Learned Large Field-of-View Imaging With Thin-Plate Optics

YIFAN PENG*, Stanford University
QILIN SUN*, King Abdullah University of Science and Technology
XIONG DUN*, King Abdullah University of Science and Technology
GORDON WETZSTEIN, Stanford University
WOLFGANG HEIDRICH, King Abdullah University of Science and Technology
FELIX HEIDE, Princeton University

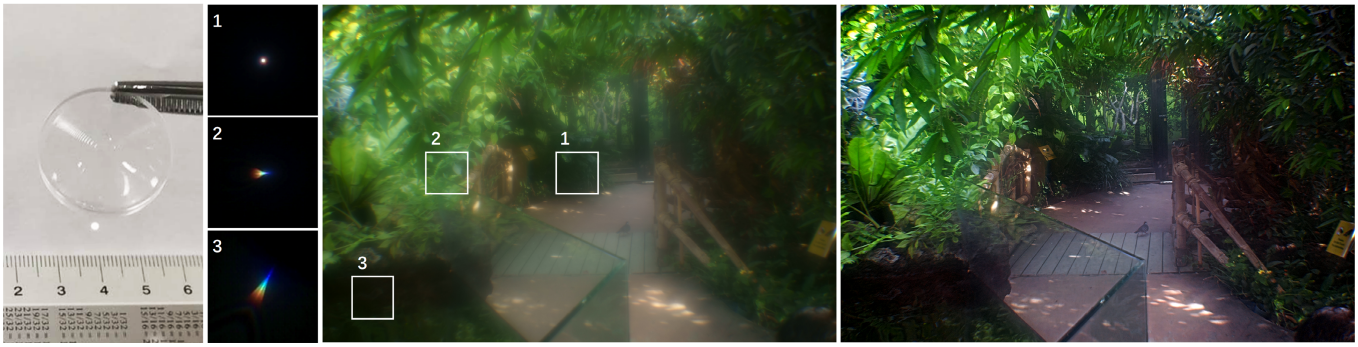

Fig. 1. Large Field-of-View Imaging With Thin-Plate Optics. We design a lens with compact form factor using one (or two) optimized refractive surfaces on a thin substrate (left). This optimization results in a dual-mixture point spread function (center-left insets), which is nearly invariant to the incident angle, exhibiting a high-intensity peak and a large, almost constant, tail. We show the sensor measurement (center) and image reconstruction (right) in natural lighting conditions, which demonstrate that the proposed deep image recovery effectively removes aberrations and haze resulting from the proposed thin-plate optics. Our prototype single element lens achieves a large field-of-view of 53° with a clear aperture of $f/1.8$ and effective aperture of $f/5.4$, see text.

Typical camera optics consist of a system of individual elements that are designed to compensate for the aberrations of a single lens. Recent computational cameras shift some of this correction task from the optics to post-capture processing, reducing the imaging optics to only a few optical elements. However, these systems only achieve reasonable image quality by limiting the field of view (FOV) to a few degrees – effectively ignoring severe off-axis aberrations with blur sizes of multiple hundred pixels.

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

CCS Concepts: • **Computing methodologies** → **Computational Imaging**; **Deep Learning**.

Additional Key Words and Phrases: thin optics, computational camera, image deblurring, deep network

ACM Reference Format:
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 Trans. Graph.* 38, 6, Article 219 (November 2019), 14 pages. https://doi.org/10.1145/3355089.3356526

1 INTRODUCTION

Modern imaging techniques have equipped us with powerful capabilities to record and interact with the world – be that in our personal devices, assistive robotics, or self-driving vehicles. Coupled with recent image processing algorithms, today's cameras are able to tackle high-dynamic range and low-light scenarios [Chen et al. 2018; Hasinoff et al. 2016]. However, while image processing algorithms have been evolving rapidly over the last decades, commercial optical systems are largely still designed following aberration theory, i.e. with the design goal of reducing deviations from Gauss's linear model of optics [Gauss 1843]. Following this approach, commercial lens systems introduce increasingly complex stacks of lens elements

*joint first authors.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
© 2019 Copyright held by the owner/author(s). Publication rights licensed to ACM.
0730-0301/2019/11-ART219 $15.00
https://doi.org/10.1145/3355089.3356526

to combat individual aberrations [Kingslake and Johnson 2009]. For example, the optical stack of the iPhone X contains more than six aspherical elements, and professional zoom optics can contain more than twenty individual elements.

Although modern lens systems are effective in minimizing optical aberrations, the depth of the lens stack is a limiting factor in miniaturizing these systems and manufacturing high-quality lenses at low cost. Moreover, using multiple optical components introduces secondary issues, such as lens flare and complicated optical stabilization, e.g. in a smartphone where the whole lens barrel is actuated. In particular, the design goals of large field-of-view (FOV, e.g. $> 50^\circ$), high numerical aperture (NA), and high resolution (e.g. 4k resolution) stand in stark contrast to a compact, simple lens system. Existing approaches address this challenge using assemblies of multiple different lenses or sensors [Brady et al. 2012; Light.co 2018; MobilEye 2018; Venkataraman et al. 2013; Yuan et al. 2017], including widely deployed dual-camera smartphones, each typically optimized for a different FOV. While providing some reduction in footprint, such spatial multiplexing increases the number of optical elements even further and requires higher bandwidth, power and challenging parallax compensation post-capture [Venkataraman et al. 2013].

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

The learned reconstruction model allows us to recover high-quality images from measurements degraded by severe aberrations. Single lens elements, such as spherical lenses or Fresnel phase plates [Peng et al. 2015], typically suffer from severe off-axis aberrations that restricts the usable FOV to around 10° [Heide et al. 2016, 2013; Peng et al. 2015]. Instead, we propose a novel lens design that offers spatially invariant PSFs, over the full FOV, which are designed to allow for aberration removal by the proposed learned reconstruction model. We achieve this by abandoning the design goal of minimal spot size, and instead balance the local contrast over the full FOV. This alternative objective allows us to build on existing optimization tools for the optics of the proposed co-design, without requiring end-to-end design. The resulting thin lens allows the reconstruction network to detect some contrast across the full FOV, invariant of the angular position, at the cost of reducing contrast in the on-axis region. As a consequence, the proposed computational optics offers an *order of magnitude larger FOV* than traditional single lenses, even with the same reconstruction network fine-tuned to such alternative designs.

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

- We propose a single free-form lens design tailored to learned image reconstruction methods for large FOV high-quality imaging. This design exhibits almost invariant aberrations across the full FOV that balance the contrast detection probability (CDP) of early network layers.
- We propose a generative adversarial model for high-resolution deconvolution for our aberrations of size ≤ 900 pixels.
- The model is trained on data acquired with a display lab setup in an automated manner, instead of painful manual acquisition in the field. We provide all models, training and validation data sets.
- We realize the optical design with two prototype lenses with effective thickness of 120 μm, aperture size of 23.4 mm, and a FOV of 53° – one with a single optical surface, the other with two optical surfaces (both sides of the same flat carrier). We experimentally validate that our approach offers high image quality for a wide range of indoor and outdoor scenes.

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

2 RELATED WORK

Optical Aberrations and Traditional Lens Design. Both monochromatic and chromatic aberrations are results of the differences of the optical path length when light travels trough different regions of a lens at different incident angles [Fowles 2012]. These aberrations manifest themselves as unwanted blur which becomes more severe with increasing numerical aperture and field-of-view [Smith 2005]. Conventional lens design aims at minimizing aberrations of all kinds by increasingly complex lens stacks [Sliusarev 1984]. This includes designing aspherical surfaces and introducing lens elements using materials with different optical properties.

State-of-the-art optical design software is a cornerstone tool for optimizing the surface profiles of refractive lens designs. However, while hyper-parameter optimization tools are becoming mature, the design process still relies on existing objectives, so-called *merit functions*, that find a compromise across a variety of criteria [Malacara-Hernández and Malacara-Hernández 2016; Shih et al. 2012], trading off the point spread function (PSF) shape across sensor locations, lens configurations (e.g. zoom levels) and target wavelength band.

Computational Optics. A large body of work on computational imaging [Dowski and Cathey 1995; Levin et al. 2009; Stork and Gill 2013, 2014] has proposed to design optics for aberration removal in post-processing. These methods often favor diffractive optical elements (DOEs) over refractive optics [Antipa et al. 2018; Heide et al. 2016; Monjur et al. 2015; Peng et al. 2016] because of their large design space. Moreover, recent work proposed caustic (holographic) designs, for projection displays or imaging lenses [Papas et al. 2012; Peng et al. 2017; Schwartzburg et al. 2014]. To simplify the inverse problem in post-processing, all of the described approaches ignore off-axis aberrations by restricting the FOV to a few degrees – existing approaches do not realize monocular imaging with a large FOV.

Several approaches to end-to-end optical imaging were recently proposed, where parametrized optics and image processing are jointly optimized for applications in extended depth of field and superresolution imaging [Sitzmann et al. 2018], monocular depth

estimation [Chang and Wetzstein 2019; Haim et al. 2018; Wu et al. 2019], and image classification [Chang et al. 2018b]. However, none of these approaches aim at large FOV imaging and all of them build on simple paraxial image formation models, which break for large fields of view. Moreover, they are limited to a single optical surface. We overcome these challenges by engineering PSFs over a large FOV, and, relying on existing optical design tools that support complex multi-surface/material designs, optimize for a well-motivated dual-mixture design tailored to deep reconstruction models.

Manufacturing Planar Optics. Various manufacturing methods exist that enable "planar" optics with low-depth optical surface, i.e. less than 1 mm. Commercial miniature form factor optics like the lenses in smartphone cameras, can be manufactured using mature injection molding techniques [Oliver et al. 2010]. Alternative fabrication methods for thin-plate lenses include diffractive optics and metalenses [Duoshu et al. 2011; Genevet et al. 2017], which require nano-fabrication methods like photolithography and nano-imprinting [Ahn and Guo 2009; Chou et al. 1996]. The UV-cure replication technique [Zoberbier et al. 2009] can facilitate manufacturing wafer-scale optical elements. Note that creating a Fresnel lens with a clear aperture diameter of 23.5 mm and a focal length of 43 mm requires, as in this work, a feature size smaller than 300 nm, which is beyond the capability of the photolithography methods used in many recent DOE works [Heide et al. 2016; Peng et al. 2016; Sitzmann et al. 2018]. Freeform lenses with a larger aperture and continuous surfaces can be manufactured using diamond turning machining [Fang et al. 2013]. The continuous surface preserves light efficiency and works under broadband illumination, while the lenses are usually thick and bulky because of the local curvature constraints.

In this work, we use high-precision diamond turning machining to prototype the proposed lenses. Instead of fabricating a freeform lens with continuous surface, e.g., as in [Sitzmann et al. 2018], we wrap the optimized surface profile using coarse wrap-around depth values instead of wavelength-scale wrapping in diffractive lens designs, see Fig. 1. This allows us to design a Fresnel-inspired free-form lens with the advantages of both refractive optics and diffractive optics: we achieve a thin form factor while reducing chromatic aberrations.

Image Quality. Imaging describes the signal chain of light being transported from a scene patch of interest to the camera, focusing in the camera optics, digitization of the focused photon flux on the sensor, and post-processing of the measured data. During each of these individual steps, information about the scene patches of interest may be lost or corrupted. Various hand-crafted image quality metrics exist that measure the cumulative error of this imaging process [Mitra et al. 2014; Wang et al. 2004], with or without known ground-truth reference [Mittal et al. 2012], or allow to individually characterize components of the imaging stack using calibration setups [EMVA Standard 2005; Estribeau and Magnan 2004]. Typical performance metrics are the signal-to-noise ratio (SNR) [Parker 2010] and modulation transfer function (MTF) [Boreman 2001; Estribeau and Magnan 2004]. While these metrics are widely reported and measurement setups are readily available, they are also not free from disadvantages due to their domain-agnostic design. For example, high SNR does not guarantee a perceptually pleasing image, which has sparked recent work on perceptual loss functions [Johnson et al. 2016]. Moreover, SNR increases in the presence of glare and quantization, which can yield inconclusive results when used as a design metric [Geese et al. 2018].

We design the proposed optical system in conjunction with the learned image reconstruction methods. To this end, we analyze the behavior of the early layers in our generator, which relate to the response of local contrast features in the scene. Relying on a probabilistic measure [Geese et al. 2018], we assess the ability to detect or miss such local features across the full FOV. This insight allows us to tailor the proposed lens design to our network-based reconstruction method.

Learned Image Reconstruction. Traditional deconvolution methods [Cho et al. 2012; Heide et al. 2013; Krishnan and Fergus 2009] using natural image priors are not robust when working with extremely large, spatially invariant blur kernels that exhibit chromatic aberrations and other challenging effects. Unfortunately, the lens design proposed in this work produces large PSFs that present a challenge to existing deconvolution methods which suffer in image quality for large aberrations, necessitating a custom image reconstruction approach. Note that computationally efficient forward models for large spatially-varying convolutions have been investigated before [Gilad and Von Hardenberg 2006]. Over the last years, a large body of work proposed data-driven approaches for image processing tasks [Schuler et al. 2013; Xu et al. 2014; Zhang et al. 2017]. Specifically addressing deconvolution, Nah et al. [2017] propose a fully connected convolutional network that iteratively deconvolves in a multi-stage approach. More recently, generative adversarial networks (GANs) have been shown to provide generative estimates with high image quality. Kupyn *et al.* [2017] demonstrate the practicability of applying GAN reconstruction methods to deblurring problems.

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

3 DESIGNING OPTICS FOR LEARNED RECOVERY

In this work, we describe an optical design tailored to learned reconstruction techniques for large field-of-view, thin-plate photography. The proposed optical system is shown in Figure 2. In contrast to state-of-the-art compound lenses, it consists of a single, almost flat, element. The two core ideas behind the proposed optical design are the following: first, to achieve a large FOV, we constrain the PSFs of our lens to be shift-invariant for the incident angle. Second, although such PSFs exhibit large spot sizes, we engineer aberrations that preserve residual contrast and hence are well-suited for learned image reconstruction.

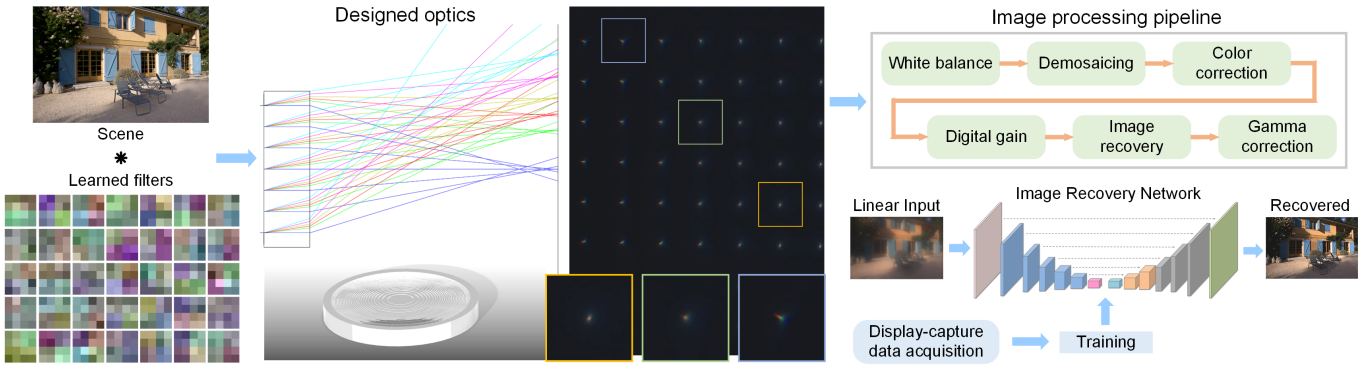

Fig. 2. Computational thin-plate lens imaging with large field-of-view. Left: learned early layers' filters applied on input scene; Center-left: optical design that preserves local contrast across full FOV. The designed optical element has a Fresnel lens surface; Center-right: calibrated PSF patches for different incident angles of our prototype lens (images gamma-tonemapped for visualization); Right: overview of the image processing pipeline and our recovery framework, which learns a mapping for the linear input to the recovered output. We introduce a learned reconstruction architecture trained using data that can be efficiently acquired in a display-capture lab setup (see details in Section 6).

Our design is motivated by a large body of work on computational photography with PSF engineering – designing PSFs invariant to a target characteristic, instead of minimizing spot size, and computation to remove the non-compact aberrations. Similar to how existing work extends the depth of field [Cossairt and Nayar 2010] or spectral range [Peng et al. 2016; Wang et al. 2016], we are the first to apply this idea to extending the field of view.

To this end, we rely on the insight that the filters of early layers of recent deep models, across applications in computer vision and imaging, have striking similarity – these early layers are gradient-like filters and respond to local contrast as essential low-level information content in the measurement. As many recent learned architectures rely on common low-level backbones, which are then transferred to different higher-level tasks [Bengio 2012; Huh et al. 2016], this transfer-learning offers an interesting opportunity for the design of imaging systems. We engineer the PSFs of the proposed optical design, shown in Figure 2, to exhibit a *peaky* distribution. While the peak contribution maximizes the probability of detecting local contrast features, the low-frequency part is extremely large (~900 pixels on the experimental sensor system covered below) and therefore leads to very low filter responses in the early layers. In contrast to conventional spherical elements, see Figure 7, this PSF exhibits the peak-preserving distribution across the full sensor which enables large FOV imaging with this single optical element.

Given a raw measurement acquired with the proposed thin-plate lens system, we recover a high-quality image using a generative adversarial network which is trained to eliminate all measurement degradations and directly outputs a deblurred, denoised, and color-corrected image, see Figure 2. To train the network in a semi-supervised fashion, using labeled and unlabeled data to learn robust loss functions along with the model parameters, we require a training dataset with ideal reference images and corresponding blurry captures. Instead of manually acquiring such a dataset, e.g., by sequentially swapping optics for a scene, we propose an automated lab setup which displays known ground-truth images on a display.

In the following, we first describe the proposed optical design in Section 4, before introducing the reconstruction architecture and training methodology in Section 5.

4 LENS DESIGN

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

4.1 Ideal Phase Profile

The phase of a lens describes the delay of the incident wave phase introduced by the lens element, at the lens plane. The geometrical (ray) optics model, commonly used in computer graphics, models light as rays of photon travel instead of waves. This model ignores diffraction, e.g. for light passing through a narrow slit. Although being an approximation to physical optics, ray optics still can provide an intuition: the perpendiculars to the waves can be thought of as rays, and, vice versa, phase intuitively describes the relative delay of photons traveling along these rays to the lens plane, as illustrated with red lines in Fig. 3. Hence, the phase of a thin lens is its height profile multiplied with the wave number and the refractive index [Goodman 2005; Hecht 1998].

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

To specify the ideal phase profile $\phi(r, \omega_i)$ for an incident ray direction i, and radial position r, see Figure 3, we assume a physical

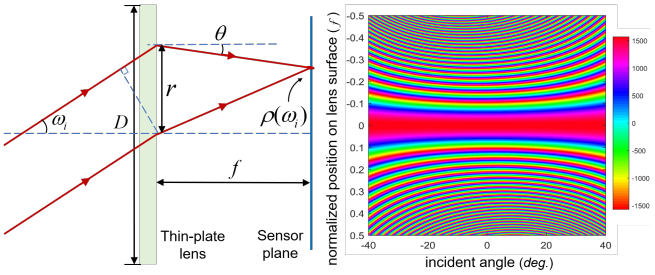

Fig. 3. Schematic of ray geometries in a radially symmetric design manner (left), and the ideal phase profile distribution subject to incident angle (right). For visualization purpose the phase map is wrapped by $1{,}000\pi$, and the vertical axis is normalized with respect to the focal length.

aperture size D, focus distance f, and set:

$$\phi(r, \omega_i) = -k \left[r \cdot \sin\omega_i - \int_0^r \sin\theta(r_1, \omega_i) dr_1 \right], \tag{1}$$

where k represents here the wave number that is specified by the wavelength, and ω_i represents the incident angle of ray direction i [Kalvach and Szabó 2016]. For this ideal lens profile, we define the output angle as:

$$\theta(r, \omega_i) = \arctan\left(\frac{\rho(\omega_i) - r}{f}\right), \tag{2}$$

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

Next, by inserting Eq. 2 into Eq. 1, we derive the target phase ϕ as:

$$\begin{aligned} \phi(r, \omega_i) &= -k \left[r \cdot \sin\omega_i - \int_0^r \frac{\rho(\omega_i) - r_1}{\sqrt{f^2 + (\rho(\omega_i) - r_1)^2}} dr_1 \right] \\ &= -k \left[r \cdot \sin\omega_i + \sqrt{f^2 + (\rho(\omega_i) - r)^2} - \sqrt{f^2 + \rho(\omega_i)^2} \right]. \end{aligned} \tag{3}$$

The ideal phase profile from Eq. 3 is visualized in Figure 3 (right). We observe a drastic variation when approaching larger incident angles. In other words, the same position on the lens aperture would need to realize different phases for different incident angles, which is not physically realizable with thin plate optics.

4.2 Aperture Partitioning

Realizing the ideal phase profile is intractable to manufacture over the full aperture, as illustrated by the large angular deviations needed in off-axis region in Figure 3. To overcome this challenge, we split the aperture into multiple sub-regions, and assign each sub-region to a different angular interval, similar to prior work [Levin et al. 2009; Zhu et al. 2013] for refractive optics. We note that this concept is also closely related to specializing optics depending on the incident ray direction in light field imaging [Ng et al. 2005], for example, tailoring optical aberrations for digital correction [Ng et al. 2012]. Specifically, we introduce a *virtual aperture* $\mathcal{A}(r, \omega_i) = circ[r - v(\omega_i)]$ to partition the incident light bundle of each direction into a peak component that we optimize for, while treating out-of-aperture components as out-of-focus blur. Here, $circ[\cdot]$ is a function representing a circular aperture, $v(\omega_i)$ indicates the axial center of the virtual aperture subject to the i^{th} incident ray direction. With this aperture partitioning, we optimize for the phase profile solving the following optimization problem:

$$[\phi_0(r), \rho, v] = \underset{\phi_0(r), \rho, v}{\arg\min} \sum_{i=1}^{N} \|\mathcal{A}(r, \omega_i)(\phi_0(r) - \phi(r, \omega_i))\|_2^2. \tag{4}$$

Note that the virtual aperture is not a physical aperture of the optical system, but is only introduced as a conceptual partitioning in the lens optimization. Figure 4 shows the virtual apertures for uniformly sampled directions superimposed on the real aperture. For every direction, we optimize only for the rays that pass through the corresponding virtual apertures; these will be focused into a sharp PSF, while all other rays from the same direction that miss $\mathcal{A}$ but pass through the full aperture D will be blurred and manifest as a low frequency "haze" in the measurement.

4.3 Fresnel Depth Profile Optimization

We solve the optimization problem from Eq. 4 using Zemax [Geary 2002]. While Eq. 4 minimizes phase differences, Zemax interprets it as minimizing the optical path difference (OPD). Zemax allows us to piggy-back on a library of parameterized surface types, and directly optimize a deep Fresnel lens profile (a deeper micro-structure than regular 2π modulation.) instead of sequentially optimizing for the phase and depth in a two-stage process. We formulate the problem from Eq. 4 using the multiple configuration function with the number of the configurations set to the discretized aperture directions (7 in this paper, uniformly sampled on half of the diagonal image size). We set the size of each virtual aperture – the effective aperture that contributes to focusing light bundles – to one third of the clear aperture. As shown in Figure 4, the center v of the virtual aperture for each direction along the clear aperture plane can be modeled by shifting a stop along the optical axis. This allows us to optimize the location of the virtual aperture by setting the stop position as an additional optimization variable. The merit (objective) function used in Zemax includes terms for minimizing the wavefront (phase) error at each sampled direction, and enforcing a desired effective focal length (EFL). We refer the reader to the supplementary document for additional details.

4.4 Aberration Analysis

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

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

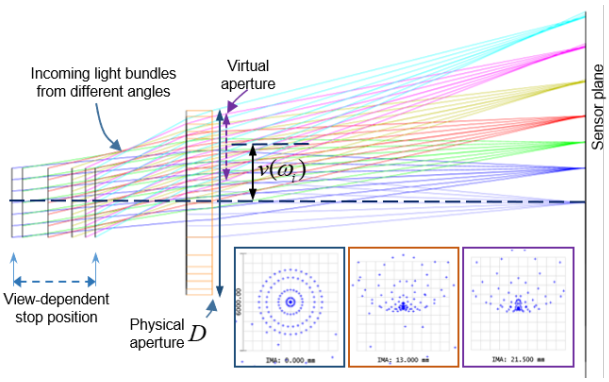

Fig. 4. Schematic of aperture partitioning approach. The spatial position of a virtual aperture (specified by the offset from the optical axis and visualized with different colors) along the radial direction is determined by the controlled position of the stop (on the left), that is further dependent on incident ray directions. The synthetic spot distributions of three directions are presented as inserts, from each pattern we observe a sharp peak that fits well to our design goal of PSFs.

the ability to detect some residual contrast, instead of completely losing contrast.

5 LEARNED IMAGE RECONSTRUCTION

In this section, we describe the forward image formation model, which models sensor measurements using the proposed optical design, and we present our learned reconstruction model which retrieves high-quality images from these measurements.

5.1 Image Formation Model

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

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

$$b_c(x,y) = \int Q_c(\lambda) \cdot [p(x,y,d,\lambda,i_c) * i_c(x,y)] d\lambda + n(x,y), \quad (5)$$

where the PSF $p(x,y,d,\lambda,i_c)$ varies with the spatial position (x,y) on the sensor, the depth d of scene, and the incident spectral distribution λ. Q_c is the color response of the sensor, and $i(x,y)$ and $n(x,y)$ represent the latent image and measurement noise, respectively. The PSF may also exhibit non-linearity in high-intensity regions, which is why the PSF p takes the latent channel i_c as further parameter. The noise may have complex characteristics, including signal-dependent shot noise as well as read noise introduced in the measurement process. We refer readers to the EMVA Standard [2005] for a detailed discussion of noise sources and calibration of the proposed model from Eq. (5).

Given a sensor measurement, conventional image processing pipelines perform a sequence of operations, each addressing an individual reconstruction problem, such as white balance, demosaicing, color calibration, digital gain, gamma compression and tone mapping [Ramanath et al. 2005]. Errors occurring during any of these operations can accumulate, adding to the ill-posedness of the overall image reconstruction problem [Brooks et al. 2018; Heide et al. 2014], that is recovering i from b by inverting Eq. 5.

To recover a latent image from the degraded image, existing methods typically perform deconvolution using optimization [Krishnan and Fergus 2009], addressing the ill-posedness of the reconstruction problem using natural image priors. We refer to the supplementary document for details on traditional deconvolution methods. However, large PSFs with hundreds of pixels in diameter and high wavelength and depth-dependency cannot be tackled by existing methods. While the scene dependency of the aberrations may be addressed with blind deconvolution approaches, these methods are currently limited to small PSF sizes of ca. 10-20 pixels in diameter [Sun et al. 2013]. Hence, existing image reconstruction methods cannot compensate for the low-frequency tail of the proposed PSF and scene-dependent PSF variation, as shown in Figure 8.

To handle the scene-dependence and non-linearities in the image formation model, i.e., PSF dependency on i in Eq. (5), we deviate from existing methods in that we do not pre-calibrate a PSF for a given illumination, and approximate the scene with broadband spectral response, but instead solve for a given image without an intermediate PSF estimate. This is done by directly learning a image-to-image mapping using a deep neural network. Next, we describe the network architecture, training methodology, and training data acquisition.

5.2 Generative Image Recovery

We propose a generative adversarial network (GAN) for the retrieval of the latent clean image i from corrupted raw sensor measurements b. Instead of relying on existing hand-crafted loss functions, which encourage overfitting as we will show below, using a GAN allows us to learn a robust loss function along with the reconstruction model. Moreover, in the learning of this loss function, we can augment training pairs for supervised training with unpaired training data from high-quality lens captures. The proposed framework is shown in Figure 5. Specifically, we adopt a variant of the U-Net architecture [Ronneberger et al. 2015], as our generative model G, referred to as *Generator* in the following, while the discriminative critic network D is referred to as *Discriminator*. During training, the generator is trained to produce latent estimates that "fool" the discriminator network into classifying the estimate as a high-quality image, while this discriminator is trained to better distinguish between images from compound lenses and the estimates produced from the generator. We use training data without blurry correspondences to augment the training of the discriminator, in a semi-supervised fashion, while the generator model is trained using a combination of a learned perceptual loss between the predicted image and the reference, and the discriminator loss using a Wasserstein generative adversarial framework.

5.2.1 *Network Architecture.*

Generator. The proposed generator network consists of a contracting path and an expansive path (Figure 5). Specifically, the contracting path consists of a 4×4 initial feature extraction layer, the repeated application of the Leaky rectified linear unit (LeakyReLU), a 4×4 convolution layer with stride 2 for downsampling,

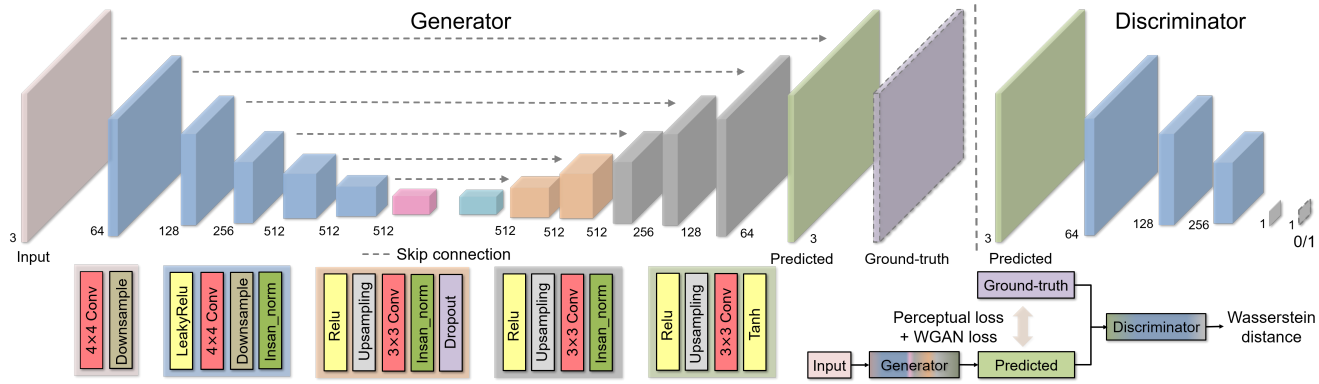

Fig. 5. Generative image reconstruction architecture. The generator model is shown on the left and the layer configurations of encoder/decoder stages are illustrated with different colored blocks (bottom-left). We apply skip connections in every decoder stage. In particular, we use a combination of a perceptual loss between the predicted image and the ground-truth, and a Wasserstein generative adversarial loss. The discriminator model for the GAN loss is similar to the encoder architecture.

and instance normalization layers. The LeakyReLU allows back-propagating the error signal to the earlier layer and the instance normalization (i.e. single batching training in this work), to avoid the crosstalk between samples in a batch. At each downsampling convolution, we double the number of feature channels. The total number of downsampling convolution steps is 7.

The expansive path consists of a stack of rectified linear units (ReLUs), the upsampling convolution layer, and the instance normalization. We use a nearest neighbor upsampling and a 3×3 convolution layer instead of the transposed convolution, that practically reduces the checkerboard artifacts caused by uneven overlaps [Odena et al. 2016]. Moreover, as shown in Figure 5, we concatenate the feature maps from the contracting path to introduce high frequencies so as to preserve fine scene details.

Discriminator. As illustrated in Figure 5, the discriminator consists of five 4×4 convolution layers with stride 2 for downsampling, where each layer is followed by a LeakyReLU activation layer and instance normalization, except for the first. We also double the number of feature channels after each downsampling layer. See Figure 5 and its caption for additional detail.

5.2.2 *Loss Functions.*

Perceptual loss. Feed-forward CNNs are often trained using a per-pixel loss (e.g. usually ℓ_1 or mean absolute error (MAE) loss and ℓ_2 or mean square error (MSE) loss) between the output and the ground-truth labels. However, this approach may lead to overly blurry outputs due to the pixel-wise average of possible optima [Ledig et al. 2017]. To obtain visually pleasing results that generalize to real data, we add a perceptual loss [Johnson et al. 2016] to our learned GAN loss. This loss component compares two images subject to the high-level representations from the pre-trained CNN. We use the VGG19 network in all our experiments. Let $\mathbb{A}_k(i)$ be the activations at the k^{th} layer of the pre-trained VGG19 network Φ with an input image i. Given a feature map $\mathbb{A}_k(i)$ with the shape of $C_k \times H_k \times W_k$, the Gram matrix, with a size of $C_k \times C_k$, can be expressed as:

$$Gram_k^{\Phi}(i) = \psi\psi^T / C_k H_k W_k, \tag{6}$$

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

$$\mathcal{L}_c = \Sigma_k \| Gram_k^{\Phi}(i) - Gram_k^{\Phi}(G(b)) \|_1. \tag{7}$$

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

Adversarial loss. We use an adversarial loss to learn a robust loss function, along with the actual generator network, which better generalizes to measured data than hand-crafted per-pixel losses. Instead of adopting a vanilla GAN [Goodfellow et al. 2014] training procedure, we rely on variant of the Wasserstein GAN [Arjovsky et al. 2017] with a gradient penalty to enforce a more robust training process with the U-Net generator in our training pipeline. The resulting adversarial loss can be expressed as:

$$\mathcal{L}_{adv} = \underbrace{\mathop{\mathbb{E}}_{i\sim\mathbb{P}_r}[D(i)] - \mathop{\mathbb{E}}_{\tilde{i}\sim\mathbb{P}_g}[D(\tilde{i})]}_{critic\ loss} + \underbrace{\lambda_g \mathop{\mathbb{E}}_{\tilde{i}\sim\mathbb{P}_{\tilde{i}}}[(\|\nabla_{\tilde{i}} D(\tilde{i})\|_2 - 1)^2]}_{gradient\ penalty}, \tag{8}$$

where $\mathbb{P}_r$ and $\mathbb{P}_g$ are distributions of data and model, respectively. Note that r contains here more sharp captured images than corresponding blurry/sharp pairs. Intuitively, the adversarial loss attempts to minimize the structural deviation between a model-generated image $\tilde{i} = G(b)$ and a real image i, penalizing missing structures, while relaxing the requirements on high color-accuracy and SNR in heavily blurred regions. We will analyze this behavior further in Sec. 8.4.

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

$$\mathcal{L}_{total} = \mathcal{L}_c + \lambda_a \mathcal{L}_{adv}. \tag{9}$$

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

6 DATASETS

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

To overcome these restrictive capture issues, we have built a *display-capture lab setup* that allows us to efficiently generate a large amount of training data without large human labor. This is realized by capturing images that are sequentially displayed on a high resolution LCD monitor (Asus PA32U), as shown in Figure 6. As a benefit of the fact that our PSF is shift-invariant, the proposed lens design does not require training over the full FOV. Instead, we train our network on a narrow field of view, which allows us to overcome prohibitive memory limitations during training with current generation GPU hardware. Moreover, this feature further aids the calibration over our large FOV. During testing we run the network on the CPU to process full-resolution measurements. We use two datasets using a Canon 5D and a Nikon D700 from the Adobe 5k set which contains in total 814 images. To cover the full FOV, we additionally select the first 200 images by name order from the 814 images and capture them by setting the monitor at large FOV. The test set is selected by name order (i.e. first 100 images) from the Canon 40D subset of the Adobe 5k set. All the images are resized to fit with the resolution of the display monitor and converted to Adobe RGB colorspace.

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

(1) We calibrated the tone curve and color reproduction of the LCD monitor using the i1 Pro calibration suite.
(2) We calibrated the system uniformity (including both the brightness uniformity of the LCD monitor and imaging vignetting of the capturing camera) by capturing a white calibration chart.
(3) We obtained coarse distortion correction parameters of the captured image and the alignment transfer matrix between the captured image and ground-truth image displayed on the monitor by capturing several known checkerboard patterns displayed on the LCD monitor [Zhang 2000].

Training Details. For training purposes, we crop both the preprocessed raw and ground-truth images into 512×512 and 1024×1024 patch pairs. These training pairs are randomly flipped and rotated to augment the training process. To preserve color fidelity, we normalize the image to range $[0, 1]$ instead of subtracting the mean and dividing its corresponding standard deviation. We choose the ADAM optimizer with $\beta_1 = 0.5$ and $\beta_2 = 0.999$, which exhibits robustness to the high noise level of our input. At first, the learning rate is initialized as 0.0001 for the first 100 epochs and linearly decayed to 0 over another 150 epochs using 512×512 patch pairs. Then, the learning rate is initialized as 0.00002 for the first 50 epochs and linearly decayed to 0 over another 50 epochs using 1024×1024 patch pairs. The batch size is set to 1 to avoid the crosstalk among samples in the batch. In all of our experiments, we set the loss weights in Eq. 9 to be $\lambda_a = 0.1$. During each training iteration, D is updated 5 times while G is updated once.

Fig. 6. Top: Illustration of our display-capture setup for preparing the training data set. Selected displayed and captured image pairs are shown as inserts; Bottom: Results on testing set images captured by our lenses. For each example we show the degraded measurement and reconstruction side-by-side.

The proposed network architecture is implemented with PyTorch 0.4, and the training process takes around 80 hours in total on a single Nvidia Tesla V100 GPU. Limited by the GPU memory that currently only allows processing up to around 12M pixels, we are unable to fit in a full resolution (i.e. 6k) image on the GPU. As an alternative solution, we solve the full resolution versions on E5-2687 CPUs which process each 6k image in 6 minutes. In addition, processing a 4k × 3k image on the GPU takes around 10s. Note, that with the rapidly emerging support of neural network computing, a hybrid memory architecture with efficient caching (e.g. GraphCore's IPU architecture) and quantization [Chang et al. 2018a] may lift this hardware limitation within the coming year.

7 PROTOTYPE

We realize the proposed lens objective, using the same optimization method, for two single element lenses, one with two optical surfaces (on both sides of the same flat carrier), the other with a single optical surface. The field of view and focal length of the lens prototypes are 53° and 43 mm with a real clear aperture size of 23.4 mm, respectively. To fabricate our lenses, we use a CNC machining system that supports 5-axis single point diamond turning (Nanotech 350FG) [Fang et al. 2013]. The substrate is polymethyl methacrylate (PMMA) with a refractive index of 1.493 at the principle wavelength of 550 nm. We use 200π phase modulation rather than regular 2π to wrap the optimized height map since our designed surface type is a

deep Fresnel surface. As a result, the final prototype lens has an effective modulation thickness of 120 μm and a total thickness of 3 mm (10 mm) including the planar substrate. The total clear aperture size of the lens is 23.4 mm with a focal length 43 mm corresponding to an f number of $f/1.8$ in the traditional sense. However, note that the effective aperture which contributes the sharp intensity peaks has a size of 8 mm yielding an effective f-number of $f/5.4$.

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

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

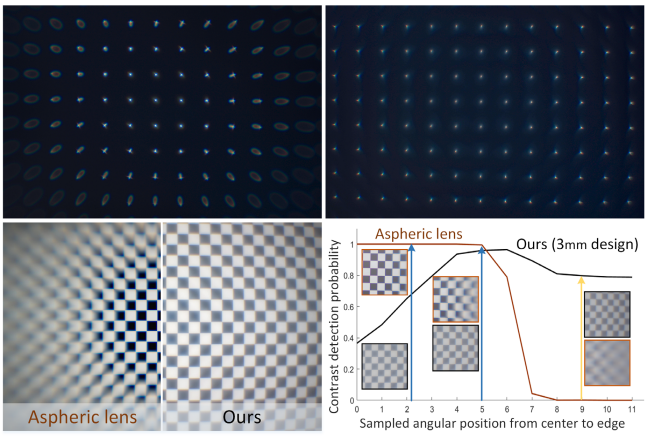

Fig. 7. PSF behavior comparison (top) and corresponding checkerboard capture comparison (bottom) between an off-the-shelf aspherical lens (ASP) and our prototype lens. Bottom-left shows the side-by-side comparison of the measurements of ASP and ours. Bottom-right shows the derived distributions of contrast detection probability. The confidence interval is set to 95% in both examples, refer to the original reference for details. Here we use a plano-convex aspherical lens (Thorlabs AL2550G) as the comparison.

8 ANALYSIS

8.1 Field of View Analysis

Figure 7 shows the spatial distribution of the aberrations and example captures of a checkerboard target across the full sensor. Our design balances the contrast detection probability [Geese et al. 2018] (CDP) across the full field of view. CDP is a probabilistic measure that allows us to characterize the ability of a higher-level processing block to detect a given contrast between two reference points after the full imaging chain.

We measure the local CDP of different measurement patches of our lens and that of an aspherical lens, see Figure 7. The reference points for this measurement are picked with 100% contrast between local patches with a lateral distance of 3 σ, with σ being the full-width-half-max (FWHM) of the peak mode of our PSF. This allows us to characterize CDP for our dual-mixture PSF without needing to vary the size of measurement patches. For our lens, a significant CDP floor of almost 50% is preserved across the full FOV, ranging from 40% at on-axis angular direction to stay above 80% at the most tilted angle. Since the PSF is not completely spatially invariant, the plot exhibits a maximum around 0.5× half-FOV where the lens focuses best. In contrast, the CDP of the aspherical lens drops drastically and approaches 0% at view directions larger than 0.5× half-FOV. The measurements agree well with our design goal that the sharp peak of our dual-mixture PSF preserves high-frequency detail and local contrast required for the feature extraction blocks in deep network models.

8.2 Generalization Analysis

The training data acquired using the proposed lab setup suffers from mismatching spectrum and tone curve, non-uniformity, etc., when compared to measurements in the wild. The most critical differences are the limited dynamic range and fixed depth of field of the monitor. Therefore, vanilla supervised learning using a per-pixel loss (i.e. MAE or MSE) overfits to these non-uniformities, hence achieving high quantitative results on a validation set displayed on the same setup, but suffers from severe artefacts on a real-world test set. The proposed semi-supervised adversarial loss, and the perceptual loss achieve robustness to this "noise" in the training data for the given approach. We validate the impact of these algorithmic components visually in Figure 8 (in large off-axis regions), and quantitatively against state-of-the-art recovery methods in Table 1.

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

To validate the proposed method against existing supervised training approaches, and assess the effect of the proposed loss functions, we train a U-net with the same structure as that of our generator on the lab-acquired data. Figure 8 and Table 1 show that vanilla supervised training overfits to the dataset acquired from the lab setup, which causes it to perform much better on captured data under the same condition (validation dataset) but to fail on real-world captures. In addition, we train pix2pix [Isola et al. 2017] as an adversarial approach while enforcing an ℓ_1 loss instead of perceptual

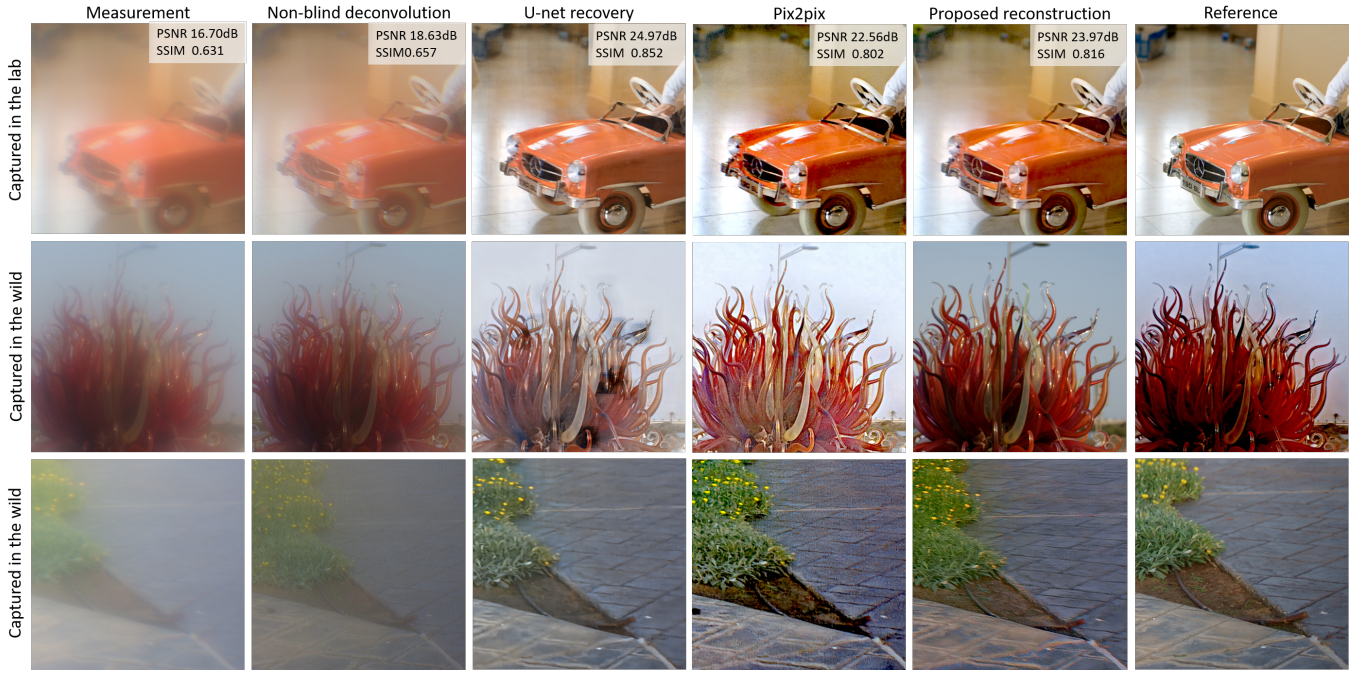

Fig. 8. Comparison of off-axis image patches recovered using different reconstruction algorithms described in Table 1. The first row presents the displayed validation from the test set of our learned reconstruction, while the remaining two rows present the data captured in the wild. Due to the mismatch of spectrum, dynamic range, and depth of field, vanilla supervised learning using a per-pixel loss may show good quantitative results while suffer from severe artefacts on real-world data. For these two examples, we present the image captured using an off-the-shelf compound lens (Sony Zeiss 50 mm $f/1.4$ Lens) as the reference. Full images are shown in the supplementary document.

Table 1. Quantitative comparison of image recovery performance of the 10 mm lens for recent deconvolution methods, including non-blind cross-channel deconvolution [Heide et al. 2016] (Cross), fully supervised U-net recovery, U-net + GAN + ℓ_1 loss (pix2pix), and our U-net + GAN + perceptual loss. We assess PSNR, SSIM, and the perceptual loss component from our model. The right-most column shows the ASP lens used in Figure 7 and Figure 9 fine-tuned for our network. Note that the fully supervised U-net in the third row does *overfit to the lab display-capture setup* and fails to generalize to real capture scenarios.

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

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

8.3 Fine-tuning for Alternative Lens Designs

To validate the efficacy of the proposed lens design, we fine-tune the described recovery method, using the same network, data and training methodology, with an aspherical lens. Compared to this alternative single-element design, the proposed design offers substantially improved sharpness in off-axis regions while trading off on-axis sharpness, as shown in Figure 9. The significant improvement in PSNR across full FOV, see last column of Table 1 and more real captures in the supplement, validate that not only the recovery algorithm is responsible for image quality but that our mixture PSF design plays an essential role in proposed computational imaging technique.

8.4 Hallucination Analysis

The evaluation and understanding of the robustness of deep networks is an active area of research. To analyze if the proposed method hallucinates image content that is not present in the measurements, we visualize the outliers with respect to perceptual and SNR metrics on a held-out validation set with known ground truth. Figure 10 plots the histograms of errors of image patches with respect to ℓ_1, ℓ_2, and the discussed perceptual loss. We show the outliers of these plots in the same figure. Other than suffering from slight blur and color inaccuracy, our recovered results do not hallucinate detail that is not present. Note that the presented image patches are the outliers with the largest error values. As the histogram mode is separated significantly from the presented outliers,

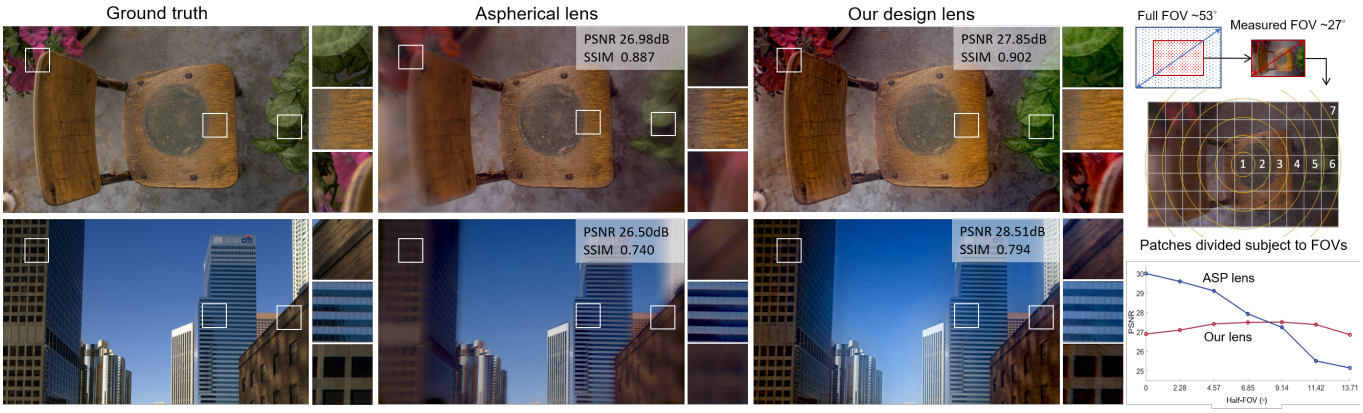

Fig. 9. Comparison of different regions on images recovered using our learned image recovery algorithm from data captured by an off-the-shelf aspherical lens (ASP) and our prototype lens. Although trading off on-axis sharpness in some sense, ours exhibits much better quality in off-axis regions. The lens parameters and settings are the same as in Figure 7. The plots on the right reveal the averaged PSNRs of patches subject to FOV over 100 validation images. Note, in this comparison we investigate only half of the full FOV of our design because the required resolution limits the FOV when using a consumer display monitor. We observe that even within this intermediate range of FOV, the recovered image quality of ASP drops drastically when the investigated half-FOV goes beyond 7°.

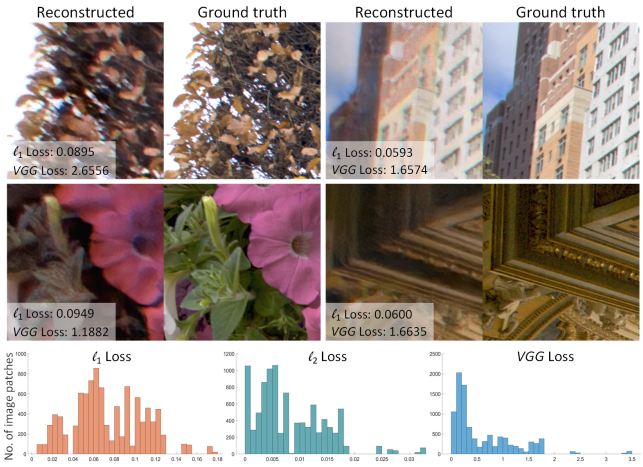

Fig. 10. Outlier analysis of reconstruction images of our deep network. For each pair we show the recovered patch (left) and its corresponding ground truth patch (right). The plots show the histograms of 9,600 evaluated image patches under three error functions, ℓ_1 loss, ℓ_2 loss, and perceptual (VGG) loss.

we conclude that the proposed reconstruction method is robust and does not hallucinate major detail. Please see the supplemental material for additional outlier visualizations.

9 EXPERIMENTAL ASSESSMENT

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

The first two rows of Figure 11 show reconstructions of indoor and outdoor scenes under both artificial illumination and natural light. Our method successfully preserves both fine details and color fidelity across full field-of-view. Note that all captures have been obtained using a clear aperture setting, i.e. $f/1.8$, and a full sensor resolution.

Single-surface Design. Next, we show results captured with a *single-surface* thin-plate lens with only the rear surface machined. The resulting optical layout and simulated optical behavior are detailed in the supplementary document. For this design, we have reduced the thickness of the substrate plate to 3 mm. The third row of Figure 11 shows results for indoor and outdoor scenes captured with this single-surface prototype.

9.1 Imaging over Large Depth Ranges and in Low Light

Figure 12 shows reconstruction results for scenes with large depth ranges and in low-light scenarios. Although we only train the proposed method with screen captures at a fixed distance, the proposed method preserves the depth-dependent defocus, i.e., bokeh, for scenes with large depth ranges. Careful readers notice that for high-intensity regions, as in the sky, our reconstruction does not recover detail. As outlined in Section 1, this is because the training data does not contain high-dynamic range captures for our low-dynamic range LCD monitor setup.

In contrast to alternative flat optical designs with wide FOV, such as pinholes with theoretically unlimited FOV, the proposed lens design allows for low-light captures due to its f-number of $f/5.4$. We demonstrate low-light and short-exposure imaging scenarios in the second two rows of Figure 12, where we compare our design against a pinhole (0.8 mm) that suppresses most aberrations over a wide FOV at the cost of very limited light throughput. The pinhole measurements are low-signal and hence corrupted with severe noise

Fig. 11. Experimental results of dual-surface lens design (first two rows) and that of single surface lens design (third row) on real word scenes. For each pair, we show the degraded measurement and the reconstruction result. The exposure time for these images are set 0.8, 125, 0.5, 1.25, 0.5, 0.4 ms with ISO 50. Refer to supplementary document for more real world results.

that results in a poor reconstruction – even with state-of-the-art low-light denoising methods [Chen et al. 2018]. Additional comparisons at different exposure levels can be found in the supplement.

10 DISCUSSION AND CONCLUSION

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

Specifically, we introduce a compact thin-plate lens design with a *dual-mixture* PSF distribution across the full FOV. Although the PSF has an extremely large spot size of ≥ 900 pixels in diameter, it preserves local contrast uniformly across the sensor plane. To recover images from such degraded measurements, we learn a deep generative model that maps captured blurry images to clean reconstructed images. To this end, we propose an automated capture method to acquire aligned training data. We tackle the mismatch between lab-captured and natural images in the wild – prohibiting vanilla supervised learning to perform well on real world scenes – by introducing a combination of adversarial and perceptual loss components. Together, the proposed network architecture, training methodology, and data acquisition, allow us to achieve image quality that makes a significant step towards the quality of commercial compound lens systems with just a single free-form lens. We have validated the proposed approach experimentally on a wide variety of challenging outdoor and indoor scenes.

While the proposed approach could enable high-quality imagery with thin and inexpensive optics in the future, on today's consumer graphics hardware, the described reconstruction method is memory-limited for models at full 24.3 Megapixel sensor resolution. Therefore, we run the post-processing on the CPU which results in low throughput on the order of minutes per image – far from that of

Fig. 12. Experimental results of large depth range imaging (top row) and low-light imaging (bottom row). The exposure time and ISO for the top two examples are set (3.125, 1) ms and ISO 50, while that for the bottom two examples are set (10 ms with ISO 500) and (20 ms with ISO 25,600).

modern image processing pipelines. The upcoming graphics hardware generation will likely overcome this memory limitation. In the meantime, a combination of cloud processing and low resolution or tile-based previews could make the method practical. The lab data acquisition is currently restricted by the dynamic range of consumer displays, which we hope to overcome in the future with alternative high-dynamic range display approaches.

Although our thin-plate lens design significantly reduces the form factor compared to complex optical systems, we validate the concept with a focal power and an aperture size comparable to existing DSLR camera lenses. To achieve the envisioned camera device form factors, a reduction in both size of the optical lens system and focal length are necessary. Miniature lens systems with short back focal length (e.g. $\leq$ 5 mm) are now possible by introducing metasurfaces or injection molding techniques to fabricate the optics, which provide feature sizes at the order of the wavelength of light and hence can diffract light at steeper angles allowing for ultra-short focal lengths.

While we designed a single-element lens in this work, dual-refractive lenses or hybrid refractive-diffractive optical systems might be interesting directions for future research. Moreover, simple optics for sensor arrays, such as the PiCam [Venkataraman et al. 2013], could be revisited with the proposed PSF design. Although this work focuses on computational photography applications, we envision a wide range of applications across computer vision, robotics, sensing and human-computer interaction, where large field-of-view imaging with simple optics and domains-specific post-processing could enable unprecedented device form factors.

ACKNOWLEDGMENTS

The authors thank Liang Xu and Xu Liu from Zhejiang University for assisting in the manufacturing of the lens prototypes.

REFERENCES

Se Hyun Ahn and L Jay Guo. 2009. Large-area roll-to-roll and roll-to-plate nanoimprint lithography: a step toward high-throughput application of continuous nanoimprinting. *ACS Nano* 3, 8 (2009), 2304–2310.

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

Martin Arjovsky, Soumith Chintala, and Léon Bottou. 2017. Wasserstein gan. *arXiv preprint arXiv:1701.07875* (2017).

Yoshua Bengio. 2012. Deep learning of representations for unsupervised and transfer learning. In *Proceedings of ICML Workshop on Unsupervised and Transfer Learning*. 17–36.

Glenn D Boreman. 2001. *Modulation transfer function in optical and electro-optical systems*. Vol. 21. SPIE press Bellingham, WA.

David J Brady, Michael E Gehm, Ronald A Stack, Daniel L Marks, David S Kittle, Dathon R Golish, EM Vera, and Steven D Feller. 2012. Multiscale gigapixel photography. *Nature* 486, 7403 (2012), 386.

Tim Brooks, Ben Mildenhall, Tianfan Xue, Jiawen Chen, Dillon Sharlet, and Jonathan T Barron. 2018. Unprocessing Images for Learned Raw Denoising. *arXiv preprint arXiv:1811.11127* (2018).

Julie Chang, Vincent Sitzmann, Xiong Dun, Wolfgang Heidrich, and Gordon Wetzstein. 2018b. Hybrid optical-electronic convolutional neural networks with optimized diffractive optics for image classification. *Scientific reports* 8, 1 (2018), 12324.

Julie Chang and Gordon Wetzstein. 2019. Deep Optics for Monocular Depth Estimation and 3D Object Detection. *arXiv preprint arXiv:1904.08601* (2019).

Li-Wen Chang, Yang Chen, Wenlei Bao, Amit Agarwal, Eldar Akchurin, Ke Deng, and Emad Barsoum. 2018a. Accelerating Recurrent Neural Networks through Compiler Techniques and Quantization. (2018).

Chen Chen, Qifeng Chen, Jia Xu, and Vladlen Koltun. 2018. Learning to See in the Dark. (2018).

Taeg Sang Cho, Charles L Zitnick, Neel Joshi, Sing Bing Kang, Richard Szeliski, and William T Freeman. 2012. Image restoration by matching gradient distributions. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)* 34, 4 (2012), 683–694.

Stephen Y Chou, Peter R Krauss, and Preston J Renstrom. 1996. Nanoimprint lithography. *Journal of Vacuum Science & Technology B: Microelectronics and Nanometer Structures Processing, Measurement, and Phenomena* 14, 6 (1996), 4129–4133.

Oliver Cossairt and Shree Nayar. 2010. Spectral focal sweep: Extended depth of field from chromatic aberrations. In *Computational Photography (ICCP), IEEE International Conference on*. IEEE, 1–8.

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

Wang Duoshu, Chongtai Luo, Yuqing Xiong, Tao Chen, Hongkai Liu, and Jizhou Wang. 2011. Fabrication technology of the centrosymmetric continuous relief diffractive optical elements. *Physics Procedia* 18 (2011), 95–99.

EMVA Standard. 2005. 1288: Standard for characterization and presentation of specification data for image sensors and cameras. *European Machine Vision Association* (2005).

Magali Estribeau and Pierre Magnan. 2004. Fast MTF measurement of CMOS imagers using ISO 12333 slanted-edge methodology. In *Detectors and Associated Signal Processing*, Vol. 5251. International Society for Optics and Photonics, 243–253.

FZ Fang, XD Zhang, A Weckenmann, GX Zhang, and C Evans. 2013. Manufacturing and measurement of freeform optics. *CIRP Annals* 62, 2 (2013), 823–846.

Grant R Fowles. 2012. *Introduction to modern optics*. Courier Dover Publications.

Carl Friedrich Gauss. 1843. *Dioptrische Untersuchungen von CF Gauss*. in der Dieterichschen Buchhandlung.

Joseph M Geary. 2002. *Introduction to lens design: with practical ZEMAX examples*. Willmann-Bell Richmond.

Marc Geese, Ulrich Seger, and Alfredo Paolillo. 2018. Detection Probabilities: Performance Prediction for Sensors of Autonomous Vehicles. *Electronic Imaging* 2018, 17 (2018), 148–1–148–14.

Patrice Genevet, Federico Capasso, Francesco Aieta, Mohammadreza Khorasaninejad, and Robert Devlin. 2017. Recent advances in planar optics: from plasmonic to dielectric metasurfaces. *Optica* 4, 1 (2017), 139–152.

Erez Gilad and Jost Von Hardenberg. 2006. A fast algorithm for convolution integrals with space and time variant kernels. *J. Comput. Phys.* 216, 1 (2006), 326–336.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *Advances in neural information processing systems*. 2672–2680.

Joseph W Goodman. 2005. *Introduction to Fourier optics*. Roberts and Company Publishers.

Hrel Haim, Shay Elmalem, Raja Giryes, Alex M. Bronstein, and Emanuel Marom. 2018. Depth Estimation From a Single Image Using Deep Learned Phase Coded Mask. *IEEE Trans. Computational Imaging* 4, 3 (2018), 298–310.

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 Transactions on Graphics (TOG)* 35, 6 (2016), 192.

Eugene Hecht. 1998. Hecht optics. *Addison Wesley* 997 (1998), 213–214.

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

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

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

Minyoung Huh, Pulkit Agrawal, and Alexei A Efros. 2016. What makes ImageNet good for transfer learning? *arXiv preprint arXiv:1608.08614* (2016).

Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. 2017. Image-to-Image Translation with Conditional Adversarial Networks. *CVPR* (2017).

Justin Johnson, Alexandre Alahi, and Li Fei-Fei. 2016. Perceptual losses for real-time style transfer and super-resolution. In *European Conference on Computer Vision*. Springer, 694–711.

Arnold Kalvach and Zsolt Szabó. 2016. Aberration-free flat lens design for a wide range of incident angles. *Journal of the Optical Society of America B* 33, 2 (2016), A66.

Rudolf Kingslake and R Barry Johnson. 2009. *Lens design fundamentals*. Academic Press.

Dilip Krishnan and Rob Fergus. 2009. Fast image deconvolution using hyper-Laplacian priors. In *Advances in Neural Information Processing Systems*. NIPS, 1033–1041.

Orest Kupyn, Volodymyr Budzan, Mykola Mykhailych, Dmytro Mishkin, and Jiri Matas. 2017. DeblurGAN: Blind Motion Deblurring Using Conditional Adversarial Networks. *arXiv preprint arXiv:1711.07064* (2017).

Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew P Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. 2017. Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network. In *CVPR*, Vol. 2. 4.

Anat Levin, Samuel W Hasinoff, Paul Green, Frédo Durand, and William T Freeman. 2009. 4D frequency analysis of computational cameras for depth of field extension. In *ACM Transactions on Graphics (TOG)*, Vol. 28. ACM, 97.

Light.co. 2018. Light L16 Camera. https://light.co

Daniel Malacara-Hernández and Zacarías Malacara-Hernández. 2016. *Handbook of optical design*. CRC Press.

Kaushik Mitra, Oliver Cossairt, and Ashok Veeraraghavan. 2014. To denoise or deblur: parameter optimization for imaging systems. In *Digital Photography X*, Vol. 9023. International Society for Optics and Photonics, 90230G.

Anish Mittal, Anush Krishna Moorthy, and Alan Conrad Bovik. 2012. No-reference image quality assessment in the spatial domain. *IEEE Transactions on Image Processing* 21, 12 (2012), 4695–4708.

MobilEye. 2018. MobileEye TriCam. https://press.zf.com/site/press/en_de/microsites/press/list/release/release_41735.html

Mehjabin Monjur, Leonidas Spinoulas, Patrick R Gill, and David G Stork. 2015. Ultra-miniature, computationally efficient diffractive visual-bar-position sensor. In *Proc. SensorComm*. IEIFSA.

Seungjun Nah, Tae Hyun Kim, and Kyoung Mu Lee. 2017. Deep multi-scale convolutional neural network for dynamic scene deblurring. In *CVPR*, Vol. 1. 3.

Ren Ng, Marc Levoy, Mathieu Brédif, Gene Duval, Mark Horowitz, Pat Hanrahan, et al. 2005. Light field photography with a hand-held plenoptic camera. (2005).

Yi-Ren Ng, Patrick M Hanrahan, Mark A Horowitz, and Marc S Levoy. 2012. Correction of optical aberrations. US Patent 8,243,157.

Augustus Odena, Vincent Dumoulin, and Chris Olah. 2016. Deconvolution and checkerboard artifacts. *Distill* 1, 10 (2016), e3.

Steve Oliver, Rick Lake, Shashikant Hegde, Jeff Viens, and Jacques Duparre. 2010. Imaging module with symmetrical lens system and method of manufacture. US Patent 7,710,667.

Marios Papas, Thomas Houit, Derek Nowrouzezahrai, Markus H Gross, and Wojciech Jarosz. 2012. The magic lens: refractive steganography. *ACM Trans. Graph.* 31, 6 (2012), 186–1.

Jim R Parker. 2010. *Algorithms for image processing and computer vision*. John Wiley & Sons.

Yifan Peng, Xiong Dun, Qilin Sun, and Wolfgang Heidrich. 2017. Mix-and-match holography. *ACM Transactions on Graphics (TOG)* 36, 6 (2017), 191.

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

Yifan Peng, Qiang Fu, Felix Heide, and Wolfgang Heidrich. 2016. The diffractive achromat full spectrum computational imaging with diffractive optics. *ACM Transactions on Graphics (TOG)* 35, 4 (2016), 31.

Rajeev Ramanath, Wesley Snyder, Youngjun Yoo, and Mark Drew. 2005. Color image processing pipeline in digital still cameras. *IEEE Signal Processing Magazine* 22, 1 (2005), 34–43.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*. Springer, 234–241.

Christian J. Schuler, Harold Christopher Burger, Stefan Harmeling, and Bernhard Scholkopf. 2013. A Machine Learning Approach for Non-blind Image Deconvolution. In *Proc. Computer Vision and Pattern Recognition*.

Yuliy Schwartzburg, Romain Testuz, Andrea Tagliasacchi, and Mark Pauly. 2014. High-contrast computational caustic design. *ACM Transactions on Graphics (TOG)* 33, 4 (2014), 74.

Yichang Shih, Brian Guenter, and Neel Joshi. 2012. Image enhancement using calibrated lens simulations. In *European Conference on Computer Vision*. Springer, 42–56.

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 (TOG)* 37, 4 (2018), 114.

Georgii Georgievich Sliusarev. 1984. Aberration and optical design theory. *Bristol, England, Adam Hilger, Ltd., 1984, 672 p. Translation.* (1984).

Warren J. Smith. 2005. *Modern lens design*. McGraw-Hill.

David G Stork and Patrick R Gill. 2013. Lensless ultra-miniature CMOS computational imagers and sensors. *Proc. SENSORCOMM* (2013), 186–190.

David G Stork and Patrick R Gill. 2014. Optical, mathematical, and computational foundations of lensless ultra-miniature diffractive imagers and sensors. *International Journal on Advances in Systems and Measurements* 7, 3 (2014), 4.

Libin Sun, Sunghyun Cho, Jue Wang, and James Hays. 2013. Edge-based blur kernel estimation using patch priors. In *Proc. International Conference on Computational Photography (ICCP)*. 1–8.

Kartik Venkataraman, Dan Lelescu, Jacques Duparré, Andrew McMahon, Gabriel Molina, Priyam Chatterjee, Robert Mullis, and Shree Nayar. 2013. Picam: An ultra-thin high performance monolithic camera array. *ACM Transactions on Graphics (TOG)* 32, 6 (2013), 166.

Peng Wang, Nabil Mohammad, and Rajesh Menon. 2016. Chromatic-aberration-corrected diffractive lenses for ultra-broadband focusing. *Scientific Reports* 6 (2016).

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.

Yicheng Wu, Vivek Boominathan, Huaijin Chen, Aswin Sankaranarayanan, and Ashok Veeraraghavan. 2019. PhaseCam3D – Learning Phase Masks for Passive Single View Depth Estimation. In *Proc. ICCP*.

Li Xu, Jimmy SJ Ren, Ce Liu, and Jiaya Jia. 2014. Deep convolutional neural network for image deconvolution. In *Advances in Neural Information Processing Systems*. 1790–1798.

Xiaoyun Yuan, Lu Fang, Qionghai Dai, David J Brady, and Yebin Liu. 2017. Multiscale gigapixel video: A cross resolution image matching and warping approach. In *2017 IEEE International Conference on Computational Photography (ICCP)*. IEEE, 1–9.

Jiawei Zhang, Jinshan Pan, Wei-Sheng Lai, Rynson WH Lau, and Ming-Hsuan Yang. 2017. Learning fully convolutional networks for iterative non-blind deconvolution. (2017).

Zhengyou Zhang. 2000. A flexible new technique for camera calibration. *IEEE Transactions on pattern analysis and machine intelligence* 22 (2000).

Jun Zhu, Tong Yang, and Guofan Jin. 2013. Design method of surface contour for a freeform lens with wide linear field-of-view. *Optics express* 21, 22 (2013), 26080–26092.

Margarete Zoberbier, Sven Hansen, Marc Hennemeyer, Dietrich Tönnies, Ralph Zoberbier, Markus Brehm, Andreas Kraft, Martin Eisner, and Reinhard Völkel. 2009. Wafer level cameras—novel fabrication and packaging technologies. In *Int. Image Sens. Workshop*.