Fisher Information Guidance for Learned Time-of-Flight Imaging

Jiaqu Li, Tao Yue, Sijie Zhao, Xuemei Hu
School of Electronic Science and Engineering, Nanjing University, Nanjing, China
jqli@smail.nju.edu.cn, yuetao@nju.edu.cn, sjzhao@smail.nju.edu.cn, xuemeihu@nju.edu.cn

Abstract

Indirect Time-of-Flight (ToF) imaging is widely applied in practice for its superiorities on cost and spatial resolution. However, lower signal-to-noise ratio (SNR) of measurement leads to larger error in ToF imaging, especially for imaging scenes with strong ambient light or long distance. In this paper, we propose a Fisher-information guided framework to jointly optimize the coding functions (light modulation and sensor demodulation functions) and the reconstruction network of iToF imaging, with the supervision of the proposed discriminative fisher loss. By introducing the differentiable modeling of physical imaging process considering various real factors and constraints, e.g., light-falloff with distance, physical implementability of coding functions, etc., followed by a dual-branch depth reconstruction neural network, the proposed method could learn the optimal iToF imaging system in an end-to-end manner. The effectiveness of the proposed method is extensively verified with both simulations and prototype experiments.

1. Introduction

Time-of-Flight (ToF) imaging can measure the depth of scenes, and has been widely applied in autonomous driving, face recognition, 3D sensing, augmented/virtual reality, *etc*. In terms of working principle, ToF imaging can be divided into direct ToF (dToF) and indirect ToF (iToF). Unlike dToF imaging, which requires high precision pulsed light source and sensors, iToF imaging encodes the depth information in the phase of the continuously modulated light, and thus of much lower cost and higher spatial resolution in practice.

Existing iToF imaging suffers from the low signal-to-noise ratio (SNR) of measurements [13, 19], especially for the cases with strong ambient light or large distance attenuations. Adopting a higher energy light source, increasing the exposure time, or capturing more measurements could help to improve the SNR of detection, while at the expense of increasing the power consumption or the acquisition time. Recently, Su *et al*. [31] proposed a ToF reconstruction network to improve the robustness to noise by post-processing, without considering the influence of coding schemes. Gupta *et al*. [8] and Gutierrez-Barragan *et al*. [9] designed a series of Hamilton coding functions that outperforms the commonly adopted sinusoid and square coding functions. However, due to the basic assumption of small noise, their performance in low SNR scenarios is still limited.

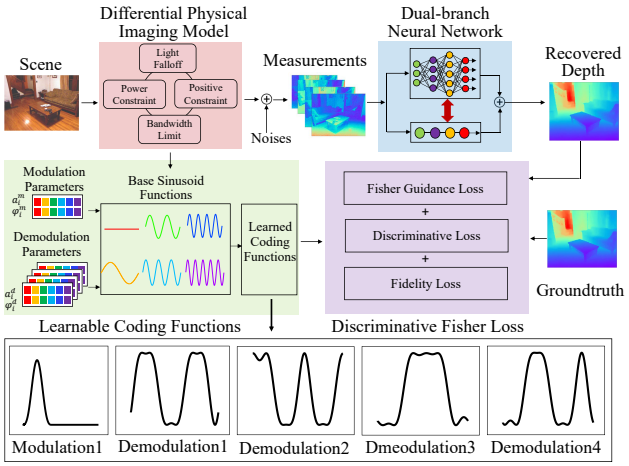

Figure 1. Overview of the proposed fisher information guided learned iToF imaging framework. We propose a differential physical imaging model with learnable coding functions and guided by the proposed discriminative fisher loss, the coding functions and dual-branch depth reconstruction neural network could be optimized simultaneously and achieve state-of-the-art performance, especially at low SNR scenarios.

In this paper, we propose an information theory guided framework to jointly optimize the coding functions and the reconstruction neural network of iToF imaging, with the proposed discriminative fisher loss, as shown in Fig. 1. Specifically, we formulate the iToF imaging process with a differential physical imaging model with learnable coding functions, taking the physical implementation constraints into consideration. Followed by the imaging model, a dual-branch depth reconstruction neural network is proposed and the proposed method could optimize the entire iToF system in an end-to-end manner. After training the iToF imaging system, we build a prototype iToF imaging system and implement the noise tolerant iToF imaging with the optimized coding functions and the reconstruction network. Through

simulation and experimental comparisons with the state-of-the-art methods, we demonstrate the superiority of the proposed learned iToF imaging.

In particular, we make the following contributions:

- We *propose* a Fisher-information guided learning framework to train the coding functions and the reconstruction neural network of the iToF imaging system in an end-to-end manner.
- We *model* the physical constraints of iToF imaging in the forward module that could learn the physical implementable modulation and demodulation functions.
- We *constraint* the coding functions with the proposed discriminative fisher loss to maximize the information about depth that could be encoded with the coding functions of iToF imaging.
- We *build* a prototype iToF imaging system with the learned optimal coding functions and verify the state-of-the-art performance of the proposed iToF imaging method, both in simulation and in real captured data.

2. Related Work

iToF imaging. Under high SNR scenarios, the conventional iToF cameras [10, 20] that adopt sinusoid or square functions as coding functions could achieve high accuracy. However, the reconstruction error with conventional iToF cameras increases as SNR decreases. Instead of improving the sensor device performance, modifying the coding functions is low-cost and easy to implement. Over the years, a variety of works have been proposed to design different coding functions for ToF imaging. Payne *et al*. [27] proposed to reduce the duty cycle of sinusoid and square functions to reduce measurement linearity error and suppress aliased harmonic components. Grootjans *et al*. [7], Kadambi *et al*. [16] proposed pseudo random binary coding sequences to reduce crosstalk, address multi-path interference. Recently, the Hamiltonian coding functions were proposed under the proposed coding space theory [8, 9], which could largely improve the depth imaging accuracy of iToF imaging. In the meantime, different corresponding ToF depth reconstruction algorithms are proposed for depth reconstruction, from the N-step phase-shift algorithm [10], the multiple frequency sinusoid algorithm [28], the zero-mean normalized cross-correlation (ZNCC) [22], to the current deep learning based reconstruction algorithms [5, 31]. While with much progress in depth imaging, the depth accuracy of existing iToF imaging schemes under low SNR is still quite limited and we propose a data-driven optimization framework based on deep neural networks that could optimize the physical-implementable ToF coding functions, especially under low SNR scenarios.

End-to-end imaging. With the development of deep learning, the end-to-end methods of jointly optimizing the imaging optics and reconstruction algorithms have gained wide attention and shown promising improvement in monocular depth estimation [3], adaptive lidar [2], high dynamic range imaging [21, 32], extended depth-of-field imaging [6, 30, 33], hyperspectral imaging [25], 3D localization microscopy [24], *etc*. In ToF imaging, Chugunov *et al*. [5] proposed to jointly learn a microlens amplitude mask pattern and encoder-decoder network to correct flying pixels in the depth map. Inspired by these works, we propose to optimize the coding functions of iToF imaging together with depth reconstruction. Through establishing a differentiable ToF imaging model with physical constraints, the modulation and demodulation functions and the CNN-based depth reconstruction algorithm can be jointly optimized to realize high depth accuracy.

Fisher information guided imaging system design. Fisher information measures the amount of information that an observable random variable χ carries about an unknown parameter θ of the distribution that models χ, so that it can be used for imaging system design. For snapshot 3D microscopic imaging with high depth accuracy, Shechtman *et al*. [29] proposed to optimize the pupil phase through maximizing the Fisher information. Chao *et al*. [4] proposed to locate the single molecule with Fisher information for microscopic imaging under low light. Wu *et al*. [34] proposed to optimize the phase mask with Fisher information, for passive single view depth estimation. Promising performance has been shown with Fisher information, which has NOT ever been explored in the field of iToF imaging.

Besides, since for iToF imaging, more than one measurements are commonly required for depth extraction, we further introduce a discrimination loss to maximize the difference between the coding functions. Through supervised by the fisher guidance loss and the discrimination loss, the coding functions can be learned to optimize the efficiency in encoding the depth information.

3. Fisher Guided Learnable iToF Framework

We propose an information theory guided optimization framework to optimize both the coding functions and the reconstruction network of a computational iToF imaging system jointly. By introducing the differentiable physical process modeling considering various real factors and constraints followed by a dual-branch neural network, the proposed method could optimize the entire iToF system in an end-to-end way.

3.1. Fisher-information Guidance

Fisher information is usually used as the metric of the amount of information included in the observation variables about the unknown parameters. Without taking the complex priors into consideration, an optimal measurement scheme should be of largest Fisher information. Therefore, it is natural to use the Fisher information as the guidance for optimizing the iToF system.

To derive the Fisher information of an iToF measurement scheme, we need to investigate the signal and noise distribution of the measurement scheme first. Commonly, the measurements of iToF system X_i on scene point p can be divided as

$$X_i(p) = S_i(p) + N_{\text{dark}} + N_{\text{readout}}, \tag{1}$$

where $S_i(p) \sim \mathcal{P}(s_i(p))$ is the depth dependent optical signal which follows the Poisson distribution, and $s_i(p)$ is the expectation of $S_i(p)$ that depends on depth, i.e. $s_i(p) = E[S_i(p)]$. $N_{\text{dark}} \sim \mathcal{P}(\lambda_{\text{d}})$ is the dark current noise with Poisson distribution, and λ_{d} is the expectation of the dark current noise. $N_{\text{readout}} \sim \mathcal{N}(0, \sigma_{\text{r}}^2)$ is Gaussian distributed readout noise, where σ_r is the standard deviation of the readout noise. In this paper, to simplify the following inference, we approximate the Poisson noise with additive Gaussian noise of vary parameters, i.e.

$$\begin{aligned} X_i(p) &\sim \mathcal{N}(\mu_i(p), \sigma_i^2(p)), \\ \mu_i(p) = s_i(p) + \lambda_{\text{d}}, \quad \sigma_i(p) &= \sqrt{s_i(p) + \lambda_{\text{d}} + \sigma_r^2}. \end{aligned} \tag{2}$$

Note that here the dark noise and readout noise are independent variables of depth $z(p)$, which only depend on the status of the detecting sensor. Typically, for each scene point in iToF imaging, multiple measurements $X(p) = [x_1(p), x_2(p), ..., x_N(p)]$ are captured with different combinations of modulation and demodulation functions. N is the number of measurement of each scene point and the probability of detecting those measurements is

$$\mathbb{P}(X(p); z(p)) = \prod_i \frac{1}{\sqrt{2\pi}\sigma_i(p)} e^{-\frac{[x_i(p) - \mu_i(p)]^2}{2\sigma_i^2(p)}}. \tag{3}$$

The Fisher information of the observed variable $X(p)$ with respect to the unknown parameter depth $z(p)$ is

$$\mathbb{I}(X(p); z(p)) = -\mathbb{E}\left[\frac{\partial^2}{\partial z^2} \log \mathbb{P}(X(p); z(p))\right]. \tag{4}$$

To derive the Fisher information, we first calculate the second derivative of the log-likelihood of Eq. 3 with the chain rule of partial derivation, i.e.

$$\begin{aligned} &\frac{\partial^2 \mathbb{P}(X(p); z(p))}{\partial z^2} \\ =&\sum_i \left\{ \frac{1}{\sigma_i^2(p)} - \frac{3\left[x_i(p) - \mu_i(p)\right]^2}{\sigma_i^4(p)} \right\} \left(\frac{\partial \sigma_i(p)}{\partial z} \right)^2 \\ &+ \left\{ \frac{\left[x_i(p) - \mu_i(p)\right]^2}{\sigma_i^3(p)} - \frac{1}{\sigma_i(p)} \right\} \frac{\partial^2 \sigma_i(p)}{\partial z^2} \\ &- \frac{1}{\sigma_i^2(p)} \left[\frac{\partial \mu_i(p)}{\partial z} \right]^2 - \frac{4\left[x_i(p) - \mu_i(p)\right]}{\sigma_i^3(p)} \frac{\partial \sigma_i}{\partial z} \frac{\partial \mu_i(p)}{\partial z} \\ &+ \frac{\left[x_i(p) - \mu_i(p)\right]}{\sigma_i^2(p)} \frac{\partial^2 \mu_i(p)}{\partial z^2} \end{aligned} \tag{5}$$

Since $x_i(p) \sim \mathcal{N}(\mu_i(p), \sigma_i^2(p))$, i.e.

$$\mathbb{E}\left\{x_i(p) - \mu_i(p); z(p)\right\} = 0, \tag{6}$$

$$\mathbb{E}\left\{[x_i(p) - \mu_i(p)]^2; z(p)\right\} = \sigma_i^2(p). \tag{7}$$

With Eqs. 4-7, the Fisher information is

$$\mathbb{I}(X(p); z(p)) = \sum_i \frac{2}{\sigma_i^2(p)} \left[\frac{\partial \sigma_i(p)}{\partial z} \right]^2 + \frac{1}{\sigma_i^2(p)} \left[\frac{\partial \mu_i(p)}{\partial z} \right]^2. \tag{8}$$

With Eq. 2, we could derive that

$$\frac{\partial \sigma_i(p)}{\partial z} = \frac{1}{2\sigma_i(p)} \frac{\partial s_i(p)}{\partial z} \quad \text{and} \quad \frac{\partial \mu_i(p)}{\partial z} = \frac{\partial s_i(p)}{\partial z}. \tag{9}$$

Finally, we could derive the Fisher information of the observed variable with respect to depth, i.e.

$$\mathbb{I}(X(p); z(p)) = \sum_i \left[\frac{1}{2\sigma_i^4(p)} + \frac{1}{\sigma_i^2(p)} \right] \left[\frac{\partial s_i(p)}{\partial z} \right]^2. \tag{10}$$

3.2. Differential Physical Modeling

The measurement function $s_i(p)$ in Eq. 10 describes the physical process of the light transmission from emitting to receiving on scene point p. In this paper, we implement this part with a differentiable physical model. Thus, the computation of discriminative fisher loss, and the induced optimization could be easily achieved. Assuming that the emitted light is modulated by a modulation function $M_i(t)$, the reflected signal of scene point p is,

$$R_i(p, t) = \alpha(p) M_i(t - \varphi(p)) + \beta\alpha(p), \tag{11}$$

where $\alpha(p)$ is the amplitude coefficient due to the reflection of the scene, β is the ambient component due to other light sources. $\varphi(p)$ is the distance dependent time delay of light propagation, i.e., $\varphi(p) = 2\frac{z(p)}{c}$, where c denotes the speed of light. Considering the light intensity falls linearly with the inverse of the square of distance, here we further formulate the light fall-off $F_{\text{falloff}}(z)$ in the measurement,

$$R_i(p, t) = F_{\text{falloff}}(z(p)) \alpha(p) M_i(t - 2\frac{z(p)}{c}) + \beta\alpha(p), \tag{12}$$

where z is the depth of light propagation.

When the reflected signal $R_i(p, t)$ reaches the sensor, the sensor demodulates $R_i(p, t)$ with the demodulation function $D_i(t)$ to derive the measurements,

$$s_i(p) = \int_0^T R_i(p, t) D_i(t) dt, \tag{13}$$

where $s_i(p)$ contains three unknowns: $\alpha(p)$, β, $z(p)$, thus commonly, it takes $N \geq 3$ measurements to reconstruct the depth $z(p)$.

Frequency decomposition for implementable bandwidth limitation. In iToF imaging devices, the bandwidth of the modulation and demodulation functions are limited by the minimum bandwidth of the signal generator, the detector, and the multiplier. To incorporate the bandwidth constraints, the modulation/demodulation functions in Eq. 13 are formulated with the summation of a set of sinusoidal waves with different frequency, below the bandwidth,

$$\begin{aligned} M_i(t) &= \sum_{i=1}^{n} a_i^m \sin(2i\pi f_0 t + \varphi_i^m) + b_m, \\ D_i(t) &= \sum_{i=1}^{n} a_i^d \sin(2\pi i f_0 t + \varphi_i^d) + b_d, \end{aligned} \tag{14}$$

where f_0 denotes the fundamental frequency of the system, $f_{\max} = nf_0$ is the bandwidth limit of the signal, a_i^m, a_i^d, φ_i^m and φ_i^d are the corresponding amplitude and phase correspond to each frequency component f_i, i.e. if_0. b_m and b_d are the corresponding DC components of the modulation function $M_i(t)$ and demodulation function $D_i(t)$. With this formulation, we could optimize the modulation and demodulation function under the constraint of system bandwidth.

Normalization base physical constraints. In theory, the higher the power of the light source, the higher the SNR of measurement. While in practice, light sources of common iToF cameras are of limited output optical power in consideration of both human eye safety standard and power consumption constraint. In order to design the modulation function of the driving light source, we add the constraint of the maximum optical power E,

$$\frac{\int_0^\tau M_i(t)dt}{\tau} \leq E, \tag{15}$$

where τ is the period of $M_i(t)$ and $D_i(t)$. Besides, $M_i(t)$ and $D_i(t)$ are required to be non-negative,

$$\begin{aligned} 0 &\leq M_i(t), \\ 0 &\leq D_i(t) \leq 1. \end{aligned} \tag{16}$$

Note that here, without loss of generality, we further constraint the amplitude of the demodulation function to be smaller than 1. This constraint could be implemented through adding normalization in the forward model before calculation of Eq. 13, i.e. $D_i(t) = \frac{D_i(t) - \min\{D_i(t)\}}{\max\{D_i(t)\} - \min\{D_i(t)\}}$.

Through setting a_i^m, a_i^b, φ_i^m and φ_i^b in Eq. 14 as learnable parameters, and consistent with the constraints in Eq. 15 and Eq. 16, the modulation and demodulation function can be optimized that meets the above physical constraints.

3.3. Depth Reconstruction Network

With the iToF imaging scheme, we could capture a set of measurements $I = [I_1, I_2, ..., I_N]^T (N \geq 3)$. To reconstruct the depth from the measurements, we propose a dual-branch multi-scale depth reconstruction network. As shown in Fig. 2, the network is mainly composed of two branches, one is composed mainly with a pyramid multi-scale neural network that could incorporate the spatial structure into the reconstruction of depth, and the other branch is composed of the pixel-based depth extraction neural network.

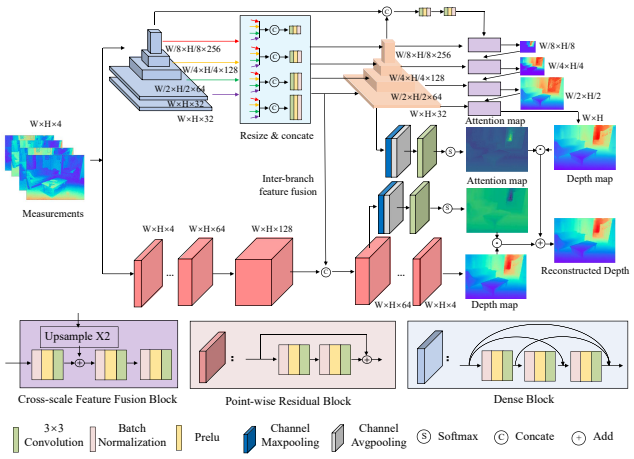

Figure 2. Dual-branch depth reconstruction network.

Spatial-structure Extraction Branch (SEB). To reconstruct the depth with the sparsity prior of the depth map, we construct a feature pyramid network to incorporate the multi-scale structural information in the space domain. Dense block is utilized as the feature encoding and decoding blocks. As for the encoder, we utilized three feature extraction blocks and the extracted multi-scale features are fused to obtain features of different scales. The decoder is constructed in a coarse-to-fine manner. First, we design the decoder to predict the smallest-scale depth map, then utilize bilinear interpolation for up-sampling, and integrate with higher-scale fusion features to generate an enlarged resolution depth map. This process is repeated to retrieve the final depth map. In this way, the small-scale depth map contains global information, and is continuously refined to finally generate more detailed large-scale depth maps.

Pixel-wise Depth Regression Branch (PDRB). The lower branch uses residual block structure with a convolution kernel of 1×1. Through the single-point direct mapping, the obtained depth map can retain the detailed features of the scene. The pixel-wise depth regression branch functions as a simulator of conventional depth extraction method of iToF imaging, i.e. extract the depth of each pixel from the corresponding measurements and guarantee the fast convergence of the learning of the coding functions. Note that, as shown in Fig. 2, for sufficient information fusion between the two branches, we introduce feature concatenation from the SEB branch to the PDRB.

The depth maps reconstructed with the two branches are then weighted by the corresponding attention map and added together to reconstruct the depth of the scene.

3.4. Loss Functions

Both the modulation/demodulation functions and the reconstruction network are learned by training with the discriminative fisher loss composed of the fisher guidance loss, fidelity loss and a multi-measurements discriminative function loss. Here we give more details about these losses.

Fisher Guidance Loss. Given the Fisher information of meansurements $s_i(p)$ on depth $z(p)$, we can derive a Fisher Guidance loss as

$$L_{\text{fisher}} = -\sum_{p} \mathcal{I}_{\text{fisher}}(p). \tag{17}$$

Fidelity Loss. To enforce the depth reconstruction fidelity, we proposed to minimize the mean absolute error (MAE) between the depth of the SEB branch z_{SEB}, the single-point mapping branch z_{PRDB}, the depth of the addition of the two branches z_* and the target depth z_{gt}, i.e.

$$\begin{aligned} L_{\text{Fidelity}} &= \frac{1}{N}\sum_{p} \|z_{\text{SEB}}(p) - z_{\text{gt}}(p)\|_1 \\ &+ \|z_{\text{PRDB}}(p) - z_{\text{gt}}(p)\|_1 + \|z_*(p) - z_{\text{gt}}(p)\|_1 \end{aligned} \tag{18}$$

Multi-measurement discriminative loss. To further improve the efficiency of different coding functions, and enforce the difference between different coding functions, here we adopt the Manhattan distance [14] as the discriminative loss, i.e.

$$L_{\text{dis}} = -\sum_{p} \sum_{0\le i\le N, 0\le j\le N} \|s_i(p) - s_j(p)\|_1. \tag{19}$$

The proposed discriminative fisher loss is constructed as the weighted combination of the fisher guidance loss, fidelity loss and discriminative loss, i.e.

$$L_{\text{DFI}} = L_{\text{Fidelity}} + \lambda_1 L_{\text{fisher}} + \lambda_2 L_{\text{dis}}. \tag{20}$$

4. Synthetic Assessment.

4.1. Implementation Details

Dataset. The train and test data used in our end-to-end network is the NYU-V2 dataset [23]. This dataset is composed of video sequences from 464 indoor scenes, containing 1449 densely labeled pairs of aligned RGB and depth images. The RGB of the scene corresponding to the depth map, can be utilized to generate the albedo map of the scene after intrinsic decomposition [15]. Without loss of generality, we choose the R channel of RGB image after intrinsic decomposition as the albedo $\alpha(p)$ in Eq. 11. Similar to [12, 17], we sample 1000 pairs of RGB-D images for training and the remaining 449 pairs for testing. We set the bandwidth limit according to the practical bandwidth that is commonly adopted, specifically, we set 50 MHz as the fundamental frequency and 250 MHz as the bandwidth limit.

Incremental Training Method. In addition to the readout noise and dark noise of the image sensor, which we assume as constant and choose to be 20 electrons in our experiment, the SNR of iToF imaging is mainly determined by the power of the incident light source, i.e. E, and the power of the ambient light β. In order to train the proposed network to realize depth reconstruction over different noise levels, we calibrate three typical noise scenarios with the setting of E and β as $(20000, 6000)$, $(14000, 6000)$, $(10000, 6000)$, that could simulate physical experiments and cover a wide range of large noise level, from low to high. To train the proposed network in dealing with different noise levels simultaneously, we adopt the incremental training strategy [1]. The network is trained with input data of different noise levels, from small to large, every other 10 epochs. When data of all noise levels are traversed, samples with random noise levels are generated and fed to the network for the convergence of the network.

We adopt ADAM [18] as the optimizer, with an initial learning rate of 0.01. The learning rate is linearly decayed with a ratio of 0.7 every 10 epochs. λ_1 and λ_2 were chosen empirically to be 1e-4 and 1e-5 initially and decayed to 1e-5 and 1e-6 after 10 epochs. Kaiming initialization [11] is utilized for the learnable modulation and demodulation functions. We implement the experiments on PyTorch [26] platform with an NVIDIA GeForce RTX 2080 GPU.

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

To demonstrate the performance superiority, we compare our method with the existing iToF imaging methods, including the conventional sinusoid and square coding functions with conventional phase shift (PS) algorithm [10], the DeepToF method [31] and the recently proposed Hamilton iToF imaging method [9]. Without loss of generality, our method uses one modulation function and four demodulation functions. Three different noisy scenarios are shown in Fig. 3, from low noise level to high, and our method performs the best. The qualitative results of different methods are also shown in Tab. 1(a), and our method enables to reconstruct depth of the lowest mean squared error (MAE), across different noise levels.

To further demonstrate the superiority of the learned coding functions, we compare our method with the other coding functions, including sinusoid, square, dual frequency sinusoid (50 MHz and 200 MHz), practical Hamilton functions [9], with the same reconstruction neural network, i.e. the proposed depth reconstruction neural network. The quantitative and qualitative results are shown in Fig. 4 and Tab. 1(b). As can be seen, the learned coding function could realize the best performance and the improvement is large especially at high noise levels, further demonstrating the noise robustness of the learned coding functions.

To demonstrate the efficiency of the proposed dual-

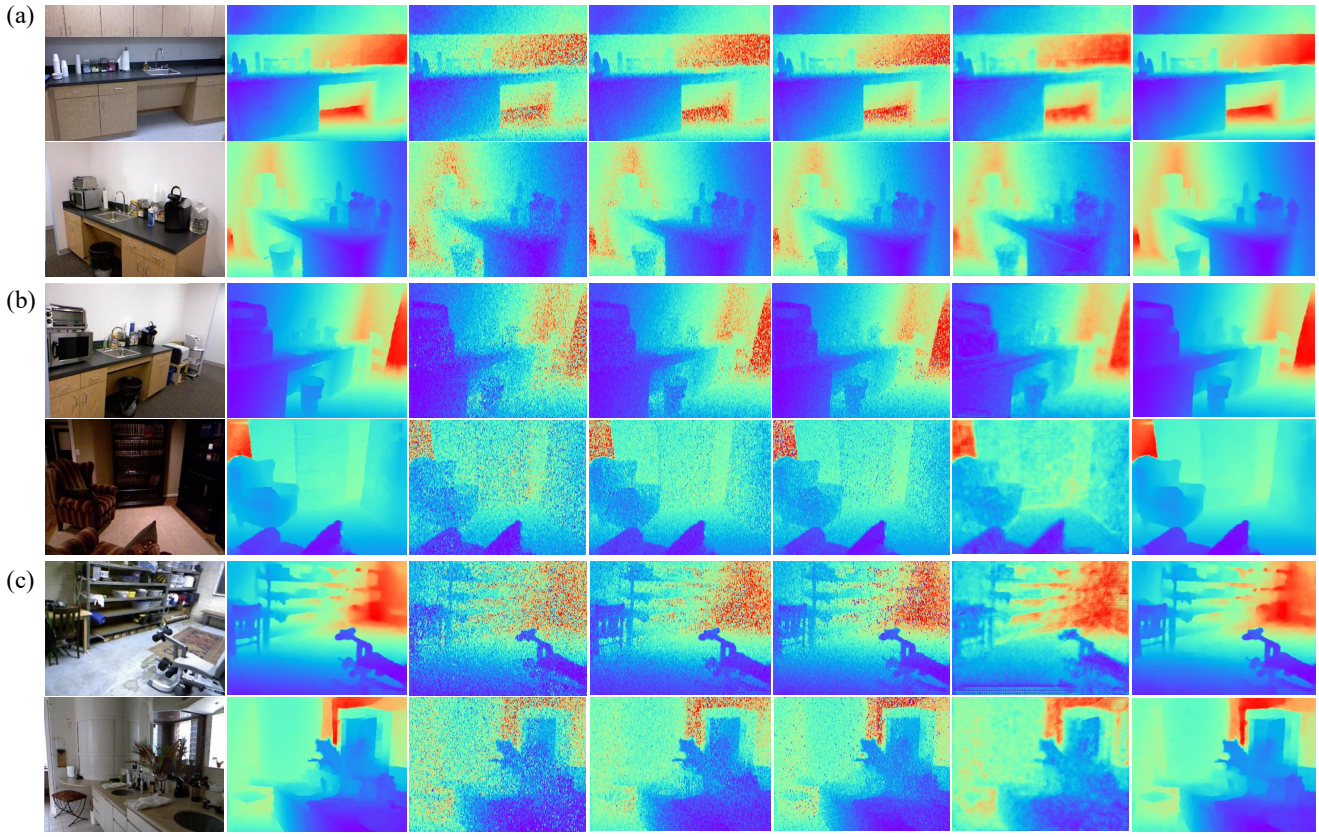

Figure 3. Overall comparisons with other iToF methods, i.e. Sine/Square + PS algorithm [10], Hamilton [9] and Sine + DeepToF [31].

branch neural network with the existing deep learning based ToF reconstruction network, i.e. DeepToF [31], and MaskToF [5]. We utilize the same learned coding functions and retrain these neural networks, the performance comparisons are shown in Fig. 5 and Tab. 1(c). Through comparison, we could demonstrate the effectiveness of the proposed dual-branch depth reconstruction neural network.

4.3. Ablation Study

Reconstruction Network. We designed ablation experiments to evaluate the proposed reconstruction network, as shown in Tab. 2. Firstly, we test the effectiveness of the proposed PDRB, which could help to improve the MAE performance in addition to the utilization of spatial structure with SEB. Then we further demonstrate the effectiveness of the proposed parallel structure with SEB and PDRB, through training and testing upon a cascaded network with PDRB and SEB, i.e. PDRB + SEB. Through comparing with the method that only uses SEB or cascaded structure, the depth reconstruction effectiveness of the proposed network is demonstrated.

Discriminative Fisher Loss. In addition to the fidelity loss, we propose the fisher guidance loss, in combination with the discriminative loss, which could greatly encourage the convergence in finding the optimal coding function and enables high quality depth precision. As shown in Tab. 2, with the combination of L_{fisher} and L_{dis}, our method obtains the highest depth accuracy. We further compare

(a) Overall Performance	MAE (mm)		
Sinusoid + PS [10]	181.192	244.679	314.793
Square + PS [10]	112.828	158.641	213.963
Hamilton [9]	97.265	148.463	208.339
DeepToF [31]	42.124	59.798	98.230
(b) Coding Function	MAE (mm)		
Sinusoid	25.024	43.438	85.957
Dual-freq Sinusoid	13.679	17.710	41.038
Square	23.231	40.731	73.994
Hamiltonian [9]	14.413	17.820	30.121
(c) Recovery Method	MAE (mm)		
MaskToF [5]	22.474	26.068	32.993
DeepToF [31]	23.858	34.076	79.862
Our method	12.857	13.763	18.264

Table 1. Quantitative comparison in terms of the overall performance, coding functions, and reconstruction methods with respect to three different noise settings from the 2nd to 4th column, i.e. (E, β) equals (20000,6000), (14000,6000) and (10000,6000).

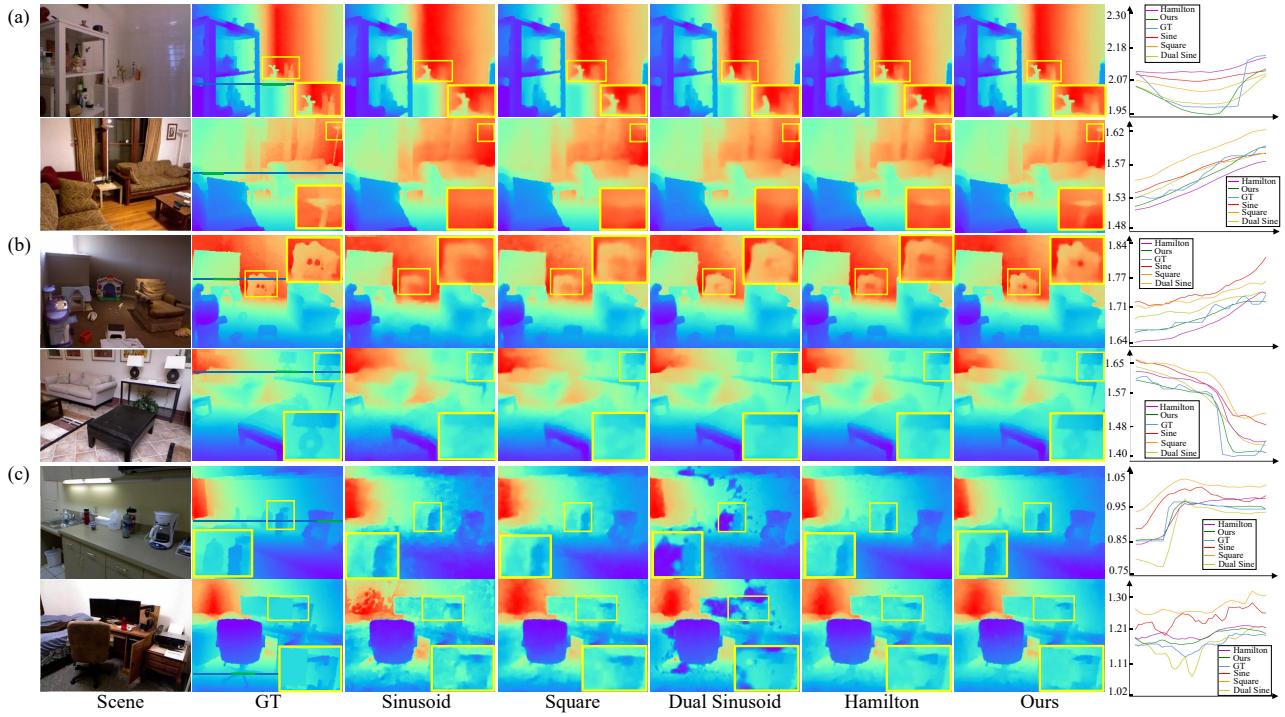

Figure 4. Depth imaging performance comparisons with different coding functions with the proposed dual-branch depth reconstruction network. The rightmost column shows the reconstructed depth value along the green line marked in the ground truth depth image.

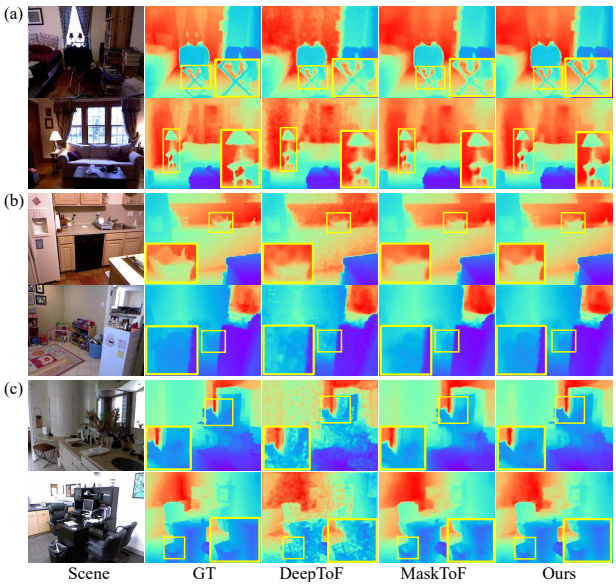

Figure 5. Comparison with other depth reconstruction networks, i.e. DeepToF [31] and MaskToF [5]

the statistical distribution of depth error of the proposed method with and without the two proposed loss functions, i.e. L_{fisher} and L_{dis}. As in Fig. 6, with only the fidelity loss, the depth error is much higher than the proposed loss especially at higher depth range, further demonstrating the effectiveness of the proposed fisher guidance.

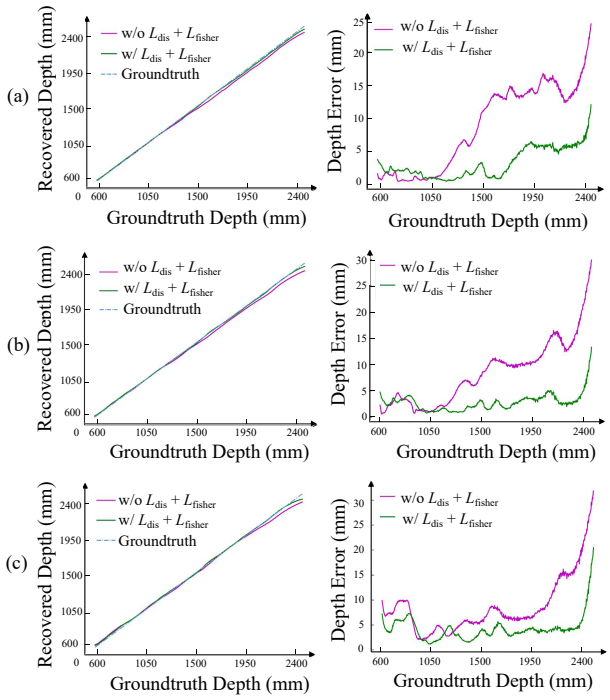

Figure 6. Ablation study of with (w/) and without (w/o) fisher-information guidance loss and discriminative loss, of noise setting (E, β): (a) (20000,6000), (b) (14000,6000), (c) (10000, 6000).

Ablation	MAE (mm)		
SEB	33.327	31.987	45.048
Cascaded Structure	23.882	32.411	44.479
$L_{\text{dis}} + L_{\text{fidelity}}$	40.997	51.537	115.435
$L_{\text{fisher}} + L_{\text{fidelity}}$	64.291	73.966	143.629
L_{fidelity}	24.990	24.119	28.818
Ours	12.857	13.763	18.264

Table 2. Quantitative Comparison of Ablation study of different noise settings, the noise setting is the same as Tab. 1.

5. Physical Experiment Results.

Prototype system. To demonstrate our method with a physical experiment, we built a prototype system that can implement iToF imaging with arbitrary coding functions. The learned modulation and demodulation functions are generated by the function signal generator (DG5252, Rigol). We adopt a 638nm laser diode with a maximum power of 200mw (L638P200, Thorlabs) as the light source. With the bias-T coupling circuit (LDM56, Thorlabs), the laser can be modulated by the learned modulation function. We scan the scene with a galvo-mirror system (GVS012, Thorlabs). The reflected light is focused by a lens (AF NIKKOR, Nikon) onto an avalanche photodiode (APD430A, Thorlabs). The APD converts reflected light into electrical signals. Then the converted signal is multiplied with the learned demodulation function with a multiplier (AD835, ADI), and the output is further amplified (OPA847, TI) and low-pass filtered (EF110, Thorlabs). Finally, the analog signals are sampled, quantized and translated into digital values by ADC (ADS112C04, TI). The scanning and acquisition of the system are controlled by the Microcontroller (STM32F103, ST), as in Fig. 7.

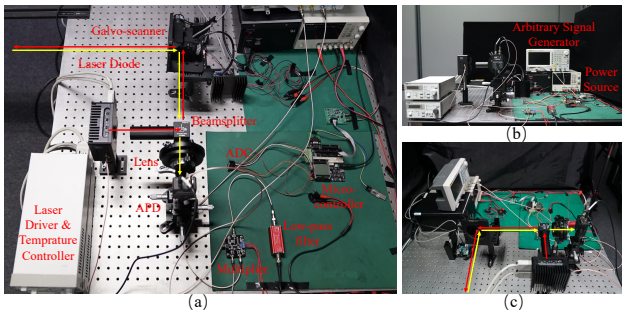

Figure 7. Experimental system, (a)-(c) top, front, side views.

Experimental Results. Through adjusting the output power of the light source and the additional ambient light source power, we capture images over a large range of noise levels. To qualitatively demonstrate the superiority of the proposed method, we first capture images of the three books, at depths 1.60 m, 1.90 m, and 2.20 m. As shown in Fig. 8, the MAE of different methods is 10.79 cm, 4.23 cm, 7.06 cm, and 1.66 cm from left to right, and the depth error of our method is the lowest, demonstrating the effectiveness of the proposed method. We further conduct depth imaging with different scenes under different noise levels as shown in Fig. 8, the depth reconstruction results are noisy due to the large noise, while the proposed method recovered the depth with much less noise effect and elegant quality in smoothness comparing with the state of the art methods.

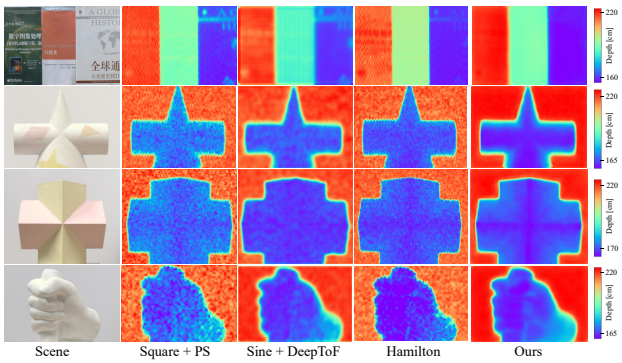

Figure 8. Performance comparisons in physical experiments.

6. Discussion

In this paper, complex lighting environment with varying ambient light source distribution has not been considered in our physical imaging model. However, with the proposed information-guided iToF imaging framework, the ambient light distribution could be directly modeled in the physical model, and through rendering training dataset with different complex ambient light distribution, the optimal iToF imaging model for complex ambient light environment could be learned end-to-end, we leave this work as our future work.

As for the on-chip implementation of our method, we will explore the DDS chip (AD9954, ADI) or DAC + FPGA, with the addition of a waveform processing circuit, to implement learned coding functions.

7. Conclusion

In conclusion, we propose an information theory guided framework to optimize the coding functions and the dual-branch reconstruction network of iToF imaging jointly. Specifically, we formulate the differential imaging model with physical implementation constraints, and propose a dual-branch deep neural network for depth reconstruction. With the proposed discriminative fisher loss and end-to-end network training, we could find the optimal coding functions and reconstruction network under large noise. The proposed method was demonstrated with extensive qualitative and quantitative results, and we further build a prototype iToF imaging system and validate our method with physical experiments.

8. Acknowledgments

This work was supported by NSFC Projects 61971465, and Fundamental Research Funds for the Central Universities, China (Grant No. 0210-14380184).

References

[1] Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *Proceedings of International Conference on Machine Learning*, pages 41–48, 2009. 5

[2] Alexander W Bergman, David B Lindell, and Gordon Wetzstein. Deep adaptive lidar: End-to-end optimization of sampling and depth completion at low sampling rates. In *Proceedings of the IEEE International Conference on Computational Photography*, pages 1–11, 2020. 2

[3] Julie Chang and Gordon Wetzstein. Deep optics for monocular depth estimation and 3d object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 10193–10202, 2019. 2

[4] Jerry Chao, E Sally Ward, and Raimund J Ober. Fisher information for emccd imaging with application to single molecule microscopy. In *IEEE The Asilomar Conference on Signals, Systems and Computers*, pages 1085–1089, 2010. 2

[5] 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. 2, 6, 7

[6] Shay Elmalem, Raja Giryes, and Emanuel Marom. Learned phase coded aperture for the benefit of depth of field extension. *Optics Express*, 26(12):15316–15331, 2018. 2

[7] R Grootjans, W Van der Tempel, D Van Nieuwenhove, C de Tandt, and M Kuijk. Improved modulation techniques for time-of-flight ranging cameras using pseudo random binary sequences. In *Proceedings of the IEEE Lasers and Electro-Optics Society*, page 217, 2006. 2

[8] Mohit Gupta, Andreas Velten, Shree K Nayar, and Eric Breitbach. What are optimal coding functions for time-of-flight imaging? *ACM Transactions on Graphics*, 37(2):1–18, 2018. 1, 2

[9] 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. 1, 2, 5, 6

[10] Miles Hansard, Seungkyu Lee, Ouk Choi, and Radu Horaud. Time of flight cameras: Principles, methods, and applications, 2012. 2, 5, 6

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1026–1034, 2015. 5

[12] 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. 5

[13] Julio Illade-Quinteiro, Víctor M Brea, Paula López, Diego Cabello, and Gines Doménech-Asensi. Distance measurement error in time-of-flight sensors due to shot noise. *MDPI Sensors*, 15(3):4624–4642, 2015. 1

[14] Herve Jegou, Hedi Harzallah, and Cordelia Schmid. A contextual dissimilarity measure for accurate and efficient image search. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2007. 5

[15] 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. 5

[16] 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. 2

[17] Beomjun Kim, Jean Ponce, and Bumsub Ham. Deformable kernel networks for joint image filtering. *International Journal of Computer Vision*, 129(2):579–600, 2021. 5

[18] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 5

[19] Robert Lange. 3d time-of-flight distance measurement with custom solid-state image sensors in cmos/ccd-technology. 2000. 1

[20] Robert Lange and Peter Seitz. Solid-state time-of-flight range camera. *IEEE Journal of Quantum Electronics*, 37(3):390–397, 2001. 2

[21] 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. 2

[22] Parsa Mirdehghan, Wenzheng Chen, and Kiriakos N Kutulakos. Optimal structured light a la carte. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6248–6257, 2018. 2

[23] Pushmeet Kohli Nathan Silberman, Derek Hoiem and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *Proceedings of the European Conference on Computer Vision*. Springer, 2012. 5

[24] 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. 2

[25] 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. 2

[26] 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. 5

[27] Andrew D Payne, Adrian A Dorrington, and Michael J Cree. Illumination waveform optimization for time-of-flight range imaging cameras. In *Videometrics, Range Imaging, and Applications XI*, volume 8085, page 80850D. International Society for Optics and Photonics, 2011. 2

[28] Andrew D Payne, Adrian PP Jongenelen, Adrian A Dorring-

ton, Michael J Cree, and Dale A Carnegie. Multiple frequency range imaging to remove measurement ambiguity. In *Optical 3-d Measurement Techniques*, 2009. 2

[29] Yoav Shechtman, Steffen J Sahl, Adam S Backer, and William E Moerner. Optimal point spread function design for 3d imaging. *Physical Review Letters*, 113(13):133902, 2014. 2

[30] 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. 2

[31] 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. 1, 2, 5, 6, 7

[32] 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. 2

[33] Shiyu Tan, Yicheng Wu, Shoou-I Yu, and Ashok Veeraraghavan. Codedstereo: Learned phase masks for large depth-of-field stereo. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7170–7179, 2021. 2

[34] Yicheng Wu, Vivek Boominathan, Huaijin Chen, Aswin Sankaranarayanan, and Ashok Veeraraghavan. Phasecam3d—learning phase masks for passive single view depth estimation. In *Proceedings of the IEEE International Conference on Computational Photography*, pages 1–12, 2019. 2