Depth Restoration in Under-Display Time-of-Flight Imaging

Xin Qiao, Chenyang Ge, Pengchao Deng, Hao Wei,
Matteo Poggi, *Member, IEEE* and Stefano Mattoccia, *Senior Member, IEEE*

Abstract—Under-display imaging has recently received considerable attention in both academia and industry. As a variation of this technique, under-display ToF (UD-ToF) cameras enable depth sensing for full-screen devices. However, it also brings problems of image blurring, signal-to-noise ratio and ranging accuracy reduction. To address these issues, we propose a cascaded deep network to improve the quality of UD-ToF depth maps. The network comprises two subnets, with the first using a complex-valued network in raw domain to perform denoising, deblurring and raw measurements enhancement jointly, while the second refining depth maps in depth domain based on the proposed multi-scale depth enhancement block (MSDEB). To enable training, we establish a data acquisition device and construct a real UD-ToF dataset by collecting real paired ToF raw data. Besides, we also build a large-scale synthetic UD-ToF dataset through noise analysis. The quantitative and qualitative evaluation results on public datasets and ours demonstrate that the presented network outperforms state-of-the-art algorithms and can further promote full-screen devices in practical applications.

Index Terms—Time-of-Flight, depth restoration, denoising, under display, CNN.

1 Introduction

Continuous-wave time-of-flight (ToF) cameras, also known as indirect ToF or iToF, are active depth sensors that have been widely used in computer vision and graphics applications due to the high-quality, low-cost, compact structure and high frame rate [1], [2]. In many of the applications, especially in smartphones, the demand for full-screen with an almost 100% screen-to-body ratio promotes manufacturers to consider locating a ToF camera Under a Display screen (UD-ToF camera) [3]. The full-screen devices can not only provide amazing visual effects but also enhance the human-computer interaction experience. While UD-ToF cameras have many exciting advantages, it also brings several imaging problems. Due to the pixels and circuits in the display [4], the signal received by the ToF sensor will be subjected to attenuation, multi-reflection, and diffraction effects, as shown in Fig. 1. Additionally, a circular polarizer is usually placed in front of the display to make a clearer and brighter image for users, further reducing the amplitude of the received signal. So the raw UD-ToF measurement will become blurry, and its SNR will drop sharply. Different from traditional cameras with digital image sensors, the depth map of a scene is obtained from the raw measurements in an indirect way (e.g., nonlinear operation). As a result, these corrupted raw measurements lead to an unreliable depth map [5]. In a word, UD-ToF depth restoration is a new task that needs to jointly solve denoising, deblurring, and raw measurements enhancement in low-sensing environments.

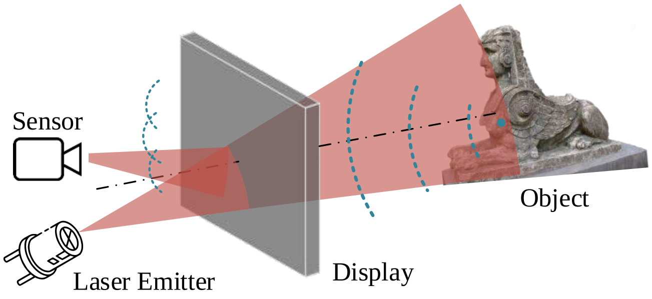


Fig. 1. Overview of the proposed under-display ToF camera. This depth camera allows for a full-screen experience while also introducing attenuation, multi-reflection and diffraction.

At present, there are mainly two types of OLED displays that can be used for full-screen devices, namely Transparent Organic Light-Emitting Diode (T-OLED) and Pentile OELD (P-OLED) [4]. Considering that the received light will be attenuated twice by the screen, it is necessary to select a display with a higher light transmission rate. Compared to the P-OLED panel with a light transmittance rate of less than 10%, the T-OLED panel with a light transmittance rate of up to 80% becomes our optimal choice. It should be noted that the display is nonactive in this paper, similarly to [4]. Since the term "ToF camera" in the following are all based on continuous-wave time-of-flight technique, without ambiguity, we refer to continuous-wave time-of-flight and ToF camera under T-OLED display as ToF and UD-ToF camera, respectively.

To acquire UD-ToF depth maps with higher precision, we can adjust the output power, integration time and modulation frequency [6]. However, the hardware adjustments will be subject to several practical constraints in these applications. Increasing the output power of laser leads to excessive power consumption and violates the human-eye safety rule, whereas utilizing more than two different

- *X. Qiao, C. Ge, P. Deng and H. Wei are with the Institute of Artificial Intelligence and Robotics, Xi'an Jiaotong university, Xi'an, China, 710049. E-mail: wudiqx@stu.xjtu.edu.cn*
- *M. Poggi and S. Mattoccia are with the Department of Computer Science and Engineering of the University of Bologna, Italy, 40136.*

Manuscript received April 19, 2005; revised August 26, 2015.

modulation frequencies may result in more severe motion blur due to its longer capture. Consequently, we turn to post-processing algorithms to improve the SNR of raw UD-ToF measurements, thereby obtaining a depth map with the same measurement accuracy as the display-free ToF camera.

In traditional processing methods [7], [8], this ill-posed problem is usually solved by a depth optimization procedure. For this depth restoration task, the accuracy of depth maps should be improved while denoising and deblurring, which is challenging because processing raw measurements may destroy their inherent physical relationship. These models are insufficient to deal with the complicated UD-ToF depth degradation problem due to information loss caused by hand-crafted feature extraction. To alleviate the problem, we propose a calibration method to eliminate the spatial noise in UD-ToF cameras, also known as Fixed-Pattern Noise (FPN), caused by the display panel. This step can highlight the image features without loss of information, making the subsequent processing easier.

Due to the powerful feature extraction and representation capability, learning-based approaches, especially deep learning algorithms, have made significant progress in 3D image processing, such as denoising [9], [10], [11], deblurring [12], [13], multi-path interference (MPI) correction [14], [15], [16], and depth super-resolution [17], [18], which inspire us to make further exploration. Considering the physical relation of raw measurements, we propose a cascaded network to solve the complex depth degradation problem. Unlike the previous approaches that improve depth quality in single raw domain or depth domain, our method performs depth optimization in both domains. Specifically, a complex-valued neural network is presented to enhance measurements in raw domain. In depth domain, we propose a Multi-Scale Depth Enhancement Block (MSDEB), which further helps refine depth maps. Moreover, some traditional ToF image processing components, such as mapping from raw measurements to depth [14], can also be utilized in our processing procedure.

For training, some synthetic and real datasets are created by previous works, such as [10], [11], [15], [19], [20]. Nevertheless, no real dataset can be employed for the UD-ToF depth restoration task. To this end, we set up a specific range-imaging device that can capture raw measurements for the same scenes in under-display and display-free modes, respectively. Furthermore, a synthetic dataset with a minor or no domain gap with the real dataset can generalize better and requires fewer efforts to collect. Nevertheless, few synthetic datasets are adeguate for our task because they do not provide noisy/clean raw measurement pairs or are not large enough. To remedy this issue, we create a synthetic dataset via noise analysis of ToF raw measurements, which is also suitable for other tasks, such as denoising, deblurring.

In this paper, we propose a learning-based approach to address the problem of UD-ToF depth degradation. Specifically, the contributions of our work are as follows:

- We present the properties of the spatial noise in UD-ToF cameras and develop a calibration method to remove it. This operation significantly improves the UD-ToF ranging accuracy.
- We propose a cascaded deep network working both in raw and depth domains. It consists of two main sub-networks: the first is the complex-valued convolutional neural network, CV-ToFNet, which jointly performs denoising, deblurring and raw measurements enhancement in the raw domain; the second performs depth refinement in the depth domain based on the proposed MSDEB.
- No dataset exists to study our problem. Hence, we create a large-scale synthetic dataset named *SUD-ToF* by analyzing the noise models, which also can be used in other issues like denoising and deblurring. Moreover, we set up a specific range-imaging device and build a real dataset, *RUF-ToF*, containing noisy/clean raw measurement pairs and depth maps. The two datasets proved to have a small domain gap.
- The experiments on our datasets validate that the proposed approach effectively restores UD-ToF depth in real time. Besides, we demonstrate that the proposed method is superior to state-of-the-art depth restoration approaches in UD-ToF depth restoration and denoising.

The paper is organized as follows: Section II discusses the related work. Then we present the principle of ToF imaging in Section III. The proposed learning-based approach is detailed in Section IV while Section V introduces the datasets including real data acquisition and synthetic data generation. Experimental comparisons and ablation study are demonstrated in Section VI. Finally, we draw a conclusion in Section VII.

2 Related Work

Depth restoration has always played an essential role in computer vision. Here, we summarize previous works related to this task from multiple aspects.

A. Image Enhancement.

The low-sensing environment is capture conditions where sensors obtain images with lower amplitude values than normal shots, which is one of the topics similar to ours. Many approaches, such as [21], [22], [23], have been presented to enhance low-sensing RGB images, which can give us inspiration on depth quality improvement.

With the recent trend of full-screen devices, which requires placing the front camera behind the screen itself, Under-Display camera restoration [24], [25], [26] emerged as a new computer vision task. Zhou et al. [4] analyze such a problem by focusing on two main types of displays. They design a system for data acquisition and a model-based pipeline to synthesize under-display images. In [27], Feng et al. present a physics-based image formation model and measure the Point Spread Function (PSF) of under-display cameras through an imaging system. They also provide a pipeline to generate synthetic under-display images, and a dynamic skip connection Network is designed to restore their quality. Kwon et al. [28] propose a controllable image restoration algorithm to alleviate the artifacts of blur and noise in under-display images.

B. Depth Restoration.

Despite advances in resolution and imaging quality, shot noise, Multi-Path Interference (MPI), pixel circuit noise, and read noise still hamper wider applications of current consumer ToF cameras [29]. For UD-ToF cameras, the images also suffer from artifacts caused by the display panel. Thus ToF depth restoration remains an open challenge.

For traditional model-based approaches, Fuchs et al. [30] investigate a multi-path model to distort depth maps, which incorporates the scene geometry. Kadambi et al. [31] regard the MPI removal as an inverse problem and use sparse deconvolution to estimate latent clean depth maps from single modulation frequency measurements. The method from Bhandari et al. [32] provides a closed-form solution based on $2K + 1$ frequency measurements and spectral estimation theory. Freedman et al. [7] conduct Sparse Reflections Analysis (SRA) to address many types of MPI and realize real-time processing. Phasor imaging for light transport analysis is proposed by Gupta et al. [33]. Georgiev et al. [34] analyze the noise model of raw measurements and remove the FPN to get accurate depth maps, which is substantial for the following process. Then they [35] apply non-local means denoising algorithm working in complex domain. Another recent work [36] by Rossi et al. formulates the depth restoration task as an optimization problem. They propose a variational framework that enforces piece-wise planar to refine depth.

Recently, an increasing number of data-driven methods for depth restoration have been introduced. Son et al. [37] first apply a neural network to MPI removal of ToF cameras and use the data collected by a structured light camera as the ground truth for training. Based on an auto-encoder architecture, Marco et al. [15] train a depth-to-depth network in two stages. The aforementioned data-driven approaches take depth maps as the input of their networks, but much useful information in raw measurements is lost after nonlinear mapping. So an end-to-end network is proposed for denoising and MPI compensation of ToF images [10]. Guo et al. [14] utilize a kernel-prediction network (KPN) working in raw domain to address the ToF artifacts, including motion, MPI, and shot noise. In [20], ToF depth maps are refined through a kernel prediction network with the help of aligned RGB images. SHARP-Net [38] introduced by Dong et al. exploits residual pyramid for ToF depth denoising in a coarse-to-fine manner. Chen et al. [39] propose a neural network to translate the corrupted ToF raw measurements to high-quality depth maps even in the conditions of extreme short exposure or low-sensing environment.

Unlike the above approaches that refine depth only in raw or depth domain, our approach containing two subnets conducts depth restoration in both domains. We take the inherently physical relationship between ToF raw measurements into account, which is equivalent to adding regularization to each network layer. Furthermore, inspired by the conventional method of multi-scale detail enhancement [21], we design a subnet to perform depth refinement in depth domain.

C. Complex-valued Network.

Due to the higher representation power of complex numbers than real-valued weights, complex-valued networks have recently demonstrated superior performance in several tasks, including speech enhancement [40], [41], action recognition [42], image classification [43] and MRI reconstruction [44], [45]. In signal processing, complex numbers can accurately delineate the amplitude and phase of specific signals, which has great potential in improving the learning speed and generalization ability of the network. For the UD-ToF depth restoration problem, the raw measurements can also construct a complex representation based on their physical relationship [35], which motivates us to devise a complex-valued network for UD-ToF restoration.

D. Synthetic ToF Dataset.

To enable training, a few of datasets for ToF imaging are proposed. Considering that real dataset collection is time-consuming and labor-intensive, most are synthesized with tools in computer graphics. Agresti et al. [19] build real and synthetic datasets with raw measurements, depth maps and amplitude maps, whereas their scale is not enough to support the training of many approaches. The synthetic dataset from DeepToF [15] consists of 25 different scenes with 1050 depth images, but ToF raw measurements are absent. The NYU-Depth V2 dataset [46] and the human body dataset introduced by [11] also have the same situation as DeepToF. In addition, Su et al. [10] propose a large-scale synthetic dataset with noisy/clean ToF raw measurements using the transient rendering technique. However, the characteristics of Gaussian noise added empirically are pretty different from real noise and cannot be generalized to real scenes. Although Guo et al. [14] establish the FLAT dataset with both raw measurements and depth maps, they apply the collected real noise on the synthetic data in the form of a lookup table, which requires much labor to capture tens of thousands of images. This is impractical to build a large-scale dataset for our task. Moreover, no existing real datasets are prepared for the UD-ToF depth restoration task. Therefore, we collect a real dataset by a specific range-imaging system and propose a large-scale synthetic dataset via noise analysis with a small domain gap from the real one.

3 Foundation of Time-of-Flight Imaging

This section presents the principle of ToF range imaging and the formation of UD-ToF sensing.

Technical foundations. To measure the distance to the target in the scene, ToF cameras emit amplitude-modulated light and measure the phase difference between the sent and received signals. Assume the emitted signal is sinusoidal of the form:

$$s(t) = \cos(\omega t) \tag{1}$$

then the reflected signal, measured and stored for a single pixel, at the same frequency can be expressed, in function of time t, as

$$r(t) = A\cos(\omega t - \phi) + B \tag{2}$$

where ω is the angular frequency, A is the amplitude of the reflected signal, B is the offset of the reflected signal due to ambient light, and ϕ is the phase shift from which the target distance can be obtained. The cross-correlation between the emitted signal and the reflected signal is formulated as:

$$C(\tau) = r \otimes s = \frac{A}{2}\cos(\omega\tau + \phi) \tag{3}$$

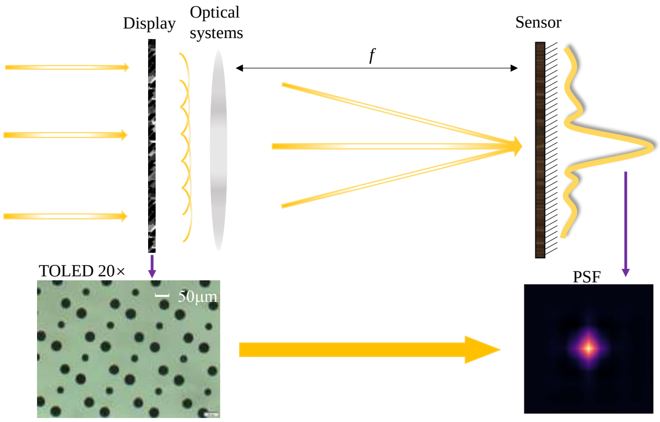


Fig. 2. The schematic diagram of the UD-ToF camera. The image on bottom left is the pattern of our display observed under $20\times$ magnification using an inverted fluorescence microscope. The image on bottom right is the measured PSF.

The ToF camera takes four samples per cycle at a certain modulation frequency f, and the phase of each sample is stepped by 90°, that is, $I_i = C(\frac{\pi}{2\omega}i), i = 0, 1, 2, 3$. Each raw measurement is generally assumed to follow the Gaussian distribution with zero mean and same variance σ, which is independent and identically distributed (i.i.d.) [35]. After the sampling process, the phase shift ϕ and the amplitude A for each pixel are determined by

$$\phi = \text{atan2}\frac{I_3 - I_1}{I_0 - I_2} \tag{4}$$

$$A = \frac{1}{2}\sqrt{(I_3 - I_1)^2 + (I_0 - I_2)^2} \tag{5}$$

Then the distance can be obtained by

$$d = \frac{c}{2f}(\frac{\phi}{2\pi} + N) \tag{6}$$

where $c \approx 3 \times 10^8 m/s$ is the speed of light and N is the number of phase wrap.

To correct the systematic errors for the ToF camera, we build a modified model based on a previous work of depth correction [47]:

$$\begin{aligned}\Delta d =& \alpha_0 + \alpha_1\tilde{d} + \alpha_2\cos(4k\tilde{d}) + \alpha_3\sin(4k\tilde{d}) + \\ & \alpha_4\cos(8k\tilde{d}) + \alpha_5\sin(8k\tilde{d}) + \alpha_6 R + \alpha_7 T\end{aligned} \tag{7}$$

where α_m is coefficient of each term in the model, $m = 0, 1, ..., 7$, $\tilde{d}$ is the measured distance, R denotes the radial distance in the image plane, T denotes the sensor temperature and $k = 2\pi f/c$ is the wavenumber. So the corrected distance can be obtained by $d_c = \tilde{d} - \Delta d$. This model matches the data accurately and can control the ranging accuracy to 1% within the maximum range.

Phase Unwrapping. When the phase delay exceeds 2π radians, due to the periodicity of the modulated light signal,it will still fall in the range of $[0, 2\pi]$. Therefore, the estimated depth map will also wrap around, and its maximum unambiguous ranging distance is $c/2f$. With only one modulation frequency, it is difficult to estimate the integer N, so we usually set $N = 0$. To handle this issue, many ToF camera manufacturers adopt the dual-frequency scheme [40], [48], [49], which can extend the maximum unambiguous range to $d_{max} = c/2f_{gcd}$, where f_{gcd} is the greatest common divisor of the two modulated frequencies f_1 and f_2.

Complex representation for ToF data. Generally, signals with positive scalar intensity and direction can be modeled using complex numbers. For ToF data, the raw measurements can be converted into two-phase maps:

$$\xi = \frac{1}{2}(I_0 - I_2), \eta = \frac{1}{2}(I_3 - I_1) \tag{8}$$

These two variables can also be represented by amplitude A and phase shift ϕ:

$$\xi = A\cos\phi, \eta = A\sin\phi \tag{9}$$

Based on the relationship between the above two equations, we can express the raw measurements in complex numbers $Z = Ae^{j\phi}$. And the variance of the two components is given by $\sigma_\xi^2 = \sigma_\eta^2 = \sigma^2/2$.

UD-ToF data formation. In UD-ToF cameras, the raw measurements are affected by several types of degradation, including sensor noise, systematic errors, and diffraction (see Fig. 2). The systematic errors can usually be corrected by model-based methods, thus the key to the UD-ToF restoration problem in portable devices lies in denoising and deblurring. Given the observed scene I and well-calibrated ToF camera, this degradation can be modeled as:

$$\tilde{I} = h(I) \otimes k + n \tag{10}$$

where I and $\tilde{I}$ are the true ToF data and observations, respectively, h denotes camera functions to observed scenes which consists of vignetting, pixel delay, modulation function and temperature drift. k is the blur kernel, also known as the Point Spread Function (PSF), $\otimes$ denotes the convolution operator and n represents signal-dependent Gaussian noise [50], [51], [52].

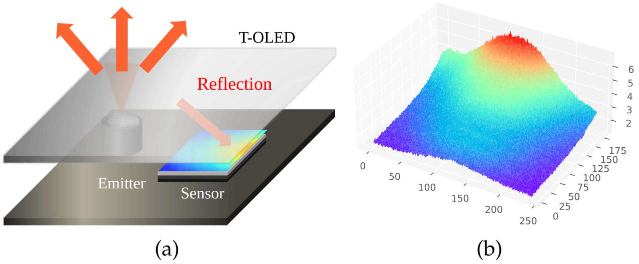
4 Our Approach

Our goal is to restore high-quality depth maps from corrupted raw measurements of UD-ToF cameras while taking into account energy efficiency and running speed. The method consists of two stages. First, we analyze the properties of FPN in the UD-ToF camera and remove it. Then we describe the presented network structure, including two cascaded subnets. The first is designed for depth denoising, deblurring, and raw measurements enhancement, while the second focuses on depth refinement. Details are given in the following subsections.

4.1 Fixed-Pattern Noise Removal

Due to the presence of the T-OLED display, the signal amplitude received by UD-ToF sensors drops dramatically, resulting in a much lower SNR than that of display-free ToF cameras. Noise sources in UD-ToF sensors can generally be classified into two types: temporal noise and spatial noise. In this part, we mainly consider removing the spatial noise from UD-ToF raw measurements while the temporal noise is processed by our network.

For display-free ToF cameras ranging in low-sensing environments, FPN is the most critical component in the spatial noise, which originates from the image non-uniformity.



(a) (b)

Fig. 3. FPN of UD-ToF is mainly caused by the reflection of the T-OLED screen. The amplitude of FPN is large enough to affect the accuracy of the depth maps. (a) Illustration of FPN formation. (b) Amplitude of FPN.

Georgiev et al. [34] build a noise model and develop an FPN removal procedure for display-free ToF cameras working in low-sensing environments. However, for UD-ToF cameras, another more crucial factor dramatically reduces the ranging accuracy. This factor is introduced by the structure of the T-OLED display. A conducting anode and a cathode, with thin organic layers between them, stack together and constitute the typical T-OLED panel. When exposed to IR light emitted by the laser, the high-reflectivity cathode will reflect it, as seen in Fig. 3a. Based on the optical path of the reflected light, it is mainly distributed in the region far away from the laser on the sensor. Furthermore, we find that the reflected light intensity is temporally constant in UD-ToF imaging. Therefore, the reflected light is received by the ToF sensor and converted into a spatially fixed electrical signal, which, together with the particular noise pattern on the ToF sensor, form the FPN of UD-ToF, as illustrated in Fig. 3b. The FPN noise causes a clear bias in depth estimation, and its effect can be seen in both raw measurements and amplitude maps, as shown in Fig. 4. We can formulate this spatial noise model as follows:

$$\tilde{I}_i(x,y) = I_i(x,y) + u_i(x,y) + n_G(x,y) \tag{11}$$

where $I_i(x,y)$ is the true intensity value output by the ToF sensor, $\tilde{I}_i(x,y)$ is the value of the observed intensity for $x, y \in \Omega$. $u_i(x,y)$ and $n_G(x,y)$ denotes FPN and additive Gaussian noise with mean $\mu = 0$ and standard deviation σ_G, respectively. Additive Gaussian noise $n_G(x,y)$ can be removed by various modern filtering techniques, which will be discussed in following subsections, so we focus on estimating the FPN here.

To this end, we place the UD-ToF camera where there are no objects within its range – since an occluder in front of the ToF sensor would cause different reflections and, thus, artefacts – then collect hundreds of frames (e.g., 300 frames) of raw measurements. Note that we collect data at night to avoid the influence of ambient light. One can also capture images in other dark environments. To verify that ambient light in dark will not affect the sensor, we compare the UD-ToF amplitude maps obtained by the camera in three poses, that is, 30° up, 30° down, and horizontal. If the ambient light does influence the UD-ToF imaging, these three amplitude maps will show apparent differences. However, the mean absolute difference and variance between them are only 0.12 and 2.8×10^{-5}, respectively. The results show that the UD-ToF sensor is not affected by ambient light at night, and we can conduct subsequent operations without considering it.

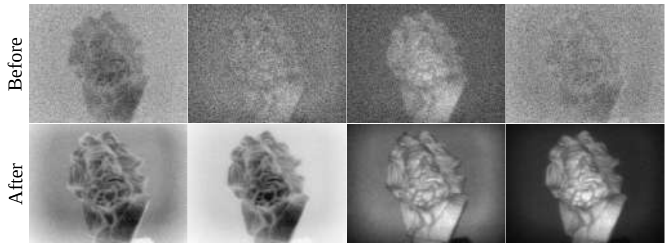


Fig. 4. Four correlated raw measurements before and after FPN removal. After FPN removal, each frame of raw measurements becomes clear, allowing the network to quickly learn the features of these images.

After averaging the hundreds of raw measurements, the mean and variance of the Gaussian noise tend to zero ($\mu_{n_G} \to 0, \sigma_G \to 0$). Therefore, under this operating condition, nothing would be captured by the UD-ToF camera except the FPN. By subtracting it from the corresponding raw measurements, the data containing only Gaussian noise can be obtained. The raw measurements after FPN removal are shown in Fig. 4. Then they are sent as input to the proposed network.

4.2 Subnet In Raw Domain

For deep learning models, it is always a challenge to deal with generalization and overfitting. Regularization methods help in overcoming this problem. Significantly, the models integrated with prior knowledge usually obtain superior results [53]. For example, conventional real-valued CNNs, treated as a special variant of Multi-Layer Perception (MLP) for images, get remarkable achievements. For ToF imaging, we apply additional prior knowledge to the subnet, which is built upon recent work in complex-valued networks [54]. Further, this is the first time using a complex-valued network to estimate phase and amplitude in ToF imaging jointly.

Rather than directly sending raw measurements to each channel of the network, we convert them into a complex representation as the input, as described in Eq.(8). We build the complex-valued model that deals with complex inputs and weights as a generalization of real-valued networks. To perform back propagation in optimization, the model must be differentiable, which is a much stronger constrain than its real counterpart. According to $\mathbb{CR}$-calculus [55], we can conduct gradient descent to update parameters if the real and imaginary components are differentiable respectively. When we perform a convolution operation on a complex matrix $\boldsymbol{Z} = \boldsymbol{X} + j\boldsymbol{Y}$ with a complex kernel $\boldsymbol{K} = \boldsymbol{W}_R + j\boldsymbol{W}_I$, the result is

$$\boldsymbol{Z} * \boldsymbol{K} = (\boldsymbol{W}_R * \boldsymbol{X} - \boldsymbol{W}_I * \boldsymbol{Y}) + j(\boldsymbol{W}_I * \boldsymbol{X} + \boldsymbol{W}_R * \boldsymbol{Y}) \tag{12}$$

where $\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{W}_R$ and $\boldsymbol{W}_I$ are real matrices. They can be represented in algebraic notation:

$$\begin{bmatrix} \Re(\boldsymbol{Z} * \boldsymbol{K}) \\ \Im(\boldsymbol{Z} * \boldsymbol{K}) \end{bmatrix} = \begin{bmatrix} \boldsymbol{W}_R & -\boldsymbol{W}_I \\ \boldsymbol{W}_I & \boldsymbol{W}_R \end{bmatrix} * \begin{bmatrix} \boldsymbol{X} \\ \boldsymbol{Y} \end{bmatrix} \tag{13}$$

Note that the modulation frequencies are set to 20/100MHz, but the subnet input is the complex ToF data at

100MHz due to the fact that ranging accuracy is positively correlated with modulation frequency [16]. Raw measurements at 20MHz only guide the unwrapping process and will not affect the ranging accuracy.

On the other hand, although the U-Net architecture was initially proposed for medical image segmentation, it has also exhibited excellent performance in image reconstruction. Our network also adopts this architecture, which is detailed in Fig. 5. Furthermore, to avoid gradient vanishing or exploding and reach a faster convergence, our model learns the intensity residual instead of latent clean raw measurements from corrupted raw measurements. In this network, the resolution of input intensity images is reduced to $\frac{1}{8}$ of its original size after three downsampling layers. The bottleneck part is composed of two consecutive residual blocks for performance improvement so that more accurate intensity images can be restored with the same size as the input after upsampling. Max-pooling operation and bilinear interpolation are replaced with strided convolution for memory efficiency and inference accuracy. Furthermore, to maintain consistency with the geometry of corresponding scenes, we add skip connections to convolution layers between the encoder and decoder.

Various activation functions have been developed to deal with complex numbers. We use the leaky $\mathbb{C}$ReLU function in this network, which applies leaky ReLUs separately to the real and imaginary parts of neuron input

$$\mathbb{C}\mathrm{ReLU}(z) = \mathrm{ReLU}(\Re(z)) + i\mathrm{ReLU}(\Im(z)) \tag{14}$$

For complex batch normalization and complex weight initialization, they are well-defined building blocks in [54]. In a previous study, the cardioid function $f(z) = \frac{1}{2}(1 + \cos \angle z)z$ yields desirable results for magnetic resonance fingerprinting (MRF) [56]. In Section VI, we will test the performance of both activation functions in the same architecture.

After raw measurements enhancement, the wrapped depth map can be obtained from Eq. (4) assuming $N = 0$. Then we employ the unwrapping method in [48] to get the unambiguous depth map, which is sent to the subnet as input in depth domain.

4.3 Subnet In Depth Domain

Through the previous processing, both the raw measurements and the depth maps have been significantly improved. Nonetheless, there is still apparent shot noise in some areas of depth maps, such as flat regions and edges, which will hinder the downstream applications of UD-ToF cameras. To suppress the noise to a lower level, we propose a small network based on formatted learning framework [23] in depth domain.

In formatted learning framework, the formatting layer (the orange part in Fig. 5) focuses on low-pass filtering while the rest of the subnet aims at extracting fine geometry in the scene. Thus, structural detail preservation and noise suppression in depth maps are well handled. However, FormNet only uses several flat convolution layers as the formatting layer, with limited representation. Inspired by [21], we design a module named Multi-Scale Depth Enhancement Block (MSDEB) to extract image features from different scales. Considering that directly utilizing kernels with different sizes will increase the computational burden, we introduce channel split into MSDEB to reduce the model size. Hence, the MSDEB allows multiple receptive fields in one layer. As shown in Fig. 6, the input first passes through a convolution layer with a kernel size of 1×1. Then we split the feature maps into four different branches evenly. Three branches are convolved by kernels of different sizes, which are $3 \times 3, 5 \times 5$ and 7×7, respectively, while the last is an identity mapping by shortcut connection. The four channels are sent to a 1×1 convolutional layer after concatenation to fuse the multi-scale information. A similar structure named multi-scale feature extraction is proposed in [57]. For better depth construction, we replace the original flat convolution with the U-net backbone of the formatting layer. The U-net applies two downsampling operations with strided convolution to have a large receptive field. Then the feature maps are restored to their original size after another two upsampling layers with deconvolution. The rest of the subnet for restoring structural details consists of three cascaded MSDEBs.

Overall, the subnet in depth domain strikes a reasonable balance between preserving high-frequency details and suppressing noise with a small number of parameters.

4.4 Loss Function

For improving the accuracy for depth estimation, we design the loss function taking into account both raw measurements and depth maps. Because of the complex representation of ToF raw measurements in our network, it is necessary to devise a loss function working in complex domain. To remove the high-frequency noise in the depth maps, we also introduce a smoothness term into a locally-smooth loss. Furthermore, the raw measurements inevitably introduce some unreliable pixels. In order to improve learning accuracy, we design a mask to remove these pixels in each frame. The confidence of received signals can characterize the noise level and reliability of raw measurements. According to this feature, we use augmented confidence [58] to evaluate the quality of each pixel, thus generating a mask to separate normal pixels from unreliable ones that are ignored in backpropagation.

Raw loss. We use an L_1 penalty term as the raw loss. The loss forces the subnet working in raw domain to minimize the pixel-wise mean absolute error between the estimated real and imaginary parts and ground truth in each frame. When the whole network is trained jointly, this term acts in the middle – supervising the first sub-network only, yet favouring the convergence of the overall architecture. The L_1 penalty term is depicted as:

$$L_{raw} = (||\tilde{\xi} - \xi||_1 + ||\tilde{\eta} - \eta||_1) \odot M_v \tag{15}$$

where $\tilde{\xi}$ and $\tilde{\eta}$ denote the real and imaginary part of the estimated raw measurement, respectively, ξ and η are the corresponding ground truth, M_v is the validity mask, and $\odot$ denotes element-wise multiplication.

Depth loss. To improve the quality of depth maps, we add another loss term *depth loss*. The depth loss is computed as the mean absolute error (MAE) on pixel depth and summed to a smoothness term, computed as the gradient

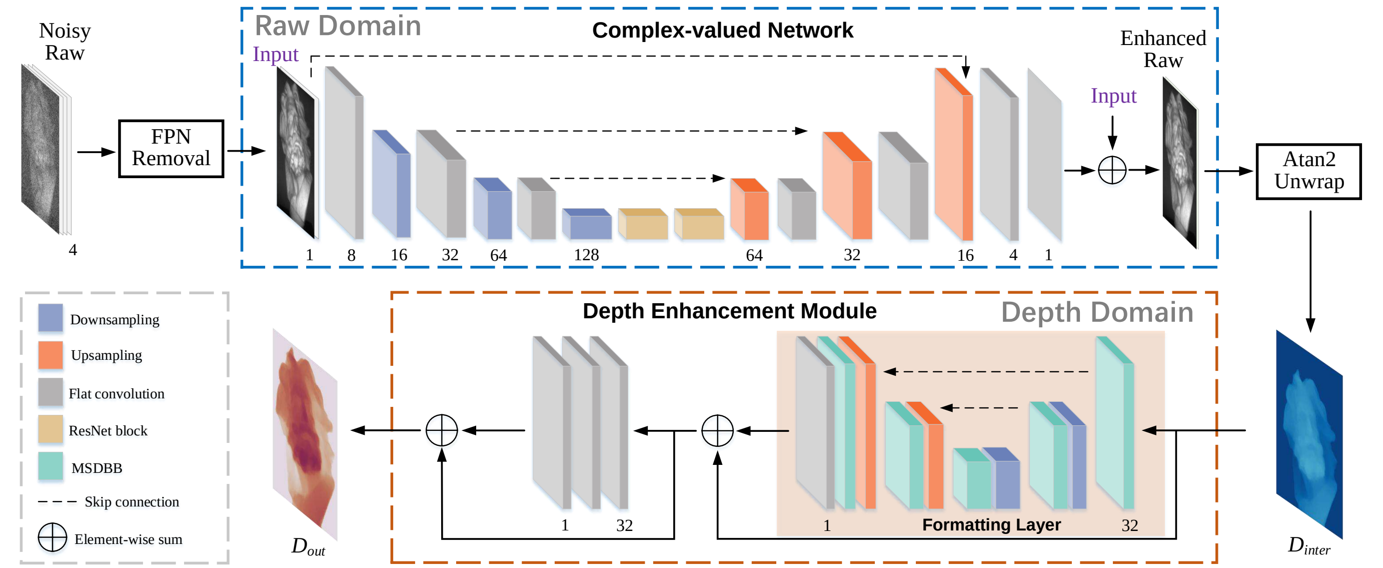


Fig. 5. Illustration of depth restoration framework with cascaded deep network, which consists of three modules, that is FPN removal, complex-valued network, and depth enhancement module.

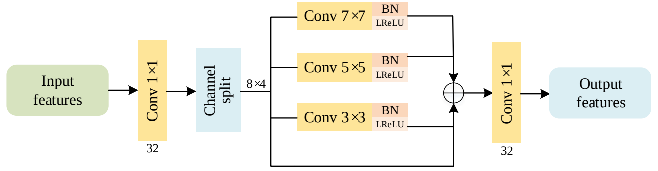


Fig. 6. Structure of the Multi-Scale Depth Enhancement Block (MSDEB), which is used to improve the geometric details of scenes in depth domain.

between the predicted depth map $\tilde{d}$ and its corresponding ground truth d:

$$L_d = (||\tilde{d} - d||_1 + \lambda_d ||\nabla \tilde{d} - \nabla d||_1) \odot M_v \qquad (16)$$

where M_v is the same mask as in Eq.(15) and λ_d is a hyperparameter. In this paper, we use $\lambda_d = 10$, as [20] does.

Total loss. The total loss is obtained as their weighted combination:

$$L_{total} = \lambda_r L_{raw} + L_d \qquad (17)$$

In all of our experiments, we set $\lambda_r = 1$ and $\lambda_d = 10$. All the equations involved in this method are differentiable, which allows them to perform backpropagation in the network.

5 Datasets

To the best of our knowledge, there are no public datasets containing noisy/clean UD-ToF raw measurements and the corresponding ground-truth depth maps for the task. So we create a real dataset by a ToF camera with variable acquisition mode and a synthetic dataset based on noise analysis. Besides, the generalization ability of synthetic data on real scenes will be demonstrated in Section VI. The two datasets will be released upon acceptance on a public repository.

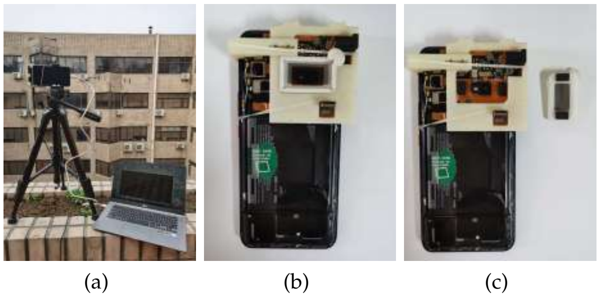

(a) (b) (c)

Fig. 7. Data acquisition device: (a) Real data collection. (b) Under-display mode. (c) Display-free mode.

5.1 Real Dataset

To collect real paired ToF raw measurements, we design a specific range-imaging device with access to raw measurements. Specifically, we use a certain type of mobile phone integrated with a ToF camera based on a Sony IMX 316 sensor for data acquisition. This camera uses dual modulation frequencies of 20 and 100MHz to image the scene, with four-phase steps and a total measurement period of 33ms (30 fps). The laser emits infrared light with a wavelength of 940nm. Under the exposure time of 500us, ToF raw measurements with a resolution of 180×240 are collected for each scene.

Owing to the development of transparent materials, the transmission of infrared light through T-OLEDs is significantly better than before. However, we can not get its transmission efficiency and other parameters from the manufacturer due to confidentiality reasons. As an alternative, by comparing the amplitude of the same scene (e.g., a white wall at 500mm) captured in under-display and display-free mode, we find that the IR transmission is around 80%. Considering that the IR light travels through the T-OLED twice, the ToF sensor only receives about 64% of photons compared with display-free amplitude maps.

With the fixture for fixing the ToF camera and detachable T-OLED display, we can capture the same scenes in the two

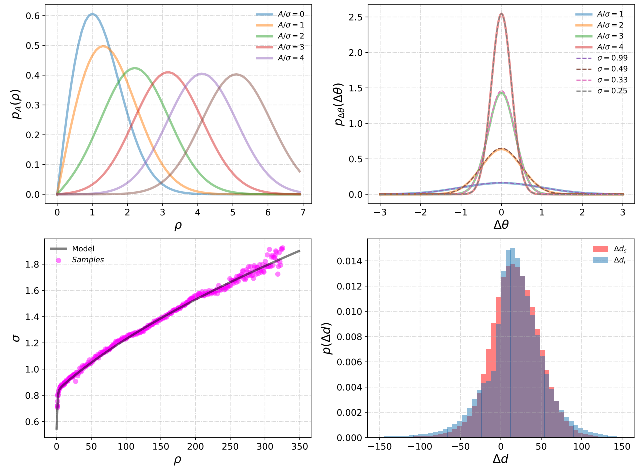


Fig. 8. Model and statistics of synthetic and real ToF data. (a) Distribution of amplitude for varying SNR. (b) Distribution of phase for varying SNR. (c) Captured samples σ and the theoretical model. (d) Error histograms of real and synthetic data for the same scenes.

modes. As shown in Fig. 7b and 7c, the T-OLED display is fixed by weak suction of two magnetic stripes, so the device can switch between the two modes by unplugging and plugging the display. We mount the device on a tripod to avoid camera shake during capturing or mode switching. However, these operations may still cause the camera to shake slightly, resulting in pixel misalignment. We solve this problem by enhanced correlation coefficient maximization (ECC) [59]. Note that since interpolation of raw measurements can change the depth values, only integer-pixel alignment in raw domain is performed.

We build the real UD-ToF dataset mainly for various indoor scenes, namely RUD-TOF, including office, laboratory, meeting room and storeroom. The data are captured in rooms without sunlight interference, so the received signal will not be affected by them. Nevertheless, common indoor lighting is allowed because it does not cover the IR spectrum by the ToF sensor [60]. Materials of the scenes are cluttered, including metals, fabric, plastics, glass and skins with different reflectivity. A total of 1171 scenes are collected for training and 105 scenes for testing. Since this TOF camera is mainly used in mobile phones and other portable devices, limited by the illumination power and the sensor pixel size, we measure the depth of scenes within 2.5m. In addition, the noise is fairly low when we take an average of ten display-free images, thus it can be used as the ground truth.

5.2 Synthetic Dataset

Due to the difficulty of collecting a large-scale real dataset, a synthetic UD-ToF dataset, namely SUD-ToF, is created to be generalized to real scenes. To synthesize ToF raw data, many of the previous works [1], [10], [14], [15], [20] chose to use transient rendering technique from computer graphics, which can well reproduce the entire ToF imaging process. We also opt for synthesizing data along this technical route. Specifically, we utilize the method by [10] as a basis for further research in data synthesis.

After the ideal ToF raw data synthesis process, we follow the degradation model to get the corrupted UD-ToF raw measurements from clean ones. We first apply systematic errors to clean raw measurements, which are then convolved with the PSF k. Next, instead of raw data, the amplitude and phase are both corrupted by noise. In order to determine the PSF that causes light diffraction and image blurring, there is a direct method using a precise collimator optic. However, this method is challenging to apply widely due to its high cost. In addition, parameterized models based on the optical system are not suitable for UD-ToF cameras. Hence, we evaluate the PSF with the help of captured images. It should be noted that all the camera settings, including luminous intensity, field of view (FOV), resolution, and range of measurement, are consistent with the one we use. In this dataset, we synthesize a total of 100K images, of which 10% are randomly selected for testing.

1) Inverse camera correction: Camera correction consists of a series of functions that correct raw measurements and depth maps, including correction of vignetting, pixel delay and temperature drift, so converting clean raw data into the corrupted version is its inverse process. For vignetting and modulation functions, we follow the method by [14] to characterize them. Through the transformation of Eq.(7), it is feasible to augment the raw data with pixel delay and temperature drift. Note that since the trigonometric function terms in Eq.(7) have little influence on the depth accuracy, they are neglected in solving the inverse function for simplicity.

2) PSF Estimation: Computing PSF in UD-ToF imaging can be regarded as a non-blind blur kernel estimation where a normal image and its under-display version are given. We compute the PSF using the approach proposed by Mosleh et al. [61]. Even if the noise level in images is very high, this method can still get the blur kernel, which is in line with our needs. Different from the previous PSF estimation performed on RGB images, we use amplitude maps instead. As shown in Fig.2, the estimated PSF is small due to the fact that the application of Indium Tin Oxide (ITO) in UD-ToF cameras dramatically increases the light transmittance of the screen.

3) Noise analysis: Previous methods mostly add Gaussian noise with the same standard deviation σ empirically to the raw measurements ξ and η. Here, we instead turn them into the polar coordinates to analyze the distribution of the variables ρ and θ, which are corresponding to A and ϕ in Eq. (9). After coordinate conversion, the noise distribution of two variables is no longer Gaussian. The probability distribution of amplitude ρ which is mentioned in communication systems [62] can be represented by:

$$p_A(\rho) = \frac{2\rho}{\sigma^2} e^{-(\rho^2+A^2)/\sigma^2} I_0(\frac{2A\rho}{\sigma^2}) \tag{18}$$

where σ denotes the variance of ToF raw measurements and I_0 denotes the modified zeroth order Bessel function of the first kind. The distribution is called Rice density. See Fig. 8(a) for the distribution with different SNR (A/σ). Only when the SNR is large, that is $A/\sigma \geq 3$, the noise distribution can be treated as a Gaussian distribution.

Since the noise in amplitude maps is signal-dependent, the amplitude can be synthesized only by getting the relationship between the amplitude and its variance, which is derived in [63]

$$\sigma_A = \sigma^2 + A^2 - \frac{\pi\sigma^2}{4} L^2(-\frac{\mathrm{SNR}^2}{2}) \tag{19}$$

where $L(x) = e^{x/2}[(1-x)\mathrm{I}_0(-\frac{x}{2}) - x\mathrm{I}_1(-\frac{x}{2})]$. Fig. 8(c) shows that the differences between the captured data and the model predicted values are small, thus verifying the effectiveness of the model. Based on the two models illustrated in Fig. 8(a) and Fig. 8(c), the amplitude maps can be simulated.

Phase images, as shown in Eq. (4), are obtained from the imaginary component η and real component ξ by computing the arctangent function of their ratio. Since this is a non-linear mapping, the distribution of phase noise does not follow a Gaussian distribution either. According to [62], the phase noise distribution $\Delta\theta = \theta - \tilde{\theta}$ is given as:

$$\begin{aligned} p_{\Delta\phi}(\Delta\theta) =& \frac{1}{2\pi} e^{-A^2/\sigma^2} [1 + \frac{2A}{\sigma}\sqrt{\pi}\cos\Delta\theta e^{A^2\cos^2\Delta\theta/\sigma^2} \\ & \cdot \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\frac{\sqrt{2}A\cos\Delta\theta}{\sigma}} e^{-x^2/2} dx] \end{aligned} \tag{20}$$

Although this expression about phase noise $\Delta\theta$ is very complicated, its corresponding curves have some interesting characteristics. When SNR is large, that is $A \gg \sigma$, the integral term in Eq. (20) is approximately equal to 1. Besides, the constant 1 in the square bracket is much smaller than the second term, so Eq. (20) can be simplified to

$$\begin{aligned} p_{\Delta\phi}(\Delta\theta) \approx& \frac{A\cos\Delta\theta}{\sigma\sqrt{\pi}} \exp\left[-A^2(1-\cos^2\Delta\theta)/2\sigma^2\right] \\ \approx& \frac{1}{\sqrt{\pi}\sigma/A} \exp\left[-\frac{\Delta\theta^2}{(\sigma/A)^2}\right] \end{aligned} \tag{21}$$

Then this model becomes the familiar Gaussian distribution. Instead, when SNR is small, this distribution can also be approximated as Gaussian, as shown in Fig. 8(b). Therefore, no matter what SNR is, this model can be approximated by Gaussian. Note that pixels will be masked out if their $\mathrm{SNR} = 0$, so we do not analyze this situation here. Considering its generalization to real data and complexity, we fix the phase noise variance to the maximum average error of the measured depth maps.

Based on the above analysis, we exhibit the error distribution of real samples Δd_r and synthetic data Δd_s, as shown in Fig. 8(d). There are 20 scenes with 50 shots in each. This figure shows that the noise characteristics of the synthetic data and the real ToF raw measurements are identical.

5.3 FLAT Dataset

In addition to the previous two datasets, we also select the FLAT dataset [14] for training and evaluation. Transient rendering technique is also utilized to mimic the ToF imaging process, including the generation of various noises. This dataset contains 1929 static scenes, which provides clean/noisy raw measurement pairs. Unlike our datasets specifications, this FLAT dataset is synthesized based on the characteristics of Kinect v2. This camera utilizes three modulation frequencies, each of which is sampled three times in a cycle, generating a total of nine raw measurements. The images have a spatial resolution of 424×512 and a working range of [0.5m, 6m].

6 Experiments and Results

In this section, we first present the training details. Then we quantitatively and qualitatively evaluate our model and compare it against state-of-the-art approaches in ToF depth restoration. Afterward, we conduct a series of ablation studies to validate the proposed method. Finally, we opt for face recognition to verify the effect of our method on downstream applications.

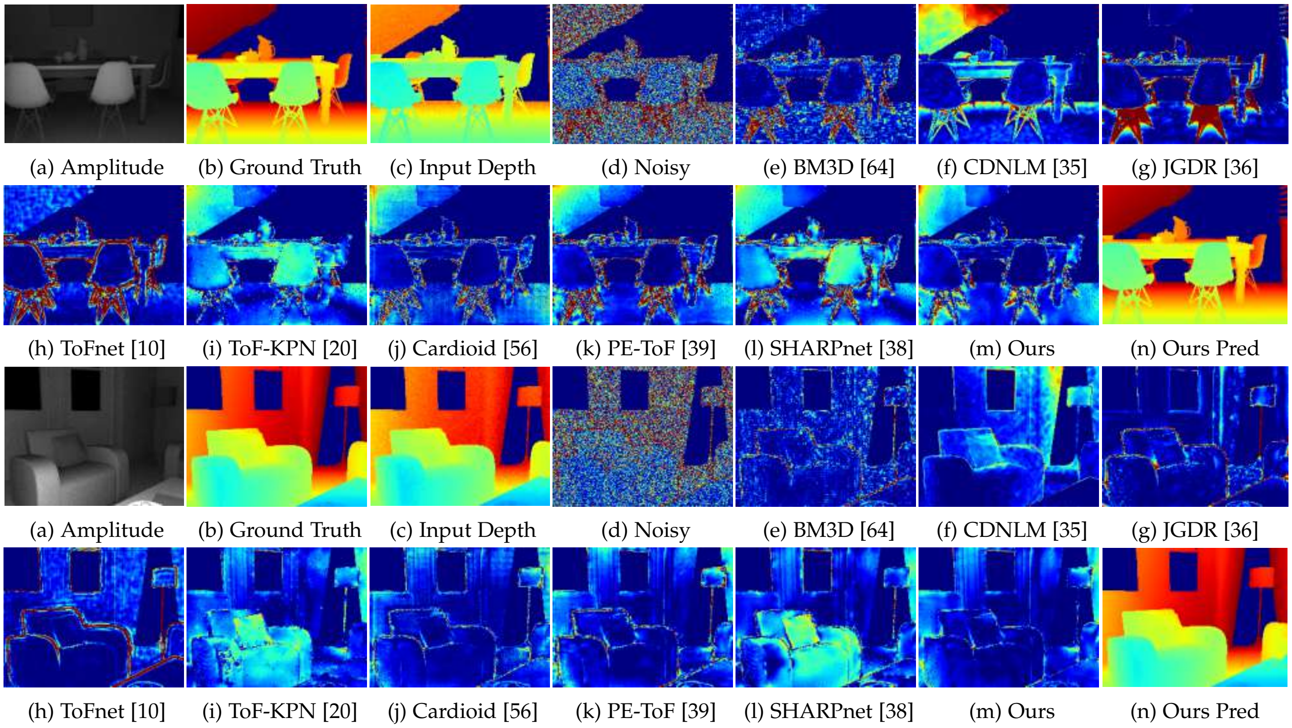
6.1 Implementation Details

We train our network on the SUD-ToF dataset and the RUD-ToF dataset for 250 and 1000 epochs, respectively. We crop out 176×240 patches on the images with a full-resolution of 180×240 to facilitate downsampling. The initial learning rates of the two subnets are both set to be 10^{-4}, and then they decay by 0.7 after every 200 epochs and 40 epochs, respectively. For the FLAT dataset, we use learning rates of 10^{-3} and 10^{-4} for the subnets, respectively. The network is trained for 2000 epochs with a decay rate of 0.5 after every 400 epochs. In all cases, we optimize utilizing Adam with a batch size of 16. The proposed network is implemented in Pytorch framework and trained on Nvidia RTX 3090 GPU.

6.2 Comparisons with State-of-the-art Methods

To evaluate the effectiveness of the algorithm, we compare it with other SOTA algorithms presented in recent years, including traditional methods BM3D [64], CDNLM [35], and JGDR [36], and learning-based methods ToFnet [10], ToF-KPN [20], Cardioid [56], PE-ToF [39], SHARPnet [38]. Among the approaches, BM3D and JGDR are general algorithms for depth restoration, and the others are designed for ToF depth restoration. For comparison fairness, we only take ToF raw depth and amplitude maps instead of RGB-D information as the input of ToF-KPN. In addition, the structure and parameters of Cardioid are the same as the proposed method except for the activation function. Furthermore, all the learning-based comparison methods employ parameters recommended by the original paper in training and evaluation. The input of these approaches is their respective modalities of UD-ToF raw measurements or depth maps after FPN removal (see Table 1, upper part).

In the following, our proposed network is compared with the methods in two aspects: (1) The state-of-the-art methods are comprehensively tested on the SUD-ToF and RUD-ToF datasets to evaluate their performance on UD-ToF depth restoration. (2) In order to further access the generalization ability of the network in other deep restoration tasks like denoising, it is evaluated on the FLAT dataset. In all the experiments, we take the commonly used MAE (Mean Absolute Error) and RMSE (Root Mean Squared Error) as the evaluation metrics. Besides, the proportion δ_{th} of pixels within a specific relative error range to total pixels is adopted as our metric also. Considering the high accuracy of the camera at short range, we follow [65] to set $th \in \{1.02, 1.05, 1.10\}$.



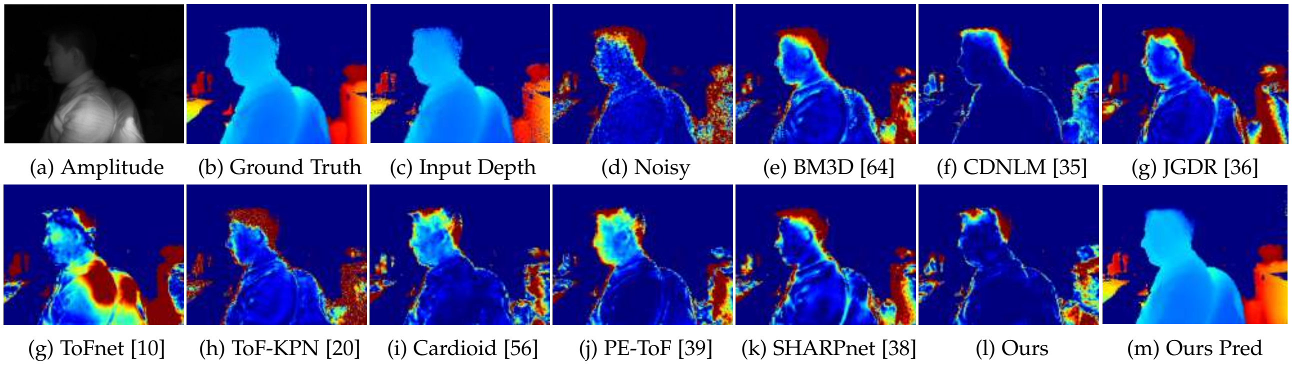
(a) Amplitude (b) Ground Truth (c) Input Depth (d) Noisy (e) BM3D [64] (f) CDNLM [35] (g) JGDR [36]

(h) ToFnet [10] (i) ToF-KPN [20] (j) Cardioid [56] (k) PE-ToF [39] (l) SHARPnet [38] (m) Ours (n) Ours Pred

(a) Amplitude (b) Ground Truth (c) Input Depth (d) Noisy (e) BM3D [64] (f) CDNLM [35] (g) JGDR [36]

(h) ToFnet [10] (i) ToF-KPN [20] (j) Cardioid [56] (k) PE-ToF [39] (l) SHARPnet [38] (m) Ours (n) Ours Pred

Fig. 9. Qualitative results on SUD-ToF data. From left to right, the first four columns are (a) Amplitude maps, (b) Ground truth, (c) Input depth maps and (d) Error maps from noisy depth maps; (e)-(l) are error maps of selected methods; the last two columns are (m) the error map and (n) the prediction generated from our method.



(a) Amplitude (b) Ground Truth (c) Input Depth (d) Noisy (e) BM3D [64] (f) CDNLM [35] (g) JGDR [36]

(g) ToFnet [10] (h) ToF-KPN [20] (i) Cardioid [56] (j) PE-ToF [39] (k) SHARPnet [38] (l) Ours (m) Ours Pred

Fig. 10. Qualitative results on RUD-ToF data. From left to right, the first four columns are (a) Amplitude maps, (b) Ground truth, (c) Input depth maps and (d) Error maps from noisy depth maps; (e)-(l) are error maps of selected methods; the last two columns are (m) the error map and (n) the prediction generated from our method.

TABLE 1

Comparison with the state-of-the-art on the SUD-ToF and RUD-ToF dataset. The best and second best results are marked in red and blue respectively. The direction of arrows in metrics represents the better trends of indicators.

Dataset	Metrics	BM3D [64]	CDNLM [35]	JGDR [36]	ToFnet [10]	ToF-KPN [20]	Cardioid [56]	PE-ToF [39]	SHARPnet [38]	ours
	Input	Depth	Raw	Depth	Raw	Depth	Raw	Raw	Depth	Raw
SUD-ToF	MAE/mm↓	8.31	33.23	9.14	10.34	13.39	8.91	9.77	14.84	8.88
	RMSE/mm↓	15.08	48.43	34.43	28.28	21.05	13.02	15.92	23.00	11.50
	$\delta_{1.02}$ ↑	94.45	51.57	95.40	92.57	86.79	95.84	95.23	80.41	99.09
	$\delta_{1.05}$ ↑	98.73	87.64	97.82	97.01	98.77	99.45	98.76	94.22	99.70
	$\delta_{1.10}$ ↑	99.56	97.01	98.66	98.29	99.57	99.89	99.53	96.82	99.94
RUD-ToF	MAE/mm↓	21.37	42.38	37.05	25.13	27.60	20.90	21.22	24.63	17.29
	RMSE/mm↓	48.01	121.61	71.36	61.50	49.94	35.20	48.76	43.68	31.11
	$\delta_{1.02}$ ↑	65.59	63.99	48.04	66.23	61.65	58.61	62.03	56.04	70.13
	$\delta_{1.05}$ ↑	87.28	84.71	80.40	87.33	81.57	86.45	87.04	79.25	90.01
	$\delta_{1.10}$ ↑	95.45	92.09	91.79	95.17	89.90	95.64	95.39	90.01	96.74

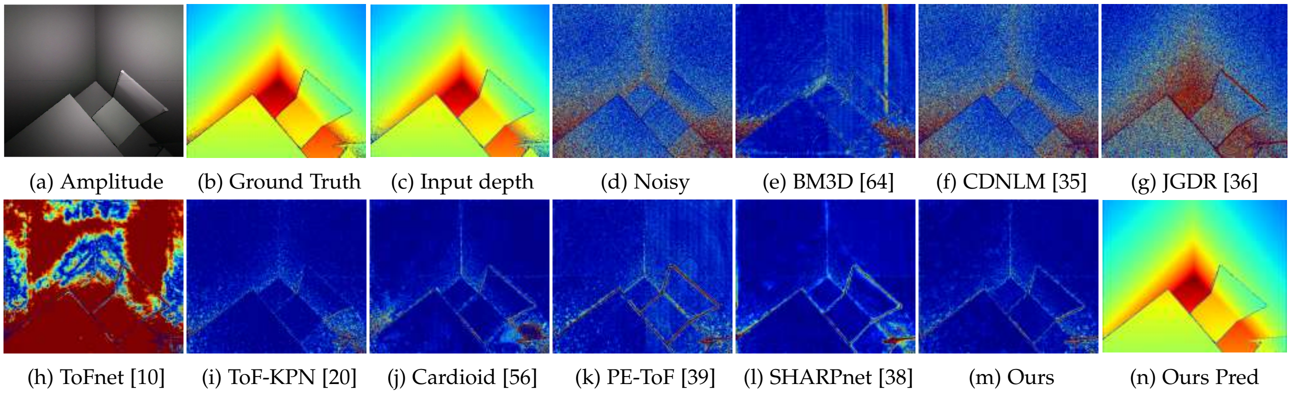


Fig. 11. Qualitative results on the FLAT dataset. From left to right, the first four columns are (a) Amplitude maps, (b) Ground truth, (c) Input depth map and (d) Error map from noisy depth map; (e)-(l) are error maps of selected methods; the last two columns are (m) the error map and (n) the prediction generated from our method.

1) UD-ToF Depth Restoration. We first quantitatively and qualitatively evaluate the performance of the proposed method against the state-of-the-arts on the SUD-ToF dataset, as shown in Table 1 and Fig. 9. In general, the traditional model-based methods perform relatively poorly compared to the data-driven approaches. CDNLM [35] does not perform well on our task because fixed parameters cannot handle all cases in the dataset. Furthermore, some outliers with extremely low confidence might significantly bias the depth estimation, as shown in Fig. 9(f). Despite the fact that BM3D [64] performs best on MAE, it is considerably inferior to our results on other metrics. The two approaches ToF-KPN [20] and SHARPnet [38], which use depth as an input, outperform traditional methods. However, the irreversible information lost in the process of mapping raw data to depth makes it difficult to achieve optimal depth estimation (see Fig. 9(i) and 9(l)). In contrast, the restored results by ToFnet [10] and PE-ToF [39] with raw data as input are better than the former two algorithms, as shown in Fig. 9(h) and 9(k). However, the two networks show deficiencies in restoring structural details (e.g., edges). As a variant of our method, the overall performance of Cardioid [56] ranks second in tests on the SUD-ToF dataset. Table 1 and Fig. 9(n) shows that our method achieve favorable performance against the state-of-the-art methods.

On the RUD-ToF dataset, we compare our cascaded network against state-of-the-arts in the same metrics. As seen in the bottom half of Table 1, the data-driven approaches outperform the model-based methods. Note that the performance of traditional methods is still less effective than that of deep learning approaches. Without the assistance of corresponding color images, JGDR [36] performs poorly on both datasets. Moreover, Cardioid [56] also shows good performance in UD-ToF restoration. Fig. 10 shows an example of human body restoration where our method performs favorably against state-of-the-art depth restoration methods.

2) Generalization Ability on Denoising. To further evaluate the robustness of our network, we conduct more comparisons on another public dataset. Specifically, we compare the denoising performance of our method with the state-of-the-arts on the FLAT dataset. As reported in Table 2 and Fig. 11, our method generally is superior to the other state-of-the-art methods. The traditional method JGDR [36] also has excellent denoising ability, while the deep learning methods still have obvious advantages over the traditional methods. However, it is worth noting that ToFnet [10] fails to restore accurate depth maps according to the results on MAE and RMSE. The main reason is that the TOF camera adopts more modulation frequencies, and one of the frequencies is not modulated by sinusoidal signal [14], which makes ToFnet unable to learn features effectively. We believe that more learning samples or larger models will help to solve this problem. Furthermore, the approaches, ToF-KPN [20] and SHARPnet [38], utilizing depth as input also produce good results due to the low noise level in this dataset. In a word, our method performs favorably against the state-of-the-arts even for different tasks and camera settings.

6.3 Ablation Study

In this section, we analyze the presented approach from six different perspectives. All the ablation studies are conducted on the SUD-ToF and RUD-ToF datasets.

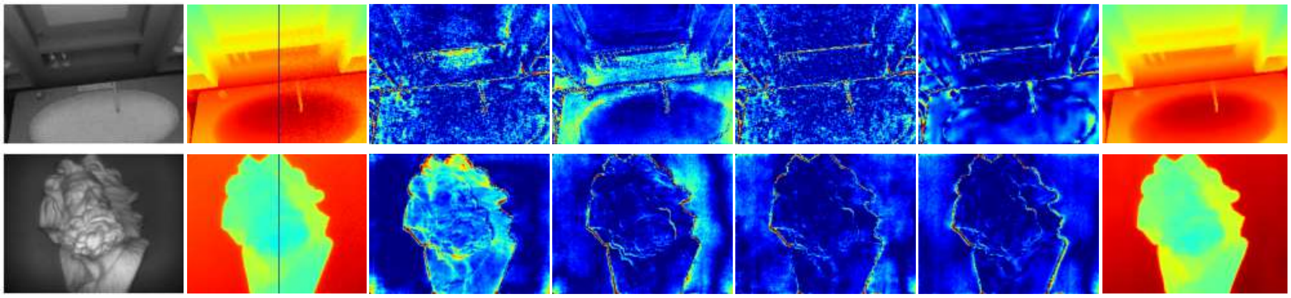
1) Effect of FPN: In this part, we show how FPN affects the performance of UD-ToF reconstruction. We compare the quality of output depth maps by our network with and without FPN removal. The quantitative results are shown in Table 3, which demonstrates that our method with FPN removal achieves better performance in all metrics. We believe this is because the network is not forced to learn the model of FPN instead of leveraging the fact that the offset of sensed intensities is spatial fixed and known.

Different from [34], even if the intensities received by the sensor are high, the influence of FPN can not be excluded from consideration, especially in the area illuminated by the reflected light of the T-OLED display. Fig. 12 shows examples from the SUD-ToF and RUD-ToF datasets. Although the object in the figure is close to the camera, the method without FPN removal cannot accurately restore the depth. Therefore, FPN removal is an essential postcapturing operation for UD-ToF reconstruction.

2) Validation on Complex-valued Network: To efficiently enhance the ToF raw measurements, we introduce the complex-valued network (referred to as *Complex*). According to the analysis in Section IV-B, our network can be viewed as a variant of a real network that exploits prior

TABLE 2
Comparison with the state-of-the-art on the FLAT dataset. The best and second best results are marked in red and blue respectively. The direction of arrows in metrics represents the better trends of indicators.

Dataset	Metrics	BM3D [64]	CDNLM [35]	JGDR [36]	ToFnet [10]	ToF-KPN [20]	Cardioid [56]	PE-ToF [39]	SHARPnet [38]	ours
FLAT	MAE/mm↓	9.89	13.86	8.86	54.33	4.65	6.74	7.93	4.62	4.41
	RMSE/mm↓	18.95	21.00	45.78	74.99	12.83	19.94	32.60	10.26	8.23
	$\delta_{1.02}$ ↑	95.57	93.87	97.90	42.40	99.65	98.29	98.02	99.46	99.52
	$\delta_{1.05}$ ↑	99.32	99.74	99.16	83.87	99.94	99.81	99.40	99.90	99.95
	$\delta_{1.10}$ ↑	99.76	99.98	99.58	99.12	99.96	99.95	99.72	99.97	99.99



(a) Amplitude (b) GT/Input (c) w FPN (d) *Real* (e) *Complex* (f) Ours (g) Ours Pred

Fig. 12. Visual comparison of results on the SUD-ToF (1st row) and RUD-ToF (2nd row) dataset, respectively. The methods with FPN and *Real* provide undesirable results. Further, *Complex* still suffers from shot noise. (a) Amplitude maps, (b) ground truth on the left and input depth maps on the right, (c)-(f) error maps encoded by color map 'jet', (g) our prediction.

TABLE 3
Effectiveness of the proposed approach on the SUD-ToF and RUD-ToF datasets.

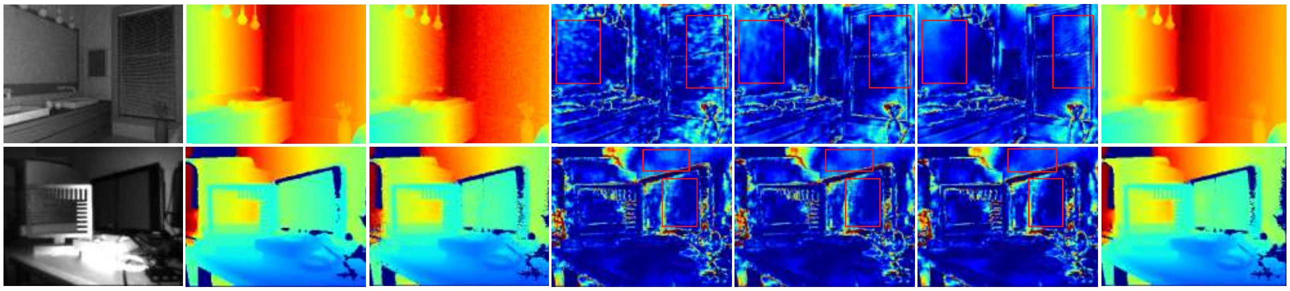
Dataset	Approach	MAE/mm↓	RMSE/mm↓	1.02↑	1.05↑	1.10↑
SUD-ToF	w FPN	10.72	14.99	91.56	99.28	99.86
	Real	11.26	14.84	89.26	99.45	99.86
	Complex	9.77	13.81	93.71	99.37	99.86
	Flat	9.07	12.07	96.69	99.53	99.88
	DUnet	9.19	12.77	96.06	99.36	99.84
	Ours	8.75	11.90	96.83	99.58	99.89
RUD-ToF	w FPN	20.05	34.04	59.39	87.42	96.13
	Real	21.20	35.98	58.68	85.16	94.82
	Complex	19.74	34.64	63.49	88.68	96.13
	Flat	19.50	34.59	65.37	88.68	96.01
	DUnet	19.52	34.55	65.50	88.69	96.00
	Ours	18.71	32.12	66.42	88.83	96.27

knowledge. We compare the evaluation results of the complex network and its corresponding real counterpart (referred to as *Real*) setting with nearly the same parameter size. For the real network, we take ξ and η as two input channels, and the network ignores the inherent relationship between them, which requires the network to learn in training. Note that the subnet in depth domain is not considered in this experiment to distinguish the two networks' performance better.

The comparison results are reported in Table 3. On the RUD-ToF dataset, it can be seen that our complex network yields better results in all metrics than the conventional real network. On the SUD-ToF dataset, the real network marginally outperforms our method in $\sigma_{1.05}$, but our method is still better overall, demonstrating the power of complex representation. In Fig. 12, we show that the complex model produces depth maps with less error, whereas the real alternative fails to restore the depth accurately. In our perspective, a primary difference between the two networks is weight sharing in convolution. The scalar multiplication with a 2×2 weight matrix in the real network can obtain four freedom degrees. Instead, for the complex version, the multiplication has only two degrees of freedom, scaling and rotation, which is easier to learn than scalar multiplication.

3) Effectiveness of the Subnet in Depth Domain: The complex network can greatly improve ToF raw data so as the depth maps. In this case, we need further to verify the necessity of the subnet in depth domain. We use the variant of our method without the subnet in depth domain to perform the comparison. The approach *Complex* in the last experiment is what we need here. From the results shown in Table 3, we can see that the presented method outperforms the *Complex* by a large margin. Especially in $\sigma_{1.10}$, our model with the subnet in depth domain exceeds its variant by 3.12% on the SUD-ToF dataset and 2.93% on the RUD-ToF dataset, which means fewer outliers exist after depth restoration. Fig. 12 further shows that following the subnet processing, shot noise is greatly reduced. It confirms that the subnet is helpful to the UD-ToF restoration. The reason is that, while the intensity of the raw data has been restored, there is still high-frequency noise in the depth map that needs to be removed, and the network can maintain as many high-frequency features as feasible while denoising.

4) Effectiveness of MSDEB: In this section, to validate the effectiveness of the MSDEB, it is compared with two other popular networks, that is flat convolution (referred to as *Flat*) and Unet for depth restoration (referred to as *DUnet*). The Formnet [23] uses the flat convolution to restore depth maps and achieves desirable performance. Since Unet



(a) Amplitude (b) Ground Truth (c) Input Depth (d) *Flat* (e) *DUnet* (f) *Ours* (g) Ours Pred

Fig. 13. Effectiveness of the MSDEB. Flat convolution and *DUnet* are ineffective at extracting features from homogeneous regions. The proposed network using MSDEB, on the other hand, can restore the depth of the low-frequency area well. (a) Amplitude maps, (b) ground depth maps, (c) Input depth maps, (e)-(f) error maps encoded by color map 'jet', (f) our prediction.

structure has been proved its power in regression problems, we choose it as one latent alternative to the formatting layer. We can see from Table 3 that our method with MSDEBs generates better outcomes than the other two networks. On the RUD-ToF dataset, our results demonstrate significant improvements in RMSE and $\sigma_{1.02}$, indicating that our proposed module can estimate more accurate depth maps. Furthermore, Fig. 13 shows two examples to compare their performance qualitatively. The results by *Flat* and *DUnet* tend to produce more high-frequency noise in homogeneous regions (red rectangle in the figure). The key reason is that the formatting layer tries to restore the homogeneous regions in depth maps, and our method extracts image characteristics on multi-scale and receptive fields with different sizes, which is conducive for depth detail boosting.

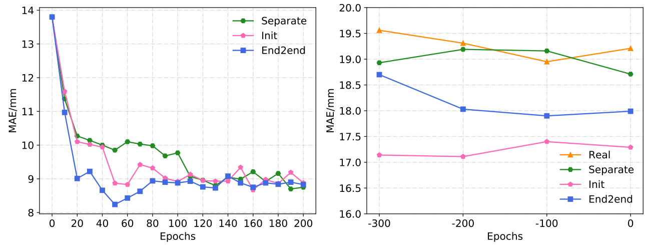
5) Domain Gap between Synthetic and Real: As mentioned in Section V-B, since collecting large-scale real data is time-consuming and labor-intensive, we build a synthetic dataset that is expected to have the same noise characteristics as the real dataset. In order to demonstrate the effectiveness of the noise model, our evaluation is carried out in the following three cases: 1) we directly employ the network parameters trained on the SUD-ToF dataset to perform evaluation on the RUD-ToF dataset (referred to as *S2R*). 2) Following case 1), fine-tuning with 200 epochs on the RUD-ToF dataset is conducted, which takes about only 80 minutes (referred to as *S2RT*). 3) Taking case 1) as pre-training, we then train the network on the RUD-ToF dataset for 1000 epochs (referred to as *Fully*). As illustrated in Table 4, the fully trained network outperforms the other two training strategies, which is also in line with our expectation. Further, we can still consider that the noise distribution of the two datasets is similar. It should be noted that the SUD-TOF dataset adopts ideal intrinsic parameters, which is different from that in the RUD-TOF dataset. In fact, even the results of *S2R* are very advantageous in comparison with other SOTA methods, such as ranking 2nd in both MAE and RMSE. As for *S2RT*, it achieves comparable performance in one-fifth the time to train *Fully*. Therefore, the results indicate that the domain gap between synthetic and real data is small. In practical applications, the specifications of mass-produced camera modules differ slightly, and our SUD-ToF dataset may greatly reduce the workload of obtaining real data.

6) Investigation of Training Fashion: For a fair comparison, we train the two subnets separately and compare their performance in previous experiments. Different training strategies, on the other hand, will have a substantial impact on the results. In this part, we estimate the increase in model accuracy caused by three training fashions on the SUD-ToF dataset: 1) the two subnets are trained separately (referred to as *Separate*). 2) The whole network is trained in an end-to-end fashion (referred to as *End2end*). 3) Before case 2, we first train the complex network alone as pre-training (referred to as *Init*). On the RUD-ToF dataset, we use the parameters trained on the SUD-ToF dataset for initialization, then train with the same procedures as above. In addition, we also implement another strategy, which is training separately just on the RUD-ToF dataset (referred to as *Real only*).

Table 5 reports the experimental results of different training strategies. It can be seen that training separately has a limited improvement on the model accuracy on both datasets. When the model is pre-trained, the result on the RUD-ToF dataset is better, showing how the SUD-ToF dataset can benefit real-world restoration. The end-to-end training further improves the model accuracy, which reduces the difficulty of tuning parameters either. For the approach *End2end* and *Init*, they perform better on the SUD-ToF and RUD-ToF datasets, respectively. Fig. 14 shows that different training manners perform similarly on the SUD-ToF dataset, whereas the method *Init* performs considerably

TABLE 4
Domain Gap Analysis between real and synthetic.

Dataset	Approach	MAE/mm↓	RMSE/mm↓	1.02↑	1.05↑	1.10↑
RUD-ToF	*S2R*	21.15	37.18	65.37	86.86	94.93
	S2RT	18.73	32.38	66.25	88.73	96.21
	Fully	18.71	32.12	66.42	88.83	96.27



(a) (b)

Fig. 14. Investigation of different training fashions. (a) Training on the SUD-ToF dataset. (b) Training on the RUD-ToF dataset.

better than the others on the RUD-ToF dataset. In general, the *Init* is more advantageous to be the final training method.

TABLE 5
Influence of Different Training Strategies.

Dataset	Approach	MAE/mm↓	RMSE/mm↓	1.02↑	1.05↑	1.10↑
SUD-ToF	*Separate*	8.75	11.90	96.83	99.58	99.89
	End2end	8.83	11.15	97.69	99.75	99.95
	Init	8.88	11.50	97.09	99.70	99.94
RUD-ToF	*Real only*	19.21	34.32	67.42	88.86	96.06
	Separate	18.71	32.12	66.42	88.83	96.27
	End2end	17.99	32.75	68.28	90.16	96.52
	Init	17.29	31.11	70.13	90.01	96.74

6.4 Downstream Application

In this section, we show the performance improvement of the UD-ToF camera restoration algorithm for downstream applications, such as face recognition. In turn, the depth-based face recognition rate can be a metric to evaluate the depth quality as well. In this experiment, we test the performance of face recognition under three conditions: display-free, before and after restoration under display. To this end, we collect a face dataset as test data in under-display and display-free mode, respectively. There are 20 people in each mode, and 100 images with different poses are collected for each person. Using the MobileFaceNet [66] as backbone, we compare their accuracy of face recognition. Moreover, the model is pre-trained on CASIA 3D Face [67] which contains 4624 images for 123 identities.

Table 6 shows the results of face recognition on our face dataset. Since random image pairing is necessary to carry out this evaluation, we perform three experiments using three different randomly paired sets. This strategy allows us to show the consistent improvement yielded by our method, independently of the specific pairing. We can see that the accuracy is increased from about 72% to over 92% after depth restoration. The results after restoration are very close to the display-free depth, reaching 99.90% accuracy on average. Note that our result is even better in the second test. However, when a face is more than $600mm$ away from the UD-ToF camera, the performance of display-free depth is significantly better than ours.

The examples of 3D face reconstruction using screened Poisson reconstruction [68] are illustrated in Fig. 15. Our method can recover high-fidelity details of a face from a single frame with inaccurate distance and high-frequency noise.

In the tests, the whole processing procedure includes FPN removal, two subnets inference and raw-to-depth mapping. Our unoptimized Python code takes an average of 14ms to run this procedure.

7 Conclusion

In this paper, we presented a learning-based approach for jointly solving denoising, deblurring, and raw measurements enhancement in UD-ToF imaging. The two critical components of UD-ToF depth restoration are the FPN removal tailored for UD-ToF cameras and the cascaded deep network work. One subnet, which works in raw domain, utilizes the inherent physical relationship between raw measurements to conduct complex convolution, while the other working in depth domain aims at depth refinement with the proposed MSDEB. To achieve supervised learning, we devised a data acquisition device for real paired images collection and built a real dataset, RUD-ToF. Additionally, a large-scale synthetic dataset named SUD-ToF based on the noise model was established to enhance the model generalization. A vast series of experiments demonstrate that the proposed method performs favorably against the state-of-the-art approaches for UD-ToF depth restoration. We intend to fuse information from more sensors, such as structured light depth sensors or RGB cameras, in the future to achieve more accurate depth reconstruction.

TABLE 6
Quantitative evaluation of the proposed method in face recognition using MobileFaceNet.

Experiment No.	TAR@FAR=1e-2		
	UD-ToF depth	After restoration	Display-free
1	71.98	92.48	92.51
2	72.20	92.90	92.79
3	71.85	92.50	92.68

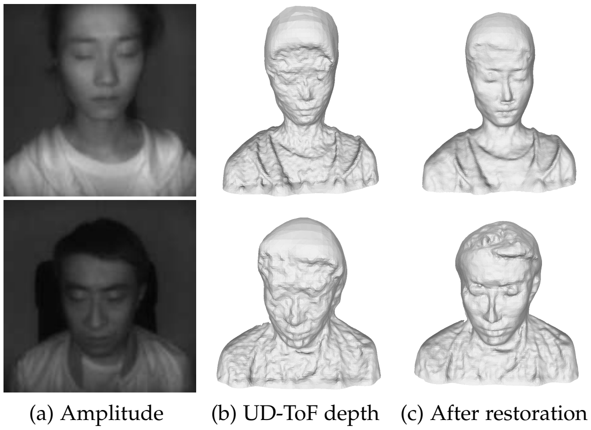

(a) Amplitude (b) UD-ToF depth (c) After restoration

Fig. 15. Qualitative results of 3D face reconstruction using screened Poisson reconstruction.

Acknowledgments

We acknowledge supports by the National Natural Science Foundation of China (No.61627811), Natural Science Foundation of Shaanxi Province(No. 2021JZ-04), Joint project of key R&D universities in Shaanxi Province(No.2021GXLH-Z-093), and the China Scholarship Council (CSC). We also thank Samsung R&D Institute China Xi'an (SRCX) for the support, as well as Sunny Optical, vivo Mobile and QingYue Optoelectronics for hardware donation.

References

[1] A. Adam, C. Dann, O. Yair, S. Mazor, and S. Nowozin, "Bayesian time-of-flight for realtime shape, illumination and albedo," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 5, pp. 851–864, 2016.

[2] S. Shrestha, F. Heide, W. Heidrich, and G. Wetzstein, "Computational imaging with multi-camera time-of-flight systems," *ACM Transactions on Graphics (ToG)*, vol. 35, no. 4, pp. 1–11, 2016.

[3] V. D. J. Evans, X. Jiang, A. E. Rubin, M. Hershenson, and X. Miao, "Optical sensors disposed beneath the display of an electronic device," 2017, uS Patent App. 15/336,620.

[4] Y. Zhou, D. Ren, N. Emerton, S. Lim, and T. Large, "Image restoration for under-display camera," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 9179–9188.

[5] P. Zanuttigh, G. Marin, C. Dal Mutto, F. Dominio, L. Minto, and G. M. Cortelazzo, "Time-of-flight and structured light depth cameras," *Technology and Applications*, pp. 978–3, 2016.

[6] J. Illade-Quinteiro, V. M. Brea, P. López, D. Cabello, and G. Doménech-Asensi, "Distance measurement error in time-of-flight sensors due to shot noise," *Sensors*, vol. 15, no. 3, pp. 4624–4642, 2015.

[7] D. Freedman, Y. Smolin, E. Krupka, I. Leichter, and M. Schmidt, "Sra: Fast removal of general multipath for tof sensors," in *European Conference on Computer Vision*. Springer, 2014, pp. 234–249.

[8] L. Xiao, F. Heide, M. O'Toole, A. Kolb, M. B. Hullin, K. Kutulakos, and W. Heidrich, "Defocus deblurring and superresolution for time-of-flight depth cameras," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 2376–2384.

[9] V. Sterzentsenko, L. Saroglou, A. Chatzitofis, S. Thermos, N. Zioulis, A. Doumanoglou, D. Zarpalas, and P. Daras, "Self-supervised deep depth denoising," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 1242–1251.

[10] S. Su, F. Heide, G. Wetzstein, and W. Heidrich, "Deep end-to-end time-of-flight imaging," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 6383–6392.

[11] S. Yan, C. Wu, L. Wang, F. Xu, L. An, K. Guo, and Y. Liu, "Ddrnet: Depth map denoising and refinement for consumer depth cameras using cascaded cnns," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 151–167.

[12] X. Tao, H. Gao, X. Shen, J. Wang, and J. Jia, "Scale-recurrent network for deep image deblurring," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 8174–8182.

[13] S. Zhou, J. Zhang, W. Zuo, H. Xie, J. Pan, and J. S. Ren, "Davanet: Stereo deblurring with view aggregation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 10 996–11 005.

[14] Q. Guo, I. Frosio, O. Gallo, T. Zickler, and J. Kautz, "Tackling 3d tof artifacts through learning and the flat dataset," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 368–383.

[15] J. Marco, Q. Hernandez, A. Munoz, Y. Dong, A. Jarabo, M. H. Kim, X. Tong, and D. Gutierrez, "Deeptof: off-the-shelf real-time correction of multipath interference in time-of-flight imaging," *ACM Transactions on Graphics (ToG)*, vol. 36, no. 6, pp. 1–12, 2017.

[16] G. Agresti, H. Schafer, P. Sartor, Y. Incesu, and P. Zanuttigh, "Unsupervised domain adaptation of deep networks for tof depth refinement," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[17] G. Riegler, M. Rüther, and H. Bischof, "Atgv-net: Accurate depth super-resolution," in *European conference on computer vision*. Springer, 2016, pp. 268–284.

[18] X. Song, Y. Dai, and X. Qin, "Deep depth super-resolution: Learning depth super-resolution using deep convolutional neural network," in *Asian conference on computer vision*. Springer, 2016, pp. 360–376.

[19] G. Agresti and P. Zanuttigh, "Deep learning for multi-path error removal in tof sensors," in *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, 2018, pp. 0–0.

[20] D. Qiu, J. Pang, W. Sun, and C. Yang, "Deep end-to-end alignment and refinement for time-of-flight rgb-d module," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 9994–10 003.

[21] Y. Kim, Y. J. Koh, C. Lee, S. Kim, and C.-S. Kim, "Dark image enhancement based onpairwise target contrast and multi-scale detail boosting," in *2015 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2015, pp. 1404–1408.

[22] C. Chen, Q. Chen, J. Xu, and V. Koltun, "Learning to see in the dark," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 3291–3300.

[23] J. Jiao, W.-C. Tu, D. Liu, S. He, R. W. Lau, and T. S. Huang, "Formnet: Formatted learning for image restoration," *IEEE Transactions on Image Processing*, vol. 29, pp. 6302–6314, 2020.

[24] Q. Yang, Y. Liu, J. Tang, and T. Ku, "Residual and dense unet for under-display camera restoration," in *Computer Vision – ECCV 2020 Workshops*, A. Bartoli and A. Fusiello, Eds. Cham: Springer International Publishing, 2020, pp. 398–408.

[25] Y. Zhou, M. Kwan, K. Tolentino, N. Emerton, S. Lim, T. Large, L. Fu, Z. Pan, B. Li, Q. Yang, Y. Liu, J. Tang, T. Ku, S. Ma, B. Hu, J. Wang, D. Puthussery, P. S. Hrishikesh, M. Kuriakose, C. V. Jiji, V. Sundar, S. Hegde, D. Kothandaraman, K. Mitra, A. Jassal, N. A. Shah, S. Nathan, N. A. E. Rahel, D. Chen, S. Nie, S. Yin, C. Ma, H. Wang, T. Zhao, S. Zhao, J. Rego, H. Chen, S. Li, Z. Hu, K. W. Lau, L.-M. Po, D. Yu, Y. A. U. Rehman, Y. Li, and L. Xing, "Udc 2020 challenge on image restoration of under-display camera: Methods and results," in *Computer Vision – ECCV 2020 Workshops*, A. Bartoli and A. Fusiello, Eds. Cham: Springer International Publishing, 2020, pp. 337–351.

[26] V. Sundar, S. Hegde, D. Kothandaraman, and K. Mitra, "Deep atrous guided filter for image restoration in under display cameras," in *Computer Vision – ECCV 2020 Workshops*, A. Bartoli and A. Fusiello, Eds. Cham: Springer International Publishing, 2020, pp. 379–397.

[27] R. Feng, C. Li, H. Chen, S. Li, C. C. Loy, and J. Gu, "Removing diffraction image artifacts in under-display camera via dynamic skip connection network," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 662–671.

[28] K. Kwon, E. Kang, S. Lee, S.-J. Lee, H.-E. Lee, B. Yoo, and J.-J. Han, "Controllable image restoration for under-display camera in smartphones," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 2073–2082.

[29] K. Wei, Y. Fu, J. Yang, and H. Huang, "A physics-based noise formation model for extreme low-light raw denoising," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 2758–2767.

[30] S. Fuchs, "Multipath interference compensation in time-of-flight camera images," in *2010 20th International Conference on Pattern Recognition*. IEEE, 2010, pp. 3583–3586.

[31] A. Kadambi, R. Whyte, A. Bhandari, L. Streeter, C. Barsi, A. Dorrington, and R. Raskar, "Coded time of flight cameras: sparse deconvolution to address multipath interference and recover time profiles," *ACM Transactions on Graphics (ToG)*, vol. 32, no. 6, pp. 1–10, 2013.

[32] A. Bhandari, M. Feigin, S. Izadi, C. Rhemann, M. Schmidt, and R. Raskar, "Resolving multipath interference in kinect: An inverse problem approach," in *SENSORS, 2014 IEEE*. IEEE, 2014, pp. 614–617.

[33] M. Gupta, S. K. Nayar, M. B. Hullin, and J. Martin, "Phasor imaging: A generalization of correlation-based time-of-flight imaging," *ACM Transactions on Graphics (ToG)*, vol. 34, no. 5, pp. 1–18, 2015.

[34] M. Georgiev, R. Bregović, and A. Gotchev, "Fixed-pattern noise modeling and removal in time-of-flight sensing," *IEEE Transactions on Instrumentation and Measurement*, vol. 65, no. 4, pp. 808–820, 2015.

[35] ——, "Time-of-flight range measurement in low-sensing environment: Noise analysis and complex-domain non-local denoising," *IEEE Transactions on Image Processing*, vol. 27, no. 6, pp. 2911–2926, 2018.

[36] M. Rossi, M. E. Gheche, A. Kuhn, and P. Frossard, "Joint graph-based depth refinement and normal estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 12 154–12 163.

[37] K. Son, M.-Y. Liu, and Y. Taguchi, "Learning to remove multipath distortions in time-of-flight range images for a robotic arm setup," in *2016 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2016, pp. 3390–3397.

[38] G. Dong, Y. Zhang, and Z. Xiong, "Spatial hierarchy aware residual pyramid network for time-of-flight depth denoising," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIV 16*. Springer, 2020, pp. 35–50.

[39] Y. Chen, J. Ren, X. Cheng, K. Qian, L. Wang, and J. Gu, "Very power efficient neural time-of-flight," in *The IEEE Winter Conference on Applications of Computer Vision*, 2020, pp. 2257–2266.

[40] H.-S. Choi, J.-H. Kim, J. Huh, A. Kim, J.-W. Ha, and K. Lee, "Phase-aware speech enhancement with deep complex u-net," in *International Conference on Learning Representations*, 2018.

[41] B. Tolooshams, R. Giri, A. H. Song, U. Isik, and A. Krishnaswamy, "Channel-attention dense u-net for multichannel speech enhancement," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 836–840.

[42] O. Hommos, S. L. Pintea, P. S. Mettes, and J. C. van Gemert, "Using phase instead of optical flow for action recognition," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 0–0.

[43] C.-A. Popa, "Complex-valued convolutional neural networks for real-valued image classification," in *2017 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2017, pp. 816–822.

[44] M. A. Dedmari, S. Conjeti, S. Estrada, P. Ehses, T. Stöcker, and M. Reuter, "Complex fully convolutional neural networks for mr image reconstruction," in *International Workshop on Machine Learning for Medical Image Reconstruction*. Springer, 2018, pp. 30–38.

[45] S. Rawat, K. Rana, and V. Kumar, "A novel complex-valued convolutional neural network for medical image denoising," *Biomedical Signal Processing and Control*, vol. 69, p. 102859, 2021.

[46] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus, "Indoor segmentation and support inference from rgbd images," in *European conference on computer vision*. Springer, 2012, pp. 746–760.

[47] T. Pattinson, "Quantification and description of distance measurement errors of a time-of-flight camera," in *DAGM Young Researchers Forum 2010*, 2010.

[48] A. P. Jongenelen, D. G. Bailey, A. D. Payne, A. A. Dorrington, and D. A. Carnegie, "Analysis of errors in tof range imaging with dual-frequency modulation," *IEEE transactions on instrumentation and measurement*, vol. 60, no. 5, pp. 1861–1868, 2011.

[49] R. Gao, N. Fan, C. Li, W. Liu, and Q. Chen, "Joint depth and normal estimation from real-world time-of-flight raw data," in *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2021, pp. 71–78.

[50] B. Mildenhall, J. T. Barron, J. Chen, D. Sharlet, R. Ng, and R. Carroll, "Burst denoising with kernel prediction networks," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[51] M. Georgiev, R. Bregović, and A. Gotchev, "Time-of-flight range measurement in low-sensing environment: Noise analysis and complex-domain non-local denoising," *IEEE Transactions on Image Processing*, vol. 27, no. 6, pp. 2911–2926, 2018.

[52] T. Brooks, B. Mildenhall, T. Xue, J. Chen, D. Sharlet, and J. T. Barron, "Unprocessing images for learned raw denoising," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[53] N. Guberman, "On complex valued convolutional neural networks," *arXiv preprint arXiv:1602.09046*, 2016.

[54] C. Trabelsi, O. Bilaniuk, Y. Zhang, D. Serdyuk, S. Subramanian, J. F. Santos, S. Mehri, N. Rostamzadeh, Y. Bengio, and C. J. Pal, "Deep complex networks," in *International Conference on Learning Representations*, 2018. [Online]. Available: https://openreview.net/forum?id=H1T2hmZAb

[55] K. Kreutz-Delgado, "The complex gradient operator and the cr-calculus," *arXiv preprint arXiv:0906.4835*, 2009.

[56] P. M. Virtue, *Complex-valued deep learning with applications to magnetic resonance image synthesis*. University of California, Berkeley, 2019.

[57] J. Li, J. Li, F. Fang, F. Li, and G. Zhang, "Luminance-aware pyramid network for low-light image enhancement," *IEEE Transactions on Multimedia*, 2020.

[58] X. Qiao, C. Ge, H. Yao, P. Deng, and Y. Zhou, "Valid depth data extraction and correction for time-of-flight camera," in *Twelfth International Conference on Machine Vision (ICMV 2019)*, vol. 11433. International Society for Optics and Photonics, 2020, p. 114332K.

[59] G. D. Evangelidis and E. Z. Psarakis, "Parametric image alignment using enhanced correlation coefficient maximization," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 10, pp. 1858–1865, 2008.

[60] G. Choe, J. Park, Y.-W. Tai, and I. So Kweon, "Exploiting shading cues in kinect ir images for geometry refinement," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 3922–3929.

[61] A. Mosleh, P. Green, E. Onzon, I. Begin, and J. Pierre Langlois, "Camera intrinsic blur kernel estimation: A reliable framework," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 4961–4968.

[62] B. P. Lathi, *Modern digital and analog communication systems*. Oxford university press, 1998.

[63] B. R. Levin, "Theoretical principles of radioengineering statistics," FOREIGN TECHNOLOGY DIV WRIGHT-PATTERSON AFB OHIO, Tech. Rep., 1968.

[64] K. Dabov, A. Foi, V. Katkovnik, and K. Egiazarian, "Image denoising by sparse 3-d transform-domain collaborative filtering," *IEEE Transactions on image processing*, vol. 16, no. 8, pp. 2080–2095, 2007.

[65] X. Cheng, P. Wang, and R. Yang, "Learning depth with convolutional spatial propagation network," *IEEE transactions on pattern analysis and machine intelligence*, vol. 42, no. 10, pp. 2361–2379, 2019.

[66] S. Chen, Y. Liu, X. Gao, and Z. Han, "Mobilefacenets: Efficient cnns for accurate real-time face verification on mobile devices," in *Chinese Conference on Biometric Recognition*. Springer, 2018, pp. 428–438.

[67] C. Zhong, Z. Sun, and T. Tan, "Robust 3d face recognition using learned visual codebook," in *2007 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2007, pp. 1–6.

[68] M. Kazhdan and H. Hoppe, "Screened poisson surface reconstruction," *ACM Transactions on Graphics (ToG)*, vol. 32, no. 3, pp. 1–13, 2013.

Xin Qiao received the bachelor's degree from the University of Electronic Science and Technology of China and the master's degree from Ordnance Science and Research Academy of China. Currently he is pursing the joint-training Ph.D degree with the Institute of Artificial Intelligence and Robotics, Xi'an Jiaotong university, and Department of Computer Science and Engineering, University of Bologna. His research interest includes depth perception and machine learning.

Chenyang Ge received the B.S., M.S. and Ph.D. degrees from Xi'an Jiaotong University, Xi'an, China, in 1999, 2002 and 2009, respectively. He is currently an associate professor in College of Artificial Intelligence, Xi'an Jiaotong University. His research interests include computational vision, depth perception and SoC design.

Pengchao Deng received the M.S degree in Software Engineering from Xi'an Jiaotong University, Xi'an, China, in 2017, and he is currently a Ph.D. candidate in Institute of Artificial Intelligence and Robotics, Xi'an Jiaotong University, Xi'an, China. His research interests focus on 3D ranging, face anti-spoofing and deepfake detection.

Hao Wei is currently a Ph.D candidate with Institute of Artifical Intelligence and Robotics at Xi'an Jiaotong University. He received his B.Sc. and M.Sc. degrees from Yangzhou University and Nanjing University of Science and Technology in 2018 and 2021, respectively. His research interests include image deblurring, image super-resolution and other low-level vision problems.

Matteo Poggi received received Master degree in Computer Science and PhD degree in Computer Science and Engineering from University of Bologna in 2014 and 2018 respectively. Currently, he is assistant professor at Department of Computer Science and Engineering, University of Bologna. His research interests include deep learning for depth estimation and embedded computer vision.

Stefano Mattoccia received a Master degree in Electronic Engineering and a Ph.D. in Computer Science Engineering from University of Bologna, respectively, in 1997 and 2002. Currently he is associate professor at the Department of Computer Science and Engineering of the University of Bologna. His research interests include computer vision, depth sensing and embedded vision.