Advancing Mobile Photography with Under-Display Cameras and Sensor Design

Submitted in partial fulfillment of the requirements for

the degree of

Doctor of Philosophy

in

Department of Electrical and Computer Engineering

Anqi Yang

Carnegie Mellon University

Pittsburgh, PA

June 2024

Acknowledgments

I am deeply grateful for having Prof. Aswin Sankaranarayanan as my PhD advisor, for his invaluable guidance, wise insights and unwavering support throughout my doctoral journey. Aswin's patience and support navigates me through research challenges. I vividly recall that in my third year when the initial attempts at phase mask design for under-display cameras yielded only negative results. Aswin recognized the value in documenting these negative findings and encouraged me to publish a paper on the theoretical aspects. His scientific spirit inspired me to embrace bold exploration and maintain a high standard of innovation, making my doctoral journey an unforgettable adventure. Reflecting on my growth over the past five years, I attribute much of my progress in research and academic skills to the privilege of having such an exceptional advisor.

I want to express my sincere gratitude for my committee members, Prof. Srinivasa Narasimhan, Prof. Vijayakumar Bhagavatula, and Prof. Jinwei Gu, for accomodating their time for the thesis meetings. Their invaluable suggestions significantly enhanced the technical depth of my work. I am also deeply grateful to Prof. Wolfgang Heidrich for attending my thesis defense and for the insightful discussions that followed. His work has been a major source of inspiration for much of my research.

I'm grateful for having Eunhee Kang and Hyong-Euk Lee as collaborators. The insightful discussions during our monthly meetings consistently sparked significant research questions and ideas. I am fortunate to have been surrounded by a group of joyful and inspiring lab members and office mates. Chia-Yin Tsai, Rick Chang, Harry Hui, Jian Wang, and Vishwa Saragadam set a high bar for research excellence. Chia-Yin, Rick, Harry, and Chao Liu offered me invaluable help as I embarked my professional career. Yi Hua, my most artistic friend, infused my PhD life with art and creations through our art store explorations. Wei-yu Chen provided fun research disussions and great help with hardware in the lab. Byeongjoo Ahn and Vishwa Saragadam, my office neighbors since pre-pandemic time, brought laughter and conversation to our shared workspace. I enjoyed many fun conversations with Haejoon about food, traveling, and lighter side of life. And I also deep grateful for the companionship and support from labmates over the years, including Michael De Zeeuw, Leron Julian, Yingsi Qin, Sagnik Ghosh, Tyler Nuanes, Natalie Janosik, Wei Chen, Kuldeep Kulkarni, Vijay Rengarajan, Manu Gopakumar, Jeremy Klotz, Carlos Taveras, and Aparajith Srinivasan. I'm thankful for the friendly help from the extended imaging group at CMU, including Dorian Chan, Benjamin Attal, Bakari Hassan, Shumian Xin, and many others. I extend my heartfelt thanks to John Shi, my very first office mate at CMU, for his enduring friendship and mutual support through the research journey.

Above all, I owe my deepest gratitude to my parents, grandparents and boyfriend Junjiao Tian for the unconditioinal love, caring, and warmness through my graduate study. My mother's boundless curiosity and perseverance in tackling challenges always inspired me. My father shared with me his insightful advice. Junjiao as well as our cat Nuomi accompanied me through demanding paper deadlines, offered solace during research setbacks, and filled my life with joy and laughter. Without the encouragement from them, I can never made through this far.

This thesis was supported by Global Research Outreach program of Samsung Advanced Institute of Technology and the NSF CAREER award CCF-1652569

Abstract

The ubiquity of mobile devices has made mobile photography an indispensable part of our daily life. Unlike standalone cameras, mobile device cameras have to adhere to unique design constraints imposed by the compact form factors and multi-functionality of these devices. In this thesis, we investigate two distinct challenges arising from current mobile device design trend and propose novel camera and sensor designs to address them.

First, the conflict between screen size and camera placement has never been more severe than it is now, driven by the demand for full-screen devices. The prevalence of organic light-emitting diode (OLED) displays, with their partial transparency, offers an exciting opportunity to place a conventional camera beneath the screen, allowing the simultaneous operation of both components.

We study under-display cameras (UDCs), an emerging type of camera that captures a scene through the micron-scale openings of an OLED display panel. Their image quality is hindered by poor signal-to-noise ratio and severe diffractive blur due to the presence of the display. Can we redesign the hardware to improve the overall image quality of UDCs? Based on Fourier optics, we find that the diffractive blur of a UDC is fundamentally determined by the shape of the display opening. Therefore, we propose a suite of modifications to the display layout, including using a random pixel tiling and optimizing the opening shape of each pixel. The proposed method significantly advances image quality by improving the invertibility of the diffractive blur. However, this requires nontrivial display redesign. As a complementary solution, we propose to *optically* modify the display opening shape by adding two phase masks, one in front of and one behind the display. The first phase mask concentrates light onto the display openings, and the other phase mask restores the original wavefront, effectively rendering the display invisible to the camera under certain assumptions. This approach improves UDCs light throughput and the conditioning of the blur, and maintains display quality.

Second, the continuous shrinking of image sensor pixels, with the potential to increase image resolution under a constrained sensor die size, presents challenges. Since small pixels collect less light, they are more susceptible to noise degradation in low-light conditions. Can we design novel computational techniques to combat noise and expand dynamic range of these sensors?

We propose two spatially varying readout techniques that adapt to local scene brightness. The first technique involves spatially varying gain. The key insight is that a larger gain or ISO setting can overcome read noise by amplifying the signal level. Conventional sensors apply a constant gain across the entire frame, limiting the use of a large gain when the scene has a wide dynamic range. In contrast, our

approach adjusts gain at small regions of interest or even individual pixels, allowing a much larger gain for dark regions while avoiding saturation in bright regions, thus effectively expanding the sensor's dynamic range. The second technique is spatially-varying binning. We investigate the optimal pixel size in terms of noise and resolution, and show that the optimal size is tightly coupled with the scene light level. We develop a simple theory that maps scene brightness to optimal pixel size, and implement this varying pixel size through binning. We demonstrate the proposed spatially varying techniques in various applications, including high dynamic range imaging, vignetting, and lens blur, and show consistently improved noise performance and effective resolution.

This thesis takes a leap forward by innovating optics and sensors to address the unique challenges in mobile photography. Interestingly, many of these challenges are fundamentally linked to classic problems in computer vision, such as mitigating blur and noise, and enhancing resolution and dynamic range. We hope that the techniques presented in this work will not only open new avenues for mobile photography but also inspire broader innovation in the field of computational imaging.

For My family

Contents

List of Figures

List of Tables

1 Introduction

Camera design has been extensively studied for over a hundred years. Nowadays, cameras have been an essential tool for diverse fields, ranging from biology, medical imaging, astronomy, to everyday photography. One particularly significant applications is the integration of cameras into mobile devices, such as smartphones, tablets, and augmented reality or virtual reality (AR/VR) headsets. Unlike standalone cameras, the design of mobile device cameras must adhere to the unique form factor constraints due to the emphasis on portability and multi-functionality. The rapid evolution of mobile device form factors necessitates innovative camera designs. In this thesis, we look at two emerging trends in mobile devices, and presents novel camera optics and sensor designs to tackle the challenges presented by these trends.

The first trend is the increasing screen-to-body ratio. As the screen gets larger, the conflict between screen and front-facing camera placement becomes more severe. Most mobile devices punch a hole or use a notch on the top edge of a screen to accommodate the cameras, compromising the aesthetics and functionality of the display. Is there a way to jointly design the display and camera to achieve high quality display and photography at the same time? Previous research has explored display-camera systems. One exciting example is BiDiScreen [Hirsch *et al.*, 2009] that interlaces photodiodes into LCD displays to form a lensless imager and captures images by using the display as an attenuation mask. Since the display has to switch between imaging mode and display mode, the refresh rate is decreased. And the image quality of the lensless camera is significantly inferior to a conventional front-facing camera.

The widespread use of organic light-emitting diode (OLED) displays in the mobile devices offers a new opportunity. As OLED displays do not require a backlit panel and can be made partially transparent, we can place a conventional lens-based camera beneath the display screen and image the scene through the openings of the OLED panel. In the first half of this thesis, we focus on the design of under-display cameras (UDCs). Despite their potential, building high-quality UDCs comes with two significant technical challenges: diffraction and low signal-to-noise ratio. The micron-scale openings on the display diffract light and produce severe diffractive blur in the captured image. We show that it is the shape of

the display openings that determines the diffractive blur, and unfortunately, the diffractive blur from off-the-shelf OLED displays are usually not conducive to invertibility. Moreover, the display panel blocks most of the light and causes extremely low signal to noise ratio, further complicating the deblurring. *Can we jointly design the shape of OLED display opening and the camera post-processing algorithms (deblurring and denoising), such that the restored image quality is better?*

To answer this question, we explore two solutions of redesigning the opening pattern of the displays, one that directly modifies the display layout, and one that *optically* modifies the displays' opening shape by placing phase masks tightly around the display.

First, we demonstrate that innovating the display opening layout can significantly improve the image quality of UDCs. Based on Fourier optics, we find that the diffractive blur of a UDC is fundamentally determined by the shape of the display opening. Therefore, we propose a suite of modifications to the display layout. First, instead of using a repetitive pixel tiling as in conventional displays, we show that a random pixel tiling improves the symmetry and invertibility of point spread functions. Second, we optimize the per-pixel opening pattern to make the point spread function more robust to deblurring. The proposed method achieves better performance than common on-the-market displays. However, changing the display opening would necessitate redesigning RGB display pixels and circuit placement, which requires significant engineering effort.

Second, to avoid directly modifying the display layout, we propose a complementary solution that *optically* modifies the display opening shape by adding two phase masks, one in front of and one behind the display. The phase mask in front of the display concentrates light onto the opening regions of the display. After light passes through the display, the other phase mask reverses this effect by recovering the original wavefront. Under certain assumptions, the display is rendered invisible from the perspective of the camera. We further show that a polarization-dependent implementation of the phase masks can leave light emitted from the display unmodulated. This approach can increase the light throughput of UDCs, improve the conditioning of the blur kernel, and maintain display quality.

In the second half of the thesis, we turn our focus to another design trend in mobile devices — continuous shrinking of pixel sizes enabled by the advancement in CMOS manufacturing capability. When the sensor die size is constrained, smaller pixels are promising to produce higher resolutional imagery. However, they are more susceptible to noise degradation when light condition is poor as smaller pixels collect less light. The increasing degradation from photon, read noise, and dynamic range are tightly coupled with the small pixel area and tackling these degradations has become an open challenge.

There is a rich literature in suppressing noise and expanding the dynamic range of an image sensor. One typical category of work tackles such problems with exposure and gain bracketing. These methods

propose computational methods to merge long exposure frame that suppresses the noise in the dark region and short exposure frame that captures bright regions without saturation. However, they tend to produce ghosting artifacts for dynamic scenes unless using algorithms with heavy computation overhead. Another major category of works look into spatial multiplexing, such as exposure or ISO encoding. Due to fixed multiplexing patterns, these methods trade the spatial resolution for higher dynamic range. More recently, there are computational sensors that adapts per-pixel exposure to the previous frame and captures high dynamic range videos. But these methods require a long exposure time to accommodate the dark regions. *Can we design other spatially-varying techniques that adapts to the brightness of the scene to improve the image quality of these sensors?*

We introduce two novel adaptive readout techniques for image sensors — spatially varying gain and binning. Essentially, all sensor pixels are exposed uniformly, but during readout, we adjust the analog signal of each pixel differently according to the local brightness of the scene. This adjustment is achieved by applying a different ISO settings or combining the signals from varying number of neighbouring pixels. The idea of varying gain is inspired by a key observation that a large ISO setting or gain suppresses read noise by enhancing the signal level. A conventionaly sensor applys a constant gain over the entire sensor, and using a large ISO would risk saturating bright regions of a scene. To address this issue, we adjust ISO settings based on the signals within small regions of interest or even individual pixels. This spatially varying gain technique avoids saturation while allowing a significant larger gain for dark regions of a scene. The motivation behind spatially varying binning is straightforward. Combining signals from neighboring pixels, or binning, improves the noise performance by equivalently creating a large pixel. However, excessive binning leads to pixellation artifacts. We show that the optimal binning size is closely related to scene light levels and develop a simple theory that maps scene brightness to optimal bin size. This theory informs our spatially-varying binning technique, which adapts pixel sizes to the scene local brightness. Both techniques enhance the signal-to-noise ratio of sensors with small pixels and effectively expand the dynamic range by approximately one magnitude.

We summarize the key challenges and contributions of this thesis in the following sections.

1.1 Key Challenges

This thesis looks into novel camera designs pertaining to two form factor constraints in mobile devices. We discuss the technical challenges for these two constraints separately.

Challenges for UDCs. Firstly, to accommodate the conflict between high-quality displays and cameras, we look into under-display camera design. There are two particular challenges in designing UDCs, both of which comes from the fact that the display screen acts as the aperture of the camera. When passing through the openings on the OLED display, incident light are diffracted and a large portion of light are blocked by the display. We elaborate the two challenges in UDCs as below.

- **Diffractive blur.** As the openings on OLED display is often at micron-scale, incident light is diffracted and creates non-negligible blur on the sensor. The smaller the opening is, the larger the diffractive blur is. Pixels with a size of tens of microns will result in a blur that spans hundreds of pixels on the display. And this diffractive blur is challenging to remove. Many state-of-the-art works design advanced neural networks to remove such diffractive blur [Feng *et al.*, 2021, Koh *et al.*, 2022, Kwon *et al.*, 2021, Zhou *et al.*, 2021]. These algorithms perform better than conventional algorithms such as Wiener deconvolution and iterative optimization with traditional priors, however, their performance is still fundamentally decided by the quality of the captured image. Re-designing the hardware of UDCs to improve the conditioning of its blur kernel becomes an open challenge.
- **Low light transmission.** Most on-the-market OLED displays block a large portion of light. For example, 80% of light are blocked after passing through Transparent-OLED and around 92% Pentile-OLED [Zhou *et al.*, 2020a]. The extremely low light transmission rate leaves the signal-to-noise ratio low in UDCs. The extreme amount of noise, coupled with large diffractive blur, make image restoration particularly challenging for UDCs.

Challenges for small sensor pitches. Secondly, mobile devices cameras with small sensors pitches tends to be noisy in low light conditions. This degradation becomes evident in scenarios such as high-dynamic range imaging, pixels towards sensor edges due to vignetting and etc. The degradations are specifically low signal-to-noise ratio and small dynamic range that comes with it.

- **Low SNR.** While small pixel size offer higher image resolution, its signal to noise ratio is worse, producing undesired image quality in dim or low light conditions. The major degradation comes from photon noise and read noise. When the overall noise is photon noise-limited, the signal-to-noise ratio (SNR) decreases with photon counts, and since smaller pixel areas receive fewer photons within a fixed exposure time, they have worse SNR. Although the read noise level stays the same across pixel sizes as it is decided by the read-out circuits' quality, the SNR of small pixels still decreases as the signal level decreases.

- **Small dynamic range.** Dynamic range of an image sensor is typically decided by the ratio between its max well capacity and read noise floor, and since smaller pixels often come with a smaller well capacity, their dynamic range is smaller compared to large pixels.

1.2 Thesis Contributions

Thesis statement. The compact form factor of smartphones poses unique challenges to the design of cameras. This thesis advances mobile photography by tackling two such challenges: (1) To alleviate the conflict between increasing screen space and placing front-facing cameras, we re-design the hardware of under-display cameras by directly and optically modifying the display pixel layouts; (2) To improve the excess noise that comes with small pixel pitches and small dynamic range, we propose spatially-varying readout techniques that adapts ISO and binning to local scene brightness.

- **Designing display layouts for UDCs.** The pattern of openings commonly found in current OLED displays are not conducive to high-quality deblurring. We redesign the layout of openings in the display to engineer a blur kernel that is robustly invertible in the presence of noise. We first provide a basic analysis using Fourier optics that indicates that the nature of the blur is critically affected by the periodicity of the display openings as well as the shape of the opening at each individual display pixel. Armed with this insight, we provide a suite of modifications to the pixel layout that promote the invertibility of the blur kernels. We evaluate the proposed layouts with photomasks placed in front of a cellphone camera, thereby emulating an under-display camera. A key takeaway is that optimizing the display layout does indeed produce significant improvements.

- **Designing phase masks for UDCs.** We incorporate phase masks on display panels to tackle both large diffractive blur and low signal-to-noise ratio in UDCs. Our design inserts two phase masks, specifically two microlens arrays, in front of and behind a display panel. The first phase mask concentrates light on the locations where the display is transparent so that more light passes throughthe display, and the second phase mask reverts the effect ofthe first phase mask. We further optimize the folding height of each microlens to improve the quality of PSFs and suppress chromatic aberration. We evaluate our design using a physically-accurate simulator based on Fourier optics. The proposed design is able to double the light throughput whileimproving the invertibility of the PSFs. Lastly, we discuss the effect of our design on the display quality and show that implementation with polarization-dependent phase maskscan leave the display quality uncompromised.

- **Spatially-varying sensor gain.** This innovation is based on the key insight that a large gain overcome readout noise by magnifying signal level. While applying a large gain uniformly across the entire image sensor would risk saturating the bright regions, we apply varying gain at each pixel location based on local scene brightness. To achieve this, we propose Region-of-Interest (ROI) based and per-pixel based implementation. ROI-based method utilizes a snapshot noisy frame to decided the optimal gain for each region, and apply the gain map in the subsequent capture and readout. Per-pixel implementation only captures and reads out an image once. It estimates the optimal gain of a subsequent pixel based on the readout of the current pixel. Both methods significantly improves the noise performance for dark regions while avoids saturating the bright regions, therefore, effectively expands the dynamic range of a sensor.

- **Spatially-varying sensor binning.** We argue that there exists an optimal binning size that best trades-off image resolution and noise performance, and the optimal binnning is tightly coupled with scene light levels. We develop a theory that maps scene brightness to the optimal pixel binning. We apply this theory to decide pixel size according to local scene brightness and implement this spatially-varying binning using ROI-based method. We analyze three common binning modes, analog average, analog additive, and digital binning. Interestingly, when a larger gain is allowed, digital binning outperforms both analog binning modes.

1.3 Organization

The thesis is organized as the following. In chapter 2, we go over prior research effort that tries to solve the two challenges in mobile photography. To accommodate the conflict between display and cameras, we explore interesting works that integrates cameras into displays, and some early works in the space of under-display cameras. To improve sensor noise performance, we look into computational sensors, that are very relevant to the techniques proposed in this thesis. In chapter 3, we introduce the formal image model of under-display cameras, and show a suite of modifications to the display layouts that improves the imaging quality of UDCs. To avoid the engineering efforts in redesigning the display pixels, in chapter 4, we propose a complementary approach to optically modify the display opening shapes. This is achieved by carefully designing phase masks and tightly place them in front of and behind the display panels. In chapter 5, we turn our attention to computational sensor designs. We introduce two spatially varying readout techniques — varying gain and binning — that improve the noise performance of sensors with small pitches. Finally, chapter 6 concludes this thesis and discuss potential future directions for mobile photography.

2 Background

In this chapter, we go over priors works that are closely related to the technical challenges we laid out in the introduction. We first look into various display-camera systems, which integrates cameras into displays without taking dedicated screen space. One of the most promising display-camera systems is under-display camras, and we explore recent works that design deep-learning based restoration algorithms to recover high-quality photographs from those captured under UDCs. We then explore prior works on computational sensors that can adapt to the content of each scene.

2.1 Display-Camera Systems

There is a rich literature on designs that seek to enhance the capabilities of a display [Masia *et al.*, 2013]; we focus on those that integrate a display with a camera. Early work in this space focuses on camera-projector systems that capture images from behind semi-transparent projection screens. TouchLight [Wilson, 2004] uses stereo cameras to track gestures and DepthTouch [Benko and Wilson, 2009] places a depth camera behind a projection screen to sense objects at different depths. BiDiScreen [Hirsch *et al.*, 2009] instead interlaces image diodes into a thin liquid-crystal display and utilizes the display as a pinhole array to form a lensless imager. However, their design requires displays that have low pixel densities and, further, the imaging quality is significantly worse than webcams and cellphone cameras.

2.2 Advances in Under-Display Cameras

Design of LED displays and under-panel cameras. OLED displays do not require a backlight panel and can transmit light through their transparent substrates and cathodes [Tsujimura, 2017]; this raises the potential for acquiring high-resolution photographs by placing cameras "under" the display panel without sacrificing the quality of displays [Cheng *et al.*, 2019, Emerton *et al.*, 2020, Zhou *et al.*, 2020b]. In particular, there has been a focus on using Transparent-OLED (T-OLED) and Pentile-OLED (P-OLED),

displays that are commonly used in commercial televisions and cellphones. Cheng *et al.* [2019] place a camera behind a T-OLED screen, simulating the point spread function (PSF), and demonstrate that the image quality is significantly worsened due to the diffractive blur introduced by the screen.

Image restoration for under-panel cameras. Recent interest in deploying under-panel cameras in smartphones has spurred interest in techniques that can restore images captured with them. In particular, many recent techniques [Feng *et al.*, 2021, Gao *et al.*, 2021, Koh *et al.*, 2022, Kwon *et al.*, 2021, Lim *et al.*, 2020, Nie *et al.*, 2020, Oh *et al.*, 2021, Sundar *et al.*, 2020, Yang *et al.*, 2020, Zhang, 2020, Zhou *et al.*, 2020a, 2021] use deep neural networks to handle the large blur and low SNR in under-panel imagery. To recover high-quality images, Zhou et al. [Zhou *et al.*, 2020b] exploit convolutional neural networks to deblur and denoise the images captured under both T-OLED and P-OLED; they demonstrate that a deep neural network, with a UNet architecture [Ronneberger *et al.*, 2015] model, produces deblurred photographs that are significantly better than simple techniques like Wiener deconvolution. Emerton et al. [Emerton *et al.*, 2020] propose to tackle degradations from diffraction using structured light with specialized frequencies to illuminate the target scenes; the need for control of scene illumination, unfortunately, places significant constraints on the use of the device. Sundar et al. [Sundar *et al.*, 2020] deblur on low-resolutional images and uses a guided filter network to restore high-resolutioinal images. Puthussery *et al.* [2020] use an encoder-decoder network and adds to each block multiple densely connected convolutional layers with different dilations. Similarly, the best-performing methods in [Zhou *et al.*, 2020a] use a UNet architecture with dense residual blocks [Zhang *et al.*, 2018c] added to each encoding and decoding unit. We refer to the network architecture as UNet-RDB, and train a model based on it for deblurring our captured photographs. Kwon *et al.* [2021] proposed a CNN that takes the degraded images, noise level, and spatially-varying blur kernels as input, and reconstructs sharp images. Feng *et al.* [2021] take into account high dynamic range and saturation, and propose a Dynamic Skip Connection Network to remove diffraction artifacts. These techniques achieve impressive image reconstruction performance for UDCs.

2.3 Computational Sensors

Focal plane sensor processors [Carey *et al.*, 2013, Nguyen *et al.*, 2022, Zarándy, 2011] have become more prevalent recently. Each pixel, containing a processing element next to the photodiode, can take in digital instructions and carry out on-chip analog and digital computation. However, due to the additional circuits at each photosite, these sensors have large pixel pitches and are typically of low resolution and not suitable for high-quality photography. Another type of computational sensor is a programmable

coded exposure sensor [Luo *et al.*, 2017, Sarhangnejad *et al.*, 2019], which adapts the pixelwise exposure time to input control signal. Ke *et al.* [2019] realize scene-adaptive coded exposure by generating exposure codes from the readout of the previous frame. To improve the SNR of the dark regions, these sensors require a longer exposure time. In contrast, our techniques fall into the category of programming gain and binning to increase noise performance, without the need of using a longer exposure time, and only involve modification to the sensor readout.

Another class of techniques work on innovating the analog to digital conversion. For example, Gulve *et al.* [2023] propose a regression-based Flux-to-Digital Conversion in place of conventional analog-to-digital conversion, a new readout strategy that occurs concurrently with exposure and avoids saturating high flux, thereby extending the dynamic range of the sensor. These works provide alternative approaches to the proposed innovation in this paper.

3 Designing Display Pixel Layouts for Under-Display Cameras

Under-display cameras provide an intriguing way to maximize the display area for a mobile device. An under-display camera images a scene via the openings in the display panel; hence, a captured photograph is noisy as well as endowed with a large diffractive blur as the display acts as an aperture on the lens. Unfortunately, the pattern of openings commonly found in current LED displays are not conducive to high-quality deblurring. This chapter redesigns the layout of openings in the display to engineer a blur kernel that is robustly invertible in the presence of noise. We first provide a basic analysis using Fourier optics that indicates that the nature of the blur is critically affected by the periodicity of the display openings as well as the shape of the opening at each individual display pixel. Armed with this insight, we provide a suite of modifications to the pixel layout that promote the invertibility of the blur kernels. We evaluate the proposed layouts with photomasks placed in front of a cellphone camera, thereby emulating an under-display camera. A key takeaway is that optimizing the display layout does indeed produce significant improvements.

3.1 Introduction

Under-display cameras provide a way to maximize the display area on a cellphone. This provides a seamless display without the wastage associated with the bezel or potential distractions such as a "notch" and a "hole punch", thereby enhancing the aesthetics of the device.

The aesthetics achieved by placing the camera beneath the display, however, also degrades the quality of the captured photographs in two distinct ways. First, a significant portion of the incident light from a scene is blocked by the display. In many existing devices, as much as three-fourth of the light is blocked [Tsujimura, 2017] and so the signal to noise ratio (SNR) of the captured photograph can be quite low, except perhaps for the brightest of scenes. Second, in addition to reducing the light levels, the display also acts as an aperture and induces a diffractive blur on the captured measurements. For the

OLED displays used currently in mobile devices, this blur can have a significant spread in hundreds of pixels. Deblurring under such a large blur, especially at low SNRs, is extremely challenging.

This chapter aims to redesign the display layout and, in particular, the pattern of openings through which the under-display camera images the scene. Our goal is to shape the blur point spread function (PSF) so as to improve the conditioning of the ensuing deblurring problem. A basic result from Fourier optics suggests that the PSF observed is the squared magnitude of the scaled Fourier transform of the aperture pattern. Specializing this result to under-display cameras, we show that the periodicity of the pixel layout as well as the specific opening at each display pixel are important factors that determine the robustness of the blur PSF to inversion.

Armed with the insights gleaned from Fourier analysis of under-display cameras, we introduce two key variations in the display layout. First, we argue that avoiding the periodic tiling of the display pixels, and replacing it with a random tiling, whose specifics we describe later, has the effect of reducing anisotropy of the blur. Second, optimizing the shape of the display opening found at a single display pixel can further improve the invertibility of the blur PSF; this optimized pattern is randomly tiled to create the display layout. Together, these innovations provide a rich design space for engineering PSFs that are superior to popularly used P-OLED and T-OLED displays.

Contributions. This chapter advances under-display camera technology by providing optimized display patterns that improve the quality of restored photographs. In this regard, we make the following contributions.

- *Analysis of the PSF.* Using Fourier optics, we analyze the properties of PSF of a camera under a typical OLED display, and connect its spread, periodicity and falloff to the repetitiveness as well as shape of the display openings.
- *Improving PSF conditioning via random tiling.* We propose a simple modification to display layout in the form of a random tiling where each of its pixels is randomly flipped or rotated by 90°. We provide detailed theoretical analysis of this random tiling and show that it improves the robustness of the PSF to inversion.
- *Improving PSF conditioning via optimization.* Finally, we improve the conditioning of the PSF by optimizing the shape of the opening at a pixel, which is kept the same across the display except for the random tiling. We explore two distinct approaches to achieving this: first, optimizing the invertibility of the PSF, and second, end-to-end optimization based on reconstruction performance over a collection of images.

Conventional TOLED display **Conventional POLED display** **Our optimized display layout**

Figure 3.1: **Improvements gained by redesigning the layout of the display pixels.** Shown above are deblurred images from three lab prototypes of under-panel cameras corresponding to (from left to right) TOLED, POLED and proposed display layouts. The insets beside each result shows the corresponding input captured photographs as well as zoomed in regions. All results emulate displays with a resolution of 300 dots per inch. The reader is encouraged to use the zoom tool to explore all three photographs.

Our proposed layouts are optimized with pre-determined constraints on the display LEDs, in terms of their size and pitch and hence, in principle, are realizable with appropriate redesign of the power/control circuitry. The contributions above are analyzed in simulations as well as real results captured from a lab prototype. The code and dataset for this work is publicly available [Yang and Sankaranarayanan, 2021a]. Figure 3.1 shows an example of the improvements that are achieved with this redesign of the display layout. We can immediately observe that the quality of the deblurred photograph is significantly enhanced with the optimized display layout.

Limitations. The contributions above come with certain limitations. Perturbing display layout, especially breaking the periodicity, is likely to complicate the fabrication of the display and the design of the power and control wiring; but given the maturity of CMOS fabrication, we expect this to be an engineering challenge and not a fundamental limitation.

3.2 Related Work

Coded apertures. The idea of using an amplitude mask to code the aperture of a lens has a long history in computational photography, including early work in lensless X-Ray and Gamma ray imaging [Feni-

more and Cannon, 1978]. Recent work in such *coded aperture* cameras has focused on robust estimation of depth from a single [Levin *et al.*, 2007, Zhou and Nayar, 2009] as well as multiple [Zhou *et al.*, 2009] images as well as estimating light fields [Veeraraghavan *et al.*, 2007] from a single coded image. Conceptually, the ideas in this chapter fall firmly under this category of coded aperture cameras. The key difference, however, is in the smallest feature size in the coded aperture. Most prior works operate with openings where the smallest feature is significantly larger, often in hundreds of microns. This permits the use of geometric optics for modeling the effect of the aperture, and further, also implies that a scene that is in focus appears sharp with little or no blur. In contrast, displays have a pixel pitch that is often smaller than 100 μm and hence, the openings have features in the scale of microns, requiring the use of tools from wave optics for modeling and analysis. This also results in a large diffractive blur even for the in-focus scene.

Coded apertures that use phase masks have also found extensive use for similar problems including extended depth of field imaging [Dowski and Cathey, 1995] as well as depth from defocus [Pavani and Piestun, 2008, Wu *et al.*, 2019]. The use of phase modulation requires that these techniques operate under a wave model, and, in this sense, are similar to the techniques used in this chapter. However, the aperture pattern in an under-display camera has to accommodate the OLED array, which necessarily blocks light and often in a periodic pattern; hence, we cannot model the resulting aperture as a pure phase mask.

Hence, both the design of coded apertures as well as the underlying modeling associated with prior work does not easily translate to under-display cameras.

3.3 Under-display Image Formation

In this section, we present the basics of image formation for an under-display camera, focusing specifically on the blur PSF and its relationship to the display layout.

3.3.1 Derivation of the Blur PSF

Setup. Figure 3.2 provides the basic setup of our display-camera system. We assume that the camera lens can be well approximated as a thin lens with focal length f_0 and with an aperture given by $p(x, y)$. The display openings are described using a binary-valued function $o(x, y)$, which is assumed to be collocated with the aperture of the thin lens without any separating distance between them; this assumption greatly simplifies the derivation and is important for analytical reasoning. Finally, the incident light is assumed to be is spatially and temporally incoherent.

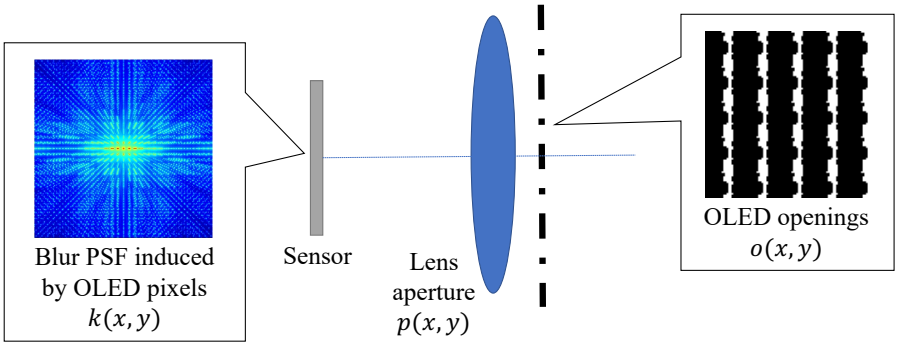


Figure 3.2: Layout of the under-panel camera. The overall aperture consists of a collocated OLED display panel and a finite lens aperture.

Spatially-invariant blur model. Let's suppose that the camera is focused on a scene at infinity. The image formed on the sensor can be written as:

$$i_{\text{blur}}(x, y) = \int_{\lambda} [i_{\text{sharp}}(x, y; \lambda) * k(x, y; \lambda)] s(\lambda) d\lambda,$$

where $*$ denotes the convolution operator, λ is the wavelength of light, $s(\lambda)$ is the camera spectral response, and $i_{\text{sharp}}(x, y; \lambda)$ is the sharp image at wavelength λ that would be formed on the sensor with an ideal thin lens, and without the display. The term $k(x, y; \lambda)$ is the blur kernel at wavelength λ, whose expressions we derive next. It is worth pointing out that the shift invariance as well as lack of interference between scene points is a consequence of the thin lens and the incoherence of light, respectively. The interested reader is referred to Chapter 3 of Goodman [Goodman, 2005].

From basic Fourier optics, the blur kernel $k(x, y; \lambda)$ can be written as the *squared magnitude* of the scaled Fourier transform of the effective aperture function. Specifically, the effective aperture function $a(x, y)$ is the product of the lens aperture $p(x, y)$ and the display openings $o(x, y)$, i.e.,

$$a(x, y) = p(x, y) o(x, y), \tag{3.1}$$

then the blur PSF is given as

$$k(x, y; \lambda) = \left| \frac{1}{j\lambda f_0} A\left(\frac{x}{\lambda f_0}, \frac{y}{\lambda f_0} \right) \right|^2, \tag{3.2}$$

where $A(u, v)$ is the Fourier transformation of the $a(x, y)$.

Specializing to under-display cameras. We now specialize the expression for the blur PSF to features commonly found in an under-panel camera. In an under-display camera, we expect the display openings to be periodic since each display pixel is identical. Let T µm be the pixel pitch of the display; this pitch

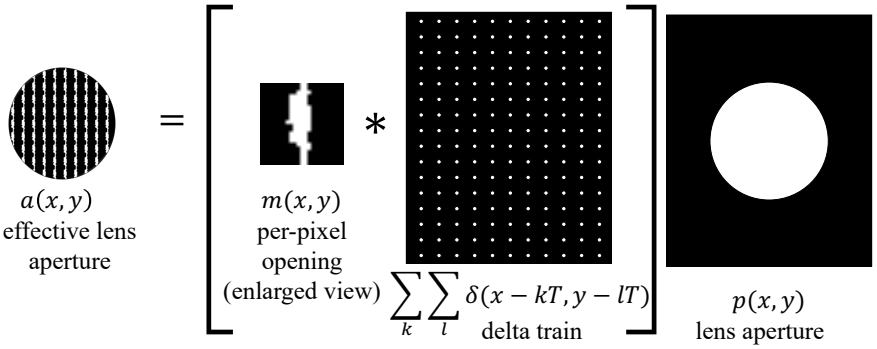


Figure 3.3: Modeling the effective aperture of an under-display camera

also determines the resolution of the display given as $25400/T$ dots per inch (DPI). If we denote $m(x, y)$ to be the opening pattern as pertaining to a single pixel, the overall display openings $o(x, y)$ can be constructed with copies of $m(x, y)$ repeating at a periodicity of T along both axes. As noted in Figure 3.3, we can mathematically express this as follows:

$$o(x, y) = m(x, y) * \sum_r \sum_c \delta(x - rT)\delta\left(y - cT\right), \tag{3.3}$$

where $\delta(x)$ is the Dirac delta function. That is, the display panel is represented as the per-pixel opening $m(x, y)$ convolved with a delta train of periodicity T along both axes.

Noting that the Fourier transform of a delta train with periodicity T µm is also a delta train, but with periodicity $1/T$ µm^{-1} and, further, multiplication in space domain leads to convolution in Fourier domain, we can write the $A(u, v)$, the Fourier transform of the effective aperture, as

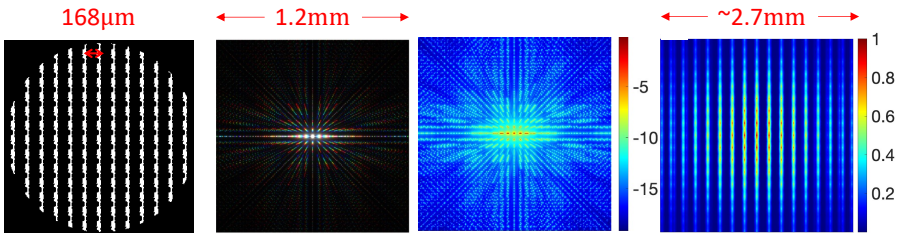
$$\begin{aligned} A(u, v) &= P(u, v) * \left[M(u, v) \sum_k \sum_l \delta\left(u - \frac{k}{T}\right) \delta\left(v - \frac{l}{T}\right) \right] \\ &= \sum_k \sum_l M\left(\frac{k}{T}, \frac{l}{T}\right) P\left(u - \frac{k}{T}, v - \frac{l}{T}\right) \end{aligned} \tag{3.4}$$

The expression above captures the dependence of the blur PSF on all the key terms that define the under-display camera; we analyze this dependence next.

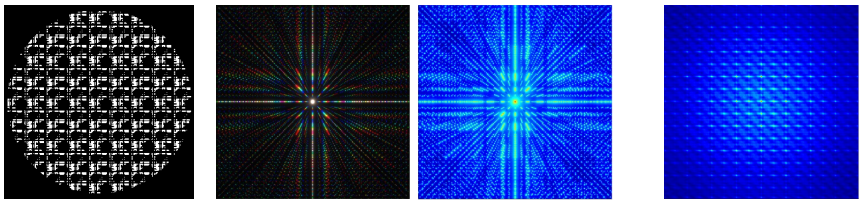
3.3.2 Properties of the Blur PSF

We can analyze the blur PSF derived in (3.2) and (3.4) and derive some of its critical properties. This allows us to make the following observations about the blur PSF.

Periodic sub-structures. The blur PSF is made of repeated copies of $P(u, v)$ — scaled locally by $M(u, v)$. Once we account for the scaling by $1/(\lambda f_0)$ of $A(u, v)$, the periodicity of P is $\lambda f_0/T$. For a display with



(a) T-OLED



(b) P-OLED

Figure 3.4: Two commonly used OLED patterns, (a) T-OLED and (b) P-OLED, and the blur induced by them. For each LED type, we show (left) the display opening pattern, (center-left) the three-color tonemapped PSF, as well as (center-right) the PSF, in log-scale, corresponding to the green channel. The PSF for each color was computed by averaging across multiple wavelengths and weighted by camera spectral response. (right) The Fourier transform of the blur PSF, which is also the scaled auto-correlation of the aperture pattern.

150 DPI, lens focal length $f_0 = 10$mm and wavelength $\lambda = 0.53$µm, this periodicty comes to 32µm, i.e., 10-15 pixels wide on the sensor. In contrast, $P(u, v)$ once scaled by $1/\lambda f_0$ is the airy disk of the open aperture and is constrained to a few pixels. Hence, we can expect the blur PSF to have *sparse* repeating structures, each shaped like an airy disk, as seen in Figure 3.4.

PSF envelope and directionality. While the local structure of the PSF is shaped by P, the overall envelope of the blur kernel, that determines it spread, is primarily determined by $M(u, v)$. This implies that the per-pixel display opening $m(x, y)$ has a dominant role in shaping the blur PSF that we observe. This happens in two distinct ways. First, since $m(x, y)$ is spatially compact and restricted to be within a square of width T μm, where T is the pixel pitch, we can expect its Fourier transform $M(x, y)$ to have a spread that is inversely proportional to T. Second, directionality or anisotropy in the shape of $m(x, y)$ leads to directionality in the shape of $M(u, v)$ and, consequently, in the PSF that we observe; an example of such anisotropy can be seen in the T-OLED display in Figure 3.4. Directional PSF implies that we would preserve detail preferentially in some directions as opposed to others. Hence, all things

considered, displays with larger pixel pitch produce smaller blur and pixel openings that are symmetric produce isotropic blur kernels.

Invertibility. We can analyze the invertibility of the blur PSF by looking at its magnitude spectra, i.e., the magnitude of its Fourier transform. Nulls and small values in the magnitude spectra are undesirable as they lead to noise amplification when we deblur the image.

Connection to the auto-correlation of the aperture. From (3.2), we can express the Fourier transform of $k(x, y; \lambda)$ as

$$K(u, v; \lambda) = \mathcal{AC}_a(\lambda f_0 u, \lambda f_0 v) \tag{3.5}$$

where $\mathcal{AC}_a$ is the auto-correlation function of $a(x, y)$; this expression comes from the fact that the power spectral density and auto-correlation are Fourier pairs, and hence $|A(u, v)|^2$ is the Fourier transform of $\mathcal{AC}_a(x, y)$. Since each color channel is a weighted sum over the visible waveband, the magnitude spectra of the blur in each color channel will be the corresponding weighted sum of $K(u, v; \lambda)$, or equivalently, $\mathcal{AC}_a(\lambda f_0 u, \lambda f_0 v)$. This smoothens the PSF as well as its Fourier transform; however, it does not change the conclusions that we draw below which are based on the monochromatic blur kernel.

Auto-correlation for periodic tiling. The auto-correlation of $a(x, y)$ depends on the lens aperture $p(x, y)$ as well as the per-pixel display opening $m(x, y)$ as given by (3.1) and (3.3). While a general expression for $\mathcal{AC}_a(x, y)$ is hard to derive, we can derive meaningful insights simply by looking at its values for small values of (x, y). Specifically, when the pitch of the display T is significantly smaller than the lens aperture, there are multiple display pixels within the aperture. In this scenario, the auto-correlation $\mathcal{AC}_a$ at small displacements (x, y) becomes repeating copies of $\mathcal{AC}_m$, the auto-correlation of $m(x, y)$, scaled by the number of copies of $m(x, y)$ within the lens aperture. Here, we directly see the effect of the per-pixel pattern $m(x, y)$ and its periodic tiling in the invertibility of the blur PSF. If $m(x, y)$ is compact along any direction, then we can expect the repeated copies of its auto-correlation to not overlap which results in nulls. Further, even if nulls are avoided, decaying tails in the $\mathcal{AC}_m$ would lead to a blur kernel that is not robust to noise. The auto-correlations associated with T/P-OLED displays are shown in Figure 3.4; we can clearly observe the periodic structures with peaks and nulls, as a consequence of the periodicity of the display tiling.

3.4 Rethinking Display Pixel Layout

We now propose new display layouts that are motivated by the analysis laid out in Section 3.3.2. In particular, we are interested in enabling robustly-invertible PSFs by shaping the pixel layout over the

aperture.

Approach. A straightforward approach is to optimize the entire pixel layout over the lens aperture under an appropriate cost on the PSF. However, any solution has to accommodate an LED array with the appropriate resolution and fill factor/LED footprint. While it is possible, in principle, to write out this as a constrained optimization problem, we adopt a different technique that makes the display design significantly simpler.

Our proposed approach relies on two key observations.

- *Random tiling.* First, the periodic tiling of $m(x, y)$ in the display layout causes its auto-correlation to have small values and nulls, which results in a non-invertible blur. This can be alleviated if we tiled the display randomly, where each pixel is randomly chosen between one of many patterns.
- *Optimizing for the per-pixel pattern $m(x, y)$.* Second, we can optimize the shape of the opening at a *single* display pixel with the goal of producing a robust PSF. This pattern is subsequently tiled, with random rotations and flips, to create a random tiling over the aperture of the lens. As a result, the number of parameters to optimize is significantly smaller than what we would have if we optimized for the entire display.

This design methodology has the added advantage of relying on a single per-pixel pattern $m(x, y)$. As long as this pattern permits the LED of a certain footprint, its tiling at the desired DPI, under the random rotation and flip, ensures a feasible LED array over the aperture. Figure 3.5 illustrates how the PSF changes with each of the two modifications.

3.4.1 Random Tiling of the Display Pixel

To understand how random tiling affects the PSF, we will perform the derivation with a 1D display and sensor; the extension to 2D is straightforward and provided in the supplemental material.

1D analysis. Let's first consider a simple 1D display where each pixel is randomly chosen between one of two patterns. Let $m_1(x)$ and $m_2(x)$ be the two potential candidates at each pixel. Also suppose that there are R display pixels over the lens aperture. With this, the overall aperture function $a_r(x)$, including the lens aperture, is given as

$$a_r(x) = \sum_{k=0}^{R-1} \frac{1 + U_k}{2} m_1(x - kT) + \frac{1 - U_k}{2} m_2(x - kT), \tag{3.6}$$

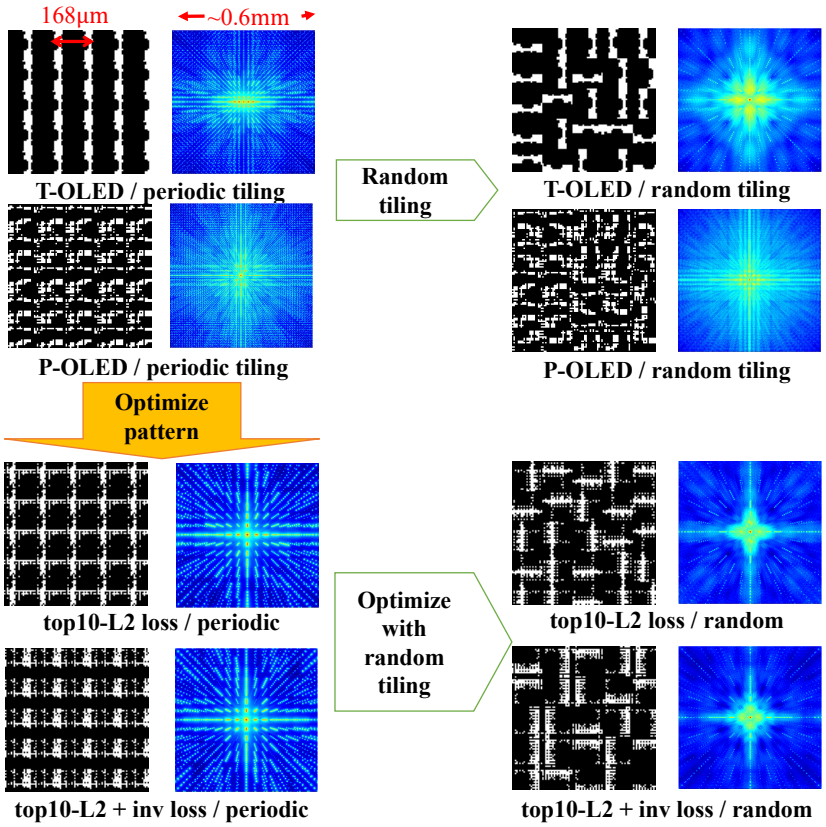


Figure 3.5: We propose to optimize pixel layout by random tiling pixels and optimizing individual pixel openings. The figure above shows how the blur PSF changes when we introduce random tiling without changing the per-pixel pattern (top row), and when we optimize for per-pixel patterns under different criteria (bottom row), both with and without random tiling.

where $\{U_k, 0 \leq k \leq R-1\}$ are iid Bernoulli random variables taking values in $\{+1, -1\}$ with equal probability; hence, U_k selects between m_1 and m_2 at the k-th pixel. By rearranging the terms involving U_k in (3.6), we get

$$a_r(x) = b_1(x) * \sum_{k=0}^{R-1} \delta(x - kT) + b_2(x) * \sum_{k=0}^{R-1} U_k \delta(x - kT), \tag{3.7}$$

where

$$b_1(x) = [m_1(x) + m_2(x)]/2, \text{ and } b_2(x) = [m_1(x) - m_2(x)]/2.$$

The term involving $b_1(x)$ is similar to the display model from before, namely, tiled copies of a pattern over the lens aperture. However, this is now modified by the second term whose effect can be studied

next.

From basic signal processing, we can write the Fourier transform of the aperture a_r as

$$A_r(u) = B_1(u) \sum_{k=0}^{R-1} e^{-j2\pi ukT} + B_2(u) \sum_{k=0}^{R-1} U_k e^{-j2\pi ukT}.$$

When the pixel pitch T is much smaller than the lens aperture, or equivalently when R is large, the observed PSF is well approximated by the *expected* value of $|A_r(u)|^2$, which can be expressed as

$$\mathrm{E}[|A_r(u)|^2] = |B_1(u)\Delta(u)|^2 + R|B_2(u)|^2. \tag{3.8}$$

In essence, when we randomly tile two patterns m_1 and m_2, the expected blur PSF is the sum of two terms: the first term $|B_1(u)\Delta(u)|^2$ that corresponds to a periodic tiling of $m_1 + m_2$ over the aperture of the lens, and the second term that is simply R times the Fourier transform of $m_1 - m_2$.

Further, as before, we can analyze the Fourier transform of the blur for invertibility and robustness. Here, there are two terms: first, the auto-correlation of $(m_1 + m_2)/2$ with its periodic tiling over the aperture — this has a behavior similar to what we get with a conventional tiled display, however, $(m_1 + m_2)/2$ is more isotropic than a single m_1 or m_2; and second, the auto-correlation of $R(m_1 - m_2)/2$ — without any tiling — which stabilizes the PSF.

Extending analysis to the 2D case. To extend the analysis to the 2D display case, we need to first identify the number of patterns that we choose from. While this is something we can choose freely, there are advantages to having a single pixel layout and simply rotating / flipping it. As a consequence, there are four distinct patterns that can appear in any pixel: the unperturbed pattern, the pattern under a 90° rotation, the pattern under a flip, and finally, the pattern under both operations.

We provide a detailed derivation of the observed PSFs of a 2D display with randomly tiled pixels. At each location, we have four candiate patterns: (1) the original pixel pattern (2) the flipped pixel pattern (3) orginal pixel rotated by $\frac{\pi}{2}$ and (4) orginal pixel rotated by $\frac{\pi}{2}$ and flipped, denoted as $m_1(\vec{x})$, $m_2(\vec{x})$, $m_3(\vec{x})$, $m_4(\vec{x})$ respectively. We have two Bernoulli random variables U_{kl}, Q_{kl} that both take values $\{-1, +1\}$ with equal probabilities. Each pixel repeat R times along both directions, and the overall aperture function $a_r(\vec{x})$ including the finite lens aperture can be written as

$$
\begin{aligned}
a_r(\vec{x}) =& m_1(\vec{x}) * \sum_{k=0}^{R-1}\sum_{l=0}^{R-1} \frac{1+U_{kl}}{2}\frac{1-Q_{kl}}{2}\delta(x-kT)\delta(y-lT) + \\
& m_2(\vec{x}) * \sum_{k=0}^{R-1}\sum_{l=0}^{R-1} \frac{1-U_{kl}}{2}\frac{1-Q_{kl}}{2}\delta(x-kT)\delta(y-lT) + \\
& m_3(\vec{x}) * \sum_{k=0}^{R-1}\sum_{l=0}^{R-1} \frac{1+U_{kl}}{2}\frac{1+Q_{kl}}{2}\delta(x-kT)\delta(y-lT) + \\
& m_4(\vec{x}) * \sum_{k=0}^{R-1}\sum_{l=0}^{R-1} \frac{1-U_{kl}}{2}\frac{1+Q_{kl}}{2}\delta(x-kT)\delta(y-lT).
\end{aligned}
$$

We subtitude the following equations into $a_r(\vec{x})$,

$$
\begin{aligned}
b_1(\vec{x}) =& \frac{1}{4}(m_3(\vec{x}) - m_4(\vec{x}) - m_1(\vec{x}) + m_2(\vec{x})) \\
b_2(\vec{x}) =& \frac{1}{4}(m_3(\vec{x}) - m_4(\vec{x}) + m_1(\vec{x}) - m_2(\vec{x})) \\
b_3(\vec{x}) =& \frac{1}{4}(m_3(\vec{x}) + m_4(\vec{x}) - m_1(\vec{x}) - m_2(\vec{x})) \\
b_4(\vec{x}) =& \frac{1}{4}(m_3(\vec{x}) + m_4(\vec{x}) + m_1(\vec{x}) + m_2(\vec{x}))
\end{aligned}
$$

and rewritten the overall aperture function $a_r(\vec{x})$ as

$$
\begin{aligned}
a_r(\vec{x}) =& b_1(\vec{x}) * \sum_{k=0}^{R-1}\sum_{l=0}^{R-1} U_{kl}Q_{kl}\delta(x-kT)\delta(y-lT) + \\
& b_2(\vec{x}) * \sum_{k=0}^{R-1}\sum_{l=0}^{R-1} U_{kl}\delta(x-kT)\delta(y-lT) + \\
& b_3(\vec{x}) * \sum_{k=0}^{R-1}\sum_{l=0}^{R-1} Q_{kl}\delta(x-kT)\delta(y-lT) + \\
& b_4(\vec{x}) * \sum_{k=0}^{R-1}\sum_{l=0}^{R-1} \delta(x-kT)\delta(y-lT).
\end{aligned}
$$

Take the Fourier transform of $a_r(\vec{x})$, we obtain

$$
\begin{aligned}
A_r(\vec{u}) = & B_1(\vec{u}) \sum_{k=0}^{R-1}\sum_{l=0}^{R-1} U_{kl}Q_{kl}e^{-j2\pi(uk+vl)T} + \\
& B_2(\vec{u}) \sum_{k=0}^{R-1}\sum_{l=0}^{R-1} U_{kl}e^{-j2\pi(uk+vl)T} + \\
& B_3(\vec{u}) \sum_{k=0}^{R-1}\sum_{l=0}^{R-1} Q_{kl}e^{-j2\pi(uk+vl)T} + \\
& B_4(\vec{u}) \sum_{k=0}^{R-1}\sum_{l=0}^{R-1} e^{-j2\pi(uk+vl)T},
\end{aligned}
$$

where $B_i(\vec{u}), i = 1, ..., 4$ is the Fourier transform of $b_i(\vec{x})$. Similar to 1D case, the observed PSF can be approximated by the *expected* value of $|A_r(\vec{u})|^2$. Since $E[U_{kl}] = 0$ and $E[Q_{kl}] = 0$, all cross terms in $|A_r(\vec{u})|^2$ is cancelled out. The observed PSF can be written as

$$
\begin{aligned}
& |A_r(\vec{u})|^2 = R^2(|B_1(\vec{u})|^2 + |B_2(\vec{u})|^2 + |B_3(\vec{u})|^2) + |B_4(\vec{u})|^2|\Delta(\vec{u})|^2 \\
& \text{where } \Delta(\vec{u}) = \mathcal{F}\{\sum_{k=0}^{R-1}\sum_{l=0}^{R-1}\delta(x - kT)\delta(y - lT)\}.
\end{aligned}
$$

As with the 1D case, the key observation is that the PSF for a randomly tiled display is made of two terms: a term that corresponds to periodic tiling and a second term that is non-repetitive.

Evaluating the efficacy of random tiling. Figure 3.5 shows how the PSF of the T-OLED and P-OLED pattern changes when we subject it to random tiling. We can observe that both the periodic sub-structures as well as the anisotropy of the original PSF are reduced significantly. While we provide a detailed quantitative evaluation of random tiling in Section 3.5.

Figure 3.6 shows how the auto-correlation and modulation transfer function (MTF) of the aperture changes when we introduce random tiling; recall that the auto-correlation is the Fourier transform of the blur PSF, and horizontal MTF corresponds to the amplitude of a slice of auto-correlation function along x-axis and intercepting y-axis at DC component. Repeatedly tiled T-OLED has a horizontal MTF that contains periodic structures. Contrast at valleys is close to zeros and features at these frequencies are extremely hard to recover. As expected, random tiling serves to smoothen the peaks and valleys of the auto-correlation function and thus raises the lower envelope of MTF. Although the vertical MTF of T-OLED becomes worse due to pixel rotations, horizontal MTF is significantly better with all the zero contrast eliminated and thus the overall PSF becomes much more robust to invert. Similar observations apply to random tiled P-OLED and our optimized patterns.

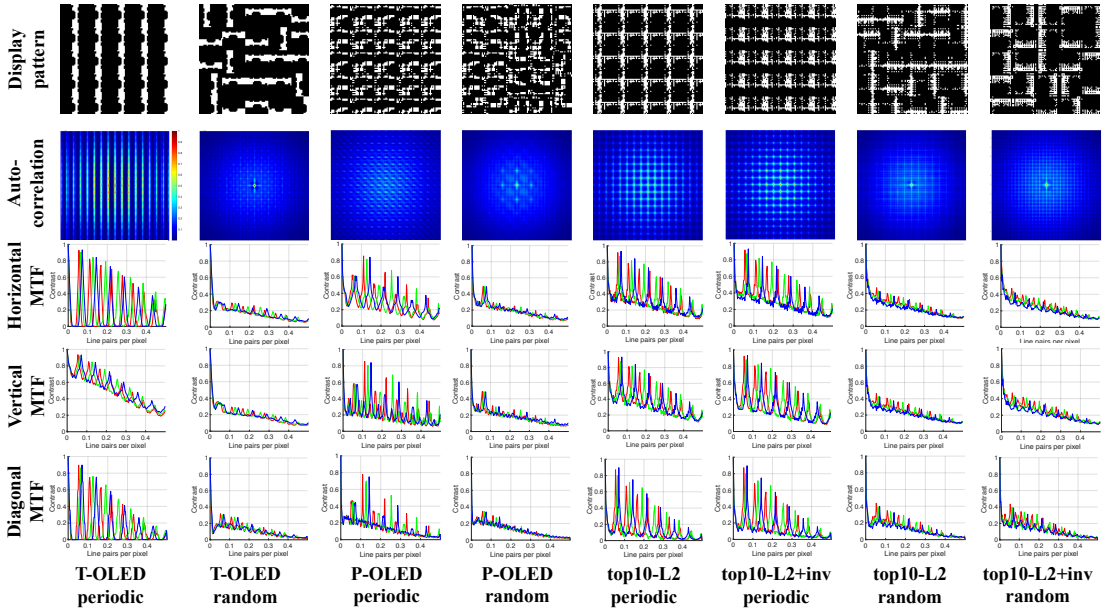
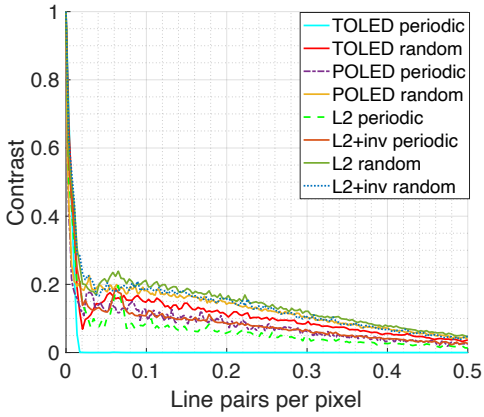


Figure 3.6: **Effect of random tiling and pixel shape optimization.** In the left four columns, we show the effect of random tiling to two common displays. In the right four columns, we show our optimized pixel opening shapes from two losses, top-10 L2 and top-10L2+invertible loss, and each with two tiling strategies during optimization.

Each curve in Figure 3.7 is generated by taking the radially minimum values of the 2D MTF. The plots are generated for a single wavelength $\lambda = 610$nm. As expected, repeatedly tiled T-OLED has multiple null values in many frequencies and randomly tiled T-OLED lifts the nulls to small values and thus stabilizes the inversion. For P-OLED pattern, a randomly-tiled aperture induces larger minimum values for all frequencies than that of the repeatedly tiled aperture and thus is more robust to inversion. Figure 3.7 also provides quantitative comparison for different display layouts by summarizing the area under the radially minimum MTF curves (AUC) of each display and listing the corresponding light transmittance rate (LTR). We can clearly see that randomly-tiled displays have higher AUCs than their periodically tiled counterparts, again indicating that random tilings are more robust to inversion.

3.4.2 Optimizing for the Per-Pixel Pattern

While random tiling provides improvement over a periodic one, we can further improve the efficacy of the blur PSF by designing the per-pixel opening $m(x, y)$. We formulate this as an optimization problem where seek a desirable PSF, as characterized by a loss/cost function, and optimize for the per-pixel



	AUC	LTR(%)
TOLED / periodic	2.43	20.7
TOLED / random	16.47	20.8
POLED / periodic	12.44	23.7
POLED / random	18.98	23.8
L2 / periodic	10.26	23.0
L2 + inv / periodic	13.51	22.7
L2 / random	21.98	22.9
L2 + inv / random	20.55	22.6

Figure 3.7: **Comparison of MTF plots.** We compare radially min MTFs of different patterns and the table summarizes the area under the MTF curve (AUC) and light transmittance rate (LTR) for each mask. Larger AUC is better.

pattern $m(x, y)$ that minimizes this cost. It is worth reemphasizing that we only optimize for a single pattern $m(x, y)$; the display layout is constructed using periodic or random tiling depending on the specifics of the design.

Optimization setup. We discretize the variable $m(x, y)$ into a 2D matrix $\vec{m} \in R^{N \times N}$. For a display pixel, each point takes value in $\{0, 1\}$, where 0 indicates closed regions that contains LEDs, control circuits and etc. and 1 represents pixel openings that allow light to comes through. Since optimizing over a binary-valued variable is not easily amenable to standard descent-based optimization, we relax $\vec{m}$ to be real-valued, and instead map its elements to $[0, 1]$ using a sigmoid function $g(\vec{m})$ to get a mask. The PSF $A(\cdot)$ is a function of the mapped pixel opening $g(\vec{m})$, as well as a number of fixed parameters that include the display pixel pitch T, the specifics of the random tiling (if used), and other parameters such as f_0 and λ.

We seek an optimal pixel opening $\vec{m}$ whose PSF is invertible with respect to Wiener deconvolution. We choose Wiener deblurring instead of a deep neural network (DNN) during optimization as in previous PSF-engineering works[Chang and Wetzstein, 2019, Metzler *et al.*, 2020, Sun *et al.*, 2020] for the following reasons. First, a simpler algorithm like Wiener deconvolution puts the emphasis entirely on the system conditioning, in terms of a mask that produces an invertible blur, so that the inadequecies of the mask are not suppressed by a powerful inverse algorithm. Second, by doing this, we also avoid being biased to specifics of the DNN that is used, and the data used in the process of training both the DNN and

our technique. Finally, Wiener deconvolution is *blindingly* fast which is very helpful in the context of optimization.

Loss #1 —- PSF-induced loss. When using Wiener deconvolution, the estimated deblurred image $\widehat{I}_{\text{sharp}}$ can be written in terms of the *ground truth* image I_{sharp} as follows,

$$\widehat{I}_{\text{sharp}} = H(\vec{m}) I_{\text{sharp}},$$

where

$$H(\vec{m}) = \frac{A(g(\vec{m}))A(g(\vec{m}))^*}{|A(g(\vec{m}))|^2 + \epsilon}. \tag{3.9}$$

Here, $H(\vec{m})$ is the overall system frequency response that characterizes blurring and recovering the desired image and it is a function of $A(g(\vec{m})$, the Fourier transform of PSF induced by the pixel opening matrix $g(\vec{m})$. When $H(\vec{m})$ is the identity operator, we obtain $\widehat{I}_{\text{sharp}} = I_{\text{sharp}}$ and, in theory, we can perfectly recover the sharp image. Hence, a good metric for optimization is to maximize the smallest value of $H(\vec{m})$. In practice, instead of taking the smallest value of $H(\vec{m})$, we take the average of the smallest thirty percent of the elements of $H(\vec{m})$ to improve the robustness of optimization. We vectorize and sort the values in $H(\vec{m})$ so that $\{H(\vec{m})\}_1 \leq \{H(\vec{m})\}_2 \leq \ldots \leq \{H(\vec{m})\}_{N^2}$. The invertible loss is defined as

$$\mathcal{L}_{\text{inv}} = -\frac{1}{N'} \sum_{i=1}^{N'} \{H(\vec{m})\}_i.$$

We use $N' = \lceil 0.3N^2 \rceil$ in all our optimizations.

The loss function $\mathcal{L}_{\text{inv}}$ is closely related to the work of Mitra et al [Mitra *et al.*, 2014], where the performance of various computational imaging systems are analyzed; here, given a forward operator A, the term $\text{Tr}(\text{inv}(A^\top A))$ is to model and analyze system conditioning. For the convolutional model, $(A^\top A)^{-1}$ is closely related to the Wiener filter. However, there are some differences in how we define the loss function. While $\text{trace}(\text{inv}(A^\top A)$ minimizes MSE, we observe that the inverse filter is *only* unstable at a small number of Fourier coefficients, and so we only optimize the worst ten percent of filter coefficients which prioritizes worst-case performance as opposed to average.

Loss #2 — Data-driven loss. While PSF-induced loss provides a data agnostic metric, deblurring performance on actual images is often the gold metric. Hence, over a small dataset of images [Zhou *et al.*, 2020a], we minimize the error between the ground-truth images and corresponding deblurred images, obtained from the Wiener deconvolution technique. Since flat regions in the image are easy to recover, the loss is dominated by the easy samples. We use hard-sample mining to penalize the largest reconstruction errors [Shrivastava *et al.*, 2016]. Specifically, we compute the residuals $\Delta I = I - \hat{I}$, vectorize

and rank the absolute values of the residuals $|\Delta\vec{i}_r|$ in descending order, take the top 10% residuals to compute L2 loss. Top-10 L2 loss is formulated as

$$\mathcal{L}_{\text{data}} = \sum_{r=1}^{R} |\Delta\vec{i}_r|^2,$$

where R is the number of 10% elements in the current batch.

Choice of tiling. In addition to the loss functions, we also have different choices in how we tile the per-pixel pattern $m(x, y)$, that we optimize for, to create the lens aperture function. The standard tiling creates a periodic pattern by repeating the pattern $m(x, y)$ till it covers the aperture of the lens. We also have the choice of random tiling where the pattern is randomly flipped and rotated to create the aperture pattern. An important point is that the sequence of random flips and rotations are randomly generated once and fixed; at optimization time, the pattern $m(x, y)$ is optimized under this specific tiling.

Target function. Combining abovementioned losses, our target function is formulated as

$$\arg\min_{\vec{m}} \alpha_{\text{inv}}\mathcal{L}_{\text{inv}} + \alpha_{\text{data}}\mathcal{L}_{\text{data}} + \alpha_{\text{area}}\left|\vec{1}^T g(\vec{m})\vec{1}/N^2 - c\right|^2.$$

The last term constraints the total opening area of the mapped pixel opening $g(\vec{m})$ to be around target ratio c. We optimize for individual pixel $m(x, y)$ of $T = 168$µm in x,y directions and under the constraint that 20% of pixel region is open, i.e. $\frac{1}{T^2}\int_{x=0}^{T}\int_{y=0}^{T} m(x, y) = 0.2$, which are parameters of a typical T-OLED pixel. We discretize the pixel into 21×21 2D matrix, i.e. $N = 21$, where each element represents a dot of width and length of 8µm. We fix the focal length as $f = 10$mm, aperture as $f/2.5$ and use wavelength of 610nm, 530nm, and 470nm. We use stochastic gradient descent with learning rate 1 and optimize for 150 epochs. The pixel opening matrix m is initialized as an all-one matrix, i.e. the pixel is all open. In the first iteration, we set the area constraint as $c = 1$ and gradually decrease it by 0.05 every five epochs until $c = 0.2$.

Optimized layouts. For purposes of evaluation, we generate four distinct display layouts by thresholding the display pattern to binary values $\{0, 1\}$ and keep the target opening area, which are shown along with their PSFs in Figure 3.5. The four sets corresponds to two distinct loss functions — top-10 L2 loss, and top-10 L2 + invertible loss — and two kinds of tiling — periodic and random. We will visualize the corresponding optimized patterns and demonstrate their performance in the simulated and real experiment sections.

3.5 Simulated Experiments

To evaluate the performance of our techniques, we quantitively compare the recovered images generated by simulating capture behind different display patterns.

Simulation setup. We utilize thirty images provided in [Zhou *et al.*, 2020b] validation set to generate degraded and ground-truth image pairs. The degraded image is generated by convolving a ground-truth image with our simulated PSFs and then adding shot noise and read-out noise to the blurry images according to the parameters of a typical cellphone camera. Specifically, we use a full well capacity of 15506 electrons and a standard deviation of read-out noise of 4.87 electrons. We simulate five different light levels with SNR varying from 24dB to 40dB and corresponding maximum number of electrons varying from 270 to 10000 (not exceeding full well capacity). To recover sharp images, we first denoise the degraded images with BM3D [Dabov *et al.*, 2007] and then deblur them using Wiener deconvolution. We measure the quality of deblurred images by comparing them with corresponding ground-truth images and compute PSNR and SSIM [Wang *et al.*, 2004].

Effect of random tiling. We first look at the effect of introducing random tiling to existing display patterns without altering the shape of individual pixel openings. Figure 3.8 reports PSNR and SSIM numbers as a function of measurement noise levels. For the T-OLED display as well as optimized ones, introducing random tiling provides improvements in both metrics; for T-OLED this improvement is very significant due to inherent anistropy of the pattern.

Effect of pixel shape optimization. We compare existing display pixels with our optimized ones. In the last four columns in Figure 3.6, we show optimized patterns from two losses, top-10 L2 loss and top-10 L2 + invertible loss, and for each of them we show two tiling strategies — periodic repeating and random tiling that is chosen and fixed prior to optimization. During testing, we use corresponding tiling strategies to form display panels. During optimization, we use 240 images from the training set of [Zhou *et al.*, 2020b] to compute the top-10 L2 loss. As shown in Figure 3.8, optimizing pixel shape with periodic repeating improves the reconstruction quality by a large margin compared to the conventional T-OLED. Incorporating random tiling as described above leads to additional improvements. The pattern optimized with "top-10 L2 + invertible" loss with random tiling has the best performance and achieves more than 8dB increase in PSNR and around 0.11 increase in SSIM over T-OLED.

Effect of display pixel density. We show the quality of reconstructed images of conventional patterns and optimized patterns under different pixel densities. For the same pixel pattern, a higher pixel density

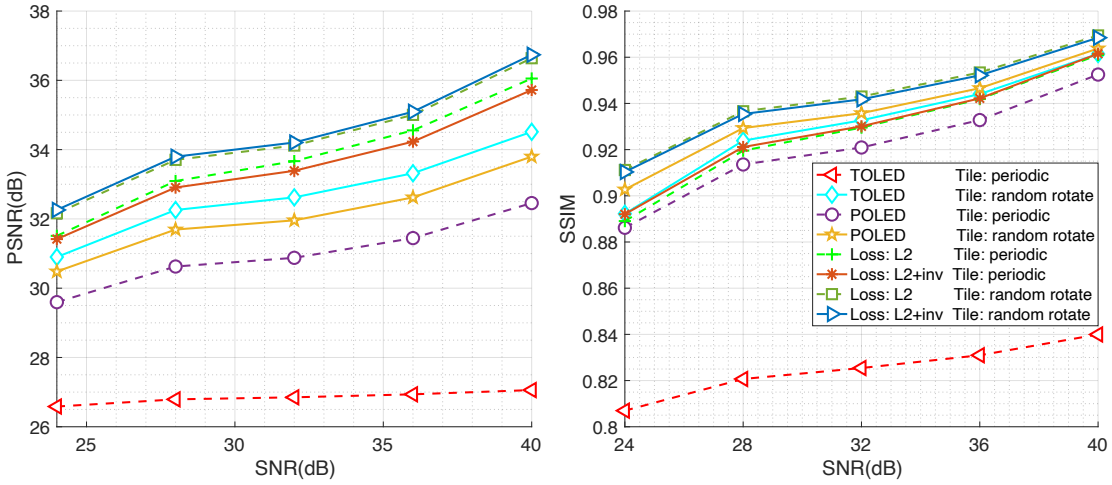


Figure 3.8: **Performance of random tiling and pixel shape optimization.** We compare six display layouts on the simulated dataset and evaluate PSNR and SSIM under varying noise levels.

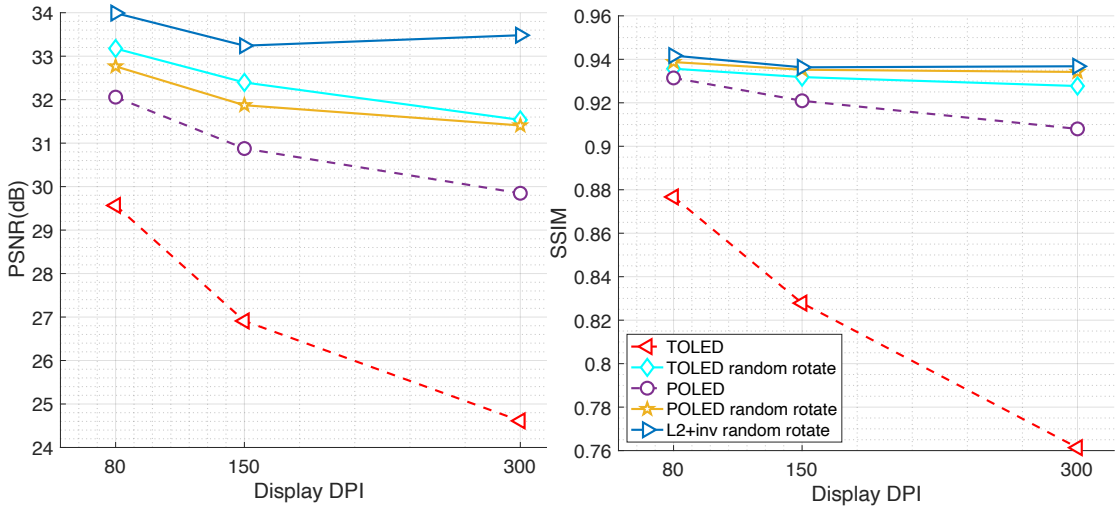


Figure 3.9: **Effect of display pixel density.** We compare four pixel openings/layouts under varying display densities. The horizontal axes are display Dot-Per-Inch(DPI), and the vertical axes are PSNRs and SSIMs of the reconstructed images under a fixed noise level.

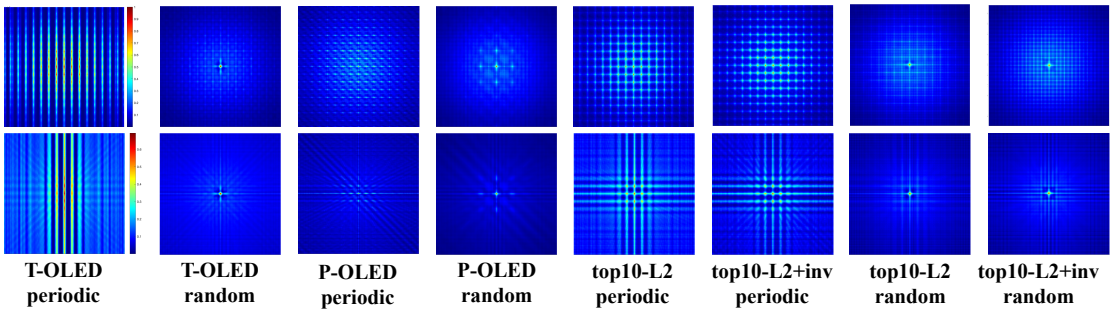


Figure 3.10: **Autocorrelation functions of mono- and multi-wavelengths blur kernels.** The first row shows autocorrelations of the green channel of the blur kernels computed from peak wavelength 0.53µm. The second row shows autocorrelations of the corresponding blur kernels computed from weighted sum of multiple wavelengths.

results in a larger blur kernel and thus is harder to recover a sharp image. We fix the noise level to be SNR= 32 dB and vary pixel density from 80 DPI to 300 DPI. As seen in Figure 3.9, the recovered image quality decreases as pixel density increases and the optimized patterns outperform two common pixels T-OLED and P-OLED under all pixel densities. Note that the improvement on P-OLED is not as significant as on T-OLED.

Mono- vs Multi-wavelengths Simulation. Recall that we design our mask under a tri-chromatic model on the spectrum of the scene, where-in each color channel is simulated as being monochromatic at the color channel's peak wavelength. Here, we test the implications of violations of this assumption, as is wont in real life, using simulations.

We compare the green channel of blur kernels computed from peak wavelength $k(x, y; \lambda)$ with $\lambda = 0.53\ \mu$m and from weighted sum of multiple wavelengths in the spectra band $k'(x, y) = \int_\lambda k(x, y; \lambda)s(\lambda)d\lambda$, where $s(\lambda)$ is the green sensor spectral response. $k'(x, y)$ is a weighted sum over blurs in different wavelengths $k(x, y; \lambda)$, and its Fourier transform $K'(u, v)$will also be corresponding weighted sum of $K(u, v; \lambda)$. As mentioned before, this weighted sum smoothens both the blur kernels as well as their Fourier transform, i.e. autocorrelation functions. As expected in Figure 3.10, Fourier transform of the multi-wavelengths blur kernels $K'(u, v)$ (second row) are smoothened Fourier transform of mono-wavelength kernels $K(u, v, \lambda)$ (first row). The smoothening effect becomes more apparent as the frequency magnitude increases. And at small values of (u, v) the two autocorrelations are consistent. Moreover, we observe simular effects for periodic tiling and random tiling in the multi-wavelength autocorrelation functions. For T/P-OLED and our optimized patterns, periodic tiling has severe repetitions

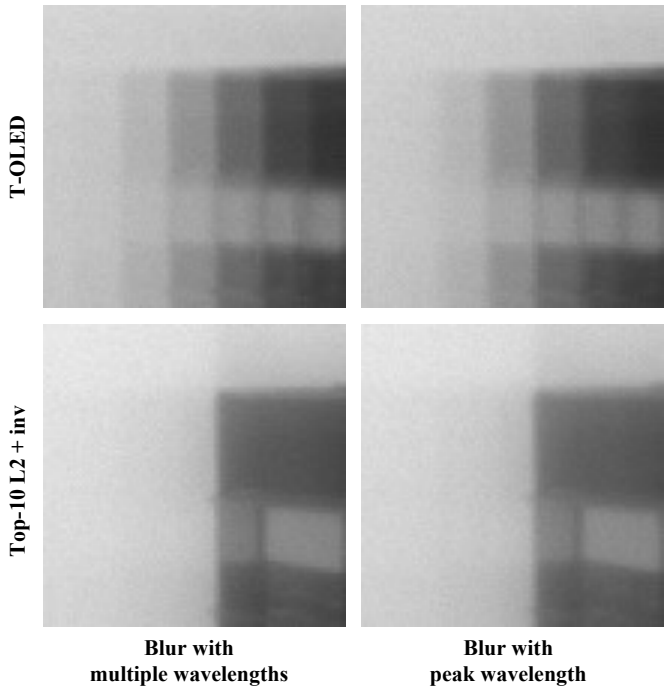
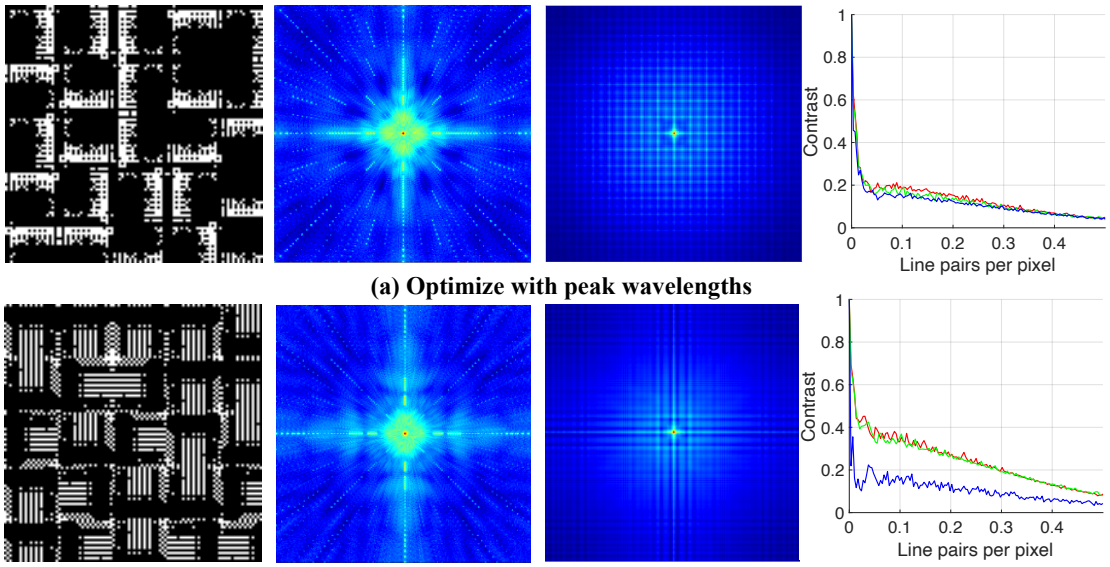


Figure 3.11: We compare to simulate blur kernels using the peak wavelength for green channel and averaging multiple wavelengths for green channel. We showcase simulated blurry image under periodic T-OLED display and randomly tiled top-10 L2 + inv display.

at small (u, v) and thus contains multiple valleys that are unstable to invert. Random tiling effectively eliminates the repititions by smoothening the peaks and valleys and the blur kernels are more robust to invert.

In Figure 3.11, we showcase a degraded image blurred by mono- and multi-wavelength blur kernels. The first row shows blurry images of periodic T-OLED display and the second row shows that of randomly tiled optimized patterns. Blurry images from mono- and multi-wavelength kernels of the same display have very similar extent of blur.

Optimization for multi-wavelengths. Finally, we test the changes in the PSF and MTF when the mask is designed under a denser sampling of wavelengths. To show the effect of optimizing blur kernels towards densely sampled wavelengths, we compute PSFs for five wavelengths around the peak wavelength of each color channel and then average these PSFs to obtain the blur kernel for one sensor channel. Using this new more realistic forward model, we optimize for the mask pattern under the "top-10 L2 loss and invertible loss" and the same parameters as described before. Figure 3.12 compares



(a) Optimize with peak wavelengths

(b) Optimize with multiple wavelengths

Figure 3.12: **Optimization for peak- versus multi-wavelengths PSFs.** The left column shows the optimized display patterns, the center-left and center-right columns show their PSFs and auto-correlation functions, and the right column shows the radially-min MTFs for R,G,B channels.

the optimized pattern, corresponding PSF, auto-correlation function, and radially-min MTFs with the results optimized for peak wavelengths. While optimizing for multiple wavelengths improves the MTFs especially for red and green channels, it leads to patterns that are not conducive to stacking OLED subpixels. During optimization, sampling multiple wavelengths requires more than 3× optimization time than sampling the peak wavelengths.

Simulated results at 150DPI. We show deblurred results from simulated captured images under six different display layouts — T-OLED, top-10 L2 optimized, top-10 L2+inv optimized, and each type with periodic and random tiling. We simulate displays that have 150 DPI and we use camera parameters $f = 10$ mm and aperture size $D = 2.5$ mm as in the simulated setup. We use BM3D [Dabov *et al.*, 2007] to denoise and then apply Wiener deconvolution to deblur the images. As shown in Figure 3.13, T-OLED yields severe one-directional artefacts, while randomly tiled T-OLED eliminates these vertical artefacts. Note that random tiled T-OLED images have low contrast compared to that of optimized patterns. Two optimized patterns with periodic tiling have better contrast but have apparent rinigng artefacts around sharp edges. Optimized patterns with random tiling yields the best performances with high contrast and less artefacts.

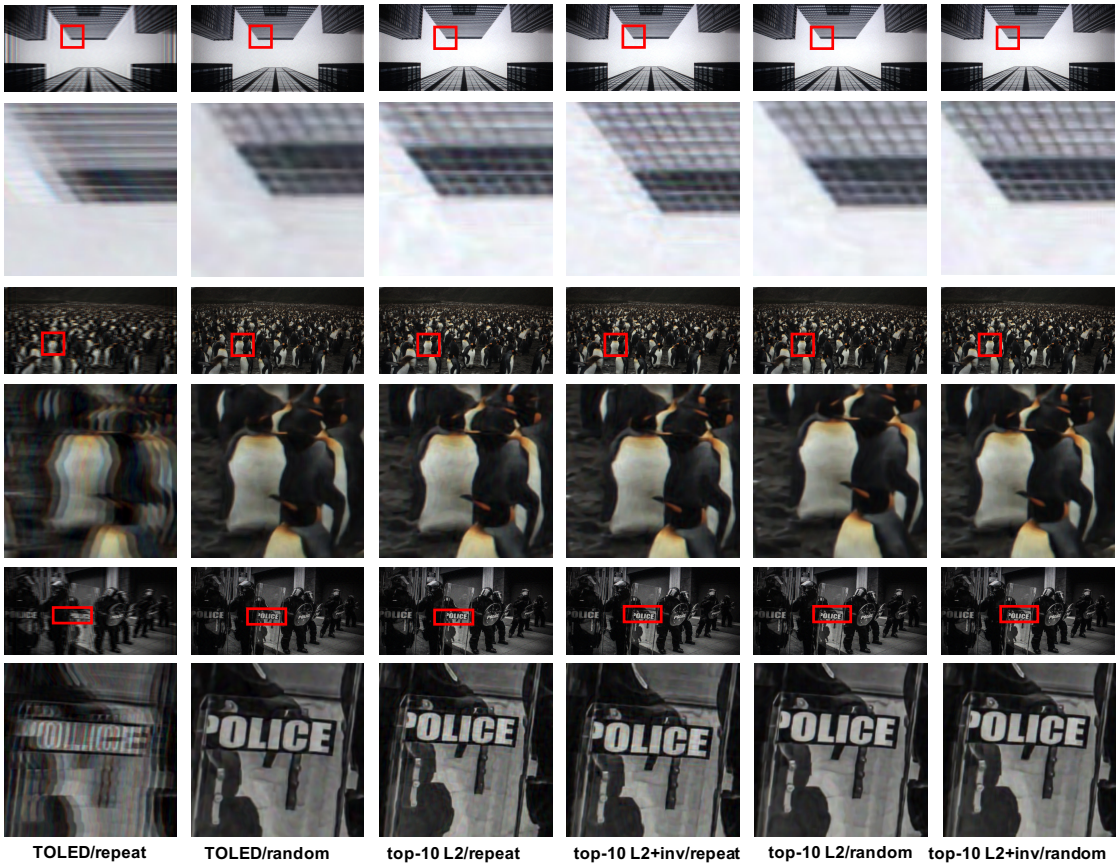


Figure 3.13: **Simulated results under six different displays.** The display masks have a pixel density of 150 DPI and we compare performance under noise level SNR=28. We show three scens and their zoom-in patch for comparison.

Ablation study for optimizing pixels. We analyze two important ingredients for optimizing a high-quality pixel opening shape — initialization and area constraint. As shown in Figure 3.14, initializing with an all-open pixel yields better performance than initializing with conventional T-OLED, since optmizing pixel pattern with area constraints is an non-convex problem and starting from all-open pixel helps avoid local minimum. And gradually decreasing the area constraints from 1 to desired opening ratio every five iterations improves the optimization results compared to fixing area constraint at the desired opening ratio.

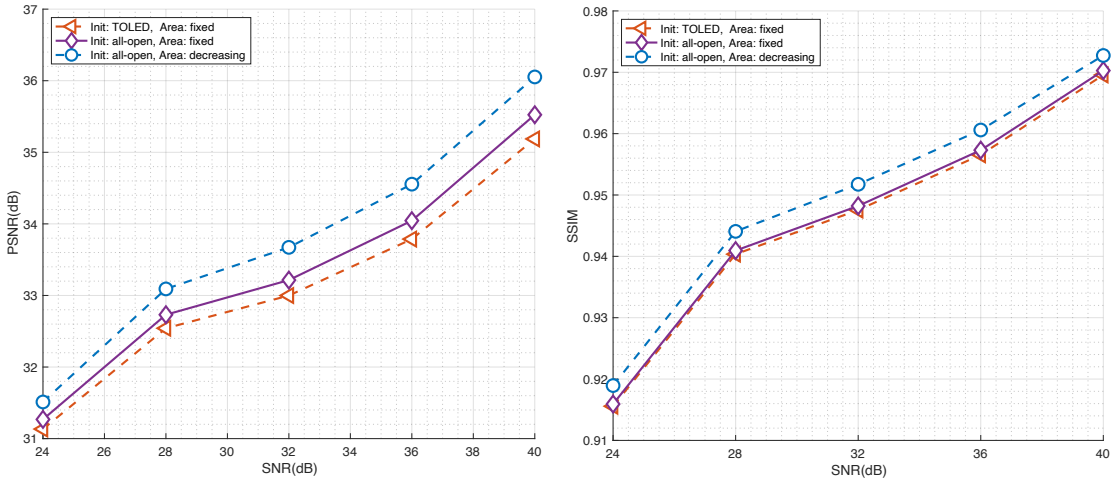
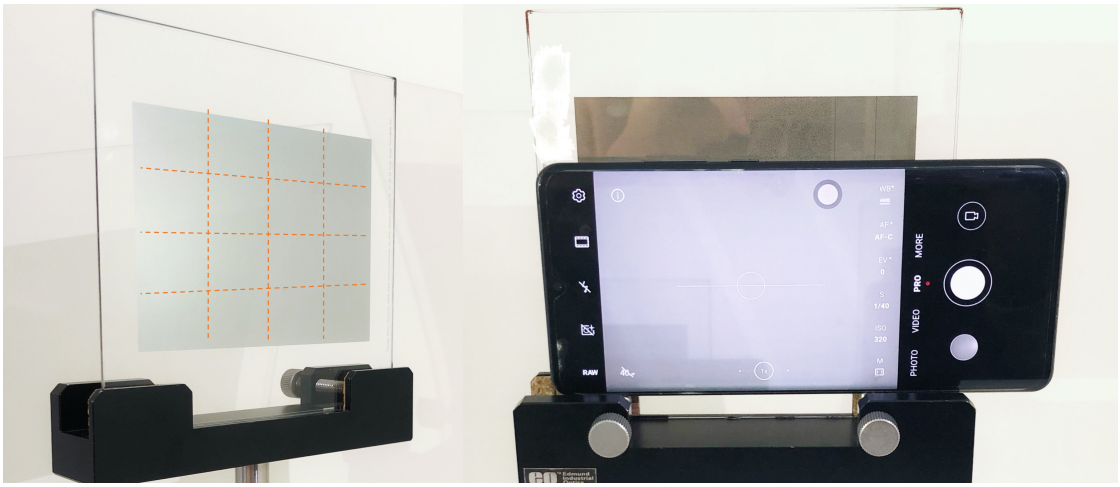


Figure 3.14: **Effect of initialization and area constraint in pixel shape optimization.** We compare two initialization and two way of constraining pixel opening area. (a) intialize with existing T-OLED pattern and fix opening area, (b) intialize with all-open pixel and fix opening area, (c) intialize with all-open pixel and gradually decrease area cosntraint from one to the desired area ratio.



(a) Photolithography mask (b) Under-panel camera prototype

Figure 3.15: Under-display camera lab prototype. (a) shows twelve photolighography masks that emulate different display designs. (b) shows our overall prototype where we place a cell-phone camera tightly against the printed mask and capture images by accessing the touch screen.

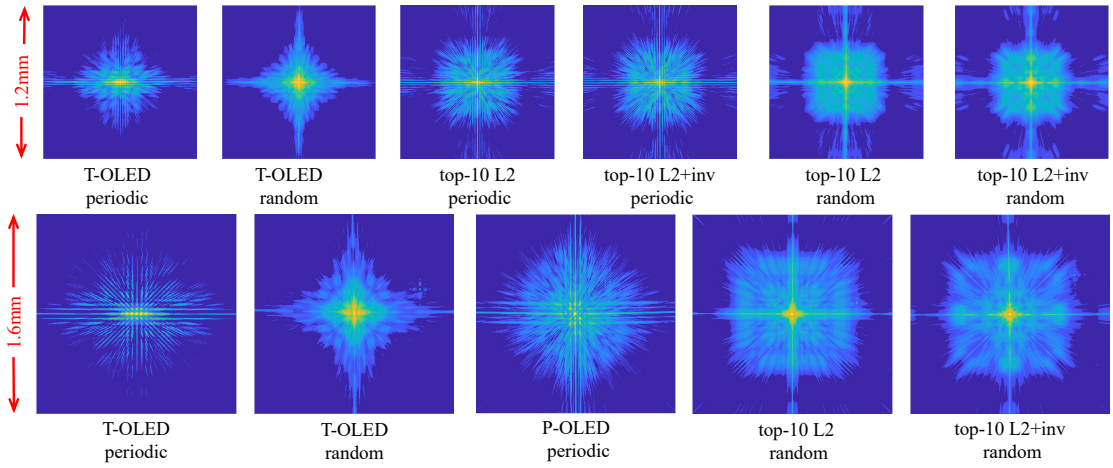


Figure 3.16: We capture PSFs of different display layouts and visualize green channel in log scale. The two rows show PSFs of displays with resolutions of 150 and 300 DPI, and of size 300×300 pixels and 400×400 pixels, respectively.

3.6 Real Experiments

We build a lab prototype to qualitatively evaluate images captured under different display designs and compare corresponding deblurring results.

Prototype. As shown in Figure 3.15, our prototype consists of a photolithography mask that emulates the display screen and an on-the-market cellphone camera. The cellphone camera has a focal length $f = 5.56$mm and an aperture of $f/1.6$.

Camera pipeline. We extract RAW images from the camera and process them using a simple pipeline. We first radiometrically calibrate the cellphone camera and use a color checkerboard, under different lighting conditions, for white balancing and color correction. This calibration is done prior to placing the photomask in front of the lens. For each RAW image, we first demosaic, spatially downsample it to 1364×1820, denoise and deblur it, and then correct color and tonemap it to obtain the final result. For Wiener deblurring, we normalize the blur kernel to sum up to 1 and set $\epsilon = 0.037$. We sufficiently pad the blurred image before deconvolution, and then crop the recovered image.

Measured PSF. Figure 3.16 showcases the PSFs measured with our prototype for different display layouts. For each display layout, we focus the camera on a white light LED that is placed far away from the camera in a dark room and capture an exposure stack. We fuse each exposure stack into an HDR image,

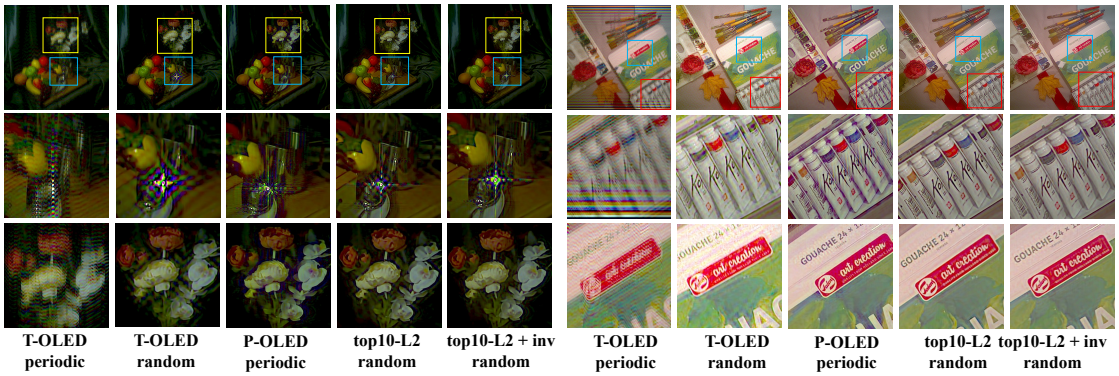


Figure 3.17: **Indoor scenes captured by our lab prototype under five different displays.** All display masks have a pixel density of 300 DPI. We show Wiener deconvolved results. Indoor scenes are close to the camera.

and crop a patch around the brightest point as PSF. Specifically, we crop a patch of 300×300 pixels for a display that has 150 DPI, and 400×400 pixels for a display that has 300 DPI. These PSFs are used both in Wiener deconvolution as well as to train DNNs for deblurring.

Capture settings. We fix ISO to be the smallest value 50 and use an exposure time of 1/125 *s* for most outdoor scenes and 1/8 *s* for indoor scenes.

Results for 300 DPI displays. We show results captured under 300 DPI display patterns for two indoor scenes in Figure 3.17 and one outdoor scene in Figure 3.18. Randomly tiled T-OLED has significantly better results than a conventional T-OLED layout. P-OLED and the optimized patterns yield relatively good performance in recovering the details such as texts and edges. However, P-OLED results contain more ringing artefacts; for example, around the toy's feet in Figure 3.18, purple halos around the flowers and ghosting around the texts on the painting tubes in Figure 3.17. All methods produce artefacts at specular regions on the spoon; this is a consequence of the non-linearity induced by saturation that violates the linear blur model. We also provide results for the outdoor scene in Figure 3.18 under 150 DPI displays in the supplementary to characterize performance of the system under lower DPI.

Results for 150 DPI displays. We compare the recovered images captured under six different display designs: T-OLED, top-10 L2 optimized patterns, and top-10 L2 + invertible loss optimized patterns, and each type with periodic and random pixel tiling. We print masks of 150 DPI containing pixels with 20% light throughput. We conduct Wiener deconvolution to recover a sharp image by capturing the

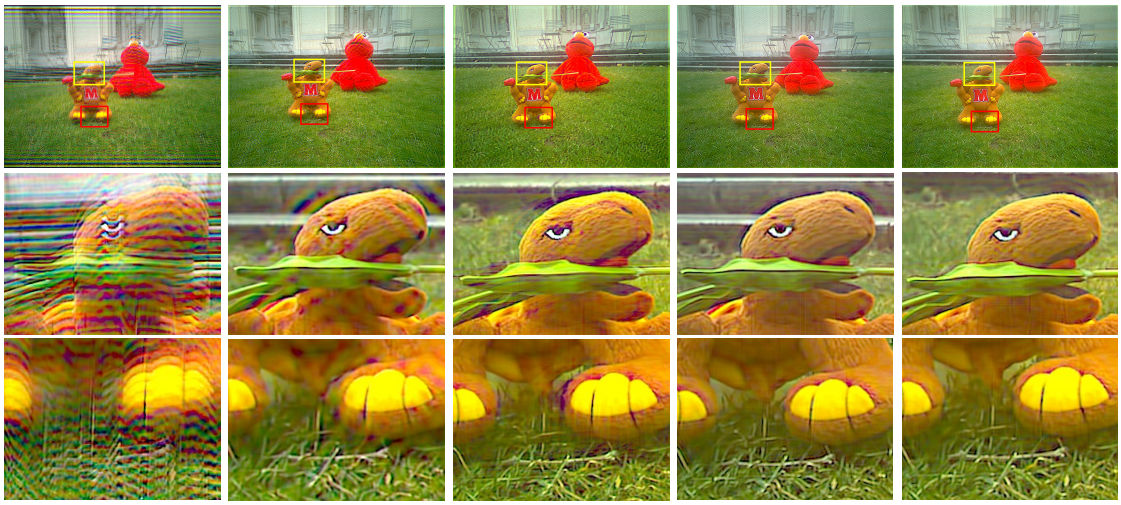


Figure 3.18: **Outdoor scenes captured by our lab prototype under five different displays.** All display masks have a pixel density of 300 DPI. Outdoor scenes are relative far-away from the camera and can better satisfy the infinity assumptions.

degraded image using our prototype and PSFs separately. In Figure 3.19 and 3.20, we show the results of these methods on an outdoor scene. T-OLED with conventional periodic tiling results in severe ringing artefacts in horizontal directions due to its one-directional openings. Introducing random tiling to T-OLED pixels suppresses the horizontal artefacts. Compared to T-OLED displays, optimized patterns with periodic tiling significantly increase the reconstruction quality. Lastly, optimized patterns with random tiling — in particular, Top 10 L2+Inv — yields the best performance, where thin structures such as texts on the clock can be easily seen from the deblurred results.

Comparison of deblurring methods. When presented with a large blur kernel, cropping introduced by the sensor has a nontrivial effect on deblurring. We show two additional deblurring methods that implicitly and explicitly handle the boundary issues — a deep neural network and deblurring with an iterative solver — on degraded images captured using our prototype.

For the neural network, which we denote as UNet-RDB, we use the same network structure as in [Zhou *et al.*, 2020a], where each layer in the downsampling subnet consists of two Residual Dense Blocks (RDB) [Zhang *et al.*, 2018c] and a 2D convolution layer, and each decoding layer has two RDBs and a transpose convolution layer. Since the network is specified to the blur kernel, we train for two networks,

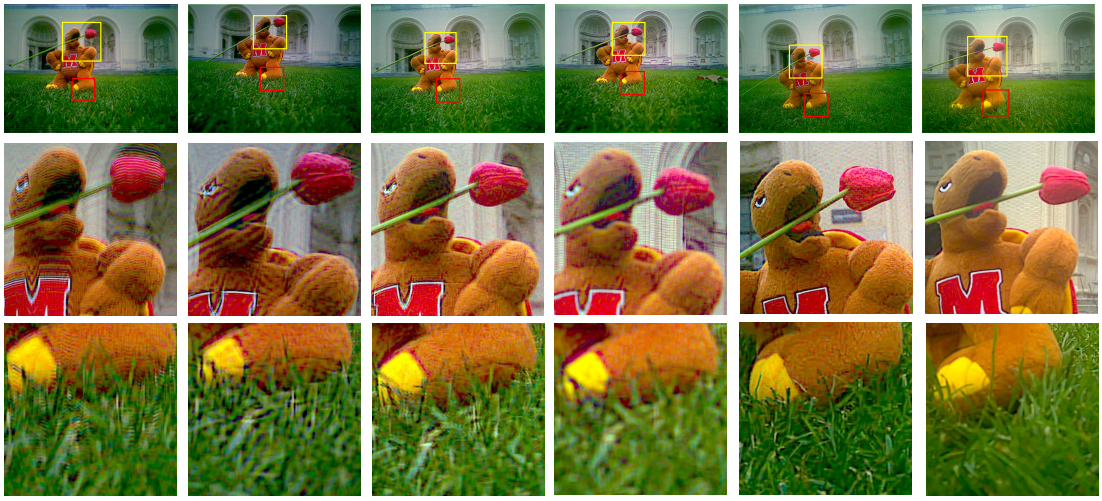


Figure 3.19: **Outdoor scenes captured by our lab prototype under six different displays.** The display masks have a pixel density of 150 DPI.

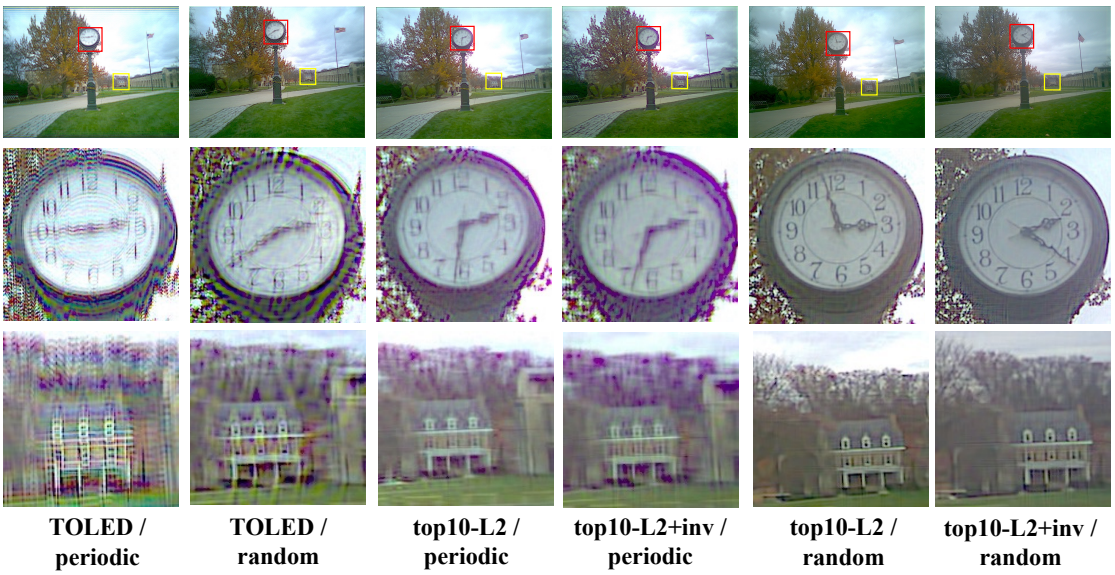


Figure 3.20: **Outdoor scenes captured by our lab prototype under six different displays.** The display masks have a pixel density of 150 DPI.

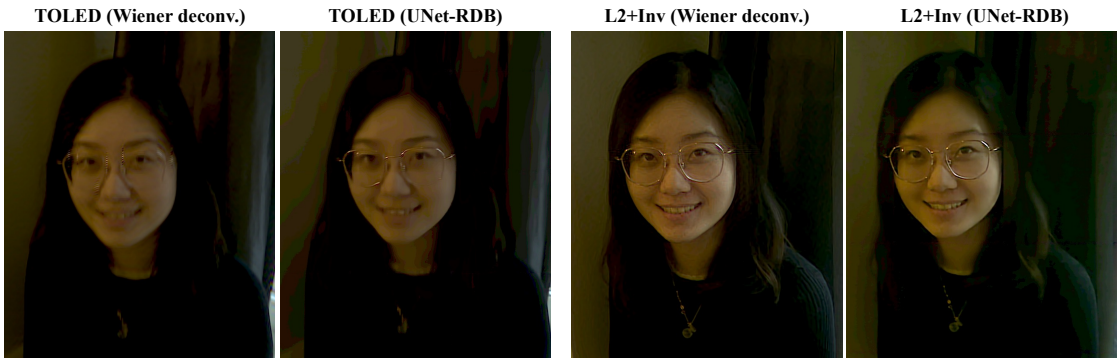


Figure 3.21: **Comparison of deblurring methods.** We compare Wiener deconvolution and UNet-RDB on a selfie captured under conventional T-OLED and our optimized display. We use display masks of 150 DPI.

one each for T-OLED with periodic tiling and L2+inv with random tiling. For each network, we construct 600 training image pairs and 30 validation pairs using images in HDR+ dataset [Hasinoff *et al.*, 2016]. We first demosaic RAW images in HDR+ [Hasinoff *et al.*, 2016] to serve as ground-truth images, and then blur the ground-truth images using captured PSFs and add noise to them. We randomly crop 256×256 patches and use a batch size of 10 in each iteration. All networks are trained for 1000 epochs with a learning rate 1e-3 at the beginning and scaled by 0.1 every 250 epochs. We use Adam optimizer with $\beta_1 = 0.9, \beta_2 = 0.999$.

As we observe in Figure 3.21, for the T-OLED pattern, the UNet-RDB significantly improves the reconstruction quality, recovering finer details with fewer ringing artefacts as compared to Wiener deconvolution. The improvements for the L2+inv pattern are less subtle, due to the inherent robustness of the PSF; there is some noise suppression, but the network also introduces some artefacts in the process.

We show additional results to compare Wiener deconvolution with UNet-RDB on degraded images captured using our prototype. As we obeserve in Figure 3.23 and Figure 3.24, the proposed pattern is consistently better than T-OLED regardless of the deblurring methods.

For the iterative solver, we model the unknown sharp image to be larger than the known blurred image, and as a convolution of the sharp image with *valid* boundary condition in MATLAB emulating convolution+sensor cropping. In Figure 3.22, we deblur the teaser images using a linear solver with Tikhonov prior on the image gradients. Compared to T-OLED and P-OLED, the proposed display layout has fewer ringing artefacts along all the edges.

Teaser. Figure 3.1 shows results on an outdoor scene for the T-OLED, P-OLED, and the Top-10 L2+Inv mask, all three with 300 DPI, and deblurred with Wiener deconvolution. The large spread of the blur in

Figure 3.22: **Deblurring with TV prior.** We recover sharp images by optimizing least square with TV prior using vanilla linear solver with *valid* boundary condition. Please zoom in to see the details.

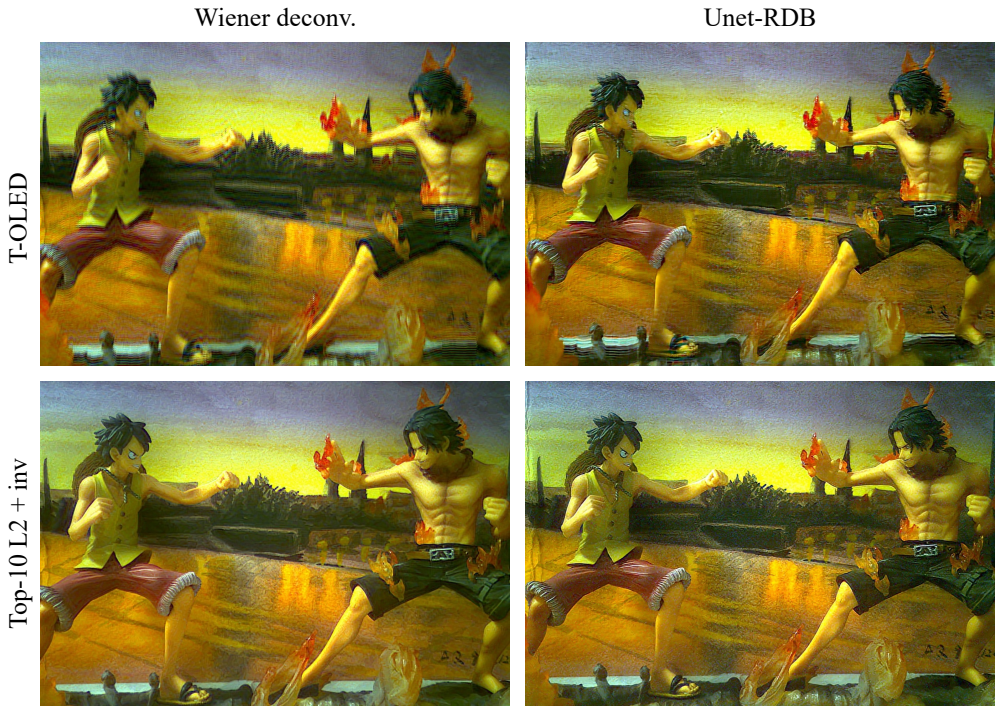
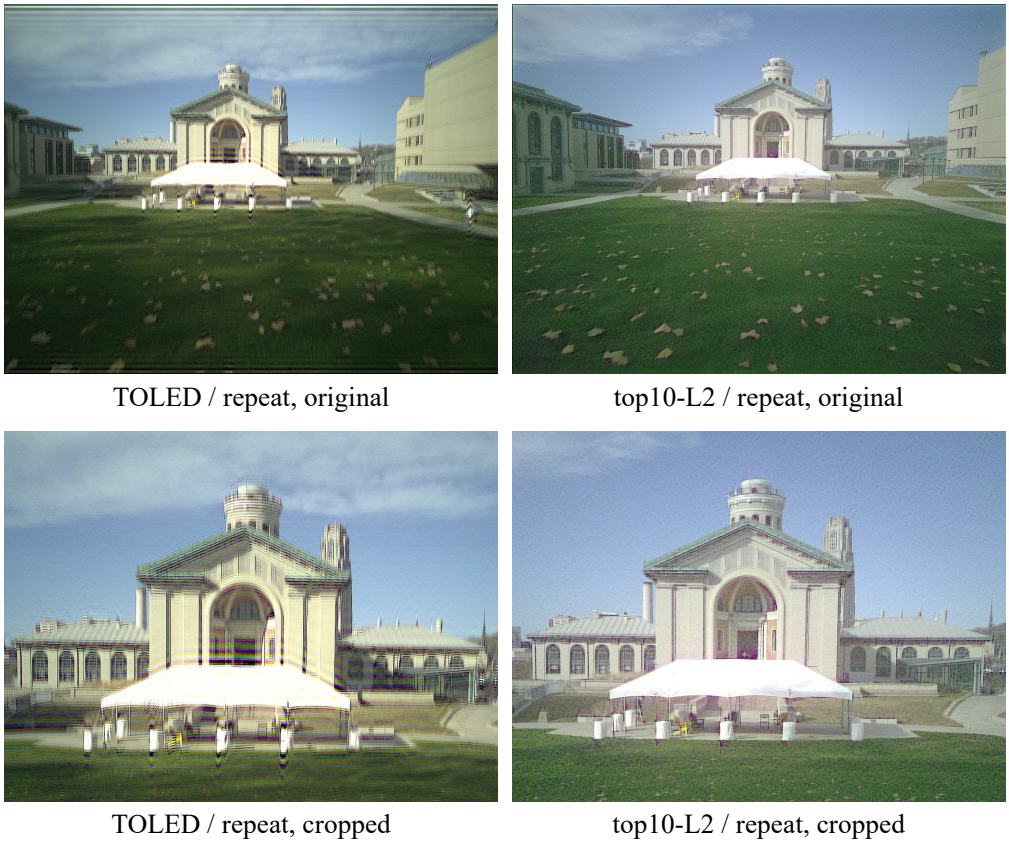


Figure 3.23: **Comparison of deblurring methods.** We compare Wiener deconvolution with UNet-RDB on an indoor scenes captured under conventional T-OLED and our optimized pattern.

Figure 3.24: **Comparison of deblurring methods.** We compare Wiener deconvolution with UNet-RDB on an outdoor scene captured under conventional T-OLED and our optimized pattern.

T-OLED along the x-axis leads to severe artefacts. These artefacts are less severe in P-OLED patterns. In contrast, the robustness enabled by our technique results in remarkably better results.

Crop vs uncrop. For many of the results shown in this chapter, we crop the edges of recovered images to better show the main subjects. To clarify the difference between a cropped deblurred image and an uncropped one, we compare these two under periodic T-OLED and one of our optimized pattern. Before cropping, T-OLED result contains severe ringing artefacts with wide range along horizontal / vertical edges. The cropped image is essentially the same image without the image edges. Note that even before cropping, our optimized pattern has very thin artefacts along the image edges. The ringing artefacts of the edges can also be observed in the Figure 3.1 and Figure 3.23 in this chapter.



TOLED / repeat, original | top10-L2 / repeat, original

TOLED / repeat, cropped | top10-L2 / repeat, cropped

Figure 3.25: **Comparison of orginal and cropped Wiener deconvolution results.** We show the orginal and cropped Wiener deconvolution results for T-OLED and our optimized patterns.

3.7 Discussions

This chapter shows that photographs obtained using under-panel cameras can be improved via careful design of the openings in the display through which the camera observes the scene. We show that introducing non-periodic pixel tilings as well as optimizing the mask openings at each pixel improves the invertibility of the diffractive blur introduced by the display; so much so that, even simple deblurring techniques like Wiener deconvolution can be successful. This indicates that designing the display layout is a promising approach for making under-display imaging practical.

OLED placement over optimized patterns. It is critical that any change in the display layout accommodate an LED array at the desired resolution, in terms of DPI. We show the RGB subpixel placement for T-OLED and P-OLED displays as well as potential subpixel placement for the proposed display pat-

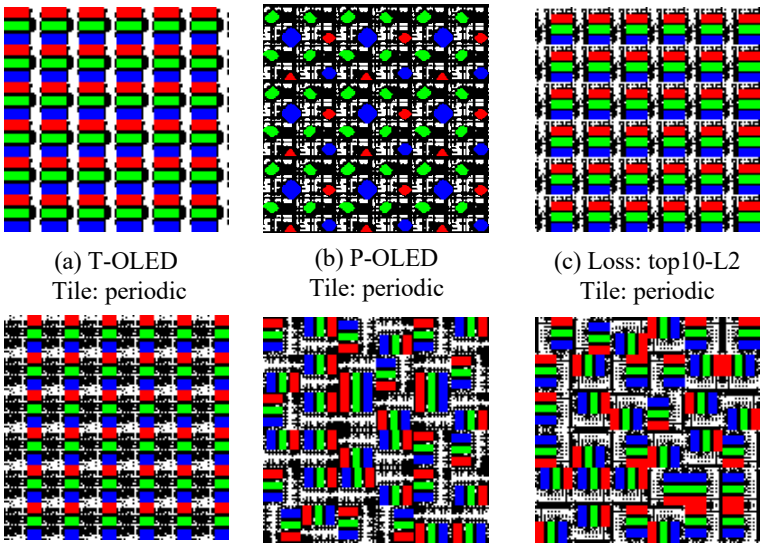


Figure 3.26: **RGB subpixel placement for different display layouts.** (a-b) show two typical RGB subpixel placement for OLED screen, and (c-f) show examples of RGB subpixel placement in out optimized display patterns. White region represents pixel openings and red, gree, blue represent regions for R, G, B subpixels.

terns in Figure 3.26. In all cases, the RGB pixels have the same footprint, in terms of area, although with an OLED placement that is no longer uniform. This non-uniform placement does run the risk of displaying content that appears aliased; however, we contend that this is minimal when operating with high-resolution displays. To validate this, we simulate images on the different display layouts in Figure 3.27. The displays are at 300 DPI and correspond to a square region with a width of 8.4mm, thereby emulating the area immediately in front of the under-panel camera. The first and the second row show T-OLED and two optimized displays, with periodic tiling and random tiling respectively. We observe that compared to the conventional periodic tiling, random tiled pixels yield reasonable display performance. Although random tiling introduces artefacts, it has a negligible visualization effect at these high resolutions. However, the randomness in OLED sub-pixel placement is likely to introduce challenges in manufacturing the display panel as well as designing the wiring pattern for data, control, and power, which we discuss next.

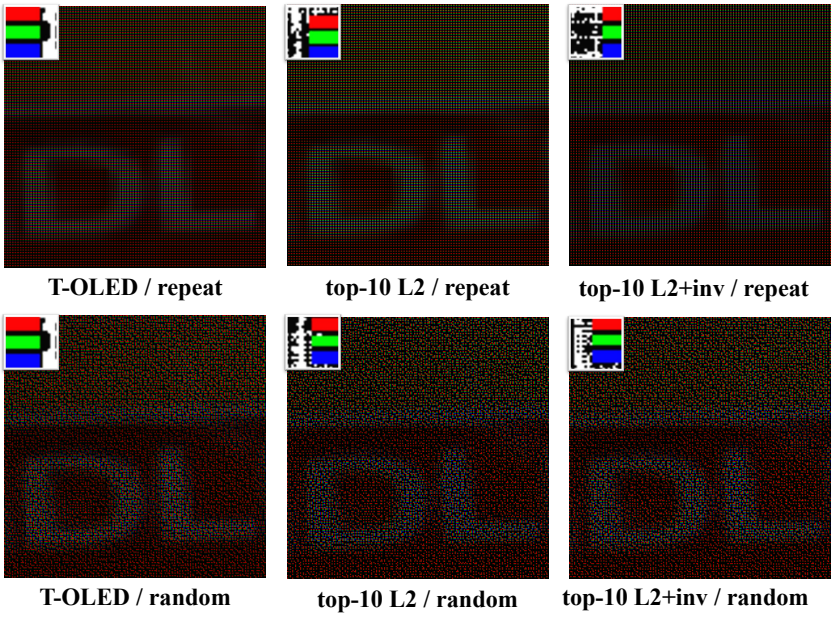


Figure 3.27: **Rendering an image using different display layouts.** This is a high-resolution rendering of the displays and each image corresponds to a display size of 8.4mm. To view it at the correct size as it would appear on a display with 300 DPI, please use 25% zoom.

Accommodating power/control wiring. Layout for the wiring required to power and control the OLEDs come in two forms: transparent and opaque [Wang *et al.*, 2020]. Transparent wiring can be overlaid under the color pixels and usually has a width of half pixels. Given its transparency, it has a negligible effect on our system modeling. Opaque wiring is narrower, taking a width of around 8μm. The simplest approach to adapt opaque wiring to our display design is to add horizontal and vertical space of 8μm around each pixel. As shown in Figure 3.28 , we add an additional 8μm spacing around each display pixel, whose size is 168μm and it has a negligible effect on the PSF. In all, given the maturity of fabrication technology, we believe that handling the randomized layout is an engineering challenge that can, in principle, be surmounted.

Spatial dependence of the blur PSF. The results shown in the chapter assume that the blur kernel is spatially invariant and that the degraded image can be modeled as a convolution between sharp image and a single PSF. In reality, the spacing between the camera lens and the display panel causes the blur PSF to be spatially varying. Further, non-idealities in the lens introduces other aberations including the Pincushion distortion seen in the PSFs shown in Figure 3.29. We show residual blurs (insets) of all PSFs

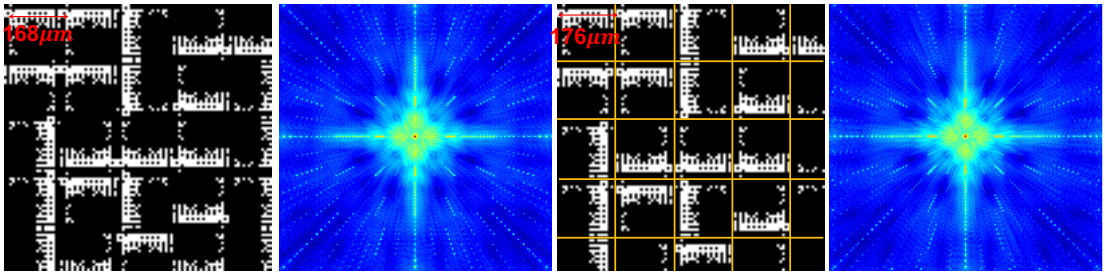


Figure 3.28: **Accomodation of opaque wiring.** The left two figures show the proposed Top10 L2+Inv random display and its PSF, and the right two figures show the same pattern with horizontal and vertical space of 8μm added around each pixel.

when deblurred by the center PSF.

Handling saturation. Non-linearities in the imaging pipeline can create a significant model mismatch to the linear model commonly assumed in deblurring approaches. For example, specular highlights in Figure 3.17 leads to jarring artefacts in the restored photographs. We can handle such scenes either by incorporating such artefacts in the training dataset for the DNN model. Alternatively, capturing an HDR image by exposure bracketing to revert back to a linear model helps, as seen in Figure 3.30.

Depth dependence. All of the results in this chapter are using blur PSFs measured under the assumption of the scene at infinity. For scenes with points close to the camera, there is a possibility that the PSF at infinity has a significant mismatch to that from a finite distance. To quantify this, we measure the PSF for our L2+Inv optimized mask when we focus on a point light source placed at different depths. This is shown in Figure 3.31. The observed blur kernel is (near) constant over the depth range that our prototype is capable of focusing on; we provide a detailed theoretical justification for this in the supplemental material. The net result is that we can successfully deblur an in-focus scene immaterial of the depth as seen in Figure 3.31, using the measured blur kernel for a scene at infinity. We also show that the defocus blur (last row of Figure 3.31) is stable when deblurred (inset), indicating a gentle bokeh on the out-of-focus regions. These observations are consistent with the deblurred textures in Figure 3.18, for both the in-focus and out-of-focus regions.

We derive the depth-dependence property of these blur PSFs, based on a standard derivation from Goodman [Goodman, 2005]. We follow notation and problem setup as stated in that book. Let us consider measuring the point spread function of our under-panel camera system. Point light source $U_0(\xi, \eta)$ is placed z_1 mm from the lens plane, a thin lens with focal length f_0 has an overall aperture function

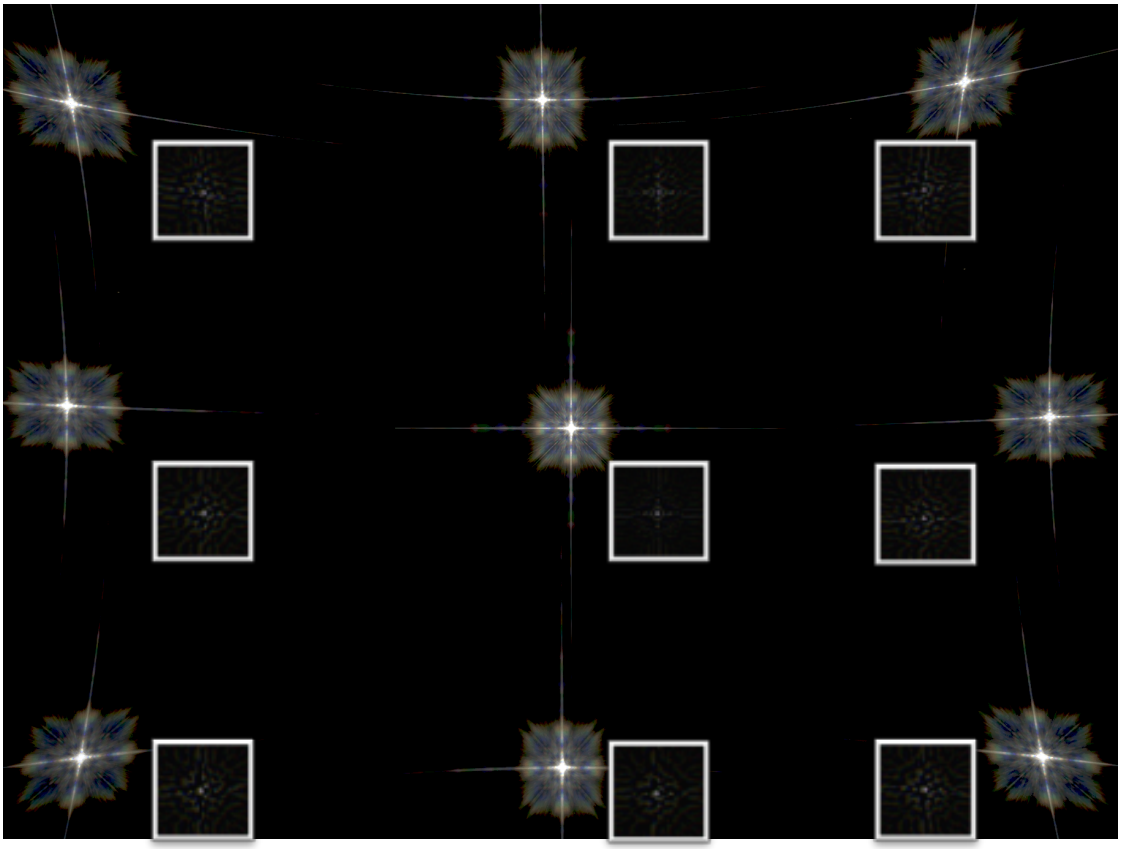


Figure 3.29: **Spatial variation of the blur kernel of the Top10-L2+Inv pattern.** We capture PSFs that appear on all corners and near edges. Insets show residuals after deblurring with the center PSF.

$a_r(x, y)$, and image sensor is placed z_2 mm from the lens. We require that the point source is in the focus of our prototype and the distances satisfy the Lens law,

$$\frac{1}{z_1} + \frac{1}{z_2} - \frac{1}{f_0} = 0.$$

The system impluse response can be given as the Fraunhofer diffraction pattern of the overall aperture $a_r(x, y)$ [Goodman, 2005],

$$h(u, v; \xi, \eta) \propto \iint_{-\infty}^{\infty} a_r(x, y) e\{-j\frac{2\pi}{\lambda z_2}[(u - M\xi)x + (v - M\eta)y]\}dxdy,$$

where $M = -z_2/z_1$. For a point ligth source $U_0(\xi, \eta)$ that lies on the optical axis, the measured point

Single exposure | Single exposure deblurred | HDR deblurred

Figure 3.30: **Handling saturation in UDCs.** Saturated pixels break the linearity of the imaging model and leads to artifacts in the deblurred photograph. We alleviate these by using HDR photography to obtain a linear blur model.

spread function can be written as

$$\begin{aligned} k(u,v;\lambda) &= \left| \iint_{-\infty}^{\infty} h(u,v;\xi,\eta) U_0(\xi,\eta) d\xi d\eta \right|^2 \\ &\propto | \iint_{-\infty}^{\infty} a_r(x,y) e\{-j\frac{2\pi}{\lambda z_2}[ux+vy]\} dxdy|^2 \\ &\propto |A_r(\frac{x}{\lambda z_2}, \frac{y}{\lambda z_2})|^2. \end{aligned}$$

Point spread function $k(u,v;\lambda)$ is the Fourier transform of the overall aperture function with a scaling factor $1/\lambda z_2$ and the scaling factor changes with the depth z_1 we focus on. In this chapter, we calibrate our system with a point at infinity (or atleast very far away), and so, $z_1 = \infty$ and $z_2 = f_0$. Note that in our camera $f = 5.56$ mm and can focus in a range of 10 cm to infinity. When the camera is focused at the nearest possible setting $z_1 = 10$ cm, the corresponding $z_2 = 5.86$ mm, which is less than 10% off from f_0 and hence the scaling that we observe on from the focused-at-infinity blur is minimal. For scenes at say $z_1 = 1$ m, $z_2 = 5.59$ mm which is extremely close to f_0 and we will observe nearly the same blur when focused on that point.

We conclude that within the depth range that our prototype is capable of focusing on, the scaling point spread function is near constant.

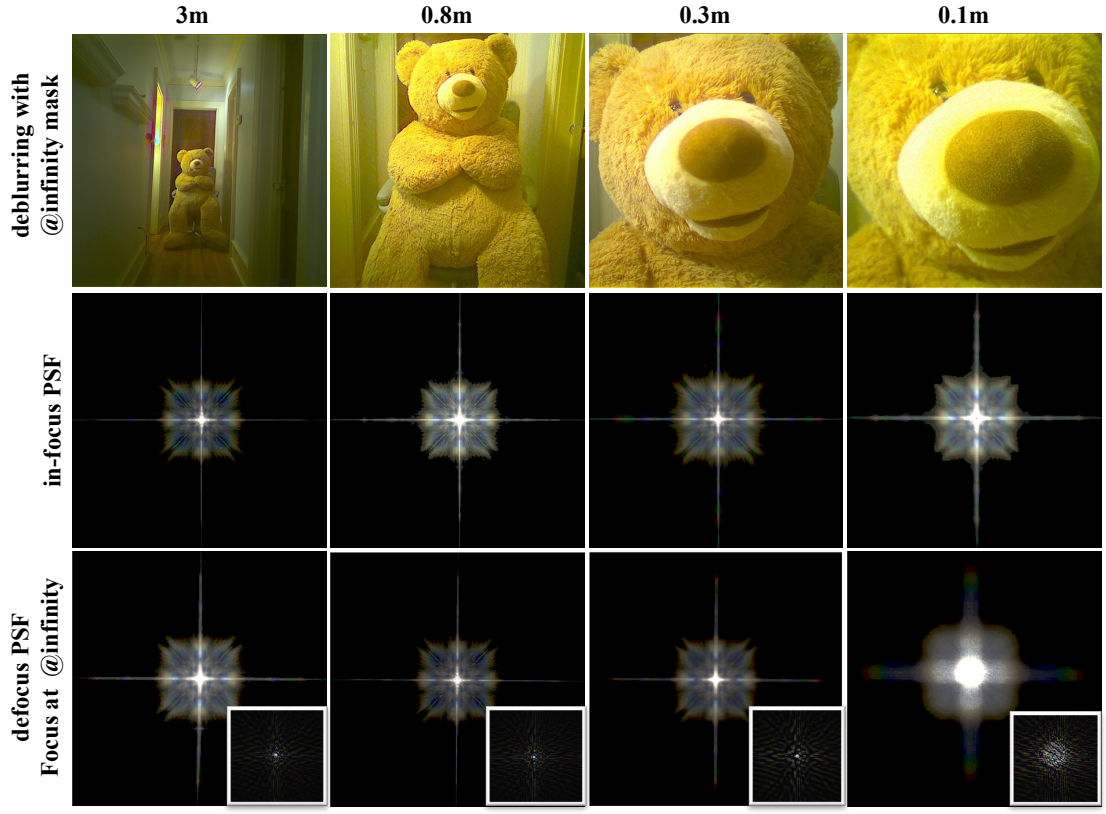


Figure 3.31: **Deblurring at different depths using the Top10-L2+Inv pattern.** (Top row) Deblurred photos for in-focus scene at different depths. In all cases, the deblurring was performed with the PSF corresponding to scene-at-infinity. (middle row) PSF of a point light source at different depths, with camera focused on the light source. (bottom) PSF of a point light source with camera focused at infinity. (bottom-inset) Residual blur after deblurring the defocus blur.

4 Designing Phase Masks for Under-Display Cameras

While optimizing the shape of the display openings produces significant improvement in the image quality of UDCs, it also requires non-trivial engineering effort to redesign the entire display. As a complementary solution, in this chapter, we incorporate phase masks on display panels to *optically* modify the display openings. Our design inserts two phase masks, specifically two microlens arrays, in front of and behind a display panel. The first phase mask concentrates light on the locations where the display is transparent so that more light passes through the display, and the second phase mask reverts the effect of the first phase mask. We further optimize the folding height of each microlens to improve the quality of PSFs and suppress chromatic aberration. We evaluate our design using a physically-accurate simulator based on Fourier optics. The proposed design is able to double the light throughput while improving the invertibility of the PSFs. Lastly, we discuss the effect of our design on the display quality and show that implementation with polarization-dependent phase masks can leave the display quality uncompromised.

4.1 Introduction

Inspired by a large body of work that enhances capability of imaging systems with phase masks [Heide *et al.*, 2016, Jeon *et al.*, 2019, Peng *et al.*, 2016, 2019, Shi *et al.*, 2022, Sitzmann *et al.*, 2018, Wu *et al.*, 2019], we propose to design phase masks to suppress diffractive blur and increase light throughput for UDCs. We first show using basic Fourier optics [Goodman, 2005] that inserting a thin phase mask at the display is ineffective in improving UDCs. To overcome the limitation of a single phase mask, we propose to use two phase masks—specifically two microlens arrays– placed *in front of* and *behind* the display; we do this for the specific case of transparent-OLEDs (TOLED), a display model commonly used in today's cellphones. The first phase mask distributes light to locations where the display is transparent, and the second phase mask recovers the original waveform. Within some limits, this allows the incident light to

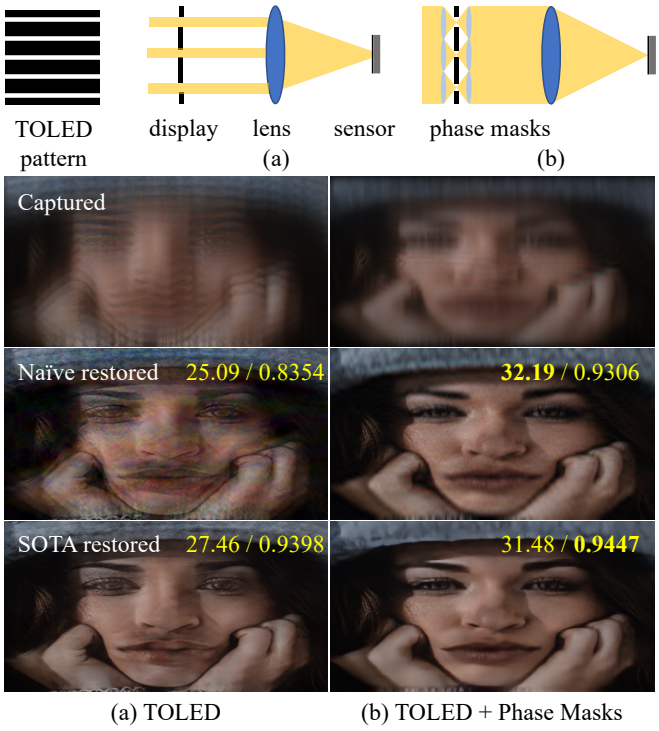


Figure 4.1: **A comparison between a UDC under a transparent-OLED display (a) without and (b) with the proposed phase masks.** (From top to bottom) Rows show the setup, images captured under each, restored images using a naive iterative solver, and using a state-of-the-art deep network [Feng *et al.*, 2021]. We show PSNR(↑) in dB and SSIM(↑) for restored images. Both UDCs have a pixel density of 600 DPI.

pass through without being blocked by the display or effectively renders the display fully transparent!

In order to prevent the microlens arrays from hindering the display quality, we propose to implement them as thin polarization-dependent phase masks. A naive implementation of microlens arrays as thin optics is to fold them at a fixed height. However, this results in severe chromatic aberrations. We instead choose a different height for each microlens through optimization, so that diffractive blur is suppressed equally at all wavelengths. Using simulations, we show that the proposed phase mask significantly increases the image quality of a UDC (see Figure 4.1). The code for this work is publicly available [Yang *et al.*, 2023].

In summary, we make the following contributions:

- We show that a phase mask placed tightly against a display is inadequate to improve the image quality

of UDCs.

- We propose to insert two microlens arrays in front of and behind a TOLED, which effectively allows more light to reach the camera and produces more invertible PSFs.
- We implement the proposed microlens arrays as thin polarization-dependent phase masks, a design that ensures light emitting from the display is not modulated and therefore guarantees high display quality.
- When implementing microlenses as thin optics, we optimize the folding height of each microlens to minimize chromatic aberrations.
- We conduct simulation based on wave optics and physically-accurate camera pipeline and demonstrate that the proposed setup outperforms the conventional UDC.

Limitations. The proposed method has two limitations. First, we show phase correction can suppress the diffractive blur of TOLED display, whose pattern is separable along x and y directions. Extending this to 2D displays is hard due to the high computational cost of simulating 2D short-distance propagation. Second, the field of view of the resulting UDC can be constrained by the use of phase masks. This is determined by the ratio of the focal length f of microlens arrays and the size of the pixel opening. Our choice of f produces a field of view of around 14°.

4.2 Background

T-OLED Displays. In this chapter, we focus on the simplest layout TOLED, whose opening pattern is shown in the upper left of Figure 4.1 and is a separable along $x-$ and $y-$directions. Along the x-direction, every display pixel has an opening of around 23.8 %; in y-direction, the aperture is fully open. The PSF produced by TOLED is also separable. Therefore, we focus on the design of one-dimensional phase masks for x-direction.

4.3 Phase Mask Design for UDCs

We explore the design space of phase masks in UDCs and look into two scenarios — first, a single phase mask placed tightly against the display and second, two phase modulations in front of and behind the display.

4.3.1 Inadequacy of Single Phase Masks

Phase masks modulate the phase of an incident wavefront and can potentially correct the wavefront to form a PSF that is easily invertible. We first examine the most common setup of placing a single phase mask at the aperture plane [Heide *et al.*, 2016, Jeon *et al.*, 2019, Peng *et al.*, 2016, 2019, Shi *et al.*, 2022, Sitzmann *et al.*, 2018, Wu *et al.*, 2019]. Unfortunately, using basic Fourier optics, we show that such a phase mask is insufficient to reduce diffraction blur in UDCs.

Lemma 1 (Inadequacy of a single phase mask)**.** *A single-sided phase mask can not improve the invertibility of the point spread function of a UDC.*

Proof. Let $a(x)$ be the aperture of a UDC, and $h(x)$ be the height map of a single-sided phase mask that is placed tightly against the display panel. We assume that the aperture and the phase mask are on the same plane, and the overall aperture function $b(x)$ can be written as

$$b(x) = a(x)e^{j\frac{2\pi}{\lambda}(n-1)h(x)} \tag{4.1}$$

where n is the refractive index of the phase mask and λ is the wavelength of the incident wavefront. The invertibility of PSF $k_b(x)$ can be measured by its amplitude spectrum $|K_b(u)|$, where values close to zero are hard to invert, and large values are robust to noise in inversion. From (3.2), $k_b(u)$ is the (scaled) power spectral density of the aperture, its Fourier transform $K_b(u) = \mathcal{AC}_b(\tau)$, where $\mathcal{AC}_b(\tau)$ is the (scaled) autocorrelation function of the aperture. We compute the autocorrelation of the overall aperture function,

$$\mathcal{AC}_b(\tau) = \int_{-\infty}^{\infty} a(x)a(x+\tau)e^{j\Delta\Phi_\tau(x)}dx, \tag{4.2}$$

and $\Delta\Phi_\tau(x) = \frac{2\pi}{\lambda}(n-1)(h(x) - h(x+\tau))$. We then compute the intensity of $\mathcal{AC}_b(\tau)$, and by triangle inequality,

$$|\mathcal{AC}_b(\tau)| = \left| \int_{-\infty}^{\infty} a(x)a(x+\tau)e^{j\Delta\Phi_\tau(x)}dx \right| \tag{4.3}$$

$$\leq \int_{-\infty}^{\infty} |a(x)a(x+\tau)|dx. \tag{4.4}$$

Since aperture function $a(x)$ is non-negative, we can further simplify the above equation,

$$|\mathcal{AC}_b(\tau)| \leq \int_{-\infty}^{\infty} a(x)a(x+\tau)dx = |\mathcal{AC}_a(\tau)|. \tag{4.5}$$

We can see that $|\mathcal{AC}_b(\tau)| \leq |\mathcal{AC}_a(\tau)|$ for all τ and

$$|K_b(u)| \leq |K_a(u)|, \tag{4.6}$$

implying that PSF produced by a display panel with a single-sided phase mask is always worse in terms of invertibility than that produced by a pure display panel. ■

4.3.2 Double Phase Masks

If inserting a thin phase mask at the display plane is ineffective to improve the image quality of a UDC, would inserting multiple phase masks help? Diffraction blur in a UDC is produced by the small openings on the display pixels that have sizes comparable to the wavelength of incident light. Smaller opening results in a more severe diffraction blur [Yang and Sankaranarayanan, 2021b]. Would it be possible to *optically* expand the size of display openings, i.e. let a larger portion of light pass through display openings?

Consider now a system with two phase masks, on either sides of a display. The first surface with a height profile $h_1(x)$ modulates light incident on the display so that, after propagating for some distance z m, most of the intensity of the wavefront is concentrated at the display openings. The second surface $h_2(x)$ modulates the diffused wavefront so as to revert the effect of the first phase mask. If successful, the display panel would be rendered invisible.

Mathematically, this can be modeled as follows: Given a wavefront $p_\theta(x;\lambda)$ that incidents from angle θ and has a wavelength of λ. The incident wavefront passes through a phase mask, a display panel, followed by another phase mask, and becomes

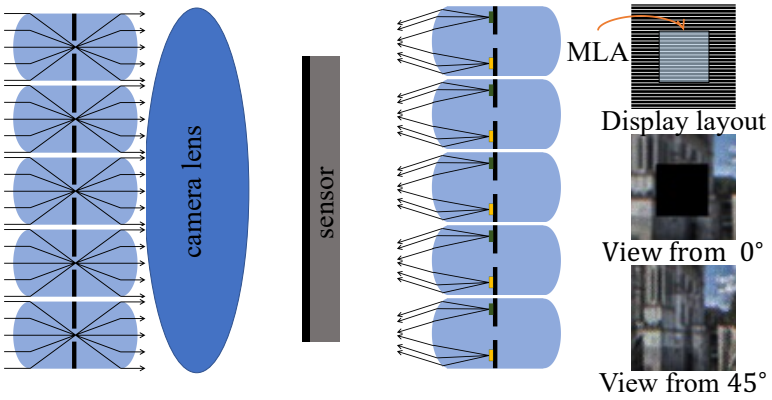
$$p'_\theta(x;\lambda) = \underbrace{(\Phi_2 \circ \mathcal{Q}_z \circ a \circ \mathcal{Q}_z \circ \Phi_1)}_{a_\Phi}(p_\theta(x;\lambda)) \tag{4.7}$$

where $a(x)$ describes the display openings, $\Phi_i(x) = \exp\{\frac{2\pi}{\lambda}(n-1)h_i(x)\}$, $i = 1, 2$ are phase modulations of the first and second height maps, and $\mathcal{Q}_z(\cdot)$ is the operator corresponding to wave propagation of z m.

Our goal is to design height maps $h_1^*(x)$, $h_2^*(x)$ and distance z^* such that the resulting aperture $a_\Phi^*(x)$ is approximately a fully-open aperture, $a_\Phi^*(x) \overset{\propto}{\sim} 1$.

4.3.3 Proposed Design: Double Microlens Arrays

In theory, height maps and thickness of an optimal double-sided phase mask $h_1^*(x)$, $h_2^*(x)$, z^* can be solved through an optimization problem. However, propagating incoherent wavefronts with a physically accurate model at each iteration is an expensive procedure, and gradient descent only allows solving for the height range that corresponds to the range of 2π modulation.



(a) UDC imaging arm (b) OLED display arm

Figure 4.2: **Proposed microlens arrays for UDCs.** In the right column, we render OLED display with MLA from two viewpoints. MLA is placed under the UDC aperture (center square). When viewed from the front (i.e, 0°), the display appears dark.

Our design is to place two microlens arrays (MLA) with equal focal lengths on either sides of the display such that the display panel lies in the focal plane of both MLAs, as shown in Figure 4.2(a). Light incident from the scene is concentrated by each microlens, passes through the display opening, and diverges to a parallel beam by the second set of microlenses. Compared to UDCs with a pure display, the proposed setup allows a larger portion of light to reach to camera main lens, and therefores improve the conditioning of incident wavefront and SNR.

However, the microlens array in front of the display also modulates light emitting from the display pixels. An illustration is shown in Figure 4.2(b). Since the display subpixels are misaligned with the optical axis of each microlens, light emitting from subpixels is rarely refracted to the direction along the optical axis. This implies the display would appear dark when users view it from orthogonal viewpoint.

4.3.4 Folding MLAs to Thin Plates

One approach to prevent microlenses from affecting the display is to implement them as polarization-dependent optics and place a pair of orthogonal linear polarizers on both sides of the display panel.

The microlens arrays only modulate the phase of light along p-polarization state. First, we examine the camera point of view. Light incident from the scene is a mixture of both states. Phase mask only modulates the p-state (shaded lines) and leaves s-state (solid color) unchanged. The polarizer behind the display selects p-states and filters out the rest. Thus the camera works in the same principle as

we described in the previous section. This polarization-dependent implementation reduces the light throughput by half, which is taken into consideration in all simulations. Second, we look into the effect of phase masks on the display quality. Due to the presence of s-polarization filter, the display RGB subpixel emits light along s-state. As our phase mask only modulates light along p-polarization state, light emitting from the display is left untouched.

Since polarization-dependent optics are only available in thin optics, either as phase spatial light modulators (SLM) or thin optical elements [Hu *et al.*, 2021, Li *et al.*, 2019]. It is necessary to fold each microlens into a thin phase plate at maximum height d_0,

$$\hat{h}(x) = \mathrm{mod}\Big(-\frac{x^2}{2(n-1)f_0}, d_0\Big). \tag{4.8}$$

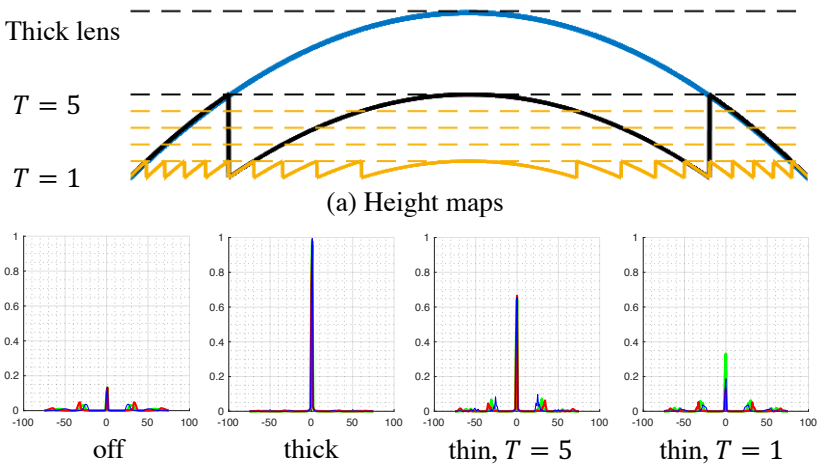
Figure 4.3 shows an example. Larger d_0 produces a phase mask that contains few phase wrappings and performs almost equally across all wavelengths; and small d_0 leads to much more phase wrappings and the resulting performance is strongly wavelength dependent. Phase plate wrapped at d_0 has preferable performance for light of a certain wavelength λ_0 over those of other wavelengths. This is because d_0 can be viewed as $\frac{T\lambda_0}{n-1}$, where T is an arbitary positive integer that coarsely controls the thickness of a phase mask and λ_0 is a wavelength that decides the exact thickness. A thick microlens can be written as $h(x) = \hat{h}(x) + c(x)\frac{T\lambda_0}{n-1}$, where $c(x) \in \mathbb{Z}$, and produces a phase modulation of $\exp\{j\frac{2\pi}{\lambda}(n-1)\hat{h}(x)\}\exp\{j\frac{2\pi}{\lambda}c(x)\frac{T\lambda_0}{n-1}\}$. For incident light of wavelength $\lambda = \lambda_0$, the modulation of thick lens is the same as that of the phase plate. For incident light of other wavelengths, the phase plate produces wrapping artifacts and thinner plates have more chromatic aberration.

A typical phase SLM is able to achieve phase modulations within a range of 2π or equivalently $T = 1$; other liquid crystal-based non-programmable optics can be implemented with larger phase retardation. Therefore, we design phase masks for two thicknesses: $T = 1$ for the phase SLM and $T = 5$ for thicker retarders.

4.3.5 Phask Masks Optimization

As mentioned in the previous section, a thin phase plate of uniform height d favors a corresponding wavelength λ and produces wrapping artifacts for other wavelengths, resulting in chromatic aberration. We propose to optimize a different height $d[l]$ for each microlens l such that the optimized inevitability of the overall PSF is the same across RGB channels to eliminate chromatic aberration.

Given a UDC with L microlenses with corresponding heights $\{d_l | l = 1, ..., L\}$, each of which takes the value in a set of heights i.e., $d_l \in \{h_j | j = 1, ..., N\}$. The set of discrete heights is created by uniformly sampling N wavelengths from 400 nm to 700 nm. The goal is to find the number of d_l with the same



(a) Height maps

(b) PSFs produced by light incident along optical axis

Figure 4.3: **Thick lens versus phase masks** wrapped at the dash lines $T = 1, 5$ and their PSFs.

height for each h_j.[1] Thus we define $m_j = \sum \mathbb{I}(d_l = h_j), \forall l = 1, .., L$ and a vector $\mathbf{m} = [m_1, ..., m_N]^\top$ for all N heights.

We calculate the invertibility of a system with different heights as a weighted combination of that of constant ones. The invertibility is measured by $v^j(\lambda)$, the region under modulation transfer function for microlens of height h_j and a specific wavelength λ, and higher scores are better. Specifically, we form form a matrix $\boldsymbol{V} \in \mathbb{R}^{N \times N}$, where $\boldsymbol{V}_{j,k} = v^j(\lambda_k)$ is the system invertibility for height h_j and wavelength λ_k. The invertibility of a new system with mixed heights $\boldsymbol{m}$ can therefore be computed by $\boldsymbol{V}_k^\top \boldsymbol{m}$.

Different wavelengths contribute to the performance of RGB channels differently, for example, wavelengths close to 470 nm, 530 nm, 610 nm matter more to the overall performance than other wavelengths, and the importance is characterized by the sensor spectral response function. We thus discretize the function into a matrix $\boldsymbol{S} = [\boldsymbol{s}_R^\top, \boldsymbol{s}_G^\top, \boldsymbol{s}_B^\top]$ and $\boldsymbol{S}^\top \boldsymbol{V}^\top \boldsymbol{x}$ computes the RGB performance under $\boldsymbol{m}$. We optimize the following problem,

$$\min_{\boldsymbol{m}} \|\boldsymbol{S}^\top \boldsymbol{V}^\top \boldsymbol{m} - \mathbf{1}\|_2^2 + \alpha \|\boldsymbol{m}\|_1 \tag{4.9}$$

$$\text{s.t.} \quad m_i \geq 0, \quad i = 1, \ldots, N. \tag{4.10}$$

[1] We show that the *ordering* of d_l has negligible effects on the performance in supplementary. Therefore, we only optimize for the counts.

The first term guarantees that performance of RGB channels is equally high, and the second term is a regularization. In optimization, $\boldsymbol{m} \in \mathbb{R}^N$ is a continuous variable, and in evaluation it is rounded up to integers. m_i is non-negative since it represents a count.

#1 Achromatic. The first term minimizes the ℓ_2 difference between the invertibilities of RGB channels and an all-one vector so that the system performance is equal across RGB channels. Note that $\boldsymbol{S}^\top \boldsymbol{V}^\top \in \mathbb{R}^{3\times N}, N \gg 3$. Therefore, the first term is an under-determined system with infinitely many solutions. Without proper regularization, the magnitudes of elements in $\boldsymbol{m}$ can be unbounded.

#2 More invertible. The second term, ℓ_1 regularization, encourages large invertibilities. The choice of ℓ_1 norm is motivated as follows. Since $\boldsymbol{m}$ represents counts of microlenses, with proper normalization, it sums to the total number of microlenses in the aperture. We let the normalized counts be $\frac{\boldsymbol{m}L}{\|\boldsymbol{m}\|_1}$, and normalized invertibilities be $\boldsymbol{S}^\top \boldsymbol{V}^\top \frac{\boldsymbol{m}L}{\|\boldsymbol{m}\|_1}$. Since $\boldsymbol{S}^\top \boldsymbol{V}^\top \boldsymbol{m}$ is constrained to be $\mathbf{1}$ by the first term, the normalized invertibilities can be simplified as $\frac{\mathbf{1}L}{\|\boldsymbol{m}\|_1}$. Therefore, minimizing $\|\boldsymbol{m}\|_1$ is equivalent to maximizing the invertibilities of RGB channels.

We use the log-barrier approach to approximate the non-negative constraint and convert the original optimization into an unconstrained problem,

$$\min_{\boldsymbol{m}} \|\boldsymbol{S}^\top \boldsymbol{V}^\top \boldsymbol{m} - \mathbf{1}\|_2^2 + \alpha \|\boldsymbol{m}\|_1 - \frac{1}{t} \sum_{j=1}^{N} \log\left(\boldsymbol{z}_j^\top \boldsymbol{m}\right) \tag{4.11}$$

where $\boldsymbol{z}_j$ is a one-hot vector that is one at j-th element. We apply the barrier method to solve this problem [Boyd *et al.*, 2004]. The algorithm is summarized in Algorithm 1.

During inference, we normalize and round up the optimal $\boldsymbol{m}$ into $\hat{\boldsymbol{m}}^* = \lfloor \frac{\boldsymbol{m}^*}{\|\boldsymbol{m}^*\|_1} L \rfloor$, such that $\sum_{j=1}^{N} \hat{m}_j^* = L$. The resulting phase mask contains $\hat{m}_j^*$ number of microlenses that have a maximum height of $h_j = \frac{T\lambda_j}{n-1}$.

Figure 4.4 shows an example of optimized phase masks. Compared to a fixed height that favors the green sensor channel and produces severe chromatic aberration, the optimized heights perform equally across RGB channels. Compared to uniformly varying heights, the optimized profile produces sharper PSFs.

Algorithm 1 Algorithm for phase mask optimization

Input: $\alpha \leftarrow 0.1, \mu \leftarrow 2, \boldsymbol{m}_{\text{init}} \leftarrow \mathbf{1}, t^{(0)} \leftarrow 0.01$

Output: $\boldsymbol{m}^*$

$k \leftarrow 0$ ▷ iteration index

$\boldsymbol{m}^{(0)} \leftarrow Newton(\mathcal{L}, \boldsymbol{m}_{\text{init}}, t^{(0)})$

while $\mu \cdot t^{(k)} \leq 10^4$ **do**

$k \leftarrow k + 1$

$t^{(k)} \leftarrow \mu \cdot t^{(k-1)}$

$\boldsymbol{m}^{(k)} \leftarrow Newton(\mathcal{L}, \boldsymbol{m}^{(k-1)}, t^{(k)})$

end while

$\boldsymbol{m}^* \leftarrow \boldsymbol{m}^{(k)}$

Function $Newton(\mathcal{L}, \boldsymbol{m}_{\text{init}}, t)$

$\boldsymbol{m} \leftarrow \boldsymbol{m}_{\text{init}}$

$\epsilon \leftarrow 1$

while $\epsilon \geq 0.001$ **do**

$\boldsymbol{m}_{\text{pre}} \leftarrow \boldsymbol{m}$

$\boldsymbol{G} \leftarrow \nabla \mathcal{L}(\boldsymbol{m}_{\text{pre}}; t)$ ▷ Gradient

$\boldsymbol{H} \leftarrow \nabla^2 \mathcal{L}(\boldsymbol{m}_{\text{pre}}; t)$ ▷ Hessian

$\boldsymbol{m} \leftarrow \boldsymbol{m}_{\text{pre}} - \boldsymbol{H}^{-1}\boldsymbol{G}$

$\epsilon \leftarrow |\mathcal{L}(\boldsymbol{m}_{\text{pre}}; t) - \mathcal{L}(\boldsymbol{m}; t)|$

end while

Return m

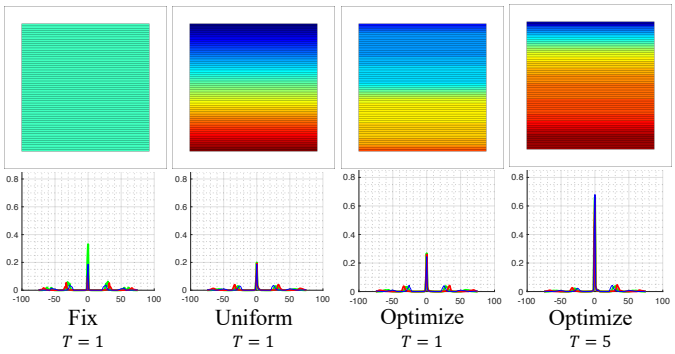


Figure 4.4: **Choice of d_0s at different locations.** The center part indicates the screen under the UDC aperture, and the white edges indicate the normal screen. The display is at 600 DPI. Colors from dark blue to red indicate d_0s determined by wavelengths from 400 nm to 700 nm.

4.4 Imaging Model and Its Characteristics

In this section, we describe the image formation model of the proposed setup and analyze its characteristics.

4.4.1 Image Formation Model

In UDCs, diffraction is usually non-negligible due to the small size of the display openings, thus we resort to wave optics in simulation. The height profiles of the first and second phase masks $h_1(x), h_2(x)$ are specified as microlens arrays as in Equation 4.8. Given a set of plane waves $p_\theta(x;\lambda)$ with unit irradiance. We can plug the $h_1(x), h_1(x), z = f_0$ into Equation 4.7, and obtain the modulated wavefront $p'_\theta(x;\lambda)$ under our design. The modulated wavefront $p'_\theta(x;\lambda)$ is then focused by the camera main lens and forms a set of blur kernels $k_\theta(x;\lambda)$, as specified in Equation 3.2. The blur kernel produced by a wide spectrum light source coming from angle θ can be computed as an integral of blur kernels with wavelength λs weighted by sensor spectral sensitivity $s(\lambda)$, $k_\theta(x) \approx \int_\lambda k_\theta(x;\lambda)s(\lambda)d\lambda$. We simulate 300 wavelengths from 400 nm to 700 nm.

Since TOLED is fully-open in y-direction and it produces a blur kernel of approximately a Dirac Delta function, the captured image can be written as

$$\boldsymbol{I}^{\text{UDC}} = \boldsymbol{K}_x \boldsymbol{I} + \boldsymbol{n} \tag{4.12}$$

where $\boldsymbol{K}_x$ is a concatenation of 1D blur kernels, $\boldsymbol{I} \in \mathbb{R}^{1024\times2048}$ is a high quality image, and $\boldsymbol{n}$ is noise. We simulate blur kernels produced by 1024 incoming directions that correspond to sensor pixel locations

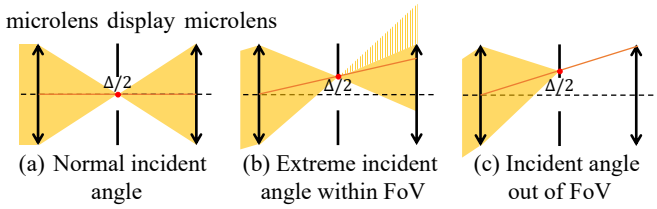


Figure 4.5: **Field of view of our design.**

along x-direction.

Reconstruction. We first apply BM3D denoiser [Dabov *et al.*, 2007] to captured images. And then we minimize the least square error between the captured $\boldsymbol{I}^{\text{UDC}}$ and estimated blurry image $\boldsymbol{K}_x\boldsymbol{I}$, and regularize the estimated $\boldsymbol{I}$ with Tikknov priors. We solve the target function using a naive iterative solver with a 'full' boundary condition and then crop the estimated $\boldsymbol{I}$ to have the same shape as $\boldsymbol{I}^{\text{UDC}}$.

4.4.2 Characterisics of Our Design

Field of View (FoV). Figure 4.5 illustrates light from different directions incident on one pair of microlenses and display pixel. Let the display pixel has an opening of Δ and the microlenses have a focal length of $\kappa\Delta$ where κ is a design choice. We choose $\kappa\Delta$ to equal display pitch, the smallest focal length if assuming spherical lenses. Normal incident light is focused to a point in the center of the display opening. As the incident angle increases, the focus point also shifts away from the center, until it reaches extreme angle. Any incident angle larger than the extreme angle is blocked by the display. The FoV of the system is

$$\text{FoV} = 2\tan^{-1}\left(\frac{\Delta/2}{\kappa\Delta}\right) \approx \frac{1}{\kappa}. \tag{4.13}$$

Light Transmission Ratio (LTR). Normal incident light passes through our setup without being blocked. As the incident angle increases, a larger portion of the light is out of the range of the second microlens. At the largest angle within the FoV, the LTR of our system is

$$\text{LTR}_{\text{min}} \approx 1 - \frac{1}{\kappa}. \tag{4.14}$$

Due to the polarization-dependent implementation of our system, the LTR is reduced by half.

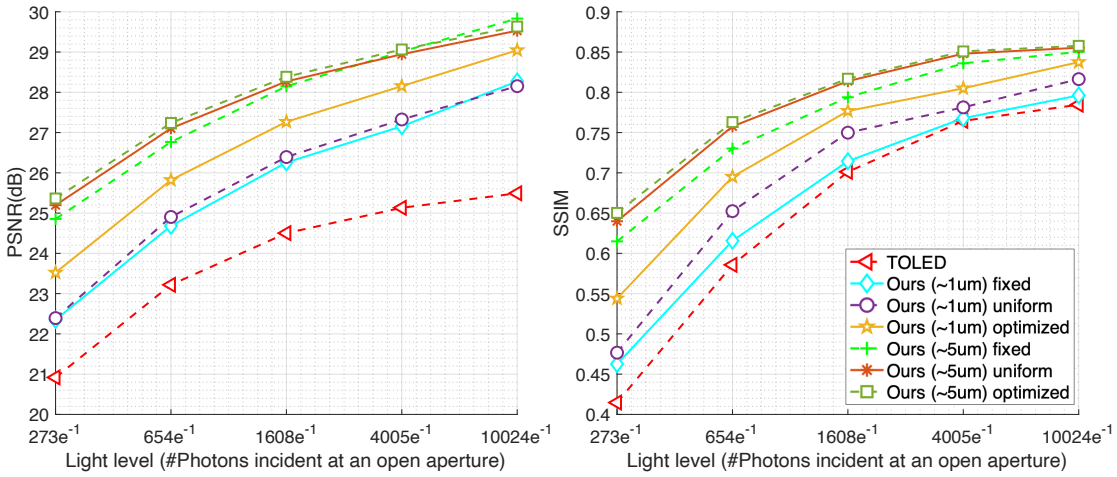


Figure 4.6: **Comparison of our setups with TOLED.**

4.5 Simulated Experiments

We design phase masks for UDCs under TOLED of pixel densities ranging from 150 to 600 DPI. All displays have an open ratio of 23.8 %. We choose the focal length to equal display pixel pitch, and thus $\kappa = 4.2$. The resulting FoV is around 14°and the LTR is between 38 % and 50 %.

We compare the proposed setup with conventional UDCs under TOLED. All displays have a pixel density of 600 DPI, equivalently a pixel pitch of 42 µm. We simulate a smartphone front camera with an aperture size of 2.3 mm and focal length of 4.67 mm. To simulate captured images, we apply spatially-varying blur kernels to ground-truth sharp images, and then add noise according to a physically-accurate noise model, and quantized to 12-bit. We emulate a sensor that has a full well capacity of 15 506 electrons and a standard deviation of 4.87 electrons, which are commonly seen in smartphone camera sensors. We set the gain to be inversely proportional to the LTR of each setup so that the captured image has consistent intensities across setups. We vary the light level by changing the number of the photons incident on an open aperture on the display from 250 to 10 000 photons. All setups are evaluated on a test set containing thirty images and using PSNR and SSIM as evaluation metrics.

Effect of Phase Masks. We compare TOLED without and with two sets of proposed phase masks that have thickness of around 1 µm and 5 µm. For each thickness, we compare three choices of wrapping heights — a fixed height determined by $\lambda_0 = 530$ nm, different heights determined by wavelengths

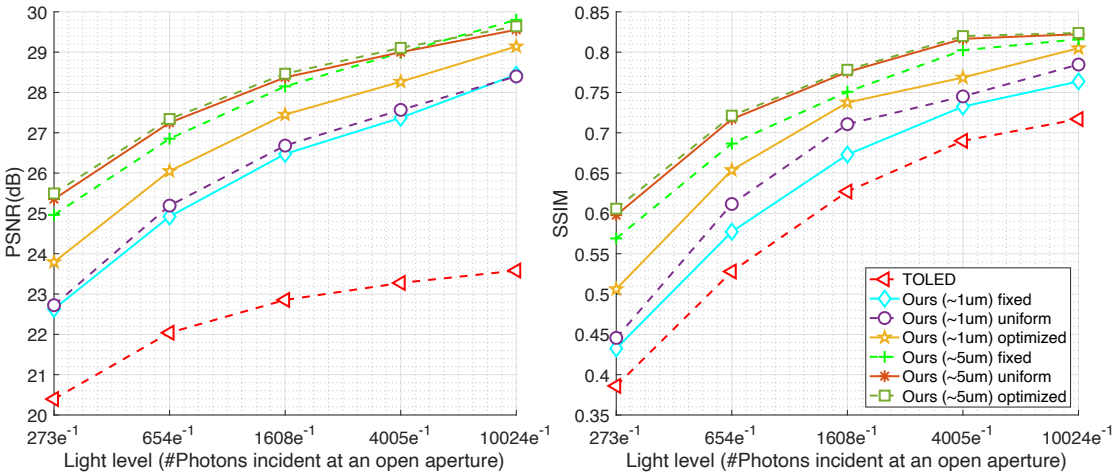


Figure 4.7: **Comparison of our setups with a traditional UDC with TOLED on validation set.**

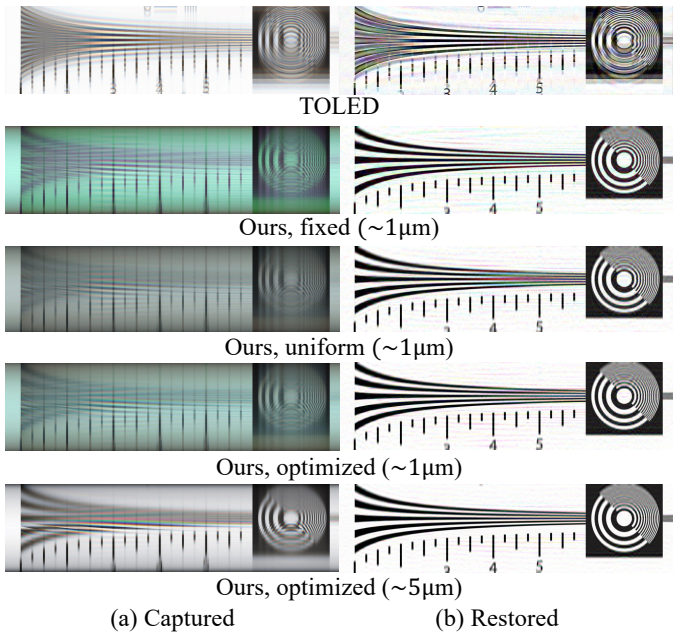


Figure 4.8: **Effect of phase mask optimization.**

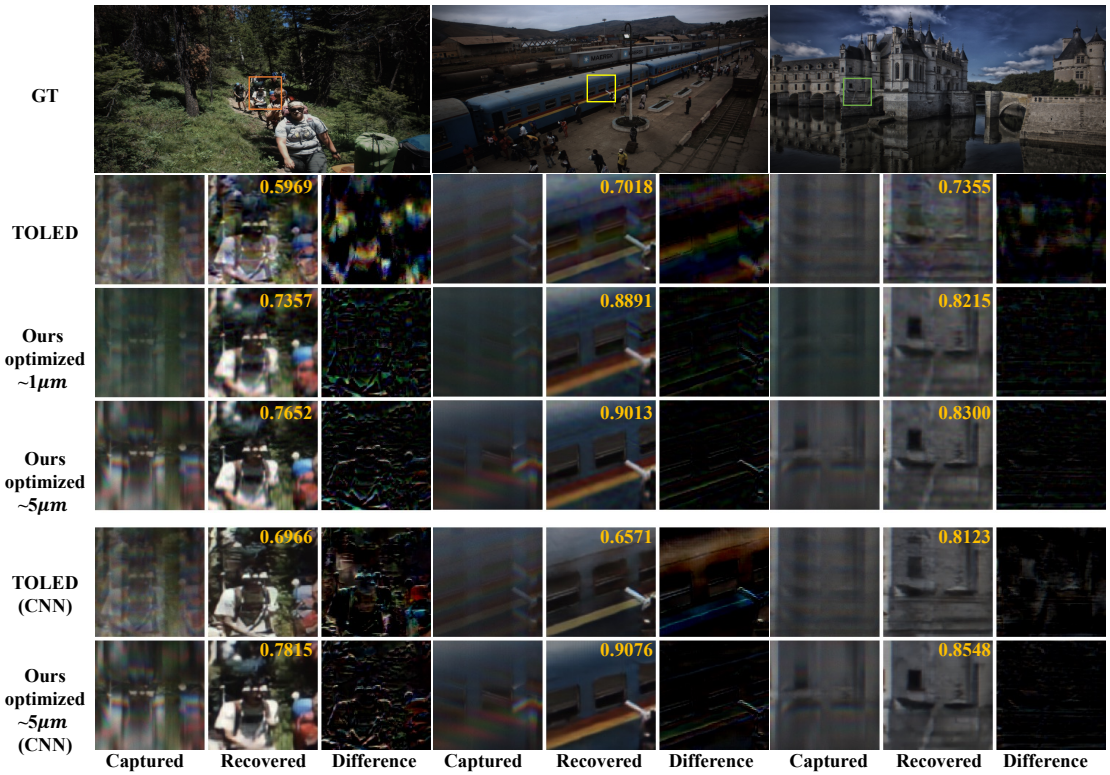


Figure 4.9: **Qualitative results from UDC under TOLED and our setups.** For each scene, columns from left to right show captured, recovered, and difference maps between the restored and ground truth. The intensities of difference maps are magnified by 2 times. We show SSIM for each restored image. Higher score means better quality.

uniformly sampled from 400 nm to 700 nm, and optimized heights. Figure 4.6 shows that the proposed setups outperform TOLED at all light levels. At 1 µm, the optimized height map largely outperforms the fixed one; while at 5 µm, different designs perform similarly. Because thinner phase masks have more phase wrappings and are more sensitive to the selection of d_0. At 5 µm, the phase mask is quite similar to a thick lens and the system performance is more consistent across different choices of d_0. Comparisons on validation set is in the supplementary.

Figure 4.7 shows additional results on the validation set comparing TOLED without and with two sets of proposed phase masks that have a thickness of around 1 µm and 5 µm. For each thickness, we compare three choices of wrapping heights. Results on the validation set are similar to those on the test set, showing that the proposed setups outperform TOLED at all light levels.

Effect of Optimization. Figure 4.8 compares qualitative results of three choices of wrapping heights on ISO 12233 resolution chart [ISO, [n.d.]]. The light level is about 10 000 photons. The restored image of TOLED contains a significant amount of ringing artifacts. The phase mask designed at a fixed height and with 1 µm thickness produces apparent chromatic aberration. The image captured under uniformly sampled heights appears less greenish. Phase masks with optimized heights further suppress chromatic artifacts and retain more details, and the one at 5 µm performs even better than the one at 1 µm. For example, it recovers more high-frequency details on the circle.

Qualitative Results. Figure 4.9 shows the qualitative results of TOLED and those of our setups. Light level is around 10 000 photons. The upper rows show results from a naive iterative solver, and the lower two rows are from the cutting-edge CNN for UDCs [Feng *et al.*, 2021]. Ours are consistently better than TOLED in SSIM. It is worth noting that TOLED results, even with CNN, contain apparent ringing artifacts. For example, in the first scene, ghosting artifacts appear on the blue and red bags, and in the third scene, there is an extra copy of the window left in the restored image.

Comparisons with Other OLED Displays. We compare our design with two display layouts commonly used in smartphone screens, TOLED and POLED [Zhou *et al.*, 2020a, 2021], and two displays layouts designed specifically for UDCs [Feng *et al.*, 2021, Yang and Sankaranarayanan, 2021b]. POLED contains a poly-amide substrate, which causes extremely low light throughput of around 8 % and produces a yellowish color shift in the captured images. The display designed by Yang and Sankaranarayanan [Yang and Sankaranarayanan, 2021b] modifies the display openings, and subsequently requires significant engineering effort to accommodate display RGB subpixels and circuits. ZTE Axon 20 phone largely reduces the display pixel density to make room for transparent regions for light to pass through. We evaluate the performance of ZTE using the PSF provided by Feng et al. [Feng *et al.*, 2021] and an estimated LTR of around 75 %. Note that reducing the pixel density results in apparent artifacts on the display.

Table 4.1 summarizes the design, LTR, and imaging performance of UDCs under various OLED displays. Ours falls into the category of requiring no change of the display openings and outperforms the other two common displays, TOLED and POLED. While Yang et al. [Yang and Sankaranarayanan, 2021b] and ZTE Axon have higher imaging quality, the modifications of the display have non-trivial negative effects on the display quality. Detailed performance at different light levels and qualitative results are in the supplementary.

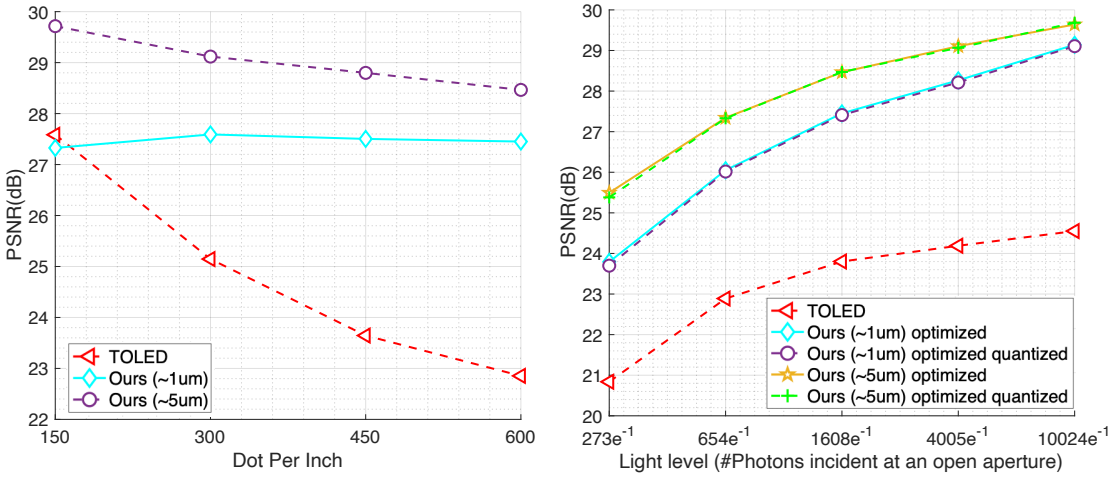
	Display Changes	LTR%	PSNR / SSIM
TOLED	–	23.8	22.42 dB / 0.59
POLED	–	8.3	26.22 dB / 0.67
Ours	–	47.6	28.01 dB / 0.75
Yang et al.	Modify layout	22.6	32.93 dB / 0.88
ZTE Axon	Low DPI	~ 75	38.24 dB / 0.96

Table 4.1: **Comparisons with other OLED displays.** TOLED, POLED, and ours do not require a change to the display openings, while Yang et al. and ZTE Axon require significant modifications to the display layout. We list averaged PSNR(↑) and SSIM(↑) across scenes from typical indoor to outdoor light levels, from 250 to 10 000 photons.

Effect of Pixel Density. Figure 4.10(a) evaluates the performance of UDCs at various pixel densities. Displays at 150 DPI are commonly used for desktop monitors and laptops; 600 DPI for high-quality cellphone displays and tablets. Light level is around 1600 photons. At 5 µm, optimized phase masks outperform TOLED at all four pixel densities, and at 1 µm ours outperform TOLED with pixel densities larger than 300 DPI. Because microlens arrays for larger pixel pitch have larger radii, and results in phase warpping artifacts when implemented as thin plates. Additional SSIM plots are shown in the supplementary. Figure 4.11(a) shows SSIM plots for UDCs at various pixel densities. The trends are similar to PSNR plots. At 5 µm, optimized phase masks outperform TOLED at all four pixel densities, and at 1 µm ours outperform TOLED with pixel densities larger than 300 DPI. Because microlens arrays for larger pixel pitch have larger radii, and results in phase wrapping artifacts when implemented as thin plates.

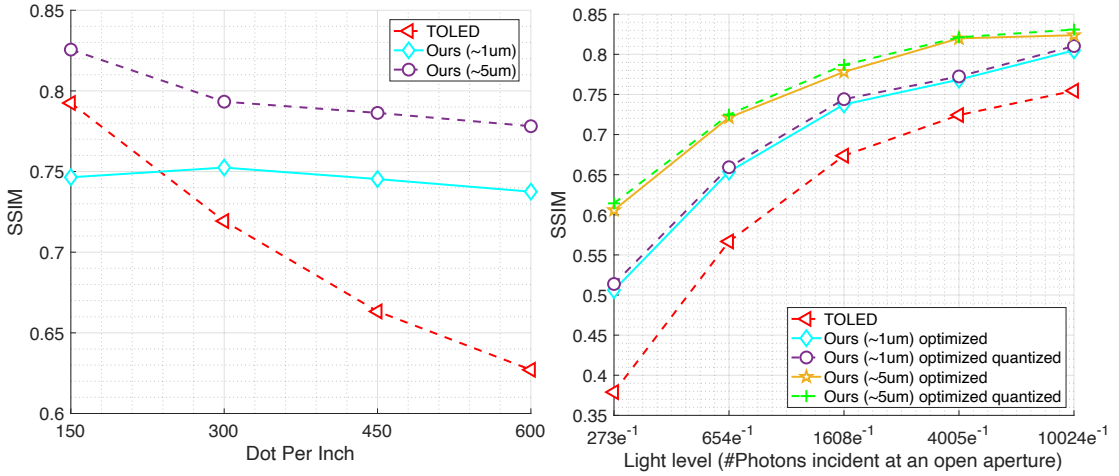
CNN-based Restoration. We adopt DISCNet [Feng *et al.*, 2021], one of the best UDC restoration networks. We utilize the 240 high-quality images in UDC dataset [Zhou *et al.*, 2020a] and simulate the captured images using the pipeline described earlier. Figure 4.1 and Figure 4.9 showcase restored images for TOLED and ours. Compared to the naive iterative solver, CNN-based restoration largely improves imaging quality. The proposed setup consistently outperforms TOLED.

Effect of phase mask quantization. In fabrication, phase masks are often quantized into discrete height maps with a step of 200 nm, for example in two-photon lithogaphy [Nanoscribe, 2007]. Figure 4.10(b) shows that quantized phase masks perform similarly as ones before quantization. Figure 4.11(b) evaluates phase masks without and with a quantization of a 200 nm step in height using SSIM. Phase masks with quantization perform similarly to ones before quantization.



(a) Effect of pixel densities (b) Effect of phase mask quantization

Figure 4.10: **Effect of (a) setups with varying display pixel densities and (b) quantization of phase masks.**



(a) Effect of pixel densities (b) Effect of phase mask quantization

Figure 4.11: **Effect of (a) setups with varying display pixel densities and (b) quantization of phase masks.**

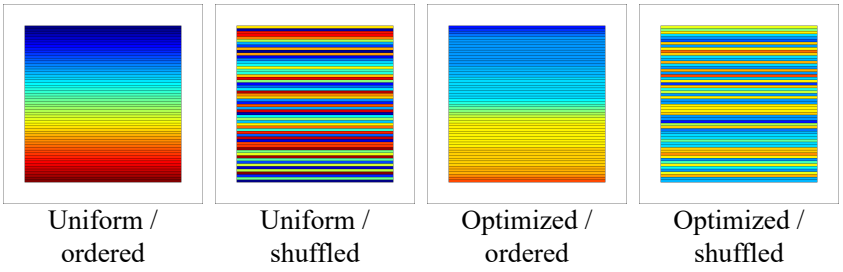


Figure 4.12: **Different ordering of d_l.**

Effect of Ordering of d_l. Given a set of folding heights d_ls for each microlens, we show the spatial ordering of d_ls has a negligible effect on the imaging performance. We compare two types of d_ls — those decided by uniformly sampled λ_0s and those decided by optimization, and for each type, we compare two orderings — sorted d_ls with an ascending order and shuffled d_ls. All phase masks are controlled by $T = 1$. Figure 4.12 illustrates two types of d_ls together with two types of orderings. Table 4.2 evaluates PSNR and SSIM of each setup at a fixed light level of around 1600 photons. Δ_{PSNR} and Δ_{SSIM} compute relative differences between the shuffled and the ordered with respect to the ordered. We can see that for each type of d_l, the ordered and shuffled have similar performance; while optimized d_ls outperform uniform d_ls. Therefore, in this chapter, we only optimize for heights d_ls and adopt an ascending order after optimization.

	PSNR	Δ_{PSNR}	SSIM	Δ_{SSIM}
uniform, *ordered*	26.69 dB	–	0.7106	-
uniform, *shuffled*	26.51 dB	−0.68 %	0.7187	1.13 %
optimized, *ordered*	27.44 dB	–	0.7368	-
optimized, *shuffled*	27.18 dB	−0.96 %	0.7396	0.38 %

Table 4.2: Effect of ordering of d_l.

Comparisons with Other OLED Displays. We compare our design with TOLED, POLED [Zhou *et al.*, 2020a, 2021], and two displays layouts designed specifically for UDCs [Feng *et al.*, 2021, Yang and Sankaranarayanan, 2021b]. Figure 4.13 shows qualitative results for scenes with an indoor light level of around 650 photons. Our design falls into the category of requiring no change to the display

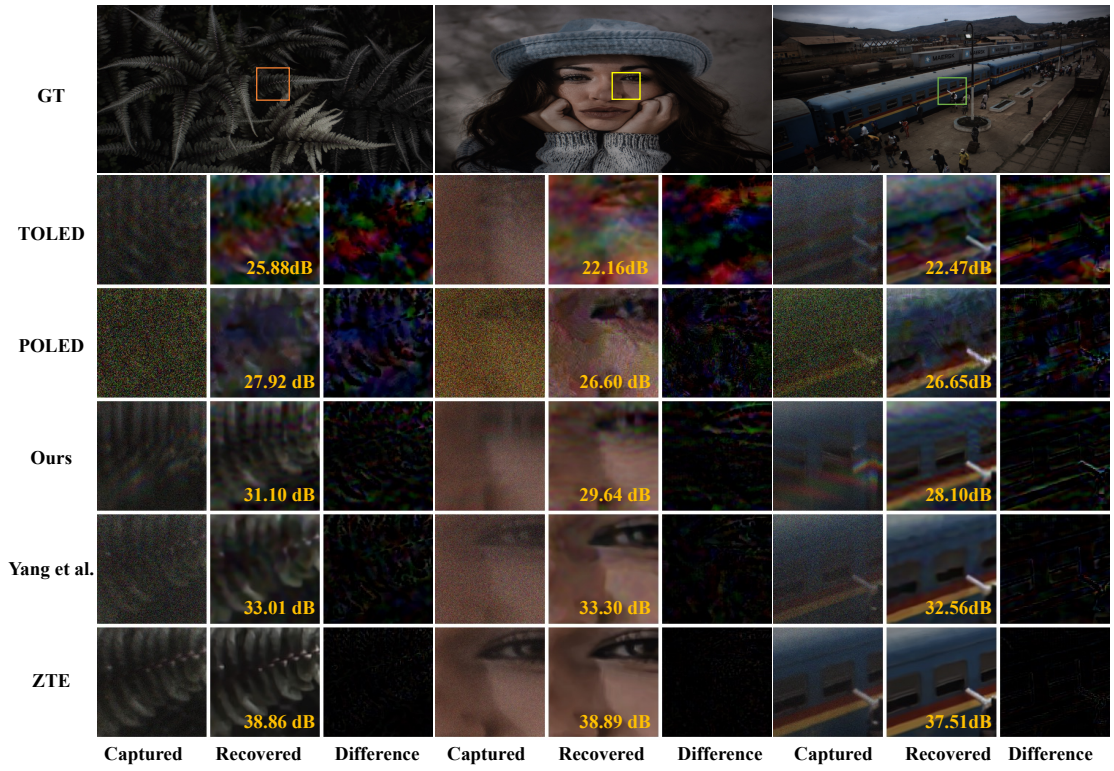


Figure 4.13: **Qualitative results comparing ours with common display layouts.** All displays are 600 DPI except for the ZTE display. ZTE display is modified to have low pixel densities to accommodate UDCs. All scenes are at an indoor light level.

openings. Compared to two other high-quality displays in this category, TOLED and POLED, ours produces significantly fewer artifacts. POLED has an LTR of around 8 % and produces photographs with the most noise. Displays designed specifically for UDCs, including Yang et al. and ZTE, have better performance than ours. However, these modifications also degrade the display quality. For example, the random tiling proposed by Yang et al. produces non-negligible visual artifacts for the display, and the ZTE display trades off pixel densities for larger transparent regions.

Figure 4.14 shows the performance of all displays for scenes ranging from indoor to outdoor light levels. Note that the performance of POLED increases fast as the light level of the scene increases, however, at a light level of 250 photons, POLED is worse than ours by around 10 dB in PSNR. This is due to its low LTR, which becomes a pronounced issue when capturing photographs of indoor scenes.

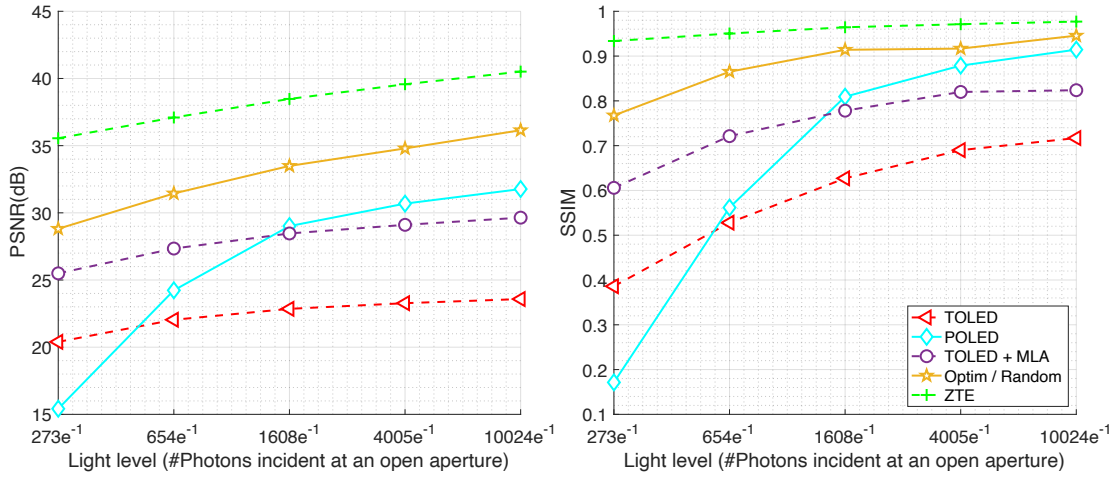


Figure 4.14: **Comparisons with other OLED displays.**

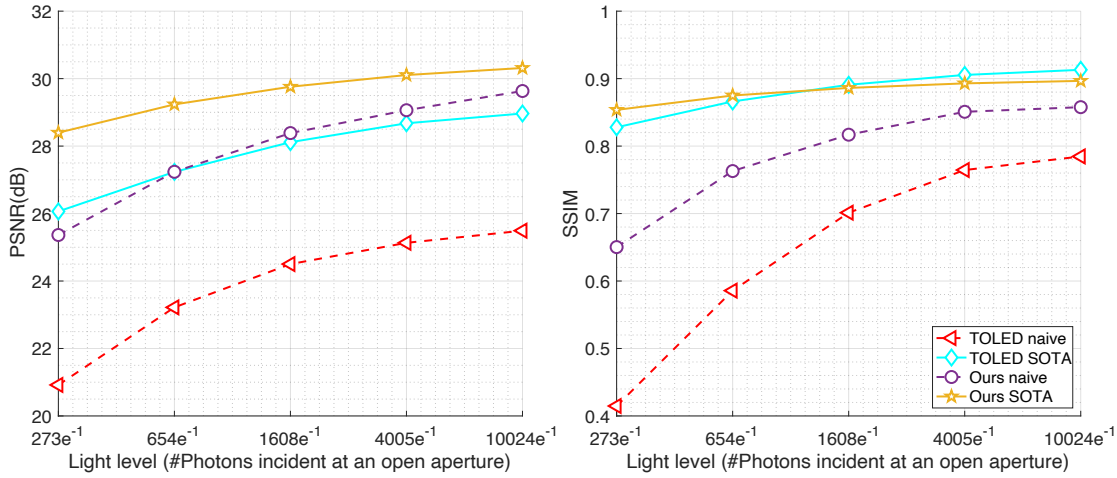
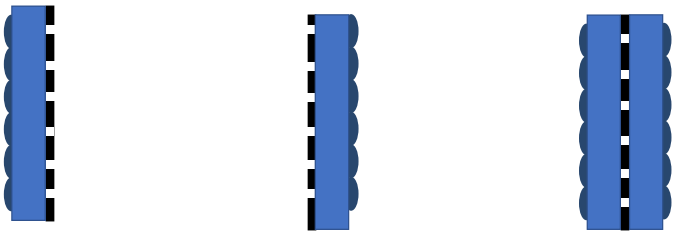


Figure 4.15: **Deblurring using an iterative solver versus using a SOTA deep neural network.**

SOTA Restoration. We compare the quantitative results from an iterative solver and from a cutting-edge deep neural network-based method for TOLED and our setup. Our setup refers to an optimized phase mask with a 5 µm thickness. First, results from SOTA deep neural network are significantly better than those from the iterative solver. For TOLED, SOTA method outperforms the iterative solver by around 4 dB; and for ours, by around 2 dB. Second, when comparing SOTA restorations for both setups, TOLED performs similarly to ours in SSIM, while worse than ours by around 2 dB in PSNR. This is because, although DISCNet recovers many sharp details for TOLED, it fails in removing widespread



Step 1: print microlens array on a photomask emulating display.

Step 2: repeat step 1.

Step 3: align the two structure by aligning the display patterns.

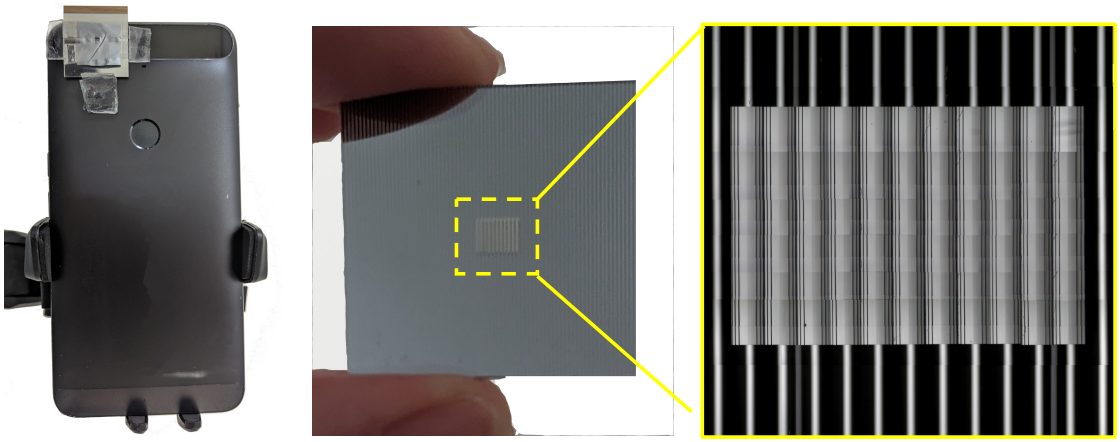
Figure 4.16: The fabrication procedure of double-sided phase masks.

ringing artifacts caused by the ill-conditioning of the PSF of TOLED. The resulting visual artifacts are ghosting effects and repetitive copies that are unfaithful to the ground-truth scene, and therefore, ours produce much more visually appealing results than TOLED.

4.6 Real Experiments

Fabrication. We use photomask with chrome patterns to emulate TOLED display that has a pitch of 336 µm, an opening of 40 µm between display pixels, and light throughput of 11.9 %. The substrate of the photomask is soda-lime and is 500 µm thick, one side of which is deposited with a thin layer of chrome and the other side is an anti-reflection (AR) coating that aids the interface finding during the fabrication of phase masks. We fold the microlens array into thin phase plates with $m = 20$ and optimize for folding heights, and the resulting thickness of microlenses is 25.28 µm. The MLA is printed on the AR coating side of the photomask. The focal length of the MLA is designed such that an incident parallel beam of light modulated by MLA is focused at the plane with chrome patterns. The total dimension is 3 mm by 3.696 mm, which approximately covers the camera aperture. As shown in fig. 4.16, we divide the fabrication into three steps: (1) print a piece of microlens array on the photomask, aligned with the chrome patterns, and (2) repeat the first step and print another piece of phase mask, (3) put two pieces together by aligning the chrome patterns. In the first and second step, two-photon lithography (Nanoscribe) is used to fabricate phase masks. The alignment is done by using the microscope built in the Nanoscribe station.

Figure 4.17 (b) shows one set of fabricated phase mask on a TOLED substrate. We align the target structure with the features on the photomask using the built-in microscope and translation stage inside Nanoscribe and print the phase mask on the photomask. In the third step, we tightly clamp the two

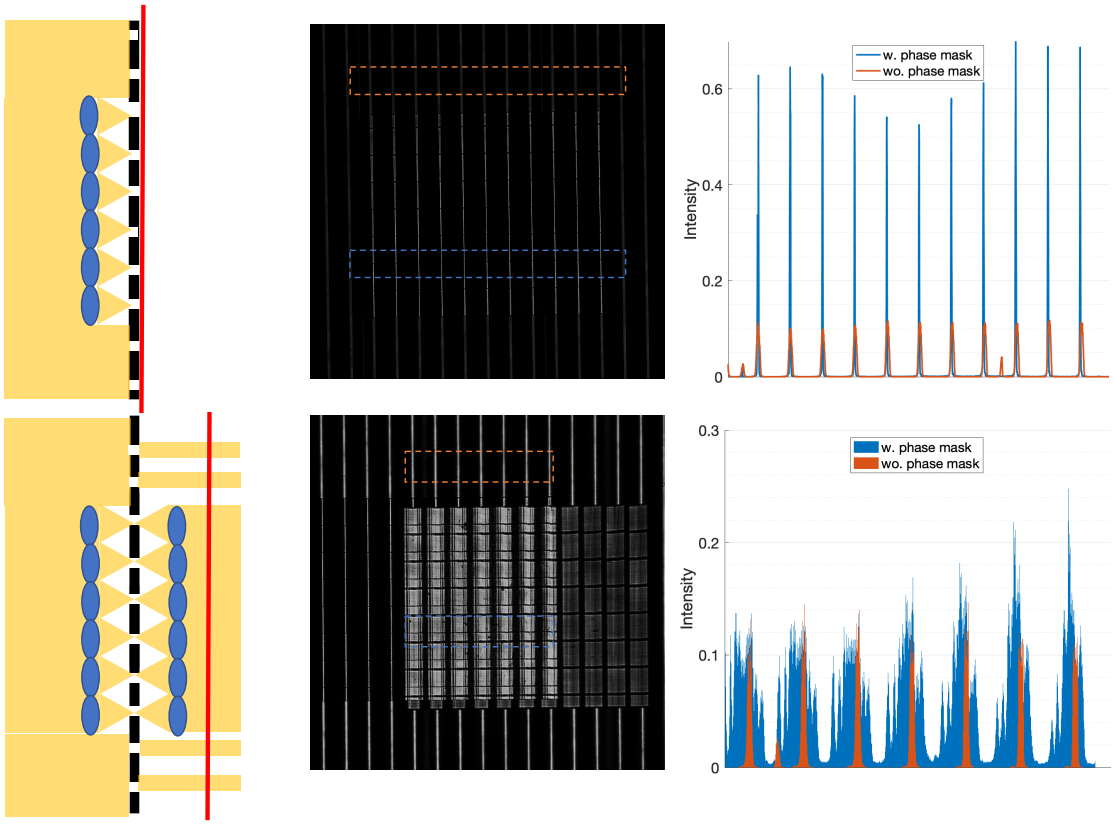


(a) Setup (b) MLA printed on a photomask that emulates TOLED display

Figure 4.17: **The hardware prototype for the proposed phase masks** We use photomasks to emulate a TOLED display and print folded microlens arrays on the photomask. The emulated display panel together with phase masks is tightly placed and taped on the rear camera of a smartphone to emulate a UDC. (a) shows the overall setup; (b) shows a microlens array aligned and printed on the TOLED display using two-photon lithography. On the right is a zoomed-in image of the printed MLA, viewed under a microscope.

photomasks and manually adjust the alignment. By maximizing the light throughput, we are able to align the two pieces with an accuracy within 4 µm along x direction.

Optical Evaluation. We use a simple optical setup to examine the quality of the fabricated phase masks. We illuminate the phase mask and TOLED display with a collimated beam of light of wavelength 530 nm, and measure propagated wavefronts using a $4f$ lens relay. In Figure 4.18 (a), the red lines showcase the target planes we measure. Upper image in (b) is the measured intensity where the center region is with the presence of one MLA. We can see that light is focused into brighter lines for the regions with MLA; while lines in other regions are dimmer. (c) is a cross-section of the two boxed regions in (b). The intensity of regions with MLA is approximately 6 times that of without MLA. Lower images in (b) and (c) are measured image and its cross-sections with the presence of two aligned MLAs. The center region is where two MLAs are aligned,, and we can see that a larger portion of light passes through the display. Examing the cross-sections, we observe that approximately 5.13 times light passes through the display with the proposed MLA. The optical evaluation confirms a single fabricated phase



(a) Target plane (b) Captured image (c) Cross section

Figure 4.18: **Evaluation of the fabricated microlens arrays.** The red line in Figure (a) illustrates the target plane we measure. (b) are captured images of the target plane, where the center part is with phase masks. Figure c are cross-sections of regions in (b) that are with and without phase masks.

mask is effective in concentrating light and two of them are able to sent a larger beam of light through the display. However, due to fabrication imperfections such as stiching effects and aberrations when approaching the field of view of objective lens, the performance of the fabricated lens is worse than ideal increase in light throughput.

Real Results. We place the fabricated phase masks and TOLED display in front of a smartphone camera. The camera has a focal length of 4.67 mm and an f-number of 2. Figure 4.19 are three indoor scenes captured with TOLED only and the proposed setup. We use the same exposure time and gain for both setups.

Images captured under proposed setup is significantly brighter than that captured with the original

(a) TOLED (b) TOLED + Phase Masks

Figure 4.19: **Captured photographs under TOLED display and our setup.** Using the proposed phase masks, the captured image is significantly brighter. This verified that the proposed phase masks guide more light through the display. However, due to the challenge in accurately align phase masks along spatial and axial directions, the phase modulation is not as well as expected in simulation.

TOLED. This verifies that the proposed phase masks successfully guide more light through the openings of the dfisplay. However, the captured images under our setup appear more blurry than simulation. We analyze that this is mainly caused by the challenge in aligning the printed phase masks and the TOLED display plane. It is fairly easy to print two (thin) phase masks; however, this would then need to be aligned with each other and with the openings on the display. Due to the small feature size of the display, the alignment tolerance is within a few micrometers along lateral dimensions. This alignment is further complicated by the precise axial distance we need to build the relay (we need 42 µm between the phase mask and the display openings for 600 DPI display). Standard substrates for phase masks have thicknesses in 200-500 µm, which is already much thicker than the required distance, let alone accurate adjustment of the distance itself to achieve the desired performance. A second challenge to resolve are artifacts from fabrication imperfections. For example, two-photon lithography with 25× objective lens has a field of view of around 0.3×0.3mm and needs to print block by block for the entire phase mask. Invariably, there are stitching artifacts as well as aberrations near the edge of each block. We believe these factors can be compensated by careful calibration of PSFs as well as more advanced post-processing algorithms and won't significantly affect the performance of the proposed setup. In summary, we believe the proposed hardware can be build if enough resources are devoted to the effort; for example, custom building a mold, aligning it to the display and curing liquid polymer to get the lenses on each side.

4.7 Discussions

In this chapter, we design phase masks to improve the image quality of UDCs. First, we show that inserting one phase mask behind the display is ineffective. Second, we propose to place two MLAs in front of and behind the display. The first MLA concentrates light to locations where the display is open, and the second recovers the original wavefront. The proposed design allows more light to reach the camera main lens and shapes the wavefront to a better condition. To ensure the display quality uncompromised, we implement microlens arrays as polarization-dependent phase masks and optimize their heights to suppress chromatic aberration. The proposed design largely improves the imaging quality of UDCs under TOLED display.

Scene at different depths. The effect of the proposed phase masks is nearly constant across scenes at different depths in the working range of selfie cameras. This is due to the small focal lengths of the proposed microlens arrays, which are at the scale of hundreds of microns.

Diffraction blur. A byproduct of inserting phase masks a short distance away from the display is that the captured images lose some details towards the edge (see Figure 4.1(b)). Similar to diffractive grating, the diffraction becomes apparent as the angle of incident light increases. In contrast, TOLED retains those details, however, with wide-spread ringing artifacts that are difficult to remove even with SOTA deep neural networks (Figure 4.1(a)). Consequently, our method yields much more visually appealing reconstructions and higher benchmark scores.

Inadequacy of Single Phase Masks: Additional Analysis In this chapter, we consider the scenario where a single phase mask is inserted tightly against the display and prove its inadequacy in improving the image quality of UDCs, as show in in Figure 4.20(a). In this section, we consider two additional scenarios shown in Figure 4.20(b)(c). We move the phase mask away from the display panel by a short distance z_0. Note that after introducing the distance z_0, plane waves that are incident on the display from different directions produce different PSFs. These spatially-varying PSFs break the convolutional imaging model, and thus prevent us from analyzing system invertibility as in Section 3.1, i.e., using the MTF as a tool for analysis. Instead, we solve for a phase mask that minimizing the difference between wavefronts observed in a UDC and a camera with fully open aperture, and examine the resulting analytical solution.

Given a UDC, we define the display as the aperture plane, and a phase mask and the camera lens are at a plane parallel to and z_0 distance away from this aperture plane. We assume wave propagation for z_0 can

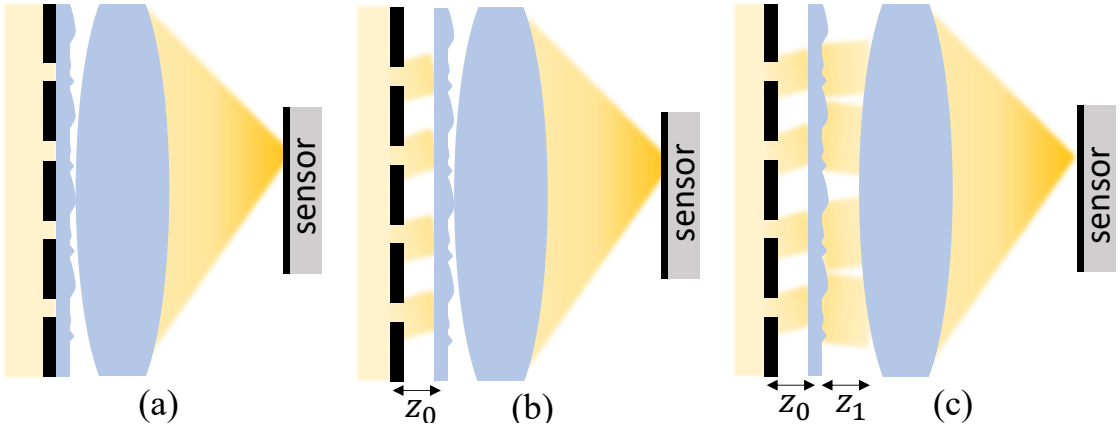


Figure 4.20: **Three scenarios where a single phase mask is inserted behind the display in UDCs.**

be well approximated by Fresnel diffraction. Now consider a plane wave incident on the display/aperture at an angle $\theta_i \in [\theta_{\min}, \theta_{\max}]$, where the bounds denote the field of view of a conventional smartphone camera. The wavefront after propagation to the phase mask, i.e., free-space propagation by a distance z_0, is denoted as u_{θ_i}; the effect of the phase mask can be denoted as a pointwise multiplication with a unit-norm phasor, and so the wavefront after the phase mask is denoted as $\phi[m]u_i[m]$, where m is a spatial index. We repeat this for a number of different incident angles $\{\theta_i, i = 1, \ldots, N\}$. Now, consider an ideal alternative, where the display (and its aperture) is not present, and we simply have the main lens z_0 distance away from the aperture plane. This ideal system provides us with a target set of wavefronts, one for each incident angle, that we denote as $T = [t_{\theta_1}, ..., t_{\theta_i}, ..., t_{\theta_N}]$.

Lemma 2 (Inadequacy of a single phase mask behind the display)**.** *Following the setup for a UDC described above, inserting a phase mask a distance away from the display panel can not decrease the Frobenius norm between the set of wavefront in the ideal camera and that in the UDC, $\|T - diag(\phi)U\|_F^2 \geq \|T - U\|_F^2$, where ϕ is the phase and amplitude modulation introduced by the phase mask.*

Proof. We solve the modulation of phase mask such that the Forbenius norm between the modulated wavefront $\text{diag}(\phi)U$ and the target wavefront T is minimized. By taking the derivative of the objective function with respect to phase modulation $\phi[m_k]$ at each location m_k and set the derivative to zero, we obtain

$$\phi[m_k] = \frac{\sum_\theta u_\theta^*[m_k]t_\theta[m_k]}{\sum_\theta u_\theta^*[m_k]u_\theta[m_k]}. \tag{4.15}$$

We can substitute wavefront incident from direction θ with that from normal direction, $u_\theta[m_k] = e^{j\frac{2\pi}{\lambda}(\theta(m_k - \frac{1}{2}\theta z))}u_0[m_k - \theta z]$ and $t_\theta[m_k] = e^{j\frac{2\pi}{\lambda}(\theta(m_k - \frac{1}{2}\theta z))}t_0[m_k - \theta z]$. The wavefront $t_0(\cdot)$ is propa-

gated from the fully-open aperture, $t_0(x) = \frac{e^{j\lambda z}}{j\lambda z}\int_{-\infty}^{+\infty} 1 \cdot e^{j\frac{k}{2z}(x-\xi)^2} d\xi = c_0$, and is thus a constant. We can simplify the expression for $\phi[m_k]$ as

$$\phi[m_k] = \frac{c_0 \sum_\theta u_0^*[m_k - \theta z]}{\|u_0[m_k - \theta z]\|_2^2}. \tag{4.16}$$

The numerator of $\phi[m]$ can be recognized as a convolution between $u_0^*[m]$ and a rectangular window running from $\theta_{\min} z$ to $\theta_{\max} z$, and the denominator is a normalization term. The periodic pixel tiling on a display panel produces a peroidic wavefront $u_0^*[m]$ with a period of pixel pitch p. When the rectangular window is significantly larger than the period of $u_o^*[m]$, $(\theta_{\max} - \theta_{\min})z = \tilde{\theta} z + np$, where $n \in \mathbb{N}$ and $\tilde{\theta} z < p$,

$$\sum\nolimits_{\theta=\theta_{\min}}^{\theta_{\max}} u_0^*[m_k - \theta z] = \sum\nolimits_{\theta=\theta_{\min}}^{\theta_{\min}+\tilde{\theta}} u_0^*[m_k - \theta z] + nc_p, \tag{4.17}$$

where $c_p = \sum_{m=0}^{p} u_0^*[m]$ is the summation of u_0^* over one period. Since $nc_p \gg \sum_{\theta=\theta_{\min}}^{\theta_{\min}+\tilde{\theta}} u_0^*[m_k - \theta z]$, (4.17) is dominated by a constant term. Thus, $\phi[m]$ is approximately a constant function.

Further, let us introduce a distance z_1 between the phase mask and lens, as is shown in scenario (c) in Figure 4.20. It is easy to see that varying distance z_1 has no effect on the PSFs, and thus the same conclusion holds. ■

Implication. The optimal phase mask that can be inserted a short distance away from the display panel is approximately a constant. Any other phase masks can only deviate wavefronts from that of an ideal camera, making the PSFs formed on the sensor less desired.

5 Spatially-Varying Gain and Binning

Pixels in image sensors have progressively become smaller, driven by the goal of producing higher-resolution imagery. However, *ceteris paribus*, a smaller pixel accumulates less light, making image quality worse. This interplay of resolution, noise and the dynamic range of the sensor and their impact on the eventual quality of acquired imagery is a fundamental concept in photography. In this chapter, we propose spatially-varying gain and binning to enhance the noise performance and dynamic range of image sensors. First, we show that by varying gain *spatially* to local scene brightness, the read noise can be made negligible, and the dynamic range of a sensor is expanded by an order of magnitude. Second, we propose a simple analysis to find a binning size that best balances resolution and noise for a given light level; this analysis predicts a spatially-varying binning strategy, again based on local scene brightness, to effectively increase the overall signal-to-noise ratio. We discuss analog and digital binning modes and, perhaps surprisingly, show that digital binning outperforms its analog counterparts when a larger gain is allowed. Finally, we demonstrate that combining spatially-varying gain and binning in various applications, including high dynamic range imaging, vignetting, and lens distortion.

5.1 Introduction

Noise and resolution are central to an image sensor, affecting the quality of the photographs acquired by it and the flavor of algorithmic post-processing required. The importance of these two factors is readily seen across a wide game: from classic problems such as denoising and super-resolution to more modern ones pertaining to (high) dynamic range. All of these challenges are routinely encountered and addressed, to some extent, *every time* a photograph is acquired. Hence, advancing the design of image sensors—the premise of this work—to combat noise and resolution can have an outsized impact on photography as well as the myriad set of applications that benefit from visual imagery.

At capture time, a sensor and its associated electronics do offer choices to a photographer to con-

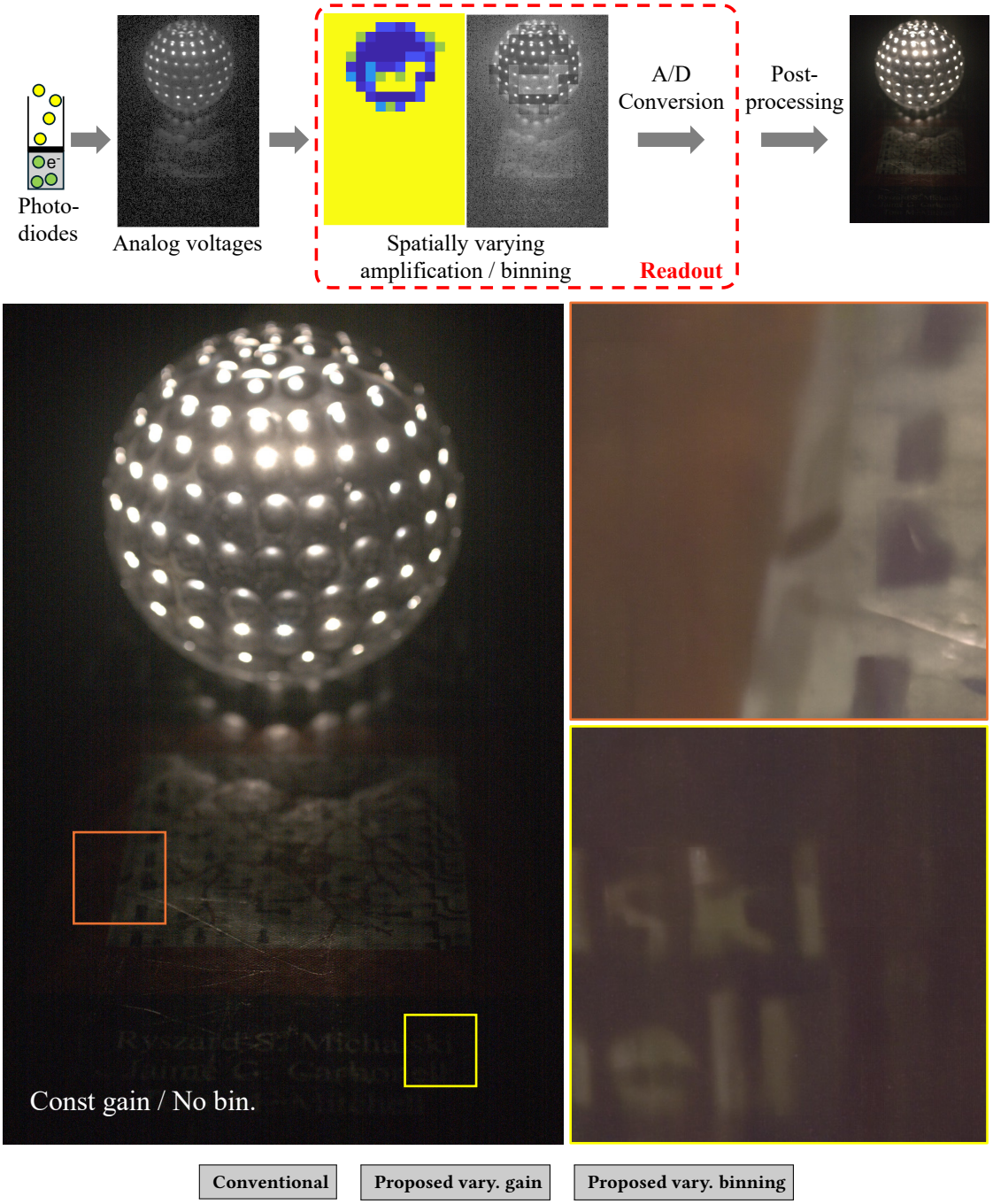


Figure 5.1: **Overview of the proposed spatially-varying readout techniques.** The upper figure is an illustration of the proposed spatially-varying readout techniques. The lower figures are captured by BFS-U3-200S6C machine vision camera and denoised by SOTA transformer-based method Restormer [Zamir *et al.*, 2022]. **Note: We kindly request readers to use Adobe Acrobat Reader to interact with the clickable buttons.** Conventional sensor uses a constant gain and no binning. Clicking between conventional and the proposed spatially-varying gain, proposed readout strategy produces much more details. Clicking between conventional and the proposed spatially-varying binning, proposed retains better contrast due to higher signal-to-noise ratio.

trol noise and resolution; this comes in the form of gain (or ISO) and binning. Gain refers to a pre-amplification of the signal before readout. Using a high gain, for example, to amplify a weak signal before digitization helps in suppressing the effects of quantization. However, a large gain also suppresses read noise, a dominant source of noise that is caused by electronics in the sensor. This improves quality in the dark regions as they are read noise dominated. However, maxing out the gain for dark regions would end up saturating the bright regions in the same scene, limiting the use of an extremely large gain. Binning, on the other hand, involves adding the charge at neighboring pixels to increase signal levels. Photon noise depends on light levels, and increasing light levels by binning increases the signal-to-noise ratio (SNR). However, binning produces larger pixels; applying the same binning size to the entire sensor would unnecessarily sacrifice fine details in the bright regions that are already resolved in high SNRs.

This chapter makes the argument for novel capabilities in image sensors in the form of *spatially-varying and scene adaptive* gain and binning. At its simplest incarnation, imagine if we had a sensor which at readout allows for each patch to be readout with a different gain and binning. The argument for a spatially-varying gain is immediately evident since, for each patch, we can select the largest gain that avoids saturation for the pixels within. Since dynamic range observed in a patch is bound to be significantly smaller than that observed in the entire image, the darkest patches will benefit from using the highest gain offered by the imager without risking saturation at brighter regions. Effectively, this expands the dynamic range of the sensor by reducing the noise floor. However, this will require some knowledge of the bright and dark patches in the scene, which we can obtain from a low-resolution snapshot. We also discuss a single-shot variant that implements a per-pixel spatially-varying gain, using the intensity observed at a previously readout pixel.

Spatially-varying binning poses a different question: can binning, which is explicitly a loss of resolution, ever improve the quality of the acquired photograph? To answer this question, we develop a simple theory that, given the light level of the scene, analyzes the optimal binning size that resolves features in the photograph. Surprisingly, a larger binning in dark regions gives better resolution, since our ability to resolve details is also strongly dependent on noise [Treibitz and Schechner, 2012]. We also analyze three binning modes: analog additive, analog average, and digital binning. We also show that digital binning can achieve better performance than both analog binning modes, when a larger gain is allowed for.

Finally, we combine spatially-varying gain and binning to reduce both read and photon noise for high dynamic range imaging, vignetting, and spatially-varying lens distortion. Figure 5.1 shows an example of the benefits to be derived using our proposed techniques.

Contributions. This chapter revisits concepts of noise, resolution, and dynamic range for image sensors, through the mindset of rethinking gain and binning.

- *Spatially-varying gain.* We propose spatially-varying gain that adapts to local scene brightness, which significantly reduces read noise for dark regions and expands sensor dynamic range.
- *Spatially-varying binning.* We establish an analysis that maps light levels to optimal binning sizes and apply it to spatially-varying binning, thereby achieving better noise-resolution tradeoffs.
- *Applications.* Proposed techniques show significant improvement in noise performance for high dynamic range imaging, vignetting, and lens distortion.

Limitations. While the proposed ROI-based techniques are relatively straight-forward to implement, per-pixel varying gain requires a modification to the readout circuits, capabilities that we are yet to implement in hardware. There is an inherent risk: changing readout circuitry could increase read noise, which might annul the improvements in the noise performance predicted by these emulations.

Impact. Our work looks into addressing the conflict between resolution, noise, and dynamic range for sensors. While increasing dynamic range for brighter parts of the scene has been studied extensively, there are few techniques that address noise floor, the limiting factor for darker regions. We show that applying spatially-varying gain, the sensor dynamic range can be expanded by an order of magnitude and resolving more signals towards the low light end. Our work also provides a way to select binning, to tradeoff resolution and noise, for general photography.

5.2 Related Work

This work touches upon noise and resolution which has been studied extensively in imaging and vision.

Noise Analysis. Early work including Clark [2016] and Healey and Kondepudy [1994] look at models for understanding noise in image sensors. Photon noise is caused due to randomness in photon arrivals at the sensor, and can be modeled as being Poisson distributed. Read noise, caused by voltage fluctuations in readout circuitry, is introduced at both pre- and post-amplifier stages. Hasinoff *et al.* [2010] provide a detailed model of read noise and the role of sensor gain in the context of high dynamic range (HDR) photography. They, and others [Martinec, 2008], point out how pre-amplifier read noise is often significantly smaller than its post-amplifier counterpart . This suggests using a larger gain to significantly suppress post-amplifier read noise. Noise and dynamic range are intricately coupled. However,

the majority of computational cameras devoted to HDR imaging [Narasimhan and Nayar, 2005, Nayar *et al.*, 2004, Sun *et al.*, 2020] focus on enhancing range at the brighter end of the light levels; this is relatively easier as it involves blocking light. A notable exception is a recent sensor [Sony, 2018] that uses microlensets of different sizes to redistribute light, providing an assorted pixel-like design, that does increase light levels at some pixels without using increased exposure times. Outside of this, there are few techniques that suppress the noise floor to enhance darker regions of the photograph—the premise of this work.

Spatially-varying gain. Hajsharif *et al.* [2014] propose a sensor where the gain is varied across pixels, with a spatial tiling that is pre-determined; for example, alternative rows of the sensor have gains of 1× or 16×, respectively. However, the gain pattern is fixed, and not adaptive to the specifics of the scene. To adapt the multiplexing patterns to different scenes, Qu *et al.* [2024] propose an enumeration method that selects the best gain and exposure pattern according to a pilot shot of the scene. Although scene adaptive, this method still uses a global multiplexing pattern for the entire image, resulting in an inherent loss of resolution at the brightest and darkest pixels. In contrast, our approach aims to avoids the loss of resolution by applying locally varying gain and binning patterns.

Pixel binning. Zhang *et al.* [2018b] propose a new pixel binning pattern for color sensors that minimizes the binning artifacts. An extension pattern design equals to a sensor interlacing original pixels with super-pixels binned from four neighbouring pixels. Jin and Hirakawa [2012] analyze the analog additive binning on color sensors and design a specialized demosaic algorithm to suppress the binning artifacts. However, these techniques cannot adaptive to the local scene brightness.

Noise and resolution. The idea that noise influences resolution has been studied formally in prior work. Treibitz and Schechner [2012] look at this interplay in the context of imaging in fog, showing how resolution loss happens not just due to loss of contrast in fog, but also due to sensor noise. We borrow the same formalism, but instead look at pixellation in place of fog. This requires certain modifications to the underlying theory, based on frequency domain methods, where the effect of pixellation and noise are readily understood.

5.3 Noise in Image Sensors

Noise in image sensors mainly consists of three types: photon noise, read noise, and dark current. The measured image i can be formed as:

$$i = \Phi\{g \cdot (l + n_D + n_{pre}) + n_{post}\} + i_0, \quad l \sim Poisson(l^*). \tag{5.1}$$

Here, l is the measured photon counts and follows a Poisson distribution with mean and variance as the expected photon arrival within the exposure time, l^*. We also assume that the sensor has a quantum efficiency of one; alternatively, we can replace the average photon arrivals with the average photo-electron arrivals, and absorb the quantum efficiency into l^*. The term n_D denotes the dark current and scales linearly with exposure time and temperature. Since we discuss photography with an exposure time of up to hundreds of milliseconds, the dark current is negligible. Finally, n_{pre} and n_{post} are pre- and post-amplifier read noise. Both are signal-independent and follow Gaussian distributions with a mean of zero and variance of σ_{pre} and σ_{post}. Note that for most sensors, σ_{post} is one or two magnitudes larger than σ_{pre}, and thus post-amplifier read noise is much more significant than pre-amplifier read noise. g is the analog gain. It usually ranges from one to hundreds. Φ denotes the analog-to-digital conversion (ADC). Since most sensors have a bit depth higher than their dynamic range, we can safely assume that the ADC noise is small. i_0 is the black level.

With the abovementioned assumption, we can simplify the noise model and estimate photon counts from the noisy digital image,

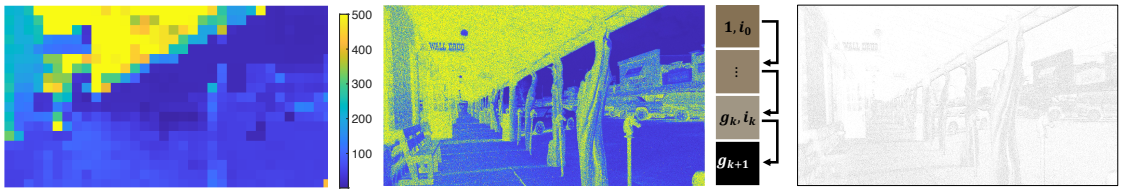
$$\hat{l} = \Phi^{-1}(i - i_0)/g = l + n_{pre} + n_{post}/g. \tag{5.2}$$

$\hat{l}$ has an expectation of l^* and a total variance of $l^* + \sigma_{pre}^2 + \sigma_{post}^2/g^2$. And SNR equals to $l^*/\sqrt{l^* + \sigma_{pre}^2 + \sigma_{post}^2/g^2}$.

We can draw two key insights from the noise model.

- *Constant gain.* A large gain g can largely reduce the post-amplifier read noise term σ_{post}^2/g^2. Since σ_{pre}^2 is much smaller than σ_{post}^2, both read noises can be suppressed. This largely benefits low-light photography where read noise standard deviation is at a similar scale as the signal l^*. As a consequence, photon noise becomes the dominant source of noise for all practical light levels. However, for a scene with a large dynamic range, increasing gain would saturate the bright regions, which limits the choice of a large gain.

- *Constant binning.* Binning neighboring pixels together increases SNR. Averaging N pixels results in an expectation of l^* and a standard deviation of $\sqrt{l^*/N}$, considering $l^* \gg \sigma_{pre}^2 + \sigma_{post}^2/g^2$. Thus the



(a)ROI-based gain map (b) Per-pixel gain map and its implementation (c) Per-pixel gain saturation map

Figure 5.2: **Implementation of spatially-varying gain.** (a) ROI-based implementation first fragments the image into multiple ROIs and then sets a gain for each ROI based on snapshot light levels. (b) per-pixel implementation adaptively sets gain according to the readout of the previous pixel. Black dots in (c) show saturated pixels with per-pixel implementation. Only around 3 % of pixel saturates.

overall SNR is increased from $\sqrt{l^*}$ to $\sqrt{Nl^*}$. However, binning comes with a side effect of pixelation and loss of resolution. The binning size N that produces the highest SNR for the low-light regions could sacrifice fine details in the bright regions.

Overview. We propose spatially-varying gain and binning to overcome read- and photon-noise limitations, and expand the dynamic range of a sensor. In section 5.4, we show that setting varying gain for varying signal levels in a single shot effectively reduces the read noise for dark scenes without saturating bright regions. In section 5.5, we propose a spatially-varying binning strategy, where pixel binning size is decided by the scene light level.

5.4 Spatially-Varying Gain

We propose to apply a spatially-varying gain, and discuss its ability to reduce read noise and expand dynamic range, as well as approaches for implementation.

Choice of gain. Given an estimated scene light level $\hat{l}, \mathbb{E}(\hat{l}) = l^*$, we aim to find the *largest* gain such that the amplified signal gl saturates with a small probability. We adopt the Gaussian-Heteroskedastic noise model [Foi *et al.*, 2008] and approximate the amplified signal with a Gaussian distribution,

$$g \cdot (l + n_{pre}) + n_{post} \sim \mathcal{N}(g\hat{l}, g^2\hat{l} + g^2\sigma_{pre}^2 + \sigma_{post}^2).$$

Note that $g\hat{l}$ is intended to be close to well-capacity l_{wc} and is much larger than the read noise variances, thus the total variance can be further simplified to $g^2\hat{l}$. To make the probability of saturation small, we

set a gain value g that satisfied

$$l_{wc} \approx g\hat{l} + \eta g\sqrt{\hat{l}}. \tag{5.3}$$

Here, when the parameter $\eta = 2$, the pixel saturates with a probability of 2.2%.

5.4.1 Design of spatially-varying gain

We provide two strategies for design for spatially-varying gain.

Two-shot ROI-based strategy. We first capture a noisy snapshot using constant gain. We fragment the snapshot into multiple region-of-interests (ROIs), so that each ROI has a smaller dynamic range, and compute the optimal gain for each ROI. The computed gain map is used to inform the subsequent capture and readout. The advance of ROI-based methods is that it only requires a minimal change to today's readout circuitry in the form of being able to select multiple ROIs, instead of one. However, it requires two images, leading to potential of motion-related artifacts (although their effects are not in the form of blur as we discuss later).

Single-shot per-pixel strategy. The per-pixel strategy sets gain adaptively during readout and only requires a *single shot*. Since natural images are typically piecewise smooth, we assume the light level of one pixel is similar to its neighbors. Therefore, we use the readout value of one pixel to set the gain for the subsequent pixel (see fig. 5.2(b)). Specifically, from the readout value and gain of the k-th pixel, we estimate its light level l_k, and use l_k as an approximate estimate for the next pixel's light level, $\widehat{l}_{k+1} \approx l_k$. By substituting $\widehat{l}_{k+1}$ into eq. (5.3), we obtain the gain g_{k+1} for $(k+1)$-th pixel and set it in the ADC circuits for readout. When $k-$th pixel saturates, we reset $g_{k+1} = 1$. Emperically, we set $\eta = 4$ and only around 3 % of pixels saturates in the captured image, as shown in Fig. 5.2(c).

5.4.2 Improving dynamic range in a single-shot

Spatially-varying gain expands the dynamic range of a sensor by one or two magnitudes within a single exposure. This is because the dynamic range of a sensor is typically decided by the ratio of its well-capacity and read noise floor, $l_{wc}/\sqrt{\sigma_{pre}^2 + \sigma_{post}^2/g^2}$. To capture a high dynamic range scene, a conventional sensor with a constant gain is limited to use a small g to avoid saturating the bright objects, and the variance of post-amplifier read noise is much larger than that of pre-amplifier read noise, thus $\sigma_{post}^2/g^2 \gg \sigma_{pre}^2$ and the dynamic range of a conventional sensor is approximately l_{wc}/σ_{post}. In contrast, the proposed spatially-varying gain uses large gain g to capture dark regions, effectively reducing post-amplifier read noise σ_{post}^2/g^2 to negligible, $\sigma_{post}^2/g^2 \ll \sigma_{pre}^2$, and the resulting dynamic range becomes

l_{wc}/σ_{pre}. Moreover, spatially-varying gain effectively reduces the read noise and leaves photon noise the bottleneck of image quality. Next, we will look into reducing photon and read noise with a proposed technique, spatially-varying binning.

5.5 Spatially-Varying Binning

Small pixels gather less lights, and one way to increase light levels is to bin pixels. This raises the following question: what is the optimal binning size that maximizes image quality? We show this optimal binning is tightly coupled with scene light levels. This is because effective resolution increases as pixel gets smaller, but decreases with increasing noise levels—a side-effect of small pixels—indicating a sweet spot that best trades off resolution and noise.

5.5.1 Analysis of the optimal pixel pitch

Given a scene light level, what is the optimal pixel size that achieves the highest effective resolution? We define *effective resolution* as the frequency whose ratio between the noise-free signal contrast and the measured noise standard deviation is greater than a predefined threshold SNR_t. The resolution with SNR equals SNR_t is the highest effective resolution, and frequencies smaller than it are all effective.

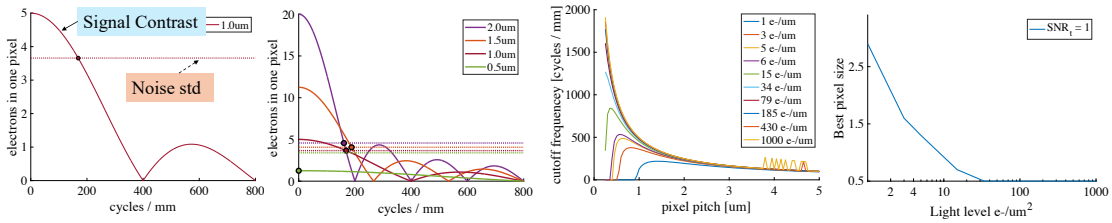
We characterize scene contents of various feature sizes by examining sinusoidal signals with varying frequencies. Consider a camera with an ideal lens and a sensor with a pixel pitch p µm. The measured signal can be modeled as a convolution between the signal and a box function induced by the pixel size. This models the expected noise-free signal measurement by incorporating the blurring effect of pixel pitch. The contrast of the noise-free measurement of a sinusoid with frequency f_0 and a light level of l_0 photons per unit area is,

$$c(f_0; l_0, p) = |\max l^* - \min l^*| = l_0 p^2 \frac{sin(\pi p f_0)}{\pi f_0}. \tag{5.4}$$

We refer readers to the appendix for detailed derivation. We plug in the expected signal l^* to the noise model shown in eq. (5.1) and obtain the noisy measurements. The total noise variance is,

$$\sigma(f_0; l_0, p) = \sqrt{\sigma_{pre}^2 + \sigma_{post}^2/g^2 + l_0 p^2/2}, \tag{5.5}$$

where read noise is independent of pixel size, and shot noise variance increases proportionally to pixel area p^2. With a specified threshold SNR_t, we can find the cutoff frequency f_{cutoff} and all features below f_{cutoff} are considered resolved.



(a)Signal contrast $c(\boldsymbol{f}; l_0, p)$ and noise std $\sigma(\boldsymbol{f}; l_0, p)$.

(b) Signal contrast $c(\boldsymbol{f}; l_0, p)$ and noise std $\sigma(\boldsymbol{f}; l_0, p)$ curves for increasing pitches p.

(c) At different light levels, the pitch p and its effective resolution.

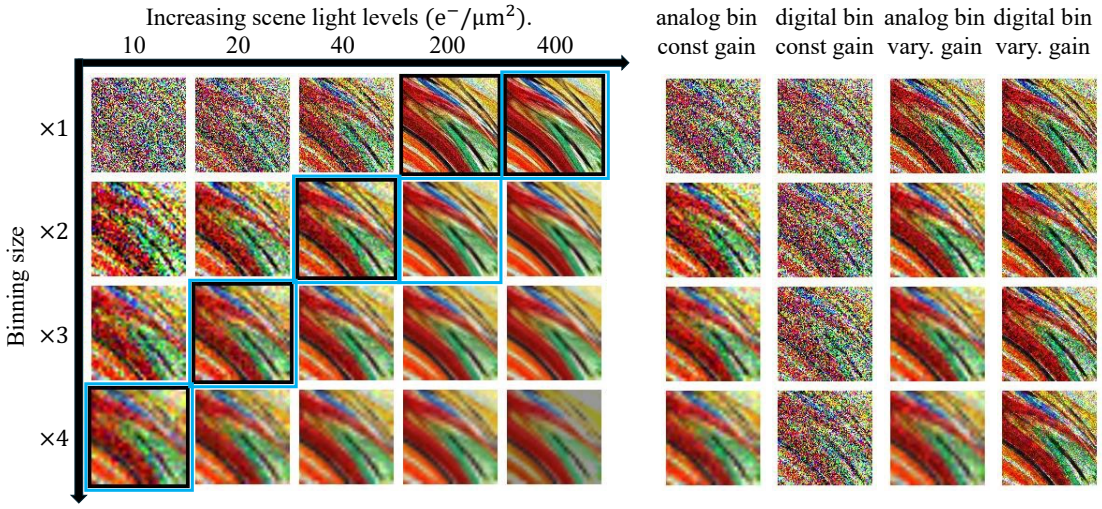
(d) A function that maps light levels to optimal pixel pitch with a given SNR_t.

Figure 5.3: **Analysis of optimal binning.** Left figure shows the signal contrast, noise floor, and cutoff frequencies for pixel pitches from 0.5 µm to 2.0 µm under a fixed light condition. Right figure shows functions that map scene light levels to optimal pixel sizes under the required SNR threshold.

Fig. 5.3(a) shows one set of signal and noise curves for pitch p_0 and at a given light level l_0. Considering $\text{SNR}_t = 1$, the two curves intersects at the cut-off frequency $f_{\text{cutoff}}(l_0, p_0)$. This means that given light level l_0, a sensor with pixel pitch p_0 can resolve features with maximum frequency $f_{\text{cutoff}}(l_0, p_0)$. As shown in fig. 5.3(b), we analyze for varying pixel pitches, from 0.5 µm to 2.0 µm, given the same light level l_0, and find their cutoff frequency $f_{\text{cutoff}}(l_0, p)$. Based on (b), we obtain the pixel and its effective resolution at light level l_0, which is shown as the blur curve in (c), and find out the optimal pitch for l_0, $p^*(l_0) = \arg\max_p f_{\text{cutoff}}(l_0, p)$. We repeat this analysis for varying light levels from extremely dark to bright. As shown in fig. 5.3(d), this allows us to analyze the optimal pixel pitch that achieves the highest effective resolution for each scene light level. Note that the optimal pixel size is a function of light level, and decreases as the scene gets brighter. This confirms the intuition that small pixels suit bright scenes and large pixels suit dark scenes. The SNR threshold SNR_t is a hyperparameter and is determined empirically by examining a set of measured image quality. We use $\text{SNR}_t = 4$ for all our experiments.

To evaluate the effectiveness of our binning theory, we simulate a texture patch captured under various light conditions and binning sizes, as shown in fig. 5.4(a). Under most light conditions, the predicted binning sizes (black boxes) match the one with best image quality (blue box) indicated by LPIPS scores [Zhang *et al.*, 2018a].

Constrasts versus SNR. Conventional MTF computes the degraded *contrasts* compared to the original contrast with varying feature sizes and the contrast is typically defined as $\frac{I_{\max} - I_{\min}}{I_{\max} + I_{\min}}$. This typically assumes noise-free signals. With the presence of noise, the intuition is that it becomes harder to differentiate the peaks and valleys in the signal. However, conventional contrasts actually become larger on noisy signals, since $I_{\max}$ is larger than clean $I^*_{\max}$ and $I_{\min}$ is smaller than clean $I^*_{\min}$. This clearly counters



(a) Comparing bin sizes under increasing light levels. (— is bin size with best LPIPS, — is the predicted bin size)

(b) Comparing analog and digital bin under light level $20e^{-}/\mu m^2$.

Figure 5.4: **Optimal bin sizes under different light levels.** (a) shows the effect of increasing bin sizes under from dim light condition to sufficient light. All are analog additive binning for unit pixel size 0.5 μm. Black box shows the predicted optimal binning, and blue box shows the binning size with best LPIPS score [Zhang *et al.*, 2018a]. (b) compares analog and digital binning with small and large gain under dim light conditions.

our intuition. What notion should we use to capture this intuition?

If we view noisy pixels at the peak and the valley as two classes, and the pixel values follow two Gaussian distributions $\mathcal{N}(I^*_{\text{max}}, \sigma^2)$ and $\mathcal{N}(I^*_{\text{min}}, \sigma^2)$. We look into existing metrics that capture the divergence of the two.

- **KL divergence.** We use Kullback–Leibler divergence to capture the difference between the two distributions, $D_{KL}(i_1||i_0) = \frac{1}{2}\frac{(I^*_{\text{max}} - I^*_{\text{min}})^2}{\sigma^2}$. Note that as noise standard deviation σ gets larger, KL divergence becomes smaller, successfully capturing the reduced difference between peaks and valleys.
- **Fisher linear discriminant rule.** Similarly, we can use Fisher linear discriminant rule to measure the intra- and inter-distribution variance, $SS = \frac{S_{\text{between}}}{S_{\text{within}}} = \frac{1}{2}\frac{(I^*_{\text{max}} - I^*_{\text{min}})^2}{\sigma^2}$. This notion also successfully captures our intuition, since SS is inversely proportional to noise variance σ^2.

Interestingly, both KL divergence and Fisher linear discriminant rule are equivalent to the signal-to-noise ratio in our case. Therefore, we use SNR as a tool to analyze effective resolution for noisy measurements and showed the above analysis.

Table 5.1: **A summary of binning modes.** Assume equal weights for neighboring pixels and normalize the combined signal to the same level.

Binning modes	Imaging model $\hat{I}$	Total noise variance
No binning	$I + n_{pre} + \frac{n_{post}}{g}$	$I^* + \sigma_{pre}^2 + \frac{\sigma_{post}^2}{g^2}$
Additive binning	$\frac{1}{N}\sum_{i=1}^{N}\{I_i + n_{i,pre}\} + \frac{n_{post}}{N\widehat{g}}$	$\frac{I^*}{N} + \frac{\sigma_{pre}^2}{N} + \frac{\sigma_{post}^2}{N^2\widehat{g}^2}$
Average binning	$\frac{1}{N}\sum_{i=1}^{N}\{I_i + n_{i,pre}\} + \frac{n_{post}}{g}$	$\frac{I^*}{N} + \frac{\sigma_{pre}^2}{N} + \frac{\sigma_{post}^2}{g^2}$
Digital binning	$\frac{1}{N}\sum_{i=1}^{N}\{I_i + n_{i,pre} + \frac{n_{i,post}}{g}\}$	$\frac{I^*}{N} + \frac{\sigma_{pre}^2}{N} + \frac{\sigma_{post}^2}{Ng^2}$

5.5.2 A sensor with varying pixel pitches through binning

We implement the optimal pixel pitches through pixel binning. We discuss three binning types that are commonly seen in sensors — analog additive, analog average, and digital binning.

- *Analog additive binning.* Both analog binning modes are conducted on the analog signals. After photoelectron counts are converted into analog voltages, the sensor *sum* up the analog voltage of N neighboring pixels.
- *Analog average binning.* The sensor weighted *average* the analog voltages of N neighbouring pixels.
- *Digital binning.* Signals are binned after read out as digital values.

We summarize the noise models and total noise variances of all binning modes in table 5.1. All three binning modes reduce the photon noise by N times. At first sight, analog additive binning reduces the most read noise, but when we take different gains into account, digital binning becomes the best. This interesting observation comes from the fact that digital binning does not raise the analog voltage while additive binning scales voltage up by around N times. To avoid saturation, additive binning uses a gain that is around N times smaller than that of digital binning. With $\widehat{g} = g/N$, read noise variance of additive binning becomes $\sigma_{post}^2/(N^2\widehat{g}^2) = \sigma_{post}^2/g^2$, and is larger than that of digital binning.

Fig. 5.4(b) compares analog and digital binning under different gain settings. When gain is restricted to be low, spatially-varying analog binning can significantly increase the SNR for dark patches, where as digital binning suffers to recover details from noise. However, if a larger gain is allowed, digital binning is superior to analog binning as it reduce noise without sacrificing resolution.

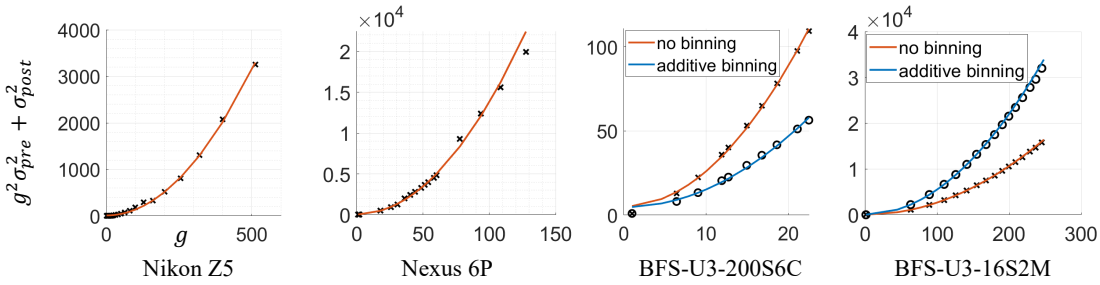


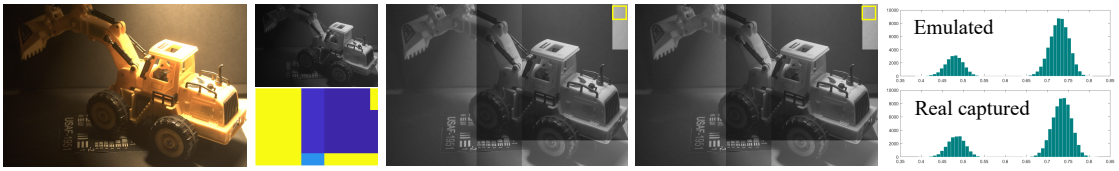
Figure 5.5: **Pre- and post-amplifier read noise calibration.** x-axis is gain and y-axis is the total noise variance. Base ISOs are normalized to 1.

Table 5.2: **A summary of sensors.** σ_{pre} and σ_{post} are in analog digit unit.

Camera	Type	Max. Gain	Binning	σ_{pre}	σ_{post}
Nikon Z5	Mirrorless	ISO 51200	—	0.11	3.53
Nexus 6P	Smartphone	ISO 7656	—	1.17	7.39
BFS-U3-200S6C	Machine Vision	27 dB	1×1	0.46	2.27
			2×2 (avg)	0.32	2.17
BFS-U3-16S2M	Machine Vision	48 dB	1×1	0.52	4.55
			2×2 (add)	0.23	1.47

5.6 Emulated Results on Real Hardware

Sensor Calibration. We calibrate and capture data with four cameras: Nikon Z5 (mirrorless camera), Nexus 6P (smartphone main camera), FLIR BFS-U3-200S6C (machine vision with Bayer color filter arrays), and FLIR BFS-U3-16S2M (machine vision monochrome). Table 5.2 summarizes their key specifications. We calibrate their pre- and post-amplifier read noise by closing the cap on sensors and capturing dark noisy frames, and capturing images with varying gain. As shown in fig. 5.5, we fit a quadratic curve between σ^2 and g and estimate the coefficients σ_{pre} and σ_{post}. As the two machine vision cameras also support analog binning, we repeat the calibration with analog binning. We summarize the calibrated noise statistics in Table 5.2. We can see that post-amplifier read noise is usually around one magnitude larger than pre-amplifier noise.



HDR scene | Snapshot + Gain map | Emulated from gain stack | Real captured by sweeping ROIs | Intensity histogram of yellow patches

Figure 5.6: **Emulator versus real-capture by windowing.** All images are captured by BFS-U3-200S6C camera. The emulated image is composed from a real captured gain stack. Real capture is obtained by sequentially setting ROIs using low-level API, integrating, and read out with the optimal gain.

Emulator. We emulate spatially-varying gain and binning by capturing a gain stack, sweeping from the lowest to highest gain settings, computing gain maps and binning maps based on the lowest gain snapshots, and compositing ROIs from corresponding bursts into one image. For mirrorless and smartphone cameras, we only capture one gain stack, and for machine vision cameras, we capture gain stacks with and without analog binning.

Note that both machine vision cameras supports random access to region-of-interests (ROIs), and allow user to specify gain and binning for each ROI. We can thus sequentially set ROI and readout each ROI using optimized parameters. In fig. 5.6, we show that for ROI-based spatially-varying techniques, emulating from a gain stack produce the same noise statistics as real-capture through windowing.

Effect of spatially-varying gain. Shown in fig. 5.7, we compare constant gain, alternating gain [Hajsharif *et al.*, 2014], and the proposed spatially-varying gain on Nikon Z5 and Nexus 6P. For each scene and from top to bottom, we show gain map, one bright patch, and one dark patch. When captured with lowest ISO, dark regions are excessively noisy and the details are largely degraded by noise. When captured with highest ISO, read noise is suppressed, but the bright regions are over-exposed. Alternating ISOs emulates a sensor with lowest and highest ISOs every alternating rows [Hajsharif *et al.*, 2014]. However, their vertical resolution is reduced by half, since only half rows are valid for the bright regions and extremely dark regions. Finally, the proposed spatially-varying gain captures details in the bright region without sacrificing resolution, and effectively reduce sensor noise in the dark regions. This reflects an expansion of the sensor dynamic range without reducing resolution.

Effect of spatially-varying binning. We evaluate the effect of spatially varying binning on BFS-U3-16S2M monochrome camera and BFS-U3-200S6C color camera. As shown in the first two columns in

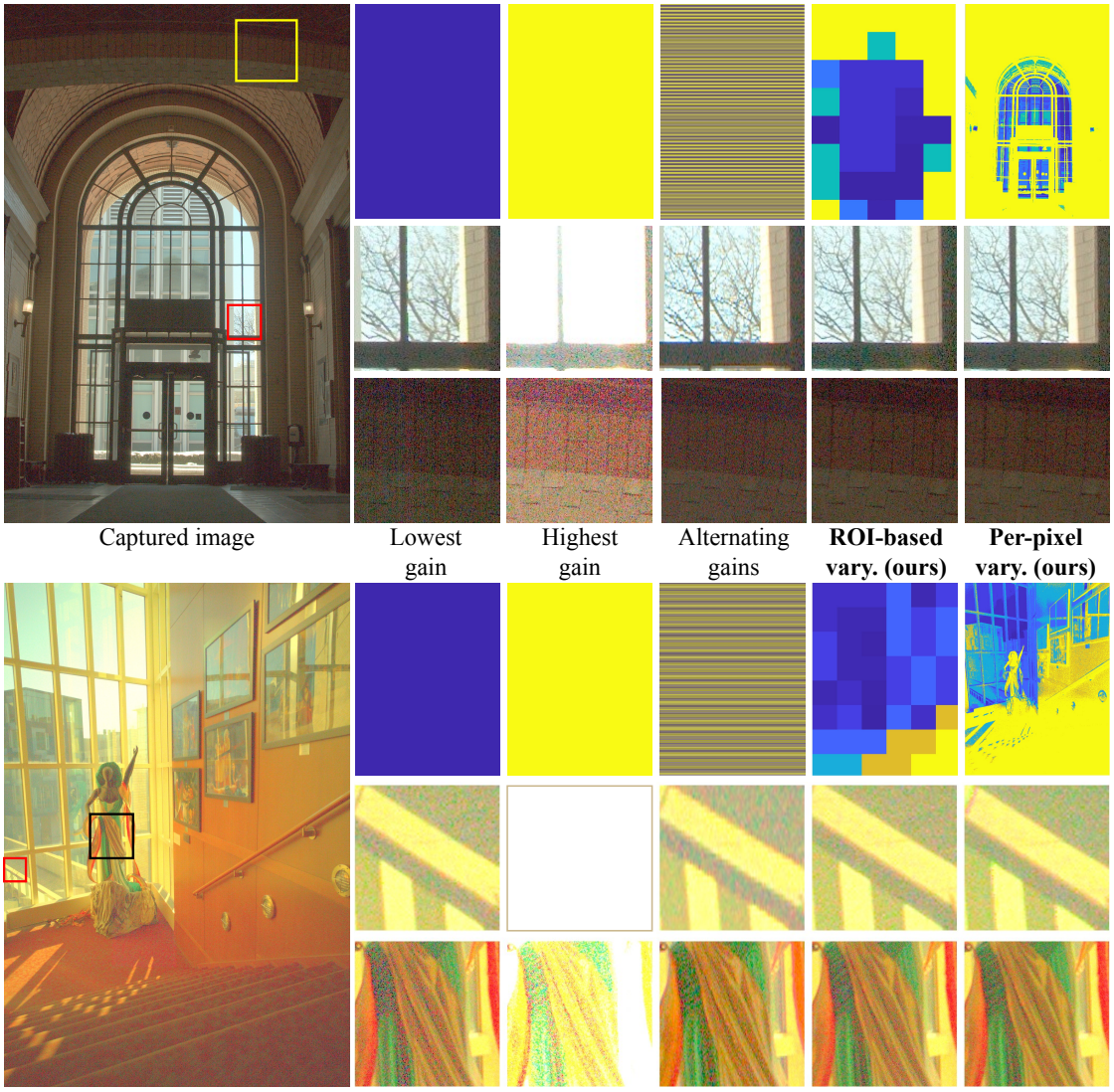


Figure 5.7: **Comparisons of gain modes in HDR.** Upper figures are captured by Nikon Z5 and lower by Nexus 6P. The lowest and highest ISOs for Nikon Z5 are 100 and 51200 and for Nexus 6P are 60 and 7656.

fig. 5.8, we capture HDR scenes with and without 2×2 analog binning. Without binning, regions in the dark are extremely noisy. With uniform analog binning, the noise is reduced, but it sacrifices the resolution in the bright regions. For example, point lights appear square-ish in the upper example, and the leaves appear blurry in the lower example. Compared with those without binning, our methods reduce noise in the dark regions, and compared to those from uniform binning, ours retain appealing details in

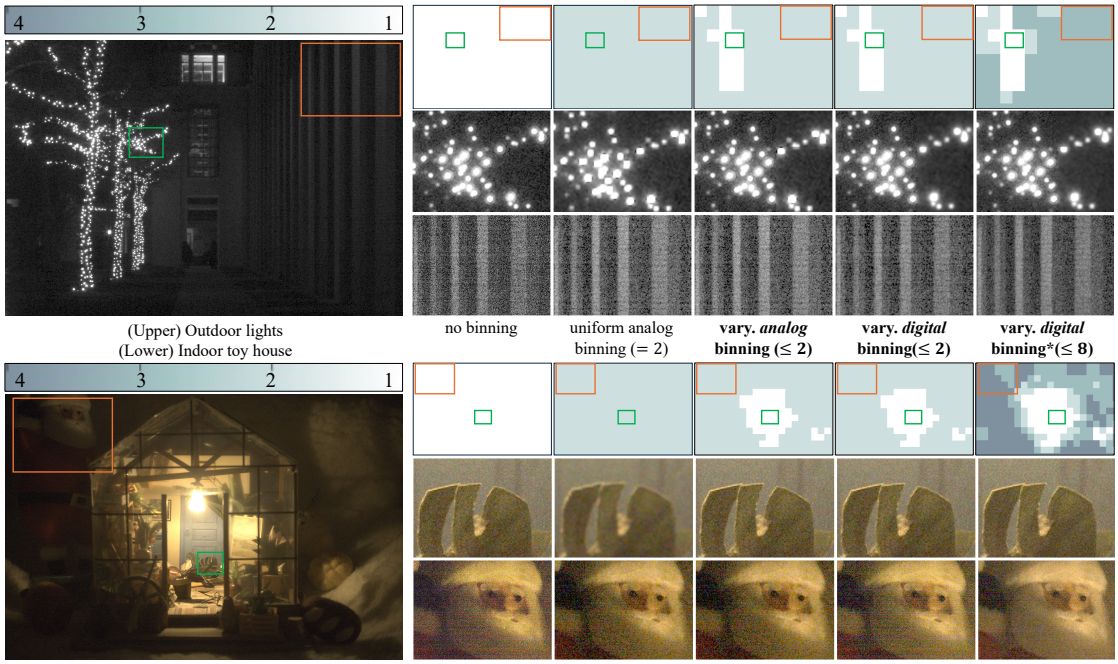


Figure 5.8: **Comparisons of binning modes in HDR.** Upper is captured by BFS-U3-16S2M and lower BFS-U3-200S6C. The first two columns are from off-the-shelf binning modes and the last three columns are proposed spatially-varying binning. Since both cameras only support analog binning up to 2×2, we demonstrate binning with larger sizes on digital binning. *The last column is digital binning under varying gain.

the bright regions. With the same gain, digital binning is slightly worse than analog additive binning, which confirms our analysis. Finally, we examine spatially-varying binning up to size eight, using digital binning since analog binning only supports up to size two; this achieves the best performance. This suggests that when light conditions is extremely low, image sensors could benefit from analog additive binning larger than two. If large gain is allowed, conducting digital binning would produce the best quality.

Effect of post-processing. Fig. 5.9 compares images denoised by Restormer, a SOTA transformer-based image restoration network [Zamir *et al.*, 2022]. We use the pretrained "Real denoising" checkpoint and denoise captured images with a patch size of 720×720 and an overlap of 32 pixels. The captured images are demosaicked and gamma corrected before fed into the network. For spatially-varying binning (up to 2×2 analog binning) images, we feed both the full resolution and downsampled images to the

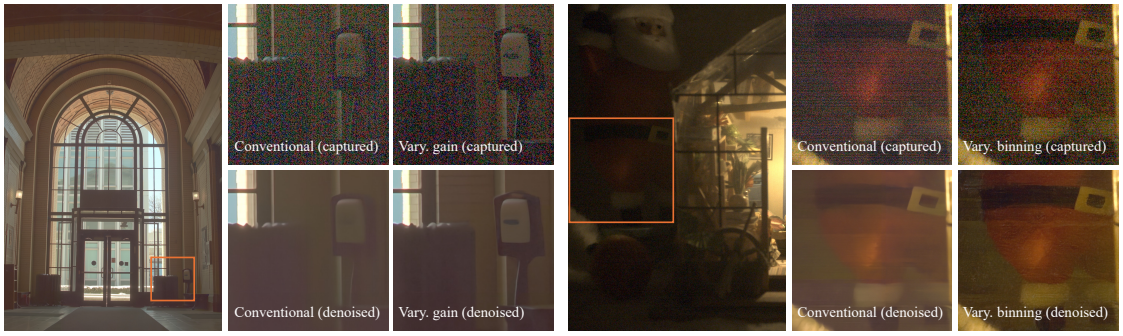


Figure 5.9: **Effect of transformer-based restoration networks.** Left example is captured by Nikon Z5 and right BFS-U3-200S6C. The upper row shows the captured image after demosaicking and gamma correction, and the lower row shows the above images denoised by Restormer [Zamir *et al.*, 2022]. (Left) Compared to conventional sensor, the proposed spatially-varying gain recovers much more object details in the dark lighting; (Right) Compared to conventional, spatially-varying binning retains better contrast and recovers sharper contours.

network, so that regions without binning and with binning are denoised separately, and then merge two denoised copies according to the binning map.

Teaser. As shown in Figure 5.1, we capture with an BFS-U3-200S6C with no binning and constant gain, with proposed varying gain, the proposed varying analog additive binning, and denoise all by Restormer [Zamir *et al.*, 2022]. The baseline image is extremely noisy, leaving the alphabets hard to tell and hazing colors in the cropped patch. Adapting gain to local brightness can effectively reduce noise and we can the details clearer. Applying varying binning reduces noise and improves the signal contrast compared to baseline.

Application on vignetting and lens blur. Vignetting and spatially-varying blur is common in photography. We show that the proposed spatially-varying gain and binning can increase the noise performance for vignetting and lens blur. As shown in fig. 5.10(a)(b), we pre-calibrated the vignetting map and spatially-varying lens blur for BFS-U3-200S6C with an 8 mm lens and fixed aperture $f/1.4$. We compute a gain map that is inversely proportional to vignetting and a binning map that bin 2×2 pixels when the PSFs are flat. We capture a scene shown in (e) and correct the intensity by inverting the vignetting map. (f) shows zoom-patches for baseline no binning and constant gain and proposed spatially-varying gain and binning. We can see that with the proposed technique, noise is reduced to a much smaller region

(a) vignetting (b) PSFs (c) gain map (d) binning map (e) captured scene (f) zoom-in patches from (e)

Figure 5.10: **Proposed techniques for vignetting and lens blur.** Images are captured by BFS-U3-200S6C camera with an 8 mm C-mount lens and $f/1.4$ aperture.

Table 5.3: **Tonemapping functions for each camera.**

Camera	Resolution	ROI size	Tonemap	Figures
Nikon Z5	2640 × 3960	512 × 512	Farbman *et al.* [2008]	Fig. 5.7 upper (2-3 row)
Nexus 6P	3024 × 4032	512 × 512	Farbman *et al.* [2008]	Fig. 5.7 lower (2-3 row)
BFS-U3-200S6C	3648 × 5472	256 × 256	$(50 \cdot i)^{1/2.2}$	Fig. 5.8 upper (2nd row)
			$(500 \cdot i)^{1/2.2}$	Fig. 5.8 upper (3rd row)
BFS-U3-16S2M	1080 × 1440	128 × 128	$(50 \cdot i)^{1/2.2}$	Fig. 5.8 lower (2nd row)
			$(500 \cdot i)^{1/2.2}$	Fig. 5.8 lower (3rd row)

towards the edge.

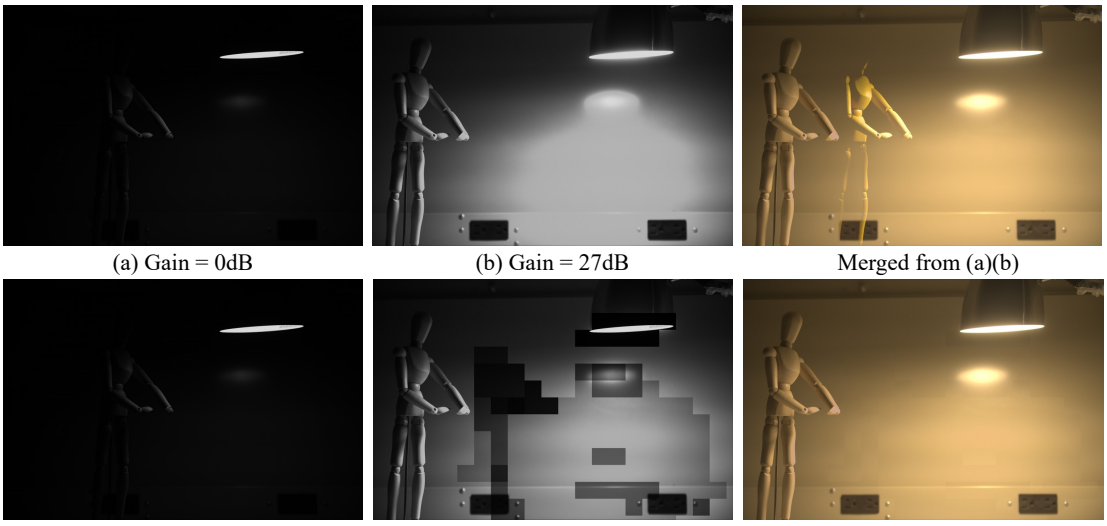
Details about tonemapping used in the chapter As shown in table 5.3, we list the tonemapping functions used in each figure. For Nikon Z5 and Nexus 6P, we use built-in Matlab function `tonemapfarbman` [Farbman *et al.*, 2008] with `RangeCompression=0.2, Saturation=2.5`. For machine vision cameras, we use power functions. In zoomed-in patches, we magnify the intensities of dark patches by 500 times so that the noise and details in the dark regions are more visible, and we only magnify bright patches by 50 to avoid saturation.

Table 5.4: **Quantitative results on simulated HDR scenes [Fairchild, 2008].** For each method and each scene, we show worst-case SSIM (larger is better) and worst-case LPIPS [Zhang *et al.*, 2018a] scores (lower is better).

Scenes	const gain no bin.	**vary. gain** no bin.	const gain **vary. bin**	**vary. gain vary. bin.**
BarHarborSunrise.exr	0.03/ 1.47	0.07/ 1.29	0.14/ 1.14	0.24 / 0.97
BloomingGorse.exr	0.47/ 0.46	0.62/ 0.37	0.47/ 0.46	0.63 / 0.37
GoldenGate.exr	0.04/ 1.39	0.07/ 1.22	0.07/ 1.08	0.09 / 1.00
JesseBrownsCabin.exr	0.04/ 1.40	0.08/ 1.20	0.08/ 1.25	0.09 / 1.12
MirrorLake.exr	0.20/ 0.93	0.48/ 0.56	0.38/ 0.66	0.57 / 0.50
NiagaraFalls.exr	0.51/ 0.60	0.74/ 0.40	0.68/ 0.45	0.80 / 0.41
RedwoodSunset.exr	0.03/ 1.38	0.08/ 1.22	0.23/ 0.81	0.26 / 0.78
TunnelView.exr	0.40/ 0.57	0.61 / 0.38	0.50/ 0.47	0.61 / 0.36
WallDrug.exr	0.03/ 1.35	0.10/ 1.12	0.12/ 1.14	0.32 / 0.82
YosemiteFalls.exr	0.06/ 1.24	0.16/ 1.01	0.10/ 1.04	0.33 / 0.72
Average	0.18/ 1.08	0.30/ 0.88	0.28/ 0.85	0.39 / 0.71

5.7 Quantitative Results

In table 5.4, we compare conventional sensor with the proposed spatially-varying gain, spatially-varying binning, and the combined. Quantitative metrics are computed on simulated images. We download high-quality HDR images from Fairchild [2008] and simulate noisy captured images based on the noise model in eq. (5.1). The simulated camera has a pixel pitch of 0.5 µm, full well capacity of 1000 e^-, pre- and post-amplifier read noise of $\sigma_{pre} = 0.33e^-, \sigma_{post} = 3.31e^-$. We set the black level to be 5% and read out 12-bit images. We normalize the captured image such that the mean intensity equals to 0.05, and gamma correct the normalized images with a power of 1/3.2. We use an ROI size of 128×128. $\text{SNR}_t = 4$ is used to decide the optimal binning. We compare the gamma corrected captured images with the ground-truth ones, and compute LPIPS [Zhang *et al.*, 2018a] and SSIM. We compute the metrics for each ROI and take the worst-case performance. Numbers to the left of slash is SSIM and right is LPIPS. LPIPS is smaller better and SSIM is larger better. We highlight the best SSIM and LPIPS using green and red box for each scene. We see that the proposed spatially-varying gain and binning is significantly better than conventional sensors.



(a) Gain = 0dB | (b) Gain = 27dB | Merged from (a)(b)

(c) Gain = 0dB (used to compute gain map) | (d) Readout with ROI-based spatially-varying gain | Processed from (d)

Figure 5.11: **Comparison between multi-shot technique and ours.** (Upper) Multi-shot method tends to produce ghosting artifacts with the presence of dynamic objects. (Lower) The proposed spatially-varying gain technique utilizes the first capture to determine gain map. The final capture is soley processed from frame (d) by normalizing gains.

5.8 Discussion

We propose two novel readout techniques for image sensors: spatially-varying gain and binning that adapt to the local scene brightness. The proposed techniques significantly improve the noise performance of the captured images for low-light regions, thereby expanding the sensor dynamic range.

Comparison with multi-shot techniques. There is a rich literature that captures high-quality HDR scenes through exposure or gain bracketing [Hasinoff *et al.*, 2010, Pérez-Pellitero *et al.*, 2022]. However, multi-shot techniques are sensitive to motion, and aligning dynamic objects across frames requires significant computation in post-processing. In contrast, our spatially-varying techniques are more robust for dynamic scenes. An example is shown in fig. 5.11. For a fair comparison, we capture two frames for both methods. In the upper row, the multi-shot technique merges low- and high-gain frames through post-processing, and is prone to produce ghosting artifacts. In contrast, our technique (lower row) uses the first frame only to compute the gain map and guide the readout of the subsequent frame. Thereby, ours is free of ghosting.

What does it take to implement in hardware? First, the proposed ROI-based techniques are partially implementable using off-the-shelf CMOS sensors [FLIR, [n.d.]] through windowing, as shown in fig. 5.6. However, existing sensors reset the cycle of integration after each ROI readout, which could lead to motion artifacts for dynamic scenes. To avoid exposing the sensor repeatly during readout, the internal timing should restart integration only after all ROIs are sequentially read out. Second, implementing the proposed per-pixel varying gain requires more engineering efforts as it requires a rapid variable amplifier and additional circuitry to set gain based on previous readout. Previous works [Cui *et al.*, 2017, Lee *et al.*, 2007] demonstrates ultra-wideband programmable variable gain amplifier. They are controled by input signal and can reach a bandwith up to 900MHz, offering a promising solution.

Is analog binning really superior to digital? Prior works emphasize the advantage of analog binning over digital binning. Our analysis shows that this conclusion is arguable with the interplay of gain. When the scene is dark and gain is restricted to a small value, analog binning is indeed better than digital binning, by combining the signal levels to overcome read noise. However, when a larger gain is allowed, either by the proposed spatially-varying gain or other dual ISO techniques, digital binning is all you need to improve the noise performance.

Detailed derivation of optimal pixel pitch We characterize scene contents of various feature sizes by examing sinusoidal signals with varying frequencies. Let us consider a scene of sinusoidal function that varies at frequency f_0 cycles/mm in $x-$direction and constant in y-direction and has the peak intensity of $l_0 e^-/\text{mm}^2$ within the exposure time.

$$i^*(x, y; f_0, l_0) = \frac{l_0}{2}\cos(2\pi f_0 x) + \frac{l_0}{2}$$

Consider a camera with an ideal lens and a sensor with a pixel pitch p µm. The measured signal can be modeled as a convolution between the signal and a box function induced by the pixel size,

$$l(x, y; f_0, l_0, p) = i^*(x, y; f_0, l_0) * b\left\{\frac{x}{p}, \frac{y}{p}\right\} + n_{shot}(x, y) + n_{read}(x, y). \tag{5.6}$$

The first part models the expected noise-free signal measurement by incorporating the blurring effect of pixel pitch. We obtain the expression of the measured signal by taking the Fourier transform of i^* and b, multiply them and taking inverse Fourier transform. The expression for noise-free component is,

$$l^*(x, y; f_0, l_0) = \frac{l_0 p}{2}\frac{\sin(\pi p f_0)}{\pi f_0}\cos(2\pi f_0 x) + \frac{l_0}{2}p^2$$

As $-1 \leq \cos(2\pi f_0 x) \leq 1$, l^* has a maximum and minimum intensity of,

$$\max l^* = \frac{l_0 p}{2}\frac{\sin(\pi p f_0)}{\pi f_0} + \frac{l_0}{2}p^2$$
$$\min l^* = -\frac{l_0 p}{2}\frac{\sin(\pi p f_0)}{\pi f_0} + \frac{l_0}{2}p^2$$

The contrast of the noise-free signal is

$$c(f_0; l_0, p) = |\max l^* - \min l^*| = l_0 p^2 \frac{sin(\pi p f_0)}{\pi f_0}. \tag{5.7}$$

Next, we examine the total noise variance. Note that we approximate the Poisson distributed shot noise by Gaussian distribution $n_{shot}(x_0, y_0) \sim \mathcal{N}(0, l^*(x_0, y_0))$ with a mean zero and variance of latent signal. Interestingly, $n_{shot}(x_0, y_0), n_{shot}(x_0 + T_0, y_0), ..., n_{shot}(x_0 + NT_0, y_0), x_0 \in [0, T_0), N \in \mathbb{Z}$ can be viewed as iid samples of the same distribution $\mathcal{N}(0, l^*(x_0, y_0))$, as the latent signal l^* is a periodic function with period T_0, $l^*(x_0, y_0) = l^*(x_0 + T_0, y_0) = ... = l^*(x_0 + NT_0, y_0), N \in \mathbb{Z}$. Therefore, the *combined* variance of $n_{shot}(x, y_0) \forall x \in [0, T_0)$ can be written as,

$$\begin{aligned}
\sigma^2_{shot} &= \int_{x=0}^{T_0} Var(n_{shot}(x, y_0))\frac{1}{T_0}dx \\
&= \int_{x=0}^{T_0} l^*(x, y_0)\frac{1}{T_0}dx \\
&= \int_{x=0}^{T_0} \left(\frac{l_0 p}{2}\frac{\sin(\pi p f_0)}{\pi f_0}\cos(2\pi f_0 x) + \frac{l_0}{2}p^2\right)\frac{1}{T_0}dx \\
&= \frac{l_0}{2}p^2.
\end{aligned}$$

n_{read} includes both pre- and post-amplifier read noise and follows a Gaussian distribution $n_{read} \sim \mathcal{N}(0, \sigma^2_{pre} + \sigma^2_{post}/g^2)$. Putting the combined variance of shot and read noise, we obtain the total noise variance as the following,

$$\sigma^2(f_0; l_0, p) = \sigma^2_{pre} + \sigma^2_{post}/g^2 + l_0 p^2/2, \tag{5.8}$$

where read noise is independent of pixel size, and shot noise variance increase proportionally to pixel area p^2. Interestingly, the combined noise variance is independent of the underlying signal frequency f_0.

With the noise-free signal contrast $c(\cdot)$ and the combined noise standard deviation $\sigma(\cdot)$, we can find out the cutoff frequency that has an SNR_t for a given pixel pitch p and at light level l_0,

$$f_{\text{cutoff}}(l_0, p) = \arg\min_{f_0} \left\| \frac{c(f_0; l_0, p)}{\sigma(f_0; l_0, p)} - \text{SNR}_t \right\|. \tag{5.9}$$

6 Conclusion and Future Work

The past decade has seen rapid development and prevalence of mobile photography. Nowadays, mobile device cameras have become an indispensible part of our daily life. How should we revolutionize the camera to further push the limit of mobile photography? This thesis advances the camera design for mobile devices from two distinct aspects. First, the introduction of under-display cameras redefine the placement of a camera unit. UDCs doesn't require any dedicated region on the display screen, therefore is free to place anywhere under the screen. Recently, UDCs have been introduced to many state-of-the art smarphone s and used as selfie cameras. It is also appealing to place UDCs in the center of the screen so that the eye gazing is more natural in video conferencing. Apart from smartphones, UDCs can also be placed in tablets, monitors, and laptops, enabling truly full-screen devices, and potentially in large TVs, VR teleconferencing stations, and front-facing cameras in VR headsets in the future. Second, combating noise has been a long-standing problem in computational imaging despite the continuous improvement in CMOS manufacturing capability. We introduce novel scene adaptive techniques to improve the noise performance of the cutting-edge sensors. The proposed techniques are not only applicable to smartphone camera sensors but also conventional DSLR, mirrorless cameras, and research cameras. We hope the proposed sensor designs could inspire more innovations in the space of computational sensors. Lastly, the ideas proposed in this thesis can be naturally extended to new directions that worth looking into in the future. We list some of these ideas in this section.

6.1 Future directions on Under-Display Cameras

6.1.1 Under-Display Cameras + 3D Imaging

UDCs don't require dedicated screen space, thus the display quality is kept the same even with more cameras beneath it. Placing multiple cameras under the screen can provide rich 3D information and can largely benefits applications such as AR/VR teleconferencing. Apart from smartphone and tablets,

under-display cameras are also particularly suitable for large screen devices such as televisions, VR teleconferencing stations, and smart mirrors. As large screen devices are viewed from afar compared to small screens and thus typically have lower pixel resolution. The large distance between display pixels improves the conditioning of the diffractive blur and increase the light throughput, thus largely benefits the quality of under-display cameras. Moreover, large screen sizes allow a much larger baseline for multi-camera systems, thus offer more robust depth estimation.

More interestingly, different display openings can be designed for multiple UDCs such that the point spread functions complement each other, producing a system more robust to inversion. A recent work [Wang *et al.*, 2024] proposes to use a pair of UDCs with one display rotated 45° from the other and show significant improvements in image quality compared to the UDC with one display pattern. Apart from hand-crafted operations such as rotation, the opening patterns of all UDCs can also be solved through optimizing the overall image quality.

6.1.2 Under-Display Cameras + Lensless Imaging

Can we turn the entire region under the OLED display to a large imager? Accommodating an image sensor as large as the entire screen would require innovative camera design as there lacks enough space to place conventional compound lenses that focuses light to the sensor.

A rich literature of research on lensless imaging provides many inspirations. The OLED display can be viewed as the attenuation mask on the lensless camera. Following similar ideas as in chapter 3, we can optimize the OLED display layout such that the quality of the restored image is the maximized. Different from BiDiScreen [Hirsch *et al.*, 2009] that uses a LCD display, this camera can operate with OLED display at the same time as OLED display doesn't require a backlit panel and is partially transparent.

Alternatively, we can place a phase mask tightly against the display panel to form a coded aperture. An intuitive phase mask profile is a lenslet array, where the lens pitch matches the display pixel size. Therefore, the opening of each display pixel acts as the aperture of each lenslet. This effectively forms a light field camera, where each image depicts a different viewpoint. Consider a display of 300 DPI and a sensor pixel pitch of 0.5 μm. Each lenslet forms an image of a resolution of around 168×168, a typical trade-off between angular resolution and spatial resolution. And the challenge of building high quality UDCs is converted from deblurring to super-resolving the low resolutional image.

Similar to the idea presented in chapter 4, we can optimize the height map of a phase mask such that the point spread function is robust to deblurring. In addition to the image quality, we can optimize the phase mask such that the point spread functions are distinct from each other at different depth, offering

depth cues during reconstruction.

6.1.3 Under-Display Camera + Active Illumination

Lots of innovation in this thesis focus on display-camera systems, especially redesigning the shape of display openings to improve camera performance. However, random pixel tiling compromises the display quality such as (chapter 3). This is because randomly flipping and rotating display pixels causes randomness in RGB subpixels and results in color leaking artifacts. It suggests that we should not only consider the shape of openings but also other light emitting units on the display. One promising approach is to jointly optimize display opening shape as well as display subpixel placement, such that both the display quality and the camera image quality is maximized.

Up till now, we view the display as the effective aperture on the camera. An natural extended idea is to use the display as an active light source to illuminate the target scene. Displaying known patterns on the display can encode information in the captured image. For example, use the display as the light source in photometric stereo for 3D reconstruction. We can jointly optimize the image used for display as well as image reconstruction algorithms for the camera to achieve desired performance.

6.1.4 Under-Display Camera + Flare Removal

Flare cause apparent artifacts in under-display cameras. In a high dynamic range scene, a bright light source produces a large diffractive blur pattern around it, occluding the textures in the surrounding dark regions. Removing the diffractive pattern of the light source and recovering the details from its surrounding regions becomes a very challenging task. Feng *et al.* [2021] incorporate saturation to the UDC imaging model and use exposure stack to capture high dynamic range scenes. The proposed restoration has less artifacts around the light source, but still struggles to recover clear details in the dark region immediately around the light source.

Our analysis shows that it is the *extremely low signal-to-noise ratio* that hinders the restoration of low light regions around the light source. Consider a high dynamic range scene spanning around 40 dB. Assume that the flare produced by a bright light source has a signal level of around 1000 photonelectrons and a background region of around 10 photonelectrons. The flare is photon-noise dominated and has noise standard deviation of around 33 photonelectrons, which is much larger than the signal level of the background regions. That is to say, the background regions has an extremely low signal-to-noise ratio and thus the texture is very hard to recover from excessive noise. How to robustly recover the details from the excessive noise with minimal hallucination is an open challenge.

6.2 Minimizing Lens Components and Light Path

Another challenge in designing mobile device camera is that the thickness of mobile devices largely constraints the focal length of the cameras. For example, a focal length equivalent to 200 mm telephoto lens would require a light path of at least x in smartphone cameras, which is much longer than the thickness of mobile devices.

Prior works propose various designs to accommodate the long focal length in thin devices. Periscopic cameras use a 45° mirror the turn the optical axis so that its length is limited by the length of the device instead of the thickness. Origami lens designs the profiles of two mirrors, so that the light bounces multiple times before reaching the sensor, and therefore folds a long focal length to a thin device.

Can we utilize the vacant space inside the entire mobile device as space for light propagation? Folding the light path inside the entire space would significantly increase the effective length of light propagation. One promising solution is to carefully design the profile and placement of mirrors or phase masks, such that they guide light incident from the aperture to the sensor with as many bounces as possible. This could be an interesting directions in mobile photography.

In additional to folding light path, reduce the size of compound lenses is also important when space is limited. Metalenses offer an exciting opportunity to replace traditional compound lenses with a thin phase masks with micron-scale thickness. Future work that focuses on improving image quality, field of view, focal length using novel optics is an appealing direction.

6.3 More on Computational Sensors

6.3.1 Focal Plane Sensor Processors

An emerging type of sensor is focal plane sensor processors [Nguyen *et al.*, 2022, Zarándy, 2011]. They support analog and digital computation at each pixelsite at the sensor plane. These new sensors can thus implement more advanced filtering or even neural networks at the sensor plane and reduce the I/O bandwith for high resolutional imagery. They are especially suitable for task-specific cameras such as edge detection, object tracking, face identification, and others. Image features can be directly extracted at the sensor plane and only low-dimensional outputs are output to the computer. End-to-end sensor plane design with specified tasks is an interesting and promising direction.

6.3.2 Adaptive Imaging

This thesis looks into sensor capabilities that adapts to the content of a scene, specifically spatially-varying gain and binning. Prior works have also explored the adaptive feature of exposure and color filter arrays [Luo *et al.*, 2017, Saragadam *et al.*, 2021, Sarhangnejad *et al.*, 2019]. Sarhangnejad *et al.* [2019] adapt the exposure at local pixels to the scene according to the previous frame, enabling high dynamic range imaging at high framerate. Saragadam *et al.* [2021] use an RGB image to guide the scene-adaptive spatial sampling, thus achieving high quality and video-rate hyperspectral imaging. Many conventional sensor features are yet to be rediscovered to adapt to the scene content to achieve extended capabilities.

Bibliography

[n.d.]. ISO 12233 Photography. https://www.graphics.cornell.edu/~westin/misc/ISO_12233-reschart.pdf.

Hrvoje Benko and Andrew D Wilson. 2009. DepthTouch: Using depthsensing camera to enable free-hand interactions on and above the interactive surface. In *IEEE Workshop on Tabletops and Interactive Surfaces*.

Stephen Boyd, Stephen P Boyd, and Lieven Vandenberghe. 2004. *Convex optimization*. Cambridge university press.

Stephen J Carey, David RW Barr, Bin Wang, Alexey Lopich, and Piotr Dudek. 2013. Mixed signal SIMD processor array vision chip for real-time image processing. *Analog Integrated Circuits and Signal Processing* 77 (2013), 385–399.

Julie Chang and Gordon Wetzstein. 2019. Deep optics for monocular depth estimation and 3d object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.

Chi Jui Cheng, Tzu Chin Huang, Wen Tsan Lin, Cheng Chih Hsieh, Peng Yu Chen, Peter Lu, and Hoang Yan Lin. 2019. Evaluation of Diffraction Induced Background Image Quality Degradation through Transparent OLED Display. In *SID Symp. Digest of Technical Papers*.

Roger N. Clark. 2016. Digital Camera Reviews and Sensor Performance Summary. https://clarkvision.com/imagedetail/digital.sensor.performance.summary/.

Shuang Cui, Tianzhao Liu, Haoran Gong, Bingyan Hu, and Yuchun Chang. 2017. A high performance switched-capacitor programmable gain amplifier design in 0.18 μm CMOS technology. In *2017 IEEE 12th International Conference on ASIC (ASICON)*. IEEE, 758–761.

Kostadin Dabov, Alessandro Foi, Vladimir Katkovnik, and Karen Egiazarian. 2007. Image denoising by sparse 3-D transform-domain collaborative filtering. *IEEE Trans. image processing* 16, 8 (2007), 2080–2095.

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

Neil Emerton, David Ren, and Tim Large. 2020. Image Capture Through TFT Arrays. In *SID Symp. Digest of Technical Papers*.

Mark Fairchild. 2008. The HDR Photographic Survey. `http://markfairchild.org/HDR.html`.

Zeev Farbman, Raanan Fattal, Dani Lischinski, and Richard Szeliski. 2008. Edge-preserving decompositions for multi-scale tone and detail manipulation. *ACM Transactions on Graphics* 27, 3 (2008), 1–10.

Ruicheng Feng, Chongyi Li, Huaijin Chen, Shuai Li, Chen Change Loy, and Jinwei Gu. 2021. Removing diffraction image artifacts in under-display camera via dynamic skip connection network. In *CVPR*.

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

FLIR. [n.d.]. Blackfly USB3 cameras. `https://www.flir.com/products/blackfly-s-usb3/`.

Alessandro Foi, Mejdi Trimeche, Vladimir Katkovnik, and Karen Egiazarian. 2008. Practical Poissonian-Gaussian noise modeling and fitting for single-image raw-data. *IEEE Transactions on Image Processing* 17, 10 (2008), 1737–1754.

KeMing Gao, Meng Chang, Kunjun Jiang, Yaxu Wang, Zhihai Xu, Huajun Feng, Qi Li, Zengxin Hu, and YueTing Chen. 2021. Image restoration for real-world under-display imaging. *Optics Express* 29, 23 (2021), 37820–37834.

Joseph W Goodman. 2005. *Introduction to Fourier optics*. Roberts & Co.

Rahul Gulve, Roberto Rangel, Ayandev Barman, Don Nguyen, Mian Wei, Motasem A Sakr, Xiaonong Sun, David B Lindell, Kiriakos N Kutulakos, and Roman Genov. 2023. Dual-Port CMOS Image Sensor with Regression-Based HDR Flux-to-Digital Conversion and 80ns Rapid-Update Pixel-Wise Exposure Coding. In *IEEE International Solid-State Circuits Conference)*. 104–106.

Saghi Hajsharif, Joel Kronander, and Jonas Unger. 2014. HDR reconstruction for alternating gain (ISO) sensor readout. In *Eurographics*.

Samuel W Hasinoff, Frédo Durand, and William T Freeman. 2010. Noise-optimal capture for high dynamic range photography. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*.

Samuel W. Hasinoff, Dillon Sharlet, Ryan Geiss, Andrew Adams, Jonathan T. Barron, Florian Kainz, Jiawen Chen, and Marc Levoy. 2016. Burst photography for high dynamic range and low-light imaging on mobile cameras. *ACM Trans on Graphics* 6 (2016).

Glenn E Healey and Raghava Kondepudy. 1994. Radiometric CCD camera calibration and noise estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 16, 3 (1994), 267–276.

Felix Heide, Qiang Fu, Yifan Peng, and Wolfgang Heidrich. 2016. Encoded diffractive optics for full-spectrum computational imaging. *Scientific reports* 6, 1 (2016), 1–10.

Matthew Hirsch, Douglas Lanman, Henry Holtzman, and Ramesh Raskar. 2009. BiDi screen: A thin, depth-sensing LCD for 3D interaction using light fields. *ACM Trans. Graphics* 28, 5 (2009), 1–9.

Jie Hu, Sankhyabrata Bandyopadhyay, Yu-hui Liu, and Li-yang Shao. 2021. A review on metasurface: from principle to smart metadevices. *Frontiers in Physics* 8 (2021), 586087.

Daniel S Jeon, Seung-Hwan Baek, Shinyoung Yi, Qiang Fu, Xiong Dun, Wolfgang Heidrich, and Min H Kim. 2019. Compact snapshot hyperspectral imaging with diffracted rotation. *ACM Transactions on Graphics* (2019).

Xiaodan Jin and Keigo Hirakawa. 2012. Analysis and processing of pixel binning for color image sensor. *EURASIP ASP* 2012 (2012), 1–15.

Huifeng Ke, Navid Sarhangnejad, Rahul Gulve, Zhengfan Xia, Nikita Gusev, Nikola Katic, Kiriakos N Kutulakos, and Roman Genov. 2019. Extending image sensor dynamic range by scene-aware pixelwise-adaptive coded exposure. In *Proc. Int. Image Sensor Workshop*. 111–114.

Jaihyun Koh, Jangho Lee, and Sungroh Yoon. 2022. BNUDC: A two-branched deep neural network for restoring images from under-display cameras. In *CVPR*.

Kinam Kwon, Eunhee Kang, Sangwon Lee, Su-Jin Lee, Hyong-Euk Lee, ByungIn Yoo, and Jae-Joon Han. 2021. Controllable Image Restoration for Under-Display Camera in Smartphones. In *CVPR*.

Hui Dong Lee, Kyung Ai Lee, and Songcheol Hong. 2007. A wideband CMOS variable gain amplifier with an exponential gain control. *IEEE Transactions on Microwave Theory and Techniques* 55, 6 (2007), 1363–1373.

Anat Levin, Rob Fergus, Frédo Durand, and William T Freeman. 2007. Image and depth from a conventional camera with a coded aperture. *ACM Trans. Graphics* 26, 3 (2007), 70–es.

Sida Li, Yueda Liu, Yan Li, Shuxin Liu, Shuyi Chen, and Yikai Su. 2019. Fast-response Pancharatnam-Berry phase optical elements based on polymer-stabilized liquid crystal. *Optics Express* 27, 16 (2019), 22522–22531.

Sehoon Lim, Yuqian Zhou, Neil Emerton, Tim Large, and Steven Bathiche. 2020. Image Restoration for Display-Integrated Camera. In *SID Symp. Digest of Technical Papers*.

Yi Luo, Derek Ho, and Shahriar Mirabbasi. 2017. Exposure-programmable CMOS pixel with selective charge storage and code memory for computational imaging. *IEEE Transactions on Circuits and Systems I: Regular Papers* 65, 5 (2017), 1555–1566.

Emil Martinec. 2008. Noise, Dynamic Range and Bit Depth in Digital SLRs. `https://homes.psd.uchicago.edu/~ejmartin/pix/20d/tests/noise/`.

Belen Masia, Gordon Wetzstein, Piotr Didyk, and Diego Gutierrez. 2013. A survey on computational displays: Pushing the boundaries of optics, computation, and perception. *Computers & Graphics* 37, 8 (2013), 1012–1038.

Christopher A Metzler, Hayato Ikoma, Yifan Peng, and Gordon Wetzstein. 2020. Deep optics for single-shot high-dynamic-range imaging. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Kaushik Mitra, Oliver S Cossairt, and Ashok Veeraraghavan. 2014. A framework for analysis of computational imaging systems: role of signal prior, sensor noise and multiplexing. *IEEE transactions on pattern analysis and machine intelligence* 36, 10 (2014), 1909–1921.

Nanoscribe. 2007. Nanoscribe. `https://www.nanoscribe.com/en/`.

S.G. Narasimhan and S.K. Nayar. 2005. Enhancing resolution along multiple imaging dimensions using assorted pixels. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 27, 4 (2005), 518–530.

Shree K Nayar, Vlad Branzoi, and Terrance E Boult. 2004. Programmable imaging using a digital micromirror array. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Cindy M Nguyen, Julien NP Martel, and Gordon Wetzstein. 2022. Learning spatially varying pixel exposures for motion deblurring. In *IEEE International Conference on Computational Photography (ICCP)*. IEEE, 1–11.

Shichao Nie, Chengconghui Ma, Dafan Chen, Shuting Yin, Haoran Wang, LiCheng Jiao, and Fang Liu. 2020. A Dual Residual Network with Channel Attention for Image Restoration. In *ECCV*.

Youngjin Oh, Gu Yong Park, Haesoo Chung, Sunwoo Cho, and Nam Ik Cho. 2021. Residual Dilated U-Net with Spatially Adaptive Normalization for the Restoration of Under Display Camera Images. In *2021 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*.

Sri Rama Prasanna Pavani and Rafael Piestun. 2008. Three dimensional tracking of fluorescent microparticles using a photon-limited double-helix response system. *Optics express* 16, 26 (2008), 22048–22057.

Yifan Peng, Qiang Fu, Felix Heide, and Wolfgang Heidrich. 2016. The diffractive achromat full spectrum computational imaging with diffractive optics. In *ACM Transactions on Graphics*. 1–2.

Yifan Peng, Qilin Sun, Xiong Dun, Gordon Wetzstein, Wolfgang Heidrich, and Felix Heide. 2019. Learned large field-of-view imaging with thin-plate optics. *ACM Transactions on Graphics* 38, 6 (2019), 219–1.

Eduardo Pérez-Pellitero, Sibi Catley-Chandar, Richard Shaw, Aleš Leonardis, Radu Timofte, Zexin Zhang, Cen Liu, Yunbo Peng, Yue Lin, Gaocheng Yu, *et al.* 2022. NTIRE 2022 challenge on high dynamic range imaging: Methods and results. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*.

Densen Puthussery, Melvin Kuriakose, Jiji C V, *et al.* 2020. Transform Domain Pyramidal Dilated Convolution Networks For Restoration of Under Display Camera Images. *arXiv preprint arXiv:2009.09393* (2020).

Xiangyu Qu, Yiheng Chi, and Stanley H Chan. 2024. Spatially Varying Exposure With 2-By-2 Multiplexing: Optimality and Universality. *IEEE Transactions on Computational Imaging* (2024).

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-Net: Convolutional networks for biomedical image segmentation. In *Intl. Conf. Medical image computing and computer-assisted intervention*.

Vishwanath Saragadam, Michael DeZeeuw, Richard G Baraniuk, Ashok Veeraraghavan, and Aswin C Sankaranarayanan. 2021. SASSI—super-pixelated adaptive spatio-spectral imaging. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 43, 7 (2021), 2233–2244.

Navid Sarhangnejad, Nikola Katic, Zhengfan Xia, Mian Wei, Nikita Gusev, Gairik Dutta, Rahul Gulve, Harel Haim, Manuel Moreno Garcia, David Stoppa, Kiriakos N. Kutulakos, and Roman Genov.

2019. Dual-tap pipelined-code-memory coded-exposure-pixel CMOS image sensor for multi-exposure single-frame computational imaging. In *IEEE International Solid-State Circuits Conference - (ISSCC)*. https://doi.org/10.1109/ISSCC.2019.8662326

Zheng Shi, Yuval Bahat, Seung-Hwan Baek, Qiang Fu, Hadi Amata, Xiao Li, Praneeth Chakravarthula, Wolfgang Heidrich, and Felix Heide. 2022. Seeing through obstructions with diffractive cloaking. *ACM Transactions on Graphics* 41, 4 (2022), 1–15.

Abhinav Shrivastava, Abhinav Gupta, and Ross Girshick. 2016. Training region-based object detectors with online hard example mining. In *CVPR*.

Vincent Sitzmann, Steven Diamond, Yifan Peng, Xiong Dun, Stephen Boyd, Wolfgang Heidrich, Felix Heide, and Gordon Wetzstein. 2018. End-to-end optimization of optics and image processing for achromatic extended depth of field and super-resolution imaging. *ACM Transactions on Graphics* 37, 4 (2018), 1–13.

Sony. 2018. Sony IMX 490 sensor. https://thinklucid.com/tech-briefs/sony-imx490-hdr-sensor-and-flicker-mitigation/.

Qilin Sun, Ethan Tseng, Qiang Fu, Wolfgang Heidrich, and Felix Heide. 2020. 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*.

Varun Sundar, Sumanth Hegde, Divya Kothandaraman, and Kaushik Mitra. 2020. Deep Atrous Guided Filter for Image Restoration in Under Display Cameras. *arXiv preprint arXiv:2008.06229* (2020).

Tali Treibitz and Yoav Y Schechner. 2012. Resolution loss without imaging blur. *JOSA A* 29, 8 (2012), 1516–1528.

Takatoshi Tsujimura. 2017. *OLED display fundamentals and applications*. John Wiley & Sons.

Ashok Veeraraghavan, Ramesh Raskar, Amit Agrawal, Ankit Mohan, and Jack Tumblin. 2007. Dappled photography: Mask enhanced cameras for heterodyned light fields and coded aperture refocusing. *ACM Trans. Graphics* 26, 3 (2007), 69.

Chengyu Wang, Jing Li, Pavan C Madhusudanarao, Jinhan Hu, Jitesh K Singh, Woojhon Choi, Seok-Jun Lee, and Hamid R Sheikh. 2024. UDAC: Under-Display Array Cameras. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 1077–1084.

Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing* 13, 4 (2004), 600–612.

Zhibin Wang, Yilu Chang, Qi Wang, Yingjie Zhang, Jacky Qiu, and Michael Helander. 2020. Self-Assembled Cathode Patterning in AMOLED for Under-Display Camera. In *SID Symposium Digest of Technical Papers*.

Andrew D Wilson. 2004. TouchLight: An imaging touch screen and display for gesture-based interaction. In *International Conference on Multimodal interfaces*.

Yicheng Wu, Vivek Boominathan, Huaijin Chen, Aswin Sankaranarayanan, and Ashok Veeraraghavan. 2019. PhaseCam3D—learning phase masks for passive single view depth estimation. In *ICCP*.

Anqi Yang, Eunhee Kang, Hyong-Euk Lee, and Aswin C. Sankaranarayanan. 2023. Design Phase Masks for Under-Display Cameras: Software. `https://github.com/Image-Science-Lab-cmu/UDC_Phase_ICCV23`.

Anqi Yang and Aswin C. Sankaranarayanan. 2021a. Design Display Pixel Layouts for Under-Panel Cameras: Software. `https://github.com/Image-Science-Lab-cmu/UPC_ICCP21_Code`.

Anqi Yang and Aswin C Sankaranarayanan. 2021b. Designing Display Pixel Layouts for Under-Panel Cameras. *IEEE TPAMI* 43, 7 (2021), 2245–2256.

Qirui Yang, Yihao Liu, Jigang Tang, and Tao Ku. 2020. Residual and Dense UNet for Under-Display Camera Restoration. In *ECCV*.

Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. 2022. Restormer: Efficient transformer for high-resolution image restoration. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 5728–5739.

Ákos Zarándy. 2011. *Focal-plane sensor-processor chips*. Springer Science & Business Media.

Jiachao Zhang, Jie Jia, Andong Sheng, and Keigo Hirakawa. 2018b. Pixel binning for high dynamic range color image sensor using square sampling lattice. *IEEE Transactions on Image Processing* 27, 5 (2018), 2229–2241.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. 2018a. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*.

Yulun Zhang, Yapeng Tian, Yu Kong, Bineng Zhong, and Yun Fu. 2018c. Residual dense network for image super-resolution. In *CVPR*.

Zhenhua Zhang. 2020. Image Deblurring of Camera Under Display by Deep Learning. In *SID Symp. Digest of Technical Papers*.

Changyin Zhou, Stephen Lin, and Shree Nayar. 2009. Coded aperture pairs for depth from defocus. In *ICCV*.

Changyin Zhou and Shree Nayar. 2009. What are good apertures for defocus deblurring?. In *ICCP*.

Yuqian Zhou, Michael Kwan, Kyle Tolentino, Neil Emerton, Sehoon Lim, Tim Large, Lijiang Fu, Zhihong Pan, Baopu Li, Qirui Yang, *et al.* 2020a. UDC 2020 challenge on image restoration of under-display camera: Methods and results. In *ECCV*.

Yuqian Zhou, David Ren, Neil Emerton, Sehoon Lim, and Timothy Large. 2020b. Image restoration for under-display camera. *arXiv preprint arXiv:2003.04857* (2020).

Yuqian Zhou, David Ren, Neil Emerton, Sehoon Lim, and Timothy Large. 2021. Image restoration for under-display camera. In *CVPR*.