一种将持续学习引入神经辐射场的新型融合方法[32]. 该类方法不显式依赖几何、语义或多传感器信息, 而是通过神经网络隐式地学习场景中的动态与静态成分, 实现记忆与遗忘的平衡. 该类方法的代表性工作如 Li 等[33] 提出的框架, 通过在线学习机制自适应地更新场景表示, 动态物体被视为应遗忘的信息, 而静态结构则被记忆下来. 这种方法避免了显式的运动分割步骤, 通过优化过程中的梯度传播隐式地实现动态抑制, 为运动分割提供了全新的思路.

伴随着动态环境下视觉 SLAM 和运动分割技术的发展, 相关综述也得到不少学者的关注[6–8, 34–38]. 不同时间段的研究侧重点和技术路线差异较大. 初期主要针对 SLAM 的基础架构做了相关研究, 当运动干扰现象明显之后, 更多的学者开始注重对运动特征的处理. 文献 [35] 按相机运动方式把动态视觉 SLAM 分成相机自身运动的 SLAM 和非相机自身运动的方式. 文献 [36] 则是从广义上的动态 SLAM 来阐述. 文献 [7] 仅是把运动分割当作是动态环境下做 SLAM 的一个子模块. 文献 [38] 针对动态环境下的视觉 SLAM 与运动恢复结构 (structure from motion, SFM) 技术所遇到的主要问题, 即在移动目标影响下如何保持定位精度, 进行了系统的论述, 提出了三类解决办法: 第一类为鲁棒视觉 SLAM, 通过运动分割剔除动态特征; 第二类为动态对象分割及 3D 跟踪识别来追踪移动物体的轨迹; 第三类为联合运动分割与重建的方法同时处理运动和静止的两种情况. 文献 [8] 针对动态视觉 SLAM 的特征选择, 分别强调了低级特征与高级特征. 低级特征采用基于点或者线的描述子匹配和几何优化方法实现高精度定位; 高级特征采用语义分割和目标检测加入先验知识来辅助运动分割与地图理解. 文献 [37] 以场景预设条件为核心逻辑划分运动分割方法, 综述了 2022 年前的工作. 简言之, 相关综述大多是围绕 SLAM 系统地展开, 尽管也出现了运动分割领域的针对性综述, 但未脱离 SLAM 系统附属模块的视角, 整体仍以 SLAM 系统为核心, 未能充分将运动分割作为独立研究领域, 对其技术突破路径、实时性瓶颈破解、多场景适配等关键难题的系统性探讨仍显不足.

1.1 视觉 SLAM 中的运动分割的难点与挑战

动态环境具有不确定性及变化的特点, 包括外部因素中的移动障碍物、天气的变化、光照的变化, 还有传感器固有的噪声以及缺失的数据. 图 1 展示了动态环境下的光照变化情况, 在真实的世界中, 除了常规的行人、车辆外, 各类不同形状、大小和运动方式的物体也在这个世界里来回地穿梭, 它们具有不同的外观、纹理以及运动轨迹, 使准确识别和分割变得非常困难. 因此, 动态环境对于视觉 SLAM 系统是一个极大的考验, 尤其是在针对运动分割这一方面的难题.

概括而言, 在动态环境中, 视觉 SLAM 系统能否以较快速度且正确地识别特征与分割是保证其实



(a) 户外自然光环境
(a) Outdoor natural light environment

(b) 封闭灯光环境
(b) Enclosed lighting environment

图 1 动态环境下光照变化示例

Fig. 1 Illustration of illumination variation in dynamic environments

时性与准确性的关键, 其难点与挑战概括如下:

1) 动态目标追踪和分割. 由于动态物体的运动轨迹是十分复杂且不可预测的, 在很多情况下, 动态物体之间会有遮挡等干扰情况出现, 如果某个对象发生了移动或其他的改变, 则很可能会导致传统的分割算法无法有效实现各个目标的区分. 特别是在多目标干扰的情况下, 应该使分割算法实现精确实时的跟踪, 而不只是一般的追踪. 对于现有的分割算法而言, 在多种复杂条件下实现对动态目标的稳定追踪仍然存在不足. 例如, 行人的动作姿态具有较大的不确定性, 车辆的姿态也会随行驶状态发生变化, 并且在运动过程中常伴随重叠遮挡等情况; 此外, 部分方法难以识别先验静止物体是否发生过运动变化. 典型实例包括: 静止不动的车辆、保持坐姿不动的行人、被他人移动的椅子以及被丢弃的球等.

2) 传感器数据噪声和融合问题. 因为视觉传感器所采集的图像易受光照强度的变化、被测物是否有遮挡、图像是否存在模糊等影响, 这些因素都会引起图像的噪声和不确定性. 例如, 在光照条件比较复杂的场景中, 图像的亮度和对比度就会发生很大的变化, 这样就使特征提取、匹配不容易实现, 图像往往达不到很好的效果, 移动目标的快速运动会使图像变得模糊, 数据就更难处理. 因此从传感器数据中提取出动态目标特征以及运动信息都非常困难, 会严重影响到动态目标的运动分割准确率.

3) 运动分割算法的实时性和精准度. 在视觉 SLAM 的实际应用场景中, 系统可实现对大量运动目标的实时分割与跟踪. 但对于提高算法精度需要更多的复杂度与模型支持, 这就导致其计算量一般较大, 达到实时性需求十分困难. 如基于深度学习的分割算法有比较高的准确率, 但是计算量大、运算速度较慢, 难以满足实时性要求.

4) 场景动态变化的适应性. 动态环境中的场景可能会随时发生变化, 如天气变化、物体的突然出现或消失等. 视觉 SLAM 系统需要能够快速适应这些变化, 及时调整运动分割策略. 然而, 现有的大多数算法对场景动态变化的适应性较差, 难以在复杂多变的环境中保持稳定的性能.

1.2 视觉 SLAM 的评估标准与数据集

1.2.1 评估标准

动态环境下 V-SLAM 的核心性能评估指标, 需针对动态场景的特殊性, 构建覆盖定位精度、运动分割性能、建图质量三大维度的评估体系, 各维度指标需兼顾算法特性与实际应用需求, 它们在准确量化算法性能、公平比较算法表现、识别潜在故障模式以及寻找突破点等方面发挥着重要作用. 在 SLAM 的性能量化中, 无论是定位精度、运动分割性能还是建图质量, 均广泛采用均方根误差 (root mean square error, RMSE) 作为核心统计指标. 一方面, RMSE 通过对误差的平方运算, 对较大误差更加敏感. 如动态物体干扰导致的帧间位姿跳变、动态掩码边缘的误判像素, 而动态场景中这类异常误差正是影响系统鲁棒性的关键, RMSE 可精准捕捉此类问题, 避免因采用平均误差等指标掩盖关键偏差. 另一方面, RMSE 具有明确的物理意义, 例如 ATE-RMSE的单位为“m”、重投影误差 RMSE 的单位为“像素”, 能直观反映算法在实际应用中的性能表现, 便于不同方法间的横向对比. 具体定义与计算逻辑如下:

1) 定位精度指标

定位精度是 V-SLAM 系统的核心性能, 动态环境下需重点衡量动态物体干扰下的位姿一致性, 常用指标包括绝对轨迹误差 (absolute trajectory error, ATE) 和相对位姿误差 (relative pose error, RPE). 这两个评价指标最早是在 TUM 数据集基准中定义的, 应用非常广泛[39].

ATE 通过系统输出轨迹与真值轨迹的欧氏距离来反映全局位姿偏差, 计算公式为:

$$ATE_{\mathrm{RMSE}} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left\| T_{\mathrm{est},\,i} \otimes T_{\mathrm{gt},\,i}^{-1} \right\|_{\mathrm{F}}^{2}} \tag{1}$$

其中, N 为轨迹的总帧数; $T_{\mathrm{est},\,i}$、$T_{\mathrm{gt},\,i}$ 分别为第 i 帧的估计位姿与真值位姿 ($SE(3)$ 空间变换矩阵); $\otimes$ 表示李群乘法; $\|\cdot\|_{\mathrm{F}}$ 为 Frobenius 范数. 该指标对长序列累积误差敏感, 适用于评估高动态场景的定位稳定性.

RPE 计算相邻 k 帧 (通常取 $k=1$ 或 $k=5$) 的相对位姿偏差来反映局部帧间位姿精度, 公式为:

$$RPE_{\mathrm{RMSE}} = \sqrt{\frac{1}{N-k}\sum_{i=1}^{N-k}\left\| E_i \right\|_{\mathrm{F}}^{2}} \tag{2}$$

其中, $E_i = \Delta T_{\mathrm{est},\,i:i+k} \otimes \Delta T_{\mathrm{gt},\,i:i+k}^{-1}$; $T_{\mathrm{est},\,i:i+k}$ 为估计的第 i 至 $i+k$ 帧相对位姿. 该指标更适合评估动态物体突发干扰下的系统响应能力.

2) 运动分割性能指标

运动分割是动态 V-SLAM 的关键前置步骤, 直接决定定位与建图质量, 需从像素级与特征级双层面评估. 像素级分割指标针对语义分割或实例分割输出的动态掩码, 采用平均交并比 (mean intersection over union, mIoU) 与平均精度 (average precision, AP) 来衡量. mIoU 计算动态区域预测掩码

与真值掩码的交并比并对所有动态类别取平均, 公式为:

$$mIoU = \frac{1}{C}\sum_{c=1}^{C}\frac{TP_c}{TP_c + FP_c + FN_c} \tag{3}$$

其中, C 为动态类别数; TP_c、FP_c、FN_c 分别为第 c 类的真阳性、假阳性、假阴性像素数. AP 则针对实例分割结果, 统计不同 IoU 阈值下的精度–召回率曲线下面积, 反映动态实例的完整识别能力, 适用于多目标动态场景.

特征级分割针对传统几何方法输出的动态特征点, 采用准确率与召回率衡量. 准确率是能正确识别的动态特征点占系统标记动态点总数的比例, 反映避免静态点误判的能力; 召回率指正确识别的动态特征点占真实动态点总数的比例, 反映捕捉动态点的能力.

3) 建图质量指标

动态环境下的建图需兼顾静态背景完整性与动态区域无干扰, 常用地图点云完整性与重投影误差 (reprojection error, RE) 衡量指标. 地图点云完整性计算系统输出的静态点云与真值点云的重叠率, 该指标反映剔除动态物体后静态场景的保留程度.

重投影误差统计静态地图点在图像平面的投影偏差, 公式为:

$$RE_{\mathrm{RMSE}} = \sqrt{\frac{1}{M}\sum_{j=1}^{M}\left\|\pi\left(T_i \times P_j\right) - u_{i,\,j}\right\|_2^2} \tag{4}$$

其中, M 为第 i 帧中参与统计的静态地图点数量; T_i 表示第 i 帧相机的位姿变换矩阵; P_j 为第 j 个静态地图点的 3D 坐标; $\pi(\cdot)$ 为相机投影函数; $u_{i,\,j}$ 为第 i 帧中 P_j 的观测像素坐标. 该指标反映静态地图与视觉观测的一致性, 动态区域修复不连贯会导致 RE 显著上升.

1.2.2　数据集

为了更加客观地评估视觉 SLAM 方法对于动态环境的适应能力, 研究人员构建了大量包含动态目标的开源数据集, 根据应用环境主要可以分为室内、室外及跨场景类型, 并且搭配了不同传感器, 包括 LiDAR、RGB 相机、事件相机 (event-based camera) 等, 以提供充足的环境信息描述. 表 1 列举了常用的动态环境下视觉 SLAM 数据集及其特点.

1) 室内动态数据集

在室内场景的数据集中, TUM 数据集[39] 是一个 RGB-D 图像序列, 其中的动态目标类别包含 9 个图像序列, 这些序列常用于室内动态环境下 SLAM 系统的运动分割研究, 涵盖了人员静坐与行走两种动态场景. Bonn 数据集[40] 则包含 24 个图像序列, 涉及 9 种室内动态场景, 其场景丰富度高于 TUM 数据集. OpenLORIS-Scene 数据集[41] 是 IROS 2019 Life Long SLAM Challenge 的官方数据集, 覆盖 5 种室内动态场景. 该数据集相较于 TUM 和 Bonn 数据集更具挑战性, 存在光照变化、视角差异以及人类生活引发的环境改变等情况, 包含 RGB-D 图像、双目图像、2D 与 3D 激光点云、IMU 及轮式里程计数据. ICL-NUIM 数据集[42] 由英国布

表 1　常用的动态环境下视觉 SLAM 数据集

Table 1　Common visual SLAM datasets in dynamic environments

数据集名称	场景	采集平台	LiDAR	RGB	MONO (单色)	Stereo	IMU	Event	GNSS
TUM[39]	室内	手持、机器人		√			√		
Bonn[40]	室内	手持		√					
OpenLORIS-Scene[41]	室内	机器人	√	√		√	√		
ICL-NUIM[42]	室内	手持		√					
Augmented ICL-NUIM[43]	室内	手持		√					
KITTI[44]	室外	汽车	√		√	√	√		√
Oxford[45]	室外	汽车	√	√	√	√			√
Mapillary Vistas[46]	室外	手持、机器人	√	√					
ApolloScape[47]	室外	汽车	√	√					
ADVIO[48]	室内、室外	手持			√		√		
M2DGR[49]	室内、室外	机器人	√	√		√	√	√	√
RAWSEEDS[50]	室内、室外	机器人	√			√	√		√
EuRoC MAV[51]	室内、室外	飞行器			√	√	√		
FusionPortableV2[52]	室内、室外	手持、机器人、汽车	√			√	√	√	√
M3DGR[53]	室内、室外	机器人	√	√			√		√

里斯托大学与诺丁汉大学的研究人员联合开发, 其视觉数据通过 RGB-D 相机在客厅和办公室两类室内环境中采集而来, 主要用于评估 3D 地图重建、SLAM 或视觉里程计算法在不同纹理、光照条件及结构变化下的性能. Augmented ICL-NUIM 数据集[43] 同样由英国布里斯托大学的 ICL-NUIM 团队构建, 该数据集是一个增强型 RGB-D 数据集, 旨在推动增强现实和虚拟现实领域的相关研究.

2) 室外动态数据集

KITTI 数据集[44] 聚焦室外道路场景, 使用了双目相机、激光雷达、IMU 等多种传感设备, 图像数据包含汽车、自行车、行人等室外道路常见动态目标. Oxford 数据集[45] 对牛津地区部分道路展开了大规模重复采集, 累计行驶里程逾 1 000 kM, 提供单目/双目图像、激光点云及 GPS/INS 组合导航数据, 图像数量超 2 000 万张. Mapillary Vistas 数据集[46] 与 ApolloScape 数据集[47] 均面向室外场景, 前者采集设备由手持终端或机器人搭载, 后者仅以机器人作为数据采集平台, 配备有 LiDAR 与 RGB 传感器, 适用于评估依赖 LiDAR 与 RGB 信息的手持设备或机器人 SLAM 算法在室外环境下的性能.

3) 跨场景与多平台数据集

ADVIO 数据集[48] 由阿尔托大学计算机科学系创建, 专注于视觉惯性里程计的实际应用. 涵盖了室内 (楼梯、扶梯、电梯、办公室) 与室外 (地铁站) 场景, 主要对象为行人. M2DGR 数据集[49] 是用于地面机器人导航的 SLAM 数据集, 采用了环视 RGB、红外、事件相机、32 线激光雷达、IMU、原始全球导航卫星系统 (global navigation satellite system, GNSS) 等各种各样的传感器及系统数据, 适用于室内、室外场景. RAWSEEDS 数据集[50] 涵盖从室内到室外的各种场景, 采集平台是机器人, 涵盖 LiDAR、MONO、Stereo、IMU、GNSS 等, 并提供各传感器采集到的数据, 用来验证 SLAM 算法在不同环境下的表现力. 该数据集包括多种场景的 RGB-D 图像序列、真实位姿及对应的真实深度信息. EuRoC MAV 数据集[51] 采用了 AscTecFirefly6 六旋翼无人机作为移动平台, 利用双目相机以及 IMU 采集数据, 包含室内、室外两种场景, 主要用于评估飞行器在不同环境下的 SLAM 算法性能. FusionPortableV2 数据集[52] 为提高泛化能力, 集成手持设备、腿式机器人、UGV、车辆这四类移动平台, 囊括了实验室、地下隧道、高速公路等共 12 种环境, 且基于高精度 RGB 点云地图和双模态真实轨迹支撑 SLAM 算法鲁棒性验证. M3DGR 数据集[53] 则专注于传感器退化鲁棒性问题, 系统性地设计了视觉遮挡、LiDAR 特征稀疏、轮式打滑、GNSS 拒止等四种失效场景, 利用双非重叠扫描 LiDAR 和全景相机等多源传感器构造的机器人传感器退化基准, 辅助进行 SLAM 算法失效恢复能力的量化评估.

上述数据集为动态环境下的视觉 SLAM 算法研究、性能评估及对比实验提供基准与实验支撑.

1.3 视觉 SLAM 的运动分割方法的分类与对比

根据对环境的预设条件, 将现有运动分割方法分为三类, 如图 2 所示, 即基于静态场景假设的运动分割方法、基于语义信息的运动分割方法和基于多传感器融合的运动分割方法. 同时, 三类运动分割方法的核心性能量化对比显示见图 3, 本量化对比依据各方法技术原理、优势与局限, 仅反映相对性能差异.

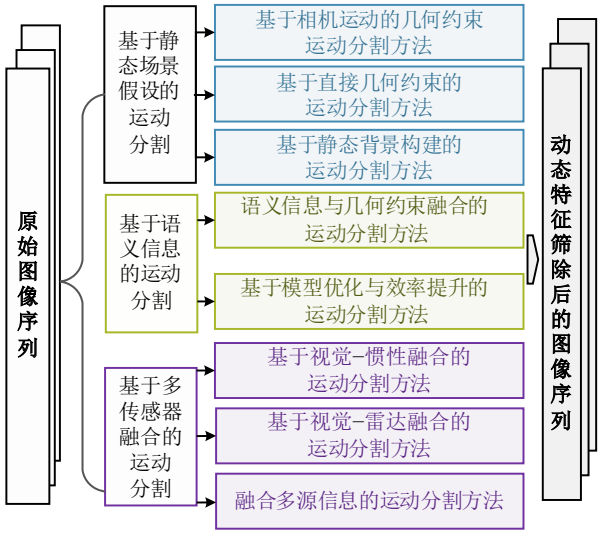


图 2 现有运动分割方法总结

Fig. 2 Summary of existing motion segmentation methods

静态场景假设方法在现实世界中除动态特征占图像主体的特殊场景外基本都可以适用, 但因为对于完全由动态目标组成的图像存在过拟合现象, 所以无法消除潜在运动物体的影响. 基于语义信息的方法无需预先假设运动特征占比, 但是利用深度学习网络模型需要进行大量的标注和训练, 在计算过程中代价较大. 随着传感器技术的发展, SLAM 任务中运用的传感器包括激光雷达、IMU、轮式里程计等, 同时联合多种传感器的方法也会有比较高的精度和鲁棒性[23]. 多传感器融合后的数据信息互补, 即不对环境预设也不依赖语义信息, 但因为加入传感器导致误差增大, 多传感器数据融合对于系统安装定位以及融合算法都有较高的要求, 此方法也无

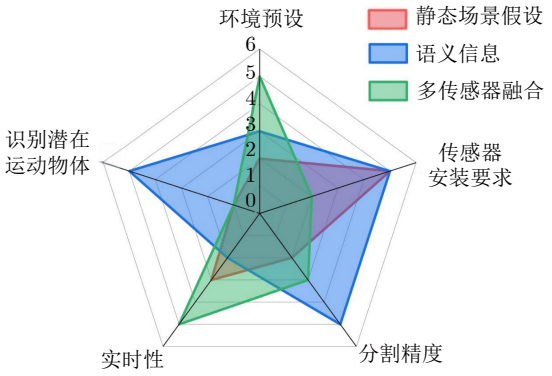


图 3　运动分割方法性能对比

Fig. 3　Performance comparison of motion segmentation methods

法去除潜在运动物体.

多传感器融合作为视觉感知的关键增强路径, 已成为提升运动分割精度与鲁棒性的核心技术, 与静态假设、语义信息方法共同构成视觉 SLAM 的运动分割范式. 因此, 本文将其与前两类方法并列阐述, 旨在完整呈现视觉 SLAM 运动分割从单视觉优化到多模态协同的技术全景. 此外, 伴随神经渲染技术的飞速发展, 特别是神经辐射场与 3D 高斯溅射的兴起, 一系列基于可微分渲染的 SLAM 系统 (如 iMAP、NICE-SLAM、Point-SLAM、SPARSE-GS 等) 为稠密建图与高保真场景重建开辟了新路径. 尽管场景表征形式发生了革命性变化, 但当前大多数基于可微分渲染的动态 SLAM 仍依赖于传统运动分割策略提供的先验信息或掩码. 简言之, 大多数可微分渲染 SLAM 系统都集成了某种形式的运动分割机制, 其本质仍未超出前述三大范式, 但其分割结果与神经场景优化过程更紧密地结合, 在高质量重建的同时对动态干扰具有良好的鲁棒性.

2　基于静态场景假设的运动分割

基于静态场景假设的运动分割方法以图像多数特征静止为依据进行运动与静止特征的区分, 并且相机帧间位姿估计和空间位姿图优化方法也建立在上述假设的基础上, 这类方法不需要太多先验语义信息和多传感器的支持, 计算更加轻量化. 然而基于静态场景假设的运动分割方法很难在高动态环境中实现较高精度的运动分割, 即存在因为大量静态特征被破坏而导致假设失效问题, 在高动态、复杂运动、遮挡、光照等因素的变化条件下, 其依靠的几何约束容易失效和变得模糊, 且难以解决潜在运动物体的问题, 即使想要提高鲁棒性也难以同时保证较好的性能. 因此, 这类基于静态假设的方法很难适用于行人、车辆密集、物体快速运动或者相机本身运动速度较高的场景.

近几年, 在改进几何约束优化效率、动态区域检测精度以及场景适应性等方面有了很大进步. 表 2 给出了基于静态场景假设方法的主要类别和特点, 这些方法多是以 ORB-SLAM 和 DVO-SLAM (dense visual odometry SLAM) 为基础. 表 3 给出了部分方法及特点, 需说明的是, 其中绝对轨迹均方根误差数据来自 TUM 数据集 fr3/walking xyz 图像序列测试结果; 表 3 中所列举方法的相机类型均为 RGB-D. 这一共性源于基于静态场景假设的运动分割方法在当前研究中多聚焦于 RGB-D 传感器的应用场景——其既能提供视觉纹理信息以支撑特征匹配, 又能通过深度数据辅助几何约束计算, 是该类方法实现动态特征检测与定位优化的常用硬件配置. “—”表示无法确定.

2.1　基于相机运动的几何约束运动分割方法

相机运动的几何约束本质上是静态点在相机运动下的投影变换规律, 其具体形式由相机模型和运动参数决定, 如对极约束、重投影误差约束、单应性约束等. 通过估计相机的运动参数 (如位姿、速度), 推断场景中物体的运动是否与相机运动一致, 从而分割动态区域. 即通过特征匹配、视觉里程计, 计算相邻帧间的相机位姿变换, 得到本质矩阵 E、基础矩阵 F 或单应矩阵 H 等几何模型参数. 然后对场景中的每个特征点, 计算其与上述几何约束的符合程度, 从而划分静态与动态, 通过设定阈值, 将残差小于阈值的点判定为满足约束, 归为静态物体; 残差大于阈值的点判定为违背约束, 归为动态物体. 然而, 在高动态环境中, 初始变换估计会受到主要动态特征的影响.

如图 4 所示, 设相机从位姿 C_1 运动到 C_2, p_1、p_2 表示特征点 P 在成像平面的投影点, 其旋转矩阵

表 2　静态场景假设方法的分类及特点

Table 2　Categories and characteristics of static-scene-assumption methods

类别	优点	缺点
基于相机运动的几何约束运动分割方法	通过几何约束区分静态与动态特征, 在低动态环境下精度较高	高动态环境下精度显著下降
基于直接几何约束的运动分割方法	无需估计相机运动, 计算量小且实时性高	动态特征占比高时分割精度低
基于静态背景构建的运动分割方法	高动态场景精度较高	计算复杂

表 3 部分基于静态场景假设的运动分割方法

Table 3 Partially static-scene-assumption-based motion segmentation methods

方法	绝对轨迹均方根误差 (m)	相机类型	运行环境	基础框架	单帧跟踪时间 (ms)
Dou 等[54]	0.01138	RGB-D	i7-12700K + 32 GB	ORB-SLAM3	14.153
SamSLAM[55]	0.19870	RGB-D	i7-12850HX + 32 GB	ORB-SLAM3	—
DZ-SLAM[56]	0.01300	RGB-D	i7-12700K + 32 GB	ORB-SLAM3	231
GMP-SLAM[57]	0.01300	RGB-D	i7-9750H + 16 GB	ORB-SLAM2	50
Wang 等[58]	0.01480	RGB-D	i5-7500HQ + 16 GB	ORB-SLAM3	54
RED-SLAM[59]	0.01480	RGB-D	i7-12700KF + 16 GB	ORB-SLAM3	26.88
Zhang 等[60]	0.01300	RGB-D	i7-13700K + 64 GB	ORB-SLAM2	169
DyGS-SLAM[61]	0.01400	RGB-D	i7-9750H + 16 GB	ORB-SLAM3	50
FD-SLAM[62]	0.01620	RGB-D	i7 + 16 GB	ORB-SLAM3	355.18
FND-SLAM[63]	0.01300	RGB-D	i7-13650HX	ESLAM	812.32
PLFF-SLAM[64]	0.08600	RGB-D	i7-13650HX	ORB-SLAM3	—
Qi 等[65]	0.01470	RGB-D	i7-12700 + RTX 4060 Ti	ORB-SLAM2	117.8
程鹏等[66]	0.01300	RGB-D	i7-13700K + 64 GB	Manhattan-SLAM	169
李泳等[67]	0.01490	RGB-D	R9 + RTX 3060	ORB-SLAM3	86

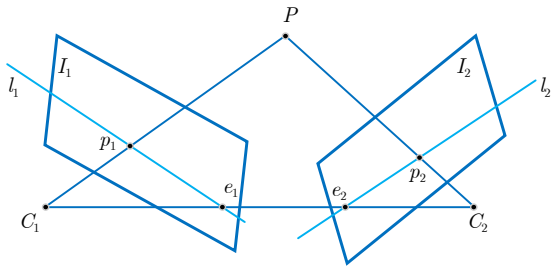


图 4 静态环境下的对极约束

Fig. 4 Epipolar constraint in static environment

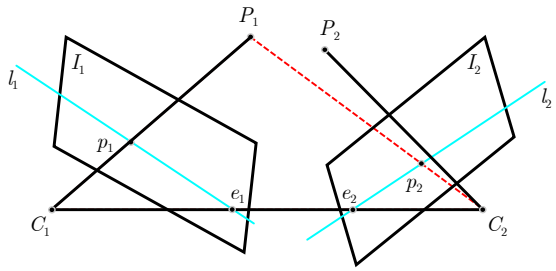


图 5 动态环境下对极约束

Fig. 5 Epipolar constraint in dynamic environment

为 $R \in SO(3)$, 平移向量为 $\boldsymbol{t} \in \mathbf{R}^3$. 根据定义, 本质矩阵 E 为:

$$E = \boldsymbol{t}^{\wedge} R \in \mathbf{R}^{3\times 3} \tag{5}$$

其中, $\boldsymbol{t}^{\wedge}$ 表示相机从 C_1 运动到 C_2 平移向量 $\boldsymbol{t}$ 的反对称矩阵.

对于本质矩阵 E, 满足对极约束:

$$\boldsymbol{x}_2^{\mathrm{T}} E \boldsymbol{x}_1 = 0 \tag{6}$$

其中, $\boldsymbol{x}_1$、$\boldsymbol{x}_2$ 表示相机 C_1 和 C_2 分别到点 P 的向量.

如图 5 所示, 相机在不同位姿 C_1、C_2 对特征点 P 观测时, 特征点 P 从 P_1 移动到 P_2 处, 其真实三维位置在两帧间发生变化, 式 (6) 中向量 $\boldsymbol{x}_1$、$\boldsymbol{x}_2$ 与 $\boldsymbol{t}$ 不共面, 因而不满足式 (5), 从而不能以此识别动态特征点.

由于在 V-SLAM 中, 相邻帧的特征点匹配常因遮挡、纹理重复等因素, 存在误匹配, 而随机抽样一致性 (random sample consensus, RANSAC) 算法能在包含大量异常值的情况下稳健估计几何模型, 因此被广泛用于剔除这些误匹配点对. 但是其算法复杂, 迭代次数过多等问题使得该算法在硬件实现及系统集成中面临挑战.

谢颖等[68] 针对双目视觉 SLAM 左右目图像关键点匹配误差导致定位精度下降的问题, 提出了一种融合投影变换的光流算法. 该算法首先通过正反 LK (Lucas-Kanade) 光流计算左目角点在右目的初始位置, 并利用 RANSAC 算法生成单应矩阵进行投影变换, 再通过光流跟踪将投影图像点映射回原始右目图像, 从而优化匹配精度. He 等[69] 利用单应矩阵描述两帧图像间的投影变换关系, 通过迭代随机采样计算模型参数, 以切比雪夫距离作为内点判定准则, 剔除动态物体的误匹配点, 将单应矩阵作为假设模型, 提出了基于单应矩阵的 RANSAC 硬件加速设计. 针对动态环境, Dou 等[54] 提出 ST-RANSAC, 通过多帧间的极线约束和时空一致性检查, 区分静态与动态特征点. Li 等[70] 研究三视张量与已知垂直方向的相对位姿估计, 利用 IMU 提供的垂直方向信息, 将旋转矩阵简化为仅含偏航角的参数形式, 提出 4 点线性解法和 3 点最小解法. 由

于所提方法所需匹配点更少, 可高效应用于 RANSAC 框架中, 用于视觉里程计中的外点剔除和位姿估计. 在算法效率与精度平衡方面, Yang 等[71] 提出 GMS-ATRANSAC 方法, 首先利用网格运动统计进行粗筛选, 通过网格单元内的特征匹配密度剔除明显误匹配, 再采用自适应阈值优化模型.

另外, 融合策略成为几何约束方法的重要补充. 杨永刚等[72] 利用光流法给出相机运动的预测值, 特征点匹配给出观测值, 将光度一致性约束和特征点的几何约束两者结合, 用卡尔曼增益进行滤波融合, 使得这两个约束结合在一起并实现了对这两类约束的协同修正. 将多源几何约束信息融合进滤波框架中能够提高几何约束方法在低纹理或存在较大光照变化场景下的鲁棒性.

基于相机运动的几何约束分割方法通过静态点必满足相机运动几何约束、动态点必违背的核心逻辑, 实现了动态场景中静态与动态物体的有效区分, 其具体的实现流程为: 首先借助 RANSAC 算法估计相邻帧间的相机运动参数, 而后依据几何约束剔除与相机运动估计不匹配的动态特征点. RANSAC 算法具有非确定性, 其通过随机抽样内点估计模型的机制, 使得最终得到的运动估计结果在概率上满足对多数特征点对的适配性. 为提升获得可靠结果的概率, 需增加算法的迭代次数. 与直接基于几何约束的方法相比, 此类方法因引入了显式的相机运动估计, 在低动态环境下能够获得更高的分割精度. 然而, 其计算成本也相应增加, 主要源于 RANSAC 算法的迭代过程对计算资源的消耗.

综上所述, RANSAC 算法的有效性依赖于场景中静止特征点占主导这一前提假设. 当动态特征点成为图像中的主要成分时, 该假设被打破, 相机运动估计易受动态特征干扰而产生偏差, 进而导致运动分割结果出现显著误差.

2.2　基于直接几何约束的运动分割方法

直接几何约束方法同样以静态场景为前提假设, 其核心特征在于不进行相机运动参数的估计, 而是直接利用几何约束准则对动态特征点进行识别, 进而实现静态与动态特征的分割. 此类方法所依据的几何约束可从极线约束、三角剖分关系、基础矩阵估计结果或重投影误差方程等多种途径推导得出, 其动态特征点的剔除流程如图 6 所示. 与基于语义信息的分割方法相比, 虽然直接几何约束方法具有无需依赖大规模标注训练数据的优势, 且不存在语义分析模型计算开销大的问题, 具备较高的运行效率和较低的计算复杂度, 但该方法也因缺乏

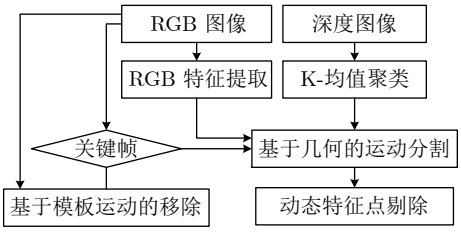


图 6　基于几何约束的动态特征点移除流程图

Fig. 6　Flowchart of dynamic feature point removal based on geometric constraints

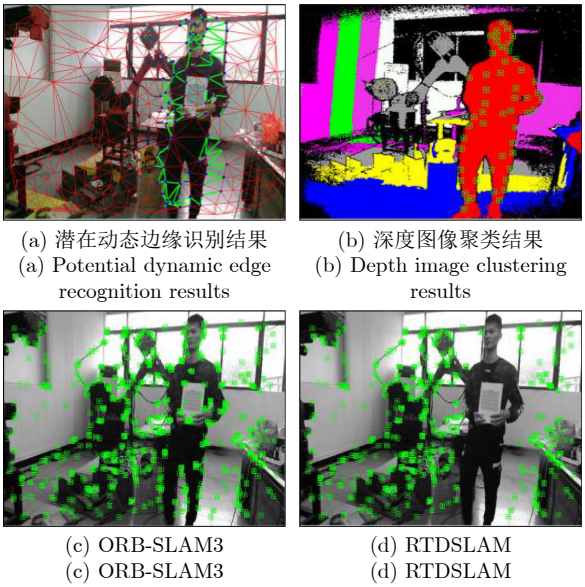
语义信息的支撑, 无法对潜在的运动物体进行识别.

Chen 等[56] 提出的 DZ-SLAM 是一种不需要任何先验知识处理环境中动态特征的方法, 首先利用 FastSAM[73] 将图像分割成多个掩码, 然后采用一种基于自适应阈值的密集光流方法 (FlowNet2) 检测图像中的动态区域, 并将获得的光流信息与 Fast-SAM 生成的掩码进行融合, 构建包含动态特征的联合掩码, 用于提取环境中的动态目标, 从而提升定位精度. Hu 等[57]设计了一种使用 3D 占用网格进行动态特征检测的算法 GMP-SLAM, 首先将相机周围 3D 空间划分为均匀体素, 以相机为中心建立网格地图, 通过 GPU 并行投影, 将体素中心投影至图像平面, 计算投影深度与实际深度的误差, 对比体素在相邻帧的投影误差, 通过欧氏距离误差识别动态点, 更新体素的占据状态, 以处理动态环境. Wang 等[58] 将静态环境视为虚拟刚体, 首先利用 Delaunay 三角测量算法剖分序列图像, 构建特征点空间关系, 计算相邻关键帧中边距变化率, 通过自适应阈值识别动态边缘, 再对深度图像进行 K-均值聚类, 将稀疏动态点投影到深度聚类区域, 定位完整动态区域, 解决了非刚体性物体分割不完整问题; 在非关键帧中, 利用模板匹配快速跟踪已识别的动态区域, 仅在关键帧更新动态模板. 图 7 给出了 RTDSLAM 系统在真实动态环境中的实验结果示意图. 图 7(a) 和图 7(b) 分别展示了潜在动态边缘的识别和深度图像上的聚类结果, 图 7(c) 和图 7(d) 分别是 ORB-SLAM3 和 RTDSLAM 提取的特征点.

这类方法无需对相机运动进行求解, 而是直接通过对比相机相邻帧来估计运动物体所在区域. 其优势在于结构轻量化、实时性能较好且计算量较小. 不过, 该类方法同样以图像中静态特征点占多数为前提假设, 因此在动态特征点在图像特征中占比较高的场景中, 其分割精度会明显下降.

2.3　基于静态背景构建的运动分割方法

基于静态背景构建的方法需借助特定假设或先验信息构建静态背景模型, 并根据输入帧与背景图



(a) 潜在动态边缘识别结果
(a) Potential dynamic edge recognition results

(b) 深度图像聚类结果
(b) Depth image clustering results

(c) ORB-SLAM3
(c) ORB-SLAM3

(d) RTDSLAM
(d) RTDSLAM

图 7 RTDSLAM 系统在真实动态环境中的实验结果

Fig. 7 Experimental results of the RTDSLAM system in a real dynamic environment

像的差别来进行阈值设置, 检测前景的运动目标. 此类方法是从当前帧里扣除掉静止背景而呈现该时刻发生的运动变化, 所以这类方法的算法结构简单且计算效率较高. 但是由于这种方法对环境的光强变化的适应性较弱, 在确定阈值的时候会受背景强度变化的影响, 对于背景和目标相互交替的场景会导致检测结果不稳定, 不能满足实际需求. 在现实场景中, 比如拍摄镜头有位移、环境中有物品晃动、光照强度变化、相机自身运动等都会产生大量动态干扰, 这极大地增加了建立并维护精准背景模型的难度. 因此, 静态背景构建方法在单独处理运动分割任务时, 其鲁棒性与效果往往受到限制, 在实际应用中通常需要与其他技术结合, 以提升整体性能.

李嘉辉等[74] 先用语义几何联合约束找到动态对象, 再采用关键帧加权映射策略, 将 RGB 图像与深度图像结合起来做修复, 最后再用区域生长算法对深度图做补充修复. Zhang 等[60] 将 Inpainting SLAM 方法加入图像修复模块恢复被动态物体遮挡的静态部分, 并且其图像修复模块基于 E2F-GVI 框架是一个利用多帧时空信息实现的视频修复方法, 能够把几帧的光流向量视作一系列对应的连续图层, 从而得到较为完整的背景场景. Luo 等[62] 提出的 FD-SLAM 方法利用基于深度学习的掩码图像修复网络 DeepFillv2 实现单帧图像的背景修复, 在移除动态物体的基础上保留完整环境信息用作 SLAM 后续步骤. Zhu 等[61] 利用 EfficientSAM 语义分割及多视图几何约束和深度重投影误差来确定动态点, 通过 YOLO-World 提供开集词典的检测框, 对分割结果进行修正. 利用修复后的静态背景图像建立 3D 高斯辐射场, 并通过可微渲染优化静态场景表示, 去除动态物体对场景表征的影响. Yang 等[63] 提出的 FND-SLAM 采用基于快速傅里叶卷积 (fast Fourier convolution, FFC) 的背景补全策略, 利用条件随机场模型生成掩码, 然后将掩码与关键帧输出合并, 形成一个四通道的张量, 将张量通过快速傅里叶卷积处理得到修复图像, 最终将掩码与修复后的图像相结合, 完成背景区域的修复.

针对动态场景中动态物体遮挡致背景缺失、修复空洞及移除不彻底等问题, 相关方法借隐式辐射场建模静态场景几何与外观特征, 实现动态物体分割与剔除. 其中, DN-SLAM[75] 利用 NeRF 新视角合成能力修复静态背景, 动态物体离开遮挡区域后, 系统依据优化后的静态辐射场参数, 以多视图一致性约束渲染未遮挡背景像素, 填补地图空洞. 实验中, 其在 TUM 数据集“fr3/walking-half”序列修复行人遮挡的桌面、墙面等区域, 静态背景完整性较传统方法提升约 35%; Bonn 数据集“crowd”序列中动态区域占比超 50% 时, 仍能通过动态掩码筛选与 NeRF 修复保持建图连续性与准确性. Xu 等[76] 提出的 DIN-SLAM, 通过过往视角的静态信息修复遮挡背景, 结合前后帧已知位置, 将先验关键帧投影到当前帧分割区域的 RGB 与深度图像, 合成无动态物体的真实图像以补充场景信息, 提升建图与跟踪性能. 对未出现或缺乏有效深度信息的未填充区域, 通过整体场景表征优化减少空洞, 在自采集大型动态场景中实现动态物体移除与分割区域良好修复. DDN-SLAM[77] 则针对跟踪漂移与重建伪影问题优化, 以 YOLOv5 获取动态目标边界框先验, 结合混合高斯模型对框内像素深度建模, 计算前景/背景后验概率实现特征点分割, 再经重投影误差检查恢复误移除的潜在静态点. 采用光学流动态判断与补全策略, 仅对 NeRF 相关关键帧做动态像素分割并填充低动态区域, 同时引入稀疏点云引导的跳采样机制提高静态表面重建质量, 有效恢复遮挡背景.

为实现高质量的运动分割效果, 模型本身的质量是关键因素之一, 而将场景深度信息和几何特征结合的运动分割方法比单独依靠几何约束的方法更加鲁棒, 但是会造成移动平台的计算量更大.

3 基于语义信息的运动分割

近几年来, 深度学习以及计算机视觉的深度融合在诸多任务中取得了优异效果, 涌现了很多优秀的深度学习网络, 如 YOLO[78]、SegNet[79]、Mask R-

CNN[80]、YOLACT[81] 等, 表 4 总结了部分具有代表性的计算机视觉算法及其贡献. 得益于深度学习在图像领域取得飞速进展, 基于语义信息的运动分割方法已被应用到动态 SLAM 中, 已有许多优秀语义视觉 SLAM 系统被研发出来, 如 DS-SLAM[82]、DynaSLAM[83]、Detect-SLAM[84]、PSPNet-SLAM[85]、SG-SLAM[86]、GSL-VO[87] 等. 与经典 ORB-SLAM2/3 相比, 其精度更好, 在动态环境下提高了传统 SLAM 系统的鲁棒性, 使 SLAM 发展向更加智能、更加精准的方向前进, 为机器人和自动驾驶车辆在复杂动态环境下的导航定位提供了新思路. 图 8 给出了通用语义增强动态 SLAM 框架, 按所采用的深度学习网络模型的角色, 基于语义信息的方法主要分为基于目标检测和基于图像分割的方法. 通过将采集到的语义信息作为判断动态目标的先验知识, 对特征点进行语义标注, 能够用语义信息去除背景当中移动的目标.

3.1　语义信息与几何约束融合的运动分割方法

语义信息与几何约束融合的运动分割方法通过将深度学习方法的语义信息和传统几何约束相结合, 可以很好地解决在高动态场景下动态物体的检

表 4　代表性计算机视觉算法

Table 4　Representative computer vision algorithms

领域	方法模型	年份	贡献
目标检测	R-CNN[88]	2014	首次将 CNN 引入目标检测领域
	Fast R-CNN[89]	2015	引入感兴趣区域池化层实现特征共享, 结合多任务损失端到端训练
	Faster R-CNN[90]	2015	提出区域建议网络替代选择性搜索, 实现候选框生成与检测一体化
	SSD[91]	2016	结合多尺度特征图预测与预设锚框机制
	YOLO[78]	2016	开创单阶段检测范式
	YOLOv5[92]	2024	集成自适应锚框计算、Mosaic 自适应数据增强及自适应图片缩放
	RT-DETRv3[93]	2025	设计了层次密集正监督方法, 通过 CNN 辅助分支和自注意力扰动策略
	YOLOv13[94]	2025	引入超图自适应相关增强机制, 通过自适应超图计算建模高阶视觉相关性
图像分割	FCN[95]	2015	首创全卷积网络结构
	SegNet[79]	2017	利用最大池化索引实现高效上采样
	Mask R-CNN[80]	2017	在 Faster R-CNN 基础上添加掩码分支并设计 RoIAlign 层消除量化误差
	DeepLabv3[96]	2017	改进空洞空间金字塔池化并引入图像级特征
	YOLACT[81]	2019	提出原型掩码生成与掩码系数预测的并行分支结构
	SegFormer[97]	2021	结合无位置编码的分层 Transformer 编码器和全 MLP 解码器
	Mask2Former[98]	2022	提出通用图像分割架构, 引入掩码注意力机制和高效的多尺度策略
	Segment Anything[99]	2023	定义可提示分割任务, 支持点/框/文本等任意提示输入
	MagNet[100]	2024	设计跨模态对齐损失函数和对齐模块, 缩小语言–图像模态差距
	OMG-Seg[101]	2024	整合多领域分割任务, 降低计算和参数开销
	GleSAM[102]	2025	利用生成式潜在空间增强提高对低质量图像的鲁棒性

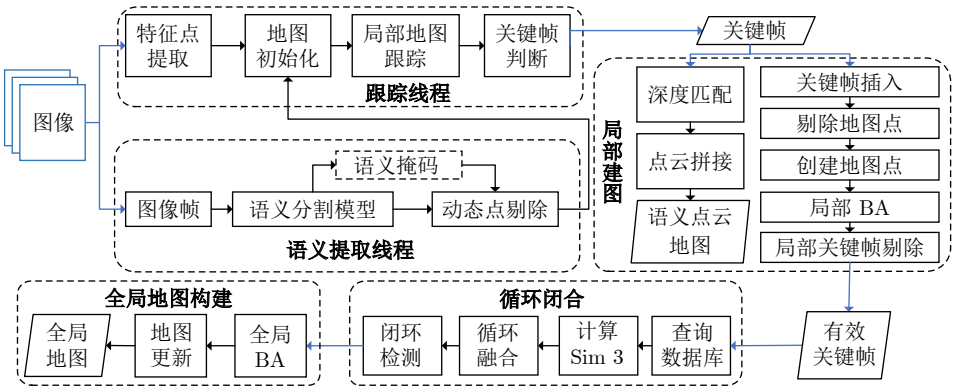


图 8　通用语义增强的动态 SLAM 框架

Fig. 8　General semantic-enhanced dynamic SLAM framework

测, 而且能在实时运行中达到有效的检测和去除静态噪声的目的. 因为这类方法利用了图像中的语义信息和几何信息, 所以相比于单靠几何信息作为检测目标会更加准确, 并且可以避免单纯依靠几何信息无法正确判定出动态物体的问题.

Yu 等[82] 提出的 DS-SLAM 系统是在 ORB-SLAM2 框架的基础上构建的, 在利用 SegNet 完成语义分割并去掉行人等动态目标之后, 再把图像划分成不同的语义区域, 最后用 RANSAC 验证了相机运动以及各个语义区域内的特征点运动的一致性, 只留下了静态区域的特征来进行定位与建图. DynaSLAM 是 Bescos 等[83] 提出的一种支持单目、双目和 RGB-D 三种输入方式的 SLAM 方法, 它在单目和双目模式下处理过程与 DS-SLAM 类似. 首先根据语义分割存在的问题获取动态区域先验, 再根据图像上的深度图对误分点进行修正; 其次基于深度图进一步扩大潜在动态区域的范围, 并移除部分可能存在的动态物体; 最后运用背景模型的信息填充动态区域, 使其能得到较好的地图效果. 类似地, 文献 [103] 也采取同样方法, 首先基于 Mask R-CNN 的语义分割得到场景的动态先验信息, 再通过图像深度信息修正分割区域边缘, 最终区分动态环境的前景特征点和背景特征点.

与此同时, Xiao 等[104] 针对 DyGS 目标检测网络存在的漏检问题, 设计了 Dynamic-SLAM 方法. 该方法基于运动目标速度保持恒定的假设, 在图像序列跟踪过程中引入目标跟踪策略, 从而有效降低漏检率: 在对图像序列中的运动目标进行跟踪时, 考虑到相邻帧之间时间间隔较小, 通过设置前后两帧运动目标之间的像素偏移阈值来判断是否发生漏检; 同时, 利用先验语义信息判别特征点的运动特性, 将运动特性不满足时间不变假设的特征点视为动态特征并予以剔除, 最终利用非动态特征完成位姿求解. 在实验配置 (i5-7300HQ, GTX 1050 Ti) 下, 该方法的平均处理速度达到 22.22 fps.

为了保证 V-SLAM 系统能在高动态场景下正常运行, Jia 等[105] 首先利用实例分割模型 FastInst 检测场景中的潜在移动目标, 并基于分割掩码滤除关键点, 以保证保留下来的特征点具有足够的稳定性用于后续跟踪; 随后, 采用一种轻量级的对象关联方法, 仅利用少量稳定特征点实现低成本跟踪, 并估计当前帧的初始相机姿态; 接着, 计算上一帧图像中 3D 对象点的深度信息, 并将其投影至当前帧图像中, 通过描述子匹配实现关键点的跨帧关联; 在此基础上, 利用运动一致性校验获取对象的真实运动状态, 并对提取的特征点进行重新筛选, 最终保留位于高质量静态对象上的特征点.

为了解决传统方法利用对极约束处理平行运动时存在失效的问题, Pan 等[106] 提出的 DEG-SLAM 引入一种退化约束机制, 并以 ORB-SLAM3 系统为基础建立模型: 首先, 通过该机制减少动态特征点的数量; 随后, 利用 YOLOv5 目标检测网络识别动态物体, 并将其语义信息传递至跟踪模块; 最后, 在跟踪过程中结合语义信息与极线约束, 共同过滤动态特征点.

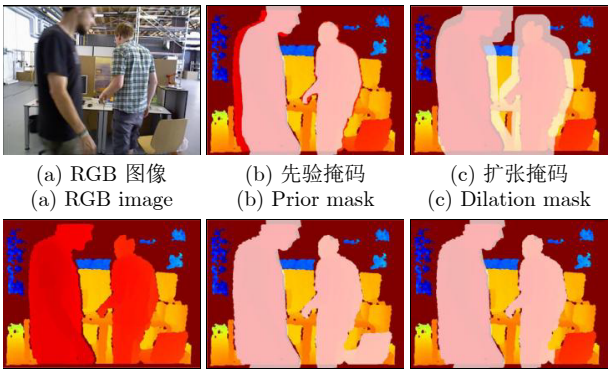
针对动态环境中移动物体导致传统 SLAM 定位精度骤降的问题, Yang 等[107] 首次融合 SOLOv2 语义分割与 DeepSort 目标跟踪方法: 通过 SOLOv2 实时生成像素级实例分割掩码, 预定义高动态类别并提取物体轮廓, 计算特征点与轮廓的空间关系, 若特征点在动态掩码内则直接删除; 对于潜在动态物体, 则利用对极约束进行验证; 随后, 通过 DeepSort 多目标跟踪漏检帧, 并在漏检帧中结合卡尔曼滤波预测目标位置, 实现动态特征点的二次剔除.

对于视觉 SLAM 在动态环境中无法过滤掉运动目标的问题, Xu 等[108] 提出一种基于特征点层次多维聚类的视觉 SLAM 系统, 简称 HMC-SLAM. 他们把特征点极线距离和深度信息融合起来对运动点进行聚类, 通过识别最强动态簇找到周围相邻的点簇进行去除, 并基于 YOLOv5 检测框建立动态滤波器. 在运动检测框内定义特征点来优化权值, 利用深度差值作为高斯分布的概率模型, 联合优化摄像机位姿和特征点置信度, 抑制潜在运动点的影响, 在 Inteli5 + RTX 4070 Ti 运行环境下平均每帧处理时间为 51.02 ms, 实时跟踪率达 94.8%, 高于 ORB-SLAM3 算法. 在 TUM 和 Bonn 数据集上的绝对轨迹误差相较 ORB-SLAM3 均有所提高. Cui 等[109] 提出的 DYMRO-SLAM 引入了新的目标检测线程, 能够识别、分割动态目标, 实现双线程处理. 前端采用 Mask R-CNN 实例分割得到动态掩码, 后端设计特征点–光流融合跟踪模块: 关键帧用 ORB 特征匹配, 非关键帧用 LK 光流跟踪, 减少描述子的运算量. 根据光流误差和重投影误差的联合目标函数, 利用 Levenberg-Marquardt 算法同步优化相机位姿、动态点权重. 在 RTX 4070 Ti 上实现了 34.67% 的速度提升, 关键帧跟踪率达到 100%, 解决了视觉 SLAM 系统掩码覆盖不准确、小区域遗漏的问题. Li 等[110] 提出了基于 YOLOv8 算法生成初始掩码的动态 SLAM 方法, 简称 DYR-SLAM. 该方法利用深度相似性函数对多帧深度信息进行融合, 然后通过权值修正掩码覆盖率, 引入了加速度

约束模型来预测动态轨迹并结合深度方差检测遮挡, 设置自适应剔除权重, 根据运动一致性动态过滤点云, 避免误删静态物体. 相比缺少融合误差项的 YOLO-SLAM, 其点云融合误差下降了 38.36%, 在 TUM 数据集的低动态场景 (fr2/desk/p) 中的均方误差为 0.065 m. 为了解决在动态环境下的特征点缺失的问题, YPL-SLAM[111] 使用 YOLOv5s 目标检测进行动态/潜在区域掩码, 对极几何约束验证潜在动态物体状态, 直接剔除动态区域特征点, 同时提取非动态区域线特征, 并利用图像动态分数和点线匹配成功数量设计加权融合策略优化点线特征匹配. 在 NVIDIA RTX 2060 硬件上, 目标检测–线特征融合–动态区域验证机制使高动态场景下的定位准确性相对于 ORB-SLAM2 最多提高了 96.1%.

可微分渲染与神经隐式表征的引入, 为语义 SLAM 提供了超越传统二维分割的、具有内在三维一致性的场景理解能力. 但现有基于 NeRF 的 SLAM (如 iMAP 或 NICE-SLAM) 虽能重建稠密地图, 但对动态对象敏感, 且缺乏实时处理能力. SemGauss-SLAM[112] 创新性地将语义特征嵌入 3D 高斯表征, 实现语义与几何的深度绑定及动态分割优化. 其核心设计包括语义高斯表征与语义引导优化两大模块: 在表征层面, 为每个 3D 高斯添加 16 维语义特征嵌入, 将语义信息与高斯的位置、颜色、透明度等几何属性关联, 形成语义–几何融合的高斯地图; 在优化层面, 引入特征损失函数, 强制渲染的语义特征与真实语义特征对齐, 同时提出语义引导约束调整策略, 利用多帧共视区域的语义一致性构建约束, 联合优化相机位姿与高斯表征, 自动剔除动态区域导致的语义不一致高斯点. 针对由于动态物体造成的地图定位漂移问题, Liu 等[113] 提出了基于三维高斯喷溅语义驱动 SLAM 系统 SDD-SLAM. SDD-SLAM 协同使用 Grounding DINO、SAM-Track 生成对象级语义掩码以及深度聚类精炼掩码.

由于深度图像中存在边缘误差, 首先通过边缘噪声滤波剔除重投影点云中的异常值, 随后结合中心偏移量与语义损失变化检测动态物体. 在此基础上, 引入一种对象级高斯密度控制策略, 仅在多视图射线交点处增加高斯椭球密度, 从而减小深度异常. 图 9 展示将精炼掩码应用于深度图像的效果.



(a) RGB 图像 (a) RGB image　(b) 先验掩码 (b) Prior mask　(c) 扩张掩码 (c) Dilation mask

(d) 深度图像 (d) Depth image　(e) 深度掩码 (e) Depth mask　(f) 精炼掩码 (f) Refined mask

图 9　基于深度聚类与形态学约束的精炼掩码生成

Fig. 9　Refined mask generation based on deep clustering and morphological constraints

3.2　基于模型优化与效率提升的运动分割方法

尽管基于语义的分割方法在动态场景下具有较高的分割精度, 但由于此类方法的计算量大, 在实际运用中面临算法如何权衡精确度与计算效率的问题. 针对此问题, 目前效率提升方法主要从两个角度来进行: 第一种是提高模型的效率; 第二种是降低语义提取频率.

提升模型效率的关键在于采用轻量化深度网络以提高分割速度. Wu 等[114] 提出的 YOLO-SLAM 系统, 构建了轻量级目标检测网络 Darknet19-YOLOv3. 该网络基于室内环境中动态物体多为人形、人体在深度图中近似平面、检测框内人体深度与周边环境差异显著这三项前提, 结合深度差与 RANSAC 算法, 分别设计了适用于 DS-SLAM 和直接映射的动态特征筛选策略. 实验结果显示, 该系统在 TUM 和 Bonn 数据集上的 RMSE 均低于 DS-SLAM, 但在无 GPU 支持的环境下无法实现实时运行. 为进一步降低运动分割的计算开销, 可让 SLAM 系统仅对部分图像帧执行语义分析, 这一策略能显著减轻系统的计算负担.

传统的动态 SLAM 方法如 DS-SLAM、Dyna-SLAM 均使用像素级语义分割网络, 造成单帧耗时较多. 其中以 DynaSLAM 为例, 其平均处理时间大于 300 ms, 难以满足 SLAM 系统的实时性需求, 因此有必要在算法上进行改进, 打破现有瓶颈. Liu 等[115] 基于 ORB-SLAM3 提出了 RDS-SLAM (real-time dynamic SLAM), 通过双线程并行架构优化效率, 语义分割线程与语义优化线程互不干扰, 不会出现阻塞主跟踪进程的情况, 应用动态概率模型将地图点移动的概率量化的思想, 动态筛选进入位姿估计及回环检测过程中的特征点. 采取语义关键帧双向调度的方式, 优先处理队列首尾关键帧, 同时满足了跟踪线程实时要求及回环检测时的语义信息完整性. 从 RDS-SLAM 系统在 TUM 数据集上的表现来看, 其定位精度维持了与一般动态 SLAM 相近水平的同时, 将处理时间缩减到了 22 ~ 30 ms. 为了进一步改善性能, RDMO-SLAM[116] 在 RDS-SLAM 的双线程并行架构基础上, 新增光流–速度估计线

程, 形成三线程协同框架. 语义分割线程通过 Mask R-CNN 生成关键帧语义标签, 避免阻塞主跟踪进程; 光流−速度估计线程利用稠密光流预测场景流, 计算地图点速度以筛选动态特征; 主跟踪线程结合前两个线程输出, 通过动态概率模型量化地图点移动概率, 实现位姿估计与动态剔除的高效协同. 该架构继承 ORB-SLAM3 的核心框架, 通过多线程解耦与信息互补, 将系统帧率提升至 30 fps, 解决了传统语义 SLAM 实时性不足的问题.

文献 [117] 以 ORB-SLAM3 为基础, 在改进的 RT-DETR 目标检测算法的基础上进行动态区域划分, 使用对极约束与光流法筛选特征点. 利用对极约束算法中计算特征点到极线距离的方式, 删除超过 2 个像素阈值的动态点. 光流法则联合金字塔 LK 光流算法, 通过非动态区光流的均值及方差为动态点的筛选设置离群点判断准则, 从而保留静止/动态物体以及背景点. 程强等[118] 基于改进的 YOLOX-S 算法提出一种动态 SLAM 方法, 将 SPP (spatial pyramid pooling) 和 AFF (adaptive feature fusion) 融合加入目标检测模块, 使用运动框自适应阈值判定算法来区分动态物体与潜在动态物体. 针对目标类别确定离群点阈值, 若动态框内离群值数量或离群值占光流跟踪特征点总数的比例超过对应阈值, 则判定为运动状态并保留动态框内的背景模型点. 黄友锐等[119] 通过引入 Ghost 模块压缩参数量、添加 SimAM 注意力机制, 设计了一种轻量化 YOLOv5s 目标检测网络; 将动态特征剔除算法用 LK 光流矢量阈值法表示, 计算动态框内特征点光流与非动态区域平均光流的距离, 若绝对值大于阈值 L 即认为是动态点. 上述三者均通过轻量化的目标检测来划分动态区域, 但是所采用的动态点筛选方法不一样.

为提高动态环境下 SLAM 的鲁棒性和低纹理环境下的特征提取能力, Cheng 等[120] 设计了 DFE-SLAM, 并且提出了三种新机制: 一是整合轻量级 YOLOv5s 动态目标检测网络以实时识别并排除动态物体, 提升动态环境下的性能; 二是采用基于灰度对比分析的自适应阈值技术, 优化 FAST (features from accelerated segment test) 特征点检测, 增强低纹理场景下的定位精度与稳定性; 三是新增基于点云库 (point cloud library, PCL) 的稠密建图线程, 利用关键帧及位姿数据生成详细的三维稠密点云地图, 克服了 ORB-SLAM3 仅能生成稀疏特征图的局限. 同样是使用轻量级的 YOLOv5s 模型, Feng 等[121] 设计的 YLS-SLAM 使用 YOLOv5s-seg 完成动态目标粗分割, 优先选择分割高动态对象, 将语义掩码边缘通过掩码膨胀运算进一步拓展动态区的范围; 同时结合边缘增强技术细化得到更为准确的动态−静态边界, 从而能够更高效地剔除残余特征点, 很好地解决了工业动态场景下语义掩码边缘信息不全造成的轨迹偏移问题. 为了解决语义分割网络计算复杂度过大的问题, Lian 等[122] 提出了轻量化端到端的 LSSMask 方案, 并设计 GDS-ECA 卷积模块将深度可分离卷积和高效的通道注意力 ECA 结合, 用 Ghost 卷积生成冗余特征以降低参数量同时增强特征表达力, 构造了 BGTNet 提取网络, 在瓶颈层嵌入 GDS-ECA 模块代替标准卷积单元, 开发了轻量级特征金字塔网络 (lightweight feature pyramid network, L-FPN) 结构, 采用与高层语义提取相同的卷积提高语义信息. 在 COCO 数据集上实现 58.82 fps 的实时分割, 模型参数量仅 4.2 M, 平均精度达 35.8%, 较 MobileNetV3 提升 1.5%, 兼顾分割速度与精度需求. Zhang 等[123] 设计 UE-SLAM, 结合多源深度信息集成框架, 利用 DINOv2 语义分割提取动态目标的信息, 融合单目深度估计、辐射场渲染深度以及 MLP 不确定性模型生成精修代理深度, 基于三平面法将几何、外观、语义信息编码到统一特征向量, 引入对称跟踪机制利用代理深度优化位姿估计. GPU 占用内存仅 3.9 G, 单帧处理耗时 26.7 ms, 在无深度传感器的情况下也可以构建高质量的语义地图.

表 5 展示了部分基于语义信息的动态 SLAM 方法及其特性. 从表 5 中能够发现, 将语义信息分割方法与几何方法相融合, 可增强系统的鲁棒性. 同时, SLAM 系统的运动分割通过优化模型以及降低语义帧提取频率, 能够减轻系统的计算负担. 总体来看, 这类方法在一定程度上缓解了实时性方面的难题, 让基于语义信息的运动分割方法更具实际应用价值. 不过, 在模型轻量化的过程中, 可能会损失部分分割精度; 在降低语义提取频率时, 如何保障语义信息在普通帧传播过程中的准确性与完整性, 仍是有待深入研究的问题.

4 基于多传感器融合的运动分割

动态环境下的视觉 SLAM 系统面临运动物体干扰导致定位漂移、建图失真以及潜在运动物体难以判别等问题. 目前常见的运动分割方法主要是基于 RGB 图像或者深度图, 对于运动背景下的目标识别比较困难, 而且单一的视觉传感器由于存在强烈的光照变化、纹理缺失、快速运动或高动态物体占据视野的情况, 其性能往往显著下降, 鲁棒性难以保障. 因此, 为了进一步提高运动分割精度, 需要融合多种传感器, 以实现综合信息提取. 这对运动

表 5　部分基于语义信息的动态 SLAM 方法

Table 5　Partially semantic-information-based dynamic SLAM methods

方法	语义帧选择	运动分割方法	相机类型	年份	运行环境	单帧跟踪时间 (ms)
DS-SLAM[82]	每帧	SegNet	RGB-D	2018	i7 + P4000	59
DynaSLAM[83]	每帧	MaskR-CNN + 几何约束	单、双目及 RGB-D	2018	Tesla M40	195
DynamicSLAM[104]	每帧	SSD	单目	2019	i5-7300HQ + GTX 1050 Ti	45
YPL-SLAM[111]	每帧	YOLOv5s	RGB-D	2024	i7-12700 + RTX 2060	50 ~ 100
SDD-SLAM[113]	关键帧滑动窗口	GroundingDINO + SAM-Track	RGB-D	2025	NVIDIA RTX 3080	—
HMC-SLAM[108]	每帧	YOLOv5	RGB-D	2025	i5 + RTX 4070 Ti	51.02
DOA-SLAM[105]	每帧	FastInst 实例分割	立体相机	2025	—	—
DYMRO-SLAM[109]	关键帧	MaskR-CNN	双目	2025	—	64.89
DYR-SLAM[110]	每帧	YOLOv8	RGB-D	2025	i7-12700K + RTX 3080	57.82
DEG-SLAM[106]	每帧	YOLOv5	RGB-D	2025	i5-8300H + GTX 1050 Ti	57.82
DHP-SLAM[107]	每帧	SOLOv2	RGB-D/双目	2025	i7-9750H + RTX 2070	76.09
YOLO-SLAM[114]	每帧	Darknet19-YOLOv3	RGB-D	2022	Intel Core i5-4288U	696.09
RDS-SLAM[115]	双向关键帧	Mask R-CNN 或 SegNet	RGB-D	2021	RTX 2080 Ti	22 ~ 30
RDMO-SLAM[116]	关键帧	Mask R-CNN	RGB-D	2021	RTX 2080 Ti	22 ~ 35
姜丽梅等[117]	每帧	RT-DETR + PP-LCNet	RGB-D	2025	i7-12650H + RTX 4060	29.86
程强等[118]	每帧	YOLOX	RGB-D/双目	2024	E5-2686 v4 + RTX 3080	37.43
黄友锐等[119]	每帧	YOLOv5s	RGB-D	2024	R7-6800H + RTX 3050 Ti	29.86
DFE-SLAM[120]	每帧	YOLOv5s	RGB-D	2024	—	—
YLS-SLAM[121]	每帧	YOLOv5s-seg	单目/RGB-D	2025	i5 + RTX 3060	17
UE-SLAM[123]	每帧	DINOv2	单目	2025	i7-12700K + RTX 3090 Ti	—

目标进行跟踪, 或是分析当前运动状态都具有一定意义. 不同的融合传感器组合对动态/潜在物体处理的方式不同, 本节将该方法分为三类进行描述.

4.1　基于视觉–惯性融合的运动分割方法

IMU 可以提供高频的自身角速度和加速度信息, 在短时间范围内能够较准确地提供相对可靠的姿态变化估计. 虽容易出现漂移现象, 但 IMU 的观测信息可用于预测当前时刻相机的姿态和平移运动, 或直接对当前时刻其他传感器的观测数据施加约束. 视觉–惯性融合的核心思想是使用 IMU 信息进行运动预测或状态约束, 以供视觉系统在动态环境下区分出相机本身的运动和外界物体的运动, 达到区别动态特征、提取动态信息的目的. 图 10 是视觉 IMU 融合动态剔除的典型框架.

经典的视觉–惯性融合运动分割方法通过位姿变换实现动态特征筛选, 借助 IMU 预积分求解相机位姿变换, 将当前时刻特征点投影至参考坐标系, 再通过残差计算判断特征运动与相机运动的一致性, 进而直接筛选出动态特征.

视觉惯性里程计依据融合策略的紧密程度, 主要分为紧耦合与松耦合两类, 其中 Vins-Mono (visual-inertial navigation system)[124] 和 MSCKF (multi-state constraint Kalman filter)[125] 是两类方法的典型代表[126]. 虽然也有先利用松耦合进行动态特征检测, 将动态特征转化为虚拟静态路标, 再通过紧耦合进行优化的方法. 不过, 在动态环境的运动分割任务中, 紧耦合方法因具备深度传感器融合机制及联合动态特征检测能力而占据主导地位. 通常而言, 紧耦合方法的精度更高, 但相应的计算开销也更大. 其核心在于建立视觉观测与惯性测量的紧耦合模型: 设相机在第 k 帧观测到特征点集 $P_k = \{\boldsymbol{p}_i\}_{i=1}^{N}$, N 为第 k 帧观测到的特征点总数, $\boldsymbol{p}_i$ 为第 i 个特征点的三维坐标, IMU 在 $[t_k,\ t_{k+1}]$ 时间间隔内提供角速度 $\boldsymbol{\omega}(\tau)$ 和加速度 $\boldsymbol{a}(\tau)$ 测量, 其中 τ 为积分时间变量, 用于表征 $[t_k,\ t_{k+1}]$ 内的任意时刻. 通过预积分计算帧间相对运动量:

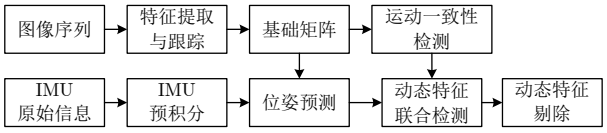


图 10　典型视觉 IMU 融合动态剔除框架

Fig. 10　Typical visual IMU fusion dynamic removal framework

$$\Delta R_{k,\,k+1} = \exp\left(\int_{t_k}^{t_{k+1}} \left(\boldsymbol{\omega}(\tau) - \boldsymbol{b}^{g}\right) \mathrm{d}\tau\right) \tag{7}$$

其中, $\Delta R_{k,\,k+1}$ 为相机/IMU 从第 k 帧到第 $k+1$ 帧的相对旋转矩阵; $\exp(\cdot)$ 为李群指数映射, 用于将旋转向量转换为旋转矩阵; $\boldsymbol{b}^g$ 为 IMU 零偏; $\int_{t_k}^{t_{k+1}} \mathrm{d}\tau$ 为时间区间 $[t_k,\ t_{k+1}]$ 内的积分运算.

$$\Delta \boldsymbol{v}_{k,\,k+1} = \int_{t_k}^{t_{k+1}} \Delta R_{k,\,\tau}\left(\boldsymbol{a}\left(\tau\right) - \boldsymbol{b}^a\right) \mathrm{d}\tau \tag{8}$$

其中, $\Delta \boldsymbol{v}_{k,\,k+1}$ 为相机/IMU 从第 k 帧到第 $k+1$ 帧的相对线速度; $\Delta R_{k,\,\tau}$ 为相机/IMU 从第 k 帧到积分时刻 τ 的相对旋转矩阵; $\boldsymbol{b}^a$ 为 IMU 零偏.

$$\Delta \boldsymbol{p}_{k,\,k+1} = \int_{t_k}^{t_{k+1}} \Delta R_{k,\,\tau}\left(\boldsymbol{a}\left(\tau\right) - \boldsymbol{b}^a\right) \mathrm{d}\tau^2 \tag{9}$$

其中, $\Delta \boldsymbol{p}_{k,\,k+1}$ 为相机/IMU 从第 k 帧到第 $k+1$ 帧的相对位置.

在紧耦合优化框架中, 动态特征点通过运动一致性残差检测, 计算特征点的运动一致性残差, 若残差不满足预设判定条件, 则判定该特征点为动态特征点, 其残差定义为:

$$r_{\mathrm{dyn}} = \boldsymbol{z}_i - \pi\left(T_{wc} T_{wb}^{-1} \Delta T T_{wb} T_{wc}^{-1} \boldsymbol{p}_i^w\right)_{\Sigma} \tag{10}$$

其中, r_{dyn} 为动态特征判断残差; $\boldsymbol{z}_i$ 为第 i 个特征点的视觉观测值; $\pi(\cdot)$ 为三维空间点到二维图像平面的投影函数; T_{wc} 为世界坐标系 (w) 到相机坐标系 (c) 的位姿变换矩阵, 包含旋转和平移信息; T_{wb} 为世界坐标系 (w) 到 IMU 坐标系 (b) 的位姿变换矩阵; T_{wb}^{-1} 为 T_{wb} 的逆矩阵, 用于实现从 IMU 坐标系到世界坐标系的变换; ΔT 为动态特征点相对于静态环境的运动变换矩阵, 表征动态特征点的自身运动; $\boldsymbol{p}_i^w$ 为第 i 个特征点在世界坐标系下的三维坐标; Σ 为视觉观测噪声的协方差矩阵.

当 $\parallel r_{\mathrm{dyn}} \parallel > \tau_{\mathrm{dyn}}$ 时判定为动态点. 其中, τ_{dyn} 为动态特征判断阈值, 用于区分静态与动态特征点, 其值通常由实验标定或基于观测噪声统计特性确定. 典型代表如 VINS-Mono[124] 通过构建视觉惯性联合优化的目标函数, 对系统状态进行最小二乘求解, 完成鲁棒的状态估计. 为提升优化精度, 算法会先剔除动态特征带来的异常观测. 系统的优化目标为最小化视觉残差与 IMU 预积分残差的马氏距离平方和, 表达式如下:

$$\min_{X} \left(\sum_{i \in V} \parallel \boldsymbol{r}_{V_i} \parallel_{\Sigma_V}^2 + \sum_{j \in I} \parallel \boldsymbol{r}_{I_j} \parallel_{\Sigma_I}^2 \right) \tag{11}$$

其中, 状态量 $X = \{T_{wb},\ \boldsymbol{v}^w,\ \boldsymbol{b}^a,\ \boldsymbol{b}^g\}$ 包含位姿、速度和 IMU 参数, $\boldsymbol{v}^w$ 为相机在世界坐标系 w 下的线速度; V 表示所有视觉观测项的索引集合; I 表示所有 IMU 预积分项的索引集合; $\boldsymbol{r}_{V_i}$ 为第 i 个视觉观测对应的视觉残差; $\boldsymbol{r}_{I_j}$ 为第 j 个 IMU 预积分对应的 IMU 残差; Σ_I 为 IMU 预积分测量噪声的协方差矩阵; Σ_V 为视觉观测过程中噪声的协方差矩阵.

在过去的研究中, 使用 IMU 数据进行状态预测、使用视觉数据进行状态观测, 并将二者嵌入同一个滤波框架中实现紧耦合, 以提高动态环境下的鲁棒性. MSCKF[125] 是一种基于扩展卡尔曼滤波 (extended Kalman filter, EKF) 的紧耦合框架, 由 Mourikis 和 Roumeliotis 提出. 该方法利用 IMU 对整个状态过程进行预测, 将当前时刻的 IMU 速度、测量偏差状态及滑动窗口内多时刻相机位姿均整合到状态向量中, 利用它们求解每个时刻的状态, 从而实现对视觉惯性里程计的运动估计.

随着更高精度和实时性能的需求, 研究者们提出了多种优化方案. Bloesch 等[127] 与 Wu 等[128] 都针对 EKF 在不可观变换下的不一致性, 先后提出了右不变误差 EKF 并建立了右不变误差模型. 其中, 文献 [127] 所提方法 RIEKF-VINS 通过 IMU 预积分提供状态预测和过程噪声, 视觉观测通过重投影残差提供状态更新, 并通过设计不变性误差表示, 确保滤波器在不可观测变换下的不变性, 避免协方差低估问题. 文献 [128] 通过李群表示保证平移与重力旋转的数学不变性, 解决了传统方法在闭环场景中 18% 的累积漂移误差. 然而, 这一方法在复杂动态环境中的表现依然存在不足, 尤其在处理快速运动或光照变化较大的场景时, 容易出现定位漂移或失锁现象. Chen 等[129] 提出了协方差交集 (covariance intersection) 滤波, 同样通过 IMU 预积分估计位姿不确定性, 并通过后向协方差传播将视觉测量不确定性转换到估计域. 在更新阶段, 采用协方差交集融合视觉与 LiDAR 位姿估计. Zhu 等[130] 提出的 PLD-VINS 结合了 RGB-D 传感器, 加入了点、线特征, 在一定程度上增强了在低纹理环境下的三维地图构建的精确度和鲁棒性. Zhang 等[131] 提出的 DP-VINS 把重投影误差与共面性约束结合起来, 设计速度自适应的动态因子权重函数, 在急加速场景下降低静态点的误判率至 24%, 将 KITTI 动态场景误检率降到 9.7%. Yin 等[132] 则采用松散耦合的 IMU 与场景流, 用 IMU 预积分计算出相邻帧之间的相对位姿; 同时根据立体光流来构建场景运动向量, 根据距离来建模场景流的不确定度, 来对地标运动进行似然度估计, 如果这个地标运动的似然度大于一定的门限值, 则将其标记成动态特征.

2024 年, Abdollahi 等[133] 提出的 Fast-MSCKF 在 MSCKF 的基础上进行了改进, 把特征的边际化、状态的修剪分别简化成测量方程和期望状态的留

出, 实现算法速度和计算效率的飞速提升. 为克服 MSCKF 在长时间估计中的漂移, Jung 等[134] 将 RGB-D 传感器和低成本 IMU 结合在一起形成 MSCKF-DVIO, 利用插值法估计未来状态、姿态更新的同时加入非完整约束提升精度. 同年, Fornasier 等[135] 通过引入对称群理论给出了一个新的方案——MSCEqF, 该方案实现了视觉、惯性数据同时使用. 2025 年, Cao 等[136] 提出了基于光流场和 IMU 预积分的运动状态估计方法, 首先通过 IMU 预积分获取相机初始位姿, 再计算动态流残差、极线残差并量化特征点的运动状态, 通过对特征点的加权降低动态物体对于位姿估计的干扰. Yu 等[137] 设计了自适应特征对应算法, 在跟踪特征少于阈值时利用 IMU 预积分预测位姿和新匹配, 在满足阈值条件下, 使用预测位姿计算基础矩阵, 通过极线约束过滤掉动态对象带来的误匹配.

该类方法通过 IMU 高频运动约束提供相对运动的短期可靠估计, 虽然会有累积漂移, 但是能够为视觉系统提供帮助, 以分清是相机自身的运动还是环境中的动态物体运动; 在视觉发生短暂失效的时候, 也能够维持系统稳定, 并且能够根据实时修正传感器的参数误差值来适应快速运动场景.

4.2　基于视觉–雷达融合的运动分割方法

视觉传感器和雷达的融合解决了单一模态感知不足的问题, 在动态环境理解方面具有明显的优势. 本节所讨论的雷达主要涵盖激光雷达 (LiDAR)、毫米波雷达 (mmWaveRadar) 及超声波雷达 (UltrasoundRadar) 等常见类型, 通过这类雷达来获取基于主动的距离/速度测量信息, 利用其环境穿透能力强的特点来克服视觉中深度估计不准确、易受光线影响、受恶劣天气影响大的问题[2, 25], 其融合框架如图 11 所示. LiDAR 提供高精度、高分辨率的 3D 点云数据, 可以为特征点提供精确的深度或空间位置先验[138]. 毫米波雷达可以提供目标或者区域的距离值以及径向速度信息; 距离信息可以用来限定视觉特征点的深度估计范围 (特别是在低纹理区域), 也可以用作校验视觉计算出来的深度是否合理; 速度信息则是判断目标的运动状态最直接的证据. 超声波雷达可以提供近距离一维距离值信息, 主要用于近场避障和简单场景下深度辅助验证.

目前, 研究者们的重点主要是通过空间对齐的方法来完成视觉和雷达的数据特征拼接. Du 等[139] 提出的 PMSB (projection mapping segmentation block) 方法, 通过先建立一种基于投影的空间块和图像像素块之间的对应关系, 再提取本地的空间关联特征以及半变差的纹理特征, 并用支持向量机做

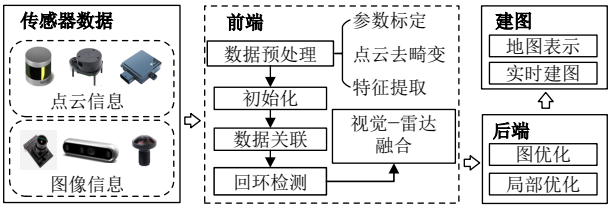


图 11　视觉–雷达融合框架

Fig. 11　Visual-radar fusion framework

多目标分类. 还有 Tan 等[140] 提出了 EPMF (efficient perception-aware multi-sensor fusion) 方法, 使用了投影对齐来对输入的数据进行优化, 用了双流网络分别提取视图域和雷达域的特征, 又引入了残差融合模块把图模态的信息添加到点云主干网, 这样就会提升分割分支对动态的目标表征能力. Song 等[141] 开发了一种基于体素–像素匹配的相机-LiDAR 融合方法, 用精校准 3D 体素和准确 2D 像素特征来补正传统 2D 投影带来的极线几何问题, 在语义分割的点云中实现高速度的动态物体检测.

随着对动态目标分割精度要求的提高, 一些学者开始尝试使用雷达提供的深度信息和图像提供的外观信息来进行分割以增强分割性能. Shi 等[142] 提出了 CMF (cross-modal fusion) 框架, 在此框架下, 借助跨模态注意力驱动的特征融合方式实现了 RGB 图像与点云的多层次融合, 并提升了语义分割的准确性. 此外, Sánchez-García 等[143] 提出的 SalsaNext + RGB 方法将 RGB 数据和雷达点云相结合, 能够提升语义分割的精度, 利用图像和点云数据的互补特性提高模型识别动态目标的能力. Zhu 等[144] 提出的 RDynaSLAM 系统将 4D 毫米波雷达点云与视觉 SLAM 相结合, 提出雷达动态聚类提取和动态掩码生成的方法, 减小动态目标对于视觉 SLAM 系统的干扰. 采用此种方法可将动态目标从相机的关键点提取中剔除, 提高动态环境下定位的精度.

这些方法旨在克服单一传感器的固有局限性, 通过融合视觉与雷达数据, 在实现动态目标分割与场景理解的同时, 面向复杂驾驶场景支持多任务协同处理, 从而提升对多种复杂动态环境的建模与处理能力.

4.3　融合多源信息的运动分割方法

融合多源信息的运动分割方法旨在构建一个鲁棒性强、智能化程度高、环境理解能力强的 SLAM 系统, 在多源异构传感器 (相机、IMU、LiDAR、毫米波雷达、GNSS、事件相机等) 信息流深度融合的基础上, 同时可以把语义理解作为提高动态环境下

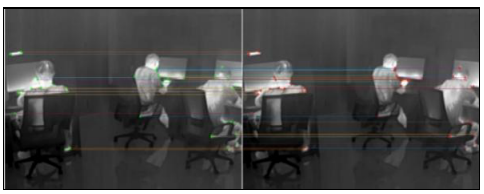
物体识别正确率和鲁棒性的关键手段, 从而提升系统对动态物体的分割能力和鲁棒性.

R3LIVE[145] 突破传统视觉-LiDAR 松散耦合的局限, 以 LiDAR 提供的精确 3D 几何结构为支撑, 为视觉数据提供稳定的深度参考与运动约束, 解决了视觉 SLAM 在无纹理环境中无特征可跟踪的核心痛点. 同时, 利用 IMU 的高频运动信息补偿视觉帧间运动畸变, 提升了视觉数据在快速运动场景下的有效性. Li 等[146] 提出了基于多 GNSS PPP (precise point positioning) 与 Stereo VINS 的半紧耦合集成框架, 通过图优化方式使本地 S-VINS 轨迹和全局 GNSS 轨迹实现互相校正, 在 GNSS 信号受到遮挡的环境下能够大幅提升定位精度; 融合 GNSS 和视觉惯性信息, 同时结合前端多传感器信息辅助判定由于非刚体运动造成的定位误差, 增强动态干扰的鲁棒性, 从而增强分割效果. Cao 等[147] 提出了 GVINS, 结合了 GNSS 与 VINS 进行紧耦合融合, 实现局部路径的精准化和全局最优的目标, 解决了 GNSS 信号缺失时的定位漂移问题. Zhang 等[148] 提出了一种面向动态环境的视觉−音频融合 SLAM 方法, 引入声源方向估计作为辅助信息定位动态目标区域, 并结合 RGB-D 视觉输入实现动态障碍物的剔除处理. 该方法利用麦克风阵列估计动态目标的方向信息, 与视觉 SLAM 估计结果进行联合分析, 实现在多机器人系统中对动态物体的高效分割与环境建图, 适用于语音交互频繁的工业协作场景. Song 等[149] 提出了 PSMD-SLAM 方法, 这是首个融合全景分割与多模态感知的 SLAM 框架, 集成了激光雷达、IMU 和图像信息, 且利用全景分割模型获得语义及实例分割的信息, 在视觉与雷达模态分别用概率传播法和 PCA (principal component analysis) 聚类的方法识别动态物体, 最终利用多模态冗余信息进行动态区域识别与剔除.

基于可微分渲染的方法以 3D 高斯或 NeRF 为多模态信息整合载体, 实现视觉、LiDAR、IMU、事件相机等数据的深度协同, 突破单一传感器的动态分割局限. 多模态与 3D 高斯溅射的结合可强化复杂场景动态分割. Lang 等[150] 提出的 Gaussian-LIC 融合 LiDAR、IMU 与相机多传感器数据, 构建多模态协同的 3D 高斯建模框架. LiDAR 点云为 3D 高斯提供高精度深度约束, 解决了视觉传感器在弱纹理区域的几何建模难题; IMU 数据提供高频运动信息, 辅助优化相机位姿的初始化与跟踪精度; 相机数据则补充场景的颜色与纹理信息, 通过可微分渲染实现多模态数据的一致性优化, 在室外复杂场景中显著提升了渲染质量与位姿估计稳定性, 实验中其渲染 PSNR (peak signal-to-noise ratio)、SSIM (structural similarity) 等指标均优于单一传感器方法. EN-SLAM[151] 作为首个事件相机-RGB-D 隐式神经 SLAM 框架, 通过跨模态共享辐射场建模、可微分 CRF (camera response function) 渲染和事件时序优化策略, 有效解决了运动模糊与光照变化下的 SLAM 挑战. 在多模态融合层面, 设计可微分相机响应函数, 将 RGB 颜色信息与事件相机的 luminance 信号映射到共享辐射场, 利用事件生成模型的时序差分特性, 精准识别快速运动的动态区域. 该方法在 DEV-Reals 数据集“motionblur”序列中, ATE-RMSE 仅 8.94 cm, 较 CoSLAM 降低约 19.7%; 在低光照场景下跟踪成功率达 100%, 而传统 RGB-D 方法均出现崩溃.

为进一步扩展多模态融合, 已有学者尝试将热红外信息、可见光相机等联合. 2012 年 Vidas 等[152] 首次实现热红外单目里程计, 一年后又利用热红外信息进行地图标注[153]. 2015 年 Mouats 等[154] 通过特征提取实现红外与可见光的立体 SLAM 系统. 2024 年 Qin 等[155] 提出的 BVT-SLAM 通过多光谱立体匹配提高了实时性和精度. 2025 年 WTI-SLAM[156] 将热红外图像作为主要输入, 用于解决低纹理、低能见度环境下的定位与感知问题, 并融合特征点提取、光流估计方法, 利用热红外图像中对于动态目标的热特征稳定性较好这一优点, 在视觉退化的场景下增强鲁棒分割能力. 如图 12 所示, 图 12(a) 中特征匹配主要集中在灰度差异显著的区域, 而差异较小的区域没有匹配; 图 12(b) 则呈现

(a) 未添加光流的特征提取实验结果
(a) Experimental results of feature extraction without optical flow



(b) 带有光流跟踪的特征提取实验结果
(b) Experimental results of feature extraction with optical flow tracking

图 12 低纹理热红外图像的特征跟踪结果

Fig. 12 Feature tracking results for low-texture thermal infrared images

特征点均匀分布的特征提取效果, 因此可知带有光流跟踪的特征提取相较于未添加光流低纹理热红外图像在特征跟踪方面有着较好的稳定性, 结合可见光、雷达等多种传感器的应用, 可以稳定地识别出复杂环境下的动态干扰因素, 呈现出多模态的跨模态协同分割的优势.

表 6 总结了部分基于多传感器信息的 SLAM 方法的特点. 由表 6 可知, 多传感器融合运动分割方法研究的发展呈现出从单一信息补偿向协同多模态理解的演进趋势; 从松耦合到紧耦合的融合策略发展; 从低级几何融合到高级语义协同发展的感知模式发展; 从静态标定到在线自适应校准的系统实现. 这些进展不仅显著提升了动态场景下的运动分割性能, 也为构建智能感知系统提供了坚实技术支撑.

总体而言, 多源信息融合运动分割方法体现了由低耦合补偿向深融合协同的技术演化趋势. 早期研究侧重于利用高频传感器辅助视觉感知进行位姿修正, 而最近几年更注重在感知阶段融合多源信息进行动态语义分割与结构补全. 多模态全景分割、跨域语义一致性建模以及声音–视觉协同感知等则是目前促进多源融合运动分割精度与实时性的重要的技术路径.

5 未来发展趋势

在动态环境中, 为实现高精度定位跟踪, 传统 V-SLAM 面临巨大困难. 运动分割是提高 V-SLAM 在动态环境下适应性的关键技术, 而如何突破其自身存在问题以及和现有技术结合解决实际问题将是运动分割后期的主要方向. 目前运动分割存在如下几个主要问题: 1) 准确性与时效性矛盾: 语义分割精度高但计算量大; 几何方法实时性好但运动检测能力弱. 2) 泛化能力差: 模型局限于某种场景或某个设备上迁移性差. 3) 多模态协同度弱: 传感器信息相互隔离, 不能协同完成动态交互的建模, 从而很难区分正常动效下的潜在运动目标. 因此, 针对基于静态场景假设的方法在高动态环境下存在精度不足且难以识别潜在运动目标的问题, 基于语义信息的方法面临实时性瓶颈, 以及多传感器融合策略受限于设备精度和数据融合复杂度, 难以有效分割潜在运动目标等问题[2], 动态环境下的运动分割技术在以下方向值得深入研究.

1) 动态场景下的实时语义–几何协同感知

在动态场景中, 物体的运动以及场景的不断变化给感知任务带来了巨大挑战. 基于静态假设的传统方法在高动态场景下无法准确地检测出潜在运动目标, 而基于深度学习的语义分割方法虽然较为准确, 但由于计算量较大、实时性较差, 因此限制了其在实时系统中的应用. 兼顾深度学习语义模型的计算效率与几何信息的实时性的思路, 即实现实时语义–几何协同感知, 是未来需突破的重要方向[7]. 通

表 6 部分基于多传感器信息的 SLAM 方法
Table 6 Partially multi-sensor-information-based SLAM methods

方法	年份	传感器	绝对轨迹均方根误差 (m)	分割方法	运行环境
MSCKF[125]	2007	单目相机 + IMU	—	多状态约束卡尔曼滤波、特征点跟踪与三角测量	T7200
VINS-Mono[124]	2018	单目相机 + IMU	0.12 ~ 0.22	关键帧选择 + 滑动窗口优化 + 特征点跟踪	i7-4790
VINS-Fusion[157]	2019	RGB (单目/双目)+ IMU + GPS	0.06	滑动窗口优化 + 因子图优化 + 多传感器因子融合	—
R3LIVE[145]	2022	LiDAR + RGB + IMU	0.085	基于点云运动一致性与视觉语义融合	i7-8550U + 8 GB
PLD-VINS[130]	2021	RGB-D 相机 + IMU	0.731 866	改进 EDLines 检测线特征, 光流法跟踪线特征	XeonE5645 + 48 GB
VA-fusion[148]	2023	RGB-D 相机 + 麦克风阵列 + IMU	0.301	声源方向投影至图像平面, 直接标记动态区域	i7 + 64 GB
DP-VINS[131]	2024	立体相机 + IMU	0.092	通过光流场聚类和残差计算估计运动状态	i7-9700K + 16 GB
FMSCKF[133]	2024	单目相机 + IMU	0.151	基于关键帧特征跟踪阈值, 结合 IMU 预积分预测匹配	i7-11800H + 32 GB
RDynaSLAM[144]	2025	4D 毫米波雷达 + 相机	0.233	通过 RANSAC 提取动态簇并生成动态掩码, 以过滤动态点	E3-1270 v2 + 8 GB
PSMD-SLAM[149]	2024	LiDAR + 相机 + IMU	2.87	概率传播 + PCA 聚类 + 全景分割辅助动态检测	5950X + RTX 3090
Ground-Fusion[158]	2024	RGB-D 相机 + IMU + 轮速计 + GNSS	0.1	运动一致性检查、深度验证 + 传感器异常检测	E3-1270 v2 + 8 GB
SFCI[137]	2025	单目相机 + IMU	0.22	IMU 预积分预测位姿生成对极几何约束, 直接剔除动态点	i9 + 16 GB RAM
EN-SLAM[151]	2024	RGB-D + 事件相机	0.159 7	利用事件数据的高动态范围和时序特性	RTX 4090
WTI-SLAM[156]	2025	热红外相机 + IMU	0.059	多尺度相位一致性特征提取 + 光流前后向跟踪	i5-12450H + 32 GB
Hybrid-VINS[159]	2025	UBSL + 单目相机 + IMU	0.11	基于深度一致性过滤动态物体	—
DVI-SLAM[160]	2024	单目/stereo 相机 + IMU	0.148	动态融合视觉 + 惯性因子优化位姿	RTX 3090

过同时获取场景中物体的语义类别信息以及精确的几何结构信息, 从而促进实现全面理解动态环境. 具体而言, 语义信息可以将物体进一步分为不同的类别, 从整个场景分割的基础上得到语义层面的不同区域的分割, 进而为动态物体的分割提供先验知识; 对于获取几何信息而言, 具有高实时性和计算时间短的优点, 且能够利用深度图、光流估计和几何约束等方法区分背景和动态目标. 总体而言, 通过分析激光雷达点云的几何特征, 如点的分布密度、法线方向等, 为相机图像的语义分割提供先验信息, 帮助模型更好地识别物体边界和类别. 将语义分割结果作为约束条件, 优化激光雷达点云的几何重建过程, 借助语义信息辅助几何重建能提高重建模型的完整性和准确性.

2) 基于时空一致性与语义建模的长时运动分割

时空信息不仅是动态感知中不可或缺的基础信息, 也是描述物体运动状态的关键信息, 时空特征与语义结合可以提高复杂动态场景的辨识度[161]. 时空特征可以通过连续帧间时间关联与运动特征变化来描述动态物体的轨迹, 在描述高速运动目标或者长时间序列跟踪时有显著作用. 而长时运动分割需解决遮挡重识别、运动突变等时变挑战, 其核心在于融合时空一致性约束与语义场景理解. 基于光流的时空一致性通过稠密光流算法计算像素运动向量, 利用运动物体像素光流向量的相似性和连贯性关联像素, 保证分割一致性. 通过设置光流向量的阈值和方向一致性条件, 可将属于同一运动物体的像素点在时间维度上进行关联, 保证长时运动分割的一致性.

此外, 对于运动物体, 其特征点在相邻帧间的匹配结果应符合物体的运动规律. 因此, 根据特征点的运动轨迹, 有利于在时间维度上保持分割结果的一致性. 通过时空一致性建模可表征运动连续性约束、保持拓扑结构等. 而语义建模不仅能将静态要素作为运动分割的参考系, 抑制相机抖动导致的误分割, 还能通过联合物体功能属性与场景物理规则实现动态语义推理, 构建环境要素的符号化表示及其关联关系, 赋予机器人对物理世界的认知理解能力. 因此, 基于时空一致性与语义建模的融合将促进克服动态场景长时运动分割的时变性与观测碎片化. 特别地, 随着神经符号系统、大语言模型的发展, 结合符号逻辑的显式规则 (如“行人不会穿墙”) 与神经网络的隐式学习, 能促进实现可解释[16] 的长时运动推理, 而利用大语言模型的常识知识库生成运动假设 (如“落叶被风吹动属于无序运动”), 能驱动分割模型的自适应能力. 基于时空一致性与语义建模的长时运动分割, 将推动运动分割从被动响应到主动预测的技术跃迁, 为 V-SLAM 系统具备可靠感知能力奠定基础.

3) 轻量化部署与边缘计算

随着嵌入式设备、移动机器人[162] 和自动驾驶系统的发展, 轻量化部署和边缘计算驱动下的运动分割是重要发展方向. 当前, 基于深度神经网络的语义分割方法由于模型较为复杂、计算需求较高, 在有限计算资源条件下的嵌入式设备上难以得到实际应用. 为了达到对动态物体进行实时准确分割的目的, 后续将进一步研究如何利用轻量化模型和边缘计算, 提升运动分割的实时性与准确性[163]. 通过网络剪枝、量化、深度可分离卷积等技术手段, 可在保证分割精度的前提下大幅降低模型开销. 基于卷积的冗余特征生成机制, 能以极低的参数量扩展特征维度; 知识蒸馏技术通过将大模型的判别能力迁移至小模型, 实现精度损失最小化–体积压缩最大化; GDS-ECA 等轻量化卷积模块结合深度可分离卷积与高效通道注意力, 进一步平衡了模型性能与计算效率. 如 LSSMask 模型参数量仅 4.2 M, 仍能达到 35.8% 的平均精度, 为端侧部署提供了可行范式. 结合边缘计算[164−165] 中使用的硬件 (GPU/NPU/FPGA 等处理器) 来加速数据的处理和传输可以极大地提高系统实时响应的能力. GPU 凭借并行计算优势加速特征提取等密集型任务, NPU 通过深度学习算子优化降低推理延迟, FPGA 则以可定制化逻辑满足低功耗场景需求. 将轻量化小模型与边缘硬件深度协同, 可显著提升数据处理与传输效率, 解决了传统语义分割在嵌入式设备上的实时性不足问题. 通过以上技术手段的协同应用, 可以使 V-SLAM 视觉系统在复杂动态环境中具有高效的实时性以及低延时的运动分割能力, 可以更好地服务于机器人导航和自动驾驶领域.

4) 跨场景、跨设备的泛化能力

现实中由于视觉 SLAM 系统部署场景和设备种类繁多, 加之不同环境与不同传感器带来的差别使得其具有较好的通用性是一个非常重要的因素. 目前已有的运动分割方法大多数是针对某些特定数据集或者固定硬件平台所做的优化. 在未知场景下, 比如换一种新的硬件设备, 就会出现较大的分割结果偏差. 因此构建具备强泛化能力的运动分割技术, 需突破“场景依赖–设备绑定”瓶颈, 而大模型与小模型的协同机制为这一目标提供了核心解决方案. 大模型的通用感知能力为跨场景泛化奠定基础. 视觉基础模型与大语言模型通过海量多场景数据训练, 已形成对通用视觉特征与物理常识的深度理解.

将大模型的通用特征提取能力与动态运动推理逻辑迁移至运动分割任务, 可减少模型对特定场景数据的依赖. 由于不同硬件设备的传感器参数存在显著差异, 直接迁移模型易导致分割精度下降. 小模型的设备适配与在线学习能力有效解决了跨设备兼容难题. 将大模型作为通用知识库, 为小模型提供跨场景、跨设备的共性约束. 小模型则基于端侧实时数据进行局部优化, 将场景特异性知识反馈至大模型的增量训练过程, 通过大小模型的协同学习与知识迁移进一步提升泛化鲁棒性. 未来, 通过大模型通用感知与小模型设备适配的深度协同, 结合无监督域自适应、元学习等技术, 可构建"一次训练–多场景多设备适配"的运动分割体系, 打破传统方法的场景与设备局限, 为 V-SLAM 系统在复杂多样的现实环境中稳定运行提供保障.

5) 神经隐式表征与持续学习

近年来, 基于神经隐式表征的可微分渲染 SLAM 方法, 正逐渐成为处理动态场景的重要前沿. 这类方法将场景表示为连续的隐式函数, 通过体渲染或高斯泼溅生成照片级真实感的新视图, 在此过程中, 动态物体会破坏多视图一致性, 从而可以被检测和分割. 其优势在于能够实现密集的、具有几何一致性的运动分割, 并同时完成高质量的静态场景重建与动态区域修复[166]. 未来研究可探索如何降低此类方法的计算开销, 使其能够实时运行, 并研究如何将其与传统的几何、语义方法更紧密地结合, 例如利用神经表征提供更准确的深度先验或运动概率. 更进一步, 如 Li 等[33] 所探索的, 将持续学习中的记忆与遗忘机制引入动态 SLAM, 使系统能够自适应地学习并更新场景中的静态元素与动态对象的先验知识, 从而实现对动态环境的长期、渐进式理解, 这将是迈向真正智能且适应性强的高层运动理解的关键. 这种解决方案摆脱了对传统分割的依赖, 提供了一种全新的基于场景表征内在学习特性的运动分割范式. 然而, 该方法目前仍面临可解释性较弱、对剧烈动态变化或长期遮挡场景处理能力不足等挑战. 尽管如此, 其为运动分割技术由显式处理向隐式学习的演进提供了可能, 并为构建更智能、更自主的 SLAM 系统提供了新的研究方向.

综上, 语义–几何协同感知是连接运动分割与运动理解的首要环节, 为运动状态解析提供精准的多维度输入. 长时运动分割通过时空特征与语义知识的融合, 解决了运动理解中交互建模的时序连贯性问题. 轻量化部署与边缘计算是运动理解从实验室走向实际场景的关键, 解决了高精度理解与低资源约束的矛盾. 跨场景、跨设备泛化能力决定了运动理解能否适应多样化动态环境, 解决了分割模型场景依赖与理解能力设备受限的问题. 神经隐式表征与持续学习的结合, 为运动理解提供"隐式学习 + 自适应更新"的全新范式, 推动其从被动解析向主动推断跃迁. 这五大方向共同支撑"感知–分析–解释"的运动理解技术路线. 最终, 通过多技术协同突破, 实现从动态/静态区分到"运动状态–模式–交互–意图"全维度认知的跨越, 为移动机器人、自动驾驶等领域提供复杂动态环境下的智能感知基础.

6 结束语

本文系统梳理了 V-SLAM 系统中运动分割技术的研究进展, 基于各类方法对场景预设条件的不同, 将现有主流技术归纳为基于静态场景假设的方法、基于语义信息的方法以及基于多传感器融合的方法, 并介绍了各种方法的技术特点、应用场景及优势和局限性. 基于静态场景假设的大部分方法无需先验语义信息或其他的传感器数据, 因此计算代价较小, 能够适用于大多数以静态为主场景下的检测和跟踪任务. 但是由于无法获得先验信息, 所以不能获取动态场景下目标之间的相对位置, 也就无法识别出各种类型的运动物体. 基于先验语义的知识可以利用目前较为成熟的深度学习技术进行训练, 可以从大规模数据集中提取场景对象, 大大提高目标的识别率. 但是由于语义信息量较大且复杂性较高, 造成其运算量大、实时性差, 很难满足 V-SLAM 系统的需求. 基于多传感器融合的方法是将视觉、IMU、激光雷达以及其他新式的传感器采集到的异构数据进行融合处理来提升整个系统精度. 但为了提升系统精度, 不但要做到精确的数据融合还需要解决大量的传感器标定、校准等问题, 所以数据融合的计算复杂度也随着传感器种类的不同而变化. 通过对动态环境下运动分割技术的深度剖析, 结合当前研究瓶颈与技术演进趋势, 未来 V-SLAM 运动分割技术将围绕"从分割到理解"的核心目标持续突破. 虽然现有的方法依然难以达到精度和实时性兼顾的效果, 运动理解的深度与广度仍有较大提升空间, 但是随着以后深度学习、时空特征融合、边缘计算、神经隐式表征以及多传感器融合等前沿技术的应用, 未来 V-SLAM 系统将能在更复杂的动态环境中实现高效、精准的运动分割, 为智能机器人、自动驾驶、增强现实等领域提供更稳定可靠的技术支持, 推动 V-SLAM 技术向更智能、更自主的方向发展.

参考文献

1 Cadena C, Carlone L, Carrillo H, Latif Y, Scaramuzza D, Neira

J, et al. Past, present, and future of simultaneous localization and mapping: Toward the robust-perception age. *IEEE Transactions on Robotics*, 2016, **32**(6): 1309−1332

2 Wang Jin-Ke, Zuo Xing-Xing, Zhao Xiang-Rui, Lv Jia-Jun, Liu Yong. Review of multi-source fusion SLAM: Current status and challenges. *Journal of Image and Graphics*, 2022, **27**(2): 368−389
(王金科, 左星星, 赵祥瑞, 吕佳俊, 刘勇. 多源融合 SLAM 的现状与挑战. 中国图象图形学报, 2022, **27**(2): 368−389)

3 Fuentes-Pacheco J, Ruiz-Ascencio J, Rendón-Mancha J M. Visual simultaneous localization and mapping: A survey. *Artificial Intelligence Review*, 2015, **43**(1): 55−81

4 Zhang Jun-Ning, Su Qun-Xing, Liu Peng-Yuan, Zhu Qing, Zhang Kai. An improved VSLAM algorithm based on adaptive feature map. *Acta Automatica Sinica*, 2019, **45**(3): 553−565
(张峻宁, 苏群星, 刘鹏远, 朱庆, 张凯. 一种自适应特征地图匹配的改进 VSLAM 算法. 自动化学报, 2019, **45**(3): 553−565)

5 Nguyen T M, Wu Q J. A consensus model for motion segmentation in dynamic scenes. *IEEE Transactions on Circuits and Systems for Video Technology*, 2016, **26**(12): 2240−2249

6 Xia L L, Cui J S, Shen R, Xu X, Gao Y P, Li X Y. A survey of image semantics-based visual simultaneous localization and mapping: Application-oriented solutions to autonomous navigation of mobile robots. *International Journal of Advanced Robotic Systems*, 2020, **17**(3): 1−17

7 Wang Y N, Tian Y B, Chen J W, Xu K, Ding X L. A survey of visual SLAM in dynamic environment: The evolution from geometric to semantic approaches. *IEEE Transactions on Instrumentation and Measurement*, 2024, **73**: Article No. 2523221

8 Xu Z W, Rong Z, Wu Y H. A survey: Which features are required for dynamic visual simultaneous localization and mapping? *Visual Computing for Industry, Biomedicine, and Art*, 2021, **4**(1): Article No. 20

9 Azzam R, Taha T, Huang S D, Zweiri Y. Feature-based visual simultaneous localization and mapping: A survey. *SN Applied Sciences*, 2020, **2**(2): Article No. 224

10 Mur-Artal R, Montiel J M M, Tardós J D. ORB-SLAM: A versatile and accurate monocular SLAM system. *IEEE Transactions on Robotics*, 2015, **31**(5): 1147−1163

11 Mur-Artal R, Tardós J D. ORB-SLAM2: An open-source SLAM system for monocular, stereo, and RGB-D cameras. *IEEE Transactions on Robotics*, 2017, **33**(5): 1255−1262

12 Campos C, Elvira R, Rodríguez J J G, Montiel J M M, Tardós J D. ORB-SLAM3: An accurate open-source library for visual, visual-inertial, and multimap SLAM. *IEEE Transactions on Robotics*, 2021, **37**(6): 1874−1890

13 Caruso D, Engel J, Cremers D. Large-scale direct SLAM for omnidirectional cameras. In: Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). Hamburg, Germany: IEEE, 2015. 141−148

14 Engel J, Koltun V, Cremers D. Direct sparse odometry. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2018, **40**(3): 611−625

15 Xu S Y, Wang T, Lang C Y, Feng S H, Jin Y. Graph-based visual odometry for VSLAM. *Industrial Robot*, 2018, **45**(5): 679−687

16 Xu B, Yang G C. Interpretability research of deep learning: A literature survey. *Information Fusion*, 2025, **115**: Article No. 102721

17 Qu H Z, Hu Z H, Zhao Y C, Lu J L, Ding K K, Liu G F, et al. Point-line feature-based vSLAM systems: A survey. *Expert Systems With Applications*, 2025, **289**: Article No. 127574

18 Luo Yuan, Shen Ji-Xiang, Li Fang-Yu. Review of visual SLAM research based on deep learning in dynamic environments. *Semiconductor Optoelectronics*, 2024, **45**(1): 1−10
(罗元, 沈吉祥, 李方宇. 动态环境下基于深度学习的视觉 SLAM 研究综述. 半导体光电, 2024, **45**(1): 1−10)

19 Lai T. A review on visual-SLAM: Advancements from geometric modelling to learning-based semantic scene understanding using multi-modal sensor fusion. *Sensors*, 2022, **22**(19): Article No. 7265

20 Zhao Yang, Liu Guo-Liang, Tian Guo-Hui, Luo Yong, Wang Zi-Ren, Zhang Wei, et al. A survey of visual SLAM based on deep learning. *Robot*, 2017, **39**(6): 889−896
(赵洋, 刘国良, 田国会, 罗勇, 王梓任, 张威, 等. 基于深度学习的视觉 SLAM 综述. 机器人, 2017, **39**(6): 889−896)

21 Huang Ze-Xia, Shao Chun-Li. Survey of visual SLAM based on deep learning. *Robot*, 2023, **45**(6): 756−768
(黄泽霞, 邵春莉. 深度学习下的视觉 SLAM 综述. 机器人, 2023, **45**(6): 756−768)

22 Zhang Rong-Fen, Yuan Wen-Hao, Li Jing-Yu, Liu Yu-Hong. Review of VSLAM research with semantic information. *Journal of Guizhou University* (*Natural Sciences*), 2022, **39**(5): 81−87
(张荣芬, 袁文昊, 李景玉, 刘宇红. 融入语义信息的 VSLAM 研究综述. 贵州大学学报 (自然科学版), 2022, **39**(5): 81−87)

23 Fan Z, Zhang L L, Wang X Y, Shen Y L, Deng F. LiDAR, IMU, and camera fusion for simultaneous localization and mapping: A systematic review. *Artificial Intelligence Review*, 2025, **58**(6): Article No. 174

24 Li X D, Dunkin F, Dezert J. Multi-source information fusion: Progress and future. *Chinese Journal of Aeronautics*, 2024, **37**(7): 24−58

25 Gao Qiang, Lu Ke-Fan, Ji Yue-Hui, Liu Jun-Jie, Xu Liang, Wei Guang-Rui. Survey on the research of multi-sensor fusion SLAM. *Modern Radar*, 2024, **46**(8): 29−39
(高强, 陆科帆, 吉月辉, 刘俊杰, 许亮, 魏光睿. 多传感器融合 SLAM 研究综述. 现代雷达, 2024, **46**(8): 29−39)

26 Qiu Z Y, Martínez-Sánchez J, Arias-Sánchez P, Rashdi R. External multi-modal imaging sensor calibration for sensor fusion: A review. *Information Fusion*, 2023, **97**: Article No. 101806

27 Tan Zhen, Niu Zhong-Yan, Zhang Jin-Pu, Chen Xie-Yuan-Li, Hu De-Wen. New opportunities in SLAM——Gaussian splatting technology. *Journal of Image and Graphics*, 2025, **30**(6): 1792−1807
(谭臻, 牛中颜, 张津浦, 陈谢沅澧, 胡德文. SLAM 新机遇——斯溅射技术. 中国图象图形学报, 2025, **30**(6): 1792−1807)

28 Yu Wei-Dong, Lu Jing, Cheng Han-Lei. Review of NeRF-based SLAM research. *Computer Systems & Applications*, 2025, **34**(4): 18−33
(喻伟东, 鲁静, 程晗蕾. 基于 NeRF 的 SLAM 研究综述. 计算机系统应用, 2025, **34**(4): 18−33)

29 Han Kai, Xu Juan. Comprehensive review of 3D scene rendering technique——Neural radiance fields. *Application Research of Computers*, 2024, **41**(8): 2252−2260
(韩开, 徐娟. 3D 场景渲染技术——神经辐射场的研究. 计算机应用研究, 2024, **41**(8): 2252−2260)

30 Mildenhall B, Srinivasan P P, Tancik M, Barron J T, Ramamoorthi R, Ng R. NeRF: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 2022, **65**(1): 99−106

31 Kerbl B, Kopanas G, Leimkühler T, Drettakis G. 3D Gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics* (*TOG*), 2023, **42**(4): Article No. 139

32 Cai Z P, Müller M. CLNeRF: Continual learning meets NeRF. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). Paris, France: IEEE, 2023. 23185−23194

33 Li B C, Yan Z K, Wu D, Jiang H Q, Zha H B. Learn to memorize and to forget: A continual learning perspective of dynamic SLAM. In: Proceedings of the 18th European Conference on Computer Vision. Milan, Italy: Springer, 2024. 41−57

34 Anthwal S, Ganotra D. An overview of optical flow-based approaches for motion segmentation. *The Imaging Science Journal*, 2019, **67**(5): 284−294

35 Wang Ke-Sai, Yao Xi-Fan, Huang Yu, Liu Min, Lu Yu-Qian. Review of visual SLAM in dynamic environment. *Robot*, 2021, **43**(6): 715−732
(王柯赛, 姚锡凡, 黄宇, 刘敏, 陆玉前. 动态环境下的视觉 SLAM 研究评述. 机器人, 2021, **43**(6): 715−732)

36 Kazerouni I A, Fitzgerald L, Dooly G, Toal D. A survey of state-of-the-art on visual SLAM. *Expert Systems With Applications*, 2022, **205**: Article No. 117734

37 Zhu Dong-Ying, Zhong Yong, Yang Guan-Ci, Li Yang. Research progress on motion segmentation of visual localization and mapping in dynamic environment. *Journal of Computer Applications*, 2023, **43**(8): 2537−2545
(朱东莹, 钟勇, 杨观赐, 李杨. 动态环境下视觉定位与建图的运动分割研究进展. 计算机应用, 2023, **43**(8): 2537−2545)

38 Saputra M R U, Markham A, Trigoni N. Visual SLAM and structure from motion in dynamic environments: A survey. *ACM Computing Surveys* (*CSUR*), 2019, **51**(2): Article No. 37

39 Sturm J, Engelhard N, Endres F, Burgard W, Cremers D. A benchmark for the evaluation of RGB-D SLAM systems. In: Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems. Vilamoura-Algarve, Portugal: IEEE, 2012. 573−580

40 Palazzolo E, Behley J, Lottes P, Giguère P, Stachniss C. ReFusion: 3D reconstruction in dynamic environments for RGB-D cameras exploiting residuals. In: Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). Macao, China: IEEE, 2019. 7855−7862

41 Shi X S, Li D J, Zhao P P, Tian Q B, Tian Y X, Long Q W, et al. Are we ready for service robots? The OpenLORIS-Scene datasets for lifelong SLAM. In: Proceedings of the IEEE International Conference on Robotics and Automation (ICRA). Paris, France: IEEE, 2020. 3139−3145

42 Handa A, Whelan T, McDonald J, Davison A J. A benchmark for RGB-D visual odometry, 3D reconstruction and SLAM. In: Proceedings of the IEEE International Conference on Robotics and Automation (ICRA). Hong Kong, China: IEEE, 2014. 1524−1531

43 Choi S, Zhou Q Y, Koltun V. Robust reconstruction of indoor scenes. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Boston, USA: IEEE, 2015. 5556−5565

44 Geiger A, Lenz P, Urtasun R. Are we ready for autonomous driving? The KITTI vision benchmark suite. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Providence, USA: IEEE, 2012. 3354−3361

45 Barnes D, Gadd M, Murcutt P, Newman P, Posner I. The Oxford radar RobotCar dataset: A radar extension to the Oxford RobotCar dataset. In: Proceedings of the IEEE International Conference on Robotics and Automation (ICRA). Paris, France: IEEE, 2020. 6433−6438

46 Neuhold G, Ollmann T, Rota Bulò S, Kontschieder P. The mapillary vistas dataset for semantic understanding of street scenes. In: Proceedings of the IEEE International Conference on Computer Vision. Venice, Italy: IEEE, 2017. 4990−4999

47 Huang X Y, Cheng X J, Geng Q C, Cao B B, Zhou D F, Wang P, et al. The ApolloScape dataset for autonomous driving. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. Salt Lake City, USA: IEEE, 2018. 954−960

48 Cortés S, Solin A, Rahtu E, Kannala J. ADVIO: An authentic dataset for visual-inertial odometry. In: Proceedings of the 15th European Conference on Computer Vision (ECCV). Munich, Germany: Springer, 2018. 419−434

49 Yin J, Li A, Li T, Yu W X, Zou D P. M2DGR: A multi-sensor and multi-scenario SLAM dataset for ground robots. *IEEE Robotics and Automation Letters*, 2022, **7**(2): 2266−2273

50 Amigoni F, Schiaffonati V. *Methods and Experimental Techniques in Computer Engineering*. Cham: Springer, 2013.

51 Burri M, Nikolic J, Gohl P, Schneider T, Rehder J, Omari S, et al. The EuRoC micro aerial vehicle datasets. *The International Journal of Robotics Research*, 2016, **35**(10): 1157−1163

52 Wei H X, Jiao J H, Hu X C, Yu J W, Xie X P, Wu J, et al. FusionPortableV2: A unified multi-sensor dataset for generalized SLAM across diverse platforms and scalable environments. *The International Journal of Robotics Research*, 2025, **44**(7): 1093−1116

53 Zhang D T, Zhang J J, Sun Y, Li T, Yin H, Xie H Z, et al. Towards robust sensor-fusion ground SLAM: A comprehensive benchmark and a resilient framework. arXiv preprint arXiv: 2507.08364, 2025.

54 Dou H X, Liu B, Jia Y H, Wang C H. Monocular initialization for real-time feature-based SLAM in dynamic environments with multiple frames. *Sensors*, 2025, **25**(8): Article No. 2404

55 Chen X H, Wang T Y, Mai H N, Yang L J. SamSLAM: A visual SLAM based on segment anything model for dynamic environment. In: Proceedings of the 8th International Conference on Robotics, Control and Automation (ICRCA). Shanghai, China: IEEE, 2024. 91−97

56 Chen Z, Zang Q Y, Zhang K H. DZ-SLAM: A SAM-based SLAM algorithm oriented to dynamic environments. *Displays*, 2024, **85**: Article No. 102846

57 Hu Z M, Fang H, Zhong R, Wei S Z, Xu B C, Dou L H. GMP-SLAM: A real-time RGB-D SLAM in dynamic environments using GPU dynamic points detection method. *IFAC-PapersOnLine*, 2023, **56**(2): 5033−5040

58 Wang K S, Yao X F, Ma N F, Jing X. Real-time motion removal based on point correlations for RGB-D SLAM in indoor dynamic environments. *Neural Computing and Applications*, 2023, **35**(12): 8707−8722

59 Liu H L, Tian L F, Du Q L, Duan R X. RED-SLAM: Real-time and effective RGB-D SLAM with spatial-geometric observations and fast semantic perception for dynamic environments. *Measurement Science and Technology*, 2025, **36**(3): Article No. 036303

60 Zhang L X, Xu B L, Chen S W, Nener B, Zhou X, Lu M L, et al. An inpainting SLAM approach for detecting and recovering regions with dynamic objects. *Journal of Intelligent & Robotic Systems*, 2025, **111**(1): Article No. 29

61 Zhu F, Zhao Y F, Chen Z Y, Jiang C M, Zhu H, Hu X X. DyGS-SLAM: Realistic map reconstruction in dynamic scenes based on double-constrained visual SLAM. *Remote Sensing*, 2025, **17**(4): Article No. 625

62 Luo Y, Rao Z R, Wu R S. FD-SLAM: A semantic SLAM based on enhanced fast-SCNN dynamic region detection and DeepFillv2-driven background inpainting. *IEEE Access*, 2023, **11**: 110615−110626

63 Yang X B, Wang T, Wang Y Y, Lang C Y, Jin Y, Li Y D. FND-SLAM: A SLAM system using feature points and NeRF in dynamic environments based on RGB-D sensors. *IEEE Sensors Journal*, 2025, **25**(5): 8598−8610

64 Huang S C, Ren W H, Li M X. PLFF-SLAM: A point and line feature fused visual SLAM algorithm for dynamic illumination environments. *IEEE Access*, 2025, **13**: 34946−34953

65 Qi H B, Chen X C, Yu Z G, Li C, Shi Y L, Zhao Q R, et al. Semantic-independent dynamic SLAM based on geometric reclustering and optical flow residuals. *IEEE Transactions on Circuits and Systems for Video Technology*, 2025, **35**(3): 2244−2259

66 Cheng Peng, Wang Ke, Deng Gan-Lin, Li Yan-Long, Li Peng. Multi-feature visual Manhattan-SLAM with improved geometric constraints. *Journal of Navigation and Positioning*, 2025, **13**(2): 172−178
(程鹏, 王珂, 邓甘霖, 李炎隆, 李鹏. 改进几何约束的多特征视觉 Manhattan-SLAM. 导航定位学报, 2025, **13**(2): 172−178)

67 Li Yong, Liu Hong-Jie, Zhou Yong-Lu, Yu Ying. Front end op-

timization method of RGB-D SLAM in indoor dynamic scenes. *Application Research of Computers*, 2023, **40**(4): 991−995
(李泳, 刘宏杰, 周永录, 余映. 一种室内动态场景下 RGB-D SLAM 的前端优化方法. 计算机应用研究, 2023, **40**(4): 991−995)

68 Xie Ying, Shi Yong-Kang. Binocular SLAM based on projection transformation and optical flow. *Electronics Optics & Control*, 2024, **31**(9): 98−103
(谢颖, 石永康. 融合投影变换与光流的双目视觉 SLAM 研究. 电光与控制, 2024, **31**(9): 98−103)

69 He W Z, Lu Z K, Liu X, Xu Z W, Zhang J S, Yang C, et al. A real-time and high precision hardware implementation of RANSAC algorithm for visual SLAM achieving mismatched feature point pair elimination. *IEEE Transactions on Circuits and Systems I: Regular Papers*, 2024, **71**(11): 5102−5114

70 Li T, Yu Z B, Guan B L, Han J L, Lv W M, Fraundorfer F. Trifocal tensor and relative pose estimation with known vertical direction. *IEEE Robotics and Automation Letters*, 2025, **10**(2): 1305−1312

71 Yang Z Y, He Y, Zhao K, Lang Q, Duan H, Xiong Y H, et al. Research on inter-frame feature mismatch removal method of VSLAM in dynamic scenes. *Sensors*, 2024, **24**(3): Article No. 1007

72 Yang Yong-Gang, Wu Chu-Jian, Yang Zheng-Quan. Research on UAV visual SLAM based on fusing improved RANSAC optical flow method. *Semiconductor Optoelectronics*, 2023, **44**(2): 277−283
(杨永刚, 武楚健, 杨正全. 基于融合改进 RANSAC 光流法的无人机视觉 SLAM 研究. 半导体光电, 2023, **44**(2): 277−283)

73 Zhao X, Ding W C, An Y Q, Du Y L, Yu T, Li M, et al. Fast segment anything. arXiv preprint arXiv: 2306.12156, 2023.

74 Li Jia-Hui, Fan Xin-Yue, Zhang Gan, Zhang Kuo. Dynamic SLAM based on background restoration. *Journal of Data Acquisition and Processing*, 2024, **39**(5): 1204−1213
(李嘉辉, 范馨月, 张干, 张阔. 基于背景修复的动态 SLAM. 数据采集与处理, 2024, **39**(5): 1204−1213)

75 Ruan C Y, Zang Q Y, Zhang K H, Huang K. DN-SLAM: A visual SLAM with ORB features and NeRF mapping in dynamic environments. *IEEE Sensors Journal*, 2024, **24**(4): 5279−5287

76 Xu Z H, Niu J W, Li Q F, Ren T, Chen C. NID-SLAM: Neural implicit representation-based RGB-D SLAM in dynamic environments. In: Proceedings of the IEEE International Conference on Multimedia and Expo (ICME). Niagara Falls, Canada: IEEE, 2024. 1−6

77 Li M R, Guo Z T, Deng T C, Zhou Y M, Ren Y X, Wang H Y. DDN-SLAM: Real time dense dynamic neural implicit SLAM. *IEEE Robotics and Automation Letters*, 2025, **10**(5): 4300−4307

78 Redmon J, Divvala S, Girshick R, Farhadi A. You only look once: Unified, real-time object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, USA: IEEE, 2016. 779−788

79 Badrinarayanan V, Kendall A, Cipolla R. SegNet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017, **39**(12): 2481−2495

80 He K M, Gkioxari G, Dollár P, Girshick R. Mask R-CNN. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). Venice, Italy: IEEE, 2017. 2961−2969

81 Bolya D, Zhou C, Xiao F Y, Lee Y J. YOLACT: Real-time instance segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). Seoul, Korea: IEEE, 2019. 9157−9166

82 Yu C, Liu Z X, Liu X J, Xie F G, Yang Y, Wei Q, et al. DS-SLAM: A semantic visual SLAM towards dynamic environments. In: Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). Madrid, Spain: IEEE, 2018. 1168−1174

83 Bescos B, Fácil J M, Civera J, Neira J. DynaSLAM: Tracking, mapping, and inpainting in dynamic scenes. *IEEE Robotics and Automation Letters*, 2018, **3**(4): 4076−4083

84 Zhong F W, Wang S, Zhang Z Q, Chen C N, Wang Y Z. Detect-SLAM: Making object detection and SLAM mutually beneficial. In: Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV). Lake Tahoe, USA: IEEE, 2018. 1001−1010

85 Long X D, Zhang W W, Zhao B. PSPNet-SLAM: A semantic SLAM detect dynamic object by pyramid scene parsing network. *IEEE Access*, 2020, **8**: 214685−214695

86 Cheng S H, Sun C H, Zhang S J, Zhang D F. SG-SLAM: A real-time RGB-D visual SLAM toward dynamic scenes with semantic and geometric information. *IEEE Transactions on Instrumentation and Measurement*, 2023, **72**: Article No. 7501012

87 Wei W H, Huang K Z, Liu X, Zhou Y F. GSL-VO: A geometric-semantic information enhanced lightweight visual odometry in dynamic environments. *IEEE Transactions on Instrumentation and Measurement*, 2023, **72**: Article No. 2522513

88 Girshick R, Donahue J, Darrell T, Malik J. Rich feature hierarchies for accurate object detection and semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Columbus, USA: IEEE, 2014. 580−587

89 Girshick R. Fast R-CNN. In: Proceedings of the IEEE International Conference on Computer Vision. Santiago, Chile: IEEE, 2015. 1440−1448

90 Ren S Q, He K M, Girshick R, Sun J. Faster R-CNN: Towards real-time object detection with region proposal networks. In: Proceedings of the 29th International Conference on Neural Information Processing Systems. Montreal, Canada: MIT Press, 2015. 91−99

91 Liu W, Anguelov D, Erhan D, Szegedy C, Reed S, Fu C Y, et al. SSD: Single shot multibox detector. In: Proceedings of the 14th European Conference on Computer Vision. Amsterdam, The Netherlands: Springer, 2016. 21−37

92 Jocher G. Ultralytics YOLOv5 [Online], available: https://github.com/ultralytics/yolov5, October 25, 2024

93 Wang S, Xia C L, Lv F, Shi Y F. RT-DETRv3: Real-time end-to-end object detection with hierarchical dense positive supervision. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). Tucson, USA: IEEE, 2025. 1628−1636

94 Lei M Q, Li S Q, Wu Y H, Hu H, Zhou Y, Zheng X H, et al. YOLOv13: Real-time object detection with hypergraph-enhanced adaptive visual perception. arXiv preprint arXiv: 2506.17733, 2025.

95 Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Boston, USA: IEEE, 2015. 3431−3440

96 Chen L C, Papandreou G, Schroff F, Adam H. Rethinking atrous convolution for semantic image segmentation. arXiv preprint arXiv: 1706.05587, 2017.

97 Xie E Z, Wang W H, Yu Z D, Anandkumar A, Alvarez J M, Luo P. SegFormer: Simple and efficient design for semantic segmentation with Transformers. In: Proceedings of the 35th International Conference on Neural Information Processing Systems. Virtual Event: Curran Associates Inc., 2021. Article No. 924

98 Cheng B W, Misra I, Schwing A G, Kirillov A, Girdhar R. Masked-attention mask Transformer for universal image segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans, USA: IEEE, 2022. 1290−1299

99 Kirillov A, Mintun E, Ravi N, Mao H Z, Rolland C, Gustafson L, et al. Segment anything. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). Paris, France: IEEE, 2023. 4015−4026

100 Chng Y X, Zheng H, Han Y Z, Qiu X C, Huang G. Mask grounding for referring image segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, USA: IEEE, 2024. 26573−26583

101 Li X T, Yuan H B, Li W, Ding H H, Wu S Z, Zhang W W, et al. OMG-Seg: Is one model good enough for all segmentation? In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, USA: IEEE, 2024. 27948−27959

102 Guo G Q, Guo Y, Yu X H, Li W B, Wang Y X, Gao S. Segment any-quality images with generative latent space enhancement. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville, USA: IEEE, 2025. 2366−2376

103 Zhang Wei-Qi, Wang Jia, Zhang Lin, Ma Zong-Fang. SUI-SLAM: A semantics and uncertainty incorporated visual SLAM algorithm towards dynamic indoor environments. *Robot*, 2024, **46**(6): 732−742
(张玮奇, 王嘉, 张琳, 马宗方. SUI-SLAM: 一种面向室内动态环境的融合语义和不确定度的视觉 SLAM 方法. 机器人, 2024, **46**(6): 732−742)

104 Xiao L H, Wang J G, Qiu X S, Rong Z, Zou X D. Dynamic-SLAM: Semantic monocular visual localization and mapping based on deep learning in dynamic environment. *Robotics and Autonomous Systems*, 2019, **117**: 1−16

105 Jia Z Q, Ma Y X, Lai J W, Wang Z G. DOA-SLAM: An efficient stereo visual SLAM system in dynamic environment. *International Journal of Control, Automation and Systems*, 2025, **23**(4): 1181−1198

106 Pan G G, Cao S Y, Lv S, Yi Y. DEG-SLAM: A dynamic visual RGB-D SLAM based on object detection and geometric constraints for degenerate motion. *Measurement Science and Technology*, 2025, **36**(2): Article No. 026302

107 Yang J M, Wang Y T, Tan X, Fang M E, Ma L Z. DHP-SLAM: A real-time visual SLAM system with high positioning accuracy under dynamic environment. *Displays*, 2025, **89**: Article No. 103067

108 Xu B W, Zheng Z X, Pan Z H, Yu L. HMC-SLAM: A robust VSLAM based on RGB-D camera in dynamic environment combined hierarchical multidimensional clustering algorithm. *IEEE Transactions on Instrumentation and Measurement*, 2025, **74**: Article No. 5020311

109 Cui H J, Zhao X, Luo G Y. DYMRO-SLAM: A robust stereo visual SLAM for dynamic environments leveraging mask R-CNN and optical flow. *IEEE Access*, 2025, **13**: 54240−54253

110 Li C, Jiang S, Zhou K Q. DYR-SLAM: Enhanced dynamic visual SLAM with YOLOv8 and RTAB-Map. *The Journal of Supercomputing*, 2025, **81**(5): Article No. 718

111 Du X W, Zhang C L, Gao K H, Liu J, Yu X F, Wang S S. YPL-SLAM: A simultaneous localization and mapping algorithm for point-line fusion in dynamic environments. *Sensors*, 2024, **24**(14): Article No. 4517

112 Zhu S T, Qin R J, Wang G M, Liu J M, Wang H S. SemGauss-SLAM: Dense semantic Gaussian splatting SLAM. arXiv preprint arXiv: 2403.07494, 2024.

113 Liu H S, Wang L, Luo H Y, Zhao F, Chen R Z, Chen Y S, et al. SDD-SLAM: Semantic-driven dynamic SLAM with Gaussian splatting. *IEEE Robotics and Automation Letters*, 2025, **10**(6): 5721−5728

114 Wu W X, Guo L, Gao H L, You Z C, Liu Y K, Chen Z Q. YOLO-SLAM: A semantic SLAM system towards dynamic environment with geometric constraint. *Neural Computing and Applications*, 2022, **34**(8): 6011−6026

115 Liu Y B, Miura J. RDS-SLAM: Real-time dynamic SLAM using semantic segmentation methods. *IEEE Access*, 2021, **9**: 23772−23785

116 Liu Y B, Miura J. RDMO-SLAM: Real-time visual SLAM for dynamic environments using semantic label prediction with optical flow. *IEEE Access*, 2021, **9**: 106981−106997

117 Jiang Li-Mei, Chen Xin-Wei. Visual SLAM algorithm based on feature point selection in dynamic scenes. *Journal of System Simulation*, 2025, **37**(3): 753−762
(姜丽梅, 陈信威. 动态场景下基于特征点筛选的视觉 SLAM 算法. 系统仿真学报, 2025, **37**(3): 753−762)

118 Cheng Qiang, Zhang You-Bing, Zhou Kui. Dynamic visual SLAM method based on improved YOLOX. *Electronic Measurement Technology*, 2024, **47**(23): 123−133
(程强, 张友兵, 周奎. 基于改进 YOLOX 的动态视觉 SLAM 方法. 电子测量技术, 2024, **47**(23): 123−133)

119 Huang You-Rui, Wang Zhao-Feng, Han Tao, Song Hong-Ping. Dynamic visual SLAM algorithm combined with lightweight YOLOv5s. *Electronic Measurement Technology*, 2024, **47**(11): 59−68
(黄友锐, 王照锋, 韩涛, 宋红萍. 结合轻量化 YOLOv5s 的动态视觉 SLAM 算法. 电子测量技术, 2024, **47**(11): 59−68)

120 Cheng G Y, Jia J F, Pang X Q, Wen J, Shi Y H, Zeng J C. DFE-SLAM: Dynamic SLAM based on improved feature extraction. In: Proceedings of the China Automation Congress (CAC). Qingdao, China: IEEE, 2024. 4584−4589

121 Feng D, Yin Z Y, Wang X H, Zhang F Q, Wang Z S. YLS-SLAM: A real-time dynamic visual SLAM based on semantic segmentation. *Industrial Robot*, 2025, **52**(1): 106−115

122 Lian X F, Kang M M, Tan L, Sun X, Wang Y L. LSSMask: A lightweight semantic segmentation network for dynamic object. *Signal, Image and Video Processing*, 2025, **19**(3): Article No. 216

123 Zhang Y Q, Jiang G G, Li M R, Feng G S. UE-SLAM: Monocular neural radiance field SLAM with semantic mapping capabilities. *Symmetry*, 2025, **17**(4): Article No. 508

124 Qin T, Li P L, Shen S J. VINS-Mono: A robust and versatile monocular visual-inertial state estimator. *IEEE Transactions on Robotics*, 2018, **34**(4): 1004−1020

125 Mourikis A I, Roumeliotis S I. A multi-state constraint Kalman filter for vision-aided inertial navigation. In: Proceedings of the IEEE International Conference on Robotics and Automation. Rome, Italy: IEEE, 2007. 3565−3572

126 Yang Guan-Ci, Wang Xiao-Yuan, Jiang Ya-Wen, Li Yang. Review of SLAM technologies based on visual and inertial sensor fusion. *Journal of Guizhou University* (*Natural Sciences*), 2020, **37**(6): 1−12
(杨观赐, 王霄远, 蒋亚汶, 李杨. 视觉与惯性传感器融合的 SLAM 技术综述. 贵州大学学报 (自然科学版), 2020, **37**(6): 1−12)

127 Bloesch M, Omari S, Hutter M, Siegwart R. Robust visual inertial odometry using a direct EKF-based approach. In: Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). Hamburg, Germany: IEEE, 2015. 298−304

128 Wu K Z, Zhang T, Su D, Huang S D, Dissanayake G. An invariant-EKF VINS algorithm for improving consistency. In: Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). Vancouver, Canada: IEEE, 2017. 1578−1585

129 Chen S M, Frémont V. A loosely coupled vision-LiDAR odometry using covariance intersection filtering. In: Proceedings of the IEEE Intelligent Vehicles Symposium (IV). Nagoya, Japan: IEEE, 2021. 1102−1107

130 Zhu Y Q, Jin R, Lou T S, Zhao L Y. PLD-VINS: RGBD visual-inertial SLAM with point and line features. *Aerospace Science and Technology*, 2021, **119**: Article No. 107185

131 Zhang L C, Yin H L, Ye W, Betz J. DP-VINS: Dynamics adaptive plane-based visual-inertial SLAM for autonomous vehicles. *IEEE Transactions on Instrumentation and Measurement*, 2024, **73**: Article No. 5036516

132 Yin H S, Li S M, Tao Y, Guo J L, Huang B. Dynam-SLAM: An accurate, robust stereo visual-inertial SLAM method in dynamic environments. *IEEE Transactions on Robotics*, 2023, **39**(1): 289−308

133 Abdollahi M R, Pourtakdoust S H, Nooshabadi M H Y, Pishkenari H N. An improved multi-state constraint Kalman filter for visual-inertial odometry. *Journal of the Franklin Institute*, 2024, **361**(15): Article No. 107130

134 Jung K, Song J, Seong S, Myung H. MSCKF-DVIO: Multi-state constraint Kalman filter based RGB-D visual-inertial odometry with spline interpolation and nonholonomic constraint. In: Proceedings of the 21st International Conference on Ubiquitous Robots (UR). New York, USA: IEEE, 2024. 558−565

135 Fornasier A, van Goor P, Allak E, Mahony R, Weiss S. MSCeqF: A multi state constraint equivariant filter for vision-aided inertial navigation. *IEEE Robotics and Automation Letters*, 2024, **9**(1): 731−738

136 Cao L, Liu J B, Lei J T, Zhang W, Chen Y S, Hyyppä J. Real-time motion state estimation of feature points based on optical flow field for robust monocular visual-inertial odometry in dynamic scenes. *Expert Systems With Applications*, 2025, **274**: Article No. 126813

137 Yu Z L. A self-adaptation feature correspondences identification algorithm in terms of IMU-aided information fusion for VINS. *Applied Intelligence*, 2025, **55**(3): Article No. 202

138 Li Yang-Ming, Song Quan-Jun, Liu Hai, Ge Yun-Jian. General purpose LIDAR feature extractor for mobile robot navigation. *Journal of Huazhong University of Science & Technology* (*Natural Science Edition*), 2013, **41**(S1): 280−283
(李阳铭, 宋全军, 刘海, 葛运建. 用于移动机器人导航的通用激光雷达特征提取. 华中科技大学学报 (自然科学版), 2013, **41**(S1): 280−283)

139 Du K Y, Meng J, Meng X H, Xiang Z H, Wang S F, Yang J H. Projection mapping segmentation block: A fusion approach of pointcloud and image for multi-objects classification. *IEEE Access*, 2023, **11**: 77802−77809

140 Tan M K, Zhuang Z W, Chen S T, Li R, Jia K, Wang Q C, et al. EPMF: Efficient perception-aware multi-sensor fusion for 3D semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024, **46**(12): 8258−8273

141 Song H M, Cho J, Ha J S, Park J, Jo K. Panoptic-FusionNet: Camera-LiDAR fusion-based point cloud panoptic segmentation for autonomous driving. *Expert Systems With Applications*, 2024, **251**: Article No. 123950

142 Shi H S, Wang X, Zhao J H, Hua X N. A cross-modal attention-driven multi-sensor fusion method for semantic segmentation of point clouds. *Sensors*, 2025, **25**(8): Article No. 2474

143 Sánchez-García F, Montiel-Marín S, Antunes-García M, Gutiérrez-Moreno R, Llamazares Á L, Bergasa L M. SalsaNext+: A multimodal-based point cloud semantic segmentation with range and RGB images. *IEEE Access*, 2025, **13**: 64133−64147

144 Zhu D Y, Yang G C. RDynaSLAM: Fusing 4D radar point clouds to visual SLAM in dynamic environments. *Journal of Intelligent & Robotic Systems*, 2025, **111**(1): Article No. 11

145 Lin J R, Zhang F. R3LIVE: A robust, real-time, RGB-colored, LiDAR-inertial-visual tightly-coupled state estimation and mapping package. In: Proceedings of the International Conference on Robotics and Automation (ICRA). Philadelphia, USA: IEEE, 2022. 10672−10678

146 Li X X, Wang X B, Liao J C, Li X, Li S Y, Lv H B. Semi-tightly coupled integration of multi-GNSS PPP and S-VINS for precise positioning in GNSS-challenged environments. *Satellite Navigation*, 2021, **2**(1): Article No. 1

147 Cao S Z, Lu X Y, Shen S J. GVINS: Tightly coupled GNSS-visual-inertial fusion for smooth and consistent state estimation. *IEEE Transactions on Robotics*, 2022, **38**(4): 2004−2021

148 Zhang T W, Zhang H Y, Li X F. Vision-audio fusion SLAM in dynamic environments. *CAAI Transactions on Intelligence Technology*, 2023, **8**(4): 1364−1373

149 Song C Q, Zeng B, Cheng J, Wu F X, Hao F S. PSMD-SLAM: Panoptic segmentation-aided multi-sensor fusion simultaneous localization and mapping in dynamic scenes. *Applied Sciences*, 2024, **14**(9): Article No. 3843

150 Lang X L, Li L J, Wu C M, Zhao C, Liu L N, Liu Y, et al. Gaussian-LIC: Real-time photorealistic SLAM with Gaussian splatting and LiDAR-inertial-camera fusion. In: Proceedings of the IEEE International Conference on Robotics and Automation (ICRA). Atlanta: IEEE, 2025. 8500−8507

151 Qu D L, Yan C, Wang D, Yin J, Chen Q Z, Xu D, et al. Implicit event-RGBD neural SLAM. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE, 2024. 19584−19594

152 Vidas S, Sridharan S. Hand-held monocular SLAM in thermal-infrared. In: Proceedings of the 12th International Conference on Control Automation Robotics & Vision (ICARCV). Guangzhou, China: IEEE, 2012. 859−864

153 Vidas S, Moghadam P, Bosse M. 3D thermal mapping of building interiors using an RGB-D and thermal camera. In: Proceedings of the IEEE International Conference on Robotics and Automation. Karlsruhe, Germany: IEEE, 2013. 2311−2318

154 Mouats T, Aouf N, Sappa A D, Aguilera C, Toledo R. Multispectral stereo odometry. *IEEE Transactions on Intelligent Transportation Systems*, 2015, **16**(3): 1210−1224

155 Qin L, Wu C, Kong X T, You Y, Zhao Z Q. BVT-SLAM: A binocular visible-thermal sensors SLAM system in low-light environments. *IEEE Sensors Journal*, 2024, **24**(7): 11599−11609

156 Li S, Ma X F, He R, Shen Y R, Guan H, Liu H Z, et al. WTI-SLAM: A novel thermal infrared visual SLAM algorithm for weak texture thermal infrared images. *Complex & Intelligent Systems*, 2025, **11**(6): Article No. 242

157 Qin T, Cao S Z, Pan J, Shen S J. A general optimization-based framework for global pose estimation with multiple sensors. arXiv preprint arXiv: 1901.03642, 2019.

158 Yin J, Li A, Xi W, Yu W X, Zou D P. Ground-fusion: A low-cost ground SLAM system robust to corner cases. In: Proceedings of the IEEE International Conference on Robotics and Automation (ICRA). Yokohama, Japan: IEEE, 2024. 8603−8609

159 Ou Y M, Fan J F, Zhou C, Zhang P J, Zeng G. Hybrid-VINS: Underwater tightly coupled hybrid visual inertial dense SLAM for AUV. *IEEE Transactions on Industrial Electronics*, 2025, **72**(3): 2821−2831

160 Peng X F, Liu Z H, Li W M, Tan P, Cho S Y, Wang Q. DVI-SLAM: A dual visual inertial SLAM network. In: Proceedings of the IEEE International Conference on Robotics and Automation (ICRA). Yokohama, Japan: IEEE, 2024. 12020−12026

161 Chen Peng-Yu, Nie Xiu-Shan, Li Nan-Jun, Li Tuo. Semi-supervised video object segmentation method based on spatio-temporal decoupling and regional robustness enhancement. *Journal of Computer Applications*, 2025, **45**(5): 1379−1386
(陈鹏宇, 聂秀山, 李南君, 李拓. 基于时空解耦和区域鲁棒性增强的半监督视频目标分割方法. 计算机应用, 2025, **45**(5): 1379−1386)

162 Yang Guan-Ci, Yang Jing, Su Zhi-Dong, Chen Zhan-Jie. An improved YOLO feature extraction algorithm and its application to privacy situation detection of social robots. *Acta Automatica Sinica*, 2018, **44**(12): 2238−2249
(杨观赐, 杨静, 苏志东, 陈占杰. 改进的 YOLO 特征提取算法及其在服务机器人隐私情境检测中的应用. 自动化学报, 2018, **44**(12): 2238−2249)

163 Zhou Y H, Sun M L. A visual SLAM loop closure detection method based on lightweight Siamese capsule network. *Scientific Reports*, 2025, **15**(1): Article No. 7644

164 Zhang Xiao-Dong, Zhang Chao-Kun, Zhao Ji-Jun. State-of-the-art survey on edge intelligence. *Journal of Computer Research and Development*, 2023, **60**(12): 2749−2769
(张晓东, 张朝昆, 赵继军. 边缘智能研究进展. 计算机研究与发展, 2023, **60**(12): 2749−2769)

165 Satyanarayanan M. The emergence of edge computing. *Computer*, 2017, **50**(1): 30−39

166 He Gao-Xiang, Zhu Bin, Xie Bo, Chen Yi. Progress in novel view synthesis using neural radiance fields. *Laser & Optoelectronics Progress*, 2024, **61**(12): Article No. 1200005
(何高湘, 朱斌, 解博, 陈熠. 基于神经辐射场的新视角合成研究进展. 激光与光电子学进展, 2024, **61**(12): Article No. 1200005)

冯嘉琪 贵州大学现代制造技术教育部重点实验室硕士研究生. 主要研究方向为视觉 SLAM, 机器人感知.
E-mail: jiaqivon20@gmail.com
(**FENG Jia-Qi** Master student at the Key Laboratory of Advanced Manufacturing Technology of Ministry of Education, Guizhou University. His research interests include visual SLAM and robotic perception.)

杨恺伦 湖南大学人工智能与机器人学院教授. 2019 年获得浙江大学测控技术与仪器专业博士学位. 主要研究方向为多模态、高维度、全视角计算光学与计算视觉.
E-mail: kailun.yang@hnu.edu.cn
(**YANG Kai-Lun** Professor at the School of Artificial Intelligence and Robotics, Hunan University. He received his Ph.D. degree in measurement and control technology and instruments from Zhejiang University in 2019. His research interests include multimodal, high-dimensional, and omnidirectional computational optics and computer vision.)

林家丞 贵州大学现代制造技术教育部重点实验室特聘教授. 2025 年获得湖南大学计算机科学与技术专业博士学位. 主要研究方向为具身机器人的场景理解, 多模态融合认知.
E-mail: jcheng_lin@hnu.edu.cn
(**LIN Jia-Cheng** Distinguished professor at the Key Laboratory of Advanced Manufacturing Technology of Ministry of Education, Guizhou University. He received his Ph.D. degree in computer science and technology from Hunan University in 2025. His research interests include scene understanding for embodied robots and multimodal fusion cognition.)

杨观赐 贵州大学现代制造技术教育部重点实验室教授. 2012 年获得中国科学院大学计算机软件与理论专业博士学位. 主要研究方向为多模态融合认知, 智能机器人, 智能机器人技能学习. 本文通信作者.
E-mail: gcyang@gzu.edu.cn
(**YANG Guan-Ci** Professor at the Key Laboratory of Advanced Manufacturing Technology of Ministry of Education, Guizhou University. He received his Ph.D. degree in computer software and theory from University of Chinese Academy of Sciences in 2012. His research interests include multimodal fusion cognition, intelligent robots, and intelligent robot skill learning. Corresponding author of this paper.)