Large-Area Fabrication-Aware Computational Diffractive Optics

KAIXUAN WEI*, King Abdullah University of Science and Technology, Saudi Arabia
HECTOR A. JIMENEZ-ROMERO*, King Abdullah University of Science and Technology, Saudi Arabia
HADI AMATA*, King Abdullah University of Science and Technology, Saudi Arabia
JIPENG SUN, Princeton University, United States of America
QIANG FU, King Abdullah University of Science and Technology, Saudi Arabia
FELIX HEIDE, Princeton University, United States of America
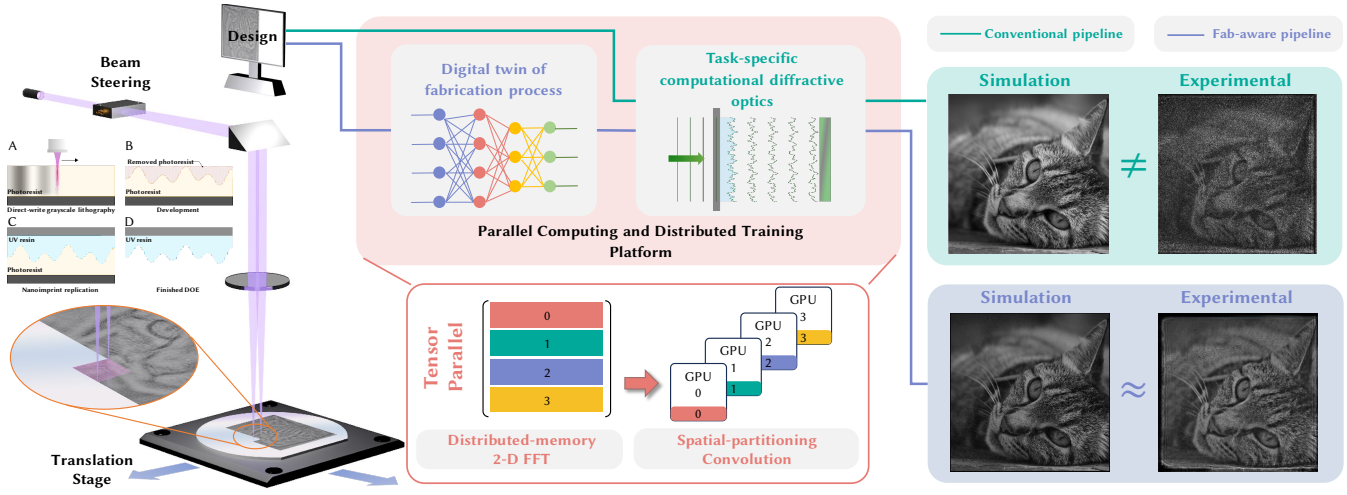WOLFGANG HEIDRICH, King Abdullah University of Science and Technology, Saudi Arabia

Fig. 1. **Fabrication-aware, End-to-end Optimization for Large-area Diffractive Optics.** We propose a fabrication-aware design method for diffractive optical elements fabricated by (left) direct-write grayscale lithography with nanoimprint replication (see the inset figures A-D for a step-by-step illustration) suited for inexpensive mass production. Enabled by tensor-parallel computing routines, our method jointly considers the fabrication 3-D geometry deformation and the downstream task-specific computational diffractive optics design. This combination of techniques allows for experimental findings with favorable quality to all tested existing methods, specifically closing the design-to-manufacturing gap in existing approaches.

Differentiable optics, as an emerging paradigm that jointly optimizes optics and (optional) image processing algorithms, has made many innovative optical designs possible across a broad range of imaging and display applications. Many of these systems utilize diffractive optical components for holography, PSF engineering, or wavefront shaping. Existing approaches have, however, mostly remained limited to laboratory prototypes, owing to a large quality gap between simulation and manufactured devices.

We aim at lifting the fundamental technical barriers to the practical use of learned diffractive optical systems. To this end, we propose a fabrication-aware design pipeline for diffractive optics fabricated by direct-write grayscale lithography followed by replication with nano-imprinting, which is directly suited for inexpensive mass-production of large area designs. We propose a super-resolved neural lithography model that can accurately predict the

3D geometry generated by the fabrication process. This model can be seamlessly integrated into existing differentiable optics frameworks, enabling fabrication-aware, end-to-end optimization of computational optical systems. To tackle the computational challenges, we also devise tensor-parallel compute framework centered on distributing large-scale FFT computation across many GPUs.

As such, we demonstrate large scale diffractive optics designs up to $32.16\,\text{mm} \times 21.44\,\text{mm}$, simulated on grids of up to 128,640 by 85,760 feature points. We find adequate agreement between simulation and fabricated prototypes for applications such as holography and PSF engineering. We also achieve high image quality from an imaging system comprised only of a single diffractive optical element, with images processed only by a one-step inverse filter utilizing the simulation PSF. We believe our findings lift the fabrication limitations for real-world applications of diffractive optics and differentiable optical design.

CCS Concepts: • **Computing methodologies → Modeling methodologies**; • **Applied computing → Engineering**.

Additional Key Words and Phrases: Computational optics, computational imaging, computational fabrication

*indicates joint first authorship.

Authors' addresses: ✉Kaixuan Wei, kaixuan.wei@kaust.edu.sa, KAUST, Saudi Arabia; Hector A. Jimenez-Romero, hector.jimenezromero@kaust.edu.sa, KAUST, Saudi Arabia; Hadi Amata, hadi.amata@kaust.edu.sa, KAUST, Saudi Arabia; Jipeng Sun, jipeng.sun@princeton.edu, Princeton University, USA; Qiang Fu, qiang.fu@kaust.edu.sa, KAUST, Saudi Arabia; Felix Heide, fheide@princeton.edu, Princeton University, USA; Wolfgang Heidrich, wolfgang.heidrich@kaust.edu.sa, KAUST, Saudi Arabia.

1 INTRODUCTION

Over the last decades, rapid advances in computational power, and photodetection devices have enabled the emergence of computational imaging and optics as a new way for designing optical systems [Bhandari et al. 2022; Mait et al. 2018]. By co-designing optics and image processing algorithms, these computational systems can produce new forms of visual information, which are otherwise difficult to capture by traditional optical systems [Nayar 2006]. Notably, diffractive optical elements (DOEs) are particularly suited for computational optics as they can encode complex optical functions, such as an arbitrary phase modulation, that are hard to achieve with refractive optics. As a result, remarkable capabilities such as snapshot high-dynamic-range [Sun et al. 2020], extended depth-of-field [Nehme et al. 2020], hyperspectral [Shi et al. 2024b] and monocular depth imaging [Ikoma et al. 2021], have been demonstrated in DOE-based computational systems, where the complex DOE design problems are addressed by the differentiable modeling of the wave optics system [Sitzmann et al. 2018], and back-propagation-based optimization popularized by the advent of deep learning (DL) era. The resulting paradigm, coined deep optics or differentiable optics, has shown to be a versatile tool to optimize computational diffractive, refractive [Sun et al. 2021; Yang et al. 2024a] or hybrid refractive-diffractive [Yang et al. 2024b] optical systems, not just for imaging but for near-eye display [Chakravarthula et al. 2019; Peng et al. 2020], with extraordinary task-specific performance.

Despite the seemingly striking results, these advanced learned diffractive optical systems have largely remained limited to laboratory prototypes, owing to the large quality discrepancies between simulation and fabricated devices [Shi et al. 2024a; Zheng et al. 2023]. Unlike refractive optical elements with smooth surfaces, DOEs rely on precise control of micron-sized structures that can quickly vary across the design space, thereby requiring sophisticated micro/nano-fabrication technology.

To understand the reason for the discrepancy between simulated designs and fabricated prototypes, we need to consider a specific fabrication process in some detail. Here we explain grayscale lithography with direct laser writing (see Fig. 1, left) [Grushina 2019]; other lithographic method such a two-photon polymerization [Wang et al. 2023] differ in the details but require similar high-level considerations. In direct laser writing, a laser beam scans across a wafer covered in thin layer photoresist in a 2D grid. Modulation of the beam intensity is used to create a spatially varying exposure map (A). The first notable effect in this process is an optical blur of the specified design, which can be modeled as the convolution of the design pattern with the point spread function of the laser beam and can, for example, result in corner rounding for small rectangular features. Next, the local exposure level induces a local chemical change in the photoresist, which can be thought of as a nonlinear local transfer function. The next step (B) is the development of the exposed resin, in which the developer decomposes the photoresist, where the rate of decomposition depends on the local exposure received by the resist. This is a complicated 3D chemical interaction that finally reveals a height field structure. Nanoimprinting can be used to transfer this shape into a UV resin (C) to form the final DOE (D). Further shape distortions are possible in this step, for example, due to volume shrinkage of the resin during curing. For mass production, only the final imprinting step needs to be repeated for each copy.

Inherent to the fabrication process is that the fabricated DOE will differ from the target shape at scales *smaller than the intended feature size*. To accurately predict the optical performance of a DOE, we not only need to have a precise digital twin of the fabrication process, but also need to conduct the optical simulation at substantially super-resolved resolutions compared to the target feature size.

A related source of error is that most inverse design pipelines for diffractive optics work on drastically *undersampled grids* out of computational necessity. Most early DOE design works use simulation grids where one grid point equals one DOE feature. More recently, simulations have started using a 2× finer grid (*e.g.*, [Yang et al. 2024b]). However, according to the Nyquist limit any sample grid can only represent sinusoidal components up to twice the grid pitch, whereas many holography and DOE designs rely on sharp step edges between neighboring features. We experimentally demonstrate this problem in Section 3.2 and Figure 3 by performing simulations on the original feature grid resolution and on an 8× super-resolved grid with nearest neighbor upsampling to force box-shaped features, highlighting drastic simulation differences even without accounting for fabrication limitations.

Several strategies for dealing with the large simulation gap have been introduced and are in common use: commercial product design often relies on a tedious iterative process between design, fabrication, and measurement [Jang et al. 2020]. In the holographic display community, camera-in-the-loop systems have recently become popular [Choi et al. 2022; Kavaklı et al. 2021; Peng et al. 2020], whereas in end-to-end designed computational imaging systems, it is common practice to fine-tune the computational module with measurements from the as-fabricated prototype (*e.g.*, [Chakravarthula et al. 2023; Peng et al. 2019; Shi et al. 2024a] etc.). However, closing the loop in this fashion is expensive and may not be practical beyond lab prototypes.

In this work, we instead desire a fabrication-aware design process that can accurately predict the final system in open-loop simulation. The core of this approach is a digital twin for the fabrication process in the form of a neural lithography model for direct-write grayscale lithography followed by nanoimprinting. This model can easily be incorporated into any existing diffractive design pipeline. Unlike the recent neural lithography work by Zheng et al. [2023], we directly target a fabrication process that is suitable for both large-scale designs as well as mass fabrication. To facilitate the resulting large-scale design processes and tackle the challenges posed by large memory requirements, we also devise a parallelization toolbox for distributing large FFTs across many GPUs. To demonstrate the efficacy and generalization capabilities of our approach, we validate the method for several applications:

- Computer-generated holograms produced open-loop (i.e. without camera in the loop) with substantially reduced noise and speckle, and excellent agreement between simulation and prototype. This includes a design of up to 32.16 mm × 21.44 mm in size, simulated on a grid of 128,640 × 85,760 feature points on 16 A-100 GPUs.

- We conduct a beam shaping experiment on the example of splitting an incident beam into a regular grid of output beams with controlled intensities, which can be used for downstream tasks such as laser material processing [Kuang et al. 2013] and 3-D sensing [Yuan et al. 2021] or visual vibrometry [Zhang et al. 2023b].
- We design a computational camera comprised of a single DOE and off-the-shelf sensor for broadband color imaging, again showing outstanding agreement between simulation and prototype, to the point that high image quality can be achieved by a one-step inverse filter utilizing the simulated PSF of the system.

The last application confirms the ability to not only eliminate fine-tuning of the image restoration module based on measured characterization of the prototype, but also demonstrates that, with a design process that includes an accurate model of the manufacturing process, image restoration does not necessarily require heavy deep neural networks but can utilize more lightweight architectures that are compatible with the computational resources of edge devices. We believe that this, by itself, is a major step forward in improving the practicality of diffractive end-to-end design beyond lab prototypes. Our code is publicly available at https://github.com/Vandermode/LAFA

2 RELATED WORK

In the following, we review work on diffractive optical elements, computational lithography, and distributed training frameworks related to the method proposed in this paper.

2.1 Diffractive Optical Elements

In 1948, Dennis Gabor introduced the optical holography, the first physical realization of DOEs through interference [Gabor 1948]. Since then, various forms of DOEs have been proposed, such as computer-generated holograms [Brown and Lohmann 1966], binary gratings [Collischon et al. 1994] and kinoform lens [Lesem et al. 1969], to name a few—see [Zhang et al. 2023a] for a comprehensive review of long history of DOEs. However, due to intrinsic difficulty in designing and manufacturing DOEs, their uses remain largely limited to well-controlled laboratory settings. Over the last decade, advances in computational imaging and the increasing capabilities of nano-fabrication technology have allowed researchers to revisit these conventional DOEs, as powerful optical encoders in hardware-software co-designed computational systems [Asif et al. 2017; Boominathan et al. 2016; Heide et al. 2016; Peng et al. 2015, 2016; Sitzmann et al. 2018; Wu et al. 2019]. These methods have been successful in a wide array of applications, including full-spectral color imaging [Heide et al. 2016; Peng et al. 2016], extended depth-of-field imaging [Tan et al. 2021], compressive lensless imaging [Asif et al. 2017; Boominathan et al. 2020], single-shot hyperspectral [Shi et al. 2024b] and depth imaging [Baek et al. 2021; Shi et al. 2024a], high-dynamic-range imaging [Metzler et al. 2020; Sun et al. 2020], seeing through obstructions [Shi et al. 2022], ultra-wide-angle holographic display [Tseng et al. 2024], and computer vision tasks [Wei et al. 2024]. Although successful, two fundamental technical obstacles remain: 1) the gap from design to manufacturing and 2) the challenge to design large-area devices at high fidelity, which are essential to compete with the widely used refractive optics.

2.2 Nano-Fabrication and Computational Lithography

Optical lithography [Dill 1975], one of the key driving forces behind Moore's Law, has made tremendous progress over the last half a century [Mack 2011; Moore 1998]. In parallel to resolution improvement via shorter-wavelength illuminator and higher numerical aperture (NA) optics [Bruning 2007], a body of work explores algorithmic resolution improvement—optical proximity correction (OPC) [Fung Chen et al. 1997], inverse lithography [Cecil et al. 2022; Pang 2021], or computational lithography [Lam and Wong 2009; Ma and Arce 2011]. These existing works are designed for the mask-based photolithography process tailored for 2-D binary patterns representative of integrated circuits, which unfortunately are not directly applicable to the fabrication of 3-D (2.5D) DOEs with continuously varying height profiles.

Since the advent of commercially available 3-D micro/nano patterning techniques, such as the two-photon polymerization (TPP) lithography [Wang et al. 2024, 2023] and the direct-write grayscale lithography [Grushina 2019], a number of works focused on the physical process modeling [Guney and Fedder 2016; Onanuga 2019; Saha et al. 2017] and thereby the structure prediction and precompensation [Chevalier et al. 2021; Lang et al. 2022; Wang et al. 2020] for 3-D micro/nano fabrication. These physical simulator-based approaches, however, largely rely on heuristics and iterative trial-and-error to precompensate the design. Recently, Zheng et al. [2023] proposed neural lithography, a differentiable neural network (NN)-based fabrication simulator that enables joint optical design and fabrication correction end-to-end, automatically guaranteeing manufacturability. However, their approach is limited to low-throughput TPP lithography for micron-sized DOE patterns. Most recently, a concurrent work to ours [Xu et al. 2025] proposed a differentiable model-based physical simulator for direct-write grayscale lithography. In contrast to our work with large-area (centimeter-scale) devices as goal, they do not address the replication challenge [Barcelo and Li 2016] while are limited in moderate-sized (millimeter-scale) devices. We also demonstrate that our data-driven neural model is more accurate than their simulation-based framework.

2.3 Parallel Computing and Distributed Training in Deep Learning

Our work also takes inspiration from recent advances in parallel computing and distributed training infrastructure [Brown et al. 2020; Narayanan et al. 2021; Radford et al. 2019]. The most common parallelization strategy is data parallelism [Hillis and Steele Jr 1986], which distributes the data across different computing nodes and operate on the data in parallel. However, this technique has a fundamental limitation in the model size it can tackle—the model must fit entirely on one worker. With the increasing size and complexity, NNs have approached the memory capacity of modern hardware accelerators. To overcome this bottleneck, *model parallelism* [Shoeybi et al. 2020] has been proposed for training billion-parameter-scale LLMs, which includes pipeline parallelism [Huang et al. 2019] and more general *tensor parallelism* [Narayanan et al. 2021]. Pipeline

Design layout (gray value)   DOE fabrication and neural lithography model   Super-resolved height profile as-fabricated   Task-specific computational diffractive optics model
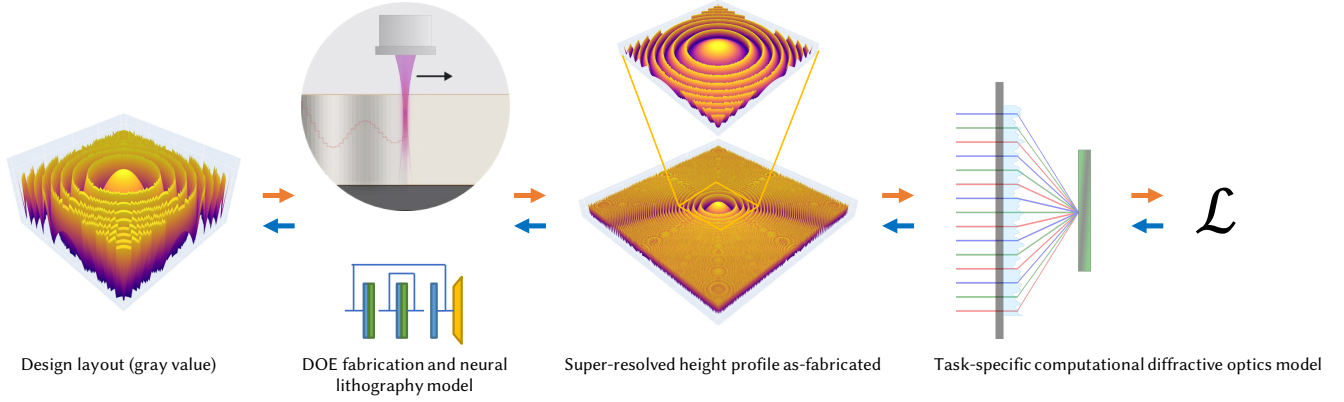
Fig. 2. **Fabrication-aware Image Formation Model.** We illustrate the proposed image formation model (see text) for computational diffractive optics. With this differentiable model in hand, we optimize design layouts (*i.e.*, inputs of the lithography machine) end-to-end informed by the proposed super-resolved neural lithography model, via backpropagation.

parallelism splits the NN pipeline into multiple stages across devices at the expense of the bubble overhead; Tensor parallelism is specialized for distributing specific atomic operators, such as the general matrix multiplication widely used in the Transformer block [Vaswani et al. 2017]. For computational optics, data parallelism supports scaling of wavelengths, incident angles, propagation distances, but cannot realize designs of large-area/scale DOEs discretized with billion-pixel-scale huge arrays [Sun et al. 2025]. As such, we leverage the tensor parallelism and devise the specialized computing routines for the large-area wave-optics model. This includes differentiable distributed-memory (D^2) FFT for free-space wave propagation as well as spatial-partitioning convolutional neural networks as the neural lithography model. Although the basic idea of distributed-memory FFT can be traced back to decades ago in the high-performance computing community [Gupta and Kumar 2002] with implementations in CPU [Frigo and Johnson 2005; Pippig 2013], GPU [Ayala et al. 2022; Gholami et al. 2016] or TPU [Lu et al. 2021] clusters based on Message Passing Interface standard (MPI) [Snir 1998], none of these methods have been designed for differentiable wave optics simulation. Our implementation follows the generalized single program multiple data (GSPMD) programming model [Shazeer et al. 2018; Xu et al. 2021], and thereby flexibly supports any tensor-dimensions to be split across any dimensions of a multi-dimensional mesh of processors—realizing arbitrary hybrid data and model (tensor) parallelism without painful code rewriting.

3 FABRICATION-AWARE END-TO-END DESIGN OF DIFFRACTIVE OPTICAL SYSTEMS

In this section, we first introduce our fabrication-aware image formation model for computational diffractive optics (Fig. 2, Section 3.1). Then, we describe the proposed super-resolved neural lithography model as a differentiable digital twin of the fabrication process (Section 3.2). The proposed computational tools for performing this large scale simulation and design tasks, *i.e.*, distributed-memory FFT and tensor-parallel convolution, for simulating and optimizing large-area/scale wave optics are described in Section 4.

3.1 Fabrication-aware Image Formation Model

We model the DOE in our forward model as a phase profile Φ_λ (with respect to a nominal wavelength λ) or, equivalently, as a height map h, which the relationship

$$\Phi_\lambda(x, y) = \frac{2\pi}{\lambda}(n_\lambda - 1)h(x, y), \tag{1}$$

where n_λ is the wavelength-dependent refractive index of the DOE material (such as resins). Given an incident wave, such as a plane or spherical wave field, $E_\lambda^{\text{in}}(x, y)$, the modulated wavefront at the DOE plane is given by

$$E_\lambda^{\text{doe}}(x, y) = E_\lambda^{\text{in}}(x, y)e^{i\Phi_\lambda(x, y)}. \tag{2}$$

Then, the destination field E_λ^{dest} in the sensor plane is given via the Rayleigh-Sommerfeld diffraction integral [Goodman 2005], implemented by a numerical diffraction propagation method such as the angular spectrum method (ASM) [Matsushima 2010; Matsushima and Shimobaba 2009; Ritter 2014; Zhang et al. 2020], *i.e.*,

$$E_\lambda^{\text{dest}} = \mathcal{F}^{-1}\left\{\mathcal{F}\left\{E_\lambda^{\text{doe}}\right\} \otimes \text{H}_\lambda\right\}, \tag{3}$$

where $\mathcal{F}\{\cdot\}$ represents the fast Fourier transform (FFT), $\otimes$ denotes the Hadamard (elementwise) product, H_λ is the transfer function associated with the propagation model. Finally, the task-specific optimization of the diffractive optical system is posed as

$$h^* = \underset{h}{\arg\min} \sum_\lambda \mathcal{L}_p\left(\|E_\lambda^{\text{dest}}(h)\|^2\right), \tag{4}$$

where $\mathcal{L}_p$ is a task-specific penalty/loss function defined using the intensity $I_\lambda = \|E_\lambda^{\text{dest}}\|^2$ of the destination field, *i.e.*, point spread functions (PSF) for imaging/sensing, or coherent hologram images for holography [Chakravarthula et al. 2019]. Note that the overall optimization objective may have additional parameters such as wavelengths, incident angles, object distances, and other propagation-related parameters. These can result in a large number of simulations according to Eq.(3). Training data and reconstruction methods can
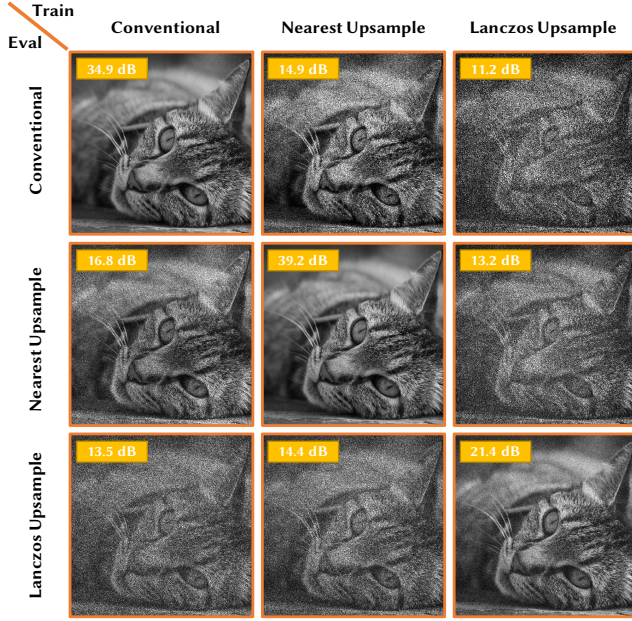
Fig. 3. **Toy Example of the "Fabrication Interpolation Kernel"**. We train and evaluate DOEs for computer-generated 2-D hologram under different settings, including 1) the conventional design at 2 μm-spacing grid; 2) design with nearest upsampling in DOE plane at 250 nm-spacing grid (*i.e.*, 8× upsampling) that meets the $\frac{\lambda}{2}$-spacing requirement of the Nyquist sampling theorem; 3) design with Lanczos upsampling in the DOE plane at 250 nm-spacing grid. Resulting PSNRs are shown in the top-left corner for each image. A comprehensive ablation study is provided in the Supplementary Material.

also be incorporated for joint optimization of optics and image processing [Sitzmann et al. 2018]. Here, we only include the wavelength λ dependency in (4) for simplicity without loss of generality.

The aforementioned image formation model and the derived optimization objective are general enough to cover a broad array of works in computational diffractive optics [Shi et al. 2024a]. However, most existing work largely ignores the manufacturing process and assume the designed DOE can be fabricated as it is, presuming a perfect indentity mapping between design layouts and manufactured devices. In reality, due to the sophisticated photolithography process, 3-D optical proximity effects as well as the complex photochemical interaction render significant deviations from design to manufacturing. We model these deviations with a lithography model $\mathcal{G}\{\cdot\}$, akin to [Zheng et al. 2023], a surrogate of the fabrication process, which maps the design layout l to the device parameterized by $h = \mathcal{G}\{l\}$. As such, the objective (4) is reformulated as

$$l^* = \underset{l}{\arg\min} \sum_{\lambda} \mathcal{L}_p \left(\|E_{\lambda}^{\text{dest}} \left(\mathcal{G}\{l\} \right) \|^2 \right), \tag{5}$$

which can be solved by back-propagation and gradient-based optimization [Kingma 2014] if all components including the lithography model $\mathcal{G}$ are implemented as differentiable operators.

3.2 Super-resolved Neural Lithography Model

Next, we motivate core concepts of the method with a toy example that demonstrates the need for fabrication-aware optimization. Then we introduce the fabrication pipeline towards large-area DOE device manufacturing under the mass-production-ready setting. Finally, we detail the calibration methods to build the proposed neural lithography model, including the contrast curve calibration and neural lithography learning.

Toy Example. A 3.6×3.6 mm^2 DOE with a feature size of 2 μm is optimized to generate a 2-D hologram image under coherent (520.6 nm wavelength) laser illumination. We use the Adam optimizer [Kingma 2014] to solve (4), where the penalty $\mathcal{L}_p$ is defined as a combination of a scale-invariant mean square error loss $\mathcal{L}_{\text{si-mse}}$ and an energy regularization $\mathcal{L}_{\text{energy}}$ [1] as

$$\mathcal{L}_p = \mathcal{L}_{\text{si-mse}} + \beta \mathcal{L}_{\text{energy}}, \tag{6}$$

where β is empirically set as 5×10^{-3}. The DOE height map is randomly initialized, and the learning rate is initially set as 10^{-2}, which follows a cosine schedule that decays to zero in 3000 iterations. Such an optimization leads to almost perfect 2-D hologram reconstruction (34.9 dB in PSNR) evaluated under this conventional setting (see the left-top image in Fig. 3). However, as explained in the introduction, the optical system simulation at 2 μm-spacing grid violates the Nyquist sampling theorem [Smith 1999], which inevitably results in aliasing artifacts degrading performance in reality. The zero-order interpolation, *i.e.*, nearest upsampling is typically employed [Kuo et al. 2023; Tseng et al. 2024] to alleviate this issue, but this assumes perfect flat feature structure can be reliably manufactured. Other smooth/low-pass interpolation kernels such as the Lanczos-3 kernel [Getreuer 2011] remain coarse approximations of the *de-facto* "fabrication kernel". Fig. 3 summarizes the results of DOEs optimized and evaluated at combinations of different settings, suggesting that 1) hologram can only be well reconstructed when train and evaluate under the same setting; 2) interpolation kernels play a critical role in high-frequency hologram feature construction, where a mismatched kernel can result in drastic quality decline. As such, this simple yet informative experiment emphasizes the importance of the "fabrication kernel" and thus motivates the need of fabrication-aware optimization.

Fabrication Pipeline. Our fabrication pipeline utilizes the advanced direct-write grayscale lithography [Grushina 2019] followed by replication with nanoimprint lithography [Barcelo and Li 2016], with a step-by-step illustration in Fig. 1 (insets A-D). Initially, a Soda-lime substrate is coated with a positive AZ® 4562 photoresist. A Heidelberg Instruments DWL 66+ mask writer is used in the direct-write grayscale lithography stage. A laser beam, modulated to vary intensity, selectively exposes the photoresist to create a continuous, three-dimensional relief pattern corresponding to the DOE design. The exposed photoresist is then developed, removing material proportional to the exposure dose, resulting in a smooth, multilevel surface profile. The resulting master template in the resist is then used in nanoimprint lithography: a thin layer of ultraviolet (UV)-curable resin is applied to a new substrate, and the master is pressed

---

[1]Detailed definitions of the loss functions are given in the Supplementary Material
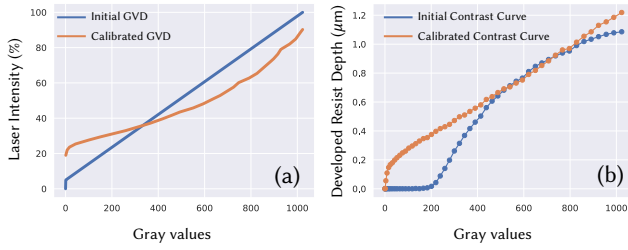
Fig. 4. **Contrast Curve Calibration.** (a) the initial and calibrated gray value distribution (GVD) for intensity control of the direct-write grayscale lithography machine. A nonlinear mapping between gray values (1024 levels) and the laser intensity is utilized after calibration. (b) the initial and calibrated contrast curve of the photoresist. After calibration, we find the developed resist depth is almost perfectly linearly proportional to the gray values.
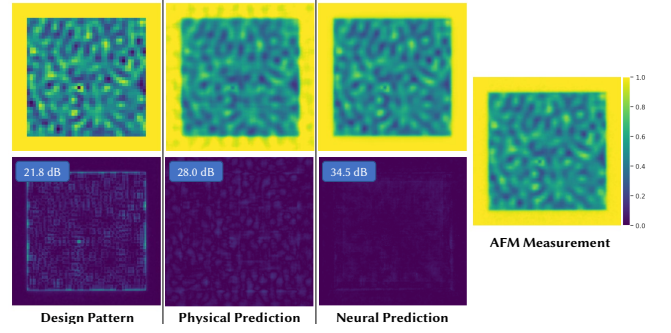
Fig. 5. **Lithography Model Evaluation**. We report here a design pattern and its corresponding AFM measurement from the constructed evalset, along with lithography model predictions—the modulation transfer function-based physical model and the proposed neural lithography model. The error maps (with respect to the AFM measurement ground truth) are also annotated with the associated PSNR values, validating the proposed model predictions.

into it to transfer the pattern. UV light cures the resin, solidifying the replicated DOE structure. The master is then released, leaving a high-fidelity replica. This process enables scalable production of large-area DOEs.

This fabrication pipeline offers several advantages over the conventional multi-level binary-mask-based photolithography process, which was widely used in DOE manufacturing in literature [Khonina et al. 2024]. Unlike the binary-mask approach, which requires multiple photolithography steps with separate masks for discrete height levels, grayscale lithography uses a single laser exposure to create continuous 3-D relief profiles, significantly reducing process complexity and eliminating alignment errors. The simplified process allows for faster prototyping and design iterations, shortening the prototyping cycle from one week to a few hours for a centimeter-scale DOE design, according to our experience in manufacturing.

Next, we describe a two-stage calibration for the proposed super-resolved neural lithography model, a differentiable digital twin of the fabrication process.

Contrast Curve Calibration. The input layout to the direct-write grayscale lithography system is a pixelated grayscale image with discrete 1024 levels (10 bits), whose values are monotonically mapped to relative laser intensities (0-100%) for exposure control. This mapping, coined gray value distribution (GVD) is realized via a lookup table (LUT) that is by default a linear mapping with a small offset (See blue curve in Fig. 4 (a)). With pre-determined laser power and development time[2], the contrast curve, delineating the relationship between gray values and developed resist depths, can be obtained by creating and measuring a test pattern, which consists of a series of uniform patches, each assigning a incrementally increasing gray value. We use a test pattern with 7×7 uniform patches, and the developed structure is then measured by a Zygo optical profilometer (NewView 7300). As shown in Fig. 4 (b), the default GVD leads to a nonlinear contrast curve owing to the nonlinear photo-response

---

[2]Laser power and development time should be pre-determined via a trial-and-error process in order to produce desirable maximum depth, corresponding to 2π phase modulation at nominal 550-nm wavelength, at maximum laser intensity

Table 1. Quantitative evaluation of two lithography models on the collected evalset. Our neural lithography model does not suffer from overfitting, while outperforming the physical model by a large margin.

	Physical [Xu et al. 2025]	Neural (Ours)
PSNR ↑	27.36 dB	35.20 dB
NRMSE ↓	7.27 %	2.45 %

of the photoresist. In practice, a linear contrast curve is always preferred as it allows for the maximum dynamic range of the design space, and simplifies the downstream neural lithography learning. By numerically inverting the measured contrast curve with linear interpolation, we obtain a new calibrated GVD that yields an almost linear contrast curve (orange one in Fig. 4 (b)) by updating the system LUT accordingly. Some uncorrected nonlinearities remains in the low-end of the calibrated contrast curve, as a result, we exclude these gray values, and only use 24 to 1023 gray values (1000 levels in total) in design.

Neural Lithography Learning. After contrast curve calibration, we design and fabricate another set of calibration patterns (spatially arranged in a single wafer) to construct the dataset for training and evaluation of the neural lithography model. Similar to prior work [Zheng et al. 2023], we randomly generate a set of 2-D patterns following a uniform distribution. A low-pass filter is then applied in the Fourier domain to limit the high-frequency components. 2×10 patterns in total are created with incrementally decreasing maximum cutoff frequency, each of which features $40\,\mu m^2$ area discretized at $1\,\mu m$-spacing grid (*c.f.,* Fig. 5). Once fabricated, the resulting sample is then measured using an atomic force microscope (AFM) [Giessibl 2003] to obtain precise 3-D profiles of the fabricated patterns, which can be used as ground truth for training and evaluating the neural lithography model. For each pattern we scan an area of $50\,\mu m \times 50\,\mu m$, with 256×256 sampling points, resulting in approximately 200 nm sampling resolution. To cope with the AFM imaging artifacts inherent in the measurement process [Ricci and Braga
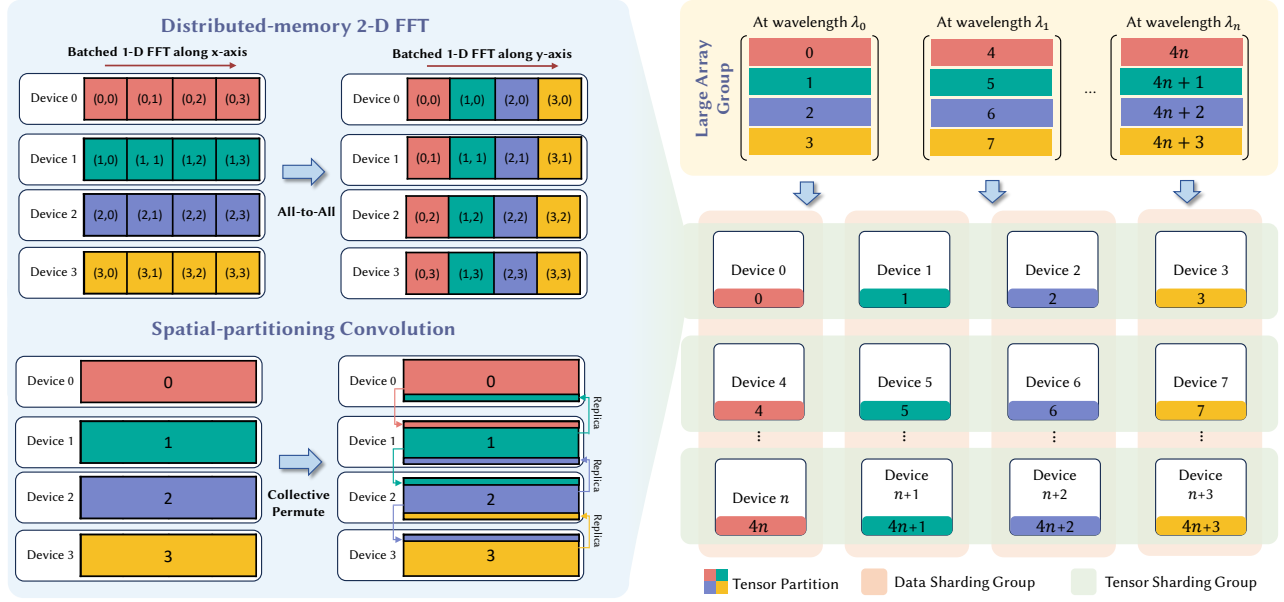
Fig. 6. **GSPMD-based Distributed Computing Framework** tailored for large-area fabrication-aware diffractive optics. We illustrate the distributed computation of the proposed D^2FFT (top-left) and the spatial-partitioning convolution (bottom-left) leveraging tensor parallelism. The (GPU) processors can be arranged into a multi-dimensional mesh to enable arbitrary combinations of hybrid data and tensor parallelism.

2004], we measure each pattern twice in two orthogonal scanning directions. The resulting raw measurements are then preprocessed, registered and fused to obtain final 3-D profiles with high fidelity.

We note the AFM measurement process is time-consuming and labor-intensive[3]—each sample might require repeated measurements due to unpredictable random errors occurring during the process, for example, the AFM probe tip might hit tiny dust atop the sample surface on occasion, leading to a failed measurement. Fortunately, we empirically find that only 10 high-quality data pairs are sufficient to learn a robust neural lithography model with good generalizability[4].

We split the collected 20 data pairs into training and evaluation sets, each of which has 10 data pairs. Then we construct a small yet effective convolutional neural network (CNN) to learn the mapping from design layouts (at 1 μm-spacing grid) to their corresponding AFM measurements (200 nm-spacing grid), which amounts to 5× super resolution. As shown in Fig. 2, the CNN consists of a few convolutional layers (2 to 4 layers) with ReLU nonlinearity followed by a pixel shuffle layer [Shi et al. 2016] for upsampling. Skip connections are also employed to prevent the gradient vanishing problem [He et al. 2016], which is especially important since the gradient flow must be appropriately propagated to the model input (design layout) for fabrication-aware optimization.

We train the CNN as our fabrication surrogate using the Adam optimizer with an initial learning rate of 10^{-3} that decays to zero in 3000 iterations following a cosine schedule. Stochastic optimization

with mini-batch size of 5 and random data augmentation such as horizontal/vertical flipping and rotations are adopted for regularizing model learning to avoid overfitting. We also calibrate a modulation transfer function (MTF) based physical model [Xu et al. 2025] using the collected trainset, by fitting a MTF that transfers nearest upsampled design patterns to their AFM measurements in Fourier domain. Qualitative and quantitative evaluations of the fitted lithography models are provided in Fig. 5 and Table 1 respectively, suggesting the effectiveness of the proposed neural lithography model for capturing structure-dependent "fabrication interpolation kernel" which is otherwise difficult to be modeled by the MTF-based linear physical model. We note that more complex NN architectures like the one used in [Zheng et al. 2023] (with an extra upsampling head) can be employed, which may achieve comparable or better results compared to the proposed simple CNN. However, we find that further reducing testing errors is meaningless considering the intrinsic random variations in both fabrication and measurement processes.

4 PARALLEL COMPUTING FOR LARGE-SCALE WAVE SIMULATION

The design framework detailed in the previous section introduces very large simulation grids that can accurately represent small-scale DOE features at wavelength scale. Unfortunately, while accurate, this pipeline also drastically increases the memory requirements for forward simulation and especially for inverse design, rendering large-scale design tasks infeasible on existing single-GPU pipelines due to the limited amount of GPU memory available.

---

[3]Each scanning of a single pattern takes more than 20 minutes and the whole process cannot be fully automated as random errors might occur during the scanning.
[4]At the early development stage of this work, we collected more AFM measurements to train the neural lithography model, but later found it was redundant and unnecessary.

We address this bottleneck by developing parallel computing tools that can distribute for example large FFT computations across GPUs (potentially located at multiple nodes).

4.1 Large-scale Wave Propagation via D^2FFT

Per the ASM in (3), the FFT lies at the heart of the (Fourier) wave optics, which turns out to be an inevitable bottleneck for high-resolution large-scale wave optics simulation—the loaded memory exceeds the GPU memory capacity which makes the model unable to fit entirely on a single GPU. To address this issue, we implement a D^2FFT, following the spirit of tensor parallelism.

Multidimensional FFT can be efficiently computed by a sequence of lower-dimensional FFTs. For instance, a 2-D FFT can be performed by first applying a 1-D FFT along the rows, followed by a 1-D FFT along the columns. Such a row-column decomposition naturally leads to a distributed-memory FFT implementation with an additional communication step. To execute a large-scale 2-D FFT, we first partition the given 2-D array into multiple sub-array chunks along the columns (Y-axis), each of which is assigned and allocated to a different (GPU) device. Under this arrangement, the row (X-axis) information is completely preserved in each device, and hence, a batched 1-D FFT along rows (X-axis) can be performed independently on each device. Then, a MPI-style primitive for collective communication, *all-to-all* communication [Alabed et al. 2025; Doi and Negishi 2010; Snir 1998], is invoked to exchange unique chunks of the distributed array between participating devices. Mathematically, this is analogous to a matrix transpose operation, such that the resulting array is now partitioned along the rows (X-axis), which is followed by a batched 1-D FFT along the Y-axis on each device to complete the 2-D FFT. Note the spatial sharding axis of the distributed array swaps after performing this distributed-memory 2-D FFT. The whole process is illustrated in Fig. 6 (top-left). We implement this operation in Jax [Bradbury et al. 2018] framework in a differentiable way, where the backward pass is realized by another distributed-memory 2-D FFT since the discrete Fourier transform can be represented by a symmetric matrix. Differentiable distributed-memory inverse FFT can be implemented similarly.

4.2 Tensor-parallel Convolutional Neural Network

With D^2FFT and thereby large-scale ASM in hand, we are almost ready to address the optimization problem in (5) at scale. However, the computational load of the neural lithography model $\mathcal{G}$ also becomes intractable for large-scale DOE designs. Despite the simplicity of the CNN we adopted as the neural lithography model, it still requires a large amount of GPU memory in dealing with high-resolution large-area design layouts (*e.g.*, a 1 cm^2 input layout at 1 μm-spacing grid is discretized as a 10,000×10,000 array as the CNN's input) which is a rather unusual case in DL. The common sliding-window strategy to circumvent this issue is only applicable at inference, since the fabrication-aware optimization requires the objective gradient to be backpropagated to the input layout at training. As such, akin to the D^2FFT, tensor parallelism must be leveraged to make the CNN scalable to large-sized inputs.

To perform the tensor-parallel spatial-partitioning convolution, we split the input array into multiple sub-arrays along the Y axis[5]. Since the convolution (in DL) is a local operation, each sub-array can be independently processed by the same convolutional kernel on each device, except for processing the sub-array boundary. To ensure consistent results around the boundary, each sub-array must be padded with the corresponding boundary information from neighboring devices, thus requiring communications. This special communication operation can be efficiently implemented by another MPI-style primitive, coined *collective permute*. By executing this operation, each sub-array has $(\frac{H}{N} + 2 \times \lfloor \frac{K}{2} \rfloor) \times W$ elements with replicated boundary values, where H, W are the height and width of the global intact array, N is the number of devices, and K denotes the kernel size (or receptive field in general) of the convolutional layer, $\lfloor \cdot \rfloor$ indicates the floor function. An illustration of this spatial-partitioning convolution is shown in Fig. 6 (bottom-left).

```
"""
A multi-dimensional processor mesh is created with 2x2 devices
where first and second axes are named wavelength ("wvl")
and tensor parallel ("tp") respectively.
"""
mesh = MeshShardingHelper([2, 2], ['wvl', 'tp'])
@partial(
    mesh.sjit,
    args_sharding_constraint=(
        PartitionSpec('wvl', 'tp', None), # for "field"
        PartitionSpec('wvl', None, 'tp'), # for "H"
    ),
    out_shardings=PartitionSpec('wvl', 'tp', None), # for "E_prop"
)
def propagate(field, H):
    U = fft_func(field)  # Perform 2D FFT (implemented elsewhere)
    E_k_prop = U * H
    E_prop = ifft_func(E_k_prop)  # Perform 2D inverse FFT
    return E_prop
```

Fig. 7. **Example GSPMD Segment** for the free-space wave propagator. An 2×2 processor mesh is created to realize hybrid data and tensor parallelism. Pure data or tensor parallelism can be achieved by simply setting the mesh size to [4, 1] or [1, 4], respectively. All tensors have three dimensions (axes), and the sharding annotations/constraints suggest the spatial sharding axis (for tensor parallelism) of "field" and "H" are the Y and X axes, respectively, because the D^2FFT would swap this sharding axis from Y to X.

4.3 GSPMD Implementation

Handcrafting computational diffractive optics systems with hardware-associated operators from above is often time-consuming and error-prone. To enable flexible configurations (such as arbitrary combinations of data and tensor parallelism for any hardware topology) we implement the framework in Jax following the GSPMD programming model, leveraging an automatic, compiler-based parallelization system [Alabed et al. 2025; Xu et al. 2021]. The essential abstractions

---

[5]The spatial sharding axis should be chosen to be compatible with the following D^2FFT to avoid unnecessary data resharding.
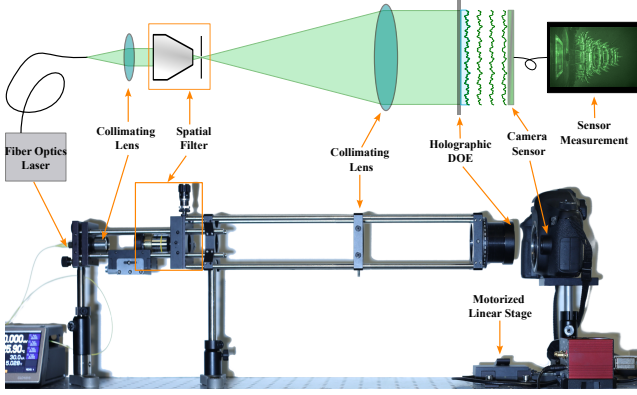
Fig. 8. **Experimental Holographic Setup.** We validate the fabrication-aware DOE designs with the experimental setup shown above. Here, a collimated laser beam is modulated by the DOE, which then propagates to the photosensor directly to form the designed hologram image.

of GSPMD are a multi-dimensional mesh of processors and a sharding annotation system. The former arranges the available computing devices into a multi-dimensional mesh with named sharding axes, while the latter annotes the partition specifications (PartitionSpec) of the input and output tensors of a given function. A simplified example of a GSPMD implementation segment of the ASM is shown in Fig. 7. By utilizing this programming model and the XLA compiler [OpenXLA Community 2025], the hardware-specific MPI-style primitive is automatically generated and invoked in the compiled program free of error-prone hard encoded low-level MPI calls.

5 APPLICATIONS AND ANALYSIS

In this section, we evaluate the proposed fabrication-aware design method for several applications, including computational display holography (Section 5.2), beam shaping (Section 5.3), and single-DOE broadband color imaging (Section 5.4). For all applications, we validate the effectiveness of our approach not only in simulation but also with an experimental prototype, resulting in 11 fabricated DOEs in total (6 for holography display, 2 for beam shaping, 3 for broadband imaging). The readers are encouraged to review the video in the Supplementary Material.

5.1 Prototype Fabrication

We follow the fabrication pipeline from Sec. 3.2 and fabricate microstructured patterned DOEs using direct-write grayscale lithography with the Heidelberg Instruments DWL 66+ mask writer, employing AZ® 4562 photoresist on soda-lime glass to create smooth 3D relief structures in a single exposure step. The process involves spin-coating, soft-baking at 120°C, writing with 1023 grayscale levels, and developing with AZ® 726 MIF for 25 seconds. These patterns are then transferred into OrmoComp, a UV-curable polymer, via room-temperature nanoimprint lithography using the Obducat Eitre 3 system, where the photoresist mold is pressed into the polymer, cured with UV light, and released to produce high-fidelity micro-optical components. After fabrication, the DOE wafers are diced

and mounted in the respective experimental setup described in the corresponding sections below.

5.2 Computational Holographic Display

Computational holographic display (CHD) [Chakravarthula et al. 2019; Peng et al. 2020] employs algorithms to simulate optical hologram creation and reconstruction. The computed holographic image is brought to life through a display, usually featuring an illuminating source and a phase-modifying element (DOEs). We focus on 2-D near-field computational holography. Specifically, we use Adam optimizer to solve the optimization problems (4) and (5), where the loss function is defined in (6), together with the training recipe in Section 3.2 (Toy Example). Three 3.6×3.6 mm^2 DOEs with 2 μm feature size are designed and fabricated to generate the same 2-D 3.6×3.6 mm^2 hologram image in 1 cm away from the DOE, for which we build the experimental setup as shown in Fig. 8. The optics in our setup are designed to produce a smooth, spatially clean, and evenly collimated beam (520.6-nm wavelength). This is achieved by using a spatial filter as an intermediate step between expander stage. The resulting beam enables plane-wave illumination for holograms of various sizes. Our physical realization does not require any additional intermediate image plane (such as those implemented with a 4-f system to filter out undesirable diffraction orders [Gopakumar et al. 2021]), thus directly assessing the patterns diffracted from DOEs. Finally, for accurate sensor positioning, we employed a motorized linear stage with a positioning precision of 10 μm. We evaluate 3.6×3.6 mm^2 DOEs with three different design approaches[6]:

- *Conventional*: conventional differentiable optics optimization solving (4) at 2 μm-spacing grid without upsampling;
- *Conventional w. Upsampling*: conventional design from above (4) at 250 nm-spacing grid with 8× nearest upsampling;
- *Fabrication-aware*: proposed fabrication-aware approach of (5) at 250 nm-spacing grid with 8× neural upsampling.

All these DOEs can be efficiently optimized under a single A-100 GPU (*c.f.,* the compute specifics in Table 2), however, to realize fabrication-aware optimization of larger-area DOE, tensor-parallel computing routines must be utilized. Fig. 9 summarizes both the simulation and experimental results for these holograms. Remarkably, our fabrication-aware approach results in an experimental realization of a nearly speckle-free, high-definition cat hologram under the coherent laser illumination. As such, we find that it largely closes the design-to-manufacturing gap evidenced by conventional approaches. We note the residual minor contrast and resolution difference between simulation and experimental results are likely attributed to the imperfection of the experimental setup itself, such as the inexact 3-D printed aperture (as evidenced by the diffraction patterns around boundary), the glass, and anti-aliasing filters atop the photosensor. Additional computer-generated hologram results (in simulation) can be found in Fig. 10, which confirm the effectiveness of the proposed method for realizing diverse and vivid hologram images with abundant details.

---

[6]Extra results of Xu et al. [2025] and our fabrication-aware approach with 2× neural upsampling are provided in the Supplementary Material.
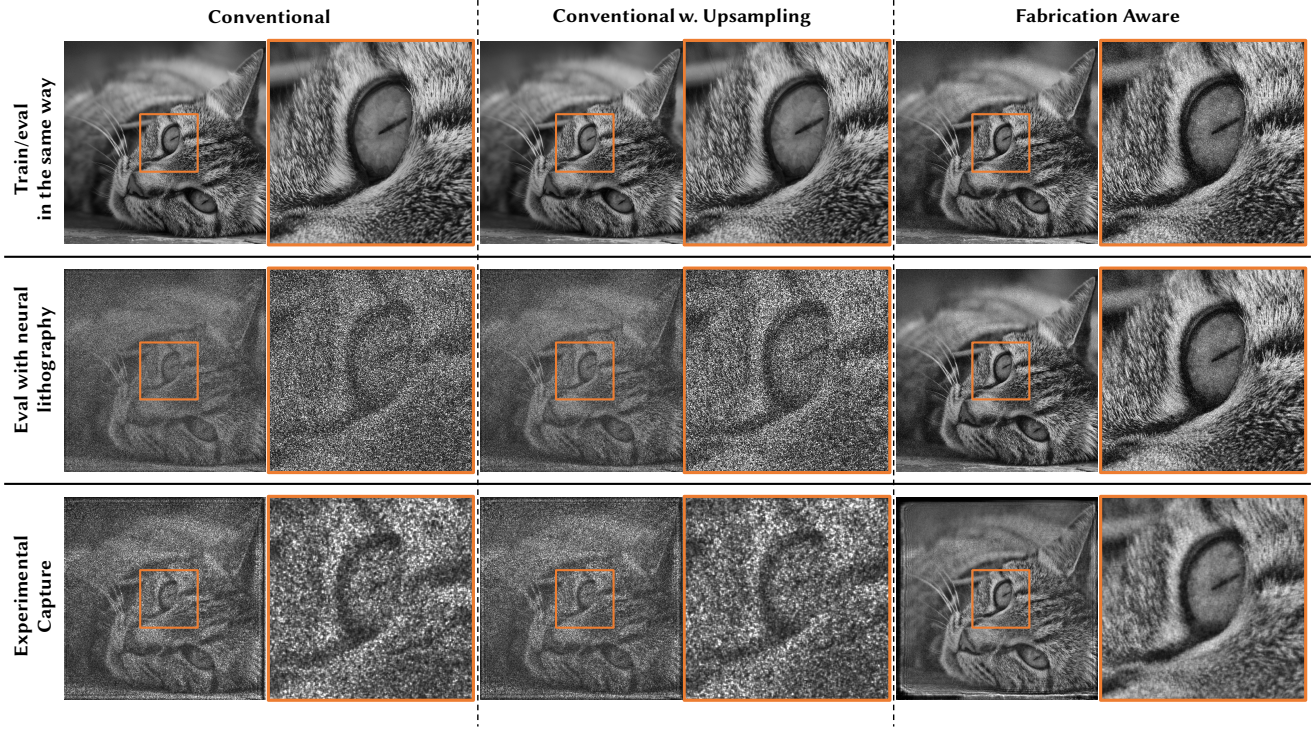
Fig. 9. **Experimental and Synthetic Evaluation of Proposed Fabrication-aware CHDs**. In simulation, 1) all approaches are successful when trained and evaluated under the same setting (row 1); 2) the quality of the conventional approaches declines drastically under the neural lithography setting (row 2) which matches the experimental captures (row 3). The fabricated DOE designed through our approach generates almost speckle-free, high-resolution coherent hologram close to the simulation, without any additional optical filtering.

Table 2. **Compute Cost for Optimizing Task DOEs.** The discrete grid characterizes the discretization of the transfer function of ASM (as well as the super-resolved DOE). The 2-D GSPMD mesh size indicates the number of devices assigned to data and tensor parallel dimension. For CHD and broadband imaging, we optimize the DOEs for 3,000 and 20,000 iterations, respectively, for which we report the total compute time of the optimization process.

DOE	Design Approach	Discrete Grid	# of GPUs	GSPMD Mesh	Compute Time
3.6mm × 3.6mm	Conventional	3,600×3,600	1	[1, 1]	42 s
CHD & Beam Shaping	Conventional w. Upsampling	28,800×28,800	1	[1, 1]	14 m
(*c.f.,* Fig. 9, 12)	Fabrication Aware	28,800×28,800	1	[1, 1]	17 m
32.16mm × 21.44mm	Conventional w. Upsampling	128,640×85,760	16	[1, 16]	13 h
CHD (*c.f.,* Fig. 11)	Fabrication Aware	128,640×85,760	16	[1, 16]	15 h
10mm × 10mm	Conventional w. Upsampling	6×40,000×40,000	12	[6, 2]	9 h
Imaging (*c.f.,* Fig. 14)	Fabrication Aware	6×40,000×40,000	12	[6, 2]	10 h

Large-Area DOEs for Ultra-definition Near-field Holography. To confirm the scalability of the proposed tensor-parallel framework tailored for large-area DOE designs, we further optimize two 2 μm feature-sized DOEs with 32.16 mm in height and 21.44 mm in width (corresponding to 4K × 6K pixels in the CanonEOS5D5 sensor we used), discretized at a 0.5 μm-spacing grid (4× super-resolution). These DOEs are optimized to produce the same-sized holograms at 6 cm away from the DOEs, given the plane-wave coherent illumination. This optimization task results in large-scale intermediary arrays with 128,640 × 85,760 pixels beyond the memory capacity of a single A-100 GPU (even a single computing node of 8 A-100 GPUs). We instead employ our GSPMD-based distributed computing framework, and deploy the DOE optimization into two nodes consisting of 16 A-100 GPUs, which readily fit this huge array spatially shard onto 16 segments. The experimental measurements (using the holographic setup in Fig. 8) for both designs (conventional and ours) are shown in Fig. 11, which clearly validates the scalability of
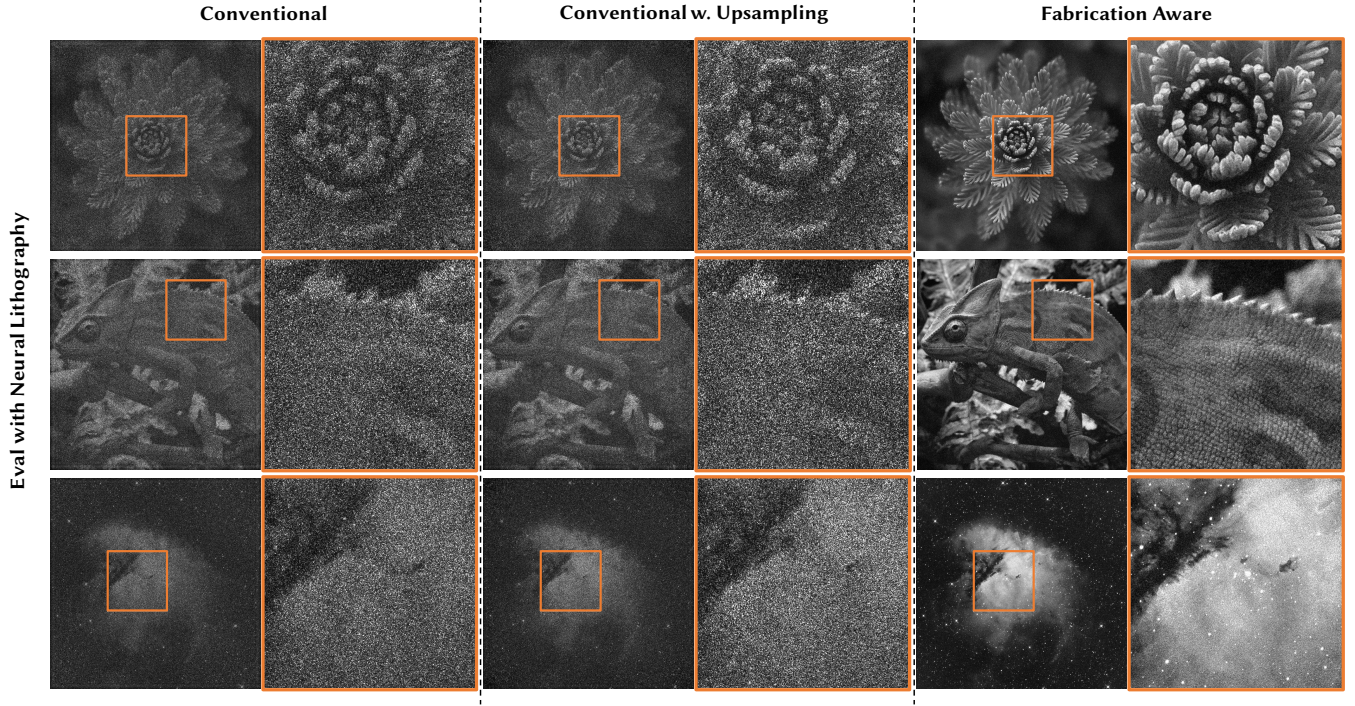
Fig. 10. **Simulation CHD Results for Diverse Scenes.** For all of the scenes, our fabrication-aware approach consistently produces high-resolution, clean holograms with complex details that are barely visible in conventional approaches.

our proposed approach in handling large-area DOE designs at an unprecedented scale.

5.3 Beam Shaping with Diffractive Beam Splitter

Beam shaping, as one of the classic use cases of DOE [Kress and Meyrueis 2000] has found downstream applications in laser material processing [Kuang et al. 2013], fast LiDAR 3-D sensing [Yuan et al. 2021] and visual vibrometry [Zhang et al. 2023b]. Here, we design DOEs as diffractive beam splitters whose goal is to split the collimated laser beam into a regular grid of beams that yields an array of spots with flat-top intensity profiles at the desired plane. This task can be viewed as a special case of computer-generated hologram—instead of generating an image, we steer the DOE to produce an array of focal spots in the destination plane. We find that beam splitter designs without considering the fabrication deformation inevitably lose diffraction efficiency and the spot array intensity uniformity. Again, we design two 3.6×3.6 mm² DOEs with 2 μm feature size, steering the incident plane-wave beam into a 8×8 array of flat-top equal-intensity spots in a focal plane located at 1 cm away from the DOE, by minimizing the loss function in (6). We reuse the experimental setup of the holographic display (Fig. 8) and measure the raw intensity patterns for fabricated beam-splitting DOEs under the same lighting condition for fair comparison of diffraction efficiency. The experimental measurements of diffracted spot-array patterns are exhibited in Fig. 12, where the overall spot

intensity of the fabrication-aware design is **53%** higher than the conventional one, indicating large diffraction efficiency gains brought by our method.

5.4 Single-DOE Broadband Color Imaging

Broadband color imaging with diffractive optics systems is fundamentally challenging due to strong wavelength-dependent dispersion [Aieta et al. 2015; Peng et al. 2016]. High-fidelity broadband reconstruction demands point-spread functions (PSFs) that are simultaneously invertible and spectrally consistent across densely sampled bands [Fröch et al. 2025; Peng et al. 2016; Sun et al. 2025], which notably increases the memory requirements of the design optimization. In addition, discrepancies between simulated and fabricated PSFs further degrade image quality [Chakravarthula et al. 2023; Shi et al. 2024b]. We validate that the distributed fabrication-aware method addresses both issues.

Using the GSPMD implementation described in Section 4.3, we shard the DOE model across two devices, sample 6 wavelengths uniformly over the visible range (400 nm to 700 nm) leveraging hybrid data and tensor parallelism (with GSPMD mesh size [6, 2]). We enforce radial symmetry on the phase profile by optimizing a 16th-order polynomial radial vector. Following the evaluation of the previous applications, two DOE variants (conventional with nearest upsampling and our fabrication-aware approach with neural upsampling) are trained with the same broadband imaging loss

$$\mathcal{L}_{\text{imaging}} = \mathcal{L}_{\text{focus}} + \mathcal{L}_{\text{consistency}} + \mathcal{L}_{\text{energy}} \tag{7}$$
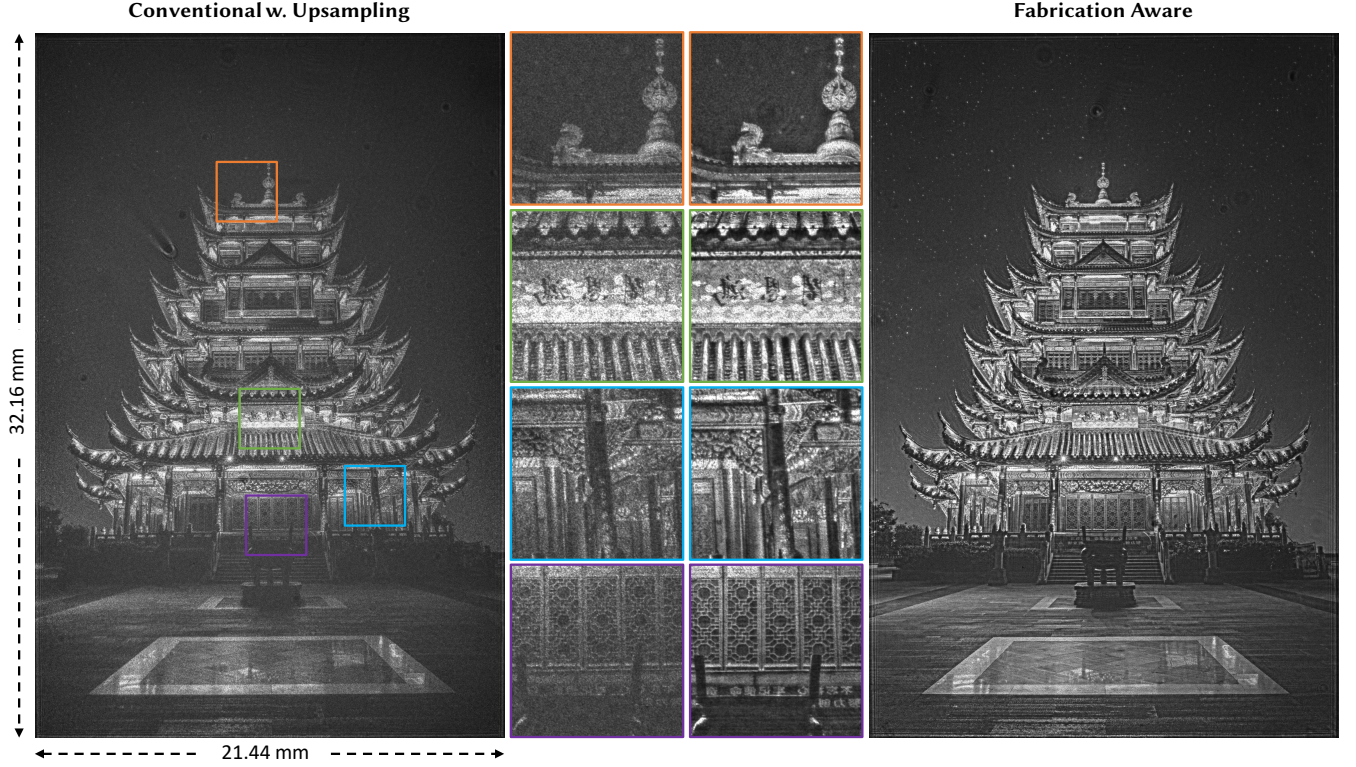
Fig. 11. **Experimental, Large-area, High-definition (4K × 6K) 2-D Hologram Reconstruction**. Two 32.16 mm × 21.44 mm DOEs with 2 μm feature size are designed and optimized at 0.5 μm-spacing grid (4× super resolution) using conventional (with nearest upsampling) and our fabrication-aware approaches, distributed across 16 A-100 GPUs through tensor parallel. We compare the resulting holograms side-by-side in real-world experiments, where our fabrication-aware approach generates sharper, clearer, high-SNR details than the conventional one.

Here, $\mathcal{L}_{\text{focus}}$ drives the PSF to concentrate its energy at the designated center pixel, ensuring a sharp focal peak; $\mathcal{L}_{\text{consistency}}$ penalizes variations in PSF shape across the sampled wavelengths, enforcing spectral uniformity; and $\mathcal{L}_{\text{energy}}$ maximizes the total PSF energy to avoid trivial, zero-energy solutions. For more details, please refer to the Supplementary Materials (Section B.3).

We fabricated both DOE variants and integrated them into the imaging rig as shown in Fig. 13. A diverse set of indoor and outdoor scenes were captured under various illumination conditions to quantitatively and qualitatively assess the performance of each lens design. A representative subset of scenes can be found in Fig. 14, column 1.

To illustrate the simulation–reality discrepancy, we first convert the captured raw image into linear-RGB image space with details in Supplementary Materials (Section B.4). We then apply Wiener deconvolution (8) to each *captured* linear–RGB measurement y_{cap} using its corresponding *simulated* on-axis PSF k_{sim}, that is

$$\hat{x} = \mathcal{F}^{-1}\left\{ \frac{\overline{K_{\text{sim}}}}{\left|K_{\text{sim}}\right|^2 + \gamma} \; Y_{\text{cap}} \right\}. \tag{8}$$

Here, capitalization denotes the Frequency domain version of the variables, and $\overline{(\cdot)}$ denotes complex conjugation.

When the fabricated PSF closely matches the design as for proposed fabrication-aware DOE (Fig. 14, column 4 and 5), the deconvolved reconstruction $\hat{x}$ closely approximates the ground truth. By contrast, the conventional, nearest-upsampling DOE (Fig. 14, columns 2 and 3) exhibits PSF mismatches that manifest as wavelength-dependent chromatic aberrations, elevated noise in poorly invertible spectral bands, and ringing artifacts under the high-SNR inversion assumption ($\gamma = 5 \times 10^{-4}$). These results confirm the effectiveness of our fabrication-aware optimization in closing the simulation-to-reality gap.

We also remark that the image quality of our neural designs is already competitive with existing single-DOE computational imaging systems, despite the basic single-step reconstruction method with simulated PSFs employed here. We find that lightweight, edge-friendly reconstruction methods can be used when an accurate optical design process is taken into account, enabled by our fabrication-aware pipeline. Additional experimental results, including the comparisons of the inverse-filtering results using simulated and experimental PSFs, can be found in Supplementary Materials (Section C.2).

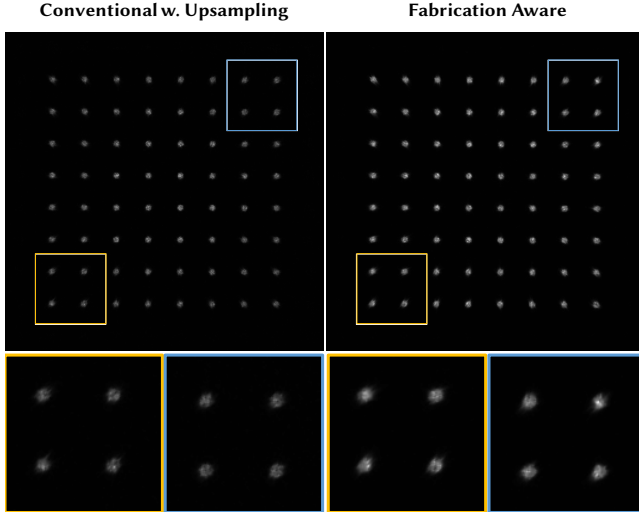**Conventional w. Upsampling**   **Fabrication Aware**

Fig. 12. **Experimental Evaluation of Diffractive Beam Splitters.** We show the experimental raw measurements captured under the same lighting condition, where we keep the exposure time, laser power constant for both measurements. The intensity of measurements directly reflects the relative diffraction efficiency of the beam splitters, validating the effectiveness of the fabrication-aware beam splitter device.
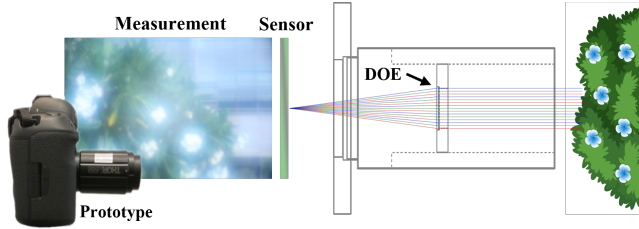

Fig. 13. **Single-DOE Broadband Imaging Setup.** For the proposed imaging application, install the designed DOE inside a telescopic lens tubes to assemble the f/8 imaging setup with design focal length. We mount the tube with an adapter on an off-she-shelf DSLR camera (left).

6 DISCUSSIONS

This section discusses the proposed framework in terms of scalability, effectiveness, and limitations.

Scaling Analysis of the Tensor-parallel Algorithms. We conduct a scaling analysis of the proposed tensor-parallel computing routines for large-area computational diffractive optics. We take the coherent computer-generated hologram as the benchmark, and test the largest system size (the total number of elements of the discrete grid of a super-resolved DOE) that can be fitted in a given number of GPU devices. The analysis is reported in Fig. 15, where we evaluate multiple device configurations from 1 to 16. We find sub-linear scaling (1.73× rather than perfect 2× linear scaling) from 1 to 2 devices, which is expected because of the communication overhead. Our method scales linearly from 2 to 16 GPU processors (up to the

available computing resources we access to), validating scalability of our distributed computing framework.

Limitations on Model/Data Uncertainty. We note that the randomness of the fabrication and measurement processes inherently limits our neural lithography model's precision (forward predictability). To analyze this limitation for our specific process, in preparation for the calibration patterns used to train the neural lithography model (Section 3.2), we deliberately place multiple identical patterns in the design layout periphery. Ideally, these patterns post-fabrication would be identical; however, as shown in Fig. 16 for two patterns, slight differences due to random fabrication and measurement error remain, setting the upper bound of prediction accuracy of the neural lithography model. Nevertheless, a significant portion of the observed random variation stems from our current academic fabrication facility, which relies on manual interventions such as the control of development time—introducing operator-dependent variability. In commercial foundries, such processes are fully automated with substantially reduced operational variance and higher repeatability.

Model Generalizability. The proprietary and diverse nature of nano/micro-fabrication recipes makes it infeasible to directly apply a model trained in one foundry to another, or even within the same foundry using different materials. Nevertheless, our methodology is universal and can be adapted to other lithography processes through appropriate calibration, rendering it highly suitable for clean-room training. Currently, AFM measurements represent the primary bottleneck in the calibration workflow; however, this requirement remains manageable, as the process can be completed within one day using only 10 relatively small patches.

Camera-in-the-loop (CITL) Approaches. Unlike CITL approaches that learn a *global* mapping for wave propagation, our digital twin of the manufacturing is designed to learn *local* shape deviations between the design and the fabricated device. These local variations correspond to process-induced effects such as optical blur, nonlinear material response, and 3D chemical reactions. Our model does not account for global effects. Based on our extensive experience with DOE fabrication, we find such global variations—for instance, those related to position on the wafer—are minimal. This sets nanofabrication apart from digital holography works employing spatial light modulators (SLM) that may suffer voltage gradients across the chip, and other non-uniformities. The focus on locality offers practical advantages: it allows the model to be trained with limited data and enables rapid recalibration for new material systems or process conditions. In contrast, a full CITL pipeline (for static DOE) with potentially hundreds of prototypes would be intractable in practice, as are AFM measurements for large area samples required to train global models.

7 CONCLUSION

This work tackles two limitations that existing computational diffractive optics struggle with: 1) the design-to-manufacturing gap and 2) the inability to simulate and optimize large-area devices. To this end, we propose a fabrication-aware, end-to-end optimization method with a super-resolved neural lithography model as a differentiable

Fig. 14. **Experimental Validation of Fabrication-aware Broadband DOE for Imaging.** We capture diverse indoor and outdoor scenes with both conventional nearest up-sampling DOE and proposed fabrication-aware DOE (column 1, 2, and 4). To visualize the discrepancy between the designed PSF and fabricated PSF, we conduct a Wiener filtering using simulated PSF on the experimental captured scenes assuming high SNR ($\gamma = 5 \times 10^{-4}$). Our proposed fabrication-aware design yields high-fidelity results, while the conventional nearest up-sampling results in images with severe chromatic aberration, ringing artifacts, and elevated noise.
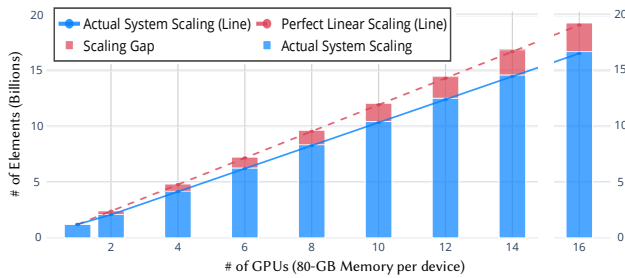

Fig. 15. **Scaling Analysis** of the tensor-parallel computing routines (D^2FFT and spatial-partitioning convolution) using the coherent computer-generated hologram as the evaluation benchmark. We report here the largest system (indicated by the number of elements of the discrete grid) scaling with the number of devices (80-GB memory per device).
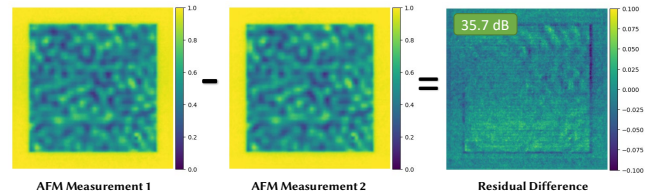

Fig. 16. **Fabrication Uncertainty Analysis** of the neural lithography pipeline. Two identical design patterns are fabricated repeatedly. Analyzing these two patterns post-fabrication yields slightly different AFM measurements, which inherently limits the precision upper bound of the neural lithography model.

fabrication surrogate. To design large area optics, we develop a tensor-parallel algorithm (*i.e.*, D^2FFT and spatial-partitioning convolution), which overcomes problem-inherent memory limitations and supports large-scale wave propagation and diffractive optics optimization at sub-micron resolution. We validate our method with direct-write grayscale lithography and nanoimprint replication, a process employed today for mass production of large-area devices. Experiments across computational display holography, beam shaping, and broadband color imaging validate that the method can produce high-quality diffractive optical systems effectively.

We believe our contributions offer a step toward bridging the gap between simulation and mass-market implementation of differentiable imaging optics. Moving forward, incorporating additional fabrication variables, such as material-specific variations, direct-write

laser beam, would be interesting areas for research. We believe these advancements provide a foundation for broader adoption of learned optical systems in real-world imaging and display applications.

ACKNOWLEDGMENTS

Author contributions: K.W., W.H. and F.H. conceived the ideas; K.W. designed, implemented and analyzed the system and framework; H.J. and K.W. built the physical setup and captured the experimental results; H.A. established the fabrication pipeline and produced all designed devices; J.S. performed the image reconstruction for the imaging task; K.W., W.H. and F.H. led the manuscript writing; H.J., H.A., J.S., Q.F. assisted in design and analysis, experiments, and manuscript writing; W.H. and F.H. supervised the project.

REFERENCES

Francesco Aieta, Mikhail A Kats, Patrice Genevet, and Federico Capasso. 2015. Multi-wavelength achromatic metasurfaces by dispersive phase compensation. *Science* 347, 6228 (2015), 1342–1345.

Sami Alabed, Daniel Belov, Bart Chrzaszcz, Juliana Franco, Dominik Grewe, Dougal Maclaurin, James Molloy, Tom Natan, Tamara Norman, Xiaoyue Pan, Adam Paszke, Norman A. Rink, Michael Schaarschmidt, Timur Sitdikov, Agnieszka Swietlik, Dimitrios Vytiniotis, and Joel Wee. 2025. PartIR: Composing SPMD Partitioning Strategies for Machine Learning. In *Proceedings of the 30th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 1.* ACM, Rotterdam Netherlands, 794–810. https://doi.org/10.1145/3669940.3707284

M. Salman Asif, Ali Ayremlou, Aswin Sankaranarayanan, Ashok Veeraraghavan, and Richard G. Baraniuk. 2017. FlatCam: Thin, Lensless Cameras Using Coded Aperture and Computation. *IEEE Transactions on Computational Imaging* 3, 3 (Sept. 2017), 384–397. https://doi.org/10.1109/TCI.2016.2593662

Alan Ayala, Stan Tomov, Miroslav Stoyanov, Azzam Haidar, and Jack Dongarra. 2022. Performance analysis of parallel FFT on large multi-GPU systems. In *2022 IEEE International Parallel and Distributed Processing Symposium Workshops (IPDPSW)*. IEEE, 372–381.

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 *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, Montreal, QC, Canada, 2631–2640. https://doi.org/10.1109/ICCV48922.2021.00265

Steven Barcelo and Zhiyong Li. 2016. Nanoimprint lithography for nanodevice fabrication. *Nano Convergence* 3, 1 (Dec. 2016), 21. https://doi.org/10.1186/s40580-016-0081-y

Ayush Bhandari, Achuta Kadambi, and Ramesh Raskar. 2022. *Computational Imaging.*

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 (Sept. 2016), 23–35. https://doi.org/10.1109/MSP.2016.2581921

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 (July 2020), 1618–1629. https://doi.org/10.1109/TPAMI.2020.2987489

James Bradbury, Roy Frostig, Peter Hawkins, Matthew James Johnson, Chris Leary, Dougal Maclaurin, George Necula, Adam Paszke, Jake VanderPlas, Skye Wanderman-Milne, et al. 2018. JAX: composable transformations of Python+ NumPy programs. (2018).

Bryon R Brown and Adolf W Lohmann. 1966. Complex spatial filtering with binary masks. *Applied optics* 5, 6 (1966), 967–969.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems* 33 (2020), 1877–1901.

John H. Bruning. 2007. Optical lithography: 40 years and holding, Donis G. Flagello (Ed.). San Jose, CA, 652004. https://doi.org/10.1117/12.720631

Thomas Cecil, Danping Peng, Daniel Abrams, Stanley J. Osher, and Eli Yablonovitch. 2022. Advances in Inverse Lithography. *ACS Photonics* (Sept. 2022), acsphotonics.2c01026. https://doi.org/10.1021/acsphotonics.2c01026

Praneeth Chakravarthula, Yifan Peng, Joel Kollin, Henry Fuchs, and Felix Heide. 2019. Wirtinger holography for near-eye displays. *ACM Transactions on Graphics* 38, 6 (Dec. 2019), 1–13. https://doi.org/10.1145/3355089.3356539

Praneeth Chakravarthula, Jipeng Sun, Xiao Li, Chenyang Lei, Gene Chou, Mario Bijelic, Johannes Froesch, Arka Majumdar, and Felix Heide. 2023. Thin On-Sensor Nanophotonic Array Cameras. *ACM Transactions on Graphics* 42, 6 (Dec. 2023), 1–18. https://doi.org/10.1145/3618398

Pierre Chevalier, Patrick Quemere, Sebastien Berard-Bergery, Jean-Baptist Henry, Charlotte Beylier, and Jerome Vaillant. 2021. Rigorous Model-Based Mask Data Preparation Algorithm Applied to Grayscale Lithography for the Patterning at the Micrometer Scale. *Journal of Microelectromechanical Systems* 30, 3 (June 2021), 442–455. https://doi.org/10.1109/JMEMS.2021.3067475

Suyeon Choi, Manu Gopakumar, Yifan Peng, Jonghyun Kim, Matthew O'Toole, and Gordon Wetzstein. 2022. Time-multiplexed neural holography: a flexible framework for holographic near-eye displays with fast heavily-quantized spatial light modulators. In *ACM SIGGRAPH 2022 Conference Proceedings.* 1–9.

M Collischon, H Haidner, P Kipfer, A Lang, John T Sheridan, Johannes Schwider, Norbert Streibl, and J Lindolf. 1994. Binary blazed reflection gratings. *Applied optics* 33, 16 (1994), 3572–3577.

F.H. Dill. 1975. Optical lithography. *IEEE Transactions on Electron Devices* 22, 7 (July 1975), 440–444. https://doi.org/10.1109/T-ED.1975.18158

Jun Doi and Yasushi Negishi. 2010. Overlapping methods of all-to-all communication and FFT algorithms for torus-connected massively parallel supercomputers. In *SC'10: Proceedings of the 2010 ACM/IEEE International Conference for High Performance Computing, Networking, Storage and Analysis.* IEEE, 1–9.

Matteo Frigo and Steven G Johnson. 2005. The design and implementation of FFTW3. *Proc. IEEE* 93, 2 (2005), 216–231.

Johannes E Fröch, Praneeth Chakravarthula, Jipeng Sun, Ethan Tseng, Shane Colburn, Alan Zhan, Forrest Miller, Anna Wirth-Singh, Quentin AA Tanguy, Zheyi Han, et al. 2025. Beating spectral bandwidth limits for large aperture broadband nano-optics. *Nature communications* 16, 1 (2025), 3025.

J Fung Chen, Tom Laidig, Kurt E Wampler, and Roger Caldwell. 1997. Optical proximity correction for intermediate-pitch features using sub-resolution scattering bars. *Journal of Vacuum Science & Technology B: Microelectronics and Nanometer Structures Processing, Measurement, and Phenomena* 15, 6 (1997), 2426–2433.

Dennis Gabor. 1948. A New Microscopic Principle. *Nature* 161 (1948), 777–778. https://doi.org/10.1038/161777a0

Pascal Getreuer. 2011. Linear methods for image interpolation. *Image Processing On Line* 1 (2011), 238–259.

Amir Gholami, Judith Hill, Dhairya Malhotra, and George Biros. 2016. AccFFT: A library for distributed-memory FFT on CPU and GPU architectures. (May 2016). http://arxiv.org/abs/1506.07933 arXiv:1506.07933 [cs].

Franz J Giessibl. 2003. Advances in atomic force microscopy. *Rev. Mod. Phys.* 75, 3 (2003).

Joseph W Goodman. 2005. Introduction to Fourier optics. (2005).

Manu Gopakumar, Jonghyun Kim, Suyeon Choi, Yifan Peng, and Gordon Wetzstein. 2021. Unfiltered holography: optimizing high diffraction orders without optical filtering for compact holographic displays. *Optics Letters* 46, 23 (Dec. 2021), 5822. https://doi.org/10.1364/OL.442851

Anya Grushina. 2019. Direct-write grayscale lithography. *Advanced Optical Technologies* 8, 3-4 (June 2019), 163–169. https://doi.org/10.1515/aot-2019-0024

M G Guney and G K Fedder. 2016. Estimation of line dimensions in 3D direct laser writing lithography. *Journal of Micromechanics and Microengineering* 26, 10 (Oct. 2016), 105011. https://doi.org/10.1088/0960-1317/26/10/105011

Anshul Gupta and Vipin Kumar. 2002. The scalability of FFT on parallel computers. *IEEE Transactions on Parallel and Distributed Systems* 4, 8 (2002), 922–932.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, Las Vegas, NV, USA, 770–778. https://doi.org/10.1109/CVPR.2016.90

Felix Heide, Qiang Fu, Yifan Peng, and Wolfgang Heidrich. 2016. Encoded diffractive optics for full-spectrum computational imaging. *Scientific Reports* 6, 1 (Sept. 2016), 33543. https://doi.org/10.1038/srep33543

W Daniel Hillis and Guy L Steele Jr. 1986. Data parallel algorithms. *Commun. ACM* 29, 12 (1986), 1170–1183.

Yanping Huang, Youlong Cheng, Ankur Bapna, Orhan Firat, Dehao Chen, Mia Chen, HyoukJoong Lee, Jiquan Ngiam, Quoc V Le, Yonghui Wu, et al. 2019. Gpipe: Efficient training of giant neural networks using pipeline parallelism. *Advances in neural information processing systems* 32 (2019).

Hayato Ikoma, Cindy M Nguyen, Christopher A Metzler, Yifan Peng, and Gordon Wetzstein. 2021. Depth from defocus with learned optics for imaging and occlusion-aware depth estimation. In *2021 IEEE International Conference on Computational Photography (ICCP)*. IEEE, 1–12.

Changwon Jang, Olivier Mercier, Kiseung Bang, Gang Li, Yang Zhao, and Douglas Lanman. 2020. Design and fabrication of freeform holographic optical elements. *ACM Transactions on Graphics* 39, 6 (Dec. 2020), 1–15. https://doi.org/10.1145/3414685.3417762

Koray Kavaklı, Hakan Urey, and Kaan Akşit. 2021. Learned holographic light transport. *Applied Optics* 61, 5 (2021), B50–B55.

S.N. Khonina, N.L. Kazanskiy, and M.A. Butt. 2024. Exploring Diffractive Optical Elements and Their Potential in Free Space Optics and imaging- A Comprehensive Review. *Laser & Photonics Reviews* 18, 12 (Dec. 2024), 2400377. https://doi.org/10.1002/lpor.202400377

Diederik P Kingma. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

Bernard Kress and Patrick Meyrueis. 2000. *Digital diffractive optics.* Wiley New York.

Zheng Kuang, Walter Perrie, Dun Liu, Stuart P Edwardson, Yao Jiang, Eamonn Fearon, Ken G Watkins, and Geoff Dearden. 2013. Ultrafast laser parallel microprocessing using high uniformity binary Dammann grating generated beam array. *Applied surface science* 273 (2013), 101–106.

Grace Kuo, Florian Schiffers, Douglas Lanman, Oliver Cossairt, and Nathan Matsuda. 2023. Multisource Holography. (Sept. 2023). http://arxiv.org/abs/2309.10816 arXiv:2309.10816 [physics].

E Y Lam and A K K Wong. 2009. Computation lithography: virtual reality and virtual virtuality. (2009).

Nicolas Lang, Sven Enns, Julian Hering, and Georg Von Freymann. 2022. Towards efficient structure prediction and pre-compensation in multi-photon lithography. *Optics Express* 30, 16 (Aug. 2022), 28805. https://doi.org/10.1364/OE.462775

LB Lesem, PM Hirsch, and JA Jordan. 1969. The kinoform: a new wavefront reconstruction device. *IBM Journal of Research and Development* 13, 2 (1969), 150–155.

Tianjian Lu, Yi-Fan Chen, Blake Hechtman, Tao Wang, and John Anderson. 2021. Large-Scale Discrete Fourier Transform on TPUs. *IEEE Access* 9 (2021), 93422–93432. https://doi.org/10.1109/ACCESS.2021.3092312

Xu Ma and Gonzalo R Arce. 2011. *Computational lithography.* John Wiley & Sons.

Chris A. Mack. 2011. Fifty Years of Moore's Law. *IEEE Transactions on Semiconductor Manufacturing* 24, 2 (May 2011), 202–207. https://doi.org/10.1109/TSM.2010.2096437

Joseph N. Mait, Gary W. Euliss, and Ravindra A. Athale. 2018. Computational imaging. *Advances in Optics and Photonics* 10, 2 (June 2018), 409. https://doi.org/10.1364/AOP.10.000409

Kyoji Matsushima. 2010. Shifted angular spectrum method for off-axis numerical propagation. *Optics Express* 18, 17 (Aug. 2010), 18453. https://doi.org/10.1364/OE.18.018453

Kyoji Matsushima and Tomoyoshi Shimobaba. 2009. Band-limited angular spectrum method for numerical simulation of free-space propagation in far and near fields. *Optics Express* 17, 22 (Oct. 2009), 19662. https://doi.org/10.1364/OE.17.019662

Christopher A. Metzler, Hayato Ikoma, Yifan Peng, and Gordon Wetzstein. 2020. Deep Optics for Single-Shot High-Dynamic-Range Imaging. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, Seattle, WA, USA, 1372–1382. https://doi.org/10.1109/CVPR42600.2020.00145

Gordon E Moore. 1998. Cramming more components onto integrated circuits. *Proc. IEEE* 86, 1 (1998), 82–85.

Deepak Narayanan, Mohammad Shoeybi, Jared Casper, Patrick LeGresley, Mostofa Patwary, Vijay Anand Korthikanti, Dmitri Vainbrand, Prethvi Kashinkunti, Julie Bernauer, Bryan Catanzaro, Amar Phanishayee, and Matei Zaharia. 2021. Efficient Large-Scale Language Model Training on GPU Clusters Using Megatron-LM. (Aug. 2021). http://arxiv.org/abs/2104.04473 arXiv:2104.04473 [cs].

S.K. Nayar. 2006. Computational Cameras: Redefining the Image. *Computer* 39, 8 (Aug. 2006), 30–38. https://doi.org/10.1109/MC.2006.258

Elias Nehme, Daniel Freedman, Racheli Gordon, Boris Ferdman, Lucien E. Weiss, Onit Alalouf, Tal Naor, Reut Orange, Tomer Michaeli, and Yoav Shechtman. 2020. DeepSTORM3D: dense 3D localization microscopy and PSF design by deep learning. *Nature Methods* 17, 7 (July 2020), 734–740. https://doi.org/10.1038/s41592-020-0853-5

Temitope Onanuga. 2019. Process modeling of two-photon and grayscale laser direct-write lithography. Ph.D. Dissertation. Dissertation, Erlangen, Friedrich-Alexander-Universität Erlangen-Nürnberg.

OpenXLA Community. 2025. XLA: Accelerated Linear Algebra. https://openxla.org/xla. (2025). Accessed: 2025-05-19.

Linyong (Leo) Pang. 2021. Inverse lithography technology: 30 years from concept to practical, full-chip reality. *Journal of Micro/Nanopatterning, Materials, and Metrology* 20, 03 (Aug. 2021). https://doi.org/10.1117/1.JMM.20.3.030901

Yifan Peng, Suyeon Choi, Nitish Padmanaban, and Gordon Wetzstein. 2020. Neural holography with camera-in-the-loop training. *ACM Transactions on Graphics* 39, 6 (Dec. 2020), 1–14. https://doi.org/10.1145/3414685.3417802

Yifan Peng, Qiang Fu, Hadi Amata, Shuochen Su, Felix Heide, and Wolfgang Heidrich. 2015. Computational imaging using lightweight diffractive-refractive optics. *Optics Express* 23, 24 (Nov. 2015), 31393. https://doi.org/10.1364/OE.23.031393

Yifan Peng, Qiang Fu, Felix Heide, and Wolfgang Heidrich. 2016. The diffractive achromat full spectrum computational imaging with diffractive optics. *ACM Transactions on Graphics* 35, 4 (July 2016), 1–11. https://doi.org/10.1145/2897824.2925941

Yifan Peng, Qilin Sun, Xiong Dun, Gordon Wetzstein, Wolfgang Heidrich, and Felix Heide. 2019. Learned large field-of-view imaging with thin-plate optics. *ACM Transactions on Graphics* 38, 6 (Nov. 2019), 1–14. https://doi.org/10.1145/3355089.3356526

Michael Pippig. 2013. PFFT: An Extension of FFTW to Massively Parallel Architectures. *SIAM Journal on Scientific Computing* 35, 3 (Jan. 2013), C213–C236. https://doi.org/10.1137/120885887

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog* 1, 8 (2019), 9.

Davide Ricci and Pier Carlo Braga. 2004. Recognizing and avoiding artifacts in AFM imaging. *Atomic force microscopy: Biomedical methods and applications* (2004), 25–37.

André Ritter. 2014. Modified shifted angular spectrum method for numerical propagation at reduced spatial sampling rates. *Optics Express* 22, 21 (Oct. 2014), 26265. https://doi.org/10.1364/OE.22.026265

Sourabh K. Saha, Chuck Divin, Jefferson A. Cuadra, and Robert M. Panas. 2017. Effect of Proximity of Features on the Damage Threshold During Submicron Additive Manufacturing Via Two-Photon Polymerization. *Journal of Micro and Nano-Manufacturing* 5, 3 (Sept. 2017), 031002. https://doi.org/10.1115/1.4036444

Noam Shazeer, Youlong Cheng, Niki Parmar, Dustin Tran, Ashish Vaswani, Penporn Koanantakool, Peter Hawkins, HyoukJoong Lee, Mingsheng Hong, Cliff Young, et al. 2018. Mesh-tensorflow: Deep learning for supercomputers. *Advances in neural information processing systems* 31 (2018).

Wenzhe Shi, Jose Caballero, Ferenc Huszar, Johannes Totz, Andrew P. Aitken, Rob Bishop, Daniel Rueckert, and Zehan Wang. 2016. Real-Time Single Image and Video Super-Resolution Using an Efficient Sub-Pixel Convolutional Neural Network. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, Las Vegas, NV, USA, 1874–1883. https://doi.org/10.1109/CVPR.2016.207

Zheng Shi, Yuval Bahat, Seung-Hwan Baek, Qiang Fu, Hadi Amata, Xiao Li, Praneeth Chakravarthula, Wolfgang Heidrich, and Felix Heide. 2022. Seeing through obstructions with diffractive cloaking. *ACM Transactions on Graphics* 41, 4 (July 2022), 1–15. https://doi.org/10.1145/3528223.3530185

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.

Mohammad Shoeybi, Mostofa Patwary, Raul Puri, Patrick LeGresley, Jared Casper, and Bryan Catanzaro. 2020. Megatron-LM: Training Multi-Billion Parameter Language Models Using Model Parallelism. (March 2020). http://arxiv.org/abs/1909.08053 arXiv:1909.08053 [cs].

Vincent Sitzmann, Steven Diamond, Yifan Peng, Xiong Dun, Stephen Boyd, Wolfgang Heidrich, Felix Heide, and Gordon Wetzstein. 2018. End-to-end optimization of optics and image processing for achromatic extended depth of field and super-resolution imaging. *ACM Transactions on Graphics* 37, 4 (Aug. 2018), 1–13. https://doi.org/10.1145/3197517.3201333

Steven W. Smith. 1999. *The scientist and engineer's guide to digital signal processing* (2nd edition ed.). California Technical Pub., San Diego (Calif.). OCLC: 493473234.

Marc Snir. 1998. *MPI–the Complete Reference: the MPI core.* Vol. 1. MIT press.

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 Transactions on Graphics* (2025). https://doi.org/10.1145/3731200

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 *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, Seattle, WA, USA, 1383–1393. https://doi.org/10.1109/CVPR42600.2020.00146

Qilin Sun, Congli Wang, Qiang Fu, Xiong Dun, and Wolfgang Heidrich. 2021. End-to-end complex lens design with differentiate ray tracing. *ACM Transactions on Graphics* 40, 4 (Aug. 2021), 1–13. https://doi.org/10.1145/3450626.3459674

Shiyu Tan, Yicheng Wu, Shoou-I Yu, and Ashok Veeraraghavan. 2021. CodedStereo: Learned Phase Masks for Large Depth-of-field Stereo. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, Nashville, TN, USA, 7166–7175. https://doi.org/10.1109/CVPR46437.2021.00709

Ethan Tseng, Grace Kuo, Seung-Hwan Baek, Nathan Matsuda, Andrew Maimone, Florian Schiffers, Praneeth Chakravarthula, Qiang Fu, Wolfgang Heidrich, Douglas Lanman, et al. 2024. Neural étendue expander for ultra-wide-angle high-fidelity holographic display. *Nature communications* 15, 1 (2024), 2907.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. *NIPS 2017* (2017), 11.

Hao Wang, Cheng-Feng Pan, Chi Li, Kishan S Menghrajani, Markus A Schmidt, Aoling Li, Fu Fan, Yu Zhou, Wang Zhang, Hongtao Wang, Parvathi Nair Suseela Nair, John You En Chan, Tomohiro Mori, Yueqiang Hu, Guangwei Hu, Stefan A Maier, Haoran Ren, Huigao Duan, and Joel K W Yang. 2024. Two-photon polymerization lithography for imaging optics. *International Journal of Extreme Manufacturing* 6, 4 (Aug. 2024), 042002. https://doi.org/10.1088/2631-7990/ad35fe

Hao Wang, Hongtao Wang, Wang Zhang, and Joel K. W. Yang. 2020. Toward Near-Perfect Diffractive Optical Elements *via* Nanoscale 3D Printing. *ACS Nano* 14, 8 (Aug. 2020), 10452–10461. https://doi.org/10.1021/acsnano.0c04313

Hao Wang, Wang Zhang, Dimitra Ladika, Haoyi Yu, Darius Gailevičius, Hongtao Wang, Cheng-Feng Pan, Parvathi Nair Suseela Nair, Yujie Ke, Tomohiro Mori, John You En Chan, Qifeng Ruan, Maria Farsari, Mangirdas Malinauskas, Saulius Juodkazis, Min Gu, and Joel K. W. Yang. 2023. Two-Photon Polymerization Lithography for Optics and Photonics: Fundamentals, Materials, Technologies, and Applications. *Advanced Functional Materials* 33, 39 (Sept. 2023), 2214211. https://doi.org/10.1002/adfm.202214211

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.

Yicheng Wu, Vivek Boominathan, Huaijin Chen, Aswin Sankaranarayanan, and Ashok Veeraraghavan. 2019. Phasecam3d—learning phase masks for passive single view depth estimation. In *2019 IEEE International Conference on Computational Photography (ICCP)*. IEEE, 1–12.

Yuanzhong Xu, HyoukJoong Lee, Dehao Chen, Blake Hechtman, Yanping Huang, Rahul Joshi, Maxim Krikun, Dmitry Lepikhin, Andy Ly, Marcello Maggioni, et al. 2021. GSPMD: general and scalable parallelization for ML computation graphs. *arXiv preprint arXiv:2105.04663* (2021).

Yunpeng Xu, Zihan Zang, Haoqiang Wang, Yanjun Han, Hongtao Li, Yi Luo, Lai Wang, Changzheng Sun, Bing Xiong, Zhibiao Hao, Jian Wang, and Lin Gan. 2025. Fabrication-integrated design for diffractive optical elements. *Optica* 12, 2 (Feb. 2025), 228. https://doi.org/10.1364/OPTICA.539824

Xinge Yang, Qiang Fu, and Wolfgang Heidrich. 2024a. Curriculum learning for ab initio deep learned refractive optics. *Nature Communications* 15, 1 (Aug. 2024), 6572. https://doi.org/10.1038/s41467-024-50835-7

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*. ACM, Tokyo Japan, 1–11. https://doi.org/10.1145/3680528.3687640

Zheng-Nan Yuan, Zhi-Bo Sun, Hoi-Sing Kwok, and Abhishek Kumar Srivastava. 2021. Fast LiDAR systems based on ferroelectric liquid crystal Dammann grating. *Liquid Crystals* 48, 10 (2021), 1402–1416.

Qiang Zhang, Zehao He, Zhenwei Xie, Qiaofeng Tan, Yunlong Sheng, Guofan Jin, Liangcai Cao, and Xiaocong Yuan. 2023a. Diffractive optical elements 75 years on: from micro-optics to metasurfaces. *Photonics Insights* 2, 4 (2023), R09. https://doi.org/10.3788/PI.2023.R09

Tianyuan Zhang, Mark Sheinin, Dorian Chan, Mark Rau, Matthew O'Toole, and Srinivasa G. Narasimhan. 2023b. Analyzing Physical Impacts using Transient Surface Wave Imaging. In *Proc. IEEE CVPR*.

Wenhui Zhang, Hao Zhang, and Guofan Jin. 2020. Frequency sampling strategy for numerical diffraction calculations. *Optics Express* 28, 26 (Dec. 2020), 39916. https://doi.org/10.1364/OE.413636

Cheng Zheng, Guangyuan Zhao, and Peter So. 2023. Close the Design-to-Manufacturing Gap in Computational Optics with a 'Real2Sim' Learned Two-Photon Neural Lithography Simulator. In *SIGGRAPH Asia 2023 Conference Papers (SA '23)*. Association for Computing Machinery, New York, NY, USA, Article 56, 9 pages. https://doi.org/10.1145/3610548.3618251