Learning Neural Scene Representation from iToF Imaging

Wenjie Chang[1], Hanzhi Chang[1], Yueyi Zhang[3], Wenfei Yang[1,2,†], Tianzhu Zhang[1,2]
[1]University of Science and Technology of China
[2]National Key Laboratoray of Deep Space Exploration, Deep Space Exploration Laboratory
[3]Miromind
changwj@mail.ustc.edu.cn

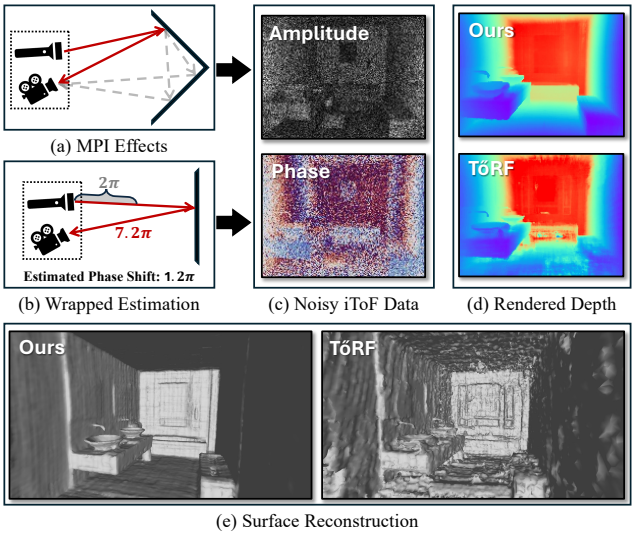
Abstract

Indirect Time-of-Flight (iToF) cameras are popular for 3D perception because they are cost-effective and easy to deploy. They emit modulated infrared signals to illuminate the scene and process the received signals to generate amplitude and phase images. The depth is calculated from the phase using the modulation frequency. However, the obtained depth often suffers from noise caused by multi-path interference, low signal-to-noise ratio (SNR), and depth wrapping. Building on recent advancements in neural scene representations, which have shown great potential in 3D modeling from multi-view RGB images, we propose leveraging this approach to reconstruct 3D representations from noisy iToF data. Our method utilizes the multi-view consistency of amplitude and phase maps, fusing information from all input views to generate an accurate scene representation. Considering the impact of infrared illumination, we propose a new rendering scheme for amplitude maps based on signed distance function (SDF) and introduce a neural lighting function to model the appearance variations caused by active illumination. We also incorporate a phase-guided sampling strategy and a wrapping-aware phase-to-depth loss to utilize raw phase information and mitigate depth wrapping. Additionally, we add a noise-weight loss to prevent excessive smoothing information across noisy multi-view measurements. Experiments conducted on synthetic and real-world datasets demonstrate that the proposed method outperforms state-of-the-art techniques.

1. Introduction

Time-of-Flight (ToF) imaging has emerged as a foundational technology in depth sensing, widely adopted across diverse fields such as autonomous driving [2, 28, 35], augmented reality [12, 25], and robotics [19, 36], owing to its capability to capture depth information by measuring the time delay between emitted and received signals. ToF imaging can be broadly classified into direct-ToF (dToF) and indirect-ToF (iToF) systems. While dToF systems offer precise depth measurements using components like single photon avalanche diodes (SPADs), they are often constrained by high hardware costs and low spatial resolution [37]. In contrast, iToF imaging emits amplitude-modulated infrared signals and calculates the phase and amplitude information, as shown in Fig. 1(c). The phase information denotes the phase difference between the emitted signals and returned signals [16], which is then transformed into depth measurements. This approach offers a more cost-effective solution and provides higher spatial resolution, making it particularly attractive for consumer applications [4].

Despite these advantages, iToF imaging faces significant challenges, such as multi-path interference, low signal-to-noise ratio (SNR), and depth wrapping [27]. As shown in



(a) MPI Effects (b) Wrapped Estimation (c) Noisy iToF Data (d) Rendered Depth (e) Surface Reconstruction

Figure 1. iToF cameras are affected by various interferences (a)-(b), resulting in noisy amplitude and phase information (c). We aim to utilize neural scene representation techniques to obtain accurate depth information (d) from multi-view iToF imaging data and achieve promising surface reconstruction (e).

[†]Corresponding Author

Fig. 1, multi-path interference leads to depth errors because the light signals captured by each pixel travel along indirect paths, affecting the accuracy of depth measurement [11]. Low SNR arises from operating under low exposure times and power consumption [4], further degrading depth accuracy. Additionally, depth wrapping can occur due to the periodic nature of phase measurements, where the measured phase ambiguously represents depth beyond a specific range, leading to significant inaccuracies, especially in scenes with high-depth variation [13]. Some researchers have been struggling to address these errors in iToF imaging. With the advancement of deep learning, many data-driven approaches [1, 7–10, 17, 20, 29, 32] have been developed, which train on simulated datasets and aim to generalize to real-world data. However, reliance on training data limits these methods' adaptability across different iToF camera systems, which hinders their generalization to scenarios beyond the training conditions.

Recently, Neural Radiance Fields (NeRF) [21] and related techniques have gained significant attention as powerful tools for 3D reconstruction from a set of input RGB images. RawNeRF [22] demonstrates that when optimizing a 3D scene representation using multi-view images with noise, the optimization process can effectively smooth out the noise through multi-view consistency, without relying on additional denoising designs. Building on this foundation, we aim to achieve precise 3D reconstruction from multi-view noisy iToF imaging data by enforcing the consistency of the amplitude and phase components across different views. Unlike data-driven methods that depend on specific training datasets, this approach functions independently of GT supervision, making it suitable for data captured by various iToF cameras at different frequencies.

When directly applying the classic volume rendering scheme to amplitude maps of iToF data, the results are often unsatisfactory due to the varying lighting conditions caused by the infrared light signals emitted by the iToF camera. To address this, we propose a rendering scheme based on Signed Distance Functions (SDF), which simulates the imaging process of iToF cameras and facilitates surface reconstruction. To model the appearance variations caused by lighting, we introduce a neural lighting function that models the reconstructed scene's response to infrared light signals in the incident direction. Moreover, the depth measurements of iToF cameras are derived from phase shift, which constrains their measurement range based on modulation frequency. This leads to phase wrapping when the actual depth exceeds the measurement limits. We propose a wrapping-aware loss that supervises reconstruction using wrapped phase estimation to mitigate the ambiguity. Furthermore, we adopt a phase-guided sampling strategy to leverage the phase information, thereby enhancing performance. To avoid excessive smoothing of noisy data from different viewpoints, we use a noise-weight regularization in the loss function to improve the reconstruction details. By integrating these components, we construct a neural scene representation directly from noisy multi-view iToF amplitude and phase information, optimizing depth estimation and outputting accurate 3D mesh reconstruction.

In summary, the contributions of this work can be summarized as follows:

- We propose a new rendering scheme based on SDF and introduce a neural lighting function to model the appearance variation caused by infrared illumination signals. These innovations enable accurate surface reconstruction from raw iToF imaging data.
- We introduce a Wrapping-Aware loss to address the ambiguity caused by phase wrapping and adapt a sampling strategy to leverage the phase information provided by iToF imaging. These methods are designed to enhance the effectiveness of the training process.
- Extensive experiments validate the effectiveness of the proposed method under both synthetic and real-world scenarios. Compared with TöRF [3], we achieved error reductions of 70.3% in the MAE metric for depth error.

2. Related Work

In this section, we briefly overview methods that are related to learning-based iToF imaging and neural scene representation (NSR) methods under active illumination.

Data-Driven ToF Imaging: Recently, learning-based methods have greatly improved depth estimation from iToF cameras. Some approaches take noisy iToF depth data as input and produce refined depth estimations. Marco *et al*. [17] created a dataset for ToF denoising and proposed a light transport model for multi-path interference effects. Agresti *et al*. [1] developed an adversarial learning strategy to address the domain shift between unlabeled real-world scenes and synthetic training data. Dong *et al*. [7] utilized the spatial structure of scenes by constructing a multi-scale depth residual pyramid. RADU [29] extended 2D ToF denoising to 3D, using 3D point CNNs for depth updating. MTDNet [8] further improved depth estimation by combining multi-frame imaging results. Other approaches enhance depth estimation by leveraging both amplitude and phase information from iToF imaging. Su *et al*. [32] directly took raw measurements as input and output denoised depth maps. Guo *et al*. [9] generated a large-scale ToF dataset and a residual-based U-Net network to remove multi-path interference and shot noise. iToF2dToF [10] interpolated frequency measurements from iToF images to estimate dToF images, offering an alternative output representation. Meng *et al*. [20] established the relationship between measurements taken at different times and phase shifts to mitigate errors caused by 3D motions in iToF imaging. Since acquiring detailed 3D ground truth data from real-world

iToF imaging is challenging and costly, these methods are predominantly trained on simulated datasets, limiting their generalizability to real-world applications.

NSR under Active Illumination: Although NeRF has demonstrated promising results [5, 21, 23, 26, 38], it has limitations when dealing with inputs under varying illumination conditions. These limitations lead to ambiguities in scenes, as the observed radiance at the same spatial location and viewing direction can vary due to variations in lighting intensity [40]. These variations are often introduced by active imaging devices, such as structured light systems and iToF cameras. To address this issue, some studies have extended NeRF to these devices by redesigning the volume rendering scheme and/or incorporating additional imaging data [6, 15, 30, 31, 39]. Structured light cameras project a known pattern onto a scene and use the captured deformation image to calculate the 3D shape of the objects. Li *et al*. [15] reconstructed 3D objects with a neural signed distance field from multi-view structured light pattern images. Shandilya *et al*. [30] adopt a density field to explicitly model raw structured light images to recover scene geometry and additional properties such as surface normals, direct and indirect lighting components. Both approaches use additional ambient images without lighting as supervision. iToF cameras emit modulated infrared light to illuminate the scene and recover depth by measuring the time it takes for the light to travel from the source to the scene and back to the sensor. TöRF [3] combined RGB images with iToF imaging data to model dynamic scenes, while F-TöRF [24] reconstructed fast-moving objects with a fixed camera position. In contrast to these works, we focus on developing an approach for high-quality static 3D reconstruction using multi-view iToF imaging data.

3. Preliminary

In this section, we introduce the fundamental principles and mathematical formulations of the iToF camera, including the transformation of the emitted infrared signal into the received signal, as well as the methods for obtaining phase and amplitude measurements. ToF cameras emit an amplitude-modulated infrared signal, represented as

$$g(t) = g_1 \cos(2\pi f t) + g_0, \tag{1}$$

where g_1 is the modulation amplitude, f is the modulation frequency, and g_0 is the direct current (DC) offset. This signal illuminates the scene, and the light reflects off objects, returning to the camera with a time delay τ_0, which depends on the distance to the object. The received signal at a camera pixel can be expressed as

$$S_{rec}(t) = r_1 \cos(2\pi f t - 2\pi f \tau_0) + r_0, \tag{2}$$

where r_1 and r_0 represent the amplitude and DC component of the signal after reflection from the object and attenuation during propagation through the medium. To measure the phase shift $2\pi f \tau_0$ and amplitude, ToF cameras process the received signal with an internal reference signal, $b\cos(2\pi f t - \phi)$, where ϕ is a programmable phase shift. The multiplication of the received signal $s(t)$ with the reference signal results in

$$i(t) = s(t) \cdot b\cos(2\pi f t - \phi). \tag{3}$$

This signal is then integrated over a long exposure time, filtering out high-frequency components and isolating the phase-related terms. By capturing measurements at multiple reference phase settings (0°, 90°, 180°, and 270°), we obtain the measurements I_ϕ, which can be expressed as:

$$I_\phi = \frac{r_1 b}{2}\cos(\phi - 2\pi f \tau_0) + a_0,\ \phi \in \{0°, 90°, 180°, 270°\}, \tag{4}$$

where a_0 includes the contributions from environmental light and DC offset. These measurements are then used to calculate the phase shift θ and amplitude A:

$$\theta = 2\pi f \tau_0 = \arctan\left(\frac{I_{90} - I_{270}}{I_{180} - I_0}\right), \tag{5}$$

$$A = \sqrt{(I_0 - I_{180})^2 + (I_{270} - I_{90})^2} = r_1 b. \tag{6}$$

The phase shift θ is directly related to the time delay of the signal and to the distance of the object, which is given by $d = c \cdot \tau_0/2$, where c is the speed of light. This equation accounts for the round-trip time of the emitted signal. The amplitude A reflects the combined effects of surface reflectivity and modulation amplitude.

4. Method

4.1. Rendering for iToF Imaging Data

The vanilla radiance integral formula from NeRF [18] is defined as:

$$\hat{C}_m^n = \int_m^n \left(c(s)\sigma(s)e^{-\int_0^s \sigma(t)dt}\right) ds, \tag{7}$$

where $\hat{C}_m^n \in \mathbb{R}^3$ denotes the integrated radiance along the segment between m and n. Here, $c(s) \in \mathbb{R}^3$ indicates the radiance emitted from that point at position s, and $\sigma(s) \in \mathbb{R}$ is the medium absorption coefficient at position s. The point $s = 0$ corresponds to the pixel at the camera. In the vanilla NeRF framework, $c(s)$ is typically assumed to be constant for a given point from the same direction, implying that the radiance emitted by each point is independent of the light source position. Following [3], we consider the light source and the signal receiver to be at the same position, and the $c(s)$ under the effects of active illumination can be further detailed as:

$$c(s) = \frac{A_0 R(s)}{s^2} \cdot e^{-\int_0^s \sigma(s)ds}, \tag{8}$$

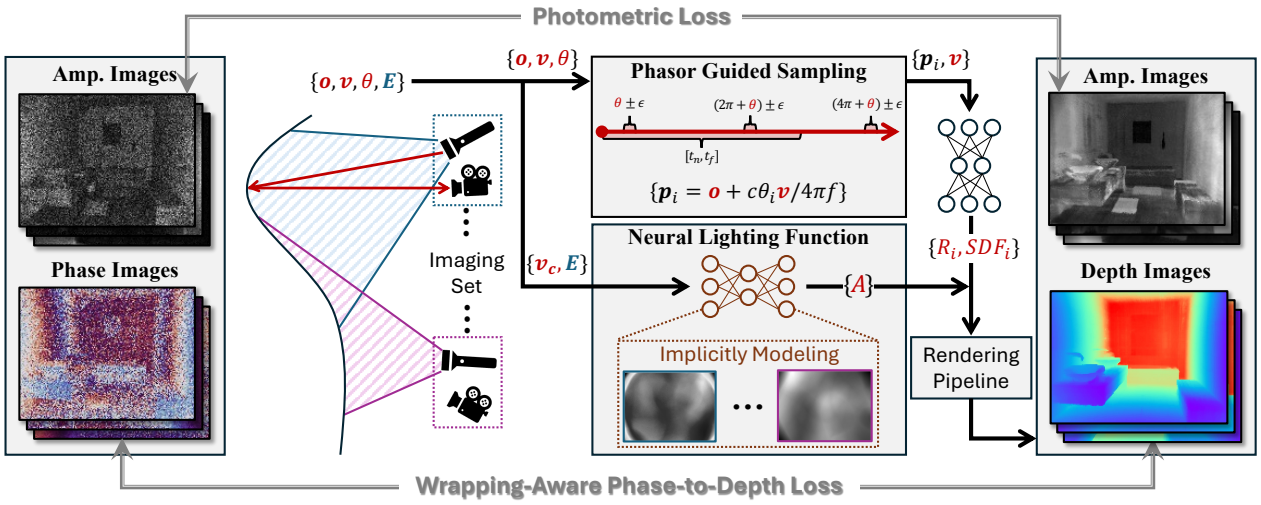


Figure 2. Our method learns 3D scene geometry from a set of iToF imaging data, including phase and amplitude information. For an emitted ray (red arrow), its optical center $\mathbf{o}$, direction $\mathbf{v}$, and measured phase information θ are used for Phasor Guided Sampling, introduced in Sec. 4.3, to obtain a series of sampling points $\mathbf{p}_i$. These sampling points are then queried through a neural network to obtain the corresponding reflectance R_i and signed distance function SDF_i. At the same time, the emitted ray's direction $\mathbf{v}$ and implicit encoding $\mathbf{E}$ are used to determine the scene's lighting response A (Sec. 4.2). Finally, the rendering pipeline introduced in Sec. 4.1 synthesizes phase and amplitude data, which are used to construct supervision along with the imaging results.

where $A_0 \in \mathbb{R}$ is the initial amplitude of the infrared light source, and $1/s^2$ models the geometric attenuation of the amplitude. The term $e^{-\int_0^s \sigma(s)ds}$ accounts for the medium-induced attenuation, while $R(s) \in \mathbb{R}$ reflects the energy loss due to surface reflections. By integrating Eq. 7 and Eq. 8, we derive the radiance integral under active illumination conditions as:

$$\hat{A}_m^n = \int_m^n \left(\frac{A_0 R(s)}{s^2} \sigma(s) e^{-\int_0^s 2\sigma(t)dt} \right) ds. \tag{9}$$

This equation is further discretized as follows:

$$\hat{A} = \sum_{i=1}^{N} e^{\left(-\sum_{j=1}^{i-1} 2\sigma_j \delta_j\right)} (1 - e^{-2\sigma_i \delta_i}) \frac{A_0 R_i}{2t_i^2}, \tag{10}$$

which represents the amplitude information of the received signal. Following NeuS [34], this formulation can be extended to incorporate Signed Distance Functions (SDF) as:

$$A = \sum_{i=1}^{N} T_i \alpha_i \frac{A_0 R_i}{2t_i^2}, \tag{11}$$

where $T_i = \prod_{j=1}^{i-1}(1 - \alpha_j)$ represents the accumulated transmittance up to point i, and the opacity $\alpha_i \in \mathbb{R}$ based on the SDF values is defined as:

$$\alpha_i = \max\left(\frac{\Phi_\beta^2(SDF(t_i)) - \Phi_\beta^2(SDF(t_{i+1}))}{\Phi_\beta^2(SDF(t_i))}, 0 \right) \tag{12}$$

where $\Phi_\beta(x) = (1 + e^{-\beta x})^{-1}$ is the logistic function that controls the smoothness of the transition. This formulation, based on the Signed Distance Representation, models the attenuation of the amplitude of the emitted signal during spatial propagation and reflection, rendering the amplitude information of iToF imaging data. Please refer to the supplementary materials for the detailed derivation process.

Particularly, given a ray $\mathbf{r}(t) = \mathbf{o} + t\mathbf{v}$, where $\mathbf{o} \in \mathbb{R}^3$ is the origin and $\mathbf{v} \in \mathbb{R}^3$ is the direction, we sample a set of points $\{\mathbf{p}_i \in \mathbb{R}^3\}_{i=1}^N$ along the ray. For each sampled point $\mathbf{p}_i$, the Signed Distance Function (SDF) value $SDF_i \in \mathbb{R}$ and the reflection attenuation coefficient R_i are predicted by the neural networks:

$$SDF_i, \mathbf{f}_i = \mathcal{G}_\phi(\mathbf{p}_i), \quad R_i = \mathcal{R}_\psi(\mathbf{p}_i, \mathbf{v}, \mathbf{f}_i), \tag{13}$$

where $\mathbf{f}_i \in \mathbb{R}^{127}$ is the feature vector predicted by the geometry network G_ϕ. The SDF value SDF_i quantifies the distance of each sampled point from the nearest surface, providing essential geometric information about the scene, while R_i models the reflection attenuation. Substituting these values into Eq. 11, we can obtain the rendered amplitude component A. Furthermore, the phase measurement $\hat{\theta}(\mathbf{r})$ can be derived from the depth estimation as:

$$\hat{\theta}(\mathbf{r}) = \left[\frac{4\pi f}{c} \cdot \hat{D}(\mathbf{r}) \right] \bmod 2\pi, \ \hat{D}(\mathbf{r}) = \sum_{i=1}^{N} T_i \alpha_i t_i, \tag{14}$$

where f is the modulated frequency of the emitted infrared signal, c is the speed of light.

4.2. Neural Lighting Function

In an ideal scenario, the infrared light signal emitted is modeled as a point light source, meaning its energy distribution is uniform in all directions [33]. Therefore, the term A_0 in Eq. 11 is considered a constant. However, in practical iToF imaging devices, the light source is often a surface light source, which leads to the inconsistency in the amplitude of the emitted signal across different directions [4]. Additionally, the amplitude of the signal is influenced by the incident angle and the normal vector of the surface. These factors affect the energy distribution in the reflection direction. To address these issues, we propose the Neural Lighting Function, which uses a neural network to implicitly model the amplitude distribution of the emitted signal. Specifically, for a ray $\mathbf{r}$ corresponding to a pixel in the iToF measurement M_j, the initial amplitude of the ray is given by:

$$A_0(\mathbf{r}) = F_\theta(\mathbf{v}_c, \mathbf{E}_j), \tag{15}$$

where $\mathbf{v}_c \in \mathbb{R}^3$ is the direction vector in the camera coordinate system, which can be obtained from the camera's intrinsic parameters; $\mathbf{E}_j \in \mathbb{R}^{32}$ is a learnable implicit encoding corresponding to the original measurement M_j. This encoding captures the specific characteristics of the light source and environmental conditions during the iToF imaging, allowing the neural network to adaptively model the initial energy distribution of the light source for each estimated scenario.

4.3. Phase Guided Sampling

In iToF imaging, the phase measurements are within the range of 0 to 2π, which corresponds to the range of distances related to the modulation frequency f. However, due to the periodicity of phase measurements as introduced in Eq. 5, the observed phase θ may not directly represent the true phase, which could be in a periodic space: $\theta_{\text{true}} = \theta + 2\pi k$, where $k \in \{0, 1, 2 \ldots\}$. To effectively leverage the prior phase information from iToF data and address noise caused by the wrapped phase, we perform uniform sampling within a phase range adjusted for potential phase wraps. Specifically, we consider the true phase values within the interval $[-\epsilon, \epsilon]$ around each periodic phase space $\theta_{\text{true}} = \theta + 2\pi k$. Finally, for the ray r, we perform uniform sampling within the range $[t_n, t_f]$ and combine it with the phase-wrapped interval sampling $\frac{c}{4\pi f}[\theta + 2\pi k - \epsilon, \theta + 2\pi k + \epsilon]$ to obtain the final sampling points. This approach helps mitigate inaccuracies in geometric representation caused by phase ambiguities and noise, ensuring a more robust and accurate modeling of the scene's depth and structure.

4.4. Loss Function

We optimize our neural scene representation by minimizing the following loss functions.

Photometric Loss.

$$\mathcal{L}_p = \frac{1}{|\mathcal{R}|} \sum_{\mathbf{r} \in \mathcal{R}} |\hat{A}(\mathbf{r}) - A(\mathbf{r})|, \tag{16}$$

where $\mathcal{R}$ is a batch of camera rays, $\hat{A}$ is the rendered amplitude and A is the reference amplitude.

Wrapping-Aware Phase-to-Depth Loss.

$$\mathcal{L}_w = \frac{1}{|\mathcal{R}|} \sum_{\mathbf{r} \in \mathcal{R}} \left| \hat{D}(\mathbf{r}) - D(\mathbf{r}) - \delta(\mathbf{r}) \right|, \tag{17}$$

$$\delta(\mathbf{r}) = d \cdot \text{round}\left(\frac{\hat{D}(\mathbf{r}) - D(\mathbf{r})}{d} \right), \tag{18}$$

where $\hat{D}$ represents the rendered depth, and D is the reference depth derived from the phase measurements, given by $D(\mathbf{r}) = c \cdot \theta(\mathbf{r}) / 4\pi f$. The parameter $d = c/2f$ defines the depth range. The round$(\cdot)$ function performs a rounding operation. This loss function is designed to measure the discrepancy between the rendered depth $\hat{D}$ and the reference depth D, with $\delta(\mathbf{r})$ incorporating a wrapping operation to mitigate the effect of periodic depth errors.

Noise-Weight Loss.

$$\mathcal{L}_n = \frac{1}{|\mathcal{R}|} \sum_{\mathbf{r} \in \mathcal{R}} w(\mathbf{r}) \cdot \left\| \hat{D}(\mathbf{r}) - D(\mathbf{r}) - \delta(\mathbf{r}) \right\|, \tag{19}$$

where $w(\mathbf{r}) = \exp(-\mu \left| \hat{D}(\mathbf{r}) - D(\mathbf{r}) - \delta(\mathbf{r}) \right|)$, and we set $\mu = 0.5$ in our experiments.

Eikonal Loss.

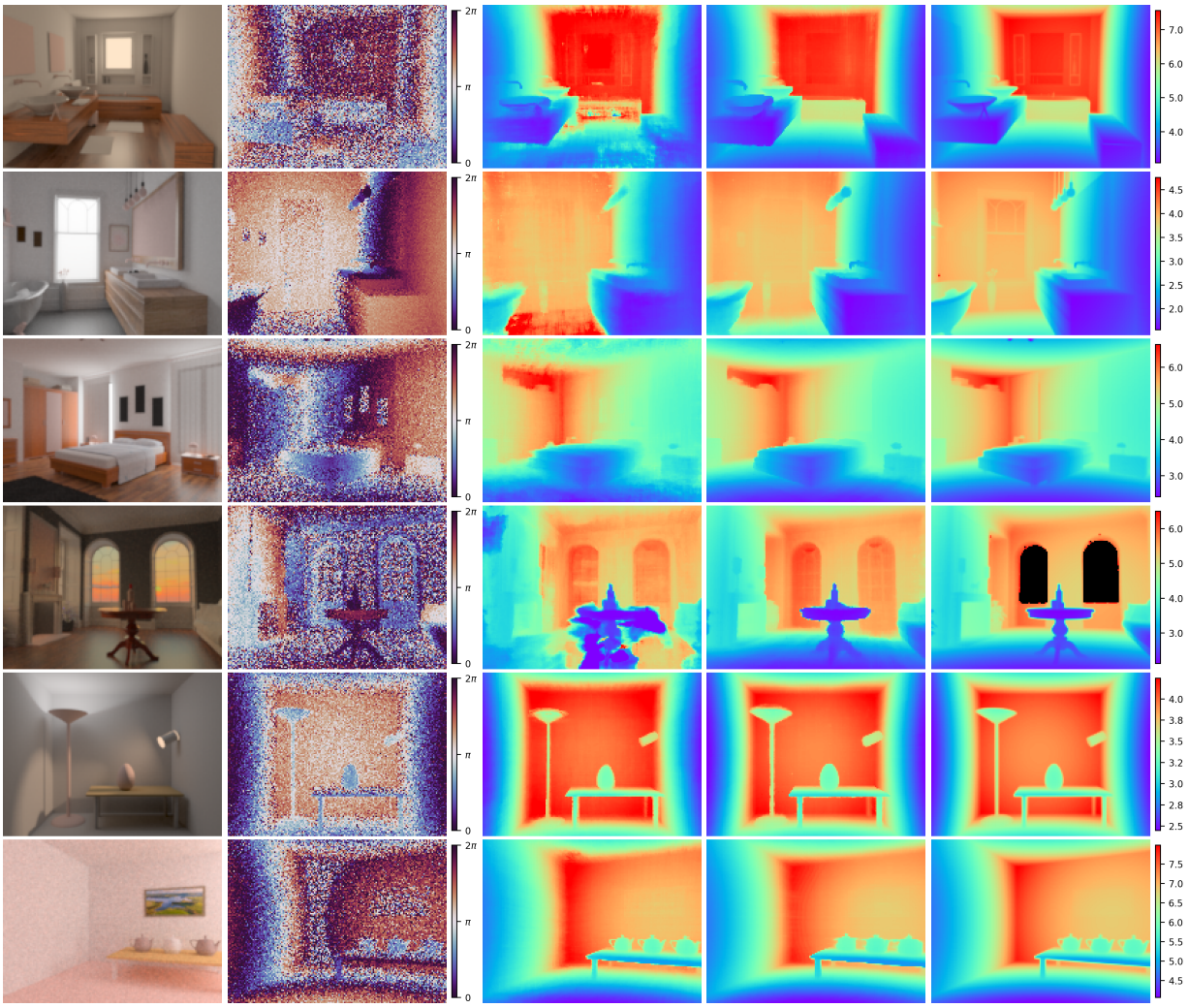
$$\mathcal{L}_r = \frac{1}{|\mathcal{P}|} \sum_{\mathbf{p} \in \mathcal{P}} \left(\|\nabla G_\phi(\mathbf{p})\|_2 - 1 \right)^2, \tag{20}$$

where $\nabla G_\phi(\mathbf{p})$ denotes the gradient of the signed distance function (SDF) at the sample points $\mathbf{p}$. This helps ensure that the SDF represents a valid distance function, improving the stability and accuracy of the model during training.

5. Experiments

5.1. Datasets

Synthetic. We constructed a multi-view reconstruction evaluation dataset based on iToF2dToF [10]. Each sample includes RGB images, ground truth depth maps, iToF frequency images at 31 frequencies (generated via MitsubaToF transient simulation), and camera poses. The scenes primarily contain diffuse materials. For evaluation, we selected six representative indoor scenes with measurements at 20, 40, and 60 MHz frequencies.



(a) RGB Reference (b) iToF Phase (c) TöRF [3] (d) Ours (e) GT Depth

Figure 3. Qualitative comparisons on the synthetic dataset. (b) represents the depth calculated from Eq. 5, visualized in the range of 0 to 2.5 meters for reference. (e), (c), and (d) are visualized within the same depth range to ensure a fair comparison.

Real-world. The dataset proposed by TöRF [3] utilizes a Texas Instruments OPT8241 sensor to capture iToF measurements at a resolution of 320×240 and 30 fps, with an unambiguous range of 5 meters. The scenes are recorded indoors in office environments, featuring a person performing dynamic actions. Since the dataset includes dynamic objects, we apply masks during training to exclude the dynamic parts of the scene, focusing on the static elements.

5.2. Implementation details

We implement our network using the PyTorch framework and train it for 20K iterations on a single NVIDIA A6000 GPU with approximately 6 GB of video memory, completing the process in around 20 minutes. We randomly sample 512 camera rays per optimization step. Specifically, during the first 5000 iterations, the network is trained with a combined photometric loss and regularization loss, defined as $\mathcal{L} = \mathcal{L}_p + 0.001\mathcal{L}_r$. For iterations 5000 to 15000, we introduce the Wrapping-Aware Phase-to-Depth Loss, updating the objective to $\mathcal{L} = \mathcal{L}_p + \mathcal{L}_w + 0.001\mathcal{L}_r$. In the final 5,000 iterations, $\mathcal{L}_w$ is replaced by $\mathcal{L}_n$.We use the Adam optimizer [14], starting with a learning rate of 5×10^{-3} that decays exponentially to 5×10^{-5}.

5.3. Evaluation

Baselines. To evaluate our approach, we conduct comparisons against following methods. iToF Depth refers to depth produced directly by the phase information from iToF measurements. TöRF [3] is a previous iToF-based approach that employs separate networks for static and dynamic parts and then blends them to reconstruct a dynamic scene. For a fair comparison, we disable TöRF's dynamic network to focus

Table 1. Evaluation results on the constructed synthetic dataset. Our method quantitatively outperforms all prior work in all scenes. TöRF† is trained with additional RGB images.

Method		MAE ↓	RMSE ↓	δ_1 ↑		MAE ↓	RMSE ↓	δ_1 ↑		MAE ↓	RMSE ↓	δ_1 ↑
iToF Depth	Bathroom2	2.4918	2.9357	0.2224	Bathroom	0.9526	1.3201	0.6130	Bedroom	1.8428	2.2191	0.3298
NeuS-ToF		2.3439	2.7788	0.2426		0.7706	1.0000	0.4649		0.5027	0.5027	0.7965
NeuS-RGB		0.5405	0.7914	0.8727		0.1629	0.2594	0.9138		0.5304	0.6683	0.6916
TöRF†		0.1470	0.2253	0.9900		0.0787	0.1483	0.9861		0.1400	0.2614	0.9727
TöRF		0.2311	0.4060	0.9424		0.2305	0.4083	0.9097		0.1328	0.2056	0.9776
Ours		**0.0427**	**0.1059**	**0.9968**		**0.0404**	**0.0900**	**0.9977**		**0.0337**	**0.0688**	**0.9984**
iToF Depth	Living-room	2.3747	2.6516	0.1695	Veach-bidir	1.4151	1.6978	0.4266	Veach-ajar	3.7801	4.0180	0.0312
NeuS-ToF		0.5742	1.1008	0.7575		0.0516	0.1284	0.9726		0.5800	1.1071	0.7794
NeuS-RGB		0.3210	0.6513	0.8752		0.5208	0.6521	0.6493		0.2766	0.4401	0.9215
TöRF†		0.3459	0.6683	0.8559		0.0977	0.3442	0.9992		0.2180	0.3887	0.9692
TöRF		0.6170	0.8930	0.7468		0.0811	0.1021	0.9993		0.1171	0.1844	0.9935
Ours		**0.0820**	**0.2846**	**0.9771**		**0.0393**	**0.0633**	**0.9994**		**0.0428**	**0.0911**	**0.9982**

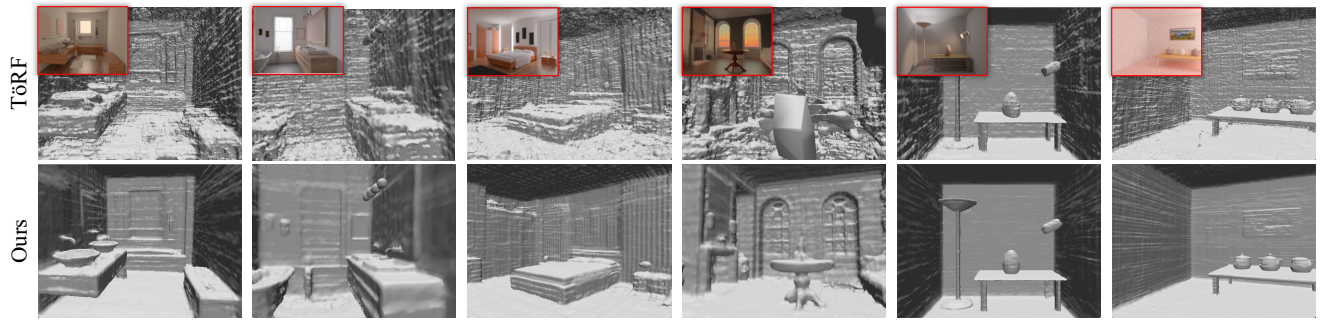


Figure 4. Comparisons of surface reconstruction on the synthetic dataset. RGB images are presented for reference.

Table 2. Experimental results on the effects of the data information. Metrics are averaged over 6 scenes from the synthetic dataset.

Supervision	MAE↓	RMSE ↓	δ_1 ↑
w/o Phase	0.3712	0.4957	0.7839
w/o Amplitude	2.9484	3.0702	0.1666
Phase & Amplitude	**0.0468**	**0.1173**	**0.9946**

on static reconstruction. Since TöRF uses both RGB and iToF information, we additionally provide results where the model is trained solely on iToF data. Additionally, we have built two baselines based on NeuS [34]. Neus-ToF uses our framework for training but applies NeuS' rendering scheme to render amplitude maps. Neus-RGB is the result of training with RGB images without iToF data.

Quantitative Results. Table 1 presents the error metrics calculated between the rendered depth and ground truth to evaluate reconstruction accuracy. All scenes were captured with the modulation frequency of 60 MHz, corresponding to a depth range of $0 \sim 2.5$ meters for the iToF camera. In these challenging scenes, our method demonstrates superior performance across all scenes. NeuS-ToF did not achieve strong results, confirming the effectiveness of our proposed rendering scheme under active illumination.

Qualitative Results. As shown in Fig. 3, we provide visual comparisons of the rendered depth maps from our method and TöRF [3]. From the first to the fourth rows, it can be observed that TöRF [3] introduces significant artifacts in complex scenes, whereas our method consistently provides accurate reconstruction results. The fourth and fifth rows illustrate the reconstruction results in simpler scenes, where our method still achieves more accurate results. Fig. 4 presents the surface construction results, and our method can better reconstruct the geometric structure of the scene. This demonstrates that our SDF-based representation outperforms TöRF in achieving better surface reconstruction. Moreover, we present the results from a real-world scenario, as shown in Fig. 5. Unlike the synthetic dataset, noise in real-world scenes often exists in local areas, which we highlight with a black box. After fusing the multi-view information, the noise is effectively removed. Additionally, the scene suffers from wrapped estimation issues (as shown in the right area of the image), which is also addressed by our method. For more visualizations in real-world scenarios, please refer to the supplementary materials.

5.4. Ablation Study

In this section, we conduct ablation studies on the synthetic dataset to evaluate the proposed methods.

Scene Representation. Since iToF imaging includes both amplitude and phase information, we investigate the impact of these components on reconstruction, as shown in Tab. 2. Without amplitude supervision, the network fails to con-

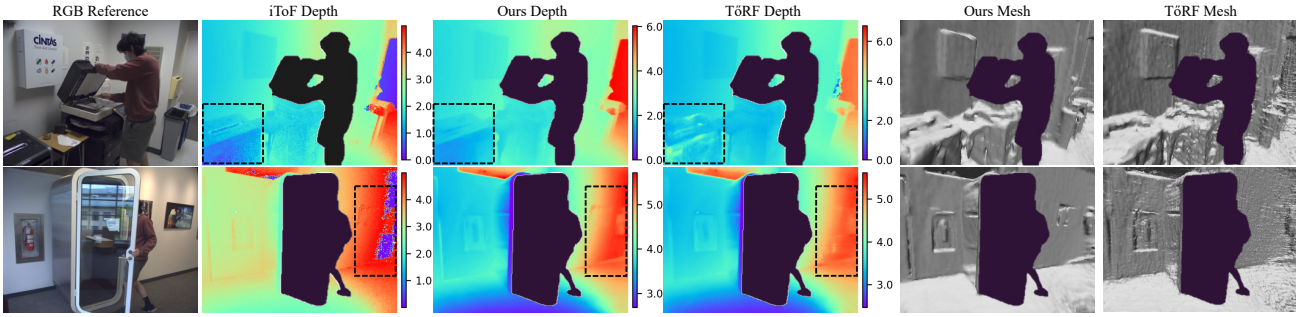


Figure 5. Visual comparisons in the real-world scenario. The scene was captured using an iToF camera with a 30 MHz modulation frequency, corresponding to a depth range of $0 \sim 5m$ meters.

Table 3. An ablation study of the techniques introduced to facilitate the optimization of the scene representation. Metrics are averaged over the 6 scenes from the synthetic dataset. $\mathcal{L}_w$ is introduced in Eq. 17. The Sampling Strategy is presented in Sec. 4.3. $\mathcal{L}_n$ is denoted with Eq. 19.

$\mathcal{L}_w$	Sampling Strategy	$\mathcal{L}_n$	MAE ↓	RMSE ↓	δ_1 ↑
-	-	-	0.9039	1.5218	0.7908
✓	-	-	0.0535	0.1307	0.9909
-	✓	-	0.3901	0.6990	0.8465
✓	✓	-	0.0510	0.1264	0.9881
✓	✓	✓	**0.0468**	**0.1173**	**0.9946**

verge. In contrast, the network can still perform moderately without phase supervision. Our approach achieves superior performance with both amplitude and phase information, featuring the importance of jointly considering these two components.

Modulation Frequency. We conduct ablation experiments to examine the impact of modulation frequency, ranging from 20 MHz to 60 MHz, which corresponds to a depth range of the iToF camera from 7.5 m to 2.5 m. The results are shown in Tab. 4. Our method achieves promising 3D reconstructions across different frequencies and leverages the higher accuracy of higher frequencies to produce more precise 3D reconstructions.

Loss Items $\mathcal{L}_w$ and $\mathcal{L}_n$. To evaluate the effectiveness of the proposed Wrapping-Aware Phase-to-Depth Loss, we replace it with a simple phase loss, formulated as $\mathcal{L} = \frac{1}{|\mathcal{R}|}\sum_{\mathbf{r}\in\mathcal{R}}\left\|\hat{\theta}(\mathbf{r}) - \theta(\mathbf{r})\right\|$. From the first row and second of Tab. 3, the proposed loss effectively recovers accurate geometric reconstruction results from wrapped estimations. Moreover, the fifth row of the Tab. 3 shows that the introduced noise-weight loss $\mathcal{L}_n$ further improves performance. Please refer to the visualization in the supplementary materials, where $\mathcal{L}_n$ helps recover more scene details.

Phase Guided Sampling. To validate the effectiveness of the proposed Phase Guided Sampling scheme, we replace it with uniform sampling in $[t_n, t_f]$. As shown in Tab. 3, the proposed sampling strategy further improves performance.

Table 4. Experiments on the effects of the modulation frequency. Metrics are averaged over 6 scenes from the synthetic dataset.

Frequency	Depth Range	MAE ↓	RMSE ↓	δ_1 ↑
20M	$0 \sim 7.5m$	0.1018	0.1735	0.9943
40M	$0 \sim 3.75m$	0.0634	0.1353	0.9914
60M	$0 \sim 2.5m$	**0.0468**	**0.1173**	**0.9946**

Table 5. Quantitative comparisons of the MAE error with the depth measurements only affected by MPI noise.

	Bathroom2	Bathroom	Bedroom	Living-Room	Veach-bidir	Veach-ajar
MPI	0.1340	0.1505	0.1868	0.1671	0.1541	0.2251
Ours	**0.0937**	**0.0700**	**0.1026**	**0.1208**	**0.1086**	**0.1148**

MPI Removal. Our work does not explicitly model the generation of MPI noise. Our motivation is that the impact of MPI varies across different viewpoints. Therefore, by constraining a unified scene representation with multi-view imaging results, we can smooth the noise and learn a more accurate expression. We conduct evaluations on data with a modulation frequency of 20M, as shown in Tab. 5. The 'MPI' represents the error of the depth measurements only affected by multi-path interference noise. Our depth errors are lower, suggesting that multi-path interference noise is reduced. Please refer to the supplementary materials for further discussion.

6. Conclusion

We propose a framework to achieve accurate 3d reconstruction from noisy amplitude and phase images of iToF camera with neural scene representation. The developed rendering scheme, lighting modeling, sampling strategy, and loss items improve the scene representation to obtain accurate measurements efficiently. Experimental results demonstrate that our method achieves state-of-the-art performance in both qualitative and quantitative ways. **Limitations**. Our method has several limitations. It relies on calibrated camera poses. The approach also struggles with strong specular reflections and complex multi-path interference.

Acknowledgement

This work was supported by National Defense Science and Technology Foundation Strengthening Program Funding (Grant. 2023-JCJQ-JJ-0219).

References

[1] Gianluca Agresti, Henrik Schaefer, Piergiorgio Sartor, and Pietro Zanuttigh. Unsupervised domain adaptation for tof data denoising with adversarial learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5584–5593, 2019. 2

[2] Guillem Alenyà, Sergi Foix, and Carme Torras. Tof cameras for active vision in robotics. *Sensors and Actuators A: Physical*, 218:10–22, 2014. 1

[3] Benjamin Attal, Eliot Laidlaw, Aaron Gokaslan, Changil Kim, Christian Richardt, James Tompkin, and Matthew O'Toole. TöRF: Time-of-flight radiance fields for dynamic scene view synthesis. *Advances in Neural Information Processing Systems*, 34:26289–26301, 2021. 2, 3, 6, 7, 5, 8, 9

[4] Cyrus Bamji, John Godbaz, Minseok Oh, Swati Mehta, Andrew Payne, Sergio Ortiz, Satyadev Nagaraja, Travis Perry, and Barry Thompson. A review of indirect time-of-flight technologies. *IEEE Transactions on Electron Devices*, 69 (6):2779–2793, 2022. 1, 2, 5

[5] 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. 3

[6] Mark Boss, Raphael Braun, Varun Jampani, Jonathan T Barron, Ce Liu, and Hendrik Lensch. Nerd: Neural reflectance decomposition from image collections. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12684–12694, 2021. 3

[7] Guanting Dong, Yueyi Zhang, and Zhiwei Xiong. Spatial hierarchy aware residual pyramid network for time-of-flight depth denoising. In *Proceedings of the European Conference on Computer Vision*, pages 35–50. Springer, 2020. 2

[8] Guanting Dong, Yueyi Zhang, Xiaoyan Sun, and Zhiwei Xiong. Exploiting dual-correlation for multi-frame time-of-flight denoising. In *Proceedings of the European Conference on Computer Vision*, pages 473–489. Springer, 2025. 2

[9] Qi Guo, Iuri Frosio, Orazio Gallo, Todd Zickler, and Jan Kautz. Tackling 3d tof artifacts through learning and the flat dataset. In *Proceedings of the European Conference on Computer Vision*, pages 368–383, 2018. 2

[10] Felipe Gutierrez-Barragan, Huaijin Chen, Mohit Gupta, Andreas Velten, and Jinwei Gu. itof2dtof: A robust and flexible representation for data-driven time-of-flight imaging. *IEEE Transactions on Computational Imaging*, 7:1205–1214, 2021. 2, 5

[11] Jonas Gutknecht and Teddy Loeliger. Multi-layer tof: Comparison of different multipath resolve methods for indirect 3d time-of-flight. In *2021 IEEE Sensors*, pages 1–4. IEEE, 2021. 2

[12] Aleksander Holynski and Johannes Kopf. Fast depth densification for occlusion-aware augmented reality. *ACM Transactions on Graphics*, 37(6):1–11, 2018. 1

[13] HyunJun Jung, Nikolas Brasch, Aleš Leonardis, Nassir Navab, and Benjamin Busam. Wild tofu: Improving range and quality of indirect time-of-flight depth with rgb fusion in challenging environments. In *International conference on 3D vision*, pages 239–248. IEEE, 2021. 2

[14] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015. 6

[15] Chunyu Li, Taisuke Hashimoto, Eiichi Matsumoto, and Hiroharu Kato. Multi-view neural surface reconstruction with structured light. In *British Machine Vision Conference*, 2022. 3

[16] Larry Li et al. Time-of-flight camera—an introduction. *Technical white paper*, (SLOA190B), 2014. 1, 2

[17] Julio Marco, Quercus Hernandez, Adolfo Munoz, Yue Dong, Adrian Jarabo, Min H Kim, Xin Tong, and Diego Gutierrez. Deeptof: off-the-shelf real-time correction of multipath interference in time-of-flight imaging. *ACM Transactions on Graphics*, 36(6):1–12, 2017. 2

[18] Nelson Max. Optical models for direct volume rendering. *IEEE Transactions on Visualization and Computer Graphics*, 1(2):99–108, 1995. 3, 1

[19] Stefan May, Bjorn Werner, Hartmut Surmann, and Kai Pervolz. 3d time-of-flight cameras for mobile robotics. In *2006 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 790–795. Ieee, 2006. 1

[20] Yu Meng, Zhou Xue, Xu Chang, Xuemei Hu, and Tao Yue. itof-flow-based high frame rate depth imaging. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4929–4938, 2024. 2

[21] 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. *Communications of the ACM*, 65(1):99–106, 2021. 2, 3, 1

[22] Ben Mildenhall, Peter Hedman, Ricardo Martin-Brualla, Pratul P. Srinivasan, and Jonathan T. Barron. NeRF in the Dark: High dynamic range view synthesis from noisy raw images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16169–16178, 2022. 2

[23] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM Transactions on Graphics*, 41 (4):1–15, 2022. 3

[24] Mikhail Okunev, Marc Mapeke, Benjamin Attal, Christian Richardt, Matthew O'Toole, and James Tompkin. Flowed time of flight radiance fields. In *Proceedings of the European Conference on Computer Vision*, pages 373–389. Springer, 2025. 3

[25] Jiamin Ping, Yue Liu, and Dongdong Weng. Comparison in depth perception between virtual reality and augmented

reality systems. In *2019 IEEE Conference on Virtual Reality and 3D User Interfaces*, pages 1124–1125. IEEE, 2019. 1

[26] 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*, pages 10318–10327, 2021. 3

[27] Xin Qiao, Matteo Poggi, Pengchao Deng, Hao Wei, Chenyang Ge, and Stefano Mattoccia. Rgb guided tof imaging system: A survey of deep learning-based methods. *International Journal of Computer Vision*, pages 1–38, 2024. 1

[28] Hazem Rashed, Mohamed Ramzy, Victor Vaquero, Ahmad El Sallab, Ganesh Sistu, and Senthil Yogamani. Fusemodnet: Real-time camera and lidar based moving object detection for robust low-light autonomous driving. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0, 2019. 1

[29] Michael Schelling, Pedro Hermosilla, and Timo Ropinski. Radu: Ray-aligned depth update convolutions for tof data denoising. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 671–680, 2022. 2

[30] 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. 3

[31] Pratul P Srinivasan, Boyang Deng, Xiuming Zhang, Matthew Tancik, Ben Mildenhall, and Jonathan T Barron. Nerv: Neural reflectance and visibility fields for relighting and view synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7495–7504, 2021. 3

[32] Shuochen Su, Felix Heide, Gordon Wetzstein, and Wolfgang Heidrich. Deep end-to-end time-of-flight imaging. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6383–6392, 2018. 2

[33] Channing P Verbeck and Donald P Greenberg. A comprehensive light-source description for computer graphics. *IEEE Computer Graphics and Applications*, 4(7):66–75, 1984. 5

[34] 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. In *Advances in Neural Information Processing Systems*, pages 27171–27183, 2021. 4, 7, 1, 2

[35] Xuan Wang, Kaiqiang Li, and Abdellah Chehri. Multi-sensor fusion technology for 3d object detection in autonomous driving: A review. *IEEE Transactions on Intelligent Transportation Systems*, 2023. 1

[36] Tao Yang, You Li, Cheng Zhao, Dexin Yao, Guanyin Chen, Li Sun, Tomas Krajnik, and Zhi Yan. 3d tof lidar in mobile robotics: A review. *arXiv preprint arXiv:2202.11025*, 2022. 1

[37] Kentaro Yoshioka. A tutorial and review of automobile direct tof lidar socs: evolution of next-generation lidars. *IEICE Transactions on Electronics*, 105(10):534–543, 2022. 1

[38] 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*, pages 4578–4587, 2021. 3

[39] Kai Zhang, Fujun Luan, Zhengqi Li, and Noah Snavely. Iron: Inverse rendering by optimizing neural sdfs and materials from photometric images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5565–5574, 2022. 3

[40] Tianyi Zhang, Kaining Huang, Weiming Zhi, and Matthew Johnson-Roberson. Darkgs: Learning neural illumination and 3d gaussians relighting for robotic exploration in the dark. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 12864–12871, 2024. 3