Neural Active Structure-from-Motion in Dark and Textureless Environment

Kazuto Ichimaru[1,2], Diego Thomas[1], Takafumi Iwaguchi[1], and Hiroshi Kawasaki[1]

[1] Kyushu University, Japan
https://www.cvg.ait.kyushu-u.ac.jp/
[2] Fujitsu Defense & National Security Limited, Japan
https://www.fujitsu.com/jp/group/fdns/

Abstract. Active 3D measurement, especially structured light (SL) has been widely used in various fields for its robustness against textureless or equivalent surfaces by low light illumination. In addition, reconstruction of large scenes by moving the SL system has become popular, however, there have been few practical techniques to obtain the system's precise pose information only from images, since most conventional techniques are based on image features, which cannot be retrieved under textureless environments. In this paper, we propose a simultaneous shape reconstruction and pose estimation technique for SL systems from an image set where sparsely projected patterns onto the scene are observed (*i.e.* no scene texture information), which we call **Active SfM**. To achieve this, we propose a full optimization framework of the volumetric shape that employs neural signed distance fields (Neural-SDF) for SL with the goal of not only reconstructing the scene shape but also estimating the poses for each motion of the system. Experimental results show that the proposed method is able to achieve accurate shape reconstruction as well as pose estimation from images where only projected patterns are observed.

1 Introduction

For decades, active 3D measurement has been widely used in the field of autonomous vehicle control, human body analysis, industrial inspection, and so on. Among them, active stereo techniques, typically structured light (SL), have been widely used because of their simple configuration and high accuracy. On the other hand, the reconstruction of large scenes by moving a 3D sensor and integrating results has gained considerable attention for AR/VR applications, such as room reconstruction by smartphones or digital map-making for underwater environments. For this purpose, an active measurement system with an external positional sensor, such as an inertial measurement unit (IMU) or a camera, is commonly used [1, 40]. Specifically, structure-from-motion (SfM) is a key technique to achieve precise localization of the system only from an image set.

However, it sometimes fails when a room consists of low-texture/uniform walls or during the exploration of extremely dark environments, such as the deep sea. To solve the problem with minimal setup, one may consider using the camera which is used for an active 3D stereo system to estimate the pose of the sensor itself as shown in Figure 1.

In this configuration, the fundamental problem is whether it is possible to estimate camera poses *from observed projected patterns by an SL onto the scene*; we name it **Active Structure-from-Motion (Active SfM)** in the paper. Since

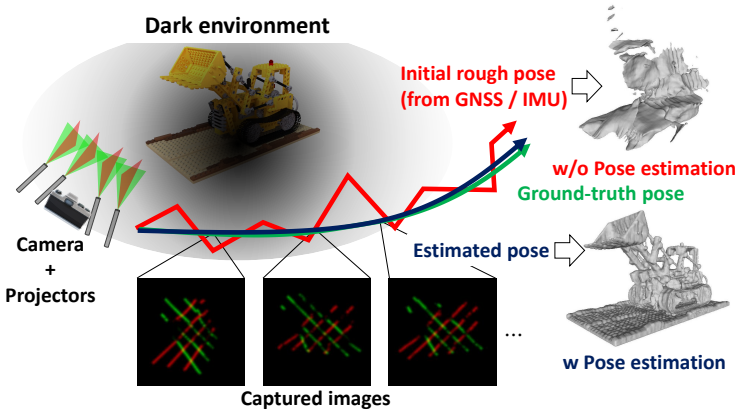


Fig. 1: Concept of Active Structure-from-Motion (Active SfM). The system consists of a camera and projectors. Images are captured in an extremely dark environment, where texture information is missing. Our goal is to recover the scene shape and the system poses with unreliable initialization from the projected patterns.

the area captured by a camera differs frame by frame, if there is no texture or dark environments, it is impossible to retrieve correspondences between frames, making it impossible to apply conventional SfM algorithms. Note that even if global navigation satellite system (GNSS) / IMU or other sensors can be used to obtain rough sensor poses, their accuracy is generally insufficient for precise shape integration.

Recently, Neural Radiance Fields (NeRF) [19] and its variants have drawn wide attention and brought breakthroughs into many computer vision tasks. By directly optimizing deep neural networks (DNN) to minimize a photometric loss in an end-to-end manner, they achieve remarkable accuracy on novel-view synthesis, 3D-shape reconstruction [16,34], super resolution [32], etc. Some also tried to integrate pattern projection into the NeRF pipeline to introduce neural fields for SL systems [15,28], however, they assume precise pre-calibration and dense reconstruction, which cannot be assumed for the sensor in motion.

In this paper, we propose a novel method to solve the Active SfM using a sparse SL system based on a NeRF variant under dark or even no illumination. Specifically, we propose a neural signed distance fields (Neural-SDF) that simultaneously estimates the 3D shape of the scene and the poses of the moving camera from the observed patterns projected by the SL system with unreliable initial poses. The technique is based on a novel volumetric rendering pipeline and hybrid encoding specialized for SL. Thanks to those proposed methods, it works in a scene where there is little texture or, in the most extreme case, no ambient illumination at all. Experiments were conducted to prove that the proposed method can solve the Active SfM from projected patterns for both synthetic and real data. Our contributions are as follows:

- We propose a novel Neural-SDF pipeline for Active SfM which enables both shape reconstruction and pose estimation of the SL system in motion from the projected pattern of SL and unreliable initial poses.
- In pursuit of fidelity and robustness, a volumetric rendering pipeline as well as a hybrid encoding technique for SL are proposed.

– Comprehensive experiments with both synthetic and real data were conducted to show the feasibility and effectiveness of the proposed method.

2 Related work

2.1 Neural Radiance Fields

Neural Radiance Fields (NeRF) is a methodology to represent a scene as a volumetric function that outputs the density of a 3D point and color from a multiple views [19]. NeRF utilizes DNN's ability of interpolation and extrapolation to accurately generate novel views from limited images, or estimate scene shape. While NeRF performs well on novel view synthesis, its performance on 3D-shape reconstruction is limited, since volumetric density function is not suitable to represent solid surfaces. VolSDF and NeuS replaced volumetric density function with signed distance function (SDF), which outputs signed distance to the closest surface [34,35,38]. Neural SDF drastically improved 3D-shape reconstruction accuracy by introducing inductive bias. Although some density function based methods achieved highly accurate reconstruction [25], SDF is suitable for various geometric down-stream tasks, such as normal reguralization by Eikonal loss [8], shape integration by TSDF [12], shape editing [31], and so on.

2.2 Structure-from-Motion

Structure-from-Motion (SfM) is a task to simultaneously estimate scene shape and camera motion. While conventional SfM methods based on image features have been successful [27], learning-based SfM and Differentiable Rendering (DR)-based SfM have been proposed in pursuit of higher accuracy and robustness [30, 41]. NeRF is also known that it can efficiently solve SfM-like problems. Yen-Chen *et al.* used NeRF as a pose estimator from known shape and rendering. Later, some found simultaneous shape reconstruction and pose estimation is possible [13, 17, 24, 36, 39]. Wang *et al.* also tried to estimate intrinsic parameters such as focal length [36], while Lin *et al.* proposed dedicated positional encoding to enhance the robustness of pose estimation [17]. However, none of them achieved shape reconstruction and pose estimation without texture information to the best of our knowledge.

One major challenge of SfM is robustness against dark and textureless environments. Conventional SfM uses image feature based localization, which is very challenging in such environments as shown in Figure 2 (Right). There are several NeRF methods that achieve NVS in dark environments, but they often require special image formats such as raw images or metadata [10,18]. Recently, LLNeRF achieved robust NVS under dark illumination environment without special image format, which is evaluated in the experiments [33]. Furthermore, there is no method that achieves pose estimation in such an environment, to the best of our knowledge.

2.3 Active Stereo for SfM

Active stereo is a methodology to estimate scene shape using active light sources. In a broad sense, active stereo may include projected pattern stereo [14] or photometric stereo [4], but we focus on structured light (SL) in this paper because of its simplicity and robustness. SL uses projected patterns as clues to find correspondences for triangulation. To eliminate the ambiguity of correspondences,

Fig. 2: Left: A failure case of ICP-based pose estimation with NeRF-Synthetic (Lego) scene with cross laser projectors. **Green arrows**: Ground-truth camera poses. **Red arrows**: Estimated camera poses via ICP with sparsely reconstructed point clouds. Note that the Ground-truth poses are used for initialization. **Right:** A failure case of SuperGlue [26] feature matching with NeRF-Synthetic (Lego) scene with little illumination (No matches are detected). Note that the contrast is enhanced for visualization.

multiple patterns have been commonly used for SL, such as Graycode [11], Phase-shifting [29]. Recently, investigations into reducing the number of patterns through optimization have been conducted [2,20]. To apply SL for dynamic scenes, such as SfM, it is preferable to use single pattern, known as oneshot scan, for example, grid-pattern [7], random dot pattern [9], colorful line pattern [5] and cross laser pattern [22,23].

To reconstruct the shapes of large-scale scenes or the entire shape of objects, it is possible to scan the scene with a one-shot scan followed by shape registration using the iterative closest point (ICP) algorithm [12]. However, in a one-shot scan, as positional information is spatially encoded into the pattern, a trade-off arises between robustness/accuracy and density in shape reconstruction. Therefore, the pursuit of greater accuracy leads to sparser reconstruction, potentially resulting in ICP algorithm unstable or even causing occasional failures as demonstrated in Figure 1 and Figure 2 (Left). Some work used multiple cross-laser projectors instead of an ordinary video projector to achieve simultaneous dense shape reconstruction and self-calibration of the system [6,22]. In [22], it is said that the problem has essentially 4DOF indeterminacy and extra constraints are required to solve the problem, proposing to use 4 cross-laser projectors, and we follow their system configuration. However, [22] uses pre-trained laser detection model, which requires labor-some annotation.

Recently, NeRF pipelines for SL systems have been proposed to incorporate into NeRF the robustness of SL systems against insufficient scene texture and illuminance. Li *et al*. combined Graycode SL [11] with Neural SDF to cope with an inter-reflection problem, which severely degrades reconstruction quality [15]. Shandilya *et al*. also integrated a Random dot pattern projector into the NeRF pipeline and proposed a model to separate ambient light, direct illumination, and indirect illumination from the light source [28]. Inspired by the current progress of NeRF / Neural SDF techniques which enabled simultaneous shape reconstruction and camera pose estimation, we propose a Neural SDF pipeline for SfM with SL, including extremely sparse patterns such as several number of line-lasers.

3 Method Overview

3.1 System configuration and environmental assumption

The proposed method assumes system configuration with a camera and an arbitrary number of projectors whose intrinsic parameters are accurately calibrated,

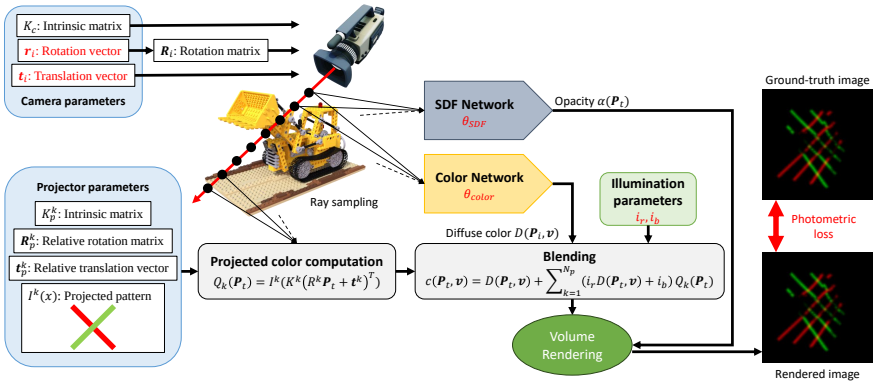


Fig. 3: Pipeline schematics of the proposed method. Parameters marked in red are optimized during training.

as shown in Figure 1. Relative transformations between the camera and the projectors are fixed and calibrated during the scan. We assume rough system poses are available for initialization from external devices such as GNSS or IMU, but their accuracy is insufficient for shape reconstruction and integration as shown in the experiments. The projectors project arbitrary patterns including very sparse patterns such as cross-line-lasers, which is used in the experiment. As for the environment, we assume the scenes are static during scan, and Lambertian reflectance is dominant on material surfaces.

3.2 Pipeline of the proposed method

The proposed method mainly follows NeuS [34] pipeline, which consists of the SDF network and the color network, except that our pipeline has **Projector parameters**, **Projected patterns** and **Illumination parameters**.Figure 3 shows the pipeline of the proposed method. Note that we omit some modules from the figure used in the pipeline, such as hierarchical sampling and variance network, which are also used by NeuS.

The training procedure of the pipeline is as follows. Bold lines are the processes for the proposed method, while others are almost identical to common NeRF pipeline.

1. Randomly sample rays casting from each camera's optical centers using camera parameters.
2. Sample 3D points on the ray from the near clip to the far clip at regular or weighted intervals.
3. Pass the 3D points to the SDF / color networks to acquire the density and albedo of the points.
4. **Render images by volumetric rendering with pattern projection**.
5. Compute photometric loss between the rendered images and the ground-truth (GT) images.
6. **Update the network parameters and the system poses to minimize the photometric loss using Adam**.

By updating the network parameters, SDF is optimized to minimize the discrepancy between the projected pattern and the GT image, *i.e.* implicitly

searching image-pattern correspondences and computing scene depth from the views. At the same time, system poses are optimized to maximize multi-view depth consistency.

4 Neural SDF for Active SfM

4.1 Ray sampling

Given an image, we randomly sample N pixel coordinates (p_x, p_y), and convert to homogeneous coordinates $\boldsymbol{p} = (p_x, p_y, 1)$. Next, we backcast the coordinates to compute the ray direction vector $\boldsymbol{P}$ and ray origin $\boldsymbol{O}$ in world coordinate system using the camera's intrinsic matrix K_c, i-th system rotation matrix $\boldsymbol{R}_i$ and translation vector $\boldsymbol{t}_i$ as follows:

$$\begin{aligned} \boldsymbol{P} &= \boldsymbol{R}_i \cdot K_c^{-1} \cdot \boldsymbol{p}^T, \\ \boldsymbol{O} &= s\boldsymbol{t}_i, \end{aligned} \tag{1}$$

where s is the predefined scaling coefficient to fit the entire scene into the unit sphere. Finally, we sample 3D points along the ray,

$$\boldsymbol{P}_u = \boldsymbol{P} \cdot d(u) + \boldsymbol{O} \tag{2}$$

where $d(u)$ returns a distance from the origin $\boldsymbol{O}$. As mentioned below, we implement the ray sampling process in a fully differentiable way to backpropagate photometric loss to the system pose parameters $\boldsymbol{R}_i$ and $\boldsymbol{t}_i$.

4.2 Volumetric rendering with pattern projection

In ordinary Neural SDF, a 2D point p is rendered using 3D points $\boldsymbol{P}_i$ $(i = 1..n)$ on the ray cast from $\boldsymbol{p}$ as Equation 3

$$C = \sum_{t=1}^{n} \left(\prod_{j=1}^{t-1} 1 - \alpha(\boldsymbol{P}_j) \right) \alpha(\boldsymbol{P}_t) c(\boldsymbol{P}_t, \boldsymbol{v}), \tag{3}$$

where $\alpha(\boldsymbol{P}_t)$ is opacity of a 3D point $\boldsymbol{P}_t$, $c(\boldsymbol{P}_t, \boldsymbol{v})$ is color of $\boldsymbol{P}_t$ viewing from direction $\boldsymbol{v}$, and C is final rendered color of the ray. Following NeuS, $\alpha(\boldsymbol{P}_t)$ is computed as Equation 4,

$$\alpha(\boldsymbol{P}_t) = \max\left(\frac{\Phi_s(f(\boldsymbol{P}_t)) - \Phi_s(f(\boldsymbol{P}_{t+1}))}{\Phi_s(f(\boldsymbol{P}_t))}, 0 \right), \tag{4}$$

where $f(\boldsymbol{P})$ is SDF value of a 3D point $\boldsymbol{P}$ and Φ_s is Sigmoid function in our implementation.

To compute $c(\boldsymbol{P}_t, \boldsymbol{v})$, the color of a 3D point with pattern projection, we follow a common linear Lambertian model [3]. Specifically, we blend albedo color of $\boldsymbol{P}_t$ (denoted as $D(\boldsymbol{P}_t, \boldsymbol{v})$) and projected color by k-th projector (denoted as $Q_k(\boldsymbol{P}_t)$) as Equation 5,

$$c(\boldsymbol{P}_t, \boldsymbol{v}) = D(\boldsymbol{P}_t, \boldsymbol{v}) + \sum_{k=1}^{N_p} (i_r D(\boldsymbol{P}_t, \boldsymbol{v}) + i_b) Q_k(\boldsymbol{P}_t), \tag{5}$$

where i_r is reflectance coefficient, and i_b is bias coefficient, integrated as learnable parameters, and N_p is the number of projectors. i_r controls reflectance level of $\boldsymbol{P}_t$, and i_b controls emissive level of $Q(\boldsymbol{P}_t)$. Once i_r and i_b are learned, we can render a high-fidelity image with pattern projection.

In our implementation, $D(\boldsymbol{P}_t, \boldsymbol{v})$ is the Color network itself, and $Q_k(\boldsymbol{P}_t)$ is computed as Equation 6,

$$Q_k(\boldsymbol{P}_t) = I^k(K^k(R^k\boldsymbol{P}_t + \boldsymbol{t}^k)^T), \tag{6}$$

where $I^k(x)$ returns the color of the pattern of k-th projector at specific point x via bilinear sampling, K^k is the intrinsic matrix and $R^k, \boldsymbol{t}^k$ are the relative transformation of k-th projector, which convert 3D points in the world coordinate system into the projector screen coordinate system.

4.3 System pose estimation

To estimate the system pose, we define the rendering pipeline as fully differentiable from the system's extrinsic parameters. The extrinsic parameters consist of rotation and translation (3 degrees of freedom, respectively). Surprisingly, such a simple modification enables pose estimation under severe conditions, such as extremely low illumination environments. We define rotations as rotation vectors to remove redundancy and convert them to rotation matrices to compute ray directions, as done in [36].

Note that we assume relative poses between the camera and the projectors are fixed and calibrated in most cases, but they can be also refined by defining them as learnable parameters. Please refer to the supplementary material for an experiment on projector pose refinement.

4.4 Hybrid encoding for robust and high-fidelity optimization

It is well-known that positional encoding plays an important role in NeRF-based methods to achieve higher accuracy and fidelity. While original NeRF and NeuS use Fourier encoding [19,34], InstantNGP and NeuS2 drastically improved reconstruction quality by introducing multi-resolution hash encoding [21, 35], computed as following,

$$\begin{aligned} e_l^h(x) &= \sum_{i=1}^{2^d} w_{i,l}\mathcal{H}_l(h(x)), \\ e^h(x) &= \left(h_1(x), h_2(x), \cdots, h_L(x)\right), \end{aligned} \tag{7}$$

where d is the dimension of the grids (usually $d = 3$), L is the number of resolutions, $\mathcal{H}_l$ is the hash entry of l-th resolution, and h is the arbitrary hash function. Multi-resolution hash encoding enables high-fidelity and extremely fast scene shape reconstruction by independently optimizing parameters for each hash entry. However, we empirically observed that using multi-resolution hash encoding in our pipeline severely degrades reconstruction and pose estimation quality due to its locality, causing a trade-off between the robustness of pose estimation and fidelity of the reconstructed shape.

To avoid such a phenomenon and still achieve higher accuracy and fidelity, we apply a modification to the positional encoding. We propose hybrid encoding,

i.e., concatenating embedded vectors from Fourier encoding and Multi-resolution hash encoding. Hybrid encoding $e(x)$ is represented as following,

$$e(x) = \left(e^f(x), e^h(x)\right), \tag{8}$$

where $e^f(x)$ is Fourier encoding, computed as follows,

$$\begin{aligned} e^f(x) = \big(&sin(2^0 x), sin(2^1 x), \cdots, sin(2^{L-1} x), \\ &cos(2^0 x), cos(2^1 x), \cdots, cos(2^{L-1} x)\big). \end{aligned} \tag{9}$$

Hybrid encoding relaxes the locality of multi-resolution hash encoding and improves reconstruction and pose estimation quality, as shown in the experiments.

4.5 Implementation details

We implemented the proposed method following NeuS implementation. We used hierarchical sampling [19], Eikonal loss [8], and mask loss [34] as well for better efficiency and accuracy. Eikonal loss is a well-known regularizer, which helps in learning a spatially consistent SDF. Overall, the objective function of the pipeline is as Equation 10

$$\mathcal{L} = \mathcal{L}_{color} + \lambda \mathcal{L}_{reg} + \beta \mathcal{L}_{mask}, \tag{10}$$

where $\mathcal{L}_{color}$ is photometric loss (L1), $\mathcal{L}_{reg}$ is Eikonal term, $\mathcal{L}_{mask}$ is mask loss term, and λ, β are balancing coefficients.

As for the hyper-parameters of training, we used learning rate $5e-4$, learning rate decay coefficient 0.05, batch size 512, Eikonal term coefficient $\lambda = 0.1$, and mask loss term coefficient $\beta = 0.1$. We trained the networks for 200k steps in the experiments.

5 Experiments

In this section, we describe the experiments with synthetic and real data to confirm the feasibility of the method.

5.1 Comparative methods

Through the experiments, we compare the proposed method with the following comparative methods. Please refer to the supplementary material for the implementation details. Note that comparison to NeuS, NeuS+Pose estim. and NeuS+SL covers ablation study of the proposed method.

- Light-sectioning: A method for active 3D shape reconstruction with sparse set of patterns, typically configured by line-laser projectors.
- LLNeRF [33]: A method for NVS under dark environments.
- NeuS [34] (Ours w/o pose estim. and pattern projection): A method for passive 3D shape reconstruction with SDF.
- NeuS+Pose estim. (Ours w/o pattern projection): A NeuS variant with the same pose estimation pipeline to the proposed method.
- NeuS+SL (Ours w/o pose estim.): A NeuS variant with structured light, which is identical to the proposed method without pose estimation.

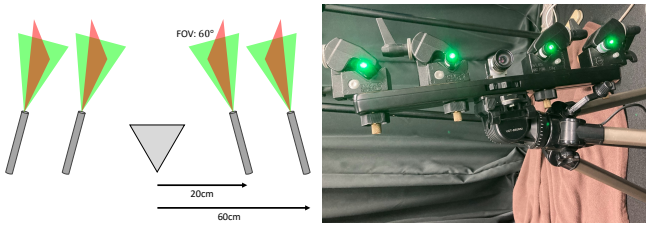
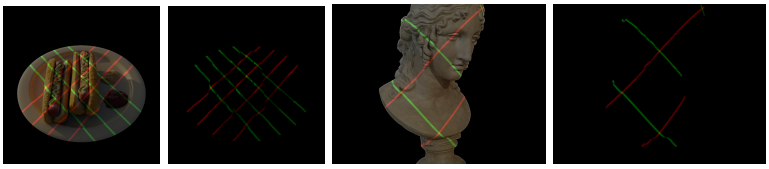


Fig. 4: System configuration of SL system in the experiments. Note that the lasers are colored red and green in the left figure to make the configuration understood easily, however, single color is sufficient as can be seen in the real system, all green (right).

Table 1: List of scenes of synthetic datasets.

Dataset	Scenes (# of images)	Baseline scalar
NeRF-Synthetic	Lego(40), Chair(40), Hotdog(40), Mic(40)	1.0
BlendedMVS	Stone(56), Dog(31), Bear(123), Sculpture(79)	0.2



Normal illum. No illum. Normal illum. No illum.

Fig. 5: Example images of the synthetic data with pattern projection. **Left**: NeRF-Synthetic. **Right**: BlendedMVS.

5.2 Evaluations with synthetic data

We conducted several experiments with synthetic data. Throughout the experiments, we used two datasets such as NeRF-Synthetic [19] and BlendedMVS [37], as shown in Table 1. For both datasets, we synthesized images with pattern projection by computing rays from virtual projectors. We used 4 virtual projectors lined up on the left and right of the camera with static cross-line-laser patterns (red and green) as shown in Figure 4 (left). The projectors are assumed to be fixed with the camera at a certain relative rotation and position, *i.e.* the projectors move with the camera. Baseline lengths of the projectors are 20cm and 60cm from the camera, and field-of-view is 60°. We synthesized each scene with different illumination conditions, such as normal illumination and no illumination. The no illumination scenes are completely dark without ambient light and have no textural information except where the patterns are projected. Figure 5 shows example images of the synthetic data with pattern projection.

Evaluation on reconstruction and pose estimation accuracy First, we evaluated reconstruction and pose estimation accuracy of the proposed method. We randomly added rotational and translational perturbations to the ground truth (GT) system poses, and measured mean L1 errors (rotation and translation) of the estimated poses to the GT poses and Chamfer distances of the reconstructed shapes to the GT shapes after training with the proposed method.

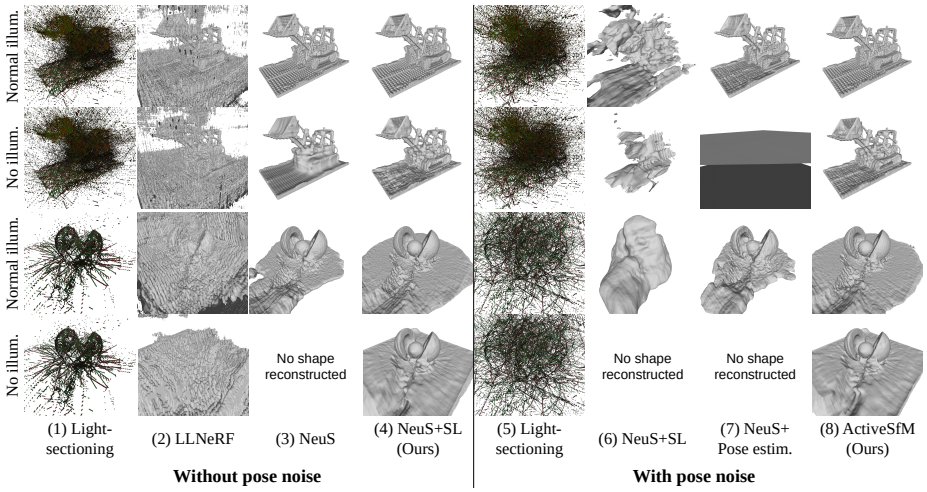


Fig. 6: Reconstructed shapes with various methods. Qualitatively, ours achieved the best accuracy under no illumination with pose noise condition. Quantitative results are shown in Table 2.

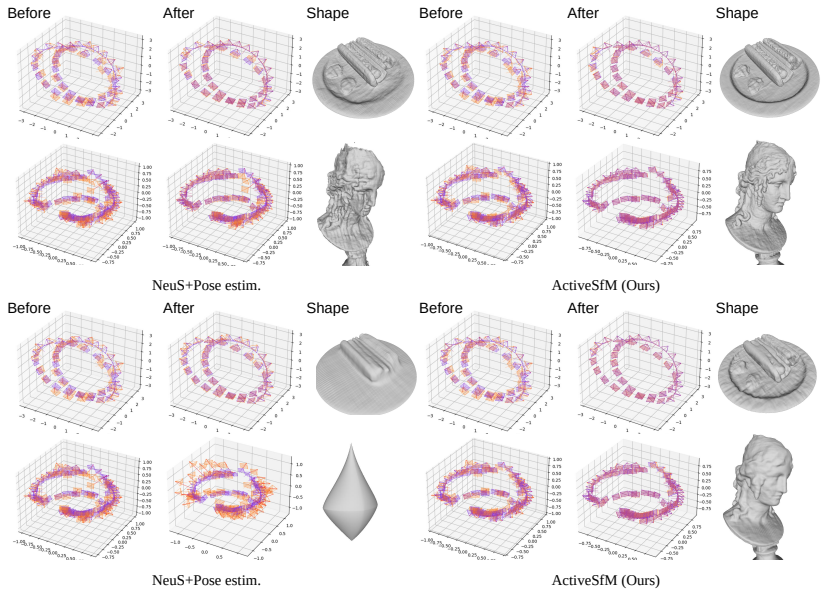


Fig. 7: Reconstruction and pose estimation results with the synthetic data. The left 3D map shows the initial system poses, and the right 3D map shows estimated system poses, for each scene. Blue frustums are the GT and orange frustum are the initial and estimated. **Top**: Normal illumination. **Bottom**: No illumination.

To remove ambiguity on global rotation, translation and scaling, we applied ICP with scaling on the reconstructed shapes to the GT shapes. As for perturba-

Table 2: Chamfer distances [mm] of the reconstructed shapes. Ours consistently outperforms comparative methods, especially when pose noise is added. "N/A" indicates no shape reconstructed. Legend: **Best**, Second best.

Pose noise	Illum	Method	NeRF-Synthetic Lego	Chair	Hotdog	Mic	BlendedMVS Stone	Dog	Bear	Sculpture
No	Normal	Light-sectioning	21.08	16.15	**14.00**	26.24	**17.40**	**6.02**	**6.95**	**4.84**
		LLNeRF	19.73	26.40	33.44	23.59	62.31	19.89	30.77	23.11
		NeuS	11.79	**9.25**	28.08	N/A	44.40	14.02	14.15	23.20
		NeuS+SL (Ours)	**11.62**	20.63	27.94	**5.82**	30.92	17.83	30.84	11.46
	No	Light-sectioning	21.08	16.15	**14.00**	26.24	**17.40**	**6.02**	**6.95**	**4.84**
		LLNeRF	23.14	32.36	40.30	24.14	87.18	42.10	50.65	24.62
		NeuS	21.74	12.12	32.27	58.73	N/A	N/A	86.05	12.52
		NeuS+SL (Ours)	**14.97**	**8.51**	26.40	**5.58**	35.35	18.30	25.63	11.21
Yes	Normal	Light-sectioning	39.88	40.56	43.72	39.90	96.97	25.31	81.32	20.80
		NeuS+Pose estim.	13.73	50.97	29.10	N/A	59.51	28.04	**23.17**	28.57
		NeuS+SL	41.47	49.33	51.36	50.34	90.85	27.62	55.49	27.64
		ActiveSfM (Ours)	**13.07**	**36.93**	**22.41**	**12.13**	**31.84**	**19.89**	23.40	**10.75**
	No	Light-sectioning	39.88	40.56	43.72	39.90	96.97	25.31	81.32	20.80
		NeuS+Pose estim.	59.75	48.45	61.16	22.69	N/A	29.31	81.40	32.26
		NeuS+SL	35.04	44.76	52.76	N/A	N/A	56.93	N/A	70.74
		ActiveSfM (Ours)	**18.50**	**18.60**	**28.47**	**6.37**	**35.36**	**21.67**	**22.99**	**11.20**

Table 3: Mean L1 errors of the estimated poses (rotation and translation). "N/A" indicates no shape reconstructed. Legend: **Best**.

Illum	Metric	Method	NeRF-Synthetic Lego	Chair	Hotdog	Mic	BlendedMVS Stone	Dog	Bear	Sculpture
Normal	Rot[°]	NeuS+Pose estim.	0.84	3.90	1.50	N/A	0.60	3.02	0.96	2.40
		ActiveSfM (Ours)	**0.33**	**1.18**	**1.29**	**0.88**	**0.52**	**0.76**	**0.38**	**0.67**
	Trans[%]	NeuS+Pose estim.	4.25	25.63	9.61	N/A	1.37	7.83	1.53	4.44
		ActiveSfM (Ours)	**1.26**	**7.29**	**5.59**	**3.38**	**0.09**	**0.62**	**0.43**	**0.56**
No	Rot[°]	NeuS+Pose estim.	6.79	7.18	7.86	77.63	N/A	7.99	7.57	5.80
		ActiveSfM (Ours)	**0.50**	**1.07**	**1.02**	**0.63**	**0.53**	**1.01**	**0.42**	**0.90**
	Trans[%]	NeuS+Pose estim.	25.80	23.86	25.20	52.39	N/A	10.71	8.60	17.77
		ActiveSfM (Ours)	**1.23**	**3.70**	**2.97**	**2.32**	**0.41**	**1.78**	**0.48**	**0.71**

tions, we used uniform noises with a constant value for rotation, and relative value to the mean baseline lengths between frames for translation, which are multiplied by baseline scaler as shown in Table 1 since BlendedMVS images are not ordered. The ranges of the uniform noises are 10° for rotation, and 50% for translation. For more evaluations with different noise ranges, and the detailed procedure of noise addition, please refer to the supplementary material.

Table 2 shows the results of the quantitative comparison, and Figure 6 shows the results of the qualitative comparison. As shown in the results, the proposed method achieved the best accuracy under no illumination with pose noise condition. Light-sectioning method is advantageous in low-light environment without pose noise, however, dense reconstruction is not possible and shapes are significantly corrupted with pose noise. LLNeRF is capable of handling low-light illumination as well, however, it produced severe floating objects since it is not dedicated for shape reconstruction. NeuS and NeuS+Pose estimation produced plausible shapes in some scenes, but they failed under no illumination scenario. NeuS+SL could reconstruct accurate shapes only when there are no pose noise (Note that NeuS+SL is identical to ActiveSfM without pose noise). Finally, ActiveSfM (Ours) successfully reconstructed accurate shapes in all scenes and

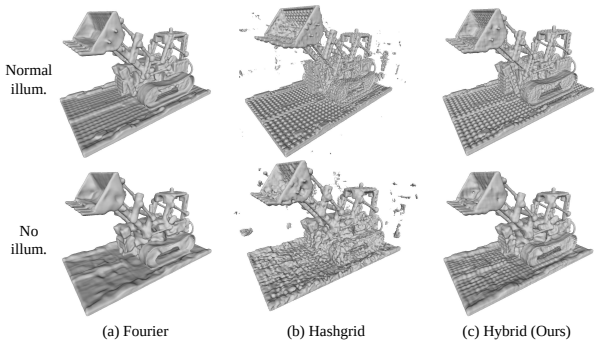


Fig. 8: Reconstructed shapes with various positional encoding with pose estimation.

scenarios, showing the feasibility of the proposed method. As for the pose estimation accuracy, the proposed method achieved higher accuracy compared to NeuS+Pose estimation in all scenes, thanks to implicit depth supervision by pattern projection (Table 3 and Figure 7).

Table 4: Results of the ablation study on positional encoding. **Best**, Second best.

Illum	Encoding	Rot[°]	Trans[%]	Shape[mm]
Normal	Fourier	0.53	**1.04**	15.84
	Hashgrid	4.73	34.98	45.63
	Hybrid	**0.33**	1.26	**13.07**
No	Fourier	0.65	1.37	22.87
	Hashgrid	1.31	1.45	23.58
	Hybrid	**0.50**	**1.23**	**18.50**

Evaluation on positional encoding Next, we conducted an ablation study on positional encoding. We compared the reconstructed shapes with Fourier encoding, multi-resolution hash encoding (Hashgrid), and the proposed hybrid encoding (Hybrid).

Figure 8 and Table 4 show quantitative and qualitative comparisons of the proposed method against other positional encoding. We observed that hybrid encoding consistently improves pose estimation and shape reconstruction in both normal and no illuminations. In particular, observing the qualitative results, we can confirm major improvements in the fidelity of reconstructed shapes and reduction of floating objects which have little impact on the Chamfer distances.

5.3 Evaluations with real data

To confirm the feasibility of the proposed method in real scenes, we captured two real sequences with pattern projection by 4 cross laser projectors as shown in Figure 4, in both normal and no illuminations. First, we used a turn table to

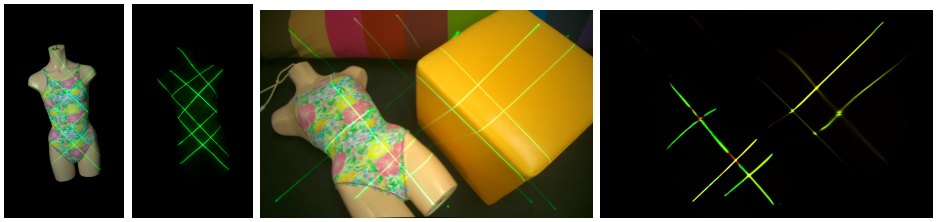


Fig. 9: Example of the real data (normal and no illumination). **Left**: Controlled sequence. **Right**: Freehand sequence. Contrast is enhanced for visualization.

rotate a mannequin (*i.e.*, virtually moving the system around the mannequin), and captured images per 10° in two illumination conditions (36 images per illumination condition, Figure 9(left)). Next, we used COLMAP to obtain the GT system poses relative to the mannequin using normal illumination (Note that we did not use the shapes obtained by COLMAP as GT). To remove a scale ambiguity, we applied a light sectioning method [22] and adjusted the scale to fit the reconstruction from light sectioning into a point cloud from COLMAP. Note that, since we captured images in two illumination conditions per viewpoint, we also have GT for the no illumination scene. Then, we evaluated the proposed method using the sequences with rotational and translational pose noises. As for the GT shape, we captured the same object using a ToF sensor and KinectFusion [12].

Similarly, we captured a scene with a freehand trajectory (78 images) to further evaluate the robustness of the method (Figure 9(right)). In the capturing process, the system was moved freely around the static scene by hand. We captured a sequence in the normal illumination as a reference to obtain the GT poses by COLMAP and synthesized a no illumination sequence by reducing textures other than laser curves. Then, we evaluated our method using the data by adding noise on rotation and translation of the poses as same as the controlled sequence.

Table 5: Qualitative comparison results of pose estimation and shape reconstruction on real data. Legend: **Best**, Second best. "-" indicates rotational and translational error does not exist since there is no pose noise added.

Pose noise	Illum	Method	Controlled sequence			Freehand sequence	
			Rot[°]	Trans[%]	Shape[mm]	Rot[°]	Trans[%]
No	Normal	Light-sectioning	-	-	13.33	-	-
		LLNeRF	-	-	15.93	-	-
		NeuS	-	-	**9.39**	-	-
		NeuS+SL(Ours)	-	-	9.55	-	-
	No	Light-sectioning	-	-	13.33	-	-
		LLNeRF	-	-	15.87	-	-
		NeuS	-	-	20.67	-	-
		NeuS+SL(Ours)	-	-	**9.39**	-	-
Yes	Normal	NeuS+Pose estim.	0.87	15.16	**9.27**	2.22	52.77
		ActiveSfM (Ours)	**0.70**	**5.27**	9.37	**1.87**	**43.69**
	No	NeuS+Pose estim.	4.94	742.14	24.41	2.18	64.60
		ActiveSfM (Ours)	**2.02**	**17.19**	**12.29**	**1.27**	**18.89**

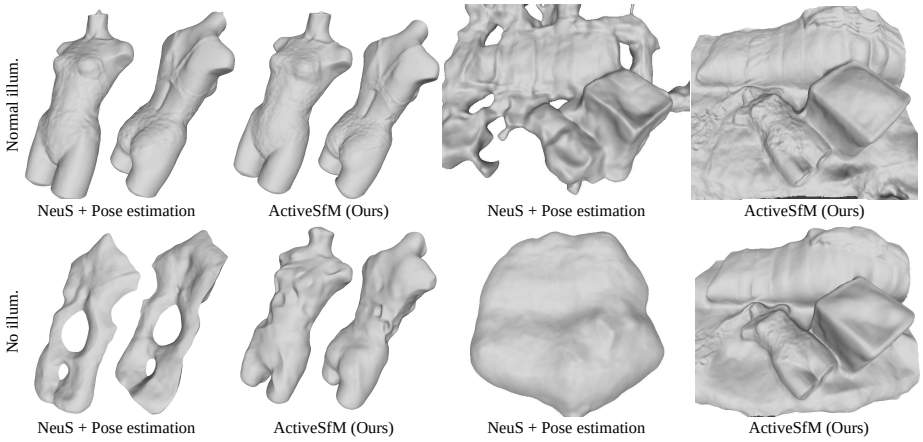


Fig. 10: Reconstructed shapes on the real scenes with pose noise. NeuS + Pose estimation suffered from severe collapse except controlled sequence in normal illumination, while ours produced plausible shapes for all scenarios.

Figure 10 and Table 5 show the result. The proposed method performed accurate shape reconstruction and pose estimation compared to NeuS+Pose estimation, especially under no illumination scenario. Although the reconstructed shapes is a bit bumpy by our method under no illumination scenario, the shapes and poses were mostly correctly estimated; we consider it is mainly because of indirect illumination effects and it is important topic for future work. From the table, translation error is considerably large under normal illumination compared to no illumination on the freehand sequence; we consider it is because small pose errors by COLMAP.

6 Conclusion

In this paper, we proposed a simultaneous shape reconstruction and pose estimation method for SL systems, which we call Active SfM, using Neural SDF. To achieve it, we proposed a volumetric rendering pipeline for SL and introduced hybrid encoding for robust pose estimation and high-fidelity shape reconstruction. Experimental results show the proposed method can efficiently recover the scene geometry only from the information of projected patterns with rough initial poses in both synthetic and real dataset. As for the future work, we are interested in whether Neural SDF can cope with other challenging conditions, such as scattering, inter-reflection, occlusion, etc. Concurrently, it is also important to address many other problems such as distortion due to refraction, attenuation, volumetric scattering to achieve an accurate deep sea vision system.

Acknowledgments This work was part supported by JST Startup JPMJSF23DR, ACT-X JPMJAX23C2 and JSPS/KAKENHI JP20H00611 and JP23H03439 in Japan.

References

1. Alzuhiri, M., Li, Z., Rao, A., Li, J., Fairchild, P., Tan, X., Deng, Y.: Imu-assisted robotic structured light sensing with featureless registration under uncertainties for pipeline inspection. NDT & E International **139**, 102936 (October 2023) 1
2. Chen, W., Mirdehghan, P., Fidler, S., Kutulakos, K.N.: Auto-tuning structured light by optical stochastic gradient descent. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2020) 4
3. Cho, S.Y., Chow, T.: A neural-learning-based reflectance model for 3-d shape reconstruction. IEEE Transactions on Industrial Electronics (2000) 6
4. Clark, J.: Active photometric stereo. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition (1992) 3
5. Fernandez, S., Salvi, J.: A novel structured light method for one-shot dense reconstruction. In: IEEE International Conference on Image Processing (2012) 4
6. Furukawa, R., Kawasaki, H.: Laser range scanner based on self-calibration techniques using coplanarities and metric constraints. Computer Vision and Image Understanding **113**(11), 1118–1129 (2009) 4
7. Furukawa, R., Mikamo, M., Sagawa, R., Kawasaki, H.: Single-shot dense active stereo with pixel-wise phase estimation based on grid-structure using cnn and correspondence estimation using gcn. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 4001–4011 (January 2022) 4
8. Gropp, A., Yariv, L., Haim, N., Atzmon, M., Lipman, Y.: Implicit geometric regularization for learning shapes. In: Proceedings of Machine Learning and Systems 2020, pp. 3569–3579 (2020) 3, 8
9. Gu, F., Song, Z., Zhao, Z.: Single-shot structured light sensor for 3d dense and dynamic reconstruction. Sensors **20**(4), 1094 (2020) 4
10. Huang, X., Zhang, Q., Feng, Y., Li, H., Wang, X., Wang, Q.: Hdr-nerf: High dynamic range neural radiance fields. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18398–18408 (2022) 3
11. Inokuchi, S., Sato, K., Matsuda, F.: Range imaging system for 3-d object recognition. In: International Conference on Pattern Recognition (1984) 4
12. Izadi, S., Kim, D., Hilliges, O., Molyneaux, D., Newcombe, R., Kohli, P., Shotton, J., Hodges, S., Freeman, D., Davison, A., Fitzgibbon, A.: Kinectfusion: Real-time 3d reconstruction and interaction using a moving depth camera. In: UIST '11 Proceedings of the 24th annual ACM symposium on User interface software and technology. pp. 559–568. ACM (October 2011), https://www.microsoft.com/en-us/research/publication/kinectfusion-real-time-3d-reconstruction-and-interaction-using-a-moving-depth-camera/ 3, 4, 13
13. Jeong, Y., Ahn, S., Choy, C., Anandkumar, A., Cho, M., Park, J.: Self-calibrating neural radiance fields. In: ICCV (2021) 3
14. Konolige, K.: Projected texture stereo. In: IEEE International Conference on Robotics and Automation (2010) 3
15. Li, C., Hashimoto, T., Matsumoto, E., Kato, H.: Multi-view neural surface reconstruction with structured light. In: The British Machine Vision Conference (BMVC) (2022) 2, 4
16. 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) 2
17. Lin, C.H., Ma, W.C., Torralba, A., Lucey, S.: Barf: Bundle-adjusting neural radiance fields. In: IEEE International Conference on Computer Vision (ICCV) (2021) 3
18. Mildenhall, B., Hedman, P., Martin-Brualla, R., Srinivasan, P.P., Barron, J.T.: NeRF in the dark: High dynamic range view synthesis from noisy raw images. CVPR (2022) 3
19. 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) 2, 3, 7, 8, 9

20. Mirdehghan, P., Chen, W., Kutulakos, K.N.: Optimal structured light a la carte. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (2018) 4
21. 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 7
22. Nagamatsu, G., Ikeda, T., Iwaguchi, T., Thomas, D., Takamatsu, J., Kawasaki, H.: Self-calibration of multiple-line-lasers based on coplanarity and epipolar constraints for wide area shape scan using moving camera. In: International Conference on Pattern Recognition (ICPR) (2022) 4, 13
23. Nagamatsu, G., Takamatsu, J., Iwaguchi, T., Thomas, D., Kawasaki, H.: Self-calibrated dense 3d sensor using multiple cross line-lasers based on light sectioning method and visual odometry. In: IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) (2021) 4
24. Park, K., Henzler, P., Mildenhall, B., Barron, J.T., Martin-Brualla, R.: Camp: Camera preconditioning for neural radiance fields. ACM Trans. Graph. (2023) 3
25. Rakotosaona, M.J., Manhardt, F., Arroyo, D.M., Niemeyer, M., Kundu, A., Tombari, F.: Nerfmeshing: Distilling neural radiance fields into geometrically-accurate 3d meshes (2023) 3
26. Sarlin, P.E., DeTone, D., Malisiewicz, T., Rabinovich, A.: SuperGlue: Learning feature matching with graph neural networks. In: CVPR (2020), https://arxiv.org/abs/1911.11763 4
27. Schönberger, J.L., Frahm, J.M.: Structure-from-motion revisited. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2016) 3
28. Shandilya, A., Attal, B., Richardt, C., Tompkin, J., O'Toole, M.: Neural fields for structured lighting. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (2023) 2, 4
29. Srinivasan, V., Liu, H.C., Halioua, M.: Automated phase-measuring profilometry of 3-d diffuse objects. Applied Optics **23**, 3105–3108 (1984) 4
30. Teed, Z., Deng, J.: DeepV2D: Video to depth with differentiable structure from motion. In: Proceedings of The International Conference on Learning Representations (ICLR) (2020) 3
31. Tzathas, P., Maragos, P., Roussos, A.: 3d neural sculpting (3dns): Editing neural signed distance functions. In: 2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). IEEE (January 2023) 3
32. Wang, C., Wu, X., Guo, Y.C., Zhang, S.H., Tai, Y.W., Hu, S.M.: Nerf-sr: High-quality neural radiance fields using supersampling. ACM International Conference on Multimedia (2022) 2
33. Wang, H., Xu, X., Xu, K., Lau, R.W.: Lighting up nerf via unsupervised decomposition and enhancement. In: ICCV (2023) 3, 8
34. 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) 2, 3, 5, 7, 8
35. 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 (ICCV) (2023) 3, 7
36. 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) 3, 7
37. 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. Computer Vision and Pattern Recognition (CVPR) (2020) 9
38. Yariv, L., Gu, J., Kasten, Y., Lipman, Y.: Volume rendering of neural implicit surfaces. In: Thirty-Fifth Conference on Neural Information Processing Systems (2021) 3
39. Zhang, J., Zhan, F., Yu, Y., Liu, K., Wu, R., Zhang, X., Shao, L., Lu, S.: Pose-free neural radiance fields via implicit pose regularization. In: Proceedings of the

IEEE/CVF International Conference on Computer Vision (ICCV). pp. 3534–3543 (October 2023) 3
40. Zhou, J., Ji, Z., Li, Y., Liu, X., Yao, W., Qin, Y.: High-precision calibration of a monocular-vision-guided handheld line-structured-light measurement system. Sensors **23**(14), 6469 (2023) 1
41. Zhou, T., Brown, M., Snavely, N., Lowe, D.G.: Unsupervised learning of depth and ego-motion from video. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2017) 3