Improving Reconstructions from Highly Multiplexed Lensless Images

A Project Report

submitted by

ADARSH V R

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

MASTER OF TECHNOLOGY

DEPARTMENT OF ELECTRICAL ENGINEERING
INDIAN INSTITUTE OF TECHNOLOGY MADRAS.

May 2019

THESIS CERTIFICATE

This is to certify that the thesis titled **Improving Reconstructions from Highly Multiplexed Lensless Images**, submitted by **Adarsh V R**, to the Indian Institute of Technology, Madras, for the award of the degree of **Master of Technology**, is a bona fide record of the research work done by him under my 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.

Dr. Kaushik Mitra
Research Guide
Assistant Professor
Department of Electrical Engineering
Indian Institute of Technology, Madras.

Place: Chennai

Date: May 05 2019

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to my project advisor Dr. Kaushik Mitra of Electrical Engineering Department for his invaluable guidance and wholehearted support during the entire course of the project. His immense knowledge and dedication towards research has not only motivated me but also helped me in evolving as a researcher.

I am greatly indebted PhD scholar Salman S Khan for his persistant help and encouragement in doing this project. This project work reflects contributions of many people with whom I had long discussions without which this would not have been possible. I would like to thank each and everyone of them.

I also thank the Head of the Electrical Engineering Department, Dr. Devendra Jalihal, for providing us the facilities required for the completion of this project. I also take this oppurtunity to thank my parents and all my friends who have helped me in times of need.

Last but not the least, I must not forget to thank Almighty for the wisdom and perseverance that He bestowed upon me during this project work, and indeed, throughout my life.

ABSTRACT

With the burgeoning of applications like Internet of Things (IoT), there is a need for cameras which have thin-form factor and less weight, that can be integrated anywhere and can be used for surveillance and distributed monitoring. Lensless imaging systems make such compact models realizable. However, reduction in the size and cost of these imagers comes at the expense of their image quality due to the high degree of multiplexing inherent in their design. This work particularly focus on FlatCam [1] , a lensless imager consisting of a coded mask placed over a bare CMOS sensor. Existing techniques for reconstructing FlatCam measurements suffer from several drawbacks including lower resolution and dynamic range than lens-based cameras. In this thesis, two methods to improve lensless reconstructions are explored. First, an end-to-end calibration free data driven method is implemented that obtain image reconstructions from lensless measurements that are more photorealistic than those currently available in the literature. Second, we also look at optimising the mask specifically for image reconstruction, making the mask more robust to noise as compared to the current mask used.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

The basic mechanism of a camera, which uses a lens to focus the light from a scene and project the image onto a photosensitive surface, was established in the 16th century. Since then, a lens has been indispensable for cameras. Although many improvements have been made to photosensitive materials for recording images, lenses remain an integral part of modern imaging systems. But with the proliferation of fields like Internet of Things (IoT), augmented reality etc, the roles of cameras have changed from merely taking photographs to being inferential inputs and such emerging applications like surveillance, drones etc impose stringent constraints on the size, weight, cost etc which cannot be met by the current lens-based imaging systems. Presence of lenses in cameras introduces a lot of constraints. Due to the large distance required between the lens and the sensor to achieve focus, cameras end up being thick, with thickness increasing at larger lens aperture sizes. Also, lenses required for wavelengths farther into the infrared and ultraviolet spectra are very expensive. Furthermore, lens-based systems require post-fabrication assembly. While a variety of devices including mobile equipment and robots have transformed themselves into thinner and more compact models, due to these limitations of lenses,there has been a limit to incorporating cameras into such thinner and more compact devices.

All the above mentioned problems can be solved by eliminating the lenses. A lensless camera is a digital camera that can take photos and video images without using any lens. It uses a mask or a permeable film, instead of a lens, to project images of the photographic subjects and reproduces pictures and images through digital processing. Recent advancements in sensor technologies and computational imaging techniques have resulted in the emergence of lensless imaging systems. These imaging systems differ from the conventional imaging system in the sense that they encode the incoming light to the sensor (instead of directly focusing it). A reconstruction algorithm is then required to decode the scene from the measurements. Lensless imaging systems provide numerous benefits over lens-based cameras. First, lensless imaging systems eliminate the need for a lens, which is the major contributor towards the size and weight of the

Figure 1.1: Lens-Based Cameras

camera. 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 [36]. Moreover, lensless cameras can be produced with traditional semiconductor fabrication technology and therefore exploit all its scaling advantages, yielding low-cost, high-performance cameras [2]. Earlier instances of using lensless coded aperture imaging systems for X-ray and gamma ray [6; 9; 3; 8; 4] are proofs that lensless imagers have better wavelength scaling as well.

However, the absence of a focusing element and the requirement of a reconstruction algorithm in lensless cameras result in two major challenges. First, lensless design results in an ill-conditioned system, yielding imperfect reconstructions. Second, poor design of the mask or reconstruction algorithm may greatly amplify noise in the images. Therefore, lensless cameras need efficient algorithms to overcome these challenges.

In this thesis, my aim is to improve the quality of reconstructions for Flatcam lensless imaging system, which consists of a coded mask placed over a bare sensor. Two approaches have been implemented in getting quality reconstructions from FlatCam measurements: First by developing a calibration-free reconstruction algorithm by using deep networks and second by optimising the mask of the camera specifically for reconstruction. The first part on end-to-end approach is a joint work with PhD scholar Salman S Khan [12]. My main contribution to this work is a calibration-free approach for the reconstruction. Other ablation studies are also done.

CHAPTER 2

Lensless Imaging and FlatCam

2.1 Lensless Imaging

In the absence of a lens, a sensor would simply record the average light intensity from the entire scene. Lensless imaging systems dispense with a lens by using other optical elements to manipulate the incoming light. The sensor records the intensity of the manipulated light, which may not appear as a focused image. However, when the system is designed correctly, the image can be recovered from the sensor measurements with the help of a computational algorithm as shown in figure 2.1.

2.2 Related Works

Pinhole Cameras: The very first cameras built centuries before the invention of lenses and photography were lensless. Pinhole Cameras, also known as camera obscura, offered simple architecture for lensless imaging that consists of a single aperture in front of a sensor. But these tiny pinholes drastically reduced the light reaching the sensors resulting in very noisy images. In fact, lenses were introduced to the light throughput, thus improving the quality of the images.

Coded Aperture Cameras: These cameras replace the tiny aperture of the pinhole cameras with a mask containing multiple apertures, thereby increasing the light efficiency.

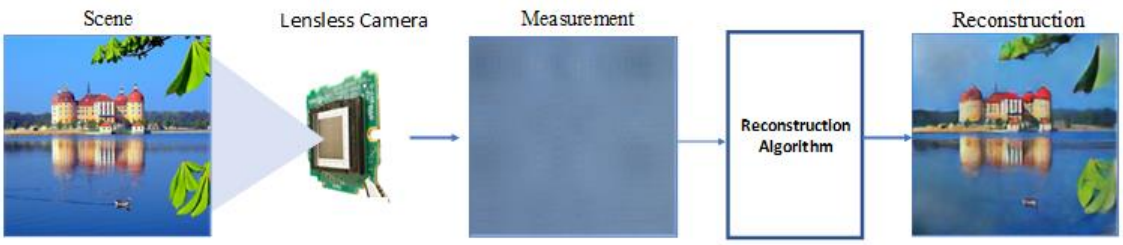

Figure 2.1: Lensless Imaging Architecture

Coded apertures were originally invented for the wavelengths of light that are not easily amenable to lens-based imaging. In a general coded aperture system, sensor measurements represent a superposition of the images formed behind each pinhole.

In contrast to a single-pinhole camera, the sensor measurements of a coded aperture camera do not resemble an image of the scene. Rather, each light source in the scene casts a unique shadow of the mask onto the sensor, encoding information about locations and intensities. We can represent the relationship between the scene and the sensor measurement as a linear system that depends on the pattern and placement of the mask. Inverting this system using an appropriate computational algorithm will recover an image of the scene.

Existing coded aperture cameras have following limitations. First, in these cameras, the masks are placed significantly far away from the sensor, thus increasing the form factor. Second, the masks employed have transparent features only in the central region, thus reducing the light throughput.

Zone Plates: A zone plate uses diffraction to focus light and form an image. It consists of concentric transparent and opaque rings. Light hitting a zone plate diffracts around the opaque regions and interferes constructively at the focal point. One advantage of zone plates is their large transparent area, which provides better light efficency.

Ultra-miniature lensless imaging with diffraction gratings: Recently, miniature cameras with integrated diffraction gratings and CMOS image sensors have been developed [27]. These cameras have been successfully demonstrated on tasks such as motion estimation and face detection. While these cameras are indeed ultra-miniature in total volume (100 Îijm sensorwidth by 200 Îijm thickness), they retain the large thicknes-to-width ratio (TWR) of conventional lens-based cameras. Because of the small sensor size, they suffer from reduced light collection ability.

Lensless Imaging with Fresnel zone aperture: Recently, Hitachi Ltd. proposed a lensless camera consisting of an image sensor and a Fresnel zone aperture (FZA). Point sources making up the subjects to be captured, cast overlapping shadows of the FZA on the sensor, which result in overlapping straight moirÃl' fringes due to multiplication of another virtual FZA in the computer. The fringes generate a captured image by two-dimensional fast Fourier transform. But the design of these cameras are fairly complex.

2.3 FlatCam

FlatCam is a lensless imaging system developed by Asif et al.[1] that consists of an amplitude mask placed above the CMOS sensor. It has extremely thin form factor as the mask is placed very close to the sensor. As the mask is made up of multiple apertures/pinholes, the resultant measurement recorded at the sensor is a superposition of the images formed due to each pinhole. In the design of FlatCam, we assume the sensor and the mask are planar and parallel to each other. separated by a distance *d*. For the sake of simplicity, we also assume that the mask used is binary: that is it contains transparent features that transmit light and opaque features that block it. Size of the transparent/opaque features is denoted by Δ and also assumption is made that the mask covers the entire sensor array.

Consider the one-dimensional (1-D) coded aperture system depicted in Fig.2.2, in which a single coded mask is placed at distance d from the sensor plane. We assume that the FOV of each sensor pixel is limited by a chief ray angle (CRA) θ_{CRA}, which implies that every pixel receives light only from the angular directions that lie within $\pm\theta_{CRA}$ with respect to its surface normal. Therefore, light rays entering any pixel are modulated by the mask pattern of length $w = 2dtan\theta_{CRA}$. As we increase (or decrease) the mask-to-sensor distance, d, the width of the mask pattern,w, also increases (or decreases). Assuming that the scene is far from the camera, the mask patterns for neighboring pixels shift by the same amount as the pixel width. If we assume that the mask features and the sensor pixels have the same width, Δ, then the mask patterns for neighboring pixels shift by exactly one mask element. If we fix $d \approx N\frac{\Delta}{2}tan\theta_{CRA}$, then exactly N mask features lie within the FOV of each pixel. If the mask is designed by repeating a pattern of N features, then the linear system that maps the light distribution in the scene to sensor measurements can be represented as a circulant system.

2.4 FlatCam : Mathematical Model

We characterize image formation using the geometric optics model. While this approach largely ignores diffraction, the resulting model is useful for the design and analysis of well-conditioned imaging architectures. Furthermore, the calibration procedure that we detail in subsequent sections can account for unmodeled diffraction effects.

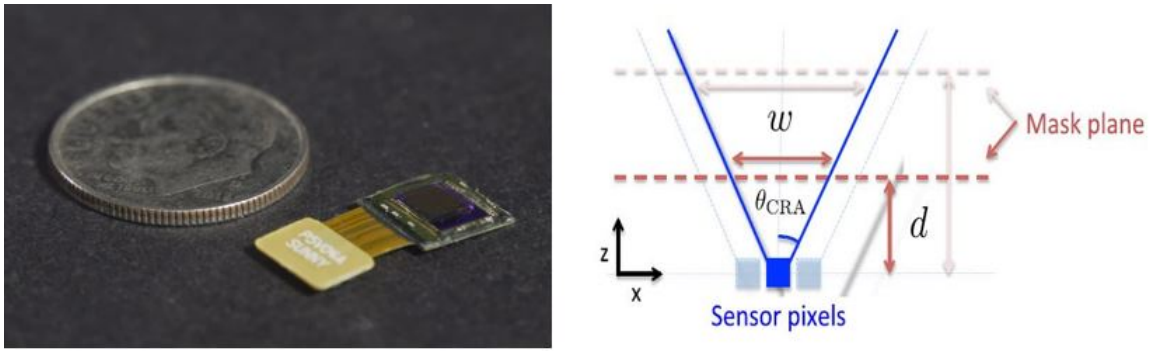

Figure 2.2: On left : A FlatCam prototype, On right:An illustration of a coded aperture system with a mask placed d units away from the sensor plane [1]

For the simplicity of notation, we assume a simplified 2-D world imaged by a one-dimensional (1-D) mask and sensor. The extension to a 3-D world imaged by a 2-D mask and sensor is straightforward except where stated otherwise. For a suitably defined scene irradiance vector X, the scene-to-sensor mapping can be described as,

$$Y = \Phi X + E. \tag{2.1}$$

where Φ is the measurement matrix, Y the image and E the measurement. This model can be interpreted in two different ways. First, each sensor measures a weighted, linear combination of light from multiple scene locations, and each row in Φ encodes the weights for the respective sensor. For a scene at infinity, the weights for two different sensor pixels simply differ by a translation of the mask pattern. As a consequence, the matrix Φ has a Toeplitz structure. Second, every light source in the scene casts a shadow of the mask on the sensor. Thus, the image formed on the sensor is a superposition of shifted and scaled versions of the mask. The shift and the scaling of the mask pattern encodes the angle and distance of the light source onto the sensor. These properties are invaluable in the design of masks that provide near-optimal recovery under noise. Given the image formation model, our tasks are to formulate an inversion algorithm that recovers the scene X from the sensed image Y and design mask patterns that achieve optimal recovery performance.

But for a megapixel scene and sensor, the Φ contains elements of order 10^{12}. This increases computational complexity. In FlatCam system, to reduce the complexity of Φ, we use a separable mask pattern. If the mask pattern is separable (i.e., an outer

product of two 1-D patterns), then the imaging system can be rewritten as

$$Y = \Phi_L X \Phi_R^T + E. \tag{2.2}$$

where Φ_L ,Φ_R denote matrices that correspond to 1-D convolution along the rows and columns of the scene, respectively, X is an N $\times$ N matrix containing the scene radiance, Y in an M $\times$M matrix containing the sensor measurements, and E denotes the sensor noise and any model mismatch. For a megapixel scene and a megapixel sensor, the calibration matrices have only 10^6 elements each, as opposed to 10^{12} elements in Φ. Here the calibration matrices have toeplitz structure.

2.5 Mask Pattern

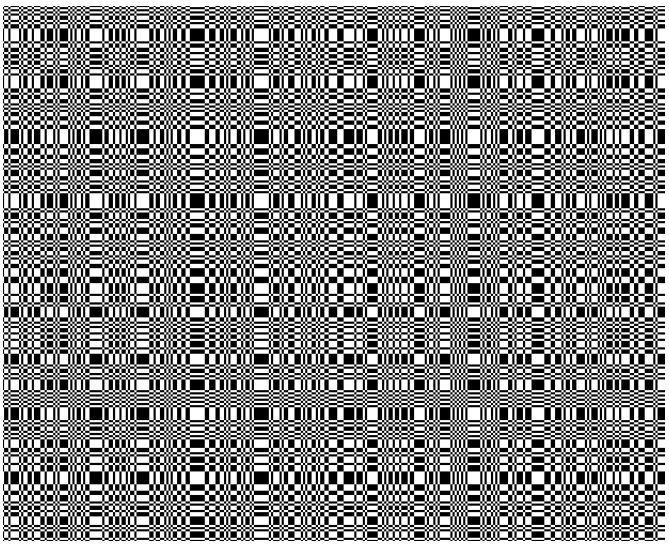

Figure 2.3: MLS mask

The design of mask patterns play an important role in the reconstruction. An ideal pattern should have well conditioned scene to sensor transfer functions, while maximising the light throughput. While designing mask for FlatCam, all these have been taken into consideration. Also for reducing computational complexity, separable mask has been designed. This means that it is an outer product of 2 1D vectors. In [1] , it is shown that mask generated using MLS pattern gives better results compared to other patterns like URA, MURA etc. The mls mask is shown in the figure 2.3.

2.6 Image Reconstruction: Current Methods

If both Φ_L and Φ_R are well-conditioned, then we can estimate X by solving a simple least-squares problem,

$$\hat{X}_{LS} = arg\min_X \left\|Y - \Phi_L X \Phi_R^T\right\|_2^2, \tag{2.3}$$

which gives a closed form solution

$$\hat{X}_{LS} = \Phi_L^+ Y \Phi_R^+, \tag{2.4}$$

where Φ_L^+ and Φ_R^+ are pseudo inverses of the calibration matrices.
If the matrices are not well conditioned or are under-determined (e.g., when we have fewer measurements M than the desired dimensionality of the scene N), some of the singular values are either very small or equal to zero. In these cases, where X_{LS} suffers from noise amplification., a regularizer term can be added to the least squares optimization.

$$\hat{X}_{LS} = arg\min_X \left\|Y - \Phi_L X \Phi_R^T\right\|_2^2 + \lambda R(X), \tag{2.5}$$

where $R(X)$ is the regularizer term and λ controls the trade-off between fidelity and regularization. In case of Tikhonov regularization, where $R(X) = ||X||_2^2$ a closed form solution can be obtained,

$$\hat{X} = V_L[(\Sigma_L^T U_L^T Y U_R \Sigma_R)./(\sigma_L \sigma_R^T + \lambda \mathbf{1}\mathbf{1}^T)]V_R^T, \tag{2.6}$$

where $\Phi_L = U_L \Sigma_L V_L^T$ and $\Phi_R = U_R \Sigma_R V_R^T$.
In many cases, exploiting the sparse or low-dimensional structure of the unknown image significantly enhances reconstruction performance. Natural images and videos exhibit a host of geometric properties, including sparse gradients and sparse coefficients in certain transform domains. Wavelet sparse models and total variation (TV) are widely used regularization methods for natural images. By enforcing these geometric properties, we can suppress noise amplification as well as obtain unique solutions. A pertinent example for image reconstruction is the sparse gradient model, which can be represented in

the form of the following TV minimization problem:

$$\hat{X}_{LS} = arg\min_{X} \left\| Y - \Phi_L X \Phi_R^T \right\|_2^2 + \lambda ||X||_{TV}, \tag{2.7}$$

The term $||X||_{TV}$ denotes the TV of the image X given by the sum of magnitudes of the image gradients. Given the scene X as a 2-D image, i.e., $X(u, v)$, we can define $G_u = D_u X$ and $G_v = D_v X$ as the spatial gradients of the image along the horizontal and vertical directions, respectively. The total variation of the image is then defined as

$$||X||_{TV} = \sum_{u,v} \sqrt{G_u(u, v)^2 + G_v(u, v)^2} \tag{2.8}$$

Minimizing the TV as in equation 2.8 produces images with sparse gradients. The optimization problem equation 2.8 is convex and can be efficiently solved using a variety of methods.

CHAPTER 3

End-to-end network for FlatCam reconstruction

3.1 Issues with current reconstruction methods

Reconstruction using current methods results in blurry outputs with vignetting effects. Also, the efficacy of these methods depends on hand crafted priors. Moreover, these methods are very sensitive to noise and results in highly noisy images. Also , 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. These methods also require calibration which is error prone and even slight mismatch can cause huge degradations in reconstructions. Hence, there is a need for a novel architecture that deals with all these issues.

3.2 Naive Approach

The Tikhonov solution of (2.6) is extremely fast to compute as shown in [1]. A naive way to obtain higher quality reconstruction from FlatCam measurements would be to obtain the Tikhonov regularized reconstruction and then use an image restoration framework to refine the reconstruction. To implement this, we pass the Tikhonov regularized reconstruction through a perceptual enhancement block (described in section 3.3.1). We use the same loss that is defined in section 3.3.3.

But using network over Tikhonov based reconstructions does not help much as the refinement network would not be able to get back any details lost in the Tikhonov based

Figure 3.1: Architecture of naive approach

reconstructions. Moreover, all those issues mentioned in section 3.1 will be applicable here also.

3.3 End-to-end network for FlatCam reconstruction

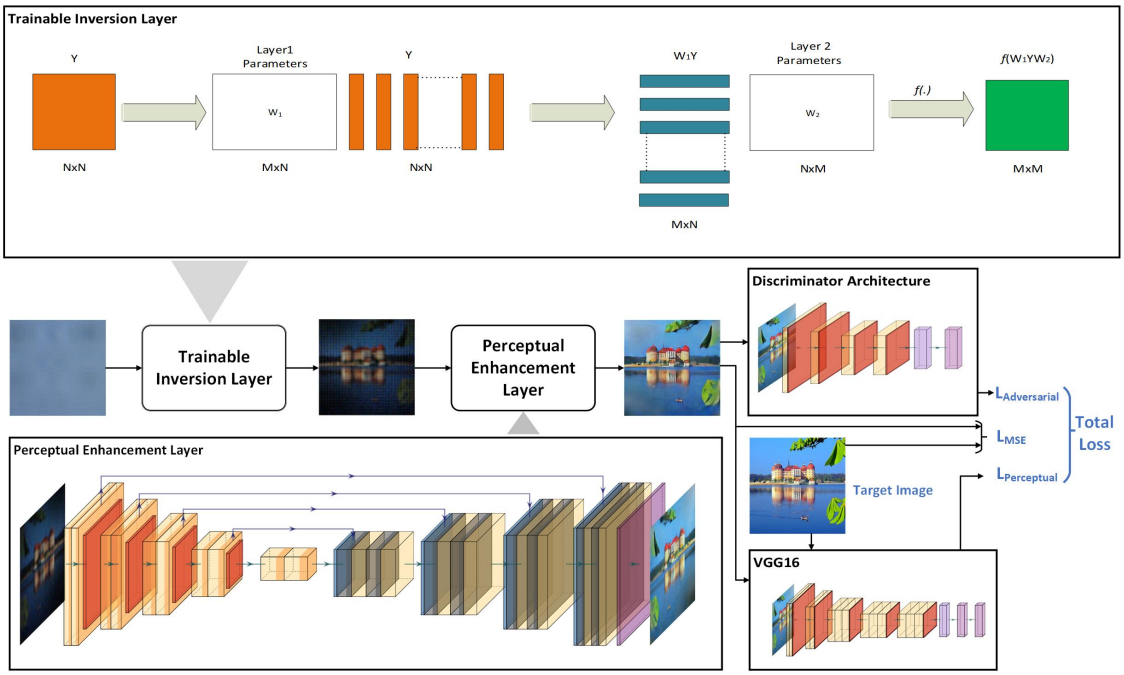

Figure 3.2: Overall architecture of the proposed system.

To address the difficulties in accurate FlatCam reconstruction, we take a data-driven approach at recovering the true scene from the highly multiplexed measurements. Following the success of Generative Adversarial Nets [10], our proposed network has two main components: a generator network that learns to output a visually meaningful reconstruction from the measurement and a discriminator network that tries to distinguish this reconstruction from real images. Both the networks are finally trained in an adversarial setup. Figure 3.2 shows the generalized block diagram for our method.

3.3.1 Generator architecture

Our generator network has two basic stages: the *trainable inversion stage* maps the FlatCam measurements into a space of intermediate reconstructions, and the *perceptual enhancement stage* refines this mapping into a semantically meaningful image.

Trainable inversion stage

In the first stage, we use two layers of trainable left and right matrix multiplications on the two-dimensional sensor measurments followed by a non-linearity. Figure 3.2 gives a diagrammatic overview of this stage. For the non-linearity, we use the leaky ReLU[16]. The dimension of the weight matrices depends on the dimension of the measurement and the scene dimension we want to recover.
It is important to initialize the weight matrices of this stage properly, so that the network does not get stuck in local minima. One way to do is to initialize our weight matrices (W_1 and W_2) with the adjoint of the calibration matrices. These calibration matrices are approximations of Φ_L and Φ_R in (2.2) physically obtained by the method described in [1]. This mode of initialization leads to faster convergence while training.

Calibration-Free Approach

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 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. The weight matrices (W_1 and W_2) are initialized with the adjoint of such constructed random matrices. We observed that the training time increased slightly for this initialization in comparison to transpose initialization.

In our experiments, we found that the products of the learned matrices W_1 and W_2

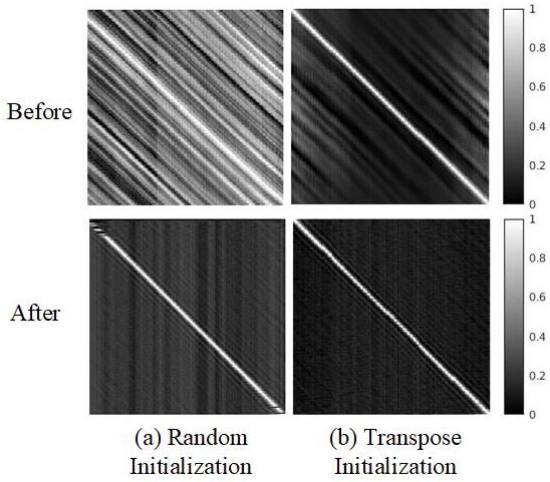

Figure 3.3: Product of the left weight matrix from the trainable inversion stage for with the calibration matrix ($W_1 \times \Phi_L$) before and after training. The top row shows the initial product at the beginning of training while the bottom row shows it after training the network. a) Random initialization. b) Transpose initialization.

with the forward model calibration matrices Φ_L and Φ_R closely resemble an identity matrices, implying that this stage has tried to invert the FlatCam forward model. This is shown in figure 3.3.

Perceptual enhancement stage

Once we obtain the output of the trainable inversion stage, which is of same dimension as that of the natural image, we use a fully convolutional network to map it into the natural image space. Owing to its large scale success in image-to-image translation problems and its multi-resolution structure, we choose a U-Net [23] 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.

3.3.2 Discriminator architecture

The trainable inversion and the perceptual enhancement stage form the generator of our architecture. We then use a discriminator framework to classify our generator'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 [21] in our discriminator.

3.3.3 Loss function

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:

$$\mathcal{L}_{\text{MSE}} = \|I_{\text{true}} - I_{\text{est}}\|_2^2 . \tag{3.1}$$

Perceptual loss: To measure the semantic difference between the estimated output and the ground truth, we use the perceptual loss introduced in [11]. We use a pre-trained VGG-16 [25] 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,

$$\mathcal{L}_{\text{percept}} = \|\phi_{22}(I_{\text{true}}) - \phi_{22}(I_{\text{est}})\|_2^2 + \|\phi_{43}(I_{\text{true}}) - \phi_{43}(I_{\text{est}})\|_2^2 . \tag{3.2}$$

Adversarial loss: Adversarial loss [10; 14] was added to further bring the distribution of the reconstructed output close to those of the real images. Given a discriminator D, this loss is given as,

$$\mathcal{L}_{\text{adv}} = -log(D(I_{\text{est}})). \tag{3.3}$$

Total loss: Our total loss is a weighted combination of the three losses and is given as,

$$\mathcal{L} = \lambda_1 \mathcal{L}_{\text{MSE}} + \lambda_2 \mathcal{L}_{\text{percept}} + \lambda_3 \mathcal{L}_{\text{adv}}. \tag{3.4}$$

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

CHAPTER 4

Amplitude Mask Optimisation

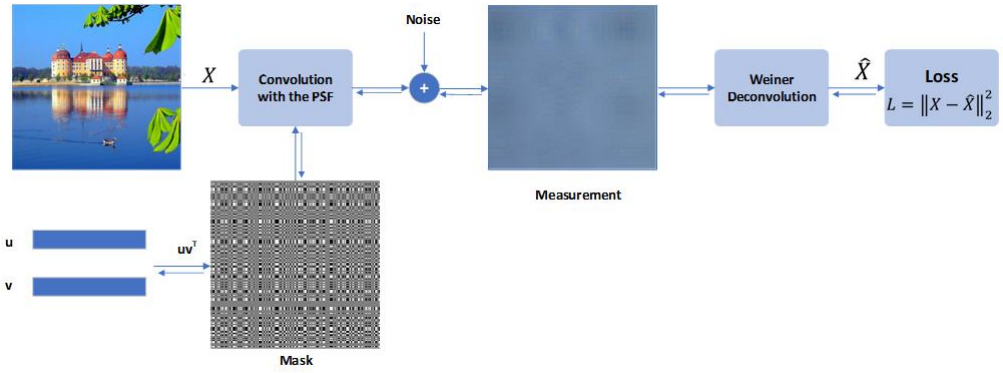

Figure 4.1: Architecture for Amplitude Mask Optimisation

Optimizing the parameters of optical elements and point spread function engineering are well-known techniques in the computational optics. Optimized optical system parameters have proven useful for extended depth of field [7], motion [22] and defocus deblurring [30], 4D light field imaging [17], super-resolved localization microscopy[19], and full-color imaging with diffractive optics[20]. Previously-proposed optimization approaches of optical elements are mainly based on heuristic cost functions applied to the PSFs, which may be a feasible approach for image deconvolution but it remains unclear how the PSF of a camera affects higher-level computer vision tasks such as image classification; second, although image processing is applied to the recorded images to remove residual aberrations or perform some inference tasks, the post-processing algorithm is usually independent of the optics design and fails to provide significant insights to guide it. In [26] , a novel perspective of joint optimization of a single diffractive or refractive element with a deconvolution post-processing step is proposed.

4.1 Mask Design

The design of a mask for a lensless imager plays an important role in the quality of images that the camera system can produce. An ideal mask should provide a well conditioned scene-to-sensor transfer function, while providing required light throuhput. In

FlatCam, three major factors were considered for mask design.

Light Throughput: Placing an amplitude-modulating mask very close to (and completely covering) the sensor results in a light collection efficiency that is a fraction of the fundamental light collection limit of the sensor. Having as many open features possible, should increase the light throughput. But according to [1], it is seen that increasing open features beyond 50% deteriorates the conditioning of the mask and MLS pattern outperforms other random patterns.

As described above, while it is true that the light collection ability of our FlatCam design is one-half of the maximum achievable with a particular sensor, the main advantage of the FlatCam design is that it allows us to use much larger sensor arrays for a given device thickness constraint, thereby significantly increasing the light collection capabilities of devices under thickness constraints.

Computational Complexity: To reduce the computational complexity, FlatCam uses separable mask pattern. This means we consider mask pattern to be an outer product of two 1D vectors. In [1], MLS pattern of length 511 is used for generating the mask.

Numerical Conditioning: The mask pattern should be chosen to make the multiplexing matrices Φ_L and Φ_R as numerically stable as possible, which ensures a stable recovery of the image X from the sensor measurements Y . Such Φ_L and Φ_R should have low condition numbers, i.e., a flat singular value spectrum. However, because of the inevitable non-idealities in our implementation, such as the limited sensor CRA and the larger than optimal sensor-mask distance due to the hot mirror, the actual Φ_L and Φ_R we obtain using a separable M-sequence based mask do not achieve a perfectly flat spectral profile.

4.2 Proposed Framework

Theoretically, MLS mask is an ideal mask. But due to the presence of sensor noise and other non-idealities of the implementation, practically, the reconstructions from MLS mask are not very good. They are blurry as well as have lot of vignetting effects. So, a practically implementable and more robust mask can be implemented if we consider the presence of sensor noise as well. This is the idea behind optimizing mask specifically for reconstructions.

In this approach, we try to optimise the mask specifically for reconstruction purpose.

Here we assume the separability condition so that mask obtained is computationally efficient. This means that the mask will be an outer product of two vectors and the rank of the mask will be one.
The proposed framework is an end-to-end differentiable pipeline architecture.

Forward Pass

In each forward pass, the PSF p of the current optical element is simulated using the FlatCam model explained in section 2.4. PSF is nothing but the shadow cast by the mask on the sensor. Therefore, in case of amplitude mask, the PSF is ideally the amplitude mask itself. The simulated PSF is then convolved with a batch of images, and noise is added to account for sensor read noise. Then, the scene is reconstructed using weiner deconvolution.
Weiner deconvolution is a method in which a weiner filter is applied before the deconvolution step in order to reduce the effect of additive white gaussian noise. It works in frequency domain, attempting to minimize the impact of deconvolved noise at frequencies which have a poor signal-to-noise ratio. It attenuates frequencies dependent on their signal-to-noise ratio. The operation of weiner deconvolution can be represented as:

$$\hat{X} = \mathcal{F}^{-1}\left\{ \frac{\bar{p}_c^*}{|\,\bar{p}_c^*|^2 + \gamma} \mathcal{F}\{Y\} \right\} \tag{4.1}$$

where $\bar{p}_c$ is the optical transfer function, Y the measurement and γ is the parameter. Finally, a differentiable loss L, mean squared error with respect to the ground-truth image, is defined on the reconstructed images.

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

Backward Pass

In the backward pass, the error is backpropagated all the way back to the PSF simulation,and finally, to the 1D vectors itself, whose outer product gives the amplitude mask. Here we also try to optimise the weiner parameter γ as well.

Mask with Rank > 1

The above mentioned architecture uses the idea of separable mask whose rank will be always one. In order to explore whether increasing the rank of mask will improve the reconstructions, we try to optimize mask for rank 10 as well. So, instead of assuming mask is an outer product of 2 random vectors, we assume that mask is an outer product of two matrices of dimension 511×10 and 10×511. This can be considered as outer product between 10 1D vectors with ten other 1D vector, thus maintaining the rank to be ten. We try to optimize these 20 vectors in order to get an optimised rank 10 mask.

CHAPTER 5

Experiments and Results

5.1 Dataset

Display capture setup

Collecting a large scale dataset of lensless measurements along with their aligned ground truth is a challenging task. To overcome this challenge, the first setup we use to capture real images is the display capture setup. In this setup, we place a monitor in front of the FlatCam and capture the images displayed on it. For the ground truth, we randomly selected 10 images from each of the ImageNet [24] classes and created a dataset of 10000 images. Out of this, we kept 9900 images from 990 classes for training and the rest 10 classes or 100 images for testing. We call these measurements display captures.

Direct capture setup

It is important to visually evaluate the performance of our reconstruction network on a direct real world setup. For collecting data for this setup, we place objects in front of FlatCam and directly capture the measurement. For this setup we do not have a corresponding ground truth. We call these measurements direct captures.

5.2 Implementation details

The FlatCam prototype used is Point Grey Flea3 camera with 1.3MP e2v EV76C560 CMOS sensor with 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.
In case of end to end reconstruction, 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

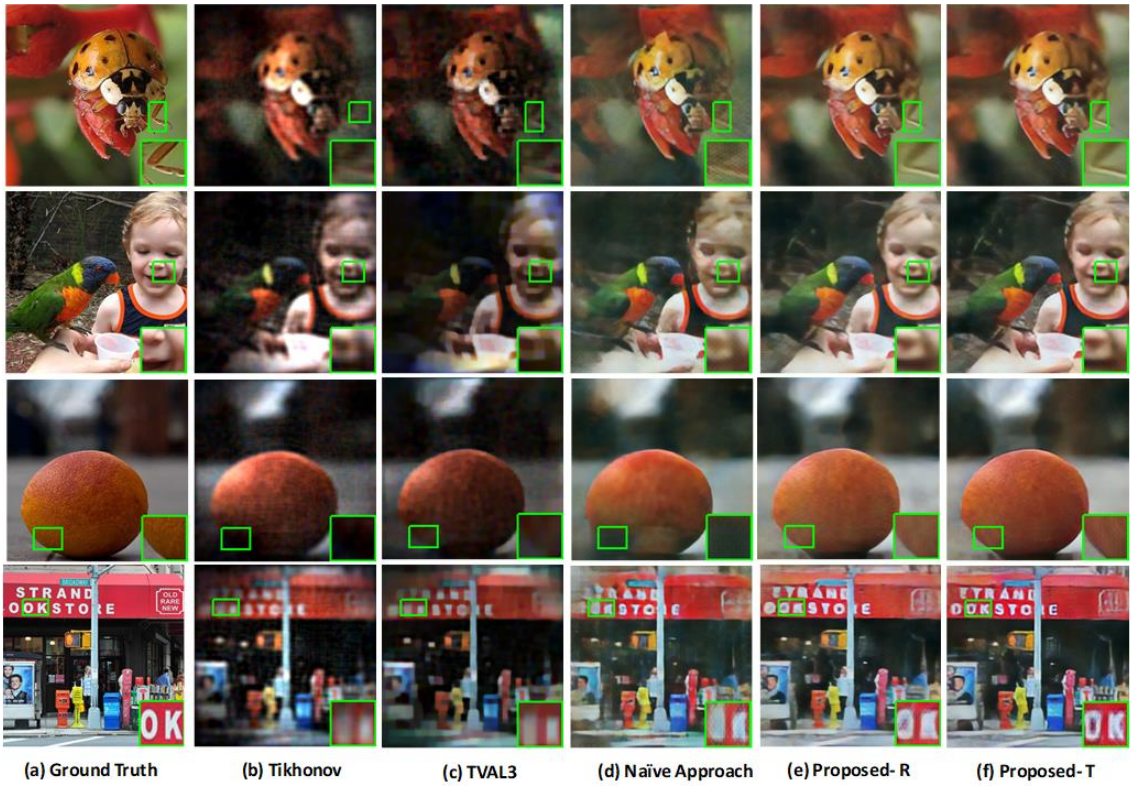

Figure 5.1: Reconstruction of display captured measurements using various approach. Proposed-R is the calibration free approach while Proposed-T is the transpose initialization.

within the network. FlatCam measurements of dimension $500 \times 620 \times 4$ 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 random initialization, we trained it for 60K iterations. The Adam [13] optimizer was used for all models. We started with a learning rate of 0.0001 and gradually reduced it by half every 5000 iterations. We train the discriminator and the generator alternatively as is done for conventional GANs [10]. We use PyTorch [18] to implement our model.

In case of amplitude mask optimisation, data were simulated using the calibration matrices. We trained for around 10K iteration with a constant learning rate of 0.0001. SGD optimizer with momentum 0.5 was used for all models.

5.3 End to end network: Experiments & Results

We present a comparison of the performance of our method with that of other techniques.

Figure 5.2: Reconstruction of direct captured measurements using various approach.

Method	PSNR (in dB)	SSIM	Perceptual score
Tikhonov	10.95	0.33	2.25
TVAL3	11.81	0.36	3.38
Naive	18.90	0.62	5.72
Proposed-R	**19.06**	**0.62**	**5.86**
Proposed-T	**19.62**	**0.64**	**6.48**
Ground Truth	-	1	8.04

Table 5.1: PSNR, SSIM and perceptual score comparison for display captured measurements.

Figure 5.1 shows the comparison of our approach with the traditional and naive approach on some of the display captured images. In all figures and tables, Proposed-R refers to the model using random initialization and Proposed-T refers to the model using transpose initialization as explained in section 3.3.1. Inset images in figure 5.1 show the preservation of finer details in our approach. Figure 5.2 shows the comparison of our approach for direct captured measurements. Table 5.1 shows the quantitative evaluation of our approach. We use PSNR, SSIM and the no-reference image quality metric of Ma [15] for signal distortion and perception evaluation.

5.3.1 Resolving Dynamic Range Issues

For a highly multiplexed lensless imager like FlatCam, 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. Figure 5.3 show that, using our proposed reconstruction algorithm, the artifacts are minimized resulting in a higher quality reconstruction of the scene.

5.3.2 Ablation Studies

Here, we try a super-resolution network (RCAN) and a denoising network (DnCNN) in place of U-Net in the perceptual enhancement layer.

RCAN[29]: Residual Channel Attention Network is the state of the art for super-resolution. We use the model with 5 residual groups and 10 residual blocks.

DnCNN[28]: Denosing convolutional neural network is a popular widely used denoising network. Different from the existing discriminative denoising models which usually train a specific model for additive white Gaussian noise (AWGN) at a certain noise level, our DnCNN model is able to handle Gaussian denoising with unknown noise level (i.e., blind Gaussian denoising). In this work, we used a 19-layered model of the popular denoiser network DnCNN for comparison. The result comparisons are given in figure 5.4.

(a) Tikhonov (b) Naive (c) Proposed

Figure 5.3: Dynamic range issues. Arrows indicate the position of LED.

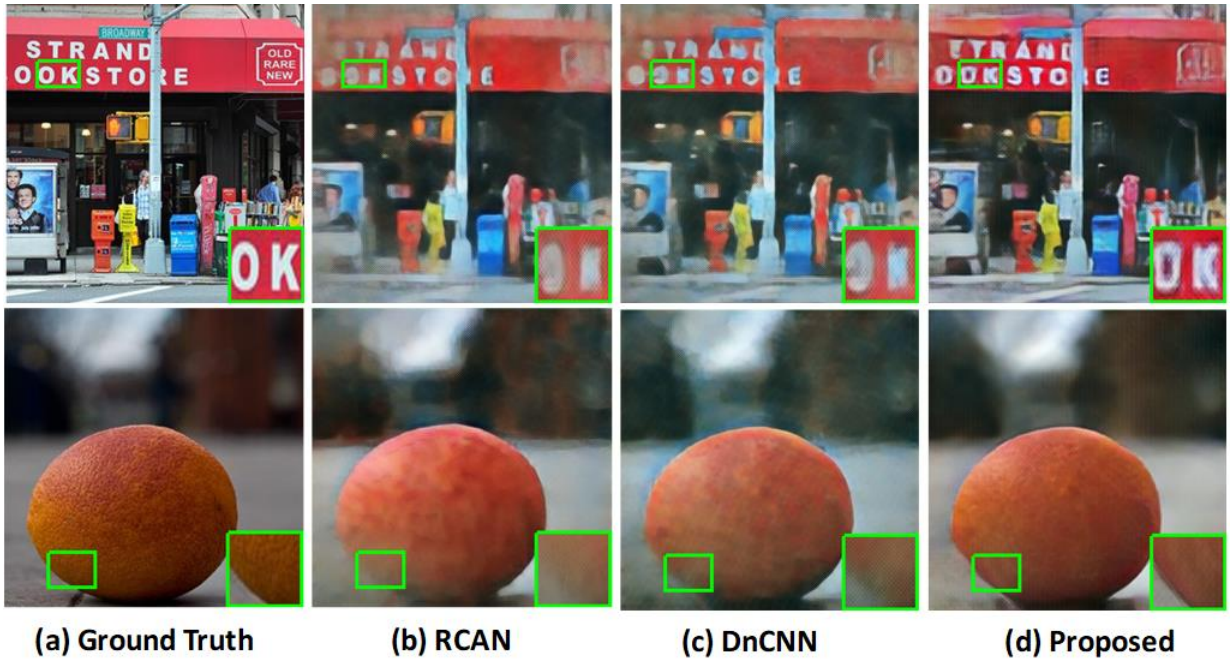

Figure 5.4: Comparison with RCAN and DnCNN in perceptual enhancement layer

5.4 Amplitude Mask Optimization

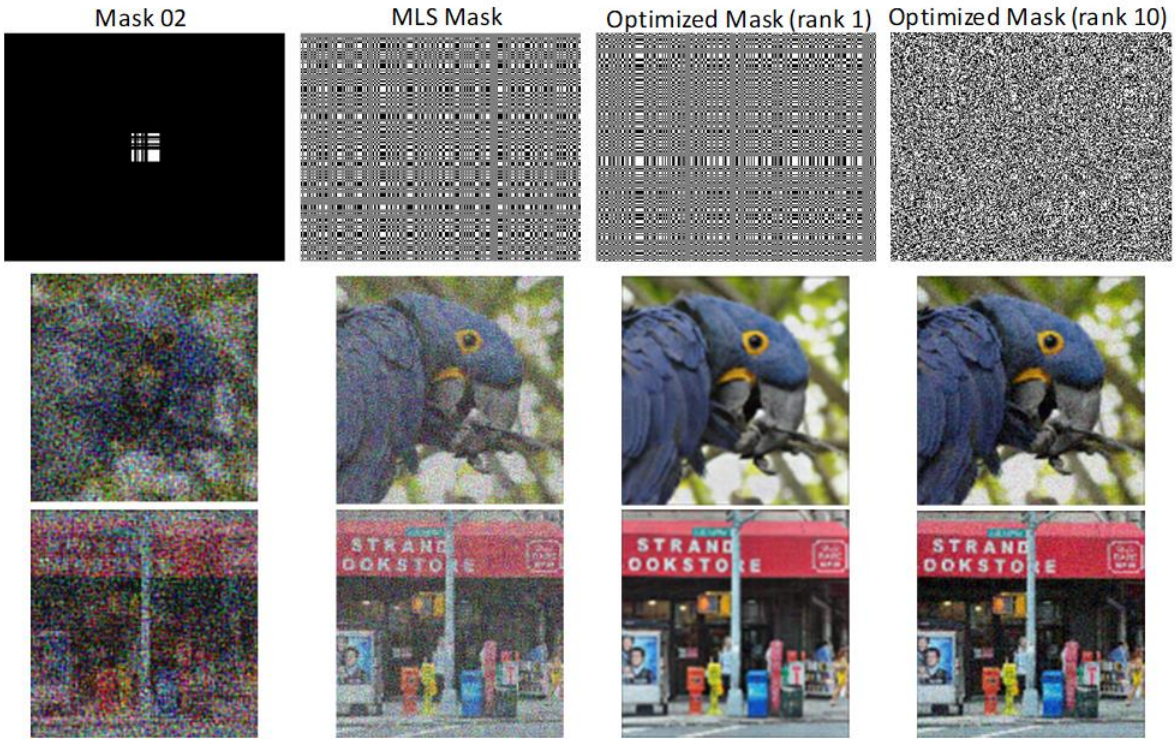

Figure 5.5: Comparison of various masks along with their reconstructions

Here we present a comparison with mask 02 and MLS mask. In all cases, we assume the distance between sensor and mask to be negligible. Our FlatCam prototype consists of a Point Grey Flea3 camera with 1.3 MP e2v EV76C560 CMOS sensor with a pixel

size of 5.3 μm and each element in mask is assumed to be 20 μm wide. We use SGD optimizer with momentum 0.5 for this model.

5.4.1 Mask 02

We generated this mask according to the specifications given in [5]. We used the following 31-element pattern: $1110001011100100001111111111111$, where 1 and 0 correspond to transparent and opaque mask features, respectively. We generated a 2-D separable mask by computing the outer product of the 31-element pattern with itself and appending additional zeros at the boundaries.

5.4.2 MLS Mask

We created the MLS mask using a 511- element M-sequence that consists of ± 1 entries. We computed the outer product to the pattern with itself and replaced every 1 entry with a 0 to obtain the mask.

5.4.3 Optimised Mask

Here, we use vectors of length 511 to initialise the mask. For rank 1 mask, we use two vectors of length while for rank 10 mask, we use two matrices of dimension 511×10 and 10×511. Since each mask element corresponds to four pixels in the sensor, we upscale the mask accordingly and then crop the required region which is used as the final mask, which is of size 500×620.

5.4.4 Comparison with existing masks

From the figure 5.5, it is clear that the optimised masks are more robust to noise as compared to the MLS mask and the 02 mask. When reconstructed using the weiner deconvolution with same parameters, it is seen that reconstructions are much better in the case of trained masks. Rank 1 mask performs better as compared to the rank 10 mask for the same number of iterations, since the latter has more number of parameters as compared to that of the former.

Method	PSNR (in dB)	Light Throughput
MLS Mask	18.68	0.501
Rank 1	22.40	0.508
Rank 10	21.20	0.507

Table 5.2: PSNR and Light Throughput comparison.

We also compared the light throughput of the mask as well as PSNR of the reconstructed images. Here, we can see that PSNR for the Rank 1 mask is almost 4 dB greater than that of MLS mask. Rank 10 mask gives the second best PNSR as well as light throughput as shown in the table 5.2.

CHAPTER 6

Conclusion and Future Works

The lensless imaging approach promises to challenge the traditional barriers of size, weight, cost, and performance in a broad range of applications spanning consumer, medical, scientific imaging, machine vision, and remote sensing. But standard optimization based methods currently used for reconstruction yield outputs that suffer from low resolution, low dynamic range and high noise sensitivity. In this work, two different approaches to improve the reconstructions are implemented. Using both mask optimization and end to end reconstruction approach, it is shown that the reconstructions can improved drastically from existing methods and the images can be made photo-realistic. Further research can be done in getting image reconstructions with mega-pixel resolution using efficient computational algorithms and intelligent mask design. Also, similar approaches can be implemented for phase mask based FlatCam as well. With these emerging applications, the future of lensless imaging indeed looks bright.

REFERENCES

[1] **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.

[2] **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.

[3] **Cannon, T.** and **E. Fenimore** (1980). Coded aperture imaging: many holes make light work. *Optical Engineering*, **19**(3), 193283.

[4] **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.

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

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

[7] **Dowski, E. R.** and **W. T. Cathey** (1995). Extended depth of field through wave-front coding. *Applied optics*, **34**(11), 1859–1866.

[8] **Durrant, P.**, **M. Dallimore**, **I. Jupp**, and **D. Ramsden** (1999). The application of pinhole and coded aperture imaging in the nuclear environment. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, **422**(1-3), 667–671.

[9] **Fenimore, E. E.** and **T. M. Cannon** (1978). Coded aperture imaging with uniformly redundant arrays. *Applied optics*, **17**(3), 337–347.

[10] **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.

[11] **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.

[12] **Khan, S.**, **V. R. Adarsh**, **V. Boominathan**, **J. Tan**, **A. Veeraraghavan**, and **K. Mitra**, Towards photorealistic reconstruction of highly multiplexed lensless images. *In The IEEE International Conference on Computer Vision (ICCV)*. Submitted 2019.

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

[14] **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.

[15] **Ma, C.**, **C.-Y. Yang**, **X. Yang**, and **M.-H. Yang** (2017). Learning a no-reference quality metric for single-image super-resolution. *Computer Vision and Image Understanding*, **158**, 1–16.

[16] **Maas, A. L.**, **A. Y. Hannun**, and **A. Y. Ng**, Rectifier nonlinearities improve neural network acoustic models. *In Proc. icml*, volume 30. 2013.

[17] **Marwah, K.**, **G. Wetzstein**, **Y. Bando**, and **R. Raskar** (2013). Compressive light field photography using overcomplete dictionaries and optimized projections. *ACM Transactions on Graphics (TOG)*, **32**(4), 46.

[18] **Paszke, A.**, **S. Gross**, **S. Chintala**, **G. Chanan**, **E. Yang**, **Z. DeVito**, **Z. Lin**, **A. Desmaison**, **L. Antiga**, and **A. Lerer** (2017). Automatic differentiation in pytorch.

[19] **Pavani, S. R. P.**, **M. A. Thompson**, **J. S. Biteen**, **S. J. Lord**, **N. Liu**, **R. J. Twieg**, **R. Piestun**, and **W. Moerner** (2009). Three-dimensional, single-molecule fluorescence imaging beyond the diffraction limit by using a double-helix point spread function. *Proceedings of the National Academy of Sciences*, **106**(9), 2995–2999.

[20] **Peng, Y.**, **Q. Fu**, **F. Heide**, and **W. Heidrich**, The diffractive achromat full spectrum computational imaging with diffractive optics. *In SIGGRAPH ASIA 2016 Virtual Reality meets Physical Reality: Modelling and Simulating Virtual Humans and Environments*. ACM, 2016.

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

[22] **Raskar, R.**, **A. Agrawal**, and **J. Tumblin**, Coded exposure photography: motion deblurring using fluttered shutter. *In ACM transactions on graphics (TOG)*, volume 25. ACM, 2006.

[23] **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.

[24] **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.

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

[26] **Sitzmann, V.**, **S. Diamond**, and **Y. Peng** (2018). End-to-end optimization of optics and image processing for achromatic extended depth of field and super-resolution imaging. *ACM Transactions on Graphics (TOG)*, **37**(4), 114.

[27] **Wang, A.**, **P. Gill**, and **A. Molnar** (2009). Angle sensitive pixels in cmos for lensless 3d imaging, 371–374.

[28] **Zhang, K.**, **W. Zuo**, **Y. Chen**, **D. Meng**, and **L. Zhang** (2017). Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising. *IEEE Transactions on Image Processing*, **26**(7), 3142–3155.

[29] **Zhang, Y.**, **K. Li**, **K. Li**, **L. Wang**, **B. Zhong**, and **Y. Fu**, Image super-resolution using very deep residual channel attention networks. *In Proceedings of the European Conference on Computer Vision (ECCV)*. 2018.

[30] **Zhou, C.** and **S. Nayar**, What are good apertures for defocus deblurring? *In 2009 IEEE international conference on computational photography (ICCP)*. IEEE, 2009.