

# Collaborative On-Sensor Array Cameras

JIPENG SUN, Princeton University, USA

KAIXUAN WEI, KAUST, Saudi Arabia

THOMAS EBOLI, Université Paris-Saclay, France

CONGLI WANG and CHENG ZHENG, Princeton University, USA

ZHIHAO ZHOU and ARKA MAJUMDAR, University of Washington, USA

WOLFGANG HEIDRICH, KAUST, Saudi Arabia

FELIX HEIDE, Princeton University, USA

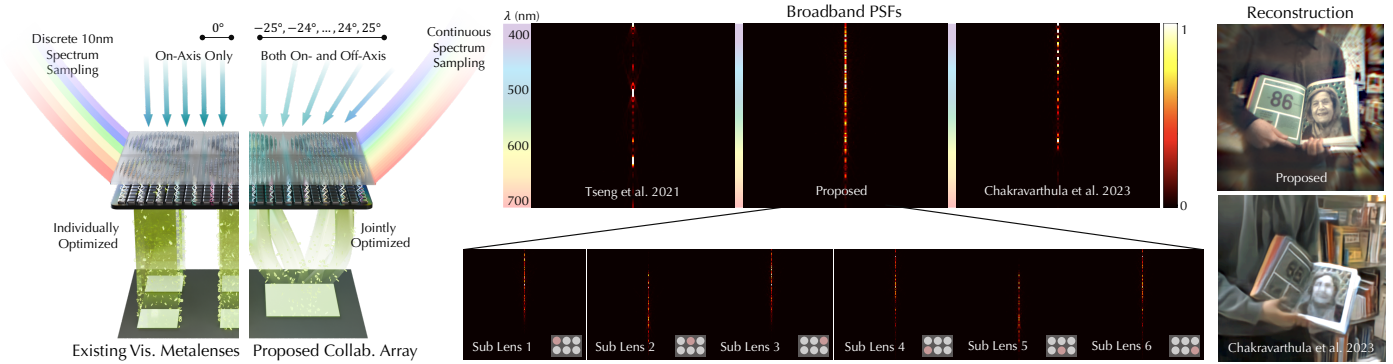


Fig. 1. We designed a collaborative metasurface array imager of  $2 \times 3$  metalenses (left) positioned 3.6 mm from the sensor plane that enables broadband focusing (see 1-D PSF sections alongside spectrum in the middle) across the full  $50^\circ \times 50^\circ$  field of view, including both axial and off-axis incident angles. The design collaboratively optimizes 100 million nanophotonic parameters through a neural structure-to-phase proxy and a distributed meta-optics training framework incorporating a parallax-aware spatially-varying method for accurate, hallucination-free image reconstruction (right). In reconstructing a book containing '86' and portrait, our collaborative metalens array accurately captures both elements, while existing flat metalens arrays [Chakravarthula et al. 2023] produces hallucinated reconstructions that more closely resemble '65'.

Modern nanofabrication techniques have enabled us to manipulate the wavefront of light with sub-wavelength-scale structures, offering the potential to replace bulky refractive surfaces in conventional optics with ultrathin metasurfaces. In theory, arrays of nanoposts provide unprecedented control over manipulating the wavefront in terms of phase, polarization, and amplitude at the nanometer resolution. A line of recent work successfully investigates flat computational cameras that replace compound lenses with a single metalens or an array of metasurfaces a few millimeters from the sensor. However, due to the inherent wavelength dependence of metalenses, in practice, these cameras do not match their refractive counterparts in image quality for broadband imaging, and may even suffer from hallucinations when relying on generative reconstruction methods.

In this work, we investigate a collaborative array of metasurface elements that are jointly learned to perform broadband imaging. To this end, we learn

a nanophotonics array with 100-million nanoposts that is end-to-end jointly optimized over the full visible spectrum—a design task that existing inverse design methods or learning approaches cannot support due to memory and compute limitations. We introduce a distributed meta-optics learning method to tackle this challenge. This allows us to optimize a large parameter array along with a learned metaatom proxy and a non-generative reconstruction method that is parallax-aware and noise-aware. The proposed camera performs favorably in simulation and in all experimental tests irrespective of the scene illumination spectrum.

CCS Concepts: • **Computing methodologies** → **Computational photography**; **Massively parallel algorithms**.

Additional Key Words and Phrases: Computational Optics, Computational Imaging, Distributed Optimization

## ACM Reference Format:

Jipeng Sun, Kaixuan Wei, Thomas Eboli, Congli Wang, Cheng Zheng, Zhihao Zhou, Arka Majumdar, Wolfgang Heidrich, and Felix Heide. 2025. Collaborative On-Sensor Array Cameras. *ACM Trans. Graph.* 44, 4 (August 2025), 18 pages. <https://doi.org/10.1145/3731200>

## 1 INTRODUCTION

The aperture of an optical system fundamentally limits its optical resolution. As such, large aperture optics are essential for applications across diverse domains, including photography, robotics, remote sensing, mobile smartphones, health, and scientific imaging.

Authors' addresses: Jipeng Sun, [jipeng.sun@princeton.edu](mailto:jipeng.sun@princeton.edu), Princeton University, USA; Kaixuan Wei, [kaixuan.wei@kaust.edu.sa](mailto:kaixuan.wei@kaust.edu.sa), KAUST, Saudi Arabia; Thomas Eboli, [thomas.eboli12@gmail.com](mailto:thomas.eboli12@gmail.com), Université Paris-Saclay, France; Congli Wang, [congli.wang@princeton.edu](mailto:congli.wang@princeton.edu), Princeton University, USA; Cheng Zheng, [chengzh@princeton.edu](mailto:chengzh@princeton.edu), Princeton University, USA; Zhihao Zhou, [zzhou99@uw.edu](mailto:zzhou99@uw.edu), University of Washington, USA; Arka Majumdar, [arka@uw.edu](mailto:arka@uw.edu), University of Washington, USA; Wolfgang Heidrich, [wolfgang.heidrich@kaust.edu.sa](mailto:wolfgang.heidrich@kaust.edu.sa), KAUST, Saudi Arabia; Felix Heide, [fheide@princeton.edu](mailto:fheide@princeton.edu), Princeton University, USA..

© 2025 Copyright held by the owner/author(s).

This is the author's version of the work. It is posted here for your personal use. Not for redistribution. The definitive Version of Record was published in *ACM Transactions on Graphics*, <https://doi.org/10.1145/3731200>.

Large optics generally result in large weight; for example, the weight of the lenses in the Hubble space telescope (aperture of  $\sim 2$  m) is on the order of  $\sim 1000$  kg. Enabled by modern nanofabrication techniques, metaoptics [Khorasaninejad and Capasso 2017] have been proposed as a potential alternative to refractive surfaces. Metaoptics consist of quasi-periodic arrays of sub-wavelength scatterers that offer access to a design space that is six orders of magnitude larger than conventional spherical optics. These degrees of freedom offer unprecedented control over the modulation of the incoming wavefront, with the potential to drastically reduce the size and weight of conventional large-aperture optics.

However, traversing this larger design space to find large-aperture optics that offer substantially smaller footprint without compromising image quality has proven challenging, even with recent learned [Chakravarthula et al. 2023; Frösch et al. 2025; Tseng et al. 2021a] design approaches. This is because diffractive optical elements (DOEs) and metasurfaces are fundamentally wavelength-dependent chromatic optics, which makes it challenging to minimize chromatic aberration while achieving high resolution and large field of view (FOV). Although techniques like dispersion engineering have been explored to address chromatic aberrations [Chen et al. 2018; Wang et al. 2018], the limited optical bandwidth and the aperture size lead to poor imaging performance under broadband illumination [Presutti and Monticone 2020]. A recent line of work has explored learning methods for optimizing metalenses or arrays of metasurfaces using proxy forward models [Chakravarthula et al. 2023; Tseng et al. 2021a]. To date, the most successful meta-optics imager for broadband imaging is inverse designed via a differentiable broadband image formation model [Chakravarthula et al. 2023]. Despite the improvements in image quality, the design limitations of the heuristic wedge phase superimposition results in compromised color fidelity and severe distortions in spatial frequencies, requiring a generative model for image recovery with the risk of hallucinations. As a direct result of the large size of the parameter space, only the center lens profile could be learned while the peripheral lenses were directly made by superimposing the center lens profile with a certain wedge phase to heuristically expand FOV. Consequently, measurements from this existing camera (especially the peripheral lenses) suffer from severe degradations in spatial frequencies and color fidelity. Although significant strides in meta-optics imaging have been made in both optics and computational imaging communities, the image quality of ultra-thin metalens cameras is still far from being competitive with traditional refractive optics.

In this work, we propose a flat camera composed of an array of metalenses which are jointly and collaboratively designed for broadband imaging. To achieve this, we learn a nanophotonic array with 100-million-nanoposts that is end-to-end optimized over the full visible spectrum – a design task existing inverse design methods or learning approaches cannot support due to memory and compute limitations. Our design is motivated by the observations that 1) a single metalens can possess favorable focusing properties within a narrow band, and 2) a sequence of measurements that are optically encoded in distinct ways can be collaboratively merged to generate high-quality results. We find that the collaborative array features per-lens optical encodings, each of which is tailored to distinct yet

complementary segments of the color spectrum. To design the array we develop a distributed, large-scale meta-optics optimization framework. We learn the array optics with a novel learned nano-structure-to-broadband-phase neural proxy. To recover images, we introduce a collaborative parallax-aware joint reconstruction model that does not rely on generative priors. We fabricate the designed metasurface array and build a prototype camera system to experimentally validate the proposed design method. We confirm that the collaborative metalens array camera consistently outperforms the existing meta-optics in terms of image quality across the entire visible spectrum, while maintaining a flat optical system on the sensor cover glass.

Our contributions are as follows:

- We introduce a collaborative on-sensor array camera, where each lens learns to capture complementary portions of the spectrum. When combined, these measurements enable high-fidelity reconstruction across the full visible spectrum range.
- We propose a learned nano-structure-to-broadband-phase neural proxy for accurate forward simulation, and a distributed large-scale meta-optics optimization method.
- We devise a parallax-aware, spatially-variant reconstruction algorithm tailored to our jointly designed lens array, deliberately avoiding reliance on generative priors.
- We evaluate our approach in simulations and in real indoor and outdoor scenes with a camera prototype, confirming favorable broadband imaging performance.

*Limitations.* Compared to conventional cameras that employ larger refractive lens assemblies to record high-quality images without computationally heavy reconstruction, the proposed flat-camera does require computational recovery to produce an image. Moreover, as a research prototype rather than a commercial product, the current camera prototype retains the thick sensor cover glass that limits inner-baffle height, precludes further focal-length reduction, and mandates a non-contiguous sublens arrangement to avoid crosstalk. These compromises enlarge the required sensor size and limit the overall light-collection efficiency.

## 2 RELATED WORK

*Flat Computational Cameras.* Designing a flat camera system with high imaging quality and low cost has been an open challenge that a large body of work in computational imaging addresses. As optical aberrations inherently limit the quality of flat optical systems, existing approaches share the same philosophy that trades off camera form factor with image quality and computational complexity [Antipa et al. 2018; Asif et al. 2016; Boominathan et al. 2016, 2020; Heide et al. 2016, 2013; Khan et al. 2020; Peng et al. 2015, 2016, 2019; Venkataraman et al. 2013]. Early attempts at flat computational camera designs compress the compound camera optics into a single refractive element (e.g., plano-convex or biconvex lenses [Kingslake and Johnson 2009]), and they correct the geometric and chromatic aberrations post-capture using variational optimization [Heide et al. 2013; Schuler et al. 2011]. To further shrink the form factor, researchers have explored DOEs building atop diffraction manipulation of light beyond refraction, including single DOE elements



(e.g., Fresnel lens [Heide et al. 2016; Peng et al. 2015, 2016]), lens-less coded apertures [Asif et al. 2016; Boominathan et al. 2020], diffusers [Kuo et al. 2017; Monakhova et al. 2020], and Fresnel zone apertures [Wu et al. 2020]. The compound cameras can achieve aberration-corrected high-quality imaging under broadband illumination, but still entails a long back-focal distance of more than 10 mm which has made *truly compact* camera design challenging. The DOE cameras, albeit allowing for ultra-thin cameras of a few millimeters in height, achieve moderate imaging quality due to the highly ill-posed nature of image reconstruction from coded measurements with global support. As such, unfortunately, no existing methods achieve flat form factors with high imaging quality.

*Flat Metasurface Optics.* Unlike conventional optical elements that rely on gradual phase shifts accumulated during light propagation to shape light beams, metasurfaces – imparting abrupt phase changes with subwavelength nanostructures – have recently been investigated for imaging [Arbabi and Faraon 2023; Dorrah and Capasso 2022; Khorasaninejad et al. 2016; Kildishev et al. 2013; Lin et al. 2014; Yu and Capasso 2014; Yu et al. 2011]. Existing metasurfaces offer flat optics, however, but suffer from severe monochromatic and chromatic aberrations when acting as imagers (i.e., metalens) [Khorasaninejad and Capasso 2017]. Although dispersion-engineered achromatic metalenses have been reported [Khorasaninejad et al. 2017; Shrestha et al. 2018; Wang et al. 2018], they are fundamentally limited to aperture sizes of tens of microns [Presutti and Monticone 2020]. To enable larger aperture size while suppressing aberrations, researchers have designed meta-optic systems together with conventional deconvolution methods [Colburn et al. 2018] and learned recovery methods [Tseng et al. 2021a] for multi-wavelength color imaging. Recent metalens camera designs have explored deep neural network reconstruction for analytical hyperbolic metalenses [Dong et al. 2024], aperture-stop optimization for three discrete wavelengths [Park et al. 2025], and folded metalens to reduce system thickness [Kim et al. 2024]. While achieving high-quality imaging results, these metasurface cameras are not collaborative designs for broadband imaging, and the number of optimized spectral bands is two orders of magnitude smaller than ours. The work most related to ours is [Chakravarthula et al. 2023], which utilizes an inverse-designed on-sensor metasurface array camera with a learned optimization algorithm for large FOV broadband imaging. While previous work used metalens arrays to expand the field of view at the cost of image quality, our approach jointly designs six complementary metalenses that collaborate for high-fidelity imaging, with a compact 3.6 mm height.

*Differentiable Optics.* While conventional camera design hand-engineers the optics in isolation for compartmentalized metrics (e.g., root-mean-square spot size), a body of work in computational imaging leverages the modern differentiable framework to jointly optimize optics and reconstruction algorithms for downstream tasks. Successful methods that follow this paradigm learn optics and/or sensors for color imaging [Chakrabarti 2016; Na et al. 2024; Sun et al. 2021; Wang et al. 2022; Yang et al. 2024a,b], microscopy [Horstmeyer et al. 2017; Kellman et al. 2019; Nehme et al. 2020; Shechtman et al. 2016], depth imaging [Chang and Wetzstein 2019; Ghanekar et al. 2024; Haim et al. 2018; Wu et al. 2019], super-resolution and extended

depth of field [Sitzmann et al. 2018; Tan et al. 2021], time-of-flight imaging [Chugunov et al. 2021; Marco et al. 2017; Su et al. 2018], high-dynamic range imaging [Martel et al. 2020; Metzler et al. 2020; Shi et al. 2024a; Sun et al. 2020], compressive sensing [Iliadis et al. 2020; Yoshida et al. 2018], active-stereo imaging [Baek and Heide 2021], hyperspectral imaging [Baek et al. 2021; Fu et al. 2020; Shi et al. 2024b], and computer vision tasks [Souza et al. 2024; Tseng et al. 2021b; Wei et al. 2024]. Our approach models differentiable optics for metasurfaces and jointly optimizes the metasurface design and the reconstruction algorithm end-to-end for general broadband imaging.

### 3 COLLABORATIVE NANOPHOTONIC ARRAY IMAGING

Our collaborative metalens camera is described in design and reconstruction phase. We first describe the design hinging on a differentiable spatially-varying image formation model with neural metasurface proxy (Section 3.1), and a joint optimization of metalens array phase profiles and PSFs for multi-image deconvolution (Section 3.2) through distributed metasurface learning (Section 3.4). For experimental reconstruction, we devise an array image alignment approach, followed by the same multi-image deconvolution method combined with a learned denoiser (Section 3.3). Implementation and training details are provided in Section 3.5. Figure 2 illustrates our collaborative nanophotonic imaging system design and reconstruction architecture.

In what follows, continuous variables (such as wavelength  $\lambda$ ) are denoted in subscript, while discrete (categorical) variables (such as the metalens index  $k$  and color channel  $c$ ) are denoted in superscript.

#### 3.1 Spatially-varying Image Formation Model

*Overview of Image Formation Model.* We start by deriving the formation model of an image by a single meta-optical element. We target broadband imaging in the visible range with wavelength  $\lambda$  from 400 nm to 700 nm. The optical imaging process at a certain wavelength can be generally modeled as the locally-varying convolution of the focal scene image  $u_\lambda$  with a point spread function (PSF)  $p_\lambda$ . This PSF kernel is a function of the wavelength-dependent phase profiles  $\Phi_\lambda$  in the lens plane<sup>1</sup>

$$p_\lambda : \Phi_\lambda \mapsto p_\lambda(\Phi_\lambda), \quad (1)$$

which can be derived by a free-space wave propagation of a wavefront (plane wave modulated by the optical phase delay) propagated to the sensor plane [Goodman 2005]. Our design leverages a  $2 \times 3$  array of  $K = 6$  metalenses, and thus has as many PSFs  $p_\lambda^k$  and related phases  $\Phi_\lambda^k$ . We chose the number of lenses to align with the sensor aspect ratio; however, the method itself is agnostic to lens number and can accommodate any array configuration.

The  $k$ -th image  $v^{(k,c)}$  at a certain color channel  $c \in \{R, G, B\}$  focused by the optics onto the sensor plane is formed by integrating PSFs over the visible spectrum weighted by the spectral sensitivities of the on-sensor color filters

$$v^{(k,c)} = \int_{\lambda \in \Lambda} \kappa_\lambda^c \left( u_\lambda \otimes p_\lambda^k \right) d\lambda + \epsilon^{(k,c)}, \quad (2)$$

<sup>1</sup>Note we omit the spatial/angle dependency of the PSF in the formula here for simplicity. The spatial/angle-dependent PSF computation will be detailed later.

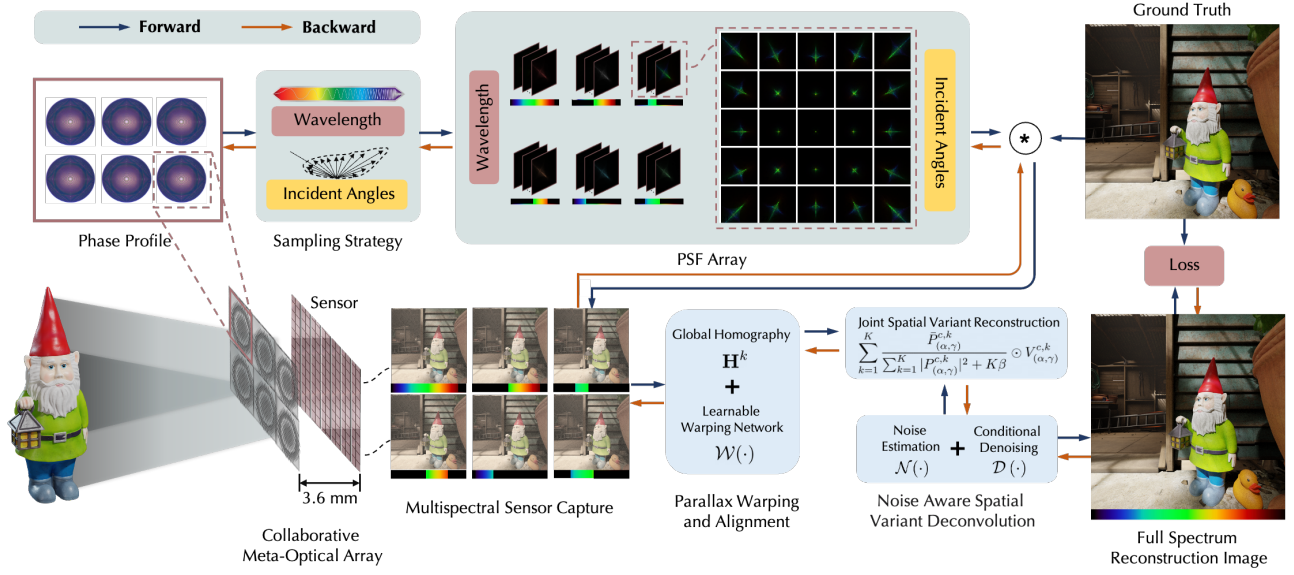


Fig. 2. **Collaborative Nanophotonic Imaging.** We propose a  $2 \times 3$  collaborative metalens array camera wherein individual metalenses in the array capture a part of the whole visible spectrum (the spectrum under each sublens image here is for illustration). After two-step alignment to compensate for parallax, we reconstruct the full visible spectrum at the resolution of the input sub-images, with a multi-image variant of Wiener deconvolution collaboratively combining wavelength spectra across the six collaborative measurements. The whole pipeline is fully differentiable, thereby enabling end-to-end optimization of the collaborative nanophotonic imaging system.

where  $\otimes$  denotes a locally-varying convolution as the PSF kernel varies across the FOV. Here,  $\kappa_{\lambda}^c$  is the camera spectral response function (a.k.a. camera sensitivity [Jiang et al. 2013; Solomatov and Akkaynak 2023]) for each color filter  $c$  atop sensor. As such,  $\epsilon^{k,c}$  denotes the random (shot and read) noise in the sensor measurements, which is modeled as a heteroscedastic Gaussian distribution [Foi et al. 2008].

Due to the nonlinear phase responses of the nano-pillars across different wavelengths, the associated PSFs can vary substantially among spectral bands [Chen et al. 2018]. Consequently, Eq. (2) cannot be approximated by a simple convolution in the RGB image space (see *Supplementary Material* for details). To accurately capture the underlying physics, we therefore optimize our optics model in the hyperspectral domain.

Given the smooth variation of the PSFs across different incident angles [Lohmann and Paris 1965], the locally-varying convolution  $\otimes$  can be approximated as spatially-uniform convolutions  $*$  of  $M \times N$  image patches localized in  $u_{\lambda}$ , with the corresponding PSFs  $p_{\lambda}^{(k,m,n)}$  computed at specific incident angles [Eboli et al. 2022; Hirsch et al. 2010; Schuler et al. 2011]. The image  $u_{\lambda} \otimes p_{\lambda}^k$  is then formed by blending these patches with a fusion windowing function  $w$  (e.g., a Hann window [Blackman and Tukey 1958]) to suppress border artifacts. This can be formulated as

$$(u_{\lambda} \otimes p_{\lambda}^k) = \sum_{m=1}^M \sum_{n=1}^N w^{(m,n)} (u_{\lambda} * p_{\lambda}^{(k,m,n)}), \quad (3)$$

where  $(m, n)$  indicates the  $(m, n)$ -th entry of an array of varying incident angles. The window  $w^{(m,n)}$  is non-zero only for the interval of the corresponding patch  $(m, n)$  in  $u_{\lambda}$ . In what follows, we

elucidate Eq. (1) in detail, which includes the metalens model and the angle-dependent PSF computation.

**Metalens Model Parameterization.** Each individual metalens is composed of silicon nitride nanoposts with a height of 1000 nm atop a fused silica substrate. These nanostructures with optimizable width  $D$  between 100-300 nm are distributed on a regular grid with  $P = 350$  nm distance, resulting in a grid of pillars with tunable width (which determines the “duty cycle”  $d = \frac{D}{P}$  between 0 to 1) [Chakravarthula et al. 2023; Tseng et al. 2021a]. Instead of working on this low-level representation that entails the computationally demanding full-wave solver, we parameterize each metalens as a radially symmetric phase profile

$$\Phi_{\lambda_0}(x, y) = \varphi(r), \quad r = \sqrt{x^2 + y^2}, \quad (4)$$

where  $(x, y)$  represents the spatial coordinates in the lens plane, and  $\varphi^k$  is a 1D radial vector of size  $L$  (i.e., the radius of the metalens), which serves as the optimization variable in our design. Note the above phase function is only defined for a specific nominal wavelength  $\lambda_0$ . We select  $\lambda_0 = 658$  nm as the nominal wavelength for our fabricated structure, as we empirically found that, there exists a one-to-one (instead of one-to-many) mapping relationship between the structure and the phase response, within a tolerance for the wavelength.

To fully characterize the broadband behavior of the metalens while maintaining the differentiability of the pipeline, we utilize two proxy functions in sequence to determine the phase profiles of the metalens across the entire visible spectrum of interest in lieu of the rigorous yet computationally expensive full-wave methods,

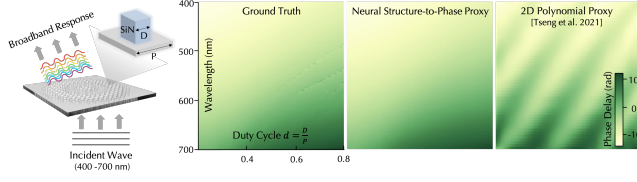


Fig. 3. **Comparison between proposed neural structure-to-phase proxy and polynomial proxy [2021a].** A proxy is needed to model the broadband response of a metalens given plane wave illumination across visible wavelengths. For a single nano-pillar shown on the left, its induced phase delay is determined by both the incident wavelength (vertical axis) and the duty cycle of the structure (horizontal axis). We plot the ground truth of the unwrapped phase delay as well as the approximated values from two differentiable proxy models. In contrast to the existing polynomial proxies [2021a], the proposed neural proxy accurately models the nonlinear relationship among nano-structure duty cycle, target wavelength, and broadband phase delay.

such as the finite-difference-time domain method (FDTD) [Kunz and Luebbers 1993; Taflov et al. 2005].

The first proxy is a precomputed polynomial that computes the metasurface geometry (duty cycle  $d$  for each nanopost) from the phase at nominal wavelength  $\lambda_0$  as  $d(\phi_{\lambda_0}) = \sum_{o=0}^O a_o (\phi_{\lambda_0} / (2\pi))^o$ , where we find that a fifth-order polynomial  $O = 5$  is sufficient to fit the phase-to-structure mapping data in experiments.

Due to the complex, nonlinear relationship among the nano-structure duty cycle, target wavelength, and the broadband phase delay—together with the high-precision requirements of broadband phase mapping—polynomial based proxy functions used in previous literature [Chakravarthula et al. 2023; Tseng et al. 2021a] fail to converge within a 1% error threshold, see Fig. 3. Therefore, we utilize a neural network proxy  $N_{proxy}$  as the second structure-to-phase mapping proxy to compute the phases at other target wavelengths  $\phi_t$  given the metasurface duty cycle  $d$ .

$$\Phi_{\lambda_t}(d) = N_{proxy}(d, \lambda_t) \bmod 2\pi. \quad (5)$$

The neural network comprises three fully connected layers with ReLU activation. Its training dataset is generated by FDTD simulations at 2 nm increments for structure sizes ranging from 80 nm to 280 nm, and at 1 nm increments across the visible spectrum. With the two proxy functions introduced, we now build a differentiable computational graph  $\Phi_{\lambda_0} \mapsto \Phi_{\lambda_t}$  from the phase profile at nominal wavelength  $\lambda_0$  to other target visible wavelengths  $\lambda_t$ . See the *Supplementary Material* for full details of the FDTD simulation.

**Angular-dependent PSF Model.** With the metalens model, we can now derive the angle-dependent PSFs of the whole optical system. Given an incident wave direction (altitude  $\alpha$  and azimuth  $\gamma$  angles), the incident plane wave phase  $\phi_{\lambda}^{(\alpha, \gamma)}$  at wavelength  $\lambda$  is given by

$$\phi_{\lambda}^{(\alpha, \gamma)}(x, y) = \frac{2\pi}{\lambda} \frac{x \tan \alpha + y \tan \gamma}{\sqrt{\tan^2 \alpha + \tan^2 \gamma + 1}}. \quad (6)$$

As explained before, the PSF smoothly varies across the FOV and thus we split the whole FOV into  $M \times N$  patches, each of which corresponds to a certain angular position ( $\alpha[m, n]$ ,  $\gamma[m, n]$ ). In practice, we discretize the FOV into a  $7 \times 7$  array, with angles spanning

equally from  $-30^\circ$  to  $30^\circ$ . As such, the modulated phase function  $\Phi_{\lambda}^{(k, m, n)}$  (with a little abuse of notations) in the lens plane is

$$\Phi_{\lambda}^{(k, m, n)}(x, y) = \Phi_{\lambda}^k(x, y) + \phi_{\lambda}^{(\alpha[m, n], \gamma[m, n])}(x, y). \quad (7)$$

Finally, the destination field (also its intensity, i.e., PSFs) in the sensor plane can be computed by Rayleigh-Sommerfeld scalar diffraction integral [Goodman 2005] of the source field  $e^{j\Phi}$  in the lens plane, numerically implemented by a (band-limited) shifted angular spectrum method (ASM) [Matsushima 2010]<sup>2</sup>.

Our image formation model is fully differentiable since all operations in Eq. (2) are differentiable. We implement this in an automatic-differentiation framework, enabling backpropagation [LeCun et al. 2015] and gradient-based optimization for the inverse design of the nanophotonic array camera.

### 3.2 Learning Collaborative Metalens Arrays

**Broadband Achromatic Initialization.** With the image formation model described in Section 3.1, we propose a broadband metalens phase initialization for the later joint optimization process. To this end, we optimize a rotationally symmetric metalens phase to maximize the center energy of the PSFs under all sampled incident wavelengths  $\Lambda_{disc}$  and angles  $\alpha_{disc}, \gamma_{disc}$ . Formally, we solve the following optimization

$$\tilde{\phi}_{\lambda_0}(r) = \arg \max_{\phi_{\lambda_0}} \sum_{\lambda_t \in \Lambda_{disc}} \sum_{\alpha \in \alpha_{disc}} \sum_{\gamma \in \gamma_{disc}} p_{\lambda_t}^{(\alpha, \gamma)}(\Delta x_{p^{(\alpha, \gamma)}}, \Delta y_{p^{(\alpha, \gamma)}}), \quad (8)$$

where

$$\begin{aligned} \Lambda_{disc} &= \{\lambda \mid \lambda = 400 + 0.5k, k \in \mathbb{Z}, 400 \leq \lambda \leq 700\}, \\ \alpha_{disc} &= \{\alpha \mid \alpha = 0 + 5k, k \in \mathbb{Z}, 0 \leq \alpha \leq 30\}, \\ \gamma_{disc} &= \{\gamma \mid \gamma = 0 + 5k, k \in \mathbb{Z}, 0 \leq \gamma \leq 30\}, \end{aligned}$$

and  $(\Delta x_{p^{(\alpha, \gamma)}}, \Delta y_{p^{(\alpha, \gamma)}})$  denote the offsets of the PSFs center under incident wave direction  $(\alpha, \gamma)$  on axis  $X$  and axis  $Y$ . By geometry, the offsets are proportional to focal length  $f$  and the incident angles

$$((\Delta x_{p^{(\alpha, \gamma)}}, \Delta y_{p^{(\alpha, \gamma)}}) = (-f \tan \alpha, -f \tan \gamma). \quad (9)$$

This starting phase ensures high initial focusing performance across the entire visible spectrum (sampled every 0.5 nm) and for both on-axis and off-axis incident angles ranging from  $0^\circ$  to  $30^\circ$  (sampled every  $5^\circ$ ).

**Joint Optimization with Multi-image Deconvolution.** Following the broadband phase initialization, we jointly optimize six metalenses in tandem with a multi-image joint deconvolution algorithm. As in Eq. (8), we employ the stochastic sampling technique from Section 3.4 to discretize the continuous visible spectrum  $\Lambda$  into the set  $\Lambda_{disc}$ . This effectively approximates the continuous image formation model for the color channel  $c$  in Eq. (2) via

$$v_{(\alpha, \gamma)}^{k, c} = \sum_{\lambda \in \Lambda_{disc}} \kappa_{\lambda}^c \left( u_{\lambda} \otimes p_{\lambda, (\alpha, \gamma)}^k \right) + \varepsilon^{(k, c)}, \quad (10)$$

<sup>2</sup>Note the shifted ASM kernel also depends on the incident angles. The details are deferred to the *Supplementary Material*.



where the hyperspectral latent images  $u_\lambda$  are convolved with the corresponding hyperspectral PSFs  $p_{\lambda,(\alpha,\gamma)}^k$ , under the incident direction  $(\alpha, \gamma)$ , plus camera noise  $\varepsilon^{(k,c)}$ .

During lens array training, we initially disregard potential parallax effects stemming from the array design to focus on improving the optical properties. Consequently, the reconstruction problem reduces to a multi-measurement, single-object scenario. In order to design an ultra-compact camera with minimal computational overhead, we adopt a multi-frame deconvolution approach [Yaroslavsky and Caulfield 1994], which jointly deconvolves the  $K$  frames and returns a single sharp result by solving for a given wavelength  $\lambda$

$$\min_{u_\lambda} \frac{1}{K} \sum_{k=1}^K \|v_{\lambda,(\alpha,\gamma)}^k - u_\lambda \otimes p_{\lambda,(\alpha,\gamma)}^k\|_2^2 + \beta \|u_\lambda\|_2^2, \quad (11)$$

with the following closed-form solution in the Fourier domain

$$U_\lambda = \sum_{k=1}^K \frac{\bar{P}_{\lambda,(\alpha,\gamma)}^k}{\sum_{\ell=1}^K |P_{\lambda,(\alpha,\gamma)}^\ell|^2 + K\beta} \odot V_{\lambda,(\alpha,\gamma)}^k, \quad (12)$$

where uppercase letters denote the Fourier transforms of the corresponding lowercase variables, and  $\bar{P}$  is the complex conjugate of  $P$ . Multiplication ( $\odot$ ) and division are applied element-wise. We set the regularization weight  $\beta$  to 0.001.

Equation (12) averages the  $K$  measurements processed with individual filters akin to the original Wiener filter [Wiener 1964] but where the denominator features the sum of the moduli of the  $K$  PSFs instead of a single one as in the original formulation. The individual filters are consequently aware of the other measurements during joint deconvolution into a single sharp result. We call *joint Wiener filter* this global filter applied to the  $K$  images  $v_{\lambda,(\alpha,\gamma)}^k$ . See the *Supplementary Material* for detailed derivation and explanations.

The joint Wiener filter enforces collaborative behavior across the lens array at each target wavelength  $\lambda_t$ . Rather than constraining each individual lens to perform uniformly well over the entire spectrum, as in Eq. (8), the contribution of each lens in the Fourier domain is weighted by its individual signal-to-noise ratio (SNR). This design relaxes the broadband capabilities of each lens and promotes mutual compensation, resulting in an improved broadband reconstruction when all lenses work collaboratively as a group.

However, the hyperspectral measurements  $V_{\lambda,(\alpha,\gamma)}^k$  required in Eq. (12) are challenging to acquire in commodity RGB color filter array (CFA) sensors. As such, we instead approximate the previous equation with the following color-channel broadband deconvolution method

$$U_{RGB}^c = \sum_{k=1}^K \frac{\bar{P}_{(\alpha,\gamma)}^{c,k}}{\sum_{\ell=1}^K |P_{(\alpha,\gamma)}^{c,\ell}|^2 + K\beta} \odot V_{(\alpha,\gamma)}^{c,k}. \quad (13)$$

We employ here instead the camera spectral response function (CSRF) weighted broadband average PSF and RGB image as

$$P_{(\alpha,\gamma)}^{c,k} = \sum_{\lambda \in \Lambda_{\text{disc}}} \kappa_\lambda^c P_{\lambda,(\alpha,\gamma)}^k \quad \text{and} \quad V_{(\alpha,\gamma)}^{c,k} = \sum_{\lambda \in \Lambda_{\text{disc}}} \kappa_\lambda^c V_{\lambda,(\alpha,\gamma)}^k. \quad (14)$$

The CSRF weights  $\kappa_\lambda^c$  from Eqs. (2) and (10) combine all the wavelengths  $\lambda$  in  $\Lambda_{\text{disc}}$  within a single joint Wiener filter.

*Spectrally Agnostic Training.* Instead of trying to estimate the spectrum from the measurements, we aim to learn a spectrally agnostic optical design with consistently good broadband performance. A way to simulate challenging spectral conditions is to sample a subset of bands during training randomly. Specifically, we randomly select  $n_v$  wavelengths from  $\Lambda_{\text{disc}}$  (denoted as  $S_{n_v}$ ) for the image formation process  $V_{(\alpha,\gamma)}^{c,k}$ , and  $n_p$  wavelengths (denoted as  $S_{n_p}$ ) for constructing the simulated deconvolution PSFs  $P_{(\alpha,\gamma)}^{c,k}$ , where  $|S_{n_p}| \gg |S_{n_v}|$ . In our experiments, we set  $|S_{n_p}| = 108$  and  $|S_{n_v}| = 3$ . We then construct a simulated measurement

$$W_{(\alpha,\gamma)}^{c,k} = \sum_{\lambda \in P_o} \kappa_\lambda^c V_{\lambda,(\alpha,\gamma)}^k, \quad (15)$$

which we recover with the approximate joint Wiener filter of Eq. (13) applied to  $W_{(\alpha,\gamma)}^{c,k}$  in place of  $V_{(\alpha,\gamma)}^{c,k}$  as synthetic input measurement. The loss function for optical training reads:

$$\mathcal{L}_{\text{optics}} = \|U_{RGB}^c - \sum_{\lambda \in S_o} \kappa_\lambda^c U_{\lambda_t}\|_2^2. \quad (16)$$

Equation (16) implicitly aligns the narrow-band RGB PSFs with their broadband counterparts while promoting the broadband collaborative PSFs that provide high reconstruction quality. This training will also make the lens robust to various illumination sources in the real world, owing to the random sampling policy.

Optimizing PSFs for a 100-million nano-post metasurface at 108 wavelengths demands significant GPU memory (around 2.1 TB intermediate tensors). To address this challenge, we leverage the distributed training framework detailed in Section 3.4. Each local device independently samples  $S_{n_v} \sim \Lambda_{\text{disc}}$  and then utilizes an all-gather operation to obtain the narrow-band PSFs from other devices, thereby forming the global broadband PSF  $P_{(\alpha,\gamma)}^{c,k}$ . Section 3.4 provides additional details on our distributed implementation.

### 3.3 Image Reconstruction Pipeline

*Alignment under Parallax.* Our metalens array captures  $K$  subimages  $v^k$  ( $k = 1, \dots, K$ ) of the same scene from slightly different viewing angles, thereby inducing parallax between the subimages. We propose a two-stage strategy to align these subimages to a common reference image  $\tilde{v}$ , which we choose to be the top middle subimage. We first coarsely align the  $v^k$ 's to a common system of coordinates with *global homographies*. We obtain the  $3 \times 3$  homography matrices  $\mathbf{H}^k$  by calibrating the  $K$  apertures with a ChArUco broad pattern. Each subimage  $v^k$  is warped to the reference system as

$$v_{\text{coarse}}^k(q) = v^k(\mathbf{H}^k q), \quad (17)$$

where  $q$  denotes pixel coordinates. Local misalignments may remain between the  $v_{\text{coarse}}^k$  and the reference due to depth variations or lens-specific distortions that a single homography cannot capture. Hence, we refine the previous coarse displacement with dense optical flows  $\mathbf{D}^k(u) \in \mathbb{R}^2$  with a pretrained RAFT model [Teed and Deng 2020], denoted by  $\mathcal{W}$ . Since the coarsely-aligned images are aberrated,  $\mathcal{W}$  trained on sharp natural images predicts inaccurate flow maps. We thus opt to first deconvolving the images  $v_{\text{coarse}}^k$  with individual



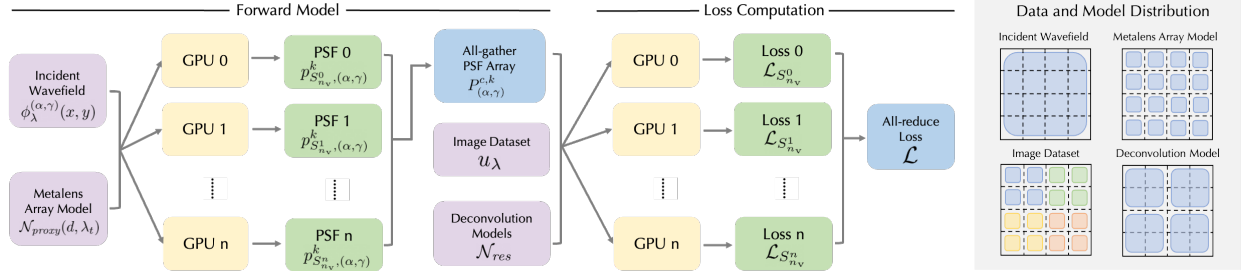


Fig. 4. **Distributed large meta-optics model optimization pipeline with Full Sharded Data Parallel (FSDP) illustration.** In the forward model, the incident wavefield is distributed across multiple GPUs. Each GPU has a complete metalens array model and computes the PSF for its assigned wavelengths, then contributes to the global PSF array through an all-gather operation. When computing loss, the image dataset is also distributed across GPUs. Depending on the size of the deconvolution model, we can also distribute the model weights. Each GPU calculates its local loss, and these losses are combined through an all-reduce operation to compute the global loss for optimization. The right figure provides a detailed illustration of data and model distribution, where the blocks represent a  $4 \times 4$  GPU cluster and each color block represents either a complete dataset or model weights. The different colors in the left bottom figure represent different image datasets.

Wiener filters reusing the Fourier transforms of the broadband PSFs in Eq. (14). Let  $s_{\text{coarse}}^k$  (resp.  $\tilde{s}$ ) be the deconvolved version of  $v_{\text{coarse}}^k$  (resp.  $\tilde{v}$ ), the estimation of the flow maps  $\mathbf{D}^k$  thus reads

$$\mathbf{D}^k = \mathcal{W}(s_{\text{coarse}}^k, \tilde{s}). \quad (18)$$

Lastly, we warp the blurry subimages to the reference following

$$v_{\text{warp}}^k(q) = v_{\text{coarse}}^k(\mathbf{D}^k(q)). \quad (19)$$

*Joint Wiener Filtering.* After having aligned the frames, we recover a single sharp image with a joint Wiener filter. To account for the spatially-varying formation model in Eq. (10), we first fragment the images  $v_{\text{warp}}^k$  into overlapping patches centered at locations  $(m, n)$ . We apply Eq. (13) to combine the  $K$  patches into a single one at each patch location  $(m, n)$  in the Fourier domain. Lastly, we merge the patches into a single restored sharp image  $\hat{u}_{\text{fused}}$ . To prevent edge artifacts during patch merge, we apply a Hann window  $w^{(m,n)}$  to the deconvolved patches before the fusion.

*Noise-Aware Image Denoising.* The joint Wiener module is a high-pass filter that removes blur, but that also boosts the noise from the measurements resulting in correlated noise in  $\hat{u}_{\text{fused}}$ . Furthermore, because of the wavelength-dependent PSFs of metalens, the boosted noise level also varies with the scene spectrum distribution.

Consequently, we develop a *noise-aware* post-fusion reconstruction module that integrates noise estimation and conditional denoising to further enhance image quality.

We employ a residual denoising neural network  $\mathcal{F}$  that denoises the fused image  $\hat{u}_{\text{fused}}$ . First, a *noise-estimation* subnetwork  $\mathcal{N}$  predicts an estimation of the variance map  $\hat{\sigma}$  as:

$$\hat{\sigma} = \mathcal{N}(\hat{u}_{\text{fused}}). \quad (20)$$

Subsequently, this noise estimate serves as guidance to a *conditional denoising* subnetwork  $\mathcal{D}$  as:

$$\hat{u} = \mathcal{D}(\hat{u}_{\text{fused}}, \hat{\sigma}), \quad (21)$$

where  $\hat{u}$  represents the final reconstructed and denoised estimate of the scene. By providing explicit noise information,  $\mathcal{D}$  becomes a non-blind denoiser that is adaptive to the input image, thus preserving

fine structures in low-noise regions and suppressing artifacts in high-noise areas.

We first derive a closed-form solution to estimate the noise variance after joint Wiener deconvolution, leveraging the hyperspectral PSFs of the lens array to train our noise estimator subnetwork  $\mathcal{N}$ . (See *Supplementary Material* for details of the derivation.) Subsequently, we train our noise-aware reconstruction model  $\mathcal{D}$  using a combination of supervision and regularization terms:

$$\mathcal{L} = \mathcal{L}_{\text{MSE}} + \alpha \mathcal{L}_{\text{TV}}, \quad (22)$$

where  $\mathcal{L}_{\text{MSE}}$  is the mean-square error between the prediction of the network and a ground-truth image, and  $\mathcal{L}_{\text{TV}}$  denotes the total variation [Rudin et al. 1992] loss,  $\alpha$  is a weighting parameter. Specifically, we adopt  $\alpha = 0.001$  to trade off fidelity and local smoothness.

### 3.4 Distributed Large Meta-Optics Optimization

Training a meta-optics array with 100 million parameters under one-hundred distinct wavefield conditions can easily require up to 2 TB of GPU memory—an order of magnitude exceeding the capacity of a single GPU. As a result, existing broadband meta-optics designs [Chakravarthula et al. 2023; Frösch et al. 2025; Tseng et al. 2021a] have been constrained to smaller aperture sizes, fewer spectral bands, and on-axis PSFs, largely due to limited GPU memory.

To relieve this constraint, we develop a distributed large meta-optics model optimization pipeline with Full Sharded Data Parallel (FSDP) [Zhao et al. 2023], as illustrated in Fig. 4. We distribute the incident wavefields, specified by the wavelength  $\lambda$  and incident angles  $(\alpha, \gamma)$  across multiple GPUs, akin to the concept of data in the machine learning context. We adopt a stochastic sampling scheme analogous to mini-batch training in deep learning. Rather than fixing a discrete set of wavelengths as in prior work [Chakravarthula et al. 2023], each optimization iteration draws a subset of wavefields randomly sampled across various  $\lambda$  and  $(\alpha, \gamma)$  values. This randomization ensures broad coverage of the underlying high-dimensional design space over the course of training, providing more robust optimization outcomes and less susceptibility to overfitting any particular set of wavefields.

Although a single GPU can generally manage the total number of trainable parameters for a single meta-optics model, the intermediate tensors generated during PSF computations for each wavefield are often two orders of magnitude larger than the model parameters themselves. To mitigate this overhead, we shard the model across multiple GPU ranks and split the wavefield inputs accordingly. Leveraging FSDP in PyTorch, we effectively partition both the model and wavefield data, ensuring that each GPU processes only a fraction of the intermediate tensors.

While FSDP manages parameter and gradient synchronization across GPUs, large-scale broadband meta-optics training introduces additional global synchronization demands. For example, broadband PSF calculation during training with Eqs. (13) and (16) necessitates an *all-gather* operation to share each rank’s PSF results with all other ranks. Similarly, when computing a broadband loss term (e.g., Eq. (8)), an *all-reduce* operation is required to accumulate partial losses from individual ranks into a global loss value. Our proposed training pipeline incorporates a suite of meta-optics-specific utilities—built on top of PyTorch distributed collectives—to address these synchronization requirements (details in *Supplementary Material*). These utilities cleanly integrate with distributed optimizers, ensuring that gradient information remains consistent with model parameters throughout the entire training process. The result is a fully end-to-end differentiable framework for large-scale meta-optics design.

### 3.5 Implementation and Training Details

**Meta-Optics Optimization.** We utilize the proposed distributed large-scale meta-optics framework in conjunction with a hyperspectral dataset [Jeon et al. 2024] to conduct optics optimization. Specifically, we deploy the framework by distributing metalens array model instances across 36 Nvidia H100 GPUs (each with 80 GB of memory) on 9 nodes (4 GPUs per node), using a data-parallel training scheme. Each rank (i.e., each GPU) receives distinct incident wavefields—characterized by unique wavelengths and incident angles—sampled via the distributed sampler. Due to the considerable volume of intermediate results generated during wave propagation, each rank can only process three incident fields at a time. Consequently, a total of 108 field conditions (36 ranks  $\times$  3 incident fields per rank) are evaluated during each training iteration.

By leveraging the hyperspectral dataset, each rank performs a shift-invariant convolution weighted by the hyperspectral illumination to generate the hyperspectral array measurement, which is then converted into RGB space using the camera’s spectral response function. Noise is subsequently introduced, and the resulting hyperspectral array measurements are passed to the joint Wiener deconvolution module to compute the loss. This loss signal is then backpropagated to update the metalens parameters, while an all-reduce operation ensures that the optimized parameters remain consistent across all ranks.

**Reconstruction Pipeline Training.** Given a jointly optimized array design, we train the reconstruction pipeline in a sequential training stage. Specifically, we aim to learn an inverse reconstruction model that jointly tackles parallax alignment, large off-axis astigmatism

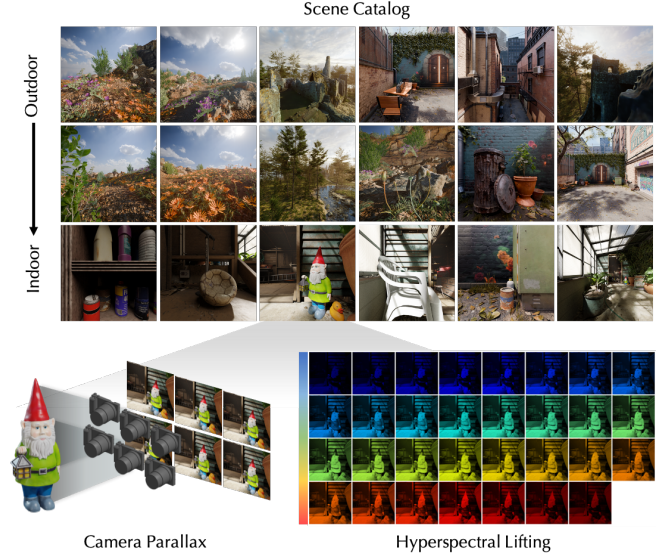


Fig. 5. **Illustration of synthetic dataset generation.** The scene catalog shows representative examples of our dataset. For each scene in the dataset, we set up a  $2 \times 3$  camera array to capture scene parallax and apply hyperspectral lifting to get images under wavelengths from 400 nm to 700 nm.

correction, and denoising on our proposed hyperspectral parallax dataset introduced in Section 4.1.

We begin by dividing the target scene into  $9 \times 9$  patches, corresponding to the sensor dimensions and the spatial arrangement of the metalens array. For each sub-aperture lens, we infer a full-spectrum (400 nm to 700 nm in 2 nm increments) PSF tensor based on the incident angle at the center of each patch. We then convolve each hyperspectral patch with the appropriate spectral PSF tensor, apply window-based blending between adjacent patches, and subsequently convert the resulting measurements back to sensor RGB space with CSRF. This process yields a realistic sensor measurement that accurately simulates parallax and angular aberrations across the entire FOV.

After alignment, see Section 3.3, we apply the patch-wise joint Wiener deconvolution followed by residual denoising (initialized with a pretrained image denoiser [Guo et al. 2019]) to warped sub-aperture images, jointly reconstructing the scene in an end-to-end manner. The overall training loss is backpropagated through the denoising, deconvolution modules, and parallax alignment, allowing the entire pipeline to be optimized for reconstruction fidelity.

## 4 DATASETS AND EXPERIMENTAL PROTOTYPE

### 4.1 Synthetic Hyperspectral Parallax Dataset

To circumvent the difficulty of capturing large amount of meta-optic array sensor measurements with corresponding high-quality ground truth, we develop a physics-based realistic rendering pipeline that can accurately synthesize realistic metalens array measurements that resemble the real-world observations. Our rendering pipeline here is inspired by today’s photorealistic spectral rendering technology developed in the computer graphics community [Jakob and Hanika 2019; Mallett and Yuksel 2019; Meng et al. 2015; Otsu et al.

2018; Van De Ruit and Eisemann 2023], yet we inject an additional rendering layer that imposes spatial-spectral-dependent PSFs to generate highly color-aberrated sensor measurements from otherwise clean images (see Fig. 2). Specifically, we use the 3-D scenes from Poly Haven [Haven 2025], a curated public 3-D asset library that contains complete Blender [Community 2025a] 3-D render-ready scenes. We set up camera positions to capture objects with different sizes, textures, colors, and distances from the camera. The sensor size and focal length of the camera are set to match the design of one single metalens camera. We offset the position of the camera according to the distance between metalenses in the array to capture the parallax in different scenes. A total of 480 sets of photorealistic linear sRGB images in an array are thus rendered using the Cycles engine [Community 2025b] under the CIE standard illuminant D65, which covers diverse indoor and outdoor scenes in cities and nature; see Fig. 5 for some representative scenes in our dataset. These linear sRGB images are then fed into a spectral uplifting method [Jakob and Hanika 2019] to synthesize hyperspectral reflectance data. Next, a spectral illuminant, randomly sampled from a spectral illumination dataset [Li et al. 2021], is blended with the spectral reflectance to synthesize realistic hyperspectral irradiance images. Finally, these hyperspectral irradiance images can be optionally convolved with the wavelength- and angle-dependent PSFs (computed from our meta-optic model)<sup>3</sup> before the spectral integration weighted by a sensor spectral response function [Jiang et al. 2013], specifically for the Sony IMX 174 sensor in our prototype, to synthesize realistic color sensor observations.

## 4.2 Experimental Setup and Datasets Capture

For real-world experimental assessment, we built a prototype to validate our collaborative metalens-based camera design in indoor and outdoor settings (Fig. 6). The setup consists of two cameras: a metalens array camera and a reference camera for comparison. The metalens camera used a CMOS sensor (Allied Vision, Prosilica GT1930C) with  $5.86 \mu\text{m}$  pixel pitch and a metalens array of  $2 \times 3$  lenses with  $3.6 \text{ mm}$  focal length,  $3.57 \text{ mm}$  lens pitch and  $1.5 \text{ mm}$  lens diameter. The fabrication details of our metasurface optic can be found in *Supplementary Material*. We aligned the metalens array parallel to the sensor using a three-axis translation stage and a two-axis rotatable mount. Each metalens produces a subimage of approximately  $3.5 \text{ mm} \times 3.5 \text{ mm}$ , covering a FOV of  $50^\circ \times 50^\circ$ . We used a custom optical baffle to prevent crosstalk and added a UV/IR cut-off filter (Edmund Optics, 54-749) to block the invisible spectrum. For reference comparison, we used a separate camera using a FLIR GS3-U3-32S4C-C sensor with a  $3.5 \text{ mm}$  C-series fixed focal length camera that has a physical length of  $\sim 50 \text{ mm}$  from front to back. A 60T/40R beamsplitter (Edmund Optics, 72-502) enabled simultaneous capture through both cameras. We developed a custom graphical user interface for visualization and synchronized camera control with adjustable exposure time and white balance. We captured 653 indoor and outdoor scenes, aligning the reference and metalens images through homography calibration using a ChArUco board pattern. See *Supplementary Material* for more details. Notably, our proposed reconstruction method does *not* rely on large-scale

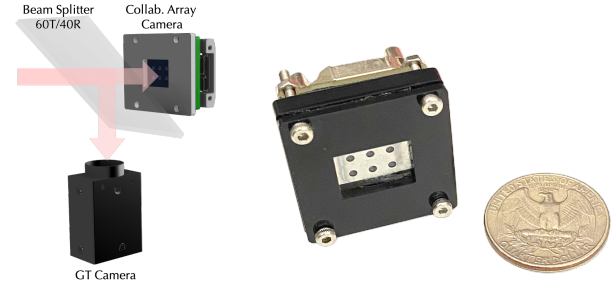


Fig. 6. **Data acquisition and experimental comparisons.** Our setup uses a 40R/60T plate beamsplitter to capture scenes with two cameras at once: a metalens array camera that receives 60 % of the light passing through the beamsplitter, and a conventional camera (reference camera) that receives 40 % of the light reflected by the beamsplitter. The cameras are aligned and calibrated to ensure their images match correctly. See the *Supplementary Material* for additional details.

training using experimentally captured datasets. Instead, it leverages a much larger amount of synthetic data generated from the rendered hyperspectral parallax dataset described earlier.

## 5 SYNTHETIC ASSESSMENT

### 5.1 Broadband PSF Analysis

We start the proposed metalens array evaluation by spectrally resolved PSF analysis. Specifically, we sample every  $1 \text{ nm}$  between  $400 \text{ nm}$  and  $700 \text{ nm}$  as the incident plane-wave wavelength,  $\Lambda_{\text{disc}}$ . For each sampled wavelength and each of the  $K$  lenses, we compute the field energy distribution on the image plane using the angular-dependent PSF model from Section 3.1, with spatial sampling matching the camera sensor model in Section 4. The result is a 4D PSF tensor  $P_{\text{disc}}$  of shape  $(K, |\Lambda_{\text{disc}}|, N_x, N_y)$ , where  $N_x$  and  $N_y$  represent the sensor pixel dimensions behind each individual lens. To visualize this high-dimensional PSF tensor, we take a radial slice  $P_{\text{slice}}$ , that is

$$P_{\text{slice}}[:, :, \frac{N_x}{2}, :]$$

and display each sublens PSF slice at the bottom of Fig. 7 after a global normalization (preserving their relative energy). For an ideal lens, this response for the full visible spectrum is a continuous centered line. We observe that the proposed joint optimization promotes a lens array in which the collective output, combined through the proposed joint reconstruction, approximates an ideal broadband lens. The sub-lenses collaborate to capture the entire broadband spectrum. The top-center “synthetic effective lens” is the weighted sum of each sublens contribution, based on their signal-to-noise ratio (SNR) as described in Eq. (12). This plot confirms the collaborative nature of the learned array elements.

We compare to existing broadband metalens imaging techniques [Chakravarthula et al. 2023; Tseng et al. 2021a] with the same analysis as described above. Figure 7 confirms that these imagers produce only small-footprint PSFs at the discrete wavelengths for which they were optimized ( $462 \text{ nm}$ ,  $511 \text{ nm}$ , and  $606 \text{ nm}$  in [Tseng et al. 2021a], and  $400 \text{ nm}$  to  $700 \text{ nm}$  in  $10 \text{ nm}$  increments for [Chakravarthula et al. 2023]). In contrast, our proposed design—with collaboratively optimized sublenses—maintains continuously sharp PSFs across the full

<sup>3</sup>PSF-based aberration rendering is disabled for the ground truth.



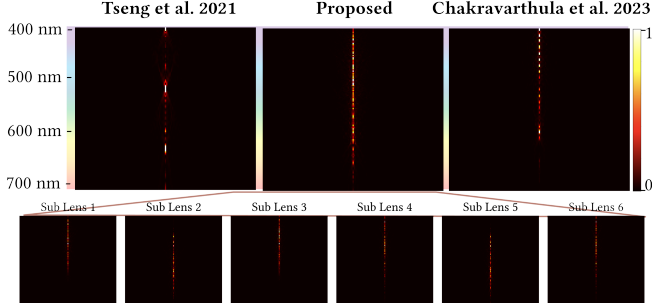


Fig. 7. **Analysis of PSFs for existing metalens array imagers.** We plot here X-Z projection of 3D stacked PSFs spanning 400 nm to 700 nm in 1 nm intervals. This visualization highlights the discretely sharp PSFs produced by [Chakravarthula et al. 2023; Tseng et al. 2021a] under broadband illumination, whereas our proposed design, with collaboratively optimized sublenses, maintains a continuously sharp PSF across the entire visible range. We note that the individual sublenses collaborate here to focus over the entire spectrum.

visible spectrum. This outcome further highlights the importance of continuous wavelength sampling and collaborative optimization for achieving high-quality broadband metalens imaging.

## 5.2 Synthetic Evaluation

**Evaluation of Broadband Performance.** We confirm the trend from the broadband PSF analysis above for image reconstruction. To this end, we simulate measurements using a real-world hyperspectral dataset [Jeon et al. 2024] as input. Specifically, the sensor measurement is generated by following the noise model in Section 3.1 with a Poissonian-Gaussian noise model. To handle any spectral mismatch between the sampled wavelength and dataset wavelength, we interpolate the input hyperspectral images. We then convert the hyperspectral ground truth image, metalens measurement, and PSF back into the RGB domain according to the camera’s spectral response function. The same joint deconvolution algorithm from Section 3.3 is applied for reconstruction except for the parallax alignment module.

First, we evaluate performance at the *discrete design wavelengths* for which other respective optics are optimized for: (462 nm, 511 nm, and 606 nm in [Tseng et al. 2021a], and 400 nm to 700 nm in 10 nm increments for [Chakravarthula et al. 2023]). To maintain consistency, we test our proposed array using the same sampling grid of [400 nm : 10 nm : 700 nm], even though it was *not explicitly optimized for this discrete wavelength set*. As reported in Fig. 8, all three imagers yield comparable reconstruction quality, apart from minor color deviations in [Tseng et al. 2021a], which uses fewer sampling wavelengths. This result validates the effectiveness of the alternate designs for their intended discrete-wavelength range.

Next, we test each imager under *continuous broadband illumination*, for which we approximate with discrete 400 to 700 nm wavelength sampling with 1 nm spacing. In agreement with the broadband PSF analysis, the two baseline methods experience a notable drop in performance, owing to the poor SNR at the unoptimized bands. In contrast, our collaborative design performs robustly across the continuous broadband illumination, where the other methods severely degrade. Notably, our design even performs better under

Table 1. **Quantitative evaluation of broadband image reconstruction with flat metasurface imagers.** We report here quantitative measurements corresponding to the experiments visualized in Fig. 8. The evaluations confirm the qualitative trends. While existing metalens imaging methods suffer from lower SNR across unoptimized bands, our collaborative design remains robust for broadband imaging, see text.

|             | Tseng<br>et al. [2021a] | Chakravarthula<br>et al. [2023] | Proposed     |
|-------------|-------------------------|---------------------------------|--------------|
| SSIM ↑      | 0.72                    | 0.80                            | <b>0.95</b>  |
| PSNR [dB] ↑ | 25.92                   | 27.34                           | <b>34.03</b> |
| 1-LPIPS ↑   | 0.55                    | 0.62                            | <b>0.87</b>  |

broadband conditions than under discrete-wavelength illumination, because the PSFs used for deconvolution align more closely with the physical system, and the overall quality is higher when more spectral bands contribute to the measurement.

**Validation of Collaboration.** In addition to the broadband PSF results shown in Fig. 7, which indicate that individual sublenses collaboratively exhibit small-footprint PSFs, we also optimized a single broadband metalens using the same methodology described in Section 3.2 to compare with the proposed design for validating the effectiveness of the collaborative array design.

To simulate realistic real-world captures under challenging illumination conditions, we utilize the hyperspectral illuminator dataset from [Li et al. 2021] in combination with the hyperspectral imaging dataset mentioned in [Jeon et al. 2024]. Specifically, we select 10 illuminators with tilted or discrete spectral characteristics and uniformly sample wavelengths in the range [400 nm : 10 nm : 700 nm]. Sensor measurements are then generated by computing a hyperspectral illumination weighted sum of the hyperspectral measurements, which are obtained via hyperspectral domain convolution between the hyperspectral data and the PSFs. Representative reconstruction outputs are shown in Fig. 9.

Although both imagers face substantial challenges due to the extreme illumination conditions, the single broadband-optimized lens shows noticeably more variation in SNR across different illuminators. In contrast, the proposed lens array design remains robust, producing consistently high-fidelity reconstructions. This underscores the spectral bandwidth limitations [Presutti and Monticone 2020] of a single metalens in maintaining strong broadband performance while highlighting the robustness of our collaborative array approach.

**Ablation Experiments.** To validate the effectiveness and all design choices of our proposed reconstruction method, we conduct ablation experiments on the synthetic hyperspectral parallax dataset introduced in Section 4.1. We compare five variants of our method: (i) the pipeline without shift-variant deconvolution, (ii) the pipeline without the parallax aligner, (iii) the pipeline without the joint deconvolution module, (iv) the pipeline without the noise estimator, and (v) the full pipeline (proposed). Qualitative and quantitative results are reported in Fig. 10 and Table 2.

All ablated methods operate with the same array measurements and PSFs. Specifically, we randomly sample 10 distinct wavelengths

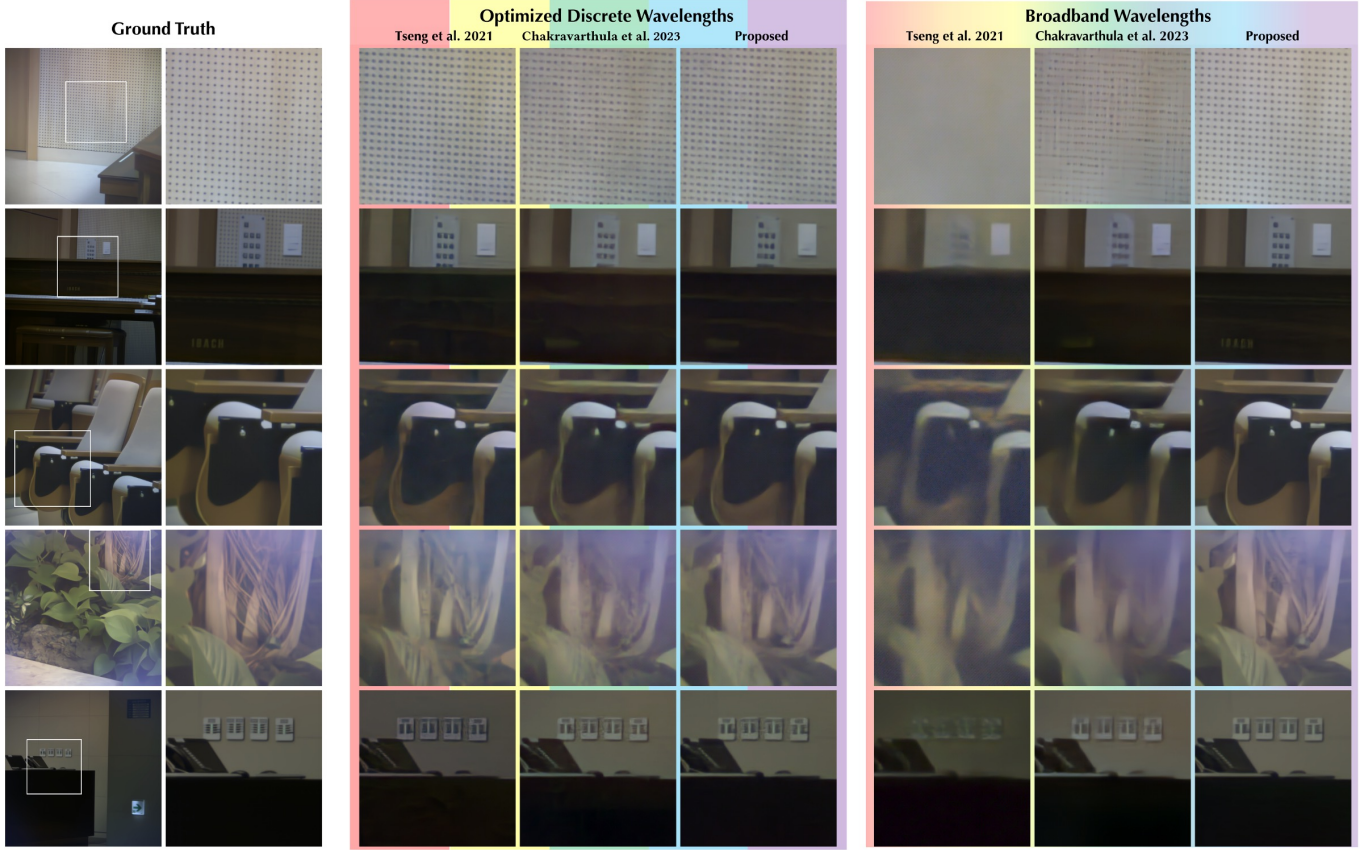


Fig. 8. **Synthetic broadband image reconstruction with flat metasurface cameras.** Due to computational and memory complexity, existing flat metalens optics [Chakravarthula et al. 2023; Tseng et al. 2021a] have been optimized for discrete wavelength sets. As such, these methods perform comparably to our proposed design at the discrete spectral bands for which they were optimized for (462 nm, 511 nm, and 606 nm for [Tseng et al. 2021a], and 400 nm:10 nm:700 nm for [Chakravarthula et al. 2023]). However, their quality deteriorates significantly under real-world continuous broadband illumination (400 nm:1 nm:700 nm). In contrast, the proposed design method, along with the array design, lifts wavelength sampling limitations, and the full-spectrum collaborative array consistently delivers superior image quality across the full spectrum.

from the visible spectrum and generate the corresponding hyperspectral ground-truth image using the hyperspectral lifting method provided by our rendered dataset. We then compute the hyperspectral PSFs of the metalens array for each sampled wavelength. The RGB lens-array measurement with parallax is synthesized by convolving the ground-truth image patches with the wavelength-specific PSFs that correspond to the incident angle of the patches, integrating the results according to the CSRF, and finally adding Poissonian-Gaussian noise [Foi et al. 2008] with parameters  $(a, b) = (4 \times 10^{-5}, 1 \times 10^{-6})$  estimated from experimental raw captures.

**Without Shift-Variant Deconvolution.** In our short focus design, the large incident angles at the periphery of the FOV induce astigmatism. When the system’s shift-variance is ignored – using only the on-axis PSF for deconvolution – resolution degrades and color artifacts appear, as shown in Fig. 10. By contrast, applying patch-wise shift-variant deconvolution accounts for off-axis aberrations and effectively corrects the degradation, resulting in improved image quality.

Table 2. **Quantitative ablation study.** We simulate RGB shift-variant array measurements using a hyperspectral parallax dataset and corresponding PSFs to assess our proposed reconstruction method. Each sub-module (spatial-variant deconvolution, parallax aligner, joint deconvolution, and SNR estimator) is removed in turn, and the resulting models are evaluated on SSIM, PSNR, and LPIPS [2018]. Our findings indicate that each sub-module serves a distinct yet essential function in the final reconstruction. By combining all sub-modules, the full reconstruction method yields the best overall performance.

|             | W/O SV<br>Deconv | W/O Parallax<br>Aligner | W/O Joint<br>Deconv | W/O SNR<br>Estimator | Proposed     |
|-------------|------------------|-------------------------|---------------------|----------------------|--------------|
| SSIM ↑      | 0.53             | 0.62                    | 0.37                | 0.61                 | <b>0.67</b>  |
| PSNR [dB] ↑ | 18.32            | 19.74                   | 18.90               | 20.18                | <b>21.14</b> |
| 1-LPIPS ↓   | 0.41             | 0.49                    | 0.35                | 0.46                 | <b>0.52</b>  |

**Without Parallax Aligner.** Removing the neural warping operations for parallax correction degrades reconstruction quality, especially for objects closer to the camera where misalignment becomes more pronounced. Although the baseline distance between the individual metalenses is relatively small, reconstruction on our 3D rendered parallax dataset shows a noticeable drop in reconstruction quality for objects located within one meter of the camera.





Fig. 9. **Comparison between a single broadband-optimized lens and the proposed collaborative array under challenging illumination spectra.** Although a single metalens can be continuously optimized for broadband illumination within our framework (left), a single lens struggles to focus the entire visible spectrum. In contrast, our method employs multiple collaboratively optimized lenses to overcome these limitations. We confirm this here qualitatively with reconstructions under challenging illumination scenarios—such as heavily tilted spectra or discrete illumination [Li et al. 2021]—demonstrating that our collaborative array design consistently outperforms the single broadband lens across a variety of illumination conditions, see text.

**Without Joint Deconvolution.** Next, we investigate replacing the joint deconvolution module with six independent deconvolution branches. The resulting per-metalens reconstructions are then averaged to form the input to the remaining pipeline. Since individual metalenses cannot maintain a consistently high SNR across a broad range of wavelengths, this independent approach produces chromatic artifacts. Furthermore, averaging these color-degraded images only compounds the blur and color inaccuracies.

**Without Noise Estimator.** Finally, we remove the neural noise estimator and retrain the network while keeping the parallax aligner

and joint deconvolution module weights fixed. Due to the wavelength-dependent SNR response of the metalens, the output from the deconvolution module exhibits varying degrees of colored noise amplification based on the MTF under different sampled wavelengths. As a result, the network without noise estimator struggles to disentangle the signal from the various noise level measurements, leading to degraded reconstructions [Guo et al. 2019].

Overall, the ablation experiments from above confirm the effectiveness of each component in the proposed reconstruction method.

## 6 EXPERIMENTAL ASSESSMENT

To validate the proposed design experimentally across diverse and challenging conditions—including variations in illumination, depth parallax, and the presence of novel objects—we conducted real-world experiments using our fabricated metasurface. Using the prototype camera design described in Section 4.2, we captured scenes featuring small text, human faces, significant depth variations, and low-contrast textures. These scenes were recorded under a wide range of illumination conditions such as low-light, artificial light, direct sunlight, and intense localized light sources. Typical exposure times were 100 ms for indoor and 30 ms for outdoor scenes, and could be reduced by removing the 60T/40R beam-splitter. Representative measurement and reconstruction results are shown in Fig. 11, alongside aligned comparison measurements from a reference 3.5 mm C-series fixed focal length camera that has a physical length of  $\sim 50$  mm. Each metalens in the array captures broadband scene measurements of objects in both indoor and outdoor scenes at the center of each metalens. The dark bands separating adjacent sub-images arise from the integrated optical baffle, which prevents stray light and suppresses crosstalk between neighboring apertures. Our reconstruction method mitigates the spatially-varying optical aberrations, improving the image contrast in the center and the periphery with finer details. See *Supplementary Material* for more results, and the *Supplementary Video* for dynamic scenes.

**Qualitative Comparison to Existing Array Cameras.** Next, in Fig. 12, we compare our camera prototype with the metalens array camera proposed in [Chakravarthula et al. 2023], whose image reconstruction technique is based on a diffusion prior. The proposed method achieves improved detail and substantially less scattering in the raw measurement: for instance, detail on the human face, the building entrance or the traffic sign in Fig. 12. These raw measurements validate densely sampling of the wavelengths and incident angles during optical optimization, as well as the accuracy of the proposed neural proxy modeling in Section 3.1.

We surpass the diffusion-based approach of [Chakravarthula et al. 2023] both in terms of accuracy and absence of ‘hallucinated content’ in the reconstructed results. We correctly recover details such as the face of the person and the polka dots on the book, the windows on the building entrance, or even the text ‘stop’ on the traffic sign, whereas the compared method instead smooths the face and mangles the dots on the book, and introduces visual artifacts on the entrance and the stop sign that is hardly readable. This is due to the generative prior of [Chakravarthula et al. 2023] trained to lift a *single* low-contrast measurement to a sharper full-color result. In contrast to the proposed method, hallucination appears in the prediction of



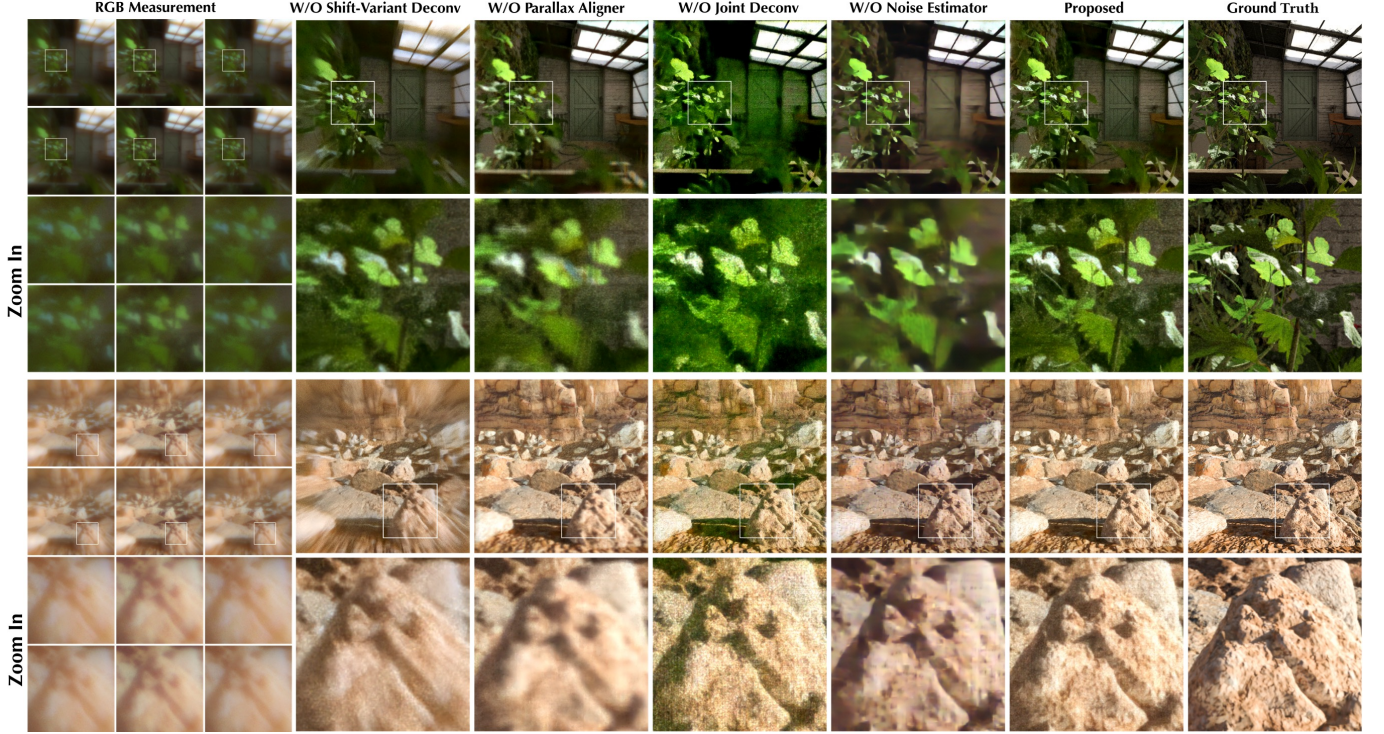


Fig. 10. **Qualitative ablation experiments.** We evaluate our reconstruction pipeline by comparing three ablated variants on the hyperspectral parallax-aware metalens array dataset: (i) replacement of the spatial-variant deconvolution with spatial-invariant deconvolution using on-axis PSF, (ii) removal of the parallax-aligner, (iii) replacement of the joint deconvolution step with independent Wiener deconvolution, and (iv) removal of the noise neural estimator. Using spatial-invariant deconvolution instead of spatial-variant deconvolution results in astigmatism at the periphery of the FOV. Methods without the parallax aligner exhibit blurry reconstructions in regions with closer object distance due to misalignments. Replacing the joint deconvolution with independent deconvolution leads to pronounced color aberrations, as individual metalenses are not broadband. Finally, omitting the noise neural estimator hampers noise removal under the metalens array’s broadband varying SNR for each scene wavelength, compromising detail preservation.

details where there is an ambiguity about the underlying sharp image. This is illustrated in Fig. 1 where the generative model has restored ‘65’ whereas the original number is ‘86’.

**Runtime Performance.** Our reconstruction runs at real-time rates. Compared to the diffusion module of [Chakravarthula et al. 2023] that takes more than a minute to process a single image, our technique is 2000× faster and runs at 35 FPS on the same test platform. Both approaches are implemented in PyTorch and do not leverage further optimizations. We rely on fast FFT-based inverse filtering with a subsequent denoising network that is lightweight since fusing six images already reduces noise<sup>4</sup>. The runtime bottleneck stems from the dense optical flow prediction with the optical flow model [Teed and Deng 2020] (which takes up about 60% of the time for reconstructing a frame with our technique).

## 7 DISCUSSION

Our collaborative metalens-array prototype delivers high-fidelity, broadband imaging under diverse illumination but it remains a research demonstrator rather than a product-grade camera due to its form factor. To facilitate prototyping, we retained the Sony IMX174

sensor cover glass and mounted a 7 mm × 10.5 mm metalens array just 1 mm above it. The array dimensions match the sensor 7.1 mm × 11 mm active area, and the resulting optics yield a 3.6 mm focal length. In an industrialized, smartphone-style implementation—with integrated inner baffles and f/2 optics on a 12 MP sensor—the array could shrink to a 2 × 2 layout of 1.5 mm elements, occupying only a 5 mm × 5 mm footprint at a 3 mm standoff while delivering a 60° field of view. This ultra-thin configuration aligns with smartphone packaging constraints—minimizing optical thickness while efficiently utilizing the PCB footprint, whereas comparable refractive solutions would either degrade into a low-quality fisheye or require an impractical, multi-element assembly. Despite the throughput trade-offs inherent to high-NA metaoptics, our prototype still supports real-time, video-rate capture.

Our reconstruction pipeline runs at 35 FPS on laptop-class GPUs and stands to gain further image-quality improvements from ring deconvolution [Kohli et al. 2022] and sub-aperture super-resolution [Eboli et al. 2022]. Pruning and fine-tuning a lightweight optical-flow network, denoiser distillation, and dedicated ISP hardware could further accelerate processing for potential on-edge deployment. The array’s angular multiplexing also enables direct depth estimation from raw captures [Venkataraman et al. 2013]. At the

<sup>4</sup>If we assume the noise to be purely Gaussian, merging 6 images corresponds to a  $\sqrt{6} \approx 2.45$  noise reduction factor.



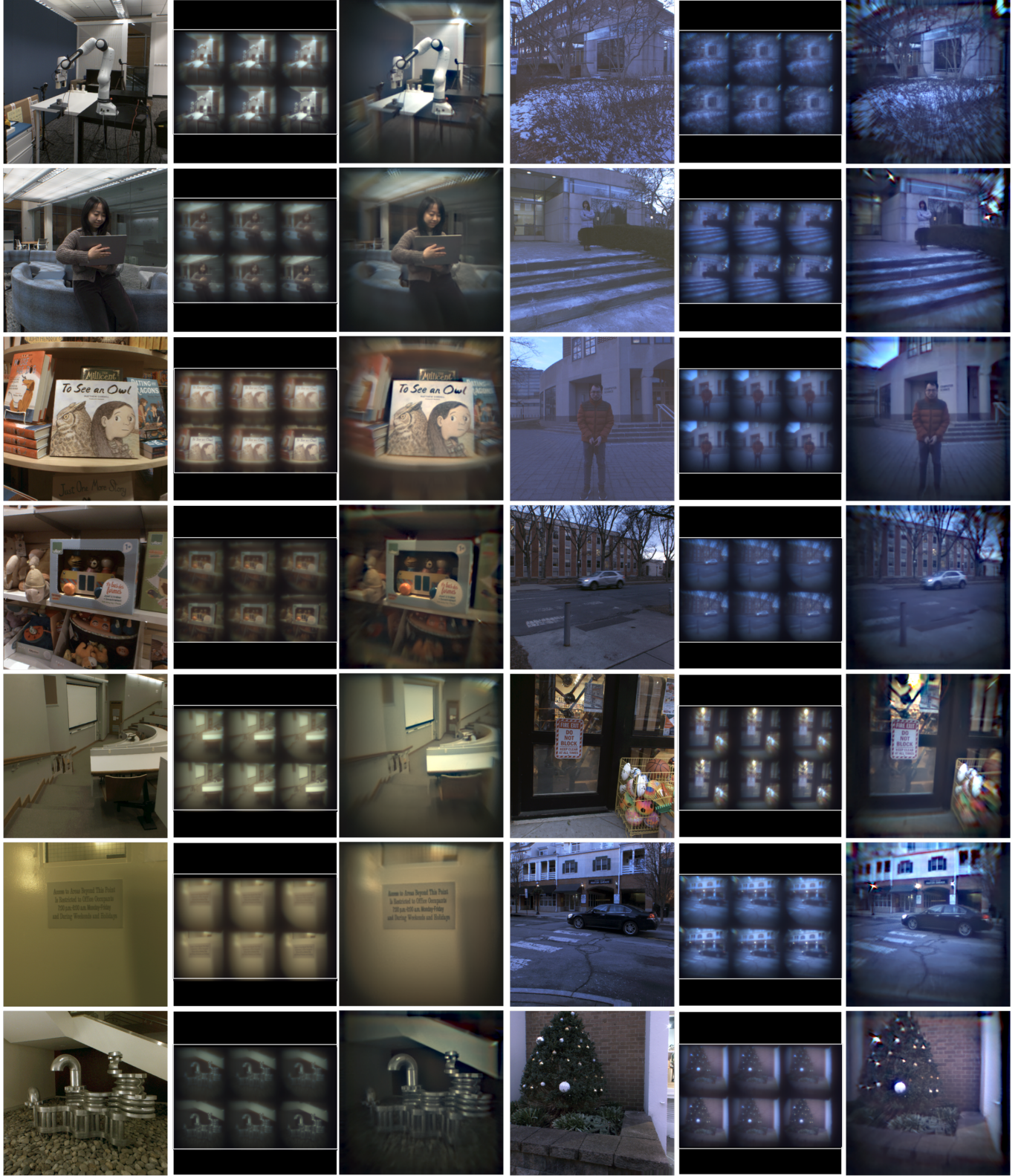


Fig. 11. **Experimental reconstructions of indoor and outdoor scenes.** The broadband-optimized metalens design yields a sharp array of image measurements with small chromatic aberration across diverse illumination conditions. The reconstructed image quality and color fidelity approach that of the reference images, captured by a conventional multi-element camera at a comparable F-number.





Fig. 12. **Experimental comparison to [Chakravarthula et al. 2023].** For similar scenes and illumination conditions, the proposed collaborative metalens array provides more accurate measurements and image reconstructions than [Chakravarthula et al. 2023]. The image zoom-ins show the proposed collaborative metalens array can capture a more broadband focused raw measurement, and a more plausible reconstruction of the scene, revealing details such as the portrait, building door windows, and a stop sign.



design level, our optimization framework supports full-angle, multi-spectral phase-response simulations for arbitrary metalens array geometries free from radial-symmetry constraints at the cost of memory overhead. By restricting the phase profile to radial symmetry under on-axis illumination, one can instead take advantage of a one-dimensional Rayleigh-Sommerfeld diffraction model to cut memory requirements [Fröch et al. 2025], trading off large-angle performance. Finally, extending our neural surrogate to capture angle- and wavelength-dependent transmission [Hazineh et al. 2022; Wirth-Singh et al. 2025] could further enhance design fidelity and unlock new high-NA meta-optic applications.

## 8 CONCLUSION

We have introduced a novel computational imager using collaborative metasurface arrays that enables broadband imaging in an ultra-thin form factor. Through the integration of a learned neural proxy for nanostructure-to-phase mapping, jointly designed collaborative reconstruction, and large-scale distributed optimization of 100 million nanoposts, our design framework achieves significant improvements to prior metalens array designs in both raw and reconstruction image quality. We validate the method with an experimental prototype that successfully captures and reconstructs broadband images in both indoor and outdoor scenes under varying illumination spectra.

In the future, one exciting avenue for research is to correct spatially-varying aberrations, which are an intrinsic difficulty of single-layer optics, through multiple metasurface elements in a thin stack, with the expanded parameter space being manageable by our framework. Looking forward, we hope that our method for ultra-thin metasurface imagers serves as a systematic tool for exploring new large-area optics applications, including large-scale manufacturing of on-chip metasurface imaging sensors, collaborative optimization for polarization-sensitive and hyperspectral metasurface imaging, and specialized meta-imaging systems. With the broadband capabilities enabled by this work, we hope that on-chip imaging sensors realize the potential to enable broad applications across the domains – from tiny camera arrays in micro-robotics to wearable flat cameras in healthcare.

## ACKNOWLEDGMENTS

We thank Dongyu Du, Jiwoon Yeom, Zheng Shi, and Johannes Fröch for their fruitful discussion on the project. Felix Heide gratefully acknowledges support from an NSF CAREER Award (2047359), a Packard Foundation Fellowship, a Sloan Research Fellowship, a Disney Research Award, a Sony Young Faculty Award, a Project X Innovation Award, and an Amazon Science Research Award. Wolfgang Heidrich acknowledges support from KAUST Individual Baseline Funding. Fabrication was carried out at the Washington Nanofabrication Facility/Molecular Analysis Facility at the University of Washington—a National Nanotechnology Coordinated Infrastructure site supported by the National Science Foundation under Grant Nos. NSF-2127235 and NSF-2223495 (EFRI-BRAID).

## REFERENCES

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.
- Amir Arbabi and Andrei Faraon. 2023. Advances in optical metalenses. *Nature Photonics* 17, 1 (2023), 16–25.
- Salman Asif, Ali Ayremlou, Aswin Sankaranarayanan, Ashok Veeraraghavan, and Richard G Baraniuk. 2016. Flatcam: Thin, lensless cameras using coded aperture and computation. *IEEE Transactions on Computational Imaging* 3, 3 (2016), 384–397.
- Seung-Hwan Baek and Felix Heide. 2021. Polka Lines: Learning Structured Illumination and Reconstruction for Active Stereo. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Seung-Hwan Baek, Hayato Ikoma, Daniel S Jeon, Yuqi Li, Wolfgang Heidrich, Gordon Wetzstein, and Min H Kim. 2021. Single-shot Hyperspectral-Depth Imaging with Learned Diffractive Optics. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 2651–2660.
- Ralph Beebe Blackman and John Wilder Tukey. 1958. The measurement of power spectra from the point of view of communications engineering—Part I. *Bell System Technical Journal* 37, 1 (1958), 185–282.
- Vivek Boominathan, Jesse K. Adams, M. Salman Asif, Benjamin W. Avants, Jacob T. Robinson, Richard G. Baraniuk, Aswin C. Sankaranarayanan, and Ashok Veeraraghavan. 2016. Lensless Imaging: A computational renaissance. *IEEE Signal Processing Magazine* 33, 5 (2016), 23–35.
- Vivek Boominathan, Jesse K Adams, Jacob T Robinson, and Ashok Veeraraghavan. 2020. Phlatcam: Designed phase-mask based thin lensless camera. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 42, 7 (2020), 1618–1629.
- Ayan Chakrabarti. 2016. Learning sensor multiplexing design through back-propagation. *Advances in Neural Information Processing Systems* 29 (2016).
- Praneeth Chakravarthula, Jipeng Sun, Xiao Li, Chenyang Lei, Gene Chou, Mario Bi-jelic, Johannes Froesch, Arka Majumdar, and Felix Heide. 2023. Thin On-Sensor Nanophotonic Array Cameras. *ACM Transactions on Graphics* 42, 6 (2023), 1–18.
- Julie Chang and Gordon Wetzstein. 2019. Deep optics for monocular depth estimation and 3D object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 10193–10202.
- 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* 13, 3 (2018), 220–226.
- Ilya Chugunov, Seung-Hwan Baek, Qiang Fu, Wolfgang Heidrich, and Felix Heide. 2021. Mask-ToF: Learning Microlens Masks for Flying Pixel Correction in Time-of-Flight Imaging. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 9116–9126.
- Shane Colburn, Alan Zhan, and Arka Majumdar. 2018. Metasurface optics for full-color computational imaging. *Science Advances* 4, 2 (2018), eaar2114.
- Blender Online Community. 2025a. *Blender - a 3D modelling and rendering package*. <https://www.blender.org> Blender Foundation, Stichting Blender Foundation, Amsterdam.
- Blender Online Community. 2025b. *Cycles - A Physically Based Production Renderer*. <https://www.cycles-renderer.org/> Blender Foundation, Stichting Blender Foundation, Amsterdam.
- Yunxi Dong, Bowen Zheng, Hang Li, Hong Tang, Huan Zhao, Yi Huang, Sensong An, and Hualiang Zhang. 2024. Achromatic single metalens imaging via deep neural network. *ACS Photonics* 11, 4 (2024), 1645–1656.
- Ahmed H Dorrah and Federico Capasso. 2022. Tunable structured light with flat optics. *Science* 376, 6591 (2022), eabi6860.
- Thomas Eboli, Jean-Michel Morel, and Gabriele Facciolo. 2022. Fast two-step blind optical aberration correction. In *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, 693–708.
- A. Foi, Mejdi Trimeche, V. Katkovnik, and K. Egiazarian. 2008. Practical Poissonian-Gaussian Noise Modeling and Fitting for Single-Image Raw-Data. *IEEE Transactions on Image Processing* 17 (2008), 1737–1754.
- Johannes E Fröch, Praneeth Chakravarthula, Jipeng Sun, Ethan Tseng, Shane Colburn, Alan Zhan, Forrest Miller, Anna Wirth-Singh, Quentin AA Tanguy, Zheyi Han, Karl F. Böhringer, Felix Heide, and Arka Majumdar. 2025. Beating spectral bandwidth limits for large aperture broadband nano-optics. *Nature Communications* 16, 1 (2025), 3025.
- Ying Fu, Tao Zhang, Yinqiang Zheng, Debing Zhang, and Hua Huang. 2020. Joint camera spectral response selection and hyperspectral image recovery. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44, 1 (2020), 256–272.
- Bhargav Ghanekar, Salman Siddique Khan, Pranav Sharma, Shreyas Singh, Vivek Boominathan, Kaushik Mitra, and Ashok Veeraraghavan. 2024. Passive Snapshot Coded Aperture Dual-Pixel RGB-D Imaging. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 25348–25357.
- Joseph W Goodman. 2005. *Introduction to Fourier optics*. Roberts and Company publishers.
- Shi Guo, Zifei Yan, Kai Zhang, Wangmeng Zuo, and Lei Zhang. 2019. Toward convolutional blind denoising of real photographs. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*. 1712–1722.
- Harel Haim, Shay Elmaleh, Raja Giryes, Alex Bronstein, and Emanuel Marom. 2018. Depth Estimation From a Single Image Using Deep Learned Phase Coded Mask. *IEEE Transactions on Computational Imaging* 4 (2018), 298–310.

- Poly Haven. 2025. A curated public 3D asset library. (2025). <https://polyhaven.com/collections>
- Dean S Hazineh, Soon Wei Daniel Lim, ZhuJun Shi, Federico Capasso, Todd Zickler, and Qi Guo. 2022. D-flat: A differentiable flat-optics framework for end-to-end metasurface visual sensor design. *arXiv preprint arXiv:2207.14780* (2022).
- Felix Heide, Qiang Fu, Yifan Peng, and Wolfgang Heidrich. 2016. Encoded diffractive optics for full-spectrum computational imaging. *Scientific Reports* 6 (2016), 33543.
- 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 (SIGGRAPH)* 32, 5 (2013), 1–14.
- Michael Hirsch, Suvrit Sra, Bernhard Schölkopf, and Stefan Harmeling. 2010. Efficient filter flow for space-variant multiframe blind deconvolution. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 607–614.
- Roarke Horstmeyer, Richard Y. Chen, Barbara Kappes, and Benjamin Judkewitz. 2017. Convolutional neural networks that teach microscopes how to image. *ArXiv abs/1709.07223* (2017).
- Michael Iliadis, Leonidas Spinoulas, and Aggelos K Katsaggelos. 2020. Deepbinarymask: Learning a binary mask for video compressive sensing. *Digital Signal Processing* 96 (2020), 102591.
- Wenzel Jakob and Johannes Hanika. 2019. A low-dimensional function space for efficient spectral upsampling. In *Computer Graphics Forum*, Vol. 38. Wiley Online Library, 147–155.
- Yujin Jeon, Eunsue Choi, Youngchan Kim, Yunseong Moon, Khalid Omer, Felix Heide, and Seung-Hwan Baek. 2024. Spectral and Polarization Vision: Spectro-polarimetric Real-world Dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 22098–22108.
- Jun Jiang, Dengyu Liu, Jinwei Gu, and Sabine Süsstrunk. 2013. What is the space of spectral sensitivity functions for digital color cameras?. In *2013 IEEE Workshop on Applications of Computer Vision (WACV)*. IEEE, 168–179.
- Michael Kellman, Emrah Bostan, Michael Chen, and Laura Waller. 2019. Data-Driven Design for Fourier Ptychographic Microscopy. In *Proceedings of the IEEE International Conference on Computational Photography (ICCP)*. IEEE, 1–8.
- Salman Siddique Khan, Varun Sundar, Vivek Boominathan, Ashok Veeraraghavan, and Kaushik Mitra. 2020. Flatnet: Towards photorealistic scene reconstruction from lensless measurements. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2020).
- Mohammadreza Khorasaninejad and Federico Capasso. 2017. Metalenses: Versatile multifunctional photonic components. *Science* 358, 6367 (2017), eaam8100.
- Mohammadreza Khorasaninejad, Wei Ting Chen, Robert C. Devlin, Jaewon Oh, Alexander Y. Zhu, and Federico Capasso. 2016. Metalenses at visible wavelengths: Diffraction-limited focusing and subwavelength resolution imaging. *Science* 352 (2016), 1190–1194.
- Mohammadreza Khorasaninejad, ZhuJun Shi, Alexander Y Zhu, Wei-Ting Chen, Vyshakh Sanjeev, Aun Zaidi, and Federico Capasso. 2017. Achromatic metalens over 60 nm bandwidth in the visible and metalens with reverse chromatic dispersion. *Nano Letters* 17, 3 (2017), 1819–1824.
- Alexander V. Kildishev, Alexandra Boltasseva, and Vladimir M. Shalaev. 2013. Planar Photonics with Metasurfaces. *Science* 339 (2013).
- Youngjin Kim, Taewon Choi, Gun-Yeal Lee, Changhyun Kim, Junseo Bang, Junhyeok Jang, Yoonchan Jeong, and Byoungcho Lee. 2024. Metasurface folded lens system for ultrathin cameras. *Science Advances* 10, 44 (2024), eadr2319.
- Rudolf Kingslake and R Barry Johnson. 2009. *Lens design fundamentals*. Academic Press.
- Amit Kohli, Anastasios N Angelopoulos, David McAllister, Esther Whang, Sixian You, Kyrolos Yanny, Federico M Gasparoli, and Laura Waller. 2022. Ring deconvolution microscopy: An exact solution for spatially-varying aberration correction. *arXiv preprint arXiv:2206.08928* (2022).
- Karl S Kunz and Raymond J Luebbers. 1993. *The finite difference time domain method for electromagnetics*. CRC Press.
- Grace Kuo, Nick Antipa, Ren Ng, and Laura Waller. 2017. DiffuserCam: Diffuser-based lensless cameras. In *Computational Optical Sensing and Imaging*. Optical Society of America, CTu3B–2.
- Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. 2015. Deep learning. *Nature* 521, 7553 (2015), 436–444.
- Yuqi Li, Qiang Fu, and Wolfgang Heidrich. 2021. Multispectral illumination estimation using deep unrolling network. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 2672–2681.
- Dianmin Lin, Pengyu Fan, Erez Hasman, and Mark L. Brongersma. 2014. Dielectric gradient metasurface optical elements. *Science* 345, 6194 (2014), 298–302.
- AW Lohmann and DP Paris. 1965. Space-variant image formation. *Journal of the Optical Society of America* 55, 8 (1965), 1007–1013.
- Ian Mallett and Cem Yuksel. 2019. Spectral Primary Decomposition for Rendering with sRGB Reflectance. *Eurographics Symposium on Rendering - DL-only and Industry Track* (2019), 7 pages. <https://doi.org/10.2312/SR.20191216> Artwork Size: 7 pages ISBN: 9783038680956 Publisher: The Eurographics Association Version Number: 009-015.
- Julio Marco, Quercus Hernandez, Adolfo Muñoz, Yue Dong, Adrián Jarabo, Min H. Kim, Xin Tong, and Diego Gutierrez. 2017. DeepToF: off-the-shelf real-time correction of multipath interference in time-of-flight imaging. *ACM Transactions on Graphics* 36 (2017), 219:1–219:12.
- Julien NP Martel, Lorenz K Mueller, Stephen J Carey, Piotr Dudek, and Gordon Wetzstein. 2020. Neural sensors: Learning pixel exposures for HDR imaging and video compressive sensing with programmable sensors. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 42, 7 (2020), 1642–1653.
- Kyoji Matsushima. 2010. Shifted angular spectrum method for off-axis numerical propagation. *Optics Express* 18, 17 (2010), 18453–18463.
- Johannes Meng, Florian Simon, Johannes Hanika, and Carsten Dachsbacher. 2015. Physically Meaningful Rendering using Tristimulus Colours. *Computer Graphics Forum* 34, 4 (July 2015), 31–40. <https://doi.org/10.1111/cgf.12676>
- Christopher A Metzler, Hayato Ikoma, Yifan Peng, and Gordon Wetzstein. 2020. Deep optics for single-shot high-dynamic-range imaging. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 1375–1385.
- Kristina Monakhova, Kyrolos Yanny, Neeraj Aggarwal, and Laura Waller. 2020. Spectral DiffuserCam: Lensless snapshot hyperspectral imaging with a spectral filter array. *Optica* 7, 10 (2020), 1298–1307.
- Mulun Na, Hector A Jimenez Romero, Xinge Yang, Jonathan Klein, Dominik L Michels, and Wolfgang Heidrich. 2024. End-to-end Optimization of Fluidic Lenses. In *SIGGRAPH Asia 2024 Conference Papers*. 1–10.
- Elias Nehme, Daniel Freedman, Racheli Gordon, Boris Ferdman, Lucien E Weiss, Onit Alalouf, Tal Naor, Reut Orange, Tomer Michaeli, and Yoav Shechtman. 2020. Deep-STORM3D: Dense 3D localization microscopy and PSF design by deep learning. *Nature Methods* 17, 7 (2020), 734–740.
- H. Otsu, M. Yamamoto, and T. Hachisuka. 2018. Reproducing Spectral Reflectances From Tristimulus Colours. *Computer Graphics Forum* 37, 6 (Sept. 2018), 370–381. <https://doi.org/10.1111/cgf.13332>
- Yeongmyeong Park, Youngjin Kim, Changhyun Kim, Gun-Yeal Lee, Hyeongyu Choi, Taewon Choi, Yoonchan Jeong, and Byoungcho Lee. 2025. End-to-End Optimization of Metalens for Broadband and Wide-Angle Imaging. *Advanced Optical Materials* 13, 9 (2025), 2402853.
- Yifan Peng, Qiang Fu, Hadi Amata, Shuochen Su, Felix Heide, and Wolfgang Heidrich. 2015. Computational imaging using lightweight diffractive-refractive optics. *Optics Express* 23 (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 (SIGGRAPH)* 35, 4 (2016), 1–11.
- Yifan Peng, Qilin Sun, Xiong Dun, Gordon Wetzstein, Wolfgang Heidrich, and Felix Heide. 2019. Learned Large Field-of-View Imaging With Thin-Plate Optics. *ACM Transactions on Graphics (SIGGRAPH Asia)* 38, 6 (2019).
- Federico Presutti and Francesco Monticone. 2020. Focusing on bandwidth: Achromatic metalens limits. *Optica* 7, 6 (Jun 2020), 624–631.
- Leonid I Rudin, Stanley Osher, and Emad Fatemi. 1992. Nonlinear total variation based noise removal algorithms. *Physica D: nonlinear phenomena* 60, 1–4 (1992), 259–268.
- Christian J Schuler, Michael Hirsch, Stefan Harmeling, and Bernhard Schölkopf. 2011. Non-stationary correction of optical aberrations. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. IEEE, 659–666.
- Yoav Shechtman, Lucien E Weiss, Adam S. Backer, Maurice Y. Lee, and W E Moerner. 2016. Multicolour localization microscopy by point-spread-function engineering. *Nature Photonics* 10 (2016), 590–594.
- Zheng Shi, Ilya Chugunov, Mario Bijelic, Geoffroi Côté, Jiwoon Yeom, Qiang Fu, Hadi Amata, Wolfgang Heidrich, and Felix Heide. 2024a. Split-Aperture 2-in-1 Computational Cameras. *ACM Transactions on Graphics (TOG)* 43, 4 (2024), 1–19.
- Zheng Shi, Xiong Dun, Haoyu Wei, Siyu Dong, Zhanshan Wang, Xinbin Cheng, Felix Heide, and Yifan Peng. 2024b. Learned Multi-aperture Color-coded Optics for Snapshot Hyperspectral Imaging. *ACM Transactions on Graphics (TOG)* 43, 6 (2024), 1–11.
- Sajan Shrestha, Adam C. Overvig, Ming Ying Lu, Aaron Stein, and Nanfang Yu. 2018. Broadband achromatic dielectric metalenses. *Light: Science & Applications* 7 (2018), 85.
- Vincent Sitzmann, Steven Diamond, Yifan Peng, Xiong Dun, Stephen P. 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 (SIGGRAPH)* 37, 4 (2018), 1–13.
- Grigory Solomatov and Derya Akkaynak. 2023. Spectral Sensitivity Estimation Without a Camera. In *Proceedings of the IEEE International Conference on Computational Photography (ICCP)*. IEEE, 1–12.
- Matheus Souza, Yidan Zheng, Kaizhang Kang, Yogeshwar Nath Mishra, Qiang Fu, and Wolfgang Heidrich. 2024. Latent Space Imaging. *arXiv preprint arXiv:2407.07052* (2024).
- Shuochen Su, Felix Heide, Gordon Wetzstein, and Wolfgang Heidrich. 2018. Deep End-to-End Time-of-Flight Imaging. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 6383–6392.



- Qilin Sun, Ethan Tseng, Qiang Fu, Wolfgang Heidrich, and Felix Heide. 2020. Learning Rank-1 Diffractive Optics for Single-shot High Dynamic Range Imaging. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Qilin Sun, Congli Wang, Qiang Fu, Xiong Dun, and Wolfgang Heidrich. 2021. End-to-end complex lens design with differentiable ray tracing. *ACM Transactions on Graphics (SIGGRAPH)* 40, 4 (2021), 1–13.
- Allen Taflove, Susan C Hagness, and Melinda Piket-May. 2005. Computational electromagnetics: the finite-difference time-domain method. *The Electrical Engineering Handbook* 3, 629–670 (2005), 15.
- Shiyu Tan, Yicheng Wu, Shou-I Yu, and Ashok Veeraraghavan. 2021. Codedstereo: Learned phase masks for large depth-of-field stereo. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 7170–7179.
- Zachary Teed and Jia Deng. 2020. RAFT: Recurrent All-Pairs Field Transforms for Optical Flow. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 402–419.
- Ethan Tseng, Shane Colburn, James Whitehead, Luocheng Huang, Seung-Hwan Baek, Arka Majumdar, and Felix Heide. 2021a. Neural Nano-Optics for High-quality Thin Lens Imaging. *Nature Communications* (December 2021).
- Ethan Tseng, Ali Mosleh, Fahim Mannan, Karl St-Arnaud, Avinash Sharma, Yifan Peng, Alexander Braun, Derek Nowrouzezahrai, Jean-Francois Lalonde, and Felix Heide. 2021b. Differentiable Compound Optics and Processing Pipeline Optimization for End-to-end Camera Design. *ACM Transactions on Graphics (SIGGRAPH)* 40, 4 (2021).
- Mark Van De Ruit and Elmar Eisemann. 2023. Metameric: Spectral Uplifting via Controllable Color Constraints. In *Special Interest Group on Computer Graphics and Interactive Techniques Conference Proceedings*. ACM, Los Angeles CA USA, 1–10. <https://doi.org/10.1145/3588432.3591565>
- 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* 32, 6 (2013), 166.
- Congli Wang, Ni Chen, and Wolfgang Heidrich. 2022. dO: A differentiable engine for Deep Lens design of computational imaging systems. *IEEE Transactions on Computational Imaging* 8 (2022), 905–916.
- Shuming Wang, Pin Chieh Wu, Vin-Cent Su, Yi-Chieh Lai, Mu-Ku Chen, Hsin Yu Kuo, Bo Han Chen, Yu Han Chen, Tzu-Ting Huang, Jung-Hsi Wang, Ray-Ming Lin, Chieh-Hsiung Kuan, Tao Li, Zhen lin Wang, Shining Zhu, and Din Ping Tsai. 2018. A broadband achromatic metalens in the visible. *Nature Nanotechnology* 13 (2018), 227–232.
- Kaixuan Wei, Xiao Li, Johannes Froech, Praneeth Chakravarthula, James Whitehead, Ethan Tseng, Arka Majumdar, and Felix Heide. 2024. Spatially varying nanophotonic neural networks. *Science Advances* 10, 45 (2024), eadp0391.
- Norbert Wiener. 1964. Extrapolation, Interpolation, and Smoothing of Stationary Time Series. (1964).
- Anna Wirth-Singh, Johannes E Fröch, Fan Yang, Louis Martin, Hanyu Zheng, Hualiang Zhang, Quentin T Tanguy, Zhihao Zhou, Luocheng Huang, Demis D John, Biljana Stamenic, Juejun Hu, Tian Gu, and Arka Majumdar. 2025. Wide field of view large aperture meta-doublet eyepiece. *Light: Science & Applications* 14, 1 (2025), 17.
- Jiachen Wu, Hua Zhang, Wenhui Zhang, Guofan Jin, Liangcai Cao, and George Barbastathis. 2020. Single-shot lensless imaging with fresnel zone aperture and incoherent illumination. *Light: Science & Applications* 9, 1 (2020), 53.
- Yicheng Wu, Vivek Boominathan, Huaijin Chen, Aswin Sankaranarayanan, and Ashok Veeraraghavan. 2019. PhaseCam3D — Learning Phase Masks for Passive Single View Depth Estimation. In *Proceedings of the IEEE International Conference on Computational Photography (ICCP)*. IEEE, 1–12.
- Xinge Yang, Qiang Fu, and Wolfgang Heidrich. 2024a. Curriculum learning for ab initio deep learned refractive optics. *Nature Communications* 15, 1 (2024), 6572.
- Xinge Yang, Matheus Souza, Kunyi Wang, Praneeth Chakravarthula, Qiang Fu, and Wolfgang Heidrich. 2024b. End-to-end hybrid refractive-diffractive lens design with differentiable ray-wave model. In *SIGGRAPH Asia 2024 Conference Papers*. 1–11.
- Leonid P Yaroslavsky and H John Caulfield. 1994. Deconvolution of multiple images of the same object. *Applied Optics* 33, 11 (1994), 2157–2162.
- Michitaka Yoshida, Akihiko Torii, Masatoshi Okutomi, Kenta Endo, Yukinobu Sugiyama, Rin-ichiro Taniguchi, and Hajime Nagahara. 2018. Joint optimization for compressive video sensing and reconstruction under hardware constraints. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 634–649.
- Nanfeng Yu and Federico Capasso. 2014. Flat optics with designer metasurfaces. *Nature Materials* 13 (2014), 139–150.
- Nanfeng Yu, Patrice Genevet, Mikhail A. Kats, Francesco Aieta, Jean-Philippe Tetienne, Federico Capasso, and Zeno Gaburro. 2011. Light Propagation with Phase Discontinuities: Generalized Laws of Reflection and Refraction. *Science* 334 (2011), 333–337.
- Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Yanli Zhao, Andrew Gu, Rohan Varma, Liang Luo, Chien-Chin Huang, Min Xu, Less Wright, Hamid Shojanazeri, Mylène Ott, Sam Shleifer, Alban Desmaison, Can Balioglu, Pritam Damania, Bernard Nguyen, Geeta Chauhan, Yuchen Hao, Ajit Mathews, and Shen Li. 2023. PyTorch FSDP: Experiences on Scaling Fully Sharded Data Parallel. *Proceedings of the VLDB Endowment* 16, 12 (Aug. 2023), 3848–3860.