2026 IEEE International Conference on Robotics and Automation (ICRA 2026)
June 1-5, 2026. Vienna, Austria

Robust SG-NeRF: Robust Scene Graph Aided Neural Surface Reconstruction

Yi Gu[1,*], Dongjun Ye[1,*], Zhaorui Wang[1,*], Jiaxu Wang[1], Jiahang Cao[1], Mingle Zhao[2], Renjing Xu[1†]

***Abstract*— Neural surface reconstruction relies heavily on accurate camera poses as input. Despite utilizing advanced pose estimators like COLMAP or ARKit, camera poses can still be noisy. Existing pose-NeRF joint optimization methods handle poses with small noise (inliers) effectively but struggle with large noise (outliers), such as mirrored poses. In this work, we focus on mitigating the impact of outlier poses. Our method integrates an inlier-outlier confidence estimation scheme, leveraging scene graph information gathered during the data preparation phase. Unlike previous works directly using rendering metrics as the reference, we employ a detached color network that omits the viewing direction as input to minimize the impact caused by shape-radiance ambiguities. This enhanced confidence updating strategy effectively differentiates between inlier and outlier poses, allowing us to sample more rays from inlier poses to construct more reliable radiance fields. Additionally, we introduce a re-projection loss based on the current Signed Distance Function (SDF) and pose estimations, strengthening the constraints between matching image pairs. For outlier poses, we adopt a Monte Carlo re-localization method to find better solutions. We also devise a scene graph updating strategy to provide more accurate information throughout the training process. We validate our approach on the SG-NeRF and DTU datasets. Experimental results on various datasets demonstrate that our methods can consistently improve the reconstruction qualities and pose accuracies. Project page: https://rsg-nerf.github.io/RSG-NeRF/.**

I. Introduction

Reconstructing the surfaces of objects from multi-view images is a fundamental challenge in both computer vision and computer graphics. Inspired by Neural Radiance Fields [1] (NeRF), recent strides [2], [3], [4], [5] have marked significant progress in neural surface reconstruction (NSR) area by leveraging implicit scene representations and volume rendering techniques. In NSR, scene geometry is encoded through a signed distance function (SDF), which is learned by a multilayer perceptron (MLP) network trained with an image-based rendering loss. Despite these promising advancements, a key challenge in NSR involves the dependency on accurate camera poses. In practice, NeRF and its variants often rely on COLMAP [6], [7], a widely-used Structure from Motion (SfM) framework, to estimate camera poses prior. Unfortunately, these pose estimations can be significantly erroneous, adversely affecting the reconstruction quality of NeRF. Consequently, recent efforts [8], [9], [10], [11], [12], [13], [14], [15], [16] have aimed to joint optimize scene representations and camera poses to minimize the impact of pose errors. Nevertheless, most of these efforts concentrate on refining relatively small pose errors (referred to as inliers). It is still a challenge to rectify noticeably incorrect camera poses (referred to as outliers). To alleviate the negative effects of outliers, SG-NeRF [17] introduces scene graphs to enhance camera pose optimization for improved geometric reconstruction. The main contribution of SG-NeRF lies in estimating the confidence of each camera pose. By prioritizing ray sampling from images with high confidence poses, SG-NeRF can recover reliable geometry, even in the presence of numerous outliers, as illustrated in the left of Fig. 1. Theoretically, the extreme of the SG-NeRF philosophy is not sampling on outliers. Thus, it is important to recognize the inliers and outliers. However, the heuristic confidence updating strategy in SG-NeRF only depends on the peak signal-to-noise ratio (PSNR) index, which can not well reflect the differences between the inliers and outliers. As shown in Fig. 2, with the wrongly estimated poses, SG-NeRF can still render images with high PSNR. This is a classical shape-radiance ambiguity problem in NeRF series [18], [19], [20]. A potential solution is to add some regularization terms to the loss function, but it may be more complicated coupled with a joint pose-NeRF optimization process.

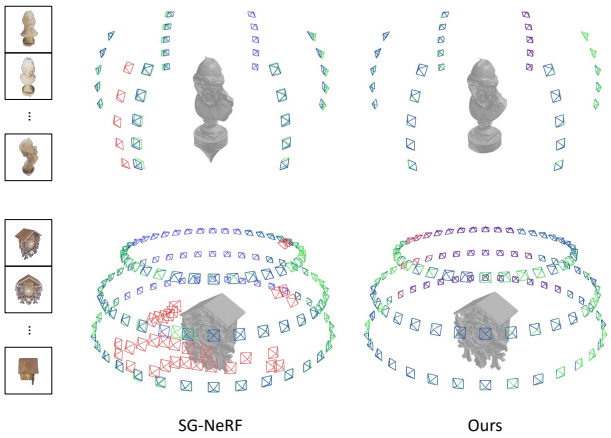

Fig. 1: Reconstruction results on the SG-NeRF [17] dataset. Both SG-NeRF [17] and our method take the same initial poses as input, including significant noises. The camera poses are also presented with optimized outlier poses, inlier poses and ground truth poses.

To address this problem, we explore an improved confi-

* Authors contributed equally to this work.
[1] The Hong Kong University of Science and Technology (Guangzhou)
[2] University of Macau
† Corresponding author renjingxu@hkust-gz.edu.cn

dence estimation method to distinguish inliers and outliers. As shown in the last column of Fig. 2, we empirically find that a color network without viewing direction as input can provide more reliable information about pose confidence, albeit possibly at the expense of rendering and geometric quality. It is important to note that our primary objective is to identify inliers and outliers based on estimated confidence. To achieve this, our framework incorporates two color networks: one that aligns with traditional NSR approaches, and another is dedicated to confidence estimation. We detach the latter from the main pose-NeRF computational graph to maintain the performance of the NSR backbone. Typically, the color network used in NSR is a shallow MLP [18], [21], [2], which means that our method does not substantially increase computational costs. This straightforward design allows us to establish a rule-based threshold to identify inliers and outliers. Subsequently, we can enhance the final results by integrating tailored designs for handling each. We present two strategies: for outliers, we employ a Monte Carlo re-localization method to provide better initialization; for inliers, we enhance constraints with re-projection and Intersection-of-Union (IoU) losses. Additionally, we devise a scene graph updating strategy based on the current SDF to eliminate incorrectly matched pairs. Experiments on the SG-NeRF [17] and the DTU [22] datasets generally show that our method not only yields high-quality 3D reconstructions but also effectively corrects outlier poses, as illustrated in the right of Fig. 1. Our contributions are summarized as follows:

- We propose a plug-and-play pose confidence estimation method that effectively identifies inliers and outliers.
- We introduce Monte Carlo re-localization to handle outliers and re-projection and IoU losses for inliers to improve geometric constraints.
- Additionally, we implement a scene graph updating strategy to enhance the training guidance.

II. Related Works

Neural Surface Reconstruction. Traditional multi-view stereo methods [23], [7] explicitly establish dense correspondences across multiple images to generate depth maps, which are subsequently fused into a global dense point cloud [24], [25]. Surface reconstruction is typically performed as a post-processing step, employing techniques such as screened Poisson surface reconstruction [26]. The processes of searching for correspondences and estimating depth have been significantly enhanced by deep learning-based approaches [27], [28]. Recently, the implicit representation has gained a lot of attention due to its continuity and capability to achieve high spatial resolutions. Building on the pioneering work of NeRF [21], many successors [29], [30], [3] integrate the signed distance function (SDF) into NeRF to enhance geometric modeling. Among these, NeuS [2] is particularly noteworthy for its ability to produce high-quality reconstructions and successfully handle scenes with severe occlusions and complex structures. Thus, in this study, we select NeuS to represent our scenes.

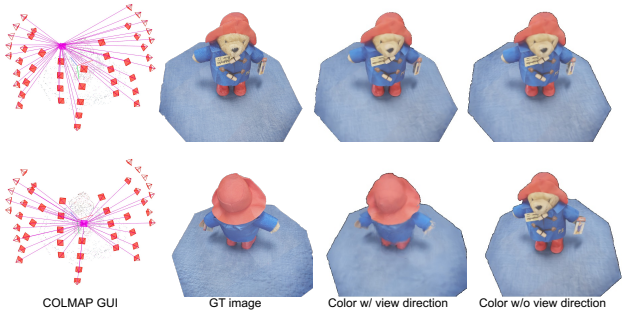

Fig. 2: The illustration of the pose ambiguity. The first row is the results from inliers and the second row presents outliers. Images in the first column come from COLMAP [6] GUI, which show that both these two poses are registered in front of the object. However, the ground truth images in the second column show the opposite phenomenon. The third column presents the rendering results of SG-NeRF [17], which use the same color network as NeuS [2] with view direction as input. As shown in the fourth column, our method incorporates an isolated color network, which can well recognize this ambiguity.

Structure from Motion (SfM) and (re-)localization. NeRF and its variants require accurate camera poses as input [31], [21], [2]. In real-world applications, Structure from Motion (SfM) [32], [6], [33], [34], [35], [36], [37] techniques are commonly employed for data pre-processing. SfM organizes a set of unstructured images by estimating camera poses and triangulating 3D scene points. An essential byproduct of this process is the scene graph, which captures information about matching pairs. However, current advanced SfM frameworks primarily depend on keypoint detection [38], [39], [40] and matching [41], [42], [43] techniques, which can be less effective in textureless or repetitive environments.

The task of (re)localization [44], [45], [46], [47] is also closely related to SfM. Given a database of posed images, the goal of this task is to estimate the camera poses of newly captured images. In the context of NeRF with re-localization, most existing studies [48], [49], [50], [51] concentrate on relocating new images within well-constructed NeRFs. In our approach, we implement the Monte Carlo re-localization method during the training phase to improve the robustness and accuracy of outlier poses.

Joint NeRF and pose optimization. NeRFmm [8] and iNeRF [51] demonstrate the potential for jointly learning or refining camera parameters alongside the NeRF framework. Following works [52], [9], [53], [54], [16] also perform different modular modifications. For example, GARF [11] and SiNeRF [55] capitalize on the inherent smoothness of non-traditional activations to mitigate the impact of noisy gradients caused by high-frequency components in positional embeddings. L2G-NeRF [13] and Invertible Neural Warp [56] tackle the camera pose representation with an overparameterization strategy. NoPe-NeRF [12] employs an external monocular depth estimation model to assist in

refining camera poses. Some works [14], [16], [10], [17] also incorporate cross-view correspondences to enhance geometry constraints. Commonly, most approaches presume that all images are properly posed initially and focus on local optimizations for pose correction. SG-NeRF [17] is the first method to employ a scene graph for guiding joint optimization. Building upon this approach, we propose an alternative strategy for confidence estimation.

III. Methods

Problem statement. Our research focuses on the object-level 3D surface reconstruction from a set of unorganized images captured in an inward-facing configuration. Specifically, given a collection of RGB images $\mathbf{I} = \{I_1, I_2, ..., I_N\}$, our objective is to reconstruct the 3D surface S of the scene. For a specific image I_i, a key output of our approach is the optimized camera pose $P_i = (R_i, t_i)$, where R_i belongs to $\mathbf{SO}(3)$ representing the rotation and t_i is a vector in $\mathbb{R}^3$ representing the translation. Additionally, each pose is assigned an inlier-outlier confidence score.

Method overview. Fig. 3 illustrates the workflow of the proposed pipeline. In the data preparation stage, we first employ a widely-used Structure-from-Motion (SfM) algorithm, specifically COLMAP [6], to obtain the initial camera poses. Given the potential inaccuracies of these poses, proceeding with a direct joint pose-NeRF optimization could be catastrophic. To mitigate this risk, we leverage scene graph information to guide the training process (Sec. III-A). We update the confidence with our tailored indicator, which can effectively distinguish inlier and outlier poses. For inliers, we introduce additional constraints to enhance the geometric consistency (Sec. III-B). For outliers, we utilize Monte Carlo re-localization to find better initializations (Sec. III-A). Additionally, We also devise a scene graph updating strategy to enhance the guidance during training (Sec. III-D).

A. Scene graph guided confidence estimation

A scene graph $\mathbf{G} = (\mathbf{V},\ \mathbf{E})$ in SfM consists of a set of nodes $\mathbf{V}$ and edges $\mathbf{E}$. Each node $v_i \in \mathbf{V}$ corresponds to an input image $I_i \in \mathbf{I}$, and an edge between two nodes contains the matching and co-visibility information about the corresponding images. We annotate all edges as $\mathbf{M} = \{M_{i,j} | v_i, v_j \in \mathbf{V},\ v_i \neq v_j\}$, where the set $M_{i,j}$ comprises all matched keypoint pairs between I_i and I_j.

The original scene graph tends to be dense and contains many incorrect matches. Following SG-NeRF [17], we set an angular threshold τ for the estimated relative rotations and remove any edges exceeding τ. Then, each node is assigned a confidence estimate based on this sparsified scene graph.

The confidence score for a node v_i is defined as the mean number of matching pairs, which can be computed as:

$$CS(v_i) = \frac{\sum_{M_{i,j} \in \mathbf{M}_i} |M_{i,j}|}{|\mathbf{M}_i|}, \tag{1}$$

where $|\cdot|$ denotes the number of elements in a set, e.g., $|M_{i,j}|$ is the total number of matching pairs of I_i and I_j and $|\mathbf{M}_i|$ is the total number of edges of v_i. A higher score indicates that the image has a better matching quality and a higher likelihood of being an inlier. We normalize this confidence score via $CS(v_i) = CS(v_i) / \sum_{v \in \mathbf{V}} CS(v)$ to form a probability distribution, which guides the training to sample more rays from poses with higher confidence. All confidence computations involve a normalization step and we omit this step in the following text for brevity.

These initial scores are derived from keypoint matches, which might lack a comprehensive understanding of the information contained in images. Thus, we adaptively update the confidence scores based on the image rendering quality. Specifically, we estimate PSNR for each image according to current image rendering loss for efficiency. Then, the confidence scores are updated by [17]:

$$CS(v_i) = CS(v_i) + \lambda_c PSNR(v_i). \tag{2}$$

However, as shown in Fig. 2, we empirically found that the PSNR of outliers can be even larger than that of inliers, which means that more outliers will be sampled. The reason behind this phenomenon comes from shape-radiance ambiguity [18], [8], [19], [20].

To solve this problem, we employ a new color network C_n which does not take viewing direction as input. To mitigate the shape-radiance ambiguity and prevent overfitting, we use the same sampling points and geometry features as the original color network C_o. We detached all relevant computations from C_n to ensure that the loss from C_n does not impact the main networks. Since we only require an indicator that can reflect the relative rendering qualities of the training images, the PSNR estimated by C_n can serve the same purpose as that by C_o. As highlighted in NeRF++ [18], most existing works [21], [2] use a shallow MLP for color network, which acts as an implicit regularization. Thus, C_n will not introduce significant computational overhead. We use the PSNR estimated from C_n (denoted as $PSNR_n$) to update the confidence score throughout the training process.

It should be noted that we also keep a record of the PSNR with C_o (donated as $PSNR_o$), which can be helpful for filtering out outliers. When $PSNR_o$ and $PSNR_n$ show a significant discrepancy, it is an indication of anomalous data. Therefore, we recognize poses with $|PSNR_o - PSNR_n| > \tau_1$ as outliers.

B. Joint optimization

We build up our framework based on NeuS [2]. The neural surface reconstruction loss function is defined as follows:

$$\mathcal{L}_{NSR} = \mathcal{L}_{color}(C_o) + \mathcal{L}_{color}(C_n) + \lambda \mathcal{L}_{reg}. \tag{3}$$

The $\mathcal{L}_{color}(C_o)$ represents calculating $\mathcal{L}_{color}$ by C_o. The $\mathcal{L}_{color}(C_n)$ is calculated by C_n, with gradients only back-propagated to C_n. The $\mathcal{L}_{color}$ is a photometric loss:

$$\mathcal{L}_{color} = \left\| \hat{\mathbf{C}} - \mathbf{C} \right\|_1, \tag{4}$$

where $\hat{\mathbf{C}}$ is obtained by volume rendering equation [21], [57] and $\mathbf{C}$ is the ground truth color.

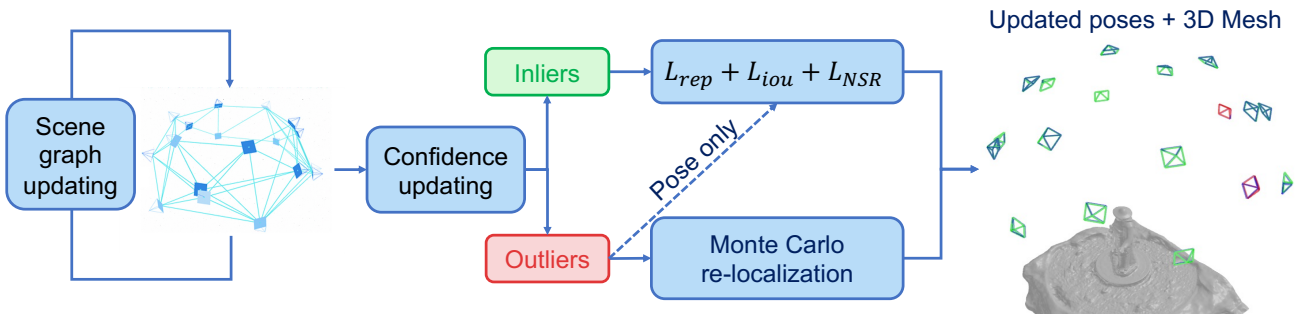

Fig. 3: An overview of the proposed pipeline. Given the initial scene graph, we apply a confidence updating strategy based on an indicator from a detached color network, which can identify inlier and outlier poses. For inliers, we utilize re-projection loss and IoU loss to enhance the geometric constraints. For outliers, we utilize Monte Carlo re-localization method to find better initializations. The scene graph is also updated based on current geometry and pose estimations. Eventually, our method can reconstruct the 3D mesh from the trained field and rectify both inlier and outlier poses with high accuracy. The coloration is same as Fig. 1.

The $\mathcal{L}_{reg}$ incorporates the Eikonal term [58] applied to the sampled points to regularize the learned SDF, which can be expressed as:

$$\mathcal{L}_{reg} = \frac{1}{k}\sum_{i=1}^{k}(\|\nabla f(p_i)\|_2 - 1)^2, \tag{5}$$

where $f(p_i)$ represents the distance estimate for each sampled 3D location along the ray.

We also utilize the Intersection-of-Union (IoU) loss $\mathcal{L}_{iou}$ and re-projection loss $\mathcal{L}_{rep}$ [59] to further improve the pose accuracy. The $\mathcal{L}_{iou}$ loss, firstly proposed by SG-NeRF [17], not only enhances geometry consistency but also accelerates convergence. Another related constraint is the epipolar loss proposed by PoRF [16]. However, we find that both epipolar and IoU losses do not handle outliers effectively. In fact, the re-projection loss can fulfill the same role as the epipolar loss [59] but is more reliable to the scene geometry. The epipolar loss does not require depth for back-projection but is invariant to the scale of the translation part. Considering the aforementioned analysis, we opt for the IoU loss and the re-projection loss in our framework.

Given a pair of matched keypoints kp_i from image I_i and kp_j from image I_j, we define the Intersection Volume as:

$$I = MoG(kp_i) \cdot MoG(kp_j), \tag{6}$$

and Union Volume as:

$$U = MoG(kp_i) + MoG(kp_j) - I, \tag{7}$$

where $MoG(\cdot)$ is a mixture of Gaussians for sampling points along a ray. The IoU loss can be computed as:

$$\mathcal{L}_{iou} = 1 - \frac{I}{U}. \tag{8}$$

With a set of points sampled from the ray corresponding to kp_i, we approximate the depth d_i of kp_i by selecting the point with maximal weights. Thus, the re-projection loss can be achieved by:

$$\mathcal{L}_{rep} = L_\delta\left(kp_i', \ kp_j\right), \tag{9}$$

where kp_i' is the re-projected point of kp_i in image I_j and L_δ represents the Huber loss.

We jointly optimize inlier-inlier pairs, while bypassing outlier-outlier pairs. For inlier-outlier pairs, we only optimize the poses of outliers and keep inliers and NeuS backbone fixed. Our overall loss is defined as:

$$\mathcal{L} = \mathcal{L}_{NSR} + \alpha\mathcal{L}_{iou} + \beta\mathcal{L}_{rep}. \tag{10}$$

C. Monte Carlo re-localization

Geometry constraints in Sec. III-B can still struggle with certain extreme cases. One intractable case comes from the mirror-symmetry ambiguity, which has been studied in SfM [60], [61], [62]. In the context of pose-NeRF joint optimization, NeRFmm [8] and LU-NeRF [15] also mentioned the same problem. LU-NeRF solves this problem by training two NeRF models, one of which uses reflected poses, requiring significant time to find the mirror poses.

Leveraging our confidence scheme, we can easily detect outliers, particularly those mirrored outliers. To maximize the use of training images, we propose to utilize Monte Carlo re-localization techniques [48], [63] to assist outliers in finding better initializations. Specifically, as we focus on inward-facing scenes, we first estimate a coarse main axis of the scene using inlier poses. The rotation around this main axis is defined as $R_{axis}(\theta)$, where θ is the angle of rotation. We then distribute the initial particles uniformly around this axis. Given a outlier pose R_o, t_o, the poses of these particles can be obtained by:

$$R_{pi} = R_{axis}(\frac{i \cdot 2\pi}{N_p}) \cdot R_o, \ \ i \in \{1, 2, \ldots, N_p\}, \tag{11}$$

$$t_{pi} = R_{axis}(\frac{i \cdot 2\pi}{N_p}) \cdot t_o, \ \ i \in \{1, 2, \ldots, N_p\}, \tag{12}$$

where (R_{p_i}, t_{p_i}) is the pose of i-th particle and N_p is the number of particles. We fix all network components and

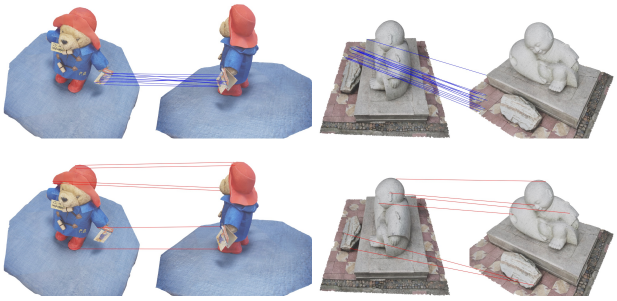

Fig. 4: The illustration of scene graph updating. We filter out the wrong matched keypoint pairs (red lines) and keep the correct pairs (blue lines). We select image pairs with relatively few matching pairs for clearer visualization.

only optimize the poses of particles. Initially, each particle is sampled equally for optimization. Subsequently, the sampling probability is adjusted based on the estimated $PSNR_n$. If the maximum $PSNR_n$ of the particles exceeds that of the current outlier at the end of the re-localization stage, we replace (R_o, t_o) with the pose of this particle.

D. Scene graph updating

The initial confidence, based on results from SfM, may be sub-optimal. Therefore, we periodically update the scene graph according to current geometry and pose estimations. Similar to Sec. III-A, we use the same angular threshold τ to remove edges from the raw graph. For remaining keypoint matching pairs, we remove those with a re-projection loss surpassing the threshold τ_{rep}, which is gradually reduced throughout the training. As illustrated in Fig. 4, our method effectively eliminates wrong matches, providing more reliable information for subsequent training iterations.

IV. Experiments

A. Experiment setup

Following SG-NeRF [17], we conduct our experiments on 8 cases from the SG-NeRF dataset and 5 cases from the DTU [22] dataset to validate our method. We assess the mesh quality with Chamfer distance (CD) and F-score metrics. The baseline methods for comparison include BARF [9], SCNeRF [10], GARF [53], L2G-NeRF [13], Joint-TensoRF [54], PoRF [16] and SG-NeRF [17]. Results with * are achieved in a two-stage manner, including official implementations and NeuS [2] with optimized poses. The initial camera poses of SG-NeRF are obtained by using Superpoint [38] and SuperGlue [41], with COLMAP [35], [45] backend optimization. As presented in SG-NeRF [17], this combination consistently outperforms the standard COLMAP but still results in a proportion of significant incorrect poses, ranging from 1/9 to 1/3. The initial poses for DTU are obtained by conventional COLMAP first. To simulate outlier poses, SG-NeRF randomly selects 1/7 to 1/4 of the images for each scene and injects random noises to their poses. For a fair comparison, all methods, including ours, use the same initial poses as input.

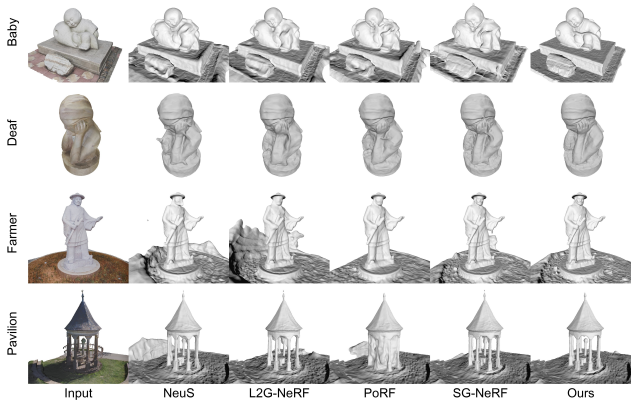

Fig. 5: Qualitative comparisons on the SG-NeRF [17] dataset. Our method can generally recover high-fidelity geometry with only one-stage training. More visual comparisons are provided in supplementary materials.

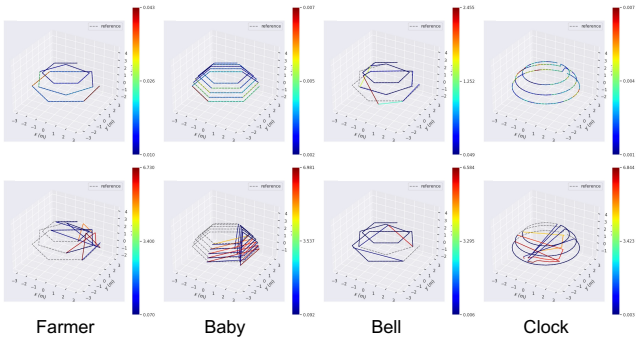

Fig. 6: Visualization of pose accuracy. The fist row presents our results and the second is SG-NeRF [17].

B. Implementation details

We implement our method based on NeuS [2]. The camera poses are parameterized with Lietorch [64], which can perform backpropagation on $\mathbf{SE}(3)$ Groups. Following SG-NeRF [17], the angular threshold τ for scene graph sparsification is set as 70 for SG-NeRF dataset [17] and 45 for DTU [22] dataset, respectively. The inlier-outlier threshold is set as $\tau_1 = 9$, which is an extremely large performance gap for $PSNR_o$ and $PSNR_n$. The weights of loss are set as $\lambda = 0.1, \alpha = 0.2$, and $\beta = 0.001$ respectively. The particle number N_p is set as 24 for efficiency. All experiments are conducted on a single NVIDIA RTX 3090 GPU. Our method runs an average of 13 hours for 150k iterations on the SG-NeRF dataset, and 22 hours for 300k iterations on the DTU dataset.

C. Comparisons

Results on SG-NeRF. The quantitative results are reported in Table I. Both NeuS and Neuralangelo degenerate severely due to significantly noisy camera poses. PoRF and SCNeRF demonstrate commendable results in certain cases, highlighting the importance of incorporating cross-view correspondences. Among the competitors, SG-NeRF achieves the best overall performance, underscoring the effectiveness of

TABLE I: Quantitative results on SG-NeRF [17]. The **red** and **blue** numbers indicate the first and second performer for each scene. † denotes that only valid values are used for the average. Methods with * are trained in a two-stage manner.

		Baby	Bear	Bell	Clock	Deaf	Farmer	Pavilion	Sculpture	Mean
Chamfer distance ↓	NeuS [2]	0.69	0.31	3.33	1.16	0.55	2.49	0.29	0.66	1.18
	Neuralangelo [3]	0.70	0.65	-	0.38	0.59	4.89	1.95	0.31	1.35†
	BARF [9]*	1.08	0.28	3.31	0.19	0.46	2.13	0.38	0.57	1.05
	SCNeRF [10]*	1.19	0.27	3.74	1.33	0.46	1.45	0.23	0.81	1.19
	GARF [53]*	2.04	2.25	3.08	2.01	0.59	1.58	0.96	0.57	1.64
	L2G-NeRF [13]*	1.15	0.29	1.26	0.24	0.40	2.18	-	4.36	1.41†
	Joint-TensoRF [54]*	3.11	-	2.49	0.36	0.88	2.51	1.35	0.70	1.63†
	PoRF [16]	**0.31**	0.49	-	-	**0.30**	3.80	2.20	-	1.42†
	SG-NeRF [17]	0.56	**0.25**	**0.98**	**0.15**	0.45	**0.87**	**0.20**	**0.22**	**0.46**
	Ours	**0.07**	**0.09**	**1.22**	**0.15**	**0.13**	**0.62**	**0.17**	**0.09**	**0.32**
F-score ↑	NeuS [2]	0.65	0.93	0.48	0.72	0.84	0.54	0.93	0.70	0.74
	Neuralangelo [3]	0.57	0.80	-	0.85	0.66	0.14	0.47	0.89	0.63†
	BARF [9]*	0.58	0.91	0.49	0.95	0.86	0.51	0.86	0.87	0.75
	SCNeRF [10]*	0.56	0.93	0.49	0.69	0.86	0.59	**0.95**	0.73	0.72
	GARF [53]*	0.18	0.21	0.50	0.27	0.78	0.57	0.41	0.83	0.47
	L2G-NeRF [13]*	0.58	0.92	0.65	0.92	0.89	0.49	-	0.21	0.67†
	Joint-TensoRF [54]*	0.20	-	0.38	0.84	0.60	0.24	0.34	0.63	0.46†
	PoRF [16]	**0.92**	0.78	-	-	**0.92**	0.39	0.35	-	0.67†
	SG-NeRF [17]	0.74	**0.93**	**0.71**	**0.96**	0.87	**0.76**	0.94	**0.92**	**0.85**
	Ours	**0.99**	**0.99**	**0.65**	**0.96**	**0.99**	**0.79**	**0.94**	**0.99**	**0.91**

TABLE II: Evaluation of pose accuracy (w.r.t. full transformation) on the SG-NeRF dataset. APE_i and RPE_i are computed w.r.t. inlier poses. And APE and RPE are calculated using all poses.

		Baby	Bear	Bell	Clock	Deaf	Farmer	Pavilion	Sculpture	Mean
APE_i ↓	SG-NeRF	0.153	0.008	0.153	0.014	0.060	0.011	0.003	0.056	0.057
	Ours	0.004	0.004	0.120	0.003	0.016	0.018	0.003	0.006	**0.022**
RPE_i ↓	SG-NeRF	0.231	0.010	0.268	0.014	0.104	0.014	0.004	0.085	0.091
	Ours	0.007	0.005	0.187	0.002	0.026	0.022	0.004	0.010	**0.033**
APE ↓	SG-NeRF	2.16	1.84	2.15	1.80	1.17	0.72	0.6	1.10	1.44
	Ours	0.004	0.004	0.37	0.003	0.016	0.018	0.003	0.006	**0.053**
RPE ↓	SG-NeRF	2.26	1.38	2.29	0.51	2.10	1.12	1.04	2.02	1.59
	Ours	0.005	0.004	0.70	0.002	0.021	0.022	0.004	0.008	**0.096**

TABLE III: Quantitative results on the DTU [22] dataset with noisy camera poses as input.

Chamfer distance ↓	24	37	40	55	63	Mean
NeuS [2]	1.07	2.80	1.52	1.30	3.20	1.98
Neuralangelo [3]	1.06	2.96	1.22	**0.42**	1.23	1.38
BARF [9]*	1.46	**1.40**	5.16	1.78	1.80	2.32
SCNeRF [10]*	1.45	2.84	2.60	0.78	1.83	1.90
GARF [53]*	1.18	2.00	2.61	2.37	8.74	3.38
L2G-NeRF [13]*	1.08	1.60	3.27	1.79	6.97	2.94
Joint-TensoRF [54]*	1.00	2.60	-	-	7.71	3.77†
PoRF [16]	1.15	2.33	0.97	0.76	1.30	1.30
SG-NeRF [17]	**0.87**	1.83	**0.88**	**0.38**	**1.13**	**1.01**
Ours	**0.80**	**1.30**	**0.61**	0.44	**1.09**	**0.85**

scene graph guidance. Our method consistently outperforms other approaches by a considerable margin, which shows the effectiveness of our framework. The visual comparison is provided in Fig. 5, where our method distinctly excels in capturing finer geometric details. However, we empirically observed that all methods, including ours, struggle with the Bell scene, likely due to the sparsity of training images.

TABLE IV: Ablation studies on the SG-NeRF (CD ↓).

	w/o Rep.	w/o S.U.	w/o M.C.	w/o detach.	SG+rules	full
baby	0.09	0.11	0.38	0.36	0.31	**0.07**
bear	0.10	0.17	0.28	0.35	0.18	**0.09**
farmer	0.65	0.69	0.88	0.86	0.92	**0.62**
sculpture	0.20	0.24	0.40	**0.09**	0.26	**0.09**
mean	0.26	0.30	0.49	0.42	0.42	**0.22**

We also utilize evo [65] to evaluate the pose accuracy of our method and SG-NeRF [17]. Due to the original SG-NeRF dataset does not provide inlier-outlier information, we utilize our indicator to filter out outliers. We align inliers to ground truth poses to get a global $\mathbf{SIM}(3)$ transformation, which is then applied to all poses. The results of absolute pose error (APE) and relative pose error (RPE) w.r.t. full transformation (including both rotation and translation parts) are reported in Table II. We also provide APE_i and RPE_i, which are computed using only inlier poses for a fair comparison. Our pose accuracy surpasses that of SG-NeRF by more than two orders of magnitude on both RPE and APE. Fig. 6 shows the visual comparison of the camera pose accuracy.

Results on DTU. The quantitative results are shown in Table III. We report a new result of SG-NeRF on Scan 37 with a better performance (originally reported as 2.39), due to the fact of our experiment. In DTU dataset, our method performs slightly better than SG-NeRF. We empirically find that our Monte Carlo re-localization has not been triggered. Thus, the experiment on the DTU dataset can be viewed as an improved version of SG-NeRF. Our method outperforms the competitors on four scans and achieves a similar performance

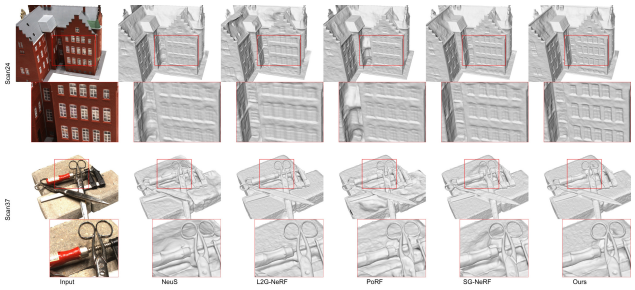

Fig. 7: Qualitative comparisons on the DTU [22] dataset. L2G-NeRF [13] is trained in a two-stage manner and others are trained in one stage with the same iterations. All methods take the same initial poses as input.

with SG-NeRF on Scan 55. The qualitative comparisons can be found in Figure 7 and more results can be found in our project page.

D. Ablation studies

We select 4 cases from the SG-NeRF [17] dataset to validate each component of our method, including the re-projection loss (w/o Rep.), Monte Carlo re-localization (w/o M.C), the scene updating strategy (w/o S.U.), and the detach operation (w/o detach). To assess different confidence strategies, we also conduct the SG+rules experiment, where the confidence reference is replaced with $PSNR_o$, and nodes with the five lowest $PSNR_o$ values are identified as outliers for the M.C. procedure. As reported in Table IV, our pipeline with the SG-NeRF confidence estimation (SG+rules) degenerates significantly, highlighting the advantage of our confidence strategy. Another problem of SG + rules lies in the re-localization process, where we can only use ambiguous color network for backpropagation. The absence of the detach operation negatively impacts the optimization process. M.C. leads to a significant improvement in the results, while the S.U. strategy also contributes to a slight but generalizable enhancement, confirming the effectiveness of these two components. Although w/o Rep. includes a cross-view constraint IoU loss, our full model still demonstrates further improvements.

V. Conclusion

This paper addresses neural surface reconstruction from image sets characterized by significant outlier poses. By leveraging the scene graph to guide training, we introduce a novel confidence updating strategy that effectively recognizes inliers and outliers. We enhance geometric constraints through the integration of Intersection-of-Union (IoU) loss and re-projection loss, while employing Monte Carlo re-localization techniques to accurately reposition outliers. These methods, combined with our scene graph updating strategy, enable our framework to achieve state-of-the-art performance on the challenging SG-NeRF dataset. One limitation of our approach is its dependency on a substantial number of inlier poses. As a promising direction for future research, incorporating prior models could make our framework more robust, especially in sparsely captured scenes.

References

[1] 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*, vol. 65, no. 1, pp. 99–106, 2021.

[2] P. Wang, L. Liu, Y. Liu, C. Theobalt, T. Komura, and W. Wang, "Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction," *NeurIPS*, 2021.

[3] Z. Li, T. Müller, A. Evans, R. H. Taylor, M. Unberath, M.-Y. Liu, and C.-H. Lin, "Neuralangelo: High-fidelity neural surface reconstruction," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[4] T. Wu, J. Wang, X. Pan, X. Xu, C. Theobalt, Z. Liu, and D. Lin, "Voxurf: Voxel-based efficient and accurate neural surface reconstruction," *arXiv preprint arXiv:2208.12697*, 2022.

[5] B. Miller, H. Chen, A. Lai, and I. Gkioulekas, "Objects as volumes: A stochastic geometry view of opaque solids," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 87–97.

[6] J. L. Schönberger and J.-M. Frahm, "Structure-from-motion revisited," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[7] J. L. Schönberger, E. Zheng, M. Pollefeys, and J.-M. Frahm, "Pixelwise view selection for unstructured multi-view stereo," in *European Conference on Computer Vision (ECCV)*, 2016.

[8] 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.

[9] C.-H. Lin, W.-C. Ma, A. Torralba, and S. Lucey, "Barf: Bundle-adjusting neural radiance fields," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 5741–5751.

[10] Y. Jeong, S. Ahn, C. Choy, A. Anandkumar, M. Cho, and J. Park, "Self-calibrating neural radiance fields," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 5846–5854.

[11] S.-F. Chng, S. Ramasinghe, J. Sherrah, and S. Lucey, "Gaussian activated neural radiance fields for high fidelity reconstruction and pose estimation," in *European Conference on Computer Vision*. Springer, 2022, pp. 264–280.

[12] 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 Conference on Computer Vision and Pattern Recognition*, 2023, pp. 4160–4169.

[13] Y. Chen, X. Chen, X. Wang, Q. Zhang, Y. Guo, Y. Shan, and F. Wang, "Local-to-global registration for bundle-adjusting neural radiance fields," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 8264–8273.

[14] P. Truong, M.-J. Rakotosaona, F. Manhardt, and F. Tombari, "Sparf: Neural radiance fields from sparse and noisy poses," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 4190–4200.

[15] Z. Cheng, C. Esteves, V. Jampani, A. Kar, S. Maji, and A. Makadia, "Lu-nerf: Scene and pose estimation by synchronizing local unposed nerfs," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 18 312–18 321.

[16] J.-W. Bian, W. Bian, V. A. Prisacariu, and P. Torr, "Porf: Pose residual field for accurate neural surface reconstruction," in *ICLR*, 2024.

[17] Y. Chen, S. Dong, X. Wang, L. Cai, Y. Zheng, and Y. Yang, "Sg-nerf: Neural surface reconstruction with scene graph optimization," in *European Conference on Computer Vision (ECCV)*, 2024.

[18] K. Zhang, G. Riegler, N. Snavely, and V. Koltun, "Nerf++: Analyzing and improving neural radiance fields," *arXiv:2010.07492*, 2020.

[19] B. Zhu, Y. Yang, X. Wang, Y. Zheng, and L. Guibas, "Vdn-nerf: Resolving shape-radiance ambiguity via view-dependence normalization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 35–45.

[20] Q. Fang, Y. Song, K. Li, and L. Bo, "Reducing shape-radiance ambiguity in radiance fields with a closed-form color estimation method," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[21] 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.

[22] R. Jensen, A. Dahl, G. Vogiatzis, E. Tola, and H. Aanæs, "Large scale multi-view stereopsis evaluation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 406–413.

[23] Y. Furukawa and J. Ponce, "Accurate, dense, and robust multiview stereopsis," *IEEE transactions on pattern analysis and machine intelligence*, vol. 32, no. 8, pp. 1362–1376, 2009.

[24] C. Zach, T. Pock, and H. Bischof, "A globally optimal algorithm for robust tv-l 1 range image integration," in *2007 IEEE 11th International Conference on Computer Vision*. IEEE, 2007, pp. 1–8.

[25] P. Merrell, A. Akbarzadeh, L. Wang, P. Mordohai, J.-M. Frahm, R. Yang, D. Nistér, and M. Pollefeys, "Real-time visibility-based fusion of depth maps," in *2007 IEEE 11th International Conference on Computer Vision*. Ieee, 2007, pp. 1–8.

[26] M. Kazhdan and H. Hoppe, "Screened poisson surface reconstruction," *ACM Transactions on Graphics (ToG)*, vol. 32, no. 3, pp. 1–13, 2013.

[27] Y. Yao, Z. Luo, S. Li, T. Fang, and L. Quan, "Mvsnet: Depth inference for unstructured multi-view stereo," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 767–783.

[28] Y. Yao, Z. Luo, S. Li, T. Shen, T. Fang, and L. Quan, "Recurrent mvsnet for high-resolution multi-view stereo depth inference," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 5525–5534.

[29] Z. Yu, S. Peng, M. Niemeyer, T. Sattler, and A. Geiger, "Monosdf: Exploring monocular geometric cues for neural implicit surface reconstruction," *Advances in neural information processing systems*, vol. 35, pp. 25 018–25 032, 2022.

[30] Q. Fu, Q. Xu, Y. S. Ong, and W. Tao, "Geo-neus: Geometry-consistent neural implicit surfaces learning for multi-view reconstruction," *Advances in Neural Information Processing Systems*, vol. 35, pp. 3403–3416, 2022.

[31] L. Yariv, Y. Kasten, D. Moran, M. Galun, M. Atzmon, B. Ronen, and Y. Lipman, "Multiview neural surface reconstruction by disentangling geometry and appearance," *Advances in Neural Information Processing Systems*, vol. 33, 2020.

[32] C. Sweeney, "Theia multiview geometry library: Tutorial & reference," http://theia-sfm.org.

[33] P. Lindenberger, P.-E. Sarlin, V. Larsson, and M. Pollefeys, "Pixel-Perfect Structure-from-Motion with Featuremetric Refinement," in *ICCV*, 2021.

[34] Z. Cui and P. Tan, "Global structure-from-motion by similarity averaging," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 864–872.

[35] C. Wu, "Visualsfm: A visual structure from motion system," *http://www. cs. washington. edu/homes/ccwu/vsfm*, 2011.

[36] N. Snavely, S. M. Seitz, and R. Szeliski, "Modeling the world from internet photo collections," *International journal of computer vision*, vol. 80, pp. 189–210, 2008.

[37] L. Pan, D. Barath, M. Pollefeys, and J. L. Schönberger, "Global Structure-from-Motion Revisited," in *European Conference on Computer Vision (ECCV)*, 2024.

[38] D. DeTone, T. Malisiewicz, and A. Rabinovich, "Superpoint: Self-supervised interest point detection and description," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2018, pp. 224–236.

[39] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International journal of computer vision*, vol. 60, pp. 91–110, 2004.

[40] M. Tyszkiewicz, P. Fua, and E. Trulls, "Disk: Learning local features with policy gradient," *Advances in Neural Information Processing Systems*, vol. 33, pp. 14 254–14 265, 2020.

[41] P.-E. Sarlin, D. DeTone, T. Malisiewicz, and A. Rabinovich, "SuperGlue: Learning feature matching with graph neural networks," in *CVPR*, 2020.

[42] P. Lindenberger, P.-E. Sarlin, and M. Pollefeys, "Lightglue: Local feature matching at light speed," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 17 627–17 638.

[43] J. Sun, Z. Shen, Y. Wang, H. Bao, and X. Zhou, "Loftr: Detector-free local feature matching with transformers," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 8922–8931.

[44] T. Sattler, B. Leibe, and L. Kobbelt, "Efficient & effective prioritized matching for large-scale image-based localization," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 9, 2016.

[45] P.-E. Sarlin, C. Cadena, R. Siegwart, and M. Dymczyk, "From coarse to fine: Robust hierarchical localization at large scale," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 12 716–12 725.

[46] E. Brachmann and C. Rother, "Visual camera re-localization from rgb and rgb-d images using dsac," *IEEE transactions on pattern analysis and machine intelligence*, vol. 44, no. 9, pp. 5847–5865, 2021.

[47] S. Dong, Q. Fan, H. Wang, J. Shi, L. Yi, T. Funkhouser, B. Chen, and L. J. Guibas, "Robust neural routing through space partitions for camera relocalization in dynamic indoor environments," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 8544–8554.

[48] D. Maggio, M. Abate, J. Shi, C. Mario, and L. Carlone, "Loc-nerf: Monte carlo localization using neural radiance fields," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2023, pp. 4018–4025.

[49] A. Moreau, N. Piasco, M. Bennehar, D. Tsishkou, B. Stanciulescu, and A. de La Fortelle, "Crossfire: Camera relocalization on self-supervised features from an implicit representation," *arXiv preprint arXiv:2303.04869*, 2023.

[50] J. Liu, Q. Nie, Y. Liu, and C. Wang, "Nerf-loc: Visual localization with conditional neural radiance field," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2023, pp. 9385–9392.

[51] 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 International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2021, pp. 1323–1330.

[52] 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.

[53] S.-F. Chng, S. Ramasinghe, J. Sherrah, and S. Lucey, "Gaussian activated neural radiance fields for high fidelity reconstruction and pose estimation," in *European Conference on Computer Vision*. Springer, 2022, pp. 264–280.

[54] B.-Y. Chen, W.-C. Chiu, and Y.-L. Liu, "Improving robustness for joint optimization of camera pose and decomposed low-rank tensorial radiance fields," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, 2024, pp. 990–1000.

[55] Y. Xia, H. Tang, R. Timofte, and L. Van Gool, "Sinerf: Sinusoidal neural radiance fields for joint pose estimation and scene reconstruction," *arXiv preprint arXiv:2210.04553*, 2022.

[56] S.-F. Chng, R. Garg, H. Saratchandran, and S. Lucey, "Invertible neural warp for nerf," in *European Conference on Computer Vision (ECCV)*, 2024.

[57] J. T. Kajiya and B. P. Von Herzen, "Ray tracing volume densities," *ACM SIGGRAPH computer graphics*, vol. 18, no. 3, pp. 165–174, 1984.

[58] A. Gropp, L. Yariv, N. Haim, M. Atzmon, and Y. Lipman, "Implicit geometric regularization for learning shapes," *arXiv preprint arXiv:2002.10099*, 2020.

[59] R. Hartley and A. Zisserman, *Multiple view geometry in computer vision*. Cambridge university press, 2003.

[60] K. E. Ozden, K. Schindler, and L. Van Gool, "Multibody structure-from-motion in practice," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 6, pp. 1134–1141, 2010.

[61] G. Schweighofer and A. Pinz, "Robust pose estimation from a planar target," *IEEE transactions on pattern analysis and machine intelligence*, vol. 28, no. 12, pp. 2024–2030, 2006.

[62] D. Oberkampf, D. F. DeMenthon, and L. S. Davis, "Iterative pose estimation using coplanar feature points," *Computer Vision and Image Understanding*, vol. 63, no. 3, pp. 495–511, 1996.

[63] F. Dellaert, D. Fox, W. Burgard, and S. Thrun, "Monte carlo localization for mobile robots," in *Proceedings 1999 IEEE international conference on robotics and automation (Cat. No. 99CH36288C)*, vol. 2. IEEE, 1999, pp. 1322–1328.

[64] Z. Teed and J. Deng, "Tangent space backpropagation for 3d transformation groups," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 10 338–10 347.

[65] M. Grupp, "evo: Python package for the evaluation of odometry and slam." https://github.com/MichaelGrupp/evo, 2017.