Multi-Agent Neural SLAM for Autonomous Robots

Tianchen Deng, Guole Shen, Xun Chen, Hongming Shen, Yanbo Wang, Jingchuan Wang, Weidong Chen

***Abstract*—Neural implicit scene representations have recently shown promising results in dense visual SLAM. However, existing implicit SLAM algorithms are constrained to single-agent scenarios, and falls difficulty in large indoor scenes and long sequences, due to their single, global radiance field with finite capacity. To this end, we propose a novel multi-agent collaborative SLAM framework with joint scene representation, distributed camera tracking, intra-to-inter loop closure, and sub-map fusion. Specifically, we propose a distributed learning framework for multi-agent neural SLAM system to improve multi-agents cooperation and communication bandwidth efficiency. A novel intra-to-inter loop closure method is designed to achieve local (single-agent) and global map consistency. Our framework supports single-agent and multi-agents operation. Furthermore, to the best of our knowledge, there is no real-world dataset for NeRF-based/GS-based SLAM that provides both continuous-time trajectories groundtruth and high-accuracy 3D meshes groundtruth. To this end, we introduce the first real-world dataset covering both single-agent and multi-agent scenarios, ranging from small rooms to large-scale environments, with high-accuracy ground truth for 3D reconstruction meshes and continuous-time camera trajectory. This dataset can advance the development of the community. Experiments on various datasets demonstrate the superiority of the proposed method in both camera tracking and mapping. The dataset and code will open-source in Github.**

I. Introduction

Dense Visual Simultaneous Localization and Mapping (SLAM) has been a fundamental challenge in robotics and computer vision, which aims to achieve 3D reconstruction of unknown environments and localization. Dense visual SLAM has wide applications such as autonomous driving [1], remote sensing, and VR/AR. Traditional visual SLAM has witnessed continuous development, achieving accurate mapping and tracking in various scenarios, such as DTAM [2], Kinectfusion [3]. They reconstruct meaningful 3D global maps. Nowadays, with the proposal of Neural Radiance Fields (NeRF) [4], there are many following works combining the implicit scene representation with SLAM framework. iMAP [5] is the first work to use a single multi-layer perceptron (MLP) to reconstruct the scene in online mapping framework. NICE-SLAM [6], ESLAM [7], Go-SLAM [8], Co-SLAM [9], and PLGSLAM [10] further improve the scene representation with hybrid feature grids, axis-aligned feature planes, joint coordinate-parametric encoding, and progressive scene representation. They can achieve promising reconstruction quality in a small indoor room. Current research on NeRF-based SLAM frameworks primarily focuses on the single-robot setting. Compared to a single robot, multi-robot systems can perform tasks more efficiently and robustly. Moreover, the distributed configuration enhances the system's resilience to individual agent failures, ensuring stability even in the presence of anomalies. CP-SLAM [11] firstly uses implicit neural mapping in multi-agent SLAM systems. It employs neural points to represent the scene which shows reasonable but limited reconstruction accuracy. All of these methods have difficulties in large-scale scenarios. We outline the key challenges for multi-agents NeRF-based SLAM in large-scale scenes: **a) Communication bandwidth restriction** agents can only exchange limited data instead of raw data in a timely manner. **b) Accumulation of errors and pose drifts** Existing frameworks struggle with accumulating errors in large-scale scenes. **c) Consensus on scene reconstruction among all agents.** The sensor observations of multiple agents are local and different, which leads to discrepancies of the final map.

To this end, we design a novel multi-agents neural SLAM framework for accurate scene reconstruction, robust pose estimation and multi-agents collaboration. Given the constraints of privacy and communication bandwidth, agents can only exchange their network weights and limited data instead of raw data in a timely manner. So, we design the distributed optimization strategy for multi-agents neural SLAM. In the

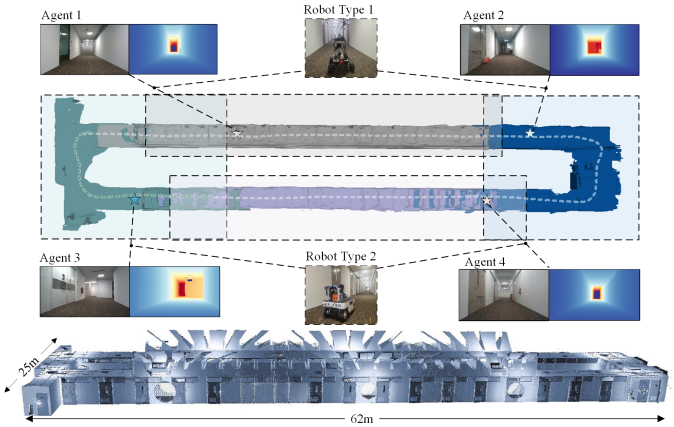


Fig. 1: We present MNE-SLAM, a novel multi-agent collaborative SLAM system with joint implicit neural scene representation, distributed camera tracking, intra-to-inter loop closure, and submap fusion. Depicted in the bottom, we demonstrate the first real-world large-scale 3D indoor dataset(>1500 m^2) with high-accuracy 3D groundtruth mesh. This scene is collected through various laser scans. In the four corner subgraphs, we demonstrate the rendered depth and color image of MNE-SLAM. The mesh of each agent is highlighted in different colors, along with their trajectories.

Tianchen Deng, Guole Shen, Yanbo Wang, Jingchuan Wang, Weidong Chen are with Institute of Medical Robotics and Department of Automation, Shanghai Jiao Tong University, and Key Laboratory of System Control and Information Processing, Ministry of Education, Shang hai 200240, China. (* The first two author are equal contribution.)

first stage, we set up independent scene representation neural networks for each agent and optimize them separately. Then, we perform loop closure, and sub-maps fusion to achieve comprehensive and consistent comprehension of multiple agents.

In multi-agents SLAM systems, all these agents optimize their scene representation independently with local and different observations. As a result, the final maps generated by individual agents only represent specific portions of the scene, leading to discrepancies. To achieve a comprehensive and consistent comprehension of the scene, we propose a novel intra-to-inter loop closure method. Intra-loop is to detect the pre-visit place of a robot, which can effectively eliminate the cumulative error with the global keyframe database. The cumulative error becomes significantly evident in large indoor scenes and long video sequences. Inter-loop is designed to detect the same scenario visited by multiple agents and register submaps generated from different agents. Inter-loop can significantly improve the global map consistency and consensus among all agents. A global consistency loss is introduced to guide the process from local-to-global optimization. The results demonstrate that each agent ultimately acquires accurate global information, enabling the construction of a comprehensive map. In practice, our method achieves SOTA performance in 3D reconstruction and camera tracking.

Furthermore, to the best of our knowledge, we find that current datasets are either virtual, such as Replica [12] or only provide trajectory ground truth without 3D ground truth, such as ScanNet [13], and RGB-D TUM dataset [14]. ScanNet++ [15] is the only real-world dataset that offers both trajectory and 3D ground truth. However, Scannet++ is primarily intended as a NeRF Training & Novel View Synthesis dataset as stated in their paper and SplaTAM [16]. ScanNet++ contains non-time-continuous trajectories with numerous abrupt jumps and teleportations, which makes it unsuitable for SLAM systems. We propose a real-world dataset with small room scenarios and large-scale environments, covering both single-agent and multi-agents scenarios. Our dataset provide high-accuracy and time-continuous trajectory and 3D mesh groudtruth, which is suitable for all neural SLAM systems, such as NeRF-based SLAM and 3DGS-based SLAM systems [17], [18]. Overall, our contributions are shown as follows:

- We design a novel multi-agent neural SLAM framework with joint neural scene representation, and distributed camera tracking, achieving accurate multi-agents pose tracking and collaborative scene reconstruction.
- An intra-to-inter loop closure method is proposed to eliminate the cumulative pose drift and register the neural sub-maps for global map consistency. We also maintain a global keyframe database to perform intra-to-inter bundle adjustment.
- We propose the first real-world dataset for all kinds of neural SLAM systems with high-accuracy ground-truth for both camera trajectory and 3D reconstruction mesh, which can foster the development of the community.

II. METHODS

The pipeline of our system is shown in Fig. 2. We propose MNE-SLAM, a collaborate multi-agent SLAM system. The input of this multi-agent system is RGB-D frames $\{I_i, D_i\}_{i=1}^{M}$ with known camera intrinsic $K \in R_{3\times3}$. Our model predicts multi-agents camera poses $\{R_i|t_i\}_{i=1}^{M}$, color $\mathbf{c}$, and implicit truncated signed distance function (TSDF) representation. The system consists of four main modules: (i) local (single-robot) mapping (Sec. II-A), (ii) distributed camera tracking via PGO (Sec. II-B), (iii) intra-to-inter loop clousre (Sec. II-C) (iv) sub-map fusion). Among these modules, distributed loop detection and intra loop closure are the only ones that involve communication between robots We will elaborate on the entire pipeline of our system in the following subsections.

A. Joint Scene Representation

Voxel grid-based architectures [19], [6] are the mainstream in NeRF-based SLAM system. However, they face challenges with cubic memory growth and real-time performance in large-scale environments. Inspired by [7], we design a parametric-coordinate joint encoding method. The parametric encoding employs tri-plane encoding, while the coordinate encoding uses one-blob encoding with an MLP. This joint scene representation architecture enables high-fidelity, smooth scene reconstruction with enhanced hole-filling capabilities. Specifically, we employ a dual-scale tri-plane representation, consisting of both coarse and fine levels, to model the scene. The tri-planes feature $\boldsymbol{T}(x)$ can be formulated as:

$$\begin{aligned} t^c(x) &= T^c_{xy}(x) + T^c_{xz}(x) + T^c_{yz}(x) \\ t^f(x) &= T^f_{xy}(x) + T^f_{xz}(x) + T^f_{yz}(x) \\ \boldsymbol{T}(x) &= Concat\left(t^c(x); t^f(x)\right) \end{aligned} \tag{1}$$

where $t^c(x), t^f(x)$ denote the coarse and fine feature form tri-planes. x is the world coordinate. $\{T^c_{xy}, T^c_{xz}, T^c_{yz}\}$ represent the three coarse geometry feature planes, and $\{T^f_{xy}, T^f_{xz}, T^f_{yz}\}$ represent the three fine geometry feature planes.

Differentiable Rendering We use a differentiable rendering process to integrate the predicted density and colors from our local scene representation. We determine a ray $r(t) = \mathbf{o} + t\mathbf{d}$ whose origin is at the camera center of projection o, ray direction r. We uniformly sample K points. The sample bound is within the near and far planes $t_k \in [t_n, t_f]$, $k \in \{1, \ldots, K\}$ with depth values $\{\mathbf{d_1}, \ldots, \mathbf{d_K}\}$ and predicted colors $\{\mathbf{c_1}, ..., \mathbf{c_K}\}$. For all sample points along rays, we query TSDF $\phi_{\boldsymbol{g}}(p_k)$ and raw color $\phi_{\boldsymbol{a}}(p_k)$ from our networks and use the SDF-Based rendering approach to convert SDF values to volume densities.

Then we define the termination probability w_k, depth $\hat{\boldsymbol{d}}$, and color $\hat{\boldsymbol{c}}$ as:

$$\begin{aligned} w_k &= \exp\left(-\sum_{m=1}^{n-1} \boldsymbol{\sigma}\left(x_m\right)\right)\left(1 - \exp\left(-\boldsymbol{\sigma}\left(x_k\right)\right)\right) \\ \hat{\boldsymbol{c}} &= \sum_{k=1}^{N} w_k \boldsymbol{\phi_a}\left(x_k\right) \quad \text{and} \quad \hat{\boldsymbol{d}} = \sum_{k=1}^{N} w_k t_k \end{aligned} \tag{2}$$

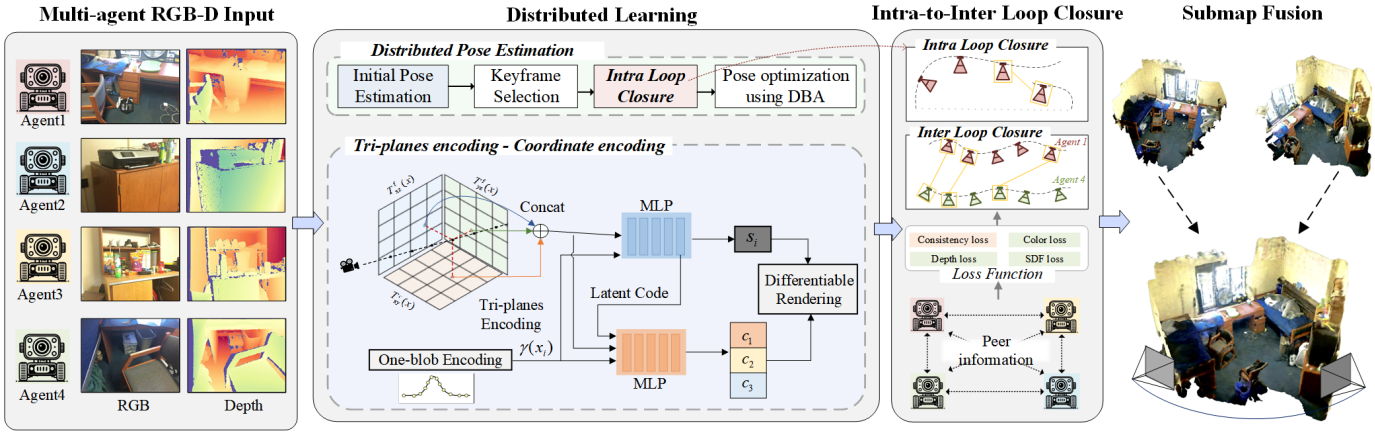


Fig. 2: **System Overview.** Our system is a multi-agent collaborative SLAM system which consists joint scene representation, distributed tracking, intra-to-inter loop closure, and submap-fusion. In distributed optimization module, each agent takes the color images and depth images as input. In addition, each agent will exchange the network weights of its peers. We carefully design consistency loss with color, depth, and SDF loss in inter-loop closure. Each agent can successively performs individual scene mapping and collaborative mapping and tracking to generate the final neural implicit map with submap-fusion.

B. Distributed Camera Tracking

In multi-agent collaboration SLAM system, distributed pose estimation are crucial under communication bandwidth restriction and accumulation of pose errors. In this work, we propose a distributed camera tracking framework with intra-loop closing and global bundle adjustment. These enhancements effectively reduce the drift and pave the way for a globally optimal map.

For each agent, our distributed camera tracking framework takes a live RGB-D video stream. It first applies a recurrent update operator based on RAFT [20] to compute the optical flow of each new frame compared to the last keyframe. If the average flow is larger than a pre-defined τ_k, a new keyframe is created and added to the global keyframe database for further refinement. We use the set of keyframes $\{KF_i\}_{i=1}^{N}$ to create a keyframe-graph structure $(\mathcal{V}, \mathcal{E})$, which represents the co-visibility between frames. An edge $(i, j) \in \mathcal{E}$ means I_i, I_j have overlapping fields of view which shared points. The keyframe graph is built dynamically with the system operation.

Afterward, we use the differentiable Dense Bundle Adjustment (DBA) layer to solve a non-linear squares optimization problem to correct the camera pose $\{R_i|t_i\}_{i=1}^{M}$.

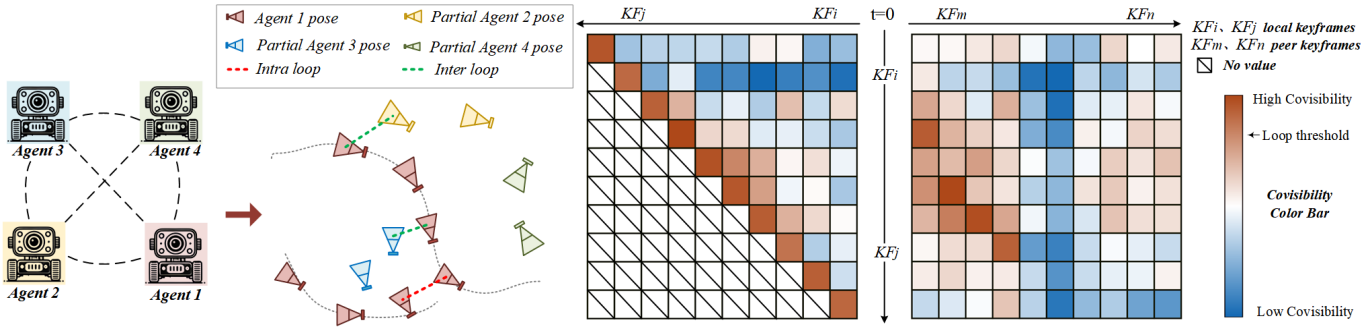
$$\mathbf{E}\left(\{\mathbf{R}|\mathbf{t}\}, \mathbf{d}\right) = \sum_{(i,j)\in\mathcal{E}} \left\| \mathbf{p}_{ij}^{*} - \Pi_c \left(\{\mathbf{R}|\mathbf{t}\}_{ij} \circ \Pi_c^{-1}\left(\mathbf{p}_i, \mathbf{d}_i\right) \right) \right\|_{\Sigma_{ij}}^{2} + \alpha \sum \left\| \hat{\mathbf{d}}_{\mathbf{i}} - \mathbf{D}_{\mathbf{i}} \right\|^{2} \tag{3}$$

where Π_c, Π_c^{-1} are the projection and back-projection functions, $\|\cdot\|_{\Sigma}$ and $\Sigma_{ij} = \operatorname{diag} \mathbf{w}_{ij}$ denotes the Mahalanobis distance which weights the error terms based on the confidence weights $\mathbf{w}_{ij}$. $\mathbf{p_i}$ is the 2D pixel position from keyframe KF_i. $\{\mathbf{R}|\mathbf{t}\}_{ij}$ is the pose transformation from KF_i to KF_j. We use $\mathbf{p}_{ij}^{*}$ to present the estimated flow. α denotes the weight of depth residual. This loss function states that we want to optimize the camera pose and per-pixel depth to maximize the compatibility with flow $\mathbf{p}_{ij}^{*}$ predicted by the recurrent update operator. We use local parameterization to linearize Eq. 3 and use the Gauss-Newton algorithm solve for updates.

C. Intra-to-inter Loop Closure

Most neural implicit SLAM frameworks suffer from accumulation of pose drifts and distortion in the reconstruction. Their tracking networks are performed via minimizing rgb loss functions with respect to learnable parameters θ to estimate the relative pose matrix $\{R_i|t_i\} \in \mathbb{SE}(3)$. With the growing cumulative error ε of pose estimation, those methods result in failure in large-scale indoor scenes and long videos. Concurrently, during multi-agent collaborative mapping in large-scale environments, the accumulation pose error of multiple agents results in a rapid degradation of the submaps. To address these problems, we propose a novel intra-to-inter loop closure method, which can eliminate pose drift and achieve global consistency across multiple submaps. We illustrate the schematic diagram of the intra-to-inter loop closure in Fig. 3.

Intra-Loop Closure For single agent (intra) loop closure, we build the keyframe-graph $(\mathcal{V}, \mathcal{E})$ with two steps: (i) detect and select keyframe pairs with high covisibility τ_{cov} in the local most recent keyframes N_{local} (ii) detect loop closure between local keyframes and historical keyframes outside the local window. Accordingly, we compute a covisibility matrix of local keyframes of size $N_{local} \times N_{KF}$ for local loop closure. We also compute a covisibility matrix of all historical keyframes of size $N_{KF} \times N_{KF}$ for global loop closure, as shown in Fig. 3 (b). The covisibility is represented by the mean rigid flow between keyframe pairs using efficient back-projection. Those keyframe pairs with low covisibility (mean flow higher than thershold (τ_{cov}) are filtered out. We build edges for these keyframe pairs. We suppress the possible neighboring edges between $\{KF_k\}_{k=i-r_{local}}^{i+r_{local}} \rightarrow \{KF_k\}_{k=i-r_{local}}^{j+r_{local}}$, where r_{local}



(a) **Intra-to-Inter** Loop Closure (b) Covisibility Matrix for **Intra**-Loop Closure (c) Covisibility Matrix for **Inter**-Loop Closure

Fig. 3: (a) presents the reconstruction of multi-agent pose graph with intra-to-inter loop closure. (b) and (c) present the covisibility matirx of intra loop closure and inter loop closure. In Figure (b), the horizontal and vertical axes represent the keyframes of the local (single-agent) system, while the horizontal and vertical axes in figure (c) represent the keyframes of the local and peer agents.

denotes a temporal radius. For local loop closure, the number of edges in the graph is linear to N_{local} with an upper bound $N_{local} \times N_{local}$. The number of edges of the global loop closure is linear to N_{KF} with an upper bound $N_{KF} \times N_{KF}$. We also suppress the redundant neighboring edges with radius r_{global}. By constructing a pose graph using global historical frames, we can leverage all information to optimize the pose of the current frame. Intra-loop closure effectively integrates local and global information which greatly improves the accuracy of camera pose optimization in large-scale indoor scenes. Through neighborhood suppression and covisibility filtering, we limit the number of edges in the keyframe-graph and ensure the efficiency of optimization of intra-loop closure.

Inter-Loop Closure For multi-agent loop closure, we detect keyframe pairs from global keyframe database $\mathcal{G}$ of different agents. Once the covisibility value of the keyframe pairs is lower than the thershold τ_{cov}, we perform inter loop closure on these frames. We present the covisibility matrix in Fig. 3(c). For the robot α, β and the detected frames A, B, we set the odometric estimates of pose i and pose j as $\{\mathbf{R}|\mathbf{t}\}^A_{\alpha_i}, \{\mathbf{R}|\mathbf{t}\}^B_{\beta_j} \in \mathbb{SE}(3)$. Then, the relative transformation between frames A, B can be formulate as:

$$\{\mathbf{R}|\mathbf{t}\}^A_{B_{ij}} \triangleq \{\mathbf{R}|\mathbf{t}\}^A_{\alpha_i}\{\mathbf{R}|\mathbf{t}\}^{\alpha_i}_{\beta_j}\left(\{\mathbf{R}|\mathbf{t}\}^B_{\beta_j}\right)^{-1} \tag{4}$$

where the subscript of $\{\mathbf{R}_i|\mathbf{t}_i\}^A_{B_{ij}}$ indicates that this estimate is computed using inter loop closure (i,j). In order to obtain a reliable estimate of the true relative transformation, we formulate the pose average problem:

$$\{\mathbf{R}|\mathbf{t}\}^A_{B_{ij}} \in \underset{\{\mathbf{R}|\mathbf{t}\}\in \mathrm{SE}(3)}{\arg\min} \sum_{(i,j)\in L_{\alpha,\beta}} \mathcal{L}\left(r_{ij}(\{\mathbf{R}|\mathbf{t}\})\right) \tag{5}$$

where $\mathcal{L}$ denotes the loss function. $L_{\alpha,\beta}$ is the set of inter-robot loop closure between robot α and β. We adopt consistency loss for multi-agent consistency.

$$\mathcal{L}_{lc}(\{\mathbf{R}|\mathbf{t}\}) = \frac{1}{n}\sum_{\{A,B\}}^{n}\left(\hat{\boldsymbol{c}_A} - \hat{\boldsymbol{c}_B}\right)^2 \tag{6}$$

$$\mathcal{L}_{ld}(\{\mathbf{R}|\mathbf{t}\}) = \frac{1}{n}\sum_{\{A,B\}}^{n}\left(\hat{\boldsymbol{d}_A} - \hat{\boldsymbol{d}_B}\right)^2 \tag{7}$$

where $\hat{\boldsymbol{c}_A}, \hat{\boldsymbol{c}_B}$ are the rendered color image from agent α, β, while $\hat{\boldsymbol{d}_A}, \hat{\boldsymbol{d}_B}$ denote the rendered depth. This consistency loss effectively aligns the trajectories and maps across agents, by comparing the rendering results of two images.

D. Distributed Optimization and Submap-Fusion

Our mapping thread is performed via minimizing our objective functions with respect to network parameters θ and camera parameters $\{\mathbf{R}|\mathbf{t}\}$. The color and depth rendering losses are used in our mapping thread:

$$\mathcal{L}_c = \frac{1}{N}\sum_{i=1}^{N}\left(\hat{\mathbf{c}}_i - \mathbf{C_i}\right)^2, \quad \mathcal{L}_d = \frac{1}{|R_i|}\sum_{i\in R_i}\left(\hat{\mathbf{d}}_\mathbf{i} - \mathbf{D_i}\right)^2 \tag{8}$$

where R_i is the set of rays that have a valid depth observation. In addition, we design SDF loss, free space loss for our mapping thread. Specifically, for samples within the truncation region, we leverage the depth sensor measurement to approximate the signed distance field:

$$\mathcal{L}_{sdf} = \frac{1}{|R_i|}\sum_{r\in R_i}\frac{1}{|X_r^{tr}|}\sum_{x\in X_r^{tr}}\left(\phi_{\boldsymbol{g}}(x)\cdot T - (\mathbf{D_i} - \mathbf{d})\right)^2 \tag{9}$$

where X_r^{tr} is a set of points on the ray r that lie in the truncation region, $|\mathbf{D_i} - \mathbf{d}| \leq tr$. We differentiate the weights of points that are closer to the surface $X_r^{tm} = \{x|x \in |\mathbf{D_i} - \mathbf{d}| \leq 0.4tr\}$ from those that are at the tail of the truncation region X_r^{tt} in our SDF loss.

$$\mathcal{L}_{sdf_m} = \mathcal{L}_{sdf}\left(X_r^{tm}\right), \quad \mathcal{L}_{sdf_t} = \mathcal{L}_{sdf}\left(X_r^{tt}\right) \tag{10}$$

For sample points that are far from the surface $|D_i - d| \geq T$:

$$\mathcal{L}_{fs} = \frac{1}{|R_i|}\sum_{r\in R_i}\frac{1}{\left|X_r^{fs}\right|}\sum_{x\in X_r^{fs}}\left(\phi_{\boldsymbol{g}}(x) - 1\right)^2 \tag{11}$$

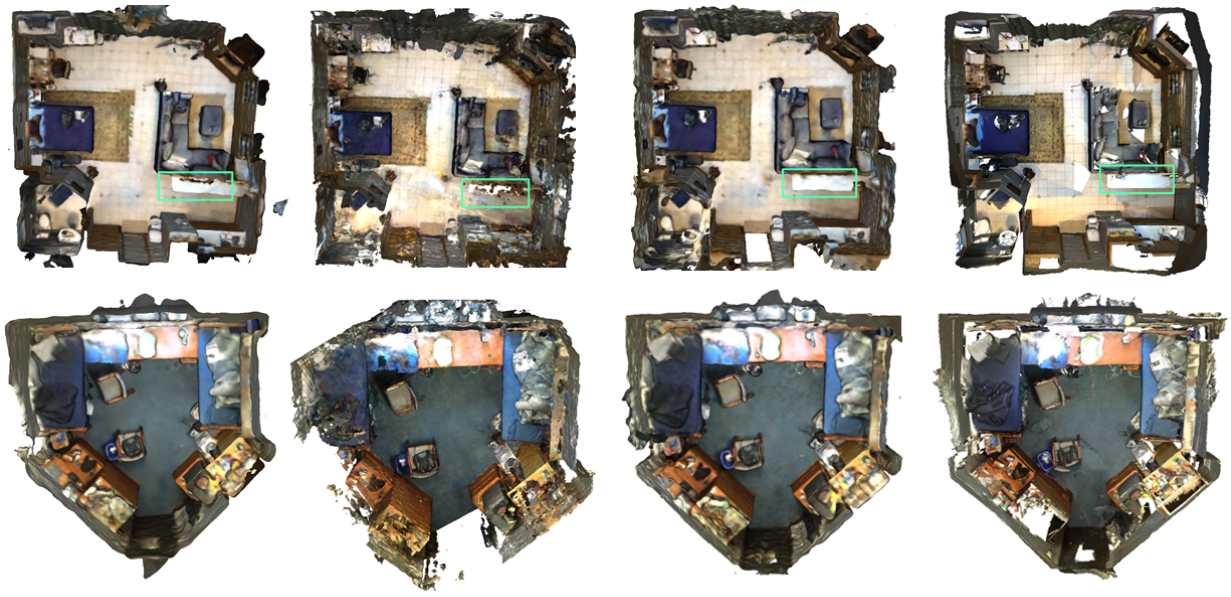


Fig. 4: Qualitative comparison of our proposed our method's surface reconstruction and localization accuracy with existing NeRF-based dense visual SLAM methods, NICE-SLAM [6], Co-SLAM [9], and ESLAM [7] on the ScanNet dataset [13].

This loss can force the SDF prediction value to be the truncated distance tr.

III. Experiments

We validate that our method outperforms existing implicit representation-based methods in surface reconstruction, pose estimation, and real-time performance.

Datasets. We evaluate PLGSLAM on a variety of scenes from different datasets.

- ScanNet dataset [13]. Real-world scenes with long sequences (more than 5000 images) and large-scale indoor scenarios. (**nearly** $7.5m \times 6.6m \times 3.5m$). We use this dataset for large-scale real-world indoor environments.
- Apartment dataset from [12]. Multi-rooms scene (**nearly** $10.8m \times 8.3m \times 3.2m$ with more than 6000 images). We use this dataset for multi-room environments.
- Our own indoor dataset for neural SLAM systems (NES dataset). This is the first real-world dataset for all kinds of neural SLAM systems with high-accuracy ground-truth for both camera trajectory and 3D reconstruction mesh. We collected single-agent and multi-agent datasets from various indoor environments, ranging from small room scenes (**nearly** $35m^2$) to large-scale scenes (>**1500** m^2), accumulating a total of 100,000 camera frames.

Metrics. We use Depth L1 (cm), Accuracy (cm), Completion (cm), and Completion ratio $(\%)$ to evaluate the reconstruction quality. Following Neural-RGBD and GO-Surf [23], [24], we perform frustum and occlusion mesh culling that removes unobserved regions outside frustum and the noisy points within the camera frustum but outside the target scene. For the evaluation of camera tracking, we adopt ATE RMSE and Mean(cm).

A. Experimental Results

Scene reconstruction and tracking. We evaluate our method on Replica [12], ScanNet [13], Apartment [12] and our own dataset. In Tab. I, we present the multi-agent reconstruction and collaborative camera tracking performance in ScanNet [13] and Apartment dataset [12]. We show the qualitative results in Fig. 4. Our method successfully achieves consistent completion as well as high-fidelity reconstruction results in large-scale indoor scenes.

B. Ablation Study

In this section, we conduct various experiments to verify the effectiveness of our method. Tab. II illustrates a quantitative evaluation with different settings.

Intra-Loop Closure We remove our intra loop closure in this experiment. Our full model leads to higher accuracy and better completion. The intra loop closure can significantly reduce the growing cumulative error and improve the robustness and accuracy of the single-agent camera tracking.

Inter-Loop Closure We remove our consistency loss of inter-loop clousre in this experiment. Inter-loop can effectively register different submaps into the global map.

Submap Fusion Submap fusion can significantly improve the integration of information among different maps, ensuring the consistency of the global map; otherwise, the map may exhibit numerous discontinuities or floating artifacts. This also suggests that handling smooth transitions between two submaps, especially when revisiting an inactive submap, is critical to the overall tracking quality.

IV. Conclusion

In this paper, we propose a novel multi-agent neural SLAM system, MNE-SLAM, which achieve accurate collaborate surface reconstruction and distributed pose estimation in small

Methods	Apartment-1			Apartment-2			Apartment-0		
	Part 1	Part 2	Average	Part 1	Part 2	Average	Part 1	Part 2	Average
	RMSE[cm]/Mean[cm]/Median[cm]			RMSE[cm]/Mean[cm]/Median[cm]			RMSE[cm]/Mean[cm]/Median[cm]		
CCM-SLAM[21]	2.12/1.94/1.74	9.31/6.36/5.57	5.71/4.15/3.66	0.51/0.45/0.40	0.48/0.43/0.38	0.49/0.44/0.39	-/-/-	-/-/-	-/-/-
ORB-SLAM3	4.93/4.65/5.01	4.93/4.04/3.80	4.93/4.35/4.41	1.35/1.05/0.65	1.36/1.24/1.11	1.36/1.15/0.88	0.67/0.58/0.47	1.46/1.11/0.79	1.07/0.85/0.63
Swarm-SLAM[22]	4.62/4.17/3.90	6.50/5.27/4.39	5.56/4.72/4.15	2.69/2.48/2.34	8.53/7.59/7.10	5.61/5.04/4.72	1.61/1.33/1.09	1.98/1.48/0.94	1.80/1.41/1.02
CP-SLAM[11]	6.21/5.56/5.27	5.67/5.37/4.67	5.94/5.46/4.97	1.45/1.43/1.39	2.48/2.32/2.27	1.97/1.88/1.83	0.62/0.47/0.30	1.28/1.17/1.37	0.95/0.82/0.84
Ours	**1.21/1.09/1.07**	**1.43/1.31/1.39**	**1.32/1.20/1.23**	**0.43/0.39/0.37**	**0.74/0.56/0.43**	**0.59/0.48/0.40**	**0.43/0.38/0.34**	**0.53/0.48/0.50**	**0.48/0.43/0.42**

TABLE I: Two-agent tracking performance in replica dataset [12]. ATE RMSE(↓), Mean(↓) and Median(↓) are used as evaluation metrics. Following the setting of [11], we quantitatively evaluated respective trajectories (part 1 and part 2) and average results of the two agents. ”-” indicates invalid results due to the failure of CCM-SLAM. Our method achieve SOTA performance compared with other existing methods.

Methods	Reconstruction[cm]			Localization[cm]	
	Acc.	Comp.	Comp.Ratio(%)	Mean	RMSE
w/o intra-loop.	18.94	4.37	72.37	6.18	6.49
w/o inter-loop	18.78	4.23	70.74	6.47	6.95
w/o submap	18.45	4.72	73.32	5.98	6.21
Ours	**17.42**	**3.94**	**76.48**	**5.72**	**5.98**

TABLE II: Ablation study. We conduct experiments on ScanNet dataset [13] to verify the effectiveness of our method. Our full model achieves better completion reconstructions and more accurate pose estimation results.

and large indoor scenes. Our distributed learning method enables our system to represent large-scale indoor scenes under communication bandwidth restriction. The intra-to-inter loop closure can effectively eliminate pose drifts and achieve global map consistency across multiple agents. Our extensive experiments demonstrate the effectiveness and accuracy of our system in both scene reconstruction, and pose estimation. The proposed neural SLAM dataset with high-accuracy 3D mesh and time-continuous camera trajectory can greatly advance the development of the community.

REFERENCES

[1] T. Deng, S. Liu, X. Wang, Y. Liu, D. Wang, and W. Chen, “Prosgnerf: Progressive dynamic neural scene graph with frequency modulated auto-encoder in urban scenes,” *arXiv preprint arXiv:2312.09076*, 2023.

[2] R. A. Newcombe, S. J. Lovegrove, and A. J. Davison, “Dtam: Dense tracking and mapping in real-time,” in *2011 international conference on computer vision*. IEEE, 2011, pp. 2320–2327.

[3] S. Izadi, D. Kim, O. Hilliges, D. Molyneaux, R. Newcombe, P. Kohli, J. Shotton, S. Hodges, D. Freeman, A. Davison *et al.*, “Kinectfusion: real-time 3d reconstruction and interaction using a moving depth camera,” in *Proceedings of the 24th annual ACM symposium on User interface software and technology*, 2011, pp. 559–568.

[4] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng, “Nerf: Representing scenes as neural radiance fields for view synthesis,” in *ECCV*, 2020.

[5] E. Sucar, S. Liu, J. Ortiz, and A. J. Davison, “imap: Implicit mapping and positioning in real-time,” in *ICCV*, October 2021, pp. 6229–6238.

[6] Z. Zhu, S. Peng, V. Larsson, W. Xu, H. Bao, Z. Cui, M. R. Oswald, and M. Pollefeys, “Nice-slam: Neural implicit scalable encoding for slam,” in *CVPR*, June 2022, pp. 12 786–12 796.

[7] M. M. Johari, C. Carta, and F. Fleuret, “Eslam: Efficient dense slam system based on hybrid representation of signed distance fields,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 17 408–17 419.

[8] Y. Zhang, F. Tosi, S. Mattoccia, and M. Poggi, “Go-slam: Global optimization for consistent 3d instant reconstruction,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 3727–3737.

[9] H. Wang, J. Wang, and L. Agapito, “Co-slam: Joint coordinate and sparse parametric encodings for neural real-time slam,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 13 293–13 302.

[10] T. Deng, G. Shen, T. Qin, J. Wang, W. Zhao, J. Wang, D. Wang, and W. Chen, “Plgslam: Progressive neural scene represenation with local to global bundle adjustment,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 19 657–19 666.

[11] J. Hu, M. Mao, H. Bao, G. Zhang, and Z. Cui, “Cp-slam: Collaborative neural point-based slam system,” *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[12] J. Straub, T. Whelan, L. Ma, Y. Chen, E. Wijmans, S. Green, J. J. Engel, R. Mur-Artal, C. Ren, S. Verma *et al.*, “The replica dataset: A digital replica of indoor spaces,” *arXiv preprint arXiv:1906.05797*, 2019.

[13] A. Dai, A. X. Chang, M. Savva, M. Halber, T. Funkhouser, and M. Niessner, “Scannet: Richly-annotated 3d reconstructions of indoor scenes,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[14] J. Sturm, N. Engelhard, F. Endres, W. Burgard, and D. Cremers, “A benchmark for the evaluation of rgb-d slam systems,” in *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2012, pp. 573–580.

[15] C. Yeshwanth, Y.-C. Liu, M. Nießner, and A. Dai, “Scannet++: A high-fidelity dataset of 3d indoor scenes,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 12–22.

[16] N. Keetha, J. Karhade, K. M. Jatavallabhula, G. Yang, S. Scherer, D. Ramanan, and J. Luiten, “Splatam: Splat track & map 3d gaussians for dense rgb-d slam,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 21 357–21 366.

[17] T. Deng, Y. Chen, L. Zhang, J. Yang, S. Yuan, D. Wang, and W. Chen, “Compact 3d gaussian splatting for dense visual slam,” *arXiv preprint arXiv:2403.11247*, 2024.

[18] B. Kerbl, G. Kopanas, T. Leimkühler, and G. Drettakis, “3d gaussian splatting for real-time radiance field rendering.” *ACM Trans. Graph.*, vol. 42, no. 4, pp. 139–1, 2023.

[19] S. Fridovich-Keil, A. Yu, M. Tancik, Q. Chen, B. Recht, and A. Kanazawa, “Plenoxels: Radiance fields without neural networks,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 5501–5510.

[20] Z. Teed and J. Deng, “Raft: Recurrent all-pairs field transforms for optical flow,” in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*. Springer, 2020, pp. 402–419.

[21] P. Schmuck and M. Chli, “Ccm-slam: Robust and efficient centralized collaborative monocular simultaneous localization and mapping for robotic teams,” *Journal of Field Robotics*, vol. 36, no. 4, pp. 763–781, 2019.

[22] P.-Y. Lajoie and G. Beltrame, “Swarm-slam: Sparse decentralized collaborative simultaneous localization and mapping framework for multi-robot systems,” *IEEE Robotics and Automation Letters*, vol. 9, no. 1, pp. 475–482, 2023.

[23] D. Azinović, R. Martin-Brualla, D. B. Goldman, M. Nießner, and J. Thies, “Neural rgb-d surface reconstruction,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022, pp. 6290–6301.

[24] J. Wang, T. Bleja, and L. Agapito, “Go-surf: Neural feature grid optimization for fast, high-fidelity rgb-d surface reconstruction,” in *2022 International Conference on 3D Vision (3DV)*, 2022, pp. 433–442.