物体级语义视觉SLAM研究综述

田　瑞, 张云洲†, 杨凌昊, 曹振中

(东北大学 信息科学与工程学院, 辽宁 沈阳 110819)

摘要: 视觉同时定位与地图构建(Visual simultaneous localization and mapping, VSLAM)是自主移动机器人、自动驾驶、增强现实(AR)等领域的关键技术. 随着深度学习的发展, 准确高效的图像语义信息在VSLAM领域得到了广泛的应用. 与传统SLAM相比, 语义VSLAM利用语义信息提升了定位精度和鲁棒性,并通过物体级重建提高了环境感知能力, 成为当前VSLAM领域的研究热点. 本文对近年来优秀的物体级语义SLAM 工作进行了阐述归纳和对比梳理, 总结了该领域的4个关键问题, 包括物体表达形式、物体初始化方法、融合语义信息的数据关联算法和融合物体级语义信息的后端优化方法. 同时, 对代表性方法进行了优缺点分析. 最后, 在现有技术成果和研究基础上, 对物体级语义VSLAM面临的挑战和未来研究方向进行了展望和分析. 当前物体级语义SLAM仍面临着物体关联不准确、物体优化框架不完善等问题. 如何有效使用和维护语义地图以应用于决策规划等任务, 以及融合多源信息以丰富视觉感知是未来的研究热点.

关键词: 视觉SLAM; 数据关联; 语义分割; 物体级地图

引用格式: 田瑞, 张云洲, 杨凌昊, 等. 物体级语义视觉SLAM研究综述. 控制理论与应用, 2023, 40(12): 2160 – 2171

DOI: 10.7641/CTA.2023.30338

Survey of object-oriented semantic visual SLAM

TIAN Rui, ZHANG Yun-zhou†, YANG Ling-hao, CAO Zhen-zhong

(College of Information Science and Technology, Shenyang Liaoning 110819, China)

Abstract: Visual simultaneous localization and mapping (VSLAM) is a key technology for autonomous robots, autonomous navigation, and AR applications. With the development of deep learning, accurate and efficient semantic information has been widely used in VSLAM. Compared with traditional SLAM, semantic SLAM leverages semantic information to improve the accuracy and robustness of localization, and enhances environmental perception ability by object-level reconstruction, which has became the trend in VSLAM research. In this survey, we provide an overview of semantic SLAM techniques with state-of-the-art object SLAM systems. Four key issues of semantic SLAM are summarized, including object representation, object initialization methods, data association methods, and back-end optimization methods integrating semantic objects. The advantages and disadvantages of the comparison methods are provided. Finally, we propose the future work and challenges of object-level SLAM technology. Currently, semantic SLAM still faces problems such as inaccurate object association and an unified optimization framework has not yet been proposed. How to effectively use and maintain semantic maps for the application of decision and planning tasks, as well as integrate multi-source information to enrich visual perception, will be future research hotspots.

Key words: visual SLAM; data association; semantic information; Semantic mapping

Citation: TIAN Rui, ZHANG Yunzhou,YANG Linghao, et.al. Survey of Object-oriented Semantic visual SLAM. *Control Theory & Applications*, 2023, 40(12): 2160 – 2171

1　引言

视觉同时定位与建图(visual simultaneous localization and mapping, VSLAM)技术通过相机实现自主定位与地图构建, 相较于激光雷达,相机具有低成本、低功耗、强感知等特点, 且二维图像的语义信息更容易通过深度学习技术获取. 结合语义信息对环境中的物体进行建模, 并利用物体的语义不变性约束提升VSLAM的定位精度和鲁棒性成为当前研究的热点. 本文着重对物体级语义VSLAM的发展和关键技术进行讨论: 首先, 阐述了物体级语义信息在SLAM中的

收稿日期: 2023−05−19; 录用日期: 2023−11−21.

†通信作者. E-mail: zhangyunzhou@mail.neu.edu.cn; Tel.: +86 13940101976.

本文责任编委: 胡德文.

国家自然科学基金项目(61973066, 61471110)资助.

Supported by the National Natural Science Foundation of China (61973066, 61471110).

重要作用; 其次, 归纳了物体级语义SLAM技术的4个关键的问题(模型表达、物体初始化、数据关联、后端优化); 最后, 对语义SLAM面临的挑战和未来发展方向进行了展望. 本文结构框图如图1所示.

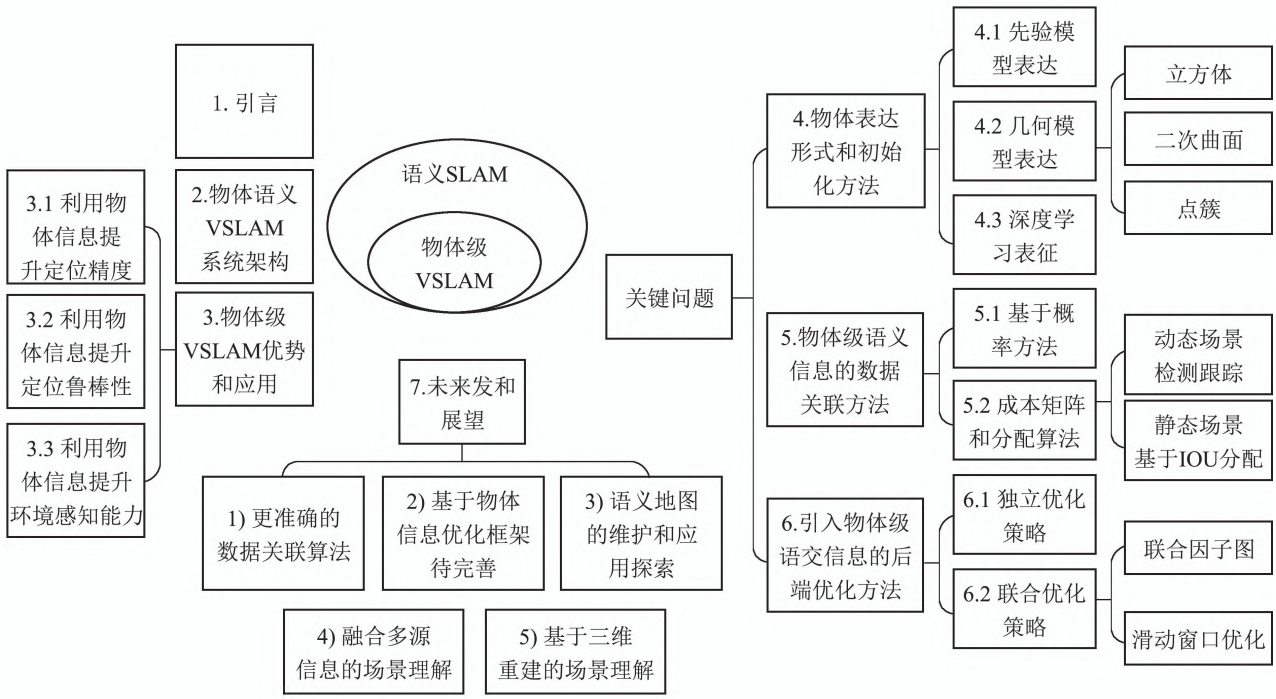


图 1　本文结构框图

Fig. 1　Structure of the survey

2　物体级语义VSLAM系统架构

物体级语义VSLAM一般采用多线程的算法架构, 分为前端和后端. 前端主要由跟踪线程和检测线程构成: 跟踪线程负责图像特征提取, 并通过帧间特征匹配和局部BA(bundle adjustment)优化求解相机位姿; 检测线程使用深度网络对输入图像进行语义信息提取,并将其送入到跟踪线程中. 图像语义信息是基于当前帧的检测结果, 因此, 使用物体数据关联对不同帧的检测信息进行处理, 并进行物体初始化. 后端优化线程负责相机和物体位姿优化, 以及对物体建模的参数进行调整. 最终, 系统构建了物体级的语义地图, 实现环境的语义感知.

语义信息的获取形式可以分为: 目标检测[1–3]、语义分割[4–8]、实例分割[9–10]. 不同的语义信息获取方式会影响算法的实时性, 通常, 语义分割网络耗时更长, 且语义分割得到的像素级分割结果存在信息冗余和误检, 目标检测网络效率更高, 但在复杂场景下容易出现漏检和误检的现象. 后端优化方式可以分为独立优化和联合优化策略, 例如, OA-SLAM(object assisted SLAM)[11]使用独立的线程来优化二次曲面参数, QuadricSLAM[12]则将物体和相机放在局部BA的统一框架下优化.

近年来, 融合目标检测和实例分割的物体级SLAM成为研究的热点, 该类方法通过多视图几何约束, 利用物体检测框重建物体模型. 重建模型可以分为二次曲面[13–20]、立方框[21]等. 实例分割可以获得更准确的物体实例掩码, 通常用于辅助物体特征提取和数据关联, 实现更准确的目标跟踪[22]. 常见的物体级VSLAM结构如图2所示.

3　物体级语义VSLAM优势和应用

传统的VSLAM一般通过点、线、面等几何元素构建地图, 例如, 稀疏点云地图[23]、稠密点云地图[24]、网格地图[25–26]、TSDF(truncated signed distance field)地图[27]等. 这些地图为自身定位和环境感知提供基础, 使得VSLAM技术得以广泛应用. 随着应用场景的增加, 人们发现传统VSLAM方法在定位精度和算法鲁棒性上具有局限性, 主要有如下原因：1) 动态干扰, 当前VSLAM算法大多基于环境静态假设, 特征匹配和优化容易受到外点干扰, 导致跟踪精度变差或者丢失. 2)光照变换, 传统的视觉特征在光照变化或者暗光条件下, 特征匹配和图像光度误差匹配失败, 导致无法实现位姿估计, 算法鲁棒性降低. 3) 高层次的语义感知需求, 传统的VSLAM在表征物体上具有局限性, 不具有语义信息, 无法满足人机交互等复杂任务的需求. 深度学习技术的引入为VSLAM定位和环境感知带了新的解决方法. 基于深度学习的特征提取技术为VSLAM在复杂光照条件下提供更稳定的匹配效

果[28–30], 实例分割或目标检测为物体的运动属性判断提供可能, 减少了VSLAM在复杂环境中受动态干扰的影响[31–32]. 通过构建的物体级地图和模型表达, 丰富了系统的环境感知能力[11, 14–19].

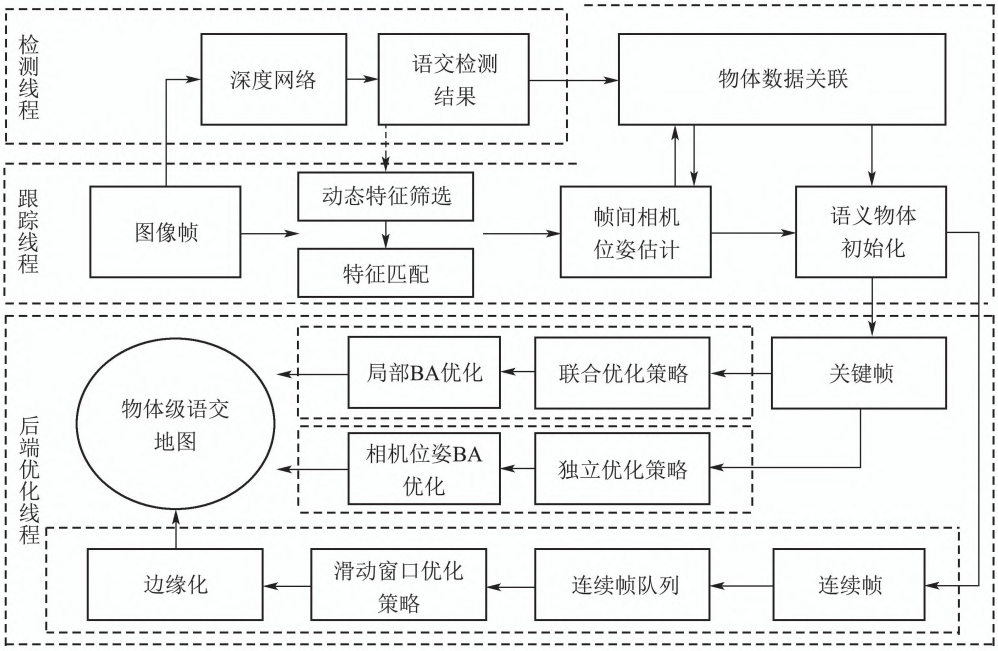


图 2 物体级语义VSLAM结构图

Fig. 2 Architecture of object VSLAM

3.1 利用物体信息提升定位精度

当前, 室内外场景下的VSLAM算法已经得到了长足的发展, 一些SLAM算法能够准确地构建环境地图, 并在一定程度上克服噪声、动态干扰和光照变化的影响. 例如, ORB-SLAM[33]、RGBD-SLAM[34]、LS-D-SLAM[35]等. 然而, 在实际应用部署中, 算法仍面临着场景动态干扰的影响. 早期的解决方案中[23], 使用运动一致性和基于外点剔除的RANSAC(random sample consensus)策略对由噪声干扰导致的错误特征匹配进行筛选, 或者在优化中引入鲁棒核函数来降低动态特征的优化权重, 例如, ORB-SLAM2[23]使用特征均匀提取和鲁棒核函数来降低错误匹配干扰.

近年来, 一些工作将物体检测结果的语义属性引入VSLAM中, 对场景中物体的动静态进行判断, 并剔除动态物体的干扰[36–40]. Detect-SLAM[41]通过目标检测剔除动态点, 并通过特征匹配和扩展区域进行运动概率传播, 在提升定位精度的同时提升了目标检测的稳定性. DS-SLAM[39]使用实例分割结果和运动一致性判断物体的运动属性, 并将动态特征进行剔除以提升定位精度. Dyna-SLAM[40]将落在运动物体掩码内的特征作为外点剔除, 从而提升其在动态场景下的定位鲁棒性. 类似的, Kaveti和Singh[42]提出了Light Field SLAM, 通过合成孔径成像技术重建被遮挡的静态场景, 不同于Bescos等人[43]的算法, 其进一步利用了重建背景的特征进行位姿跟踪以实现更好的定位性能.

针对基于深度学习的动态物体检测通常存在漏检和错检问题, Ballester 等人[44]提出了DOT-SLAM, 结合实例分割和多视图几何来生成动态物体掩码, 并通过最小化光度误差进行跟踪. 这种方法不仅提高了定位精度, 还提高了语义分割的精度. 上述工作的重点是通过剔除动态信息来提升自身定位的鲁棒和准确性, 但忽略了对场景中移动物体状态的感知.

作为VSLAM对动态场景理解的扩展, 结合运动跟踪的VSLAM成为当前研究的热点. Wang等人[45]首先提出了带有运动物体跟踪的SLAM, 将自身位姿估计和动态物体位姿估计分解为两个独立的状态估计问题. Kundu等人[46]结合SfM(structure from motion)和运动物体跟踪来解决运动场景下的SLAM问题, 该方法将系统输出统一到包含静态结构和运动物体轨迹的三维动态地图中. Huang 等人[47] 提出了Cluster-VO, 能够进行多个物体的运动估计. 该方法提出了一种多层概率关联机制来高效地跟踪物体特征, 利用异构条件随机场(conditional random filed, CRF)聚类方法进行物体关联, 最后在滑动窗口内优化物体的运动轨迹. Bescos 等人[43]将运动物体与自身状态估计问题紧耦合到统一框架中, 对跟踪点集使用主成分分析(principal component analysis, PCA)聚类和立方框建模, 并使用动态路标点对自身位姿进行约束.

考虑到场景的先验约束, Twist SLAM[48]使用机械关节约束来限制物体在特定场景位姿估计的自由度, 结合3D目标检测获得先验物体估计, 使用语义信息来构建物体点簇地图, 并利用静态簇(道路和房屋)来估计相机位姿. 动态簇则通过速度的变化进行跟踪和约束. VDO-SLAM[49]使用聚类点的形式对物体进行状态估计, 使用实例分割和稠密场景流, 提高了动态物体观测的数量和关联质量, 该方法将动态和静态结构集成到统一的估计框架中, 实现了对相机位姿和物体位姿的联合估计.

3.2　利用物体信息提升定位鲁棒性

传统的视觉定位大多采用手工描述了, 如ORB[50], SIFT[51]等特征, 并使用基于视觉词袋(bag of words, BOW)进行定位, 当图像视角变化或者光照发生明显改变时, 该方案的视觉定位会失效. 物体语义信息能有效克服大视角变换以及光照变换等情况, 为VSLAM提供更鲁棒的定位.

实时的物体级单目SLAM算法SLAM++[52]利用了一个大型物体数据库, 使用单词袋来识别对象, 实现鲁棒定位. Zins 等[11]提出的OA-SLAM利用重建的物体级语义地图进行相机重定位. 该方案结合了特征描述子和场景物体的重投影观测, 利用物体的相对位置关系约束, 在视角变化剧烈的场景下实现定位, 提升了视觉定位的鲁棒性. Liu等[53]提出基于物体级描述符的定位方法. 文献[54]提出基于深度网络的物体描述符定位方法.

CubeSLAM[55]利用物体立方框和当前帧的目标检测约束, 提升系统在无纹理场景下的定位鲁棒性. QuadricSLAM[12]提出基于二次曲面的物体观测约束, 首次使用3D椭球作为路标, 同时使用一个联合优化框架, 将相机位姿和二次曲面联合优化. 文献[56]利用单目视觉构建的物体级路标和物体先验大小约束, 减少了单目定位的尺度漂移, 提升了单目视觉的定位精度和鲁棒性. 类似的方案如文献[57–58], 采用物体先验尺度约束单目定位漂移. EAO-SLAM[21]则使用物体立方框约束构建观测误差, 减少了定位漂移.

可以看出, 融合物体语义信息已经成为了提高视觉定位精度和鲁棒性的有效途径之一. 语义信息已经广泛应用于SLAM系统的初始化、后端优化、重定位和闭环检测等阶段. 因此, 有效地处理和利用语义信息是提高定位精度的关键.

3.3　利用物体信息提升系统环境感知能力

VSLAM构建的地图可以分为: 稀疏点云地图[23]、稠密地图[27]、半稠密地图[24]、结构地图[59–60]、平面地图[61–65]、物体级地图[13–19,52]等. 点云地图中仅具有点云结构信息, 通常用于为SLAM提供定位约束. 半稠密和稠密地图可以更精细地表达环境. 结构地图和平面地图通过抽象的场景点线面的结构, 为场景提供轻量级的地图表达. 然而, 上述的地图表达形式缺少对环境的高层次语义感知能力.

近年来, 随着自动驾驶、人机交互等领域的兴起, 环境的语义感知越来越受到研究者的重视. 语义信息的融入为SLAM的地图提供更为丰富的感知信息.

早期的物体SLAM, 例如, SLAM++[52]利用物体CAD模型构建语义地图, 通过目标检测和识别, 将先验物体数据库的物体加载在地图中. 文献[37]将语义标签信息融合到稠密点云地图中, 构建了稠密语义地图. CubeSLAM[55]和EAO-SLAM[21]通过立方框构建物体级地图.

文献[13–19]构建了物体的二次曲面地图, 同时估计了物体的大小、旋转和位置. 相比于二次曲面和立方体的包络, 超二次曲面可以通过调节二次模型参数适应不同形状的物体, 丰富环境物体的表达. 文献[66]使用超二次曲面构建室内场景的物体级地图. 一些工作将抽象的语义标识加入到地图表达中, AVP- SLAM[67]通过检测道路的车道线, 交通标识等信息构建了轻量级的语义地图, 用于实现准确的室外场景定位.

另外, 一些研究者将运动物体的感知信息加入到SLAM中, 提出了SLAM-MOT[22,47,68], 在构建场景稀疏点云地图的同时, 表达物体的运动轨迹, 构建包含运动信息的物体地图. 例如, VDO-SLAM[49]提出利用语义信息构建环境结构, 跟踪刚性物体的运动并估计其三维运动轨迹, 其地图表示如图3所示.

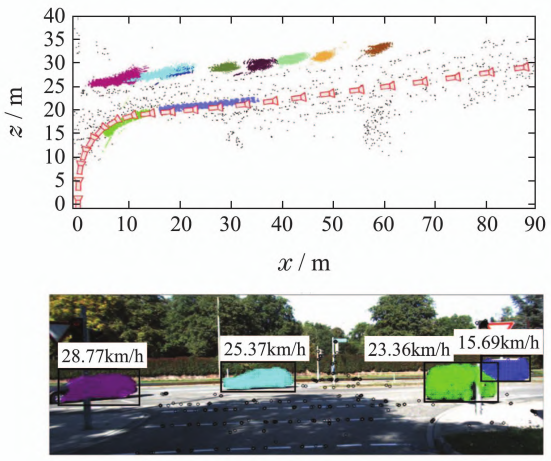


图 3　VDO-SLAM系统可视化地图[49], 包含运动物体跟踪和三维轨迹

Fig. 3　Visualization of Object tracking and trajectory estimation of VDO-SLAM[49]

可以看出, 融合语义信息后, VSLAM的地图表达形式更加丰富. 构建的物体级地图包含场景的高层次

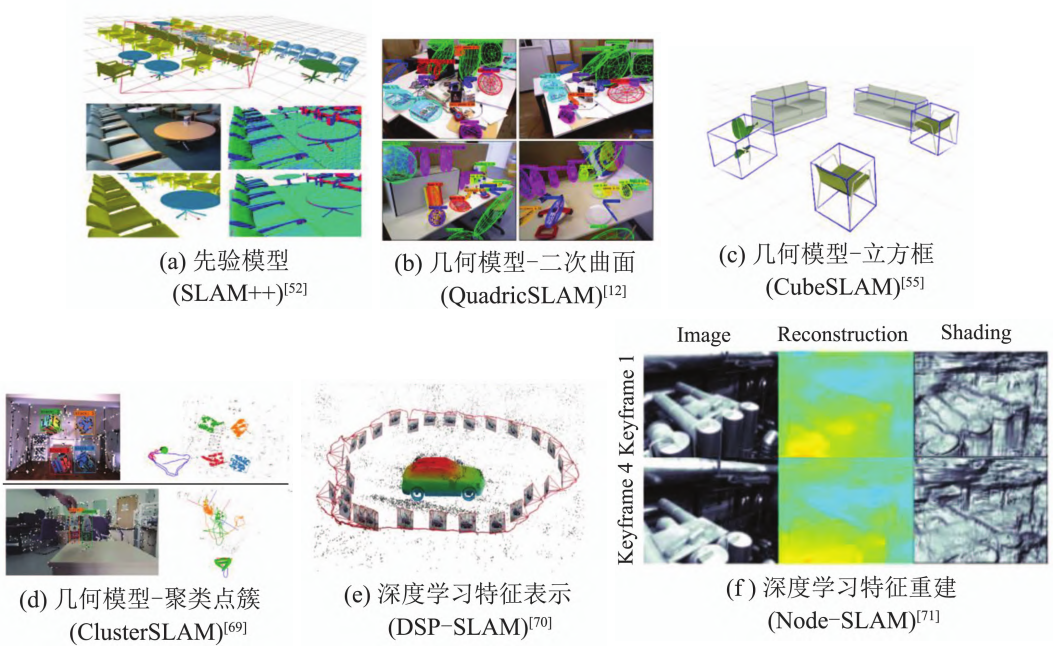
语义信息, 而且通过动态跟踪和联合位姿估计, 可以获得动态物体的速度和运动轨迹估计, 使得VSLAM可以实时估计环境物体的运动轨迹, 具有更丰富的环境感知能力.

4 物体语义的表达形式和初始化方法

物体表达形式是物体级语义SLAM进行环境感知的重要环节, 传统的SLAM算法使用几何特征, 例如点、线、面等元素构建环境地图. 这些几何特征能为SLAM提供定位约束, 并在一定程度上表征场景的感知信息, 但缺少语义信息.

SIFT[51], SURF[50]和ORB[50]是最常用的特征. 利用稀疏点表达环境的视觉SLAM方法[23,33]已经在三维场景重建领域取得了巨大的成功. 然而, 这类地图由三维空间中稀疏分布的点集构成, 缺乏对物体位姿和边界的准确描述. 因此, 稀疏点云地图不能应用于复杂的任务, 如路径规划、避障等.

近年来, 得益于深度学习检测技术的发展, SLAM的地图构建已经由传统的几何表征转为语义描述, 特别是物体级的描述. 在物体表达上, 可以分为: 先验模型、几何模型、深度学习表征等. 这些物体表达提升了SLAM的语义感知能力, 不同物体的表达如图4所示[12,52,55,69–71].



(a) 先验模型 (SLAM++)[52] (b) 几何模型–二次曲面 (QuadricSLAM)[12] (c) 几何模型–立方框 (CubeSLAM)[55]

(d) 几何模型–聚类点簇 (ClusterSLAM)[69] (e) 深度学习特征表示 (DSP–SLAM)[70] (f) 深度学习特征重建 (Node–SLAM)[71]

图 4 物体语义的表达形式

Fig. 4 Object representation method of object VSLAM

4.1 先验模型表达

先验模型表达使用预先建立的先验数据库, 通过检测–匹配的方式加载物体. 如图4(a)所示, 先验模型表达的代表为SLAM++[52]. 文献[72]提出使用检测立方框与先验CAD模型进行ICP匹配, 通过物体路标约束, 实现缺乏纹理的地下停车场定位. 文献[73]使用预先集成或预定义的模型来进行对象跟踪, 该工作的目标是建立一个具有物体标识的环境地图, 并使用预集成的对象模型辅助定位, 其结合了两种不同的深度网络输出结果来联合物体检测和对象的姿态估计.

4.2 几何模型表达

几何模型通过参数化的二次曲面或者立方框实现, 如图4(b)–4(c)所示. Nicholson等人[12]提出了Quadric-SLAM, 首次将二次曲面作为路标引入到SLAM 中, 详细推导了如何利用多帧不同视角的目标检测观测数据构建约束, 求解物体的二次曲面参数. 并提出二次曲面投影观测模型, 使得二次曲面参与位姿优化成为可能. 后续的大多数基于二次曲面的SLAM方案都是基于这个思路的延续[74]. Hosseinzadeh等人[75]提出了Structure Aware SLAM, 在二次曲面路标的基础上加入了平面约束, 使得二次曲面的建模精度进一步提高. Ok等人[14]使用室外物体前向运动假设, 提出了一种利用目标检测框、图像纹理以及语义尺度先验估计二次曲面参数的方法, 降低了二次曲面初始化的难度, 然而, 该方法只能对车辆进行建模.

Liao等人[76]引入对称性假设, 提出了物体感知S-LAM, 利用物体对称性补全物体点云, 进而根据物体点云拟合二次曲面. Chen等人[77]针对物体前向平移运动假设, 提出了一种基于物体凸包和目标检测的二次曲面初始化方法, 为二次曲面初始化提供了新的思路. 为了解决二次曲面初始化对噪声敏感的问题, Ti-

an等人[19]提出了一种参数分离的二次曲面初始化方法, 将旋转和平移估计解耦估计, 提升了初始化对检测框噪声的鲁棒性. 利用物体对称性可以实现快速二次曲面初始化, Liao等人[78]提出的SO-SLAM是一种新颖的单目物体级语义SLAM, 该方法使用三种具有代表性的空间约束, 包括比例比例约束、对称纹理约束和平面支撑约束实现单帧视角下的二次曲面初始化.

立方框表达的代表作是CubeSLAM[55], 将物体模型参数化为三维立方框. EAO-SLAM[21] 使用立方框和椭球对室内物体进行空间描述. 然而, 相比于立方框, 二次曲面具有完备的数学模型表达和射影几何描述, 更易于通过二次曲面重投影约束融合到SLAM的后端优化框架中, 因此受到研究者的青睐. 另外, 一些物体模型表达方案采用物体聚类点描述, Cluster-SLAM[69]及后续的ClusterVO[47]均使用物体聚类点簇进行物体位姿估计和表达, 如图4(d)所示.

4.3　深度学习表征

粗略的几何模型往往不能表示物体的精确体积, 而稠密点云需要大量的内存占用来存储地图. 最近一些工作使用基于深度学习的特征进行模型表达, 结合学习表征的物体级路标实现室外定位[79].

DSP-SLAM[70]使用DeepSDF(signed distance function)网络[80]提取物体特征, 并通过网络参数和表面重建损失函数进行物体表面恢复, 构建场景的物体地图, 如图4(e)所示. SceneCode[81]和Node-SLAM[71]则使用了深度网络中间层特征来表征物体. 利用这些深度提取的特征和表面渲染误差函数, 可以恢复物体的几何形状, 如图4(f)所示.

以上可知, 物体的初始化表征方法决定了物体SLAM的地图表达形式, 深度学习需要高算力的计算设备, 且系统的实时性无法保证, 几何模型可以准确描述物体的大小、旋转和位置, 能完整表达物体的占据信息, 且地图占用小, 已经成为当前研究的热点.

5　物体级语义信息的数据关联方法

基于深度学习的语义提取方法大多关注于单帧检测, 而VSLAM在定位和建图环节均需要考虑时间和空间上的数据关联. 针对物体级语义SLAM, 解决不同帧之间的语义观测关联问题, 确定同一语义对象在连续帧的关联性, 是后续实现多帧优化的前提条件. 当前数据关联方法可以分为两类: 基于概率关联的方法和基于分配算法的关联方法.

5.1　基于概率关联方法

该方法将属于物体的观测约束建模为概率分布模型, 根据模型分布关系来确定帧间物体关联. Beipeng Mu等人[82]使用实例分割掩码的中心深度表征物体观测, 并利用Dirichlet分布对观测进行建模, 通过DP means算法和最大似然估计(maximum likelihood estimation, MLE)迭代结果确定物体的数据关联. Bowman等人[83]使用期望最大化(expectation-maximization, EM)算法对物体路标进行软关联, 并将物体路标作为约束因子与几何观测进行融合. 文献[84]使用概率数据关联的方式解决动态环境下的物体关联. Iqbal和Gans等人[85]分析了不同物体点云深度分布之间的区别, 使用层次密度聚类算法和非参数检验方法对物体进行关联.

5.2　基于分配算法的关联方法

基于分配算法的关联方法能利用多帧观测解决帧间漏检等问题, 为系统提供稳定的物体关联结果. 文献[86]使用物体词袋方法构建成本矩阵, 通过分配算法实现关联. OA-SLAM[11]使用目标检测结果和物体路标重投影的交并比(intersection over union, IoU)构建成本矩阵, 并使用KM(kuhn-munkres,KM)算法进行分配. 然而, 由于有限的观测视角以及观测帧数, 上述方法对于动态场景下的物体数据关联表现并不理想. 为了解决上述问题, 一些工作采用检测跟踪算法(track-by-detection)实现物体数据关联.

Bewley等人[87]使用卡尔曼滤波器对检测框进行状态预测和更新, 通过计算预测和检测结果的2D IoU来度量匹配相似度, 并使用匈牙利算法求解指派问题. 针对单源相似度的局限性, Deep SORT (deep simple online and realtime tracking)[88]融入了外观信息, 使用重识别网络提取的特征, 增强了匹配性能, 同时, 其在匹配策略上增加了级联匹配模块, 根据轨迹相似性进行关联, 降低了遮挡目标ID切换的频率. Hosseinzadeh等人[89]采用检测框内特征点投影匹配数量作为度量, 该方法能够在一定程度上克服跟踪时的遮挡问题.

可以看到, 当前数据关联方法主要通过融合多源特征构建成本矩阵, 并通过分配算法求解实现. 然而, 数据关联结果依赖于语义提取模块精度, 当检测精度降低时会对关联结果产生影响, 进而影系统的定位精度和鲁棒性. 稳定可靠的数据关联方法是提升系统表现的有效途径之一.

6　融合物体级语义信息的后端优化方法

在物体完成初始化后, 需要利用后续观测信息对地图中的重建物体进行优化, 根据物体是否参与相机位姿优化, 后端优化策略可以分为独立优化策略和联合优化策略. 根据是否需要跟踪场景中的动态物体, 联合优化策略的因子图也有不同的形式. 后端优化策略示意图如图5所示.

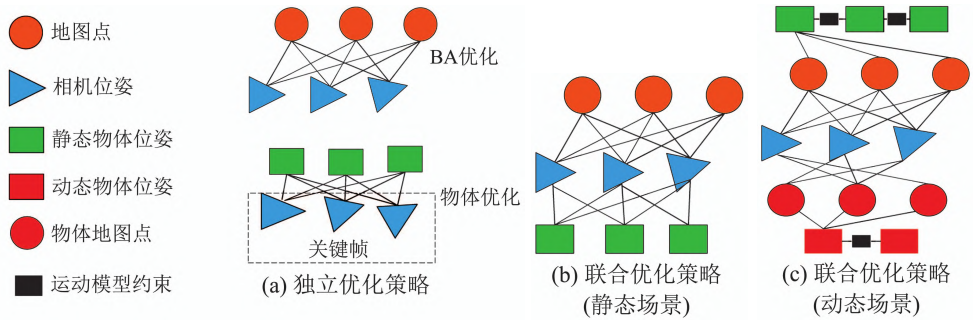


图 5 融合物体信息的语义VSLAM 后端优化方法

Fig. 5 Back-end of object VSLAM with object observations

6.1 独立优化策略

如图5(a)所示, 独立优化策略下, 物体的位姿和模型参数单独进行优化, 物体模型利用跟踪线程中提供的相机初始位姿进行优化. OA-SLAM[11]使用连续帧的目标检测结果对椭球参数单独优化, 并在后端优化中使用优化后的物体路标对相机位姿进行优化. CubeSLAM[55]使用采样得分初始化立方框, 并独立估计相机位姿和立方框参数, 从而确保相机位姿估计的准确性.

独立优化关注于物体重建, 在进行物体位姿优化调整时无法对相机定位结果进行修正, 当相机定位失败时, 系统无法实现准确的自身定位和语义地图构建, 没有充分利用语义信息辅助定位.

6.2 联合优化策略

1) 联合因子图, 该方案将物体参数和位姿估计放在统一因子图中进行优化, 并根据是否需要对动态物体进行位姿估计分别采用不同的因子图.

静态场景的联合优化因子图如图5(b)所示, 该方法通常适用于静态场景或采用动态特征剔除策略的SLAM算法. QuadricSLAM[12]将二次曲面参数和相机位姿优化优化放在联合优化中, 构建了室内场景的语义地图. Tian等[19]提出的方法将初始化椭球和关键帧位姿放在统一优化因子图中进行优化, 提升了室外场景下的定位精度和二次曲面建图准确性.

动态场景的因子图如图 5(c) 所示, 引入了动态物体位姿估计和模型参数优化的误差因子. VDO- SLAM[49]使用物体语义信息和基于场景光流的特征关联, 实现刚性物体位姿估计, 将动态和静态结构放在统一的后端优化框架中. 后续研究如[14–15, 17, 20]也将物体位姿优化放在局部建图线程中以实现联合优化. 近年来, 融合二次曲面路标观测的VSLAM成为了研究的热点[12, 19, 75, 78].

2) 滑动窗口优化策略. 相比于静态场景, 动态场景下的物体观测容易受到漏检、遮挡等因素的干扰, 基于关键帧的关联方案不能为动态物体提供准确的数据关联信息. 为了克服这些问题, 一些基于滑窗的优化方式被提出[43, 47–48]. 滑动窗口由固定帧数的观测队列组成, 当新的帧观测加入队列时, 位于时序最早的帧观测被移出, 同时, 其维护的状态也通过滑窗边缘化的方式进行求解, 如图6所示.

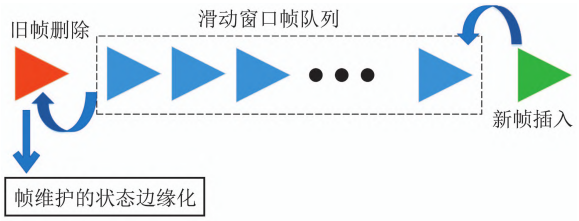


图 6 滑动窗口优化结构示意图

Fig. 6 Sliding window based optimization method

滑动窗口优化将物体位姿和相机位姿放在统一优化框架中, 由于运动物体的特点, 使用滑窗优化可以有效利用连续帧的特征信息. DynaSLAM2[43]将场景静态结构, 相机位姿以及动态物体运动轨迹维护在一个紧耦合的局部BA进行优化, 通过目标检测的二维检测框构建物体位姿约束, 使用舒尔补加速稀疏矩阵边缘化求解, 解决滑窗优化的计算效率问题. Cluster-VO[47]使用点聚类的方式, 将物体点和背景点放在滑窗内进行优化. 该方法使用时间和空间双通道的关键帧管理策略保证计算效率, 同时对遮挡的运动物体进行预测和跟踪.

可以看出, 滑窗的方式具有快速响应、参数优化更准确的特点, 适用于动态物体的跟踪和位姿估计. 基于因子图的联合优化可以有效利用关键帧信息, 对室内场景的物体优化更准确.

7 未来发展和展望

利用语义信息, SLAM可以适应动态和复杂环境下的定位, 并通过物体级语义地图提升系统的环境感知能力. 其技术可以应用于无人驾驶、机器人导航、智慧城市等领域. 未来, 融合语义信息的高层次信息可以为增强现实(AR)和虚拟现实(VR)提供更丰富的

交互体验, 或与其他技术如自然语言处理, 深度学习大模型等结合, 实现更丰富和智能的应用.

尽管当前物体级语义SLAM已经取得了很多优秀的成果, 但仍存在着一些问题需要解决:

1) 物体数据关联方法有待改进.

当前大多数的语义SLAM系统的数据关联都基于静态环境假设, 使用检测IoU构建分配问题, 但是对于运动的物体, 这种策略容易造成错误关联, 导致错误的物体初始化结果, 进而影响定位. 为了解决上述问题, 针对动态场景一般使用检测跟踪(Track-by-Detection)的思路, 即基于物体运动模型的卡尔曼滤波器跟踪检测框, 并通过分配算法实现关联, 如SORT[90]. 一些工作引入外观相似性作为关联误差, 文献[49]引入了光流构建约束, 文献[53]引入词袋构建关联. 这些方法在一定程度上解决了单纯依赖IoU关联的局限性.

近年来, 一些基于深度学习的关联方法被提出, 得益于Transformer等网络的快速发展, 一些方法将其引入到物体数据关联的特征提取中, 以提升多物体跟踪的稳定性和鲁棒性.

DeepSORT[91], Tracktor[92], TransTrack[93]以及TrackFormer[94]引入Transformer的特征外观相似性用于物体数据关联和跟踪. MOTR[95] 提出Transformer的端到端的多物体实时跟踪网络, 其不依赖于IoU匹配. TransMOT[96]使用时序图网络(spatial-temporal-graph transformer) 构建关联.

如何将具有跟踪能力的检测网络引入到语义V-SLAM中, 以此来改善物体关联准确性和鲁棒性, 进而提升语义SLAM建模和定位精度是当前需要考虑的问题.

2) 物体的优化框架不完善.

利用语义信息提升自身定位精度已经成为了当前研究的热点, 当前的语义信息利用方案主要分为紧耦合和松耦合的方案, 两种方案都有适应的场景, 一些工作如文献[19–20], 也表明紧耦合的方案不一定会显著提升定位精度, 相反还会对精度产生影响, 因此统一的框架还没有形成. 在设计损失函数时, 如何减少观测误差的影响, 对误差进行量化和建模是当前研究的方向. 文献[97]提出建模误差, 对数据关联结果的误差进行建模, 从而调整优化时的权重, 提升系统的定位精度.

3) 语义地图的维护和后续的应用.

当前研究主要关注于如何构建具备语义信息的地图, 并在建图的同时提升定位, 但是如何利用已经构建的语义地图实现定位仍是需要解决的问题. 文献[98]提出使用视觉传感器在具有语义标注的激光点云地图上实现准确定位的方案, 然而, 其先验地图构建需要依赖于激光雷达等高精建图设备. 如何利用低成本的视觉传感器构建可用于定位的语义地图是目前亟需解决的问题.

OA-SLAM[11]提出利用物体级语义地图实现相机重定位, 但是该方案主要适用于室内场景. 文献[99]通过语义和视觉编码的联合相似度来衡量查询图像与物体路标的相似度来实现室内场景下的鲁棒定位, 然而, 该算法在室外场景下的关联效果并不理想. 可见, 在室外使用语义地图进行定位仍是挑战[100–102]. 由于高精语义地图的制作成本高昂, 需要大量的标注和数据处理, 如何平衡计算资源和成本也是实际部署和应用中需要考虑的问题.

4) 融合多源信息的场景理解

单一的视觉感知具有一定的局限性, 容易受到雨雾、光照以及曝光的影响. 近年来, 融合多源信息的多模态算法, 如BEVFusion[103]等,成为自动驾驶领域的热点. 这类方法将各种传感器进行多层次、多空间的信息互补和优化组合处理, 弥补视觉传感器缺点. BEV Fusion将激光雷达和视觉图像分别编码为BEV空间特征, 并通过网络融合特征提升感知精度. AutoAlignV2[104]提出稀疏的采样点进行跨模态建模, 通过深度特征提取以及深度网络提升聚合速度. DETR-3D[105]采用基于连续深度预测和3D点云的方法, 将物体查询通过Transformer提取特征并投影在图像上获得3D检测框. 可见, BEV空间下的多模态融合充分利用了各类传感器的信息, 进一步提升了BEV空间下的物体感知精度. 如何有效引入多源信息进行场景物体感知是当前的热点方向.

5) 基于三维重建的场景理解

当前以神经辐射场(neural radiance field,NeRF)[106]为代表的三维重建方法已经成为研究的热点, 相比于传统基于“前端–后端”模式的SLAM, 通过引入三维重建网络的SLAM 可以实现完整统一的端到端框架, 其重建地图更加稠密且能够实现像素级别的误差优化.

近年来, 出现了一些基于NeRF的SLAM算法, 例如Nice-SLAM[107], NERF-SLAM[108], Vox-Fusion[109]等, 展示了其应用在视觉定位和三维场景理解的巨大潜力. 然而, 当前基于NeRF 的方法具有计算开销大, 优化时间长, 难以保证系统实时性的问题. 如何解决算法耗时长和训练复杂, 提出轻量级的三维重建方法是当前亟需解决的问题.

8 结束语

语义信息的融合提升了VSLAM的定位精度和鲁棒性, 同时丰富了VSLAM的环境感知能力. 本文讨论

了VSLAM领域的热点问题: 物体数据关联问题、物体模型表示、物体初始化方法以及后端优化方法. 详细对比和分析了物体级语义VSLAM的代表性研究成果, 并对其关键问题进行了归纳, 为物体级VSLAM的研究提供有益的参考.

参考文献:

[1] REDMON J, FARHADI A. YOLO9000: Better, faster, stronger. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2017: 7263 – 7271.

[2] REDMON J, FARHADI A. Yolov3: An incremental improvement. *ArXiv Preprint*, 2018, arXiv:1804. 02767.

[3] BOCHKOVSKIY A, WANG C Y, LIAO H Y M. Yolov4: Optimal speed and accuracy of object detection. *ArXiv Preprint*, 2020, arXiv: 2004. 10934.

[4] GIRSHICK R. Fast r-cnn. *Proceedings of the IEEE International Conference on Computer Vision*. 2015: 1440 – 1448.

[5] REN S, HE K, GIRSHICK R, et al. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in Neural Information Processing Systems*, 2015: 28.

[6] RONNEBERGER O, FISCHER P, BROX T. U-net: Convolutional networks for biomedical image segmentation. *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015: The 18th International Conference.* Munich, Germany: Springer International Publishing, 2015: 234 – 241.

[7] KENDALL A, BADRINARAYANAN V, CIPOLLA R. Bayesian segnet: Model uncertainty in deep convolutional encoder-decoder architectures for scene understanding. *ArXiv Preprint*, 2015, arXiv: 1511. 02680.

[8] BADRINARAYANAN V, HANDA A, CIPOLLA R. Segnet: A deep convolutional encoder-decoder architecture for robust semantic pixel-wise labelling. *ArXiv preprint*, arXiv:1505. 07293, 2015.

[9] HE K, GKIOXARI G, DOLLAR P, et al. Mask r-cnn*Proceedings of the IEEE International Conference on Computer Vision*. 2017: 2961 – 2969.

[10] BOLYA D, ZHOU C, XIAO F, et al. Yolact: Real-time instance segmentation. *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2019: 9157 – 9166.

[11] ZINS M, SIMON G, BERGER M O. OA-SLAM: Leveraging objects for camera relocalization in visual SLAM. *International Symposium on Mixed and Augmented Reality (ISMAR)*. Barcelona, Spain, Munich, Germany: IEEE, 2022: 720 – 728.

[12] NICHOLSON L, MILFORD M, SüNDERHAUF N. QuadricSLAM: Dual quadrics from object detections as landmarks in object-oriented SLAM. *IEEE Robotics and Automation Letters*, 2018, 4(1): 1 – 8.

[13] HOSSEINZADEH M, LATIF Y, PHAM T, et al. Structure aware SLAM using quadrics and planes. *Computer Vision – ACCV 2018: 14th Asian Conference on Computer Vision, Perth, Australia*. Springer International Publishing, 2019: 410 – 426.

[14] OK K, LIU K, FREY K, et al. Robust object-based SLAM for high-speed autonomous navigation*2019 International Conference on Robotics and Automation (ICRA)*. Montreal, QC, Canada: IEEE, 2019: 669 – 675.

[15] LIAO Z, WANG W, QI X, et al. Object-oriented SLAM using quadrics and symmetry properties for indoor environments. ArXiv Preprint, 2020, arXiv:2004. 05303.

[16] LIAO Z, WANG W, QI X, et al. RGB-D object SLAM using quadrics for indoor environments. *Sensors*, 2020, 20(18): 5150.

[17] CHEN S, SONG S, ZHAO J, et al. Robust dual quadric initialization for forward-translating camera movements. *IEEE Robotics and Automation Letters*, 2021, 6(3): 4712 – 4719.

[18] ZHEN W, YU H, HU Y, et al. Unified representation of geometric primitives for graph-slam optimization using decomposed quadrics. *The International Conference on Robotics and Automation (ICRA)*. Philadelphia, PA, USA: IEEE, 2022: 5636 – 5642.

[19] TIAN R, ZHANG Y, FENG Y, et al. Accurate and robust object SLAM with 3D quadric landmark reconstruction in outdoors. *IEEE Robotics and Automation Letters*, 2021, 7(2): 1534 – 1541.

[20] CAO Z Z, ZHANG Y, TIAN R, et al. Object-aware SLAM based on efficient quadric initialization and joint data association. *IEEE Robotics and Automation Letters*, 2022, 7(4): 9802 – 9809.

[21] WU Y, ZHANG Y, ZHU D, et al. EAO-SLAM: Monocular semi-dense object SLAM based on ensemble data association. *International Conference on Intelligent Robots and Systems (IROS)*. Las Vegas, NV, USA: IEEE, 2020: 4966 – 4973.

[22] BALLESTER I, FONTáN A, CIVERA J, et al. DOT: Dynamic object tracking for visual SLAM. *International Conference on Robotics and Automation (ICRA)*. Xi'an, China: IEEE, 2021: 11705 – 11711.

[23] MUR-ARTAL R, TARDOS J. Orb-slam2: An open-source slam system for monocular, stereo, and rgbd cameras. *IEEE Transactions on Robotics*, 2017, 33(5): 1255 – 1262.

[24] FORSTER C, PIZZOLI M, SCARAMUZZA D. Svo: Fast semi-direct monocular visual odometry. *International Conference on Robotics and Automation (ICRA)*. Hong Kong, China: IEEE, 2014: 15 – 22.

[25] WOLF D, PRANKL J, VINCZE M. Fast semantic segmentation of 3D point clouds using a dense CRF with learned parameters. *International Conference on Robotics and Automation (ICRA)*. IEEE, 2015: 4867 – 4873.

[26] RIEMENSCHNEIDER H, BODIS-SZOMORúA, WEISSENBERG J, et al. Learning where to classify in multi-view semantic segmentation. *Computer Vision – ECCV 2014: The 13th European Conference, Zurich, Switzerland*. Springer International Publishing, 2014: 516 – 532.

[27] IZADI S, KIM D, HILLIGES O, et al. Kinectfusion: Real-time 3D reconstruction and interaction using a moving depth camera. *Proceedings of the 24th Annual ACM Symposium on User Interface Software and Technology*. 2011: 559 – 568.

[28] LI D, SHI X, LONG Q, et al. Dxslam: A robust and efficient visual SLAM system with deep features. *International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2020: 4958 – 4965.

[29] ZHOU W, LIU C, LEI J, et al. Hfnet: Hierarchical feedback network with multilevel atrous spatial pyramid pooling for RGB-D saliency detection. *Neurocomputing*, 2022, 490: 347 – 357.

[30] DETONE D, MALISIEWICZ T, RABINOVICH A. Superpoint: Self-supervised interest point detection and description. *The Conference on Computer Vision and Pattern Recognition Workshops (CVPRW),* Salt Lake City, UT, USA, 2018: 337 – 33712.

[31] YU C, LIU Z, LIU X J, et al. DS-SLAM: A semantic visual SLAM towards dynamic environments. *International Conference on Intelligent Robots and Systems (IROS)*. Madrid, Spain: IEEE, 2018: 1168 – 1174.

[32] BRASCH N, BOZIC A, LALLEMAND J, et al. Semantic monocular SLAM for highly dynamic environments. *International Conference on Intelligent Robots and Systems (IROS)*. Madrid, Spain: IEEE, 2018: 393 – 400.

[33] CAMPOS C, ELVIRA R, RODRIGUEZ J J G, et al. Orb-slam3: An accurate open-source library for visual, visual – inertial, and multimap SLAM. *IEEE Transactions on Robotics*, 2021, 37(6): 1874 – 1890.

[34] LI S, LEE D. RGB-D SLAM in dynamic environments using static point weighting. *IEEE Robotics and Automation Letters*, 2017, 2(4): 2263 – 2270.

[35] ENGEL J, SCHPS T, CREMERS D. LSD-SLAM: Large-scale direct monocular SLAM. *Computer Vision – ECCV 2014: The 13th European Conference, Zurich, Switzerland*. Springer International Publishing, 2014: 834 – 849.

[36] AN L, ZHANG X, GAO H, et al. Semantic segmentation – aided visual odometry for urban autonomous driving. *International Journal of Advanced Robotic Systems*, 2017, 14(5): 1729881417735667.

[37] RUNZ M, BUFFIER M, AGAPITO L. Maskfusion: Real-time recognition, tracking and reconstruction of multiple moving objects. IEEE International Symposium on Mixed and Augmented Reality (ISMAR). IEEE, 2018: 10 – 20.

[38] BEI B, VALADA A. Dynamic object removal and spatio-temporal RGB-D inpainting via geometry-aware adversarial learning. *IEEE Transactions on Intelligent Vehicles*, 2022, 7(2): 170 – 185.

[39] YU C, LIU Z, LIU X J, et al. DS-SLAM: A semantic visual SLAM towards dynamic environments. *International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2018: 1168 – 1174.

[40] BESCOS B, FACIL J M, CIVERA J, et al. DynaSLAM: Tracking, mapping, and inpainting in dynamic scenes. *IEEE Robotics and Automation Letters*, 2018, 3(4): 4076 – 4083.

[41] ZHONG F, WANG S, ZHANG Z, et al. Detect-SLAM: Making object detection and SLAM mutually beneficial. *Winter Conference on Applications of Computer Vision (WACV)*. Lake Tahoe, NV, USA: IEEE, 2018: 1001 – 1010.

[42] KAVETI P, SINGH H. A Light Field Front-end for Robust SLAM in Dynamic Environments. *ArXiv Preprint ,* 2020, arXiv: 2012. 10714.

[43] BESCOS B, CAMPOS C, TARDOS J D, et al. DynaSLAM II: Tightly-coupled multi-object tracking and SLAM. *IEEE Robotics and Automation Letters*, 2021, 6(3): 5191 – 5198.

[44] BALLESTER I, FONTAN A, CIVERA J, et al. DOT: Dynamic object tracking for visual SLAM. *International Conference on Robotics and Automation (ICRA)*. IEEE, 2021: 11705 – 11711.

[45] WANG C C, THORPE C, THRUN S, et al. Simultaneous localization, mapping and moving object tracking. *The International Journal of Robotics Research*, 2007, 26(9): 889 – 916.

[46] KUNDU A, KRISHNA K M, JAWAHAR C V. Realtime multibody visual SLAM with a smoothly moving monocular camera*International Conference on Computer Vision*. Barcelona, Spain: IEEE, 2011: 2080 – 2087.

[47] HUANG J, YANG S, MU T J, et al. Clustervo: Clustering moving instances and estimating visual odometry for self and surroundings. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020: 2168 – 2177.

[48] GONZALEZ M, MARCHAND E, KACETE A, et al. Twistslam: Constrained slam in dynamic environment. *IEEE Robotics and Automation Letters*, 2022, 7(3): 6846 – 6853.

[49] ZHANG J, HENEIN M, MAHONY R, et al. VDO-SLAM: A visual dynamic object-aware SLAM system. *ArXiv Preprint, 2020, arXiv*:2005. 11052.

[50] RUBLEE E, RABAUD V, KONOLIGE K, et al. ORB: An efficient alternative to SIFT or SURF. *International Conference on Computer Vision*. Barcelona, Spain: IEEE, 2011: 2564 – 2571.

[51] LOWE G. SIFT-The Scale Invariant Feature Transform. Int. J, 2004, 2(91-110): 2.

[52] SALAS-MORENO R F, NEWCOMBE R A, STRASDAT H, et al. SLAM++: Simultaneous localisation and mapping at the level of objects. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2013: 1352 – 1359.

[53] LIU Y, PETILLOT Y, LANE D, et al. Global localization with object-level semantics and topology. *International Conference on Robotics and Automation (ICRA)*. Montreal, QC, Canada, Montreal, QC, Canada: IEEE, 2019: 4909 – 4915.

[54] SUCAR E, WADA K, DAVISON A. NodeSLAM: Neural object descriptors for multi-view shape reconstruction. *International Conference on 3D Vision (3DV)*. Fukuoka, Japan: IEEE, 2020: 949 – 958.

[55] YANG S, SCHERER S. Cubeslam: Monocular 3-D object SLAM. *IEEE Transactions on Robotics*, 2019, 35(4): 925 – 938.

[56] SONG S, ZHAO J, FENG T, et al. Scale Estimation with Dual Quadrics for Monocular Object SLAM. *International Conference on Intelligent Robots and Systems (IROS)*. Kyoto, Japan: IEEE, 2022: 1374 – 1381.

[57] SUCAR E, HAYET J B. Bayesian scale estimation for monocular SLAM based on generic object detection for correcting scale drift. *IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2018: 5152 – 5158.

[58] SHARMA A, DONG W, KAESS M. Compositional and scalable object SLAM. *IEEE International Conference on Robotics and Automation (ICRA)*. Xi'an, China, IEEE, 2021: 11626 – 11632.

[59] TAGUCHI Y, JIAN Y D, RAMALINGAM S, et al. Point-plane SLAM for hand-held 3D sensors. *IEEE International Conference on Robotics and Automation*. IEEE, 2013: 5182 – 5189.

[60] KIM P, COLTIN B, KIM H J. Linear RGB-D SLAM for planar environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 2022,44(11): 8403 – 8419.

[61] YANG S, SONG Y, KAESS M, et al. Pop-up SLAM: Semantic monocular plane SLAM for low-texture environments. *International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2016: 1222 – 1229.

[62] KAESS M. Simultaneous localization and mapping with infinite planes. *IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2015: 4605 – 4611.

[63] LI B, ZOU D, SARTORI D, et al. TextSLAM: Visual SLAM with planar text features. *International Conference on Robotics and Automation (ICRA)*. IEEE, 2020: 2102 – 2108.

[64] MCCORMAC J, CLARK R, BLOESCH M, et al. Fusion++: Volumetric object-level SLAM. *International Conference on 3D Vision (3DV)*. IEEE, 2018: 32 – 41.

[65] ZHANG X, WANG W, QI X, et al. Stereo plane SLAM based on intersecting lines. *International Conference on Intelligent Robots and Systems (IROS)*. Prague, Czech Republic: IEEE, 2021: 6566 – 6572.

[66] HAN X, YANG L. SQ-SLAM: Monocular Semantic SLAM Based on Superquadric Object Representation. *ArXiv Preprint, 2022, arXiv*:2209. 10817.

[67] QIN T, CHEN T, CHEN Y, et al. AVP-SLAM: Semantic visual mapping and localization for autonomous vehicles in the parking lot. *International Conference on Intelligent Robots and Systems (IROS)*. Las Vegas, NV, USA: IEEE, 2020: 5939 – 5945.

[68] REDDY N D, SINGHAL P, CHARI V, et al. Dynamic body VSLAM with semantic constraints. *International Conference on Intelligent Robots and Systems (IROS)*. Hamburg, Germany: IEEE, 2015: 1897 – 1904.

[69] HUANG J, YANG S, ZHAO Z, et al. Clusterslam: A slam backend for simultaneous rigid body clustering and motion estimation. *Proceedings of the IEEE/CVF International Conference on Computer Vision*. Seoul, Korea (South), 2019: 5875 – 5884.

[70] WANG J, RüNZ M, AGAPITO L. DSP-SLAM: Object-oriented SLAM with deep shape priors. *International Conference on 3D Vision (3DV)*. IEEE, 2021: 1362 – 1371.

[71] SUCAR E, WADA K, DAVISON A. NodeSLAM: Neural object descriptors for multi- view shape reconstruction. *International Conference on 3D Vision (3DV)*. Fukuoka, Japan: IEEE, 2020: 949 – 958.

[72] DONG Y, WANG S, YUE J, et al. A novel texture-less object-oriented visual SLAM system. *IEEE Transactions on Intelligent Transportation Systems*, 2019, 22(1): 36 – 49.

[73] FEI X, SOATTO S. Visual-inertial object detection and mapping. *Proceedings of the European Conference on Computer Vision (ECCV)*. 2018: 301 – 317.

[74] HOSSEINZADEH M, LI K, LATIF Y, et al. Real-time monocular object-model aware sparse SLAM. *International Conference on Robotics and Automation (ICRA)*. Montreal, QC, Canada: IEEE, 2019: 7123 – 7129.

[75] HOSSEINZADEH M, LATIF Y, PHAM T, et al. Structure aware SLAM using quadrics and planes. *Computer Vision – ACCV 2018: 14th Asian Conference on Computer Vision. Perth, Australia*. Springer International Publishing, 2019: 410 – 426.

[76] LIAO Z, WANG W, QI X, et al. Object-oriented SLAM using quadrics and symmetry properties for indoor environments. *ArXiv Preprint,* 2020 arXiv:2004. 05303.

[77] CHEN S, SONG S, ZHAO J, et al. Robust dual quadric initialization for forward-translating camera movements. *IEEE Robotics and Automation Letters*, 2021, 6(3): 4712 – 4719.

[78] LIAO Z, HU Y, ZHANG J, et al. SO-SLAM: semantic object SLAM with scale proportional and symmetrical texture constraints. *IEEE Robotics and Automation Letters*, 2022, 7(2): 4008 – 4015.

[79] SHAN M, FENG Q, ATANASOV N. ORCVIO: Object Residual Constrained Visual-Inertial Odometry. *International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2020: 5104 – 5111.

[80] PARK J J, FLORENCE P, STRAUB J, et al. Deepsdf: Learning continuous signed distance functions for shape representation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Long Beach, CA, USA, 2019: 165 – 174.

[81] ZHI S, BLOESCH M, LEUTENEGGER S, et al. Scenecode: Monocular dense semantic reconstruction using learned encoded scene representations. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Long Beach, CA, USA, 2019: 11768 – 11777.

[82] MU B, LIU S Y, PAULL L, et al. Slam with objects using a nonparametric pose graph *International Conference on Intelligent Robots and Systems (IROS).* Korea (South): IEEE, 2016: 4602 – 4609.

[83] BOWMAN S L, ATANASOV N, DANIILIDIS K, et al. Probabilistic data association for semantic slam. *International Conference on Robotics and Automation (ICRA)*. Singapore: IEEE, 2017: 1722 – 1729.

[84] STRECKE M, STUCKLER J. Em-fusion: Dynamic object-level slam with probabilistic data association*Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2019: 5865 – 5874.

[85] IQBAL A, GANS N R. Localization of classified objects in slam using nonparametric statistics and clustering. *International Conference on Intelligent Robots and Systems (IROS)*. Madrid, Spain: IEEE, 2018: 161 – 168.

[86] QIAN Z, PATATH K, FU J, et al. Semantic slam with autonomous object-level data association. *International Conference on Robotics and Automation (ICRA)*. Madrid, Spain: IEEE, 2021: 11203 – 11209.

[87] BEWLEY A, GE Z, OTT L, et al. Simple online and realtime tracking. *International Conference on Image Processing (ICIP)*. IEEE, 2016: 3464 – 3468.

[88] WOJKE N, BEWLEY A. Deep cosine metric learning for person re-identification. *Winter Conference on Applications of Computer Vision (WACV)*. Lake Tahoe, NV, USA: IEEE, 2018: 748 – 756.

[89] HOSSEINZADEH M, LI K, LATIF Y, et al. Real-time monocular object-model aware sparse SLAM. *International Conference on Robotics and Automation (ICRA)*. Montreal, QC, Canada: IEEE, 2019: 7123 – 7129.

[90] BEWLEY A, GE Z, OTT L, et al. Simple online and realtime tracking. *International Conference on Image Processing (ICIP)*. Phoenix, AZ, USA: IEEE, 2016: 3464 – 3468.

[91] WOJKE N, BEWLEY A, PAULUS D. Simple online and realtime tracking with a deep association metric. *IEEE International Conference on Image Processing (ICIP)*. IEEE, 2017: 3645 – 3649.

[92] BERGMANN P, MEINHARDT T, LEAL-TAIXE L. Tracking without bells and whistles. *Proceedings of the IEEE/CVF International Conference on Computer Vision*. Seoul, Korea (South), 2019: 941 – 951.

[93] SUN P, CAO J, JIANG Y, et al. Transtrack: Multiple object tracking with transformer. *ArXiv Preprint*, 2020, arXiv:2012. 15460.

[94] MEINHARDT T, KIRILLOV A, LEAL-TAIXE L, et al. Trackformer: Multi-object tracking with transformers. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022: 8844 – 8854.

[95] ZENG F, DONG B, ZHANG Y, et al. Motr: End-to-end multiple-object tracking with transformer. *European Conference on Computer Vision*. Cham: Springer Nature Switzerland, 2022: 659 – 675.

[96] CHU P, WANG J, YOU Q, et al. Transmot: Spatial-temporal graph transformer for multiple object tracking. *ArXiv Preprint,* 2021 arXiv:2104. 00194.

[97] MERRILL N, GUO Y, ZUO X, et al. Symmetry and uncertainty-aware object SLAM for 6DOF object pose estimation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Orleans, LA, USA, 2022: 14881 – 14890.

[98] LIANG S, ZHANG Y, TIAN R, et al. SemLoc: Accurate and robust visual localization with semantic and structural constraints from prior Maps. *International Conference on Robotics and Automation (ICRA)*. Philadelphia, PA, USA, IEEE, 2022: 4135 – 4141.

[99] BESCOS B, CADENA C, NEIRA J. Empty cities: A dynamic-object-invariant space for visual SLAM. *IEEE Transactions on Robotics*, 2020, 37(2): 433 – 451.

[100] GAWEL A, DEL DON C, SIEGWART R, et al. X-view: Graph-based semantic multi-view localization. *IEEE Robotics and Automation Letters*, 2018, 3(3): 1687 – 1694.

[101] LI J, MEGER D, DUDEK G. Semantic mapping for view-invariant relocalization. *International Conference on Robotics and Automation (ICRA)*. Montreal, QC, Canada: IEEE, 2019: 7108 – 7115.

[102] QIN T, ZHENG Y, CHEN T, et al. A light-weight semantic map for visual localization towards autonomous driving. *International Conference on Robotics and Automation (ICRA)*. Xi'an, China: IEEE, 2021: 11248 – 11254.

[103] LIU Z, TANG H, AMINI A, et al. Bevfusion: Multi-task multi-sensor fusion with unified bird's-eye view representation. *International Conference on Robotics and Automation (ICRA)*. Xi'an, China: IEEE, 2023: 2774 – 2781.

[104] CHEN Z, LI Z, ZHANG S, et al. Autoalignv2: Deformable feature aggregation for dynamic multi-modal 3D object detection. *ArXiv Preprint,2022* arXiv: 2207, 2022.

[105] WANG Y, GUIZILINI V C, ZHANG T, et al. Detr3d: 3d object detection from multi-view images via 3D-to-2D queries. *Conference on Robot Learning*. PMLR, 2022: 180 – 191.

[106] WANG Z, WU S, XIE W, et al. NeRF: Neural radiance fields without known camera parameters. *ArXiv Preprint*, 2021 arXiv:2102. 07064.

[107] ZHU Z, PENG S, LARSSON V, et al. Nice-slam: Neural implicit scalable encoding for slam. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. New Orleans, LA, USA,2022: 12776 – 12786.

[108] ROSINOL A, LEONARD J J, CARLONE L. Nerf-slam: Real-time dense monocular slam with neural radiance fields. *ArXiv Preprint,* 2022 arXiv:2210. 13641.

[109] YANG X, LI H, ZHAI H, et al. Vox-Fusion: Dense tracking and mapping with voxel-based neural implicit representation. *International Symposium on Mixed and Augmented Reality (ISMAR)*. Singapore, Singapore: IEEE, 2022: 499 – 507.

作者简介:

田 瑞 博士研究生, 硕士毕业于东北大学, 研究方向为视觉SLAM、三维重建、语义SLAM, E-mail: tianruineu@qq. com;

张云洲 教授, 博士生导师, 研究方向为智能移动机器人和计算机视觉, E-mail: zhangyunzhou@mail.neu. edu.cn;

杨凌昊 硕士研究生, 研究方向为视觉SLAM, E-mail: yang linghaoneu@163.com;

曹振中 硕士研究生, 研究方向为视觉SLAM, E-mail: 3322 12807@qq.com.