SG-NeRF: Neural Surface Reconstruction with Scene Graph Optimization

Yiyang Chen[1,2*] Siyan Dong[3*] Xulong Wang[1] Lulu Cai[1,4]
Youyi Zheng[1†] Yanchao Yang[3,4†]

[1] State Key Lab of CAD&CG, Zhejiang University [2] Chohotech Co. ltd.
[3] Institute of Data Science, The University of Hong Kong
[4] Department of Electrical and Electronic Engineering, The University of Hong Kong

Abstract. 3D surface reconstruction from images is essential for numerous applications. Recently, Neural Radiance Fields (NeRFs) have emerged as a promising framework for 3D modeling. However, NeRFs require accurate camera poses as input, and existing methods struggle to handle significantly noisy pose estimates (i.e., outliers), which are commonly encountered in real-world scenarios. To tackle this challenge, we present a novel approach that optimizes radiance fields with scene graphs to mitigate the influence of outlier poses. Our method incorporates an adaptive inlier-outlier confidence estimation scheme based on scene graphs, emphasizing images of high compatibility with the neighborhood and consistency in the rendering quality. We also introduce an effective intersection-over-union (IoU) loss to optimize the camera pose and surface geometry, together with a coarse-to-fine strategy to facilitate the training. Furthermore, we propose a new dataset containing typical outlier poses for a detailed evaluation. Experimental results on various datasets consistently demonstrate the effectiveness and superiority of our method over existing approaches, showcasing its robustness in handling outliers and producing high-quality 3D reconstructions. Our code and data are available at: `https://github.com/Iris-cyy/SG-NeRF`.

Keywords: surface reconstruction · pose optimization · scene graph

1 Introduction

3D mapping and reconstruction from multi-view images is crucial for a wide range of applications, such as virtual and augmented reality. Given a set of unorganized images captured around an object, most pipelines proceed in two stages for obtaining the reconstruction. Firstly, Structure-from-Motion (SfM) techniques [19, 40] are employed to estimate camera poses of the images and

* Equal contributions (chen_yy@zju.edu.cn, siyan3d@hku.hk).
† Corresponding authors.
[2] This work was done during the author's internship at Chohotech Co. ltd..

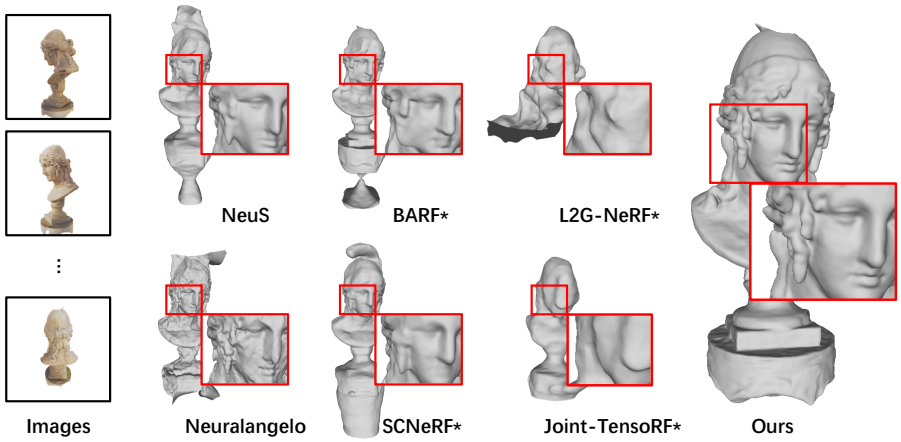

Fig. 1: 3D surface reconstruction (meshes) from images with camera poses that present significant noise. Directly training radiance fields with noisy poses can lead to incorrect structures (NeuS [49] and Neuralangelo [24]). Recent approaches that focus on optimizing camera poses (BARF [25]*, SCNeRF [22]*, L2G-NeRF [6]*, and Joint-TensoRF [7]*, where * denotes their integration of NeuS for surface modeling) also fall short in handling significant pose errors, leading to unsatisfactory reconstruction. Our method works effectively and can produce high-quality 3D reconstructions.

produce a sparse scene representation. Consecutively, the 3D scene geometry can be recovered using Multi-View Stereo (MVS) algorithms [17, 18, 41, 48, 57], which assume accurate camera poses and derive the dense reconstruction.

Recently, Neural Radiance Fields (NeRFs) [29] have been proposed for high-quality image synthesis. The core idea of NeRFs is to implicitly encode a set of posed images into the weights of a neural network. The resulting implicit scene representations [1, 45] have demonstrated photo-realistic rendering quality and the capability for novel view synthesis (NVS). Given the connection between NVS and 3D reconstruction, NeRFs have been further adapted to develop surface-aware representations [24, 49], enabling high-quality 3D reconstruction.

Despite the promising results, one tension of scene reconstruction with NeRFs is the dependence on accurate camera pose estimates. In practice, NeRF and many variants employ COLMAP [40], a widely used SfM framework, to estimate camera poses before training the scene representation. Unfortunately, the poses obtained could have significant errors affecting NeRF's reconstruction quality. Therefore, recent works [3, 8, 52, 56] perform joint optimization of the scene representation and camera poses to mitigate the effect. However, these works still struggle with outlier poses, e.g., they either focus on forward-facing cameras with relatively small baselines or require proper initial poses for water-tight meshes in inward-facing scenarios [6, 7, 22, 25, 47].

Given the context, an outlier refers to an image with a noticeably incorrect camera pose estimation, which can hardly be rectified via local optimiza-

tion. Outlier images can happen when repetitive patterns or textureless regions are present, resulting in SfM failures. On the other hand, SfM systems provide byproducts such as sparse 3D reconstructions and scene graphs besides camera poses. Several works [11, 36] have shown that leveraging the sparse reconstruction can lead to faster training using fewer images. However, to the best of our knowledge, previous research has yet to explore using scene graphs to further optimize camera poses for better geometry reconstruction.

In this paper, we propose a novel framework that jointly optimizes the neural radiance field with a scene graph to alleviate the influence of outliers. We also initialize the scene graph with an SfM system [12,38,40]. To facilitate the joint training, we introduce an adaptive inlier-outlier confidence estimation mechanism to account for the influence of outlier images/poses. Besides NeRF's photometric loss, we propose an intersection-over-union (IoU) loss across paired images in the scene graph, whose selection is coupled with the estimated confidence, to further optimize camera poses. Additionally, we apply a coarse-to-fine strategy to ensure the stability and efficiency of the training process. To perform a thorough evaluation, we collect a new dataset of 8 challenging scenes containing various levels of incorrect camera poses, as estimated by our SfM system. We conduct experiments on both the proposed and existing datasets from the literature. The experiments demonstrate the effectiveness of our method, highlighting its ability to produce high-quality 3D reconstruction results while remaining robust in the presence of outlier images or poses (e.g., Figure 1). To summarize:

- We investigate a practical problem of NeRF-based 3D reconstruction from images with significant pose errors. In contrast to previous works that assume moderate pose errors, we aim for a more challenging yet practical scenario. The images are casually captured without being carefully selected, which can lead to failures of state-of-the-art SfM systems.
- Accordingly, we propose a novel method that performs a joint optimization of the radiance field and the scene graph initialized by an SfM. The proposed can reconstruct 3D surface under significant camera pose noise with an adaptive inlier-outlier confidence estimation, an IoU loss that efficiently leverages the confidence for pose correction, and a coarse-to-fine strategy that effectively promotes the training.
- Besides showing better reconstruction performance on existing datasets, e.g., DTU [21], our method also achieves state-of-the-art results on a newly proposed dataset for multiview 3D that presents significant outlier camera poses.

2 Related Work

Neural radiance fields (NeRFs). The central concept of NeRFs [1, 29, 32, 45, 55, 61, 62] is to represent scenes as implicit fields. Such representation is obtained by separately training for each scene with posed images. Integrating the signed distance function (SDF) [10, 34, 60] into NeRFs [5, 16, 24, 33, 49–51, 54, 59] allows the representation to learn implicit surfaces, thus enabling 3D mesh reconstruction. Due to simplicity and efficiency, we choose NeuS [49] as our NeRF representation.

Since NeRFs require known camera poses as input, SfM techniques are usually applied to register the images before NeRF training. Besides camera poses, SfMs produce scene graphs and sparse 3D reconstructions. Several works [11,36] indicate that the sparse reconstructions can offer several enhancements compared to plain NeRFs. However, previous research has not yet investigated the usage of scene graphs for camera pose optimization.

There are also works that perform joint NeRF and camera pose optimization with modular modifications [2, 7, 8, 20, 25, 35, 52] and cross-view correspondences [22, 47]. Most assume that all images have poses initialized properly and aim at local optimization for pose correction, e.g., L2G-NeRF [6] presents a local-to-global registration pipeline to alleviate the suboptimality. However, they still suffer from the presence of outlier images. In this paper, we leverage scene graphs and introduce an adaptive inlier-outlier confidence scoring mechanism to mitigate the influence of outlier images in scene reconstruction. In contrast to [22, 47], we propose an intersection-over-union (IoU) loss, further enhancing the geometry quality of the implicit surface. Existing works also perform joint camera tracking and scene reconstruction in a sequential fashion [3,56,63], which go beyond the scope of our work and we omit due to limited space.

Structure-from-Motion (SfM) and (re)localization. SfM techniques [19] are widely used as data pre-processing for NeRF reconstruction. Given a set of unorganized images, SfM systems [26,27,40,43,44,53] organize the images by estimating camera poses and triangulating 3D scene points. Global SfM approaches [9, 44] are time efficient, yet can be sensitive to outliers. Incremental SfM systems [40, 42] are more commonly used in the literature on NeRFs. Following previous research [22,25], we utilize COLMAP [40] as a basic component of our SfM module. It estimates camera poses along with a graph structure, i.e., scene graph, where each node represents an image and each edge indicates the presence of estimated matching points in the connected images. Despite the usage of geometric verification [15,46] and outlier filtering [40] in SfMs, it is highly possible for the scene graphs to still contain significantly incorrect pose estimations (i.e., outliers). The presence of outlier poses can heavily affect NeRF training, resulting in incorrect geometry and blurred appearance.

The task of localization [37, 39] is closely related to incremental SfM. Given a set of posed images as a database, it aims to estimate the camera poses of new images captured from different viewpoints. Recent studies [30, 31] have introduced NeRFs for the visual localization task. However, their pose accuracy still falls behind. Although there are implicit representations [4, 13] that can achieve better pose estimation, they can hardly be extended to optimize the poses within the database. In the proposed method, we leverage the recent advances of hloc [37, 38], an SfM-based visual localization toolbox, to supplement our SfM module with SuperPoint [12] and SuperGlue [38].

3 Method

In the following, we first detail the problem setting and provide an overview of the proposed pipeline. Then, we elaborate on the key technical designs in the subsequent sections.

Problem statement. We aim for 3D surface reconstruction of object-level scenes from unorganized image sets. We assume that the camera's intrinsic parameters are known and that there are no distortions in images. We focus on inward-facing scenes, a common scenario for object scanning in practice. Specifically, for each scene, the input is a set of RGB images $\mathbf{I} = \{I_1, I_2, ..., I_n\}$, and the output is a 3D surface reconstruction S of the scene. The byproduct of our method is the optimized camera pose $P_i = (R_i, t_i)$ for each training image, where $R_i \in SO(3)$ and $t_i \in \mathbb{R}^3$ denoting the rotation and translation respectively. Each pose is assigned an inlier-outlier confidence score. One can also synthesize novel view images from the trained radiance field.

Method overview. Figure 2 illustrates the workflow of the proposed pipeline. Given the training images, we first apply a widely used Structure-from-Motion (SfM) algorithm, i.e., COLMAP [40], to construct an initial scene graph of the images, where the keypoint descriptor and matching are provided by SuperPoint [12] and SuperGlue [38], respectively. The scene graph created through SfM usually contains outlier poses. Therefore, we refine the graph and allocate an inlier-outlier confidence score to each node. Following this, we train a Neural Radiance Field (NeRF) using the refined scene graph. The training process is essentially a scene-specific joint optimization. It involves alternating between adjusting the radiance field and updating the scene graph. In particular, the radiance field learns to recover the 3D densities and RGB colors in the scene. Simultaneously, the scene graph optimizes camera poses and their confidence scores, gradually eliminating the influence of estimated outliers. After training, we extract the 3D scene mesh from the density of the optimized radiance field.

Next, we describe how to initialize the scene graph (Sec. 3.1). Then, we present our joint optimization method for training the radiance field and updating the scene graph (Sec. 3.2). Lastly, we introduce a coarse-to-fine training strategy to ensure an efficient and stable training process (Sec. 3.3).

3.1 Scene Graph

A scene graph $G = (V, E)$ in SfM consists of a set of nodes V and edges E. Each node $v_i \in V$ corresponds to an input image $I_i \in \mathbf{I}$, and an edge between two nodes indicates that the connected images share a co-visible region of the scene. More explicitly, the nodes record the camera poses $\{P_1, P_2, ..., P_n\}$ of the corresponding images, and the edges record the keypoint matches $\mathbf{M} = \{M_{i,j} | I_i, I_j \in \mathbf{I}\}$ of the paired images I_i, I_j, i.e., $M_{ij} = \{(kp_i^{(n)}, kp_j^{(n)}) | kp_i^{(n)}, kp_j^{(n)} \in \mathbb{R}^2\}$ is the set of matched keypoint locations $kp_i^{(n)}, kp_j^{(n)}$ (n-th match between I_i and I_j).

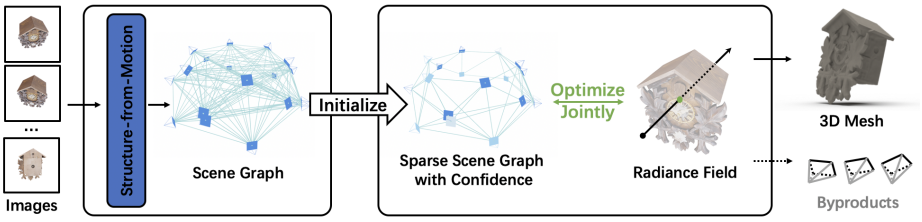

Fig. 2: An overview of the proposed joint learning pipeline. Given a set of images, we first apply a Structure-from-Motion (SfM) algorithm to construct an initial scene graph (left), within which, each node represents a posed image. An edge between two nodes suggests that the involved images are to share overlapped regions. Next, the initial scene graph is sanctified. Each node is then assigned a confidence score based on the number of matching points among neighboring nodes. Then, we train a Neural Radiance Field (NeRF) using the confidence-aware scene graph and images. The training process alternates between fitting the radiance field and updating the scene graph. Eventually, we can extract the 3D scene mesh from the trained field.

Initialization. The scene graph is initially constructed with the employed SfM module [40], which contains two major steps: a) correspondence search, and b) incremental registration and reconstruction. In the first step, we apply pre-trained SuperPoint [12] to extract keypoints from images, and exhaustively match every pair of images with keypoints using pre-trained SuperGlue [38] model. As a result, we obtain a set of successfully matched image pairs. Since there could be incorrect matches, the SfM pipeline validates each pair by estimating a relative pose via solving the essential matrix [19] with RANSAC [14] from a set of putative matches. If a valid essential matrix exists that maps a sufficient number of keypoints between the two images, the pair is considered to have passed the verification. The initial scene graph is constructed by setting the images as nodes and the verified pairs as edges.

The second step begins with selecting an appropriate image pair to initialize a metric reconstruction as well as the global coordinate system of the scene, which triangulates 3D coordinates from keypoints. Subsequently, the SfM system alternates between image registration and scene reconstruction. An image I_i is registered by estimating an absolute camera pose P_i in the scene coordinate system, which is then recorded in the corresponding node v_i. This is achieved by solving the Perspective-n-Point (PnP) problem [23] using RANSAC [14]. Once an image is registered, it expands the 3D reconstruction by triangulating new keypoints. Following registration and triangulation, the SfM system performs bundle adjustment and employs a filter to detect and remove outlier keypoints. The matches after the filtering process are recorded to the edges E.

Pruning and confidence estimation. Despite the outlier filtering in the SfM process, the scene graph obtained may still contain errors that affect the reconstruction quality. As shown in Figure 3, there can be false positive edges resulting from incorrect matches. Empirically, a pair of images is less reliable when there is a

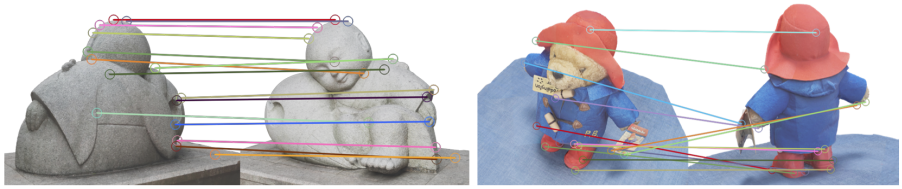

Fig. 3: Visualization of matches that are falsely established as correspondences from non-overlapping regions. The results are obtained using COLMAP [40] with SuperPoint [12] and SuperGlue [38]. Note that most of the estimations are incorrect and can heavily affects the reconstruction quality.

larger relative rotation or fewer matches. Therefore, we set an angular threshold τ for the estimated relative rotations and remove any edges exceeding τ. This sparsification effectively prunes low-quality image pairs (i.e., edges) and reduces redundancy. Furthermore, we assign a confidence estimate to each node based on keypoint matches, which helps us determine whether it is likely an inlier or outlier. The confidence score for a node v_i is computed as:

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

where $\mathbf{M}_i \subset \mathbf{M}$ denotes the set of edges connected to v_i and $|\cdot|$ is the number of elements in a set. Then, the confidence score is normalized among all the nodes via $CS(v_i) = CS(v_i) / \sum_{v_i \in V} CS(v_i)$. As such, a higher score indicates that the image has more keypoint matches and a higher probability of being an inlier. The confidence scores form a probability distribution and it is used to guide the sampling of training data for the consecutive radiance field training process.

The sparsified scene graph, along with its confidence estimates, are jointly optimized during the upcoming radiance field training process. Explicitly, the confidence scores will be updated, the camera poses will be optimized, and the graph structure will remain fixed. Details are elaborated on in the next section.

3.2 Joint Optimization

Once we obtain a scene graph with confidence scores, we train a neural radiance field to fit the 3D scene representation. To perform 3D reconstruction, we utilize a neural implicit surface representation, NeuS [49], as our backbone. Below, we first briefly review the radiance field representation and then introduce our joint optimization scheme.

Radiance field. The neural radiance fields (NeRFs) [29] represent a scene by modeling the occupancy and color of a point in the 3D space. This is achieved by training a neural network to fit each individual scene. The network takes a 3D location and viewing direction as input and generates the corresponding density and RGB color (i.e., radiance) as output. The volume rendering technique allows

the synthesis of an image by integrating radiance along the viewing ray. During the training process, the goal is to minimize the difference between synthesized pixels and those in real images as an L1 photometric loss:

$$\mathcal{L}_{photo} = |\hat{I}_i - I_i|. \tag{2}$$

In order to extract high-quality surfaces from the field, NeuS [49] enhances the density estimation with a signed distance function f (SDF) and introduces an additional regularization loss on each viewing ray:

$$\mathcal{L}_{reg} = \frac{1}{k}\sum_{i=1}^{k}(||\nabla f(p_i)||_2 - 1)^2, \tag{3}$$

where $f(p_i)$ represents the distance estimate for each sampled 3D location along the ray. We utilize NeuS as the backbone for our radiance field and leverage the two aforementioned loss terms during our training process.

Joint optimization of the field and scene graph. Our training process simultaneously fits the radiance field parameters and optimizes the scene graph. It consists of several training epochs. In each epoch, we alternate between two steps: optimizing the parameters of the radiance field and camera poses (*field-pose step*), and updating the confidence scores (*confidence step*).

During the *field-pose step,* we start by creating a temporary training set that includes images and their camera pose estimates. This set is established by *sampling with replacement based on the confidence scores*, which effectively loops in the adaptive confidence estimation into the NeRF training process. A higher confidence score indicates a larger probability of the corresponding image being selected. After selection, for the training images, besides photometric (Eq. 2) and regularization (Eq. 3) loss terms, we propose an intersection-over-union (IoU) loss. As illustrated in Figure 4, it is computed on top of keypoint matches. For each keypoint in a match, we project a ray from the camera center. We then fit a mixture of Gaussians (MoG) using the points sampled along this ray. The IoU loss aims to maximize the intersection-over-union between the two MoGs that correspond to the matched keypoints. It is calculated as:

$$\mathcal{L}_{iou} = 1 - \frac{MoG(kp_i) \cdot MoG(kp_j)}{MoG(kp_i) + MoG(kp_j)}. \tag{4}$$

Given a training image as a source, we traverse through all the paired reference images to compute the IoU loss. Finally, our total training loss is defined as:

$$\mathcal{L} = \mathcal{L}_{photo} + \alpha\mathcal{L}_{reg} + \beta\mathcal{L}_{iou}\,. \tag{5}$$

The photometric loss helps optimize the radiance field's color, geometry, and the camera poses, while the regularization terms further bias the training towards more robust and geometrically meaningful reconstructions.

In the *confidence step,* the confidence scores are updated adaptively according to the actual image reconstruction quality. Since the initial scores are derived

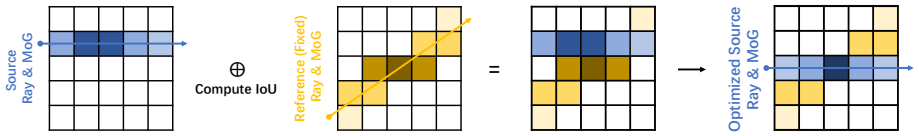

Fig. 4: Illustration of the two-view intersection-over-union (IoU) loss in 2D that can be easily extended into 3D. Given a pair of matched keypoints from source and reference images, in order to maximize the IoU between the two rays, both the camera pose of the source image and the estimated density in the radiance field have to be optimized.

from keypoint matches, they might lack a comprehensive understanding of the information contained in an image. Especially, we observe that the synthesized image from an outlier viewpoint tends to exhibit significant artifacts, due to bad alignment with the majority of the properly posed images. Therefore, we calculate the peak signal-to-noise ratio (PSNR) for each image in the scene graph against its rendering from the radiance field. The PSNR values are normalized among all the nodes, serving as a fidelity of the matching, and fused into the confidence scores as

$$CS(v_i) = CS(v_i) + \lambda PSNR(v_i). \tag{6}$$

Finally, we re-normalize the scores to maintain a probability distribution, thereby guiding the sampling process in the next epoch. As training progresses, the confidence scores are gradually scaled before the aforementioned re-normalization step to promote convergence.

3.3 Coarse-to-Fine Strategy

During training, we further apply a coarse-to-fine strategy to improve robustness and efficiency. Firstly, we remove high-frequency details from the input images by a smoothing, which allows for a quick fitting of a coarse radiance field representation of the scene. As discussed in the recent study [25], removing high-frequency bands also helps to avoid getting stuck in local minimas during training. Moreover, it facilitates the proposed PSNR-based confidence update, because low PSNR can be observed even in accurately posed images containing a large amount of high-frequency details. Thus, removing high-frequency details of images during the initial training epochs tends to put more optimization capacity on poses. As the training progresses, we gradually recover the high-frequency details. More specifically, the coarse-to-fine strategy is implemented by applying a Gaussian filter to the original input images at the beginning of each epoch. We initially set a large standard deviation σ for the Gaussian kernel (coarsest scale), which gradually decreases as the training proceeds. When $\sigma < 1$ (pixel), we stop the Gaussian filtering and use the original images as input (finest scale). Next, we validate the proposed pipeline and designs.

4 Experiments

We evaluate the effectiveness of our method through extensive experiments on various datasets, which includes a new inward-facing dataset containing significant outlier camera poses produced by the SfM system (Sec. 4.1). Next, we describe the implementation details of the proposed (Sec. 4.2). We then report the comparisons with state-of-the-art methods on both the proposed dataset and a widely used benchmark, DTU dataset [21] (Sec. 4.3). Furthermore, we perform a series of ablation studies and analyses to verify the effectiveness of each proposed component (Sec. 4.4).

4.1 Proposed Dataset

To demonstrate the generality of the stated problem setting and test the efficiency of the proposed method, we collect 3D meshes from BlendedMVS [58] and construct a new inward-facing dataset by uniformly sampling camera viewpoints in the hemisphere around each mesh. We select 8 representative scenes, and each of the selected scenes contain 18-45 training images, except the scene *Clock*, which comprises 108 training images.

We calculate the initial camera poses for these training views with our SfM module. The initial reconstruction result of each scene contains significant incorrect poses, with a proportion ranging from 1/9 to 1/3. Most of these poses tend to come with a large angular deviation and cannot be rectified through local optimization. Due to limited space, please refer to the supplementary material for a detailed description and statistics of the dataset.

4.2 Implementation Details

We employ COLMAP [40] as our SfM framework. Following hloc [37], we replace the keypoints and the matching module with SuperPoint [12] and SuperGlue [38], respectively. We employ the off-the-shelf weights from their official repository. After SfM, to prune the scene graph, we set the angular threshold to $\tau = 70$ degrees for our dataset. For the DTU dataset, we set $\tau = 45$ degrees because the viewpoints are more densely sampled. We set λ (Eq. 6) as 1.0 to balance the initial and updated confidence scores.

We implement our radiance field based on NeuS [49]. We follow the hierarchical sampling strategy in NeuS and set the batch size to 512, among which, we select 16 matched keypoints and use them to calculate the IoU loss. The MoG of each ray is calculated using 8 points with the highest densities. Specifically, we assign a 3D Gaussian to each point with the point coordinates as the mean and a fixed value of 0.1 as the covariance. Then, we fuse the Gaussians along the ray to form a MoG by a normalized weighted sum, where the weights come from the field densities estimated by NeuS. We discretize the MoG with a resolution of $64 \times 64 \times 64$. We set the loss weights (Eq. 5) to $\alpha = 0.1$ and $\beta = 0.2$.

We initialize the coarse-to-fine parameter σ as $max(H, W) \times 0.02$, where H and W represent the height and weight of the input images. After training, we

use the marching cube algorithm [28] to extract a 3D mesh from the radiance field. All of the experiments are conducted on NVIDIA RTX 3090 GPUs. Our method runs in average 11 hours for 150k iterations on the proposed dataset, and 18 hours for 300k iterations on the DTU dataset.

4.3 Comparisons

We compare our method with existing approaches on the proposed dataset and the DTU dataset. We follow the evaluation protocol in the literature [24,49] and report Chamfer distance and F-score for evaluating the mesh quality.

Table 1: Quantitative results on our dataset. The **red** and **blue** numbers indicate the first and second performer for each scene. † denotes that only valid values are used for the average. Overall, our method achieves the best reconstruction results.

		Baby	Bear	Bell	Clock	Deaf	Farmer	Pavilion	Sculpture	Mean
Chamfer distance ↓	NeuS [49]	0.69	0.31	3.33	1.16	0.55	2.49	0.29	0.66	1.18
	Neuralangelo [24]	0.70	0.65	-	0.38	0.59	4.89	1.95	**0.31**	1.35†
	BARF [25]*	1.08	0.28	3.31	**0.19**	0.46	2.13	0.38	0.57	**1.05**
	SCNeRF [22]*	1.19	**0.27**	3.74	1.33	0.46	**1.45**	**0.23**	0.81	1.19
	GARF [8]*	2.04	2.25	3.08	2.01	0.59	1.58	0.96	0.57	1.64
	L2G-NeRF [6]*	1.15	0.29	**1.26**	0.24	**0.40**	2.18	-	4.36	1.41†
	Joint-TensoRF [7]*	3.11	-	2.49	0.36	0.88	2.51	1.35	0.70	1.63†
	PoRF [2]	**0.31**	0.49	-	-	**0.30**	3.80	2.20	-	1.42†
	SG-NeRF (Ours)	**0.56**	**0.25**	**0.98**	**0.15**	0.45	**0.87**	**0.20**	**0.22**	**0.46**
F-score ↑	NeuS [49]	0.65	**0.93**	0.48	0.72	0.84	0.54	0.93	0.70	0.74
	Neuralangelo [24]	0.57	0.80	-	0.85	0.66	0.14	0.47	**0.89**	0.63†
	BARF [25]*	0.58	0.91	0.49	**0.95**	0.86	0.51	0.86	0.87	**0.75**
	SCNeRF [22]*	0.56	**0.93**	0.49	0.69	0.86	**0.59**	**0.95**	0.73	0.72
	GARF [8]*	0.18	0.21	0.50	0.27	0.78	0.57	0.41	0.83	0.47
	L2G-NeRF [6]*	0.58	**0.92**	**0.65**	0.92	**0.89**	0.49	-	0.21	0.67†
	Joint-TensoRF [7]*	0.20	-	0.38	0.84	0.60	0.24	0.34	0.63	0.46†
	PoRF [2]	**0.92**	0.78	-	-	**0.92**	0.39	0.35	-	0.67†
	SG-NeRF (Ours)	**0.74**	**0.93**	**0.71**	**0.96**	0.87	**0.76**	**0.94**	**0.92**	**0.85**

Competitors. We compare our method with our backbone model NeuS [49] and the state-of-the-art Neuralangelo [24]. We integrate our backbone with state-of-the-art camera pose optimization methods to create a series of primary competitors. Specifically, we select BARF [25], SCNeRF [22], GARF [8], L2G-NeRF [6], and Joint-TensoRF [7] as the baselines. For each of them, we adopt the official implementation to optimize camera poses, and then apply the optimized poses to train NeuS. The resulting competitors are respectively labeled as BARF*, SC-NeRF*, GARF*, L2G-NeRF*, and Joint-TensoRF*. Note that for all methods,

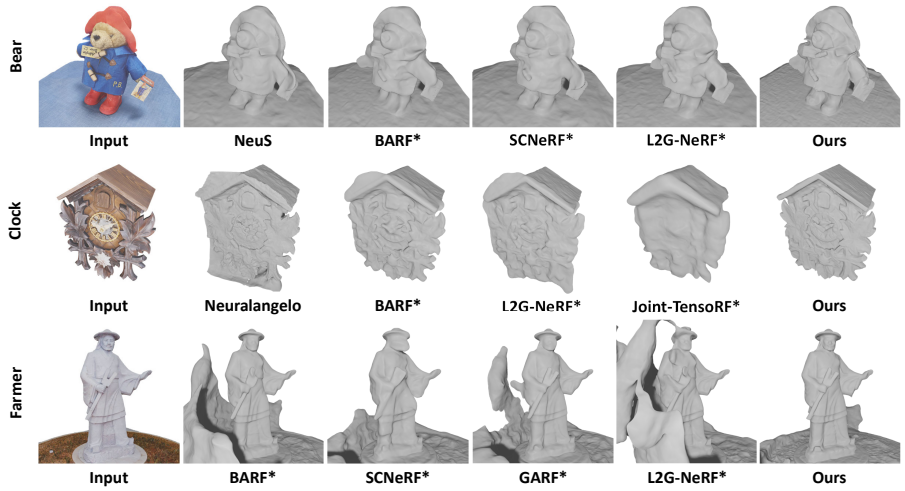

Fig. 5: Qualitative comparisons on the proposed dataset. As shown, our method is more robust to outlier poses, producing less distortion and better geometric detail. For the sake of space, we display the five top-performing results for each scene.

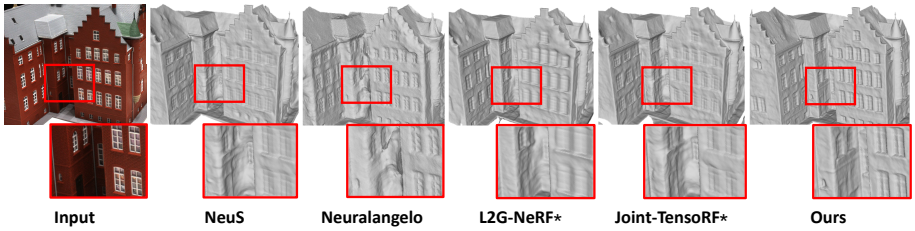

Fig. 6: Qualitative comparison on the DTU dataset (top five). Neuralangelo shows detailed windows but struggles with the gap area between buildings. In contrast, our method produces more accurate details.

we utilize the initial camera poses derived from our SfM module for a fair comparison. We also compare with recent work PoRF [2] on our dataset. It jointly optimizes camera poses and neural surface reconstruction.

Results on our dataset. The quantitative results are reported in Table 1. Regarding NeuS and Neuralangelo, due to significantly noisy camera poses, their 3D reconstructions present significant flaws, as shown in Figure 5. In contrast, our method shows robustness to pose errors and outperforms NeuS by 61% in Chamfer distance and by 15% in F-score.

Interestingly, among the competitors, the early one, BARF*, delivers the best overall performance. But it still lags behind our method. The subpar performance of the competitors is due to their pose optimization processes, namely, local optimizations, which cannot rectify the poses with significant errors. These

incorrect poses can introduce large deviations during the optimization process, affecting the correct poses, and leading to an overall decline in the pose accuracy. On the contrary, thanks to the confidence-based sampling scheme, our method is more robust to outlier poses. Please also note that there are several failure cases from the competitors indicating completely incorrect reconstruction. For example, Neuralangelo in *Bell* and L2G-NeRF in *Pavilion*, while our method consistently achieves reasonable results across all scenes.

Results on the DTU dataset. The quantitative results are shown in Table 2. To simulate outlier poses, we randomly select $1/7$ to $1/4$ of the images for each scene, and inject random noise ($\epsilon_t \in [0, 90]$ degrees to the direction of the translation vector, and $\epsilon_r \in [0, 20]$ degrees to the rotation matrix) to the corresponding poses. Since the ground-truth 3D models in DTU are obtained from structured light scanning, they have fine-grained details.

Table 2: Quantitative results on the DTU dataset with noisy camera poses as input.

Chamfer distance ↓	24	37	40	55	63	Mean
NeuS [49]	1.07	2.80	1.52	1.30	3.20	1.98
Neuralangelo [24]	1.06	2.96	**1.22**	**0.42**	**1.23**	**1.38**
BARF [25]*	1.46	**1.40**	5.16	1.78	1.80	2.32
SCNeRF [22]*	1.45	2.84	2.60	0.78	1.83	1.90
GARF [8]*	1.18	2.00	2.61	2.37	8.74	3.38
L2G-NeRF [6]*	1.08	**1.60**	3.27	1.79	6.97	2.94
Joint-TensoRF [7]*	**1.00**	2.60	-	-	7.71	3.77†
SG-NeRF (Ours)	**0.87**	2.39	**0.88**	**0.38**	**1.13**	**1.13**

As observed, Neuralangelo produces the second-best results. While BARF* achieves the best results in scene 37, it is more likely to impose negative impact on camera poses, thereby has worse performance in most scenes. Compared to the competitors, our method achieves the best overall performance among all the methods. A qualitative comparison can be found in Figure 6.

4.4 Analyses

Ablation studies. We select three representative scenes from the proposed dataset and conduct ablation studies to evaluate the effectiveness of each component. Specifically, the experiments aim to validate the contribution of the scene graph (joint optimization), the IoU loss, and the coarse-to-fine strategy.

The results are reported in Table 3. To evaluate the effectiveness of the joint optimization, we directly train our method using the original scene graph obtained from SfM without further refinement. A noticeable performance drop is observed when compared to our full method (as indicated by w/o τ in the table). By disabling the use of the confidence score (w/o CS), an even larger performance drop is observed. Similar performance drop is also observed when we remove the IoU term from the loss function (w/o IoU). These highlight the importance of employing the proposed components in the joint optimization framework. Finally, the usefulness of the coarse-to-fine strategy is verified by comparing with the second last column (w/o C2F).

Why our method is robust to outliers. Here we provide an examination of how our method can help mitigate the influence of outliers. In Figure 7, we show

Table 3: Quantitative results of our ablation studies. We individually remove the use of sparsification by thresholding (w/o τ), confidence estimation (w/o CS), Intersection-over-Union loss (w/o IoU), and coarse-to-fine optimization strategy (w/o C2F) from our full method. The **bold** numbers indicate the best results.

	Chamfer distance ↓					F-score ↑				
	w/o τ	w/o CS	w/o IoU	w/o C2F	Full	w/o τ	w/o CS	w/o IoU	w/o C2F	Full
Bell	1.27	1.32	1.68	1.15	**0.98**	0.56	0.64	0.40	0.67	**0.71**
Pavilion	0.27	0.30	0.26	0.21	**0.20**	0.90	0.88	0.91	0.93	**0.94**
Sculpture	0.26	0.38	0.23	0.37	**0.22**	0.91	0.89	0.91	0.85	**0.92**
Mean	0.60	0.67	0.72	0.58	**0.47**	0.79	0.80	0.74	0.82	**0.86**

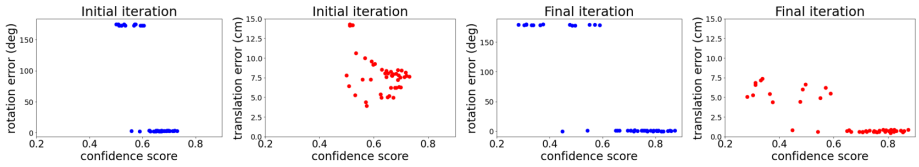

Fig. 7: The correlation between confidence scores and actual camera pose errors. As the training progresses, images with lower pose errors tend to have higher confidence scores, leading to a major improvement of the inlier poses.

the confidence score distribution in the scene *Bear* and the pose error. The left two subplots display the initial distributions, while the right two depict the distributions after the optimization. Notably, images with larger pose errors exhibit lower scores, and this gap increases as the refinement progresses, during which, the inlier poses are sampled and refined, resulting in the observed improvement. Nevertheless, the observation in Figure 7 is just an aspect of the working mechanism of our full pipeline, which leverages a synergy of the scene graph optimization with an IoU loss and a coarse-to-fine strategy.

5 Conclusion

This paper explores the task of neural surface reconstruction from image sets containing significant outlier poses. We introduce a novel method that jointly optimizes the neural radiance field with a scene graph. The key idea is to adaptively estimate the proposed inlier-outlier confidence scores and reduce the influence of outlier poses during reconstruction. In addition, we propose an Intersection-over-Union (IoU) loss and a coarse-to-fine strategy to facilitate the optimization process. Even though our method can greatly refine the inlier poses, the improvement on outlier poses is moderate (whose effect is still largely alleviated with the proposed confidence scheme), which we deem a limitation. Future explorations on rectifying outlier poses with visual (re)localization would be a promising direction, according to our study in this work.

Acknowledgments. We thank the anonymous reviewers for their valuable feedback. This work is supported by the Early Career Scheme of the Research Grants Council (grant # 27207224), the HKU-100 Award, and in part by NSF China (No. 62172363). Siyan Dong would also like to thank the support from HKU Musketeers Foundation Institute of Data Science for the Postdoctoral Research Fellowship.

References

1. Barron, J.T., Mildenhall, B., Verbin, D., Srinivasan, P.P., Hedman, P.: Mip-nerf 360: Unbounded anti-aliased neural radiance fields. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5470–5479 (2022)
2. Bian, J.W., Bian, W., Prisacariu, V.A., Torr, P.: Porf: Pose residual field for accurate neural surface reconstruction. In: ICLR (2024)
3. Bian, W., Wang, Z., Li, K., Bian, J.W., Prisacariu, V.A.: Nope-nerf: Optimising neural radiance field with no pose prior. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4160–4169 (2023)
4. Brachmann, E., Rother, C.: Visual camera re-localization from rgb and rgb-d images using dsac. IEEE transactions on pattern analysis and machine intelligence **44**(9), 5847–5865 (2021)
5. Cai, B., Huang, J., Jia, R., Lv, C., Fu, H.: Neuda: Neural deformable anchor for high-fidelity implicit surface reconstruction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8476–8485 (2023)
6. Chen, Y., Chen, X., Wang, X., Zhang, Q., Guo, Y., Shan, Y., Wang, F.: Local-to-global registration for bundle-adjusting neural radiance fields. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8264–8273 (2023)
7. Cheng, B.Y., Chiu, W.C., Liu, Y.L.: Improving robustness for joint optimization of camera poses and decomposed low-rank tensorial radiance fields. arXiv preprint arXiv:2402.13252 (2024)
8. Chng, S.F., Ramasinghe, S., Sherrah, J., Lucey, S.: Gaussian activated neural radiance fields for high fidelity reconstruction and pose estimation. In: European Conference on Computer Vision. pp. 264–280. Springer (2022)
9. Cui, Z., Tan, P.: Global structure-from-motion by similarity averaging. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 864–872 (2015)
10. Curless, B., Levoy, M.: A volumetric method for building complex models from range images. In: Proceedings of the 23rd annual conference on Computer graphics and interactive techniques. pp. 303–312 (1996)
11. Deng, K., Liu, A., Zhu, J.Y., Ramanan, D.: Depth-supervised NeRF: Fewer views and faster training for free. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2022)
12. DeTone, D., Malisiewicz, T., Rabinovich, A.: Superpoint: Self-supervised interest point detection and description. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops. pp. 224–236 (2018)
13. Dong, S., Fan, Q., Wang, H., Shi, J., Yi, L., Funkhouser, T., Chen, B., Guibas, L.J.: 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. pp. 8544–8554 (2021)

14. Fischler, M.A., Bolles, R.C.: Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. Communications of the ACM **24**(6), 381–395 (1981)
15. Frahm, J.M., Pollefeys, M.: Ransac for (quasi-) degenerate data (qdegsac). In: 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06). vol. 1, pp. 453–460. IEEE (2006)
16. Fu, Q., Xu, Q., Ong, Y.S., Tao, W.: Geo-neus: Geometry-consistent neural implicit surfaces learning for multi-view reconstruction. Advances in Neural Information Processing Systems **35**, 3403–3416 (2022)
17. Furukawa, Y., Hernández, C., et al.: Multi-view stereo: A tutorial. Foundations and Trends® in Computer Graphics and Vision **9**(1-2), 1–148 (2015)
18. Gu, X., Fan, Z., Zhu, S., Dai, Z., Tan, F., Tan, P.: Cascade cost volume for high-resolution multi-view stereo and stereo matching. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 2495–2504 (2020)
19. Hartley, R., Zisserman, A.: Multiple view geometry in computer vision. Cambridge university press (2003)
20. Heo, H., Kim, T., Lee, J., Lee, J., Kim, S., Kim, H.J., Kim, J.H.: Robust camera pose refinement for multi-resolution hash encoding. arXiv preprint arXiv:2302.01571 (2023)
21. Jensen, R., Dahl, A., Vogiatzis, G., Tola, E., Aanæs, H.: Large scale multi-view stereopsis evaluation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 406–413 (2014)
22. Jeong, Y., Ahn, S., Choy, C., Anandkumar, A., Cho, M., Park, J.: Self-calibrating neural radiance fields. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5846–5854 (2021)
23. Lepetit, V., Moreno-Noguer, F., Fua, P.: Ep n p: An accurate o (n) solution to the p n p problem. International journal of computer vision **81**, 155–166 (2009)
24. Li, Z., Müller, T., Evans, A., Taylor, R.H., Unberath, M., Liu, M.Y., Lin, C.H.: Neuralangelo: High-fidelity neural surface reconstruction. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2023)
25. Lin, C.H., Ma, W.C., Torralba, A., Lucey, S.: Barf: Bundle-adjusting neural radiance fields. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5741–5751 (2021)
26. Lindenberger, P., Sarlin, P.E., Larsson, V., Pollefeys, M.: Pixel-Perfect Structure-from-Motion with Featuremetric Refinement. In: ICCV (2021)
27. Liu, S., Yu, Y., Pautrat, R., Pollefeys, M., Larsson, V.: 3d line mapping revisited. In: Computer Vision and Pattern Recognition (CVPR) (2023)
28. Lorensen, W.E., Cline, H.E.: Marching cubes: A high resolution 3d surface construction algorithm. In: Seminal graphics: pioneering efforts that shaped the field, pp. 347–353 (1998)
29. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: Nerf: Representing scenes as neural radiance fields for view synthesis. In: ECCV (2020)
30. Moreau, A., Piasco, N., Bennehar, M., Tsishkou, D., Stanciulescu, B., de La Fortelle, A.: Crossfire: Camera relocalization on self-supervised features from an implicit representation. arXiv preprint arXiv:2303.04869 (2023)
31. Moreau, A., Piasco, N., Tsishkou, D., Stanciulescu, B., de La Fortelle, A.: Lens: Localization enhanced by nerf synthesis. In: Conference on Robot Learning. pp. 1347–1356. PMLR (2022)

32. Müller, T., Evans, A., Schied, C., Keller, A.: Instant neural graphics primitives with a multiresolution hash encoding. ACM Trans. Graph. **41**(4), 102:1–102:15 (Jul 2022). https://doi.org/10.1145/3528223.3530127, https://doi.org/10.1145/3528223.3530127
33. Oechsle, M., Peng, S., Geiger, A.: Unisurf: Unifying neural implicit surfaces and radiance fields for multi-view reconstruction. In: International Conference on Computer Vision (ICCV) (2021)
34. Park, J.J., Florence, P., Straub, J., Newcombe, R., Lovegrove, S.: Deepsdf: Learning continuous signed distance functions for shape representation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 165–174 (2019)
35. Park, K., Henzler, P., Mildenhall, B., Barron, J.T., Martin-Brualla, R.: Camp: Camera preconditioning for neural radiance fields. ACM Transactions on Graphics (TOG) **42**(6), 1–11 (2023)
36. Roessle, B., Barron, J.T., Mildenhall, B., Srinivasan, P.P., Nießner, M.: Dense depth priors for neural radiance fields from sparse input views. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12892–12901 (2022)
37. Sarlin, P.E., Cadena, C., Siegwart, R., Dymczyk, M.: From coarse to fine: Robust hierarchical localization at large scale. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12716–12725 (2019)
38. Sarlin, P.E., DeTone, D., Malisiewicz, T., Rabinovich, A.: SuperGlue: Learning feature matching with graph neural networks. In: CVPR (2020)
39. Sattler, T., Leibe, B., Kobbelt, L.: Efficient & effective prioritized matching for large-scale image-based localization. IEEE transactions on pattern analysis and machine intelligence **39**(9), 1744–1756 (2016)
40. Schönberger, J.L., Frahm, J.M.: Structure-from-motion revisited. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2016)
41. Schönberger, J.L., Zheng, E., Pollefeys, M., Frahm, J.M.: Pixelwise view selection for unstructured multi-view stereo. In: European Conference on Computer Vision (ECCV) (2016)
42. Snavely, N., Seitz, S.M., Szeliski, R.: Photo tourism: exploring photo collections in 3d. In: ACM siggraph 2006 papers, pp. 835–846 (2006)
43. Snavely, N., Seitz, S.M., Szeliski, R.: Modeling the world from internet photo collections. International journal of computer vision **80**, 189–210 (2008)
44. Sweeney, C.: Theia multiview geometry library: Tutorial & reference. http://theia-sfm.org
45. Tancik, M., Weber, E., Ng, E., Li, R., Yi, B., Wang, T., Kristoffersen, A., Austin, J., Salahi, K., Ahuja, A., et al.: Nerfstudio: A modular framework for neural radiance field development. In: ACM SIGGRAPH 2023 Conference Proceedings. pp. 1–12 (2023)
46. Torr, P.H.: An assessment of information criteria for motion model selection. In: Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition. pp. 47–52. IEEE (1997)
47. Truong, P., Rakotosaona, M.J., Manhardt, F., Tombari, F.: Sparf: Neural radiance fields from sparse and noisy poses. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4190–4200 (2023)
48. Wang, F., Galliani, S., Vogel, C., Pollefeys, M.: Itermvs: Iterative probability estimation for efficient multi-view stereo. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8606–8615 (2022)

49. Wang, P., Liu, L., Liu, Y., Theobalt, C., Komura, T., Wang, W.: Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. NeurIPS (2021)
50. Wang, Y., Han, Q., Habermann, M., Daniilidis, K., Theobalt, C., Liu, L.: Neus2: Fast learning of neural implicit surfaces for multi-view reconstruction. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3295–3306 (2023)
51. Wang, Y., Skorokhodov, I., Wonka, P.: Hf-neus: Improved surface reconstruction using high-frequency details. Advances in Neural Information Processing Systems **35**, 1966–1978 (2022)
52. Wang, Z., Wu, S., Xie, W., Chen, M., Prisacariu, V.A.: Nerf–: Neural radiance fields without known camera parameters. arXiv preprint arXiv:2102.07064 (2021)
53. Wu, C.: Visualsfm: A visual structure from motion system. http://www. cs. washington. edu/homes/ccwu/vsfm (2011)
54. Wu, T., Wang, J., Pan, X., Xu, X., Theobalt, C., Liu, Z., Lin, D.: Voxurf: Voxel-based efficient and accurate neural surface reconstruction. arXiv preprint arXiv:2208.12697 (2022)
55. Xu, X., Yang, Y., Mo, K., Pan, B., Yi, L., Guibas, L.: Jacobinerf: Nerf shaping with mutual information gradients. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16498–16507 (2023)
56. Yan, Q., Wang, Q., Zhao, K., Chen, J., Li, B., Chu, X., Deng, F.: Cf-nerf: Camera parameter free neural radiance fields with incremental learning. arXiv preprint arXiv:2312.08760 (2023)
57. Yao, Y., Luo, Z., Li, S., Fang, T., Quan, L.: Mvsnet: Depth inference for unstructured multi-view stereo. European Conference on Computer Vision (ECCV) (2018)
58. Yao, Y., Luo, Z., Li, S., Zhang, J., Ren, Y., Zhou, L., Fang, T., Quan, L.: Blendedmvs: A large-scale dataset for generalized multi-view stereo networks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 1790–1799 (2020)
59. Yariv, L., Gu, J., Kasten, Y., Lipman, Y.: Volume rendering of neural implicit surfaces. In: Thirty-Fifth Conference on Neural Information Processing Systems (2021)
60. Yariv, L., Kasten, Y., Moran, D., Galun, M., Atzmon, M., Ronen, B., Lipman, Y.: Multiview neural surface reconstruction by disentangling geometry and appearance. Advances in Neural Information Processing Systems **33** (2020)
61. Zhang, K., Riegler, G., Snavely, N., Koltun, V.: Nerf++: Analyzing and improving neural radiance fields. arXiv:2010.07492 (2020)
62. Zhu, B., Yang, Y., Wang, X., Zheng, Y., Guibas, L.: Vdn-nerf: Resolving shape-radiance ambiguity via view-dependence normalization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 35–45 (2023)
63. Zhu, Z., Peng, S., Larsson, V., Xu, W., Bao, H., Cui, Z., Oswald, M.R., Pollefeys, M.: Nice-slam: Neural implicit scalable encoding for slam. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12786–12796 (2022)