HHS Public Access

Author manuscript

IEEE Trans Pattern Anal Mach Intell. Author manuscript; available in PMC 2023 April 01.

Published in final edited form as:

IEEE Trans Pattern Anal Mach Intell. 2022 April ; 44(4): 1934–1948. doi:10.1109/TPAMI.2020.3033882.

FlatNet: Towards Photorealistic Scene Reconstruction from Lensless Measurements

Salman S. Khan,
Department of Electrical Engineering, Indian Institute of Technology, Madras, TN, India, 600036

Varun Sundar,
Department of Electrical Engineering, Indian Institute of Technology, Madras, TN, India, 600036

Vivek Boominathan,
Rice University, Houston, TX, USA, 77005

Ashok Veeraraghavan,
Rice University, Houston, TX, USA, 77005

Kaushik Mitra
Department of Electrical Engineering, Indian Institute of Technology, Madras, TN, India, 600036

Abstract

Lensless imaging has emerged as a potential solution towards realizing ultra-miniature cameras by eschewing the bulky lens in a traditional camera. Without a focusing lens, the lensless cameras rely on computational algorithms to recover the scenes from multiplexed measurements. However, the current iterative-optimization-based reconstruction algorithms produce noisier and perceptually poorer images. In this work, we propose a non-iterative deep learning-based reconstruction approach that results in orders of magnitude improvement in image quality for lensless reconstructions. Our approach, called *FlatNet*, lays down a framework for reconstructing high-quality photorealistic images from mask-based lensless cameras, where the camera's forward model formulation is known. FlatNet consists of two stages: (1) an inversion stage that maps the measurement into a space of intermediate reconstruction by learning parameters within the forward model formulation, and (2) a perceptual enhancement stage that improves the perceptual quality of this intermediate reconstruction. These stages are trained together in an end-to-end manner. We show high-quality reconstructions by performing extensive experiments on real and challenging scenes using two different types of lensless prototypes: one which uses a separable forward model and another, which uses a more general non-separable cropped-convolution model. Our end-to-end approach is fast, produces photorealistic reconstructions, and is easy to adopt for other mask-based lensless cameras.

Keywords

lensless imaging; image reconstruction

sk39@smail.iitm.ac.in .
S. S. Khan and V. Sundar contributed equally to this work.

1 INTRODUCTION

Emerging applications such as wearables, augmented reality, virtual reality, biometrics, and many others are driving an acute need for highly miniaturized imaging systems. Unfortunately, current-generation cameras are based on lenses – and these lenses typically account for more than 90% of the cost, volume and weight of cameras. While lenses and optics have been miniaturized by two orders of magnitude, over the last century, we are inching up against fundamental laws (diffraction limit and Lohman's scaling law [3]) precluding further miniaturization.

Over the last decade, lensless imaging systems have emerged as a potential solution for light-weight, ultra-compact, inexpensive imaging. The basic idea in lensless imaging is to replace the lens with an amplitude [1] or a phase mask [2], [4]; typically placed quite close to the sensor. These lensless imaging systems provide numerous benefits over lens-based cameras. The need for a lens, which is a major contributor towards the size and weight of a camera, is eliminated. In addition, a lensless design permits a broader class of sensor geometries, allowing sensors to have more unconventional shapes (e.g. spherical or cylindrical) or to be physically flexible [5]. Moreover, lensless cameras can be produced with traditional semiconductor fabrication technology and therefore exploit all of its scaling advantages - yielding low-cost, high-performance cameras [6].

Due to the absence of any focusing element, the sensor measurements recorded in a lensless imager are no longer photographs of the scene but rather highly multiplexed measurements. Reconstruction algorithms are needed to undo the effects of this multiplexing and produce photographs of the scene being imaged. However, the design of a recovery algorithm for lensless cameras is a challenging task mainly because of the large support of the Point Spread Functions (PSFs) inherent to lensless design. In particular, the recovery algorithms face the following challenges. First, large support of PSFs result in large linear systems which makes such systems difficult to store and invert. Second, large PSFs also result in a very high degree of global multiplexing. Conventional data-driven methods like convolutional neural networks which are designed for natural images are not suited to handle this amount of multiplexing due to their limited receptive field. Third, lensless design results in ill-conditioned systems which affect the quality of reconstruction as well as noise characteristic of such systems. The poor reconstruction quality can be observed in the Tikhonov regularized reconstructions shown in Figure 1. Therefore, lensless cameras need robust and efficient algorithms to overcome these challenges.

Keeping the above challenges in mind, we propose a feed-forward deep neural network for photorealistic lensless reconstruction, which we refer to as *FlatNet*. FlatNet learns a direct mapping from lensless measurements to scene outputs. FlatNet consists of two stages: the first stage is a learnable inversion stage that brings the multiplexed measurements back to image space. This stage depends on the camera model. The second stage enhances this intermediate reconstruction using a fully convolutional network. It should be noted that the two stages are trained in an end-to-end fashion. It was shown in [2] that separable lensless mask based lensless cameras have inferior characteristics as compared to their existing non-separable counterparts. In our previous work [7], we had demonstrated FlatNet's

effectiveness for separable lensless model. But it cannot be trivially used for non-separable mask based lensless cameras. Here we extend the previous work to handle non-separable lensless model. In particular, we propose an efficient implementation of the learnable intermediate mapping for non-separable lensless model which is based on Fourier domain operations. We also propose an initialization scheme for this learnable intermediate stage that doesn't require explicit PSF calibration. We show that the intermediate mapping is robust for cases where the lensless model is non-circulant. This happens when the sensor size is smaller than the full measurement size required for deconvolution. Finally, to verify the robustness and efficiency of FlatNet, we perform extensive experiments on challenging real scenes captured using separable mask based lensless camera called FlatCam [1] and the non-separable mask based lensless camera called PhlatCam [2]. To summarize, the key contributions of this paper are:

- We propose an efficient implementation for the learnable intermediate stage of non-separable or general lensless model. In [7], we had only shown this for the separable lensless model. Here we non-trivially extend it to the general lensless case.
- We verify the robustness of the proposed learnable intermediate mapping for the non-separable lensless model on challenging scenarios where the lensless system does not follow a full convolutional or circulant assumption.
- We propose an initialization scheme for the non-separable lensless model that doesn't require explicit PSF calibration.
- Similar to the display and direct captured measurements collected using the separable mask FlatCam and described in our previous work [7], we collect corresponding datasets for the non-separable mask PhlatCam [2].
- We also collect a dataset of unconstrained indoor lensless measurements paired with corresponding unaligned webcam images which is finally used to finetune our proposed FlatNet to robustly deal with unconstrained real-world scenes.
- Our method outperforms previous traditional and deep learning based lensless reconstruction methods.

1.1 Related work

1.1.1 Lensless imaging—Lensless imaging involves capturing an image of a scene without physically focusing the incoming light with a lens. It has been widely used in the past for X-ray and gamma ray imaging for astronomy [8], [9], but its use for visible spectrum applications has only recently been studied. In a lensless imaging system, the scene is captured either directly on the sensor [10] or after being modulated by a mask element. Types of masks that have been used include phase gratings [11], random diffusers [4], designed phasemasks [2], amplitude masks [1], [12], compressive samplers [13], [14] and spatial light modulators [15], [16]. Replacing lens with the above masks result in multiplexed sensor capture that lacks any resemblance to the scene imaged. A recognizable image is then recovered using a computational reconstruction algorithm. In this paper, we

develop a deep learning based reconstruction algorithm for both separable and non-separable mask based lensless cameras.

1.1.2 Image reconstruction—Image reconstruction is a core aspect of most computational imaging problems [1], [2], [4], [17], [18]. In general, image reconstruction for computational imaging is ill-posed and requires regularization. Traditional methods for image reconstruction involve solving regularized least squares problems. Numerous regularizers based on heuristics have been developed in the past. These include the sparsity in gradient domain [2], [4], [19], wavelet/frequency domain sparsity [20], etc. However, these methods suffer from the fact that often the resulting cost function doesn't have a closed-form minima and an iterative approach has to be taken to solve it. Moreover, the regularizers are based on heuristics and may not be ideal for the specific task at hand.

Deep neural network have also been designed to solve image reconstruction problems in computational imaging systems. A class of deep learning based solution involves learning of regularizers or proximal mapping stage and then iteratively solving a MAP problem. Methods like [21], [22], [23] fall under this category. Another class of algorithm is designed as a feed-forward deep neural network that has either been trained in a supervised or self-supervised manner. Works on compressive image recovery [24], [25], [26], Fourier Ptychography [27], lensless recovery [28] fall under this category. Among these feed-forward networks, [26], [28] are inspired by the physics of the imaging model and are unrolled versions of traditional optimization frameworks. Although these methods provide interpretability, the drawbacks they offer include increased computation and higher memory consumption due to large number of unrolled iterations. The proposed method and its preliminary version [7] fall under the category of physics inspired deep neural network as well. However, they don't involve any unrolling thereby avoiding large computational and memory cost.

2 MASK BASED LENSLESS IMAGING

Mask based lensless imagers, unlike their lens-based counterparts, measure a global linear multiplexed version of the scene. This multiplexing is a function of the mask placed in front of the sensor. Mathematically, this is given as:

$$y = \Phi x + n, \tag{1}$$

where x and y are the vectorized representations of the scene and measurement respectively, Φ represents the generalized linear transformation, and $\mathbf{n}$ is the additive noise. In general, Φ has a large memory footprint, and hence, storing and computing with Φ is computationally intractable. Reconstructing a scene with $O(N^2)$ pixels from a sensor measurement of $O(N^2)$ pixels requires Φ with $O(N^4)$ elements. For example, a 1-megapixel scene and a 1-megapixel sensor requires Φ with $\sim 10^{12}$ elements. However, by careful design of masks and using a forward model derived from physics, the computational complexity can be greatly reduced.

The modulation performed by the mask characterizes the linear matrix Φ. By using a low-rank separable mask pattern, the huge Φ can be broken down into smaller matrices [1], [29]. Specifically, in [1], the single-separable lensless forward model reduces to:

$$Y = \Phi_L X \Phi_R^T + N, \tag{2}$$

where, Φ_L and Φ_R are the separable breakdown of Φ, X is the 2D scene irradiance, Y is the 2D recorded measurement, and N models additive noise.

By adding a small enough aperture over a non-separable mask and thereby ensuring that the off-axis shifted PSF stays within the sensor, [2] showed that the lensless forward model can be written as a convolutional model:

$$Y = P * X + N, \tag{3}$$

where P is PSF of the system. PSF of a lensless camera is the pattern projected by the mask on the sensor when illuminated by a single point source [2]. The PSF shifts when the point source moves laterally, and for a general scene, the sensor measurement is the weighted sum of various shifted PSFs, leading to a convolutional model.

If the sensor isn't large enough compared to the PSF, the PSF can shift out of the sensor for an oblique angled scene point. In such a case, [4] uses a cropped convolution model:

$$Y = C(P * X) + N, \tag{4}$$

where C is the sensor cropping operation. Such a system described by Equation 4 is no longer circulant. For a separable mask, the cropping is already incorporated in the model matrices Φ_L and Φ_R.

In this work, we will be primarily focusing on two prototypes of lensless cameras, (a) FlatCam [1] that has a separable mask and, (b) PhlatCam [2] that has a non-separable mask. We explore a data-driven approach that incorporates the lensless imaging models to produce photorealistic reconstructions from the above cameras. We also explore an alternate approach to sensor cropping for PhlatCam by preprocessing the sensor measurement [30].

3 FLATNET

To address the challenges involved in lensless image reconstruction, we take a data-driven approach for scene recovery. We model our reconstruction framework into a two stage fully trainable deep network. This two stage network is then jointly trained in an adversarial setup.

Trainable camera inversion.

The first stage of FlatNet is a learnable intermediate mapping called the *Trainable Camera Inversion* stage that learns to invert the lensless forward model obtaining intermediate reconstructions from globally multiplexed lensless measurements. We implement separate formulations of this trainable inversion stage for separable and non-separable lensless models exploiting the properties of the forward model for each type of these lensless systems.

Perceptual enhancement.

The second stage of FlatNet, called the *Perceptual Enhancement* stage, is a fully convolutional network that enhances the intermediate reconstruction obtained from the trainable inversion stage giving it more photorealistic appearance.

3.1 Trainable camera inversion—In the first stage of our network, we learn to invert the forward operation of the lensless camera model. This allows us to obtain an intermediate representation with local structures intact. To implement this, we follow a separate approach for separable and non-separable lensless camera models. Owing to the computational simplicity of a separable model, we will first describe the implementation of the inversion stage for the separable model.

3.1.1 Separable model: Given the lensless model described in Equation 2, we learn two layers of left and right trainable matrices that act directly on 2-D measurements. This can be mathematically represented as,

$$X_{\text{interm}} = f(W_1 Y W_2), \tag{5}$$

where X_{interm} is the output of this stage, f is a pointwise nonlinearity (which in our case is a leaky ReLU), Y is the input measurement, and W_1 and W_2 are the corresponding weight matrices for this stage. The dimension of the weight matrices depends on the dimension of the measurement and the scene dimension we want to recover i.e. the dimension of W_1 is the same as the dimension of the transpose of Φ_L while the dimension of W_2 is the same as the dimension of Φ_R. Eventually, these matrices learn to invert the forward matrices Φ_L and Φ_R. We refer to this version of FlatNet for separable lensless model as FlatNet-sep. It is important to initialize the weight matrices of this stage properly, so that the network does not get stuck in local minima. This can be done in two ways.

Calibrated initialization.

For this approach, we initialize our weight matrix W_1 with the transpose of Φ_L and W_2 with Φ_R, akin to back-projection. These calibration matrices (Φ_L and Φ_R) in (2) are physically obtained by the method described in [1]. This mode of initialization leads to faster convergence while training.

Uncalibrated initialization.

Calibration of FlatCam require careful alignment with display monitor [1], which can be a time consuming and inconvenient process especially for large volumes of FlatCams. Even a small error in calibration can lead to severe degradation in the performance of the reconstruction algorithm. To overcome the problems involved in calibration, we also propose a calibration-free approach by initializing the weight matrices with carefully designed pseudo-random matrices.

Initializing with any pseudo-random matrices of appropriate size does not yield successful reconstruction. To carefully design the random initialization, we make the following two observations regarding the FlatCam forward model: the calibration matrices have a ‘toeplitz-

like' structure and the slope of constant entries in the 'toeplitz-like' structure can be approximately determined using the FlatCam geometry, in particular the distance between the mask and the sensor and the pixel pitch. As the FlatCam's geometry is known apriori, we can construct the pseudo-random 'toeplitz-like' matrices with appropriate slope, and size, thereby making our approach calibration free. We discuss the generation of these pseudo-random matrices in more detail in the supplementary. The weight matrix W_1 is initialized with the adjoint of the random matrix constructed corresponding to Φ_L, while the matrix W_2 is initialized with the random matrix constructed corresponding to Φ_R. We observed that the training time increased slightly for this initialization in comparison to transpose initialization.

3.1.2 Non-separable model—Unlike in the separable model, it is infeasible to implement the trainable inversion stage in the non-separable model as a matrix multiplication layer owing to the extremely large dimension of Φ. However, one can still implement it in the Fourier domain. In order to implement the inversion stage efficiently, we analyze the forward model given in Equations 1 and 3. Following the observation that the forward model is purely convolutional for an appropriate sensor dimension i.e. the forward operation is described by Equation 3, we model our trainable inversion stage for the non-separable case in the form of a learned inverse implemented as Hadamard product in Fourier domain. This stems from the fact that the inverse of a circulant system given by Equation 3 is also circulant and can be diagonalized by Fourier transform.

Mathematically, this operation is given as,

$$X_{\text{interm}} = \mathcal{F}^{-1}(\mathcal{F}(W) \odot \mathcal{F}(Y)), \tag{6}$$

where X_{interm} is the output of this stage and Y is the measurement, $\mathcal{F}(.)$ and $\mathcal{F}^{-1}(.)$ are the DFT and the Inverse DFT operations, W is the filter that is learned (akin to W_1 and W_2 in the separable model) and $\odot$ refers to Hadamard product. For a $N \times M$ dimensional measurement, the dimension of W is $N \times M$. We found that using nonlinearity such as ReLU has no noticeable effect on the final output and as a result we did not include it in the non-separable model. The convolutional model of Equation 3 would require a large sensor as the PSF's in lensless systems have large spatial dimension and in some scenarios it would be infeasible to use such a large sensor. Such a case would require the lensless model to follow Equation 4. Of course, we cannot accurately represent the inverse of the system described by Equation 4 through a convolutional filter as the system is no longer circulant. As a result, one could ask if the proposed trainable inversion stage will still be valid if a smaller sensor was used? To answer this question, we show in Section 4.4.2, that with a small modification to the trainable inversion stage described in Equation 6, we can handle these cropped-convolutional or non-circulant cases without significant drop in the performance. We refer to this version of FlatNet for non-separable lensless model as FlatNet-Gen.

Calibrated initialization.

Like the separable model, initialization of W is important for convergence of the training process. Assuming we have a calibrated PSF and H is the Fourier transform of this PSF, in

our experiments, we initialize W using $\mathcal{F}^{-1}(\frac{H*}{K+|H|^2})$, i.e the regularized pseudo-inverse of the PSF or the well-known Wiener filter. In this expression, K is a regularization parameter.

Uncalibrated initialization.

We also propose an initialization scheme that doesn't require explicit PSF calibration. Given the mask pattern and the camera geometry, one can simulate the PSF of the lensless systems. Specifically, for PhlatCam, given the height profile of the mask, we use Fresnel propagation to simulate the PSF as described in [2]. This initialization scheme is particularly useful for cases where the PSF exceeds the sensor size (see Section 4.4.2). It should be noted here that this mode of initialization can be used for cases where we have access to height profile, for example in [2]. For cases where getting a rough estimate of the height profile is not possible, for example when random diffusers are used, calibrated mode of initialization should be preferred.

3.2 Perceptual enhancement—Once we obtain the output of the trainable inversion stage, which is of same dimension as that of the natural image we want to recover, we use a fully convolutional network to map it to the perceptually enhanced image. Owing to its large scale success in image-to-image translation problems and its multi-resolution structure, we choose a U-Net [31] to map the intermediate reconstruction to the final perceptually enhanced image. We keep the kernel size fixed at 3×3 while the number of filters is gradually increased from 128 to 1024 in the encoder and then reduced back to 128 in the decoder. In the end, we map the signal back to 3 RGB channels.

For the non-seperable case, we deal with slightly larger dimensional scenes. Similar to [35], we find it useful to employ Pixel-Shuffle [36] to downsample intermediate image before U-Net. By allowing U-Net to operate on a smaller spatial resolution (as a result bigger contextual area), we recover finer details for the increased image dimensions. Moreover, downsampling by Pixel-Shuffle doesn't throw away pixels and hence can be inverted exactly unlike other downsampling methods.

3.3 Discriminator architecture—We train FlatNet-sep and FlatNet-gen in an adversarial setup. We use a discriminator framework to classify FlatNet's output as real or fake. We find that using a discriminator network improves the perceptual quality of our reconstruction. We use 4 layers of 2-strided convolution followed by batch normalization and the swish activation function [37] in our discriminator. Same discriminator architecture was used for both FlatNet-sep and FlatNet-gen.

3.4 Loss function—An appropriate loss function is required to optimize our system to provide the desired output. Pixelwise losses like mean absolute error (MAE) or mean squared error (MSE) have been successfully used to capture signal distortion. However, they fail to capture the perceptual quality of images. As our objective is to obtain high quality photorealistic reconstructions from lensless measurements, perceptual quality matters. Thus, we use a weighted combination of signal distortion and perceptual losses. The losses used for our model are given below:

Mean squared error: We use MSE to measure the distortion between the ground truth and the estimated output. Given the ground truth image I_{true} and the estimated image I_{est}, this is given as:

$$\mathscr{L}_{\text{MSE}} = \|I_{\text{true}} - I_{\text{est}}\|_2^2. \tag{7}$$

Perceptual loss: To measure the semantic difference between the estimated output and the ground truth, we use the perceptual loss introduced in [32]. We use a pretrained VGG-16 [33] model for our perceptual loss. We extract feature maps between the second convolution (after activation) and second max pool layers, and between the third convolution (after activation) and the fourth max pool layers. We call these activations ϕ_{22} and ϕ_{43}, respectively. This loss is given as,

$$\begin{aligned}\mathscr{L}_{\text{percept}} = \|\phi_{22}(I_{\text{true}}) - \phi_{22}(I_{\text{est}})\|_2^2 + \\ \|\phi_{43}(I_{\text{true}}) - \phi_{43}(I_{\text{est}})\|_2^2\end{aligned} \tag{8}$$

Adversarial loss: Adversarial loss [34], [38] was added to further bring the distribution of the reconstructed output close to those of the real images. Given the discriminator D described in Section 3.3, this loss is given as,

$$\mathscr{L}_{\text{adv}} = -\log(D(I_{\text{est}})). \tag{9}$$

Our discriminator, consisting of 4 layers of 2-strided convolution followed by batch normalization and ReLU activation function, classifies the generator output as real or fake.

Total generator loss: Our total loss for the FlatNet while training is a weighted combination of the three losses and is given as,

$$\mathscr{L} = \lambda_1 \mathscr{L}_{\text{MSE}} + \lambda_2 \mathscr{L}_{\text{percept}} + \lambda_3 \mathscr{L}_{\text{adv}}. \tag{10}$$

where, λ_1, λ_2 and λ_3 are weights assigned to each loss.

Discriminator loss: Given I_{est}, I_{true} and discriminator D, the discriminator was trained using the following loss,

$$\mathscr{L}_{\text{disc}} = -\log(D(I_{\text{true}})) - \log(1 - D(I_{\text{est}})). \tag{11}$$

Contextual Loss: For finetuning FlatNet-gen on unaligned PhlatCam and webcam pairs (described in Section 4.5), we use only contextual loss as proposed in [39]. Denoting output image features ($\phi_{44}(I_{est})$) as $\{p_i\}_{i=1}^{N}$, target image features ($\phi_{44}(I_{true})$) as $\{q_j\}_{j=1}^{N}$ and number of pixels in each of these feature maps as N, contextual loss finds the nearest neighbour feature match $q = \text{argmin}_q \mathbb{D}(p, q_j)_{j=1}^{N}$ for each p. We then minimize the summed

distance of all such feature pairs. The distance metric we adopt here is cosine-distance, although it could also be L_1, L_2, etc. This loss term is given by:

$$\mathscr{L}_{\text{contextual}} = \frac{1}{N}\sum_{i=1}^{N} \min_{j \in [N]} \mathbb{D}\left(p_i, q_j\right) \tag{12}$$

We found ϕ_{44} to be a suitable feature extractor based on the computational cost and sharpness of reconstruction.

4 EXPERIMENTS AND RESULTS

In this section, we describe all our experiments. We perform all our experiments on real data. We will refer to the FlatNet for separable model as FlatNet-sep and for the non-separable model as FlatNet-gen. They will further be suffixed by -C and -UC to indicate calibrated or uncalibrated method of initialization respectively. Unless specifically mentioned, simply using FlatNet-gen or FlatNet-sep would indicate FlatNet-gen-C or FlatNet-sep-C i.e. FlatNets initialized with the calibrated method of initialization.

4.1 Dataset

Supervised training of deep neural networks require large scale labelled dataset. However, collecting a large scale dataset for lensless images is a challenging task. One could use the known lensless model to simulate measurements from the available natural image datasets. This, however, will sometimes fail to mimic the true imaging model due to several non-idealities. To overcome this challenge, we collect a large dataset by projecting images on monitors and capturing this projection using lensless cameras. This not only takes care of the true imaging model for lensless camera, it also helps us collect a labelled dataset for lensless images. We follow the same dataset collection procedure for both FlatCam [1] and PhlatCam [2]. For our work, we use a subset of ILSVRC 2012 [40]. Specifically, we used 10 random images from each class as our ground truth. Of the 1000 classes, we kept 990 classes for training and the rest for testing. So in total, we used 9900 images for training and 100 images for testing. Before capturing the dataset, we resize the images displayed on monitor so as to cover the entire field of view (FoV) of camera. We call this dataset the Display Captured Dataset. For this dataset, the ground truth images are the ones that were projected on the monitor screen. The monitor was kept beyond the hyperfocal distance of the cameras to avoid the variation of the PSF with depth. The hyperfocal distance for the FlatCam prototype is around a foot and for the PhlatCam prototype is around 16 inches. To test the FlatNet on real scenes, we also capture measurements of objects placed directly in front of the camera. Using FlatCam we collect 15 such measurements while using PhlatCam we collect 20 such measurements. We call this dataset Direct Captured Dataset. This dataset doesn't have corresponding ground truths for the measurements. To demonstrate the effectiveness of FlatNet-gen on unconstrained indoor scenarios, we collect a dataset of unaligned PhlatCam and webcam captures using the setup described in Figure 13. This dataset consists of 475 training samples and 25 test samples. We call this dataset the Unconstrained Indoor Dataset. Samples from our datasets can be seen in Figure 3. We will release this dataset upon acceptance of this manuscript

4.2 Implementation details

The FlatCam prototype uses a Point Grey Flea3 camera with 1.3MP e2v EV76C560 CMOS sensor and a pixel size of 5.3 μm. All the ground truth images were resized to 256 × 256 as the FlatCam is calibrated to produce 256 × 256 output images. This ensures that there is no misalignment among the input and ground truth pairs. We directly used the Bayer measurements, split into 4 channels (R,Gr,Gb,B), as our input to the network and convert them into 3 channel RGB within the network. FlatCam measurements of dimension 512×640×4 in batches of 4 were used as inputs for training. A smaller batch size was used due to memory constraints. We set λ_1 as 1, λ_2 to be 1.2 and λ_3 to be 0.6. For transpose initialization, we trained our model for 45K iterations while for random initialization, we trained it for 60K iterations. The Adam [41] optimizer was used for all models. We started with a learning rate of 10^{-4} and gradually reduced it by half every 5000 iterations. The PhlatCam prototype used is a Basler Ace4024–29uc with 12.2MP Sony IMX226 sensor with a pixel size of 1.85μm. All the ground truth images were resized to 384 × 384 which is equal to the FoV of the prototype. We directly used the Bayer measurements, split into 4 channels (R,Gr,Gb,B), as our input to the network and convert them into 3 channel RGB within the network. We used the same set of λ_i's as that for FlatNet-sep. The full measurements used were of dimension 1280 × 1408 × 4. For the small sensor experiments of Section 4.4.2, we use measurements of dimension 608 × 864 × 4.

4.3 Comparison with other approaches

4.3.1 Separable lensless model—In this subsection, we show results for the amplitude mask FlatCam that follows a separable model.

We compare FlatNet-sep with the closed form Tikhonov reconstruction described in [1] and a total variation based reconstruction implemented using TVAL3 [19].

Qualitative discussion.: In Figure 4, we compare our methods, FlatNet-sep-UC with uncalibrated initialization and FlatNet-sep-C with calibrated initialization, with traditional methods, Tikhonov and TVAL3. As can be observed from the reconstructions, the Tikhonov regularized reconstructions are prone to severe vignetting effects which is somewhat reduced in the TVAL3 results. Inset images in Figure 4 show the preservation of finer details in our approach. Figure 5 shows the performance of the various methods for direct captured measurements. Tikhonov regularization has a tendency to suppress low signal values and as a result has difficulty restoring the poorly illuminated background for most of the scenes in Figure 5. The performance of TVAL3 [19] is also similar. FlatNet-sep, on the other hand, produces higher quality photorealistic reconstruction. Note that our uncalibrated model FlatNet-sep-UC gives similar performance to that of the calibrated model FlatNet-sep-C. Thus, our method does not require explicit calibration unlike the rest of the approaches.

Quantitative discussion.: We present the quantitative performance of FlatNet for separable mask FlatCam in Table 1. For evaluation, we use PSNR, SSIM and the recently proposed LPIPS [42]. Higher PSNR and SSIM score indicate better performance while lower LPIPS indicates better perceptual quality. It can be clearly seen that our approach using transpose initialization (FlatNet-sep-C) outperforms all the other reconstruction techniques

for FlatCam. The next best approach is the FlatNet-sep using random initialization (FlatNet-sep-UC), which unlike other methods, is a calibration-free technique. We also compare the inference time for various approaches in the same table. The Tikhonov and TVAL3 [19] regularized reconstructions are computed on Intel Core i7 CPU with 16 GB RAM while the rest of the approaches are evaluated on Nvidia GTX 1080 Ti GPU.

4.3.2 Non-separable lensless model—For experiments on the non-separable model, we compare FlatNet-gen with traditional and learning based approaches. We describe these approaches below.

Traditional approaches.: In traditional method, we compare FlatNet-gen with traditional Tikhonov regularized reconstruction implemented in Fourier domain (as Wiener restoration filter) and total variation regularized reconstruction implemented using ADMM [4]

Learning based approaches.: For learning based approach, we use the unrolled deep network described in [28]. However, for fairness, we use the five stage unrolled ADMM followed by our perceptual enhancement stage.

Qualitative discussion.: Figure 6 shows the display captured reconstruction for PhlatCam. We can clearly see higher quality reconstruction for FlatNet-gen in comparison to traditional Tikhonov regularized reconstruction or Wiener deconvolution and ADMM based method. It also results in better quality reconstruction than the Le-ADMM model. This trend in performance is also observed in the direct captured reconstructions in Figure 7. It should also be noted that Le-ADMM, despite having fewer parameters, is extremely memory and computation intensive due to the large number of intermediates/primal and dual variables calculated at each stage of the unrolled ADMM. It is due to this significant increment in memory consumption, that it becomes infeasible to implement this model on the captured PhlatCam measurements without downsampling. In our comparison, we downsample the measurements by a factor of 4 (similar to [28]) before passing them through the Le-ADMM network. Unless explicitly mentioned, we will refer to this downsampled Le-ADMM model as Le-ADMM. Downsampling operation leads to compromise in the reconstruction resolution resulting in the lack of sharpness observed in the final reconstruction. On the other hand, the FlatNet-gen has significantly lower memory requirement that doesn't require any downsampling pre-processing thereby preventing any loss of sharpness or resolution. We also provide comparison for FlatNet-gen initialized with uncalibrated PSF in the supplementary material. We call this model FlatNet-gen-UC.

Quantitative discussion.: The quantitative results are provided in Table 2. Along with the uncalibrated FlatNet-gen model, we also provide the performance of uncalibrated version of Le-ADMM in this table. It is referred to as Le-ADMM-UC. The consistency with visual results is maintained in the quantitative metrics. It can be clearly seen that FlatNet-gen outperforms all other methods quantitatively. FlatNet-gen-UC performs almost at par with FlatNet-gen-C and outperforms Le-ADMM-UC. It should be noted that the difference between FlatNet-gen-C and FlatNet-gen-UC is smaller as compared to Le-ADMM-C and Le-ADMM-UC. This is primarily due to the stronger dependence of Le-ADMM on the true PSF while FlatNet-gen requires the knowledge of PSF only for better initialization and

learns to converge to a better inverse after training. We also provide the runtime for the methods compared. For Wiener and TV-based ADMM, we report the speed on CPU while for others we report the speed for a forward pass in GPU.

Assuming the true measurement is of dimension 1280 × 1408, we additionally compare FlatNet-gen's trainable inversion stage with the unrolled ADMM block of Le-ADMM (without the U-Net) in terms of memory and computation in Table 3. We provide the memory consumption (in Megabytes, computed on Nvidia GTX 1080 Ti GPU) and computations (in FLOPs, computed theoretically) required to process one image using the two methods. We unroll the ADMM for 5 iterations. In the table, Le-ADMM-Full refers to the unrolled ADMM without any downsampling while Le-ADMM-Downsampled refers to the case where the PSF and the scene were downsampled by a factor of 4. It can be observed that a full resolution Le-ADMM requires significant amount of memory which would have negative implications if deployment is considered. Moreover, appended with dense CNNs like U-Net, Le-ADMM-Full is difficult to implement on a conventional GPU, thereby necessitating the downsampling of the measurements which in turn leads to the degradation of the reconstruction quality. One should also note the amount of computations performed in the unrolled ADMM block for the particular dimensions of PSF and scene. Due to a series of intermediate estimates that depend on Fourier and Inverse Fourier transforms, this computation blows up for Le-ADMM-Full. FlatNet-gen provides a better trade-off for resolution, and memory and computational requirements which is essential for lensless systems which, by design, suffer from poor reconstruction resolution.

4.4 Further analysis

4.4.1 Effect of learning the inversion stage—In this section, we highlight the importance of the end-to-end learning strategy of FlatNet. We compare FlatNet with a network with just the perceptual enhancement block. We train this network with Tikhonov regularized reconstructions. For training this network, we use the same loss as defined in Equation 10. We call this method Tikh+U-Net. We implement this approach for both separable and non-separable lensless models. Top row of Figure 8 compares the reconstruction quality of FlatNet-sep with Tikh+U-Net. We can easily observe the improved quality of reconstruction obtained from FlatNet-sep compared to Tikh+U-Net. Tikh+U-Net suffers from blurrier reconstructions with amplified artifacts. We also compare the performance of FlatNet-gen with its corresponding Tikh+U-Net in the bottom row of Figure 8. FlatNet-gen provides sharper reconstructions over Tikh+U-Net.

Table 4 provides a quantitative flavor to the above analysis. We can see that FlatNet outperforms Tikh+U-Net for both separable and non-separable models in terms of PSNR and LPIPS.

One may notice that the difference between FlatNet-gen and Tikh+U-Net is not as significant as between FlatNet-sep and its corresponding Tikh+U-Net. This is due to the higher quality of Tikhonov reconstruction in the case of PhlatCam compared to FlatCam [2]. However, one should note that Tikh+U-Net is strictly based on convolutional assumption for the forward model, and performs poorly when this assumption is violated as will be verified in Section 4.4.2.

4.4.2 Performance on cropped measurements—As we have already seen in Section 2, the forward operation in a mask-based lensless camera is no longer convolutional if the size of the sensor is small compared to the true measurement size i.e. the forward model is given by Equation 4. This coupled with large PSFs, makes lensless reconstruction challenging for traditional reconstruction approaches which rely on the circulant or convolutional assumptions (e.g. Wiener deconvolution). This naturally leads to a question: Will the proposed trainable inversion layer of FlatNet-gen, which is based on learned Fourier domain inversion, be robust against cases where the deviation from the circulant assumption is significant? In other words, will FlatNet-gen be able to deal with measurements from which a significant amount of pixels have been thrown away due to the finite sensor size and fully open aperture? In this section, we show that we can deal with the small sensor size case without losing much in terms of reconstruction quality and perform better than Le-ADMM which explicitly tries to deal with the cropped out pixels. For our experiments, we take a central crop of size 608 × 864 from our 7MP full sensor measurement. Effectively, this can be thought as using a 2MP sensor instead of the 7MP sensor.

2

Following the observation in [30], we replicate pad our cropped measurements as a pre-processing step. To smooth the discontinuities due to padding, we multiply this padded measurement with a gaussian filtered box. The effectiveness of our method of padding can be observed in Figure 9. Mathematically, the trainable inversion stage changes to,

$$X_{\text{interm}} = \mathcal{F}^{-1}(\mathcal{F}(W) \odot \mathcal{F}(pad(Y))) . \tag{13}$$

This is a modification to Equation 6 to account for the cropped measurement. $pad(.)$ refers to the padding and smoothing operation described above. The same padding and smoothing procedure is also followed for Tikh+U-Net applied on the cropped measurements. Figure 10 shows the reconstruction quality for the display captured cropped measurement compared with full measurement for Tikh+U-Net, Le-ADMM and FlatNet. Even after padding the measurements, there are artifacts in the Wiener restored images that cannot be effectively removed using Tikh+U-Net. Le-ADMM performs slightly better than Tikh+U-Net due to its intermediate stage that approximately estimates the uncropped measurement. However, it is not as robust to crop as FlatNet-gen is. Similarly, in Figure 11, we show the reconstructions for direct captured cropped measurement. It can be clearly seen that Tikh+U-Net and Le-ADMM suffer from significant color artifacts. These artifacts are however not significant in the FlatNet-gen reconstructions. Table 5 gives the comparison of average scores for each model on the display captured dataset.

It should be noted that for the model used to obtain Figures 10 and 11 and Table 5, the PSF size (608 × 870) exceeds the assumed sensor size (606 × 864). In such a case, estimation of the true PSF is a tedious process and one can use the uncalibrated FlatNet-gen-UC. From Table 5, we can see that FlatNet-gen outperforms all other learned methods. FlatNet-gen-UC has a comparable performance to FlatNet-gen, while Tikh+U-Net-UC and Le-ADMM-UC breakdown: indicating that accurate PSF calibration is required for these

methods. The visual results for FlatNet-gen-UC for cropped measurements are provided in the supplementary material.

Apart from the crop size mentioned above, we also show the performance of the learning based approaches for various different crop sizes in Figure 12. Here, we normalize the size of the cropped measurements with respect to the full measurements. It can be seen that FlatNet-gen consistently outperforms Le-ADMM and Tikh+U-Net for all crop sizes.

It should also be noted that FlatNet-sep is, by design, robust to non-circulant scenarios as it involves learned inversion in the spatial domain.

4.5 Performance on unconstrained indoor scenes

In the previous sections, we performed all our experiments using FlatNets trained on display captured dataset. However, real measurements captured in the wild differs from the dispay captured measurements for the following reasons: a) real world captures have significantly higher amount of noise compared to display captured measurements, b) in an unconstrained setup, bright scene points beyond the FoV described by the Chief Ray Angle (CRA) can also influence the captured measurement which is not the case with display captured measurements captured with monitors filling the whole of CRA defined FoV. To take these differences into account and make our FlatNet robust to real world scenarios, we finetune FlatNet using a real world dataset we captured called the Unconstrained Indoor Dataset. This dataset consists of unaligned webcam and PhlatCam captures collected using the setup described in Figure 13. We collected 500 pairs of such data, keeping 475 pairs for training and 25 for testing. We finetune the entire network with a small learning rate (10^{-12} for the trainable inversion stage and 10^{-6} for the perceptual enhancement stage). To account for misalignment between PhlatCam and webcam captures, we only use Contextual Loss [39] which was previously proposed for unaligned data. Figure 13 shows some of our reconstruction results with and without finetuning along with webcam captures for reference. It can be observed that finetuning results in more photorealistic reconstructions. In the supplementary material, we show reconstructions from cropped unconstrained indoor measurements.

5 DISCUSSION AND CONCLUSION

In this paper, we propose an end-to-end trainable deep network called FlatNet for photorealistic scene reconstruction from lensless measurements. Despite the numerous promises that lensless imaging provides, it is somewhat restricted by the quality of the reconstructed image. In this paper, we have attempted to bridge this gap between the promise of lensless imaging and its performance. FlatNet leverages the physics of the forward model (through the trainable camera inversion) and the success of data-driven approaches to learn a photorealistic mapping from the highly multiplexed lensless captures to the estimated scene. Unlike unrolling based networks [28], it has the advantage of low memory and computational requirements which are desirable criteria for stand-alone devices. We also show that by finetuning FlatNet trained on display captured measurements, using unaligned Webcam-PhlatCam indoor scenes, we can recover photorealistic images in the wild using these ultra-thin sensors.

It should also be noted that like most GAN based approaches, FlatNet reconstructions suffer from hallucination artifacts that favor photorealism over high-fidelity. Therefore, FlatNet should be used with caution when the task at hand is critical to these hallucination artifacts (for example medical imaging). Nevertheless, in such critical systems, one can still use the trainable camera inversion of FlatNet and make modifications to the perceptual enhancement and the losses appropriately.

In future, it would be interesting to look into the co-design of mask or PSF and reconstruction algorithm for mask-based lensless cameras.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGMENTS

This work was supported in part by NSF CAREER: IIS-1652633, NSF EXPEDITIONS: CCF-1730574, DARPA NESD: HR0011-17-C0026, NIH Grant: R21EY029459 and the Qualcomm Innovation Fellowship India.

Biography

Salman Siddique Khan is currently a Ph.D. student in the Department of Electrical Engineering, IIT Madras, India. He received the B.Tech degree in Electronics and Instrumentation Engineering from the National Institute of Technology, Rourkela, India in 2018. His research interests include computational imaging and computer vision.

Varun Sundar is presently an undergraduate in the Department of Electrical Engineering at IIT Madras, India. He is also an incoming PhD student at the University of Wisconsin Madison. At IIT Madras, he is associated with the Computational Imaging lab, where he worked on lensless imaging systems.

Vivek Boominathan received the B.Tech degree in Electrical Engineering from the Indian Institute of Technology Hyderabad, Hyderabad, India, in 2012, and the M.S. and Ph.D. degree in 2016 and 2019, from the Department of Electrical and Computer Engineering, Rice University, Houston, TX, USA. He is currently a Postdoctoral Associate with Rice University, Houston, TX. His research interests lie in the areas of computer vision, signal processing, wave optics, and computational imaging.

Ashok Veeraraghavan received the bachelor's degree in electrical engineering from the Indian Institute of Technology, Madras, Chennai, India, in 2002 and the M.S. and Ph.D. degrees from the Department of Electrical and Computer Engineering, University of Maryland, College Park, MD, USA, in 2004 and 2008, respectively. He is currently a Professor of Electrical and Computer Engineering, Rice University, Houston, TX, USA. Before joining Rice University, he spent three years as a Research Scientist at Mitsubishi Electric Research Labs, Cambridge, MA, USA. His research interests are broadly in the areas of computational imaging, computer vision, machine learning, and robotics. Dr. Veeraraghavan's thesis received the Doctoral Dissertation Award from the Department of Electrical and Computer Engineering at the University of Maryland. He is the recipient of

the National Science Foundation CAREER Award in 2017. At Rice University, he directs the Computational Imaging and Vision Lab.

Kaushik Mitra received the Ph.D. degree from the Department of Electrical and Computer Engineering, University of Maryland, College Park, MD, USA. He is currently an Assistant Professor with the Department of Electrical Engineering, Indian Institute of Technology Madras, Chennai, India. Before joining IIT Madras, he was a Postdoctoral Research Associate with the Department of Electrical and Computer Engineering, Rice University, Houston, TX, USA. His research interests include computational imaging, computer vision, and machine learning. His contributions to computational imaging include proposing a theoretical framework for analysis and design of novel computational imaging systems, development of novel imaging systems, such as hybrid light field camera and assorted camera array, and using machine learning techniques, such as dictionary learning and deep learning for improving the performance of computational imaging systems.

REFERENCES

[1]. Asif MS, Ayremlou A, Sankaranarayanan A, Veeraraghavan A, and Baraniuk RG, “Flatcam: Thin, lensless cameras using coded aperture and computation,” IEEE Transactions on Computational Imaging, vol. 3, no. 3, pp. 384–397, 2017.

[2]. Boominathan V, Adams J, Robinson J, and Veeraraghavan A, “Phlatcam: Designed phase-mask based thin lensless camera,” IEEE transactions on pattern analysis and machine intelligence, 2020.

[3]. Lohmann AW, “Scaling laws for lens systems,” Applied optics, vol. 28, no. 23, pp. 4996–4998, 1989. [PubMed: 20555989]

[4]. Antipa N, Kuo G, Heckel R, Mildenhall B, Bostan E, Ng R, and Waller L, “Diffusercam: lensless single-exposure 3d imaging,” Optica, vol. 5, no. 1, pp. 1–9, 2018.

[5]. Tremblay EJ, Stack RA, Morrison RL, and Ford JE, “Ultrathin cameras using annular folded optics,” Applied optics, vol. 46, no. 4, pp. 463–471, 2007. [PubMed: 17230237]

[6]. Boominathan V, Adams JK, Asif MS, Avants BW, Robinson JT, Baraniuk RG, Sankaranarayanan AC, and Veeraraghavan A, “Lensless imaging: A computational renaissance,” IEEE Signal Processing Magazine, vol. 33, no. 5, pp. 23–35, 2016.

[7]. Khan SS, Adarsh V, Boominathan V, Tan J, Veeraraghavan A, and Mitra K, “Towards photorealistic reconstruction of highly multiplexed lensless images,” in Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 7860–7869.

[8]. Dicke R, “Scatter-hole cameras for x-rays and gamma rays,” The astrophysical journal, vol. 153, p. L101, 1968.

[9]. Caroli E, Stephen J, Di Cocco G, Natalucci L, and Spizzichino A, “Coded aperture imaging in x-and gamma-ray astronomy,” Space Science Reviews, vol. 45, no. 3–4, pp. 349–403, 1987.

[10]. Kim G, Isaacson K, Palmer R, and Menon R, “Lensless photography with only an image sensor,” Applied optics, vol. 56, no. 23, pp. 6450–6456, 2017. [PubMed: 29047934]

[11]. Stork DG and Gill PR, “Lensless ultra-miniature cmos computational imagers and sensors,” Proc. SENSORCOMM, pp. 186–190, 2013.

[12]. Shimano T, Nakamura Y, Tajima K, Sao M, and Hoshizawa T, “Lensless light-field imaging with fresnel zone aperture: quasi-coherent coding,” Applied optics, vol. 57, no. 11, pp. 2841–2850, 2018. [PubMed: 29714287]

[13]. Huang G, Jiang H, Matthews K, and Wilford P, “Lensless imaging by compressive sensing,” in 2013 IEEE International Conference on Image Processing. IEEE, 2013, pp. 2101–2105.

[14]. Satat G, Tancik M, and Raskar R, “Lensless imaging with compressive ultrafast sensing,” IEEE Transactions on Computational Imaging, vol. 3, no. 3, pp. 398–407, 2017.

[15]. Chi W and George N, “Optical imaging with phase-coded aperture,” Optics express, vol. 19, no. 5, pp. 4294–4300, 2011. [PubMed: 21369259]

[16]. DeWeert MJ and Farm BP, “Lensless coded-aperture imaging with separable doubly-toeplitz masks,” Optical Engineering, vol. 54, no. 2, p. 023102, 2015.

[17]. Duarte MF, Davenport MA, Takhar D, Laska JN, Sun T, Kelly KF, and Baraniuk RG, “Single-pixel imaging via compressive sampling,” IEEE signal processing magazine, vol. 25, no. 2, pp. 83–91, 2008.

[18]. Antipa N, Oare P, Bostan E, Ng R, and Waller L, “Video from stills: Lensless imaging with rolling shutter,” in 2019 IEEE International Conference on Computational Photography (ICCP). IEEE, 2019, pp. 1–8.

[19]. Li C, Yin W, Jiang H, and Zhang Y, “An efficient augmented lagrangian method with applications to total variation minimization,” Computational Optimization and Applications, vol. 56, no. 3, pp. 507–530, 2013.

[20]. Reddy D, Veeraraghavan A, and Chellappa R, “P2c2: Programmable pixel compressive camera for high speed imaging,” in CVPR 2011. IEEE, 2011, pp. 329–336.

[21]. Dave A, Vadathya AK, Subramanyam R, Baburajan R, and Mitra K, “Solving inverse computational imaging problems using deep pixel-level prior,” IEEE Transactions on Computational Imaging, vol. 5, no. 1, pp. 37–51, 2018.

[22]. Dave A, Kumar A, Mitra K et al., “Compressive image recovery using recurrent generative model,” in 2017 IEEE International Conference on Image Processing (ICIP). IEEE, 2017, pp. 1702–1706.

[23]. Rick Chang J, Li C-L, Poczos B, Vijaya Kumar B, and Sankaranarayanan AC, “One network to solve them all–solving linear inverse problems using deep projection models,” in Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 5888–5897.

[24]. Kulkarni K, Lohit S, Turaga P, Kerviche R, and Ashok A, “Reconnet: Non-iterative reconstruction of images from compressively sensed measurements,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 449–458.

[25]. Mousavi A, Patel AB, and Baraniuk RG, “A deep learning approach to structured signal recovery,” in 2015 53rd Annual Allerton Conference on Communication, Control, and Computing (Allerton). IEEE, 2015, pp. 1336–1343.

[26]. Zhang J and Ghanem B, “Ista-net: Interpretable optimization-inspired deep network for image compressive sensing,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 1828–1837.

[27]. Boominathan L, Maniparambil M, Gupta H, Baburajan R, and Mitra K, “Phase retrieval for fourier ptychography under varying amount of measurements,” arXiv preprint arXiv:1805.03593, 2018.

[28]. Monakhova K, Yurtsever J, Kuo G, Antipa N, Yanny K, and Waller L, “Learned reconstructions for practical mask-based lensless imaging,” Optics express, vol. 27, no. 20, pp. 28 075–28 090, 2019.

[29]. Adams JK, Boominathan V, Avants BW, Vercosa DG, Ye F, Baraniuk RG, Robinson JT, and Veeraraghavan A, “Single-frame 3d fluorescence microscopy with ultraminiature lensless flatscope,” Science advances, vol. 3, no. 12, p. e1701548, 2017. [PubMed: 29226243]

[30]. Reeves SJ, “Fast image restoration without boundary artifacts,” IEEE Transactions on image processing, vol. 14, no. 10, pp. 1448–1453, 2005. [PubMed: 16238051]

[31]. Ronneberger O, Fischer P, and Brox T, “U-net: Convolutional networks for biomedical image segmentation,” in International Conference on Medical image computing and computer-assisted intervention. Springer, 2015, pp. 234–241.

[32]. Johnson J, Alahi A, and Fei-Fei L, “Perceptual losses for real-time style transfer and super-resolution,” in European conference on computer vision. Springer, 2016, pp. 694–711.

[33]. Simonyan K and Zisserman A, “Very deep convolutional networks for large-scale image recognition,” arXiv preprint arXiv:1409.1556, 2014.

[34]. Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, and Bengio Y, “Generative adversarial nets,” in Advances in neural information processing systems, 2014, pp. 2672–2680.

[35]. Gu S, Li Y, Gool LV, and Timofte R, “Self-guided network for fast image denoising,” in Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 2511–2520.

[36]. Shi W, Caballero J, Huszár F, Totz J, Aitken AP, Bishop R, Rueckert D, and Wang Z, “Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 1874–1883.

[37]. Ramachandran P, Zoph B, and Le QV, “Searching for activation functions,” arXiv preprint arXiv:1710.05941, 2017.

[38]. Ledig C, Theis L, Huszár F, Caballero J, Cunningham A, Acosta A, Aitken A, Tejani A, Totz J, Wang Z et al., “Photorealistic single image super-resolution using a generative adversarial network,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 4681–4690.

[39]. Mechrez R, Talmi I, and Zelnik-Manor L, “The contextual loss for image transformation with non-aligned data,” in Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 768–783.

[40]. Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, Huang Z, Karpathy A, Khosla A, Bernstein M, Berg AC, and Fei-Fei L, “ImageNet Large Scale Visual Recognition Challenge,” International Journal of Computer Vision (IJCV), vol. 115, no. 3, pp. 211–252, 2015.

[41]. Kingma DP and Ba J, “Adam: A method for stochastic optimization,” arXiv preprint arXiv:1412.6980, 2014.

[42]. Zhang R, Isola P, Efros AA, Shechtman E, and Wang O, “The unreasonable effectiveness of deep features as a perceptual metric,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 586–595.

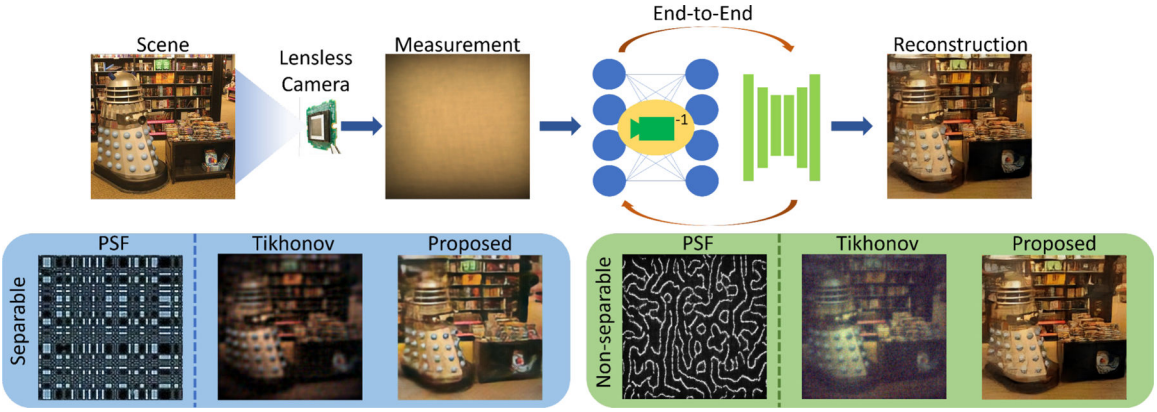

Fig. 1. Lensless imaging.

Lensless cameras require computation to recover the true scene from measurements. In this work we propose a deep learning based lensless reconstruction algorithm for both separable [1] and non-separable mask [2] based lensless cameras that produce photorealistic reconstructions for real and challenging scenarios.

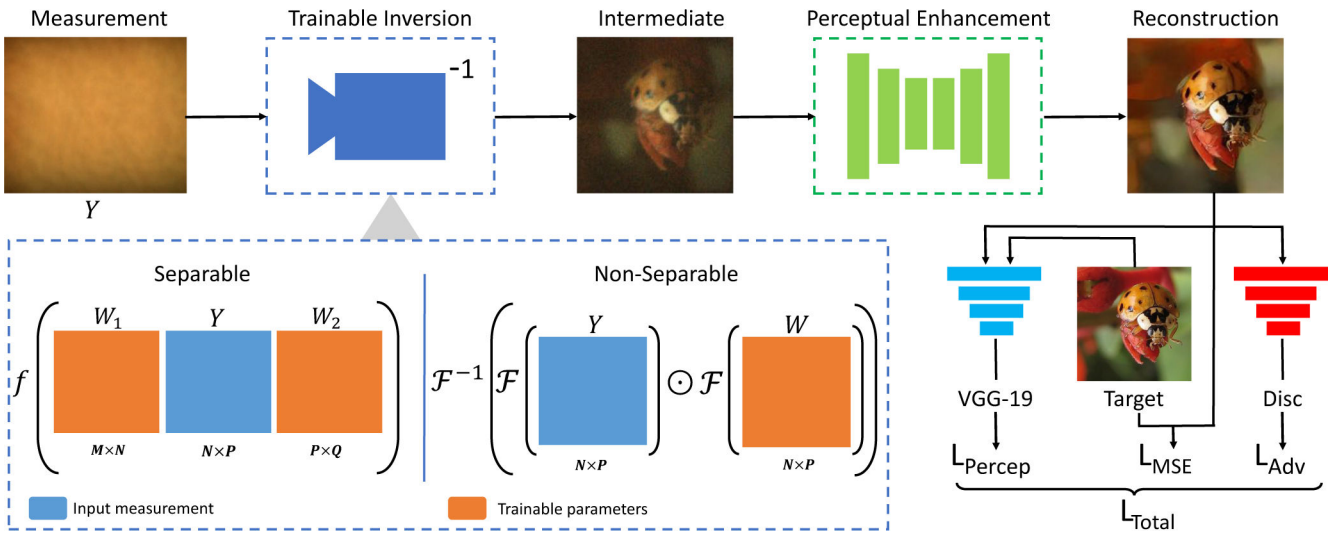

Fig. 2. Overall architecture of the FlatNet.

The lensless camera measurement is first mapped into an intermediate image space using a trainable camera inversion layer. This stage is implemented separately for the separable and the non-separable case. A U-Net [31] then enhances the perceptual quality of the intermediate reconstruction. We use a weighted combination of three losses in training our network: a perceptual loss [32] using a VGG16 network [33], mean-square error (MSE), and adversarial loss using a discriminator neural network [34].

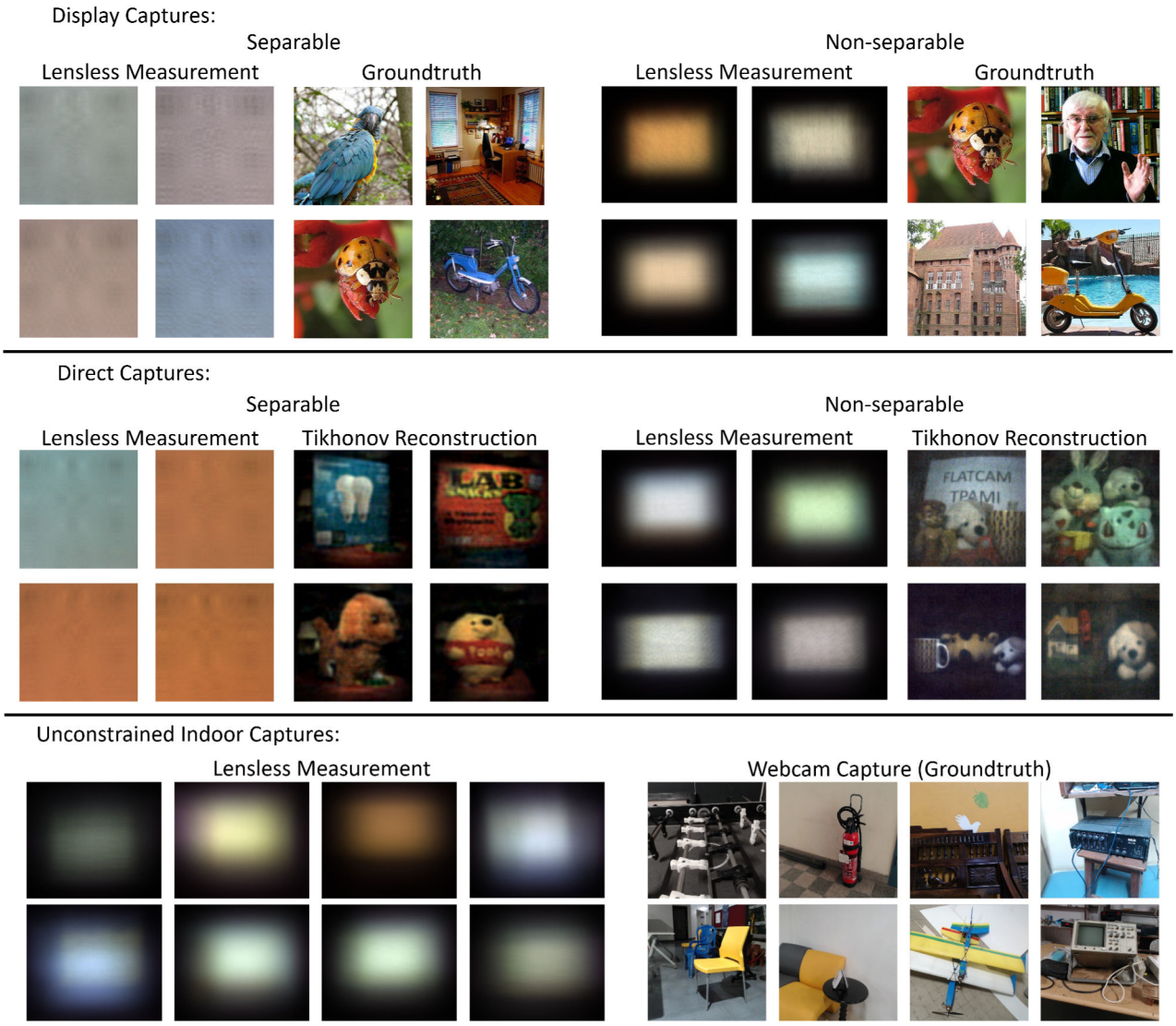

Fig. 3. Samples from our collected datasets.

All our experiments are conducted on real data captured using lensless prototypes. We collect Display Captured Dataset using both separable and non-separable prototypes to train FlatNet-sep and FlatNet-gen, respectively. We also collect Direct Captured Dataset by placing objects in front of the lensless cameras under controlled illumination. Finally, to improve the robustness of FlatNet, we collect a dataset of Unconstrained Indoor Scenes using PhlatCam and Webcam pairs.

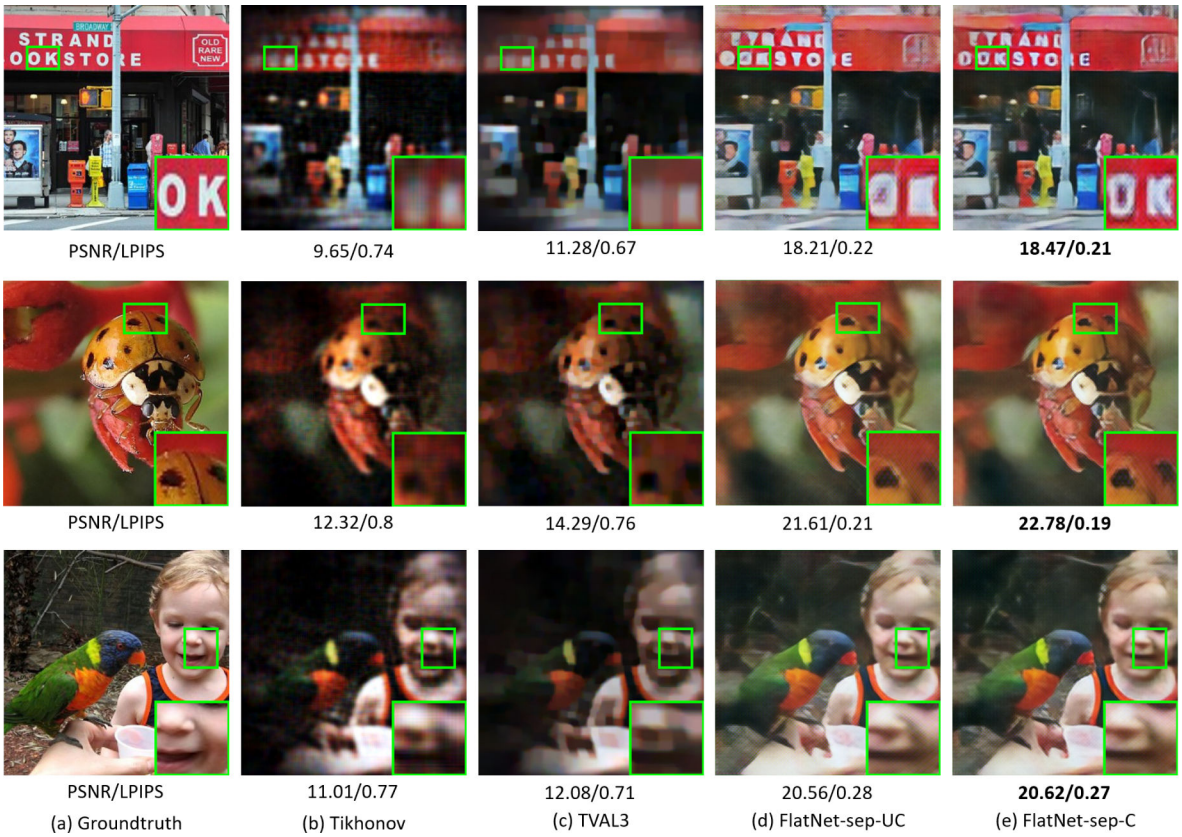

Fig. 4. Display Captured Reconstructions for FlatCam.

Ground truth images are shown in a). Finer details like the text in the first image and spots on the insect in the second image are lost in b) Tikhonov regularized and c) TVAL3 reconstruction. Finer details are better preserved in FlatNet-sep for both d) uncalibrated and e) calibrated initializations.

Fig. 5. Direct Captured Reconstructions for FlatCam:
a) Details in the border and darker regions are lost in the Tikhonov regularized reconstructions. b) TVAL3 reconstructs the border but is unable to restore the sharpness. The proposed end-to-end models for both c) random and d) transpose initializations produce the best reconstructions. These methods are robust to noise and does not contain any regularization parameters.

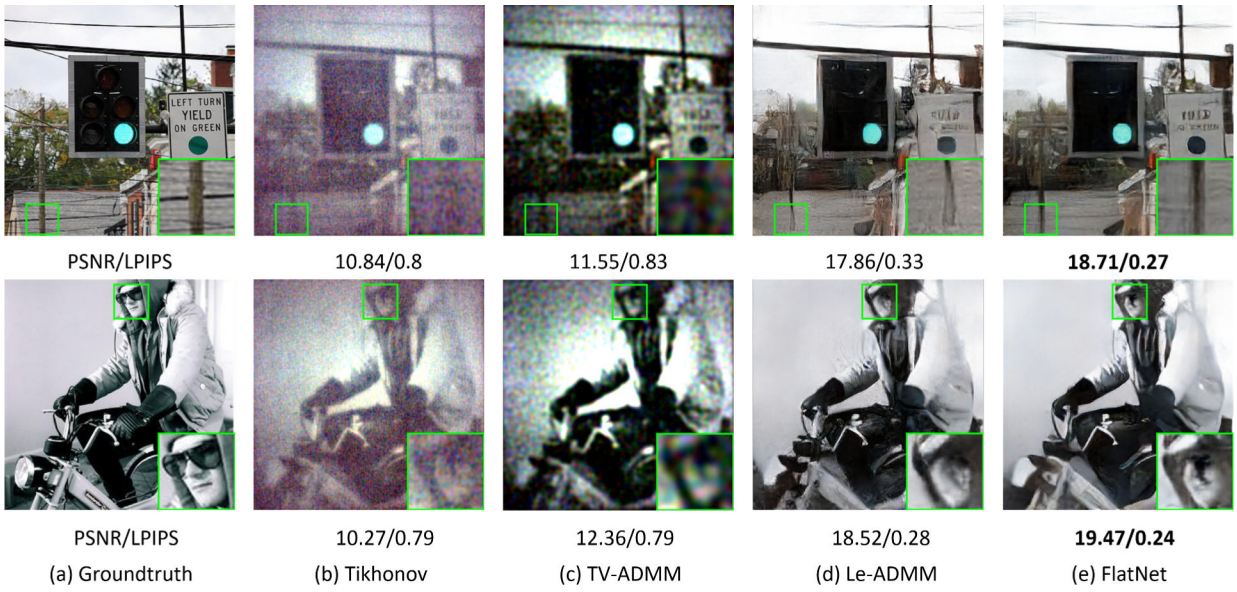

Fig. 6. Display Captured Reconstructions for PhlatCam.

While the learning based methods clearly outperform traditional methods like Tikhonov and TV-based ADMM, FlatNet-gen has superior performance in terms of reconstructing finer details.

Fig. 7. Direct Captured Reconstructions for PhlatCam.

FlatNet-gen has fewer artifacts while Le-ADMM suffers from blurry reconstructions and hallucinated artifacts.

Fig. 8. Comparison of FlatNet with Tikh+U-Net.

Top row shows the comparison of FlatNet-sep with Tikh+U-Net while the bottom row shows the comparison of FlatNet-gen with Tikh+U-Net. FlatNet provides sharper and more photorealistic reconstructions compared to Tikh+U-Net for both separable and non-separable models.

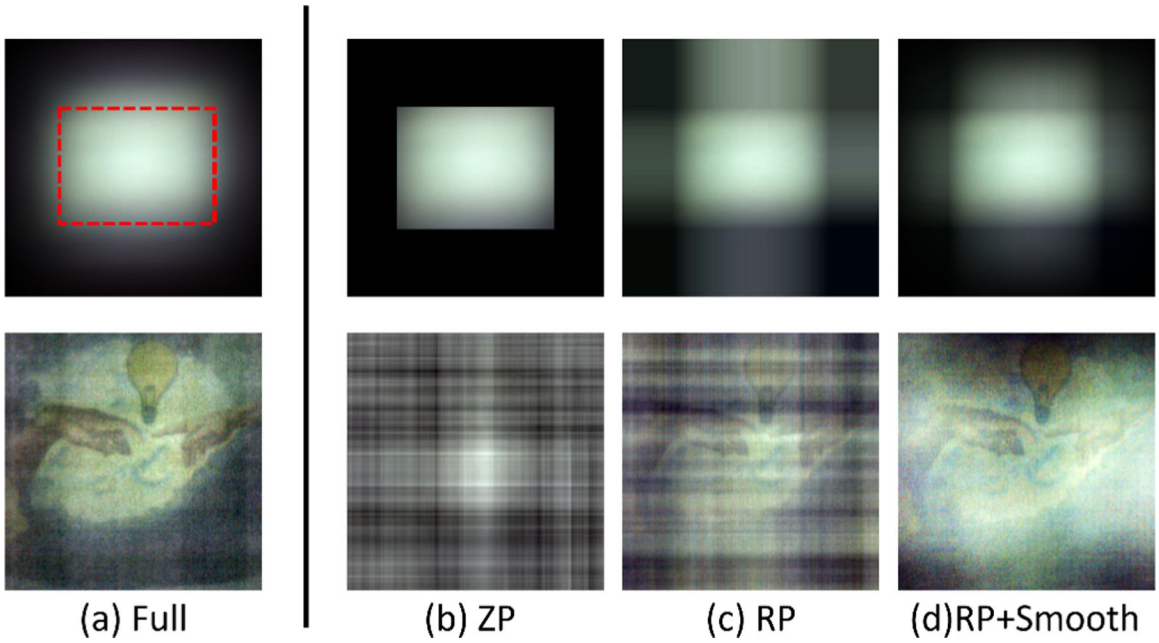

Fig. 9. Effect of padding on Wiener deconvolution for cropped measurement.

Top row shows the measurement while the bottom row shows the corresponding Wiener reconstruction. (a) Full measurement. Red box indicates the cropped out region. (b) Zero padded measurement and the corresponding reconstruction. (c) Replicate padded measurement and the corresponding reconstruction. (d) Smoothened replicate padded measurement along with the corresponding reconstruction. Line artifacts are significantly reduced in (d) which is used in this work.

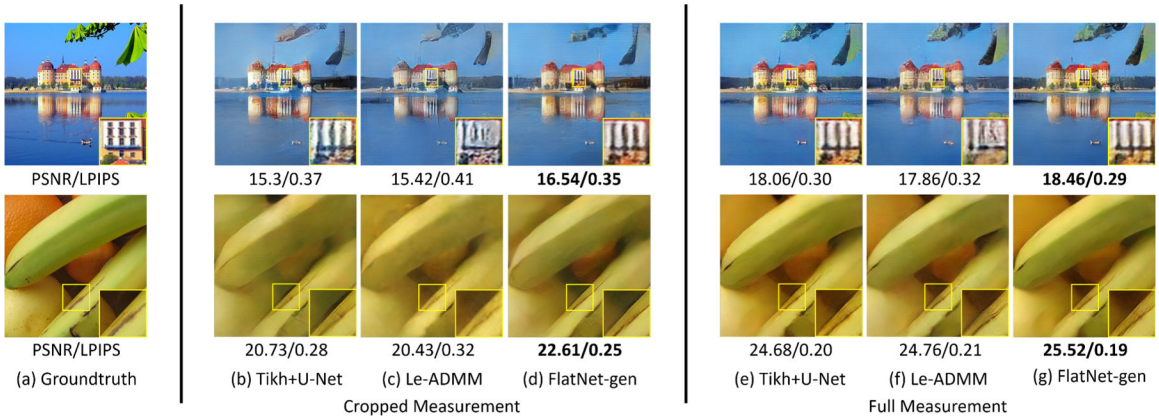

Fig. 10. Display Captured Reconstructions for cropped PhlatCam measurements.

The difference observed in the performance of FlatNet for cropped and full measurements is small. This difference is, however, large for both Le-ADMM and Tikh+U-Net.

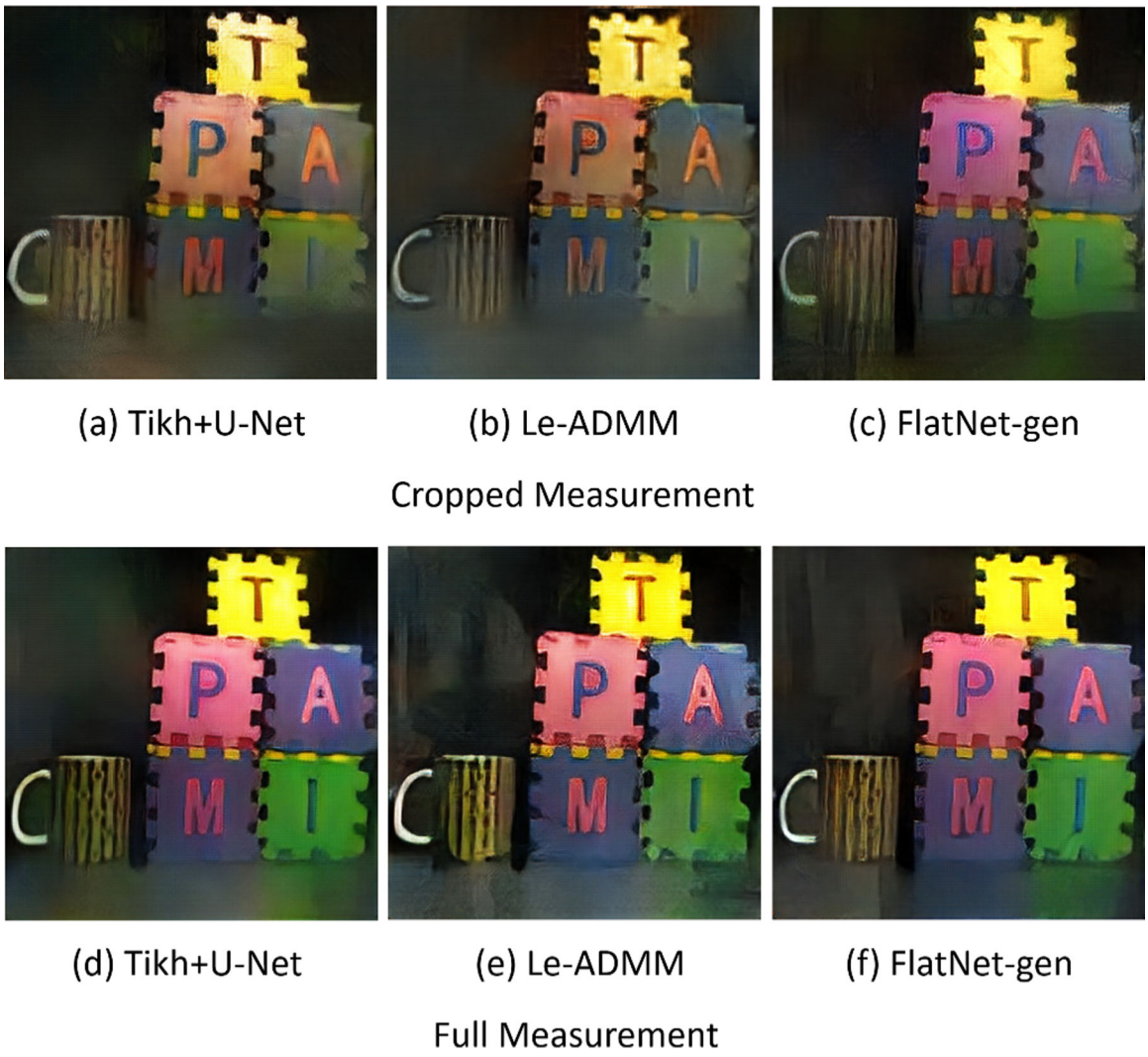

Fig. 11. Direct Captured Reconstructions for cropped PhlatCam measurements.
We can see FlatNet-gen performs reasonably well while both Le-ADMM and Tikh+U-Net breakdown. This can be observed through the colour of the letters and hazy appearance especially around the borders in Tikh+U-Net and Le-ADMM.

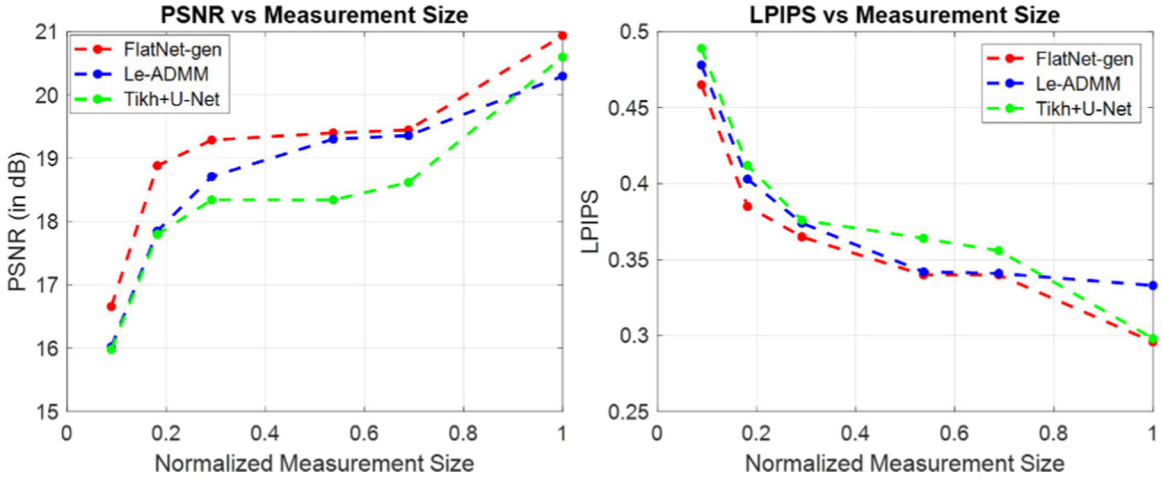

Fig. 12. Performance of learning based techniques for various amount of crops.

We plot the PSNR and LPIPS of FlatNet-gen, LeADMM and Tikh+U-Net under various measurement sizes normalized with respect to full measurement size. We can see FlatNet-gen consistently outperforms other learning based methods for all crop sizes.

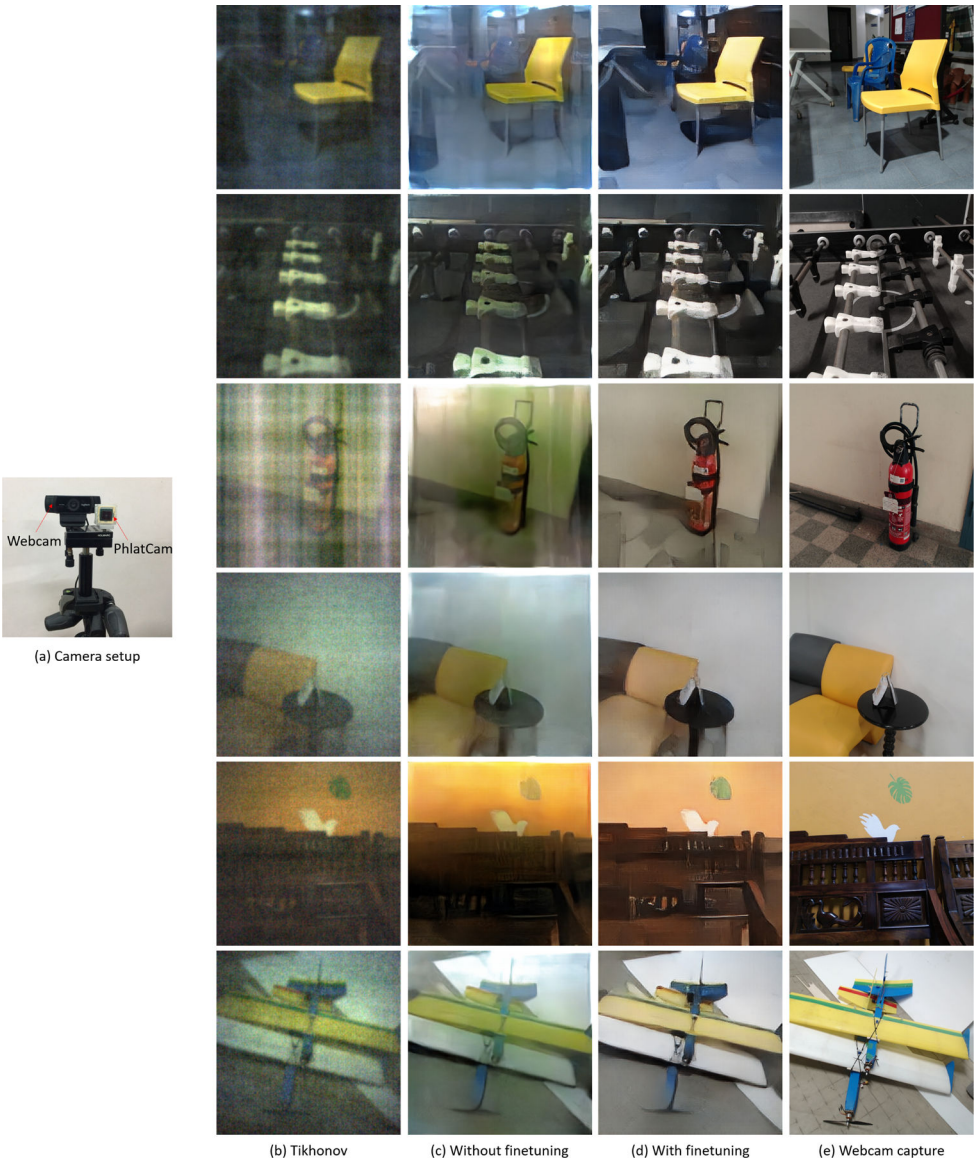

Fig. 13. Photorealistic reconstruction for unconstrained indoor scenes.

(a) The PhlatCam-Webcam setup to capture the dataset for finetuning FlatNet-gen. (b) Tikhonov reconstruction. (c) Reconstructions from FlatNet-gen trained just on display captured data. (d) Reconstructions using FlatNet-gen finetuned on unconstrained indoor captures. (e) Webcam image for reference. Finetuning makes the reconstructions more realistic.

TABLE 1

Average Metrics on Display Captured FlatCam measurements.

FlatNet-sep with transpose initialization (FlatNet-sep-C) gives the best result. Comparable performance of FlatNet-sep-UC indicates that our approach can be used for situations where careful calibration isn't possible.

Method	PSNR (in dB)	SSIM	LPIPS	Inference Time (in sec)
Tikhonov	10.95	0.33	0.795	0.03
TVAL3	11.81	0.36	0.752	45.28
FlatNet-sep-UC	19.06	0.62	0.274	0.006
FlatNet-sep-C	**19.62**	**0.64**	**0.256**	**0.006**

TABLE 2

Average Metrics on Display Captured PhlatCam measurements.

FlatNet-gen produces higher quality results without compromising on the inference time for both the real PSF case (FlatNet-gen-C) and the simulated PSF case (FlatNet-gen-UC). Le-ADMM shows larger difference in quality between the real and simulated PSF cases owing to its stronger dependence on the PSF.

Method	PSNR (in dB)	SSIM	LPIPS	Inference Time (in sec)
Tikhonov	12.67	0.25	0.758	0.03
TV-ADMM	13.51	0.26	0.755	180
Le-ADMM-UC	18.35	0.49	0.407	0.08
Le-ADMM-C	20.29	0.51	0.333	0.08
FlatNet-gen-UC	20.53	0.54	0.318	0.03
FlatNet-gen-C	**20.94**	**0.55**	**0.296**	**0.03**

TABLE 3

Memory and FLOP comparison.

Comparison of memory consumption and FLOPs for five unrolled iterations of the ADMM block in Le-ADMM (full and 4X downsampled versions) and the trainable inversion stage of our proposed FlatNet-gen.

Method	Memory (in MB)	Computation (in MFLOP)
Le-ADMM-Full	6300	1290
Le-ADMM-Downsampled	1000	65
FlatNet-gen	**990**	**53**

TABLE 4

Comparison of FlatNet with Tikh+U-Net.

The top half compares FlatNet-sep with Tikh+U-Net for separable lensless model while the bottom half compares FlatNet-gen with the corresponding Tikh+U-Net. FlatNet outperforms Tikh+U-Net for both separable and non-separable models because it learns an end-to-end mapping.

Methods	PSNR (in dB)	LPIPS
Separable Model		
Tikh+U-Net	18.90	0.322
FlatNet	**19.62**	**0.256**
Non-separable Model		
Tikh+U-Net	20.60	0.298
FlatNet	**20.94**	**0.296**

TABLE 5

Average Metrics on cropped Display Captured PhlatCam measurements.

FlatNet-gen performs consistently better than other learned approaches for both real (FlatNet-gen-C) and simulated PSF case(FlatNet-gen-UC). It should be noted that FlatNet-gen-UC performs as good as Le-ADMM based on real PSF.

Method	PSNR(in dB)	SSIM	LPIPS
Tikh+U-Net-UC	17.53	0.45	0.438
Tikh+U-Net-C	18.34	0.48	0.376
Le-ADMM-UC	17.94	0.45	0.410
Le-ADMM-C	18.72	0.48	0.371
FlatNet-gen-UC	18.72	0.48	0.375
FlatNet-gen-C	**19.29**	**0.50**	**0.365**