智能系统学报
CAAI TRANSACTIONS ON INTELLIGENT SYSTEMS

融合深度学习与神经隐式表征的视觉SLAM系统

张含笑, 邢向磊

引用本文:
张含笑, 邢向磊. 融合深度学习与神经隐式表征的视觉SLAM系统[J]. 智能系统学报, 2026, 21(1): 120-131.
ZHANG Hanxiao, XING Xianglei. Deep-learning-enhanced visual SLAM with neural implicit scene representation[J]. *CAAI Transactions on Intelligent Systems*, 2026, 21(1): 120-131.

在线阅读 View online: https://dx.doi.org/10.11992/tis.202505029

DOI: 10.11992/tis.202505029
网络出版地址: https://link.cnki.net/urlid/23.1538.tp.20251220.1153.002

融合深度学习与神经隐式表征的视觉 SLAM 系统

张含笑, 邢向磊
(哈尔滨工程大学 智能科学与工程学院, 黑龙江 哈尔滨 150001)

摘　要: 近年来, 神经辐射场在三维重建任务中展现出卓越性能。然而, 应用在视觉同时定位与地图构建 (simultaneous localization and mapping, SLAM) 中因缺乏全局优化机制容易导致系统定位精度不足以及重建失败。针对该问题, 本文提出一种融合深度学习位姿估计与神经隐式表征的视觉 SLAM 系统。通过稠密束调整层以及高效的全局优化机制对相机位姿和深度进行像素级的循环迭代, 并基于神经辐射场方法更新全局一致的隐式重建表面, 使得系统在精准定位的同时能够重建高保真场景, 并且在此基础上引入语言查询机制, 增强系统的交互能力。在 EuRoC 和 Replica 数据集上进行大量实验, 在不同的输入条件下, 分别与 3 类基准方法进行对比, 结果表明该系统在跟踪鲁棒性和重建精度方面相较于现有方法表现更优。本方法可为后续基于神经辐射场的视觉 SLAM 方法提供参考。

关键词: 神经辐射场; 视觉 SLAM; 回环检测; 位姿估计; 深度学习; 三维重建; 语义嵌入; 轨迹预测
中图分类号: TP391.41　　**文献标志码**: A　　**文章编号**: 1673−4785(2026)01−0120−12

中文引用格式: 张含笑, 邢向磊. 融合深度学习与神经隐式表征的视觉 SLAM 系统 [J]. 智能系统学报, 2026, 21(1): 120–131.
英文引用格式: ZHANG Hanxiao, XING Xianglei. Deep-learning-enhanced visual SLAM with neural implicit scene representation[J]. CAAI transactions on intelligent systems, 2026, 21(1): 120–131.

Deep-learning-enhanced visual SLAM with neural implicit scene representation

ZHANG Hanxiao, XING Xianglei
(College of Intelligent Science and Engineering, Harbin Engineering University, Harbin 150001, China)

Abstract: In recent years, neural radiation fields have demonstrated strong capability in high-fidelity three-dimensional scene reconstruction. However, visual simultaneous localization and mapping(SLAM) systems that employ neural radiance fields still face challenges in localization accuracy and the flexibility of explicit scene representation. To address these limitations, this work proposes a visual SLAM system that integrates deep-learning-based pose estimation with neural implicit scene representation. Through dense bundle adjustment layers and efficient global optimization mechanisms, the camera pose and depth are iteratively optimized at the pixel level, and a globally consistent implicit reconstruction surface is incrementally updated based on neural radiation fields, enabling the system to reconstruct high-fidelity scenes while achieving accurate localization. Furthermore, a language query mechanism was introduced to enhance the system's interactive capability. Extensive experiments were conducted on the EuRoC and Replica datasets, and the results were compared with those of three benchmark methods under different input conditions. The results showed that the proposed system outperformed existing methods in terms of tracking robustness and reconstruction accuracy, providing a reference for subsequent visual SLAM methods based on neural radiation fields.

Keywords: neural radiation field; visual SLAM; loop detection; pose estimation; deep learning; 3D reconstruction; semantic embedding; trajectory prediction

同时定位与地图构建 (simultaneous localization and mapping, SLAM) 技术是移动机器人实现自主定位和导航的关键[1]。随着计算机技术和人工智能的突破性发展, 对高保真三维物体和场景重建的需求不断增长。然而, 尽管三维重建技术取得了重大进展, 但在不牺牲精度和空间分辨率的情况下实时获得高质量表征仍然具有挑战性。与传统的视觉 SLAM 算法相比, 深度学习的引入能够在保持实时性的同时提高 SLAM 系统的定

收稿日期: 2025−05−28.　**网络出版日期**: 2025−12−23.
基金项目: 国家自然科学基金项目 (62076078, 61703119); 中央高校基本科研业务费项目 (3072024LJ0403).
通信作者: 邢向磊. E-mail: xingxl@hrbeu.edu.cn.

位和建图精度, 许多工作集中在针对特定的子问题上, 如特征检测[2-3]、特征匹配和异常值剔除[4-5]、定位[6-7] 和地图语义理解[8-9] 等。同时, 一些视觉 SLAM 系统利用深度学习方法[10-14] 进行单目三维重建, 但是通常使用点云表征方法进行重建, 在形状提取方面缺乏灵活性, 抑制了高保真重建。其中, DROID-SLAM[14] 完全依赖于深度学习模型, 在输入一系列连续的红绿蓝 (red green blue, RGB) 图像的条件下, 利用 RAFT(recurrent all-pairs field Transforms) 算法[15] 提取光流特征并形成 4D 关联体, 采用双线性插值从关联体中检索值, 将不同分辨率下的检索结果级联起来, 得到最终的特征向量, 充分利用了像素信息来捕捉场景中的全局结构和局部细节, 但是仅在相机跟踪结束后离线执行全局光束法平差, 很难消除累积误差。

为了提升对可见与遮挡区域的高质量渲染, 神经辐射场[16](neural radiation field, NeRF) 被应用到视觉 SLAM 系统中。iMAP[17] 是第一个使用 NeRF 进行建图和相机跟踪的统一 SLAM 管道, 根据输入的 RGB-D 图像, 利用隐式神经网络将三维坐标映射成颜色和体密度并进行渲染, 从而联合优化网络参数和相机位姿, 但是每次输入新的图像都需要更新整个多层感知机 (multilayer perceptron, MLP), 导致系统存在训练优化时间过长以及严重的遗忘问题。在此基础上, vMAP[18]、NICE-SLAM[19] 和 Co-SLAM[20] 等方法不断降低深度 MLP 查询的代价并加快建图速度, 实现了小规模场景的精确三维重建, 但是由于缺乏回环检测和全局优化机制, 在大场景的重建过程中容易崩溃, 并且需要依赖 RGB-D 相机的显式深度信息来实现辐射场的快速收敛。

针对上述问题, 本文提出了一种融合深度学习位姿估计与神经隐式建图的视觉 SLAM 系统模型, 实现相机精准定位的同时对三维场景进行高保真隐式重建。本文的主要贡献如下:

1) 提出一种基于深度学习的全局位姿优化系统。采用完全可微分的端到端架构, 结合经典视觉 SLAM 框架和深度神经网络的表达能力, 实现相机姿态和深度图的联合优化。

2) 设计一种基于共视帧图的高效对齐策略, 实现实时回环检测和全局结构校正, 降低内存需求, 提高系统在资源受限环境下的运行效率。

3) 提出一种结合神经辐射场的即时建图方法。根据全局优化系统中的相机姿态和深度信息, 实现全局一致的稠密三维重建, 同时引入语言查询机制, 拓展系统的交互能力与应用场景。

1　视觉 SLAM 系统整体结构

本文提出的基于神经隐式表征的深度学习视觉 SLAM 系统模型框架如图 1 所示, 以单目、双目或 RGB-D 视频流作为输入, 主要包括前端跟踪、后端跟踪以及即时建图 3 个部分。前端跟踪线程进行特征提取与关联选取出关键帧之后, 迭代更新当前关键帧的位姿和深度, 并执行回环检测; 在后端跟踪线程中, 对前端所有的历史关键帧进行全局束调整 (bundle adjustment, BA) 以优化位姿和深度; 最后, 建图线程基于神经辐射场方法构建高保真地图, 实时地适应全局优化位姿和深度的连续变化, 并根据提示文本提取出有效的语义信息。

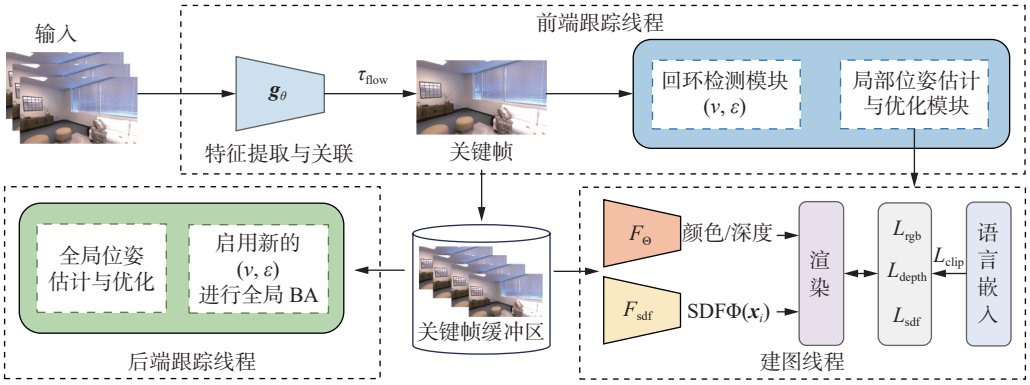


图 1　系统总体框架

Fig. 1　Overall framework of the system

2　基于深度学习的跟踪线程

2.1　特征提取与关联

对于输入的每帧图像 $\boldsymbol{I}_i \in \mathbf{R}^{H\times W\times 3}(i=1,2,\cdots,N)$ 使用特征网络 $\boldsymbol{g}_\theta(\cdot)\in\mathbf{R}^{H\times W\times D}$ 提取特征, 该网络由 6 个残差块和 3 个下采样层组成, 在输入图像分辨率的 1/8 处生成稠密特征图。计算输出的特征图 $\boldsymbol{g}_\theta(\boldsymbol{I}_i)$ 和 $\boldsymbol{g}_\theta(\boldsymbol{I}_j)$ 中所有点对之间的点积, 来衡量两帧图像之间的共视性, 形成 4 维共视特征图。

$$\boldsymbol{C}^{ij}_{u_1v_1u_2v_2}=\left\langle \boldsymbol{g}_\theta(\boldsymbol{I}_i)_{u_1v_1},\boldsymbol{g}_\theta(\boldsymbol{I}_j)_{u_2v_2}\right\rangle\in\mathbf{R}^{H\times W\times H\times W}$$

式中：(u,v)为图像中的像素坐标，H为图像的高度，W为图像的宽度。

为了进一步提取不同尺度下的共视信息，通过将共视特征图的最后两维进行平均池化，形成4层共视金字塔$\boldsymbol{C}^k\in\mathbf{R}^{H\times W\times H/2^k\times W/2^k}(k=1,2,3,4)$，在保持高分辨率信息的同时，有效地捕捉不同尺度的位移信息，从而更精确地恢复快速移动的小物体的运动。

此外，本文引入查找算子[15]，以一个$H\times W$的坐标网格作为输入，通过双线性插值在共视特征图中进行检索。将该查找算子用于共视金字塔的每一层共视特征图，并通过级联每层的结果计算最终的光流特征。如果平均光流大于预定义的阈值τ_{flow}，则将当前帧创建为一个新的关键帧，并添加到关键帧缓冲区中以便后续的操作。

2.2　位姿估计与优化

前端位姿估计和优化的核心部件为位姿估计与优化模块，如图2所示。

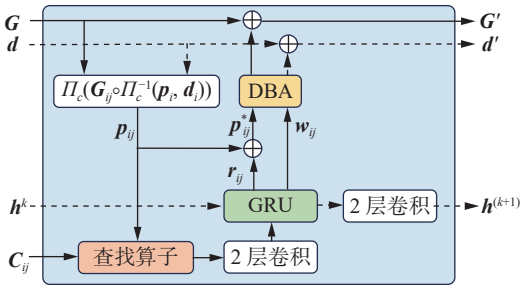


图 2　位姿估计与优化模块

Fig. 2　Pose estimation and optimization module

本文采用共视帧图$(\mathcal{V},\mathcal{E})$来表示帧之间的共视关系，帧图中的一条边$(i,j)\in\mathcal{E}$表示图像$\boldsymbol{I}_i$和$\boldsymbol{I}_j$之间有重叠的视场。在每次位姿更新之前，利用当前位姿和深度的估计值计算帧图中每条边对应的点云集合，并将其转化为像素坐标，得到每条边的对应场，具体公式为

$$\boldsymbol{p}_{ij}=\Pi_c(\boldsymbol{G}_{ij}\circ\Pi_c^{-1}(\boldsymbol{p}_i,\boldsymbol{d}_i)),\boldsymbol{G}_{ij}=\boldsymbol{G}_j\circ\boldsymbol{G}_i^{-1}$$

式中：$\boldsymbol{p}_i\in\mathbf{R}^{H\times W\times 2}$为图像$\boldsymbol{I}_i$的像素坐标，$\boldsymbol{d}_i$为图像$\boldsymbol{I}_i$当前的深度估计值，$\Pi_c$为世界坐标系映射到图像坐标系的相机模型，$\Pi_c^{-1}$为$\Pi_c$的反变换，$\boldsymbol{G}_{ij}$为图像$\boldsymbol{I}_i$和$\boldsymbol{I}_j$之间的位姿变换，$\boldsymbol{p}_{ij}$为使用估计的位姿和深度将图像$\boldsymbol{I}_i$中的像素$\boldsymbol{p}_i$映射到图像$\boldsymbol{I}_j$坐标系得到的对应坐标。

从而相机运动引起的光流特征为$\boldsymbol{p}_{ij}-\boldsymbol{p}_j$，并根据$\boldsymbol{p}_{ij}$从4维共视特征图中检索共视特征，将两种特征作为位姿估计与优化模块的输入。共视特征提供了$\boldsymbol{p}_{ij}$邻域内视觉相似性的信息，使网络能够学习对齐视觉相似的图像区域，光流特征则提供了额外的补充信息，增强网络的鲁棒性。在特征处理过程中，首先对光流特征和共视特征分别通过两个卷积层进行映射；接下来，输入到门控循环单元[21](gated recurrent unit, GRU)中，GRU通过两个额外的卷积层来映射生成新的隐藏状态$\boldsymbol{h}^{(k+1)}$，并将对应场$\boldsymbol{p}_{ij}$进行修正，生成与共视性相关的置信度权重$\boldsymbol{w}_{ij}\in\mathbf{R}^{H\times W\times 2}$以及修正的残差流场$\boldsymbol{r}_{ij}\in\mathbf{R}^{H\times W\times 2}$；最后，将残差流场与对应场相加，得到修正后的对应场为$\boldsymbol{p}^*_{ij}=\boldsymbol{r}_{ij}+\boldsymbol{p}_{ij}$。

位姿估计与优化模块利用修正后的对应场$\boldsymbol{p}^*_{ij}$和置信度权重$\boldsymbol{w}_{ij}$，通过可微的密集束调整层[14](dense bundle adjustment, DBA)来优化相机姿态$\boldsymbol{G}\in\text{SE}(3)$和关键帧的逆深度$\boldsymbol{d}\in\mathbf{R}^{H\times W}$。本文定义位姿更新的损失函数为

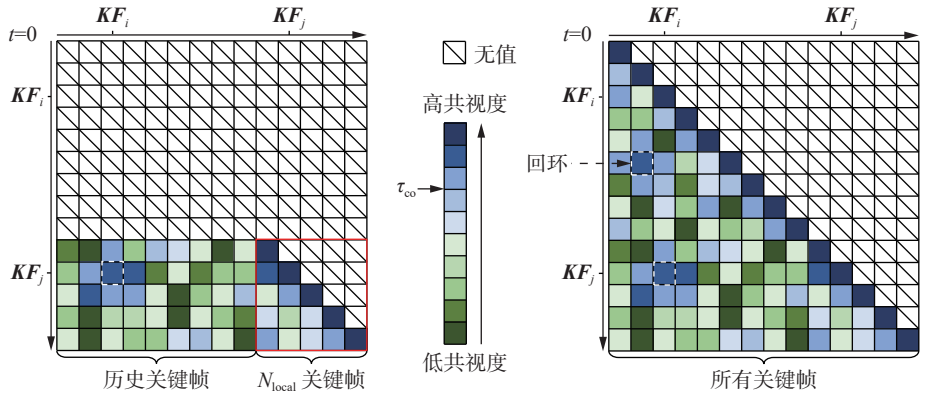
$$L=\sum_{(i,j)\in\mathcal{E}}\left\|\boldsymbol{p}^*_{ij}-\boldsymbol{p}_{ij}\right\|^2_{\Sigma_{ij}}$$

式中$\sum_{ij}=\text{diag}\boldsymbol{w}_{ij}$，$\|\cdot\|_\Sigma$为基于置信度权重$\boldsymbol{w}_{ij}$对误差项进行加权的马氏距离。

最终，更新相机姿态和深度的问题转化为解决一个非线性最小二乘问题，最小化损失函数，使得重投影点和位姿估计与优化模块预测的对应场$\boldsymbol{p}^*_{ij}$相匹配。本文采用局部参数化的方法损失函数进行线性化处理，并使用高斯–牛顿算法进行求解。

2.3　回环检测

为了能够实现实时的回环检测和全局姿态的校正，本文提出了一种有效的对齐策略。如图3所示，根据上述特征提取与关联模块中的关键帧选取方法，可以得到目前为止创建的关键帧集合$\{\boldsymbol{KF}_k\}_{k=1}^{N_{KF}}$以及关键帧的共视帧图$(\mathcal{V},\mathcal{E})$，选择最近的$N_{\text{local}}$个局部关键帧建立高共视度连接，并在局部窗口外，检测局部关键帧和历史关键帧之间的回环。首先，在N_{local}张局部关键帧和所有的N_{KF}张历史关键帧之间建立大小为$N_{\text{local}}\times N_{KF}$的共视帧图，如图3(a)所示；接下来，在关键帧对之间进行反投影计算获得每条边的平均光流，并过滤掉平均流大于τ_{co}的边，从而获得具有高共视度的关键帧对；最后，将保留下来的关键帧对建立边链接，形成回环检测的候选区域。为了避免冗余，一旦在关键帧图中添加了边链接$\boldsymbol{KF}_i\leftrightarrow\boldsymbol{KF}_j$，则抑制半径$r_{\text{local}}$内的所有可能边连接。根据上述步骤，将局部共视帧图中未探索的部分按照共视度进行降序采样边缘，并用半径r_{loop}抑制相邻边。为了接受一个回环候选，回环检测模块连续检测3个回环候选，如果它们的平均流小于τ_{co}，则判定为检测到一个回环。



(a) 局部共视帧图　　(b) 全局共视帧图

图 3　回环检测模块示意

Fig. 3　Schematic diagram of loop detection module

2.4　后端跟踪线程

在后端跟踪模块中同时优化整个历史关键帧的计算成本非常昂贵。为了解决这个问题, 本文将完整的 BA 优化在线运行在一个单独的线程中, 允许系统继续跟踪新的帧并进行回环检测。与前端的局部回环检测模块类似, 启动一个新的关键帧图, 并插入具有高共视度的关键帧对以及时间相邻的关键帧, 如图 3(b) 所示。当建立新的边时, 用半径 r_{global} 来抑制冗余的相邻边。由于最新关键帧的姿态已经在回环检测时进行了全局几何校正, 因此缓解了对全局 BA 的实时性要求。

3　基于神经辐射场的建图线程

3.1　关键帧选取

本文提出的基于神经辐射场的建图线程主要实现实时更新全局三维重建和语义查询功能。为了平衡全局一致性与实时性, 本文提出了一种高效的关键帧选择策略。在三维重建的每次更新开始之前, 建图线程首先对跟踪到的所有关键帧的位姿和深度进行快照, 以确保重建期间的几何一致性。然后, 在所有关键帧中选出用于重建更新的关键帧, 包括 3 个部分: 第 1 个部分是选取最新的 2 个关键帧和还没有经过映射重建的关键帧; 第 2 个部分是将所有关键帧按照当前和上一次更新状态之间的位姿差降序排列, 选择前 10 个关键帧; 第 3 个部分是通过分层抽样从所有可用的关键帧中选取 10 个关键帧, 以防止几何信息遗忘问题。

3.2　渲染

对于选取出的关键帧, 根据 2.2 节中的位姿估计与优化模块, 可以获得每个关键帧的图像 $\boldsymbol{I}$、姿态 $\boldsymbol{G}$ 和深度 $\boldsymbol{d}$, 在每张关键帧上随机采样 M 个像素点进行训练。对采样的像素从相机光心 $\boldsymbol{o}$ 发射射线:

$$\boldsymbol{r}(t)=\boldsymbol{o}+t\boldsymbol{v}$$

式中: $\boldsymbol{r}(t)$ 为第 t 条射线, $\boldsymbol{v}$ 为观测方向。

沿射线采样 $N_{\text{ray}}=N_{\text{start}}+N_{\text{imp}}$ 个点, 其中 N_{start} 为使用分层抽样进行采样的点数, N_{imp} 为在深度值附近采样的点数。对射线上每个采样的 3D 点 $t_i(i=1,2,\cdots,N_{\text{ray}})$ 进行哈希编码[22], 将得到高维的位置信息 $\boldsymbol{x}_i$ 输入到预测网络中获取采样点的对应信息并进行渲染。

3.2.1　颜色和深度信息

将采样点的位置信息 $\boldsymbol{x}$ 和观测方向 $\boldsymbol{v}=(\theta,\phi)$ 输入到预测网络 F_Θ, 其中 F 由 2 层 MLP 网络构成, Θ 为网络权重。F_Θ 输出该采样点的颜色值 $\boldsymbol{c}=(r,g,b)$ 和密度值 σ, 则从光线起点到采样点 t_i 之间的累计透射率可以用公式表示为

$$T_i=\exp\left(-\sum_{j=1}^{i-1}\sigma_j\delta_j\right)$$

式中: σ_i 为采样点 t_i 的密度, $\delta_i=t_{i+1}-t_i$ 为相邻采样点之间的距离。

通过体渲染可以计算出每条射线所对应的颜色值和深度值, 即像素级的颜色和深度的预测结果为

$$\hat{\boldsymbol{I}}=\sum_{i=1}^{N_{\text{ray}}}T_i(1-\exp(-\sigma_i\delta_i))\boldsymbol{c}_i$$

$$\hat{\boldsymbol{d}}=\sum_{i=1}^{N_{\text{ray}}}T_i(1-\exp(-\sigma_i\delta_i))t_i$$

3.2.2　SDF 信息

为了监督重建表面, 本文模型使用单层 MLP 构成预测网络 F_{sdf} 来获得采样点预测的符号距离函数 $\Phi(\boldsymbol{x}_i)$, 并对预测的 SDF 进行正则化约束。为了获得近似采样点真实的 SDF, 本文计算采样点 t_i 到关键帧深度 $\boldsymbol{d}$ 之间的距离, 公式表示为

$$b(\boldsymbol{x}_i)=d_m-t_i$$

式中 d_m 为在真值图像上采样的像素点对应的深

度。则符号距离函数需要满足$|\Phi(\boldsymbol{x}_i)| \leqslant |b(\boldsymbol{x}_i)|, \forall \boldsymbol{x}_i$。

3.2.3　语言嵌入信息

为了将语言信息嵌入到三维场景中，采用和神经辐射场渲染三维场景类似的处理方式。以采样点为中心定义一个物理尺度，构建出采样点对应的立方体，公式表示为

$$s(t_i) = s_{\text{init}} \times f / t_i$$

式中：s_{init}为在像素平面上固定的初始尺度，f为相机焦距。

将编码后的位置信息和上述定义的物理尺度$s(t_i)$输入到语言嵌入网络F_{lang}中，从而获取到该采样点的语言嵌入信息$\boldsymbol{\xi}_i \in \mathbf{R}^d$。利用神经辐射场输出的密度值对语言嵌入信息进行渲染。

$$\boldsymbol{\xi} = \sum_{i=1}^{N_{\text{ray}}} T_i (1 - \exp(-\sigma_i \delta_i)) \boldsymbol{\xi}_i$$

为了确保特征空间的一致性，对渲染后的语言嵌入信息进行归一化处理，从而得到最终穿过该射线的像素所对应的语言嵌入信息为

$$\hat{\boldsymbol{\xi}} = \frac{\boldsymbol{\xi}}{\|\boldsymbol{\xi}\|}$$

3.3　损失函数

为了优化 3D 渲染网络，以关键帧图像$\boldsymbol{I}$和深度$\boldsymbol{d}$作为真值，在图像上选取M个像素构建 RGB 损失L_{rgb}和深度损失L_{depth}。

$$L_{\text{rgb}} = \frac{1}{M} \sum_{m=1}^{M} |\hat{\boldsymbol{I}}_m - \boldsymbol{I}_m|$$

$$L_{\text{depth}} = \frac{1}{M} \sum_{m=1}^{M} |\hat{\boldsymbol{d}}_m - \boldsymbol{d}_m|$$

另外，对 SDF 值进行进一步约束，由于满足$|\Phi(\boldsymbol{x}_i)| \leqslant |b(\boldsymbol{x}_i)|$，则对于接近重建表面的点，设置损失函数为

$$L_{\text{near}} = |\Phi(\boldsymbol{x}_i) - b(\boldsymbol{x}_i)|$$

而对于其他自由空间上的点，定义一个较为宽松的损失函数为

$$L_{\text{free}} = \max\left(\mathrm{e}^{-\beta \Phi(\boldsymbol{x}_i)} - 1, \Phi(\boldsymbol{x}_i) - b(\boldsymbol{x}_i), 0\right)$$

式中β为超参数，当预测的$\Phi(\boldsymbol{x}_i)$在自由空间中为负时，L_{free}起约束作用。

总体上 SDF 损失可以定义为

$$L_{\text{sdf}} = \frac{1}{M N_{\text{ray}}} \sum_{m,i} \begin{cases} L_{\text{near}}, & \text{if } |b(\boldsymbol{x}_i)| \leqslant \tau_{\text{trunc}} \\ L_{\text{free}}, & \text{其他} \end{cases}$$

式中τ_{trunc}为超参数，表示截断阈值。

为了对渲染的语言嵌入信息进行监督，本文通过对比性语言–图像预训练 (contrastive language-image pre-training, CLIP) 图像编码器[23]构建出多尺度金字塔。但是，经过随机采样的像素点不一定会落在预先构建的 CLIP 多尺度金字塔的中心位置，所以根据采样像素点的s_{init}值，在上下相邻的两个裁剪尺度下，各自找出与采样像素点最近的 4 个子图像的 CLIP 嵌入信息，通过三线性插值进行融合，从而得到采样像素点在金字塔中的 CLIP 嵌入信息$\boldsymbol{\xi}_{\text{clip}}$。把$\boldsymbol{\xi}_{\text{clip}}$作为语言特征的参考嵌入，构建出关于语言嵌入信息和 CLIP 嵌入信息相关的损失函数。

$$L_{\text{clip}} = \lambda_{\text{clip}} \hat{\boldsymbol{\xi}} \cdot \boldsymbol{\xi}_{\text{clip}}$$

式中：λ_{clip}为缩放常数，$\hat{\boldsymbol{\xi}} \cdot \boldsymbol{\xi}_{\text{clip}}$为语言嵌入信息之间的内积。

由于场景外观会随着视角的不同而发生变化，但是同一个物体的语义信息应该保持不变，所以语言嵌入信息的优化单独进行。最终，本文在三维重建过程中的总体损失函数定义为

$$L = \lambda_{\text{rgb}} L_{\text{rgb}} + \lambda_{\text{depth}} L_{\text{depth}} + \lambda_{\text{sdf}} L_{\text{sdf}}$$

式中λ_{rgb}、λ_{depth}和λ_{sdf}为各个损失项的加权因子，用于平衡各项损失函数对优化过程的贡献，实现在选定关键帧的所有采样像素上持续优化场景重建。

4　实验结果及分析

4.1　实验设置

4.1.1　数据集

EuRoC 数据集[24]由苏黎世大学和苏黎世联邦理工学院联合推出，被广泛应用于 SLAM 和自主机器人导航研究。该数据集通过配备高精度传感器的微型无人机采集数据，提供了双目相机图像、惯性测量单元信息以及激光跟踪仪采集的毫米级精度地面真值轨迹，可以用于评估 SLAM 算法的鲁棒性和定位精度。本文使用 9 个场景 (MH01~MH05、V101、V102、V201、V202) 的双目相机图像来评估算法模型，所有的图像都被下采样为 512 像素×384 像素分辨率。

Replica 数据集[25]是由 Facebook Reality 实验室发布的高质量室内三维重建数据集，主要应用于 SLAM、3D 视觉和增强现实等研究。该数据集使用高精度激光扫描仪采集 18 个多样化室内场景的毫米级精度三维模型，提供了多视角的单目图像和深度图像、精确的相机位姿、密集网格模型以及语义和实例分割标签等丰富的场景信息。本文采用 640 像素×320 像素的分辨率，在 8 个场景 (Room0~Room2、Office0~Office4) 中分别使用单目输入和 RGB-D 输入进行实验，并对提取的语言嵌入信息进行评估。

4.1.2　评价指标

定位任务主要使用绝对轨迹误差 (absolute

trajectory error, ATE) 作为评价指标, 衡量的是估计轨迹与真实轨迹之间的全局偏差, 数值越小表示性能越好。建图任务中使用的评价指标包括精度、完成度和完成率。精度计算的是重建结果与真实场景之间的平均距离误差, 反映重建模型的几何准确性, 数值越小, 几何准确性越高; 完成度主要衡量重建结果是否完整捕捉了真实场景的几何结构, 对真实场景中的每个点, 计算其到重建表面的最近距离, 取所有点的平均距离, 数值越小越好; 完成率统计的是真实场景中到重建表面距离小于 5 cm 的点的百分比, 数值越大, 表明重建模型越完整。

4.1.3　参数设置

在跟踪线程中, 选取关键帧的共视度阈值 $\tau_{\text{flow}}=4$, 并使用 DROID-SLAM[14] 预训练好的权重降低系统运行的时间。对于回环检测模块, 局部关键帧窗口大小 $N_{\text{local}}=75$, 平均光流阈值 $\tau_{\text{co}}=25$, 邻域半径 $r_{\text{local}}=1$, $r_{\text{loop}}=1$。后端跟踪线程中的邻域半径 $r_{\text{global}}=5$。在建图线程中的参数设置具体包括: 沿射线的采样点数 $N_{\text{start}}=24$, $N_{\text{imp}}=48$; 像素采样点数量 $M=4400$; 约束系数 $\beta=5$; 截断阈值 $\tau_{\text{trunc}}=16$; 损失权重 $\lambda_{\text{rgb}}=1$, $\lambda_{\text{depth}}=1$, $\lambda_{\text{sdf}}=0.1$; 神经辐射场初始学习率为 0.001, 每 10 个周期衰减 0.8。

4.2　定位结果与分析

在 EuRoC 数据集上, 本文模型使用双目图像作为输入, 与 SVO[26]、ORB-SLAM2[27]、ORB-SLAM3[28] 和 DROID-SLAM[14] 方法进行比较, 在 9 个场景下的 ATE 结果如表 1 所示。从表 1 中可以看出, 在双目输入的情况下, EuRoC 数据集上实验结果均优于其他方法, 例如, MH03 场景下本文模型的 ATE 为 0.019, 比 SVO 降低了 92.96%, 比 DROID-SLAM 降低了 45.71%; V202 场景的 ATE 为 0.009, 比 ORB-SLAM2 降低了 74.29%, 比 ORB-SLAM3 降低了 67.86%。从整体的平均表现来看, 本文模型的 ATE 比 SVO 降低了 78.09%, 比 ORB-SLAM2 降低了 49.61%, 比 ORB-SLAM3 降低了 42.31%, 整体的绝对轨迹误差有显著的降低。实验结果表明, 本文模型能够在双目输入条件下有效完成定位任务。

表 1　双目输入下不同模型在 EuRoC 数据集上的 ATE 结果
Table 1　ATE results of different models on EuRoC under binocular input　　m

场景	MH01	MH02	MH03	MH04	MH05	V101	V102	V201	V202	平均
SVO[26]	0.040	0.070	0.270	0.170	0.120	0.040	0.040	0.050	0.090	0.099
ORB-SLAM2[27]	0.035	0.018	0.028	0.119	0.060	0.035	0.020	0.037	0.035	0.043
ORB-SLAM3[28]	0.029	0.019	0.024	0.085	0.052	0.035	0.025	0.041	0.028	0.038
DROID-SLAM[14]	0.015	0.013	0.035	0.048	**0.040**	0.037	0.011	0.018	0.015	0.026
本文模型	**0.012**	**0.012**	**0.019**	**0.043**	**0.040**	**0.034**	**0.010**	**0.016**	**0.009**	**0.022**

注: 加黑代表最优结果。

在 Replica 数据集上, 本文模型分别使用单目图像和 RGB-D 图像作为输入, 在 8 个场景下的 ATE 结果如表 2 所示。在单目输入情况下, 与基于传统方法的 ORB-SLAM2 和 COLMAP[29] 相比, ORB-SLAM2 和 COLMAP 虽然在小部分场景中取得了较好的定位性能, 但是整体鲁棒性较差, 本文模型的平均 ATE 比 ORB-SLAM2 降低了 86.84%, 比 COLMAP 降低了 87.90%, 表明本文模型在跟踪模块通过引入深度学习方法, 有效提升了系统定位的精度。与基于深度学习方法的 DROID-SLAM 相比, 本文模型的整体定位性能较好, 并且在大部分场景下的定位结果优于 DROID-SLAM, 例如, Office0 场景中本文模型的 ATE 为 0.29, 比 DROID-SLAM 降低了 72.64%。与基于神经辐射场方法的 NICER-SLAM[30] 相比, 本文模型在所有场景下的定位精度均有优势, 平均 ATE 比 NICER-SLAM 降低了 77.47%。实验结果表明, 在单目输入下, 本文模型通过融合深度学习和神经辐射场方法有效提升了系统的定位性能。

表 2　不同模型在 Replica 数据集上的 ATE 结果
Table 2　ATE results of different models on Replica　　cm

输入	方法	Room0	Room1	Room2	Office0	Office1	Office2	Office3	Office4	平均
单目	ORB-SLAM2[27]	**0.30**	0.42	**0.25**	0.43	0.30	12.2	0.39	11.4	3.21
	COLMAP[29]	0.62	23.7	0.39	0.33	**0.24**	0.79	**0.14**	1.73	3.49
	DROID-SLAM[14]	0.58	0.58	0.38	1.06	0.40	0.70	0.53	1.33	0.70
	NICER-SLAM[30]	1.36	1.60	1.14	2.12	3.23	2.12	1.42	2.01	1.88
	本文模型	0.41	**0.34**	0.29	**0.29**	0.42	**0.38**	0.60	**0.65**	**0.42**

续表 2

输入	方法	Room0	Room1	Room2	Office0	Office1	Office2	Office3	Office4	平均
RGB-D	NICE-SLAM[19]	1.69	2.04	1.55	0.99	0.90	1.39	3.97	3.08	1.95
	Vox-Fusion[31]	**0.27**	1.33	0.47	0.70	1.11	**0.46**	**0.26**	0.58	0.65
	本文模型	0.64	**0.44**	**0.34**	**0.36**	**0.33**	0.47	0.49	**0.53**	**0.45**

注：加黑代表最优结果。

RGB-D 输入情况下，本文模型与 NICE-SLAM[19] 和 Vox-Fusion[31] 方法进行比较。从单个场景来看，本文模型在大多数场景中的表现更为精确，例如，在 Room1 场景中，本文模型的 ATE 为 0.44，比 NICE-SLAM 降低了 78.43%，比 Vox-Fusion 降低了 66.92%。从整体的平均表现来看，本文模型的定位性能提高了 76.94%。实验表明，在 RGB-D 输入下，本文模型在定位任务中依然具有优越性。

为了进一步验证模型性能，图 4 给出了在 2 种数据集上不同输入条件下的场景轨迹可视化结果。图 4 中的红色轨迹表示数据集提供的地面真值，蓝色轨迹表示本文模型的预测轨迹。从可视化的结果中可以观察到，在长序列的复杂场景下，本文模型的预测轨迹能够与地面真值轨迹保持高度一致。此外，本文模型在单目、双目和 RGB-D 等多种输入条件下均能够保持整体轨迹的准确性，表明本文模型的泛化能力较强，可以用于不同传感器配置的实际应用场景。

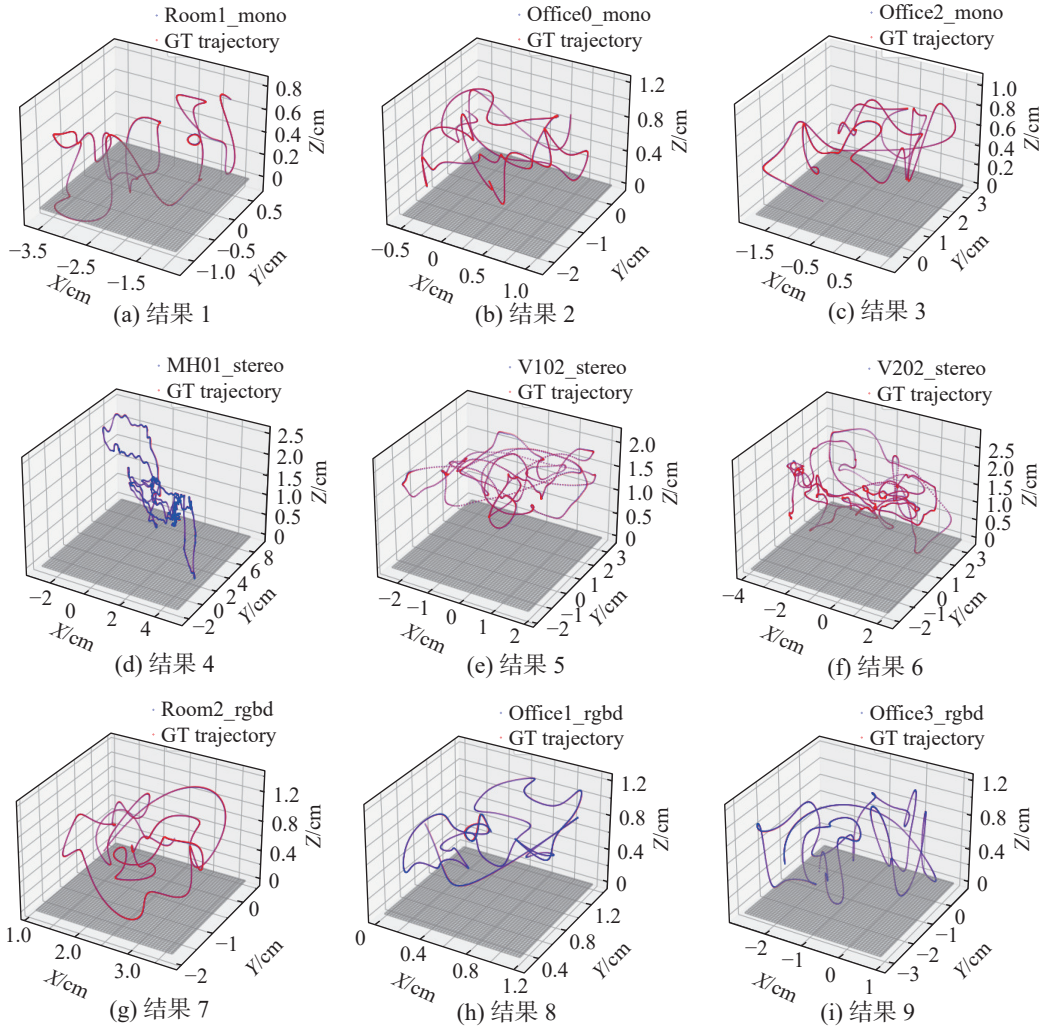


图 4　轨迹预测可视化结果

Fig. 4　Visualization results of trajectory prediction

4.3　重建结果与分析

本文在 Replica 数据集上对模型的建图性能进行量化评估。在单目输入条件下，本文模型与 DROID-SLAM[14]、COLMAP[29] 和 NICER-SLAM[30] 方法进行对比，结果如表 3 所示。与基于深度学习方法的 DROID-SLAM 和 COLMAP 相比，本文模型在大多数场景下的精度具有优势，例如，在 Office0 场景下，本文模型的重建精度为 2.83 cm，

比 DROID-SLAM 提高了 5.98%, 比 COLMAP 提高了 45.68%。整体上, 8 个场景的平均重建精度分别提升了 29.44% 和 55.33%。从完成度和完成率指标来看, 本文模型在所有场景下的性能均优于 DROID-SLAM 和 COLMAP。实验表明, 本文模型使用基于神经辐射场的稠密建图方法在几何完整性和场景覆盖能力上显著优于点云建图方法。相比于同样使用神经辐射场进行建图的 NICER-SLAM 方法, 本文模型在各个性能指标上的表现各有优劣。NICER-SLAM 通过使用更复杂的神经辐射场网络结构在重建精度和完成度上略有优势, 而本文模型在场景的完成率上表现略好, 实验结果表明本文模型能够在重建质量和场景覆盖率之间进行有效平衡。

表 3　单目输入下不同模型在 Replica 数据集上的重建定量结果

Table 3　Quantitative reconstruction results of different models on Replica under monocular input

方法	指标	Room0	Room1	Room2	Office0	Office1	Office2	Office3	Office4	平均
DROID-SLAM[14]	精度/cm	12.18	8.35	**3.26**	3.01	**2.39**	5.66	4.49	4.65	5.50
	完成度/cm	8.96	6.07	16.01	16.19	16.20	15.56	9.73	9.63	12.29
	完成率<5 cm/%	60.07	76.20	61.62	64.19	60.63	56.78	61.95	67.51	63.62
COL-MAP[29]	精度/cm	3.87	27.29	5.41	5.21	12.69	4.28	5.29	5.45	8.69
	完成度/cm	4.78	23.90	17.42	12.98	12.35	4.96	16.17	4.41	12.12
	完成率<5 cm/%	83.08	22.89	64.47	72.59	69.52	**81.12**	64.38	82.92	67.62
NICER-SLAM[30]	精度/cm	**2.53**	**3.93**	3.40	5.49	3.45	**4.02**	**3.34**	**3.03**	**3.65**
	完成度/cm	**3.04**	**4.10**	**3.42**	6.09	4.42	**4.29**	**4.03**	**3.87**	**4.16**
	完成率<5 cm/%	**88.75**	76.61	**86.10**	65.19	77.84	74.51	**82.01**	**83.98**	79.37
本文模型	精度/cm	4.40	4.13	3.67	**2.83**	3.12	4.81	4.38	3.70	3.88
	完成度/cm	4.45	4.52	6.29	**2.80**	**2.58**	4.79	4.84	4.12	4.30
	完成率<5 cm/%	80.23	**82.80**	82.10	**81.29**	**82.76**	75.59	75.88	80.24	**80.11**

注: 加黑代表最优结果。

在 RGB-D 输入条件下, 本文模型与 iMAP[17] 和 NICE-SLAM[19] 方法进行比较, 结果如表 4 所示。从单个场景来看, 本文模型在大多数场景下的重建性能都具有优势, 例如, 在 Office1 场景下, 本文模型的各项指标分别为 1.97、2.38 cm 和 93.08%, 与 iMAP 相比, 重建精度提高了 46.90%, 完成度提高了 54.75%, 完成率提高了 16.86%。相比于 NICE-SLAM, 重建精度提高了 41.19%, 完成度提高了 40.94%, 完成率提高了 13.33%。从整体的平均表现来看, 本文模型的各项指标均优于 iMAP, 完成率比 NICE-SLAM 高出 3.77%。定量的实验结果表明, 本文模型通过对基于神经辐射场的视觉 SLAM 系统结构进行优化, 能够在 RGB-D 输入下重建出更精确的稠密地图。

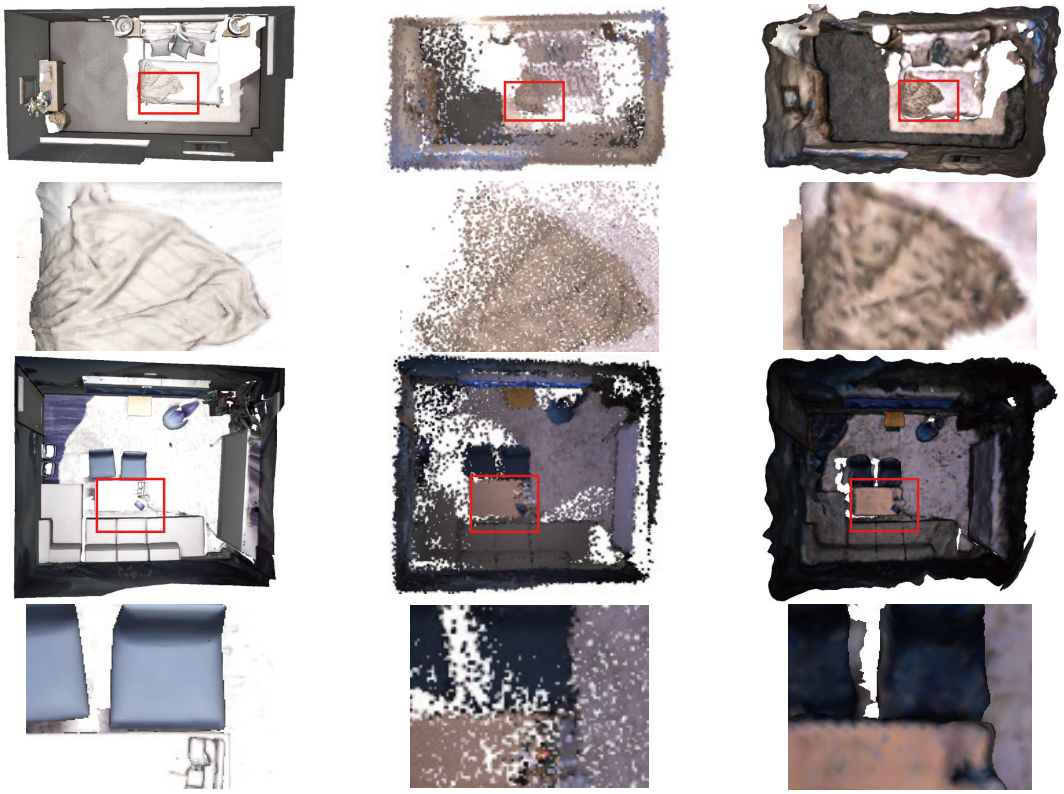
表 4　RGB-D 输入下不同模型在 Replica 数据集上的重建定量结果

Table 4　Quantitative reconstruction results of different models on Replica under RGB-D input

方法	指标	Room0	Room1	Room2	Office0	Office1	Office2	Office3	Office4	平均
iMAP[17]	精度/cm	3.58	3.69	4.68	5.87	3.71	4.81	4.27	4.83	4.43
	完成度/cm	5.06	4.87	5.51	6.11	5.26	5.65	5.45	6.59	5.56
	完成率<5 cm/%	83.91	83.45	75.53	77.71	79.64	77.22	77.34	77.63	79.05
NICE-SLAM[19]	精度/cm	**3.53**	3.60	**3.03**	5.56	3.35	4.71	**3.84**	**3.35**	3.87
	完成度/cm	**3.40**	3.62	**3.27**	4.55	4.03	**3.94**	**3.99**	4.15	**3.87**
	完成率<5 cm/%	**86.05**	80.75	**87.23**	79.34	82.13	80.35	**80.55**	82.88	82.41
本文模型	精度/cm	3.73	**2.53**	3.09	**2.19**	**1.97**	**3.94**	4.09	3.34	**3.11**
	完成度/cm	4.01	**2.32**	7.20	**2.30**	**2.38**	3.95	4.59	**4.02**	3.85
	完成率<5 cm/%	81.12	**94.05**	82.60	**92.27**	**93.08**	**84.00**	78.55	**83.79**	**86.18**

注: 加黑代表最优结果。

为了进一步验证模型性能, 图 5 给出了在单目输入条件下不同模型的定性实验结果。其中, 第 1 行和第 2 行分别为 Room1 场景的可视化渲染结果及其局部放大图, 第 3 行和第 4 行为在 Office0 场景中的实验结果。从整体的重建效果来看, 本文模型的重建结果更接近于真值, 能够较为完整地还原房间的整体布局, 而 DROID-SLAM 的重建结果为离散的点云, 存在较多的结构丢失。从局部的重建细节来看, 本文模型在细节处理上更为精细, 能够较好地恢复出物体表面的细节信息。

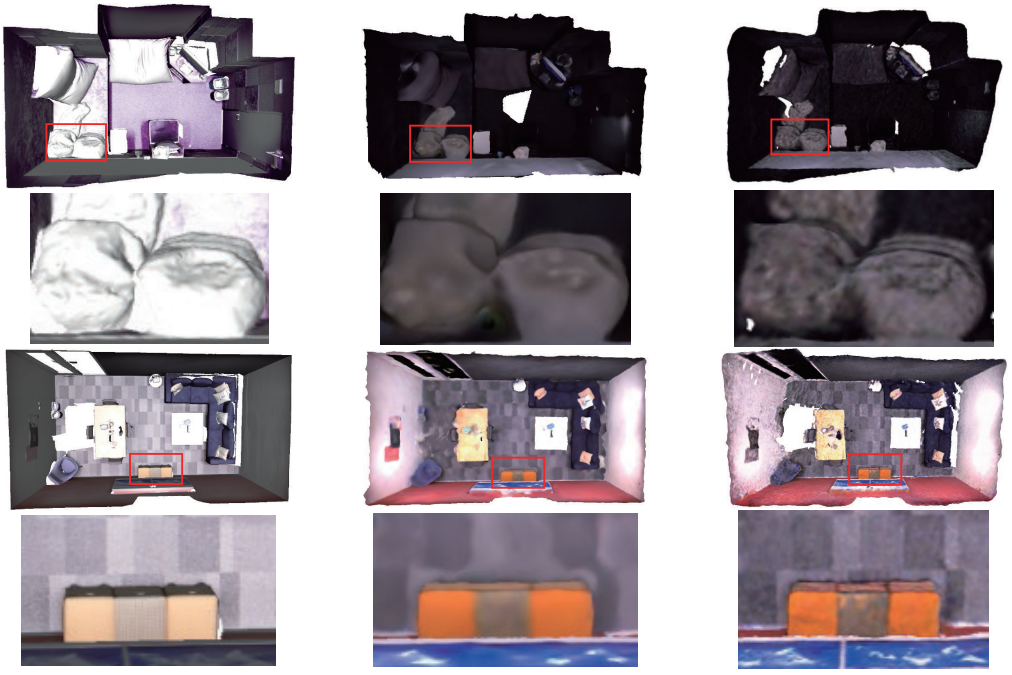


(a) 真值　(b) DROID-SLAM[14]　(c) 本文模型

图 5　单目输入下不同模型在 Replica 数据集上的定性实验结果

Fig. 5　Qualitative experimental results of different models on Replica under monocular input

图 6 给出了在 RGB-D 输入条件下，Replica 数据集中 Office1 和 Office2 两个场景下不同模型的定性对比实验结果。相比之下，本文模型在几何结构的细节重建方面更具有优势，能够对细节纹理进行准确建模。例如，在 Office1 场景中，对枕头褶皱纹理的建模清晰可见，更加贴合实际，Office2 场景中的地板木纹周期性图案、柜门把手等细节也能够进行高度还原。而 NICE-SLAM 的重建效果过于平滑，对高频细节的捕捉上相对不足。



(a) 真值　(b) NICE-SLAM[30]　(c) 本文模型

图 6　RGB-D 输入下不同模型在 Replica 数据集上的定性实验结果

Fig. 6　Qualitative experimental results of different models on Replica under RGB-D input

图 7 给出了在双目输入条件下, 本文模型和 DROID-SLAM 方法的重建对比结果。实验结果进一步说明利用点云进行重建难以对三维场景中的结构进行精确的理解, 例如, 对 MH01 和 V201 场景的重建结果存在明显的噪声。相比之下, 本文模型在场景结构完整性以及应对复杂场景的鲁棒性等方面具有明显优势, 能够准确重建出三维场景。

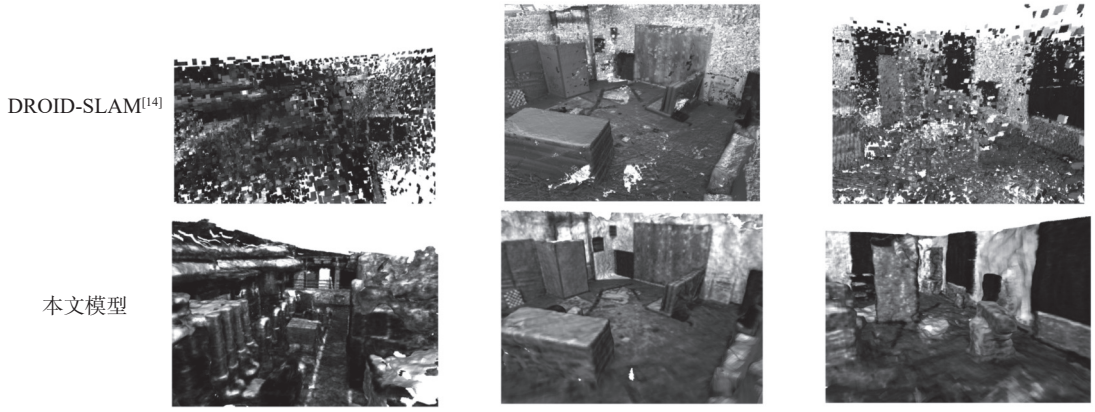


图 7　双目输入下不同模型在 EuRoC 数据集上的定性实验结果

Fig. 7　Qualitative experimental results of different models on EuRoC under binocular input

4.4　语言查询结果与分析

图 8 给出了在 Room0、Room2 和 Office3 这 3 个场景上的部分语言查询的可视化实验结果。从图 8 中可以看出, 本文模型对于常见物体和经典材质的语言查询表现较好, 能够较为准确地定位到相关 3D 位置, 例如, 在 Room0 场景中, 通过查询“pillows on the couch”, 模型能够在沙发区域上显示出枕头所在的位置, 表明本文模型具有对物体组合及位置关系的理解能力; 根据文本“wood”进行查询时, 模型也能够识别出木材材质及其所在物体, 表明本文模型具备一定的材质感知能力。

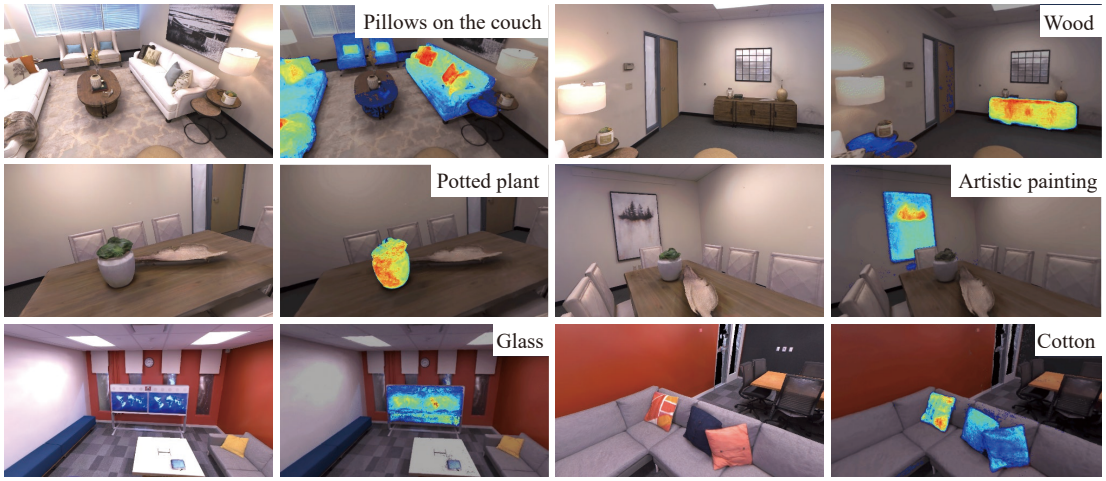


图 8　Replica 数据集上语言查询可视化实验结果

Fig. 8　Visualization experiment results of language queries on Replica

4.5　消融实验

本文模型在跟踪线程设置了回环检测和完整的全局 BA 优化, 为验证两个环节对 SLAM 系统定位性能的影响, 在 Replica 数据集上使用单目输入进行消融实验, 计算不同设置下 8 个场景的平均 ATE 值, 结果如表 5 所示。实验结果表明, 通过引入全局 BA, 本文模型对位姿估计的性能提高了 25.78%, 表明回环检测环节则有效缓解了 SLAM 系统的累积漂移问题, 为系统在复杂环境中运行提供了可靠性。与未引入回环检测和全局 BA 相比, 本文模型的定位性能提高了 65.11%。

表 5　Replica 数据集上针对定位的消融实验结果

Table 5　Results of ablation experiments targeting localization on Replica

回环检测	全局BA	ATE/cm
√	√	**0.43**
√	×	0.54

续表 5

回环检测	全局BA	ATE/cm
×	√	0.49
×	×	0.71

注：加黑代表最优结果。

为验证建图模块中各个损失函数对系统发挥的作用，同样在 Replica 数据集的 8 个场景中使用单目输入进行消融实验，结果如表 6 所示。在仅依靠颜色损失L_{rgb}进行优化时，模型的几何重建性能显著下降，完成率仅为 32.74%，表明颜色损失虽然能够对纹理进行约束，但是无法保证三维结构的准确性。对于深度损失L_{depth}来说，与仅使用颜色损失相比，重建质量明显提升，表明深度损失能够有效利用输入图像的深度信息，对场景几何进行优化。而通过进一步引入 SDF 损失L_{sdf}，本文模型的整体性能达到最优，表明通过对重建表面进行约束，可以有效减少重建噪声，使得重建结果更加鲁棒。针对各项损失函数的消融实验结果表明，L_{rgb}、L_{depth}和L_{sdf}共同提升了重建的精度和完整性。

表 6　Replica 数据集上针对重建的消融实验结果

Table 6　Experimental results of ablation for reconstruction on Replica

L_{rgb}	L_{depth}	L_{sdf}	精度/cm	完成度/cm	完成率<5 cm/%
√	√	√	**3.89**	**4.27**	**80.17**
√	√	×	3.95	4.34	78.91
√	×	√	3.93	4.31	79.39
√	×	×	9.53	10.46	32.74

注：加黑代表最优结果。

5　结束语

本文提出的融合深度学习位姿估计与神经隐式表征的视觉 SLAM 系统，通过在跟踪模块引入基于共视帧图的对齐策略，集成高效的回环检测与全局优化机制，并结合神经辐射场构建出具有语言查询功能的高保真地图，有效提升了系统的定位精度、重建质量和实用性。在公共数据集上的大量实验证明了本文算法在定位和重建方面的准确性与鲁棒性。在未来工作中，将关注系统在更复杂环境中的适应性，提高系统性能。

参考文献：

[1] 黄泽霞, 邵春莉. 深度学习下的视觉 SLAM 综述[J]. 机器人, 2023, 45(6): 756−768.
HUANG Zexia, SHAO Chunli. A survey of visual SLAM under deep learning[J]. Robot, 2023, 45(6): 756−768.

[2] DETONE D, MALISIEWICZ T, RABINOVICH A. SuperPoint: self-supervised interest point detection and description[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. Salt Lake City: IEEE, 2018.

[3] LUO Zixin, SHEN Tianwei, ZHOU Lei, et al. GeoDesc: learning local descriptors by integrating geometry constraints[C]//European Conference on Computer Vision. Munich: ECVA, 2018.

[4] SARLIN P E, DETONE D, MALISIEWICZ T, et al. SuperGlue: learning feature matching with graph neural networks[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020.

[5] RANFTL R, KOLTUN V. Deep fundamental matrix estimation[C]//European Conference on Computer Vision. Munich: ECVA, 2018.

[6] VON STUMBERG L, WENZEL P, YANG Nan, et al. LM-reloc: levenberg-marquardt based direct visual relocalization[C]//2020 International Conference on 3D Vision. Fukuoka: IEEE, 2020.

[7] SARLIN P E, UNAGAR A, LARSSON M, et al. Back to the feature: learning robust camera localization from pixels to pose[C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville: IEEE, 2021.

[8] MCCORMAC J, HANDA A, DAVISON A, et al. SemanticFusion: dense 3D semantic mapping with convolutional neural networks[C]//2017 IEEE International Conference on Robotics and Automation. Singapore: IEEE, 2017.

[9] YU Chao, LIU Zuxin, LIU Xinjun, et al. DS-SLAM: a semantic visual SLAM towards dynamic environments[C]//2018 IEEE/RSJ International Conference on Intelligent Robots and Systems. Madrid: IEEE, 2018.

[10] TATENO K, TOMBARI F, LAINA I, et al. CNN-SLAM: real-time dense monocular SLAM with learned depth prediction[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017.

[11] ZHOU Huizhong, UMMENHOFER B, BROX T. DeepTAM: deep tracking and mapping[C]//European Conference on Computer Vision. Munich: ECVA, 2018.

[12] BLOESCH M, CZARNOWSKI J, CLARK R, et al. CodeSLAM-learning a compact, optimisable representation for dense visual SLAM[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018.

[13] CZARNOWSKI J, LAIDLOW T, CLARK R, et al. Deep-

Factors: real-time probabilistic dense monocular SLAM[J]. IEEE robotics and automation letters, 2020, 5(2): 721−728.

[14] TEED Z, DENG J. Droid-slam: Deep visual slam for monocular, stereo, and RGB-D cameras[C]//Proceedings of the 38th Annual Conference on Neural Information Processing Systems. Vancouver: NeurIPS, 2021.

[15] TEED Z, DENG Jia. RAFT: recurrent all-pairs field transforms for optical flow[C]//European Conference on Computer Vision. ONLINE: ECVA, 2020.

[16] MILDENHALL B, SRINIVASAN P P, TANCIK M, et al. NeRF: representing scenes as neural radiance fields for view synthesis[C]//European Conference on Computer Vision. online: ECVA, 2020.

[17] SUCAR E, LIU Shikun, ORTIZ J, et al. iMAP: implicit mapping and positioning in real-time[C]//2021 IEEE/CVF International Conference on Computer Vision. Montreal: IEEE, 2021.

[18] KONG Xin, LIU Shikun, TAHER M, et al. vMAP: vectorised object mapping for neural field SLAM[C]//2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver: IEEE, 2023.

[19] ZHU Zihan, PENG Songyou, LARSSON V, et al. NICE-SLAM: neural implicit scalable encoding for SLAM[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans: IEEE, 2022.

[20] WANG Hengyi, WANG Jingwen, AGAPITO L. Co-SLAM: joint coordinate and sparse parametric encodings for neural real-time SLAM[C]//2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver: IEEE, 2023.

[21] CHO K, VAN MERRIENBOER B, GULCEHRE C, et al. Learning phrase representations using RNN encoder–decoder for statistical machine translation[C]//Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing. Doha: USAACL, 2014.

[22] 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): 1−15.

[23] RADFORD A, KIM W J, HALLACY C, et al. Learning transferable visual models from natural language supervision[EB/OL]. (2021−02−26)[2025−04−20]. https://arxiv.org/abs/2103.00020.

[24] 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.

[25] STRAUB J, WHELAN T, MA L N, et al. The replica dataset: a digital replica of indoor space[EB/OL]. (2019−06−13)[2025−04−20]. https://arxiv.org/abs/1906.05797.

[26] FORSTER C, PIZZOLI M, SCARAMUZZA D. SVO: fast semi-direct monocular visual odometry[C]//2014 IEEE International Conference on Robotics and Automation. Hong Kong: IEEE, 2014.

[27] MUR-ARTAL R, TARDÓS J D. ORB-SLAM2: an open-source SLAM system for monocular, stereo, and RGB-D cameras[J]. IEEE transactions on robotics, 2017, 33(5): 1255−1262.

[28] 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.

[29] SCHÖNBERGER J L, FRAHM J M. Structure-from-motion revisited[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016.

[30] ZHU Zihan, PENG Songyou, LARSSON V, et al. NICER-SLAM: neural implicit scene encoding for RGB SLAM[C]//2024 International Conference on 3D Vision. Davos: IEEE, 2024.

[31] YANG Xingrui, LI Hai, ZHAI Hongjia, et al. Vox-fusion: dense tracking and mapping with voxel-based neural implicit representation[C]//2022 IEEE International Symposium on Mixed and Augmented Reality. Singapore: IEEE, 2022.

作者简介:

张含笑, 硕士, 主要研究方向为计算机视觉。E-mail: 2682706067@qq.com。

邢向磊, 教授, 博士生导师, 主要研究方向为模式识别与计算机视觉。获得黑龙江省高校科学技术奖 (自然科学类) 一等奖, 获《智能系统学报》优秀论文奖。发表学术论文 60 余篇。E-mail: xingxl@hrbeu.edu.cn。