# End-to-end Optics Design for Computational Cameras

Dissertation by

Qilin Sun

In Partial Fulfillment of the Requirements

For the Degree of

Doctor of Philosophy

King Abdullah University of Science and Technology

Thuwal, Kingdom of Saudi Arabia

September, 2021

# EXAMINATION COMMITTEE PAGE

The dissertation of Qilin Sun is approved by the examination committee

Committee Chairperson: Prof. Wolfgang Heidrich
Committee Members: Prof. Bernard Ghanem, Prof. Dominik Michels, Prof. Ashok Veeraraghavan

# ABSTRACT

End-to-end Optics Design for Computational Cameras

Qilin Sun

Imaging systems have long been designed in separated steps: the experience-driven optical design followed by sophisticated image processing. Such a general-propose approach achieves success in the past but left the question open for specific tasks and the best compromise between optics and post-processing, as well as minimizing costs. Driven from this, a series of works are proposed to bring the imaging system design into end-to-end fashion step by step, from joint optics design, point spread function (PSF) optimization, phase map optimization to a general end-to-end complex lens camera.

To demonstrate the joint optics application with image recovery, we applied it to flat lens imaging with a large field of view (LFOV). In applying a super-resolution single-photon avalanche diode (SPAD) camera, the PSF encoded by diffractive optical element (DOE) is optimized together with the post-processing, which brings the optics design into the end-to-end stage. Expanding to color imaging, optimizing PSF to achieve DOE fails to find the best compromise between different wavelengths. Snapshot HDR imaging is achieved by optimizing a phase map directly. All works are demonstrated with prototypes and experiments in the real world.

To further compete for the blueprint of end-to-end camera design and break the limits of a simple wave optics model and a single lens surface. Finally, we propose a general end-to-end complex lens design framework enabled by a differentiable ray tracing image formation model. All works are demonstrated with prototypes and experiments in the real world. Our frameworks offer competitive alternatives for the design of modern imaging systems and several challenging imaging applications.

# ACKNOWLEDGEMENTS

Firstly, I would like to thank my Ph.D. advisor, Prof. Dr. Wolfgang Heidrich, for his valuable guidance, scholarly inputs, and consistent encouragement throughout the research work. A good supervisor, and more than thousands of books. Wolfgang is not only the disseminator of knowledge and is a life-long friend, an engineer of my soul, and a model. This dissertation would not have been possible without his continued patience and endless support. I am grateful to KAUST for treating the students like family and for the words "Your success is the success of KAUST." I am very grateful to the Ph.D. committee member, Prof. Bernard Ghanem and Prof. Dominik Michels, for their valuable guidance and patient academic support that helped me realize my lifelong dream of completing a Ph.D. degree as expected. Furthermore, I would like to thank Bernard Ghanem, Gordon Weinstein for cooperating with my research work. Their very thoughtful comments and feedback, such as the suggestion to SPAD cameras and LFOV imaging, are encouraging.

In addition, I would like to thank Felix Heide for his deep guidance on most of my research during my Ph.D. life. I would also like to thank Dr. Qiang Fu and Dr. Xiong Dun for many fruitful discussions on optics and optimization. It is so enjoyable to work and collaborate with them. Dr. Xiong Dun was involved in all the projects presented in this dissertation, and I am really lucky to work on these interesting projects with him during my Ph.D. studies.

Thanks to Congli Wang, Yifang Peng, Ethan Tseng for their deep cooperation and fruitful discussions on optics, paperwork, and image reconstructions.

# TABLE OF CONTENTS

# LIST OF ABBREVIATIONS

| | |
|---|---|
| ASP | aspherical lens |
| CDP | contrast detection probability |
| DOE | diffractive optical element |
| DOF | depth of field |
| EDOF | extended depth of field |
| FOV | field of view |
| FWHM | full-width-half-max |
| GS | Gerchberg-Saxton |
| HDR | high dynamic range |
| HR | low-resolution |
| LFOV | large field of view |
| LR | low-resolution |
| MAE | mean absolute error |
| MSE | mean square error |
| MTF | modulation transfer functions |
| OLPF | optical low-pass filter |
| PC | polycarbonate |
| PMMA | polymethyl methacrylate |
| PSF | point spread function |
| PSNR | peak signal-to-noise ratio |
| RMSE | root mean square error |
| SPAD | single-photon avalanche diode |
| SR | super-resolution |
| SSIM | structural similarity index measure |
| TCSPC | Time-correlated single photon counting |
| TOF | time-of-flight |

# LIST OF FIGURES

# LIST OF TABLES

# Chapter 1

# Introduction

Cameras are designed with a complicated trade-off between image quality (e.g., sharpness, contrast, color fidelity), and practical considerations such as cost, form factor, and weight. High-quality imaging systems require a stack of multiple optical elements to combat aberrations of all kinds. At the heart of the design process are tools like ZEMAX and Code V, which rely on merit functions to trade off the shape of the PSF over different image regions, depth, or zoom settings. Such a design process requires significant user knowledge and experience and the emphasis on PSF shaping neglects any subsequent image processing operations, specific application scenarios, or the desire to encode extra information in the image.

Therefore, domain-specific computational imaging has attracted researchers' attention in the past several decades. Enabling the co-design of optics with post-processing, computational cameras have achieved impressive results in extended depth of field (EDOF) [1, 2, 3, 4], high dynamic range (HDR) [5, 6, 7, 8], and image resolution [9, 10, 11]. Nevertheless, all those older methods are either heuristic or use some proxy metric on the PSF rather than considering the imaging quality after post-processing. Therefore, finding a joint optimal solution for both imaging optics and image reconstruction for a given task remains an unsolved problem in general.

Over the past few years, co-design of optics and image processing [12, 13], or even data-driven end-to-end design [14] have emerged to bridge the gap between optical design and algorithm development. Co-design of optics and post-processing algorithms has achieved a superior performance for domain specific tasks such as

depth estimation [15], LFOV imaging [13], EDOF [16], optimal sampling [17], and HDR imaging [18, 19].

Data-driven optimization of all the parameters in a complex lens assembly is challenging. On the one hand, the optical surfaces' varying parameters cause scaling and distortion during the optimization process. On the other hand, a naive implementation will consume enormous computational resources due to the differentiable ray tracing engine [20]. Recently, an end-to-end differentiable compound lens design [21] through building a PSF dictionary using Zemax has been investigated. In the meantime, we proposed a fully differentiable complex lens model [22] based on differentiable rendering, which achieved the milestone of the end-to-end computational camera design that considers aberrations of all kinds and supports complex lens design using data-driven methods.

This dissertation introduces the story of how we realized the end-to-end designing of optics and post-processing step by step. Inspired by the insight that the filters of early layers of recent deep models have a striking similarity, the joint design of optics and image recovery network is investigated firstly. To build a direct relationship between optics design and post-processing bottle-necked by the image sensor, the PSF is optimized according to the sensor property. The optical element is optimized according to the optimal PSF. To compromise between different wavelengths, directly optimizing the phase profile of optics is investigated, and it is applied to snap-shot HDR imaging. Unfortunately, these works' differentiable lens models have been too limited to describe complex optical assemblies and have only been allowed to optimize a single optical with a single material. Finally, we seek to build an optics model that can support complex lens design and optimize off-axis aberrations to meet the scope of consumer-level cameras.

## 1.1 Overview

This dissertation focuses on how to realize end-to-end optics design step by step and corresponding applications.

**Joint optics design with image recovery** (Chapter 3) is the first step that investigated to build the relationship between the imaging optics and image recovery network. This work is motivated by a large body of work on computational photography with PSF engineering–designing PSFs variant to a target characteristic instead of minimizing spot size and computation to remove the non-compact aberrations.

To achieve a LFOV with thin plate optics, the PSFs of our lens are constrained to be shift-invariant for the incident angle during the optimization process. Although such PSFs exhibit large spot sizes, the aberrations are engineered to preserve residual contrast and are well-suited for learned image reconstruction. A learned generative reconstruction model, a lens design tailored to this model, and a lab data acquisition approach that does not require the painful acquisition of real training images in the wild are realized in this work.

**End-to-end encoding through optimizing PSF** (Chapter 4) can support directly optimize the optics together with post-processing. This chapter takes one more step to bring optics design into end-to-end fashion.

SPAD sensor suffers from both low resolution and low fill-factor. An optical low-pass filter (OLPF) is introduced to suppress aliasing while preserving as much information as possible for super-resolution image reconstruction. In our framework, this filter and the matching reconstruction network are *jointly* learned in an end-to-end pipeline. After training, we extract the optimal PSF and then apply a Gerchberg-Saxton (GS)-based phase retrieval algorithm to derive the phase mask, which acts as an optical coder installed at the front focal plane of a regular lens to generate the

optimal PSF for later implementations.

**End-to-end encoding through optimizing phase mask** (Chapter 5) overcomes the limitation of single wavelength for the application of SR SPAD camera. It avoids relying on GS based phase retravel method. This chapter takes one more step to end-to-end camera design and considers color information as refractive index varies with the wavelength.

HDR imaging is an essential imaging modality for a wide range of applications, especially in uncontrolled environments like autonomous, robotics, and mobile phone cameras. However, existing HDR techniques struggle with dynamic scenes due to multi-shot acquisition and post-processing time. This chapter introduces a snapshot HDR imaging method that learns an optical encoding mask that maps saturated highlights into neighboring unsaturated areas. Novelly, a rank-1 parameterization of the DOEs drastically reduces the optical search space while allowing high-frequency encoding. Followed with stagelized recovery networks, this method can effectively recover HDR images in the real world. In addition, the recovery method is demonstrated to be effective in removing glare from in-the-wild automotive optics with windshield-induced streaks.

**Differentiable complex lens design** (Chapter 6) breaks the limitation of simple wave optics models such as Fourier transform or similar paraxial models. Previous models only support the optimization of a single lens surface, which limits the achievable image quality.

In this chapter, a general end-to-end complex lens design framework enabled by a differentiable ray tracing image formation model is proposed. A novel configurable and differentiable complex lens model that can simulate aberrations of all kinds is given, and it offers greater design freedom than the previous optics models. The differentiable complex lens model relies on the differentiable ray-tracing engine to render

optical images in the full field by considering all on/off-axis aberrations governed by the theory of geometric optics. This model supports the end-to-end optimization of optics with the image recovery algorithm for a specific imaging task and reaches the scope of consumer-level image quality. Finally, we demonstrate the effectiveness of the proposed method on two typical applications, including LFOV imaging and EDOF imaging. This framework offers a competition for the design of modern imaging systems.

## 1.2 Dissertation structure

In the remainder of this dissertation, Chapter 2 gives the basic concepts, diffractive and refractive optics model, aberrations and PSFs, imaging model, and post-processing model. Then, Chapter 3 presents the first step to end-to-end optics design, which jointly optimize the optics and image recovery network to realize LFOV imaging with thin plate optics. In Chapter 4, the story steps into a new chapter that enables a directly end-to-end PSF optimization with the recovery network. The optimal optics is then obtained through GS based phase retrieval algorithm to achieve a phase mask that can be fabricated. The fabricated phase mask success in the scenario of the SR of a low fill factor and low pixel counts SPAD array camera. Next, Chapter 4 further release the designing freedom of optics that allows direct optimize of the height map of the optics and find the best compromise between color channels. Followed with more advanced stagelized image recovery networks, it can effectively recover HDR images in the real world. To further release the designing freedom, we expand the optics model to the differentiable complex lens design, which not only breaks the limitation of a single optics surface but also considers off-axis aberrations. Beyond the optics model, we introduce a novel configurable and differentiable complex lens model based on differentiable ray-tracing, and this model can simulate aberrations of all kinds. Our framework offers a competitive alternative for the design of modern

imaging systems.

Finally, we conclude this dissertation with further research points in Chapter 7.

# Chapter 2

# Background and Related Work

In this chapter, we review the background techniques and the recent work related to the end-to-end optics model and the imaging model of the camera design and present the research background on which the work presented in this proposal has been based. In particular, it introduces the reader fundamental Optics terms, followed by introducing imaging models and image processing pipelines. We will refer to these concepts when we discuss algorithms and methods in the following chapters.

## 2.1 Luminance

The physical light measurement most suitable for images are photometric luminance units, and we will define it based on [23]. Luminance is an integrated spectral radiance across the range of visible wavelengths with the weighting function $W(\lambda)$:

$$Y = \int_{380\ nm}^{770\ nm} L(\lambda)W(\lambda)d\lambda. \tag{2.1}$$

Function $W(\lambda)$, called the spectral luminous efficiency curve [24], gives more weight to the wavelengths that are more sensitive to the human visual system(HVS). The function $W(\lambda)$ is different for daylight vision (photopic) and night vision (scotopic), and it is linked to our perception of brightness. Luminance, Y, is generally given in $cd/m^2$ or $nit$ equivalent units.

## 2.2 Differentiable Optics

### 2.2.1 Ray Optics

The lens model based on ray optics are usually spherical and aspherical. Given a Cartesian coordinate system $(x, y, z)$, the $z$-axis coincides with the optical axis, while $(x, y)$ forms the transverse plane. Let $r = \sqrt{x^2 + y^2}$ and $\rho = r^2$. Then the height of the aspheric surface and its derivative are defined as:

$$h(\rho) = \frac{c\rho}{1 + \sqrt{1 - \alpha\rho}} + \sum_{i=2}^{n} a_{2i}\rho^i, \tag{2.2}$$

$$h'(\rho) = c\frac{1 + \sqrt{1 - \alpha\rho} - \alpha\rho/2}{\sqrt{1 - \alpha\rho}\left(1 + \sqrt{1 - \alpha\rho}\right)^2} + \sum_{i=2}^{n} a_{2i} i \rho^{i-1}, \tag{2.3}$$

where $c$ is the curvature, $\alpha = (1 + \kappa)c^2$ with $\kappa$ being the conic coefficient, and $a_{2i}$'s are higher-order coefficients. The implicit form $f(x, y, z)$ and its spatial derivatives $\nabla f$ are:

$$f(x, y, z) = h(\rho) - z, \tag{2.4}$$

$$\nabla f = \big(2h'(\rho)x, 2h'(\rho)y, -1\big). \tag{2.5}$$

Note that spherical surfaces are special cases of asperical surfaces when $\kappa = 0$ and $a_{2i} = 0\,(i = 2, \cdots, n)$.

### 2.2.2 Diffractive Optics

Despite using refractive optics as the imaging or projection lens, diffractive optical element (DOE) is a flexible alternative. DOEs bring new image posiblities such as depth estimation [15] large field-of-view imaging [13], EDOF [16], optimal sampling [17] and high dynamic range (HDR) imaging [18, 19].

DOEs are operated by means of interference and diffraction to produce arbitrary

distributions of light , and have the following inherent advantages. First, DOE can be fabricated on a thin sheet; second, a single DOE can perform multiple optical operations simultaneously to act as an efficient light modulation platform; third, DOE is easily to be easily written as a differentiable model as when DOE plays as the role of the aperture of the optical system , the relationship of the DOE and the PSF can be simplied as a Fourier Transform.

To design a DOE in an end-to-end fashion, the optical model usually contains the following parts.

**Point Light Source.** The optical model begins with a point light source placed a certain distance in front of the DOE plane. Like most camera systems, the PSF for our optical model is depth-dependent. We chose a 5 m focal point as a compromise for near-infinite scene depths.

The point source generates a spherical wave. Upon the arrival of the wavefront to the DOE plane, the phase of the wavefront can be expressed as

$$\mathbf{u}_{-} = A_0 e^{jk\sqrt{x'^2+y'^2+z^2}}, \tag{2.6}$$

where $A_0$ is the amplitude, $k = 2\pi/\lambda$ is the wavenumber, and $z$ is the distance from the point source and DOE center. When the pint light source is far enough from the camera, the wave can be approximated as a plane wave.

**DOE Layer.** We then use a DOE layer to modulate the incident wave and set the DOE plane as the aperture $\mathcal{A}(x', y')$ of the whole optical system. The modulated field can be expressed as

$$\mathbf{u}_{+} = \mathcal{A}(x', y')\mathbf{u}_{-} e^{jk(n_\lambda - 1)\mathbf{h}(x', y')}, \tag{2.7}$$

where $n_\lambda$ is the wavelength-dependent refractive index of the DOE and $\mathbf{h}(x', y')$ is the height map of the DOE.

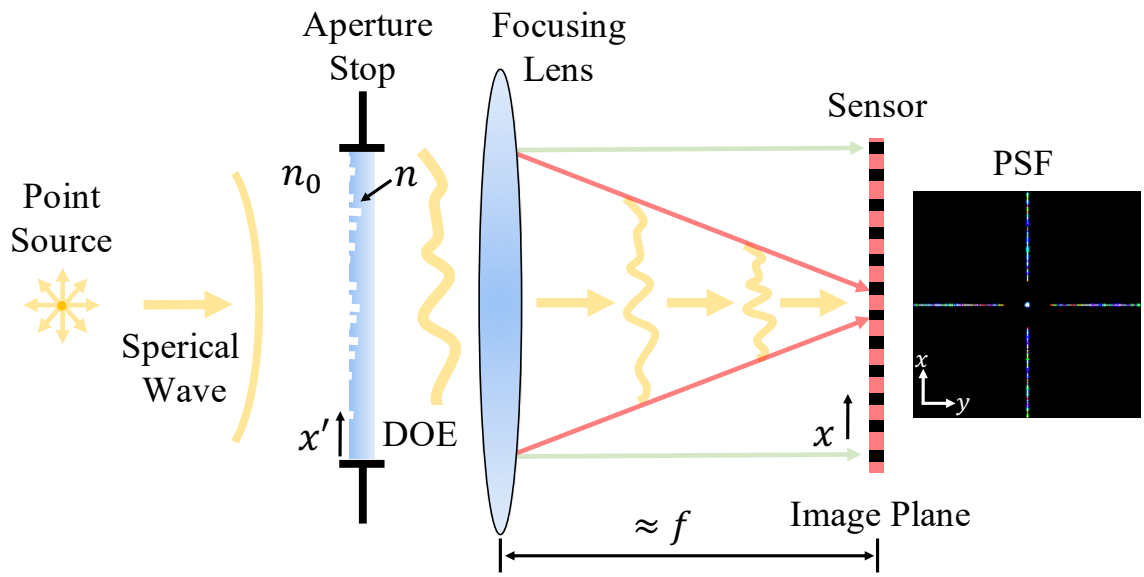


Figure 2.1: The optical forward model consists of a point light source which generates a spherical wave that is modulated by the DOE and focusing lens before being captured by the sensor. The corresponding PSF is used to simulate images.

The DOE layer brings a huge design space, but this model is based on paraxial approximation. In other words, the field of view (FOV) of the final optical system is limited. In addition, researchers usually take the DOE plane as a pixelwised variable map [14, 15, 25, 26, 27]. However, such kind of scenario has a huge variable map that is hard to train and converge to a good local minimum.

**Focusing-Lens Layer.** To free the design space of DOE, a focusing lens can be added. This lens is responsible for focusing the image, and allows the DOE to be purely optimized for the HDR encoding without also requiring the focusing operation for broadband illumination. The wave field $\mathbf{u}_l$ can be expressed as

$$\mathbf{u}_l = \mathbf{u}_+ e^{jk(f-\sqrt{x'^2+y'^2+f^2})}. \tag{2.8}$$

**Fresnel Propagation Layer.** We use the Fresnel approximation here to describe the field propagation from the focusing-lens to sensor. Specifically, the field $\mathbf{u}_s$ at the

sensor plane can be expressed as

$$\mathbf{u}_s = \mathcal{F}^{-1}\{\mathcal{F}\{\mathbf{u}_l\}\mathcal{H}\}, \tag{2.9}$$

where $\mathcal{H}(f_x, f_y) = e^{jkL}e^{-j\pi\lambda L(f_x^2+f_y^2)}$, with $f_x = 1/2\Delta x'$ and $f_y = 1/2\Delta y'$, is the Fresnel propagation kernel and $L$ is the distance between the normal lens and sensor plane. Finally, the PSF corresponding to the entire image formation model is given by $\mathbf{p} \propto |\mathbf{u}_s|^2$.

## 2.3 Optical Aberrations and PSF

Optical systems suffer from both monochromatic and chromatic aberrations [28], which are derived from the optical path deviations when light travels through different regions of an objective lens (i.e., having a certain aperture size and thickness) from different incident angles.

**Monochromatic aberrations.** Monochromatic aberrations describe the deviations from the idealized linear Gaussian optics [29] occur even for a single wavelength. Five primary aberrations are classified into this type, including spherical aberration, coma, astigmatism, field curvature, and distortion. The aberrations of a real lens are interpreted as the fact that the beam emitted from a point source may be focused at many points that vary in the spatial domain.

**Chromatic aberrations.** Chromatic aberrations result from the wavelength-dependency of the refractive index of the material. When imaging a white point light source, the shift or spread of the imaged spot along the optical axis is denoted as axial chromatic aberration. The remaining component in the image plane is denoted as lateral chromatic aberration [30]. Chromatic aberrations present one of the critical contributions for preserving color fidelity in image signal processing.

All these aberrations, including both monochromatic and chromatic aberrations, are visualized as the unwanted spatial and spectral variant blur of the image that may become more severe for a lens having a relatively large numerical aperture or a small f-number. The aberration theory was firstly introduced by Seidel in 1857 [31]. From this, lens design in optics attempts to eliminate aberrations by designing increasingly complex lens structures stacked with many elements [32]. This involves designing aspherical surfaces as well as finding materials with better optical properties. Therefore, lens design is a heuristic compromise among various optical evaluation criteria [33, 34]. Accordingly, state-of-the-art commercial optical design toolbox, like Zemax, Code-V, etc., has been intensively used in designing refractive lens profiles for a wide range of applications. Intuitively, a high-level insight in the design space is to enforce the focusing contribution to be as clear as possible subject to field-of-views and wavelengths. Researchers recognized that optical aberrations could be deliberately minimized until recent decades but designed to exhibit other behavior for obtaining extra information of the scene.

Our physical world is a 3D world, with multiple different light sources which vary with wavelengths, polarization state, and intensity distributions. Objects in our world have their shapes and materials. Each material has its reflectance property which can be described as Bidirectional Reflectance Distribution Function and spectrum reflectance. Our camera, with the assistance of optimized illumination, optimized optics, optimized sensors, optimized post-processing, and optimized visualization. It is possible to rebuild a virtual physical world through computational photography!

## 2.4 Ditital Cameras, Imaging Model and Key Trade-off

Most modern digital cameras use CCD (charge-coupled device) or CMOS (complementary metal-oxide-semiconductor) chips to record radiant energy. Their role is to convert incident radiant energy into readable digital numbers. Each photodetector

is called a pixel, which has a p–n junction that converts light photons into the current. The ratio of the number of emitted photoelectrons and absorbed photons are quantum efficiency (QE), which is related to the photon energy of the light.

The shutter is a photographic device that administers the exposure by limiting the time $T$ over which light is admitted. For the usual digital sensor, the accumulated photoelectrons are read out as an amplified analog signal through an analog-to-digital converter (ADC). As a result, the unprocessed raw image can be expressed as:

$$I_c(x', y') = \int Q_c(\lambda) \cdot [p(x', y', d, \lambda, s_c) * s_c(x', y', d)] d\lambda + n(x', y'), \tag{2.10}$$

where the PSF $p(x', y', d, \lambda, s_c)$ is a function with spatial position $(x', y')$ on the sensor, the depth $d$ of scene, and the incident spectral distribution $\lambda$. $Q_c$ is the color response of the sensor, and $s_c(x', y', d)$ and $n(x', y')$ represent the latent scene and measurement noise (white Gaussian noise), respectively. The PSF $\mathbf{p}(x', y', d, \lambda, s_c)$ is controllable by optics design and is the key trade-off in end-to-end camera design. .

## 2.5 Image Recovery Model

Image reconstruction is another critical stage for an end-to-end computational camera. Usually, the reconstruction is formulated as an optimization problem of a data fidelity term with an additional regularization term:

$$\boldsymbol{I} = \mathbf{arg\,min}_{\boldsymbol{I}} \frac{1}{2} \|\boldsymbol{p} * \boldsymbol{I} - \boldsymbol{I}_s\|_2^2 + \beta \|\Phi(\boldsymbol{I})\|_1, \tag{2.11}$$

where $\Phi(\cdot)$ denotes the transform coefficients of the ground truth $\boldsymbol{I}$ with respect to some transform $\Phi$ that can be either linear or optimized non-linear. Sparsity in the transform space $\Phi(\boldsymbol{I})$ is encouraged by the $\ell_1$ norm with $\beta$ being a regularization parameter.

Usually, natural images are non-stationary in classic domains like DCT, gradients,

and wavelets, which may result in an ill-posed problem under such an imaging model. If the image degeneration model is fixed, the image reconstruction can be directly solved by optimizing the cost function 4.2. For example, with a fixed PSF $\boldsymbol{p}$ like a gaussian kernel, the transform $\Phi$ represents the gradient operator. Then we can solve this cost function through the well-known alternating direction method of multipliers (ADMM) method or some other optimization methods such as the proximal operator. Another condition is that the PSF varies with the scene, like the incident spectral distribution, spatial position, etc. Such kind of situation becomes hard to solve as the degeneration model is not fixed. With the development of deep learning, image recovery for the scene-dependent condition become possible. The key to solving this inverse problem is to simulate the corruption well to train the image recovery network mapping to the target result. The image recovery network can be tailored to a specific application.

# Chapter 3

# Joint Optics Design with Image Recovery: Learned Large FOV Imaging

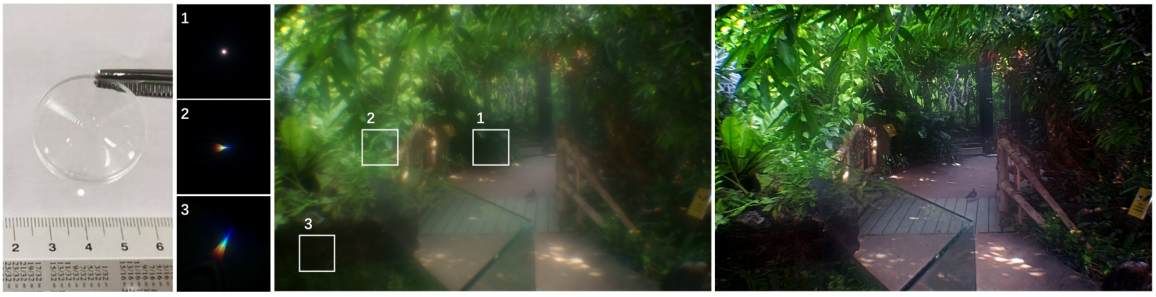


Figure 3.1: Results of LFOV imaging with thin-plate optics. We design a lens with compact form factor using one (or two) optimized refractive surfaces on a thin substrate (left). This optimization results in a dual-mixture point spread function (center-left insets), which is almost shift-invariant to the incident angle, exhibiting a high-intensity peak and a large, almost constant, tail. We show the sensor measurement (center) and image reconstruction (right) in natural lighting conditions, which demonstrate that the proposed deep image recovery effectively removes aberrations and haze resulting from the proposed thin-plate optics. Our prototype single element lens achieves a large field-of-view of 53° with a clear aperture of F1.8 and effective aperture of F5.4, see text.

This chapter initializes the joint optics design through the application of large FOV imaging. Typical camera optics consist of a system of individual elements designed to compensate for the aberrations of a single lens. Recent computational cameras shift this correction task from optics to image signal processing, reducing the imaging optics to only a few optical elements. However, these systems only achieve reasonable image quality by limiting the FOV to a few degrees – effectively ignoring severe off-axis aberrations with blur sizes of multiple hundred pixels.

In this chapter, we present a lens design and learned reconstruction architecture

that lifts this limitation and provides an order of magnitude increase in FOV using only a single thin-plate lens element. Specifically, we design a lens to produce spatially shift-invariant point spread functions over the full FOV tailored to the proposed reconstruction architecture. We achieve this with a dual-mixture PSF, consisting of a peak and a low-frequency component, which provides residual contrast instead of a small spot size as in traditional lens designs. To perform the reconstruction, we train a deep network on captured data from a display lab setup, eliminating the need for manual acquisition of training data in the field. We assess the proposed method in simulation and experimentally with a prototype camera system. We compare our system against existing single-element designs, including an aspherical lens and a pinhole, and we compare against a complex multi-element lens, validating high-quality large field-of-view (i.e., 53°) imaging performance using only a single thin-plate element.

## 3.1 Introduction

Modern imaging techniques have equipped us with powerful capabilities to record and interact with the world – be that in our devices, assistive robotics, or self-driving vehicles. Coupled with recent image processing algorithms, today's cameras can tackle high-dynamic range and low-light scenarios [35, 36]. However, while image processing algorithms have been evolving rapidly over the last decades, commercial optical systems are primarily still designed following aberration theory, i.e., with the design goal of reducing deviations from Gauss's linear model of optics [29]. Following this approach, commercial lens systems introduce increasingly complex stacks of lens elements to combat individual aberrations [30]. For example, the optical stack of the iPhone X contains more than six aspherical elements, and professional zoom optics can contain more than twenty individual elements.

Although modern lens systems are effective in minimizing optical aberrations, the

depth of the lens stack is a limiting factor in miniaturizing these systems and manufacturing high-quality lenses at low cost. Moreover, using multiple optical components introduces secondary issues, such as lens flare and complicated optical stabilization, e.g., in a smartphone where the whole the lens barrel is actuated. In particular, the design goals of large field-of-view (FOV, e.g., $> 50°$), high numerical aperture (NA), and high resolution (e.g., 4k resolution) stand in stark contrast to a compact, simple lens system.Existing approaches address this challenge using assemblies of multiple different lenses or sensors [37, 38, 39, 40, 10], including widely deployed dual-camera smartphones, each typically optimized for a different FOV. While providing some reduction in footprint, such spatial multiplexing increases the number of optical elements even further and requires higher bandwidth, power, and challenging parallax compensation post-capture [40].

In this work, we deviate from traditional lens design goals and demonstrate high-quality, monocular *large*-FOV imaging using a single deep Fresnel lens, i.e., a thin lens with a microstructure allowing for larger than $2\pi$ modulation. Specifically, we propose a learned generative reconstruction model, a lens design tailored to this model, and a lab data acquisition approach that does not require the painful acquisition of real training images in the wild.

The learned reconstruction model allows us to recover high-quality images from measurements degraded by severe aberrations. Single-lens elements, such as spherical lenses or Fresnel phase plates [41], typically suffer from severe off-axis aberrations that restricts the usable FOV to around $10°$ [42, 41, 43]. Instead, we propose a novel lens design that offers spatially invariant PSFs, over the full FOV designed to allow aberration removal by the proposed learned reconstruction model. We achieve this by abandoning the design goal of minimal spot size and instead balance the local contrast over the full FOV. This alternative objective allows us to build on existing optimization tools for the optics of the proposed co-design, without requiring end-to-

end design. The resulting thin lens allows the reconstruction network to detect some contrast across the full FOV, invariant of the angular position, at the cost of reducing the contrast in the on-axis region. As a consequence, the proposed computational optics offers an *order of magnitude larger FOV* than traditional single lenses, even with the same reconstruction network fine-tuned to such alternative designs.

The following technical contributions enable large FOV imaging using thin, almost planar, optics:

- We propose a single free-form lens design tailored to learned image reconstruction methods for large FOV high-quality imaging. This design exhibits almost invariant aberrations across the full FOV that balance the contrast detection probability (CDP) of early network layers.

- We propose a generative adversarial model for high-resolution deconvolution for our aberrations of size $\leq 900$ pixels.

- The model is trained on data acquired with a display lab setup in an automated manner, instead of painful manual acquisition in the field. We provide all models, training and validation data sets.

- We realize the optical design with two prototype lenses with effective thickness of 120 $\mu$m, aperture size of 23.4 mm, and a FOV of 53° – one with a single optical surface, the other with two optical surfaces (both sides of the same flat carrier). We experimentally validate that our approach offers high image quality for a wide range of indoor and outdoor scenes.

**Overview of Limitations** We note that, compared to conventional digital cameras, the proposed reconstruction method requires more computational resources. Although our thin-plate lens design reduces the form factor compared to complex optical systems, its back focal length is comparable to conventional optics.

## 3.2 Related Work

### 3.2.1 Aberrations and Traditional Lens Design.

Both monochromatic and chromatic aberrations are results of the differences of the optical path length when light travels through different regions of a lens at different incident angles [44]. These aberrations manifest themselves as unwanted blur, which becomes more severe with increasing numerical aperture and field-of-view [45]. Conventional lens design aims at minimizing aberrations of all kinds by increasingly complex lens stacks [32]. This includes designing aspherical surfaces and introducing lens elements using materials with different optical properties.

State-of-the-art optical design software is a cornerstone tool for optimizing the surface profiles of refractive lens designs. However, while hyper-parameter optimization tools are becoming mature, the design process still relies on existing objectives, so-called *merit functions*, that find a compromise across a variety of criteria [34, 46], trading off the point spread function (PSF) shape across sensor locations, lens configurations (e.g., zoom levels) and target wavelength band.

### 3.2.2 Computational Optics.

A large body of work on computational imaging [1, 47, 48, 4] has proposed to design optics for aberration removal in post-processing. These methods often favor diffractive optical elements (DOEs) over refractive optics [49, 50, 43, 51] because of their large design space. Moreover, recent work proposed caustic (holographic) designs, for projection displays or imaging lenses [52, 53, 54]. To simplify the inverse problem in post-processing, all of the described approaches ignore off-axis aberrations by restricting the FOV to a few degrees – existing approaches do not realize monocular imaging with a large FOV.

Several approaches to end-to-end optical imaging were recently proposed, where

parametrized optics and image processing are jointly optimized for applications in EDOF and superresolution imaging [14], monocular depth estimation [25, 55, 16], and image classification [56]. However, none of these approaches aim at large FOV imaging, and all of them build on simple paraxial image formation models, which break for large fields of view. Moreover, they are limited to a single optical surface. We overcome these challenges by engineering PSFs over a large FOV and, relying on existing optical design tools that support complex multi-surface/material designs, optimize for a well-motivated dual-mixture design tailored to deep reconstruction models.

### 3.2.3 Manufacturing Planar Optics.

Various manufacturing methods enable "planar" optics with the low-depth optical surface, i.e., less than 1 mm. Commercial miniature form factor optics like the lenses in smartphone cameras, can be manufactured using mature injection molding techniques [57]. Alternative fabrication methods for thin-plate lenses include diffractive optics and metalenses [58, 59], which require nano-fabrication methods like photolithography and nano-imprinting [60, 61]. The UV-cure replication technique [62] can facilitate manufacturing wafer-scale optical elements. Note that creating a Fresnel lens with a clear aperture diameter of 23.5 mm and a focal length of 43 mm requires, as in this work, a feature size smaller than 300 nm, which is beyond the capability of the photolithography methods used in many recent DOE works [43, 51, 14]. Freeform lenses with a larger aperture and continuous surfaces can be manufactured using diamond turning machining [63]. The continuous surface preserves light efficiency and works under broadband illumination, while the lenses are usually thick and bulky because of the local curvature constraints.

In this work, we use high-precision diamond turning machining for prototyping the proposed lenses. Instead of fabricating a freeform lens with a continuous surface,

e.g., as in [14], we wrap the optimized surface profile using coarse wrap-around depth values instead of wavelength-scale wrapping in diffractive lens designs, see Fig. 4.1. This allows us to design a Fresnel-inspired free-form lens with the advantages of both refractive optics and diffractive optics: we achieve a thin form factor while reducing chromatic aberrations.

### 3.2.4 Image Quality.

Imaging describes the signal chain of light being transported from a scene patch of interest to the camera, focusing on the camera optics, digitization of the focused photon flux on the sensor, and post-processing of the measured data. During each of these individual steps, information about the scene patches of interest may be lost or corrupted. Various hand-crafted image quality metrics exist that measure the cumulative error of this imaging process [64, 65], with or without known ground-truth reference [66], or allow to individually characterize components of the imaging stack using calibration setups [67, 68]. Typical performance metrics are the signal-to-noise ratio. (SNR) [69] and modulation transfer function (MTF) [70, 67]. While these metrics are widely reported, and measurement setups are readily available, they are also not free from disadvantages due to their domain-agnostic design. For example, high SNR does not guarantee a perceptually pleasing image, which has sparked recent work on perceptual loss functions [71]. Moreover, SNR increases in the presence of glare and quantization, which can yield inconclusive results when used as a design metric [72].

We design the proposed optical system in conjunction with the learned image reconstruction methods. To this end, we analyze the behavior of the early layers in our generator, which relate to the response of local contrast features in the scene. Relying on a probabilistic measure [72], we assess the ability to detect or miss such local features across the whole FOV. This insight allows us to tailor the proposed lens

design to our network-based reconstruction method.

### 3.2.5 Learned Image Reconstruction.

Traditional deconvolution methods [73, 74, 42] using natural image priors are not robust when working with extremely large, spatially invariant blur kernels that exhibit chromatic aberrations and other challenging effects. Unfortunately, the lens design proposed in this work produces large PSFs that present a challenge to existing deconvolution methods which suffer in image quality for large aberrations, necessitating a custom image reconstruction approach. Note that computationally efficient forward models for large spatially-varying convolutions have been investigated before [75].

Over the last years, a large body of work proposed data-driven approaches for image processing tasks [76, 77, 78]. Specifically addressing deconvolution, Nah et al. [79] propose a fully connected convolutional network that iteratively deconvolves in a multi-stage approach. More recently, generative adversarial networks (GANs) have been shown to provide generative estimates with high image quality. Kupyn *et al.* [80] demonstrate the practicability of applying GAN reconstruction methods to deblurring problems.

All of these approaches have in common that they require either accurate PSF calibration or large training data that has been manually acquired. In contrast, we propose a lab capture process to generate a large training corpus with the PSF encoded in the captured data. Note that the large aberrations make training on very small image patches prohibitive. The proposed automated acquisition approach allows for supervised training on a very large training set of full-sized images, which are needed to encode large scene-dependent blur. The training approach, together with the proposed model and loss function, allows us to tackle the large scene-dependent blur, color shift, and contrast loss of our thin-plate lens design.

## 3.3 Designing Optics for Learned Recovery

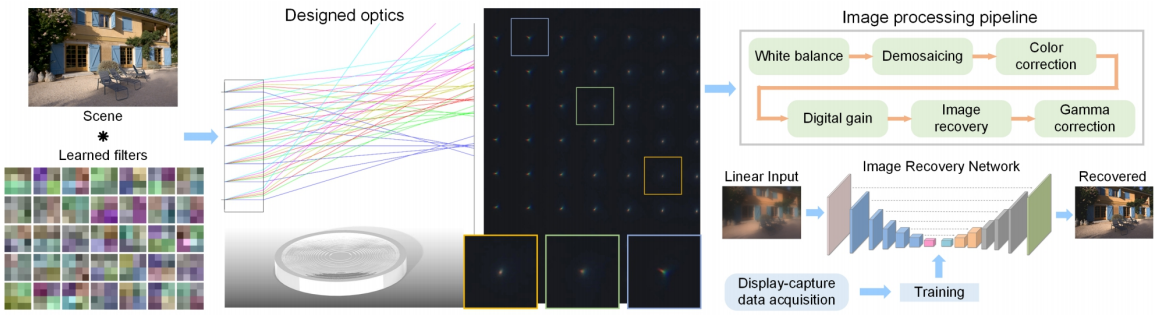


Figure 3.2: Computational thin-plate lens imaging with large field-of-view. Left: learned early layers' filters applied on input scene; Center-left: optical design that preserves local contrast across full FOV. The designed optical element has a Fresnel lens surface; Center-right: calibrated PSF patches for different incident angles of our prototype lens (images gamma-tonemapped for visualization); Right: overview of the image processing pipeline and our recovery framework, which learns a mapping for the linear input to the recovered output. We introduce a learned reconstruction architecture trained using data that can be efficiently acquired in a display-capture lab setup (see details in Section 6).

In this work, we describe an optical design tailored to learned reconstruction techniques for large field-of-view, thin-plate photography. The proposed optical system is shown in Figure 3.2. In contrast to state-of-the-art compound lenses, it consists of a single, almost flat, element. The two core ideas behind the proposed optical design are the following: first, to achieve a large FOV, we constrain the PSFs of our lens to be shift-invariant for the incident angle. Second, although such PSFs exhibit large spot sizes, we engineer aberrations that preserve residual contrast and hence are well-suited for learned image reconstruction.

Our design is motivated by a large body of work on computational photography with PSF engineering – designing PSFs invariant to a target characteristic, instead of minimizing spot size, and computation to remove the non-compact aberrations. Similar to how existing work EDOF [81] or spectral range [82, 51], we are the first to apply this idea to extending the FOV.

To this end, we rely on the insight that the filters of early layers of recent deep

models, across applications in computer vision and imaging, have striking similarity – these early layers are gradient-like filters and respond to local contrast as essential low-level information content in the measurement. As many recent learned architectures rely on common low-level backbones, which are then transferred to different higher-level tasks [83, 84], this transfer-learning offers an interesting opportunity for the design of imaging systems.

We engineer the PSFs of the proposed optical design, shown in Figure 3.2, to exhibit a *peaky* distribution. While the peak contribution maximizes the probability of detecting local contrast features, the low-frequency part is extremely large ($\sim$900 pixels on the experimental sensor system covered below) and therefore leads to very low filter responses in the early layers. In contrast to conventional spherical elements, see Figure 3.7, this PSF exhibits the peak-preserving distribution across the full sensor which enables large FOV imaging with this single optical element.

Given a raw measurement acquired with the proposed thin-plate lens system, we recover a high-quality image using a generative adversarial network which is trained to eliminate all measurement degradations and directly outputs a deblurred, denoised, and color-corrected image, see Figure 3.2. To train the network in a semi-supervised fashion, using labeled and unlabeled data to learn robust loss functions along with the model parameters, we require a training dataset with ideal reference images and corresponding blurry captures. Instead of manually acquiring such a dataset, e.g., by sequentially swapping optics for a scene, we propose an automated lab setup which displays known ground-truth images on a display.

In the following, we first describe the proposed optical design in Section 3.4, before introducing the reconstruction architecture and training methodology in Section 3.5.

## 3.4 Lens Design

Throughout the rest of this chapter, we consider rotationally symmetrical designs. Although our approach can be generalized to rotationally asymmetrical profiles, rotational symmetry facilitates manufacturing using turning machines.

### 3.4.1 Ideal Phase Profile

The phase of a lens describes the delay of the incident wave phase introduced by the lens element, at the lens plane. The geometrical (ray) optics model, commonly used in computer graphics, models light as rays of photon travel instead of waves. This model ignores diffraction, e.g. for light passing through a narrow slit. Although being an approximation to physical optics, ray optics still can provide an intuition: the perpendiculars to the waves can be thought of as rays, and, vice versa, phase intuitively describes the relative delay of photons traveling along these rays to the lens plane, as illustrated with red lines in Fig. 3.3. Hence, the phase of a thin lens is its height profile multiplied with the wave number and the refractive index [85, 86].

We design the proposed lens by first specifying an ideal phase profile for perfect, spatially invariant PSFs over the full FOV, i.e., mapping incident rays from one direction to one single point. Because it will turn out intractable to manufacture this ideal lens, we propose an aperture partitioning strategy as an approximation. The deviation of this partitioned phase profile to the ideal profile is a large low-frequency component which is independent of the incident angle. Together with the peak-component, which preserves local contrast over the full FOV, these two components make up the desired spatially invariant dual-mixture PSF.

To specify the ideal phase profile $\phi(r, \omega_i)$ for an incident ray direction $i$, and radial position $r$, see Figure 3.3, we assume a physical aperture size $D$, focus distance $f$,

and set:

$$\phi(r, \omega_i) = -k \left[ r \cdot \sin\omega_i - \int_0^r \sin\theta(r_1, \omega_i) dr_1 \right], \tag{3.1}$$

where $k$ represents here the wave number that is specified by the wavelength, and $\omega_i$ represents the incident angle of ray direction $i$ [87]. For this ideal lens profile, we define the output angle as:

$$\theta(r, \omega_i) = \arctan\left( \frac{\rho(\omega_i) - r}{f} \right), \tag{3.2}$$

since the ideal lens design maps the incident rays from one direction $\omega_i$ to a single point with spatial position $\rho(\omega_i)$ on the image plane.

Next, by inserting Eq. 3.2 into Eq. 3.1, we derive the target phase $\phi$ as:

$$\begin{aligned} \phi(r, \omega_i) &= -k \left[ r \cdot \sin\omega_i - \int_0^r \frac{\rho(\omega_i) - r_1}{\sqrt{f^2 + (\rho(\omega_i) - r_1)^2}} dr_1 \right] \\ &= -k \left[ r \cdot \sin\omega_i + \sqrt{f^2 + (\rho(\omega_i) - r)^2} - \sqrt{f^2 + \rho(\omega_i)^2} \right]. \end{aligned} \tag{3.3}$$

The ideal phase profile from Eq. 3.3 is visualized in Figure 3.3 (right). We observe a drastic variation when approaching larger incident angles. In other words, the same position on the lens aperture would need to realize different phases for different incident angles, which is not physically realizable with thin plate optics.

### 3.4.2 Aperture Partitioning

Realizing the ideal phase profile is intractable to manufacture over the full aperture, as illustrated by the large angular deviations needed in off-axis region in Figure 3.3. To overcome this challenge, we split the aperture into multiple sub-regions, and assign each sub-region to a different angular interval, similar to prior work [4, 88] for refractive optics. We note that this concept is also closely related to specializing optics depending on the incident ray direction in light field imaging [89], for example,

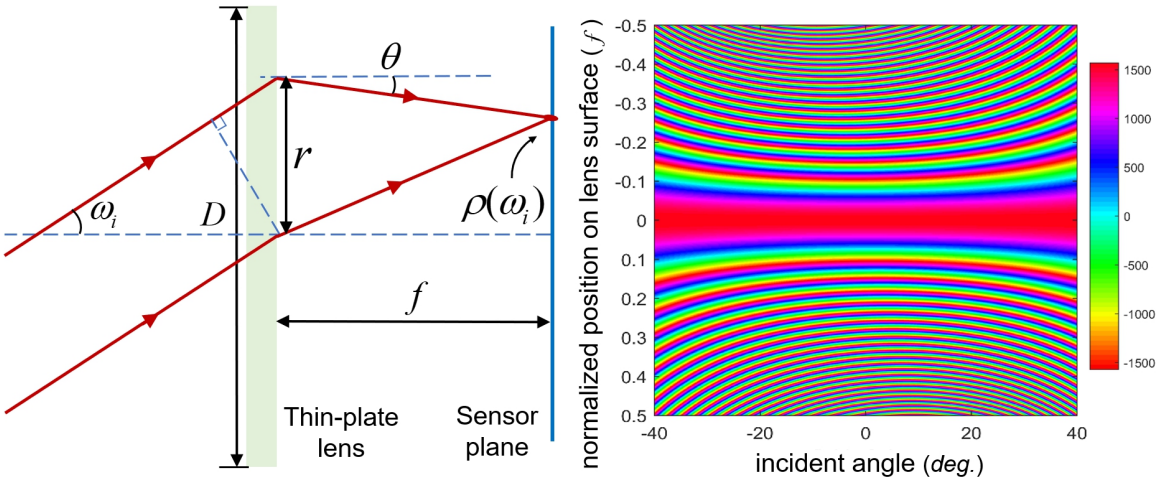


Figure 3.3: Schematic of ray geometries in a radially symmetric design manner (left) , and the ideal phase profile distribution subject to incident angle (right). For visualization purpose the phase map is wrapped by 1,000$\pi$, and the vertical axis is normalized with respect to the focal length.

tailoring optical aberrations for digital correction [90]. Specifically, we introduce a *virtual aperture* $\mathcal{A}(r, \omega_i) = circ[r - \nu(\omega_i)]$ to partition the incident light bundle of each direction into a peak component that we optimize for, while treating out-of-aperture components as out-of-focus blur. Here, $circ[\cdot]$ is a function representing a circular aperture, $\nu(\omega_i)$ indicates the axial center of the virtual aperture subject to the $i^{\text{th}}$ incident ray direction. With this aperture partitioning, we optimize for the phase profile solving the following optimization problem:

$$[\phi_0(r), \rho, \nu] = \underset{\phi_0(r),\rho,\nu}{\arg\min} \sum_{i=1}^{N} \|\mathcal{A}(r, \omega_i)(\phi_0(r) - \phi(r, \omega_i))\|_2^2. \tag{3.4}$$

Note that the virtual aperture is not a physical aperture of the optical system, but is only introduced as a conceptual partitioning in the lens optimization. Figure 3.4 shows the virtual apertures for uniformly sampled directions superimposed on the real aperture. For every direction, we optimize only for the rays that pass through the corresponding virtual apertures; these will be focused into a sharp PSF, while all other rays from the same direction that miss $\mathcal{A}$ but pass through the full aperture $D$

will be blurred and manifest as a low frequency "haze" in the measurement.

### 3.4.3 Fresnel Depth Profile Optimization

We solve the optimization problem from Eq. 3.4 using Zemax [33]. While Eq. 3.4 minimizes phase differences, Zemax interprets it as minimizing the optical path difference (OPD). Zemax allows us to piggy-back on a library of parameterized surface types, and directly optimize a deep Fresnel lens profile (a deeper micro-structure than regular $2\pi$ modulation.) instead of sequentially optimizing for the phase and depth in a two-stage process. We formulate the problem from Eq. 3.4 using the multiple configuration function with the number of the configurations set to the discretized aperture directions (7 in this work, uniformly sampled on half of the diagonal image size). We set the size of each virtual aperture – the effective aperture that contributes to focusing light bundles – to one third of the clear aperture. As shown in Figure 3.4, the center $\nu$ of the virtual aperture for each direction along the clear aperture plane can be modeled by shifting a stop along the optical axis. This allows us to optimize the location of the virtual aperture by setting the stop position as an additional optimization variable. The merit (objective) function used in Zemax includes terms for minimizing the wavefront (phase) error at each sampled direction, and enforcing a desired effective focal length (EFL). We refer the reader to the supplementary document for additional details.

### 3.4.4 Aberration Analysis

The optical aberrations of the proposed design have the following properties. The chromatic variation is small because a deep Fresnel surface results in only small focal length differences in the visible wavelength region. Off-axis variation (i.e. spatial intensity variation of PSFs across FOV) are small since we only control a part of light of each direction to focus into the sharp peak (see Figures 2 and 7).

For each viewing direction, the PSF exhibits two components, a high-intensity

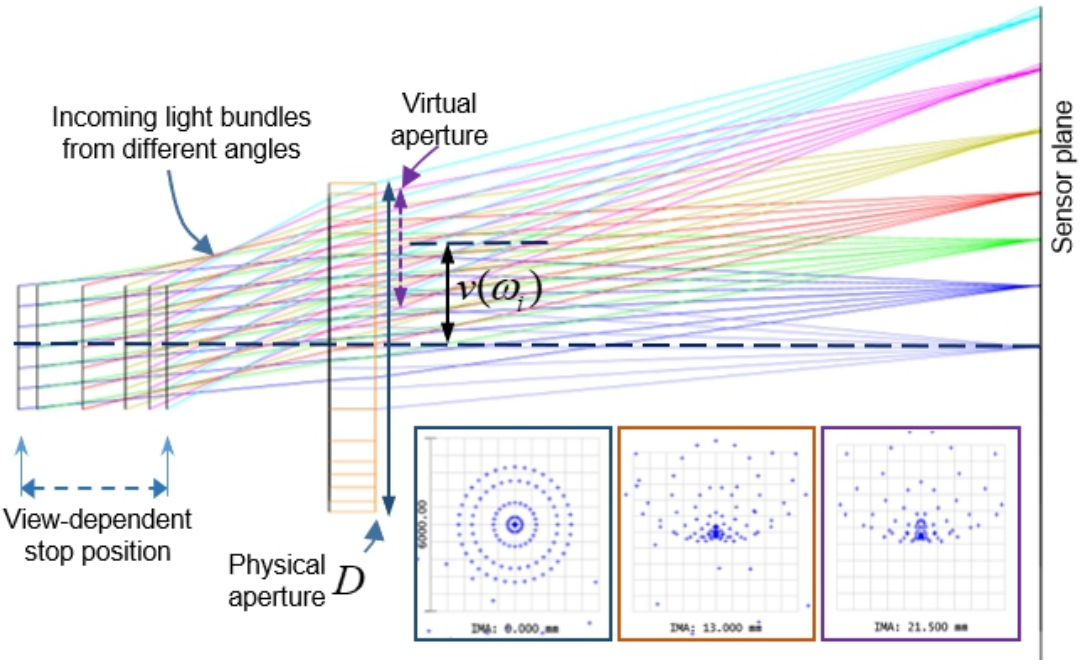


Figure 3.4: Schematic of aperture partitioning approach. The spatial position of a virtual aperture (specified by the offset from the optical axis and visualized with different colors) along the radial direction is determined by the controlled position of the stop (on the left), that is further dependent on incident ray directions. The synthetic spot distributions of three directions are presented as inserts, from each pattern we observe a sharp peak that fits well to our design goal of PSFs.

peak, which preserves local contrast, and a large low-frequency component. We note that this property differs from conventional spherical or aspherical singlets with the same NA whose field curvature can be severe. Although the low-frequency PSF component reduces contrast, it does so uniformly across the FOV. In contrast to conventional single element optics, which have very poor contrast in regions far from the optical axis (required for wide-FOV imaging), it is this design which allows us to preserve the ability to detect some residual contrast, instead of completely losing contrast.

## 3.5 Learned Image Reconstruction

In this section, we describe the forward image formation model, which models sensor measurements using the proposed optical design, and we present our learned recon-

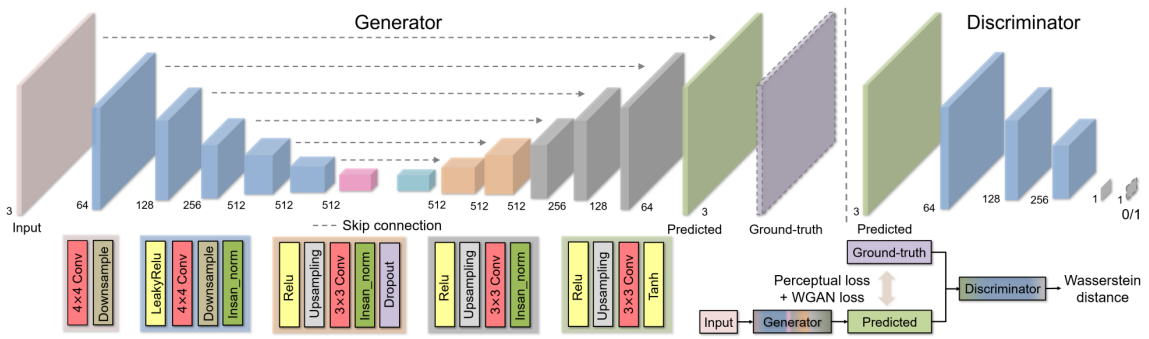


Figure 3.5: Generative image reconstruction architecture. The generator model is shown on the left and the layer configurations of encoder/decoder stages are illustrated with different colored blocks (bottom-left). We apply skip connections in every decoder stage. In particular, we use a combination of a perceptual loss between the predicted image and the ground-truth, and a Wasserstein generative adversarial loss. The discriminator model for the GAN loss is similar to the encoder architecture.

struction model which retrieves high-quality images from these measurements.

### 3.5.1 Image Formation Model

Modern digital imaging consists of two main stages: a first stage which records scene information in measurement via optics and a sensor, and a second stage which extracts this information from the measurements using computational post-processing techniques.

In the recording stage, a sensor measurement $b_c$ for a given color channel $c$ can be expressed as:

$$b_c(x,y) = \int Q_c(\lambda) \cdot [p(x,y,d,\lambda,i_c) * i_c(x,y)] d\lambda + n(x,y), \tag{3.5}$$

where the PSF $p(x,y,d,\lambda,i_c)$ varies with the spatial position $(x,y)$ on the sensor, the depth $d$ of scene, and the incident spectral distribution $\lambda$. $Q_c$ is the color response of the sensor, and $i(x,y)$ and $n(x,y)$ represent the latent image and measurement noise, respectively. The PSF may also exhibit non-linearity in high-intensity regions, which is why the PSF $p$ takes the latent channel $i_c$ as further parameter. The noise

may have complex characteristics, including signal-dependent shot noise as well as read noise introduced in the measurement process. We refer readers to the EMVA Standard [68] for a detailed discussion of noise sources and calibration of the proposed model from Eq. (3.5).

Given a sensor measurement, conventional image processing pipelines perform a sequence of operations, each addressing an individual reconstruction problem, such as white balance, demosaicing, color calibration, digital gain, gamma compression and tone mapping [91]. Errors occurring during any of these operations can accumulate, adding to the ill-posedness of the overall image reconstruction problem [92, 93], that is recovering $i$ from $b$ by inverting Eq. 3.5.

To recover a latent image from the degraded image, existing methods typically perform deconvolution using optimization [74], addressing the ill-posedness of the reconstruction problem using natural image priors. We refer to the supplementary document for details on traditional deconvolution methods. However, large PSFs with hundreds of pixels in diameter and high wavelength and depth-dependency cannot be tackled by existing methods. While the scene dependency of the aberrations may be addressed with blind deconvolution approaches, these methods are currently limited to small PSF sizes of ca. 10-20 pixels in diameter [94]. Hence, existing image reconstruction methods cannot compensate for the low-frequency tail of the proposed PSF and scene-dependent PSF variation, as shown in Figure 8.

To handle the scene-dependence and non-linearities in the image formation model, i.e., PSF dependency on $i$ in Eq. (3.5), we deviate from existing methods in that we do not pre-calibrate a PSF for a given illumination, and approximate the scene with broadband spectral response, but instead solve for a given image without an intermediate PSF estimate. This is done by directly learning a image-to-image mapping using a deep neural network. Next, we describe the network architecture, training methodology, and training data acquisition.

### 3.5.2 Generative Image Recovery

We propose a generative adversarial network (GAN) for the retrieval of the latent clean image $i$ from corrupted raw sensor measurements $b$. Instead of relying on existing hand-crafted loss functions, which encourage overfitting as we will show below, using a GAN allows us to learn a robust loss function along with the reconstruction model. Moreover, in the learning of this loss function, we can augment training pairs for supervised training with unpaired training data from high-quality lens captures. The proposed framework is shown in Figure 3.5. Specifically, we adopt a variant of the U-Net architecture [95], as our generative model $G$, referred to as *Generator* in the following, while the discriminative critic network $D$ is referred to as *Discriminator*. During training, the generator is trained to produce latent estimates that "fool" the discriminator network into classifying the estimate as a high-quality image, while this discriminator is trained to better distinguish between images from compound lenses and the estimates produced from the generator. We use training data without blurry correspondences to augment the training of the discriminator, in a semi-supervised fashion, while the generator model is trained using a combination of a learned perceptual loss between the predicted image and the reference, and the discriminator loss using a Wasserstein generative adversarial framework.

#### Network Architecture

**Generator.** The proposed generator network consists of a contracting path and an expansive path (Figure 3.5). Specifically, the contracting path consists of a 4×4 initial feature extraction layer, the repeated application of the Leaky rectified linear unit (LeakyReLU), a 4×4 convolution layer with stride 2 for downsampling, and instance normalization layers. The LeakyReLU allows back-propagating the error signal to the earlier layer and the instance normalization (i.e. single batching training in this work), to avoid the crosstalk between samples in a batch. At each downsam-

pling convolution, we double the number of feature channels. The total number of downsampling convolution steps is 7.

The expansive path consists of a stack of rectified linear units (ReLUs), the upsampling convolution layer, and the instance normalization. We use a nearest neighbor upsampling and a 3×3 convolution layer instead of the transposed convolution, that practically reduces the checkerboard artifacts caused by uneven overlaps [96]. Moreover, as shown in Figure 3.5, we concatenate the feature maps from the contracting path to introduce high frequencies so as to preserve fine scene details.

**Discriminator.** As illustrated in Figure 3.5, the discriminator consists of five 4×4 convolution layers with stride 2 for downsampling, where each layer is followed by a LeakyReLU activation layer and instance normalization, except for the first. We also double the number of feature channels after each downsampling layer. See Figure 3.5 and its caption for additional detail.

## Loss Functions

**Perceptual loss.** Feed-forward CNNs are often trained using a per-pixel loss (e.g. usually $\ell_1$ or mean absolute error (MAE) loss and $\ell_2$ or mean square error (MSE) loss) between the output and the ground-truth labels. However, this approach may lead to overly blurry outputs due to the pixel-wise average of possible optima [97]. To obtain visually pleasing results that generalize to real data, we add a perceptual loss [71] to our learned GAN loss. This loss component compares two images subject to the high-level representations from the pre-trained CNN. We use the VGG19 network in all our experiments. Let $\mathbb{A}_k(i)$ be the activations at the $k^{\text{th}}$ layer of the pre-trained VGG19 network $\Phi$ with an input image $i$. Given a feature map $\mathbb{A}_k(i)$ with the shape of $C_k \times H_k \times W_k$, the Gram matrix, with a size of $C_k \times C_k$, can be expressed as:

$$Gram_k^{\Phi}(i) = \psi\psi^T / C_k H_k W_k, \tag{3.6}$$

where $\psi$ presents the reshaped $\mathbb{A}_k(i)$ with a size of $C_k \times H_k W_k$. As a result, our content loss is described as:

$$\mathcal{L}_c = \Sigma_k \| Gram_k^{\Phi}(i) - Gram_k^{\Phi}(G(b)) \|_1. \tag{3.7}$$

Specifically, we choose the $k = 15$ layer (i.e. relu3_2) after ReLU operations of the pre-trained VGG19 network to generate the feature map of the input image $i$.

**Adversarial loss.** We use an adversarial loss to learn a robust loss function, along with the actual generator network, which better generalizes to measured data than hand-crafted per-pixel losses. Instead of adopting a vanilla GAN [98] training procedure, we rely on variant of the Wasserstein GAN [99] with a gradient penalty to enforce a more robust training process with the U-Net generator in our training pipeline. The resulting adversarial loss can be expressed as:

$$\mathcal{L}_{adv} = \underbrace{\mathop{\mathbb{E}}_{i \sim \mathbb{P}_r}[D(i)] - \mathop{\mathbb{E}}_{\bar{i} \sim \mathbb{P}_g}[D(\bar{i})]}_{critic\ loss} + \underbrace{\lambda_g \mathop{\mathbb{E}}_{\bar{i} \sim \mathbb{P}_{\bar{i}}}[(\|\nabla_{\bar{i}} D(\bar{i})\|_2 - 1)^2]}_{gradient\ penalty}, \tag{3.8}$$

where $\mathbb{P}_r$ and $\mathbb{P}_g$ are distributions of data and model, respectively. Note that $r$ contains here more sharp captured images than corresponding blurry/sharp pairs. Intuitively, the adversarial loss attempts to minimize the structural deviation between a model-generated image $\bar{i} = G(b)$ and a real image $i$, penalizing missing structures, while relaxing the requirements on high color-accuracy and SNR in heavily blurred regions. We will analyze this behavior further in Sec. 3.8.4.

**Overall loss.** We use a weighted combination of both loss functions:

$$\mathcal{L}_{total} = \mathcal{L}_c + \lambda_a \mathcal{L}_{adv}. \tag{3.9}$$

During training, *Generator* $G$ and *Discriminator* $D$ alternate in a way that $G$ gradually refines the latent image to "convince" $D$ the result is a real image free of degradations, while $D$ is trying to distinguish between real and generated samples, including corresponding and non-corresponding real captures, by minimizing the Wasserstein distance.

## 3.6 Datasets

**Data Acquisition** To be successful, the supervised training set component of the proposed architecture requires corresponding sharp ground truth images, and blurry captures using the proposed optical system. Manual acquisition of this dataset in the wild, e.g., changing optics in a sequential fashion per capture, would require complicated robotic systems to ensure identical positions, and captures of various sceneries. Alignment of nearby placed cameras also poses a major hurdle due to the severe aberrations in the prototype, which make alignment in parallax areas very challenging.

To overcome these restrictive capture issues, we have built a *display-capture lab setup* that allows us to efficiently generate a large amount of training data without large human labor. This is realized by capturing images that are sequentially displayed on a high resolution LCD monitor (Asus PA32U), as shown in Figure 3.6. As a benefit of the fact that our PSF is shift-invariant, the proposed lens design does not require training over the full FOV. Instead, we train our network on a narrow field of view, which allows us to overcome prohibitive memory limitations during training with current generation GPU hardware. Moreover, this feature further aids the calibration over our large FOV. During testing we run the network on the CPU to process full-resolution measurements. We use two datasets using a Canon 5D and a Nikon D700 from the Adobe 5k set which contains in total 814 images. To cover the full FOV, we additionally select the first 200 images by name order from the 814

images and capture them by setting the monitor at large FOV. The test set is selected by name order (i.e. first 100 images) from the Canon 40D subset of the Adobe 5k set. All the images are resized to fit with the resolution of the display monitor and converted to Adobe RGB colorspace.

Before starting the image capture procedure we calibrate the setup as follows:

1. We calibrated the tone curve and color reproduction of the LCD monitor using the i1 Pro calibration suite.

2. We calibrated the system uniformity (including both the brightness uniformity of the LCD monitor and imaging vignetting of the capturing camera) by capturing a white calibration chart.

3. We obtained coarse distortion correction parameters of the captured image and the alignment transfer matrix between the captured image and ground-truth image displayed on the monitor by capturing several known checkerboard patterns displayed on the LCD monitor.

**Training Details** For training purposes, we crop both the pre-processed raw and ground-truth images into 512×512 and 1024×1024 patch pairs. These training pairs are randomly flipped and rotated to augment the training process. To preserve color fidelity, we normalize the image to range $[0, 1]$ instead of subtracting the mean and dividing its corresponding standard deviation. We choose the ADAM optimizer with $\beta_1 = 0.5$ and $\beta_2 = 0.999$, which exhibits robustness to the high noise level of our input. At first, the learning rate is initialized as 0.0001 for the first 100 epochs and linearly decayed to 0 over another 150 epochs using 512×512 patch pairs. Then, the learning rate is initialized as 0.00002 for the first 50 epochs and linearly decayed to 0 over another 50 epochs using 1024×1024 patch pairs. The batch size is set to 1 to avoid the crosstalk among samples in the batch. In all of our experiments, we set the


Figure 3.6: Top: Illustration of our display-capture setup for preparing the training data set. Selected displayed and captured image pairs are shown as inserts; Bottom: Results on testing set images captured by our lenses. For each example we show the degraded measurement and reconstruction side-by-side.

loss weights in Eq. 3.9 to be $\lambda_a = 0.1$. During each training iteration, $D$ is updated 5 times while $G$ is updated once.

The proposed network architecture is implemented with PyTorch 0.4, and the training process takes around 80 hours in total on a single Nvidia Tesla V100 GPU. Limited by the GPU memory that currently only allows processing up to around 12M pixels, we are unable to fit in a full resolution (i.e. 6k) image on the GPU. As an alternative solution, we solve the full resolution versions on E5-2687 CPUs which process each 6k image in 6 minutes. In addition, processing a 4k $\times$ 3k image on the GPU takes around 10s. Note, that with the rapidly emerging support of neural network computing, a hybrid memory architecture with efficient caching (e.g. GraphCore's IPU architecture) and quantization [100] may lift this hardware limitation within the coming year.

## 3.7 Prototype

We realize the proposed lens objective, using the same optimization method, for two single element lenses, one with two optical surfaces (on both sides of the same flat carrier), the other with a single optical surface. The field of view and focal length of the lens prototypes are 53° and 43 mm with a real clear aperture size of 23.4 mm, respectively. To fabricate our lenses, we use a CNC machining system that supports 5-axis single point diamond turning (Nanotech 350FG) [63]. The substrate is polymethyl methacrylate (PMMA) with a refractive index of 1.493 at the principle wavelength of 550 nm. We use $200\pi$ phase modulation rather than regular $2\pi$ to wrap the optimized height map since our designed surface type is a deep Fresnel surface. As a result, the final prototype lens has an effective modulation thickness of 120 $\mu$m and a total thickness of 3 mm (10 mm) including the planar substrate. The total clear aperture size of the lens is 23.4 mm with a focal length 43 mm corresponding to an $f$ number of $f/1.8$ in the traditional sense. However, note that the effective

aperture which contributes the sharp intensity peaks has a size of 8 mm yielding an effective $f$-number of $f/5.4$.

We note that the accuracy of the fabrication method is limited by the turning tool which has a rounded tip with 16 $\mu$m radius, prohibiting the reproduction of discontinuities in the profile. The light loss and haze caused by this prototyping constraint accounts for some artifacts we will observe in the experimental result section. We discuss this limitation in depth in the supplementary document.

To validate the proposed approach experimentally, we use a Sony A7 full-frame camera system with 6,000×4,000 pixels with a pixel pitch of 5.96 $\mu$m, resulting in a diagonal FOV of 53°. To collect reference data on real scenes as comparisons, we use an off-the-shelf well-corrected lens (Sony Zeiss 50 mm $f/1.4$ Lens). This compound reference lens has been designed with more than a dozen refractive optical elements to minimize aberrations for a large FOV. To evaluate the proposed approach against alternative single-element designs, we compare our lens against a single plano-convex aspherical lens (Thorlabs AL2550G) with a focal length of 50 mm and a thickness of 6 mm. In contrast to a spherical lens, this aspherical lens (ASP) eliminates severe on-axis aberrations. Note that a (phase-wrapped) diffractive Fresnel lens is equivalent to an ASP at one designated wavelength, ignoring wrapping errors and fabrication errors. Hence, we consider the ASP the state-of-the-art single lens alternative to the proposed design.

## 3.8 Analysis

### 3.8.1 Field of View Analysis

Figure 3.7 shows the spatial distribution of the aberrations and example captures of a checkerboard target across the full sensor. Our design balances the contrast detection probability (CDP) [72] across the full field of view. CDP is a probabilistic measure that allows us to characterize the ability of a higher-level processing block to detect

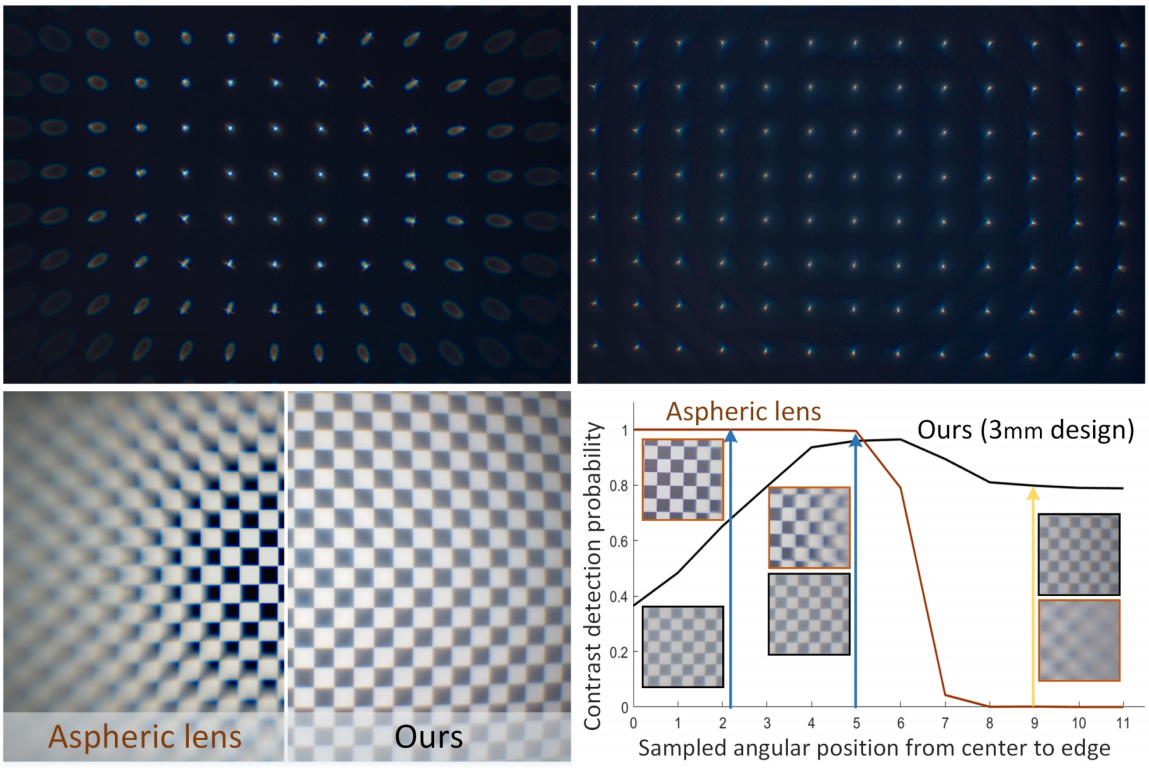


Figure 3.7: PSF behavior comparison (top) and corresponding checkerboard capture comparison (bottom) between an off-the-shelf aspherical lens (ASP) and our prototype lens. Bottom-left shows the side-by-side comparison of the measurements of ASP and ours. Bottom-right shows the derived distributions of contrast detection probability. The confidence interval is set to 95% in both examples, refer to the original reference for details. Here we use a plano-convex aspherical lens (Thorlabs AL2550G) as the comparison.

a given contrast between two reference points after the full imaging chain.

We measure the local CDP of different measurement patches of our lens and that of an aspherical lens, see Figure 3.7. The reference points for this measurement are picked with 100% contrast between local patches with a lateral distance of 3 $\sigma$, with $\sigma$ being the full-width-half-max (FWHM) of the peak mode of our PSF. This allows us to characterize CDP for our dual-mixture PSF without needing to vary the size of measurement patches. For our lens, a significant CDP floor of almost 50% is preserved across the full FOV, ranging from 40% at on-axis angular direction to stay above 80% at the most tilted angle. Since the PSF is not completely spatially invariant, the plot exhibits a maximum around 0.5× half-FOV where the lens focuses best. In contrast,

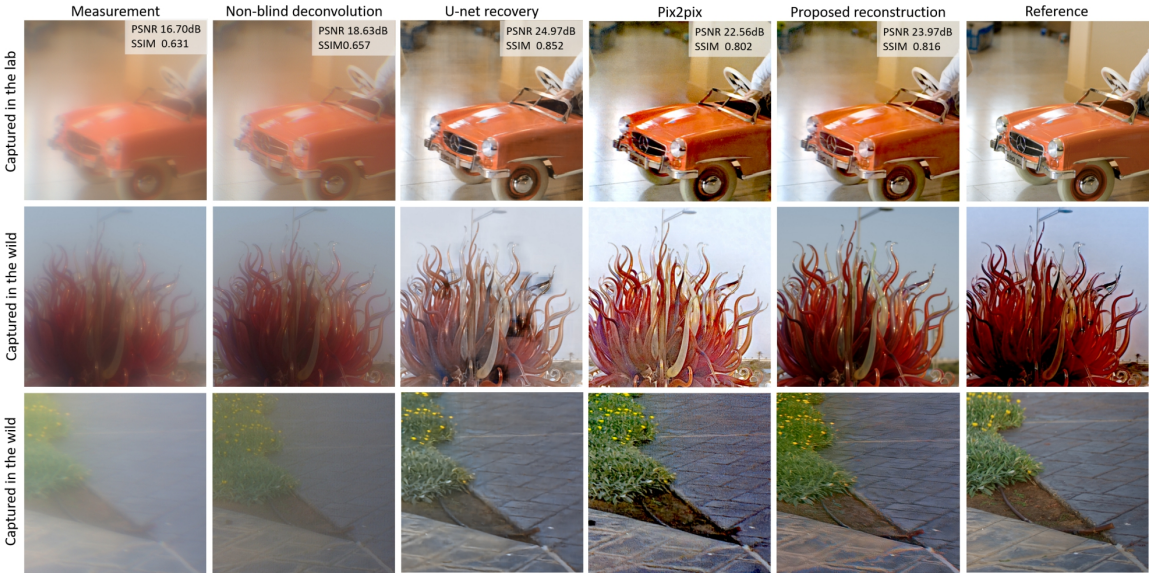


Figure 3.8: Comparison of off-axis image patches recovered using different reconstruction algorithms described in Table 3.1.The first row presents the displayed validation from the test set of our learned reconstruction, while the remaining two rows present the data captured in the wild. Due to the mismatch of spectrum, dynamic range, and depth of field, vanilla supervised learning using a per-pixel loss may show good quantitative results while suffer from severe artefacts on real-world data. For these two examples, we present the image captured using an off-the-shelf compound lens (Sony Zeiss 50 mm $f/1.4$ Lens) as the reference. Full images are shown in the supplementary document.

the CDP of the aspherical lens drops drastically and approaches 0% at view directions larger than 0.5× half-FOV. The measurements agree well with our design goal that the sharp peak of our dual-mixture PSF preserves high-frequency detail and local contrast required for the feature extraction blocks in deep network models.

### 3.8.2 Generalization Analysis

The training data acquired using the proposed lab setup suffers from mismatching spectrum and tone curve, non-uniformity, etc., when compared to measurements in the wild. The most critical differences are the limited dynamic range and fixed depth of field of the monitor. Therefore, vanilla supervised learning using a per-pixel loss (i.e. mean absolute error (MAE) or mean square error (MSE)) overfits to these non-

uniformities, hence achieving high quantitative results on a validation set displayed on the same setup, but suffers from severe artefacts on a real-world test set. The proposed semi-supervised adversarial loss, and the perceptual loss achieve robustness to this "noise" in the training data for the given approach. We validate the impact of these algorithmic components visually in Figure 3.8 (in large off-axis regions), and quantitatively against state-of-the-art recovery methods in Table 3.1.

Existing deconvolution methods recover the latent sharp images to some degree but suffer from severe artefacts across the full FOV, which manifests as noticeable haze and low contrast. The size and scene-dependence of the aberrations of the proposed lens make it extremely challenging for prior-based optimization algorithms to recover fine detail and remove apparent haze.

Table 3.1: Quantitative comparison of image recovery performance of the 10 mm lens for recent deconvolution methods, including non-blind cross-channel deconvolution [43] (Cross), fully supervised U-net recovery, U-net + GAN + $\ell_1$ loss (pix2pix), and our U-net + GAN + perceptual loss. We assess peak signal-to-noise ratio (PSNR), structural similarity index measure (SSIM), and the perceptual loss component from our model. The right-most column shows the ASP lens used in Figure 3.7 and Figure 3.9 fine-tuned for our network. Note that the fully supervised U-net in the third row does *overfit to the lab display-capture setup* and fails to generalize to real capture scenarios.

| | Input | Cross | U-net | pix2pix | Ours | ASP |
|---|---|---|---|---|---|---|
| PSNR | 21.28 | 21.46 | 29.70 | 22.20 | **25.89** | **22.53** |
| SSIM | 0.79 | 0.73 | 0.91 | 0.77 | **0.86** | **0.84** |
| Perceptual Loss | 0.87 | 0.80 | 0.57 | 0.93 | **0.47** | **0.65** |

To validate the proposed method against existing supervised training approaches, and assess the effect of the proposed loss functions, we train a U-net with the same structure as that of our generator on the lab-acquired data. Figure 3.8 and Table 3.1 show that vanilla supervised training overfits to the dataset acquired from the lab setup, which causes it to perform much better on captured data under the same condition (validation dataset) but to fail on real-world captures. In addition, we train pix2pix [101] as an adversarial approach while enforcing an $\ell_1$ loss instead of

perceptual loss. By introducing this adversarial loss, the recovery performs better on real world data but still suffers from non-trivial and visually unpleasant artifacts, e.g. the high intensity sky and low intensity ground in the patches. In other words, pix2pix is not robust enough to resolve the mismatch of dynamic range and depth of field. By introducing a perceptual loss rather than a per pixel loss in the proposed method, our approach outperforms existing baselines for real world experimental captures while preserving local contrast and detail, that fits well with the scope of building consumer level cameras.

### 3.8.3 Fine-tuning for Alternative Lens Designs

To validate the efficacy of the proposed lens design, we fine-tune the described recovery method, using the same network, data and training methodology, with an aspherical lens. Compared to this alternative single-element design, the proposed design offers substantially improved sharpness in off-axis regions while trading off on-axis sharpness, as shown in Figure 3.9. The significant improvement in PSNR across full FOV, see last column of Table 3.1 and more real captures in the supplement, validate that not only the recovery algorithm is responsible for image quality but that our mixture PSF design plays an essential role in proposed computational imaging technique.

### 3.8.4 Hallucination Analysis

The evaluation and understanding of the robustness of deep networks is an active area of research. To analyze if the proposed method hallucinates image content that is not present in the measurements, we visualize the outliers with respect to perceptual and SNR metrics on a held-out validation set with known ground truth. Figure 3.10 plots the histograms of errors of image patches with respect to $\ell_1$, $\ell_2$, and the discussed perceptual loss. We show the outliers of these plots in the same figure. Other than

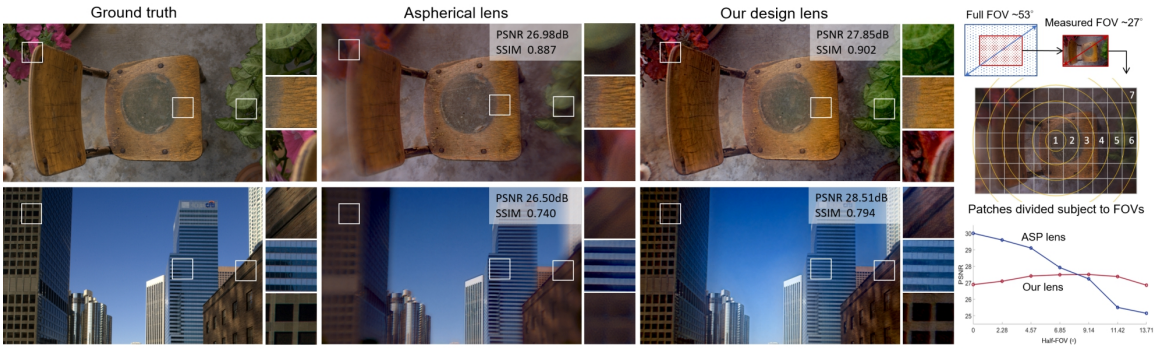


Figure 3.9: Comparison of different regions on images recovered using our learned image recovery algorithm from data captured by an off-the-shelf aspherical lens (ASP) and our prototype lens. Although trading off on-axis sharpness in some sense, ours exhibits much better quality in off-axis regions. The lens parameters and settings are the same as in Figure 3.7. The plots on the right reveal the averaged PSNRs of patches subject to FOV over 100 validation images. Note, in this comparison we investigate only half of the full FOV of our design because the required resolution limits the FOV when using a consumer display monitor. We observe that even within this intermediate range of FOV, the recovered image quality of ASP drops drastically when the investigated half-FOV goes beyond 7°.

suffering from slight blur and color inaccuracy, our recovered results do not hallucinate detail that is not present. Note that the presented image patches are the outliers with the largest error values. As the histogram mode is separated significantly from the presented outliers, we conclude that the proposed reconstruction method is robust and does not hallucinate major detail. Please see the supplemental material for additional outlier visualizations.

## 3.9 Experimental Assessment

**Dual-surface Design.** We first show results for our *dual-surface* thin-plate lens where both surfaces are configured as target depth profiles to be optimized. The resulting optics layout and simulated optical behavior are reviewed in the supplementary document. For this design, to mitigate the possible pressure distortion because of the hard contact turning fabrication, we use a plastic substrate plate that has a thickness of 10 mm as a proof-of-concept. In mass manufacturing, this substrate can

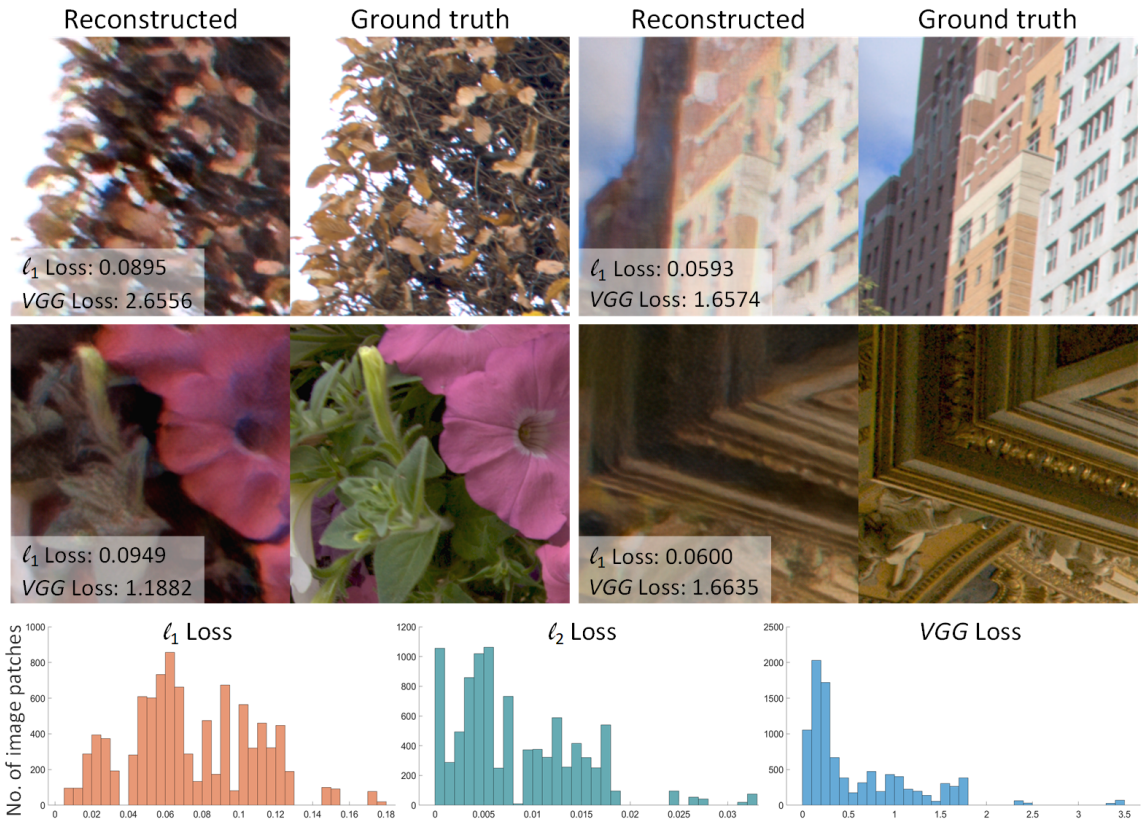


Figure 3.10: Outlier analysis of reconstruction images of our deep network. For each pair we show the recovered patch (left) and its corresponding ground truth patch (right). The plots show the histograms of 9,600 evaluated image patches under three error functions, $\ell_1$ loss, $\ell_2$ loss, and perceptual (VGG) loss.

be reduced to a thinner and more solid structure using glass substrates.

The first two rows of Figure 3.11 show reconstructions of indoor and outdoor scenes under both artificial illumination and natural light. Our method successfully preserves both fine details and color fidelity across full field-of-view. Note that all captures have been obtained using a clear aperture setting, i.e. $f/1.8$, and a full sensor resolution.

**Single-surface Design.** Next, we show results captured with a *single-surface* thin-plate lens with only the rear surface machined. The resulting optical layout and simulated optical behavior are detailed in the supplementary document. For this design, we have reduced the thickness of the substrate plate to 3 mm. The third

Figure 3.11: Experimental results of dual-surface lens design (first two rows) and that of single surface lens design (third row) on real word scenes. For each pair, we show the degraded measurement and the reconstruction result. The exposure time for these images are set 0.8, 125, 0.5, 1.25, 0.5, 0.4 ms with ISO 50. Refer to supplementary document for more real world results.

row of Figure 3.11 shows results for indoor and outdoor scenes captured with this single-surface prototype.

### 3.9.1 Imaging over Large Depth Ranges and in Low Light

Figure 3.12 shows reconstruction results for scenes with large depth ranges and in low-light scenarios. Although we only train the proposed method with screen captures at a fixed distance, the proposed method preserves the depth-dependent defocus, i.e., bokeh, for scenes with large depth ranges. Careful readers notice that for high-intensity regions, as in the sky, our reconstruction does not recover detail. As outlined in Section 1, this is because the training data does not contain high-dynamic range

captures for our low-dynamic range LCD monitor setup.


Figure 3.12: Experimental results of large DOF (top row) and low-light imaging (bottom row). The exposure time and ISO for the top two examples are set (3.125, 1) ms and ISO 50, while that for the bottom two examples are set (10 ms with ISO 500) and (20 ms with ISO 25,600).

In contrast to alternative flat optical designs with wide FOV, such as pinholes with theoretically unlimited FOV, the proposed lens design allows for low-light captures due to its $f$-number of $f/5.4$. We demonstrate low-light and short-exposure imaging scenarios in the second two rows of Figure 3.12, where we compare our design against a pinhole (0.8 mm) that suppresses most aberrations over a wide FOV at the cost of very limited light throughput. The pinhole measurements are low-signal and hence corrupted with severe noise that results in a poor reconstruction – even with state-of-the-art low-light denoising methods [36]. Additional comparisons at different exposure levels can be found in the supplement.

## 3.10 Discussion and Conclusion

We have demonstrated that it is viable to realize *high-quality, large field-of-view* imaging with only a *single thin-plate* lens element. We achieve this by designing deep Fresnel surface optics for a learned image reconstruction algorithm.

Specifically, we introduce a compact thin-plate lens design with a *dual-mixture* PSF distribution across the full FOV. Although the PSF has an extremely large spot

size of $\geq 900$ pixels in diameter, it preserves local contrast uniformly across the sensor plane.

To recover images from such degraded measurements, we learn a deep generative model that maps captured blurry images to clean reconstructed images. To this end, we propose an automated capture method to acquire aligned training data. We tackle the mismatch between lab-captured and natural images in the wild – prohibiting vanilla supervised learning to perform well on real world scenes – by introducing a combination of adversarial and perceptual loss components. Together, the proposed network architecture, training methodology, and data acquisition, allow us to achieve image quality that makes a significant step towards the quality of commercial compound lens systems with just a single free-form lens. We have validated the proposed approach experimentally on a wide variety of challenging outdoor and indoor scenes.

While the proposed approach could enable high-quality imagery with thin and inexpensive optics in the future, on today's consumer graphics hardware, the described reconstruction method is memory-limited for models at full 24.3 Megapixel sensor resolution. Therefore, we run the post-processing on the CPU which results in low throughput on the order of minutes per image – far from that of modern image processing pipelines. The upcoming graphics hardware generation will likely overcome this memory limitation. In the meantime, a combination of cloud processing and low resolution or tile-based previews could make the method practical. The lab data acquisition is currently restricted by the dynamic range of consumer displays, which we hope to overcome in the future with alternative high-dynamic range display approaches.

Although our thin-plate lens design significantly reduces the form factor compared to complex optical systems, we validate the concept with a focal power and an aperture size comparable to existing DSLR camera lenses. To achieve the envisioned camera device form factors, a reduction in both size of the optical lens system

and focal length are necessary. Miniature lens systems with short back focal length (e.g. $\leq$ 5 mm) are now possible by introducing metasurfaces or injection molding techniques to fabricate the optics, which provide feature sizes at the order of the wavelength of light and hence can diffract light at steeper angles allowing for ultra-short focal lengths.

While we designed a single-element lens in this work, dual-refractive lenses or hybrid refractive-diffractive optical systems might be interesting directions for future research. Moreover, simple optics for sensor arrays, such as the PiCam [40], could be revisited with the proposed PSF design. Although this work focuses on computational photography applications, we envision a wide range of applications across computer vision, robotics, sensing and human-computer interaction, where large field-of-view imaging with simple optics and domains-specific post-processing could enable unprecedented device form factors.

# Chapter 4

# End-to-End Encoding Through Optimizing PSF: Super-resolution SPAD Camera

In the previous chapter, we have realized the joint design of optics and image reconstruction network. It is a big step to improve the imaging capability for large FOV imaging, which cannot be realized before with a single element. However, we still cannot directly optimize the optics together with post-processing. In this chapter, we take one more step that optimizing the PSF together with the post-processing, bring the imaging system design into an end-to-end fashion.

SPADs have recently received a a lot of attention in imaging and vision applications due to their excellent performance in low-light conditions, as well as their ultra-high temporal resolution. Unfortunately, like many evolving sensor technologies, image sensors built around SPAD technology currently suffer from a low pixel count.

In this work, we investigate a simple, low-cost, and compact optical coding camera design that supports high-resolution image reconstructions from raw measurements with low pixel counts. We demonstrate this approach for regular intensity imaging, depth imaging, as well transient imaging.

We adopt the differentiable diffractive optics model described in Section 2.2.2 Our method uses an end-to-end framework to optimize simultaneously the optical design and a reconstruction network for obtaining super-resolved images from raw measurements. The optical design space is that of an engineered point spread function (implemented with diffractive optics), which can be considered an optimized anti-aliasing

filter to preserve as much high-resolution information as possible despite imaging with a low pixel count, low fill-factor SPAD array. We further investigate a deep network for reconstruction. The effectiveness of this joint design and reconstruction approach is demonstrated for a range of different applications, including high-speed imaging and time of flight depth imaging, as well as transient imaging. While our work specifically focuses on low-resolution SPAD sensors, similar approaches should prove effective for other emerging image sensor technologies with low pixel counts and low fill-factors.

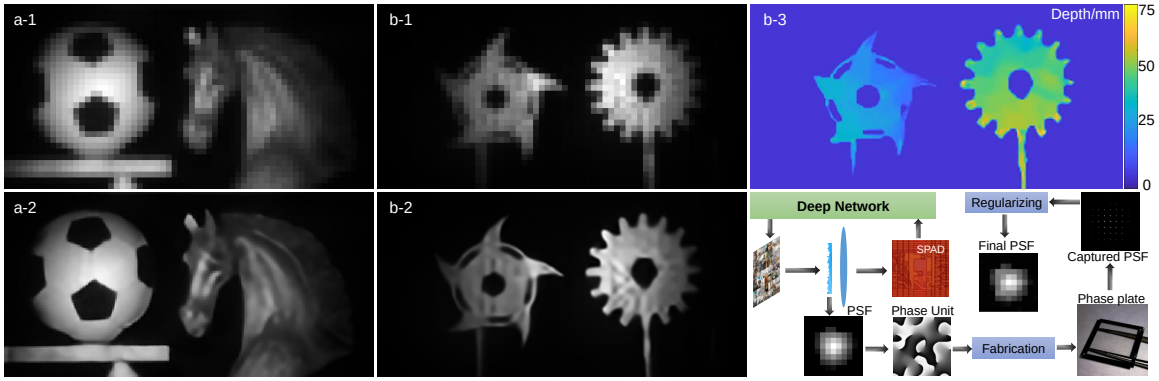


Figure 4.1: Overview of our optically coded computational SR SPAD camera. We computationally design phase plates that can suppress aliasing while preserving as much information as possible for SR image reconstruction (right bottom). Fabricated using photolithography technique, this optimized phase plate produces the target PSF at the image plane. In this figure, we demonstrate two representative applications of our optically coded SR SPAD camera: regular intensity imaging, as well as depth estimation, where we obtain high-quality super-resolved (4×) images (a-2) from raw data (a-1) modulated by our phase mask, and super-resolved (4×) intensity (b-2) and depth images (b-3) from the noisy raw data (b-1).

## 4.1 Introduction

Arrays of Single Photon Avalanche Diode (SPAD) have recently emerged as an alternative hardware solution to photomultiplier tubes (PMT) and streak cameras [102, 103]. Features such as single-photon light sensitivity and sub-nanosecond time resolution make this technology promising for many photon-starved applications like time-

of-flight [104], transient imaging [105, 106], fluorescence lifetime imaging [107, 108] and positron emission tomography [109].

Unfortunately, image sensors built upon SPAD technologies still suffer from low spatial resolution (e.g., $64 \times 32$) and low fill-factor, i.e., the fact that the light-sensitive area of a pixel is only a small fraction of the pixel's total area (e.g., 3.14% in the MPD-SPC3 SPAD camera used in our experiments). Although recent research prototypes of SPAD arrays have substantially higher pixel counts (e.g. up to $512 \times 512$ pixels [110]), they still fall short of the resolution of conventional image sensors. Therefore our method is relevant to the latest generation of prototype SPAD image sensors as well as all commercially available SPAD arrays. Both the limited pixel count (e.g., [104, 12]) and the limited fill-factor and its associated loss in light efficiency [111, 112] have been targeted by recent research. However, no definitive solution is available at this time.

To date, computational imaging has achieved tremendous success in the fields of spatial resolution enhancement [113, 12] and defocus deblurring SR [114]. Via point spread function (PSF) engineering [115, 116], researchers have succeeded in localizing microscopic point emitters in a 3D volume by inserting either a spatial light modulator (SLM) or a physical phase plate.

Although optimizing the parameters of DOEs for a computational camera has been studied intensively, state-of-the-art PSF engineering methods still, for the most part, do not consider the optical design together with the sensor performance and the reconstruction algorithm in a full end-to-end fashion. A notable exception is a recent work by Sitzmann et al. [14], that employed an end-to-end optimization that jointly considers optics and image processing to extract optimal PSFs for the purposes of superresolution and depth of field extension. Although this work takes a significant step towards full end-to-end design of cameras, the reconstruction method used is quite simple and with only fixed blocks, for example, the Wiener deconvolution. In

our work, we extend this concept by *jointly* optimizing both the PSF design for the sampling model and the reconstruction algorithm, particularly in the context of a deep neural network.

Putting these pieces together, we aim to overcome the essential the spatial resolution limit of SPAD sensors by developing an optically encoded SR SPAD camera with only a single-shot capture procedure. This is achieved by a combination of an optical system that encodes the incident light and a deep neural network that faithfully decodes the high-resolution image. The optical encoding is interpreted as an engineered PSF, acting as an anti-aliasing filter that helps preserve as much information as possible, given the specific sampling the pattern of SPAD sensors. We demonstrate significant improvements gained by our prototype when imaging natural scenes. While our method can, in principle, be applied in any imaging system that employs SPAD array sensors, we focus in particular on three applications: regular intensity imaging (including high speed imaging), depth imaging, and transient (i.e. light-in-flight) imaging.

Our main technical contributions are as follows:

- We exploit an end-to-end design paradigm for computational superresolution camera systems, incorporating both PSF design, imaging model, and deep network reconstruction. The system finds optimized compromises between sharpness and anti-aliasing for a given pixel fill-factor.

- We develop a novel single-shot optically coded SPAD camera that achieves an aggressive spatial resolution enhancement of $4\times$. By simply applying an ultra-thin phase plate that can be easily fabricated and assembled, we achieve an almost zero budget enhancement of hardware configuration.

- We build a prototype with a general phase plate being easily assembled in front of a regular lens. We validate our claims of resolving high-resolution

images through simulations and real experiments in normal imaging, high-speed imaging, and time-of-flight (TOF)/transient imaging.

## 4.2 Related work

Computational imaging has been applied in both low-level vision tasks like artifact removal [13], and higher-level imaging applications like depth estimation [117, 4]. Particularly, a large amount of work has studied image enhancement using the end-to-end method for applications such as haze removal [118], motion deblur [119], and time-of-flight imaging [120]. In the following, we focus on a few more narrow categories of research that are most relevant to our work.

### 4.2.1 Image SR

For target applications like high speed imaging, fluorescent lifetime imaging, time-of-flight depth or transient imaging, achieving an aggressive resolution enhancement is highly desirable. A large body of work is based on learning the mapping from low-resolution (LR) to low-resolution (HR) images, using techniques such as dictionary learning [121, 122], local linear regression [123, 124], random forests [125], and CNNs [126, 127, 128]. Alternatively, one can employ a sparse coding based network to fully explore the sparsity of natural images [129].

Ongoing research efforts have attempted to improve the SR quality using deeper networks [130, 131]. Alternative work includes a Laplacian Pyramid SR network [132] and an enhanced deep SR network [133] that removes unnecessary modules in conventional residual networks [134]. More recently, Haris et al. proposed a deep back-projection network [135], exploiting iterative up and down sampling layers and providing an error feedback mechanism for projection errors at each stage.

The mentioned approaches take a traditional image processing approach, whereby the imaging hardware is given and not part of the design decision. Computational

imaging approaches, where the imaging hardware and the reconstruction method are *co-designed*, promise improved system performance. This is the approach we take in this work, specifically with the design of an optimal sampling strategy for low pixel count, small fill factor SPAD image sensors.

### 4.2.2 PSF Engineering for Computational Imaging

The optics and computational imaging communities have widely investigated the deliberate design of (non-Dirac) PSFs with favorable properties for specific applications. One of the earliest approaches was *wavefront coding*, a method to make the PSF depth-invariant in an attempt to EDOF [1, 136]. Recently, the utility of PSF engineering was expanded to 3D to realize a 3D super-resolution effect [137]. Encoding the aperture of the optical system not only enables recovery of depth information with great fidelity but also generates a high resolution image image [117, 138]. Furthermore, coded aperture techniques have been intensively incorporated into compressive sensing [139, 140, 141].

Instead of inserting a (usually binary) coded aperture, we investigate the link between the aperture and the image plane in the domain of diffractive optics. By introducing a phase modulation diffractive optical element into the aperture, one has greater flexibility to design the desired PSF in the image plane. There have been a wide range of optimization-based algorithms capable of generating desirable phase or amplitude distributions in both the spatial and the spectral domain. To this end, iterative methods based on GS search, simulated annealing, or direct binary search, have been applied to design both monochromatic and broadband DOEs [142, 143].

Another related avenue of investigation is the design of DOEs to serve as replacements for refractive lenses in imaging systems. Peng et al.'s work on achromatic DOE lenses [51] started a sequence of DOE design works with similar methodology [43, 144, 145]. Instead of automated end-to-end design, the PSF design and re-

construction method are developed separately with a human in the loop. Some recent works [146, 147] have explored the role of anti-aliasing filters in image super-resolution; however, they use analytical filters (Butterworth and Gaussian, respectively), instead of end-to-end learned ones.

### 4.2.3 Imaging with SPAD Sensors

Time-correlated single photon counting (TCSPC) [148] is a common technique for pico-second rate recording of photon events using SPAD arrays. It has been widely applied for example in fluorescence lifetime imaging [108, 149]. By repeatedly measuring the time duration between a laser pulse and the corresponding transient photon arrival, one can achieve typically sub-nanosecond resolution. Starting with first photon imaging [150], several approaches have been proposed to abstract the correct temporal information like temporal deconvolution [12], pile-up compensation [151, 152] and non-line-of-sight imaging [153, 154].

To overcome the limitations of low fill-factor and low spatial resolution, researchers have used 2D translation setups to shift a 2D SPAD array with a fixed lens [104], or used a galvo mirror setup to scan a 1D line SPAD camera [106, 155]. An alternative approach is the use of DMD-based focal plane spatial modulation to enable a compressive sensing design with SPAD arrays [12]. This method requires high precision mechanics and additional imaging optics. Other works have focused primarily on improving the fill-factor of SPAD arrays [111, 112].

Although state-of-the-art methods have yielded a reasonable spatial resolution, they are significantly complicating the camera design, and/or require multi-shot image acquisitions, which makes it impossible to image non-repeatable phenomena. We seek a computational super-resolution imaging solution that can maintain all the advantages of SPAD sensors including the snapshot capability, i.e. super resolution reconstruction from a single image capture.

### 4.2.4 End-to-End Computational Cameras

Motivated by recent advances in hardware as well as optimization methods, researchers have started to investigate joint optimization over optics like binary masks [156] for compressive sensing and even sensor structure like a color filter array [157]. More recently, an end-to-end optimization [14] over more complicated phase modulation elements was reported. In work parallel to ours, full end-to-end pipelines have been shown recently for the design of depth-encoding PSFs in shape-from-defocus applications [158, 16].

In addition to conventional imaging applications, diffractive optical elements can also be used as convolutional layers in neural networks [159] to speed up the process. Instead, we are inspired to simulate our imaging model for SPAD sensor using convolutional layer. Taking the convolutional layer into a physical world, we are able to realize the difficult super-resolution task for low fill-factor and low resolution SPAD sensor by incorporating both optics and deep reconstruction networks.

## 4.3 End-to-end Diffractive Optics Design and Image Reconstruction

We aim to realize super-resolution imaging over a SPAD sensor that suffers from both low resolution and low fill-factor. These two problems will result in significant spatial aliasing and the associated reconstruction artifacts [160]. To address this issue, we introduce an OLPF into the optical system of the camera. The OLPF acts as an *anti-aliasing filter*, that is specially designed to suppress aliasing while preserving as much information as possible for SR image reconstruction.

In our framework, this filter and the matching reconstruction network are *jointly* learned in an end-to-end sense, as illustrated in Figure 6.2. Specifically, we first synthesize the low resolution input using a convolutional layer $conv(11, 1)^1$, representing the PSF and the sensor sampling model, followed by a feature extraction step to

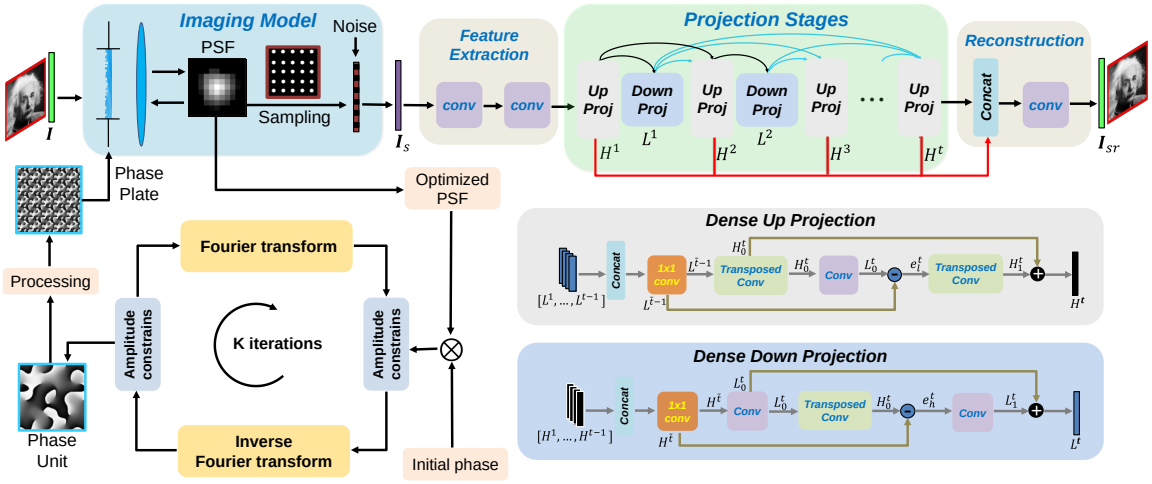


Figure 4.2: Framework of end-to-end optics design and reconstruction for our super-resolution SPAD camera. The anti-aliasing filter (PSF) for the low fill-factor SPAD array is learned using our design paradigm. In each forward pass, the synthetic PSF is convolved with a batch of images, and Poisson noise is added to account for sensor's counting noise after the interval sampling process. After obtaining the optimized PSF, we apply a GS based phase retrieval algorithm to derive the phase mask. The reconstruction network is composed of three main parts: initial feature extraction, back projection stages, and reconstruction step. The back-projection stage (right bottom), alternating between reconstruction of $H^t$ and $L^t$, consists of $T$ up projection stages and $T-1$ down projection stages. Each unit is connected with the outputs of all previous units.

generate LR feature maps. Then, at the projection stages a mapping between the LR feature maps and the HR feature maps is built. Finally, a reconstruction step is added to convert the HR feature maps into high resolution images.

After training, we extract the optimal PSF from the weights of $conv(11,1)$ and then apply a GS-based phase retrieval algorithm to derive the phase mask (see left-bottom of Figure 6.2), which acts as an optical coder installed at the front focal plane of a regular lens to generate the optimal PSF for later implementations. In order to account for differences between the design and the fabrication of the phase mask, the real-world PSF of the mask can be calibrated, and the reconstruction network can be fine-tuned through re-training.

[1]For convenience, we denote a convolutional layer as $conv(f,n)[.]$ and a transposed convolutional layer as $conv^T(f,n)[.]$ where $f$ is the filter size and $n$ is the number of filters.

In the following, we first detail the image formation model, incorporating the anti-aliasing filter applied to the sampling model of SPAD array and the phase mask optimization to generate the learned PSF combined with a regular imaging lens. Next, we present the deep neural network reconstruction and the time profile sharpening strategy.

### 4.3.1 Image Formation

#### Anti-aliasing filtering and image sampling

As mentioned, the fill factor of most current SPAD imaging sensors is very low, that is, the light sensitive area of the pixel is much smaller than the total area occupied by the pixel structure. For example, the SPAD array used in our experiments (MPD-SPC3) has a pixel pitch of 150 $\mu$m horizontally and vertically, however the active area is only 30 $\mu$m in each dimension. The physical low pixel count and small fill-factor severely degrade the image quality, creating the desire for super-resolved image reconstruction. To avoid aliasing, the image signal should be pre-filtered with a low pass filter of the appropriate cut-off frequency, followed by a down-sampling process [160]. Again, the goal is to trade-off sharpness and aliasing, so as to find a good compromise that preserves most details of interest.

Due to the low resolution of the sensor array, we can reasonably neglect off-axis aberrations like coma. Image formation becomes a shift-invariant convolution of a latent image with a kernel. To this end, we jointly learn the optimal anti-aliasing filter (e.g. the convolved kernel) and the reconstruction network to eventually preserve the finest details of natural images so as to realize a SR enhancement. The quantitative evaluation of applying this desired OLPF is detailed in Section **??**.

At the position $(x, y)$ on the sensor, the detected signal $I_s(x, y)$ is expressed as:

$$I_s(x, y) = \mathcal{P}(\mathcal{S}(\boldsymbol{p}_\lambda * \boldsymbol{I})), \tag{4.1}$$

where $\mathcal{S}$ is a 2D sampling operator corresponding to the physical structure of SPAD sensor, $\boldsymbol{I}$ is the latent image formed on the sensor, $\boldsymbol{p}_\lambda$ is the kernel (or PSF) realized by the optical system, and $\mathcal{P}$ represents a generator of the Poisson noise, which is the appropriate noise model for low light scenarios that are typical for SPAD imaging.

### Learning optimal PSF using end-to-end design

To obtain the optimal PSF $\boldsymbol{p}_{\lambda\text{opt}}$ using our end-to-end framework, we model our PSF as well as the low-resolution sampling process of the SPAD array as a convolutional layer $conv(11, 1)$. In each forward pass, the synthetic PSF (convolutional layer) is convolved with a batch of images, and Poisson noise is added to account for photon shot noise after the interval sampling process. In other words, we represent both the PSF and the sampling process as layers in our neural network during training, and then physically realize the learned result as a custom DOE for our SPAD camera (see Section 4.3.3).

To determine the size of the kernel, we take a large kernel $21 \times 21$ at the beginning and then we found only an $11 \times 11$ region of the filter had non-zero values. Therefore, we take $11 \times 11$ as the kernel size of the PSF whose physical dimension is $412.5 \times 412.5\mu m^2$.

## 4.3.2 Image Reconstruction

Image reconstruction is the final stage for applications like regular intensity imaging or high speed imaging, and the second last stage for applications like depth and transient imaging. For our camera the reconstruction is formulated as an optimization problem of a data fitting term with an additional regularization term:

$$\min_{\boldsymbol{I}} \frac{1}{2}\|\mathcal{S}(\boldsymbol{p}_{\lambda\text{opt}} * \boldsymbol{I}) - \boldsymbol{I}_s\|_2^2 + \beta\|\Phi(\boldsymbol{I})\|_1, \tag{4.2}$$

where $\Phi(\cdot)$ denotes the transform coefficients of $\boldsymbol{I}$ with respect to some transform $\Phi$ that can be either linear or optimized non-linear. Sparsity in the transform space $\Phi(\boldsymbol{I})$ is encouraged by the $\ell_1$ norm with $\beta$ being a regularization parameter.

Usually, natural images are non-stationary in classic domains like DCT, gradients, and wavelets, which may result in an ill-posed problem under such an imaging model. Although an optimized PSF model can preserve a large amount of spatial information, conventional optimization-based methods fail to faithfully reconstruct good quality results when the sampling ratio is very low (e.g. in our case with a sampling ratio only 3.14%). To this end, a trainable architecture for SR with powerful learning ability for features meets our strict requirements as our learned PSF itself encodes features. We choose the state-of-the-art method— dense deep back-projection networks (D-DBPN) [135] as our reconstruction network, as shown in Figure 6.2. The D-DBPN framework introduces an iterative error correcting feedback mechanism to characterize the features in previous layers. More importantly, it addresses the mutual dependency by taking the back-projection from HR domain to LR domain.

## Framework architecture

As shown in Figure 6.2, the end-to-end framework to obtain the optimal filter and reconstruction network can be divided into four parts:

**a. Imaging model.** As we have already discussed in Sec. 4.3.1, we take the physical imaging model as the first part of our end-to-end framework. The joint framework is used to learn the optimal anti-aliasing filter. After fabricating the filter we then refine the learning process of the reconstruction network with additional training to account for fabrication errors. For more details, please refer to Section 4.5.

**b. Initial feature extraction.** The initial feature maps $L^0$ are constructed using a $conv(3, n_0)$ layer to extract features and a $conv(1, n_R)$ layer to pool the features

and reduce the dimension from $n_0$ to $n_R$. In the experiments, $n_0$ is set as 256 and $n_R$, which is the number of filters used in each projection unit, is set as 64.

**c. Back-projection.** As illustrated in Figure 6.2, at $t^{th}$ stage ($T = 7$ stages in total), the LR feature maps $[L^1, L^2, \cdots, L^{t-1}]$ and HR feature maps $[H^1, H^2, \cdots, H^t]$ are concatenated to be used as input for up- and down-projection units respectively. In each projection unit, we use a $conv(1, n_R)$ to merge all previous outputs from each unit after the shown concatenation process.

The up-projection is defined as follows:

$$\begin{aligned}
&\text{scale up} && H_0^t = conv^T(f_p, n_R)[L^{t-1}] \\
&\text{scale down} && L_0^t = conv(f_p, n_R)[H_0^t] \\
&\text{residual:} && e_t^l = L_0^t - L^{t-1} \\
&\text{scale residual up:} && H_1^t = conv^T(f_p, n_R)[e_t^l] \\
&\text{output feature map:} && H^t = H_0^t + H_1^t
\end{aligned} \tag{4.3}$$

The down-projection is defined as follows:

$$\begin{aligned}
&\text{scale down} && L_0^t = conv(f_p, n_R)[H^t] \\
&\text{scale up} && H_0^t = conv^T(f_p, n_R)[L_0^t] \\
&\text{residual:} && e_t^h = H_0^t - H^t \\
&\text{scale residual down:} && L_1^t = conv(f_p, n_R)[e_H^l] \\
&\text{output feature map:} && L^t = L_0^t + L_1^t
\end{aligned} \tag{4.4}$$

**d. Reconstruction.** Finally, we take the concatenated HR feature maps $[H^1, H^2, \cdots, H^t]$ as input and use a $conv(3, 1)$ layer to reconstruct the target HR image.

## Training details

To train the network, we use the MSE loss function. In the stated framework, we use an 8×8 convolutional layer with a stride of four and a padding of two. All convolutional and transposed convolutional layers are followed by a parametric rectified linear unit. We trained our network using the high resolution images from the DIV2K dataset, using a batch size of 64. For convenience, the LR image resolution was 32×32 (half the size of our SPAD array), and the HR image size was 128×128. We take a convolution layer $conv(11, 1)$ as our PSF following the sampling model of the SPAD sensor to simulate the LR images from HR images. We use ADAM as the optimizer with momentum set to 0.9 and weight decay set to $10^{-4}$. The learning rate is initialized to $10^{-4}$ for all layers and decayed by a factor of 10 for every half of total epochs. All experiments were conducted using Pytorch on a single NVIDIA TITAN Xp GPU. For learning the optimal PSF, we trained the whole framework with 50 epochs taking around 40 hours. After calibrating the PSF generated by the fabricated phase mask, we take the weights of the network trained above as initialization and continue to train the reconstruction network with 11 epochs taking around 8 hours.

### 4.3.3 Phase Mask Generation

After obtaining the optimal PSF with our framework, we establish the relationship between the PSF and the phase mask. We first analyze the propagation of light from the phase mask to the image plane, and then present the details of phase mask design.

## Optical model

As shown in Figure 4.3, the mask is placed at the front focal plane of the lens, and acts as the pupil of whole system. For modeling the light propagation, we apply scalar diffraction theory [86] to approximate the paraxial incident wave. The phase of a complex-valued incident wave is delayed by a phase profile $\phi(x', y')$ proportionally

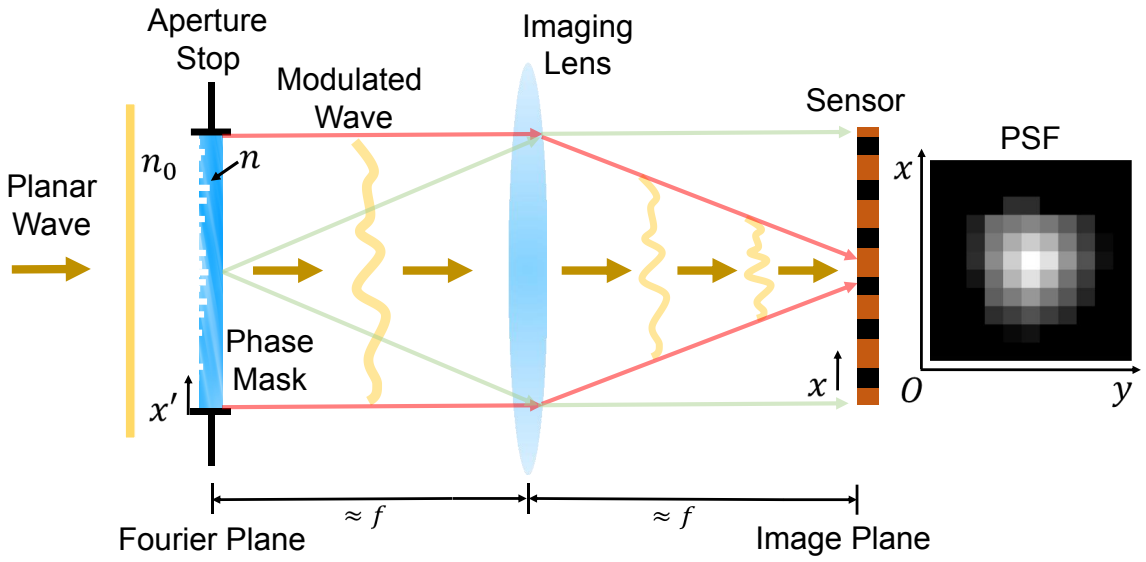


Figure 4.3: Illustration of light propagation and desired PSF. The phase mask (i.e. DOE) is set at the equivalent Fourier plane of imaging lens to modulate the incident light and produces the desired PSF on the sensor.

to the height map of a diffractive optical element $h(x', y')$:

$$\phi(x', y') = \Delta n \frac{2\pi}{\lambda} h(x', y'), \tag{4.5}$$

where $\lambda$ is the wavelength, $(x', y')$ is the location on the phase mask plane , and $\Delta n = n - n_0$ represents the refractive index difference between air $(n_0)$ and the substrate material $(n)$. Placed at the front focal plane of a lens together with our customized limited stop, the phase mask acts as the complex pupil function.

The incident wave field $U_\lambda(x', y', z = 0_-) = A(x', y')\phi_d(x', y')$ is modulated by the phase mask, shown as:

$$U_\lambda(x', y', z = 0_+) = U_\lambda(x', y', z = 0_-) \cdot e^{\mathbf{i}\phi(x', y')}, \tag{4.6}$$

where we use the notation $z = 0_-$ and $z = 0_+$ to denote positions just before and just after the mask, respectively.

Using the Fresnel approximation, the light propagates through a lens with a focal

length $f$ to the image plane is then formulated as:

$$\begin{aligned} U_\lambda(x,y) =& \frac{e^{\mathbf{i}kf}}{\mathbf{i}\lambda f}\int\int_{\sum} U_\lambda(x',y',z=f)e^{-\frac{\mathbf{i}k}{2f}(x'^2+y'^2)} \\ & e^{\frac{\mathbf{i}k}{2f}\left[(x-x')^2+(y-y')^2\right]}dx'dy' \\ =& \int\int_{\sum}\phi(x',y')e^{-i2\pi\frac{x'x+y'y'}{\lambda f}}dx'dy', \end{aligned} \tag{4.7}$$

where $k = 2\pi/\lambda$ is the wave number, $(x, y)$ is the location on the image plane, and $e^{-\frac{\mathbf{i}k}{2f}(x'^2+y'^2)}$ represents the optical transfer function of the lens. Note that Equation (4.7) represents essentially a Fourier transform (FT).

For an imaging system, the diffractive PSF on the image plane is eventually obtained as:

$$p_\lambda(x,y) \propto \|(\mathcal{F}\{\phi(x',y')\}\|^2. \tag{4.8}$$

## Phase retrieval

After deriving the relationship between PSF and the phase mask, we can design a physical height profile $h(x', y')$ on a substrate of refractive index $n$ to implement an image-plane PSF $\boldsymbol{p}_\lambda$ using the GS [161] phase retrieval algorithm based on Equation (4.5).

The core of the phase retrieval is shown on the bottom left of Figure 6.2. In the beginning, a random phase distribution serves as the initial estimate subject to the amplitude of the PSF. Then, using the initial phase and the amplitude constraint (between 0 and 1) of learned PSF, we apply an inverse Fourier transform on this synthesized complex field function. The resulting phase part of the discrete complex field is preserved while the amplitude part is discarded. In the next round, this preserved phase is plugged into the forward propagation procedure of applying a Fourier transform to update the amplitude estimate of the complex field on the image plane. Eventually, the process is repeated with a finite number times to converge to

an optimal phase profile. For more details, please refer to the work by Morgan et al. [162]. Since we optimize the phase plate for only one wavelength (that of our picosecond laser), we are guaranteed to obtain a phase plate that can generate the optimal PSF we desire. As shown in Figure 4.4, the correlation coefficient between the PSF generated by phase plate and the learned PSF is 0.9996, and the root mean square error (RMSE) between them is 0.0061. This all means the optimal PSF is accurately realized by the phase mask.

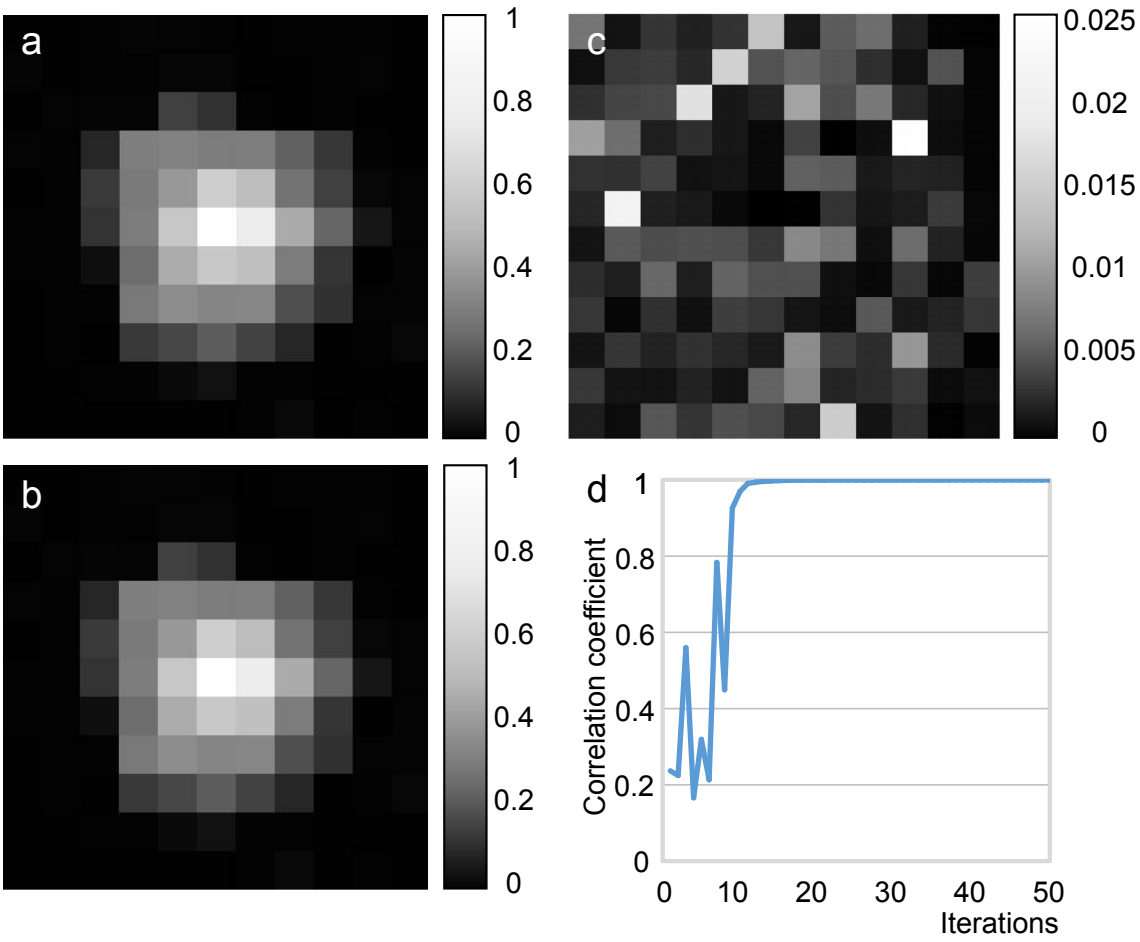


Figure 4.4: Efficiency illustration of GS phase retrieval method for our design. a) Learned PSF; b) Simulated PSF using the phase profile optimized by GS method; c) The absolute error between a) and b), and the RMSE is 0.0061; d) The correlation coefficient of the learned PSF and the PSF generated by phase plate, and finally it converges to 0.9996.

## Phase mask tiling

As shown in Figure 5.7, a subpixel on the learned PSF has a size of $l_p = 37.5$ $\mu$m. Accordingly, the size of phase profile obtained using Equation 4.7 is $l_u = \lambda f / l_p = 0.8733$ mm, which would make for a very small, square aperture. To design optical systems with larger apertures, one could over-parameterize the design space to optimize the phase profile over a defined larger aperture. This would require a re-design of the pattern for each aperture size, and rule out the use of the aperture stop diaphragm in the main camera lens.

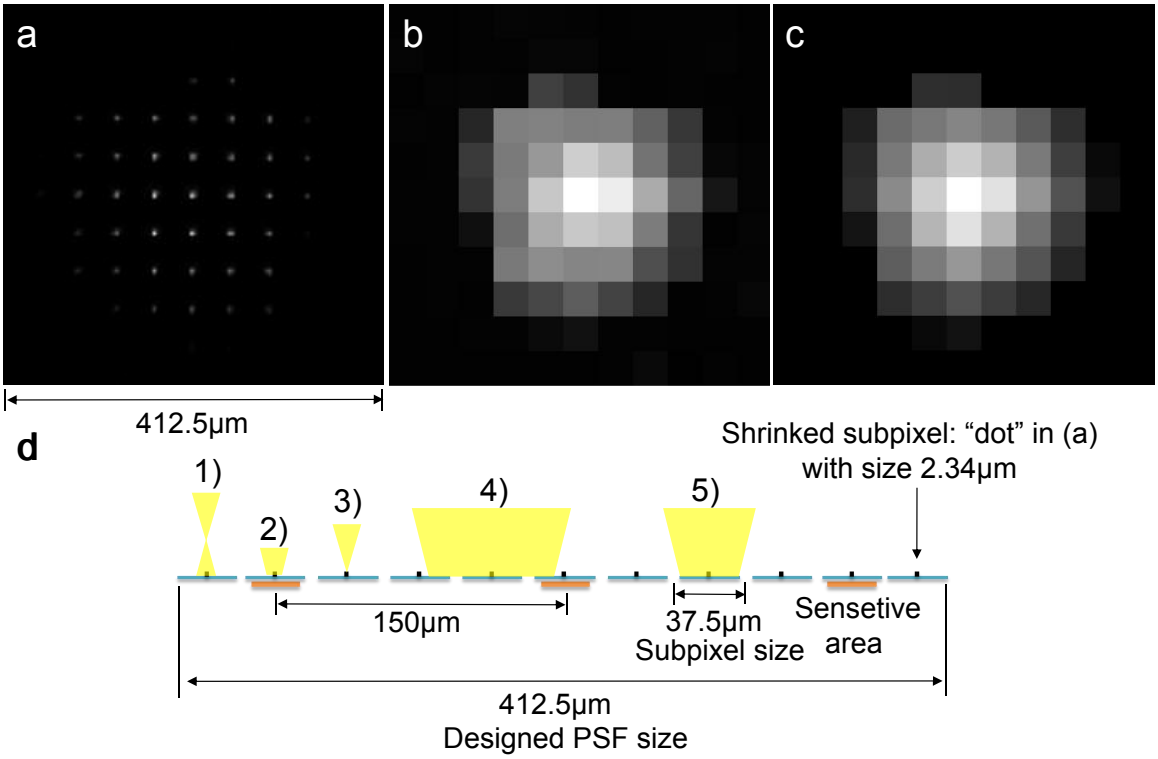


Figure 4.5: Calibrating the PSF generated by our fabricated phase plate. a) Captured PSF; b) Synthetic learned PSF; c) Effective PSF as a result of combining PSF a) with the SPAD pixel sampling pattern; d) Illustrating the effect of focus on the dot pattern from a) – see text for details.

A simple alternative that overcomes these issues, is to side-by-side replicate the small optimized phase pattern described above in order to tile the aperture. In our prototype, we tile a square area of edge length $L = 14$ mm, which defines a maximum

aperture that can be further stopped down using the lens diaphragm. The tiling has the effect of creating a discrete dot pattern instead of a continuous PSF in the image plane. At a size of $l_p l_u / L = 2.34$ $\mu$m, the individual dots are significantly smaller than a sub-pixel and their center-to-center spacing is exactly the sub-pixel pitch, which also matches the edge length of the light sensitive area of a SPAD pixel. Therefore, as the SPAD sensor integrates spatially over the light sensitive area, it integrates over exactly one of the dots in the dot pattern, which is equivalent to implementing the continuous version of the PSF designed above.

As an added benefit, the dot pattern simplifies the alignment process in the assembly of the optical system. As illustrated in Figure 5.7,(d1)-(d3), slight defocus does not spread the energy out of the subpixel block. If we were to instead employ a large, non-repeating mask, a slight slight defocus would spread energy to neighboring subpixels, equivalent to an additional low-pass filter, as illustrated in Figure 5.7,(d4)-(d5).

### 4.3.4 Temporal Sharpening for Depth and Transient Imaging

To extract temporal information from our reconstructed images, we use a recent reported temporal PSF model [12] for SPAD sensors to sharpen our reconstructed 3D data. For depth and transient imaging, our SPAD sensor works in time-correlated single photon counting (TCSPC) mode.

This model is useful for precise temporal localization of Gaussian laser pulses from an observed time profile at each pixel $\boldsymbol{I_i}$, using a model of the temporal response of the SPAD pixel, $\mathbf{\Pi}(t)$. The gate signal $\mathbf{\Pi}(t)$ is not a simple rectangular pulse, but is distorted according to a resistor-capacitor (RC) circuit response (also compare Figure 4.6 bottom left). We band limit this RC model with a small Gaussian filter ($\sigma_f = 100$ ps in the experiments), see Figure 4.6 bottom center.

The observed time profile at each pixel $\boldsymbol{I_i}$ is then modeled as a convolution of

this gate model $\mathbf{\Pi}(t)$ with the Gaussian laser pulse $\boldsymbol{G}(t; A, \mu) = A\boldsymbol{e}^{-\frac{(t-\mu)^2}{2\sigma^2}}$, where the parameters $A$ and $\mu$ of the Gaussian are initially unknown. They can be determined by solving the following minimization problem for each pixel:

$$\min_{A,\mu} \|\mathbf{G}(t; A, \mu) * \mathbf{\Pi}(t) - \boldsymbol{I_i}\|_2^2, \tag{4.9}$$

where $*$ denotes the convolution. Please refer to the original paper of Sun et al. [12] for technical details. Instead of using a Gaussian model for the laser pulse, we note that it would be straightforward to substitute other models such as an exponentially modified Gaussian [163] to estimate parameters for inter-reflection, subsurface scattering, or fluorescent lifetime imaging (FLIM).

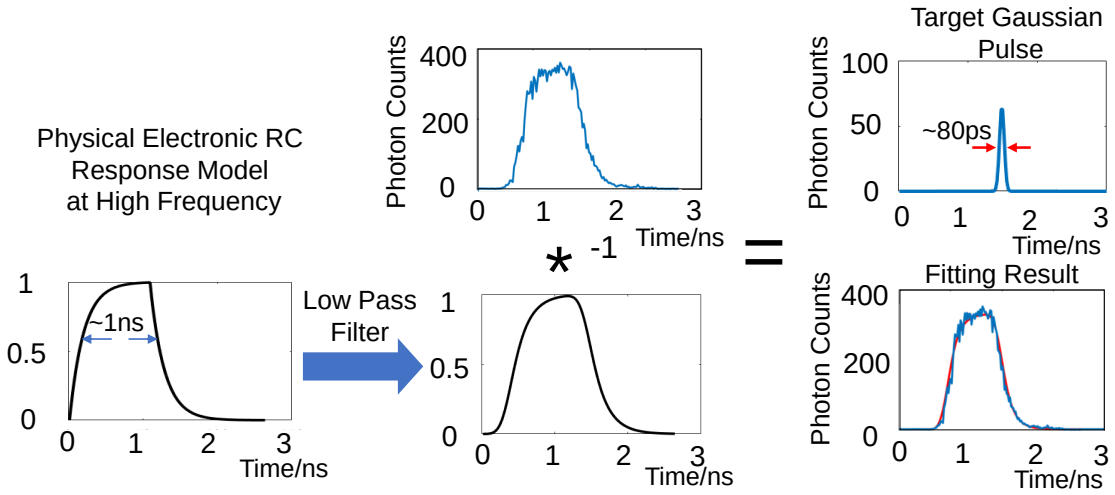


Figure 4.6: Modeling the temporal PSF of the system as the convolution of distorted SPAD gate signal and a Gaussian laser pulse profile [12]. The data of the histogram is selected from location (45, 167) in Figure 4.1 b-3.

## 4.4 Evaluation in Simulation

We first present a quantitative comparison of some of state-of-the-art SR methods like VSDR [131] and SRCNN [128]. Table 4.1 shows that although that these kinds of methods perform well on a conventional SR problems, they fail in the low fill-factor

case. In this table, each of the methods, including our own reconstruction network was trained using the low fill-factor model (i.e. without an anti-aliasing filter) on the same DIV2K dataset. We also tried VSDR and SRCNN on the optical design obtained with our method, but the resulting SNR and SSIM results are slightly worse than in the low fill factor case shown in the table.

Table 4.1: Quantitative assessment of current SR methods over the low fill-factor sampling model in PSNR and SSIM (grayscale).

| Methods | Set5 | Set14 | BSDS100 |
|---|---|---|---|
| Bicubic | 24.26/0.8336 | 21.51/0.7589 | 20.83/0.7175 |
| SRCNN | 25.27/0.8620 | 22.34/0.7812 | 21.58/0.7397 |
| VSDR | 25.45/0.8717 | 22.57/0.7915 | 21.74/0.7481 |
| **Ours** | **27.17/0.9019** | **23.97/0.8066** | **23.82/0.7691** |

Next, we present a quantitative comparison of applying our reconstruction network to *four* different sampling models: (1) Low fill-factor sampling model that considers the SPAD sensor model without the phase mask; (2) Full fill-factor sampling model that is common for other imaging sensors; (3) Low fill-factor sampling model that considers the SPAD sensor with a Gaussian PSF of standard deviation $\sigma_N = \sqrt{3\log 2}/\pi \approx 0.459$, corresponding to a least-squares fit of the sinc function that corresponds to the ideal low pass filter; (4) Our sampling model that considers the SPAD sensor model with setting the phase mask at the front focal plane of imaging lens.

To make a fair comparison, we use the same training dataset and parameters to retrain the network for the low fill-factor model, full fill-factor model, and a low fill-factor model with a Gaussian PSF. We then assess on 3 well-known datasets: Set5 [164], Set14 [165], and BSDS100 [166]. Table 4.2 summarizes the averaged PSNR and SSIM scores. We observe that, without the aid of our phase mask, the original low fill-factor model exhibits significantly worse performance than ours both in terms of of PSNR and SSIM. Concerning the Gaussian PSF, even in comparison to a perfectly

shaped Gaussian diffuser (which would need to be carefully designed, manufactured, and aligned for a specific sensor geometry, and would certainly not be an off-the-shelf part), the scores and recovered image detail (see Figure 4.8) are still worse than that of our end-to-end system. In addition, we also evaluate a hypothetical full fill-factor model that might be feasible with alternative sensor designs. The results show a clear advantage of our end-to-end design over all alternatives sampling patterns on all datasets.

Table 4.2: Quantitative comparison of 4× SR under different sampling models in PSNR and SSIM (grayscale). $\sigma_N$ is chosen to best approximate the ideal low pass filter with a Gaussian (see text).

| Model | Set5 | Set14 | BSDS100 |
|---|---|---|---|
| Low fill-factor | 27.17/0.9019 | 23.97/0.8066 | 23.82/0.7691 |
| Full fill-factor | 29.77/0.9317 | 26.13/0.8442 | 25.59/0.8069 |
| Gaussian (optimal) | 30.41/0.9360 | 26.68/0.8498 | 26.05/0.8157 |
| Gaussian (w./o. re-training) | 20.46/0.8087 | 19.64/0.7268 | 20.07/0.7016 |
| **Ours** | **30.76/0.9399** | **26.91/0.8557** | **26.23/0.8198** |

Figure 4.7 visualizes several examples selected from the test dataset. Sampling of a low fill-factor sensor destroys most information, thereby the reconstructed results suffer from noticeable artifacts and distortions. These artifacts are alleviated by our proposed method. For instance, the texture on the butterfly is well preserved, but is in comparison, corrupted by artifacts in the low fill-factor case without the phase mask. The full fill-factor sensor shows lightly better performance than that of the low fill-factor sensor, since it averages our the information at all frequencies across the full pixel block. Instead, our sampling model preserves the most desired information, showing reconstruction results closer to ground truth (GT). To this end, we believe our anti-aliasing filtering design contributes to preserving interesting details while suppressing other artifacts.

## 4.5 Prototype and Assessments

In this section, we assess the modulation transfer function (MTF) of our imaging system and present the prototype results of three application scenarios. Before detailing the experimental assessments, we briefly summarize the fabrication of the phase masks and the calibration of the PSFs.

### 4.5.1 Prototype

**Fabrication.** The phase mask is discretized into eight levels which can then be realized by repeatedly applying photo-lithography and reactive ion etching (RIE) 3 times [162, 51] on a 0.5 mm Fused Silica substrate. The principal wavelength is 655 nm, and a $2\pi$ phase modulation is used to wrap the heightmap. Refer to the supplemental document for fabrication details.

We use a FLIR mono sensor GS3-U3-50S5M with a pixel pitch of 3.45 $\mu$m to calibrate the PSF of the fabricated phase plate. The phase plate is placed at the front focal plane of a Canon 50 mm lens. A point light source with a 655 nm/10 nm bandpass filter is set 1.35 m away from the sensor. Figure 5.7a shows the calibrated PSF of our fabricated phase mask (see Section 4.3.3). The sparse dot pattern structure is due to the tiling of the phase plate as described in Section 4.3.3.

### 4.5.2 MTF Analysis

We use the slanted edge method [167] to assess the modulation transfer functions (MTF)s of our results and that of the low-resolution reference, as shown in Figure 4.9. We observe outliers larger than 1 in the plot of the SR image without phase mask (orange plot). In contrast, the MTF of our super-resolution camera is closer to the desired MTF in optical systems: smoothly and monotonously decreasing from an amplitude of 100% for the DC term to ca. 10% at the Nyquist limit of the SR image, with no erroneous maxima for higher frequencies. This result is enabled by better

preservation of super-resolution information in our learned PSFs. Here we remind the reader that MTFs are intended to characterize linear systems, and may not be the best metric of assessing non-linear computational imaging systems such as ours.

### 4.5.3 Intensity Imaging

**Experimental setup.** The prototype of normal intensity imaging is illustrated in Figure 4.10. We use an MPD-SPC3 SPAD array as the detector. The phase mask is optimized for imaging daily scenes and human activity. The SPAD array is operated in snapshot mode with the integration time set as 52 $\mu$s. We sum up 100 frames before read-out, corresponding to a total integration time of around 5.2 ms.

**Results of intensity imaging.** To validate the practicability of the proposed optically coded single-shot super-resolution design, we employ the fabricated phase mask on a normal imaging setup that acts as the basis of alternative applications, for example, depth and transient imaging, as well as low light imaging. A sequence of raw images (upsampled to the size of the reconstructed images for ease of comparison) is shown in Figure 4.11-1. The advantages of generating the optimal PSF specifically designed for the SPAD sensor's low fill-factor structure are significant. The reconstructed super-resolution results (i.e. Figure 4.11-2 faithfully preserve many details as without introducing artifacts. Therefore, for such a kind of low fill-factor sensor structure, our method succeeds in preserving the spatial information.

**Results of reference experiments.** To further demonstrate that our phase mask works as designed, we performed a reference experiment for the same scenes without phase mask. Figure 4.11-3 presents the raw images without phase mask. The visualization of the raw images contains more high frequencies compared with those with phase mask. These undesirable high frequencies only result in the loss of fine details we want to preserve, but also introduce strong artifacts as illustrated in Figure 4.11-4.

In comparison, our phase mask can preserve the most useful information while suppressing aliasing, consistent with the simulation results as described in Section **??**.

### 4.5.4 High Speed Imaging

**Experimental setup.** We use the same camera setup described above. The SPAD array is operated in snapshot mode at a frame rate of 1,250 fps with the integration time set as 80 $\mu$s. In this example we sum up 10 frames before read-out. As illustrated in Figure 4.10b, we use a CPU fan as a high speed spinning object. One of the blades is marked black as a position tracker, as shown in Figure 4.10c.

**Results of high speed imaging.** The optically coded single-shot super-solution camera fits well with unsynchronized and non-repeatable conditions, where time-sequential spatial resolution enhancement methods like compressive sensing with a DMD, 2D mechanical scanning, or 1D line scanning are not applicable. As illustrated in Figure 4.12, we successfully capture and reconstruct the frames of a high-speed rotating fan (roughly calculated at 3,750 rpm from the shown frames). Figure 4.12a presents the captured raw data with darkcounts and background noise removed. Figure 4.12b presents the reconstructed 4$\times$ super-resolved frames. We can distinguish the fine details of fan and football. For more details, please refer to the supplemental video.

### 4.5.5 Depth and Transient Imaging

**Experimental setup.** Figure 4.13 illustrates the experimental setup for depth and transient imaging and the corresponding scenes. We use a 655nm picosecond laser (PicoQuant LDH P-650) with an average power of around 1 mW as the illumination source. The FWHM of the laser pulses is around 80 ps, and the repetition rate is 50 MHz. To illuminate the scene smoothly, we scatter the laser beam using a diffuser

and use an 80 mm plano-convex lens to re-concentrate the overly scattered beam.

We operate the SPAD camera in TCSPC mode with a 200 ps gate width and a 20 ps phase shift per cycle. The integration time is set to 52 $\mu$s, and 1,500 frames are summed up before read-out. In total, the capture process lasts around 9.8 s.

During the capture, the SPAD array sends the synchronizing signal to trigger the laser driver and then counts the arrival photons with a fixed phase offset of the gate. After sufficient integration, the SPAD camera shifts the gate window (i.e., 20 ps delay) and captures another frame until covering all designed phase offsets.

**Fabrication Details** As mentioned in main text, the designed phase plates are discretized to 8 height levels, which can then be realized by repeatedly applying photolithography and reactive ion etch (RIE) process 3 times [162]. The core of fabrication (i.e. photolithography and RIE) is shown in Figure 4.14. The thickness of Cr layer is 100nm and the etching depths for each circle are 178nm, 356nm, and 712nm, sequentially.

The microscope images of fabricated diffractive phase plates are shown in Figure 4.15. Each phase unit is designed as 0.873mm$\times$0.873mm, and we repeat 16$\times$16 units for the final diffractive phase plate. With this kind of repeatable designing scheme, our phase plates fit well with different sizes of aperture, exhibiting an aperture invariant PSF behaviour. In addition, our designed phase unit has a smooth and continuous profile. This releases the requirement of a very high precision fabrication method. For commercial use, one can use state-of-the-art micro-stamp methods [168] to fabricate the phase plates in a low budget.

**Results of depth imaging.** In this experiment, we demonstrate the ability to resolve the geometric details of several objects (fans, horse, wooden toys, etc.) in the scene depicted in Figure 4.16. As shown, the reconstructed intensity (Figure 4.16b) and depth images (Figure 4.16c) exhibit details that are hardly distinguishable in

raw data, for instance the edges of fans and wooden toys. From Figure 4.16a we observe that the raw images obtained by summing over the time axis remains very noisy although the dark counts and background noise have been mostly removed. Compared to the raw data of intensity imaging, i.e. Figure 4.11a), the summed pixel values show a considerably larger uncertainty, which makes it challenging to reconstruct good quality results. This is because the output power of our laser is very low with an average output only around 1 mW. Furthermore, the light is scatted to illuminate the entire scene. Consequently, only a few photons can be collected by our camera after bouncing back.

**Results of transient imaging.** Figure 4.17 presents the selected results of reconstructed transient frames. A mirror is placed near the objects to reflect the light. In Figure 4.17a, the light pulse starts hitting the objects, resulting in a gradual increase and then a gradual decrease of the illumination. Later, the reflected light from the objects propagates to the mirror. Similarly, the reflected image (left part) shows the same phenomenon as the objects that the illumination gradually increases and then gradually decreases. The results in Figure 4.17b show a similar process. Thus, we have successfully captured and reconstructed high resolution transient phenomena from the low resolution raw data.

## 4.6 Discussion

**Fabrication feasibility and generalization.** Our optimized PSFs are relatively small, and which means that the phase plate only needs to diffract the light slightly, which can be achieved with relatively large feature sizes (5 $\mu$m in our experiments). This easily fits within the fabrication capability of inexpensive mass-production methods like micro-imprinting. In practice, the assembling accuracy (rotation $\pm 4^{\circ}$, displacement $\pm 2$ mm) shows a minimal impact on reconstruction results. It is viable

to design systems where the phase plate can be easily switched by end users, simple as switching a regular lens, to maximize the performance for different application scenarios. We believe the proposed design paradigm can be generalized to alternative low fill-factor and low resolution sensors like on-board pixel processing circuits [169], 3D cameras, fluorescent analyzers, thermal cameras, etc.

**Limitations for depth and transient imaging.** We reasonably ignore the multipath effect at the stage of proof-of-concept since current illumination region is constrained within a level of a few decimetres. But there are several limitations that affect the reconstruction quality of depth and transient imaging. On the one hand, the picosecond laser used in our experiments has a power of only 1 mW. On the other hand, current photon detection efficiency (PDE) is only 12% at the wavelength of 655 nm. These two essential hardware constraints, in tandem with the need of diffusing laser beam into a 2D space to illuminate the whole scene, result in a fact that only a few reflected photons can be collected by the sensor. In contrast, line SPAD-based scanning methods [106, 155] scatter the laser beam only into a line and use the spectra of 450 nm, corresponding to a SPAD detection PDE around 50%. Therefore, currently the relatively lower light efficiency of our method adds difficulties to tackle the strong noise in the reconstruction.

**Future Work.** In depth and transient imaging applications increased illumination power always improves measurement range and robustness to ambient light. However, safety and cost concerns set tight limits to the laser power in many scenarios. To overcome this problem, using an intensity-modulated continuous laser, similar to amplitude modulated continuous wave (AMCW) time-of-flight sensors, can be a good alternative. A future direction of research would be to build a counting and digital version of AMCW TOF sensors using continuous wave illumination. This can be achieved by replacing the two capacitors that collect the charge of a photodiode with

two counting units that count the photons of SPAD. In this way the SPAD-PMD device can lower the requirements on illumination while exhibit more robustness to ambient light. SPAD arrays are a particularly promising technology for the field of fluorescent lifetime imaging, where state-of-the-art hardware solutions either suffer from low resolution or require complex and time-consuming mechanical scanning. To this end, optimizing a phase mask can enable a fast, high resolution, and scanning-free fluorescent lifetime imaging system.

## 4.7 Conclusion

In conclusion, we present a general design paradigm to realize an optically coded single-shot super-resolution camera for low fill-factor sensors. This is achieved by incorporating optical design, sensor modeling, and deep network reconstruction. We build a high-resolution SPAD camera and demonstrate its viability in the application scenarios of intensity, high speed, and depth/transient imaging. Our approach for the first time overcomes the spatial resolution limit of existing SPAD sensor arrays with a single-shot capture, without the need of any mechanical scanning or repeatable measurement. The hardware improvement requires only a relatively inexpensive phase mask to the front focal plane of an existing optical system. We envision a wide range of applications across computer vision, sensing, and microscopic imaging.

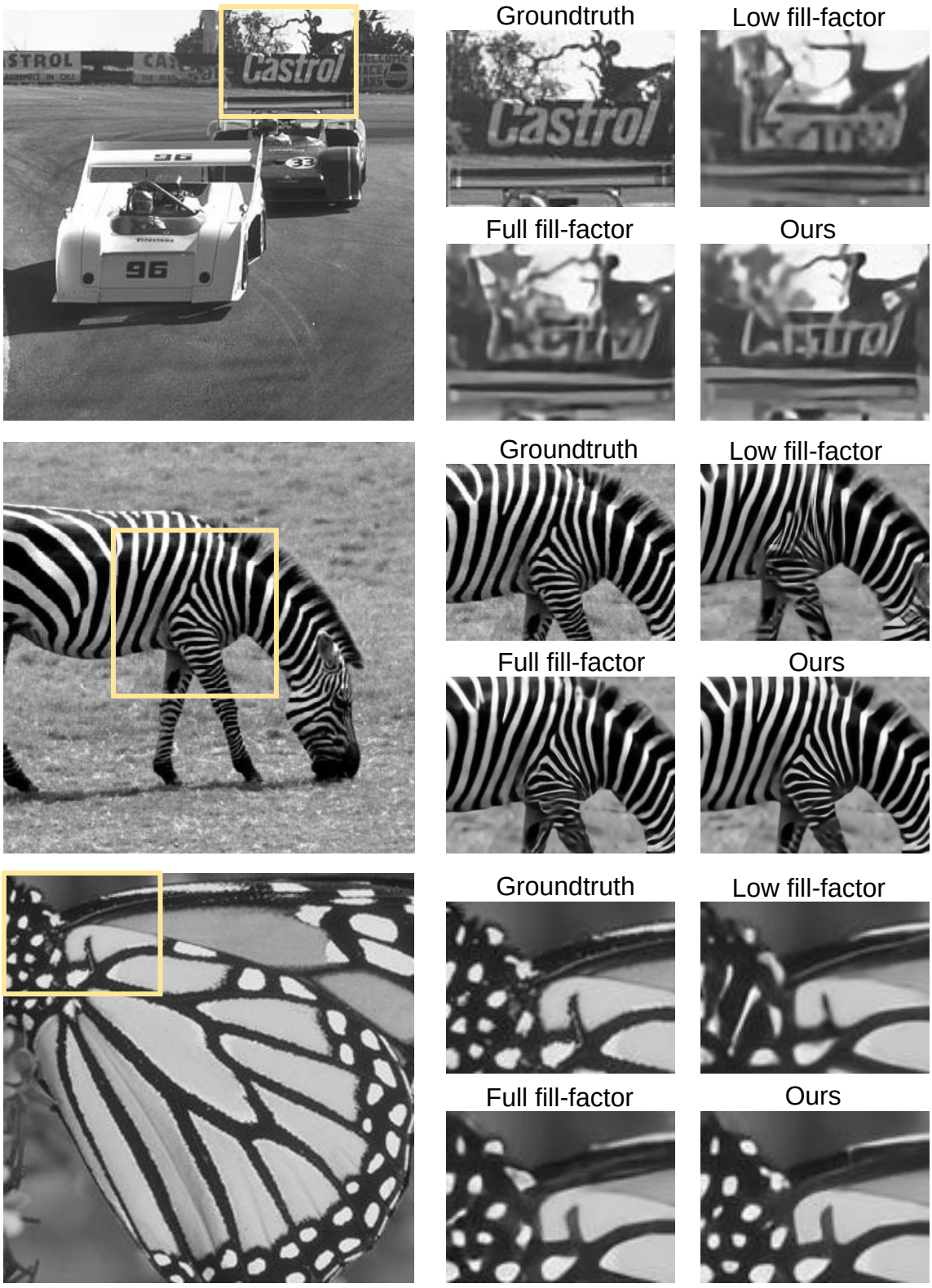


Figure 4.7: Selected examples of 4× SR under different sampling models. For the low fill-factor case, we direct apply the low fill-factor model of SPAD to sample the high resolution images to obtain 1/4 resolution images. For the full fill-factor case, we average the 4×4 pixel area to obtain 1/4 resolution images. For our method, we apply the low fill-factor sampling model of SPAD with pre-filtering using our learned PSF kernel.

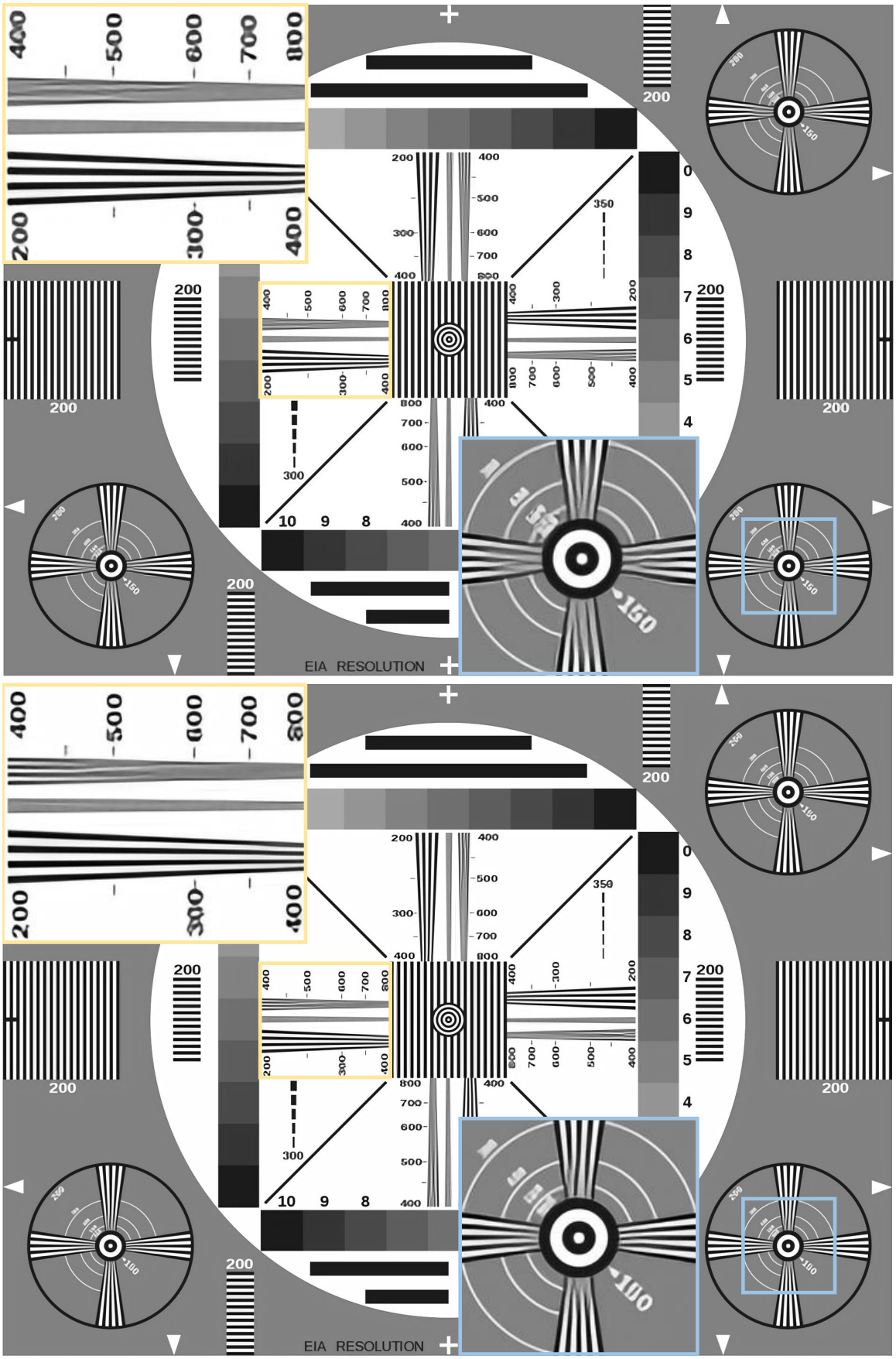


Figure 4.8: Imaging performance of the best Gaussian PSF (top) and our end-to-end learned PSF (bottom). Our end-to end learned approach does show significantly better preservation of details above the Nyquist limit, see insets.

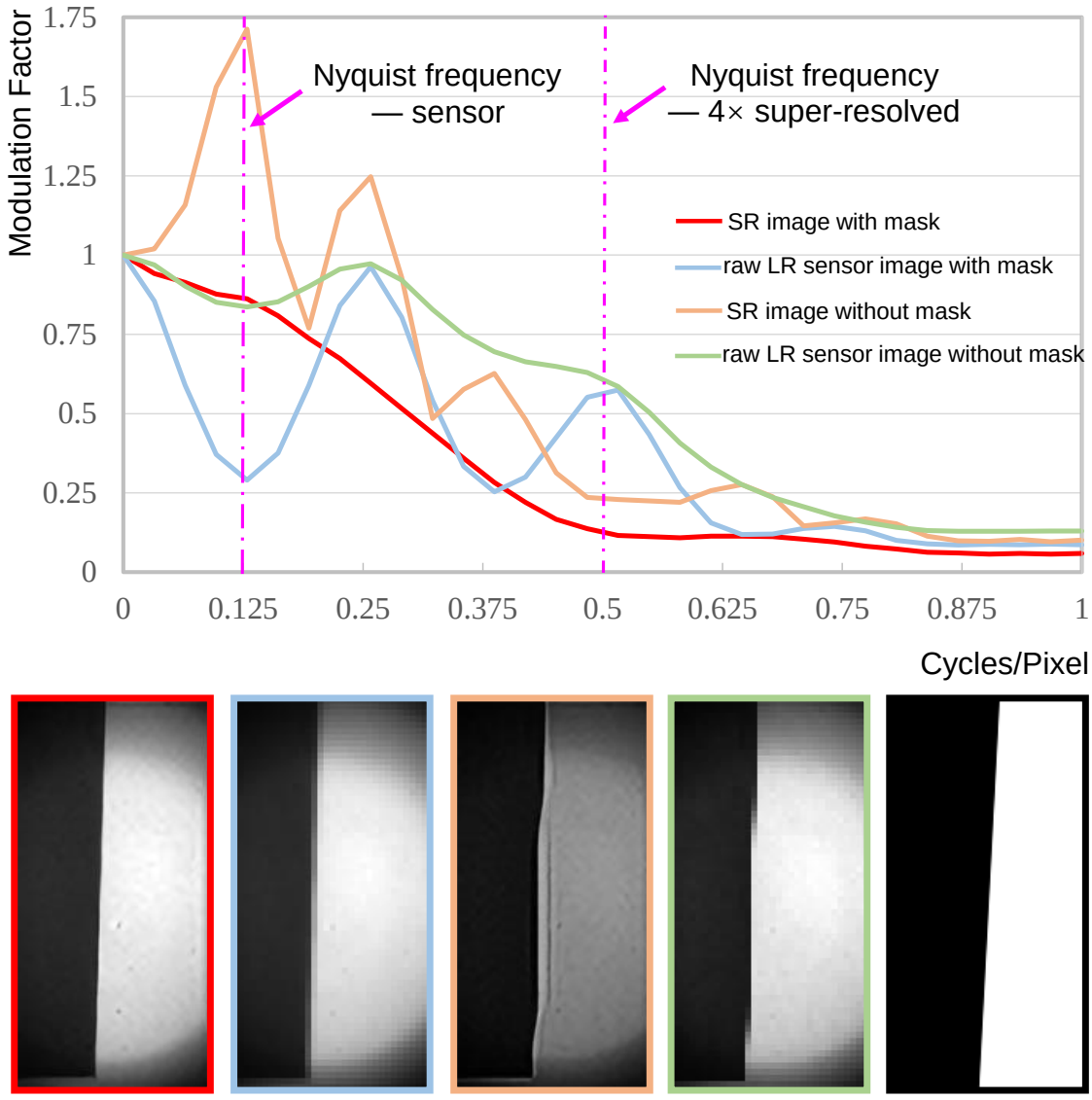


Figure 4.9: MTFs derived from experimental results, including raw LR sensor image and 4× super-resolved SR image with and without phase mask, respectively. The corresponding images are revealed with different color plots and the ideal 4× SR image is marked by black color.

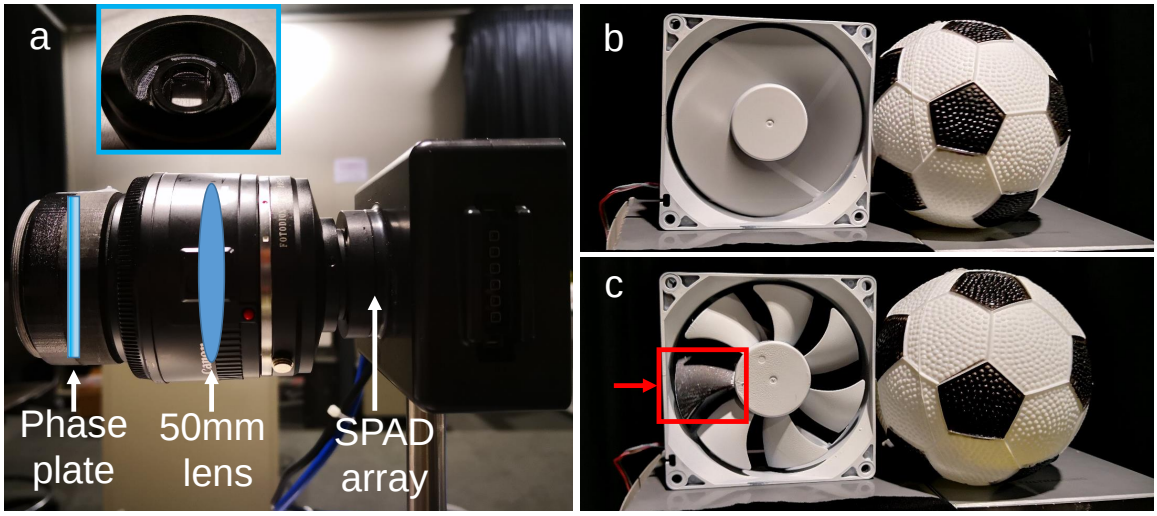


Figure 4.10: Prototype for normal/high speed imaging and the scene: a) The prototype of normal imaging and high speed imaging. b) The scene of running fans captured with a regular RGB sensor. c) Static states of the scene shown in b), and the red marked area are manually set as black to mark the rotating position.

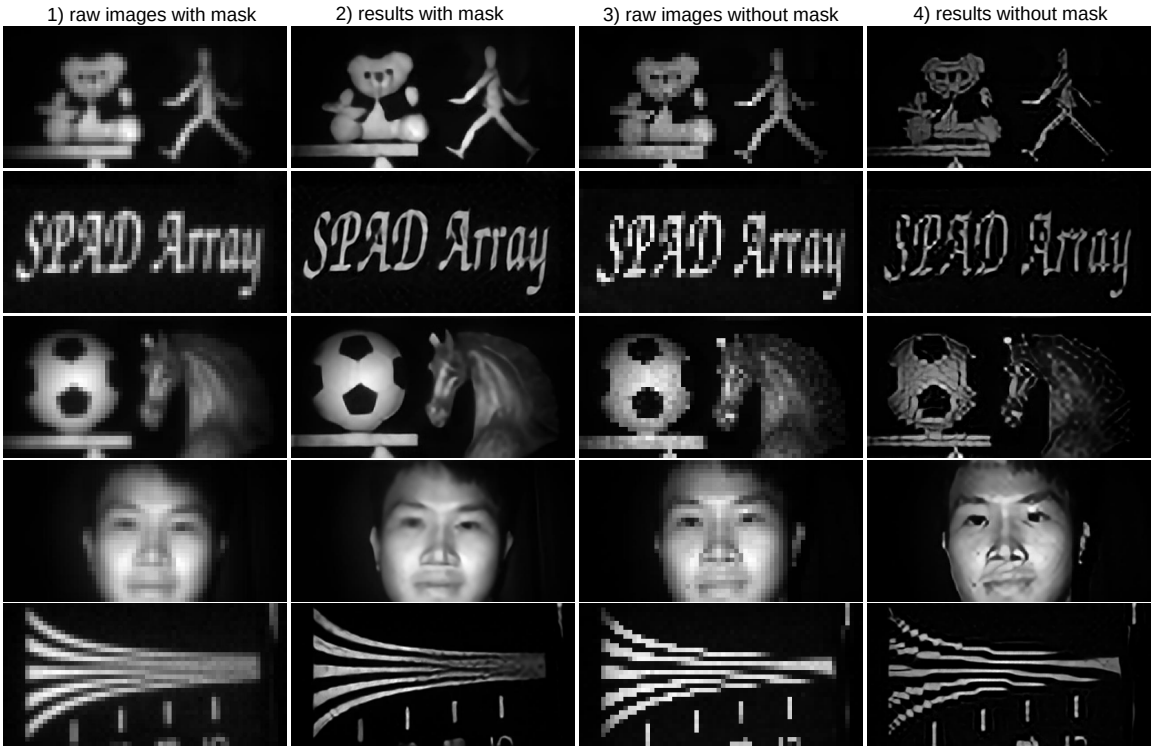


Figure 4.11: Results of normal imaging. 1) Captured raw images with phase mask and with dark counts and background noise removed. 2) Results with phase mask. 3) Captured raw images without phase mask and with dark counts and background noise removed. 4) Results without phase mask.

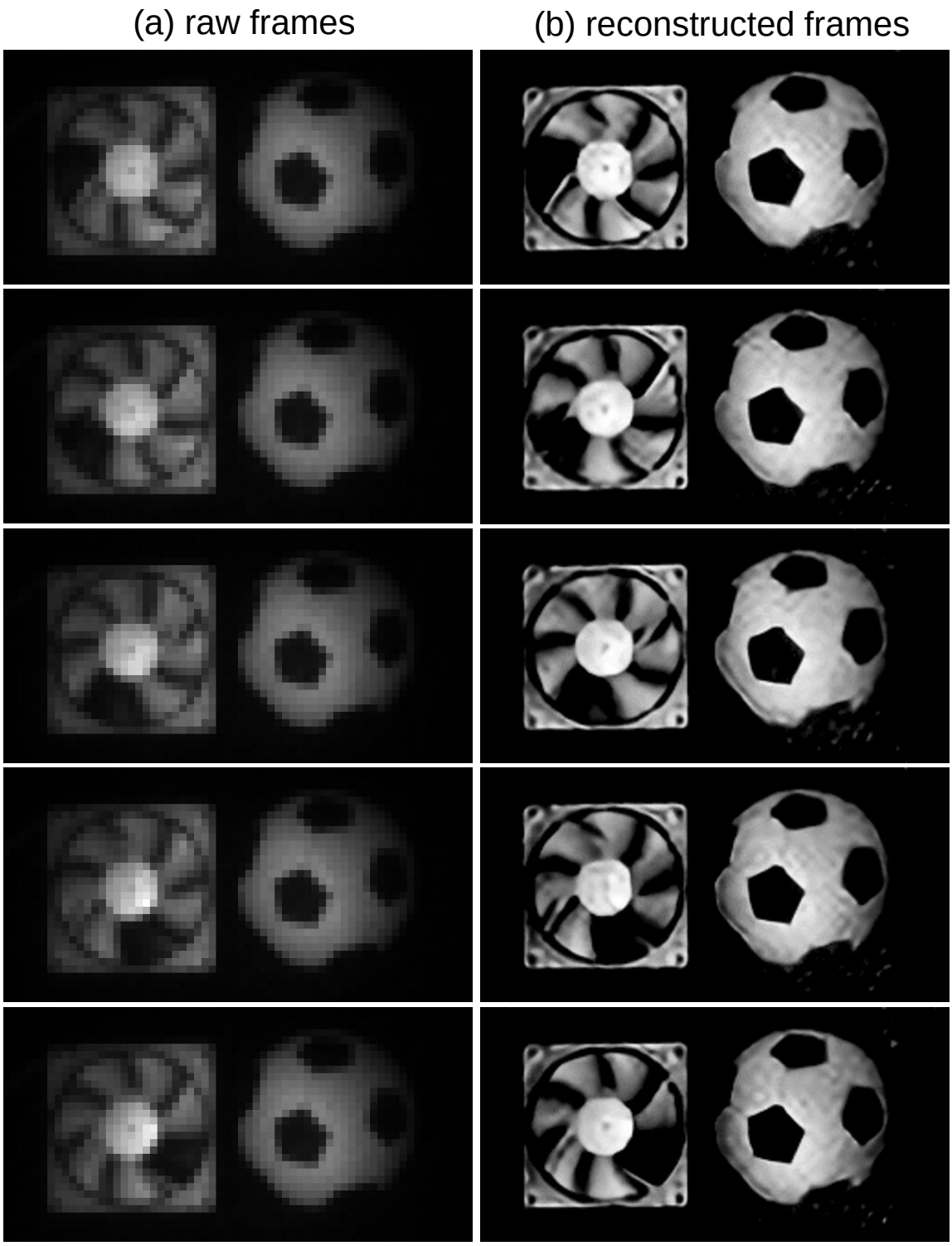


Figure 4.12: Results of high speed imaging. The displayed data is selected for every five frames and we set the frame rates around 1,250 fps. (a) Selected raw frames with darkcounts and background noise removed. (b) Reconstructed high resolution frames.

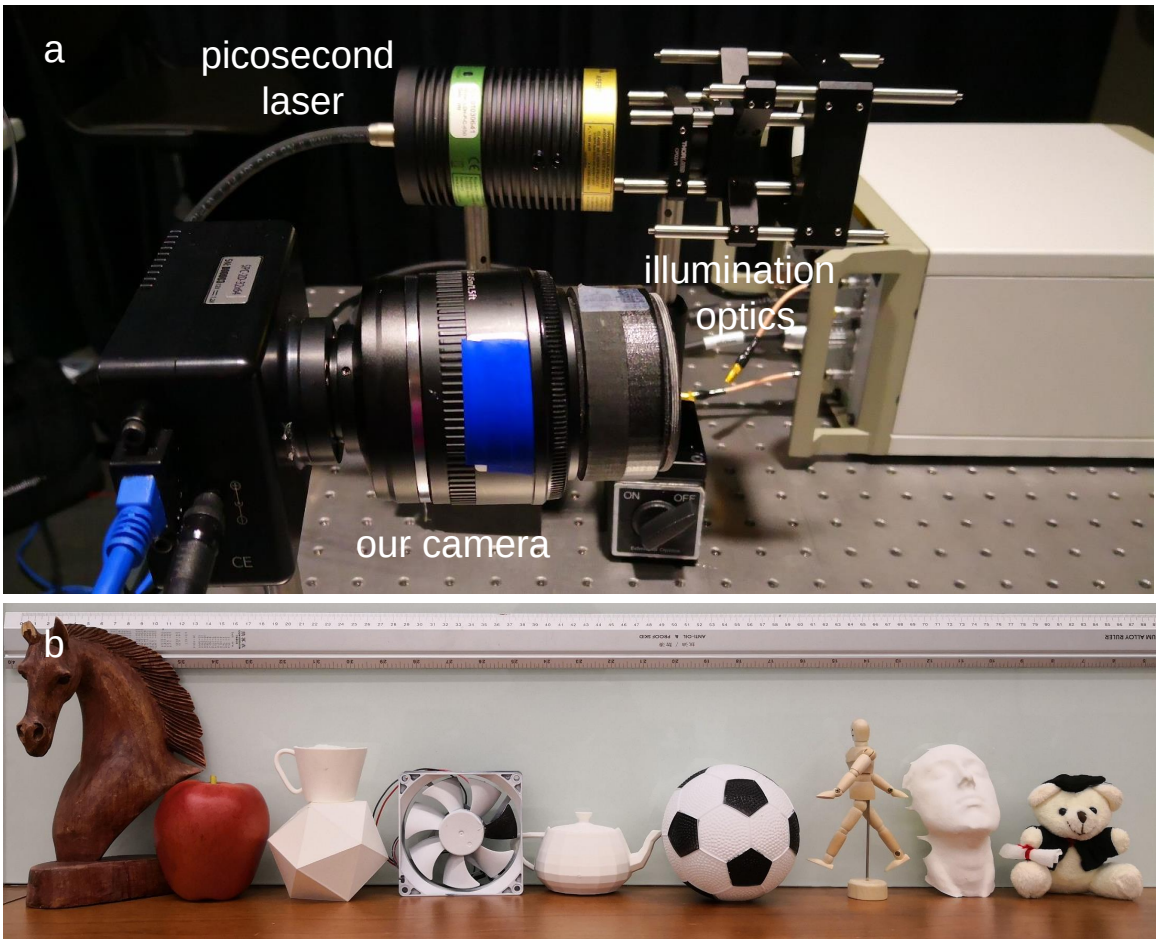


Figure 4.13: Photograph of hardware set-up of depth and transient imaging (a); and the scenes used in experiments (b).

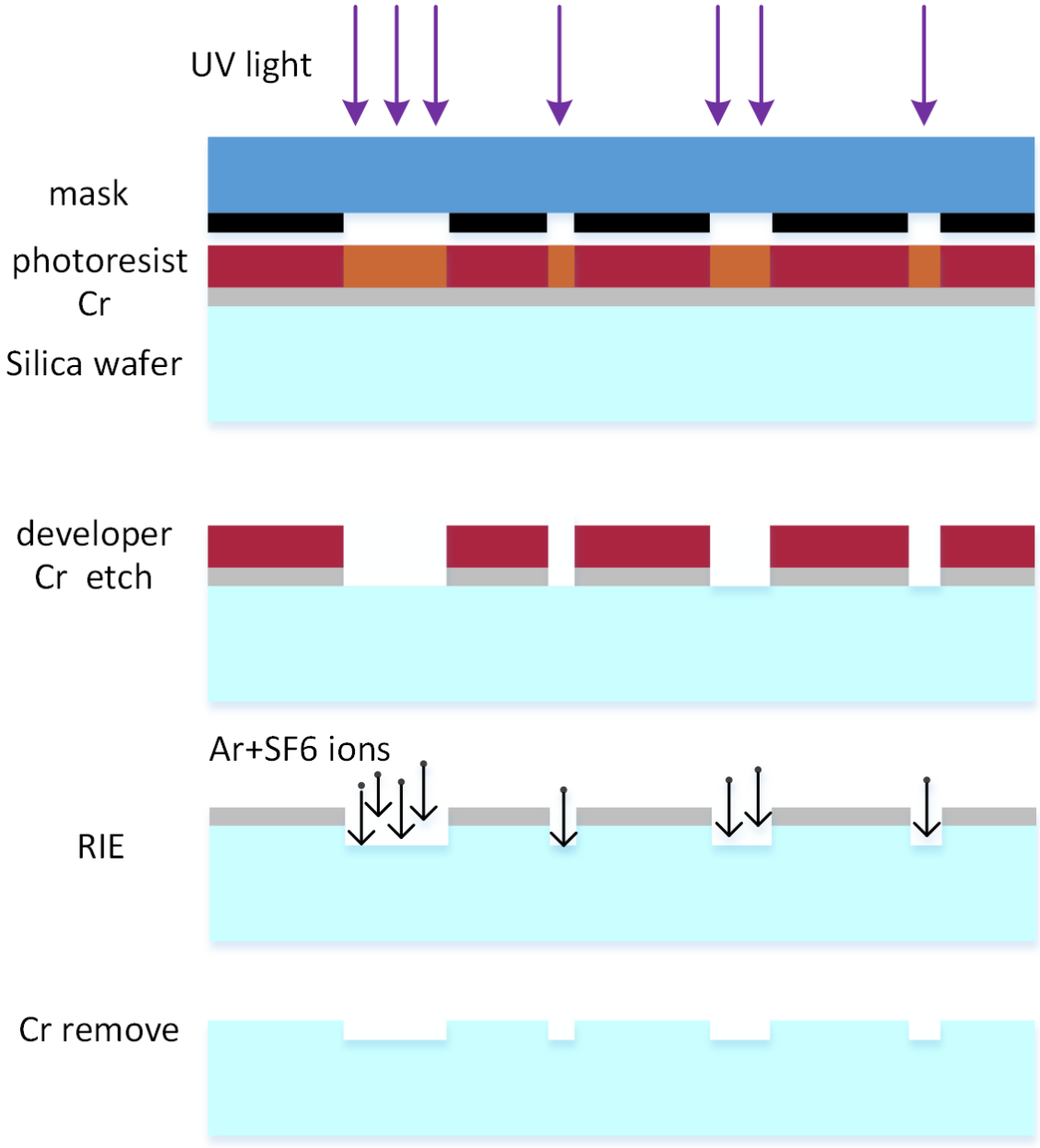


Figure 4.14: Steps of DOE fabrication. First, patterns are transferred from masks to photoresist on Fused Silica wafer through the exposure under UV illumination and the following develop process. Then, the transferred patterns are converted to binary profiles on the wafer by Cr etching and reactive ion beam (Ar + SF6) bombardment. The final binary profile is obtained by removing the Cr layer.

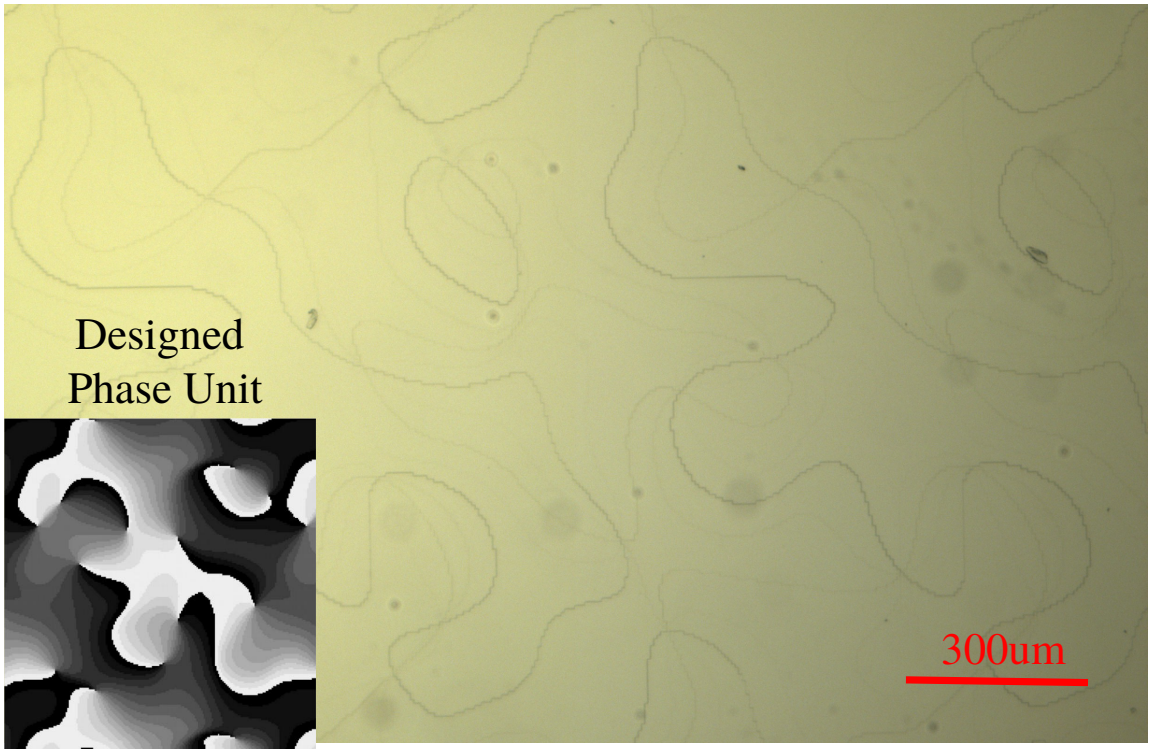


Figure 4.15: 5X microscope images of a fabricated DOE.

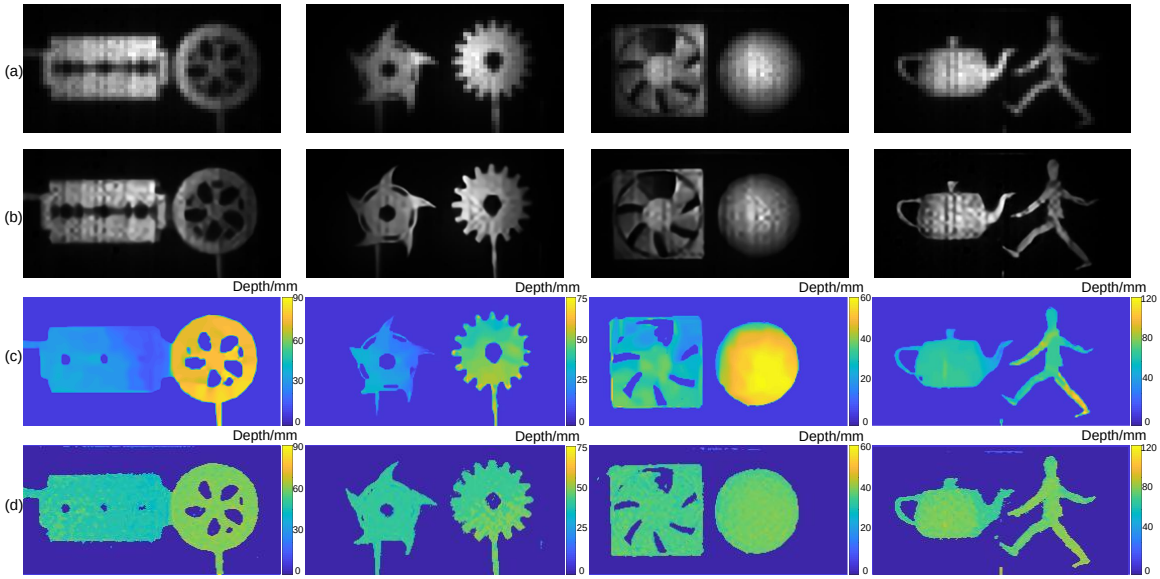


Figure 4.16: Results of depth imaging: (a) raw image(with darkcounts and background noise removed) summed over the time dimension; (b) reconstructed intensity image according to (a); (c) reconstructed depth image; (d) reconstructed depth image without temporal deconvolution.


Figure 4.17: Results of transient imaging. From left to right and from top to bottom are the selected frames of our reconstructed transient video stream. We here present one frame out of every 9 frames with an visualized interval of 180 ps. The right bottom of a) shows the captured scene containing a cup, a polyhedron, and a large mirror. The right bottom of b) is the captured scene containing a wooden skeleton and a large mirror.

# Chapter 5

# End-to-End Encoding Through Optimizing Phase Mask: Single-shot HDR Imaging

In the previous chapter, we take one more step that optimizing the PSF together with the post-processing, bring the imaging system design into end-to-end fashion and successfully applied it to enable the super-resolution of low fill-factor SPAD camera. However, it is based on the assumption of a single wavelength such that we can find a fixed relationship of optics and PSFs through GS based method. When applying to color imaging or multispectrum imaging, we need to consider that the refractive index varying with the wavelength. Therefore, in this chapter, we build a differentiable DOE based on fresnel approximation to compromise color channels.

HDR imaging is an essential imaging modality for a wide range of applications in uncontrolled environments, including autonomous driving, robotics, and mobile phone cameras. However, existing HDR techniques in commodity devices struggle with dynamic scenes due to multi-shot acquisition and post-processing time, e.g. mobile phone burst photography, making such approaches unsuitable for real-time applications. In this chapter, we propose a method for snapshot HDR imaging by learning an optical HDR encoding in a single image which maps saturated highlights into neighboring unsaturated areas using a DOE. We propose a novel rank-1 parameterization of the DOE which drastically reduces the optical search space while allowing us to efficiently encode high-frequency detail. We propose a reconstruction network tailored to this rank-1 parametrization for the recovery of clipped information from the encoded measurements. The proposed end-to-end framework is validated through


Figure 5.1: Due to pixel saturation, image detail in bright regions is lost in a single snapshot LDR image. Our camera, with the learned optic prototype (left), captures LDR images where high-intensity image content is encoded through a series of streaks (center left). This allows us to reconstruct the lost highlights (center, center right) with a specialized two-stage CNN.

simulation and real-world experiments and improves the PSNR by more than 7 dB over state-of-the-art end-to-end designs.

## 5.1 Introduction

HDR imaging has become a commodity imaging technique as evident by its applications across many domains, including mobile consumer photography, robotics, drones, surveillance, content capture for display, driver assistance systems, and autonomous driving. The pixels in conventional CMOS and CCD image sensors act as potential wells that saturate when the well capacity is reached. Unlike film, which provides a gradual compression of high intensities, digital image sensors thus suffer from a hard cutoff at some peak intensity, so that information about the saturated bright regions is irrevocably lost. By reducing the exposure time, brighter regions can be recovered, but at the cost of under-exposing, i.e., reducing signal photons in darker image regions.

As a result, single captures of conventional sensors provide high fidelity only for low-contrast scenes, while struggling for high-contrast scenes at night with both low- and high-flux scene content.

Although existing HDR imaging methods in widely deployed consumer smartphone devices [35, 170, 93] successfully overcome this limitation by acquiring bursts

of captures, the combined capture and processing time of multiple seconds [35] is prohibitive for many applications in robotics and autonomous driving that demand real-time feeds.

Faster multi-capture imaging methods [171, 172, 173], relying on only 2-3 low-dynamic range exposures and hardware exposure fusion, fail for higher dynamic scenes typical in automotive and robotics applications. As an alternative, emerging sensor designs multiplex exposures on the sensor [174, 175, 176], however, at the cost of spatial resolution required for spectral or spatial information. Optical splitting methods using multiple sensors [177] are often not practical in an application due to their cost and footprint. To tackle this issue, a line of recent work explores the hallucination of HDR images [178, 179] from single low-dynamic range (LDR) captures. These methods can only rely on semantic context but no measurement signal to recover the clipped HDR regions. In an alternative direction, Rouf et al. [180] proposed a hand-crafted star filter attachment to optically encode lost information by spreading out saturated highlights to nearby regions. Unfortunately, their approach only achieves low image quality far below that of recent hallucination approaches.

In this chapter, we revisit this idea, but by learning an optical HDR encoding in an end-to-end optimization. To this end, we jointly design the optical PSF together with the inverse reconstruction method, i.e., the post-capture processing that recovers the latent HDR scene from the input measurement, which we formulate as an (RAW-)image-to-image neural network. However, we found that applying existing end-to-end methods [14, 181] easily finds a local minimum of the vast design space, parameterized by an unconstrained diffractive phase plate optic. Instead, we parameterize the diffractive element in the proposed optical design with a rank-1 phase pattern. This constrained PSF design spaces makes allows us to tailor the architecture of the reconstruction network to the recovery from such streak-encoded measurement. We optimize the diffractive optic and reconstruction algorithm jointly

in an end-to-end optimization which finds a local minimum that outperforms vanilla end-to-end designs with similar network capacity by more than 7 dB PSNR.

We demonstrate the proposed approach outperforms the state-of-the-art snapshot HDR methods in simulation. We prototype our design camera system by fabricating the DOE and demonstrate on a broad set of experimental in-the-wild captures, that this method generalizes to unseen scenarios, outperforming existing optical designs. Our method is most effective for recovering concentrated high-intensity light sources such as street lamps. In addition, we also show that the proposed network is effective in removing glare from in-the-wild automotive optics with windshield-induced streaks.

Specifically, we make the following contributions:

- We introduce a novel rank-1 parameterized optical design that learns to encode saturated information with a streak-like PSF.
- We co-design a tailored reconstruction network which first splits the unsaturated regions from the coded information and then recovers the saturated highlights from the encodings.
- We validate the proposed method in simulation and on real-world measurements acquired with a fabricated prototype system. The proposed method outperforms existing designs by over 7 dB in simulation.

## 5.2 Related Work

### 5.2.1 Multi-exposure HDR Imaging

Traditionally, HDR imaging is performed by sequentially capturing LDR images for different exposures and then combining them through exposure bracketing [6, 182, 183, 184, 185, 186]. This approach is unsuitable for handling highly dynamic scenes and for fast captures necessary for real-time applications. More rapid HDR imaging can be realized with burst HDR acquisition [35, 170, 93]. However, these techniques

still suffer from motion artifacts and require seconds for capture and processing for a single acquisition.

To alleviate motion artifacts, prior work has employed HDR stitching [187, 188], optical flow [189], patch matching [190, 191, 192, 193, 194], and deep learning [195, 196]. These techniques have even enabled HDR videography, but the post-processing cost makes them impractical for fast capture. Ultimately, these approaches attempt to find a trade-off between densely sampling different exposures and post-processing computation time.

### 5.2.2 HDR Snapshot Reconstruction

A large body of work has explored reconstructing HDR content from a single LDR image, a process referred to as inverse tone-mapping. Early work in this domain utilized heuristic approaches [197, 198, 199, 200], but often does not provide satisfying HDR reconstructions [201, 202]. Building upon these works, deep learning has been used to hallucinate HDR content from LDR images [178, 179, 203, 204, 205, 206, 207, 208, 209, 210, 211]. These approaches generate plausible reconstructions of low-light regions but fail to recover saturated details accurately.

Several approaches encode information into the captured LDR image to allow for better estimation of highlights. This can be achieved by modifying the sensor architecture through spatially varying pixel exposures [174, 212], convolutional sparse coding [213], compressed sensing [214], or modulo cameras [215]. Drawbacks of these approaches include the need for expensive custom cameras and loss of detail in the low dynamic range. Furthermore, recovering highlights in scenes involving very large dynamic ranges is still a challenge for these approaches. Instead of modifying the sensor, other approaches place optical components in front of conventional cameras to affect the captured LDR image. Hirakawa et al. [216] utilized color filters to avoid saturation of any single color channel. Rouf et al. [180] proposed to use a known

optical element to spread saturated information content into unsaturated regions. Although this allows for high fidelity estimation of highlights, these techniques leave noticeable artifacts in the unsaturated areas.

### 5.2.3 End-to-end Optics Design

Joint optimization of optics and reconstruction has demonstrated superior performance over traditional heuristic approaches in color image restoration [217], microscopy[218, 219, 220, 221], monocular depth imaging [16, 25, 27, 222], super-resolution and extended depth of field [14], and time-of-flight imaging [223, 120].

We propose an end-to-end optimization framework for single-shot HDR imaging. Drawing inspiration from Rouf et al. [180], our optimized optic is a DOE that encodes clipped highlights into specific unsaturated regions. The ample design space of DOEs allows for rich optical encodings but has the unintended consequence of being challenging to optimize. As such, investigating a suitable parameterized model of the DOE becomes a critical design step. Recent work, in parallel to ours, explores end-to-end optimization of optics for HDR imaging by directly learning a heightmap for the DOE [181]. This approach causes the DOE to produce shifted scaled copies of saturated content that allow for HDR reconstruction but that are difficult to remove from the unsaturated regions. Another approach used by Sitzmann et al. [14] is to represent the DOE with Zernike polynomials, but this only allows the DOE to affect low frequencies and is inadequate for capturing high-frequency detail in HDR scenes.

In this chapter, we found that by constraining the DOE height map model to a rank-1 phase pattern, our DOE learns to produce streak patterns that are easy to remove from the unsaturated regions but still allow for high fidelity reconstruction of saturated image content. We employ a structured multi-stage CNN, instead of a single-stage U-Net as in [181], to perform these tasks step by step.

## 5.3 Image Formation Model

Our image formation model is illustrated in Figure 2.1 in Section 2.2.2, and the basic pipeline is illustrated in Chapter 3. In the following, we describe the details that distinguished to the forward optics model, and it can later be used for end-to-end optimization.

### 5.3.1 DOE Layer and Rank-1 Factorization.

Existing end-to-end frameworks have used an unconstrained height map model for the DOE, where each location in the $m \times m$ height map is a learnable parameter [14]. We found that this model has a tendency to produce local minima in the form of very local encodings such as shifted and scaled copies of highlights, as also shown in parallel work [181], but rarely produces non-local encodings such as streaks. Using these local encodings provides lower quality HDR reconstructions as they are difficult to separate from the unsaturated areas in the close neighborhood. We note that alternative parameterizations such as a truncated Zernike basis [14] are also not suitable for our application, because even though it can model non-local encodings, it is only suitable for low spatial frequencies and cannot encode high-frequency content.

To tackle this challenge, we propose a novel rank-1 decomposition of the 2D height map which not only can encode high frequencies but also reduces the number of parameters touched in training. The height map at location $(x', y')$ is given by

$$\mathbf{h}(x', y') = h_{\max}\, \sigma\left(\mathbf{v}\mathbf{q}^\top\right), \tag{5.1}$$

where $\mathbf{v} \in \mathbb{R}^{m\times 3}$ and $\mathbf{q} \in \mathbb{R}^{m\times 3}$ are trainable variable basis pairs whose outer product describes the DOE height map, $\sigma$ is the sigmoid function, and $h_{\max} = 1.125$ $\mu$m is the maximum height that corresponds to $2\pi$ phase modulation at $\lambda = 550$ nm for fused silica. The sigmoid function $\sigma(x) = 1/(1 + e^{-x})$ is applied element-wise to $\mathbf{v}\mathbf{q}^\top$

to clamp the range to [0,1].

Our rank-1 parameterized model encourages global optical encodings, such as streaks, while still permitting local encodings, such as peaks. Furthermore, this parameterization produces a grating-like height map which is more suitable for DOE fabrication. In addition to the rank-1 parameterization, we also use an additional constraint to assist the framework in finding proper optical encodings. To ensure that highlights are encoded without severely affecting low-light regions, we adopt a regularization loss to constrain 94% of the energy into the center of the PSF and to spread the remaining 6% into the surroundings. We found that if we take our converged height map and continue optimizing without our rank-1 parameterization, then the optimized height profile is still maintained, which suggests that we do indeed find a good local optimum. Please refer to the Supplemental Material for details.

**Sensor Model.** The image captured by the sensor $\mathbf{I}_s$ is given by

$$\mathbf{I}_s = \mathbf{s}(\mathbf{I} * \mathbf{p} + \eta), \tag{5.2}$$

where $\mathbf{I}$ is the high dynamic range ground truth image, $\mathbf{p}$ is the point-spread function of the optical system, $\eta$ is sensor noise, and $\mathbf{s}(\cdot)$ is the camera response function that clips to [0, 1]. Note that $\mathbf{I}_s$ and $\mathbf{I}$ are both continuous variables.

## 5.4 End-to-end Design and Reconstruction

The proposed end-to-end imaging system consists of three main parts: a differentiable optical model, a robust network for recovering and separating the unsaturated image $\mathbf{I}_{\mathbb{U}}$ (i.e., pixel values in $[0, 1]$) from the residual information $\mathbf{I}_r$ encoded by the saturated image $\mathbf{I}_{\mathbb{S}}$ (i.e., pixel values in $[1, 2^8]$), and a reconstruction network for inferring $\mathbf{I}_{\mathbb{S}}$ from $\mathbf{I}_r$. In a final step, the recovered unsaturated component $\mathbf{I}_{\mathbb{U}}$ and the recovered highlight component $\mathbf{I}_{\mathbb{S}}$ are combined using a fusion network to predict the latent

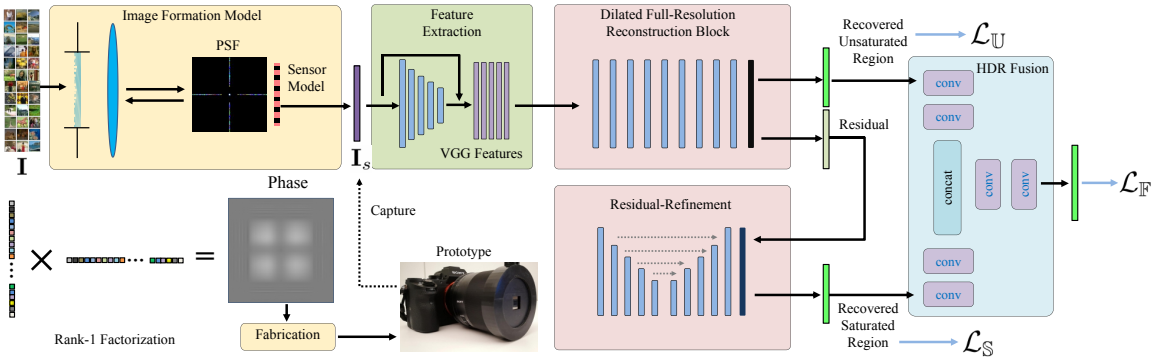


Figure 5.2: Our end-to-end pipeline consists of the image formation model and CNN reconstruction. Our CNN is divided into several stages that focus on separating the encoding from the captured LDR image, recovering the highlights, and fusing the recovered unsaturated and saturated regions to produce the final HDR prediction. After fabrication our image formation model is replaced by real-world captures.

HDR image $\mathbf{I}$.

**Differentiable Optical Model** We implement the optical model as described in Section 5.3.

**Residual Splitting Network.** We first discuss the network for reconstructing $\mathbf{I}_{\mathbb{U}}$ and separating this unsaturated part from $\mathbf{I}_r$. Our residual splitting network $f_{\mathbb{U}}$ takes in the coded LDR sensor capture $\mathbf{I}_s$ and outputs a prediction $\hat{\mathbf{I}}_{\mathbb{U}}$ for the unsaturated image and a prediction $\hat{\mathbf{I}}_r$ for the residual information:

$$\hat{\mathbf{I}}_{\mathbb{U}}, \hat{\mathbf{I}}_r = f_{\mathbb{U}}\left(\mathbf{I}_s\right). \tag{5.3}$$

Inspired by recent work on separation of reflection from transmission in single-shot images [26], the network first uses a pre-trained VGG model to extract feature maps. Specifically, we used the pre-trained VGG-19 network to extract "conv1_2", "conv2_2", "conv3_2", "conv4_2" and "conv5_2" feature maps and bilinearly upsampled them to the input image size. These feature maps, along with the input image, are then compressed to 64 channels by using a $1 \times 1$ convolution layer before being

fed through seven $3 \times 3$ dilated convolution layers with dilation rates from 1 to 64 (Dilated Full-Resolution Reconstruction Block in Fig 5.2). Each dilated convolution layer has 64 channels. Finally, a $1 \times 1$ convolution layer is used to reduce to six channels, three of which correspond to $\hat{\mathbf{I}}_{\mathbb{U}}$ and the other three correspond to $\hat{\mathbf{I}}_r$. Each dilated convolution layer is followed by a Leaky ReLu activation (slope = 0.2) and an instance normalization layer. The loss on the unsaturated pixels $\mathcal{L}_{\mathbb{U}}$ as shown in Figure 5.2 forces this network to effectively split streaks from the unsaturated image $\hat{\mathbf{I}}_{\mathbb{U}}$.

**Highlight Reconstruction Network.** After splitting the unsaturated image from the residual encoding we then use the residual to reconstruct highlights. Since the residual encoding was produced by convolving the highlights with our designed PSF, reconstructing the highlights becomes a deconvolution problem. Our network $f_{\mathbb{S}}$ thus takes in the residual prediction $\hat{\mathbf{I}}_r$ and outputs a prediction $\hat{\mathbf{I}}_{\mathbb{S}}$ of the highlights:

$$\hat{\mathbf{I}}_{\mathbb{S}} = f_{\mathbb{S}}\left(\hat{\mathbf{I}}_r\right). \tag{5.4}$$

We rely on a variation of the U-Net architecture [95] to deal with this deconvolution task. Specifically, our U-Net has five scales with four consecutive downsamplings (maxpools) and four consecutive upsamplings (nearest neighbor upsampling following by a $3 \times 3$ convolution layer). Each layer uses a $3 \times 3$ kernel window except for the first layer with $7 \times 7$ and the last layer with $1 \times 1$. Since the coded information is in the range $[0, 1]$ while the values to reconstruct are in the range $[1, 2^8]$, we avoid using normalization in the last two convolution layers to allow the network to learn a large range of output values. Similar to the splitting network, the loss $\mathcal{L}_{\mathbb{S}}$ encourages high-fidelity highlight reconstructions.

**Fusion Network.** In order to avoid boundary artifacts caused by combining $\hat{\mathbf{I}}_{\mathbb{U}}$ and $\hat{\mathbf{I}}_{\mathbb{S}}$ into a single image, we adopt a light-weight fusion network $f_{\mathbb{F}}$ to combine them and create the final predicted HDR image $\hat{\mathbf{I}}_{\mathbb{F}}$:

$$\hat{\mathbf{I}}_{\mathbb{F}} = f_{\mathbb{F}}\left(\hat{\mathbf{I}}_{\mathbb{U}}, \hat{\mathbf{I}}_{\mathbb{S}}\right). \tag{5.5}$$

The fusion network applies two $3 \times 3$ convolution layers with 64 feature channels to $\hat{\mathbf{I}}_{\mathbb{U}}$ and $\hat{\mathbf{I}}_{\mathbb{S}}$ separately, concatenates the feature maps together, and then applies two $3 \times 3$ convolution layers with 32 and 3 feature channels to produce the final predicted output $\hat{\mathbf{I}}_{\mathbb{F}}$.

### 5.4.1 Loss Functions

Our loss functions consist of two intermediary losses $\mathcal{L}_{\mathbb{U}}$ and $\mathcal{L}_{\mathbb{S}}$ which we apply to the intermediate outputs of the residual splitting network and the highlight reconstruction network respectively. We also apply a final loss $\mathcal{L}_{\mathbb{F}}$ to the final output of our network. The total loss that we minimize when training our network is given by

$$\mathcal{L}_{\text{Total}} = \mathcal{L}_{\mathbb{U}} + \mathcal{L}_{\mathbb{S}} + \mathcal{L}_{\mathbb{F}}. \tag{5.6}$$

**Loss on Unsaturated Regions** We adopt a perceptual loss as a metric of difference between the intermediate unsaturated image prediction $\hat{\mathbf{I}}_{\mathbb{U}}$ and the ground truth unsaturated image $\mathbf{I}_{\mathbb{U}}$. Our perceptual loss is defined using the pre-trained VGG-19 network and is given by

$$\mathcal{L}_{\text{VGG}}(\hat{x}, x) = \sum_{l} \nu_l \|\phi_l(\hat{x}) - \phi_l(x)\|_1, \tag{5.7}$$

where $\{\nu_l\}$ are loss balancing weights and $\phi_l$ are the feature maps from the $l$-th layer of pre-trained VGG-19. Specifically, we use the "conv2_2", "conv3_2", and

"conv4_2" layers of the VGG-19 network.

To better separate the unsaturated region prediction from the residual prediction, we apply an exclusion loss [26] $\mathcal{L}_{\text{excl}}$ during network fine-tuning. We assume that the edges of the unsaturated image and the edges of the residual encoding are unlikely to overlap, and we apply this assumption through an exclusion loss that penalizes correlation between the predicted unsaturated image and the residual in the gradient domain. The exclusion loss is defined to be

$$\mathcal{L}_{\text{excl}} = \|\tanh(\eta_{\mathbb{U}} \mid \nabla\hat{\mathbf{I}}_{\mathbb{U}} \mid) \odot \tanh(\eta_r \mid \nabla\hat{\mathbf{I}}_r \mid)\|_F, \tag{5.8}$$

where $\eta_{\mathbb{U}} = \sqrt{\|\hat{\mathbf{I}}_r\|_F / \|\hat{\mathbf{I}}_{\mathbb{U}}\|_F}$ and $\eta_r = \sqrt{\|\hat{\mathbf{I}}_{\mathbb{U}}\|_F / \|\hat{\mathbf{I}}_r\|_F}$ represent normalization factors, and $\|\cdot\|_F$ represents the Frobenius norm.

In conclusion, the loss that is applied to the intermediate output of the residual splitting network is

$$\mathcal{L}_{\mathbb{U}} = \alpha_1 \mathcal{L}_{\text{VGG}}(\hat{\mathbf{I}}_{\mathbb{U}}, \mathbf{I}_{\mathbb{U}}) + \alpha_2 \mathcal{L}_{\text{excl}}(\hat{\mathbf{I}}_{\mathbb{U}}, \hat{\mathbf{I}}_r). \tag{5.9}$$

**Loss on Saturated Regions** To extract information and perform deconvolution from the residual artifacts we use the same VGG loss given in Eq 5.7 for the intermediate prediction of the saturated highlights:

$$\mathcal{L}_{\mathbb{S}} = \beta \mathcal{L}_{\text{VGG}}(\hat{\mathbf{I}}_{\mathbb{S}}, \mathbf{I}_{\mathbb{S}}). \tag{5.10}$$

**Loss on Fused Output** We applied a Huber loss with $\gamma = 1/2$ to the ground truth HDR image $\mathbf{I}$ and final network prediction $\hat{\mathbf{I}}_{\mathbb{F}}$:

$$\mathcal{L}_{\mathbb{F}} = \mathcal{L}_{\text{Huber}}\left((\hat{\mathbf{I}} + \epsilon)^{\gamma}, (\mathbf{I} + \epsilon)^{\gamma}\right). \tag{5.11}$$

**Regularization** As described above we apply a regularization loss during training. This loss is applied by using the energy distribution mask shown in Figure 5.3c and keeping 94% of the energy in the center and 6% in the line-like satellite regions. Our regularizer is formally given by

$$\mathcal{L}_{\text{reg}} = \tau_{\mathbf{c}}|0.94 - \mathbf{p} \odot \mathbf{M_c}| + \tau_{\mathbf{s}}|0.06 - \mathbf{p} \odot \mathbf{M_s}| \tag{5.12}$$

where $\mathbf{p}$ is the PSF, $\mathbf{M_c}$ is the energy mask corresponding to the center, and $\mathbf{M_s}$ is the energy mask corresponding to the satellite regions. In our experiments we used $\tau_{\mathbf{c}} = 0.05$ and $\tau_{\mathbf{s}} = 0.1$ .

We also performed an experiment where we trained using our optical model but without regularization. We found that the final PSF converges to a Dirac point instead of spreading out energy from the saturated area and the performance is only 36.9 dB PSNR and 61.45 points on HDR-VDP 2 [224] on the test set. This experiment illustrates the importance of our regularizer.

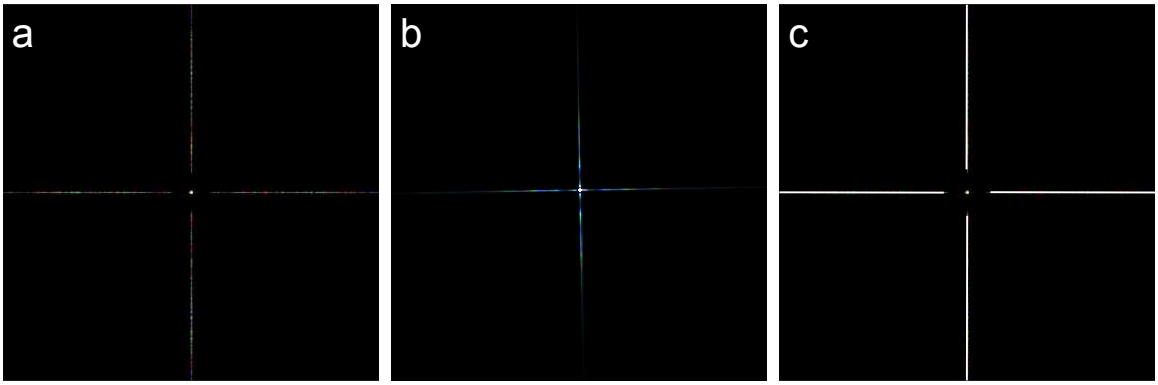


Figure 5.3: PSF corresponding to Rank-1 height map parameterization. (a) Simulated PSF. (b) PSF captured in the real-world. (c) Energy distribution mask used for regularization.

**Continued optimization without rank-1 height map factorization and without regularization** To validate that we achieve a good local optimum with our optical design we continued to train without our rank-1 factorization and without

our regularizer for 25 epochs. That is, we take our learned height map profile and continue to train as an unconstrained height map where each location is a learnable parameter. For this experiment the starting learning rate of the optical model is lowered from 1e−3 to 1e−6 while all other hyperparameters are the same as the original model training process. We observe that after 25 epochs the height map has changed insignificantly, as illustrated in Figure 5.4. This suggests that we do indeed find a good local optimum.

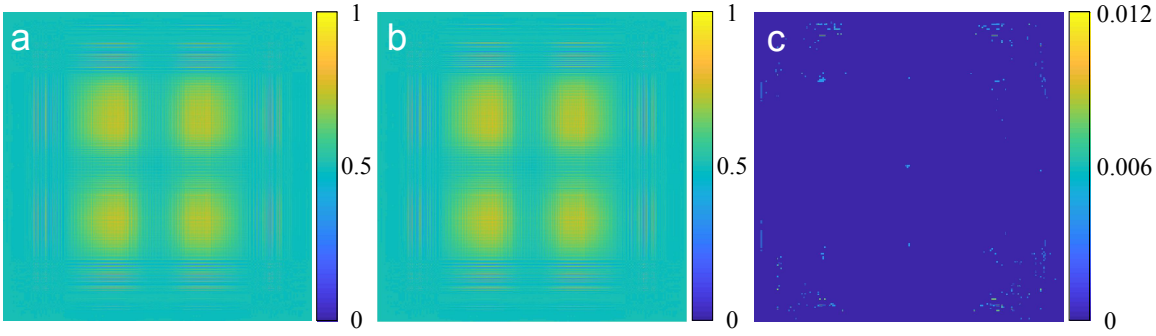


Figure 5.4: Height map comparison. (a) Our learned height map. (b) Continued training height map without rank-1 factorization and without regularization. (c) Absolute difference between (a) and (b). All figures are normalized by the maximum fabrication height $h_{\max} = 1.125\ \mu$m

### 5.4.2 Implementation and Training

We implement our rank-1 DOE height map model and reconstruction network in TensorFlow. Our network assumes inputs are in the range $[0, 1]$, and outputs are in the range $[0, 2^8]$. The model is jointly optimized using the Adam optimizer with polynomial learning rate decay. For more details, please refer to the Supplemental Material.

### 5.4.3 Details for reconstruction network

**Details for residual splitting network** Our residual splitting network configuration is shown in Table 5.1. As described in Section 4 of the chapter, we use the

| Layer | Convolution layer | Activation | Normalization |
|---|---|---|---|
| 0 | conv-n64-k1-d1 | Leaky Relu | nm |
| 1 | conv-n64-k3-d1 | Leaky Relu | nm |
| 2 | conv-n64-k3-d2 | Leaky Relu | nm |
| 3 | conv-n64-k3-d4 | Leaky Relu | nm |
| 4 | conv-n64-k3-d8 | Leaky Relu | nm |
| 5 | conv-n64-k3-d16 | Leaky Relu | nm |
| 6 | conv-n64-k3-d32 | Leaky Relu | nm |
| 7 | conv-n64-k3-d64 | Leaky Relu | nm |
| 8 | conv-n64-k3-d1 | Leaky Relu | nm |
| 9 | conv-n6-k1-d1 | - | - |

Table 5.1: Configuration of residual splitting network. In the table, "conv-n($a$)-k($b$)-d($c$)" represents a convolution layer with $a$ output channels, using a $b \times b$ kernel, and using a dilation rate $c$. Each "Leaky Relu" has slope 0.2 and $\text{nm}(x) = w_0 x + w_1 \text{Instance\_norm}(x)$, where $w_0$ and $w_1$ are trainable variables.

pre-trained VGG-19 model to extract feature maps (1472 channels in total) and upsample them to the same size as the input image (3 channels). Then we concatenate them together (1475 channels) and feed into our residual splitting network. We use a skip connection so that the output unsaturated image estimate $\hat{\mathbf{I}}_{\mathbb{U}}$ is given by the sum of the first three channels of the output of our residual splitting network and the input image $\mathbf{I}_s$. The last three channels of the output of our residual splitting network gives the encoded residual estimate $\hat{\mathbf{I}}_r$. We clip $\hat{\mathbf{I}}_{\mathbb{U}}$ to $[0, 1]$ and $\hat{\mathbf{I}}_r$ to $[0, 2^4]$.

**Details for highlight reconstruction network** Our highlight reconstruction network configuration is shown in Table 5.2. We avoid using normalization in the last two-layers '9_1' and '9_2', and the last layer '10' to allow the network to output a larger range of values. The output of the network $\hat{\mathbf{I}}_{\mathbb{S}}$ is clipped to $[1, 2^8]$.

**Details for fusion network** As shown in Table 5.3, we first convolve inputs $\hat{\mathbf{I}}_{\mathbb{U}}$ and $\hat{\mathbf{I}}_{\mathbb{S}}$ separately using two convolution layers for each input. Then we concatenate the feature maps together and generate the final HDR prediction using another two convolution layers. In addition, we mask out the unsaturated region for $\hat{\mathbf{I}}_{\mathbb{S}}$ before it

is sent to the fusion network.

### 5.4.4 Dataset

To ensure that our model accurately reconstructs highlights, we gathered HDR images that contain large dynamic ranges with small saturated regions. These include a mix of urban and rural scenes at night as well as indoor scenes from 19 different sources, see Supplemental Material for a complete list of dataset sources. To accommodate different image sizes we manually took $512 \times 512$ crops of the images specifically located at where the saturated regions were. After preprocessing, we had a total of 2039 images for training and 36 images for testing.

As part of the sensor simulation, we saturate a random percentage of pixels during training. That is, we multiply images by a scale factor such that 1% to 3% of pixels are larger than 1. After the scaling, we clip extreme pixel values, any pixel values larger than $2^8$ are set to $2^8$. We also augment the images using random rotations and flips. During testing, we saturate exactly 1.5% of the pixels in all test images and again clip pixel values larger than $2^8$.

## 5.5 Evaluation and Comparisons

We evaluate our approach in simulation against recent state-of-the-art single-shot HDR methods [181, 180, 178]. For HDR-CNN we used their pre-trained model. For Rouf et al.'s glare filter method, we applied an 8-point star PSF to the image using their experimentally obtained glare filter. For Deep Optics, we used the authors' PSF and trained their network on our dataset. Table 5.4 displays quantitative results on the test set. PSNR is calculated in the linear domain with a maximum value of $2^8$. HDR-VDP Version 2.2.1 was used with default settings except for pixels per degree which was computed using 24 inch diagonal display size, $512 \times 512$ resolution, and 1 m viewing distance. We report the Quality Correlation score. Figure 5.5 shows

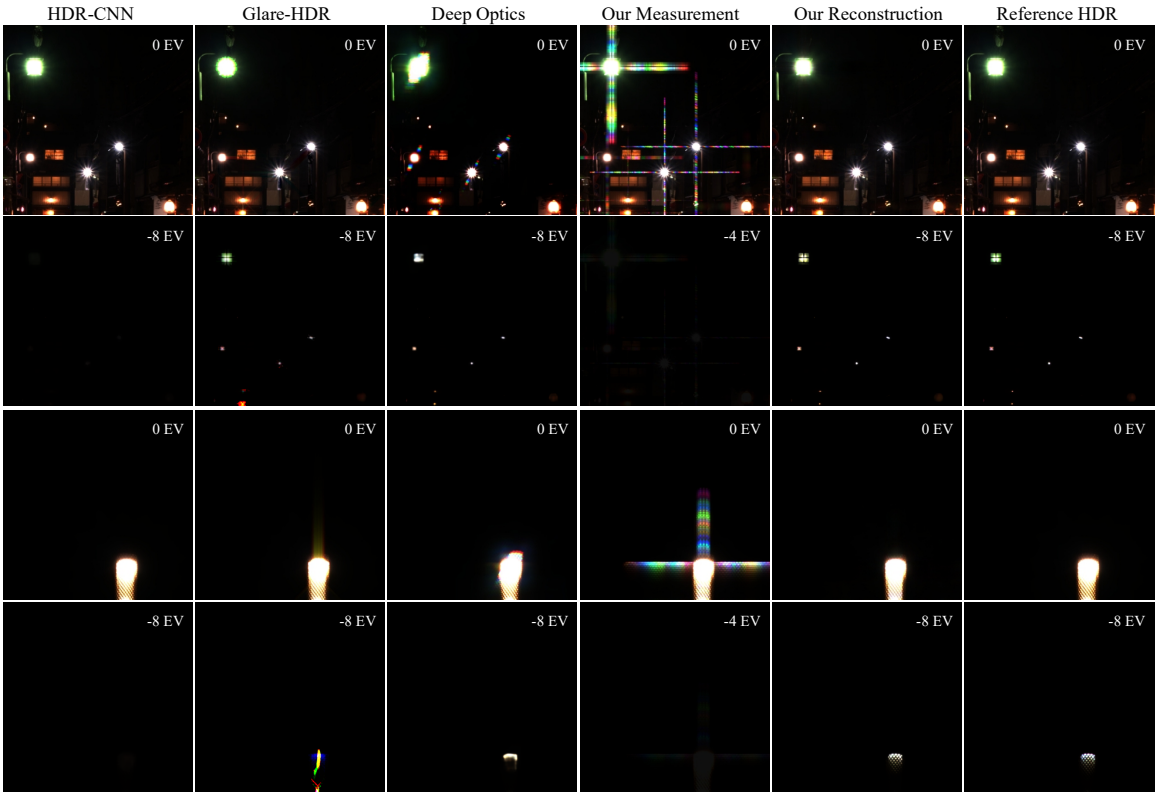


Figure 5.5: Visual comparison of different snapshot HDR methods in simulation. HDR-CNN [178] severely underestimates the intensity of saturated regions. Glare-HDR [180] often leaves artifacts and fails to estimate highlights correctly. The copied peaks for Deep Optics [181] sometimes overlap with the saturated areas and consequently are ineffective for highlight reconstruction. Please zoom in to view image details, such as the fine structures within the saturated areas.

qualitative comparisons of our approach against others.

### 5.5.1 Ablation Study

We performed an ablation study to illustrate the benefits of our proposed optical design and reconstruction network. Table 5.5 compares performance when using different reconstruction networks. We found that our network was best suited to HDR recovery with our learned PSF. Table 5.5 also shows performance when using different PSFs with our reconstruction network. For these experiments, only the network was optimized, and the PSF remains fixed. We observed that our PSF outperforms alternative PSF designs.

### 5.5.2 Scene depth experiments

Our optical model assumes that the point light source is placed 5 m away from the DOE plane. However, the PSF varies with different scene depth. As such, we investigate the robustness of our reconstruction network for handling PSFs corresponding to different scene depths in simulation. We change the position of the point light source from 1 m to infinity (while adjusting the distance between the sensor and focusing lens accordingly). Figure 5.6 shows our PSNR results on the test set in simulation. Our reconstruction network does best for 5 m depth as expected, and the performance is slightly degraded for other depths.

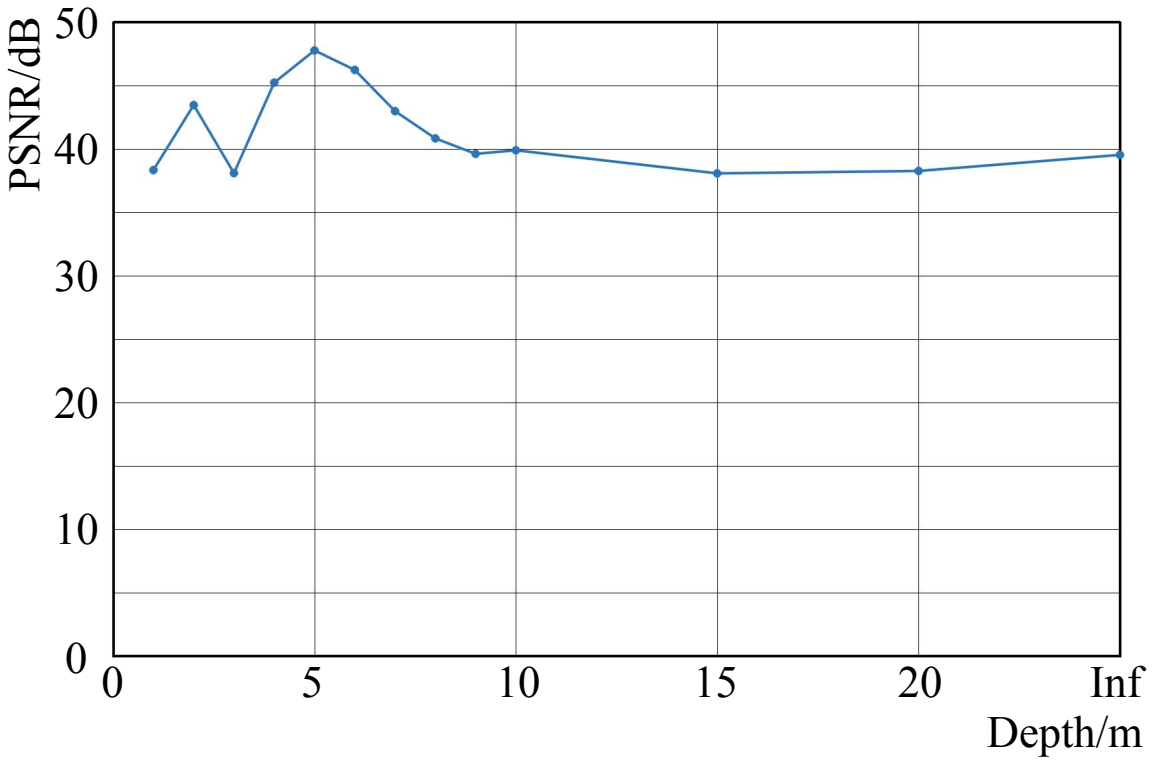


Figure 5.6: PSNR performance in simulation over different depths.

## 5.6 Experimental HDR Captures

We fabricate the optimized DOE using multilevel photolithography techniques [41, 43]. Due to fabrication limitations, we first slice the continuous phase mask into four layers with $2^4 = 16$ phase levels. This results in a high diffraction efficiency (theoretically more than 90%) [225]. By repeating the photolithography and reactive ion etching (RIE) for four iterations, we fabricated the phase mask on a 0.5 mm fused silica substrate with aperture size 9.16 mm and feature size 6 $\mu$m. Please refer to the Supplemental Material for further fabrication details.

Our imaging pipeline uses a Sony A7 with a pixel pitch of 5.97 $\mu$m, and the phase mask is closely placed in front of a Zeiss 50 mm f/1.4 lens (recall that we do not model the propagation between DOE and standard lens). Figure 5.7 shows that the real-world PSF matches the simulated PSF with slight contrast loss due to manufacturing imperfections and model approximations. Therefore, we perform a PSF calibration step where we capture the real-world PSF and then use it to fine-tune our reconstruction network. The real-world PSF is obtained by placing a white point light source 5 m away from the sensor, taking multi-exposure (five) snapshots at 3 EV intervals, and then fusing the snapshots in linear space using MATLAB's HDR toolbox.

### 5.6.1 Results

Figures 6.1 and 5.8 show real-world captures and reconstructions with our setup and reconstruction procedure. Reference images were taken by the same camera without the DOE (same aperture and position) using exposure fusion as described above for the PSF capture. In Figure 6.1, we correctly reconstruct highlights in the illuminated letters while removing most of the encoding streaks. In Figure 5.8 the left images show a brick wall where details are lost due to the light sources. Our method recovers these details, including color and structure. Note that our method succeeds despite

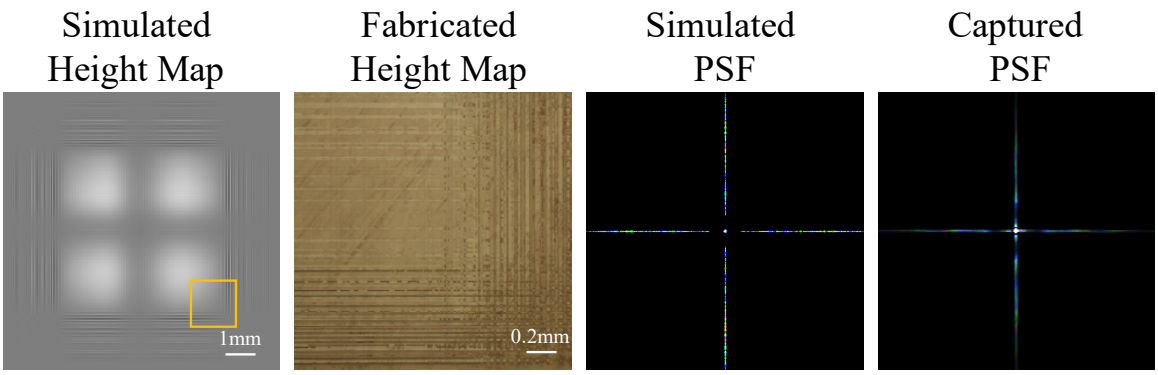


Figure 5.7: Comparison of simulation versus the real-world for the heightmap and PSF. Fine-tuning was performed with the captured PSF.

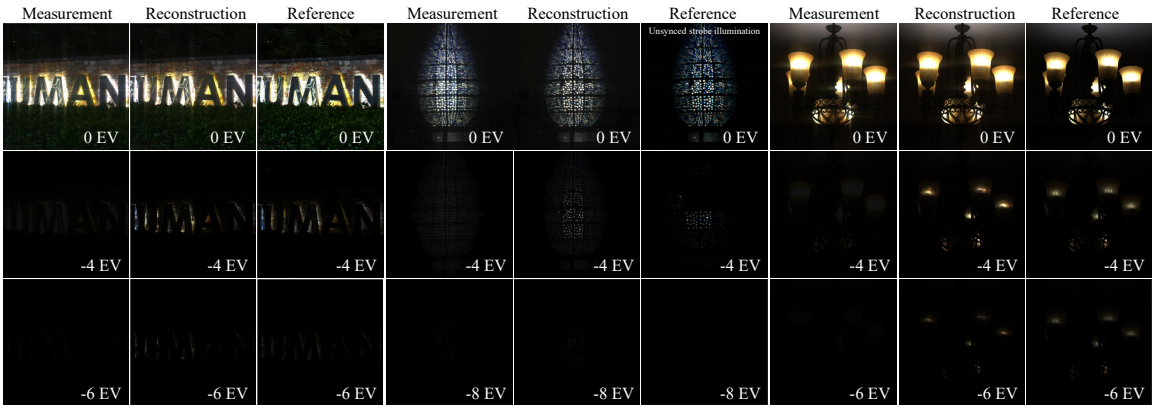


Figure 5.8: Real-world captures using fabricated DOE and reconstruction results. Note that the middle image is of a strobe light array with 50 Hz frequency. The reference images of $-4$ EV, $-6$ EV and $-8$ EV are taken by the same camera without the DOE (same aperture and position) by reducing the exposure time to $1/2^4$, $1/2^6$ and $1/2^8$ respectively. Please zoom in to view image details.

interference between the background image and our streak encodings. The middle images show that our method also works for dynamic scenes with of flashing strobe lights, which are challenging for burst HDR methods as the bursts would not be synchronized with the strobe. The right side shows that detailed reconstructions can be obtained for high-intensity lamp regions.

The presented reconstruction results and additional results in the Supplemental Material validate the proposed method for various scene types, including high-contrast night time urban environments and indoor settings. However, it is important to use

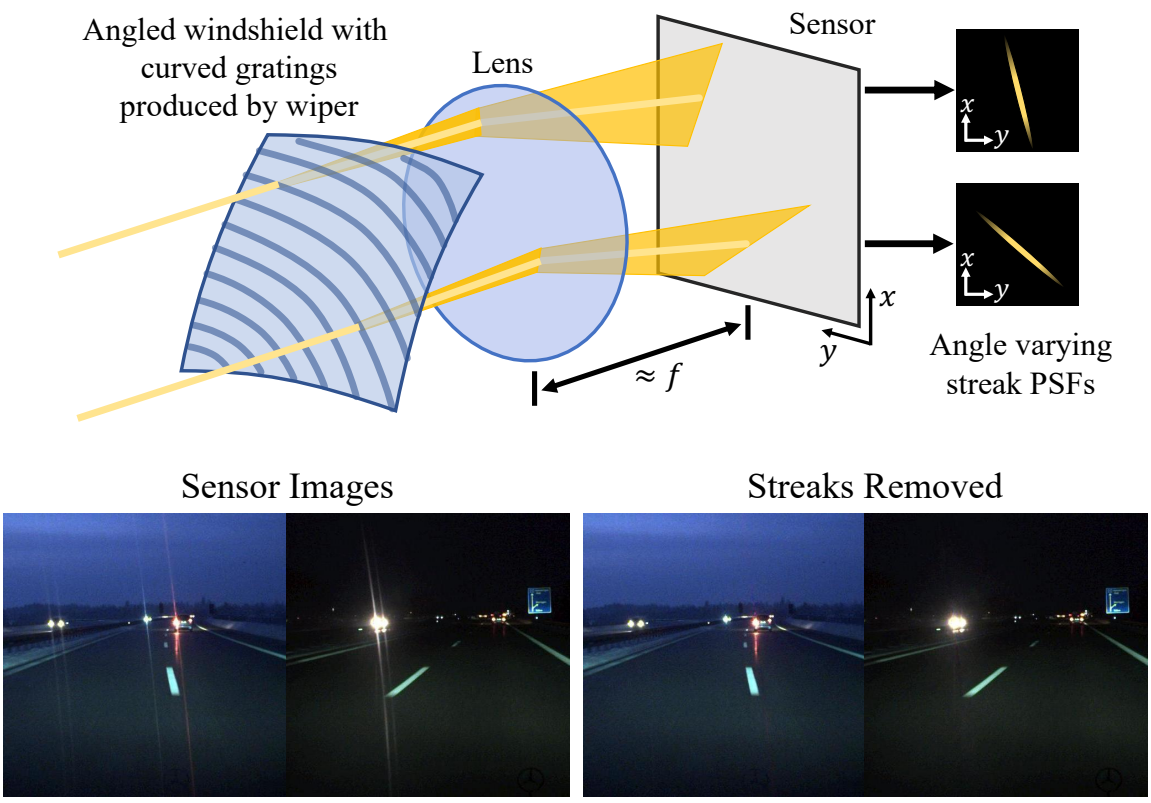


Figure 5.9: Automotive streaks are caused by grating-like patterns on the windshield. Applying our residual splitting network to the corresponding PSFs allows us to remove them.

a low exposure time as our method fails when the streaks are overexposed.

To evaluate real-time applicability we benchmarked the reconstruction latency. Our unoptimized network in TensorFlow takes 530 ms on an Nvidia Titan RTX to process a single LDR capture. After TensorRT optimization and network pruning our network takes 85 ms with fp32 precision and 44 ms with fp16 precision on the same GPU.

## 5.7 Grating Optics In-the-Wild

This section explores reconstruction without a designed optic, but with grating-like optics in the wild. As such, front-facing automotive cameras suffer from glare induced by thin lines of dust and dirt remaining on the windshield after wiping [226],

see Fig. 5.9. These thin streaks of dust are oriented perpendicular to the windshield wiping orientation on a typically curved windshield. As a result, they scatter light along streaks with varying orientation, which negatively impacts the imaging systems of autonomous vehicles during night time driving. Removing these streaks could improve performance for display applications, such as digital mirrors, as well as downstream computer vision tasks.

Although the PSFs corresponding to these streaks are different from our learned PSF, we can still apply our residual splitting network for removing these streaks. To demonstrate this, we collected several night time driving video sequences. We modeled the streak PSF in these videos using a 2-point star PSF, and we trained the residual splitting network using the same dataset from Section 5.4.4 and the unsaturated loss from Eq 5.9. To account for variations in the rotation angle of the 2-point star PSF, we uniformly sampled the rotation angle within $(-8, -2.5) \cup (2.5, 8)$ radians where 0 radians refers to the 2-point star PSF being parallel to a vertical line. Example snapshots along with removed glare results can be seen in Figure 5.9. For additional qualitative results, please refer to the Supplemental Material.

### 5.7.1 Automotive streak removal

We model streaks using a 2-point star PSF with the same parameterization from Rouf et al. [180]. We set $\alpha = 1.0$, $\beta = 0.00025$, $\gamma = 0$, $m = 0.014$ in order to closely approximate the streaks seen in the video sequence. To remove the streaks we train our residual splitting network with the unsaturated loss $\mathcal{L}_{\mathbb{U}}$ described in Section 4.1 and use $\hat{\mathbf{I}}_{\mathbb{U}}$ as the output. We do not use the highlight reconstruction network or the fusion network for this task. Figure 5.10 shows additional qualitative results for automotive streak removal.

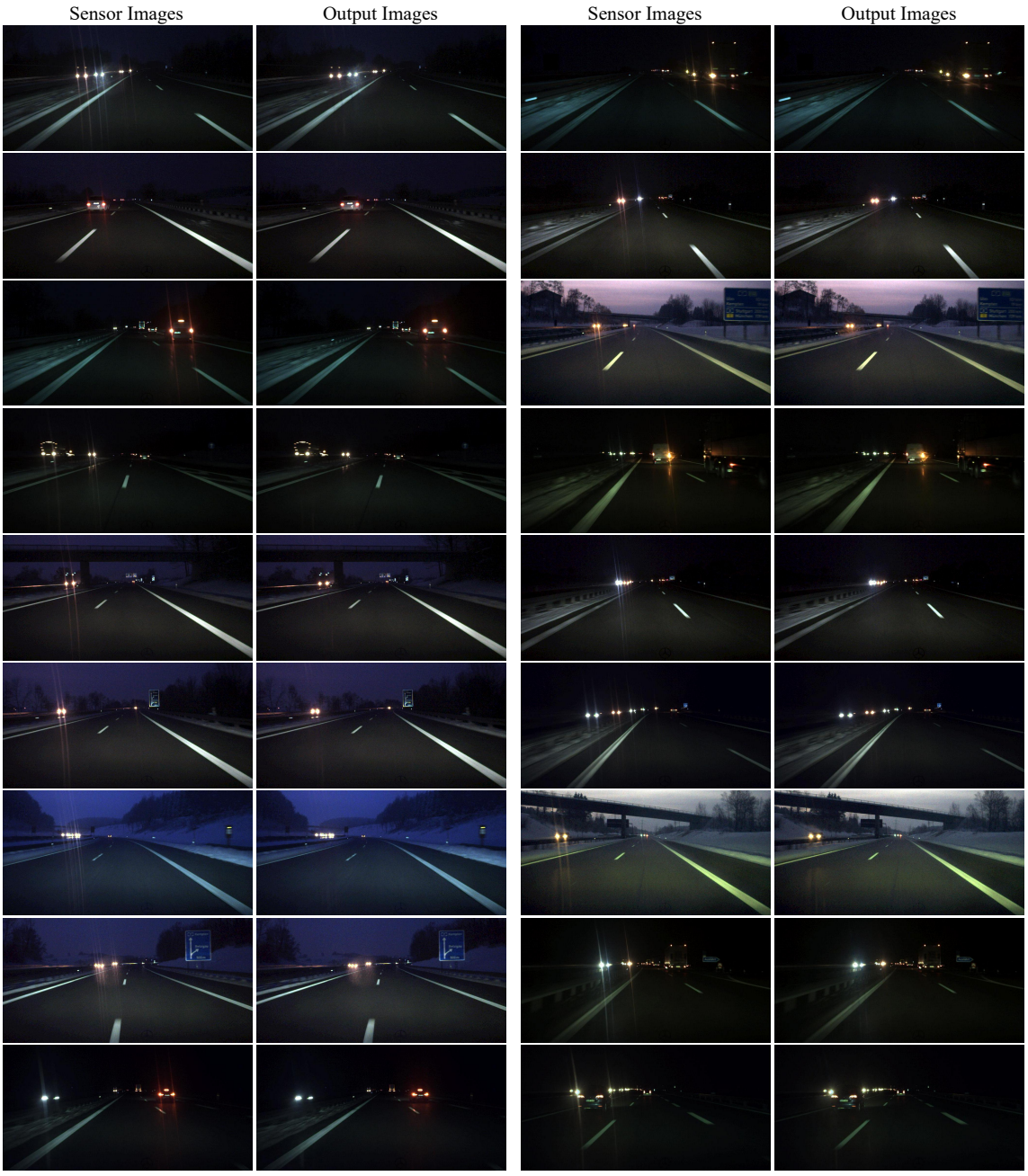


Figure 5.10: Additional qualitative results for automotive streak removal.

### 5.7.2 Automotive highlight reconstruction

Highlight reconstruction can also be performed with the automotive streaks. Figures 5.11 and 5.12 show highlight reconstruction results when training our full network on the same glare streaks described in Section 5.7.1.

## 5.8 Discussion

**Limitations** Like other optical encoding methods, our method requires that the encoding streaks themselves are not saturated. While we can ensure this for applications where small, saturated regions are expected (e.g night time driving and indoor navigation) our method does struggle with larger saturated regions. Please refer to the Supplemental Matrial for further discussion and failure examples.

State-of-the-art GPUs allow us to achieve real-time latencies, requiring multiple GPUs for high sensor resolutions, but are impractical for low-power consumer applications. Porting to dedicated hardware, such as power-efficient ASICs or FPGAs, is an important next step.

**Large saturated regions** Our method is most effective for recovering highly saturated, small area regions, but struggles like other optical encoding methods when the saturated regions are larger in area.

We performed simulation experiments on scenes with larger saturated regions to further illustrate the capabilities and limitations of our method, which can be seen in Fig. 5.13. The left images contain large saturated regions which causes some of the encoding streaks to be saturated. In spite of this, our method is still able to recover lost details and remove the encoding artifacts. The middle images show that our method is able to accurately detect and recover highlights of different intensities even if they all lie within the same saturated region. Specifically, the high intensity ceiling lights are correctly determined to be of higher intensity than the reflected light from

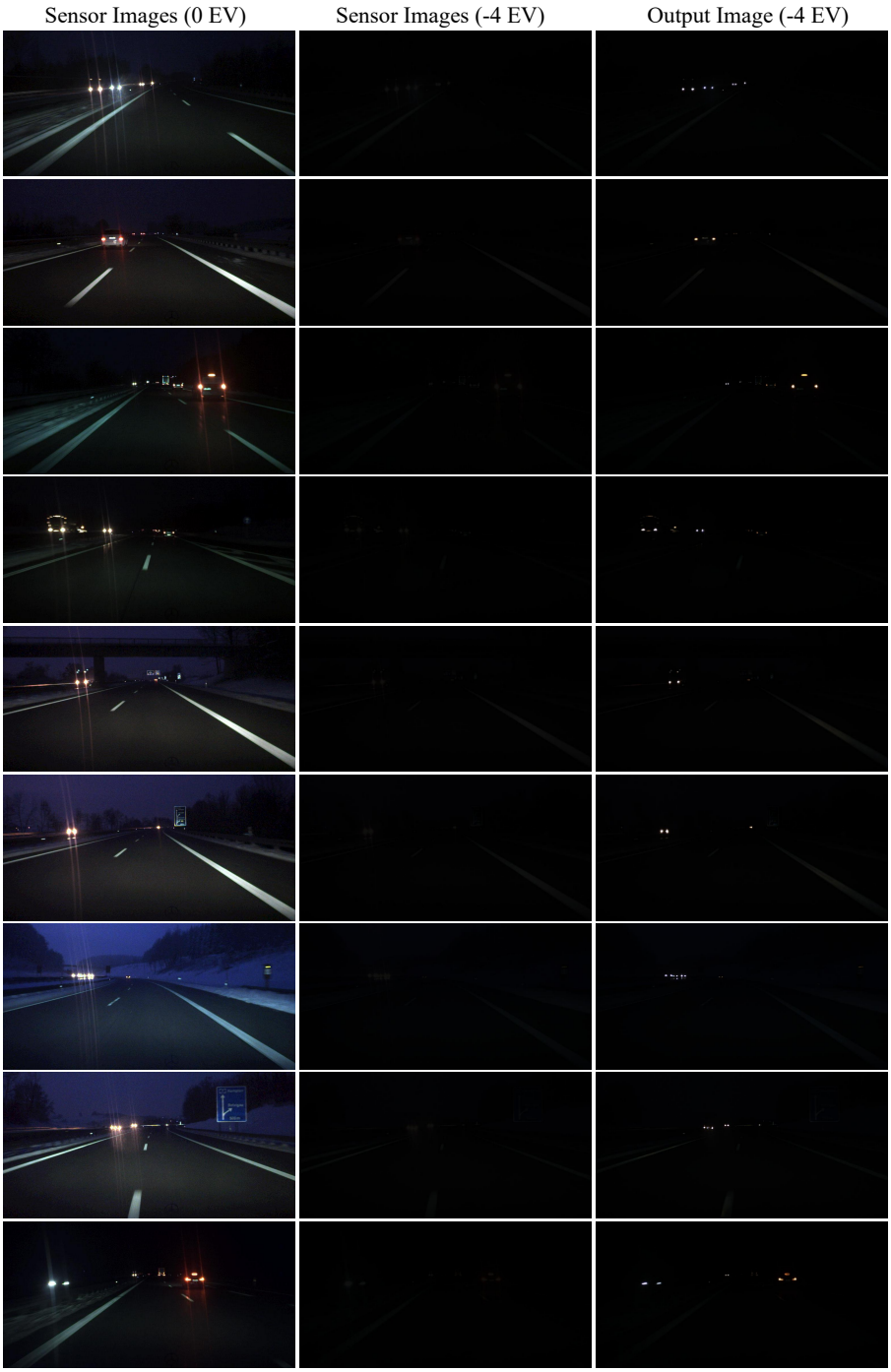


Figure 5.11: Qualitative results A for HDR imaging from windshield streaks.

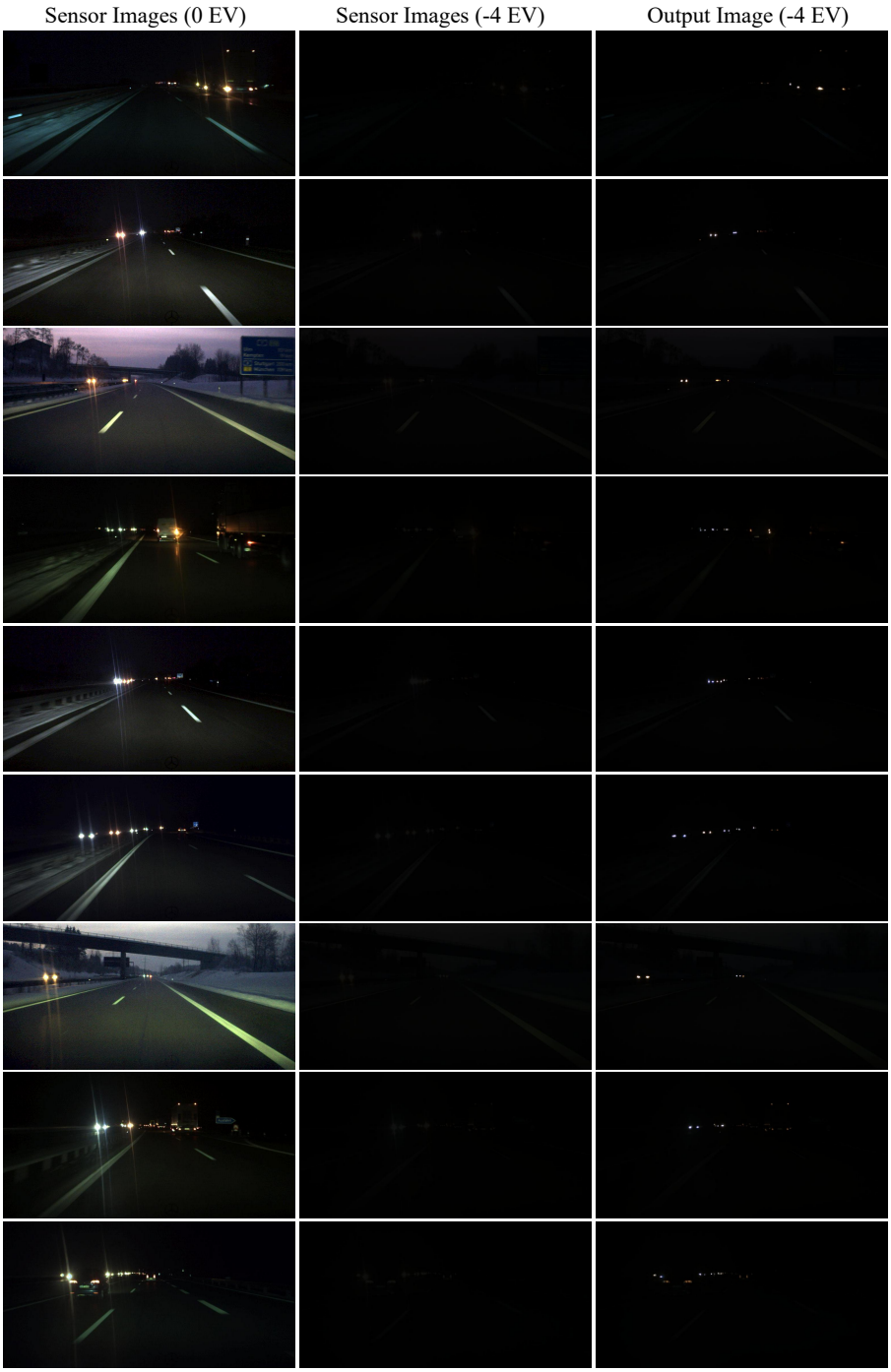


Figure 5.12: Qualitative results B for HDR imaging from windshield streaks.

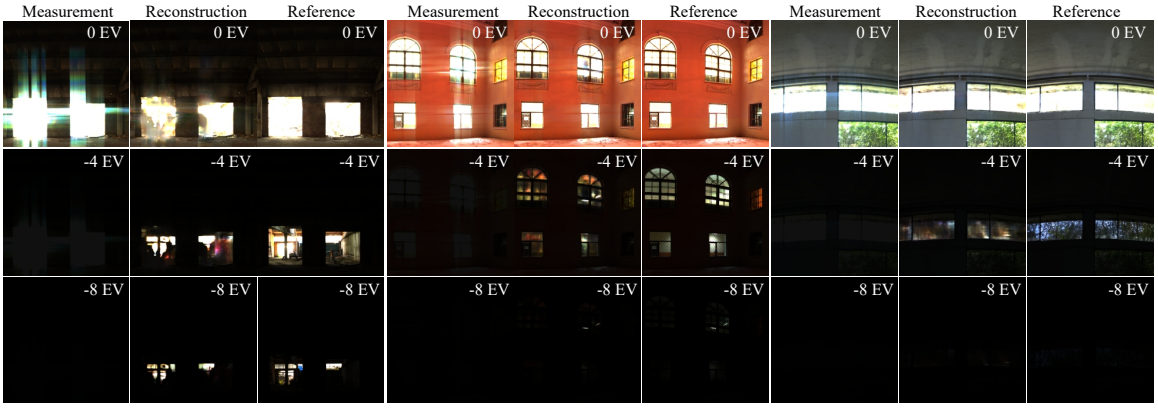


Figure 5.13: Additional simulation results on area with larger saturation regions. These larger saturated highlights are more difficult to recover, but we are still able to provide accurate reconstructions. Note that different highlight intensities within saturated regions are reconstructed with the correct intensity, for example, the ceiling lights in the middle image have higher intensity than light reflected from the windows.

the windows, even though both are saturated in the LDR measurement. The right image illustrates a failure mode consisting of a very complex scene within a large saturated region. Nevertheless, our method is still able to recover details and with correct intensity levels.

**Conclusion** We present a novel approach tackling the challenge of estimating HDR images from single-shot LDR captures. To this end, we propose a rank-1 DOE encoding of HDR content and a catered reconstruction network, which when jointly optimized allow for snapshot HDR captures that outperform previous state-of-the-art methods. Going forwards, we envision making snapshot HDR capture truly practical by extending our optical model to handle greater scene information, such as depth and multispectral data, as well as designing our algorithms for specialized hardware for low-power processing at the edge.

| Layer | Convolution layer | Activation | Normalization |
|---|---|---|---|
| 1_1 | conv-n32-k7-d1 | Leaky Relu | Instance |
| 1_2 | conv-n32-k3-d1 | Leaky Relu | Instance |
| | Max Pooling | | |
| 2_1 | conv-n64-k3-d1 | Leaky Relu | Instance |
| 2_2 | conv-n64-k3-d1 | Leaky Relu | Instance |
| | Max Pooling | | |
| 3_1 | conv-n128-k3-d1 | Leaky Relu | Instance |
| 3_2 | conv-n128-k3-d1 | Leaky Relu | Instance |
| | Max Pooling | | |
| 4_1 | conv-n256-k3-d1 | Leaky Relu | Instance |
| 4_2 | conv-n256-k3-d1 | Leaky Relu | Instance |
| | Max Pooling | | |
| 5_1 | conv-n512-k3-d1 | Leaky Relu | Instance |
| 5_2 | conv-n512-k3-d1 | Leaky Relu | Instance |
| | Upsampling & Concat | | |
| 6_1 | conv-n256-k3-d1 | Leaky Relu | Instance |
| 6_2 | conv-n256-k3-d1 | Leaky Relu | Instance |
| | Upsampling & Concat | | |
| 7_1 | conv-n128-k3-d1 | Leaky Relu | Instance |
| 7_2 | conv-n128-k3-d1 | Leaky Relu | Instance |
| | Upsampling & Concat | | |
| 8_1 | conv-n64-k3-d1 | Leaky Relu | Instance |
| 8_2 | conv-n64-k3-d1 | Leaky Relu | Instance |
| | Upsampling & Concat | | |
| 9_1 | conv-n32-k3-d1 | Leaky Relu | - |
| 9_2 | conv-n32-k3-d1 | Leaky Relu | - |
| 10 | conv-n3-k1-d1 | - | - |

Table 5.2: Configuration of highlight reconstruction network. In the table, "conv-n($a$)-k($b$)-d($c$)" represents a convolution layer with $a$ output channels, using a $b \times b$ kernel, and using a dilation rate $c$. Each "Leaky Relu" has slope 0.2 and "Max Pooling" represents a max pooling layer with a $2 \times 2$ kernel and a stride of 2. Each "Upsampling" represents nearest neighbor upsampling with a factor 2 followed by a convolution layer with a $3 \times 3$ kernel.

| Layer | Convolution layer | Activation | Normalization |
|---|---|---|---|
| $1\_1_{\mathbb{U}}$ | conv-n64-k3-d1 | Leaky Relu | - |
| $1\_2_{\mathbb{U}}$ | conv-n64-k3-d1 | Leaky Relu | - |
| $1\_1_{\mathbb{S}}$ | conv-n64-k3-d1 | Leaky Relu | - |
| $1\_2_{\mathbb{S}}$ | conv-n64-k3-d1 | Leaky Relu | - |
| Concat | | | |
| 2_1 | conv-n32-k3-d1 | Leaky Relu | - |
| 2_2 | conv-n3-k3-d1 | Leaky Relu | - |

Table 5.3: Configuration of fusion network. In the table, "conv-n($a$)-k($b$)-d($c$)" represents a convolution layer with $a$ output channels, using a $b \times b$ kernel, and using a dilation rate $c$. Each "Leaky Relu" has slope 0.2 and $\text{nm}(x) = w_0 x + w_1 \text{Instance\_norm}(x)$, where $w_0$ and $w_1$ are trainable variables. $1\_1_{\mathbb{U}}$, $1\_2_{\mathbb{U}}$ are applied to $\hat{\mathbf{I}}_{\mathbb{U}}$ while $1\_1_{\mathbb{S}}$, $1\_2_{\mathbb{S}}$ are applied to $\hat{\mathbf{I}}_{\mathbb{S}}$.

Table 5.4: Quantitative comparison across single-shot HDR methods.

| Methods | PSNR | HDR-VDP 2 |
|---|---|---|
| **Ours** | **48.26** | **74.47** |
| Deep Optics [181] | 40.30 | 67.96 |
| Glare-HDR [180] | 32.23 | 56.76 |
| HDR-CNN [178] | 34.06 | 54.34 |
| LDR | 33.57 | 52.43 |

Table 5.5: Ablation study with different PSFs and reconstruction networks.

| PSF | Network | PSNR | HDR-VDP 2 |
|---|---|---|---|
| **Ours** | **Ours** | **48.26** | **74.47** |
| Ours | Deep Optics [181] | 37.91 | 61.30 |
| Ours | HDR-CNN [178] | 33.51 | 52.66 |
| Dual Peak PSF [181] | Ours | 43.08 | 70.25 |
| Star PSF [180] | Ours | 42.62 | 68.03 |
| Dirac PSF | Ours | 37.25 | 63.45 |

# Chapter 6

## Differentiable Complex Lens Design: A General End-to-end Optics Design Pipeline

Our primary focus was on building a differentiable diffractive optics model that can be jointly optimized with post-processing in the previous chapters. We first built a differentiable optics model combined with the GS-based algorithm to realize an optimal PSF physically through DOE. Then we expand the differentiable diffractive optics model to a general fresnel propagation problem to optimized the PSF on the sensor. To solve the problem that it is easy for a vast variable map of DOE to be stuck in a local minimum, we first proposed a rank-1 factorization model to express the DOE layer with fewer parameters. As a result, we applied this model to snapshot HDR imaging and achieved state-of-the-art results.

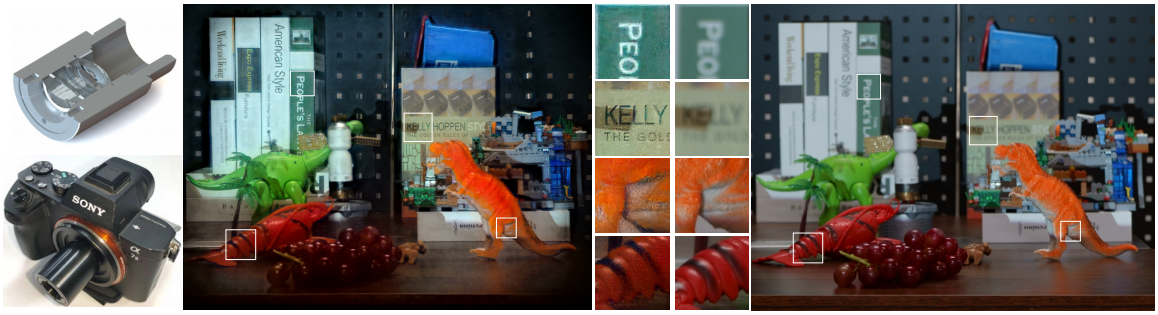


Figure 6.1: An exemplary triplet design for EDOF imaging optimized by our end-to-end differentiable complex lens design framework. Top left: 3D model of the optimized triplet lens design ($50mm/F4$). Bottom left: prototype fabricated by single-point diamond turning. Middle: final image that is captured by our end-to-end designed lenses and processed by our algorithm. Right: the same scene captured by a Sony $28-70mm$ zoom lens at $50mm/F4.5$. The objects are placed in the range from around $0.8m$ to $1.8m$ from the lenses. Our prototype succeeds in obtaining the all-in-focus image, while the conventional lens shows a narrow depth of field.

However, those differentiable diffractive optics models used in the previous chapters have been quite simplistic, built either on simple wave optics models such as Fourier transform or similar paraxial models. Such models only support the optimization of a single lens surface, which limits the achievable image quality.

To overcome these challenges, we propose a general end-to-end complex lens design framework enabled by a differentiable ray tracing image formation model. Specifically, our model relies on the differentiable ray tracing rendering engine to render optical images in the full field by taking into account all on/off-axis aberrations governed by the theory of geometric optics. Our design pipeline can jointly optimize the lens module and the image reconstruction network for a specific imaging task. We demonstrate the effectiveness of the proposed method on two typical applications, including large field-of-view imaging and EDOF imaging. Both simulation and experimental results show superior image quality compared with conventional lens designs. Our framework offers a competitive alternative for the design of modern imaging systems.

## 6.1 Introduction

Cameras are designed with a complicated tradeoff between image quality (e.g. sharpness, contrast, color fidelity), and practical considerations such as cost, form factor, and weight. High-quality imaging systems require a stack of multiple optical elements to combat aberrations of all kinds. At the heart of the design process are tools like ZEMAX and Code V, which rely on merit functions to trade off the shape of the PSF over different image regions, depth, or zoom settings. Such a design process requires significant user knowledge and experience, and the emphasis on PSF shaping neglects any subsequent image processing operations, specific application scenarios, or the desire to encode extra information in the image.

Therefore, domain-specific computational imaging has attracted researchers' attention in the past several decades. Enabling the co-design of optics with post-

processing, computational cameras have achieved impressive results in EDOF (EDOF) [1, 2, 3, 4], high dynamic range imaging (HDR) [5, 6, 7, 8], and image resolution [9, 10, 11]. Nevertheless, all those older methods are either heuristic or use some proxy metric on the point spread function (PSF) rather than considering the imaging quality after post-processing. Therefore, finding a joint optimal solution for both imaging optics and image reconstruction for a given task remains an unsolved problem in general.

Over the past few years, co-design of optics and image processing [12, 13], or even data-driven end-to-end design [14] have emerged to bridge the gap between optical design and algorithm development. Co-design of optics and post-processing algorithms has achieved a superior performance for domain specific tasks such as depth estimation [15], large field-of-view imaging [13], EDOF [16], optimal sampling [17], and high dynamic range (HDR) imaging [18, 19]. Unfortunately, the differentiable lens models used in these works have been too limited to describe complex optical assemblies, and have instead only allowed to optimize a single optical surface with a single material. This narrow design space limits the final image quality compared to commercial consumer-level or industrial-level cameras. Furthermore, existing models are based on the paraxial approximation and ignore off-axis aberrations, which degrades the quality for large field of view imaging.

Data-driven optimization of all the parameters in a complex lens assembly is challenging. On the one hand, the varying parameters of the optical surfaces cause scaling and distortion that change during the optimization process. On the other hand, a naive implementation will consume huge computational resources due to the differentiable ray tracing engine [20]. In this work, we overcome these challenges, and achieve the first end-to-end optimization system for complex lens assemblies. We propose a unique differentiable and configurable optical model that not only overcomes the limitation of a single optical surface and a single material, but also supports optimizing off-axis regions. In addition, we propose an end-to-end framework

configurable for a given task with a tailored recovery algorithm, loss function, and data. As a result, we are able to directly render the images with aberrations of all kinds. That means we can optimize the complex lens model while accounting for the continuous variation of the PSF across the image plane. Beyond the goal of capturing a sharp and clear image on the sensor, the proposed method offers huge design flexibility that can not only find a compromise between optics and post-processing, but also opens up the design space for optical encoding.

It must be stressed, however, that our approach does not completely eliminate the need for an experienced user. Specifically, since lens design is a highly non-convex problem, we can not initialize the parameter space randomly; instead, we initialize the system with a coarse design that has the desired number of lens elements, and is roughly focused along the optical axis. This optical system is then improved and adapted to a specific imaging scenario using end-to-end optimization. In this chapter we demonstrate both large field-of-view and large depth-of-field as the two application scenarios. The proposed approach outperforms the state-of-the-art complex lens design (by ZEMAX) in both simulation and experiments. We prototype our designs with complex lenses manufactured by a CNC machining system that supports point diamond turning. The experiments are carried out in-the-wild as conventional cameras. Our results show significantly improved performance over conventional designs on the above-mentioned applications.

Specifically, we make the following contributions:

- We introduce a novel configurable and differentiable complex lens model based on differentiable ray-tracing, and this model can simulate aberrations of all kinds. We allow users to easily define the initial optics design, including lens surface profile, positions, and materials.

- Our differentiable complex lens model is the first in end-to-end design to consider off-axis performance, and offers a greater design freedom compared to

existing end-to-end optics models.

- We propose an end-to-end pipeline that can jointly optimize the lens model and the recovery network. The reconstruction network and loss functions can be tailored to a given computational imaging task.

- We successfully apply our model and pipeline to large field-of-view imaging and EDOF imaging using designs that are compact and low-budget, but high-quality. We validate them in both simulations and on real-world measurements captured by our assembled and fabricated aspherical lens group and verify that the experimental results agree with the simulations.

### 6.1.1 Optical Aberrations and Traditional Lens Design.

The most common monochromatic aberrations are defocusing, spherical aberration, coma, astigmatism, field curvature, and distortion, while the chromatic aberrations are typically axial and lateral chromatic aberration. Both types of aberrations are the result of the differences in the optical path length when light travels through different regions of a lens at different incident angles [44]. These aberrations manifest themselves as an unwanted blur, which becomes more severe with increasing depth of field (DOF), numerical aperture, and FOV [45].

Conventional lens design is a semi-automated process, in which a rough initial design is chosen by an experienced designer, and then optimized with software like CODE V and ZEMAX. These typically use either the Levenberg-Marquardt algorithm or damped least squares (DLS) to optimize the optical system including spherical and aspherical lenses, hybrid optical elements [227, 228], and lens elements with different material properties. These tools are the cornerstone of lens design and rely on existing aberrations objectives, so-called merit functions, to find a compromise across a variety of criteria [34, 46], trading off the PSFs across sensor locations, lens configurations

(e.g., zoom levels), and target wavelength band.

However, critically the established merit functions only operate on the PSFs, trading off their footprint over different configurations. This approach is agnostic to any intended usage case or image reconstruction approach. As a result, it is hard to co-design the optics and post-processing together for domain-specific cameras [14] since they can not use the final imaging performance criteria as an optimization object. Thinking beyond the traditional complex lens design for a given task, we seek to investigate a differentiable complex lens model and end-to-end optimization framework to bring the complex lens design into an end-to-end era.

### 6.1.2 Computational Optics.

Many works on computational imaging [229, 230, 47, 48] have proposed designing optics for aberration removal in post-processing. These methods often favor DOEs [231, 49, 50, 43], or even metasurfaces [232] over refractive optics because of their large design space or ultra thin form factor [233]. To simplify the inverse problem in post-processing, all of the described approaches ignore off-axis aberrations by restricting the FOV to a few degrees – existing methods do not realize monocular and chromatic imaging with a large FOV. The state-of-the-art joint designing of optics and post-processing [13] firstly enables a large FOV imaging with a single lens. However, their model is still to design the optics and image processing algorithm separately to include the FOV in the design process. Moreover, they need a complicated and time-consuming dataset acquisition from the monitor.

In addition to minimizing optical aberrations optics, computational imaging also aims to improve the basic capabilities of a camera by including optical coding, such as DOF [1, 2, 3, 4], dynamic range [5, 6, 7, 8] and image resolution [9, 10, 11].

Our proposed end-to-end complex lens design framework could be applied to many of these applications. It introduces a general design paradigm for computational

cameras that optimizes directly for the post-processed output with respect to a chosen quality metric and domain-specific dataset.

### 6.1.3 End-to-end Optics Design.

Co-designing of optics and post-processing has demonstrated superior performance over traditional heuristic approaches in single-lens color imaging [217, 13], HDR imaging [18, 19], single image depth estimation [15, 25, 26, 55, 234, 235, 236, 237], microscopy imaging [218, 219, 220, 221].

In computer vision, the emergence of deep learning has led to rapid progress in several challenging tasks and the state-of-the-art results for well-established problems [76, 77, 78]. For example, a deep approach for deconvolution by including a fully connected convolutional network [79] has been proposed. Generative adversarial networks (GANs) are shown to provide generative estimates with high image quality. Kupyn *et al.* [80, 238] demonstrated the practicability of applying GAN reconstruction methods to deblurring problems. Those approaches have been demonstrated to obtain state-of-the-art results in many computational photography tasks but not take one step further to optimize the optics together. G. Côté *et al.* [239, 240] utilize deep learning to get lens design databases to produce high-quality starting points from various optical specifications. However, they generally focused on the design of starting points, and the designing space is limited to spherical surfaces.

The deep optics [14] approach involves joint design of optics and image recovery for a specific task in an end-to-end fashion. Based on this model, a series of applications have been investigated in the last two years like hyperspectral imaging [241, 242], high dynamic range imaging [19], full-spectrum imaging [231] and depth estimation [16]. These works are inherently limited to designing only a single optical surface, and therefore the image quality of their final designs does not reach the level of regular consumer camera optics. A solution to this problem has been to utilize a commercial

lens but add a single, co-designed element for a specific purpose. This approach has been applied to super-resolution SPAD cameras [17] and high dynamic range imaging [18]. However, none of these approaches can deal with large FOVs as their image formation model relies on simple paraxial approximation in addition to the single-surface restriction. Co-designing complex optics with the image reconstruction was not addressed until the work on learned large FOV imaging [13]. They overcame the limitation of FOV by separating the optical design and image processing but not in an end-to-end fashion.

In conclusion, existing end-to-end methods work in a very restricted setting, including only a single optical surface and a simple paraxial image formation model (small FOV), or rely on existing optical design tools. They also have in common that they require either accurate PSF calibration or extensive training data. We propose a general configurable and differentiable complex lens model and an end-to-end framework with tailored recovery networks for different tasks. Drawing inspiration from the state-of-the-art differentiable rendering technique [20, 243, 244, 245], our complex lens model offers a great design freedom where the number of elements, lens surface profiles and positions can be configurable. The ample design space of our proposed lens model allows for rich optical encodings and the end-to-end pipeline achieves optimal synergy with the image reconstruction algorithm. Our complex lens model can optimize the lens parameters and simulate all kinds of aberrations without considering spatial and depth varying PSF. This property makes it easier for the later reconstruction network retraining and fine-tuning stage to get a highly accurate simulated dataset. Finally, our solution overcomes the limitation for large FOV and makes it possible to design a high-quality consumer-level lens in an end-to-end manner.

## 6.2 End-to-end Optimization of Complex Lens and Image Recovery

### 6.2.1 Image Formation Model

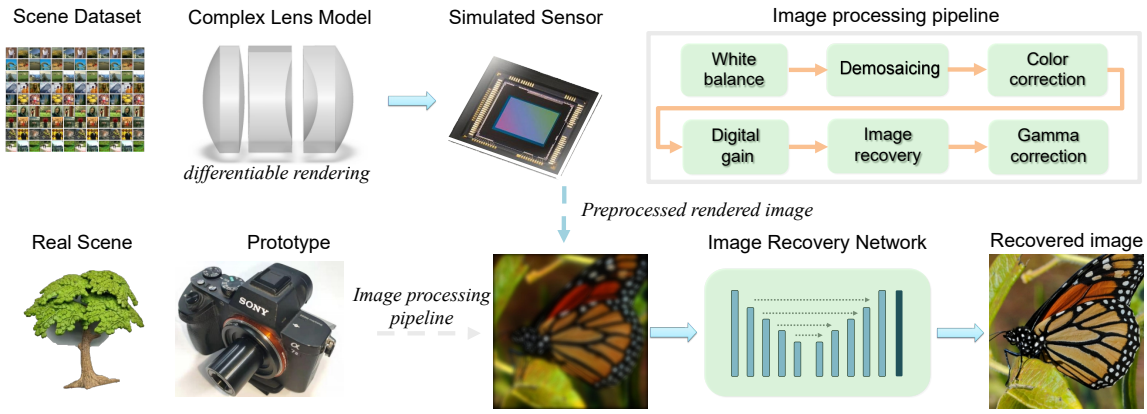


Figure 6.2: Framework for end-to-end designing of differentiable complex lens model and reconstruction. In each forward pass, we set up one scene from a certain point-of-view and render the simulated sensor image through the differentiable complex lens model. Then, the simulated images are sent to the image reconstruction network and we train the whole framework simultaneously. For the experimental stage, we directly send the preprocessed real-captures to the pre-trained network. Notice that the scene setup, initial lens design and image recovery network can be tailored to specific applications.

Our end-to-end framework consists of an optical simulation stage with the lens model and a trimmed recovery network as a reconstruction stage to achieve the best results by employing a generative adversarial network (GAN). As in most existing complex lens systems, the refraction is usually generated by either spherical or aspherical surfaces. Throughout the rest of this chapter, we consider rotationally symmetric lens profile designs, which can be manufactured using diamond turning machines.

Note, however, that our lens model could be easily applied to rotationally asymmetric profiles such as Zernike basis functions.

## Differentiable Lens Model and Ray Tracer

We implement our differentiable lens tracer following [246], based on Mitsuba2 [20]. The framework is fully differentiable, including the configurable and differentiable complex lens model and the image recovery network. Each part of this pipeline can

be easily configured to tailor for specific tasks.

After training, we take the optimized parameters like radius, conic coefficients and high order coefficients of the lens profiles to fabricate the lens. To account for better reconstruction with the image processing pipeline and the recovery network, it can be tailored and fine-tuned through re-training after the lens parameters are fixed.

In the following, we show how to efficiently integrate the differentiable ray-tracing into our lens designing pipeline.

**Aspherical Lenses** Our lens model is based on a standard representation of an aspherical lens as a spherical component with a polynomial correction factor. Given a Cartesian coordinate system $(x, y, z)$, the $z$-axis coincides with the optical axis, while $(x, y)$ forms the transverse plane. Let $r = \sqrt{x^2 + y^2}$ and $\rho = r^2$. Then the height of the aspheric surface and its derivative is defined as:

$$h(\rho) = \frac{c\rho}{1 + \sqrt{1 - \alpha\rho}} + \sum_{i=2}^{n} a_{2i}\rho^i, \tag{6.1}$$

$$h'(\rho) = c\frac{1 + \sqrt{1 - \alpha\rho} - \alpha\rho/2}{\sqrt{1 - \alpha\rho}\left(1 + \sqrt{1 - \alpha\rho}\right)^2} + \sum_{i=2}^{n} a_{2i} i \rho^{i-1}, \tag{6.2}$$

where $c$ is the curvature, $\alpha = (1 + \kappa)c^2$ with $\kappa$ being the conic coefficient, and $a_{2i}$'s are higher-order coefficients. The implicit form $f(x, y, z)$ and its spatial derivatives $\nabla f$ are:

$$f(x, y, z) = h(\rho) - z, \tag{6.3}$$

$$\nabla f = \big(2h'(\rho)x, 2h'(\rho)y, -1\big). \tag{6.4}$$

Note that spherical surfaces are special cases of aspheric surfaces when $\kappa = 0$ and $a_{2i} = 0\,(i = 2, \cdots, n)$.

In the following, we derive a differentiable ray-tracing based image formation

model which simulates all kinds of aberration at the same time. For each surface in the lens, its profile is directly described by (6.1) and the lens materials are predefined according to the prior knowledge of optical design to cancel the chromatic aberrations.

**Ray-surface Intersection by Newton's Method** To use the above lens model in a ray-tracer, we need to be able to compute the intersection point $(x, y, z)$ and ray marching distance $t$ for intersecting surface $f(x, y, z) = 0$ (implicit form), given a ray $(\mathbf{o}, \mathbf{d})$ of origin $\mathbf{o} = (o_x, o_y, o_z)$ and direction $\mathbf{d} = (d_x, d_y, d_z)$ of unit length (i.e. $\|\mathbf{d}\| = 1$). Mathematically, this is a root finding problem, i.e. we need to determine $t > 0$ such that

$$f(x, y, z) = f(\mathbf{o} + t\mathbf{d}) = 0. \tag{6.5}$$

Since there is no analytical solution for this problem for the aspherical lens model, we solve the problem numerically using Newton's method. At iteration $k + 1$, we update $t^{(k+1)}$ from previous estimate $t^{(k)}$ as:

$$\begin{aligned} t^{(k+1)} &\leftarrow t^{(k)} - \frac{f(\mathbf{o} + t^{(k)}\mathbf{d})}{f'(\mathbf{o} + t^{(k)}\mathbf{d})} \\ &\leftarrow t^{(k)} - \frac{f(\mathbf{o} + t^{(k)}\mathbf{d})}{\nabla f \cdot \mathbf{d}}, \end{aligned} \tag{6.6}$$

where $f'$ and $\nabla f$ denote derivatives w.r.t. $t$ and $(x, y, z)$, respectively. A coarse (non-singular) initialization is $t^{(0)} = (z - o_z)/d_z$, and the iteration stops when the difference is smaller than tolerance.

**Dispersion by Cauchy's equation** To model dispersion, we extend Mitsuba2 by formulating the lens material refractive index using Cauchy's equation [247]:

$$n(\lambda) = A + \frac{B}{\lambda^2} + \frac{C}{\lambda^4} + \cdots . \tag{6.7}$$

In practice, we found it sufficient to use only the first two terms (with parameters $A$ and $B$) in the equation. When the central wavelength $n_D$ and Abbe numbers $V$ are given, $A$ and $B$ are computed as:

$$A = n_D - \frac{B}{\lambda_D^2} \qquad \text{and} \qquad B = \frac{n_D - 1}{V(\lambda_F^{-2} - \lambda_C^{-2})}, \tag{6.8}$$

where $\lambda_D = 589.3\,\text{nm}$, $\lambda_F = 486.1\,\text{nm}$, and $\lambda_C = 656.3\,\text{nm}$.

## Optics simulation

End-to-end computational imaging consists of simulated optics used to generate simulated image data with all aberrations present, as well as software reconstruction pipeline. For joint design, both of these modules should be fully differentiable so that gradient update become possible across both components.

With the optimal trade-off between the simulation stage and the reconstruction stage, the PSF usually varies within the field of view, and across scene depth and spectrum. For a given color channel $c$, the recorded sensor measurement $I_c$ can be expressed as:

$$I_c(x', y') = \int Q_c(\lambda) \cdot [p(x', y', d, \lambda) * s_c(x', y', d)] d\lambda + n(x', y'), \tag{6.9}$$

where the PSF $p(x', y', d, \lambda)$ is a function with spatial position $(x', y')$ on the sensor, the depth $d$ of scene, and the incident spectral distribution $\lambda$. $Q_c$ is a function of the color response of the sensor, and $s_c(x', y', d)$ and $n(x', y')$ represent the latent scene and measurement noise (white Gaussian noise), respectively. The operator $*$ represents convolution.

We use Monte Carlo sampling in the rendering engine. At each pixel, rays are sampled starting from the sensor plane, with the wavelengths, sub-pixel origin shift, and direction sampled by a uniform random number generator without any impor-

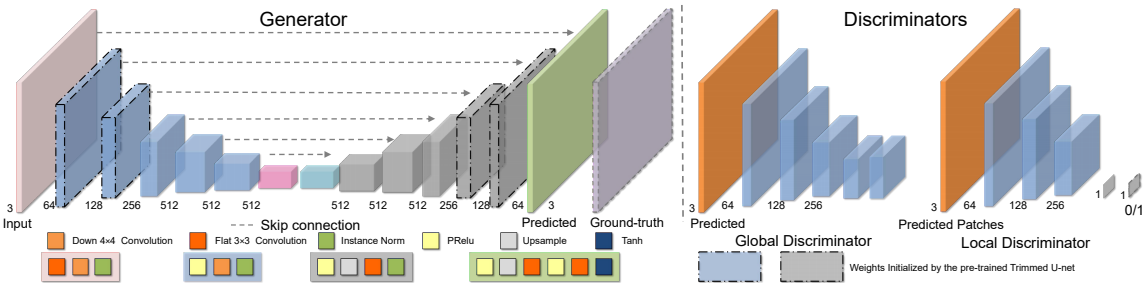


Figure 6.3: Image reconstruction architecture. The generator model is a U-net architecture that has seven scales with six consecutive downsampling and upsampling operations. We adopt a global and a local discriminator to incorporate both full spatial contexts and local details. In addition, the layers marked by the dashed line is the trimmed U-net for the end-to-end designing stage, and they are initialized with the results of the designing stage.

tance sampling. These sampled rays are then traced sequentially through each of the refractive surfaces following Snell's law. Rays are marked as invalid and do not contribute to the final rendered image when the intersections are outside of the lens geometry or when total internal reflection takes place.

Unfortunately, the number of samples per pixel (SPP) is limited by the GPU memory, resulting in Monte Carlo rendering noise. To overcome this issue, we first render several passes and average them to get a clean estimate, then replace the PyTorch variable with the clean estimation to calculate the gradient multiple times to get the averaged clean gradients. After this processing, the obtained images and gradients are clean enough, and the Monte Carlo sampling noise can be ignored [248] compared to added white Gaussian noise.

## Image alignment during training

Another challenge for end-to-end optical design is that in the initial stages of the optimization, the simulated image is both distorted and scaled compared to the desired reference. This misalignment makes it hard to accurately calculate a meaningful loss between the reference image and the rendered simulations. To solve the problem of pixelwise alignment, we first forward trace 16 points that are uniformly distributed

from the center of the texture plane to the border and obtain the points $\mathbf{r}_d$ intersected with the sensor plane. Then we set corresponding ideal points on the sensor as $\mathbf{r}$ and the relation between the point pairs can be expressed as:

$$\mathbf{r}_d = \xi\mathbf{r}(1 + k_1\mathbf{r}^2 + k_2\mathbf{r}^4 + k_3\mathbf{r}^6) \tag{6.10}$$

To simulate the distortion and magnification, we only consider the radial distortion and solve a least-squares problem as:

$$\min_{\mathbf{K}} \|[\mathbf{r}, \mathbf{r}^3, \mathbf{r}^5, \mathbf{r}^7]\mathbf{K}^T - \mathbf{r}_d\|_2^2, \tag{6.11}$$

where $\mathbf{K} = \xi(1, k_1, k_2, k3)$ represents the current distortion coefficients along with a magnification coefficient $\xi$. Then we distort and resize the reference ground truth to match the currently rendered simulation pixel-to-pixel. Once the lens parameters are fixed, we undistort the captured image in the experiments.

### 6.2.2 Image Reconstruction

**End-to-end lens design.** As shown in Figure 6.2, we connect a U-net like architecture [36] with deep layers trimmed (only use the marked layers in Figure 6.3 in designing stage) but its early layers filters that can encode the information on sensor [13]. This setup speeds up the training process and provides sufficient degrees of freedom to encode the simulated information for the end-to-end design. Specifically, the trimmed U-net architecture in the design stage has three scales with two max pool operations for downsampling and two transposed convolutions for upsampling. At the bottleneck, we adopt two flat convolutional layers. Each convolutional layer is followed by a parametric rectified linear unit (PRelu). The trained weights in this trimmed U-net network are then taken to initialize the corresponding layers for the final fine reconstruction. Refer Figure 6.3 for details.

**Generator.** At the final reconstruction stage with fixed lens parameters, we adopt a GAN as shown in Figure 6.3 to recover the corrupted sensor image $I$ from the estimate $\hat{I}$. The generator $G$ is a U-net architecture with seven scales and six down-sampling and upsampling stages. We compute the loss between the prediction $\hat{I}$ and the corresponding ground truth $I_{ref}$ by

$$\mathcal{L}_c(I_{ref}, \hat{I}) = \nu_1 \|\phi_l(\hat{I}) - \phi_l(I_{ref})\|_2 + \nu_2 \|\hat{I} - I_{ref}\|_1, \tag{6.12}$$

where $\nu_1 = 0.5$ and $\nu_2 = 0.006$ are loss balancing weights and $\nu_2$ is added to keep the color fidelity, and $\phi_l$ extracts the feature maps from the $l$-th layer of pre-trained VGG-19. Specifically, we use the "conv3_3" layer of the VGG-19 network.

**Discriminators.** As illustrated in Figure 6.3, we adopt a global discriminator to incorporate full spatial context and a local discriminator based on PatchGAN [101] to take advantage of local features. We adopt the relativistic "warping" on the least square GAN named RaGAN-LS loss [238] for a discriminator $D$ can be expressed as:

$$\begin{aligned} \mathcal{L}_{adv}(x, z) = &\mathop{\mathbb{E}}_{x \sim \mathbb{P}_x} [(D(x) - \mathop{\mathbb{E}}_{z \sim \mathbb{P}_z} [(D(G(z)) - 1)^2] \\ &+ \mathop{\mathbb{E}}_{z \sim \mathbb{P}_z} [(D(G(z)) - \mathop{\mathbb{E}}_{\hat{x} \sim \mathbb{P}_{\hat{x}}} [(D(x) - 1)^2], \end{aligned} \tag{6.13}$$

where $\mathbb{P}_x$ and $\mathbb{P}_z$ are the distributions of the data and model, respectively. This proved faster and more stable than WGAN-GP [99] in minimizing a model-generated image $z$ and the ground truth $x$. The resulting total loss can be expressed as:

$$\mathcal{L}_{total} = \mathcal{L}_c(I_{ref}, \hat{I}) + \sigma_g L_{adv-g}(I_{ref}, \hat{I}) + \sigma_l L_{adv-l}(I_{ref}, \hat{I}), \tag{6.14}$$

where $L_{adv-g}$ and $L_{adv-l}$ represents global and local adversarial loss and $\sigma_g = \sigma_l = 0.01$.

## 6.3 Implementation and Prototypes


Figure 6.4: Prototypes and the rendered section views of our designed lenses. The top left shows our fabricated lens for LFOV (left) and EDOF (right) imaging, and the corresponding structures are shown in the medium/bottom left and medium/bottom right, respectively.

The top right shows the assembled prototype with the camera body.

### 6.3.1 Datasets and Training details

For the training details of the end-to-end designing stage, please refer to Section 6.4 and Section 6.5 according to the requirements of the application LFOV and EDOF. With the lens parameters fixed, we train and finetune the image recovery network for both applications as follows. First, we rendered simulations with the full DIV2K dataset, and the texture plane are set according to the applications. We reserve the first 100 images in the DIV2K dataset [249] for quantitative comparisons, and use the remainder for training. Then we calibrate the lens distortion and find the homography to align the rendered result with the ground truth image. We use ADAM as the optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.999$. The learning rate is initialized to $10^{-4}$ for the first 50 epochs and linearly decayed to 0 over another 100 epochs using 256×256 patch pairs. All experiments were conducted using a single Nvidia RTX Titan GPU and each design takes around 20 hours.

### 6.3.2 Prototypes

**Fabrication.** Once the parameters $c$, $\kappa$ and $c_k$ of each lens profile are fixed after the end-to-end design, we fabricate each lens element with a coarse CNC machining process followed by a single-point diamond turning process. First, each lens blank was machined using a CNC machine with a precision of 0.05mm to prepare it for turning. Then we used a CNC machining system that supports 3-axis single point diamond turning (Nanotech 450) [63]. We use two substrates: PMMA with a refractive index of 1.493, and polycarbonate (PC) with a refractive index 1.5892, both measured at a principal wavelength of 550 nm. These materials represent a set of low index/low dispersion and high index/high dispersion materials that is required for designing achromatic optics.

We consider two applications, LFOV and EDOF. For the LFOV application, we design a lens system with two lens elements, made from PPMA and PC, respec-

tively. For the EDOF application, we use three elements and six design surfaces, the corresponding materials are PMMA, PC, and PC.

**System Integration.** To demonstrate the proposed framework experimentally, we use a Sony A7 camera with 6,000×4,000 pixels and a pixel pitch of 5.96 $\mu$m. The equivalent focal length for both lens designs is 50mm, with aperture sizes of 12mm and 12mm for LFOV and EDOF, respectively. Correspondingly, both of the lens designs have f-numbers of about $F4$. The fabricated lenses are mounted by our custom-designed lens tubes, and both of them have a standard C-mount as shown in 6.4. Finally, both of the two lens tubes are mounted to the camera with a C/E mount adapter.

## 6.4 Large Field-of-View Imaging

A modern complex system is effective in minimizing optical aberrations but the depth of the lens stack limiting in manufacturing high-quality LFOV lens with a low cost and will introduce additional issues, such as lens flare and complicated optical stabilization and assembling [10, 250, 40, 37]. In the last year, Peng *et al.* [13] proposed the state-of-the-art of large FOV imaging with a thin-plate optics which adopts a virtual aperture design of two aspherical surfaces, reconstructed by a generative network. However, limited by the designing space of a single element, the PSFs at different FOVs are typically larger than 900 pixels yielding strong hazing and blurring artifacts recorded on the sensor. In addition, the optics and reconstruction network are not designed fully end-to-end, the recovered image left visible artifacts even with a powerful GAN recovery network.

With our proposed differentiable complex lens model, we can design a lens with multiple elements with an aspherical profile according to the needs of the applications and find the best compromise between the complexity of lens and image quality. To

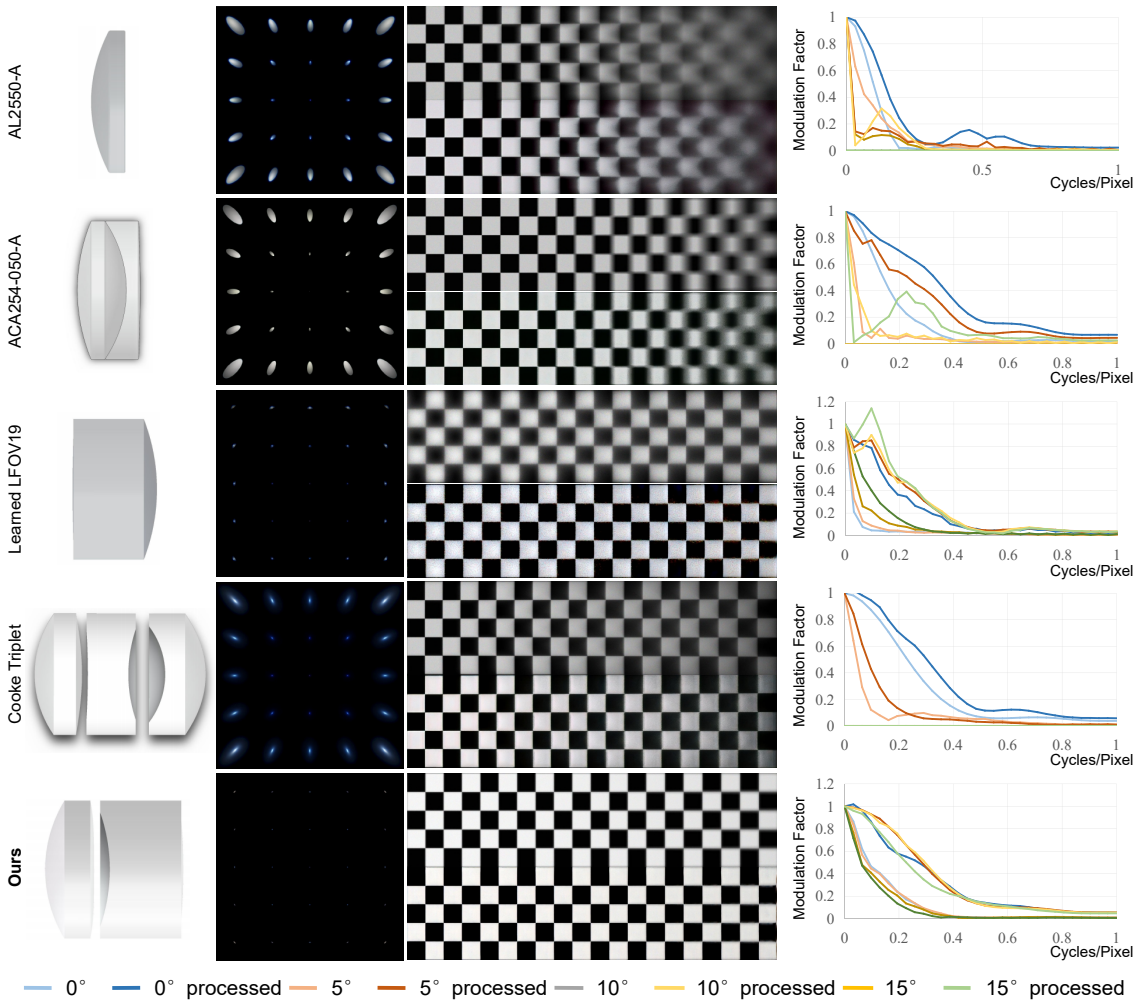


Figure 6.5: Evaluation of LFOV imaging in simulation. We compare the performance of the state-of-the-art commercially available aspherical lens Thorlabs AL2550-A, lens pairs ACA254-050-A, jointly designed optics without modulation to flat lens [13], Cooke triplet and our end-to-end designed camera. All the texture planes are located at 1m away from the camera, and the simulations are based on ray optics without considering diffraction. The first column shows the section view of each lens, and the second column shows the corresponding PSFs at different angles up to 30˙. The third column shows the simulated sensor image (top) and recovered image (bottom). The fourth column shows the MTFs of each lens at different angles. The PSFs and rendered simulation of AL2550-A and ACA254-050-A lenses show a strong blurring at large angles. LFOV19 lens performs better in balancing PSF but left significant artifacts in both measurements and reconstructions. Cooke triplet performs better than AL2550-A and ACA254-050-A but still fails at a large FOV. Instead, our design shows a better PSF distribution, and the results have fewer artifacts. Notice that all lenses are adjusted to F4.

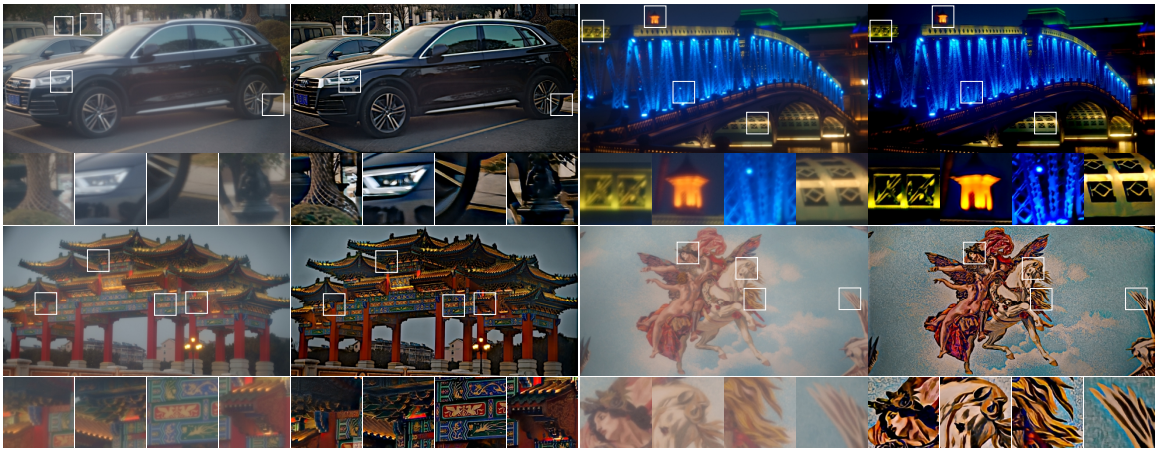

Figure 6.6: Experimental results of LFOV imaging with two elements and four surfaces design. For each pair, we give the sensor measurement by our prototype camera and the reconstructed results. Please zoom in to see more details.

apply our framework to LFOV imaging ( $\geq 30°$ ), we set a texture plane at 1m away from the camera with the size around 437mm×54.5mm and set the sensor resolution to 4096×512 pixels to cover the full designed FOV. As the lens is designed rotationally sycemmatric, the full FOV should be calculated as $2\arctan(0.437/2 * \sqrt{2}) = 34.3°$. The pixel size in the simulation is defined as $6\mu$m, matching the camera sensor used in the experiments. Limited by the GPU memory, we set the SPP to 64 for each rendering pass and average ten passes for a single scene to get a clean rendered image and corresponding accurate gradients. Refer to Section 6.2.1 for more details. To simulate a larger field of view with limited resources and reduce the time consumption, we align the sensor's left bottom corner with the optical center and simulate the image only in the first quadrant as the lens is symmetric.

We initialize the system with an initial lens design made of two lenses that are brought in focus on the optical axis. The materials are chosen as PMMA and PC for better cancelling of chromatic aberrations. To train the lens parameters for LFOV, we set all the conic coefficients $\kappa$ of each surface as variables. As illustrated in Figure 6.2, a trimmed U-net architecture $\mathbf{G}_t$ connected to the end-to-end framework. The initial learning rate is set to 0.08 and 0.0008 for the lens parameters and network parameters,

and both of them are decayed by a factor of 0.8 at each epoch. Our loss function is described as:

$$\mathcal{L}_c = \|\mathbf{G}_t(I_c) - I_{ref}\|_1. \tag{6.15}$$

Where the $I_{ref}$ represents the ground truth. In addition, once the lens parameters are fixed, we take the parameters of the network to replace the corresponding layers (marked in Figure 6.3) of the image recovery network as the initialization and train the image recover network as described in Section 6.2 to process the images captured in the real experiments. In addition, the reference images are pre-distorted and resized at the beginning of each optimization step by the method described in Section 6.2.1 to make the image pairs matched pixel-to-pixel.

### 6.4.1 Evaluation in Simulation

Figure 6.5 shows a qualitative comparison of high-quality commercial available lenses and the state-of-the-art LFOV imaging lens (LFOV19) [13]. We also show the MTFs of each lens before and after the post-processing. Notice that some data in those MTF charts are missing due to the observed edges are heavily blurred and becomes uncalculatable. We first compare against the high-quality commercial available aspherical lens Thorlabs AL2550-A, which is optimized for focusing light incident on the aspherical side of the lens with minimal spherical aberration. Then we compared against an air-spaced doublet design ACA254-050-A, which provides superior spherical and chromatic aberration correction. As illustrated in Figure 6.5, the simulated PSFs by Zemax of AL2550-A and ACA254-050-A are well focused at the center FOV while corrupted when reaching a FOV 20°. The whole FOVs of simulated images shown in Figure 6.5 are all up to 30°. The Cooke triplet performs better compared with AL2550-A and ACA254-050-A but still has a noticeable blurry at a large angle. Refer to the supplementary material for more details. The state-of-the-art dual-surface aspherical lens design named LFOV19 has a better performance than the commer-

cially available lenses as they balanced the aberrations of different FOV to achieve a larger FOV. However, this design yielding a very large PSF ($\geq$ 900 pixels) that overly degrades the image and has noticeable artifacts even with powerful generative post-processing. Our design, which is also compact and low-cost, introduces a differentiable ray-tracing based complex lens model that can directly optimize the lens parameters according to the tasks. The second column of Figure 6.5 illustrates that ours performs better from the PSF across the FOV. The third and fourth columns in Figure 6.4 give two examples of the cropped rendered simulation and corresponding reconstructed results, which use the same model and were retrained according to the lens. We show the cropped part of rendered simulation from a full FOV 0° (left side) to 30° (right side). Obviously, the results of AL2550-A and ACA254-050-A has a good performance at a small FOV but suffer from heavy blurring in off-axis regions. The LFOV19 shows an almost equal performance across the FOV but left noticeable artifacts. Ours has a better PSF behavior across the FOV, yielding better sensor measurements and reconstruction results.

We also show the quantitative comparisons in simulation in Table 6.1. Obviously, our lens performs better in both PSNR and SSIM compared with the others over a FOV from 0° to 30°. Note that the training data and recovery network are re-rendered and retrained for each lens.

Table 6.1: Quantitative comparison of image recovery performance of different lenses. We compare PSNR values in dB and SSIM values over a FOV from 0° to 30°. Notice that all lenses are adjusted to F4.

| | AL2550-A | ACA254-050A | LFOV19 | Cooke | Ours |
|---|---|---|---|---|---|
| PSNR | 16.96 | 19.03 | 16.86 | 15.724 | **22.8** |
| SSIM | 0.478 | 0.499 | 0.314 | 0.422 | **0.719** |

### 6.4.2 Experimental Results

To validate the practicability of the proposed differentiable complex lens model and the end-to-end framework, we fabricated the lens elements using single-point diamond turning and assembled them with the custom designed lens tube as shown in Figure 6.4. Figure 6.6 shows the pairs of "in-the-wild" captured raw sensor data (left) and corresponding reconstructed results (right). The exposure times for each image are 33ms, 167ms, 167ms, 100ms with ISO 50. With our end-to-end designed imaging lens and reconstruction, we achieve a high-quality LFOV imaging with minor artifacts with only two lens elements. Notice the sensor measurements show haze artifacts, which is mainly introduced by the surface roughness, scratch, and low transparency of PMMA and PC in experiments. With our generative image reconstruction, we obtain clean results with fine details, as shown in Figure 6.6. Our lens design is compact and low-cost compared to commercial bulky lens and can get comparable results with the help of our differentiable complex lens model and end-to-end framework.

## 6.5 Extended Depth of Field

Computational EDOF cameras usually design an approximately depth-invariant PSF for one wavelength and then employ a simple deconvolution to the sensor capture to obtain an all-in-focus image [1, 81, 251]. Recently, researchers proposed an end-to-end pipeline for diffractive optics or Zernike Basis [14] and applied it to achromatic EDOF imaging. However, their optics model is based on the paraxial approximation, which is only a simple Fourier transform and can only deal with a single optical surface. With the proposed differentiable complex lens model and our end-to-end framework, we relax the designing space from a single surface to multiple surfaces for EDOF imaging.

To apply our end-to-end framework to EDOF imaging, we start with an initial triple-lens design with six surfaces where the second surface of the first and third ele-

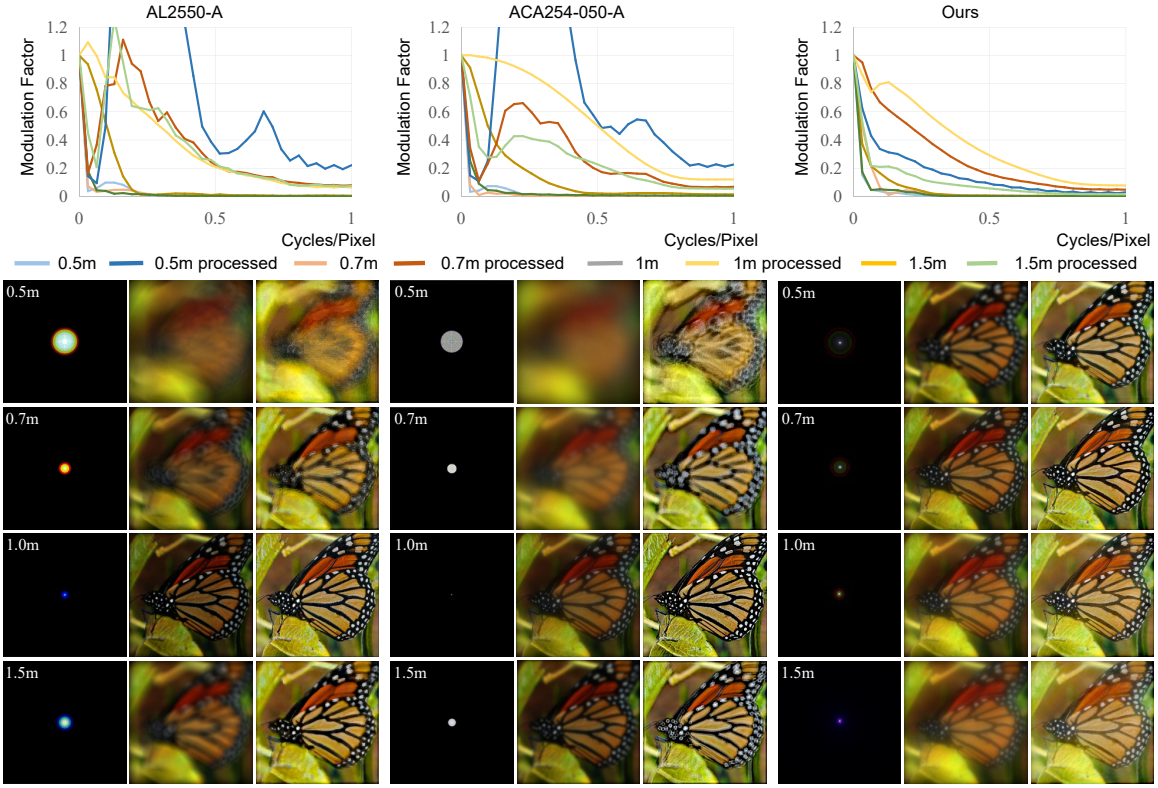


Figure 6.7: Evaluation of EDOF imaging in simulation. We compare the performance of the state-of-the-art commercially available aspherical lenses, including Thorlabs AL2550-A and ACA254-050-A. The first row shows the MTFs of each lens before and after post-processing at different depths. All the texture planes are placed 1m away from the camera, and the simulation is based on ray optics without considering diffraction. The second row shows the corresponding PSFs at the selected depth. The third column shows the simulated sensor image. Obviously, the PSFs of rendered simulation of AL2550-A and ACA254-050-A lenses exhibit a strong blur when out of focus. Instead, our design shows an almost depth invariant PSF and results with fewer artifacts. Additional results are available in the supplementary material. Notice that all lenses are adjusted to F4.

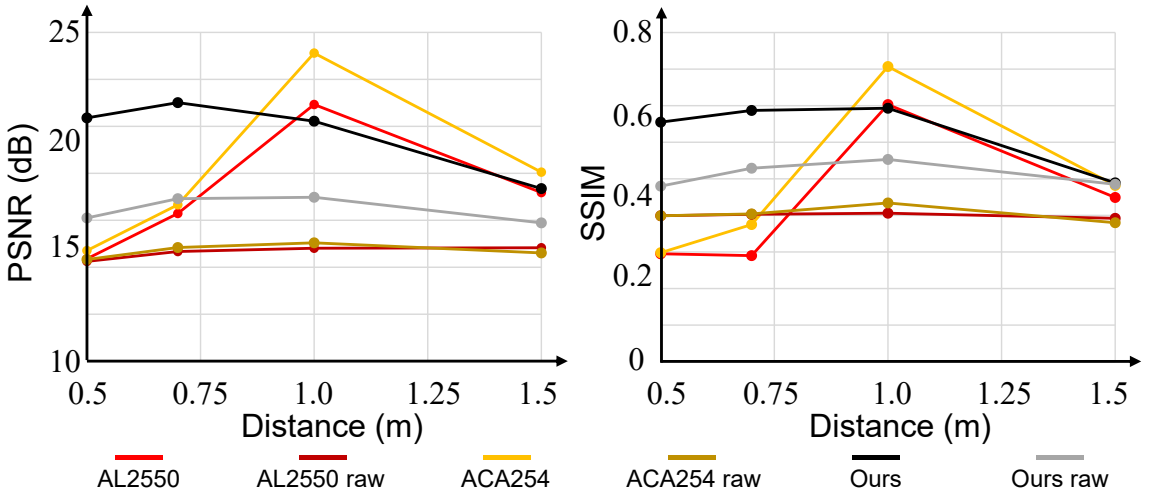


Figure 6.8: Quantitative comparison of image recovery performance of different lenses. We compare PSNR values in dB and SSIM values at 0.5m, 0.7m, 1.0m, and 1.5m. Notice that all lenses are adjusted to F4.

ments are aspherical. This design is brought in good focus near the optical axis. The materials for the three lens elements are selected as PMMA, PC, and PC for better canceling of chromatic artifacts and easier fabrication. Refer to the supplementary material for more details. To obtain clean rendered images and corresponding accurate gradients despite the limited GPU memory, we use 10 rendering passes with 128 samples per pixel each. Please refer to Section 6.2.1 for more details.

We place the texture plane at 0.5m, 0.7m, 1m, and 1.5m away from the camera in simulation and try to find the best compromise between the different depths. The pixel size in the simulation is set to 6$\mu$m and the sensor resolution is set 256×256 pixels while the texture plane sizes are set to 13.82mm×13.82mm, 20.51mm×20.51mm, 30.72mm×30.72mm and 46.08mm×46.08mm, respectively. To train the lens parameters to achieve a EDOF camera, we set the conic coefficients $\kappa$ of the spherical surface as the variable (four surfaces in total). The initial learning rate is set to 0.08 for the lens parameters, and they are decayed by a factor of 0.8 at each epoch. Our loss

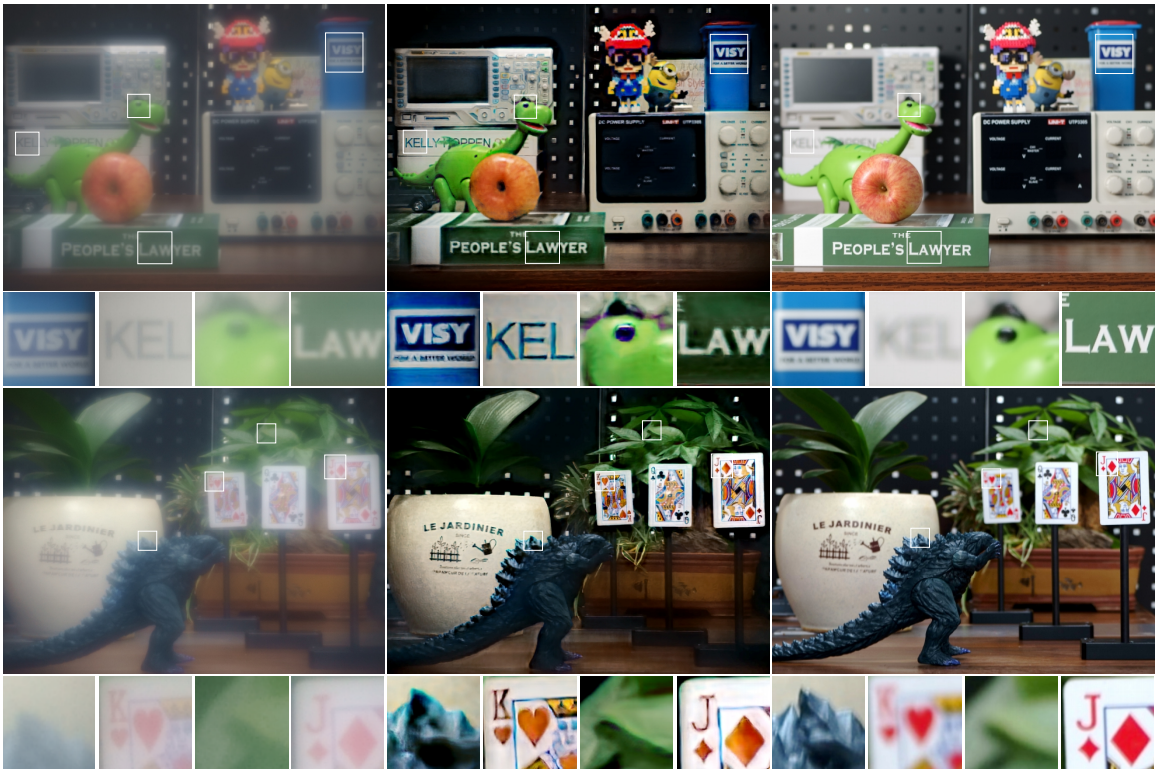


Figure 6.9: Experimental results of EDOF with three elements and six surfaces design. The left column shows the raw sensor data from our design, the center column shows our reconstruction result and the right column shows images captured by a commercial Sony 28-70mm zoom lens adjusted to 50mm/F4.5. The objects shown in these two figures are placed from around 0.8m to 1.8m, and we succeed in obtaining the all-in-focus image. Please zoom in to see more details.

function can be expressed as:

$$\mathcal{L}_c = \Sigma_j \zeta_i \| I_{ci} - I_{c2} \|_1 + \Sigma_j 3(1 - SSIM(I_c, I_{ref})), \tag{6.16}$$

where $\zeta_i$ represents the weights for different depths, and we set them to $\{3, 3, 0.3, 1\}$ corresponding to the depths mentioned above, respectively. $I_{c2}$ represents the sensor measurement with the texture plane placed a distance of 1m, and we take it as a reference to balancing the blurring amount over different depths. For each depth, we adopt a SSIM loss between the sensor measurement and the corresponding clean reference as the brightness and gamma might mismatch with the reference. In addition, the reference images are pre-distorted and resized at the beginning of each optimization step by the method described in Section 6.2.1 to align the image pairs pixel-to-pixel.

### 6.5.1 Evaluation in Simulation

We first validate our lens design in simulation and compared our lens with the high-quality commercially available lenses, including AL2550-A and ACA254-050-A. We focus all the lenses at 1m away from the camera. As illustrated in Figure 6.7, the simulated center PSFs by Zemax of AL2550-A and ACA254-050-A behave well when in focus. However, their PSF becomes unacceptably large when out of focus. In contrast, our design has an almost depth invariant PSF behavior compared with the others. We first validate our lens design in simulation and compared our lens with the high-quality commercially available lenses, including AL2550-A and ACA254-050-A. We focus all the lenses at 1m away from the camera. As illustrated in Figure 6.7, the simulated center PSFs by Zemax of AL2550-A and ACA254-050-A behave well when in focus. However, their PSF becomes unacceptably large when out of focus. In contrast, our design has an almost depth invariant PSF behavior compared with the

others. Besides, the MTFs in Figure 6.7 show that the MTF of our optimized lens is closer to the desired MTF in optical systems: smoothly and monotonously decreasing from an amplitude of 100% for the DC term to ca. 10% at the Nyquist limit of the recovered image, with no erroneous maxima for higher frequencies. Instead, the others show an obvious outlier for the post-processed data and worse performance before processing.

We further rendered the scene at different depths for each lens to provide further evidence that our end-to-end design has a larger DOF. We rendered the whole dataset for each rendered scene and retrained the recovery network for fair comparison for each rendered scene and recovered estimation pairs. As illustrated in Figure 6.7, AL2550-A and ACA254-050-A have better performance when in focus for both rendered results and corresponding recoveries but break when out of focus. In contrast, our design has a depth-balanced performance in both sensor measurements and reconstructed images.

We also show the quantitative comparisons in simulation in Figure 6.8. Our lens performs better balancing over depth in both PSNR and SSIM compared with the others. For a fair comparison, all lenses are adjusted with an aperture of F4. Note that the training data and recovery network are re-rendered and retrained according to each lens. Furthermore, the rendered images' energy distribution might vary with the lenses and make them different from the reference images, causing a relatively low matric value and less accuracy.

### 6.5.2 Experimental Results

To demonstrate the practicability of our approach in EDOF, we fabricated and assembled the lenses with the custom-designed lens tube as shown in Figure 6.4. Figure 6.9 shows the captured raw sensor measurement (left), reconstructed results (middle), and the reference image captured by a Sony 28-70mm standard zoom lens adjusted

to 50mm/F4.5. The exposure times for each image are all set to 200ms with ISO 50. Figure 6.9 illustrates that we achieved good performance and high image quality over a large DOF. Compared with our lens (F4), the Sony 28-70mm standard zoom lens has a larger f-number but worse DOF performance. Notice the sensor measurements show haze artifacts, which has been discussed in Section 6.4.2 and Section 6.6.

## 6.6 Discussion and Conclusion

### 6.6.1 Discussion

**Stablity and efficiency.** We have introduced a differentiable complex lens model that can be connected with tailored image reconstructions. Compared to conventional lens design, which requires much experience in setting up merit functions to affect the desired design characteristics, our approach reduces the need for human involvement. However, both methods can converge to a local minimum if the starting point of the design is too far off from a feasible solution. Like traditional lens design, which requires a proper initialization, our approach still requires a good initial structure that can then be further optimized automatically. Our data-driven optics do not yet take into account many standard tasks of optical design, such as zoom and focus changes, or design aspects such as tolerancing or anti-reflective coatings. We believe, however, that such extensions will be easy to add to the framework in the future.

Our approach traces hundreds of rays for each pixel as for the computational efficiency, resulting in millions of rays to compute the gradients. This is less time-efficient even with the help of the state-of-the-art ray tracing cores and has a vast space to optimize. In future work, we would like to introduce a patch wised rendering strategy instead of tracing the whole FOV to improve the computational efficiency during the designing stage.

**Fabrication.** To demonstrate our differentiable optics model and end-to-end the pipeline, we fabricated two prototypes for different applications using single-point diamond turning from PC and PMMA material. PC and PMMA are easy to manufacture, but the stability, transparency, and the easily transformed make it hard to achieve a good imaging quality for a complex lens system. Besides, the center alignment of the lenses and surfaces is a challenging task during machining and assembling. As a result, the real captured sensor images have haze artifacts compared with the simulations. However, many of these issues can likely be overcome in mass production, such as injection molding fabrication processes like the ones already used in the manufacture of cell phone cameras. We also would like to fabricate lenses from optical glass with coatings to reduce the stray light in the future.

### 6.6.2 Conclusion

We proposed a novel differentiable complex model that provides a new approach for optics design and an end-to-end framework that can be tailored for specifics tasks. We demonstrated our model and pipeline on two applications, including LFOV and EDOF imaging with compact lens designs, and tested both in real-world experiments. In addition, our model can not only be applied to the end-to-end design of optics but also offers a new approach for simulating lens' aberrations, which makes it less cumbersome to obtain a large, well-aligned training dataset for the image recovery training stage. While the proposed approach enables practical, high-quality imagery with compact designs, stability to initialization, and computational efficiency need to be further investigated in future work.

In the future, it might also be interesting to explore hybrid refractive/diffractive optical systems, and to incorporate features like coatings and other optical effects. Furthermore, building a knowledge graph that contains a large library of classic designs are an exciting direction to get rid of human involvement in initializing our

system and making the design process fully automatic.

# Chapter 7

# Concluding Remarks

## 7.1 Summary

Staring with a background introduction in Chapter 3, this dissertation formulates a path to end-to-end optics design framework step by step. Driven from the insight that the filters of early layers of recent deep models are gradient-like filters and respond to local contrast as essential low-level information, we jointly design the optics and corresponding recover network to realize LFOV imaging. This is the anterior plot of the end-to-end optics design story. To further enable the end-to-end computational camera design, Chapter 4 learns the PSF of a phase mask placed at the front pupil of the optical system to generate a proper optical encoding of the spatial information for the low fill-factor SPAD sensor model. This is the first milestone that enables end-to-end optics deigning. To free the designing freedom for different color channels, Chapter 5 gives a differentiable diffractive optics model that can directly optimize the heightmap of the optics. By introducing a rank-1 factorization of the heightmap, we drastically reduce the optical search space while allowing high-frequency encoding. This makes the story has the robustness to face challenging tasks with a stable solution.

To further bridge the gap in an end-to-end fashion, bring the simple wave optics models such as Fourier transform, or on similar paraxial models into a general, complex lens system. We build a differentiable complex lens model based on a differentiable ray-tracing rendering engine. Finally, we applied this model to two classical

computational photography tasks: LFOV imaging and EDOF imaging.

## 7.2 Future Research Work

We may further explore the application range and make all the scenarios to solve the practical imaging problems. For diffractive optics models, we can further optimize the heightmap to reduce the fabrication difficulty to make it possible for volume production. For refractive complex lens model, we can find what can be used in industry like mobile phone who wants to have a large DOF for a certain camera. Besides, we can generalize a proper optics model to reduce the costs of certain imaging products with fewer lenses but higher quality.

Besides, as the optimization process of the complex lenses is highly non-convex, the whole optimizing process is not stable enough to meet the requirement of industry or academic using without relying on an experienced optics art designer. A combination with a traditional lens design that not only renders aberrations of all kinds but also takes a serious of optical contain into consideration could be an interesting direction at the current stage.

# REFERENCES


[1] E. R. Dowski and W. T. Cathey, "Extended depth of field through wave-front coding," *Applied optics*, vol. 34, no. 11, pp. 1859–1866, 1995.

[2] S. C. Tucker, W. T. Cathey, and E. R. Dowski, "Extended depth of field and aberration control for inexpensive digital microscope systems," *Optics Express*, vol. 4, no. 11, pp. 467–474, 1999.

[3] W. T. Cathey and E. R. Dowski, "New paradigm for imaging systems," *Applied Optics*, vol. 41, no. 29, pp. 6080–6092, 2002.

[4] A. Levin, S. W. Hasinoff, P. Green, F. Durand, and W. T. Freeman, "4d frequency analysis of computational cameras for depth of field extension," in *ACM Trans. Graph. (TOG)*, vol. 28, no. 3. ACM, 2009, p. 97.

[5] P. E. Debevec and J. Malik, "Recovering high dynamic range radiance maps from photographs," in *Proceedings of the 24th Annual Conference on Computer Graphics and Interactive Techniques*, ser. SIGGRAPH '97. USA: ACM Press/Addison-Wesley Publishing Co., 1997, p. 369–378.

[6] S. Mann and R. W. Picard, "Being 'undigital' with digital cameras: extending dynamic range by combining differently exposed pictures," 1994.

[7] E. Reinhard and K. Devlin, "Dynamic range reduction inspired by photoreceptor physiology," *IEEE Transactions on Visualization and Computer Graphics*, vol. 11, no. 1, pp. 13–24, 2005.

[8] M. Rouf, R. Mantiuk, W. Heidrich, M. Trentacoste, and C. Lau, "Glare encoding of high dynamic range images," ser. CVPR '11. USA: IEEE Computer Society, 2011.

[9] S. Nayar, V. Branzoi, and T. Boult, "Programmable Imaging using a Digital Micromirror Array," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. I, Jun 2004, pp. 436–443.

[10] D. J. Brady, M. E. Gehm, R. A. Stack, D. L. Marks, D. S. Kittle, D. R. Golish, E. Vera, and S. D. Feller, "Multiscale gigapixel photography," *Nature*, vol. 486, no. 7403, p. 386, 2012.

[11] O. S. Cossairt, D. Miau, and S. K. Nayar, "Gigapixel computational imaging," in *2011 IEEE International Conference on Computational Photography (ICCP)*, 2011, pp. 1–8.

[12] Q. Sun, X. Dun, Y. Peng, and W. Heidrich, "Depth and transient imaging with compressive spad array cameras," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[13] Y. Peng, Q. Sun, X. Dun, G. Wetzstein, and W. Heidrich, "Learned large field-of-view imaging with thin-plate optics," in *ACM Transactions on Graphics (Proc. SIGGRAPH Asia)*, vol. 38, no. 6. ACM, 2019.

[14] V. Sitzmann, S. Diamond, Y. Peng, X. Dun, S. Boyd, W. Heidrich, F. Heide, and G. Wetzstein, "End-to-end optimization of optics and image processing for achromatic extended depth of field and super-resolution imaging," *ACM Transactions on Graphics (TOG)*, vol. 37, no. 4, p. 114, 2018.

[15] J. Chang and G. Wetzstein, "Deep optics for monocular depth estimation and 3d object detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

[16] ——, "Deep optics for monocular depth estimation and 3d object detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

[17] Q. Sun, J. Zhang, X. Dun, B. Ghanem, Y. Peng, and W. Heidrich, "End-to-end learned, optically coded super-resolution spad camera," *ACM Trans. Graph.*, vol. 39, no. 2, Mar. 2020.

[18] Q. Sun, E. Tseng, Q. Fu, W. Heidrich, and F. Heide, "Learning rank-1 diffractive optics for single-shot high dynamic range imaging," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[19] C. A. Metzler, H. Ikoma, Y. Peng, and G. Wetzstein, "Deep optics for single-shot high-dynamic-range imaging," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 1375–1385.

[20] M. Nimier-David, D. Vicini, T. Zeltner, and W. Jakob, "Mitsuba 2: A retargetable forward and inverse renderer," *Transactions on Graphics (Proceedings of SIGGRAPH Asia)*, vol. 38, no. 6, Dec. 2019.

[21] E. Tseng, A. Mosleh, F. Mannan, K. St-Arnaud, A. Sharma, Y. Peng, A. Braun, D. Nowrouzezahrai, J.-F. Lalonde, and F. Heide, "Differentiable compound op-

tics and processing pipeline optimization for end-to-end camera design," *ACM Transactions on Graphics (TOG)*, vol. 40, no. 4, 2021.

[22] Q. Sun, C. Wang, F. Qiang, D. Xiong, and H. Wolfgang, "End-to-end complex lens design with differentiable ray tracing," *ACM Transactions on Graphics (TOG)*, vol. 40, no. 4, 2021.

[23] T. O. Aydin, R. Mantiuk, K. Myszkowski, and H.-P. Seidel, "Dynamic range independent image quality assessment," *ACM Trans. Graph.*, vol. 27, no. 3, p. 1–10, Aug. 2008.

[24] A. R. Robertson, "Recent cie work on color difference evaluation," in *Review and Evaluation of Appearance: Methods and Techniques*. ASTM International, 1986.

[25] H. Haim, S. Elmalem, R. Giryes, A. Bronstein, and E. Marom, "Depth estimation from a single image using deep learned phase coded mask," *IEEE Transactions on Computational Imaging*, vol. 4, pp. 298–310, 2018.

[26] X. Zhang, R. Ng, and Q. Chen, "Single image reflection separation with perceptual losses," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[27] L. He, G. Wang, and Z. Hu, "Learning depth from single images with deep neural network embedding focal length," *IEEE Transactions on Image Processing*, vol. 27, pp. 4676–4689, 2018.

[28] G. R. Fowles, *Introduction to modern optics*. Courier Corporation, 1975.

[29] C. F. Gauss, *Dioptrische Untersuchungen von CF Gauss*. in der Dieterichschen Buchhandlung, 1843.

[30] R. Kingslake and R. B. Johnson, *Lens design fundamentals*. Academic Press, 2009.

[31] L. Seidel, "Ueber die theorie der fehler," *mit welchen die durch optische Instrumente gesehenen Bilder behaftet sind, und über die mathematischen Bedingungen ihrer Aufhebung. Abhandlungen der Naturwissenschaftlich-Technischen Commission bei der Königl. Bayerischen Akademie der Wissenschaften in München. Cotta*, vol. 2, p. 4, 1857.

[32] G. G. Sliusarev, "Aberration and optical design theory," *Bristol, England, Adam Hilger, Ltd., 1984, 672 p. Translation.*, 1984.

[33] J. M. Geary, *Introduction to lens design: with practical ZEMAX examples*. Willmann-Bell Richmond, 2002.

[34] D. Malacara-Hernández and Z. Malacara-Hernández, *Handbook of optical design*. CRC Press, 2016.

[35] S. W. Hasinoff, D. Sharlet, R. Geiss, A. Adams, J. T. Barron, F. Kainz, J. Chen, and M. Levoy, "Burst photography for high dynamic range and low-light imaging on mobile cameras," *ACM Transactions on Graphics (TOG)*, vol. 35, no. 6, p. 192, 2016.

[36] C. Chen, Q. Chen, J. Xu, and V. Koltun, "Learning to see in the dark," 2018.

[37] X. Yuan, L. Fang, Q. Dai, D. J. Brady, and Y. Liu, "Multiscale gigapixel video: A cross resolution image matching and warping approach," in *2017 IEEE International Conference on Computational Photography (ICCP)*. IEEE, 2017, pp. 1–9.

[38] Light.co, "Light l16 camera," 2018.

[39] MobilEye, "Mobileeye tricam," 2018.

[40] K. Venkataraman, D. Lelescu, J. Duparré, A. McMahon, G. Molina, P. Chatterjee, R. Mullis, and S. Nayar, "Picam: An ultra-thin high performance monolithic camera array," *ACM Transactions on Graphics (TOG)*, vol. 32, no. 6, p. 166, 2013.

[41] Y. Peng, Q. Fu, H. Amata, S. Su, F. Heide, and W. Heidrich, "Computational imaging using lightweight diffractive-refractive optics," *Optics express*, vol. 23, no. 24, pp. 31 393–31 407, 2015.

[42] F. Heide, M. Rouf, M. B. Hullin, B. Labitzke, W. Heidrich, and A. Kolb, "High-quality computational imaging through simple lenses," *ACM Transactions on Graphics (TOG)*, vol. 32, no. 5, p. 149, 2013.

[43] F. Heide, Q. Fu, Y. Peng, and W. Heidrich, "Encoded diffractive optics for full-spectrum computational imaging," *Scientific Reports*, vol. 6, 2016.

[44] G. R. Fowles, *Introduction to modern optics*. Courier Dover Publications, 2012.

[45] W. J. Smith, *Modern lens design*. McGraw-Hill, 2005.

[46] Y. Shih, B. Guenter, and N. Joshi, "Image enhancement using calibrated lens simulations," in *European Conference on Computer Vision*. Springer, 2012, pp. 42–56.

[47] D. G. Stork and P. R. Gill, "Lensless ultra-miniature cmos computational imagers and sensors," *Proc. Sensorcomm*, pp. 186–190, 2013.

[48] ——, "Optical, mathematical, and computational foundations of lensless ultra-miniature diffractive imagers and sensors," *International Journal on Advances in Systems and Measurements*, vol. 7, no. 3, p. 4, 2014.

[49] M. Monjur, L. Spinoulas, P. R. Gill, and D. G. Stork, "Ultra-miniature, computationally efficient diffractive visual-bar-position sensor," in *Proc. SensorComm*. IEIFSA, 2015.

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

[51] Y. Peng, Q. Fu, F. Heide, and W. Heidrich, "The diffractive achromat full spectrum computational imaging with diffractive optics," *ACM Trans. Graph. (SIGGRAPH)*, vol. 35, no. 4, p. 31, 2016.

[52] M. Papas, T. Houit, D. Nowrouzezahrai, M. H. Gross, and W. Jarosz, "The magic lens: refractive steganography." *ACM Trans. Graph.*, vol. 31, no. 6, pp. 186–1, 2012.

[53] Y. Schwartzburg, R. Testuz, A. Tagliasacchi, and M. Pauly, "High-contrast computational caustic design," *ACM Transactions on Graphics (TOG)*, vol. 33, no. 4, p. 74, 2014.

[54] Y. Peng, X. Dun, Q. Sun, and W. Heidrich, "Mix-and-match holography," *ACM Trans. Graph. (SIGGRAPH Aisa)*, vol. 36, no. 6, p. 191, 2017.

[55] Y. Wu, V. Boominathan, H. Chen, A. Sankaranarayanan, and A. Veeraraghavan, "Phasecam3d – learning phase masks for passive single view depth estimation," in *Proc. ICCP*, 2019.

[56] J. Chang, V. Sitzmann, X. Dun, W. Heidrich, and G. Wetzstein, "Hybrid optical-electronic convolutional neural networks with optimized diffractive optics for image classification," *Scientific reports*, vol. 8, no. 1, p. 12324, 2018.

[57] S. Oliver, R. Lake, S. Hegde, J. Viens, and J. Duparre, "Imaging module with symmetrical lens system and method of manufacture," May 4 2010, uS Patent 7,710,667.

[58] W. Duoshu, C. Luo, Y. Xiong, T. Chen, H. Liu, and J. Wang, "Fabrication technology of the centrosymmetric continuous relief diffractive optical elements," *Physics Procedia*, vol. 18, pp. 95–99, 2011.

[59] P. Genevet, F. Capasso, F. Aieta, M. Khorasaninejad, and R. Devlin, "Recent advances in planar optics: from plasmonic to dielectric metasurfaces," *Optica*, vol. 4, no. 1, pp. 139–152, 2017.

[60] S. H. Ahn and L. J. Guo, "Large-area roll-to-roll and roll-to-plate nanoimprint lithography: a step toward high-throughput application of continuous nanoimprinting," *ACS Nano*, vol. 3, no. 8, pp. 2304–2310, 2009.

[61] S. Y. Chou, P. R. Krauss, and P. J. Renstrom, "Nanoimprint lithography," *Journal of Vacuum Science & Technology B: Microelectronics and Nanometer Structures Processing, Measurement, and Phenomena*, vol. 14, no. 6, pp. 4129–4133, 1996.

[62] M. Zoberbier, S. Hansen, M. Hennemeyer, D. Tönnies, R. Zoberbier, M. Brehm, A. Kraft, M. Eisner, and R. Völkel, "Wafer level cameras—novel fabrication and packaging technologies," in *Int. Image Sens. Workshop*, 2009.

[63] F. Fang, X. Zhang, A. Weckenmann, G. Zhang, and C. Evans, "Manufacturing and measurement of freeform optics," *CIRP Annals*, vol. 62, no. 2, pp. 823–846, 2013.

[64] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004.

[65] K. Mitra, O. Cossairt, and A. Veeraraghavan, "To denoise or deblur: parameter optimization for imaging systems," in *Digital Photography X*, vol. 9023. International Society for Optics and Photonics, 2014, p. 90230G.

[66] A. Mittal, A. K. Moorthy, and A. C. Bovik, "No-reference image quality assessment in the spatial domain," *IEEE Transactions on Image Processing*, vol. 21, no. 12, pp. 4695–4708, 2012.

[67] M. Estribeau and P. Magnan, "Fast mtf measurement of cmos imagers using iso 12333 slanted-edge methodology," in *Detectors and Associated Signal Processing*, vol. 5251. International Society for Optics and Photonics, 2004, pp. 243–253.

[68] EMVA Standard, "1288: Standard for characterization and presentation of specification data for image sensors and cameras," *European Machine Vision Association*, 2005.

[69] J. R. Parker, *Algorithms for image processing and computer vision*. John Wiley & Sons, 2010.

[70] G. D. Boreman, *Modulation transfer function in optical and electro-optical systems*. SPIE press Bellingham, WA, 2001, vol. 21.

[71] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *European Conference on Computer Vision*. Springer, 2016, pp. 694–711.

[72] M. Geese, U. Seger, and A. Paolillo, "Detection probabilities: Performance prediction for sensors of autonomous vehicles," *Electronic Imaging*, vol. 2018, no. 17, pp. 148–1–148–14, 2018.

[73] T. S. Cho, C. L. Zitnick, N. Joshi, S. B. Kang, R. Szeliski, and W. T. Freeman, "Image restoration by matching gradient distributions," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 34, no. 4, pp. 683–694, 2012.

[74] D. Krishnan and R. Fergus, "Fast image deconvolution using hyper-laplacian priors," in *Advances in Neural Information Processing Systems*. NIPS, 2009, pp. 1033–1041.

[75] E. Gilad and J. Von Hardenberg, "A fast algorithm for convolution integrals with space and time variant kernels," *Journal of Computational Physics*, vol. 216, no. 1, pp. 326–336, 2006.

[76] C. J. Schuler, H. Christopher Burger, S. Harmeling, and B. Scholkopf, "A machine learning approach for non-blind image deconvolution," in *Proc. Computer Vision and Pattern Recognition*, 2013.

[77] L. Xu, J. S. Ren, C. Liu, and J. Jia, "Deep convolutional neural network for image deconvolution," in *Advances in Neural Information Processing Systems*, 2014, pp. 1790–1798.

[78] J. Zhang, J. Pan, W.-S. Lai, R. W. Lau, and M.-H. Yang, "Learning fully convolutional networks for iterative non-blind deconvolution," 2017.

[79] S. Nah, T. H. Kim, and K. M. Lee, "Deep multi-scale convolutional neural network for dynamic scene deblurring," in *CVPR*, vol. 1, no. 2, 2017, p. 3.

[80] O. Kupyn, V. Budzan, M. Mykhailych, D. Mishkin, and J. Matas, "Deblurgan: Blind motion deblurring using conditional adversarial networks," *arXiv preprint arXiv:1711.07064*, 2017.

[81] O. Cossairt and S. Nayar, "Spectral focal sweep: Extended depth of field from chromatic aberrations," in *IEEE International Conference on Computational Photography (ICCP)*. IEEE, 2010, pp. 1–8.

[82] P. Wang, N. Mohammad, and R. Menon, "Chromatic-aberration-corrected diffractive lenses for ultra-broadband focusing," *Scientific Reports*, vol. 6, 2016.

[83] M. Huh, P. Agrawal, and A. A. Efros, "What makes imagenet good for transfer learning?" *arXiv preprint arXiv:1608.08614*, 2016.

[84] Y. Bengio, "Deep learning of representations for unsupervised and transfer learning," in *Proceedings of ICML Workshop on Unsupervised and Transfer Learning*, 2012, pp. 17–36.

[85] E. Hecht, "Hecht optics," *Addison Wesley*, vol. 997, pp. 213–214, 1998.

[86] J. W. Goodman, *Introduction to Fourier optics*. Roberts and Company Publishers, 2005.

[87] A. Kalvach and Z. Szabó, "Aberration-free flat lens design for a wide range of incident angles," *Journal of the Optical Society of America B*, vol. 33, no. 2, p. A66, 2016.

[88] J. Zhu, T. Yang, and G. Jin, "Design method of surface contour for a freeform lens with wide linear field-of-view," *Optics express*, vol. 21, no. 22, pp. 26 080–26 092, 2013.

[89] R. Ng, M. Levoy, M. Brédif, G. Duval, M. Horowitz, P. Hanrahan *et al.*, "Light field photography with a hand-held plenoptic camera," 2005.

[90] Y.-R. Ng, P. M. Hanrahan, M. A. Horowitz, and M. S. Levoy, "Correction of optical aberrations," Aug. 14 2012, uS Patent 8,243,157.

[91] R. Ramanath, W. Snyder, Y. Yoo, and M. Drew, "Color image processing pipeline in digital still cameras," *IEEE Signal Processing Magazine*, vol. 22, no. 1, pp. 34–43, 2005.

[92] T. Brooks, B. Mildenhall, T. Xue, J. Chen, D. Sharlet, and J. T. Barron, "Unprocessing images for learned raw denoising," *arXiv preprint arXiv:1811.11127*, 2018.

[93] F. Heide, M. Steinberger, Y.-T. Tsai, M. Rouf, D. Pajak, D. Reddy, O. Gallo, J. Liu, W. Heidrich, K. Egiazarian *et al.*, "Flexisp: A flexible camera image processing framework," *ACM Transactions on Graphics (TOG)*, vol. 33, no. 6, p. 231, 2014.

[94] L. Sun, S. Cho, J. Wang, and J. Hays, "Edge-based blur kernel estimation using patch priors," in *Proc. International Conference on Computational Photography (ICCP)*, 2013, pp. 1–8.

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

[96] A. Odena, V. Dumoulin, and C. Olah, "Deconvolution and checkerboard artifacts," *Distill*, vol. 1, no. 10, p. e3, 2016.

[97] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. P. Aitken, A. Tejani, J. Totz, Z. Wang *et al.*, "Photo-realistic single image super-resolution using a generative adversarial network," in *CVPR*, vol. 2, no. 3, 2017, p. 4.

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

[99] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein gan," *arXiv preprint arXiv:1701.07875*, 2017.

[100] L.-W. Chang, Y. Chen, W. Bao, A. Agarwal, E. Akchurin, K. Deng, and E. Barsoum, "Accelerating recurrent neural networks through compiler techniques and quantization," 2018.

[101] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," *CVPR*, 2017.

[102] A. Velten, T. Willwacher, O. Gupta, A. Veeraraghavan, M. G. Bawendi, and R. Raskar, "Recovering three-dimensional shape around a corner using ultrafast time-of-flight imaging," *Nature communications*, vol. 3, p. 745, 2012.

[103] A. Velten, D. Wu, A. Jarabo, B. Masia, C. Barsi, C. Joshi, E. Lawson, M. Bawendi, D. Gutierrez, and R. Raskar, "Femto-photography: capturing and visualizing the propagation of light," *ACM Trans. Graph. (ToG)*, vol. 32, no. 4, p. 44, 2013.

[104] D. Shin, F. Xu, D. Venkatraman, R. Lussana, F. Villa, F. Zappa, V. K. Goyal, F. N. Wong, and J. H. Shapiro, "Photon-efficient imaging with a single-photon camera," *Nature Communications*, vol. 7, 2016.

[105] G. Gariepy, N. Krstajić, R. Henderson, C. Li, R. R. Thomson, G. S. Buller, B. Heshmat, R. Raskar, J. Leach, and D. Faccio, "Single-photon sensitive light-in-fight imaging," *Nature Communications*, vol. 6, 2015.

[106] M. O'Toole, F. Heide, D. B. Lindell, K. Zang, S. Diamond, and G. Wetzstein, "Reconstructing transient images from single-photon sensors," in *Proc. Computer Vision and Pattern Recognization (CVPR)*. IEEE, 2017, pp. 2289–2297.

[107] D. E. Schwartz, E. Charbon, and K. L. Shepard, "A single-photon avalanche diode array for fluorescence lifetime imaging microscopy," *IEEE journal of solid-state circuits*, vol. 43, no. 11, pp. 2546–2557, 2008.

[108] D.-U. Li, J. Arlt, J. Richardson, R. Walker, A. Buts, D. Stoppa, E. Charbon, and R. Henderson, "Real-time fluorescence lifetime imaging system with a 32× 32 0.13 $\mu$m cmos low dark-count single-photon avalanche diode array," *Optics Express*, vol. 18, no. 10, pp. 10 257–10 269, 2010.

[109] M. V. Nemallapudi, S. Gundacker, P. Lecoq, E. Auffray, A. Ferri, A. Gola, and C. Piemonte, "Sub-100 ps coincidence time resolution for positron emission tomography with lso: Ce codoped with ca," *Physics in Medicine & Biology*, vol. 60, no. 12, p. 4635, 2015.

[110] A. C. Ulku, C. Bruschini, I. M. Antolović, Y. Kuo, R. Ankri, S. Weiss, X. Michalet, and E. Charbon, "A 512× 512 spad image sensor with integrated gating for widefield flim," *IEEE Journal of Selected Topics in Quantum Electronics*, vol. 25, no. 1, pp. 1–12, 2018.

[111] J. M. Pavia, M. Wolf, and E. Charbon, "Measurement and modeling of microlenses fabricated on single-photon avalanche diode arrays for fill factor recovery," *Optics express*, vol. 22, no. 4, pp. 4202–4213, 2014.

[112] G. Intermite, A. McCarthy, R. E. Warburton, X. Ren, F. Villa, R. Lussana, A. J. Waddie, M. R. Taghizadeh, A. Tosi, F. Zappa *et al.*, "Fill-factor improvement of si cmos single-photon avalanche diode detector arrays by integration of diffractive microlens arrays," *Optics Express*, vol. 23, no. 26, pp. 33 777–33 791, 2015.

[113] H. Chen, M. S. Asif, A. C. Sankaranarayanan, and A. Veeraraghavan, "Fpa-cs: Focal plane array-based compressive imaging in short-wave infrared," in *Proc. Computer Vision and Pattern Recognization (CVPR)*. IEEE, 2015, pp. 2358–2366.

[114] L. Xiao, F. Heide, M. O'Toole, A. Kolb, M. B. Hullin, K. Kutulakos, and W. Heidrich, "Defocus deblurring and superresolution for time-of-flight depth cameras," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2015, pp. 2376–2384.

[115] S. R. P. Pavani, M. A. Thompson, J. S. Biteen, S. J. Lord, N. Liu, R. J. Twieg, R. Piestun, and W. Moerner, "Three-dimensional, single-molecule fluorescence imaging beyond the diffraction limit by using a double-helix point spread function," *Proceedings of the National Academy of Sciences*, vol. 106, no. 9, pp. 2995–2999, 2009.

[116] Y. Shechtman, S. J. Sahl, A. S. Backer, and W. Moerner, "Optimal point spread function design for 3d imaging," *Physical review letters*, vol. 113, no. 13, p. 133902, 2014.

[117] A. Levin, R. Fergus, F. Durand, and W. T. Freeman, "Image and depth from a conventional camera with a coded aperture," *ACM Trans. Graph. (TOG)*, vol. 26, no. 3, p. 70, 2007.

[118] B. Cai, X. Xu, K. Jia, C. Qing, and D. Tao, "Dehazenet: An end-to-end system for single image haze removal," *IEEE Transactions on Image Processing*, vol. 25, no. 11, pp. 5187–5198, 2016.

[119] D. Gong, J. Yang, L. Liu, Y. Zhang, I. D. Reid, C. Shen, A. Van Den Hengel, and Q. Shi, "From motion blur to motion flow: A deep learning solution for removing heterogeneous motion blur." in *Proc. Computer Vision and Pattern Recognization (CVPR)*, vol. 1, no. 2. IEEE, 2017, p. 5.

[120] S. Su, F. Heide, G. Wetzstein, and W. Heidrich, "Deep end-to-end time-of-flight imaging," in *Proc. Computer Vision and Pattern Recognization (CVPR)*. IEEE, 2018, pp. 6383–6392.

[121] J. Yang, J. Wright, T. Huang, and Y. Ma, "Image super-resolution as sparse representation of raw image patches," in *Proc. Computer Vision and Pattern Recognization (CVPR)*. IEEE, 2008, pp. 1–8.

[122] J. Yang, J. Wright, T. S. Huang, and Y. Ma, "Image super-resolution via sparse representation," *IEEE transactions on image processing*, vol. 19, no. 11, pp. 2861–2873, 2010.

[123] C.-Y. Yang and M.-H. Yang, "Fast direct super-resolution by simple functions." IEEE, 2013, pp. 561–568.

[124] R. Timofte, V. De Smet, and L. Van Gool, "A+: Adjusted anchored neighborhood regression for fast super-resolution." Springer, 2014, pp. 111–126.

[125] S. Schulter, C. Leistner, and H. Bischof, "Fast and accurate image upscaling with super-resolution forests," in *Proc. Computer Vision and Pattern Recognization (CVPR)*. IEEE, 2015, pp. 3791–3799.

[126] W. Shi, J. Caballero, F. Huszár, J. Totz, A. P. Aitken, R. Bishop, D. Rueckert, and Z. Wang, "Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2016, pp. 1874–1883.

[127] C. Dong, C. C. Loy, and X. Tang, "Accelerating the super-resolution convolutional neural network," in *European Conference on Computer Vision*. Springer, 2016, pp. 391–407.

[128] C. Dong, C. C. Loy, K. He, and X. Tang, "Image super-resolution using deep convolutional networks," *IEEE transactions on pattern analysis and machine intelligence*, vol. 38, no. 2, pp. 295–307, 2016.

[129] Z. Wang, D. Liu, J. Yang, W. Han, and T. Huang, "Deep networks for image super-resolution with sparse prior." IEEE, 2015, pp. 370–378.

[130] J. Kim, J. Kwon Lee, and K. Mu Lee, "Accurate image super-resolution using very deep convolutional networks," in *Proc. Computer Vision and Pattern Recognization (CVPR)*. IEEE, 2016, pp. 1646–1654.

[131] ——, "Deeply-recursive convolutional network for image super-resolution," in *Proc. Computer Vision and Pattern Recognization (CVPR)*. IEEE, 2016, pp. 1637–1645.

[132] W.-S. Lai, J.-B. Huang, N. Ahuja, and M.-H. Yang, "Deep laplacian pyramid networks for fast and accurate super-resolution," in *Proc. Computer Vision and Pattern Recognization (CVPR)*. IEEE, 2017, pp. 624–632.

[133] B. Lim, S. Son, H. Kim, S. Nah, and K. M. Lee, "Enhanced deep residual networks for single image super-resolution," in *Proc. Computer Vision and Pattern Recognization (CVPR)Workshops*, vol. 1, no. 2. IEEE, 2017, p. 3.

[134] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. Computer Vision and Pattern Recognization (CVPR)*. IEEE, 2016, pp. 770–778.

[135] M. Haris, G. Shakhnarovich, and N. Ukita, "Deep back-projection networks for super-resolution," *arXiv*, 2018.

[136] N. George and W. Chi, "Extended depth of field using a logarithmic asphere," *Journal of Optics A: Pure and Applied Optics*, vol. 5, no. 5, p. S157, 2003.

[137] L.-H. Yeh and L. Waller, "3d super-resolution optical fluctuation imaging (3d-sofi) with speckle illumination," in *Computational Optical Sensing and Imaging*. Optical Society of America, 2016, pp. CW5D–2.

[138] C. Zhou, S. Lin, and S. K. Nayar, "Coded aperture pairs for depth from defocus and defocus deblurring," *International journal of computer vision*, vol. 93, no. 1, pp. 53–72, 2011.

[139] R. F. Marcia, Z. T. Harmany, and R. M. Willett, "Compressive coded aperture imaging," in *Computational Imaging VII*, vol. 7246. International Society for Optics and Photonics, 2009, p. 72460G.

[140] P. Llull, X. Liao, X. Yuan, J. Yang, D. Kittle, L. Carin, G. Sapiro, and D. J. Brady, "Coded aperture compressive temporal imaging," *Optics express*, vol. 21, no. 9, pp. 10 526–10 545, 2013.

[141] G. R. Arce, D. J. Brady, L. Carin, H. Arguello, and D. S. Kittle, "Compressive coded aperture spectral imaging: An introduction," *IEEE Signal Processing Magazine*, vol. 31, no. 1, pp. 105–115, 2014.

[142] G. Kim, J. A. Domínguez-Caballero, and R. Menon, "Design and analysis of multi-wavelength diffractive optics," *Optics Express*, vol. 20, no. 3, pp. 2814–2823, 2012.

[143] W. Qu, H. Gu, H. Zhang, and Q. Tan, "Image magnification in lensless holographic projection using double-sampling fresnel diffraction," *Applied Optics*, vol. 54, no. 34, pp. 10 018–10 021, 2015.

[144] M. Petrov, S. Bibikov, Y. Yuzifovich, R. Skidanov, and A. Nikonorov, "Color correction with 3d lookup tables in diffractive optical imaging systems," *Procedia Engineering*, vol. 201, pp. 73–82, 2017.

[145] Y. Peng, X. Dun, Q. Sun, F. Heide, and W. Heidrich, "Focal sweep imaging with multi-focal diffractive optics," in *International Conference on Computational Photography (ICCP)*. IEEE, 2018, pp. 1–8.

[146] C. Zhao, A. Carass, B. E. Dewey, J. Woo, J. Oh, P. A. Calabresi, D. S. Reich, P. Sati, D. L. Pham, and J. L. Prince, "A deep learning based anti-aliasing self super-resolution algorithm for mri," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2018, pp. 100–108.

[147] S. Datta, N. Chaki, and K. Saeed, "Minimizing aliasing effects using faster super resolution technique on text images," in *Transactions on Computational Science XXXI*. Springer, 2018, pp. 136–153.

[148] D. O'Connor, *Time-correlated single photon counting*. Academic Press, 2012.

[149] D. D.-U. Li, S. Ameer-Beg, J. Arlt, D. Tyndall, R. Walker, D. R. Matthews, V. Visitkul, J. Richardson, and R. K. Henderson, "Time-domain fluorescence lifetime imaging techniques suitable for solid-state imaging sensor arrays," *Sensors*, vol. 12, no. 5, pp. 5650–5669, 2012.

[150] A. Kirmani, D. Venkatraman, D. Shin, A. Colaço, F. N. Wong, J. H. Shapiro, and V. K. Goyal, "First-photon imaging," *Science*, vol. 343, no. 6166, pp. 58–61, 2014.

[151] A. K. Pediredla, A. C. Sankaranarayanan, M. Buttafava, A. Tosi, and A. Veeraraghavan, "Signal processing based pile-up compensation for gated single-photon avalanche diodes," *arXiv preprint arXiv:1806.07437*, 2018.

[152] F. Heide, S. Diamond, D. B. Lindell, and G. Wetzstein, "Sub-picosecond photon-efficient 3d imaging using single-photon sensors," *arXiv*, 2018.

[153] F. Heide, M. O'Toole, K. Zang, D. B. Lindell, S. Diamond, and G. Wetzstein, "Non-line-of-sight imaging with partial occluders and surface normals," *ACM Trans. Graph.*, 2019.

[154] D. B. Lindell, G. Wetzstein, and M. O'Toole, "Wave-based non-line-of-sight imaging using fast f-k migration," *ACM Trans. Graph. (SIGGRAPH)*, vol. 38, no. 4, p. 116, 2019.

[155] D. B. Lindell, M. O'Toole, and G. Wetzstein, "Single-photon 3d imaging with deep sensor fusion," *ACM Trans. Graph. (SIGGRAPH)*, no. 4, 2018.

[156] M. Iliadis, L. Spinoulas, and A. K. Katsaggelos, "Deepbinarymask: Learning a binary mask for video compressive sensing," *arXiv*, 2016.

[157] A. Chakrabarti, "Learning sensor multiplexing design through back-propagation," in *Advances in Neural Information Processing Systems*, 2016, pp. 3081–3089.

[158] Y. Wu, V. Boominathan, H. Chen, A. Sankaranarayanan, and A. Veeraraghavan, "Phasecam3d—learning phase masks for passive single view depth estimation," in *Computational Photography (ICCP), 2019 IEEE International Conference on*. IEEE, 2019, pp. 1–8.

[159] J. Chang, V. Sitzmann, X. Dun, W. Heidrich, and G. Wetzstein, "Hybrid optical-electronic convolutional neural networks with optimized diffractive optics for image classification," *Scientific Reports*, 2018.

[160] M. Parker, *Digital Signal Processing 101, Second Edition: Everything You Need to Know to Get Started*, 2nd ed. Newton, MA, USA: Newnes, 2017.

[161] R. W. Gerchberg and W. O. Saxton, "A practical algorithm for the determination of phase from image and diffraction plane pictures," *Optik*, vol. 35, p. 237, 1972.

[162] B. Morgan, C. M. Waits, J. Krizmanic, and R. Ghodssi, "Development of a deep silicon phase fresnel lens using gray-scale lithography and deep reactive ion etching," *Journal of microelectromechanical systems*, vol. 13, no. 1, pp. 113–120, 2004.

[163] F. Heide, L. Xiao, A. Kolb, M. B. Hullin, and W. Heidrich, "Imaging in scattering media using correlation image sensors and sparse convolutional coding," *Optics Express*, vol. 22, no. 21, pp. 26 338–26 350, 2014.

[164] M. Bevilacqua, A. Roumy, C. Guillemot, and M. L. Alberi-Morel, "Low-complexity single-image super-resolution based on nonnegative neighbor embedding," 2012.

[165] R. Zeyde, M. Elad, and M. Protter, "On single image scale-up using sparse-representations," in *International conference on curves and surfaces*. Springer, 2010, pp. 711–730.

[166] P. Arbelaez, M. Maire, C. Fowlkes, and J. Malik, "Contour detection and hierarchical image segmentation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 33, no. 5, pp. 898–916, 2011.

[167] P. D. Burns and D. Williams, "Refined slanted-edge measurement for practical camera and scanner testing," in *IS AND TS PICS CONFERENCE*. Society for Imaging Science and Technology, 2002, pp. 191–195.

[168] D. Qin, Y. Xia, and G. M. Whitesides, "Soft lithography for micro-and nanoscale patterning," *Nature protocols*, vol. 5, no. 3, pp. 491–502, 2010.

[169] S. Donati, G. Martini, and M. Norgia, "Microconcentrators to recover fill-factor in image photodetectors with pixel on-board processing circuits," *Optics express*, vol. 15, no. 26, pp. 18 066–18 075, 2007.

[170] B. Mildenhall, J. T. Barron, J. Chen, D. Sharlet, R. Ng, and R. Carroll, "Burst denoising with kernel prediction networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 2502–2510.

[171] A. Darmont and S. of Photo-optical Instrumentation Engineers, "High dynamic range imaging: sensors and architectures." SPIE Washington, 2012.

[172] U. Seger, "Hdr imaging in automotive applications," in *High Dynamic Range Video*. Elsevier, 2016, pp. 477–498.

[173] T. Mertens, J. Kautz, and F. Van Reeth, "Exposure fusion: A simple and practical alternative to high dynamic range photography," in *Computer graphics forum*, vol. 28, no. 1. Wiley Online Library, 2009, pp. 161–171.

[174] S. K. Nayar and T. Mitsunaga, "High dynamic range imaging: Spatially varying pixel exposures," in *Proceedings IEEE Conference on Computer Vision and Pattern Recognition. CVPR 2000 (Cat. No. PR00662)*, vol. 1. IEEE, 2000, pp. 472–479.

[175] T. Willassen, J. Solhusvik, R. Johansson, S. Yaghmai, H. Rhodes, S. Manabe, D. Mao, Z. Lin, D. Yang, O. Cellek *et al.*, "A 1280× 1080 4.2 $\mu$m split-diode pixel hdr sensor in 110 nm bsi cmos process," in *Proceedings of the International Image Sensor Workshop, Vaals, The Netherlands*, 2015, pp. 8–11.

[176] A. Morimitsu, I. Hirota, S. Yokogawa, I. Ohdaira, M. Matsumura, H. Takahashi, T. Yamazaki, H. Oyaizu, Y. Incesu, M. Atif *et al.*, "A 4m pixel full-pdaf cmos image sensor with 1.58 $\mu$m 2× 1 on-chip micro-split-lens technology," in *ITE Technical Report 39.35*. The Institute of Image Information and Television Engineers, 2015, pp. 5–8.

[177] M. D. Tocci, C. Kiser, N. Tocci, and P. Sen, "A versatile hdr video production system," in *ACM Transactions on Graphics (TOG)*, vol. 30, no. 4. ACM, 2011, p. 41.

[178] G. Eilertsen, J. Kronander, G. Denes, R. K. Mantiuk, and J. Unger, "Hdr image reconstruction from a single exposure using deep cnns," *ACM Transactions on Graphics (TOG)*, vol. 36, no. 6, p. 178, 2017.

[179] K. Fotiadou, G. Tsagkatakis, and P. Tsakalides, "Snapshot high dynamic range imaging via sparse representations and feature learning," *IEEE Transactions on Multimedia*, 2019.

[180] M. Rouf, R. Mantiuk, W. Heidrich, M. Trentacoste, and C. Lau, "Glare encoding of high dynamic range images," *CVPR 2011*, pp. 289–296, 2011.

[181] C. A. Metzler, H. Ikoma, Y. Peng, and G. Wetzstein, "Deep optics for single-shot high-dynamic-range imaging," *arXiv preprint arXiv:1908.00620*, 2019.

[182] P. E. Debevec and J. Malik, "Recovering high dynamic range radiance maps from photographs," in *SIGGRAPH '08*, 1997.

[183] E. Reinhard, G. Ward, S. Pattanaik, P. E. Debevec, W. Heidrich, and K. Myszkowski, "High dynamic range imaging: Acquisition, display, and image-based lighting," 2010.

[184] M. D. Grossberg and S. K. Nayar, "High dynamic range from multiple images: Which exposures to combine?" 2003.

[185] T. Mertens, J. Kautz, and F. V. Reeth, "Exposure fusion: A simple and practical alternative to high dynamic range photography," *Comput. Graph. Forum*, vol. 28, pp. 161–171, 2009.

[186] S. W. Hasinoff, F. Durand, and W. T. Freeman, "Noise-optimal capture for high dynamic range photography," *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 553–560, 2010.

[187] S. B. Kang, M. Uyttendaele, S. A. J. Winder, and R. Szeliski, "High dynamic range video," *ACM Trans. Graph.*, vol. 22, pp. 319–325, 2003.

[188] E. A. Khan, A. O. Akyüz, and E. Reinhard, "Ghost removal in high dynamic range images," *2006 International Conference on Image Processing*, pp. 2005–2008, 2006.

[189] C. Liu, "Exploring new representations and applications for motion analysis," 2009.

[190] O. Gallo, N. Gelfandz, W.-C. Chen, M. Tico, and K. Pulli, "Artifact-free high dynamic range imaging," *2009 IEEE International Conference on Computational Photography (ICCP)*, pp. 1–7, 2009.

[191] M. Granados, K. I. Kim, J. Tompkin, and C. Theobalt, "Automatic noise modeling for ghost-free hdr reconstruction," *ACM Trans. Graph.*, vol. 32, pp. 201:1–201:10, 2013.

[192] J. Hu, O. Gallo, K. Pulli, and X. Sun, "Hdr deghosting: How to deal with saturation?" *2013 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1163–1170, 2013.

[193] N. K. Kalantari, E. Shechtman, C. Barnes, S. Darabi, D. B. Goldman, and P. Sen, "Patch-based high dynamic range video," *ACM Trans. Graph.*, vol. 32, pp. 202:1–202:8, 2013.

[194] P. Sen, N. K. Kalantari, M. Yaesoubi, S. Darabi, D. B. Goldman, and E. Shechtman, "Robust patch-based hdr reconstruction of dynamic scenes," *ACM Trans. Graph.*, vol. 31, pp. 203:1–203:11, 2012.

[195] N. K. Kalantari and R. Ramamoorthi, "Deep high dynamic range imaging of dynamic scenes," *ACM Trans. Graph.*, vol. 36, pp. 144:1–144:12, 2017.

[196] ——, "Deep hdr video from sequences with alternating exposures," *Comput. Graph. Forum*, vol. 38, pp. 193–205, 2019.

[197] F. Banterle, P. Ledda, K. Debattista, and A. Chalmers, "Inverse tone mapping," in *GRAPHITE*, 2006.

[198] P. Didyk, R. Mantiuk, M. Hein, and H.-P. Seidel, "Enhancement of bright video features for hdr displays," *Comput. Graph. Forum*, vol. 27, pp. 1265–1274, 2008.

[199] L. Meylan, S. J. Daly, and S. Süsstrunk, "The reproduction of specular highlights on high dynamic range displays," in *Color Imaging Conference*, 2006.

[200] A. G. Rempel, M. Trentacoste, H. Seetzen, H. D. Young, W. Heidrich, L. A. Whitehead, and G. Ward, "Ldr2hdr: on-the-fly reverse tone mapping of legacy video and photographs," in *SIGGRAPH 2007*, 2007.

[201] A. O. Akyüz, R. W. Fleming, B. E. Riecke, E. Reinhard, and H. H. Bülthoff, "Do hdr displays support ldr content?: a psychophysical evaluation," in *SIGGRAPH 2007*, 2007.

[202] B. Masiá, S. Agustin, R. W. Fleming, O. Sorkine-Hornung, and D. Gutierrez, "Evaluation of reverse tone mapping through varying exposure conditions," *ACM Trans. Graph.*, vol. 28, p. 160, 2009.

[203] K. Moriwaki, R. Yoshihashi, R. Kawakami, S. You, and T. Naemura, "Hybrid loss for learning single-image-based HDR reconstruction," *arXiv preprint arXiv:1812.07134*, 2018.

[204] Y. Endo, Y. Kanamori, and J. Mitani, "Deep reverse tone mapping," *ACM Transactions on Graphics (Proc. of SIGGRAPH Asia)*, vol. 36, no. 6, p. 177, 2017.

[205] J. Zhang and J. Lalonde, "Learning high dynamic range from outdoor panoramas," *CoRR*, vol. abs/1703.10200, 2017. [Online]. Available: http://arxiv.org/abs/1703.10200

[206] S. Lee, G. H. An, and S.-J. Kang, "Deep chain hdri: Reconstructing a high dynamic range image from a single low dynamic range image," *IEEE Access*, vol. 6, pp. 49 913–49 924, 2018.

[207] S. Lee, G. Hwan An, and S.-J. Kang, "Deep recursive hdri: Inverse tone mapping using generative adversarial networks," in *The European Conference on Computer Vision (ECCV)*, September 2018.

[208] C. Wang, Y. Zhao, and R. Wang, "Deep inverse tone mapping for compressed images," *IEEE Access*, vol. 7, pp. 74 558–74 569, 2019.

[209] D. Marnerides, T. Bashford-Rogers, J. Hatchett, and K. Debattista, "Expandnet: A deep convolutional neural network for high dynamic range expansion from low dynamic range content," *CoRR*, vol. abs/1803.02266, 2018. [Online]. Available: http://arxiv.org/abs/1803.02266

[210] S. Ning, H. Xu, L. Song, R. Xie, and W. Zhang, "Learning an inverse tone mapping network with a generative adversarial regularizer," *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1383–1387, 2018.

[211] H. Jang, K. Bang, J. Jang, and D. Hwang, "Inverse tone mapping operator using sequential deep neural networks based on the human visual system," *IEEE Access*, vol. 6, pp. 52 058–52 072, 2018.

[212] S. Hajisharif, J. Kronander, and J. Unger, "Adaptive dualiso hdr reconstruction," *EURASIP Journal on Image and Video Processing*, vol. 2015, pp. 1–13, 2015.

[213] A. Serrano, F. Heide, D. Gutierrez, G. Wetzstein, and B. Masiá, "Convolutional sparse coding for high dynamic range imaging," *Comput. Graph. Forum*, vol. 35, pp. 153–163, 2016.

[214] W. Guicquero, A. Dupret, and P. Vandergheynst, "An algorithm architecture co-design for cmos compressive high dynamic range imaging," *IEEE Transactions on Computational Imaging*, vol. 2, pp. 190–203, 2016.

[215] H. Zhao, B. Shi, C. Fernandez-Cull, S.-K. Yeung, and R. Raskar, "Unbounded high dynamic range photography using a modulo camera," *2015 IEEE International Conference on Computational Photography (ICCP)*, pp. 1–10, 2015.

[216] K. Hirakawa and P. M. Simon, "Single-shot high dynamic range imaging with conventional camera hardware," *2011 International Conference on Computer Vision*, pp. 1339–1346, 2011.

[217] A. Chakrabarti, "Learning sensor multiplexing design through back-propagation," *ArXiv*, vol. abs/1605.07078, 2016.

[218] R. Horstmeyer, R. Y. Chen, B. Kappes, and B. Judkewitz, "Convolutional neural networks that teach microscopes how to image," *ArXiv*, vol. abs/1709.07223, 2017.

[219] M. Kellman, E. Bostan, M. Chen, and L. Waller, "Data-driven design for fourier ptychographic microscopy," in *2019 IEEE International Conference on Computational Photography (ICCP)*. IEEE, 2019, pp. 1–8.

[220] E. Nehme, D. Freedman, R. Gordon, B. Ferdman, T. Michaeli, and Y. Shechtman, "Dense three dimensional localization microscopy by deep learning," 2019.

[221] Y. Shechtman, L. E. Weiss, A. S. Backer, M. Y. Lee, and W. E. Moerner, "Multicolour localization microscopy by point-spread-function engineering." *Nature photonics*, vol. 10, pp. 590–594, 2016.

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

[223] J. Marco, Q. Hernandez, A. Muñoz, Y. Dong, A. Jarabo, M. H. Kim, X. Tong, and D. Gutierrez, "Deeptof: off-the-shelf real-time correction of multipath interference in time-of-flight imaging," *ACM Trans. Graph.*, vol. 36, pp. 219:1–219:12, 2017.

[224] R. Mantiuk, K. J. Kim, A. G. Rempel, and W. Heidrich, "Hdr-vdp-2: A calibrated visual metric for visibility and quality predictions in all luminance conditions," *ACM Transactions on graphics (TOG)*, vol. 30, no. 4, pp. 1–14, 2011.

[225] R. E. Fischer, B. Tadic-Galeb, P. R. Yoder, and R. Galeb, *Optical system design*. McGraw Hill New York, 2000.

[226] M. J. Allen, "Automobile windshields, surface deterioration," SAE Technical Paper, Tech. Rep., 1970.

[227] A. Flores, M. R. Wang, and J. J. Yang, "Achromatic hybrid refractive-diffractive lens with extended depth of focus," *Appl. Opt.*, vol. 43, no. 30,

pp. 5618–5630, Oct 2004. [Online]. Available: http://ao.osa.org/abstract.cfm?URI=ao-43-30-5618

[228] Z. Liu, A. Flores, M. R. Wang, and J. J. Yang, "Diffractive infrared lens with extended depth of focus," *Optical Engineering*, vol. 46, no. 1, pp. 1 – 9, 2007.

[229] A. Levin, R. Fergus, F. Durand, and W. T. Freeman, "Image and depth from a conventional camera with a coded aperture," *ACM Trans. Graph.*, vol. 26, no. 3, p. 70–es, Jul. 2007.

[230] A. Levin, "Analyzing depth from coded aperture sets," in *Computer Vision – ECCV 2010*, K. Daniilidis, P. Maragos, and N. Paragios, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2010, pp. 214–227.

[231] X. Dun, H. Ikoma, G. Wetzstein, Z. Wang, X. Cheng, and Y. Peng, "Learned rotationally symmetric diffractive achromat for full-spectrum computational imaging," *Optica*, vol. 7, no. 8, pp. 913–922, Aug 2020.

[232] S. Colburn, A. Zhan, and A. Majumdar, "Metasurface optics for full-color computational imaging," *Science Advances*, vol. 4, no. 2, 2018.

[233] S. S. Khan, A. V. R. , V. Boominathan, J. Tan, A. Veeraraghavan, and K. Mitra, "Towards photorealistic reconstruction of highly multiplexed lensless images," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

[234] Y. Wu, V. Boominathan, H. Chen, A. Sankaranarayanan, and A. Veeraraghavan, "Phasecam3d â€" learning phase masks for passive single view depth estimation," in *2019 IEEE International Conference on Computational Photography (ICCP)*. Los Alamitos, CA, USA: IEEE Computer Society, may 2019, pp. 1–12.

[235] A. Kotwal, A. Levin, and I. Gkioulekas, "Interferometric transmission probing with coded mutual intensity," vol. 39, no. 4, Jul. 2020.

[236] Y. Wu, F. Li, F. Willomitzer, A. Veeraraghavan, and O. Cossairt, "Wished: Wavefront imaging sensor with high resolution and depth ranging," in *2020 IEEE International Conference on Computational Photography (ICCP)*, 2020, pp. 1–10.

[237] V. Boominathan, J. K. Adams, J. T. Robinson, and A. Veeraraghavan, "Phlatcam: Designed phase-mask based thin lensless camera," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 7, pp. 1618–1629, 2020.

[238] O. Kupyn, T. Martyniuk, J. Wu, and Z. Wang, "Deblurgan-v2: Deblurring (orders-of-magnitude) faster and better," in *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2019.

[239] G. Côté, J.-F. Lalonde, and S. Thibault, "Extrapolating from lens design databases using deep learning," *Opt. Express*, vol. 27, no. 20, pp. 28 279–28 292, Sep 2019.

[240] ——, "Deep learning-enabled framework for automatic lens design starting point generation," *Opt. Express*, vol. 29, no. 3, pp. 3841–3854, Feb 2021.

[241] D. S. Jeon, S.-H. Baek, S. Yi, Q. Fu, X. Dun, W. Heidrich, and M. H. Kim, "Compact snapshot hyperspectral imaging with diffracted rotation," *ACM Transactions on Graphics (Proc. SIGGRAPH 2019)*, vol. 38, no. 4, pp. 117:1–13, 2019.

[242] S.-H. Baek, H. Ikoma, D. S. Jeon, Y. Li, W. Heidrich, G. Wetzstein, and M. H. Kim, "End-to-end hyperspectral-depth imaging with learned diffractive optics," *arXiv preprint arXiv:2009.00463*, 2020.

[243] C. Zhang, B. Miller, K. Yan, I. Gkioulekas, and S. Zhao, "Path-space differentiable rendering," *ACM Trans. Graph.*, vol. 39, no. 4, pp. 143:1–143:19, 2020.

[244] C. Zhang, L. Wu, C. Zheng, I. Gkioulekas, R. Ramamoorthi, and S. Zhao, "A differential theory of radiative transfer," *ACM Trans. Graph.*, vol. 38, no. 6, pp. 227:1–227:16, 2019.

[245] S. Bangaru, T.-M. Li, and F. Durand, "Unbiased warped-area sampling for differentiable rendering," *ACM Trans. Graph.*, vol. 39, no. 6, pp. 245:1–245:18, 2020.

[246] C. Kolb, D. Mitchell, and P. Hanrahan, "A realistic camera model for computer graphics," in *Proceedings of the 22nd annual conference on Computer graphics and interactive techniques*, 1995, pp. 317–324.

[247] F. A. Jenkins and H. E. White, *Fundamentals of optics*. Tata McGraw-Hill Education, 2018.

[248] Q. Guo, I. Frosio, O. Gallo, T. Zickler, and J. Kautz, "Tackling 3d tof artifacts through learning and the flat dataset," in *The European Conference on Computer Vision (ECCV)*, September 2018.

[249] E. Agustsson and R. Timofte, "Ntire 2017 challenge on single image super-resolution: Dataset and study," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, July 2017.

[250] S. W. Hasinoff and K. N. Kutulakos, "Light-efficient photography," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 11, pp. 2203–2214, 2011.

[251] O. Cossairt, C. Zhou, and S. Nayar, "Diffusion Coding Photography for Extended Depth of Field," *ACM Transactions on Graphics (TOG)*, Aug 2010.

## Publications

This dissertation is based on the following works.

- **Qilin Sun**, C Wang, F Qiang, X Dun, W Heidrich, "End-to-end Complex Lens Design with Differentiable Ray-Tracing" ACM Transactions on Graphics (Proc. SIGGRAPH), 2021

- **Qilin Sun**, Ethan Tseng, Qiang Fu, Wolfgang Heidrich, and Felix Heide "Learning Rank-1 Diffractive Optics for Single-shot High Dynamic Range Imaging" In Proc. Computer Vision and Pattern Recognition (CVPR oral), IEEE, 2020.

- **Qilin Sun**, Jian Zhang, Xiong Dun, Bernard Ghanem, Yifan Peng, and Wolfgang Heidrich " End-to-End Learned, Optically Coded Super-resolution SPAD Camera." ACM Transactions on Graphics, 2020

- Yifan Peng*, **Qilin Sun***, Xiong Dun*, Gordon Wetzstein, Wolfgang Heidrich, and Felix Heide (*Joint first authors) "Learned Large Field-of-View Imaging With Thin-Plate Optics" ACM Transactions on Graphics (Proc. SIGGRAPH Asia), 2019

Follows are publications toward other projects that are not directly related to this dissertation.

- **Qilin Sun**, Xiong Dun, Yifan Peng, and Wolfgang Heidrich "Depth and Transient Imaging with Compressive SPAD Array Cameras" In Proc. Computer Vision and Pattern Recognition (CVPR), IEEE, 2018.

- **Qilin Sun**, Xiong Dun, Yifan Peng, and Wolfgang Heidrich "Optical coding of SPAD array and its application in compressive depth and transient imaging" Optoelectronic Imaging and Multimedia Technology, 2019

- Yifan Peng, Xiong Dun, **Qilin Sun**, Felix Heide, and Wolfgang Heidrich "Focal sweep imaging with multi-focal diffractive optics" Internal Conference on Computational Photography (ICCP), IEEE, 2018.

- Yifan Peng, Xiong Dun, **Qilin Sun**, and Wolfgang Heidrich "Mix-and-match holography" ACM Transactions on Graphics (Proc. SIGGRAPH Asia), 2017.