UNIVERSITY OF CALIFORNIA
RIVERSIDE

Sensing Approaches and Reconstruction Methods for Improved Lensless Imaging

A Dissertation submitted in partial satisfaction
of the requirements for the degree of

Doctor of Philosophy

in

Electrical Engineering

by

Yucheng Zheng

June 2023

Dissertation Committee:

Dr. M. Salman Asif, Chairperson
Dr. Samet Oymak
Dr. Hyoseung Kim

The Dissertation of Yucheng Zheng is approved:

Committee Chairperson

University of California, Riverside

Acknowledgments

First and foremost, I would like to express my deepest gratitude to my advisor, Dr. Salman Asif, for his unwavering support throughout my entire program. Over the years, I have learned an immense amount from him, not only in terms of research and knowledge, but also in approaching problems with a more critical and logical mindset. Beyond research, Dr. Asif has consistently shown patience and friendliness since the beginning of the program, offering invaluable encouragement. I would like to thank Dr. Aswin Sankaranarayanan for his guidance and advice in my research and experiments. I am also grateful to Dr. Samet Oymak and Dr. Hyoseung Kim for their generous contributions to my PhD committee and their assistance with my dissertation and defense.

I am grateful to all of my friends, both at UCR and abroad, for their warm advice, support, and enjoyable conversations. Their companionship has made this journey far more enjoyable. Finally, I would like to express my special gratitude to my parents for their unconditional love, support, and encouragement at all times.

To my parents for their unconditional love and support.

ABSTRACT OF THE DISSERTATION

Sensing Approaches and Reconstruction Methods for Improved Lensless Imaging

by

Yucheng Zheng

Doctor of Philosophy, Graduate Program in Electrical Engineering
University of California, Riverside, June 2023
Dr. M. Salman Asif, Chairperson

Lensless imaging systems present a novel alternative to traditional camera designs by replacing the camera lens with a thin, flat, and potentially inexpensive optical encoder, often referred to as a coded mask. This dissertation establishes a framework for the lensless imaging model and explores computational algorithms to extract image and depth information from lensless imaging measurements. The dissertation is organized around three primary areas of focus:

Lensless Imaging for 3D Reconstruction Traditionally, estimating the disparity map of a scene requires either capturing a sequence of focus-stacked images or utilizing binocular views with a baseline. However, in lensless imaging, the optical encoder positioned above the camera sensor generates a complex and structured point spread function (PSF) that varies significantly depending on the depth of the light source. This unique property enables the retrieval of depth information from a single frame of measurements. In order to estimate depth maps more accurately on a continuous domain, we formulate the forward imaging model of the system, such that the PSF is represented as a continuous function of depth. Based on this imaging model, we design an alternating update algorithm to jointly estimate the RGB and depth information from the measurements.

Improved 3D Lensless Imaging With Learned Programmable Masks Existing methods for lensless imaging can recover both depth and intensity of a scene, but they necessitate solving computationally-intensive inverse problems. Additionally, these methods encounter difficulties when attempting to recover dense scenes with significant depth variations. To address these challenges, we propose a lensless imaging system that captures a limited number of measurements using distinct patterns on a programmable mask. We introduce a fast recovery algorithm, which can be parallelized, to recover textures on a fixed number of depth planes from multiple measurements. Furthermore, we examine the mask design problem for programmable lensless cameras and offer a design template aimed at optimizing mask patterns to enhance depth estimation.

Coded Illumination for Improved Lensless Imaging In space-limited situations, such as under-display imaging, the imaging system is often ill-conditioned, and the number of sensing elements may be constrained. However, by employing coded illumination patterns, we can capture more potentially uncorrelated measurements, thereby improving the conditioning of the original linear system. In the proposed framework, the scene is illuminated by a sequence of simple coded illumination patterns while the lensless camera captures sensor measurements. We introduce a fast and low-complexity recovery algorithm that leverages the properties of the illumination patterns and the system. Ultimately, the images reconstructed by combining measurements from the illuminated scene demonstrate a significant improvement over the original camera system.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Lensless cameras provide novel designs for extreme imaging conditions that require small, thin form factor, large field-of-view, or large-area sensors [3, 6, 12, 14]. Compared to conventional lens-based cameras, lensless cameras can be flat, thin, light-weight, and potentially flexible because the physical constraints imposed by a lens are relaxed. In lensless imaging, the measurements formed on the sensor with a coded mask is a linear superposition of shifted and scaled versions of the point spread function. To recover the scene image from the sensor measurements, we need to solve a linear inverse problem. The quality of the recovered image depends on the conditioning of the linear system; especially in the absence of any prior knowledge about the scene.

The main features of lensless camera are its small-size, light weight and scalable FOV depending only on the area of sensor, such flat lensless camera has wide potential application under space-limited circumstances, such as under-display sensing, microscopy, or installed on AR/VR device. In addition to the physical advantages, due to its complex and structured point spread function pattern, lensless camera is amenable to compressive sensing approaches. In contrast to a traditional

lens-based camera, the point spread function of our proposed lensless camera is more complex and structured. The PSF shifts when the point moves parallel to the sensor plane and expands/shrinks when the point source moves toward/away from the sensor plane. The measurements recorded on the sensor thus represent a superposition of shifted and scaled versions of the mask shadows corresponding to light sources in different directions and depths. One main track of this dissertation focuses on exploiting this property and retrieving RGB and depth information from lensless measurements.

Despite its advantages in space-saving, FOV, and compressive sensing, lensless camera suffer from its ill-conditioned system and low light throughput due to the lack of lens. Therefore, the imaging quality of lensless camera is often poor compared to conventional camera, especially in constrained conditions, such as space-limited applications where the mask is put very close the sensor or the number of sensing elements is constrained. Fortunately, we may improve such situation with the assistance of external light source. The second main track of this dissertation focuses on improving the 2D and 3D imaging quality with the assistance of a sequence of coded illumination patterns.

This dissertation is organized as follows, we develop the a general framework and imaging model for RGB and depth estimation from single-frame lensless measurements in Chapter 2. To further improve the depth quality, we propose in Chapter 3 to capture measurements with a sequence of learned programmable mask patterns. To improve the stability and imaging quality of lensless camera, we propose a novel imaging model with the application of coded illumination patterns in Chapter 4. In the end, we also explore its application in 3D reconstruction in Chapter 5.

1.1 Lensless Imaging for 3D Reconstruction

Depth estimation is an important and challenging problem that arises in a variety of applications including computer vision, robotics, and autonomous systems. Existing depth estimation systems use stereo pairs of conventional (lens-based) cameras or time-of-flight sensors [44, 45, 88]. These cameras can be heavy, bulky, and require large space for their installation. Therefore, their adoption in portable and lightweight devices with strict physical constraints is still limited.

In fact, depth estimation is an integral part of lensless imaging. The basic principle of a lensless camera is that the sensor records the summation of the measurements associated with the scene points. Each scene point casts its own unique sensor measurement, depending on its three-dimensional (3D) spatial location and radiance. Thus, the 3D scene information is encoded in the sensor measurements, but its recovery requires us to solve a nonlinear inverse problem. Further, it is important to model this depth dependence in the measurements especially since ignoring it results in significant reduction in the quality of reconstructions.

1.1.1 Joint Image and Depth Estimation From Single-Frame Lensless Measurements

Despite the nature of lensless camera in encoding 3D information with its measurements, existing depth recovery algorithms either assume that the scene consists of a small number of depth planes or solve a sparse recovery problem over a large 3D volume. Both these approaches fail to recover the scenes with large depth variations. In chapter 2, we propose a new approach for depth estimation based on an alternating gradient descent algorithm that jointly estimates a continuous depth map and light distribution of the unknown scene from its lensless measurements.

To accurately estimate the depth of the scene pixels, we exploit the relationship between the depth and PSF and formulate the imaging model as following

$$\mathbf{y} = \mathbf{\Psi}(\boldsymbol{\alpha})\mathbf{l}, \tag{1.1}$$

where $\mathbf{y} \in \mathbb{R}^{M^2}$ is denotes sensor measurements, $\mathbf{l} \in \mathbb{R}^{N^2}$ represents RGB intensity from all the angles, and $\mathbf{\Psi}$ is a matrix with all the basis functions parameterized by the inverse depth α.

To jointly estimate the depth and light distribution, we propose a two-step approach that consists of an initialization step and an alternating gradient descent step to minimize our objective. To preserve sharp edges in the image intensity and depth map, we include an adaptive regularization penalty in our objective function.

1.1.2 Improved 3D Lensless Imaging With Learned Programmable Masks

Existing methods for 3D lensless imaging can be divided into two categories. One category of methods estimate the 3D scene with a single image measurement [1, 3, 5, 13, 122]. These methods jointly estimate the image and depth map of a 3D scene by solving an optimization problem using iterative techniques. Since the number of variables to be estimated is much larger than the number of measurements, the recovery problem is severely under-determined and the methods rely on some prior knowledge about the 3D scene. For instance, [3] assumes that 3D volume is sparse and solves an ℓ_1 norm-based optimization problem to estimate a 3D image. Another drawback of these optimization-based methods is their large computational cost and run time. The second category of methods capture multiple measurements, each with a different mask, which makes the 3D recovery of dense scenes possible [48, 114].

In chapter 3, we propose a novel framework that reconstructs the 3D scene using lensless measurements captured with multiple learned programmable masks. The most related work that concerns designing mask patterns for multiple measurements is SweepCam [48], which captures multiple measurements of the same scene using a programmable, shifting mask and estimates one plane in the 3D scene at a time. The recovery of a single plane is much faster than joint recovery of the entire 3D scene, but the number of mask patterns needed to achieve artifact-free reconstruction of a single depth plane can be large (in the range of 100mm–400mm). Our proposed method falls in the second category, but instead of estimating a single plane, we recover the 3D scene at multiple depth planes jointly from a smaller number of measurements, using a simple algorithm.

The proposed algorithm exploits the well-known fact that convolutional systems are diagonalized in the Fourier domain. Given that the measurement operator associated with each scene depth is well-approximated as a convolution, we show that the joint depth and image recovery problem can be significantly simplified when we model the scene intensities in the spatial frequency domain. Hence, instead of solving a single large linear system for image intensity across all of the depth planes, we solve several small linear systems in parallel, with each system associated with a different frequency coefficient.

1.2 Lensless Imaging With Coded Illumination

Despite recent progress in lensless cameras, the quality of images recovered from the lensless cameras is often poor due to the ill-conditioning of the underlying measurement system, especially in space-limited applications where the mask is put very close the sensor or the number of sensing elements is constrained.

1.2.1 Coded Illumination for Improved Lensless Imaging

In chapter 4, we propose a novel framework that combines coded illumination with mask-based lensless cameras to improve the quality of recovered images. A 2D target scene is illuminated with multiple coded patterns during data acquisition. We capture sensor measurements for each illumination pattern and use a fast recovery algorithm to reconstruct the scene using all the measurements.

We can formulate the sensor measurements $Y_{i,j}$ for each of the projected illumination pattern $P_{i,j}$ as the following

$$Y_{i,j} = \Phi_L(P_{i,j} \odot X)\Phi_R^\top + E_{i,j}, \tag{1.2}$$

where Φ_L, Φ_R are forward imaging matrices, X is the target image, $\odot$ denotes element-wise multiplication operator and $E_{i,j}$ denotes measurement noise. By exploiting the separability of the mask and illumination pattern, we can solve for the underlying X in a fast and low-complexity closed-form solution.

1.2.2 Binocular 3D Lensless Imaging With Coded Illumination

Existing methods for 3D reconstruction from lensless measurements suffer from poor spatial and depth resolution. This is partially due to the system ill conditioning that arises because the point-spread functions (PSFs) from different depth planes are very similar. In this chapter, we propose to capture multiple measurements of the scene under a sequence of coded illumination

patterns to improve the 3D image reconstruction quality. In addition, we put the illumination source at a distance away from the camera. With such baseline distance between the lensless camera and illumination source, the camera observes a slice of the 3D volume, and the PSF of each depth plane becomes more resolvable from each other.

Chapter 2

Lensless Imaging for 3D Reconstruction

Mask-based lensless cameras replace the lens of a conventional camera with a custom mask. These cameras can potentially be very thin and even flexible. Recently, it has been demonstrated that such mask-based cameras can recover light intensity and depth information of a scene. Existing depth recovery algorithms either assume that the scene consists of a small number of depth planes or solve a sparse recovery problem over a large 3D volume. Both these approaches fail to recover the scenes with large depth variations. In this chapter, we propose a new approach for depth estimation based on an alternating gradient descent algorithm that jointly estimates a continuous depth map and light distribution of the unknown scene from its lensless measurements. We present simulation results on image and depth reconstruction for a variety of 3D test scenes. A comparison between the proposed algorithm and other method shows that our algorithm is more robust for natural scenes with a large range of depths. We built a prototype lensless camera and present experimental results for reconstruction of intensity and depth maps of different real objects.

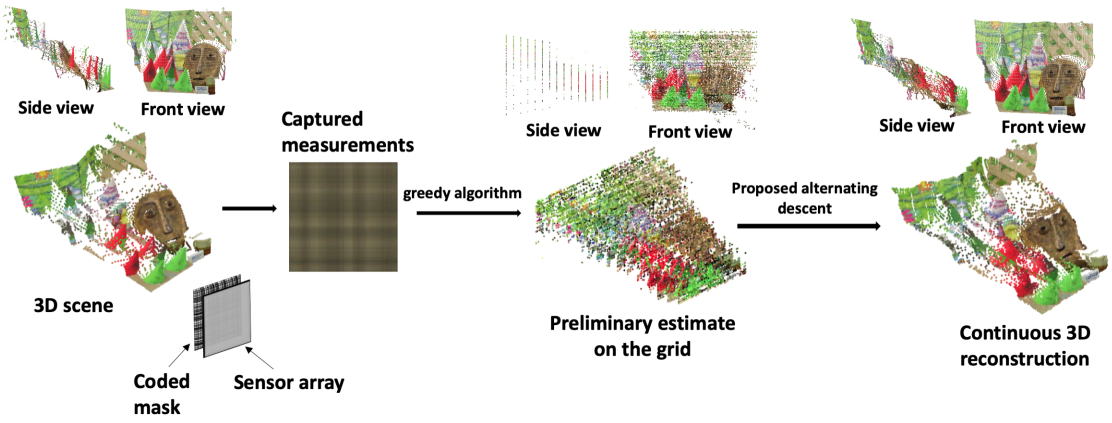


Figure 2.1: An overview of the proposed intensity and depth estimation framework. Consider a natural scene as a 3D point cloud, where each point represents a light source located at a different depth. The camera consists of a fixed, coded mask placed on top of an image sensor. Every point in the scene casts a shadow of the mask on the sensor plane. Each sensor pixels records a linear combination of the scene modulated by the mask pattern. The recovery algorithm consists of two steps. (1) Initialization using a greedy depth selection method. (2) An alternating gradient descent-based refinement algorithm that jointly estimates the light distribution and depth map on a continuous domain.

2.1 Introduction

An overview of the reconstruction framework is illustrated in Figure 2.1. In this chapter, we use the same sampling framework proposed in [5]. We initialize the estimate of the depth map by selecting a single plane or solving the greedy algorithm proposed in [5]. The greedy method assumes that the scene consists of a small number of depth planes and fails to recover scene with continuous depth variations. The method proposed in this chapter can estimate continuous depth by minimizing an objective function with respect to image intensity and depth via alternating gradient descent. We present extensive simulation and real experimental results with different objects. The main contributions of this chapter are as follows.

- We propose a new computational framework for joint estimation of light intensity and depth maps from a single image of a mask-based lensless camera. In contrast to other methods, our method estimates the depth map on a continuous domain. Our algorithm consists of a careful initialization step based on greedy pursuit and an alternating minimization step based on gradient descent.
- The problem of joint image and depth recovery is highly nonconvex. To tackle this issue, we present different regularization schemes that offer robust recovery on a diverse dataset.
- We present simulation results on standard 3D datasets and demonstrated a significant improvement over existing methods for 3D imaging using coded mask-based lensless cameras.
- We built a hardware prototype to capture measurements of real objects. We present image and depth reconstruction results of these real objects using our proposed algorithm and a comparison with existing methods.

2.2 Background and Related Work

A pinhole camera, also known as *camera obscura*, is the simplest example of a mask-based lensless camera. Even though a pinhole can easily provide an image of the scene onto a sensor plane, the image quality is often severely affected by noise because the amount of light collected is limited by the pinhole aperture [118]. Coded aperture-based lensless cameras avoid this problem by increasing the number of pinholes and allowing more light to reach the sensor [3, 6, 12, 19, 23, 34]. In contrast to a pinhole camera where only one inverted image of the scene is obtained through a single pinhole, the measurements captured through a coded-mask are a linear combination of all the pinhole images under every mask element. To recover an image of the scene, we need to solve a computational image recovery problem [3, 6, 34].

Recent work on mask-based lensless imaging broadly falls into two categories. FlatCam [5] uses a separable mask aligned with the sensor such that the sensor measurements corresponding to a plane at a fixed depth from the sensor can be written as a separable system. DiffuserCam [3] assumes that the mask size and angular span of the object are small enough so that the sensor measurements of a plane can be modeled as a convolution of the mask pattern with image intensity at that plane. The convolutional model can be computationally efficient if the object falls within a small angular range because we can use fast Fourier transform to compute convolutions. The separable model does not require a small angular range assumption. A number of methods based on deep learning have also been developed recently for both separable and convolutional imaging models to recover images at a fixed depth plane [30, 52, 73].

A coded aperture system offers another advantage by encoding light from different directions and depths differently. The depth-dependent imaging capability in coded aperture systems is known since the pioneering work in this domain [10, 34]. However, the classical methods usually assume that the scene consists of a single plane at known depth. In this chapter, we assume that the depth map is arbitrarily distributed on a continuous domain and the true depth map is unknown at the time of reconstruction.

The 3D lensless imaging problem has also recently been studied in [1, 3, 5, 48]. These methods can broadly be divided into two categories. In the first category, the 3D scene is divided into a finite number of voxels. To recover the 3D light distribution, these methods solve an ℓ_1 norm-based recovery problem under the assumption that the scene is very sparse [1, 3]. In the second category, the 3D scene is divided into an intensity map and multiple depth planes such that each pixel is assigned one intensity and depth. To solve the intensity and depth recovery problem, these

methods either sweep through the depth planes [48] or assign depth to each pixel using a greedy method [5]. Our proposed method belongs to the second category in which we model the image intensity and depth separately and assume that the depth values of the scene are distributed on a continuous domain. To recover the 3D scene, we jointly estimate the image intensity and depth map from the available sensor measurements.

Joint estimation of image intensity and depth map can be viewed as a nonlinear inverse problem in which the sampling function is dependent on scene depth. Similar inverse problem also arises in many other fields such as direction-of-arrival estimation in radar [102], super-resolution [17] and compressed sensing [9, 76, 107]. Similar to the joint estimation of image intensity and depth, the solution approaches to these problems consists of two main steps: identification of signal bases and the estimation of signal intensities based on the identified bases. The problem of identifying the signal bases from continuously varying candidates is often called off-the-grid signal recovery. The methods for solving the off-the-grid signal recovery problems can be divided into two main types. The first approach formulates the problem as a convex program on a continuous domain and solves it using an atomic norm minimization approach [27, 103]. The second approach linearizes the problem for the optimization parameter using a first-order approximation at every iteration [17, 116]. Our proposed algorithm is inspired by the second approach.

Mask-based lensless cameras have traditionally been used for imaging light at wavelengths beyond the visible spectrum [19, 23]. Other examples related to mask-based cameras include controllable aperture and employing coded-mask for compressed sensing and computational imaging [100, 127], distributed lensless camera [121], single pixel camera [32] and external mask setting [82].

Coded masks have also recently been used with conventional lens-based cameras to estimate depth and lightfield [47, 56, 66, 108]. Recently, a number of data-driven methods have been proposed to design custom phase masks and optical elements to estimate depth from a single image [28, 113]. An all-optical diffractive deep neural network is proposed in [60, 67], which can perform pattern recognition tasks such as handwritten digits classification using optical mask layers. Such networks can literally process images at a lightning-fast pace with near-zero energy cost.

2.3 Forward Lensless Imaging Model

We divide the 3D scene under observation into $N \times N$ uniformly spaced directions. We use θ_i and θ_j to denote the angular directions of a light source with respect to the center of the sensor. The intensity and depth of the light source are denoted using $l_{i,j}$ and $z_{i,j}$ respectively. Figure 2.2 depicts the geometry of such an imaging model. A planar coded-mask is placed on top of a planar sensor array at distance d. The $M \times M$ sensor array captures lights coming from the scene modulated by the coded-mask.

Every light source in the scene casts a shadow of the mask on the sensor array, which we denote using basis functions ψ. We use s_u and s_v to index a pixel on the rectangular sensor array. The shadow cast by a light source with unit intensity at $(\theta_i, \theta_j, z_{i,j})$ can be represented as the following basis or point spread function:

$$\psi_{i,j}(s_u, s_v) = \text{mask}\left[\alpha_{i,j} s_u + d\tan(\theta_i), \alpha_{i,j} s_v + d\tan(\theta_j)\right], \tag{2.1}$$

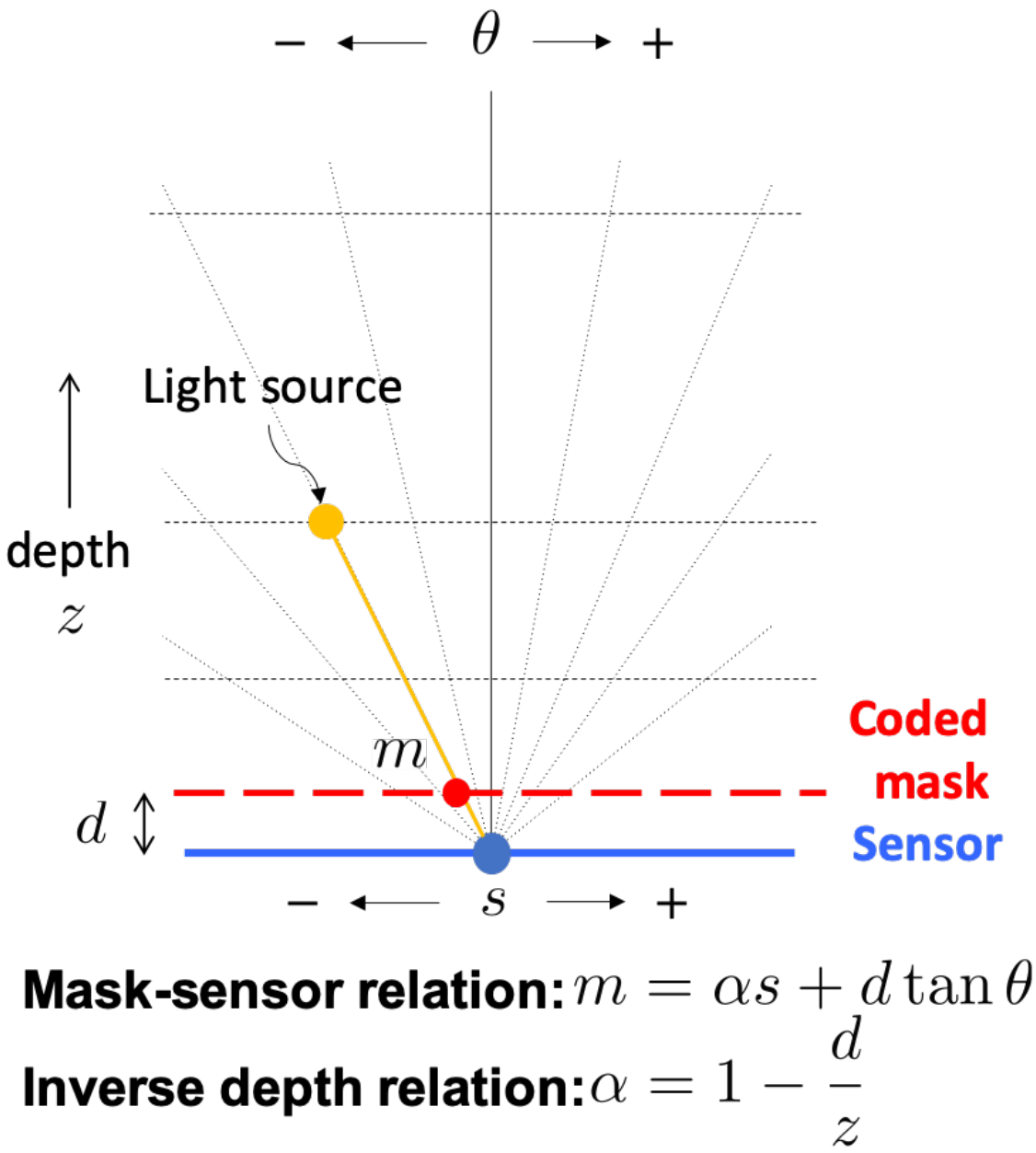


Figure 2.2: 1D imaging model for a planar sensor with a coded mask placed at distance d. Light rays from a light source at location (θ, z) are received by all the sensor pixels. A light ray that hits sensor pixel s passes through mask at location m.

where mask$[u, v]$ denotes the transmittance of the mask pattern at location (u, v) on the mask plane and $\alpha_{i,j}$ is a variable that is related to the physical depth $z_{i,j}$ with the following inverse relation:

$$\alpha_{i,j} = 1 - \frac{d}{z_{i,j}}, \tag{2.2}$$

If the 3D scene consists of only a single point source at (θ_i, θ_j) with light intensity $l_{i,j}$, the measurement captured at sensor pixel (s_u, s_v) would be

$$y(s_u, s_v) = \psi_{i,j}(s_u, s_v) l_{i,j}. \tag{2.3}$$

The measurement recorded on any sensor pixel is the summation of contributions from each of the point sources in the 3D scene. The imaging model for a single sensor pixel can be represented by

$$y(s_u, s_v) = \sum_{i=1}^{N} \sum_{j=1}^{N} \psi_{i,j}(s_u, s_v) l_{i,j}. \tag{2.4}$$

We can write the imaging model for the entire sensor in a compact form as

$$\mathbf{y} = \mathbf{\Psi}(\boldsymbol{\alpha})\mathbf{l} + e, \tag{2.5}$$

where $\mathbf{y} \in \mathbb{R}^{M^2}$ is a vectorized form of an $M \times M$ matrix that denotes sensor measurements, $\mathbf{l} \in \mathbb{R}^{N^2}$ is a vectorized form of an $N \times N$ matrix that denotes light intensity from all the locations $(\theta_i, \theta_j, \alpha_{i,j})$, and $\mathbf{\Psi}$ is a matrix with all the basis functions corresponding to $\theta_i, \theta_j, \alpha_{i,j}$. The basis functions in (2.5) are parameterized by the unknown $\boldsymbol{\alpha} \in \mathbb{R}^{N^2}$ and e denotes noise and other nonidealities in the system.

We can jointly estimate light distribution ($\mathbf{l}$) and inverse depth map ($\boldsymbol{\alpha}$)[1] using the following optimization problem:

$$\underset{\boldsymbol{\alpha},\mathbf{l}}{\text{minimize}} \ \frac{1}{2}\|\mathbf{y} - \boldsymbol{\Psi}(\boldsymbol{\alpha})\mathbf{l}\|_2^2. \tag{2.6}$$

Note that if we know the true values of $\boldsymbol{\alpha}$ (or we fix it to something), then the problem in (2.6) reduces to a linear least-squares problem that can be efficiently solved via standard solvers. On the other hand, if we fix the value of $\mathbf{l}$, the problem remains nonlinear with respect to $\boldsymbol{\alpha}$. In the next few sections we discuss our approach for solving the problem in (2.6) via alternating minimization.

2.4 Technical Approaches

2.4.1 Matching Pursuit Algorithm

Since the minimization problem in (2.6) is not convex, a proper initialization is often needed to ensure convergence to a local minimum close to the optimal point. non-convex and non-smooth because of its dependence on the depth. The mask pattern ψ we use in (2.1) is not necessarily differentiable with respect to depth either. Therefore, solving the optimization problem in (2.6) from scratch without a good initialization will be difficult. To overcome this problem, we need a good preliminary estimate for initialization.

[1] $\boldsymbol{\alpha}$ has an inverse relation with the depth map (2.2); therefore we refer to it as inverse depth map throughout the paper.

Depth Scanning A naïve approach is to initialize all the point sources in the scene at the same depth plane. To select an initial depth plane, we sweep through a set of candidate depth planes and perform image reconstruction on one depth plane at a time by solving the following linear least squares problem:

$$\underset{\mathbf{l}}{\text{minimize}} \ \frac{1}{2}\|\mathbf{y} - \mathbf{\Psi}(\boldsymbol{\alpha})\mathbf{l}\|_2^2. \tag{2.7}$$

We evaluate the loss value for all the candidate depth planes and picked the one with the smallest loss as our initialized depth. The mask basis function in (2.1) changes as we change $\boldsymbol{\alpha}$, which has an inverse relation with the scene depth. We select candidate depth corresponding to uniformly sampled values of $\boldsymbol{\alpha}$, which yields non-uniform sampling of the physical scene depth. The single-depth initialization approach is computationally simple and provides a reasonable initialization of light distribution to start with, especially when the scene is far from the sensor.

Matching Pursuit Our second approach for initialization is the greedy method proposed in [5]. Greedy algorithms are widely used for sparse signal recovery [9,76,107]. Based on these algorithms, [5] proposed a greedy depth pursuit algorithm for depth estimation from FlatCam [6]. The algorithm works by iteratively updating the depth surface that matches the observed measurements the best.

The depth pursuit method assumes that the scene consists of a small number of predefined depth planes. We start the program by initializing all the pixels at a single depth plane and the estimation of light intensities $\mathbf{l}$ based on the initialized depth map. The first step is to select new candidate values for $\boldsymbol{\alpha}$. The new candidates are selected using the basis vectors that are mostly correlated with the current residual of the estimate. In the second step, new candidates for $\boldsymbol{\alpha}$ are

appended to the current estimate. We solve a least squares problem using the appended $\boldsymbol{\alpha}$. In the third step, we prune the $\boldsymbol{\alpha}$ by selecting $\alpha_{i,j}$ as the value corresponding to the largest magnitude of $\mathbf{l}_{i,j}$. Although this method may not estimate the off-grid point sources well, it produces a good preliminary estimate of the scene.

2.4.2 Alternating Gradient Descent for Continuous Image and Depth Estimation

To solve the minimization problem in (2.6), we start with the preliminary image and depth estimates from the initialization step and alternately update depth and light distribution via gradient descent. The main computational task in gradient descent method is computing the gradient of the loss function w.r.t. $\boldsymbol{\alpha}$. To compute that gradient, we expand the loss function in (2.6) as

$$L = \frac{1}{2} \sum_{u,v=1}^{M} (y(s_u, s_v) - \sum_{i,j=1}^{N} \psi_{i,j}(s_u, s_v) l_{i,j})^2 \tag{2.8}$$

We define $R_{u,v} = y(s_u, s_v) - \sum_{i,j=1}^{N} \psi_{i,j}(s_u, s_v) l_{i,j}$ as the residual approximation error at location (s_u, s_v). The derivatives of the loss function with respect to the $\alpha_{i,j}$ is given as

$$\frac{\partial L}{\partial \alpha_{i,j}} = \sum_{u,v=1}^{M} R_{u,v} \frac{\partial R_{u,v}}{\partial \alpha_{i,j}} = -l_{i,j} \sum_{u,v=1}^{M} R_{u,v} \frac{\partial \psi_{i,j}(s_u, s_v)}{\partial \alpha_{i,j}}. \tag{2.9}$$

We compute the derivatives of sensor value with respect to the $\alpha_{i,j}$ using the total derivative [2] as follows.

$$\begin{aligned}\frac{\partial \psi_{i,j}(s_u, s_v)}{\partial \alpha_{i,j}} &= \frac{\partial \psi_{i,j}(s_u, s_v)}{\partial u_{i,j}} \frac{\partial u_{i,j}}{\partial \alpha_{i,j}} + \frac{\partial \psi_{i,j}(s_u, s_v)}{\partial v_{i,j}} \frac{\partial v_{i,j}}{\partial \alpha_{i,j}} \\ &= \frac{\partial \psi_{i,j}(s_u, s_v)}{\partial u_{i,j}} s_u + \frac{\partial \psi_{i,j}(s_u, s_v)}{\partial v_{i,j}} s_v. \end{aligned} \tag{2.10}$$

$u_{i,j} = \alpha_{i,j} s_u + d\tan(\theta_i)$ and $v_{i,j} = \alpha_{i,j} s_v + d\tan(\theta_j)$ denote two dummy variables that also correspond to the specific location on the mask where a light ray from a point source at angle (θ_i, θ_j) and depth $\alpha_{i,j}$ and sensor pixel at (s_u, s_v) intersects with the mask plane. The terms in $\frac{\partial \psi_{i,j}(s_u, s_v)}{\partial u_{i,j}}, \frac{\partial \psi_{i,j}(s_u, s_v)}{\partial v_{i,j}}$ can be viewed as the derivatives of mask pattern along the respective spatial coordinates and evaluated at $u_{i,j}, v_{i,j}$. We compute these derivatives using finite-difference of $\psi_{i,j}(s_u, s_v)$ over a fine grid and linear interpolation.

2.4.3 Algorithm Analysis

To solve the non-linear least squares problem in (2.6) in our algorithms, we compute the gradient derived in (2.10) and use it as input of a optimization solver. Suppose ψ_i and ψ_j denote the basis function vectors evaluated on a 1D mask as

$$\begin{aligned}\psi_i(s_u) &= \text{mask}\left[\alpha_{i,j} s_u + d\tan(\theta_i)\right] \\ \psi_j(s_v) &= \text{mask}\left[\alpha_{i,j} s_v + d\tan(\theta_j)\right]. \end{aligned} \tag{2.11}$$

[2] Recall that the total derivative of a multivariate function $f(x, y)$ is $\frac{\partial f(x,y)}{\partial x} dx + \frac{\partial f(x,y)}{\partial y} dy$.

If we use a separable mask pattern, then the 2D mask function $\psi_{i,j}$ in (2.1) can be computed as the outer product of two vectors given as $\psi_{i,j} = \psi_i \psi_j^T$. Similarly, we define 1D sub-gradient function g as

$$g_i(s_u) = \frac{\partial \psi_{i,j}(s_u, s_v)}{\partial u_{i,j}}$$
$$g_j(s_v) = \frac{\partial \psi_{i,j}(s_u, s_v)}{\partial v_{i,j}}, \tag{2.12}$$

Similar to (2.10), the functions $\frac{\partial \psi_{i,j}(s_u,s_v)}{\partial u_{i,j}}$ and $\frac{\partial \psi_{i,j}(s_u,s_v)}{\partial v_{i,j}}$ are the sub-gradient functions along the 1D mask. It takes non-negative values at locations where mask pattern value changes and takes zero value at the other places. Using the derivation in (2.10), the matrix contains $\frac{\partial \psi_{i,j}(s_u,s_v)}{\partial \alpha_{i,j}}$ at all (s_u, s_v) can be computed using the following sum of two vector outer products.

$$\frac{\partial \psi_{i,j}}{\partial \alpha_{i,j}} = g_i \psi_j^T + \psi_i g_j^T \tag{2.13}$$

Using the derivations in (2.9), the derivative of loss function with respect to depth value can be computed using the following matrix multiplications, where R refers to the matrix of residual $R_{u,v}$ at all (s_u, s_v)

$$\frac{\partial L}{\partial \alpha_{i,j}} = g_i^T R \psi_j + \psi_i^T R g_j \tag{2.14}$$

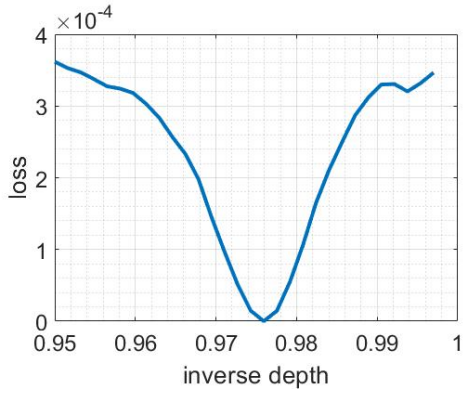
Suppose we have $M \times M$ pixels on sensor array. The computation in (2.14) takes $2M^2 + 2M$ multiplications. We then feed our gradients to `minfunc` solver [93] with L-BFGS algorithm [61] to solve the non-linear optimization problem in (2.6).

2.4.4 Regularization Approaches

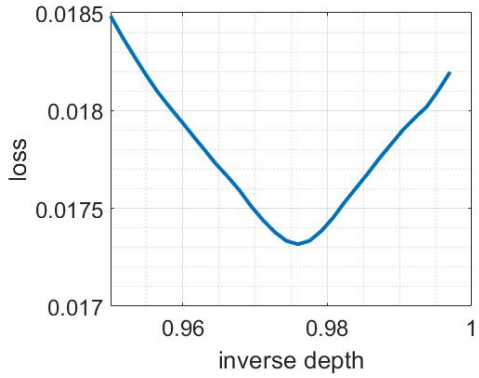
ℓ_2 Regularization on Spatial Gradients. The optimization problem in (2.6) is highly non-convex and contains several local minima; therefore, the estimate often gets stuck in some local minima and the estimated intensity and depth maps are coarse. To improve the performance of our algorithm for solving the non-convex problem in (2.6), we seek to exploit additional structures in the scene. A standard assumption is that the depth of neighboring pixels is usually close, which implies that the spatial differences of (inverse) depth map are small. To incorporate this assumption in our model, we add a quadratic regularization term on the spatial gradients of the inverse depth map to our loss function. The quadratic regularization term is defined on an $N \times N$ inverse depth map matrix $\boldsymbol{\alpha}$ and can be written as

$$\begin{aligned} R(\boldsymbol{\alpha}) &= \sum_{i,j=1}^{N} (\alpha_{i,j} - \alpha_{i+1,j})^2 + (\alpha_{i,j} - \alpha_{i,j+1})^2 \\ &= \|\nabla_r \boldsymbol{\alpha}\|_F^2 + \|\nabla_c \boldsymbol{\alpha}\|_F^2, \end{aligned} \tag{2.15}$$

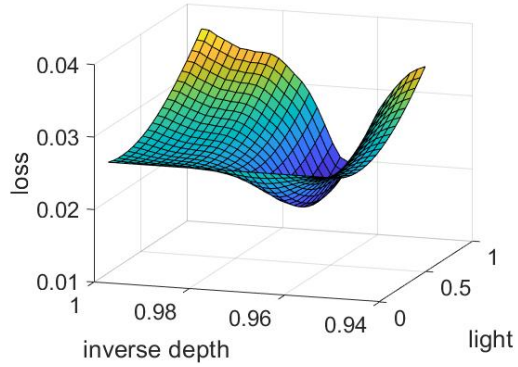
where the operators ∇_r, ∇_c compute spatial differences along rows and columns, respectively. We call this regularization an ℓ_2 norm-based total variation (TV-ℓ_2) in this chapter. Figure 2.3 illustrates the effect of the depth regularization. From Figure 2.3, we observe that smoothness regularization improves the loss function by removing several local minima. We also observed this effect in our simulations for a high-dimensional depth recovery problem, which is not very sensitive to initialization with depth regularization.



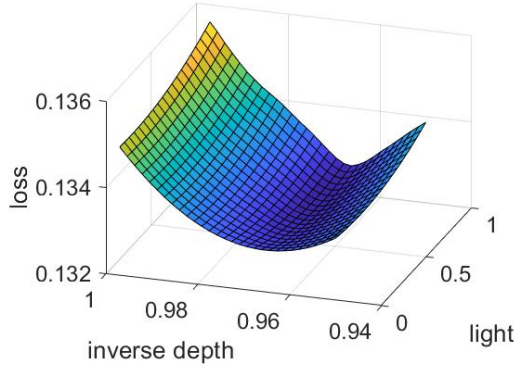
(a) Without smooth regularization, the loss curve is highly non-convex and contains several local minima.



(b) With the smooth regularization, the loss curve is smooth and several local minima are removed.



(c) Similar to 1D case, loss surface contains many local minima without smooth regularization.



(d) With smooth regularization, many local minima are removed from loss surface.

Figure 2.3: A comparison between objective loss functions without and with smooth regularization. The inverse depth axis refers to the value of α.

Weighted ℓ_2 Regularization on Spatial Gradients. Even though smoothness regularization on the inverse depth map removes some local minima and helps with converge, it does not respect the sharp edges in the depth map. To preserve sharp discontinuities in the (inverse) depth map, we used the following adaptive weighted regularization inspired from [62]:

$$R_W(\boldsymbol{\alpha}) = \sum_{i,j=1}^{N} W^c_{i,j}(\alpha_{i,j} - \alpha_{i+1,j})^2 + W^r_{i,j}(\alpha_{i,j} - \alpha_{i,j+1})^2, \tag{2.16}$$

where $W_{i,j}^{r,\alpha}$ and $W_{i,j}^{c,\alpha}$ denote weights for row and column differences, respectively. We aim to select these weights to promote depth similarity for neighboring pixels, but avoid smoothing the sharp edges. To promote this, we selected weights with exponential decay in our experiments that we compute as

$$W_{i,j}^{r} = \exp\left(-\frac{(\alpha_{i,j} - \alpha_{i+1,j})^2}{\sigma}\right)$$
$$W_{i,j}^{c} = \exp\left(-\frac{(\alpha_{i,j} - \alpha_{i,j+1})^2}{\sigma}\right). \tag{2.17}$$

Such a weighted regularization forces pixels that have depth within a small range of one another to be smooth and does not penalize the points that have larger gap in depth (which indicates the presence of an edge). This helps preserve sharp edges in the reconstructed depth estimates. This weighting approach is analogous to bilateral filtering approach for image denoising [33, 106].

The regularized estimation problem for image and depth can be written in the following form:

$$\underset{\boldsymbol{\alpha},\mathbf{l}}{\text{minimize}}\ \frac{1}{2}\|\mathbf{y} - \boldsymbol{\Psi}(\boldsymbol{\alpha})\mathbf{l}\|_2^2 + \lambda R_W(\boldsymbol{\alpha}). \tag{2.18}$$

We call this regularization approach weighted TV-ℓ_2 and solve it by alternately updating the inverse depth map $\boldsymbol{\alpha}$ and light intensity $\mathbf{l}$. A pseudocode of the algorithm is presented at Algorithm 1.

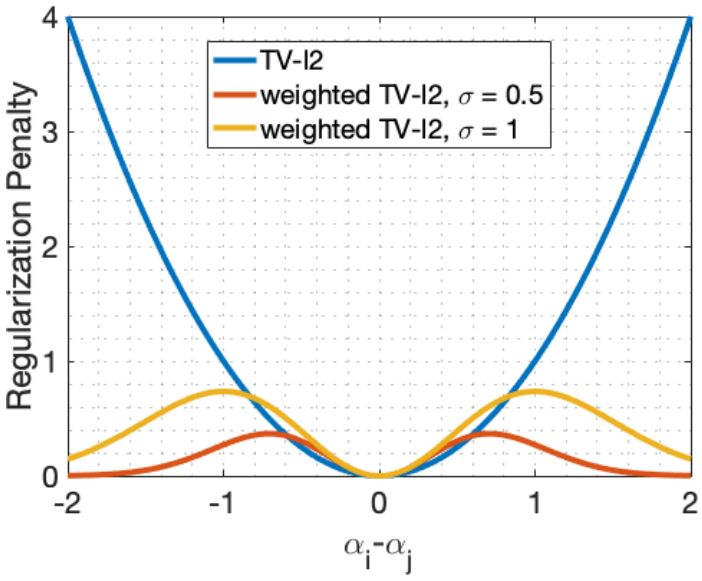


Figure 2.4: The weighted regularization function penalizes depth values that are within a small distance of one another and does not penalize those values that are above certain threshold. The smooth range can be changed by tuning the parameter σ. In contrast to the TV-ℓ_2, a weighted TV-ℓ_2 regularization term does not penalize neighboring pixels with large depth disparity, which tends to preserve the sharpness of the edges in the depth estimation.

ℓ_1 Regularization on Spatial Gradients. It is well-known that the ℓ_1 norm regularization enforces the solution to be sparse. We add an ℓ_1-based total variation norm [87] of the depth to our optimization problem. By enforcing the sparsity of spatial gradients, the edges of (inverse) depth map can be preserved. The ℓ_1 norm-based TV regularization term is given as

$$\begin{aligned} R_{TV}(\boldsymbol{\alpha}) &= \sum_{i,j=1}^{N} |\alpha_{i,j} - \alpha_{i+1,j}| + |\alpha_{i,j} - \alpha_{i,j+1}| \\ &= \|\nabla_r \boldsymbol{\alpha}\|_1 + \|\nabla_c \boldsymbol{\alpha}\|_1. \end{aligned} \tag{2.19}$$

Algorithm 1 Weighted TV-ℓ_2 regularized optimization

Input: Sensor measurements: $\mathbf{y}$
Output: Light distribution and inverse depth map: $\mathbf{l}, \boldsymbol{\alpha}$
Initialization via greedy algorithm:
Compute $\boldsymbol{\alpha}$ and $\mathbf{l}$ with depth pursuit algorithm in [5].
Refinement via alternating gradient descent:
for $k = 1 : k_{\max}$ **do**
$\widehat{\boldsymbol{\alpha}}^k = \underset{\boldsymbol{\alpha}}{\operatorname{argmin}} \frac{1}{2}\|\mathbf{y} - \boldsymbol{\Psi}(\boldsymbol{\alpha})\mathbf{l}^{k-1}\|_2^2 + \lambda R_W(\boldsymbol{\alpha})$
$\widehat{\mathbf{l}}^k = \underset{\mathbf{l}}{\operatorname{argmin}} \frac{1}{2}\|\mathbf{y} - \boldsymbol{\Psi}(\boldsymbol{\alpha}^k)\mathbf{l}\|_2^2$
end for
return $\widehat{\mathbf{l}}$ and $\widehat{\boldsymbol{\alpha}}$

To solve the nonlinear optimization problem with ℓ_1 norm regularization, we write the optimization problem as

$$\begin{aligned} \underset{\boldsymbol{\alpha},\mathbf{l}}{\text{minimize}}\ & \frac{1}{2}\|\mathbf{y} - \boldsymbol{\Psi}(\boldsymbol{\alpha})\mathbf{l}\|_2^2 + \lambda(\|\mathbf{d}_r\|_1 + \|\mathbf{d}_c\|_1) \\ \text{s.t. } & \mathbf{d}_r = \nabla_r \boldsymbol{\alpha},\ \mathbf{d}_c = \nabla_c \boldsymbol{\alpha}. \end{aligned} \tag{2.20}$$

We solve this problem (2.20) using a split-Bregman method [37]. A pseudocode is presented in Algorithm 2, where μ is the factor for the penalty function of the equation constraints in the unconstrained form of the optimization problem (2.20). soft represents soft-thresholding operator [109]. Every step in Algorithm 2 can be solved either in a closed form or using a gradient descent-based solver as in Algorithm 1.

Algorithm 2 TV-ℓ_1 regularized optimization

Input: Sensor measurements $\mathbf{y}$, intial light intensity $\mathbf{l}$, and inverse depth map $\boldsymbol{\alpha}$
Output: Light distribution and inverse depth map: $\mathbf{l}, \boldsymbol{\alpha}$
Initialization: Set $\mathbf{d}_r, \mathbf{d}_c, \mathbf{b}_x, \mathbf{b}_y$ to zero
Refinement via Split-Bregman steps:
for $k = 1 : k_{\text{max}}$ **do**

$$\widehat{\boldsymbol{\alpha}}^k = \underset{\boldsymbol{\alpha}}{\text{argmin}} \, \|\mathbf{y} - \boldsymbol{\Psi}(\boldsymbol{\alpha})\mathbf{l}^{k-1}\|_2^2 + \ldots$$
$$+ \mu(\|\mathbf{d}_r^{k-1} - \nabla_r \boldsymbol{\alpha}^k - \mathbf{b}_r^{k-1}\|_2^2 + \|\mathbf{d}_c^{k-1} - \nabla_c \boldsymbol{\alpha}^k - \mathbf{b}_c^{k-1}\|_2^2)$$

$$\widehat{\mathbf{l}}^k = \underset{\mathbf{l}}{\text{argmin}} \, \|\mathbf{y} - \boldsymbol{\Psi}(\boldsymbol{\alpha}^k)\mathbf{l}\|_2^2$$
$$\mathbf{d}_r^k = \text{soft}(\nabla_r \boldsymbol{\alpha}^k + \mathbf{b}_r^{k-1}, \tfrac{2\lambda}{\mu})$$
$$\mathbf{d}_c^k = \text{soft}(\nabla_c \boldsymbol{\alpha}^k + \mathbf{b}_c^{k-1}, \tfrac{2\lambda}{\mu})$$
$$\mathbf{b}_r^k = \mathbf{b}_x^{k-1} + (\nabla_r \boldsymbol{\alpha}^k - \mathbf{d}_r^k)$$
$$\mathbf{b}_c^k = \mathbf{b}_y^{k-1} + (\nabla_c \boldsymbol{\alpha}^k - \mathbf{d}_c^k)$$

end for
return $\widehat{\mathbf{l}}$ and $\widehat{\boldsymbol{\alpha}}$

2.5 Simulation Results

In this section, we present simulation results to evaluate the performance of our methods under different noise levels and sensor sizes. We also present a comparison of our proposed method with two existing methods for 3D imaging with lensless cameras.

2.5.1 Simulation Setup

To validate the performance of the proposed algorithm, we simulate a lensless imaging system using a binary planar mask with a separable maximum length sequence (MLS) pattern [63] that is placed 4mm away from a planar sensor array. We used an MLS sequence of length 1024 and converted all the -1s to 0s to create a separable binary pattern. We used square mask features, each of which is 30μm wide. Since we optimize the objective function in (2.6) with respect to α and need to compute the gradient in (2.9), we require the mask function to be smooth and differentiable with respect to α. Therefore, we convolved the binary pattern with a Gaussian blur kernel of length

15μm and standard deviation 5. In our simulations, we do not explicitly model the diffraction blur. However, the Gaussian blur kernel that we apply to the mask function can be viewed as an approximation of the diffraction blur. The sensor contains 512×512 square pixels, each of which is 50μm wide. The chief ray angle of each sensor pixel is $\pm 18^\circ$. We assume that there is no noise added to the sensor measurements. In our experiments for continuous depth estimation, we fixed all the parameters to these default values and analyze the performance with respect to a single parameter.

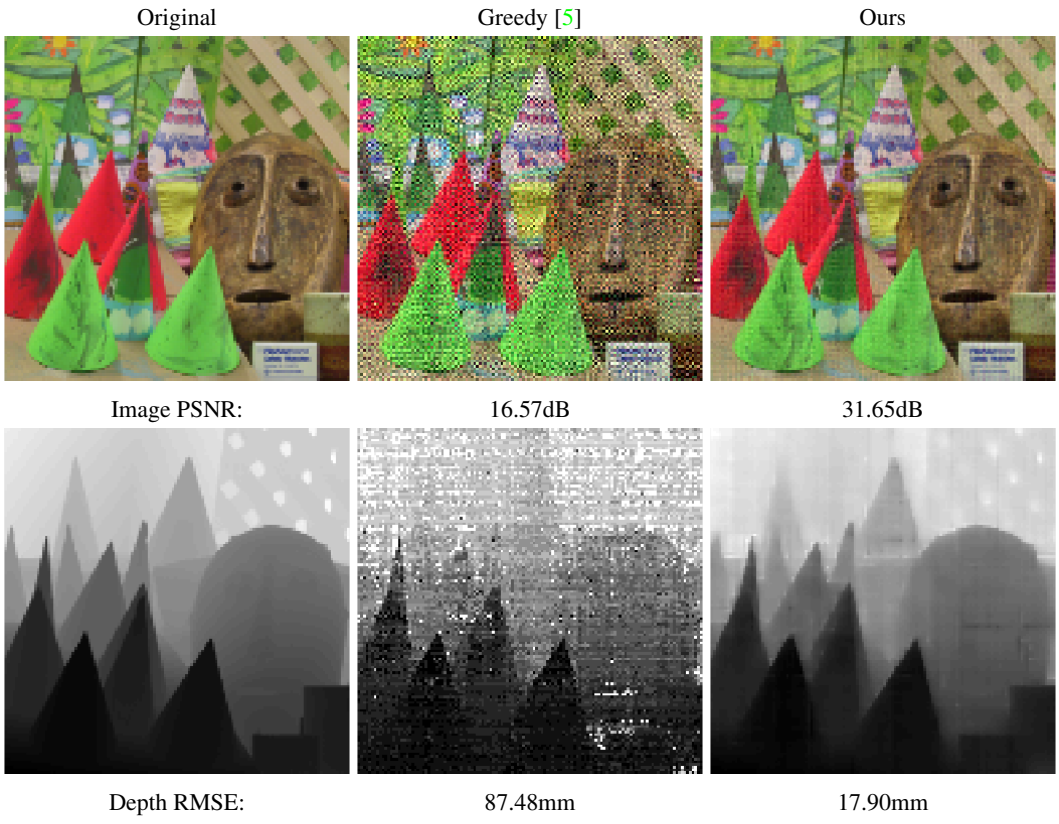


Figure 2.5: Left to right: original image and depth of the Cones scene; image and depth initialized via greedy algorithm [5]; depth estimation using weighted ℓ_2-based regularization. The depth in this scene varies from around 0.99m to 1.7m.

2.5.2 Reconstruction of Scenes With Continuous Depth

Depth Datasets We performed all our experiments on 3D images created using light intensities and depth information from Middlebury [91], Make3D [89, 90] and NYU Depth [74], the test scenes and their depth ranges are listed in Table 2.1.

Test datasets	Min depth (m)	Max depth (m)
Sword	0.65	0.95
Playtable	1.47	3.75
Cones	0.99	1.70
Corner	3.93	10.60
Whiteboard	1.08	2.90
Playroom	1.62	2.93
Moebius	0.74	1.23
Books	0.73	1.27

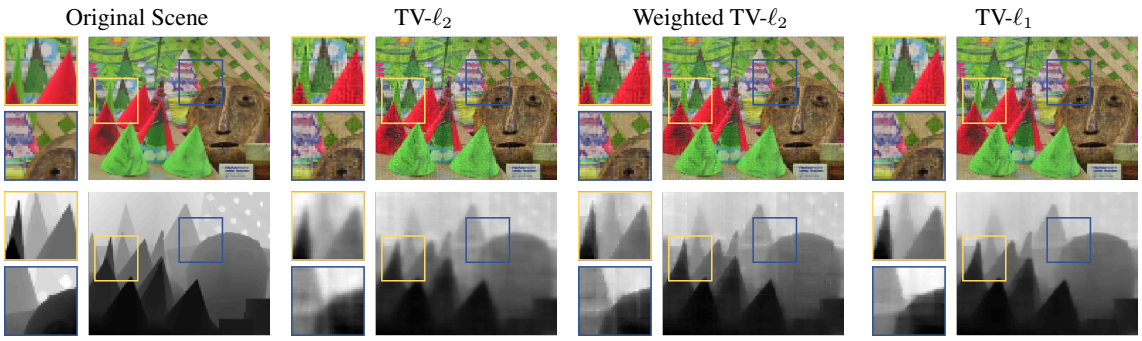
Table 2.1: Analysis experiments are performed on multiple scenes picked from Middlebury [91], Make3D [89, 90] and NYU Depth [74]. Results of the two scenes above line are presented within the main text, while the rest of them are reported in the supplementary material.

Initialization via Greedy Method Let us further discuss our simulation setup using the *Cones* scene, for which the results are presented in Figure 2.5. We simulated the 3D scene using depth data from Middlebury dataset [91]. We sample the scene at uniform angles to create a 128×128 image and its (inverse) depth map with the same size. We can compute the physical depth from $\boldsymbol{\alpha}$ using (2.2). In our simulation, the depth of this scene ranges from around 0.99m to 1.7m. We used depth pursuit greedy algorithm in [5] as our initialization method. We selected 15 candidate depths by uniformly sampling the inverse depth values $\boldsymbol{\alpha}$ from 0.996 to 0.9976, which gives an effective

depth in the same range as the original depth. Since we are trying to gauge the performance for off-the-grid estimate of depth, the candidate values of $\boldsymbol{\alpha}$ are not exactly the same as the true values of $\boldsymbol{\alpha}$ in our simulations. The output of the initialization algorithm is then fed into the alternating gradient descent method.

Performance Metrics We evaluate the performance of recovered image intensity and depth independent of each other. We report the peak signal to noise ratio (PSNR) of the estimated intensity distribution and root mean squared error (RMSE) of the estimated depth maps for all our experiments. The estimates for image intensity and depth maps for the initialization and our proposed weighted TV-ℓ_2 method are shown in Figure 2.5, along with the PSNR and RMSE. We can observe that both image and depth estimation from greedy method [5] contain several spikes because of the model mismatch with the predefined depth grid. In contrast, many of these spikes are removed in the estimations from the proposed algorithm with weighted TV-ℓ_2 while the edges are preserved.

Comparison of Regularization Methods Here we present a comparison between three different regularization approaches. We reconstruct image intensity and (inverse) depth map using the same measurements with TV-ℓ_2, weighted TV-ℓ_2, and TV-ℓ_1 regularization. The results are shown in Figure 2.6. Compared to the TV-ℓ_2 method, we observe that both weighted TV-ℓ_2 and TV-ℓ_1 preserve the sharp edges in image and depth estimates. Overall, in our experiments, weighted TV-ℓ_2 provided the best results. Therefore, we used that as our default method for the rest of the paper.



(a) Image and depth of the original scene. The selected Cones scene is taken from Middlebury dataset [91]. The range of depth is from 0.99 to 1.7 meters.

(b) Image and depth reconstruction from isotripic total variation. PSNR of image is 29.69dB and depth RMSE is 25.21mm.

(c) Image and depth reconstruction from weighted ℓ_2 total variation. The PSNR of image is 31.65dB and the RMSE of depth is 17.90mm. The edges of depth are preserved better.

(d) Image and depth reconstruction from TV-ℓ_1. The PSNR of image is 30.82dB and the depth RMSE is 19.56mm. The edges of depth are preserved better .

Figure 2.6: Comparison between reconstructions using three different regularization approaches from the same measurements.

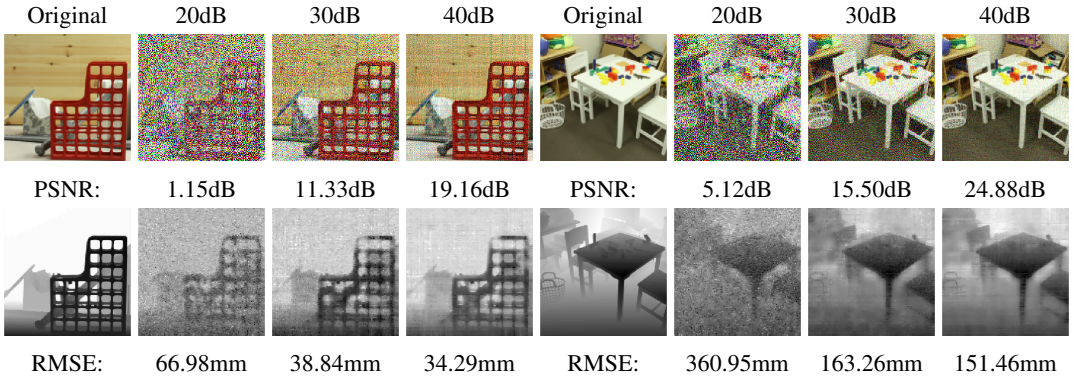
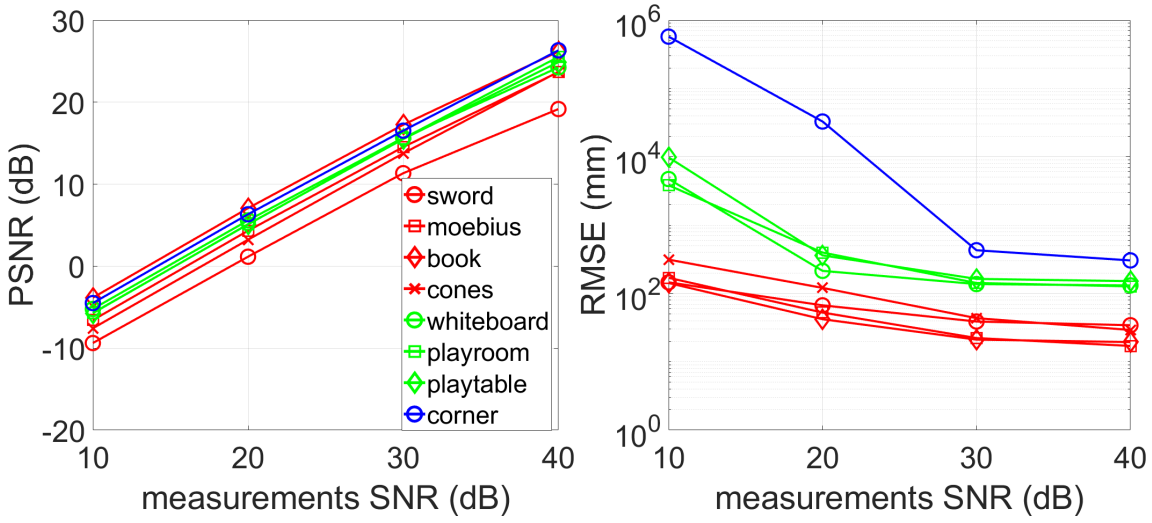


Figure 2.7: Effects of noise: Reconstruction from the measurements with signal-to-noise ratio (SNR) at 20dB, 30dB and 40dB, along with the **PSNR** of reconstructed image and **RMSE** of reconstructed depth map. As expected, the quality of reconstructed image and depth improves as the noise level is reduced. The sequence in left is for *Sword*, right is *Playtable*.



(a) Image PSNR for different noise levels

(b) Depth RMSE for different noise levels

Figure 2.8: Reconstruction from measurements with different levels of Gaussian noise on multiple scenes. Both of the image Peak Signal-Noise Ratio and depth Root mean squared error are improved as the noise is reduced. The reconstruction quality degrades if the scene is placed farther from the camera.

2.5.3 Effects of Noise

Sensor noise exists widely in any observation process. The amplitude of noise depends on the intensities of sensor measurements and can adversely affect the reconstruction results. To investigate the effect of noise on our algorithm, we present simulation results for the reconstruction of scenes from the same sensor measurements under different levels of additive white Gaussian noise. The experiments are performed on multiple 3D scenes listed in Table 2.1. Some examples of reconstruction with different levels of noise are shown in Figure 2.7.

The plots recording PSNR of image intensities and RMSE of depth maps over a range of measurement SNR values are presented in Figure 2.8. As we can observe from the curves that the quality of both estimated image and depth improve when the measurements have small noise (high SNR) and the quality degrades as we add more noise in the measurements (low SNR). Another observation we can make is that the scenes that are farther away have higher RMSE. This aspect is understandable because as the scenes move farther, α of the scene pixels all get very close to 1 and we cannot resolve fine depth variations in the scene.

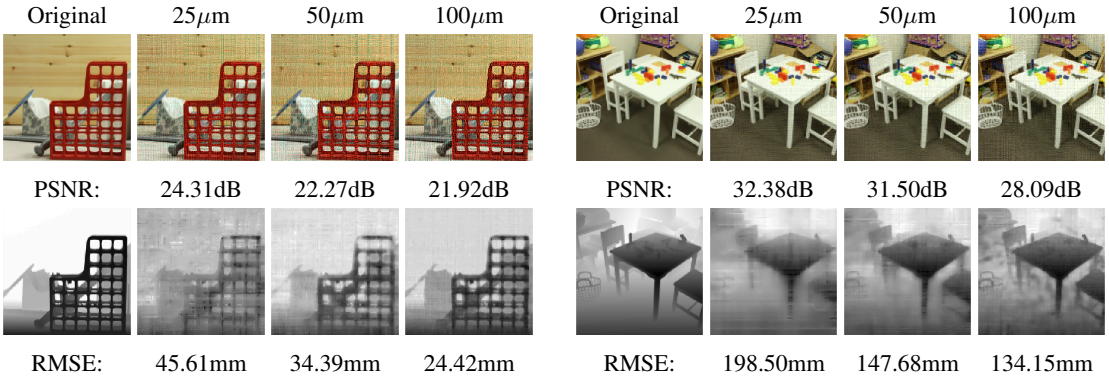


Figure 2.9: Reconstructions from measurements with different sizes of sensor pixels. The number of sensor pixels is fixed as 512×512. We compare the results in metric image PSNR and depth RMSE. The quality of depth reconstruction improves as we increase the size of sensor pixels.

2.5.4 Size of Sensor

In conventional disparity-based depth estimation method [44], the quality of reconstructed depth depends on the disparity between frames captured from multiple camera views. Larger distance between camera viewing positions results in better depth estimation accuracy. In a lensless imaging system, we can think of each pinhole on the mask and the sensor area behind the mask

as a tiny pinhole camera. The analogy only goes this far, because we do not record images from these tiny pinhole cameras separately; instead, we record a multiplexed version of all the views. The disparity between different points on the sensors, however, does affect our ability to resolve the depth of the scene, which is determined by the size of sensor.

To analyze the effect of disparity in our system, we performed experiments with three different sizes of sensor pixels from 25μm, 50μm, and 100μm. For comparison, the number of sensor pixels and other parameters are set to the default settings as described earlier. No noise is included in this experiment. Results in terms of reconstructed image and depth maps are presented in Figure 2.9, where we observe that the quality of depth reconstruction improves as we increase the size of sensor pixels. The results in Figure 2.9 demonstrate that increasing the disparity of viewing points increases the depth reconstruction quality.

2.5.5 Comparison With Existing Methods

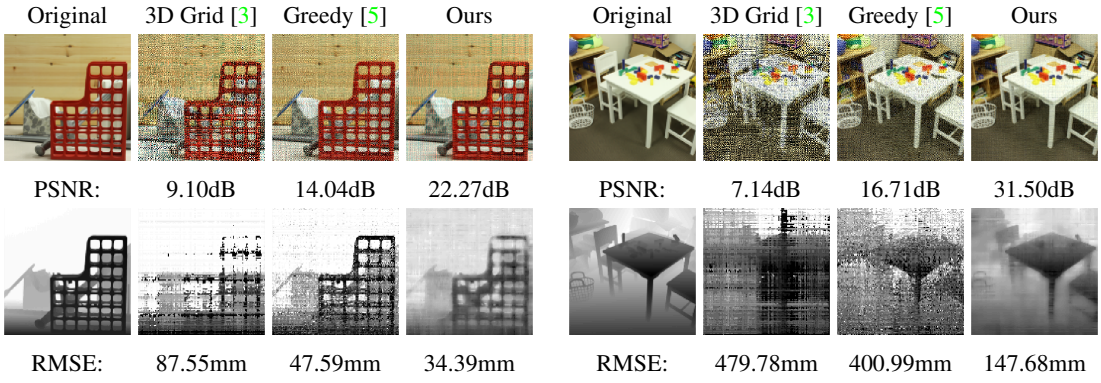


Figure 2.10: Comparison of existing 3D recovery methods for lensless imaging, 3D grid method from [1,3] and greedy method from [5], with our proposed method in metric image PSNR and depth RMSE. 3D grid method provides a 3D volume with multiple depth planes; therefore, we pick the depth with the largest light intensity along any angle for comparison.

Finally, we present a comparison of our proposed algorithm and two other methods for 3D recovery with lensless cameras. In our method, we estimate light intensity and a depth map over continuous domain. The greedy method in [5] also estimates intensity and depth separately, but the depth map for any angle is restricted to one of the predetermined planes. Three-dimensional recovery using lensless cameras for 3D fluorescence microscopy was presented in [3] and [1], which estimate the entire 3D volume of the scene sampled over a predetermined 3D grid. Since the unknown volume scene in microscopy is often very sparse, the 3D scene recovery problem is solved as a sparse recovery problem for the light intensity over all the grid voxels. The result is a light distribution over the entire 3D space. We call this method 3D Grid and use the code provided in [3] to solve the 3D recovery problem using the forward model and measurements from our simulation setup.

The imaging experiments in [3] and [1] are aimed at fluorescence imaging in which objects are mostly transparent and all the points in the 3D volume can contribute to the sensor measurements without occluding one another. In contrast, we consider natural photographic scenes, where objects are usually opaque and block light from objects behind them along the same angular direction. We can model such scenes as having only one voxel along any angle to be nonzero; however, that will be a nonconvex constraint and to enforce that we will have to resort to some heuristic similar to the one in [5]. For the sake of comparison, we solve the ℓ_1 norm-based sparse recovery problem as described in [3], but then we pick the points with the maximum light intensity at each angle to form the reconstructed image and (inverse) depth map.

A comparison of different recovery methods with the same imaging setup is shown in Figure 2.10. For the same scene, we reconstruct the same measurements using the three methods. As we can observe that our proposed algorithm offers a significant improvement compared to existing methods in all the test scenes.

The time and storage complexity of our proposed method and the other two methods depend on different factors; such as whether the imaging model is separable or convolutional and the sampling density along the depth. Since the main computational complexity of all the methods arises from the applications of the forward and adjoint operators, we will just discuss the complexity of those operators for different methods. The imaging operator in the greedy algorithm uses a separable mask and assumes that the scene consists of D depth planes. The computational complexity of the operator is $O(DMN^2)$ when we have $M \times M$ sensor pixels to reconstruct $N \times N$ image at D predefined depth planes. The convolutional model can be implemented using a fast Fourier transform and its complexity for a 3D volume with D depth planes is $O(DN^2 \log(N)$. The time complexity of forward imaging operator in the proposed method is $O(M^2N^2)$ because we assign independent depth values to each of the angles.

2.6 Experimental Results

To demonstrate the performance of our proposed method in the real world, we built a FlatCam prototype to capture images of different objects with different depth profiles. Below we discuss the details of our experiments and present reconstructed intensity and depth maps for some real objects.

2.6.1 Prototype Setup

Image Sensor. We used a Sony IMX249 CMOS color sensor that came inside a point grey camera (model BFLY-U3-23S6C-C). The sensor has 1920×1200 pixels and the size of each pixel is 5.86μm. The physical size of the sensor is approximately 11.2mm$\times$7mm.

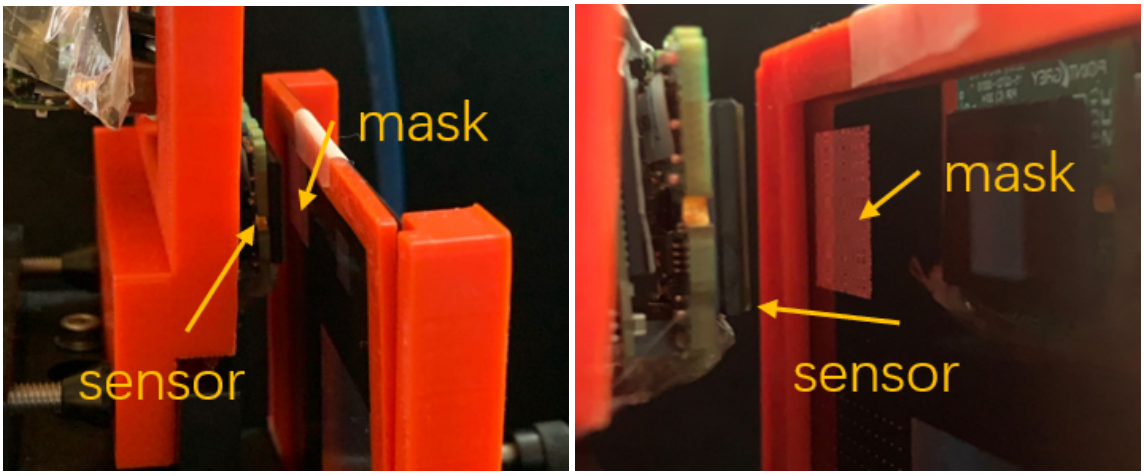


Figure 2.11: Camera prototype. The side view of the sensor and mask assembly. The sensor and mask are placed at a large distance for this image, but their distance (d) is approximately 4mm in our experiments. The mask pattern is binary and separable, and the physical size of each feature is 60μm.

Mask Pattern. We printed a binary mask pattern on a plastic sheet. The mask pattern was created by computing an outer product of two 255-length MLS vectors and setting all the -1 entries to 0. The physical size of each mask feature is 60μm. The physical size of the generated mask pattern is approximately 15.3mm$\times$15.3mm.

Sensor and Mask Placement. We placed the mask and the bare sensor on two optical posts such that the mask-to-sensor distance (d) is approximately 4mm; we attached kinematic platforms on top of the optical posts so that we can align the sensor and mask. Pictures of our sensor and mask setup are shown in Figure 2.11.

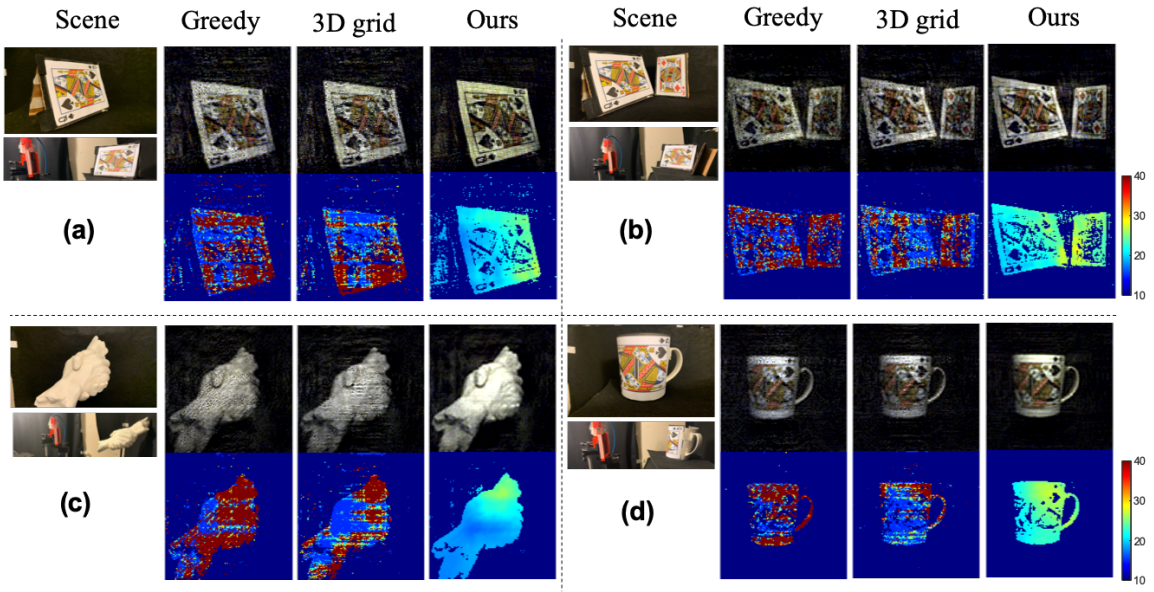


Figure 2.12: Experiments on real objects. **(a)** A slanted card; the depth range is 18–28cm **(b)** Two slanted cards; the depth range of left card is 18–28cm and the right card is 26–29cm. **(c)** Hand sculpture; depth range is 15–30cm. **(d)** A mug with card texture; depth range is 24–27cm. We divide each group of real scenes into four columns, the first column is front view and side view of the scene, the second column is the result from greedy algorithm in [5], the third column is the output of sparse 3D grid recovery algorithm proposed in [3] and [1], and the last column is the image intensity and depth map estimated using our proposed algorithm.

Data Acquisition and Processing. In our experiments, we calibrated the system by capturing sensor measurements while moving an LED flashlight at different locations in front of the camera. We performed all our experiments by uniformly illuminating the object with a table lamp. We reconstructed depth map and colored images at 128×128 pixel resolution from 512×512 sensor measurements. The sensor provides 1920×1200 pixels; we first resize the sensor measurement into 960×600 pixels by binning 2×2 pairs, and then we crop a 512×512 area in the center.

2.6.2 Calibration Procedure of the Prototype Camera

We use a separable mask pattern and align the mask and sensor assembly such that the response of any point source on the sensor is a rank-one image after mean subtraction [6]. Let us denote the Hadamard patterns as an $n \times n$ orthogonal matrix $\mathbf{H} = [\mathbf{h}_1, \ldots, \mathbf{h}_n]$, where each $\mathbf{h}_i$ represents a Hadamard pattern of length n. To calibrate the system for an $n \times n$ pixel grid, we project n horizontal and n vertical (rank-one) Hadamard patterns on a flat surface in the scene and record their response on the sensor. Every horizontal pattern can be represented as an $n \times n$ rank-one matrix $X_i = \mathbf{h}_i \mathbf{1}^\top$, where $\mathbf{1}$ denotes a vector of all ones. The corresponding sensor response can be represented as a rank-one matrix $Y_i = \Phi_L(\mathbf{h}_i \mathbf{1}^\top)\Phi_R^\top \equiv \mathbf{u}_k \mathbf{v}^\top$, where $\mathbf{u}_i = \Phi_L \mathbf{h}_i$ and $\mathbf{v} = \Phi_R \mathbf{1}$. We can concatenate all the $\mathbf{u}_i$ as columns in a matrix as $\mathbf{U} = [\mathbf{u}_1 \ \cdots \ \mathbf{u}_n] = \Phi_L \mathbf{H}$ and estimate $\Phi_L = \mathbf{U}\mathbf{H}^\top$. We can repeat the same procedure with vertical Hadamard patterns $\mathbf{1}\mathbf{h}_i^\top$ and estimate Φ_R. Instead of calibrating the separable system matrices for different depth planes, we calibrated the mask pattern function at one depth and evaluated the point spread function at arbitrary depth and angle according to (2.1). Because our mask pattern is bigger than the sensor, we captured sensor measurements for LED flashlight at 9 different angles at the same depth and merged them to estimate the mask function at that depth.

In our experiments, we captured the sensor measurements by placing an LED at $z = 42cm$ away from the sensor, which corresponds to the mask function in (2.1) evaluated at $\alpha = 1 - d/z = 0.9905$ for $d = 4mm, z = 42cm$. We first resized the calibrated mask function to compute the mask function corresponding to $\alpha = 1$.

2.6.3 Reconstruction of Real Objects

We present results for four objects in Figure 2.12, (a) slanted card has depth range from 18cm to 28cm, (b) two slanted cards have depth ranges from 18cm to 28cm and 26cm to 29cm, (c) hands sculpture has depth range from 15cm to 30cm, and (d) mug with card texture depth is from 24cm to 27cm. The figure is divided into four boxes. In each box, we present a front- and side-view of the object along with estimated scene intensity and depth maps for three methods. the greedy algorithm in [5], the sparse 3D volume recovery method from [1, 3], and our proposed method. For the greedy and 3D grid method, we generated 15 candidate depth planes by uniformly sampling the inverse depth values α between 0.96 and 0.9905 (corresponding to the depth of 10cm and 42cm, respectively).

All the objects in our experiments are placed in front of the black background and the depth values for dark pixels are not meaningful. We can observe that in all these experiments, our proposed method provides a continuous depth map that is consistent with the real depth of the object in the scene. In comparison, both the greedy algorithm [5] and the sparse 3D volume recovery algorithm [1, 3] produce coarse and discretized depth maps. The intensity map recovered by our method is also visually better compared to other methods.

Even though our proposed algorithm produces better intensity and depth maps compared to the greedy and 3D grid methods, we observed that the estimated depth has some errors in the darker parts of the objects. For instance, the left part of the mug is darker than the right part because the object was illuminated from a lamp on the right side. The left part appears to have errors in the depth estimate as several pixels are assigned small depth values, but that part is in fact farther from the sensor. We also observe a similar effect in other experiments, where depth estimates for darker parts of the scene appear to have larger errors.

2.7 Conclusion

We presented a new algorithm to jointly estimate the image and depth of a scene using a single snapshot of a mask-based lensless camera. Existing methods for 3D lensless imaging either estimate scene over a predefined 3D grid (which is computationally expensive) or a small number of candidate depth planes (which provides a coarse depth map). We divide the scene into an intensity map at uniform angles and a depth map on a continuous domain, which allows us to estimate a variety of scenes with different depth ranges using the same formulation. We jointly estimate the image intensity and depth map by solving a nonconvex problem. We initialize our estimates using a greedy method and add weighted regularization to enforce smoothness in the depth estimate while preserving the sharp edges. We demonstrated with extensive simulations and experiments with real data that our proposed method can recover image and depth with high accuracy for a variety of scenes. We evaluated the performance of our methods under different noise levels, sensor sizes, and numbers of sensor pixels and found the method to be robust. We presented a comparison with existing methods for lensless 3D imaging and demonstrated both in simulation and real experiments

that our method provides significantly better results. We believe this work provides a step toward capturing complex scenes with lensless cameras, where depth estimation is a feature as well as a compulsion because if the depth information is unavailable or inaccurate, that will cause artifacts in the recovered images.

Chapter 3

Improved 3D Lensless Imaging With Learned Programmable Masks

Lensless cameras provide a framework to build thin imaging systems by replacing the lens in a conventional camera with an amplitude or phase mask near the sensor. Existing methods for lensless imaging can recover the depth and intensity of the scene, but they require solving computationally-expensive inverse problems. Furthermore, existing methods struggle to recover dense scenes with large depth variations. In this chapter, we propose a lensless imaging system that captures a small number of measurements using different patterns on a programmable mask. In this context, we make three contributions. First, we present a fast recovery algorithm to recover textures on a fixed number of depth planes in the scene. Second, we consider the mask design problem, for programmable lensless cameras, and provide a design template for optimizing the

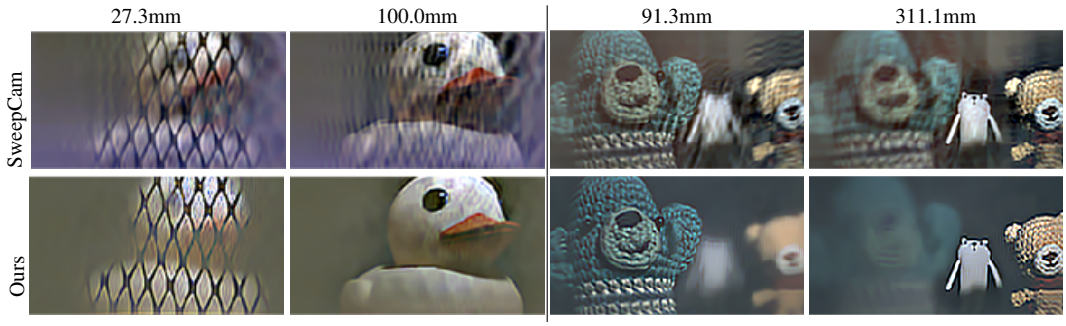


Figure 3.1: Examples of two 3D scenes reconstructed at different depth planes from eight sensor measurements using SweepCam [48] and our proposed method.

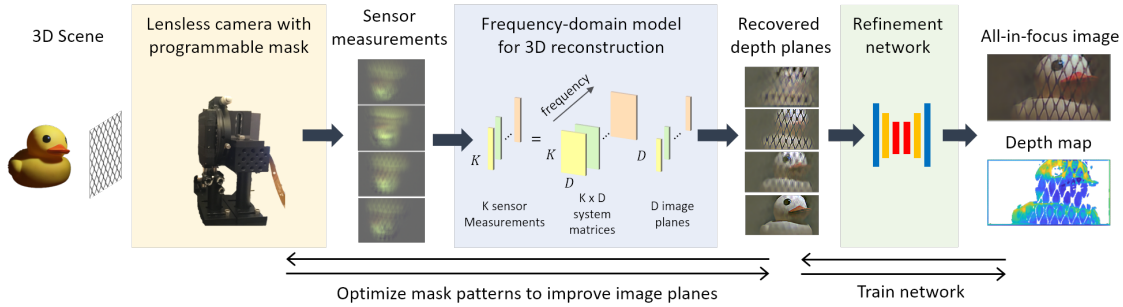


Figure 3.2: An overview of the proposed method. The lensless camera captures multiple measurements with a programmable mask. We reconstruct multiple image planes in the 3D scene by solving multiple small systems of equations in parallel (one for each frequency component). Using the recovery algorithm as a differentiable function, we learn the mask patterns to improve the estimates of image planes. We further refine the estimated image planes and convert them into an all-in-focus image and a depth map using a trained refinement network.

mask patterns with the goal of improving depth estimation. Third, we use a refinement network as a post-processing step to identify and remove artifacts in the reconstruction. These modifications are evaluated extensively with experimental results on a lensless camera prototype to showcase the performance benefits of the optimized masks and recovery algorithms over the state of the art.

3.1 Introduction

Contributions. The main contributions of this chapter are as follows.

- We present a fast algorithm to recover multi-plane images for the 3D scene from lensless measurements in the Fourier domain.
- We implement a multi-plane lensless camera and Fourier-domain recovery algorithm as a differentiable network to optimize mask patterns for the lensless camera.
- We train a neural network to map the estimated multi-plane images into an all-in-focus image and a continuous-valued depth map.
- We built a prototype lensless camera with a programmable mask and report several experiments to validate our proposed methods and a comparison with existing methods.

Limitations. Despite the advantages listed above, the mask-based lensless camera that we propose here has some limitations. One main limitation, common to many lensless cameras, is the light throughput and limited dynamic range of the sensor. A second set of limitations stem from the assumption of the convolutional model, a common assumption in 3D lensless imaging work. The convolutional model fails to account for small sensor area that crop the measurements, non-Lambertian scenes, and the sensor's angular response. In our experiments, we observe some artifacts in those scenarios, but they are usually local; for example, artifacts from sensor cropping only affect image boundaries and specular reflections only produce artifacts in local regions, and do not affect other parts of images. Finally, we assume the scene is static in the duration that multiple measurements are captured. However, this work significantly reduces the number of measurement required, and hence the duration in which the scene is required to be static, compared to previous methods [48].

3.2 Background and Related Work

A pinhole camera is a classical example of lensless imaging. FlatCam [6] is an extended version of a pinhole camera that uses a mask that has multiple pinholes. The sensor captures linear measurements of the scene and the system requires careful calibration and computational algorithm to reconstruct the target scene. The lensless imaging system underlying FlatCam has been used in face detection [101], privacy protection [78] and fluorescence microscopy [1]. Recently, deep learning-based recovery methods have also been introduced for FlatCam in [53].

In contrast to FlatCam, where an amplitude mask is used, DiffuserCam [3] places a phase mask on top of the sensor. The imaging system exploits the shift-invariant property of the PSF and models the measurements as the convolution between the scene and corresponding PSF. DiffuserCam has also been used to recover other types of high-dimensional signals, such as video [4] and hyperspectral images [72]. Methods for learning-based approaches have been introduced in [73].

Algorithms for depth estimation or 3D image reconstruction with a single mask-based FlatCam and DiffuserCam can be divided into two categories. One approach solves a sparse recovery problem based on a 3D grid [1, 3]. These methods are often used to recover sparse images in fluorescence microscopy where all the light point sources in the field of view can be detected by the sensor. The main drawback of these iterative 3D reconstruction algorithms is that they often require large time and space complexity for recovery. The other approach jointly estimates the image intensity and depth of every point/angle in the image [5, 122]. These methods assume angular sampling of 3D scenes in which only one light source exists at every angle and a greedy or nonlinear optimization algorithm to estimate the intensity and the depth map of the 3D scene.

SweepCam [48] is an extension to FlatCam that reconstructs image texture and depth map using multiple sensor measurements captured by shifting the mask pattern. The shifting mask pattern provides a depth-dependent disparity of different planes in the 3D scene for every measurement instance, which makes it possible to recover a focused image for a specific depth plane. SweepCam reconstructs one depth at a time and generates depth map using local contrast information of reconstructed depth planes. The main disadvantage of SweepCam is that it requires a very large number of sensor measurements, often in hundreds, to recover the depth planes without artifacts. In this chapter, we use a similar imaging setup as SweepCam [48] but use only eight to ten sensor measurements to recover the depth planes jointly.

Several methods for joint design of optics and reconstruction algorithms have been recently proposed in [13, 28, 68, 98, 99, 111, 113]. The main design principle is to represent sensor measurements as a differentiable function of the scene image and reconstruction of image as a differentiable neural network, after which parameters for the optical elements and the neural network can be learned jointly in an end-to-end manner. In our proposed framework, we recover 3D scene as a regularized least squares problem that has a closed-form solution. The main motivation for using a simple solver instead of training a deep network for reconstruction is to build a well-conditioned system that is generalizable to arbitrary scenes. We then train a separate refinement network that maps the estimated multi-plane images to an all-in-focus image and a continuous-valued depth map.

In this chapter, we propose a framework that (1) Instead of estimating one depth at a time, we recover multiple depth planes jointly using a computationally simple algorithm in the Fourier domain. (2) We implement our imaging model and recovery algorithm as a differentiable function and optimize the mask patterns to recover images planes accurately. (3) Finally, we train a network to remove any artifacts from the estimated image planes and recover an all-in-focus image and continuous-valued depth map as a post-processing step.

3.3 Technical Approches

3.3.1 Multi-Plane Lensless Imaging Model

We introduce the model underlying our lensless imager, where 3D scenes are represented using multi-plane images. Under some mild conditions on the scene and sensor sizes, we can represent sensor measurements as a summation of every scene plane convolved with its respective depth-dependent point spread function (PSF). The PSF depends on the mask pattern, which we can change to capture multiple sensor measurements and recover the 3D scene.

Let us assume that our lensless camera consists of a programmable amplitude mask placed at a distance d from a planar sensor array. Let us further assume that the sensor is placed at the origin of a Cartesian coordinate system indexed by (u, v, z), where (u, v) denotes the horizontal and vertical coordinates and z denotes the depth. In a mask-based lensless camera, every sensor pixel records a linear combination of light coming from all points in the 3D scene. Let us assume that light rays originating from scene point (u, v, z) have same effective brightness $l(u, v, z)$ with respect to every sensor pixel. Suppose we use an amplitude mask whose attenuation at (u, v) in the

mask plane is represented as $\varphi(u, v)$. Under these assumptions, the sensor measurement recorded at pixel (u, v) is given as

$$y(u,v) = \int_z \left[\int_{u_0,v_0} l(u_0, v_0, z) \varphi\left(\tilde{u}, \tilde{v}\right) du_0 dv_0 \right] dz, \tag{3.1}$$

where $\tilde{u} = \frac{z-d}{z}u + \frac{d}{z}u_0$, $\tilde{v} = \frac{z-d}{z}v + \frac{d}{z}v_0$. A simple derivation shows that the inner integration in (5.1) for a fixed z can be represented as a convolution of two functions [3, 48]:

$$\begin{aligned} y(u,v) &= \int_z \left[\int_{u_0,v_0} l_z(u_0, v_0) \varphi_z(u - u_0, v - v_0) du_0 dv_0 \right] dz \\ &= \int_z [l_z * \varphi_z](u, v) dz, \end{aligned} \tag{3.2}$$

where $l_z(u, v) = l(\frac{d-z}{d}u, \frac{d-z}{d}v)$ denotes the scene plane at depth z and $\varphi_z(u, v) = \varphi(\frac{z-d}{z}u, \frac{z-d}{z}v)$ denotes the PSF for any point at depth z.

Let us assume that the sensor array has M pixels on a uniform grid. By discretizing the system in (3.2) along (u, v, z), we can write the sensor measurements as a summation of 2D linear convolutions:

$$\mathbf{y} = \sum_z \varphi_z * \mathbf{l}_z + \mathbf{e}, \tag{3.3}$$

where $\mathbf{y}$ denotes an array with M sensor measurements $y(u, v)$, $\mathbf{l}_z$ denotes image-plane intensity values $l_z(u, v)$, φ_z denotes the PSF for depth z, and $\mathbf{e}$ denotes the sensor noise. For the ease of exposition, we ignore the additive noise term in the equations below, but we include noise in our experiments. If the sensor size is large enough so that the response of every light source in the scene

is completely recorded by the sensor, then we can treat the linear convolution in (3.3) as circular convolution. We can diagonalize the system in (3.3), using the convolution theorem, as

$$\mathcal{F}(\mathbf{y}) = \mathcal{F}(\varphi_z) \odot \mathcal{F}(\mathbf{l}_z) \Rightarrow \mathbf{Y} = \sum_z \Phi_z \odot_z, \tag{3.4}$$

where $\mathcal{F}(\cdot)$ denotes the 2D Fourier transform operator, $\mathbf{Y}$, Φ_z, and $_z$ denote 2D Fourier transforms of $\mathbf{y}$, φ_z and $\mathbf{l}_z$, respectively, and $\odot$ denotes an element-wise multiplication operator. All the arrays in (3.4) have M entries.

Since we use a programmable mask, we can capture multiple measurements of the scene using different mask patterns. Let us represent the sensor measurements captured using kth mask pattern as

$$\mathbf{y}^k = \sum_z \varphi_z^k * \mathbf{l}_z \quad \xrightarrow{\mathcal{F}} \quad \mathbf{Y}^k = \sum_z \Phi_z^k \odot_z . \tag{3.5}$$

3.3.2 Fast Algorithm for Multi-Plane Reconstruction

In this section, we discuss a Fourier-domain algorithm to recover 3D scene using multiple lensless measurements in (3.5). This method is an adaptation of the classical frequency-domain multi-channel deconvolution method [36] to the lensless imaging setup. Let us assume that we are given K sensor measurements as $\mathbf{y}^k$ for $k = 1, \ldots, K$. Let us further assume that the 3D scene consists of D planes represented as $\mathbf{l}_1, \ldots, \mathbf{l}_D$. To simplify the notation, we will use $\varphi_1, \ldots, \varphi_D$ to denote PSFs for different depths. To recover the image planes in the Fourier domain, we can solve the following regularized least-squares problem:

$$\min_{1,\ldots,D} \sum_k \|\mathbf{Y}_k - \sum_z \Phi_z \odot_z \|_F^2 + \tau \sum_z \|_z\|_F^2, \tag{3.6}$$

which involves solving for MD unknowns using MK measurements and can be computationally expensive to solve for large values of M, K, and D.

Fortunately, we can simplify the recovery problem by separating the optimization problem in (3.6) into M independent problems, each defined over a single frequency coefficient in all the depth planes. Note that because of the diagonal structure in the Fourier domain, we can separate the measurements for a frequency coefficient ω_m as

$$\begin{bmatrix} \mathbf{Y}^1(\omega_m) \\ \vdots \\ \mathbf{Y}^K(\omega_m) \end{bmatrix} = \begin{bmatrix} \Phi_1^1(\omega_m) & \dots & \Phi_D^1(\omega_m) \\ \vdots & \ddots & \vdots \\ \Phi_1^K(\omega_m) & \dots & \Phi_D^K(\omega_m) \end{bmatrix} \begin{bmatrix} {}_1(\omega_m) \\ \vdots \\ {}_D(\omega_m) \end{bmatrix}. \tag{3.7}$$

We can rewrite this system in a compact form as

$$\mathbf{Y}_{\omega_m} = \Phi_{\omega_m \omega_m}, \tag{3.8}$$

$\mathbf{Y}_{\omega_m}$ is a complex vector of length K, Φ_{ω_m} is a $K \times D$ complex matrix, and ${}_{\omega_m}$ is the unknown vector of length D. The original optimization problem can be written as a summation of M independent optimization problems as

$$\min_{\omega_1,\dots,\omega_M} \sum_{\omega_m} \|\mathbf{Y}_{\omega_m} - \Phi_{\omega_m \omega_m}\|_F^2 + \tau \|_{\omega_m}\|_F^2. \tag{3.9}$$

The solution for any ω_m can be written in a closed-form as

$$\tilde{}_{\omega_m} = (\Phi_{\omega_m}^* \Phi_{\omega_m} + \tau I)^{-1} \Phi_{\omega_m}^* \mathbf{Y}_{\omega_m}, \tag{3.10}$$

which we can compute by either directly inverting the $D \times D$ matrices or using an iterative method such as conjugate gradients [38]. Since the computations for all the ω_m are independent of one another, we can solve (3.10) in parallel, which provides a fast recovery algorithm. To recover the 3D image planes, we can apply inverse Fourier transform on reconstructed frequency-domain depth planes. In practice, we adjust τ for each ω_m according to the Frobenius norm of the corresponding system matrix. Our method recovers multiple image planes from multiple measurements using a closed-form solution. In fact, if $K = D = 1$, then our method is equivalent to the standard Wiener deconvolution.

Practical Considerations. To ensure stable recovery of *arbitrary* 3D scenes, we desire the matrices in (3.10) to be invertible for all frequencies ω_m. In practice, we can resolve this issue in two ways. First, the energy of the Fourier coefficients for natural images is mostly concentrated at a small number of frequencies; therefore, we can recover the image planes reliably as long as the matrices corresponding to the significant frequencies are well-conditioned. In Sec. 3.3.3, we discuss an approach to improve the conditioning of the system and the estimation of the image planes by optimizing the mask patterns. Second, even though the invertibility condition requires the number of depth planes (D) to be at most the number of mask patterns (K), we have the flexibility to choose which depth planes to recover. We can adjust the location of depth planes according to the scene or select the depth planes that provide the best recovery performance. In our experiments, we observed that sampling depth planes uniformly in $\alpha = 1 - \frac{d}{z}$ parameter provides best reconstruction, where $\alpha \in [0, 1]$ maps to $z \in [d, \infty]$.

Relation to Existing Methods. In our proposed method, we use multiple mask patterns to recover multiple depth planes in a 3D scene. Recovery of a 3D scene using a single mask pattern ($K = 1$) is possible, but it remains a challenging problem. Existing methods for 3D lensless imaging from a single sensor image either assume sparse prior on the 3D scene and solve an ℓ_1-regularized problem over the 3D volume [3, 12] or solve a nonlinear inverse problem to jointly recover the intensity and depth of scene [5, 122]. Both of which are computationally expensive. SweepCam [48] recover a single depth plane using a "focusing" operation. We can show that the focusing operation for a shifting mask pattern in SweepCam is equivalent to solving the system in (3.7) for one depth at a time in the frequency domain. Mathematically, it is equivalent to estimating a single frequency ω_m for plane at depth z, which can be written as the following scalar equation:

$$\widetilde{}_z(\omega_m) = \frac{\sum_k \Phi_z^k(\omega_m)^* \mathbf{Y}^k(\omega_m)}{(\sum_k \Phi_z^k(\omega_m)^* \Phi_z^k(\omega_m) + \tau)}. \tag{3.11}$$

In the experiment section, we present a detailed comparison between the performance of our proposed method, depth pursuit in [5], and SweepCam [48]; our results demonstrate that our method outperforms existing methods.

3.3.3 Learning Mask Patterns

To improve the quality of estimated image planes, we optimize the mask patterns by implementing the multi-plane image recovery algorithm as a differentiable network. We build a computation graph that implements the imaging model in (3.5) and the fast recovery algorithm, as illustrated in Figure 3.2. The optimization variable is a $K \times P_u \times P_v$ tensor that has K mask patterns each of size $P_u \times P_v$ (we used $P_u = P_v = 63$ in our experiments). We minimize the mean squared error (MSE) between the input and reconstructed image planes with respect to the mask patterns via backpropagation. We use 50 scenes from NYU [74] dataset as training data to optimize the masks.

To implement the imaging model in (3.5), we first perform linear interpolation to compute the PSF of every mask for D predefined depth planes that are uniformly sampled along $\alpha = 1 - \frac{d}{z}$. This provides us a $K \times D \times M_u \times M_v$ tensor, where $M_u \times M_v$ is the size of the sensor (i.e., $M = M_u M_v$). Then we generate sensor measurements for given training image planes using the convolution model in (3.5). We add independent instances of Gaussian noise in the sensor measurements during mask optimization. The reconstruction operator provides D estimated image planes from the simulated measurements by solving the problem in (3.9). As explained in Sec 3.3.2, we can solve independent least-squares problems for all the frequencies and then reconstruct the image planes with an inverse Fourier transform. We optimize the masks for 300 epochs using Adam optimizer [54] with the learning rate of 0.01.

We use a liquid crystal on Silicon (LCoS) spatial light modulator (SLM) as a programmable mask and restrict the mask patterns to be binary during training. The phase retardation in LCoS has a strong spectral dependence, which makes it hard to implement a desired continuous-valued pattern consistently across a span of wavelengths. This is not an issue for binary patterns since we can sat-

urate phase retardation across all wavelengths. We represent the mask as a zero-centered sigmoid function of a continuous-valued optimization variable, which keeps the mask values in the range [-1,1] during optimization. We increase the slope of the sigmoid function at every epoch, which pushes the mask values closer to -1 or 1, and we finally set them to -1/1 at the end of optimization.

3.3.4 Refinement Network and Post-Processing

To further enhance the image quality and depth accuracy of the estimated planes, we train a neural network using the U-Net architecture [86] that maps the estimated multi-plane images to an all-in-focus image and a continuous-valued depth map. The U-Net accepts all the color channel of the multi-plane image stack and generates an RGBD tensor [28, 113]. To train the U-Net parameters, we generated synthetic multi-plane images using NYU depth dataset [74]. We first scale the depth of scenes in the NYU dataset linearly in the 35mm to 200mm range and then quantize the scenes to create 100 depth planes. Then we simulate the sensor measurements using the imaging model in (3.5) and reconstruct images at $D = 8$ depth planes by solving the problem in (3.9). The reconstructed image planes are fed into the U-Net as a single tensor with D RGB planes, and U-Net provides an output RGBD tensor. The U-Net loss function is defined as the mean squared error between the ground-truth and the output RGBD tensors. We used 380 scenes to train the network for 200 epochs. In our experiments, we compare the U-Net results against a model-based approach in which we first denoise the estimated image at every depth using BM3D [29], and then assign each pixel the depth value that provides maximum local contrast in the chosen depth plane.

3.4 Simulation Results

Simulation Setup. To validate the proposed algorithm and learn the mask patterns, we simulated an imaging system for the camera prototype discussed in Sec. 3.5. We used a 256×256 sensor array with a mask placed 10.51mm away. The number of features in each mask pattern is 63×63 and each feature size is 36. We evaluate the algorithm on five scenes selected from Middlebury dataset (cones, books, piano, playroom, and playtable) [91]. The spatial resolution of each image is 128×128. We rescale the depth values in every scene to 35–380mm range and quantize the resulting depth into eight planes. We add Gaussian noise at 40dB SNR to the sensor measurements. After we reconstruct the image planes of the 3D scene using the proposed algorithm, we generate an all-in-focus image and depth map using local contrast in which we pick the depth with the largest local contrast among all planes for every pixel. We evaluate the quality of recovered all-in-focus images using structural similarity index (SSIM). We evaluate the quality of the estimated depth map using depth accuracy that we define as the ratio of pixels that are assigned the correct depth to those assigned incorrect depth.

Number and Types of Mask Patterns. We evaluate the performance of our proposed method for different types and number of mask patterns (K). We test four different types of mask patterns: random binary masks, separable MLS masks [6], shifted MLS masks as used in SweepCam [48], and learned mask patterns that are optimized according to the method discussed in Sec. 3.3.3. We test $K = \{4, 6, 8, 10\}$ for each mask pattern with five scenes, each of which have $D = 8$ depth planes. Our test includes both the cases for under-determined $(K < D)$ and over-determined $(K > D)$ systems. We use sensor measurements with eight masks to reconstruct the multi-plane image

stacks as described in (3.10). Then we convert image planes to an all-in-focus image and a depth map using local contrast. We present examples of reconstructed images and average performance curves over five test scenes in terms of SSIM and depth accuracy in Figure 3.3. The reconstruction quality of image and depth improves as K increases. We observe that the learned masks provide significantly better reconstruction for intensity and depth estimates compared to MLS, random, and shifted MLS. Incorrect depth estimates cause model mismatch, which in turn cause artifacts in the reconstructed intensity images that can be seen in Figure 3.3(a). Additional examples of reconstructed all-in-focus images and depth maps are available in the supplementary material.

3.5 Experiments With Camera Prototype

3.5.1 Camera Prototype

To evaluate the performance of our proposed algorithms and mask patterns, we built a camera prototype (as shown in Figure 3.2 and the supplementary material). The camera consists of an image sensor and an LCoS display that acts as the programmable amplitude mask. Our LCoS is HOLOEYE LC2012 transmissive spatial light modulator; it has a pixel pitch of 36 and fill factor 58%; and it is sandwiched between a pair of cross polarizers. The distance between LCoS and the sensor is 10.51mm. We use a Sony IMX183 sensor on a board level Blackfly S, which has 5472×3648 pixels with 2.4 pitch and sensor area of 1.31cm $\times$ 0.88cm. In our experiments, we used 228×342 sensor measurements by binning 16×16 adjacent pixels, which gives an effective sensor pixel of size 38.4. The reconstructed image planes have 148×274 pixels because we crop the remaining pixels.

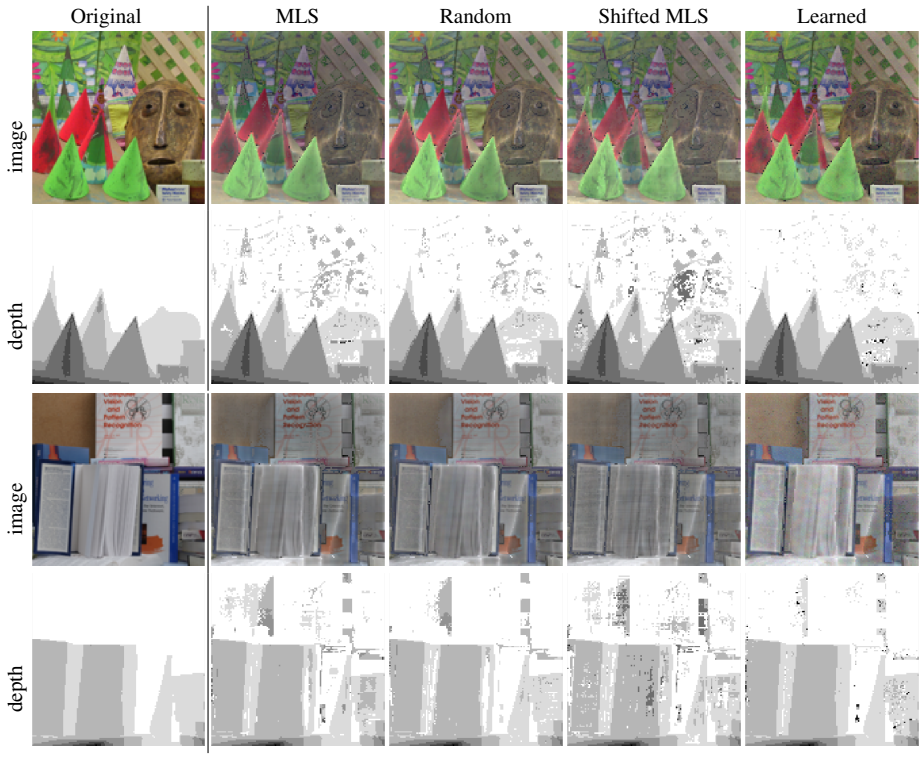
We capture measurements for multiple scenes with varying depths in the range from 30mm to 600mm. We calibrate the PSF of the camera at different depths. For each scene, we capture eight measurements using eight mask patterns with +1/-1 entries. We capture measurements for positive and negative parts of the masks separately and subtract the measurements for the negative part from the measurements for the positive part. Each mask has 63×63 pixels. We test the algorithms on four types of masks in our experiments: Shifted MLS masks: shifted versions of a single separable MLS mask [48]. The shifting distance is from 0 to 48 LCoS pixels. MLS masks are separable MLS masks generated from different random seeds. Random masks are non-separable random ± 1 masks. Learned masks are the he optimized masks generated from data-driven method discussed in Sec. 3.3.3. To present the images with the same dynamic range, we applied the same brightness and color contrast adjustment to all the images.

3.5.2 Reconstruction Using Learned Masks

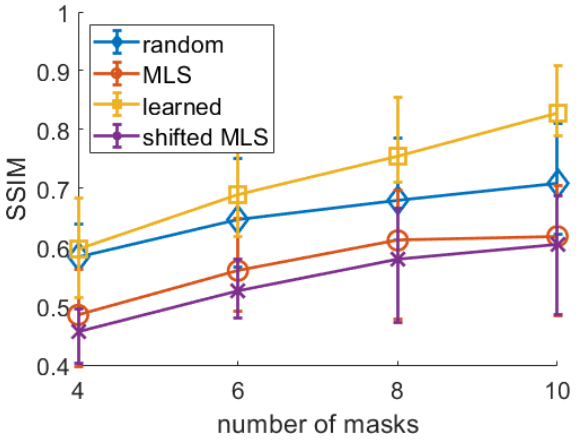
We first present some example scenes and the reconstructed image planes using our proposed method. For each scene, we capture sensor measurements using eight learned mask patterns. Then we recover eight depth planes using the proposed algorithm in Sec. 3.3.2. We sampled the depth planes uniformly in $[1 - \frac{d}{z_{\text{min}}}, 1 - \frac{d}{z_{\text{max}}}]$, where we input rough estimate of $z_{\text{min}}, z_{\text{max}}$ for each scene. Figure 3.4 presents three of the estimated planes for each scene. The first scene has a net mesh nearly 30mm from the camera and a rubber duck nearly 100mm behind the net. The near-depth plane image is focused on the net and duck is blurred, while the far-depth plane is focused on the rubber duck and the net has almost disappeared. The third scene consists of 2 cards, the red card (on left) is placed at 35mm and the yellow card (on right) is placed at 200mm. The near plane

has the left card in focus and the far plane has the right card in focus. The last scene has 3 toys, the blue toy is placed at 80mm, yellow toy at 200mm, and white toy at 600mm. We observe that the blue, yellow, and white toys appear sharpest in near, middle, and far planes, respectively. In lensless imaging systems, depth estimates of farther and darker objects are not as accurate as that of the closer and brighter objects; therefore, we removed the depth values of low-intensity pixels from the displayed depth maps.

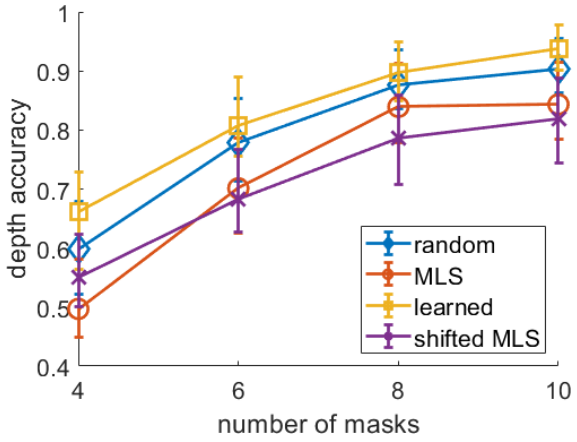
The scene shown in Figure 3.5 consists of a fork and a plate and demonstrates the limitation of our imaging system. Both of these objects have non-Lambertian surfaces and include specular reflections, which violates our convolution assumption. For this reason, the reconstructed images show artifacts.



(a) Reconstructed all-in-focus images and depth maps for cones with $K = 8$.



(b) SSIM for all-in-focus image



(c) Accuracy for estimated depth

Figure 3.3: Comparison of different types and number of masks. (a) Reconstructed all-in-focus images and depth maps for cones with $K = 8$ measurements. (b,c) Average SSIM of recovered all-in-focus images and accuracy of estimated depth for five test scenes. Quality of reconstruction improves as the number of masks increases, and learned mask patterns outperform other mask patterns.

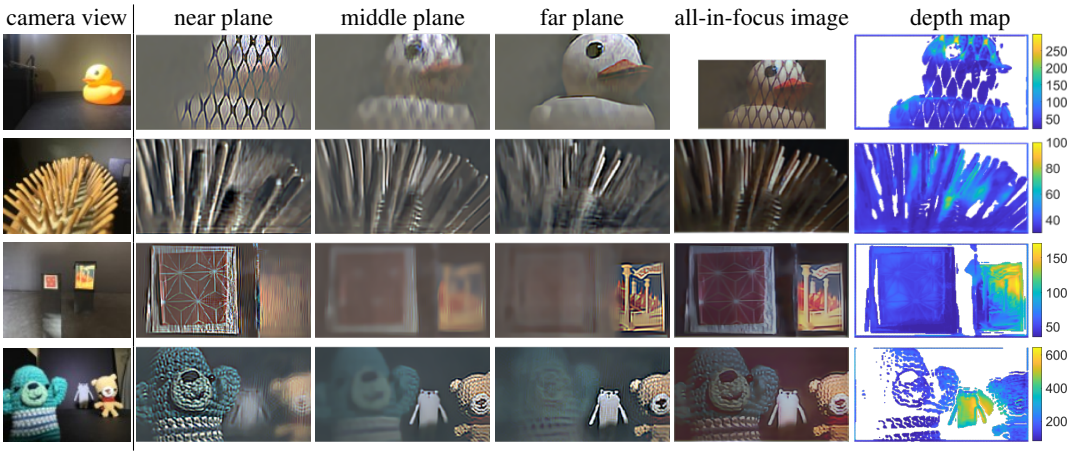


Figure 3.4: Reconstruction of depth planes for different scenes using our proposed fast recovery algorithm with learned masks. Objects outside the recovered plane almost disappear. We also show all-in-focus image and depth map (in mm) created after passing the estimated multi-plane images through the trained U-net based refinement network, as discussed in Sec. 3.3.4. Depth values of pixels with low intensity (e.g., mesh in the first row) are usually unrealiable and therefore removed. Results of additional depth planes and scenes are available in the supplementary material.

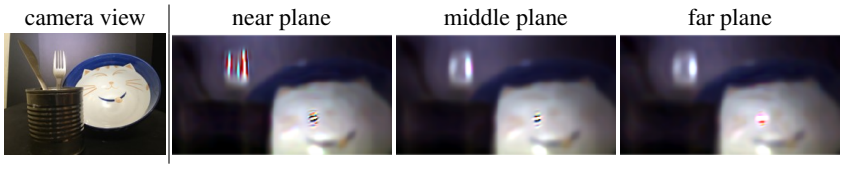


Figure 3.5: Reconstructions of a scene with specular reflections that our method fails to recover.

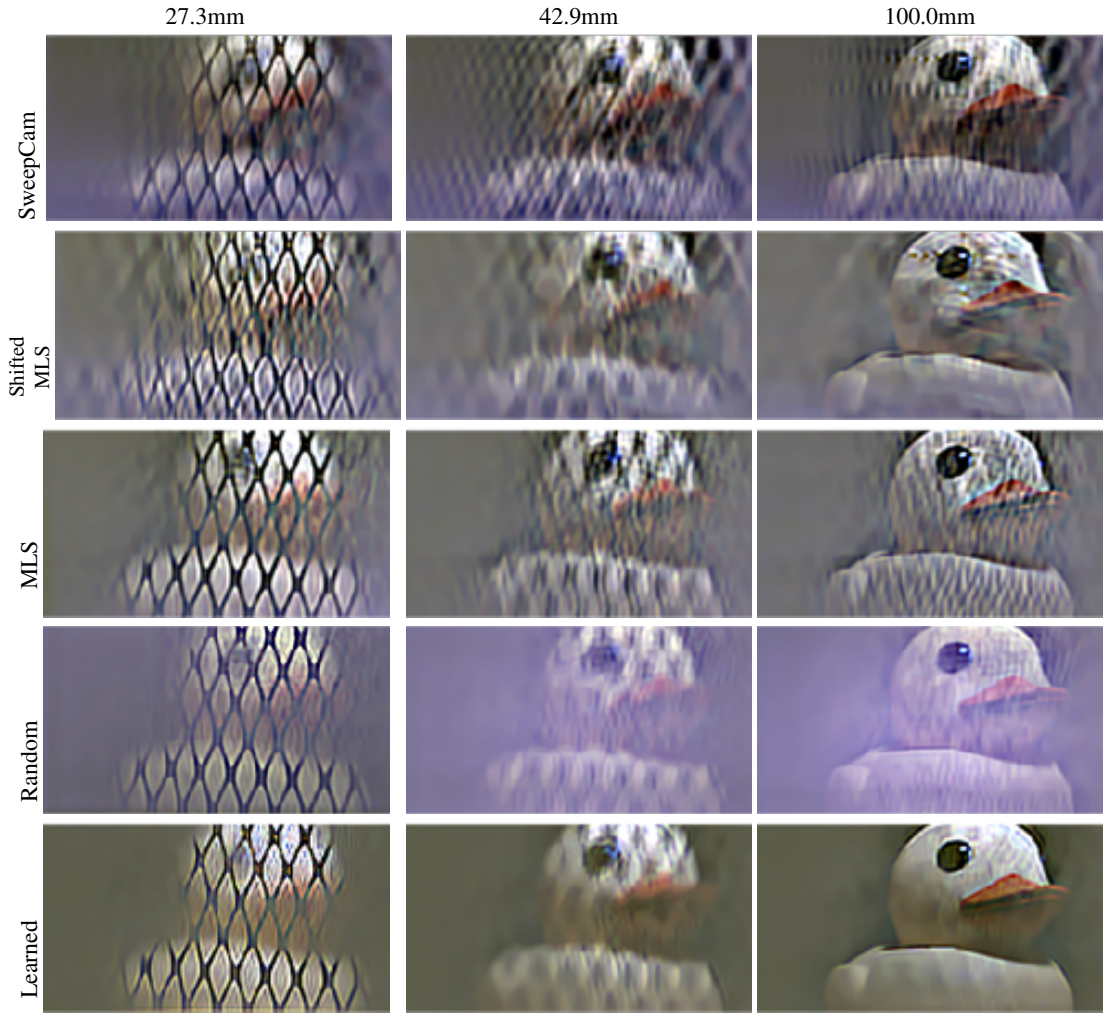


Figure 3.6: Reconstructions of depth planes for recovered depth planes using SweepCam [48] and our proposed methods using shifted MLS masks, MLS masks, random masks, and learned masks. We observe that the learned masks outperform all the other masks.

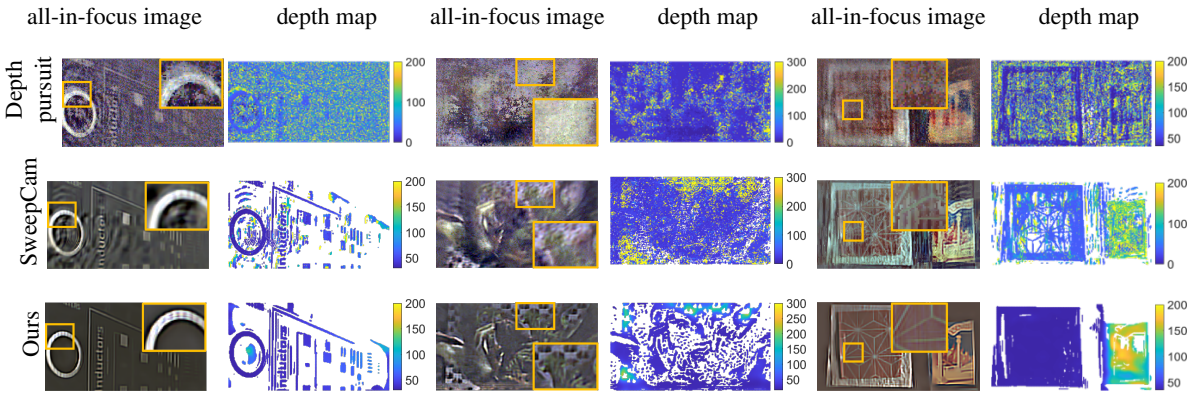


Figure 3.7: Comparison of the depth pursuit algorithm [5], SweepCam [48], and our proposed method. The details in the results from our proposed method are cleaner and sharper than the results from other methods.

3.5.3 Comparison of Mask Patterns

We show the comparison between different mask patterns and the learned masks. We present the reconstruction results in Figure 3.6. We compare SweepCam method with shifted MLS masks [48] and our method using shifted MLS masks, MLS masks, random masks, and learned masks. Figure 3.6 shows that SweepCam fails to recover depth planes accurately because we only used eight mask patterns. The reconstruction from learned masks using our method provides the best quality among all the mask types; far-depth plane separates the net and duck clearly. Compared to the learned masks, all the other mask patterns carry artifacts in their reconstructions (the shifted MLS and MLS masks give poor reconstruction for image at 100mm and random mask reconstruction exhibit haze artifacts). Additional experiments are available in the supplementary material.

3.5.4 Comparison With Existing Methods

We compare the performance of our proposed method with the greedy depth pursuit algorithm in [5] and the depth sweep method in [48]. In our experiments, the depth pursuit algorithm uses a single MLS mask, SweepCam method uses 8 shifted MLS masks, and our proposed method uses 8 learned masks. We present the estimated all-in-focus images and depth maps from the three methods in Figure 3.7, which shows that our method provides a better all-in-focus image and depth map compared to the other methods.

3.5.5 Refinement and Post-Processing

We show a comparison between two post-processing approaches for converting estimated image planes to all-in-focus image and continuous-valued depth map in Figure 3.8. In the first approach, we apply a BM3D denoiser [29] on every estimated plane; then we perform a local contrast analysis to select the depth for every pixel that has maximum local contrast. In the second approach, we use a trained U-Net [86] following the approach outlined in Sec. 3.3.4. Figure 3.8 shows that the U-Net results have less artifacts but all-in-focus image looks blurry. The results from local contrast-based method have more artifacts but images look sharp.

3.5.6 Computational Complexity and Time

To reconstruct an $M_u \times M_v \times D$ volume with K mask patterns, we solve $M = M_u M_v$ least-squares problems, each of size $K \times D$ as in (3.10). The computational complexity of every least-squares problem is $\mathcal{O}(D^3 + KD^2)$ [38]. Since all the least-square problems can be solved independently of each other, we can solve them in parallel to accelerate the algorithm. In practice,

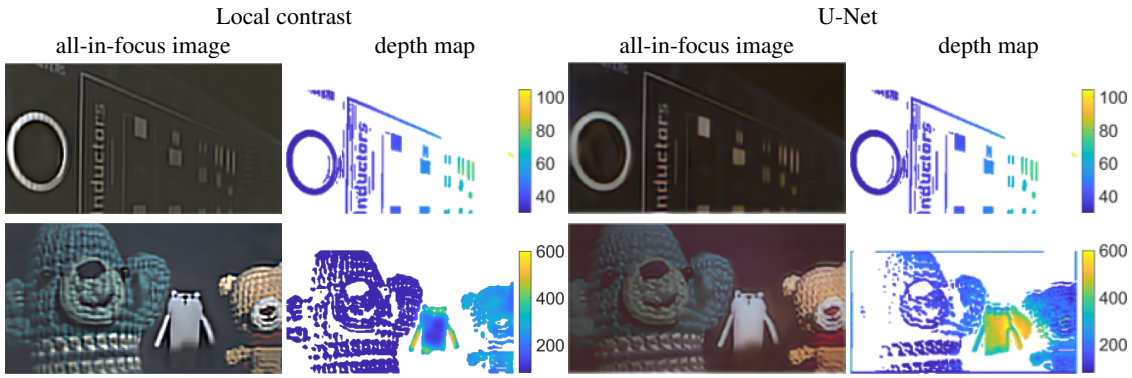


Figure 3.8: The all-in-focus images and depth maps generated by the local contrast-based method and the trained U-Net. Local contrast method uses 47.85 seconds and U-Net uses 0.0043 seconds on average.

the average running time to reconstruct eight depth planes with a single mask using the greedy depth pursuit algorithm [5] is 138.16 seconds. The average running time to reconstruct eight depth planes with eight programmable masks are 0.82 seconds for SweepCam [48] and 0.33 seconds for our method. In the post-processing step, which converts image planes to all-in-focus image and depth map, BM3D followed by local contrast method requires an average of 47.85 seconds, while the trained U-Net requires 0.0043 seconds.

3.6 Conclusion

In this chapter, we present a new framework to recover 3D scenes using a lensless camera with a programmable mask. Our proposed method can recover multiple depth planes in the 3D scene using a computationally efficient algorithm that solves multiple small linear systems in parallel in the frequency domain. To further improve the quality of 3D scene recovery, we optimized the mask patterns and trained a U-Net that converts estimated image planes to all-in-focus image and

continuous-valued depth map. Our experimental results demonstrate that the proposed method can reliably recover dense 3D scenes with a small number of sensor measurements and outperform existing methods. The reconstruction quality of our proposed method, like other lensless imaging systems, drops for scenes with specular reflections and large occlusions.

Chapter 4

Coded Illumination for Improved Lensless Imaging

Mask-based lensless cameras can be flat, thin, and light-weight, which makes them suitable for novel designs of computational imaging systems with large surface areas and arbitrary shapes. Despite recent progress in lensless cameras, the quality of images recovered from the lensless cameras is often poor due to the ill-conditioning of the underlying measurement system. In this chapter, we propose to use coded illumination to improve the quality of images reconstructed with lensless cameras. In our imaging model, the scene/object is illuminated by multiple coded illumination patterns as the lensless camera records sensor measurements. We designed and tested a number of illumination patterns and observed that shifting dots (and related orthogonal) patterns

provide the best overall performance. We propose a fast and low-complexity recovery algorithm that exploits the separability and block-diagonal structure in our system. We present simulation results and hardware experiment results to demonstrate that our proposed method can significantly improve the reconstruction quality.

4.1 Introduction

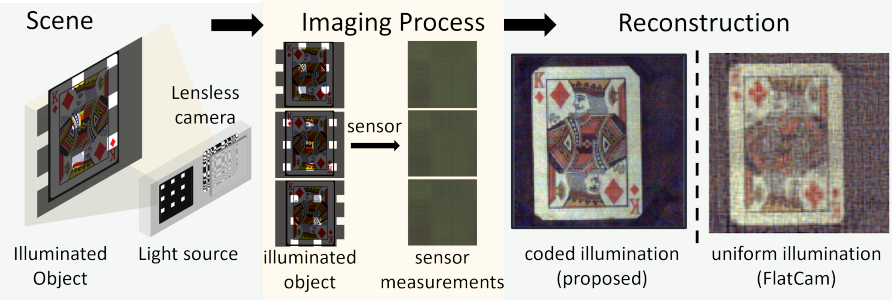


Figure 4.1: An overview of our proposed method. We project a sequence of binary illumination patterns onto the object and capture the sensor measurements corresponding to each illumination pattern. The reconstruction result using multiple coded illumination patterns significantly outperforms the conventional method where the scene is illuminated by uniform illumination.

In this chapter, we propose a new method that combines coded illumination with mask-based lensless cameras (such as FlatCam) to improve the quality of recovered images, illustrated in Fig. 5.1. A 2D scene is illuminated with multiple coded patterns during data acquisition. We capture sensor measurements for each illumination pattern and use a fast recovery algorithm to reconstruct the scene using all the measurements. The main contributions of this chapter are as follows

- We propose a new framework to combine coded illumination with lensless imaging.

- We propose a fast and low-complexity algorithm that exploits separability of the mask and illumination patterns and avoids storing all the measurements or creating large system matrices during reconstruction.
- We design shifting dots and orthogonal patterns that provide best overall performance in terms of quality and computational complexity.
- We present simulation results to show that our method can improve the image reconstruction quality under different system conditions.
- We present real experiments using a prototype we built to evaluate the performance of our imaging system and algorithm in the real environments.

Our main objective is to demonstrate that we can significantly improve the conditioning of a lensless imaging systems and the quality of the reconstruction by using coded illumination. Our experiments show that the image quality improves in almost all cases as we increase the number of illumination patterns. Our method also has an inherent trade-off between imaging speed and the quality of the recovered images. In this chapter, we mainly implement and discuss the design of FlatCam [6] with separable masks and illumination patterns. Nevertheless, the ideas presented here can be used to improve the imaging performance of other lensless systems, such as DiffuserCam [3], fresnel zone aperture [97], or phlatcam [13].

The proposed framework can be used to build large-area, high-resolution, and compact cameras that can potentially be used for under-the-display fingerprint and vein imaging. Such cameras can use the display panels for coded illumination [115,119,120]. Lensless microscopes [1,3,24] can also benefit by combining coded illumination with data capture. Another potential application for such lensless cameras is in different wearable devices. In particular, our inspiration came from a

project in which we seek to embed lensless sensors on a soft robot that will be used to assist infants in rehabilitation. The goal is to track infant arm motion using lensless sensors on a wearable sleeve while they reach objects. In all these cases, the objects in the scene can potentially be illuminated with a built-in display, a mini projector, or an add-on light source combined with a shifting mask,.

4.2 Background and Related Work

The mask-based lensless camera such as FlatCam [6] is an extended version of pinhole cameras. Although a pinhole camera is able to image the scene directly on its sensor, it often suffers from noise [118]. Coded aperture-based cameras alleviate this problem by using multiple pinholes placed in a designed pattern [6, 12, 19, 23, 34]. In contrast to conventional lens-based cameras that capture images of the scene directly, coded mask-based cameras capture linear measurements of the scene and perform reconstruction by solving a linear inverse problem. Coded aperture-based cameras can also recover the depth information of a scene [1, 3, 5, 48, 56, 122, 125]. The main advantage of FlatCam is the thin and flat form factor, which also makes the system ill-conditioned and affects the quality of reconstructed images.

Signal recovery from ill-conditioned and under-determined systems is a classical and long-standing problem in signal processing. A number of methods have been proposed to tackle these problems over the decades [21, 38, 87]. An ill-conditioned system is unstable as its solution can change dramatically with tiny perturbations; thus, such a system rarely generates good results in a signal recovery problem. An under-determined system has fewer measurements than the number of unknowns; thus, it admits infinitely many solutions, and one can hardly determine the true solution using only the measurements. A standard approach to deal with ill-conditioned and

under-determined systems is to add a signal-dependent regularization term in the recovery problem, which constrains the range of solutions. Popular methods include adding sparse and low-rank priors on the signals [8, 21, 31, 81, 87] and natural-image-like generators prior [15, 43, 49]. Another approach is to capture multiple, diverse measurements of the scene that makes the modified imaging system well-condition and the reconstruction more accurate [48, 125]. Recently, a number of methods have been proposed that use deep networks to reconstruct images from lensless measurements [13, 52, 71, 73]. Some of these methods provide an exceptional improvement over traditional optimization-based methods. Nevertheless, deep learning-based methods in general, and end-to-end methods in particular, provide a huge variation in performance for simulated and real data (mainly because of mismatch in the simulated/actual mask-sensor-projector configuration and scenes). In contrast to deep learning methods, our method seeks to improve the conditioning of the underlying linear system and offer better generalization and robust results for arbitrary scenes without the need for any learning from data.

Multiplexed illumination analysis is discussed in [92], which shows how overall reconstruction SNR can be improved using coded illumination with Hadamard codes. Plenoptic imaging and noise analysis for reconstruction from multiplexed measurements are discussed in [110]. Even though this chapter is focused on lensless imaging with coded illumination, the noise analysis we discuss in Sec. 4.6.1 follows similar arguments as [92, 110] to show the relationship between reconstruction error and overall system conditioning and singular values.

Our proposed approach is an active imaging approach combining coded modulation or structured illumination method with coded aperture imaging [35, 42, 75]. Structured illumination schemes are commonly used for imaging beyond diffraction in microscopy. These schemes use multiple structured illumination patterns to down-modulate high spatial frequencies in a sample into a low-frequency region that can be captured by the microscope [41, 42, 46]. Structured light is also widely used in multiplexing scene recovery to improve image quality and SNR [40, 70, 80, 92]. Other active imaging approach includes time-of-flight sensors [45, 88] that estimate the 3D scene by sending out infrared light and measuring its traveling time in reflection. Coded diffraction imaging is used to recover complex-valued wavefront from Fourier measurements [22, 96]. In coded diffraction imaging, the signal of interest gets modulated by a sequence of illumination patterns before the K-space measurements were captured [20, 50, 69]. Ptychography is another related method for capturing high-resolution microscopy images by capturing multiple images of the scene using a sequence of coded illumination patterns [85, 104]. Dual photography [94] and compressive sensing [32, 95] schemes also use coded illumination to sample the scenes. A dual photography system can create a high-resolution image of the scene by scanning the entire scene one pixel at a time, but it requires a fast laser projector. Single-pixel camera collects thousands of multiplexed measurements of the scene and solves a regularized optimization problem to reconstruct the image. In our method, we use lensless camera to capture tens of sensor frames with shifting coded illumination patterns and get better reconstruction with improved system conditioning.

A random illumination patterns-based lensless imaging method with simulations was presented in [126]. In contrast, we design shifting dots and orthogonal patterns that are significantly superior to random patterns in terms of quality of reconstruction and computational complexity. We provide detailed simulations and experimental results on real data captured with a custom-built prototype.

4.3 Technical Approaches

4.3.1 Separable Imaging Model

FlatCam [6] consists of an amplitude mask placed on top of a bare sensor, and every sensor pixel records a linear combination of the entire scene. Suppose the sensor plane is at the origin of the 3D Cartesian coordinates (u, v, z) and an amplitude mask is placed parallel to the sensor at distance d. We can model the measurement recorded at sensor pixel (u, v) as

$$y(u,v) = \int x(u',v',z)\,\varphi(u',v',u,v,z)\,du'dv'dz, \tag{4.1}$$

where $x(u',v',z)$ represents the intensity and $\varphi(u',v',u,v,z)$ represents the sensor response of a point source at (u',v',z). In this chapter, we assume that the scene consists of a single plane at a known depth; therefore, we can ignore the depth parameter and represent the sensor measurements as

$$\begin{aligned} y(u,v) &= \int x(u',v')\,\varphi(u',v',u,v)\,du'dv' \\ &\Rightarrow \quad \mathbf{y} = \mathbf{\Phi x}, \end{aligned} \tag{4.2}$$

where $\mathbf{x}$ denotes the scene intensity vector, $\mathbf{\Phi}$ denotes the system matrix, and $\mathbf{y}$ denotes the sensor measurement vector. The computational and memory complexity of the general imaging model in (4.2) makes it unsuitable for systems with a large number of scene and sensor pixels. We can overcome this challenge in a number of ways; for instance, we can use a separable model as in FlatCam [1, 6] or a convolutional model as in DiffuserCam [3]. We use a separable system in this chapter.

A separable mask pattern that is aligned with the sensor grid yields a separable imaging system, which can be represented as

$$y(u,v) = \int\int x(u',v')\,\varphi_L(u',u)\,du'\,\varphi_R(v',v)\,dv'. \tag{4.3}$$

The product of $\varphi_L(u',v)$ and $\varphi_R(v',v)$ represents the separable system response for point sources along u, v axes. Let us assume that X represents an $n \times n$ image of the scene intensities at a fixed plane and Y denotes $m \times m$ sensor measurements, then we can represent the separable system in (4.3) as

$$Y = \Phi_L X \Phi_R^\top, \tag{4.4}$$

where Φ_L, Φ_R denote the system matrices for u, v axes, respectively. We assume square shapes for the scene and sensor to keep our discussion simple, but the ideas can be extended to arbitrary shapes.

4.3.2 Coded Illumination and Reconstruction

The effect of illumination can be modeled as an element-wise product between the scene and the illumination patterns. In our experiments, we use a laser projector placed next to the lensless camera to illuminate the object. In other applications, such as under-the-display cameras, an LED screen can be used for illumination. To simplify the recovery process, we further assume that the illumination patterns are separable and drawn from columns of $n \times k$ matrices P_L and P_R. Let us denote a pattern as $P_{i,j} = p_{Li} p_{Rj}^\top$, where p_{Li} and p_{Rj} are ith and jth columns of P_L and P_R, respectively. We can describe sensor measurements for any given illumination pattern $P_{i,j}$ as

$$Y_{i,j} = \Phi_L (P_{i,j} \odot X) \Phi_R^\top + E_{i,j}, \tag{4.5}$$

where $\odot$ represents element-wise multiplication operator and $E_{i,j}$ denotes measurement noise.

To recover image X from the sensor measurements $Y_{i,j}$, we can solve the following ℓ_2-regularized least-squares problem:

$$\underset{X}{\operatorname{argmin}} \sum_{i,j} \|Y_{i,j} - \Phi_L (P_{i,j} \odot X) \Phi_R^\top\|_2^2 + \lambda \|X\|_2^2, \tag{4.6}$$

where $\lambda > 0$ is a regularization parameter. An optimal solution of (4.6) must satisfy the following conditions (which can be derived by setting the gradient to zero) with $\mathbf{Q} = \sum_{i,j}(\Phi_L^\top Y_{i,j}\Phi_R) \odot P_{i,j}$:

$$\mathbf{Q} = \underbrace{(\Phi_L^T\Phi_L \odot P_L P_L^T)}_{\mathbf{A}_L} X \underbrace{(\Phi_R^T\Phi_R \odot P_R P_R^T)}_{\mathbf{A}_R} + \lambda X,$$

$$\Rightarrow \mathbf{Q} = \mathbf{A_L} X \mathbf{A_R} + \lambda X, \tag{4.7}$$

where $\mathbf{A_L}$, and $\mathbf{A_R}$ are $n \times n$ matrices. The solution of (4.7) can be written in closed form using the eigen-decomposition of $\mathbf{A}_L, \mathbf{A}_R$ [6] as

$$\hat{X} = \mathbf{V_L}[(\mathbf{V_L}^\top \mathbf{Q} \mathbf{V_R})./(\mathbf{s_L}\mathbf{s_R}^\top + \lambda \mathbf{1}\mathbf{1}^\top)]\mathbf{V_R}^\top, \tag{4.8}$$

where $\mathbf{V_L}, \mathbf{V_R}$ denote the eigenvectors and $\mathbf{s_L}, \mathbf{s_R}$ denote the eigenvalues of $\mathbf{A_L}, \mathbf{A_R}$, respectively, $./$ denotes element-wise division of entries in two matrices, and $\mathbf{1}$ denotes a vector with all ones.

A naïve approach would require storing all the measurements $Y_{i,j}$, which increases the storage complexity of the system proportional to the number of illumination patterns. The procedure described above avoids this cost, as we can recursively update an estimate of all the matrices and vectors needed for image recovery without any additional storage overhead. Thus, the storage cost of our method remains constant regardless of the number of illumination patterns. Every captured sensor measurement requires some processing, so the computational cost per recovered image increases linearly with the number of illumination patterns.

The required storage space for all the parameters is $\mathcal{O}(n^2)$ because $\mathbf{Q}$, $\mathbf{A_L}$, and $\mathbf{A_R}$ are $n \times n$ matrices. We only need to compute $\mathbf{A_L}$, $\mathbf{A_R}$ once, each of which costs $O(mn^2+kn^2)$. Eigen-decomposition of $n \times n$ matrices is $O(n^3)$. The most expensive step in our method is computing $\mathbf{Q}$, which we can perform by in-place addition of $(\Phi_L^\top Y_{i,j} \Phi_R) \odot P_{i,j}$ as we acquire measurements for all i, j. In this manner, we never need to store any of the captured measurements. The complexity of updating $\mathbf{Q}$ is $\mathcal{O}(k^2(nm^2 + mn^2))$.

4.3.3 Choice of Illumination Patterns

One of our goals is to select the $n \times k$ illumination pattern matrices P_L, P_R that maximize the quality of reconstruction for fixed Φ_L, Φ_R. The quality of reconstruction in (4.8) directly depends on the conditioning of the $\mathbf{A_L}$, $\mathbf{A_R}$ matrices in (4.7), which in turn depends on the mask and illumination patterns. One possible approach to improve the conditioning of $\mathbf{A_L}, \mathbf{A_R}$ is to make them diagonal or diagonally dominant [38], which we can achieve by enforcing the same structures in $P_L P_L^\top, P_R P_R^\top$.

In principle, we can make $P_L P_L^\top$ diagonal or even identity by using P_L as an identity matrix, which requires $k = n$. This would be equivalent to scanning the entire scene by illuminating one pixel at a time. We can also make $P_L P_L^\top$ identity by selecting P_L as any orthogonal matrix, which also requires $k = n$. In a practical scenario, we can only use a small number of illumination patterns; therefore, $k \ll n$. Below we discuss how we can get diagonally dominant $\mathbf{A_L}$, $\mathbf{A_R}$ using small values of k.

We propose to use illumination patterns that constitute an orthogonal basis over $k \times k$ blocks and repeat the same patterns across the entire scene. The simplest example of such patterns is a dot pattern in which two adjacent dots are placed k scene pixels apart. We can then shift the dot pattern across horizontal and vertical directions, one pixel at a time, to capture k^2 shifting dots patterns. These shifting dots patterns are separable and orthogonal over every $k \times k$ block. More generally, we can use any sequence of orthogonal separable patterns over $k \times k$ blocks. Let us assume the separable illumination patterns can be drawn from P_L, P_R that are defined as

$$P_L = P_R = \begin{bmatrix} \psi_k \\ \vdots \\ \psi_k, \end{bmatrix} \Rightarrow P_L P_L^\top = P_R P_R^\top = \begin{bmatrix} I_k & \dots & I_k \\ \vdots & \ddots & \vdots \\ I_k & \dots & I_k \end{bmatrix}, \tag{4.9}$$

where ψ_k and I_k denote $k \times k$ orthogonal and identity matrices, respectively. The resulting $P_L P_L^\top$, $P_R P_R^\top$ matrices (shown above) will be block matrices with $k \times k$ identity blocks, and the $\mathbf{A_L}$, $\mathbf{A_R}$ matrices (shown in Fig. 4.2) will be block matrices with $k \times k$ diagonal blocks.

Recall that $\mathbf{A_L}, \mathbf{A_R}$ are system matrices for the linear system we need to solve in (4.7). We can permute the rows and columns of $\mathbf{Q}$, which is equivalent to permuting rows of $\mathbf{A_L}$ and columns of $\mathbf{A_R}$, without affecting the solution of (4.7). Let us represent the resulting permuted equations as

$$\widetilde{\mathbf{Q}} = \widetilde{\mathbf{A}}_\mathbf{L} X \widetilde{\mathbf{A}}_\mathbf{R} + \lambda X. \tag{4.10}$$

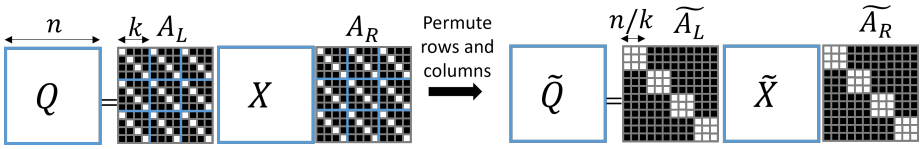


Figure 4.2: Illustration of system in (4.7) when the illumination patterns form orthogonal basis over $k \times k$ image patch. (Left) A_L and A_R are $n \times n$ block matrices with diagonal blocks of size $k \times k$. (Right) Permuting rows and columns results in block diagonal matrices with block size $\frac{n}{k} \times \frac{n}{k}$. The recovery performance of this system depends on the conditioning of each block. We can solve the block diagonal system by recovering $\frac{n}{k} \times \frac{n}{k}$ patches in X independently, in parallel.

We illustrate the permuted system in Fig. 4.2, where $\widetilde{\mathbf{A}}_\mathbf{L}, \widetilde{\mathbf{A}}_\mathbf{R}$ represent $n \times n$ block diagonal matrices, with k blocks along the diagonal each of size $\frac{n}{k} \times \frac{n}{k}$. As we increase the value of k, the system matrices $\widetilde{\mathbf{A}}_\mathbf{L}, \widetilde{\mathbf{A}}_\mathbf{R}$ become diagonally dominant and the overall conditioning of the system improves.

We can exploit the block diagonal structure of the system matrices to solve the system in (4.10) in a reliable and computationally efficient manner. Note that the separable, block diagonal system can be divided into k^2 independent systems, each involving an $\frac{n}{k} \times \frac{n}{k}$ patch in $\mathbf{X}$. We can solve all these systems in parallel to speed up recovery. The overall complexity of the inversion also reduces from $O(n^3)$ to $O(n^3/k)$. Furthermore, the conditioning of the overall system now depends on the conditioning of each $\frac{n}{k} \times \frac{n}{k}$ block in $\widetilde{\mathbf{A}}_\mathbf{L}, \widetilde{\mathbf{A}}_\mathbf{R}$. As long as all the blocks are well-conditioned, we can recover the underlying signal accurately.

4.4 Simulations

4.4.1 Simulation Setup

To validate the performance of the proposed algorithm, we simulate a lensless imaging system where a coded-mask is placed on top of an image sensor. We use a separable maximum length sequence (MLS) mask pattern. The size of each mask feature is 60μm, and the sensor-mask distance is 2mm. The sensor pitch in the simulation is 11.72μm and the total number of pixels on the sensors is 512×512. We simulate a 128×128 planar scene that is 40cm away from the sensor, and the height/width of the scene is 12cm. The simulated sensor noise includes photon noise and read noise [48], and the noisy sensor measurements can be described as

$$\mathbf{Y}_n = \frac{G}{F}(\text{Poisson}(\frac{F}{G}\mathbf{Y}) + N(0, \sigma^2)), \tag{4.11}$$

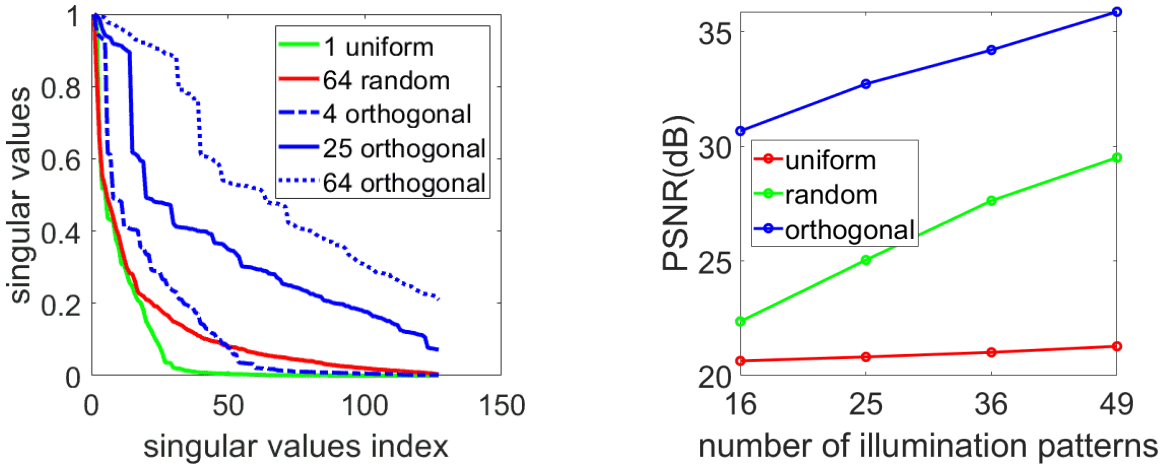
where $\mathbf{Y}$ and $\mathbf{Y}_n$ refers to original and noisy measurements, F stands for the full-well capacity, and G represents the gain value. The variance $\sigma = F \times 10^{-R/20}$ and R is the dynamic range. We show the reconstruction results on a few example scenes using different illumination patterns; additional results can be found in the supplementary material.

4.4.2 Effect of Illumination on Reconstruction

We first evaluate the conditioning of different illumination patterns by observing the singular values of the system matrices in (4.7). The matrices that have flat singular value spectrum provide better recovery performance [3, 6, 38]. We tested different types of binary, separable illumination patterns for this experiment. We generate different instances of matrices P_L, P_R and use

outer products of all pairs of columns to generate the illumination patterns. We ensure that the union of all the patterns should illuminate all the pixels (i.e., if we add columns of P_L, P_R, they should be nonzero everywhere). **Uniform:** One pattern that illuminates all the pixels simultaneously; P_L, P_R are vectors of all ones. **Random:** P_L and P_R are $k \times n$ binary random matrices that generate k^2 patterns. **Orthogonal:** We tested two types of orthogonal patterns (shifting dots and repeated Hadamard) that yield identical system matrices in (4.10). **Shifting dots:** P_L, P_R are $k \times n$ matrices, each of which consists of $k \times k$ identity matrices stacked on top of each other (as described in (4.9)). The base illumination pattern consists of dots separated by k pixels along the horizontal and vertical directions. We generate a total of k^2 illumination patterns, each of which is a shifted version of the base pattern. The summation of all the patterns will give us a uniform illumination pattern. **Repeated Hadamard:** As an extension to shifting dots, P_L, P_R are $k \times n$ matrices, each of which consists of the same $k \times k$ orthogonal Hadamard matrix stacked on top of each other. We can use grayscale or color patterns to illuminate the scene. In real experiments, the calibration of the projector and nonlinearity of color/intensity ranges pose additional challenges.

To evaluate the effect of illumination on the lensless imaging system, we observe the decay of singular values of the system matrices in (4.7) as we increase the number of illumination patterns. Figure 4.3a plots the singular values for different illumination patterns. The singular values of the original system matrix with one uniform illumination decay sharply. We tested various patterns and found that the shifting dots or orthogonal patterns provide the best overall conditioning (flat SVD curve) for a given budget of measurements. As we increase the number of illumination



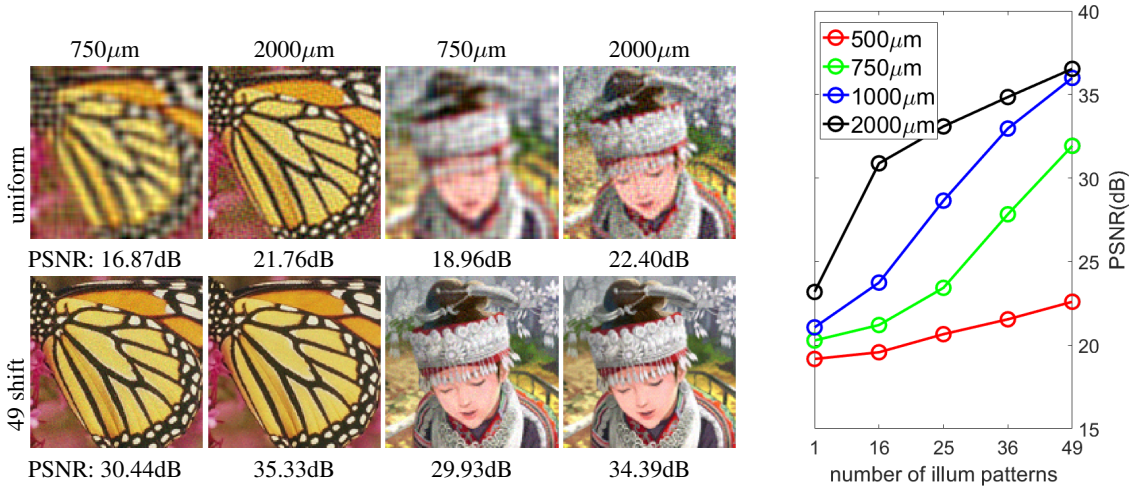
(a) System singular values.

(b) Average PSNR values.

Figure 4.3: Recovery performance of the imaging system. (a) Singular values of the system matrices with uniform, 64 random, and 4, 25, and 64 orthogonal shifting dots patterns. (b) Average PSNR of 8 test images reconstructed with different numbers of illumination patterns. Red dashed line shows results with uniform illumination. Reconstruction quality improves as we increase the number of illumination patterns. The orthogonal patterns outperform random and uniform patterns.

patterns, the singular values spectrum becomes flatter, which corresponds to a system that is nearly orthogonal. In principle, we can use a dot projector to create shifting dots patterns, but we need a programmable projector for Hadamard-like patterns. In our experiments, we use a laser projector for illumination.

We present simulation results for reconstruction of 8 test images using different types and number of illumination patterns in Fig. 5.2. We observe that orthogonal patterns outperform other patterns in terms of PSNR. Additional simulation results are available in the supplementary material.



(a) Examples of reconstructed images with sensor-to-mask distance at 750μm and 2000μm.

(b) Average PSNR of all test images.

Figure 4.4: Simulation results for reconstruction with different sensor-to-mask distances in uniform illumination and shifting dots illumination patterns. The reconstruction quality improves as the sensor-to-mask distance increases.

4.4.3 Effect of Sensor-to-Mask Distance

The sensor-to-mask distance of a lensless imaging system greatly influences the conditioning of the system and the quality of reconstruction. We present simulation results to evaluate the performance with different sensor-to-mask distances. We keep the sensor size fixed at 512×512 pixels, and test the sensor-to-mask distance at 500μm, 750μm, 1000μm, and 2000μm. The reconstruction results for two test images are presented in Fig. 4.4, along with the PSNR plot for different numbers of shifting dots patterns. We observe that multiple coded illumination patterns outperform the results with a single uniform pattern at all the sensor-to-mask distances. Even if the sensor-to-mask distance is 750μm, the results with 25 coded illumination patterns are comparable to the uniform illumination results at 2000μm. Additional simulation results are available in the supplementary material.

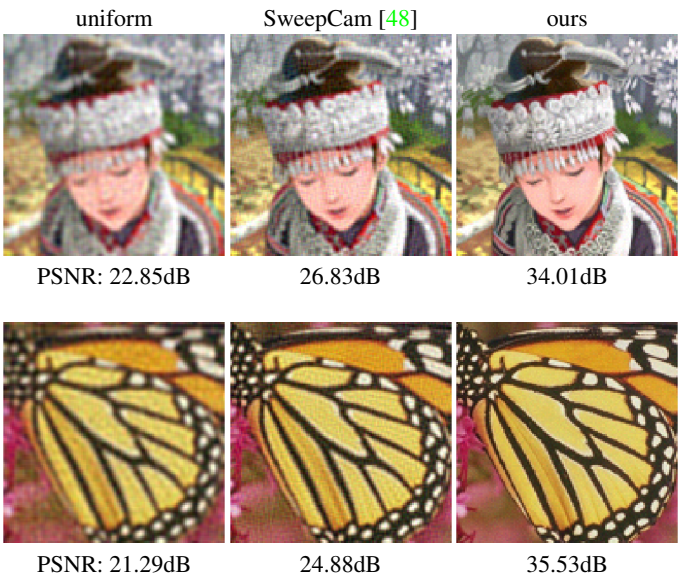
4.4.4 Comparison With Multishot Lensless Methods

Multishot lensless methods are also applied in [48, 125], where multiple frames of lensless measurements are captured with a shifting or programmable mask on top of the image sensor. We present the simulation results comparing with a shifting mask-based multishot lensless imaging system [48] in Fig. 4.5. We keep the sensor size fixed at 512×512 and sensor-to-mask distance at 2000μm. The mask feature size is 60μm in all cases. We simulate the imaging process for SweepCam [48] with the mask at multiple shifting positions between -15 and $+15$ mask feature pixels (i.e., -900μm to $+900\mu$m physical shift). The number of captured frames for SweepCam [48] are the same as our method with coded illumination patterns. The simulation results in Fig. 4.5 show that our method with the same number of coded illumination patterns provides significantly better results compared to SweepCam.

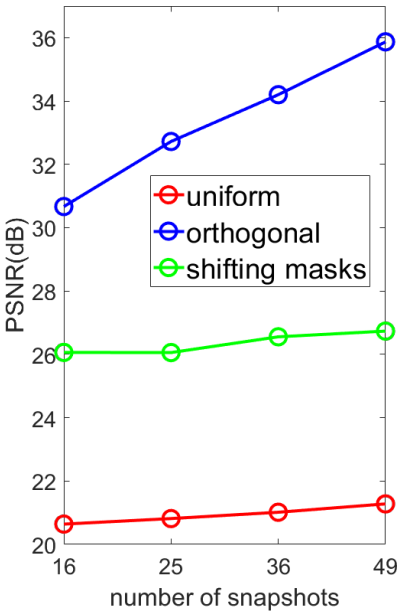
4.5 Experiments

4.5.1 Experiment Setup

We build a prototype with a lensless camera and a Sony MP-CL1 laser projector, shown in Fig. 4.6. The lensless camera prototype consists of an image sensor with a coded mask on top of it. The mask has a separable MLS pattern described in [6]. The mask has 511×511 square features, each of length/width 60μm. The sensor-mask distance is 2mm. We use a Sony IMX249 sensor that has 1920×1200 pixels with 5.86μm pixel pitch. We bin 2×2 sensor pixels and record 512×512 measurements from the center of the sensor. The effective sensor pitch is 11.72μm and the effective



(a) Example results with uniform illumination, 49 shifting masks, and 49 shifting dots.



(b) Average PSNR of all test images.

Figure 4.5: Simulation results for reconstruction using uniform illumination, shifting masks from SweepCam [48] and shifting dots illumination patterns using our method at different number of measurements instances. Our method outperform uniform the other methods.

sensor area is nearly 6×6 mm. The target objects are 40cm away from the camera, and the projector illuminates 12×12 cm area on the scene plane. Finally, we reconstruct 128×128 pixels in the illuminated area, which results in the effective sampling interval of 120mm/128 = 0.93mm per pixel in the reconstructed images.

In our experiment, the pixel grid of the scene, illumination patterns P_L, P_R, the system matrices Φ_L, Φ_R must be correctly aligned; otherwise, we will get artifacts in the reconstruction. To avoid any grid mismatch, we use the same projector to calibrate the system matrices and generate the illumination patterns in our experiments. We estimate the system matrices Φ_L, Φ_R following the Hadamard pattern-based calibration procedure in [6]. Let us denote the Hadamard patterns as an $n \times n$ orthogonal matrix $\mathbf{H} = [\mathbf{h}_1, \ldots, \mathbf{h}_n]$, where each $\mathbf{h}_i$ represents a Hadamard pattern of

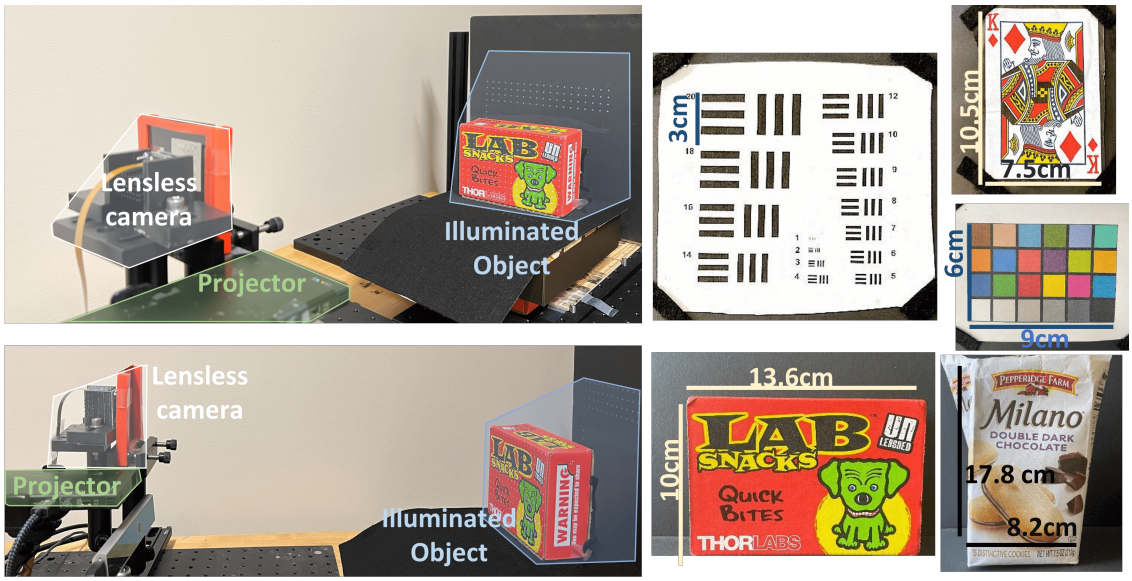


Figure 4.6: The experiment setup and five test scenes (annotated and scaled to proportional size). The projector is placed right next to the lensless camera. The target scenes/objects are 40cm away from the camera.

length n. To calibrate the system for an $n \times n$ pixel grid, we project n horizontal and n vertical (rank-one) Hadamard patterns on a flat surface in the scene and record their response on the sensor. Every horizontal pattern can be represented as an $n \times n$ rank-one matrix $X_i = \mathbf{h}_i \mathbf{1}^\top$, where $\mathbf{1}$ denotes a vector of all ones. The corresponding sensor response can be represented as a rank-one matrix $Y_i = \Phi_L(\mathbf{h}_i \mathbf{1}^\top)\Phi_R^\top \equiv \mathbf{u}_k \mathbf{v}^\top$, where $\mathbf{u}_i = \Phi_L \mathbf{h}_i$ and $\mathbf{v} = \Phi_R \mathbf{1}$. We can concatenate all the $\mathbf{u}_i$ as columns in a matrix as $\mathbf{U} = [\mathbf{u}_1 \ \cdots \ \mathbf{u}_n] = \Phi_L \mathbf{H}$ and estimate $\Phi_L = \mathbf{U}\mathbf{H}^\top$. We can repeat the same procedure with vertical Hadamard patterns $\mathbf{1}\mathbf{h}_i^\top$ and estimate Φ_R. Finally, we conduct the experiments with coded illumination while the position and angle of the projector are fixed. This ensures that the pixel grids of the projector and the transfer matrices are identical.

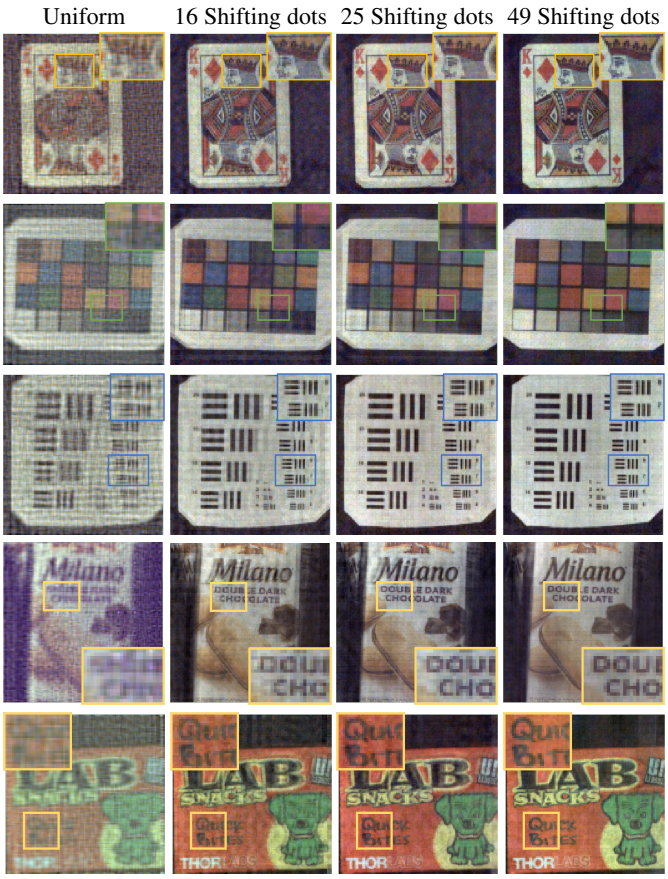


Figure 4.7: Experimental results on five test scenes with different numbers of shifting dots patterns. The reconstruction results using multiple coded illumination patterns outperform the standard lensless camera with uniform illumination. The quality of reconstruction gradually increases as the number of illumination patterns increases.

4.5.2 Effect of Illumination Patterns

We present experimental results of our method on five different scenes in Fig. 4.7 and Fig. 4.8. The first three rows are printed card, color board, and a resolution target on sheets of paper. Since these scenes are planar, they fit the 2D scene assumption of our model perfectly. The last two rows are real objects with depth variations and may cause depth mismatch in the reconstruction.

From the results in Fig. 4.7, we observe that a single uniform illumination provides poor reconstruction. As we increase the number of illumination patterns, the resolution of the reconstruction images improves significantly. For example, in the third row, the horizontal and vertical features that cannot be resolved with the uniform illumination are easily resolved with the 49 patterns. Also, we observe in the fourth row that, the letters on the cookie bag that are completely unrecognizable with uniform illumination become very clear with 49 illumination patterns. The last two rows also suggest that our method is robust to small variations in depth. From the results in Fig. 4.8, the repeated orthogonal patterns(shifting dots and Hadamard) outperform other patterns, as expected from the singular values in Fig. 4.3a.

Note that the neighboring pixels have a similar response on the sensor in the lensless imaging system; therefore, neighboring pixels are harder to resolve compared to two pixels that are far from each other. The shifting dots pattern is a dot array where the illuminated pixels are maximally separated; therefore, two neighboring pixels in the scene are not illuminated at the same time. Also, as we increase the number of shifting dots patterns, the distance between adjacent illuminated pixels also increases, and that provides better reconstruction. Hadamard patterns provide similar separation because of their orthogonal structure.

Note that if we increase the number of measurements with coded illumination, the system conditioning and the quality of reconstructed images improve. On the other hand, if we increase the number of measurements for the uniform illumination pattern, the system conditioning and the quality of reconstructed images remains unchanged. To show this effect, we provide experimental results comparing the quality of reconstruction with multiple measurements captured under uniform and coded illumination. We present the results using 49 uniform and 49 shifting dots illumination

patterns in Fig. 4.9. Note that the 49 uniform and 49 shifting dots illuminations use the same amount of exposure/capture time. We observe that shifting dots illumination provides significantly better reconstruction compared to uniform illumination. We observe that capturing 49 shots with uniform illumination reduces noise slightly compared to one uniform illumination, but 49 shifting dots provide significantly superior results. Note that multiple shots with uniform illumination can provide slightly better SNR because noise variance reduces due to averaging. The improvement with shifting dots patterns is the result of the better conditioning of the system that multiple uniform illumination shots cannot provide.

In summary, the ill-conditioned system matrices with uniform illumination pattern cause various artifacts. Capturing measurements from coded illumination improves the conditioning of the overall system and the resulting reconstructed images contain much fewer artifacts and noise. More illumination patterns demand more acquisition time, which enforces a trade-off between the quality of reconstruction and data acquisition time.

4.5.3 Effective Resolution and MTF

We use the resolution target image in Fig. 4.10 to analyze the spatial resolution of our method empirically. The distance between two printed white stripes in group 20 (upper-left) is 7.5mm (0.13 lp/mm) and the distance between two white stripes in group 5 is 1.9mm (0.52 lp/mm). The resolution of our imaging system is directly determined by the sampling interval of the illumination patterns. We place the target scene 40cm away from the camera and projector. The resolution of the MP-CL1 projector is 1280×720 and the overlap between the camera FOV and the maximum illuminating area at 40cm throw distance is 22×13cm. The projector pixel pitch is 0.17mm, which

determines the achievable resolution of our method with the MP-CL1 projector. In our experiments, the width of every reconstructed pixel is 0.93mm, and the angular sampling of our system is 0.27°. We can select smaller sampling intervals pitch by dividing the illuminating area into more pixels, but the minimal size cannot be smaller than projector pixel.

We present the modulation transfer function (MTF) of the reconstruction from different numbers of illumination patterns in Fig. 4.10. MTF measures how well we can discern the intensity of bright and dark pixels in line pairs under different conditions [112]. To plot the MTF, we manually select every group of horizontal and vertical line pairs after subtracting a fixed DC background, then average every group along the columns and rows, respectively, and finally compute the contrast percentage as $\frac{I_{\max}-I_{\min}}{I_{\max}+I_{\min}} \times 100$. We observe from the MTF plots that the contrast ratio for all the resolution groups improves as we increase the number of illumination patterns. We also plot the intensity values of the horizontal line stripes from multiple groups in Fig. 4.10, where the peaks and valleys correspond to the white and black stripes of the horizontal line pairs. We observe that the line pairs of group 12 can be distinguished clearly while the line pairs of group 3 cannot be distinguished.

4.5.4 Compressive Sensor Measurements

One potential application of coded illumination with lensless imaging is to expand the space of sensor measurements without increasing the number of sensor pixels. This can be especially useful when sensors capture fewer measurements than the number of pixels we seek to recover. Compressive sensing often deals with such scenarios where the number of sensor pixels is

smaller than the number of unknown scene pixels [21, 31]. The system becomes under-determined and the reconstruction quality becomes worse. In lensless imaging systems, the number of sensor measurements is often larger than the number of reconstruction pixels, but that comes at the expense of low resolution of reconstructed images or larger and more costly sensors.

To validate the robustness of our method in the case of an under-determined system, we performed an experiment by binning the sensor measurements at increasing factors. The results are presented in Fig. 4.11. In our experiments, we capture 512×512 measurements and bin them to 256×256, 128×128, or 64×64 pixels by averaging the neighboring pixels in post-processing. The reconstruction image has 128×128 pixels, which implies the 64×64 sensor measurements yield an under-determined system with one illumination. We present the result of two scenes with different binning factors. We observe that as we bin a larger number of pixels, the system with a single uniform illumination becomes ill-conditioned and the quality of reconstruction degrades. On the other hand, the system with 49 shifting dots patterns provides stable reconstruction even when the measurements are binned to 64×64 pixels.

4.5.5 Deep Network-Based Denoising vs Reconstruction

Deep learning-based methods have been widely used for image recovery and enhancement tasks [28, 52, 73]. For instance, UNet [86] is used for image denoising and removing artifacts from reconstructed images [52,125]. Deep learning is also used as the priors to improve the reconstruction results [15], especially when the number of measurements is small. UNet-based networks have been used to recover photorealistic images from lensless measurements in [52]. In our experiments, we observed that deep learning-based methods that learn to recover images directly from sensor

measurements perform very well on simulated data but provide catastrophic results on real data. In contrast, the simple least-squares method we discussed in (4.8) provides stable results in all cases because coded illumination provides a well-conditioned system. We present simulation and experimental results below.

We performed some experiments to compare the performance of four methods with coded and uniform illumination: least squares (LS) in (4.8), LS with a trained UNet that is used as a refinement network, LS with pretrained refinement network in FlatNet [52], and end-to-end trained FlatNet. We present results for simulated sensor measurements in Fig. 4.12 and on real data captured using our prototype in Fig. 4.13.

The description of the four methods is as follows. We used 1000 natural RGB images from [25] to train the deep networks.

- **LS.** Images are reconstructed using the least-square (LS) method with ℓ_2-regularization (4.8).
- **LS + Trained UNet.** We trained a UNet using 1000 training images that learns to map the LS reconstruction into a refined image.
- **Trained FlatNet.** FlatNet [52] contains two steps: inversion step and refinement step that are trained jointly in an end-to-end manner. The inversion step maps the sensor measurements into a coarse image estimate, while the refinement step maps the coarse estimate to a cleaner image. We train both steps of the FlatNet in an end-to-end manner with simulated measurements for uniform and coded illuminations. We created the synthetic data by simulating the noisy lensless measurements for training images using the calibrated transfer matrices Φ_L, Φ_R described in Sec.4.5.1. Note that we need to train a separate network for different types/number of illumination patterns.

- **LS + Pretrained FlatNet.** The refinement network in FlatNet has same architecture as our trained UNet. For the sake of comparison, we also used the pretrained refinement network from FlatNet to refine the LS solution.

Figure 4.12 shows images reconstructed from simulated measurements of selected test images. We observe that LS, LS+UNet, and trained FlatNet provide good reconstruction for simulated data using uniform illumination. In particular, images from trained FlatNet have best visual appearance simulations Fig. 4.12(d). For coded illumination, the quality of reconstruction improves for all the methods. The pretrained refinement network from FlatNet provides overly smoothed images and distorts spatial details.

Figure 4.13 shows images reconstructed from measurements captured by real prototype. Trained FlatNet fails to recover images from real data captured with uniform illumination in Fig. 4.13. In contrast, LS and LS+UNet provide a low-resolution reconstruction with uniform illumination. For coded illumination, trained FlatNet reconstructions have high quality that is similar to the results provided by LS and LS + trained UNet. Figure 4.13 further shows that UNet can partially reduce the artifacts and noise in the images reconstructed from the uniform illumination, but fails to improve the spatial resolution. In contrast, coded illuminations significantly improves the resolution and quality of the reconstructed images. Additional results can be found in the supplementary material.

Our main takeaway from these experiments is that deep networks cannot always recover missing details if the system is ill conditioned. In particular, deep network-based refinement/denoising can improve the appearance of images but cannot recover missing details. End-to-end trained networks, which recover images directly from multiplexed sensor measurements, perform well when train and test settings are identical, but provide catastrophic results in case of any mismatch. Coded illuminations improve the conditioning of the system, which benefits all the recovery methods.

4.6 Discussion

We propose a framework for combining coded illumination with lensless imaging. We present extensive simulation and real experiment results to demonstrate that we can get significantly improved reconstruction with multiple coded illumination patterns compared to original uniform illumination.

Advantages

- Qualitative experiment results show that our method has robust performance even if the original system is severely ill-conditioned (small sensor-mask distance and small number of measurements).
- In space-limited applications such as under-the-display sensing, where the sensor-to-mask distance has to be small, our proposed method can offer significantly better reconstruction compared to uniform illumination.

- Our proposed reconstruction algorithm is efficient in both space and time. For shifting dots (and any other orthogonal) pattern over $k \times k$ blocks, the system can be transformed into k^2 independent small problems that can be solved in parallel for faster and better reconstruction.
- Shifting dots patterns have a simple structure but provide the best results in our experiments. Shifting dots patterns can be easier to implement because all the patterns are simply the shifted copies of the same base pattern.
- Programmable orthogonal patterns would provide similar gains as shifting dots with potentially better performance in strong ambient light (at the expense of complex hardware).

Limitations

- The resolution of reconstructed images are limited by the resolution of illumination patterns.
- Coded illumination hardware can bring additional cost and complexity to the system design.
- Our current system cannot capture dynamic scenes because our model requires multiple measurements of every scene under different illumination patterns.
- The shifting dots patterns are simple and easy to implement, but they reduce the light throughput.
- The camera and projector are co-located in our setup; therefore coded illumination does not provide a direct advantage toward depth estimation. Reconstruction of 3D and dynamic scenes with this framework is possible and requires additional work [123].
- Similar to other active illumination systems, the performance of our method will suffer under strong ambient lights (see supplementary).
- Our imaging model does not account for non-Lambertian objects, and any such object can cause artifacts in the reconstruction (see supplementary).

4.6.1 Error Analysis in Multi-Shot System

In this section, we discuss the effects of noise and system conditioning on the reconstruction error. In our coded illumination method (especially, with shifting dots), each frame of the multi-shot system can have low SNR compared to single-shot system, given the same capture time. While the SNR of each frame is worse, the multi-shot systems offer much better conditioning, which in turn provides better reconstruction.

Let us represent the measurements captured by a reference lensless camera under uniform illumination in the following vectorized form: $y = Ax + e$, where A denotes the system matrix, x denotes the unknown image, and e denotes the measurement noise. We denote the measurement signal to noise ratio of this reference system as SNR_{ref}. Note that under the assumption that photon count can be approximated as $\mathcal{N}(\lambda t, \lambda t)$, $\text{SNR}_{ref} \propto \sqrt{\lambda t}$, where λ represents rate of photon arrival and t represents exposure time).

Suppose A is full rank but ill-conditioned with singular values $s_1, \ldots, s_N$, and the least-squares solution is $\hat{x}$. If we capture T measurements with uniform illumination (or increase exposure time of a single measurement by a factor of T), then the measurement SNR will increase by a factor of $\sqrt{T}$ but the singular values remain unchanged. A simple derivation (see details in [92]) can show that mean signal to mean reconstruction error ratio (SER) can be written as

$$\text{SER}_{unif} = \frac{\mathbb{E}\|x\|_2}{\mathbb{E}\|x - \hat{x}\|_2} = \frac{\sqrt{T}\,\text{SNR}_{ref}}{\sqrt{\text{Tr}(A^\top A)^{-1}}} = \frac{\sqrt{T}\,\text{SNR}_{ref}}{\rho_{ref}}, \tag{4.12}$$

where $\rho_{ref} = \text{Tr}(A^\top A)^{-1} = \sqrt{\sum_{i=1}^{N} 1/s_i^2}$. The reconstruction error is dominated by the small singular values as they will increase the denominator.

On the other hand, using coded illuminations will provide us a modified system matrix $\tilde{A}$ with singular values $\tilde{s}_1, \ldots, \tilde{s}_N$ that decay at a slower rate (as shown in Fig. 4.3a). The SNR of coded-illumination measurements can be different from uniform illumination; for example, if we turn on every kth pixel in the illumination pattern, the measurement SNR will reduce by a factor of $\sqrt{k}$. In our equivalent design with shifting dots, we would capture $T = k$ measurements with shifting dots. Therefore, the mean signal to mean reconstruction error ratio (SER) with T shifting dots illumination patterns can be written as

$$\text{SER}_{dots} = \frac{\mathbb{E}\|x\|_2}{\mathbb{E}\|x - \hat{x}\|_2} = \frac{\text{SNR}_{ref}}{\rho_{dots}}, \tag{4.13}$$

$\rho_{dots} = \text{Tr}(\tilde{A}^\top \tilde{A})^{-1} = \sqrt{\sum_{i=1}^{N} 1/\tilde{s}_i^2}$. Note that if we use T Hadamard patterns instead of shifting dots, then the SNR per measurement would not reduce, and the error bound above will not have the $\sqrt{T}$ factor in the denominator.

In summary, suppose we use same total exposure time (or number of equal-exposure measurements) for uniform and shifting dots patterns, then

$$\frac{\text{SER}_{dots}}{\text{SER}_{unif}} = \frac{1}{T}\frac{\rho_{ref}}{\rho_{dots}}. \tag{4.14}$$

In practice, the coded illumination-based system offers significantly better conditioning compared to reference system; that is, $\rho_{ref} \gg \rho_{dots}$.

We also show experimental results for multi-shot coded illumination system and uniform pattern system with the same amount of exposure time in Fig. 4.9(c,d). We can observe that the 49 shifting dots outperform the reconstruction using uniform illumination.

4.6.2 Coded Illumination With Nonseparable Systems

All the lensless cameras can be modeled as a linear system; therefore, the broader finding that adding coded illumination improves the conditioning of the system is valid across all the lensless designs. Our particular choice of separable masks and illumination patterns is indeed specific to FlatCam-inspired designs, where we gain computational advantages as the solution can be written in a separable closed form.

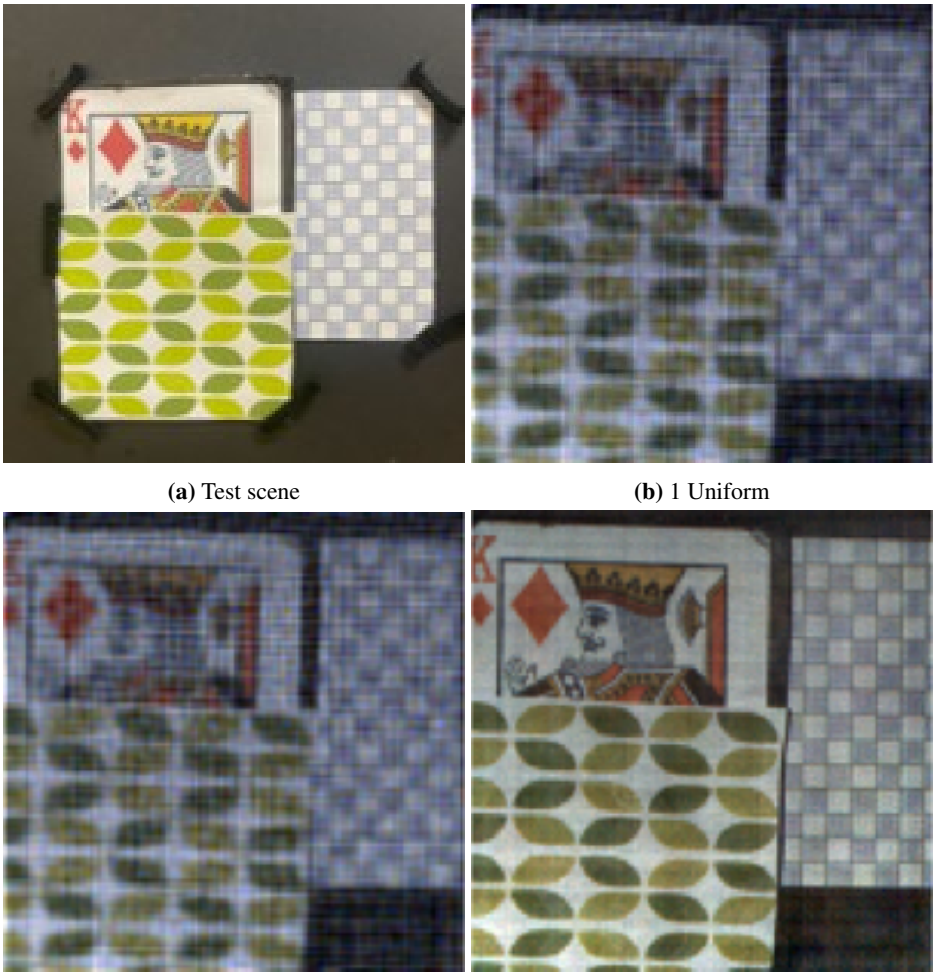
In principle, we can add coded illumination to convolutional models such as DiffuserCam as $\mathbf{Y}_i = \varphi * (\mathbf{P}_i \odot \mathbf{X})$, where φ denotes PSF for the mask, $\mathbf{X}$ denotes the unknown image, $\mathbf{P}_i$ denotes ith illumination pattern, and $\mathbf{Y}_i$ denotes the corresponding sensor measurements. We do not have a closed-form solution for such an imaging model; therefore, reconstruction involves using iterative solver.

To validate our argument, we also provide simulation results in Fig. 4.14 combining our coded illumination system with the convolution model proposed in DiffuserCam [3]. The PSFs used in the simulation are also from the DiffuserCam data. We observe in Fig. 4.14 that the coded illumination pattern also improves the condition of the DiffuserCam.

Combining a convolution model with coded illumination has some additional practical challenges that do not arise in the separable model we presented earlier in the paper. Two main challenges we encounter are (1) alignment of spatial grids and (2) reconstruction complexity. The alignment process is necessary to avoid any model mismatch. The sampling grid in the convolutional model cannot be arbitrary, as it is deteremined by the PSF and sensor pixel pitch. The alignment requires additional (tedious) calibration steps to estimate relative sizes/displacements of illumination pixels and scene pixels. We leave the calibration and hardware implementation tasks

for future work. In contrast, in our paper, we used a separable model that can be calibrated using the illumination pattern itself (as long as illumination grid yields a separable response on the sensor). Coded illumination-based convolutional models do not have a simple closed-form solution because the combined operator cannot be diagonalized with Fourier transform. We can implement the forward and adjoint operators using Fourier transform, but the overall reconstruction procedure is much slower than the separable model we used in our paper, as shown in Fig. 4.14c.

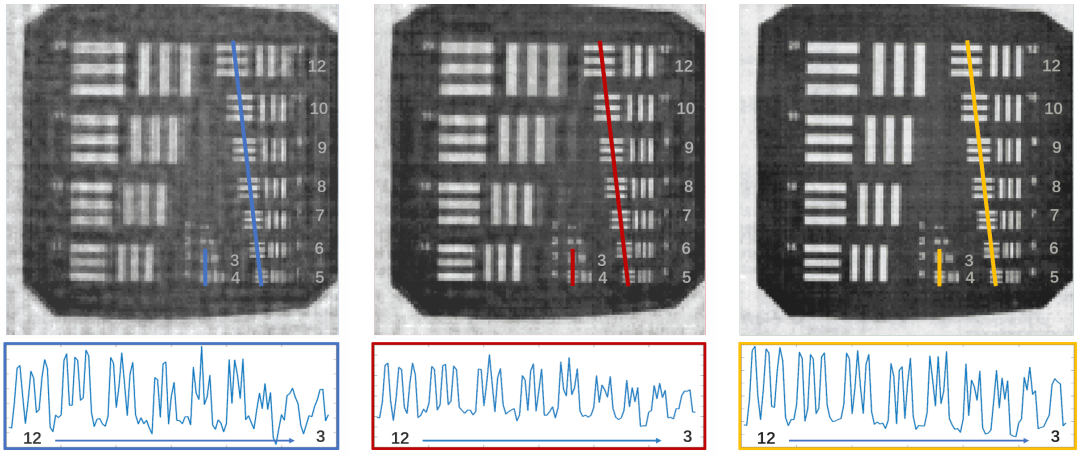
Figure 4.8: Experimental results on five test scenes with different types of illumination patterns. The orthogonal patterns (shifting dots, hadamard) outperform random patterns and uniform illumination.



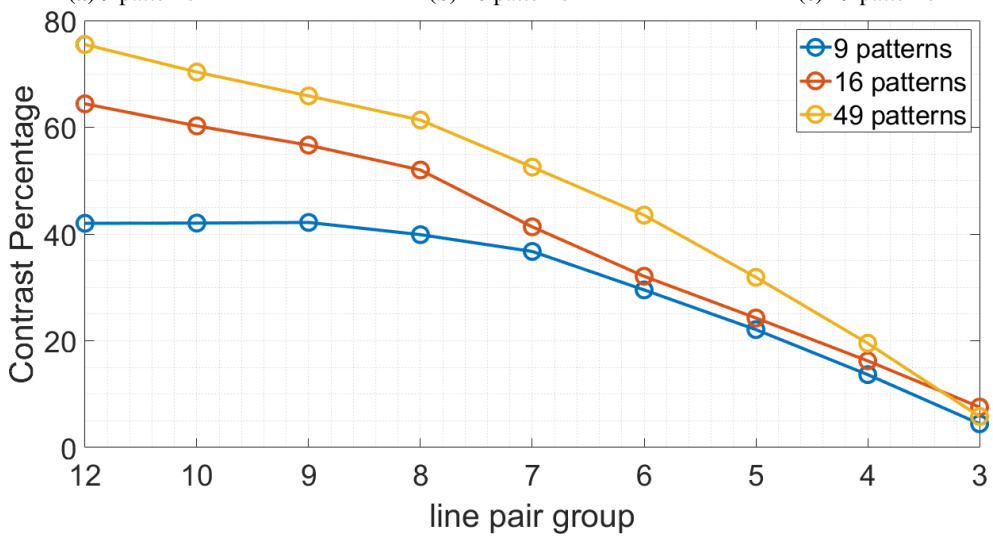
(a) Test scene **(b)** 1 Uniform

(c) 49 Uniform **(d)** 49 Shifting dots

Figure 4.9: Sample results for imaging performance with one uniform, 49 uniform, and 49 shifting dots illumination patterns. Capturing 49 shots requires the same data acquisition time, but the results for 49 shifting dots patterns are significantly better than 49 uniform patterns.



(a) 9 patterns **(b)** 16 patterns **(c)** 49 patterns



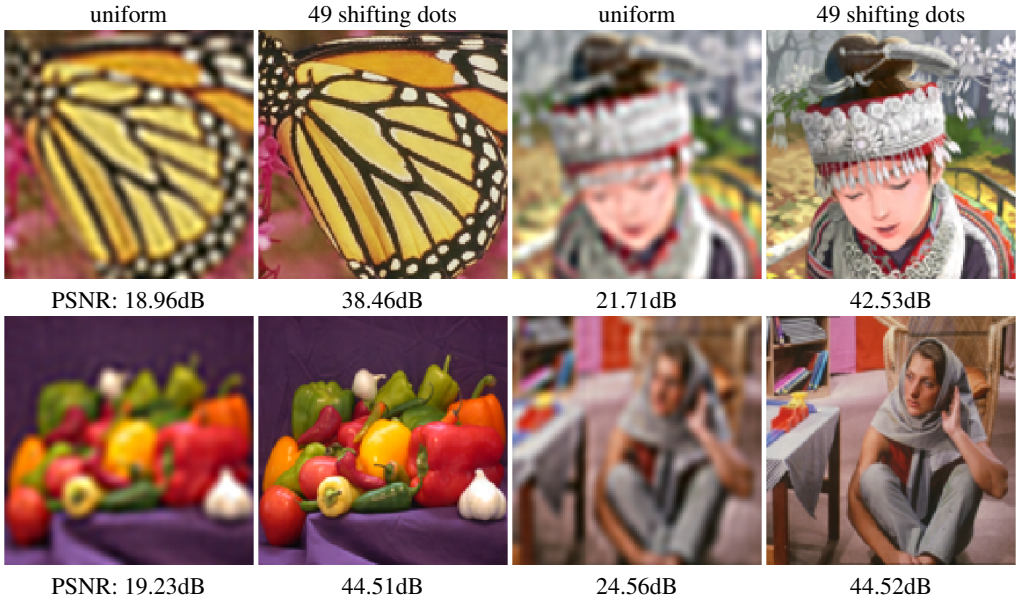
(d) Modulation transfer function plot. The vertical axis shows the contrast ratio in percentage.

Figure 4.10: Resolution analysis of coded illumination. Top images in (a,b,c) show resolution target reconstructed using 9, 16, and 49 shifting dots patterns. Bottom plots in (a,b,c) show the intensity of a line from group 12 to group 3. The MTF (modulation transfer function) plot for different numbers of shifting dots illumination patterns is shown in (d). To compute MTF, we manually select each group of horizontal and vertical line pairs after subtracting a fixed DC background, then average every group along the columns and rows, respectively, and finally compute the contrast ratio.

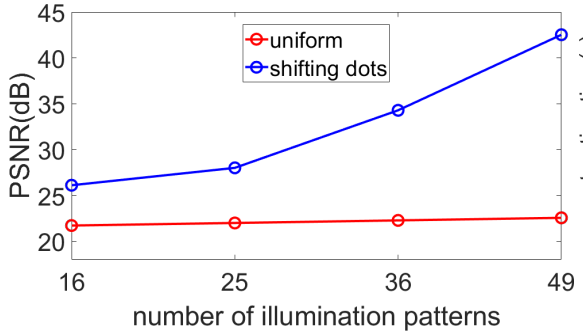
Figure 4.11: Experimental results for reconstructing the 128×128 images from 64×64 and 128×128 measurements. Single (uniform) illumination-based method fails to recover images as the number of measured pixels reduce. Our method with 49 shifting dots patterns recovers near-perfect reconstruction at different levels of binning (compression) factors.

Figure 4.12: Reconstruction results for simulated measurements with 49 uniform and shifting dots illumination patterns. Images in four columns show (a) LS solution, (b) LS solution with trained UNet refinement, (c) LS solution with pretrained FlatNet refinement, and (d) trained FlatNet that reconstructs image directly from measurements. For each image, we show the SSIM value underneath.

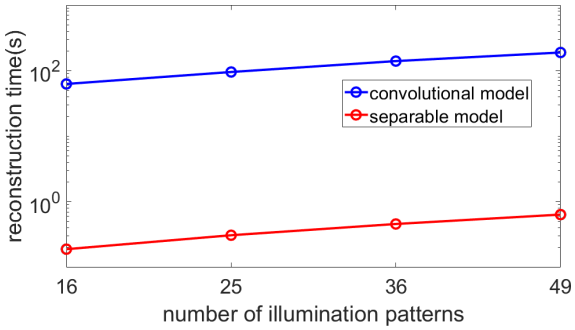
Figure 4.13: Reconstruction results for real-hardware measurements with 49 uniform and shifting dots illumination patterns. Images in four columns show (a) LS solution, (b) LS solution with trained UNet refinement, (c) LS solution with pretrained FlatNet refinement, and (d) trained FlatNet that reconstructs image directly from measurements.



(a) Example reconstructed images using DiffuserCam model [3].



(b) Average PSNR values.



(c) Average reconstruction time.

Figure 4.14: Simulation results with uniform illumination and 49 shifting dots coded illumination patterns using the convolution model in DiffuserCam [3]. (a) Sample reconstructed images in (a) and average reconstruction PSNR vs number of patterns in (b) show that coded illumination patterns improve the imaging performance of DiffuserCam. (c) Average reconstruction time for the convolutional model is an order of magnitude larger than the separable model that has a simple closed-form solution.

Chapter 5

Binocular 3D Lensless Imaging With Coded Illumination

Mask-based lensless cameras offer a novel design for imaging systems by replacing the lens in a conventional camera with a layer of coded mask. Each pixel of the lensless camera encodes the information of the entire 3D scene. Existing methods for 3D reconstruction from lensless measurements suffer from poor spatial and depth resolution. This is partially due to the system ill conditioning that arises because the point-spread functions (PSFs) from different depth planes are very similar. In this chapter, we propose to capture multiple measurements of the scene under a sequence of coded illumination patterns to improve the 3D image reconstruction quality. In addition, we put the illumination source at a distance away from the camera. With such baseline distance

between the lensless camera and illumination source, the camera observes a slice of the 3D volume, and the PSF of each depth plane becomes more resolvable from each other. We present simulation results along with experimental results with a camera prototype to demonstrate the effectiveness of our approach.

5.1 Introduction

Lensless cameras provide novel designs for extreme imaging conditions that require small, thin form factor, large field-of-view, or large-area sensors [3, 6, 12, 14]. Compared to conventional lens-based cameras, lensless cameras are flat, thin, light-weight, and geometry flexible. Depth estimation with lensless imaging has been a challenging problem [1, 3, 122]. The primary reason is that the sensor responses for different depth planes have small differences, which makes the 3D reconstruction an ill-conditioned problem.

In this chapter, we propose a new method that combines coded illumination with mask-based lensless cameras (such as FlatCam [6]) to improve the quality of recovered 3D scenes. We project a sequence of coded illumination patterns onto the 3D scene and capture multiple frames of lensless measurements. We then solve an inverse problem to recover the 3D scene volume using all the coded measurements. Coded illumination-based measurements provide a better-conditioned system and improve the quality of 3D reconstruction. The illumination source is separated from the camera by a baseline distance, which ensures that the depth-dependent point spread functions (PSFs) of each depth plane is different from one another. The choice and design of the illumination source depend on the application of the imaging system. We use a projector installed next to the lensless camera as the illumination source.

The main contributions of this chapter are as follows.

- We propose a novel framework to capture lensless measurements under a sequence of coded illumination patterns and improve the 3D reconstruction results.
- We show that the baseline between projector and camera cause depth-dependent shifts of PSF and enhance the 3D performance at large distances.
- We provide simulation and experimental results to validate the proposed method. Our experiments show that the quality of 3D reconstruction improves significantly with coded illumination.

5.2 Background and Related Work

Mask-based lensless cameras, such as FlatCam [6], can be viewed as extended versions of pinhole cameras. Although a pinhole camera is able to image the scene directly on a sensor, it often suffers from severe sensor noise [118]. Coded aperture-based cameras alleviate this problem by using multiple pinholes arranged in a designed pattern [6, 12, 19, 23, 34]. The scene is reconstructed by solving an inverse problem using the linear multiplexed lensless measurements. With the small baseline between the pinholes on the mask, the coded aperture-based cameras are also able to capture the depth information of the scene [1, 3, 5, 48, 56, 122, 125]. 3D reconstruction using a single snapshot of a lensless camera is an under-determined and highly ill-conditioned problem [122].

Signal recovery from ill-conditioned and under-determined systems is a long-standing problem in signal processing. A standard approach to deal with ill-conditioned and under-determined systems is to add a signal-dependent regularization term in the recovery problem, which constrains the range of the solutions. Popular methods include adding sparse and low-rank priors [8, 21, 31, 81, 87] or natural image prior [15, 43, 49]. Recently, a number of methods have been proposed that use deep networks to reconstruct or post-process the images from lensless measurements [13, 52, 71, 73]. Some of these methods provide exceptional improvement over traditional optimization-based methods. Nevertheless, deep learning-based methods in general, and end-to-end methods in particular, provide a huge variation in performance for simulated and real data (mainly because of mismatch in the simulated/actual mask-sensor-projector configuration and scenes). In contrast to deep learning methods, our method seeks to improve the conditioning of the underlying linear system and offer better generalization and robust results for arbitrary scenes without the need for learning from data [48, 125].

Our proposed approach can be viewed as an active imaging approach combining coded modulation or structured illumination method with coded aperture imaging [35, 42, 75]. Structured illumination schemes are commonly used for imaging beyond diffraction in microscopy. These schemes use multiple structured illumination patterns to down-modulate high spatial frequencies in a sample into a low-frequency region that can be captured by the microscope [41, 42, 46]. Coded illumination for lensless imaging of 2D scenes was recently presented in [124]. Another active imaging approach uses time-of-flight sensors [45, 88] that estimate the 3D scene by sending out infrared light pulses and measuring the traveling time of their reflections.

5.3 Technical Approaches

5.3.1 Imaging Model

Mask-based lensless cameras replace the lens with a layer of coded mask and capture linear multiplexed measurements with an image sensor. The mask pattern can be placed parallel to the sensor plane at distance d, as illustrated in Figure 5.1. In general, we can model the measurement recorded at a sensor pixel (s_u, s_v) as a linear function of the scene intensity as

$$y(s_u, s_v) = \int I(x, y, z)\, \varphi(s_u, s_v; x, y, z)\, dxdydz, \tag{5.1}$$

where $I(x, y, z)$ denotes the image intensity at 3D location (x, y, z) and $\varphi(s_u, s_v; x, y, z)$ denotes the point spread function (PSF) or the sensor response recorded at (s_u, s_v) in the sensor plane for a point source at (x, y, z).

The general system in (5.1) can be simplified depending on the system design and placement and pattern of the mask. In our proposed method, we use a separable model proposed in [6], where we use a rank-1 matrix as the amplitude mask. With the separable mask placed parallel to the image sensor, the PSF of an arbitrary point, $\varphi(s_u, s_v; x, y, z)$, will be a rank-1 matrix, and the general model in (5.1) can be written in a simpler form as a separable system.

Suppose we discretize the continuous scene $I(x, y, z)$ into D depth planes $\mathbf{I_1}, \ldots, \mathbf{I_D}$, each with $N \times N$ pixels. The separable system can be represented in the following compact form:

$$\mathbf{Y} = \sum_k \mathbf{\Phi_k} \mathbf{I_k} \mathbf{\Phi_k^T}. \tag{5.2}$$

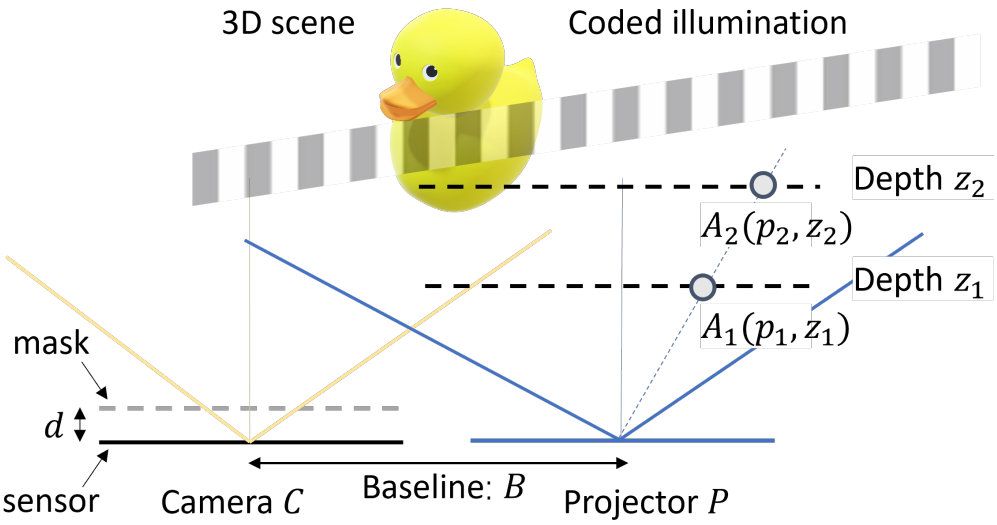


Figure 5.1: Illustration of a lensless camera with coded illumination. Camera and projector are separated by baseline distance B. The 3D scene is illuminated by a sequence of coded illumination patterns from the projector, and observed by the camera sensor beneath the coded mask. Rays that receive same illumination in projector coordinates appear at different angles in camera coordinates that provides different depth-dependent PSFs.

$\mathbf{Y}$ represents $M \times M$ sensor measurements and Φ_k represents the system matrix for the k-th depth plane.

5.3.2 Coded Illumination

We use a projector separated by baseline distance B from the camera to illuminate the scene with a sequence of coded illumination patterns (as illustrated in Figure 5.1. The effect of coded illumination can be modelled as an element-wise product between the illumination pattern and the scene. We divide the field-of-view (FOV) cone of the projector into $N \times N$ angles, which also determines the spatial discretization of the scene. We generate a sequence of illumination

patterns and capture the corresponding measurements on the sensor. The measurements captured for i-th pattern $\mathbf{P_i}$ can be represented as

$$\mathbf{Y}_i = \sum_k \mathbf{\Phi_k}(\mathbf{P_i} \odot \mathbf{I_k})\mathbf{\Phi_k^T}. \tag{5.3}$$

Note that we assume the same illumination pattern for every depth plane at a time. This is because we use the projector to determine the scene discretization at every depth plane.

To recover the 3D scene as a stack of D planes, $\mathbf{I} = \{\mathbf{I_1}, \ldots, \mathbf{I_D}\}$, we solve the following regularized least-squares problem:

$$\min_{\mathbf{I_1},\ldots,\mathbf{I_D}} \sum_i \|\mathbf{Y_i} - \sum_k \mathbf{\Phi_k}(\mathbf{P_i} \odot \mathbf{I_k})\mathbf{\Phi_k^T}\|_2^2 + \lambda\|\mathbf{D}(\mathbf{I})\|_2. \tag{5.4}$$

$\mathbf{D}(\mathbf{I})$ represents a finite difference operator that computes local gradients of the 3D volume $\mathbf{I}$ along spatial and depth directions. The ℓ_2 norm of the local differences provides the 3D total variation function that we use as the regularization function. The total variation function constrains the magnitude of the local variation in the reconstruction and is widely used in ill-conditioned image recovery problems [58, 87]. The optimization problem in (5.4) can be solved using an iterative least-squares solvers; we used the TVreg package [51].

5.3.3 Effect of Baseline on Depth-Dependent PSFs

As discussed in previous work on 3D lensless imaging [1,3,48,122], the points at different depth in the scene provide a scaled version of the mask pattern as the sensor response. However, if the object is far from the lensless camera, the depth-dependent differences in the sensor reponse become almost negligible.

Coded illumination in our proposed system provides robust 3D reconstruction for two main reasons: (1) Coded illumination selects a subset of scene points that contribute to each sensor measurement. (2) Spatial separation between camera and projector (i.e., baseline) maps depth variations in scene points into depth-dependent shifts in the sensor response. Since shifted versions of the the mask pattern can be easily resolved compare to the scaled versions, the baseline plays a critical role in quality of 3D reconstruction.

Let us consider the 1D case of our proposed framework, the projector P is placed at a baseline distance B away from the camera C, as shown in Figure 5.1. For an arbitrary point at (p, z) in the coordinate system of C, its measurements on camera C can be written as

$$\phi(s; p, z) = \text{mask}\left[(1 - \frac{d}{z})s + d\frac{p}{z}\right], \tag{5.5}$$

where s denotes the coordinates on the camera sensor and d denotes the sensor-to-mask distance. The coordinates of two arbitrary points $A_1(p_1, z_1)$ and $A_2(p_2, z_2)$ on the same ray of the projector in the coordinate system of P become $(p_1 + B, z_1)$, $(p_2 + B, z_2)$ in the coordinate system of camera C (because of the the baseline between camera and projector). We can represent the camera response or PSF corresponding to each of these points as

$$\begin{aligned}\phi(s; p_1, z_1) &= \text{mask}\left[(1 - \frac{d}{z_1})s + \frac{dp_1}{z_1} - \frac{dB}{z_1}\right] \\ &= \text{mask}\left[(1 - \frac{d}{z_1})(s - \frac{dB}{z_1 - d}) + d\frac{p_1}{z_1}\right] \end{aligned} \tag{5.6}$$

$$\begin{aligned}\phi(s; p_2, z_2) &= \text{mask}\left[(1 - \frac{d}{z_2})s + \frac{dp_2}{z_2} - \frac{dB}{z_2}\right] \\ &= \text{mask}\left[(1 - \frac{d}{z_2})(s - \frac{dB}{z_2 - d}) + d\frac{p_2}{z_2}\right]. \end{aligned} \tag{5.7}$$

Note that $\frac{p_1}{z_1} = \frac{p_2}{z_2}$ because the two points are at different depths on the same ray angle. Therefore, the PSF of points A_1 and A_2 differ from each other with a scaling factor $1 - \frac{d}{z}$ and a depth-dependent shift $\frac{dB}{z-d}$.

Specifically, when the object is far from the camera, we can often ignore the difference in depth scaling factor $1 - \frac{d}{z}$, and the difference of the depth-dependent shift becomes $|\frac{dB}{z_1 - d} - \frac{dB}{z_2 - d}|$. When the baseline is zero, two point light sources on the same ray are the scaled versions of each other, and the scaling factor becomes almost the same when the object distance (z) is large.

However, by separating the camera and project by baseline distance B, the camera observes a shifted 3D grid; two points on the shifted grid provide an angular difference with respect to the camera. Therefore, the depth resolvability of the system improves. This effect was previously discussed in [5] for more general geometries with multiple cameras.

In general, the projector P and camera C can be separated laterally and axially. Lateral separation provides depth-dependent shifts of the PSF, which we discussed in (5.6) and (5.7). Axial separation would provide depth-dependent scaling and shifts of the PSF, which can also be deduced from (5.6) and (5.7). For instance, if the camera and projector are separated axially by Δz, the depth-dependent shifts can be calculated by replacing z_1, z_2 with $z_1 + \Delta z, z_2 + \Delta z$, respectively. Since these terms appear in the denominator, their influence on the PSF shift will be small compared to lateral baseline B.

5.4 Simulation Results

To validate the performance of the proposed algorithm, we simulate a lensless imaging system where a coded-mask is placed on top of an image sensor. We use a separable maximum length sequence (MLS) mask pattern [6, 19]. The size of each mask feature is 60μm, and the sensor-mask distance is 2mm. The sensor pitch in the simulation is 4.8μm and the total number of pixels on the sensors is 512×512. We simulate a multi-plane 3D scene with $128 \times 128 \times 10$

voxels. The simulated sensor noise consists of photon noise and read noise, and the noisy sensor measurements can be described as

$$\mathbf{Y}_n = \frac{G}{F}(\text{Poisson}(\frac{F}{G}\mathbf{Y}) + N(0, \sigma^2)), \tag{5.8}$$

where $\mathbf{Y}$ and $\mathbf{Y}_n$ refers to original and noisy measurements, where F stands for the full-well capacity of the sensor, and G represents the gain value. The variance $\sigma = F \times 10^{-R/20}$ and R is the dynamic range.

5.4.1 Effect of Illumination Patterns

We test different types of binary illumination patterns for the simulation. The patterns are designed to be binary to keep the model simple and to avoid the effect of non-linearity caused by the Gamma curve of the projector..

Uniform. One pattern that illuminates all the pixels simultaneously;

Random. A sequence of separable binary random matrices. We ensure that the union of all the patterns should illuminate all the pixels (i.e., if we add up all the illumination patterns, they should not have zero entries anywhere).

Shifting Dots Array. The base illumination pattern consists of dots separated by k pixels along the horizontal and vertical directions. We then generate a total of k^2 illumination patterns, each of which is a shifted version of the base pattern. The summation of all the patterns will give us a uniform illumination pattern.

Shifting Lines. Similar to shifting dots array, the base illumination patterns consist of horizontal lines separated by k pixels along vertical axis and vertical lines separated along horizontal axis. We then generate shifted version of these two base patterns. The summation of all the patterns is a uniform illumination pattern.

We present simulation results with different number and types of illumination patterns in Figure 5.2. The simulated test scene is taken from NYU depth dataset [74]. The depth of scene is rescaled into the range from 40 to 60 and discretized into 50 depth planes to simulate the sensor measurements. The camera setup and the baseline between camera and projector are fixed during the simulation. The shifting lines and shifting dots outperform the uniform pattern in terms of depth RMSE. Also, the depth RMSE drops as we increase the number of illumination patterns.

5.4.2 Effects of Baseline

The baseline between the lensless camera and the projector affects the depth resolvability of the system. Shifting the lensless camera by a distance, the camera observes the scene from a side view and transfers the depth difference of two points into angular difference. We present simulation results in Figure 5.3 to demonstrate the effect of camera-projector baselines. The number of illumination patterns for all the simulation are the same. We then fix the baseline along axial direction to 0cm and the baseline along lateral direction as $B = \{0, 2.5, 5, 7.5\}$. As shown in Figure 5.3, the depth RMSE is decreased as we increase the baseline between the camera and the projector. When the baseline is zero, which means the camera and projector are overlapped, we barely distinguish any depth.

One important consideration for our method is that the target object should lie within the intersection of the sensor FOV cone and the projector illumination cone. As we increase the baseline, the intersection of the two cones is pushed farther from the sensor. Therefore, we should determine the maximum baseline based on the object distance, sensor FOV, and projector cone. If we increase the baseline beyond the maximum limit, then the reconstruction quality can decrease.

5.4.3 Comparison With an Ideal Pinhole Camera

In existing structured illumination methods [41, 42, 46], a lens-based camera is used to capture the scene from the side view of the projector and depth map can be accurately reconstructed by triangulation. We can model the lens-based camera as an ideal pinhole camera (ignoring photon or sensor noise) for the sake of comparison with our method. We present simulation results comparing a pinhole-based camera with structured illumination in Figure 5.4. The baseline between the projector and the camera is fixed at 5cm in all the simulations. Results in Figure 5.4 show that the pinhole mask (that represents an ideal lens-based camera) provides better results compared to the MLS mask. Compared to mask-based lensless camera where the sensor measurements are multiplexed, a lens-based system can offer better conditioning and depth reconstruction. Nevertheless, a lens-based camera imposes additional burden in terms of device thickness, weight, and geometry.

5.4.4 Comparison With Multishot Lensless System

In our proposed method, multiple frames of measurements are captured, which introduce additional limitations such as long capture time and low frame rate. In Figure 5.5, we present simulation results comparing our method with another multi-shot lensless imaging system called SweepCam [48]. SweepCam captures multiple frames of sensor measurements while translating the mask laterally. The translation of the mask offers a perspective shift in the measurements that depends on the depth of objects in the scene. In our simulations, the SweepCam mask is translated to 48 positions within a range of 2.88mm $\times$ 2.88mm. However, since the translating distance of the mask is limited by the sensor area, the SweepCam method fails to resolve the depths when the scene is farther than 30cm.

5.5 Experimental Results

To validate our proposed method, we built a prototype with a lensless camera and a Sony MP-CL1 laser projector, shown in Figure 5.6. The lensless camera prototype consists of an image sensor and a coded amplitude mask on top of it. We employ the outer-product of two MLS vectors as our mask pattern. The mask has 511×511 square features. The pixel pitch is 60μm and the sensor-to-mask distance is 2mm. We use a Sony IMX183 sensor and bin 2×2 sensor pixels, which yields the effective sensor pitch close to 4.8μm. We record 512×512 measurements from the sensor and the effective sensor size is 2.46mm $\times$ 2.46mm. We place the test 3D objects within 40cm and 60cm depth range with respect to the camera. Finally, we reconstruct $128 \times 128 \times 10$ voxels in

the illuminated area. In our method, the lensless camera and the projector are separated by a 55mm baseline. We first reconstruct the depth planes by solving the regularized least-squares problem in (5.4). Then we create an all-in-focus image and depth map by selecting the pixel with the maximum amplitude along each light ray.

In our experiments, the pixel grid of the scene, illumination patterns P_i , the system matrices at each depth $\mathbf{\Phi_k}$ must be correctly aligned; otherwise, we will get artifacts in the reconstruction. To avoid any grid mismatch, we use the same projector to calibrate the system matrices and generate the illumination patterns in our experiments.

5.5.1 Effect of Illumination Patterns

We present experimental results of 3D reconstruction with our proposed method for real objects in Figures 5.7 and 5.9. We show the results of reconstructed depth planes, estimated all-in-focus images and depth maps using uniform, shifting lines, and shifting dots patterns. For comparison, we captured the original image and depth map for each scene using Intel RealSense D415 depth camera, where the baseline between the lens-based camera and the projector is 55mm.

The results in Figure 5.7 compare 3D reconstruction with uniform and 48 shifting lines. In the first two rows, the scene is a slanted box, containing continuous depth varying from 40cm to 60cm. In the last two rows, the scene contains a red toy located at 40cm and a green toy lying from 50cm to 60cm. The results in the first three columns in Figure 5.7 represent three depth planes at 58cm, 62cm, and 66cm. The results show that the correct depth can be easily distinguished in

images reconstructed with 48 shifting lines pattern, whereas depth planes reconstructed with the uniform illumination pattern show incorrect depth and intensity. The estimated all-in-focus image and depth maps for 48 shifting lines also appear significantly better than those from the uniform illumination patterns.

The results in Figure 5.8 compare different number and types of illumination patterns. We observe that the uniform illumination pattern barely recover any depth. The illumination pattern with 16 and 49 shifting dots provide better results than uniform illumination. 16 shifting lines provide slightly better results compared to shifting dots and 48 shifting lines provide significanlty better image and depth map.

In summary, the ill-conditioned system matrices with uniform illumination pattern cause various artifacts in 3D reconstruction. Capturing measurements from coded illumination improves the conditioning of the overall system and the reconstructed images have better spatial and depth resolution. Increasing the number of illumination patterns provides better reconstruction. More illumination patterns would require longer acquisition time as well, which enforces a trade off between the quality of reconstruction and data acquisition time.

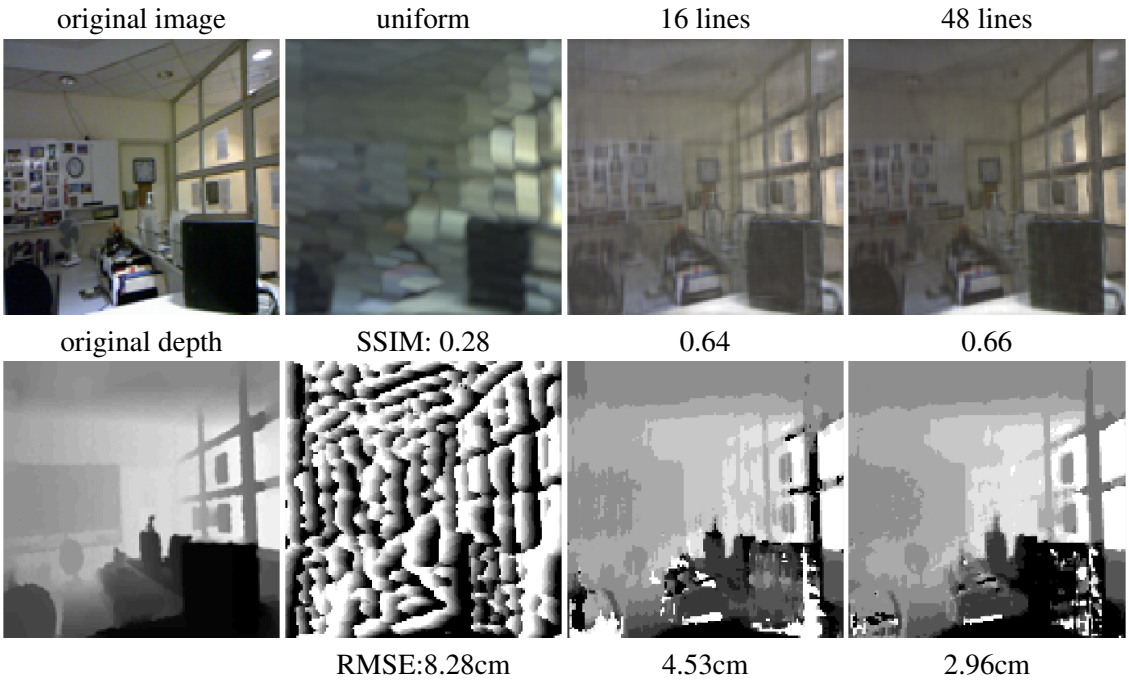
5.5.2 Effect of Baselines

We show experimental results for different baselines in Figure 5.9. We captured the same scene with 5.5 and 10.5 baseline and performed 3D reconstruction with the respective measurements. The results in Figure 5.9 show that 10.5cm baseline offers finer depth resolution (indicated as narrow depth of field) compared to the reconstruction with 5.5cm baseline. The improvement is small, and this effect was observed in the simulation results in Figure 5.3b that show the depth RMSE of the system tapers off as we increase the baseline between the camera and the projector.

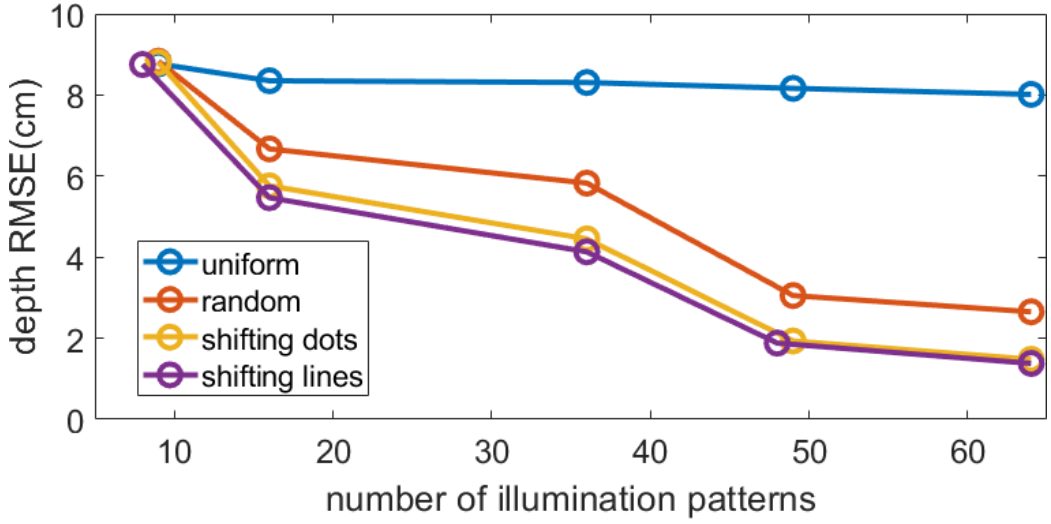
5.6 Conclusion and Discussion

We propose a framework for combining coded illumination with lensless imaging for 3D lensless imaging. We present simulation and real experiment results to demonstrate that our proposed method can achieve significantly improved 3D reconstruction with multiple coded illumination compared to uniform illumination. Such a mask-based lensless camera can be useful in space-limited applications such as under-the-display or large-area sensing, where installing a lens-based camera can be challenging. Our proposed setup can also be useful for distributed lensless sensors (in different shapes and geometries), where we may want to image over a large area, large field-of-view, but keep the devices flat, thin, and lens-free.

Limitaitons. Our current setup can add extra cost and complexity to the system design because of the illumination source. The need to capture multiple shots can also increase the data acquisition time and restrict the usage for static or slow-moving objects.

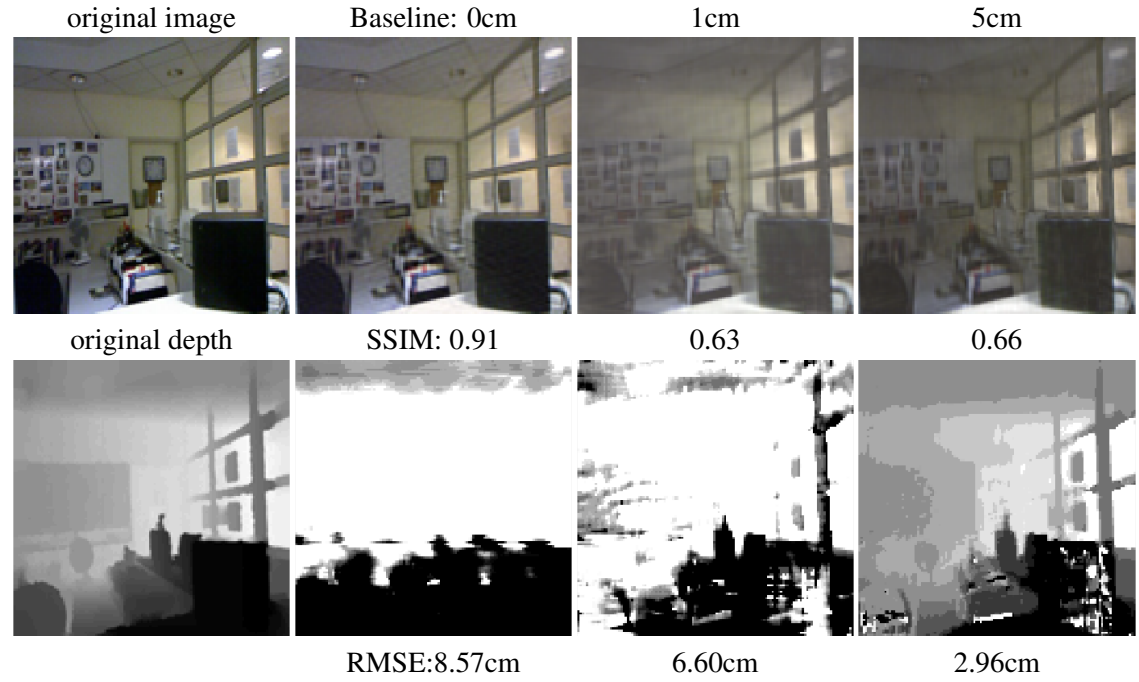


(a) Reconstruction of synthetic 3D test scene for different numbers of illumination patterns with the same baseline. Top row represents the estimated all-in-focus images. Bottom row represents the estimated depth maps.

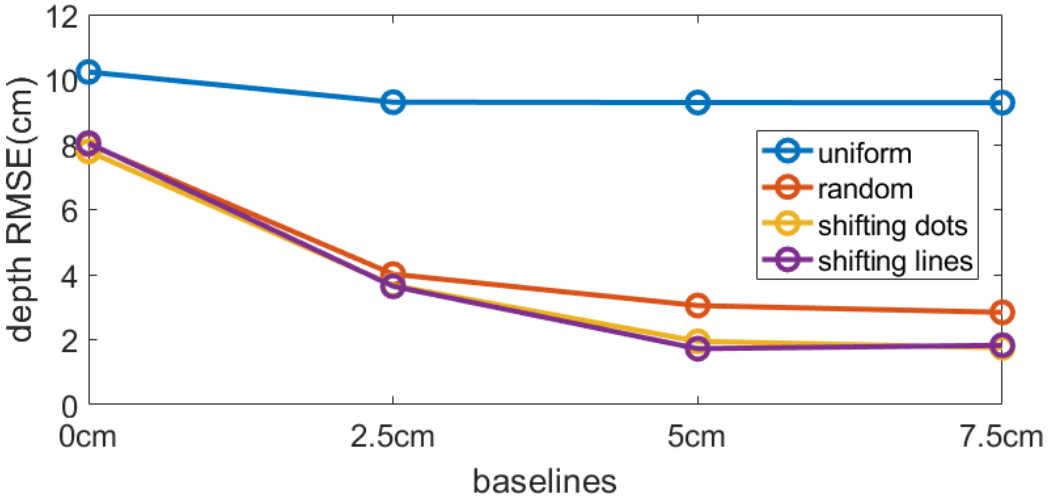


(b) Averaged depth RMSE of all test scenes.

Figure 5.2: Reconstruction and averaged depth RMSE for different number and types of illumination patterns. The baseline is fixed at 5cm during simulation. We observe that performance improves as we increase the number of illumination patterns.

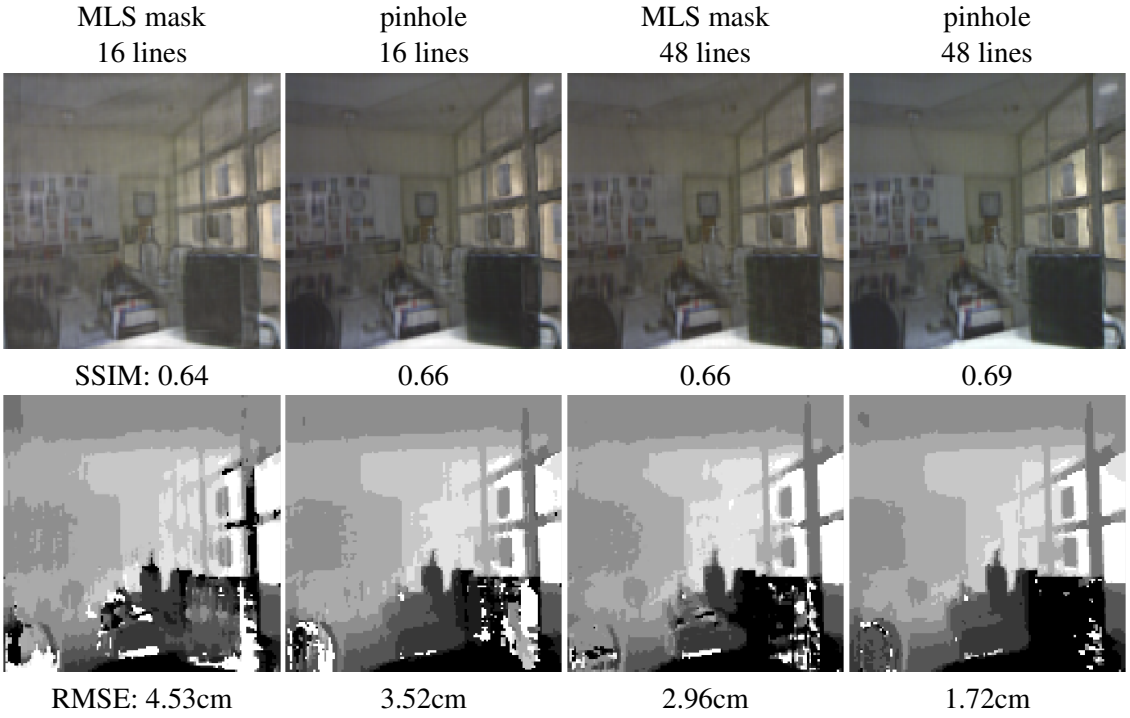


(a) Reconstruction of synthetic 3D test scene for different baselines with 48 shifting lines patterns. Top row represents the estimated all-in-focus images. Bottom row represents the estimated depth maps.

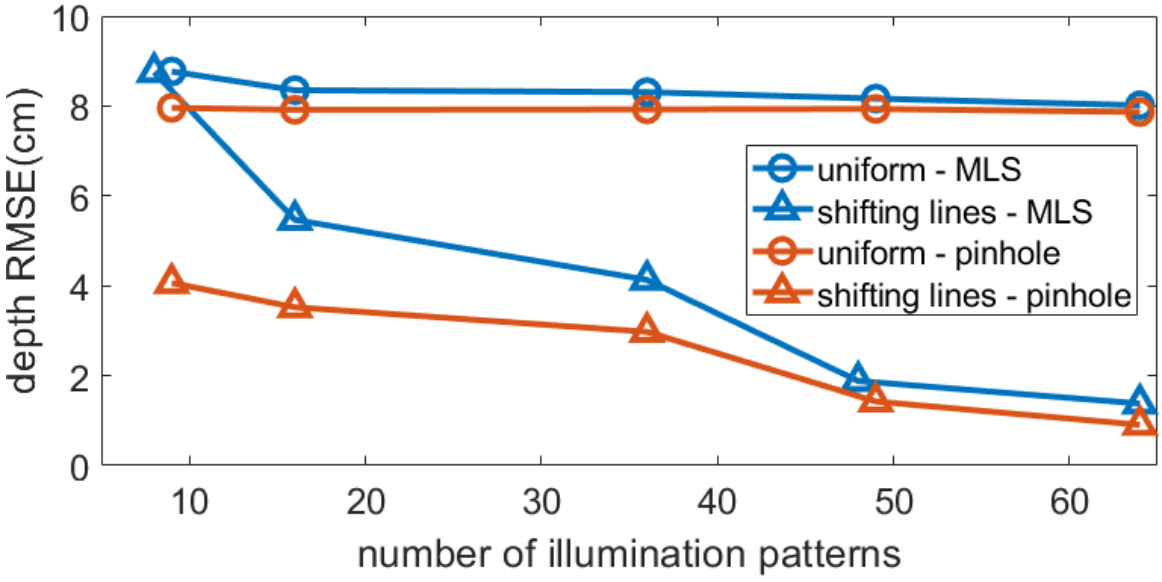


(b) Averaged depth RMSE of all test scenes.

Figure 5.3: Reconstruction and averaged depth RMSE for different values of baseline distance. The number of illumination patterns is fixed for all tests. We observe that larger baselines provide better 3D reconstruction.



(a) Reconstruction of synthetic 3D test scene. Top row represents the estimated all-in-focus images. Bottom row represents the estimated depth maps.



(b) Averaged depth RMSE of all test scenes.

Figure 5.4: Comparison of the ideal pinhole-based and MLS mask-based camera models with coded illumination patterns. The pinhole-based model performs better due to its better system conditioning.

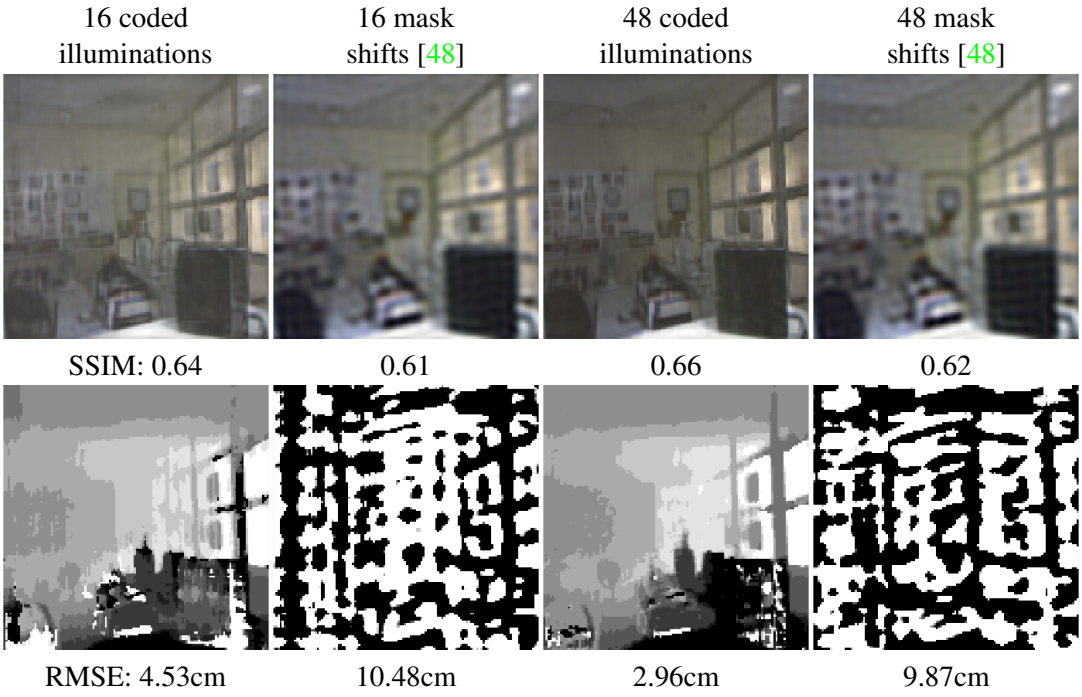


Figure 5.5: Comparison of the proposed coded illumination-based reconstruction with shifting mask-based reconstruction in SweepCam [48]. The SweepCam fails to resolve objects that are far from the camera.

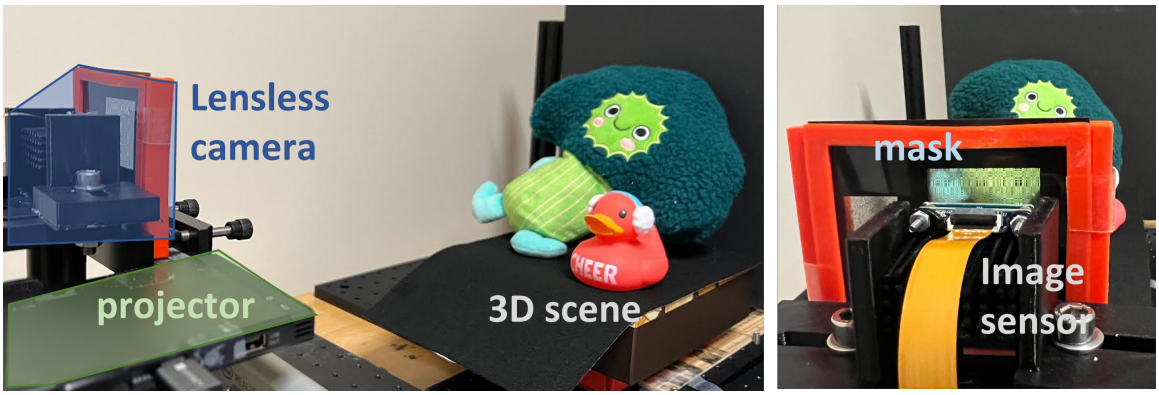


Figure 5.6: Camera and projector setup used in our experiments. The projector is placed next to the camera. The scene objects are placed ranging from 40cm to 60cm. We capture multiple frames of sensor measurements under a sequence of coded illumination patterns from the projector to improve the 3D image reconstruction quality.

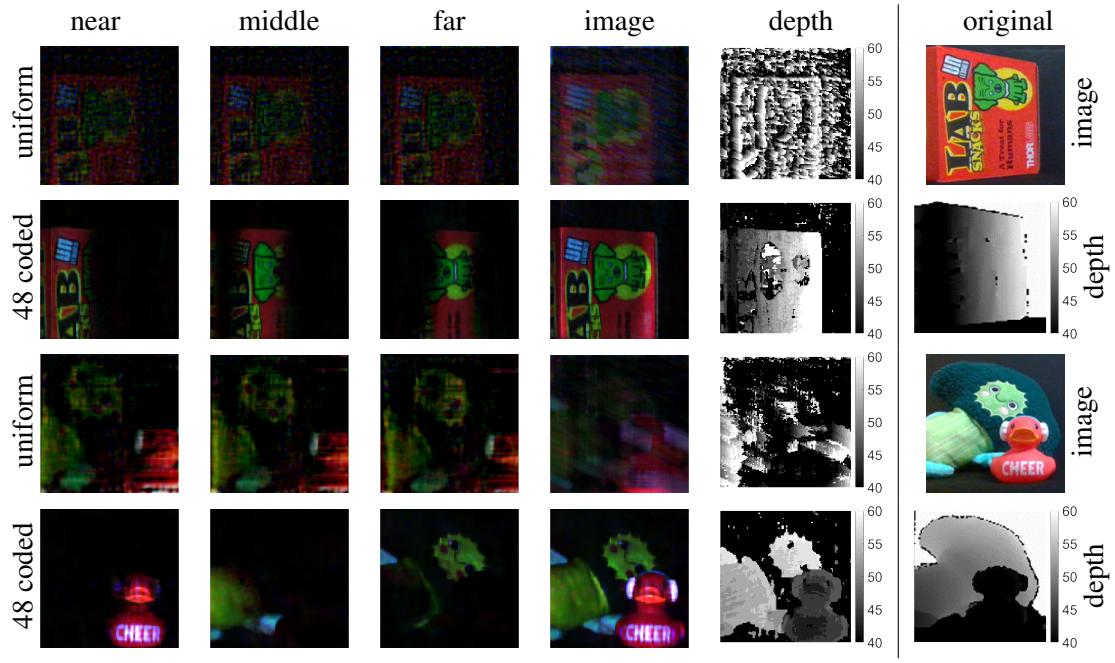


Figure 5.7: Reconstructed images at three selected depth planes, all-in-focus images, and depth maps using uniform and 48 coded shifting lines patterns. The depth maps of the real scenes and the estimated depth maps are all plotted in grayscale to show range from 40cm to 60cm. We observe that uniform illumination-based system fails to recover correct depth planes whereas the coded illumination-based system can recover depth planes and entire 3D image with high quality.

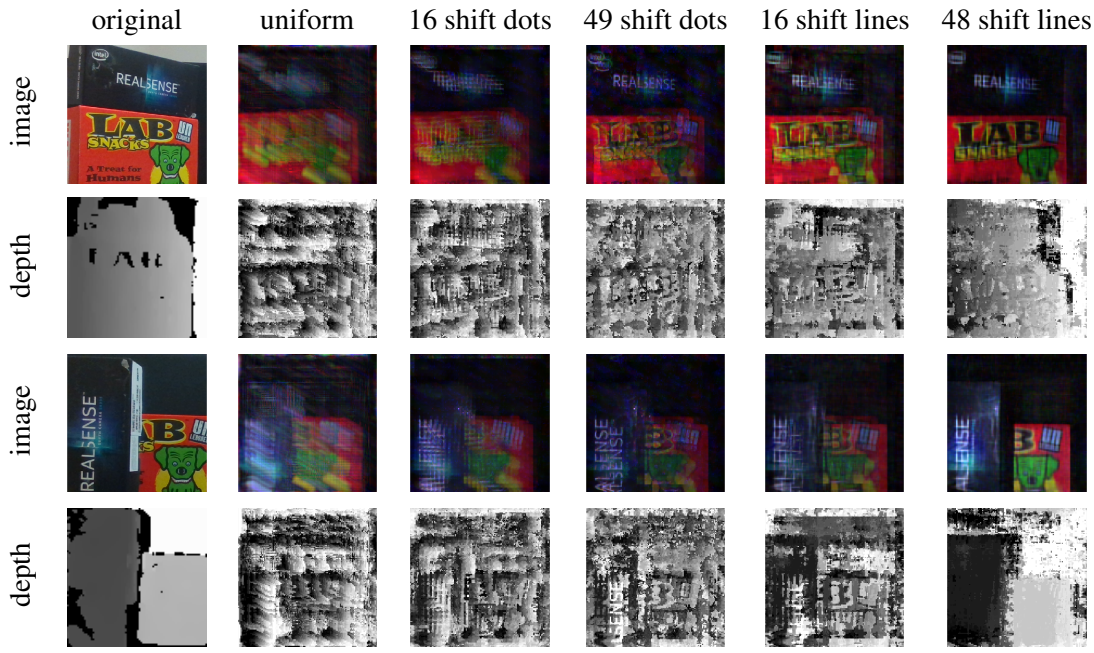


Figure 5.8: Reconstruction results with original, uniform, 16 and 49 shifting dots patterns, 16 and 48 shifting lines patterns. We show estimated depth maps and all-in-focus images by selecting the pixel with the maximum light magnitude along each ray. The depth maps of the real scenes and the estimated depth maps are all plotted in grayscale to show range from 40cm to 60cm. We observe that 48 shifting lines provides high-quality spatial and depth resolution.

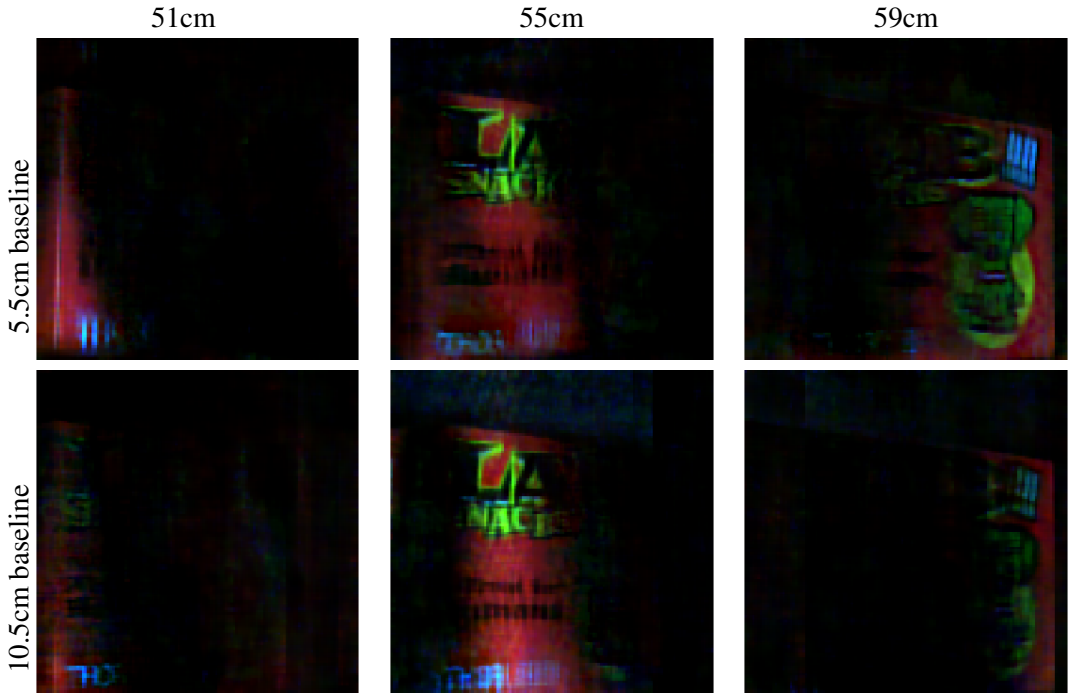


Figure 5.9: Reconstructed depth planes with 5.5cm and 10.5cm baselines within the adjustable range of hardware. The 10.5cm baseline results have better depth resolvability. The patterns are fixed with 48 shifting lines. We observe that larger baseline offers better depth resolvability.

Chapter 6

Conclusions

As conclusion, we present the summary of the contribution of each chapter and the discussion of potential future directions. Lensless cameras of a flat optical encoder(coded mask) and its camera sensor underneath. The imaging process of lensless cameras is formulated as linear equations and the underlying images or 3D scenes are solved with linear algebra techniques. In this dissertation, we explore the joint design of sensing approaches, such as coded mask and external light source, along with reconstruction algorithm. With sophisticated designed sensing approaches and reconstruction algorithm, the quality of the reconstruction can be improved significantly, the forward imaging model and computation complexity can also be simplified.

In Chapter 2, we presented a new algorithm to jointly estimate the image and depth of a scene using a single snapshot of a mask-based lensless camera. Existing methods for 3D lensless imaging either estimate scene over a predefined 3D grid (which is computationally expensive) or a small number of candidate depth planes (which provides a coarse depth map). We divide the scene into an intensity map at uniform angles and a depth map on a continuous domain, which

allows us to estimate a variety of scenes with different depth ranges using the same formulation. We jointly estimate the image intensity and depth map by solving a nonconvex problem. We initialize our estimates using a greedy method and add weighted regularization to enforce smoothness in the depth estimate while preserving the sharp edges. We demonstrated with extensive simulations and experiments with real data that our proposed method can recover image and depth with high accuracy for a variety of scenes. We evaluated the performance of our methods under different noise levels, sensor sizes, and numbers of sensor pixels and found the method to be robust. We presented a comparison with existing methods for lensless 3D imaging and demonstrated both in simulation and real experiments that our method provides significantly better results.

In Chapter 3, we present a new framework to recover 3D scenes using a lensless camera with a programmable mask. Our proposed method can recover multiple depth planes in the 3D scene using a computationally efficient algorithm that solves multiple small linear systems in parallel in the frequency domain. To further improve the quality of 3D scene recovery, we optimized the mask patterns and trained a U-Net that converts estimated image planes to all-in-focus image and continuous-valued depth map. Our experimental results demonstrate that the proposed method can reliably recover dense 3D scenes with a small number of sensor measurements and outperform existing methods. The reconstruction quality of our proposed method, like other lensless imaging systems, drops for scenes with specular reflections and large occlusions.

In Chapter 4, we propose a framework for combining coded illumination with lensless imaging. We present extensive simulation and real experiment results to demonstrate that we can get significantly improved reconstruction with multiple coded illumination patterns compared to original uniform illumination. In space-limited applications such as under-the-display sensing, where

the sensor-to-mask distance has to be small, our proposed method can offer significantly better reconstruction compared to uniform illumination. We also explore that any orthogonal illumination patterns(such as simple shifting dots patterns) can be practical in improving the system conditioning and imaging quality.

In Chapter 5, we propose a framework for combining coded illumination with lensless imaging for 3D lensless imaging. We present simulation and real experiment results to demonstrate that our proposed method can achieve significantly improved 3D reconstruction with multiple coded illumination compared to uniform illumination. Such a mask-based lensless camera can be useful in space-limited applications such as under-the-display or large-area sensing, where installing a lens-based camera can be challenging. Our proposed setup can also be useful for distributed lensless sensors (in different shapes and geometries), where we may want to image over a large area, large field-of-view, but keep the devices flat, thin, and lens-free.

As for future directions, extending our method to dynamic scenes is a natural direction for future work. We also need to further explore if some other illumination patters can offer better 3D reconstruction for scenes with different depth profiles. Co-design of illumination patterns, mask pattern/placement, and overall system arrangement can further improve the quality of 3D reconstruction. On the algorithmic side, the recovery algorithm can be improved by including more sophisticated priors for the 3D scenes.

Data-driven methods such as deep learning techniques has been demonstrated effective in many aspects of image processing and computer vision. However, due to its ill interpretability, the outputs of trained networks for lensless imaging may vary drastically depending on the lensless cameras and light conditions. One interesting future direction is explore whether we can fine-tune only part of the network parameters to make it work with different lensless camera systems and light conditions.

Bibliography

[1] Jesse K. Adams, Vivek Boominathan, Benjamin W. Avants, Daniel G. Vercosa, Fan Ye, Richard G. Baraniuk, Jacob T. Robinson, and Ashok Veeraraghavan. Single-frame 3d fluorescence microscopy with ultraminiature lensless flatscope. *Science Advances*, 3(12), 2017.

[2] Manya V. Afonso, José M. Bioucas-Dias, and Mário A. T. Figueiredo. Fast image recovery using variable splitting and constrained optimization. *IEEE Transactions on Image Processing*, 19(9):2345–2356, 2010.

[3] Nick Antipa, Grace Kuo, Reinhard Heckel, Ben Mildenhall, Emrah Bostan, Ren Ng, and Laura Waller. Diffusercam: lensless single-exposure 3d imaging. *Optica*, 5(1):1–9, Jan 2018.

[4] Nick Antipa, Patrick Oare, Emrah Bostan, Ren Ng, and Laura Waller. Video from stills: Lensless imaging with rolling shutter. In *IEEE International Conference on Computational Photography (ICCP)*, pages 1–8, Los Alamitos, CA, USA, may 2019. IEEE Computer Society.

[5] M Salman Asif. Lensless 3d imaging using mask-based cameras. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6498–6502. IEEE, April 2018.

[6] M. Salman Asif, Ali Ayremlou, Aswin Sankaranarayanan, Ashok Veeraraghavan, and Richard G. Baraniuk. Flatcam: Thin, lensless cameras using coded aperture and computation. *IEEE Transactions on Computational Imaging*, 3(3):384–397, 2017.

[7] R. G. Baraniuk, V. Cevher, M. F. Duarte, and C. Hegde. Model-based compressive sensing. *IEEE Transactions on Information Theory*, 56(4):1982–2001, April 2010.

[8] Richard G. Baraniuk. Compressive sensing [lecture notes]. *IEEE Signal Processing Magazine*, 24(4):118–121, 2007.

[9] Richard G Baraniuk, Volkan Cevher, Marco F Duarte, and Chinmay Hegde. Model-based compressive sensing. *IEEE Transactions on Information Theory*, 56(4):1982–2001, 2010.

[10] H. H. Barrett, D. T. Wilson, G. D. DeMeester, and H. Scharfman. Fresnel zone plate imaging in radiology and nuclear medicine. *Optical Engineering*, 12(1):8 – 12 – 5, 1973.

[11] JONATHAN BARZILAI and JONATHAN M. BORWEIN. Two-point step size gradient methods. *IMA Journal of Numerical Analysis*, 8(1):141–148, 1988.

[12] Vivek Boominathan, Jesse K Adams, M Salman Asif, Benjamin W Avants, Jacob T Robinson, Richard G Baraniuk, Aswin C Sankaranarayanan, and Ashok Veeraraghavan. Lensless imaging: A computational renaissance. *IEEE Signal Processing Magazine*, 33(5):23–35, 2016.

[13] Vivek Boominathan, Jesse K. Adams, Jacob T. Robinson, and Ashok Veeraraghavan. Phlatcam: Designed phase-mask based thin lensless camera. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(7):1618–1629, 2020.

[14] Vivek Boominathan, Jacob T Robinson, Laura Waller, and Ashok Veeraraghavan. Recent advances in lensless imaging. *Optica*, 9(1):1–16, 2022.

[15] Ashish Bora, Ajil Jalal, Eric Price, and Alexandros G Dimakis. Compressed sensing using generative models. In *International Conference on Machine Learning*, pages 537–546. PMLR, 2017.

[16] Ashish Bora, Ajil Jalal, Eric Price, and Alexandros G Dimakis. Compressed sensing using generative models. In *International Conference on Machine Learning*, pages 537–546. PMLR, 2017.

[17] N. Boyd, G. Schiebinger, and B. Recht. The alternating descent conditional gradient method for sparse inverse problems. In *IEEE 6th International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP)*, pages 57–60, Dec 2015.

[18] C. Broyden. The convergence of a class of double-rank minimization algorithms 1. general considerations. *IMA Journal of Applied Mathematics*, 6(1):76–90, 1970.

[19] Axel Busboom, Harald Elders-Boll, and Hans D. Schotten. Uniformly redundant arrays. *Experimental Astronomy*, 8(2):97–123, Jun 1998.

[20] Emmanuel J. Candès, Xiaodong Li, and Mahdi Soltanolkotabi. Phase retrieval from coded diffraction patterns. *Applied and Computational Harmonic Analysis*, 39(2):277–299, 2015.

[21] Emmanuel J. Candès, Justin K. Romberg, and Terence Tao. Stable signal recovery from incomplete and inaccurate measurements. *Communications on Pure and Applied Mathematics*, 59(8):1207–1223, 2006.

[22] Emmanuel J. Candès, Thomas Strohmer, and Vladislav Voroninski. Phaselift: Exact and stable signal recovery from magnitude measurements via convex programming. *Communications on Pure and Applied Mathematics*, 66(8):1241–1274, 2013.

[23] Thomas M. Cannon and Edward E. Fenimore. Coded Aperture Imaging: Many Holes Make Light Work. *Optical Engineering*, 19:283, June 1980.

[24] Nadya Chakrova, Rainer Heintzmann, Bernd Rieger, and Sjoerd Stallinga. Studying different illumination patterns for resolution improvement in fluorescence microscopy. *Opt. Express*, 23(24):31367–31383, Nov 2015.

[25] Giorgi Chaladze and Levan kalatozishvili. Linnaeus 5 dataset, Nov 2017.

[26] Rohan Chandra, Tom Goldstein, and Christoph Studer. Phasepack: A phase retrieval library. In *2019 13th International conference on Sampling Theory and Applications (SampTA)*, pages 1–5, 2019.

[27] Venkat Chandrasekaran, Benjamin Recht, Pablo A. Parrilo, and Alan S. Willsky. The convex geometry of linear inverse problems. *Foundations of Computational Mathematics*, 12(6):805–849, Dec 2012.

[28] Julie Chang and Gordon Wetzstein. Deep optics for monocular depth estimation and 3d object detection. In *Proc. IEEE ICCV*, 2019.

[29] Kostadin Dabov, Alessandro Foi, Vladimir Katkovnik, and Karen Egiazarian. Image denoising by sparse 3-d transform-domain collaborative filtering. *IEEE Transactions on Image Processing*, 16(8):2080–2095, 2007.

[30] Akshat Dave, Anil Kumar Vadathya, Ramana Subramanyam, Rahul Baburajan, and Kaushik Mitra. Solving inverse computational imaging problems using deep pixel-level prior. *IEEE Trans. Computational Imaging*, 5(1):37–51, 2019.

[31] David L. Donoho. Compressed sensing. *IEEE Transactions on Information Theory*, 52(4):1289–1306, April 2006.

[32] M. F. Duarte, M. A. Davenport, D. Takhar, J. N. Laska, T. Sun, K. F. Kelly, and R. G. Baraniuk. Single-pixel imaging via compressive sampling. *IEEE Signal Processing Magazine*, 25(2):83–91, March 2008.

[33] Frédo Durand and Julie Dorsey. Fast bilateral filtering for the display of high-dynamic-range images. *ACM Trans. Graph.*, 21(3):257–266, July 2002.

[34] Edward E. Fenimore and Thomas M. Cannon. Coded aperture imaging with uniformly redundant arrays. *Appl. Opt.*, 17(3):337–347, Feb 1978.

[35] David Fofi, Tadeusz Sliwa, and Yvon Voisin. A comparative survey on invisible structured light. In *IST/SPIE Electronic Imaging*, 2004.

[36] Nikolas P. Galatsanos, Aggelos K. Katsaggelos, Roland T. Chin, and Allen D Hillery. Least squares restoration of multichannel images. *IEEE Transactions on Signal Processing*, 39(10):2222–2236, 1991.

[37] Tom Goldstein and Stanley Osher. The split bregman method for l1-regularized problems. *SIAM J. Img. Sci.*, 2(2):323–343, April 2009.

[38] Gene H. Golub and Charles F. Van Loan. *Matrix Computations (3rd Ed.)*. Johns Hopkins University Press, Baltimore, MD, USA, 1996.

[39] Ryan D. Gow, David Renshaw, Keith Findlater, Lindsay Grant, Stuart J. McLeod, John Hart, and Robert L. Nicol. A comprehensive tool for modeling cmos image-sensor-noise performance. *IEEE Transactions on Electron Devices*, 54(6):1321–1329, 2007.

[40] Jinwei Gu, Toshihiro Kobayashi, Mohit Gupta, and Shree K. Nayar. Multiplexed illumination for scene recovery in the presence of global illumination. In *2011 International Conference on Computer Vision*, pages 691–698, 2011.

[41] Mats G L Gustafsson. Surpassing the lateral resolution limit by a factor of two using structured illumination microscopy. *Journal of microscopy*, 198 Pt 2:82–7, 2000.

[42] Mats G L Gustafsson, Lin Shao, Peter M Carlton, C J Rachel Wang, Inna N Golubovskaya, W Zacheus Cande, David A Agard, and John W Sedat. Three-Dimensional Resolution Doubling in Wide-Field Fluorescence Microscopy by Structured Illumination. *Biophysical Journal*, 94:4957–4970, June 2008.

[43] Paul Hand, Oscar Leong, and Vlad Voroninski. Phase retrieval under a generative prior. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 9136–9146. Curran Associates, Inc., 2018.

[44] Richard Hartley and Andrew Zisserman. *Multiple view geometry in computer vision*. Cambridge university press, 2003.

[45] Felix Heide, Matthias B Hullin, James Gregson, and Wolfgang Heidrich. Low-budget transient imaging using photonic mixer devices. *ACM Transactions on Graphics (ToG)*, 32(4):45, 2013.

[46] Rainer Heintzmann and Christoph G. Cremer. Laterally modulated excitation microscopy: improvement of resolution by using a diffraction grating. In Irving J. Bigio, Herbert Schneckenburger, Jan Slavik, Katarina Svanberg M.D., and Pierre M. Viallet, editors, *Optical Biopsies and Microscopic Techniques III*, volume 3568, pages 185 – 196. International Society for Optics and Photonics, SPIE, 1999.

[47] M. Hirsch, S. Sivaramakrishnan, S. Jayasuriya, A. Wang, A. Molnar, R. Raskar, and G. Wetzstein. A switchable light field camera architecture with angle sensitive pixels and dictionary-based sparse coding. In *2014 IEEE International Conference on Computational Photography (ICCP)*, pages 1–10, 2014.

[48] Yi Hua, Shigeki Nakamura, M. Salman Asif, and Aswin C. Sankaranarayanan. Sweepcam — depth-aware lensless imaging using programmable masks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(7):1606–1617, 2020.

[49] Rakib Hyder, Viraj Shah, Chinmay Hegde, and M. Salman Asif. Alternating phase projected gradient descent with generative priors for solving compressive phase retrieval. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7705–7709, 2019.

[50] Gauri Jagatap, Zhengyu Chen, Seyedehsara Nayer, Chinmay Hegde, and Namrata Vaswani. Sample efficient fourier ptychography for structured data. *IEEE Transactions on Computational Imaging*, 6:344–357, 2020.

[51] Tobias Lindstrøm Jensen, Jakob Heide Jørgensen, Per Christian Hansen, and Søren Holdt Jensen. Implementation of an optimal first-order method for strongly convex total variation regularization. In *BIT Numerical Mathematics*, volume 52, pages 329–356, 2012.

[52] S. Khan, V. Sundar, V. Boominathan, A. Veeraraghavan, and K. Mitra. Flatnet: Towards photorealistic scene reconstruction from lensless measurements. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, oct 2020.

[53] Salman Khan, Adarsh R, Vivek Boominathan, Jasper Tan, Ashok Veeraraghavan, and Kaushik Mitra. Towards photorealistic reconstruction of highly multiplexed lensless images. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7859–7868, 2019.

[54] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, 2015.

[55] LaserComponents. Pattern Generators for FLEXPOINT Laser Modules.

[56] Anat Levin, Rob Fergus, Frédo Durand, and William T. Freeman. Image and depth from a conventional camera with a coded aperture. *ACM Trans. Graph.*, 26(3), July 2007.

[57] Chengbo Li. An efficient algorithm for total variation regularization with applications to the single pixel camera and compressive sensing. 2010.

[58] Chengbo Li, Wotao Yin, Hong Jiang, and Yin Zhang. An efficient augmented lagrangian method with applications to total variation minimization. *Computational Optimization and Applications*, 56(3):507–530, 2013.

[59] Chengbo Li, Wotao Yin, and Yin Zhang. Tval3: Tv minimization by augmented lagrangian and alternating direction algorithms. 2009.

[60] Xing Lin, Yair Rivenson, Nezih T. Yardimci, Muhammed Veli, Yi Luo, Mona Jarrahi, and Aydogan Ozcan. All-optical machine learning using diffractive deep neural networks. *Science*, 361(6406):1004–1008, 2018.

[61] Dong C. Liu and Jorge Nocedal. On the limited memory bfgs method for large scale optimization. *Mathematical Programming*, 45(1):503–528, Aug 1989.

[62] Yan Liu, Jianhua Ma, Yi Fan, and Zhengrong Liang. Adaptive-weighted total variation minimization for sparse data toward low-dose x-ray computed tomography image reconstruction. *Physics in Medicine and Biology*, 57(23):7923, 2012.

[63] F. J. MacWilliams and N. J. A. Sloane. Pseudo-random sequences and arrays. *Proceedings of the IEEE*, 64(12):1715–1729, Dec 1976.

[64] S. Mallat. *A Wavelet Tour of Signal Processing*. 01 2009.

[65] Donald W Marquardt. An algorithm for least-squares estimation of nonlinear parameters. *Journal of the society for Industrial and Applied Mathematics*, 11(2):431–441, 1963.

[66] K. Marwah, G. Wetzstein, Y. Bando, and R. Raskar. Compressive Light Field Photography using Overcomplete Dictionaries and Optimized Projections. *ACM Trans. Graph. (Proc. SIGGRAPH)*, 32(4):1–11, 2013.

[67] D. Mengu, Y. Luo, Y. Rivenson, and A. Ozcan. Analysis of diffractive optical neural networks and their integration with electronic neural networks. *IEEE Journal of Selected Topics in Quantum Electronics*, 26(1):1–14, Jan 2020.

[68] Christopher A. Metzler, Hayato Ikoma, Yifan Peng, and Gordon Wetzstein. Deep optics for single-shot high-dynamic-range imaging. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[69] Jianwei Miao, Tetsuya Ishikawa, Qun Shen, and Thomas Earnest. Extending x-ray crystallography to allow the imaging of noncrystalline materials, cells, and single protein complexes. *Annual Review of Physical Chemistry*, 59(1):387–410, 2008. PMID: 18031219.

[70] Kaushik Mitra, Oliver S. Cossairt, and Ashok Veeraraghavan. A framework for analysis of computational imaging systems: Role of signal prior, sensor noise and multiplexing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(10):1909–1921, 2014.

[71] Kristina Monakhova, Vi Tran, Grace Kuo, and Laura Waller. Untrained networks for compressive lensless photography. *Opt. Express*, 29(13):20913–20929, Jun 2021.

[72] Kristina Monakhova, Kyrollos Yanny, Neerja Aggarwal, and Laura Waller. Spectral diffusercam: lensless snapshot hyperspectral imaging with a spectral filter array. *Optica*, 7(10):1298–1307, Oct 2020.

[73] Kristina Monakhova, Joshua Yurtsever, Grace Kuo, Nick Antipa, Kyrollos Yanny, and Laura Waller. Learned reconstructions for practical mask-based lensless imaging. *Opt. Express*, 27(20):28075–28090, Sep 2019.

[74] Pushmeet Kohli Nathan Silberman, Derek Hoiem and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *ECCV*, 2012.

[75] Shree K. Nayar and Mohit Gupta. Diffuse structured light. In *2012 IEEE International Conference on Computational Photography (ICCP)*, pages 1–11, April 2012.

[76] Deanna Needell and Joel A. Tropp. Cosamp: Iterative signal recovery from incomplete and inaccurate samples. *Commun. ACM*, 53(12):93–100, December 2010.

[77] M. A. A. Neil, R. Juskaitis, and T. Wilson. Method of obtaining optical sectioning by using structured light in a conventional microscope. *Opt. Lett.*, 22(24):1905–1907, Dec 1997.

[78] Thuong Nguyen Canh and Hajime Nagahara. Deep compressive sensing for visual privacy protection in flatcam imaging. In *IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, Oct 2019.

[79] Ramesh Raskar, Amit Agrawal, and Jack Tumblin. Coded exposure photography: Motion deblurring using fluttered shutter. *ACM Trans. Graph.*, 25:795–804, 07 2006.

[80] Netanel Ratner, Yoav Y. Schechner, and Felix Goldberg. Optimal multiplexed sensing: bounds, conditions and a graph theory link. *Opt. Express*, 15(25):17072–17092, Dec 2007.

[81] Benjamin. Recht, Maryam. Fazel, and Pablo A. Parrilo. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM Review*, 52(3):471–501, 2010.

[82] D. Reddy, J. Bai, and R. Ramamoorthi. External mask based depth and light field camera. In *2013 IEEE International Conference on Computer Vision Workshops*, pages 37–44, Dec 2013.

[83] D. Reddy, A. Veeraraghavan, and R. Chellappa. P2c2: Programmable pixel compressive camera for high speed imaging. In *CVPR 2011*, pages 329–336, June 2011.

[84] Ilya Reshetouski, Hideki Oyaizu, Kenichiro Nakamura, Ryuta Satoh, Suguru Ushiki, Ryuichi Tadano, Atsushi Ito, and Jun Murayama. Lensless imaging with focusing sparse ura masks in long-wave infrared and its application for human detection. In *European Conference on Computer Vision (ECCV)*, pages 237–253, 2020.

[85] John M. Rodenburg. Ptychography and related diffractive imaging methods. volume 150 of *Advances in Imaging and Electron Physics*, pages 87–184. Elsevier, 2008.

[86] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, pages 234–241, 2015.

[87] Leonid I. Rudin, Stanley Osher, and Emad Fatemi. Nonlinear total variation based noise removal algorithms. *Phys. D*, 60(1-4):259–268, November 1992.

[88] Cyrus Bamji S. Burak Gokturk, Hakan Yalcin. A time-of-flight depth sensor - system description, issues and solutions. In *Conference on Computer Vision and Pattern Recognition Workshop*, pages 35–35, June 2004.

[89] A. Saxena, M. Sun, and A. Y. Ng. Make3d: Learning 3d scene structure from a single still image. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(5):824–840, May 2009.

[90] Ashutosh Saxena, Sung H. Chung, and Andrew Y. Ng. Learning depth from single monocular images. In Y. Weiss, B. Schölkopf, and J. C. Platt, editors, *Advances in Neural Information Processing Systems 18*, pages 1161–1168. MIT Press, 2006.

[91] D. Scharstein, R. Szeliski, and R. Zabih. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. In *Proceedings IEEE Workshop on Stereo and Multi-Baseline Vision*, pages 131–140, Dec 2001.

[92] Y. Y. Schechner, P. N. Belhumeur, and S. K. Nayar. Multiplexing for optimal lighting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(08):1339–1354, aug 2007.

[93] Mark Schmidt. minfunc: unconstrained differentiable multivariate optimization in matlab. 2005.

[94] Pradeep Sen, Billy Chen, Gaurav Garg, Stephen R. Marschner, Mark Horowitz, Marc Levoy, and Hendrik P. A. Lensch. Dual photography. In *ACM SIGGRAPH 2005 Papers*, SIGGRAPH '05, page 745–755, New York, NY, USA, 2005. Association for Computing Machinery.

[95] Pradeep Sen and Soheil Darabi. Compressive dual photography. In *Computer Graphics Forum*, volume 28, pages 609–618. Wiley Online Library, 2009.

[96] Yoav Shechtman, Yonina C. Eldar, Oren Cohen, Henry Nicholas Chapman, Jianwei Miao, and Mordechai Segev. Phase retrieval with application to optical imaging: A contemporary overview. *IEEE Signal Processing Magazine*, 32(3):87–109, 2015.

[97] Takeshi Shimano, Yusuke Nakamura, Kazuyuki Tajima, Mayu Sao, and Taku Hoshizawa. Lensless light-field imaging with fresnel zone aperture: quasi-coherent coding. *Appl. Opt.*, 57(11):2841–2850, Apr 2018.

[98] Vincent Sitzmann, Steven Diamond, Yifan Peng, Xiong Dun, Stephen Boyd, Wolfgang Heidrich, Felix Heide, and Gordon Wetzstein. End-to-end optimization of optics and image processing for achromatic extended depth of field and super-resolution imaging. *ACM Transactions on Graphics (SIGGRAPH)*, 2018.

[99] Qilin Sun, Ethan Tseng, Qiang Fu, Wolfgang Heidrich, and Felix Heide. Learning rank-1 diffractive optics for single-shot high dynamic range imaging. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[100] Dharmpal Takhar, Jason N. Laska, Michael B. Wakin, Marco F. Duarte, Dror Baron, Shriram Sarvotham, Kevin F. Kelly, and Richard G. Baraniuk. A new compressive imaging camera architecture using optical-domain compression. *Proc.SPIE*, 6065:6065 – 6065 – 10, 2006.

[101] Jasper Tan, Li Niu, Jesse K. Adams, Vivek Boominathan, Jacob T. Robinson, Richard G. Baraniuk, and Ashok Veeraraghavan. Face detection and verification using lensless cameras. *IEEE Transactions on Computational Imaging*, 5(2):180–194, June 2019.

[102] Z. Tan, P. Yang, and A. Nehorai. Joint sparse recovery method for compressed sensing with structured dictionary mismatches. *IEEE Transactions on Signal Processing*, 62(19):4997–5008, Oct 2014.

[103] G. Tang, B. N. Bhaskar, P. Shah, and B. Recht. Compressed sensing off the grid. *IEEE Transactions on Information Theory*, 59(11):7465–7490, Nov 2013.

[104] Lei Tian, Xiao Li, Kannan Ramchandran, and Laura Waller. Multiplexed coded illumination for fourier ptychography with an led array microscope. *Biomed. Opt. Express*, 5(7):2376–2389, Jul 2014.

[105] T. Tieleman and G. Hinton. Lecture 6.5—RmsProp: Divide the gradient by a running average of its recent magnitude. COURSERA: Neural Networks for Machine Learning, 2012.

[106] C. Tomasi and R. Manduchi. Bilateral filtering for gray and color images. In *Sixth International Conference on Computer Vision (IEEE Cat. No.98CH36271)*, pages 839–846, Jan 1998.

[107] J. A. Tropp and A. C. Gilbert. Signal recovery from random measurements via orthogonal matching pursuit. *IEEE Transactions on Information Theory*, 53(12):4655–4666, Dec 2007.

[108] Ashok Veeraraghavan, Ramesh Raskar, Amit Agrawal, Ankit Mohan, and Jack Tumblin. Dappled photography: Mask enhanced cameras for heterodyned light fields and coded aperture refocusing. In *ACM transactions on graphics (TOG)*, volume 26, page 69. ACM, 2007.

[109] Y. Wang, J. Yang, W. Yin, and Y. Zhang. A new alternating minimization algorithm for total variation image reconstruction. *SIAM Journal on Imaging Sciences*, 1(3):248–272, 2008.

[110] Gordon Wetzstein, Ivo Ihrke, and Wolfgang Heidrich. On plenoptic multiplexing and reconstruction. *International Journal of Computer Vision*, 101:384–400, 2012.

[111] Gordon Wetzstein, Aydogan Ozcan, Sylvain Gigan, Shanhui Fan, Dirk Englund, Marin Soljačić, Cornelia Denz, David A B Miller, and Demetri Psaltis. Inference in artificial intelligence with deep optics and photonics. *Nature*, 588(7836):39—47, December 2020.

[112] Charles S. Williams and Orville A. Becklund. *Introduction to the Optical Transfer Function*. 1989.

[113] Y. Wu, V. Boominathan, H. Chen, A. Sankaranarayanan, and A. Veeraraghavan. Phasecam3d — learning phase masks for passive single view depth estimation. In *2019 IEEE International Conference on Computational Photography (ICCP)*, pages 1–12, May 2019.

[114] Keita Yamaguchi, Yusuke Nakamura, Kazuyuki Tajima, Toshiki Ishii, Koji Yamasaki, and Takeshi Shimano. Lensless 3d sensing technology with fresnel zone aperture based light-field imaging. In *Industrial Optical Devices and Systems*, volume 11125, page 111250F. International Society for Optics and Photonics, 2019.

[115] Anqi Yang and Aswin C. Sankaranarayanan. Designing display pixel layouts for under-panel cameras. *IEEE Trans. Pattern Analysis and Machine Intelligence (TPAMI) / Special Issue of ICCP*, 43(7):2245–2256, July 2021.

[116] Z. Yang, L. Xie, and C. Zhang. Off-grid direction of arrival estimation using sparse bayesian inference. *IEEE Transactions on Signal Processing*, 61(1):38–43, Jan 2013.

[117] Kyrollos Yanny, Kristina Monakhova, Richard W. Shuai, and Laura Waller. Deep learning for fast spatially varying deconvolution. *Optica*, 9(1):96–99, Jan 2022.

[118] Adam Yedidia, Christos Thrampoulidis, and Gregory Wornell. Analysis and optimization of aperture design in computational imaging. *IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 4029–4033, April 2018.

[119] Ping-Hung Yin, Chih-Wen Lu, Jia-Shyang Wang, Keng-Li Chang, Fu-Kuo Lin, Chia-Jung Chang, and Gen-Chiuan Bai. A 368 × 184 optical under-display fingerprint sensor with global shutter and high-dynamic-range operation. In *2020 IEEE Custom Integrated Circuits Conference (CICC)*, pages 1–4, 2020.

[120] Tomoyuki Yokota, Kenjiro Fukuda, and Takao Someya. Recent progress of flexible image sensors for biomedical applications. *Advanced Materials*, 33(19):2004416, 2021.

[121] Yucheng Zheng and M. Salman Asif. Imaging with distributed lensless line sensors. In *2019 53rd Asilomar Conference on Signals, Systems, and Computers*, pages 1289–1293, Nov 2019.

[122] Yucheng Zheng and M. Salman Asif. Joint image and depth estimation with mask-based lensless cameras. *IEEE Transactions on Computational Imaging*, 6:1167–1178, 2020.

[123] Yucheng Zheng and M Salman Asif. Coded illumination for 3d lensless imaging. *IEEE Open Journal of Signal Processing*, 3:432–439, 2022.

[124] Yucheng Zheng and M. Salman Asif. Coded illumination for improved lensless imaging. *IEEE Transactions on Computational Imaging*, 9:172–184, 2023.

[125] Yucheng Zheng, Yi Hua, Aswin C Sankaranarayanan, and M Salman Asif. A simple framework for 3d lensless imaging with programmable masks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2603–2612, 2021.

[126] Yucheng Zheng, Rongjia Zhang, and M. Salman Asif. Coded illumination and multiplexing for lensless imaging. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 9250–9253, 2020.

[127] A. Zomet and S. K. Nayar. Lensless imaging with a controllable aperture. In *IEEE Computer Vision and Pattern Recognition*, volume 1, pages 339–346, June 2006.