

Di-NeRF: Distributed NeRF for Collaborative Learning with Unknown Relative Poses

Mahboubeh Asadi*, Kourosh Zareinia*, Sajad Saeedi*

Abstract—Collaborative mapping of unknown environments can be done faster and more robustly than a single robot. However, a collaborative approach requires a distributed paradigm to be scalable and deal with communication issues. This work presents a fully distributed algorithm enabling a group of robots to collectively optimize the parameters of a Neural Radiance Field (NeRF). The algorithm involves the communication of each robot’s trained NeRF parameters over a mesh network, where each robot trains its NeRF and has access to its own visual data only. Additionally, the relative poses of all robots are jointly optimized alongside the model parameters, enabling mapping with unknown relative camera poses. We show that multi-robot systems can benefit from differentiable and robust 3D reconstruction optimized from multiple NeRFs. Experiments on real-world and synthetic data demonstrate the efficiency of the proposed algorithm. See the website of the project for videos of the experiments¹.

I. INTRODUCTION

There is an increasing demand for robots to collaborate on complex tasks, such as mapping an unknown environment. Centralized approaches for the coordination of the robots or processing data face scalability issues, vulnerability to central node failures, and the risk of communication blackouts. A true collaborative robotic system needs to work in a distributed manner [1]. Further, building a high-quality representation in collaborative mapping is needed to make an informed decision. The representations also need to be compact for easy sharing. Neural Radiance Field (NeRF) [2] enables such a representation for a single robot, leveraging advancements in neural networks. Extending the capabilities of NeRF to facilitate multi-robot NeRF in a distributed manner emerges as a natural progression. This extension allows for the collaborative development of high-quality representations of unknown environments without dependence on a central node.

Performing distributed NeRF requires addressing several problems inherent to multi-robot systems [3]. Firstly, determining the relative poses of robots, preferably without requiring them to rendezvous. This is needed to enable the fusions of the local maps (i.e. maps of individual robots) into a global map. The relative poses can be determined either via line-of-sight rendezvous [4], which necessitates motion coordination, or by identifying overlaps within local maps [5], [6], requiring additional processing. In large-scale applications, avoiding line-of-sight rendezvous is preferred to prevent further planning constraints. The second challenge is deciding what information to share among robots. While sharing unprocessed raw visual data is resource-intensive and raises privacy concerns, sharing compact neural models is a more viable option. However, sharing maps introduces

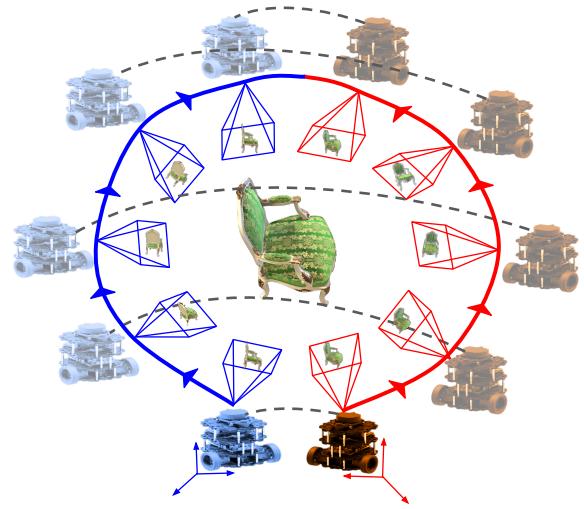


Figure 1. Di-NeRF allows robots to cooperatively optimize local copies of a neural network model without explicitly sharing visual data. In this figure, two robots use Di-NeRF to cooperatively optimize a unified NeRF. Each robot only sees part of the chair, and robots do not know their relative poses. The robots communicate over a wireless network (gray dashed lines) to cooperatively optimize the final network and relative poses.

complexity to the subsequent challenge. The third challenge is merging representations, requiring the development of a method to unify local neural maps into a global map. Merging neural representations is more intricate than merging geometric representations. These challenges must be systematically addressed in a distributed manner to ensure scalability and robustness against node failure.

There are many methods for distributed/centralized geometric multi-robot mapping and localization [7], [4], [8]. For learned methods, federated [9] and distributed learning [1] are common for overcoming the issues associated with centralized learning. In these methods, the training process is decentralized. Each robot performs its training, and the results are aggregated in a central node [9]. The need for a central node still constrains the scalability of such methods. A recent work using multiple agents for NeRF is Block-NeRF [10], composed of independent local maps, i.e. NeRF block. In Block-NeRF, the relative poses of the blocks are known to a degree and the task is done in a centralized manner. Another closely related work is DiNNO [11], where distributed mapping is performed using NeRF. DiNNO performs 2D mapping, and the assumption is that the relative poses of the robots are known.

In this paper, we present Di-NeRF (Distributed Neural Radiance Fields), an algorithm that builds on the Consensus Alternating Direction Method of Multipliers (C-ADMM) as a versatile distributed optimization method for multi-robot systems. In Di-NeRF, each robot starts building its NeRF by

Toronto Metropolitan University, mahboubeh.asadi@torontomu.ca

¹<https://sites.google.com/view/di-nerf/home>

relying on its own data. By integrating NeRFs from other robots, each robot will build a global NeRF model with a quality comparable to a fully centralized method. The advantages of Di-NeRF include the robots do not know their initial poses, and there is no robot rendezvous. The robots do not send raw data over wireless networks, protecting privacy and optimizing wireless bandwidth usage, as illustrated graphically in Fig. 1. In Di-NeRF, robots alternate between local optimization of an objective function and communication of intermediate network weights over the wireless network. Di-NeRF can consider different communication graphs (e.g. fully connected, circular, and ring connectivity). The key contributions of Di-NeRF include i) developing fully distributed optimization for 3D reconstruction with RGB images as input, and a NeRF as the backend, enabling fusing NeRFs in training not the rendering process, and ii) optimization of the relative poses of the robots, eliminating the need for a global coordinate system for all the robots and removing the need to know the prior relative poses.

The structure of the paper is as follows. In Sec. II, related work is presented. In Sec. III, we present Di-NeRF, followed by results and conclusions in Sec. IV and V.

II. LITERATURE REVIEW

Multi-robot localization and mapping algorithms utilize various representations such as sparse landmarks [12], [13], dense geometric maps [14], object classes [15], [16], and semantics [17]–[19]. With the advent of neural radiance field [2] and its variants [20]–[25], there are efforts to use neural maps for collaborative localization/mapping [10], [26], [27].

Emerging radiance fields, including non-neural representations such as Plenoxels [28], and neural representations, such as NeRF [2] have revolutionized localization and mapping algorithms. Nice SLAM [29], iMAP [30], NeRF-SALM [31], and [32] are using such representations. The early versions of NeRF representations [23], [25], [33]–[37] required pose information often generated via Structure-from-Motion (SfM) algorithms, e.g. COLMAP [38]. This limitation was later addressed in [20], [24], [39]–[42], by jointly optimizing camera poses and the scene. It was also shown that a camera view can be localized in a NeRF map, as shown in iNeRF [43].

Extending NeRF techniques to multi-robot scenarios has also been proposed in Block-NeRF [10], NeRFuser [44], and NeRFusion [45]. These methods use blending techniques to render an image, and often this is done in a centralized manner, performing inference on multiple NeRF models. Furthermore, it is assumed that there is prior knowledge about the relative pose of the robots. A similar research trend is federated learning [26], [27], [46], where the training is done in a decentralized manner, and each node learns a common NeRF in parallel. Then the weights are transferred to a server for aggregation. A closely related work is DiNNO [11], a distributed algorithm that enables multiple robots to collaboratively optimize the parameters of a neural network to learn a 2D map. Each robot maintains its version of a neural network model and has access only to its data. They communicate the models to achieve a consensus; however, prior knowledge of each robot's trajectory

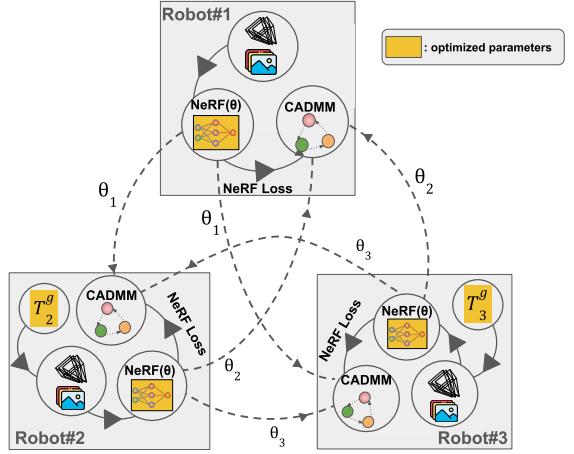


Figure 2. Three connected robots share their local NeRF to build a global NeRF. Each robot adjusts its weight compared to receiving weights using C-ADMM, and all robots except robot R^g (Robot#1, the robot whose coordinate is considered as the global coordinate) optimize for the relative pose T_i^g .

is needed for the algorithm to update the models. The core of the algorithm is a distributed optimization algorithm known as C-ADMM [1], [47].

III. DISTRIBUTED NEURAL RADIANCE FIELD (DI-NERF)

Here the problem is defined followed by the proposed method for relative pose estimation and distributed NeRF. The overall pipeline for Di-NeRF is shown in Fig. 2. Di-NeRF is a fully distributed algorithm for 3D reconstruction using RGB images. We assume the relative poses of the robots are unknown. With Di-NeRF, each robot has its local coordinate and only shares the learned models with other robots to achieve a consensus while optimizing for the relative poses.

A. Problem Statement

Given a set of images captured from a few viewpoints of a scene with the associated local camera parameters, the goal of Di-NeRF is to build a scene representation that enables all the robots to generate realistic images from novel viewpoints. Furthermore, since the poses of each robot are defined in a local coordinate system, it is necessary to perform relative pose optimization to build a consistent model that accurately represents the entire scene. The problem is expressed as:

$$\theta, \mathbf{T} = \underset{\theta, \mathbf{T}}{\operatorname{argmin}} \sum_{i \in \mathcal{V}} L_i(\theta, T_i^g, I_i, \pi_i), \quad \mathbf{T} = \{T_1^g, \dots, T_N^g\}, \quad (1)$$

where L_i represent the NeRF loss function corresponding to the robot i , θ denotes the parameters of the 3D reconstruction model [25]. The relative poses are denoted by \mathbf{T} , where the pose of the robot i relative to the global coordinate system is represented by $T_i^g = [R|t]$ within the space of $SE(3)$; g stands for global coordinate, which in this paper is the first robot coordinate (T_1^g is the identity matrix). N robots are connected with a communication graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ consisting of vertices $\mathcal{V} = \{1, \dots, N\}$ and edges \mathcal{E} which pairwise communication can occur. Here, the NeRF representation in [25] is used. Any other variants can also be used.

B. Relative Camera Pose Optimization

In multi-robot NeRF, robots do not send raw data over the wireless network. Therefore, using an SfM pipeline to calculate the relative poses for all robots is not easy. Here, the camera poses of each robot are expressed based on a local coordinate system, and gradually all poses transfer to the robot R^g coordinate system (throughout this paper, the robot R^g will be the arbitrary reference agent.). The optimization of the relative poses is performed jointly with the model parameters. Relative camera poses of multi-robot systems can be expressed as a camera-to-world transformation matrix (similar to each robot's local camera pose) $T_i^g = [R|t] \in SE(3)$, where $R \in SO(3)$ and $t \in \mathbb{R}^3$ show camera rotation and translation, respectively [20]. Optimizing the translation vector t involves the designation of trainable parameters, given its definition in Euclidean space. For camera rotation, which is in $SO(3)$ space, the axis-angle representation is chosen: $\phi := \alpha\omega, \phi \in \mathbb{R}^3$, where ω and α are a normalized rotation axis and a rotation angle, respectively. The rotation matrix R can be expressed from Rodrigues' formula:

$$R = I + \frac{\sin(\alpha)}{\alpha} \phi^\wedge + \frac{1 - \cos(\alpha)}{\alpha^2} (\phi^\wedge)^2, \quad (2)$$

where $(.)^\wedge$ is the skew operator that converts a vector ϕ to a skew matrix. The relative camera poses for each robot i are optimized by trainable parameters ϕ_i and t_i .

C. Distributed Formulation

In the problem presented in (1), each agent i has its local cost function L_i , indicating that the optimization must be done in a distributed manner on each agent. However, θ is a global variable and is shared among all agents. This necessitates that the agents should achieve a consensus on the optimal value of θ . In addition to the global variable θ , agents have their local variables, such as T_i^g . To express that the optimization distributed, (1) is rewritten as

$$\begin{aligned} \theta, \mathbf{T} &= \underset{\theta, \mathbf{T}}{\operatorname{argmin}} \sum_{i \in \mathcal{V}} L_i(\theta_i, T_i^g, I_i, \pi_i), \\ \text{s.t. } \theta_i &= z_{ij}, \forall j \in \mathcal{N}_i \end{aligned} \quad (3)$$

where each agent solves for its version of the global variable θ , θ_i , a relaxed local variable or primal variable. Assuming \mathcal{N}_i is the set of neighbors of robot i , to achieve a consensus between agents i and $\forall j \in \mathcal{N}_i$, an auxiliary variable z_{ij} is introduced. This is also known as the ‘complicating variable’ which ensures that the agents achieve a consensus via ‘complicating constraints’: i.e. $\theta_i = z_{ij}$ and $\theta_j = z_{ij}$. The new problem in (3) is then solved by introducing augmented Lagrangian and redefining the cost function to \mathcal{L}_i :

$$\mathcal{L}_i(\theta_i) = L_i(\theta_i) + (\theta_i - z_{ij})^\top y_i + \frac{\rho}{2} \sum_{j \in \mathcal{N}_i} \|\theta_i - z_{ij}\|_2^2, \quad (4)$$

where y_i is the Lagrangian multiplier or dual variable and ρ is penalty factor. The last two terms enforce that constraints are satisfied, but the penalty function ensures that around the optimal point, the objective function is quadratic.

There are various techniques to optimize (4), such as auxiliary problem principle (APP) [48] and alternating direction method of multipliers (ADMM) [49]. ADMM has improved convergence properties [49]. We use a version of ADMM that is suitable for distributed systems, known as consensus ADMM (C-ADMM) [50] described next.

D. Distributed Optimization of NeRF

The optimization through ADMM alternates between variables, and C-ADMM allows the agents to share the intermediate optimized variables to achieve a consensus. The optimization process is composed of three steps, in each step, other variables are treated as constants (superscript (k) denotes the step number in updating the variable):

- θ_i -minimization (z_{ij} and y_i are constants):

$$\min_{T_i^{g(k)}, \theta_i^{(k)}} L_i(\theta_i^{(k)}, T_i^{g(k)}) + \theta_i^{(k)\top} y_i^{(k)} + \frac{\rho}{2} \sum_{j \in \mathcal{N}_i} \|\theta_i^{(k)} - z_{ij}^{(k)}\|_2^2, \quad (5)$$

- z_{ij} -minimization (L_i , θ_i and y_i are constants):

$$\min_{z_{ij}^{(k)}} -z_{ij}^{(k)\top} y_i^{(k)} + \frac{\rho}{2} \sum_{j \in \mathcal{N}_i} \|\theta_i^{(k)} - z_{ij}^{(k)}\|_2^2, \quad (6)$$

- y_i -update (or dual variable update):

$$y_i^{(k)} \leftarrow y_i^{(k-1)} + \rho \sum_{j \in \mathcal{N}_i} (\theta_i^{(k)} - z_{ij}^{(k)}), \quad (7)$$

The optimization step alternates between minimizing the updated local objective function with respect to primal variables (θ_i) and maximizing the updated local objective function with respect to the dual variable. The z_{ij} -minimization step can be simplified to the following update (see appendix A):

$$z_{ij}^{(k+1)} = \frac{1}{N} \sum_{i \in \mathcal{V}} \theta_i^{(k)} := \bar{\theta}^{(k)}, \quad (8)$$

Therefore the θ_i -minimization becomes

$$\min_{T_i^{g(k)}, \theta_i^{(k)}} L_i(\theta_i^{(k)}, T_i^{g(k)}) + \theta_i^{(k)\top} y_i^{(k)} + \frac{\rho}{2} \sum_{j \in \mathcal{N}_i} \|\theta_i^{(k)} - \bar{\theta}^{(k)}\|_2^2. \quad (9)$$

Algorithm 1 summarizes the optimization procedures for local primal, dual variables, and relative poses for each robot. In line 1, The initial values for the NeRF parameters $\theta_i^{(0)}$ and the dual variable $y_i^{(0)}$ are both set to zero. The initial value for the relative poses - three rotation angles R and three translation displacements t - are explained in Sec. III-B. The first robot undergoes the NeRF training process during the first iteration (approximately 200 steps (lines 4-11)). Subsequently, it shares the learned weights with other robots. The remaining robots engage in collaborative optimization, simultaneously refining NeRF weights and adjusting relative camera poses in a distributed training framework based on the robot R^g weights (lines 12-19). This iterative process of communication and training persists until uniformity is achieved in both the weights and optimized relative camera poses across all robots. The relative pose between robot i coordinate system and robot 1 coordinate system (global coordinate) $T_i^g \in SE(3)$ is optimized jointly in the distributed training (line 14).

Algorithm 1 Di-NeRF Algorithm

```

1: Initialization:  $k \leftarrow 0, \theta_i^{(0)} \in \mathbb{R}^n, y_i^{(0)} = 0, T_i^g \in SE(3)$ 
2: Internal variables:  $y_i^{(k)}$ 
3: Public variables:  $\mathcal{Q}_i^{(k)} = \theta_i^{(k)}$ 
4: procedure ROBOT  $R^g$ 
5:   for  $i$  in  $\mathcal{V}$  do
6:      $\theta_i^{(k+1)} = \operatorname{argmin}_{\theta_i} \left\{ L_i(\theta_i) + \theta_i^\top y_i^{(k)} \right. \\ \left. + \rho \sum_{j \in \mathcal{N}_i} \left\| \theta_i - \frac{1}{2} (\theta_i^{(k)} + \theta_j^{(k)}) \right\|_2^2 \right\}$ 
7:     Communicate  $\mathcal{Q}_i^{(k)}$  to all  $j$  in  $\mathcal{N}_i$ 
8:     Receive  $\mathcal{Q}_j^{(k)}$  from all  $j$  in  $\mathcal{N}_i$ 
9:      $y_i^{(k+1)} = y_i^{(k)} + \rho \sum_{j \in \mathcal{N}_i} (\theta_i^{(k+1)} - \theta_j^{(k+1)})$ 
10:     $k \leftarrow k + 1$ 
11:   while stopping condition not satisfied do
12:   procedure OTHER ROBOTS
13:     for  $i$  in  $\mathcal{V}$  do
14:        $\theta_i^{(k+1)}, T_i^{g(k+1)} = \operatorname{argmin}_{\theta_i, T_i^g} \left\{ L_i(\theta_i, T_i^g) + \theta_i^\top y_i^{(k)} \right. \\ \left. + \rho \sum_{j \in \mathcal{N}_i} \left\| \theta_i - \frac{1}{2} (\theta_i^{(k)} + \theta_j^{(k)}) \right\|_2^2 \right\}$ 
15:       Communicate  $\mathcal{Q}_i^{(k)}$  to all  $j$  in  $\mathcal{N}_i$ 
16:       Receive  $\mathcal{Q}_j^{(k)}$  from all  $j$  in  $\mathcal{N}_i$ 
17:        $y_i^{(k+1)} = y_i^{(k)} + \rho \sum_{j \in \mathcal{N}_i} (\theta_i^{(k+1)} - \theta_j^{(k+1)})$ 
18:        $k \leftarrow k + 1$ 
19:   while stopping condition not satisfied do

```

E. Convergence Properties

The convergence of C-ADMM and its variants typically requires the dual variables' sum to converge to zero, a condition challenging in unreliable networks. Moreover, the nonlinearity and nonconvexity of neural networks specially NeRF, preclude guaranteed global solutions and linear rates [11], [50]. Despite these issues, Algorithm 1 has proven effective in practice for distributed NeRF training, as shown in Sec. IV, converging to solutions equivalent to centralized methods.

IV. EXPERIMENTS

This section presents the experimental results on synthetic and real-world datasets, including 1) comparing Di-NeRF and a centralized model based on several criteria, 2) examining the impact of the number of robots and communication graphs, 3) analyzing the impact of the amount of overlap between the robots' trajectories on convergence, and 4) evaluation using Waymo Block-NeRF dataset [10]. The architecture of NeRF is based on the instant neural graphics primitives [25]. Technical details are described in Sec IV-E.

For the relative poses, the initial estimates for the three translational components were set to 50 cm, and the three rotational components were offset by 10 degrees from the ground truth. These represent the maximum deviations that Di-NeRF has been capable of rectifying.

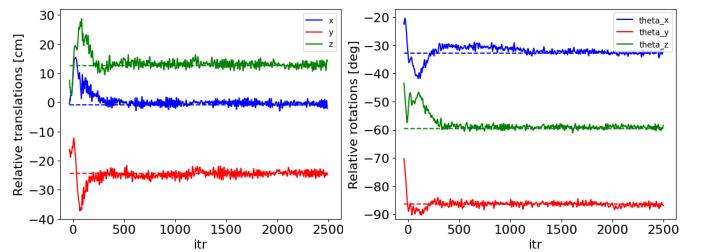


Figure 3. Optimizing the relative poses for two robots for chair from synthetic NeRF dataset. The relative translation and rotation estimates are plotted. The desired translation/rotation is marked by a dash line.

A. Di-NeRF vs Centralized NeRF on Synthetic Dataset

In this experiment, we use Di-NeRF to learn the RGB and density field of a three-dimensional environment in a distributed manner and compare the results with a centralized NeRF. The robots do not have access to a global coordinate frame, thus the relative poses are unknown. The datasets used are from the synthetic NeRF dataset [2], including Chair, Hotdog, and Lego sequences. We also use the long sequence of Barn from the real-world dataset of Tanks and Temples Benchmark [51]. For all the permutations, the configurations for centralized and Di-NeRF are the same, including the number of iterations (5000 steps). The number of images used in the centralized setup is around 100 to 150 for synthetic data and 300 for the Barn dataset. For multiple robots, the images are divided among the robots, in a controlled manner, such that the overlaps are controlled. In all experiments, we follow the official pre-processing procedures and train/test splits.

The metrics used to compare the performance of Di-NeRF with the centralized solution are the average rendering i) PSNRs, ii) SSIM, and iii) camera pose accuracy (in degrees and cm). The accuracy is determined by comparing the ground truth and the optimized pose values.

For the synthetic sequences, two robots are involved, and the dataset is split into two segments. For instance, in the case of the Chair sequence, robot R^1 exclusively observes the front of the chair. In contrast, robot R^2 focuses solely on the back (without any overlap in the robots' trajectories). A similar pattern is applied to other synthetic sequences. The whole data are divided into two parts for synthetic dataset and with COLMAP the poses were estimated for each segment. Therefore, the coordinate system for each robot is different, and the relative pose is estimated by running Di-NeRF. In Fig. 3, the convergence of the relative camera pose for chair dataset can be seen.

Qualitative results illustrating this setup can be found in Fig. 4. In Fig. 4-(first column), a sample of raw input images for two robots is shown. Since robot R^1 has no image from the back of the chair, it fails to reconstruct the back. The same is true for the robot R^2 reconstructing the front of the chair. This can be seen in Fig. 4-(middle column). Di-NeRF enables all robots to render the whole scene, see Fig. 4-(last column). Additionally, we present a comparison of PSNR and L2 loss in Fig. 5. The PSNR and L2 convergences for Di-NeRF and

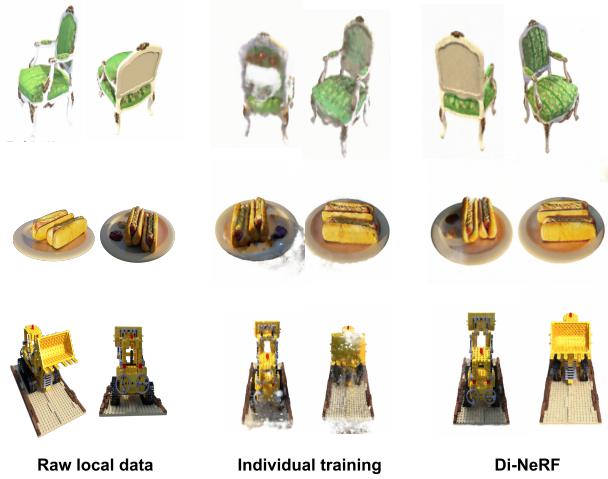


Figure 4. Di-NeRF can reconstruct the entire scene via the collaboration of two robots, where each robot only sees a part of the scene. Individual training results in poor reconstruction quality for some areas of the scene, whereas Di-NeRF maintains good quality throughout. In each column, the left image is for robot R^1 and the right one is for robot R^2 .

centralized training exhibit nearly identical behavior. Table I presents the quantitative results. In this table, the metrics separated by the / sign show the values of the metric for each robot. Overall, our distributed joint optimization model achieves similar quality compared to the centralized one.

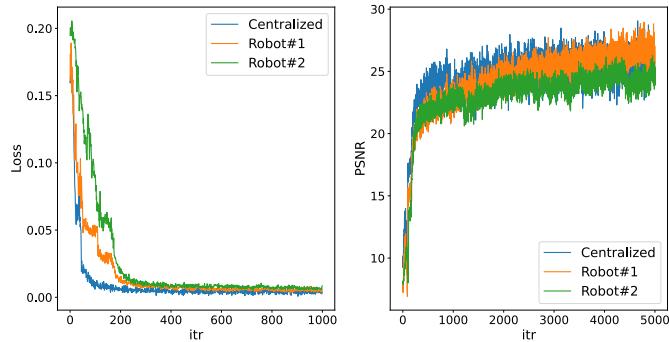


Figure 5. NeRF validation Loss and PSNR for chair dataset - Di-NeRF compared with centralized for two robots

Table I
IMAGE AND RELATIVE CAMERA POSE METRIC PERFORMANCES FOR TWO ROBOTS, LOWER/HIGHER BETTER, IS INDICATED BY \downarrow/\uparrow RESPECTIVELY.
PSNR AND SSIM ARE REPORTED FOR EACH ROBOT. THE RELATIVE POSE ACCURACY IS SHOWN FOR R^2 . R^1 IS THE GLOBAL FRAME.

	Di-NeRF	Cntr.	Di-NeRF	Cntr.	Rel. Pose Err. \downarrow	
	PSNR \uparrow		SSIM \uparrow		δR° δtcm	
Chair	31.25/31.31	33.02	0.922/0.926	0.952	0.70	1.02
Hotdog	29.40/29.32	32.63	0.924/0.914	0.963	0.61	0.93
Lego	28.15/28.17	29.67	0.929/0.935	0.941	0.55	0.89
Barn	28.13/28.13	29.94	0.879/0.869	0.894	0.63	0.95

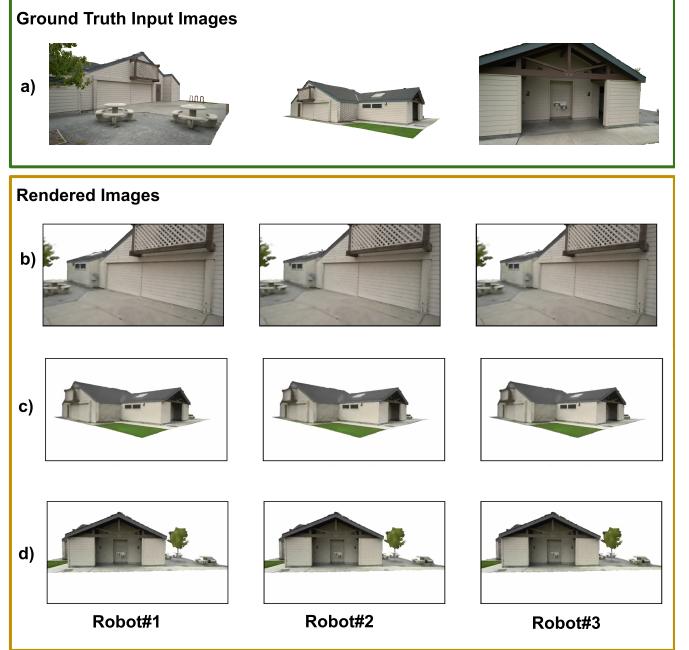


Figure 6. In row (a), input images, sourced from the Tanks and Temples dataset (Barn) are shown, which is provided to each robot. Rows (b, c, d) showcase rendered images generated by Di-NeRF. In these rows, the robots collaboratively communicate to reconstruct the complete scene. Notably, the views depicted in rows b, c, and d are strategically chosen—each is only visible in the raw data of a specific robot but can be rendered by all robots collectively in the final reconstruction.

B. Di-NeRF vs Centralized NeRF on Real-world Dataset

This experiment shows the performance of Di-NeRF not only on real-world data but also when the number of robots increases from two to five. This scenario applies to large-scale environments. Here, the Barn dataset from Tanks and Temples is used. There are 300 images in this sequence. Images are divided by the number of robots such that there is no common image in the local data of each robot, and the robot trajectories have zero overlaps. Fig. 6 shows the rendered images from each robot’s distributed trained NeRF. Each robot can render a novel view from any point. For instance, Robot R^3 has never seen the large bay doors but it is able to render an image when requested, shown in Fig. 6 (b).

Next, the number of robots in this experiment is altered. In Fig. 7, the trajectory for each robot is shown alongside the rendered images for each robot after collaborative training is completed. Each color represents the trajectory of a different robot. Table II presents the quantitative results. The rendering qualities and accuracy of relative camera pose in comparison to a centralized approach are evaluated. As anticipated with a distributed method, the outcomes closely resemble those achieved with the centralized approach.

In Table III, we experiment with different connectivity graphs for robots. Five robots are used with the Barn sequence from Tanks and Temples. In the star connection, R^1 is in the center. In Table III, the average result for 5 robots is shown.

Table II

IMAGE AND RELATIVE CAMERA POSE METRIC PERFORMANCES FOR DIFFERENT NUMBERS OF ROBOTS FOR BARN (TANKS AND TEMPLES DATASET). LOWER/HIGHER BETTER, IS INDICATED BY \downarrow/\uparrow RESPECTIVELY. PSNR AND SSIM ARE REPORTED FOR EACH ROBOT. THE RELATIVE POSE OPTIMIZATION ACCURACY FOR ALL ROBOTS IS SHOWN, EXCEPT FOR ROBOT R^1 WHICH IS CONSIDERED AS THE GLOBAL COORDINATE.

No. Robots	Di-NeRF		Centralized	Di-NeRF		Centralized	Relative Camera Pose Error	
	PSNR \uparrow	SSIM \uparrow		PSNR \uparrow	SSIM \uparrow		Δrot (deg) \downarrow	Δtran (cm) \downarrow
2	28.13/28.13	29.94		0.879/0.869	0.894		0.63	0.95
3	28.23/28.25/28.25	29.94		0.868/0.857/0.861	0.894		0.68/0.69	0.97/0.94
4	28.75/28.59/28.67/28.57	29.94		0.874/0.808/0.812/0.854	0.894		0.56/0.51/0.60	0.84/0.86/0.87
5	28.47/28.48/28.50/28.51/28.47	29.94		0.867/0.842/0.853/0.847/0.850	0.894		0.74/0.71/0.75/0.81	0.90/0.91/0.93/0.89

Table III

MEAN VALIDATION RESULT FOR DI-NERF AND DIFFERENT CONNECTIVITY GRAPHS. BARN FROM TANKS AND TEMPLES DATASET AND 5 ROBOTS ARE CHOSEN IN ALL CONNECTIVITY TYPES. DI-NERF WORKS WELL EVEN IF THE NETWORK OF ROBOTS IS NOT FULLY CONNECTED.

Communication	PSNR	SSIM	$\delta R^\circ \downarrow$	δt (cm) \downarrow
Fully Connected	28.49	0.867	0.74	0.90
Ring	28.28	0.847	0.84	0.87
Star	28.31	0.859	0.75	0.88
Line	28.17	0.855	0.91	0.97

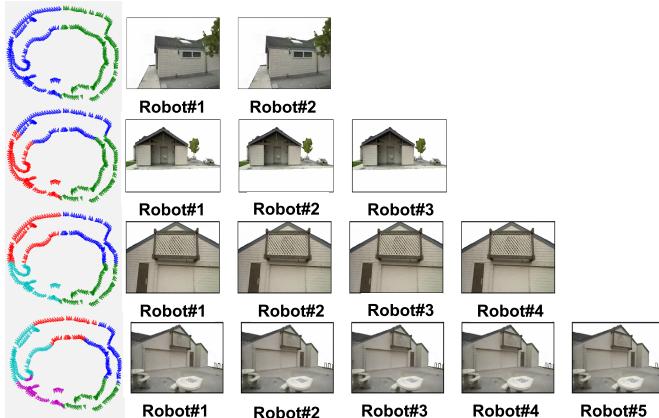


Figure 7. Di-NeRF for different numbers of robots. All robots are fully connected, and the relative poses and NeRFs are trained jointly. On the left side of the image, the allocation of frames and poses from the dataset to the different robots are shown in different colors. None of the robots have common images, but in the end, all robots can render the whole scene.

C. Trajectory Overlap Analysis

In this experiment, the sensitivity of the convergence of the relative poses of the robots with respect to the overlaps between the views of the robots is analyzed. For this experiment, two robots are considered. There is no common images between the robots, and the overlaps are only in the trajectories. Fig. 8 shows the top-down view of the overlaps, ranging from 5% to 40%, i.e. [5, 10, 20, 30, 40]%. The overlaps are determined using x and y coordinates of camera poses, assuming the poses are known in a coordinate frame. To divide the views among N robots, first, the $x - y$ plane containing the desired object (the Chair, in this case) is divided into N sectors, centered on the object. This will generate a 0% overlap for N robots, which then is extended to increase the overlap, as shown in Fig. 8.

For two robots, once the overlaps are determined, the poses of robot R^2 are manipulated to create a relative displacement with respect to R^1 , by translating them for 100 cm along each axis and rotating them for 30° around each axis. Once the data is created, optimization is performed, assuming an initial estimate for the relative pose with 60 cm for each translation component and 20° for each rotation.

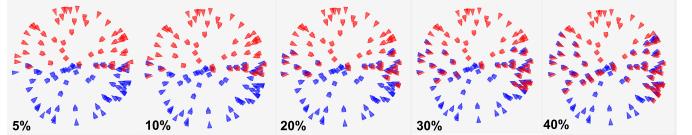


Figure 8. Segmenting the Chair sequence with different overlaps.

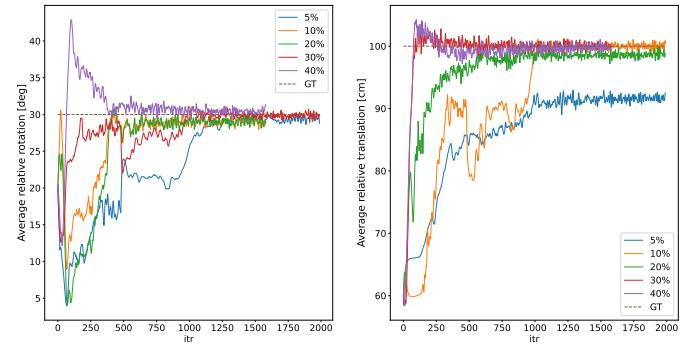


Figure 9. Optimizing the relative poses for two robots with varying overlaps. The average of translation and rotation estimates are plotted. The desired translation/rotation is marked by GT. Large overlaps improve the convergence.

Fig. 9 shows the convergence results for different overlaps, for 2000 steps of Di-NeRF training. The ground-truth relative poses are marked with GT. For each overlap, the plot shows the average estimates of translation and rotation components. The training configurations are the same for all overlaps. It is observed that large overlaps improve the convergence.

D. Waymo dataset - Unbounded Scenes

The San Francisco Mission Bay Dataset [10] is collected in an urban landscape with dynamic objects and reflective surfaces. It records 12,000 images with an array of 12 cameras, with different viewpoints, on a car for 100 seconds, over a distance of 1.08 km. For this experiment, a segment of the data, including 233 images from one camera (one that provides a complete surround view from the roof of the car),

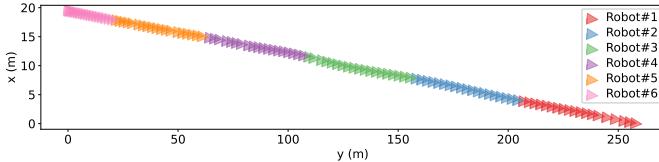


Figure 10. A sequence of the Mission Bay Dataset trajectories is divided into six segments, i.e. robots, and used to evaluate Di-NeRF.

measuring approximately 286 meters was selected and shown in Fig. 10. The sequence was divided into six segments to resemble a multi-agent setup. The primary objective of this experiment is not to surpass the state-of-the-art of Block-NeRF [10]. The goal here is to achieve comparable results between a distributed and a centralized system. In Fig. 11 the rendering result for this setup is shown for centralized and Di-NeRF training. All robots are able to render a view that they have never seen directly. The average values for the PSNR and SSIM metric are 25.10 and 0.814 over 6 robots and 25.21 and 0.825 for the centralized setup, In Table IV, the values for each robot are presented.

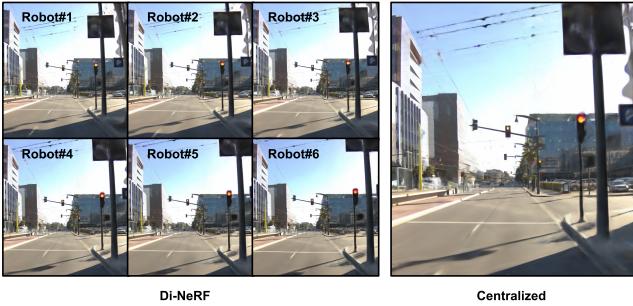


Figure 11. Rendered images from Di-NeRF on Mission Bay Dataset, all six robots demonstrated the capability to render similar images with quality comparable to that achieved through centralized processing.

Table IV

DI-NERF AND CENTRALIZED IMAGE METRIC PERFORMANCES FOR MISSION BAY DATASET FOR 6 ROBOTS, LOWER/HIGHER BETTER, IS INDICATED BY ↓/↑ RESPECTIVELY. PSNR AND SSIM ARE REPORTED FOR EACH ROBOT AND RELATIVE POSE OPTIMIZATION ACCURACY FOR ALL ROBOTS EXCEPT ROBOT R^1 WHICH IS CONSIDERED AS THE GLOBAL COORDINATE.

	R#1	R#2	R#3	R#4	R#5	R#6	Cntr.
PSNR↑	25.15	25.18	25.07	25.08	25.13	24.97	25.21
SSIM ↑	0.814	0.816	0.814	0.817	0.815	0.809	0.825
$\delta R^\circ \downarrow$	-	0.85	0.81	0.81	0.87	0.88	-
$\delta t(cm) \downarrow$	-	3.1	1.9	2.6	2.9	3.1	-

E. Technical Details of the Experiments

In this work, three datasets are used to evaluate Di-NeRF: The synthetic dataset, the Tanks and Temples, and the Waymo dataset. Consistency was maintained in these experiments concerning several parameters:

Batch Size: For all datasets is 2048.

Learning Rate: For all experiments is 0.01.

Regularization Parameter: The value of ρ in (9) is 0.001, which is the weight for the quadratic term and is the step size in the gradient ascent for the dual variable optimization.

Communication Rounds: The number of steps before a communication round is 10 for all robots.

Hardware and Software Configuration: The training was conducted on one NVIDIA RTX A5000 GPU with 24 GB of memory. We utilized PyTorch as our deep learning framework.

NeRF Network Configuration: The network F_θ has 8 layers, each with 256 channels. The hashmap size is 2^{19} .

Dataset Specifications: Image size for the Synthetic Dataset is 800×800 pixels, for Tanks and Temples is 1920×1080 pixels, and for the Waymo Dataset is 1217×1096 pixels. The rendering time for a fully connected graph is summarized in table V.

Table V
DI-NERF AND CENTRALIZED TRAINING TIME COMPARISON.

Dataset	Di-NeRF (sec/itr)	Centralized (sec/itr)	N_Robots
Synthetic Dataset	4.27	0.94	2
Tank and Temple	4.78	0.93	5
Mission Bay	4.73	0.95	6

V. CONCLUSION AND FUTURE WORK

We introduced the Di-NeRF algorithm, the first fully distributed NeRF used for multi-robot systems and 3D scenes, facilitating high-performance distributed training of Neural Radiance Fields (NeRF). We demonstrate its versatility across synthetic and real datasets. Di-NeRF harnesses the advantages of NeRF in distributed learning, relying solely on RGB input for relative pose estimation. Our analyses reveal its applicability to varying numbers of robots, diverse connection types, and distinct trajectory overlaps.

In this work, the relative extrinsic camera parameters are optimized. Optimizing the intrinsic camera parameters becomes particularly advantageous when dealing with various camera settings on different robots. Furthermore, the base NeRF model for each robot has limitations in handling unbounded scenes. A more advanced NeRF model, specially designed for unbounded scenes, can be seamlessly integrated into this project. Finally, in this study, we assume that the local camera poses are known but with different origins. By jointly optimizing NeRF and local camera poses, it becomes possible to eliminate the requirement for prior camera pose information.

VI. ACKNOWLEDGMENTS

We acknowledge the support of the Natural Sciences and Engineering Research Council of Canada (NSERC).

REFERENCES

- [1] T. Halsted, O. Shorinwa, J. Yu, and M. Schwager, “A survey of distributed optimization methods for multi-robot systems,” *arXiv preprint arXiv:2301.11361*, 2023.

- [2] 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,” *Communications of the ACM*, 2021.
- [3] S. Saeedi, M. Trentini, M. Seto, and H. Li, “Multiple-robot simultaneous localization and mapping: A review,” *JFR*, vol. 33, pp. 3–46, 2016.
- [4] R. Murai, J. Ortiz, S. Saeedi, P. H. J. Kelly, and A. J. Davison, “A robot web for distributed many-device localization,” *IEEE T-RO*, vol. 40, pp. 121–138, 2024.
- [5] N. Stathopoulos, A. Koval, A.-a. Agha-mohammadi, and G. Nikolakopoulos, “FRAME: Fast and robust autonomous 3d point cloud mapmerging for egocentric multi-robot exploration,” in *2023 IEEE ICRA*, 2023, pp. 3483–3489.
- [6] S. Saeedi, L. Paull, M. Trentini, M. Seto, and H. Li, “Group mapping: A topological approach to map merging for multiple robots,” *IEEE RA Magazine*, vol. 21, no. 2, pp. 60–72, 2014.
- [7] R. Murai, I. Alzugaray, P. H. Kelly, and A. J. Davison, “Distributed simultaneous localisation and auto-calibration using gaussian belief propagation,” *IEEE RA-L*, pp. 1–8, 2024.
- [8] Y. Tian and J. P. How, “Spectral sparsification for communication-efficient collaborative rotation and translation estimation,” *IEEE T-RO*, vol. 40, pp. 257–276, 2024.
- [9] T. Li, A. K. Sahu, A. Talwalkar, and V. Smith, “Federated learning: Challenges, methods, and future directions,” *IEEE SPM*, vol. 37, no. 3, pp. 50–60, 2020.
- [10] M. Tancik, V. Casser, X. Yan, S. Pradhan, B. P. Mildenhall, P. Srinivasan, J. T. Barron, and H. Kretzschmar, “Block-NeRF: Scalable large scene neural view synthesis,” in *2022 IEEE/CVF CVPR*, 2022, pp. 8238–8248.
- [11] J. Yu, J. A. Vincent, and M. Schwager, “DiNNO : Distributed neural network optimization for multi-robot collaborative learning,” *IEEE RA-L*, vol. 7, no. 2, pp. 1896–1903, 2022.
- [12] A. Cunningham, M. Paluri, and F. Dellaert, “DDF-SAM: Fully distributed slam using constrained factor graphs,” in *2010 IEEE/RSJ IROS*. IEEE, 2010, pp. 3025–3030.
- [13] A. Cunningham, V. Indelman, and F. Dellaert, “DDF-SAM 2.0: Consistent distributed smoothing and mapping,” in *2013 IEEE ICRA*. IEEE, 2013, pp. 5220–5227.
- [14] M. J. Schuster, K. Schmid, C. Brand, and M. Beetz, “Distributed stereo vision-based 6d localization and mapping for multi-robot teams,” *JFR*, vol. 36, no. 2, pp. 305–332, 2019.
- [15] S. Choudhary, L. Carbone, C. Nieto, J. Rogers, H. I. Christensen, and F. Dellaert, “Distributed mapping with privacy and communication constraints: Lightweight algorithms and object-based models,” *arXiv preprint arXiv:1702.03435*, 2017.
- [16] V. Tchouiev and V. Indelman, “Distributed consistent multi-robot semantic localization and mapping,” *IEEE RA-L*, vol. 5, no. 3, p. 4649–4656, Jul. 2020.
- [17] Y. Yue, C. Zhao, M. Wen, Z. Wu, and D. Wang, “Collaborative semantic perception and relative localization based on map matching,” in *2020 IEEE/RSJ IROS*. IEEE, 2020, pp. 6188–6193.
- [18] Y. Chang, Y. Tian, J. P. How, and L. Carbone, “Kimera-Multi: a system for distributed multi-robot metric-semantic simultaneous localization and mapping,” in *2021 IEEE ICRA*. IEEE, May 2021.
- [19] Y. Tian, Y. Chang, F. H. Arias, C. Nieto-Granda, J. P. How, and L. Carbone, “Kimera-multi: Robust, distributed, dense metric-semantic slam for multi-robot systems,” *IEEE T-OR*, vol. 38, no. 4, 2022.
- [20] Z. Wang, S. Wu, W. Xie, M. Chen, and V. A. Prisacariu, “NeRF—: Neural radiance fields without known camera parameters,” *arXiv preprint arXiv:2102.07064*, 2021.
- [21] L. Yen-Chen, P. Florence, J. T. Barron, A. Rodriguez, P. Isola, and T.-Y. Lin, “iNeRF: Inverting neural radiance fields for pose estimation,” in *2021 IEEE/RSJ IROS*. IEEE, 2021, pp. 1323–1330.
- [22] A. Yu, V. Ye, M. Tancik, and A. Kanazawa, “pixelNeRF: Neural radiance fields from one or few images,” in *IEEE/CVF CVPR*, 2021, pp. 4578–4587.
- [23] K. Park, U. Sinha, J. T. Barron, S. Bouaziz, D. B. Goldman, S. M. Seitz, and R. Martin-Brualla, “Nerfies: Deformable neural radiance fields,” in *Proceedings of the IEEE/CVF ICCV*, 2021, pp. 5865–5874.
- [24] W. Bian, Z. Wang, K. Li, J.-W. Bian, and V. A. Prisacariu, “Nope-NeRF: Optimising neural radiance field with no pose prior,” in *Proceedings of the IEEE/CVF CVPR*, 2023, pp. 4160–4169.
- [25] T. Müller, A. Evans, C. Schied, and A. Keller, “Instant neural graphics primitives with a multiresolution hash encoding,” *ACM Trans. Graph.*, vol. 41, no. 4, pp. 102:1–102:15, Jul. 2022.
- [26] L. Holden, F. Dayoub, D. Harvey, and T.-J. Chin, “Federated neural radiance fields,” *arXiv preprint arXiv:2305.01163*, 2023.
- [27] T. Suzuki, “Federated learning for large-scale scene modeling with neural radiance fields,” *arXiv preprint arXiv:2309.06030*, 2023.
- [28] Sara Fridovich-Keil and Alex Yu, M. Tancik, Q. Chen, B. Recht, and A. Kanazawa, “Plenoxels: Radiance fields without neural networks,” in *CVPR*, 2022.
- [29] 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 *IEEE/CVF CVPR*, June 2022.
- [30] E. Sucar, S. Liu, J. Ortiz, and A. Davison, “iMAP: Implicit mapping and positioning in real-time,” in *ICCV*, 2021.
- [31] A. Rosinol, J. J. Leonard, and L. Carbone, “NeRF-SLAM: Real-time dense monocular slam with neural radiance fields,” in *2023 IEEE/RSJ IROS*. IEEE, 2023, pp. 3437–3444.
- [32] A. L. Teigen, Y. Park, A. Stahl, and R. Mester, “RGB-D mapping and tracking in a plenoxel radiance field,” in *IEEE/CVF WACV*, 2024, pp. 3342–3351.
- [33] S. J. Garbin, M. Kowalski, M. Johnson, J. Shotton, and J. Valentin, “FastNeRF: High-fidelity neural rendering at 200fps,” in *Proceedings of the IEEE/CVF ICCV*, 2021, pp. 14 346–14 355.
- [34] K. Wadhwan and T. Kojima, “SqueezeNeRF: Further factorized Fast-NeRF for memory-efficient inference,” in *Proceedings of the IEEE/CVF CVPR*, 2022, pp. 2717–2725.
- [35] J. Chibane, A. Bansal, V. Lazova, and G. Pons-Moll, “Stereo radiance fields (SRF),” in *Proceedings of the IEEE/CVF CVPR*, 2021, pp. 7911–7920.
- [36] L. Liu, J. Gu, K. Zaw Lin, T.-S. Chua, and C. Theobalt, “Neural sparse voxel fields,” *NeurIPS*, vol. 33, pp. 15 651–15 663, 2020.
- [37] A. Yu, R. Li, M. Tancik, H. Li, R. Ng, and A. Kanazawa, “Plenotrees for real-time rendering of neural radiance fields,” in *Proceedings of the IEEE/CVF ICCV*, 2021, pp. 5752–5761.
- [38] J. L. Schönberger and J.-M. Frahm, “Structure-from-Motion Revisited,” in *CVPR*, 2016.
- [39] Y. Jeong, S. Ahn, C. Choy, A. Anandkumar, M. Cho, and J. Park, “Self-calibrating neural radiance fields,” in *ICCV*, 2021.
- [40] C.-H. Lin, W.-C. Ma, A. Torralba, and S. Lucey, “BARF: Bundle-adjusting neural radiance fields,” in *IEEE/CVF ICCV*, 2021, pp. 5741–5751.
- [41] S. Chen, Y. Zhang, Y. Xu, and B. Zou, “Structure-aware NeRF without posed camera via epipolar constraint,” 2022.
- [42] K. Park, P. Henzler, B. Mildenhall, J. T. Barron, and R. Martin-Brualla, “CampP: Camera preconditioning for neural radiance fields,” *ACM TOG*, vol. 42, no. 6, pp. 1–11, 2023.
- [43] L. Yen-Chen, P. Florence, J. T. Barron, A. Rodriguez, P. Isola, and T.-Y. Lin, “iNeRF: Inverting neural radiance fields for pose estimation,” in *IEEE/RSJ IROS*, 2021.
- [44] J. Fang, S. Lin, I. Vasiljevic, V. Guizilini, R. Ambrus, A. Gaidon, G. Shakhnarovich, and M. R. Walter, “NeRFuser: Large-scale scene representation by nerf fusion,” *arXiv preprint arXiv:2305.13307*, 2023.
- [45] X. Zhang, S. Bi, K. Sunkavalli, H. Su, and Z. Xu, “NeRFusion: Fusing radiance fields for large-scale scene reconstruction,” in *IEEE/CVF CVPR*, 2022, pp. 5449–5458.
- [46] J. Hu, M. Mao, H. Bao, G. Zhang, and Z. Cui, “CP-SLAM: Collaborative neural point-based slam system,” *arXiv preprint arXiv:2311.08013*, 2023.
- [47] D. McGann, K. Lassak, and M. Kaess, “Asynchronous distributed smoothing and mapping via on-manifold consensus ADMM,” *arXiv preprint arXiv:2310.12320*, 2023.
- [48] A. Losi and M. Russo, “On the Application of the Auxiliary Problem Principle,” *Journal of Optimization Theory and Applications*, vol. 117, no. 2, pp. 377–396, 2003.
- [49] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, “Distributed optimization and statistical learning via the alternating direction method of multipliers,” *Foundations and Trends in Machine Learning*, vol. 3, no. 1, pp. 1–122, 2011.
- [50] O. Shorinwa, T. Halsted, J. Yu, and M. Schwager, “Distributed optimization methods for multi-robot systems: Part II—a survey,” *arXiv preprint arXiv:2301.11361*, 2023.
- [51] A. Knapitsch, J. Park, Q.-Y. Zhou, and V. Koltun, “Tanks and Temples: Benchmarking large-scale scene reconstruction,” *ACM TOG*, vol. 36, no. 4, 2017.

- [52] S. Boyd, N. Parikh, E. Chu, B. Peleato, J. Eckstein *et al.*, ‘‘Distributed optimization and statistical learning via the alternating direction method of multipliers,’’ *Foundations and Trends® in Machine learning*, vol. 3, no. 1, pp. 1–122, 2011.

APPENDIX

A. The Auxiliary parameter in Di-NeRF [52]

Considering the distributed optimization in Di-NeRF that is subject to the constraint $\theta_i = z_{ij}$, the following sub-problem optimization can be written for each robot as follows using the auxiliary variable z_{ij} :

$$\begin{aligned} \theta_i^{(k+1)}, T_i^{g(k+1)} &= \underset{\theta_i, T_i^g}{\operatorname{argmin}} \left\{ L_i(\theta_i, T_i^g) + \theta_i^\top y_i^{(k)} \right. \\ &\quad \left. + \rho \sum_{j \in \mathcal{N}_i} \left\| \theta_i - z_{ij}^{(k)} \right\|_2^2 \right\}, \end{aligned} \quad (10)$$

where L_i is the local loss function for each robot and θ_i is the network parameters of the robot i which is communicating with the robot j . The parameter ρ is the weight for the quadratic term $\sum_{j \in \mathcal{N}_i} \|\theta_i - z_{ij}\|_2^2$ and the step size in the gradient ascent of the dual variable y_i . \mathcal{N}_i is the set of neighbours of robot i , and T_i^g is the relative pose of robot i with respect to the global coordinate. According to (10), the local loss function for each robot can be updated according to the following:

$$\mathcal{L}_i = L_i(\theta_i, T_i^g) + \theta_i^\top y_i + \frac{\rho}{2} \sum_{j \in \mathcal{N}_i} \|\theta_i - z_{ij}\|_2^2, \quad (11)$$

where \mathcal{L}_i is the updated loss function for each robot based on Di-NeRF. The dual variable y_i update according to C-ADMM is as follows:

$$y_i^{(k+1)} = y_i^{(k)} + \rho(\theta_i^{(k+1)} - z_{ij}^{(k+1)}), \quad (12)$$

To update the auxiliary parameter z_{ij} given $\theta_i - z_{ij} = 0$ and (10), the z update subproblem is as follows:

$$\min_{z^{(k+1)}} - \sum_{i \in \mathcal{N}_i} y_i^{(k)} z^{(k+1)} + \frac{1}{2} \sum_{i \in \mathcal{N}_i} \left\| \theta_i^{(k+1)} - z^{(k+1)} \right\|_2^2 \quad (13)$$

By taking the first derivative from (13), the following equation can be driven for z_{ij} :

$$z^{(k+1)} = \frac{1}{N} \sum_{i \in \mathcal{N}_i} \left(\theta_i^{(k+1)} + \frac{1}{\rho} y_i^{(k)} \right) \quad (14)$$

Dual variable update is an equality, not an optimization problem and is called a central collector or fusion center. Equation (14) can be simplified further by writing it as:

$$z^{(k+1)} = \bar{\theta}^{(k+1)} + \frac{1}{\rho} \bar{y}^{(k)} \quad (15)$$

Substitution of (15) into the average value of y_i over \mathcal{N}_i in (12) (i.e. $\bar{y}_i^{(k+1)} = \bar{y}_i^{(k)} + \rho(\bar{\theta}_i^{(k+1)} - z^{(k+1)})$), yields $\bar{y}_i^{(k+1)} = 0$. Further substitution of this result into (15) leads to the following final equation:

$$z_{ij}^{(k+1)} = \frac{1}{N} \sum \theta_i^{(k)} := \bar{\theta}^{(k)}, \quad (16)$$

During each iteration, k , every robot, indexed by i , independently solves its own subproblem to determine the value of the global variable $\theta_i^{(k)}$. The robot incurs a penalty proportional to the deviation of its variable from the mean value of the global variable, as computed from all robots in the preceding iteration.