

Learning Neural Scene Representation from iToF Imaging

Supplementary Material

8. Rendering Scheme

Based on [18, 21, 34], we introduce the derivation of the classic volume rendering equation and extend it to render iToF measurements based on the Signed Distance Function (SDF). First, we present the absorption process. This process describes how light is attenuated as it passes through a medium containing particles. The number of particles in a small volume of length Δs is given by $\rho E \Delta s$, where ρ represents the density of particles and E is the area. The fraction of light passing through the volume is given by $\frac{P \rho E \Delta s}{E} = P \rho \Delta s$, where P is the projected area of one particle. Thus, the change in intensity due to absorption is:

$$A_{out} - A_{in} = -P \rho \Delta s A_{in}. \quad (21)$$

Taking the differential form, we have:

$$\frac{dA}{ds} = -\rho(s) P A(s) = -\sigma(s) A(s) \quad (22)$$

where $\sigma(s) = \rho(s) P$ represents the rate of light attenuation per unit distance.

Emission describes how light is added to the system from the particles themselves. The light emitted over a small segment Δs is given by $\frac{C \rho P E \Delta s}{E}$, where $C(s)$ is the emission coefficient. Thus, the emission differential equation is:

$$\frac{dA}{ds} = C(s) \rho(s) P = C(s) \sigma(s). \quad (23)$$

The combined effect of absorption and emission is:

$$\begin{aligned} \frac{dA}{ds} &= C(s) \sigma(s) - \sigma(s) A(s) \\ \frac{dA}{ds} + \sigma(s) A(s) &= C(s) \sigma(s) \end{aligned} \quad (24)$$

To simplify, we recognize the equation in a form where the product rule applies:

$$\frac{d}{ds} \left(A(s) e^{\int_0^s \sigma(x) dx} \right) = \frac{dA}{ds} e^{\int_0^s \sigma(x) dx} + A(s) \sigma(s) e^{\int_0^s \sigma(x) dx} \quad (25)$$

The Eq. 24 can be rewritten as:

$$\frac{d}{ds} \left(A(s) e^{\int_0^s \sigma(x) dx} \right) = C(s) \sigma(s) e^{\int_0^s \sigma(x) dx} \quad (26)$$

We integrate this equation from the start of the light path $s = 0$ to the camera pixel at $s = D$:

$$A(D) e^{\int_0^D \sigma(x) dx} - A(0) = \int_0^D \left(C(s) \sigma(s) e^{\int_0^s \sigma(x) dx} \right) ds \quad (27)$$

Multiplying both sides by $e^{-\int_0^D \sigma(x) dx}$, we obtain:

$$A(D) = A(0) e^{-\int_0^D \sigma(x) dx} + \int_0^D \left(C(s) \sigma(s) e^{\int_0^s \sigma(x) dx - \int_0^D \sigma(x) dx} \right) ds. \quad (28)$$

the contribution from ambient light, denoted as $A(0)$, is considered to be zero. Thus, the equation simplifies further.

$$A(D) = \int_0^D \left(C(s) \sigma(s) e^{-\int_s^D \sigma(x) dx} \right) ds \quad (29)$$

While NeRF models particles as self-emitting light sources that are constant from the same view direction, in active illumination, $C(s)$ is regarded as the combination of the light source contribution $L(s)$ and the reflectivity $R(s)$. For a light source located at $s = D$, referring to Eq. 21, the intensity attenuation of the light source is:

$$\begin{aligned} L(s + \delta) - L(s) &= \sigma(s) L(s) \\ \int_D^s \frac{dL}{L(s)} &= \int_D^s \sigma(x) dx \\ \ln |L(s)| - \ln |L(D)| &= \int_D^s \sigma(x) dx \\ \frac{|L(s)|}{|L(D)|} &= e^{\left(\int_D^s \sigma(x) dx \right)} \\ L(s) &= L(D) e^{\left(\int_D^s \sigma(x) dx \right)} \end{aligned} \quad (30)$$

Integrating both geometric attenuation and reflectance attenuation, Eq. 29 can be written as:

$$A(D) = \int_0^D \left(\frac{L(D) R(s) \sigma(s)}{(D-s)^2} e^{-\int_s^D 2\sigma(x) dx} \right) ds \quad (31)$$

To isolate the contribution of light from a specific segment $[m, n]$ along the path:

$$A_m^n = \int_m^n \left(\frac{L(D) R(s) \sigma(s)}{(D-s)^2} e^{-\int_s^D 2\sigma(x) dx} \right) ds \quad (32)$$

We denote $t^* = 0.5(m+n)$, Assuming $(D-s)^2$ and $R(s)$ and $\sigma(s)$ are constants over this segment

$$A_m^n = \frac{L(D) R \sigma}{(D-t^*)^2} \int_m^n e^{-\int_s^D 2\sigma(x) dx} ds \quad (33)$$

$$= \frac{L(D) R \sigma}{(D-t^*)^2} e^{-\int_n^D 2\sigma(x) dx} \int_m^n e^{-2\sigma(n-s)} ds \quad (34)$$

$$= \frac{L(D) R \sigma}{(D-t^*)^2} e^{-\int_n^D 2\sigma(x) dx} \left(\frac{e^{-2\sigma(n-s)}}{2\sigma} \right) \Big|_m^n \quad (35)$$

$$= \frac{L(D) R}{(D-t^*)^2} e^{-\int_n^D 2\sigma(x) dx} (1 - e^{-2\sigma(n-m)}) \quad (36)$$

denote $\delta = n - m$

$$A_i = \frac{L(D)R}{(D - t^*)^2} e^{-\int_n^D 2\sigma(x)dx} (1 - e^{-2\sigma\delta_i}) \quad (37)$$

Next, we perform the substitution $D - t^* = t$ and $A_0 = L(D)$ that represents the initial light intensity. This allows us to rewrite the formula as:

$$A_i = \frac{L(D)R}{t^2} e^{-\int_0^{D-t} 2\sigma(t)dt} (1 - e^{-2\sigma\delta_i}). \quad (38)$$

By summing the contributions from each segment, we obtain the rendered phase measurement:

$$\hat{A} = \sum_{i=1}^N e^{(-\sum_{j=1}^{i-1} 2\sigma_j\delta_j)} (1 - e^{-2\sigma_i\delta_i}) \frac{A_0 R_i}{2t_i^2} \quad (39)$$

$$= \sum_{i=1}^N T_i \alpha_i \frac{A_0 R_i}{2t_i^2} \quad (40)$$

where $\alpha_i = (1 - e^{-2\sigma\delta_i})$, $T_i = \prod_{j=1}^{i-1} (1 - \alpha_j)$ represents the accumulated transmittance. From NeuS [34], the density is represented with the SDF value as:

$$\sigma(t) = \max\left(\frac{-\frac{d}{dt}\Phi_s(SDF(t))}{\Phi_s(SDF(t))}, 0\right) \quad (41)$$

$$\int_{t_i}^{t_{i+1}} \frac{-\frac{d}{dt}\Phi_s(SDF(t))}{\Phi_s(SDF(t))} dt = \ln\left(\frac{\Phi_s(SDF(t_i))}{\Phi_s(SDF(t_{i+1}))}\right) \quad (42)$$

We substitute the above conclusion into $\alpha_i = (1 - e^{-2\sigma\delta_i})$. When $\frac{-\frac{d}{dt}\Phi_s(SDF(t))}{\Phi_s(SDF(t))} > 0$, we have:

$$\alpha_i = 1 - e^{-\int_{t_i}^{t_{i+1}} 2 \max\left(\frac{-\frac{d}{dt}\Phi_s(SDF(t))}{\Phi_s(SDF(t))}, 0\right) dt} \quad (43)$$

$$= 1 - e^{-2 \int_{t_i}^{t_{i+1}} \frac{-\frac{d}{dt}\Phi_s(SDF(t))}{\Phi_s(SDF(t))} dt} \quad (44)$$

$$= 1 - e^{\left(-2 \ln\left(\frac{\Phi_s(SDF(t_i))}{\Phi_s(SDF(t_{i+1}))}\right)\right)} \quad (45)$$

$$= 1 - \frac{\Phi_s^2(SDF(t_{i+1}))}{\Phi_s^2(SDF(t_i))} \quad (46)$$

When $\frac{-\frac{d}{dt}\Phi_s(SDF(t))}{\Phi_s(SDF(t))} \leq 0$, we have:

$$\alpha_i = 1 - e^{-\int_{t_i}^{t_{i+1}} 2 \max\left(\frac{-\frac{d}{dt}\Phi_s(SDF(t))}{\Phi_s(SDF(t))}, 0\right) dt} \quad (47)$$

$$= 1 - e^{-2 \int_{t_i}^{t_{i+1}} 0 dt} \quad (48)$$

$$= 0 \quad (49)$$

In summary, we have derived the relationship between α_i and the SDF value under active illumination:

$$\alpha_i = \max\left(\frac{\Phi_s^2(SDF(t_i)) - \Phi_s^2(SDF(t_{i+1}))}{\Phi_s^2(SDF(t_i))}, 0\right) \quad (50)$$

9. iToF Imaging

In this section, we provide a detailed explanation of the camera model used in indirect Time-of-Flight (iToF) cameras, as described in [9, 16]. ToF cameras emit amplitude-modulated infrared signals. These signals can be mathematically represented as:

$$g(t) = g_1 \cos(2\pi ft) + g_0 \quad (51)$$

Here, g_1 represents the modulation amplitude, which controls the strength of the emitted signal's modulation. The symbol f denotes the modulation frequency. The term g_0 is the DC offset, representing the constant baseline of the emitted signal. This modulated signal illuminates the scene. Light reflected from objects returns to the camera with a time delay σ_0 , which depends on the distance to the object. The camera receives this reflected signal, which can be described by:

$$S_{rec}(t) = R \frac{g_1 \cos(2\pi ft - 2\pi f\tau_0) + g_0}{c^2 \tau_0^2} + e_0, \quad (52)$$

where R represents the reflectivity of the surface. Factors like material properties, color, and angle of view influence it. The constant c denotes the speed of light, accounting for the signal's travel speed. The parameter τ_0 indicates the time delay of the reflected signal, which is directly related to the object's distance. The term e_0 represents the contribution from environmental light. It adds a constant component to the overall signal received by the camera. We simplify the components of the received signal as the amplitude r_1 and the DC component r_0 . Their corresponding relationships are expressed as:

$$r_1 = R \frac{g_1}{c^2 \tau_0^2}, \quad r_0 = R \frac{g_0}{c^2 \tau_0^2} + e_0. \quad (53)$$

The iToF camera determines the phase shift $2\pi f\tau_0$ and amplitude by multiplying the received signal $s(t)$ with an internal reference signal $b \cos(2\pi ft - \phi)$. Here, ϕ is a programmable phase shift. By substituting $s(t) = r_1 \cos(2\pi ft - 2\pi f\tau_0) + r_0$ into the multiplication, we obtain the resulting signal:

$$i(t) = (r_1 \cos(2\pi ft - 2\pi f\tau_0) + r_0) \cdot b \cos(2\pi ft - \phi). \quad (54)$$

Expanding this expression, we have:

$$\begin{aligned} i(t) &= b[r_1 \cos(2\pi ft - 2\pi f\tau_0) \cos(2\pi ft - \phi) + r_0 \cos(2\pi ft - \phi)] \\ &= \frac{br_1}{2} \cos(\phi - 2\pi f\tau_0) + \frac{br_1}{2} \cos(2\pi ft - 2\pi f\tau_0 - \phi) \\ &\quad + br_0 \cos(2\pi ft - \phi). \end{aligned}$$

From the above equation, we can see that the resulting signal $i(t)$ contains a DC term $\frac{br_1}{2} \cos(\phi - 2\pi f\tau_0)$ that is related to the phase ϕ . It also includes high-frequency compo-

nents. By integrating over time, we can filter out the oscillatory terms, leaving the phase-related components for subsequent phase and amplitude calculations. The raw measurements I_ϕ captured at programmed phase settings (0° , 90° , 180° , and 270°) can be expressed as:

$$I_\phi = \frac{r_1 \cdot b}{2} \cos(\phi - 2\pi f \tau_0) + a_0, \quad (55)$$

where a_0 represents the combined contribution from environmental light. Substituting the programmed phase settings into this equation, we get:

$$\begin{aligned} I_0 &= \frac{r_1 \cdot b}{2} \cos(-2\pi f \tau_0) + a_0 = \frac{r_1 \cdot b}{2} \cos(2\pi f \tau_0) + a_0 \\ I_{90} &= \frac{r_1 \cdot b}{2} \cos\left(\frac{\pi}{2} - 2\pi f \tau_0\right) + a_0 = \frac{r_1 \cdot b}{2} (-\sin(2\pi f \tau_0)) + a_0 \\ I_{180} &= \frac{r_1 \cdot b}{2} \cos(\pi - 2\pi f \tau_0) + a_0 = \frac{r_1 \cdot b}{2} (-\cos(2\pi f \tau_0)) + a_0 \\ I_{270} &= \frac{r_1 \cdot b}{2} \cos\left(\frac{3\pi}{2} - 2\pi f \tau_0\right) + a_0 = \frac{r_1 \cdot b}{2} \sin(2\pi f \tau_0) + a_0 \end{aligned}$$

From these measurements, the phase shift $2\pi f \tau_0$ can be computed as:

$$\theta = 2\pi f \tau_0 = \arctan\left(\frac{I_{90} - I_{270}}{I_{180} - I_0}\right). \quad (56)$$

By substituting the expressions for I_{90} , I_{270} , I_{180} , and I_0 , we obtain::

$$\theta = \arctan\left(\frac{\left(\frac{r_1 \cdot b}{2} (-\sin(2\pi f \tau_0)) + a_0\right) - \left(\frac{r_1 \cdot b}{2} \sin(2\pi f \tau_0) + a_0\right)}{\left(\frac{r_1 \cdot b}{2} (-\cos(2\pi f \tau_0)) + a_0\right) - \left(\frac{r_1 \cdot b}{2} \cos(2\pi f \tau_0) + a_0\right)}\right)$$

Simplifying further, we get:

$$\begin{aligned} \theta &= \arctan\left(\frac{-r_1 \cdot b \sin(2\pi f \tau_0)}{-r_1 \cdot b \cos(2\pi f \tau_0)}\right) \\ &= \arctan\left(\frac{\sin(2\pi f \tau_0)}{\cos(2\pi f \tau_0)}\right). \end{aligned}$$

Then, the object's distance is calculated using:

$$d = \frac{c \cdot \tau_0}{2} = \frac{c\theta}{4\pi f}. \quad (57)$$

The amplitude, which encodes information about reflectivity and illumination, is computed as:

$$\begin{aligned} A &= \sqrt{(I_0 - I_{180})^2 + (I_{270} - I_{90})^2} \\ &= \sqrt{r_1 b \cos^2(2\pi f \tau_0) + r_1 b \sin^2(2\pi f \tau_0)} \\ &= r_1 \cdot b = R \frac{g_1}{c^2 \sigma_0^2} \cdot b. \end{aligned}$$

From the above equations, we observe that the amplitude of the iToF signal is influenced by the scene's reflectivity R and the object's distance. This relationship is also evident in the rendering formula Eq. 11 we derived.

10. Depth Metrics

We use the following metrics to evaluate the performance: mean absolute error (MAE), root mean square error of linear measures (RMSE) and relative accuracy δ_1 (the fraction of pixels where the relative error is within a threshold of 1.25). All errors are calculated in meters. The mathematical expressions of the evaluation metrics are presented in the following:

$$\text{Mean Absolute Error: } MAE = \frac{1}{n} \sum_p^n \frac{|y_p - \hat{y}_p|}{\hat{y}_p}$$

$$\text{Root Mean Square Error: } MSE = \sqrt{\frac{1}{n} \sum_p^n (y_p - \hat{y}_p)^2}$$

$$\text{Threshold: } \delta = \% \text{ of } y_p \text{ s.t. } \max\left(\frac{y_p}{\hat{y}_p}, \frac{\hat{y}_p}{y_p}\right) = \delta < 1.25$$

where n denotes the number of valid pixels of a panorama image, \hat{y}_p denotes the depth value in ground truth and y_p is the depth value predicted from networks.

11. Network Details

The Multi-Layer Perceptron (MLP) structure we use is shown in Fig. 6. Our model consists of simple MLPs: \mathcal{G}_ϕ is a six-layer MLP, and both \mathcal{R}_ψ and F_θ are two-layer MLPs, each with a hidden dimension of 128.

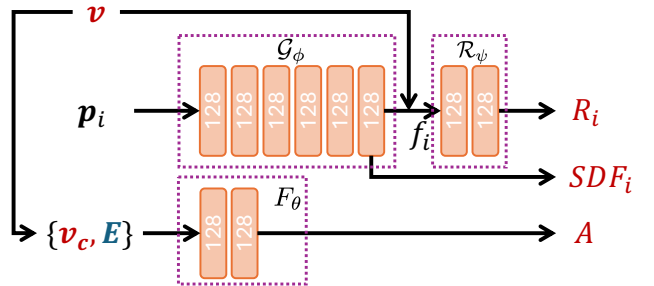


Figure 6. Network Structure

12. Effectiveness of the Noise-Weight Loss

We conducted an experiment to evaluate the impact of using and not using the noise-weight loss term while training. As shown in Figure 7, the first row is the ground truth depth, the second row is the depth map without the loss term, and the third row is the depth map with the loss term. The results show that using the noise-weight loss term improves the object details in the depth map.

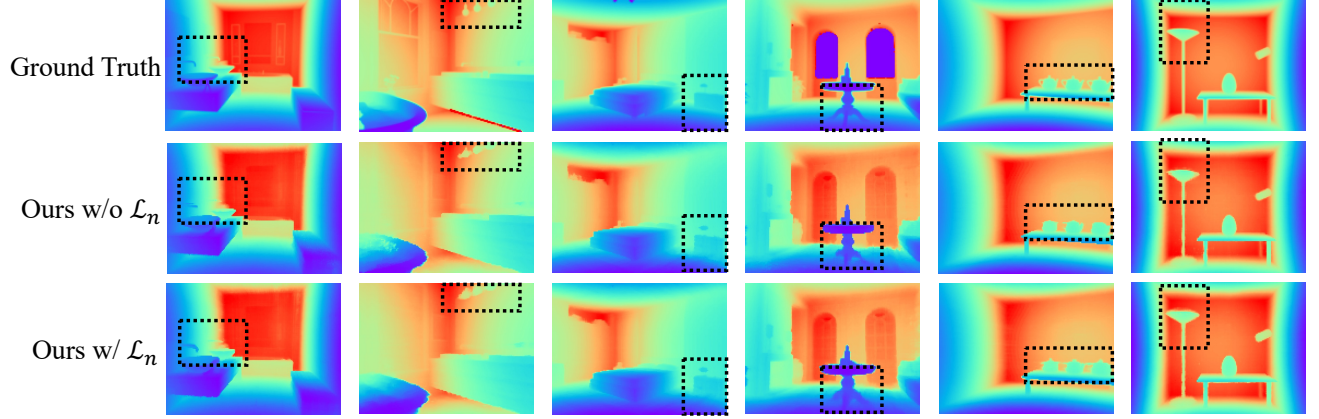


Figure 7. Visualization for ablation study on loss term \mathcal{L}_n .



Figure 8. Examples from the synthetic dataset.

13. Details of the Datasets

For the synthetic dataset, we use random selection combined with manual screening for each scene, resulting in different numbers of images for each scene as shown in Tab. 6, and various viewpoints as depicted in Fig. 8.

Table 6. Number of images used in each scenes.

Scenes	bathroom2	bedroom	living-room	bathroom	veach-bidir	veach-ajar
Images	34	41	38	40	42	56

14. Robustness Analysis Under Noisy Pose Conditions

To comprehensively evaluate the robustness of our method against pose estimation errors, we conducted experiments by introducing controlled Gaussian noise to the input camera poses. We added zero-mean Gaussian noise with vary-

ing standard deviations (STD) ranging from 0.00 to 0.30 meters to the camera positions. As shown in Table 7 the evaluation was performed on our synthetic dataset where ground-truth poses are available, allowing us to isolate the impact of pose noise on reconstruction quality.

Table 7. Reconstruction quality (MAE in meters) under different levels of pose noise. Our method maintains robust performance with STD below 0.05m, which covers typical pose estimation errors in practice.

STD	0.00	0.01	0.02	0.05	0.10	0.15	0.20	0.30
MAE	0.0468	0.0525	0.0650	0.1136	0.1922	0.2629	0.3307	0.5385

15. Comparisons with the Additional Baseline

We provide additional quantitative comparisons with the state-of-the-art 'Neural RGB-D Surface Reconstruction' method on six indoor scenes from our dataset.

Table 8. Depth reconstruction error comparison. Our method significantly outperforms Neural RGB-D across all test scenes.

Method	Bathroom2	Bathroom	Bedroom	Living-Room	Veach-bidir	Veach-ajar
RGBD	0.1616	0.0924	0.1361	0.2044	0.2830	0.1195
Ours	0.0427	0.0404	0.0337	0.0820	0.0393	0.0428

16. Ablation Study on Light Source Modeling

To validate our design choice for light representation, we conducted ablation studies comparing our model against a simplified point light source parameterization. We implemented a baseline using a single learnable scalar parameter $A_0 \in \mathbb{R}^1$ to represent constant point light intensity. As shown in Fig. 9, we tested various initialization values ranging from 0.1 to 10.0, as the optimization landscape for such simplified representations can be highly non-convex.

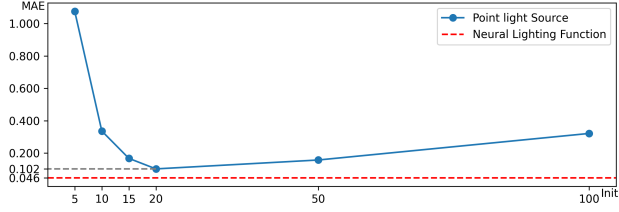


Figure 9. Reconstruction error across different initialization values for point light source representation.

17. Visualization of Phase Wrapping

We add wrapping maps shown below to visually demonstrate the significance of wrapping issues in our evaluation dataset. Fig. 10 shows iToF depth (top), our results (lower left), and wrapping periods (lower right).

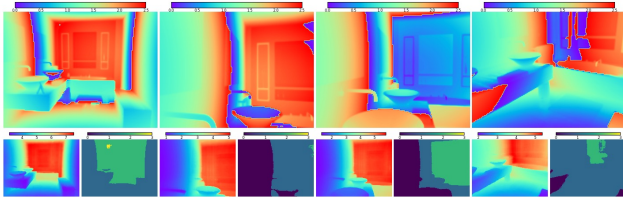


Figure 10. Visualization of phase wrapping effects.

18. Performance Analysis Under Varying SNR

Signal-to-noise ratio (SNR) is a critical factor affecting iToF depth quality, as it directly impacts the accuracy of phase measurements. We systematically evaluate our method's performance under different SNR conditions combined with three common modulation frequencies (20 MHz, 40 MHz, and 60 MHz).

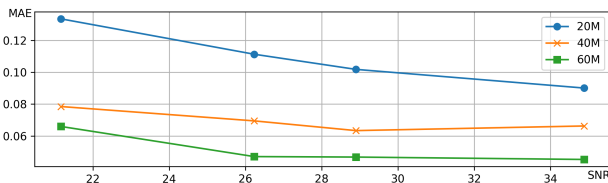


Figure 11. Depth reconstruction error (MAE) under varying SNR levels and modulation frequencies.

19. Enhanced Visualization of Multi-path Interference Effects

Multi-path interference (MPI) occurs particularly severe in indoor environments with concave structures, corners, and reflective surfaces. We provide enhanced visualizations comparing our method's handling of MPI against the baseline [3] in Fig. 12, with clear highlighting. All error maps are visualized in the same scale.

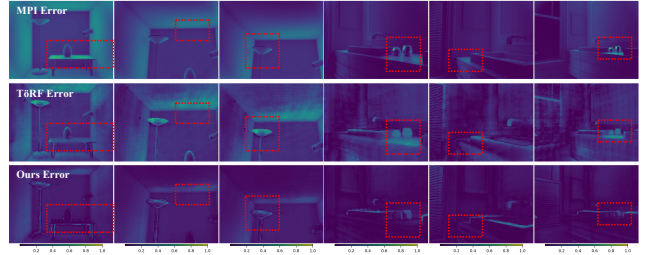


Figure 12. MPI Error Reduction Highlighted.

20. Discussion on the multi-path interference removal

In our paper, we emphasize that our work does not explicitly model the generation of multi-path interference noise. Our motivation is that the impact of multi-path interference varies across different viewpoints. By constraining a unified scene representation with multi-view imaging results, we can smooth the noise and achieve a more accurate representation. We provide some visual support for our viewpoint in Figure 13. The 'MPI depth' refers to depth measurements affected only by multi-path interference noise. The 'MPI error' represents the MAE error map calculated with respect to the ground truth. All error maps are visualized at the same scale. Red dots in the multi-path interference error map highlight the same positions in the scene where errors vary across different viewpoints. It can be seen that multi-path interference error changes with the viewpoint, being smaller when the target is closer or directly facing the camera. Therefore, we can average the multi-path interference noise by constraining a unified scene representation with multi-view imaging results. The 'Ours Error' shows the MAE error map between our method and the GT, where multi-path interference errors, particularly at floor and wall corners, are reduced.

21. Modulation Frequency

In this section, we present the ablation study on modulation frequency. Figure 14 and Table 9 display experimental results for scenes captured at 20 MHz. Figure 15 and Table 10 show results for scenes captured at 40 MHz.

22. Experiments on Real World Dataset

We present more visualizations from the real-world data captured by [3]. We manually masked the moving objects in the iToF imaging. All methods are trained on the processed dataset. Figure 16 shows RGB images for reference, which is not strictly aligned with the viewpoint of the iToF camera. Figure 17 presents visualization results on the real dataset. It displays the rendered depth maps and reconstructed meshes. Due to the lack of ground truth depth for reference, the scales of the visualized depth maps are not the same.

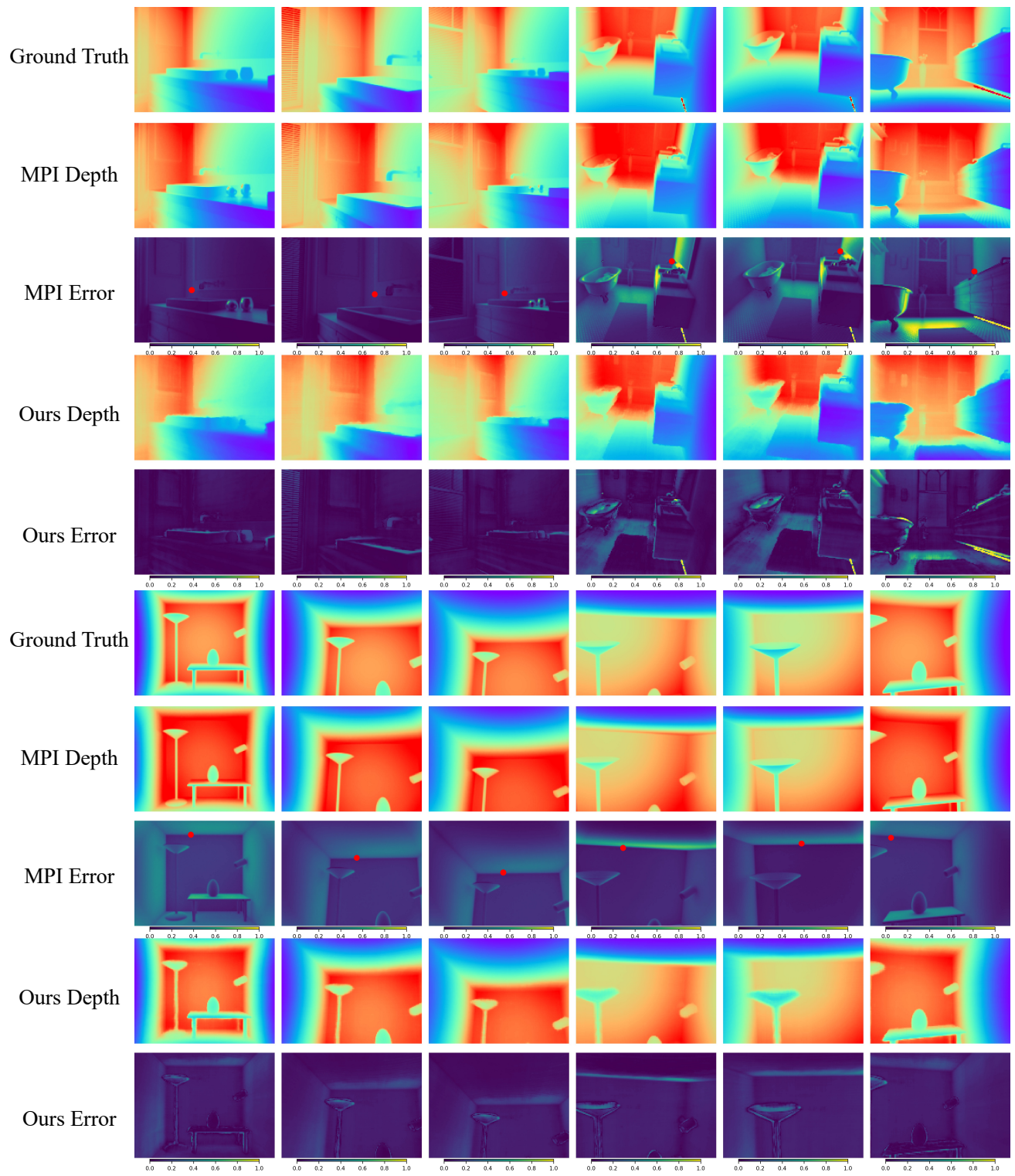


Figure 13. Visualization on the multi-path interference errors.

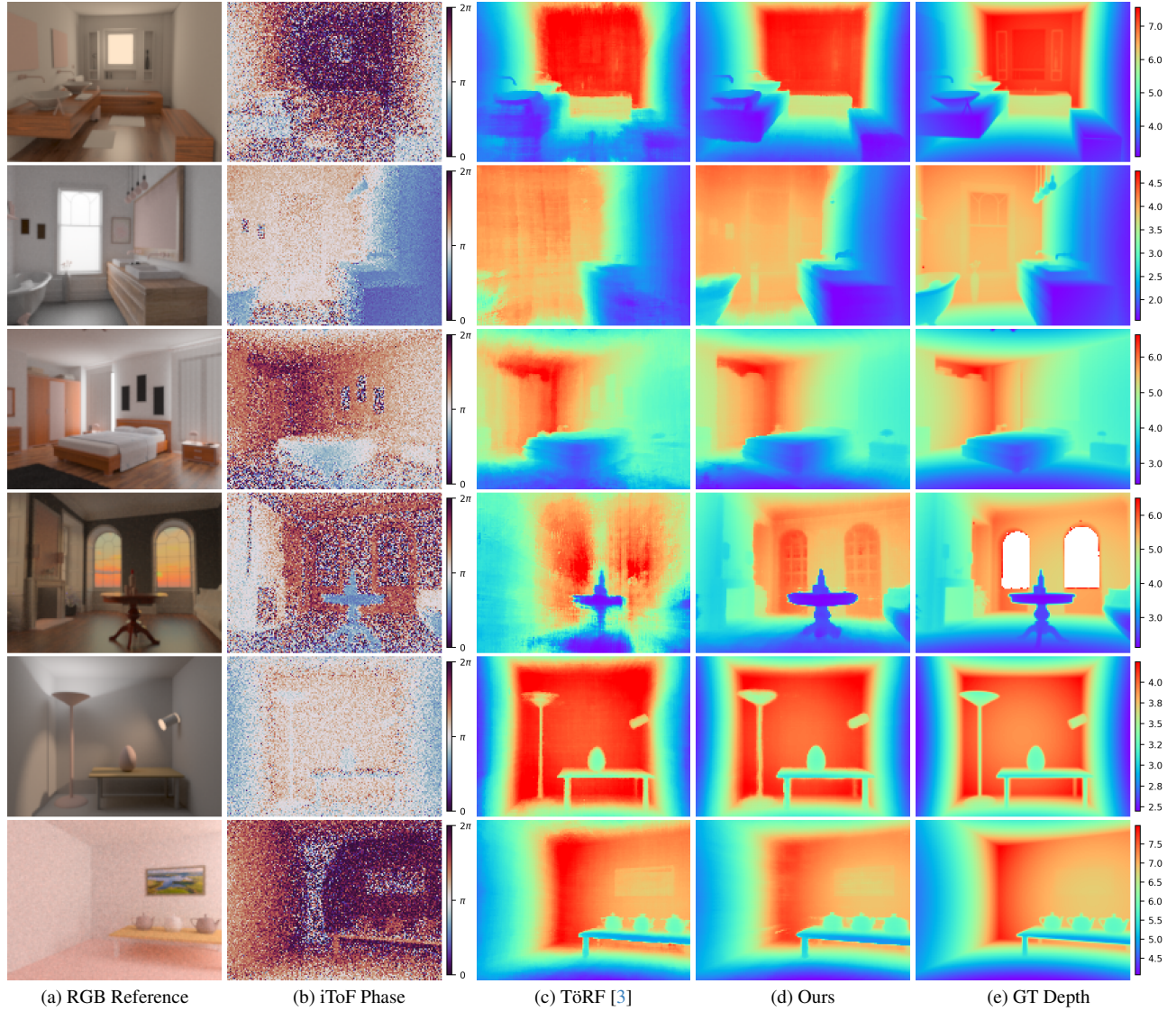


Figure 14. Qualitative comparisons of scenes captured with modulation signals at the 20 MHz frequency.

Table 9. Evaluation results on the scenes captured with the modulation frequency of 20 MHz. Our method quantitatively outperforms all prior work in all scenes.

Method		MAE ↓	RMSE ↓	δ_1 ↑		MAE ↓	RMSE ↓	δ_1 ↑		MAE ↓	RMSE ↓	δ_1 ↑
TöRF [3]	Bathroom2	0.1780	0.2420	0.9926	Bathroom	0.1690	0.2505	0.9293	Bedroom	0.1960	0.2522	0.9804
Ours		0.0937	0.1479	0.9980		0.0700	0.1337	0.9918		0.1026	0.1600	0.9979
TöRF [3]	Living-room	0.3174	0.5622	0.9017	Veach-bidir	0.1766	0.201	0.9831	Veach-ajar	0.1983	0.2533	0.9974
Ours		0.1208	0.3070	0.9845		0.1086	0.1293	0.9942		0.1148	0.1632	0.9992

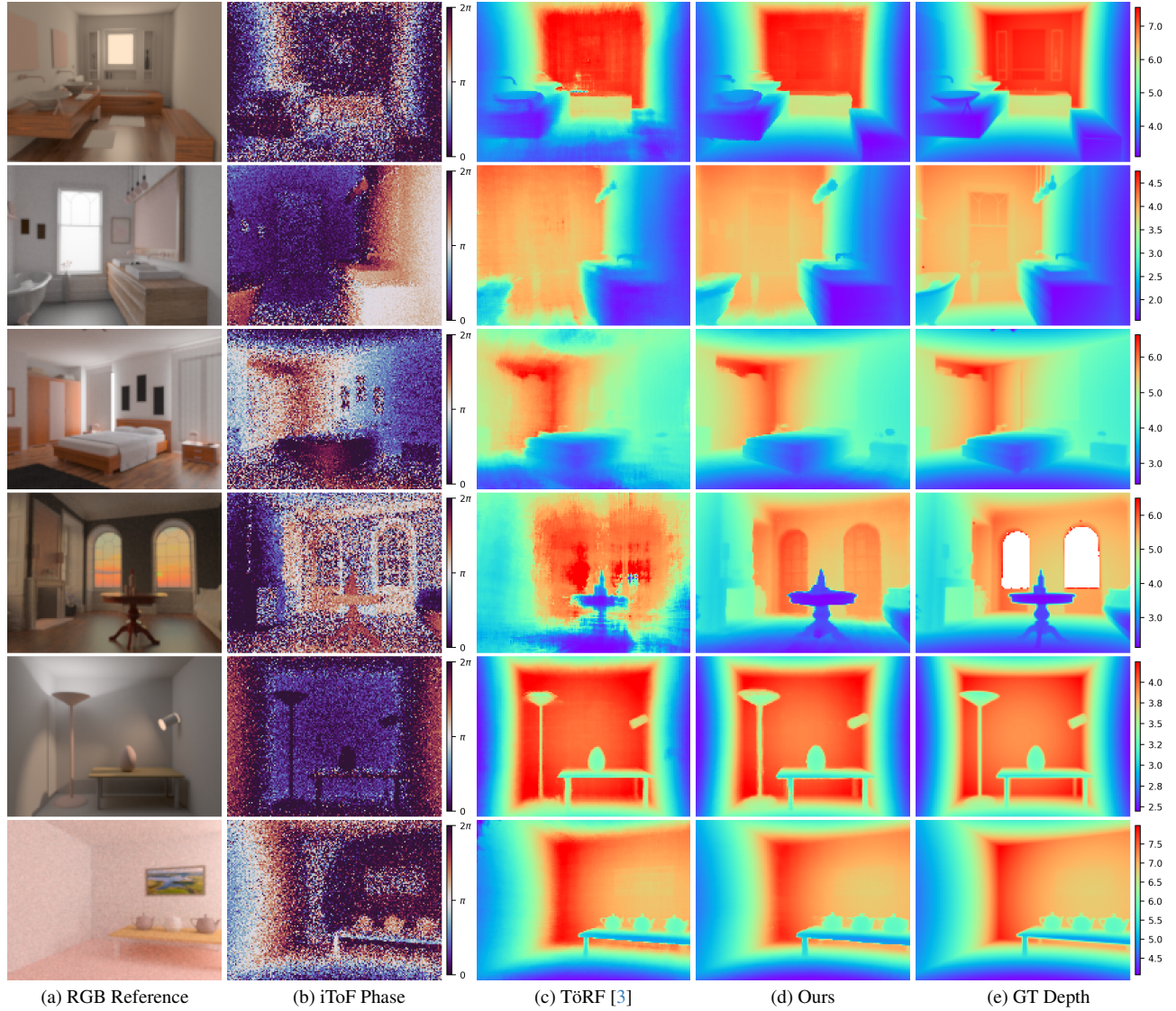


Figure 15. Qualitative comparisons of scenes captured with modulation signals at the 40 MHz frequency.

Table 10. Evaluation results on the scenes captured with the modulation frequency of 40 MHz. Our method quantitatively outperforms all prior work in all scenes.

Method		MAE ↓	RMSE ↓	δ_1 ↑		MAE ↓	RMSE ↓	δ_1 ↑		MAE ↓	RMSE ↓	δ_1 ↑
TöRF [3]	Bathroom2	0.1390	0.2780	0.9850	Bathroom	0.0872	0.1553	0.9856	Bedroom	0.1925	0.2759	0.9678
Ours		0.0585	0.1381	0.9957		0.0452	0.0949	0.9970		0.0537	0.0915	0.9986
TöRF [3]	Living-room	0.3251	0.5663	0.9167	Veach-bidir	0.1014	0.1231	0.9981	Veach-ajar	0.2403	0.3675	0.9681
Ours		0.0941	0.2880	0.9605		0.0636	0.0888	0.9976		0.0654	0.1108	0.9988



Figure 16. Reference images from real world data.

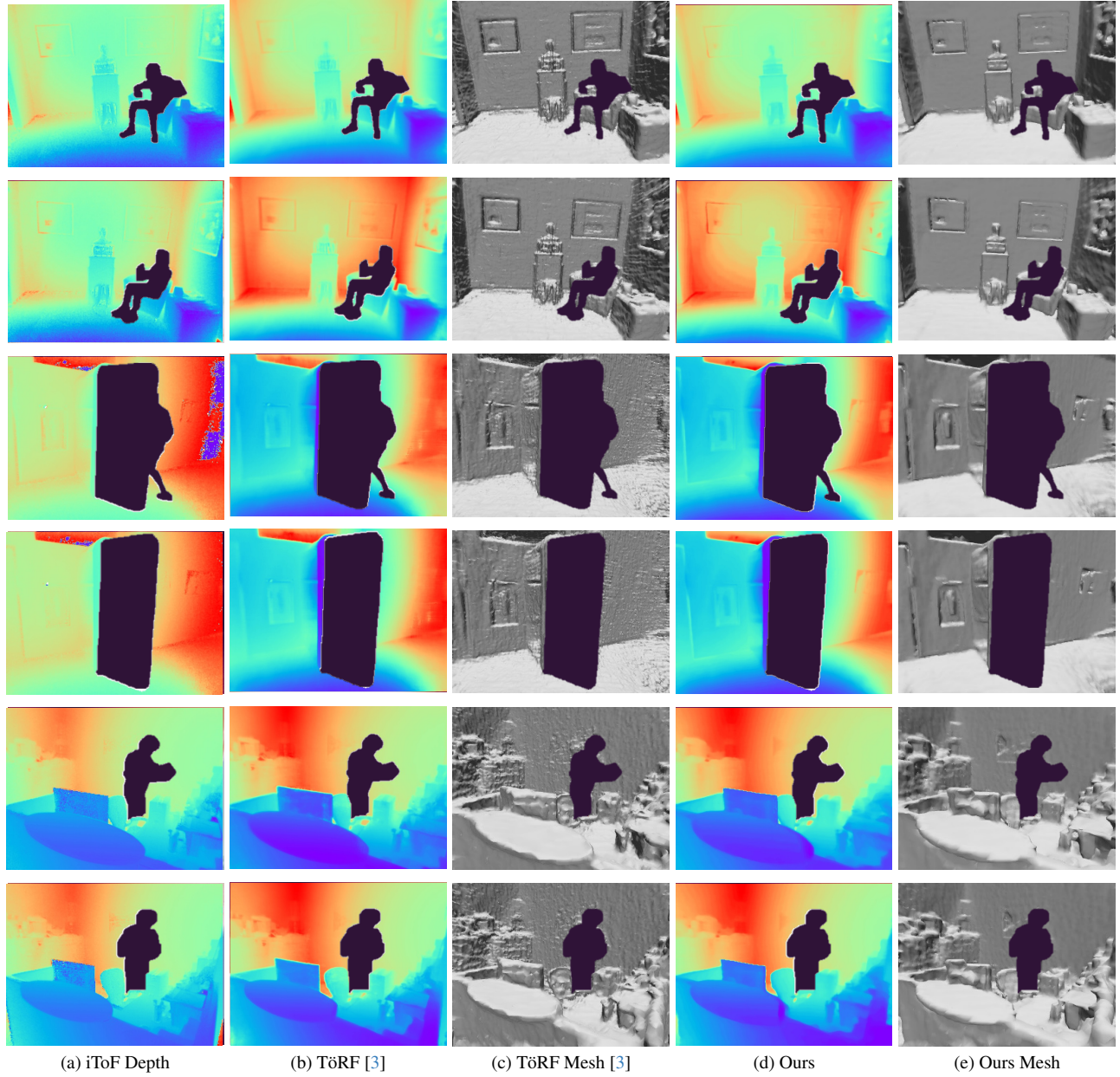


Figure 17. Qualitative comparisons on the real world data.