结合隐式建图的视觉 SLAM 技术综述

郭恺悦, 刘越*

(北京理工大学光电学院　北京　100081)
(liuyue@bit.edu.cn)

摘　要: 同步定位与地图构建(simultaneous localization and mapping, SLAM)技术能够在陌生环境中定位自身位置的同时构建周围环境, 已经成为机器人、无人驾驶和虚拟现实等领域非常重要的基础技术. 隐式建图方法对于场景未观测区域具有一定补全预测能力, 可以实现对遮挡或稀疏观测区域的孔填充, 近年来, 将该方法融入 SLAM, 以提高其系统性能逐渐成为 SLAM 领域的研究热点. 文中首先总结应用于视觉 SLAM 中的隐式建图方法并基于地图存储载体对其进行分类; 然后基于建图渲染速度提高、大规模场景扩展方法、建图鲁棒性提高、前端功能的改进和回环检测的补充等改进方向对结合隐式建图的视觉 SLAM 进行分类说明, 并梳理了面向语义建图、动态场景和多传感器融合等特定场景的隐式建图 SLAM 系统; 随后介绍隐式建图 SLAM 系统常用的数据集和评价标准, 并基于相同数据集和评价标准对多个 SLAM 系统进行对比和分析; 最后总结隐式建图视觉 SLAM 系统提高自身性能的改进方式, 剖析系统现存的计算量大和遗忘严重等短板, 并与其他技术对比展望未来发展趋势.

关键词: 同步定位与地图构建; 隐式建图; 体素网格; 多层感知器
中图法分类号: TP391.41　　**DOI:** 10.3724/SP.J.1089.2024-00269

Review on Visual SLAM Combined with Implicit Mapping

Guo Kaiyue and Liu Yue*

(*School of Optoelectronics, Beijing Institute of Technology, Beijing*　100081)

Abstract: Simultaneous localization and mapping (SLAM) technology can locate itself and construct the surrounding environment in an unfamiliar environment. It has become an important basic technology in fields such as robotics, autonomous driving and virtual reality. The implicit mapping methods have a certain ability to complete and predict the unobserved areas of the scene and can realize hole filling in occluded or sparsely observed areas. Recently, it has gradually become a research hotspot that integrating the implicit mapping methods into SLAM system. This paper firstly summarized the implicit mapping methods applied in visual SLAM and classified them based on the map storage carrier. Then, it classified and explained the visual SLAM combined with implicit mapping based on improvement directions such as improving the mapping rendering speed, methods for large-scale scene expansion, enhancing the mapping robustness, improving the front-end functions and supplementing loop closure detection. It also sorted out the implicit mapping SLAM systems for specific scenarios such as semantic mapping, dynamic scenes and multi-sensor fusion. Subsequently, it introduced the commonly used datasets and evaluation criteria of implicit mapping SLAM systems and compared multiple SLAM systems based on the same datasets and evaluation criteria. Finally, it summarized the improvement methods of implicit mapping visual SLAM systems to improve their own performance, analyzed the existing shortcomings

收稿日期: 2024-05-24; 修回日期: 2025-02-10. **基金项目**: 国家科技创新 2030“新一代人工智能”重大项目(2022ZD0115901); 国家自然科学基金(61960206007, 62332003). **郭恺悦**(2001—), 女, 硕士研究生, 主要研究方向是视觉 SLAM、神经辐射场建图、3D 高斯建图; **刘越**(1968—), 男, 博士, 教授, 博士生导师, CCF 会员, 论文通信作者, 主要研究方向为虚拟现实与增强现实、新型三维显示、计算机视觉等.

such as large computational load and serious forgetting and compared with the other technology to look ahead to the future trends.

Key words: simultaneous localization and mapping; implicit mapping; voxel grid; multi-layer perceptron

同步定位与地图构建(simultaneous localization and mapping, SLAM)技术最早应用于机器人领域, 能够让机器人在陌生的环境中进行自我定位的同时构建出周围环境的地图. 随着各类传感器的发展和优化理论的更新, SLAM 技术在智能家居、增强现实、虚拟现实、军用探测和无人航行器等领域同样得到了广泛的应用, 在较为特殊的生产场景(水下、山洞、隧道等)中, 可辅助进行救援、巡航等任务, 发挥了极其重要的作用. 目前, 根据使用的传感器类型, 主流的 SLAM 技术可以分为激光 SLAM 和视觉 SLAM. 其中, 激光 SLAM 使用激光雷达传感器采集周围环境的信息, 发展较为成熟且精度较高, 但在成本较高和应用场景中受到限制; 视觉 SLAM 利用单目相机、双目相机或 RGB-D 相机等采集信息, 具有成本较低、硬件轻量、应用场景丰富等优点, 日益受到国内外研究人员的关注.

视觉 SLAM 系统大多通过处理连续视频帧计算当前视频对应的相机运动轨迹, 并结合相机位姿与对应视频帧实现地图构建, 导致视觉 SLAM 在推断脱离相机运动轨迹的、未观测视角下的场景几何形状和颜色信息有些困难. 近年来, 以神经辐射场(neural radiance fields, NeRF)为代表的隐式建图方法被引入视觉 SLAM 系统中, 以实现未观测区域几何形状的预估, 并完成高质量的新视图渲染. 隐式建图方法的引入可实现对场景的连续表面建模, 以及对遮挡或观测稀疏的区域进行孔填充, 使得 SLAM 系统对于未观测场景具有一定的补全预测能力, 并逐渐成为视觉 SLAM 领域的研究热点.

目前, 针对视觉 SLAM 技术的综述[1-3]对众多视觉 SLAM 系统进行了相对全面的阐述, 但关于结合隐式建图的 SLAM 系统的综述文献仍然较为匮乏. 鉴于此, 本文对视觉 SLAM 中常用的隐式场景表示及其体渲染方法进行总结, 并基于不同的地图存储载体(简称为存储载体)对建图方法进行分类; 重点分析隐式重建方法对 SLAM 系统建图环节的改进, 同时梳理隐式建图 SLAM 系统中的前端改进方法和回环检测补充方法, 详述面向特定场景(语义地图、动态场景和多传感器融合)的应用; 还讨论了常用的评价标准和数据集, 并基于相同数据集和评价标准对比和分析多个隐式建图视觉 SLAM 系统; 最后进行总结, 并展望未来的发展方向.

1 用于视觉 SLAM 技术的隐式建图方法

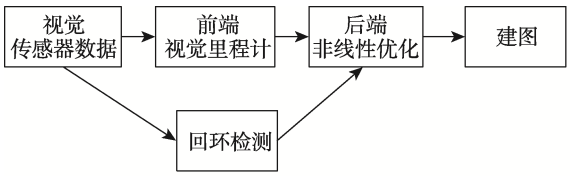
地图构建是 SLAM 系统的重要功能之一, 为众多应用任务提供初始环境信息. SLAM 的建图环节有不同的场景表示方法、渲染方式和存储载体. 本节重点介绍应用在 SLAM 系统中的隐式场景表示和光线投射体渲染方法, 并根据存储载体对建图方法进行分类.

1.1 视觉 SLAM 系统的基本组成和经典系统

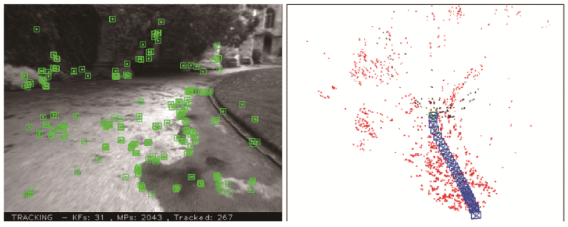
1986 年, Smith 等[4]提出 SLAM 系统的概念. MonoSLAM[5]是较早出现的视觉 SLAM 系统, 其证明了使用单目相机可完成 SLAM 任务. 随着研究的不断深入, 视觉 SLAM 系统已经具有较为成熟的框架, 包括前端、后端、回环检测和建图 4 个组成部分, 如图 1a 所示. 其中, 前端通过处理传感器获取的信息计算图像帧间的相机运动关系, 并完成相机位姿估计, 生成局部地图; 后端优化前端得到的结果, 利用卡尔曼滤波器等多种线性滤波器或者光束平差法(bundle adjustment, BA)[6]等非线性优化方法, 得到最优的相机位姿估计组合和精度更高的地图; 回环检测的作用是判断相机是否回到曾经到达过的位置, 通常利用词袋模型[7-8]等方法进行检测, 若检测到回环, 则对相机轨迹和已构建的地图进行调整并消除累积误差; 建图指系统根据自身定位与环境信息感知的结果构建出需要的地图. 通常, SLAM 系统的跟踪和建图双线程进行[9], 以提高运行效率. 已有大量视觉 SLAM 技术的研究成果, 包括 ORB-SLAM 系列[10-12], DTAM[13]和 LSD-SLAM[14]等经典系统, 随着深度学习的出现和发展, 研究者开始将各类神经网络引入 SLAM 系统[15-17], 以提升系统性能.

1.2 隐式场景表示及其体渲染方法

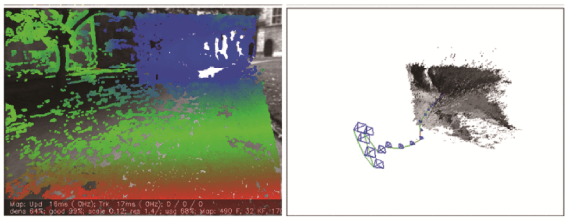
隐式场景表示无法直接反映场景的几何结构, 通常需要结合体渲染技术获取场景中某坐标点相对相机位姿的深度和颜色等信息. 结合隐式建图的 SLAM 系统中, 常用的隐式场景表示包括符号距离函数(signed distance function, SDF)、占用和密度, 通过结合对应的光线投射体渲染方法[18], 能



a. 组成部分



b. ORB-SLAM[10]



c. LSD-SLAM[14]

图 1　SLAM 组成部分和经典系统

够产生质量较高的渲染结果.

三维世界的场景存在连续或近乎连续的表面, 理论上可以用含有 3 个变量的函数表示, 但很难显式地表达某个具体场景表面形状的曲面函数. SDF[19]通过定义空间点和场景表面的空间关系, 隐式地描述连续场景表面, 能够表达具有复杂几何结构的场景. 某空间点的 SDF 值指该空间点和场景表面之间的最小距离, 点在场景表面外部时符号取正, 在表面内部时符号取负. 在场景表面上时取 0, SDF 的绝对值计算公式为

$$\left|f\left(p,\Omega\right)\right|=\min_{x\in\Omega}\left\|x-p\right\|.$$

其中, p 表示采样空间点; Ω 表示场景表面隐函数; x 表示属于场景表面的点的集合. 由于大多数情况下靠近场景表面的空间点是更重要的, 因此通常使用截断 SDF(truncated SDF, TSDF)提高计算速度, 即对于距离场景表面远于截断距离的空间点, 使用截断距离代替其 SDF 值.

占用值指空间中的某位置点被场景表面占用的概率值, 场景中所有空间点的占用值构成占用场. 与 SDF 不同, SDF 值没有取值范围限制, 而某空间位置点占用值的取值范围是[0,1]. 实际系统中通常定义一个阈值, 当占用值高于该阈值时, 认为该空间点在场景表面上; 否则, 认为该空间点不在场景表面上. 2019 年, Mescheder 等[20]提出可适用于视觉 SLAM 的占位网络算法, 其利用神经网络对空间中每个点的二值占位情况进行预测, 即对三维空间训练一个二分类网络, 推动了占用场在视觉 SLAM 中的应用.

密度值表示光线在特定点与场景的体积“介质”相互作用的微分概率, 用于处理光散射和光吸收不均匀物质的重建问题, 如烟雾和一些流体等, 场景中所有空间点的密度值构成密度场. 在实际应用场景中, 光的吸收或散射系数在空间中是相互独立地变化的, 吸收系数越高, 则体积就越不透明. 研究者通常将场景的散射和吸收系数设置为常数值, 使用密度场调制场景在某视图下的渲染图像.

隐式场景表示对应的体渲染方法的核心是光线投射[18]. 从相机中心 $\boldsymbol{o}$ 沿着每个像素的标准化视图方向 $\boldsymbol{r}$ 投射光线穿过重建场景, 每个像素的光线方向 $\boldsymbol{r}$ 互不相同, 沿着每条光线采样 N 个点. 在像素 i 发出的光线上, 采样点 $\boldsymbol{p}_i$ 的位置为 $\boldsymbol{p}_i=\boldsymbol{o}+t_i\boldsymbol{r}$. 其中, t_i 表示 $\boldsymbol{p}_i$ 沿该光线的深度, 通常被限制在一定范围内. 不同 SLAM 系统往往使用不同的方法决定 N 和 t_i .

确定采样点后, 利用其空间位置获取对应的隐式场景表示和颜色值 $\boldsymbol{c}_i$, 该光线对应像素 i 的深度信息 D 和颜色信息 C 由该光线上所有采样点的 t_i 和 $\boldsymbol{c}_i$ 进行累加计算得到, 它们的计算公式为

$$\begin{cases}D=\sum_{i=1}^{N}\omega_i t_i\\ C=\sum_{i=1}^{N}\omega_i\boldsymbol{c}_i\end{cases}.$$

其中, 权重 ω_i 由采样点处的隐式场景表示计算得出, 不同的隐式场景表示需采用不同的权重计算方法. 记点 $\boldsymbol{p}_i$ 处的占用值为 o_i, 则占用场的权重计算公式为 $\omega_i=o_i\prod_{j=1}^{i-1}\left(1-o_j\right)$.

计算密度场时, 首先将 $\boldsymbol{p}_i$ 处的密度值 σ_i 转化为 o_i, 然后用 o_i 计算 ω_i. 转化公式为

$$o_i=1-\exp\left(-\Delta_i\sigma_i\right).$$

其中, $\Delta_i=t_i-t_{i-1}$.

SDF 值 ϕ_i 的权重计算方法较多. Azinović 等[21]提出直接进行计算的方法, 即

$$\omega_i=\mathrm{Sigmoid}\left(\frac{\phi_i}{t}\right)\cdot\mathrm{Sigmoid}\left(-\frac{\phi_i}{t}\right).$$

其中, t 表示 TSDF 的截断距离. VolSDF[22]将 ϕ_i 转化成 σ_i 后, 再利用密度场计算权重. ϕ_i 转化成 σ_i 的

公式为

$$\sigma_i=\begin{cases}\dfrac{\alpha}{2}\exp\left(\dfrac{\phi}{\beta}\right), & \text{if } \phi\leqslant 0\\ \alpha\left(1-\exp\left(\dfrac{\phi}{\beta}\right)\right), & \text{if } \phi>0\end{cases} \tag{1}$$

StyleSDF[23]中提出的转化公式为

$$\sigma_i=\beta\cdot\mathrm{Sigmoid}\left(-\beta\cdot\phi_i\right) \tag{2}$$

式(1)(2)中，α 和 β 都是控制转化的参数. SDF 的权重计算较为复杂，需要根据具体的 SLAM 系统进行选择和调整.

1.3 存储载体

在 SLAM 系统中，隐式场景表示以及位置信息、法向量信息、拓扑信息等显式场景表示都需要存储载体进行信息存储. 存储载体定义为 SLAM 系统中地图几何信息实际存储的载体形式，包括点云、多边形网格、体素网格和多层感知器(multilayer perceptron，MLP)参数等. 存储载体可以存储显式场景表示、隐式场景表示和提取出的特征，其中，特征通常需要经过 MLP 解码器解码出隐式场景表示[24-25]. 随着计算能力的提升，目前，一种存储载体已经能够存储不同形式的几何信息，如点云早期只具有位置信息，但后续能够通过存储可被解码的特征实现更多场景细节的存储[26-27]. NeRF 方法[28]具有结构精简和渲染效果好的优势，其本质是以 MLP 参数为存储载体计算密度场的建图方法. 图 2 所示为依据不同的存储载体及其包含的不同形式的几何信息，对显式建图方法、隐式建图方法和混合式建图方法进行分类.

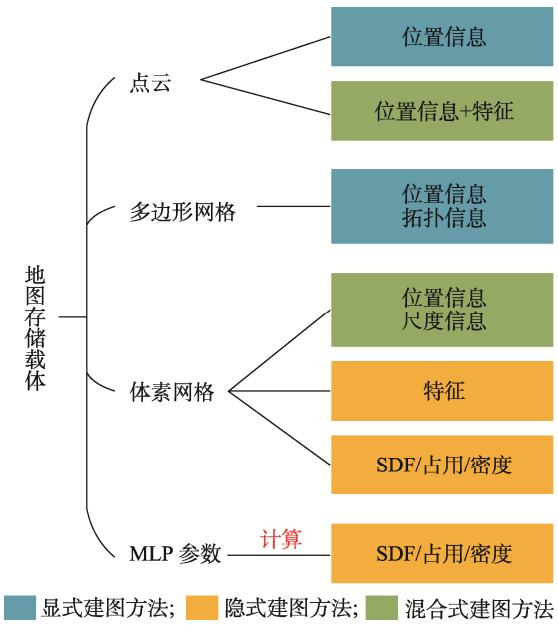


图 2 基于存储载体的建图方法分类

本文介绍的 SLAM 系统中，采用的隐式建图方法大多以体素网格或 MLP 参数为存储载体，以隐式场景表示或场景中提取出来的特征作为几何信息. 如图 3 所示，以 MLP 参数作为存储载体，是指将某采样点的空间位置等信息输入 MLP，计算出隐式场景表示；以体素网格为存储载体，是指在体素网格顶点处直接存储特征或隐式场景表示，并通过插值等方法获取采样点位置的信息.

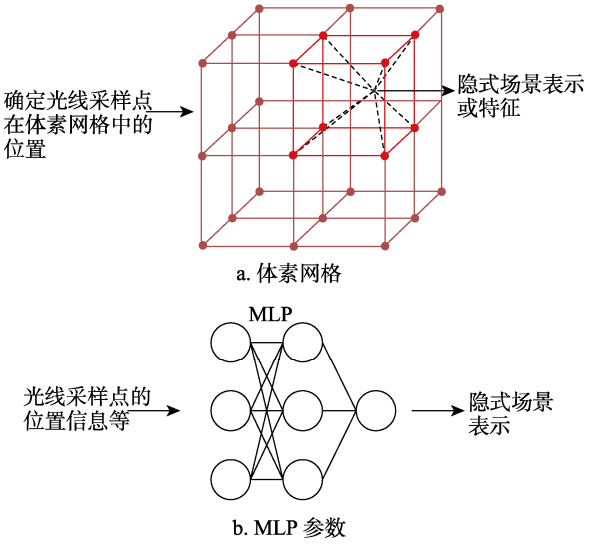


图 3 隐式建图方法常用存储载体

隐式建图方法能够在重建三维场景有部分空洞的情况下渲染出比较逼真的结果，但并非所有的隐式场景表示都能融入 SLAM 系统. 本节主要分析了适用于 SLAM 系统的隐式场景表示及其体渲染方法，针对存储载体对显式建图、隐式建图和混合式建图方法进行划分；同时，对本文介绍的 SLAM 系统的存储载体(体素网格和 MLP 参数)进行具体描述.

2 隐式建图 SLAM 的建图环节改进方法

与显式建图方法相比，将隐式建图方法结合光线投射体渲染通常能达到非常逼真的新视角渲染效果，填补重建时可能出现的孔洞. iMAP[29]较早将隐式建图融入视觉 SLAM 系统，其结合经典方法和深度学习方法，在采用传统 SLAM 关键帧策略的同时，基于 NeRF 方法用单个 MLP 进行场景表示；建立包含深度几何误差和颜色光度误差的损失函数，对 MLP 参数和相机位姿进行联合优化，其中，MLP 参数是在没有任何先验数据的前提下，在新场景内直接进行训练的. 为了保持迭代速度，iMAP 会在关键帧集内筛选出子集参与当前迭

代过程. SDFMAP[30]将 MLP 的决策面逼近场景表面, 利用 MLP 参数计算 TSDF 值获取场景信息. SDFMAP 无需任何预训练, 通过深度和颜色 2 个分支处理TSDF: 在深度分支中, MLP预测的TSDF被转换为场景深度并被输入的深度图监督; 在颜色分支中, 通过累积分布函数(cumulative distribution function, CDF)将 TSDF 截断范围中的颜色权重转换为像素颜色并由输入图像监督, 实现对 MLP 参数和相机位姿的优化.

iMAP[29]和 SDFMAP[30]将隐式建图方式直接引入 SLAM 建图环节进行实验, 在证明该类方法可行的同时也暴露出了较多问题: MLP 参数在新场景中的训练和渲染占据大量资源并耗费大量计算时间, 使得系统很难满足实时性要求; MLP 参数迭代导致地图遗忘问题, 使得系统无法适应不断拓展的地图; 建图的鲁棒性仍有待提高等. 针对这些问题, 研究者提出一些针对建图环节的改进方法[29-45], 以提高 SLAM 系统的性能, 如表 1 所示.

表 1　SLAM 系统的建图环节改进方法

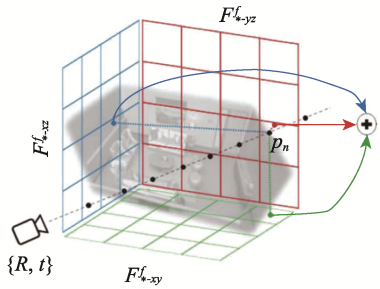
改进方向	SLAM 系统	传感器	几何信息形式	存储载体
基础	iMAP[29]	RGB-D	密度	MLP 参数
	SDFMAP[30]	RGB-D	SDF	MLP 参数
建图渲染速度提高	NICE-SLAM[31]	RGB-D	占用	体素网格
	ESLAM[32]	RGB-D	TSDF	三平面网格
	Vox-Fusion[33]	RGB-D	SDF	体素网格
	Co-SLAM[34]	RGB-D	TSDF	体素网格
	Plenoxel-SLAM[35]	RGB-D	密度*	体素网格
大规模场景扩展	MeSLAM[36]	RGB-D	密度	MLP 参数
	NISB-Map[37]	RGB-D	密度	MLP 参数
	MIPS-Fusion[38]	RGB-D	TSDF	MLP 参数
	NEWTON[39]	RGB-D/RGB	密度	球面网格
	FD-SLAM[40]	RGB-D	TSDF*	体素
	Multiple-SLAM[41]	RGB-D	SDF	体素网格
建图鲁棒性提高	UncLe-SLAM[42]	RGB-D(可加 1 个深度传感器)	占用	体素网格
	DI-Fusion[43]	RGB-D	概率 SDF*	体素
	MLM-SLAM[44]	RGB-D	占用	体素网格
	iMODE[45]	RGB	密度	MLP 参数

注. *表示体素网格直接存储.

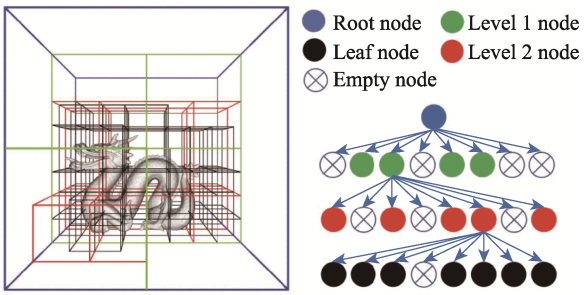
2.1　建图渲染速度提高方法

iMAP[29]通过建立特定相机位姿下的输入信息和已建图部分的渲染结果之间的损失函数, 实现对相机位姿和建图参数的联合优化. 在优化过程中, 参与迭代的建图参数数量和渲染采样点的隐式场景表示计算都会影响系统的实时性.

NICE-SLAM[31]通过减少参与迭代的建图参数数量的方法提高系统运行时的建图渲染速度, 使用多尺度体素网格作为存储载体, 并在网格顶点处存储特征; 在渲染时, 解码器对局部多尺度特征网格进行解码, 获取采样点处颜色信息和尺度不一的占用值. 其中, 由于占用值解码器参数已通过预训练确定, 因此每次迭代时只需要更新颜色解码器的参数和采样点周围参与解码的局部网格特征参数, 有效地减少了迭代参数量. 该方法在所有体素网格中进行特征存储, 充分地保留场景信息, 但也使得不包含场景表面信息的体素网格也参与了优化迭代, 同时, 迭代优化 MLP 解码器也部分地影响了系统运行速度. 为进一步降低迭代参数量, 基于 Chan 等[46]提出的三平面特征网格方法, ESLAM[32]建立了多尺度轴对齐三平面特征网格存储场景信息, 如图4a所示. 在渲染时, 首先将光线采样点投影到不同尺度的特征平面上, 并通过最近邻插值获取投影点的特征; 然后将三平面投影点的特征相加获得当前尺度下的特征; 最后通过连接不同尺度的特征输入几何解码器, 获得该空间点的 TSDF 值. ESLAM 针对位于截断距离内外的采样点建立不同的损失函数, 迭代优化特征平面参数、相机位姿和 MLP 解码器参数. 由体素网格到三平面网格的地图载体变化有效地降低了迭代参数量, 并且地图可随着分辨率的提高而高效扩展, 但会造成某些空间位置细节信息的丢失.



a. 三平面特征网格[32]



b. SVO[33]

图 4 建图渲染速度提高方法

渲染采样点的隐式场景表示计算包括采样点的定位、特征计算和特征解码等环节，因此，加速渲染采样点的定位是提高建图渲染速度的有效方法. 去除不包含场景表面信息的体素网格和引入位置编码[47]可有效地提高采样点定位速度. Vox-Surf[48]将空间划分为有限的稀疏体素，采用渐进式训练逐步剔除空体素；神经稀疏体素场[49]则采用稀疏体素八叉树(sparse voxel octree, SVO)跳过不包含场景表面信息的体素；Vox-Fusion[33]在引入如图 4b 所示的 SVO 划分场景的同时使用 Morton 编码体素坐标，可快速地检索采样点所在的体素网格，并解码相关特征获取连续 SDF；Co-SLAM[34]采用 Hash 编码多分辨率特征体素网格存储场景，使用 OneBlob[50]坐标编码提供场景平滑度和一致性先验. 为了提高关键帧的插入频次和实现更大关键帧库的维护，Co-SLAM 采用关键帧的一个随机采样像素子集代表该关键帧，在未大幅损失建图性能的前提下达到了 10~17 Hz 的运行速度，具有较好的实时性；Plenoxel-SLAM[35]基于 Plenoxels 算法[51]进行建图，通过规避 MLP 特征解码器的特征解码和迭代更新过程提高建图渲染速度. 在稀疏体素网格的每个顶点，Plenoxel-SLAM 采用固定的向量格式存储几何信息和颜色信息，包括 1 个密度值和每个颜色通道各 9 个球谐函数系数值，同时给出向量优化的分析导数方程，实现了对由朗伯曲面和镜面反射引起的视图相关关系建模. 与对特征进行解码后再渲染的方法相比，使用三线性插值直接产生像素的密度和颜色的速度更快，但复杂的球谐系数颜色表示部分提高了参与迭代的参数量.

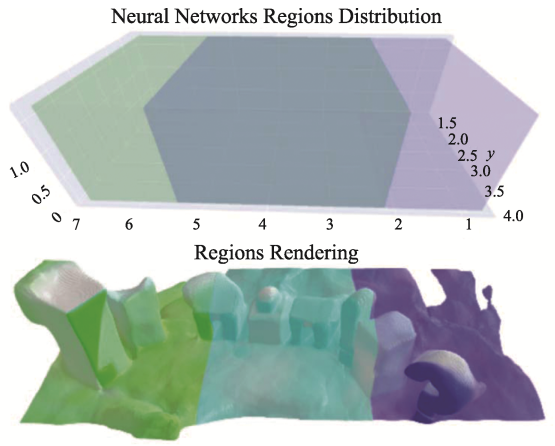
由上述分析可知，隐式建图视觉 SLAM 系统主要通过降低迭代参数量和加速渲染采样点的隐式场景表示计算提高建图渲染速度，以满足 SLAM 系统实时性要求.

2.2 大规模场景扩展方法

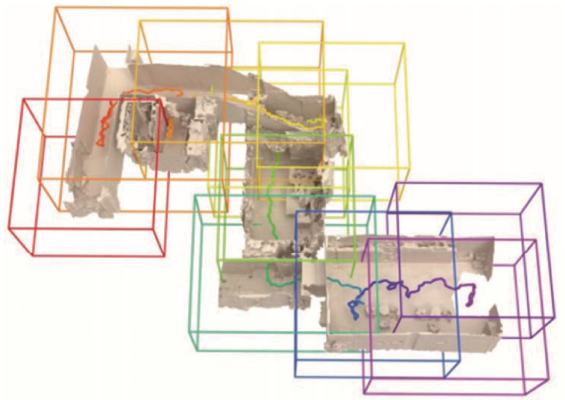
iMAP[29]每次优化都会更新全部参数，这个过程导致较为严重的地图遗忘问题，使其无法应用于大规模复杂场景. 一种常见的解决方法是将大规模场景划分成众多局部场景并建立局部子地图，再通过拼接子地图完成整个场景的重建，因此子地图的创建方式和子地图间的拼接方法，是将隐式建图视觉 SLAM 系统应用于大规模场景的关键问题.

一种子地图的创建方式是基于坐标位置划分整个场景. KiloNeRF[52]利用多个微小 MLP 取代原 NeRF 方法中的单个大型 MLP 的思路，将多个微小 MLP 在分解原本单个 MLP 权重空间的同时还能够并行构建，与原始 NeRF 模型相比，渲染速度提高了 3 个数量级；Block-NeRF[53]将场景分解为多个 block，其中每个 block 由单独一个 NeRF 进行场景表示，使得渲染时间不再与整个场景的规模大小直接相关，有效地扩大可渲染的场景规模，最终实现了城市级的 NeRF 重建. 基于以上思路，MeSLAM[36]使用如图 5a 所示的将整个场景基于坐标位置以固定大小进行子地图划分的方法，对划分出的每个子地图使用一个 MLP 进行 NeRF 建图. 为了完成子地图间的拼接，MeSLAM 使子地图之间具有固定比例的重叠，并且在重叠区域联合优化相机位姿与 NeRF 建图，最终，子地图拼接结果满足了没有明显变形、扭曲重叠和位移的地图拼接基本要求；但是，不同区域之间的建图仍存在明显的密度不连续的问题，且部分牺牲了建图的准确性和精度. 因为 MeSLAM 仅能对既有场景进行划分，所以难以通过跟随相机移动实现地图动态扩展. 为了实现地图可动态扩展的目标，NISB-Map[37]中提出具有固定立方体大小的 NeRF 场景表示——可扩展神经隐式空间块(neural implicit spatial blocks, NISB). 基于具有相机位姿信息的 RGB-D 图像序列，NISB-Map 根据当前帧的相机位姿建立视锥，并在视锥内搜索 NISB；若未搜索到，则创建新的 NISB，实现场景规模不断扩大的情况下的 NeRF 子地图动态创建. 为了解决相邻 NISB 之间密度不连续导致

的边界伪影问题，NISB-Map 设置相邻 NISB 具有部分重叠，并将知识蒸馏应用于重叠区域以确保几何连续性. NISB-Map 实现了场景动态扩展，但基于坐标系的 NISB 搜索会花费大量时间，同时用于处理重叠区域的知识蒸馏方法也增加了计算资源消耗.



a. 基于坐标划分已有场景[36]



b. 沿相机轨迹创建[38]

图 5　子地图创建策略

与基于坐标位置进行子地图划分和搜索的解决思路不同，SLAM 系统的关键帧策略提供了子地图划分的另一种方案. MIPS-Fusion[38]通过沿相机运动轨迹增量分配并动态地学习多个子地图实现地图的扩展，在当前关键帧相对上一个关键帧的移动大于阈值时，如图 5b 所示创建一个新的子地图；其将建立新子地图时最早出现的关键帧定义为锚定关键帧，同时将锚定关键帧对应的相机位姿定义为当前子地图在世界坐标系下的锚点. 为了保证子地图之间的连续性，相邻的子地图间共享至少一个关键帧. 在每个子地图内，MIPS-Fusion 通过 MLP 参数获取 TSDF 进行隐式建图，同时结合基于梯度优化和随机优化的相机位姿跟踪方法，可适应相机快速运动的场景. NEWTON[39]也采用类似的思路建立局部坐标系，与 MIPS-Fusion 不同，NEWTON 引入 ORB-SLAM2[11]进行相机跟踪和回环检测，并采用可以表示无界场景的球面坐标系多尺度特征网格表征子地图. MIPS-Fusion 和 NEWTON 都是沿着相机轨迹创建子地图，但缺乏鲁棒的回环检测，因此相机跟踪的漂移误差影响了子地图的对齐和最终的建图质量.

创建或划分子地图后，如何高质量地完成子地图拼接也是大规模场景隐式建图 SLAM 的重要问题. 只依靠子地图之间的重叠部分进行拼接易出现衔接处不平滑或产生伪影的问题，同时，多个子地图拼接时坐标调整也比较困难. FD-SLAM[40]采用特征提取网络提取用于回环检测和重定位的 ORB 特征，使用全局 BA 联合优化特征点的位置和子地图坐标系的位姿，实质上仍是依赖于传统的 SLAM 跟踪组件. Multiple-SLAM[41]提出采用浮点 SVO 存储地图信息提升子地图的拼接速度，解决了大多数稀疏 SVO 只支持整数坐标访问而造成的点坐标变换受限制的问题. Multiple-SLAM 计算浮点 SVO 根节点的 3 个正交方向量，使用根节点体素网格某一顶点和正交量的线性组合表示内部体素，由于对顶点及正交量进行变换相当于对局部地图内所有点进行变换，因此子地图的拼接速度大大提升. 该方法还对融合后的体素网格进行冗余删除，在确保地图融合的一致性和完整性的同时缓解了伪影问题. 然而，Multiple-SLAM 只在较大室内场景进行了测试，未能其在说明室外大规模场景的效果.

因此，隐式建图 SLAM 系统在重建大规模场景时聚焦于子地图的划分策略和拼接方法，在完成大范围场景重建的同时保证子地图融合的一致性.

2.3　建图鲁棒性提高方法

SLAM 的前端接收和处理的信息通常携带噪声和外点，通过引入不确定性可以筛除异常信息，放大更为准确的信息权重，使 SLAM 系统更加稳定. 在 NICE-SLAM[31]的基础上，UncLe-SLAM[42]将特征网格减少为 2 个尺度，利用计算出的法线图与光线投射之间的夹角，联合深度图输入不确定性解码器预测深度不确定性，预测的深度不确定性可以对用于联合优化相机位姿跟踪和建图的损失函数重新进行权重分配. 该系统还能处理除 RGB-D 数据外其他深度传感器数据输入的情况. UncLe-SLAM 的不足是解码器接受的特征过于简单，难以计算精细细节的不确定性. DI-Fusion[43]中提出概率局部隐

式体素(probabilistic local implicit voxel, PLIVox), 可以对每个体素网格进行精细的不确定性计算. DI-Fusion 中, 参考 CodeSLAM[16]将深度作为先验信息, 利用预训练的 MLP 编码器-解码器向体素内存储, 其本质是正则高斯分布的概率 SDF, 可同时完成对场景几何和几何不确定性的记录; 由于其编码器和解码器之间存在潜在特征向量, 在潜在特征向量域上执行几何积分, 可以在保证场景重建质量的情况下提高渲染速度. 但是, DI-Fusion 中每个 PLIVox 相互独立, 导致场景重建的空间连续性可能会被破坏.

除了引入不确定性, 将输入信息经过多种方式处理的结果进行相互印证和补充, 也是提高系统鲁棒性的有效方法. MLM-SLAM[44]对 MLP 解码器进行改进, 创建了一个用于多尺度特征网格的多 MLP 可微渲染框架, 可在没有任何先验信息的情况下记录场景的细节; 4 个解码器按对应的体素网格尺度大小被分为大型解码器和小型解码器, 其中, 2 个大型解码器不直接参与渲染, 而是为小型解码器提供残差; 小型解码器接受包括残差等众多场景信息并进行渲染, 可以实现更丰富的高频细节的呈现. 但是, 复杂的解码器结构延长了 MLM-SLAM 的数据处理时间. iMODE[45]提供一种适用于多尺度场景的基于 RGB 的视觉 SLAM 系统方案, 在较小的室内场景和较大的室外场景都能完成跟踪建图任务; 其主要贡献是利用频域处理, 将用于渲染的采样点通过不同位置编码函数映射到高频和低频的频率集中, 再将低频信息输入 MLP 后将输出结果输入密度解码器, 并将同一输出结果联合高频信息输入颜色解码器, 实现了大规模场景重建和小于 10 cm 的细节重建渲染. 然而, iMODE 的跟踪与建图耦合松散, 无法通过联合优化进一步提高建图精度.

由上述分析可知, 隐式建图视觉 SLAM 系统通过引入不确定性和多种处理结果的相互补充和印证, 提高了建图的鲁棒性.

3 隐式建图 SLAM 的其他环节改进方法

将隐式建图融入视觉 SLAM 后, 研究者通过改进前端提高系统跟踪精度和建立基于 RGB 数据的隐式建图视觉 SLAM 系统, 同时, 通过补充回环检测进一步提高跟踪建图性能, 如表 2 所示.

3.1 前端改进方法

提高相机位姿跟踪精度是 SLAM 系统前端改进的重要目标. iMAP[29]利用与跟踪建图联合优化过程相同的损失函数和优化器直接优化相机位姿, 但这类方法的精度和速度通常不如经典特征法 SLAM. iDF-SLAM[54]中提出将特征法的前端与隐式建图方法相结合的思路, 使用基于特征的神经跟踪器作为前端跟踪相机轨迹, 同时结合单个 MLP 计算 TSDF 的建图方法实现了端到端的 SLAM 系统; 其神经网络跟踪器在 ScanNet 数据集[63]上完成预训练并在实际运行中进行调整, 实现当前场景的特定特征的提取. iDF-SLAM 的不足是神经跟踪器对于剧烈运动和模糊图像的处理存在困难. SVR-Net[55]也是结合特征法建立的端到端 SLAM 系统, 其采用稀疏三维卷积匹配网络进行特征匹配, 输入为一对 RGB 帧, 输出的是帧间的相对位姿和对应的局部体素 TSDF; 该系统的前端引入语义编码器对场景尺度等信息进行编码, 并通过在迭代中嵌入 Gauss-Newton 算法施加几何约束, 以提高位姿估计的精度. 但是, SVR-Net 只使用相邻帧的信息导致累积误差的产生, 限制了定位精度的进一步提高.

表 2 隐式建图 SLAM 系统的其他环节改进方法

改进方向	SLAM 系统	传感器	几何信息形式	存储载体
前端改进	iDF-SLAM[54]	RGB-D	TSDF	MLP 参数
	SVR-Net[55]	RGB	TSDF	MLP 参数
	DIM-SLAM[56]	RGB	密度	体素网格
	NICER-SLAM[57]	RGB	SDF	体素网格
	HI-SLAM[58]	RGB	SDF	体素网格
回环检测补充	Orbeez-SLAM[59]	RGB	密度*	体素网格
	NGEL-SLAM[60]	RGB-D	占用	体素网格
	NeRF-SLAM[61]	RGB	密度	MLP 参数
	GO-SLAM[62]	单目/双目/RGB-D	SDF	MLP 参数

注. *表示体素网格直接存储.

鉴于大部分的隐式建图视觉 SLAM 系统仍是基于 RGB-D 数据实现, 研究者通过改进前端, 使得隐式建图 SLAM 能够基于 RGB 数据实现运行. DIM-SLAM[56]基于多视图立体视觉建立了以图像块为单位的结构相似性指数(structural similarity, SSIM)损失函数, 以计算并优化深度信息, 是较早尝试基于 RGB 进行隐式建图的 SLAM 系统, 其效果与 RGB-D 系统 NICE-SLAM[31]类似. 除了利用立体视觉, 还可以借助神经网络获取深度信息, 如图 6 所示. NICER-SLAM[57]直接在前端引入已有的神经网络模型进行单目深度估计和单目法线估计, 采用多尺度特征体素网格存储场景, 并用 MLP 解码器

解码 SDF；将由 SDF 计算的深度和法线结果，以及神经网络的估计结果都作为损失函数的一部分．该系统集中了颜色、法线、深度、光流和光度等众多信息，虽然几何信息丰富，但多个神经网络模型和复杂的损失函数大幅降低了 NICER-SLAM 的运行速度，无法满足实时性要求．在 DROID-SLAM[17]前端的基础上，HI-SLAM[58]通过引入训练完成的单目深度法线估计神经网络，实现几何信息估计．DROID-SLAM 是基于光流估计方法 RAFT(recurrent all-pairs field transforms)算法[64]的可处理单目、双目和 RGB-D 数据的 SLAM 系统，它以循环迭代的方式实现了比较优秀的相机位姿跟踪．HI-SLAM[58]基于 RGB 数据，采用特征体素网格解码 SDF 进行场景表征，通过联合优化深度和尺度解决了单目 RGB 中固有的尺度模糊问题，可以恢复更多的场景几何细节，并提高表面平滑度；在深度和尺度联合调整模块中，已有深度先验的尺度和偏移被作为变量纳入 BA 优化中，以保持深度先验的尺度一致性．为了提高优化速度、减轻相机位姿漂移和尺度漂移，HI-SLAM 采用基于 Sim(3)的位姿图束调整(pose graph bundle adjustment, PGBA)方法代替全局 BA，其中 Sim(3)是指基于 3 对点计算相似变换的方法．

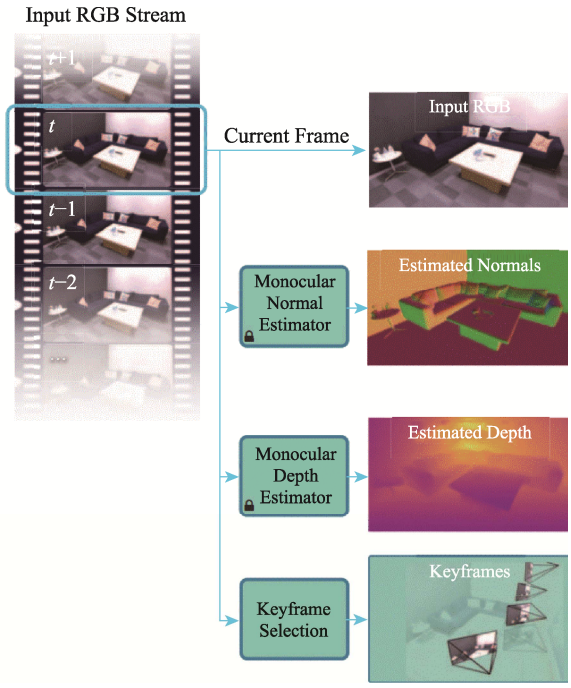


图 6　神经网络为单目图像补充几何信息[57]

通常，改进前端的目的是进一步提高跟踪精度，为此，研究者引入基于特征的前端跟踪方法．此外，隐式建图 SLAM 系统通过改进前端获取深度信息或直接引入成熟 SLAM 系统的前端，可实现基于 RGB 数据的同步跟踪与建图．

3.2　回环检测方法补充

回环检测可以进一步提高 SLAM 跟踪与建图的精度，但其在早期的隐式建图 SLAM 中建图渲染速度较慢，引入回环检测会进一步降低系统运行速度．2022 年，Müller 等[47]提出即时神经图形原件(instant neural graphics primitives, Instant-NGP)，利用新颖的位置参数编码机制和低计算复杂度的 Hash 算法，将场景 MLP 的训练学习完成时间压缩到了秒级；该建图方法的高速度为将回环检测引入隐式建图 SLAM 提供了基础．但是，由于利用隐式场景表示或体素网格内特征直接建立回环比较困难，因此通常采用传统方法进行回环检测．

ORB-SLAM 系列[10-12]在前端提取出的 ORB 特征被用于回环检测中．Orbeez-SLAM[59]结合 ORB-SLAM2[11]的前端和 Instant-NGP 建图方法，在无需进行任何预训练的前提下，将关键帧集和相机位姿输入 MLP 训练其内部参数，通过渲染将前端产生的稀疏点云上采样成密集点云．基于 Orbeez-SLAM[59]框架，NGEL-SLAM[60]引入子地图方法并补充回环检测，利用 ORB-SLAM3[12]进行相机位姿估计，并通过局部 BA 对当前子地图进行优化，将前端提取出的 ORB 特征用于回环检测；其子地图策略与 MIPS-Fusion[38]相同，即由构建该子地图的第 1 个关键帧作为锚定关键帧．在利用 ORB 特征检测到回环位置时，NGEL-SLAM 对所有的锚定关键帧进行 BA 优化，子地图将跟随锚定关键帧进行整体移动，实现了全局地图的调整．

NeRF-SLAM[61]将 DROID-SLAM[17]的前端和 Instant-NGP[47]建图方法相结合，通过引入相机位姿和深度的不确定性为建图提供更加丰富的信息．基于 NeRF-SLAM 的框架，GO-SLAM[62]添加了基于共视矩阵的回环检测环节．共视矩阵是利用反向投影计算关键帧对之间的平均刚体光流得到的．GO-SLAM 首先将接收到的新关键帧与时间上相邻的一些关键帧联合进行局部 BA，然后计算该新关键帧与历史上所有关键帧的共视性，最终建立所有关键帧之间的共视矩阵，如图 7 所示．可以看出，当 2 个关键帧之间的共视性超过设定的阈值时，GO-SLAM 将选取极大值处建立 2 个关键帧之间的回环关系．该系统同时对周围关键帧进行共视性的非极大值抑制，最后通过多个回环实现全局 BA 以提高精度；但是，所有关键帧之间共视矩阵的建立导致了计算工作量的大幅提高．

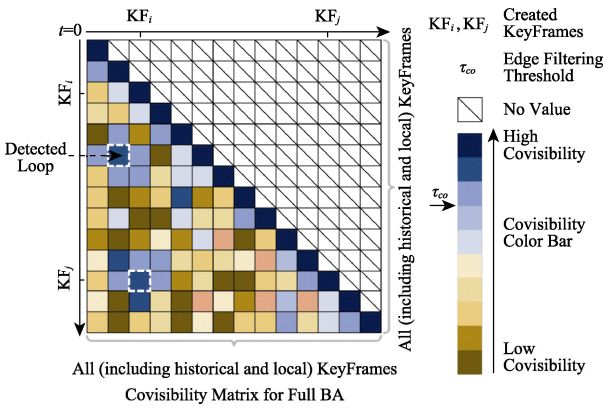


图 7 基于共视性矩阵的回环检测[62]

由上述分析可知，隐式建图 SLAM 系统回环检测的建立大多基于已有的方法，如特征法和共视性. 隐式场景表示本身的属性，在一定程度上限制了基于隐式建图场景的回环检测.

4 面向特定应用方向的隐式建图 SLAM

语义地图、动态场景和多传感器融合是视觉 SLAM 系统应用的重要方向，研究者将隐式建图视觉 SLAM 系统进行有针对性的改造并证明了其在这些领域的可行性，如表 3 所示.

表 3 面向特定应用方向的隐式建图 SLAM

应用方向	SLAM 系统	传感器	几何信息形式	存储载体
语义建图	Structerf-SLAM[65]	RGB-D	占用	体素网格
	FR-Fusion[66]	RGB-D	密度	MLP 参数
	vMAP[67]	RGB-D	占用	MLP 参数
	RO-MAP[68]	RGB	密度	MLP 参数
动态场景	DN-SLAM[69]	RGB-D	密度	MLP 参数
多传感器融合	SimpleMapping[70]	RGB+IMU	TSDF*	体素网格
	TOW-SLAM[71]	RGB-D+IMU	密度	球面网格
	ToF-SLAM[72]	RGB+低分辨率深度信息	SDF	体素网格

注. *表示体素网格直接存储.

4.1 语义建图

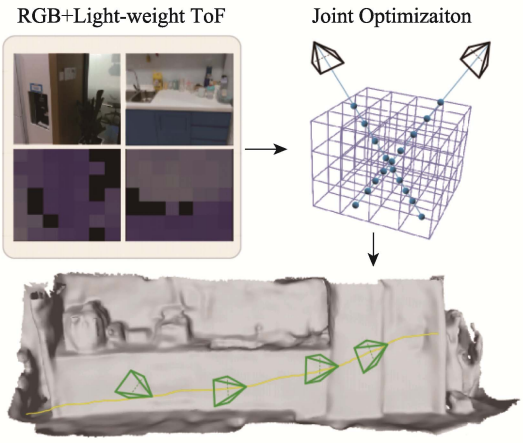
在语义地图方面的应用方向上，隐式建图视觉 SLAM 系统包括场景内平面识别与分割、像素级语义建图和物体级语义建图等. Structerf-SLAM[65]使用超像素分割算法[73]对每幅图像提取超像素，并将面积大于预定义阈值的区域视为平面；然后将平面抽象为平面特征，用于对跟踪和建图添加约束. 该系统在跟踪阶段添加平面特征匹配约束，通过将三维平面的法向量和平面深度进行数据关联，克服了纹理信息不足带来的问题；在建图阶段添加一致性深度约束，较好地拟合渲染出的平面深度，实现精度更高的地图重建；在低帧率和稀疏数据的情况下也能保持较高的稳定性，但其处理速度与 iMAP[29]相差不大，无法适应高速视频流信息.

为了实现像素级语义建图，FR-Fusion[66]选用与语义属性关系更密切的高维抽象特征图进行场景表示，基于 iMAP[29]引入潜在特征体渲染技术. 为了解决提取后生成的特征图空间分辨率通常低于原始图像分辨率的问题，采用 Semantic NeRF[74]中用于特征空间插值的 MLP 稀疏监督方法，使得语义预测结果能够继承形状和颜色重建的几何一致性，确保语义区域在稀疏注释下也可以准确地拟合对象的形状. 然而，FR-Fusion 在遇到未训练到的标签类型时，需要用户为少数像素手动提供标签，造成了使用时的不便.

在物体级语义建图方面，基于 Object-NeRF[75]，vMAP[67]中提出将场景分为物体和背景的隐式建图方法，用较大的神经网络表示背景，并且用多个结构相同的 MLP 对应表示场景中的多个物体，同时通过并行训练减少参数训练所需时间. vMAP 采用数据集提供的分割结果或利用二维实例分割网络获取语义分割结果，通过结合新的关键帧和相机位姿的输入不断地迭代优化三维物体边界，最终在单个场景中实现了约 50 个单独物体的优化. 然而，该系统缺乏合适的全局约束，导致较难处理语义分割错误产生的模糊问题. 在不依赖三维先验的前提下，RO-MAP[68]通过处理单目 RGB 图像和实例分割图，实现了场景中物体对象的定位和重建，利用基于 ORB-SLAM2[11]改造的轻量级前端处理实例分割的语义信息和与物体对象关联的稀疏点云，完成物体位姿计算和尺度估计；建图时，每个对象实例都由一个单独的 MLP 表示，利用估计出的物体位姿、尺度和原始图像等多种信息并行训练多个 MLP，如图 8a 所示，最终实现了物



a. 物体级语义建图[68]



b. 多传感器融合[72]

图 8　面向特定应用方向的隐式建图 SLAM 系统

体级密集语义地图的建立.

4.2　动态场景

DN-SLAM[69]是将 NeRF 应用于动态场景的视觉 SLAM 系统, 其结合 ORB-SLAM3[12]和 Instant-NGP[47], 使用光流估计方法获得动态信息, 使用粗略和精细 2 个尺度的语义分割网络进行图像特征分割, 并检测具有潜在运动可能性的对象, 同时利用静态区域特征点信息进行建图, 修复被动态对象遮挡的背景. 由于场景信息会被遮挡或模糊, DN-SLAM 在修复被遮挡的场景信息时仍存在较大误差.

4.3　多传感器融合

在多传感器融合方面, 惯性测量单元(inertial measurement unit, IMU)通常能够为视觉 SLAM 的相机位姿跟踪提供更多信息. SimpleMapping[70]利用 ORB-SLAM3[12]的前端处理 RGB 图像并结合 IMU 数据完成相机位姿跟踪, 生成具有三维稀疏点的局部地图; 在建图时, 首先将关键帧及对应二维稀疏深度图输入到稀疏点辅助多视图立体网络(sparse point aided multi-view stereo neural network, SPA-MVSNet)中恢复密集深度图, 然后通过 Voxel Hashing[76]逐步融合到存储 TSDF 的体素网格中. SimpleMapping 的不足是处理低纹理场景和存在反射或遮挡的场景比较困难. TOW-SLAM[71]基于 NICE-SLAM[31]融入 IMU 数据, 同时引入球形网格表示无边界场景; 利用 IMU 数据建立损失函数, 以优化相机位姿跟踪, 同时, 通过将多个 IMU 数据预处理成相对运动增量(relative motion increment, RMI), 约束建图时使用的关键帧间相机位姿变化.

ToF-SLAM[72]是基于单目 RGB 相机和一个轻量级 ToF(time of flight)传感器进行密集场景重建的 SLAM 系统. 与 RGB-D 相机相比, 轻量级 ToF 深度传感器硬件质量轻且成本较低, 但其深度测量非常稀疏且有噪声, 所以很少被用于密集几何重建. ToF-SLAM 使用的 ToF 传感器同样只能返回低分辨率的深度分布, 即只能获取较大区域的深度平均值和方差. 为了实现深度上采样, ToF-SLAM 首先将新观测到的 ToF 信号和从原有重建场景中初始化的 ToF 信号进行融合以实现去噪; 然后联合 RGB 信号输入 DELTAR[77]中, 获取逐像素的深度预测结果. ToF-SLAM 采用多尺度特征网格存储场景, 将 Mip-NeRF[78]中提出的集成位置编码(integrated positional encoding, IPE)理论推广到基于特征体素网格中, 实现了低分辨率信号(ToF 信号)的区域级渲染, 以及如图 8b 所示的高分辨率信号(RGB 和像素级深度)的像素级渲染.

从上述分析可知, 在像素级和物体级语义建图 SLAM 中, 引入了有针对性地改造后的隐式建图方法; 在动态环境 SLAM 中, 采用传统方法进行动态点去除; 进行多传感器融合的系统实现时, 则需要根据传感器性质对隐式建图 SLAM 进行针对性改造.

5　评价标准与结果分析

评价标准的统一和数据集的建立为 SLAM 系统的性能量化奠定了基础. 本节讨论隐式建图 SLAM 系统常用的定量评价标准和数据集, 同时, 基于相同的数据集比较并分析多个隐式建图视觉 SLAM 系统的性能表现.

5.1　评价标准

理想的 SLAM 标准通常包括稳定的相机跟踪、精准的场景建模、实时性能、对大型场景的可扩展性, 以及对噪声数据的鲁棒性等. 通常, 通过对场景相机位姿跟踪、地图重建和地图渲染, 以及设立

定量评价标准评估视觉 SLAM 系统的性能.

5.1.1 相机跟踪

在相机位姿跟踪方面, 绝对轨迹误差(absolute trajectory error, ATE)是衡量 SLAM 系统相机位姿跟踪水平的常用指标, 其值越小越好. 计算 ATE 时, 首先对齐 SLAM 系统生成的完整相机轨迹与数据集提供的真值, 然后计算评估二者之间的绝对位姿差距. ATE 反映了系统的跟踪精度和轨迹的全局一致性, 大部分视觉 SLAM 系统采用 ATE 衡量相机位姿跟踪性能的指标. 假设系统给出的一条估计轨迹第 i 时刻的相机位姿为 $T_{\mathrm{esti},i}$, 参考真值轨迹第 i 时刻的相机位姿为 $T_{\mathrm{gt},i}$, 其中, $i=1,2,\cdots,N$, 则 ATE 可以定义为

$$\mathrm{ATE}=\sqrt{\frac{1}{N}\sum_{i=1}^{N}\left\|\log\left(T_{\mathrm{gt},i}^{-1}T_{\mathrm{esti},i}\right)^{\vee}\right\|_2^2}.$$

其中, $\log(\cdot)^{\vee}$ 表示将位姿变换矩阵转换为李代数形式.

5.1.2 地图重建

地图重建指标反映 SLAM 所建三维地图几何形状与场景真值的差距, 评价标准包括 Depth L_1, Accuracy, Completion 和 F_1 (F1-Score). 其中, Depth L_1 是基于随机采样计算的重建地图深度图和对应真值的深度图之间的 L_1 损失函数值, 其值越小越好; Accuracy 表示重建网格采样点与最近的地面真值点之间的平均距离, 其值越小越好; Completion 表示真值网格采样点与最近重建点之间的平均距离, 其值越小越好, Completion Ratio 是 Completion 值小于某阈值的点占所有真值采样点的比例, 该阈值通常设置为 5 cm. Depth L_1, Accuracy, Completion 和 Completion Ratio 是隐式建图 SLAM 系统中常用的地图重建评价标准. F_1 是精确度 P 和召回率 R 的调和平均值, 计算公式为

$$F_1=\frac{2\times P\times R}{P+R}.$$

其中, P 表示重建网格采样点与真值网格采样点的距离小于阈值 d 的点占所有重建网格采样点的比例; R 表示真值网格采样点和重建网格采样点的距离小于阈值 d 的点占所有真值网格采样点的比例.

5.1.3 地图渲染

地图渲染评估指标通过比较基于数据集的真值网格渲染出的图像和 SLAM 系统产生的渲染图像, 定量说明 SLAM 系统的地图渲染水平, 包括峰值信噪比(peak signal-to-noise ratio, PSNR)、结构相似性指数(structural similarity index, SSIM)和神经网络图像块感知相似性(learned perceptual image patch similarity, LPIPS).

PSNR 可以评估一幅图像与原始图像之间的相似度, 通常用于图像重建和压缩. 在隐式建图 SLAM 系统中, 常用 PSNR 作为建图渲染结果评价标准, 其值越高, 表示 2 幅图像之间的相似度越高, 渲染质量越高. PSNR 的计算公式为

$$\mathrm{PSNR}=10\cdot\lg\left(\frac{M^2}{\mathrm{MSE}}\right).$$

其中, M 表示图像像素值的最大可能取值(如对于 8 位图像, 该值为 255); MSE 表示 2 幅图像的逐像素均方误差(mean squared error, MSE).

SSIM 考虑了亮度、对比度和结构信息等多项因素, 也可用于衡量 2 幅图像之间的相似度. SSIM 的值范围是在–1~1, 越接近 1, 表示 2 幅图像越相似, 渲染质量越高. 对于 2 幅图像 x 和 y, SSIM 的计算公式为

$$\mathrm{SSIM}(x,y)=\frac{\left(2\mu_x\mu_y+C_1\right)\left(2\sigma_{xy}+C_2\right)}{\left(\mu_x^2+\mu_y^2+C_1^2\right)\left(\sigma_x^2+\sigma_y^2+C_2^2\right)}.$$

其中, μ_x 和 μ_y 分别表示图像 x 和 y 的亮度均值; σ_x^2 和 σ_y^2 分别表示图像 x 和 y 的亮度方差; σ_{xy} 表示图像 x 和 y 之间的亮度协方差; C_1 和 C_2 是常数, 用于稳定计算, 通常设置较小的正数值.

LPIPS[79]是利用深度学习方法测量 2 幅图像之间感知相似性的指标, 能够更好地模拟人类视觉感知, 由经过大规模训练的深度神经网络模型实现, 分数越低, 说明 2 幅图像越相似, 渲染质量越高.

5.2 数据集

数据集是比较各类研究方法的基础, 当前, 隐式建图 SLAM 常用的数据集有 ScanNet[63], TUM RGB-D[80], Replica[81], ScanNet++[82]和 EuRoC[83]等. 其中, ScanNet[63], TUM RGB-D[80]和 Replica[81]是隐式建图视觉 SLAM 系统常用的数据集.

TUM RGB-D 数据集[80]由慕尼黑工业大学的视觉组于 2012 年公开, 包含多组由微软 Kinect 相机采集的室内图像序列及与其时间同步的高精度相机位姿真值; 视频序列包含分辨率为 640×480 像素的颜色和深度 ToF 图像, 视频帧率为 30 Hz. 该数据集包括办公环境和工业大厅共 39 个图像序列, 涵盖了各种摄像机运动方式和多种场景, 按照场景中运动物体出现的频率和运动的剧烈程度, 将所有视频序列分为低动态场景和高动态场景 2 类.

Replica 数据集[81]是由 Facebook Reality Labs

在 2019 年推出的高质量室内场景三维重建合成数据集, 包含 18 个根据真实世界数据合成的高真实感室内场景重建数据, 每个场景包括密集网格、高分辨率高动态范围纹理、众多基本语义类及其实例信息, 部分场景包含平面反射镜和玻璃反射镜. 该数据集致力于辅助研究基于视觉、几何或语义的真实世界场景模型机器学习生成方法.

ScanNet[63]是斯坦福大学提供的数据集, 共包含针对 707 个室内环境的 1 513 次 RGB-D 扫描数据, 同时提供了对应的相机位姿、表面重建、纹理网格和语义重建的注释及其中部分扫描序列的 CAD 建模. 由于该数据集的数据真值是由 Bundle-Fusion 算法[84]计算获取而不是利用各类传感器和设备采集, 因此其准确度和可靠性存疑.

ScanNet++[82]是由慕尼黑工业大学采集并建立的数据集, 包含 460 个室内场景的高分辨率重建和密集实例语义标注; 数据集的所有场景共包含 28 万幅单反相机高质量图像和超过 370 万帧的 iPhone 商品级 RGB-D 数据. 其中, 单反相机图像适用于基准测试和具体方法的通用性能测试, iPhone 捕捉的 RGB-D 流提供了具有运动模糊和包含噪声相机位姿的训练测试数据. 为了提供场景更全面细致的语义信息, 在场景重建中采用多标签注释并明确标注模糊场景中的各类标签.

EuRoC 数据集[83]是基于欧洲机器人挑战赛(European robotics challenge, EuRoC)背景产生的数据集, 用于评估参赛者的微型飞行器的 SLAM 系统性能或三维重建能力, 包含同步立体图像、IMU 数据和 6 自由度位姿真值. 其中包括 2 类数据, 一类是在工业环境中收集的、可用于评测系统在真实场景中性能的数据, 另一类是在室内环境收集的、评测精确环境重建的数据. 该数据集包含良好视觉条件下的慢速飞行到具有运动模糊和较差照明的动态飞行等各类场景, 在 SLAM 研究中常被用于测试与 IMU 结合的视觉 SLAM 系统性能.

5.3　对比与分析

本节基于 TUM RGB-D[80], ScanNet[63]和 Replica[81]数据集, 比较多个隐式建图视觉 SLAM 系统的相机位姿跟踪和地图重建结果. 选取部分隐式建图视觉 SLAM 系统, 在同一硬件平台(单个 NVIDIA 3090 GPU)下运行并测量其运行显存占用和建图速度, 并对结果进行分析.

5.3.1　跟踪性能

在 TUM RGB-D 数据集[80]上, 相机跟踪结果如表 4 所示. 其中, fr1/desk, fr2/xyz 和 fr3/office 分别表示该数据集内的场景序列. 在 ScanNet 数据集[63]上, 相机跟踪结果如表 5 所示. 其中, 0000, 0059, 0106, 0169, 0181 和 0207 表示该数据集内的场景序列. 在 Replica 数据集[81]上, 相机跟踪结果如表 6 所示. 其中, R0, R1, R2, O0, O1, O2, O3 和 O4 表示该数据集内的场景序列.

表 4　在 TUM RGB-D 数据集[80]上相机跟踪结果

数据类型	SLAM 系统	ATE/cm			
		fr1/desk	fr2/xyz	fr3/office	平均
RGB-D	ORB-SLAM2[11]	1.6	0.4	1.0	1.0
	iMAP[29]	4.9	2.0	5.8	4.2
	SDFMAP[30]	3.6	1.9	3.3	2.9
	NICE-SLAM[31]	2.7	1.8	3.0	2.5
	ESLAM[32]	2.5	1.1	2.4	2.0
	Co-SLAM[34]	2.7	1.9	2.6	2.4
	MeSLAM[36]	6.0	6.5	7.5	6.7
	MIPS-Fusion[38]	3.0	1.4	4.6	3.0
	MLM-SLAM[44]	2.4	1.7	2.9	2.3
	NGEL-SLAM[60]	1.5	0.5	1.0	1.0
	GO-SLAM[62]	1.5	0.6	1.3	1.1
RGB	ORB-SLAM2[11]	1.9	0.6	2.4	1.6
	DROID-SLAM[17]	1.8	0.5	2.8	1.7
	DIM-SLAM[56]	2.0	0.6	2.3	1.6
	Orbeez-SLAM[59]	1.9	0.3	1.0	1.1
语义	vMAP[67]	2.6	1.6	3.0	2.4

从表 4~表 6 可以看出:

(1) 无论是基于 RGB-D 数据或是 RGB 数据, 与经典视觉 SLAM 系统(如 ORB-SLAM2[11])相比, 大多数隐式建图视觉 SLAM 系统在跟踪精度方面仍有一定差距.

(2) 通过引入 ORB-SLAM 系列[10-12]和 DROID-SLAM[17]等较成熟的前端方法, 并基于前端方法建立相应的回环检测, 能够帮助隐式建图视觉 SLAM 系统提高跟踪性能. 其中, NGEL-SLAM[60]引入 ORB-SLAM3[12]系列, HI-SLAM[58]和 GO-SLAM[62]则引入了 DROID-SLAM[17]的前端, 并且都具有回环检测环节, 这些隐式建图视觉 SLAM 系统在多个数据集上都能取得较为出色的跟踪结果.

将表 6 与表 4 和表 5 对比可以看出, 隐式建图视觉 SLAM 系统在 Replica 数据集[80]的跟踪误差普遍小于在 TUM RGB-D[80]和 ScanNet 数据集[63]上的跟踪误差, 原因在于 Replica[80]是合成数据集, 与真实采集的数据集相比几乎没有噪声和外点, 使得 SLAM 系统在 Replica[80]上的跟踪结果与现实场景实际跟踪结果之间可能存在差距.

表 5 在 ScanNet 数据集[63]上相机跟踪结果

数据类型	SLAM 系统	ATE/cm						
		0000	0059	0106	0169	0181	0207	平均
RGB-D	DROID-SLAM[17]	5.4	7.7	7.1	8.0	7.0		
	iMAP[29]	56.0	32.1	17.5	70.5	32.1	11.9	36.7
	NICE-SLAM[31]	8.6	12.3	8.1	10.3	12.9	5.6	9.6
	ESLAM[32]	7.3	8.5	7.5	6.5	9.0	5.7	7.4
	Vox-Fusion[33]	7.5		3.2	8.4	8.8	2.9	
	Co-SLAM[34]	7.2	12.3	9.6	6.6	13.4	7.1	9.4
	MIPS-Fusion[38]	7.9	10.7	9.7	9.7	14.2	7.8	10.0
	MLM-SLAM[44]	6.9	9.1	7.4	3.1	8.6		
	NGEL-SLAM[60]	7.2	7.0	8.0	6.1	10.1	6.3	7.4
	GO-SLAM[62]	5.4	7.5	7.0	7.7	6.8		
RGB	DROID-SLAM[17]	5.5	9.0	6.8	7.9	7.4		
	HI-SLAM[58]	6.4	7.2	6.5	8.5	7.6	8.4	7.4
	Orbeez-SLAM[59]	7.2	7.2	8.1	6.6	15.8	7.2	8.7
	GO-SLAM[62]	5.9	8.3	8.1	8.4	8.3		

表 6 在 Replica 数据集[81]上相机跟踪结果

数据类型	SLAM 系统	ATE/cm								
		R0	R1	R2	O0	O1	O2	O3	O4	平均
RGB-D	iMAP[29]	3.8	6.8	3.2	3.3	2.8	3.8	3.8	4.0	3.9
	SDFMAP[30]	1.7	2.0	1.8	1.2	1.7	2.3	2.4	2.0	1.9
	NICE-SLAM[31]	1.7	2.1	2.0	1.1	1.0	1.8	3.6	3.4	2.1
	ESLAM[32]									0.6
	Vox-Fusion[33]	0.3	1.3	0.5	0.7	1.1	0.5	0.3	0.6	0.7
	MIPS-Fusion[38]	1.1	1.2	1.1	0.7	0.8	1.3	2.2	1.1	1.2
	iDF-SLAM[54]	1.8	5.9	2.6	1.6	2.1	1.8	1.9	2.1	2.5
	GO-SLAM[62]									0.3
RGB	DROID-SLAM[17]									0.4
	DIM-SLAM[56]	0.8	1.2	0.7	0.9	0.7	1.1	0.8	0.9	0.9
	NICER-SLAM[57]	1.4	1.6	1.1	2.1	3.2	2.1	1.4	2.0	1.9
	GO-SLAM[62]									0.4

5.3.2 建图性能

在 Replica 数据集[81]上, 部分 SLAM 系统的地图重建结果如表 7 所示. 可以看出: (1) 在 RGB-D 隐式建图 SLAM 系统中, 地图重建性能较好的是 ESLAM[32]和 Co-SLAM[34]; 在 RGB 隐式建图 SLAM 系统中, NICER-SLAM[57]和 HI-SLAM[58]的地图重建性能较好: 这些 SLAM 系统的共同点是采取以体素网格(或类体素网格)作为存储载体, 以 SDF 或 TSDF 为几何信息形式的建图方式. (2) 以 MLP 参数作为存储载体或以占用或密度作为几何信息形式的建图方式, 需要渲染深度信息才能完成场景几何重建; 相比之下, 在三维层面体素网格结合 SDF/TSDF 的建图方式, 通过 Marching Cubes 算法[85]直接获取几何一致性较高的多边形网格结果, 有效地提高了地图重建的效果.

5.3.3 运行性能

在同一硬件平台(单张 NVIDIA 3090 GPU)、同一数据集场景(TUM RGB-D[80] fr1/desk)下, 多个开源隐式建图视觉 SLAM 系统运行后, 获取的运行显存占用和建图速度对比如表 8 所示. 其中, 运行显存占用指运行过程中平均显示内存占用, 建图速度指参与建图优化的帧总数除以建图优化时间. 表 8 中实际测试的 Orbeez-SLAM[59]的运行速度包含了跟踪与建图 2 个过程, 故其建图速度应高于该数

表 7　在 Replica 数据集[81]上地图重建结果对比

数据类型	SLAM 系统	Depth L_1	Accuracy/cm	Completion/cm	Completion Ratio <5 cm/%
RGB-D	iMAP[29]		4.4	5.6	79.1
	SDFMAP[30]		3.1	5.1	83.4
	NICE-SLAM[31]	3.5	2.9	3.0	89.3
	ESLAM[32]	1.2	1.0	1.1	98.6
	Vox-Fusion[33]		3.4	2.6	90.7
	Co-SLAM[34]	1.5	2.1	2.1	93.4
	MLM-SLAM[44]	2.3	3.6	2.8	90.4
	iDF-SLAM[54]		4.9	3.1	86.3
	GO-SLAM[62]	3.4	2.5	3.7	88.1
RGB	iMODE[45]		8.9	13.9	37.1
	DIM-SLAM[56]		4.3	5.5	77.4
	NICER-SLAM[57]		3.7	4.2	79.4
	HI-SLAM[58]	3.6	3.6	4.6	80.6
	GO-SLAM[62]	4.4	3.8	4.8	78.0
语义	vMAP[67]		3.2	2.4	93.0

表 8　运行显存占用和建图速度对比

SLAM 系统	运行显存占用/GB	建图速度/(帧·s^{-1})
iMAP[29]	9.10	0.02
NICE-SLAM[31]	10.87	0.10
ESLAM[32]	7.72	9.47
Co-SLAM[34]	4.95	12.18
Orbeez-SLAM[59]	3.15	>20.02

据. 可以看出, 与 iMAP[29]相比, NICE-SLAM[31]中提出的以体素网格为存储载体的方法有效地提高了运行速度, 但同时增加了运行时的显存占用; 基于 NICE-SLAM[31]和 ESLAM[32]采用多尺度三平面网格存储特征, 同时实现了显存占用降低和建图速度提高; Co-SLAM 则应用 Hash 编码体素网格降低显存占用, 同时引入位置编码加速渲染采样点的定位; Instant-NGP[47]建图方法出现后, Orbeez-SLAM 将其与 ORB-SLAM2[11]的前端相结合, 采用 CUDA 计算机语言编写 SLAM 系统, 实现了较高的建图速度和较低的显存占用.

6　结语与展望

结合隐式建图的视觉 SLAM 对观测区域具有一定补全预测能力且通常能产生更好的渲染效果, 近年来已经成为该领域的研究热点. 在视觉 SLAM 的建图环节, 存储载体的选择会影响跟踪和建图的精度、渲染质量和运行时间. 早期, iMAP[29]等方法使用 MLP 参数作为存储载体, 可实现紧凑且连续的场景建模, 但由于每次需要更新 MLP 全部参数, 系统难以进行实时重建. 为了解决这类问题, 后续研究大多数采用体素网格作为存储载体, 每次只需优化部分参数, 通过减少参数量提高运行速度; 将稀疏八叉树和位置编码等方法引入体素网格, 通过加速渲染采样点的定位进一步提高运行速度. 但是, 受到分辨率的限制, 导致目前体素网格隐式建图 SLAM 对精细细节的重建能力仍有限.

为了进一步提高隐式建图 SLAM 的性能, 研究者对各个环节进行改进: (1) 引入其他 SLAM 系统已采用的解决方案, 如引入不确定性提高建图鲁棒性、引入神经网络为 RGB 系统补充深度信息、通过特征匹配或共视性判断回环检测等; (2) 将改进后的隐式建图方法融入 SLAM 系统中, 如将大规模场景划分成子地图并分别用 MLP 进行建图的方法, 以及物体级语义 SLAM 的建立; (3) 将前 2 种方法进行结合, 如在应用于大规模场景时, 结合子地图划分方法和 SLAM 本身的关键帧策略, MIPS-Fusion[38]提出了根据锚定关键帧建立子地图的方法. 除了各个环节内部的改进, 研究者还提出将经典 SLAM 系统(如 ORB-SLAM 系列[10-12])的前端与隐式建图方法相结合[60,62], 以获得更精确的相机位姿跟踪结果和更高质量的隐式建图渲染结果. 当前, 不同环节的各类改进方法的协调和融合, 使得隐式建图 SLAM 已经能够达到较好的性能.

然而, 隐式建图 SLAM 仍在许多方面存在不足. (1) 在 SLAM 组成环节方面, 当前隐式建图视觉 SLAM 系统的回环检测主要依赖于较成熟的视

觉 SLAM 系统的跟踪组件, 适合于隐式建图方式的回环检测方法仍然缺乏, 需要进一步探索研究. (2) 在跟踪建图性能方面, 主要依赖 MLP 的隐式建图视觉 SLAM 系统面临着过度平滑的问题; 地图遗忘问题仍然严重, 虽然采取建立子地图和重播历史关键帧等方法可以有效地缓解这些问题, 但随着建图规模扩大, 可能会因达到上限导致无法继续扩展; 在训练过程中, 光线投射式体渲染和反向迭代优化产生了无法规避的的高计算量, 限制了 SLAM 系统运行速度的进一步提升. (3) 在特定应用场景方面, 因为运动模糊和深度噪声等影响相机跟踪和建图精度, 动态场景尤其是高度动态场景下的隐式建图 SLAM 系统仍然缺乏, 且对被动态物体遮挡的静态环境进行信息补充比较困难, 通常会导致重建场景中出现缺失和孔洞. 如何实现动态场景下的高性能的隐式建图视觉 SLAM 是亟待解决的问题. 此外, 渲染和反向迭代优化都需要较大的计算量, SLAM 系统的显存占用要求和计算资源要求使其难以迁移至移动端应用, 因此系统的存储轻量化和计算轻量化是改进的重要方向.

近期, 出现了渲染效果较好的基于三维高斯飞溅(3D gaussian splatting, 3DGS)[86]的显式建图方法, 随之出现了结合 3DGS 建图的视觉 SLAM 系统[87-88]. 与大多数隐式建图采用的光线投射式体渲染相比, 3DGS 建图直接通过对 3DGS 建图基元的可微光栅化实现视图渲染, 显式的建图形式和直接的梯度传递大大加速了渲染过程, 赋予了 3DGS 建图较高的训练速度, 使得结合 3DGS 的视觉 SLAM 系统具有较快的运行速度. 当前, 3DGS 视觉 SLAM 逐渐获得更多关注, 但其本身亦有一定局限性, 如初始化情况会对运行结果产生重大影响, 较难实现大规模场景的扩展建图, 未观测区域 3DGS 建图相对不可控, 会在渲染视图中产生伪影; 并且与隐式建图视觉 SLAM 系统相比, 3DGS 视觉 SLAM 系统缺乏直接的多边形网格提取算法, 可能会对地图重建结果产生影响. 隐式建图视觉 SLAM 系统和 3DGS 视觉 SLAM 系统各有优劣, 需根据实际情况作相应的选择, 或对二者进行改进和融合以进行优势互补, 实现视觉 SLAM 系统性能的进一步提高.

综上, 本文首先介绍视觉 SLAM 中常用的隐式建图和基于存储载体的建图分类方法; 然后基于 SLAM 系统的建图方法改进、前端改进、回环检测补充以及特定场景应用等, 分类说明了各种结合隐式建图的视觉 SLAM 技术; 再针对隐式建图视觉 SLAM 系统的跟踪建图结果和运行性能进行对比和分析; 最后总结并分析了该研究方向当前仍存在的挑战和未来的发展趋势.

参考文献(References):

[1] Liu Haomin, Zhang Guofeng, Bao Hujun. A survey of monocular simultaneous localization and mapping[J]. Journal of Computer-Aided Design & Computer Graphics, 2016, 28(6): 855-868(in Chinese)
(刘浩敏, 章国锋, 鲍虎军. 基于单目视觉的同时定位与地图构建方法综述[J]. 计算机辅助设计与图形学学报, 2016, 28(6): 855-868)

[2] Abaspur Kazerouni I, Fitzgerald L, Dooly G, *et al*. A survey of state-of-the-art on visual SLAM[J]. Expert Systems with Applications, 2022, 205: Article No.117734

[3] Zhang Y, Wu Y Q, Tong K, *et al*. Review of visual simultaneous localization and mapping based on deep learning[J]. Remote Sensing, 2023, 15(11): Article No.2740

[4] Smith R C, Cheeseman P. On the representation and estimation of spatial uncertainty[J]. The International Journal of Robotics Research, 1986, 5(4): 56-68

[5] Davison A J, Reid I D, Molton N D, *et al*. MonoSLAM: real-time single camera SLAM[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2007, 29(6): 1052-1067

[6] Triggs B, McLauchlan P F, Hartley R I, *et al*. Bundle adjustment—a modern synthesis[C] //Proceedings of the International Workshop on Vision Algorithms Corfu. Heidelberg: Springer, 1999: 298-372

[7] Sivic J, Zisserman A. Video Google: a text retrieval approach to object matching in videos[C] //Proceedings of the 9th IEEE International Conference on Computer Vision. Los Alamitos: IEEE Computer Society Press, 2003: 1470-1477

[8] Cummins M, Newman P. Appearance-only SLAM at large scale with FAB-MAP 2.0[J]. The International Journal of Robotics Research, 2011, 30(9): 1100-1123

[9] Klein G, Murray D. Murray. Parallel tracking and mapping for small AR workspaces[C] //Proceedings of the 6th IEEE and ACM International Symposium on Mixed and Augmented Reality. Los Alamitos: IEEE Computer Society Press, 2007: 225-234

[10] Mur-Artal R, Montiel J M M, Tardós J D. ORB-SLAM: a versatile and accurate monocular SLAM system[J]. 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[J]. IEEE Transactions on Robotics, 2017, 33(5): 1255-1262

[12] Campos C, Elvira R, Rodríguez J J G, *et al*. ORB-SLAM3: an accurate open-source library for visual, visual-inertial, and multimap SLAM[J]. IEEE Transactions on Robotics, 2021, 37(6): 1874-1890

[13] Newcombe R A, Lovegrove S J, Davison A J. Davison. DTAM: dense tracking and mapping in real-time[C] //Proceedings of the International Conference on Computer Vision. Los Alamitos: IEEE Computer Society Press, 2011: 2320-2327

[14] Engel J, Schöps T, Cremers D. LSD-SLAM: large-scale direct monocular SLAM[C] //Proceedings of the 13th European Con-

ference on Computer Vision. Heidelberg: Springer, 2014: 834-849

[15] Tateno K, Tombari F, Laina I, *et al*. CNN-SLAM: real-time dense monocular SLAM with learned depth prediction[C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2017: 6565-6574

[16] Bloesch M, Czarnowski J, Clark R, *et al*. CodeSLAM— learning a compact, optimisable representation for dense visual SLAM[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2018: 2560-2568

[17] Teed Z, Deng J. DROID-SLAM: deep visual SLAM for monocular, stereo, and RGB-D cameras[C] //Proceedings of the 35th International Conference on Neural Information Processing Systems. Red Hook: Curran Associates Inc, 2021: Article No.1266

[18] Appel A. Some techniques for shading machine renderings of solids[C] //Proceedings of the Spring Joint Computer Conference. New York: ACM Press, 1968: 37-45

[19] Curless B, Levoy M. A volumetric method for building complex models from range images[C] //Proceedings of the 23rd Annual Conference on Computer Graphics and Interactive Techniques. New York: ACM Press, 1996: 303-312

[20] Mescheder L, Oechsle M, Niemeyer M, *et al*. Occupancy Networks: learning 3D reconstruction in function space[C] // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2019: 4455-4465

[21] Azinović D, Martin-Brualla R, Goldman D B, *et al*. Neural RGB-D surface reconstruction[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2022: 6280-6291

[22] Yariv L, Gu J T, Kasten Y, *et al*. Volume rendering of neural implicit surfaces[C] //Proceedings of the 35th International Conference on Neural Information Processing Systems. Red Hook: Curran Associates Inc, 2021: Article No.367

[23] OrEl R, Luo X, Shan M Y, *et al*. StyleSDF: high-resolution 3D-consistent image and geometry generation[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2022: 13493-13503

[24] Peng S Y, Niemeyer M, Mescheder L, *et al*. Convolutional occupancy networks[C] //Proceedings of the 16th European Conference on Computer Vision. Heidelberg: Springer, 2020: 523-540

[25] Jiang C Y, Sud A, Makadia A, *et al*. Local implicit grid representations for 3D scenes[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2020: 6000-6009

[26] Hu J R, Mao M, Bao H J, *et al*. CP-SLAM: collaborative neural point-based SLAM[C] //Proceedings of the 37th International Conference on Neural Information Processing Systems. Red Hook: Curran Associates Inc, 2023: Article No.1712

[27] Sandström E, Li Y, van Gool L, *et al*. Point-SLAM: dense neural point cloud-based SLAM[C] //Proceedings of the IEEE/CVF International Conference on Computer Vision. Los Alamitos: IEEE Computer Society Press, 2023: 18387-18398

[28] Mildenhall B, Srinivasan P P, Tancik M, *et al*. NeRF: representing scenes as neural radiance fields for view synthesis[J]. Communications of the ACM, 2021, 65(1): 99-106

[29] Sucar E, Liu S K, Ortiz J, *et al*. iMAP: implicit mapping and positioning in real-time[C] //Proceedings of the IEEE/CVF International Conference on Computer Vision. Los Alamitos: IEEE Computer Society Press, 2021: 6209-6218

[30] Liu S F, Zhu J K. SDFMAP: neural signed distance fields for mapping and positioning in real-time[C] //Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems. Los Alamitos: IEEE Computer Society Press, 2023: 9590-9597

[31] Zhu Z H, Peng S Y, Larsson V, *et al*. NICE-SLAM: neural implicit scalable encoding for SLAM[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2022: 12776-12786

[32] Johari M M, Carta C, Fleuret F. ESLAM: efficient dense SLAM system based on hybrid representation of signed distance fields[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2023: 17408-17419

[33] Yang X R, Li H, Zhai H J, *et al*. Vox-Fusion: dense tracking and mapping with voxel-based neural implicit representation[C] //Proceedings of the IEEE International Symposium on Mixed and Augmented Reality. Los Alamitos: IEEE Computer Society Press, 2022: 499-507

[34] Wang H Y, Wang J W, Agapito L. Co-SLAM: joint coordinate and sparse parametric encodings for neural real-time SLAM[C] // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2023: 13293-13302

[35] Teigen A L, Park Y, Stahl A, *et al*. RGB-D mapping and tracking in a plenoxel radiance field[C] //Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. Los Alamitos: IEEE Computer Society Press, 2024: 3330-3339

[36] Kruzhkov E, Savinykh A, Karpyshev P, *et al*. MeSLAM: memory efficient SLAM based on neural fields[C] //Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics. Los Alamitos: IEEE Computer Society Press, 2022: 430-435

[37] Xiang B C, Sun Y X, Xie Z Q, *et al*. NISB-Map: scalable mapping with neural implicit spatial block[J]. IEEE Robotics and Automation Letters, 2023, 8(8): 4761-4768

[38] Tang Y J, Zhang J Z, Yu Z N, *et al*. MIPS-Fusion: multi-implicit-submaps for scalable and robust online neural RGB-D reconstruction[J]. ACM Transactions on Graphics, 2023, 42(6): Article No.246

[39] Matsuki H, Tateno K, Niemeyer M, *et al*. NEWTON: neural view-centric mapping for on-the-fly large-scale SLAM[J]. IEEE Robotics and Automation Letters, 2024, 9(4): 3704-3711

[40] Yang X R, Ming Y H, Cui Z P, *et al*. FD-SLAM: 3-D reconstruction using features and dense matching[C] //Proceedings of the International Conference on Robotics and Automation. Los Alamitos: IEEE Computer Society Press, 2022: 8040-8046

[41] Liu S F, Zhu J K. Efficient map fusion for multiple implicit SLAM agents[J]. IEEE Transactions on Intelligent Vehicles, 2024, 9(1): 852-865

[42] Sandström E, Ta K, van Gool L, *et al*. UncLe-SLAM: uncertainty learning for dense neural SLAM[C] //Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops. Los Alamitos: IEEE Computer Society Press, 2023: 4539-4550

[43] Huang J H, Huang S S, Song H X, *et al*. DI-Fusion: online implicit 3D reconstruction with deep priors[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2021: 8928-8937

[44] Li M R, He J M, Wang Y Y, *et al*. End-to-end RGB-D SLAM with multi-MLPs dense neural implicit representations[J]. IEEE Robotics and Automation Letters, 2023, 8(11): 7138-7145

[45] Matsuki H, Sucar E, Laidow T, *et al*. iMODE: real-time incremental monocular dense mapping using neural field[C] //Proceedings of the IEEE International Conference on Robotics and Automation. Los Alamitos: IEEE Computer Society Press, 2023: 4171-4177

[46] Chan E R, Lin C Z, Chan M A, *et al*. Efficient geometry-aware 3D generative adversarial networks[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2022: 16102-16112

[47] Müller T, Evans A, Schied C, *et al*. Instant neural graphics primitives with a multiresolution hash encoding[J]. ACM Transactions on Graphics, 2022, 41(4): Article No.102

[48] Li H, Yang X R, Zhai H J, *et al*. Vox-Surf: voxel-based implicit surface representation[J]. IEEE Transactions on Visualization and Computer Graphics, 2024, 30(3): 1743-1755

[49] Liu L J, Gu J T, Zaw Lin K, *et al*. Neural sparse voxel fields[C] //Proceedings of the 34th International Conference on Neural Information Processing Systems. Red Hook: Curran Associates Inc, 2020: Article No.1313

[50] Müller T, McWilliams B, Rousselle F, *et al*. Neural importance sampling[J]. ACM Transactions on Graphics, 2019, 38(5): Article No.145

[51] Fridovich-Keil S, Yu A, Tancik M, *et al*. Plenoxels: radiance fields without neural networks[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2022: 5491-5500

[52] Reiser C, Peng S Y, Liao Y Y, *et al*. KiloNeRF: speeding up neural radiance fields with thousands of tiny MLPs[C] //Proceedings of the IEEE/CVF International Conference on Computer Vision. Los Alamitos: IEEE Computer Society Press, 2021: 14315-14325

[53] Tancik M, Casser V, Yan X C, *et al*. Block-NeRF: scalable large scene neural view synthesis[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2022: 8238-8248

[54] Ming Y H, Ye W C, Calway A. Calway. iDF-SLAM: end-to-end RGB-D SLAM with neural implicit mapping and deep feature tracking[OL]. [2024-05-24]. https://arxiv.org/abs/ 2209.07919

[55] Lang R L, Fan Y, Chang Q. SVR-Net: a sparse voxelized recurrent network for robust monocular SLAM with direct TSDF mapping[J]. Sensors, 2023, 23(8): Article No.3942

[56] Li H, Gu X D, Yuan W H, *et al*. Dense RGB SLAM with neural implicit maps[OL]. [2024-05-24]. https://arxiv.org/abs/2301.08930

[57] Zhu Z H, Peng S Y, Larsson V, *et al*. NICER-SLAM: neural implicit scene encoding for RGB SLAM[C] //Proceedings of the International Conference on 3D Vision. Los Alamitos: IEEE Computer Society Press, 2024: 42-52

[58] Zhang W, Sun T C, Wang S, *et al*. HI-SLAM: monocular real-time dense mapping with hybrid implicit fields[J]. IEEE Robotics and Automation Letters, 2024, 9(2): 1548-1555

[59] Chung C M, Tseng Y C, Hsu Y C, *et al*. Orbeez-SLAM: a real-time monocular visual SLAM with ORB features and NeRF-realized mapping[C] //Proceedings of the IEEE International Conference on Robotics and Automation. Los Alamitos: IEEE Computer Society Press, 2023: 9400-9406

[60] Mao Y X, Yu X, Zhang Z Q, *et al*. NGEL-SLAM: neural implicit representation-based global Consistent low-latency SLAM System[C] //Proceedings of the IEEE International Conference on Robotics and Automation. Los Alamitos: IEEE Computer Society Press, 2024: 6952-6958

[61] Rosinol A, Leonard J J, Carlone L. NeRF-SLAM: real-time dense monocular SLAM with neural radiance fields[C] //Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems. Los Alamitos: IEEE Computer Society Press, 2023: 3437-3444

[62] Zhang Y M, Tosi F, Mattoccia S, *et al*. GO-SLAM: global optimization for consistent 3D instant reconstruction[C] //Proceedings of the IEEE/CVF International Conference on Computer Vision. Los Alamitos: IEEE Computer Society Press, 2023: 3704-3714

[63] Dai A, Chang A X, Savva M, *et al*. ScanNet: richly-annotated 3D reconstructions of indoor scenes[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2017: 2432-2443

[64] Teed Z, Deng J. RAFT: recurrent all-pairs field transforms for optical flow[C] //Proceedings of the 16th European Conference on Computer Vision. Heidelberg: Springer, 2020: 402-419

[65] Wang H C, Cao Y L, Wei X Y, *et al*. Structerf-SLAM: neural implicit representation SLAM for structural environments[J]. Computers & Graphics, 2024, 119: 103893

[66] Mazur K, Sucar E, Davison A J. Davison. Feature-realistic neural fusion for real-time, open Set scene understanding[C] //Proceedings of the IEEE International Conference on Robotics and Automation. Los Alamitos: IEEE Computer Society Press, 2023: 8201-8207

[67] Kong X, Liu S K, Taher M, *et al*. vMAP: vectorised object mapping for neural field SLAM[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2023: 952-961

[68] Han X, Liu H X, Ding Y C, *et al*. RO-MAP: real-time multi-object mapping with neural radiance fields[J]. IEEE Robotics and Automation Letters, 2023, 8(9): 5950-5957

[69] Ruan C Y, Zang Q Y, Zhang K H, *et al*. DN-SLAM: a visual SLAM with ORB features and NeRF mapping in dynamic environments[J]. IEEE Sensors Journal, 2024, 24(4): 5279-5287

[70] Xin Y Y, Zuo X X, Lu D Y, *et al*. SimpleMapping: real-time visual-inertial dense mapping with deep multi-view stereo[C] // Proceedings of the IEEE International Symposium on Mixed and Augmented Reality. Los Alamitos: IEEE Computer Society Press, 2023: 273-282

[71] Lisus D, Holmes C, Waslander S. Towards open world NeRF-based SLAM[C] //Proceedings of the 20th Conference on Robots and Vision. Los Alamitos: IEEE Computer Society Press, 2023: 37-44

[72] Liu X Y, Li Y J, Teng Y B, *et al*. Multi-modal neural radiance field for monocular dense SLAM with a light-weight ToF sensor[C] //Proceedings of the IEEE/CVF International Conference on Computer Vision. Los Alamitos: IEEE Computer Society Press, 2023: 1-11

[73] Felzenszwalb P F, Huttenlocher D P. Efficient graph-based image segmentation[J]. International Journal of Computer Vision, 2004, 59(2): 167-181

[74] Zhi S F, Laidlow T, Leutenegger S, *et al*. In-place scene labelling and understanding with implicit scene representation[C] // Proceedings of the IEEE/CVF International Conference on Computer Vision. Los Alamitos: IEEE Computer Society Press, 2021: 15818-15827

[75] Yang B B, Zhang Y D, Xu Y H, *et al*. Learning object-compositional neural radiance field for editable scene rendering[C] //Proceedings of the IEEE/CVF International Conference on Computer Vision. Los Alamitos: IEEE Computer Society Press, 2021: 13759-13768

[76] Nießner M, Zollhöfer M, Izadi S, *et al*. Real-time 3D reconstruction at scale using voxel hashing[J]. ACM Transactions on Graphics, 2013, 32(6): Article No.169

[77] Li Y J, Liu X Y, Dong W Q, *et al*. DELTAR: depth estimation from a light-weight ToF sensor and RGB image[C] //Proceedings of the 17th European Conference on Computer Vision. Heidelberg: Springer, 2022: 619-636

[78] Barron J T, Mildenhall B, Tancik M, *et al*. Mip-NeRF: a multiscale representation for anti-aliasing neural radiance fields[C] //Proceedings of the IEEE/CVF International Conference on Computer Vision. Los Alamitos: IEEE Computer Society Press, 2021: 5835-5844

[79] Zhang R, Isola P, Efros A A, *et al*. The unreasonable effectiveness of deep features as a perceptual metric[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2018: 586-595

[80] Sturm J, Engelhard N, Endres F, *et al*. A benchmark for the evaluation of RGB-D SLAM systems[C] //Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems. Los Alamitos: IEEE Computer Society Press, 2012: 573-580

[81] Straub J, Whelan T, Ma L N, *et al*. The Replica dataset: a digital replica of indoor spaces[OL]. [2024-05-24]. https://arxiv.org/ abs/1906.05797

[82] Yeshwanth C, Liu Y C, Nießner M, *et al*. ScanNet++: a high-fidelity dataset of 3D indoor scenes[C] //Proceedings of the IEEE/CVF International Conference on Computer Vision. Los Alamitos: IEEE Computer Society Press, 2023: 12-22

[83] Burri M, Nikolic J, Gohl P, *et al*. The EuRoC micro aerial vehicle datasets[J]. The International Journal of Robotics Research, 2016, 35(10): 1157-1163

[84] Dai A, Nießner M, Zollhöfer M, *et al*. BundleFusion: real-time globally consistent 3D reconstruction using on-the-fly surface reintegration[J]. ACM Transactions on Graphics, 2017, 36(3): Article No.24

[85] Lorensen W E, Cline H E. Marching cubes: a high resolution 3D surface construction algorithm[J]. ACM SIGGRAPH Computer Graphics, 1987, 21(4): 163-169

[86] Kerbl B, Kopanas G, Leimkuehler T, *et al*. 3D Gaussian splatting for real-time radiance field rendering[J]. ACM Transactions on Graphics, 2023, 42(4): Article No.139

[87] Keetha N, Karhade J, Murthy Jatavallabhula K, *et al*. SplaTAM: splat, track & map 3D Gaussians for dense RGB-D SLAM[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2024: 21357-21366

[88] Matsuki H, Murai R, Kelly P H J, *et al*. Gaussian splatting SLAM[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2024: 18039-18048