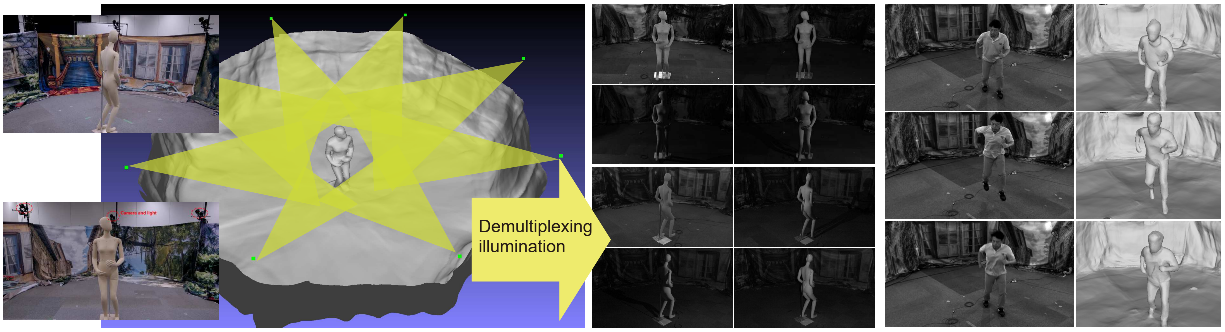
Neural 4D Scene Reconstruction with Multiple One-Shot Scanning Systems

Ryusuke Sagawa
AIST

Kota Nishihara
Kyushu University

Takafumi Iwaguchi
Kyushu University

Hiroshi Kawasaki
Kyushu University



Observation of whole scene with sparse cameras and light sources
Illuminating a target simultaneously with lights

Multi-channel images of the target
captured with different illumination

Shape reconstruction of the whole
dynamic scene from multi-channel images

Figure 1. Our method can reconstruct dynamic scene (jumping human) using only eight cameras and eight point light sources.

Abstract

Recently, 3D reconstruction from multiview stereo (MVS) has advanced significantly with the introduction of neural implicit representation methods, which estimate voxel densities or signed distance fields (SDFs) to describe the 3D structure of a scene. Although such neural-based methods typically require a large number of captured images to estimate dense volumetric information during training, developing systems that can recover the 3D shape of moving objects using only a small number of stationary cameras remains highly demanding and challenging. To address the issue of sparse views, various active lighting techniques have been proposed. However, the problem remains inherently difficult, particularly when attempting to capture the complete shape of an object with a wide baseline. In this paper, we propose a novel approach that combines active lighting with photometric stereo (PS) using neural representations. Additionally, we introduce a multiplexed illumination technique that captures the entire shape of an object in a single shot. Although this results in a low signal-to-noise ratio (SNR), our method also addresses this issue. The advantages of our technique are demonstrated through real-world experiments, showcasing its ability to capture a 4D scene.

1. Introduction

Multiview stereo (MVS) is widely used to capture 3D information of a large scene or the complete shape of an object [11, 15, 49]. MVS methods involve observing a scene from multiple viewpoints, establishing correspondences between images captured from different angles, and reconstructing surface points through triangulation. The advantage of MVS lies in its ability to observe a point in the scene from several viewpoints, which helps mitigate reconstruction errors by leveraging information from multiple views.

MVS methods are generally categorized into two types: moving-camera approaches [47, 48, 52] and stationary multi-camera setups [26, 55]. The former, often referred to as SfM/V-SLAM, has been extensively studied and has recently surged in popularity through neural-based methods such as NeRF, NeuS, and 3DGS [17, 24, 32, 60]. However, these methods assume static scenes and cannot, in principle, capture moving objects (4D scenes). In contrast, stationary multi-camera methods have mainly been developed for dynamic scene capture [22, 30, 58]. However, achieving dense and accurate reconstructions of 4D scenes typically requires a large number of cameras, leading to extremely high system setup costs as well as computational overhead. To reduce the number of cameras and computational demands, active-light-based approaches have been proposed [19, 44, 45]. These methods project high-frequency patterns, known as structured light (SL), onto objects to enhance surface texture and enable dense correspondence estimation. However, in practice, such attempts often fail, as it is nearly impossible to design projected patterns that are both dense and distinguishable at the same time.

In this paper, we address this issue by capturing scenes illuminated virtually by one light source at a time, thereby constraining the relationship between the image intensities observed from multiple viewpoints and the irradiance of the light source. When a scene is captured under varying illumination conditions, richer information about each surface point is obtained -— a concept known as photometric stereo (PS) [19]. Recently, this setting has gained attention as multi-view photometric stereo (MVPS) and has demonstrated promising results [67, 75]. However, most existing methods are prediction-based that require images captured under a large number of light sources and assume static scenes for shape integration. In contrast, our method is designed for 4D scenes and requires only a limited number of light sources by leveraging neural representations. Furthermore, we introduce a multiplexed illumination technique [46], which increases the number of independent light sources while avoiding mutual interference.

We demonstrate the effectiveness of the proposed system through extensive experiments on synthetic datasets, comparing against prior techniques, and further validate its feasibility in real-world large-scale environments with both static and dynamic objects (*e.g*., moving humans).

2. Related Work

MVS methods have recently seen significant improvements through neural-based approaches [1, 2, 32], which use multilayer perceptrons (MLPs) to predict density and color for a given 3D point and viewing direction. These approaches have been extended to improve training/rendering speed [35], camera parameter estimation [29, 63], surface mesh reconstruction via SDF [60], and reflection parameter estimation [31, 57]. Although NeRF was originally developed for view synthesis, it lacks regularization. Wide-baseline stereo, however, requires regularization for effective view interpolation. To address this, we generate a view-independent surface rather than synthesized views as a form of regularization. In addition, several NeRF-based methods have been proposed for visualizing dynamic objects [38, 39, 51], which use multiple frames to regularize deformation. In contrast, our work focuses on reconstructing shapes from only a few images at a single moment.

Sparse Views: MVS and NeRF-based methods find correspondences between 3D scene points and 2D image points by minimizing color differences, but they require images from dense viewpoints. Using many fixed cameras to capture dense views is costly, so some NeRF-based methods address this with regularization when input data is limited. One approach pre-trains the model on multiview datasets to use as a scene prior [5, 6, 41, 61, 71]. Diet-NeRF [20] incorporates CLIP-based similarity constraints, and RegNeRF [36] introduces geometry and photometry constraints, the latter trained on JFT-300M [53]. These methods rely on large training datasets and similarity between the training set and the target scene. More recently, feed-forward sparse-view reconstruction methods [21, 54, 59, 62] have shown promise; however, there is a trade-off between accuracy and convenience compared to measurement-based approaches.

The second approach assumes object shape distributions. NeRS [73] assumes spherical topology, while DS [13] optimizes a mesh generated from object masks. InfoNeRF [25] minimizes ray entropy, assuming each ray intersects the surface once, and Cerkezi and Favaro [4] use object-centric ray sampling, combining mesh and volume representations but requiring objects to be scene-centered. FreeNeRF [68] adds frequency and occlusion regularization during training, which is applied during training and combined with the approach proposed in this paper.

The third approach for NeRF-based 3D reconstruction from limited images uses additional data like depth, which provides crucial density information along camera rays. Depth from SfM (*e.g*., COLMAP [47]) has been applied as a constraint in works such as [8, 42, 64]. For RGB-D data, neural fields are trained by synthesizing images [72]. Similar methods [7, 9, 56, 66] leverage depth from 3D Gaussian splatting (3DGS) [24] using monocular [12, 69] or multiview depth [62] for sparse view reconstruction. Although these methods explicitly rely on depth, our approach achieves a similar effect implicitly through active lighting based on the Photometric Stereo (PS) technique.

Active Lighting and PS: Several methods for 3D reconstruction based on neural representations combined with active lighting have been proposed. When correspondences between camera images and projected patterns are available, this shape information can be utilized to optimize the neural representation [27]. If these correspondences are not provided beforehand, they can be determined by optimizing the neural representation [10, 18, 40, 50]. While these methods leverage pattern information to achieve high reconstruction accuracy, the requirement of multiple known patterns is impractical, and a large image set is needed to train the MLP. In contrast, a method using simple point light sources as active lighting to leverage the Photometric Stereo (PS) technique has been proposed [16, 28, 37]. PS-NeRF [70] proposed an approach where normal vectors calculated by photometric stereo are used to regularize the prediction of the neural radiance field. Kaya *et al*. [23] used normal vectors estimated by photometric stereo as input to an MLP to predict color. Recently, multi-view photometric stereo (MVPS) has gained significant attention [3, 67, 75]; however, it still requires multiple images captured by a single camera under varying light source positions, and surface normal estimation remains prediction-based. In contrast, we propose using the light power prior to surface reflection as a loss function for the neural SDF, which implicitly facilitates object shape estimation through optimization with a limited number of views and light sources.

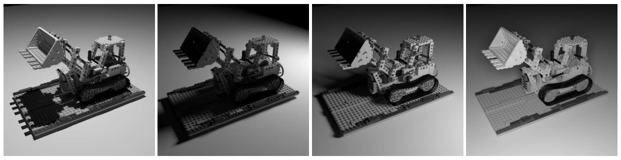


Figure 2. A scene is illuminated by one of multiple lights, respectively. In this example, scene is illuminated by seven different positions of point light sources.

Multiplexed Illumination: Multiplexed illumination decomposes images of a scene lit by multiple light sources into images for each individual source. Schechner *et al*. [46] used Hadamard-based multiplexing to enhance SNR by illuminating an object with multiple sources. Mukaigawa *et al*. [34] extended this to estimate an object's BRDF by illuminating from various directions. While typically applied to static scenes, Wenger *et al*. [65] introduced motion compensation via optical flow for moving subjects. Sagawa *et al*. [43] later applied signal processing to capture illumination on fast-moving objects under strong external light. By combining this approach, the proposed method can realize the 3D reconstruction of moving objects by using the images illuminated by different lights at the same time.

3. Method

3.1. Multiview Stereo with Multiple Light Sources

In this paper, we assume that a scene illuminated by multiple light sources is observed from multiple viewpoints. At each viewpoint, the images that the scene is illuminated only by one of the lights are captured. Since the same scene is observed only by changing the illumination, a multichannel image can be formed by stacking the images with different illumination. Fig. 2 shows an example of a scene illuminated by seven light sources. In this case, the light source positions are roughly calibrated and assumed to be known. Although each captured image can be any type of image, it is assumed to be single-channel image for simplicity. Therefore, an image of seven channels is formed from the images in Fig. 2. Changing the illumination contributes to improving the features to find the correspondence for MVS without increasing the number of viewpoints.

3.2. Estimating a Scene based on Volumetric Representation

NeRF-based methods estimate the information of a scene by fitting the parameters of MLP by minimizing the loss between captured images and the images generated by volume rendering. In the manner of volume rendering, the density at each 3D point is used to represent the shape of a scene.

As one of the types of volumetric representation, a signed distance function (SDF) $f(\boldsymbol{p})$ is often employed. Given a 3D position $\boldsymbol{p}$, the function returns a scalar value indicating the distance from the closest surface. If $\boldsymbol{p}$ lies outside of objects, $f(\boldsymbol{p})$ is positive; otherwise, it is negative. The surface shape is denoted by the set of points $\boldsymbol{p}$ where $f(\boldsymbol{p}) = 0$. The density can be calculated by the opaque density function $\rho(\boldsymbol{p}) = \phi(f(\boldsymbol{p}))$ from the SDF based on the derivative of the Sigmoid function defined as $\phi(x) = se^{-sx}/(1+se^{-sx})^2$, where s is a scale parameter.

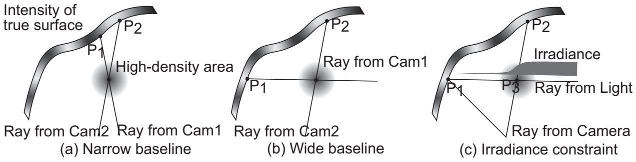


Figure 3. The difference in ray projection onto the true surface occurs when the high-density area is offset from the surface in the case of (a) narrow baseline, (b) wide baseline and (c) camera and light.

Now, let's assume that N light sources illuminate a scene from various positions, and M camera images are captured, and we consider the intensity $C(\boldsymbol{o}_k, \boldsymbol{v}_{kx})$ of a point x in the k-th camera image, where the camera position is $\boldsymbol{o}_k$ and the view direction of the point is $\boldsymbol{v}_{kx}$. The intensity of an image point is calculated as the integral along the ray of the point x as follows:

$$C(\boldsymbol{o}_k, \boldsymbol{v}_{kx}) = \int_0^{+\infty} w_k(\boldsymbol{p}_x(t)) c(\boldsymbol{p}_x(t), \boldsymbol{v}_{kx}) dt, \tag{1}$$

where $w(t)$ is the weight at the point $\boldsymbol{p}_x(t)$, which is a 3D point whose distance from the camera position is t along the ray. In our implementation, the weight is calculated based on NeuS [60] by

$$\begin{aligned} & w_k(\boldsymbol{p}(t)) = T_k(\boldsymbol{p}(t))\rho(\boldsymbol{p}(t)), \\ & \text{where } T_k(\boldsymbol{p}(t)) = \exp\left(-\int_0^t \rho(\boldsymbol{p}(u))du\right). \end{aligned} \tag{2}$$

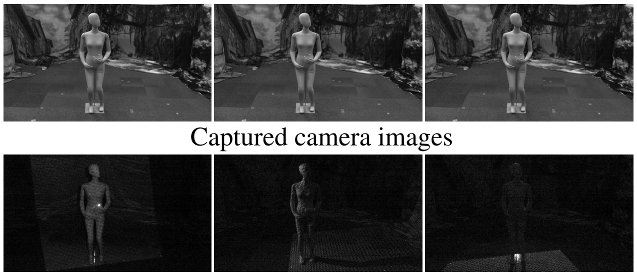
The functions $f(\boldsymbol{p})$ and $c(\boldsymbol{p}(t), \boldsymbol{v}_{kx})$ are represented by MLPs, and their parameters are estimated by minimizing the below loss between the captured images and the images generated by volume rendering and captured images:

$$\mathcal{L}_C = \sum_k \sum_x \parallel C(\boldsymbol{o}_k, \boldsymbol{v}_{kx}) - \hat{C}_{kx} \parallel, \tag{3}$$

where $\hat{C}_{kx}$ is the intensity at the pixel x of k-th captured image. The loss is calculated across all channels of the camera image, including those illuminated by different light sources.

3.3. Constraint on the Irradiance of Projected Lights

In this section, we assume that the input images are multichannel images captured under different illuminations, and we consider one of the channels that is captured only with the j-th light source. If the lighting condition is known, a


Captured camera images

Demultiplexed images

Figure 4. The images in the first row are the captured camera images, and those in the second row are demultiplexed images, which are illuminated by one of the light source.

new constraint can be introduced for the irradiance at each 3D point.

The direction $\boldsymbol{v}_j$ from the j-th light position $\boldsymbol{o}_j$ to a 3D point $\boldsymbol{p}(t)$ is

$$\boldsymbol{v}_{j\boldsymbol{p}} = \frac{\boldsymbol{p}(t) - \boldsymbol{o}_j}{\parallel \boldsymbol{p}(t) - \boldsymbol{o}_j \parallel}. \tag{4}$$

If the irradiance of the light source along $\boldsymbol{v}_j$ at a unit distance is I_j, the irradiance $i_j(\boldsymbol{p}(t))$ at the point $\boldsymbol{p}(t)$ is attenuated as

$$i_j(\boldsymbol{p}(t)) = \frac{I_j T_j(\boldsymbol{p}(t))}{\parallel \boldsymbol{p}(t) - \boldsymbol{o}_j \parallel^2}, \tag{5}$$

where $T_j(\boldsymbol{p}(t))$ is the accumulated transmittance at the point $\boldsymbol{p}$ along the ray from the j-th light source.

If the object surface exists at $\boldsymbol{p}(t)$ and the reflectance property is assumed as Lambertian surface, the intensity at $\boldsymbol{p}(t)$ is independent of the surface orientation with respect to the camera as follows:

$$c_j(\boldsymbol{p}(t), \boldsymbol{v}_{j\boldsymbol{p}}) = i_j(\boldsymbol{p}(t)) r(\boldsymbol{p}(t), \boldsymbol{v}_{j\boldsymbol{p}}), \tag{6}$$

where $r(\boldsymbol{p}(t), \boldsymbol{v}_{j\boldsymbol{p}})$ is the reflectance property at $\boldsymbol{p}(t)$.

To calculate the intensity of an image point, Eq. (1) needs to integrate along the view directions of both the camera $\boldsymbol{v}_k$ and the light source $\boldsymbol{v}_j$ to calculate T_k and T_j. Instead, a new constraint can be introduced from Eq. (6) that $c_j(\boldsymbol{p}(t), \boldsymbol{v}_{j\boldsymbol{p}})$ is proportional to the irradiance that depends on the transmittance and the attenuation from the j-th light source, since the reflectance property $r(\boldsymbol{p}(t), \boldsymbol{v}_{j\boldsymbol{p}})$ is independent of the irradiance. The loss to minimize is defined as follows:

$$\mathcal{L}_g = \sum_j \sum_{\boldsymbol{p}} \int_0^{+\infty} \left\| g_j(\boldsymbol{p}(t), \boldsymbol{v}_{j\boldsymbol{p}}) - \mathrm{sg}\left(\frac{\bar{g}_j}{\bar{T}_j} T_j(\boldsymbol{p}(t)) \right) \right\|^2 dt, \tag{7}$$

where $g_j(\boldsymbol{p}(t), \boldsymbol{v}_{j\boldsymbol{p}}) = c_j(\boldsymbol{p}(t), \boldsymbol{v}_{j\boldsymbol{p}}) \parallel \boldsymbol{p}(t) - \boldsymbol{o}_j \parallel^2$, and $\bar{g}_j$ and $\bar{T}_j$ are the averages of $g_j(\boldsymbol{p}(t), \boldsymbol{v}_{j\boldsymbol{p}})$ and $T_j(\boldsymbol{p}(t))$ along the ray from the viewpoint of the j-th light source, respectively. $\mathrm{sg}(\cdot)$ indicates a stop-gradient operation that blocks gradients from flowing into its argument. For the channel captured under the illumination only with j-th light source, the functions to be estimated are the SDF $f(\boldsymbol{p}_x(t))$ and the intensity $c_j(\boldsymbol{p}(t), \boldsymbol{v}_{j\boldsymbol{p}})$, which are realized as MLPs. Since the transmittance is monotonically decreasing according to the distance from the light source, $g_j(\boldsymbol{p}(t), \boldsymbol{v}_{j\boldsymbol{p}})$ is regularized via Eq. (7) to encourage the same characteristics. Although inter-reflections can violate this assumption when light sources are close to the object, our practical setup places the sources sufficiently far away to render near-field deviations negligible.”

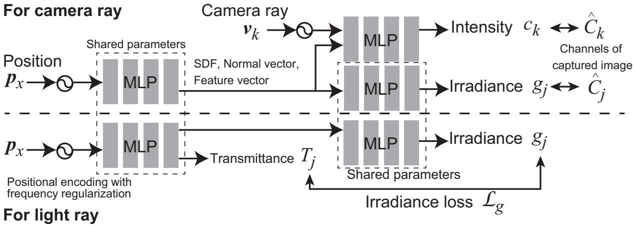


Figure 5. The workflow of the proposed method: the MLP to calculate the SDF is used for both camera and light rays. The view direction is used as one of the input of the MLPs to calculate the intensity only for camera rays.

3.4. Application to Wide-Baseline Views

In volumetric rendering, the high-density area contributes to the calculation of $C(\boldsymbol{o}_k, \boldsymbol{v}_{kx})$ in Eq. (1). If the high-density area is offset from the true surface, $\hat{C}_{kx}$ corresponds to the different positions of the surface from multiple viewpoints. If the baseline between the cameras is narrow as shown in Fig. 3(a), the projected points, P_1 and P_2, are close to each other and their intensities can be similar, which makes the convergence of the neural representation easy. Otherwise as shown in Fig. 3(b), P_1 and P_2 are far from each other, but their colors can be similar. Consequently, the high-density area is consistent with the intensities of the projected points. Since the intensity change between them will not be monotonic, it is difficult to make the high-density area close to the true surface by gradient descent. It indicates that the minimization based on Eq. (3) requires the images captured by the cameras densely located in the scene.

The advantage of the constraint by Eq. (7) is that it can be applied for any relative angle between the rays from a camera and a light source, since it is not necessary to calculate the similarity between two images. Now, we consider the channel illuminated by one of the light sources. In Fig. 3(c), both P_1 and P_2 are high intensity in the camera image but the irradiance is monotonically decreasing along the ray from the light source. If the density and $c_j(P_3, \boldsymbol{v}_{jP_3})$ are high at the point P_3, the intensity calculation by Eq. (1) matches with the intensity of P_2 observed from the camera. However, the light that reaches to the point P_1 is decreased due to the high density at P_3, which is inconsistent with the high intensity at P_1. The irradiance constraint works to resolve this inconsitency by encforcing the monotonicity to $c_j(P_1, \boldsymbol{v}_{jP_1})$ along the light ray.

3.5. Observing Moving Targets with Multiplexed Illumination

The proposed approach assumes that the scene illuminated by each light source is captured from every viewpoint. If a target object is moving during the observation, it is difficult to capture the same situation by the sequential acquisition of the scene under different illumination. To solve this issue, we combine the proposed approach with multiplexed illumination [46], in which multiple light sources illuminate the scene simultaneously. N images are captured to demultiplex them into images in which the scene is illuminated only by each light source, if N light sources illuminate the scene. However, the motion during the acquisition violates the assumption of multiplexing that the scene is static. Since the artifact in demultiplexing is caused by the motion, the method proposed in [43] introduced high-pass filtering to the captured images, and demodulates the signal after removing the low-frequency component. Fig. 4 shows an example of the images of a target that demultiplexed from the input images under external lights. In this example, the scene is illuminated by eight controlled light sources and ceiling lights as an external light. The illumination from each light source is extracted successfully, and they form nine channels of an input image captured from a viewpoint. Since the power of the controlled light sources is very weak in this case, it is hard to recognize the illumination in the camera images. Therefore, the resulting images are demultiplexed from 256 images per frame captured by high-speed cameras to achieve 30 fps while improving the signal-to-noise ratio

3.6. Training

The constraint defined by Eq. (7) enforces that the intensity $c_j(\boldsymbol{p}(t), \boldsymbol{v}_{j\boldsymbol{p}})$ along the ray direction from a camera becomes consistent with the transmittance $T_j(\boldsymbol{p}(t))$ along the ray from the light source. It realizes calculating Eq. (3) and Eq. (7) in the same manner with volume rendering. The network parameters are estimated by minimizing the following loss function:

$$\mathcal{L} = \mathcal{L}_C + \alpha \mathcal{L}_g + \beta \mathcal{L}_{reg}, \tag{8}$$

where $\mathcal{L}_{reg}$ is the Eikonal term to regularize the SDF [14]:

$$\mathcal{L}_{reg} = \sum_k \sum_{\boldsymbol{p}} \int_0^{+\infty} (\| \nabla f(\boldsymbol{p}(t)) \|_2 - 1)^2 dt. \tag{9}$$

α and β are the weight parameters for training. $\mathcal{L}_C$ is calculated for all channels along the rays of the input images. The workflow of the proposed method is shown in Fig. 5. It is based on NeuS and the components for the light ray are added to introduce the constraints. Additionally, since it is reported that starting from generating low-frequency shape is effective in the case of sparse views, the positions and directions are encoded by sinusoidal functions with frequency regularization proposed in FreeNeRF [68].

$\mathcal{L}_g$ is calculated only for the channel along the ray from the light source that is the sole illuminator of the channel. Since the lights are assumed as point light sources in this paper, they can be geometrically regarded as perspective cameras. Therefore, $\mathcal{L}_g$ is calculated by using the accumulated transmittance during volume rendering from the viewpoint of the light sources. As described in Section 3.4, it is expected to work if the high-density area is offset from the surface, which will happen at the beginning of the iteration. Since the precise correspondence should be found by minimizing $\mathcal{L}_C$, the weight α is gradually reduced during the iteration.

4. Experiments

4.1. Simulation

First, the proposed method is evaluated by using the images generated by simulation. To show the advantage of the proposed method, the situations that are difficult to find correspondences are tackled. Namely, the cameras are sparsely located and the baselines between them are very wide. Fig. 6 shows some of the input images. In this simulation, six cameras and six calibrated light sources are located around an object. Since the objects are illuminated by external light in addition, each input image has seven channels as shown in Fig. 2. The 3D models provided in the Blender dataset [32] are used as the samples. Since the shadow gives information about the irradiance, we chose two objects and added a planar surface under the objects in this experiment.

Fig. 7 shows the results by the tested methods. The ground truths (GT) of the objects are shown in (a). As the methods to compare, (b) shows the results of NeuS [60] and (c) shows the results of NeuS with frequency regularization, which is added to the NeuS code by regularizing the positional encoding based on the implementation of FreeNeRF [68]. The key difference between this method and the proposed approach is whether it utilizes the channels of the controlled lights or not. From the results, it is confirmed that the reconstructed shapes are over-smoothed.

We also compared with NeRF-based methods, such as (d) DS [13], (e) NeRS [73], and (f) PS-NeRF [70] using the back ground masks, which are not required for our method. Since NeRF is not designed to reconstruct object shapes, the quality of the reconstructed shapes differs significantly from the GT.

Next, we compared with 3DGS-based methods, such as (g) InstantSplat [9], (h) DUSt3R [62] + 2DGS [17], which generates mesh surfaces, with the point cloud as the initial guess, and (i) SparseGS [66]. The meshes in (g) and (i) are extracted based on the implementation of 2DGS, which first renders depth maps by Gaussian splatting and then integrate them by utilizing truncated signed distance fusion (TSDF) using Open3D [76]. Since 3DGS cannot handle shading effects of multiple light sources correctly, reconstructed shapes are mostly wrong.

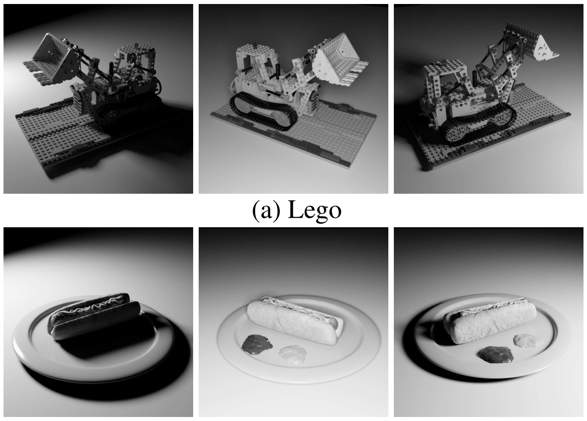
Finally, we compared with MVPS, such as (j) IRON [74] and (k) SuperNormal [67] as well as two structured light (SL) based methods, such as (l) ActiveNeus [18] and (m) TurboSL [33]. Since the number of light sources is limited, shapes could not be reconstructed correctly.

Results of our method are shown in (n). While the number of viewpoints is limited, the shapes of objects are successfully reconstructed. Although no background mask is used in the proposed method, the planar surface is successfully reconstructed as the background object.

The accuracy of the shapes is compared to the ground truth of the simulation models by Chamfer distance. Since some of the compared methods generate floaters due to wrong correspondences or other reasons, they are manually pruned in the figure and not included for calculating the error metrics. Table 1 shows the results of numerical comparison. Although the Neural-based methods, (b), (c), (d), (e) and (f) extract smooth point clouds, the shapes are far from the ground truth. The errors are much larger than that of the proposed method. The point clouds extracted by 3DGS-based methods, (g) and (i), are close to the ground truth. However, the mesh surfaces are not smooth. Althogh the result in (h) is smoother than (g) and (i), the error is worse than them by Chamfer distance. MVPS methods, (j) and (k), extracted noisy point clouds and SL based method (l) and (m) could not recover detail of the objects. On the other hand, by successfully establishing correspondences in a wide-baseline setup, the proposed method achieves better results than other approaches according to the Chamfer distance metric, indicating that the information from multiple light sources is effectively utilized.

4.2. Ablation Study

As the ablation study of the proposed method, we test the components of the proposed method concerning two aspects: the contribution of the irradiance loss $\mathcal{L}_g$ introduced in the proposed method and the contribution of the images captured under the controlled illumination. Fig. 8 shows the results under different conditions. (a) shows the result without using the irradiance loss $\mathcal{L}_g$. The results of reducing the number of cameras are shown in (b) and (c). Additionally, we tested reducing the number of lights. The numerical comparison between the different conditions is shown in Table 2. In the case of not using the constraint of $\mathcal{L}_g$, the distances increase because some artifacts remain in the shape, while most of the part is similar to the case with using $\mathcal{L}_g$. The reason can be considered that it is difficult to converge to the correct correspondence since the initial guess of the density is far from the ground truth. If the number of cameras and lights is reduced, the accuracy is degraded, but the shapes are still better than those without the information of active lighting shown in Table 1.



(a) Lego

(b) Hotdog

Figure 6. Example input images for the simulation: the objects are observed from six viewpoints, each illuminated by a calibrated light source colocated with the corresponding camera. Consequently, in this simulation, six images are combined into seven channels for each viewpoint.

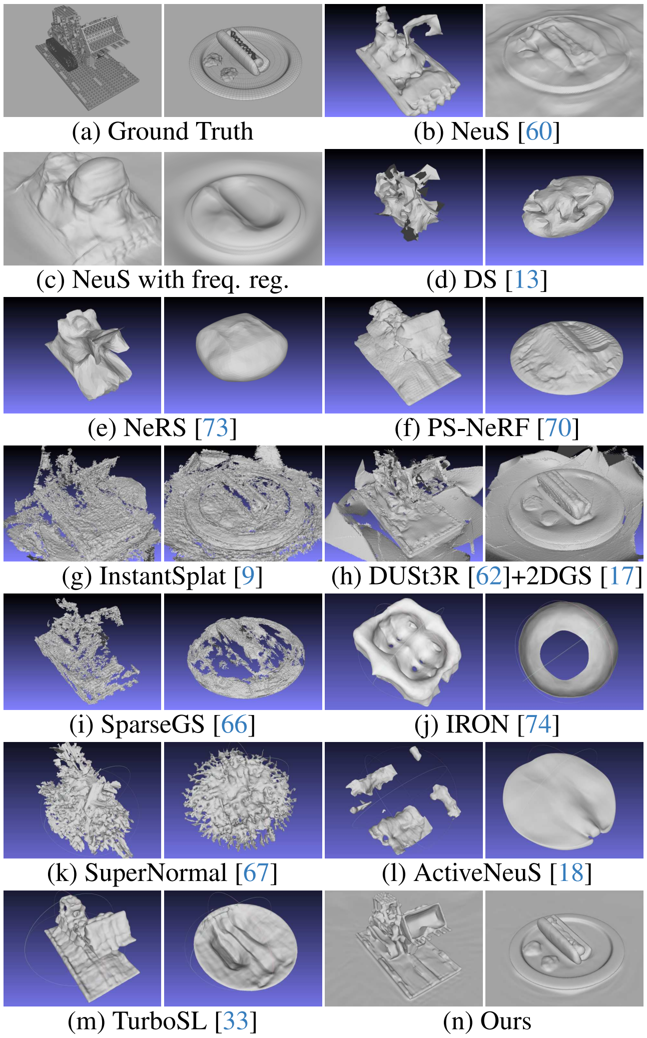
4.3. Real Experiments

Next, we tested the proposed method by capturing images of real objects. The capturing system, as shown in Fig. 1, consists of eight cameras and eight light sources positioned around a target object. Since the objects are further illuminated by room lights, the input images comprise a total of nine channels. The scene is surrounded by a background curtain with a diameter of about five meters. Examples of input images from the real experiments are shown in Fig. 4 and Fig. 9(the third row). Laser light sources equipped with diffractive optical elements (DOE) are used in the experiment to emit a fixed pattern. Although the pattern information is not utilized in the proposed method, it may aid in finding correspondences during optimization. Due to the weaker power of the laser light sources compared to room lights, the channels of the laser lights exhibit some noise. In this experiment, each light source is positioned next to a camera, and therefore the light source positions are assumed to coincide with those of the cameras.

Fig. 9 illustrates the result of 3D reconstruction from the captured images, along with comparisons to existing methods, IRON [74], SuperNormal [67], ActiveNeuS [18] and TurboSL [33]. The area not visible from the cameras is removed from the result in this figure. The shape of the mannequin (1.7m in height) and the plaster object (0.6m in height), both placed at the center of the scene, are successfully reconstructed despite their relatively small sizes compared to the entire scene (5m in diameter) with our tech-

Table 1. Chamfer distances calculated for the results reconstructed by the compared methods.

	NeuS	Freq. Reg.	DS	NeRS	PS-NeRF	InstantSplat	DUSt3R+2DGS	SparseGS	IRON	SuperNormal	ActiveNeuS	TurboSL	Ours
Lego	0.299	0.415	0.385	0.262	0.090	0.061	0.087	0.092	0.299	0.185	0.496	0.081	**0.023**
Hotdog	0.340	0.072	0.517	0.419	0.154	0.055	0.106	0.115	0.800	0.113	0.195	0.312	**0.016**



(a) Ground Truth (b) NeuS [60]

(c) NeuS with freq. reg. (d) DS [13]

(e) NeRS [73] (f) PS-NeRF [70]

(g) InstantSplat [9] (h) DUSt3R [62]+2DGS [17]

(i) SparseGS [66] (j) IRON [74]

(k) SuperNormal [67] (l) ActiveNeuS [18]

(m) TurboSL [33] (n) Ours

Figure 7. Results of 3D reconstruction using the compared methods and our proposed method.

Table 2. Chamfer distances calculated for the results under the different conditions as the ablation study. N/A indicates that a comparable shape is not generated in the setting.

	w/o irradiance loss $\mathcal{L}_g$	4 cams, 6 lights	4 cams, 4 lights	3 cams, 6 lights	3 cams, 3 lights	2 cams, 6 lights
Lego	0.02397	0.03052	0.03343	0.10623	0.13774	N/A
Hotdog	0.01640	0.0343	0.03444	N/A	N/A	0.02755

Table 3. Chamfer distances between the reference and measured shapes of the plaster object shown in Fig. 9(bottom).

Ours	IRON [74]	SuperNormal [67]	ActiveNeuS [18]	TurboSL [33]
3.27mm	98.3mm	87.6mm	10.3mm	14.8mm

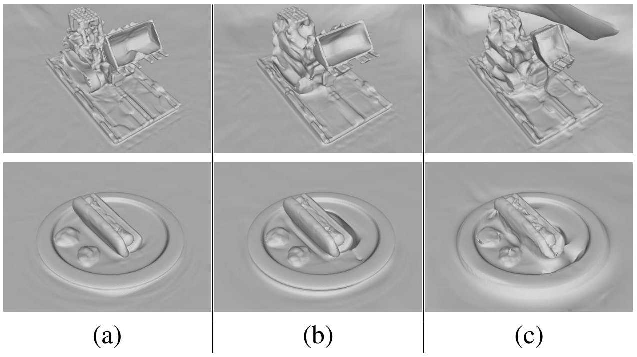


Figure 8. The results are reconstructed under the following conditions: (a) without using $\mathcal{L}_g$ with 6 cameras and 6 lights, (b) using 4 cameras and 6 lights, (c) 3 cameras and 6 lights for Lego, 2 cameras and 6 lights for Hotdog.

nique. As shown in Fig. 9, MVPS-based methods, such as IRON [74] and SuperNormal [67], failed to recover meaningful shapes, primarily due to the limited number of viewpoints and light sources. By contrast, structured light–based methods, such as ActiveNeuS [18] and TurboSL [33], were able to reconstruct comparatively better shapes; however, high-frequency details were lost due to ambiguous correspondences and calibration errors, thereby highlighting the strength of our method.

Quantitative evaluations are conducted using Chamfer distance errors between the reference shapes and the reconstructed results, as shown in Table 3. The reference shapes were obtained with a handheld 3D scanner as ground truth. From the table, it is evident that our method outperforms previous approaches.

Finally, we apply the proposed method to a dynamic scene. Fig. 10 depicts a situation where a person is standing up. The images for each illumination are demultiplexed for each frame. The experimental setup is the same as for the mannequin case. We reconstructed the shapes for 90 frames, sampling 512 rays for each iteration, and conducting 210K iterations. The reconstruction time for each frame is approximately 12 hours using an NVIDIA V100 GPU. The entire shape of the person and the background is successfully reconstructed using eight cameras, as shown in Fig. 11. In addition to the results of standing-up and jumping motion shown in Fig. 1, Fig. 12 shows the results for two motions: walking and sitting down. Please refer to the supplemental video for the 4D scene reconstruction results. One limitation of the approach is the reconstruction of occluded areas, which can occur more easily compared to methods using a larger number of cameras. For example, in the case of sitting down, the shape of the seat surface is not accurately reconstructed due to occlusion by the person.

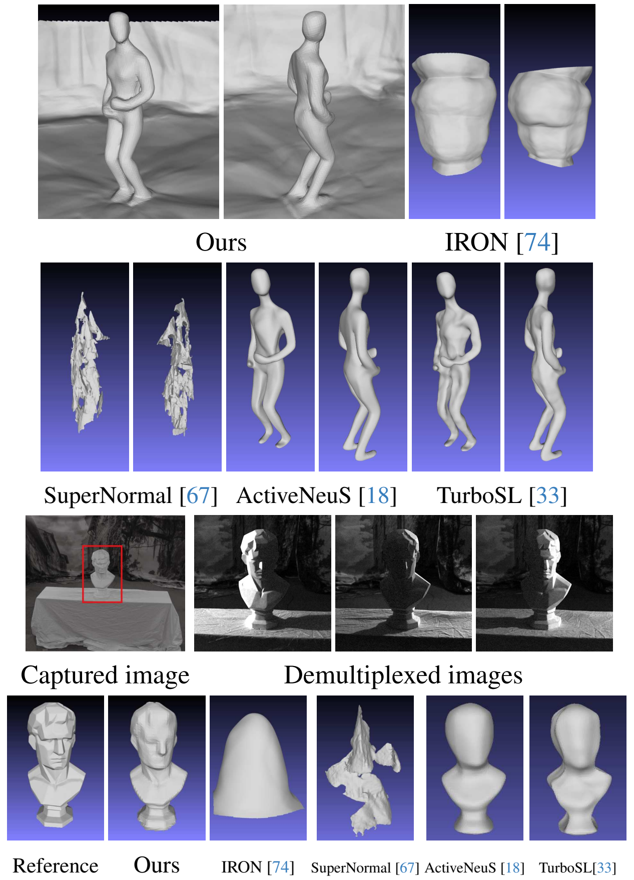


Figure 9. The results of the 3D reconstruction using the proposed method are shown. The 1st and 2nd rows: a 1.7m-tall mannequin is reconstructed using eight cameras. The 3rd row: Captured and demultiplexed images of a plaster object (0.6m in height) placed at the center of the scene. Thr 4th row: A reference and reconstructed shapes of a plaster object. The reference shape is captured by a handheld 3D scanner as the ground truth for comparison.

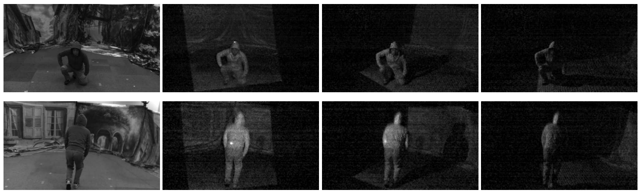


Figure 10. The two frames from the demultiplexed images of a person in motion.

Improving robustness against occlusion is our important future work.

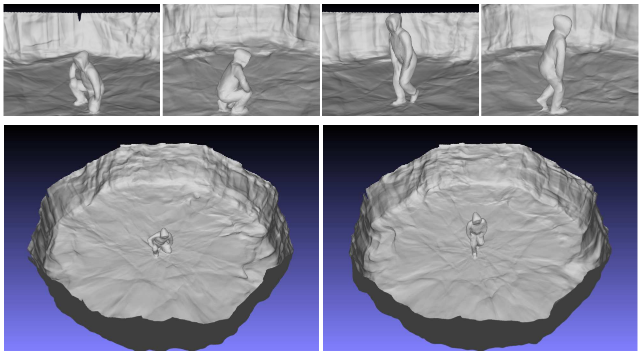


Figure 11. The two frames from the results of a person in motion.

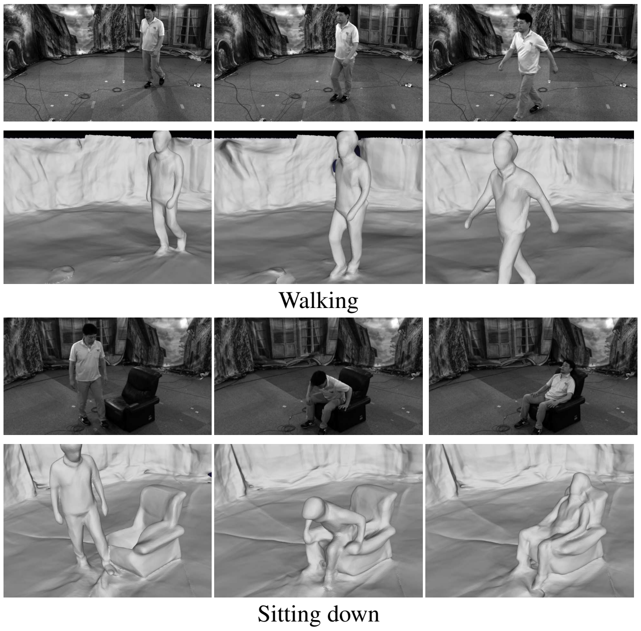


Figure 12. The results of 3D reconstruction of additional two motions. Please refer to the supplemental video for the 4D scene reconstruction results.

5. Conclusion

In this paper, we have proposed a method for 3D reconstruction using multiview stereo based on neural representation. One of the challenges in reconstructing shapes is the requirement to capture images from densely distributed viewpoints, which becomes particularly problematic when reconstructing moving objects due to the high cost associated with preparing many cameras to capture them simultaneously. To address this issue, we have introduced a new approach that combines active lighting to acquire additional information for estimating the 3D shape of a scene. This approach leverages constraints derived from multichannel images and the relationship between image intensity and irradiance from the light source. Our experiments demonstrated that the information provided by active lighting was effective in finding correspondences, resulting in successful shape reconstruction. Additionally, by combining this approach with multiplexed illumination techniques, we were able to achieve 3D reconstruction of moving objects. In future work, we plan to introduce constraints between frames, as the current approach reconstructs frames individually.

Acknowledgment

This work was supported by JST PARKS2025_04, and JSPS/KAKENHI JP25K22821 and 22H00545 in Japan.

References

[1] Jonathan T Barron, Ben Mildenhall, Matthew Tancik, Peter Hedman, Ricardo Martin-Brualla, and Pratul P Srinivasan. Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5855–5864, 2021. 2

[2] Jonathan T Barron, Ben Mildenhall, Dor Verbin, Pratul P Srinivasan, and Peter Hedman. Mip-nerf 360: Unbounded anti-aliased neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5470–5479, 2022. 2

[3] Baptiste Brument, Robin Bruneau, Yvain Quéau, Jean Mélou, François Lauze, Jean-Denis Durou, and Lilian Calvet. Rnb-neus: Reflectance and normal-based multi-view 3d reconstruction. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 2

[4] Llukman Cerkezi and Paolo Favaro. Sparse 3d reconstruction via object-centric ray sampling. *arXiv preprint arXiv:2309.03008*, 2023. 2

[5] Anpei Chen, Zexiang Xu, Fuqiang Zhao, Xiaoshuai Zhang, Fanbo Xiang, Jingyi Yu, and Hao Su. Mvsnerf: Fast generalizable radiance field reconstruction from multi-view stereo. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14124–14133, 2021. 2

[6] Julian Chibane, Aayush Bansal, Verica Lazova, and Gerard Pons-Moll. Stereo radiance fields (srf): Learning view synthesis from sparse views of novel scenes. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2021. 2

[7] Jaeyoung Chung, Jeongtaek Oh, and Kyoung Mu Lee. Depth-regularized optimization for 3d gaussian splatting in few-shot images. *arXiv preprint arXiv:2311.13398*, 2023. 2

[8] Kangle Deng, Andrew Liu, Jun-Yan Zhu, and Deva Ramanan. Depth-supervised nerf: Fewer views and faster training for free. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12882–12891, 2022. 2

[9] Zhiwen Fan, Wenyan Cong, Kairun Wen, Kevin Wang, Jian Zhang, Xinghao Ding, Danfei Xu, Boris Ivanovic, Marco Pavone, Georgios Pavlakos, Zhangyang Wang, and Yue Wang. Instantsplat: Unbounded sparse-view pose-free gaussian splatting in 40 seconds, 2024. 2, 5, 7

[10] Ryo Furukawa, Ryusuke Sagawa, Shiro Oka, and Hiroshi Kawasaki. Nerf-based multi-frame 3d integration for 3d endoscopy using active stereo. In *2024 46nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. IEEE, 2024. 2

[11] Yasutaka Furukawa, Carlos Hernández, et al. Multi-view stereo: A tutorial. *Foundations and Trends® in Computer Graphics and Vision*, 9(1-2):1–148, 2015. 1

[12] Clément Godard, Oisin Mac Aodha, and Gabriel J. Brostow. Unsupervised monocular depth estimation with left-right consistency. In *CVPR*, 2017. 2

[13] Shubham Goel, Georgia Gkioxari, and Jitendra Malik. Differentiable stereopsis: Meshes from multiple views using differentiable rendering. *arXiv preprint arXiv:2110.05472*, 2021. 2, 5, 7

[14] Amos Gropp, Lior Yariv, Niv Haim, Matan Atzmon, and Yaron Lipman. Implicit geometric regularization for learning shapes. In *Proceedings of Machine Learning and Systems 2020*, pages 3569–3579. 2020. 5

[15] Richard Hartley and Andrew Zisserman. *Multiple view geometry in computer vision*. Cambridge University Press, Cambridge, UK; New York, 2003. 1

[16] Carlos Hernandez, George Vogiatzis, and Roberto Cipolla. Multiview photometric stereo. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(3):548–554, 2008. 2

[17] Binbin Huang, Zehao Yu, Anpei Chen, Andreas Geiger, and Shenghua Gao. 2d gaussian splatting for geometrically accurate radiance fields. In *SIGGRAPH 2024 Conference Papers*. Association for Computing Machinery, 2024. 1, 5, 7

[18] Kazuto Ichimaru, Takaki Ikeda, Diego Thomas, Takafumi Iwaguchi, and Hiroshi Kawasaki. Activeneus: Neural signed distance fields for active stereo. In *2024 International Conference on 3D Vision (3DV)*, pages 539–548. IEEE, 2024. 2, 6, 7, 8

[19] Katsushi Ikeuchi, Yasuyuki Matsushita, Ryusuke Sagawa, Hiroshi Kawasaki, Yasuhiro Mukaigawa, Ryo Furukawa, and Daisuke Miyazaki. *Active lighting and its application for computer vision*. Springer, 2020. 1, 2

[20] Ajay Jain, Matthew Tancik, and Pieter Abbeel. Putting nerf on a diet: Semantically consistent few-shot view synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5885–5894, 2021. 2

[21] Lihan Jiang, Yucheng Mao, Linning Xu, Tao Lu, Kerui Ren, Yichen Jin, Xudong Xu, Mulin Yu, Jiangmiao Pang, Feng Zhao, et al. Anysplat: Feed-forward 3d gaussian splatting from unconstrained views. *ACM Transactions on Graphics (TOG)*, 44(6):1–16, 2025. 2

[22] Hanbyul Joo, Hao Liu, Lei Tan, Lin Gui, Bart Nabbe, Iain Matthews, Takeo Kanade, Shohei Nobuhara, and Yaser Sheikh. Panoptic studio: A massively multiview system for social motion capture. In *The IEEE International Conference on Computer Vision (ICCV)*, 2015. 1

[23] Berk Kaya, Suryansh Kumar, Francesco Sarno, Vittorio Ferrari, and Luc Van Gool. Neural radiance fields approach to deep multi-view photometric stereo. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1965–1977, 2022. 2

[24] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42 (4), 2023. 1, 2

[25] Mijeong Kim, Seonguk Seo, and Bohyung Han. Infonerf: Ray entropy minimization for few-shot neural volume rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12912–12921, 2022. 2

[26] Vincent Leroy, Jean-Sebastien Franco, and Edmond Boyer. Volume sweeping: Learning photoconsistency for multi-view shape reconstruction. *International Journal of Computer Vision*, 129:1–16, 2021. 1

[27] Chunyu Li, Taisuke Hashimoto, Eiichi Matsumoto, and Hiroharu Kato. Multi-view neural surface reconstruction with structured light. *arXiv preprint arXiv:2211.11971*, 2022. 2

[28] Min Li, Zhenglong Zhou, Zhe Wu, Boxin Shi, Changyu Diao, and Ping Tan. Multi-view photometric stereo: A robust solution and benchmark dataset for spatially varying isotropic materials. *IEEE Transactions on Image Processing*, 29:4159–4173, 2020. 2

[29] Chen-Hsuan Lin, Wei-Chiu Ma, Antonio Torralba, and Simon Lucey. Barf: Bundle-adjusting neural radiance fields. In *IEEE International Conference on Computer Vision (ICCV)*, 2021. 2

[30] Yebin Liu, Qionghai Dai, and Wenli Xu. A point-cloud-based multiview stereo algorithm for free-viewpoint video. *IEEE transactions on visualization and computer graphics*, 16(3):407–418, 2009. 1

[31] Yuan Liu, Peng Wang, Cheng Lin, Xiaoxiao Long, Jiepeng Wang, Lingjie Liu, Taku Komura, and Wenping Wang. Nero: Neural geometry and brdf reconstruction of reflective objects from multiview images. 2023. 2

[32] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *ECCV*, 2020. 1, 2, 5

[33] Parsa Mirdehghan, Maxx Wu, Wenzheng Chen, David B. Lindell, and Kiriakos N. Kutulakos. Turbosl: Dense accurate and fast 3d by neural inverse structured light. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 25067–25076, 2024. 6, 7, 8

[34] Y. Mukaigawa, K.Sumino, and Y.Yagi. Multiplexed illumination for measuring brdf using an ellipsoidal mirror and a projector. In *Proc. of Asian Conference on Computer Vision*, pages 246–257, 2007. 3

[35] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM Trans. Graph.*, 41(4):102:1–102:15, 2022. 2

[36] Michael Niemeyer, Jonathan T. Barron, Ben Mildenhall, Mehdi S. M. Sajjadi, Andreas Geiger, and Noha Radwan. Regnerf: Regularizing neural radiance fields for view synthesis from sparse inputs. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2

[37] Jaesik Park, Sudipta N Sinha, Yasuyuki Matsushita, Yu-Wing Tai, and In So Kweon. Robust multiview photometric stereo using planar mesh parameterization. *IEEE transactions on pattern analysis and machine intelligence*, 39(8): 1591–1604, 2016. 2

[38] Keunhong Park, Utkarsh Sinha, Peter Hedman, Jonathan T. Barron, Sofien Bouaziz, Dan B Goldman, Ricardo Martin-Brualla, and Steven M. Seitz. Hypernerf: A higher-dimensional representation for topologically varying neural radiance fields. *ACM Trans. Graph.*, 40(6), 2021. 2

[39] Albert Pumarola, Enric Corona, Gerard Pons-Moll, and Francesc Moreno-Noguer. D-NeRF: Neural Radiance Fields for Dynamic Scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020. 2

[40] Rukun Qiao, Hiroshi Kawasaki, and Hongbin Zha. Depth reconstruction with neural signed distance fields in structured light systems. In *2024 International Conference on 3D Vision (3DV)*, pages 770–779. IEEE, 2024. 2

[41] Konstantinos Rematas, Ricardo Martin-Brualla, and Vittorio Ferrari. Sharf: Shape-conditioned radiance fields from a single view. In *ICML*, 2021. 2

[42] Barbara Roessle, Jonathan T Barron, Ben Mildenhall, Pratul P Srinivasan, and Matthias Nießner. Dense depth priors for neural radiance fields from sparse input views. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12892–12901, 2022. 2

[43] Ryusuke Sagawa and Yutaka Satoh. Illuminant-camera communication to observe moving objects under strong external light by spread spectrum modulation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5097–5105, 2017. 3, 5

[44] Ryusuke Sagawa, Yuichi Ota, Yasushi Yagi, Ryo Furukawa, Naoki Asada, and Hiroshi Kawasaki. Dense 3d reconstruction method using a single pattern for fast moving object. In *ICCV*, pages 1779–1786, 2009. 1

[45] Ryusuke Sagawa, Kazuhiro Sakashita, Nozomu Kasuya, Hiroshi Kawasaki, Ryo Furukawa, and Yasushi Yagi. Grid-based active stereo with single-colored wave pattern for dense one-shot 3D scan. In *3DIMPVT*, pages 363–370, 2012. 1

[46] Y.Y. Schechner, S.K. Nayar, and P.N. Belhumeur. A theory of multiplexed illumination. In *IEEE International Conference on Computer Vision*, pages 808–815, 2003. 2, 3, 5

[47] Johannes L Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4104–4113, 2016. 1, 2

[48] Thomas Schops, Johannes L Schonberger, Silvano Galliani, Torsten Sattler, Konrad Schindler, Marc Pollefeys, and Andreas Geiger. A multi-view stereo benchmark with high-resolution images and multi-camera videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3260–3269, 2017. 1

[49] Steven M Seitz, Brian Curless, James Diebel, Daniel Scharstein, and Richard Szeliski. A comparison and evaluation of multi-view stereo reconstruction algorithms. In *2006 IEEE computer society conference on computer vision and pattern recognition (CVPR'06)*, pages 519–528. IEEE, 2006. 1

[50] Aarrushi Shandilya, Benjamin Attal, Christian Richardt, James Tompkin, and Matthew O'Toole. Neural fields for structured lighting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3512–3522, 2023. 2

[51] Chonghyuk Song, Gengshan Yang, Kangle Deng, Jun-Yan Zhu, and Deva Ramanan. Total-recon: Deformable scene reconstruction for embodied view synthesis. In *IEEE International Conference on Computer Vision (ICCV)*, 2023. 2

[52] Soohwan Song, Daekyum Kim, and Sungho Jo. Active 3d modeling via online multi-view stereo. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 5284–5291. IEEE, 2020. 1

[53] Chen Sun, Abhinav Shrivastava, Saurabh Singh, and Abhinav Gupta. Revisiting unreasonable effectiveness of data in deep learning era. In *Proceedings of the IEEE international conference on computer vision*, pages 843–852, 2017. 2

[54] Zhenggang Tang, Yuchen Fan, Dilin Wang, Hongyu Xu, Rakesh Ranjan, Alexander Schwing, and Zhicheng Yan. Mv-dust3r+: Single-stage scene reconstruction from sparse views in 2 seconds. *arXiv preprint arXiv:2412.06974*, 2024. 2

[55] Tony Tung, Shohei Nobuhara, and Takashi Matsuyama. Complete multi-view reconstruction of dynamic scenes from probabilistic fusion of narrow and wide baseline stereo. In *2009 IEEE 12th International Conference on Computer Vision*, pages 1709–1716, 2009. 1

[56] Matias Turkulainen, Xuqian Ren, Iaroslav Melekhov, Otto Seiskari, Esa Rahtu, and Juho Kannala. Dn-splatter: Depth and normal priors for gaussian splatting and meshing, 2024. 2

[57] Dor Verbin, Peter Hedman, Ben Mildenhall, Todd Zickler, Jonathan T. Barron, and Pratul P. Srinivasan. Ref-NeRF: Structured view-dependent appearance for neural radiance fields. *CVPR*, 2022. 2

[58] Daniel Vlasic, Pieter Peers, Ilya Baran, Paul Debevec, Jovan Popović, Szymon Rusinkiewicz, and Wojciech Matusik. Dynamic shape capture using multi-view photometric stereo. In *ACM SIGGRAPH Asia 2009 papers*, pages 1–11. 2009. 1

[59] Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025. 2

[60] Peng Wang, Lingjie Liu, Yuan Liu, Christian Theobalt, Taku Komura, and Wenping Wang. Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. *arXiv preprint arXiv:2106.10689*, 2021. 1, 2, 3, 5, 7

[61] Qianqian Wang, Zhicheng Wang, Kyle Genova, Pratul Srinivasan, Howard Zhou, Jonathan T. Barron, Ricardo Martin-Brualla, Noah Snavely, and Thomas Funkhouser. Ibrnet: Learning multi-view image-based rendering. In *CVPR*, 2021. 2

[62] Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. Dust3r: Geometric 3d vision made easy. In *CVPR*, 2024. 2, 5, 7

[63] Zirui Wang, Shangzhe Wu, Weidi Xie, Min Chen, and Victor Adrian Prisacariu. NeRF−−: Neural radiance fields without known camera parameters. *arXiv preprint arXiv:2102.07064*, 2021. 2

[64] Yi Wei, Shaohui Liu, Yongming Rao, Wang Zhao, Jiwen Lu, and Jie Zhou. Nerfingmvs: Guided optimization of neural radiance fields for indoor multi-view stereo. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5610–5619, 2021. 2

[65] A. Wenger, A. Gardner, C. Tchou, J. Unger, T. Hawkins, and P. Debevec. Performance relighting and reflectance transformation with time-multiplexed illumination. In *SIGGRAPH*, 2005. 3

[66] Haolin Xiong, Sairisheek Muttukuru, Rishi Upadhyay, Pradyumna Chari, and Achuta Kadambi. Sparsegs: Realtime 360° sparse view synthesis using gaussian splatting. *Arxiv*, 2023. 2, 6, 7

[67] Cao Xu and Taketomi Takafumi. Supernormal: Neural surface reconstruction via multi-view normal integration. In *CVPR*, 2024. 2, 6, 7, 8

[68] Jiawei Yang, Marco Pavone, and Yue Wang. Freenerf: Improving few-shot neural rendering with free frequency regularization. 2023. 2, 5

[69] Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. In *CVPR*, 2024. 2

[70] Wenqi Yang, Guanying Chen, Chaofeng Chen, Zhenfang Chen, and Kwan-Yee K. Wong. Ps-nerf: Neural inverse rendering for multi-view photometric stereo. In *European Conference on Computer Vision (ECCV)*, 2022. 2, 5, 7

[71] Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. pixelnerf: Neural radiance fields from one or few images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4578–4587, 2021. 2

[72] Yu-Jie Yuan, Yu-Kun Lai, Yi-Hua Huang, Leif Kobbelt, and Lin Gao. Neural radiance fields from sparse rgb-d images for high-quality view synthesis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. 2

[73] Jason Zhang, Gengshan Yang, Shubham Tulsiani, and Deva Ramanan. Ners: Neural reflectance surfaces for sparse-view 3d reconstruction in the wild. *Advances in Neural Information Processing Systems*, 34:29835–29847, 2021. 2, 5, 7

[74] Kai Zhang, Fujun Luan, Zhengqi Li, and Noah Snavely. Iron: Inverse rendering by optimizing neural sdfs and materials from photometric images. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2022. 6, 7, 8

[75] Dongxu Zhao, Daniel Lichy, Pierre-Nicolas Perrin, Jan-Michael Frahm, and Soumyadip Sengupta. Mvpsnet: Fast generalizable multi-view photometric stereo. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12525–12536, 2023. 2

[76] Qian-Yi Zhou, Jaesik Park, and Vladlen Koltun. Open3D: A modern library for 3D data processing. *arXiv:1801.09847*, 2018. 6