Robust Reconstruction With Deep Learning to Handle Model Mismatch in Lensless Imaging

Tianjiao Zeng and Edmund Y. Lam, *Fellow, IEEE*

***Abstract*—Mask-based lensless imaging is an emerging imaging modality, which replaces the lenses with optical elements and makes use of computation to reconstruct images from the multiplexed measurements. Most existing reconstruction algorithms are implemented assuming that the forward imaging process is a convolution operation, with a point spread function based on the system model. In reality, there is model mismatch, leading to inferior image reconstruction results. In this paper, we investigate the impact of model mismatch in mask-based lensless imaging and for the first time, illustrate the accumulated artifacts and information loss due to mismatch error in the state-of-the-art approaches, which perform model-based reconstruction and learning-based enhancement in separate stages. To overcome this, we develop a novel physics-informed deep learning architecture that aims at addressing such mismatch error. The proposed hybrid reconstruction network consists of both unrolled model-based optimization to apply system physics and deep learning layers for model correction. Besides a cascaded enhancement network, we introduce a data-driven branch in parallel, making use of both input measurement and all intermediate outputs from the model-based layers to correct the bias and compensate for the information loss due to model mismatch. The effectiveness and robustness of the proposed model mismatch compensation network, referred to as MMCN, is demonstrated on real lensless images. Experimental results show noticeably better performance for MMCN compared with the alternative methods.**

***Index Terms*—Computational imaging, deep learning, image reconstruction, inverse problems, lensless imaging, optimization.**

I. Introduction

LENSLESS imaging arises due to the desire for smaller and more lightweight imaging systems [1]. The basic idea is to replace the lenses with a modulation mask, which maps each point source in the field-of-view (FoV) into multiple pixels on the sensor (Fig. 1), forming a unique high-contrast point spread function (PSF) [2]–[5]. Unlike the point-to-point imaging structure of a conventional camera, a lensless imaging system encodes the information into a highly multiplexed diffraction pattern and then reconstructs an image through computational recovery algorithms.

Manuscript received April 22, 2021; revised August 3, 2021 and September 15, 2021; accepted September 16, 2021. Date of publication September 22, 2021; date of current version October 14, 2021. This work was supported in part by the Research Grants Council of Hong Kong under Grants GRF 17201818, 17200019, and 17201620, and in part by the University of Hong Kong (104005864). The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Hajime Nagahara. *(Corresponding author: Edmund Y. Lam.)*

The authors are with the Department of Electrical and Electronic Engineering, The University of Hong Kong, Pokfulam, Hong Kong (e-mail: tjzeng99@connect.hku.hk; elam@eee.hku.hk).

Digital Object Identifier 10.1109/TCI.2021.3114542

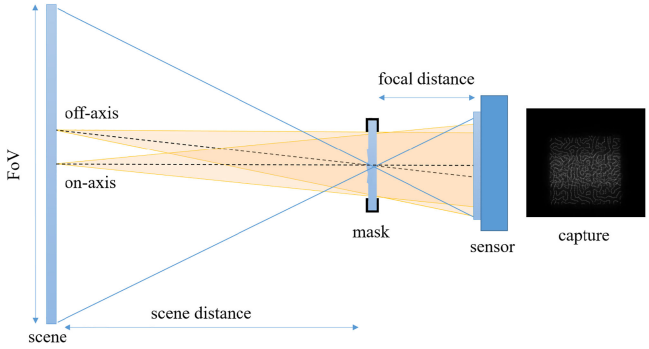
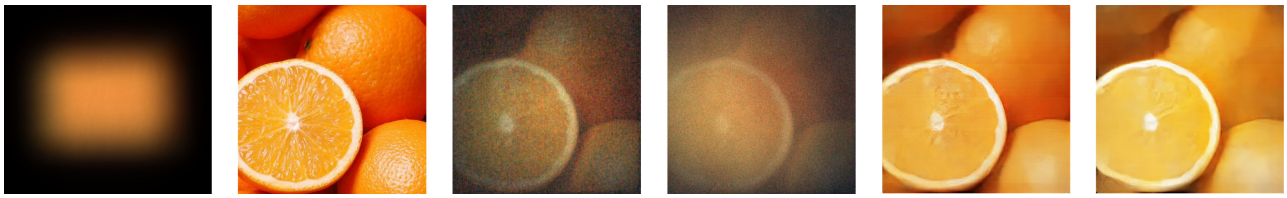


Fig. 1. General geometry of lensless imaging setup. The point sources in the scene are imaged as the superposition of corresponding PSFs on the sensor.

Despite advances in such algorithms, the image quality has much room for improvement. The main culprit is model mismatch. This is due to incomplete or erroneous knowledge in modeling the forward imaging process [6]. The limiting factors that cause model mismatch in lensless imaging are diverse and complex, involving different kinds of imperfections in the optical devices, physical system, experimental procedure/environment, and also mathematical approximations. For example, there are various situations where a convolutional model fails to describe the actual imaging process, such as recording scenes with specular objects or occlusion. In three-dimensional (3D) imaging, the wide depth range can also invalidate the oversimplified convolution model. For this work, we mainly focus on two aspects that result in model mismatch. One is the assumption of lateral shift-invariant PSFs, also known as memory effect [7], [8] in imaging through scattering media. In practice, the variation not only exists but noticeably increases when the angle of incident light gets larger, where the marginal areas would suffer from low reconstructed signal-to-noise ratio (SNR) [9]–[11]. As quantitatively illustrated in [12], the similarity between the on-axis and off-axis PSFs drops to only 0.75 at an angle of 37°. This degrades the resolution particularly around the edge of the FoV. Another aspect is that the on-axis mask pattern used as the system PSF for reconstruction is erroneous, caused by either insufficient accuracy in simulation or misalignment when measuring it experimentally [13]. These problems compound with the ill-conditioning nature of solving the inverse problem in reconstruction, leading to severe artifacts, an example of which is depicted in Fig. 2.

Current existing reconstruction methods in lensless imaging either ignore the model mismatch, or apply learning-based



(a) Raw (b) Latent image (c) Wiener (C-PSF) (d) Wiener (S-PSF) (e) MMCN (C-PSF) (f) MMCN (S-PSF)

Fig. 2. Illustration of the impact of model mismatch on lensless reconstructions. From left to right are (a) raw lensless measurement captured by PhlatCam [3], (b) latent clean image; reconstructed images using Wiener deconvolution with (c) calibrated PSF (C-PSF) and (d) simulated PSF (S-PSF), and also reconstructed ones using proposed model mismatch compensation network (MMCN) with (e) calibrated PSF (C-PSF) and (f) simulated PSF (S-PSF). There is severe degradation of image quality in physics-based reconstructions when leveraging simulated PSF compared to that using experimentally calibrated PSF. In addition, it can be clearly observed that the resolution gradually degrades from central areas to the edges of both physics-based reconstructions, indicating the impact of increasing difference between on-axis and off-axis PSFs. In comparison, the proposed method presents obvious improvement and better robustness in reconstruction performance for both cases.

methods without incorporating any knowledge of the forward imaging model. Although deep learning has achieved significant results in various problems in computational imaging [14]–[22], including image reconstruction, the lack of a physics prior leads to difficulty for pure deep learning networks to restore details from such highly multiplexed patterns. Recently, there are efforts to combine classical methods and deep learning networks [23]–[26]. In mask-based lensless imaging, classic optimization algorithms such as alternating direction method of multipliers (ADMM) [27] and half quadratic splitting (HQS) have been unrolled into networks as a direct attempt of embedding physics-based optimization process in a learning-based framework [28]–[30]. To further improve the performance, recent approaches consider the reconstructed estimate from model-based layers as the preliminary result, and add a CNN-based network such as a U-Net [31] at the end as a learned denoiser for additional perceptual enhancement [28], [32]. These state-of-the-art physics-embedded networks divide the reconstruction of lensless measurements into preliminary reconstruction and enhancement steps, implemented by two separate trainable networks in cascade. Intuitively, this can be seen as mapping the data from the measurements to an intermediate reconstruction based on the forward model of the system, and then to final reconstruction after enhancement by a CNN-based network. However, the preliminary reconstruction operation achieved by classical model-based reconstruction algorithms is not a lossless procedure. Such approaches are susceptible to the model mismatch problem discussed above, since commonly used regularizers, such as total variation (TV) [33], [34], cannot handle the model error effectively. Physical constraints, which impose priors such as smoothness or sparsity, may harm the preservation of crucial details, resulting in information loss in the physics stage. Adding a data-driven refinement stage later may help suppress artifacts, but it is difficult to handle irreversible degradation in physics-based reconstructions. Thus, the cascade of separate operations without taking the model error into account, will feed images with undesired artifacts and information loss into the CNN denoiser, and thus deteriorate the ultimate performance.

To overcome this, we develop a novel hybrid architecture that learns the mapping from the multiplexed measurements to image reconstructions with specific treatment on model mismatch problem, which we call model mismatch compensation network (MMCN). The architecture consists of a model-based block and a deep neural network with an additional compensation branch. This branch formed by multiple CNN streams serves as a correction on model-based optimization reconstruction. For illustration, unrolled ADMM is used to embed the physical model in the network architecture. The features extracted from the intensity updates of each ADMM layer, together with the raw input, are concatenated in the data-driven branch and pass through an expansive path to reconstruct the final image.

This paper describes a model-based reconstruction architecture with correction on mismatch error in lensless imaging. The main contributions are:

- We illustrate how lensless imaging is susceptible to model mismatch and present the first work to do model mismatch compensation for lensless imaging in a completely data-driven fashion without additional calibration or estimation of off-axis parameters.
- We propose a solution in addressing model mismatch error by introducing an additional data-driven compensation branch with multiple CNN streams leveraging all intermediate intensity updates from the unrolled optimization blocks.
- We carry out experiments on real lensless datasets to validate the robustness of the proposed architecture, MMCN, on full-size/cropped measurements given calibrated/simulated PSFs.

II. Related Work

In this section, we overview related studies along three aspects, namely, lensless imaging, image reconstruction and model mismatch. As the model mismatch problem has not been specifically studied in mask-based lensless imaging, works that address this problem in other areas like scattering and image deblurring will be reviewed here.

A. Lensless Imaging

Lensless imaging has been traditionally used for imaging at wavelengths beyond the visible spectrum [35]–[37]. Driven by the growing interest in more compact and lightweight cameras, applications in the visible spectrum have aroused attention and developed rapidly in recent years [38]. By encoding

the incoming light from the scene onto the sensors via modulation masks, lensless cameras eliminate the need for lens and record scene information in non-photorealistic measurements [39]. Mask-based lensless imaging has been demonstrated for various encoding elements including compressive samplers [40], phase gratings [41], spatial light modulators [42], and diffractive masks [3], [12], [43]. The raw measurements captured on the sensor will then be computationally processed to generate photorealistic images resembling the original scene. By introducing this extra computational step to the imaging, the burden brought by unwieldy, expensive hardware can be alleviated and transferred to computation, yielding advantages of smaller size, lighter weight, easier fabrication, and lower cost. These advantages enable lensless imaging to be applied in fields such as vivo imaging, wearables, mobile platforms, and many others requiring ultra-miniature designs [44]. Moreover, the mask-based lensless designs have also been implemented in hyperspectral imaging [45], 3D fluorescence microscopy [46], and refocusable photography [47].

B. Image Reconstruction

Image reconstruction in computational imaging is usually formulated as an ill-posed inverse problem. Typically, the classical reconstruction algorithms involve solving an optimization problem with a data fidelity term and a regularization term exploiting the prior information of the images. These approaches broadly fall into two categories, single-step reconstructions [47]–[50] and iterative reconstructions [3], [51]. Despite fast computations, single-step methods require more stringent restrictions on the imaging including mask fabrication, and would be infeasible for optimization problems without closed-form solutions. Iterative methods are more commonly used in ill-posed inverse problems with generally better performance, but are usually too slow due to a large number of iterations before convergence.

With the development of deep learning, more and more works involving neural networks have been designed to solve image reconstruction problems in computational imaging systems. The deep networks such as CNNs are trained to serve as the inverse operator, which maps raw measurements to latent images [6], [52]. Compared to traditional methods, the learning-based approaches lack physics knowledge and are hard to interpret. Therefore, reconstruction algorithms in the middle ground between the traditional and deep learning methods have been proposed. In [53], [54], neural networks are used to learn the regularizers or proximal mapping in iterative forms. Meanwhile in [28], [55], the neural networks are designed as the unrolled versions of the traditional optimization frameworks. Both increase interpretability by incorporating forward imaging models. Another work [56] leverages the knowledge of the forward model only in the training stage and generalizes the reconstruction model for different imagers with multiple PSFs.

C. Model Mismatch

The model mismatch in lensless imaging is due to the imperfect modeling of the imaging process. Based on the assumption of shift invariance, also known as angular "memory effect" (ME), measurements are commonly regarded as a convolution of the system PSF and intensity pattern of the scene. However, both the convolution model and the system PSF are not free of error, leading to model mismatch in image reconstructions. As this problem has not been specially studied in the literature of mask-based lensless imaging, we review some related work in scattering imaging and image deblurring tasks. The ME-based methods usually model the scattering medium as a linear system under the assumption that the angular span of the scene to be imaged is small enough. As the impinging angles become larger, the mismatch between the mathematical model and exact imaging process becomes non-neglectable, leading to overlapping speckles for objects beyond the ME region. Works have been done to address this problem and expand the angular range for scattering imaging, including stitching multi-view measurements with diverse precalibrated PSFs [57], exploiting prior knowledge of the object [10] and multi-target imaging [58]. In image deblurring, the mismatch of the estimated kernel and true blurry kernel is a common issue and has been shown to produce ringing artifacts in the results. Various approaches have been introduced to handle such error [59]–[63]. Works such as [59], [62] address the kernel error by adding an extra regularizer or an estimation step in the iterative methods. Unrolled networks are also investigated with deep neural networks learning the regularizers and hyper-parameters to handle residual noise in kernels [61], [62].

III. Model Mismatch in Lensless Imaging

In this section, we present the mathematical formulations of lensless imaging, as well as the model mismatch problem, and theoretically explain why image reconstruction in lensless imaging is susceptible to model mismatch.

Different from the point-to-point imaging model of conventional cameras, a lensless imaging system does not seek to capture a spatially accurate reproduction of a scene, but records a multiplexed measurement by applying an encoding optical element in front of the sensor. Assuming that the point sources making up the entire scene are incoherent with each other, we can model the measurements captured by the sensor to be a linear summation of the mask pattern intensities generated by different points in the scene. Such patterns are the point spread functions (PSFs) of the imaging system. This can be mathematically expressed as a matrix-vector multiplication with the generalized forward model matrix $\mathbf{H}$, such that

$$\boldsymbol{b} = \mathbf{H}\boldsymbol{s} + \boldsymbol{n}, \tag{1}$$

where $\boldsymbol{s}$ and $\boldsymbol{b}$ denote the vectorized scene intensity and lensless measurement, respectively, and $\boldsymbol{n}$ represents the additive noise. The PSFs, which can vary with the positions of point sources, are in the columns of $\mathbf{H}$. Supposedly, it would require extensive calibration work on points all over the scene to obtain the exact matrix.

Commonly, though, these PSFs are approximated to be shift-invariant, such that the patterns generated by off-axis point sources are assumed to be the same as the on-axis PSF with a lateral shift. The computational complexity, as well as the

burden on storage and calibration, can be significantly alleviated by modeling the imaging process with a convolution model. The system matrix $\mathbf{H}$ in Eq. (1) can be simplified as a Toeplitz matrix embedding the convolution operation with the on-axis PSF. The reconstruction can then be computed with standard inverse imaging approaches [64].

This approximation allows for computationally tractable modeling of the problem, at the expense of using physically exact PSFs. Moreover, the on-axis PSF, either obtained by calibration during the experiment or simulation based on the mask pattern and imaging geometry, is not error-free. Hence, such errors cause a mismatch between the estimated system and the true imaging system. This can cause severe artifacts in reconstruction quality due to the ill-conditioned nature of a lensless imaging system.

Let the deviation between the biased and true PSF be denoted as $\Delta\mathbf{H}$. We have

$$\boldsymbol{b} = (\tilde{\mathbf{H}} + \Delta\mathbf{H})\boldsymbol{s} + \boldsymbol{n}, \tag{2}$$

where $\tilde{\mathbf{H}}$ represents the biased forward PSF operator due to model mismatch, such that the true operator should be $\mathbf{H} = \tilde{\mathbf{H}} + \Delta\mathbf{H}$. Suppose the model is invertible. The estimated reconstruction can be derived by Taylor series expansion on direct inversion as

$$\begin{aligned}\tilde{\boldsymbol{s}} &= \tilde{\mathbf{H}}^{-1}\boldsymbol{b} \\ &= (\mathbf{H} - \Delta\mathbf{H})^{-1}(\mathbf{H}\boldsymbol{s} + \boldsymbol{n}) \\ &= (\mathbf{I} - \mathbf{H}^{-1}\Delta\mathbf{H})^{-1}(\boldsymbol{s} + \mathbf{H}^{-1}\boldsymbol{n}) \\ &= \boldsymbol{s} + \mathbf{H}^{-1}\Delta\mathbf{H}\boldsymbol{s} + (\mathbf{I} + \mathbf{H}^{-1}\Delta\mathbf{H})\mathbf{H}^{-1}\boldsymbol{n} + \mathcal{O}(\|\Delta\mathbf{H}\|_F^2),\end{aligned} \tag{3}$$

where $\|\cdot\|_F$ is the Frobenius norm. Due to the ill-conditioned system matrix $\mathbf{H}$, the error resulted from model mismatch $\mathbf{H}^{-1}\Delta\mathbf{H}\boldsymbol{s}$ as well as the measurement noise $(\mathbf{I} + \mathbf{H}^{-1}\Delta\mathbf{H})\mathbf{H}^{-1}\boldsymbol{n}$ will be significantly amplified and result in poor recovery [59], [62]. The inverse problem is usually formulated as a regularized optimization problem with the generalized form [64]

$$\hat{\boldsymbol{s}} = \arg\min_{\boldsymbol{s}\geq 0}\left\{\frac{1}{2}\|\boldsymbol{b} - \mathbf{H}\boldsymbol{s}\|_2^2 + \lambda\mathcal{R}(\boldsymbol{s})\right\}, \tag{4}$$

where $\mathcal{R}(\cdot)$ is a regularization function with tuning parameter λ. Conventional regularizations such as Tikhonov and TV are not capable of addressing model mismatch errors [60]. As illustrated in Fig. 2, considerable artifacts exist in the images, indicating how subtle model mismatch is magnified by the ill-conditioned lensless imaging model. In addition, the image resolution gradually degrades from the central areas to the edges. Hence, reconstruction methods with more robust performance to model mismatch problem is essential in lensless imaging.

IV. Methodology

Here, we present our detailed solution to the model mismatch problem using deep learning. The proposed hybrid network combines both physics-based optimization approaches, implemented by unrolled ADMM layers, with CNN-based data-driven methods in a novel structure. In particular, to address the challenges discussed above, we introduce a compensation branch leveraging all intermediate outputs as well as the raw input, which are usually discarded after obtaining the physics-based reconstruction. We first illustrate the implementation of unrolled ADMM layers and then mathematically demonstrate the accumulation of model mismatch error in the physics-based reconstruction stage. Next, we present concrete details on our methodology for dealing with this issue. At last, we give a detailed description of the diagram and also the example configuration used in the following experiments.

A. Unrolled ADMM Blocks

Realistically, a sensor is often not large enough to capture the entire multiplexed measurements. The measurements are therefore cropped to the finite size of the sensor. Let the crop operation be denoted as $\mathbf{\Phi}$ (identity operator for full-size measurements) and the regularization term be set as TV regularization with sparsifying transform $\mathbf{\Psi}$, i.e., $\mathcal{R}(\boldsymbol{s}) = \|\mathbf{\Psi}\boldsymbol{s}\|_1$. The optimization problem is then described as

$$\hat{\boldsymbol{s}} = \arg\min_{\boldsymbol{s}\geq 0}\left\{\frac{1}{2}\|\boldsymbol{b} - \mathbf{\Phi}\mathbf{H}\boldsymbol{s}\|_2^2 + \lambda\|\mathbf{\Psi}\boldsymbol{s}\|_1\right\}. \tag{5}$$

Due to the information loss caused by the cropping operation, it is infeasible to directly invert the mapping and thus iterative optimization is generally needed. By introducing auxiliary variables, the above optimization can be reformulated and solved by iterative ADMM. Following the variable splitting strategy, we can write

$$\begin{aligned}&\text{minimize } \frac{1}{2}\|\boldsymbol{b} - \mathbf{\Phi}\boldsymbol{x}\|_2^2 + \lambda\|\boldsymbol{z}\|_1, \\ &\text{such that } \boldsymbol{x} = \mathbf{H}\boldsymbol{s}, \\ &\qquad\quad \boldsymbol{z} = \mathbf{\Psi}\boldsymbol{s}, \\ &\qquad\quad \boldsymbol{y} = \boldsymbol{s},\ \boldsymbol{y} \geq 0\end{aligned} \tag{6}$$

where $\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}$ are auxiliary variables for ADMM. Using the augmented Lagrangian [65], this can be split into several simplified sub-problems, and each variable is updated by optimizing its corresponding sub-problem with other variables fixed in each iteration [66], [67]. Specifically, the solutions to the sub-problems for the auxiliary variables and intensity $\boldsymbol{s}$ at iteration k are given by [27]

$$\boldsymbol{x}^{(k)} = (\mathbf{\Phi}^\top\mathbf{\Phi} + \rho_x\mathbf{I})^{-1}(\mathbf{\Phi}^\top\boldsymbol{b} + \boldsymbol{u}_x^{(k-1)} + \rho_x\mathbf{H}\boldsymbol{s}^{(k-1)}) \tag{7}$$

$$\boldsymbol{y}^{(k)} = \max\left(\boldsymbol{s}^{(k-1)} + \frac{1}{\rho_y}\boldsymbol{u}_y^{(k-1)}, 0\right) \tag{8}$$

$$\boldsymbol{z}^{(k)} = \mathcal{S}_{\frac{\lambda}{\rho_z}}\left(\mathbf{\Psi}\boldsymbol{s}^{(k-1)} + \frac{1}{\rho_z}\boldsymbol{u}_z^{(k-1)}\right) \tag{9}$$

$$\begin{aligned}\boldsymbol{s}^{(k)} = (\rho_x\mathbf{H}^\top\mathbf{H} + \rho_z\mathbf{\Psi}^\top\mathbf{\Psi} + \rho_y I)^{-1}[\mathbf{H}^\top(\rho_x\boldsymbol{x}^{(k)} \\ - \boldsymbol{u}_x^{(k-1)}) + \boldsymbol{r}^{(k)}],\end{aligned} \tag{10}$$

where $\boldsymbol{r}^{(k)} = \boldsymbol{\Psi}^\top(\rho_z \boldsymbol{z}^{(k)} - \boldsymbol{u}_z^{(k-1)}) + (\rho_y \boldsymbol{y}^{(k)} - \boldsymbol{u}_y^{(k-1)})$. Note that $\boldsymbol{u}_x$, $\boldsymbol{u}_y$, $\boldsymbol{u}_z$ are the dual variables, ρ_x, ρ_y, ρ_z denote positive penalty parameters, and $\mathcal{S}(\cdot)$ is a soft-thresholding operator with parameter λ/ρ_z [68]. We follow the same way to unroll ADMM into a network as introduced in [28]. Each iteration is modeled as a layer of the model-based branch, termed an ADMM block, and the above equations are interpreted as the functions in each block with trainable parameters λ and ρ_x, ρ_y, ρ_z. Therefore, each ADMM block contains three types of sub-layers: auxiliary variable layers to update $\boldsymbol{x}$, $\boldsymbol{y}$, $\boldsymbol{z}$, a reconstruction layer to update $\boldsymbol{s}$, and dual update layers for $\boldsymbol{u}_x$, $\boldsymbol{u}_y$, $\boldsymbol{u}_z$, respectively.

B. Accumulated Mismatch Error

However, these updates are derived under ideal circumstances towards exact modeling of the imaging process, where model mismatch is not taken into consideration. In practice, only the estimated PSF operator $\tilde{\mathbf{H}}$ is available for the optimization problem. To compute the deviation on the ADMM updates in a single iteration resulting from the model error, we suppose biased PSF is only used at iteration k and the updates for the previous step are unbiased. The least-square updates $\tilde{\boldsymbol{x}}^{(k)}$, $\tilde{\boldsymbol{s}}^{(k)}$ based on the biased model matrix $\tilde{\mathbf{H}}$ at iteration k can be written as

$$\tilde{\boldsymbol{x}}^{(k)} = (\boldsymbol{\Phi}^\top\boldsymbol{\Phi} + \rho_x\mathbf{I})^{-1}(\boldsymbol{\Phi}^\top\boldsymbol{b} + \boldsymbol{u}_x^{(k-1)} + \rho_x\tilde{\mathbf{H}}\boldsymbol{s}^{(k-1)}) \quad (11)$$

$$\tilde{\boldsymbol{s}}^{(k)} = (\rho_x\tilde{\mathbf{H}}^\top\tilde{\mathbf{H}} + \rho_z\boldsymbol{\Psi}^\top\boldsymbol{\Psi} + \rho_y\mathbf{I})^{-1}[\tilde{\mathbf{H}}^\top(\rho_x\tilde{\boldsymbol{x}}^{(k)} - \boldsymbol{u}_x^{(k-1)}) + \boldsymbol{r}^{(k)}]. \quad (12)$$

Therefore, corrections should be made to obtain the unbiased updates, given by

$$\begin{aligned}\boldsymbol{x}^{(k)} &= (\boldsymbol{\Phi}^\top\boldsymbol{\Phi} + \rho_x\mathbf{I})^{-1}(\boldsymbol{\Phi}^\top\boldsymbol{b} + \boldsymbol{u}_x^{(k-1)} + \rho_x(\tilde{\mathbf{H}} + \Delta\mathbf{H})\boldsymbol{s}^{(k-1)}) \\ &= (\boldsymbol{\Phi}^\top\boldsymbol{\Phi} + \rho_x\mathbf{I})^{-1}(\boldsymbol{\Phi}^\top\boldsymbol{b} + \boldsymbol{u}_x^{(k-1)} + \rho_x\tilde{\mathbf{H}}\boldsymbol{s}^{(k-1)} \\ &\quad + \boxed{\rho_x\Delta\mathbf{H}\boldsymbol{s}^{(k-1)}}) \quad (13)\end{aligned}$$

$$\begin{aligned}\boldsymbol{s}^{(k)} &= (\rho_x(\tilde{\mathbf{H}} + \Delta\mathbf{H})^\top(\tilde{\mathbf{H}} + \Delta\mathbf{H}) + \rho_z\boldsymbol{\Psi}^\top\boldsymbol{\Psi} + \rho_y\mathbf{I})^{-1} \\ &\quad [(\tilde{\mathbf{H}} + \Delta\mathbf{H})^\top(\rho_x\boldsymbol{x}^{(k)} - \boldsymbol{u}_x^{(k-1)}) + \boldsymbol{r}^{(k)}] \\ &= (\mathbf{W}_1 + \boxed{\rho_x\delta_{\mathbf{H}}})^{-1}[\tilde{\mathbf{H}}^\top(\rho_x\boldsymbol{x}^{(k)} - \boldsymbol{u}_x^{(k-1)}) + \boldsymbol{r}^{(k)}] \\ &\quad + \boxed{(\mathbf{W}_1 + \rho_x\delta_{\mathbf{H}})^{-1}\Delta\mathbf{H}^\top(\rho_x\boldsymbol{x}^{(k)} - \boldsymbol{u}_x^{(k-1)})}, \quad (14)\end{aligned}$$

where $\mathbf{W}_1 = \rho_x\tilde{\mathbf{H}}^\top\tilde{\mathbf{H}} + \rho_z\boldsymbol{\Psi}^\top\boldsymbol{\Psi} + \rho_y\mathbf{I}$ and $\delta_{\mathbf{H}} = \tilde{\mathbf{H}}^\top\Delta\mathbf{H} + \Delta\mathbf{H}^\top\tilde{\mathbf{H}} + \Delta\mathbf{H}^\top\Delta\mathbf{H}$. The errors from the biased PSF operator are highlighted in red and would accumulate for each iteration. By substituting Eq. (13) for $x^{(k)}$ in Eq. (14), the reconstruction estimate can be written as a function of the biased estimate $\tilde{\boldsymbol{s}}^{(k)}$, original measurements $\boldsymbol{b}$ and updates from the previous step. Thus, we have

$$\boldsymbol{s}^{(k)} = (\mathbf{W}_1 + \rho_x\delta_{\mathbf{H}})^{-1}\mathbf{W}_1\tilde{\boldsymbol{s}}^{(k)} + \mathbf{W}_2\boldsymbol{\Phi}^\top\boldsymbol{b} + \boldsymbol{\epsilon}^{(k-1)}, \quad (15)$$

where $\mathbf{W}_2 = (\mathbf{W}_1 + \rho_x\delta_{\mathbf{H}})^{-1}\Delta\mathbf{H}^\top\rho_x(\boldsymbol{\Phi}^\top\boldsymbol{\Phi} + \rho_x\mathbf{I})^{-1}$ and $\boldsymbol{\epsilon}^{(k-1)}$ denotes the combination of terms involving updates from the $(k-1)^{th}$ iteration. Therefore the unbiased intensity updates in all steps can be intuitively expressed as

$$\begin{aligned}\boldsymbol{s}^{(1)} &= \phi(\boldsymbol{s}^{(0)}, \tilde{\boldsymbol{s}}^{(1)}, \boldsymbol{b}) = g_1(\tilde{\boldsymbol{s}}^{(1)}, \boldsymbol{b}) \\ \boldsymbol{s}^{(2)} &= \phi(\boldsymbol{s}^{(1)}, \tilde{\boldsymbol{s}}^{(2)}, \boldsymbol{b}) = g_2(\tilde{\boldsymbol{s}}^{(1)}, \tilde{\boldsymbol{s}}^{(2)}, \boldsymbol{b}) \\ \boldsymbol{s}^{(3)} &= \phi(\boldsymbol{s}^{(2)}, \tilde{\boldsymbol{s}}^{(3)}, \boldsymbol{b}) = g_3(\tilde{\boldsymbol{s}}^{(1)}, \tilde{\boldsymbol{s}}^{(2)}, \tilde{\boldsymbol{s}}^{(3)}, \boldsymbol{b}) \\ &\vdots \\ \boldsymbol{s}^{(k)} &= \phi(\boldsymbol{s}^{(k-1)}, \tilde{\boldsymbol{s}}^{(k)}, \boldsymbol{b}) = g_k([\tilde{\boldsymbol{s}}^{(1)}, \tilde{\boldsymbol{s}}^{(2)}, \ldots, \tilde{\boldsymbol{s}}^{(k)}], \boldsymbol{b}), \quad (16)\end{aligned}$$

where $\phi(\cdot)$ and $g_k(\cdot)$ denote the corresponding mapping functions to obtain $\boldsymbol{s}^{(k)}$.

As described in earlier sections, in state-of-the-art networks that employ forward imaging models, lensless measurements are reconstructed through a model-based learning algorithm followed by a data-driven enhancement network sequentially in separate phases. They rely on the cascaded CNN to address all artifacts in the preliminary reconstruction while failing to consider the information loss in the prior session. As seen in Eq. (15), a significant bias on $\tilde{\boldsymbol{s}}^{(k)}$ can be observed in each step of ADMM and would accumulate for all iterations. Moreover, the image sparsity assumption as the prior on $\tilde{\boldsymbol{s}}^{(k)}$ smooths out valuable details. Similarly, the nonlinear function in Eq. (8) enforcing non-negativity may lead to irreversible degradation as well. All these cause information loss in the physics-based reconstruction stage. Taking the output of this stage as a noisy image and stacking a typical CNN denoiser such as Unet at the end does help with suppressing artifacts to some extent and improves the perceptual quality of the reconstructed images. However, it is difficult to correct the model mismatch errors this way and further enhancement would also be problematic with incomplete information provided.

C. Compensation Branch

As illustrated in Fig. 3, we propose to learn a compensation branch parallel to the cascaded enhancement network, with multiple convolution streams making use of observation $\boldsymbol{b}$ and intermediate intensity update in each ADMM block, to estimate the correction in the presence of model mismatch. With the number of unrolled ADMM blocks set to be K, a sequence of reconstructions are generated $\{\tilde{\boldsymbol{s}}^{(k)}\}_{k=1}^{K}$. The cascaded Unet can be described as the composition of an encoder $f(\cdot)$ for feature extraction and a decoder $d(\cdot)$ for reconstruction. Due to information loss in the unrolled ADMM blocks, the feature vector $\tilde{\boldsymbol{h}}$ extracted from $\tilde{\boldsymbol{s}}^{(k)}$ deviates from the feature learnt under the unbiased situation. Therefore, we use a data-driven approach to compensate such a difference, and the CNN-based correction process can be expressed as

$$\mathcal{P}(\cdot): \begin{pmatrix} \tilde{\boldsymbol{s}}^{(K)} \\ \boldsymbol{b}, [\tilde{\boldsymbol{s}}^{(1)}, \tilde{\boldsymbol{s}}^{(2)}, \ldots, \tilde{\boldsymbol{s}}^{(K-1)}] \end{pmatrix} \longrightarrow \hat{\boldsymbol{h}} \longrightarrow \hat{\boldsymbol{s}}, \quad (17)$$

where $\mathcal{P}$ implements the Unet with two contracting branches. The additional contracting path is the compensation branch,

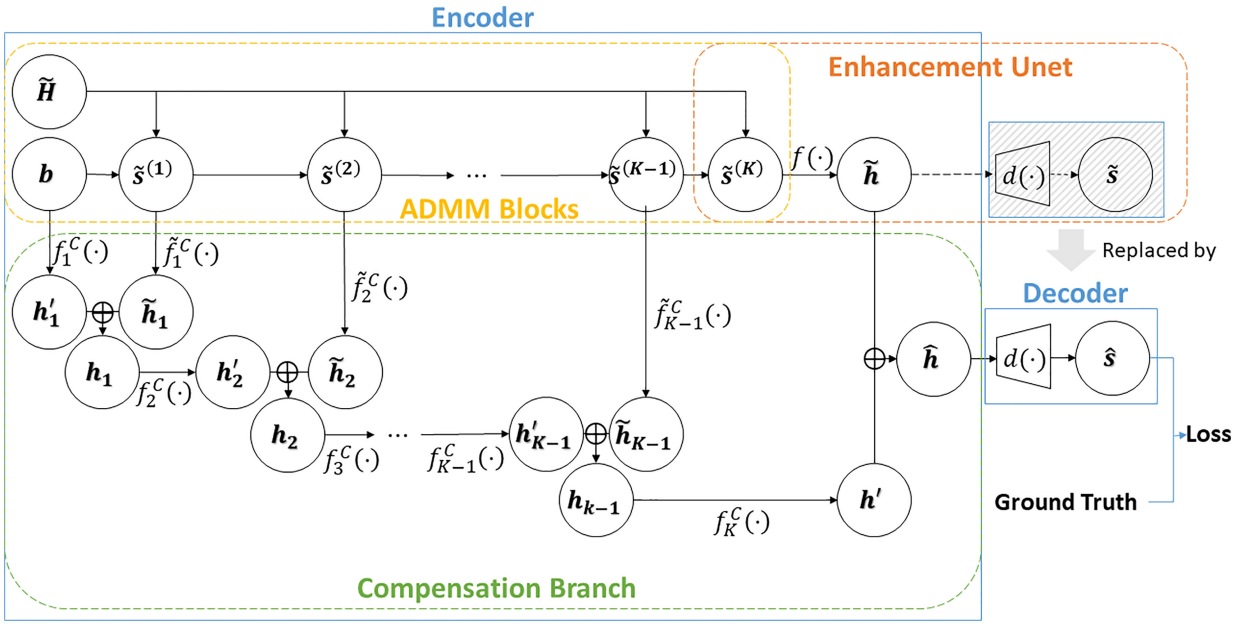


Fig. 3. Schematic diagram of the proposed MMCN for reconstruction from lensless measurements in the presence of model mismatch. The intermediate intensity updates $\{\tilde{\boldsymbol{s}}^{(k)}\}_{k=1}^{K-1}$ from the unrolled ADMM blocks based on biased operator $\tilde{\mathbf{H}}$, together with lensless measurement $\boldsymbol{b}$, are incorporated in the compensation branch (C-branch) to correct errors resulting from model mismatch. The feature representation $\boldsymbol{h}_k$ at stage k is the concatenation of feature vectors extracted from $\tilde{\boldsymbol{s}}^{(k)}$ and from the last stage $\boldsymbol{h}_{k-1}$ towards contracting convolutional layers $\tilde{f}_k^C(\cdot)$ and $f_k^C(\cdot)$. The final representation $\boldsymbol{h}'$ is stacked with $\tilde{\boldsymbol{h}}$, which is extracted by the encoder $f(\cdot)$ from the Unet denoiser, and then fed into a decoder for reconstruction. The shaded parts are decoder and reconstruction $\tilde{\boldsymbol{s}}$ of the enhancement Unet without correction on the feature representations, which are replaced by the expansive path $d(\cdot)$ and reconstruction $\hat{\boldsymbol{s}}$ in the blue solid box. The parameters are updated via backpropogation with respect to the loss between the estimated $\hat{\boldsymbol{s}}$ and the ground truth.

which learns the correction for model mismatch with the combination of feature maps from multiple inputs $(\boldsymbol{b}, [\tilde{\boldsymbol{s}}^{(1)}, \tilde{\boldsymbol{s}}^{(2)}, \ldots, \tilde{\boldsymbol{s}}^{(K-1)}])$. Let $\boldsymbol{h}'_k$, $\tilde{\boldsymbol{h}}_k$ be a paired feature representation extracted from the feature vector $\boldsymbol{h}_{k-1}$ at step $k-1$ and intermediate update $\tilde{\boldsymbol{s}}^{(k)}$ at the k^{th} ADMM block such that $\boldsymbol{h}'_k = f_k^C(\boldsymbol{h}_{k-1})$, $\tilde{\boldsymbol{h}}_k = \tilde{f}_k^C(\tilde{\boldsymbol{s}}_k)$. We have

$$\boldsymbol{h}_1 = f_1^C(\boldsymbol{b}) \oplus \tilde{f}_1^C(\tilde{\boldsymbol{s}}_1) \tag{18}$$

$$\boldsymbol{h}_2 = f_2^C(\boldsymbol{h}_1) \oplus \tilde{f}_2^C(\tilde{\boldsymbol{s}}_2) \tag{19}$$

$$\vdots$$

$$\boldsymbol{h}_{K-1} = \boldsymbol{h}'_{K-1} \oplus \tilde{\boldsymbol{h}}_{K-1} = f_{K-1}^C(\boldsymbol{h}_{K-1}) \oplus \tilde{f}_{K-1}^C(\tilde{\boldsymbol{s}}_{K-1}) \tag{20}$$

$$\hat{\boldsymbol{h}} = \boldsymbol{h}' \oplus \tilde{\boldsymbol{h}} = f_K^C(\boldsymbol{h}_{K-1}) \oplus f(\tilde{\boldsymbol{s}}_K), \tag{21}$$

where $\oplus$ denotes the concatenation operator. The feature extractors $f_k^C(\cdot)$ and $\tilde{f}_k^C$ are sets of downsampling CNN filters to generate feature map $\boldsymbol{h_k}$.

An example of detailed configurations under the proposed structure is illustrated in Fig. 4. A total of five ADMM blocks ($K = 5$) are utilized, representing five iterations of ADMM and each block consists of sub-layers corresponding sub-problems in each ADMM iteration step to update measurement intensity, auxiliary variables and dual multipliers. The Unet stacked at the end of architecture is of standard form with five repeated applications of two 3×3 convolutions followed by a 2×2 max-pooling operation as the encoder, and symmetric structure with 2×2 up-convolutions for the decoder. Regarding the compensation branch for correction, we utilize double convolutional layers followed by a max-pooling layer as feature extractor $f_k^C(\cdot)$ and a set of residual blocks as $\tilde{f}_k^C(\cdot)$, each consisting of a skip connection, two convolutional layers and a max-pooling. After each $\tilde{f}_k^C(\cdot)$, a concatenation with the corresponding representation extracted by $\tilde{f}_k^C(\cdot)$ is then applied. The output of compensation branch is then stacked to the feature map of Unet encoder and together fed into a bottleneck convolutional layer before the upscaling decoder. The channel number of feature representation is increased from 24 to 512 in the encoder ($24 \rightarrow 64 \rightarrow 128 \rightarrow 256 \rightarrow 512$) and decreased to 24 in the decoder. Notice that each convolutional layer, unless otherwise stated, is followed by a batch normalization (BN) layer and a rectified linear unit (ReLU) activation function. The loss function adopted here is a weighted combination of mean square error (MSE) and perceptual loss, quantifying the distortion and perceptual quality of reconstructed images with respect to the ground truths. The perceptual loss used here to generate more photo-realistic images is obtained by Learned Perceptual Image Patch Similarity metric (LPIPS) [69] that measures the perceptual distance between ground truth and estimated images.

V. Experiments and Results

We make use of datasets captured by two mask-based lensless imaging systems using different phase masks, namely, DiffuserCam [12] with an off-the-shelf diffuser, and PhlatCam [3] with a designed phase mask. Both datasets are collected by capturing

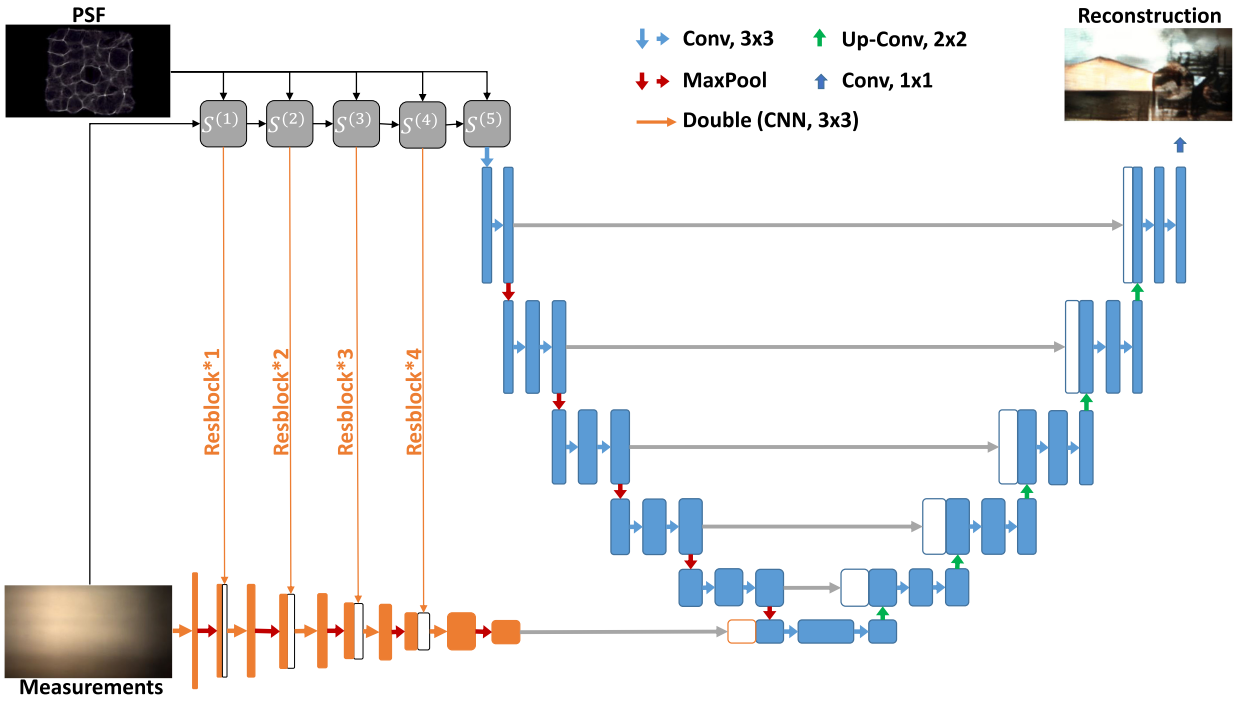


Fig. 4. An example of detailed configurations under the proposed structure. A total of five ADMM blocks are utilized and the Unet stacked at the end of architecture is of standard form. The input measurement and intermediate updates from ADMM blocks are fed into the data-driven compensation branch via double CNNs and residual blocks.

projections of images on the monitor. We have also evaluated the proposed MMCN on real-world datasets, i.e, measurements of unconstrained indoor scenes using PhlatCam. Experiments for all datasets are implemented using Pytorch framework on NVIDIA V100 GPUs (32 GB, SMX2). Codes and trained models will be made available[1].

A. *For DiffuserCam*

The DiffuserCam Lensless Mirflickr Dataset (DLMD) [28] consists of 25,000 aligned image pairs (24,000 for training and 1,000 for testing) taken by a lensed ground truth camera and a mask-based lensless camera simultaneously via a beamsplitter. For both cameras, Basler Dart (daA1920-30uc) sensors are used to capture data with 1080×1920 pixels (restricted by the sensor), which is then downsampled to 270×480 to reduce moire fringes and cropped to 210×380 to remove areas beyond the screen borders for final display. The calibrated PSF is captured by placing an LED point source at the center of the monitor. Due to the random diffuser mask, it is not possible to obtain a rough estimate of the mask, and thus simulated PSF is not available for DLMD. The network is trained with a batch size of 4 for 100 epochs using Adam optimizer [70]. The initial learning rate is 10^{-4} and is then decreased linearly to 10^{-6}. After pretraining with only MSE loss, the weights for MSE and LPIPS loss are then set to be 1 and 1.2, respectively.

To illustrate the performance gain brought by different components of the structure, experiments are conducted to compare the proposed MMCN with purely deep network Unet [31], learnable unrolled ADMM (Le-ADMM) [28], and its derived network Le-ADMM-U [28], namely Le-ADMM followed by a CNN denoiser (Unet) for further enhancement. Besides, an additional ablation study is performed focusing on the performance gain brought by the introduction of information from raw measurements and intermediate reconstructions of ADMM blocks. Specifically, we train the network with the same settings as we proposed, while only feeding the final update of ADMM blocks to the compensation branch. This downgraded MMCN model for ablation study is denoted as MMCN-DG. In this way, MMCN-DG shares the same number of parameters with the MMCN model, while replacing the intermediate reconstructions and raw measurements incorporated in MMCN with five replicas of the output from the last ADMM block $S^{(5)}$. The objective assessments used to evaluate the reconstructed performance are peak signal-to-noise ratio (PSNR), structural similarity index (SSIM), and perceptual metric LPIPS based on AlexNet. As presented in Table I, the proposed C-branch improves the image quality on all the evaluation metrics at the expense of a longer run time. It can be observed that the downgraded model MMCN-DG, by adding an additional branch itself, has made improvements in PSNR and produced comparable results in SSIM and LIPIS values compared with Le-ADMM-U. Nevertheless, it gives noticeably inferior performance in all metrics compared with our proposed MMCN leveraging both raw measurement and intermediate updates. Specifically, the proposed model provides a noticeable performance gain, more than 3 dB in PSNR and 0.02 in SSIM, as compared with Le-ADMM-U. Comparison has also been made on numerical performance among various

[1][Online]. Available: Link: https://github.com/tianjiaozeng/MMCN

TABLE I
ABLATION STUDY OF THE PROPOSED RECONSTRUCTION NETWORK ON THE DIFFUSERCAM LENSLESS MIRFLICKR DATASET. MMCN-DG STANDS FOR THE DOWNGRADED MMCN MODEL OF THE SAME STRUCTURE WITHOUT RAW MEASUREMENTS AND INTERMEDIATE RECONSTRUCTIONS

Method	Component				Metric			Inference Time (sec)
	Physics prior	CNN denoiser	C-branch	Raw measurements + Intermediate reconstructions	PSNR (dB)	SSIM	LPIPS	
Unet		✓			19.22	0.76	0.2461	0.010
Le-ADMM	✓				13.73	0.59	0.4434	0.071
Le-ADMM-U	✓	✓			21.92	0.84	0.1954	0.075
MMCN-DG	✓	✓	✓		23.64	0.84	0.1951	0.091
MMCN	✓	✓	✓	✓	25.69	0.86	0.1897	0.091

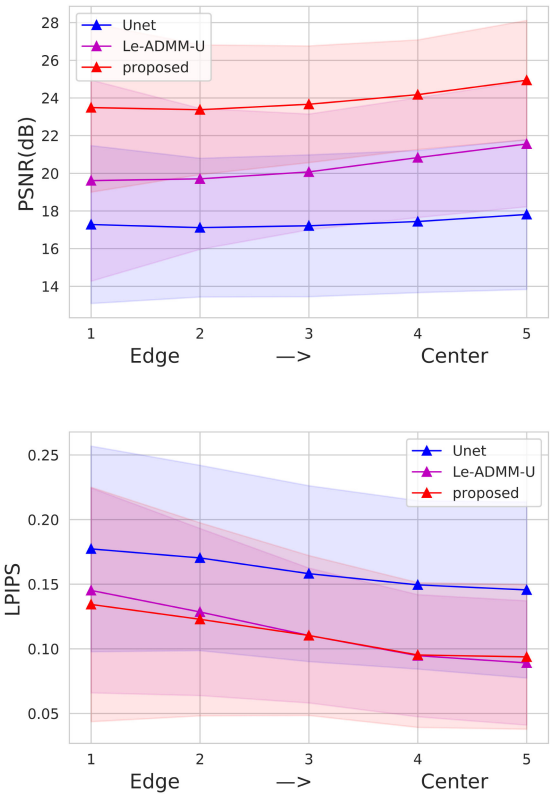


Fig. 5. Comparison on the reconstruction quality between areas among the edges and the center of the FoV. The PSNR and LPIPS values of five patches obtained on locations from the edge to the central areas are plotted. We see that the performance gain of the proposed model compared with Le-ADMM-U turns to be larger on patches near the edge of the image. The difference becomes smaller as moving towards central areas. Unet presents comparably most stable performance across different areas.

locations within FOV for Unet, Le-ADMM-U and proposed method (Fig. 5). The reconstructed images from edge to center are equally cut into five slices, referred to as image patches 1 to 5. As the index increases from 1 to 5, the corresponding location gets nearer to the central areas, where patch 5 is located in the center and patch 1 is closest to the edge. Fig. 5 shows the PSNR and LPIPS values for each of the five regions. Mean and standard deviations across samples in the testing set are plotted. All three models have present more or less superior performance in areas near the center. Specifically, the steepest slope is observed in the curve of Le-ADMM-U in both metrics. The PSNR gain of the proposed model compared with Le-ADMM-U gets larger near the edge, indicating more robustness to the mismatch between on-axis and off-axis PSF. The LPIPS loss of our network also increases slower than Le-ADMM-U as moving towards the edge. Among all compared models, Unet exhibits relatively smallest differences in both metrics across regions as expected, since it is purely data-driven and is free from the influence of model mismatch error. Despite the most robust performance, the overall reconstruction quality of Unet is inferior to the other methods.

The visual inspection of the images recovered by the proposed method, the ablation model MMCN-DG, Le-ADMM-U, Unet, Le-ADMM and ground truth images are shown in Fig. 6. The reconstruction from our method presents the most perceptually appealing image quality compared with the other two methods. It is worth noticing that the proposed network provides more appropriate reconstruction both in the colors and texture details. Compared with Le-ADMM-U, it can be clearly seen that the distortion is less and more details are preserved with the compensation branch added to the network, especially around the edges of the FoV. As depicted in Fig. 6, patches around marginal areas are zoomed out for better comparison, where the proposed architecture shows obviously better reconstruction quality. The ablation model MMCN-DG has presented generally superior performance in most of the sample reconstruction images compared with Le-ADMM-U and has improved on some of the marginal areas. However, noticeable enhancement can be observed in the proposed model when leveraging with raw measurement and intermediate reconstructions. It indicates that the compensation branch effectively improves the off-axis resolution, which is severely affected by the approximation of circulant convolution. As for reconstructions of Unet, some severe artifacts can be seen visually, indicating the difficulty for the data-driven network alone to learn appropriate inverse process without the help of system priors.

B. For PhlatCam

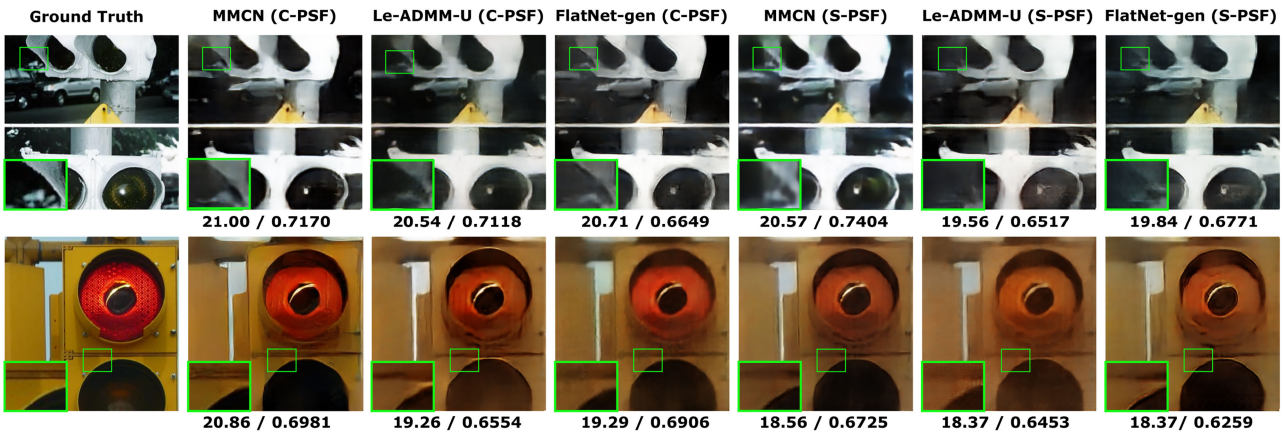
For the PhlatCam dataset [32], a total of 10,000 images randomly selected from 1,000 classes (10 images from each class) of ImageNet ILSVRC 2012 [71] are resized to 384×384 and displayed on monitor for imaging. Basler Ace4024-29uc with 12.2MP Sony IMX226 sensor is used to capture lensless measurements of size 1280×1480. For the conditions that the sensor is not big enough for full measurements, the images are cropped to 608×864. The dataset is then divided into 9,000 training images and 1,000 testing images. Besides calibrated PSF, an uncalibrated PSF is simulated based on the mask pattern and the camera geometry. Different from DLMD, there is no

Fig. 6. Visual inspection on the reconstruction performance for data from DLMD using MMCN, MMCN-DG (no raw measurement and intermediate reconstructions from ADMM blocks), Le-ADMM-U (no compensation branch), Unet (pure data-driven network without physics prior) and unrolled ADMM network (Top to bottom: raw measurements, ground truth images, and reconstructed results of MMCN, MMCN-DG, Le-ADMM-U, Unet and Le-ADMM). The PSNR and SSIM values for each reconstruction are presented. The visual reconstruction quality of the proposed model with compensation branch leveraging both raw measurement and intermediate reconstructions from ADMM blocks, noticeably outperforms other methods including the downgraded version MMCN-DG which only uses output from the final ADMM block. Compared with Le-ADMM-U (no compensation branch), our model presents superior performance, especially in restoring correct colors and details around marginal areas.

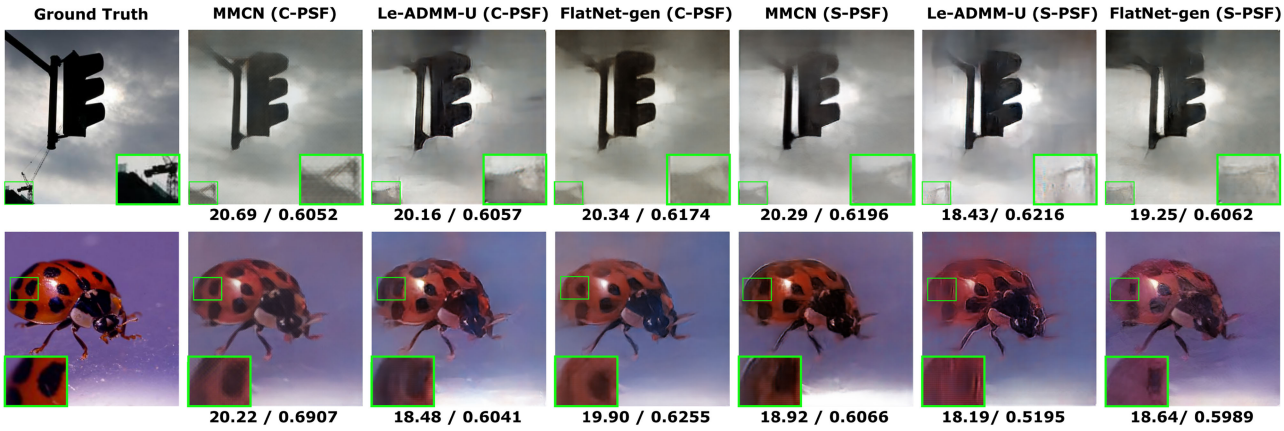
lensed camera during imaging and thus the displayed ImageNet images are directly taken as the ground truth images. Since the lensed ground truths aligned with lensless measurements are not available, the adversarial loss is added to enhance the reconstruction quality [72]. The discriminator used for adversarial learning consists of four convolutional layers with batch normalization and the swish activation function [73]. After pretraining with only MSE, LIPIS and adversarial losses are then added with weights of 1.2 and 0.6, respectively.

We evaluate the performance on PhlatCam measurements focusing on the comparison of model robustness between the reconstruction results with calibrated PSF (C-PSF) and simulated PSF (S-PSF) used in the network. The experiments are divided into two conditions, full-size measurements and cropped measurements, corresponding respectively to the ideal circumstances where the sensor is large enough to record the complete data of the imaged scene and more common circumstances where captured measurements are cropped. The proposed method with calibrated and simulated PSFs is compared with learning-based methods, Le-ADMM-U and FlatNet-gen [32] via metrics PSNR, SSIM and LPIPS based AlexNet. FlatNet-gen replaces the unrolled ADMM in the physics stage with single-step trainable inversion and then followed by a CNN-based enhancement network. All compared methods are trained using the same Unet as presented in Fig. 4 in the enhancement stage for fairness.

The numerical results on reconstructions of full-size measurements are shown in Table II. The proposed method MMCN with calibrated PSF, as well as simulated PSF, outperforms all other models. Notice that the smaller drop in numerical



(a) Full measurements



(b) Cropped measurements

Fig. 7. Visual inspection on the reconstruction performance for Phlatcam data using our proposed MMCN, FlatNet-gen, and Le-ADMM-U. The PSNR and SSIM values for each reconstruction are presented. Both calibrated PSF and simulated PSF are used for comparison. (a) Ground truth images and reconstructions for full-size measurements. (b) Ground truth images and reconstructions for cropped measurements.

TABLE II
COMPARISON OF RECONSTRUCTED PERFORMANCE ON PHLATCAM MEASUREMENTS OF FULL SIZE. THE CALIBRATED AND SIMULATED PSF USED IN THE RECONSTRUCTION ARE DENOTED AS C-PSF AND S-PSF

Method	PSNR (dB)	SSIM	LPIPS	Time (sec)
FlatNet-gen (S-PSF)	19.53	0.50	0.378	0.033
FlatNet-gen (C-PSF)	19.98	0.51	0.355	0.033
Le-ADMM-U (S-PSF)	19.05	0.47	0.413	0.083
Le-ADMM-U (C-PSF)	19.95	0.51	0.352	0.083
MMCN (S-PSF)	19.66	0.50	0.363	0.097
MMCN (C-PSF)	20.39	0.52	0.345	0.097

TABLE III
COMPARISON OF RECONSTRUCTED PERFORMANCE ON CROPPED PHLATCAM MEASUREMENTS

Method	PSNR (dB)	SSIM	LPIPS
FlatNet-gen (S-PSF)	18.24	0.47	0.408
FlatNet-gen (C-PSF)	18.79	0.49	0.375
Le-ADMM-U (S-PSF)	17.94	0.44	0.429
Le-ADMM-U (C-PSF)	18.72	0.48	0.381
MMCN (S-PSF)	18.46	0.47	0.387
MMCN (C-PSF)	19.32	0.50	0.371

results between calibrated and simulated PSF can be seen in the proposed method compared with Le-ADMM-U, indicating that the additional compensation branch improves the robustness in the presence of bias in PSF. While FlatNet-gen shows the smallest difference between using calibrated and simulated PSF, the overall performance is inferior to our models. It might due to the weaker dependence on PSF which makes the performance robust to different PSFs but results in inadequate reconstructions. Fig. 7(a) shows reconstructions from five sample measurements given calibrated and simulated PSF. For both cases, the proposed architecture produces better image quality than the other two methods overall. It can be seen that the reconstructed images of our model are visually more similar to the ground truth ones, especially in preserving colors and details.

The numerical results for reconstructions from cropped lensless measurements are displayed in Table III. Due to the information loss in measurements, the direct inversion in the physics stage of FlatNet-gen is infeasible as introduced earlier. Thus a preprocessing step is applied to the raw measurements before

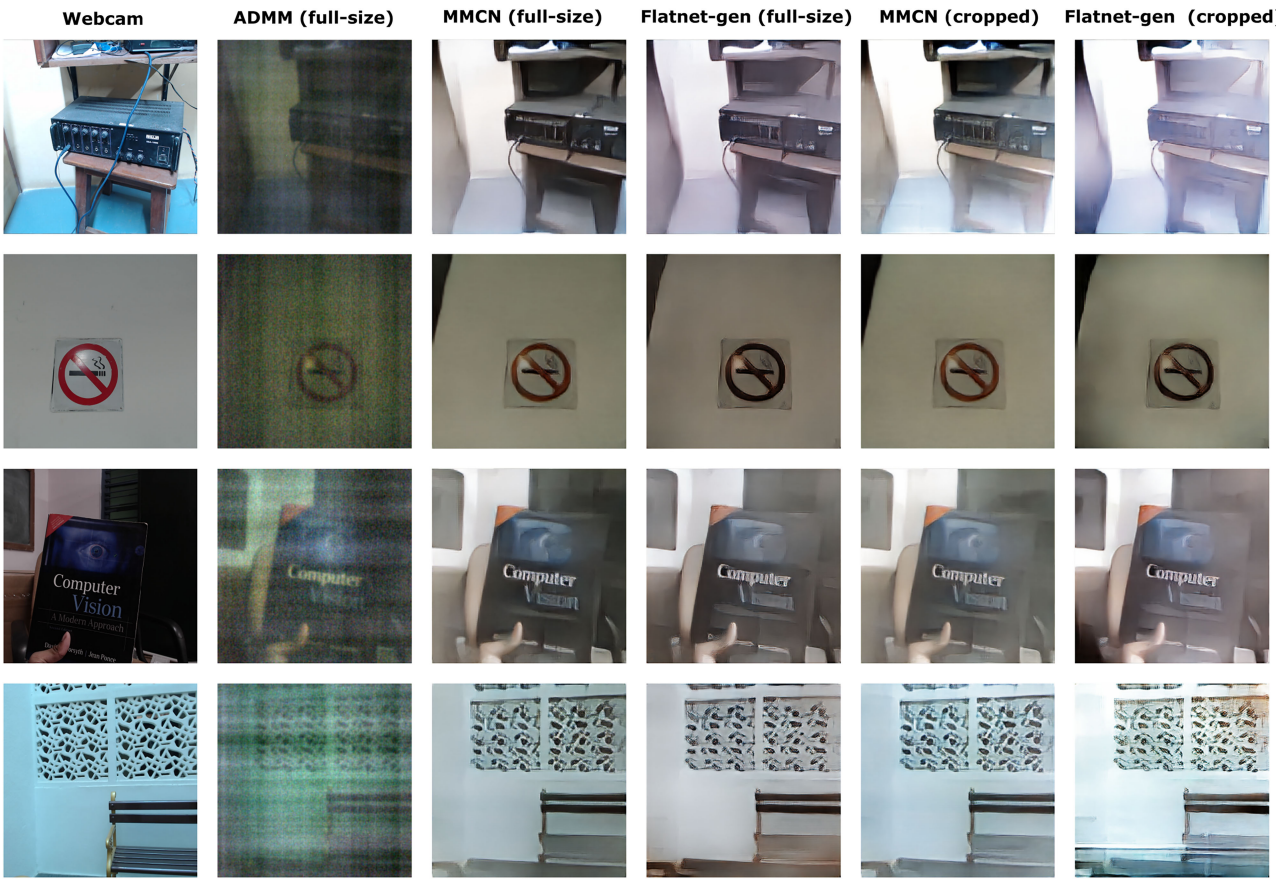


Fig. 8. Visual inspection on the reconstruction performance for real-world measurements captured by PhlatCam using the proposed architecture and FlatNet-gen for full-size and also cropped measurements.

feeding them into the networks. The images are first padded to full size through replicate padding and then multiplied with a Gaussian filter mask for smoothing [74]. In spite of an overall drop in evaluation metrics as compared with reconstructions of full-size images, our method achieves a comparably less drop in PSNR than the other state-of-the-art methods. The reconstruction quality can also be visually inspected in Fig. 7(b). As opposed to full-size measurements, the degradation of image quality around the edges can be observed in the reconstructions. The proposed architecture shows noticeable improvement in reconstruction quality around these areas and produces more similar images to the ground truth ones.

Besides datasets captured by displaying images on the monitor, real-world measurements recording indoor scenes via PhlatCam have also been utilized for testing the robustness of the proposed method. Due to the unconstrained acquisition process with significantly more severe noise compared with displayed measurements, the reconstruction is much more challenging. This dataset consists of 500 image pairs captured by PhlatCam and a webcam, including 475 training pairs and 25 testing pairs. Despite the availability of ground truth captures using webcam, there exists non-neglectable deviation between imaging angles of two cameras. Thus, the models pretrained on the display datasets are fine-tuned on real-world data using contextual loss [75], which is specially designed for unaligned image pairs. The images reconstructed by the proposed model MMCN and Flatnet-gen for both full-size and cropped measurements are presented in Fig. 8, as well as captures of webcam and ADMM reconstructions. As can be seen in reconstructed sample images, the proposed method presents significant improvements on reconstructions of the traditional ADMM approach. It also shows comparable or better results for full-size and cropped measurements compared with FlatNet-gen. It can be observed that some reconstructions of Flatnet-gen contain obvious color distortion and blurring artifacts, while MMCN exhibits more robust performance with better preservation in details.

C. Limitations and Future Work

Despite the promising performance of our proposed method, there are still limitations that require further work. First, the performance of deep learning relies heavily on the quality of the training dataset, which includes both raw measurements and ground truth images. Once the acquisition is done, the capability of error correction is bounded by the training image pairs fed into the network. The joint optimization of both experimental setups and learning-based reconstruction algorithms would be a promising direction to improve performance. We can make the acquisition process a trainable model where the imaging system with a lensless mask is designed by joint training with the

numerical reconstruction network. Moreover, the robustness of the proposed method has only been demonstrated in addressing erroneous PSFs and inaccurate mathematical model due to large impinging angles. As mentioned in previous sections, there exist many other kinds of model mismatch errors in lensless imaging, stemming from imperfect sensors, wide depth range of 3D objects and scenarios with occlusions or reflective surfaces. Extending the proposed method to these cases is of great interest. It may also be more desirable to investigate networks with customized designs corresponding to different target situations.

VI. Conclusion

In this work, we introduce a model-based reconstruction network for lensless imaging that is robust in the presence of model mismatch error. Despite the benefits brought by incorporating the physics prior, the existing lensless reconstruction approaches fail to consider the existence of model mismatch and thus result in artifacts and information loss. By leveraging all intermediate updates and raw measurements, a data-driven branch is developed to serve as a correction and compensation process on the model-based reconstruction. Extensive experiments have been conducted on real data collected by lensless imaging systems with two different phase masks. The results show that the proposed network provides superior performance in the reconstruction task for lensless imaging than other state-of-the-art approaches, and the introduced data-driven compensation branch effectively improves the robustness of reconstruction approach in dealing with model error.

References

[1] A. Ozcan and E. McLeod, "Lensless imaging and sensing," *Annu. Rev. Biomed. Eng.*, vol. 18, pp. 77–102, Jul. 2016.

[2] 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. Imag.*, vol. 3, no. 3, pp. 384–397, Jul. 2016.

[3] 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.*, vol. 42, no. 7, pp. 1618–1629, Apr. 2020.

[4] X. Xie *et al.*, "Extended depth-resolved imaging through a thin scattering medium with PSF manipulation," *Sci. Rep.*, vol. 8, no. 1, pp. 1–8, Mar. 2018.

[5] 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.*, vol. 42, no. 7, pp. 1606–1617, Apr. 2020.

[6] A. Sinha, J. Lee, S. Li, and G. Barbastathis, "Lensless computational imaging through deep learning," *Optica*, vol. 4, no. 9, pp. 1117–1125, Sep. 2017.

[7] O. Katz, P. Heidmann, M. Fink, and S. Gigan, "Non-invasive single-shot imaging through scattering layers and around corners via speckle correlations," *Nat. Photon.*, vol. 8, no. 10, pp. 784–790, Oct. 2014.

[8] E. Edrei and G. Scarcelli, "Memory-effect based deconvolution microscopy for super-resolution imaging through scattering media," *Sci. Rep.*, vol. 6, no. 1, pp. 1–8, Sep. 2016.

[9] V. Boominathan *et al.*, "Lensless imaging: A computational renaissance," *IEEE Signal Process. Mag.*, vol. 33, no. 5, pp. 23–35, Sep. 2016.

[10] X. Wang, X. Jin, J. Li, X. Lian, X. Ji, and Q. Dai, "Prior-information-free single-shot scattering imaging beyond the memory effect," *Opt. Lett.*, vol. 44, no. 6, pp. 1423–1426, Mar. 2019.

[11] C. Guo *et al.*, "Imaging through scattering layers exceeding memory effect range by exploiting prior information," *Opt. Commun.*, vol. 434, pp. 203–208, Mar. 2019.

[12] N. Antipa *et al.*, "Diffusercam: Lensless single-exposure 3D imaging," *Optica*, vol. 5, no. 1, pp. 1–9, Jan. 2018.

[13] Z. Zhang *et al.*, "Mask-modulated lensless imaging with multi-angle illuminations," *APL Photon.*, vol. 3, no. 6, Jun. 2018, Art. no. 060803.

[14] Z. Ren, Z. Xu, and E. Y. Lam, "Learning-based nonparametric autofocusing for digital holography," *Optica*, vol. 5, no. 4, pp. 337–344, Apr. 2018.

[15] K. H. Jin, M. T. McCann, E. Froustey, and M. Unser, "Deep convolutional neural network for inverse problems in imaging," *IEEE Trans. Image Process.*, vol. 26, no. 9, pp. 4509–4522, Jun. 2017.

[16] N. Meng, Z. Ge, T. Zeng, and E. Y. Lam, "Lightgan: A deep generative model for light field reconstruction," *IEEE Access*, vol. 8, pp. 116052–116063, Jun. 2020.

[17] G. Barbastathis, A. Ozcan, and G. Situ, "On the use of deep learning for computational imaging," *Optica*, vol. 6, no. 8, pp. 921–943, Aug. 2019.

[18] Z. Ren, Z. Xu, and E. Y. Lam, "End-to-end deep learning framework for digital holographic reconstruction," *Adv. Photon.*, vol. 1, no. 1, Jan. 2019, Art. no. 016004.

[19] J. Zhao *et al.*, "Deep-learning cell imaging through anderson localizing optical fiber," *Adv. Photon.*, vol. 1, no. 6, Nov. 2019, Art. no. 066001.

[20] T. Zeng, H. K.-H. So, and E. Y. Lam, "Redcap: Residual encoder-decoder capsule network for holographic image reconstruction," *Opt. Exp.*, vol. 28, no. 4, pp. 4876–4887, Feb. 2020.

[21] J. Li, D. Mengu, Y. Luo, Y. Rivenson, and A. Ozcan, "Class-specific differential detection in diffractive optical neural networks improves inference accuracy," *Adv. Photon.*, vol. 1, no. 4, Aug. 2019, Art. no. 046001.

[22] Y. Zhu, C. H. Yeung, and E. Y. Lam, "Microplastic pollution monitoring with holographic classification and deep learning," *J. Phy., Photon.*, vol. 3, no. 2, Mar. 2021, Art. no. 024013.

[23] H. K. Aggarwal, M. P. Mani, and M. Jacob, "Modl: Model-based deep learning architecture for inverse problems," *IEEE Trans. Med. Imag.*, vol. 38, no. 2, pp. 394–405, Aug. 2018.

[24] T. Zeng, H. K.-H. So, and E. Y. Lam, "Computational image speckle suppression using block matching and machine learning," *Appl. Opt.*, vol. 58, no. 7, pp. B39–B45, Mar. 2019.

[25] K. Zhang, W. Zuo, S. Gu, and L. Zhang, "Learning deep CNN denoiser prior for image restoration," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 3929–3938.

[26] W. Dong, P. Wang, W. Yin, G. Shi, F. Wu, and X. Lu, "Denoising prior driven deep neural network for image restoration," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 10, pp. 2305–2318, Oct. 2018.

[27] S. Boyd, N. Parikh, and E. Chu, *Distributed Optimization and Statistical Learning Via the Alternating Direction Method of Multipliers*. Boston, MA, USA: Now Publishers Inc, 2011.

[28] K. Monakhova, J. Yurtsever, G. Kuo, N. Antipa, K. Yanny, and L. Waller, "Learned reconstructions for practical mask-based lensless imaging," *Opt. Exp.*, vol. 27, no. 20, pp. 28075–28090, Sep. 2019.

[29] H. Zhou, H. Feng, Z. Hu, Z. Xu, Q. Li, and Y. Chen, "Lensless cameras using a mask based on almost perfect sequence through deep learning," *Opt. Exp.*, vol. 28, no. 20, pp. 30248–30262, Sep. 2020.

[30] Y. Yang, J. Sun, H. Li, and Z. Xu, "ADMM-CSNet: A deep learning approach for image compressive sensing," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 3, pp. 521–538, Nov. 2020.

[31] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.- Assist. Intervention*, 2015, pp. 234–241.

[32] 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.*, to be published, doi: 10.1109/TPAMI.2020.3033882.

[33] Q. Shan, J. Jia, and A. Agarwala, "High-quality motion deblurring from a single image," *ACM Trans. Graph.*, vol. 27, no. 3, pp. 1–10, Aug. 2008.

[34] S. Mallat, *A Wavelet Tour of Signal Processing: The Sparse Way*. New York, NY, USA: Academic Press, 2008.

[35] A. Busboom, H. Elders-Boll, and H. Schotten, "Uniformly redundant arrays," *Exp. Astron.*, vol. 8, no. 2, pp. 97–123, 1998.

[36] T. Cannon and E. Fenimore, "Coded aperture imaging: Many holes make light work," *Opt. Eng.*, vol. 19, no. 3, 1980, Art. no. 193283.

[37] E. Caroli, J. Stephen, G. Di Cocco, L. Natalucci, and A. Spizzichino, "Coded aperture imaging in X-and gamma-ray astronomy," *Space Sci. Rev.*, vol. 45, no. 3-4, pp. 349–403, 1987.

[38] Y. Zheng and M. S. Asif, "Joint image and depth estimation with mask-based lensless cameras," *IEEE Trans. Comput. Imag.*, vol. 6, pp. 1167–1178, 2020, doi: 10.1109/TCI.2020.3010360.

[39] J. K. Adams *et al.*, "Single-frame 3D fluorescence microscopy with ultraminiature lensless flatscope," *Sci. Adv.*, vol. 3, no. 12, Dec. 2017, Art. no. e1701548.

[40] G. Huang, H. Jiang, K. Matthews, and P. Wilford, "Lensless imaging by compressive sensing," in *Proc. IEEE Int. Conf. Image Process.*, 2013, pp. 2101–2105.

[41] D. G. Stork and P. R. Gill, "Lensless ultra-miniature CMOS computational imagers and sensors," in *Proc. Int. Conf. Sensor Technol. Appl.*, 2013, pp. 186–190.
[42] W. Chi and N. George, "Optical imaging with phase-coded aperture," *Opt. Exp.*, vol. 19, no. 5, pp. 4294–4300, 2011.
[43] G. Kuo, F. L. Liu, I. Grossrubatscher, R. Ng, and L. Waller, "On-chip fluorescence microscopy with a random microlens diffuser," *Opt. Exp.*, vol. 28, no. 6, pp. 8384–8399, Mar. 2020.
[44] J. Tan *et al.*, "Face detection and verification using lensless cameras," *IEEE Trans. Comput. Imag.*, vol. 5, no. 2, pp. 180–194, Dec. 2018.
[45] K. Monakhova, K. Yanny, N. Aggarwal, and L. Waller, "Spectral diffusercam: Lensless snapshot hyperspectral imaging with a spectral filter array," *Optica*, vol. 7, no. 10, pp. 1298–1307, 2020.
[46] K. Yanny *et al.*, "Miniscope3D: Optimized single-shot miniature 3D fluorescence microscopy," *Light, Sci. Appl.*, vol. 9, no. 1, pp. 1–13, 2020.
[47] K. Tajima, T. Shimano, Y. Nakamura, M. Sao, and T. Hoshizawa, "Lensless light-field imaging with multi-phased fresnel zone aperture," in *Proc. IEEE Int. Conf. Comput. Photography*, 2017, pp. 1–7.
[48] M. S. Asif, A. Ayremlou, A. Veeraraghavan, R. Baraniuk, and A. Sankaranarayanan, "Flatcam: Replacing lenses with masks and computation," in *Proc. IEEE Int. Conf. Comput. Vis. Workshop*, 2015, pp. 663–666.
[49] J. Tanida *et al.*, "Thin observation module by bound optics (TOMBO): Concept and experimental verification," *Appl. Opt.*, vol. 40, no. 11, pp. 1806–1813, 2001.
[50] P. R. Gill *et al.*, "Thermal escher sensors: Pixel-efficient lensless imagers based on tiled optics," in *Proc. Imag. Appl. Opt.*, Optical Society of America, 2017, pp. CTu3B.3.
[51] D. G. Stork and P. R. Gill, "Optical, mathematical, and computational foundations of lensless ultra-miniature diffractive imagers and sensors," *Int. J. Adv. Syst. Meas.*, vol. 7, no. 3, pp. 201–208, 2014.
[52] J. Wu, L. Cao, and G. Barbastathis, "DNN-FZA camera: A deep learning approach toward broadband fza lensless imaging," *Opt. Lett.*, vol. 46, no. 1, pp. 130–133, Jan. 2021.
[53] A. Dave, A. K. Vadathya, R. Subramanyam, R. Baburajan, and K. Mitra, "Solving inverse computational imaging problems using deep pixel-level prior," *IEEE Trans. Comput. Imag.*, vol. 5, no. 1, pp. 37–51, Mar. 2019.
[54] J. Rick Chang, C.-L. Li, B. Poczos, B. Vijaya Kumar, and A. C. Sankaranarayanan, "One network to solve them all-solving linear inverse problems using deep projection models," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 5888–5897.
[55] J. Zhang and B. Ghanem, "Ista-net: Interpretable optimization-inspired deep network for image compressive sensing," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 1828–1837.
[56] J. D. Rego, K. Kulkarni, and S. Jayasuriya, "Robust lensless image reconstruction via PSF estimation," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis.*, 2021, pp. 403–412.
[57] H. Zhuang, H. He, X. Xie, and J. Zhou, "High speed color imaging through scattering media with a large field of view," *Sci. Rep.*, vol. 6, no. 1, pp. 1–7, 2016.
[58] W. Li, J. Liu, S. He, L. Liu, and X. Shao, "Multitarget imaging through scattering media beyond the 3D optical memory effect," *Opt. Lett.*, vol. 45, no. 10, pp. 2692–2695, 2020.
[59] D. Ren, W. Zuo, D. Zhang, J. Xu, and L. Zhang, "Partial deconvolution with inaccurate blur kernel," *IEEE Trans. Image Process.*, vol. 27, no. 1, pp. 511–524, Oct. 2017.
[60] Y. Nan and H. Ji, "Deep learning for handling kernel/model uncertainty in image deconvolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 2388–2397.
[61] S. Vasu, V. R. Maligireddy, and A. Rajagopalan, "Non-blind deblurring: Handling kernel uncertainty with CNNS," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 3272–3281.
[62] H. Ji and K. Wang, "Robust image deblurring with an inaccurate blur kernel," *IEEE Trans. Image Process.*, vol. 21, no. 4, pp. 1624–1634, Nov. 2011.
[63] D. Ren, W. Zuo, D. Zhang, L. Zhang, and M.-H. Yang, "Simultaneous fidelity and regularization learning for image restoration," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 1, pp. 284–299, Jan. 2021.
[64] M. Bertero and P. Boccacci, *Introduction to Inverse Problems in Imaging*. Boca Raton, FL, USA: CRC Press, Aug. 2020.
[65] J. Nocedal and S. Wright, *Numerical Optimization*. New York, NY, USA: Springer, 2006.
[66] M. V. Afonso, J. M. Bioucas-Dias, and M. A. Figueiredo, "Fast image recovery using variable splitting and constrained optimization," *IEEE Trans. Image Process.*, vol. 19, no. 9, pp. 2345–2356, Apr. 2010.
[67] M. S. C. Almeida and M. Figueiredo, "Deconvolving images with unknown boundaries using the alternating direction method of multipliers," *IEEE Trans. Image Process.*, vol. 22, no. 8, pp. 3074–3086, Apr. 2013.
[68] Y. Wang, J. Yang, W. Yin, and Y. Zhang, "A new alternating minimization algorithm for total variation image reconstruction," *SIAM J. Imag. Sci.*, vol. 1, no. 3, pp. 248–272, 2008.
[69] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 586–595.
[70] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. 3rd Int. Conf. Learn. Representations (ICLR)*, 2014, pp. 1–41.
[71] O. Russakovsky *et al.*, "Imagenet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, Dec. 2015.
[72] I. Goodfellow *et al.*, "Generative adversarial nets," *Adv. Neural Info. Process. Syst.*, vol. 27, pp. 2672–2680, 2014.
[73] P. Ramachandran, B. Zoph, and Q. V. Le, "Searching for activation functions," 2017, *arXiv:1710.05941*.
[74] S. J. Reeves, "Fast image restoration without boundary artifacts," *IEEE Trans. Image Process.*, vol. 14, no. 10, pp. 1448–1453, Sep. 2005.
[75] R. Mechrez, I. Talmi, and L. Zelnik-Manor, "The contextual loss for image transformation with non-aligned data," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 768–783.

Tianjiao Zeng received the B.S. degree in electrical engineering from the University of Electronic Science and Technology of China, and the M.S. degree in electrical and computer engineering from Rutgers University. She is currently working toward the Ph.D. degree with the Department of Electrical and Electronic Engineering, The University of Hong Kong. Her research interests include pattern recognition, computational imaging, and machine learning.

Edmund Y. Lam (Fellow, IEEE) received the B.S., M.S., and Ph.D. degrees in electrical engineering from Stanford University. He was a Visiting Associate Professor with the Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, and is currently a Professor of Electrical and Electronic Engineering with The University of Hong Kong. He also is the Computer Engineering Program Director, and a Research Program Director in the AI Chip Center for Emerging Smart Systems. His research interest includes computational imaging algorithms and systems, particularly holographic microscopy and light field imaging. He is also a Fellow of OSA, SPIE, and IS&T.