Programmable Optics for Computational Photography

Dissertation

der Mathematisch-Naturwissenschaftlichen Fakultät

der Eberhard Karls Universität Tübingen

zur Erlangung des Grades eines

Doktors der Naturwissenschaften

(Dr. rer. nat.)

vorgelegt von

Jieen Chen
aus Weifang/V.R. China

Tübingen
2021

Gedruckt mit Genehmigung der Mathematisch-Naturwissenschaftlichen Fakultät
der Eberhard Karls Universität Tübingen.

Tag der mündlichen Qualifikation:	25.04.2022
Dekan:	Prof. Dr. Thilo Stehle
1. Berichterstatter:	Prof. Dr.-ing. Hendrik P. A. Lensch
2. Berichterstatter:	Prof. Dr. Bernhard Eberhardt

καὶ τὸ φῶς ἐν τῇ σκοτίᾳ φαίνει, καὶ ἡ σκοτία αὐτὸ οὐ κατέλαβεν

Abstract

Programmable optics is generating considerable interest in terms of computational photography. Recent developments regarding spatial light modulators (SLMs) have made optical coding in photography possible, which bears the promise of solving imaging tasks more cost-effectively. This thesis examines the joint design of optics and digital processing regarding both optical wavefront and ray encoding and inverse solvers and neural networks.

We start with an introduction to the fundamentals of image formation, the design paradigm, the details of the facilitated SLM, and the basics of image reconstruction. We then identify the point spread functions (PSFs) as a link between the optics and digital modules of the imaging chain and set up a computational camera, allowing user-defined PSFs through the phase-coded aperture. Once the camera is verified, we realize snapshot multispectral imaging by inspecting the chromatic dispersion of PSFs. A model-based framework is constructed to encode per-pixel spectra on a single image and then to decode multiple full-resolution spectral images with a regularizer-assisted inverse solver. While hand-tuned PSFs are effective, a holistic approach takes advantage of the large amounts of image data to learn the optical phase pattern. Next, we constitute and validate a data-driven joint-design method that optimizes physical parameters and digital reconstruction in an end-to-end fashion on a classical imaging problem: extended depth of field.

In addition, the joint-design scheme sheds light to solve an imaging conundrum: how to achieve an optically true hybrid zooming with a single lens on a single camera? We approach the solution by establishing a computational camera that modifies the chief ray distribution; ergo, it enables a programmable field of view.

Kurzfassung

Die programmierbare Optik stößt auf großes Interesse in Bezug auf die computergestützte Fotografie. Jüngste Entwicklungen bezüglich räumlicher Lichtmodulatoren (SLMs) haben dazu geführt, dass eine optische Kodierung in der Fotografie möglich ist, die das Versprechen birgt, Abbildungsaufgaben kostengünstiger zu lösen. Diese Arbeit untersucht die gemeinsame Gestaltung von Optik und digitaler Verarbeitung sowohl hinsichtlich optischer Wellenfront- und Strahlencodierung als auch Inverser-Solver und neuronaler Netze.

Wir beginnen mit einer Einführung in die Grundlagen der Bildentstehung, das Gestaltungsparadigma, die Details des erleichterten SLM und die Grundlagen der Bildrekonstruktion. Anschließend identifizieren wir die Punktspreizfunktionen (PSF) als Bindeglied zwischen optischen und digitalen Modulen der Abbildungskette und bauen eine computerbasierte Kamera auf, die benutzerdefinierte PSF durch die phasenkodierte Blende ermöglicht. Sobald die Kamera verifiziert ist, realisieren wir eine Schnappschuss-Multispektral-Bildgebung, indem wir die chromatische Dispersion der PSFs untersuchen. Ein modellbasierter Rahmen wird konstruiert, um Spektren pro Pixel in einem einzigen Bild zu kodieren und anschließend mehrere Spektralbilder in voller Auflösung mit einem inversen Solver zu dekodieren. Während handabgestimmte PSFs effektiv sind, nutzt ein ganzheitlicher Ansatz die großen Mengen an Bilddaten, um das optische Phasenmuster zu lernen. Als nächstes wird eine datengetriebene Joint-Design-Methode entwickelt und validiert, die die physikalischen Parameter und die digitale Rekonstruktion in einer End-to-End-Methode für ein klassisches Bildgebungsproblem optimiert: die erweiterte Tiefenschärfe.

Darüber hinaus wirft die Joint-Design-Methode Licht auf die Lösung eines Abbildungsproblems: Wie kann man ein optisch echtes Hybrid-Zoom mit einem einzigen Objektiv an einer einzigen Kamera erreichen? Wir nähern uns der Lösung, indem wir eine computerbasierte Kamera einrichten, die die Hauptstrahlverteilung modifiziert, sodaß ein programmierbares Sichtfeld ermöglicht wird.

Acknowledgments

First and foremost, I would like to thank GOD through Jesus Christ and in the Holy Spirit, who created me and saved me from my sins. I am deeply grateful for HIS word that gave me strength during my studies: "one who is faithful in a very little is also faithful in much, and one who is dishonest in a very little is also dishonest in much" (Luke 16:10). I am thankful for HIS forbearance despite my meager endeavors and failures.

I want to convey my gratitude toward Prof. Dr.-ing. Hendrik P.A. Lensch, who provided this valuable opportunity and a positive research environment and patiently instructed me with technical support, original ideas, and advice on scientific practices. I appreciate his insights, spirit of scientific exploration, and rigorous methodology. I want to thank Prof. Dr. Bernhard Eberhardt and Prof. Dr. Andreas Schilling for giving me this opportunity to pursue this research. It is also important to thank Dr. Michael Hirsch for his constant help in the first two projects. His expertise in imaging and digital deconvolution significantly contributed to the definition of the research direction. Our collaboration greatly supported my development as a scientist. I want to thank Andreas Engelhardt, who cooperated with me in the last two projects and allowed me to use materials from our collective projects. I appreciate his enthusiasm for computational photography, brilliant teamwork, and extensive knowledge of machine learning. I treasure the memory of cooperative problem-solving and those moments of reaching paper submission deadlines.

I thank all the collaborators in the projects. I thank Dr. Uli Wannek for generously lending a spectrometer that assisted the multispectral imaging project. I also thank my colleagues: Sebastian Herholz, Raphael Braun, Dr. Benjamin Resch, Fabian Groh, Dr.-ing. Christian Fuchs, Dr. Katharina Schwarz, Dr. Jian Wei, Dr. Andreas Karge, Dr. Patrick Wieschollek, Mark Boss, Arijit Mallick, Dennis Bukenberger, and Lukas Ruppert, for their friendship, thoughtful discussions, creativity, and technical support in the laboratory. I cherish all the joyful moments of overcoming research obstacles and sharing life. I also thank Violaine Le Guily, Sibylle Hasse, Manuela Di Paolo, and Birgit Grieg, who processed many of my documents for the university and various offices so that I could focus on the research.

I owe special thanks to my good friend Juan Purcalla Arrufi, whose research and life inspire me. I thank my encouraging friends who were researchers in the university: Dr. Jacolien van Rij and Dr. Elisabeth Früh. I appreciate the patient spiritual support of Dr. James Louis Kautt and Barbara Black. I also appreciate the long-lasting loyal love of my parents, Yuliang Chen and Hui Zhu. Last but not least, I thank my wife, Dr. Julia Louttit Chen. "An excellent wife is the crown of her husband" (Proverbs 12:4), and it is true in

scientific research too.

I thank GOD for the journey. Apart from HIM I can do nothing (John 15:5).

Papers Included in This Thesis

Papers included in this thesis are listed below:

[CHH+17] Jieen Chen, Michael Hirsch, Rainer Heintzmann, Bernhard Eberhardt, and Hendrik Lensch. A phase-coded aperture camera with programmable optics. *Electronic Imaging*, 2017(17):70–75, 2017.

[CHEL18] Jieen Chen, Michael Hirsch, Bernhard Eberhardt, and Hendrik PA Lensch. A computational camera with programmable optics for snapshot high-resolution multispectral imaging. In *Asian Conference on Computer Vision*, pages 685–699. Springer, 2018.

Contents

Chapter 1

Introduction

Computational photography utilizes digital techniques to capture and produce new types of photographic images, including high dynamic range (HDR) and light-field imaging, which have become standard operations. One breakthrough of computational photography is the invention of computational cameras, which combine optical and digital processing to generate unconventional images and thus measure object properties that are overseen by traditional cameras [ZN11]. They effectively solve the weak link in the imaging chain and extend the imaging capacity by firstly exploring the information connection between two modules–the optical and the digital; secondly, by joining the two with the help of novel optics and digital processing techniques. For instance, the combination of diffusive optics and image restoration enabled non-invasive imaging through scattering media as well as looking around corners [KHFG14]. Computational photography is a paradigm shift from both optical design and computer vision, where the former focused on measuring radiance and the latter on processing data. The recent success of deep learning also contributes to this shift.

The differentiator between optics and digital processing is the rich computational resources of the latter. Much potential is promised by programmable optics, such as spatial light modulators (SLMs), commonly used in computer-generated holograms. The optical coding, such as diffractive wavefront modulation and ray distortion map, will assist reconstruction and be optimized along with image restoration. Integrating programmable optics will make cameras holistically computational and open them up to novel computational models to optimize physical and digital parameters. Such computational cameras will solve traditional imaging problems with less energy consumption by taking fewer images for more information and serve as a platform for computational photography research. Therefore, our objective is to set up the computational camera hardware and build up its software framework for optical programming and digital computing.

To realize this goal, one must identify the mechanism of information transmission between optics and digital processing. This link is manifested in optical and digital image processing similarities, including blur and deblurring, optical and digital Fourier transforms, and ray geometry. The first similarity brings us to PSF-engineering in the optical domain and deconvolution in the digital domain. For example, previous works have explored the depth information encoded in the PSFs through amplitude-coded-aperture

techniques and promoted the digital deconvolution for the extended depth of field (EDoF) [LFDF07]. In addition to depth, PSFs also contain spectral features, for example, in the chromatic aberrations. Although aberrations are traditionally unwanted in lens design, they can transmit information for digital processing.

The outline of this thesis and the research questions are as follows. Chapter 2 introduces light and image formation fundamentals, the joint design paradigm, and hardware and software tools. We also describe the SLM parameters and review a classic joint design example. Next, we examine **research question 1:** *how to establish a computational camera that integrates the optical and digital modules in the imaging chain?* In Chapter 3, we incorporate the SLM with an off-the-shelf camera. Using phase-coded aperture, we empower the computational camera to allow the user to define PSF spatial distributions. The forming of PSFs is simulated through a hologram-generating technique that takes advantage of the optical Fourier transforms. The imaging properties of the SLM are examined. An application of refocusing is demonstrated through encoding Fresnel lenses.

Chapter 4 advances this computational camera regarding **research question 2:** *can the joint design of optical coding and image reconstruction realize single-shot spectral imaging?* The critical idea is to generate images with chromatically modified PSFs. The optical multiplexing encodes the spectra of each pixel as 2D "rainbows," and the optimizer decodes the spectra through image deconvolution, that is, the restoration of multiple spectral images from a snapshot. The PSFs are modulated to be spectrally variant and cover a wide range of chromatic bands. The joint-design framework fuses the inverse problem solver that consists of the image formation model with the knowledge of the PSFs. Due to the highly ill-posed nature, intra- and inter-band regularizers are designed and employed in the optimization process using the L-BFGS-B solver. A proof-of-concept is made in simulation as well as in the real world. Besides, the performance of diverse PSF designs is tested and compared. The supplementary results are shown in Appendix A.

After the success of spectral imaging with hand-crafted PSFs, the next step is to answer **research question 3:** *is there a data-driven framework that optimizes optical parameters and image reconstruction all at once?* We employ convolutional neural networks (CNNs) to learn both the PSFs and image processing. Inspired by the end-to-end network, in Chapter 5, we adopt an autoencoder that comprises a differential renderer as the optical encoder and a UNet-based decoder for the EDoF application. The purpose is to verify the prototype of learning-based joint-design and discover novel optical phase coding and restoration. The optical encoder replicates the computational camera's real-world image formation while also parametrizing the SLM phase pattern as a grayscale image to be learned. The output of the encoder is the intermediate image to be captured by the camera. Next, we capture real-world images as inputs to the decoder. The training uses image inputs located at randomly distributed depths. The decoder then recovers the sharp image. Both simulated and real-world images are recovered. The learning-based joint-design has the potential to solve one critical problem in computational photography–

optical-digital hybrid zooming using a single shot with one unifocal lens, which could drastically reduce the energy consumption of multi-camera solutions and the expense of zooming lenses while keeping physical fidelity.

With the learning-based scheme installed, we pursue a novel application driven by **research question 4:** *is the computational camera with the programmable optics capable of achieving hybrid zoom with a single lens?* In Chapter 6, a chief-ray modulation computational camera is built to modify the distortion map of the field of view.

Chapter 2

Background and Overview

Fundamentally, the task of photography is to measure the radiant energy from the objects in a scene. This process uses optics, image sensors, and digital image processing to infer the properties of the objects. By exploring the collection of light, optical engineering has assimilated the mechanism of human eyes to build cameras that capture images. Computer vision research has enabled machines to understand images by mimicking the human organization of visual information. In this thesis, we investigate the co-dependence of these two components of the imaging chain through extending the programmability to the optics and transmitting optical information to the digital. Since a good understanding of the underlying physics is key for system modeling and optimization, this chapter is devoted to these foundations.

We start by describing the theories of light and the mathematical model of image formation of common optical systems in section 2.1. Next, in section 2.2, the concept and the mathematical model of the imaging chain of the digital camera system are presented. We highlight and concisely justify the idea of joint design of optics and image processing. Considering the programmable optical device–the HOLOEYE PLUTO liquid crystal on silicon (LCoS) spatial light modulator (SLM), we specify its physical characteristics in section 2.3. In section 2.4, we then introduce some image reconstruction algorithms. Last but not least, we demonstrate some joint-design examples of optical and digital image processing in section 2.5.

2.1 Theories of Light and Image Formation

2.1.1 Modelling Light

To understand image formation, one must investigate the nature of light. Light is a form of electromagnetic radiation. Therefore the theoretical principles of electromagnetic radiation govern the phenomena of light. Quantum optics explains nearly all phenomena of light. However, within various scales and conditions, approximations can be initiated. When confined in the classical range, namely when light is treated as an electromagnetic wave, electromagnetic optics provides a complete explanation of light. If the physical condition of scalar approximation is fulfilled, phenomena of light can be well described

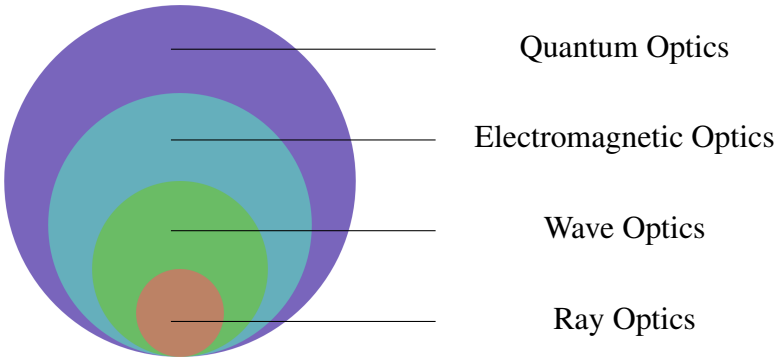


Figure 2.1: Theories of light. Quantum optics interprets virtually all optical phenomena. When quantum effects are negligible, electromagnetic optics explains light as an electromagnetic wave. When electromagnetic effects such as polarization are trivial, wave optics using the scalar wave function characterize the interference and diffraction faithfully. If, when compared to the environment, the wavelength is close to zero, ray optics is an efficient tool to describe light propagation.

through scalar wave optics or wave optics. If the wavelength is close to infinitely small compared to the environment, ray optics characterize light propagation sufficiently well. Figure 2.1 illustrates the hierarchy of theories being used in various conditions to model light.

Here we introduce these theories from micro- to macro-scale. In quantum optics, light is treated as photons that carry electromagnetic energy, momentum, and angular momentum associated with polarization. Photons also carry the wave principles of interference and diffraction. The photon-number statistics are helpful to understand and reduce the Poisson noise present in digital images. Regarding a photodetector, quantum efficiency specifies the ratio between charged carriers and incoming photons.

Governed by the vector wave theory, electromagnetic optics describes light as a kind of electromagnetic radiation with the form of coupled electric and magnetic vector waves. The starting point is the macroscopic (without considering atomic scale charges and quantum phenomena) Maxwell's equations,

$$\nabla \times \vec{E} = \frac{\partial \vec{B}}{\partial t} \tag{2.1}$$

$$\nabla \times \vec{H} = \vec{j}_m + \frac{\partial \vec{D}}{\partial t} \tag{2.2}$$

$$\nabla \cdot \vec{D} = \rho_{ext} \tag{2.3}$$

$$\nabla \cdot \vec{B} = 0 \tag{2.4}$$

where $\vec{E}$ is the electric field, $\vec{B}$ is the magnetic flux density, $\vec{D}$ is the dielectric flux density,

$\vec{H}$ is the magnetic field, ρ_{ext} is the external charge density, and $\vec{j_m}$ is the macroscopic current density. The auxiliary equations are,

$$\vec{D} = \varepsilon_0 \vec{E} + \vec{P} \tag{2.5}$$

$$\vec{H} = \frac{1}{\mu_0}[\vec{B} - \vec{M}] \tag{2.6}$$

where $\vec{P}$ is the dielectric polarization, and $\vec{M}$ is the magnetic polarization. $\times$ and $\cdot$ represent a vector cross product and a vector dot product, respectively.

In linear, isotropic, homogeneous, and non-dispersive media (e.g., free space), wave optics is sufficient for explanations. The electric and magnetic fields fulfill the wave equation as follows,

$$\nabla^2 u - \frac{n^2}{c^2}\frac{\partial^2 u}{\partial t^2} = 0 \tag{2.7}$$

n is the refractive index, and c is the speed of light in free space. The wave equation represents any scalar components u of the electromagnetic wave. Despite the inability to describe the phenomena on the boundaries of dielectric media and polarization, it explains the manifestation of the waveform, such as interference and diffraction. The wavefront is the locus where the wave shares the same phase of the sinusoid. This concept is significant to model the propagation of light.

Ray optics is a further simplified model based on the rough approximation that the wavelength is close to 0, where no wave is considered. Rays can be regarded as the direction of the energy flow governed by the Poynting vector,

$$\vec{S} = \vec{E} \times \vec{H} \tag{2.8}$$

Concerning the wave optics, the rays are perpendicular to the wavefronts. Fermat's principle governs optical rays traveling between two points: light rays travel along the path of least time. When rays transmit through the boundary of two media, Snell's law describes refraction,

$$n_1 \sin(\theta_1) = n_2 \sin(\theta_2) \tag{2.9}$$

where n_1 and n_2 are the refractive indices of the two media. θ_1 is the incidence angle, and θ_2 is the refraction angle.

In this thesis, we use these models to comprehend diverse phenomena. For the purpose of image formation, in chapter 3, 4, and 5, we use wave optics to engineer the PSFs, and in chapter 6, we employ ray optics to program the field of view. To drive the LCoS SLM, we exploit electromagnetic optics to understand the light-matter interaction of incoming light and liquid crystal cells. Due to the quantization of photon energy, we consider quantum optics to recognize Poisson noises.

To illustrate the models of light propagation, in Figure 2.2, we show a point source that is imaged by an optical system depicted as a black box. The light source emanates infinite

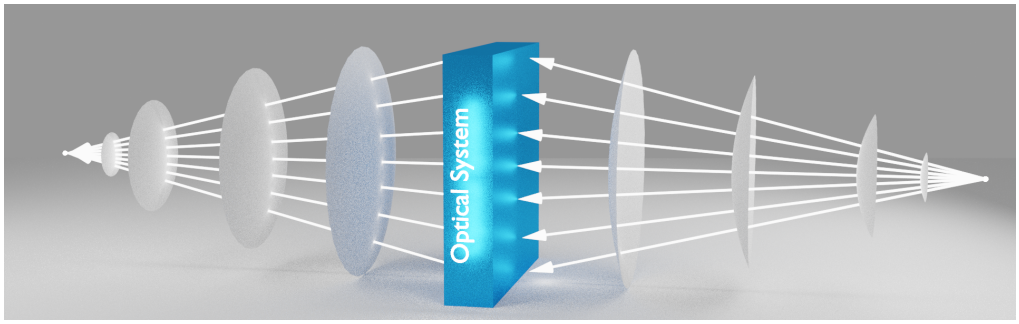


Figure 2.2: Theories used in this thesis for image formation: ray and wave optics. A point source emits light from the right side. The spheres are the wavefronts at various locations, and the arrows indicate the rays. An optical system redirects light to form a focused point on the left side.

amounts of light rays bundled by a spherical wave. The optical system then collects and redirects the rays or deforms the wavefront to generate a focused point image.

2.1.2 Image Formation

Image formation [BM13] is the process of redistributing light from the object space to the image space. This process is mostly a deterministic mapping defined by a mapping operator, which is to say that, for each object, there is a defined image. However, for one image, there could be multiple possible objects that produce the same image. Additionally, noise has to be considered because repeated images of the same object feature non-identical noise. The image formation mechanism can be formulated as,

$$\mathbf{O} = \mathbf{H}\mathbf{I} + \mathbf{n} \tag{2.10}$$

where $\mathbf{I}$ is the matrix representation of the latent image(s), $\mathbf{O}$ is the captured image, $\mathbf{H}$ is the mapping operator, and $\mathbf{n}$ is the noise. Traditionally, image formation seeks to produce a replica of the object as the image. However, this undermines some information embedded in the light, such as color spectra and ray angles. In this thesis, we expand the mapping operation and use the joint-design paradigm for image reconstruction.

The image formed by the optics on the sensor plane is the distribution of optical intensity. To understand image formation using visible light in our computational cameras, we must understand light propagation using wave optics and ray optics.

Propagation of Light Using Wave Optics

A monochromatic wave at location P and time t can be written as,

$$u(P,t) = A(P)\exp[i\phi(P,t)] \tag{2.11}$$

where $A(P)$ is the amplitude and $\phi(P,t)$ is the phase. The task of image formation using wave optics is to predict this intensity distribution through determined wavefront relationships.

The scalar diffraction foretells the propagation of the optical wavefront. While the stops in optical systems obstruct the wavefront, diffraction describes the wavefront propagation phase, energy, and density distribution. Following the Huygens-Fresnel principle (that is, the optical field of any point is a superposition of wavelets from every unobstructed point of a wavefront), general theories such as Kirchhoff and Rayleigh-Sommerfeld yield accurate estimation. However, approximations can be performed through simple mathematical formulation in some circumstances. The Fraunhofer and Fresnel approximations are commonly used in assessing optical systems.

If the object locates close to the optical axis, namely under the paraxial condition, we can apply the Fresnel approximation,

$$U(x,y) = \frac{\exp(ikz)}{i\lambda z} \iint_{-\infty}^{\infty} U(\xi,\eta) \exp\{i\frac{k}{2z}[(\xi - x)^2 + (\eta - y)^2]\} d\xi d\eta \tag{2.12}$$

where (ξ,η) represents the coordinates of the wavefront before propagation, λ is the wavelength, and $k = \frac{2\pi}{\lambda}$ is the wave number. If the object is in the far-field where the so-called Fresnel number $N_f \lesssim 0.1$ with $N_f = \frac{a^2}{\lambda z}$ and a being the aperture radius, we can further simplify the approximation by the Fraunhofer formula [Goo05]. The condition indicates that the object is far from the aperture.

$$U(x,y) = \frac{\exp[ikz + i\frac{k}{2z}(x^2+y^2)]}{i\lambda z} \iint_{-\infty}^{\infty} U(\xi,\eta) \exp[-i\frac{2\pi}{\lambda z}(x\xi + y\eta)] d\xi d\eta \tag{2.13}$$

Analytically or numerically computing all wavefront propagation through each optical element can be inefficient due to its complexity. Fortunately, lenses, as the most commonly used elements, have deterministic effects. Optical Fourier transform refers to the Fourier transform effects seen when using lenses. The optical phase transformation of a thin lens is written as,

$$t_l(x,y) = \exp[-i\frac{k}{2f}(x^2+y^2)] \tag{2.14}$$

where f is the focal length. To estimate the effects of a thin lens on an incident disturbance is to compute the field distribution after the lens. In a $2f$ system (the distance between the lens and incident wavefront, as well as the distance between the lens to the image plane, are both the focal length), the complex amplitude of the wavefront at the back focal plane of a lens is the complete Fourier transform of that at the front focal plane.

When a wavefront is close to the front surface of the lens, on the focal plane, a quadratic phase curvature appears in the phase term. However, the amplitude remains

the Fourier transform of the wavefront.

$$U_f(x,y) = \frac{\exp[i\frac{k}{2f}(x^2+y^2)]}{i\lambda f} \iint_{\infty}^{\infty} U_l(\xi,\eta)\exp[-i\frac{2\pi}{\lambda f}(x\xi+y\eta)]dxdy \tag{2.15}$$

Despite the phase term, the amplitude of the incoming wavefront and the output amplitude distribution are related by the Fourier transform. Since the lens is the last optical element in our imaging task, we take this arrangement as the basis for our system before we measure the optical intensity.

As we have seen, the phase component and the wave nature of light are convenient in evaluating the wavefront. However, the frequency response differs with coherent and incoherent illuminations. In photography, one frequently uses natural or studio lighting, which confines us to incoherent imaging. Incoherent imaging employs incoherent light for imaging and differs from coherent imaging. Photographic lighting sources for artists' or consumers' purposes, such as daylight, flash, or continuous artificial lighting, are incoherent. This lighting has both advantages and disadvantages. The advantages are the simple setup and insensitivity to coherent diffractive artifacts caused by dust. The disadvantages are the limits of using an interference technique to subtract wavefronts and the achievable resolution. The distinction in image formation is the PSF and the convolution operation. In coherent imaging, the image intensity is,

$$O(x,y) = \|p(x,y;\xi,\eta) \otimes U_g(\xi,\eta)\|^2 \tag{2.16}$$

where $O(x,y)$ is the image intensity, $p(x,y;\varepsilon,\eta)$ is the complex spread function with both amplitude and phase components, and $U_g(\varepsilon,\eta)$ is the emitting wavefront from the object. The incoherent imaging [Goo05] is formulated as,

$$O_i(x,y) = \|p(x,y;\xi,\eta)\|^2 \otimes \|U_g(\xi,\eta)\|^2 \tag{2.17}$$

Considering Equation (2.23), the PSF and the complex spread function in incoherent imaging are related as,

$$h(x,y;\xi,\eta) = \|p(x,y;\xi,\eta)\|^2 \tag{2.18}$$

For a photographic lens, the convolution operator is $\mathbf{H}$ in Equation (4.1), which is $h(x,y;\xi,\eta)$ after vectorization.

The coherent and incoherent imaging frequency content can be quite different; this is often observed at the edges of images. There is no interference with light from other points in the scene, and the intensities are simply added. Therefore, the incoherent imaging system is shift-invariant.

The problem for PSF engineering, in this thesis, is to find the phase modulation function given the amplitude PSF. This algorithmic process is called phase retrieval. Phase retrieval algorithms must satisfy other constraints such as the stop size, the distance between the wavefronts, and the wavelength. Because of the lightness and small size of the

diffractive optical elements (DOEs), phase retrieval is commonly used to cooperate with DoEs in coherent diffraction imaging, ptychography, and lensless imaging.

Propagation of Light Using Ray Optics

Ray tracing [MS15] renders 3D scenes and produces a 2D image using image-order or object-order algorithms by finding the object seen at the pixel positions in the image from a particular viewpoint. A ray tracer computes the geometry of the viewing rays and ray-plane intersections and incorporates shading models and BRDFs in rendering.

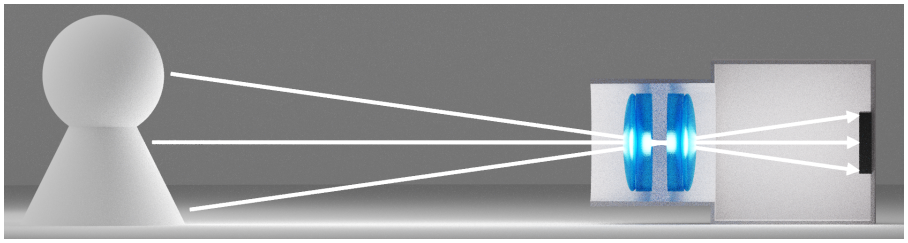


Figure 2.3: Ray tracing from the object through the lens to the sensor.

Marginal and chief rays are crucial for lens design because they are reference rays that show aberration properties. The field of view (FoV) of a photography lens is defined by the chief rays originating from the edge of an object and running through the center of the aperture. In optical design, the intercepts of the chief rays and the focal plane or the spot centroid are often referenced to define the FoV.

2.2 The Imaging Chain, Computational Photography and Joint-Design Approach

2.2.1 The Imaging Chain and Computational Photography

The imaging chain [Fie10] outlines the chain of physical events that form, process, and display an image, which is perceived in the final step. It consists of the light source, the objects of a scene, optics that collects and redistributes light, the photosensor that converts photons into digital counts, digital image processing that enhances or restores the bitmap, displays, and the interpretation and perception of the image.
The mathematical model of the key elements of the end-to-end imaging chain for digital camera systems is the linear shift-invariant (LSI) model. The LSI contains both properties of linear systems and shift-invariant systems. The imaging system can be summarized as,

$$O[f(x,y)] = g(x',y') \tag{2.19}$$

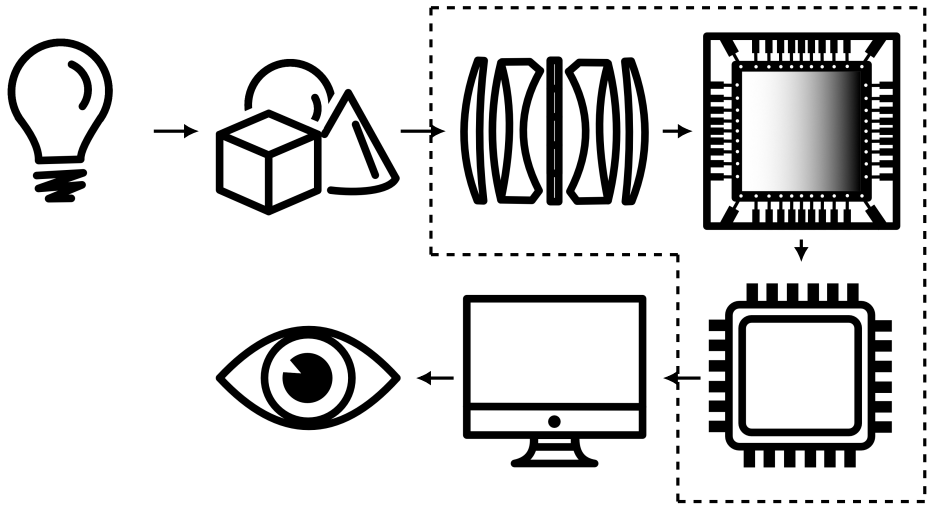


Figure 2.4: The imaging chain. From beginning to end: lighting, objects, optics, sensor, digital processing, displays, and perception. The joint design paradigm optimizes the optics and the processing as a unison with sensor in between these two modules.

where O is the operation function that transforms the input $f(x,y)$ to the output $g(x',y')$. A linear system contains two properties,

$$O[f_1(x,y)+f_2(x,y)] = g_1(x',y')+g_2(x',y') \tag{2.20}$$
$$O[\alpha f(x,y)] = \alpha g(x',y') \tag{2.21}$$

Equation (2.20) and (2.21) are the additivity and homogeneity properties.
The shift-invariant system is characterized as,

$$O[f(x+\triangle x, y+\triangle y)] = g(x'+\triangle x', y'+\triangle y') \tag{2.22}$$

The response of a single point, namely the impulse response function of an imaging system, is known as the point spread function (PSF).

$$O[\delta(x-x_0, y-y_0)] = h(x'-x'_0, y'-y'_0) \tag{2.23}$$

where $h(x,y)$ represents the PSF.

In camera design, the links between these modules play significant roles in producing quality images. Since the quality of the images is defined by the weakest link, the vital engineering task is to seek and enhance the dominant weak link.

In recent years, the imaging chain has been extended by the development of computational photography [Nay06]. It seeks to capture and process images by enabling the computational power of individual modules of the imaging chain, such as programmable lighting, optics, as well as the optimized color filter array of the sensors. This has gen-

erated various novel solutions to 3D measurement with a light-field camera, denoising, demosaicing, deblurring, computational microscopy, spectral imaging, super-resolution, and HDR imaging. In this thesis, we facilitate the programmability of the imaging optics and conjoin the digital processing.

2.2.2 Joint-Design Paradigm

Due to the kinship of optical and digital image processing, we propose the idea of joint design that takes these two separate modules as one conjunction. The two components are shown in the dashed area in Figure 2.4. We empower the optics with a programmable optical device to uphold this model. Data representations can accordingly be shared to manipulate the incoming light.

What Links Optical and Digital Image Processing?

At least three mathematical models link optical and digital image processing: PSF, Fourier transform, and ray geometry. The notion of PSF is arguably the most vital in optical imaging. Defined in Equation (2.23), the PSF is the response of a point input of a lens. The PSF can be a Dirac delta function in an ideal system (i.e., aberration free and no diffraction). In reality, a real-world PSF has a spatial distribution. When the PSF is invariant under translations, the image on the sensor is subsequently the result of a convolution operation,

$$O_i(x,y) = h(x,y;\xi,\eta) \otimes I(\xi,\eta) \tag{2.24}$$

where $I(\varepsilon,\eta)$ is the latent image, and $O_i(x,y)$ is the image on the sensor plane. The filtering kernel in digital image processing similarly functions as a convolution operation. In many cases, the PSF varies across the spatial domain. Therefore variance has to be modeled, measured, and calibrated.

The second link is the Fourier transform. Both optical and digital image processing use the Fourier theory to operate on the image spectrum. Optical waveforms can be treated as the superposition of plane waves. The discipline of Fourier optics is dedicated to the use of Fourier theory to analyze optical waves. In digital image processing, discrete Fourier transforms (DFT) are utilized to process images in the frequency domain. The Fourier transform in one dimension is,

$$F(\xi) = \mathcal{F}\{f(x)\} = \int_{-\infty}^{\infty} f(x) e^{-2\pi i \xi x} dx \tag{2.25}$$

The third link is ray geometry (e.g. light field [LH96]). Resonating ray optics, the light field uses the plenoptic function, a vector function that depicts the directional information of rays flowing in space, which provides key information of depth. Additionally, image formation can be simulated through ray tracing.

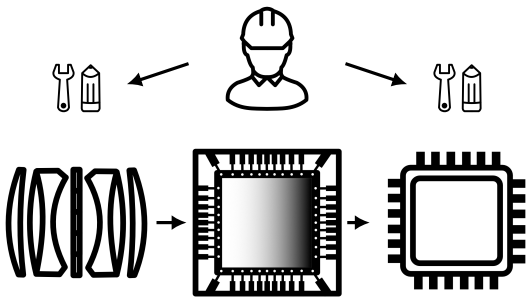


Figure 2.5: The model-based approach.

How to Link Optical and Digital Image Processing

The joint design paradigm requires the design of optics and digital image processing. The corresponding aspects, according to Yaroslavsky [Yar11], are:

- discrete representation of analog convolution and Fourier transforms for sampled signals;
- characterization of digital filtering and Fourier analysis;
- building continuous image models, image recovery from sparse data, signal differentiation and integration;
- digital-to-analog conversion in computer-generated holography.

In this thesis, we consider all the aspects above in chapter 3, 4, and5, and additionally the ray or geometric optics aspect in chapter 6, where the field of view is estimated by sampling the chief ray.

There are two approaches for the design scheme: model-based and data-driven. As illustrated in Figure 2.5, the model-based method demands user-defined linking functions (i.e., PSFs) to encode the desired optical information in the image formation process. As is shown in Figure 2.6, a classical inverse problem solver is utilized to estimate the latent images. On the other hand, the data-driven approach takes advantage of the image datasets and end-to-end convolutional neural network (CNN) to optimize the linking functions and the reconstruction networks all at once. The optical characteristics are parametrized as part of the end-to-end CNN to be learned. This is an indispensable process to ensure that simulation conforms to real physics.

Information from Multiple Aspects: the Optical and the Digital

The definition of information is one of the central questions in the philosophy of information and the philosophy of science. There are many views about the definition of information [FEM05, Bas17]. Shannon [Sha48] recognized that the nature of information

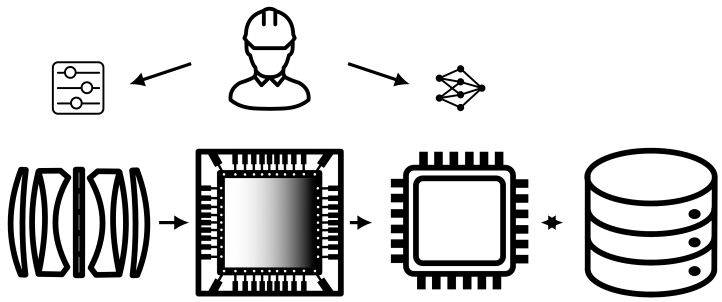


Figure 2.6: The data-driven approach.

and the physical thermodynamic notion of entropy are related, while Wiener [Wie19] argues that information cannot be reduced to matter and energy, and information is related to order, a capacity for useful work.

Regarding light, the photons originate from atoms. Through their propagation in space and time, one can infer information carried by light such as the spatial location, spectrum, and material properties. Digital images containing finite, 2D-discretized quantities are characterized by probability distributions. With the links between optical and digital information, it is legitimate to encode physically and decode numerically.

Alavi and Leidner [AL01] state that "data is raw numbers and facts, information is processed data, and knowledge is authenticated information." If so, both optical and digital image processing of the imaging chain denote stages of data processing. Especially with the continuous enlargement of image datasets, the inherent correlation between the later stage of digital processing (formative aspect) and the earlier stages of optical processing (physical aspect) can be exploited through deep learning. It is reasonable to draw and inject information from multiple aspects: the optical and the digital.

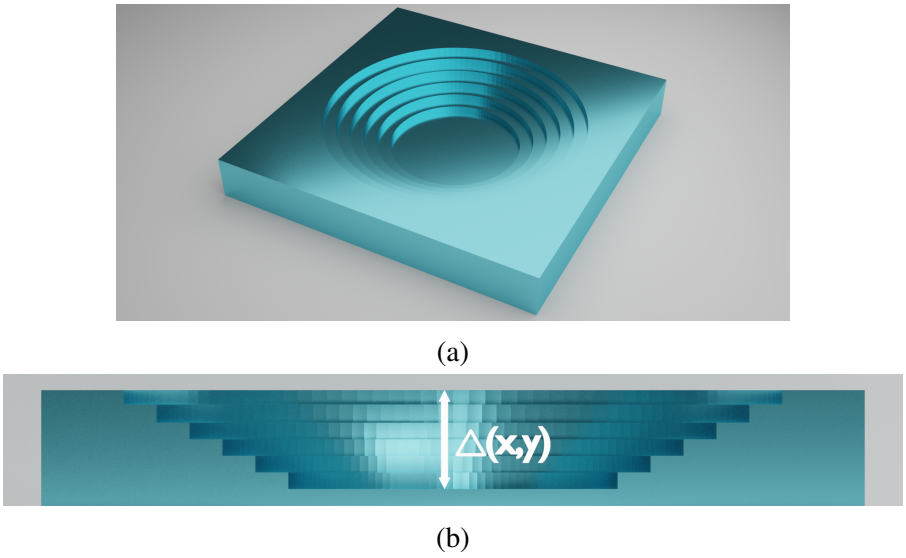
2.3 Programmable Optics

To capacitate the programmability of the image formation, we adopt the LCoS SLM to control the phase component of the optical wavefront. Similar to an LCD panel, the LCoS SLM displays an entire image. The grayscale value represents the refractive index, and therefore the phase modulation function for a given wavelength. Through programming the phase modulation function, we enable the imaging optics to be parametrized.

Diffractive Optics Elements (DOEs)

DOEs [Goo05] use diffraction instead of commonly used refraction or reflection to control light distribution. They are phase elements with thin micro-structures. A DOE can reshape light to almost any designed distribution through diffraction and wavefront propagation. The shape of a wavefront is changed without varying other parameters such as

polarization, and wavelength.



(a)

(b)

Figure 2.7: Binary DOE. (a) is the full view of a conceptual binary DOE, whose surface relief profile is designed with concentric rings with different depths. (b) shows the phase modulation is exercised by tuning the height of each location at (x,y).

One developed DOE employs binary optics. The binary optics are manufactured using photolithography and micromachining. The optical phase is manipulated by its surface relief profile, such as in Figure 2.8. This can be formulated as,

$$\phi(x,y) = 2\pi\frac{n(\lambda)\triangle(x,y)}{\lambda} \tag{2.26}$$

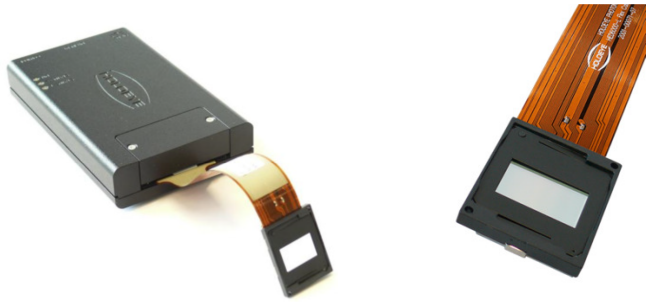
where $\triangle(x,y)$ is the height at location (x,y). DOEs can be highly dispersive, depending on the material used and the phase modulation itself. Ideally, applying a continuous phase function would fulfill the phase modulation purpose. However, this has to be approximated by a thickness function with a limited step height in practice.

This thesis uses the LCoS phase-only SLM, a liquid crystal-based microdisplay with each pixel's refractive index tunable according to the driving unit. The height of all pixels in the display panel remains the same. Instead of controlling the height $\triangle(x,y)$, the phase function is modulated through varying the refractive indices $n(x,y;\lambda)$. The input to the SLM is, therefore, a grayscale image.

HOLOEYE PLUTO LCoS SLM

The HOLOEYE PLUTO LCoS SLM is a plug-and-play, phase-only SLM with an HDMI interface. The SLM consists of a driver that enables it to function as an external display device. The resolution is HD (1920×1080). The video signal is communicated through

the green color channel. 8-bit grayscale images are used for addressing the SLM. Each grayscale value corresponds to the addressing voltage through a look-up table (LUT). The vendor provides SLMs optimized for different bandwidths. Our imaging tasks operate under visible light. Therefore we use the PLUTO-VIS-006-A (420nm-720nm) version with an anti-reflection coating with a front reflection of less than 0.5%. Its related parameters are shown in Table 2.1. The maximum phase shift is shown in Table 2.2.



(a) (b)

Figure 2.8: Driving unit and the LCoS panel of HOLOEYE PLUTO-VIS-006-A. (a) is the driving unit and the panel with the liquid crystal pixel array. (b) is the zoomed-in view of the LCoS panel. Image copyright: HOLOEYE

Name	**Characteristics**
Type	LCoS (reflective), active matrix LCD
Drive scheme	Digital pulse width modulation
Mode	Nematic
Phase levels	256 levels
Active area	15.36 mm $\times$ 8.64 mm
Resolution nominal	HD (1920 $\times$ 1080)
Pixel pitch	8.0 µm
Fill factor	87%
Image frame rate	60Hz

Table 2.1: Parameters of HOLOEYE PLUTO-VIS-006-A.

Wavelength	**Maximum phase shift**	**Average reflectivity**
452nm	9.0π	60%
532nm	6.7π	60%
633nm	5.4π	60%

Table 2.2: Maximum phase shift and average reflectivity of PLUTO-VIS-006-A.

2.4 Image Reconstruction

2.4.1 Mathematical Problem Description

After the optically encoded images are captured, the digital processing module continues the imaging chain. We formulate the reconstruction problem and give a short introduction to the general methods and the ideas we developed in this thesis. As shown in Equation (4.1), the image formation as the forward model can be vectorized. The reconstruction aims to restore the latent image(s) from the optically encoded image. The inversion that recovers the latent images can be formulated as a standard minimization,

$$\arg\min_{\mathbf{I}} = J(\mathbf{I}) = \arg\min_{\mathbf{I}}[\|A(\mathbf{I}) - \mathbf{O}\|_2^2 + R(\mathbf{I})] \tag{2.27}$$

where $A(\cdot)$ is the programmed optical image formation model that can consist of a convolution operation with PSFs, mosaicing, vignetting, noises, and FoV mapping. $\mathbf{O}$ is the captured image and $\mathbf{I}$ is the latent image(s). Due to the highly ill-posed nature of the problem, there is no unique solution; prior knowledge of the restored dataset, such as edge sharpness and sparsity, is critical for producing convincing results. $R(\cdot)$ is the regularization term that penalizes unrealistic restoration. The goal of the inverse problem in imaging is to infer the properties of the incoming light from the captured observation.

2.4.2 Deconvolution

To illustrate the inverse problem in imaging, we make an example of deconvolution. A commonly observed photograph artifact is camera motion, which is caused by the motion-blurred PSFs. To recover the sharp image, in the case of $A(\cdot)$ being the convolution operation with a blur kernel in matrix form $\mathbf{H}$ in Equation (4.1), the task is called deconvolution. We call the problem non-blind deconvolution if $\mathbf{H}$ is known and blind deconvolution if not.

Image priors are employed to bring forth knowledge of the optimized images. Commonly, the latent images bear specific natural image statistical characteristics, such as the distribution of the histogram of image gradients [FSH+06], or image smoothness (L2) or boundary continuity (L1).

Since convolution is multiplication in Fourier space, as a non-blind deconvolution method, the Wiener deconvolution was developed to apply the Wiener filter [Wie49] to suppress noise amplification in division in Fourier space,

$$\tilde{I}(\omega_x, \omega_y) = \frac{H^*(\omega_x, \omega_y) O(\omega_x, \omega_y)}{|H(\omega_x, \omega_y)|^2 + K} \tag{2.28}$$

(ω_x, ω_y) denote the coordinates in Fourier domain. $H^*(\omega_x, \omega_y) = H(-\omega_x, -\omega_y)$ is the conjugate of $H(\omega_x, \omega_y)$. K is a constant indicating the estimated signal-to-noise ratio.

After an inverse Fourier transform, the restored **I** is obtained.

Using the Bayesian approach, the observation of the original image has a particular probability,

$$p(\mathbf{I}|\mathbf{O}) = \frac{p(\mathbf{O}|\mathbf{I})p(\mathbf{I})}{p(\mathbf{O})} \tag{2.29}$$

where $p(\mathbf{O})$ is the prior model of the input **I**, $p(\mathbf{O}|\mathbf{I})$ is the probability of observing **O** given **I**. Our goal $p(\mathbf{I}|\mathbf{O})$ is to reconstruct **I** given the observation **O**. With $p(\mathbf{O}|\mathbf{I})$ being the convolution and Poisson noise, Richardson-Lucy deconvolution [Ric72a] iteratively solves the problem in the spatial domain.

$$I^{n+1}(x,y) = I^n(x,y)\left[h^*(x,y) \otimes \frac{O(x,y)}{(I^n \otimes h)(x,y)}\right] \frac{1}{1 - w\Gamma(x,y)} \tag{2.30}$$

where $\Gamma(x,y)$ is a potential regularizer (i.e., Tikhonov regularizer) and w is its weight. I^n is the nth iteration of the restored image.

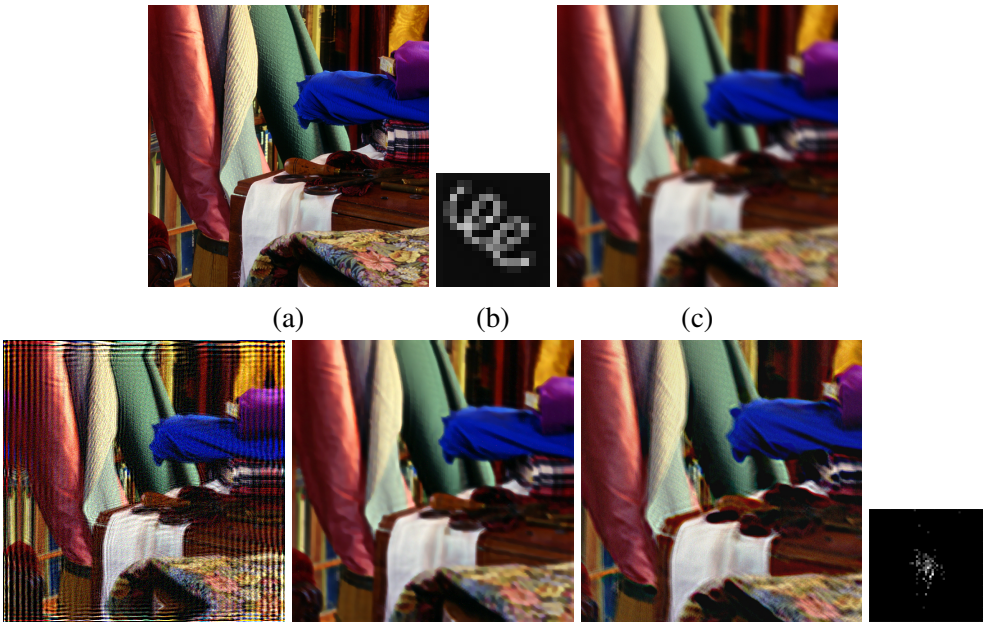
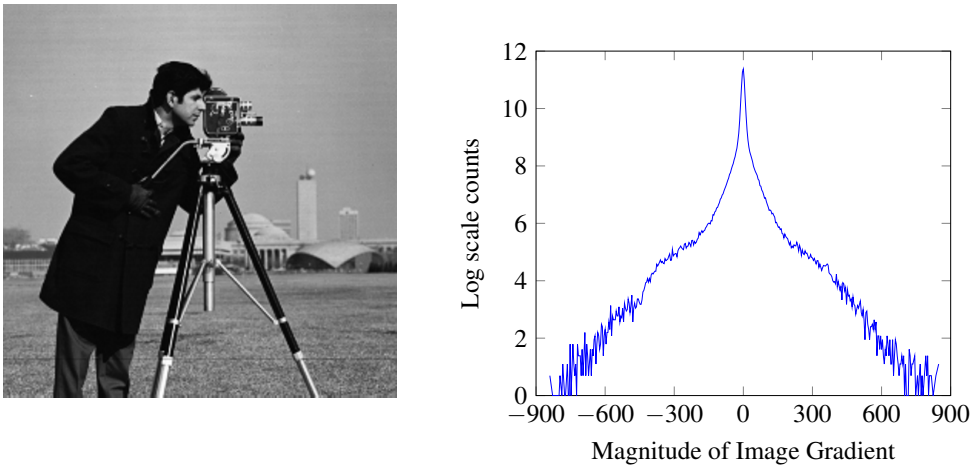


Figure 2.9: Deconvolution. (a) is the original image (500×500). (b) is the blur kernel (16×16). (c) is the blurred image with Gaussian noise. (d) is the Wiener reconstruction. (e) is the Richardson-Lucy reconstruction. (f)(g) are the blind deconvolution [FSH$^+$06] and its estimated blur kernel.

The blind deconvolution estimates the kernel and the restored image at the same time. To remove the blur from a single shot, Fergus *et al.* [FSH$^+$06] employ natural statistics and find the blur kernel in a patch in the gradient domain by Bayesian approach,



(a) (b)

Figure 2.10: Natural image statistics by Fergus *et al.* [FSH+06]. (a) is the original image, whose image gradient in both vertical and horizontal directions are computed. (b) is the histogram of the image gradient magnitude. The x-axis is the magnitude of the image gradients, and the y-axis is the log-scale counts. It shows the heavy-tails of the magnitude histogram and a Gaussian-mixture model [FSH+06] to approximate such distributions.

then a standard deconvolution algorithm follows to restore the image. Since the kernel is unknown, multiple combinations of intrinsic images and kernels can result in the same output. Natural statistics is the histogram of the image gradient. The sparsity and positivity of the blur kernel are enforced as the regularizers. A variational Bayesian method is used to estimate the kernel in multiple scales. In the end, a Richardson-Lucy deconvolution is utilized to generate the final output. An example of natural statistics is shown in Figure 2.10.

2.4.3 Optimization Methods

Since the image formation model can be more complex than convolution, a more general solution to the inverse problem is desired. Therefore, we come to numerical optimization, which is covered by numerous instances of [NW06]. Optimization methods seek the maximization or minimization of an objective function with or without constraints on its variables (i.e., the positivity of the optical intensity, and boundary values of the image) Basic unconstrained methods are line search methods, trust-region methods, conjugate gradient methods, quasi-Newton methods, least-square problem solvers, and non-linear equations. The constrained methods are linear programming, quadratic programming, Lagrangian methods, and interior-point methods. Since the goal is to find out the minima, when the objective function $J(\mathbf{I})$ in Equation (2.27) is smooth, we can find out if $\mathbf{I}$ is the local minimum by inspecting its gradient $\nabla J(\mathbf{I})$ and the Hessian $\nabla^2 J(\mathbf{I})$. Taylor's

theorem is the fundamental tool for this analysis.

Two basic strategies of optimization are line search and trust region. The line search finds a direction and a step length to approach new iterates with lower function values. The trust region approximates the objective function with a model and finds the minimum in a restricted region with an estimated step. This model is usually defined as the quadratic function,

$$M(\mathbf{I}_k + \mathbf{q}_k) = J_k + \mathbf{q}_k^T \nabla J_k + \frac{1}{2}\mathbf{q}_k^T \mathbf{B}_k \mathbf{q}_k \tag{2.31}$$

where $\mathbf{B}_k$ is the Hessian matrix or its approximation. The difference between line search and trust region is the order of finding direction and distance in each iteration. The line search starts with a fixed direction and looks for a optimal distance, however, the trust region fixes a distance and then identifies a direction.

The simplest solver is the steepest descent, where the direction is $-\nabla J(\mathbf{I})$ and a step length is chosen. It accurately traces towards the local minimum despite the severe slowness. If the direction is derived from the second-order derivative, we obtain the Newton method. We find the descent direction $\mathbf{q}_k$ by setting the $M(\mathbf{I}_k + \mathbf{q}_k)$ to be zero when $\mathbf{B}_k$ is the Hessian matrix $\nabla^2 J(\mathbf{I})$. As long as the second-order derivative exists and is smooth, and Equation (2.31) is a close approximation of $J(\mathbf{I})$, the convergence is guaranteed with a fast rate. The main shortcoming is the evaluation of the Hessian matrix. This optimization can be a heavy computational process and is inclined to miscalculation. The quasi-Newton method keeps the convergence rate and reduces the burden of computing the Hessian matrix by approximating it iteratively using $\mathbf{B}_k$. With each step, additional knowledge is accounted for to update this variable.

One of the most popular quasi-Newton methods is the BFGS algorithm named after the four authors. The basic principle of passing the knowledge from the previous iteration to the next is that the gradient of M_k should match both $\mathbf{I}_k$ and $\mathbf{I}_{k+1}$. Following the analytical solution of $\mathbf{B}_k$, the search direction and the step size are computed with each iteration until the convergence tolerance is reached. The L-BFGS by Byrd *et al.* [BNS94] is the extension to a limited end by avoiding storing the approximated $\mathbf{B}_k$ at each step. Each step begins with the assumption that the $\mathbf{B}_{k-1}$ is an identity matrix and stores some vectors to update $\mathbf{B}_k$.

The L-BFGS-B [ZBLN97] extends the L-BFGS method further to a constrained optimization subject to simple bounds of the variables,

$$\mathbf{l} \leq \mathbf{I} \leq \mathbf{u} \tag{2.32}$$

where the vectors $\mathbf{l}$ and $\mathbf{u}$ represent the lower and upper bounds on the variables. Since the inverse problems in imaging are constraints with bounds such as the non-negativity of radiance, the finite dynamic range of the sensor, and the color filter array bandwidth limitations due to spectral sampling, the L-BFGS-B is a powerful optimizer to solve these problems with known bounds.

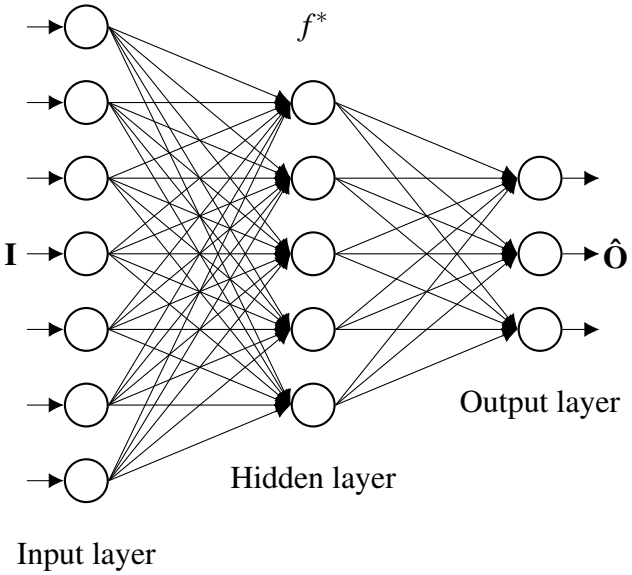


Figure 2.11: Multilayer perceptron.

2.4.4 Convolutional Neural Networks

While numerical optimization strictly follows the image formation model $A(\cdot)$ and the image prior in numerical optimization limits the solutions to a support region, the model mismatch, calibration errors, heuristic regularizers, and the manually tuned parameters could lead to reconstruction artifacts. Another approach does not use a hand-picked regularizer $R(\mathbf{I})$ or $p(\mathbf{I})$ but learns the estimator from ground truth data, and forms the images by artificial neural networks (ANNs) [GBC16]. In practice, this data-driven approach learns the image statistics from a large dataset of captured photos.

Popularized by the study of cybernetics and inspired by the biological brain, neural networks are composed of neurons that have connections with weights. The values of the weights denote an excitation of inhibition. After a combination of the inputs, an activation function controls the output. The quintessential deep learning model is the multilayer perceptron, also known as feedforward neural network, where there are no feedback connections from the output.

Shown in Figure 2.11, the multilayer perceptron maps a set of input values to the output values by forming several simple functions with a network. It consists of an input layer, at least one hidden layer, and an output layer. The input layer contains the variables we observe. The hidden layers contain data values not given in the input, rather, abstract features explain the relationships in the data such as edges, corners, and contours. This network architecture features nodes and operations, which is described by a computational graph. This graph is defined by,

$$\hat{\mathbf{O}} = f(\mathbf{I}, \theta) = f^n(\dots f^3(f^2(f^1(\mathbf{I}, \theta_1), \theta_2), \theta_3) \dots, \theta_n) \tag{2.33}$$

where $\mathbf{I}$ is the input layer, $\hat{\mathbf{O}}$ is the output layer, and θ contains the parameters to be learned. The functions f^* are the hidden layers. The layers are basic building blocks with input and parameters. A loss function (i.e., a mean square error) must be defined to measure the error E against the ground truths. Usually, the parameters are randomly initialized and a gradient-based optimizer is used to force the cost function to a low value. To compute the gradient of the model, we need to enable the information to flow from the cost function backwards through the network body. This process is called back-propagation, which is frequently used to train a neural network. The back-propagation is computed by the chain-rule,

$$\mathbf{g}_i = \frac{\partial E}{\partial \theta_i} = \frac{\partial E}{\partial f_n(\mathbf{I})} \frac{\partial f_n(\mathbf{I})}{\partial f_{n-1}(\mathbf{I})} \frac{\partial f_{n-1}(\mathbf{I})}{\partial f_{n-2}(\mathbf{I})} \cdots \frac{\partial f_i(\mathbf{I})}{\partial \theta_i} \tag{2.34}$$

Each layer's parameters are then updated iteratively.

Gradient descent optimizers [Rud16] are commonly used to compute the updates. The simplest batch gradient descent calculates the gradients of each training sample as follows.

$$\theta_i^{m+1} = \theta_i^m - \beta \mathbf{g}_i^m \tag{2.35}$$

where β is the weighting constant. This optimizer guarantees to converge, yet it is slow and requires large memory. The stochastic gradient descent (SGD) performs on randomly selected samples to avoid redundancy and speed up this process. Since one bottleneck the optimizers face is to have proper learning rates for each iteration, several methods are developed correspondingly. The Adagrad method [DHS11] adapts the learning rates by normalizing the root mean square of previous gradients.

$$\theta_i^{m+1} = \theta_i^m - \frac{\beta}{\sqrt{\sum_{t=1}^{m} (\mathbf{g}_i^t)^2}} \mathbf{g}_i^m \tag{2.36}$$

To solve the problem of fast decaying learning rate, the RMSprop method [HSS12] uses the average of the gradients for this adaptation.

$$\theta_i^{m+1} = \theta_i^m - \frac{\beta}{\sqrt{\mathsf{E}[(\mathbf{g})^2]^m + \varepsilon}} \mathbf{g}_i^m \tag{2.37}$$

where $\mathsf{E}[(\mathbf{g})^2]^m$ is the mean value of the gradients at iteration m and ε is a constant. Among the SGD methods, the commonly used Adam optimizer [KB14] combines these two methods with a weighting factor.

The fully connected networks are memory-consuming to process 2D image data. One well-performed method is the convolutional neural networks (CNNs) [LBD$^+$89]. CNNs use at least one convolution operation in place of matrix multiplication in its layers. (The relevance of the convolution operation for optical and digital image processing has been shown repeatedly.) By exploring the sparse connectivity, the CNNs need much less

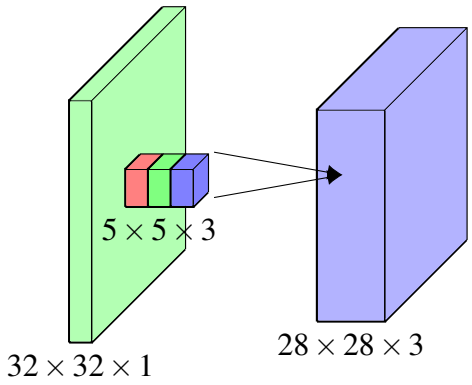


Figure 2.12: Convolution layer. The convolution operation is shown where a $28 \times 28 \times 3$ layer is computed from a $32 \times 32 \times 1$ layer by a convolution $5 \times 5 \times 3$ kernel.

memory compared to a fully connected network. CNNs are particularly beneficial in image processing because typical features, such as edges, only occupy small areas.

A typical convolution layer has three stages: convolution, detection, and pooling. The convolution yields at least one linear activation layer. Figure 2.12 shows an example of a $32 \times 32 \times 1$ layer operated by a $5 \times 5 \times 3$ kernel into a $28 \times 28 \times 3$ layer. To describe a convolution operation, we set the dimension of the first layer to be $n \times n \times c$, the kernel size to be $k \times k \times c_k$, the stride length to be s, and the padding size to be $m \times m$. The output dimension is,

$$(\frac{n+2m-k}{s}+1) \times (\frac{n+2m-k}{s}+1) \times c_k \tag{2.38}$$

The detection then runs this layer through some non-linear activation function (i.e., sigmoid, rectified linear unit (ReLU), and leaky ReLU). Afterward, the output is adjusted by pooling functions such as max pooling, where the maximum output within a rectangular neighborhood is selected.

Since the inception of the ConvNet to classify written letters [LBD+89], the CNN architectures have developed rapidly for the past few decades [KSZQ20], which generated the LeNet [LBBH98], AlexNet [KSH12], GoogleNet [SLJ+15], ResNet [HZRS16], VGG [SZ14], autoencoder (AE) [HS06], U-Net [RFB15], SegNet [BKC17], DenseNet [IMK+14], and PolyNet [ZLCLL17].

In particular, the autoencoder networks, as shown in Figure 2.13, copy the input to the output by an encoder and a decoder. This processing is described as,

$$J(\mathbf{I}, g(f(\mathbf{I}))) \tag{2.39}$$

where $J(\cdot)$ is the objective function such as the mean square error and regularizers in Equation (2.27). The bottleneck enforces the network to concentrate on latent information. In other words, this network is more able to process than classify.

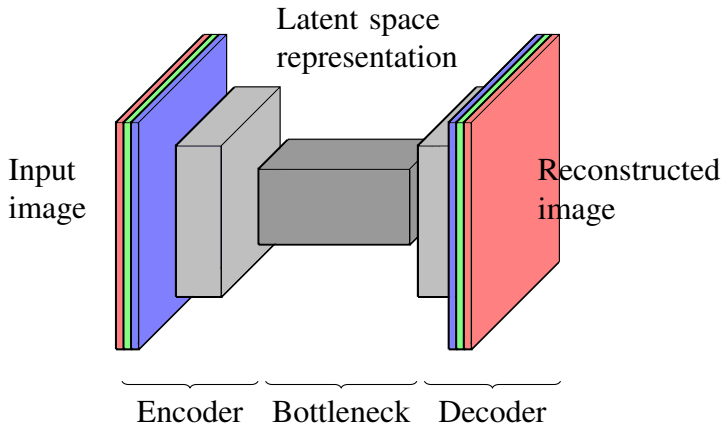


Figure 2.13: Autoencoder. The autoencoder is comprised of two parts: an encoder and an decoder. It copies the input to the output to process images. A bottleneck typically locates in the middle of this contracting and expanding network, where the latent space representation presides.

One widely used architecture is the U-Net, whose scheme is shown in Figure 2.14. The U-Net was initially developed for segmentation for biological images. It comprises a contracting path and an expanding path. The contracting path captures the context. The convolution in the same sampling level is 3×3 operations. The downsampling then follows, with a rectified linear unit and a 2×2 max pooling operation with stride 2. The symmetric expanding path enables the inverse operation–localization. The upsampling step is through a 2×2 convolution followed by two 3×3 convolution operations, in turn, followed by a ReLU. For each scale of one upsampling level, the cropped feature map is concatenated with the input by skip connections. Finally, a 1×1 convolution follows in mapping the classes of segmentation.

The encoder-decoder is also similar to the imaging forward and inverse process described in previous sections. if, for example, we have the encoder that follows the image formation function $A(\mathbf{I})$ in Equation (4.1), the latent image $\mathbf{I}$ is the input data to the encoder-decoder. The output of the encoder is the intermediate image that we expect to capture. The encoder reads the encoded image and reconstructs the latent image. The mapping process is then $p_{encoder}(\mathbf{O}|\mathbf{I})$ and $p_{decoder}(\mathbf{I}|\mathbf{O})$. In the real world, the encoder can be substituted by a computational camera as an optical encoder, where optical properties are parametrized within the neural network.

Despite the powerful performance, CNNs have some limits, such as the difficulty of incorporating image formation, being hard to interpret, and having no convergence guarantees.

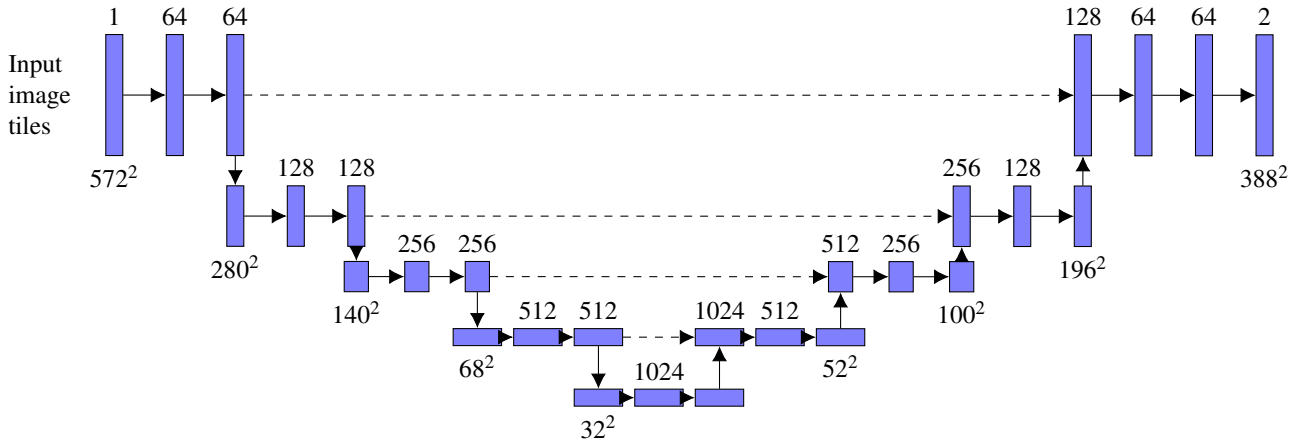


Figure 2.14: U-Net architecture by Ronneberger *et al.* [RFB15]. The blue boxes are multi-channel feature maps. On top of each box are the numbers of channels, and on the bottom are the image dimensions. The down and up arrows are convolution operations. The dashed arrows are skip connections of copies and crops.

2.5 A Joint-Design Example: Cubic Phase Plate for EDoF

As a fundamental example, we outline the classic joint-design paradigm–wavefront coding for extended depth of field (EDoF) [DJC95, CD02]. This optical-digital incoherent imaging system employs a phase mask that modifies the incoming wavefront to engineer depth-invariant PSFs. The PSFs are altered so that they are invariant of the misfocus term in the aberration functions. A photosensor captures the modulated intermediate image. A subsequent image deconvolution operation decodes the sharp latent image in the full field of view with an extended depth of field.

To design the phase mask, the authors develop the theory of wavefront coding. They first analyze the misfocus by the so-called ambiguity function. The ambiguity function is shown to represent the polar display [BLOC83] of the optical transfer function (OTF, the Fourier transform of the PSF) of an imaging system. This relation in 1D is described as,

$$H(u,\phi) = B(u, u\frac{2}{\lambda}W_{20}) \tag{2.40}$$

where $H(\cdot)$ is the optical transfer function, $B(\cdot)$ is the ambiguity function, W_{20} is the misfocus term, and u is the normalized coordinate of the spatial frequency. Thus, for a given amount of defocusing (W_{20}), the OTF associates with the ambiguity function by a horizontal (in focus) or slanted (defocused) line cross the polar display. Regarding the ambiguity function, to avoid the out-of-focus OTF lines of going through zeros points

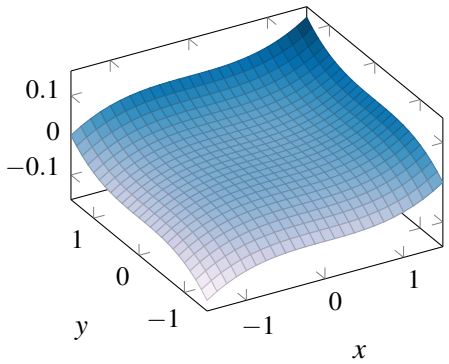
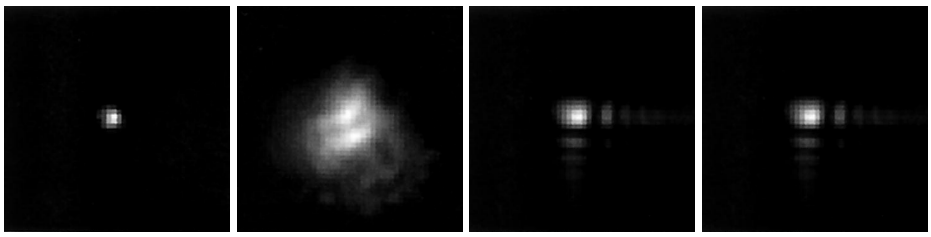


Figure 2.15: Relief profile of the cubic phase plate $\triangle(x,y) = \gamma(x^3 + y^3)$. The plot is made at $\gamma = 0.02$.

(phase shifts) and to spread the OTFs over a larger region of misfocus, the authors add a cubic phase function to the aperture stop to modulate the ambiguity function. Recall that the phase modulation in Equation (2.26), the surface profile $\triangle(x,y)$ is described as,

$$\triangle(x,y) = \gamma(x^3 + y^3) \tag{2.41}$$

where γ is a constant for tuning the range of the depth of field. The cubic phase plate is shown in Figure 2.15. The depth invariant PSFs are displayed in Figure 2.16.



(a) (b) (c) (d)

Figure 2.16: The engineered PSFs by Cathey and Dowski [CD02]. (a) and (c) are PSFs without modulation in focus and out of focus. (b) and (d) are the PSFs with the cubic phase plate in focus and out of focus.

The cubic phase plate is not the unique solution to the problem of extended depth of field, but it provides computational efficiency for the following deconvolution because it is rectangularly separable. The rectangularly separable processing independently restores the rows and columns with two 1D filters. The speed of processing the intermediate image is preferable to nonseparable 2D processing. An example of the intermediate image and the reconstructed results are shown in Figure 2.17. A depth mask (Figure 2.17b) differentiates the blur kernels of the foreground (Figure 2.16a2.16c) and background (Figure 2.16b and 2.16d). The non-modulated intermediate image is shown

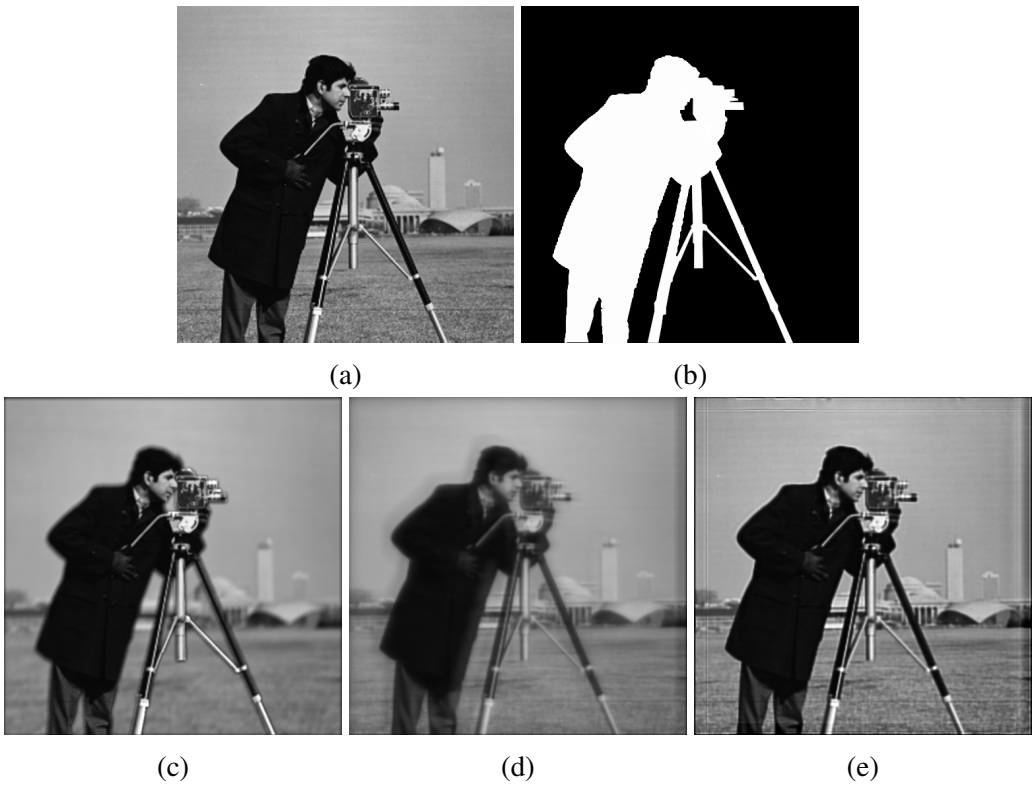


Figure 2.17: Extended depth of field (EDoF) with the cubic phase plate and deconvolution. (a) is the original camera man image. (b) is a depth map to differentiate the foreground and background scenes. (c) is the blurred image using non-modulated PSFs in Figure 2.16a 2.16b. (d) is the PSF-engineered intermediate image by the cubic phase mask. Its PSFs are shown in Figure 2.16c and 2.16d. (e) is the reconstruction from (d) with the Richardson-Lucy algorithm.

in Figure 2.17c where the far scene is blurred due to limited depth of field. The PSF-engineered intermediate image shown in Figure 2.17d is formed by convolving the depth invariant PSFs. A deconvolution by the Richardson-Lucy method using a known PSF outputs the extended depth of field (Figure 2.17e).

This joint-design imaging system gives insight into the internal link of optical and digital image processing. In this thesis, we boost the strength of wavefront coding by programmable optics to encode phase functions with more flexibility and advance the algorithms to reconstruct essential information and optical design.

Chapter 3

A Computational Camera with Programmable Optics

The material of this chapter is based on the following publication:

[CHH+17] Jieen Chen, Michael Hirsch, Bernhard Eberhardt, and Hendrik PA Lensch. A computational camera with programmable optics for snapshot high-resolution multispectral imaging. In *Asian Conference on Computer Vision*, pages 685–699. Springer, 2018.

We propose a novel computational imaging system that enables the generation of point spread functions (PSFs) of user-specified geometry. The key ingredient of our system is a phase-coded aperture which manipulates the phase distribution of the pupil function by inserting a phase modulator. We use a reflective phase-only liquid crystal-based spatial light modulator (SLM) for phase modulation. By encoding a grayscale image on the SLM, the refractive index of each cell can be altered. Phase patterns of PSFs with different shapes are optimized by the Gerchberg-Saxton algorithm. A number of non-trivial, complex-shaped PSFs has been captured. We further demonstrate how such a system can realize refocusing through encoding a Fresnel lens phase pattern to shift the focal plane.

3.1 Introduction

Imaging systems under partially coherent illumination can be described as linear systems and as such are fully characterized by their reponse to a point light source. This response is known as the point spread function (PSF) since the image of a point light source would typically get dispersed and extend over several pixels on the imaging plane.

While in traditional imaging systems, the PSF – being a consequence of the optical design – is considered fixed, in computational imaging systems, ways are studied that enable deliberate PSF manipulation.

Pupil plane coding is often used for modulating PSFs of an imaging system. For example, amplitude-coded aperture techniques modify the transmittance of the pupil to explore the PSF pattern. Especially when defocus exists, the amplitude-coded aperture

can be employed for depth map measurement. However, the energy distribution cannot be fully controlled. In contrast, phase-coded aperture techniques manipulate the phase of an optical wavefront. By modulating the phase at different spatial locations, light experiences different degrees of delay. By doing so, the wavefront phase distribution can be controlled and therewith its corresponding PSF. Phase-coded aperture techniques explore this property to control PSFs. More applications could potentially arise by providing freedom to manipulate PSFs into arbitrary shapes.

We propose a phase-coded aperture setup for PSF engineering using a spatial light modulator (SLM). By placing the SLM at the pupil plane of a camera lens, we explore the Fraunhofer diffraction relation between a PSF and its complex pupil function. We generate phase patterns encoded on the SLM using a standard phase retrieval algorithm to produce PSFs of arbitrary shape.

Our described system can: (1) produce a PSF following a target shape; (2) use the temporal ability of the SLM to obtain a better fit by altering the pattern sent to the SLM over time; and (3) even allow for changing the PSF of the system during capture. As one particular application, we show how our setup can be used to realize refocusing by encoding a phase pattern of a Fresnel lens on the SLM.

3.2 Related Work

A **point spread function (PSF)** is the intensity impulse response of an optical system. In computational photography applications, the illuminant light is in most cases incoherent to a level far beyond the size of the detection pupil. As a consequence, the PSF is determined by the square of the amplitude impulse response function [Goo05]. At the same time, the captured image can be computed by convolution of the PSF with the emitted intensity. The PSF is often used as a measure for the quality of an imaging system since the PSF captures the deviations from an ideal optical system, i.e. optical aberrations [Smi07] such as chromatic aberrations and misfocus.

Pupil plane coding is a computational approach for engineering PSFs [ZN11] by placing optical elements at the pupil plane. The pupil plane is the surface where all the chief rays from different object points cross the optical axis and pivot about. This allows for a uniform modulation of the incoming wavefront. A well-known approach is coded aperture, which uses an optimized occluder pupil pattern to preserve high frequency in case of defocus. Levin *et al.* [LFDF07] insert a patterned occluder within the aperture of the lens of a conventional camera, which creates corresponding pre-designed PSFs at different depths. By estimating the deconvolution filter and introducing a sparse prior, image and depth are simultaneously recovered. Cossairt *et al.* [CZN10] places an optical diffuser at the pupil plane. Due to light scattering, the PSFs of their system are invariant to the focal plane, which preserves high frequencies across different depths. A subsequent deconvolution is implemented to recover an image with extended depth of field.

Phase-coded aperture, often referred to as wavefront coding, places a phase modulator

at the pupil, e.g. a plate of glass with a 3D profile or a spatial light modulator (SLM) at, or close to the pupil plane of a photographic lens. The complex pupil function can be modified by optical aberration theory [Smi07] or phase retrieval algorithms [Fie82]. Herewith the shape of PSFs can be modulated.

The best known technique is wavefront coding for extended depth of field [DJC95]. By designing a cubic shape 3D profile of the phase plate, the incoherent wavefront is altered and the PSFs become independent of the misfocus function. An image with extended depth of field is then recovered by a deconvolution of the intermediate image. Chi *et al.* [CG09] propose a lensless phase-coded aperture imaging system by combining a phase-only screen and a detector array. Peng *et al.* [PFA+15] propose a phase-coded aperture lens realized by a diffractive optical element to shape the PSFs.

A **spatial light modulator (SLM)** is a real-time device containing a microscopic pixel array that is capable of spatially modifying the incoming wavefront in response to optical or electrical control signals. Amplitude, polarization, and phase of the complex optical wave distribution can be modified using SLMs. SLMs are widely used in display, holography, adaptive optics, and optical computing. There are various types of SLMs, which operate by varying the pixel cell height or the refractive index of the cell [Goo05]. With the latter type, pure optical phase modulation can be achieved by placing a linear polarizer in front. Phase profiles can then be encoded to manipulate the wavefront. For this reason, SLMs are widely used in adaptive optics applications in astronomy and biology to correct optical aberrations [Tys15, Kub13]. In [MPCRR14], a programmable diffractive lens phase profile is encoded on a SLM for ophthalmic applications. The reflectance display of Glasner *et al.* [GZL14] uses a liquid crystal on silicon(LCoS) SLM to fabricate BRDFs. Carles *et al.* [CMBH10] use an SLM as an adaptable phase mask for wavefront coding.

3.3 PSF Engineering

PSF engineering can be achieved by modulating the pupil function. We use the phase-coded aperture approach to manipulate the PSF distribution by employing a phase-only SLM as a phase mask. This phase-modulating SLM in the pupil plane acts similarly to an optical lens: however, instead of spatially varying the thickness of a glass plate, we encode the SLM with a grey level image to program the refractive indices of the cell array. We use a standard phase retrieval algorithm to compute the phase pattern that is deployed on the SLM. Before we describe our phase estimation in more detail, we will first revise the relation between the pupil function that we are going to modulate and the resulting PSF.

3.3.1 Fraunhofer Approximation

The PSF of an incoherent imaging system is the squared magnitude of the amplitude response function. Therefore, the phase term of the response function provides an additional degree of freedom to optimize the PSF distribution.

According to the Fraunhofer approximation [Goo05] the PSF of a diffraction limited optical system is obtained as the Fourier transform of the exit pupil function. The PSF of this system can be formulated as

$$I_f(x,y;\lambda) = \frac{A^2}{\lambda^2 f^2} \| \iint_{-\infty}^{\infty} \mathcal{P}(x_0,y_0) \exp\left[-i\frac{2\pi}{\lambda f}(xx_0+yy_0)\right] dx_0 dy_0 \|^2 \tag{3.1}$$

where A is a constant denoting the amplitude, f is the focal length, and $\mathcal{P}(x_0,y_0)$ is the complex pupil function. The complex generalized pupil function of a modulated system is

$$\mathcal{P}(x,y) = t_A(x,y)\exp[i\phi(x,y)] \tag{3.2}$$

where $\phi(x,y)$ is the wavefront deformation function introduced by the SLM, and $t_A(x,y)$ is the amplitude transmittance associated with the limited size of the lens pupil.

3.3.2 Phase Modulation

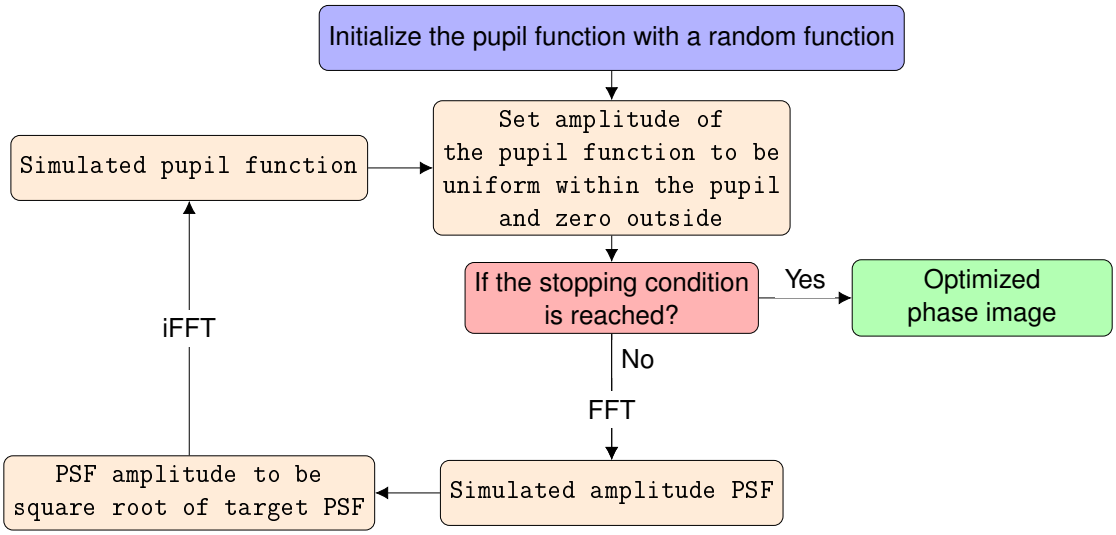


Figure 3.1: Overview of our optimization procedure for the pupil phase function.

The PSF modelling problem can be seen as one of designing a proper function $\phi(x,y)$. For phase estimation, we use the commonly used Gerchberg-Saxton method [Ger72]. An overview of our processing pipeline is shown in Figure 3.1. We initialize the pupil

function with a random phase and amplitude function. Within a single iteration, first the pupil function amplitude is set to be uniform within the pupil and zero outside. And the phase function remains as initialization. Then the wavefront is propagated to the image plane applying the Fraunhofer diffraction theory by Fourier transformation. At the image plane, the result of this transformation is calculated. We calculate the target amplitude function from the intensity pattern of the target PSF and subsititute the calculated amplitude function with the target. The phase term of the wavefront distribution remains the same. Then it propagates back to the pupil plane by inverse Fraunhofer propagation through inverse Fourier transformation. After multiple iterations, we obtain the phase distribution of the pupil function. The algorithm error decreases due to Parseval's theorem.

We found both in the simulation and experiment that the generated single-phase pattern does not produce a smooth PSF pattern. The main reason for this is that the iteration algorithm only finds approximated results with speckles. To alleviate the abovementioned effects, we generate an entire sequence of phase frames. In particular, we generate 60 frames of phase images with different initialized random pupil functions and combine them into a one-second video featuring a frame rate of 60Hz. The video is then played back on the SLM and repeated during image capture.

3.4 Experimental Setup

To achieve PSF modulation, we use a liquid crystal on silicon PLUTO-VIS-016-HR by HOLOEYE. It consists of a 1920×1080 refractive liquid crystal cells on a silicon reflective layer with 8 µm cell pitch. A linear polarizer must be added in front of the SLM, because phase-only modulation applies only for polarized light. The maximum modulation range of the SLM is 6.7π at 532 nm. The SLM operates as a display device employing a frame rate of 60Hz. By displaying an 8-bit grayscale image, the refractive index of each cell is changed by digital pulse code addressing. Each gray value of the displayed image relates to an addressing voltage, which drives the molecular orientation of the liquid crystal cell to switch its refractive index. We use a voltage-to-phase mapping for gamma correction. The limitation of the SLM is the addressing scheme. It is pulse code-based and hence produces temporal discrete electronic pulses to accumulate a certain averaged Z orientation of the parallel-aligned liquid crystal molecules. Therefore, a phase-jittering effect exists due to timing and orientation mismatch between different molecular orientations.

We designed a simple imaging setup having the SLM placed in front of a telephoto lens. We set the object at the far field, and use the small field of view of the telephoto lens to make the non-uniform phase coding negligible.

Our experimental setup is illustrated in Figure 3.2. We put an artificial star with an aperture diameter of 70 µm at a distance of 1.72 m from the SLM surface to fulfill the far field condition of the Fraunhofer approximation [Goo05]. We place a spectral filter

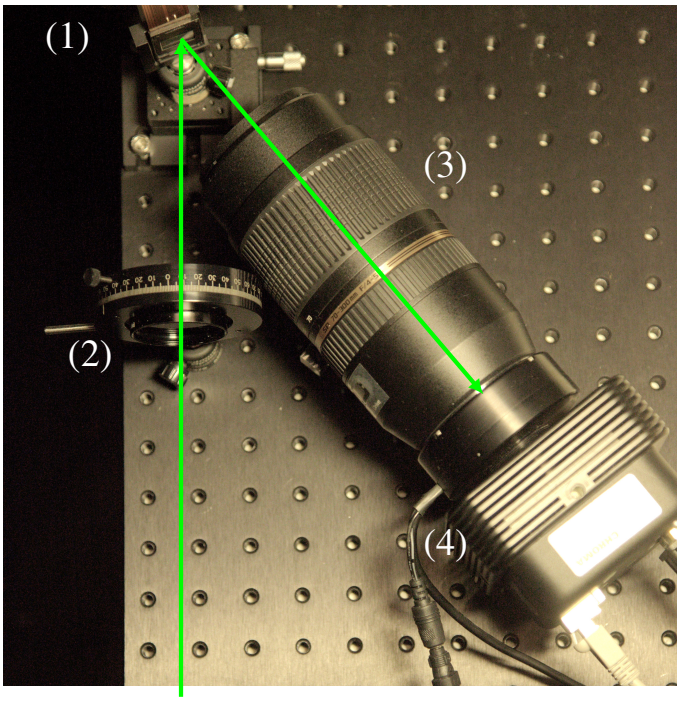


Figure 3.2: Imaging setup with SLM and a telephoto lens. (1) is the SLM. (2) is a linear polarizer. (3) is a telephoto lens. (4) is the camera body.

in front of the star aperture to produce monochromatic illumination. The spectral filter is a VariSpec VIS-07-35. It has a 35 mm aperture and 33 color bands from 400 nm to 720 nm with a step-size of 10 nm and a bandwidth of 7 nm. In our experiments, we used the 550 nm band. A linear polarizer is then set between the star target and the SLM surface to enable phase-only modulation. The modulated light is captured by a camera with a telephoto lens. We use a Tamron 70-300 mm $f/4$-5.6 zoom lens. We employ a DVC4000C camera with a Bayer color filter array for image capturing. We pick the focal length 125 mm to magnify the spiral PSF with a width less than 100 pixel. Finally, we capture all images with an HDR pipeline following Granados *et al.* [GAW+10] which uses an optimal weighting function under the assumption of compound-Gaussian noise.

3.5 Results

We now demonstrate results for generating PSFs of various shapes and geometry. We show the optimized phase pattern and illustrate the limitation of displaying a static phase image rather than playing back an entire sequence of phase images. We also include a robustness analysis by comparing the PSFs obtained by using phase videos at different frame rates. We demonstrate a refocusing application with this system.

3.5.1 Captured PSFs and Optimized Phase Pattern

In Figure 3.3 (a)-(e), we show various images of PSFs generated through phase modulation with our imaging setup. Note the large variety and the complex shapes that can be realized with our setup. All of these images have been logarithmically scaled for better visibility. The prominent central peak is caused by zeroth order diffraction of the SLM cell array and the non-modulated reflection. If necessary, it could be removed by subtracting an image with neutral phase pattern. The vertical strikes are caused by the diffraction of the non-modulating reflective SLM frame. Figure 3.3 (g) shows an optimized phase pattern for a spiral PSF obtained by the Gerchberg-Saxton algorithm. In Figure 3.3 (f), we present an image of a real scene with Lego bricks captured with the spiral-shaped PSF depicted in Figure 3.3 (e). The Lego scene is monochromatically illuminated by the same color filter illuminated from a broadband light source filtered by the same color filter that has been used for PSF capture. At careful inspection, one can read off the spiral PSF at isolated highlights.

We can also vary the PSF pattern in a sequence of image capturing. We present an animation of an animated spiral PSFs in Figure 3.4[1]. We produce a phase video for generating each frame of the PSFs. One can observe the smoothly shrinking sequence of the PSFs.

3.5.2 PSF Stability Analysis

Experimentally, we observed that the captured image of a generated PSF exhibits unwanted discontinuities as shown in Figure 3.5 (a). These discontinuities are highly dependent on the initial phase function of the pupil function. Therefore, one can increase the PSF stability by averaging the energy distribution caused by the phase modulation. Deploying the advantage of the SLM to play back videos, we generate a sequence of images to produce phase images produced by varying the initialization. During image capture, we repeatedly play back the video on the SLM. The resulting PSF can be thought of as a temporal average over the individual played-back video frames. Figure 3.5 (b) shows the PSF obtained by playback of a phase video containing 60 frames over a duration of one second. Note the much improved smoothness in appearance through temporal averaging.

To analyze the stability of this procedure, we created multiple phase videos of different frame rates. We captured PSF images and compared the PSF smoothness. To this end, we chose the 60fps phase video to be the ground truth PSF, which is depicted in Figure 3.5 (b). We compute the mean square error between the ground truth PSF and the resulting PSFs obtained by phase videos with a varying frame rate of 5fps, 10fps, 15fps, ..., 55fps. Integration time was one second in all cases. We show the result of this comparison in Figure 3.5 (c). The discontinuities decrease quickly with an increased frame rate and level off at a frame rate of 30fps.

[1] Please use Adobe Acrobat Reader for display of the animated image.

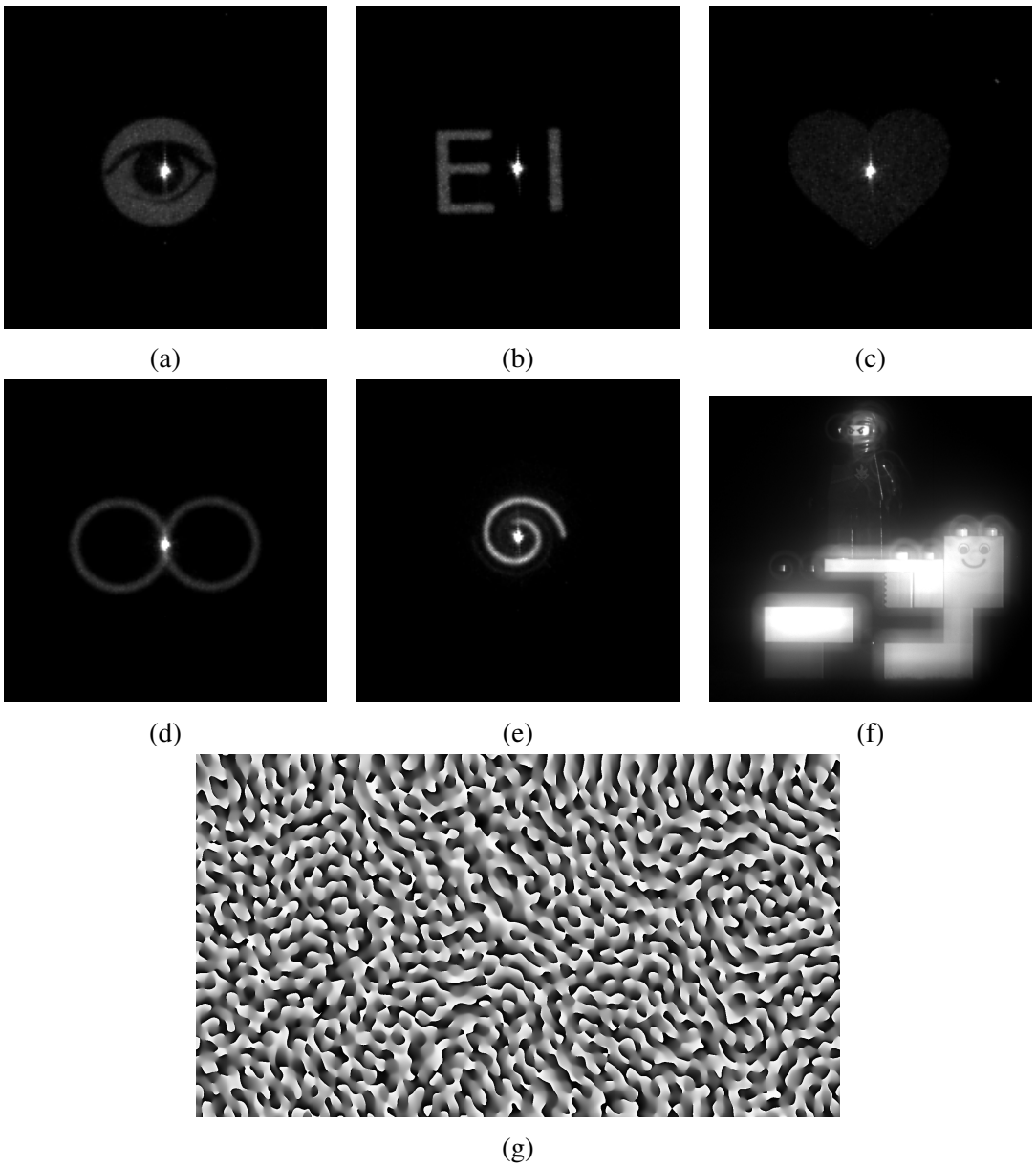


Figure 3.3: Captured PSFs. (a) to (e) are pre-designed PSFs: eye, EI logo, heart, infinity symbol, spiral. (f) is the real scene captured with the spiral PSFs in (e). (g) is the optimized phase pattern for the spiral PSF. Note the spatial discontinuities.

3.5.3 Refocusing

By modulating the phase term of an optical wavefront, its propagation will be redirected. We reduce a quadratic phase transformation function modulo the phase modulation range to get a Fresnel lens phase pattern as shown in Figure 3.6 (c). By adding a Fresnel

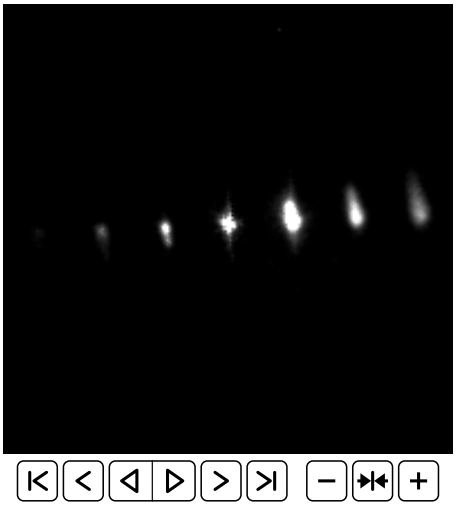


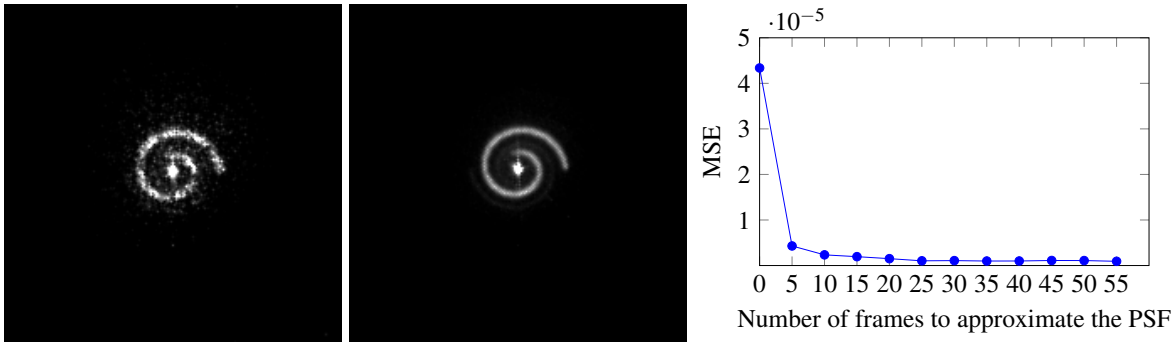
Figure 3.4: Time-varying PSF in form of a animated spiral. Please use a PDF viewer that allows for animated image playback such as Adobe Acrobat Reader.

lens phase pattern, the incoming wave is converged to a different focal plane from its original focal plane. In this experiment, we compare the images captured by displaying an uniform grayscale image and a Fresnel lens grayscale image on the SLM.

In practice, our refocussing technique is demonstrated in Figure 3.6. Three Lego bricks are placed at three different distances from the SLM. The camera is manually focused on the farthest object. We encode a uniform level grayscale image on the SLM to reflect the incoming light without any modulation. Figure 3.6 (a) shows a captured image by encoding an uniform zero level gray image to the SLM. We then encode a Fresnel lens on the SLM to shift the focal plane to the Lego figure. We show the result in Figure 3.6 (b). One can observe, especially from the arm and the head of the Lego figure, a desired high frequency edge is reproduced by refocusing. One can also observe a halo effect on each object. This is caused by the residual non-modulated light.

3.6 Discussion and Future Work

The main limitation of our current technique is twofold: the PSF contains speckles and a central peak. Smooth PSF patterns can be achieved by encoding a phase video to perform temporal averaging, of course at the cost of additional exposure time that is needed for image capture. The central peak of the PSFs are caused by the non-modulated reflection, zeroth order diffraction. The observed speckles are produced due to existence of inteference. However, the temporal cost can be reduced by using the residual PSF as the target to compensate the PSF discontinuities in the future. In the future, one could try to obtain a smoother SLM pattern by exploiting transport theory as done for the design of caustics pattern [STTP14]. The central peak response can be eliminated by capturing



(a) (b) (c)

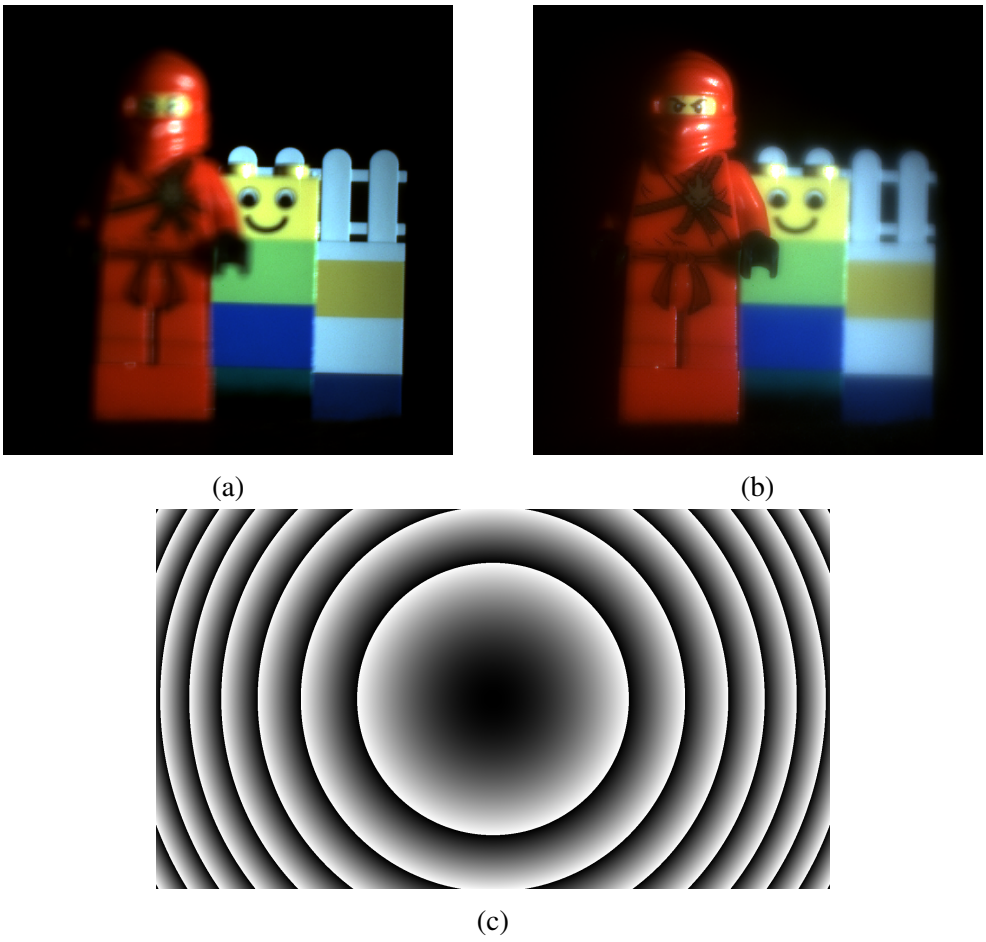
Figure 3.5: PSF stability analysis. (a) shows the PSF when displaying a single static phase on the SLM. (b) demonstrates the PSF when averaged over 60 phase patterns displayed on the SLM. (c) is the PSF reconstruction error vs number of temporally averaged frames.

PSFs by encoding uniform level phase images onto the SLM and with a subtraction thereafter. The possibility of generating PSFs of user-specified geometry might result in novel deconvolution applications such as spatial-spectral multiplexing or motion blur removal. Even light field applications could be generated by encoding micro-lens array patterns on the SLM.

3.7 Conclusion

We introduced a novel phase-coded aperture technique that allows the generation of user-specified PSF geometry on a photographic camera. Using the standard Gerchberg-Saxton method, we produced 8-bit grayscale images to encode the phase pattern on the SLM to realize phase modulation. The resulting system allows for full control of the PSF even over time. This ability might enable a number of interesting novel imaging applications. We demonstrated its practical use in the case of digital refocusing. The possibility of generating PSFs of user-specified geometry might result in novel deconvolution applications such as spatial-spectral multiplexing or motion blur removal. Even light-field applications could be generated by encoding micro-lens array patterns on the SLM.

This work has laid a foundation for the future research on phase-coded aperture. This sets the stage for the next step to investigate single-shot spectral imaging by PSF design and correlated inverse problem solving. The flexibility of user-defined PSFs will be extended to comprise a wider chromatic range. A tailored optimizer will encompass the knowledge of image formation to restore spectral image cubes from snapshots.



(a) (b)

(c)

Figure 3.6: Refocusing by implementing a Fresnel lens phase pattern. (a) is the defocused Lego figure. (b) is the refocused Lego figure. (c) is the phase pattern of a Fresnel lens.

Chapter 4

Joint Design of PSFs and Image Processing for Multispectral Imaging from Single Shot

The material of this chapter is based on the following publication:

[CHEL18] Jieen Chen, Michael Hirsch, Rainer Heintzmann, Bernhard Eberhardt, and Hendrik Lensch. A phase-coded aperture camera with programmable optics. *Electronic Imaging*, 2017(17):70–75, 2017.

Spectral imaging has many uses in the field of conservation of cultural heritage, and medical imaging. It collects spectral information at each location of an image plane as an image cube. Among various approaches, snapshot multispectral imaging techniques measure the cube within one integration period. Previous work has addressed the issue of optical design, while recent developments have shifted the focus towards computation. In this chapter, we present a snapshot multispectral imaging technique with a computational camera and a corresponding image restoration algorithm. The main characteristics are: (1) transferring spectral information to the spatial domain by engineering user-defined PSFs; (2) measuring spectral images by computationally inverting the image formation. The design of our computational camera is based on a phase-coded aperture technique to generate spatial and spectral variant PSFs. The corresponding algorithm is designed by adapting single-channel and cross-channel priors. We show experimentally the viability of our technique: it reconstructs high-resolution multispectral images from a snapshot. We further validate that the role of PSF design is critical.

4.1 Introduction

Computational cameras use controllable optical systems followed by computational decoding to produce new types of images. Among computational photography applications, spectral imaging is a branch that captures the spectra at each location of the image plane as a 3D dataset. Conceptually, there are two approaches to acquire multispec-

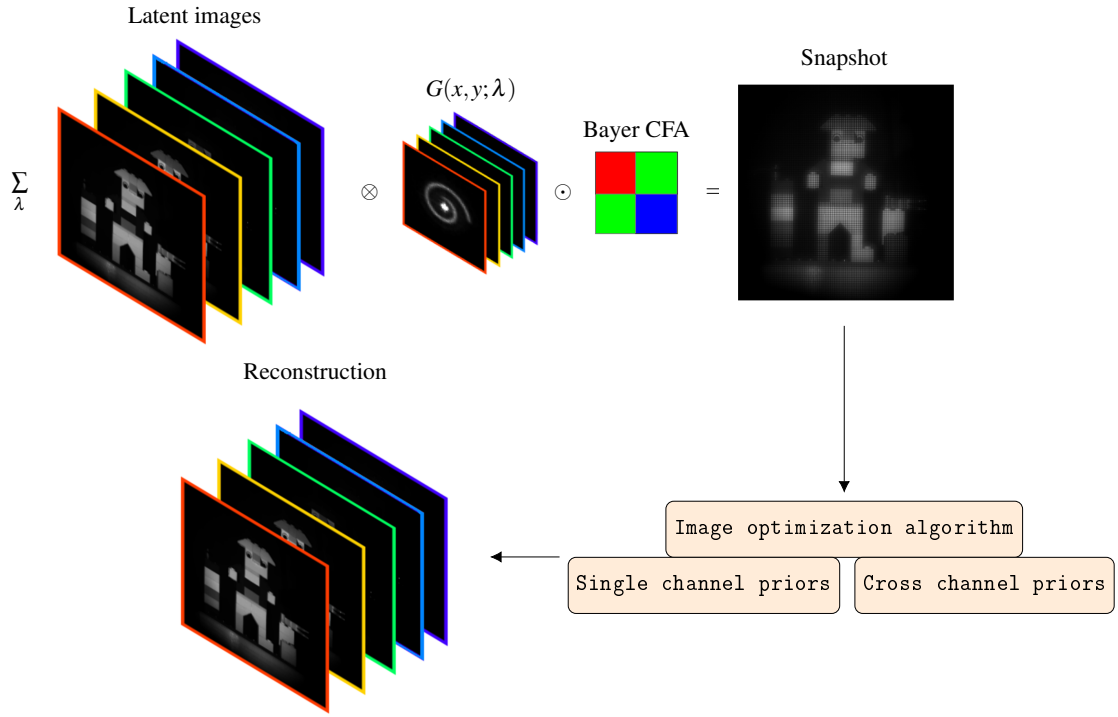


Figure 4.1: Pipeline overview. PSF engineering by computational camera and reconstruction of multispectral images.

tral data: scanning and snapshot imaging. Scanning spectral imagers measure time-sequential 2D slices, e.g. using color filters. Snapshot spectral imagers measure all elements of the 3D dataset simultaneously and decode them in postprocessing. However, the optical design of snapshot imagers is typically of a rather high complexity in order to boost light collection capacity while guaranteeing a specific reconstruction quality. Various astronomical or biomedical applications [HK13] employ complicated setups with mirrors, fiber arrays, beam splitters, and multiple color filters.

Recent advances in computational imaging have shifted the workload from optics to algorithms. Image optimization algorithms have been explored for HDR, denoising, demosaicing, deconvolution, and multispectral imaging. In particular, deconvolution is related to optical design through Point Spread Function (PSF) engineering. As the fingerprint of an imaging system, the PSF is the spatial response of a point light source. We address the problem of snapshot multispectral imaging by multiplexing spectral information to the spatial domain through wavelength-dependent PSFs. A linear disperser distributes colors along one dimension, which produces an overlap of spectra of neighboring pixels. Without a strong regularizer, this leads to error-prone restoration. However, by generating PSFs with two-dimensional color dispersion, the spectral information can be converted into a spatial code.

Computational cameras with wavelength-dependent PSFs for multispectral imaging have been developed in the past few years [WSVM16, SWB+16]. Current computational cameras do not provide flexibility to tune the spatial and spectral distribution of the PSFs.

In this chapter, we aim to overcome the limitations of snapshot multispectral imaging by introducing a computational technique that combines a programmable optics device and a computational reconstruction pipeline. We employ an SLM as the programmable optics device to generate user-defined spatially and spectrally variant PSFs. Phase patterns are generated by a standard phase retrieval algorithm to encode the pupil function. An optimization pipeline is then implemented with TV, L2, and a cross-channel regularizer. An overview of our technique is shown in Figure 4.1. The cross-channel regularizer enforces elimination of color fringing and properly locates edges across spectral bands. Our technique provides flexibility as a platform to computationally tune the PSF both spatially and spectrally. We examine the significance of appropriate PSF design in multispectral reconstruction. The technical contributions are as follows:

- Generating spatially and spectrally variant PSFs with a computational camera consisting of an off-the-shelf camera and a programmable optical device.
- A corresponding multispectral image reconstruction technique with Sobolev, TV, and cross-channel regularizers.

4.2 Related Work

Snapshot multispectral imaging is a technique to capture fine color spectrum information for each image pixel within a single shot. Compared to the scanning imaging spectrometer architectures, such as using a tunable filter camera [PA01], snapshot multispectral imaging allows light collection within a single integration time. An informative survey on snapshot multispectral imaging is presented by Hagen and Kudenov [HK13]. One of the common shortcomings in this area is the high setup complexity of the imaging system. Integral field spectrometry [Bow38] uses prisms or glass plates to slice the optical beam into a long slit with multiple spectral images.

Coded Aperture Snapshot Spectral Imager (CASSI) [WJWB08] replaces the entrance slit with a coded aperture in order to measure the multispectral data cube. It takes advantage of compressive sensing theory to reconstruct data termed to be insufficiently sampled by the Nyquist limit. An over-complete dictionary learning for sparse reconstruction is presented by Lin *et al.* [LLWD14] based on CASSI. Wang *et al.* [WXG+15] investigated a dual-camera system constructed with one low frame rate CASSI camera and a panchromatic high frame rate camera to capture high-speed multispectral video. The method suffers from the complexity of an imaging setup with a prism, as well as the attenuation caused by the amplitude-coded aperture design. In contrast, our approach uses a phase-coded aperture setup for flexible design of PSFs which avoids this issue. A diffractive filter can also be produced following the experimental phase profile.

A multi-aperture filtered camera [HIII94] measures the full spectral band using an array of imaging elements, such as a color-filtered lenslet array. Computational tomography-imaging spectrometry [OTY93, BV92] projects the spectral cube by a 2D dispersor at the aperture of the spectrometer. The main shortfall is the reduction of resolution due to angular projection. However, our approach enables full resolution reconstruction because of the convolution nature of PSF modulation.

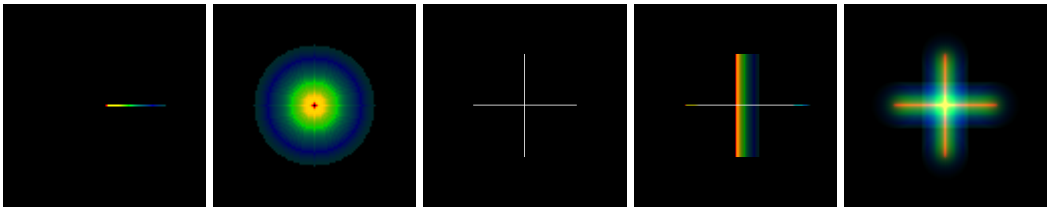
A compact snapshot hyperspectral imaging system is proposed by Baek *et al.* [BKGK17], which equips an ordinary prism with a DSLR camera. With linear dispersion produced by the prism, the spectral information is estimated from sparse dispersion information especially from edges. Our approach, instead of linear dispersion, enhances spectral encoding by 2D dispersion. Wang *et al.* [WSVM16] introduce a diffractive filter. The authors build a diffractive filter to generate spatially, spectrally variant PSFs. Our approach provides more flexibility using an SLM that is able to generate arbitrary user-defined PSFs. Another technique that encodes color in the image by exploiting chromatic dispersion through a design of new phase masks is proposed by Shechtman *et al.* [SWB$^+$16]. The phase masks produce controllable PSFs for different wavelengths. While this approach provides a multispectral imaging solution in microscopy, our method applies in the photography domain. Chen *et al.* [CHH$^+$17] present a phase coded-aperture setup with programmable optics to control PSFs and refocus. In this chapter, we employ the phase coded-aperture setup to generate spectrally, spatially variant PSFs.

Image optimization finds the solution with minimal energy using optimization algorithms. Examples include both blind and non-blind deconvolution, demosaicing, image denoising, and inpainting. It is demonstrated by Heide *et al.* [HST$^+$14] that a subset of low-level image processing problems can be solved through a single framework. This is presented by Heide *et al.* [HDN$^+$16] as a domain-specific language and compiler for image optimization problems using proximal operators as fundamental building blocks that make it easy to experiment with different problem formulations and algorithm choices. The proximal operator of the regularization [MMHC17] can be replaced by a denoising neural network to solve image deconvolution and demosaicing problems. In this research, we employ demosaicing and non-blind deconvolution as a global optimization problem to reconstruct multispectral images.

4.3 Multispectral Imaging with PSF Engineering

The principal advantage of our approach is the ability to measure spectral images from a snapshot with spectral information multiplexed in the spatial domain. The key ingredient is to modulate spectral-dependent PSFs that produce 2D dispersion.

This solution improves on traditional methods using linear gratings or prisms. An example of a comparison with a traditional prism is illustrated in Figure 4.2. The PSFs of a linear dispersor smear across a fixed direction (Figure 4.2(a)(c)). In particular, in spectral imaging, the spectrum is overlapped along the dispersion direction. However, with a 2D



(a) (b) (c) (d) (e)

Figure 4.2: Comparison of dispersion of prism and our approach. (a) is the linear dispersion PSF. (b) is a ring-shaped PSF. (c) is a cross target without dispersion. (d) is the cross target with a linear dispersion. (e) is the cross target with the ring-shaped PSF dispersion.

dispersion, each spectrum has a unique spatial distribution as a fingerprint. This is a cue for spectral image reconstruction. The phase images are optimized regarding one narrow band to generate PSFs with user-defined spatial distribution. The PSF-engineering becomes complete when the dispersive nature of the SLM helps to produce PSFs with different sizes for varying wavelengths. Once PSFs are measured, a reconstruction algorithm is modelled to invert the image formation.

4.3.1 Image Formation Model

The incoherent imaging process is modeled as the convolution of the latent image and the intensity PSF [Goo05, PFHH16]. The PSFs are modulated in a wavelength-dependent fashion. Since a conventional RGB camera is used, the camera spectral response is also considered. Our spectral image formation is therefore modeled as,

$$O(x,y;c) = \int_{\lambda} S(x,y;c,\lambda)[I(x,y;\lambda) \otimes G(x,y;\lambda)]\, d\lambda + N(x,y) \tag{4.1}$$

where $\otimes$ represents convolution; $I(x,y;\lambda)$ is the latent spectral image at wavelength λ; $G(x,y;\lambda)$ is the intensity PSF; $S(x,y;c,\lambda)$ is the spectral response of each RGB pixel of the camera sensor ($c \in R,G,B$); and $N(x,y)$ is the image noise. $O(x,y;c)$ is the captured image, whose channel c is defined by the layout of the color filter array. Each spectral image is firstly convolved by its PSF and then filtered by mosaic patterns on the camera sensor.

4.3.2 Generating Spatially and Spectrally Variant PSFs with Programmable Optics

We introduce the design of spatial and spectral variant PSFs by the computational camera. The phase image is optimized under monochromatic conditions using the Gerchberg-Saxton algorithm [Ger72, CHH$^+$17] to produce PSFs with user-defined spatial distribu-

tion.

The SLM phase modulation varies with different wavelengths. Liquid crystal cells have a dispersion property that is caused by refractive indices of different wavelengths. With the SLM used in our setup, the phase shift ability is as follows: 633 nm has 5.4π, 532 nm has 6.7π and 452 nm has 9.0π. By exploiting the use of this property, we are able to generate spectral PSFs with different sizes. In Figure 4.4, spatial and spectral variant PSFs are presented.

4.3.3 Design Spatial Distribution and Phase Profile of PSFs

We take advantage of the Fourier relation between the pupil function and the PSFs, which assumes that the object is located at the far field. By the Gerchberg-Saxton phase retrieval algorithm [Ger72], the phase images for encoding programmable optics can be optimized. The spatial distribution of the PSF is defined by the user.

The PSF design problem can be phrased in terms of the strength of 2D dispersion, as well as the frequency coverage. This gives intuitive design cues for PSFs to be shapes including spiral, spread dots, and triangles. Having the designed spatial distributions, we can optimize the phase profile to produce PSFs using a standard phase retrieval algorithm. Several example PSFs are shown in Figure 4.4. The example of a ring-shaped PSF is shown in Figure 4.3a. Its phase image is shown in Figure 4.3b. Due to the limited modulation depth, the optimized phase pattern has to be folded. Six monochromatic PSFs are shown in Figure 4.3. One can clearly observe the PSF size variation with wavelength.

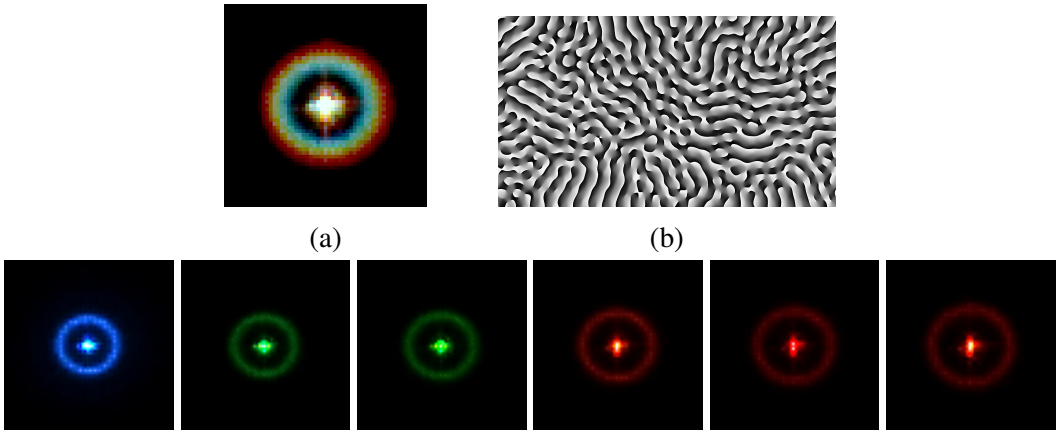
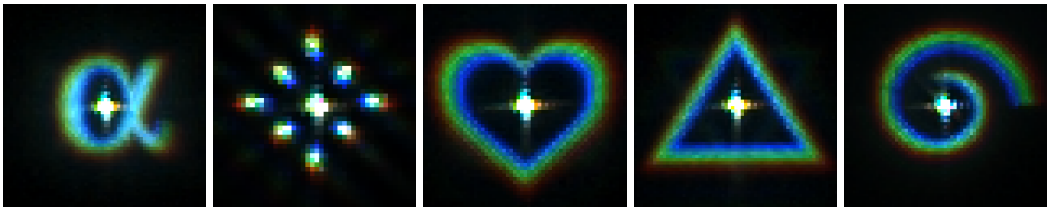


Figure 4.3: Example of user-defined ring PSFs. Note the varying size for the different wavelengths. (a) is the color ring PSF. (b) is the phase pattern of ring PSFs. (c) to (h) are the monochromatic PSFs (colored for display) in 450 nm, 500 nm, 550 nm, 600 nm, 650 nm, and 700 nm.



(a) (b) (c) (d) (e)

Figure 4.4: Spatial and spectral variant PSFs captured by the computational camera. (a) to (e) are various pre-designed spatial distributions of PSFs.

4.4 Reconstruction of Spectral Images

In order to obtain multispectral images from a single PSF-modulated image, we introduce an inverse process with the knowledge of PSFs and the camera spectral response function. This problem can be outlined in terms of the image formation model. Combining our prior knowledge of hue consistency and image smoothness, we formulate an objective function with data fitting, single-channel, and cross-channel regularizers.

4.4.1 Image Formation Operation

Consider the multispectral latent images with resolution $X \times Y \times \Lambda$. The image formation can be discretized from Equation (4.1) as,

$$\mathbf{O} = \sum_{n=1}^{N} \mathbf{S}_n \mathbf{G}_n \mathbf{I}_n + \mathbf{N} \tag{4.2}$$

where $\mathbf{S}_n$ is the Bayer mosaicing operator at band n, $\mathbf{G}_n$ is a wavelength dependent convolution kernel, and N is the total number of spectral bands. $\mathbf{O}$ is the single shot image captured by the computational camera with modulated PSFs. $\mathbf{I}_n$ is the latent spectral image at band n.

4.4.2 Reconstruction of Spectral Information

Our algorithm seeks the solution of the following objective function,

$$f = \arg\min_{\mathbf{I}_n} \| \sum_{n=1}^{N} \mathbf{S}_n \mathbf{G}_n \mathbf{I}_n - \mathbf{O} \|_2^2 + \Gamma_n(\mathbf{I}_n) + \Gamma_c(\mathbf{I}_n) \tag{4.3}$$

where the first term is a standard least-square data fitting term. The second and third terms are regularization terms where image priors are applied. We employ a non-blind deconvolution scheme with the prior knowledge of the modulated PSFs.

Although unique PSFs are produced for each spectral band, the inverse problem is highly ill-posed because of the multiplexing nature of the image formation. One advantage is the existence of the zeroth order diffraction–central peak in the PSFs, which preserves some amount of image edges. We enforce intraband and interband prior knowledge for spectral reconstruction. Through the use of single-channel image priors, homogeneous reconstruction with edges is enforced intraband. In addition, edge information is shared across different color bands through a cross-channel prior.

Two priors are chosen in order to recover intraband information: the total variation (TV) and Sobolev. The TV prior is capable of recovering blocky images from noisy data. The Sobolev prior is a quadratic term that preserves uniform smoothness. The single-channel regularization is thus formatted as,

$$\Gamma_n(\mathbf{I}_n) = \sum_{n=1}^{N} \alpha \|\nabla \mathbf{I}_n\|_2^2 + \beta \|\nabla \mathbf{I}_n\|_1^1 \tag{4.4}$$

where α and β are weights of the priors.

The key part of regularization is a cross-channel prior that borrows edge information from other color channels to benefit the reconstruction. By producing PSFs with their size variation with wavelengths, we intentionally increase the chromatic aberration of the optical system, while traditionally, lens designers minimize it. During reconstruction, having cross-channel priors shares information between the individual reconstructed spectral images. We use the cross-channel regularizer employed by Heide *et al.* [HRH+13], which is based on the assumption that hue remains constant along edges. In other words, edges share the same location across all channels. This is formulated in an L1 fashion as follows,

$$\nabla \mathbf{I}_k \cdot \mathbf{I}_l \approx \nabla \mathbf{I}_l \cdot \mathbf{I}_k \tag{4.5}$$

where l and k represent two spectral bands. The cross-channel regularizer is described as,

$$\Gamma_c(\mathbf{I}_n) = \sum_{n=1}^{N} \sum_{n \neq i} \sigma_{ni} \|\nabla \mathbf{I}_n \cdot \mathbf{I}_i - \nabla \mathbf{I}_i \cdot \mathbf{I}_n\|_1 \tag{4.6}$$

where σ_{ni} is the weight for the prior. Having a strong cross-channel regularizer enables us to solve this highly ill-posed problem, because it emphasizes the inverse algorithm to search for regions that have constant hues, and rejects strong chromatic aberrations. However, we avoid using regularizer based on assumption of spectral smoothness.

4.5 Experimental Results and Discussion

To capture PSF-modulated snapshots, we build a computational camera with reference to Chen *et al.* [CHH+17]. Our computational camera shown in Figure 4.5 is built in a phase coded-aperture fashion with an LCoS SLM.

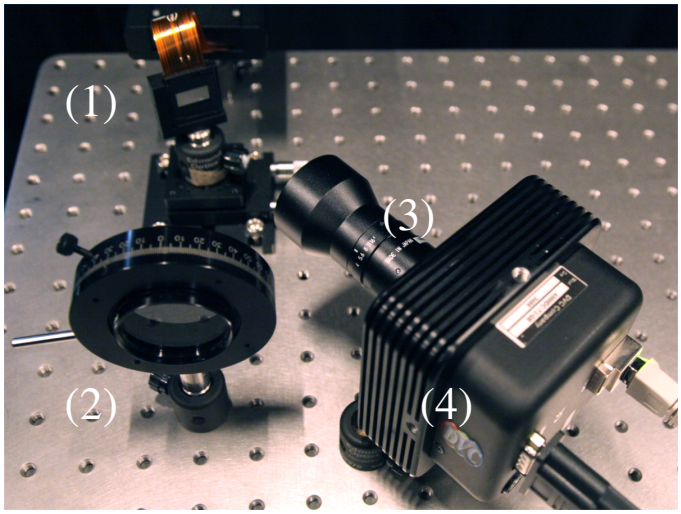


Figure 4.5: Computational camera setup with a telephoto lens and SLM. (1) is the LCoS SLM. (2) is a linear polarizer to assure phase modulation. (3) is a telephoto lens. (4) is the camera body.

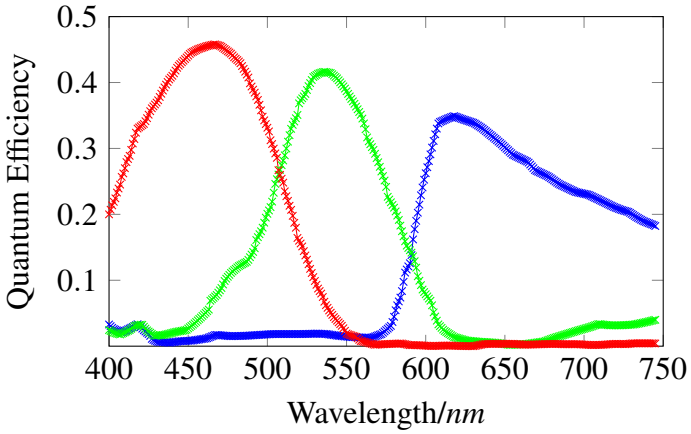


Figure 4.6: Camera spectral response function of the red, green, and blue pixels.

A 100 mm compact fixed focal lens is mounted on a conventional RGB camera. The small field of view (FoV) of the telephoto lens makes the non-uniform phase coding caused by magnification negligible. The programmable optics we use in the setup is the PLUTO VIS-006-A (420 - 700 nm) HR version SLM by HOLOEYE, which offers 87% fill factor. All images are captured with a DVC4000C camera whose sensor has a Bayer color filter array. An HDR pipeline is selected for image acquisition. We insert a tunable VariSpec color filter (400 - 720 nm, bandwidth 10 nm) in front of a halogen light source Osram 64655 HLX to produce monochromatic illumination.

We measure the ground truth dataset by encoding a uniform gray image onto the SLM. Each scene is illuminated monochromatically while capturing the ground truth data. Similarly, monochromatic PSFs are measured by capturing images of a pinhole light source. A debayering algorithm is used to find the optimum of both the ground truth and PSF estimation. The optimization is formulated as,

$$f = \arg\min_{\mathbf{I}_{pn}} \|\mathbf{S}_n\mathbf{I}_{pn} - \mathbf{I}_{cpn}\|_2^2 + \gamma\|\nabla\mathbf{I}_{pn}\|_2^2 \tag{4.7}$$

where $\mathbf{I}_{cpn}$ is the captured PSF image at band n, $\mathbf{I}_{pn}$ is the latent PSF image, and γ is the weight for the Sobolev regularizer. We access the camera spectral response function $\mathbf{S}_n$ from datasets provided by vendor shown in Figure 4.6.

Due to our filtered monochromatic illumination, PSF-modulated snapshots and ground truth images are captured under different illumination spectra. In order to validate the multispectral images with the ground truths, we measure the scaling factor of the VariSpec filter using a Konica-Minolta CS2000A spectroradiometer.

Validation. The performance of our method is validated using both synthetic and real-world data with ground truth. For the synthetic images, we use the multispectral database by Yasuma *et al.* [YMIN10]. In Figure 4.7, we show our results using a flower scene. The PSFs are designed to have 2D dispersion in ring shapes as are shown in Figure 4.7m. We opted for ring-shaped PSFs on the basis of their symmetric sampling. Six channels of PSFs are used during testing. The results show that without loss of resolution, our inverse approach is able to reconstruct all six multispectral images. The peak signal to noise ratio (PSNR) of each image are 27.6, 28.9, 28.6, 28.4, 26.7 and 24.9 dB. Results show degrees of blur due to the paramter choices of the priors.

More interestingly, we show the real-world implementation. Multiple scenes are captured that contain a ColorChecker, plastic, organic and fabric materials. Three different designs of PSFs–ring, dots and sprial, are used and encoded with our computational camera in accordance with the optimized phase profiles. We successfully reconstruct six spectral bands with full resolution. Consequently, single RGB images are restored using these spectral images. We show our experimental results of a ColorChecker, when the ring PSFs are modulated in Figure 4.8. The individual spectral images closely match the ground truth data. The image contrast within each band is faithfully restored. In the

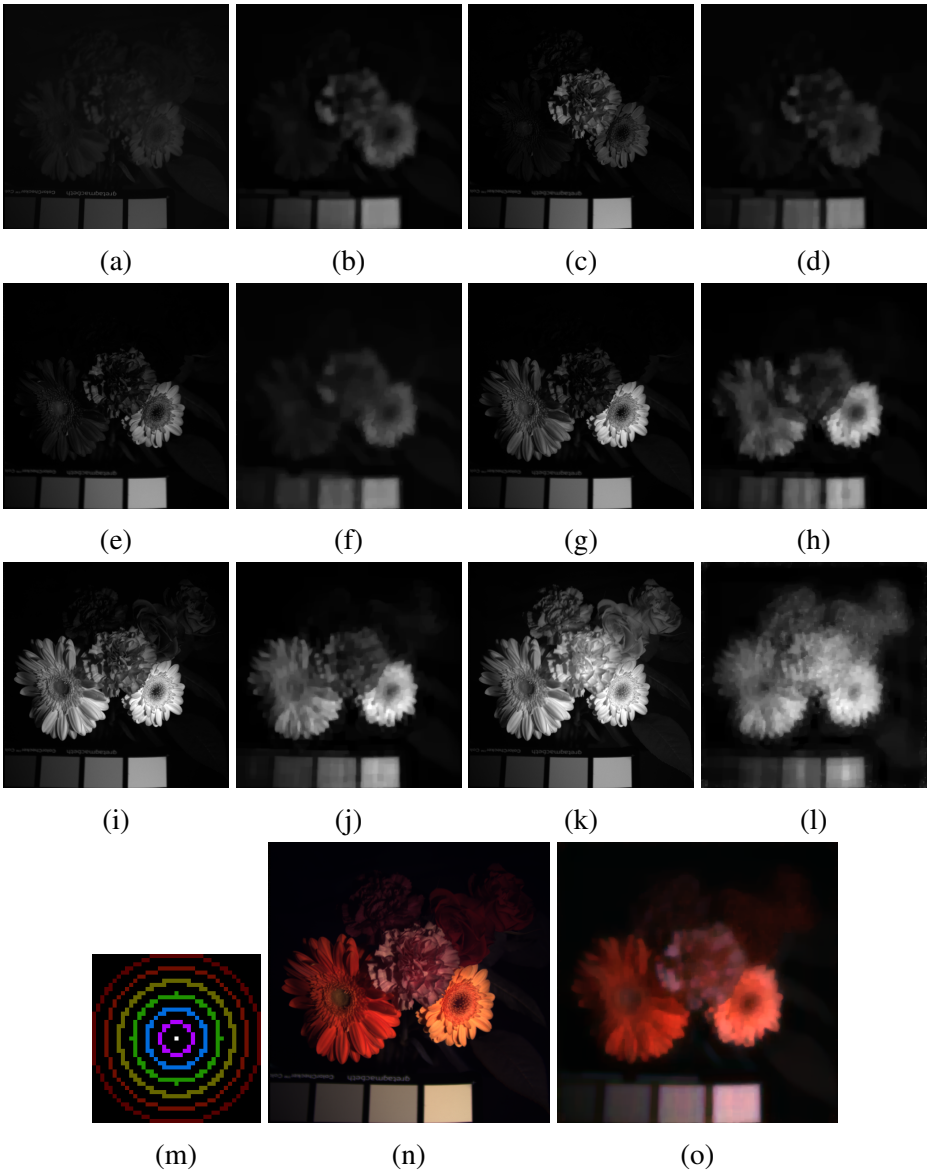


Figure 4.7: Synthetic results comparison using CAVE multispectral data. GT and RS stand for ground truth and results. The image contents are as follows, (a) 450 nm GT, (b) 450 nm RS, (c) 500 nm GT, (d) 500 nm RS, (e) 550 nm GT, (f) 550 nm RS, (g) 600 nm GT, (h) 600 nm RS, (i) 650 nm GT, (j) 650 nm RS, (k) 700 nm GT, (l) 700 nm RS, (m) ring PSFs, (n) RGB ground truth, and (o) snapshot with ring PSFs.

color images, the patches show clear spectral proximity to the ground truth. Please zoom in to see more details. We further analyze reconstructions of images with other scenes. The RGB image results are displayed in Figure 4.9. These tests reveal the reconstruction stabiliy of our approach. Even with an organic object that has a low image gradient and

homogeneous spectra, our method is still able to restore the spectral images. Inevitably, there are some discrepancies such as the highlight on the shoulder of the Lego figure due to limited number of PSF bands.

The image resolution of the ColorChecker, cloth, Lego and lemon are 805×805, 805×805, 800×800 and 804×804. It is worth noting that all images are restored without compromising resolution. Please refer to the supplementary materials for results of diverse scenes using different PSFs.

Table 4.1: PSNRs of real-world results

	ColorChecker			Lemon		
	Ring	Spiral	Dots	Ring	Spiral	Dots
Mean	21.5473	22.0312	22.2183	28.1851	21.7892	23.8086
	Lego			Cloth		
	Ring	Spiral	Dots	Ring	Spiral	Dots
Mean	22.9578	22.4835	21.3271	18.4300	19.0417	18.7232

We run our image reconstruction experiments, as well as ground truth and PSF estimation with the L-BFGS-B optimizer. We use the MATLAB implementation [Car14]. The real-world reconstruction parameters are empirically set to be $\alpha = 0.3 \times 10^{-2}$, $\beta = 0.8 \times 10^{-2}$, $\sigma_{ni} = 0.5$ and $\gamma = 1.0$. We employ strong cross-channel prior to reject chromatic aberrations, as well as strong Sobolev in PSF restoration for sharp spectral image reconstruction.

Comparison with reconstruction from linear dispersion. We further compare results by designed PSFs and linear dispersion. PSFs with linear dispersion is shown in Figure 4.11. A peak response is left in the center of the PSFs to mimic zeroth order diffraction. We synthesize the snapshot using these PSFs to avoid any noises during capturing. The ground truth ColorChecker spectral images are used for simulation. We show the reconstructed results in Figure 4.11. The results indicate that multispectral imaging with simple linear dispersion is severly ill-posed. It is hardly usable for recovering multispectral images even with the strong cross-channel prior. It is verified that a 2D design of PSFs is necessary.

Comparison of results with different PSFs. We show in Figure 4.10 the reconstructed results of a ColorChecker using different user-designed PSFs. Multiple captured PSFs in Figure 4.4 are used for this test. The analysis shows that PSFs with different spatial distributions generate discrete levels of performance. Ring PSFs yield the best results compared to PSFs shaped as dots or spirals. There is a positive correlation between the specific halo artifact and the PSF spatial distribution. For example, in Figure 4.10(c),

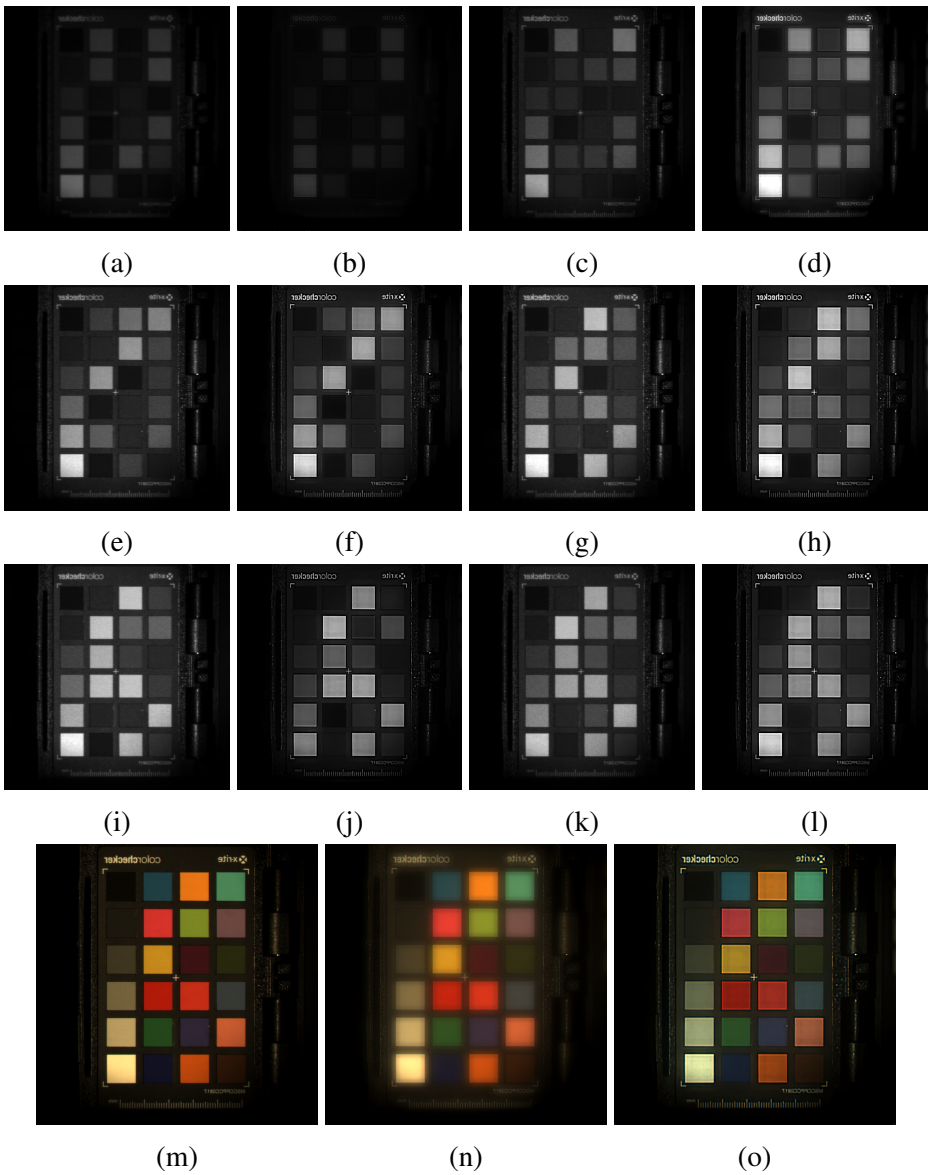


Figure 4.8: Real-world ColorChecker scene comparison of ground truth and reconstruction. GT and RS stand for ground truth and results. The image contents are as follows, (a) 450 nm GT, (b) 450 nm RS, (c) 500 nm GT, (d) 500 nm RS, (e) 550 nm GT, (f) 550 nm RS, (g) 600 nm GT, (h) 600 nm RS, (i) 650 nm GT, (j) 650 nm RS, (k) 700 nm GT, (l) 700 nm RS, (m) RGB single shot ground truth, (n) snapshot with ring PSFs, and (o) restored RGB single shot.

asymmetric halos appear on each patch due to the particular non-uniform sampling of spiral PSFs. Similarly, Figure 4.10(d) shows periodic halos surrounding each patch. This finding confirms that the PSF design is critical in snapshot multispectral imaging.

Noise analysis. We evaluate reconstruction on real data with different level of Gaussian white noises added to the single shot input in Table 4.2. The influence of noise is reduced by the knowledge of the PSFs and the employed regularizers, especially the cross-channel regularizer shares edge information across less determined channels.

Table 4.2: PSNRs of a real ColorChecker with ring PSFs results under noise

σ	0	0.01	0.05	0.1
Ring	22.3361	22.3747	17.7970	14.7672

Limitations. We acknowledge that our research may have two limitations. The first is that the number of spectral bands is still limited. It is possible that a trade-off exists between spatial and spectral resolution due to the ill-posed nature of this method. Careful investigation must be exercised regarding the maximal number of spectral bands. The second is the limited freedom in desiging 2D PSFs. Current PSF design is based on optimizing phase profile of a single monochromatic band. However, it will be beneficial to design PSFs of each band separately. SLM-based color-multiplexing technique may be an interesting avenue to explore this design.

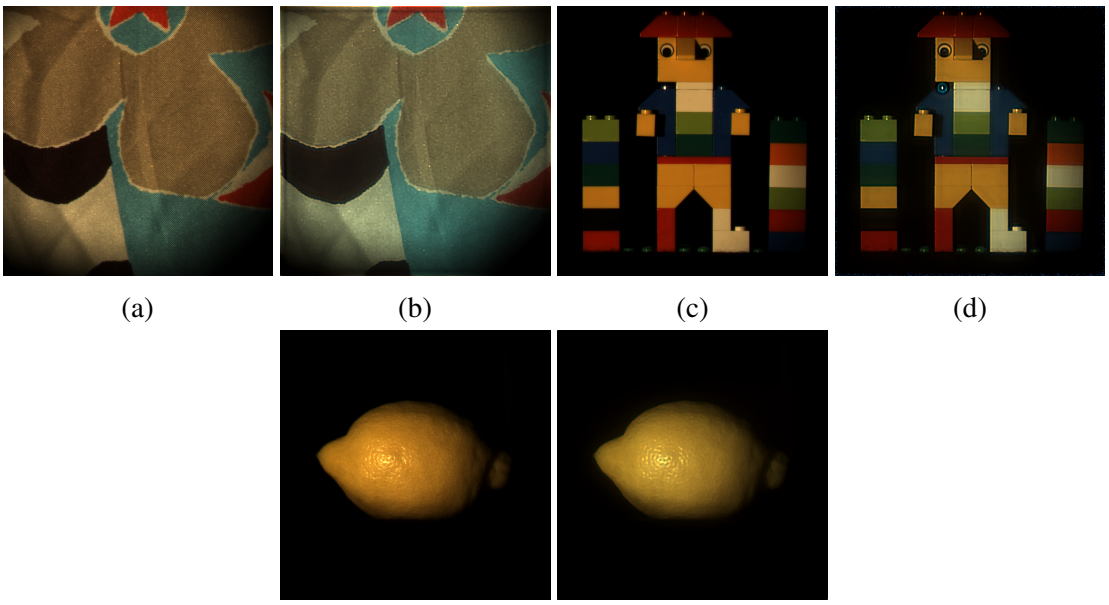
4.6 Conclusion

In this chapter, we presented a novel snapshot multispectral imaging technique with a computational camera, equiped with an SLM as the programmable optical device, enabling user-defined spatial and spectral variant PSFs. The 2D-dispersed PSFs thus multiplex spectral information to spatial domain. We have built up and calibrated the computational camera using off-the-shelf devices. A reconstruction strategy is also devised based on the inverse of the image formation model, adapting the single-channel and cross-channel priors to carefully locate edges. We have demonstrated multispectral results with six bands using both synthetic and real-world data without loss of image resolution. The strength of PSF design is highlighted by comparison of reconstruction using various PSFs.

The present study is limited by the number of reconstructed bands and the need for modulation of 2D dispersion. Further studies, which take into account optimal PSF design, will need to be undertaken. This approach could be applied to print a simple specialized optical filter with the optimal phase profile for the purpose of convenient low-cost snapshot multispectral imaging.

This work serves as a base for the combination of PSF design and reconstruction. The spectral information is transmitted through the hand-crafted PSFs to the digital end. The

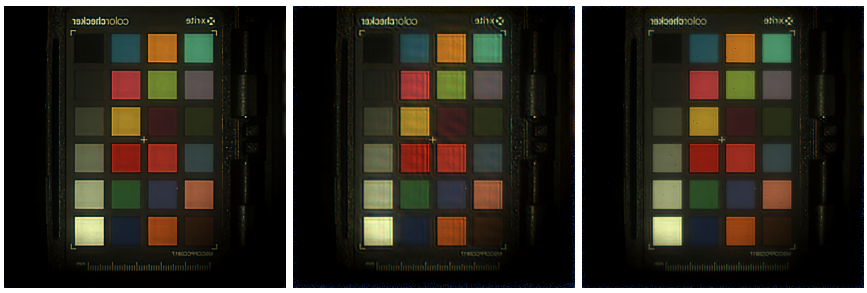
next sensible step is to optimize the PSFs using a deep learning framework by taking advantage of the image datasets. It is to this goal we turn in the next chapter.



(a) (b) (c) (d)

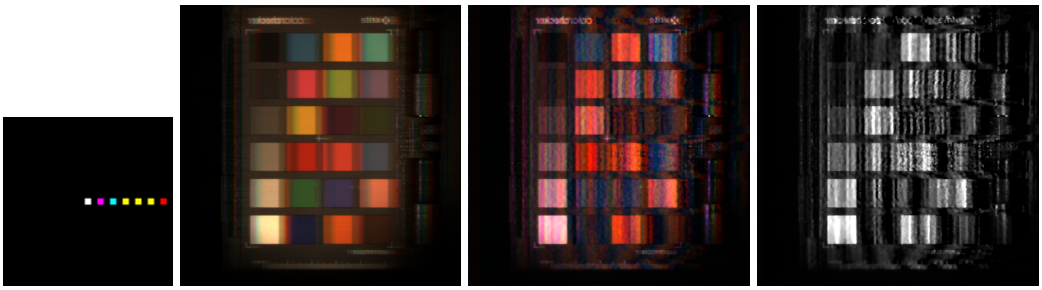
(e) (f)

Figure 4.9: Reconstruction with diverse scenes using ring PSFs: cloth, Lego, and lemon. (a) cloth ground truth (b) cloth reconstruction (c) Lego ground truth (d) Lego reconstruction (e) lemon ground truth (f) lemon reconstruction



(a) (b) (c)

Figure 4.10: Comparison of reconstruction from different PSFs. (a) ring PSF reconstruction (b) spiral PSF reconstruction (c) dot PSF reconstruction



(a) (b) (c) (d)

Figure 4.11: Reconstruction with linear dispersion. (a) Linear dispersion PSFs (b) Simulated snapshot (c) RGB image restored from reconstructed spectral image (d) Spectral image at 600nm. The failed reconstruction proves the advantage of our spatial and spectral variant PSFs.

Chapter 5

A Learning-Based Joint Design of Optics and Image Processing for EDoF

As computational imaging and optics converge, new application-specific camera designs emerge with a high-dimensional parameter space. The properties of point spread functions (PSFs), as the critical link between the optics and image processing, are vital to imaging performance. The last chapter has demonstrated the imaging power of joining heuristic-designed PSFs and inverse solvers. However, the best practice is to find the optimal PSF design from the feedbacks of the image processing module.

In this chapter, using the deep learning toolbox, we present a joint-design approach of parametrizing the optical formation of images and the reconstruction process using an end-to-end framework. We solve a classic problem, extended depth of field (EDoF), by producing a learned optical phase mask. The computational camera hardware verifies the results with a spatial light modulator (SLM). Our framework lays the foundation for learning-based joint-design with a non-regular phase mask and neural networks for image reconstruction.

5.1 Introduction

Driven by technological and manufacturing achievements, the generalization of digital photography is changing multiple economic, scientific, and social aspects [Maî17]. The software has become an integral part of photography products to satisfy the demand for new features and improved image quality. There are increasingly more digital imaging applications, from computer vision for autonomous systems to entertainment on smartphones, each with unique challenges and requirements. The discipline of computational photography uses a combination of optics and processing to acquire richer scene data compared to traditional camera approaches. Innovations overcome traditional imaging concepts, which mostly assimilate the visual system of mammals [SDP+18]. For example, extended depth of field (EDoF) captures a strong representation despite grades of blur from depth variation. Usually, the computational photography systems consist of unconventional optics that generate a coded image on the sensor, which is then decoded to be a visually recognizable representation using a processing algorithm [ZN11, Gre14].

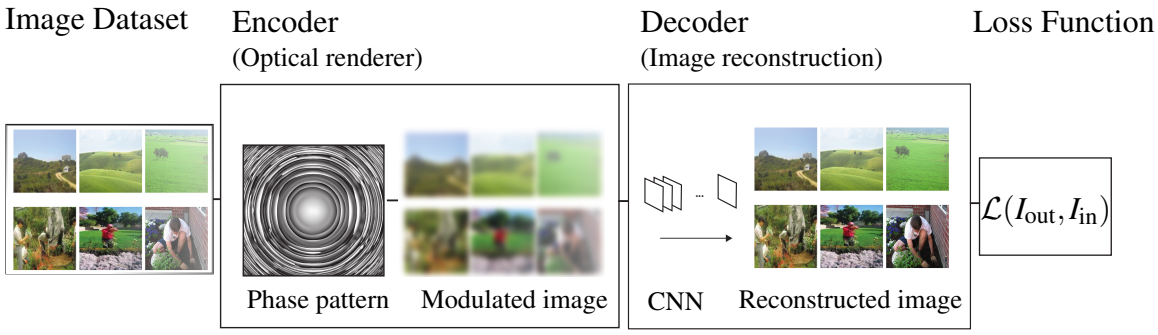


Figure 5.1: Learning-based joint-design framework. The encoder simulates the image formation based on wave optics theory. The decoder reconstructs the scene information from the intermediate, encoded data. The decoder uses a CNN to minimize the application specific loss function.

As is shown in Chapter 4, a joint design of camera hardware with manually tuned optics and processing algorithm is an effective method, but its heuristic nature bypasses the optimal solution. Considering the dimension of the design space and relations between different parameters, finding the optimal physical parameters poses a challenging task. In recent years, convolutional neural networks (CNNs) have achieved superior results in image restoration [XXC12, MSY16] and deblurring [XRLJ14, NKL17]. At the same time, spatial light modulators (SLMs), especially phase-only SLMs, based on liquid crystals, improved drastically over the past years in terms of resolution and cost, which makes them an attractive technology for future camera designs. Using an SLM for wavefront coding and learning-based reconstruction, our approach improves the design process and creates imaging applications considering the imaging chain.

Figure 5.1 illustrates our joint-design framework. The imaging process is simulated based on wave optics. The differentiable optical simulation uses Fresnel diffraction to parametrize optical configurations and calculate light propagation to adapt to the end-to-end framework. The image formation is modeled for convolution operations on monochromatic images from a large data set, with the simulated PSF incorporating phase modulation by the SLM. This intermediate image is passed to the decoder accompanied by a PSF. One PSF is used for images formed from all distances. It is computed taking the camera focus setting into account. The decoder module consists of a CNN. It is based on a U-Net architecture [RFB15]. During the training, the pixel values of the phase image and the network parameters are optimized to minimize the distance between the CNN output and the target images.

This chapter contributes to computational photography in the following ways,

- We introduce an end-to-end machine learning framework featuring a optical encoder for image formation and a U-Net as decoder.

- We present new phase masks for the classical task of EDoF.
- We build a computational camera setup incorporating an SLM and transfer the optimized phase patterns to validate the results.

5.2 Related Work

This work is built on the idea of end-to-end optimization in the context of computational imaging with wavefront coding and deep learning. Several works have proposed learning-based imaging system design interpreted as constraint auto-encoder networks. Chakrabarti [Cha16] and Henz *et al.* [HGO18] used a machine learning framework to optimize the color filter array in front of a camera sensor with a CNN for optimal color reproduction. Both achieved notably better quality than the traditional Bayer pattern in simulation. Nie *et al.* [NGZ+18] presented a similar joint optimization approach for hyperspectral image reconstruction. They also fabricated some of the learned filters and constructed a camera. Jiang *et al.* [JTFW17] and Schwartz *et al.* [SGB18] focused on the image processing part while still using an end-to-end approach to optimize the post-processing of raw sensor data. For example, Schwartz *et al.* [SGB18] used a CNN architecture to learn low- and high-level processing of image data in a smartphone camera successfully. Sitzmann *et al.* [SDP+18] developed a fully differentiable simulation of the image formation that can be used to optimize optical elements jointly with the reconstruction filter (i.e., the regularization parameter). They demonstrated two applications: EDoF and super-resolution. They optimized a single optical element and successfully manufactured it for each application. We extend this concept to a computational camera with an SLM, which needs no hardware modification once installed. Besides, it has the potential for future research in video-rate PSF engineering. Instead of basic deconvolution using Tikhonov-regularized least-squares, we employ a deep CNN for reconstruction.

Deep neural networks and CNNs have played a significant role in the context of low-level computer vision and graphics tasks [WT14]. For example, Pan *et al.* [PLS+18] and Liu *et al.* [LPY16] presented general neural network architectures that were trained to perform a variety of tasks like denoising, edge-preserving filtering, and super-resolution. Gao and Grauman [GG16] generated application-specific training data using a feedback mechanism to learn inpainting, interpolation, image deblurring, and denoising. High-performance results were achieved for both blind and non-blind image deblurring using CNNs [JMFU17, NKL17, WHSL17]. In our case, the reconstruction module is inspired by CNNs for both image reconstruction as well as deblurring.

Higher-level applications of deep learning can be found in the field of computational photography. CNNs are used to enhance light field acquisition [YJY+15, WZK+17] or to recover high dynamic range radiance maps from low dynamic range images [EKD+17].

Overall, a large number of camera designs have been proposed for computational photography [Nay11, ZN11, GIDW11]. The camera design is based on the work of

Chen *et al.* [CHH+17], which is a wavefront coding camera. The term wavefront coding [DC95, CD02] denotes optical coding by modifying the phase of a wavefront from incoherent illumination. For example, by introducing wavefront (or phase) coding with a glass plate of a specific shape, EDoF was achieved through digital deconvolution. The cubic phase mask [DC95, CD02] and other shapes have been proposed for the task of EDoF, such as the rotational symmetric approach by Ohta *et al.* [OSSS15] or even a combination of phase and color-coding [CZL+17]. These masks are mostly optimized for fabrication as optical plates, whereas our approach targets the SLM, which comes with a different set of constraints.

As mentioned above, the joint-design concept is inspired by the application-oriented framework [SDP+18, HGO18, Cha16]. This architecture combines both optimizations of optical phase functions and image decoding, which is shown in Figure 5.1.

We interpret the digital image processing pipeline as an auto-encoder. The encoder renders the optically modified image with actual physical parameters of the camera hardware. The decoder reconstructs the final image. In other words, the encoder models the image formation by simulating the propagation of light from the object to the sensor. Subsequently, the decoder contains convolution layers for image processing to produce interpretable results for human vision. For a given task, it enables the joint optimization of optical properties to form the coded image on the sensor, which can be seen as a learned feature map, and the filter parameters for image processing, which is interpreted as a chain of filter operations.

5.3 Optical Encoder

The encoder is defined as a differentiable renderer that simulates the sensor image. It takes the latent image data and the camera parameters as input. These physical parameters, including aperture size, focal length, image sensor pitch, compose the optical wave propagation, in other words, the image formation. The encoder is mainly made up of several modules that simulate optical diffraction. They generate the PSF, which characterizes the optical configuration, including the phase modulation to be learned. A convolution of the PSF follows the modules with the latent images to produce the simulated camera captures. A wave optics model [Goo17] is used to approximate the diffraction and the wavelength-dependent effects.

The scheme of the image formation is shown in Figure 5.2. An incoming light wave propagates from the object at the distance of d to the SLM. The programmable SLM modulates the wavefront with an 8-bit encoded image pattern. The modulated wave then continually proceeds by distance z_1 to the lens. Finally, the lens transforms the wave to the sensor plane. The analysis of this process is as follows.

From object to SLM. A point light source is located at the distance d from the SLM.

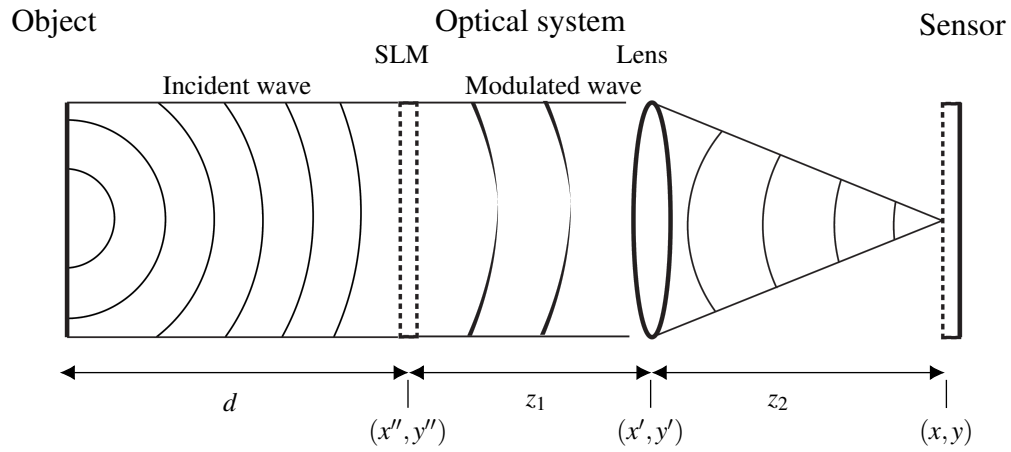


Figure 5.2: Optical encoder. The spherical incident wave is emitted by a point source located at d from the SLM. The optical wave passes through the SLM and is modulated. It then propagates distance z_1 to a fixed lens element. At last, it travels a distance z_2 until it reaches the sensor. Propagation is modeled using Fresnel diffraction approximation.

The SLM is illuminated by the spherical wave radiated from the source,

$$\begin{aligned} U_0(x'',y'';\lambda) &= \frac{A_0}{r_{01}} \exp[i\,\phi_d(x'',y'';\lambda)] \\ &= \frac{A_0}{r_{01}} \exp(ikr_{01}) \end{aligned} \tag{5.1}$$

where r_{01} describes the distance between the source and the pixels on the SLM. x'' and y'' are the 2D coordinates at the SLM plane.

$$r_{01} = \sqrt{x''^2 + y''^2 + d^2} \tag{5.2}$$

ϕ_d is the phase function of the spherical wavefront. A_0 is the amplitude of the wavefront. The wavenumber k is defined as $k = \frac{2\pi}{\lambda}$.

From SLM to lens. The modulation of the pupil function is denoted as ϕ_S, which is the HD resolution image to be learned by the network. The modulated wavefront is,

$$\tilde{U_0}(x'',y'';\lambda) = U_0(x'',y'';\lambda) \exp[\phi_S(x'',y'';\lambda)] \tag{5.3}$$

The propagation from the SLM to the lens is simulated using Fresnel approximation (see

Equation (2.12) in Chapter 2 for more information about the approximation) as follows,

$$U_1(x',y';\lambda) = \frac{\exp(ikz_1)}{i\lambda z_1} \iint_{-\infty}^{\infty} \tilde{U}_0 \exp(i\frac{k}{2z_1}[(x'-x'')^2 + (y'-y'')^2]\,dx''\,dy'' \tag{5.4}$$

From lens to sensor. The phase delay caused by the fixed lens can be derived from a thin lens model in paraxial approximation as,

$$\begin{aligned}\tilde{U_1}(x',y';\lambda) &= U_1 A(x',y') \exp[\phi_l(x',y';\lambda)] \\ &= U_1 A(x',y') \exp[\frac{-\pi}{\lambda f_l}(x'^2,y'^2)]\end{aligned} \tag{5.5}$$

where f_l is the focal length. ϕ_l is the phase modulation by a diffraction-limited lens, which is $\frac{-\pi}{\lambda f_l}(x'^2,y'^2)$. $A(x',y')$ is the complex transmittance representing the effect of the lens aperture. The last step from the lens to the sensor is again calculated via Fresnel propagation,

$$U_s(x,y;\lambda) = \frac{\exp(ikz_1)}{i\lambda z_1} \iint_{-\infty}^{\infty} \tilde{U}_1 \exp\{i\frac{k}{2z_1}[(x-x')^2 + (y-y')^2]\}\,dx'\,dy' \tag{5.6}$$

The Fresnel approximation is evaluated numerically using Fourier transforms, which is described in more detail in Section 5.6. The PSF under incoherent illumination is then obtained as the squared magnitude of the amplitude spread function $U_s(x,y)$, i.e. the field amplitude produced by the unit-amplitude point source,

$$P(x,y;\lambda) = \|U_s(x,y;\lambda)\|^2 \tag{5.7}$$

A convolution of the input with the generated PSF produces the optical image on the sensor. The last step is to sample this image concerning the sensor parameters. Thus, we add noises and clip the optical image to the dynamic range of the sensor. Since a PSF varies by wavelength, we consider the spectral dependency of the monochromatic sensor as $S(\lambda)$. The image noise is denoted as $N(x,y)$. In the optical renderer, the noise is formulated according to a Gaussian distribution. The simulated capture $I_s(x,y)$ is then as follows,

$$I_s(x,y) = \int S(\lambda)[I(x,y;\lambda) \otimes P(x,y;\lambda)]\,d\lambda + N(x,y) \tag{5.8}$$

Differential renderer. The image formation is formulated by simulating stages of optical wavefront propagation. In the encoder, it is modeled as a chain of multiplicative operations. This chain indicates basic operations using pre-defined gradient operations in machine learning libraries and the accumulation of the overall gradient based on the

chain rule. To this end, we build a differential renderer to simulate a computational camera incorporating phase modulation.

5.4 Decoder

The coded image is handed over to a CNN for decoding. The idea is employed to assimilate the imaging model by transmitting physically available information from one module to the next. It gives the decoder freedom in learning the optimal use of the data. This general optimization framework suggests a generic CNN architecture as an initial design for the decoder. The CNN-base decoder is a modified U-Net [RFB15] as shown in Figure 5.3. A general introduction to U-Net is presented in Figure 2.14 in Chapter 2. The U-Net architecture is proven to reach outstanding performance on several image-to-image tasks, such as super-resolution and deblurring [NGZ+18, NKL17, JMFU17, IZZE17]. It consists of several levels of downsampling to save memory. Meanwhile, it increases the receptive field and incorporates residual connections between levels of abstraction.

Our decoder encompasses residual blocks [HZRS15] with minimally three convolutional layers and three scale levels. Variants with more levels are modified depending on the available memory and task. Comparing with the original formulation, we use concatenation only for the topmost skip-connection and summation. This design saves memory and suits the deblurring tasks as is described by Nah *et al.* [NKL17]. We employ variations of the U-Net with three (U-Net 3) and four (U-Net 4) scale levels. If the scale levels are few, we utilize *à trous* convolution to increase the receptive field, which is beneficial for deblurring tasks [SBHS13, Cao15]. Additionally, an output block is adjoined without residual connection to perform post-processing like denoising. The beginning of each layer contains a ReLU as is advised by He *et al.* [HZRS15]. It follows a convolutional or transposed convolutional operation with trainable filters. The dimension of regular convolutions is three by three, with a stride of two for downsampling. For upsampling, we apply transposed convolution with a fractional stride. To include more local context in the upsampling process, we utilize four by four kernels.

5.5 Training

A set of roughly eighty thousand high-resolution images are facilitated in training in the back-propagation fashion. Both image dimensions are larger than two thousand. We obtain the data from the Inria Holidays dataset [JDS08] and a Google image search crawl. To cover a wide range of image statistics, we pursue various scenes using the ImageNet classes as search queries. To avoid violating copyrights, we obtain only images with creative commons licenses. The threshold for selection is a minimum resolution and the image data is stored in random order. During training, patches of the are cropped

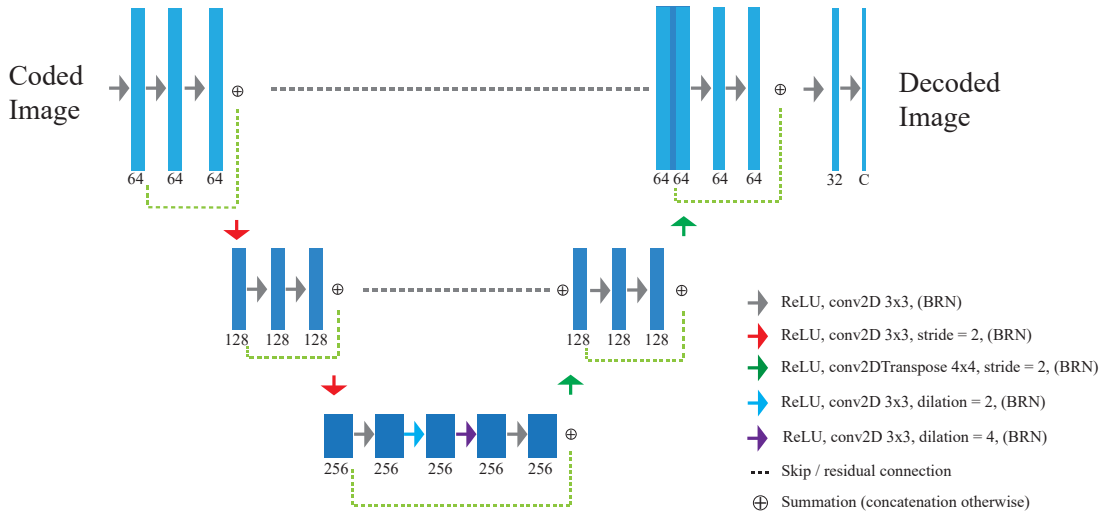


Figure 5.3: Decoder (U-Net 3). This is a modified U-Net architecture with five residual blocks at three scale levels. The scheme can be expanded by compensating with additional scale levels. The U-Net 4 is structured one level deeper. BRN represents batch renormalization.

randomly, scaled, and randomly rotated by multiple steps of 90°. Additional basic augmentation includes flipping and flopping.

ϕ_s is defined in Section 5.3 as the trainable variable in the optical encoder. It is initialized with a dark frame. To stay faithful to the physical constraints of the SLM, we clip the variables to a range between zero and one, which corresponds to a phase shift between zero and 2π. A linear curve of the SLM image-phase correspondence is presumed during the training. The actual response of the SLM is linearized using a look-up table during camera calibration.

The loss function in training f_t is defined as the mean square error (MSE) between the decoder output $\hat{\mathbf{I}}_i$ and the input latent image $\mathbf{I}_{oi}$, e.g., the sharp image in the case of EDoF,

$$f_t = \frac{1}{m}\sum_{i=1}^{m} ||\hat{\mathbf{I}}_i - \mathbf{I}_{oi}||^2 \tag{5.9}$$

Noisy PSFs can lead to low contrast of reconstruction. To prevent this effect, we enforce sparsity on the PSFs with L1 and total variation regularization for ϕ_s. Furthermore, axial symmetry can be enforced to increase the training speed by optimizing only a quarter of the SLM. The phase pattern ϕ_s can then be duplicated and mirrored. We employ the ADAM optimizer for training with the proposed settings by Kingma and Ba [KB14]. The learning rate begins as a small learning rate, which enables the reconstruction network to stabilize. Afterward, the rate is increased to accelerate the generation of the phase pattern

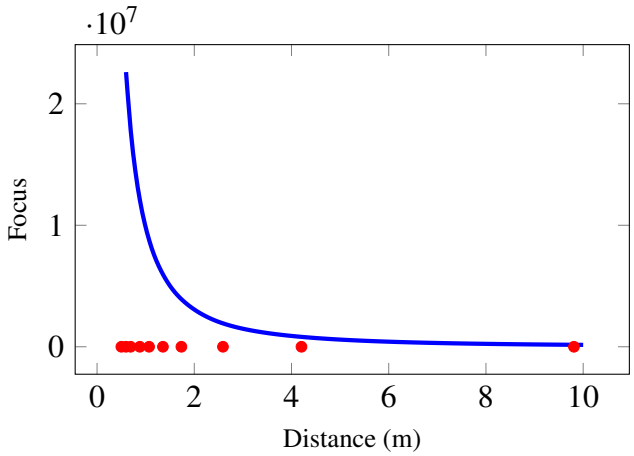


Figure 5.4: Distribution of distance values for the EDoF training. The values are identified according to the change of focus. The blue curve represents the location of the focus point versus the object distance. The red dots are sampled locations. For smaller distances, the spacings are smaller between samples.

until it can be decreased again for fine-tuning at the last stage. The statistics change drastically between batches because of the random depth (d in Figure 5.2) assigned to each training image.

5.6 Applications

Implementation. We implement the framework with TensorFlow and Tensorpack for Python [AAB+15, Wu18]. Our method of evaluating Fresnel diffraction integral [TF82, Rob86]uses the angular spectrum method with various grid sizes of the propagation planes. The Fourier transforms are computed as fast Fourier transforms (FFT). The general difficulty of the Fresnel diffraction integral is the quadratic phase factors since they are not band-limited. In discrete Fourier transform, if the grid spacing and the number of samples are not well-defined, severe aliasing can result in unreliable simulation. In an attempt to reduce aliasing, the optical field is upsampled and padded with zeros during the simulation of the optics. It is generally practical to scale the optical elements, with each grid spacing determined by the number of samples. Convolution is computed as multiplication in the frequency domain, and the complex exponents are computed using Euler's formula.

EDoF in simulation. As stated above, our goal is to reconstruct sharp images from blurred images caused by a wide range of depth without varying the lens's focus in the camera. Therefore, the best practice is to locate the input images representing object radiance at different distances d. The value of d is randomly selected per batch from a

range between optical infinity and 0.5 m as the target in-focus range. The distribution of distances for generating training data significantly impacts the system performance and the training speed. In order to reach consistent performance over a wide range of depth, training data should include close objects, where the focus shifts stronger than distant scenes. We formulate the focus shift according to the thin-lens model and the object distance, as a Fourier relation ($h(d) = \mathcal{F}\{e^{-i(\phi_d+\phi_l)}\}$) using the Fraunhofer approximation [Goo17],

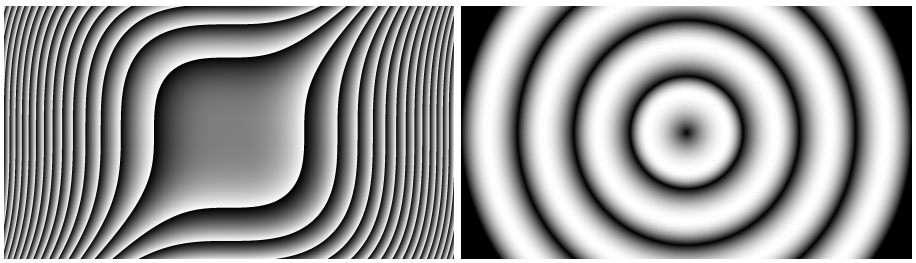
$$C = \frac{h(d_2) - h(d_1)}{d_2 - d_1} \tag{5.10}$$

Equation (5.10) corresponds to the mean value of the derivative of $h(d)$ over the interval $[d_1, d_2]$ of two adjacent depths. Importance sampling of the equation yields a sample distribution as shown in Figure 5.4.

Regarding the fixed camera focus, we define z_2 in Figure 5.2 to be either infinity or 2 m according to the real-world hardware setting. The simulation is performed under monochromatic illumination at 520 nm band. Likewise, the focal length of the fixed lens is set to be 100 mm, and the aperture is F5.6. The dimension of SLM is also applied with the physically real values: 15.36 mm by 8.64 mm with a pixel pitch of 3.45 µm. To enable the modeled phase pattern with flickering, we add uniformly distributed noise in the range between -0.015 to 0.015, with zero as the mean. Besides, a random standard deviation between 0.005 and 0.02 is added to the coded image. We pre-process the expected sharp image without adding noise by firstly converting linear data to sRGB space, then taking the green channel as the ground truth. Our camera setup sees a peak in the PSF's center. This peak is the zeroth-order diffraction plus the reflection of non-modulated light. We simulate the final PSF using a weighted average of a modulated and non-modulated PSF for image convolution, with the weight of the non-modulated light between 5 to 10%. In an effort to model the limited frequency range of the real camera, a low pass filter is applied to the PSF. On an Nvidia GTX1080 Ti, the average training speed is 1.2 iterations/s. The training is performed in parallel on two GPUs with the batch of one. A pattern is learned after 80,000 iterations. Marginal improvement is seen after 160,000 iterations.

To evaluate the outcome, we utilize two classic phase patterns as a comparison. They are the cubic phase mask [DC95] and the annular phase mask (APM) [OSSS15] (shown in Figure 5.5). To investigate the learned reconstruction algorithm, we compare with a basic Wiener filter for deconvolution using an optimized regularization parameter and the Richardson-Lucy deconvolution approach [Luc74, Ric72b].

Camera prototype. We build a computational camera and acquire images after encoding the learned phase patterns from the end-to-end network mentioned above to validate our approach. The camera setup is demonstrated in Figure 5.6. The PSFs are measured with a 3D-printed pinhole mask with a diameter of 1 mm. The illumination is from a halogen light source Osram 64655 HLX. A spectral bandpass filter (Rosco 4490 CalColor Green 90) is installed with a peak wavelength of 530 nm. The camera is designed to be reimple-



(a) Cubic phase mask (folded) (b) APM phase mask

Figure 5.5: Cubic and APM pase mask. They are modified according to HD 16:9 aspect ratio. The cubic phase mask is truncated within the phase modulation range.

mented with alignment tools such as lasers, alignment targets, and mirrors.We employ a reflective installation of scenes. A focus test chart and some images from our test data are printed as grayscales on A4 papers with a Konica-Minolta C368 laser printer. Afterward, we measure the distance between the test image and the surface of the SLM. The camera focus is tuned at 2 m. The test targets are located at five distances 1 m, 1.5 m, 2 m, 2.5 m and 3 m. Experiments are compared with phase patterns for each location: constant phase (no modulation), cubic phase function, and APM (eight rings). A well-known setup for aperture coding is the $4f$ setup, where two lenses are used to extend the pupil plane. However, a key problem is the difficulty of transferring this setup to a commercial photographic device. Thus, we use a telephoto lens and insert an SLM in front of the lens. The simulation of image formation is in line with the light propagates from the object to the SLM and then is reflected to the lens.

Each pixel of the SLM is addressed with a voltage level according to the gray value, which alters its refractive index. To assure the scale of optical modification by the phase pattern, we configure the voltage-to-modulation level of the SLM by applying a look-up table. The modulation range is tuned within 0 to 2π, corresponding to the grayscale image values 0 to 255. In each iteration of the PSF acquisition, we upload the phase patterns to the SLM. It is operated as a second display device of the workstation. We align the color-filtered pinhole light to the center of the region of interest (RoI) and capture snapshots with the Oryx camera. In a similar way, the targets are illuminated by color-filtered light while each phase pattern is uploaded to the SLM.

5.7 Results and Evaluation

The first set of results are the optimized phase patterns. It evaluates the learned PSFs with depth-invariant properties. Furthermore, the reconstruction results of the simulation are investigated to analyze the capacity of the end-to-end framework itself. Finally, we show the real-world reconstruction results of EDoF, which verifies the framework in the real scenario.

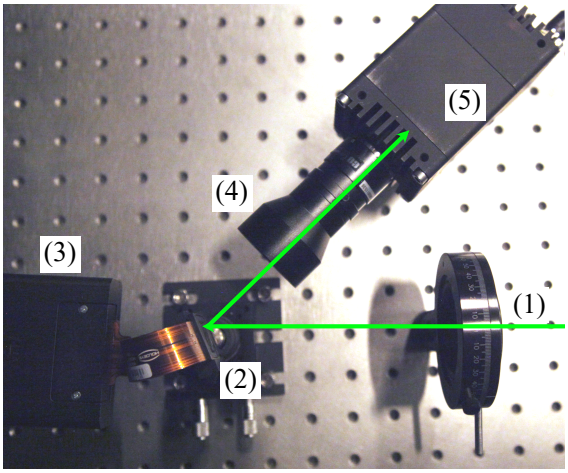
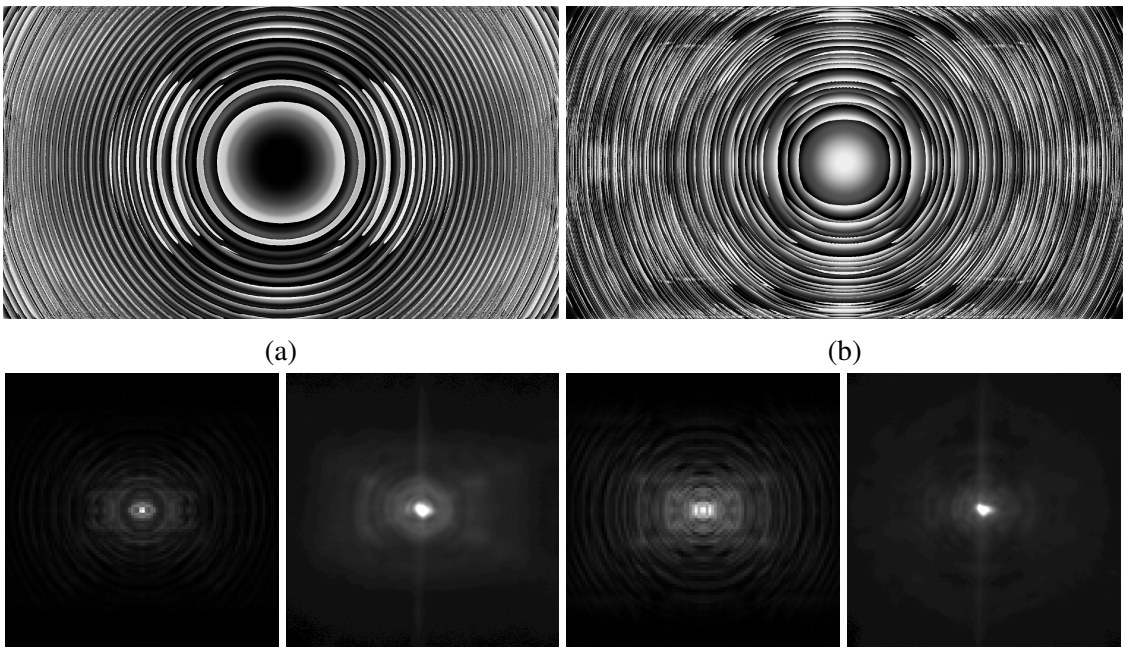


Figure 5.6: Camera setup with programmable optics (SLM). (1) is the linear polarizer. (2) is the cube beam splitter. (3) is the HOLOEYE PLUTO SLM. (4) is an SLM driver. (5) is a 100 mm telephoto lens. (6) is the FLIR Oryx camera (ORX-10G-123S6M-C).

PSF simulation. To evaluate the EDoF application, we first examine the simulated PSFs in contrast with the real PSFs. The results of the PSF comparison are shown in Figure 5.7. Figure 5.7(c)(e) show features of rings and rectangles in the center, and (d)(f) match in the real world except for the blurriness. In Figure 5.7(a)(b), the generation of the phase patterns is enforced with axial symmetry. The patterns resemble a combination of different frequencies of rings similar to the annular phase mask (Figure 5.5b). Although the PSFs are not strictly depth-wise constant, the motif of rings and rectangles are cues for the network to identify similarities of PSFs to extend the depth of field. The selection of distances assigned to the training images strongly influences the generated patterns. The high frequency noise in Figure 5.7 appears when we include object depths under 1 m. The high-frequency discontinuities improve results in simulation. However, they generate cloudy artifacts in real PSFs because of the influence of neighboring pixels in the SLM.

To demonstrate the depth-invariance of the PSFs, we show simulated PSFs across multiple depths from Figure 5.7(a) in Figure 5.8. From 1 m to 50 m, the PSFs are strictly identical, but they contain repetitive rings, which serve as cues for the decoder to detect depth invariance.

EDoF simulated results. The next step is to investigate the simulated results. Figure 5.9 shows two sets of reconstructed images using the phase pattern from Figure 5.7(a). The results from three targets with three d values are compared with the ground truth. We show the average peak signal to noise ratio (PSNRs) in Table 5.1 over a test set evaluated for nine locations, which includes multiple untrained locations to explore the versatility of the network. Our findings show that the system can generate sharp images for a wide range of depth. In addition, they affirm the system performance within the depth range



(a) (b)

(c) (d) (e) (f)

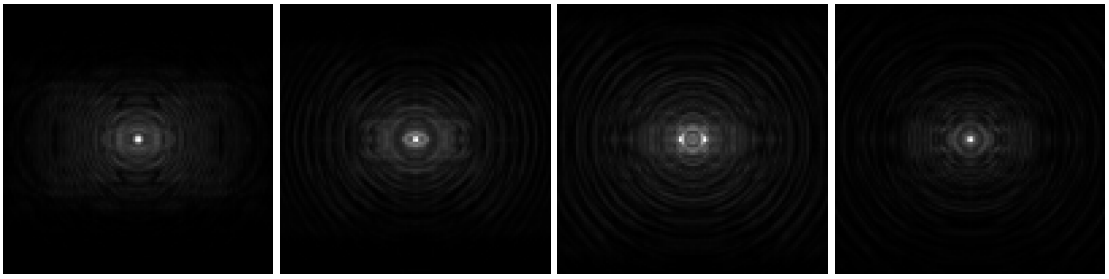
Figure 5.7: Optimized phase patterns and their PSFs. (a) is the phase pattern with axial symmetry constraints. It is trained by focusing at 2 m with an aperture of F5.6. (b) is the simulated PSF of (a). (c) is the real PSF of (a). (d) consists of the phase pattern with axial symmetry constraints. It is trained by focusing on infinity. (e) is the simulated PSF of (d). (f) is the real PSF of (d). The results are produced by *Zeroth-order diffraction and PSF filter with U-Net 4 and low noise* in Table 5.2.

even without training on a specific depth.

Table 5.1: PSNR means of nine distances in simulation

Distance(m)	1	2	3	4
PSNR	32.9012	31.3133	29.0155	34.2097
5	6	7	8	9
33.6177	31.8464	32.8652	26.3337	23.3679

Table 5.2 compares simulation results on a test dataset with 500 images. The images are randomly selected high-resolution images from the Google Open Images Dataset v4 test set [KDA+17]. We compare different reconstruction approaches, optical configurations with different noise settings or symmetry constraints, and distance distributions during training. Two levels of Gaussian noises are adopted: a high-level noise with σ between 0.02 and 0.05; and a low noise level with σ between 0.001 and 0.015. In each

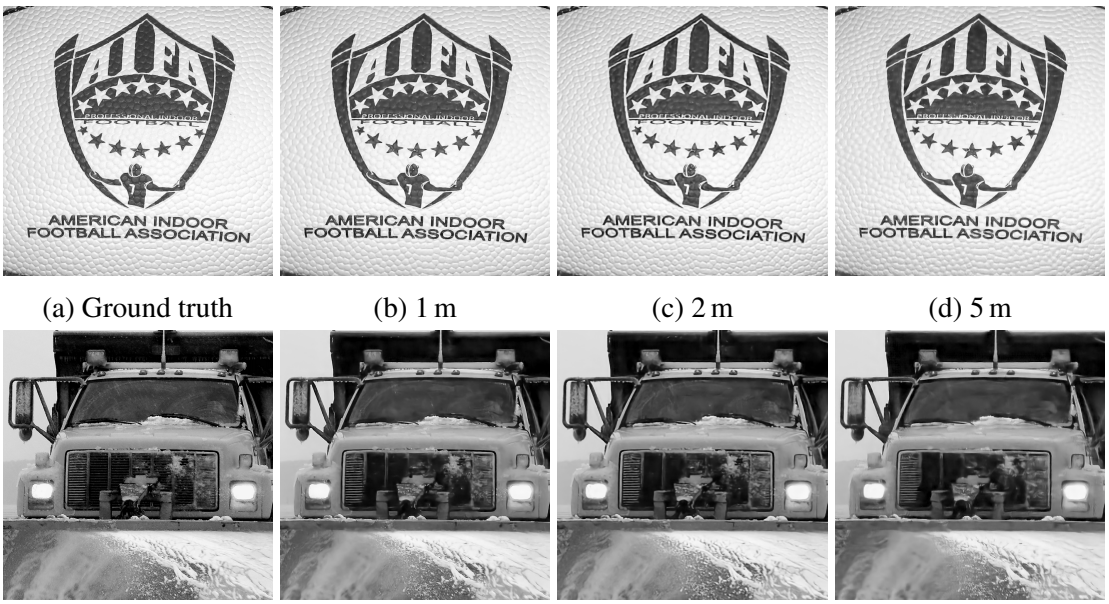


(a) 1 m (b) 2 m (c) 5 m (d) 50 m

Figure 5.8: Simulated PSFs across different distances using the phase pattern of axial symmetry constraints focused at 2 m (Figure 5.7(a)). Note that depth invariant features remain while blur increases, despite the PSF being split into two peaks at 5 m.

iteration, they are chosen randomly. Furthermore, some configurations apply a low pass filter on the PSF. In addition, the central peak of the PSFs is considered and tested independently. Regardless of the training distance distribution, tested distances in [m] are

$$\{0.5, 0.62, 0.75, 0.8, 1, 1.5, 2, 3.5, 5, 50\} \tag{5.11}$$



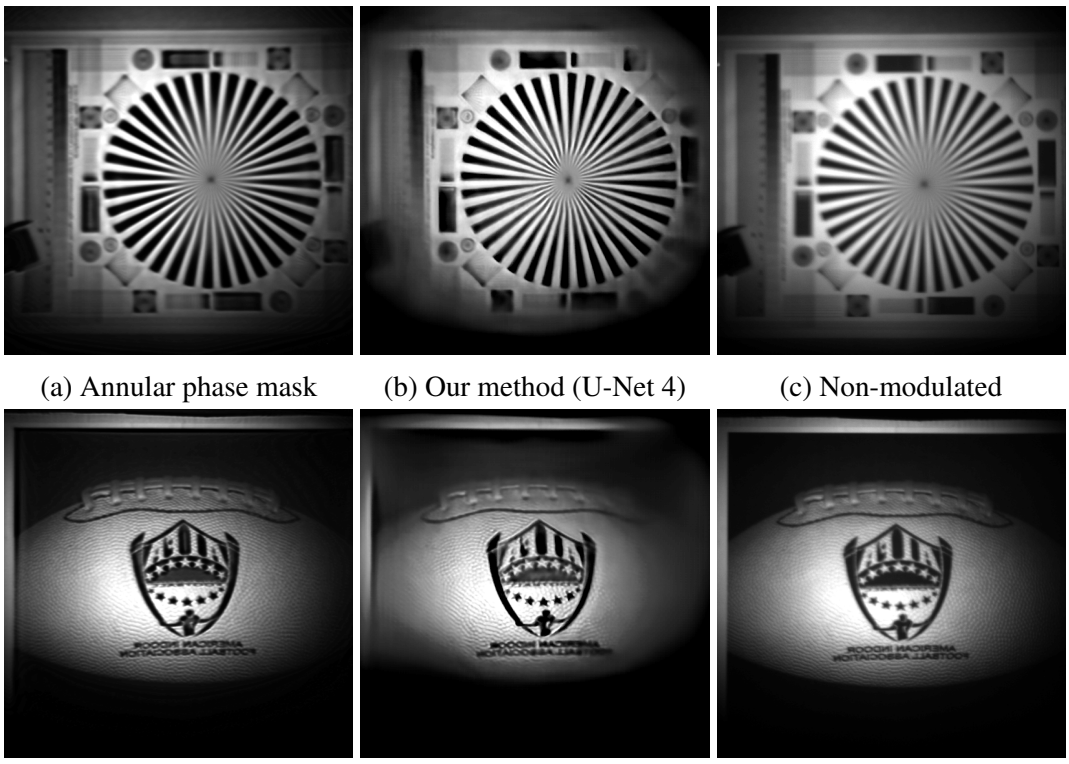
(a) Ground truth (b) 1 m (c) 2 m (d) 5 m

(e) Ground truth (f) 1 m (g) 2 m (h) 5 m

Figure 5.9: Simulated EDoF results using the phase pattern Figure 5.7(a). The results are produced by *Zeroth-order diffraction and PSF filter with U-Net 4 and low noise* in Table 5.2.

Table 5.2: Quantitative comparison of various configurations for EDoF tested on five hundred images using average PSNRs and structural similarity index measure (SSIMs). The high noise level denotes a Gaussian noise with σ between 0.05 and 0.02. The low noise level indicates a σ between 0.001 and 0.015. "PSF filter" abbreviates the application of a low pass filter on the PSF, and "zeroth-order diffraction" refers to the simulation of the central peak of the PSFs.

Method	Initial focus	PSNR	SSIM	Noise
[SDP+18] SLM with U-Net 4	inf	29.34	0.80	low
[SDP+18] lens with Wiener Filter	inf	24.1	0.67	low
Cubic phase with Richardson-Lucy after 30 iterations	inf	18.5	0.61	low
Cubic phase with Wiener filter after single iteration	inf	21.6	0.425	low
Cubic phase with Extended U-Net CNN	inf	22.7	0.67	low
Cubic phase with U-Net	inf	18.3	0.57	low
4 rings APM phase with Richardson-Lucy after 30 iterations	inf	17.5	0.54	low
4 rings APM phase with Wiener filter after 3 iterations	inf	21.6	0.43	low
Ours: SLM coded aperture 50 mm lens with U-Net 3	inf	31.28	0.84	low
Ours: Unconstrained SLM with U-Net 3	inf	**31.84**	0.85	low
Ours: Zeroth-order diffraction and high noise with U-Net 4	2.5 m	30.73	0.82	low
Ours: Zeroth-order diffraction and high noise with U-Net 4	2.5 m	30.00	0.80	high
Ours: Zeroth-order diffraction, high noise, and PSF filter with U-Net 4	2.5 m	29.95	0.80	high
Ours: Zeroth-order diffraction and PSF filter with U-Net 4	2.5 m	30.24	0.81	low
Ours: Symmetric PSF filter with U-Net 4	inf	29.77	0.80	low



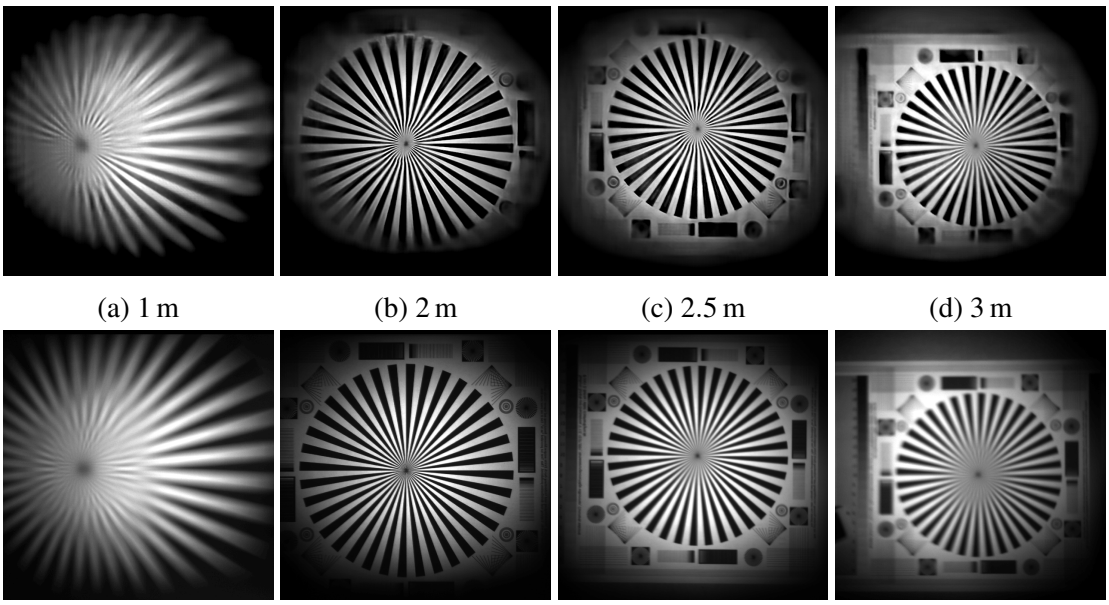
(a) Annular phase mask (b) Our method (U-Net 4) (c) Non-modulated

(d) Annular phase mask (e) Our method (U-Net 4) (f) Non-modulated

Figure 5.10: Results from real-world camera captures at 2.5 m. The results are produced by *Zeroth-order diffraction and PSF filter with U-Net 4 and low noise* in Table 5.2.

Our results achieve better scores in all configurations in simulation compared with other approaches like reconstruction with a regularized Wiener filter. Our system also extends the experiment of [SDP+18] with the SLM and a CNN-base reconstruction. In practice, we select the *Zeroth-order diffraction and PSF filter with U-Net 4 and low noise* in Table 5.2 to demonstrated simulated and real-world results, because it is closest to the camera model.

EDoF real-world results. The decoded images from the real camera are presented in Figure 5.10 and Figure 5.11. Qualitative analysis of the real-world data reveals the effectiveness when compared with the non-modulated acquisitions. The quality of reconstruction is similar to that of the APM mask. However, it succeeds in retaining more high-frequency content. It is critical to note from the test chart that the details of the central spikes are more comprehensible than the APM approach. This result supports the previous simulated prediction. However, we admit that the visual quality does not reach the level of the simulation. Contrary to expectations, the image margins are smeared. The vignetting effects due to PSF spatial variation may likely have caused the error. Nevertheless, we can still state that our end-to-end method is competent in the paraxial area.



(a) 1 m (b) 2 m (c) 2.5 m (d) 3 m

(e) 1 m (f) 2 m (g) 2.5 m (h) 3 m

Figure 5.11: Reconstruction of test chart across different distances. The first row is composed of the restored images by our approach. The second row presents the non-modulated images. The results are produced by *Zeroth-order diffraction and PSF filter with U-Net 4 and low noise* in Table 5.2.

The system achieves EDoF by encoding a learned phase modulation to modify an intermediate image, which a trained CNN subsequently decodes. The framework demonstrates stability by its capacity to decode images from the untrained depth and the commonality of various learned phase patterns. The execution of the trained network reaches up to six frames per second on an Nvidia GTX1080 Ti GPU, which is potentially adaptable to embedded hardware in the future.

Achromatic EDoF. Our framework is not limited to monochromatic photography. RGB images are associated with three discrete spectral responses in the standard color mode; thus, up to three channels can be processed in a single iteration in training. For more than three channels, we propose a stochastic approach to select spectral bands like the object distance in a similar way with [SDP+18]. As a proof of concept, we demonstrate in Figure 5.12 with simulated EDoF results with RGB output. Chromatic aberrations are barely observable from the results. The simplified configuration is utilized, consisting of a 50 mm lens and the phase modulator located at the pupil plane. The object is positioned at 0.8 m away.

Limitations. Optical aberrations, especially concerning the spatial variance of PSFs, are not modeled in simulations, which can be incorporated in the optical encoder in the future. In our framework, the image formation is a presumed paraxial approximation. Nevertheless, as long as the distance between the optics and scene is not over the ap-

(a) Ground truth (b) Non-modulated (c) PSF-modulated (d) Reconstructed

Figure 5.12: Achromatic EDoF. The image is placed at 0.8 m away from the camera. The results are produced by *Zeroth-order diffraction and PSF filter with U-Net 4 and low noise* in Table 5.2.

proximation limits, this approach is valid [ANNW16, SDP+18]. The scene is currently modeled only with plane objects. We show a failed reconstruction with three objects at three distances in Figure 5.13. To conform to the physical model, it is advisable to improve the encoder as an optical raytracer, which demands more GPU memory as a trade-off. In addition, the ignorance of the spectral response of the SLM constrains the achromatic EDoF in a real-world application. The spectral discrepancy is to be investigated and corrected. High-power illumination and displaying of test targets are key to experiments. The actual noise profile is to be investigated. Multiple factors, such as SLM and cameras, should be tested and adopted in the network.

The image metrics do not always represent visual quality. High scores of PSNR do not signify content fidelity. In the future, content-specific metrics should be investigated along with decoder parameter design.

Regarding sceneries with multiple objects, the experiments did not show remarkable improvement over APM masks. Figure 5.13 pinpoints the failure of restoring edges when various depth is present. This result is not unexpected due to the mentioned training scheme solely using plane objects. Despite this limitation of our method, our findings do confirm the concept of learning-based computational camera design.

5.8 Conclusions and Future Work

The presented joint-design framework is a paradigm to adopt deep learning in computational camera design. It incorporates a differential renderer as the encoder and a U-Net as a decoder in an end-to-end fashion.

We have outlined the theoretical foundation and design concept, followed by the structure of the optical encoder, which mainly approximates the PSF through optical wavefront propagation. A convolution operation and several preprocessing functions then define the coded image. The U-Net decoder is adapted to restore the intermediate image

(a) Non-modulated (b) Reconstructed by APM (c) Reconstructed by our method

Figure 5.13: A failed case containing reconstruction of a scene with objects at various depth. The results are produced by *Zeroth-order diffraction and PSF filter with U-Net 4 and low noise* in Table 5.2.

to the target image. A training procedure is devised using many high-resolution images, careful arrangement of object locations, and observance of physical constraints. An MSE is employed as the objective function while the convergence of the SLM phase pattern is inspected.

We managed to validate the EDoF application in both simulations and on an actual camera prototype. The results indicate that the system benefits from optimizing network parameters and phase modulation in unison. Cues of depth-invariance are visible from the simulated PSFs. The simulation yields higher scores than those using fixed phase masks with a deconvolution solver. The optimized pattern is transferred to the real camera for image acquisition. The real PSFs resemble the simulated ones. The real EDoF results improve on previous methods. The limitation of our approach is the training on plane objects. In addition, color EDoF reconstruction can be realized if the spectral responses of the sensor and SLM are measured, as is demonstrated by Lizana *et al.* [LMM+08]. We present a proof-of-concept of achromatic EDoF in simulation.

Depending on the future application, other CNN architectures and loss functions are worth exploring. An adversarial learning approach with additional feedback to the system can improve performance. Besides, physically real model-based learning is essential for future research. Future work should investigate a full-fledged optical renderer, including PSF spatial variance, with various object depths in one scene. Our method has important implications in substituting complex traditional lenses for high-quality imaging with low cost. A learning-based approach could also enable completely novel optical design and imaging applications.

With the feasibility of the learning-based joint-design framework validated, one reasonable next step is to challenge a crucial problem in computational photography: optically accurate zooming using a single camera with a unifocal lens. The thickness of

the camera has been a defining factor in the design of mobile phones. It is not yet established whether, in mobile photography, a single camera can achieve optical or hybrid zooming without bulky varifocal lenses. We propose a computational camera platform for learning hybrid zooming in the next chapter.

Chapter 6

A Computational Camera for Hybrid Zooming

Zooming has many roles in the field of photography, observation, and medical imaging. It varies the field of view (FoV) by changing the size of the region of interest. Previous work has focused chiefly on the separate design of optical systems–optical zoom lenses, solid-state zoom systems, or digital zooming systems. With the advancement of the mobile camera industry, multi-camera digital-optical hybrid zooming with lenses of various focal lengths emerged to estimate an image of arbitrary magnification with both fidelity of real-world radiance and data processing power. Recent developments regarding computational photography have led to the joint design of optics and image processing, which finds optimal parameters for both optics design and CNN-based image reconstruction and drove a series of novel solutions. However, it has not yet been established whether the joint-design approach is suitable for non-mechanical single-camera zooming, which could drastically reduce the energy cost of the multiple cameras and eliminate the size of the objective lens. We build a computational camera for a single camera and a single-shot hybrid zooming. We insert a phase-only SLM in the intermediate image plane to alter the chief rays and then form the image on the sensor with a relay lens. The SLM functions as a programmable optical device whose encoded phase image is to be determined.

6.1 Introduction

Mobile photography accounts for much of the progress of computational photography that integrates capturing and processing techniques. One remarkable feature of mobile photography is its compactness and computational power. It is increasingly desirable to capacitate mobile phones for professional photography. As an essential member of photographic lenses, zoom lenses cover a wide range of focal lengths with a portable size for the professional end. However, their volume remains too bulky for mobile photography. Besides, the camera size in a smartphone is a vital factor that accounts for its thickness. Thus it is significant to investigate zooming in mobile photography while maintaining a small space for imaging equipment.

In the history of zooming on smartphones, studies were made in digital, optical, or

hybrid methods, where the first focuses on manipulating ready-captured image data; the second seeks to design micro zoom lenses; and the third fuses both methods, often with multiple cameras. Notable shortcomings are the loss of physical fidelity of the digital method, the sacrifice of compactness of the optical method, and the multiplied energy cost of the hybrid one. It is a dilemma to maintain optical faithfulness while preserving the small space and little energy consumption.

This work aims to overcome this dilemma by a hybrid zooming method, optically faithful while only using a single camera with a unifocal lens. We identify the chief rays, which define the field of view (FoV), to be engineered. The chief rays originate from the object and then pass through the center of the aperture stop. The intersection between the chief ray and the image plane thus defines the FoV. We inactivate the FoV's programmability by integrating an SLM to an intermediate image plane between a primary imaging lens and a pinhole camera, where each SLM pixel registers a fixed bundle of chief rays, therefore manipulating their refraction. The SLM is encoded by an 8-bit image whose pixel values represent the optical phase modulation, thus carrying out the camera image distortion by acting as a freeform lens-like modulator, the only variable being the refractive indices. The strengths of this imaging device lie not only in the maintenance of its compact size and simple optical design but also in the framework's openness to data-driven learning approaches.

The contribution of this work is the computational camera, integrated with programmable optics to encode the FoV, which provides a platform for training deep learning-based hybrid zooming solutions.

6.2 Related Work

In recent years, there has been growing interest in joint design of optics and image processing, particularly using deep learning. This section reviews the body of literature that investigated this scheme and specifically hybrid optical-digital zoom.

Joint design of optics and image processing. Much work on computational photography has improved the capabilities of cameras, such as super-resolution [BEZN05], extended depth of field (EDoF) [LHG+09], digital refocusing [ANNW16], etc. While the traditional camera is an imitation of the model of a human eye with a lens, sensor, and casing, other animal eyes [LN12] also consist of mirrors and compound lenses. It is inspiring that their eyes and neural systems are highly mutually adapted. In the last years, optical information processing has been exploited to facilitate the physically-based image processing for novel applications, such as mask-based lensless imaging [AKH+18][KHFG14, MYK+19] and computational photography with diffractive optics elements [PSD+19, CHH+17, CHEL18], descattering and seeing around corners [KHFG14], transient imaging [HHGH13], light transport probing [GLDZ15], etc. With the development of a deep learning toolkit that automates differentiation, it became feasible to incorporate hardware parameters in an end-to-end fashion, e.g., Henz *et*

al. [HGO18] introduce a deep joint design of the color filter array and demosaicing. For instance, several end-to-end optimization frameworks that join optics and image processing are proposed. Sitzmann *et al.* [SDP+18] introduce a joint-design framework with a diffractive optical element for extended depth of field and super-resolution. Optical phase masks can be learned for depth estimation [WBC+19, CW19]. Sun *et al.* [SZD+19] compensate the low pixel count of the single-photon avalanche photodiodes by learning the optical design. Metzler *et al.* [MIPW20] solve the single-shot HDR problem by learning the lens surface as well as the HDR image reconstruction. Among these works, PSF engineering was used as a common link between the optical design, often using diffractive optical elements (DoEs) and image processing. We modify a photography camera with a spatial light modulator (SLM) as a programmable optical device based on the framework by Chen *et al.* [CHEL18].

Hybrid zooming. Zoom lenses take up a high percentage of modern photography lenses [Bar17]. They consist of multiple optical elements and are usually very bulky. The optical design [GBA08] provides variation in magnification and reduces aberrations while varying focal lengths. In computer vision, numerous super resolution algorithms [Sze10] were developed to achieve the zoom effect. In recent studies, image super-resolution has advanced with learning-based methods [WCH20]. Recasens *et al.* [RKS+18] introduced a salience-based distortion layer for a spatial sampling of input data. Lim *et al.* [LSK+17] remove unnecessary modules in conventional residual networks to achieve at least 4x upscaling with a compact design. Zhang *et al.* [ZLL+18] propose the very deep residual channel attention networks (RCAN) with a residual structure and an attention mechanism on the feature maps to learn high-frequency information for high accuracy. Zhang *et al.* [ZCNK19] operate on raw sensor data for super resolution while obtaining ground truth data via optical zoom. In mobile photography, we lack the space for a zoom lens with moving elements. Therefore, hybrid zoom systems with multiple cameras have been developed, e.g., using multiple aligned cameras with different field of views to zoom optically and electronically [GL13, MSL19]. Multiple smartphones from Light Co., Apple, Huawei, and Samsung with powerful zoom are available on the market. The energy consumption and spacing of the multi-camera design are bottlenecks due to the nature of this approach. Hardware with multiple cameras and beam splitters is also used for defocus video matting [MMP+05]. Another zoom system design without mechanical movement is called the solid-state zoom system [GBA08]. A minimum of two varifocal lenses are required. It has been designed with fluidic lenses [ZJL05, PCZZ07] and SLMs [MWR01, MWP+04].

An alternative approach to extend FoV without hardware changes is foveated imaging, which has pre-designed projection curves to capture multi-views [KKS+95, MWR01, TAG+17]. Carles *et al.* [CBW+17] propose a novel approach to foveated imaging by dual-aperture optics, which superimpose two images of different magnification on one sensor. Instead of blending multiple captures from single or multiple cameras, our method employs a single camera with two lenses and encodes the intermediate image plane, where the modulator redirects the chief ray of the second lens according to the

network-optimized warp functions.

6.3 Image Formation through Chief Ray Modulation

The chief rays are identified as the critical variable to control the FoV due to their association with distortion aberration to establish a programmable distortion map. To begin the imaging process, one must find where to insert the programmable optical device (SLM) in the imaging path. Using the Helmholtz reciprocity, we design the image formation from back to front. The first step is to replace the object plane of a pinhole camera (realized by a small aperture lens) with the SLM and register each SLM pixel to a chief ray. Afterward, the SLM refracts the chief ray to the object. A large aperture lens is put between the object and the SLM to increase the radiance and collect photons from a monitor at the object plane. The displayed images are from image datasets. In other words, the SLM is inserted into the object plane of the pinhole camera. The captured image is the modified intermediate image after the SLM. Thus the design of the image formation consists of two lenses (one with a large aperture and the other with a small one) and one SLM, such as shown in Figure 6.1.

In the forward image formation process, rays emitted from the objects are collected by the first lens (in the following text, we refer to it as the primary lens) then propagated to the SLM. The SLM redirects incoming light rays to align with the chief ray bundle of the pinhole camera. The SLM is composed of a thin liquid crystal pixel array on a silicon mirror. Thus it is treated as a thin lens. The chief rays are controlled by encoding image patterns to the SLM. The refractive indices are modified, which is equivalent to generating a freeform lens truncated to a fixed thickness. The image pattern, which is closely related to the distortion map, is parametrized and ready to be optimized through the neural network along with a decoder.

This hybrid zooming approach is not in the conventional sense of manipulating the magnification globally but generates and decodes modified images that have compressed or enlarged continual image fragments from various viewing angles. In other words, the camera produces a pixel-dependent viewing.

6.4 Validation of the Computational Camera

The first step to verify the imaging model is the simulation using a raytracer. We chose a succinct model by combining an SLM and a pinhole camera because it explicitly demonstrates the pixel-dependent FoV modulation. This model is equivalent to the setup in Figure 6.1 without the primary lens. We use a brute-force technique that traces one ray for each camera sensor pixel towards the plane object. The key role is played by the SLM, which redirects the chief rays by the encoded grayscale image. In practice, the SLM thickness is constant, but the refractive indices are altered; therefore, the image

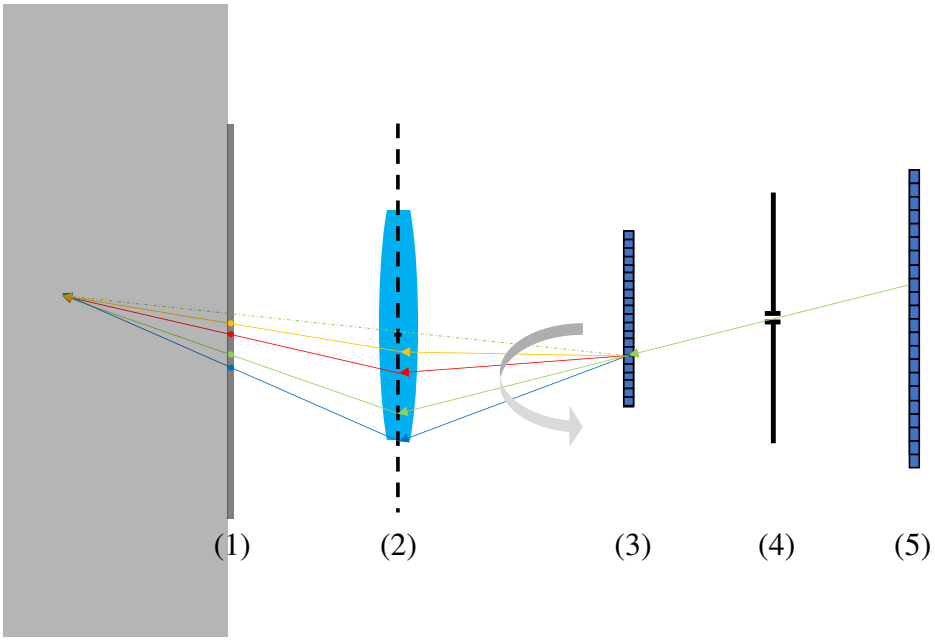


Figure 6.1: The scheme of image formation. (1) is the object plane. In the verification, a monitor is placed here. (2) is the primary lens. (3) is the LCoS SLM, whose pixels redirect the incoming light. (4) is a lens with a small aperture. (5) is the camera sensor. (4) and (5) combine as a pinhole camera. The ray tracing illustrates the image formation from the sensor to the object reversely. First, the intersection of chief rays and the sensor plane of the pinhole camera define the FoV. Each sensor pixel registers with maximally one SLM pixel. The camera's resolution should be higher than the SLM's to guarantee proper sampling. One chief ray without modulation is illustrated as the green ray. When the modulation is switched on, this ray is directed along a different outward direction of the same hemisphere (illustrated by other colors), thus reaching a guided object point. Due to the focus of the primary lens, all modulated rays pivot around the point on which the SLM Pixel is mapped. The pixel-dependent FoV is constructed by encoding an image to the SLM, with its image gradient prescribing the ray trajectories.

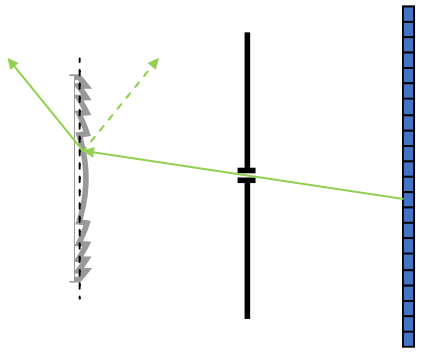


Figure 6.2: The scheme of the micro-mirror model of the SLM. The SLM is modeled as a programmable mirror. In the simulation, the ray is traced from the sensor across the perspective origin onto the SLM. The image gradient of the encoded pattern defines the pixel-wise surface normals. A reflected ray is then computed and flopped (from the ray represented by the dashed arrow) to point towards the object.

gradients encoded on the SLM determine the ray refraction. Since the vendor does not reveal the SLM thickness, we simulate the refraction by a reflective micro-mirror model shown in Figure 6.2. The thickness is set to be zero, and the refractive indices alter the surface normals. The reflected rays are reset to point towards the object. Since the angles are discrete, the discrepancy can be corrected by a look-up table. The surface normal is calculated from the image gradients of the SLM encoding as follows,

$$\vec{n} = (0,0,1) + (w \cdot g_y, 0, 0) + (0, w \cdot g_x, 0) \tag{6.1}$$

g_x and g_y are the image gradient value of horizontal and vertical direction, respectively, and w is the modulation factor, manually chosen as 0.0005. The new ray is then calculated by specular reflection,

$$\vec{r} = (0,0,-1) \circ (\vec{d} - 2(\vec{d} \cdot \vec{n})\vec{n}) \tag{6.2}$$

$\circ$ is the Hadamard product, the vector $(0,0,-1)$ indicates the refraction instead of reflection.

The rendering function of the ray tracer is shown in Algorithm 14. We first initialize the camera parameters and load the SLM image and the object image. Then perspective projection is defined for each chief ray by the camera pixel, the perspective center, and the pixel location. Then, the SLM intersection and this ray are calculated to identify the SLM pixel and modulation value. Afterward, the ray is redirected by SLM modulation and traced towards a guided object point. It is noteworthy that the refracted ray has a new perspective origin. The color is then assigned according to the camera pixel index.

We show simulated results of the central region of the ISO 12233 chart, a checkerboard

```
1  Function Render():
2      load the radiance map of object plane ;
3      initialize camera parameters, e.g. resolution, focal length, sensor pitch, center
         of perspective or original point O ;
4      load SLM image pattern ;                    // strength of modification
5      enter the factor of modulation ;
6      for camera pixels do
7          compute chief ray by sensor pixel location and O ;
8          compute intersection of ray and SLM ;
9          modify the ray direction by the SLM pixel ;
10         find the perspective origin of the new direction ;
11         compute intersection of redirected ray and the object ;
12         record color ;
13     end
14     return rendered image ;
```

Algorithm 1: Ray-tracing of the computational camera with an SLM

pattern, and one test image by Asuni and Giachetti [AG14] in Figure 6.4 to 6.9. We encode 8-bit grayscale images with the following patterns: low and high-frequency sine wave patterns, a Fresnel lens pattern, a pattern with randomly generated patches, and an axicon glass pattern.

The simulated results shown in Figure 6.4 to 6.9 confirm the model that pixel-dependent FoV modulation can be generated through programmable optics. According to the SLM image gradients, compared with the non-modulated results and each SLM pattern, the pixel colors of the results are sampled individually. This pixel-wise operation generates a global distortion with a user-defined map. This map can realize unconventional distortion, such as the axicon pattern that compresses the center and magnifies the marginal; or the sine wave pattern that enlarges targeted rows. A traditional magnification effect can be produced by encoding the Fresnel lens pattern shown in Figure 6.7, whose image gradient is linear except the ring effect generated by the strong viewing angle of the sharp edges. This effect is further illustrated in the result with the random patch pattern. The results vindicate the usefulness of the phase-only SLM that, in effect, guide the chief ray purposefully. They demonstrate the critical capacity to alter the ray-object intersection per pixel. In the future, the optical encoding task of hybrid-zooming is to learn the distortion map through a data-driven approach. The simulation limit is the discrepancy between the micro-mirror model and the real-world SLM, which is to be calibrated by a look-up table.

The next step is to set up a pinhole camera. We adopt a 50 mm lens and close the aperture to F22, which is approximately a 2.27 mm pinhole. The lens is combined with a monochromatic camera body (FLIR Oryx 10GigE ORX-10G-12S6M-C). We employ

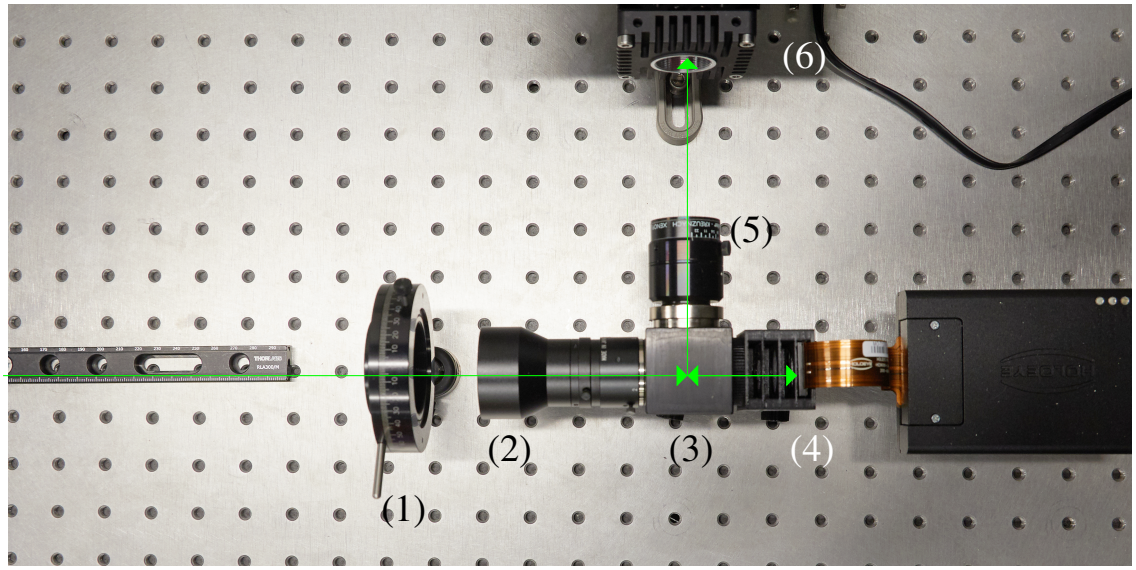


Figure 6.3: The camera prototype. (1) is a linear polarizer that enables the SLM. (2) is the primary imaging lens with 100 mm focal length. (3) is the beamsplitter which has C-mountings for the two lenses. (4) is the phase-only SLM. (5) is the lens with a small aperture (around 2.27 mm). (6) is the image sensor. (5) and (6) combine as a pinhole camera.

a 100 mm telephoto lens with an F1.8 aperture to collect a sufficient amount of photons. As is shown in Figure 6.3, both lenses are mounted on a beam splitter, which has four C-mount sockets. The HOLOEYE PLUTO SLM, operated by an external driver and server to load the look-up table and image, follows directly after the imaging path of the primary lens and bounces light back to the beam splitter.

We install a macro-photography setup, with an object located 140 mm away from the frontal lens surface. To obtain the ground truth is the key for verifying the network in the future, yet this is not without difficulties in this specified setup to measure the radiance of a broader or magnified (even more difficult) view. Furthermore, the correct projection needs to be calibrated. We use an LCD as an object with the image data as prior knowledge to overcome these drawbacks.

The real-world results shown in Figure 6.10 to 6.13 are consistent with the idea and the simulation, where the distortion matches the SLM image. For example, in Figure 6.11b, one observes the magnification of the third row (from top), and in Figure 6.12, the strong compression of the rows demonstrates the view-guiding by the SLM pattern. The result with random patches in Figure 6.13 has generated view distortions similar to the simulated in Figure 6.8. Note that the brightness difference is due to post-processing leveling, not noise caused by the SLM. The results limit is the vignetting, where the pinhole camera captures only part of the SLM. The non-linearity in the actual scenario should also be investigated.

6.5 Conclusion and Outlook

This study presented a computational camera platform to realize a programmable pixel-dependent field of view (FoV). We aim to investigate a computational camera for a deep learning-based hybrid zooming approach. Such a camera would preserve actual optical signal while using only one camera with non-moving optical components, which overcomes the trade-off between energy costs (i.e., methods using multiple cameras) and fidelity of image data.

The findings of this research highlight a novel computational camera platform for hybrid zooming through controlling the pixel-dependent viewing direction by the SLM's image pattern.

This work gives an account of the chief ray modulation scheme where the pixels of the programmable optics register the chief rays of the imaging camera. We built a computational camera consisting of a phase-only SLM, a photographic lens, and a pinhole camera. We have demonstrated the results of both simulation and real-world experiments, which indicate successful control of pixel-wise viewing. The camera ensures the programmability of the FoV for deep learning methods.

The next step is to construct an autoencoder network that consists of a differential optical renderer and a decoder. The formulation of the matrix operations can be described as follows,

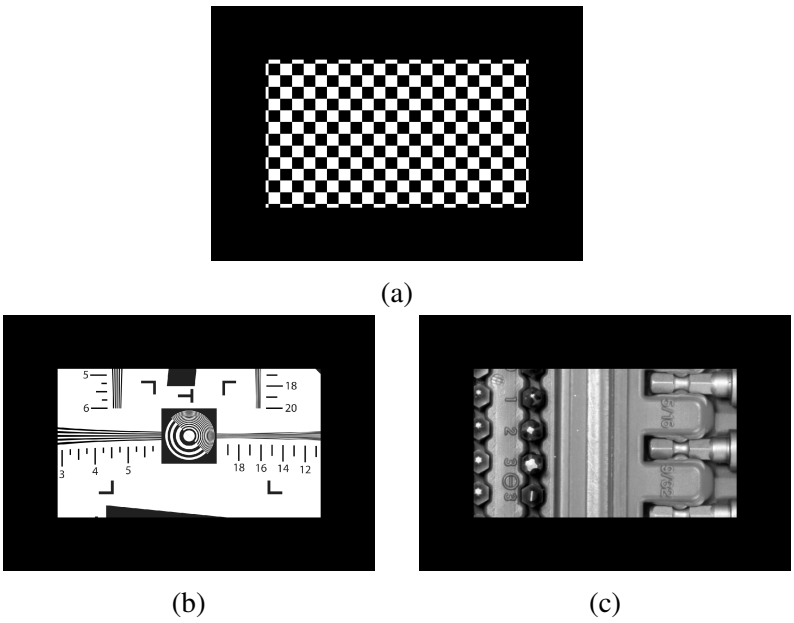
$$\mathbf{c} = \mathbf{P}(l(\mathbf{I})\mathbf{a}) \tag{6.3}$$

where $\mathbf{a}$ is the object, $\mathbf{I}$ is the encoded image on the SLM to be learned, $l(\cdot)$ is the conversion function of the SLM pattern to light transport matrix, $\mathbf{P}$ is the image formation by the pinhole camera, and $\mathbf{c}$ is the captured image.

The forward model can be implemented as the encoder and a decoder that solves the inverse problem to recover the intended view. A monitor can display image datasets, and the network can be trained to learn pre-designed FoVs, which could lead to non-line-of-axis zooming due to its flexibility.

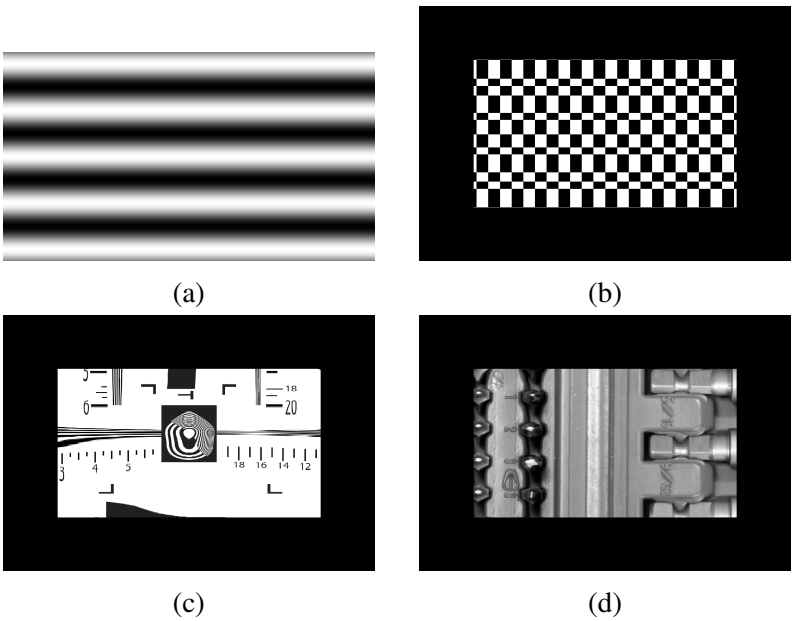
Though the training must be limited to a photo laboratory, it is more interesting to test the camera in outdoor scenarios, especially for landscape photography, opthalmological photography, or remote sensing. With the advancement of SLM technology, we anticipate the emergence of phase-only transmissive SLMs with strong modulation power. These SLMs can be a substitute for the current one, thus further reducing the camera's size.

The present study has not examined the non-linearity of view modulation shown in real-world experiments such as Figure 6.12a. The chromatic aberrations caused by the SLM will be investigated in the future too. The captured images only show a part of the entire SLM. One possible explanation is the strong vignetting due to the lens' and beam splitter's dimensions. The findings might not be transferable to 3D objects located at a close distance. However, this limitation is trivial in landscape photography or remote sensing, in which the camera working distance is long.



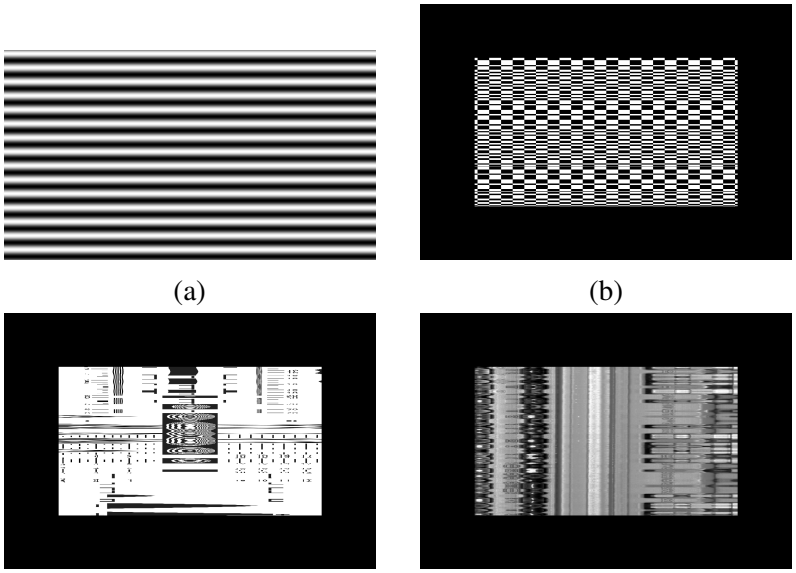
(a) (b) (c)

Figure 6.4: Raytraced images without modulation. The views of (a) is the checkerboard object; and (b) is from the ISO 12233 chart; and (c) is a test image by Asuni and Giachetti [AG14].



(a) (b) (c) (d)

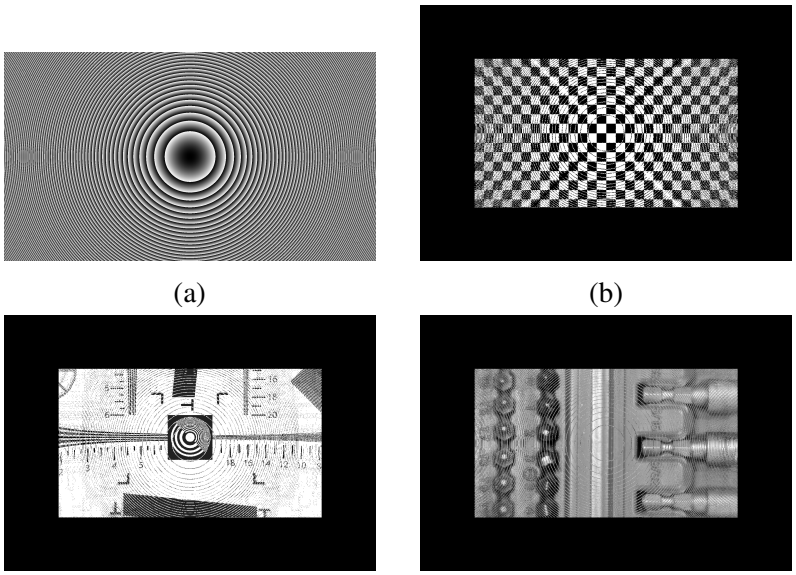
Figure 6.5: Raytraced images with a low-frequency sine wave modulation. (a) is the sine wave pattern loaded on the SLM. (b), (c), and (d) are the rendered images.



(a) (b)

(c) (d)

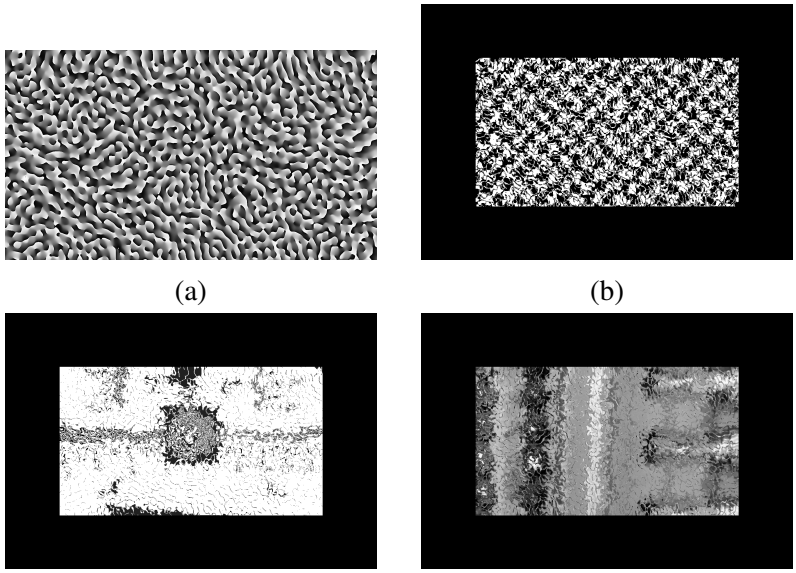
Figure 6.6: Raytraced images with a high-frequency sine wave modulation. (a) is the sine wave pattern loaded on the SLM. (b), (c), and (c) are the rendered images.



(a) (b)

(c) (d)

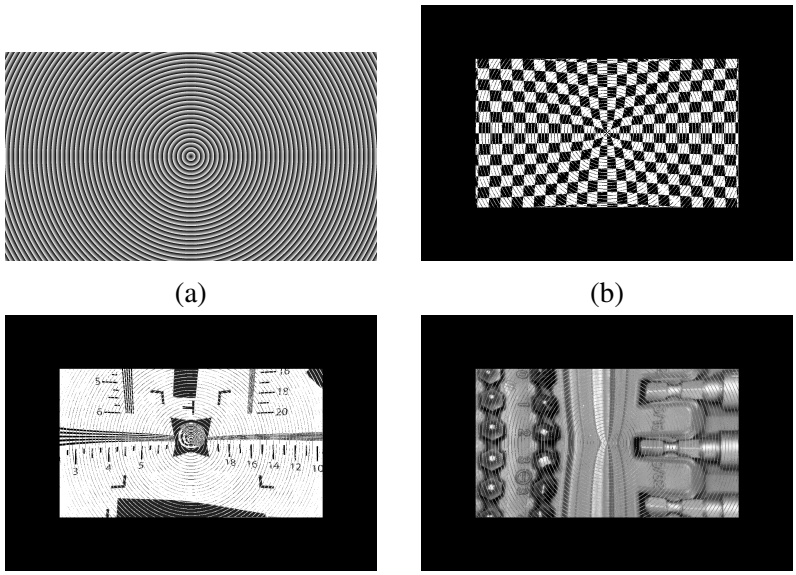
Figure 6.7: Raytraced images with a Fresnel lens pattern modulation. (a) is the Fresnel lens pattern loaded on the SLM. (b), (c), and (d) are the rendered images.



(a) (b)

(c) (d)

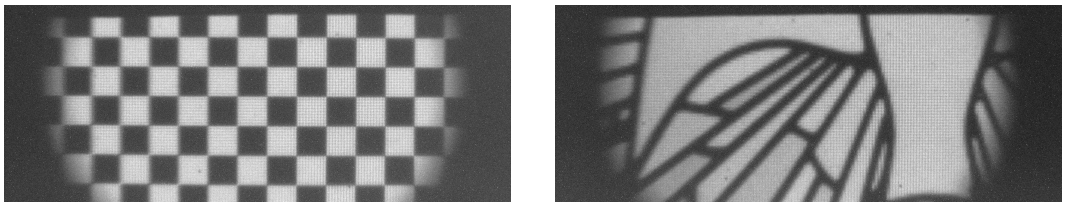
Figure 6.8: Raytraced images with a pattern of randomly distributed patches. (a) is the image pattern loaded on the SLM. (b), (c), and (d) are the rendered images.



(a) (b)

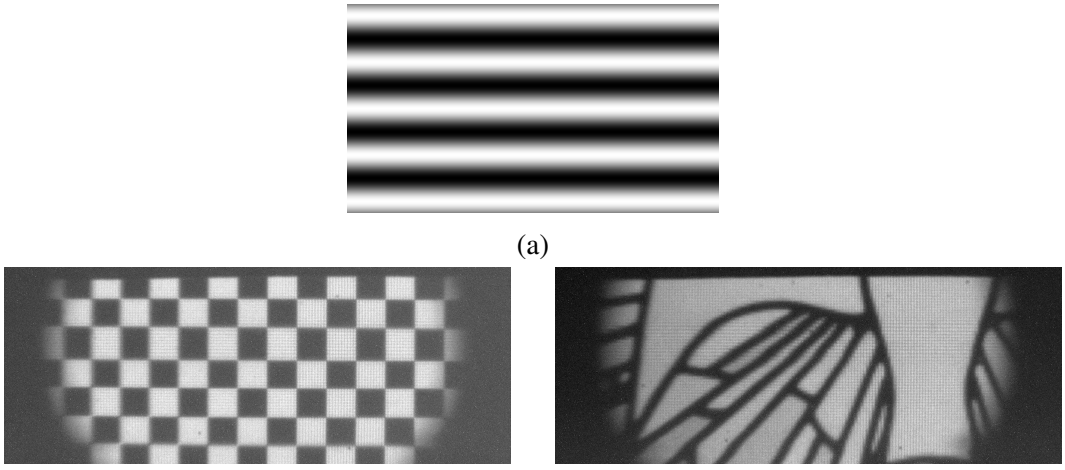
(c) (d)

Figure 6.9: Raytraced images with an axicon pattern. (a) is the image pattern loaded on the SLM. (b), (c), and (d) are the rendered images.



(a) (b)

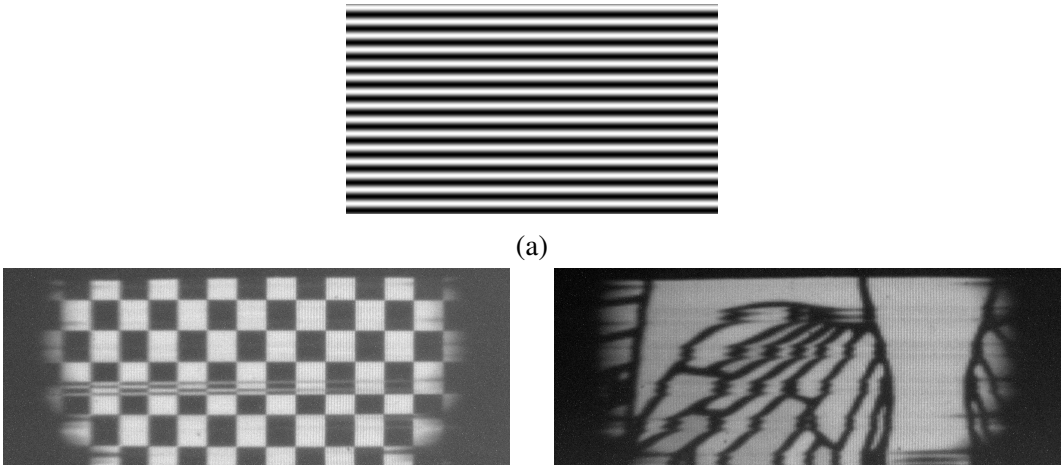
Figure 6.10: Real-world images without modulation. A mobile phone display is located at the object plane. (a) shows a checker pattern. (b) is a crop of a stained glass picture.



(a)

(b) (c)

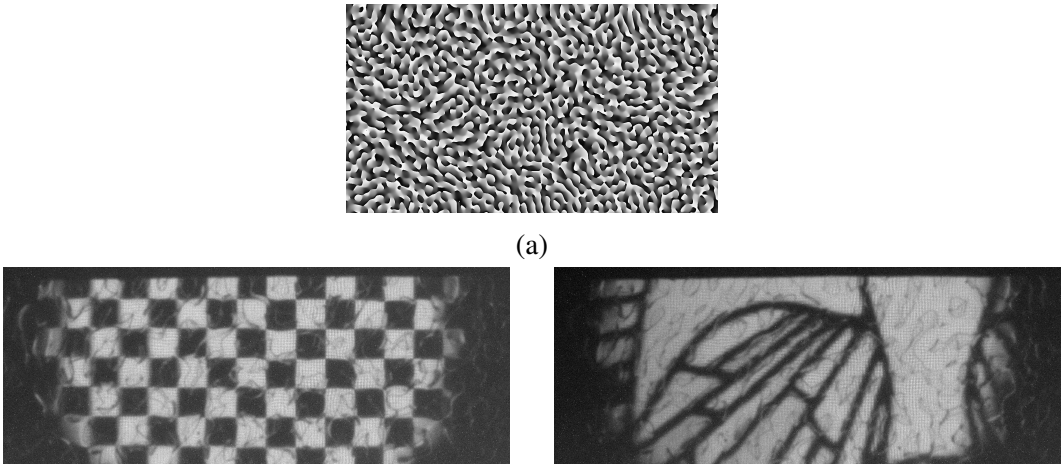
Figure 6.11: Real-world images with a low-frequency sine wave modulation. This is a counterpart of the simulated images in Figure 6.5. (a) is the sine wave pattern loaded on the SLM. (b) and (c) are the rendered images.



(a)

(b) (c)

Figure 6.12: Real-world images with a high-frequency sine wave modulation. (a) is the pattern loaded on the SLM. (b) and (c) are the rendered images.



(a)

(b) (c)

Figure 6.13: Real-world images with a pattern with randomly distributed patches. (a) is the pattern loaded on the SLM. (b) and (c) are the rendered images.

Chapter 7

Conclusion and Outlook

In the imaging chain, optics and digital processing have much in common. With the emergence of programmable optics and the development of computational photography, a framework that conjoins both modules is promising for energy-saving and efficient solutions through sharing critical information and opens the way for algorithmic development. This thesis designs and establishes a computational photography platform that integrates programmable optics to design optics and image processing for various photographic applications cooperatively.

The thesis covers the joint-design framework's conception, the construction of computational camera hardware, and algorithmic solutions. We saw in Chapter 2 that the dependency of image formation and digital processing is the foundation for modeling the framework's programmability. We identify the information link through light propagation, interpreted by both wave optics and ray optics. The process of image formation is characterized by the point spread functions (PSFs) in wave optics, and the field of view is defined by the chief rays in ray optics. We utilize a spatial light modulator (SLM) to integrate programmable optics into the camera system. This scheme facilitates optical encoding that avails optical information overseen by traditional cameras. The image reconstruction inverses the optical procedures with regularizers or using convolution neural networks (CNNs). The formulation outlines a joint-design paradigm that can incorporate both physical parameters and image processing.

The initial step is to prove the feasibility of integrating the programmable optics in a computational camera. Chapter 3 uses a phase-coded aperture setup and the Gerchberg-Saxton algorithm to produce complex-shaped user-defined PSF. The stability of PSF engineering is verified in the temporal domain. In Chapter 4, optical phase PSF engineering as the driving idea of the computational camera is implemented for spectral imaging. In traditional optical imaging, chromatic aberration degrades image quality. Thus, it was seen to be suppressed. However, this is under the presumption that per-pixel spectra cannot be detected. We achieve spectral variant PSFs and spectrally encoded images with the pixel-wise spectra spread on the sensor plane with our computational camera. An inverse solver is built to take up the image formation with single-channel and cross-channel regularizers to restore multiple spectral images from a single shot.

We observe that the heuristic design of the PSFs is potent. However, the optimal de-

signs are to be found. The obvious next step is to have a general pipeline that utilizes the feedback from the image processing to assist the PSF design. Therefore, we use a learning-based approach to adopt abundant image data for a joint end-to-end framework. The encoder is an optical renderer that simulates the image formation by light propagation from a plane object to the sensor. The encoder is tuned to match the real-world computational camera. The decoder is a U-Net that receives the intermediate image and a fixed PSF for all distances to restore the target images. To validate the expediency of the method, we test this architecture on an extended depth of field (EDoF) application, both in simulation and in the actual scenario.

One bottleneck in mobile photography is the lack of cost-effective hybrid zooming. Pure digital zooming is powerful, but it fabricates radiance data. Optical zooming, on the other hand, is bulky and costly. Multi-camera photography as a compelling alternative solution can triple the energy cost. Chapter 6 forms a camera setup that employs the SLM to programmatically guide the camera's chief ray distribution, consequently encoding the field of view. We justify the camera design with a raytracer and compare it with actual image acquisition. Future research will continue towards an end-to-end solution based on this setup.

The proposed methods are an attempt to extend the horizons of the computational camera through programmable optics. As mentioned above, the aim is imaging efficiency, energy reduction, and novel applications through jointly controlling parameters of both optics and processing. It will be intriguing to see how the design methodology and mechanism are used in consumer photography, remote sensing, or ophthalmological device.

Some future projects could be snapshot spectral imaging without knowledge of PSFs, differential renderers with non-paraxial diffraction, verification of computational cameras in outdoor scenarios, and lensless imaging using Ptychography. In addition, the discrepancy between the camera's physical properties and the digital encoding must be reduced for practical applications. Meanwhile, we hope to see the advancement of higher resolution phase-modulating SLMs for more dynamic optical modulation.

To conclude, the paradigm shift brought by computational photography has re-oriented imaging towards image data acquisition and restoration. Therefore, programmable optics and their integration in the camera are becoming a keystone of imaging.

Appendix A

Supplementary Materials for Multispectral Imaging from Single-shot

In this supplementary material, we provide the design model of the point spread function (PSF) engineering, and validation of our reconstruction model. For the design model, we compare three different PSF designs with respect to their dispersion powers and frequency domain analysis. For the reconstruction, we show the synthetic twelve channel reconstruction in detail as well as different real-world reconstruction with various PSFs.

A.1 Design Model of Spatial and Spectral Variant PSFs

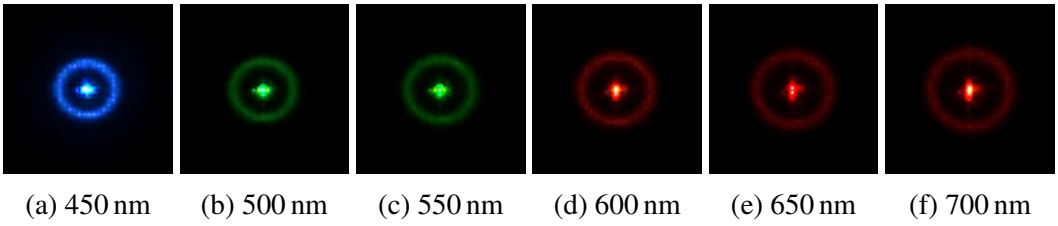
We analyze three different user-designed PSFs (ring, dots and spiral) used in Chapter 4 for real-world scene capture. The key ingredient of PSF design is described in Section 3.3. Firstly, in order to maximize spectral resolution, PSFs of each spectral bands should be discriminable. We formulate the dispersion power as in Equation (A.1), where the intensity of each PSF is weighted by the distance to its recent neighbors.

$$D = \frac{1}{N}\sum_{2}^{N-1}\frac{\|\mathbf{I}_n\|_1}{\|\mathbf{I}_{n+1}-\mathbf{I}_n\|_1+\|\mathbf{I}_n-\mathbf{I}_{n-1}\|_1} \tag{A.1}$$

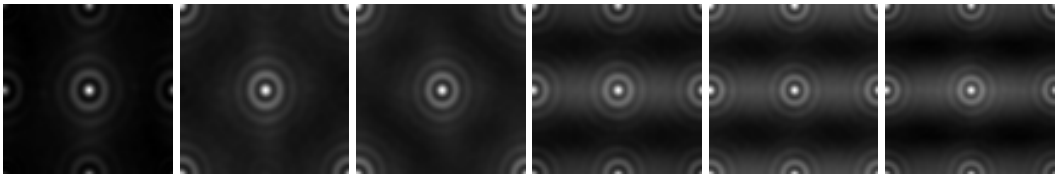
where D indicates the dispersion power, N is the total number of spectral bands. $\mathbf{I}_n$ is the PSF image at band n for $n \in [2, N-1]$. Secondly, PSFs should have similar frequency coverage across channels, therefore we analyze each PSF in Fourier space. Particularly, we verify the significance of isotropic sampling of PSFs. We validate the PSF designs through reconstruction of real-world data. The multispectral PSFs and their Fourier transforms are shown in Figure A.1 to A.3. The corresponding dispersion powers are presented in Table A.1.

Table A.1: Dispersion power of ring, dots and spiral PSFs

PSFs	Ring	Dots	Spiral
Dispersion power	20.2007	7.6048	8.1108

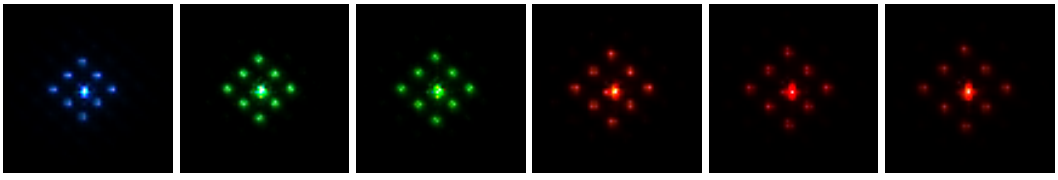


(a) 450 nm (b) 500 nm (c) 550 nm (d) 600 nm (e) 650 nm (f) 700 nm

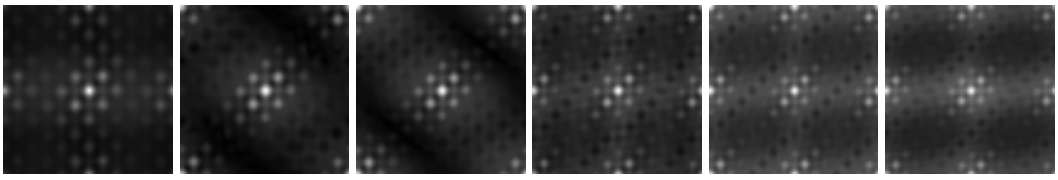


(g) 450 nm (h) 500 nm (i) 550 nm (j) 600 nm (k) 650 nm (l) 700 nm

Figure A.1: Spatial and spectral variant *ring* PSFs. Illustrations (a) to (f) are PSFs in spatial domain, and (g) to (l) are the corresponding Fourier transforms.

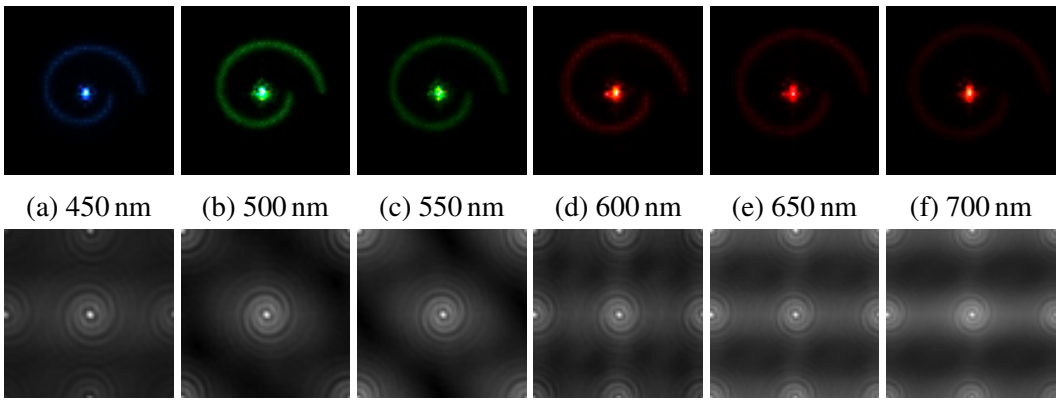


(a) 450 nm (b) 500 nm (c) 550 nm (d) 600 nm (e) 650 nm (f) 700 nm



(g) 450 nm (h) 500 nm (i) 550 nm (j) 600 nm (k) 650 nm (l) 500 nm

Figure A.2: Spatial and spectral variant *dots* PSFs. Illustrations (a) to (f) are PSFs in spatial domain, and (g) to (l) are the corresponding Fourier transforms.



(a) 450 nm (b) 500 nm (c) 550 nm (d) 600 nm (e) 650 nm (f) 700 nm

(g) 450 nm (h) 500 nm (i) 550 nm (j) 600 nm (k) 650 nm (l) 700 nm

Figure A.3: Spatial and spectral variant *spiral* PSFs. Illustrations (a) to (f) are PSFs in spatial domain, and (g) to (l) are the corresponding Fourier transforms.

A.2 Reconstruction

Our reconstruction pipeline inverts the image formation by data fitting using single-channel and cross-channel priors. The data fitting contains the mosaicing operator and the convolution kernels.

The convolution kernels are measured according to the PSFs, which limits the number of spectral bands and governs the reconstruction quality. We show additional materials to prove that our method is able to reconstruct twelve channel spectral images. We also provide the six channel reconstruction results of different real-world scenes using various PSFs.

A.2.1 Synthetic Twelve Channel Reconstruction

In our reconstruction, the number of restored spectral images are limited by the amount of measured spectral PSFs. We validate the reconstruction of twelve spectral bands using synthetic data. Firstly, ring PSFs are simulated with twelve spectral bands. Then we generate PSF-modulated snapshot images according to Equation 2. Following our optimization, the results are shown in Figure 4 to 6 using public available data [YMIN10], and in Figure 7 using our captured ColorChecker data. The corresponding average PSNRs are shown in the Table A.2.

A.2.2 Real-world Reconstruction with Various PSFs

We demonstrate reconstruction of various scenes such as ColorChecker, lemon, Lego, and cloth, with different PSF modulation. In Figure A.8 to A.18, we provide reconstructed spectral images with their ground truths as well as the restored RGB snapshots

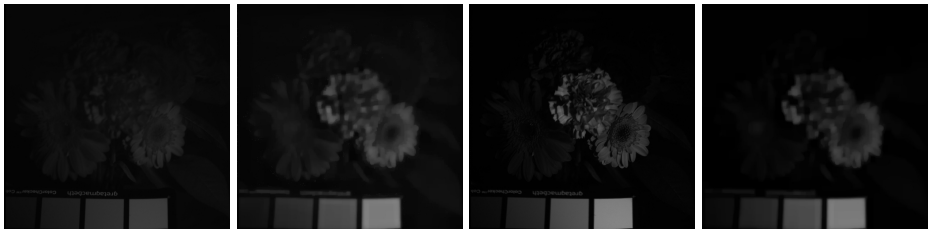
Table A.2: Dispersion power of ring, dots, and spiral PSFs

Scene	Flowers	Glass tiles	Paints	ColorChecker
PSNR	29.0181	23.8982	23.5897	35.4812

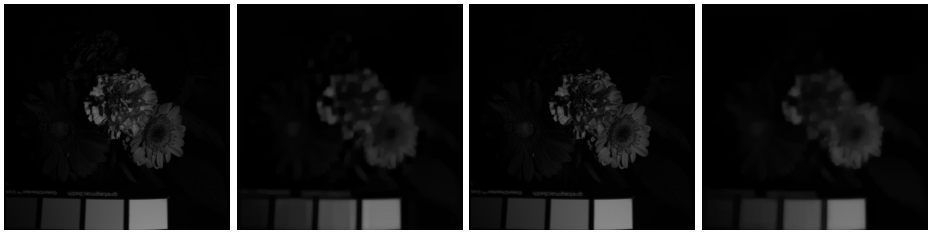
from the spectral images. The average full-resolution PSNRs of the spectral images are shown in Table A.3.

Table A.3: Average PSNRs of ColorChecker, lemon, lego and cloth with different PSFs

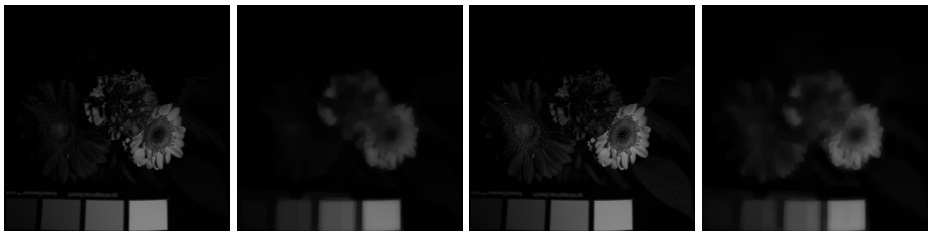
	ColorChecker	Lemon	Lego	Cloth
Ring	21.5473	28.1851	22.9578	18.4300
Spiral	22.0312	21.7892	22.4835	19.0417
Dots	22.2183	23.8086	21.3271	18.7232



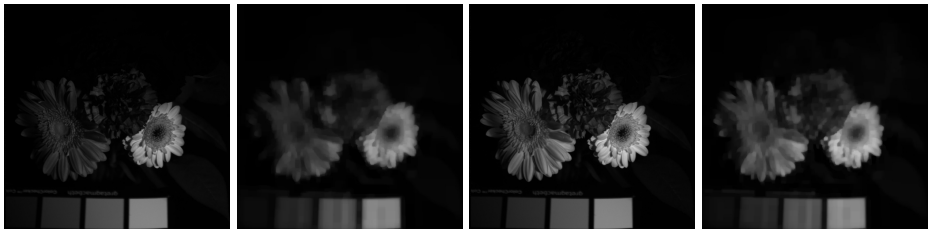
(a) 410 nm GT (b) 410 nm RS (c) 440 nm GT (d) 440 nm RS



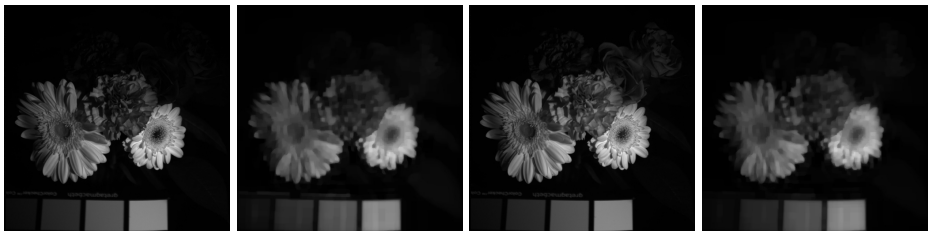
(e) 470 nm GT (f) 470 nm RS (g) 500 nm GT (h) 500 nm RS



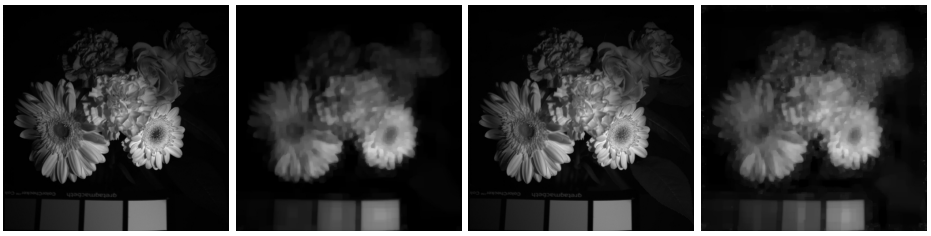
(i) 530 nm GT (j) 530 nm RS (k) 560 nm GT (l) 560 nm RS



(m) 580 nm GT (n) 580 nm RS (o) 610 nm GT (p) 610 nm RS

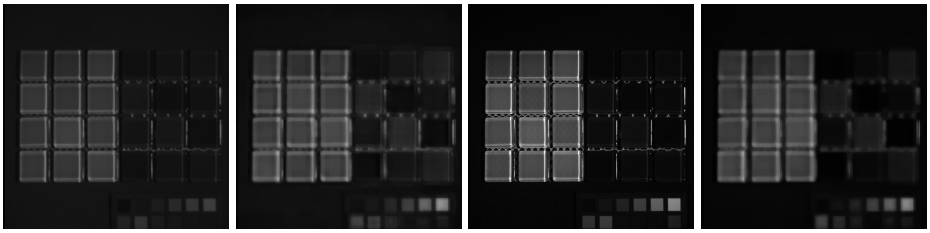


(q) 640 nm GT (r) 640 nm RS (s) 670 nm GT (t) 670 nm RS

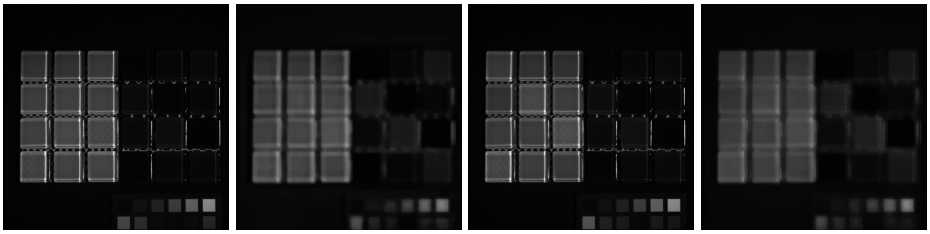


(u) 700 nm GT (v) 700 nm RS (w) 720 nm GT (x) 720 nm RS

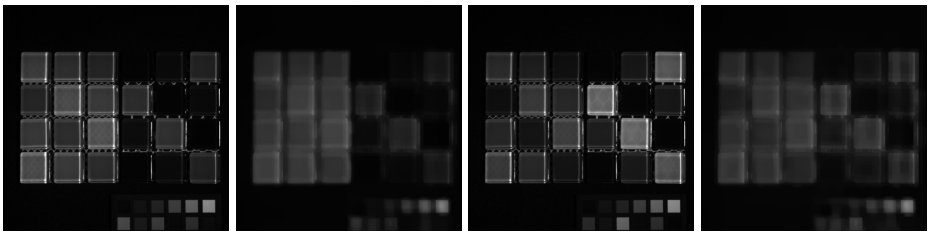
Figure A.4: Simulated reconstruction of 12 channel spectral images. GT stands for ground truth. RS means results. Ground truth data is from the work of Yasuma *et al.* [YMIN10].



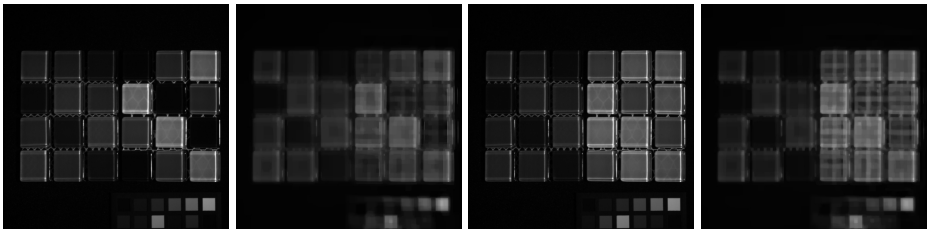
(a) 410 nm GT (b) 410 nm RS (c) 440 nm GT (d) 440 nm RS



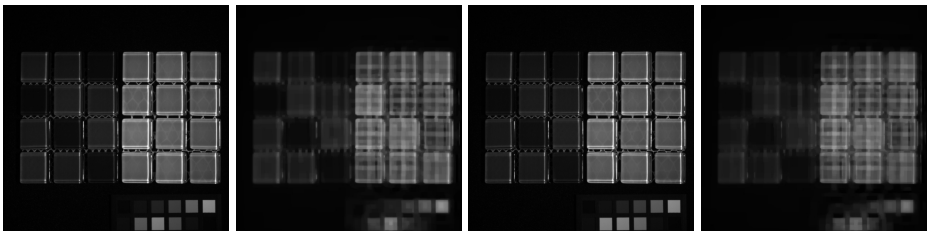
(e) 470 nm GT (f) 470 nm RS (g) 500 nm GT (h) 500 nm RS



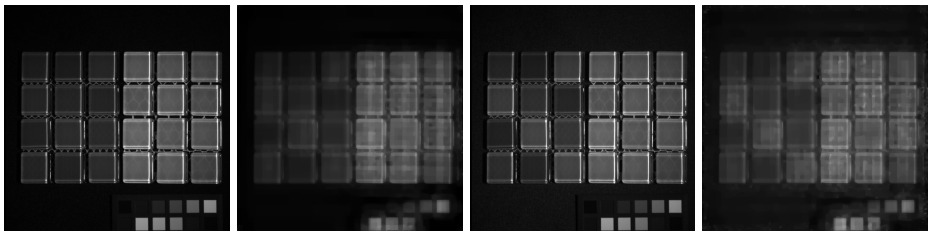
(i) 530 nm GT (j) 530 nm RS (k) 560 nm GT (l) 560 nm RS



(m) 580 nm GT (n) 580 nm RS (o) 610 nm GT (p) 610 nm RS

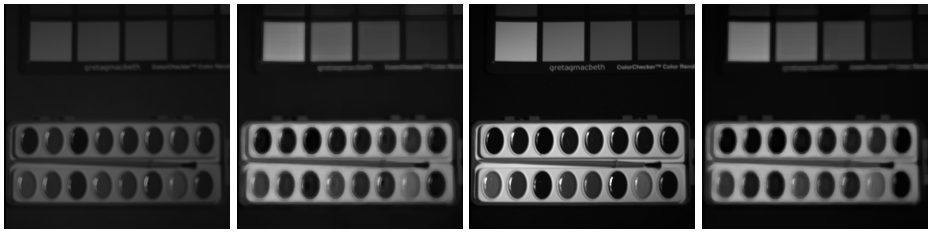


(q) 640 nm GT (r) 640 nm RS (s) 670 nm GT (t) 670 nm RS

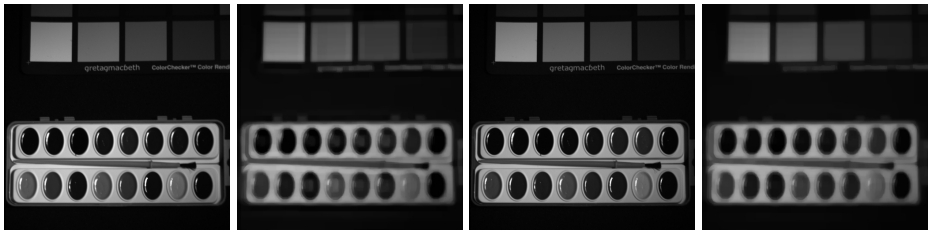


(u) 700 nm GT (v) 700 nm RS (w) 720 nmGT (x) 720 nm RS

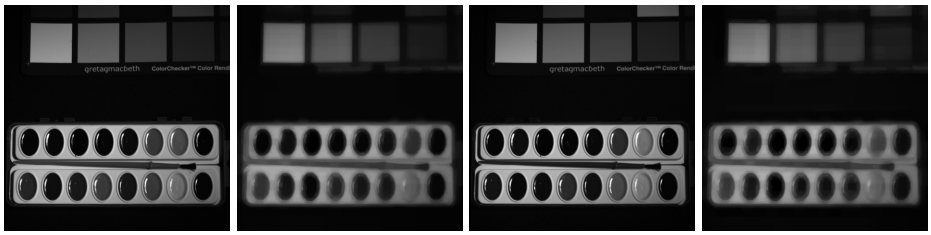
Figure A.5: Simulated reconstruction of 12 channel spectral images. GT stands for ground truth. RS means results. Ground truth data is from the work of Yasuma *et al.* [YMIN10].



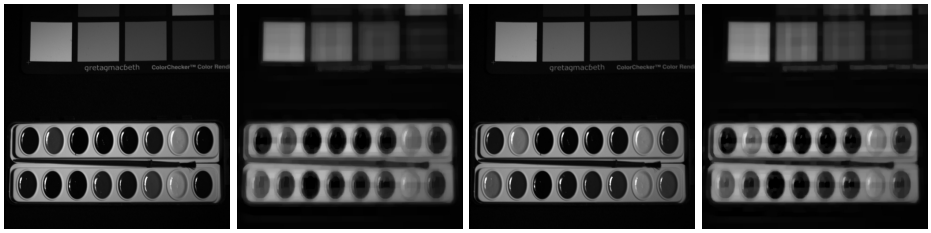
(a) 410 nm GT (b) 410 nm RS (c) 440 nm GT (d) 440 nm RS



(e) 470 nm GT (f) 470 nm RS (g) 500 nm GT (h) 500 nm RS



(i) 530 nm GT (j) 530 nm RS (k) 560 nm GT (l) 560 nm RS

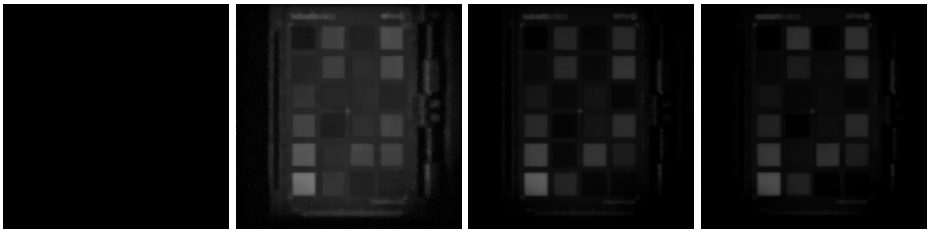


(m) 580 nm GT (n) 580 nm RS (o) 610 nm GT (p) 610 nm RS

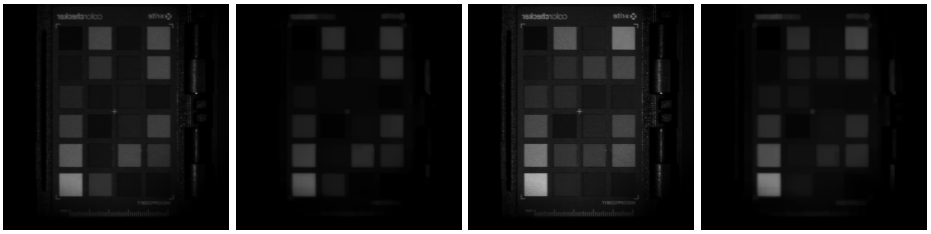
(q) 640 nm GT (r) 640 nm RS (s) 670 nm GT (t) 670 nm RS

(u) 700 nm GT (v) 700 nm RS (w) 720 nm GT (x) 720 nm RS

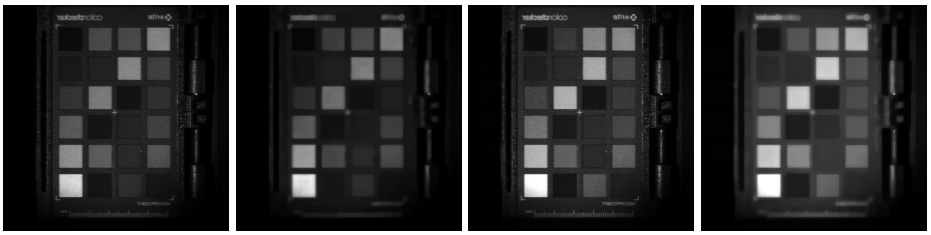
Figure A.6: Simulated reconstruction of 12 channel spectral images. GT stands for ground truth. RS means results. Ground truth data is from the work of Yasuma *et al.* [YMIN10].



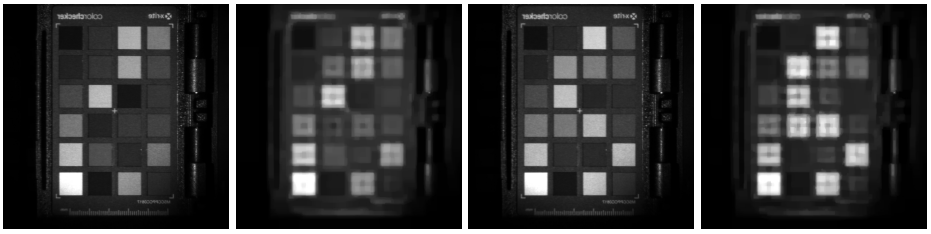
(a) 410 nm GT (b) 410 nm RS (c) 440 nm GT (d) 440 nm RS



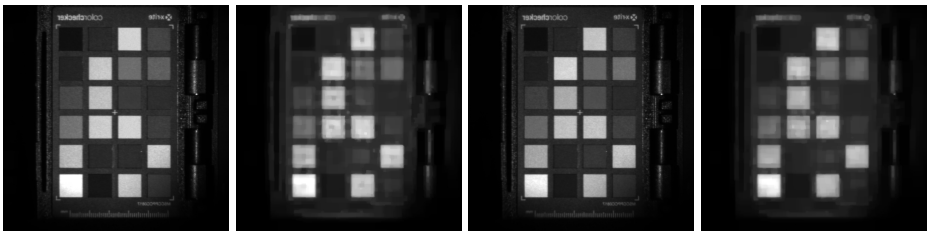
(e) 470 nm GT (f) 470 nm RS (g) 500 nm GT (h) 500 nm RS



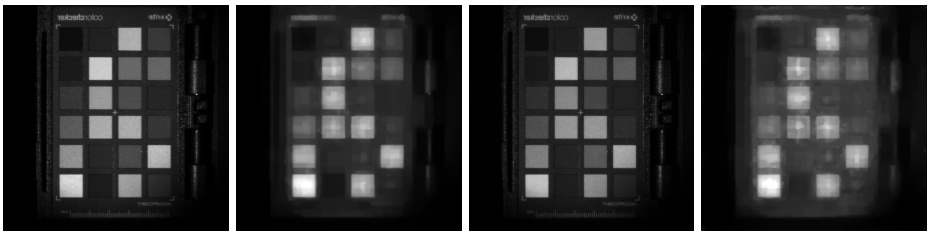
(i) 530 nm GT (j) 530 nm RS (k) 560 nm GT (l) 560 nm RS



(m) 580 nm GT (n) 580 nm RS (o) 610 nm GT (p) 610 nm RS



(q) 640 nm GT (r) 640 nm RS (s) 670 nm GT (t) 670 nm RS



(u) 700 nm GT (v) 700 nm RS (w) 720 nm GT (x) 720 nm RS

Figure A.7: Simulated reconstruction of 12 channel spectral images. GT stands for ground truth. RS means results. Ground truths are captured by our computational camera.

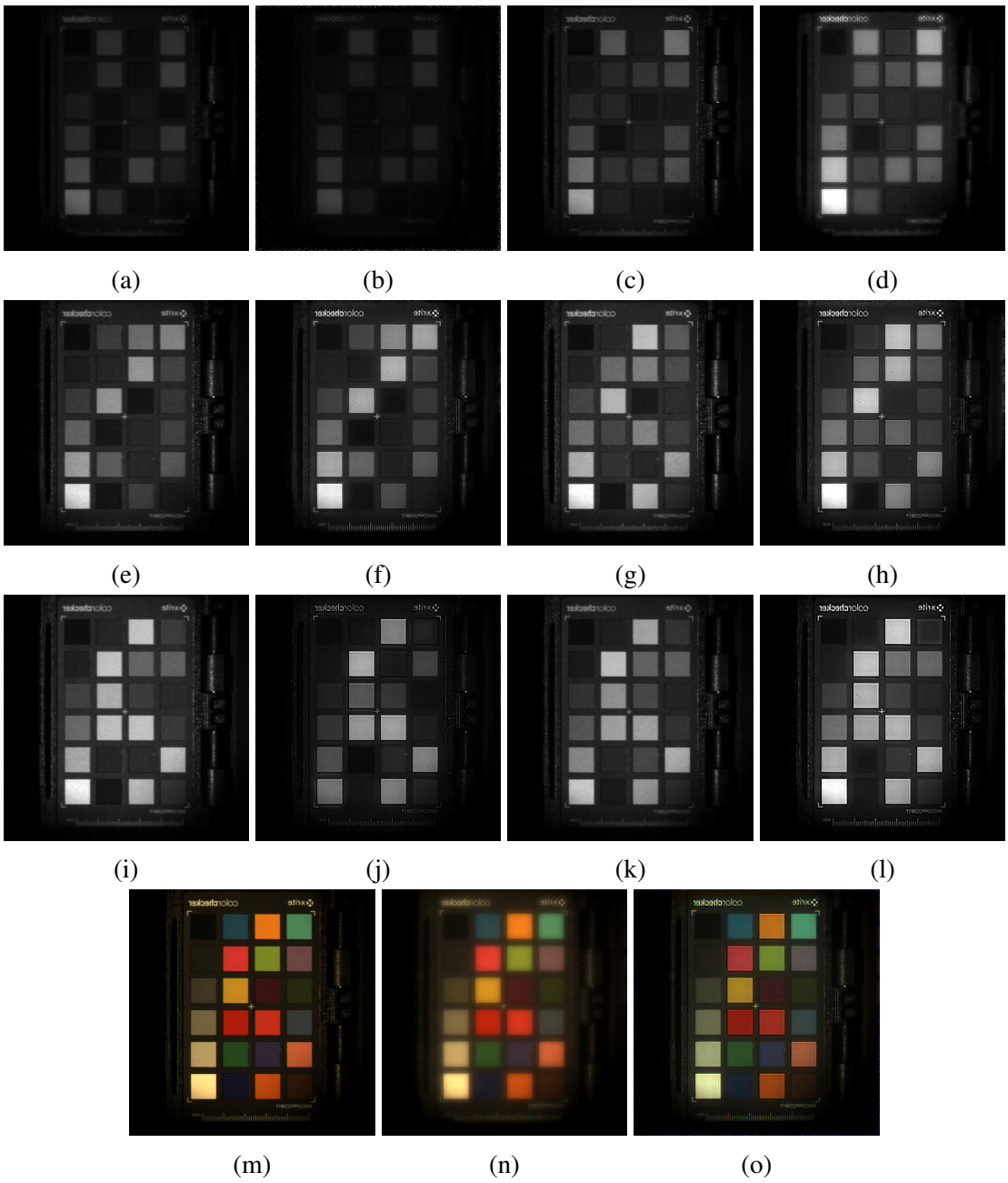


Figure A.8: Reconstruction from snapshot with *dots* PSFs of real-world ColorChecker scene. GT stands for ground truth. RS means results. (a) to (l) are comparisons of individual spectral images: (a) 450 nm GT, (b) 450 nm RS, (c) 500 nm GT, (d) 500 nm RS, (e) 550 nm GT, (f) 550 nm RS, (g) 600 nm GT, (h) 600 nm RS, (i) 650 nm GT, (j) 650 nm RS, (k) 700 nm GT, (l) 700 nm RS, and (m) is the ground truth single shot image without PSF modulation. (n) is the snapshot with designed PSFs. (o) is the restored RGB image from spectral information.

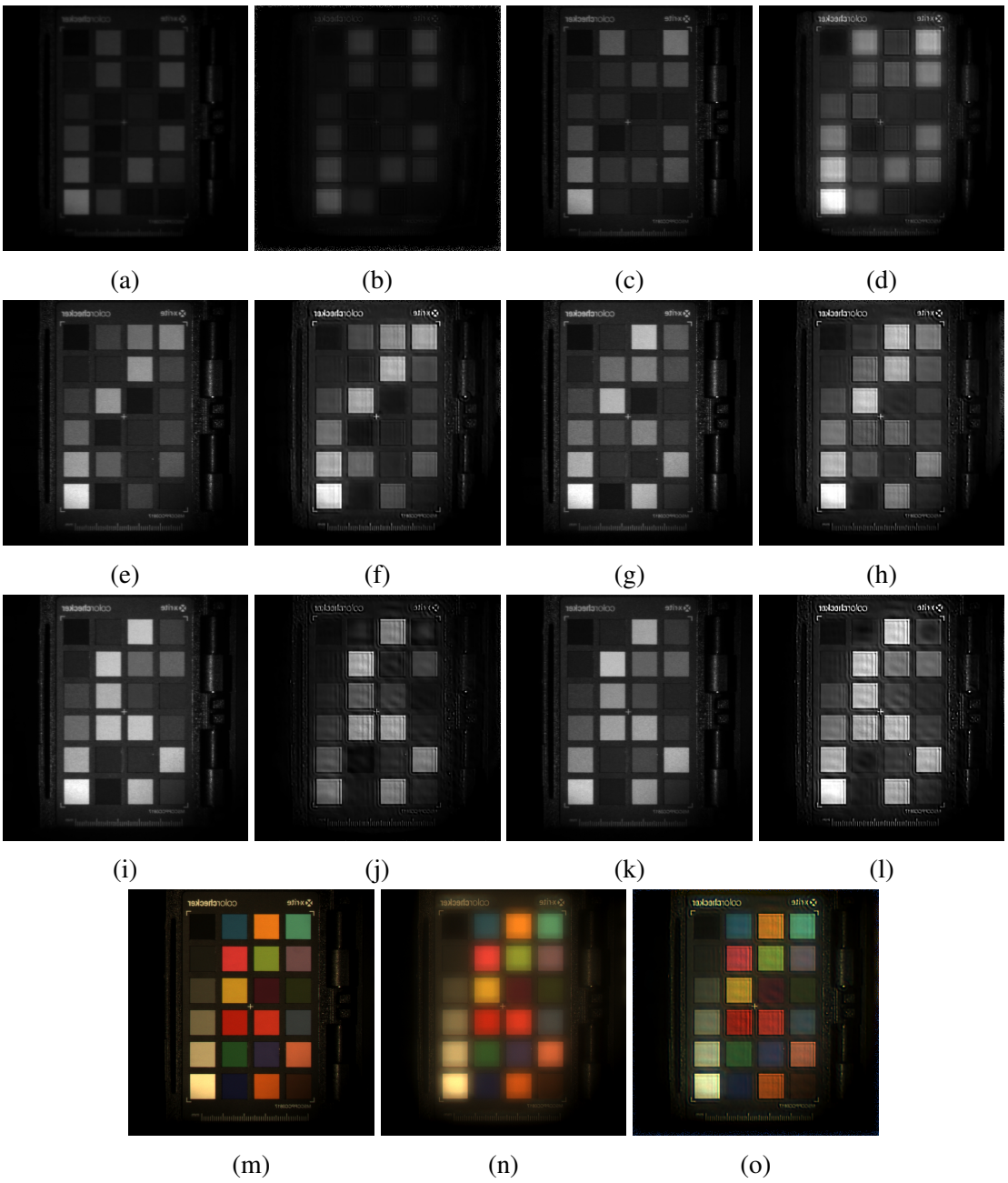


Figure A.9: Reconstruction from snapshot with *spiral* PSFs of real-world ColorChecker scene. GT and RS stand for ground truth and results. Illustrations (a) to (l) are comparisons of individual spectral images: (a) 450 nm GT, (b) 450 nm RS, (c) 500 nm GT, (d) 500 nm RS, (e) 550 nm GT, (f) 550 nm RS, (g) 600 nm GT, (h) 600 nm RS, (i) 650 nm GT, (j) 650 nm RS, (k) 700 nm GT, (l) 700 nm RS, and (m) is the ground truth single shot image without PSF modulation. (n) is the snapshot with designed PSFs. (o) is the restored RGB image from spectral information.

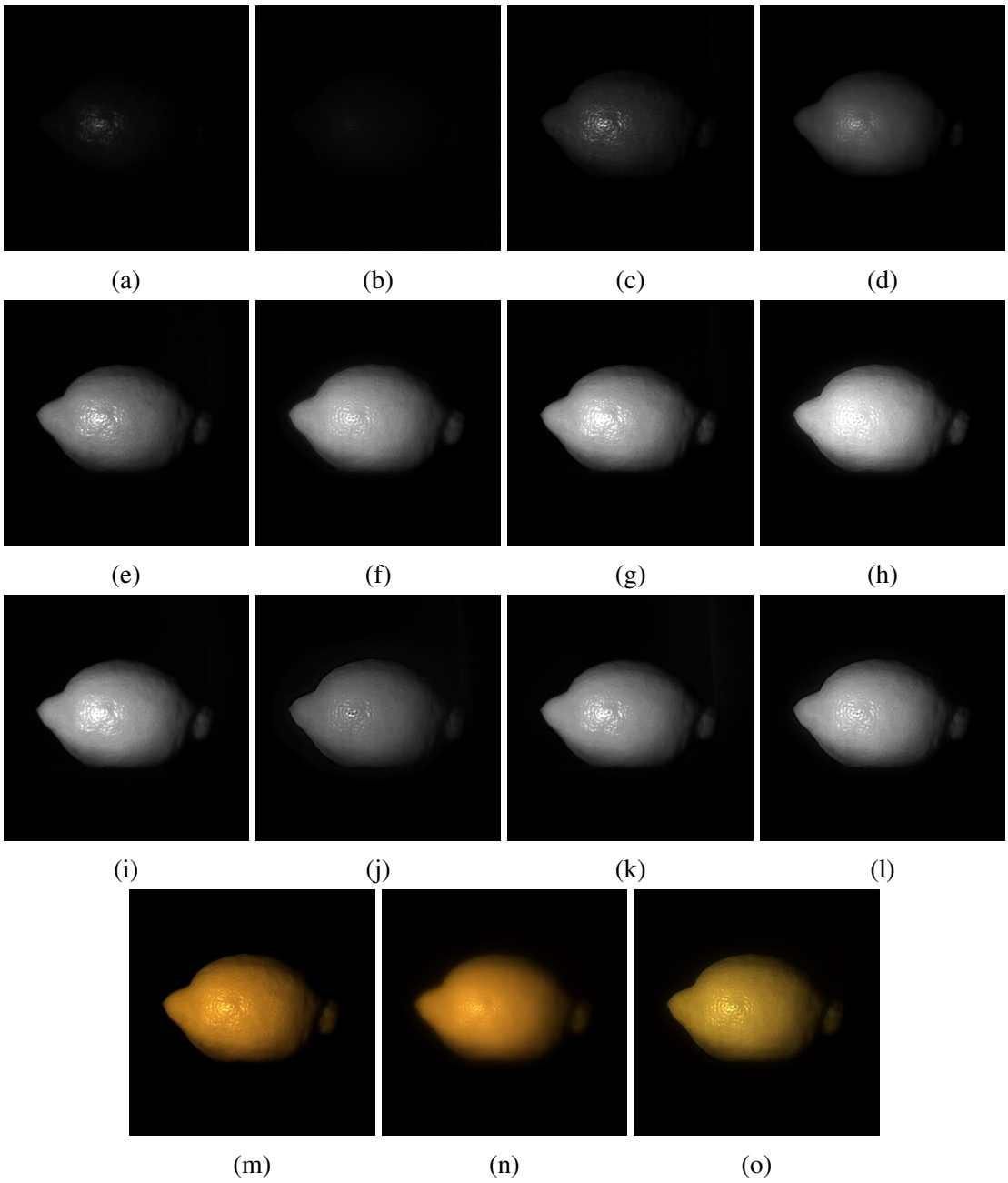


Figure A.10: Reconstruction from snapshot with *ring* PSFs of real-world lemon scene. GT stands for ground truth. RS means results. Illustrations (a) to (l) are comparisons of individual spectral images: (a) 450 nm GT, (b) 450 nm RS, (c) 500 nm GT, (d) 500 nm RS, (e) 550 nm GT, (f) 550 nm RS, (g) 600 nm GT, (h) 600 nm RS, (i) 650 nm GT, (j) 650 nm RS, (k) 700 nm GT, (l) 700 nm RS, and (m) is the ground truth single shot image without PSF modulation. (n) is the snapshot with designed PSFs. (o) is the restored RGB image from spectral information.

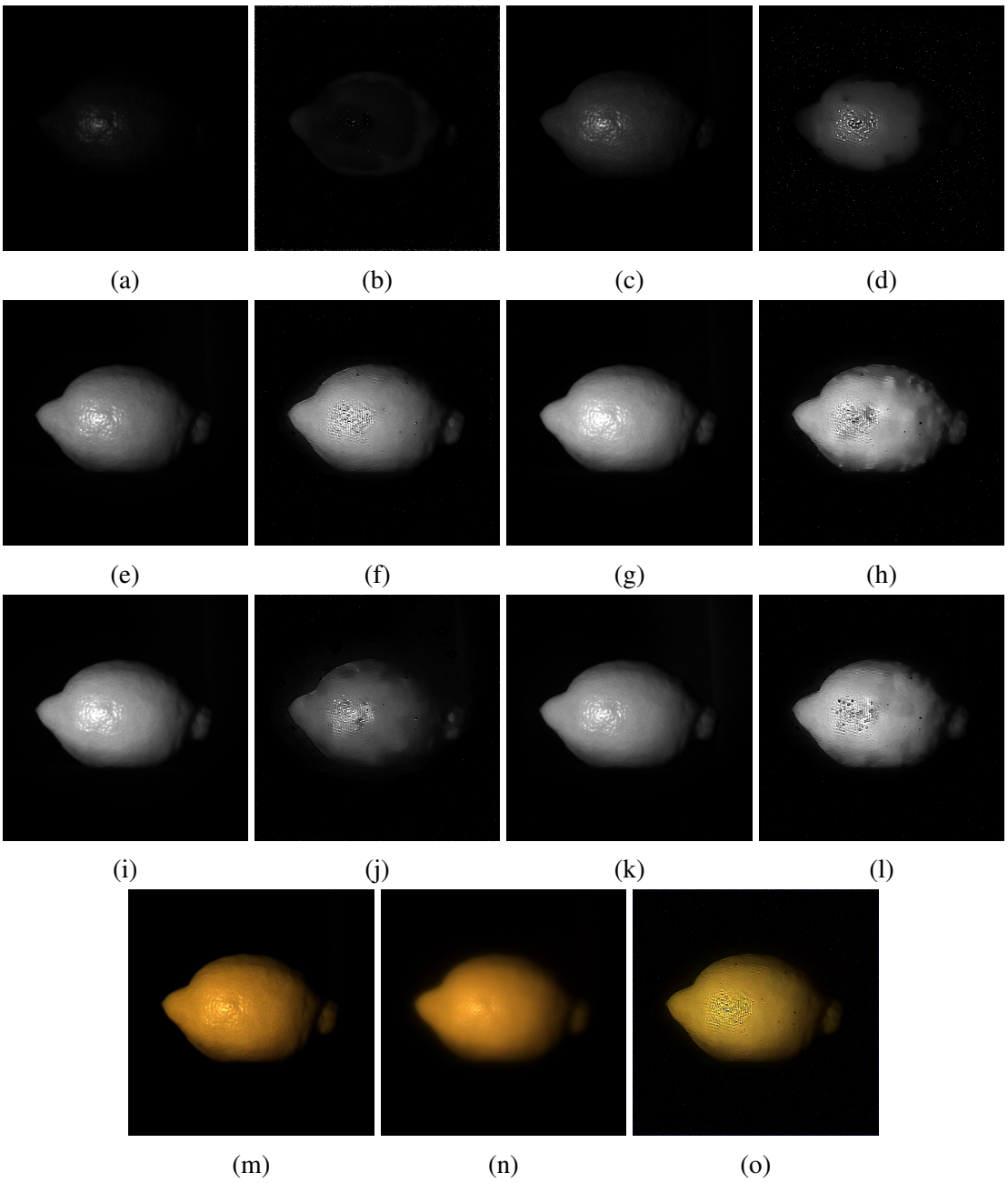


Figure A.11: Reconstruction from snapshot with *dots* PSFs of real-world lemon scene. GT stands for ground truth. RS means results. Illustrations (a) to (l) are comparisons of individual spectral images: (a) 450 nm GT, (b) 450 nm RS, (c) 500 nm GT, (d) 500 nm RS, (e) 550 nm GT, (f) 550 nm RS, (g) 600 nm GT, (h) 600 nm RS, (i) 650 nm GT, (j) 650 nm RS, (k) 700 nm GT, (l) 700 nm RS, and (m) is the ground truth single shot image without PSF modulation. (n) is the snapshot with designed PSFs. (o) is the restored RGB image from spectral information.

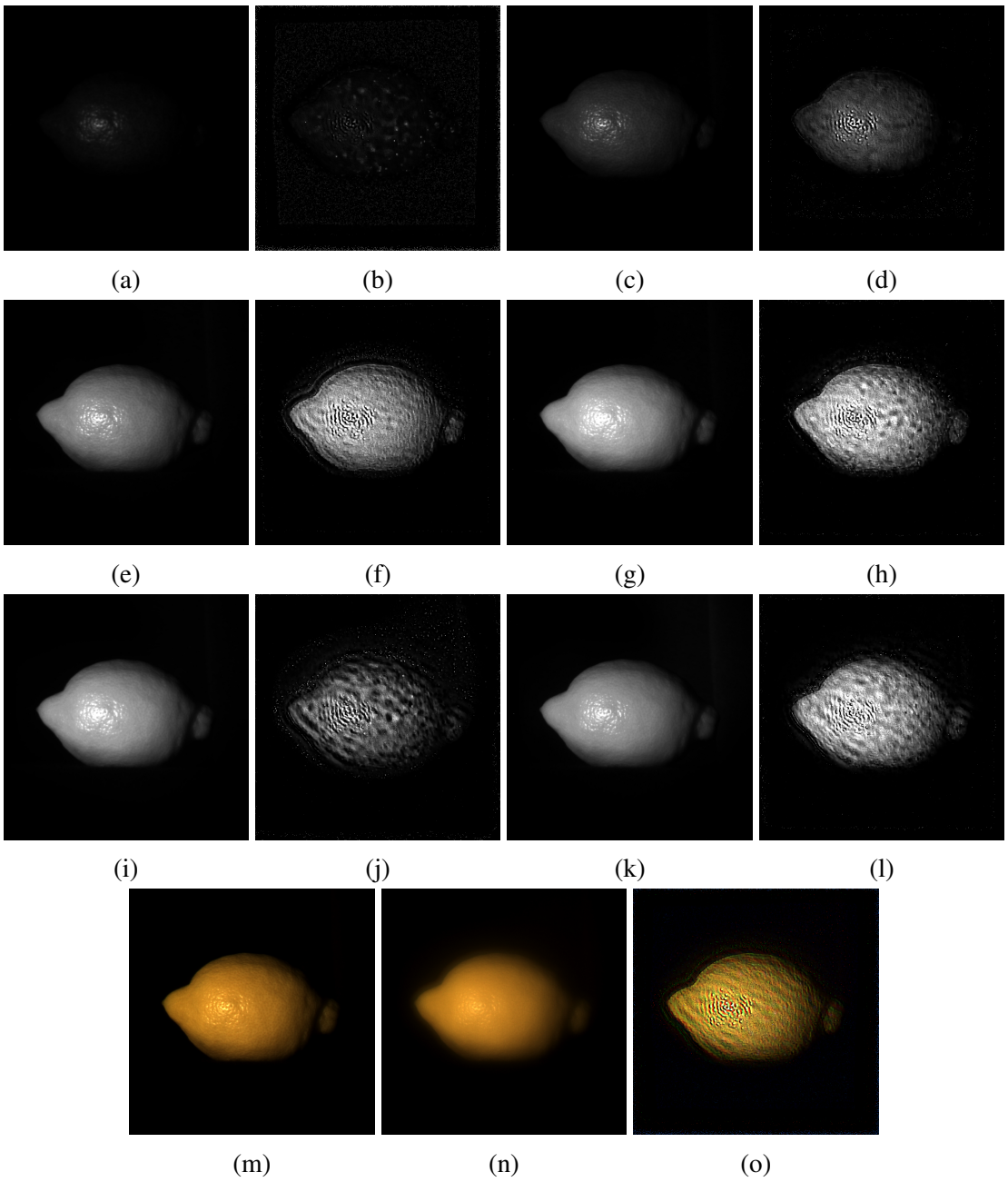


Figure A.12: Reconstruction from snapshot with *spiral* PSFs of real-world lemon scene. GT stands for ground truth. RS means results. Illustrations (a) to (l) are comparisons of individual spectral images: (a) 450 nm GT, (b) 450 nm RS, (c) 500 nm GT, (d) 500 nm RS, (e) 550 nm GT, (f) 550 nm RS, (g) 600 nm GT, (h) 600 nm RS, (i) 650 nm GT, (j) 650 nm RS, (k) 700 nm GT, (l) 700 nm RS, and (m) is the ground truth single shot image without PSF modulation. (n) is the snapshot with designed PSFs. (o) is the restored RGB image from spectral information.

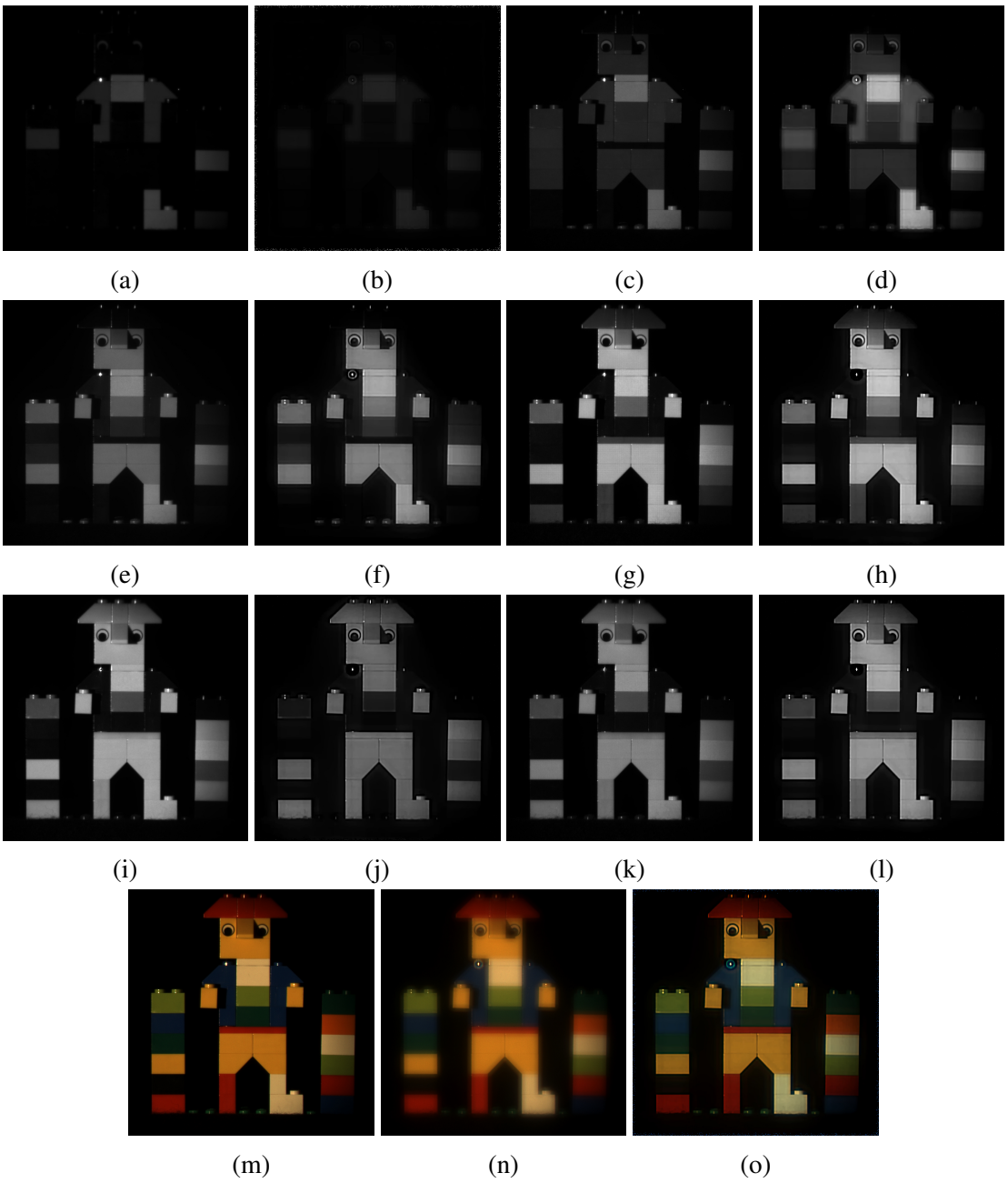


Figure A.13: Reconstruction from snapshot with *ring* PSFs of real-world lego scene. GT stands for ground truth. RS means results. Illustrations (a) to (l) are comparisons of individual spectral images: (a) 450 nm GT, (b) 450 nm RS, (c) 500 nm GT, (d) 500 nm RS, (e) 550 nm GT, (f) 550 nm RS, (g) 600 nm GT, (h) 600 nm RS, (i) 650 nm GT, (j) 650 nm RS, (k) 700 nm GT, (l) 700 nm RS, and (m) is the ground truth single shot image without PSF modulation. (n) is the snapshot with designed PSFs. (o) is the restored RGB image from spectral information.

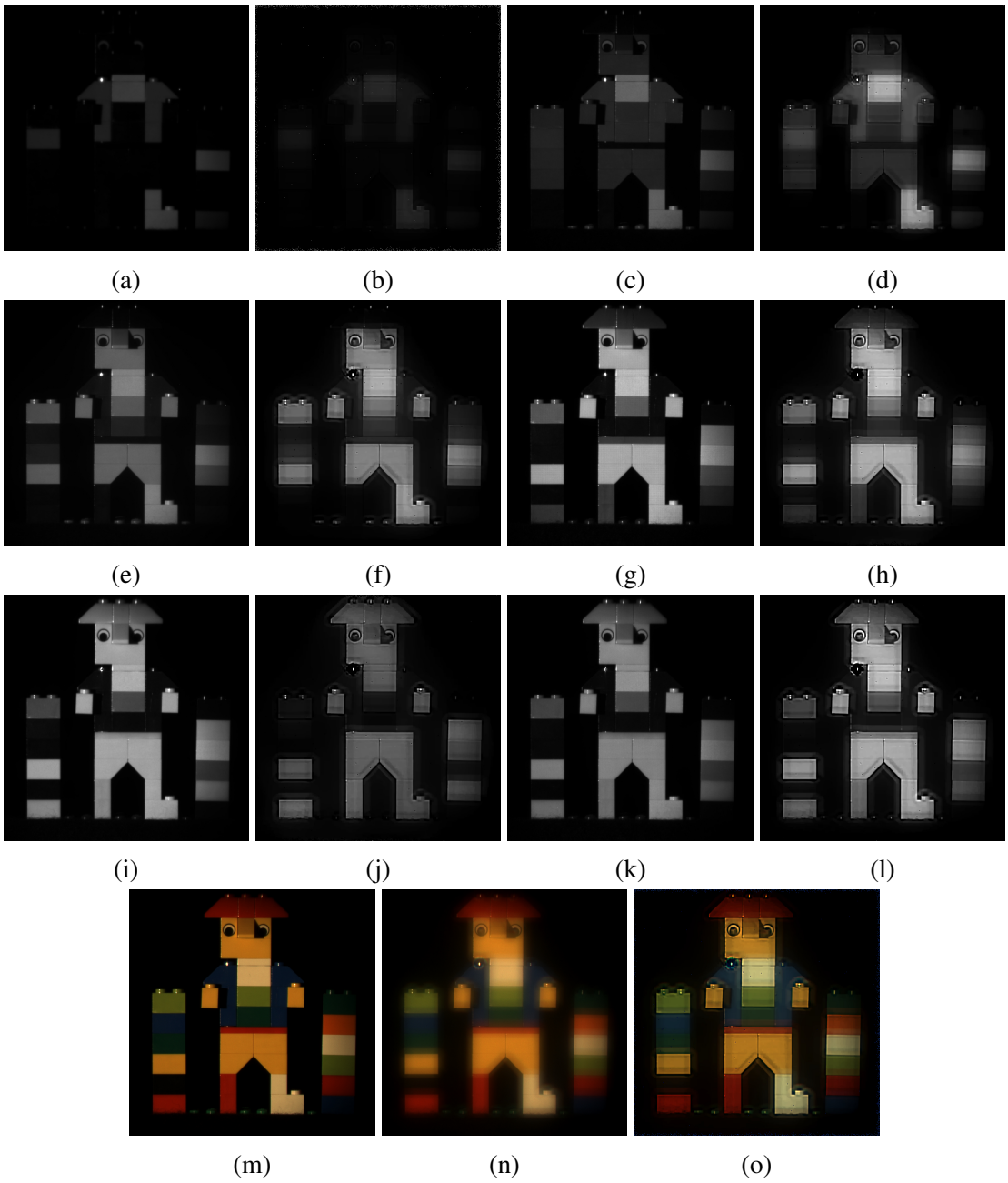


Figure A.14: Reconstruction from snapshot with *dots* PSFs of real-world lego scene. GT stands for ground truth. RS means results. Illustrations (a) to (l) are comparisons of individual spectral images: (a) 450 nm GT, (b) 450 nm RS, (c) 500 nm GT, (d) 500 nm RS, (e) 550 nm GT, (f) 550 nm RS, (g) 600 nm GT, (h) 600 nm RS, (i) 650 nm GT, (j) 650 nm RS, (k) 700 nm GT, (l) 700 nm RS, and (m) is the ground truth single shot image without PSF modulation. (n) is the snapshot with designed PSFs. (o) is the restored RGB image from spectral information.

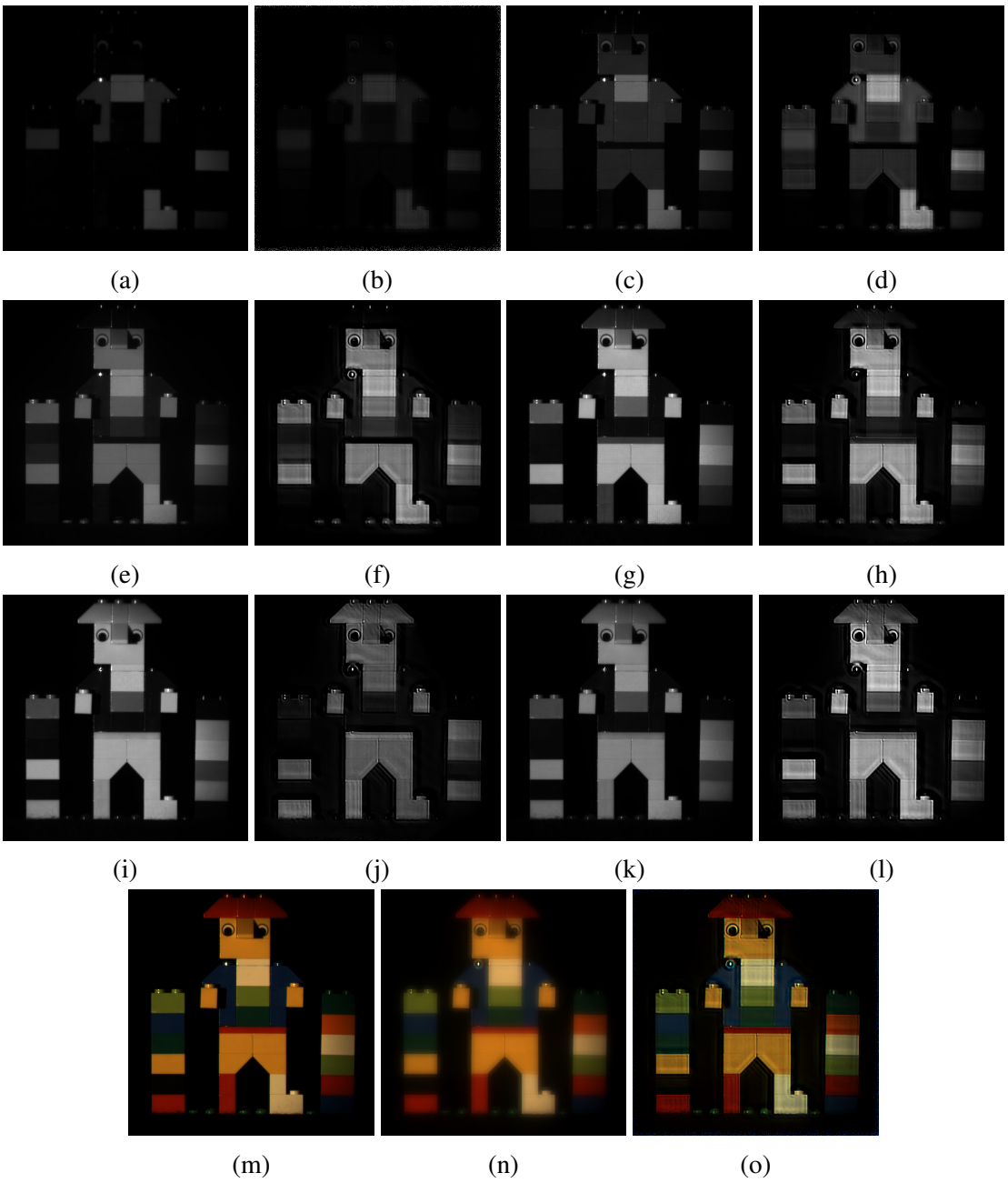


Figure A.15: Reconstruction from snapshot with *spiral* PSFs of real-world lego scene. GT stands for ground truth. RS means results. Illustrations (a) to (l) are comparisons of individual spectral images: (a) 450 nm GT, (b) 450 nm RS, (c) 500 nm GT, (d) 500 nm RS, (e) 550 nm GT, (f) 550 nm RS, (g) 600 nm GT, (h) 600 nm RS, (i) 650 nm GT, (j) 650 nm RS, (k) 700 nm GT, (l) 700 nm RS, and (m) is the ground truth single shot image without PSF modulation. (n) is the snapshot with designed PSFs. (o) is the restored RGB image from spectral information.

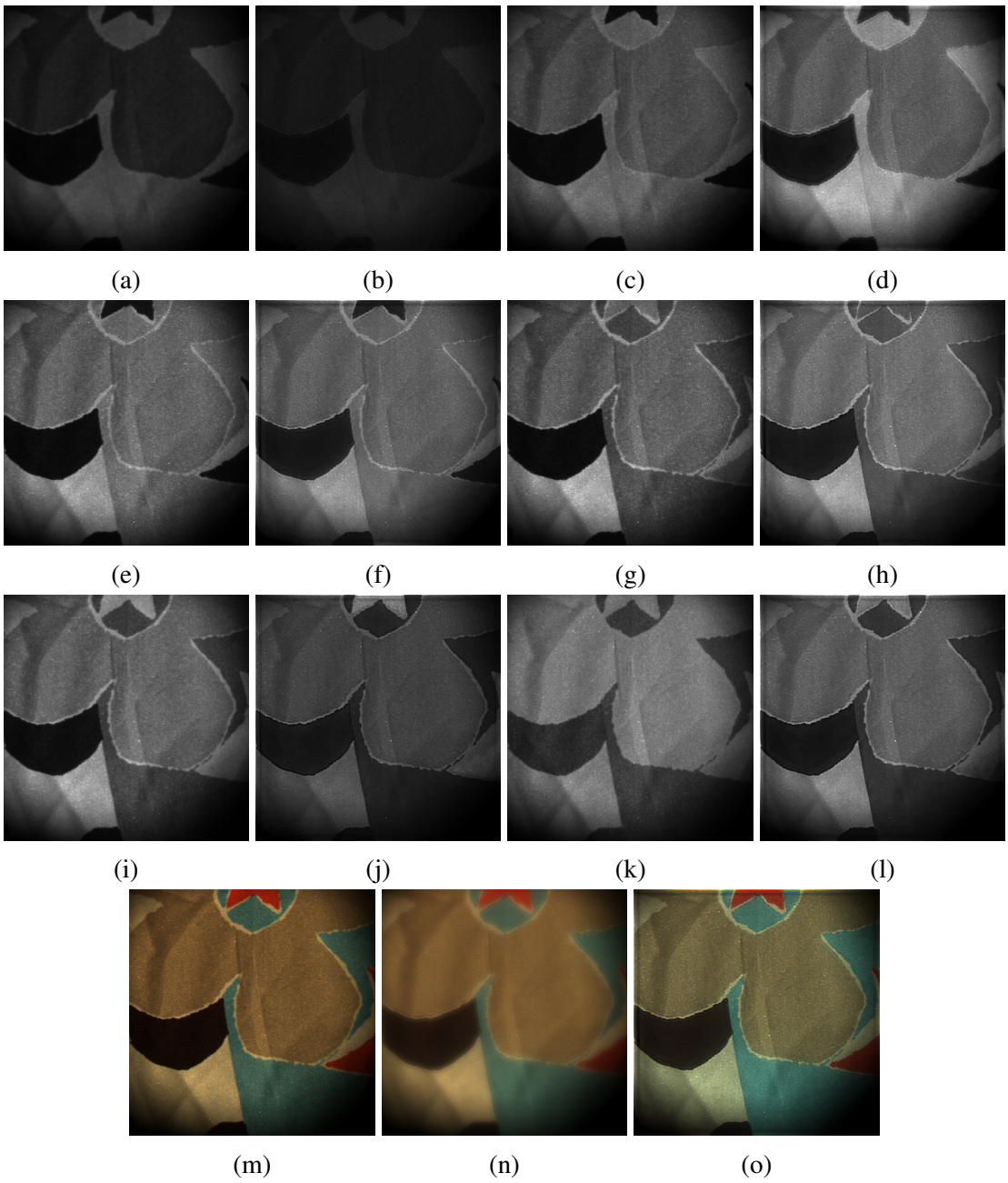


Figure A.16: Reconstruction from snapshot with *ring* PSFs of real-world cloth scene. GT stands for ground truth. RS means results. Illustrations (a) to (l) are comparisons of individual spectral images: (a) 450 nm GT, (b) 450 nm RS, (c) 500 nm GT, (d) 500 nm RS, (e) 550 nm GT, (f) 550 nm RS, (g) 600 nm GT, (h) 600 nm RS, (i) 650 nm GT, (j) 650 nm RS, (k) 700 nm GT, (l) 700 nm RS, and (m) is the ground truth single shot image without PSF modulation. (n) is the snapshot with designed PSFs. (o) is the restored RGB image from spectral information.

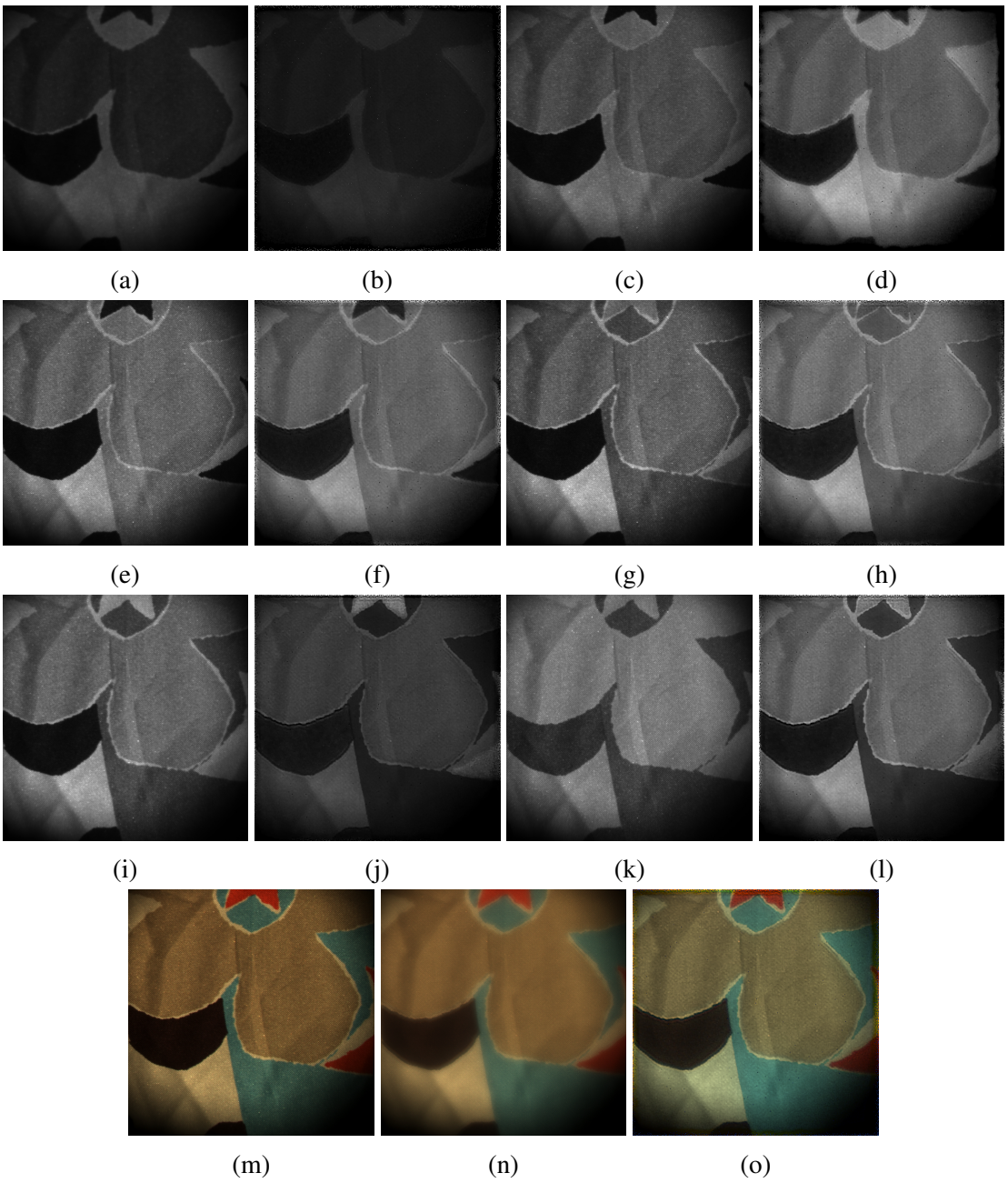


Figure A.17: Reconstruction from snapshot with *dots* PSFs of real-world cloth scene. GT stands for ground truth. RS means results. Illustrations (a) to (l) are comparisons of individual spectral images: (a) 450 nm GT, (b) 450 nm RS, (c) 500 nm GT, (d) 500 nm RS, (e) 550 nm GT, (f) 550 nm RS, (g) 600 nm GT, (h) 600 nm RS, (i) 650 nm GT, (j) 650 nm RS, (k) 700 nm GT, (l) 700 nm RS, and (m) is the ground truth single shot image without PSF modulation. (n) is the snapshot with designed PSFs. (o) is the restored RGB image from spectral information.

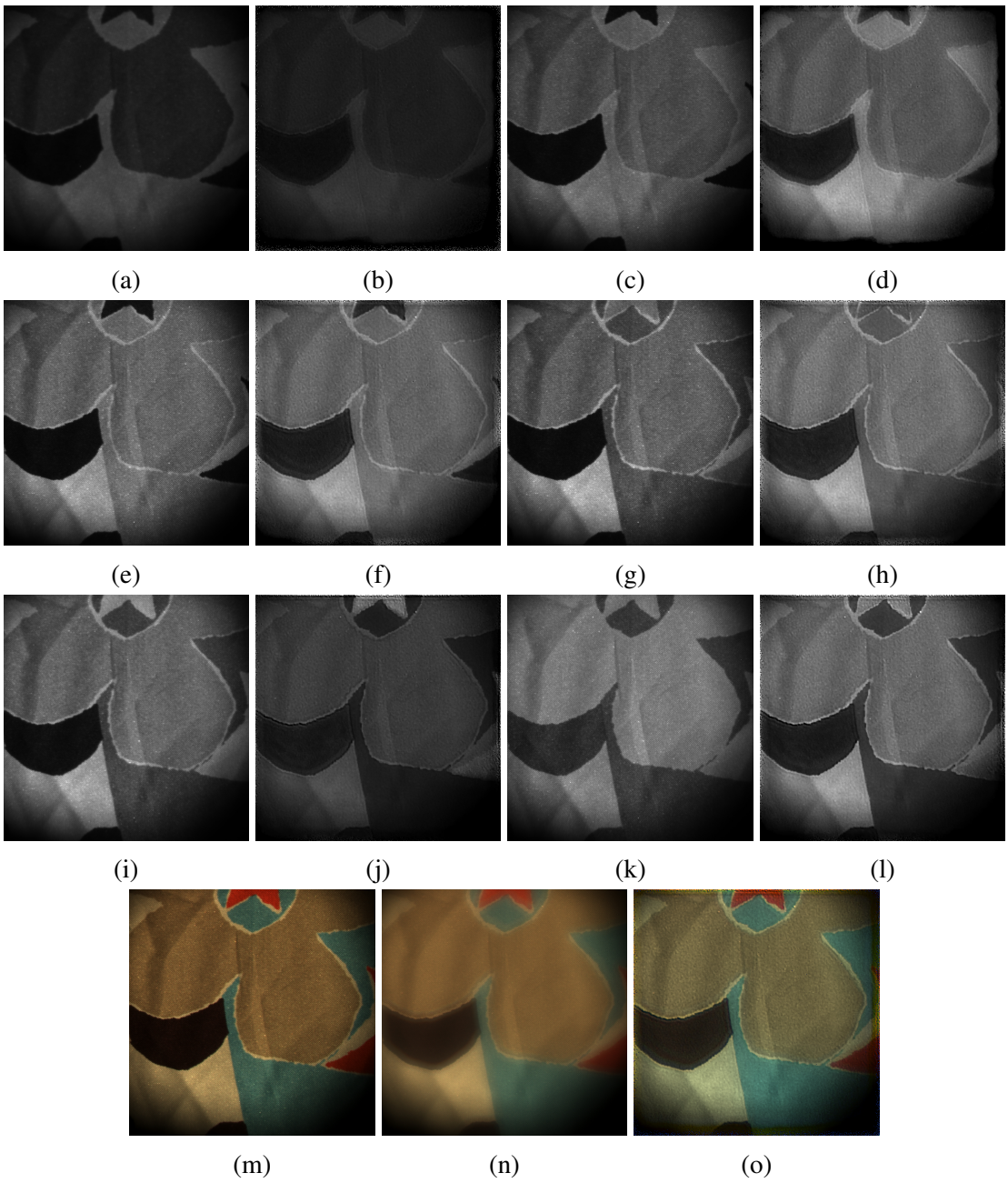


Figure A.18: Reconstruction from snapshot with *spiral* PSFs of real-world cloth scene. GT stands for ground truth. RS means results. Illustrations (a) to (l) are comparisons of individual spectral images: (a) 450 nm GT, (b) 450 nm RS, (c) 500 nm GT, (d) 500 nm RS, (e) 550 nm GT, (f) 550 nm RS, (g) 600 nm GT, (h) 600 nm RS, (i) 650 nm GT, (j) 650 nm RS, (k) 700 nm GT, (l) 700 nm RS, and (m) is the ground truth single shot image without PSF modulation. (n) is the snapshot with designed PSFs. (o) is the restored RGB image from spectral information.

Nomenclature

Matrices and vectors

α, β, n, w	scalar values
θ^n	nth iteration of θ
i	the imaginary unit
$\vec{E}, \vec{H}, \vec{r}$	row vectors
$\mathbf{p}, \mathbf{n}$	column vectors
g_x	the x component of a vector
$\mathbf{I}, \mathbf{O}$	matrices

Other symbols

$f(n)$	value of a function f at n
exp	exponential function
$\int$	integral operator
$\|\mathbf{v}\|$	norm of a vector or matrix $\mathbf{v}$
$\sum$	summation
$\arg\min$	the arguments of minima
∇, $\nabla\cdot$, $\nabla\times$	gradient, divergence, and curl operators
$\bigotimes$	convolution
$\mathcal{F}$	Fourier transform

Abbreviations

CNN	convolutional neural network	23
DOE	diffractive optical element	15
DFT	digital Fourier transform	13
EDoF	extended depth of field	26
FFT	fast Fourier transform	65
FoV	field of view	11
HDR	high dynamic range	1
LCoS	liquid crystal on silicon	15
LSI	linear shift-invariant	11
PSF	point spread function	30
PSNR	peak signal to noise ratio	50
SLM	spatial light modulator	16
SSIM	structural similarity index measure	71
TV	total variation	48

Bibliography

[AAB+15] Martin Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, G.s Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, and Xiaoqiang Zheng. *TensorFlow : Large-Scale Machine Learning on Heterogeneous Distributed Systems*. January 2015.

[AG14] Nicola Asuni and Andrea Giachetti. Testimages: a large-scale archive for testing visual devices and basic image processing algorithms. In *STAG*, pages 63–70, 2014.

[AKH+18] 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, 2018.

[AL01] Maryam Alavi and Dorothy E Leidner. Knowledge management and knowledge management systems: Conceptual foundations and research issues. *MIS quarterly*, pages 107–136, 2001.

[ANNW16] N. Antipa, S. Necula, R. Ng, and L. Waller. Single-shot diffuser-encoded light field imaging. In *2016 IEEE International Conference on Computational Photography (ICCP)*, pages 1–11, May 2016.

[Bar17] Bruce Barnbaum. *The Art of Photography: A Personal Approach to Artistic Expression*. Rocky Nook, Inc., 2017.

[Bas17] Andrew Basden. *The foundations of information systems: Research and practice*. Routledge, 2017.

[BEZN05] Moshe Ben-Ezra, Assaf Zomet, and Shree K Nayar. Video super-resolution using controlled subpixel detector shifts. *IEEE transactions on pattern analysis and machine intelligence*, 27(6):977–987, 2005.

[BKC17] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 39(12):2481–2495, 2017.

[BKGK17] Seung-Hwan Baek, Incheol Kim, Diego Gutierrez, and Min H Kim. Compact single-shot hyperspectral imaging using a prism. *ACM Transactions on Graphics (TOG)*, 36(6):217, 2017.

[BLOC83] K-H Brenner, AW Lohmann, and Jorge Ojeda-Castañeda. The ambiguity function as a polar display of the otf. *Optics Communications*, 44(5):323–326, 1983.

[BM13] Harrison H Barrett and Kyle J Myers. *Foundations of image science*. John Wiley & Sons, 2013.

[BNS94] Richard H Byrd, Jorge Nocedal, and Robert B Schnabel. Representations of quasi-newton matrices and their use in limited memory methods. *Mathematical Programming*, 63(1-3):129–156, 1994.

[Bow38] IS Bowen. The image-slicer a device for reducing loss of light at slit of stellar spectrograph. *The Astrophysical Journal*, 88:113, 1938.

[BV92] Theodor V Bulygin and Gennady N Vishnyakov. Spectrotomography: a new method of obtaining spectrograms of two-dimensional objects. In *Analytical Methods for Optical Tomography*, pages 315–322. International Society for Optics and Photonics, 1992.

[Cao15] Xudong Cao. A practical theory for designing very deep convolutional neural networks. page 6, 2015.

[Car14] Peter Carbonetto. A matlab interface for l-bfgs-b, March 2014.

[CBW+17] Guillem Carles, James Babington, Andrew Wood, Jason F Ralph, and Andrew R Harvey. Superimposed multi-resolution imaging. *Optics Express*, 25(26):33043–33055, 2017.

[CD02] W. Thomas Cathey and Edward R. Dowski. New paradigm for imaging systems. *Applied Optics*, 41(29):6080–6092, October 2002.

[CG09] Wanli Chi and Nicholas George. Phase-coded aperture for optical imaging. *Optics Communications*, 282(11):2110–2117, 2009.

[Cha16] Ayan Chakrabarti. Learning Sensor Multiplexing Design Through Backpropagation. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, pages 3089–3097, USA, 2016. Curran Associates Inc.

[CHEL18] Jieen Chen, Michael Hirsch, Bernhard Eberhardt, and Hendrik PA Lensch. A computational camera with programmable optics for snapshot high-resolution multispectral imaging. In *Asian Conference on Computer Vision*, pages 685–699. Springer, 2018.

[CHH+17] Jieen Chen, Michael Hirsch, Rainer Heintzmann, Bernhard Eberhardt, and Hendrik Lensch. A phase-coded aperture camera with programmable optics. *Electronic Imaging*, 2017(17):70–75, 2017.

[CMBH10] G Carles, G Muyo, S Bosch, and AR Harvey. Use of a spatial light modulator as an adaptable phase mask for wavefront coding. *Journal of Modern Optics*, 57(10):893–900, 2010.

[CW19] Julie Chang and Gordon Wetzstein. Deep optics for monocular depth estimation and 3d object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 10193–10202, 2019.

[CZL+17] Zhaolou Cao, Chunjie Zhai, Jinhua Li, Fenglin Xian, and Shixin Pei. Combination of color coding and wavefront coding for extended depth of field. *Optics Communications*, 392:252–257, June 2017.

[CZN10] O. Cossairt, C. Zhou, and S.K. Nayar. Diffusion Coding Photography for Extended Depth of Field. *ACM Trans. on Graphics (also Proc. of ACM SIGGRAPH)*, August 2010.

[DC95] Edward R. Dowski and W. Thomas Cathey. Extended depth of field through wave-front coding. *Applied Optics*, 34(11):1859, April 1995.

[DHS11] John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of machine learning research*, 12(7), 2011.

[DJC95] Edward R Dowski Jr and W Thomas Cathey. Extended depth of field through wave-front coding. *Applied Optics*, 34(11):1859–1866, 1995.

[EKD+17] Gabriel Eilertsen, Joel Kronander, Gyorgy Denes, RafałK. Mantiuk, and Jonas Unger. HDR Image Reconstruction from a Single Exposure Using Deep CNNs. *ACM Trans. Graph.*, 36(6):178:1–178:15, November 2017.

[FEM05] Karen E Fisher, Sanda Erdelez, and Lynne EF McKechnie. *Theories of information behavior*. Information Today, Inc., 2005.

[Fie82] James R Fienup. Phase retrieval algorithms: a comparison. *Applied optics*, 21(15):2758–2769, 1982.

[Fie10] Robert D Fiete. *Modeling the imaging chain of digital cameras*. SPIE press Bellingham, 2010.

[FSH+06] Rob Fergus, Barun Singh, Aaron Hertzmann, Sam T Roweis, and William T Freeman. Removing camera shake from a single photograph. In *ACM SIGGRAPH 2006 Papers*, pages 787–794. 2006.

[GAW+10] Miguel Granados, Boris Ajdin, Michael Wand, Christian Theobalt, Hans-Peter Seidel, and Hendrik PA Lensch. Optimal HDR reconstruction with linear digital cameras. In *CVPR*, 2010.

[GBA08] Herbert Gross, Fritz Blechinger, and Bertram Achtner. Zoom systems. *Handbook of Optical Systems: Volume 4: Survey of Optical Instruments*, 4:445–539, 2008.

[GBC16] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016.

[Ger72] Ralph W Gerchberg. A practical algorithm for the determination of phase from image and diffraction plane pictures. *Optik*, 35:237, 1972.

[GG16] Ruohan Gao and Kristen Grauman. On-Demand Learning for Deep Image Restoration. *arXiv:1612.01380 [cs]*, December 2016.

[GIDW11] Wetzstein Gordon, Ihrke Ivo, Lanman Douglas, and Heidrich Wolfgang. Computational Plenoptic Imaging. *Computer Graphics Forum*, 30(8):2397–2426, October 2011.

[GL13] John D Griffith and Jisoo Lee. Dual sensor camera, September 2013. US Patent 8,542,287.

[GLDZ15] Ioannis Gkioulekas, Anat Levin, Frédo Durand, and Todd Zickler. Micron-scale light transport decomposition using interferometry. *ACM Transactions on Graphics (ToG)*, 34(4):1–14, 2015.

[Goo05] Joseph W Goodman. *Introduction to Fourier optics*. Roberts and Company Publishers, New York, fourth edition edition, 2005.

[Goo17] Joseph W. Goodman. *Introduction to Fourier optics*. W.H. Freeman, Macmillan Learning, New York, fourth edition edition, 2017.

[Gre14] Samuel Greengard. Computational photography comes into focus. *Communications of the ACM*, 57(2):19–21, February 2014.

[GZL14] Daniel Glasner, Todd Zickler, and Anat Levin. A reflectance display. *ACM Trans. Graph.*, 33(4):61, 2014.

[HDN+16] Felix Heide, Steven Diamond, Matthias Nießner, Jonathan Ragan-Kelley, Wolfgang Heidrich, and Gordon Wetzstein. Proximal: Efficient image optimization using proximal algorithms. *ACM Transactions on Graphics (TOG)*, 35(4):84, 2016.

[HGO18] Bernardo Henz, Eduardo S. L. Gastal, and Manuel M. Oliveira. Deep Joint Design of Color Filter Arrays and Demosaicing. *Computer Graphics Forum*, 37(2), 2018.

[HHGH13] 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):1–10, 2013.

[HIII94] Akiako Hirai, Takashi Inoue, Kazuyoshi Itoh, and Yoshiki Ichioka. Application of measurement multiple-image fourier of fast phenomena transform spectral imaging to measurement of fast phenomena. *Optical Review*, 1(2):205–207, 1994.

[HK13] Nathan Hagen and Michael W Kudenov. Review of snapshot spectral imaging technologies. *Optical Engineering*, 52(9):090901–090901, 2013.

[HRH+13] Felix Heide, Mushfiqur Rouf, Matthias B Hullin, Bjorn Labitzke, Wolfgang Heidrich, and Andreas Kolb. High-quality computational imaging through simple lenses. *ACM Transactions on Graphics (TOG)*, 32(5):149, 2013.

[HS06] Geoffrey E Hinton and Ruslan R Salakhutdinov. Reducing the dimensionality of data with neural networks. *science*, 313(5786):504–507, 2006.

[HSS12] Geoffrey Hinton, Nitish Srivastava, and Kevin Swersky. Neural networks for machine learning lecture 6a overview of mini-batch gradient descent. *Cited on*, 14(8):2, 2012.

[HST+14] Felix Heide, Markus Steinberger, Yun-Ta Tsai, Mushfiqur Rouf, Dawid Pająk, Dikpal Reddy, Orazio Gallo, Jing Liu, Wolfgang Heidrich, Karen Egiazarian, et al. Flexisp: A flexible camera image processing framework. *ACM Transactions on Graphics (TOG)*, 33(6):231, 2014.

[HZRS15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. 7, December 2015.

[HZRS16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. 7:770–778, December 2016.

[IMK+14] Forrest Iandola, Matt Moskewicz, Sergey Karayev, Ross Girshick, Trevor Darrell, and Kurt Keutzer. Densenet: Implementing efficient convnet descriptor pyramids. *arXiv preprint arXiv:1404.1869*, 2014.

[IZZE17] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. Image-to-Image Translation with Conditional Adversarial Networks. In *2017 IEEE*

Conference on Computer Vision and Pattern Recognition (CVPR), pages 5967–5976, Honolulu, HI, July 2017. IEEE.

[JDS08] Hervé Jégou, Matthijs Douze, and Cordelia Schmid. Hamming embedding and weak geometric consistency for large scale image search. pages 304–317, January 2008.

[JMFU17] Kyong Hwan Jin, Michael T. McCann, Emmanuel Froustey, and Michael Unser. Deep Convolutional Neural Network for Inverse Problems in Imaging. *IEEE Transactions on Image Processing*, 26(9):4509–4522, September 2017.

[JTFW17] Haomiao Jiang, Qiyuan Tian, Joyce Farrell, and Brian Wandell. Learning the image processing pipeline. *IEEE Transactions on Image Processing*, 26(10):5032–5042, October 2017.

[KB14] Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. *arXiv:1412.6980 [cs]*, December 2014.

[KDA+17] Ivan Krasin, Tom Duerig, Neil Alldrin, Vittorio Ferrari, Sami Abu-El-Haija, Alina Kuznetsova, Hassan Rom, Jasper Uijlings, Stefan Popov, Shahab Kamali, Matteo Malloci, Jordi Pont-Tuset, Andreas Veit, Serge Belongie, Victor Gomes, Abhinav Gupta, Chen Sun, Gal Chechik, David Cai, Zheyun Feng, Dhyanesh Narayanan, and Kevin Murphy. Openimages: A public dataset for large-scale multi-label and multi-class image classification. *Dataset available from https://storage.googleapis.com/openimages/web/index.html*, 2017.

[KHFG14] Ori Katz, Pierre Heidmann, Mathias Fink, and Sylvain Gigan. Non-invasive single-shot imaging through scattering layers and around corners via speckle correlations. *Nature photonics*, 8(10):784–790, 2014.

[KKS+95] K Kuniyoshi, Nobuyuki Kita, Kazuhide Sugimoto, Shin Nakamura, and Takashi Suehiro. A foveated wide angle lens for active vision. In *Proceedings of 1995 IEEE International Conference on Robotics and Automation*, volume 3, pages 2982–2988. IEEE, 1995.

[KSH12] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[KSZQ20] Asifullah Khan, Anabia Sohail, Umme Zahoora, and Aqsa Saeed Qureshi. A survey of the recent architectures of deep convolutional neural networks. *Artificial Intelligence Review*, pages 1–62, 2020.

[Kub13] Joel A Kubby. *Adaptive Optics for Biological Imaging*. CRC press, 2013.

[LBBH98] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[LBD+89] Yann LeCun, Bernhard Boser, John S Denker, Donnie Henderson, Richard E Howard, Wayne Hubbard, and Lawrence D Jackel. Backpropagation applied to handwritten zip code recognition. *Neural computation*, 1(4):541–551, 1989.

[LFDF07] Anat Levin, Rob Fergus, Frédo Durand, and William T Freeman. Image and depth from a conventional camera with a coded aperture. In *ACM Transactions on Graphics (TOG)*, volume 26, page 70. ACM, 2007.

[LH96] Marc Levoy and Pat Hanrahan. Light field rendering. In *Proceedings of the 23rd annual conference on Computer graphics and interactive techniques*, pages 31–42, 1996.

[LHG+09] Anat Levin, Samuel W Hasinoff, Paul Green, Frédo Durand, and William T Freeman. 4d frequency analysis of computational cameras for depth of field extension. In *ACM Transactions on Graphics (TOG)*, volume 28, page 97. ACM, ACM New York, NY, USA, 2009.

[LLWD14] Xing Lin, Yebin Liu, Jiamin Wu, and Qionghai Dai. Spatial-spectral encoded compressive hyperspectral imaging. *ACM Transactions on Graphics (TOG)*, 33(6):233, 2014.

[LMM+08] A. Lizana, A. Marquez, I. Moreno, C. Iemmi, J. Campos, and M. J. Yzuel. Wavelength dependence of polarimetric and phase-shift characterization of a liquid crystal on silicon display. *Journal of the European Optical Society - Rapid publications*, 3(0), March 2008.

[LN12] Michael F Land and Dan-Eric Nilsson. *Animal eyes*. Oxford University Press, 2012.

[LPY16] Sifei Liu, Jinshan Pan, and Ming-Hsuan Yang. Learning Recursive Filters for Low-Level Vision via a Hybrid Neural Network. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *Computer Vision – ECCV 2016*, volume 9908, pages 560–576. Springer International Publishing, Cham, 2016.

[LSK+17] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 136–144, 2017.

[Luc74] L. B. Lucy. An iterative technique for the rectification of observed distributions. *The Astronomical Journal*, 79:745, June 1974.

[Maî17] Henri Maître. *From Photon to Pixel: The Digital Camera Handbook*. John Wiley & Sons, February 2017.

[MIPW20] Christopher A Metzler, Hayato Ikoma, Yifan Peng, and Gordon Wetzstein. Deep optics for single-shot high-dynamic-range imaging. In *Proc. CVPR*, 2020.

[MMHC17] Tim Meinhardt, Michael Möller, Caner Hazirbas, and Daniel Cremers. Learning proximal operators: Using denoising networks for regularizing inverse imaging problems. *ArXiv e-prints, Apr*, 2017.

[MMP+05] Morgan McGuire, Wojciech Matusik, Hanspeter Pfister, John F. Hughes, and Frédo Durand. Defocus video matting. *ACM Trans. Graph.*, 24(3):567–576, 2005.

[MPCRR14] María S Millán, Elisabet Pérez-Cabré, Lenny A Romero, and Natalia Ramírez. Programmable diffractive lens for ophthalmic application. *Optical Engineering*, 53(6):061709–061709, 2014.

[MS15] Steve Marschner and Peter Shirley. *Fundamentals of computer graphics*. CRC Press, 2015.

[MSL19] Scott W Miller, Shashank Sharma, and Simon S Lee. Mobile zoom using multiple optical image stabilization cameras, April 2019. US Patent 10,264,188.

[MSY16] Xiaojiao Mao, Chunhua Shen, and Yu-Bin Yang. Image Restoration Using Very Deep Convolutional Encoder-Decoder Networks with Symmetric Skip Connections. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 2802–2810. Curran Associates, Inc., 2016.

[MWP+04] Ty Martinez, David V Wick, Don M Payne, Jeffrey T Baker, and Sergio R Restaino. Non-mechanical zoom system. In *Sensors, Systems, and Next-Generation Satellites VII*, volume 5234, pages 375–378. International Society for Optics and Photonics, 2004.

[MWR01] Ty Martinez, David V Wick, and Sergio R Restaino. Foveated, wide field-of-view imaging system using a liquid crystal spatial light modulator. *Optics Express*, 8(10):555–560, 2001.

[MYK+19] Kristina Monakhova, Joshua Yurtsever, Grace Kuo, Nick Antipa, Kyrollos Yanny, and Laura Waller. Learned reconstructions for practical mask-based lensless imaging. *Optics express*, 27(20):28075–28090, 2019.

[Nay06] S. K. Nayar. Computational Cameras: Redefining the Image. *Computer*, 39(8):30–38, August 2006.

[Nay11] Shree K. Nayar. Computational Cameras: Approaches, Benefits and Limits. 2011.

[NGZ+18] Shijie Nie, Lin Gu, Yinqiang Zheng, Antony Lam, Nobutaka Ono, and Imari Sato. Deeply Learned Filter Response Functions for Hyperspectral Reconstruction. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[NKL17] Seungjun Nah, Tae Hyun Kim, and Kyoung Mu Lee. Deep Multi-scale Convolutional Neural Network for Dynamic Scene Deblurring. pages 257–265. IEEE, July 2017.

[NW06] Jorge Nocedal and Stephen Wright. *Numerical optimization*. Springer Science & Business Media, 2006.

[OSSS15] Mitsuhiko Ohta, Koichi Sakita, Takeshi Shimano, and Akito Sakemoto. Rotationally symmetric wavefront coding for extended depth of focus with annular phase mask. *Japanese Journal of Applied Physics*, 54(9S):09ME03, September 2015.

[OTY93] Takayuki Okamoto, Akinori Takahashi, and Ichirou Yamaguchi. Simultaneous acquisition of spectral and spatial intensity distribution. *Applied Spectroscopy*, 47(8):1198–1202, 1993.

[PA01] Sofya Poger and Elli Angelopoulou. Multispectral sensors in computer vision. *Stevens Institute of Technology Technical Report CS Report 2001*, 3, 2001.

[PCZZ07] Runling Peng, Jiabi Chen, Cheng Zhu, and Songlin Zhuang. Design of a zoom lens without motorized optical elements. *Optics express*, 15(11):6664–6669, 2007.

[PFA+15] Yifan Peng, Qiang Fu, Hadi Amata, Shuochen Su, Felix Heide, and Wolfgang Heidrich. Computational imaging using lightweight diffractive-refractive optics. *Optics express*, 23(24):31393–31407, 2015.

[PFHH16] Yifan Peng, Qiang Fu, Felix Heide, and Wolfgang Heidrich. The diffractive achromat full spectrum computational imaging with diffractive optics.

In *SIGGRAPH ASIA 2016 Virtual Reality meets Physical Reality: Modelling and Simulating Virtual Humans and Environments*, page 4. ACM, 2016.

[PLS+18] Jinshan Pan, Sifei Liu, Deqing Sun, Jiawei Zhang, Yang Liu, Jimmy Ren, Zechao Li, Jinhui Tang, Huchuan Lu, Yu-Wing Tai, and Ming-Hsuan Yang. Learning Dual Convolutional Neural Networks for Low-Level Vision. *arXiv:1805.05020 [cs]*, May 2018.

[PSD+19] Yifan Peng, Qilin Sun, Xiong Dun, Gordon Wetzstein, Wolfgang Heidrich, and Felix Heide. Learned large field-of-view imaging with thin-plate optics. *ACM Transactions on Graphics (TOG)*, 38(6):219, 2019.

[RFB15] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional Networks for Biomedical Image Segmentation. *arXiv:1505.04597 [cs]*, May 2015.

[Ric72a] William Hadley Richardson. Bayesian-based iterative method of image restoration. *JOSA*, 62(1):55–59, January 1972.

[Ric72b] William Hadley Richardson. Bayesian-Based Iterative Method of Image Restoration*. *JOSA*, 62(1):55–59, January 1972.

[RKS+18] Adria Recasens, Petr Kellnhofer, Simon Stent, Wojciech Matusik, and Antonio Torralba. Learning to zoom: a saliency-based sampling layer for neural networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 51–66, 2018.

[Rob86] Phillip H Roberts. A wave optics propagation code. *Rep. TR-760 (the Optical Sciences Company, Anaheim, Calif., 1986)*, 1986.

[Rud16] Sebastian Ruder. An overview of gradient descent optimization algorithms. *arXiv preprint arXiv:1609.04747*, 2016.

[SBHS13] C. J. Schuler, H. C. Burger, S. Harmeling, and B. Schölkopf. A Machine Learning Approach for Non-blind Image Deconvolution. In *2013 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1067–1074, June 2013.

[SDP+18] 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 SIGGRAPH*, 2018.

[SGB18] Eli Schwartz, Raja Giryes, and Alex M. Bronstein. DeepISP: Learning End-to-End Image Processing Pipeline. *arXiv:1801.06724 [cs, eess]*, January 2018.

[Sha48] Claude E Shannon. A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423, 1948.

[SLJ+15] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015.

[Smi07] W. Smith. *Modern Optical Engineering, 4th Ed.* McGraw Hill professional. McGraw-Hill Education, 2007.

[STTP14] Yuliy Schwartzburg, Romain Testuz, Andrea Tagliasacchi, and Mark Pauly. High-contrast computational caustic design. *ACM Transactions on Graphics (SIGGRAPH)*, 2014.

[SWB+16] Yoav Shechtman, Lucien E Weiss, Adam S Backer, Maurice Y Lee, and WE Moerner. Multicolour localization microscopy by point-spread-function engineering. *Nature Photonics*, 2016.

[SZ14] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[SZD+19] Qilin Sun, Jian Zhang, Xiong Dun, Bernard Ghanem, Yifan Peng, and Wolfgang Heidrich. End-to-end learned, optically coded super-resolution spad camera. 2019.

[Sze10] Richard Szeliski. *Computer vision: algorithms and applications*. Springer Science & Business Media, 2010.

[TAG+17] Simon Thiele, Kathrin Arzenbacher, Timo Gissibl, Harald Giessen, and Alois M Herkommer. 3d-printed eagle eye: Compound microlens system for foveated imaging. *Science advances*, 3(2):e1602655, 2017.

[TF82] GA Tyler and DL Fried. A wave optics propagation algorithm. *Rep. TR-451 (the Optical Sciences Company, Anaheim, Calif., 1982)*, 1982.

[Tys15] R.K. Tyson. *Principles of Adaptive Optics, Fourth Edition*. CRC Press, 2015.

[WBC+19] Yicheng Wu, Vivek Boominathan, Huaijin Chen, Aswin Sankaranarayanan, and Ashok Veeraraghavan. Phasecam3d—learning phase masks for passive single view depth estimation. In *2019 IEEE International Conference on Computational Photography (ICCP)*, pages 1–12. IEEE, 2019.

[WCH20] Zhihao Wang, Jian Chen, and Steven CH Hoi. Deep learning for image super-resolution: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

[WHSL17] P. Wieschollek, M. Hirsch, B. Schölkopf, and H. Lensch. Learning Blind Motion Deblurring. In *Proceedings IEEE International Conference on Computer Vision (ICCV)*, pages 231–240, Piscataway, NJ, USA, October 2017. IEEE.

[Wie49] Norbert Wiener. *Extrapolation, interpolation, and smoothing of stationary time series*, volume 2. MIT press Cambridge, MA, 1949.

[Wie19] Norbert Wiener. *Cybernetics or Control and Communication in the Animal and the Machine*. MIT press, 2019.

[WJWB08] Ashwin Wagadarikar, Renu John, Rebecca Willett, and David Brady. Single disperser design for coded aperture snapshot spectral imaging. *Applied optics*, 47(10):B44–B51, 2008.

[WSVM16] Peng Wang, Eyal Shafran, Fernando G Vasquez, and Rajesh Menon. Snapshot high-resolution hyper-spectral imager based on an ultra-thin diffractive filter. In *Imaging Systems and Applications*, pages IW1E–1. Optical Society of America, 2016.

[WT14] Ruxin Wang and Dacheng Tao. Recent Progress in Image Deblurring. *CoRR*, abs/1409.6838, 2014.

[Wu18] Yuxin Wu. tensorpack: A Neural Net Training Interface on TensorFlow, May 2018. original-date: 2015-12-25.

[WXG+15] Lizhi Wang, Zhiwei Xiong, Dahua Gao, Guangming Shi, Wenjun Zeng, and Feng Wu. High-speed hyperspectral video acquisition with a dual-camera architecture. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4942–4950, 2015.

[WZK+17] Ting-Chun Wang, Jun-Yan Zhu, Nima Khademi Kalantari, Alexei A. Efros, and Ravi Ramamoorthi. Light Field Video Capture Using a Learning-Based Hybrid Imaging System. *ACM Transactions on Graphics (Proceedings of SIGGRAPH 2017)*, 36(4), 2017.

[XRLJ14] Li Xu, Jimmy S. J. Ren, Ce Liu, and Jiaya Jia. Deep Convolutional Neural Network for Image Deconvolution. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'14, pages 1790–1798, Cambridge, MA, USA, 2014. MIT Press.

[XXC12] Junyuan Xie, Linli Xu, and Enhong Chen. Image Denoising and Inpainting with Deep Neural Networks. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 341–349. Curran Associates, Inc., 2012.

[Yar11] Leonid P Yaroslavsky. Linking analog and digital image processing. *Optical and Digital Image Processing: Fundamentals and Applications*, pages 397–418, 2011.

[YJY+15] Youngjin Yoon, Hae-Gon Jeon, Donggeun Yoo, Joon-Young Lee, and In So Kweon. Learning a Deep Convolutional Network for Light-Field Image Super-Resolution. pages 57–65. IEEE, December 2015.

[YMIN10] Fumihito Yasuma, Tomoo Mitsunaga, Daisuke Iso, and Shree K Nayar. Generalized assorted pixel camera: postcapture control of resolution, dynamic range, and spectrum. *IEEE transactions on image processing*, 19(9):2241–2253, 2010.

[ZBLN97] Ciyou Zhu, Richard H Byrd, Peihuang Lu, and Jorge Nocedal. Algorithm 778: L-bfgs-b: Fortran subroutines for large-scale bound-constrained optimization. *ACM Transactions on Mathematical Software (TOMS)*, 23(4):550–560, 1997.

[ZCNK19] Xuaner Zhang, Qifeng Chen, Ren Ng, and Vladlen Koltun. Zoom to learn, learn to zoom. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3762–3770, 2019.

[ZJL05] De-Ying Zhang, Nicole Justis, and Yu-Hwa Lo. Fluidic adaptive zoom lens with high zoom ratio and widely tunable field of view. *Optics communications*, 249(1-3):175–182, 2005.

[ZLCLL17] Xingcheng Zhang, Zhizhong Li, Chen Change Loy, and Dahua Lin. Polynet: A pursuit of structural diversity in very deep networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 718–726, 2017.

[ZLL+18] Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu. Image super-resolution using very deep residual channel attention networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 286–301, 2018.

[ZN11] Changyin Zhou and Shree K Nayar. Computational cameras: convergence of optics and processing. *Image Processing, IEEE Transactions on*, 20(12):3322–3340, 2011.