Learnable Burst-Encodable Time-of-Flight Imaging for High-Fidelity Long-Distance Depth Sensing

Manchao Bao[1]
manchaobao@smail.nju.edu.cn

Shengjiang Fang[1,2]
shengjiangfang@smail.nju.edu.cn

Tao Yue[1]
yuetao@nju.edu.cn

Xuemei Hu[1]
xuemeihu@nju.edu.cn

[1]Nanjing University

[2]Nanjing Electronic Devices Institute

Abstract

Long-distance depth imaging holds great promise for applications such as autonomous driving and robotics. Direct time-of-flight (dToF) imaging offers high-precision, long-distance depth sensing, yet demands ultra-short pulse light sources and high-resolution time-to-digital converters. In contrast, indirect time-of-flight (iToF) imaging often suffers from phase wrapping and low signal-to-noise ratio (SNR) as the sensing distance increases. In this paper, we introduce a novel ToF imaging paradigm, termed Burst-Encodable Time-of-Flight (BE-ToF), which facilitates high-fidelity, long-distance depth imaging. Specifically, the BE-ToF system emits light pulses in burst mode and estimates the phase delay of the reflected signal over the entire burst period, thereby effectively avoiding the phase wrapping inherent to conventional iToF systems. Moreover, to address the low SNR caused by light attenuation over increasing distances, we propose an end-to-end learnable framework that jointly optimizes the coding functions and the depth reconstruction network. A specialized double well function and first-order difference term are incorporated into the framework to ensure the hardware implementability of the coding functions. The proposed approach is rigorously validated through comprehensive simulations and real-world prototype experiments, demonstrating its effectiveness and practical applicability. The code is available at: `https://github.com/ComputationalPerceptionLab/BE-ToF`.

1 Introduction

Achieving high-precision depth imaging over long distances has remained a fundamental objective in fields such as computer vision, robotics, and autonomous systems. Time-of-flight (ToF) imaging [1, 2, 3], as a key approach to depth imaging, can be further categorized into direct ToF (dToF) and indirect ToF (iToF) based on differences in working principles. Direct ToF imaging [4] estimates depth by directly measuring the round-trip time of light, enabling high-precision and long-range sensing. Despite its advantages, this approach requires ultra-short pulsed light sources and high-resolution time-to-digital converters (TDCs), imposing stringent hardware demands that increase system complexity and cost, thereby limiting its practicality for widespread deployment. Indirect ToF systems [5, 6, 7, 8, 9], in contrast, emit amplitude-modulated continuous wave (AMCW) signals and infer depth by analyzing the phase shift between the transmitted and received signals. Due to their relatively lower hardware complexity and cost, iToF systems offer a more practical and hardware-friendly solution. Nevertheless, existing iToF technologies face significant challenges in long-range imaging, primarily due to phase wrapping [10] and low signal-to-noise ratio (SNR) resulting from optical attenuation [11]. To address the phase wrapping, dual-frequency modulation

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

techniques [12, 13] have been proposed, albeit at the cost of increased computational complexity and stricter hardware synchronization requirements. Alternative approaches have sought to mitigate phase wrapping under single-frequency modulation by incorporating scene priors [5, 14], however, these methods do not fundamentally resolve the intrinsic ambiguity introduced by periodic modulation.

In this paper, we propose a novel ToF imaging paradigm termed Burst-Encodable Time-of-Flight (BE-ToF). Our BE-ToF system operates in a low-frequency burst mode for light pulse modulation and demodulation, such that the phase of the reflected signal sweeps the entire range $[0, 2\pi]$ within a single, long burst period. This facilitates high-fidelity, long-distance depth imaging using only single frequency modulation. Moreover, considering the significant variation in SNRs caused by the light-falloff, we propose an end-to-end learnable framework that jointly optimizes the coding functions and the depth reconstruction network, thereby ensuring high-precision depth estimation. In particular, we incorporate constraints based on double well function and first-order difference to ensure the hardware implementability of the learned coding functions. We evaluate our method on a synthetic dataset and compare it with conventional iToF approaches, including single-frequency and multi-frequency modulation techniques. Finally, we built a prototype system to prove the effectiveness of our method in real-world experiments.

In general, we make the following contributions:

- We present a novel Burst-Encodable Time-of-Flight imaging system that enables high-fidelity long-distance depth sensing using only a single modulation frequency, thereby fundamentally mitigating the issue of phase wrapping inherent in traditional iToF systems.
- We propose an end-to-end learnable framework that jointly optimizes the coding functions and the depth reconstruction network to ensure high-precision depth estimation across varying distances.
- We uniquely incorporate double well function and first-order difference as loss function to ensure the hardware implementability of the learned coding functions.
- We develop a prototype of our BE-ToF system and demonstrate its superior performance on both simulated datasets and real-world scenarios.

2 Related Work

ToF imaging. ToF imaging has emerged as a prevalent and effective technique for depth acquisition. Direct ToF imaging enables long-range depth estimation by measuring the round-trip time of light pulses [15]. However, achieving high-precision depth measurements with dToF imposes stringent requirements on the pulsed light source, typically necessitating pulse widths in the nanosecond or picosecond range [4, 16, 17]. In addition, the system requires TDCs with tens-of-picoseconds time resolution and low timing jitter [18, 19]. These demanding hardware specifications present substantial challenges for practical implementation and large-scale deployment. In contrast, indirect ToF imaging leverages cost-effective CMOS sensors to deliver high-resolution depth estimation. However, iToF often faces phase ambiguity caused by phase wrapping when performing long-range depth imaging. A viable approach to address this issue is the use of multi-frequency modulation [12, 20, 21], where low frequencies are employed to extend the maximum unambiguous range, while high frequencies ensure precise depth measurements. Hanto *et al.* [22] developed a novel ToF LiDAR range finder based on dual-modulation frequency switching to achieve depth imaging with an extended range. Su *et al.* [23] propose an end-to-end time-of-flight imaging framework that enables high-quality depth reconstruction under multi-frequency modulation. However, multi-frequency modulation often results in increased hardware complexity and computational cost. In addition, various approaches such as amplitude correction [5], surface normal constraints [14] and RGB fusion [24] have been proposed to enable phase unwrapping from single-frequency measurements, which extremely rely on the scene prior. In this paper, we propose Burst-Encodable Time-of-Flight Imaging to fundamentally solve the phase wrapping issue in iToF, enabling high-fidelity, long-distance depth estimation.

End-to-end learning. End-to-end learning is a method aimed at jointly optimizing optical systems and reconstruction algorithms. Metzler *et al.* [25, 26] obtain high dynamic range (HDR) images from a single shot by jointly optimizing the optical encoder and the electronic decoder. Nie

et al. [27] leverage an end-to-end network for hyperspectral reconstruction, enabling simultaneous learning of optimized camera spectral response functions and a mapping for spectral reconstruction. For dense 3D localization microscopy, Nehme *et al.* [28] proposed a deep STORM-based method to achieve end-to-end optimization of point spread function engineering and accurate 3D localization. To extend the depth of field (EDOF), Sitzmann *et al.* [29] proposed jointly optimizing the optical system and the reconstruction algorithm's parameters to achieve achromatic EDOF imaging. Guo *et al.* [30] put forward an end-to-end network framework capable of jointly optimizing the encoding function and exposure time to improve the accuracy of fluorescence lifetime imaging. Besides, in iToF imaging, Chugunov *et al.* [31] proposed to jointly learn a microlens amplitude mask and an encoder-decoder network to reduce flying pixels in depth captures. Li *et al.* [11] put forward a Fisher-information guided framework for the joint optimization of the coding functions and the reconstruction network. Given the remarkable potential of end-to-end learning in elevating imaging performance, we propose an end-to-end learnable framework that jointly optimizes the coding functions and depth reconstruction network of our BE-ToF, ensuring high-quality depth performance across varying distances.

3 Learnable Burst-Encodable Time-of-Flight Imaging

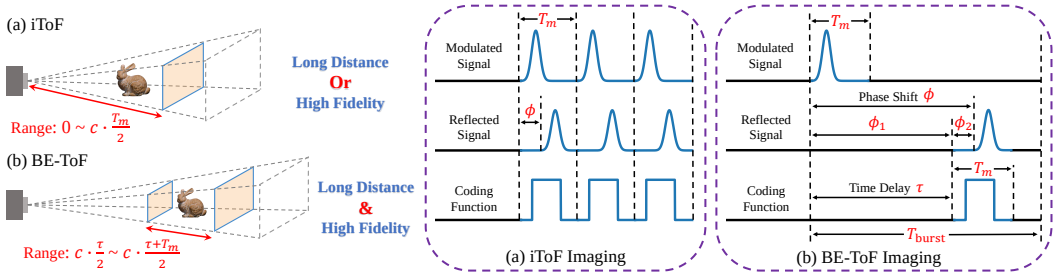

Figure 1: Comparison between iToF and BE-ToF. (a) Principle of iToF imaging, which suffers from a trade-off between sensing distance and precision; (b) Principle of BE-ToF imaging, enabling long-distance and high-fidelity depth sensing through modulation and demodulation in burst mode.

In this section, we first introduce the working principle of our BE-ToF. As shown in Fig. 1(a), conventional iToF is fundamentally constrained by a trade-off between maximum unambiguous range and depth precision, governed by the modulation period T_m. To handle this, our BE-ToF performs short-period light pulse modulation/demodulation in a low-frequency burst mode. As illustrated in Fig. 1(b), within each long burst period T_{burst}, a single modulated signal is emitted. When the reflected signal returns with a phase shift ϕ, it can be demodulated by coding functions with controllable time delay τ. Specifically, the total phase shift ϕ can be decomposed into two components: ϕ_1, which is primarily determined by the controllable time delay τ, and ϕ_2, which can be recovered using demodulation techniques like 4-step phase shift [5] or deep learning [23, 11]. In summary, the depth d can be defined as Eq. 1

$$d = \frac{c\,(\phi_1 + \phi_2)\,T_{\text{burst}}}{4\pi} = \frac{c\tau}{2} + \mathcal{D}(\phi_2)\,, \tag{1}$$

where c is the light speed and $\mathcal{D}(\phi_2)$ represents the demodulation process of ϕ_2.

Thus, in our BE-ToF system, the maximum unambiguous range d_{mur} is primarily determined by burst period T_{burst}, as defined in Eq. 2

$$d_{mur} = \frac{c}{2f_{burst}} = \frac{c \cdot T_{burst}}{2}\,. \tag{2}$$

Regarding depth precision, since we divide the phase delay ϕ into two components ϕ_1 and ϕ_2, where ϕ_1 is entirely determined by the time delay τ. Consequently, the depth error in our BE-ToF system mainly arises during the demodulation of ϕ_2. Thus, the depth error ϵ_d of our BE-ToF system can be represented as Eq. 3

$$\epsilon_d = \frac{c \cdot \epsilon_{\phi_2}}{4\pi f_m} = \frac{c \cdot \epsilon_{\phi_2} \cdot T_m}{4\pi}\,, \tag{3}$$

where ϵ_{ϕ_2} is the phase error due to several factors like photon noise, readout noise, and multi-path interference. For fixed phase error, the depth error is chiefly governed by the modulation/demodulation period T_m.

Based on the above analysis, BE-ToF substantially extends the maximum unambiguous range while maintaining the same depth precision as conventional iToF, thereby enabling long distance and high-fidelity depth imaging. Moreover, the depth sensing range of our BE-ToF spans from $\frac{c\cdot\tau}{2}$ to $\frac{c\cdot(\tau+T_m)}{2}$, which can be flexibly adjusted by tuning the time delay τ.

3.1 Differential BE-ToF Imaging Model

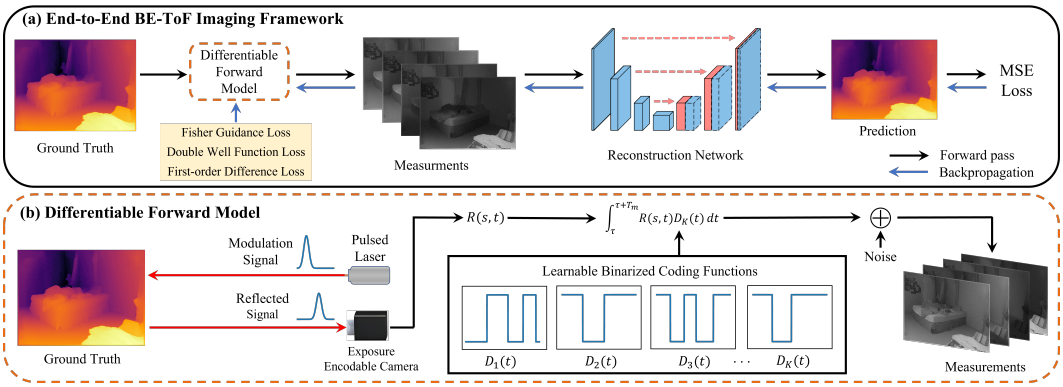

Figure 2: (a) End-to-end BE-ToF imaging framework for jointly optimizing the demodulation function and reconstruction network, (b) The differentiable physical model of the BE-ToF system.

Based on the operating principle of BE-ToF, which can be realized with a pulsed laser and an exposure-encodable camera [32], we propose an end-to-end imaging framework to ensure high-quality reconstruction across varying distances and SNRs. As shown in Fig. 2(a), the framework comprises two components: a differentiable forward model that synthesizes BE-ToF measurements in our simulation pipeline, and a reconstruction network that estimates depth from multiple measurements. By jointly optimizing the coding functions and the reconstruction network, the system delivers accurate depth reconstructions under challenging conditions.

In this section, we first establish the differentiable forward model of our BE-ToF for end-to-end optimization, as illustrated in Fig. 2(b). Assuming $M(t)$ is the modulated signal emitted by pulse laser, the reflected signal of scene point $s \in \mathbb{R}^3$ can be defined as Eq. 4

$$R(s,t) = \rho_s M(t - 2\frac{d(s)}{c}) + I_{amb}\,, \tag{4}$$

where ρ_s is the inherent reflectance of the scene point s, I_{amb} is theambient light, $d(s)$ denotes the depth value of point s. Furthermore, considering the attenuation of light intensity with distance during propagation, we incorporate the attenuation function into our model as Eq. 5

$$R(s,t) = \mathcal{F}_{d(s)}\rho_s M(t - 2\frac{d(s)}{c}) + I_{amb}\,, \tag{5}$$

where $\mathcal{F}_{d(s)}$ is the attenuation coefficient of the emitted light $M(t)$ at depth $d(s)$, which is typically inversely proportional to the square of the distance [33]. Finally, the whole BE-ToF imaging process can be formulated as Eq. 6

$$I_i(s) = \int_{\tau}^{\tau+T_m} R(s,t)D_i(t)\,dt, \quad i \in 1, ..., K\,, \tag{6}$$

where $I_i(s)$ is the measurement value of the camera, $D_i(t)$ denotes the coding functions and K denotes the number of measurements. Taking into account the inherent noise of the sensor, the final measurement can be expressed as Eq. 7

$$X_i(s) = I_i(s) + n_d + n_r, \quad n_d \sim \mathcal{P}(\mathbb{E}(n_d)),\ n_r \sim \mathcal{N}(0, \sigma_r^2)\,, \tag{7}$$

where n_d is the dark noise following the Poisson distribution with expectation $\mathbb{E}(n_d)$ and n_r is the readout noise following Gaussian distribution with standard deviation σ_r.

Considering that $X_i(s)$ contains three unknowns: ρ_s, I_{amb}, $d(s)$. Therefore, at least $K \geq 3$ measurements are required to solve for the depth $d(s)$.

3.2 Reconstruction Network

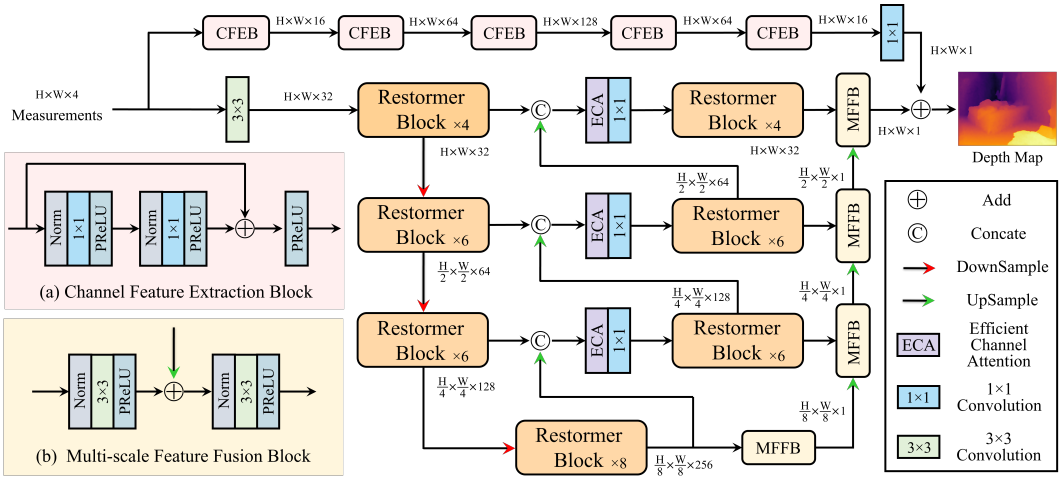

Figure 3: Architecture of the Restormer-based Spatial-Channel Fusion Network(RSCF-Net), (a) Channel Feature Extraction Block(CFEB), (b) Multi-scale Feature Fusion Block(MFFB).

With the proposed differentiable forward model, we can simulate the K measurements of the BE-ToF imaging process. To recover high-fidelity depth map from this set of measurements, we propose a Restormer-based Spatial-Channel Fusion Network(RSCF-Net). As shown in Fig. 3, our network adopts Restormer [34] as the backbone, featuring a four-level encoder-decoder structure in which each level comprises multiple Restormer blocks. In contrast to the conventional skip connections used in the original Restormer, we integrate an Efficient Channel Attention (ECA) module [35] to enhance the fusion of features between encoder and decoder branches. Furthermore, recognizing the inherent differences between depth reconstruction and the image restoration tasks for which Restormer was originally designed, we augment our network with two additional components: the Channel Feature Extraction Block (CFEB) and the Multi-scale Feature Fusion Block (MFFB). The CFEB is composed of multiple residual-connected 1×1 convolutional layers, designed to extract inter-channel relationships across multiple per-pixel measurements. On the other hand, the MFFB emphasizes spatial structure by performing preliminary depth estimation at each decoder level and progressively integrating features from multiple scales in a coarse-to-fine manner. The outputs of CFEB and MFFB are subsequently fused to produce the final high-fidelity depth map.

3.3 Loss Function

During the training process, we jointly optimize the coding functions and the reconstruction network. Given that our encodable exposure camera supports only binarized coding functions, we enforce hardware implementability by applying constraints based on a double well function and first-order difference. Additionally, Fisher information is incorporated into the loss to improve reconstruction quality, while mean squared error (MSE) is used as the objective to guide the final output. Here we give more details about these losses.

Mean Squared Error Loss. We employ MSE as the fidelity loss to supervise the predicted depth map, as defined in Eq. 8

$$\mathcal{L}_{MSE} = \sum_{s} \|d_{pre}(s) - d_{gt}(s)\|_2^2 . \tag{8}$$

Fisher Guidance Loss. The SNR is one of the key factors influencing the quality of ToF imaging. Inspired by [11], we introduce the fisher guidance loss to enhance the quality of our depth reconstruction, which can be summarized as Eq. 9

$$\mathcal{L}_{fisher} = -\sum_{s}\sum_{i=1}^{K}\left[\frac{1}{2\sigma_i^4(s)} + \frac{1}{\sigma_i^2(s)}\right]\left[\frac{\partial \mathbb{E}(I_i(s))}{\partial d}\right]^2, \tag{9}$$

where $\mathbb{E}(I_i(s))$ is the expectation of $I_i(s)$ and $\sigma_i(s) = \sqrt{\mathbb{E}(I_i(s)) + \mathbb{E}(n_d) + \sigma_r^2}$.

Double Well Function Loss. To enable the optimization of binarized coding functions within the differentiable physical model. We introduce the double well function from quantum mechanics [36], which is formulated in Eq. 10

$$f_{dw}(x) = 4(x-0.5)^4 - 2(x-0.5)^2\,. \tag{10}$$

As shown in Fig. 4, this function has two valleys at $x = 0$ and $x = 1$, thereby encouraging the coding functions to converge toward binary states during the optimization process. Therefore, our double well function loss can be defined as Eq. 11

$$\mathcal{L}_{dw} = \sum_{i=1}^{K}\sum_{j=1}^{M} f_{dw}(D_i(t_j))\,, \tag{11}$$

where M is the sampling points on each coding function.

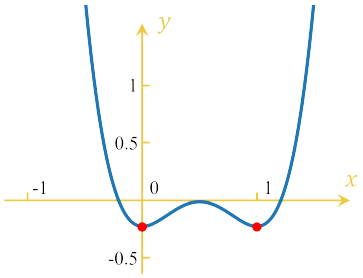

Figure 4: Demonstration of the double well function with two identical minima located at $x = 0$ and $x = 1$.

First-order Difference Loss. Although the double well function effectively constrains the coding functions to a binary state, we observe that the learned functions often exhibit extremely narrow peaks, which pose challenges for practical hardware implementation. To mitigate this issue, we introduce a first-order difference loss, as defined in Eq. 12. By minimizing the first-order difference loss, narrow peaks can be effectively suppressed, thus ensuring feasibility for hardware implementation.

$$\mathcal{L}_{1st} = \sum_{i=1}^{K}\sum_{j=1}^{M-1} |D_i(t_{j+1}) - D_i(t_j)|\,. \tag{12}$$

Finally, our complete loss can be summarized as Eq. 13

$$\mathcal{L} = \mathcal{L}_{MSE} + \gamma_1\mathcal{L}_{fisher} + \gamma_2\mathcal{L}_{dw} + \gamma_3\mathcal{L}_{1st}\,, \tag{13}$$

where γ_1, γ_2 and γ_3 are loss balance coefficients.

4 Synthetic Assessment

4.1 Implementation Details

Dataset. We use the NYU-V2 dataset [37] to train and test our end-to-end framework. The NYU-V2 dataset is a high-quality RGB-D dataset captured by Kinect with a resolution of 640×480. It contains a total of 1449 pairs of precisely aligned RGB and depth images collected from 464 indoor scenes, which enables its extensive application in academic research. For each RGB-D pair, we first apply intrinsic image decomposition [38] to the RGB image to obtain reflectance and ambient light maps. Subsequently, as detailed in Sec. 3.1, given the reflectance ρ_s, ambient light I_{amb}, and depth $d(s)$, we can synthesize multiple BE-ToF measurements. We divide the dataset in detail, using 1000 pairs of data as the training set and the remaining 449 pairs as the test set [39, 40].

Incremental Training Method. In our BE-ToF system, the SNR varies not only with distance but also significantly under the same distance due to ambient light I_{amb}. Therefore, we introduce an incremental training strategy [41] to ensure robust depth estimation of our network under varying SNR levels. Specifically, for each distance, we define three distinct SNR scenarios arranged from high to low. The network is trained with input data of varying SNRs, progressively transitioning from high to low every 10 epochs. When data of all SNRs are traversed, samples with random SNR are generated and fed to the network for the convergence of the network.

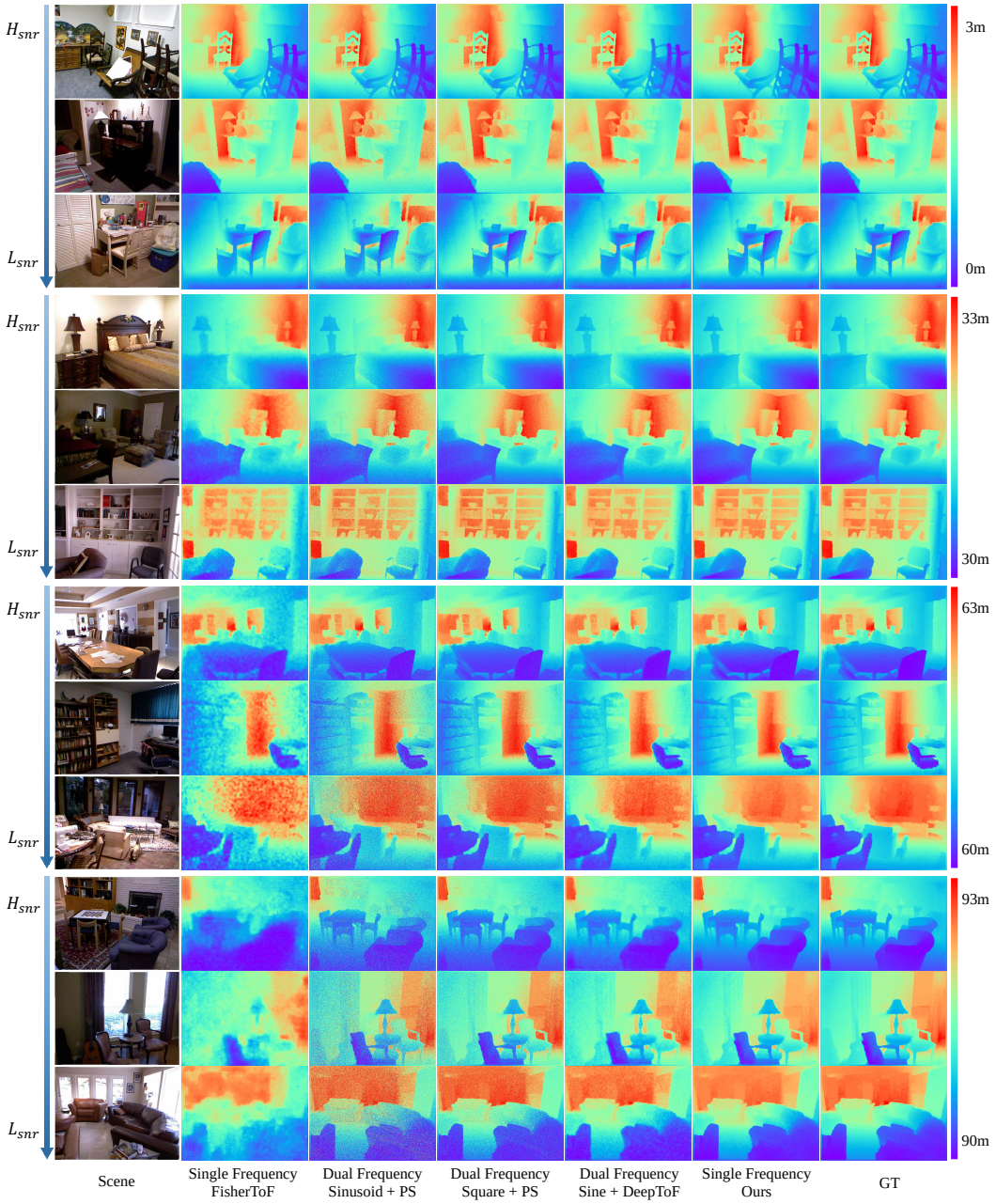

Figure 5: Overall comparisons with traditional iToF methods under various distances and SNRs, including FisherToF [11] under single frequency modulation; Sine/Square + PS algorithm [5] and Sine + DeepToF [23] under dual frequency modulation.

Training Parameters. We choose K in Eq. 6 as 4 and M in Eq. 11 as 1000. The number of restormer blocks in the network is set to $[4, 6, 6, 8]$. We train the network for 200 epochs using the ADAM optimizer [42] with a batch size of 20. The learning rate is initialized at 0.01 and decays by a factor of 0.7 every 10 epochs. The loss balance coefficients γ_1 and γ_2 are empirically set to 5e-4 and 5e-2 initially, and are updated to 5e-5 and 1 after 40 epochs. γ_3 is always set to 5. Xavier initialization is used for the learnable coding functions. All experiments are conducted on the PyTorch platform [43], using an NVIDIA GeForce RTX 4090 GPU.

Table 1: Quantitative comparison of overall performance, coding schemes, and reconstruction networks. All metrics are reported as MAE (mm).

	0-3m			30-33m			60-63m			90-93m		
	H_{snr}	M_{snr}	L_{snr}	H_{snr}	M_{snr}	L_{snr}	H_{snr}	M_{snr}	L_{snr}	H_{snr}	M_{snr}	L_{snr}
(a) Overall Performance												
FisherToF [11]	7.19	10.46	16.91	20.80	24.54	31.61	34.58	42.43	54.73	75.19	77.68	138.11
Sine+PS [5]	43.26	58.08	78.09	56.96	77.89	107.25	79.20	111.49	158.40	117.21	170.76	244.29
Square+PS [5]	33.21	40.64	51.29	40.06	51.19	66.90	51.90	69.12	93.62	72.16	100.12	140.98
DeepToF [23]	17.76	19.19	29.51	26.54	29.25	37.49	31.50	35.02	45.54	42.64	45.12	56.97
(b) Coding Scheme												
Square	12.66	15.18	21.35	16.82	22.10	29.05	20.85	24.54	32.77	25.51	30.08	44.47
(c) Reconstruction Network												
DeepToF [23]	14.76	16.20	20.32	18.25	23.30	34.12	24.95	31.50	37.90	28.12	38.17	48.55
MaskToF [31]	11.73	12.89	17.41	14.72	19.44	26.73	15.94	21.38	33.55	25.71	27.86	38.60
FisherToF [11]	6.94	8.10	16.26	10.22	15.71	21.68	14.62	18.31	29.73	22.60	24.89	31.83
Ours	**5.90**	**6.95**	**12.71**	**8.03**	**12.25**	**18.29**	**11.93**	**16.60**	**26.08**	**18.96**	**21.99**	**29.58**

4.2 Comparison with the State-of-the-art Methods

To demonstrate the superiority of our method, we conduct a detailed comparison with traditional iToF approaches, including single frequency modulation and dual frequency modulation. The scenarios encompass multiple distance ranges (0-3m, 30-33m, 60-63m, and 90-93m) combined with varying SNRs, specifically high (H_{snr} = 5.23 dB), medium(M_{snr} = 3.68 dB) and low (L_{snr} = 2.22 dB). As shown in Fig. 5, we first compare our method with FisherToF [11] under single frequency modulation. While FisherToF achieves precise depth reconstruction at close range, it still suffers from the rapid decline in imaging quality over distance. We then compare our method with a variety of dual frequency modulation approaches, including sinusoid and square coding functions with Phase Shift (PS) algorithm [5] and the learning-based DeepToF [23] method. Our method achieves the best performance across various distances and SNRs, using only single frequency modulation. We present a detailed quantitative comparison in Tab. 1 (a), with Mean Absolute Error (MAE) as the evaluation criterion.

We further substantiate the superiority of our method through an analysis of the learnable coding function and the proposed RSCF-Net. As for the coding function, considering the practical hardware implementability, we compare our learnable coding functions with the square coding function with the same RSCF-Net. The quantitative results presented in Tab. 1 (b) prove that our learnable coding functions provides superior depth reconstruction and enhanced robustness to noise. Additionally,

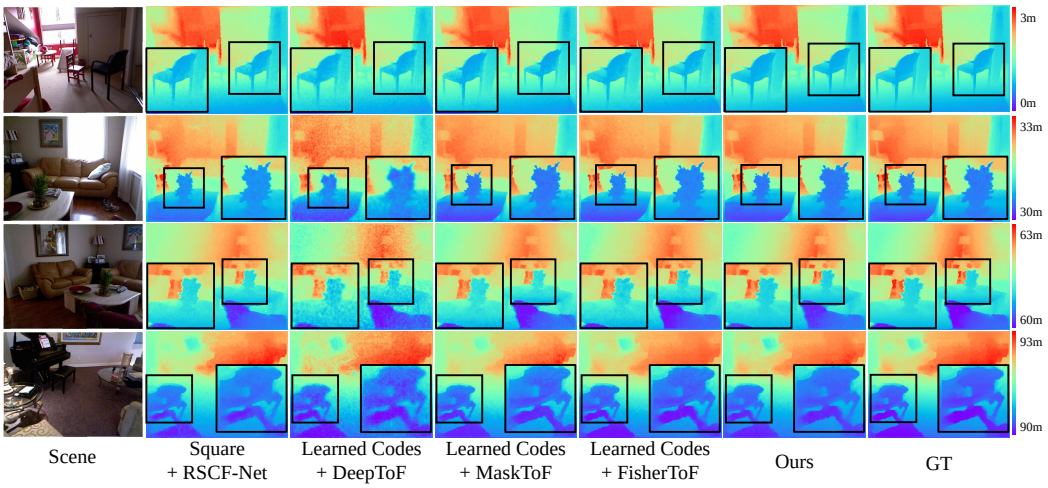

Figure 6: Comparisons with different coding scheme and reconstruction networks.

we perform a thorough comparison of our RSCF-Net with existing depth reconstruction networks with the same learned coding functions, including DeepToF [23], MaskToF [31] and FisherToF [11]. The quantitative results in Tab. 1 (c) confirm the effectiveness of our network. Fig. 6 presents the visual results of different methods across four distances under low SNR conditions, intuitively demonstrating the advantages of our approach.

4.3 Ablation Study

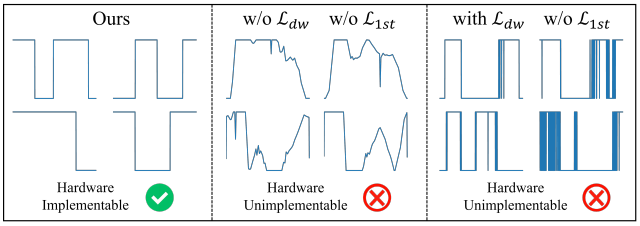

Figure 7: Visual ablations on $\mathcal{L}_{dw}$ and $\mathcal{L}_{1st}$.

Table 2: Quantitative ablations with MAE(mm) as the evaluation metric.

Distance(m)	0-3	30-33	60-63	90-93
w/o $\mathcal{L}_{fisher}$	15.36	19.41	27.02	37.45
w/o CFEB	58.69	73.64	78.56	86.94
w/o MFFB	9.8	15.77	23.36	28.40
w/o ECA	10.22	13.01	21.41	26.06
Ours	**8.52**	**12.86**	**18.20**	**23.51**

We first perform ablations on the proposed double well function loss $\mathcal{L}_{dw}$ and first-order difference loss $\mathcal{L}_{1st}$. Since the learned coding functions are primarily used to control the camera's exposure, they must be strictly binary. As illustrated in Fig. 7, the absence of $\mathcal{L}_{dw}$ and $\mathcal{L}_{1st}$ results in coding functions that are entirely impractical to implement in hardware. With only the $\mathcal{L}_{dw}$, the coding functions do converge to binary states; however, the proliferation of narrow peaks still makes them impossible to implement on real hardware.

In the next, we present a quantitative analysis to evaluate the impact of the fisher guidance loss and different network blocks. The values in Tab. 2 represent the average MAE measured under different SNRs at the same distance. The experimental results demonstrate the effectiveness of the introduced Fisher loss in guiding the network to learn an optimal coding functions. The ablation studies on different network blocks further validate the significant improvement in reconstruction quality brought by the proposed CFEB and MFFB.

5 Physical Experiment Results

Hardware Prototype. As shown in Fig. 8, to validate the effectiveness of our BE-ToF approach in real world scenarios, we built a prototype system comprising a solid-state pulsed laser and an exposure-encodable ICMOS sensor. The laser operates at 532 nm with a 5 ns pulse width, a fixed 1 kHz repetition rate, and up to 1 mJ single-pulse energy. To realize area illumination, we homogenize the beam with a diffuser and expand it using a beam expander. The ICMOS is fitted with a zoom lens (300-800 mm) and supports a minimum exposure gate of 3 ns. Timing synchronization is provided by a fast photodiode that detects each laser pulse and issues a hardware trigger to a signal generator, which then drives the ICMOS with the learned coding functions. This hardware chain achieves picosecond-scale synchronization, ensuring high-quality imaging.

Experimental Results. As shown in Fig. 9, we evaluate our approach across diverse indoor and outdoor scenarios, including a hand model and a kettle indoors and a stone model and a stair out-

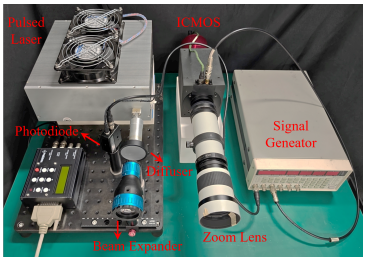

Figure 8: System Prototype.

Table 3: Quantitative results on real world experiments with MAE(cm) as the evaluation metric.

	Square + PS	Square + RSCF-Net	Learned Codes + DeepToF	Learned Codes + FisherToF	Ours
Hand	10.57	3.96	4.73	2.47	1.78
Kettle	9.13	3.51	4.14	1.99	1.41
Stone	12.16	4.76	5.94	3.78	2.50
Stair	15.79	5.92	6.70	4.26	3.83

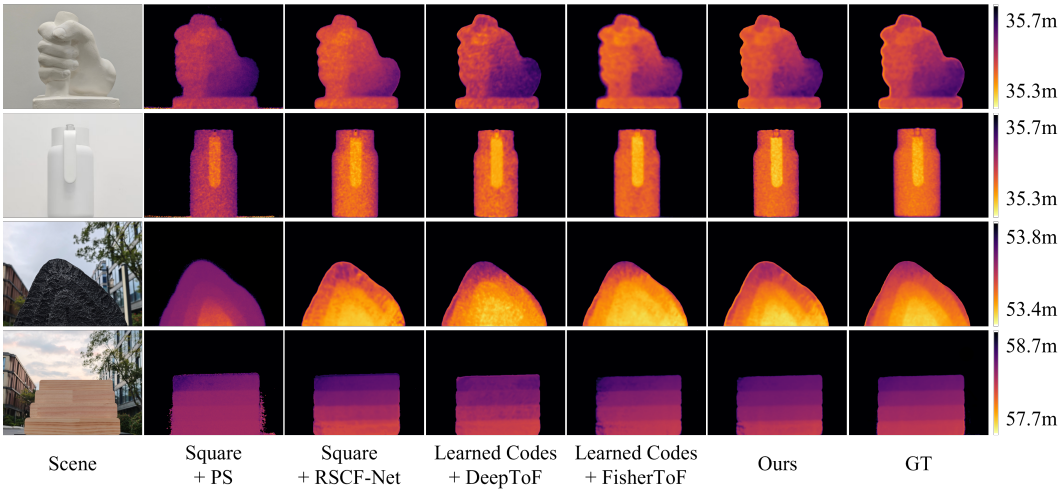

Figure 9: Real world experiments across indoor and outdoor scenarios.

doors. All experiments use the same settings as in simulation: we apply the coding functions learned in simulation and reconstruct with RSCF-Net. The modulation period T_m is fixed at 20 ns, and the burst period T_{burst} is set to 1ms, corresponding to the laser repetition rate of 1 kHz. We perform a detailed comparison against other methods, including square coding function and several reconstruction networks. Quantitative results are summarized in Tab. 3. Ground truth is acquired via a time-delay scan at the minimum exposure time (3 ns) with a 1 ns step. Both qualitative and quantitative results demonstrate that our system consistently achieves centimeter-level depth accuracy across these scenarios and outperforms existing methods.

6 Conclusion, Limitations, and Broader Impact

In conclusion, we propose a novel ToF imaging paradigm, termed BE-ToF. The BE-ToF system enables long-distance high-fidelity depth imaging by modulating and demodulating pulsed signals in burst mode using only single-frequency modulation. Additionally, we introduce a learnable end-to-end framework that jointly optimizes binarized coding functions and the reconstruction network to effectively handle varying SNRs across different distances, achieving state-of-the-art performance.

Limitations. Despite achieving both long-distance and high-fidelity depth imaging, our BE-ToF system is subject to limitations in its imaging range. As shown in Fig. 1, the operational range is confined between $\frac{c\cdot\tau}{2}$ and $\frac{c\cdot(\tau+T_m)}{2}$, with higher precision resulting in a narrower imaging range. We are currently exploring several promising directions to mitigate these limitations. First, we can exploit BE-ToFs flexible time-delay control to perform temporal scanning and synthesize a wide-range depth map. Second, because temporal scanning can incur significant latency, we favor a coarse-to-fine strategy: first capture a wide-range, low-resolution depth map, then use BE-ToF to selectively acquire high-precision depth in regions of interest (ROIs).

Broader Impact. The proposed BE-ToF system demonstrates strong potential for applications such as autonomous driving and topographic surveying, offering enhanced reconstruction quality and improved processing efficiency. However, its ability to perform long-distance depth imaging raises potential privacy concerns, particularly in scenarios where individuals may be unknowingly captured. Addressing these concerns responsibly is essential for real-world deployment.

Acknowledgments and Disclosure of Funding

This work was supported by the National Key Research and Development Program of China No. 2022YFA1207200, National Natural Science Foundation of China No. 62522113, 62505132, the Fundamental Research Funds for Central Universities No. 021014380227, 021014380260, and Open Research Project of Suzhou Laboratory No. SZLAB-1508-2024-TS015.

References

[1] Ruixuan Chen, Haowen Shu, Bitao Shen, Lin Chang, Weiqiang Xie, Wenchao Liao, Zihan Tao, John E Bowers, and Xingjun Wang. Breaking the temporal and frequency congestion of lidar by parallel chaos. *Nature Photonics*, 17(4):306–314, 2023.

[2] Ricardo Roriz, Jorge Cabral, and Tiago Gomes. Automotive lidar technology: A survey. *IEEE Transactions on Intelligent Transportation Systems*, 23(7):6282–6297, 2021.

[3] Simone Zennaro, Matteo Munaro, Simone Milani, Pietro Zanuttigh, Andrea Bernardi, Stefano Ghidoni, and Emanuele Menegatti. Performance evaluation of the 1st and 2nd generation kinect for multimedia applications. In *IEEE International Conference on Multimedia and Expo*, pages 1–6. IEEE, 2015.

[4] David T Delpy, Mark Cope, Pieter van der Zee, Simon Arridge, Susan Wray, and JS Wyatt. Estimation of optical pathlength through tissue from direct time of flight measurement. *Physics in Medicine & Biology*, 33(12):1433, 1988.

[5] Miles Hansard, Seungkyu Lee, Ouk Choi, and Radu Patrice Horaud. *Time-of-flight cameras: principles, methods and applications*. Springer Science & Business Media, 2012.

[6] Larry Li et al. Time-of-flight cameraan introduction. *Technical White Paper*, (SLOA190B), 2014.

[7] Mario Frank, Matthias Plaue, Holger Rapp, Ullrich Köthe, Bernd Jähne, and Fred A Hamprecht. Theoretical and experimental error analysis of continuous-wave time-of-flight range cameras. *Optical Engineering*, 48(1):013602–013602, 2009.

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

[9] Yu Meng, Tao Yue, and Xuemei Hu. Alignment-free 3d motion compensation for itof imaging via local linear transfer-enhanced joint optimization. *Information Fusion*, page 103636, 2025.

[10] Radu Horaud, Miles Hansard, Georgios Evangelidis, and Clément Ménier. An overview of depth cameras and range scanners based on time-of-flight technologies. *Machine vision and applications*, 27(7):1005–1020, 2016.

[11] Jiaqu Li, Tao Yue, Sijie Zhao, and Xuemei Hu. Fisher information guidance for learned time-of-flight imaging. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 16334–16343, 2022.

[12] Stephane Poujouly and Bernard Journet. A twofold modulation frequency laser range finder. *Journal of Optics A: Pure and Applied Optics*, 4(6):S356, 2002.

[13] Adrian PP Jongenelen, Donald G Bailey, Andrew D Payne, Adrian A Dorrington, and Dale A Carnegie. Analysis of errors in tof range imaging with dual-frequency modulation. *IEEE Transactions on Instrumentation and Measurement*, 60(5):1861–1868, 2011.

[14] Ryan Crabb and Roberto Manduchi. Fast single-frequency time-of-flight range imaging. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 58–65, 2015.

[15] Aongus McCarthy, Gregor G Taylor, Jorge Garcia-Armenta, Boris Korzh, Dmitry V Morozov, Andrew D Beyer, Ryan M Briggs, Jason P Allmaras, Bruce Bumble, Marco Colangelo, et al. High-resolution long-distance depth imaging lidar with ultra-low timing jitter superconducting nanowire single-photon detectors. *Optica*, 12(2):168–177, 2025.

[16] Mamadou Diop and Keith St. Lawrence. Improving the depth sensitivity of time-resolved measurements by extracting the distribution of times-of-flight. *Biomedical Optics Express*, 4(3):447–459, 2013.

[17] Shinzo Koyama, Motonori Ishii, Shigeru Saito, Masato Takemoto, Yugo Nose, Akito Inoue, Yusuke Sakata, Yuki Sugiura, Manabu Usuda, Tatsuya Kabe, et al. A 220 m-range direct time-of-flight 688$\times$ 384 cmos image sensor with sub-photon signal extraction (spse) pixels using vertical avalanche photodiodes and 6 khz light pulse counters. In *IEEE Symposium on VLSI Circuits*, pages 71–72. IEEE, 2018.

[18] Matteo Perenzoni, Daniele Perenzoni, and David Stoppa. A 64$\times$64-pixels digital silicon photomultiplier direct tof sensor with 100-mphotons/s/pixel background rejection and imaging/altimeter mode with 0.14% precision up to 6 km for spacecraft navigation and landing. *IEEE Journal of Solid-State Circuits*, 52(1):151–160, 2016.

[19] Tang Xu, Qianyu Chen, Dajing Bian, and Yue Xu. A near-infrared single-photon detector for direct time-of-flight measurement using time-to-amplitude-digital hybrid conversion method. *IEEE Transactions on Instrumentation and Measurement*, 73:1–9, 2023.

[20] David Droeschel, Dirk Holz, and Sven Behnke. Multi-frequency phase unwrapping for time-of-flight cameras. In *IEEE International Conference on Intelligent Robots and Systems*, pages 1463–1469. IEEE, 2010.

[21] 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.

[22] Dwi Hanto, Hari Pratomo, Agitta Rianaris, Andi Setiono, Sartika Sartika, Mohamad Syahadi, Eko Joni Pristianto, Dayat Kurniawan, Dwi Bayuwati, Hendra Adinanta, et al. Time of flight lidar employing dual-modulation frequencies switching for optimizing unambiguous range extension and high resolution. *IEEE Transactions on Instrumentation and Measurement*, 72:1–8, 2023.

[23] 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.

[24] 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.

[25] Christopher A Metzler, Hayato Ikoma, Yifan Peng, and Gordon Wetzstein. Deep optics for single-shot high-dynamic-range imaging. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1375–1385, 2020.

[26] Qilin Sun, Ethan Tseng, Qiang Fu, Wolfgang Heidrich, and Felix Heide. Learning rank-1 diffractive optics for single-shot high dynamic range imaging. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1386–1396, 2020.

[27] Shijie Nie, Lin Gu, Yinqiang Zheng, Antony Lam, Nobutaka Ono, and Imari Sato. Deeply learned filter response functions for hyperspectral reconstruction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4767–4776, 2018.

[28] Elias Nehme, Daniel Freedman, Racheli Gordon, Boris Ferdman, Lucien E Weiss, Onit Alalouf, Tal Naor, Reut Orange, Tomer Michaeli, and Yoav Shechtman. Deepstorm3d: dense 3d localization microscopy and psf design by deep learning. *Nature Methods*, 17(7):734–740, 2020.

[29] Vincent Sitzmann, Steven Diamond, Yifan Peng, Xiong Dun, Stephen Boyd, Wolfgang Heidrich, Felix Heide, and Gordon Wetzstein. End-to-end optimization of optics and image processing for achromatic extended depth of field and super-resolution imaging. *ACM Transactions on Graphics*, 37(4):1–13, 2018.

[30] Ziyi Guo, Jiaqu Li, Kanghui Wang, Tao Yue, and Xuemei Hu. End-to-end fluorescence lifetime imaging with optimized encoding and exposure allocation. In *IEEE International Conference on Computational Photography*, pages 1–12. IEEE, 2024.

[31] Ilya Chugunov, Seung-Hwan Baek, Qiang Fu, Wolfgang Heidrich, and Felix Heide. Mask-tof: Learning microlens masks for flying pixel correction in time-of-flight imaging. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9116–9126, 2021.

[32] Amit Agrawal and Yi Xu. Coded exposure deblurring: Optimized codes for psf estimation and invertibility. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2066–2073. IEEE, 2009.

[33] Miao Liao, Liang Wang, Ruigang Yang, and Minglun Gong. Light fall-off stereo. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, 2007.

[34] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Restormer: Efficient transformer for high-resolution image restoration. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5728–5739, 2022.

[35] Qilong Wang, Banggu Wu, Pengfei Zhu, Peihua Li, Wangmeng Zuo, and Qinghua Hu. Eca-net: Efficient channel attention for deep convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 11534–11542, 2020.

[36] V Jelic and F Marsiglio. The double-well potential in quantum mechanics: a simple, numerically exact formulation. *European Journal of Physics*, 33(6):1651, 2012.

[37] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *Proceedings of the European Conference on Computer Vision*, pages 746–760. Springer, 2012.

[38] Junho Jeon, Sunghyun Cho, Xin Tong, and Seungyong Lee. Intrinsic image decomposition using structure-texture separation and surface normals. In *Proceedings of the European Conference on Computer Vision*, pages 218–233. Springer, 2014.

[39] Beomjun Kim, Jean Ponce, and Bumsub Ham. Deformable kernel networks for joint image filtering. *International Journal of Computer Vision*, 129(2):579–600, 2021.

[40] Lingzhi He, Hongguang Zhu, Feng Li, Huihui Bai, Runmin Cong, Chunjie Zhang, Chunyu Lin, Meiqin Liu, and Yao Zhao. Towards fast and accurate real-world depth super-resolution: Benchmark dataset and baseline. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9229–9238, 2021.

[41] Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *Proceedings of the International Conference on Machine Learning*, pages 41–48, 2009.

[42] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[43] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017.

[44] Felipe Gutierrez-Barragan, Syed Azer Reza, Andreas Velten, and Mohit Gupta. Practical coding function design for time-of-flight imaging. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1566–1574, 2019.

[45] Achuta Kadambi, Refael Whyte, Ayush Bhandari, Lee Streeter, Christopher Barsi, Adrian Dorrington, and Ramesh Raskar. Coded time of flight cameras: sparse deconvolution to address multipath interference and recover time profiles. *ACM Transactions on Graphics*, 32(6):1–10, 2013.

[46] Katrin Honauer, Ole Johannsen, Daniel Kondermann, and Bastian Goldluecke. A dataset and evaluation methodology for depth estimation on 4d light fields. In *Asian Conference on Computer Vision*, pages 19–34. Springer, 2017.

[47] Shuran Song, Samuel P Lichtenberg, and Jianxiong Xiao. Sun rgb-d: A rgb-d scene understanding benchmark suite. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 567–576, 2015.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The main claims in the abstract and introduction are consistent with the contributions of this paper and align with both the simulation and real-world experiments.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discuss the limitations of our proposed method in detail in Sec. 6.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: All the theories in the paper are provided with complete and correct proofs.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We provide a detailed description of our experimental setup in Sec. 4.1, which will facilitate the reproducibility of our main experimental results by others.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We will provide our dataset and code with sufficient instructions.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so No is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: We provide a detailed description of training and test details in Sec. 4.1.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: The error bars reported in this paper are suitably and correctly defined.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)

- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Experiments compute resources are detailed descried in Sec. 4.1.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: Research conducted in the paper conforms to the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We discuss the broader impacts of the proposed method in detail in Sec. 6.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: Our paper poses no such risks.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The original authors of the code are properly credited, and the dataset used in this paper is properly cited.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: New assets introduced in the paper are well documented and accompanied by appropriate documentation.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: This paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: This paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.

A Technical Appendices and Supplementary Material

A.1 Comparison with Other Coding Functions

To fully highlight the advantages of our learnable binarized coding functions, we conduct a detailed comparison against alternative coding functions, including sinusoid, square, Hamiltonian [44], and M-sequence [45]. To ensure a fair comparison of coding functions, we employ the same imaging setup (with K = 4 measurements) and reconstruction network, varying only the coding functions. The quantitative results, evaluated using Mean Absolute Error (MAE) in millimeters, are summarized in Tab. 4. It can be observed that, under the same reconstruction network, our coding functions delivers the best performance. Notably, although we compare multiple coding functions in simulation, only the square coding function and our learnable binarized coding functions are implementable on actual hardware. The sinusoidal coding function is excluded due to its non-binary nature, while Hamiltonian codes and M-sequences contain narrow peaks that are impractical to implement given hardware constraints on minimum exposure time.

To further demonstrate the advantages of optimizing coding functions using neural networks, we use Fisher information to evaluate the quality of each coding function. As discussed in [11], Fisher information can be used as a metric to assess the optimality of different coding schemes a higher Fisher Information value indicates a more optimal coding scheme. Therefore, we list the Fisher Information of different coding functions in Tab. 5 for a straightforward comparison. It can be observed that our learnable binarized coding functions achieves the highest Fisher information, which demonstrates its optimality.

Table 4: Quantitative comparison with other coding functions.

	0-3m	30-33m	60-63m	90-93m
Sinusoid	24.67	31.12	39.80	45.39
Square	16.40	22.66	26.05	33.35
Hamiltonian [44]	11.28	14.53	21.74	27.11
M-sequence [45]	15.19	21.24	28.33	38.81
Ours	**8.52**	**12.86**	**18.20**	**23.51**

Table 5: Quantitative comparison of Fisher Information with other coding functions.

	Sinusoid	Square	Hamiltonian	M-sequence	**Ours**
Fisher Information	1.27×10^6	2.29×10^6	2.92×10^6	2.18×10^6	$\mathbf{3.93 \times 10^6}$

A.2 Experiments with Fewer Measurements

As discussed in Sec. 3.1, at least $K \geq 3$ measurements are required to recover depth. In our work, we choose $K = 4$ to ensure high reconstruction quality and robustness to noise. In Tab. 6, we present quantitative results for cases with $K \leq 4$, evaluated using Mean Absolute Error (MAE) in millimeters. When $K < 3$, the reconstruction quality significantly degrades, which is reasonable given the limited information available for depth recovery. With $K = 3$, depth can be reasonably reconstructed, though still slightly inferior to $K = 4$, where the additional measurement improves robustness to noise and other perturbations.

Table 6: Quantitative comparison with different measurements.

	0-3m	30-33m	60-63m	90-93m
$K = 1$	133.22	166.90	156.18	198.19
$K = 2$	37.84	42.94	51.96	49.18
$K = 3$	14.58	17.59	21.64	29.88
$K = 4$	8.52	12.86	18.20	23.51

A.3 Ablations on Loss Balance Coefficients

In the loss defined in Eq. 13, the MSE term is the principal objective driving accurate depth reconstruction, while the three additional regularization terms serve as auxiliary constraints that guide the learning of the binarized coding functions. To validate our choice of loss weights, we conduct a comprehensive ablation over the coefficients and report the results. All experiments are performed at distances of 0-3 m.

For coefficient γ_1 before 40 epochs, as shown in Tab. 7, setting $\gamma_1 < 5 \times 10^{-6}$ makes its effect too weak for the network to learn effective coding functions, leading to performance drop. When it exceeds 5×10^{-4}, it disrupts the double-well and first-order difference losses, resulting in coding functions unimplementable for hardware. Thus, we set γ_1 to 5×10^{-4} during the first 40 epochs.

Table 7: Ablations on γ_1 before 40 epochs.

γ_1 Before 40 Epochs	5×10^{-7}	5×10^{-6}	5×10^{-5}	5×10^{-4}	5×10^{-3}	5×10^{-2}
MAE(mm)	17.23	12.47	9.76	8.52	Hardware Unimplementable	Hardware Unimplementable

For coefficient γ_1 in epochs after 40, as shown in Tab. 8, we find that the value of γ_1 has little impact on the final performance. However, setting it too high can slow down the convergence of the network. Therefore, we set γ_1 to 5e-5 after 40 epochs to balance performance and convergence speed.

Table 8: Ablations on γ_1 after 40 epochs.

γ_1 After 40 Epochs	5×10^{-7}	5×10^{-6}	5×10^{-5}	5×10^{-4}	5×10^{-3}	5×10^{-2}
MAE(mm)	9.94	8.73	8.52	14.58	11.62	10.17
Convergence Epochs	107	103	112	123	133	137

For coefficient γ_2 in epochs before 40, we set γ_2 to a small value so that the first-order difference loss dominates and helps suppress narrow peaks. As shown in Tab. 9, when γ_2 exceeds 1, the learned coding functions exhibit narrow peaks and become unsuitable for hardware implementation. Thus, we set γ_2 to 5×10^{-2} during the first 40 epochs of training.

Table 9: Ablations on γ_2 before 40 epochs.

γ_2 Before 40 Epochs	5×10^{-4}	5×10^{-3}	5×10^{-2}	5×10^{-1}	1	5
MAE(mm)	11.98	12.35	8.52	12.67	10.27	Hardware Unimplementable

For coefficient γ_2 in epochs after 40, we increase the value of γ_2 to encourage the coding functions to converge more rapidly to a binary state. As shwon in Tab. 10, setting the coefficient below 5×10^{-2} prevents the coding functions from reaching a binary state, while values above 30 cause noticeable performance degradation. Therefore, we set γ_2 to 1 after 40 epochs.

Table 10: Ablations on γ_2 after 40 epochs.

γ_2 After 40 Epochs	5×10^{-3}	5×10^{-2}	5×10^{-1}	1	12	20	30	40
MAE(mm)	Hardware Unimplementable	10.26	9.02	8.52	9.29	11.41	10.68	17.92

For coefficient γ_3, as shown in Tab. 11, when the coefficient is too small, narrow peaks appear, making hardware unimplementable. Conversely, a large coefficient results in degraded depth recon-

struction quality. A balanced performance is achieved with values between 0.05 and 10; we set it to 5 in our experiments.

Table 11: Ablations on γ_3.

γ_3	5×10^{-4}	5×10^{-3}	5×10^{-2}	1	5	10	20	30
MAE(mm)	Hardware Unimplementable	Hardware Unimplementable	9.02	8.98	8.52	8.53	16.88	23.76

A.4 Simulation Method for the Long-Range Indoor Dataset

We use the NYU Depth V2 RGB-D dataset to train and evaluate our network. First, we scale the depth values to a fixed maximum range of 3 m (corresponding to T_m=20ns), and apply this setting consistently across all simulation experiments for fairness. As discussed in Sec. 3.1, ToF imaging is principally affected by ambient light, scene reflectance, and depth. To model these factors, we perform intrinsic image decomposition [38] on the RGB images to separate ambient light and reflectance components, and we calibrate a distance-dependent attenuation curve under long-range conditions within the simulation environment. Because ambient illumination and scene reflectance are depth-invariant in this model, variations across distance are primarily attributed to attenuation. Accordingly, given a fixed emitted signal, we apply distance-dependent attenuation coefficients to the reflected signal and then synthesize the corresponding ToF measurements. In addition, we explicitly model sensor noise as in Eq. 7, incorporating both dark current and readout noise to obtain more realistic measurements.

It is worth noting that our setting still differs from real long-range outdoor scenarios in several respects: (i) the current simulation does not model atmospheric effects, which can significantly influence ToF imaging outdoors; and (ii) beam divergence remains a key factor affecting image quality. To mitigate the latter, we employ a laser beam expander to generate area illumination, though some deviation from ideal uniform lighting persists. In future work, we plan to incorporate atmospheric effects into the simulation framework to better align with real-world experiments; meanwhile, spatial filtering is used to further improve illumination uniformity.

A.5 Precision of Binarized Coding Function Optimization

In this work, we adopt a differentiable double well function to drive the coding functions toward 0 or 1, thus ensuring that our end-to-end framework is fully differentiable. While the optimized coding functions are effectively binary, they do not attain exact binary values in floating-point arithmetic. Empirically, the learned coding functions lie extremely close to the binary extremes (e.g., around 0.0001 or 0.999), and we regard such deviations as negligible for our network. To validate our conclusion, we apply a round function during testing to convert coding functions into strict binary states(0 or 1) and compare results without it. The quantitative results in Tab. 12 show that strict binarization does not cause performance degradation; on the contrary, it slightly improves performance.

Table 12: Quantitative comparison of coding functions with/without round function with MAE(mm) as the evaluation metric.

	0-3m	30-33m	60-63m	90-93m
Without Round Function	8.52	12.86	18.20	23.51
With Round Function	8.51	12.78	17.97	21.97

A.6 Additional Results on Other Datasets

To rigorously assess the generalization capability of our method, we conduct experiments on the 4D Light Field [46] and SUN RGB-D [47] datasets, selecting 16 scenes from the former and 298 scenes from the latter as the test sets. Both datasets are processed according to the approach detailed in the main manuscript: each RGB-D pair undergoes intrinsic image decomposition to separate it into an albedo map and a shading map. Specifically, the R-channel of the albedo map is utilized as the

albedo component, while the average of its three RGB channels serves as the ambient illumination. The network, trained exclusively on the NYU-V2 [37] dataset, is then evaluated on these datasets without any fine-tuning. For comparison, we include several baseline methods: FisherToF [11] under single frequency modulation, and sinusoidal and square coding functions combined with the Phase Shift (PS) algorithm [5], as well as the learning-based DeepToF [23], all under dual-frequency modulation. The quantitative results are summarized in Tab. 13. As shown, our method exhibits strong generalization performance across both datasets and achieves the highest reconstruction accuracy among all compared approaches. Additionally, we present qualitative results on the 4D Light Field and SUN RGB-D datasets in Fig. 10 and Fig. 11, respectively. The visualizations further confirm that our method delivers robust reconstruction performance, even under challenging conditions such as long-range scenes or low signal-to-noise ratios.

Table 13: Quantitative comparison of overall performance on 4D Light Field [46] and SUN RGB-D [47] dataset.

	0-3m			30-33m			60-63m			90-93m		
	H_{snr}	M_{snr}	L_{snr}	H_{snr}	M_{snr}	L_{snr}	H_{snr}	M_{snr}	L_{snr}	H_{snr}	M_{snr}	L_{snr}
(a) Overall Performance on 4D Light Field Dataset												
Sine+PS [5]	43.39	58.20	78.10	57.07	78.01	106.93	79.37	111.08	157.91	116.95	170.50	247.06
Square+PS [5]	33.15	40.58	51.20	39.99	51.11	66.80	51.83	69.04	93.32	72.05	99.72	140.14
DeepToF [23]	19.16	23.27	40.73	34.10	36.91	45.53	38.48	42.08	52.98	54.44	55.25	66.11
FisherToF [11]	8.83	12.81	**19.19**	27.19	31.99	39.04	44.72	52.95	65.41	83.79	98.05	140.15
Ours	**6.79**	**11.06**	19.64	**11.35**	**16.29**	**26.75**	**15.11**	**21.13**	**32.25**	**26.78**	**28.78**	**34.17**
(b) Overall Performance on SUN RGB-D Dataset												
Sine+PS [5]	40.46	54.10	72.39	53.06	72.20	98.64	73.39	102.39	144.44	107.51	155.64	224.36
Square+PS [5]	31.90	38059	48.25	38.06	48.15	62.44	48.79	64.46	86.58	67.20	92.42	129.06
DeepToF [23]	16.47	23.45	30.71	25.12	32.67	36.78	27.91	35.80	40.99	36.07	45.11	55.05
FisherToF [11]	8.08	11.80	18.96	21.35	25.39	31.83	33.34	40.62	51.86	74.61	76.92	132.77
Ours	**6.63**	**7.53**	**12.11**	**9.00**	**12.55**	**18.03**	**11.59**	**16.85**	**29.20**	**17.93**	**21.62**	**31.70**

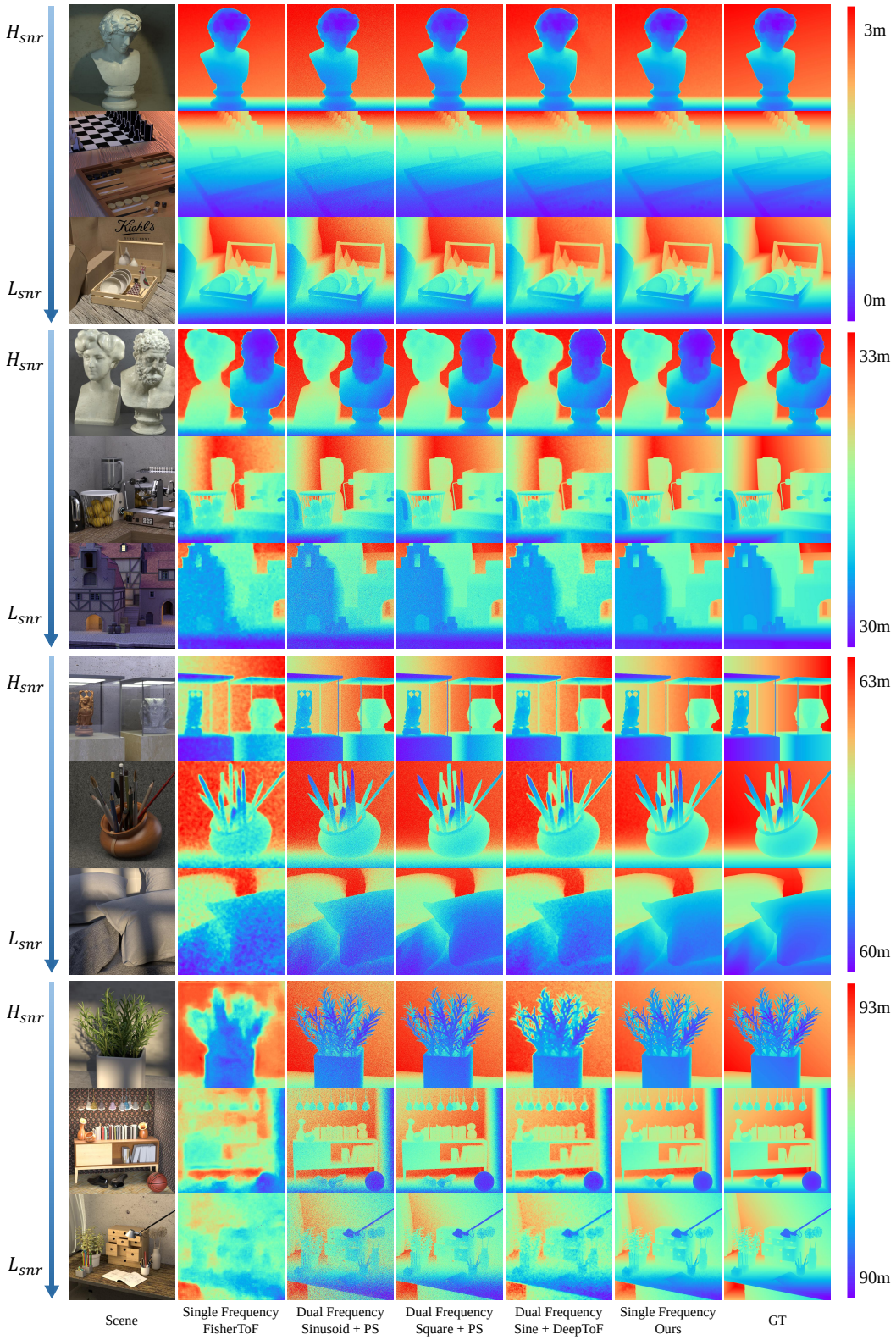

Figure 10: Overall comparisons with traditional iToF methods under various distances and SNRs on 4D Light Field dataset, including FisherToF [11] under single frequency modulation; Sine/Square + PS algorithm [5] and Sine + DeepToF [23] under dual frequency modulation.

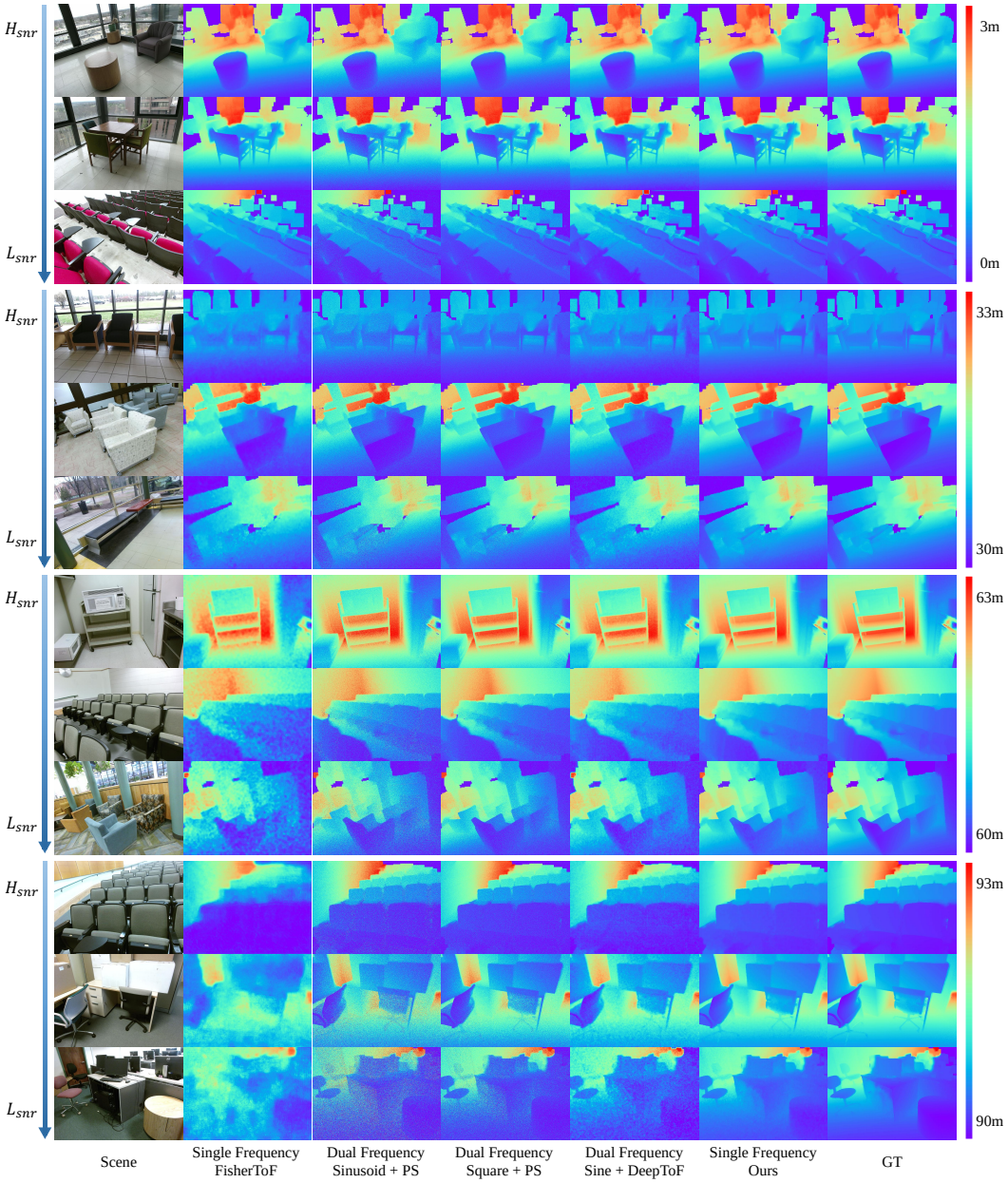

Figure 11: Overall comparisons with traditional iToF methods under various distances and SNRs on SUN RGB-D dataset, including FisherToF [11] under single frequency modulation; Sine/Square + PS algorithm [5] and Sine + DeepToF [23] under dual frequency modulation.