FlatNet: Photorealisitc Scene Recovery for Mask Based Lensless Imaging

A Project Report

submitted by

VARUN SUNDAR

in partial fulfilment of the requirements
for the award of the degree of

BACHELOR OF TECHNOLOGY

DEPARTMENT OF ELECTRICAL ENGINEERING
INDIAN INSTITUTE OF TECHNOLOGY MADRAS.

June 2020

THESIS CERTIFICATE

This is to certify that the thesis titled **FlatNet: Photorealisitc Scene Recovery for Mask Based Lensless Imaging**, submitted by **Varun Sundar**, to the Indian Institute of Technology, Madras, for the award of the degree of **Bachelors of Technology**, is a bona fide record of the research work done by him under our supervision. The contents of this thesis, in full or in parts, have not been submitted to any other Institute or University for the award of any degree or diploma.

Prof. Kaushik Mitra
Research Guide
Assistant Professor
Department of Electrical Engineering
IIT-Madras, 600 036

Place: Chennai

Date: June 17, 2020

ACKNOWLEDGEMENTS

FlatCam (and it's variants) has provided an exciting, intriguing and sometimes even confounding project for me to work on over the past year. In many ways, it has led me to appreciate the intricacies in lens based design, various tradeoffs in imaging systems and promising avenues in thin multiplexing cameras. This work has also greatly influenced my interests in the fields of Computer Vision and Computational Photography, which I have chosen to further pursue in graduate school.

I would like to thank my advisor Prof. Kaushik Mitra for providing me this opportunity and for his guidance throughout this journey. I would like to express my sincere gratitude to him for taking me through the ups and downs of research, and also advising me in crucial junctions such as applying to PhD programmes. I am also particularly grateful to Salman Siddique of the *Computational Imaging Group* for the mentorship, brain storming, and the numerous (*last minute!*) experiments and analysis we conducted. This thesis and the work we accomplished would not have been possible without him.

I would also like to acknowledge the support of our collaborators at Rice University- Dr. Vivek Boominathan and Prof. Ashok Veeraragahavan without whom this project would not be possible. I would also like to acknowledge the role IIT Madras (or 'insti') has played in providing a conducive environment for holistic growth. The lifelong friends I have made here were instrumental in making this a memorable journey. Finally, I would acknowledge the support of my family, who have been pillars of strength throughout my growth as an individual.

ABSTRACT

KEYWORDS: Lensless imaging, Image reconstruction, Computational Imaging.

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 *PhlatCam* which uses a 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.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABBREVIATIONS

CRA	Chief Ray Angle.
FoV	Field of View.
FLOP	Floating Point Operation.
IITM	Indian Institute of Technology, Madras.
LPIPS	Learned Perceptual Image Patch Similarity metric.
MSE	Mean Squared Error.
PSF	Point Spread Function.
PSNR	Peak Signal to Noise Ratio.
SSIM	Structure Similarity Index.
-C	Calibrated model.
-UC	Uncalibrated model, utilises mask profile to initialise model instead.
-Gen	General (Non-Separable) FlatNet model.
-Sep	Separable FlatNet model.

NOTATION

$\mathcal{F}$	Discrete Fourier Transform.
$\mathcal{F}^{-1}$	Inverse Discrete Fourier Transform.
Φ	Multiplexing matrix, which linearly transforms the scene into measurements. For a scene of $N \times N$ dimensions and measurements of $M \times M$ dimensions, Φ is a $M^2 \times N^2$ matrix.
$\Phi_{L,R}$	Left and Right Φ matrices when Φ is separable. Each has MN elements.
P	Point Spread Function, when Φ is a circulant matrix.
X	2D Scene of dimension $N \times N$.
x	Lexicographic flattened 1D scene of dimension N^2.
Y	2D Measurement of dimension $M \times M$.
y	Lexicographic flattened 1D dimension M^2.
$\mathbf{C}(.)$	Crop operator.
$\mathbf{pad}(.)$	Pad operator.
$*$	2D Convolution, unless indicated otherwise.
$\otimes$	Hadamard Product (element-wise), unless indicated otherwise.

CHAPTER 1

INTRODUCTION

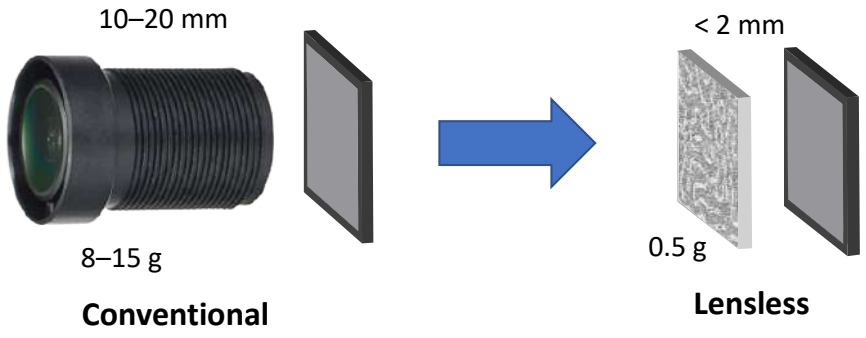

Figure 1.1: **Lensless cameras greatly facilitate miniaturisation.** Such miniaturisation is crucial for applications in emerging fields such as IoT, Medical Imaging and Surviellance. Mask based lensless cameras can be fabricated to a given thickness and are free from common lens associated limitations. In comparison solutions which use either a lens or a lens like focussing element are always limited by their focal length and depth of field constraints. See Boominathan *et al.* (2020) for more details.

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) 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 a mask (an amplitude (Asif *et al.* (2017)) or a phase mask (Antipa *et al.* (2018); Boominathan *et al.* (2020))), 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. This is illustrated in figure 1.1, which demonstrates over an order of magnitude reduction in size and weight of lensless cameras in comparison to conventional ones. 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 (Tremblay *et al.* (2007)). 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 (Boominathan *et al.* (2016)).

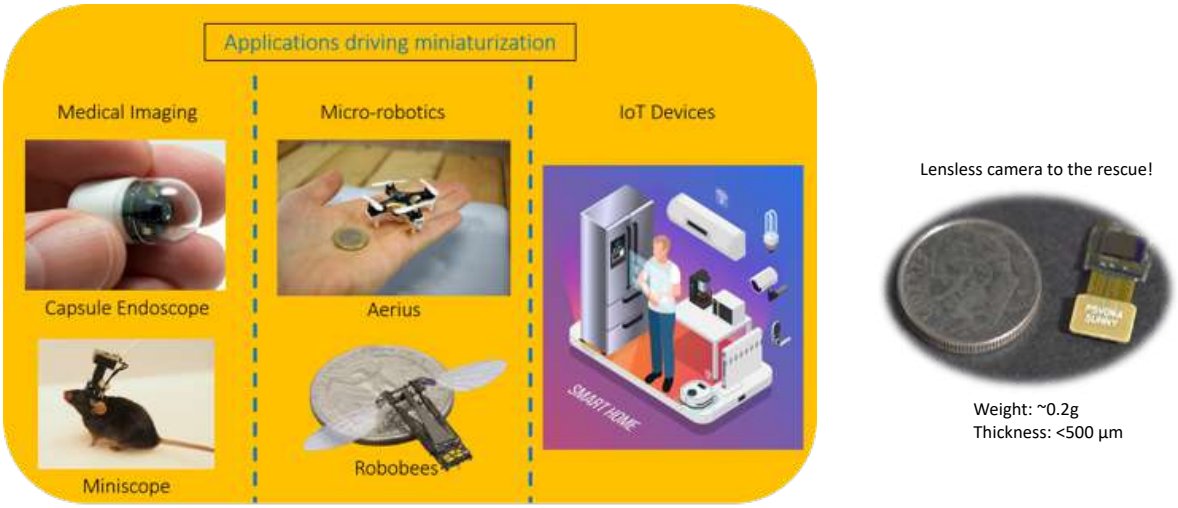

Figure 1.2: **A few applications of lensless cameras.** These applications are mainly driven by the utility of camera miniaturisation and cameras as an inferential device. We expand more on this idea of cameras as an inferential device under our concluding notes in chapter 6.

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. 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 1.3. Therefore, lensless cameras need robust and efficient algorithms to overcome these challenges.

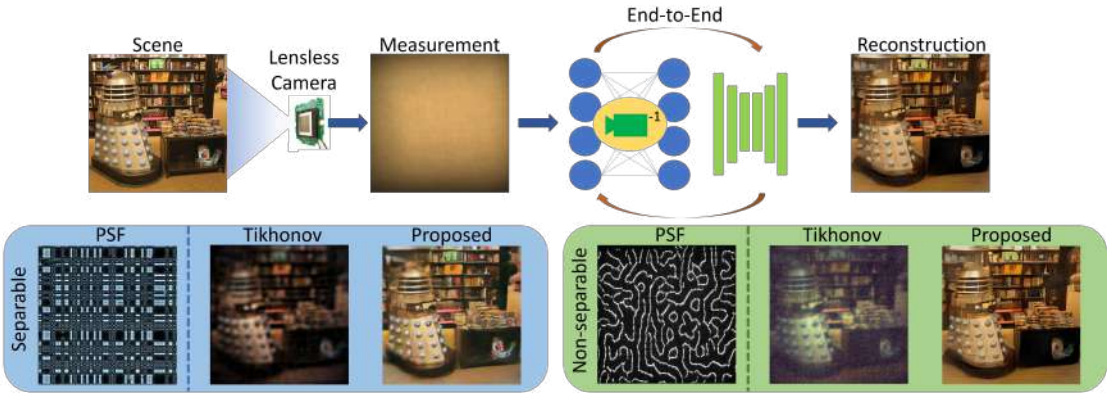

Figure 1.3: **Lensless cameras require computation to recover the true scene from measurements.** In this work we propose a deep learning based lensless reconstruction algorithm for both separable (Asif *et al.* (2017)) and non-separable mask (Boominathan *et al.* (2020)) based lensless cameras that produce photorealistic reconstructions for real and challenging scenarios.

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 Boominathan *et al.* (2020) that separable lensless mask based lensless cameras have inferior characteristics as compared to their existing non-separable counterparts. Khan *et al.* (2019) had demonstrated such a model for separable lensless systems. 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 (Asif *et al.* (2017))

and the non-separable mask based lensless camera called PhlatCam Boominathan *et al.* (2020). To summarize, the key contributions in this work are:

- We propose an efficient implementation for the learnable intermediate stage of non-separable or general lensless model. Khan *et al.* (2019) 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 (Khan *et al.* (2019)), we collect corresponding datasets for the non-separable mask PhlatCam (Boominathan *et al.* (2020)).
- 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.

CHAPTER 2

BACKGROUND AND PROBLEM SETUP

2.1 Masks as a Linear Multiplexing Element

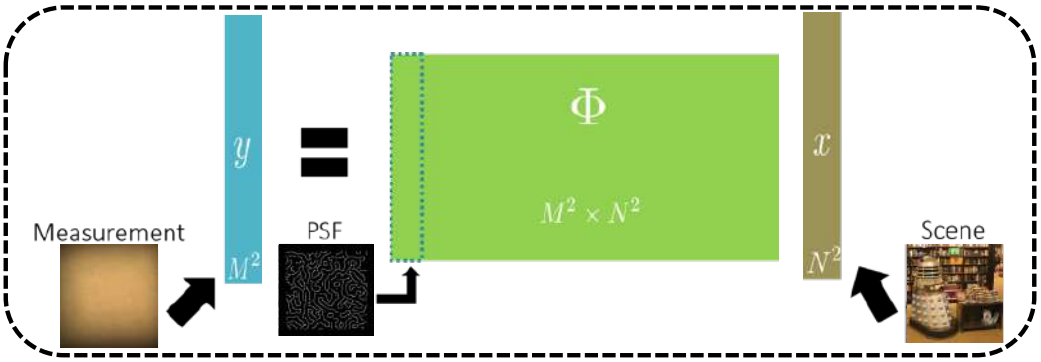

Figure 2.1: **Image formation in lensless cameras interpreted via linear systems.** The Φ matrix in the above system contains M^2N^2 elements, where the scene is M and the measurement is N. For a megapixel scene and a megapixel sensor, this results in an order of 10^12 elements. Inverting this via conventional linear algebra methods is not straightforward.

In this chapter, we develop and motivate the forward model used behind lensless cameras- which can mainly be categorised as separable or non-separable. In this work, we formulate an efficient reconstruction algorithm for *PhlatCam* which uses a non-separable mask. However, for completeness, we shall also cover the separable forward model here.

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:

$$y = \Phi x + n, \tag{2.1}$$

where **y** is the measurement obtained at the sensor vectorized in lexicographic manner, Φ represents the generalized linear transformation, and **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. This aspect is depicted diagrammatically in Figure 2.1. 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 (Asif *et al.* (2017); Adams *et al.* (2017)). Specifically, in Asif *et al.* (2017), the single-separable lensless forward model reduces to:

$$Y = \Phi_L X \Phi_R^T + N, \tag{2.2}$$

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. This model is followed by *FlatCam* (Asif *et al.* (2017)).

By adding an aperture over a non-separable mask, Antipa *et al.* (2018); Boominathan *et al.* (2020) showed that the lensless forward model can be written as a convolutional model:

$$Y = P * X + N, \tag{2.3}$$

where P is the point-spread-function (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 (Boominathan *et al.* (2020)). 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, Antipa *et al.* (2018) uses a cropped convolution model:

$$Y = C(P * X) + N, \tag{2.4}$$

where C is the sensor cropping operation. Such a system described by Equation 2.4 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 *PhlatCam* (Boominathan *et al.* (2020)) 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 (Reeves (2005)).

2.2 Inversion Stage for Separable Models

In this section we briefly cover the inversion layer used by Khan *et al.* (2019) for separable lensless systems.

Given the lensless model described in Equation 2.2, we learn two layers of left and right trainable matrices that act directly on 2-D measurements. This can be mathematically represented as,

$$X_{\text{interm}} = f(W_1 Y W_2), \tag{2.5}$$

where X_{interm} is the output of this stage, f is a pointwise nonlinearity, 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. they have the same dimension as the transpose of the forward matrices. 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, weight matrices (W_1 and W_2) are initialized with the adjoint of the calibration matrices, akin to back-projection. These calibration matrices are approximations of Φ_L and Φ_R in (2.2) physically obtained by the method described in Asif *et al.* (2017). This mode of initialization leads to faster convergence while training.

Uncalibrated initialization. Calibration of FlatCam require careful alignment with display monitor (Asif *et al.* (2017)), 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, Khan *et al.* (2019) also proposes a calibration-free approach by initializing the weight matrices with carefully designed pseudo-random matrices.

2.3 Fourier Optics Analysis

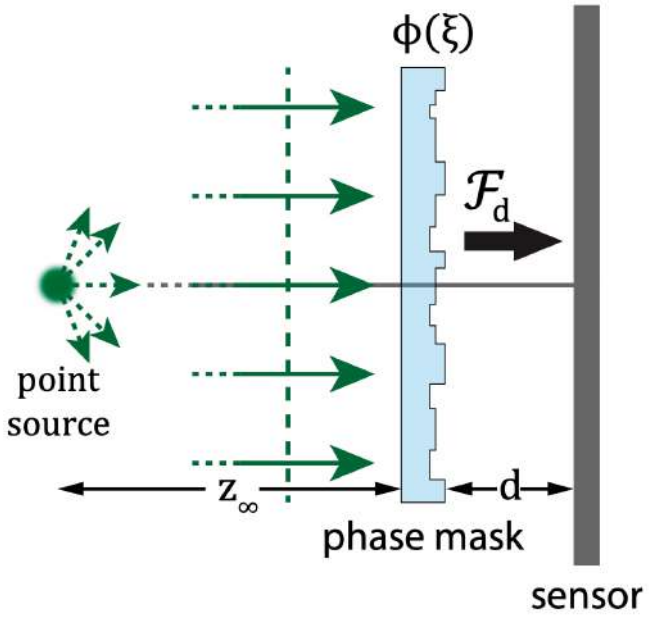

Figure 2.2: **Convolutional model derived via Fourier (Wave) Optics.** A similar treatment may be found in Boominathan *et al.* (2020). Figure courtesy: Boominathan *et al.* (2020).

An alternative method to arrive at the convolutional based (or non-separable) model is to resort to a treatment via Fourier Optics. This method also sheds light on a few properties of the large support PSFs that characterise lensless cameras and the range of validity of these properties.

Consider a phase mask with height profile $\phi(\xi)$. The following analysis can be carried out for a general mask too, by replacing $\exp(j\phi(\xi))$ by any transmittance function $t_l(\xi)$. The former represents the phase delay on account of the thin phase mask. We shall also carry out the Fourier Propagation in 1D for simplicity, and the same results can easily be extended to 2D. For a starting reference on the equations used for the Fresnel Propagation operator and other Fourier Optics relations, one may refer Good-

man (2005).

If we illuminate this phase mask with a coherent collimated light, the intensity pattern $p(x)$ observed is:

$$
\begin{aligned}
p(x) &= |\mathcal{F}_d\big(e^{j\phi(\xi)}\big)|^2 \\
&= |\frac{1}{\sqrt{\lambda d}} \int e^{j\phi(\xi)} \exp[j\frac{\pi}{\lambda d}(x-\xi)^2] d\xi|^2 \\
&= |\int e^{j\phi(\xi)} \exp[j\frac{\pi}{\lambda d}(\xi^2 - 2x\xi)] d\xi|^2 \quad \text{Neglect the constant distance } d
\end{aligned}
$$

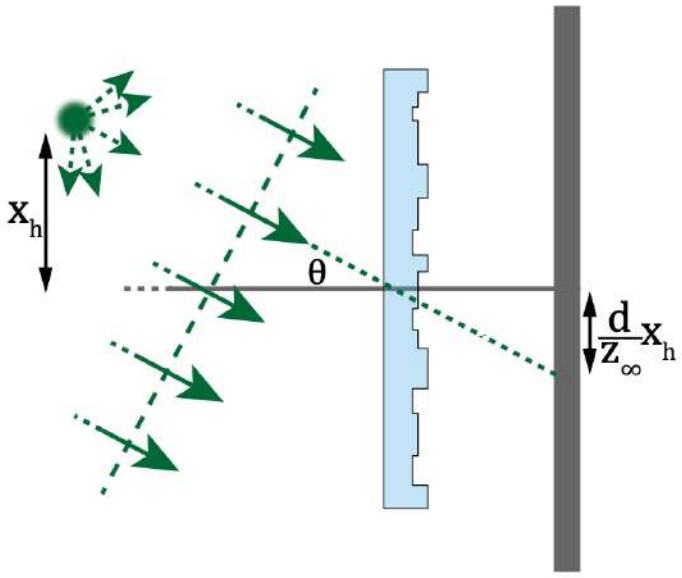

Figure 2.3: **Shift-invariance Property of Lensless PSFs.** We find the illumination pattern for an off-axis, but distant point source. Figure courtesy: Boominathan *et al.* (2020).

The collimated light (or planar waves) can be thought of as originating from an on-axis point source at infinite distance. If this point were instead off axis by some angle θ (as depicted in Figure 2.3), then:

$$
\begin{aligned}
Y_\theta(x) &= |\int e^{j\frac{2\pi \sin\theta}{\lambda}\phi(\xi)} \exp[j\frac{\pi}{\lambda d}(\xi^2 - 2x\xi)]d\xi|^2 \\
&= |\int e^{j\phi(\xi)} \exp[j\frac{\pi}{\lambda d}(\xi^2 - 2(x - d\sin\theta)\xi)]d\xi|^2 \\
&= p(x - d\sin\theta) \\
&\approx p(x - \frac{d}{z_\infty}x_h) \quad \text{Using paraxial approximation.}
\end{aligned}
$$

Hence, under the paraxial approximation, the PSF is shift invariant. In other words, the forward model is convolutional and can be written as:

$$
\begin{aligned}
Y &= P_z * X \quad \text{Where X is the scene, Y is the resultant intensity.} \\
&= P * X \quad \text{Assume scene is far enough.}
\end{aligned}
$$

In the preceding equation, we have also implicitly assumed that the scene consists of incoherent, distant point sources. Hence, the system becomes linear in intensity response (see Chapter 6 in Goodman (2005)). We exploit this fact in Section 4.1, where we reduce the forward model of a large dimensional linearly multiplexed system into a much more tractable convolutional model.

We end this section by examining the depth-dependence of the PSF. For a point source at distance z from the mask, the spherical waves cannot be assumed to be planar. However, under the paraxial approximation, they impart an additional phase factor $e^{j\frac{\pi}{\lambda z}\xi^2}$, with the result:

$$
\begin{aligned}
Y_z(x) &= |\int e^{j\phi(\xi) + j\frac{\pi}{\lambda z}\xi^2} \exp[j\frac{\pi}{\lambda d}(\xi^2 - 2x\xi)]d\xi|^2 \\
&= |\int e^{j\phi(\xi)} \exp[j\frac{\pi}{\lambda}(\frac{1}{d} + \frac{1}{z})(\xi^2 - 2\frac{x}{1 + d/z}\xi)]d\xi|^2 \\
&\approx |\int e^{j\phi(\xi)} \exp[j\frac{\pi}{\lambda d}(\xi^2 - 2\frac{x}{1 + d/z}\xi)]d\xi|^2 \\
&= p(\frac{x}{1 + d/z})
\end{aligned}
$$

The last approximation is valid since the mask sensor distance (which is typically in millimetres for lensless cameras) is much smaller than scene depths. This is depicted in Figure 2.4.

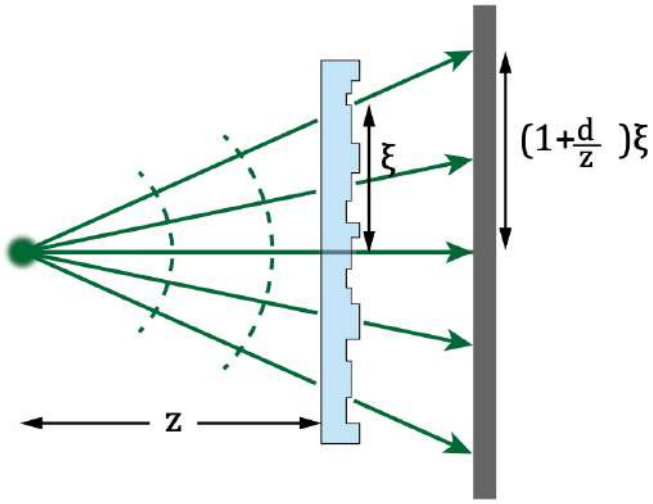

Figure 2.4: **PSF Scaling Property of Lensless Cameras.** As the point moves closer, the PSF scales in size. The depth dependence however saturates beyond a certain distance, leading to a 2D scene approximation. Figure courtesy: Boominathan *et al.* (2020).

This implies that the forward model is actually,

$$Y = \sum_{z} P_z * X$$

But for large distances (an order more than d), $p(\frac{x}{1+d/z}) \approx p(x)$ and we can neglect the PSF scaling. Thus, for most scene depths, lensless cameras perceive a 2D scene. PSF scaling is exploited in applications such as microscopy (Adams *et al.* (2017); Antipa *et al.* (2018)) where distances are much smaller.

CHAPTER 3

RELATED WORK

3.1 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 (Dicke (1968); Caroli *et al.* (1987)), 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 (Kim *et al.* (2017)) or after being modulated by a mask element. Types of masks that have been used include phase gratings (Stork and Gill (2013)), random diffusers (Antipa *et al.* (2018)), designed phase-masks (Boominathan *et al.* (2020)), amplitude masks (Shimano *et al.* (2018); Asif *et al.* (2017)), compressive samplers (Huang *et al.* (2013); Satat *et al.* (2017)) and spatial light modulators (Chi and George (2011); DeWeert and Farm (2015)). 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.

3.2 Image Reconstruction for Computational Imaging

Image reconstruction is a core aspect of most computational imaging problems (Duarte *et al.* (2008); Antipa *et al.* (2019); Asif *et al.* (2017); Antipa *et al.* (2018); Boominathan *et al.* (2020)). 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 (Li *et al.* (2013); Boominathan *et al.* (2020); Antipa *et al.* (2018)), wavelet or frequency domain sparsity (Reddy *et al.* (2011)), 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 Dave *et al.* (2018, 2017); Rick Chang *et al.* (2017) 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 (Kulkarni *et al.* (2016); Mousavi *et al.* (2015); Zhang and Ghanem (2018)), Fourier Ptychography (Boominathan *et al.* (2018)), lensless recovery (Monakhova *et al.* (2019)) fall under this category. Among these feed-forward networks, Monakhova *et al.* (2019); Zhang and Ghanem (2018) 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 Khan *et al.* (2019) 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.

CHAPTER 4

PROPOSED METHOD

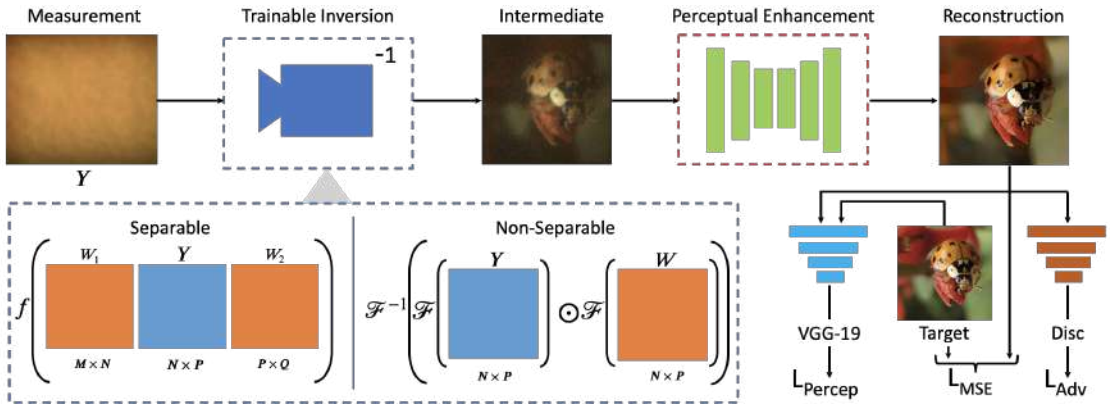

Figure 4.1: **Overall architecture of the FlatNet.** The lensless camera measurement is first mapped into an intermediate image space using a trainable camera inversion layer. This stage is implemented separately for the separable and the non-separable case. A U-Net (Ronneberger *et al.* (2015)) then enhances the perceptual quality of the intermediate reconstruction. We use a weighted combination of three losses in training our network: a perceptual loss (Johnson *et al.* (2016)) using a VGG16 network (Simonyan and Zisserman (2014)), mean-square error (MSE), and adversarial loss using a discriminator neural network (Goodfellow *et al.* (2014)). The separable case was proposed in prior work (Khan *et al.* (2019)), but we include it in this figure for completeness and comparison to our implementation.

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.

4.1 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. The separable model was formulated in Khan *et al.* (2019), and we have restated this in Section 2.2. Our implementation, which extends the idea in Khan *et al.* (2019) to non-separable models is elaborated here.

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 2.1 and 2.3.

Following the observation that the forward model is purely convolutional for an appropriate sensor dimension i.e. the forward operation is described by Equation 2.3, 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 2.3 is also circulant and can be diagonalized by Fourier transform.

Mathematically, this operation is given as,

$$X_{\text{interm}} = \mathcal{F}^{-1}(\mathcal{F}(W) \odot \mathcal{F}(Y)), \tag{4.1}$$

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 $N \times M$. The convolutional model

of Equation 2.3 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 2.4. Of course, we cannot accurately represent the inverse of the system described by Equation 2.4 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 5.4.2, that with a small modification to the trainable inversion stage described in Equation 4.1, 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 Boominathan *et al.* (2020). This initialization scheme is particularly useful for cases where the PSF exceeds the sensor size (see Section 5.4.2).

4.2 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 (Ronneberger *et al.* (2015)) to map the intermediate reconstruction to the final perceptually enhanced image. We keep the kernel size fixed at 3x3 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 Gu *et al.* (2019), we find it useful to employ Pixel-Shuffle (Shi *et al.* (2016)) 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.

4.3 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 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 (Ramachandran *et al.* (2017)) in our discriminator. Same discriminator architecture was used for both FlatNet-sep and FlatNet-gen.

4.4 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:

$$\mathcal{L}_{\text{MSE}} = ||I_{\text{true}} - I_{\text{est}}||_2^2. \tag{4.2}$$

Perceptual loss: To measure the semantic difference between the estimated output and the ground truth, we use the perceptual loss introduced in Johnson *et al.* (2016). We use a pre-trained VGG-16 (Simonyan and Zisserman (2014)) 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,

$$\mathcal{L}_{\text{percept}} = ||\phi_{22}(I_{\text{true}}) - \phi_{22}(I_{\text{est}})||_2^2 + ||\phi_{43}(I_{\text{true}}) - \phi_{43}(I_{\text{est}})||_2^2. \tag{4.3}$$

Adversarial loss: Adversarial loss (Ledig *et al.* (2017); Goodfellow *et al.* (2014)) 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 4.3, this loss is given as,

$$\mathcal{L}_{\text{adv}} = -\log(D(I_{\text{est}})). \tag{4.4}$$

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,

$$\mathcal{L} = \lambda_1 \mathcal{L}_{\text{MSE}} + \lambda_2 \mathcal{L}_{\text{percept}} + \lambda_3 \mathcal{L}_{\text{adv}}. \tag{4.5}$$

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,

$$\mathcal{L}_{\text{disc}} = -\log(D(I_{\text{true}})) - \log(1 - D(I_{\text{est}})). \tag{4.6}$$

Contextual Loss: For finetuning FlatNet-gen on unaligned PhlatCam and webcam pairs (described in Section 5.4.4), we use only contextual loss as proposed in Mechrez *et al.* (2018). 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 $q = \arg\min_q \mathbb{D}(p, q_j)_{j=1}^N$ 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:

$$\mathcal{L}_{\text{contextual}} = \frac{1}{N} \sum_{i=1}^{N} min_{j \in [N]} \mathbb{D}(p_i, q_j) \tag{4.7}$$

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

CHAPTER 5

EXPERIMENTS AND ANALYSIS

In this chapter, 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 (proposed in Khan *et al.* (2019)) 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.

5.1 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 a similar dataset collection procedure as Khan *et al.* (2019) for PhlatCam (Boominathan *et al.* (2020)). For our work, we use a subset of ILSVRC 2012 (Russakovsky *et al.* (2015)). 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.

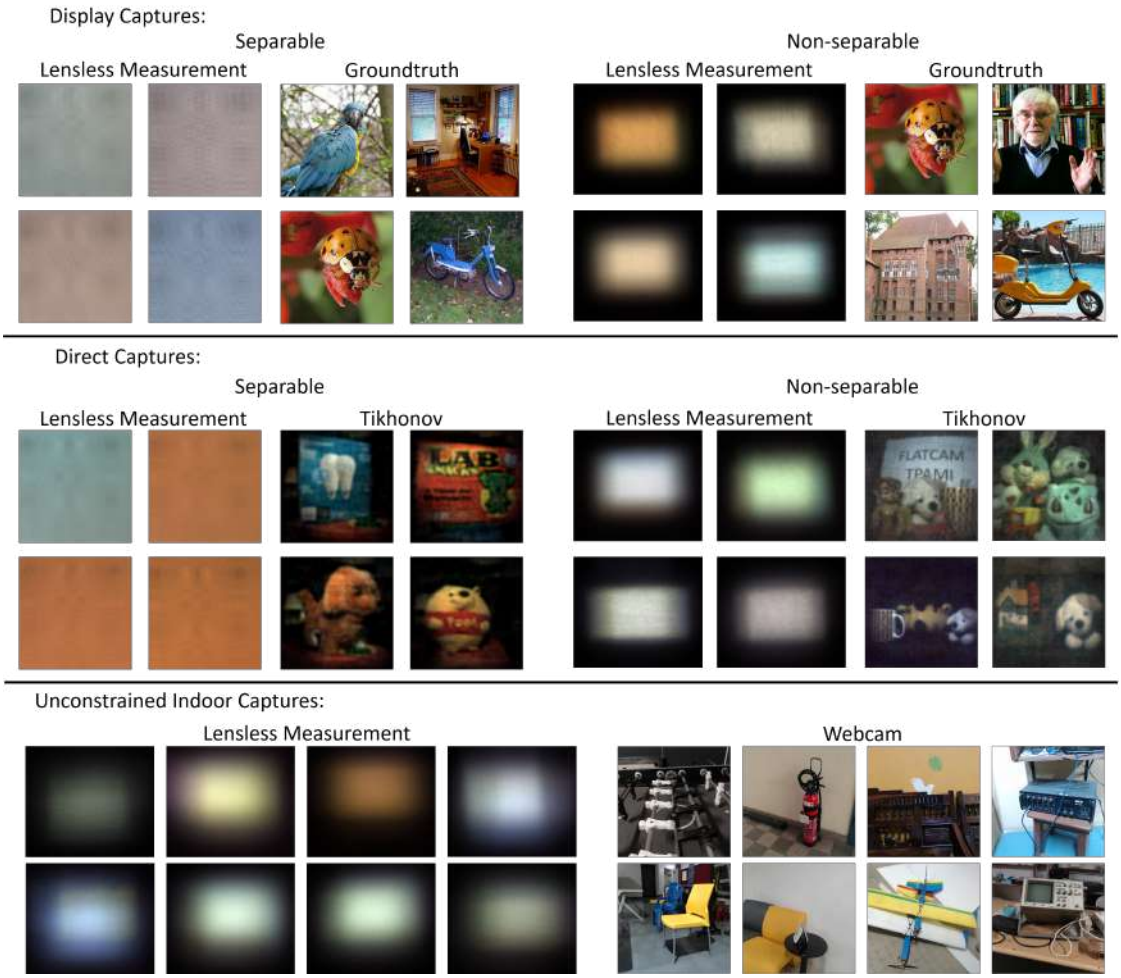

Figure 5.1: **Samples from our collected datasets**. All our experiments are conducted on real data captured using PhlatCam (Boominathan *et al.* (2020)). We collect Display Captured Dataset using PhlatCam, a non-separable prototype, to train FlatNet-gen. We also collect Direct Captured Dataset by placing objects in front of the lensless camera under controlled illumination. Finally, to improve the robustness of FlatNet, we collect a dataset of Unconstrained Indoor Scenes using PhlatCam and Webcam pairs. For comparison, we have also shown samples used by Khan *et al.* (2019) in training FlatNet-sep, which is already publicly available at https://siddiquesalman.github.io/flatcam_iccv.html.

To test the FlatNet on real scenes, we also capture measurements of objects placed directly in front of the camera. 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 5.14. 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 5.1.

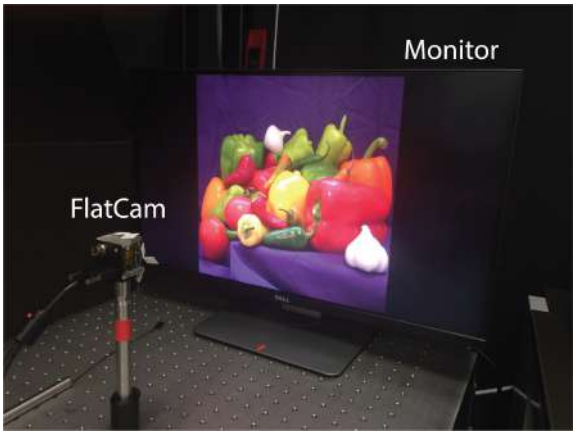

Figure 5.2: **Display Capture Setup**. The monitor is placed such that the entire field of view (FoV) of the camera is occupied. This is aligned by using a checkerboard pattern and aligning the monitor with the Tikhonov reconstruction. Both cameras are adjusted for white balance.

5.2 Implementation details

5.2.1 Camera Prototype

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 of 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 $1280 \times 1408 \times 4$. This size of measurement was selected to account for the non-zero nature of the bright points beyond the Chief Ray Angle (CRA) of the sensor. For the small sensor experiments of Section 5.4.2, we use measurements of dimension $608 \times 864 \times 4$.

We set the loss coefficients in Equation 4.5 as: λ_1 to be 1, λ_2 to be 1.2 and λ_3 to be 0.6. The Adam (Kingma and Ba (2014)) optimizer was used for all models. We started with a learning rate of 10^{-4} and gradually reduced it by half every 5000 iterations.

5.2.2 Display Capture Setup

To capture a display-captured image using PhlatCam (Boominathan *et al.* (2020)), the image is resized so as to occupy the biggest central square on a 24-inch monitor using

Figure 5.3: **Direct Capture Setup**. We collect measurements in a dark room, with the only source of illumination being a controllable light source. These measurements are more similar in distribution to monitor acquired images (display capture). Indoor measurements in the wild can however differ significantly from this.

bicubic interpolation. The monitor was placed at appropriate distance so that the image occupied the field of view of the cameras. For PhlatCam this was 8 inches. This setup is fixed for all image captures such that the alignment of the monitor pixels to the camera pixels is uniform throughout both training and test. The white balance setting was estimated once before the capture began by capturing a demo picture. The exposure was set at 10000 microseconds. Figure 5.2 shows the setup for FlatCam capture. The setup for PhlatCam is similar.

5.2.3 Direct Capture Setup

We follow the same white balance offsetting procedure for both cameras as mentioned in the preceding paragraph. As seen in Figure 5.3, we make use of a dark room with just a single light used to control illumination. This makes direct capture quite similar to monitor based acquisition (display capture). Indoor measurements in the wild, however, can differ from display capture. For this reason, as detailed in Sections 5.1 and 5.4.4, we collect an unconstrained capture dataset.

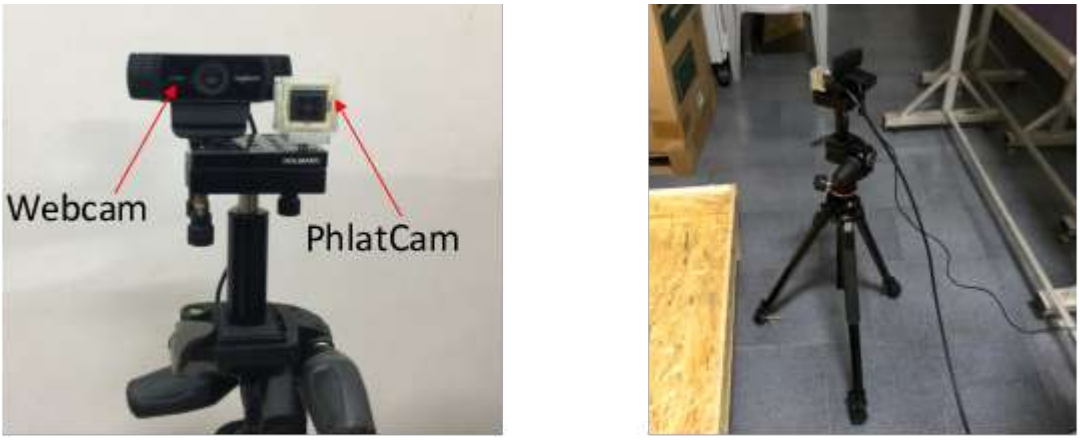

Figure 5.4: **Unconstrained Capture Setup**. We capture Webcam-PhlatCam pairs to finetune FlatNet on indoor scenes with varying, uncontrolled lighting. *Left.* Arrangement used to mount PhlatCam and Webcam. *Right.* Sample measurement collection of an indoor scene.

5.2.4 Unconstrained Capture Setup

As we elaborate in Section 5.4.4, real measurements may differ from display capture due to issues like stray light. To offset this, we collect Webcam-PhlatCam pairs and finetune FlatNet. The setup used for collecting the measurements and a sample scene collection can be seen in Figure 5.4.

5.3 Comparison with other approaches

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 (Antipa *et al.* (2018)).
- **Learning based approaches.** For learning based approach, we use the unrolled deep network described in Monakhova *et al.* (2019). However, for fairness, we use the five stage unrolled ADMM followed by our perceptual enhancement stage.

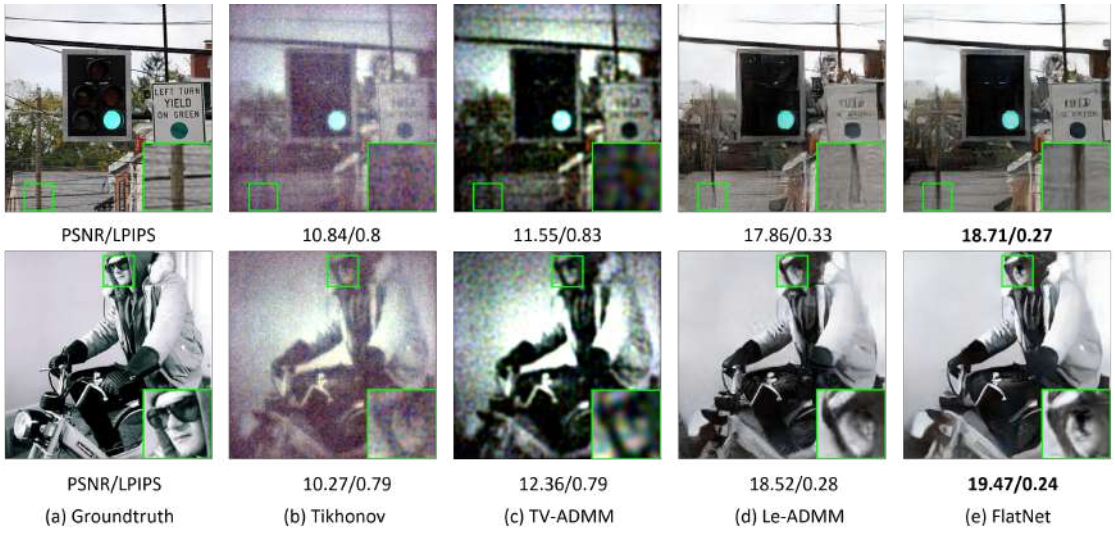

Figure 5.5: **Display Captured Reconstructions for PhlatCam**. While the learning based methods clearly outperform traditional methods like Tikhonov and TV-based ADMM, FlatNet-gen has superior performance in terms of reconstructing finer details.

5.3.1 Qualitative discussion.

Figure 5.5 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 5.6. 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 Monakhova *et al.* (2019)) 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.

Figure 5.6: **Direct Captured Reconstructions for PhlatCam**. FlatNet-gen has fewer artifacts while Le-ADMM suffers from blurry reconstructions and hallucinated artifacts.

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 5.1: **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.

We also provide comparison for FlatNet-gen initialized with uncalibrated PSF in Section 5.4.3. We call this model FlatNet-gen-UC.

5.3.2 Quantitative discussion.

The quantitative results are provided in Table 5.1. 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

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

Table 5.2: **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. We show here for 5 unrolled iterations of ADMM. Despite using 4x downsampled measurements, Le-ADMM suffers from higher computational and memory requirements in comparsion to the trainable inversion stage of FlatNet-gen.

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 5.2. 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 par-

ticular 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.

5.4 Further analysis

5.4.1 Effect of learning the inversion stage

(a) Tikhonov (b) Tikh+U-Net (c) FlatNet

Figure 5.7: **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.

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 4.5. We call this method Tikh+U-Net. We compare the performance of FlatNet-gen with its corresponding Tikh+U-Net in Figure 5.7. FlatNet-gen provides sharper reconstructions over Tikh+U-Net. Tikh+U-Net suffers from blurrier reconstructions with amplified artefacts.

Table 5.3 provides a quantitative flavor to the above analysis. We can see that FlatNet outperforms Tikh+U-Net for both in terms of PSNR and LPIPS. One may notice that the difference between FlatNet-gen and Tikh+U-Net is not very significant, due to the higher quality of Tikhonov reconstruction in the case of PhlatCam (Boominathan

Methods	PSNR (in dB)	LPIPS
Tikh+U-Net	20.60	0.298
FlatNet	**20.94**	**0.296**

Table 5.3: **Comparison of FlatNet with Tikh+U-Net**. FlatNet outperforms Tikh+U-Net because it learns an end-to-end mapping.

et al. (2020)). 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 5.4.2.

5.4.2 Performance on cropped measurements

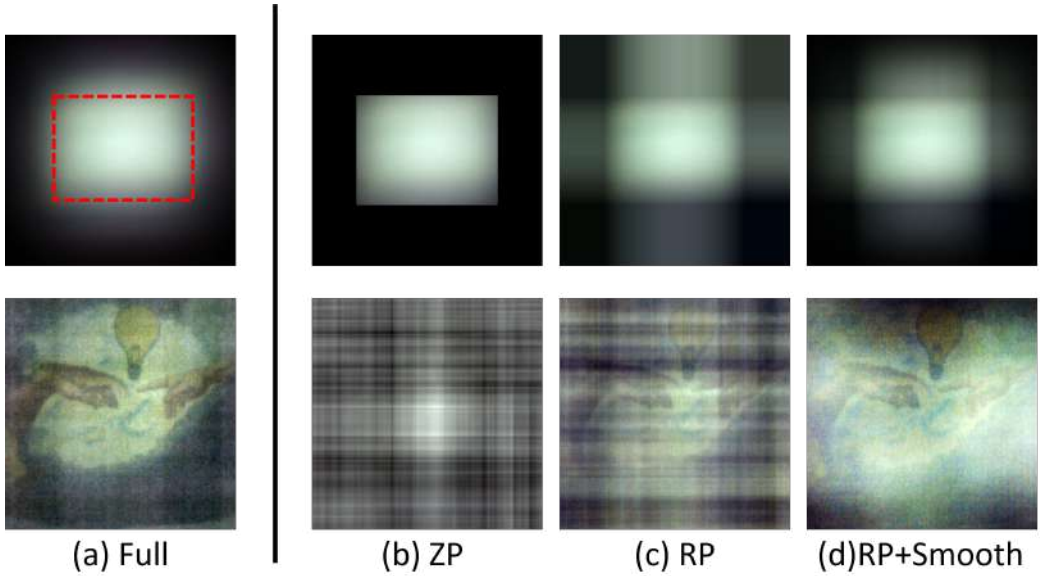

Figure 5.8: **Effect of padding on Wiener deconvolution for cropped measurement**. Top row shows the measurement while the bottom row shows the corresponding Wiener reconstruction. (a) Full measurement. Red box indicates the cropped out region. (b) Zero padded measurement and the corresponding reconstruction. (c) Replicate padded measurement and the corresponding reconstruction. (d) Smoothened replicate padded measurement along with the corresponding reconstruction. Line artefacts are significantly reduced in (d) which is used in this work.

As we have already seen in Section 4.1, 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 2.4. 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.

It was previously observed in Reeves (2005) that estimating the cropped out pixels followed by a Wiener deconvolution performed very closely to the Wiener deconvolution applied on replicate padded measurement. Following this observation in Reeves (2005), 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. However, in contrast to the edge-tapering operation, our smoothening operation is significantly cheaper as it doesn't involve any convolution with the large lensless PSF.

The effectiveness of our method of padding can be observed in Figure 5.8. Mathematically, the trainable inversion stage changes to,

$$X_{\text{interm}} = \mathcal{F}^{-1}(\mathcal{F}(W) \odot \mathcal{F}(pad(Y))). \tag{5.1}$$

This is a modification to Equation 4.1 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.

We can see that FlatNet-gen with replicate padding followed by smoothing led to fewer line artifacts. Figure 5.9 shows the reconstruction quality for the display captured cropped measurement compared with full measurement for Tikh+U-Net, Le-ADMM

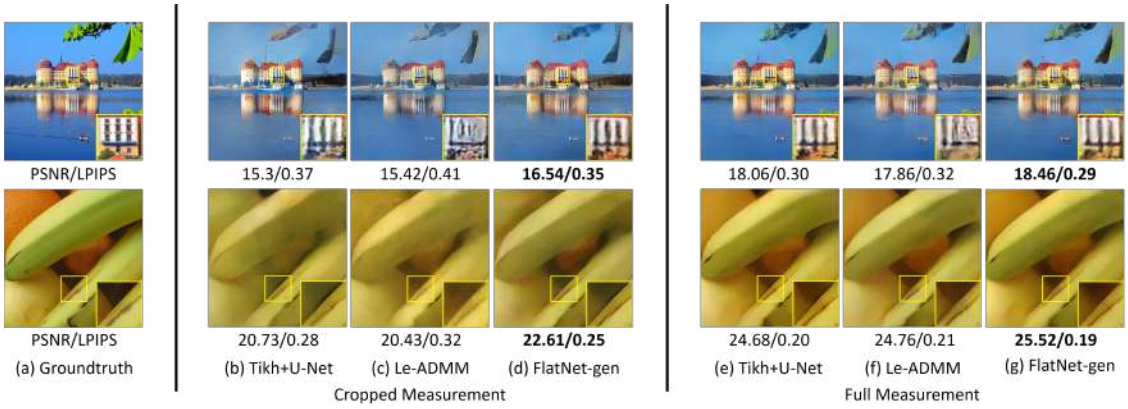

Figure 5.9: **Display Captured Reconstructions for cropped PhlatCam measurements**. The difference observed in the performance of FlatNet for cropped and full measurements is small. This difference is, however, large for both Le-ADMM and Tikh+U-Net.

and FlatNet. It should be noted that for applying Tikh+U-Net approach on cropped measurement, we apply the Wiener deconvolution on the padded measurement i.e. replicate padded followed by smoothing operation.

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 5.10, 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 5.4 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 5.9 and 5.10 and Table 5.4, the PSF size (608×870) exceeds the assumed sensor size (606×864). This is barely 30% of the pixels present in the full measurement. In such a case, estimation of the true PSF is a tedious process and one can use the uncalibrated FlatNet-gen-UC. From Table 5.4, 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

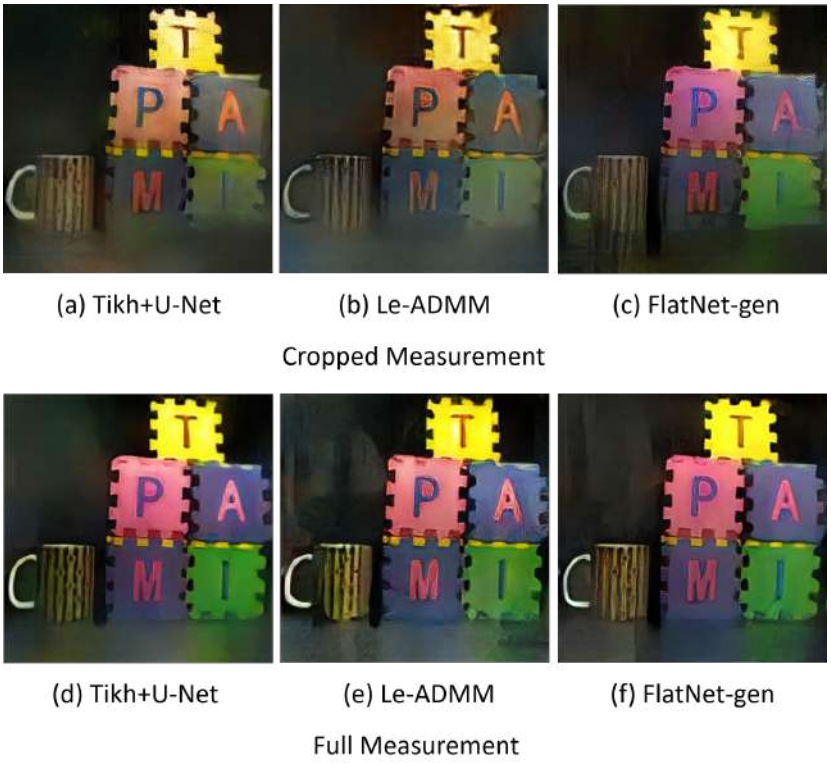

Figure 5.10: **Direct Captured Reconstructions for cropped PhlatCam measurements**. We can see FlatNet-gen performs reasonably well while both Le-ADMM and Tikh+U-Net breakdown. This can be observed through the colour of the letters and hazy appearance especially around the borders in Tikh+U-Net and Le-ADMM.

provided in Section 5.4.3.

Apart from the crop size mentioned above, we also show the performance of the learning based approaches for various different crop sizes in Figure 5.11. 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.

5.4.3 Qualitative Comparison for Uncalibrated PSF Case

In Sections 5.3 and 5.4.1, we provided the quantitative comparison for FlatNet-gen with Le-ADMM and Tikh+U-Net. In this section, we provide the visual results for the uncalibrated versions of the same. In particular, we use PSF simulated using the method described in Section 4.1 and use this PSF for learning Le-ADMM, Tikh+U-Net

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**

Table 5.4: **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.

and FlatNet-gen. We provide the comparison for both full measurement in Figure 5.12 and cropped measurement in Figure 5.13. We can see clearly that the performance of FlatNet-gen-UC is very close to its calibrated counterpart i.e. FlatNet-gen-C. However, this is not the case with Le-ADMM and Tikh+U-Net, demonstrating the effectiveness of the trainable inversion layer to work effectively with even an inexact copy of the Point Spread Function.

This can prove to be quite beneficial in practical settings, when lensless cameras are fabricated on a large scale. Even with tight control on mask lithography and sensor mounting techniques, the Point Spread Function (PSF) between two cameras may not be identical. Notice that Le-ADMM which assumes an explicity calibrated PSF being used as an input parameter cannot effectively handle such discrepancies.

5.4.4 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.

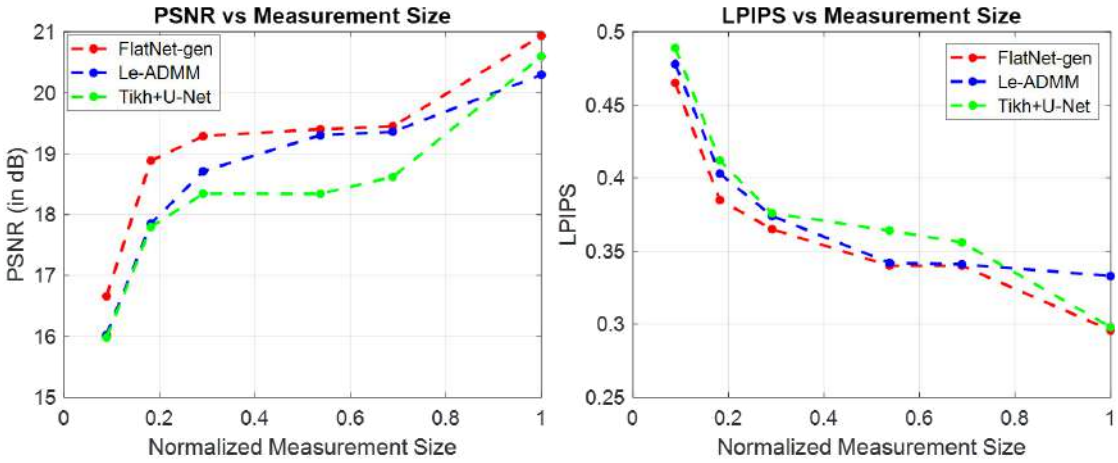

Figure 5.11: **Performance of learning based techniques for various amount of crops**. We plot the PSNR and LPIPS of FlatNet-gen, LeADMM and Tikh+U-Net under various measurement sizes normalized with respect to full measurement size. We can see FlatNet-gen consistently outperforms other learning based methods for all crop sizes.

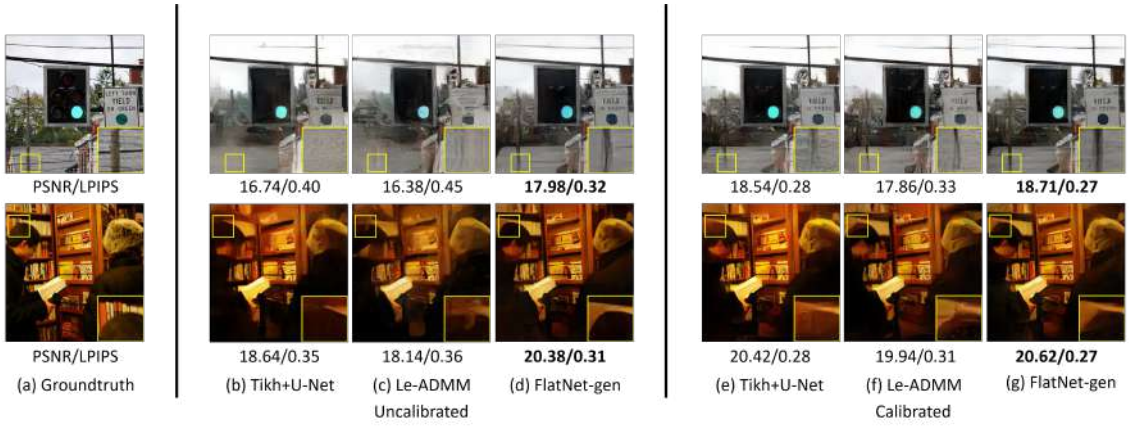

Figure 5.12: **Comparison between uncalibrated and calibrated learning based approaches for full PhlatCam measurement**. Tikh+U-Net and Le-ADMM rely on accurate estimation of PSF while FlatNet-gen relies on PSF only for initialization and rather learns the inverse of the PhlatCam forward model. FlatNet-gen higher quality reconstructions with finer details for both calibrated and uncalibrated case. This is not the case for Le-ADMM or Tikh+U-Net.

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 5.14. We collected 500 pairs of such data, keeping 475 pairs for training and 25 for testing. We finetune the entire

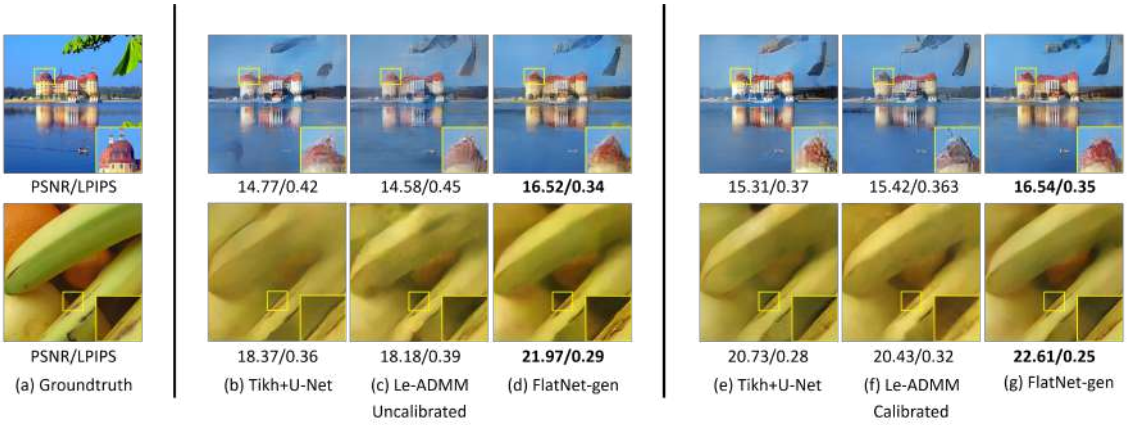

Figure 5.13: **Comparison between uncalibrated and calibrated learning based approaches for cropped PhlatCam measurement**. FlatNet-gen provides higher quality reconstruction for both calibrated and uncalibrated case even when the measurement is extensively cropped. This indicates that FlatNet-gen can be used for small sensor setup without accurately estimating the PSF.

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 (Mechrez *et al.* (2018)) which was previously proposed for unaligned data. We note that we did unsuccessfully attempt to use a feature based (such as ORB or SURF) registration method- where we would use baseline FlatNet reconstructions and Webcam captures to find a homography. This, however, was not successful, since the artefacts found in the FlatNet reconstructions lead to poor alignment results.

Figure 5.14 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. The strong line artifacts observed in the reconstructed fire extinguisher image for FlatNet-gen without finetuning indicates that the signal outside the field of view described by the CRA is significant and as a result, the full measurement extends beyond the assumed spatial dimension of 1280×1408. These line artifacts are however suppressed due to finetuning.

It is interesting to observe the effectiveness of the finetuned FlatNet for cropped unconstrained indoor scenes. In Figure 5.15, we provide visual comparison for the reconstructions from cropped measurement and full measurement along with the web-

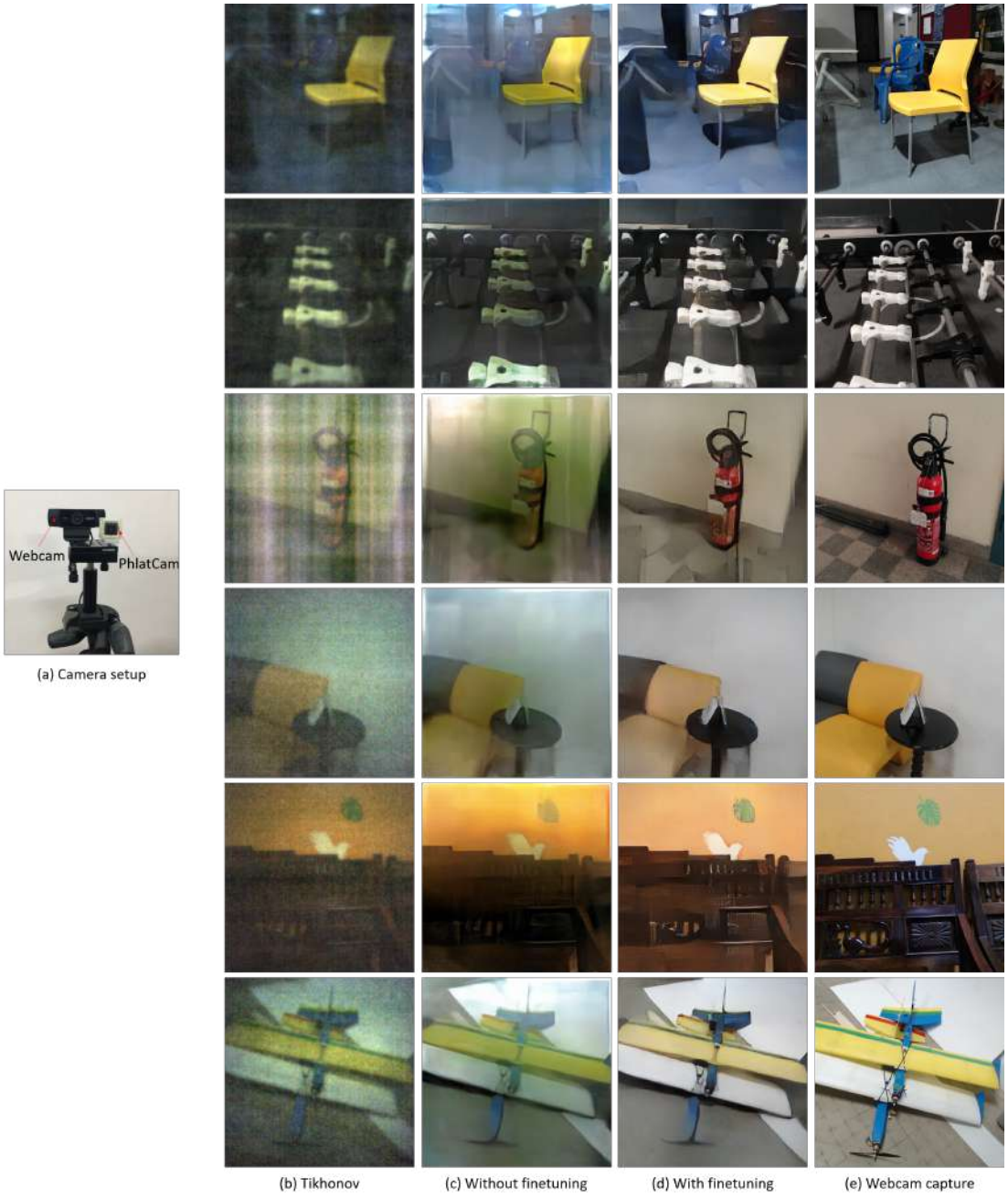

Figure 5.14: **Photorealistic reconstruction for unconstrained indoor scenes**. (a) The PhlatCam-Webcam setup to capture the dataset for finetuning FlatNet-gen. (b) Tikhonov reconstruction. (c) Reconstructions from FlatNet-gen trained just on display captured data. (d) Reconstructions using FlatNet-gen finetuned on unconstrained indoor captures. (e) Webcam image for reference. Finetuning makes the reconstructions more realistic.

cam capture. We show result for crop sizes of 990×1254. It should be noted that in an unconstrained setup, there may be large signals (due to bright objects) outside the field of view described by the CRA which would result in strong line artifacts in the reconstructions produced by model without finetuning.

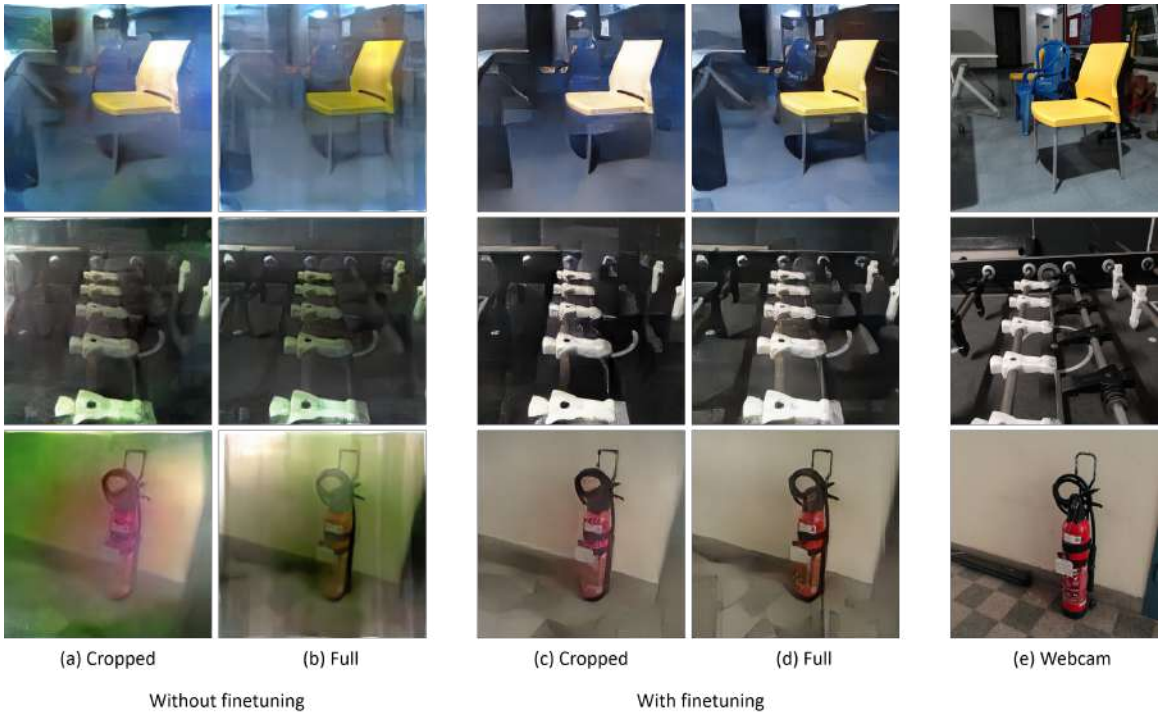

Figure 5.15: **Cropped measurements for Unconstrained Indoor Scenes**. We can observe that FlatNet-gen finetuned on unconstrained scenes provides reasonable reconstruction quality even for cropped measurements

5.4.5 Effect of Bright Object

For a highly multiplexed lensless imager, every pixel receives light from every point in the scene. Hence, if there is any really bright object (like a highly reflective object or a lamp) in the scene, the light from the object can dominate the pixel intensities and result in severe reconstruction artifacts on the dimmer objects. We show that, using FlatNet, the artifacts are minimized resulting in a higher quality reconstruction of the scene.

We show the bright object problem by introducing an LED into the scene. Figure 5.16 shows the reconstruction for PhlatCam (Boominathan *et al.* (2020)). We can observe that FlatNet-gen reconstructions have significantly fewer artifacts than other traditional and learning based approaches.

While such an experiment is not a reliable analysis of high dynamic scene capture using lensless cameras, it represents an early indicator that learning based techniques can (to a certain extent) offset sensor saturation. In the future, it would be interesting to see if learning based techniques such as HDRNet (Gharbi *et al.* (2017)) can be adapted to lensless systems.

Figure 5.16: **Reconstruction of scenes with bright objects (LED) using PhlatCam**. Artifacts occuring in Tikhonov reconstructions are amplified by Tikh+U-Net reconstruction. While Le-ADMM performs slightly better than Tikh+U-Net for PhlatCam, they are outperformed by FlatNet-gen

CHAPTER 6

Conclusion

In this paper, we propose end-to-end trainable deep network for photorealistic scene reconstruction from lensless measurements. Despite the numerous promises that lensless imaging provides, it is restricted by the quality of image that can recovered using such a thin and cheap camera. In this paper, we have attempted to bridge this gap between the promise of lensless imaging and its performance. Our reconstruction algorithm provides significant advantage over existing approaches including some deep learning based approaches. This is naturally reflected in the high quality of reconstructions we get for both display and real captures under both large and small sensor scenarios. Finally, we show that by finetuning our model trained on display captured measurements, using unaligned Webcam-PhlatCam indoor scenes, we can recover extremely photorealistic images from these tiny cameras.

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

REFERENCES

1. **Adams, J. K.**, **V. Boominathan**, **B. W. Avants**, **D. G. Vercosa**, **F. Ye**, **R. G. Baraniuk**, **J. T. Robinson**, and **A. Veeraraghavan** (2017). Single-frame 3d fluorescence microscopy with ultraminiature lensless flatscope. *Science advances*, **3**(12), e1701548.

2. **Antipa, N.**, **G. Kuo**, **R. Heckel**, **B. Mildenhall**, **E. Bostan**, **R. Ng**, and **L. Waller** (2018). Diffusercam: lensless single-exposure 3d imaging. *Optica*, **5**(1), 1–9.

3. **Antipa, N.**, **P. Oare**, **E. Bostan**, **R. Ng**, and **L. Waller**, Video from stills: Lensless imaging with rolling shutter. *In 2019 IEEE International Conference on Computational Photography (ICCP)*. IEEE, 2019.

4. **Asif, M. S.**, **A. Ayremlou**, **A. Sankaranarayanan**, **A. Veeraraghavan**, and **R. G. Baraniuk** (2017). Flatcam: Thin, lensless cameras using coded aperture and computation. *IEEE Transactions on Computational Imaging*, **3**(3), 384–397.

5. **Boominathan, L.**, **M. Maniparambil**, **H. Gupta**, **R. Baburajan**, and **K. Mitra** (2018). Phase retrieval for fourier ptychography under varying amount of measurements. *arXiv preprint arXiv:1805.03593*.

6. **Boominathan, V.**, **J. Adams**, **J. Robinson**, and **A. Veeraraghavan** (2020). Phlatcam: Designed phase-mask based thin lensless camera. *IEEE transactions on pattern analysis and machine intelligence*.

7. **Boominathan, V.**, **J. K. Adams**, **M. S. Asif**, **B. W. Avants**, **J. T. Robinson**, **R. G. Baraniuk**, **A. C. Sankaranarayanan**, and **A. Veeraraghavan** (2016). Lensless imaging: A computational renaissance. *IEEE Signal Processing Magazine*, **33**(5), 23–35.

8. **Caroli, E.**, **J. Stephen**, **G. Di Cocco**, **L. Natalucci**, and **A. Spizzichino** (1987). Coded aperture imaging in x-and gamma-ray astronomy. *Space Science Reviews*, **45**(3-4), 349–403.

9. **Chi, W.** and **N. George** (2011). Optical imaging with phase-coded aperture. *Optics express*, **19**(5), 4294–4300.

10. **Dave, A.**, **A. Kumar**, **K. Mitra**, *et al.*, Compressive image recovery using recurrent generative model. *In 2017 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2017.

11. **Dave, A.**, **A. K. Vadathya**, **R. Subramanyam**, **R. Baburajan**, and **K. Mitra** (2018). Solving inverse computational imaging problems using deep pixel-level prior. *IEEE Transactions on Computational Imaging*, **5**(1), 37–51.

12. **DeWeert, M. J.** and **B. P. Farm** (2015). Lensless coded-aperture imaging with separable doubly-toeplitz masks. *Optical Engineering*, **54**(2), 023102.

13. **Dicke, R.** (1968). Scatter-hole cameras for x-rays and gamma rays. *The astrophysical journal*, **153**, L101.

14. **Duarte, M. F.**, **M. A. Davenport**, **D. Takhar**, **J. N. Laska**, **T. Sun**, **K. F. Kelly**, and **R. G. Baraniuk** (2008). Single-pixel imaging via compressive sampling. *IEEE signal processing magazine*, **25**(2), 83–91.

15. **Gharbi, M.**, **J. Chen**, **J. T. Barron**, **S. W. Hasinoff**, and **F. Durand** (2017). Deep bilateral learning for real-time image enhancement. *ACM Transactions on Graphics (TOG)*, **36**(4), 118.

16. **Goodfellow, I.**, **J. Pouget-Abadie**, **M. Mirza**, **B. Xu**, **D. Warde-Farley**, **S. Ozair**, **A. Courville**, and **Y. Bengio**, Generative adversarial nets. *In Advances in neural information processing systems*. 2014.

17. **Goodman, J. W.** (2005). Introduction to fourier optics. *Introduction to Fourier optics, 3rd ed., by JW Goodman. Englewood, CO: Roberts & Co. Publishers, 2005*, **1**.

18. **Gu, S.**, **Y. Li**, **L. V. Gool**, and **R. Timofte**, Self-guided network for fast image denoising. *In Proceedings of the IEEE International Conference on Computer Vision*. 2019.

19. **Huang, G.**, **H. Jiang**, **K. Matthews**, and **P. Wilford**, Lensless imaging by compressive sensing. *In 2013 IEEE International Conference on Image Processing*. IEEE, 2013.

20. **Johnson, J.**, **A. Alahi**, and **L. Fei-Fei**, Perceptual losses for real-time style transfer and super-resolution. *In European conference on computer vision*. Springer, 2016.

21. **Khan, S. S.**, **V. Adarsh**, **V. Boominathan**, **J. Tan**, **A. Veeraraghavan**, and **K. Mitra**, Towards photorealistic reconstruction of highly multiplexed lensless images. *In Proceedings of the IEEE International Conference on Computer Vision*. 2019.

22. **Kim, G.**, **K. Isaacson**, **R. Palmer**, and **R. Menon** (2017). Lensless photography with only an image sensor. *Applied optics*, **56**(23), 6450–6456.

23. **Kingma, D. P.** and **J. Ba** (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

24. **Kulkarni, K.**, **S. Lohit**, **P. Turaga**, **R. Kerviche**, and **A. Ashok**, Reconnet: Non-iterative reconstruction of images from compressively sensed measurements. *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2016.

25. **Ledig, C.**, **L. Theis**, **F. Huszár**, **J. Caballero**, **A. Cunningham**, **A. Acosta**, **A. Aitken**, **A. Tejani**, **J. Totz**, **Z. Wang**, *et al.*, Photo-realistic single image super-resolution using a generative adversarial network. *In Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017.

26. **Li, C.**, **W. Yin**, **H. Jiang**, and **Y. Zhang** (2013). An efficient augmented lagrangian method with applications to total variation minimization. *Computational Optimization and Applications*, **56**(3), 507–530.

27. **Mechrez, R.**, **I. Talmi**, and **L. Zelnik-Manor**, The contextual loss for image transformation with non-aligned data. *In Proceedings of the European Conference on Computer Vision (ECCV)*. 2018.

28. **Monakhova, K.**, **J. Yurtsever**, **G. Kuo**, **N. Antipa**, **K. Yanny**, and **L. Waller** (2019). Learned reconstructions for practical mask-based lensless imaging. *Optics express*, **27**(20), 28075–28090.

29. **Mousavi, A.**, **A. B. Patel**, and **R. G. Baraniuk**, A deep learning approach to structured signal recovery. *In 2015 53rd Annual Allerton Conference on Communication, Control, and Computing (Allerton)*. IEEE, 2015.

30. **Ramachandran, P.**, **B. Zoph**, and **Q. V. Le** (2017). Searching for activation functions. *arXiv preprint arXiv:1710.05941*.

31. **Reddy, D.**, **A. Veeraraghavan**, and **R. Chellappa**, P2c2: Programmable pixel compressive camera for high speed imaging. *In CVPR 2011*. IEEE, 2011.

32. **Reeves, S. J.** (2005). Fast image restoration without boundary artifacts. *IEEE Transactions on image processing*, **14**(10), 1448–1453.

33. **Rick Chang, J.**, **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 Proceedings of the IEEE International Conference on Computer Vision*. 2017.

34. **Ronneberger, O.**, **P. Fischer**, and **T. Brox**, U-net: Convolutional networks for biomedical image segmentation. *In International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015.

35. **Russakovsky, O.**, **J. Deng**, **H. Su**, **J. Krause**, **S. Satheesh**, **S. Ma**, **Z. Huang**, **A. Karpathy**, **A. Khosla**, **M. Bernstein**, **A. C. Berg**, and **L. Fei-Fei** (2015). ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, **115**(3), 211–252.

36. **Satat, G.**, **M. Tancik**, and **R. Raskar** (2017). Lensless imaging with compressive ultrafast sensing. *IEEE Transactions on Computational Imaging*, **3**(3), 398–407.

37. **Shi, W.**, **J. Caballero**, **F. Huszár**, **J. Totz**, **A. P. Aitken**, **R. Bishop**, **D. Rueckert**, and **Z. Wang**, 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.

38. **Shimano, T.**, **Y. Nakamura**, **K. Tajima**, **M. Sao**, and **T. Hoshizawa** (2018). Lensless light-field imaging with fresnel zone aperture: quasi-coherent coding. *Applied optics*, **57**(11), 2841–2850.

39. **Simonyan, K.** and **A. Zisserman** (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

40. **Stork, D. G.** and **P. R. Gill** (2013). Lensless ultra-miniature cmos computational imagers and sensors. *Proc. SENSORCOMM*, 186–190.

41. **Tremblay, E. J.**, **R. A. Stack**, **R. L. Morrison**, and **J. E. Ford** (2007). Ultrathin cameras using annular folded optics. *Applied optics*, **46**(4), 463–471.

42. **Zhang, J.** and **B. Ghanem**, Ista-net: Interpretable optimization-inspired deep network for image compressive sensing. *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018.

LIST OF PAPERS BASED ON THESIS

1. Salman Siddique Khan*, **Varun Sundar***, Vivek Boominathan, Kaushik Mitra and Ashok Veeraraghavan "FlatNet: Towards Photorealistic Reconstructions of Lensless Measurements" *Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, under peer-review.

2. Salman Siddique Khan*, **Varun Sundar***, Vivek Boominathan, Kaushik Mitra and Ashok Veeraraghavan "FlatNet: Towards Photorealistic Reconstructions of Lensless Measurements" *CVPR CCD 2020 Spotlight Video*, to be made live on June 19th 2020.