DOI: 10. 19650/j. cnki. cjsi. J2412684

基于稠密点云的神经辐射场 NeRF 在视觉 SLAM 建图任务中的应用研究*

陈久朋[1,2],陈治帆[1],伞红军[1,2],徐 贝[1]
(1. 昆明理工大学机电工程学院 昆明 650500; 2. 云南省先进装备智能制造技术重点实验室 昆明 650500)

摘 要:基于点云等显式场景表达的传统 SLAM 技术在精度和鲁棒性上已经较为成熟,但在地图纹理和语义信息还原方面存在不足。为了提高 SLAM 技术在纹理和语义信息获取方面的性能,本文将具有可微渲染能力的神经辐射场(NeRF)引入到传统视觉 SLAM 系统中,提出了一种新型视觉 SLAM 方法 DRM-SLAM。该方法使用 ORB-SLAM3 进行相机位姿估计,并结合关键帧的 RGB 信息和深度信息生成稠密点云,在动态体素网格的基础上,根据点云数据提供的三维几何信息在体素网格中进行采样减少 NeRF 调用多层感知机的频率。同时,该方法结合利用了多分辨率哈希编码和 CUDA 框架的 NeRF 实现,显著提升了 NeRF 的训练速度。在 TUM、WHU-RSVI、Replica 和 STAR 数据集上对本文提出的方法进行建图精度、完整度以及实时性测试的结果表明:DRM-SLAM 利用稠密点云和 NeRF 体渲染技术填补了点云中的空洞,保留了传统的 SLAM 方法在位姿估计精度上的优势,提升了地图的纹理和材质的连续性。DRM-SLAM 算法在 Replica 数据集上的帧率为 22. 3,该值远大于 NICE-SLAM、iMap 和 Co-SLAM 算法,证明了所提算法具有较高的实时性。在相同的场景下进行消融实验,基于稠密点云进行 NeRF 渲染比传统的 NeRF 的方法帧率提升了 3 倍,进一步证明了稠密点云可以加速 NeRF 收敛,充分展示了 DRM-SLAM 在地图重建方面的性能。

关键词: 移动机器人;DRM-SLAM;视觉 SLAM;稠密点云;神经辐射场
中图分类号: TH14 **文献标识码**: A **国家标准学科分类代码**: 460. 4020

Research on the application of NeRF based on dense point clouds in visual SLAM mapping tasks

Chen Jiupeng[1,2], Chen Zhifan[1], San Hongjun[1,2], Xu Bei[1]
(*1. Faculty of Mechanical and Electrical Engineering, Kunming University of Science and Technology, Kunming 650500, China;*
2. Key Laboratory of Advanced Equipment Intelligent Manufacturing Technology of Yunnan Province, Kunming 650500, China)

Abstract: Traditional SLAM technologies based on explicit scene representations, such as point clouds, have matured in accuracy and robustness but fall short in capturing the texture and semantic information of the map. To address this limitation, this paper introduces neural radiance fields (NeRF) with differentiable rendering capabilities into the traditional visual SLAM system, proposing a novel visual SLAM method: DRM-SLAM (dense radiance mapper-SLAM). This method uses ORB-SLAM3 for camera pose estimation and combines the RGB and depth information of keyframes to generate dense point clouds. By utilizing a dynamic voxel grid, the method samples within the grid according to the three-dimensional geometric information provided by the point cloud data, thereby reducing the frequency of NeRF calling the multilayer perceptron (MLP). Additionally, the method incorporates multi-resolution hash coding and the CUDA framework's NeRF implementation, significantly accelerating NeRF training speed. Tests on the TUM, WHU-RSVI, Replica, and STAR datasets demonstrate that DRM-SLAM effectively uses dense point clouds and NeRF volume rendering technology to fill gaps in point clouds, maintaining the pose estimation accuracy of traditional SLAM methods while enhancing texture and material continuity in the map. The DRM-SLAM algorithm achieves a frame rate of 22. 3 on the Replica dataset, which is significantly higher than NICE-SLAM, iMap, and Co SLAM algorithms, showcasing its high real-time performance. Ablation experiments in the same scenario show that NeRF rendering based on dense point clouds increases the frame rate threefold compared to traditional NeRF methods, further proving that

收稿日期:2024-04-03 Received Date:2024-04-03
*基金项目:云南省科技厅基础研发计划-青年基金(202301AU070059)项目资助

dense point clouds can accelerate NeRF convergence and demonstrating the effectiveness of DRM-SLAM in map reconstruction.
Keywords: mobile robots; DRM-SLAM; visual SLAM; dense point cloud; neural radiance fields

0 引　　言

同步定位和建图(simultaneous localization and mapping,SLAM)是一项使得自主移动设备在能够在完全陌生的三维环境中理解环境并完成既定任务的技术。目前,SLAM 技术在定位方面已取得显著进展,建图领域仍待深入研究。在传统 SLAM 建图中,地图主要分为稀疏和稠密两种类型。稠密视觉 SLAM 常用的地图表示形式包括点云[1-2]、面元[3]、符号距离场[4]和体素网格[5-7]等。这些技术虽已在实际应用中展现成效,但仍存在局限性,如点云等显式地图能直接表达场景的三维几何特征,却难以精确还原纹理和语义信息,且无法合成未观测到的新视角[8]。

神经辐射场[9](neural radiance fields,NeRF)技术的出现,推动了基于神经网络的隐式场景表达方法的发展[10-13]。这些基于 NeRF 的 SLAM 方法[14-15]利用可微渲染技术获取稠密光照信息,并结合输入图像构建损失函数,以生成高保真度的地图。在 NeRF 之前,符号距离场(signed distance field,SDF)是实现相对保真地图的一种方式。特别是,基于学习的 Tandem 算法[16]通过使用截断符号距离函数(truncated signed distance function,TSDF)来融合并生成稠密的 3D 地图,实现了地图的实时构建。然而,该方法中的深度估计模块需要预先训练,这与在陌生环境中进行实时建图的目标不符。相比之下,NeRF 在训练过程中无需深度监督,能够更准确地捕捉纹理和语义信息,有效解决了实时建图的挑战。随着 NeRF 技术的进步,iMAP[17]成为首个将 NeRF 应用于 SLAM 中作为地图表示的工作。它通过 NeRF 的光度损失进行反向传播,优化相机姿态。在 iMAP[17]的基础上,NICE-SLAM[18]进一步发展,引入了分层特征网格模块,扩大了场景处理的规模。然而,iMAP 和 NICE-SLAM 都仅通过神经网络优化相机姿态,没有融入视觉里程计(visual odometer,VO),这可能导致初始位姿估计的不精确性。

NeRF 技术通过在从像素发出的射线上执行多次采样并对每次采样点的渲染结果进行积分,能够生成高质量的场景表示。然而,这一过程的收敛速度相对较慢,难以满足实时训练的需求。为了解决这一问题,相关工作[19-24]中采用了多种策略来推进 NeRF 技术的实用化。这些策略包括在内部网络中通过连续学习减少信息遗忘、更好地捕获场景的语义信息和增强确定性,以及采用高效的多分辨率哈希编码和改进的损失函数等方法。这些改进显著提升了 NeRF 技术在实际应用中的性能和效率。此外,Point-SLAM[25-26]技术通过利用神经点云和执行带有特征插值的体积渲染,进一步提高了 3 D 重建的性能。尽管如此,Point-SLAM 的体积射线采样的效率仍然受到限制,这在一定程度上影响了其实时处理能力。

综上所述,尽管 NeRF 技术在处理速度上存在局限,但通过不断的技术创新和改进,研究者们已经取得了显著的进展,使得 NeRF 及其衍生技术如 Point-SLAM 在三维场景理解和重建方面展现出巨大的潜力。未来的研究将继续探索如何进一步提升这些技术的处理速度和效率,以满足实时应用的严格要求。

为了解决传统 SLAM 建图技术在纹理和语义信息重构和新视角合成的局限性,并提高基于 NeRF 技术的实时性能,本研究在 ORB-SLAM3[27]的基础上,融合了基于稠密点云的神经辐射场,创新性地提出了 DRM-SLAM(dense radiance mapper-SLAM)算法。DRM-SLAM 技术有效地结合了 NeRF 的体积渲染优势和传统显式场景表达的快速收敛与易于编辑的特点,实现了地图的快速构建和高精度纹理信息的准确提取。本研究的主要创新点包括:

1)通过将神经辐射场(NeRF)集成到传统视觉 SLAM 框架中,本研究克服了传统 SLAM 在地图纹理和语义信息重建方面的不足,从而提高了地图的保真度和实用性。

2)利用稠密点云为 NeRF 训练提供了三维几何先验,减少了对多层感知机(multilayer perceptron,MLP)的调用频率,从而显著加快了 NeRF 的训练速度,提升了地图的实时渲染能力。

3)本研究结合了 ORB-SLAM3 在位姿估计、地图扩展和编辑方面的优势,以及基于学习的 SLAM 方法在地图保真度方面的先进性,提出了创新的 DRM-SLAM 框架。

1　总体研究思路

本文采用经典的视觉里程计 ORB-SLAM3[27]来估计输入图像的位置和姿态。该方法利用关键帧的 RGB 信息和深度信息来生成初始的稠密点云,为后续处理提供了基础。进一步地,选取位于像素到相机光心连线上的点云作为 NeRF 的采样点,并根据这些点云动态构建八叉树体素网格。在包含点云的体素网格内进行采样。对于少数未能查询到点云的光线,采用 coarse-to-fine 的策略进行采样,以确保采样的全面性和准确性。采样结果随后被用于 Instant-ngp[28]进行 NeRF 的训练。Instant-

ngp[28]引入了多分辨率哈希编码和并行计算框架(compute unified device architecture,CUDA)[29],有效地提高了 NeRF 训练的效率,一定程度上解决了 NeRF 训练速度较慢的问题。本研究中采用 CPU 进行初始点云的创建,这一步骤为 NeRF 提供了丰富的三维几何先验,显著降低了在训练过程中对 MLP 的调用频率。同时,NeRF 的计算过程由 GPU 并行执行,并行计算的方式有利于提升系统的实时性。整体框架的设计如图 1 所示,展示了从图像输入到 NeRF 训练的完整流程。通过这种结合传统视觉 SLAM 和新兴 NeRF 技术的方法,DRM-SLAM 能够在保持实时性的同时,提升三维场景重建的精度和质量。

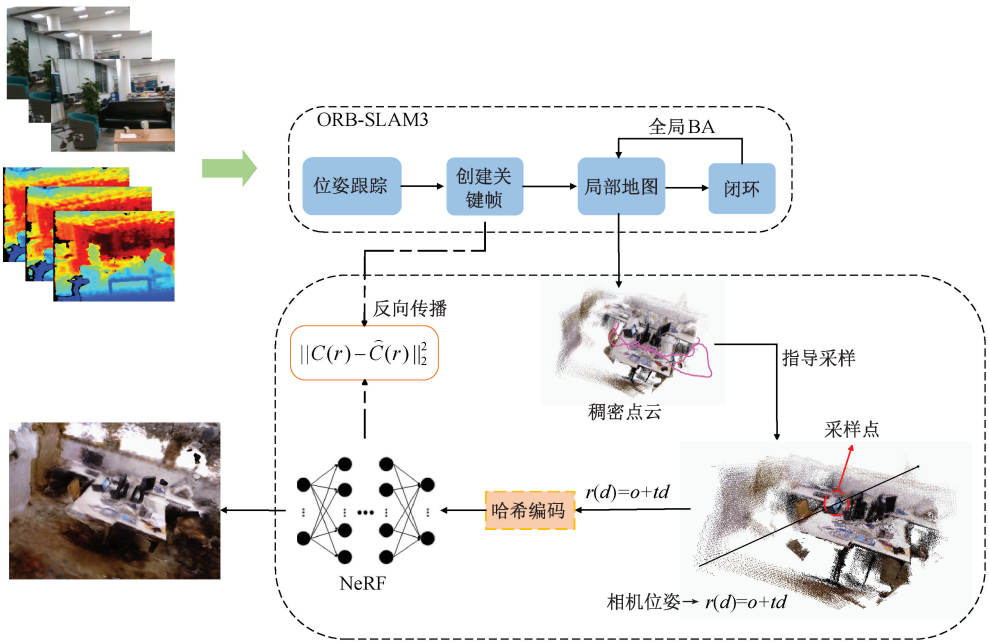


图 1 DRM-SLAM 整体流程图

Fig. 1 Overview of DRM-SLAM

1.1 ORB-SLAM3 系统框架介绍

ORB-SLAM3 由跟踪线程、局部建图线程和回环检测线程 3 大线程组成,如图 2 所示。跟踪线程负责处理连续的图像帧,通过 ORB 算法提取关键帧中的特征点,并在相邻帧之间进行特征匹配。这些匹配的特征点经过三角化过程生成三维地图点,从而构建或更新地图。在算法的初始化阶段,使用对极几何方法从第 1、2 帧图像中估计出初始位姿,并将第 1、2 帧设为参考关键帧。对后续进入的关键帧使用恒速度模型和光束法平差(bundle adjustment,BA)对位姿进行进一步的估计和优化。

局部建图线程在初始化地图的基础上,继续处理新传入的图像帧,进行特征匹配和三维地图点的生成。根据当前帧的质量、时间间隔和重定位历史等条件,决定是否创建新的关键帧。新创建的关键帧会与共视程度最高的关键帧进行地图点融合,以优化地图的质量和精度。

回环检测线程则负责检测环境中的闭环,即当相机返回到先前访问过的位置时,通过词袋模型进行闭环候选关键帧的检测和匹配。一旦检测到闭环,算法会通过相似变换群(sim3/SE3)变换和 BA 优化对地图进行矫正,以消除累积误差并提高地图的全局一致性。此外,算法在处理过程中还会不断优化和更新共视图和本质图,这两个图结构分别记录了关键帧之间的共视关系和高权重的共视关系,对于跟踪局部地图的扩大搜索范围、局部建图中关键帧之间的新建地图点、闭环检测和重定位检测等方面起着至关重要的作用。

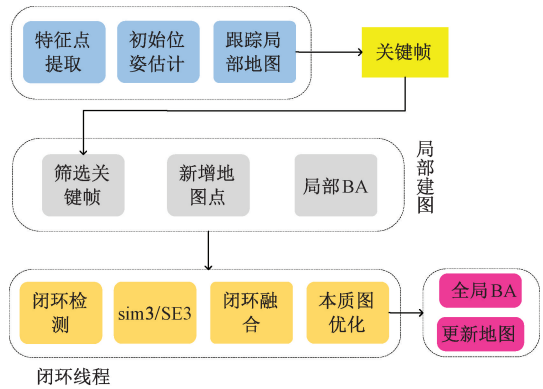


图 2 ORB-SLAM3 算法框架

Fig. 2 ORB-SLAM3 Algorithm Framework

1.2　实时位姿估计

采用深度相机捕获 RGB 图像及其对应的深度图像。对于初始阶段,当连续两帧图像中提取出足够数量的特征点(超过设定的阈值)时,首先从本质矩阵 E 中提取这两帧图像的初始相对位姿。利用相机的内参,精确地计算出这些特征点在三维空间中的坐标,从而构建初始的三维地图点。对于后续的每一帧图像,对于后续进入的每一帧图像,使用 Huber 函数[30]构造式(1)所示的最小二乘来最小化三维地图点的重投影误差获得相应的位姿:

$$\{\hat{\boldsymbol{T}}_{iw}\} = \underset{\hat{\boldsymbol{T}}_{iw}}{\operatorname{argmin}} \sum_i \left\| u_k^i - \frac{1}{s_k} M(\hat{\boldsymbol{T}}_{iw} \boldsymbol{P}_k^w) \right\|_{\Sigma}^2 \tag{1}$$

其中,$\boldsymbol{\Sigma}$ 是与特征点 $u_k{}^i$ 相关联的协方差矩阵,$\boldsymbol{T}_{iw}$ 表示第 i 帧图像在世界坐标系下的位姿变换矩阵;$u_k{}^i$ 表示第 i 帧图像中的第 k 二维特征点,$P_k{}^w$ 是世界坐标系下与特征点 $u_k{}^i$ 对应的三维地图点;S_k 表示三维地图点 $P_k{}^w$ 在 Z 轴上的深度,因为在三维向二维的映射中会丢失深度,因此像素坐标中归一化 Z 轴坐标为 1,M 表示相机模型如式(2) 所示。

$$M\left(\begin{bmatrix} X \\ Y \\ Z \end{bmatrix}\right) = \begin{bmatrix} f_x \dfrac{X}{Z} + c_x \\ f_y \dfrac{Y}{Z} + c_y \end{bmatrix} \tag{2}$$

其中,f_x, f_y, c_x, c_y 都是相机内参,在相机标定之后都是已知的。在优化结束时,使用 95% 的 χ^2 检验对重投影误差进行离群值过滤[31]。在跟踪阶段结束后,进入局建图阶段将进行 Bundle adjustment 优化,将局部地图中的关键帧位姿 $\hat{\boldsymbol{T}}_{iw}$ 和地图点 $\hat{\boldsymbol{P}}_k^w$ 一起进行优化:

$$\{\hat{\boldsymbol{T}}_{iw}, \hat{\boldsymbol{P}}_k^w\} = \underset{\hat{\boldsymbol{T}}_{iw}, \hat{\boldsymbol{P}}_k^w}{\operatorname{argmin}} \sum_i \sum_k \left\| u_k^i - \frac{1}{s_k} M(\hat{\boldsymbol{T}}_{iw} \hat{\boldsymbol{P}}_k^w) \right\|_{\Sigma}^2 \tag{3}$$

1.3　点云地图的拼接

根据关键帧位姿和相机模型(包含相机的内参),将相机坐标系下的三维点依次投影映射到世界坐标系中完成点云的拼接,具体操作如式(4)所示。

$$\boldsymbol{P}_k^w = s_k \boldsymbol{T}_{wi} M^{-1}(u_k^i) \tag{4}$$

拼接过程中每生成一些新的地图点或检测到闭环时,局部地图会不断被优化,且稠密程度将急剧增加。因为传感器本身会存在大量噪声,所以在完成点云拼接后对点云进行滤波。为了最大限度保留点云密度用于后续的 NeRF 体渲染,本文在点云拼接的过程中没有进行剔除和融合。稠密点云的建图效果如图 3 所示。

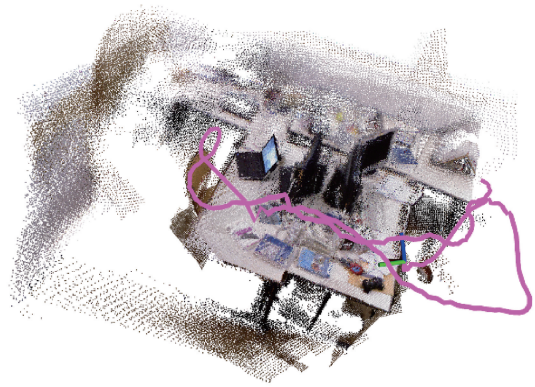


图 3　稠密点云地图

Fig. 3　Dense point cloud map

2　基于稠密点云的 NeRF 建图

2.1　神经辐射场建图策略

神经辐射场是一种通过神经网络模拟光线在三维空间中传播的技术。它通过计算从相机光心出发的射线上各点的颜色和密度并进行积分来渲染出每个像素的颜色。一个连续的场景可以表示为一个多层感知机 F。使用 x 表示相机在三维空间中的位置, $d = (\mathrm{d}x, \mathrm{d}y, \mathrm{d}z)$ 表示相机方向 (θ, ϕ) 的三维笛卡尔单位向量,c 表示网络输入的 (R, G, B),σ 表示体积密度,则整个神经辐射场可以表示为:

$$F(\boldsymbol{x}, \boldsymbol{d}) = (\boldsymbol{c}, \sigma) \tag{5}$$

通过给出一个关键帧 $(\boldsymbol{K}, [R\,|\,t])$ 和一个像素坐标 $[u, v]$,其中 $\boldsymbol{K}$ 表示内参矩阵,$[R\,|\,t]$ 表示相机位姿。可由式(6)可得从相机光心到任意像素 $[u, v]$ 发出的光线:

$$\boldsymbol{r}(\boldsymbol{d}) = (-\boldsymbol{R}^{\mathrm{T}}\boldsymbol{t} + \boldsymbol{d}\boldsymbol{R}^{\mathrm{T}}\boldsymbol{K}^{-1}[u, v, 1]^{\mathrm{T}}) \tag{6}$$

根据上述定义,对于任何世界坐标系下的射线 $r(t) = o + td$(o 为相机在世界坐标中的位置),在近端和远端区间 $[tn, tf]$ 内与视角相关的辐射亮度 $C(r)$ 可以通过体渲染式(7)积分得到,积分的内容包括光线上的透射率 $T(t)$、体积密度 σ_t 和 RGB 值:

$$C(\boldsymbol{r}) = \int_{t_n}^{t_f} T(t)\sigma(\boldsymbol{r}(t))c(\boldsymbol{r}(t), \boldsymbol{d})\,\mathrm{d}t \tag{7}$$

$$T(t) = \exp\left(-\int_{t_n}^{t} \sigma(\boldsymbol{r}(s))\,\mathrm{d}s\right) \tag{8}$$

对一条光线上的无穷多个点采样不切实际,因此基于神经网络实现 NeRF 过程需通过在区间 $[t_n, , t_f]$ 之间采样 N 个离散点 $t_i (i \in [1, N])$,并将每个采样点的渲染结果进行求和,以此代替在光线上的积分运算,求和运算如式(9)所示。

$$\hat{C}(r) = \sum_{i=1}^{N} T_i(1 - \exp(-\sigma_i(t_{i+1} - t_i)))c_i \tag{9}$$

$$T_i = \exp\left(-\sum_{j=1}^{i-1} \sigma_j(\mathrm{t}_{i+1} - \mathrm{t}_i)\right) \tag{10}$$

采用该方法可以创建连续且可微的辐射场,其渲染过程通过梯度下降进行优化。最后使用最小化预测颜色与采样点云对应像素的真实颜色之间的均方误差来拟合

一个 MLP,网络结构如图 4 所示。

$$L_C = \|C(r) - \hat{C}(r)\|_2^2 \tag{11}$$

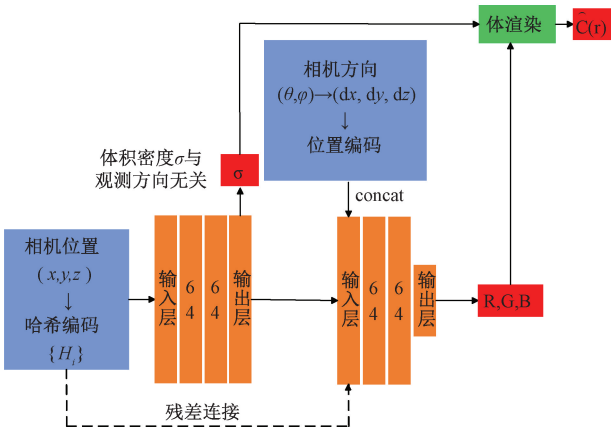


图 4 网络结构

Fig. 4 Network architecture

2.2 基于稠密点云的 NeRF 建图过程

传统的 NeRF 虽然能够渲染出连续的像素级三维场景,并能够合成输入图像中未包含的新视角,但是对密集光线上的大量采样点频繁调用并优化 MLP 导致基于像素的 NeRF 收敛速度太慢,难以满足 SLAM 任务对于实时性的要求。其次隐式状态下的场景信息被编码进 MLP 中无法直接用于拓展和编辑。采用点云作为载体不仅可以加速 NeRF 收敛过程,而且场景可拓展和编辑。本文作者借鉴文献[18]使用的体素网格,利用点云来指导 NeRF 采样。与文献[18]不同的是本文所提出的 DRM-SLAM 算法是在世界坐标系中根据点云提供的三维信息动态分配体素网格。根据点云数据的空间范围动态定义网格尺寸,创建一个三维网格框架(体素网格),这个框架由一系列均匀分布的立方体单元组成,每个单元都是一个体素,体素网格的每个顶点都位于点云数据的包围盒内,遍历点云中的每个点,根据点的空间位置将其映射到最近的体素单元中,计算点到体素网格中每个体素中心的距离,从而找到最近的体素。如果一个体素内存在点则认为该体素被占用,随着 SLAM 前端位姿估计模块不断的运行,点云将会被不断扩充,体素网格也将会随之动态拓展。动态拓展过程将对应于未观测场景区域的叶节点设置为空,当新的点云拼接进来后,为没有落在现有体素中的点分配新的体素。在后续的采样中只需跳过空体素,查询被占用的体素,并在其中进行采样。通过这种方式使得地图易于拓展且可编辑,同时大幅降低了内存消耗,八叉树体素网格的结构如图 5 所示。

为了实现点云在渲染过程中的准确对应,首先需要将点云从世界坐标系投影到相机坐标系中。在视觉里程计中估计的相机位姿误差以及通过单目、双目三角化或

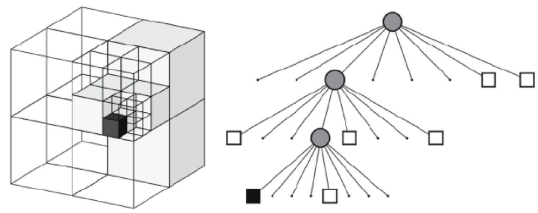


图 5 八叉树体素网格

Fig. 5 Octree voxel grid

深度相机获取的点云中的噪声,都可能导致点云无法精确地与投影光线对齐。为了解决这一问题,预先设定了一个阈值 ε,对于每个像素,在从相机光心发出的光线上,沿着半径为 ε 的范围搜索点云。由于场景表面更可能位于点云周围,因此在搜索过程中跳过未被点云占据的体素,仅在点云所在的体素内继续采样足够数量的点来渲染该光线,如图 6 所示。对于在阈值 ε 范围内仍然没有点云穿过的光线,采用 NeRF 中的 coarse-to-fine 两阶段采样策略。在粗略(coarse)阶段,沿着光线均匀地进行采样。而在精细(fine)阶段,利用粗略阶段得到的密度权重来指导采样,偏向于选择更接近潜在表面的位置。粗略阶段的采样采用分层均匀采样法,即将光线分割成等长的小段,并在每个小段内进行均匀采样。精细采样阶段则使用与粗略阶段相同的网络,并结合粗略采样阶段累积得到的权重 ω_i 式(12)所示来指导采样。这种结合预先定义阈值和两阶段采样策略的方法,能够在保持渲染效率的同时,提高渲染质量,确保点云与光线的精确对应。

$$\omega_i = T_i(1 - \exp(-\sigma_i(t_{i+1} - t_i))) \tag{12}$$

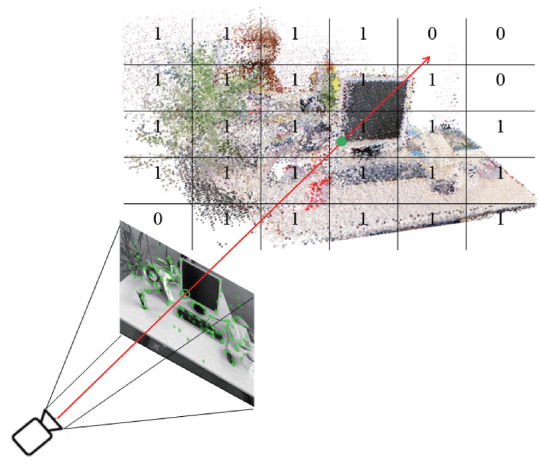


图 6 在有点云占据的体素中采样

Fig. 6 Sampling within voxels occupied by point clouds

3 实验验证与分析

为了验证所提 DRM-SLAM 算法的实时性和建图质

量,本文在公开的 4 个数据集中的多个视频序列上测试 DRM-SLAM 的性能,因为 DRM-SLAM 算法是基于 NeRF 的算法,因此需要与基于 NeRF 的其他 SLAM 算法进行对比验证。本文选取了 3 种较为典型基于 NeRF 的 SLAM 算法进行对比实验,其中 iMap[17] 是第 1 个将神经辐射场引入 SLAM 任务中的算法;NICE-SLAM 是在 NeRF 的基础上引入分层场景表示实现大型室内场景的重建的优秀算法。Co-SLAM[32] 引入了 one-blob 编码和哈希网格,大幅度提升了 NeRF 的训练速度。在完成数据集测试的基础上,将 DRM-SLAM 算法用于实际的工作场景中测试其在真实环境中稳定运行的能力。所有的测试均在相同的硬件条件下进行,并且所有的对比方法的参数配置均参考原始论文中给出的最佳配置。

3.1　公开数据集的选择

为了验证所提算法的有效性,并使实验不失一般性,本文分别在 TUM-RGB-D[33]、Replica[34]、WHU-RSVI[35] 和 STAR[36] 4 个公开的 RGB-D 数据集中挑选出了具有代表性的多个视频序列,这些视频序列包含了广泛的实际工作场景中的 RGB 图像、深度图像和相机轨迹,并在 SLAM 研究中被广泛使用。

3.2　硬件条件配置

本文使用了如图 7 所示的基于机器人操作系统(robot operating system,ROS)的移动机器人 TurtleBot2 作为实验平台,在 TurtleBot2 上架设了 RGB 相机 Intel RealSense D455,使用这台相机采集单目的 RGB 图像和 ROS 驱动程序实时输出的视差图作为输入 SLAM 系统的原始视频序列。使用一台 Laptop PC 对 TurtleBot2 进行控制,PC 搭载了 Intel i9-13900HX 处理器,主频为 5.40 GHz,和 NVIDIA RTX 3090ti GPU,配备了 16.00 GB 的运行内存,并运行 Ubuntu 20.04 操作系统。其中视觉传感器使用的是 Intel RealSense D455,该相机的操作范围为 0.6~6 m,视野深度 86×57,分辨率设置为 848×480,频率为 30 Hz。PC 与相机之间的连接采用的 USB3.1。TurtleBot2 的平移速度最高为 70 cm/s,最大旋转速度为 180°/s。

3.3　位姿估计精度测试

在 TUM-RGB-D、Replica、WHU-RSVI 和 STAR 4 个公开的 RGB-D 数据集中,TUM 数据集和 WHU-RSVI 数据集主要用于评估位姿估计精度。选取 TUM-RGB-D 数据集和 WHU-RSVI 数据集中的共 7 个视频序列作为测试素材,将 ORB-SLAM3、iMap、NICE-SLAM 以及 Co-SLAM 四种 SLAM 方法与本文所提的 DRM-SLAM 方法进行对比验证,测试结果如表 1 所示,表 1 中绝对轨迹误差数值可以反映定位精度的好坏。从表 1 中各指标数据可知,本

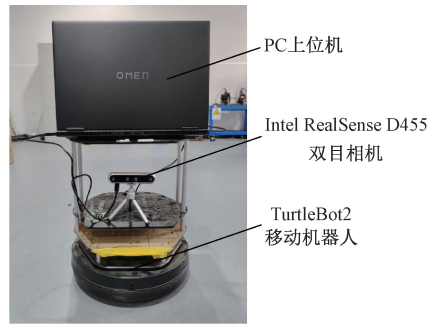


图 7　TurtleBot2 移动机器人

Fig. 7　TurtleBot2 mobile robot

文提出的 DRM-SLAM 算法在绝对轨迹误差上与 ORB-SLAM3 相差不大,这是因为该算法是基于 ORB-SLAM3 作为定位基础提出的。iMap、NICE-SLAM 和 Co-SLAM 算法相对于本文所提算法来说绝对轨迹误差较大,其中 iMap 算法和 NICE-SLAM 在定位精度上表现较为糟糕,Co-SLAM 次之。

3.4　实时性对比分析

目前基于 NeRF 的 SLAM 方法亟需解决的难题是其渲染速度难以匹配上相机实时采集图像的速度,为了提升渲染速度,本文提出的 DRM-SLAM 使用实时构建的点云地图作为先验,结合文献[28]提出的利用了多分辨率哈希编码和 CUDA 框架 Instant-ngp 加速神经辐射场的收敛。相比于其他基于学习的 SLAM 方法,DRM-SLAM 只需使用点云渲染地图,而省去了不断迭代更新 MLP 来估计位姿和地图的步骤。表 2 是在 STAR 和 TUM 两个数据集上部分视频序列的每秒传输帧数(frames per second,FPS)的数值大小。从表 2 可以看出,DRM-SLAM 相比于 iMap、NICE-SLAM、Co-SLAM 3 种常见的先进的 SLAM 算法具有较高的帧率,其中 NIC-SLAM,iMap 2 种算法在实时性上远不如本文所提算法,说明在没有牺牲重建质量的条件下 DRM-SLAM 性能更为优越。

表 3 为在上述视频序列上的消融实验结果,使用每次迭代所需的时间和 FPS 作为评价收敛速度的指标,从表 3 可以看出,DRM-SLAM 每次迭代所需时间和帧率均最小,实时性最高,这证明了使用稠密点云作为先验可加速 NeRF 的收敛。

3.5　重建质量对比分析

NeRF 将三维空间中的点和视角映射成颜色和体积密度,依据在这个过程中采样的若干点的体积密度来确定场景表面位置并最终输出渲染后的三维网格,使用本文方法在 TUM 数据集的 6 个视频序列上重建出的稠密点云地图和使用 NeRF 最终渲染出的三维网格地图的定性结果如图 8 所示,图 8(a)展示的是基于对应视频序列创建的稠密点云,图 8(b)展示的是基于对应稠密点云

表 1 绝对轨迹误差对比

Table 1 Comparison of absolute trajectory error

序号	视频序列名称	绝对轨迹误差(ape)	DRM-SLAM	ORB-SLAM3	iMap	NICE-SLAM	Co-SLAM
1	tum_fr1_desk	Max ape	1.647 2	1.832 9	4.765 6	4.397 9	3.937 9
		Mean ape	0.443 5	0.472 3	3.717 5	3.069 1	1.740 0
		Rmse	0.512 3	0.706 1	3.641 1	2.259 8	3.191 3
2	tum_fr2_xyz	Max ape	0.714 7	0.488 6	4.120 6	2.754 5	1.973 1
		Mean ape	0.383 2	0.283 2	2.997 6	2.915 5	1.392 0
		Rmse	0.221 0	0.254 7	3.609 6	2.459 6	1.333 2
3	tum_fr3_office	Max ape	1.219 4	0.932 6	3.786 1	3.401 7	2.952 2
		Mean ape	0.645 4	0.470 1	3.067 9	2.100 5	2.668 7
		Rmse	0.525 0	0.323 8	3.679 3	2.607 5	1.987 5
4	tum_fr2_slam3	Max ape	1.182 9	1.159 8	4.568 0	3.268 6	2.256 6
		Mean ape	0.638 1	0.433 4	3.286 1	3.240 3	1.595 4
		Rmse	0.475 3	0.503 4	3.500 3	3.267 2	2.251 1
5	tum_fr1_xyz	Max ape	0.325 4	0.294 9	3.297 4	2.878 7	2.633 4
		Mean ape	0.281 5	0.180 7	2.751 2	2.755 2	1.569 2
		Rmse	0.005 6	0.022 5	3.498 5	2.480 7	1.332 3
6	WHU_rsvi_t1_fast	Max ape	5.454 4	5.477 9	9.546 1	8.680 5	8.748 0
		Mean ape	2.615 4	2.294 8	5.765 1	4.105 6	4.925 9
		Rmse	2.743 6	2.753 5	5.451 7	4.407 6	5.049 9
7	WHU_rsvi_t2	Max ape	6.688 6	6.762 9	9.126 7	9.244 1	8.459 6
		Mean ape	2.808 1	2.829 3	6.329 5	5.676 1	4.447 6
		Rmse	3.274 4	3.390 0	6.430 2	4.801 7	4.799 5

表 2 在 STAR 和 TUM 数据上的帧率

Table 2 FPS on STAR and TUM datasets

(帧·s^{-1})

视频序列	DRM-SLAM	Co-SLAM	NICE SLAM	iMap
STAR_Handheld_normal	18.9	9.6	0.56	0.048
STAR_Wheeld_normal	19.5	10.4	0.69	0.033
fr1_desk	19.2	11.3	0.044	0.065
fr2_xyz	22.1	12.8	0.071	0.038
fr3_office3	18.8	9.1	0.056	0.048

表 3 消融实验

Table 3 Ablation study

指标	Instant-ngp	DRM-SLAM
每次迭代所需时间/ms	25.4	18.2
FPS	12.4	19.8

进行本文所提的 NeRF 渲染后得到的三维网格地图。从图 8 中可以明显看出,相比稠密点云,NeRF 渲染出的地图不仅具有连续的连续纹理和材质效果,并且能最大程度的填充点云中的空洞和背景,能够直接用于后续的语义分割等下游任务,这对于移动设备理解环境是至关重要。此外,通过神经辐射场生成的三维网格使得评估点云的指标如 Accuracy 和 Completion 可被应用于评估 NeRF 重建的质量。精度(Accuracy)表示对于每个渲染出来的离散点寻找在一定阈值内的真实值点,最终匹配上的平均距离即为精度,完整度(Completeness)是指对于每个真值中的三维点寻找在一定阈值内最近的渲染出来的点,最终可以匹配上的平均距离即为完整度。由于 TUM、WHU-RSVI 和 STAR 3 个数据集仅包含 RGB 和深度图像以及对应的真实轨迹,没有提供对应场景的三维信息,因此以上数据集仅能用于评估位姿估计的精度,不能用于定量评估地图重建质量。Replica 数据集提供了场景的真实三维网格数据,所以本文采用 Replica 数据集评估地图重建质量。所提 DRM-SLAM 使用具有深度测距功能的相机采集的 RGB 图像和对应深度图像作为输入,因此无需进行深度预测。

(a) 稠密点云
(a) Dense point cloud

(b) 本文方法
(b) Methodology of this paper

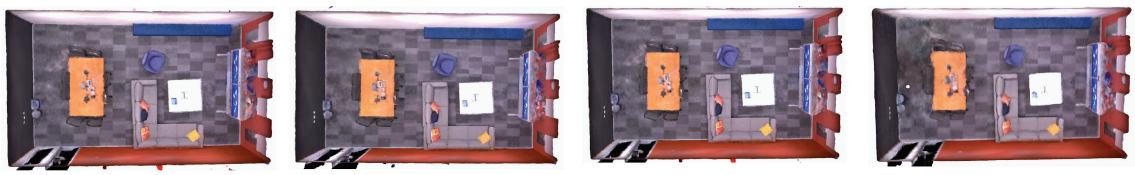
图 8　在 TUM 数据集上的空洞填充和纹理材质重建效果

Fig. 8　Hole filling and texture material reconstruction effects on the TUM dataset

采用基于 NeRF 的 iMap、NICE-SLAM、Co-SLAM 以及本文提出的 DRM-SLAM 四种 SLAM 算法在 Replica 数据集的 office_3 场景中的建图效果如图 9 所示。从图 9 中可以看出 4 种 SLAM 算法均能完成地图的重建。

4 种 SLAM 算法在 Replica 数据集的 room_0 场景中建图效果的细节对比如图 10 所示,矩形框中的部分是场景的局部放大图,从矩形框中可以看出 DRM-SLAM 能有效的还原了物体的纹理和材质(如沙发、窗户、电视以及光照效果),重建出比 iMap 和 NICE-SLAM 具有更清晰和纹理和形状。但是相比于目前较为建图效果较为优越的 Co-SLAM 算法,本文所提算法存在微弱的差距。这侧面表明本文所提的 DRM-SLAM 算法具有一定的建图精度和完整度,但在建图效果并不是最优的。对于一个算法来说,不可能对所有性能指标均能面面俱到,因此在获得位姿精度提升以及实时性提升的同时会对削弱其建图的能力。

为了全面评估本文提出的 DRM-SLAM 算法在重建精度、重建完整度、实时性 3 个关键性能指标上的表现,选择了 Replica 数据集中的 4 个具有代表性的场景,即 office_2,office_3,room_0 以及 room_1,作为测试环境。在这些场景中使用 DRM-SLAM 与 iMap、NICE-SLAM、Co-SLAM 等当前主流的 SLAM 算法进行了对比测试。各算法在上述 4 个场景中的性能测试结果的平均值已在表 4 中给出。由表 4 结果得出 DRM-SLAM 在 Replica 数据集上的帧率(FPS)为 22.3,该值远大于 NICE-SLAM、iMap 、Co-SLAM 算法,这表明所提算法实时性远高于其他算法,相比于其他方法均表现出了显著的优势。从建图效果上



(a) DRM-SLAM算法
(a) DRM-SLAM algoithm

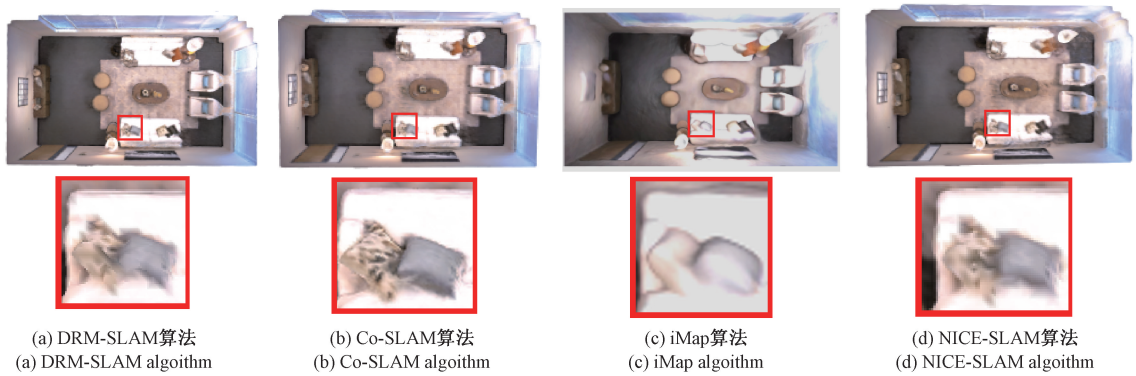
(b) Co-SLAM算法
(b) Co-SLAM algoithm

(c) NICE-SLAM算法
(c) NICE-SLAM algoithm

(d) iMap算法
(d) iMap algoithm

图 9　在 Replica 数据集 office3 序列上的建图效果

Fig. 9　The mapping effect on the office 3 sequence of the Replica datasets



(a) DRM-SLAM算法
(a) DRM-SLAM algoithm

(b) Co-SLAM算法
(b) Co-SLAM algoithm

(c) iMap算法
(c) iMap algoithm

(d) NICE-SLAM算法
(d) NICE-SLAM algoithm

图 10　建图效果细节对比

Fig. 10　Comparison of mapping details

看,DRM-SLAM 在重建精度和重建完整度上略低于目前主流的建图算法 Co-SLAM,但两者在数值上差异不大,这是因为在 DRM-SLAM 中仅使用视觉里程计创建的关键帧进行采样渲染,而 Co-SLAM 由于缺乏关键帧的筛选策略固定每隔 5 帧采样 2 048 条光线进行渲染,这大大增加了采样的频率,例如在 fr2_desk 视频序列中总共有 2 965 帧图像,DRM-SLAM 仅使用其中的 256 个关键帧进行采样,而 Co-SLAM 会使用 593 帧进行采样,这种高频采样策略是以牺牲算法实时性为代价换取了微弱的渲染质量,这对于需要实时运行的 SLAM 算法来说并不可取。

表 4 在 Replica 数据集上 4 种 SLAM 算法的对比分析

Table 4 Comparative analysis of four SLAM algorithms on the Replica dataset

算法	重建精度/cm	重建完整度/cm	FPS/(帧·s^{-1})
DRM-SLAM	2.45	2.74	22.3
NICE-SLAM	2.85	3.00	1.1
iMap	4.43	5.56	1.5
Co-SLAM	2.10	2.08	15.2

DRM-SLAM 的建图质量相比 iMap 和 NICE-SLAM 算法均有显著提升,证明了本文提出的 DRM-SLAM 算法用于建图的可行性。结合 3.3 节的位姿估计精度测试结果和 3.4 节的帧率测试结果可以得出本文提出的 DRM-SLAM 方法能够重建出高保真度三维环境地图,重建质量接近当前最先进的方法。综上分析可知,在满足建图要求的前提下,本文提出的 DRM-SLAM 算法具备更高的实时性和更好的位姿估计精度。因 DRM-SLAM 是基于稠密点云的神经辐射场应用算法,所提算法主要解决实时性和位姿估计两个方面的需求,因此本文所提算法具备较大的应用前景和价值。

3.6 真实环境中实验测试

上述实验证明了本文所提算法在公开数据集上的性能,其在定位和建图方面相比于传统的 SLAM 算法在定位精度、实时性和完整度上均有较大的性能提升。为了验证本文所提算法在真实场景中的建图效果,将 DRM-SLAM 系统部署到了图 7 所示的硬件平台上。使用 Intel RealSense D455 相机作为传感器,将采集的左目 RGB 图像和其 ROS 驱动程序生成的双目视差图作为输入,控制 TurtleBot2 自主移动机器人在 3D 空间内运行,其重建效果如图 11 所示,由图 11 可知,机器人能根据所提算法快速实现地图构建,证明了所提算法的有效性。

(a) 真实场景 (a) Real scene (b) 真实场景 (b) Mapping results

图 11 真实场景的建图效果

Fig. 11 Mapping results in real scenarios

4 结 论

为了提高 SLAM 技术在纹理和语义信息获取方面的性能,同时加快建图的效率,本文将具有可微渲染能力的神经辐射场(NeRF)引入到传统视觉视觉 SLAM 系统中,提出了一种新型视觉 SLAM 方法:DRM-SLAM。该方法继承了传统 SLAM 的快速收敛的优势,同时结合了 NeRF 可渲染高保真地图的特性,显著提升了建图任务的性能。

在公开的多种 RGB-D 数据集上进行了广泛测试并在真实环境中进行实验验证,结果显示本文所提的 DRM-SLAM 在多个方面表现出优越的性能。具体体现如下:DRM-SLAM 用于位姿估计精度上保留了传统的 SLAM 方法在定位上的优势,通过在 TUM 和 WHU-RSVI 数据集上测得的绝对轨迹误差(APE)显著优于其他基于学习的 SLAM 方法。在重建质量方面,DRM-SLAM 利用稠密点云和 NeRF 的可微渲染能力,有效地填充了点云中的空洞并提高了地图的纹理和材质连续性,在 Replica 数据集上的精度、完整度和完整率均接近或优于当前较为先进的 SLAM 方法。此外,DRM-SLAM 通过引入基于稠密点云的动态体素网格和多分辨率哈希编码,显著提高了 NeRF 训练的速度,实现了在 STAR 和 TUM 数据集上的较高的帧率,证明了其在所提算法的实时性。此外,消融实验结果进一步证实了稠密点云可以加速 NeRF 收敛。本文提出的 DRM-SLAM 在理论上具有一定的创新性,因其在位姿估计精度、重建质量以及实时性方面的优异表现,可为视觉 SLAM 建图提供一定的参考价值。未来工作可以在此基础上进一步探索如何提高系统的鲁棒性,扩展到更大规模的场景,并探索与其他传感器的融合可能性,以实现更加高效和可靠的自主导航和环境理解。

参考文献

[1] 佟国峰, 杨宇航, 彭浩, 等. 基于视觉语义与激光点云交融构建的 SLAM 算法[J]. 控制与决策, 2024, 39(1): 103-111.
TONG G F, YANG Y H, PENG H, et al. SLAM algorithm based on visual semantics and laser point cloud fusion construction[J]. Control and Decision, 2024,

39(1): 103-111.

[2] CAO Y P, KOBBELT L, HU SH M. Real-time highaccuracy three-dimensional reconstruction with consumer RGB-D cameras[J]. ACM Transactions on Graphics, 2018, 37(5): 171.

[3] SCHOPS T, SATTLER T, POLLEFEYS M. BAD SLAM: Bundle adjusted direct RGB-D SLAM[C]. 2019 IEEE/CVIP Conference on Computer Vision and Pattern Recognition, 2019: 134-144.

[4] NEWCOMBE R A, IZADI S, HILLIGES O, et al. Kinectfusion: Real-time dense surface mapping and tracking[C]. 2011 10th IEEE International Symposium on Mixed and Augmented Reality, 2011:127-136.

[5] HOLLAND L V, STOTKO P, KRUMPEN S, et al. Efficient 3D reconstruction, streaming and visualization of static and dynamic scene parts for multi-client live-telepresence in large-scale environments [C]. 2023 IEEE/CVF International Conference on Computer Vision, 2023: 4258-4272.

[6] STEINBRUCKER F, KERL C, CREMERS D. Large-scale multi-resolution surface reconstruction from RGB-D sequences[C]. 2013 IEEE International Conference on Computer Vision, 2013: 3264-3271.

[7] DURVASULA S, KIGURU R, MATHUR S, et al. VoxelCache: Accelerating online mapping in robotics and 3D reconstruction tasks [C]. Proceedings of the International Conference on Parallel Architectures and Compilation Techniques, 2022: 239-251.

[8] SARLIN P E, DUSMANU M, SCHONBERGER J L, et al. Lamar: Benchmarking localization and mapping for augmented reality[C]. European Conference on Computer Vision, 2022: 686-704.

[9] MILDENHALL B, SRINIVASAN P P, TANCIK M, et al. Nerf: Representing scenes as neural radiance fields for view synthesis[C]. Computer Vision-ECCV 2020, 2020: 405-421.

[10] AZINOVIC D, MARTIN-BRUALLA R, GOLDMAN D B, et al. Neural RGB-D surface reconstruction[C]. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022: 6280-6291.

[11] LIN CH H, MA W CH, TORRALBA A, et al. Barf: Bundle-adjusting neural radiance fields[C]. 2021 IEEE/CVF International Conference on Computer Vision, 2021: 5721-5731.

[12] WANG Z, WU SH ZH, XIE W D, et al. NeRF: Neural radiance fields without known camera parameters[J]. ArXiv preprint, arXiv:2102.07064v1,2021.

[13] LI ZH Q, NIKLAUS S, SNAVELY N, et al. Neural scene flow fields for space-time view synthesis of dynamic scenes[C]. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021: 6494-6504.

[14] SANDSTROM E, LI Y, GOOL L V, et al. Point-slam: Dense neural point cloud-based slam[C]. 2023 IEEE/CVF International Conference on Computer Vision, 2023: 18387-18398.

[15] ZHU Z H, PENG S Y, LARSSON V, et al. Nice-slam: Neural implicit scalable encoding for slam[C]. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022: 12776-12786.

[16] KOESTLER L, YANG N, ZELLER N, et al. Tandem: Tracking and dense mapping in real-time using deep multi-view stereo[C]. Proceedings of Machine Learning Research, 2021: 34-45.

[17] SUCAR E, LIU SH K, ORTIZ J, et al. iMAP: Implicit mapping and positioning in real-time[C]. 2021 IEEE/CVF Conference on Computer Vision, 2021: 6209-6218.

[18] ZHU Z H, PENG S Y, LARSSON V, et al. Nice-slam: Neural implicit scalable encoding for slam[C]. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022: 12776-12786.

[19] HAN X, LIU H, DING Y, et al. New findings from university of electronic science and technology of china in the area of robotics and automation reported (Ro-map: Real-time multi-object mapping with neural radiance fields)[J]. Robotics & Machine Learning Daily News, 2023,15:13-14.

[20] KONG X, LIU SH 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, 2023: 952-961.

[21] LI H, GU X D, YUAN W H, et al. Dense RGB slam with neural implicit maps[J]. ArXiv preprint, arXiv: 2301.08930v2,2023.

[22] MING Y H, YE W C, CALWAY A. Idf-slam: End-to-end RGB-D slam with neural implicit mapping and deep feature tracking [J]. ArXiv preprint, arXiv: 2209.07919v1,2022.

[23] ROSINOL A, LEONARD J J, CARLONE L. Nerfslam: Real-time dense monocular slam with neural radiance fields[J]. ArXiv preprint, arXiv:2210.13641v1,2022.

[24] SANDSTROM E, TA K, GOOL L V, et al. Uncle-slam: Uncertainty learning for dense neural slam[J]. 2023

IEEE/CVF International Conference on Computer Vision Workshops, 2023: 4539-4550.

[25] SANDSTROM E, LI Y, GOOL L V, et al. Point-slam: Dense neural point cloud-based slam[C]. 2023 IEEE/CVF International Conference on Computer Vision, 2023: 18387-18398.

[26] XU Q G, XU Z X, PHILIP J, et al. Point-nerf: Point-based neural radiance fields [C]. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022: 5438-5448.

[27] 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.

[28] MULLER T, EVANS A, SCHIED C, et al. Instant neural graphics primitives with a multiresolution hash encoding[J]. Acm Transactions on Graphics, 2022, 41(4): 1021-1015.

[29] MULLER T. Tiny CUDA neural network framework[J/OL]. [2021] https://github.com/nvlabs/tiny-cuda-nn.

[30] HARTLEY R, ZISSERMAN A. Multiple view geometry in computer vision [C]. Cambridge Univer-sity Press, 2000.

[31] FERRERA M, EUDES A, MORAS J, et al. OV2 SLAM: A fully online and versatile visual SLAM for real-time applications [J]. IEEE Robotics and Automation Letters, 2021, 6(2): 1399-1406.

[32] WANG H Y, WANG J W, AGAPITO L. Co-slam: Joint coordinate and sparse parametric encodings for neural real-time slam[J]. ArXiv preprint, ArXiv: 2304.14377, 2023.

[33] STURM J, ENGELHARD N, ENDRES F, et al. A benchmark for the evaluation of RGB-D SLAM systems[C]. 2012 IEEE/RSJ International Conference on Intelligent Robots and Systems, 2012:573-580.

[34] STRAUB J, WHELAN T, MA L N, et al. The replica dataset: A digital replica of indoor spaces[J]. ArXiv preprint, ArXiv:1906.05797v1, 2019.

[35] 武汉大学先进机器人与智能控制实验室. WHU-RSVI 数据集[OL]. 2024, http://aric.whu.edu.cn/公共资源/whu-rsvi-dataset/.
Advanced Robotics and Intelligent Control Laboratory of Wuhan University. WHU-RSVI dataset [OL]. 2024, http://aric.whu.edu.cn/公共资源/whu-rsvi-dataset/.

[36] Shanghai Tech Automation and Robotics Center. STAR dataset[OL]. 2024, Private Seafile (shanghaitech.edu.cn).

作者简介

陈久朋,2016 年于昆明理工大学获得学士学位,2018 年于昆明理工大学获得硕士学位,2021 年于昆明理工大学获得博士学位,现为昆明理工大学讲师,主要研究方向为机器人技术及应用、机械电子工程、机械设计。
E-mail:18314490225@163.com

Chen Jiupeng received his B.Sc. degree in 2016 from Kunming University of Science and Technology, received his M.Sc. degree in 2018 from Kunming University of Science and Technology, received his Ph.D. degree in 2021 from Kunming University of Science and Technology. Now he is lecturer in Kunming University of Science and Technology. His main research interests include robot technology and applications, mechanical, electronic engineering, mechanical design.

陈治帆,2022 年于昆明理工大学获得学士学位。现为昆明理工大学硕士研究生,主要研究方向为机器人技术及应用、机器视觉、机械设计。
E-mail:zhifanczf@163.com

Chen Zhifan received his B.Sc. degree in 2022 from Kunming University of Science and Technology. Now he is a M.Sc. candidate at Kunming University of Science and Technology. His main research interests include robot technology and applications, machine vision, and mechanical design.

伞红军(通信作者),2000 年于东北农业大学获得学士学位,2003 年于哈尔滨工业大学获得硕士学位,2009 年于哈尔滨工业大学获得博士学位,现为昆明理工大学副教授,主要研究方向为机器人技术及应用、机械设计。
E-mail:sanhjun@163.com

San Hongjun (Corresponding author) received his B.Sc. degree in 2000 from Northeast Agricultural University, received his M.Sc. degree in 2003 from Harbin Institute of Technology, received his Ph.D. degree in 2009 from Harbin Institute of Technology. Now he is associate professor in Kunming University of Science and Technology. His main research interests include robot technology and applications, and mechanical design.

徐贝,2022 年于江西理工大学获得学士学位,现为昆明理工大学硕士研究生,主要研究方向为机器人技术及应用、深度学习、机械设计。
E-mail:xxubei@163.com

Xu Bei received her B.Sc. degree in 2022 from Jiangxi University of Science and Technology. Now she is a M.Sc. candidate at Kunming University of Science and Technology. Her main research interests include robot technology and applications, deep learning, and mechanical design.