End-to-end Optimization of Optics and Image Processing for Achromatic Extended Depth of Field and Super-resolution Imaging

VINCENT SITZMANN*, Stanford University, USA
STEVEN DIAMOND*, Stanford University, USA
YIFAN PENG*, The University of British Columbia, Canada and Stanford University, USA
XIONG DUN, King Abdullah University of Science and Technology, Saudi Arabia
STEPHEN BOYD, Stanford University, USA
WOLFGANG HEIDRICH, King Abdullah University of Science and Technology, Saudi Arabia
FELIX HEIDE, Stanford University, USA
GORDON WETZSTEIN, Stanford University, USA

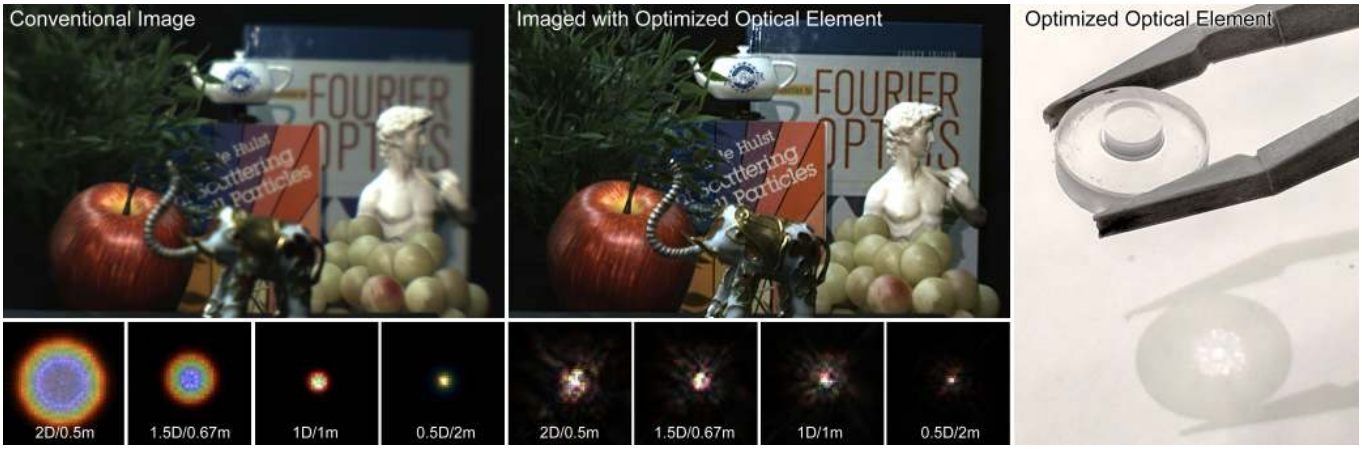

Fig. 1. One of the applications of the proposed end-to-end computational camera design paradigm is achromatic extended depth of field. When capturing an image with a regular singlet lens (top left), out-of-focus regions are blurry and chromatic aberrations further degrade the image quality. With our framework, we optimize the profile of a refractive optical element that achieves both depth and chromatic invariance. This element is fabricated using diamond turning (right) or using photolithography. After processing an image recorded with this optical element using a simple Wiener deconvolution, we obtain an all-in-focus image with little chromatic aberrations (top center). Point spread functions for both the regular lens and the optimized optical element are shown in the bottom. In this paper, we explore several applications that demonstrate the efficacy of our novel approach to domain-specific computational camera design.

Authors' addresses: Vincent Sitzmann*, Stanford University, Department of Electrical Engineering, Stanford, CA, 94305, USA, sitzmann@cs.stanford.edu; Steven Diamond*, Stanford University, Department of Computer Science, Stanford, CA, 94305, USA, diamond@cs.stanford.edu; Yifan Peng*, The University of British Columbia, Department of Computer Science, Vancouver, V6T 1Z4, Canada, Stanford University, Department of Electrical Engineering, Stanford, CA, 94305, USA, evanpeng@cs.ubc.ca; Xiong Dun, King Abdullah University of Science and Technology, Visual Computing Center, Thuwal, 23955, Saudi Arabia, xiong.dun@kaust.edu.sa; Stephen Boyd, Stanford University, Department of Electrical Engineering, Stanford, CA, 94305, USA, boyd@stanford.edu; Wolfgang Heidrich, King Abdullah University of Science and Technology, Visual Computing Center, Thuwal, 23955, Saudi Arabia, wolfgang.heidrich@kaust.edu.sa; Felix Heide, Stanford University, Department of Electrical Engineering, Stanford, CA, 94305, USA, fheide@stanford.edu; Gordon Wetzstein, Stanford University, Department of Electrical Engineering, Stanford, CA, 94305, USA, gordon.wetzstein@stanford.edu.

In typical cameras the optical system is designed first; once it is fixed, the parameters in the image processing algorithm are tuned to get good image reproduction. In contrast to this sequential design approach, we consider joint optimization of an optical system (for example, the physical shape of the lens) together with the parameters of the reconstruction algorithm. We build a fully-differentiable simulation model that maps the true source image to the reconstructed one. The model includes diffractive light propagation, depth and wavelength-dependent effects, noise and nonlinearities, and the image post-processing. We jointly optimize the optical parameters and the image processing algorithm parameters so as to minimize the deviation between the true and reconstructed image, over a large set of images. We implement our joint optimization method using autodifferentiation to efficiently compute parameter gradients in a stochastic optimization algorithm. We demonstrate the efficacy of this approach by applying it to achromatic extended depth of field and snapshot super-resolution imaging.

0730-0301/2018/8-ART114 $15.00
https://doi.org/10.1145/3197517.3201333

CCS Concepts: • **Computing methodologies → Computational photography**; **Reconstruction**;

Additional Key Words and Phrases: computational optics

ACM Reference Format:
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 Trans. Graph.* 37, 4, Article 114 (August 2018), 13 pages. https://doi.org/10.1145/3197517.3201333

1 INTRODUCTION

The visual systems of animals are often highly adapted to their environments [Land and Nielsson 2002]. In spite of being used for a diverse range of applications, digital imaging systems, on the other hand, have been engineered to mimic only one of these systems: the human eye. While such a general-purpose approach to imaging is sometimes successful, it leaves an important question unanswered: *What is the optimal camera design for a given task?* To address this question, domain-specific computational cameras have emerged over the last two decades [Nayar 2006]. By co-designing camera optics and image processing algorithms, computational cameras have the potential to optimize task-specific performance over conventional, general-purpose imaging systems in a wide range of applications.

To date, computational cameras have demonstrated new imaging capabilities, such as extended depth of field [Cossairt and Nayar 2010; Cossairt et al. 2010; Dowski and Cathey 1995], super-resolution [Ben-Ezra et al. 2004], and high dynamic range [Debevec and Malik 1997; Mann et al. 1995] imaging. Optical elements have also been optimized, for example to localize microscopic point emitters in a 3D volume via point spread function (PSF) engineering [Pavani et al. 2009; Shechtman et al. 2014] or to optimize the focusing performance of a diffractive optical element across the color spectrum [Peng et al. 2016]. Yet, all of these approaches are either heuristic or use some proxy metric on the PSF rather than considering the image quality after post-processing. Without a true end-to-end approach for jointly optimizing parameters of the image-forming optics and the algorithm processing the data, being able to find an optimal computational camera for a given task remains elusive.

What does optimizing a domain-specific computational camera entail? First, the task has to be defined and appropriate quality metrics devised to assess a camera's performance. Generative image processing tasks include denoising, deconvolution, or other forms of image reconstruction; their quality is often measured as peak signal-to-noise-ratio (PSNR). Discriminative tasks, on the other hand, would use a very different quality metric, such as classification accuracy for image classification. Second, it may be helpful to characterize the input data for a specific task. Natural images, for example, follow certain statistics that can be exploited as priors for generative tasks. But it may not always be obvious what good priors for domain-specific datasets actually are. Third, post-processing algorithms may vary drastically between different tasks or even for the same task, but in different settings. Conventional approaches to computational camera design, such as PSF engineering, do not offer the flexibility of addressing all of these challenges simultaneously.

In this paper, we introduce a new paradigm for computational camera design: end-to-end optimization of a refractive or diffractive optical element with respect to the output of a reconstruction algorithm, using stochastic gradient methods. We build a fully-differentiable wave optics image formation model that is used to jointly optimize the optical parameters and the image processing algorithm parameters for domain-specific computational cameras.

Specifically, our contributions are

- We introduce a framework for end-to-end optimization of an optical element with respect to the output of a reconstruction algorithm, using stochastic gradient methods. The framework includes a wave optics image formation model, object depth and wavelength-dependent effects, sensor noise and nonlinearities, and the image processing. The source code is publicly available [1].
- We validate this framework in simulation for the applications of achromatic extended depth of field and snapshot super-resolution imaging.
- We fabricate the optimized optical elements, using photolithography for diffractive elements and diamond turning for refractive lenses, and verify that experimental results from a prototype camera setup match the simulations.

Scope. In principle, the proposed framework for end-to-end optimization of optics and image processing generalizes to various low-level and high-level algorithms. Deep convolutional neural networks, for example, could be used to optimize the lens of a camera tailored to image classification or other high-level tasks. Exploring this large space of application- and domain-specific computational cameras is an exciting vision towards which we take first steps in this paper. We believe that the insights provided by our work on developing fully-differentiable wave optics image formation models, inverting them robustly with tools like TensorFlow, and actually fabricating the optimized optical elements are invaluable for the emerging field of computational optics.

2 RELATED WORK

Computational Cameras. Much work on computational photography has focused on improving basic capabilities of a camera, such as depth of field [Cossairt and Nayar 2010; Cossairt et al. 2010; Dowski and Cathey 1995], dynamic range [Debevec and Malik 1997; Mann et al. 1995; Reinhard et al. 2005; Rouf et al. 2011], and image resolution [Ben-Ezra et al. 2004; Brady et al. 2012; Cossairt et al. 2011]. Computational photography has also been used for tasks as diverse as motion deblurring [Raskar et al. 2006], defocus deblurring [Zhou et al. 2009; Zhou and Nayar 2009], depth estimation [Levin et al. 2007, 2009], multispectral imaging [Wagadarikar et al. 2008], light field imaging [Marwah et al. 2013; Ng et al. 2005; Veeraraghavan et al. 2007], and lensless imaging [Antipa et al. 2016; Asif et al. 2017]. Many of these approaches use either optical coding, multiplexing, burst photography [Hasinoff et al. 2016], or multi-shot approaches to capture high-dimensional visual data [Wetzstein et al. 2011].

*These authors contributed equally.

[1] https://vsitzmann.github.io/deepoptics

The proposed end-to-end optimization framework could be applied to many of these applications, as it introduces a general design paradigm for computational cameras that optimizes directly for the post-processed output with respect to a chosen quality metric and domain-specific dataset.

Deep Computational Photography. In computer vision, natural language processing, and many other fields, the emergence of deep learning has lead to rapid progress in a number of challenging tasks and state-of-the-art results for well-established problems. The computational photography community too is at the cusp of adopting tools from the deep learning community, such as convolutional neural networks. For example, high dynamic range image estimation from a single low dynamic range photograph was recently demonstrated to achieve unprecedented image quality [Eilertsen et al. 2017; Kalantari and Ramamoorthi 2017; Zhang and Lalonde 2017]. The task of super-resolving a single image has also been approached via deep learning [Dong et al. 2016b; Shi et al. 2016]. Finally, it was recently shown that it is possible to learn the mapping from a single image to a light field [Srinivasan et al. 2017] and to produce light field video clips from a hybrid camera [Wang et al. 2017].

All of these approaches have demonstrated state-of-the-art results for various computational photography applications. Yet, most of them only consider the algorithm processing the data. We go one step further and ask whether it is possible to optimize the co-design of optics and image processing for domain-specific computational cameras in an end-to-end fashion. Although we demonstrate the efficacy of our approach with applications that rely on relatively simple reconstruction algorithms, in principle, our approach could also be used to optimize optical elements leveraging more advanced deep computational photography algorithms.

Optimizing Optical Elements. Optimizing the parameters of optical elements and *point spread function engineering* are well-known techniques in the computational optics and visual computing communities. Optimized optical system parameters have proven useful for extended depth of field [Dowski and Cathey 1995; Flores et al. 2004; Liu 2007], motion [Raskar et al. 2006] and defocus [Zhou and Nayar 2009] deblurring, 4D light field imaging [Marwah et al. 2013], super-resolved localization microscopy [Pavani et al. 2009; Shechtman et al. 2014], and full-color imaging with diffractive optics [Heide et al. 2016; Peng et al. 2016]. Optimization of optical models has also been proposed for multi-element systems to either arrive at novel arrangements of off-the-shelf lenses [Sun et al. 2015] or to allow precise calibration of models [Shih et al. 2012] of these systems. Two observations remain. First, previously-proposed optimization approaches of optical elements are mainly based on heuristic cost functions applied to the PSFs, which may be a feasible approach for image deconvolution but it remains unclear how the PSF of a camera affects higher-level computer vision tasks such as image classification; second, although image processing is applied to the recorded images to remove residual aberrations or perform some inference tasks, the post-processing algorithm is usually independent of the optics design and fails to provide significant insights to guide it.

We also optimize the point spread function of an optical system, but do so in an end-to-end manner using a data-driven approach that applies a cost function on the reconstructed image, not on the PSF. Our approach is motivated by recent advances in hardware, autodifferentiation tools, and optimization algorithms for deep learning. While some recent work investigates joint optimization of either binary masks or color filter arrays with neural network post-processing for video compressed sensing or demosaicking [Chakrabarti 2016; Iliadis et al. 2016], they do not consider the optimization of phase-modulating optical elements such as lenses and do not consider diffraction in their forward model. With this work, we thus take first steps towards utilizing the full potential of end-to-end optimization for computational camera design.

Achromatic Extended Depth of Field. Extended depth of field (EDOF) is a classic application of computational imaging. An extension is the design of a single optical element that combines EDOF with achromaticity, yielding all-in-focus images with minimal chromatic aberrations. Depth and wavelength are closely coupled in the image formation, such that the seminal wavefront-coding approach to EDOF, the cubic phase plate [Dowski and Cathey 1995], displays increased achromaticity, and achromatic elements, such as the diffractive achromat [Peng et al. 2016], display some extended depth of field. However, this duality breaks down at the extremes of either wavelength or depth. One possible solution are metalenses, which have recently enabled the fabrication of single ultrathin optical elements that encode wavelength-dependent phase patterns onto the incoming light, thereby achieving physical achromaticity without the need for post-processing [Chen et al. 2018; Yang et al. 2017] and, in combination with a digital filter, achromatic EDOF [Colburn et al. 2018]. However, metalenses are currently difficult and costly to fabricate, and usually only support small numerical apertures with low light efficiency. While compound systems are well-known for their ability to correct chromatic aberrations, they increase the device footprint. Hybrid elements for achromatic EDOF have been proposed [Flores et al. 2004; Liu 2007] as a compromise that reduces chromatic aberrations by combining a single diffractive and a single refractive optical element, but add complexity to the manufacturing process.

With this work, we propose a novel perspective of joint optimization of a single diffractive or refractive element with a deconvolution post-processing step, affording achromatic EDOF with standard diamond-turning or lithography manufacturing techniques, without increasing device footprint, and with no hand-crafted losses applied to the PSF. We note that this approach does not preclude more sophisticated optical elements which may offer more degrees of freedom in the optics design. Future work may thus extend the proposed end-to-end optimization framework to other optical elements.

3 END-TO-END OPTIMIZATION OF OPTICS AND RECONSTRUCTION

In the following, we derive a wave-based image formation model that accounts for diffraction and wavelength-dependent effects when imaging natural scenes. We assume spatially incoherent light, meaning that light reflected from an object point interferes with

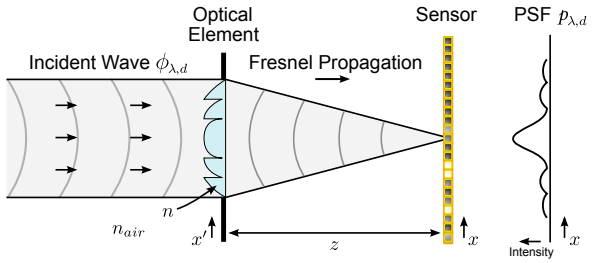

Fig. 2. Differentiable wave-based PSF simulation module. A light wave $\phi_{\lambda,d}$ with a given wavelength λ and a curvature appropriate for a point source at a distance d is incident on the aperture plane containing the optical element (refractive index n) to be optimized. The optical element shifts the phase of the incident wavefront. The resulting wavefront is propagated by the aperture-sensor distance z via Fresnel propagation. The intensities of the sensor-incident wavefront define the PSF $p_{\lambda,d}$. Optical elements can be parameterized both as height maps or using a basis representation, such as Zernike polynomials.

other light reflected from that same point along a different path, but that it does not interfere with light from other points in the scene. The image formation model is based on Fourier optics [Goodman 2017], and we show how to efficiently integrate it into the workflow of modern deep learning tools.

3.1 Image Formation Model

3.1.1 Wave-based Point Spread Function Model. Consider a single refractive or diffractive optical element, such as a thin lens. This element delays the phase of a complex-valued wave field proportionally to its thickness h

$$\phi\left(x', y'\right) = \frac{2\pi\Delta n}{\lambda} h\left(x', y'\right). \tag{1}$$

Here, λ is the wavelength and Δn is the refractive index difference between air and the material of the optical element.

A wave field U_λ with amplitude A and phase ϕ_d incident on the optical element will be affected as

$$U_\lambda\left(x', y', z=0\right) = A\left(x', y'\right) e^{i\left(\phi_d\left(x', y'\right) + \phi\left(x', y'\right)\right)}, \tag{2}$$

such that $U_\lambda\left(x', y', 0\right)$ is the wave field right after it passed through the optical element. As illustrated in Figure 2, the field propagates in free space by a distance z as

$$U_\lambda\left(x, y, z\right) = \frac{e^{ikz}}{i\lambda z} \iint U\left(x', y', 0\right) e^{\frac{ik}{2z}\left((x-x')^2 + (y-y')^2\right)} dx' dy'. \tag{3}$$

This formulation uses the Fresnel propagation operator, which is an accurate model for near and far distances when $\lambda \ll z$. The wavenumber is $k = 2\pi/\lambda$.

To derive a point spread function (PSF) p of the optical element, we model a plane wave, representing a single point at optical infinity, propagating along the optical axis through the element before reaching a sensor at a distance z from the element as

$$p_\lambda\left(x, y\right) \propto \left| \mathcal{F}\left\{ A\left(x', y'\right) e^{i\phi\left(x', y'\right)} e^{i\frac{\pi}{\lambda z}\left(x'^2 + y'^2\right)} \right\} \right|^2. \tag{4}$$

Note that the PSF is wavelength dependent and that its intensity is given by the squared magnitude of the complex-valued wave field $\left|U_\lambda\left(x, y, z\right)\right|^2$. Detailed derivations of these formulations can be found in the textbook by Goodman [2017].

3.1.2 From PSF to Image. Although Equation 4 only considers the image formation of a single point source with a specific wavelength, we can extend this model to account for a natural scene. In this context, the point spread functions from different scene points add incoherently, i.e. in intensity, on the sensor. Assuming that the paraxial approximation is valid, this image formation is a shift-invariant convolution of the image and the PSF; consequently, off-axis aberrations like coma or chromatic off-axis aberration are neglected. We further account for the wavelength sensitivity κ_c of the sensor for each of the three color channels R, G, B

$$I_c\left(x, y\right) = \int \left(I_\lambda * p_\lambda\right)\left(x, y\right) \kappa_c\left(\lambda\right) d\lambda, \tag{5}$$

where the index c denotes the color channel. Note that this image formation is physically accurate for spatially incoherent scenes that are observed at a distance far from the optical element, because each scene point is effectively modeled as a plane wave. The benefit of this simplification is that the image formation becomes computationally very efficient, but the limitation is that only scenes at optical infinity are modeled correctly.

To account for this limitation, we can lift the restriction imposed by plane waves and model point sources at different distances by modeling spherical waves with the appropriate curvature using the term ϕ_d in Equation 2. In this case, the image formation ceases to be a shift-invariant convolution. Nevertheless, for sufficiently large distances between the scene and the optical element, a shift-invariant image formation may still be a good approximation [Antipa et al. 2016].

3.1.3 Sensor. The image I_c formed on the sensor is integrated over the sensor pixels and corrupted by noise, yielding a measurement y_c given by

$$y_c = S(I_c) + \eta, \tag{6}$$

where S is the pixel integration and sampling operator and $\eta \sim \mathcal{N}(0, \sigma^2)$ is Gaussian read noise.

3.1.4 Reconstruction. The final stage of the proposed model is image reconstruction. We reconstruct the estimate $\widetilde{I_c}$ of the source image I_λ for RGB wavelengths by solving the Tikhonov regularized least-squares problem

$$\min_{\{\widetilde{I}\}} \quad \|y_c - S(p_c * \widetilde{I_c})\|_2^2 + \gamma \|\widetilde{I_c}\|_2^2, \tag{7}$$

where $\widetilde{I}$ is the unknown variable, p_c is the PSF p_λ integrated over the wavelength sensitivity κ_c in a narrow band around RGB wavelengths, and $\gamma > 0$ is an (optimized) regularization parameter.

When the PSF discretization size matches the sensor pixel size, the pixel integration operator S is the identity. We can then solve problem (7) in closed form with Wiener filtering under the simplifying assumption of circular boundary conditions, which we approximate by symmetric padding of y_c. The Wiener filtering operation is given

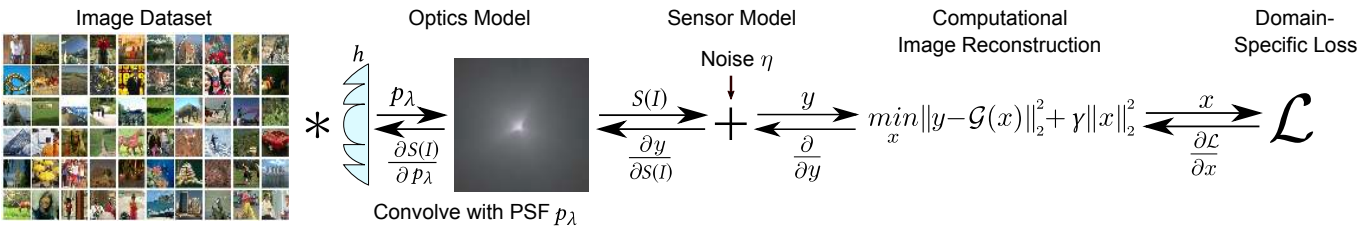

Fig. 3. The proposed framework is an end-to-end differentiable pipeline architecture. In each forward pass, the PSF p of the current optical element is simulated using the proposed wave-based image formation module (see Figure 2). The simulated PSF is then convolved with a batch of images, and noise η is added to account for sensor read noise. A post-processing algorithm solves the Tikhonov-regularized least-squares problem for image reconstruction with the image formation model $\mathcal{G}$. Finally, a differentiable loss $\mathcal{L}$, such as mean squared error with respect to the ground-truth image, is defined on the reconstructed images. In the backward pass, the error is backpropagated all the way back to the PSF simulation, through the diffraction, to the optical element itself.

by

$$\widetilde{I}_c = \mathcal{F}^{-1}\left\{\frac{\overline{p}_c^*}{|\overline{p}_c|^2 + \gamma}\,\mathcal{F}\left\{y_c\right\}\right\}, \tag{8}$$

where $\overline{p}$ is the optical transfer function, or Discrete Fourier Transform of p, and multiplication and division are element-wise. When the pixel size is larger than the discretization of the PSF, we use a fixed number of conjugate gradient steps to solve Equation 7.

We use Tikhonov-regularized least-squares as a reconstruction method because it is a computationally efficient approach that directly parameterizes the reconstruction with the PSF, so that updates to the PSF during optimization are immediately propagated to the reconstruction method. In future work, more powerful, differentiable deconvolution methods [Diamond et al. 2017b] could be used in this model or image reconstructions could potentially be replaced with a high-level convolutional neural network for image analysis [Diamond et al. 2017a].

3.2 End-to-end Optimization Framework

We develop a framework for optimizing a computational camera with an optical element in TensorFlow using stochastic gradient methods. We express each stage of the model described in the previous subsection as a differentiable module. The optical height map h is an optimization variable. The optical element size, sensor pixel size, propagation distance z, and sensor read noise level are hyperparameters.

Models are optimized on a dataset of RGB images (defining I_λ). The optimization variables in the optical element and reconstruction method are optimized with respect to the expected mean-squared error loss

$$\mathcal{L}(\widetilde{I}_c, I_\lambda) = \sum_{c\in\{R,G,B\}} \|\widetilde{I}_c - I_c\|_2^2, \tag{9}$$

over the dataset. The full optimization pipeline is shown in Figure 3. The approach easily generalizes to other data fidelity losses as well as losses on semantic content of images, such as cross-entropy loss for image classification.

The key challenges in developing the proposed optimization framework were to satisfy manufacturing constraints, finding stable optimization algorithms, and fitting models within memory limits. We discuss insights that allowed us to tackle these challenges in the following.

Feasibility Constraints of Fabrication. The constraints of a specific manufacturing process impose constraints on the optimization. In the context of this paper, we consider two different options: manufacturing (i) diffractive optical elements (DOEs) with photolithography and (ii) refractive optical elements with diamond turning. DOEs are flat lenses that rely on the small phase delays induced by ultra-small features to diffract light in a desired manner. The manufacturing tolerances of the photolithography process and required discretization of the fabricated height map, however, may make it challenging to precisely fabricate a continuous height map at sub-wavelength resolution. Diamond turning refractive elements is a common process, but the constrained toolpath of the mill makes this process best suited for fabricating smooth, continuous surfaces. To address these constraints, we can use a basis representation of the height map h and optimize for the respective basis coefficients. A natural choice would be a representation using Zernike polynomials, which is given by $h = \sum_{j=1}^{n} \alpha_j Z_j$, where Z_j is the j$^{\text{th}}$ Zernike polynomial in Noll notation and α_j is the corresponding coefficient [Born and Wolf 1999]. Similarly, we can represent the height map h as a sum of weighted Fourier basis functions, which would implicitly constrain the resulting phase plate to lower-frequency components. We chose a Zernike representation for all refractive optical elements fabricated with diamond turning and the Fourier representation for the DOEs. Lastly, during optimization, we add random uniform noise in the range ±20 nm to the height map before simulating the PSF, to increase robustness to manufacturing imperfections. Both diamond turning and photolithography manufacturing processes are discussed in detail in Section 6 and in the Supplemental Information.

Optimization algorithms. We experimented with several variants of stochastic gradient descent (SGD)-based algorithms, which have recently been used extensively in the deep learning community. We found that for the Zernike base representation, the sensitivity of the optical element to single coefficients varied significantly. This led to instabilities when optimizing using SGD variants with static step sizes, such as vanilla SGD or SGD with Nesterov acceleration [Nesterov 1983]. We consequently found that the Adadelta optimization algorithm [Zeiler 2012] was most stable in experiments using the Zernike representation, since it dynamically adapts step sizes for each parameter during optimization. The direct height map parameterization and the Fourier basis representation were accessible to various optimizers.

Fig. 4. Optimizing a simple optical element. We optimize a lens to focus an image at a distance of $z = 25$ mm without image reconstruction for a wavelength of $\lambda = 550$ nm. As expected, the optimized height profile (left) converges to a structure that looks similar to a Fresnel lens. As seen in the point spread function (PSF, center inset), the element is capable of focusing the green channel but residual blur remains in the other color channels because a single optical element cannot be achromatic. After deconvolving each color channel with their respective PSF, we obtain the result shown (right).

Memory constraints. Fitting models within memory limits was difficult due to the fine discretization of the optical elements necessary for high fidelity Fresnel propagation. Typical models discretize 5 mm apertures in a 2 µm grid, yielding a height map with over 6 million elements. The many intermediate values computed for Fresnel propagation must all be stored for backpropagation, and the total memory is again multiplied by the number of depths and wavelengths considered. An essential insight for keeping memory usage tractable was to take advantage of the stochastic optimization scheme by randomly binning the depths and wavelengths present in each image into a smaller set. For instance, a given image can be placed at a random depth drawn from a distribution over the depths of interest. Over the whole optimization process, the SGD algorithm will sample a wide variety of depths.

Simple lens. To validate the proposed framework, we first consider the simple test case of designing a lens to focus light of a single wavelength from objects at optical infinity onto a sensor. We remove the reconstruction method, so the measured sensor image y_c is the final output of the model fed into the loss function. The standard diffractive design for this imaging task is a Fresnel lens with 2π phase wrapping, which produces a PSF as close as possible to a Dirac peak [Smith 2007].

We optimize over the dataset proposed by Jegou et al. [2008], which consists of a selection of high-resolution images, mainly of outdoor scenes. We consider a lens-to-sensor distance of 25 mm and an aperture of 5 mm, leading to an f-number of 5. We model a sensor with a resolution of $1,248 \times 1,248$ and a pixel size of 4 µm. The sensor read noise has a standard deviation drawn from a uniform distribution between 0.001 and 0.02. The optimization variable is the (discretized) optical height map h with no constraints. We optimize the pipeline using SGD with Nesterov momentum with a step size of 5×10^{-3} and the momentum parameter set to 0.5. Intuitively, the optimized profile converges to an optical element that closely matches a Fresnel lens (Figure 4, left).

4 ACHROMATIC EXTENDED DEPTH OF FIELD

The idea of extended depth of field (EDOF) imaging includes the design of an approximately depth-invariant PSF for one wavelength and then applying a shift-invariant deconvolution to the recorded images to obtain an all-in-focus image [Cossairt and Nayar 2010; Cossairt et al. 2010; Dowski and Cathey 1995]. Achromatic imaging, on the other hand, aims at designing a spectrally-invariant PSF for one depth to achieve full-color imaging with a single optical element, which would otherwise not be able to focus light of different wavelengths at the same depth [Peng et al. 2016]. With the proposed end-to-end optimization approach, we can formulate a wide variety of cost functions for a computational camera. In this section, we report results for combining achromatic imaging using a single optical element and extended depth of field.

To apply our framework to achromatic EDOF imaging, we discretize the full spectrum to three wavelengths (460, 550, and 640 nm) and five depths (0, 0.5, 1, 1.5, 2 diopters or, similarly, ∞, 2, 1, 0.67, and 0.5 m). We then sample images from a dataset of 1,244 high-resolution images [Jegou et al. 2008] and assign one of the five depth settings randomly while considering all three wavelengths simultaneously. After applying the wave-optics image formation model, we also include Gaussian read noise in the measurements, with a random magnitude between 0.1–1%. We found this noise level to be appropriate for our experimental sensor, though we also explored a noise level of 2% with similar results. Finally, the image is reconstructed by deconvolution with a PSF at a single depth, where the specific depth is an optimization variable. This encourages convergence to a depth-invariant, invertible PSF. We note that while for spherical incoming waves (scene not at infinity), the image formation model ceases to be a shift-invariant convolution, this is still a good approximation for the considered scene depths [Antipa et al. 2016]. For this application, we optimize optical elements for both a diffractive Fourier coefficient parameterization and a Zernike basis representation (cf. Equation 1). The simulated optical setup matches that of our prototype (see Section 6) and entails an aperture diameter of 5 mm and a propagation distance of $z = 35.5$ mm between optical element and sensor. The exact parameters of the optical model, the parameters of the Zernike and Fourier coefficient parameterizations as well as the optimization hyperparameters can be found in Appendix A.1. The resulting profiles of these optimized optical elements are shown in Figure 5 (top).

4.1 Evaluation in Simulation

We show a qualitative comparison of results achieved with the proposed optics and alternative approaches in Figure 5. The baseline for all comparisons is the Fresnel lens (e.g. [Smith 2007]), which focuses light of one wavelength to one specific depth. Although compound lenses, for example used in commercial cameras, can focus a range of wavelengths to the same depth, this is generally not possible with a single refractive or diffractive optical element. We chose parameters to allow the Fresnel lens to focus one of our three target wavelengths, i.e. 550 nm, to a distance of 1 m. A multi-focal lens creates an approximately depth-invariant PSF over the full target depth range of 0–2 D for a wavelength of 550 nm. Similarly, a cubic phase profile [Dowski and Cathey 1995] combined with the phase profile of a standard focusing lens also provides an approximately depth-invariant PSF over the target range. A diffractive achromat [Peng et al. 2016], providing some EDOF due to its design as an achromatic

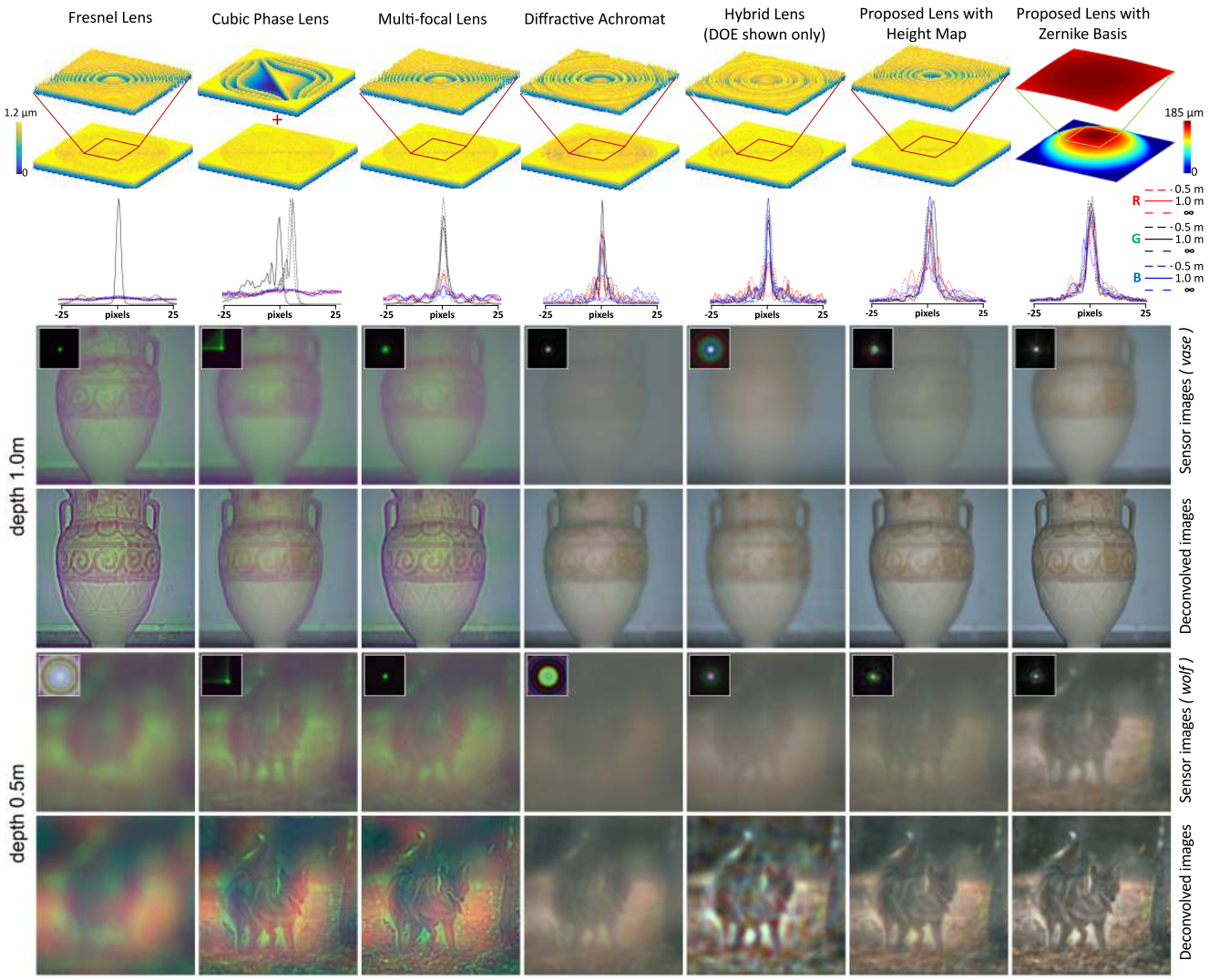

Fig. 5. Evaluation of achromatic extended depth of field imaging in simulation. We compare the performance of a Fresnel lens optimized for one of the target wavelengths (left), a cubic phase plate combined with the phase of a focusing lens (second column), a multi-focal lens optimized for all five target depths (third column), a diffractive achromat optimized for all three target wavelengths at 1 m (fourth column), a hybrid diffractive-refractive element (fifth column), optics optimized end-to-end with Wiener deconvolution with a height map parameterization (sixth column), and optics optimized end-to-end with Wiener deconvolution with a Zernike basis representation (right). All methods produce flat height profiles (top row), except for the Zernike representation and the refractive-diffractive hybrid lens, for which we show only the diffractive element. The resulting point spread functions are shown for all target wavelengths and depths (second row). Sensor images and PSFs of two example scenes (rows 3 and 5) exhibit strong chromatic aberrations for the Fresnel lens, the multi-focal lens, and the cubic phase plate, whereas the diffractive achromat, the hybrid element and the proposed optics mitigate wavelength-dependent image variations much better. After deconvolution (rows 4 and 6), residual color artifacts are observed for all lenses but the diffractive achromat and the proposed lenses. The PSF of the hybrid element is not readily invertible, so that no sharp image can be achieved without the introduction of deconvolution artifacts. Both the diffractive achromat and the hybrid element fail to capture scene content at varying depths adequately, whereas our approach allows for high-fidelity reconstructions at various wavelengths and scene depths. Additional results can be found in the Supplemental Information.

element, is optimized to create a wavelength-invariant PSF at one depth, here 1 m. Lastly, we optimize a diffractive-refractive hybrid element for achromatic EDOF, following the design specifications proposed by Flores et al. [2004]. All images are then deconvolved with the PSF of the respective imaging system at a depth of 1 m. We note that all baselines are examples of optical elements that were optimized separately from the post-processing step, while our approach is the only one that jointly optimizes post-processing and optical element. As discussed in Section 3, the proposed end-to-end optimization approach optimizes the height map of an optical element parameterized either with Fourier coefficients or Zernike basis functions. We show the output of both options in Figure 5 (top

Fig. 6. Experimental results for achromatic extended depth of field imaging with DOEs. Left: an image captured through a Fresnel lens. Center, right: an image captured through a diffractive optical element (DOE) optimized with the proposed framework without (center) and with (right) Wiener deconvolution as a post-processing step. Over the scene depth range of 0.5 m to 2 m, the Fresnel lens displays significant out-of-focus blur at the extremes of the depth range and further suffers from significant chromatic aberrations. In contrast, the proposed optical element succeeds in focusing a wide range of depths for all three color channels, leading to an all-in-focus scene with little chromatic aberrations.

row) along with 1D slices of corresponding PSFs at all depths for all target wavelengths (second row).

The simulations in Figure 5 show that the sensor images of the Fresnel lens, the cubic phase lens, and the multi-focal lens suffer from severe chromatic aberrations. While the hybrid element performs better, color variance in the PSF across depths still leads to chromatic artifacts after deconvolution at the extreme end of the depth range. Furthermore, the PSF of the hybrid element is not readily invertible, and no sharp image can thus be formed without the introduction of deconvolution artifacts even at a distance of 1 m. This may be caused by the extreme depth range we consider, as Flores et al. [2004] reported results on a much smaller depth range. The diffractive achromat produces an image that can be recovered with a reasonable quality at one depth, i.e. 1 m, but it fails to generalize across depths. With the proposed lenses and reconstruction, we show the best results across wavelengths and depths for both parameterizations of the phase profile. Notably, the optimized refractive element performs best, as the diffractive element still suffers from some degree of residual chromatic aberrations that are typical for diffraction-based elements. Nevertheless, the diffractive element also significantly outperforms baseline approaches. Generally, the PSFs found by the optimization were (approximately) invertible but not necessarily focused. This is intuitive because the cost function we optimized only considers the final reconstructed image, not the PSFs or images formed on the sensor.

In addition to these qualitative results, we also show a detailed quantitative evaluation implemented on the 100 test images of the BSDS500 dataset [Martin et al. 2001] in Table 1. For this purpose, we average the mean squared error (MSE) of reconstructions over all five target depths and three target wavelengths for each individual scene. Table 1 outlines peak signal-to-noise ratios (PSNR) computed on the average MSE per scene for the two examples of Figure 5 and also the average PSNR of all 100 test scenes. All simulated sensor images include 0.2% Gaussian noise and the deconvolution for all methods is performed with the respective wavelength-dependent PSF calibrated at 1 m. Our method significantly outperforms other approaches for the task of simultaneous depth and wavelength-invariant imaging with a single optical element. Similarly, our method outperforms all baselines for intermediate depths (0.585, 0.835, 1.5 and 3 m) that were not explicitly optimized for, as well as for a higher noise level of 2%. Additional sensor images, deconvolved results, PSFs for all settings, as well as the quantitative results for the higher noise level of 2% and intermediate depths are shown in the Supplemental Information.

Table 1. Quantitative comparison of achromatic extended depth of field. We report PSNR values in dB for a Fresnel lens, a multi-focal lens (MFL), the cubic phase plate combined with the phase of a focusing lens (CPP), a diffractive achromat (DA), the diffractive-refractive hybrid lens (Hybrid), and the proposed method optimized for a height map h or for a Zernike basis representation Z. The proposed method outperforms the best alternative approach by a large margin, on average 4.1 dB for the Zernike basis representation.

	Scene 1 (vase)	Scene 2 (wolf)	Avg. 100 scenes
Fresnel	23.87	20.40	17.95
MFL	24.22	20.73	18.32
CPP	24.38	20.70	18.33
DA	26.74	22.31	20.20
Hybrid	25.50	21.07	18.92
End-to-end with h	**29.57**	**24.70**	**22.69**
End-to-end with Z	**31.13**	**26.40**	**24.30**

4.2 Experimental Results

To demonstrate the practical viability of the proposed, end-to-end optimized optics, we fabricate an optimized diffractive optical element and also a reference Fresnel lens with 4-layer photolithography and another refractive element using diamond turning. Figure 1 compares an image taken through a conventional refractive lens to an image taken through the optimized diamond-turned element, which is subsequently deconvolved using Wiener deconvolution. The scene covers a depth range from 0.5 m (the elephant sculpture) to 1.5 m (the textbook). Both optical elements share the same f-number. The benefits of the proposed optical design are clearly visible, with the whole scene perfectly sharp, where the conventional lens displays significant blurring for all but its focus plane of 1 m.

Figure 6 compares an image captured through a diffractive Fresnel lens with the optimized DOE. Again, the scene depth ranges from 0.5 m (the cherries) to 2.0 m. The optimized DOE succeeds in displaying the complete image in-focus, where the Fresnel lens

displays significant blur at the extremes of the depth range. Furthermore, the Fresnel lens displays significant chromatic aberrations (being only optimized for a single wavelength), while the proposed design was optimized for three wavelengths of the visible spectrum and thus significantly reduces chromatic aberrations. We also show a demonstration of video-rate processing of the optimized refractive element over a continuous depth range in the supplemental video.

5 SNAPSHOT SUPER-RESOLUTION IMAGING

Optical zoom in cameras is typically achieved by increasing the distance between lens and sensor to magnify the recorded image. In many scenarios, such as cell phone cameras, it may be impossible to further increase the device form factor of the camera, making optical zoom impractical. Digital zoom is an alternative, but this approach uses either simple upsampling methods in software or deep learning-based single-image super-resolution methods to hallucinate image details that were not actually recorded. Another approach is the addition of extra camera modules or a single sensor with multiple subapertures. However, this makes the camera more costly, bulky, or decreases the diffraction-limited resolution. We ask whether it may be possible to design a single lens that maintains a constant physical footprint while facilitating image super-resolution. In this scenario, the standard non-zoom lens is swapped laterally with an optimized ultrathin lens. Post-processing then achieves computational zoom, while the sensor-lens distance remains constant.

To investigate this possibility, we use the proposed end-to-end optimization framework to design a diffractive lens that optically encodes information in a sensor image that may make it possible to recover a 2× super-resolved image by solving Equation 7. We assume a monochromatic sensor and single wavelength images with $\lambda = 550\,\text{nm}$, located at optical infinity. While the framework generalizes to multiple wavelengths in a straight-forward manner, we choose a monochromatic sensor to simplify analysis of experimental results. We sample images from a dataset of 30 images [Xu et al. 2014], converted from RGB to monochrome. The simulated optical setup matches that of our prototype (see Section 6) and entails an aperture diameter of 5 mm and a propagation distance of $z = 35.5\,\text{mm}$ between optical element and sensor. We use 5 iterations of the conjugate gradient method for the reconstruction (which is not closed form due to pixel integration, cf. Eq. 6). We fix the regularization $\gamma = 2 \times 10^{-4}$ and optimize for a Fourier coefficient representation of the height map h. Detailed optimization parameters can be found in Appendix A.2.

5.1 Evaluation in simulation

The optimized optical height profile along with a photograph of the fabricated element casting caustic patterns that resemble the PSF are shown in Figure 7. We observe that the optimized element resembles the shape of three separate lenses, which is verified by the caustic patterns consisting of several strong peaks. Such a PSF, when convolved with an image, results in several optical copies of the input image on the sensor, as seen in Figure 9 (top right) for an experimentally captured image. It is intuitive that the optimization could result in such a PSF, because it is well-known from early work on super-resolution [Ben-Ezra et al. 2004] that multiple

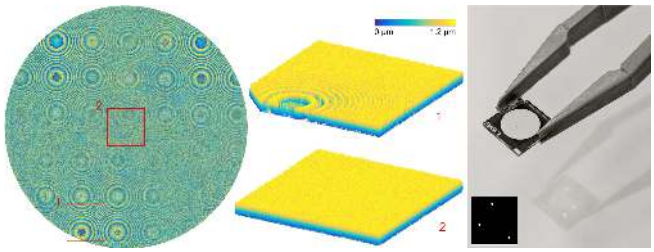

Fig. 7. Optimized phase profile of diffractive optical element (left) for super-resolution. The phase profile appears to create three interleaved lens profiles; one of these is shown in the closeup 3D rendering (center top). Note that the middle portion of the DOE is not perfectly flat (center bottom). These variations are minute and can be considered noise in the optimization that gets blurred out in the point spread function. We also show a photograph of the manufactured DOE with a caustic that shows the three peaks, which create multiple image copies on the sensor (right, with PSF inset, blurred to increase peak radius for better visibility).

sub-pixel-shifted images can be used to recover a super-resolved image. The proposed end-to-end optimization achieves the same result in a single shot by multiplexing these sub-pixel-shifted image copies on the sensor. We note that while the optimized phase profile is intuitive, the optimization determined shape, number, and placement of these sub-PSFs without supervision, in a manner that would minimize interference of the three sub-PSFs as well as respecting manufacturing constraints, while optimizing for the fidelity of the final, reconstructed image. These would traditionally be hand-crafted parameters in a large design space.

Even the simple conjugate gradient based reconstruction method in the framework is capable of recovering the target image very well, as shown in Table 2 and Figure 8. Note that our simulations assume that only a part of the sensor image is used for conventional imaging, whereas the PSF optimization has sufficient degrees of freedom to spread the recorded signal out over a larger area on the sensor, thus creating non-overlapping copies of the image. The image resolution in these simulations are also limited by the pixel size of the sensor, not the diffraction limit. Sub-diffraction limited imaging is not possible with the proposed approach; although image copies could also be created with a diffraction-limited optical system, the respective copies would contain the exact same image information. Similar to previous super-resolution methods [Ben-Ezra et al. 2004], our method relies on aliasing in these optical copies, which is created by sub-pixel shifts.

For the simulated result shown in Figure 8, we downsample the target image by a factor of 2 in each dimension using area interpolation. Bicubic upsampling is not capable of restoring fine image details. We also apply a state-of-the art deep learning approach for single-image super-resolution [Lai et al. 2017] to the low-resolution image. This method hallucinates high-frequency details, but it is not capable of adequately restoring image details, which our method is able to recover. We verify that these results generalize to other images with the extensive quantitative evaluation summarized in Table 2. Values for peak signal-to-noise ratio (PSNR) and structural similarity index (SSIM) of all approaches other than ours are adopted

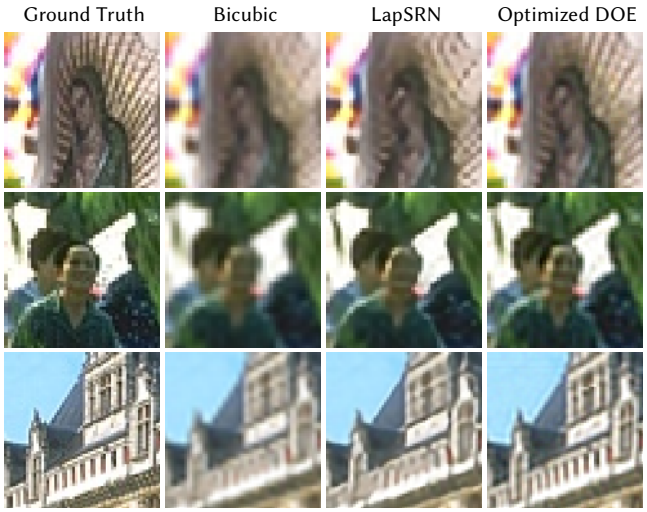

Fig. 8. Qualitative comparison of 2× super-resolution imaging in simulation. We downsample a target image (left) by a factor of 2 in each dimension. Neither bicubic upsampling nor the state-of-the-art single-image super-resolution method proposed by Lai et al. [2017] (LapSRN) achieve a high image quality for the reconstruction. Similar to other single-image super-resolution methods, LapSRN hallucinates high-frequency content and, in the process, introduces aliasing. Our approach optimizes a diffractive optical element (DOE) that, together with a simple conjugate gradient solver, recovers a super-resolved image with a high quality.

Table 2. Quantitative comparison of 2× super-resolution methods. Peak signal-to-noise ratio (PSNR) and structural similarity index (SSIM) values comparing a number of different approaches for several datasets are adopted from Lai et al. [2017]. Our end-to-end optics and image reconstruction approach outperforms all of these methods in SSIM. Deep digital zoom methods achieve higher PSNR for the Urban100 dataset by exploiting image priors to interpolate textured regions.

	Set14 PSNR/SSIM	BSDS100 PSNR/SSIM	Urban100 PSNR/SSIM
Bicubic	30.34/0.870	29.56/0.844	26.88/0.841
A+ [Timofte et al. 2014]	32.40/0.906	31.22/0.887	29.23/0.894
SRCNN [Dong et al. 2016b]	32.29/0.903	31.36/0.888	29.52/0.895
FSRCNN [Dong et al. 2016a]	32.73/0.909	31.51/0.891	29.87/0.901
SelfExSR [Huang et al. 2015]	32.44/0.906	31.18/0.886	29.54/0.897
RFL [Schulter et al. 2015]	32.36/0.905	31.16/0.885	29.13/0.891
SCN [Wang et al. 2015]	32.42/0.904	31.24/0.884	29.50/0.896
VDSR [Kim et al. 2016a]	32.97/0.913	31.90/0.896	**30.77**/0.914
DRCN [Kim et al. 2016b]	32.98/0.913	31.85/0.894	30.76/0.913
LapSRN [Lai et al. 2017]	33.08/0.913	31.80/0.895	30.41/0.910
Optimized DOE	**33.88/0.933**	**32.84 /0.933**	30.39/**0.919**

from Lai et al. [2017]. We also show additional qualitative comparisons of these methods, as well as an additional comparison with a naive multiplexing baseline, in the Supplemental Information. The optimized diffractive optical system outperforms the state-of-the-art in SSIM on all datasets, and on PSNR for the Set14 and BSDS100 datasets. Deep learning-based super-resolution approaches achieve higher PSNR on the Urban100 dataset, which features many images with regular texture that can be easily interpolated using learned image priors.

5.2 Experimental Results

To verify that we can achieve super-resolution in practice, we fabricated the diffractive optical element (DOE) shown in Figure 7 using photolithography. Details on the fabrication process are found in Section 6 and in the Supplement. An image captured through this DOE is shown in Figure 9 (top right). This photograph clearly shows the individual image copies, as well as a slight haze due to imperfections in the fabrication and limited diffraction efficiency of the DOE, which limits the contrast of our measurements compared to a conventional lens (Figure 9, top left). Due to the fact that our method relies on aliasing in these copies, which are effects that are smaller than a single pixel, any deviations of the optical PSF from the simulated PSF make this a very challenging experiment. Yet, as shown in Figure 9, the proposed DOE and reconstruction restores image detail that is lost with digital zoom techniques such as bicubic upsampling or a state-of-the-art LapSRN super-resolution network.

6 FABRICATING CUSTOM OPTICS

We have employed two fabrication methods to manufacture our optimized lenses. The resulting products are noted as *diffractive optical elements* and *freeform lenses* in the following.

Fabricating Diffractive Optical Elements (DOEs). We repeatedly apply 4 rounds (i.e. 16-phase-level structures) of photolithography and reactive iron etching techniques [Morgan et al. 2004] to fabricate optics that are optimized with Fourier coefficient parameterizations of the discretized height map. The substrate is a 0.5 mm thick, fused silica wafer with a refractive index of 1.459 at the principle wavelength of 550 nm. We use 2π phase modulation to wrap the height map to a uniform maximum height. Please refer to the supplement for details of the fabrication procedure as well as microscope images of our lenses. We note that this kind of micro-fabrication technique involves repeated procedures at micrometer level alignment accuracy. Such precise alignment makes the fabrication procedure relatively complex, but opens up a large design space by allowing many small features and high-frequency detail on the DOE [Peng et al. 2016].

Fabricating Refractive Freeform Lenses. In addition, we use a CNC machining platform that supports 5-axis single point diamond turning (Nanotech 350FG), similar to [Damberg and Heidrich 2015; Schwartzburg et al. 2014; Wu et al. 2013], to fabricate lenses that are parameterized using a Zernike polynomial basis. The substrate is polymethyl methacrylate (PMMA) with a refractive index of 1.493 at the principle wavelength of 550 nm. The downside of freeform lenses relative to DOEs manufactured with photolithography is a larger form factor and less design freedom due to the need for a smooth surface to mill, but the upside is that the freeform lens has less inherent color dispersion, much higher light efficiency, and much lower production cost.

System Integration. We use two sensors, one chromatic sensor (FLIR GS3-23S6C-C) that has 1,920×1,200 pixels with a pixel pitch of 5.86 µm, and one monochromatic sensor (FLIR GS3-91S6M-C) that has 3,376×2,704 pixels with a pixel pitch of 3.69 µm in our experiments. The former is used for AEDOF imaging (Section 4) while the latter is used for snapshot super-resolution imaging (Section 5). The focal distance of the AEDOF and superresolution optical elements is

35.5 mm with an aperture size of 5 mm. This yields an f-number of $f/7.1$. As discussed in Section 3, our non-blind image reconstruction method requires us to calibrate the PSF in advance. Accordingly, we use a white LED light source with a 35 µm pinhole attached in front to calibrate the PSFs of our custom lenses.

7 DISCUSSION

In summary, we demonstrate that the co-design of camera optics and reconstruction is feasible using a fully-differentiable pipeline that includes a wave optics model for the image formation and a regularized least-squares image reconstruction. We explore different parameterizations for the optimized optical elements, including height maps or Zernike polynomials, and we verify the principle of operation of our optimized elements with fabricated optical elements. We demonstrate state-of-the-art results of the proposed framework for applications in achromatic extended depth of field and snapshot super-resolution imaging.

The primary benefit of the proposed methodology is that optical elements can be jointly optimized with post-processing algorithms to minimize differentiable losses that only consider the performance of the joint model, with no optimization of intermediate steps such as point spread function engineering. The proposed framework takes advantage of hardware, tools and algorithms developed in the deep learning community, allowing easy customization and profiting from hardware and software progress in that field.

There are several limitations of our current approach. First, the algorithms we used for image reconstructions are simple. We either use a Wiener filter or a truncated set of conjugate gradient iterations. While this is adequate for the presented applications, it seems insufficient for more advanced tasks, such as high dynamic range imaging, depth from defocus, image classification, semantic segmentation, etc. Second, we currently approximate the tolerances of the respective fabrication methods by simply optimizing the optical elements in one of two basis representations, a Fourier basis or a Zernike basis. These tolerances should be better quantified and modeled as constraints in the optimization. Third, even though our image formation model includes depth variation, it does not handle occlusion boundaries between objects at different depths appropriately. We optimize the extended depth of field phase profiles by randomly sampling depth values of the planar input images during optimization. A more precise image formation model, such as ray tracing, may further improve results and enable new applications.

Future Work. In future work, we would like to explore more sophisticated differentiable reconstruction methods, such as convolutional neural networks. Advanced computational camera designs, for example tailored to higher-level vision tasks, likely require deep algorithmic frameworks. We would also like to explore otherwise inaccessible parts of the camera design spectrum, for example by minimizing the device form factor or overcoming fundamental limits of conventional cameras. Finally, designing multiple sensors jointly could open new research directions as the proposed end-to-end framework naturally extends to such systems with an appropriate image formation model.

Fig. 9. Experimental result for 2× super-resolution. We capture an image with a conventional 35.5 mm lens (top left) and with the fabricated diffractive optical element (DOE) obtained with the proposed end-to-end optimization framework (top right, with PSF inset, blurred to increase peak radius for better visibility. Three peaks multiplexing the image are clearly visible.). As expected, the image recorded with the DOE creates multiple aliased copies of the scene. Bicubic upsampling of the scene (row 2) does not restore the fine details observed in the target image (row 5). Applying Lai et al.'s [2017] method (row 3) or other single-image super-resolution methods allows for high-frequency details to be hallucinated but these may be incorrect (i.e., hand, sword, ornaments). The proposed lens and reconstruction restores some of the true image detail better, even with the limited optical quality offered by our custom DOEs.

8 CONCLUSION

End-to-end optimization is an emerging design paradigm for computational cameras. Although the idea of jointly optimizing camera optics, sensing, and algorithms has been at the heart of the computational photography community for years, leveraging modern tools of the deep learning community for this problem opens new research directions and has the potential to make unprecedented camera designs possible. With our work, we demonstrate the efficacy of the end-to-end computational camera design paradigm for addressing challenging imaging problems. These results encourage the exploration of more advanced end-to-end frameworks, for example using convolutional neural networks, in future computational cameras.

ACKNOWLEDGMENTS

The authors would like to thank Xu Liu and Liang Xu from the State Key Lab of Modern Optical Instrumentation, Zhejiang University, and the KAUST Visual Computing Center for support in designing and prototyping of DOEs. This project was supported by an NSF CAREER award (IIS 1553333), an NSF Graduate Research Fellowship (DGE-114747), a Sloan Fellowship, a Terman Faculty Fellowship, a Stanford Graduate Fellowship, the Intel Compressive Sensing Alliance, and by the KAUST Office of Sponsored Research through the Visual Computing Center CCF grant.

REFERENCES

N. Antipa, S. Necula, R. Ng, and L. Waller. 2016. Single-shot diffuser-encoded light field imaging. In *Proc. IEEE ICCP*. 1–11.

M. S. Asif, A. Ayremlou, A. Sankaranarayanan, A. Veeraraghavan, and R. G. Baraniuk. 2017. FlatCam: Thin, Lensless Cameras Using Coded Aperture and Computation. *IEEE Trans. Computational Imaging* 3, 3 (2017), 384–397.

M. Ben-Ezra, A. Zomet, and S.K. Nayar. 2004. Jitter Camera: High Resolution Video from a Low Resolution Detector. In *Proc. CVPR*. 135–142.

Max Born and Emil Wolf. 1999. *Principles of Optics: Electromagnetic Theory of Propagation, Interference and Diffraction of Light* (7 ed.). Cambridge University Press.

D. J. Brady, M. E. Gehm, R. A. Stack, D. L. Marks, D. S. Kittle, D. R. Golish, E. M. Vera, and S. D. Feller. 2012. Multiscale Gigapixel Photography. *Nature* 486 (2012), 386–389.

Ayan Chakrabarti. 2016. Learning sensor multiplexing design through back-propagation. In *Advances in Neural Information Processing Systems*. 3081–3089.

Wei Ting Chen, Alexander Y Zhu, Vyshakh Sanjeev, Mohammadreza Khorasaninejad, Zhujun Shi, Eric Lee, and Federico Capasso. 2018. A broadband achromatic metalens for focusing and imaging in the visible. *Nature nanotechnology* (2018), 1.

Shane Colburn, Alan Zhan, and Arka Majumdar. 2018. Metasurface optics for full-color computational imaging. *Science Advances* 4, 2 (2018), eaar2114.

O. Cossairt, D. Miau, and S.K. Nayar. 2011. Gigapixel Computational Imaging. In *Proc. ICCP*.

Oliver Cossairt and Shree Nayar. 2010. Spectral focal sweep: Extended depth of field from chromatic aberrations. In *Proc. ICCP*. 1–8.

Oliver Cossairt, Changyin Zhou, and Shree Nayar. 2010. Diffusion Coded Photography for Extended Depth of Field. *ACM Trans. Graph. (SIGGRAPH)* 29, 4 (2010), 31:1–31:10.

Gerwin Damberg and Wolfgang Heidrich. 2015. Efficient freeform lens optimization for computational caustic displays. *Optics Express* 23, 8 (2015), 10224–10232.

Paul E. Debevec and Jitendra Malik. 1997. Recovering High Dynamic Range Radiance Maps from Photographs. In *ACM SIGGRAPH*. 369–378.

Steven Diamond, Vincent Sitzmann, Stephen Boyd, Gordon Wetzstein, and Felix Heide. 2017a. Dirty pixels: Optimizing image classification architectures for raw sensor data. *arXiv preprint arXiv:1701.06487* (2017).

Steven Diamond, Vincent Sitzmann, Felix Heide, and Gordon Wetzstein. 2017b. Unrolled Optimization with Deep Priors. (2017). arXiv:1705.08041

C. Dong, C. Loy, K. He, and X. Tang. 2016b. Image Super-Resolution Using Deep Convolutional Networks. *IEEE Trans. PAMI* 38, 2 (2016), 295–307.

Chao Dong, Chen Change Loy, and Xiaoou Tang. 2016a. Accelerating the super-resolution convolutional neural network. In *Proceedings of the European Conference on Computer Vision*. 391–407.

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

Gabriel Eilertsen, Joel Kronander, Gyorgy Denes, Rafal Mantiuk, and Jonas Unger. 2017. HDR image reconstruction from a single exposure using deep CNNs. *ACM Trans. Graph. (SIGGRAPH Asia)* 36, 6 (2017).

Angel Flores, Michael R Wang, and Jame J Yang. 2004. Achromatic hybrid refractive-diffractive lens with extended depth of focus. *Applied optics* 43, 30 (2004), 5618–5630.

Joseph Goodman. 2017. *Introduction to Fourier Optics* (4 ed.). W.H. Freeman.

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. Graph. (SIGGRAPH)* 35, 6 (2016), 192:1–192:12.

Felix Heide, Qiang Fu, Yifan Peng, and Wolfgang Heidrich. 2016. Encoded diffractive optics for full-spectrum computational imaging. *Scientific Reports* 6, 33543 (2016).

Jia-Bin Huang, Abhishek Singh, and Narendra Ahuja. 2015. Single image super-resolution from transformed self-exemplars. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 5197–5206.

Michael Iliadis, Leonidas Spinoulas, and Aggelos K Katsaggelos. 2016. Deepbinarymask: Learning a binary mask for video compressive sensing. *arXiv preprint arXiv:1607.03343* (2016).

Herve Jegou, Matthijs Douze, and Cordelia Schmid. 2008. Hamming Embedding and Weak Geometric Consistency for Large Scale Image Search. In *Proc. ECCV*. 304–317.

Nima Khademi Kalantari and Ravi Ramamoorthi. 2017. Deep High Dynamic Range Imaging of Dynamic Scenes. *ACM Trans. Graph. (SIGGRAPH)* 36, 4 (2017), 144:1–144:12.

Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. 2016a. Accurate image super-resolution using very deep convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1646–1654.

Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. 2016b. Deeply-recursive convolutional network for image super-resolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1637–1645.

Wei-Sheng Lai, Jia-Bin Huang, Narendra Ahuja, and Ming-Hsuan Yang. 2017. Deep Laplacian Pyramid Networks for Fast and Accurate Super-Resolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Michael L. Land and Dan-Eric Nielsson. 2002. *Animal Eyes*. Oxford University Press.

Anat Levin, Rob Fergus, Frédo Durand, and William T. Freeman. 2007. Image and Depth from a Conventional Camera with a Coded Aperture. *ACM Trans. Graph. (SIGGRAPH)* 26, 3 (2007).

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. *ACM Trans. Graph. (SIGGRAPH)* 28, 3 (2009), 97:1–97:14.

Zhiqiang Liu. 2007. Diffractive lens with extended depth of focus and its applications. (2007).

Mann, Picard, S. Mann, and R. W. Picard. 1995. On Being 'undigital' With Digital Cameras: Extending Dynamic Range By Combining Differently Exposed Pictures. In *Proceedings of IS&T*. 442–448.

D. Martin, C. Fowlkes, D. Tal, and J. Malik. 2001. A Database of Human Segmented Natural Images and its Application to Evaluating Segmentation Algorithms and Measuring Ecological Statistics. In *ICCV*, Vol. 2. 416–423.

Kshitij Marwah, Gordon Wetzstein, Yosuke Bando, and Ramesh Raskar. 2013. Compressive Light Field Photography Using Overcomplete Dictionaries and Optimized Projections. *ACM Trans. Graph. (SIGGRAPH)* 32, 4 (2013), 46:1–46:12.

Brian Morgan, Christopher M Waits, John Krizmanic, and Reza Ghodssi. 2004. Development of a deep silicon phase Fresnel lens using gray-scale lithography and deep reactive ion etching. *Journal of microelectromechanical systems* 13, 1 (2004), 113–120.

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

Y. Nesterov. 1983. A method of solving a convex programming problem with convergence rate $O(1/k^2)$. *Soviet Mathematics Doklady* 27 (1983), 372–376.

Ren Ng, Marc Levoy, Mathieu Bredif, Gene Duval, Mark Horowitz, and Pat Hanrahan. 2005. Light Field Photography with a Hand-Held Plenoptic Camera. Tech Report CSTR 2005-02.

Sri Rama Prasanna Pavani, Michael A. Thompson, Julie S. Biteen, Samuel J. Lord, Na Liu, Robert J. Twieg, Rafael Piestun, and W. E. Moerner. 2009. Three-dimensional, single-molecule fluorescence imaging beyond the diffraction limit by using a double-helix point spread function. 106, 9 (2009), 2995–2999.

Yifan Peng, Qiang Fu, Felix Heide, and Wolfgang Heidrich. 2016. The Diffractive Achromat Full Spectrum Computational Imaging with Diffractive Optics. *ACM Trans. Graph. (SIGGRAPH)* 35, 4 (2016), 31:1–31:11.

Ramesh Raskar, Amit Agrawal, and Jack Tumblin. 2006. Coded Exposure Photography: Motion Deblurring Using Fluttered Shutter. *ACM Trans. Graph. (SIGGRAPH)* 25, 3 (2006), 795–804.

Erik Reinhard, Greg Ward, Sumanta Pattanaik, and Paul Debevec. 2005. *High Dynamic Range Imaging: Acquisition, Display, and Image-Based Lighting*. Morgan Kaufmann.

Mushfiqur Rouf, Rafał Mantiuk, Wolfgang Heidrich, Matthew Trentacoste, and Cheryl Lau. 2011. Glare encoding of high dynamic range images. In *Proc. CVPR*. 289–296.

Samuel Schulter, Christian Leistner, and Horst Bischof. 2015. Fast and accurate image upscaling with super-resolution forests. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3791–3799.
Yuliy Schwartzburg, Romain Testuz, Andrea Tagliasacchi, and Mark Pauly. 2014. High-contrast computational caustic design. *ACM Trans. Graph. (SIGGRAPH)* 33, 4 (2014), 74.
Yoav Shechtman, Steffen J. Sahl, Adam S. Backer, and W. E. Moerner. 2014. Optimal Point Spread Function Design for 3D Imaging. *Phys. Rev. Lett.* 113 (2014), 133902. Issue 13.
Wenzhe Shi, Jose Caballero, Ferenc Huszar, Johannes Totz, Andrew P. Aitken, Rob Bishop, Daniel Rueckert, and Zehan Wang. 2016. Real-Time Single Image and Video Super-Resolution Using an Efficient Sub-Pixel Convolutional Neural Network. In *Proc. CVPR*.
Yichang Shih, Brian Guenter, and Neel Joshi. 2012. Image enhancement using calibrated lens simulations. In *European Conference on Computer Vision*. Springer, 42–56.
Warren J. Smith. 2007. *Modern Optical Engineering* (4 ed.). McGraw-Hill.
Pratul P. Srinivasan, Tongzhou Wang, Ashwin Sreelal, Ravi Ramamoorthi, and Ren Ng. 2017. Learning to Synthesize a 4D RGBD Light Field from a Single Image. In *Proc. IEEE ICCV*.
Libin Sun, Neel Joshi, Brian Guenter, and James Hays. 2015. Lens Factory: Automatic Lens Generation Using Off-the-shelf Components. *CoRR* abs/1506.08956 (2015). arXiv:1506.08956 http://arxiv.org/abs/1506.08956
Radu Timofte, Vincent De Smet, and Luc Van Gool. 2014. A+: Adjusted anchored neighborhood regression for fast super-resolution. In *Proceedings of Asian Conference on Computer Vision*. 111–126.
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. Graph. (SIGGRAPH)* 26, 3 (2007).
Ashwin Wagadarikar, Renu John, Rebecca Willett, and David Brady. 2008. Single disperser design for coded aperture snapshot spectral imaging. *OSA Appl. Opt.* 47, 10 (2008), B44–B51.
Ting-Chun Wang, Jun-Yan Zhu, Nima Khademi Kalantari, Alexei A. Efros, and Ravi Ramamoorthi. 2017. Light Field Video Capture Using a Learning-Based Hybrid Imaging System. *ACM Trans. Graph. (SIGGRAPH)* 36, 4 (2017).
Zhaowen Wang, Ding Liu, Jianchao Yang, Wei Han, and Thomas Huang. 2015. Deep networks for image super-resolution with sparse prior. In *Proceedings of the IEEE International Conference on Computer Vision*. 370–378.
Gordon Wetzstein, Ivo Ihrke, Douglas Lanman, and Wolfgang Heidrich. 2011. Computational Plenoptic Imaging. *Computer Graphics Forum* 30, 8 (2011), 2397–2426.
Rengmao Wu, Liang Xu, Peng Liu, Yaqin Zhang, Zhenrong Zheng, Haifeng Li, and Xu Liu. 2013. Freeform illumination design: a nonlinear boundary problem for the elliptic Monge–Ampére equation. *Optics letters* 38, 2 (2013), 229–231.
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.
Jie Yang, Jiafu Wang, Mingde Feng, Yongfeng Li, Xinhua Wang, Xiaoyang Zhou, Tiejun Cui, and Shaobo Qu. 2017. Achromatic flat focusing lens based on dispersion engineering of spoof surface plasmon polaritons. *Applied Physics Letters* 110, 20 (2017), 203507.
Matthew D Zeiler. 2012. ADADELTA: an adaptive learning rate method. *arXiv preprint arXiv:1212.5701* (2012).
Jinsong Zhang and Jean-François Lalonde. 2017. Learning High Dynamic Range from Outdoor Panoramas. In *Proc. IEEE ICCV*.
C. Zhou, S. Lin, and S. K. Nayar. 2009. Coded Aperture Pairs for Depth from Defocus. In *Proc. ICCV*.
C. Zhou and S. Nayar. 2009. What are good apertures for defocus deblurring?. In *Proc. IEEE ICCP*. 1–8.

A OPTIMIZATION PARAMETERS

In the following, we describe the exact parameters used to optimize the optical elements described in this work.

A.1 Achromatic Extended Depth of Field

A.1.1 Zernike parameterization. We simulate a sensor with a pixel size of 3.69 µm and a resolution of $1,356 \times 1,356$ pixels. We consider the first 350 Zernike coefficients in Noll notation. The optical element is initialized as a standard collimator lens, with the fourth Zernike coefficient (the defocus term) initialized such that the lens has a focal length of 35.5 mm. The optical element is discretized with a 3.69 µm feature size on a $1,356 \times 1,356$ grid. In the learning phase, which includes optimizing the optical element and finding the optimal regularization parameter γ for the reconstruction, we use the Adadelta optimizer with a step size of 1. The optimization phase is run for 8 epochs, which takes approximately 6 hours on a single NVIDIA TITAN X Pascal GPU.

A.1.2 Fourier coefficient parameterization. We simulate a sensor with a pixel size of 4 µm and a resolution of $1,248 \times 1,248$ pixels. We set the 37.5% highest frequencies to zero as a smoothness prior. All Fourier coefficients are initialized to zero at the beginning of the optimization. The optical element is discretized with a 2 µm feature size on a $2,496 \times 2,496$ grid. In the learning phase, which includes optimizing the optical element and finding the optimal regularization parameter γ for the reconstruction, we use a step size of 5×10^{-1} with a stochastic gradient descent solver using a Nesterov momentum term of 0.5. The optimization phase is run for 64 epochs, which takes approximately 4 hours on a single NVIDIA TITAN X Pascal GPU.

A.2 Snapshot Super-Resolution Imaging

We simulate a sensor with a pixel size of 3.69 µm and a resolution of $1,356 \times 1,356$ pixels, which we downsample by a factor of 2× using area interpolation to simulate larger pixels. The optical element is discretized with a 2.46 µm feature size on a $2,034 \times 2,034$ grid. The PSF is subsequently downsampled to a 3.69 µm resolution following Fresnel propagation for computational efficiency. We set the 37.5% highest frequency Fourier coefficients to zero as a smoothness prior. All Fourier coefficients are initialized to zero at the beginning of the optimization. We use Adadelta with a step size of 1 to optimize the model. The optimization phase is run for 50,000 iterations with batch size 1, which takes approximately 20 hours on a single NVIDIA TITAN X Pascal GPU.