Title

Learned lensless 3D camera.

Journal

Optics Express, 30(19)

Authors

Tian, Feng

Yang, Weijian

Publication Date

2022-09-12

DOI

10.1364/OE.465933

Peer reviewed

Research Article | Vol. 30, No. 19 / 12 Sep 2022 / Optics Express 34479
Optics EXPRESS

Learned lensless 3D camera

FENG TIAN AND WEIJIAN YANG*

Department of Electrical and Computer Engineering, University of California, Davis, CA 95616, USA
**wejyang@ucdavis.edu*

Abstract: Single-shot three-dimensional (3D) imaging with compact device footprint, high imaging quality, and fast processing speed is challenging in computational imaging. Mask-based lensless imagers, which replace the bulky optics with customized thin optical masks, are portable and lightweight, and can recover 3D object from a snap-shot image. Existing lensless imaging typically requires extensive calibration of its point spread function and heavy computational resources to reconstruct the object. Here we overcome these challenges and demonstrate a compact and learnable lensless 3D camera for real-time photorealistic imaging. We custom designed and fabricated the optical phase mask with an optimized spatial frequency support and axial resolving ability. We developed a simple and robust physics-aware deep learning model with adversarial learning module for real-time depth-resolved photorealistic reconstructions. Our lensless imager does not require calibrating the point spread function and has the capability to resolve depth and "see-through" opaque obstacles to image features being blocked, enabling broad applications in computational imaging.

1. Introduction

Mask-based lensless camera is a special type of cameras where the bulk lenses between the object and the image sensor are replaced by a thin layer of optical modulator. The optical modulator could encode multi-dimensional object information (e.g. three-dimensional space, wavelengths, time etc.) onto a two-dimensional (space) image, enabling compressed imaging through a thin and light-weight device [1–14]. They have great potentials in various applications such as microscopy, wearable and implantable devices, photography, machine vision etc. A common strategy to recover high-dimensional information from the snap-shot two-dimensional (2D) image is iterative optimization [15,16]. Typically, the point spread function (PSF) of the system (i.e. the system matrix in the forward model) needs to be calibrated experimentally, and the reconstruction is slow and computationally expensive [4,5,7]. Deep learning based reconstruction can operate much faster, but it usually requires a large amount of training data and may not guarantee a satisfactory reconstruction quality [17]. For lensless imaging, most of the existing deep learning approaches focus on 2D imaging applications [18,19], and require calibration of the PSF to initialize the network parameters in order to achieve stable performance. In this work, we demonstrate a lensless camera and incorporate physics-aware neural networks as learnable reconstruction models to resolve 3D objects at multiple distances instantly (**Fig. 1**). With the embedded physical model, the neural network is easy to train. We cascaded adversarial generative network with customized comprehensive loss functions to perform photorealistic enhancement of the reconstruction results. Through a joint design of the imaging optics and deep learning reconstruction models, our imager does not require system matrix calibration nor the time-consuming iterative regression algorithms to resolve objects. We experimentally demonstrate the ability of our camera in 3D imaging and imaging objects behind opaque obstacles. To the best of our knowledge, we are the first to demonstrate deep learning data-driven 3D photorealistic reconstruction without system calibration and initializations, and we are the first to demonstrate imaging objects behind obstacles using lensless imagers.

#465933 https://doi.org/10.1364/OE.465933
Journal © 2022 Received 5 Jun 2022; revised 28 Jul 2022; accepted 30 Jul 2022; published 7 Sep 2022

Optics EXPRESS

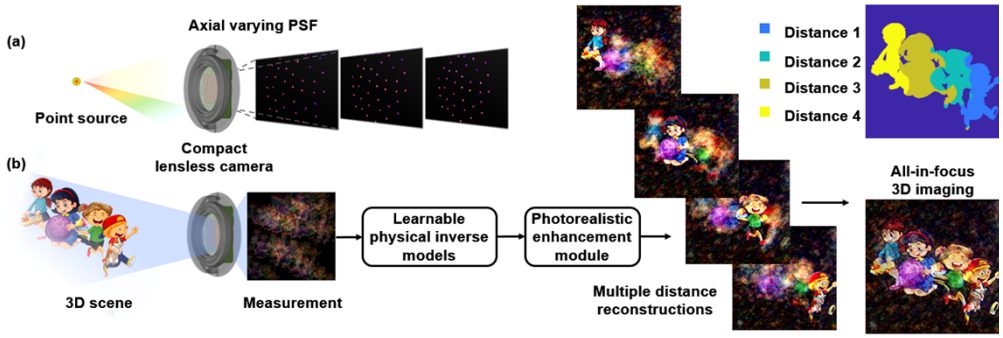

Fig. 1. Overview of the learned 3D lensless imager. (a) Axial varying PSF of a thin layer of microlens array. (b) Image formation and reconstruction pipeline. The 3D scene is captured by the microlens array and reconstructed by trained physical inverse models and photorealistic / perceptual enhancement networks. The all-in-focus 3D image and depth map are synthesized through post-processing. License was obtained from VectorStock to use the cartoon characters.

2. Design of the 3D lensless camera

2.1. Design of the PSF and microlens array

Various types of optical masks, such as amplitude masks like multi-aperture array [4,8,11,20–22] and phase masks such as diffuser [5], have been reported for lensless imaging. In our work, we choose microlens array [7,10,12,14] as the optical mask, as it has a high optical transmission coefficient (compared with amplitude masks [4,8,11,20–22]), and its PSF is locally confined and spatially sparse which reduces image background and thus facilitates reconstruction (compared with the phase masks producing spatially dispersed PSF [5,6]). We designed each lens unit's location to be random so the PSF has a low self-correlation [10,12], which reduces artifacts in reconstruction. We optimized the geometry of the lens unit and the lens pattern so the PSF has a wide frequency support (Fig. 2). We designed the focal length of the lens units to be 15 mm so the PSF has a large enough axial scaling to resolve object depths within the desirable distance range (10~60 cm). The microlens array contains 37 lens units, each being 3 mm in diameter, randomly distributed within an aperture of 12 mm in diameter. The separation distance between adjacent lens units follows the continuous uniform distribution between a lower bound and an upper bound. The fill factor of the microlens array is ~ 0.7. The gaps in between the lens units are covered with a thin layer of opaque polymer. The microlens array is positioned ~ 15 mm from the surface of the image sensor. **Table 1** shows the full specification of the microlens array. See Appendix A for a discussion on how to choose the number of lens units.

2.2. Reconstruction algorithms

The forward model of the imaging system can be described as the image b at the sensor plane being the sum of the convolution between the point spread function PSF_z and the object x_z at each distance z, where the PSF_z is transversely invariant in paraxial approximation:

$$b = \sum_z PSF_z \otimes x_z \tag{1}$$

A common approach to reconstruct the object is to solve the optimization problem, typically with some priors, for example to constrain the estimated object or its gradient to be sparse. This approach requires a known forward model, which can be estimated analytically from design or calibrated experimentally by measuring PSF_z. The object reconstruction may have a single

Optics EXPRESS

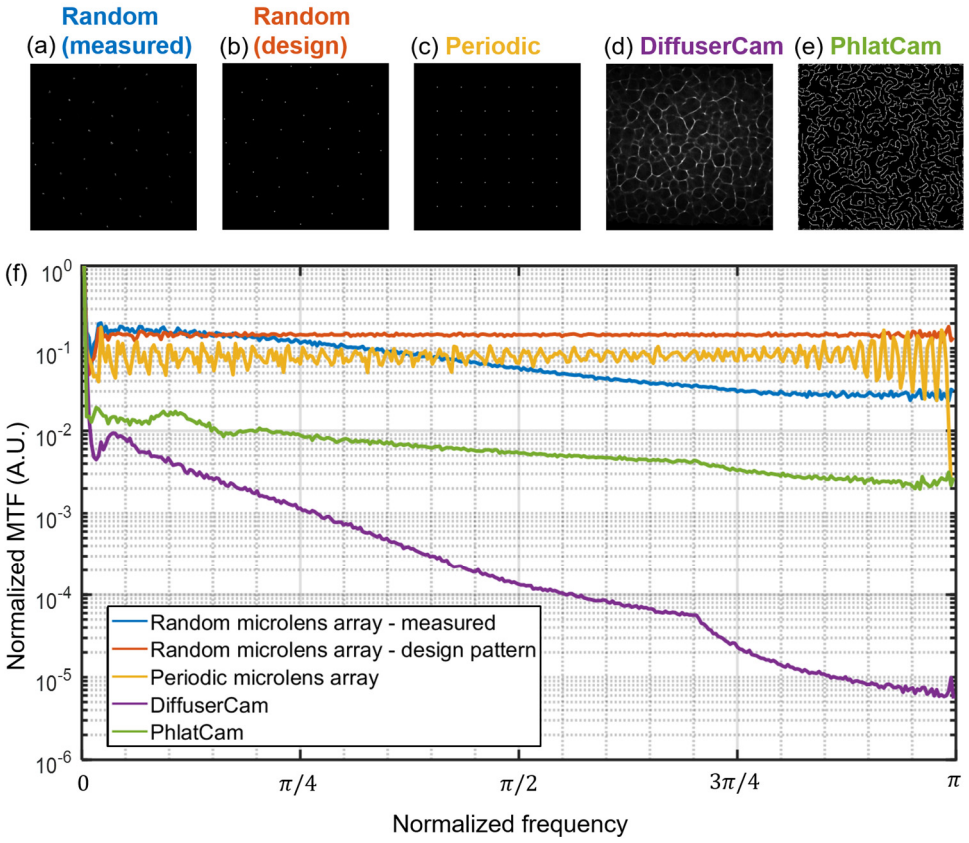

Fig. 2. Comparison of PSF and MTF of different phase masks. (a) Experimentally measured PSF of our microlens array where the positions of individual lens units are randomly set with the constrain that the separation distance between adjacent lens units follows the continuous uniform distribution between two bounds. The total number of lens units is 37. (b) Noise-free, ideal PSF of the designed microlens array. (c) Simulated PSF of a periodic microlens array with 36 lens units. (d) Simulated PSF of DiffuserCam [5]. (e) Simulated PSF of PhlatCam [6]. (f) Normalized MTF of our microlens array, the periodic microlens array, DiffuserCam, and PhlatCam. The PSF in (a)-(c) are Gaussian blurred for better visibility.

Table 1. Design parameters of the microlens array (MLA)

Working distance	10 cm ~ infinity
Lens array area	36 mm × 24 mm (FX format camera)
Lens unit radius	1.5 mm
Refractive index	1.43
MLA to camera sensor distance	~15 mm
Number of microlens units	37
Focal length of the microlens unit	15 mm
Radius of curvature of the microlens unit	6.5 mm
Fill factor	~0.7

Optics EXPRESS

step solution such as through the Wiener filter, or can be solved iteratively using algorithms such as Richardson-Lucy, FISTA [23], or ADMM [24]. The iterative approach results in higher reconstruction quality but takes significantly longer time. Regardless of the exact approach, these iterative algorithms are in general slow and heavily rely on the forward model. If the forward model is not calibrated precisely or some experimental factors are not described well by the convolutional model, the reconstruction quality will be degraded.

We recently developed a parallelized ray tracing algorithm to project each image pixel back to contributive voxels based on the PSF [10] to reconstruct 3D objects in lensless imaging. This can be effectively formulated as transposed convolution in spatial domain, i.e. the object at distance z is approximated as the correlation between the image b and PSF_z. Using Fourier domain analysis, it can be expressed as

$$\hat{x}_z = \mathcal{F}^{-1}[B \odot OTF_z^*] \tag{2}$$

where $\hat{x}_z$ is the refocused object at distance z, B is the Fourier transform of the image b, OTF_z^* is the complex conjugate of the optical transfer function at distance z, $\odot$ represents the elementwise multiplication (Hadamard product), and $\mathcal{F}^{-1}$ represents the inverse Fourier transform. Implementing Eq. (2) in Fourier domain is operationally simple, but a simple transpose convolution alone could result in strong background and artifacts in the reconstruction, particularly when there are overlaps between different sub-images on the camera sensor. Furthermore, it again requires measuring the OTF (or PSF), which could be easily corrupted by noise and thus the reconstruction quality could be degraded.

Here, to overcome the limitations of the above approaches, we implement a physics-aware neural network to reconstruct the object, which not only operates in high speed but also eliminates the complicated process to measure the system PSF. Compared to the typical deep neural network, we embed the physics of the inverse model in the network, which can reduce the required training data and training time. Our network is inspired by the model of Eq. (2). It contains a pair of learnable Hadamard layers, and can be described in Eq. (3)

$$\hat{x}_z = \mathcal{F}^{-1}\left[real(B) \odot W_{zr} + \sqrt{-1}imaginary(B) \odot W_{zi}\right] \tag{3}$$

where W_{zr} and W_{zi} are two learnable filters to reconstruct the object at depth z, and *real* and *imaginary* represents the operator to extract the real and imaginary part of B respectively (Fig. 3). As our PSF is sharp in spatial domain, our inverse process can be more robustly learned in frequency domain. We construct the neural network for each reconstruction distance. For color images, we construct three separate Hadamard layer pairs for R, G, B channels. The reconstructed RGB channel signals is then concatenated into color image (Fig. 4(a)).

There are a few noteworthy features of our method. Firstly, our neural network adopts the simplicity and underlying physics of transpose convolution. It operates efficiently in Fourier domain. We use two different Hadamard layers to handle the real and imaginary operation of the network. Such a strategy avoids complex number operation and thus simplifies the network operation and makes it easier to train. Once the network is trained, it can reconstruct object scenes beyond the field of view used in the training process (Fig. 4(a)–(b)). Furthermore, if the camera sensor is shifted with respect to the microlens array, a high quality reconstruction can still be obtained through the same trained network (Fig. 4(c)). All these indicate the network indeed learns the image inverse process. Secondly, compared to the pure transpose convolution which could result in high background and strong artifacts in the reconstruction, the neural network is data driven and it can learn to suppress the background and artifacts. We extract the learned W_{zr} and W_{zi} in Fourier domain and approximate the learned kernel of the transpose convolution in spatial domain as $\mathcal{F}^{-1}\left(W_{zr} + \sqrt{-1}W_{zi}\right)$ (Fig. 5). This learned kernel resembles the overall system PSF of the microlens array, including the aberration pattern of the PSF of individual microlens unit. This again confirms that our neural network does capture the underlying principle

Optics EXPRESS

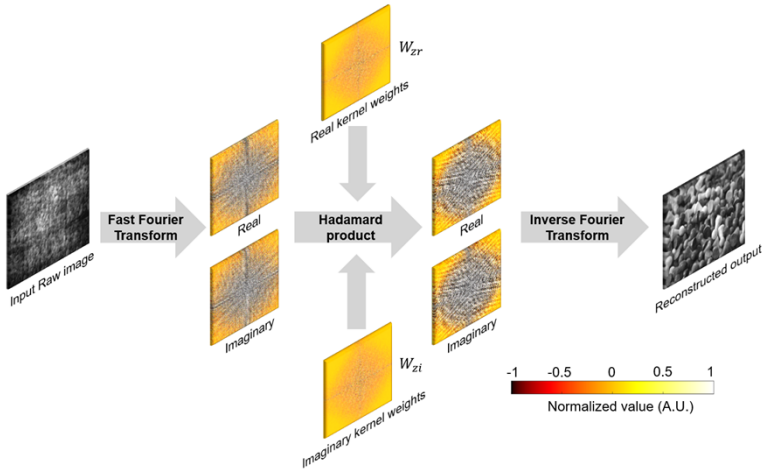

Fig. 3. Schematic of the learnable Hadamard layers in the reconstruction module. Raw image is Fourier transformed to the frequency space, with the real and imaginary component separated. Elementwise multiplication (i.e. Hadamard product) is performed between the learnable kernel weights and the real and imaginary part of the Fourier components of the raw image, respectively. The product is then inverse Fourier transformed to spatial space to obtain the reconstructed output. the shown imge sample is from the PxHere database [25].

of transpose convolution. However, the detailed structure of the individual lobes of the learned kernel is different from the PSF of individual microlens units. In particularly, there appears to be negative pixels in the individual lobes of the kernel, which putatively could be used to suppress the background and artifacts during transpose convolution. In other words, our neural network can capture the features of imaging formation process which cannot be easily modeled by the analytical forward model. A comparison between the reconstruction through pure transpose convolution and our neural network can be found in Sec. 2.4. Thirdly, we trained the RGB channel independently. By using a pixel-wise loss function during training, the learnable Hadamard layers can adapt to the chromatic aberrations of the fabricated microlens array in the three color channels. Fourthly, using the microlens array as the optical mask, we do not need to initialize the network model from PSF measurement, and the network parameter can converge to the physical inverse of the imager fast and stably. This is advantageous compared with other networks that share similar network structures but require specific initialization models of the network parameters [18,26]. We attribute this to the sharp and sparse PSF from the microlens array, which well balances the different frequency domain parameters in the Hadamard layer and makes the network easy to train in the frequency domain.

2.3. Photorealistic enhancement

For photography applications, it is desirable to have high dynamic range and natural rendering of the object scenes. Similar to other works [18,27], we cascaded a photorealistic enhancement module with an adversarial learning model and custom designed loss functions after the basic reconstruction module (Fig. 6). We used a U-net [28] to implement the enhancement module with four loss functions (Appendix B): mean squared error and perceptual loss to measure respectively the distortion and semantic difference between the target image and enhancement module output, adversarial loss to push the distribution of the generated reconstructions close to the target images, and total variation loss to reduce the artifacts generated in the reconstruction module and smooth the sharp gradient changes across the field of view in all color channels.

Optics EXPRESS

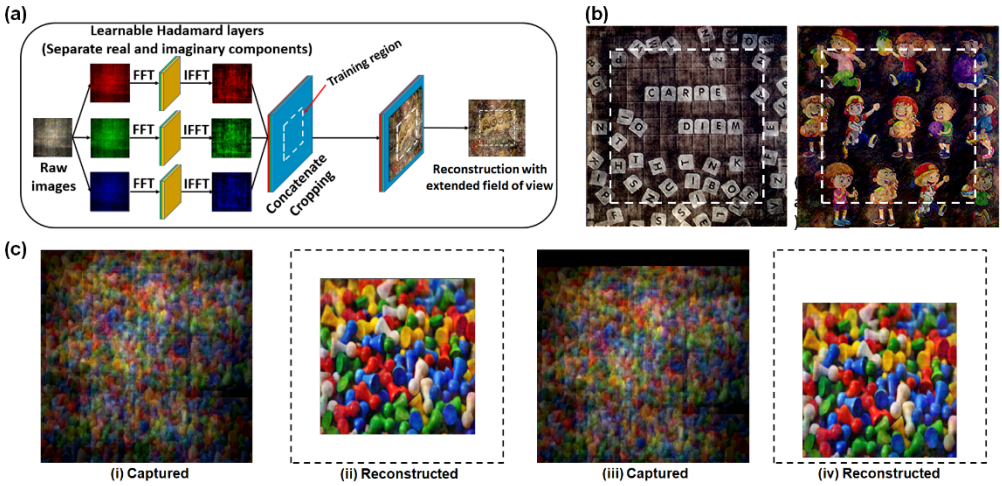

Fig. 4. The trained reconstruction module captures the underlying physics of the imaging system and performs a physics-aware inverse to reconstruct color images. (a) The network architecture of the reconstruction module where three pairs of learnable Hadamard layers are used to reconstruct the R, G, B channels of the color images. Once trained, the network can reconstruct a larger field of view compared to the field of view used in the training (marked as a white dashed box), without adjusting any parameter in the Hadamard layers. (b) Additional examples of reconstructing an extended field of view beyond that in the training. The white dashed box indicates the field of view used in the training. The image sample on the left is from PxHere database [25]. License was obtained from VectorStock to use the cartoon characters for the image sample on the right. In the reconstruction results shown in (a-b), enhancement module (described in Sec. 2.3) is further used to enhance the photorealisticity of the reconstruction. (c) Examples illustrating the trained network can tolerate shifts between the camera sensor with respect to the microlens array. Though the raw captured image on the right panel (iii) is shifted with respect to the left (i), high quality reconstructions (ii and iv) could be obtained using the same trained network. The image sample is from PxHere database [25].

2.4. *Network training and simulation results*

We evaluated the performance of our reconstruction algorithms through simulation (Fig. 7). The reconstruction module and enhancement module are first trained separately, and then together. Compared with the results directly reconstructed from Eq. (2) without training, the outputs of both the reconstruction module and enhancement module have a significantly improved quality, as quantified by the peak signal-to-noise ratio (PSNR), structural similarity index (SSIM), learned perceptual image patch similarity (LPIPS) [29] and the image sharpness (Fig. 7). Pure transposed convolution in frequency domain alone by Eq. (2) results in high background levels and artifacts. By allowing the elements in the Hadamard layers to be trainable, we could suppress the background and artifacts in the reconstructed object, especially in the cases where different sub-images from the microlens array have a large overlap. While the enhancement module may slightly reduce the PSNR and SSIM, it reduces the residue artifacts from the reconstruction module and increase the image sharpness, resulting in an improved LPIPS. Here, we calculated the sharpness as the averaged magnitude of gradients in the reconstructed images.

Optics EXPRESS

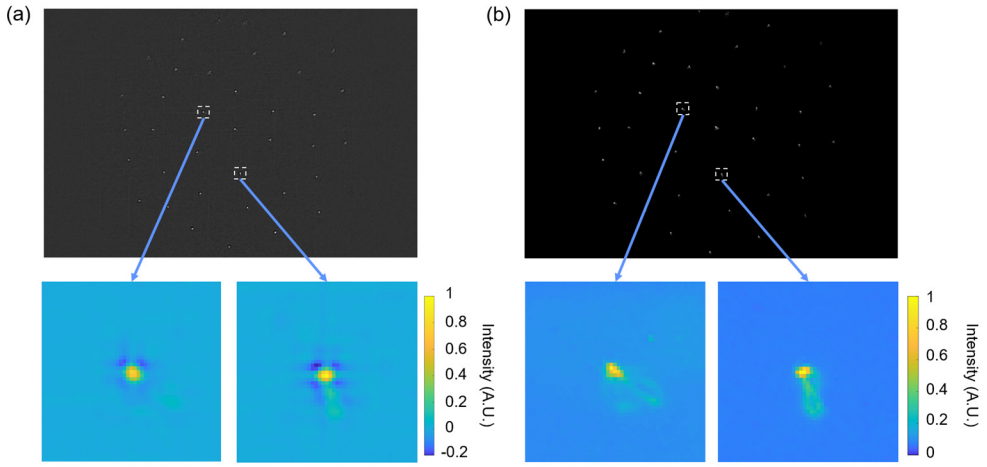

Fig. 5. Comparison of (a) the learned kernel of the Hadamard layer in spatial domain (real component for one color channel) and (b) the PSF of the microlens array. The bottom panel shows the zoom-in view of the represented lobes of the kernel and PSF in the top panel [note the different scale bar for (a) and (b)]. There appears to be negative pixels in the individual lobes in the learned kernel, which putatively function to suppress the background and artifacts in transpose convolution. The results in both (a) and (b) are obtained from experiment (Sec. 3). The kernel and PSF in the top panel are Gaussian blurred for better visibility.

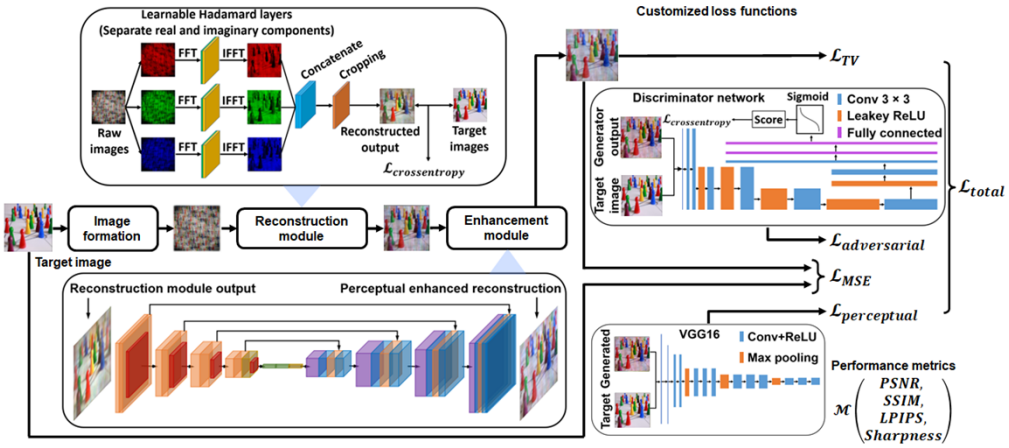

Fig. 6. The overall reconstruction pipeline with reconstruction module and enhancement module. The real target is imaged by the microlens array and then reconstructed and refocused at a specific distance through the reconstruction module using learnable Hadamard layers. Multiple reconstruction modules could be used to reconstruct the target at different distances. Intermediate results are then processed by the enhancement module which is a U-net trained with adversarial learning and four customized loss functions to enhance photorealistic properties. The shown image sample is from PxHere database [25].

Optics EXPRESS

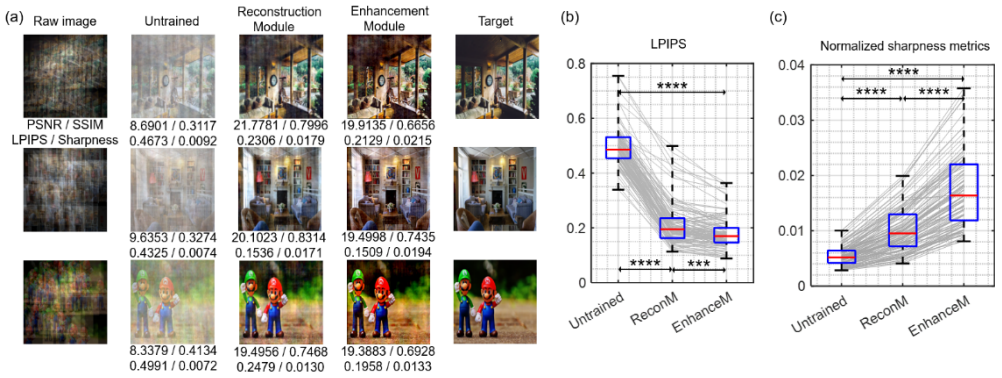

Fig. 7. Simulation results of 2D imaging and reconstruction. (a) Comparison between untrained (Eq. (2)), trained [through reconstruction module (ReconM)] and enhanced [through both reconstruction module and enhancement module (EnhanceM)] reconstructions of 2D images, illustrated by three exemplary images. Compared to the untrained reconstruction, the trained reconstruction module improves the performance metric of PSNR, SSIM, LPIPS and feature sharpness. The enhancement module further improves the LPIPS and feature sharpness. (b) Comparison of LPIPS among untrained, trained and enhanced reconstruction, for a total of 103 samples. (c) Comparison of sharpness among untrained, trained and enhanced reconstruction, for a total of 103 samples. ***, $p<0.001$, ****, $p<0.0001$, in one-way ANOVA. The shown image sample is from PxHere database [25].

3. Experimental results

We fabricated the microlens array using optical transparent polydimethylsiloxane (PDMS) with a negative 3D-printed mold. We mounted the microlens array in close proximity to an image sensor (IMX309BQJ). For each object distance, we trained a reconstruction module where the training dataset [30] is generated by capturing the images from a display monitor placed at the corresponding distance from the lensless camera. In all reconstruction modules, the network parameters are initialized randomly. We trained 14 object distances over ~ 60 cm object distance. For each object distance, we used 8 images for training. All the reconstruction modules share the same enhancement module, where we used 64 images for training. Once the entire network is trained, reconstructing an object takes 46 ms per object depth (including both reconstruction and enhancement module and all three color channels).

3.1. 2D imaging

We first tested the system through 2D images [103] displayed by a monitor placed in front of the lensless camera. We reconstructed the images using the trained networks and compared the results from the untrained method directly computed from Eq. (2) (Fig. 8). Similar to the simulation results, the trained modules reconstructed the objects with a reduced background, higher dynamic range, PSNR and SSIM, improved LPIPS, and increased feature sharpness. The enhancement module further increases the LPIPS and the sharpness of the reconstructed object features. The average PSNR, SSIM and LPIPS of the reconstructed images (after enhancement module) is 16.7159, 0.6270, 0.4969, respectively, for a total of 46 samples.

3.2. 3D imaging

To test the refocusing and 3D imaging capability of the lensless camera, we placed several toy characters in a line (Fig. 9(a)) and captured the raw image through the microlens array (Fig. 9(b)). We used multiple trained reconstruction modules to reconstruct the object scenes at different

Optics EXPRESS

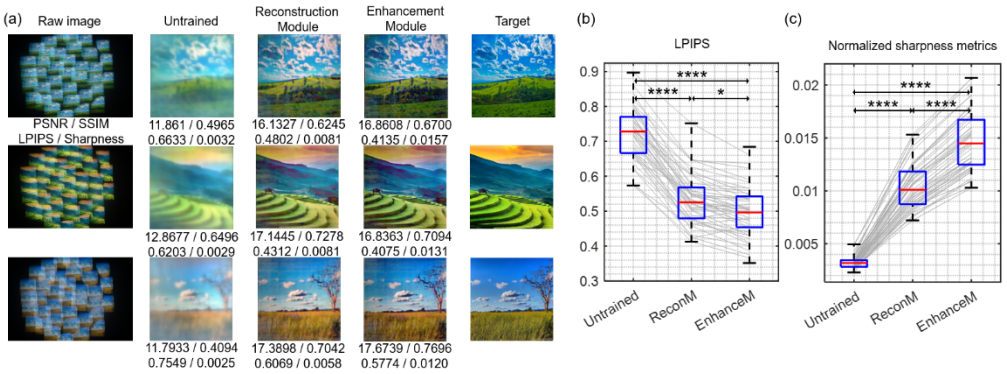

Fig. 8. Experimental results on imaging and reconstructing 2D displays of a monitor in front of the lensless camera. (a) Comparison between untrained (Eq. (2)), trained [through reconstruction module (ReconM)] and enhanced [through both reconstruction module and enhancement module (EnhanceM)] reconstructions of 2D images, illustrated by three exemplary images. (b) Comparison of LPIPS among untrained, trained and enhanced reconstruction, for a total of 46 samples. (c) Comparison of sharpness among untrained, trained and enhanced reconstruction, for a total of 46 samples. *, $p<0.05$, ****, $p<0.0001$, in one-way ANOVA. Due to a color mismatched between the digital images and those displayed on the monitor, an exponentiation operator is used right after the reconstruction module to improve the color saturation and visibility. The image samples are from [30].

distances (Fig. 9(c)–(h)). The reconstructed object scene correctly focuses on the toy characters at the corresponding distance, and blurs those at the foreground and background. This demonstrates the ability of our lensless imager and that the physics-aware reconstruction network indeed learns the depth dependent reconstruction kernel of the microlens array, and can correctly refocus objects at desirable distance from a single exposure. The enhancement module further increases the sharpness of the reconstructed object features. We note that one of the toy characters has a dark face which reflects less amount of light and makes the imaging more challenging. Our reconstruction module and enhancement module could still recover good images, demonstrating the robustness of our reconstruction algorithm.

Based on the refocus object scenes, we rendered an all-in-focus image and a distance map. We applied an edge filter on the reconstructed object scenes at each distance to calculate the local contrast maps. Using the local contrast as weights, we averaged the reconstructed object scenes at different distance to synthesize an all-in-focus image (Fig. 10(a)). For each pixel of the all-in-focus image, we selected the distance with the highest local contrast to construct a distance map of the scenes, followed by thresholding to remove the obvious noise in the background. The distance map (Fig. 10(b)) shows a good agreement with the real object locations.

3.3. *Imaging objects behind obstacles*

In addition to 3D imaging, thanks to the microlens array with a large lateral extension, our lensless camera can image objects behind opaque obstacles. We note that this is a challenging task that conventional cameras with a single stack of lenses cannot typically handle. Here, the key mechanism is that individual microlens unit on the array has different fields of view and can image the objects from different perspectives. As long as there are some perspective views where the objects are not blocked by the obstacles, our lensless camera will be able to capture and reconstruct the objects behind the obstacles from the multiplexed image formation.

To demonstrate this special feature, we placed four opaque objects in front of the lensless camera, where the two objects closer to the camera partially blocked the two objects further

Optics EXPRESS

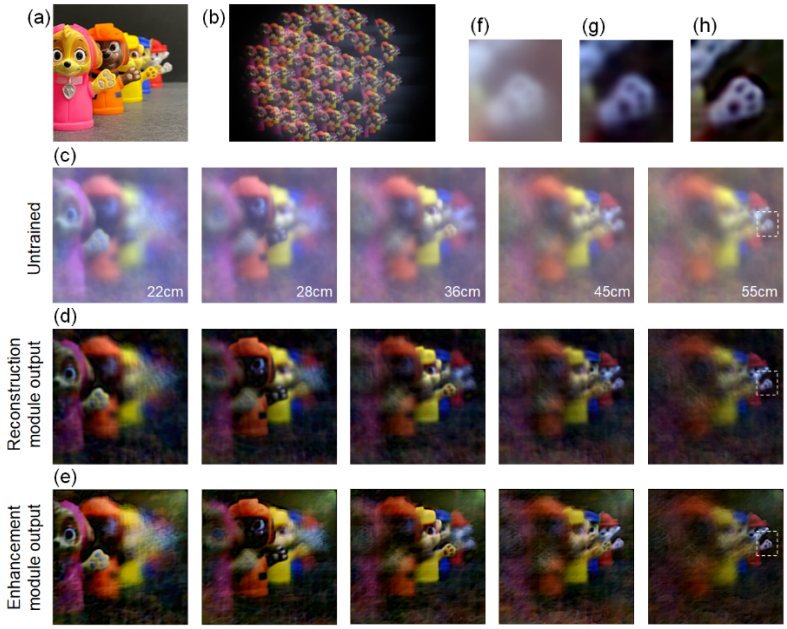

Fig. 9. 3D imaging experiment. (a) Image of the objects lined up at multiple distances, captured by a cell phone camera. (b) Captured image by the lensless 3D camera using microlens array. (c)-(e), untrained, trained and enhanced reconstructions at multiple refocused distances. (f)-(h), zoom-in views from the white dashed box in (c)-(e). The refocused distances are labeled in panel (c). The toy figurines are characters from Nickelodeon and manufactured by Ginsey Home Solutions. Permissions were obtained from Nick.com and Ginsey Home Solutions of publishing images of these toy figurines.

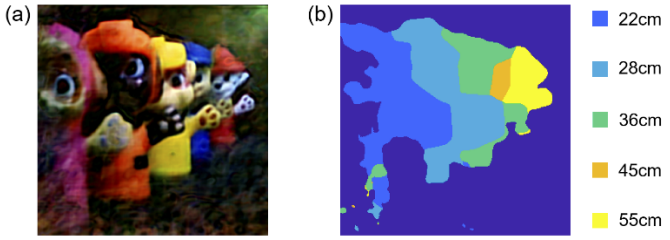

Fig. 10. All-in-focus reconstruction of 3D object scene. (a) All-in-focus image synthesized by weighted average of the refocused scenes based on local contrast. (b) Distance map calculated from the local contrast maps of each refocused scene. The toy figurines are characters from Nickelodeon and manufactured by Ginsey Home Solutions. Permissions were obtained from Nick.com and Ginsey Home Solutions of publishing images of these toy figurines.

Optics EXPRESS

away from the camera, viewed along the direction of the optical axis (Fig. 11(a)–(b)). The raw sub-images from different microlens units show different perspectives of the object scene (Fig. 11(c)–(d)). Though each single sub-image may not contain features of all four objects, our reconstruction algorithms can synthesize the projections from different microlens units, refocus onto individual objects and resolve their features (Fig. 11(e)–(f)).

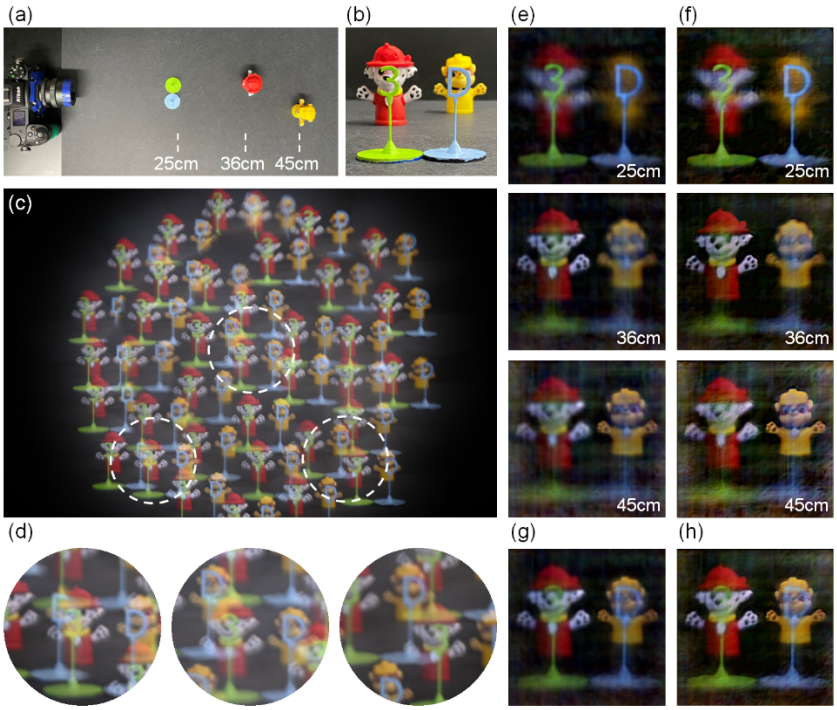

Fig. 11. "See through obstacles" experiment using the lensless camera. (a) Four opaque targets placed in front of the lensless camera where the two targets further away from the camera are partially blocked by the two targets closer to the camera, viewed from the optical axis direction. (b) Front-view image of the four objects at multiple distances, captured by a cell phone camera. (c) Multiplexed captured images where raw sub-images from different microlens units show different perspective views. (d) Zoom-in view of three patches of the sub-images from the white dashed circles in (c). (e-f) The four targets are reconstructed and resolved clearly at their corresponding refocusing distances. (e) shows the results after the reconstruction module, and (f) shows those after the enhancement module. (g-h) All-in-focus reconstruction of the four targets. (g) shows the results after the reconstruction module and (h) shows those after the enhancement module. The refocused distances are labeled in panel (e) and (f). The toy figurines are characters from Nickelodeon and manufactured by Ginsey Home Solutions. Permissions were obtained from Nick.com and Ginsey Home Solutions of publishing images of these toy figurines.

Using the edge filter method that is similar in the 3D imaging demonstration, we can combine the four objects where they are focused all together (Fig. 11(g)–(h)). This all-in-focus result agrees with the prediction where the scene is captured by a camera using a single lens with the same size as the entire microlens array, provided that the depth of focus is large enough to focus onto individual objects simultaneously. To the best of our knowledge, this is the first demonstration of using lensless imager to resolve objects behind obstacles. This is attributed to the large field of view and strong axial resolving capability of our lensless imager.

Optics EXPRESS

4. Discussion

In summary, we built a 3D lensless imager by using a customized microlens array (Fig. 2) to form PSF patterns optimized for depth rendering. We built a physics-aware learnable network to reconstruct objects in frequency domain (Fig. 3 –5), and the algorithm runs in fast speed and requires no measurement of the PSF. We trained a convolutional neural network with adversarial learning to perform photorealistic enhancement (Fig. 6) of the reconstructed 3D scenes (Fig. 9,10). Our lensless camera has a large field of view, and is capable to "see through obstacles" where the target objects are in close distance from the obstacles (Fig. 11).

Our custom designed microlens array performs better than other types of optical modulators used in lensless imagers, in terms of 3D resolving power and image resolution. Compared with amplitude modulators in lensless imagers [4,8,11,20–22], which form images through the shadows of the mask, our phase mask allows higher light transmission and a better axial resolving power in photography application. To form a sharp image, the amplitude modulators typically need to be placed close to the image sensor, rendering a small PSF variation in response to object distance. Our mask allows a larger distance to the image sensor, which increases the PSF dependence on object distance and thus the axial resolving power. Programmable amplitude masks can better resolve distance information [22], however, multiple exposures with different mask patterns are required, which reduces temporal resolution and increases the system complexity. Compared with other phase modulators such as diffuser (represented by DiffuserCam) [5] and contour mask (represented by PhlatCam) [6], our microlens array has a better support in high spatial frequency components, and is better suited to resolve sharp features (Fig. 2). Furthermore, our PSF is sparser so the image is less multiplexed; this allows us to simplify the reconstruction complexity, achieve a faster convergence in training, and reach a higher quality reconstruction (Fig. 12).

Our reconstruction algorithm is co-designed with the microlens array and does not require calibration of the PSF. The network employs the frequency domain reconstruction and is constructed with a simple pair of Hadamard product layers, which embeds an underlying physics model. Similar methods have been reported [18,26], but either the PSF measurement or specific initialization method of the network parameter is required. Benefited by the locally sharp PSF, flat supported spectrum and frequency domain learning strategy, our neural network is easy to train and converges fast. We do not need initialization and thus the PSF calibration in the reconstruction. This reduces the operation complexity. Once the reconstruction network is trained, our microlens array is portable to other sensors without further training. For example, if the camera sensor is shifted with respect to the microlens array, a high quality reconstruction can still be obtained through the same trained network (Fig. 4(c)).

The physics-aware learnable Hadamard product layers are highly efficient. Compared with other neural network such as convolutional neural network that does not account for the underlying image formation process, our method could use smaller amount of network parameters and training time, while achieving an improved PSNR, SSIM and LPIPS (Appendix C), and can handle less sparse objects where there are large sub-image overlaps on the sensor. Though our approach has a slightly worse reconstruction quality compared with the iterative optimization approaches such as ADMM (Appendix D), our neural network reduces the reconstruction time in orders of magnitude. At present, we use different reconstruction modules to reconstruct object scenes at different distances. As each reconstruction module is highly efficient, the overall reconstruction is fast and in real time (~46 ms per depth). The all-in-focus image with the distance map is synthesized through a simple post-processing. Future work could explore a single neural network that can directly output the all-in-focus image and distance map [31], which may further increase the overall reconstruction speed. Finally, future work could also explore an entire end-to-end trained design of forward model and reconstruction neural network [31,32], which could further enhance the overall performance.

Optics EXPRESS

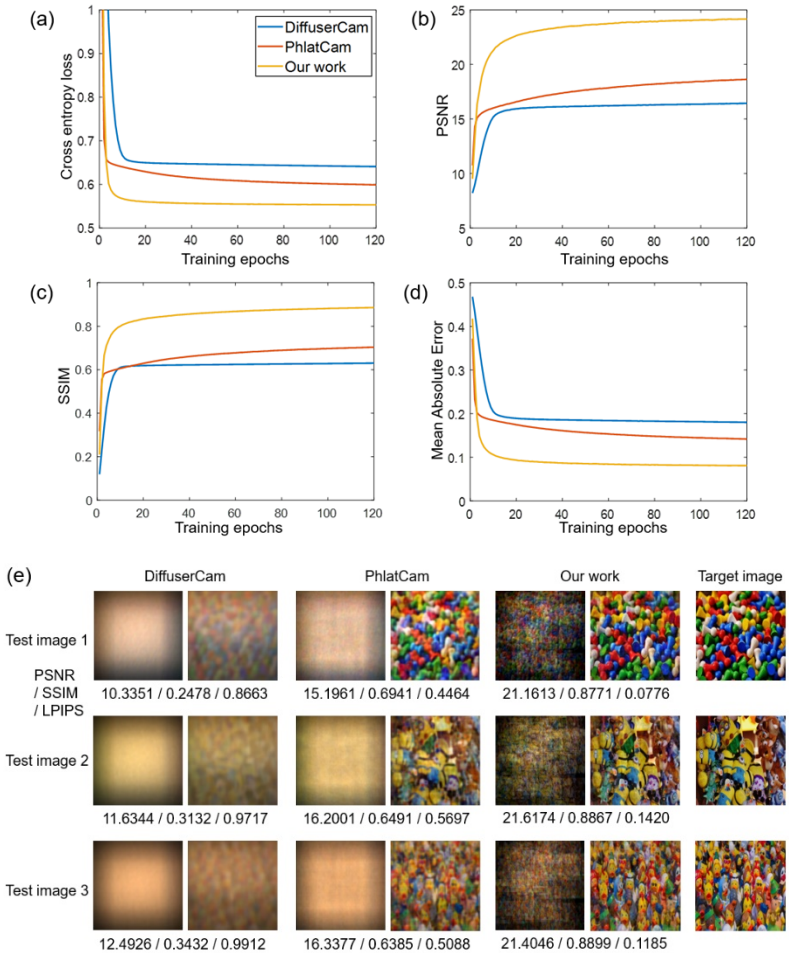

Fig. 12. Comparison of (a-d) the performance metrics of reconstruction module during the training process and (e) the reconstruction results between two existing lensless imagers (DiffuserCam [5] and PhlatCam [6]) and ours, under the same reconstruction network (without parameters initialization) and training dataset, performed by simulation. The loss function in (a) shows faster convergence and smaller converged loss in our work. The other metric functions (b) PSNR, (c) SSIM and (d) MAE all show a higher performance of our imager compared with DiffuserCam and PhlatCam. (e) Raw measurements (left) and reconstruction results (right) of exemplary test images for each lensless imager. Our learnable lensless camera using microlens array shows a better reconstruction quality (PSNR, SSIM and LPIPS) than the DiffuserCam and PhlatCam. The image samples are from PxHere database [25].

Appendix A: Number of microlens units

In our imager using the microlens array, the object is projected onto the sensor by a multiplexed encoding mechanism. Thus there could be substantial overlap between the sub images from each lens unit, and the reconstruction will be ill-posed. From the perspective of ray tracing, we define an occupancy parameter V to describe the percentage of the camera sensor area being illuminated when an object occupies the entire field of view at a single depth [10]:

$$V = \|N \times R \times M^2\|$$

Optics EXPRESS

where M is the system magnification, R is the ratio between the field of view and the sensor area, and N is the total sub image number which is equal to the number of microlens lens units in our application. In general, a small V promotes a higher reconstruction quality under the same reconstruction method and computational complexity, but results in a reduced axial resolving power. We simulate the reconstruction quality for different microlens units and thus the occupancy parameter (Fig. 13). Based on these results, we set the number of microlens unit N to be 37 and a corresponding occupancy parameter ~3 in our real imaging configuration.

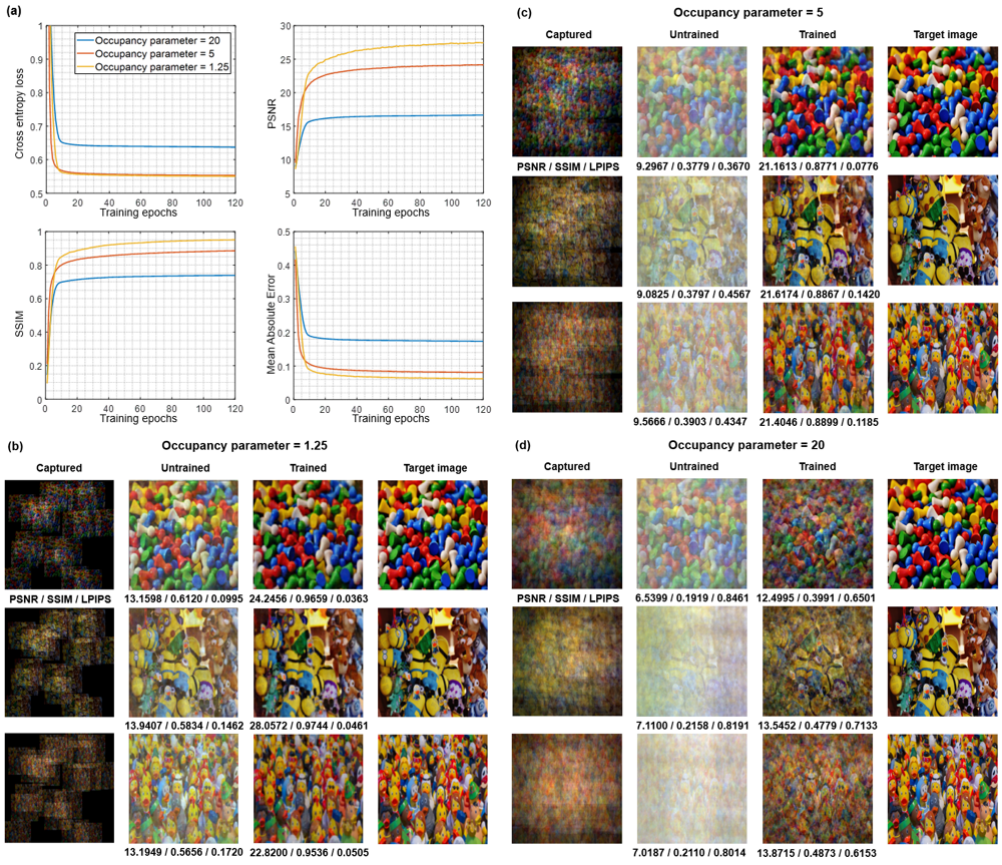

Fig. 13. Comparison of (a) the performance metrics during the training process and (b-d) the reconstruction results of exemplary test images for microlens array with different occupancy parameters. Overall, the reconstruction quality degrades as the occupancy parameters increase. No enhancement module is used in this comparison. The image samples are from PxHere database [25].

Appendix B: Enhancement module and loss functions

We built an enhancement module after the reconstruction module to enhance the reconstruction quality. The enhancement module is realized by a U-net with flexible input size as a generator, shown in Fig. 6 in the main text. A convolutional network is used as a discriminator, which is trained together with the U-net to form an adversarial learning model.

The adversarial learning model is trained with four custom designed loss functions, each with an experiential weight. We summarize the loss functions as follows:

Optics EXPRESS

- Mean squared error (MSE). The MSE loss $\mathcal{L}_{MSE}$ measures the distortion between the target image I_{true} and the enhancement module output I_{pred}. At the beginning of the training when the enhancement module had not yet output reasonable reconstruction or tonic of the image, we assigned a dominant weight of the MSE loss over others. This prevented the convolutional neural network from making up features and being misled by other loss functions such as total variation or perceptual loss.

$$\mathcal{L}_{MSE} = \left\| I_{true} - I_{pred} \right\|_2^2$$

- Perceptual loss. The perceptual loss $\mathcal{L}_{Percept}$ measures the semantic difference between the target image and the enhancement module output. We use a pre-trained VGG-16 [33] on ImageNet to calculate perceptual loss by extracting features on generated output and target image pairs.

$$\mathcal{L}_{Percept} = \left\| \phi(I_{true}) - \phi(I_{pred}) \right\|_2^2$$

  where ϕ represents the VGG-16 network before fully connected layers.

- Adversarial loss. Adversarial loss $\mathcal{L}_{adv}$ differentiates between the target images and the generated reconstruction. It pushes the distribution of the generated reconstruction towards that of the target images. The architecture of the discriminator is shown in Fig. 6. It starts with a resizing layer to resize the input from the generator output (which can have a flexible size) to $256 \times 256 \times 3$ before the convolution layers. Each convolution is $2\times$ strided down sampling, followed by ReLU activation function. The last two layers of the network are fully connected layers. The sigmoid score from 0 to 1 was used to predict the similarity to real target images.

$$\mathcal{L}_{adv} = D(I_{pred})$$

  where D represents the discriminator network.

- Total variation loss. The output of the Hadamard reconstruction module may contain horizontal or vertical fringe artifacts as the matrix is rectangular. Different color channels may also contain small "spikes" of the feature intensity after the inverse Fourier transform. We use total variation loss $\mathcal{L}_{TV}$ to suppress these artifacts and make the image look more natural in all color channels.

$$\mathcal{L}_{TV} = \left\| G(I_{pred}) \right\|_2^2$$

  where G represent the gradient of generated images.

The total loss function of the generator in the enhancement module is thus expressed as below, where λ represents scalar empirical weights

$$\mathcal{L}_{Gen} = \lambda_{MSE}\mathcal{L}_{MSE} + \lambda_{Percept}\mathcal{L}_{Percept} + \lambda_{adv}\mathcal{L}_{adv} + \lambda_{TV}\mathcal{L}_{TV}$$

The loss function of the discriminator network is the binary cross entropy loss $\mathcal{L}_{cross-entropy}$:

$$\mathcal{L}_{cross-entropy} = -\log D(I_{true}) - \log(1 - D(I_{pred}))$$

We used RMSProp optimizer [34] to train both the generator and discriminator. We first trained the reconstruction modules (Hadamard layer). We then froze the Hadamard layer reconstruction modules, and trained the enhancement module (U-net generator and the adversarial learning model). Finally, we unfroze the Hadamard layers and trained both reconstruction modules and the enhancement module together. The full model was trained on NVIDIA Tesla A10G GPU (24G RAM). In experiments, it took ~30 seconds to train each reconstruction module (i.e. for a single refocusing depth) and ~2 hours for the entire enhancement module.

Optics EXPRESS

Appendix C: Comparison between conventional convolutional neural network (CNN) and our physics-aware reconstruction method

We compare the reconstruction quality of a conventional convolutional neural network (CNN) and our physics-aware neural network. We built the CNN using a U-net [28]. It starts with 16 channels in the first convolutional layer. Between each down-sampling or up-sampling layers, there are two convolutional layers. The input is the raw captured image from the lensless camera and the output is the reconstructed target image, same as the physics-aware neural network using Hadamard layers.

To achieve a reasonable reconstruction quality, the U-net uses nearly 10× more parameters and requires >100× more time (>1 hours) in training per reconstruction depth, when compared with our Hadamard reconstruction module, though the pixel numbers of the input and output images are the same between the two networks. In U-net, the convolutional layer focuses on the locality property between the input and output. It cannot capture the global relations between the input and output, which can be learned much easier in frequency domain in the physics-aware Hadamard reconstruction module. As a result, the learning efficiency of U-net is much lower compared to our physics-aware learning method. The reconstruction results of U-net also have a worse PSNR, SSIM and LPIPS compared to the physics-aware method (Fig. 14).

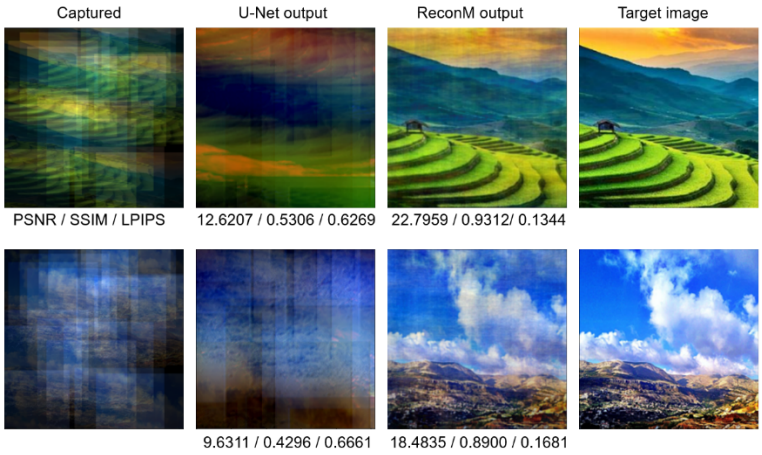

Fig. 14. Comparison of the reconstruction results between a convolutional neural network (U-net) reconstruction model and our physics-aware reconstruction model. Our physics-aware reconstruction model takes much less time and resource to train and converge robustly, while the U-net requires larger number of parameters for sampling and leaves residues from the multiplexed sub images in the reconstruction. In this simulation, the number of parameters in our physics-aware learning module and the U-net is 5.2×10^5 and 3.7×10^6 respectively, with the same input and output data scale. The image samples are from [30].

Appendix D: Reconstruction through iterative optimization approach

We also reconstructed the images using the conventional iterative optimization approaches such as ADMM [24] (Fig. 15). The reconstruction results are better than our approach (Fig. 12), but the reconstruction time is orders of magnitude higher. Our approach could achieve real-time reconstruction whereas the iterative optimization approaches typically take minutes to hours to reconstruct.

Funding. Burroughs Wellcome Fund (CASI 1015761); National Eye Institute (R21EY029472).

Optics EXPRESS

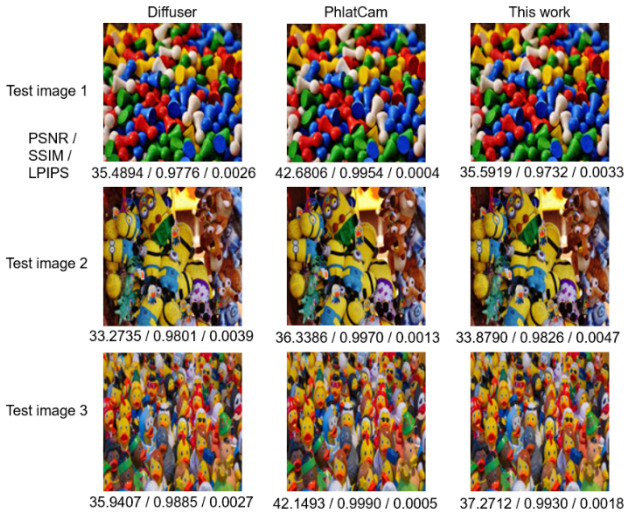

Fig. 15. Reconstruction results of Diffuser [5], PhlatCam [6] and our learnable 3D camera using iterative optimization through ADMM. Compared to our method using the Hadamard layers, the performance of the ADMM approach is higher, but the computational cost and time is substantially larger. The image samples are from [25].

Disclosures. The authors declare no conflict of interest.

Data availability. Data underlying the results presented in this paper are not publicly available at this time but may be obtained from the authors upon reasonable request.

References

1. V. Boominathan, J. K. Adams, M. S. Asif, B. W. Avants, J. T. Robinson, R. G. Baraniuk, A. C. Sankaranarayanan, and A. Veeraraghavan, "Lensless Imaging A computational renaissance," IEEE Signal Process. Mag. **33**(5), 23–35 (2016).
2. V. Boominathan, J. T. Robinson, L. Waller, and A. Veeraraghavan, "Recent advances in lensless imaging," Optica **9**(1), 1–16 (2022).
3. J. N. Mait, G. W. Euliss, and R. A. Athale, "Computational imaging," Adv. Opt. Photonics **10**(2), 409–483 (2018).
4. M. S. Asif, A. Ayremlou, A. Sankaranarayanan, A. Veeraraghavan, and R. G. Baraniuk, "FlatCam: Thin, Lensless Cameras Using Coded Aperture and Computation," IEEE Trans. Comput. Imaging **3**, 384–397 (2017).
5. N. Antipa, G. Kuo, R. Heckel, B. Mildenhall, E. Bostan, R. Ng, and L. Waller, "DiffuserCam: lensless single-exposure 3D imaging," Optica **5**(1), 1–9 (2018).
6. V. Boominathan, J. K. Adams, J. T. Robinson, and A. Veeraraghavan, "PhlatCam: Designed Phase-Mask Based Thin Lensless Camera," IEEE Trans. Pattern Anal. Mach. Intell. **42**(7), 1618–1629 (2020).
7. Y. Xue, I. G. Davison, D. A. Boas, and L. Tian, "Single-shot 3D wide-field fluorescence imaging with a Computational Miniature Mesoscope," Sci. Adv. **6**(43), eabb7508 (2020).
8. J. C. Wu, H. Zhang, W. H. Zhang, G. F. Jin, L. C. Cao, and G. Barbastathis, "Single-shot lensless imaging with fresnel zone aperture and incoherent illumination," Light: Sci. Appl. **9**(1), 53 (2020).
9. Z. Cai, J. Chen, G. Pedrini, W. Osten, X. Liu, and X. Peng, "Lensless light-field imaging through diffuser encoding," Light: Sci. Appl. **9**(1), 143 (2020).
10. F. Tian, J. Hu, and W. Yang, "GEOMScope: Large Field-of-view 3D Lensless Microscopy with Low Computational Complexity," Laser Photonics Rev. **15**(8), 2100072 (2021).
11. J. K. Adams, V. Boominathan, B. W. Avants, D. G. Vercosa, F. Ye, R. G. Baraniuk, J. T. Robinson, and A. Veeraraghavan, "Single-frame 3D fluorescence microscopy with ultraminiature lensless FlatScope," Sci. Adv. **3**(12), e1701548 (2017).
12. G. Kuo, F. Linda Liu, I. Grossrubatscher, R. Ng, and L. Waller, "On-chip fluorescence microscopy with a random microlens diffuser," Opt. Express **28**(6), 8384–8399 (2020).
13. J. K. Adams, D. Yan, J. Wu, V. Boominathan, S. Gao, A. V. Rodriguez, S. Kim, J. Carns, R. Richards-Kortum, C. Kemere, A. Veeraraghavan, and J. T. Robinson, "In vivo lensless microscopy via a phase mask generating diffraction patterns with high-contrast contours," Nat. Biomed. Eng. **6**(5), 617–628 (2022).
14. Y. Xue, Q. Yang, G. Hu, K. Guo, and L. Tian, "Computational Miniature Mesoscope V2: A deep learning-augmented miniaturized microscope for single-shot 3D high-resolution fluorescence imaging," arXiv:2205.00123 (2022).

Optics EXPRESS

15. S. P. Boyd and L. Vandenberghe, *Convex optimization* (Cambridge University Press, 2004).
16. M. S. C. Almeida and M. A. T. Figueiredo, "Deconvolving Images With Unknown Boundaries Using the Alternating Direction Method of Multipliers," IEEE Trans. on Image Process. **22**(8), 3074–3086 (2013).
17. J. D. Rego, K. Kulkarni, and S. Jayasuriya, "Robust Lensless Image Reconstruction via PSF Estimation," *IEEE Winter Conference on Applications of Computer Vision*, 403–412 (2021).
18. S. S. Khan, V. Sundar, V. Boominathan, A. Veeraraghavan, and K. Mitra, "FlatNet: Towards Photorealistic Scene Reconstruction from Lensless Measurements," IEEE Trans. Pattern Anal. Mach. Intell. **44**, 1934–1948 (2022).
19. K. Monakhova, J. Yurtsever, G. Kuo, N. Antipa, K. Yanny, and L. Waller, "Learned reconstructions for practical mask-based lensless imaging," Opt. Express **27**(20), 28075–28090 (2019).
20. E. E. Fenimore and T. M. Cannon, "Coded Aperture Imaging with Uniformly Redundant Arrays," Appl. Opt. **17**(3), 337–347 (1978).
21. H. Zhou, H. J. Feng, Z. X. Hu, Z. H. Xu, Q. Li, and Y. T. Chen, "Lensless cameras using a mask based on almost perfect sequence through deep learning," Opt. Express **28**(20), 30248–30262 (2020).
22. Y. Hua, S. Nakamura, M. S. Asif, and A. C. Sankaranarayanan, "SweepCam - Depth-Aware Lensless Imaging Using Programmable Masks," IEEE Trans. Pattern Anal. Mach. Intell. **42**(7), 1606–1617 (2020).
23. A. Beck and M. Teboulle, "A Fast Iterative Shrinkage-Thresholding Algorithm for Linear Inverse Problems," SIAM J. Imaging Sci. **2**(1), 183–202 (2009).
24. S. P. Boyd, *Distributed optimization and statistical learning via the alternating direction method of multipliers* (Now Publishers Inc., 2011).
25. PxHere, 2022, https://pxhere.com
26. K. Yanny, K. Monakhova, R. W. Shuai, and L. Waller, "Deep learning for fast spatially varying deconvolution," Optica **9**(1), 96–99 (2022).
27. S. S. Khan, V. R. Adarsh, V. Boominathan, J. Tan, A. Veeraraghavan, and K. Mitra, "Towards Photorealistic Reconstruction of Highly Multiplexed Lensless Images," *IEEE/CVF International Conference on Computer Vision*, 7859–7868 (2019).
28. O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," Lect Notes Comput Sc **9351**, 234–241 (2015).
29. R. Zhang, P. Isola, A.A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in IEEE Conference on Computer Vision and Pattern Recognition (2018). pp.586–595.
30. A. Rougetet, "Datasets of pictures of natural landscapes," Kaggle Inc. (2022), https://www.kaggle.com/datasets/arnaud58/landscape-pictures.
31. Y. C. Wu, V. Boominathan, H. J. Chen, A. Sankaranarayanan, and A. Veeraraghavan, "PhaseCam3D-Learning Phase Masks for Passive Single View Depth Estimation," *IEEE International Conference on Computational Photography* (2019).
32. K. Zhang, J. Hu, and W. Yang, "Deep Compressed Imaging via Optimized-Pattern Scanning," Photonics Res. **9**(3), B57–B70 (2021).
33. K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition,"parXiv:1409.1556, (2014).
34. S. Ruder, "An overview of gradient descent optimization algorithms," arXiv:1609.04747 (2016).