Learnable Polarization-multiplexed Modulation Imager for Depth from Defocus

Zhiwei Huang, Mingyou Dai, Tao Yue, and Xuemei Hu

Abstract—Estimating depth from a single snapshot image with defocus information is still a tricky problem for the ill-posedness introduced by the limited depth cues implied in the defocus images. This paper proposes a Polarization-multiplexed Modulation Imager (PoMI) to fully utilize the multiplexed polarization channels for capturing more depth cues with a single snapshot image. The polarization-dependent modulator, i.e., Liquid Crystal Spatial Light Modulator (LC-SLM), is applied to modulate the depth information into polarization channels. A differentiable polarization-dependent modulation camera model is proposed, combined with the Polarization-Driven Attention Network, to enable the joint system optimization by end-to-end training. Extensive tests have been applied to the synthetic datasets to verify the effectiveness of the proposed method. A system prototype is built to conduct real experiments demonstrating the feasibility of the proposed method for natural scenes.

Index Terms—Computational Photography, Polarization-multiplexed, Depth from Defocus

1 Introduction

Acquiring a high-quality, dense depth map of natural scenes from single images with defocus information is a long-standing problem and has shown great potential for automatic driving [1], [2], [3], [4], augmented reality [5], [6], 3D reconstruction [7] and other high-level vision tasks. Although the learning-based methods [8], [9], [10], [11] have been proven to achieve significant improvements in monocular defocus-based depth estimation, the inherent ill-posedness caused by the limited depth cues from defocus blur still impedes the further performance improvement.

To improve the performance of Depth from Defocus (DfD), series of approaches [12], [13], [14], [15] are proposed to encode more depth information into the defocus blur with the coded aperture modulations. One of the recent development is to replace the hand-crafted modulation schemes with the end-to-end training of the optical system and reconstruction network [16], [17], [18], [19]. However, the information deficiency problem still plagues researchers in this field.

Recently, the multi-PSF based techniques [20], [21] are proposed to improve the reconstruction quality in applications of 3D localization in microscopy, demonstrating that the multi-channel information could greatly reduce ambiguities of depth information and improving the quality of reconstruction. However, the complexity of the optical system still impedes the wide application of these techniques on consumer-level applications. Ghanekar *et al*. [22] proposed a polarization-based multiplexing system to achieve single-shot 3D sensing. Nevertheless, the two polarization branches have to be modulated separately at different spatial regions of the aperture, limiting the modulation capabilities of the system and resulting the inconformity of angle of view between two channels.

- *The authors are with the School of Electrical Science and Engineering, Nanjing University, Nanjing 210023, China.*
E-mail: {zhiweihuang, MingyouDai}@smail.nju.edu.cn, {yuetao, xuemeihu}@nju.edu.cn.

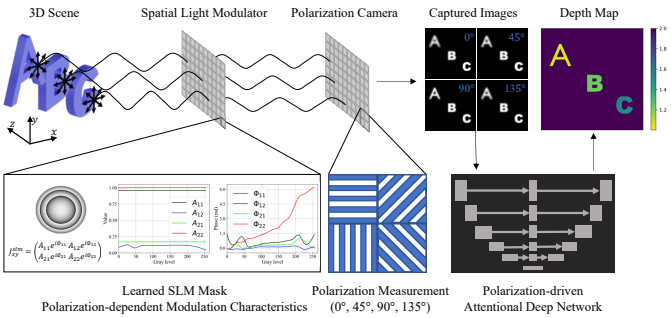

Fig. 1. Schematics for the proposed Polarization-multiplexed Modulation Imager (PoMI) system. The linearly polarized light distributed in the natural scene is modulated by the polarization-dependent modulator, which is modeled by Jones function. The depth information multiplexed into different polarization states is captured by a polarization camera and then reconstructed by PADNet.

In this paper, we propose a Polarization-multiplexed Modulation Imager (PoMI) system for depth imaging of natural scene, which multiplex the depth information into different polarization states with a single polarization-dependent modulator. A differentiable polarization-dependent image formation model is proposed with an end-to-end optimization framework, enabling the joint optimization of the modulation mask and the network for depth estimation.

Specifically, the proposed system comprises a polarization-dependent liquid crystal spatial light modulator (LC-SLM) and a linear polarization camera for the polarization-dependent modulating and sensing respectively. The polarization-dependent LC-SLM is placed in the aperture plane, achieving different modulation effects on different polarization states. The polarization camera equipped with a polarization filter array on the image sensor (Sony Polarsens) is able to capture polarization images with spatially aligned and timing-synchronized channels corresponding to the four linear polarization angles (0°, 45°, 90°, and 135°).

To optimize the modulation effect, we proposed a differentiable Polarization-multiplexed Modulation Camera Model (PoMCaM) to compute the polarization states, modulations, and propagations of the proposed system with a tensor-wise representation. Specifically, the polarization response property of each pixel of the LC-SLM is modeled by a Jones matrix function. The elements of the Jones matrix varies with the input gray value of the pixel and could be described by complex response curves, as shown in Fig. 1. Combined with the phase modulation effect of the objective lens, the polarization-dependent system modulation function can be represented by a Jones tensor. The polarized optical transfer function of the system can be derived by propagating the Jones tensor from the aperture to the sensor by the Fouier/Hankel transform. For handling the natural light, the Jones tensors are transformed to the Muller-based representation, so that the full Stokes response of the system can be computed. A simple transform is required to computed the captured linearly polarized PSFs from the full Stokes response of the system. By delicately designing the input of the polarization-dependent LC-SLM modulator, the polarization channels could efficiently capture complementary depth cues within four polarization channels by a single snapshot. For reconstruction, we propose a Polarization-driven Attentional Deep Network (PADNet) to recover the depth from the captured multi-channel polarization images.

With the proposed differential forward model, i.e. PoMCaM, and depth reconstruction network, i.e. PADNet, we propose to learn the polarization-multiplexed modulation and reconstruction network in an end-to-end manner, achieving the joint optimization of the modulation mask and the reconstruction network.

The main contributions of this paper are as follows.

- We introduce a novel imaging system for depth estimation called polarization-multiplexed modulation imager, based on the polarization-dependent modulation characteristic of LC-SLM.
- We develop a differentiable PoMCaM model to compute the polarization states, modulation, and propagation of the proposed system with a tensor-wise representation, enabling the joint optimization of the system in an end-to-end manner.
- To extract informative polarization features for depth estimation, we design the Polarization-driven Attentional Deep Network (PADNet) to process the coded polarization images. The analysis and evaluation on two datasets (FlyingThings3D, NYU Depth v2) shows that our method could outperform the state-of-the-art DfD approaches.
- We build a prototype imaging system and load the LC-SLM with the optimized modulation mask to conduct real experiments, demonstrating the effectiveness of the proposed method for natural scenes.

2 Related Work

2.1 Computational Imaging for Depth Estimation

To alleviate the ambiguity of estimating depth from a single 2D image, methods in computational imaging usually rely on designed optical systems to make an indirect measurement of depth information, which is then recovered by subsequent image processing. Many variants of DfD approaches based on the conventional imaging system are proposed [23], [24], [25], which determine the clear relationship between depth and defocus by capturing two or more images. While the other category of DfD approaches engineer the PSF to be varying more distinctively with the scene depth by inserting amplitude-coded [12], [14] or phase-coded apertures [26], [27]. Recently a major manifestation of the combination of deep learning and optics is deep optics, which introduce joint optimization of optics and depth estimation networks. These type of works [17], [18], [19], [28] replace the hand-crafted aperture designs with end-to-end optimized phase design and achieve considerable performance improvements. Recently, multi-PSF-based depth information encoding is proposed, and demonstrated efficient in improving depth accuracy [20], [21]. In this paper, we propose an end-to-end multi-PSF-based depth imaging method, with a polarization-multiplexing imaging scheme. Through utilizing the polarization-dependent response property of SLM and a polarization sensing sensor, the defocus blur in different polarization channels can be optimized and high quality depth information are demonstrated to be achieved.

2.2 Aperture Modulations

Aperture modulation-based systems place spatial light modulators with designed modulation masks in the aperture plane to modulate the optical properties of incident light, obtaining image data containing informative features in visual tasks including depth estimation [12], compressive sensing [29], [30], lensless imaging [31], [32], etc. However, each point in the aperture plane of most existing aperture modulation systems are only formulated as tunable variables of intensity or phase, in a scalar form, without taking the polarization dimension into considering. In this paper, we introduce a polarization-multiplexed modulation system, which utilizes the polarization-dependent SLM as the aperture modulator. Each point in the aperture plane is formulated in a complex tensor function, with respect to the gray level. Based upon the polarization dependent aperture modulator, the forward imaging model could be formulated in a tensor-wise form, enabling higher flexibility and efficiency for depth information encoding in different polarization channels.

2.3 Polarization Imaging

Extensive research has been conducted in polarization imaging since the 1970s, resulting in significant advancements. The implementation of polarization imaging can be categorized into rotary polarization imaging with rotating polarizers [33], electronically controlled polarization imaging with polarization-dependent devices [34], [35], fractional focal plane polarization imaging with micro-polarized optical elements [36]. In recent years, several consumer-level polarization camera become off-the-shelf, facilitating the application of polarization imaging in other vision tasks. In this paper, to capture the polarization dependent signal, we adopt the consumer-level polarization camera, i.e., the Triton Polarization Camera (with Sony IMX250MZR sensor), as the polarization image sensor in our PoMI system. With the polarization camera, the complexity of the proposed

polarization-multiplexing imaging system could be largely reduced, and spatially aligned and time-synchronized images of different polarization channels could be captured in a snapshot way.

2.4 Application of Polarization Imaging in Vision Tasks

As for the application, polarization imaging can provide more information and feasibility for many vision tasks. The polarization properties of scenes are proved to be useful in 3D sensing [37], [38], [39], object segmentation [40], [41], [42], dehazing [43], [44], etc. To take better advantage of the polarization information, physical models considering polarimetric characteristic are proposed to solve recognition in complex scenes originally [45], [46], [47]. With the development of deep learning, numerous multi-branch CNN architectures are introduced to effectively leverage the extra information of polarization images [40], [41].

It is worth noting that Ghanekar *et al*.propose a Polarized Spiral Point Spread Function system [22] to achieve accurate 3D sensing with polarization information. The proposed method is distinct from their method on the following three main points,

- The proposed method use a polarization-dependent modulator to achieve the polarization-multiplexed modulation without spatial division of the aperture.
- A differentiable tensor-wise imaging model is proposed based on the Fourier optics taking the polarization states into consideration.
- An end-to-end training framework is applied to jointly optimize the optical modulation mask and the following reconstruction network.

In all, we propose a PoMI system to achieve the polarization-multiplex modulation, and build the corresponding differentiable forward imaging model to enable the end-to-end learning of the entire system. Based on the proposed polarization multiplexing and end-to-end optimization framework, more accurate depth estimation performance from single snapshot can be achieved.

3 Polarization-multiplexed Modulation Imaging System for Snapshot Depth Sensing

In this section, we present the proposed Polarization-multiplexed Modulation Imager (PoMI) system to estimate the depth map from a single snapshot polarized image, as shown in Fig. 2. The PoMI system comprises a polarization-dependent LC-SLM for polarization-multiplexed light modulation and a polarization camera for capturing the four-channel polarization images. We build a differentiable Polarization-multiplexed Modulation Camera Model (PoMCaM) based on the Jones matrix tensors to model the PoMI system with polarization states. A PADNet is presented to estimate the depth map from the polarization images captured by the PoMI system. Combined with the differentiable PoMCaM model, the entire system could be optimized in an end-to-end manner, as shown in Fig. 2.

3.1 Polarization-multiplexed Modulation Camera Model

In the following, we model the polarization-multiplexed modulation imaging system based on Fourier optics [48], taking polarization modulation effects into consideration.

3.1.1 Jones Function for Polarization-dependent Modulator

Traditionally, the fully polarized light could be modeled by the 2D complex Jones vector, $\boldsymbol{e} = [E_x, E_y]^T$, where the complex elements $E_x = E_{0x}e^{i\Phi_x}$ and $E_y = E_{0y}e^{i\Phi_y}$. E_{0x}, E_{0y}, Φ_x and Φ_y represent the amplitudes and phases of the electronic field in x and y directions, respectively.

When the light incident on the surface of polarization-dependent devices such as the LC-SLM, its polarization states change. The Jones matrix is used to model the effect of optical elements, e.g., lenses, beam splitters, mirrors, on the polarization states as an operator that act on the Jones vectors of the incident light.

In our paper, we describe the Jones matrix function of the polarization-dependent modulator, i.e. the LC-SLM, based on the complexed tensor form,

$$J_{xy}^{\mathrm{SLM}} = \begin{pmatrix} A_{11}e^{i\Phi_{11}} & A_{12}e^{i\Phi_{12}} \\ A_{21}e^{i\Phi_{21}} & A_{22}e^{i\Phi_{22}} \end{pmatrix}, \tag{1}$$

where A and Φ are the amplitude and phase transfer function between Jones vectors of the incident light and those of the emergent light.

To mitigate the impact of the dispersion of liquid crystal, we assume that the incident light is monochromatic of wavelength λ. At the fixed operating wavelength λ, LC-SLM has a variable electro-optic response to the input gray levels g ranging from 0 to 255. Thus each pixel located at the spatial coordinate (m, n) of LC-SLM has a varying Jones matrix function $J_{mnxy\lambda}^{\mathrm{SLM}}(g)$.

3.1.2 Polarized Tensor Model for PoMI System

To handle the entire PoMI system, the Jones tensor of the system, including the modulation function of both the objective lens and the LC-SLM, can be derived by

$$J_{mnxy\lambda}^{\mathrm{sys}}(g) = \tau_{mn\lambda}^{\mathrm{lens}} * J_{mnxy\lambda}^{\mathrm{SLM}}(g), \tag{2}$$

where $*$ denotes the Hadamard products between two tensors, m-n denotes the aperture plane, $\tau_{mn\lambda}^{\mathrm{lens}}$ is the transmission tensor of objective lens. In our system, the objective lens with focal length f is polarization-independent, so that the lens can be presented by $\tau_{mn\lambda}^{\mathrm{lens}} = A^{lens}e^{i\phi_{mn\lambda}}$. A^{lens} is the amplitude transmission rate that usually can be regarded as constant 1 within the aperture. $\phi_{mn\lambda} = -\frac{k}{2f}(m^2+n^2)$ is the phase modulation term, where $k = \frac{2\pi}{\lambda}$ is the wavenumber.

Given the Jones tensor of the PoMI system, the response in the sensor plane could be approximately computed by the Fresnel diffraction integral with 2D Fourier transform in m-n plane:

$$\begin{aligned} R_{m'n'xyz\lambda}(g) =& \frac{e^{iks}}{\lambda s} e^{\frac{ik}{2s}(m'^2+n'^2)} \\ & \mathcal{F}_{m\to\omega_m, n\to\omega_m}(T_{mnz\lambda} J_{mnxy\lambda}^{\mathrm{sys}}(g) e^{\frac{ik}{2s}(m^2+n^2)}), \end{aligned} \tag{3}$$

where m'-n' denotes the sensor plane, $\omega_m = \frac{m'}{\lambda s}$ and $\omega_n = \frac{n'}{\lambda s}$ are spatial frequencies, s is the distance between the aperture plane and the sensor plane, and $T_{mnz\lambda} = \frac{e^{ikz}}{\lambda z} e^{\frac{ik}{2z}(m^2+n^2)}$ is the corresponding transport term for an incident point source located a distance z in front of the objective lens.

To satisfy the computational memory requirement and enable rapid system prototyping, instead of optimizing the

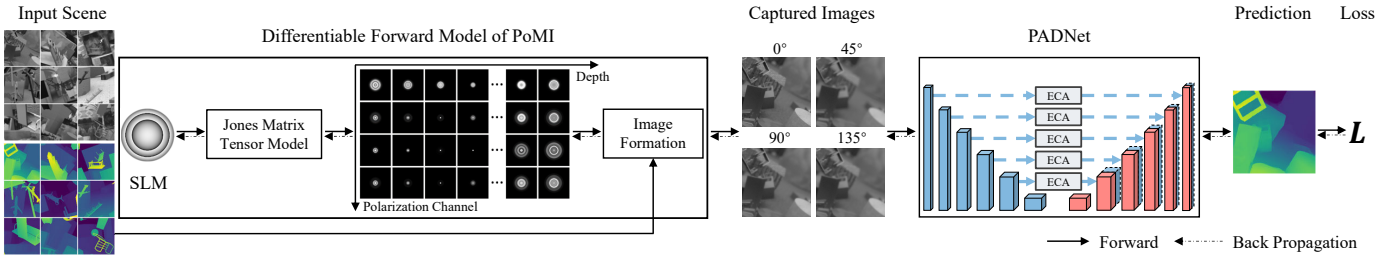

Fig. 2. Overview of the proposed end-to-end learnable PoMI system. As for training, the proposed differentiable PoMCaM model is utilized to simulate the physics of the acquisition process. The resulting polarization modulated images with four linear polarization channels are fed into PADNet to reconstruct the depth maps. The calculated loss between the estimated depth and ground truth depth is back propagated to update the modulation gray level of SLM and the parameters of PADNet jointly.

two-dimension modulator in an unconstrained way, the widely used radial symmetry constraint [49] is used to reduce the modulation degree of freedom to one dimension by replacing (m, n) with the radius $r = \sqrt{m^2 + n^2}$. By leveraging the radial symmetry, the two-dimensional Fourier transform for propagation in Eq. 3 can be simplified as a one-dimensional manipulation called Hankel transform of order zero [48]:

$$\mathcal{H}_{r\to r'}(T_{rz\lambda}J^{\text{sys}}_{rxy\lambda}(g)) = 2\pi\int_0^\infty rT_{rz\lambda}J^{\text{sys}}_{rxy\lambda}(g)B_0(2\pi r'r)dr, \tag{4}$$

where $B_0(\cdot)$ is the zero-th order Bessel function of the first kind, and $r' = \sqrt{m'^2 + n'^2}$ is the radial coordinate in the sensor plane. Thus, the response tensor in the sensor plane of our PoMI system becomes

$$R_{r'xyz\lambda}(g) = \frac{e^{iks}e^{\frac{ikr'^2}{2s}}}{\lambda s}\mathcal{H}_{r\to r'}(T_{rz\lambda}J^{\text{sys}}_{rxy\lambda}(g)). \tag{5}$$

It is worth noting that the above model is based on the Jones tensors, which can be only applied for cases with fully polarized incident light. Therefore, we boot the response tensor to the Mueller based tensor to handle the natural or partially polarized light in our scenario, i.e., depth estimation of natural scene without active illumination. Specifically, the Muller-based response tensor can be calculated by

$$P^{\text{Muller}}_{r'xyz\lambda}(g) = \Gamma \times_{xy} (R_{r'xyz\lambda}(g) \otimes_{xy} R_{r'xyz\lambda}(g)^*) \times_{xy} \Gamma^{-1}, \tag{6}$$

where $(\cdot)^*$ indicates the complex conjugate, $\otimes_{xy}$ represents the 2D Kronecker product in x-y plane. The size of the results of the Kronecker product in x-y plane becomes 4×4. $\times_{xy}$ denotes the 2D matrix product in x-y dimension as well. Γ is a 4×4 transform matrix providing the change of basis from the Pauli basis to the standard Cartesian basis (See [50] for details) in the form of

$$\Gamma = \begin{pmatrix} 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & -1 \\ 0 & 1 & 1 & 0 \\ 0 & i & -i & 0 \end{pmatrix}. \tag{7}$$

Given the Muller-based response function tensor $P^{\text{Muller}}_{r'xyz\lambda}(g)$, the Stokes vectorized Point Spread Function (SPSF) can be computed by

$$S^{\text{PSF}}_{r'xz\lambda}(g) = P^{\text{Muller}}_{r'xyz\lambda}(g) \times_y S^{\text{in}}, \tag{8}$$

where $S^{\text{PSF}}_{r'xz\lambda}$ means the Stokes tensor PSF with four Stokes channels indexed by x, r' is the one dimensional coordinate of the PSF in radial dimension, λ is the wavelength of the light, $\times_y$ is the matrix product on y dimension. S^{in} is the Stokes vector of the incident light, for nature light $S^{\text{in}} = [1, 0, 0, 0]^T$.

3.1.3 Sensing by Polarization Camera

Currently, since the consumer-level single-snapshot full-Stokes polarization camera is still unavailable in practice, we use the linear polarization camera with four linearly polarized channels, i.e., $(0°, 45°, 90°, 135°)$. According to the relationships between the Stokes parameters and linear polarization channels,

$$P^{\text{PSF}}_{r'\theta z\lambda}(g) = T^{\text{Stokes}\to\text{linear}}_{\theta x} \times_x S^{\text{PSF}}_{r'xz\lambda}(g), \tag{9}$$

where $P^{\text{PSF}}_{r'\theta z\lambda}$ is the captured PSF tensor in four linearly polarized channels $\theta \in \{0°, 45°, 90°, 135°\}$, $T^{\text{Stokes}\to\text{linear}}_{\theta x}$ is the transfer matrix projecting the full Stokes vector to the linearly polarized subspaces, which can be express by

$$T^{\text{Stokes}\to\text{linear}}_{\theta x} = \begin{pmatrix} 1/2 & 1/2 & 0 & 0 \\ 1/2 & -1/2 & 0 & 0 \\ 1/2 & 0 & 1/2 & 0 \\ 1/2 & 0 & -1/2 & 0 \end{pmatrix}. \tag{10}$$

3.1.4 Image Formation

Given the differentiable model of the proposed PoMI, we could simulate the transfer function of the system, and calculate the PSFs at different depths. To further derive the polarization images with defocus effects, we simply extend the differentiable image formation method in [19] to multi-channel case for handling the defocus blur at depth discontinuities. For the input with linear polarization channel θ, the depth map is quantized into K depth layers to compose alpha masks $\alpha_\theta(k)$ [51]. Then the monochromatic image is quantized into K depth layers $l_\theta(k)$. Then, the defocused images can be calculated as

$$I_\theta = \sum_{k=0}^{K-1} \widetilde{l}_\theta(k) \prod_{k'=k+1}^{K-1} (1 - \widetilde{\alpha}_\theta(k')) + \eta, \tag{11}$$

where $\widetilde{l}_\theta(k) = (\text{PSF}_\theta(k) * l_\theta(k))/E_\theta(k)$, $\widetilde{\alpha}_\theta(k') = (\text{PSF}_\theta(k) * \alpha_\theta(k))/E_\theta(k)$ and η is the additive noise. To keep the continuity of brightness at the transition of depth layers, a normalization with a factor is applied as $E_\theta(k) = \text{PSF}_\theta(k) * \sum_{k'=0}^{k} \alpha_\theta(k')$.

3.2 Polarization-driven Attentional Deep Network

After the PoMI system, the captured/simulated multi-channel images are sent to a Polarization-driven Attentional Deep Network (PADNet) for reconstructing the depth maps, as shown in Fig. 2. The details of the PADNet and the corresponding loss function are introduced in the following.

3.2.1 Architecture

The PADNet consists of two main blocks: 1) an encoder-decoder depth regression network built on a pretrained EfficientNet B3 [52] encoder and a standard decoder with five consecutive upsampling operations; 2) Efficient Channel Attention (ECA) module used to extract informative polarization features. The ECA modules placed before the skip connections extract different levels of rich contextual features, further promoting the reconstruction performance. The details of the ECA module are given in the following description.

3.2.2 ECA Module

The input of PADNet is the modulated polarization-multiplexing images with more depth cues implied in the channel-wise features. In order to gather informative features selectively from the channel-wise information, the ECA modules based on channel attention [53] are applied before skip connections, as shown in Fig. 3. The module produces more informative features by exploring the inter-channel relationship of features extracted by the encoder.

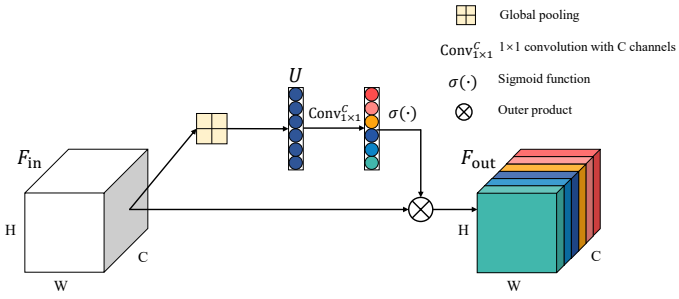

Fig. 3. Efficient Channel Attention (ECA) module. We apply ECA module before skip connection to extract informative polarization features.

Assuming the input feature map is $F_{\text{in}} \in \mathbb{R}^{C\times H\times W}$, we first apply global average pooling, to have the output $U \in \mathbb{R}^{C\times 1\times 1}$, where C denotes the number of feature channels, H, W denote the height and width of feature maps respectively. The c-th ($c \in [1, C]$) element of U can be expressed as

$$U^c = \frac{1}{H\times W}\sum_{i=1}^{H}\sum_{j=1}^{W} {F_{\text{in}}}^c_{i,j}. \tag{12}$$

Then the weighted feature map after the ECA module can be expressed as

$$F_{\text{out}} = F_{\text{in}} \otimes \sigma[\varphi(U)], \tag{13}$$

where $\otimes$ denotes the outer product, σ denotes the sigmoid function and $\varphi(\cdot)$ denotes a 1×1 convolution layer with the same number of channels as U.

3.2.3 Loss Function

For training our end-to-end PoMI framework, we define the total loss as the weighted sum of three loss functions:

$$\mathcal{L} = \mathcal{L}_{\text{depth}} + \beta_{\text{grad}}\mathcal{L}_{\text{grad}} + \mathcal{L}_R. \tag{14}$$

The first loss term $\mathcal{L}_{\text{depth}}$ is the L1 loss defined on the ground truth depth image d and the reconstructed depth image $\hat{d}$:

$$\mathcal{L}_{\text{depth}} = \frac{1}{N}\sum_{i=0}^{N}|d_i - \hat{d}_i| \tag{15}$$

The second loss term $\mathcal{L}_{\text{grad}}$ is the L1 loss defined over the image gradient of the ground truth depth image d and the reconstructed depth image $\hat{d}$:

$$\mathcal{L}_{\text{grad}} = \frac{1}{N}\sum_{i=0}^{N}|g_x(d_i) - g_x(\hat{d}_i)| + |g_y(d_i) - g_y(\hat{d}_i)|, \tag{16}$$

where $g_x(\cdot)$ and $g_y(\cdot)$ denote the x and y components for the gradients of depth image. We set $\beta_{\text{grad}} = 10$ empirically.

The last term $\mathcal{L}_R$ is the loss for preventing the energy leakage during propagation in Eq. 5. We also introduce the transfer PSF regularization [19] in our system to constrain the energy of PSFs focus on a limited area by

$$\mathcal{L}_R = \sum_{k=0}^{K-1}\sum_{r'>r_{\text{target}}}\sum_{i=0}^{1}\sum_{j=0}^{1}\left|R_{r'ijk\lambda}(g)\right|^2, \tag{17}$$

where the target radius r_{target} is 32 pixels.

4 Experiments

In this section, we present the details of the PoMI system, e.g., the calibration method, initialization and training details, and the implementation details of the prototype system. Qualitative and quantitative results are demonstrated on the synthetic datasets. We conduct extensive comparisons with state-of-the-art methods to verify the effectiveness and superiorities of the proposed method. Finally, the experiment results on the real captured images are present to demonstrate the feasibility of the proposed system.

4.1 System Implementation

We implement the end-to-end optimization framework on the PyTorch [54] platform and train the system on the synthetic datasets. The prototype system is implemented with a reflective LC-SLM, which is calibrated beforehand. To reduce the chromatic aberration, the proposed PoMI system captures monochromatic images corresponding to four polarization channels. Considering the operating wavelength of our LC-SLM, a narrow bandpass filter whose center wavelength is 532±2 nm is placed in front of the LC-SLM. The implementation details are discussed in the following.

4.1.1 Calibrating the Jones Matrix of Modulators

To achieve the desired modulation effect, it is necessary to accurately calculate all four complex elements of the Jones matrix in Eq. 1. Many calibration methods have been proposed to determine the Jones matrix function of LC-SLM. We adopt the methods [55], [56], [57] that do not require any prior knowledge about the SLM fabrication materials and

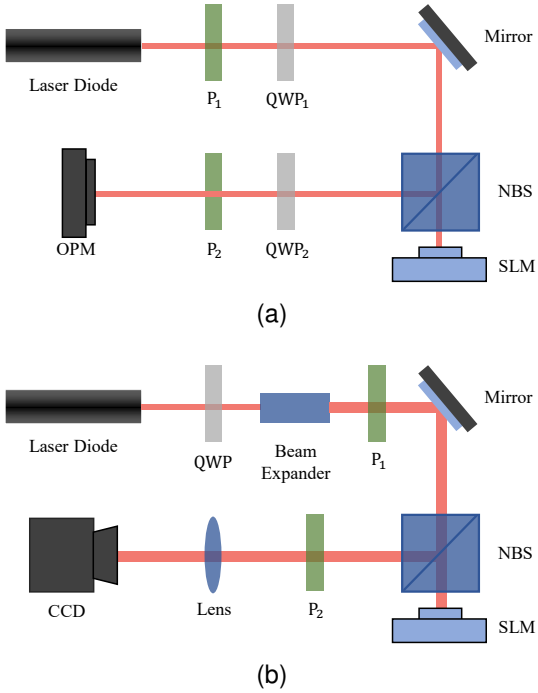

Fig. 4. The experimental setups for calibrating the Jones matrix of LC-SL: irradiance (a) and phase (b) measurements. P: linear polarizers; QWP: quarter-wave plates; NBS: Non-polarizing beam splitter; OPM: optical power meter.

techniques. It is required to measure the optical properties of the transmitted light for different polarization configurations, which can be adjusted by changing the orientation of polarizing optical elements in the calibration setups. The elements of the Jones matrix can be computed from the measurements of different configurations.

Fig. 4 shows two schemes of the experimental setups. A laser diode with a 532 nm center wavelength is employed to illuminate the LC-SLM in our setups. Through the setup in Fig. 4a, we derive the amplitude transfer function A by measuring the transmitted irradiance for ten polarization configurations. The setup in Fig. 4b is a modification of a Twyman-Green interferometer [58]. We set the left half of the LC-SLM as a reference with a constant gray level value equal to 0. The right half of the LC-SLM addresses different gray levels for measuring phase transfer function Φ. Since the phase modulation provided by each half is different, a phase shift Φ' is generated and then translated into a fringe displacement on the interference plane, which is captured by a CCD camera. The value of phase shift Φ' determined by Φ can be calculated by,

$$\Phi = \mathcal{Y}(\Phi') = \mathcal{Y}(2\pi\frac{\Delta s}{L}), \tag{18}$$

where $\mathcal{Y}$ represents the mapping relations from the phase shift Φ' to parameters Φ (see [55], [56] for details), Δs represents the left-right fringe displacement and L is the period of the interference fringes. Fig. 5 shows a typical pattern of interference fringes captured and its corresponding vertical intensity distribution. We employ the fast Fourier transform (FFT) filter to smooth the noisy fringes and then precisely obtain Δs and L. Therefore, the phase transfer function Φ can be derived from the above phase measurements.

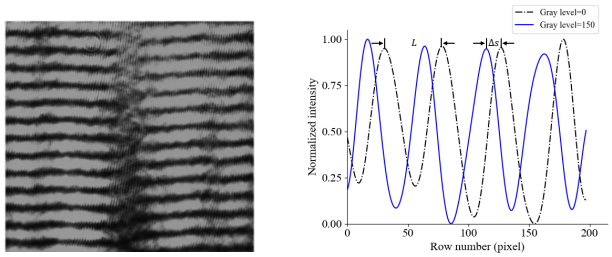

Fig. 5. Interference fringes (a) and the corresponding intensity distribution (b) for gray levels 0 and 150.

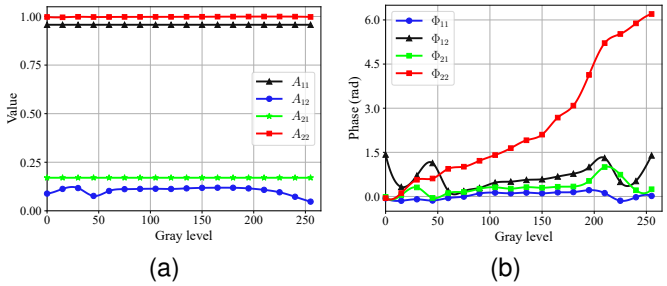

Fig. 6. Calibrated parameters of Jones matrix function of LC-SLM with respect to the gray level. (a) Amplitude elements; (b) Phase elements.

Calibration Results. We use the optical system setups shown in Fig. 4 to calibrate the Jones matrix function of the LC-SLM (Meadowlark 1920 x 1152 XY Phase Series SLM) in 532 nm wavelength. First, we calculate Jones matrix elements for 18 sets of gray levels sampled uniformly from 0 to 255. The result is shown in Fig. 6. According to the discrete measurements, we employ the polynomial fitting method to fit the nonlinear relationship between the Jones matrix function elements and the corresponding gray level. Then the fitted continuous functions are applied to the end-to-end optimization framework, which enables gradients to flow backward to update the gray level of the modulation mask projected on the LC-SLM.

4.1.2 *Initialization of PoMI System*

The calibration results shown in Fig. 6b indicate that the phase modulation effect of our LC-SLM is quite different for $0°$ and $90°$ polarization channels. This characteristic enables us to enrich the system response diversity among linearly polarized channels. Since focus variations are proven to be useful in DfD approaches [59], [60], [61], an SLM mask that shifts the focus of the PoMI system in $90°$ polarization channel is designed as the initial mask. As shown in Fig. 2, the objective lens is focused at 1.7 m with an f-number of 6.3 while PoMI system in $90°$ polarization channel is set to be focused at 1.27m empirically.

4.1.3 *Datasets*

Two datasets with 3D/depth information are used for generating the training and testing datasets.

FlyingThings3D. This synthetic dataset has been widely used for depth estimation in recent years because it contains high-quality texture and depth. We use the cleanpass

version of the FlyingThings3D [62], [63] for training, which contains 22K training samples and 8K testing samples. The training dataset is divided into two sets with 18K and 4K samples for training and validation, respectively. We train our model on a random crop of 384×384 with random horizontal/vertical flipping.

NYU Depth v2. To verify the generalization ability of our system on natural scenes, we train and test the system on the real captured dataset, i.e., NYU Depth v2 [64] as well. The NYU Depth v2 dataset consists of 464 indoor scenes of size 640×480, which are split into 249 scenes for training and 215 scenes for testing. We adopt the processed version [65] that contains 50K training samples and 654 samples. The training samples are divided into 42K and 8K samples for training and validation. The processing of data during training is the same as above.

It is assumed that the system is applied to the unpolarized scene, so the monochromatic images with four polarization angles look the same before being modulated. To simulate the input, we select the green channel of the RGB images and generate four copies in the channel dimension as the original scene image. This four-channel image is quantized into 12 depth layers according to its corresponding depth map. Following the camera configuration in [19], the target depth range in both datasets is set to 1.0 m to 5.0 m. Afterward, the depth map is resampled with the inverse perspective sampling scheme [66] during training.

4.1.4 Training Details

The model is trained with the Adam optimizer ($\beta_1 = 0.9$, $\beta_2 = 0.999$, $max_lr = 0.01$). For FlyingThings3D, the total number of epochs is set to 100 with a batch size of 12. For NYU Depth v2, 50 epochs are used for training with a batch size of 12. To avoid overfitting, we use the exponential learning rate decay with $\gamma = 0.98$. The model achieving the lowest validation loss after training is used for evaluating the test samples. It is noted that the evaluation is performed in the absolute depth space, which is transformed from the inverse perspective sampling domain.

4.1.5 System Prototype

We built a system prototype shown in Fig. 7. The system consists of a primary lens (50 mm f/1.4 HS5018J), a 532 nm band-pass filter, two relay lenses of 50 mm focal length, one non-polarizing beam splitter, an LC-SLM (Meadowlark 1920 x 1152 XY Phase Series) and a polarization camera (TRI050S-PC, resolution 2448 x 2048, 9.2 µm square pixels). A learned mask with 860 × 860 pixels is projected on the central region of the LC-SLM to encode the incoming lights from a natural scene. Finally, the polarization camera captures modulated monochromatic images corresponding to four linearly polarized channels.

4.2 Experiments on Synthetic Data

4.2.1 Comparison to the State-of-the-art

We consider prior works that try to improve the depth estimation effect through the encoding of the optical system to be our primary competitors. Image formation models in these works were reimplemented with a U-Net style network [67] by Ikoma *et al*. [19].

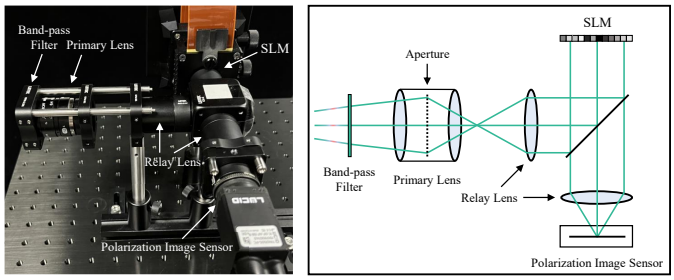

Fig. 7. The prototype and corresponding optical diagram of the proposed PoMI system.

To make the comparison more convincing, we additionally train a model consisting of the proposed polarization-multiplexed modulation and the same U-Net style network in [19] as a baseline for demonstrating the superiority of the proposed PoMI system.

FlyingThings3D. We present the quantitative and qualitative comparisons on the FlyingThings3D testset in Tab. 1 and Fig. 8. Obviously, the proposed method, i.e., including the polarization-multiplexed modulation and PDANet, outperforms all the state-of-the-art methods in all metrics. Compared with Ikoma *et al*. [19] w/o pinv, our baseline method, which exactly uses the same reconstruction network in [19] and the proposed polarization-multiplexed modulation, achieves significant improvement both quantitatively and qualitatively, demonstrating that our system has a higher degree of modulation freedom and information encoding capacity.

NYU Depth v2. Since most of the prior works did not conduct experiments on NYU Depth v2, we reimplemented several image formation models, e.g., all in focus (AiF), non-learned defocus (DFD), and learned defocus model proposed by Ikoma *et al*. [19], for comparison. Tab. 2 and Fig. 9 show the comparison results on the official NYU Depth v2 test set. We can see that our method outperforms the state-of-the-art as well, demonstrating the generalization ability of the proposed system for natural scenes.

4.2.2 Robustness Test at Different Noise Levels

Since the four polarization channels are not completely independent, the corresponding PSFs are correlated. To verify the effectiveness of the other two channels (45° and 135°), we conduct robustness test on models utilizing four channels and two orthogonal channels (0° and 90°) on NYU Depth V2 dataset. It is noted that the noise level (standard deviation) for training ranges from 0 to 0.01, while the one for testing ranges from 0 to 0.05. Results in Fig. 10 indicate that the superiorities of the proposed four-channel model are gradually expanding with the noise level increases, demonstrating that the other two channels could help improve the robustness to different noise levels.

4.2.3 Ablation Study on Modulation Schemes

For our ablation study, we evaluate the influence of different modulation schemes on reconstruction results. Based on the fixed imaging system settings, we train the PADNet by simulating the captured images modulated by several

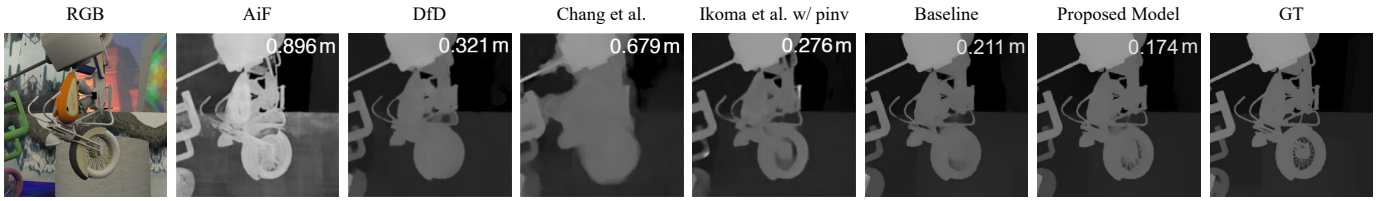

Fig. 8. Qualitative comparison of our proposed method on the FlyingThings3D dataset against prior works. RMSE of the depth map are shown on the top right.

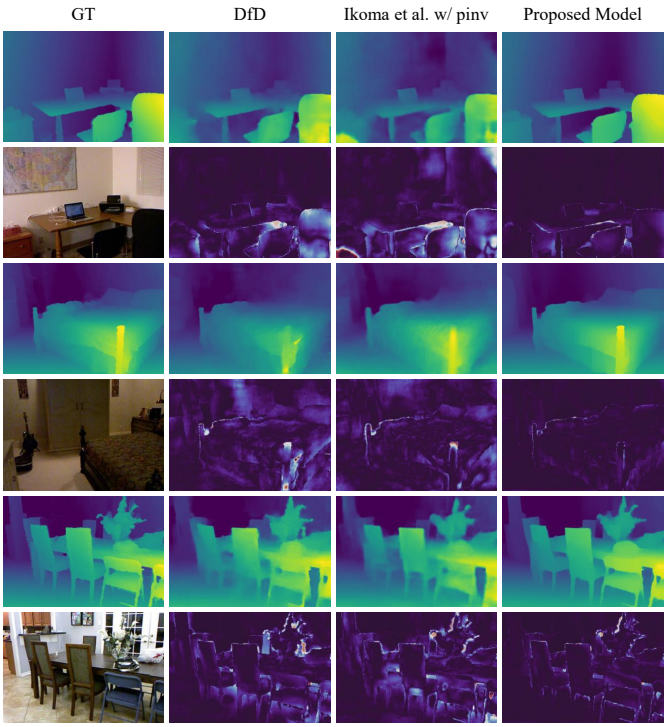

Fig. 9. Qualitative comparison of our proposed method on the NYU Depth V2 dataset against prior works. For each sample, the RGB image in the lower-left corner represents the captured scene. The first rows are the estimated depth maps and the second rows show the corresponding absolute error maps (bright color denotes large errors).

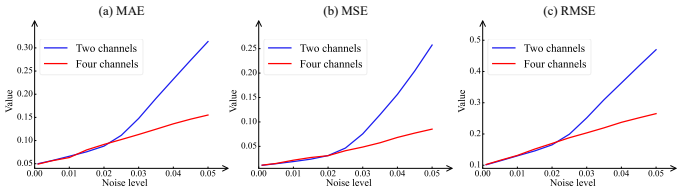

Fig. 10. Quantitative comparison of two models at different noise levels.

alternative modulation schemes. The quantitative and qualitative results shown in Tab. 3 and Fig. 11 indicate that the proposed modulation is superior to other modulation schemes in our scenario.

4.3 Experiments with Prototype

We also perform physical experiments using a system prototype to verify the simulation results. The details about the system prototype and experimental results are described in the section.

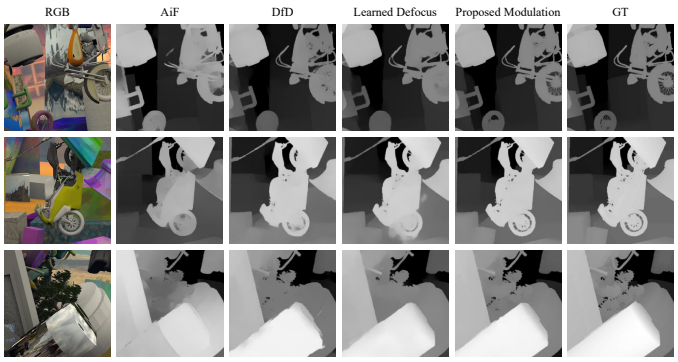

Fig. 11. Qualitative comparison of our proposed modulation against other modulation schemes.

4.3.1 *Model Refinement with PSF Calibration*

After projecting the optimal modulation mask on the LC-SLM, we captured the depth-dependent PSFs of our proposed system using a 532 nm LED Emitter with a 50 µm pinhole. As shown in Fig. 12, the difference of PSFs among the four linearly polarized channels is consistent in the simulation and practical experiment. However, the energy distribution inside the captured PSFs is slightly different from the optimized ones due to misalignment of the light path, deviations in parameters of the optical instrument, and other optical aberrations.

In order to compensate for this difference and achieve better experimental results, we retrain PADNet with the captured PSFs for the depth estimation in natural scenes. The NYU Depth v2 dataset is used for the model refinement.

4.3.2 *Experimental Results on Real Capture Data*

After the PSF calibration and model refinement, we capture the modulated polarization images of natural scenes using the system prototype. Due to the monochromatic design, we utilize a green LED as the fill light for indoor scenes, and natural light for outdoor scenes. For the reconstruction, we first apply the sRGB conversion to produce the sensor images in linear color space and then feed them into retrained PADNet. For comparison, we additionally train a modified PADNet with captured images in 0° polarization channel. As shown in Fig. 13, our system has demonstrated a strong proficiency in estimating the depth of textured areas and edges. The discrepancy in depth estimation between polarization-multiplexed and single channel modulation indicates that the learned polarimetric PSFs modulate more informative depth cues than single learned PSF, which facilitates the accuracy of depth estimation. The main limitation of our system is that it performs poorly on

TABLE 1
Comparison of performances on the FlyingThings3D dataset. Best results are in bold, second best are underlined.

Model	Depth					
	MAE ↓	RMSE ↓	$\log_{10}$ ↓	$\delta < 1.25$ ↑	$\delta < 1.25^2$ ↑	$\delta < 1.25^3$ ↑
All in focus (AiF)	0.357	0.500	0.099	0.658	0.807	0.874
Non-learned defocus (DfD)	0.097	0.228	0.039	0.929	0.965	0.979
Hain *et al*. [16]	0.297	0.635	0.109	0.803	0.879	0.923
Wu *et al*. [17]	0.207	0.521	0.090	0.865	0.918	0.945
Chang *et al*. [18]	0.205	0.490	0.077	0.888	0.945	0.968
Ikoma *et al*. [19] w/o pinv	0.104	0.237	0.041	0.925	0.963	0.977
Ikoma *et al*. [19] w/ pinv	0.089	0.191	0.034	0.941	0.970	0.981
Baseline (Ours)	<u>0.049</u>	<u>0.109</u>	<u>0.009</u>	<u>0.946</u>	<u>0.971</u>	<u>0.983</u>
Proposed model (Ours)	**0.038**	**0.084**	**0.007**	**0.956**	**0.978**	**0.986**

TABLE 2
Comparison of performances on the NYU Depth v2 dataset. Best results are in bold, second best are underlined.

Model	Depth					
	MAE ↓	RMSE ↓	$\log_{10}$ ↓	$\delta < 1.25$ ↑	$\delta < 1.25^2$ ↑	$\delta < 1.25^3$ ↑
All in focus (AiF)	0.655	0.830	0.106	0.725	0.881	0.939
Non-learned defocus (DfD)	0.131	0.252	0.022	0.953	0.975	0.984
Ikoma *et al*. [19] w/ pinv	0.125	0.219	0.020	0.963	0.983	0.990
Baseline (Ours)	<u>0.060</u>	<u>0.120</u>	<u>0.009</u>	<u>0.986</u>	<u>0.994</u>	<u>0.996</u>
Proposed model (Ours)	**0.050**	**0.098**	**0.008**	**0.987**	**0.995**	**0.997**

TABLE 3
Performances for different modulations with PADNet on FlyingThing3D dataset. All in focus: ground truth (GT) all-in-focus image without any modulations. Non-learned defocus: a conventional defocus blur with the same f-number as our setting. Learned defocus: a variant of defocus blur engineered by a learned phase modulated aperture [19]. Best results are in bold, second best are underlined.

Modulation	Depth					
	MAE ↓	RMSE ↓	$\log_{10}$ ↓	$\delta < 1.25$ ↑	$\delta < 1.25^2$ ↑	$\delta < 1.25^3$ ↑
All in focus	0.331	0.427	0.068	0.669	0.810	0.866
Non-learned defocus	<u>0.071</u>	0.145	0.024	0.925	0.963	0.976
Learned defocus	0.075	<u>0.130</u>	<u>0.012</u>	<u>0.933</u>	<u>0.968</u>	<u>0.981</u>
Proposed modulation	**0.038**	**0.084**	**0.007**	**0.956**	**0.978**	**0.986**

areas with specular reflections, when the unpolarized scene assumption is not satisfied.

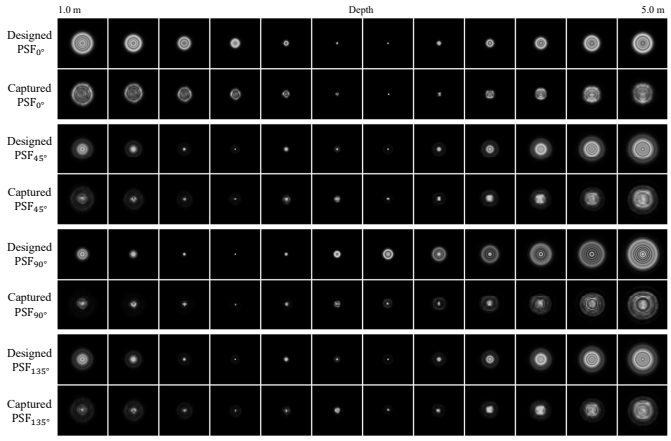

Fig. 12. Point spread functions (PSFs) of PoMI system. For each linearly polarized channel, the first row shows the PSF designed by our end-to-end optimization framework and the second row shows the PSF captured using our prototype.

In addition, we perform the quantitative experiment to validate the depth accuracy of our system prototype. In this experiment, we reconstruct the depth of a barrel positioned at different known distances. For quantitative analysis, we calculate the mean value and root mean square error (RMSE) corresponding to the region of barrel in each estimated depth map. The evaluation results shown in Fig. 14 prove that the depth accuracy of our system prototype is in good agreement with the ground truth.

5 Conclusion and Discussion

This paper proposes a Polarization-multiplexed Modulation Imager system, which could encode the depth information into multiple polarization channels with a single polarization-dependent modulator. A differentiable tensor-wise polarization-dependent modulation camera model is built based on the Fourier optics with the polarization states considered, enabling the forward simulation and back-propagation of the proposed system. The paper conducts extensive experiments to verify the effectiveness and the generalization of the proposed method quantitatively and

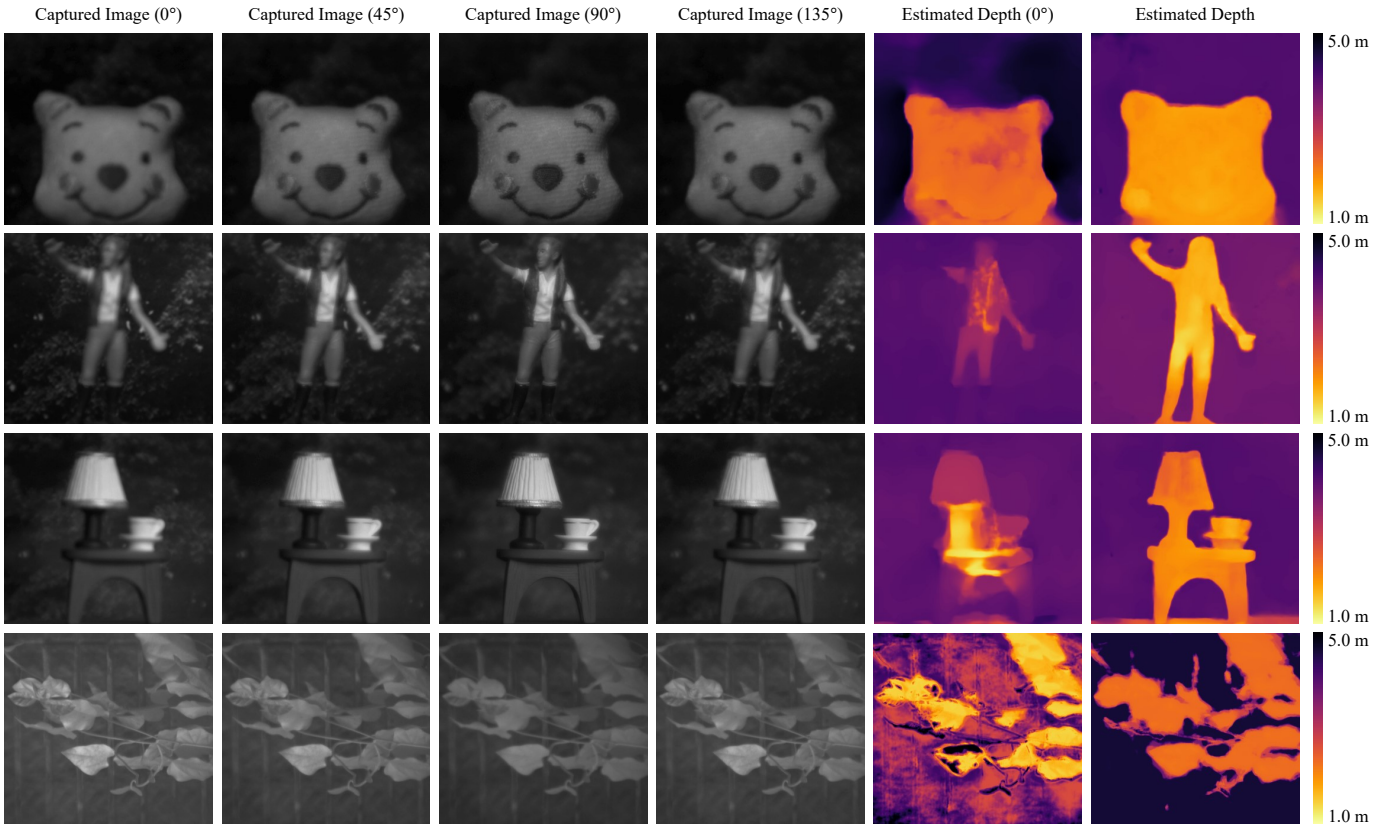

Fig. 13. Experimental results on real capture data. For each scene, the first four columns are the captured images in four polarization channels. The fifth column is depth map estimated by PADNet trained with captured images in 0° polarization channel. The sixth column is depth map estimated by PADNet trained with multiplexed polarization channels. To keep the captured scenes relatively centered in the field of view, the size of foreground objects in the indoor scenes and outdoor scenes is set to be about 15 $\times$ 15 cm and 20 $\times$ 20 cm.

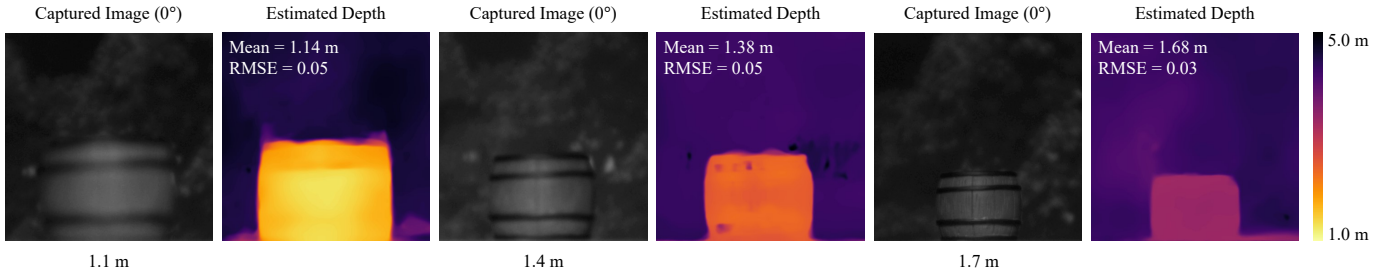

Fig. 14. Quantitative analysis of physical experiments by our system prototype. For each distance, mean value and root mean square error in the upper-left corner of the estimated depth map represents the depth accuracy quantitatively.

qualitatively and build a prototype system to prove the feasibility of the proposed method for natural scenes.

At the same time, there are still some limitations in our work. Here, we summarize these limitations and future work as follows: 1.*Unpolarized scene assumption*: In our work, the assumption that the captured scene is unpolarized causes a poor performance in areas with specular reflections, which is polarized. To overcome this deficiency, the analytic physical prior utilized in shape from polarization [68] could be introduced to help infer the depth information of polarized areas. 2. *System complexity and low portability*: A promising option is to replace the reflective LC-SLM by meta-surface elements, so as to convert the scheme of our system from relay optics to transmission-mode, thereby largely improving the system compactness. 3. *Performance improvement:* To further improve our system's reliability and stability, we clarify potential solutions in terms of modulation freedom and refinement procedure. Theoretically, the four linear polarization channels utilized in the proposed system are not completely independent, which limits the modulation freedom. It would be an ideal prospect to capture four independent channels by using a full-Stokes polarization camera. With four independent channels, we are able to modulate more depth cues. In addition, we will introduce modifications such as data augmentation for illuminance, extra physical constraints and robustness testing to the refinement procedure. Besides, we will also extend the PoMI system to other domain-specific tasks, such as snapshot compressive sensing, light field imaging, and particle localization and tracking applications in microscopy.

ACKNOWLEDGMENTS

This work was supported by National Key Research and Development Program of China (2022YFA-1207200), NSFC Projects 61971465, and Fundamental Research Funds for the Central Universities, China (Grant No. 0210-14380184).

REFERENCES

[1] N. Metni, T. Hamel, and F. Derkx, "Visual tracking control of aerial robotic systems with adaptive depth estimation," in *Proceedings of IEEE Conference on Decision and Control*, 2005, pp. 6078–6084.

[2] J. Stowers, M. Hayes, and A. Bainbridge-Smith, "Altitude control of a quadrotor helicopter using depth map from microsoft kinect sensor," in *Proceedings of IEEE International Conference on Mechatronics*, 2011, pp. 358–362.

[3] Y. Wang, W.-L. Chao, D. Garg, B. Hariharan, M. Campbell, and K. Q. Weinberger, "Pseudo-lidar from visual depth estimation: Bridging the gap in 3d object detection for autonomous driving," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 8445–8453.

[4] X. Chen, H. Ma, J. Wan, B. Li, and T. Xia, "Multi-view 3d object detection network for autonomous driving," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1907–1915.

[5] M. Kalia, N. Navab, and T. Salcudean, "A real-time interactive augmented reality depth estimation technique for surgical robotics," in *Proceedings of International Conference on Robotics and Automation*, 2019, pp. 8291–8297.

[6] F. El Jamiy and R. Marsh, "Distance estimation in virtual reality and augmented reality: A survey," in *Proceedings of IEEE International Conference on Electro Information Technology*, 2019, pp. 063–068.

[7] S. Izadi, D. Kim, O. Hilliges, D. Molyneaux, R. Newcombe, P. Kohli, J. Shotton, S. Hodges, D. Freeman, A. Davison *et al.*, "Kinectfusion: real-time 3d reconstruction and interaction using a moving depth camera," in *Proceedings of Proceedings of ACM Symposium on User Interface Software and Technology*, 2011, pp. 559–568.

[8] A. Agarwal and C. Arora, "Attention attention everywhere: Monocular depth prediction with skip attention," in *Proceedings of IEEE Winter Conference on Applications of Computer Vision*, 2023, pp. 5861–5870.

[9] S. F. Bhat, I. Alhashim, and P. Wonka, "Adabins: Depth estimation using adaptive bins," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2021, pp. 4009–4018.

[10] R. de Queiroz Mendes, E. G. Ribeiro, N. dos Santos Rosa, and V. Grassi Jr, "On deep learning techniques to boost monocular depth estimation for autonomous navigation," *Robotics and Autonomous Systems*, vol. 136, p. 103701, 2021.

[11] H. Fu, M. Gong, C. Wang, K. Batmanghelich, and D. Tao, "Deep ordinal regression network for monocular depth estimation," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 2002–2011.

[12] A. Levin, R. Fergus, F. Durand, and W. T. Freeman, "Image and depth from a conventional camera with a coded aperture," *ACM Transactions on Graphics*, vol. 26, no. 3, pp. 70–es, 2007.

[13] A. Veeraraghavan, R. Raskar, A. Agrawal, A. Mohan, and J. Tumblin, "Dappled photography: Mask enhanced cameras for heterodyned light fields and coded aperture refocusing," *ACM Trans. Graph.*, vol. 26, no. 3, pp. 69–es, 2007.

[14] C. Zhou, S. Lin, and S. Nayar, "Coded aperture pairs for depth from defocus," in *Proceedings of IEEE International Conference on Computer Vision*, 2009, pp. 325–332.

[15] P. A. Shedligeri, S. Mohan, and K. Mitra, "Data driven coded aperture design for depth recovery," in *Proceedings of IEEE International Conference on Image Processing*, 2017, pp. 56–60.

[16] H. Haim, S. Elmalem, R. Giryes, A. M. Bronstein, and E. Marom, "Depth estimation from a single image using deep learned phase coded mask," *IEEE Transactions on Computational Imaging*, vol. 4, no. 3, pp. 298–310, 2018.

[17] Y. Wu, V. Boominathan, H. Chen, A. Sankaranarayanan, and A. Veeraraghavan, "Phasecam3d-learning phase masks for passive single view depth estimation," in *Proceedings of IEEE International Conference on Computational Photography*, 2019, pp. 1–12.

[18] J. Chang and G. Wetzstein, "Deep optics for monocular depth estimation and 3d object detection," in *Proceedings of IEEE International Conference on Computer Vision*, 2019, pp. 10 193–10 202.

[19] H. Ikoma, C. M. Nguyen, C. A. Metzler, Y. Peng, and G. Wetzstein, "Depth from defocus with learned optics for imaging and occlusion-aware depth estimation," in *Proceedings of IEEE International Conference on Computational Photography*, 2021, pp. 1–12.

[20] C. Roider, A. Jesacher, S. Bernet, and M. Ritsch-Marte, "Axial super-localisation using rotating point spread functions shaped by polarisation-dependent phase modulation," *Optics Express*, vol. 22, no. 4, pp. 4029–4037, 2014.

[21] H. Ikoma, T. Kudo, Y. Peng, M. Broxton, and G. Wetzstein, "Deep learning multi-shot 3d localization microscopy using hybrid optical–electronic computing," *Optics Letters*, vol. 46, no. 24, pp. 6023–6026, 2021.

[22] B. Ghanekar, V. Saragadam, D. Mehra, A.-K. Gustavsson, A. C. Sankaranarayanan, and A. Veeraraghavan, "Ps ^{2} f: Polarized spiral point spread function for single-shot 3d sensing," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–12, 2022.

[23] S. K. Nayar and Y. Nakagawa, "Shape from focus," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 16, no. 8, pp. 824–831, 1994.

[24] M. Subbarao and G. Surya, "Depth from defocus: A spatial domain approach," *International Journal of Computer Vision*, vol. 13, no. 3, pp. 271–294, 1994.

[25] H. Tang, S. Cohen, B. Price, S. Schiller, and K. N. Kutulakos, "Depth from defocus in the wild," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 2740–2748.

[26] Y. Shechtman, S. J. Sahl, A. S. Backer, and W. E. Moerner, "Optimal point spread function design for 3d imaging," *Physical review letters*, vol. 113, no. 13, p. 133902, 2014.

[27] S. R. P. Pavani, M. A. Thompson, J. S. Biteen, S. J. Lord, N. Liu, R. J. Twieg, R. Piestun, and W. E. Moerner, "Three-dimensional, single-molecule fluorescence imaging beyond the diffraction limit by using a double-helix point spread function," *Proceedings of National Academy of Sciences*, vol. 106, no. 9, pp. 2995–2999, 2009.

[28] E. Nehme, D. Freedman, R. Gordon, B. Ferdman, L. E. Weiss, O. Alalouf, T. Naor, R. Orange, T. Michaeli, and Y. Shechtman, "Deepstorm3d: dense 3d localization microscopy and psf design by deep learning," *Nature methods*, vol. 17, no. 7, pp. 734–740, 2020.

[29] K. Marwah, G. Wetzstein, Y. Bando, and R. Raskar, "Compressive light field photography using overcomplete dictionaries and optimized projections," *ACM Transactions on Graphics*, vol. 32, no. 4, pp. 1–12, 2013.

[30] H. Arguello and G. R. Arce, "Colored coded aperture design by concentration of measure in compressive spectral imaging," *IEEE Transactions on Image Processing*, vol. 23, no. 4, pp. 1896–1908, 2014.

[31] Y. Hua, S. Nakamura, M. S. Asif, and A. C. Sankaranarayanan, "Sweepcamdepth-aware lensless imaging using programmable masks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 7, pp. 1606–1617, 2020.

[32] Y. Zheng, Y. Hua, A. C. Sankaranarayanan, and M. S. Asif, "A simple framework for 3d lensless imaging with programmable masks," in *Proceedings of IEEE International Conference on Computer Vision*, 2021, pp. 2603–2612.

[33] Y. Han, K. Zhao, and Z. You, "Development of rapid rotary polarization imaging detection devices," *Optics and Precision Engineering*, vol. 26, no. 10, pp. 7–16, 2018.

[34] S. Alali, T. Yang, and I. A. Vitkin, "Rapid time-gated polarimetric stokes imaging using photoelastic modulators," *Optics Letters*, vol. 38, no. 16, pp. 2997–3000, 2013.

[35] O. Arteaga, J. Freudenthal, B. Wang, and B. Kahr, "Mueller matrix polarimetry with four photoelastic modulators: theory and calibration," *Applied Optics*, vol. 51, no. 28, pp. 6805–6817, 2012.

[36] D. Rebhan, M. Rosenberger, and G. Notni, "Principle investigations on polarization image sensors," in *Proceedings of Photonics and Education in Measurement Science*, vol. 11144, 2019, pp. 50–54.

[37] Y. Ba, A. Gilbert, F. Wang, J. Yang, R. Chen, Y. Wang, L. Yan, B. Shi, and A. Kadambi, "Deep shape from polarization," in *Proceedings of European Conference Computer Vision*, 2020, pp. 554–571.

[38] C. Lei, C. Qi, J. Xie, N. Fan, V. Koltun, and Q. Chen, "Shape from polarization for complex scenes in the wild," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2022, pp. 12 632–12 641.

[39] Y. Verdié, J. Song, B. Mas, B. Busam, A. Leonardis, and S. McDonagh, "Cromo: Cross-modal learning for monocular depth es-

timation," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2022, pp. 3927–3937.
[40] A. Kalra, V. Taamazyan, S. K. Rao, K. Venkataraman, R. Raskar, and A. Kadambi, "Deep polarization cues for transparent object segmentation," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2020, pp. 8602–8611.
[41] K. Xiang, K. Yang, and K. Wang, "Polarization-driven semantic segmentation via efficient attention-bridged fusion," *Optics Express*, vol. 29, no. 4, pp. 4802–4820, 2021.
[42] H. Mei, B. Dong, W. Dong, J. Yang, S.-H. Baek, F. Heide, P. Peers, X. Wei, and X. Yang, "Glass segmentation using intensity and spectral polarization cues," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2022, pp. 12 622–12 631.
[43] C. Zhou, M. Teng, Y. Han, C. Xu, and B. Shi, "Learning to dehaze with polarization," vol. 34, pp. 11 487–11 500, 2021.
[44] S. Fang, X. Xia, X. Huo, and C. Chen, "Image dehazing using polarization effects of objects and airlight," *Optics Express*, vol. 22, no. 16, pp. 19 523–19 537, 2014.
[45] J. Tan, J. Zhang, and Y. Zhang, "Target detection for polarized hyperspectral images based on tensor decomposition," *IEEE Geoscience and Remote Sensing Letters*, vol. 14, no. 5, pp. 674–678, 2017.
[46] H. Zhao, Z. Ji, Y. Zhang, X. Sun, P. Song, and Y. Li, "Mid-infrared imaging system based on polarizers for detecting marine targets covered in sun glint," *Optics Express*, vol. 24, no. 15, pp. 16 396–16 409, 2016.
[47] R. Wu, J. Suo, F. Dai, Y. Zhang, and Q. Dai, "Scattering robust 3d reconstruction via polarized transient imaging," *Optics Letters*, vol. 41, no. 17, pp. 3948–3951, 2016.
[48] J. W. Goodman, *Introduction to Fourier optics*. Roberts and Company publishers, 2005.
[49] X. Dun, H. Ikoma, G. Wetzstein, Z. Wang, X. Cheng, and Y. Peng, "Learned rotationally symmetric diffractive achromat for full-spectrum computational imaging," *Optica*, vol. 7, no. 8, pp. 913–922, 2020.
[50] J. J. Gil and R. Ossikovski, *Polarized light and the Mueller matrix approach*. CRC press, 2022.
[51] S. W. Hasinoff and K. N. Kutulakos, "A layer-based restoration framework for variable-aperture photography," in *Proceedings of IEEE International Conference on Computer Vision*, 2007, pp. 1–8.
[52] Q. Xie, M.-T. Luong, E. Hovy, and Q. V. Le, "Self-training with noisy student improves imagenet classification," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2020, pp. 10 687–10 698.
[53] X. Hu, K. Yang, L. Fei, and K. Wang, "Acnet: Attention based network to exploit complementary features for rgbd semantic segmentation," in *Proceedings of IEEE International Conference on Image Processing*, 2019, pp. 1440–1444.
[54] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga *et al.*, "Pytorch: An imperative style, high-performance deep learning library," in *Proceedings of Advances in Neural Information Processing Systems*, vol. 32, 2019.
[55] I. Moreno, P. Velásquez, C. Fernández-Pousa, M. Sánchez-López, and F. Mateos, "Jones matrix method for predicting and optimizing the optical modulation properties of a liquid-crystal display," *Journal of Applied Physics*, vol. 94, no. 6, pp. 3697–3702, 2003.
[56] B. Ma, B. Yao, T. Ye, and M. Lei, "Prediction of optical modulation properties of twisted-nematic liquid-crystal display by improved measurement of jones matrix," *Journal of Applied Physics*, vol. 107, no. 7, p. 073107, 2010.
[57] J. del Hoyo, L. M. Sanchez-Brea, and A. Soria-Garcia, "Calibration method to determine the complete jones matrix of slms," *Optics and Lasers in Engineering*, vol. 151, p. 106914, 2022.
[58] M. Bass, *Handbook of optics: volume I-geometrical and physical optics, polarized light, components and instruments*. McGraw-Hill Education, 2010.
[59] P. Favaro and S. Soatto, "A geometric approach to shape from defocus," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 3, pp. 406–417, 2005.
[60] P. Favaro, S. Soatto, M. Burger, and S. J. Osher, "Shape from defocus via diffusion," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 3, pp. 518–531, 2008.
[61] C. Zhou, O. Cossairt, and S. Nayar, "Depth from diffusion," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2010, pp. 1110–1117.
[62] N. Mayer, E. Ilg, P. Hausser, P. Fischer, D. Cremers, A. Dosovitskiy, and T. Brox, "A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 4040–4048.
[63] E. Ilg, N. Mayer, T. Saikia, M. Keuper, A. Dosovitskiy, and T. Brox, "Flownet 2.0: Evolution of optical flow estimation with deep networks," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 2462–2470.
[64] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus, "Indoor segmentation and support inference from rgbd images," in *Proceedings of European Conference on Computer Vision*, 2012, pp. 746–760.
[65] J. Hu, M. Ozay, Y. Zhang, and T. Okatani, "Revisiting single image depth estimation: Toward higher resolution maps with accurate object boundaries," in *Proceedings of IEEE Winter Conference on Applications of Computer Vision*, 2019, pp. 1043–1051.
[66] R. Garg, N. Wadhwa, S. Ansari, and J. T. Barron, "Learning single camera depth estimation using dual-pixels," in *Proceedings of IEEE International Conference on Computer Vision*, 2019, pp. 7628–7637.
[67] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Proceedings of Medical Image Computing and Computer-Assisted Intervention*, 2015, pp. 234–241.
[68] S. Rahmann and N. Canterakis, "Reconstruction of specular surfaces using polarization imaging," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2001.

Zhiwei Huang received the BS degree from the School of Electronic Engineering, Xidian University, Xi'an, China, in 2021. He is currently working toward the MS degree with the School of Electronic Science and Engineering, Nanjing University, China. His research interests include computational imaging and image processing.

Mingyou Dai received the BS degree from the School of Physical and Electronic Sciences, Hubei University, Hubei, China, in 2021. He is currently working toward the PHD degree with the School of Electronic Science and Engineering, Nanjing University, China. His research interests include optical microscopy and computational imaging.

Tao Yue received the BS degree in automation from Northwestern Polytechnical University, Xi'an, China, in 2009, and the PhD degree from Tsinghua University, Beijing, China, in 2015. He is currently an Associate Professor with Nanjing University, China. His research interests include computer vision, image processing, and computational imaging.

Xuemei Hu received the BS degree from the Department of Automation, Tsinghua University, China, in 2013, and the PhD degree from Tsinghua University, in 2018. She is currently an associate researcher with Nanjing University, China. Her research interests include computational imaging and image processing.