Dissertation

submitted to the

Combined Faculty of Mathematics, Engineering and Natural Sciences

of Heidelberg University, Germany

for the degree of

Doctor of Natural Sciences

Put forward by

Sai Nikhilesh, Murty Kottapalli

born in: Bhawanipatna, India

Oral examination: 25 July 2025

Optical Wavefront Engineering and Optoelectronic Techniques for Neuromorphic Computing

Referees: Prof. Dr. Peer Fischer

Prof. Dr. Wolfram Pernice

ABSTRACT

Machine learning has emerged as a powerful tool for solving complex problems in various fields, including optics. The scalability of compute resources and energy consumption for larger AI models is, however, a major concern. Optics can be used to implement matrix-vector-multiplications, essential for neural networks, in an energy-efficient manner. However, optical non-linearities are difficult to implement with traditional optical methods. This work demonstrates an optoelectronic device that implements an optical MVM and an electronic non-linearity without the need for high intensities. The device operation is experimentally demonstrated and a scaled-up version is proposed that demonstrates an order of magnitude higher energy efficiency when compared to conventional hardware. In addition to improving ML implementations with optics, ML can be used to solve problems in optics, including in holography. Conventional holography uses a phase hologram to generate a target intensity pattern upon diffraction. A novel alternative is introduced: holography using only polarization. Conventional phase retrieval algorithms are insufficient to optimize such a hologram. This work demonstrates the use of gradient based optimization of neural networks incorporating a differentiable numerical model of polarized light propagation to optimize for a target intensity distribution as well as the joint optimization for a target intensity and polarization distribution post diffraction.

ZUSAMMENFASSUNG

Maschinelles Lernen (ML) hat sich als leistungsstarkes Werkzeug zur Lösung komplexer Probleme in verschiedenen Bereichen, u.a. in der Optik, etabliert. Die Skalierbarkeit von Rechenressourcen und der Energieverbrauch größerer KI-Modelle stellen allerdings ein erhebliches Problem dar. Optische Systeme können zur energieeffizienten Implementierung von Matrix-Vektor-Multiplikationen, die für neuronale Netze essenziell sind, eingesetzt werden. Allerdings ist es schwierig optische Nichtlinearitäten zu implementieren. Diese Arbeit beschreibt ein optoelektronisches Bauelement das eine optische Matrix-Vektor-Multiplikation und eine elektronische Nichtlinearität bei niedrigen Lichtintensitäten implementiert. Die Funktion des Bauelements wird experimentell gezeigt, und es wird eine mögliche Hochskalierung vorgestellt, die die Energieeffizienz im Vergleich zu konventioneller Hardware um eine Größenordnung verbessert. Optik kann ML-Implementierungen verbessern und Maschinelles Lernen kann wiederum zur Lösung von Problemen in der Optik, beispielsweise in der Holographie, eingesetzt werden. Typischerweise verwendet die Holographie ein Phasenhologramm, um nach der Beugung das gewünschte Intensitätsmuster zu erzeugen. Als neuartige Alternative wird die Holographie mittels reiner Polarisationshologramme vorgestellt. Konventionelle Phasenrekonstruktionsalgorithmen sind jedoch unzureichend für die Optimierung solcher Hologramme. Diese Arbeit demonstriert die Anwendung Gradientenbasierter Optimierung neuronaler Netze. Diese Netze beinhalten ein differenzierbares numerisches Modell der Ausbreitung polarisierten Lichts und dienen der Optimierung einer Intensitätsverteilung sowie der gemeinsamen Optimierung einer Zielintensitäts- und Polarisationsverteilung nach der Beugung.

Dedicated to my father, whose love, wisdom, and strength
continue to guide me, even in his absence.

I miss you deeply and hope I have made you proud

Sasank Murty Kottapalli
(1956-2025)

ACKNOWLEDGMENTS

I would like to begin by thanking my family, without whom none of this would have been possible. My father and mother have always been my rock, standing by me during good times and bad, offering the stability I often lacked. I am eternally grateful for their unconditional love and support. I dearly wish my father could have been here today to witness this milestone, but I find comfort knowing he would be proud of the person I have become. I would also like to thank my brother and sister-in-law, who have been a constant source of strength, encouragement, and friendship throughout my journey.

I want to express my sincere gratitude to Peer for giving me an opportunity at a time when I felt lost and confused. Even after all these years of working together, I continue to be in awe of the remarkable person you are. There is still so much I have yet to learn from you, not only in science but also about becoming a better person. Thank you deeply for your invaluable guidance and unwavering support throughout my PhD.

To Alex—I am incredibly thankful for your patience and for putting up with me all these years. I genuinely could not have reached this milestone without your help and support. Your friendship has been invaluable, and I am grateful beyond words for everything.

I would also like to thank all my friends within the group—without you, this journey would have been impossible. The countless moments of laughter, joy, and friendship we shared made this challenging path enjoyable. Thank you for always being there for me when I needed it most.

Similarly, I want to extend my gratitude to my friends outside the group, who have been unwavering sources of encouragement and support. Your friendship has made me a better person and given me strength during challenging times.

Finally, I would like to sincerely thank the members of my committee—Prof. Wolfram Pernice, Prof. Tristan Bereau, and Prof. Johannes Schemmel—for taking time from their busy schedules to engage with my work and be here today. I greatly appreciate your interest, insightful feedback, and encouragement.

Thank you all, from the bottom of my heart.

CONTENTS

1 PREFACE

Machine learning (ML) has rapidly evolved from a niche research area into an omnipresent tool and part of modern life. Fundamentally, ML can be understood from a statistical perspective as the application of a numerical algorithm to learn a function directly from data rather than relying on explicitly programmed instructions [1]. This capability allows computers to detect patterns and structures in diverse data forms, such as spoken language [2], complex protein sequences [3], or vast astronomical images [4], often revealing insights difficult to discern otherwise. The modern success of ML can be explained by the availability of meticulously labeled large datasets (e.g., MNIST or CIFAR-10) that provide the empirical basis for supervised learning approaches [5, 6], the availability of powerful computational infrastructures capable of performing a large number of floating-point operations per second [7], and continuous algorithmic advancements that make optimizing highly complex, high-dimensional problems tractable [8].

Artificial neural networks (ANNs) are predominant in the field. These networks consist of layers of relatively simple, differentiable mathematical operators, referred to as neurons, composed together to form highly flexible and universal function approximators [9]. Deriving inference in such a network typically involves a feed-forward evaluation of layer-wise matrix-vector products interleaved with nonlinear transformations. Conversely, learning involves adjusting the weights within these matrices. Broadly, learning in ANNs falls into three generic paradigms [10, Sec. 1.2]:

- **Supervised learning** minimizes a loss function between the model's predictions and the ground-truth values via gradient optimization techniques. This thesis predominantly adopts this method in both parts because it provides well-defined optimization objectives [1].
- **Unsupervised learning** discovers latent structure within data without labels [11].
- **Reinforcement learning** aims to optimize a policy that dictates an agent's actions given a state, to maximize an expected long-term reward R within a specific environment [12].

Although unsupervised and reinforcement methods are central to areas like reinforcement learning agents and generative modeling, supervised learning forms the basis of many frequently used model types, including those for solving inverse problems in scientific applications [13]. We therefore restrict the subsequent discussion to this type, which is also the focus of this thesis. Classification remains the benchmark task for comparing different ML architectures, as metrics such as accuracy, precision, and recall on publicly available datasets can rigorously quantify performance. MNIST, a dataset of handwritten digits, serves as a classic benchmark

for evaluating various fully connected and convolutional neural network architectures' performance and is widely used, including in this work [5].

Training of a neural network is done through back-propagation, which uses the chain rule to evaluate the gradient of the loss function with respect to the weights [14, Sec. 4.3]. Stochastic Gradient Descent (SGD) then iteratively updates these weights [14, Sec. 5.9]. The choice of loss function (e.g., mean-squared-error loss, cross-entropy loss, or even custom-defined loss functions as discussed in Part II of this thesis), the learning-rate schedule, and applied regularization techniques shape the optimization landscape's specific characteristics.

However, the scale of modern ML presents challenges. As models scale into the multi-billion parameter regime, exemplified by systems like modern transformer-based large language models such as GPT-4.1, even a single forward pass can require an extraordinary number of multiply-accumulate (MAC) operations, on the order of $\mathcal{O}(10^{17})$, which leads to large energy consumption on conventional hardware [15]. The substantial carbon footprint from such energy demands has become a significant public and scientific concern. Large-scale models already accelerate scientific discovery in diverse areas, including protein-fold prediction [3], exa-scale climate simulation [16], and high-energy-physics triggers. Nevertheless, their ecological impact compels us to rethink fundamental aspects of architectures, compilation efficiency, and hardware design. This thesis proposes optics, with its inherent parallelism and potential for near-zero inference energy cost [17], as a complementary computational substrate that can offer favorable throughput-per-watt compared to conventional electronic accelerators.

Meanwhile optics serves as the foundation for a wide range of areas of science and technology. A critical research area within optics is wavefront engineering: manipulating amplitude, phase and polarization to shape radiation precisely with components such as spatial light modulators (SLMs) or static diffractive optical elements. These components typically impart a spatially varying modulation to phase and/or amplitude across an incident wavefront. Upon diffraction, this modulated wavefront yields a desired target output. While traditional holography has primarily concentrated on shaping only light intensity, the more advanced field of vectorial holography seeks to control all three key properties—amplitude, phase, and polarization—simultaneously.

The advanced control offered by vectorial holography forms the central focus of Part I of this thesis. Machine learning (ML) techniques are frequently used to tackle complex, often ill-posed, inverse problems in optics [18]. These problems include recovering object properties from the scattered field [19], compensating for aberrations when imaging through turbid media [20], or designing optical masks that generate a specific target light field [21]. While traditional optimization methods like the Gerchberg-Saxton algorithm [22] or those based on direct search [23] or on alternating projections [24], gradient based optimization is a newer alternative which involves embedding a differentiable Helmholtz solver within the optimization framework [25]. Part I of this thesis adopts precisely this strategy, utilizing back-propagation through an angular-spectrum propagator to learn spatially varying polarization masks that can implement complex optical fields using polarization holograms. The inverse optical design problems explored in Part I uti-

lize physics-informed differentiable models and structural-similarity (SSIM) losses, which often better align with perceptual quality [26].

As previously discussed, the substantial energy demands of large-scale ML models and their ecological impact compel the research community to develop new approaches to AI hardware. Consequently, Part II of this thesis explores new hardware approaches to reduce inference energy consumption. One of the approaches that is particularly relevant in this context is that of Optical Neural Networks (ONNs) . ONNs can be broadly classified into two types, namely photonic and free-space systems, depending on whether the light in the system is confined to a photonic chip or whether light diffracts freely in 3D space. The overarching goal of both of these implementations is to replace digital MACs with passive optical MACs [27]. Examples of photonic implementations include Mach-Zehnder Interferometer (MZI) meshes [28], which can perform unitary transforms [29], and microring resonators, which can weight Wavelength Division Multiplexing (WDM) channels [30]. Free-space approaches, such as deep diffractive neural networks (DDNN), have proven effective on tasks like classification without intermediate electronic calculation [31]. Such free-space methods are scalable and can implement large matrix-vector-multiplications with dense matrices for comparatively low energy. Nonetheless, challenges such as the implementation of low-energy non-linearities and the scalability of free-space approaches persist [27].

This thesis consists of two parts with each part addressing an aspect of the two approaches combining machine learning and optics as outlined above. The first part focuses on using machine learning techniques to optimize polarization holograms for achieving arbitrary intensity and polarization distributions on the target plane. The second part demonstrates a scalable and energy-efficient opto-electronic neural network (OENN) that implements a fully-connected neural network with a non-linear activation function. Together, these parts aim to develop novel tools and approaches that can be utilized in conjunction with or that aid machine learning.

PUBLICATIONS OF THE AUTHOR

Parts of this thesis have been published or submitted to a journal or are in preparation for publication

[1] S. N. Murty Kottapalli, A. Song, and P. Fischer, "Engineering wavefronts with machine learned structured polarization," arXiv:2203.11185v4, Submitted 2025.

[2] A. Song [†], S. N. Murty Kottapalli[†], R. Goyal, B. Schölkopf, and P. Fischer, "Low-power scalable multilayer optoelectronic neural networks enabled with incoherent light," Nature Communications **15**, [†]These authors contributed equally., 10692 (2024).

The following oral presentations were given by the author at conferences and workshops during the course of the PhD:

1. **European Optics Society Annual Meeting (EOSAM)**, Sep. 2023, Dijon, France *Image classification with a fully connected opto-electronic neural network*
2. **SPIE Photonics West**, Jan. 2024, San Francisco USA: *Fully connected optoelectronic neural network for image classification*
3. **SPIE Photonics West**, Jan. 2024, San Francisco, USA: *Polarization-based nonlinear deep diffractive neural networks*

In addition, the author has contributed to the following publications that are not part of the present thesis:

[1] D. Singh, A. Domínguez, U. Choudhury, S. Kottapalli, M. N. Popescu, S. Dietrich, and P. Fischer, "Interface-mediated spontaneous symmetry breaking and mutual communication between drops containing chemically active particles," Nature communications **11**, 2210 (2020).

[2] V. M. Kadiri, J.-P. Günther, S. N. Kottapalli, R. Goyal, F. Peter, M. Alarcón-Correa, K. Son, H.-N. Barad, M. Börsch, and P. Fischer, "Light-and magnetically actuated FePt microswimmers," The European Physical Journal E **44**, 74 (2021).

[3] T. Zinn, T. Narayanan, S. N. Kottapalli, J. Sachs, T. Sottmann, and P. Fischer, "Emergent dynamics of light-induced active colloids probed by XPCS," New Journal of Physics **24**, 093007 (2022).

DECLARATION OF GENERATIVE AI USE

I hereby declare that generative AI tools were used in the preparation of this thesis. Specifically, tools such as Elicit, Perplexity, and Google Gemini Deep Research were used for literature survey, while ChatGPT and Gemini were used for proofreading and assessing the readability of the text. No AI-generated text or figures have been included in this thesis. All text and figures presented are entirely original and created by me or cited appropriately when reproduced/adapted from an external source.

Part I

MACHINE LEARNING FOR POLARIZATION HOLOGRAM OPTIMIZATION

The following chapters are primarily based on the work presented in [32]. The conceptualization and experimental work were carried out in collaboration with Dr. Alexander Song, while I was responsible for the simulations and data analysis. I would also like to acknowledge Lennart Schlieder, Dr. Valentin Volchkov, and Prof. Bernhard Schölkopf for their valuable discussions and insights.

2

INTRODUCTION

Wavefront engineering, the control of a light field's spatial structure (amplitude, phase, polarization) [33, 34], is fundamental to modern optics. While traditional holography often relies on phase manipulation for intensity control [35], this work focuses on using spatially structured polarization modulation for wavefront shaping and generating desired amplitude patterns.

Engineered wavefronts enable diverse applications, including super-resolution microscopy and imaging through scattering media [36], optical trapping of micro/nano-objects [37], precision laser manufacturing [38], spatial multiplexing in communications [39], quantum information encoding [40], and probing light-matter interactions [41]. This control allows overcoming limitations of diffraction by creating by creating non-diffracting Bessel beams or vector beams.

However, achieving versatile dynamic wavefront control solely via polarization modulation faces challenges. Early methods offered limited control [42, 43], and while specific structures like vortices [44] or focal patterns [45] were generated, full dynamic reconfigurability lagged behind phase-based systems [46]. Static elements like metasurfaces provide sophisticated polarization dependent shaping [47–49] but lack dynamic control and involve complex fabrication. Continued progress requires advancing both static and dynamic modulation techniques [14, 50], leveraging recent improvements in dynamic modulators (SLMs, DMDs) and algorithms [51, 52]. This thesis introduces an approach using dynamically addressable liquid crystal devices to address this gap.

This chapter provides the necessary background, covering fundamental properties of wavefront engineering with polarized light, modulation hardware like liquid crystal SLMs, and methods for generating structured light (e.g., Bessel, vector beams) via spatially varying polarization. Our methodology extends previous work on polarization phenomena [44], offering a general framework to synthesize arbitrary intensity patterns using polarization masks.

2.1 CONVENTIONAL WAVEFRONT MODULATION TECHNIQUES

Light can be shaped by modulating its amplitude, phase, or polarization. Traditional methods of wavefront shaping, which are well-established techniques in optics, primarily involve amplitude and phase modulation. This section provides a brief introduction to these traditional methods. Subsequently, it details the generation of Bessel beams as an example application. This will serve as a comparison for the alternative polarization-based modulation technique introduced later.

The most direct approach, amplitude modulation, involves modifying the transmittance or reflectance across the beam's profile, thereby shaping its intensity dis-

tribution. Mathematically, this operation multiplies the incident field $E_{in}(x, y)$ by a transmission function $T(x, y)$, where $0 \leqslant |T(x, y)| \leqslant 1$:

$$E_{out}(x, y) = T(x, y)E_{in}(x, y) \tag{1}$$

Amplitude modulation is commonly achieved using either static, prefabricated binary amplitude masks or dynamic devices such as Digital Micromirror Devices (DMDs) [53] or Spatial Light Modulators (SLMs). However, a key limitation of amplitude modulation is energy inefficiency: any region where $|T(x, y)| < 1$ results in partial absorption or scattering of light, leading to loss in total optical power [50]. Despite this drawback, amplitude masks remain useful, particularly in applications requiring high spatial resolution or binary contrast.

Phase modulation is a more common alternative where the phase profile $\phi(x, y)$ of the propagating wavefront is modulated while ideally preserving its amplitude. A phase-only transmission function takes the form:

$$E_{out}(x, y) = \exp[i\phi(x, y)]\, E_{in}(x, y) \tag{2}$$

Phase modulation is physically implemented by inducing spatial variations in the optical path length (OPL) when the wavefront passes through a medium. Widely used optics such as lenses and gratings modulate the wavefront by introducing a spatially varying phase shift. It is also possible to design a diffractive optical element (DOE) to create more complex wavefronts [54]. However, these elements are typically static and require complex fabrication processes. For dynamic applications such as computer-generated holography, electronically tunable LC-SLMs are used, resulting in a tunable phase shift [55]. Wavefront shaping using phase modulation is more energy efficient than amplitude methods due to its potential for high transmission efficiency [50]. Consequently, it plays a central role in beam shaping, adaptive optics, holography, and generating structured beams.

2.2 BESSEL BEAMS

Among the structured beams enabled by wavefront modulation, Bessel beams are an important class of non-diffracting optical wavefronts. They are distinguished by their ability to maintain their transverse intensity profile over extended propagation distances, unlike conventional diffracting Gaussian beams [56]. Their unique transverse intensity profile is mathematically described by a Bessel function of the first kind, $J_0(k_\perp \rho)$, where $k_\perp$ is the transverse wavevector component and ρ is the radial coordinate. A further key characteristic is their self-healing property: the ability to reconstruct their profile when a small part of the propagating beam is blocked. [57].

Ideal Bessel beams formed by the interference of waves composing a cone of infinite extent, require infinite energy and are not physically realizable. Therefore,

practical applications implement *pseudo-Bessel beams*, which are generated using finite apertures [58]. These beams exhibit the characteristic Bessel-like properties (non-diffraction, J_0 profile) over a finite axial range, denoted z'_{max}. Common generation methods include axicons [59], SLM-based phase masks [60], metasurfaces [61], and amplitude masks like ring gratings [62]. The polarization modulation method explored in this thesis, as will be detailed, is analogous in principle to using amplitude or phase masks to create the necessary interference conditions. The following subsections delve into the physical principles underlying their formation.

2.2.1 *Interference and Non-Diffracting Profiles*

This section closely follows and reproduces parts of the discussion in Ref. [63, Section 6.4.3.1]. To understand the origin of the non-diffracting nature of Bessel beams, it is instructive to consider the role of interference. We can illustrate this with a simplified one-dimensional (1D) case involving two coherent point sources located at $(\pm x_0, 0, 0)$ emitting light with wavelength λ and wavenumber $k = 2\pi/\lambda$. In the far-field ($z \gg x, x_0$) and paraxial ($x \ll z$) regime, the superposition of the spherical waves yields a complex amplitude $E(x, z)$. The resulting intensity $I = |E|^2$ exhibits a sinusoidal modulation (Similar to Ref. [63, Eq. 6.35]):

$$I(x, z) \propto \frac{1}{z^2} \cos^2\left(\frac{kxx_0}{z}\right) \propto \frac{1}{z^2}\left[1 + \cos\left(\frac{2kxx_0}{z}\right)\right] \tag{3}$$

Here, the term kx_0/z represents the transverse wavevector component $k_x = k\sin(\theta) \approx k(x_0/z)$, where θ is the angle each wavevector makes with the propagation axis [64]. The intensity profile $I(x) \propto 1 + \cos(2k_x x)$ thus possesses a transverse structure determined by k_x. Crucially, within the paraxial approximation, this structure maintains its form along the propagation direction z (aside from the overall $1/z^2$ intensity scaling). This simple example illustrates how specific interference patterns generated by appropriately arranged sources can resist diffraction. Bessel beams, in essence, arise from a continuous superposition of waves corresponding to sources distributed on a ring, leading to a conical interference pattern responsible for their non-diffracting and self-healing properties [56].

2.2.2 *Bessel Beam Formation via Diffraction from Circular Structures*

This section closely follows the discussion in Ref. [56, 57].

Building on the interference principle, a Bessel beam can be formed by generating an optical field whose angular spectrum consists of plane waves with wave vectors lying on the surface of a cone. This cone is characterized by a constant transverse wavevector magnitude $k_\perp = k\sin(\theta)$. Diffracting a plane wave from structures possessing circular symmetry (such as ring gratings or axicons) is a common method to produce such an angular spectrum.

Specifically, using scalar diffraction theory in the Fraunhofer regime (considering propagation to a distance z, or equivalently, observation in the back focal plane of a lens where $z = f$), the far-field complex amplitude $E(\rho, \phi, z)$ is proportional to the two-dimensional (2D) Fourier Transform (FT) of the field $E_{ap}(r', \phi')$ immediately after the diffracting element [56][65, Sec. 10.5]:

$$E(\rho, \phi, z) \propto \frac{e^{ikz}}{i\lambda z} e^{i\frac{k}{2z}\rho^2} \mathcal{F}\{E_{ap}(r', \phi')\}\Big|_{f_\rho = \rho/(\lambda z)} \tag{4}$$

where $f_\rho = \sqrt{f_x^2 + f_y^2}$ represents the radial spatial frequency. For instance, if the element is a circular grating with period P, diffraction orders occur at specific spatial frequencies. By selecting the m^{th} diffraction order, we isolate components with $k_\perp = |m|(2\pi/P) = k\sin(\theta)$. In the far-field, these components manifest as an annular ring at a radius $\rho \approx z\sin(\theta) = z(k_\perp/k)$.

This ring of light then acts as the effective source for the subsequent propagation. The field $E_B(\rho, \phi, z')$ propagating downstream from this ring results from the coherent superposition of these conical waves. Assuming a uniform amplitude A_0 around the ring, this superposition evaluates to the zeroth-order Bessel beam [37, Eq. 2]:

$$E_B(\rho, \phi, z') \propto A_0 J_0(k_\perp \rho) \exp(ik_z z') \tag{5}$$

Here, $\mathbf{k}(\phi')$ represents a wave vector on the cone with transverse component $k_\perp$ and longitudinal component $k_z = \sqrt{k^2 - k_\perp^2}$, and $\mathbf{r}' = (\rho, \phi, z')$ is the position vector in cylindrical coordinates. The term in the exponent is the dot product $\mathbf{k}(\phi') \cdot \mathbf{r}' = k_\perp \rho\cos(\phi' - \phi) + k_z z'$. The evaluation of the integral in Eq. (5), under the assumption of uniform amplitude $A(\phi')$ (absorbed into A_0), relies on the integral representation of the zeroth-order Bessel function of the first kind:

$$J_0(x) = \frac{1}{2\pi}\int_0^{2\pi} e^{ix\cos(\theta - \alpha)}\, d\theta \tag{6}$$

Identifying $x = k_\perp \rho$, the integration variable $\theta = \phi'$, and $\alpha = \phi$, the integral over ϕ' in Eq. (5) yields $2\pi J_0(k_\perp \rho)$, resulting in the proportionality shown.

The resulting intensity profile is thus ([56]:

$$I(\rho) \propto |A_0|^2 [J_0(k_\perp \rho)]^2 = |A_0|^2 \left[J_0\left(\frac{2\pi|m|}{P}\rho\right)\right]^2 \tag{7}$$

It is worth reiterating that real beams generated with finite apertures (e.g., of characteristic size w) are quasi-Bessel beams, exhibiting this characteristic profile over a finite propagation distance $z'_{max} \approx w/\tan(\theta)$.

2.2.3 *Generation via Spatially Varying Polarization Rotation*

Having established the principle that Bessel beams arise from a conical angular spectrum, we now discuss how this spectrum can be generated using spatially varying polarization modulation via an SLM, the primary technique investigated in this thesis. The objective is to encode a specific radial structure into the polarization state of an incident beam, such that its subsequent diffraction pattern yields the desired annular ring in the far-field.

To achieve this, we first calculate a target radial cosine pattern $M(r) \propto \cos(k_{mask} r)$, where the spatial frequency k_{mask} is chosen to correspond to the desired transverse wavevector $k_\perp$ of the Bessel beam. This pattern is then used to drive the SLM to impart a spatially varying polarization rotation $\theta(r)$ to the incident light. Assuming a linear response of the SLM [66] within an aperture of radius R_{max}, we can approximate the imparted rotation by its fundamental component:

$$\theta(r) \approx \theta'_{max} \cos(k_{mask} r) \cdot \mathrm{circ}(r/R_{max}) \tag{8}$$

Where, circ is a 2D circle function that can be expressed as [14, Page 4]:

$$\mathrm{circ}(r/R_{max}) = \begin{cases} 1 & r < R_{max} \\ 0 & r \geqslant R_{max} \end{cases} \tag{9}$$

Consider linearly polarized input light, represented by the Jones vector $\mathbf{E}_{in} = E_0[1, 0]^T$. The SLM acts as a spatially varying rotator, described by the Jones matrix $\mathbf{R}(\theta(r))$ [67, Page 343]:

$$\mathbf{E}_{in} = E_0 \begin{pmatrix} 1 \\ 0 \end{pmatrix} \tag{10}$$

$$\mathbf{R}(\theta(r)) = \begin{pmatrix} \cos\theta(r) & -\sin\theta(r) \\ \sin\theta(r) & \cos\theta(r) \end{pmatrix} \tag{11}$$

The output field immediately after the SLM (at $z = 0$) is consequently a vector beam with a radially varying linear polarization state:

$$\mathbf{E}_{out}(r) = \mathbf{R}(\theta(r))\mathbf{E}_{in} = E_0 \begin{pmatrix} \cos\theta(r) \\ \sin\theta(r) \end{pmatrix} \cdot \mathrm{circ}(r/R_{max}) \tag{12}$$

The Cartesian components of this field within the aperture are $E_{out,x}(r) = E_0 \cos(\theta(r))$ and $E_{out,y}(r) = E_0 \sin(\theta(r))$.

To determine the resulting far-field pattern, we apply the Fraunhofer diffraction integral (Eq. (4)) to each Cartesian component. Owing to the radial symmetry of the problem, this calculation transforms into the Hankel transform of order zero ($\mathcal{H}_0$) [14, Sec. 2.1.4 and Sec. 2.1.5] for more information). The far-field components at a radial spatial frequency $q = \rho/(\lambda z)$ are thus given by:

$$E_{far,x}(q) \propto \mathcal{H}_0\{E_0 \cos[\theta(r)] \cdot \mathrm{circ}(r/R_{max})\} \tag{13}$$

$$E_{far,y}(q) \propto \mathcal{H}_0\{E_0 \sin[\theta(r)] \cdot \mathrm{circ}(r/R_{max})\} \tag{14}$$

While evaluating these transforms analytically can be complex, the crucial point is that the underlying radial modulation frequency k_{mask} embedded in the polarization rotation $\theta(r)$ causes the energy of both far-field components, $E_{far,x}$ and $E_{far,y}$, to become concentrated around the spatial frequency $q \approx k_{mask}/(2\pi)$. This corresponds precisely to the desired transverse wavevector $k_\perp \approx k_{mask}$. Consequently, the total far-field intensity, given by the sum of the squared magnitudes of the components:

$$I_{far}(q) \propto |E_{far,x}(q)|^2 + |E_{far,y}(q)|^2 \tag{15}$$

exhibits an annular ring at the target radial wavevector. This ring signifies the presence of the conical angular spectrum necessary for forming the pseudo-Bessel beam profile $J_0(k_\perp \rho)$ upon further propagation. This specific approach to Bessel beam generation is experimentally verified in Sec. 3.3.1.

2.3 LIQUID CRYSTAL BASED SPATIAL LIGHT MODULATORS

Liquid crystals (LCs) exhibit properties intermediate between conventional liquids and solid crystals, characterized notably by the long-range orientational order of their constituent molecules. This order is described by an average molecular orientation, the director **n**, and leads to optical anisotropy, specifically birefringence. Light polarized parallel and perpendicular to **n** experiences extraordinary (n_e) and ordinary (n_o) refractive indices, respectively. The difference, $\delta n = n_e - n_o$, results in a relative phase retardation for polarized light traversing the material [51].

For Spatial Light Modulator (SLM) applications, the LC director, and thus the effective birefringence, is controlled by an external electric field. This control stems from the dielectric anisotropy ($\delta\epsilon = \epsilon_\parallel - \epsilon_\perp$) of LC molecules. SLMs typically utilize LCs with positive dielectric anisotropy ($\delta\epsilon > 0$), causing their long axes to align parallel to an applied field [68, Page 11]. Applying a voltage across an LC layer reorients the molecules; an electrostatic torque overcomes internal elastic and surface anchoring forces, thereby changing the director orientation [68, Section 5.1.3]. Pixelated electrodes allow for spatial variation of this voltage, leading to spatially modulated birefringence and thus pixel-wise control over the phase and/or polarization of an incident wavefront.

The specific LC cell design, including electrode configuration and alignment layers, dictates the SLM's function. Alignment layers determine the initial LC configuration (e.g., a nematic phase arrangement) in the absence of an electric field [51]. This thesis primarily utilizes two types of LC-SLMs: Twisted Nematic LC-SLMs (TNLC-SLMs) for polarization modulation and Liquid Crystal on Silicon SLMs (LCoS-SLMs) for phase modulation. Their distinct operating principles and applications are detailed in the following sections.

2.3.1 *Twisted Nematic Liquid Crystal Spatial Light Modulators*

Most of the discussion presented here has been adapted from Ref. [68, Section 5.1] and reproduced here.

TNLC-SLMs, used in this work primarily for polarization rotation, feature a nematic LC layer situated between two transparent, electrode-coated substrates. Orthogonal alignment layers on these substrates create a helical twist of the LC director, typically 90 degrees, through the cell's thickness [69, 70]. In the absence of an applied field (zero-voltage state), linearly polarized light with its polarization axis aligned parallel to the input LC director undergoes polarization rotation. This occurs due to adiabatic following, or waveguiding, of the light's polarization vector along the twisted LC structure, provided the Mauguin condition ($\delta n \cdot d \gg 0.5\lambda$, where d is cell thickness and λ is the light wavelength) is satisfied [68, Sec. 5.1.2]. For a standard 90° twisted cell, this results in a 0.5π rotation of the polarization plane.

An applied voltage generates an electric field perpendicular to the substrates. As detailed in Sec. 2.3 (for LCs with $\delta\epsilon > 0$), this field tilts the molecules towards the field direction, thereby disrupting the helical twist and reducing the effective twist angle experienced by light [69, 70]. Consequently, the polarization rotation becomes voltage-dependent. At zero voltage, full rotation occurs. Increasing voltage reduces the effective twist and thus the rotation, until at sufficiently high voltages, the twist is nearly eliminated, and the light's polarization state remains largely unchanged. Importantly, this voltage-induced tilting also alters the effective refractive index due to LC birefringence. This coupling inevitably leads to a phase modulation effect occurring alongside the primary polarization rotation, which can be a limiting factor for applications requiring pure polarization control.

The propagation of polarized light through a TNLC-SLM can be described by Jones calculus. The Jones matrix for a TNLC cell relates the output electric field $\mathbf{E}_{out}$ to the input field $\mathbf{E}_{in}$, encapsulating the combined effect of the effective twist angle (ϕ) and phase retardation (Γ), both of which are functions of the applied voltage V [68, 71]:

$$\mathbf{E}_{\text{out}} = \begin{pmatrix} \cos\phi & -\sin\phi \\ \sin\phi & \cos\phi \end{pmatrix} \times \begin{pmatrix} \cos\left(\frac{\Gamma}{2}\right) - i\sin\left(\frac{\Gamma}{2}\right)\cos(2\theta_0) & -i\sin\left(\frac{\Gamma}{2}\right)\sin(2\theta_0) \\ -i\sin\left(\frac{\Gamma}{2}\right)\sin(2\theta_0) & \cos\left(\frac{\Gamma}{2}\right) + i\sin\left(\frac{\Gamma}{2}\right)\cos(2\theta_0) \end{pmatrix} \mathbf{E}_{\text{in}} \tag{16}$$

Here, θ_0 represents the angle of the director at the input face. The first matrix represents rotation by $\phi(V)$, while the second describes retardation effects due to $\Gamma(V)$. (For unpolarized or partially polarized light, a Mueller matrix treatment is necessary.) The TNLC-SLM used in this work was operated under conditions optimized for polarization rotation, allowing retardation effects to be largely ignored for the primary analysis.

2.3.2 *Liquid Crystal on Silicon Spatial Light Modulators*

LCoS-SLMs are reflective devices, predominantly employed for phase modulation, and were used for such purposes in this research. An LCoS device consists of an LC layer sandwiched between a transparent glass cover coated with a continuous electrode, and a silicon backplane. This backplane, fabricated using CMOS processes, contains a high-density array of individually addressable pixel electrodes that also act as reflective mirrors [72]. Light enters through the cover glass, traverses the LC layer, reflects off a pixel mirror, and passes through the LC layer again before exiting. As described in Sec. 2.3, the voltage applied to each pixel controls the LC orientation within that pixel.

For phase-only modulation, a parallel-aligned nematic (PAN) LC configuration is typically used, where the LC director is initially aligned parallel to the electrode planes. The incident wavefront's polarization must be aligned with this director axis. An applied voltage tilts the LC molecules (as per Sec. 2.3), directly altering the effective refractive index $n_{\text{eff}}(V)$ experienced by the light polarized along the director. This tilt reduces $n_{\text{eff}}(V)$ from its maximum (n_e) towards its minimum (n_o), inducing a controllable phase shift $\Delta\phi$ [73]:

$$\Delta\phi = \frac{2\pi}{\lambda}\Big(2d_{LC}\Delta n_{\text{eff}}(V)\Big) = \frac{4\pi d_{LC}}{\lambda}\Delta n_{\text{eff}}(V) \tag{17}$$

Here, d_{LC} is the LC layer thickness, $\Delta n_{\text{eff}}(V) = n_{\text{eff}}(V) - n_o$ is the voltage-dependent change in the effective refractive index relative to the ordinary index, and λ is the light wavelength. The factor of 2 in $2d_{LC}$ accounts for the double pass due to reflection. Device parameters are chosen to achieve a phase modulation range typically up to or exceeding 2π radians. The relationship between the applied grayscale value (controlling pixel voltage) and the resultant phase shift is generally non-linear and is typically addressed through calibration data or look-up tables provided by the manufacturer or determined empirically [71]. The Holoeye Pluto LCoS-SLM was utilized in this work for experiments requiring precise phase modulation [74].

2.4 MEASURING WAVEFRONTS

Beyond intensity measurements, characterizing the phase profile of a propagating wavefront is crucial for comprehensive wavefront engineering. This is particularly important for evaluating the performance of optical elements like Spatial Light Modulators (SLMs), where a coupled polarization and phase modulation is obtained. The Shack-Hartmann Wavefront Sensor (SHWS) is a widely adopted instrument for such phase measurements.

The SHWS, building upon earlier concepts such as Hartmann's screen [75] and further developed for modern applications, employs a microlens array conjugate to the target wavefront plane, positioned before a position-sensitive detector. All lenslets in the array share an identical focal length. A planar wavefront incident on this array produces a regular grid of focal spots on the detector, as each lenslet focuses light onto its optical axis. Conversely, an aberrated wavefront presents varying local slopes across the microlenses. These local tilts displace the focal spots from their reference (planar wave) positions. This displacement vector is directly proportional to the average local wavefront slope across the lenslet aperture and the focal length f of the lenslet [76–78]:

$$\Delta i(x_j) = f \cdot \partial_i W(x_j) \tag{18}$$

Here, $\Delta i(x_j)$ represents the i-th component ($i = x$ or y) of the focal spot displacement for the lenslet centered at position x_j, f is the lenslet focal length, and $\partial_i W(x_j)$ denotes the local wavefront slope component $\partial W/\partial i$ averaged over the j-th lenslet [78]. Measuring these displacements $\Delta i(x_j)$ for all lenslets yields a map of local wavefront slopes, from which the complete wavefront phase profile W can be reconstructed.

The wavefront phase profile W is reconstructed from these measured slopes, typically by decomposing it into a sum of orthogonal basis functions. Zernike polynomials (Z_n) are most commonly used for this purpose as they effectively represent various optical aberration modes [78]:

$$\partial_i W(x_j) \approx \sum_{n=0}^{N-1} a_n \, \partial_i Z_n(x_j) \tag{19}$$

Fitting the measured slope data (derived from $\Delta i(x_j)$) using the known derivatives of these Zernike polynomials (as shown conceptually in Eq. (19)) allows determination of the modal coefficients a_n. This process leads to a complete reconstruction of the incident wavefront phase profile W.

While the SHWS technique is broadly applied in optical system design, characterization, and adaptive optics [79, 80], in the context of this work, it is primarily employed to measure the phase of the wavefront after modulation by an SLM. The specific SHWS used is a commercial device, the Thorlabs WFS20-K1/M. This sensor system is factory-calibrated and includes software to reconstruct the complete

wavefront using a Zernike polynomial expansion based on the measured slopes, consistent with the approach described.

2.5 COMPUTATIONAL METHODS FOR WAVEFRONT SHAPING

Beyond hardware, effective wavefront shaping relies heavily on computational methods. Precisely modeling, predicting, and controlling wavefront propagation is crucial in modern optics and photonics research, enabling significant advances in diverse areas such as high-resolution imaging, optical manipulation, and quantum technologies. In the context of this work, the ability to simulate the propagation of polarized light is essential for further optimization of the generated wavefront.

Light propagation modeling can be broadly categorized into scalar and vector theories. Scalar diffraction theories offer a simplification by treating the light wave as a single scalar field, neglecting its vector polarization state. These theories are computationally efficient and often provide sufficient accuracy when the diffracting structures are large compared to the wavelength and when polarization effects are minor [81]. However, when light interacts with structures comparable in size to the wavelength (subwavelength structures), propagates through high numerical aperture (NA) systems, or in situations where the polarization state itself is critical, vector diffraction theories that adhere more closely to Maxwell's equations become indispensable [65, 82].

2.5.1 *Diffraction Theory*

The discussion of diffraction theory presented in this section is reproduced in part from Ref. [65, Chapter 10].

While Maxwell's equations provide the complete and fundamental description of the wave nature of light, obtaining solutions from the full set of coupled vector equations for complex propagation scenarios is often computationally prohibitive. As a consequence, scalar diffraction theories are widely employed. These theories simplify the problem by treating light as a scalar quantity, thereby focusing on the spatial distribution of the wave's amplitude and phase while neglecting its vector polarization state. This simplification is a practical compromise for many optical analyses. A foundational concept within scalar diffraction is the Huygens-Fresnel principle. This principle states that every point on a given wavefront can be regarded as a source of secondary spherical wavelets. The subsequent superposition of these wavelets then determines the form of the wavefront at a later time. This principle is mathematically formulated as the Huygens-Fresnel diffraction integral [65, Eq. 10.1]:

$$E(x,y,z) = -\frac{i}{\lambda}\iint_{\text{aperture}} E(x',y',0)\frac{e^{ikR}}{R}dx'dy' \tag{20}$$

In this expression, R denotes the distance from a point (x', y') on the aperture to the observation point (x, y, z), and is given by [65, Eq. 10.2]:

$$R = \sqrt{(x - x')^2 + (y - y')^2 + z^2} \tag{21}$$

The term $E(x', y', 0)$ within the integral represents the initial field distribution at the aperture plane (taken as $z = 0$), and $k = 2\pi/\lambda$ is the wave number corresponding to a wavelength λ. Gustav Kirchhoff subsequently provided a more rigorous derivation of a similar diffraction formula, which originates from the scalar Helmholtz equation ($\nabla^2 E + k^2 E = 0$). Kirchhoff's formulation, known as the Fresnel-Kirchhoff diffraction formula, introduces an obliquity factor that accounts for the directionality of the emitted wavelets. Despite this increased rigor, the resulting integral often remains challenging to evaluate analytically for arbitrary aperture shapes, and it also depends on certain approximations at the boundary[65, Page 266]. Nevertheless, for many common forward propagation scenarios, the obliquity factor is approximately unity. In these situations, the Huygens-Fresnel formula (Eq. (20)) serves as a widely accepted and practical starting point for diffraction analysis.

To further simplify the diffraction integral, particularly under specific geometrical conditions, additional approximations are introduced. These lead to the Fresnel and Fraunhofer diffraction regimes.

The Fresnel approximation is applicable to near-field diffraction scenarios. These are typically characterized by an observation distance z in the same range as the dimensions of the aperture and the wavelength. This approximation simplifies Eq. (20) by taking $R \approx z$ in the denominator term, and by utilizing a quadratic expansion for R in the exponent of the phase term e^{ikR} [65, Eq. 10.12]:

$$R \approx z \left[1 + \frac{(x - x')^2 + (y - y')^2}{2z^2}\right] \quad \text{(for the exponent)} \tag{22}$$

Substituting these approximations into the Huygens-Fresnel formula (Eq. (20)) results in the Fresnel diffraction integral [65, Eq. 10.13]:

$$E(x, y, z) \approx -\frac{ie^{ikz} e^{i\frac{k}{2z}(x^2 + y^2)}}{\lambda z} \iint_{\text{aperture}} E(x', y', 0) e^{i\frac{k}{2z}(x'^2 + y'^2)} e^{-i\frac{k}{z}(xx' + yy')} dx' dy' \tag{23}$$

Although this integral form remains complex, it can be identified as a convolution. This mathematical structure is advantageous as it allows for efficient numerical computation, often performed using algorithms based on the Fast Fourier Transform (FFT).

In contrast, the Fraunhofer approximation is employed under far-field conditions. This regime applies when the observation distance z is significantly larger

than the characteristic dimensions of the aperture relative to the wavelength (specifically, when $z \gg \frac{k}{2}(\text{aperture radius})^2$). Under these conditions, a further simplification is made to the phase term involving the source coordinates within the Fresnel integral (Eq. (23)), as given by [65, Eq. 10.17]:

$$e^{i\frac{k}{2z}(x'^2+y'^2)} \approx 1 \tag{24}$$

Applying this simplification yields the Fraunhofer diffraction integral [65, Eq. 10.19]:

$$E(x,y,z) \approx -\frac{ie^{ikz}e^{i\frac{k}{2z}(x^2+y^2)}}{\lambda z}\iint_{\text{aperture}} E(x',y',0)e^{-i\frac{k}{z}(xx'+yy')}dx'dy' \tag{25}$$

The Fraunhofer integral shows that the far-field complex amplitude distribution is proportional to the two-dimensional Fourier transform of the aperture function $E(x',y',0)$. This transform is evaluated at spatial frequencies $f_x = x/(\lambda z)$ and $f_y = y/(\lambda z)$. This direct relationship with the Fourier transform also facilitates efficient computation using FFT-based numerical methods.

More general propagation methods are often necessary for more complex scenarios or when higher accuracy is required across different diffraction regimes. For situations requiring broader applicability and computational efficiency, particularly in iterative optimization contexts, the Angular Spectrum Method (ASM) has become an important technique [83]. This work primarily employs the ASM for simulating wavefront propagation, as detailed in the following section.

2.5.2 *Angular Spectrum Method*

The discussion in this section is reproduced in part from Ref. [63, Chapter 6].

The Angular Spectrum Method (ASM) offers an accurate and computationally efficient approach for simulating wave propagation, well-suited for numerical implementation. This method is based on the principle of decomposing a complex scalar field, $U(x,y,0)$, at an initial plane (conventionally $z = 0$) into an infinite sum of plane waves. Each constituent plane wave has unique spatial frequency components (k_x, k_y) and a corresponding longitudinal component k_z. This decomposition is mathematically realized by applying a two-dimensional Fourier transform to the initial field $U(x,y,0)$, yielding its angular spectrum, denoted as $A(k_x,k_y;0)$ [63, Eq. 6.2]:

$$A(k_x,k_y;0) = \iint_{-\infty}^{\infty} U(x,y,0)e^{-i(k_x x+k_y y)}\,dx\,dy \tag{26}$$

This operation can also be concisely expressed as $A(k_x,k_y;0) = \mathcal{F}\{U(x,y,0)\}$, where $\mathcal{F}$ denotes the Fourier transform operator. Here, k_x and k_y represent

the transverse angular spatial frequencies. The longitudinal component of the wavevector, k_z, is constrained by the Helmholtz equation, which in a homogeneous medium with wavenumber $k = 2\pi/\lambda$ (where λ is the wavelength in the medium) dictates the dispersion relation $k^2 = k_x^2 + k_y^2 + k_z^2$. Thus, k_z is determined as [63, Eq. 6.8]:

$$k_z = \sqrt{k^2 - k_x^2 - k_y^2} \tag{27}$$

If $k_x^2 + k_y^2 > k^2$, k_z becomes imaginary. Such components correspond to evanescent waves, which experience rapid attenuation along the z-axis and do not propagate into the far-field.

The propagation of the angular spectrum over a distance z into the medium is described by multiplying each plane wave component $A(k_x, k_y; 0)$ by a phase factor, which constitutes the ASM transfer function for free-space propagation, $H(k_x, k_y, z)$:

$$H(k_x, k_y, z) = e^{ik_z z} = e^{i\sqrt{k^2 - k_x^2 - k_y^2}\, z} \tag{28}$$

The complex scalar field $U(x, y, z)$ at the new plane z is then reconstructed by performing an inverse Fourier transform on the propagated angular spectrum $A(k_x, k_y; z) = A(k_x, k_y; 0) \cdot H(k_x, k_y, z)$ (Adapted from [65, Eq. 6.11]):

$$\begin{aligned} U(x, y, z) &= \frac{1}{(2\pi)^2} \iint_{-\infty}^{\infty} A(k_x, k_y; 0) H(k_x, k_y, z) e^{i(k_x x + k_y y)}\, dk_x\, dk_y \\ &= \mathcal{F}^{-1}\left[A(k_x, k_y; 0) \cdot H(k_x, k_y, z)\right] \end{aligned} \tag{29}$$

The factor of $1/(2\pi)^2$ is the standard normalization constant for the inverse Fourier transform when using angular frequencies.

For numerical implementation, the Angular Spectrum Method (ASM) uses the computationally efficient Fast Fourier Transform (FFT) algorithm. The process comprises of: (1) a forward FFT of the initial field $U(x, y, 0)$ to acquire its discrete angular spectrum $A(k_x, k_y; 0)$; (2) pointwise multiplication of this spectrum by the discrete transfer function $H(k_x, k_y, z)$; and (3) an inverse FFT to reconstruct the propagated field $U(x, y, z)$. This approach achieves a favorable $O(N^2 \log N)$ computational scaling for an $N \times N$ grid, significantly outperforming the $O(N^4)$ complexity of direct diffraction integral evaluations, thereby benefiting computationally intensive tasks like iterative optimization [84, 85]. However, it is important to be careful with sampling to avoid aliasing errors, which can come from rapid phase oscillations in $H(k_x, k_y, z)$, particularly for large propagation distances z or high spatial frequencies [86]. Zero-padding $U(x, y, 0)$ pre-FFT is a common mitigation strategy, enhancing frequency domain sampling of $A(k_x, k_y; 0)$ and H, though it increases array sizes and computational load, necessitating a balance between accuracy and resources.

The ASM is derived directly from the Rayleigh-Sommerfeld diffraction formula, an exact solution to the scalar Helmholtz equation under appropriate boundary conditions [87]. With adequate sampling, ASM generally offers superior accuracy and a broader range of validity than the more restrictive Fresnel and Fraunhofer approximations [88]. Consequently, its combination of computational efficiency and accuracy establishes ASM as a preferred method for simulating scalar wave propagation [63, 86].

2.5.3 *Extension to Polarized Light*

While scalar ASM effectively models propagation of a scalar wavefront, we would like to simulate the propagation of vector light beams for this work. This requires explicitly accounting for the polarization state across the wavefront. Although rigorous vector diffraction theories (e.g., Stratton-Chu theory [89], Richards-Wolf theory for high-NA focusing [90]) and direct numerical solutions of Maxwell's equations (e.g., FDTD, RCWA) offer high accuracy, their computational cost is typically prohibitive for iterative optimization routines [91]

Consequently, this work adopts a practical approach by extending the scalar ASM framework to simulate vectorial wavefronts. This is suitable as we primarily deal with linearly polarized light propagating through free space or isotropic optical elements. Under these conditions, orthogonal linear polarization components (e.g., E_x and E_y) propagate independently without coupling. The source electric field $\mathbf{E}_{\text{source}}(x, y)$ is represented as a Jones vector:

$$\mathbf{E}_{\text{source}}(x, y) = \begin{pmatrix} E_{x,\text{source}}(x, y) \\ E_{y,\text{source}}(x, y) \end{pmatrix} \tag{30}$$

Each scalar component is then propagated independently using the scalar ASM operator, denoted $\mathcal{AS}$:

$$\mathbf{E}_{\text{propagated}}(x, y, z) = \begin{pmatrix} \mathcal{AS}\{E_{x,\text{source}}(x, y)\} \\ \mathcal{AS}\{E_{y,\text{source}}(x, y)\} \end{pmatrix} = \begin{pmatrix} E_{x,\text{propagated}}(x, y, z) \\ E_{y,\text{propagated}}(x, y, z) \end{pmatrix} \tag{31}$$

Subsequent interactions of these propagated components with optical elements like polarizers are modeled using standard Jones matrix multiplication, incurring minimal additional computational overhead. Finally, the resultant intensity $I(x, y, z)$ for intensity-sensitive detectors is obtained by coherently summing the squared magnitudes of the independently propagated vector components:

$$I(x, y, z) = |E_{x,\text{propagated}}(x, y, z)|^2 + |E_{y,\text{propagated}}(x, y, z)|^2 \tag{32}$$

This component-wise scalar propagation approach provides a computationally tractable means to simulate the essential aspects of vector beam propagation, maintaining sufficient accuracy for the iterative optimization tasks involving polarized light undertaken in this research.

2.6 OPTIMIZATION FOR WAVEFRONT SHAPING

The ability to computationally model wavefront propagation, as discussed in the previous section, leads naturally to the inverse problem of optimization for wavefront shaping. The inverse problem consists of calculating a diffracting element which on illumination from a known light source leads to desired light field distribution at the target plane. While this inverse problem is challenging even for scalar fields (where the goal might be to control only intensity or phase), it becomes substantially more complex when dealing with vectorial beams, where the goal is to control the full polarization state across the field profile.

The problem for this specific instance can be formalized as follows: For a given target vector field, $E_t(x', y')$, specified by its complex amplitude distribution for each orthogonal polarization component (e.g., $E_{t,x}$ and $E_{t,y}$) on a designated target plane, we need to calculate the required modulation properties of a hologram mask located at the hologram plane. This task differs significantly from the conventional scalar holography problem, which typically targets only an intensity distribution $|E_t|^2$. As a result, the hologram mask must be designed to impart highly specific, spatially varying modulation to both the amplitude and phase of the incident light's two orthogonal polarization components.

Iterative algorithms are commonly used to solve such inverse problems in the context of wavefront shaping. In a traditional phase hologram, a randomly initiated phase modulation mask is iteratively improved until a criteria for the target intensity distribution is met [92]. In general, the process involves numerically propagating a wavefront between two or more related domains that are linked by a known linear transform, such as the Fourier transform [93, 94]. There are various phase retrieval algorithms such as the Gerchberg-Saxton (GS) algorithm [22] and the Fienup algorithm [95]. The following section introduces the GS algorithm as a slightly modified version which is used in this thesis to optimize a polarization hologram.

2.6.1 *Gerchberg–Saxton Algorithm*

The Gerchberg–Saxton (GS) algorithm solves the classic phase-retrieval problem in optics, where one knows only the intensity (or amplitude) of a wavefield in two planes linked by a propagation operator $\mathcal{P}$—for example, between a hologram plane and its far-field diffraction pattern [22]. Because intensity measurements discard the phase, GS iteratively reconstructs a phase distribution in the hologram plane, $E_h = A_h e^{i\phi_h}$, so that after propagation the resulting field in the target plane, $E_t = A_t e^{i\phi_t}$, matches a desired amplitude A_t.

Starting from an initial guess $\phi_h^{(0)}$ (often random), each iteration alternates between the two planes. First, one applies the forward transform $\mathcal{P}$ to the current hologram estimate $A_h e^{i\phi_h^{(k)}}$, then replaces its amplitude by A_t while keeping the computed phase. In the reverse step, this modified field is propagated back with $\mathcal{P}^{-1}$ and its amplitude is reset to A_h, retaining the updated phase. Symbolically, one full update reads:

$$\phi_h^{(k+1)} = \arg\Big(\mathcal{P}^{-1}\big[A_t\, e^{\,i\,\arg(\mathcal{P}[A_h\, e^{i\phi_h^{(k)}}])}\big]\Big). \tag{33}$$

By enforcing the known amplitudes in each domain and shuttling the phase back and forth, GS steadily reduces the mismatch between the generated and desired target intensities. In practice, one implements $\mathcal{P}$ and $\mathcal{P}^{-1}$ via FFTs for speed.

Despite its simplicity and efficiency, the GS algorithm can stall in local minima, especially for intricate target patterns [96]. More critically for polarization holography, its scalar formulation handles only single component amplitudes and cannot manage coupled vectorial constraints between orthogonal polarization channels [97]. However, we try to modify the traditional GS algorithm to optimize a polarization hologram, as discussed in Sec. 3.4. These limitations motivate more advanced, often machine-learning-based, strategies for designing true polarization holograms.

2.7 MACHINE LEARNING FOR WAVEFRONT SHAPING

Limitations of conventional phase retrieval methods, such as the Gerchberg-Saxton algorithm, have led to the adoption of machine learning (ML) techniques for solving the holographic inverse problems described earlier. Some of the earliest applications of ML in this context used Convolutional Neural Networks (CNNs) for tasks such as generating phase holograms and denoising holographic outputs [98, 99]. More recently, the use of differentiable physical propagation models, especially the angular spectrum algorithm, combined with neural network pipelines has enabled significant progress in complex wavefront shaping tasks where traditional methods fail [100, 101]. These advances have led to non-iterative solutions that match or even exceed the performance of classical algorithms.

ML-based optimization techniques have opened up new possibilities for phase retrieval problems that are difficult or impossible to solve with traditional methods. These include simultaneous optimization for multiple target patterns [21, 102–104] and solving for multiple conflicting objectives within a single process [105], significantly expanding the range of achievable applications [83]. These approaches have also proven effective in designing metasurfaces for generating vectorial holograms [106–108]. In the context of this work, ML has also been shown to support true vectorial wavefront shaping by jointly optimizing both phase and polarization masks [109].

In this work, we use ML techniques to optimize polarization modulation masks for wavefront engineering. The physics-informed optimization framework developed for this purpose explicitly models diffraction using the ASM (Sec. 2.5.2) and uses the standard backpropagation algorithm to learn the optimal mask parameters. The effectiveness of this optimization depends strongly on the choice of loss function. Standard metrics like Mean Squared Error (MSE) or cross-entropy often fall short for complex optical tasks. Custom loss functions, tailored to the specific application, are frequently necessary. In particular, image quality metrics such as

the Structural Similarity Index Measure (SSIM) [26], which captures structural information, luminance, and contrast, often provide better results than generic losses. These perceptually aligned metrics help ensure that the optimized output better matches human visual expectations or structural targets. Further details on the differentiable model and the specific loss function used are provided in Sec. 3.5.

Having established the theoretical basis for wavefront modulation, measurement, propagation modeling, and optimization in this chapter, we now turn to the next chapter, which presents the experimental setup used for this work and discusses the corresponding results.

3

MACHINE LEARNING OPTIMIZED POLARIZATION MODULATION FOR BEAM MANIPULATION

This chapter details how modulating the polarization across a propagating wavefront can be utilized to optimize that wavefront upon reaching a target plane. While traditional holographic methods often rely on amplitude or phase modulation, this work explores the potential of polarization. It is well-established that modulating the phase across a propagating wavefront allows the resulting diffraction pattern to be controlled for obtaining desired target intensity patterns. Indeed, numerous algorithms exist to optimize the required phase distribution for a given target amplitude distribution. It is also fundamental that any aperture in a beam path induces diffraction, an effect famously used to demonstrate the wave nature of light in experiments such as Young's double-slit experiment. This same effect occurs when placing an aperture in the path of a coherent light beam, where such an aperture essentially acts as a binary amplitude mask.

Building upon this principle, the concept can be extended to exploit diffraction resulting from an inhomogeneous polarization distribution across the wavefront. Since a polarized light beam can be decomposed into two orthogonal polarization states, spatially modulating the polarization state acts analogously to applying two distinct amplitude masks, one corresponding to each orthogonal component. Consequently, the resulting pattern observed after propagation is a superposition of the patterns generated by each component propagating independently. To demonstrate this, this chapter first presents an initial experimental test exploring the diffractive nature of spatially inhomogeneous polarization distributions. Subsequently, methods that can optimize such polarization modulation for specific applications. Furthermore, it demonstrates the effectiveness of this technique in achieving a desired joint distribution of both amplitude and polarization on the target plane, using only polarization as the control parameter. Although various techniques exist for modulating polarization, this work specifically employs a spatial light modulator (SLM) for this purpose. The specific experimental setup used is described in detail in the following section.

3.1 EXPERIMENTAL AND NUMERICAL TECHNIQUES

This section introduces the experimental setup and the numerical simulation techniques employed to obtain the results presented in this chapter.

3.1.1 *Experimental Techniques*

Fig. 1 provides a schematic illustration of this experimental setup. The wavefront originating from a 532 nm Gaussian laser source is modulated as follows. The initial laser beam width is 2.5 mm and then expanded by a telescopic beam expander to approximately 15 mm. The expanded beam is then directed onto a spatial light modulator (SLM) (Holoeye LC2012) [66], which serves as the polarization modulation element.

As detailed in Sec. 2.3.1, the specific SLM used (Holoeye LC2012) is of the twisted nematic liquid crystal (TNLC) type. In this device, liquid crystal molecules inherently form a helical twist when no voltage is applied. When a uniform linearly polarized wavefront, with its polarization parallel to the input substrate's director, passes through the SLM at zero voltage, the polarization state adiabatically follows the molecular twist, resulting in a rotation of 90°. This leads to a uniformly 90° rotated polarization across the entire beam. However, applying a voltage pattern across the device's pixelated electrodes disrupts this twist non-uniformly. This disruption, further detailed in Sec. 2.3.1, leads to a spatially varying polarization distribution across the wavefront. The resulting wavefront, now possesses a spatially inhomogeneous polarization distribution as it propagates over a predefined distance towards the detector. Finally, a scientific CMOS (sCMOS) camera (FLIR ORX-10G-71S7M-C) with 3208×2200 pixel resolution captures the resulting output intensity distribution.

The camera records the total intensity distribution I_0 at the output plane. To determine the polarization state at this plane, an analyzer (which is a linear polarizer) is oriented and placed immediately before the camera. Measuring the intensity distribution I transmitted through this analyzer allows for the determination of the local polarization state. For improved statistical robustness, measurements were repeated for several distinct analyzer orientations. Specifically, using Malus's law, we can calculate the local polarization angle θ relative to the known analyzer axis from the intensities measured with the analyzer (I) and without it (I_0):

$$I = I_0 \cos^2 \theta \tag{34}$$

Rearranging this equation gives $\theta = \arccos(\sqrt{I/I_0})$. By performing this calculation pixel-wise for multiple analyzer angles, the spatially varying polarization state across the entire wavefront can be accurately determined. A key advantage of employing polarization modulation is its compatibility with traditional wavefront shaping methods like phase or amplitude modulation.

Indeed, Section 3.6 demonstrates an application that benefits significantly from combining both phase and polarization modulation. In that setup, a liquid crystal on silicon (LCoS) SLM (Holoeye PLUTO) is utilized to implement the required phase modulation. As described in Sec. 2.3.2, applying a voltage pattern to the LCoS SLM modulates its birefringence, thereby enabling precise control over the phase profile of the propagating wavefront [74]. The relative positioning of the phase (LCoS) and polarization (TNLC) SLMs can be flexible, provided that any

diffraction effects occurring over the propagation distances between them are properly accounted for in the design or analysis. In the specific setup used for the combined modulation experiments, the phase SLM first modulates the wavefront's phase profile, and the polarization SLM subsequently modulates the polarization distribution of this phase-shaped wavefront. Care is taken during alignment to ensure that the beam incident on the phase SLM is polarized along its intended operational axis (typically the slow axis for achieving phase-only modulation). Finally, the fully modulated wavefront propagates to the sCMOS camera for measurement, similarly to the polarization-only configuration.

3.2 DIFFRACTION OF POLARIZATION MODULATED WAVEFRONT

The diffractive effects of light have been recognized for centuries [110]. Thomas Young's seminal double-slit experiment in 1801 provided compelling early experimental evidence for the wave nature of light, demonstrating observed interference patterns which are a direct manifestation of diffraction. In general, diffraction occurs whenever a coherent light beam encounters an obstruction or aperture, which can be implemented, for instance, as an amplitude mask or a phase mask [111]. The study of diffraction resulting specifically from polarization masks, however, remains less explored compared to amplitude or phase effects. Conceptually, a polarization mask can be understood as creating two complementary grayscale amplitude masks, with one acting on each of the two orthogonal polarization components of the incident light. These orthogonal wavefront components then propagate independently and interact separately with any subsequent optical elements. To illustrate this principle and highlight potential advantages of utilizing polarization masks, this section employs an analog of Young's double-slit experiment.

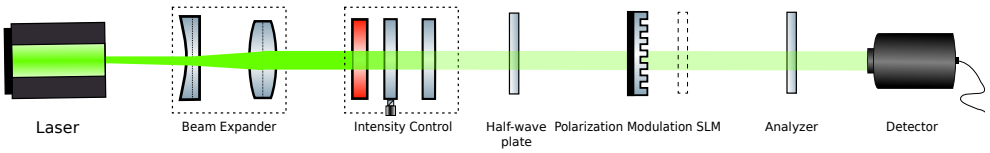


Figure 1: **Schematic of the experimental setup for the Young's double-slit analog experiment.** A Gaussian laser beam is expanded using a lens pair. A TNLC-SLM imparts a polarization pattern analogous to double slits. An analyzer can be placed either near the SLM (dashed representation) or near the camera (solid representation) before the diffracted wavefront is detected. The diagram illustrates the independent propagation of orthogonal polarization components and their interaction with the analyzer.

In our Young's double-slit analog, the TNLC-SLM is used to project a polarization mask corresponding pattern featuring two slit-like regions. This mask imparts a 0.5π ($90°$) polarization rotation specifically in the areas equivalent to the slits (illustrated in Fig. 2a). The light passing through these spatially separated regions, now possessing orthogonal polarization states relative to the background,

propagates until reaching an analyzer, which collapses both states onto a single polarization axis for interference. Subsequently, the sCMOS camera captures the resulting intensity distribution (Fig. 2b). This observed pattern strongly resembles the characteristic interference pattern obtained from a traditional Young's double-slit experiment using an amplitude mask.

To experimentally verify the equivalence between the polarization mask approach and a conventional amplitude mask, we conducted a control experiment by placing the analyzer immediately after the SLM (position 1 in Fig. 1). In this particular configuration, the SLM and the analyzer together function directly as an amplitude mask, effectively allowing only light passing through the "slit" regions with polarization aligned to the analyzer's axis to propagate further and diffract. The resulting intensity pattern measured by the camera is shown in Fig. 2d. A direct comparison between Fig. 2b (analyzer placed near the camera) and Fig. 2d (analyzer placed near the SLM) confirms that both scenarios produce nearly identical diffraction patterns. This result clearly demonstrates the principle that spatially varying polarization, followed by interaction with a polarizer, can effectively replicate the diffractive effects of amplitude modulation, even when the amplitude is homogeneous across the entire wavefront.

This experiment also served as an opportunity to validate our numerical simulation method, which was introduced conceptually in Sec. 2.5.3 (and is based on the Angular Spectrum Method detailed in Sec. 2.5.2), specifically designed for simulating vectorial light propagation. The corresponding simulation results for the two analyzer positions (near camera and near SLM) are presented in Fig. 2c and 2e, respectively. As can be seen, the simulation results show good agreement with the experimental measurements for both configurations.

While the final intensity pattern observed after using a polarization mask and an analyzer mirrors the result from a traditional amplitude mask, the intermediate independent propagation of the orthogonal polarization components offers unique flexibility. Fig. 3 explicitly illustrates this independence. The simulated (Fig. 3b, c) and experimentally measured (Fig. 3d, e) intensity patterns corresponding to each orthogonal component (which are selected by placing the analyzer appropriately relative to the background polarization) confirm their independent propagation characteristics.

An advantage offered by polarization modulation is the ability to use intermediate polarization rotation angles, not restricted to 0 or 0.5π. This capability allows for the creation of grayscale-like amplitude masks upon analysis, contrasting with the inherently binary nature of traditional physical slits or simple amplitude masks. Fig. 4 demonstrates this concept by employing a polarization mask with two slit-equivalents that impart different rotation angles: 0.25π (45°) and 0.5π (90°), respectively (Fig. 4a). The resulting diffraction pattern captured experimentally (Fig. 4b) clearly shows an interference pattern featuring an asymmetric intensity envelope: specifically, the side corresponding to the 0.5π rotation slit appears brighter than the side corresponding to the 0.25π rotation. The corresponding simulation result (Fig. 4c) accurately matches this experimental measurement.

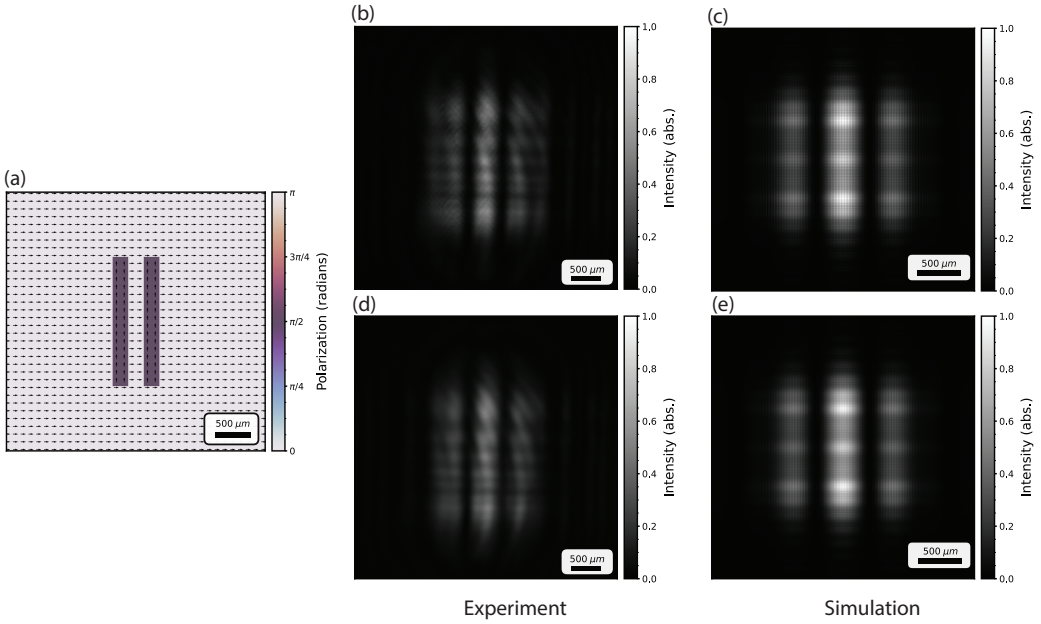


Figure 2: **Young's double-slit experiment with 0.5π polarization rotation at slit-equivalent positions.** (a) The polarization mask applied to the SLM, showing 0.5π rotation in the two slit regions. (b) Experimental intensity distribution with the analyzer near the camera. (d) Experimental intensity distribution with the analyzer near the SLM. (c) and (e) Corresponding simulation results for analyzer near camera and near SLM, respectively. The similarity between experimental results (b, d) and their agreement with simulations (c, e) demonstrate the equivalence principle and validate the numerical method.

Collectively, these results clearly demonstrate the potential to exploit diffraction effects arising from polarization modulation as a means for creating desired target wavefronts. However, it is important to acknowledge that the TNLC-SLM used can potentially introduce coupled phase modulation alongside the intended polarization rotation (as described by Eq. 16). Although we operate the SLM in a manufacturer-specified regime designed to minimize such phase effects and utilize a simplified rotation matrix model in our simulations, it remains crucial to confirm that polarization modulation, rather than unintended phase effects (which could include geometric phase contributions), is the dominant mechanism responsible for the observed diffraction patterns. It is known that polarization rotation during propagation can inherently introduce a geometric phase, often referred to as the Berry phase. Therefore, to verify that the observed diffraction patterns were not primarily caused by such phase effects, we conducted a control experiment where the sCMOS camera was replaced with a Shack-Hartmann wavefront sensor (SHWS), described in Sec. 2.4. This allowed us to directly measure the phase profile of the wavefront at the detection plane.

In our control experiment, we applied a sawtooth polarization rotation pattern (varying linearly from zero to the maximum achievable rotation and then resetting periodically) to the SLM (shown in Fig. 5a) and measured the resulting phase distribution using the SHWS (Fig. 5b). The measurement results indicate a total added phase variation (likely a combination of device-specific phase coupling and

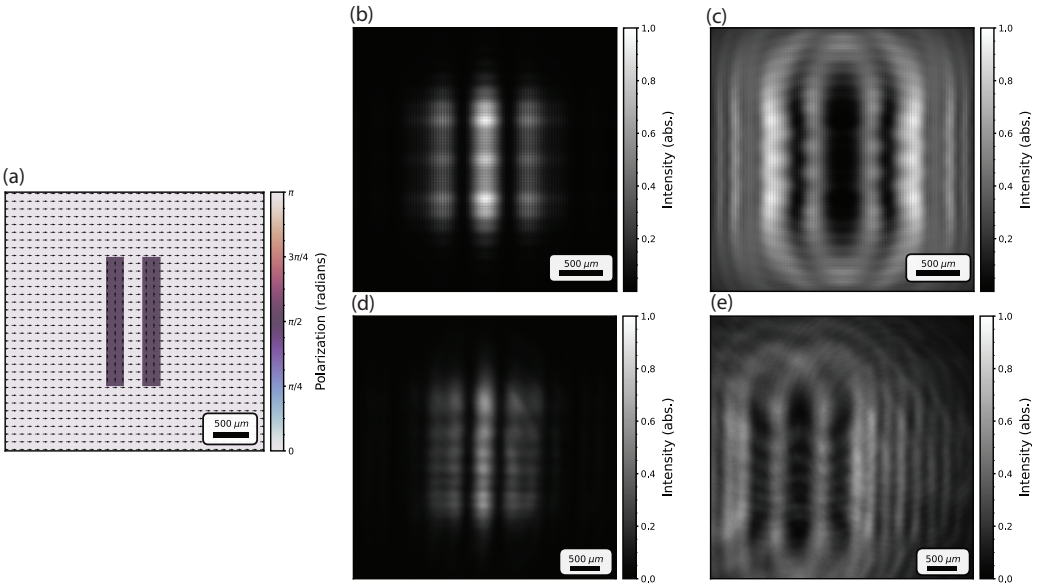


Figure 3: **Complementarity of the two orthogonal polarization states.** (a) Young's double-slit analog polarization pattern. (b) Simulated intensity distribution for the component polarized parallel to the analyzer axis after propagation. (c) Simulated intensity for the orthogonal component. (d) Experimental result corresponding to (b). (e) Experimental result corresponding to (c). The good match between simulations and experiments confirms the independent propagation and complementarity of the orthogonal states.

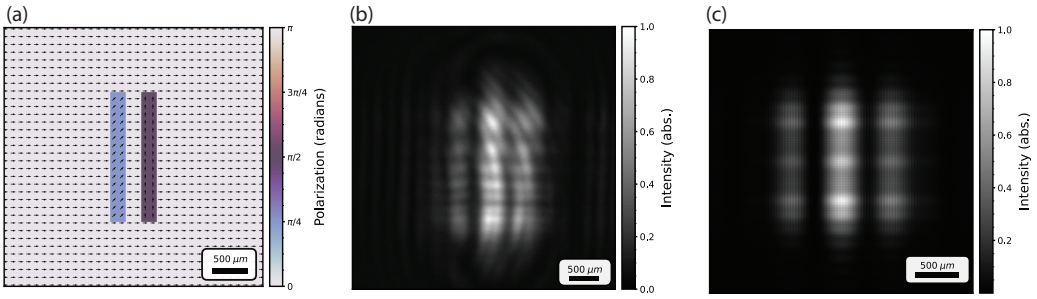


Figure 4: **Young's double-slit experiment with unequal polarization rotation at slit-equivalent positions.** (a) Polarization mask with 0.5π rotation in one slit region and 0.25π in the other. (b) Experimentally measured intensity distribution, showing an asymmetric diffraction pattern envelope. (c) Corresponding simulation result, matching the experiment.

any induced geometric phase) of approximately 0.25π radians across the full range of the applied polarization pattern. To assess whether a phase variation of this magnitude could solely explain the observed diffraction patterns in the YDSE analog experiments, we simulated the diffraction pattern that would result from the YDSE mask structure if it were treated purely as a phase mask, scaled to this limited 0.25π phase modulation range (Fig. 5c). The resulting simulated diffraction pattern (Fig. 5d) exhibits only minimal diffraction features compared to the pattern obtained when simulating the full polarization modulation effect (which, if interpreted purely as phase modulation, would correspond to a much larger ef-

fective phase shift, as shown for comparison in Fig. 5e). This comparison strongly demonstrates that the measured parasitic phase variation is insufficient on its own to produce the strong diffraction effects observed experimentally, thus confirming that polarization modulation is indeed the dominant underlying mechanism.

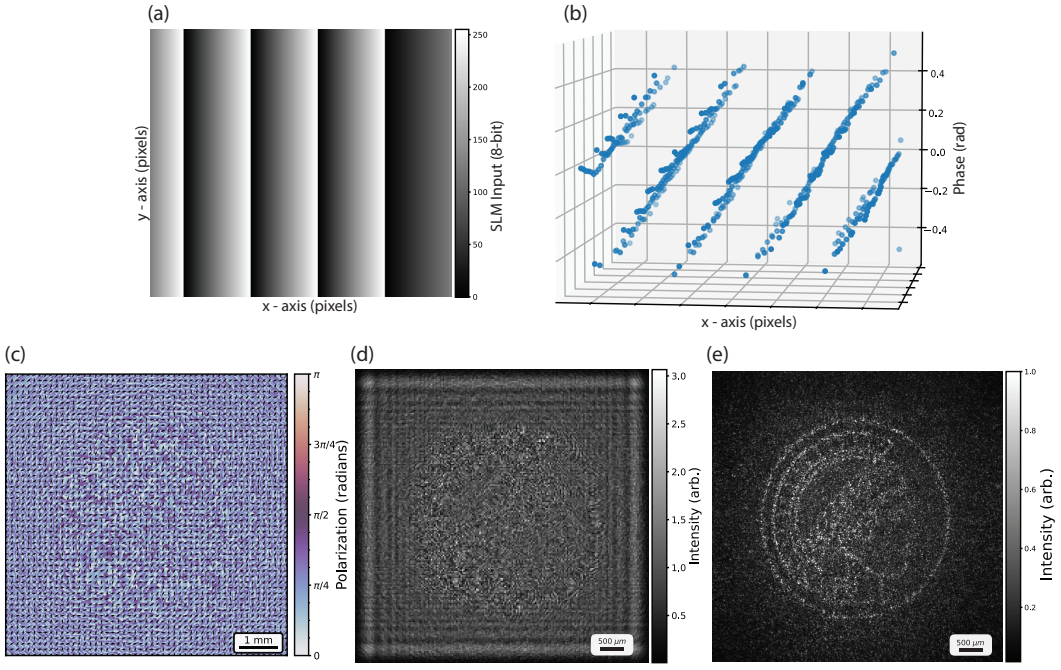


Figure 5: **Control experiment for the geometric/coupled phase contribution.** (a) Input sawtooth polarization rotation pattern applied to the TNLC-SLM. (b) Phase distribution measured by SHWS at the detection plane, showing a maximum variation of approx. 0.25π radians. (c) Scaled phase mask representing this measured phase variation applied to the YDSE structure. (d) Simulated diffraction pattern using only the scaled phase mask (c), showing weak diffraction. (e) Simulated diffraction pattern considering the full polarization modulation effect (from Fig. 2c, shown for comparison), exhibiting strong diffraction. This confirms that the observed diffraction is primarily due to polarization modulation, not unintended phase effects.

The Young's double-slit analog described here serves as a proof-of-concept experiment, demonstrating the possibilities offered by using pixel-wise polarization rotation (effectively acting like a spatially variable half-wave plate array via the SLM) to modulate propagating wavefronts. Having established this principle, we now extend its application to more complex beam shaping tasks. Specifically, the next section describes the generation of Bessel beams using a carefully designed polarization mask.

3.3 POLARIZATION HOLOGRAPHY ENABLED BY DIFFRACTION OF POLARIZATION MODULATED WAVEFRONTS

As detailed in previous background sections (e.g., Sec. 2.6), holography has emerged as a powerful numerical and experimental technique for realizing com-

plex optical wavefronts. Traditionally, these methods rely predominantly on modulating either the phase or the amplitude of the light field. Polarization holography, which utilizes the polarization state as the modulated parameter, however, remains a relatively less explored domain. This section investigates the use of polarization holography, driven by the diffraction from polarization-modulated wavefronts, to generate specific complex wavefronts. The experimental setup employed for these demonstrations, shown schematically in Fig. 6, primarily consists of the expanded laser source, the TNLC-SLM configured for polarization modulation, and the sCMOS camera detector. The individual components were described earlier in Section 3.1.1. A key challenge in polarization holography is devising a suitable numerical framework capable of optimizing the required polarization modulation pattern to achieve a desired target wavefront [112]. As an initial test of the technique's capabilities, we first aim to generate a well-known and practically useful wavefront—a non-diffracting Bessel beam—without resorting to complex optimization algorithms initially.

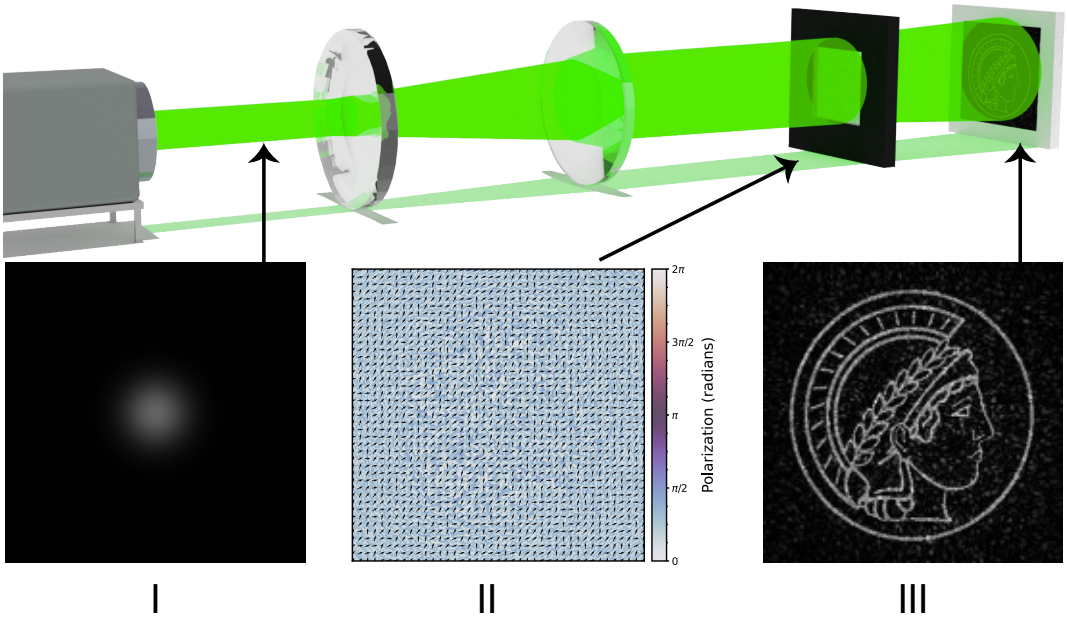


Figure 6: **Experimental setup for polarization modulation in wavefront engineering.** (I) Gaussian laser source, expanded. (II) TNLC-SLM configured for polarization modulation, imparting a spatially varying polarization distribution. (III) Modulated wavefront propagates to the target plane for intensity capture by the sCMOS camera. Figure adapted from [32]

3.3.1 *Generating Pseudo-Bessel Beams using Polarization Modulation*

As described in the background (Sec. 2.2), Bessel beams represent a class of non-diffracting beams [113]. While approaches exist for generating approximations known as pseudo-Bessel beams (detailed in Sec. 2.2), the use of phase masks is a common method. Here, however, we explore the alternative approach of using polarization modulation for this purpose.

We experimentally demonstrate the creation of pseudo-Bessel beams using polarization modulation with the setup illustrated previously in Fig. 6. In this configuration, the SLM modulates the incident beam's polarization distribution according to a specific pattern, and the sCMOS camera captures the resulting intensity distribution after propagation. The polarization mask applied to the SLM consists of a radially symmetric pattern featuring sinusoidal modulation, specifically designed according to the principles outlined in Sec. 2.2.3 to generate a zeroth-order Bessel beam at the output (the mask pattern is shown in Fig. 7a). As shown in Eq. 12, the resultant field after the SLM can be expressed as

$$\mathbf{E}_{out}(r) = \mathbf{R}(\theta(r))\mathbf{E}_{in} = E_0 \begin{pmatrix} \cos\theta(r) \\ \sin\theta(r) \end{pmatrix} \cdot \mathrm{circ}(r/R_{max}) \tag{35}$$

To verify the beam's characteristics, we captured the resulting intensity distribution at various propagation distances downstream from the SLM. Since true Bessel beams are non-diffracting, their intensity profile should ideally remain relatively constant during propagation. Figures 7b, c, e, and f present the experimental measurements recorded at different distances, confirming the strong visual similarity of the intensity distributions over the tested range.

To quantify this observation more rigorously, the captured images were processed computationally using Python along with standard image processing and numerical libraries (e.g., OpenCV, NumPy, Matplotlib). For each grayscale image captured at a different distance, the center of the primary bright spot was identified using thresholding followed by contour analysis (specifically, calculating the centroid of the main contour). Subsequently, the vertical intensity profile passing directly through this identified center was extracted for each image. These individual intensity profiles were then aligned relative to their respective centroids and plotted together for direct comparison (Fig. 7d). As expected for a Bessel beam, the profiles clearly exhibit the characteristic pattern: a central bright spot surrounded by concentric rings of diminishing intensity. Crucially, the similarity of these profiles across the different propagation distances confirms the beam's non-diffracting characteristic within the tested range, thereby validating the successful generation of a pseudo-Bessel beam via polarization modulation.

While this experiment successfully demonstrates the generation of a known, structured beam by extending principles analogous to amplitude mask diffraction, the broader goal is to create arbitrary wavefronts using polarization modulation alone. Achieving this requires more sophisticated methods for determining the necessary polarization mask. Therefore, the next section describes the optimization techniques developed and employed for this purpose.

3.4 MODIFIED GERCHBERG-SAXTON ALGORITHM FOR OPTIMIZING POLARIZATION DISTRIBUTION

The Gerchberg-Saxton (GS) algorithm, detailed in Sec. 2.6.1, is a well-established iterative phase retrieval algorithm based on the principle of alternating projections

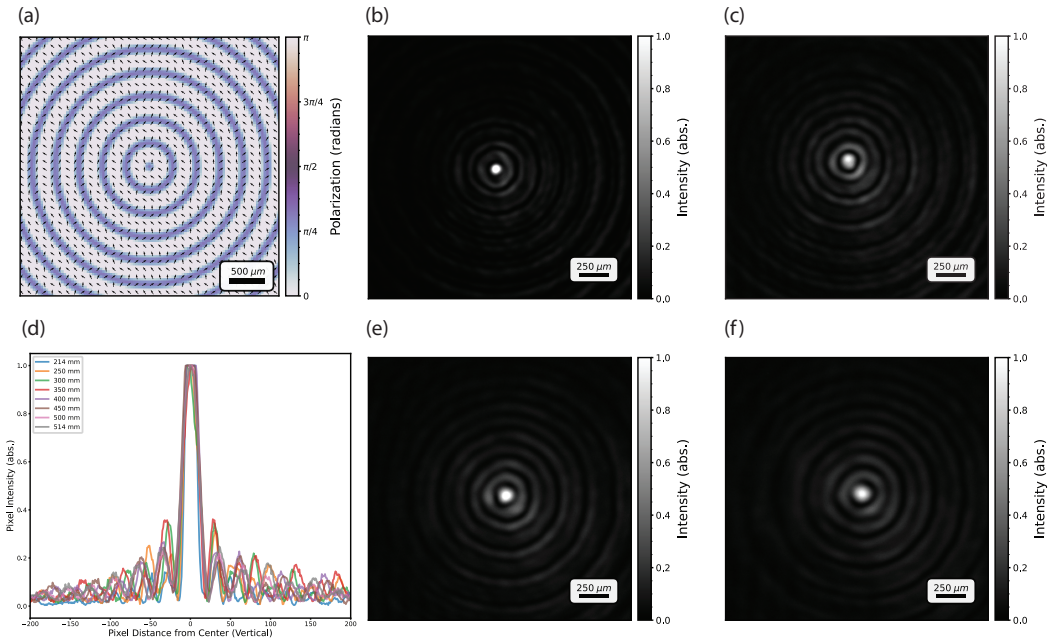


Figure 7: **Generation of pseudo-Bessel beams using polarization modulation.** (a) Radially symmetric polarization mask applied to the SLM. (b, c, e, f) Experimental intensity distributions at propagation distances 214 mm, 300 mm, 400 mm, and 500 mm, respectively. (d) Comparison of vertical intensity profiles through the beam center at the different distances. The consistent ringed structure and profiles confirm the generation of a non-diffracting pseudo-Bessel beam.

between spatial and Fourier domains. It is typically employed for optimizing phase masks designed to produce specific target amplitude distributions. To apply this algorithm to our context, we need to modify its standard implementation so that it yields a polarization distribution that, when imparted to an input Gaussian beam via the SLM, produces the desired target amplitude distribution after free-space propagation. The resulting modified algorithm, shown schematically in Fig. 8a, closely follows the logic of the original GS algorithm but incorporates a crucial final transformation step. The specific steps involved are as follows:

1. **Source Field Initialization:** First, initialize a complex field representing the input Gaussian beam at the plane of the SLM, ensuring it matches the experimental characteristics (such as beam waist and wavelength). Assume a uniform phase front and a defined input polarization state.
2. **Random Phase Initialization:** Create an initial random phase mask $\phi^{(0)}(x, y)$ with the same spatial dimensions as the SLM display area. This serves as the starting point for the iterative optimization.
3. **Hologram Field Construction (Iteration k):** Combine the known source amplitude $A_s(x, y)$ with the current phase estimate $\phi^{(k)}(x, y)$ to construct the complex field at the SLM plane for the current iteration k: $E_h^{(k)} = A_s \exp(i\phi^{(k)})$.
4. **Forward Propagation:** Numerically propagate the field $E_h^{(k)}$ from the SLM plane to the target plane using the Angular Spectrum Method (detailed in

Sec. 2.5.2). This yields the propagated field $E_t'^{(k)} = \mathcal{P}\{E_h^{(k)}\}$, where $\mathcal{P}$ denotes the propagation operator.

5. **Target Amplitude Constraint:** At the target plane, replace the amplitude of the propagated field $|E_t'^{(k)}|$ with the desired target amplitude $A_t(x', y')$, while importantly retaining the calculated phase of the propagated field: $E_t''^{(k)} = A_t \exp(i \arg(E_t'^{(k)}))$.

6. **Inverse Propagation:** Propagate the constrained field $E_t''^{(k)}$ back from the target plane to the SLM plane using the inverse propagator $\mathcal{P}^{-1}$: $E_h'^{(k)} = \mathcal{P}^{-1}\{E_t''^{(k)}\}$.

7. **Phase Update:** Extract the phase of this back-propagated field to obtain the updated phase estimate for the next iteration: $\phi^{(k+1)} = \arg(E_h'^{(k)})$. Note that the amplitude constraint at the hologram plane (i.e., using the known source amplitude A_s) is implicitly reapplied in step 3 of the subsequent iteration. Repeat steps 3 through 7 for a preset number of iterations or until convergence criteria are met.

8. **Phase-to-Polarization Transformation:** Finally, transform the optimized phase mask obtained after the iterations, ϕ_{opt}, into a corresponding polarization rotation mask $\theta(x, y)$. This represents the most significant modification to the standard GS algorithm. A simple linear mapping $\theta(x, y) = c \cdot \phi_{opt}(x, y)$ is used heuristically, where the constant c scales the phase range (typically 0 to 2π) to the desired polarization rotation range achievable by the SLM (e.g., 0 to $\pi/2$ for emulating a half-wave plate array). While phase directly affects the imaginary part of the field's exponent and polarization rotation affects the vector components, their influence on the resulting *diffracted intensity pattern* exhibits qualitative similarities. This similarity allows the heuristic transformation to yield useful, albeit not rigorously optimal, results.

We applied this modified GS process to optimize a polarization modulation mask intended to generate a target amplitude distribution resembling the Minerva logo of the Max Planck Society (Fig. 8b), assuming a propagation distance of 500 mm. The resulting optimized polarization mask calculated by the algorithm is shown in Fig. 8c. Subsequently, we projected this mask using the TNLC-SLM within the experimental setup (Fig. 6) and captured the resulting intensity distribution after propagation using the sCMOS camera (Fig. 8d). As observed, the experimental result matches the overall shape of the target amplitude distribution reasonably well.

However, upon closer inspection, it is evident that finer details within the logo are poorly represented, and a significant amount of background speckle noise is present in the experimental result. This outcome is likely attributable to the fact that the standard GS algorithm, even with our heuristic phase-to-polarization transformation, is not inherently designed for the direct optimization of polarization holograms. Its core mechanism revolves around phase retrieval, making the final transformation step somewhat indirect and potentially suboptimal for minimizing errors in the polarization-driven diffraction process. Motivated by these limitations, we now explore the application of machine learning techniques to directly

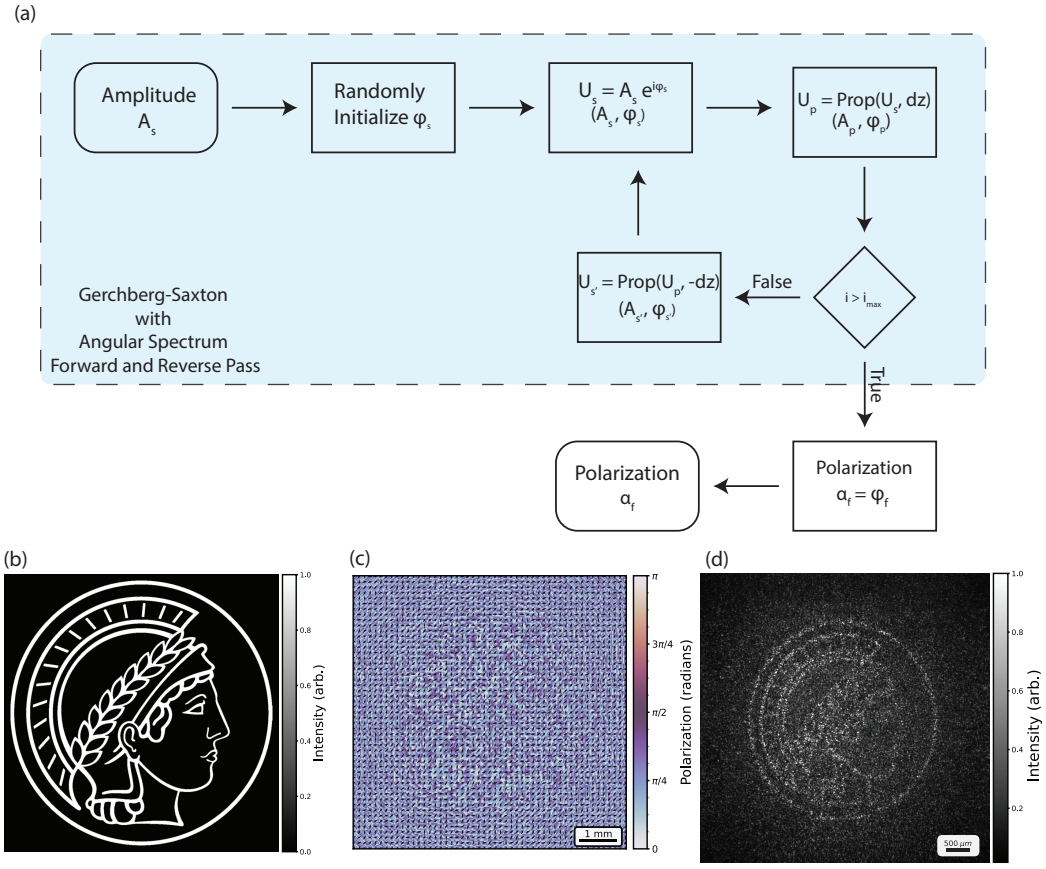


Figure 8: **Optimization of a polarization modulation mask using the Gerchberg-Saxton algorithm.** (a) Flowchart of the modified GS algorithm. (b) Target amplitude distribution (Minerva logo). (c) Optimized polarization modulation mask obtained via modified GS. (d) Experimental intensity result using the mask (c), showing resemblance to the target but with noticeable background noise/speckle image. Figure adapted from [32].

optimize the polarization distribution, aiming to achieve improved performance and fidelity without relying on intermediate heuristic transformations.

3.5 MACHINE LEARNING OPTIMIZATION FOR POLARIZATION MODULATION

Machine learning methodologies, particularly deep learning approaches that utilize differentiable models, have recently demonstrated significant success when applied to challenging optimization problems within physics and optics (as discussed in Sec. 2.7). Building on this trend, this section explores the application of ML techniques specifically for optimizing polarization modulation masks.

The first critical step in developing an ML-based optimization framework is defining a differentiable forward model that accurately reflects the physical experiment. This model must encompass the entire process: an initial Gaussian laser source beam is expanded, its polarization is modulated by an SLM imparting a spatially varying rotation, the modulated beam propagates through free space to

a target plane, and its intensity is finally detected by an sCMOS camera. We represent each component of this process in a computationally differentiable manner:

- **Input Beam:** Modeled as a complex field possessing a Gaussian amplitude profile that matches the experimentally measured expanded beam incident on the SLM. The input polarization state is assumed to be uniform and linear.
- **Propagation:** Free-space propagation is simulated using the Angular Spectrum Method (ASM), described in detail in Sec. 2.5.2. The core ASM propagation formula (repeated here for clarity) is given by:

$$U(x, y, z) = \mathcal{F}^{-1}\{\mathcal{F}\{U(x, y, 0)\}H(k_x, k_y; z)\} \tag{36}$$

  This involves Fourier transforms ($\mathcal{F}, \mathcal{F}^{-1}$) and element-wise multiplication (by the transfer function H). Crucially, all of these operations are inherently differentiable. The extension of ASM to handle polarized light (Sec. 2.5.3) simply involves applying this propagation operator independently to each orthogonal polarization component of the Jones vector, thereby preserving the overall differentiability of the propagation step.
- **SLM (Polarization Modulation):** The TNLC-SLM is operated in a regime intended to primarily induce polarization rotation (as detailed in Sec. 3.1.1). An ideal polarization rotation by a spatially varying angle $\theta(x, y)$ is modeled by multiplying the input Jones vector $\mathbf{E}_{in}$ by the corresponding rotation matrix $\mathbf{R}(\theta(x, y))$ (Eq. 11). The elements of the rotation matrix $\mathbf{R}$, namely $\cos\theta$ and $\sin\theta$, are differentiable functions with respect to the rotation angle θ. The polarization mask $\theta(x, y)$ itself constitutes the set of trainable parameters that the ML algorithm will optimize.
- **Camera Detection:** The sCMOS camera measures the optical intensity. This detection process is modeled computationally as the squared absolute value of the complex electric field amplitude incident on the detector plane, summed over both polarization components:

$$I(x, y) = |E_x(x, y, z)|^2 + |E_y(x, y, z)|^2 \tag{37}$$

  This operation, involving the sum of squares of the real and imaginary parts of the field components, is also fully differentiable with respect to those field components.

Given that all individual components are modeled using differentiable operations, the entire physics-based forward model, tracing the path from the input SLM mask parameters $\theta(x, y)$ to the final predicted intensity $I(x, y)$ on the camera plane, is end-to-end differentiable. A schematic representation of this forward model network is conceptually similar to the one shown in Fig. 9a. This end-to-end differentiability is the key property that allows the use of gradient-based optimization methods, specifically backpropagation, to iteratively adjust the polarization

mask parameters $\theta(x, y)$ by minimizing a suitably defined loss function that compares the predicted intensity P_{int} (the output of the forward model) to the desired target intensity T_{int}.

Choosing an appropriate loss function is critical for the success of the optimization process. While standard loss functions like Mean Squared Error (MSE) are available, custom-designed loss functions often yield significantly better performance for specific physics-based tasks like holographic optimization [114]. As introduced in the background (Sec. 2.7), incorporating metrics that reflect human perceptual quality or specific physical constraints can be highly beneficial. The loss function employed in this work is a composite function that combines several terms to address different aspects of image fidelity and quality:

1. **Intensity Fidelity Loss ($\mathcal{L}_{int}$):** This primary component aims to ensure that the predicted intensity pattern P_{int} closely matches the target pattern T_{int}. It combines the Mean Absolute Error ($\mathcal{L}_{L1}$), which focuses on pixel-wise accuracy, with the Structural Similarity Index Measure (SSIM) loss ($\mathcal{L}_{SSIM}$), which prioritizes the preservation of structural information and is known to correlate well with perceived image quality [26, 115, 116]:

$$\mathcal{L}_{int} = \alpha \cdot \mathcal{L}_{L1}(P_{int}, T_{int}) + (1 - \alpha) \cdot \mathcal{L}_{SSIM}(P_{int}, T_{int}) \tag{38}$$

   Here, α is a hyperparameter (typically set around 0.8) that balances the contribution of the L1 and SSIM terms.

2. **Dark Region Contrast Loss ($\mathcal{L}_{dark}$):** This term specifically penalizes any non-zero predicted intensity values occurring in regions where the target intensity is effectively zero (i.e., below a small threshold T_{dark}). Its purpose is to actively reduce the background noise and speckle artifacts that are often prevalent in results obtained using purely iterative methods like GS. It is calculated as the mean squared intensity within these target-dark regions, denoted by $\mathcal{D}$:

$$\mathcal{L}_{dark} = \frac{1}{|\mathcal{D}|} \sum_{(i,j)\in\mathcal{D}} (P_{int,i,j})^2 \quad \text{where } \mathcal{D} = \{(i, j) \mid T_{int,i,j} < T_{dark}\} \tag{39}$$

   The threshold T_{dark} is typically chosen to be a small value (e.g., 0.1 on a normalized scale).

These two components are then combined into the final contrast-aware loss function $\mathcal{L}_{contrast}$:

$$\mathcal{L}_{contrast} = \mathcal{L}_{int} + w_{dark}\mathcal{L}_{dark} \tag{40}$$

The hyperparameter w_{dark} controls the relative weight assigned to the contrast enhancement term (e.g., a value of $w_{dark} = 2.0$ was found effective for the Minerva logo target).

Employing this tailored loss function $\mathcal{L}_{contrast}$ in conjunction with the differentiable forward model, we trained the polarization mask parameters $\theta(x, y)$ using

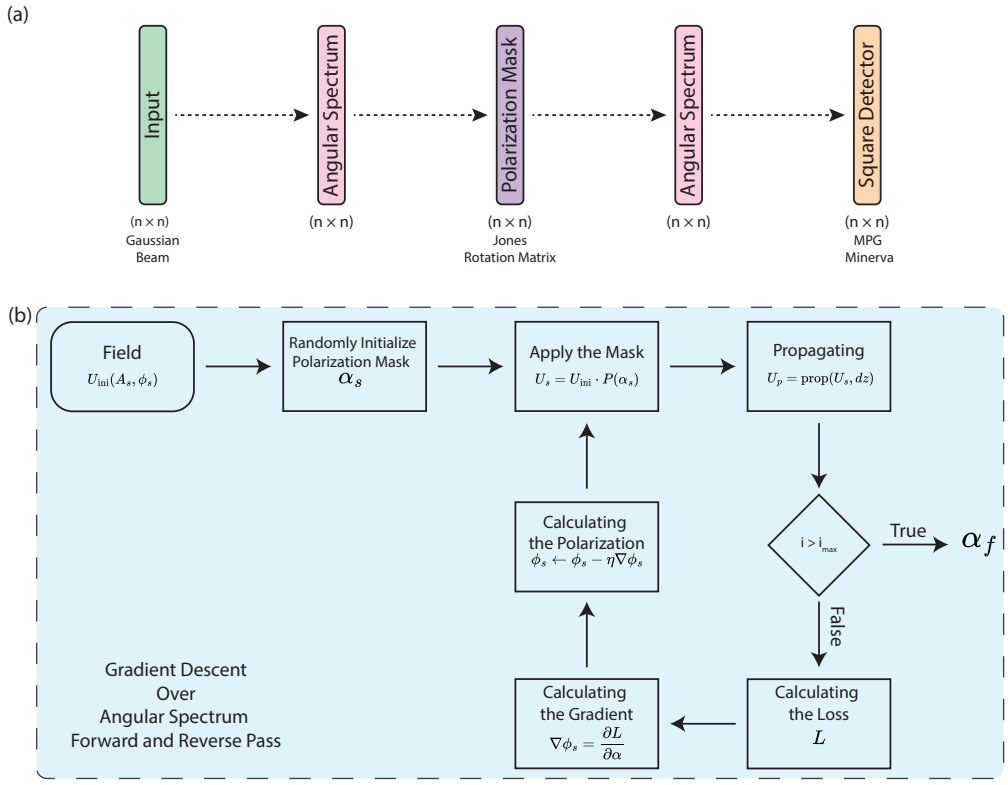


Figure 9: **Optimization of a polarization modulation mask using machine learning.** (a) Schematic of the differentiable physics-based model used in the optimization framework. (b) Flowchart of the training loop: Initialize random polarization mask $\theta(x, y)$, compute predicted intensity P_{int} using the forward model (input beam -> SLM modulation -> propagation -> camera detection), calculate loss $\mathcal{L}_{contrast}$ between P_{int} and target T_{int}, compute gradients via backpropagation, update $\theta(x, y)$ using an optimizer (e.g., Adam), repeat. Figure adapted from [32].

backpropagation and an Adam optimizer (as depicted schematically in Fig. 9b) [117]. This process directly optimized the mask to produce the target amplitude distribution (the Minerva logo, Fig. 10a), thereby bypassing the indirect, heuristic approach required by the modified GS algorithm. The final optimized polarization mask obtained through this ML process is shown in Fig. 10b. This optimized mask was then uploaded to the SLM in the experimental setup, and the resulting output intensity distribution was measured (Fig. 10c). Comparing this experimental result to the one obtained using the modified GS method (Fig. 8d), the ML-optimized result clearly exhibits a closer match to the target intensity distribution and possesses significantly improved background contrast with much less speckle noise. This successful outcome demonstrates the distinct superiority of the direct ML optimization method for this specific task, offering not only better fidelity but also greater flexibility in tailoring the optimization process through the design of the loss function.

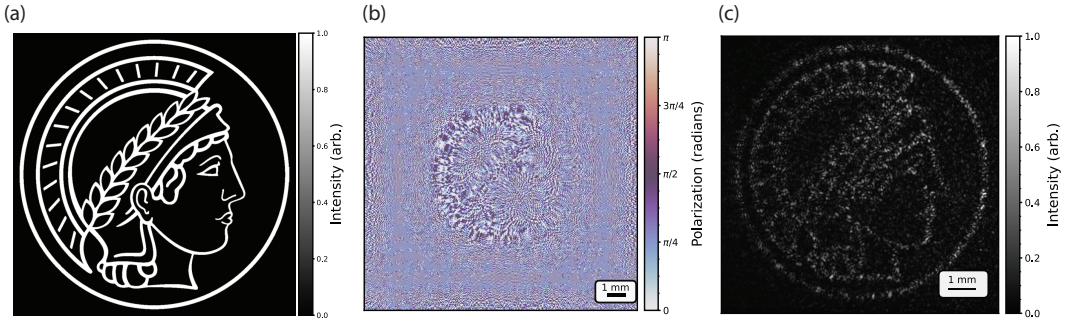


Figure 10: **Experimental results for machine-learned polarization modulation mask.** (a) Target intensity distribution (Minerva logo). (b) Optimized polarization modulation mask obtained using the ML framework. (c) Experimentally recorded intensity distribution using mask (b), showing high fidelity and contrast compared to the GS method (Fig. 8d). Adapted from [32].

3.5.1 *Joint Optimization for Target Amplitude and Polarization*

The previous section successfully demonstrated that employing an ML optimization framework yields results superior to traditional methods when optimizing a polarization mask for a specific target amplitude. This ML approach not only enhances accuracy and contrast but also inherently offers greater flexibility, making it suitable for tackling scenarios that are challenging for conventional algorithms. One such challenging scenario involves the joint optimization of multiple output parameters using only a single control mechanism. Specifically, the ML framework developed here allows for the simultaneous optimization of **both** the amplitude **and** the polarization state at the output plane, utilizing only the input polarization mask applied by the SLM as the trainable parameter set.

To achieve this joint optimization, the framework is extended by defining a more sophisticated compound loss function. This loss function is designed to penalize deviations in both the predicted intensity (or amplitude) P_{amp} and the predicted polarization state P_{pol} relative to their respective desired targets, T_{amp} and T_{pol}. The total loss, $\mathcal{L}_{total}$, is constructed by combining distinct loss components calculated independently for each output channel (amplitude and polarization):

1. **Intensity Loss Component ($\mathcal{L}_{amp}$):** For matching the intensity, the previously defined contrast-aware loss $\mathcal{L}_{contrast}$ (Eq. 40) is used directly. This component effectively handles the intensity matching requirement, with its inherent capability to improve results in low-signal or dark regions: $\mathcal{L}_{amp} = \mathcal{L}_{contrast}(P_{amp}, T_{amp})$.
2. **Polarization Loss Component ($\mathcal{L}_{pol}$):** Matching the polarization state requires high fidelity. Therefore, this component combines two terms. The first term is the Mean Absolute Error (L1 loss) calculated between the predicted (P_{pol}) and target (T_{pol}) polarization values (which could represent, for example, the polarization angle) across all N pixels in the output plane:

$$\mathcal{L}_{\text{pol-L1}} = \frac{1}{N} \sum_i \left| P_{\text{pol},i} - T_{\text{pol},i} \right| \tag{41}$$

The second term introduces an exact match penalty ($\mathcal{L}_{\text{pol-exact}}$) designed to punish deviations $d_i = |P_{\text{pol},i} - T_{\text{pol},i}|$ that exceed a predefined small tolerance threshold T_{exact}. This penalty applies a weighted quadratic cost only when the deviation d_i surpasses this tolerance:

$$\mathcal{L}_{\text{pol-exact}} = \frac{1}{N} \sum_i \begin{cases} (d_i \cdot w_{\text{exact}})^2, & \text{if } d_i > T_{\text{exact}}, \\ 0, & \text{if } d_i \leqslant T_{\text{exact}}. \end{cases} \tag{42}$$

Here, w_{exact} serves as a weighting factor for this penalty term. The total polarization loss is then the sum of these two parts:

$$\mathcal{L}_{\text{pol}} = \mathcal{L}_{\text{pol-L1}} + \mathcal{L}_{\text{pol-exact}} \tag{43}$$

These individual loss components for amplitude and polarization are combined into the final total loss function using a hyperparameter $\alpha \in [0, 1]$. This hyperparameter acts as a weighting factor to balance the relative importance assigned to achieving the target intensity versus achieving the target polarization state:

$$\mathcal{L}_{\text{total}} = \alpha \cdot \mathcal{L}_{\text{amp}} + (1 - \alpha) \cdot \mathcal{L}_{\text{pol}} \tag{44}$$

By adjusting the value of α, one can effectively tune the optimization process to prioritize matching one channel more closely than the other if needed.

To experimentally test this joint optimization capability, we first chose a relatively simple target configuration: a rounded square shape for the intensity distribution (Fig. 11b) combined with an azimuthally varying polarization pattern (similar to the pattern generated by an axicon, exhibiting uniform rotation around the center) spanning the entire frame (Fig. 11c). Running the ML optimization framework with the joint loss function yielded the optimized polarization mask shown in Fig. 11a. Experimentally implementing this mask produced the intensity distribution captured in Fig. 11d, which closely matches the target rounded square shape. The corresponding output polarization state was measured using the analyzer-based method described in Sec. 3.1.1, yielding the distribution shown in Fig. 11e. This measured polarization pattern also matches the target azimuthal variation reasonably well, although some degree of non-uniformity is visually apparent. This non-uniformity likely arises from the inherent coupling between amplitude and polarization control when attempting to optimize both simultaneously using only a single modulation mask.

To further demonstrate the flexibility of the joint optimization framework, we then aimed for a more complex polarization target while maintaining the same

intensity target. The target intensity remained a rounded square (Fig. 12b), but the target polarization pattern was modified to exhibit rotation confined primarily to the perimeter of the square, completing two full sweeps (e.g., from 0 to $\pi/2$ radians, repeated twice) around the loop (Fig. 12c). The ML optimization produced the mask shown in Fig. 12a. The corresponding experimental results for intensity (Fig. 12d) and polarization (Fig. 12e) again show good agreement with their respective complex targets. This result confirms that a single polarization modulation mask, when optimized using the ML framework and an appropriate joint loss function, can indeed simultaneously control both the amplitude and the polarization state at the output plane, even for non-trivial target patterns.

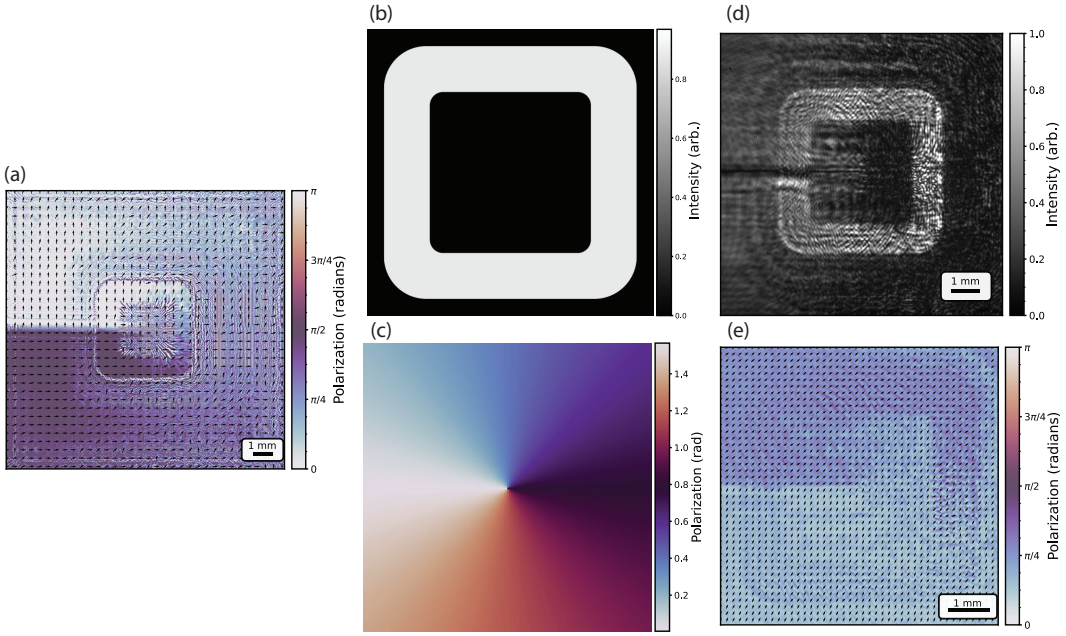


Figure 11: **Results of joint optimization for simple amplitude and polarization targets.** (a) Optimized polarization modulation mask. (b) Target intensity (rounded square). (c) Target polarization (azimuthal variation). (d) Measured intensity result. (e) Measured polarization result. Both results show good agreement with the targets. Adapted from [32].

It is important to note, however, that this joint control method using only a single polarization mask does face limitations, particularly when targeting complex wavefronts with high spatial frequency content. This is because the output amplitude and polarization patterns are inherently coupled through the diffraction process originating from that single mask. This coupling can manifest itself as deviations from the intended targets, such as the faint imprint of the intensity shape (the rounded square) that is visible on the measured polarization distribution in Fig. 11e. Despite these inherent constraints, the technique clearly demonstrates significant potential. Combining this approach with independent phase modulation, for instance, could potentially decouple these effects to some extent and unlock even further capabilities in wavefront engineering.

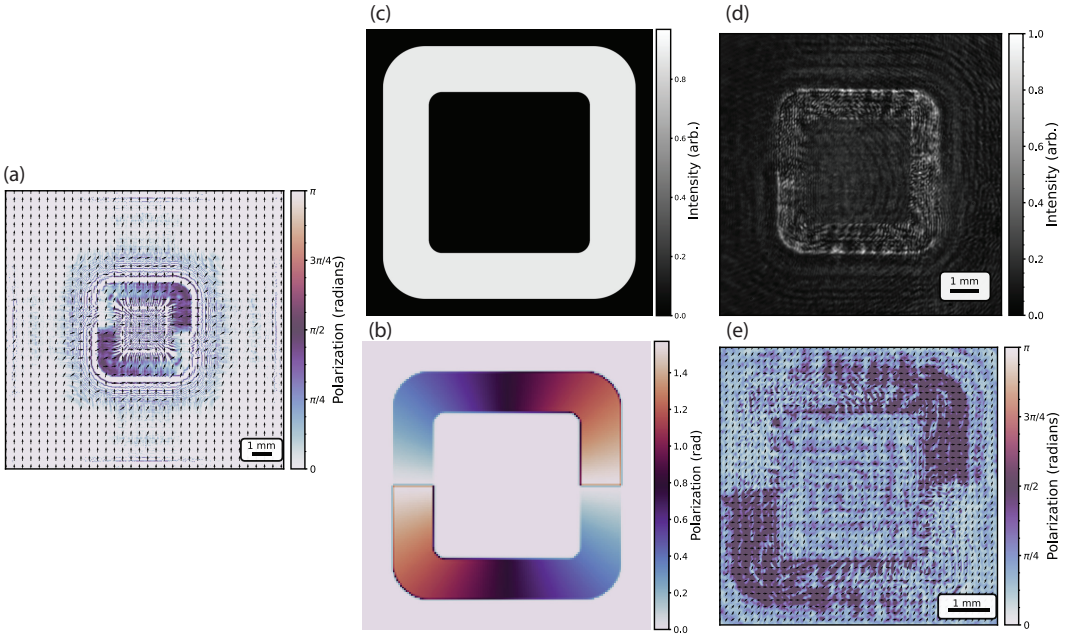


Figure 12: **Results of joint optimization for complex polarization targets.** (a) Optimized polarization modulation mask. (b) Target intensity (rounded square). (c) Target polarization (two full sweeps around the perimeter). (d) Measured intensity result. (e) Measured polarization result. Good agreement achieved even for complex targets. Adapted from [32].

3.6 COMBINING POLARIZATION MODULATION WITH PHASE

While polarization modulation alone, as demonstrated in the preceding sections, offers a novel and viable framework for wavefront shaping, combining it strategically with conventional phase modulation can greatly enhance the overall capabilities, potentially enabling more complex and versatile control over the final output optical field. One promising potential application area where such combined modulation could be advantageous is in the development of high-speed, non-mechanical optical point scanning systems. Such systems are crucial for applications like confocal microscopy and laser materials processing, offering a potentially faster alternative to established methods based on slower mechanical components (like galvanometer mirrors) or other complex electro-optical systems (such as MEMS mirrors or acousto-optic deflectors).

To explore this possibility, this section demonstrates a proof-of-concept experimental setup that utilizes static phase and polarization masks to achieve rapid selection or modulation of predefined points in space. The setup, illustrated schematically in Fig. 13, incorporates the essential components: the laser source, the TNLC-SLM (primarily acting as a spatially variable polarization rotator), and an additional LCoS SLM dedicated to phase modulation. In this configuration, the phase mask displayed on the LCoS SLM defines the spatial locations of the target points in the output plane, while the polarization mask displayed on the TNLC SLM applies a distinct, spatially varying polarization rotation to the portion of the light beam directed towards each specific point. An analyzer placed before the camera

then serves to select a specific polarization state for detection. Consequently, by dynamically rotating this analyzer (or, equivalently, by electronically controlling the global input polarization state before the masks), different points defined by the phase mask can be selectively illuminated or extinguished on the detector plane based on the polarization state imparted to them by the TNLC mask.

In this specific demonstration, we employed two different static phase masks programmed onto the LCoS SLM: one designed to generate five collinear points, and another generating five arbitrarily positioned points in the output plane (the resulting combined intensity patterns are shown in Fig. 14a and d, respectively). Concurrently, a relatively simple sinusoidal polarization pattern was applied using the TNLC-SLM (Fig. 13b). The "scanning" or selection process was achieved in this proof-of-concept by manually rotating the analyzer through a full 0 to 2π range. As expected, the intensity of each individual point varied sinusoidally as the analyzer rotated (as shown by the traces in Fig. 14b and e). However, crucially, due to the spatially varying polarization imparted by the TNLC-SLM pattern, the phase of this sinusoidal intensity modulation differed for each of the five points. As a direct consequence, each point reached its maximum intensity at a distinct analyzer angle (these angles of maximum intensity are summarized in Fig. 14c and f for the collinear and arbitrary cases, respectively). It is important to note that in a practical high-speed implementation, the manual analyzer rotation would be replaced by a fast electro-optic modulator (such as a Pockels cell) capable of electronically controlling the global polarization state without requiring any moving parts.

Experimental measurements, obtained by recording the average intensity within small regions of interest centered around each target point while systematically changing the analyzer angle, confirmed this predicted differential intensity modulation behavior. While a simple sinusoidal polarization mask was sufficient for this demonstration, utilizing more complex polarization patterns could potentially achieve more intricate intensity variations across the points, adding further flexibility to the system.

This observed differential modulation effectively enables the sequential addressing or intensity modulation of a set of predefined points simply by controlling the projected polarization state. Therefore, this demonstration highlights the feasibility of realizing non-mechanical scanning systems by combining static phase modulation (which defines the geometric pattern of points) with dynamic polarization modulation (which provides intensity control or selection). Furthermore, the ability to completely redesign the scan pattern simply by loading a different phase mask onto the LCoS SLM adds significant versatility. This proof-of-concept system could potentially be extended through the use of more complex phase and polarization masks, coupled with dynamic electronic control, to enable advanced applications in imaging and beam delivery.

3.7 CONCLUSION

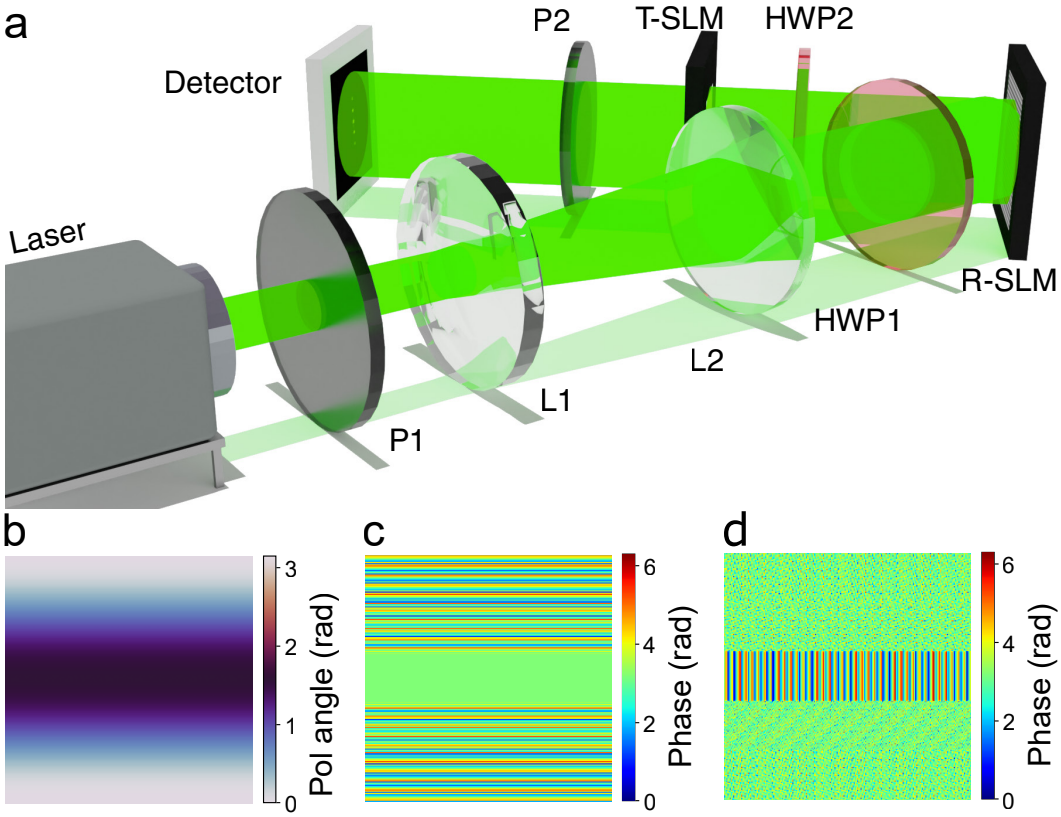


Figure 13: **Experimental setup for combining phase and polarization modulation.** (a) Schematic: Laser, Polarizer (P1), Beam Expander, Half-Wave Plates (HWP1, HWP2 for alignment), Phase SLM (R-SLM, LCoS), Polarization SLM (T-SLM, TNLC), Analyzer (P2), Camera. (b) Sinusoidal polarization mask on T-SLM. (c) Phase mask for collinear points on R-SLM. (d) Phase mask for arbitrary points on R-SLM.

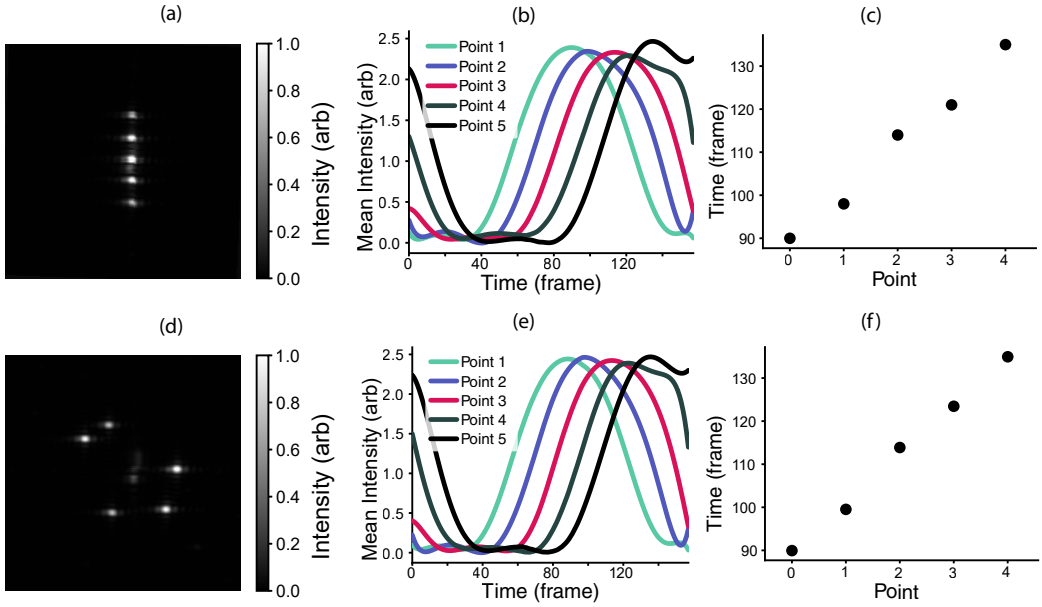


Figure 14: **Results of combining polarization and phase modulation for point scanning.** (a) Measured intensity of all five collinear points combined. (d) Measured intensity of all five arbitrary points combined. (b, e) Evolution of mean intensity for each of the five points (different colors) as analyzer angle rotates for collinear and arbitrary cases, respectively. (c, f) Analyzer angle at which each point reaches maximum intensity, demonstrating sequential addressing for collinear and arbitrary cases.

Part II

SCALABLE LOW POWER OPTOELECTRONIC NEURAL NETWORKS

The following chapters are primarily based on the work presented in [118, 119]. The project was conceptualized in collaboration with Dr. Alexander Song, Prof. Bernhard Schölkopf, and Prof. Peer Fischer. The development of optical and optomechanical hardware, as well as the experimental data collection and analysis, were carried out in collaboration with Dr. Alexander Song. Dr. Song was responsible for training the neural networks. I designed, simulated, and fabricated the electronic components used in this project, with assistance from Dr. Rahul Goyal. I would like to thank Lennart Schlieder, Dr. Valentin Volchkov, Steffen Epple and Dr. Zili Yu for their valuable discussions and insights.

4

INTRODUCTION

Artificial intelligence (AI) models have become integral to modern society, powering a wide range of applications, from image recognition to natural language processing. The rapid proliferation of these technologies has been driven by developing increasingly large and sophisticated deep learning models. However, this advance has incurred a significant cost: a sharp increase in demand for computational resources and, consequently, the energy required for their development and deployment.

The energy demands of AI can be broadly grouped into two categories: those due to training and those involving inference. Training requires optimizing a model's parameters using very large datasets, which requires substantial computational resources. For instance, training a state-of-the-art model such as GPT-4 consumes considerable amounts of energy over extended periods, typically running on large clusters of Graphics Processing Units (GPUs), estimated to have cost about 10 million USD in training energy costs [120]. Although energy-intensive, this is generally a one-time cost. In contrast, inference—the process of using a trained model to generate predictions—is performed continuously during model deployment. Every use of the model requires inference, leading to ongoing significant energy consumption. This challenge is especially pronounced in energy-constrained edge-computing scenarios, such as autonomous vehicles, where real-time inference is essential, but power availability is inherently limited.

A fundamental contributor to the energy inefficiency of conventional computing hardware is its architectural design. Traditional computers are based on the von Neumann architecture, where memory and processing units are physically separate and connected by a data bus. This separation necessitates constant data movement between memory and the processor, creating a "memory bottleneck" that limits computational speed and incurs significant energy costs from frequent read and write operations. Neural networks, inspired by the parallel processing capabilities of biological nervous systems—where interconnected neurons store and process information—are inherently ill-suited to the sequential nature of von Neumann machines. In particular, the core operations of neural networks, primarily matrix-vector multiplications (MVM), are severely impacted by this data transfer bottleneck. The need to read parameters and write computation results to memory further compounds the problem, leading to a significant energy inefficiency.

In-memory computing, or Compute-in-Memory (CIM), has emerged as a promising paradigm to address this architectural limitation. CIM aims to perform computations directly within the memory array, thereby minimizing data movement and reducing the associated energy overhead. This approach is particularly advantageous for AI applications, where processing vast datasets is common. By reducing data transfer, CIM can achieve significant energy savings. Developing

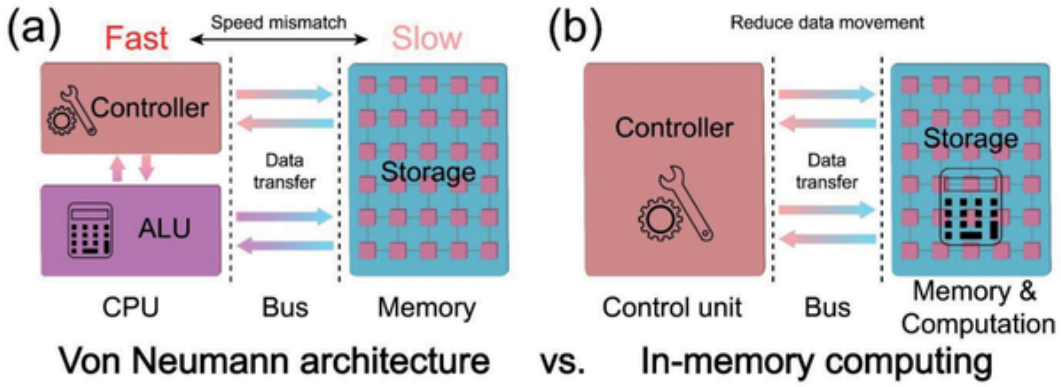


Figure 15: **Comparison between von Neumann architecture and in-memory computing.** (a) illustrates the conventional von Neumann architecture, where memory and processing units are physically separate. In this design, neural networks are implemented using matrix-vector multiplication (MVM), with individual elements stored at different memory locations while computations occur in the Arithmetic and Logic Unit (ALU). A controller manages data movement and issues instructions. However, the data transfer between the ALU and memory via the bus creates a bottleneck, leading to significant energy consumption due to frequent read-write operations. In contrast, (b) shows in-memory computing, where computations are performed directly within the memory array. The controller only needs to issue instructions, significantly reducing data transfer and lowering overall energy consumption. Figure reproduced in its entirety from [121].

hardware solutions that embody CIM principles, even if initially optimized for inference tasks, is therefore critical for improving the energy efficiency of AI computing.

To tackle AI's energy efficiency challenges, new classes of hardware are being developed. Custom Application-Specific Integrated Circuits (ASICs), such as Google's Tensor Processing Unit (TPU) [122] and Microsoft's Maia 100 [123], are designed specifically for AI workloads, offering optimized performance and efficiency. Specialized accelerators, including purpose-built ASICs and Field-Programmable Gate Arrays (FPGAs), provide further improvements. For instance, custom ASICs and FPGA implementations have been tailored for specific applications, such as sensor processing, using hardware-software co-design to bypass the inefficiencies of general-purpose hardware [124].

Hybrid architectures are also being explored. Some ASICs integrate digital multiply-accumulate (MAC) arrays with analog AiMC (Analog-in-Memory Computing) units, allowing for dynamic selection of the most efficient processing cores depending on the neural network layer [125]. Alternative designs, such as NeuRRAM, employ Resistive Random-Access Memory (RRAM)-based architectures that implement CIM principles, significantly reducing data movement within the chip. This approach has demonstrated 5-8x better energy efficiency for edge AI applications [126]. For large-scale networks like transformers, heterogeneous CIM designs use dense 2D mesh architectures that combine analog CIM tiles with digital cores, offering scalability without sacrificing efficiency [127].

Progress in semiconductor technology is also critical for improving energy efficiency in conventional electronic systems. For example, 3nm process nodes using FinFET multiport SRAM have demonstrated a scaleup in throughput for lower energy consumption [128]. In parallel, researchers are exploring ferroelectric materials to enhance memristor stability, providing reliable analog-AI systems [127, 129].

Memristor-based systems represent an alternative to traditional digital processing. Analog-AI crossbars using memristors can perform low-power matrix operations and, when combined with hybrid digital/analog systems-on-chip (SoC), achieve enhanced precision and efficiency [125]. Neuromorphic hardware further advances this concept by mimicking the brain architecture, enabling efficient parallel computation. For example, chips that emulate synaptic plasticity can significantly reduce energy consumption for parallel computations [130–132]. A more abstract implementation of brain function is represented by Spiking Neural Networks (SNNs), such as Intel's Loihi, which use event-driven processing for energy-efficient temporal tasks [127, 133].

In contrast to these electronics-based approaches, optical computing provides an alternative pathway. Optical systems can perform neural network operations, including MVM, using the inherent parallelism of light. This approach bypasses the data movement bottleneck of traditional architectures, offering a fundamentally different means of achieving efficient computation. In this work, we explore using optics for performing neural network operations, aiming to leverage the advantages of optical parallelism and energy efficiency.

4.1 OPTICAL NEURAL NETWORKS

Researchers have explored optical approaches for implementing AI, and especially neural networks, also referred to as optical neural networks (ONN), since the early days of AI [27]. The earliest work in the 1980s demonstrated implementations of optical fan-in and fan-out, showing the feasibility of optical interconnects [134]. Subsequently, researchers demonstrated an optical implementation of the Hopfield model used for a pattern recognition task [135]. This work showed the feasibility of using optics to implement content-addressable associative memory through optical processing while leveraging the inherent parallelism and massive interconnection capabilities of optical systems. Holographic techniques were also explored to assist in implementing ONNs [136, 137]. Scientists investigated photorefractive crystals, as well as various spatial light modulation techniques, to physically realize the parameters in the models [138–141]. Optoelectronic methods were also explored to leverage the best features of both optics and electronics [142]. Training larger neural networks remained a limitation; therefore, optical training schemes were also proposed for implementations with photorefractive crystals [143, 144]. More details about the early work in ONNs can be found in [145].

However, progress in the field had stalled because of a general lack of interest in AI study during those periods, also referred to as "AI winter" [147]. The falling cost of traditional computing and the emergence of general-purpose graphics pro-

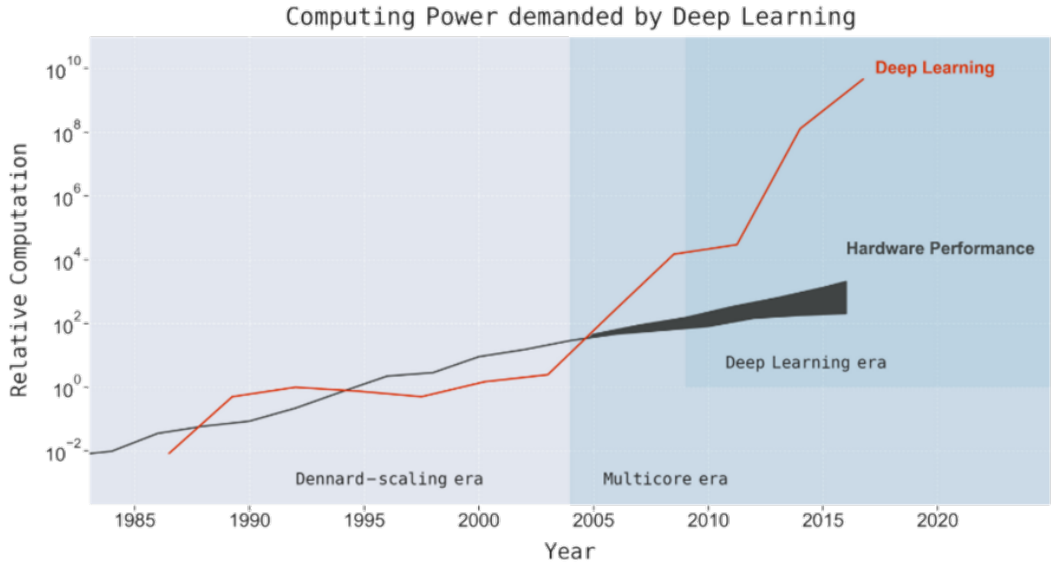


Figure 16: **Compute cost trends for AI models.** The figure illustrates the exponential growth in compute requirements for training state-of-the-art AI models over the years while hardware performance has not been able to keep up. This trend highlights the increasing energy demands and the need for alternative hardware solutions to address the associated challenges. Figure reproduced in its entirety from [146].

cessing unit (GP-GPU) computing, particularly NVIDIA CUDA, made it practical to implement larger models on traditional von Neumann computers, leading to a resurgence in the field. However, conventional computer hardware has been unable to keep up with the increasing size of AI models over the past few years, notwithstanding the resultant sharp increase in energy consumption and costs [146]. Fig. 16 shows the relative increase in compute requirements for training state-of-the-art AI models over the years as opposed to the available compute from conventional hardware. Consequently, there has been a renewed interest in ONNs as a potential solution to build scalable, energy-efficient AI co-processors [148, 149]. The entire field of ONNs can be divided into two broad categories: photonic and free-space approaches [150].

Integrated photonic approaches represent a significant step towards realizing compact, stable, and potentially mass-manufacturable ONN co-processors [27]. These approaches leverage established semiconductor fabrication techniques to create complex optical circuits on small chips. Several key methods have emerged in this domain: One prominent method involves using on-chip Mach-Zehnder Interferometer (MZI) meshes [151–153]. MZIs, composed of interconnected waveguides, phase shifters, and couplers, can be configured to perform arbitrary unitary transformations, which are fundamental to linear operations in neural networks [154, 155]. By cascading these MZIs in programmable arrays, complex matrix multiplications can be realized directly on the photonic chip [151].

Another significant approach utilizes on-chip Micro-Ring Resonator (MRR) weight banks [30, 156]. MRRs are wavelength-selective components whose transmission characteristics can be finely tuned (e.g., thermally or electro-optically) to

represent synaptic weights [30]. Arrays of MRRs, often coupled with wavelength-division multiplexing (WDM), can perform vector-matrix multiplications by modulating the power of different wavelength channels [157, 158]. Phase Change Materials (PCMs) have also been integrated with MRRs to create non-volatile weights and introduce nonlinear activation functions [158, 159].

On-chip diffractive metasurfaces extend the concept of diffractive deep neural networks (D^2NNs) to the integrated platform [160, 161]. These metasurfaces consist of precisely engineered subwavelength nanostructures fabricated on a chip (often a Complementary Metal-Oxide-Semiconductor (CMOS) substrate or within slab waveguides) that sculpt the light propagating through them to perform neural computations [162–164]. This approach offers the potential for very high neuron density and parallel processing [162]. Beyond these, various other on-chip optical components and architectures are being explored. These include using optical scattering units optimized via inverse design techniques [165], three-dimensional (3D) integrated waveguides for dense interconnectivity [166], and specialized photonic tensor cores designed for efficient processing of multidimensional data by exploiting spatial, wavelength, and even radio-frequency modulation of photonic signals. These integrated photonic platforms often aim to reduce energy consumption by performing computations in the optical domain, minimizing costly optoelectronic conversions, and leveraging the inherent parallelism of light [27, 151, 167, 168]. However, despite these advances, integrated photonic approaches face challenges, including the difficulty in simultaneously achieving efficient nonlinearity and reconfigurability on a single platform, managing issues like thermal crosstalk in densely packed active components, overcoming limitations in parallel input dimensions for many waveguide-based designs, and enabling robust scalability when cascading multiple on-chip ONNs. This fundamental aim of leveraging optical computation underscores the drive to overcome such challenges for enhanced energy efficiency [27].

Beyond integrated photonics, free-space optical neural networks (FSONNs) offer an alternative framework for optical computing implementations [27, 169–171]. Instead of confining light within waveguides, FSONNs perform computations by manipulating light as it propagates through free space, interacting with a sequence of engineered optical elements that spatially structure one of the fundamental properties of light. A key advantage of this approach is the ability to directly process incoming optical wavefronts carrying rich information about a scene or object—including spatial amplitude and phase, polarization state, spectral content, and orbital angular momentum (OAM)—often without needing complex preprocessing or optoelectronic conversions [31]. This characteristic makes FSONNs particularly appealing for applications involving visual information processing, computational imaging, and sensing.

The core components of FSONNs are typically engineered diffractive surfaces [31], although other approaches exist. One major category uses elements structured at the wavelength scale ($\geqslant \lambda/2$) to modulate the light. These systems, often referred to as D^2NNs, usually consist of multiple passive layers placed one after another [31]. Light diffracts from one layer to the next, and the "learning" is encoded in the physical structure of these layers, typically by varying the thickness

or refractive index of the material across the surface to impart specific phase shifts onto the wavefront [31]. Because of the larger feature sizes involved, D^2NNs operating at lower frequencies (like THz) can often be rapidly prototyped using 3D printing techniques [31]. Fabricating D^2NNs for visible wavelengths requires more advanced nanofabrication, such as lithography [172]. These networks have successfully demonstrated tasks ranging from object classification and computational imaging to implementing fundamental logic operations [31, 173–175].

Pushing the feature size down to the subwavelength scale ($< \lambda/2$) is possible with metasurfaces within FSONNs [176–178]. Metasurfaces are composed of densely packed arrays of 'meta-atoms' (e.g., plasmonic nanoparticles or dielectric nanopillars) whose individual geometry and arrangement allow for control over the properties of light. Unlike simpler diffractive layers that might primarily modulate phase, metasurfaces can be designed to independently control amplitude, phase, polarization state, spectral response, and even OAM [179–181]. While plasmonic metasurfaces offer design flexibility, they often suffer from higher optical losses; dielectric metasurfaces (using materials like TiO_2, Si, or GaN) generally provide higher efficiency but present greater fabrication challenges [176, 182, 183]. The sophisticated light manipulation offered by metasurfaces opens possibilities for highly parallel processing, potentially encoding multiple computational tasks within a single device based on the input light's properties (e.g., wavelength or polarization multiplexing) [184–186].

For applications requiring adaptability, Spatial Light Modulators (SLMs) serve as reconfigurable diffractive elements in FSONNs. Phase-only SLMs (often using liquid crystals) and amplitude-only SLMs, like Digital Micromirror Devices (DMDs), can be programmed electronically to shape the wavefront, allowing for dynamic modification of the network's function without physical fabrication [69, 187, 188]. This versatility makes SLM-based systems ideal experimental platforms, facilitating research into more complex architectures, including integrating active components to introduce optical nonlinearity — a key challenge in passive optical systems [189]. However, one drawback of SLMs is their relative slow speeds. Demonstrations have included incorporating elements like magneto-optical traps or image intensifiers between SLM layers to act as nonlinear activation functions [189, 190]. A specific free-space architecture worth noting is the 4f optical system. This setup uses a pair of lenses to create a Fourier plane where spatial filtering or modulation can be easily performed, often used for implementing optical convolutions [191].

However, FSONN systems also carry certain disadvantages, such as sensitivity to alignment, fabrication complexity for diffractive layers (for visible wavelengths as well), lack of reconfigurability with static diffractive layers, lower energy efficiency because of optical effects, and the difficulty of implementing optical nonlinearity in low-power implementations [189, 192]. Although using spatial light modulators can address some of these challenges, more progress is needed to address the inefficiency problem and to enable flexible nonlinearity implementation within these systems. One way of addressing these challenges is by combining the strengths and weaknesses of both optical and electronic computing systems. In this context, researchers have explored hybrid optoelectronic systems that combine free-space optical processors for handling computationally intensive linear operations (like

large-scale matrix multiplications or convolutions) at high speed and low power, with conventional electronic hardware performing tasks such as control and processing [191, 193–195]. In this work, we aim to implement an energy-efficient optoelectronic processor that also incorporates operational nonlinearity. This synergistic approach aims to leverage the best of both worlds for optimal performance.

4.2 PRINCIPLE OF OPERATION

The fundamental operational schematic of the multilayer optoelectronic neural network involves a sequence of alternating optical and electronic layers, responsible for executing MVM and implementing nonlinear activation functions, respectively. The system's light source comprises a two-dimensional (2D) array of incoherent light-emitting diodes (LEDs), where each LED's intensity represents an individual neuron's activation state. As this incoherent light propagates, it passes through a spatially encoded amplitude mask. This mask modulates the light, effectively encoding the neural network layer's synaptic weights. A 2D array of photodiodes (PDs) subsequently detects the modulated optical signals. These photodetectors integrate with analog electronic circuitry that performs functions such as converting optical signals to electrical signals and implementing nonlinear operations through differential signal processing. This architecture synergistically combines optics' inherent advantages, particularly its capability for highly parallel and energy-efficient MVM, with electronics' strengths in performing low-energy analog signal processing necessary for realizing nonlinear activation functions, such as the rectified linear unit (ReLU).

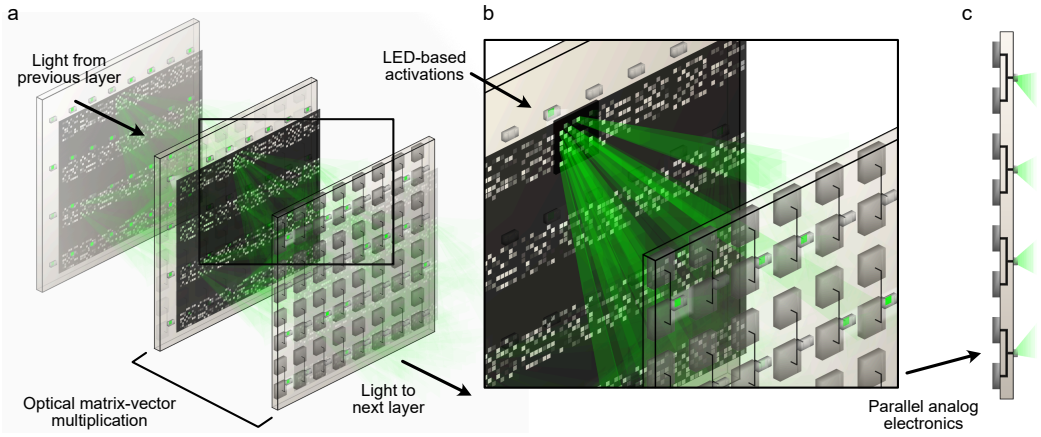


Figure 17: **Conceptual illustration of the multilayer optoelectronic neural network.** (a) The system architecture features interleaved optical layers for matrix-vector multiplication (MVM) and electronic layers for nonlinear processing. (b) Optical MVM is achieved using an array of incoherent light-emitting diodes (LEDs) whose outputs are modulated by an amplitude mask encoding weights, and then projected onto a photodiode (PD) array. (c) The electronic layer consists of neuron units with paired photodiodes for differential input, enabling the implementation of nonlinear activation functions. Figure reproduced from [118]

The following sections briefly introduce the numerical techniques essential for simulating the optical and electronic response of the components used in the multilayer optoelectronic neural network.

4.3 NUMERICAL MODELLING

The system's operation involves two main components; consequently, simulating the multilayer network is also divided into two parts: electronic and optical simulation. We use the numerical framework, "Simulation Program with Integrated Circuit Emphasis" (SPICE), to simulate how electronic circuits function before testing on a matrix-board and subsequently fabricating printed circuit boards (PCBs). Optical simulation techniques simulate the spread of incoherent light through the system and model its interaction with the mask.

4.3.0.1 *Electronic Simulation*

The discussion in this section is based on [196]. Designing and verifying modern electronic circuits heavily rely on simulation tools to predict electrical behavior prior to fabrication. Simulation Program with Integrated Circuit Emphasis (SPICE) is a foundational software standard for such analog circuit simulation. SPICE enables us to model circuits using representations of electronic components and their interconnections, then analyze performance through various simulated tests. During this project, we designed the source, difference ReLU, and detector PCBs to build the system. Each PCB consisted of multiple, independently operating, parallelly connected circuits that needed to meet specific requirements. We designed and then tested these circuits in LTspice, a SPICE simulator developed by Analog Devices.

At its core, SPICE operates on two main principles: detailed device modeling and robust numerical solution techniques.

DEVICE MODELING SPICE utilizes mathematical models to accurately describe the current-voltage (I-V) characteristics of electronic components. These models range in complexity, from basic Level 1 models to more sophisticated, semi-empirical Level 3 models, with advanced models also available for specialized scenarios. The purpose of these models is to translate the physical properties of electronic components into mathematical equations that capture their electrical behavior.

In this section, while our circuit designs incorporate a variety of components including operational amplifiers and Metal-Oxide-Semiconductor Field-Effect Transistors (MOSFETs), we focus on the MOSFET as an illustrative example. Specifically, the MOSFET drain current (I_{DS}) is a function of its terminal voltages and intrinsic device parameters. For a MOSFET operating in the saturation region (when $V_{GS} > V_{th}$), a Level 2 model provides a more accurate representation of the drain current, expressed as:

$$I_{DS} = \frac{1}{2}\mu_n C_{ox}\frac{W}{L}(V_{GS} - V_{th})^2(1 + \lambda V_{DS}) \tag{45}$$

Here, μ_n is the carrier mobility, C_{ox} is the oxide capacitance per unit area, W and L are the channel width and length, V_{GS} is the gate-source voltage, V_{th} is the threshold voltage, V_{DS} is the drain-source voltage, λ is the channel-length modulation factor and $V_{de} = \min(V_{ds}, V_{dsat})$. This model effectively captures the non-linear characteristics of the MOSFET, providing a robust basis for circuit analysis and simulation.

NUMERICAL SIMULATION ENGINE To analyze a complete circuit, SPICE solves a system of non-linear differential equations derived from Kirchhoff's laws and the device model equations. The Newton-Raphson (N-R) iterative method is commonly employed to solve these non-linear equations, particularly for determining DC operating points or for each time-step in a transient analysis. SPICE simulators facilitate several analysis types, including DC operating point, AC small-signal frequency response, and time-domain transient analysis, allowing for comprehensive circuit characterization.

For this project, we optimized the circuit design, component selection, and compensation techniques using the simulator. Subsequently, these optimized designs were first tested on an electronic matrix board before fabricating PCBs. While SPICE provides a robust framework for simulating the electronic response, different modelling approaches are needed to simulate incoherent light. The next section details the optical simulation techniques employed, such as ray tracing and modified angular spectrum method, to accurately model incoherent light propagation and its interaction with the amplitude mask.

4.3.1 *Optical Simulation*

Simulating light propagation through the system employs two main strategies, selected based on the scale of optical features relative to the wavelength of light.

4.3.2 *Ray Tracing for Geometrically Large Features*

When characteristic dimensions of optical elements (like mask patterns) are significantly larger than the light's wavelength (λ), raytracing based on geometrical optics provides an accurate model of light spread and is useful for modelling effects of LED geometry, lenses and reflections in the light path. The light source is an array of N_{LED} LEDs. Each LED, say the j-th LED at position $\mathbf{r}_{LED,j}$, is treated as a point source. It emits N_{rays} individual light rays, each carrying an initial power, $P_{ray,init}$. The probability of a ray being emitted at an angle θ relative to the LED's surface normal is often described by its angular emission profile, $I_{emission}(\theta)$. We model the LED's emission properties based on the datasheet description provided

by the manufacturer. The Monte Carlo method is employed to sample ray directions based on this probability distribution.

The interaction with the semi-transparent mask is a key step. A rigorous treatment of a ray's interaction with each interface of the mask involves the Fresnel equations. When a ray is incident from a medium of refractive index n_1 onto the mask material of refractive index n_2 at an angle of incidence θ_i (measured from the normal to the interface), the angle of refraction θ_t (also measured from the normal) into the mask is determined by Snell's Law:

$$n_1 \sin(\theta_i) = n_2 \sin(\theta_t) \tag{46}$$

The Fresnel equations define the reflectances for s-polarized (R_s) and p-polarized (R_p) light components, which depend on n_1, n_2, θ_i, and θ_t. From these, the average transmittance T_{Fresnel} for unpolarized light across a single interface is given by:

$$T_{\text{Fresnel}} = 1 - \frac{R_s + R_p}{2} \tag{47}$$

This T_{Fresnel} represents the fraction of incident unpolarized light transmitted into the mask material at the first interface, thus relating the external incident intensity to $I_{\text{inc\_int}}$. A similar calculation applies for transmission out of the mask at its second interface. Internal absorption within the mask material of thickness d_{mask} is described by the Beer-Lambert law:

$$I_{\text{transmitted}} = I_{\text{inc\_int}} e^{-\mu_t d_{\text{mask}}} \tag{48}$$

where $I_{\text{inc\_int}}$ is the intensity just inside the mask, and μ_t is the attenuation coefficient combining absorption and scattering effects. However, we consider a simplified model of transmittance for this work. We use the transmittance as a result of the angle of incidence, derived from (47), $T_f(r_{\text{mask}})$ and a local transmittance probability, $T_{\text{mask}}(\mathbf{r}_{\text{mask}})$. If a ray with power P_{ray} strikes the mask at position $\mathbf{r}_{\text{mask}}$, its power after passing through is:

$$P'_{\text{ray}} = P_{\text{ray}} \cdot T_{\text{mask}}(\mathbf{r}_{\text{mask}}) \cdot T_f(r_{\text{mask}}) \tag{49}$$

This approach primarily accounts for the amplitude modulation by the mask, neglecting refraction's effect.

After potentially passing through the mask, the power from all transmitted rays reaching a specific pixel p on the detector plane is summed. This process is repeated for each of the N_{LED} light sources. The total power $P_{\text{det},p}$ at pixel p is the incoherent sum of contributions from all LEDs:

$$P_{det,p} = \sum_{j=1}^{N_{LED}} \left(\sum_{k \in \substack{\text{rays from LED } j \\ \text{on pixel } p}} P'_{ray,jk} \right) \tag{50}$$

where $P'_{ray,jk}$ is the power of the k-th ray from the j-th LED after passing through the mask and reaching pixel p. This sum represents the final simulated intensity distribution on the detector plane.

4.3.3 *Modified Angular Spectrum Method for Diffraction-Limited Features*

When feature sizes on the mask are comparable to or smaller than the wavelength λ, diffraction becomes significant. We adapt the angular spectrum method (ASM) to model this. The ASM describes the propagation of a scalar, monochromatic optical field, $U(x, y, 0)$, from an initial plane ($z = 0$) to a subsequent plane ($z > 0$). A detailed introduction to the angular spectrum method is provided in 2.5.2.

To simulate spatially incoherent light using ASM, we modify this procedure. A set of N_{wf} (e.g., 100) distinct initial wavefronts, $U_m(x, y, 0)$, are generated. These wavefronts share a common amplitude profile, $|U_{source}(x, y, 0)|$, derived from the light source characteristics, but each is endowed with a statistically independent, random phase distribution, $\phi_m(x, y)$:

$$U_m(x, y, 0) = |U_{source}(x, y, 0)| e^{i\phi_m(x,y)} \tag{51}$$

Each of these N_{wf} wavefronts is propagated independently using the ASM as described above, resulting in a set of propagated complex fields $\{U_m(x, y, z)\}_{m=1}^{N_{wf}}$.

The final intensity distribution, $I_{out}(x, y, z)$, at the output plane is obtained by averaging the individual intensities of these propagated wavefronts:

$$I_{out}(x, y, z) = \frac{1}{N_{wf}} \sum_{m=1}^{N_{wf}} |U_m(x, y, z)|^2 \tag{52}$$

This summation of intensities, rather than complex amplitudes, effectively models spatially incoherent light's behavior by averaging out interference patterns that would arise from fixed phase relationships. This approach accounts for spatial incoherence. While temporal incoherence is not explicitly modeled, this simulation is considered sufficient for this work's applications.

In the following chapter, we describe the experimental implementation and the results built up using the principles described in this chapter.

5

LAB SCALE IMPLEMENTATION OF THE OPTOELECTRONIC NEURAL NETWORK

The objective is to physically implement deep neural network computations in an energy-efficient manner, aiming for lower power consumption compared to conventional computing hardware. Data read-in and read-out operations create a bottleneck that significantly contributes to high energy usage in traditional deep neural network computations. This work demonstrates a multilayer optoelectronic architecture designed to address this issue. The core concept, illustrated schematically in Fig. 18a, involves alternating electronic and optical processing layers. Optical layers perform matrix-vector multiplication (MVM) operations, which are computationally and energetically intensive on conventional hardware.

An LED array generates incoherent light for the MVM. A National Instruments (NI) digital-to-analog converter (DAC) PXIe-6739 reads data from a computer. The system amplifies the analog signal, and an array of 64 LEDs, arranged in an 8×8 grid, emits light with intensity proportional to the analog input. The propagated light passes through an amplitude mask, as Fig. 18b shows, modulating the intensity at the mask plane. Trained neural network weights determine the features on the mask, and principles from ray optics guide their positioning. The resulting incoherent light distribution is incident on a 10×10 photodiode (PD) array in the first hidden layer. The system applies a Rectified Linear Unit (ReLU) nonlinearity by subtracting signals from pairs of photodiodes, amplifying the result, and re-emitting it from a 5×10 LED array on the same layer, as shown in Fig. 18c and Fig. 18d. Because light intensity can only be positive, subtracting the signal from a pair of photodiodes implements negative weights and ReLU nonlinearity. This process repeats through another hidden layer before an 8×8 photodiode array detects the signal at the output layer. An NI analog-to-digital converter (ADC) PXIe-6355 amplifies and reads the output signals into the computer. Custom Python code, utilizing NIDAQmx libraries for NI hardware communication and the Holoeye Python library to drive the spatial light modulator (SLM), controls the entire system.

This chapter provides an overview of a lab-scale prototype of the optoelectronic neural network (OENN) system. First, we discuss the optical principle for implementing MVMs optically, followed by the electronic implementation of amplification and nonlinearity. Next, we discuss the system's calibration and the experimental setup. Finally, the chapter discusses the system's performance on classification tasks.

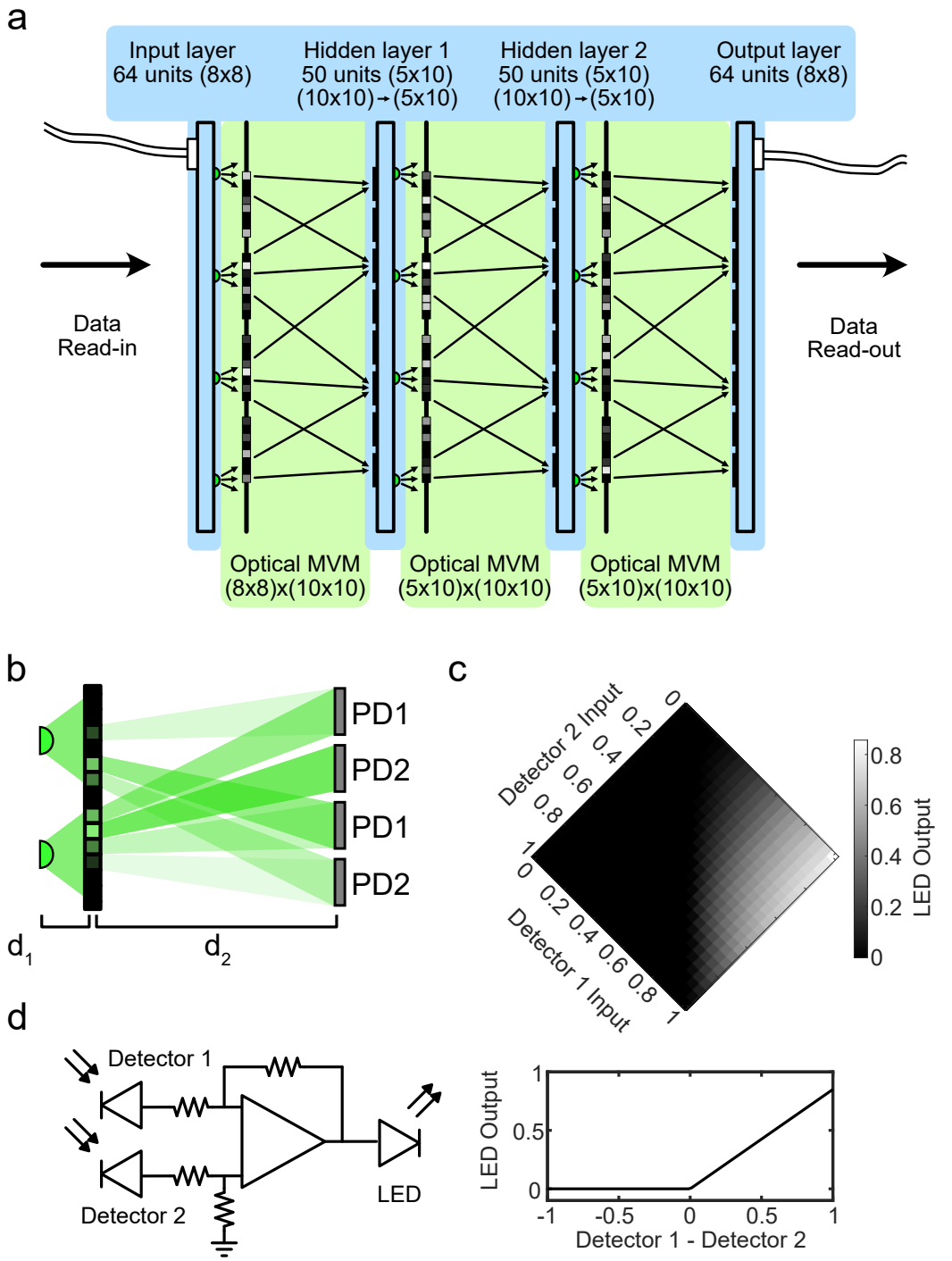


Figure 18: **Components of the Optoelectronic Neural Network (OENN).** (a) Schematic representation of the system architecture. The input board receives data from a computer via a digital-to-analog converter (DAC) and drives an 8×8 grid of 64 input LEDs. The emitted light from these LEDs is optically mapped to a 10×10 photodiode array in the subsequent hidden layer, performing an MVM operation. Signals from pairs of photodiodes are combined, amplified, and drive a 5×10 LED array, which is then optically mapped to the next hidden layer. This process continues through the network until the output layer, where a 10×10 photodiode array detects the final signals. The output photodiodes connect to a readout board that digitizes the signals for further processing. (b) Ray-traced illustration of the MVM implemented in the system. (c) Circuit representation for implementing negative weights, enabling the Rectified Linear Unit (ReLU) nonlinearity shown in (d). Figure adapted from [118].

5.1 OPTICAL IMPLEMENTATION

The optical subsystem of the OENN performs MVM operations. As discussed in Chap. 4, MVM is one of the most energy-intensive operations in traditional electronic computers. In contrast, we implement MVM optically by using either passive optical components or active elements with negligible electrical power consumption. We use the inherent spreading behavior of incoherent light and its interaction with an amplitude modulation mask to perform computational tasks. Specifically, a grayscale mask modulates the intensity-encoded information from a light-emitting diode (LED) array based on local transmission values. This mask is designed from trained neural network weights, adhering to ray optics principles. Figure 19 schematically illustrates the key factors influencing this optical mapping, which the following section further details.

5.1.1 *Optical Matrix-Vector Multiplication*

The amplitude modulation mask implementing the MVM operation consists of an array of smaller transparent blocks. As the neural network we implement is fully-connected, each output from the preceding LED layer maps to every photodiode (PD) in the subsequent layer. If the LED array is arranged in an $(m \times n)$ grid and the photodiode array in a $(p \times q)$ configuration, the mask comprises an $(m \times n)$ array of subarrays, where each subarray contains $(p \times q)$ modulation blocks. Light emitted from an LED at position $(1, 1)$ spreads spatially and is modulated by an individual $(p \times q)$ section of the mask corresponding to the subsequent PDs. This process repeats for each LED source, resulting in an intensity sum on each PD from multiple contributing LEDs. Therefore, precisely calculating the spread from each mask element is important. Figure 19 schematically represents the main parameters that require careful design consideration. Given the scale of system components relative to the optical wavelength, ray optics can accurately model the spread of light from incoherent LED sources, with diffraction effects being negligible. Simulations using a modified angular spectrum method, detailed in Sec. 2.5.2, support this conclusion. Because the angular spectrum method assumes coherent fields, we adapted it for incoherent light by initializing the field with a random phase distribution. We repeated this random initialization over one hundred iterations to compute a statistical average of the light intensity at the output plane, effectively simulating a quasi-incoherent field. The results, shown in Fig. 21, demonstrate a smooth and continuous light intensity distribution at the output plane, validating the use of geometric optics for modeling the system.

We approximate each LED as a point source and trace its rays from the emission plane to the mask and subsequently onto the photodiode plane, following simple geometrical principles. Realizing MVM within this optoelectronic system's optical layers depends on accurately calculating the spread of incoherent LED light through an amplitude mask and onto a detector array. This approach deliberately avoids using traditional refractive or diffractive elements. Within the operational regime defined by the experimental parameters, where component dimensions

and separations are substantially larger than the optical wavelength (λ), geometric optics can effectively describe light's behavior, allowing us to ignore diffraction phenomena. Figure 19 provides a visual guide to these foundational geometric principles.

Light emission originates from a planar array of LEDs. Each LED acts as an independent source, and its emitted light travels a distance d_1 to intersect an intermediate plane containing the amplitude mask. This mask imposes a spatially varying transmission function, effectively multiplying the incident light intensity by the desired matrix weight values W^{ij}. The modulated light then propagates an additional distance d_2 before impinging upon the PD array at the detection plane. The ratio of the total propagation distance ($d_1 + d_2$) to the initial LED-mask distance (d_1) defines the geometric magnification factor, $M = (d_1 + d_2)/d_1$. This factor dictates the transverse scaling of any light pattern as it projects from the mask plane to the detector plane. Consequently, a feature located at a transverse position (x', y') on the mask will appear at position (Mx', My') on the detector plane relative to the projection axis, consistent with the relationship $d_{y2} = M \cdot d_{y1}$ shown in Fig. 19(a).

Crucially, light arriving at the detector corresponding to a single weight element is not a point but forms a spot of finite extent. This spatial spread arises physically from the non-zero dimensions of the originating LED emitter, W_{LED}, and the finite size of the specific aperture on the amplitude mask, w_{amp}, through which light passes, both geometrically projected and scaled onto the detector plane. The LED source's contribution effectively creates a convolution of the LED die with the mask with a characteristic size proportional to $(M-1)W_{LED}$, while the mask aperture projects to a size proportional to Mw_{amp} (related to components 'a' and 'b' in Fig. 19(b)). In the geometric limit, the combination of these two effects determines the total spot extent. This finite spot size is fundamentally linked to crosstalk between adjacent photodiodes, making it a critical parameter influencing system performance. Therefore, maintaining a spot extent comfortably less than the photodiode pitch S_{PD} is essential to clearly reflect the expected results from in-silico inference.

These geometric projection rules tightly couple the physical layout of components across the different planes, as Fig. 19 schematically suggests. The spacing between individual weight elements on the mask, s_{amp}, must be precisely $1/M$ times the photodiode spacing S_{PD} (i.e., $S_{PD} = M \cdot s_{amp}$, Fig. 19(c)) to ensure correct mapping. Furthermore, the mask itself is structured into distinct zones, each corresponding to an input LED. The spatial separation S_{MM} required between the centers of these zones on the mask relates to the LED pitch S_{LED} via the relationship $S_{MM} = S_{LED} \cdot (M-1)/M$ (Fig. 19(e)). Adhering to these geometric scaling laws ensures that the intensity contribution from each LED_i, modulated by the appropriate weight transmittance W^{ij} on the mask, arrives predominantly at the intended PD_j. The total intensity measured at PD_j then approximates the desired MVM result, $O^j_{PD} \approx \sum_i I^i_{LED} \cdot W^{ij}$, where I^i_{LED} is the intensity of the i-th LED. The design process thus involves selecting d_1, d_2, and component dimensions to optimize this process, minimizing crosstalk by managing the spot spread relative to S_{PD}, while simultaneously maximizing collected optical power for system effi-

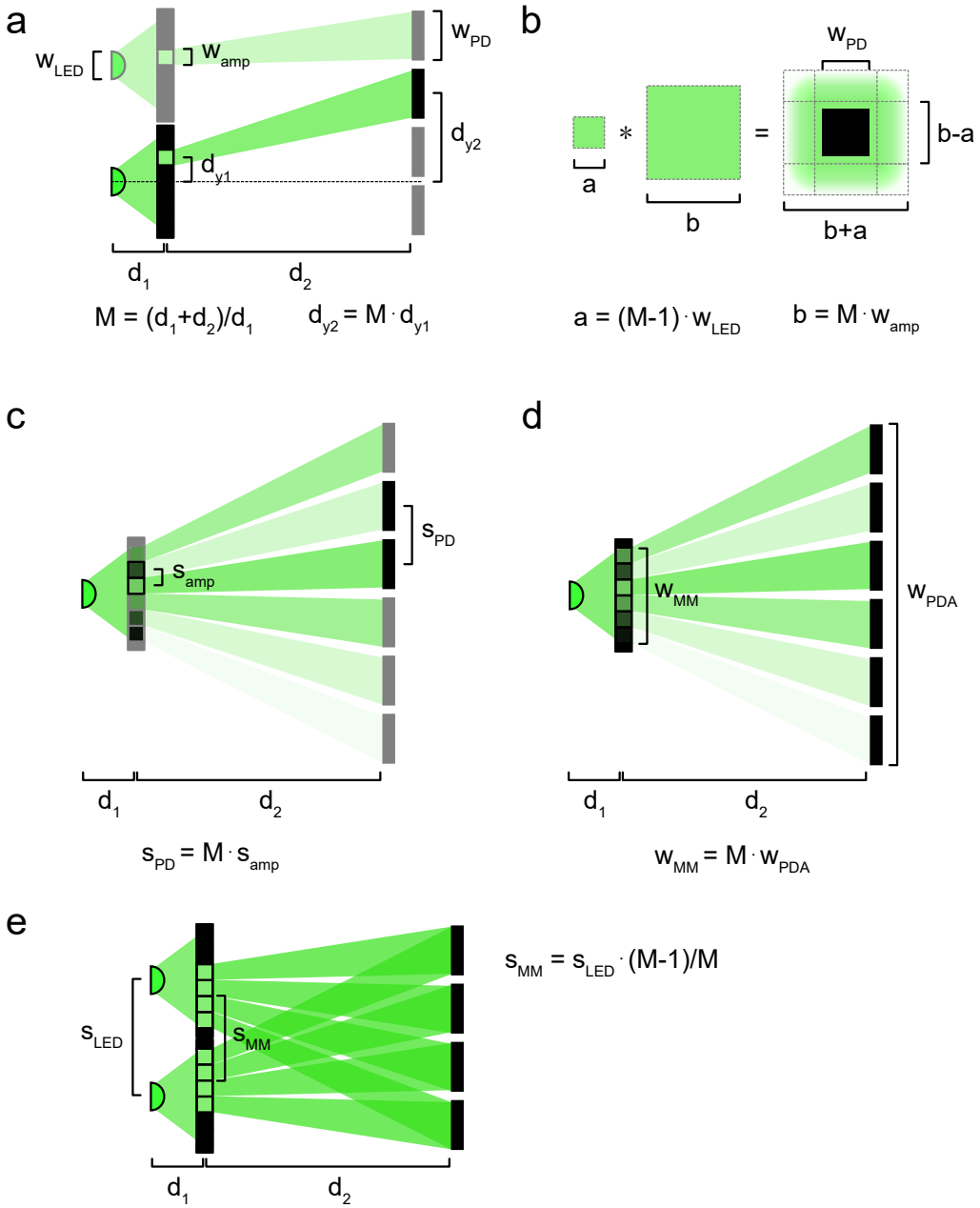


Figure 19: **Ray-tracing representation of the optical operation in the OENN.** This figure demonstrates implementing a fully-connected optical MVM operation in the system. The principles of ray tracing design the amplitude mask. Each LED in (a) is multiplied by a sub-array on the mask whose feature size magnification, M, determines. The spot size made by a feature in (b) depends on the LED emitter size, shape of the mask feature, and magnification factor. The photodiode spacing in (c) depends on the magnification and the spacing between mask features. (d) The magnification factor M scales the total output region. (e) Each LED is associated with a submask that encodes the weights for the MVM operation. Magnification and LED spacing can evaluate the spacing between neighboring masks. Figure reproduced from [118].

ciency. The following section describes two different techniques to physically implement the MVM mask detailed here.

5.1.2 *Physical Realization of the MVM Amplitude Mask*

The critical step of encoding matrix weights, denoted by W^{ij}, requires precise positioning of the transparency blocks that constitute the amplitude mask. This mask defines the transmission function $T(x, y)$ at each point (x, y) along a plane intersecting the propagating light. Two principal strategies exist for realizing such masks: static masks and dynamic masks.

Static Amplitude Masks

Static amplitude masks involve creating a fixed physical structure whose local transmission $T(x, y)$ encodes the desired matrix weights. Various fabrication techniques can fabricate such a mask, with inkjet printing on transparent sheets being the simplest method. Although this method suits rapid prototyping, it typically suffers from limitations in achievable spatial resolution Δx, optical contrast ratio (T_{max}/T_{min}), and mechanical stability [197]. Imperfections in the transparent film substrate or insufficient rigidity may also induce phase aberrations $\phi(x, y)$ in the transmitted wavefront, degrading system accuracy.

For applications demanding high precision, standard microfabrication techniques, such as photolithographic patterning of optically dense materials (e.g., chromium) onto high-quality substrates (e.g., quartz or fused silica), are preferable. Such methods provide excellent spatial resolution, high contrast, superior surface flatness, and long-term mechanical durability, making them suitable for this application. However, conventional lithographic processes are often optimized for binary patterning, where transmission is either 0 or 1, in contrast to the continuous-valued weights W^{ij} neural networks typically require.

We use spatial dithering techniques to solve this challenge, effectively creating an analog transmission response using only binary (0 or 1) substructures. The principle relies on modulating the local density of transparent features across a subregion Ω so that the spatial average approximates the desired transmission T_{target} [198]:

$$\langle T_{binary}(x, y)\rangle_{\Omega} = \frac{1}{|\Omega|}\iint_{\Omega} T_{binary}(x, y)\, dx\, dy \approx T_{target}. \tag{53}$$

In our implementation, we subdivide the area corresponding to a single W^{ij} into an 5×5 array of sub-pixels, each of which a dithering algorithm individually assigns as transparent or opaque. Techniques such as error diffusion sequentially quantize each sub-pixel, calculate the resulting quantization error $e = T_{target} - T_{binary}$, and redistribute this error to neighboring pixels, thereby ensuring the local spatial average converges toward T_{target}. This strategy allows encoding

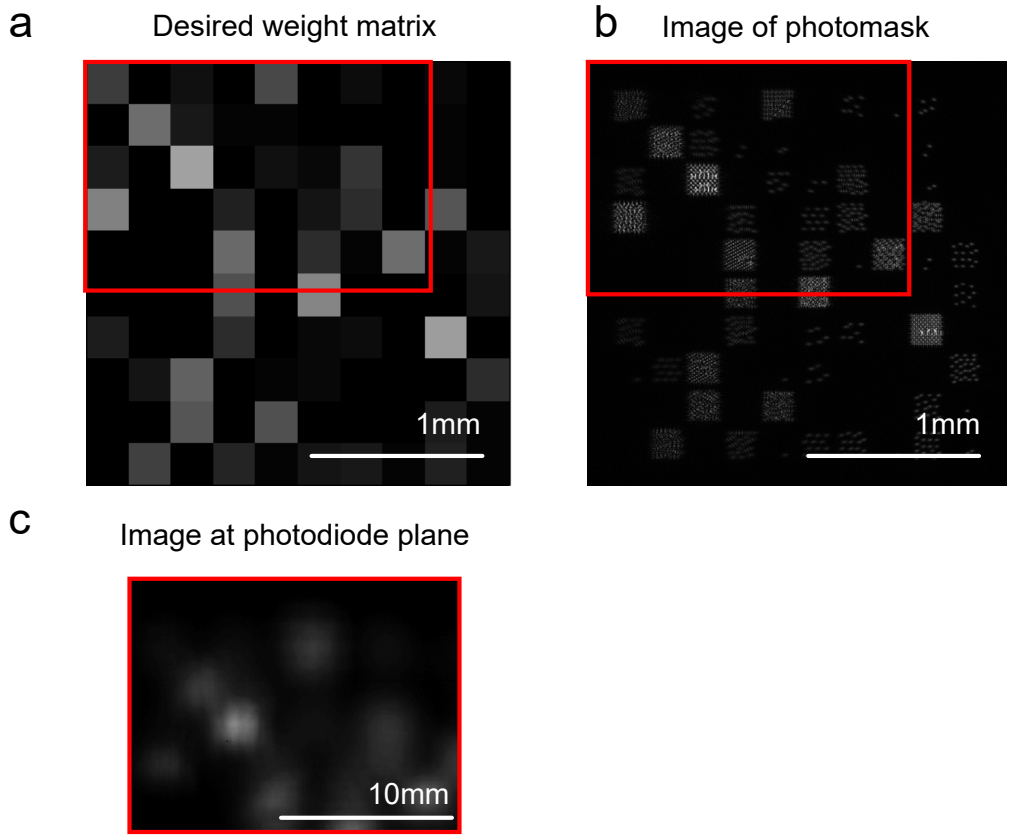


Figure 20: **Implementation of amplitude masks**. For a given desired amplitude mask (a), two methods can implement the mask. The first method uses a dithered binary mask, shown in (b). The resultant intensity distribution after propagation shows a smooth intensity profile (c). Image adapted from [118].

approximately $O(n^2)$ effective transmission levels. The lithographic toolchain then receives the precise geometric coordinates of transparent sub-pixels for mask fabrication.

Experimental validation appears in Fig. 20: Panel (a) depicts the intended continuous-valued weight matrix W^{ij}; Panel (b) shows a micrograph of the fabricated high-resolution binary mask; Panel (c) presents the measured optical intensity distribution $I(x, y)$ after incoherent illumination and free-space propagation through the mask. The inherent optical low-pass filtering smooths high spatial frequencies, recovering a continuous-tone intensity profile that faithfully represents the original weight values. These results confirm that spatial dithering can achieve continuous amplitude encoding.

Dynamic Amplitude Masks

Alternatively, twisted nematic liquid crystal spatial light modulators (TNLC-SLMs) can realize dynamically reconfigurable amplitude masks [70]. In this configuration, a voltage-controlled liquid crystal panel is sandwiched between two polariz-

ers. The applied voltage V_k to each pixel modulates the twisted nematic LC cell's twist angle, altering the transmitted light's polarization state. An analyzer converts this polarization modulation into amplitude modulation, resulting in a pixel transmission function $T_k = T(V_k)$, typically controlled with 8-bit digital precision ($k = 0, \ldots, 255$).

The primary advantage of dynamic masks is their reconfigurability. The encoded weight matrix W^{ij}, represented by pixel transmissions T_k, can be electronically updated without physically replacing the mask, allowing for a more flexible alignment procedure and rapid iteration over different weight masks. This flexibility speeds up the testing process. However, this reconfigurability has certain limitations. The SLM's pixel pitch p constrains the achievable spatial resolution, imposing a Nyquist limit $\sim 1/(2p)$ on the highest spatial frequencies that can be faithfully represented. Consequently, dynamic masks may not achieve the same spatial density or optical fidelity as static masks. Moreover, SLMs are active devices requiring continuous electrical power P_{SLM}, including contributions from driving electronics and potential backlighting. Although the liquid crystal layer itself may draw low power for static patterns due to its capacitive nature, the total consumption is non-negligible and must be considered when evaluating overall system energy efficiency.

Thus, choosing between static and dynamic amplitude masks involves balancing trade-offs among adaptability, resolution, and energy consumption. Static, lithographically patterned masks offer superior optical performance and passive operation but lack flexibility. Dynamic SLM-based masks provide valuable reprogrammability at the cost of reduced resolution and active power demands. In the present experiments, we use dynamic masks to rapidly prototype and test different trained networks. However, static masks could replace them in the future for a more energy-efficient and integrated implementation.

In addition to the choice of masking technology, the free-space propagation geometry connecting the source, mask, and detection planes equally governs the overall effectiveness of the optical MVM layer. Carefully optimizing these geometric parameters is essential to minimize optical crosstalk and maximize signal fidelity, especially when scaling to larger network sizes.

5.2 ELECTRONIC IMPLEMENTATION

The electronic circuits form the OENN's active core, allowing us to generate input signals, detect and amplify signals from the previous layer, implement analog signal subtraction (see Fig. 18), and enable interfacing with external control and measurement hardware. We had three objectives when designing the circuits: achieve an operational bandwidth capable of supporting signal modulation frequencies approaching 1 MHz (with 800 kHz demonstrated experimentally); minimize electrical power consumption per computational channel, making the system energy-efficient; and maintain low noise levels throughout the signal path, thereby ensuring accurate analog computation and maximizing each channel's effective signal-to-noise ratio (SNR).

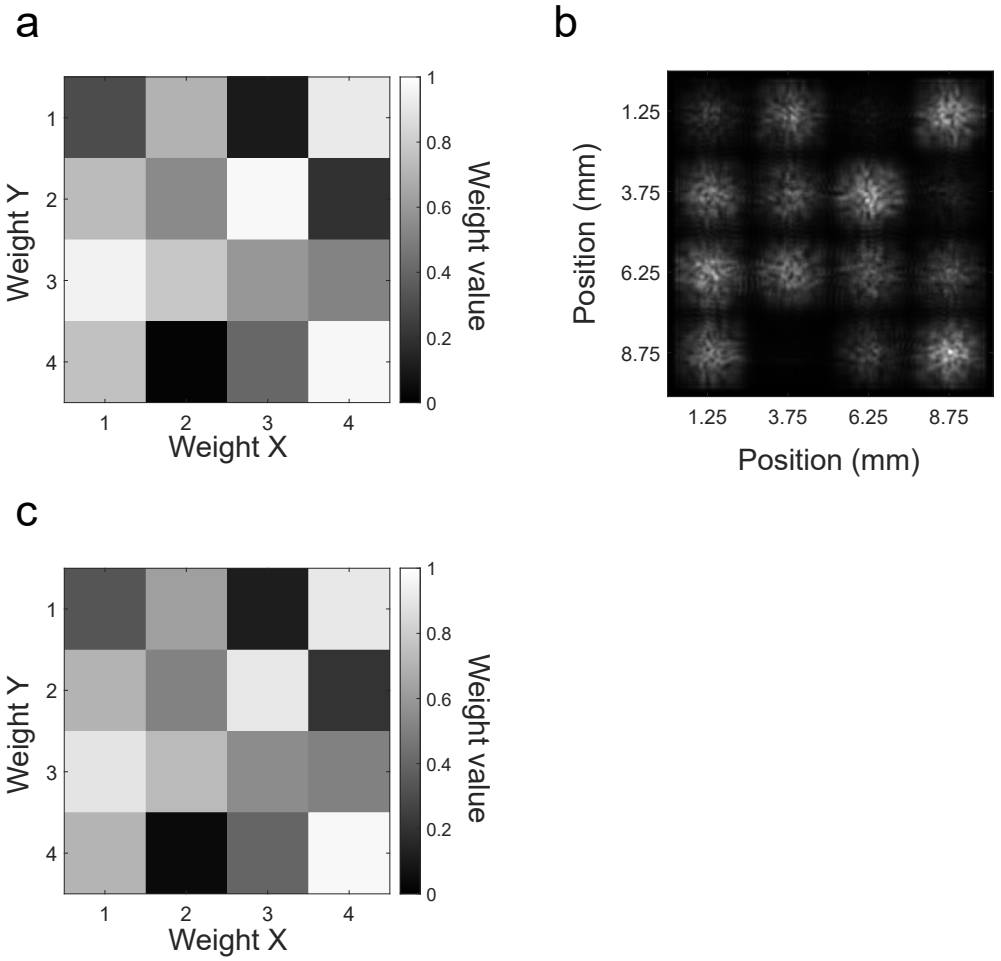


Figure 21: **Impact of Diffraction on Optical Propagation.** Simulated results using a modified angular spectrum propagation method for LED light passing through amplitude masks. The setup assumes a photodiode spacing of 2.5 mm and an LED die size of 200 µm. (a) Target amplitude mask weights. (b) Resulting intensity distribution at the output plane. (c) Histogram of output values from the central region of (b), grouped into bins for analysis. Figure reproduced from [118].

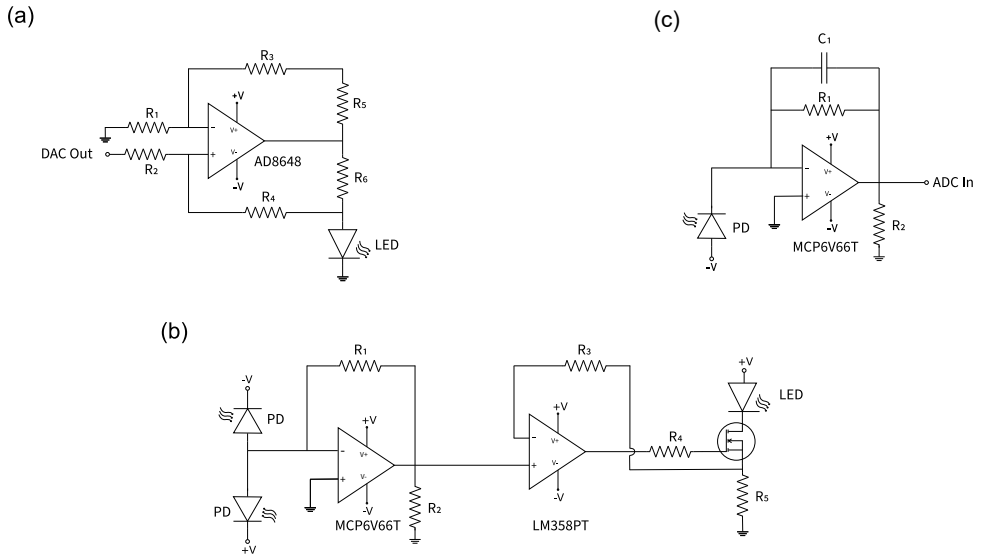


Figure 22: **Operational amplifier circuits used in the OENN electronic modules.** (a) Circuit diagram for driving input LEDs based on digital signals. (b) Readout circuit employed to capture and digitize signals from the final photodiode array. (c) Intermediate circuit responsible for detecting photodiode signals, computing differences between pairs, amplifying the result, and driving subsequent LEDs. Figure reproduced from [118].

The computational system consists of three distinct types of circuit boards, corresponding to the different functional layers within the OENN architecture: an Input board for driving the initial LED array, Intermediate boards for implementing hidden layer functionality (detection, differencing, amplification, and re-emission), and an Output board for final signal detection and readout.

- **Input Board:** Converts digital control signals into modulated light intensities at the first LED array (circuit details in Fig. 22a).
- **Intermediate Board:** Implements core hidden neuron functionality, including photodetection, differential amplification (realizing ReLU activation), and driving the subsequent LED array (Figure 22b).
- **Output Board:** Performs final optical signal detection and amplification to produce analog voltages for external digitization (Fig. 22c).

The circuit development followed a standard electronic design flow: initial concepts were modeled and simulated in LTSPICE (see Sec. 4.3.0.1) to evaluate bandwidth, gain, noise, and stability; prototypes were constructed and tested experimentally (using matrix boards) to validate functionality before final implementation on custom-designed printed circuit boards (PCBs). Figure 22 shows the schematics for each of the three circuit boards.

5.2.1 *Circuit Operation*

5.2.1.1 *Input Board LED Driver (Fig. 22a):*

The Input board interfaces the controlling computer with the OENN's optical input. Parallel analog voltage signals V_{DAC}, generated by a DAC (National Instruments PXIe-6739), provide input activations $V_{DAC}^{(i)}$ for each i-th neuron. Each input channel circuit converts its corresponding $V_{DAC}^{(i)}$ into a proportional LED drive current $I_{LED}^{(i)}$, suitable for each individual LED (Würth 150040GS73220) within the 8×8 array. The driver circuit employs an operational amplifier (AD8648) in a voltage-to-current converter topology. The emitted optical power, $P_{opt}^{(i)}$, is proportional to the forward current:

$$P_{opt}^{(i)} \approx \eta_{ext} \frac{h\nu}{q} I_{LED}^{(i)} \tag{54}$$

where η_{ext} is the external quantum efficiency, $h\nu$ is the energy of each emitted photon, q is the elementary charge and $I_{LED^{(i)}}$ is the injected electrical current through the i-th LED. Design challenges included achieving the target modulation bandwidth (up to 1 MHz) while accounting for the LED's nonlinear current-voltage (I-V) and light-current (L-I) characteristics. The LED forward current follows the ideal diode equation:

$$I_{LED} \approx I_s \left(\exp\left(\frac{V_D}{n V_T} \right) - 1 \right) \tag{55}$$

where I_s is the reverse saturation current, V_D is the diode voltage, n is the ideality factor, and V_T is the thermal voltage. This behavior requires either circuit linearization strategies or calibration. As we discuss in Sec. 5.3, we experimentally characterize our circuit boards' responses, fit the measured data, and use this model for driving the LEDs.

5.2.1.2 *Intermediate Board Neuron Circuit (Fig. 22b):*

The Intermediate board implements a hidden layer in the OENN, with each board comprising $5 \times 10 = 50$ replicated neuron circuits. Each neuron receives differential optical inputs and realizes an effective ReLU activation before driving a subsequent LED. A pair of photodiodes (SFH2704) are arranged in a balanced configuration to detect optical inputs. This arrangement generates photocurrents flowing in opposite directions at the operational amplifier's inverting input, creating an algebraic difference ($I_{PD+} - I_{PD-}$) at the virtual ground node.

The initial transimpedance amplification (TIA) stage, built around the first op-amp (MCP6V66T), converts this differential photocurrent into a voltage, with a

gain set by the feedback resistor R_1 (in Ω) [199]. The photodiodes are reverse-biased to minimize junction capacitance, significantly enhancing the circuit's response speed and linearity. The second amplifier stage, utilizing another op-amp (LM358), operates as a voltage-controlled constant current sink [200] that modulates an output transistor's (BSS138PW MOSFET) gate voltage. The op-amp continuously adjusts the gate drive to maintain input voltage equality, causing the voltage across R_5 to track the non-inverting input voltage. This mechanism controls the output current I_{LED} through the subsequent LED (Wurth 150040GS73220) [201].

The ReLU nonlinearity, $f(x) = \max(0, x)$, is intrinsically approximated by the combined behavior of the current sink and the LED's forward voltage threshold. Light emission occurs only when the differential input photocurrent produces a positive output from the first amplification stage sufficient to forward-bias the LED, resulting in:

$$P_{out} \propto \max(0, V_{diff} - V'_{th}) \equiv \max(0, A_d(I_{PD+} - I_{PD-}) - V_{th}), \tag{56}$$

where V_{th} is the input-referred threshold voltage, and A_d denotes the transimpedance gain (in V/A).

Although not a perfect realization, the circuit achieves a good approximation of the ReLU function, exhibiting some nonlinearity in the transition region. Careful component selection and adding explicit compensation capacitors in feedback paths maintain circuit stability, which can introduce stability challenges at higher frequencies [199]. The overall speed of the neuron circuit is influenced by multiple factors, including photodiode junction capacitance, op-amp input capacitance, op-amp slew rate limitations, and MOSFET gate capacitance.

5.2.1.3 *Output Board Readout Circuit (Figure 22c):*

The Output board captures the final optical outputs via an 8×8 photodiode array and converts them into analog voltages V_{out} suitable for external digitization (ADC: National Instruments PXIe-6355). Each channel uses a dedicated TIA based on an op-amp (MCP6V66T), converting photocurrent I_{PD} into an output voltage:

$$V_{out} = -I_{PD} \cdot R_f, \tag{57}$$

where R_f is the feedback resistor. Critical performance metrics include gain, bandwidth, and noise. The -3 dB bandwidth can be approximated as [202]:

$$f_{-3dB} \approx \frac{1}{2\pi R_f C_{tot}}, \tag{58}$$

where C_{tot} includes the photodiode, op-amp input, and parasitic capacitances. Adding a feedback capacitance stabilized the circuit [199]. Key noise sources include the thermal noise of R_f, and the op-amp's voltage and current noise (e_n, i_n). The output must remain within the linear dynamic range of both the op-amp and the ADC.

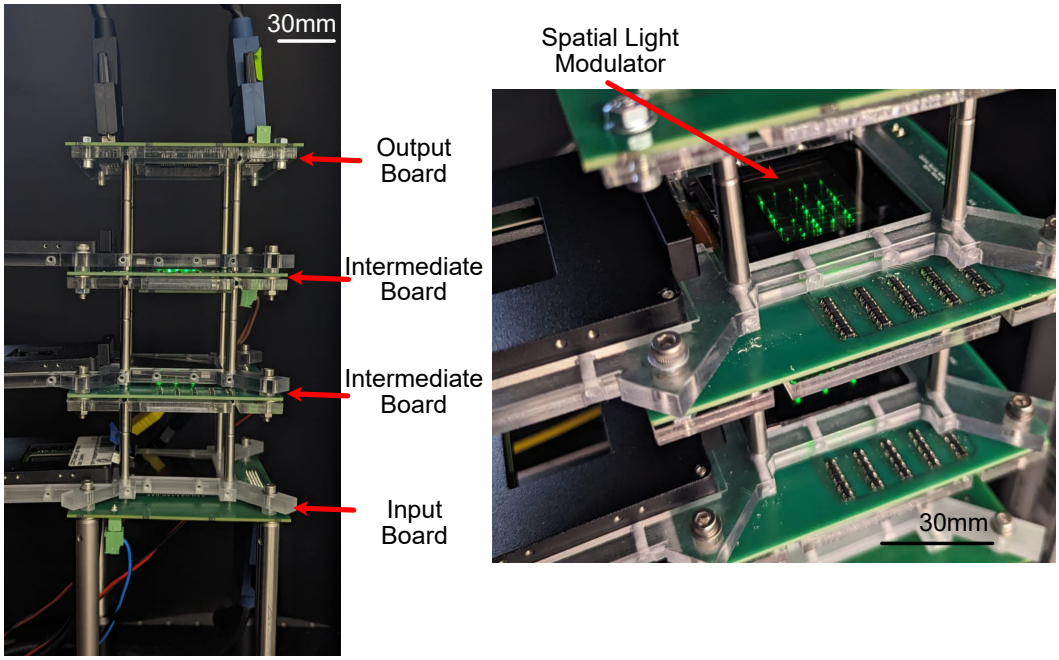


Figure 23: **Overview of the experimental hardware setup for the OENN.** Photograph showing the full optoelectronic network layout. Signal processing starts at the input board (base), proceeds through two successive intermediate layers, and concludes at the output layer (top). A spatial light modulator and polarizer pair are used in combination to dynamically encode optical weights.

The electronic system was intentionally modular, employing distinct PCB types for input, intermediate, and output functionalities. This modular approach facilitated independent layer testing, simplified system assembly, and allowed for straightforward upgrades or modifications, offering considerable practical advantages during research and development.

Figure 23 shows a photograph of the assembled OENN, with the input board at the base, two intermediate layers in the middle, and the output layer at the top. SLMs are positioned between each LED and photodiode layer, with a pair of orthogonally placed sheet polarizers used to control light intensity. A computer drives the SLM via a USB interface, while the input and output boards connect to a PXI chassis for data acquisition and control. We developed a mechanical mounting system using custom 3D-printed parts and 30mm cage rods from Thorlabs. The entire system is housed inside a light-blocking frame to prevent noise from ambient light. A custom Python script, interfacing with the PXI chassis and SLM, controls the entire system. The script allows real-time control of input signals, SLM patterns, and data acquisition from the output board. This modular design enables easy reconfiguration and testing of different neural network architectures and weight configurations. We then calibrate the system under the same conditions as the final experiments to ensure proper alignment and expected functioning of optical and electronic components.

5.3 CHARACTERIZATION AND CALIBRATION

As described in Sec. 5.2, the OENN consists of multiple electronic circuits on PCBs with components such as op-amps, photodiodes, and LEDs. While component datasheets summarize individual performance, real-world behavior often deviates due to manufacturing variability. Additionally, the system exhibits inherent nonlinear responses that must be characterized to transfer trained neural networks onto the hardware effectively.

First, we characterize the LEDs' nonlinear response, focusing on the observed nonlinear relationship between applied input voltage and emitted light intensity. This behavior is intrinsic to semiconductor LEDs and arises from manufacturing variability. If uncorrected, it can introduce systemic errors in neuronal activations that propagate through the network.

To address this, we measure the output intensity as a function of input voltage for each LED. Figure 24 shows the responses for a selected subset. The data clearly show the nonlinear relationship. Notably, one malfunctioning LED exhibits atypical behavior, likely due to a defect in its electronic circuit. However, because each circuit operates independently, such malfunctions do not significantly affect the board's overall functionality if the specific circuit is unused. We fit the measured responses using a model, and the resulting fit parameters calibrate the input voltages, effectively linearizing the LED response during operation.

We also characterize the response of the electronic circuit implementing the difference-ReLU operation, which ideally computes $\mathrm{LED}_{\mathrm{output}} = \mathrm{ReLU}(I_1 - I_2)$, where I_1 and I_2 are photocurrents from a pair of photodiodes. In practice, deviations from this ideal behavior arise from mismatched photodiode sensitivities and offset terms introduced by operational amplifiers, leading to a more accurate empirical model: $\mathrm{LED}_{\mathrm{output}} = \mathrm{ReLU}(c_1 I_1 - c_2 I_2 + c_3)$.

These non-idealities arise from several sources: the previously characterized nonlinear LED emission response (Fig. 24); residual nonlinearity in photodiodes, which are operated in photoconductive mode but remain subject to nonlinear responsivity and junction capacitance at low or high light levels; and op-amp limitations. The first op-amp (MCP6V66T), configured as a TIA, and the second (LM358), acting as a current sink, are both optimized for small-signal linearity but are susceptible to input offset voltages and bias currents. These offsets become significant due to the circuit's high overall gain (≈ 30000), potentially shifting the LED output. Combined, these effects introduce neuron-specific deviations: unequal input weighting ($c_1 \neq c_2$) and an offset shift (c_3), both degrading activation accuracy.

To quantify these deviations, we systematically vary I_1 and I_2, record the corresponding LED output (Fig. 25a), and fit measurements to the model above. The fit captures the circuit response well (Fig. 25b). Figure 25c shows the distribution of offset terms c_3, and Fig. 25d summarizes aggregated neuron responses. This procedure repeats for all neurons, and the resulting fit parameters (c_1, c_2, c_3) calibrate the amplitude mask weights, compensating for circuit-specific nonlinearities.

Another important aspect of device calibration involves characterizing the system's temporal response, specifically signal propagation delay through each op-

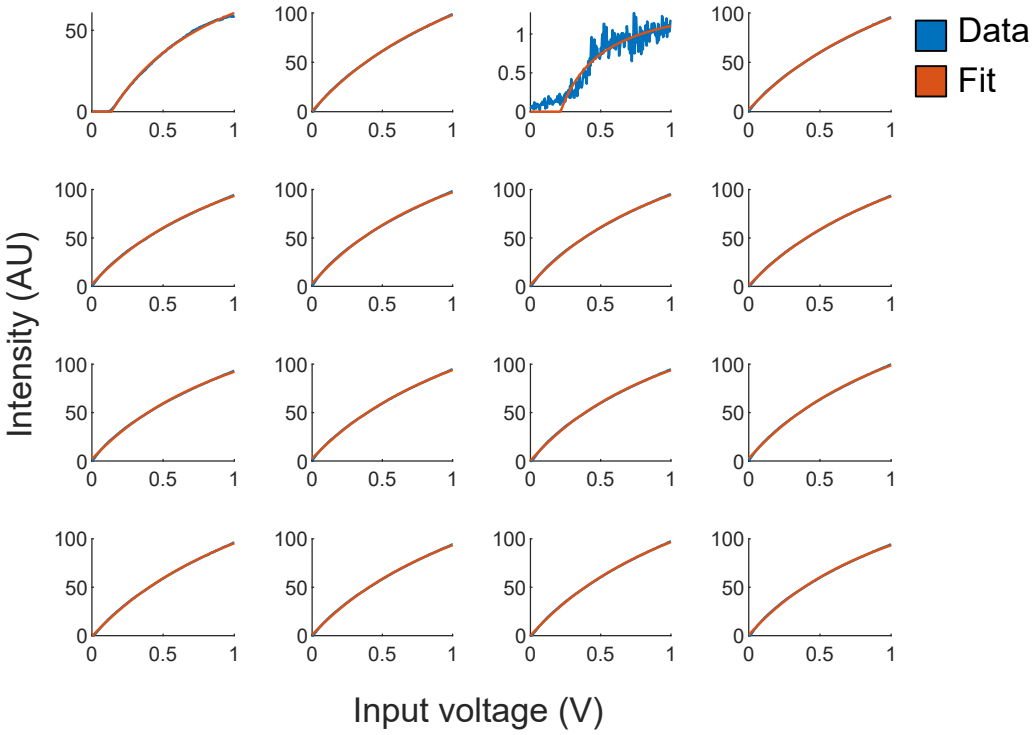


Figure 24: **Characterization of individual LED performance on the input board.** Measured output intensity versus applied input voltage for selected LEDs. One malfunctioning LED (row 1, column 3) exhibits atypical behavior. However, because each LED operates independently, the malfunction does not propagate to neighboring units. Figure reproduced from [118].

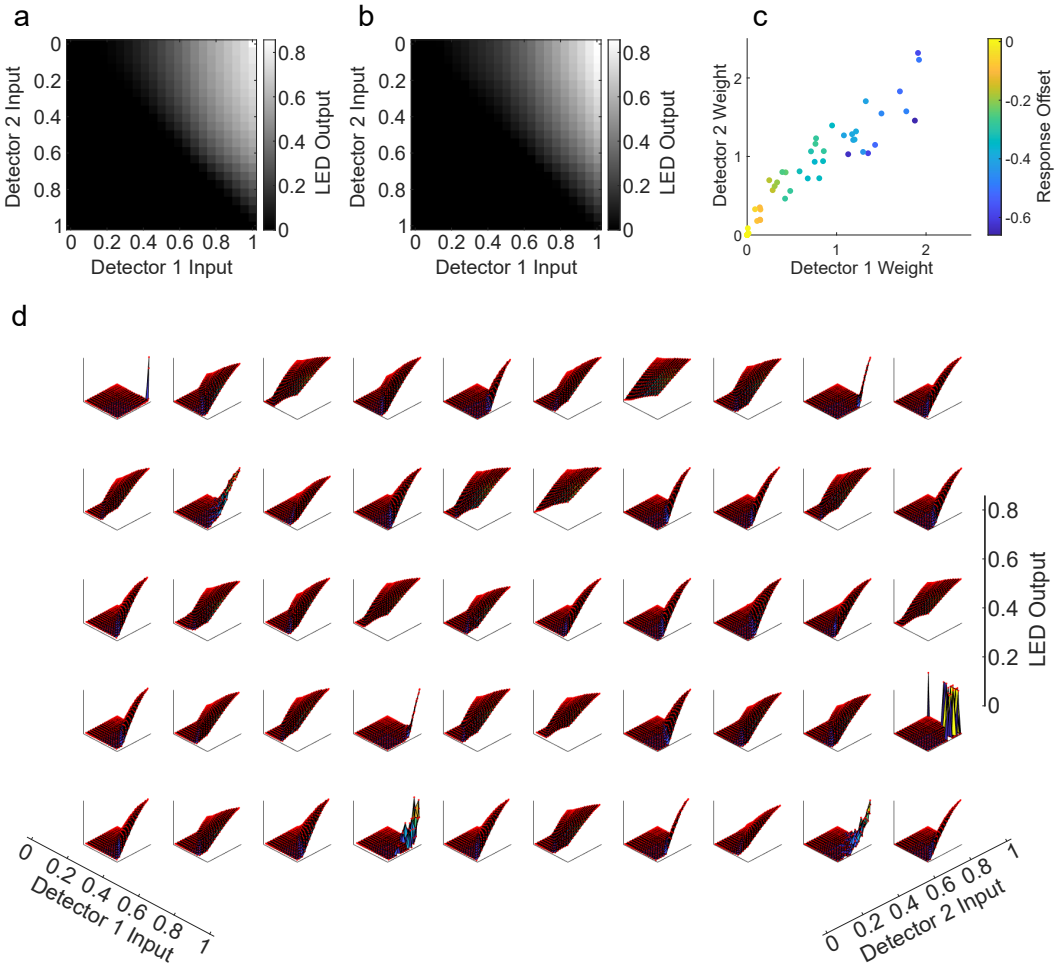


Figure 25: **Fitting of experimental neuron responses to difference-ReLU behavior.** (a) Measured output current from difference circuits as a function of photodiode inputs. (b) Fitted model matching experimental data to the function $\mathrm{LED}_{\mathrm{output}} = \mathrm{ReLU}(c_1 I_1 - c_2 I_2 + c_3)$. (c) Offset added to light emission resulting from offsets throughout the electronic circuit. (d) Aggregated measured response curves from all 50 neurons implemented on a representative intermediate board. Figure adapted from [118].

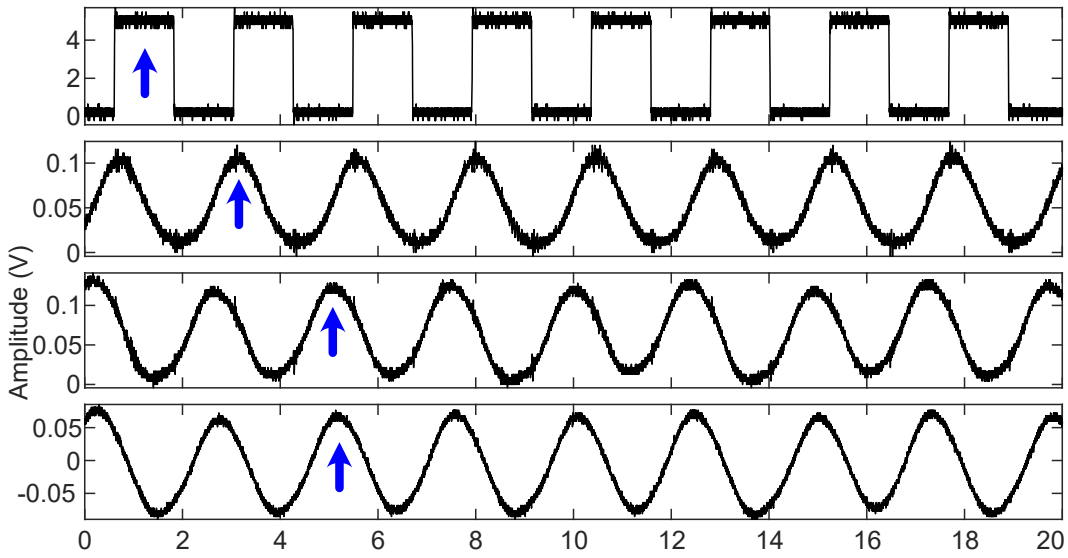


Figure 26: **Temporal response of optoelectronic neural network operation.** (a) Temporal signal traces as an 800 kHz square wave sequentially propagates through two intermediate optoelectronic layers and the output photodiode, highlighting cumulative delay at each stage. Figure adapted from [118].

toelectronic layer. While individual components like photodiodes and LEDs can operate at much higher frequencies—as Fig. 27a demonstrates, where an LED is driven at 10 MHz—the gain-bandwidth product and slew rate limitations of discrete operational amplifiers used in hidden layer circuitry constrain overall system speed. This delay imposes a ceiling on the maximum operational frequency and, by extension, OENN throughput. Moreover, cumulative delays across multiple layers can impact synchronization, particularly in recurrent architectures.

To quantify this, we apply an 800 kHz square wave to the input board's LED array and measure the signal as it propagates through the first and second hidden layers before reaching the output board, as shown in Fig. 26a. The recorded signals, with blue arrows highlighting the tracked wavefront, reveal that each hidden layer introduces an approximate 2 µs delay, while the output detector responds almost instantaneously. Recognizing this bottleneck is integral to system-level calibration; although it currently limits operating speed, future designs can mitigate this constraint by employing faster, potentially integrated, electronic circuits, as discussed in Chap. 7.

In addition to electronics, spatial components such as the SLM also require calibration. The SLM operates by modulating light polarization through spatial variations in liquid crystal molecule orientation, with a pair of polarizers converting this modulation into amplitude control. While ideally the SLM would offer uniform and precise transmission control from 0% to 100%, the physical device exhibits non-ideal behavior. These include spatially varying maximum transmission across the display area and a finite extinction ratio, both degrading encoded weight fidelity.

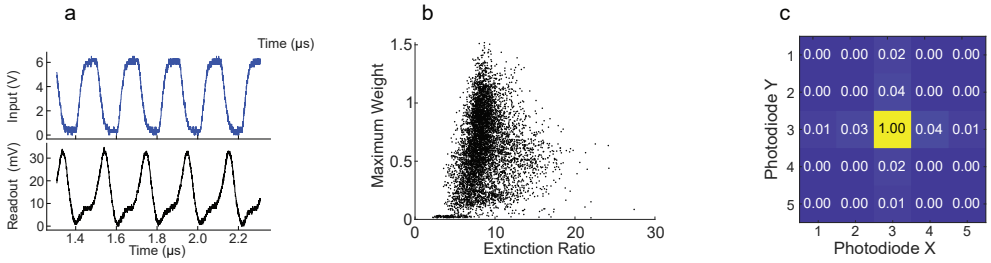


Figure 27: **Characterization of high-speed operation and spatial components.** (a) Measured frequency response of a representative LED, driven by a 10 MHz square wave (blue), with a photodiode recording the output (black). (b) Spatial distribution of maximum achievable transmission (related to extinction ratio) across the SLM area used for encoding weights, highlighting non-uniformity. (c) Measured average optical crosstalk distribution, showing light intensity spread from intended weight locations onto neighboring areas on the detector plane. Figure adapted from [118].

Variations in the liquid crystal layer, non-uniformities in driving electronics, and imperfections in polarizer alignment are the underlying causes. Consequently, a given grayscale value can produce different transmission levels depending on its spatial location, introducing errors in the realized weight matrix W^{ij}. These inaccuracies accumulate across layers, reducing optical MVM precision and degrading system performance. Moreover, the finite extinction ratio prevents implementing true zero weights, further limiting accuracy. To account for these effects, we experimentally characterize the SLM's transmission response, measuring both the spatial variation in achievable transmission and the extinction ratio (defined as the ratio of maximum to minimum transmission), as shown in Fig. 27b. This spatial calibration is incorporated during the mapping of trained weights to physical amplitude masks, as described in Sec. 5.4, to compensate for device-specific imperfections.

Finally, we also calibrate the system for optical crosstalk within the free-space MVM. This error arises from unintended illumination of a photodiode by light originating from an optical channel other than its designated one. Improperly selected length scales, including weight size and spacing, and separation between multiple layers, are primary causes of crosstalk. Diffraction due to the finite size of weights encoded on the SLM is an additional contributing factor, an effect that becomes more significant in scaled-up designs with smaller feature sizes.

Crosstalk's effect is a corrupted signal at each photodiode, which receives a sum of its intended signal and unwanted contributions from neighboring channels. This reduces computed matrix-vector product accuracy and degrades overall network performance, especially as errors accumulate across layers. We characterize this effect by measuring the average spatial distribution of light intensity that spills over from designated weight locations onto adjacent detector areas, as shown in Fig. 27c. Optimization strategies to minimize crosstalk include careful selection of geometric parameters, such as LED and PD spacing, component sizes, and propagation distances (d_1, d_2), guided by ray-tracing simulations. This characterization allows incorporating residual crosstalk into the system model, improving inference accuracy.

The detailed characterization and calibration procedures described in this section—addressing nonlinearities, timing constraints, spatial variations, and crosstalk—enable constructing a comprehensive model that accurately describes the OENN's physical behavior. This hardware-informed model is essential for transferring trained weights or for training them directly using a more accurate computational model of the system. The following section describes the methods used for this.

5.4 WEIGHT IMPLEMENTATION STRATEGIES INCORPORATING HARDWARE CALIBRATION

We comprehensively characterized the physical system in Sec. 5.3 to model deviations from ideal behavior observed in the hardware. The hardware model accounts for LED nonlinearities (Fig. 24), electronic circuit behavior including offsets and gain mismatches (Fig. 25), SLM spatial variations and finite extinction (Fig. 27b), and optical crosstalk (Fig. 27c).

Neural networks can be implemented on the OENN hardware using two strategies: (1) transferring weights from a pre-trained network and applying calibration-based corrections, or (2) training the network directly using a hardware-aware model. The following sections describe both approaches.

5.4.1 *Calibrated Direct Weight Transfer*

First, a fully connected neural network is trained digitally with input, hidden layer, and output sizes matching the experimental setup. The optimized digital weights, denoted $\mathbf{W}_{\mathrm{dig}}$, are then mapped onto the SLM-based amplitude mask by compensating for measured hardware non-idealities after training. This approach enables efficient inference by deploying pre-trained networks onto the physical hardware. The implementation consists of two main steps:

1. **Hardware-Constrained Digital Training:** Initial training occurs within a standard PyTorch training environment on a computer but includes constraints derived from hardware characterization. Specifically, weights W_{dig}^{ij} are clamped during optimization to match the SLM's operational range, defined by the measured minimum transmission W_{min} and average maximum transmission $\langle W_{\mathrm{max}} \rangle$ (Fig. 27b):

2. $$W_{\mathrm{dig}}^{ij} \leftarrow \mathrm{clamp}(W_{\mathrm{dig}}^{ij}, W_{\mathrm{min}}, \langle W_{\mathrm{max}} \rangle) \tag{59}$$

   In addition, hidden layer simulation incorporates the average electronic offset $\langle c_3 \rangle$, as characterized in Fig. 25. This offset is added before the ReLU activation in the digital model:

$$y_j = f\left(\sum_i W_{dig}^{ji} x_i + \langle c_3 \rangle\right) \tag{60}$$

where y_j represents the j-th hidden neuron's activation. Gaussian noise is introduced for every forward pass in the training to make the network robust to experimental inaccuracies. While trained weights reflect these baseline hardware constraints, additional corrections are still needed to compensate for effects not explicitly included in training.

3. **Calibration-Based Normalization and Loading:** After training, calibration data refines weights before hardware deployment. This step corrects for residual non-idealities not accounted for during training. Once the calibration is complete, the response of the system to an input matches the ideal case.

   - **Positioning of weights:** Position of the weights is then calibrated using an iterative process where the weights are shifted from the idealized position one pixel at a time to maximize the response of the neuron on the subsequent layer.
   - **LED Input Linearization:** Each LED's nonlinear voltage-to-intensity response (Fig. 24) is inverted to determine the required input voltage $V_{in,i}$ for a desired intensity x_i, ensuring linear optical output during operation.
   - **Circuit Response Compensation:** Fitted parameters (c_{1j}, c_{2j}, c_{3j}) for each neuron's difference-ReLU circuit (Fig. 25) adjust incoming weights W^{ij}. This compensates for gain imbalance between positive (P_j^+) and negative (P_j^-) inputs and corrects for local offsets.
   - **SLM Weight Normalization:** Each digital weight W_{dig}^{ij} is adjusted according to the measured maximum achievable transmission $M^{ij} = W_{max}^{ij}$ and minimum W_{min} at that location (Fig. 27b). These corrections compute the control signal V_{SLM}^{ij}, ensuring the realized optical weight W_{opt}^{ij} lies within the device's physical limits $[W_{min}, M^{ij}]$.

   This post-training correction ensures that the physical system closely reproduces the optimized digital model's behavior.

5.4.2 *Hardware-Aware Digital Training*

An alternative approach incorporates physical hardware characteristics directly into the training process. Instead of applying corrections post-training, this method builds a model of hardware non-idealities into the simulation environment and trains the network accordingly. The resulting weights are inherently robust to hardware imperfections.

This is achieved by modifying the forward pass during training to reflect hardware-calibrated behavior:

- **Simulated Circuit Behavior:** The empirically fitted difference-ReLU model (Fig. 25) replaces the ideal hidden layer function. This can include neuron-specific parameters (c_{1j}, c_{2j}, c_{3j}) if available, applied as $y_j^{sim} = \text{ReLU}(c_{1j}P_j^+ - c_{2j}P_j^- + c_{3j})$.
- **Simulated SLM Normalization:** Digital weights $\mathbf{W}_{dig}$ are quantized to the SLM's native bit depth and clamped based on the measured range $[W_{min}, W_{max}^{ij}]$ (Fig. 27b).
- **Simulated Optical Crosstalk:** Crosstalk between channels (Fig. 27c) is simulated using a calibrated crosstalk matrix $\mathbf{C}$. The optical MVM is modified as $\mathbf{P} = \mathbf{C}(\mathbf{W}_{dig}\mathbf{x})$ before further processing.
- **Other Simulated Effects:** Additional effects, such as the LED nonlinear response (Fig. 24) and system noise, were also integrated into the simulation.

The total response of the circuit and optical weights are collectively combined in the calibration curves for the individual weights for the final experiments. in the intermediate exp. this model was used. optical crosstalk was difficult to fit and the system performed well with only adding crosstalk noise in the forward mode

The circuit response as well as the response from the optical implementation of the weights are collectively combined in the calibration curves. By training under these realistic conditions, the network learns weights that perform reliably when deployed on actual hardware.

Both approaches aim to mitigate the impact of hardware non-idealities on system performance. Calibrated transfer enables high-performance inference using pre-trained networks by applying post-hoc corrections based on detailed hardware characterization. Hardware-aware training, on the other hand, produces more robust weights by integrating hardware models directly into the training loop. These weights can generalize better across similar systems but typically underperform compared to the calibrated transfer approach. In our experiments, the effect of optical crosstalk was difficult to fit in this case, so an additional noise source representing optical cross-talk was introduced. An alternative method—using real hardware outputs instead of simulated forward passes during training—offers greater accuracy but at the cost of experimental complexity. Although we do not explore this method in this work, it represents a promising direction for future research.

Given its performance and flexibility, we adopt the calibrated transfer method for the remainder of this work. To evaluate it, we apply the method to a benchmark classification task using the MNIST dataset. Figure 28 shows an example of a trained and hardware-adapted amplitude mask.

5.5 RESULTS

To evaluate OENN hardware performance and the effectiveness of previously discussed weight implementation strategies, we tested the system on standard machine learning classification tasks. We begin with the widely recognized MNIST dataset of handwritten digits, followed by a more challenging nonlinear spiral dataset to specifically test the contribution of nonlinearity introduced in this work.

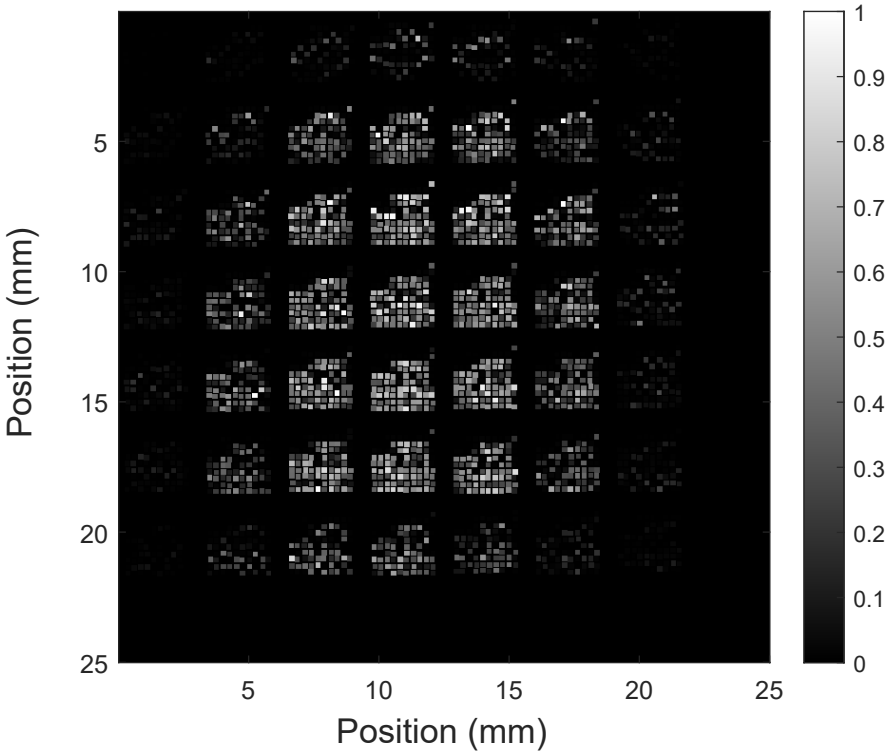


Figure 28: Example amplitude mask encoding weights for a network trained on the MNIST digit classification dataset. Individual weights have been shifted to account for exact LED and PD positions. Figure reproduced from [118].

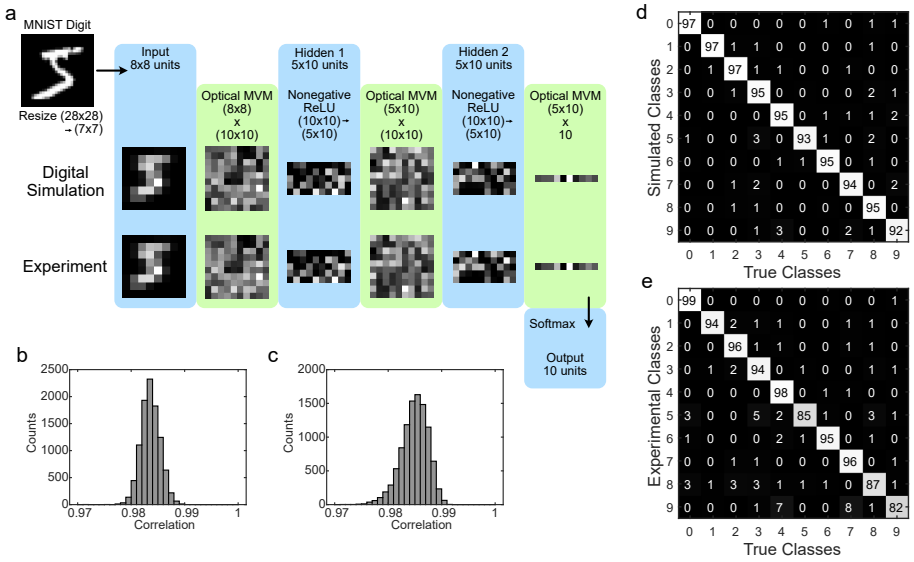


Figure 29: **MNIST digit classification with the three-layer OENN.** (a) Example propagation trace showing experimental (bottom row) versus simulated (top row) neuron activations for an input digit '4' through the input layer, first optical MVM, first hidden layer (ReLU output), second optical MVM, second hidden layer (ReLU output), and final optical MVM (output layer). (b, c) Correlation plots comparing experimental and simulated neuron activations in Hidden Layer 1 and Hidden Layer 2, respectively, across multiple MNIST test digits. (d, e) Confusion matrices showing classification performance for digital simulation and experimental hardware, respectively. Experimental accuracy reaches 92.3%, closely matching the simulated 95.4%. Figure reproduced from [118].

5.5.1 *MNIST Handwritten Digit Classification*

The MNIST dataset provides a standard benchmark for image classification and consists of 60000 examples of handwritten digits between 0 and 9 [5]. For compatibility with our OENN prototype dimensions (8×8 input layer), original 28×28 pixel images were downscaled and padded to an 8×8 shape, forming a 64-element input vector for each digit image. The network, configured with two hidden layers (50 neurons each) and a 10-neuron output layer corresponding to digit classes (0–9), was trained on a computer as a standard fully connected network. Trained weights were then implemented on the OENN hardware using the calibrated direct weight transfer approach.

Figure 29 shows results obtained on this classification task using the OENN system. Panel (a) displays an example of a downscaled MNIST digit ('4') propagating through successive OENN layers. Experimentally measured activations at each stage are compared to corresponding values from a digital simulation of the trained network using the same input. This close match is not limited to the digit '4' but holds for a wide range of input images. Figure 30 shows additional examples.

A high degree of correlation is observed between experimental neuron activations and simulated values in hidden layers, as shown in Fig. 29b and Fig. 29c for Hidden Layer 1 and Hidden Layer 2, respectively. This strong correlation indi-

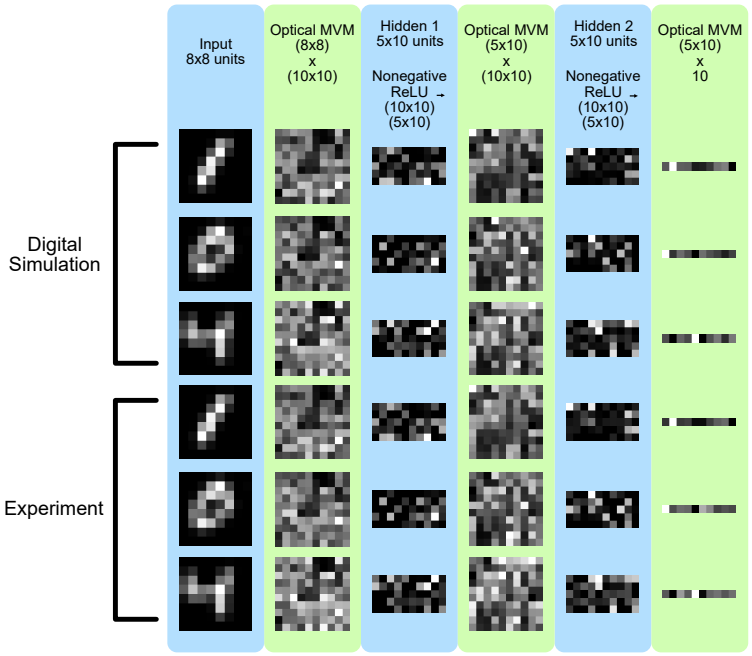


Figure 30: **Additional MNIST digit propagation examples.** Visual comparison between digital simulation (top rows within pairs) and experimental measurements (bottom rows within pairs) for different input digits propagating through network layers, complementing Fig. 29a. Figure reproduced from [118].

cates that signal transformations—optical MVM and electronic nonlinearity—are performed as intended and that the calibration procedure described in Sec. 5.3 is effective. This enables the OENN hardware to reliably implement deeper networks with multiple cascaded layers.

As operations performed in the OENN are analog, deviations from the ideal digital model invariably arise. Figure 31 quantifies this deviation by comparing normalized experimental activations with activations in the digital network across all network stages. The analysis reveals relative standard deviations ranging from approximately 0.05 to 0.19 across layers. While these errors reflect inherent noise and imperfections in the analog system, their bounded nature shows that error accumulation does not significantly degrade performance at this network depth. The system maintains sufficient computational fidelity despite these analog effects.

This operational fidelity directly translates into strong classification performance, as shown by the confusion matrices in Fig. 29d and e. The OENN hardware achieves a 92.3% test accuracy, closely matching the 95.4% obtained from the digital simulation of the same network. This result validates the OENN architecture and shows that, when combined with calibration and weight transfer procedures, it effectively implements pre-trained digital models. Moreover, both experimental and simulated accuracies significantly exceed the 82.4% achieved by an optimized linear classifier. This clearly demonstrates the contribution of ReLU nonlinearities: the OENN is not a linear system reducible to a single layer (See. Chap. 1) but a gen-

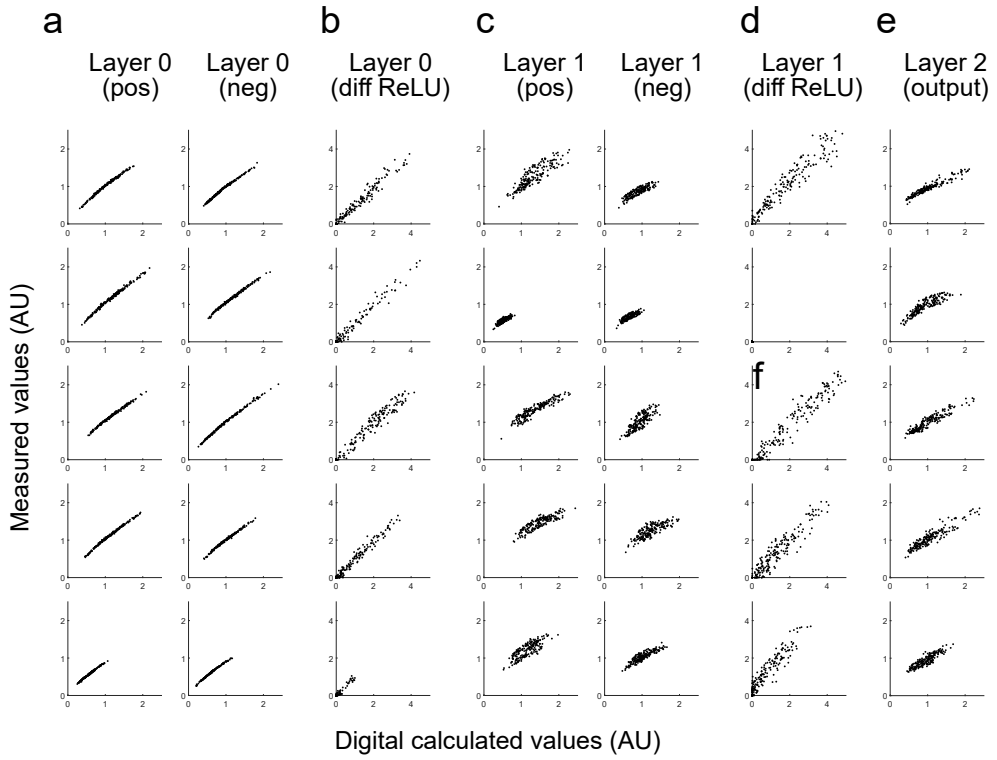


Figure 31: **Layer-wise comparison of experimental and simulated neuron activations for MNIST task.** Scatterplots show normalized experimental activation versus corresponding digital simulation values after the (a) first optical MVM (positive/negative components shown separately), (b) first differential ReLU, (c) second optical MVM, (d) second differential ReLU, and (e) third optical MVM (output layer). The relative standard deviation (σ_{rel}) of the difference between experimental and simulated values, normalized by the standard deviation of simulated activations, is reported for each stage: (a) $\sigma_{rel} = 0.048$, (b) $\sigma_{rel} = 0.152$, (c) $\sigma_{rel} = 0.145$, (d) $\sigma_{rel} = 0.191$, and (e) $\sigma_{rel} = 0.154$. Figure reproduced from [118].

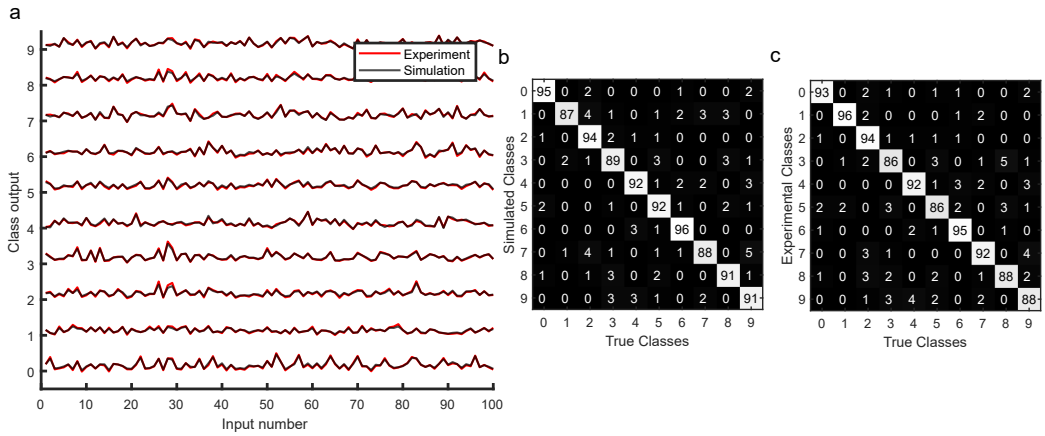


Figure 32: **MNIST classification with simultaneous multilayer operation.** Performance evaluation when all OENN layers operate continuously without intermediate digitization. (a) Comparison between simulated and experimental output layer activations for multiple test digits. (b) Confusion matrix for the digital simulation under these conditions (Test Accuracy: 91.2%). (c) Confusion matrix for the experimental hardware with all three OENN layers implemented simultaneously (Test Accuracy: 91.1%). Figure reproduced from [118].

uinely nonlinear multilayer network capable of solving complex tasks like MNIST classification.

While we have theoretically shown that deeper multilayer networks can be implemented on the OENN, results so far were obtained using a test-bench setup with a source layer, a single OENN layer, and a detection layer. Sequential readout and emission emulated multilayer behavior. We then extended the experiment to construct a three-layer MOENN device, enabling signals to propagate continuously through all optical and electronic stages without intermediate computer readout. This mode reflects the intended use as a true hardware accelerator designed to minimize data movement. Results under these conditions (Fig. 32) remain consistent, with experimental accuracy at 91.1%, closely matching the simulated 91.2%. This successful demonstration of continuous, fully analog multilayer computation is a key outcome, reinforcing the architecture's potential to reduce read-in/read-out overhead compared to single-layer optical accelerators.

5.5.2 *Nonlinear Spiral Classification*

While MNIST demonstrates accuracy on a standard task, the four-class spiral dataset provides a more stringent test of the OENN's nonlinear processing capabilities. This dataset (Fig. 33a) consists of intertwined classes that linear decision boundaries cannot separate, making it inherently difficult for simpler models.

The results in Fig. 33 highlight the OENN's capability in this nonlinear regime. A linear classifier fails on this task, achieving only 30.1% accuracy (Fig. 33c). In contrast, the OENN, using the same multilayer architecture with ReLU nonlinearities, achieves an 86.0% experimental accuracy (Fig. 33e). This large improvement shows

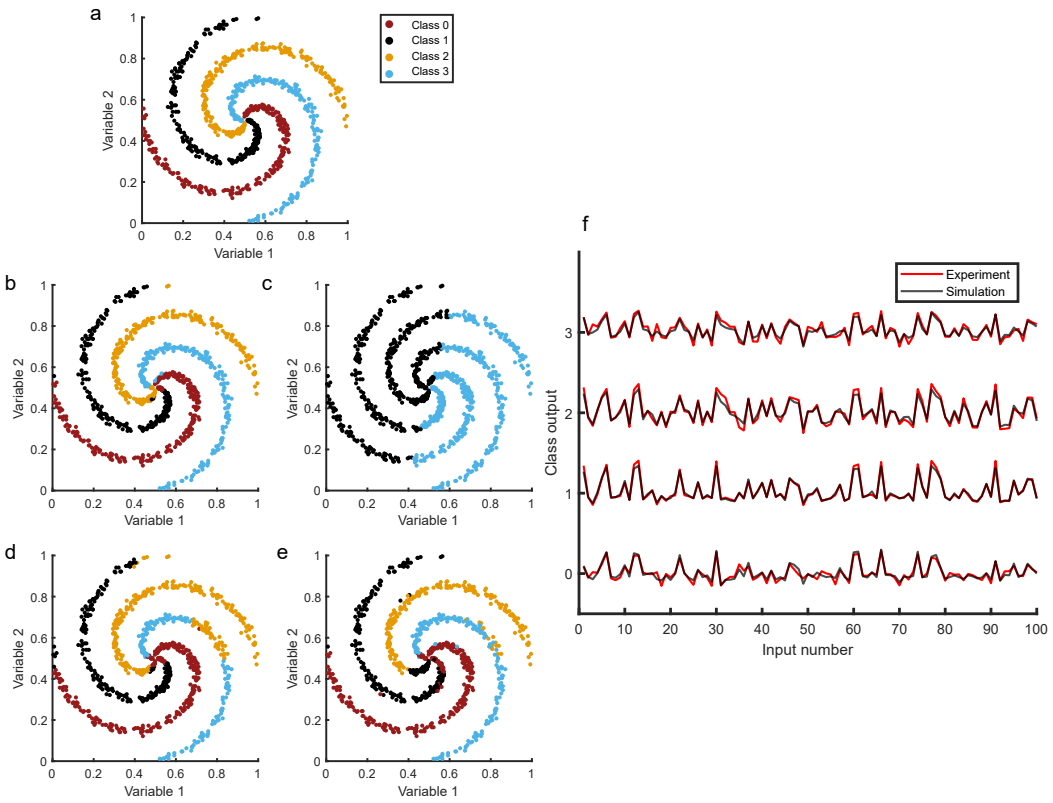


Figure 33: **Classification of the nonlinear four-class spiral dataset.** (a) Visualization of the dataset with four intertwined classes in a 2D input space. (b) Decision boundaries learned by a digitally trained network with ideal parameters (Accuracy: 96.1%). (c) Decision boundary of the best linear classifier (Accuracy: 30.1%). (d) Classification performance of a digital simulation using weights constrained by hardware limits (Accuracy: 87.8%). (e) Experimentally measured classification performance of the OENN hardware (Accuracy: 86.0%). (f) Direct comparison of simulated versus experimental output values for the four classes across multiple input samples. Figure reproduced from [118].

that the hardware successfully implements nonlinear functions required to learn complex decision boundaries. The experimental accuracy also aligns closely with simulations incorporating hardware constraints, which reach 87.8% (Fig. 33d). The close match between experimental and simulated outputs across the four classes (Fig. 33f) further supports this result. Successfully classifying the spiral dataset demonstrates that the multilayer OENN, with integrated optical MVMs and electronic ReLU nonlinearities, is a viable hardware platform for solving nonlinear problems typical in real-world machine learning applications.

5.6 POWER CONSUMPTION IN THE PROTOTYPE CIRCUIT

A critical aspect of evaluating any electronic implementation is understanding its power consumption characteristics. In the OENN system, intermediate layers—responsible for performing the core difference-ReLU operation—are primary contributors to electrical power draw. For this analysis, we exclude contributions from the SLM and detection circuit, as their power consumption is negligible compared to that of the intermediate circuit. Moreover, SLM power draw is largely independent of the input signal and can be eliminated entirely by replacing it with passive dithered masks as described earlier.

Each intermediate board consists of an array of independent neuron circuits, each implementing photodetection, differencing, amplification, and light emission.

Table 1: Component list for implementing one intermediate neuron circuit in the experimental prototype.

Component Type	Part Number / Value	Quantity
Operational Amplifier	MCP6V66T-E/OT	1
	LM358	1
MOSFET	BSS138PW	1
Photodiode	SFH2704	2
Light Emitting Diode	150040GS73220	1
Resistors	12.5 Ω	1
	10 kΩ	3
	330 kΩ	1

Table 1 lists the specific components used in the prototype's intermediate neuron circuit. Key active components include SFH2704 photodiodes for detection, MCP6V66T and LM358 operational amplifiers for transimpedance and differencing stages, a BSS138PW MOSFET, and a Wurth 150040GS73220 green LED for light emission. These parts were selected for availability and functional compatibility, though not optimized for performance or efficiency.

Power consumption was experimentally measured on a representative neuron circuit built on a matrix breadboard. The total power draw (P_{Total}) was broken down into contributions from photodiode biasing (P_{PD}), the first op-amp stage

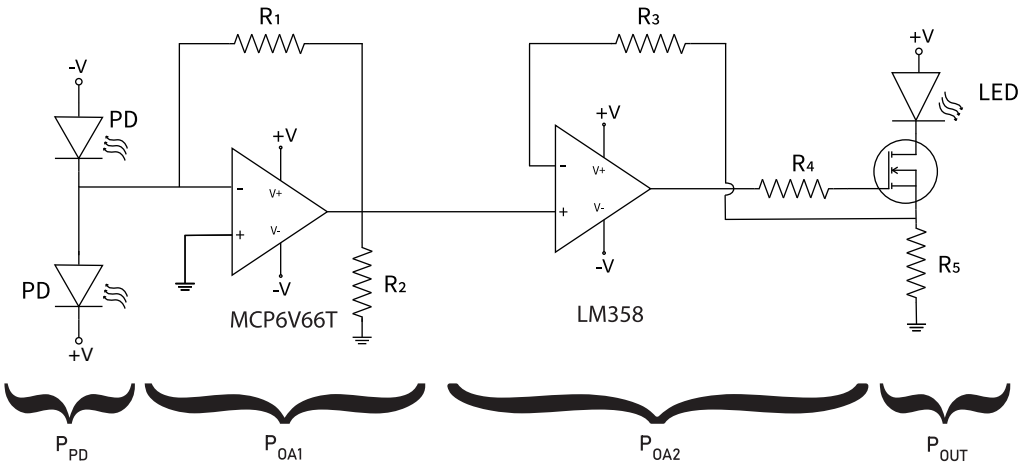


Figure 34: **Power draw components in the prototype intermediate circuit.** Schematic of the intermediate neuron circuit used in the experimental prototype, illustrating the conceptual breakdown for power consumption analysis. Considered contributions include: P_{PD} (photodiode biasing/dark current), P_{OA_1} (first op-amp stage), P_{OA_2} (second op-amp stage), and P_{OUT} (output stage driving the LED). Figure reproduced from [118].

(P_{OA_1}), the second op-amp stage (P_{OA_2}), and the output stage (P_{OUT}) that drives the LED, as shown in Fig. 34. Power consumed by photodiode dark current (P_{PD}) is negligible (in the nanowatt range).

A two-step measurement method, depicted in Fig. 35, estimated power for the two op-amp stages. In each step, one stage was selectively powered, and microammeters measured current at supply terminals ($I_{OA_1\pm}$, $I_{OA_2\pm}$), assuming negligible current into op-amp inputs. Output stage power (P_{OUT}) was calculated based on current through load resistor R_5. Table 2 shows a representative power breakdown under a high illumination condition (200 mW/cm^2). Considering the 32 active neuron circuits in the prototype's intermediate layer (excluding input drivers and output detectors), the total power consumption is estimated as $P_{\text{Total, expt}} = 32 \times 4.6 \text{ mW} = 147 \text{ mW}$. Importantly, this excludes power consumption of non-fundamental peripheral components used for experimental convenience, such as the SLM (4W) and DAC/ADC modules, which more efficient alternatives could replace in a dedicated system.

Table 2: Compilation of power drawn by different stages in the prototype intermediate neuron circuit for an illumination intensity of 200 mW/cm^2.

Stage	Power Draw (mW)
P_{PD} (Photodiode Bias/Dark Current)	400×10^{-9}
P_{OA_1} (First Op-Amp Stage)	0.2653
P_{OA_2} (Second Op-Amp Stage)	7.0853
P_{OUT} (Output Stage)	1.6

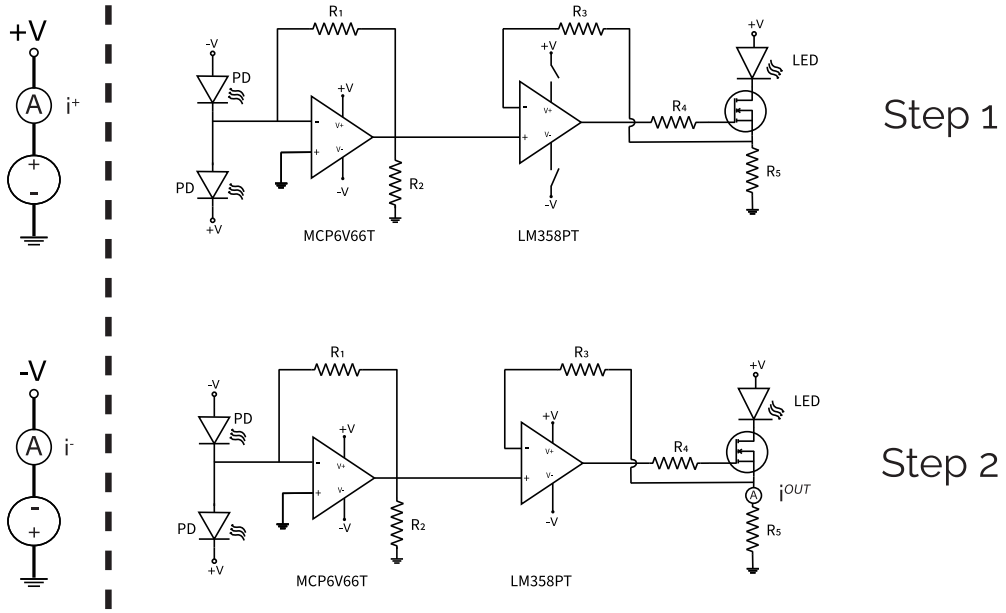


Figure 35: **Stages of measurement of prototype circuit power draw.** Schematic illustrating the two-step experimental method used to estimate power drawn by each operational amplifier stage (MCP6V66T/LM358 based) in the prototype's intermediate neuron circuit. Current measurements at supply terminals in each step allow calculating power consumed by individual stages. Figure reproduced from [118].

To estimate total electrical power consumed by the full intermediate layer, per-neuron values in Tab. 2 are scaled by the number of neurons on the prototype board. With 50 neuron circuits, total power consumption at an illumination intensity of 200 mW/cm^2 is approximately:

$$P_{\text{total}}^{\text{board}} = 50 \times (0.2653 + 7.0853 + 1.6)\,\text{mW} \approx 495\,\text{mW} \tag{61}$$

Note that P_{PD} is negligible and omitted from this estimate.

Circuit power consumption also varies with incident optical power, as this affects photocurrents and signal levels handled by amplifiers and the output stage. To capture this dependency, we characterized a single neuron circuit's power draw across a range of illumination levels using a calibrated light source. Table 3 summarizes these results, providing a baseline reference for evaluating the energy efficiency of a scaled design discussed in Sec. 6.5.2.

Table 3: Compilation of measured power drawn by a prototype intermediate neuron circuit at typical experimental illumination levels.

Net Optical Input (mW)	Input Intensity (mW/cm^2)	Circuit Power (mW)
0.59	39	5.3
0.26	17	4.9
0.14	9	4.6
0.03	2	4.4
0.00	0	4.1

Successful demonstration of classification tasks on both MNIST and spiral datasets validates the multilayer OENN prototype's functionality, demonstrating integrated optical MVM and electronic nonlinear operations within a single device. With this proof-of-concept in place, the next key question is whether this architecture can scale effectively to support larger, more complex neural networks. Additionally, the prototype circuit was not optimized for power consumption and therefore cannot yet compete with conventional or emerging hardware platforms in energy efficiency. The following chapter explores OENN approach scalability, evaluating projected performance improvements and addressing physical and architectural challenges of increasing network size and operational speed.

6

SCALING THE MULTILAYER OENN ARCHITECTURE

6.1 INTRODUCTION

The successful demonstration of the multilayer Optoelectronic Neural Network (OENN) prototype confirmed the core principle that incoherent light can perform optical matrix-vector multiplication for implementing neural networks. We additionally demonstrate the novel implementation of electronic non-linear activation. Chapter 5 details the experimental setup, design challenges, and results. However, translating this proof-of-concept into a practical and potentially competitive computing platform requires a more rigorous study of its scalability, in terms of increasing the number of neurons per layer, enhancing spatial operation density, achieving higher operational speed, and improving energy efficiency. Simply replicating the prototype design at larger physical scales or higher frequencies is insufficient. This limitation arises because of the increased physical footprint and higher implementation cost. Conversely, scaling down the device introduces significant diffraction effects that impact performance.

In this chapter, we analyze the scaling potential of our optoelectronic approach for building a cost-competitive, energy-efficient, and scalable hardware platform. Our analysis relies on simulations, validated with experiments where necessary. We begin by examining the primary optical challenges, focusing particularly on the unavoidable role of diffraction and its impact on interconnect fidelity as component sizes decrease. We then address electronic speed considerations and conclude with projected estimates of computational throughput and energy efficiency for a high-performance scaled design.

6.2 OPTICAL SCALING: CHALLENGES AND DESIGN CONSIDERATIONS

The prototype experiment described in the previous chapter consisted of an 8×8 Light Emitting Diode (LED) array on the source board and 10×10 on the intermediate board implementing the difference Rectified Linear Unit (ReLU) operation. We aim to scale the Printed Circuit Boards (PCBs) to implement more operations simultaneously. This involves scaling the PCBs to contain larger emitter arrays (e.g., 32×32), which necessitates a significant increase in the density of optical interconnections within a comparable physical volume. We can use raytracing simulations to study the complexity of the required interconnects for the prototype versus a scaled system. Figure 36 shows this comparison and illustrates the dense pattern of optical interconnects required at scale. We can accomplish this scale-up using the same emitters and similar electronic components with the same separation. However, this would mean that the resulting device would have a very large physical

footprint and higher manufacturing costs for the board. Hence, the most practical way to accommodate this increase in density requires reducing the physical size of emitters (LEDs) and the features representing weights on the amplitude mask.

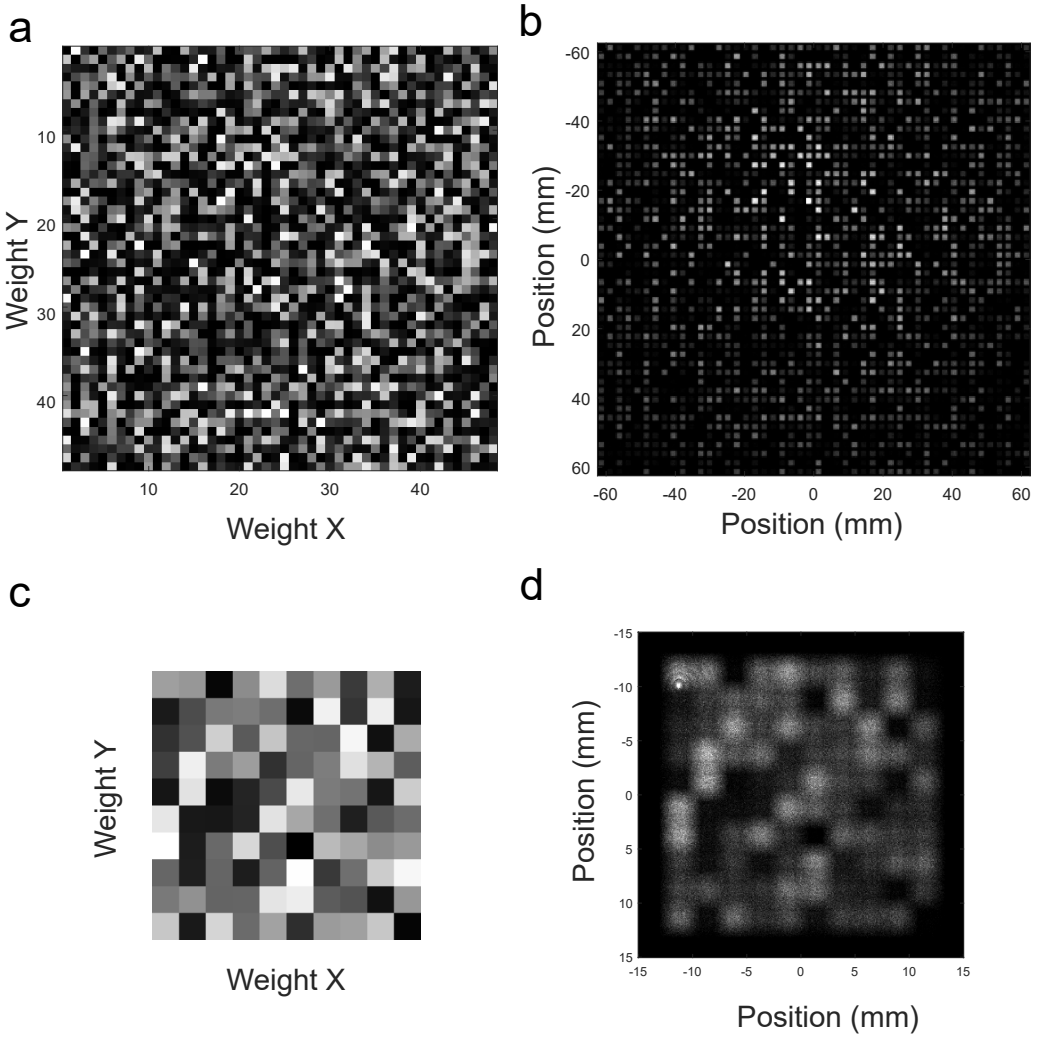


Figure 36: **Raytracing simulation comparison for prototype and scaled systems.** (a) and (b) show the target weights and ray-tracing output for the scaled-up system. In comparison, (c) and (d) display the expected and propagated weights for the smaller experimental setup. The results illustrate the increased density of optical interconnects achieved through scaling. Figure reproduced from [118].

As these feature sizes shrink, the physical phenomenon of diffraction becomes critically important, deviating strongly from the geometric optics approximations valid for the prototype (see Fig. 21). Figure 37 illustrates this fundamental effect using a Rayleigh-Sommerfeld simulation, showing how light from a point source diffracts and spreads when passing through small Gaussian apertures (10 μm spots separated by 25 μm) onto the photodiode (PD) plane over experimental propagation distances. We choose these Gaussian apertures to avoid edge diffraction effects in simulations and they are representative of a denser amplitude mask. This phenomenon becomes especially significant when multiple such apertures are close together. This diffraction poses a primary challenge to optical scaling, as it

leads to increased optical crosstalk—light intended for one detector spilling onto others—which can distort the intended intensity distribution at the PD plane.

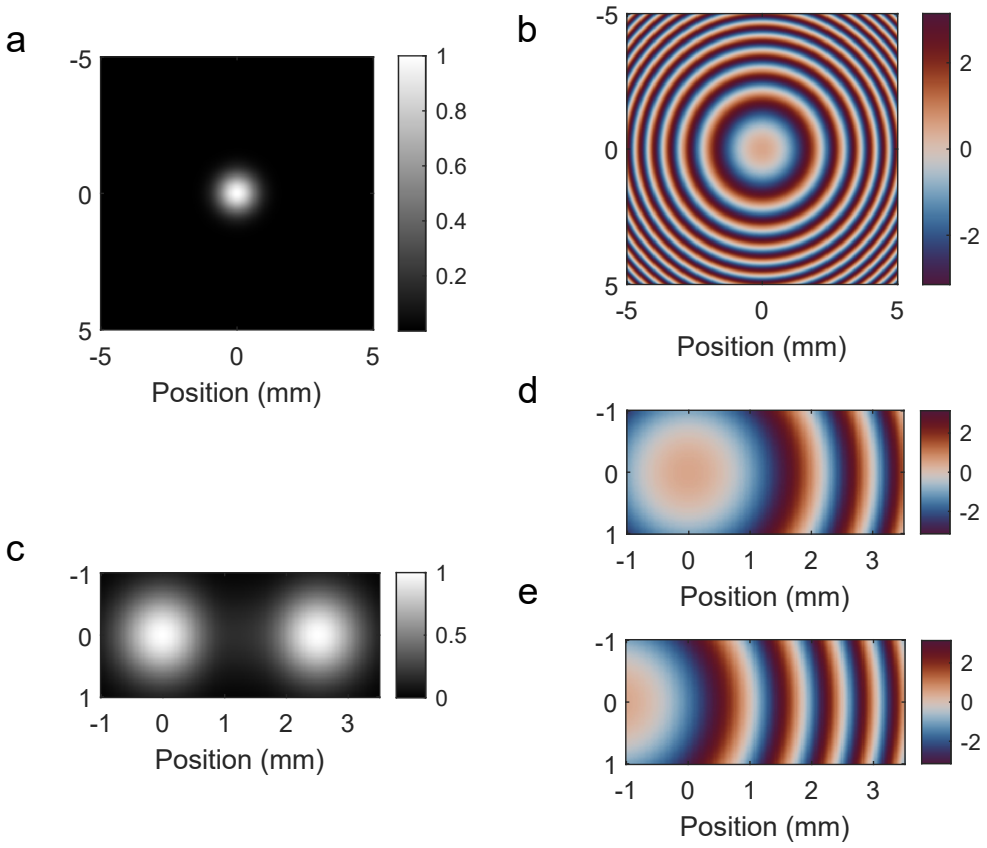


Figure 37: **Simulated diffraction through Gaussian apertures.** Visualization of light intensity patterns resulting from diffraction when a point source illuminates a single (a) and multiple closely spaced (c) Gaussian apertures, representing weight elements in the scaled system. The corresponding phase distributions for the single (b) and multiple aperture cases (d, e) are also shown. Figure reproduced from [118].

Because the scaled-down system is sensitive to diffraction, mitigating these effects requires carefully optimizing the optical system's geometry. Key parameters influencing these trade-offs include the propagation distances, notably the distance d_2 between the mask and the detector, and the lateral displacement (weight offset) of mask elements relative to the direct optical axis. We now introduce an analytical model to estimate the relationship between these parameters, signal strength, and spread.

6.2.1 *Analytical Modeling of Spread*

This analytical model describes the propagation of light from an idealized point source (representing an LED) located at $z = 0$, through a single Gaussian aperture of characteristic width σ_1 (representing a weight) positioned at (x_1', y_1') on the

mask plane at $z = d_1$. Light then propagates to the PD plane located at $z = d_1 + d_2$. To account for off-axis apertures ($x_1', y_1' \neq 0$), we rotate the coordinate system so that the effective propagation axis aligns with the source-to-aperture vector. This transformation introduces a geometric scaling factor η, defined as:

$$\eta = 1 + \frac{x_1'^2 + y_1'^2}{d_1^2} \tag{62}$$

In the rotated frame, the effective propagation distances become $\Delta z_1 = d_1\sqrt{\eta}$ and $\Delta z_2 = d_2\sqrt{\eta}$. The Gaussian aperture function $T(x', y')$ is anisotropically scaled, appearing narrower along the offset direction (assumed to be x' here):

$$T(x', y') = \exp[-(\eta(x' - x_1')^2 + (y' - y_1')^2)/\sigma_1^2] \tag{63}$$

Using the Fresnel diffraction approximation, the complex amplitude field from the point source just before the aperture is $U_{in}(x', y') \propto \exp[\frac{ik}{2\Delta z_1}(x'^2 + y'^2)]$. We compute the field at the PD plane, $U(x, y)$, by propagating the field after the aperture, $U_{aperture} = U_{in} \times T$, over the distance Δz_2 using a diffraction integral:

$$U(x, y) \propto \iint U_{aperture}(x', y') \times \text{Propagator}(x, y; x', y')\, dx'\, dy' \tag{64}$$

The propagator accounts for the propagation distance Δz_2, and the integral separates into x' and y' components, each forming a complex Gaussian integral.

Solving these integrals reveals that the complex amplitude $U(x, y)$ at the PD plane retains a Gaussian profile. The resulting intensity distribution is characterized by $1/e^2$ widths (spreads) that differ along the directions parallel (s_{ampx}) and orthogonal (s_{ampy}) to the lateral offset vector. These are given by:

$$s_{ampx} = \sqrt{\frac{4(c_1^2 + c_{3x}^2)}{c_1^2 c_2^2}} \quad \text{and} \quad s_{ampy} = \sqrt{\frac{4(c_1^2 + c_{3y}^2)}{c_1^2 c_2^2}} \tag{65}$$

where $c_1 = k/(2\Delta z_1)$, $c_2 = k/\Delta z_2$, $k = 2\pi/\lambda$, $c_{3x} = \eta/\sigma_1^2$, and $c_{3y} = 1/\sigma_1^2$.

These analytical results, applied to the scaled-up system using parameters listed in Tab. 4, are visualized in Fig. 38. The figure illustrates key trade-offs between minimizing optical spot size (to reduce crosstalk) and maximizing collected signal strength (for high signal-to-noise ratio and efficiency), as a function of propagation geometry and lateral aperture offsets.

A central design consideration is the spatial extent—or spot spread—of light at each photodiode. Panels (a) and (b) of Fig. 38 show how this spread varies with propagation distance and lateral offset, based on Eq. (65). Spot size increases with distance, but more critically, increases sharply for larger lateral offsets. The η

scaling factor (Eq. (62)) and the elongated effective path length drive this, making crosstalk increasingly problematic at larger array sizes.

In addition to spread, maximizing the collected optical signal is vital. The analytical model provides the intensity distribution $I(x, y) = |U(x, y)|^2$ at the PD plane. Panels (c) and (d) of Fig. 38 display derived signal metrics. Panel (d) shows the useful signal strength as the integrated intensity over the finite PD area, while panel (c) shows the optical power transmitted through the Gaussian aperture. Both metrics decrease with increased propagation distance (due to divergence) and fall off sharply with greater lateral offsets, as diffraction and angular effects cause light to miss the PD area.

Thus, Fig. 38 highlights the core trade-off: while larger propagation distances and lateral offsets may ease layout constraints or enhance mapping flexibility, they reduce both signal fidelity and intensity. This analysis informs the selection of optimal geometry, such as $d_1 = 2.5\,\text{mm}$ and $d_2 = 84.2\,\text{mm}$, used in the scaled design (Tab. 4). Additional refinements can incorporate finite LED source size and spectral bandwidth for more precise modeling.

Table 4: Optical parameters used for analytical modeling and simulation of the scaled-up system.

Parameter	Symbol	Value
LED die width	w_0	$10\,\mu\text{m}$
Width ($1/e$ amplitude) of optical weight	σ_1	$25\,\mu\text{m}$
Weight separation (center-to-center)	δ_1	$75\,\mu\text{m}$
Photodiode separation (center-to-center)	δ_2	$2.5\,\text{mm}$
Distance from LED plane to weight plane	d_1	$2.5\,\text{mm}$
Distance between weight plane and PD plane	d_2	$84.2\,\text{mm}$
Photodiode width (square side assumed)	σ_2	$1.2\,\text{mm}$
Wavelength range	λ	$500 - 540\,\text{nm}$

6.3 ELECTRONIC SCALING: DESIGN FOR HIGH-SPEED OPERATION

A substantial increase in circuit operation speed is needed to complement the optical scaling strategies discussed previously and achieve a computational throughput competitive with existing technologies. The prototype system, while demonstrating the core principles, was inherently limited by the response times and slew rates of the discrete operational amplifiers chosen in the intermediate layers (as characterized in Fig. 26). Overcoming this bottleneck is crucial for realizing a high-performance OENN accelerator.

To establish the feasibility of operating the core electronic functions at significantly higher frequencies, we designed and simulated a circuit for 10 MHz operation using LTSPICE. The circuit diagram, presented in Fig. 39, retains the fundamental topology of the difference-ReLU function implemented in the prototype

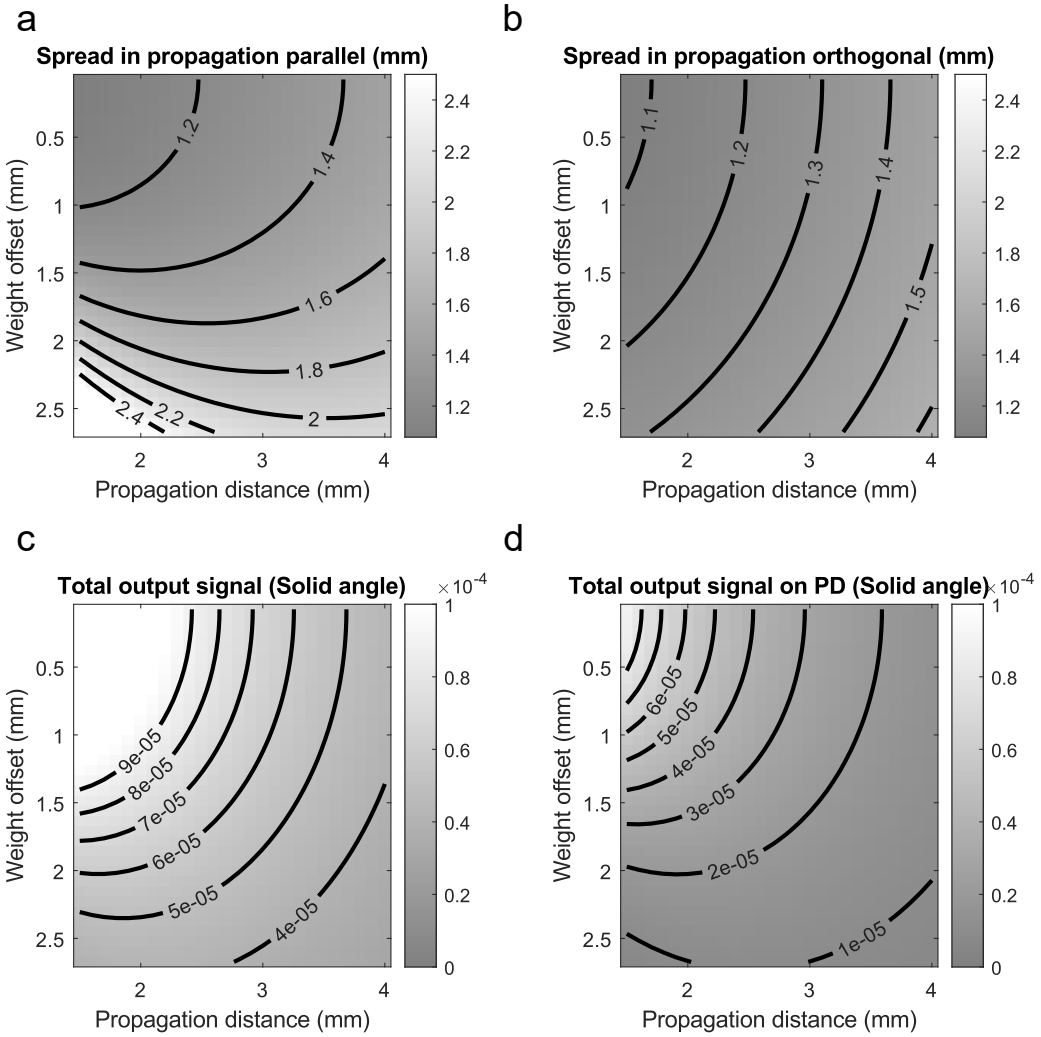


Figure 38: **Analysis of optical trade-offs with propagation distance and lateral offset.** Performance metrics as a function of the mask-to-photodiode distance d_2 and lateral weight aperture offset from the LED-photodiode axis: (a) Spot spread parallel to the offset. (b) Spot spread orthogonal to the offset. (c) Solid angle of light transmitted per unit area. (d) Solid angle collected by the target photodiode per unit area. Figure reproduced from [118].

(compare with Fig. 22c). However, it incorporates faster components, specifically using OPA818 op-amps known for higher bandwidth and slew rates compared to the MCP6V66T and LM358 used previously. The design includes appropriate resistor and capacitor values carefully selected to ensure stable and accurate operation at the target 10 MHz frequency, compensating for potential high-frequency instabilities inherent in faster amplifier designs. The output stage also uses a high-speed switching transistor (2N2369) to drive the LED (150040GS73220).

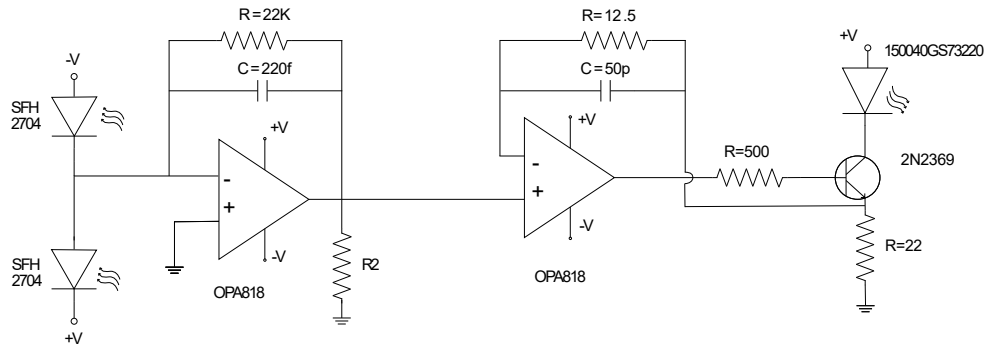
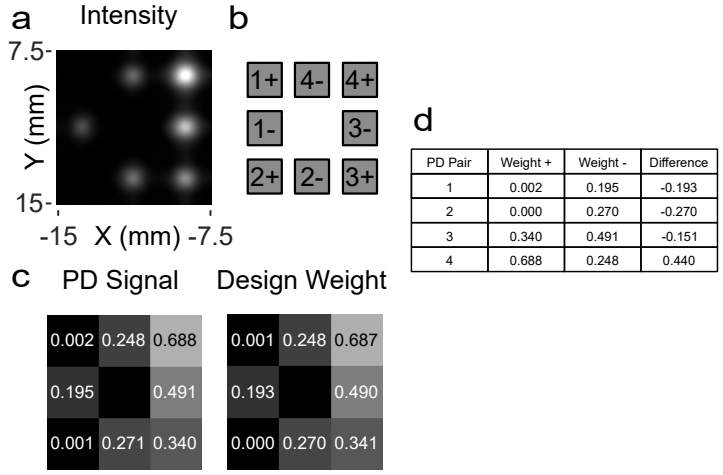


Figure 39: **Electronic circuit for high-speed operation.** Circuit diagram for a design enabling 10 MHz operation. It retains the same topology as the circuit used in the experimental setup (see Fig. 22c), with carefully selected components and compensation for potential instabilities. Figure reproduced from [118].

Designing the detector layout to use PCB space optimally is a key aspect of implementing the difference operation electronically in the scaled-up architecture. Since the optical layer performs a non-negative matrix multiplication, pairs of photodiodes are needed to represent the positive and negative components required for a signed neuron activation before the ReLU non-linearity. Figure 40 illustrates this concept for a representative 4-neuron subunit within the larger 32×32 neuron output layer, receiving input from a subsection of the 48×48 PD array. As shown in Fig. 40a, light corresponding to different weights illuminates an array of eight PDs arranged in a 3×3 grid with the central position left empty. The signals from these eight PDs are electronically routed to form the inputs for four distinct output neurons (Fig. 40b). Each neuron receives a designated positive (+) input signal (summed current from its assigned positive PDs) and a negative (-) input signal (summed current from its assigned negative PDs). The simulated signals detected by the individual PDs (Fig. 40c, left) closely correspond to the target design weights encoded in the optical mask (Figure 40c, right), demonstrating the optical fidelity. The electronic circuit then performs the subtraction, yielding the final effective signed weights for each neuron (Fig. 40d) before the inherent rectification by the LED driver.

While the SPICE simulation using discrete components validates the circuit's functionality and potential for high-speed operation, realizing a practical, large-scale OENN with optimal performance, particularly in terms of power efficiency and physical footprint, requires moving towards custom integrated circuits.



PD Pair	Weight +	Weight -	Difference
1	0.002	0.195	-0.193
2	0.000	0.270	-0.270
3	0.340	0.491	-0.151
4	0.688	0.248	0.440

Figure 40: **Photodiode grouping for difference operation in the scaled model.** (a) Simulated intensity pattern on a 3×3 photodiode (PD) array subsection. (b) Conceptual mapping of these 8 PDs to provide positive (+) and negative (-) inputs for 4 distinct differential neuron pairs. (c) Comparison of simulated PD signals (left) derived from the intensity pattern in (a) versus the target design weights (right) for this subsection. (d) Table showing the final effective signed weights computed as the difference between the corresponding positive and negative PD signals from (c) for each of the 4 neurons.

Application-Specific Integrated Circuits (ASICs) offer the potential to co-integrate photodiodes and highly optimized analog processing circuitry, significantly reducing parasitic capacitances, power consumption, and physical size compared to PCB-based implementations. Figure 41 presents a conceptual schematic for such an ASIC-based neuron. This design envisions integrated photodiodes directly feeding into specialized transimpedance amplifiers (TIAs). TIAs are well-suited for converting the small photocurrents into voltages with high gain and low noise. The outputs of the TIAs corresponding to the positive and negative inputs would then be processed by a differential amplifier stage, potentially an Operational Transconductance Amplifier (OTA), which performs the subtraction and converts the resulting voltage into a current suitable for directly driving the output LED, thereby completing the difference-ReLU operation within a compact, efficient integrated unit. Such ASIC implementations are essential for leveraging the full potential of the OENN architecture in terms of energy efficiency and computational density, as projected in subsequent analyses.

6.4 SIMULATED PERFORMANCE OF A SCALED-UP OENN MODEL

Building upon the optical and electronic design considerations outlined above, we constructed a detailed simulation model of one OENN layer, scaled to $N = 1024$ neurons (32×32 array size) and operating at $f = 10$ MHz, to provide concrete performance projections. The model incorporates the optimized optical parameters (Tab. 4) and the high-speed electronic circuit design (Sec. 6.3).

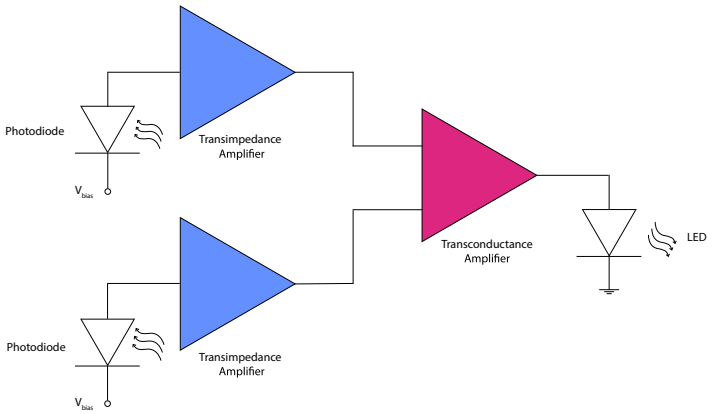


Figure 41: **Conceptual schematic for an ASIC implementation.** Proposed integrated circuit design architecture for a single difference-ReLU neuron. It features integrated photodiodes coupled to low-noise transimpedance amplifiers (TIAs) for photocurrent-to-voltage conversion. A differential transconductance amplifier (OTA) or similar stage takes the TIA outputs, performs subtraction, and provides the driving current for the output LED, inherently implementing the rectification. Figure reproduced from [118].

6.4.1 *Simulated Optical Performance*

We evaluated the optical performance of the scaled-up OENN layer, designed for $N = 1024$ neurons (32×32) operating at 10 MHz, through detailed simulations that incorporate diffractive effects, which become significant at the smaller feature sizes required for high-density operation. The simulation used optimized optical parameters derived from analytical modeling (summarized in Tab. 4) and employed a modified angular spectrum propagation method, introduced in Sec. 4.3.1, to accurately model incoherent light propagation.

Figure 42 shows the physical layout and simulated light propagation. A 32×32 LED array is positioned at $d_1 = 2.5$ mm from the amplitude mask, which contains individually addressable weights for each connection. Each LED illuminates a 3.6×3.6 mm submask region that encodes $48 \times 48 = 2304$ weights connecting it to the PD array. These weights are implemented as Gaussian amplitude profiles with a width of $\sigma_1 = 25$ µm and a center-to-center spacing of $\delta_1 = 75$ µm. The Gaussian profile helps minimize edge diffraction compared to sharp-edged apertures. Light then propagates an additional $d_2 = 84.2$ mm to reach the 48×48 PD array, resulting in an overall magnification of $M \approx 34$ (Fig. 42d). Figure 42b and Fig. 42e provide side views of the simulations, illustrating the optical path from the LED through the submask and onto the PD plane. Figure 42c,f visualizes the phase and intensity immediately after the mask.

The key output of the simulation is the intensity distribution across the 48×48 PD array at the detector plane, shown in Fig. 43a. This pattern represents the optical Matrix-Vector Multiplication (MVM) result, including the effects of diffraction and crosstalk. To estimate the effectively implemented optical weights, we integrate the intensity over each PD's active area. Figure 43b zooms in on a 3×3

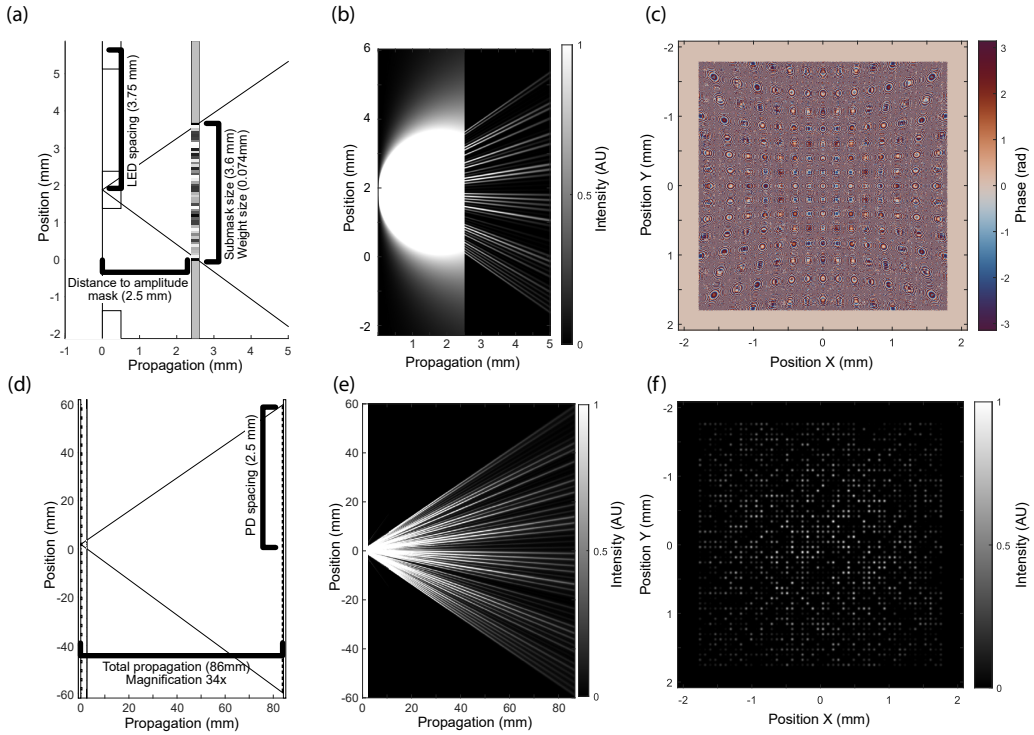


Figure 42: **Optical design and simulation for the scaled-up model.** (a, d) Schematics detailing the optical path geometry. Panel (a) shows the LED-to-mask segment ($d_1 = 2.5$ mm), indicating LED spacing (3.75 mm) and mask parameters (submask size 3.6 mm, weight size 0.074 mm spacing, 25 μm Gaussian width). Panel (d) shows the full geometry, including the mask-to-PD distance ($d_2 = 84.2$ mm), PD spacing (2.5 mm), and total magnification ($M \approx 34$). (b, e) Side-view simulations of light propagation from a single LED to the mask (b) and from the mask to the PD plane (e), calculated via modified angular spectrum method. (c, f) Simulated complex optical field immediately after the amplitude mask plane, showing the phase (c) and intensity (f) patterns representing the encoded weight information. Figure adapted from [118].

PD region, comparing the intensity pattern (left), the estimated weights (middle), and the ideal design weights (right). Despite visible blurring from diffraction, the measured intensity distribution representing the weights remains largely intact. The full spatial comparison of the estimated weights (Fig. 43d) and the original design weights (Fig. 43e) further confirms this. A scatter plot of the difference between estimated and target weights across all connections (Fig. 43c) shows that most deviations are small, indicating good overall fidelity, with minor errors due to diffraction and crosstalk.

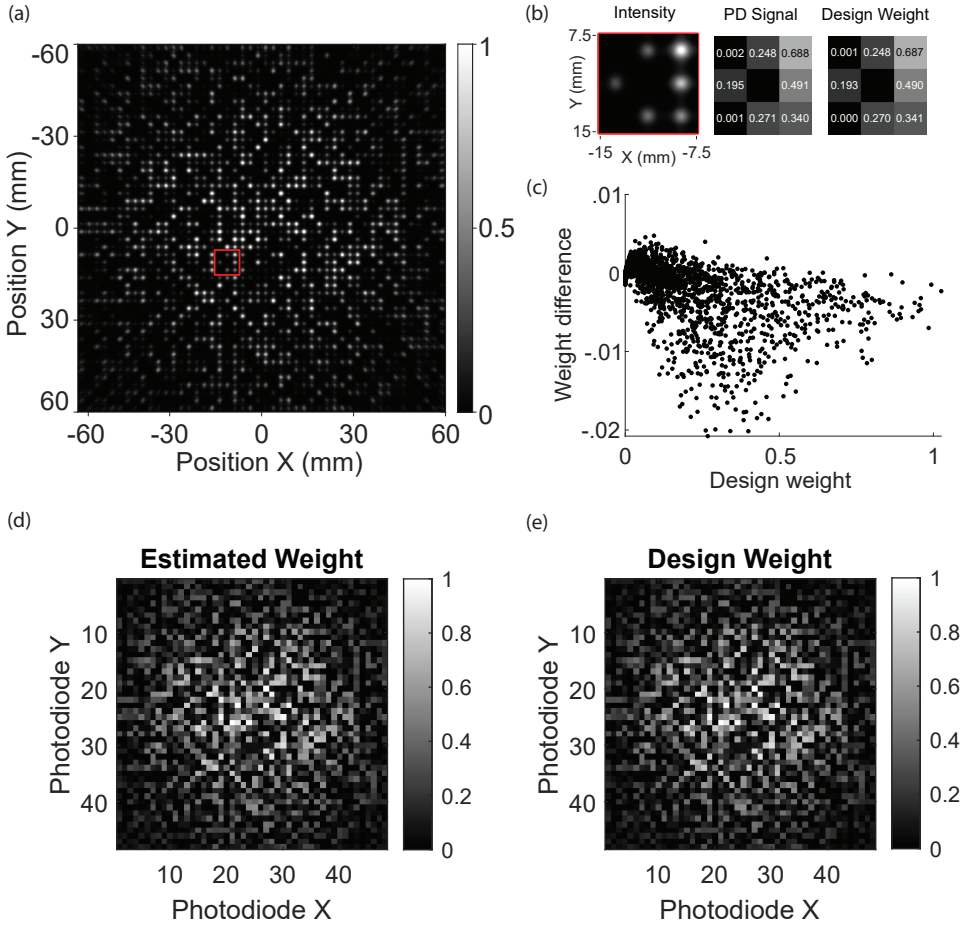


Figure 43: **Simulation of optical performance and weight accuracy for the scaled-up model.** (a) Simulated intensity distribution across the full 48×48 photodiode array, calculated using modified angular spectrum propagation. (b) Zoomed-in view of the 3×3 region highlighted in red in (a), comparing the intensity pattern (left), the estimated weights derived from integrating intensity over PD areas (middle), and the corresponding target design weights (right). (c) Scatter plot illustrating the distribution of the difference (error) between estimated and design weights for all connections in the $32 \times 32 \rightarrow 48 \times 48 \rightarrow 32 \times 32$ layer. (d, e) Full spatial maps comparing the matrix of estimated optical weights (d) with the target design weight matrix (e). Figure adapted from [118].

We further explore optical crosstalk, particularly for connections with large lateral offsets between the LED and PD. The risk of crosstalk is higher in these cases because of longer propagation paths and stronger diffraction, as the analytical

model predicts (Fig. 42). Figure 44 shows results from using Rayleigh-Sommerfeld diffraction from a point source through a 3×3 array of randomized Gaussian weights. Figures 44a and 44b show the relation between weight error and design weight, with data color-coded by the lateral offset of the photodiode. These plots reveal that both the magnitude and spread of error grow with offset distance. Figure 44c focuses on corner photodiodes, which experience the highest offsets and crosstalk, confirming this trend.

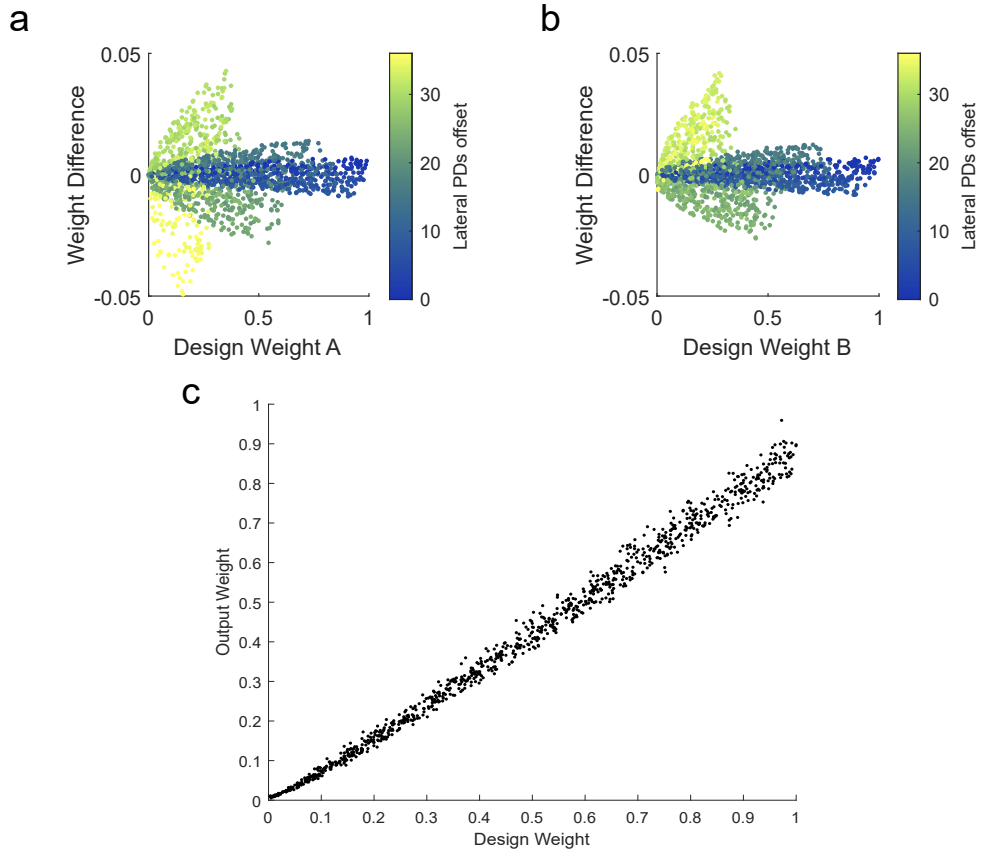


Figure 44: **Analysis of optical crosstalk dependency on position in scaled-up model simulations.** (a, b) Scatter plots showing the difference between simulated and design weights versus the design weight value for two different sets of random weights (A and B). Points are color-coded based on the lateral offset of the target photodiode from the optical axis of the point source, indicating increased error variance for larger offsets. (c) Simulated weight versus design weight specifically for connections terminating at a corner photodiode, representing conditions of maximum lateral offset and potential crosstalk. Simulations employed Rayleigh-Sommerfeld diffraction from a point source illuminating a 3×3 grid of adjacent randomized Gaussian weights. Figure adapted from [118].

Together, these simulations show that while diffraction plays a more significant role in the scaled design compared to the larger-feature-size prototype, careful co-design of the optical geometry guided by analytical and numerical modeling enables realizing a fully connected 32×32 OENN layer with acceptable MVM fidelity. The chosen design balances high interconnect density with control over diffractive and crosstalk effects.

6.4.2 *Simulated Electronic Performance*

With the optical performance of the scaled-up OENN model successfully demonstrated, we now analyze the electronic performance of the difference-ReLU circuit designed for high-speed operation. The goal is to validate whether the circuit can support the target operating frequency of 10 MHz while maintaining the required functionality and signal fidelity. We implemented the circuit in LTSPICE and performed both transient simulations and DC sweeps to assess its dynamic response and steady-state input-output characteristics.

Figure 45a presents the transient simulation results over a 1 µs time window. The top trace shows the optical input power densities ($\mu W/mm^2$) incident on the positive (Pos PD) and negative (Neg PD) photodiodes. The middle trace depicts the intermediate voltage (V_{OA}) generated by the op-amp stages after differencing and amplification. The bottom trace shows the resulting optical power (mW) emitted by the output LED, driven by the circuit's final stage. The simulation confirms that the circuit responds as expected at 10 MHz input frequencies. The LED output power tracks the rectified difference between the positive and negative inputs without significant distortion, ringing, or delay.

To confirm the accurate implementation of the ReLU non-linearity, we also simulated the circuit's steady-state behavior across a range of input intensities. Figure 45b shows this characteristic, mapping PD+ intensity (x-axis) and PD intensity (y-axis) to the LED's steady-state output power (color map). The circuit behaves as desired: the output is effectively zero when the negative input equals or exceeds the positive input. As the positive input increases beyond the negative input, the LED output rises approximately linearly with the input difference (PD_+ Intensity $-$ PD_- Intensity). This ReLU-like behavior is preserved across the input range, confirming that the circuit accurately performs the required nonlinear activation.

With both the optical and electronic performance of the scaled-up system validated through simulation, we now turn to the central question: how does this approach compare to existing hardware in terms of efficiency and performance? The following section addresses this.

6.5 PROJECTED THROUGHPUT AND ENERGY EFFICIENCY

Having validated the optical fidelity and electronic speed of the scaled-up OENN model through simulations in Sec. 6.4, we now project its key performance metrics. These projections, particularly in terms of computational throughput and energy efficiency, are critical for benchmarking the OENN against both conventional and emerging computing platforms.

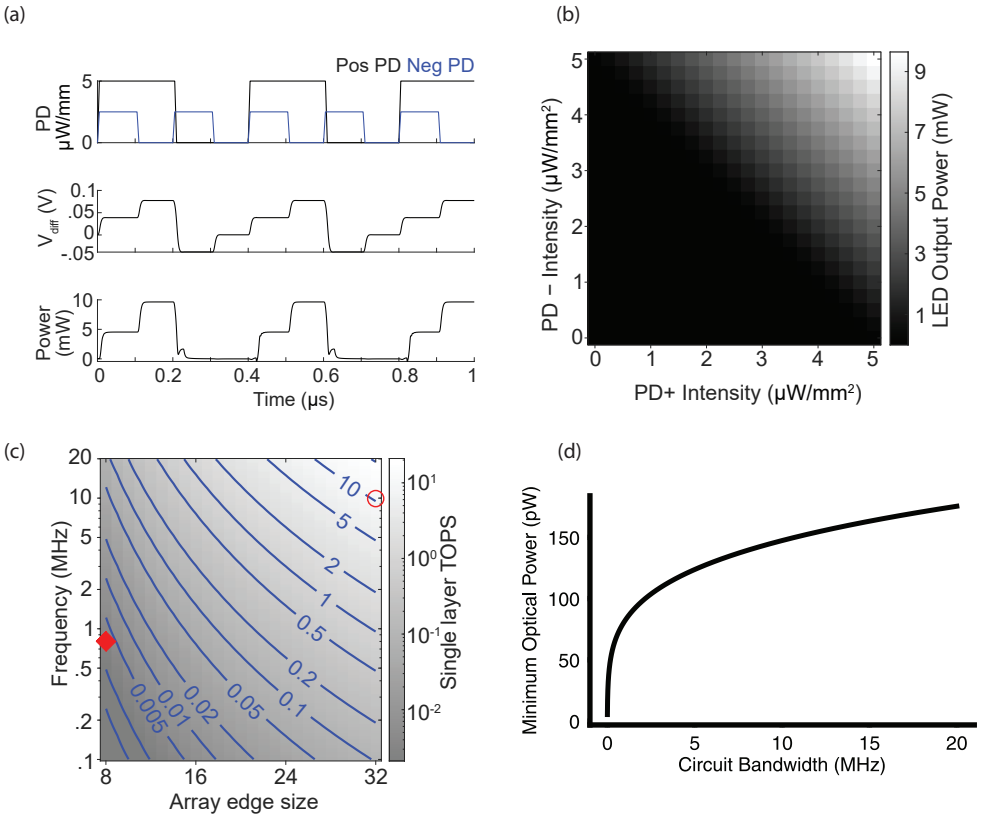


Figure 45: **Simulated electronic performance of the scaled design.** (a) SPICE simulation showing the temporal response of the difference-ReLU circuit (Figure 39) to a 10 MHz differential optical input. Traces show input power density on positive (Pos PD) and negative (Neg PD) photodiodes, intermediate voltage (V_{OA}), and output LED power over 1 µs. (b / e) Simulated steady-state input-output characteristic, mapping PD+ Intensity and PD- Intensity to LED Output Power, demonstrating the ReLU-like non-linear activation function. (c / f) Projected single-layer computational throughput (Tera-Operations Per Second, TOPS) as a function of operating frequency and neuron array edge size (n for an $n \times n$ array). The red diamond marks the experimental prototype performance (8×8, ≈ 0.8 MHz), while the red circle marks the target scaled design point (32×32, 10 MHz) validated by these simulations. (d) Minimum required optical power per detector to overcome noise, shown as a function of the electronic circuit bandwidth. Figure adapted from [118].

6.5.1 *Throughput*

Computational throughput quantifies the number of operations performed per second and is typically measured in Tera-Operations Per Second (TOPS) for high-performance systems. In the OENN architecture, each operation includes a multiply-accumulate (MAC) step during MVM, followed by a non-linear activation implemented electronically. The overall throughput depends on the number of parallel processing units and the operating frequency (f).

For the scaled model, which uses $N = 1024$ neurons arranged in a 32×32 grid and operates at $f = 10$ MHz, the per-layer throughput is given by:

$$\text{Throughput} = f \cdot (N + 1)N \tag{66}$$

Substituting the model parameters ($N = 1024$, $f = 10 \times 10^6$ Hz), we compute:

$$\text{Throughput} = (10 \times 10^6 \text{ Hz}) \cdot (1024 + 1) \cdot 1024 \approx \mathbf{10.5} \text{ TOPS} \tag{67}$$

This value represents the projected computational throughput for a single layer of the scaled OENN. Figure 45c illustrates this result, showing throughput as a function of operating frequency and array edge size n. The red circle marks the 10.5 TOPS point corresponding to the 32×32 array at 10 MHz. The figure also highlights the favorable scaling properties of the architecture: throughput increases linearly with frequency (f) and quadratically with neuron count (N^2), resulting in an overall n^4 scaling for an $n \times n$ array.

While high throughput is important, energy efficiency—measured in operations per unit power—is often a more critical metric for assessing the real-world viability of a hardware platform. The following section discusses this further.

6.5.2 *Energy Efficiency*

Beyond computational speed, energy efficiency, quantified as throughput per unit power (TOPS/W or Giga-Operations Per Second (GOPS)/W), is one of the most critical metrics for assessing the practical viability of computing hardware. In this section, we evaluate the energy efficiency of the OENN architecture. We first establish an experimental baseline based on the prototype system and then project the performance of the scaled-up model using the optimized optical design (Sec. 6.4.1) and electronic simulations (Sec. 6.4.2). The core methodology involves dividing the calculated throughput (Sec. 6.5.1) by the measured or estimated power consumption.

6.5.2.1 *Experimental Baseline Efficiency*

As described in Sec. 5.6, we measured the total power drawn by the experimental prototype system to be 147 mW. When combined with its computed throughput of approximately 1.7 GOPS (from Sec. 6.5.1), the resulting energy efficiency is:

$$\text{Efficiency}_{\text{expt}} = \frac{1.7 \times 10^9 \text{ OPS}}{0.147 \text{ W}} \approx \mathbf{11.5 \text{ GOPS/W}} \tag{68}$$

6.5.2.2 *Projected Efficiency of Scaled Model*

To estimate the potential energy efficiency of the scaled and optimized OENN architecture, we perform a bottom-up power analysis for the $N = 1024$, 10 MHz model. This estimate includes two primary contributors: the electrical power required to drive the LEDs (P_L) and the power consumed by the analog processing electronics (P_A), assuming an ASIC implementation.

The minimum optical power needed at the PD plane to overcome dominant noise sources (primarily shot noise, see Sec. 6.2.1) and achieve 8-bit signal precision determines the LED driving power (P_L). Experimental noise characterization in the prototype (Fig. 46) confirms that this level of precision is realistic, showing low error accumulation across layers. At 10 MHz, we calculate the photocurrent required for 8-bit resolution to be 84 nA, corresponding to an optical intensity of 166 nW/mm^2 on the detector. Accounting for collection efficiency, the number of LEDs, average weight values, and estimated LED wall-plug efficiency, we estimate the electrical power per LED as $P_l = 160$ µW. For the full $N = 1024$ layer, the total LED driving power is $P_L = 1024 \times P_l = 163$ mW.

Table 5: Comparison of performance requirements for an idealized scaled-up model ASIC implementation with available literature examples for Transimpedance (TIA) and Transconductance (TCA) amplifiers.

Property	Requirement	TIA [203]	TCA [204]	Combined Est.
Operating Frequency	10 MHz	10 MHz	10.88 MHz	10 MHz
Supply Voltage	No limitation	1.8 V	±400 mV	–
Amplification	73 dB	90 dBΩ	1 Ω^{-1}	90 dB
Power Consumption	–	36 µW	62 µW	134 µW

We estimated the analog electronics power (P_A) using published low-power ASIC designs for TIAs [203] and TCAs [204], selected to meet the required bandwidth (10 MHz), gain, and drive characteristics. Figure 41 shows the corresponding ASIC schematic, with performance summarized in Tab. 5. Based on these designs, the power consumption per neuron circuit—comprising two TIAs and one TCA—is $P_a = 134$ µW. For all 1024 neurons, the total analog electronics power is $P_A = 1024 \times P_a = 137$ mW.

Thus, the total projected power consumption for the scaled layer is:

$$P_{tot} = P_L + P_A = 163\ \text{mW} + 137\ \text{mW} = 300\ \text{mW}.$$

Using this and the projected throughput of 10.5 TOPS from Sec. 6.5.1, the energy efficiency of the scaled model is:

$$\text{Efficiency}_{\text{scaled}} = \frac{10.5 \times 10^{12}\ \text{OPS}}{0.300\ \text{W}} = \mathbf{35\ TOPS/W}$$

These projected efficiency figures serve as a reference point for comparing the OENN architecture to state-of-the-art digital processors and other optical computing platforms, which the next section discusses.

6.5.3 *Performance Comparison*

These projections are summarized in Tab. 6, which compares the projected throughput and energy efficiency of the OENN architecture with a range of conventional and emerging computing platforms. The results suggest that the scaled OENN could offer class-leading energy efficiency—comparable to or surpassing modern Graphics Processing Units (GPUs) and Tensor Processing Units (TPUs)—while also delivering high throughput suitable for demanding acceleration tasks.

6.6 SCALABILITY PROSPECTS AND LIMITATIONS

The analysis presented here supports the scalability of the proposed multilayer OENN architecture. Simulations confirm that optical diffraction effects in downscaled systems can be managed through geometric co-design, maintaining acceptable MVM fidelity up to the 32×32 scale investigated (Sec. 6.4.1). Additionally, we have shown that electronic operation at 10 MHz is feasible using improved circuit design (Sec. 6.4.2), with the potential for further gains through ASIC integration. The architecture exhibits favorable scaling characteristics, with throughput increasing quadratically with neuron count (N^2) (Sec. 6.5.1) and projected energy efficiency levels that are competitive with current digital hardware (Sec. 6.5.2).

A major factor contributing to this efficiency is the multilayer architecture itself (Fig. 47b,c). By retaining intermediate results within the analog domain, it reduces the need for frequent data read-in and read-out operations, which are typically energy-intensive in single-layer accelerators (Fig. 47a). This benefit, measured as the number of operations performed per read-in event, increases significantly with both the network depth and the array size (Fig. 47d).

However, there are fundamental limitations to further scaling. The current lens-free optical design may face constraints because of diffraction and crosstalk at higher densities or smaller footprints (Sec. 6.2). Overcoming these limitations will likely require alternative optical approaches, such as lenslet arrays or diffractive

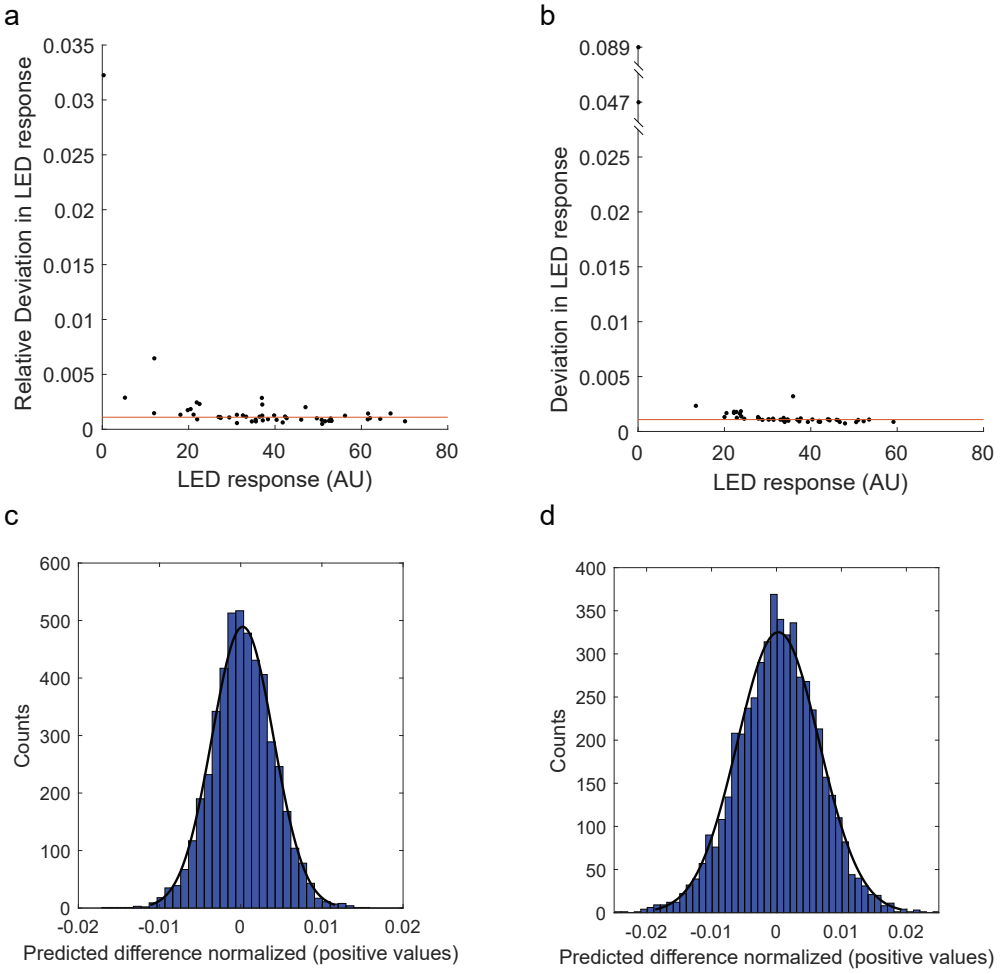


Figure 46: **Experimental characterization of noise and model fidelity across intermediate layers.** (a, b) Relative deviation in measured LED brightness after the first (a) and second (b) layers across repeated measurements or varying inputs. Red line indicates the median deviation ($\approx$ 0.1%). (c, d) Histograms showing normalized differences between measured LED outputs and predictions from a calibrated linear weighting + difference-ReLU model using randomized inputs and weights. Gaussian fits (black lines) yield standard deviations of $\sigma = 0.0038$ and $\sigma = 0.0063$ for the first and second layers, respectively. This informs the noise assumptions used in the power estimation for the scaled model. Figure adapted from [118].

Table 6: Performance comparison of our approach to conventional computing systems and other optical/opto-electronic approaches. Note: The numbers for NVIDIA B200* and RTX 4090** represent the performance for thousands of cores.

Technique	**Approach**	**Throughput** (TOPS)	**Efficiency** (Expt, TOPS/W)	**Efficiency** (Proj, TOPS/W)	**Precision** (bit)	**Reference**
NVIDIA B200	GPU	$144 \cdot 10^{3*}$	10.01	–	4	[205]
		$57 \cdot 10^{3*}$	5.03	–	8	
NVIDIA RTX 4090	GPU	660.60**	0.78	–	8	[206]
Google TPUv4	ASIC	275	1.62	–	8	[207] [208]
Photonic WD-M/PCM in-memory computing	Photonic	0.65	0.50	7.00	5	[209]
Image Intensifier	Incoherent Free Space	5.76×10^{-7}	3.03×10^{-7}	66.67	8	[190]
Photonic Conv. Accelerator	Photonic	0.48	1.26	–	8	[210]
Free Space OENN	Incoherent Free Space	1.6×10^{-3}	11.45×10^{-3}	35.09	8	This Work

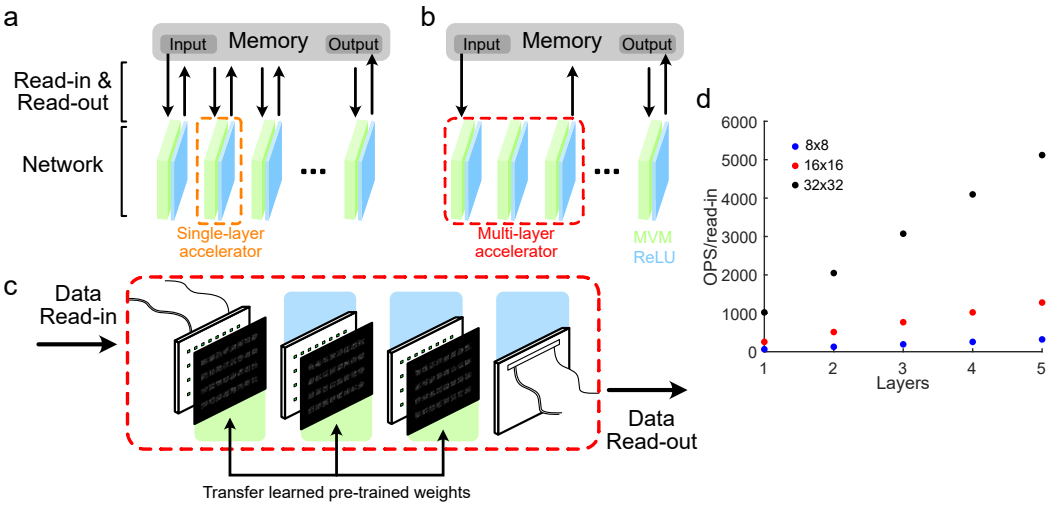


Figure 47: **Conceptual illustration of multilayer OENN advantages.** (a) Data flow in a typical single-layer accelerator scenario, requiring external data read-in and read-out for each layer processed. (b) Data flow in the multilayer OENN architecture, where intermediate results pass directly between layers within the system, minimizing external memory access. (c) Diagram representing the implemented three-layer MVM + ReLU architecture capable of utilizing transferred weights. (d) Plot illustrating the scaling advantage: the number of compute operations performed per external data read-in increases significantly as the number of layers processed within the accelerator grows, or as the array size (N) increases. Figure adapted from [118].

optics, which could also support convolutional architectures. Higher-density incoherent light sources, such as micro-LED displays, offer one potential solution. Achieving the projected 35 TOPS/W efficiency also depends critically on developing low-power, high-speed analog ASICs (Fig. 41).

Despite these challenges, the demonstrated scalability and architectural strengths position the OENN as a promising candidate for energy-efficient AI hardware. Fully realizing this potential will require further research into alternative optical designs and integrated electronics. The next chapter explores key development directions—including custom ASIC strategies, the implementation of recurrence and convolution, and the use of advanced techniques like differentiable ray tracing for incoherent diffractive networks. These advancements will be essential for translating the OENN architecture to practical, real-world applications.

7

CONCLUSION AND FUTURE OUTLOOK

This doctoral thesis explores the potential of using machine learning (ML) techniques to enable the use of a novel optical effect to shape the optical wavefront. In addition, it demonstrates a new opto-electronic hardware implementation for neural network inference with greater energy efficiency compared to conventional hardware. The work presented herein is accordingly divided into two primary parts: The first part introduces a novel optical effect that allows for shaping the amplitude and polarization of a wavefront on diffraction, solely by spatially modulating the polarization across the wavefront. The limits of traditional phase retrieval algorithms for this optimization problem is discussed and a gradient-based optimization technique to optimize these polarization holograms is introduced. To the best of my knowledge, this is the first such demonstration of polarization modulation-based shaping of a wavefront to form diffraction images purely because of polarization. The second part proposes and demonstrates a novel opto-electronic neural network architecture that combines optical matrix-vector multiplication (MVM) with electronic non-linear activation functions to achieve scalable, energy-efficient neural network computations. Using existing components and a combination of LEDs and photodetectors an energy efficient multi-layer NN has been realized. The following sections summarize the key findings and describe future research directions in each of these two topics.

PART I: MACHINE LEARNING FOR POLARIZATION HOLOGRAM OPTIMIZATION

The manipulation of propagating optical wavefronts by specifically exploiting the diffractive effects that arise from applying a spatially inhomogeneous polarization distribution is demonstrated. This approach offers an alternative to traditional holographic techniques based purely on phase or amplitude modulation. We first demonstrated the underlying fundamental principle using an analog of Young's double-slit experiment, showing how polarization patterns can produce effects traditionally achieved with amplitude masks. Subsequently, we generated a non-diffracting pseudo-Bessel beam solely with a tailored radial polarization mask, demonstrating the potential of this phenomenon for a classical beam shaping application.

However, the true potential of this phenomenon lies in its application to generate more general wavefronts. Traditional phase retrieval algorithms perform poorly on this optimization problem. A key contribution of this thesis was the development of a differentiable physics-based model that can be directly optimized for the requisite polarization modulation mask using gradient-based optimization. Achieving high-fidelity results necessitated the creation of custom compound loss functions that incorporated relevant image quality metrics like SSIM alongside contrast-

enhancing penalty terms. Using this ML framework, we successfully demonstrated the optimization of polarization masks for complex amplitude targets.

Furthermore, we extended this technique to optimize a polarization modulation mask to simultaneously obtain a target amplitude as well as a target polarization distribution on the target plane on diffraction. This result is notable as it allows for explicit control of two output parameters using a single input degree of freedom. This required designing a more complex joint loss function capable of balancing the fidelity requirements for both output channels. Successful experimental generation of such joint amplitude/polarization targets clearly demonstrate the potential of this approach, while simultaneously highlighting the inherent coupling limitations that arise when controlling multiple output parameters with a single input modality.

The technique introduced here demonstrates beam shaping using only the polarization degree of freedom. However, we also demonstrate a proof-of-concept experiment that combines polarization modulation with conventional phase modulation. This combined approach was applied to realize a non-mechanical point scanning system, illustrating the potential extensibility of this technique to realize more complex optical systems not directly achievable with either method alone.

Collectively, this work shows that spatially varying polarization is a viable and, importantly, optimizable degree of freedom for wavefront shaping applications. The application of machine learning optimization techniques, particularly when coupled with differentiable physical models and carefully tailored loss functions has been very successful. The core principles established here—namely, treating polarization as a controllable parameter within a differentiable optical system amenable to gradient-based optimization—suggest potential directions for future research.

Outlook

Outlook: Beam manipulation and shaping form a core part of many other fields of research, such as material processing [211], microscopy [109], and display technology [212]. Traditional vectorial holography methods typically combine phase and polarization modulation to achieve full vectorial control of the wavefront on the target plane. The technique developed in this work can simplify the generation of arbitrary vectorial wavefronts dynamically, as it relies solely on polarization. Although we have only demonstrated simultaneous control over amplitude and polarization on the target plane, the optimization framework can be extended for full vectorial control over the wavefront on the detection plane. However, this extension is non-trivial as amplitude, phase, and polarization are all coupled in diffraction theory, and only one degree of freedom, namely, polarization modulation, is available for optimization.

The techniques presented in this part of the thesis are also relevant for designing a polarization-based deep diffractive neural network (DDNN) implementation. The advantage of this approach lies in its ability to implement non-linearities in diffractive neural networks, which has been an Achilles' heel for the field, as de-

scribed in Chap. 1. Fig. 48 demonstrates a single layer of the proposed polarization-based DDNN using the principles developed in this thesis. The inputs and the trained weights are implemented as a polarization mask, unlike in a traditional diffractive neural network where the weights are implemented as a phase mask. The polarization-encoded wavefront diffracts before passing through a polarizer, which is mathematically equivalent to a sinusoidal projection of the spatial polarization distribution onto the amplitude domain. The wavefront propagates further before passing through an amplitude-to-polarization converter.

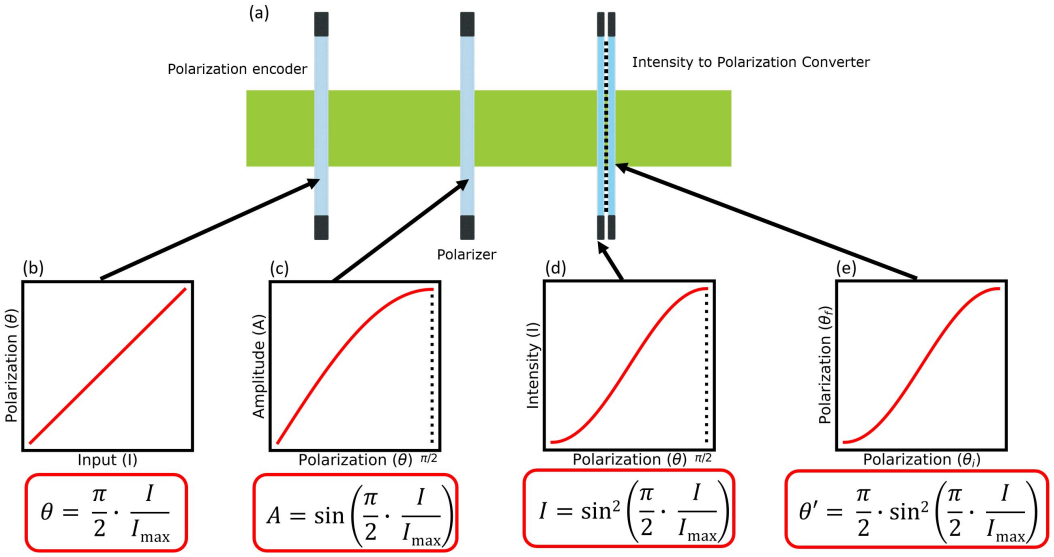


Figure 48: **Implementation of non-linearities in a polarization-based diffractive neural network.** A schematic of the proposed polarization-based non-linearity in a polarization-based diffractive neural network. The information from the source/previous layer is encoded in the polarization state of the light. The encoded weights are optimized using the differentiable model of the system. Subsequently, after propagation, the light passes through a polarizer where a sinusoidal projection is applied to the incident wavefront. The spatially varying amplitude wavefront propagates further onto an intensity-to-polarization converter, where a square detection non-linearity is applied, and the wavefront is converted to a spatially polarization and amplitude varying wavefront. The effective result of this is a sinusoidal squared non-linearity that is applied to the input wavefront. The output of this layer can be used as the input to the next layer.

This device is modeled based on a widely used liquid crystal device called the liquid crystal light valve (LCLV) or optically addressable spatial light modulator (OASLM) [213], which in the past has also been applied to the field of optical computing [213–215]. The LCLV has a pixelated photosensitive square-law intensity detection material, such as hydrogenated amorphous silicon [216], that converts the light intensity into a field distribution. Depending on the type and geometry of the liquid crystal cell, this spatially varying field can be used to modulate the polarization of the light passing through it. This constitutes a single layer in the proposed polarization-based optical neural network that effectively implements a sinusoidal squared non-linearity.

The proposed approach is a promising future research direction as it allows for the implementation of low-energy non-linearities in a diffractive neural network using existing technology. In addition, phase weights can also be implemented in the network, providing greater flexibility for optimizing a specific problem at hand.

PART II: SCALABLE LOW POWER OPTOELECTRONIC NEURAL NETWORKS

This thesis has shown that the proposed multilayer optoelectronic neural network (MOENN) architecture presents a scalable and energy-efficient alternative to conventional computers for neural network inference. The MOENN architecture takes advantage of the parallelism offered by optical matrix-vector-multiplication (MVM) and implements an energy-efficient difference-ReLU non-linear activation. A proof-of-concept lab-scale system was developed and experimentally validated on the classification task for the MNIST dataset. Subsequently, the system was benchmarked on the classification of a custom-developed spiral dataset, demonstrating its ability to generalize over non-linear decision boundaries.

Building on this experimental foundation, we investigated the scalability of the architecture through analytical modeling and simulation. A faster circuit implementation was proposed and we verified the system operation at 10MHz using electronic simulations. Optical simulations, such as ray tracing and a modified angular spectrum method, were used to show that the system can be scaled up to 48×48 neurons in the same physical footprint as the proof-of-concept system before diffraction effects become significant. For a system of this size, we projected a computational throughput of 10.5 TOPS and an energy efficiency of 35 TOPS/W, the latter surpassing that of state-of-the-art electronic accelerators.

While these results are promising, transforming the MOENN concept into a practical, high-performance computing platform requires addressing several fundamental challenges that arise with scaling. These challenges span both optical and electronic domains. As the number of operations increases within a fixed physical footprint, higher interconnect density leads to stronger diffraction effects and a greater risk of optical crosstalk. Advancing the optical design is therefore critical—particularly in overcoming the physical limits that emerge when attempting to increase interconnect density. As feature sizes shrink to accommodate larger networks within compact footprints, diffraction becomes a dominant concern, threatening the fidelity of the optical MVM. The following sections will first focus on these optical challenges and possible solutions, before addressing issues related to electronic scaling.

Outlook: Advances in optics

The optical design of the MOENN architecture is a key factor in its performance and scalability. With scaling, as the number of neurons and consequently interconnects increases, the density of these interconnects also rises if the physical footprint is to remain constant. As a result, maintaining high fidelity in optical MVM

becomes increasingly challenging because of diffraction effects for light sources and Gaussian apertures at such a small scale. This increases crosstalk between the optical interconnects, which can lead to significant degradation in the neural network's performance. Therefore, addressing this challenge is crucial to ensure the successful scaling of the MOENN architecture.

One promising approach to mitigate diffraction-imposed limitations, particularly the strict requirements of point-to-point mappings in fully connected optical layers, involves implementing convolutional layers in optoelectronic neural networks. Convolutional layers, which apply shared filters to extract local features from input data, are inherently more robust to spatial distortions and require fewer parameters for a given computation. This spatial invariance, coupled with the kernel-sharing mechanism, makes them well-suited for optical hardware, where diffraction and alignment tolerances are critical considerations [217].

In optical systems, Fourier optics can efficiently implement convolutions. Specifically, a 4f configuration enables convolution operations by placing a filter in the Fourier plane of a lens, offering a compact and efficient optical analog to electronic convolution operations. To further advance this concept within the framework of a multi-kernel optoelectronic neural network (MOENN) architecture, we propose an optical convolutional scheme that utilizes microlens arrays (MLAs) in a spatially structured manner [218].

In our approach, a microlens array is positioned between a structured light emitter array and an amplitude mask containing convolution kernels, as shown in Fig. 49. Each light emitter interacts with a local group of microlenses, creating focused light spots designed to illuminate the exact locations of the corresponding kernel elements on the mask. Emitters in the same sub-array are spatially configured so that their output light fields align onto the same kernel, ensuring coherent sampling of identical feature regions. After passing through the mask, the light travels through free space, spreading to form the distinct, shifted patterns of the output channels on the sensor, which correspond to different convolutional channels. Independent emitter sub-arrays are assigned to distinct kernel regions on the mask, enabling multiple convolutions to be implemented in parallel.

Validation simulations of this approach, modeling each emitter as a Light Emitting Diode (LED) with a defined angular emission profile, have shown the potential. Photon propagation through the system, including the microlens array, amplitude mask, and free-space region, was modeled using ray tracing; Fig. 49 illustrates how the system was tested by convolving sparse input data, characterized by an exponential distribution, with randomly generated kernels. The ray-tracing simulation produced results that showed strong agreement with ideal digital convolution, indicated by a high Pearson correlation (0.902). These findings validate the potential of using structured light emission and microlens-mediated multiplexing for implementing convolutional operations in optical hardware. Key challenges for incorporating this method include achieving high-precision alignment of the MLA and mask components at scale and further optimizing energy efficiency [119].

However, employing the computational technique of differentiable ray tracing to design arbitrary diffractive layers for incoherent light can further mitigate these

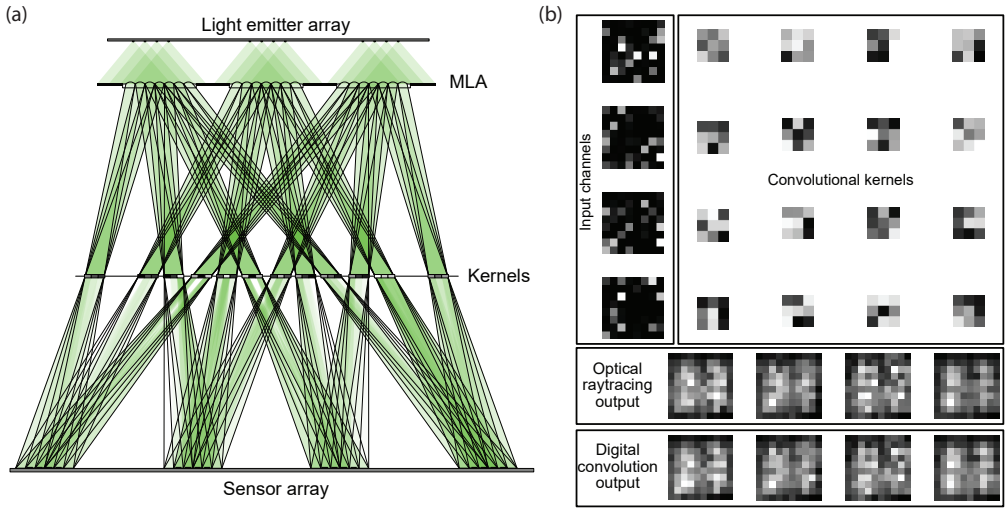


Figure 49: **Multi-channel free-space optical convolution approach.** (a) The schematic illustrates the optical setup where light from emitters passes through a microlens array (MLA) and an amplitude mask containing the convolutional kernels before reaching the sensor. (b) Simulated performance compares the optical raytracing output against ideal digital convolution for multiple input channels and kernels, demonstrating high fidelity with a Pearson correlation of 0.902. Figure adapted from [119].

challenges. Differentiable ray tracing is well-established in computer graphics [219], but in recent years, it has also become widely adopted in the optics and optical design community [220–222]. The resulting optimized phase masks can be physically implemented using transmissive phase spatial light modulators (SLMs) [223] or custom-printed diffractive optical elements (DOEs), which can be fabricated using lithography techniques [224].

Future work within the MOENN project will focus on integrating such optical convolutional layers, potentially replacing or complementing the dense MVM layers, and experimentally characterizing their performance within the full system, thereby supporting the development of scalable, low-power optical neural network coprocessors.

Outlook: Advances in electronics

The MOENN prototype shown uses discrete electronic components on printed circuit boards (PCBs) allowing for rapid iterative prototyping. However, this approach is neither scalable nor energy efficient as a practical implementation. This restricts the throughput of the lab-scale implementation, diminishing its real-world applicability. Application-specific integrated circuits (ASICs) solve this problem through an optimized implementation of the required analog electronic functions in a compact and energy-efficient manner. In addition, tight integration of electronic functions inside an ASIC unlocks the possibility of implementing additional useful features. Fig. 50 shows a proposed block diagram of a custom ASIC, pro-

posed by our collaborators at IMS Chips Stuttgart, that can be used to implement the OENN described in this thesis.

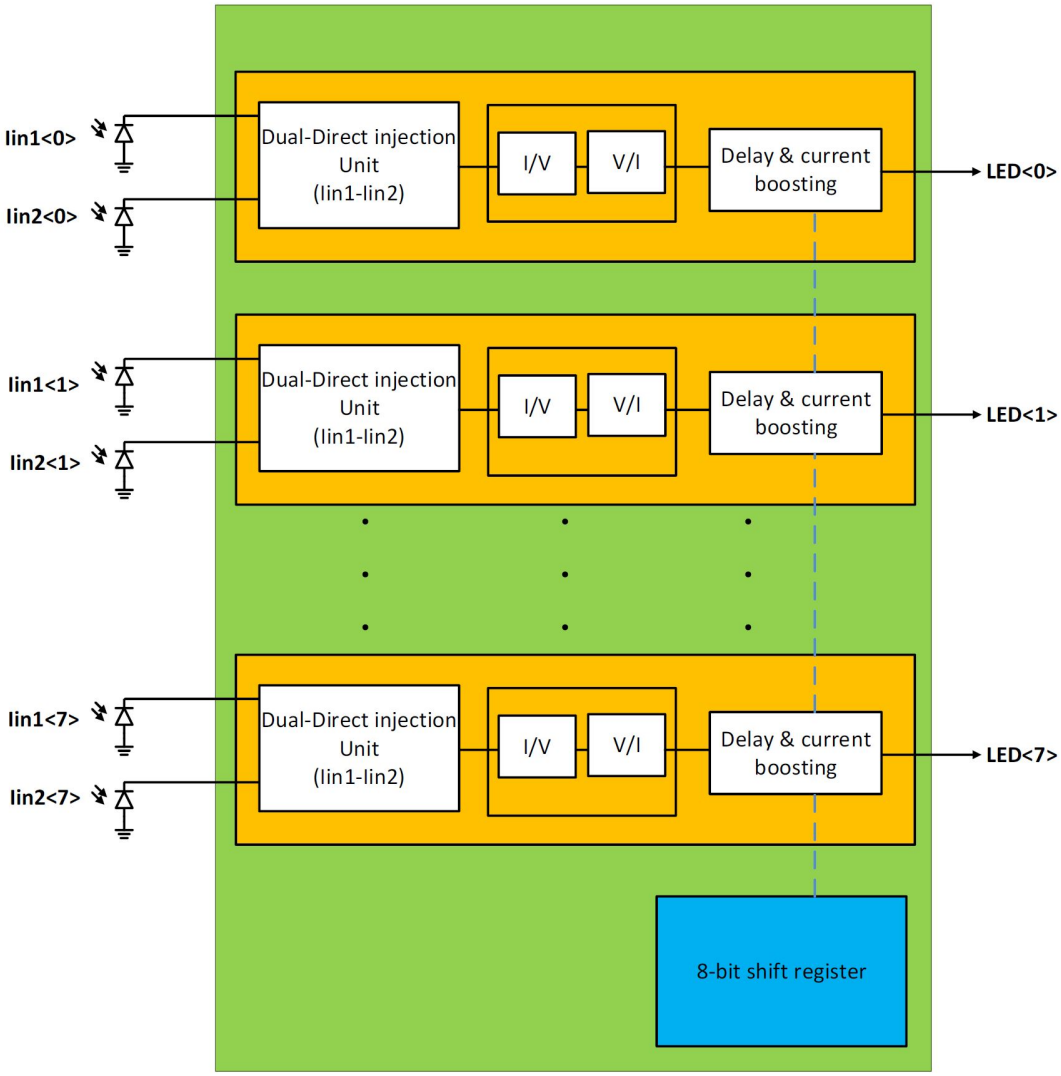


Figure 50: **Proposed ASIC for optoelectronic neural networks.** Block-diagram showing the implementation of custom ASIC, featuring eight parallel analog channels for photodiode current subtraction, ReLU activation, programmable amplification, and LED driving with the ability to introduce a delay. Figure reproduced from an internal communication with IMS Chips Stuttgart.

The proposed ASIC performs all of the same functions as the prototype system, but it additionally implements variable amplification of LED output as well as electronic memory. These supplementary features allow for a more reliable operation of the system and allow for the implementation of neural network architectures that would benefit from memory, such as a recurrent neural network (RNN). In addition to the analog electronics, CMOS processes used for fabricating the ASICs also open up the possibility of implementing photodetectors on the same chip, as is commonly done with CMOS image sensors (CIS) [225]. Traditionally, it has been challenging to implement light emitters in the same CMOS process node [226]; however, recent progress in microLED technology has opened up a roadmap for

implementing a small, fast, and energy-efficient light source that is compatible with the same manufacturing process as the rest of the electronics [227, 228]. Combining all the possible technological integrations mentioned here opens a path for a fully integrated optoelectronic neural network on a single chip.

An energy-efficient and compact MOENN chip would be a significant help in many practical applications that require a low-latency, compact, and energy-efficient system while working natively with optical inputs. Two such applications are autonomous vehicles [229] and untethered robotics [80]. In both of these applications, the vision input is natively optical, and both of these systems need to make real-time inference on the scene around. The MOENN architecture is well-suited for this as the input signal is directly processed without digitization and digital processing. In addition, these applications have limited power budget to spare for computing and would, thus, benefit from the energy-efficient nature of the MOENN architecture. The proposed ASIC implementation would allow for a more compact and energy-efficient system that can be integrated into these applications.

BIBLIOGRAPHY

[1] B. K. Spears, "Contemporary machine learning: a guide for practitioners in the physical sciences," arXiv:1712.08523.

[2] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is All you Need," in Advances in Neural Information Processing Systems, Vol. 30 (2017).

[3] J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, et al., "Highly accurate protein structure prediction with AlphaFold," Nature **596**, 583 (2021).

[4] K. Akiyama, A. Alberdi, W. Alef, K. Asada, R. Azulay, A.-K. Baczko, D. Ball, M. Baloković, J. Barrett, D. Bintley, et al., "First M87 event horizon telescope results. IV. Imaging the central supermassive black hole," The Astrophysical Journal Letters **875**, L4 (2019).

[5] L. Deng, "The MNIST database of handwritten digit images for machine learning research," IEEE signal processing magazine **29**, 141 (2012).

[6] A. Krizhevsky, Learning multiple layers of features from tiny images, tech. rep. (University of Toronto, 2009).

[7] D. Schneider, "The Exascale Era is Upon Us: The Frontier supercomputer may be the first to reach 1,000,000,000,000,000,000 operations per second," IEEE Spectrum **59**, 34–35 (2022).

[8] F. Benning, "High Dimensional Optimization through the Lens of Machine Learning," arXiv:2112.15392.

[9] M. T. Augustine, "A Survey on Universal Approximation Theorems," arXiv:2407.12895.

[10] A. Jung, Machine learning: the basics (Springer Nature, 2022).

[11] S. Kotsiantis and P. Pintelas, "Recent advances in clustering: A brief survey," in WSEAS Transactions on Information Science and Applications, Vol. 1 (2004), p. 73.

[12] A. Nowe and T. Brys, "A gentle introduction to reinforcement learning," in Scalable Uncertainty Management: 10th International Conference (Springer, 2016), p. 18.

[13] S. Kamyab, Z. Azimifar, R. Sabzi, and P. W. Fieguth, "Survey of Deep Learning Methods for Inverse Problems," arXiv:2111.04731.

[14] J. W. Goodman, Introduction to Fourier optics (Roberts and Company publishers, 2005).

[15] M. Shoeybi, M. Patwary, R. Puri, P. LeGresley, J. Casper, and B. Catanzaro, "Megatron-LM: Training Multi-Billion Parameter Language Models Using Model Parallelism," arXiv:1909.08053.

[16] T. Nguyen, J. Brandstetter, A. Kapoor, J. K. Gupta, and A. Grover, "Climax: A foundation model for weather and climate," arXiv:2301.10343.

[17] T. Wang, S.-Y. Ma, L. G. Wright, T. Onodera, B. C. Richard, and P. L. McMahon, "An optical neural network using less than 1 photon per multiplication," Nature Communications **13**, 123 (2022).

[18] T. Gahlmann and P. Tassin, "Deep neural networks for the prediction of the optical properties and the free-form inverse design of metamaterials," Physical Review B **106**, 085408 (2022).

[19] N. Borhani, E. Kakkava, C. Moser, and D. Psaltis, "Learning to see through multimode fibers," Optica **5**, 960 (2018).

[20] B. Y. Feng, H. Guo, M. Xie, V. Boominathan, M. K. Sharma, A. Veeraraghavan, and C. A. Metzler, "NeuWS: Neural wavefront shaping for guidestar-free imaging through static and dynamic scattering media," Science Advances **9** (2023).

[21] J. Zhang, N. Pegard, J. Zhong, H. Adesnik, and L. Waller, "3D Computer-generated holography by non-convex optimization," Optica **4**, 1306 (2017).

[22] R. Gerchberg and W. Saxton, "A practical algorithm for the determination of phase from image and diffraction plane pictures," SPIE milestone series MS **94**, 646 (1994).

[23] P. J. Christopher, R. Mouthaan, M. El Guendy, and T. D. Wilkinson, "Linear-time algorithm for phase-sensitive holography," Optical Engineering **59**, 085104 (2020).

[24] B. K. Jennison, J. P. Allebach, and D. W. Sweeney, "Iterative approaches to computer-generated holography," in Optical Engineering, Vol. 28 (1989).

[25] T. Harte, G. D. Bruce, J. Keeling, and D. Cassettari, "Conjugate gradient minimisation approach to generating holographic traps for ultracold atoms," Optics Express **22**, 26548 (2014).

[26] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," IEEE Transactions on Image Processing **13**, 600 (2004).

[27] T. Fu, J. Zhang, R. Sun, Y. Huang, W. Xu, S. Yang, Z. Zhu, and H. Chen, "Optical neural networks: progress and challenges," Light: Science & Applications **13**, 263 (2024).

[28] Y. Shen, N. C. Harris, S. Skirlo, M. Prabhu, T. Baehr-Jones, M. Hochberg, X. Sun, S. Zhao, H. Larochelle, D. Englund, et al., "Deep learning with coherent nanophotonic circuits," Nature Photonics **11**, 441 (2017).

[29] J. Carolan, C. Harrold, C. Sparrow, E. Martin-Lopez, N. J. Russell, J. W. Silverstone, P. J. Shadbolt, N. Matsuda, M. Oguma, M. Itoh, et al., "Universal linear optics," Science **349**, 711 (2015).

[30] A. N. Tait, A. X. Wu, T. F. De Lima, E. Zhou, B. J. Shastri, M. A. Nahmias, and P. R. Prucnal, "Microring weight banks," IEEE Journal of Selected Topics in Quantum Electronics **22**, 312 (2016).

[31] X. Lin, Y. Rivenson, N. T. Yardimci, M. Veli, Y. Luo, M. Jarrahi, and A. Ozcan, "All-optical machine learning using diffractive deep neural networks," Science **361**, 1004 (2018).

[32] S. N. M. Kottapalli, A. Song, and P. Fischer, "Diffraction of non-uniformly polarized beams enables beam manipulation," arXiv:2203.11185.

[33] A. Forbes, M. De Oliveira, and M. R. Dennis, "Structured light," Nature Photonics **15**, 253 (2021).

[34] F. M. Dickey, Laser beam shaping: theory and techniques (CRC press, 2018).

[35] P. Hariharan, Optical Holography: Principles, techniques and applications (Cambridge University Press, 1996).

[36] D. G. Grier, "A revolution in optical manipulation," Nature **424**, 810 (2003).

[37] K. Dholakia and T. Cizmar, "Shaping the future of manipulation," Nature Photonics **5**, 335 (2011).

[38] B. Mills, D. Heath, M. Feinaeugle, J. Grant-Jacob, and R. Eason, "Laser ablation via programmable image projection for submicron dimension machining in diamond," Journal of Laser Applications **26** (2014).

[39] A. E. Willner, H. Huang, Y. Yan, Y. Ren, N. Ahmed, G. Xie, C. Bao, L. Li, Y. Cao, Z. Zhao, et al., "Optical communications using orbital angular momentum beams," Advances in Optics and Photonics **7**, 66 (2015).

[40] R. Fickler, R. Lapkiewicz, W. N. Plick, M. Krenn, C. Schaeff, S. Ramelow, and A. Zeilinger, "Quantum entanglement of high angular momenta," Science **338**, 640 (2012).

[41] Y. Koo, T. Moon, M. Kang, H. Joo, C. Lee, H. Lee, V. Kravtsov, and K.-D. Park, "Dynamical control of nanoscale light-matter interactions in low-dimensional quantum materials," Light: Science & Applications **13**, 30 (2024).

[42] M. Fratz, D. M. Giel, and P. Fischer, "Digital polarization holograms with defined magnitude and orientation of each pixel's birefringence," Optics Letters **34**, 1270 (2009).

[43] L. J. Salazar-Serrano, D. A. Guzman, A. Valencia, and J. P. Torres, "Demonstration of a highly-sensitive tunable beam displacer with no use of beam deflection based on the concept of weak value amplification," Optics Express **23**, 10097 (2015).

[44] M. R. Dennis, K. O'Holleran, and M. J. Padgett, "Singular optics: optical vortices and polarization singularities," Progress in Optics **53**, 293 (2009).

[45] H. Chen, S. Tripathi, and K. C. Toussaint, "Demonstration of flat-top focusing under radial polarization illumination," Optics Letters **39**, 834 (2014).

[46] T. G. Brown and A. M. Beckley, "Stress engineering and the applications of inhomogeneously polarized optical fields," Frontiers of Optoelectronics **6**, 89 (2013).

[47] E. Arbabi, S. M. Kamali, A. Arbabi, and A. Faraon, "Vectorial holograms with a dielectric metasurface: ultimate polarization pattern generation," ACS Photonics **6**, 2712 (2019).

[48] N. A. Rubin, A. Zaidi, A. H. Dorrah, Z. Shi, and F. Capasso, "Jones matrix holography with metasurfaces," Science Advances **7**, eabg7488 (2021).

[49] J. Cai, F. Zhang, M. Zhang, Y. Ou, and H. Yu, "Simultaneous polarization filtering and wavefront shaping enabled by localized polarization-selective interference," Scientific Reports **10**, 14477 (2020).

[50] B. E. Saleh and M. C. Teich, Fundamentals of photonics (Wiley, 2008).

[51] C. Rosales-Guzman and A. Forbes, How to shape light with spatial light modulators (Society of Photo-Optical Instrumentation Engineers (SPIE), 2017).

[52] P. Yu, Y. Liu, Y. Wu, Z. Wang, Y. Li, and L. Gong, "Dynamic polarization holographic projection enabled by a scattering material-based reconfigurable hologram," ACS Photonics **9**, 3712 (2022).

[53] Y.-X. Ren, R.-D. Lu, and L. Gong, "Tailoring light with a digital micromirror device," Annalen der Physik **527**, 447 (2015).

[54] Q. Zhang, Z. He, Z. Xie, Q. Tan, Y. Sheng, G. Jin, L. Cao, and X. Yuan, "Diffractive optical elements 75 years on: from micro-optics to metasurfaces," Photonics Insights **2**, R09 (2023).

[55] W. Osten and N. O. Reingand, Optical Imaging and Metrology: Advanced Technologies (John Wiley & Sons, 2012).

[56] J. Durnin, J. Miceli Jr, and J. H. Eberly, "Diffraction-free beams," Physical Review Letters **58**, 1499 (1987).

[57] D. McGloin and K. Dholakia, "Bessel beams: diffraction in a new light," Contemporary Physics **46**, 15 (2005).

[58] S. Jimenez-Gambin, N. Jimenez, J. M. Benlloch, and F. Camarena, "Generating Bessel beams with broad depth-of-field by using phase-only acoustic holograms," Scientific Reports **9**, 20104 (2019).

[59] R. Tudor, G. A. Bulzan, M. Kusko, C. Kusko, V. Avramescu, D. Vasilache, and R. Gavrila, "Multilevel spiral axicon for high-order Bessel–Gauss beams generation," Nanomaterials **13**, 579 (2023).

[60] S. Fu, S. Zhang, and C. Gao, "Bessel beams with spatial oscillating polarization," Scientific Reports **6**, 30765 (2016).

[61] H. Liu, H. Xue, Y. Liu, and L. Li, "Generation of multiple pseudo Bessel beams with accurately controllable propagation directions and high efficiency using a reflective metasurface," Applied Sciences **10**, 7219 (2020).

[62] D. Fan, L. Wang, and Y. Ekinci, "Nanolithography using Bessel beams of extreme ultraviolet wavelength," Scientific Reports **6**, 31301 (2016).

[63] A. Konijnenberg, A. J. Adam, and P. Urbach, BSc Optics (TU Delft OPEN Publishing, 2021).

[64] D. Caprini et al., "Generation, dynamics and control of microbubbles in microdevices," PhD thesis (Universita degli Studi di Roma La Sapienza, 2019).

[65] J. Peatross and M. Ware, Physics of Light and Optics (Brigham Young University, 2015).

[66] LC 2012 Spatial Light Modulator (transmissive) - HOLOEYE Photonics AG, 2025.

[67] F. L. Pedrotti, L. M. Pedrotti, and L. S. Pedrotti, Introduction to optics (Cambridge university press, 2017).

[68] P. Yeh, "Optics of liquid crystal displays," in 2007 Conference on Lasers and Electro-Optics-Pacific Rim (IEEE, 2007), p. 1.

[69] Y. Yang, A. Forbes, and L. Cao, "A review of liquid crystal spatial light modulators: devices and applications," Opto-Electronic Science **2**, 230026 (2023).

[70] M. Cepic, "Liquid Crystals In Education–The Basics," European Journal of Physics Education **3** (2012).

[71] K. Dev, V. R. Singh, and A. Asundi, "Full-field TN-LCSLM phase modulation characterization using digital holography," in Liquid Crystals XIV, Vol. 7775 (SPIE, 2010), p. 216.

[72] A. Marquez and A. Lizana, "Special issue on liquid crystal on silicon devices: modeling and advanced spatial light modulation applications," Applied Sciences **9**, 3049 (2019).

[73] F. Difato, M. D. Maschio, R. Beltramo, A. Blau, F. Benfenati, and T. Fellin, "Spatial light modulators for complex spatiotemporal illumination of neuronal networks," Neuronal Network Analysis: Concepts and Experimental Approaches **67**, 61 (2012).

[74] PLUTO-2.1 LCOS Spatial Light Modulator - HOLOEYE Photonics AG, en-US, 2025.

[75] J. Schwiegerling and D. R. Neal, "Historical development of the Shack-Hartmann wavefront sensor," Robert Shannon and Roland Shack: Legends in Applied Optics **1**, 132 (2005).

[76] R. Paschotta, Shack-Hartmann wavefront sensors, 2025.

[77] D. R. Neal, J. Copland, and D. A. Neal, "Shack-Hartmann wavefront sensor precision and accuracy," in Advanced Characterization Techniques for Optical, Semiconductor, and Data Storage Components, Vol. 4779 (2002), p. 148.

[78] R. Mazzoleni, F. Gonte, I. Surdej, C. Araujo, R. Brast, F. Derie, P. Duhoux, C. Dupuy, C. Frank, R. Karban, et al., "Design and performances of the Shack-Hartmann sensor within the Active Phasing Experiment," in Ground-based and Airborne Telescopes II, Vol. 7012 (2008), p. 1246.

[79] B. Schafer, J. Gloger, U. Leinhos, and K. Mann, "Photo-thermal measurement of absorptance losses, temperature induced wavefront deformation and compaction in DUV-optics," Optics Express **17**, 23025 (2009).

[80] R. M. Abdelazeem, M. M. A. Ahmed, S. Hassab-Elnaby, and M. Agour, "Improving the phase modulation of spatial light modulator using Shack-Hartmann wavefront sensor," in Optical Measurement Systems for Industrial Inspection XIII (2023).

[81] J. D. Schmidt, Numerical simulation of optical wave propagation with examples in MATLAB (SPIE, 2010).

[82] Y. Zhang, H. An, D. Zhang, G. Cui, and X. Ruan, "Diffraction theory of high numerical aperture subwavelength circular binary phase Fresnel zone plate," Optics Express **22**, 27425 (2014).

[83] K. Melde, H. Kremer, M. Shi, S. Seneca, C. Frey, I. Platzman, C. Degel, D. Schmitt, B. Scholkopf, and P. Fischer, "Compact holographic sound fields enable rapid one-step assembly of matter in 3D," Science Advances **9**, eadf6182 (2023).

[84] X. Zeng and R. J. McGough, "Evaluation of the angular spectrum approach for simulations of near-field pressures," The Journal of the Acoustical Society of America **123**, 68 (2008).

[85] D. M. Cottrell and J. A. Davis, "Computational methods in applying the angular spectrum algorithm to optical fibers," OSA Continuum **3**, 1346 (2020).

[86] K. Matsushima, "Shifted angular spectrum method for off-axis numerical propagation," Optics Express **18**, 18453 (2010).

[87] K. Matsushima and T. Shimobaba, "Band-limited angular spectrum method for numerical simulation of free-space propagation in far and near fields," Optics Express **17**, 19662 (2009).

[88] S. Mehrabkhani and T. Schneider, "Is the Rayleigh-Sommerfeld diffraction always an exact reference for high speed diffraction algorithms?" Optics Express **25**, 30229 (2017).

[89] L. Palfalvi, Z. T. Godana, and J. Hebling, "Electromagnetic Field Distribution and Divergence-Dependence of a Radially Polarized Gaussian Vector Beam Focused by a Parabolic Mirror," arXiv:2406.00795.

[90] N. Lindlein, M. Sondermann, R. Maiwald, H. Konermann, U. Peschel, and G. Leuchs, "Focusing light with a deep parabolic mirror," arXiv:1905.05997.

[91] P. Varga and P. Torok, "Focusing of electromagnetic waves by paraboloid mirrors. I. Theory," Journal of the Optical Society of America A **17**, 2081 (2000).

[92] Y. Ishii, T. Shimobaba, D. Blinder, T. Birnbaum, P. Schelkens, T. Kakue, and T. Ito, "Optimization of phase-only holograms calculated with scaled diffraction calculation through deep neural networks," Applied Physics B **128**, 22 (2022).

[93] A. Rundquist, A. Efimov, and D. H. Reitze, "Pulse shaping with the Gerchberg–Saxton algorithm," Journal of the Optical Society of America B **19**, 2468 (2002).

[94] G. Huang, D. Wu, J. Luo, L. Lu, F. Li, Y. Shen, and Z. Li, "Generalizing the Gerchberg–Saxton algorithm for retrieving complex optical transmission matrices," Photonics Research **9**, 34 (2020).

[95] J. R. Fienup, "Phase retrieval algorithms: a comparison," Applied Optics **21**, 2758 (1982).

[96] G. Shang, H. Li, Z. Wang, K. Zhang, S. N. Burokur, J. Liu, Q. Wu, X. Ding, and X. Ding, "Coding metasurface holography with polarization-multiplexed functionality," Journal of Applied Physics **129** (2021).

[97] G. Shang, H. Li, Z. Wang, K. Zhang, S. N. Burokur, J. Liu, Q. Wu, X. M. Ding, and X. Ding, "Coding metasurface holography with polarization-multiplexed functionality," Journal of Applied Physics **129**, 035304 (2021).

[98] R. Horisaki, R. Takagi, and J. Tanida, "Deep-learning-generated holography," Applied Optics **57**, 3859 (2018).

[99] K. Zhang, W. Zuo, Y. Chen, D. Meng, and L. Zhang, "Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising," IEEE transactions on image processing **26**, 3142 (2017).

[100] Y. Rivenson, Y. Zhang, H. Gunaydin, D. Teng, and A. Ozcan, "Phase recovery and holographic image reconstruction using deep learning in neural networks," Light: Science & Applications **7**, 17141 (2018).

[101] H. Ren, W. Shao, Y. Li, F. Salim, and M. Gu, "Three-dimensional vectorial holography based on machine learning inverse design," Science Advances **6**, eaaz4261 (2020).

[102] L.-Y. Yu and S. You, "High-fidelity and high-speed wavefront shaping by leveraging complex media," Science Advances **10**, eadn2846 (2024).

[103] J. Xi, J. Shen, M. T. Chow, T. Li, J. Ng, and J. Li, "Deep-learning assisted polarization holograms," Advanced Optical Materials **12**, 2202663 (2024).

[104] T. Liu, K. de Haan, B. Bai, Y. Rivenson, Y. Luo, H. Wang, D. Karalli, H. Fu, Y. Zhang, J. FitzGerald, et al., "Deep learning-based holographic polarization microscopy," ACS Photonics **7**, 3023 (2020).

[105] Y. Zhang and E. Y. Lam, "Robust holographic imaging for real-world applications with joint optimization," Optics Express **33**, 5932 (2025).

[106] H. Wei, X. He, and W. Cao, "Spin-multiplexed metasurface inverse-design based on a bi-directional deep neural network for terahertz wavefront control," Optica **12**, 505 (2025).

[107] D. Yang, W. Seo, H. Yu, S. I. Kim, B. Shin, C.-K. Lee, S. Moon, J. An, J.-Y. Hong, G. Sung, et al., "Diffraction-engineered holography: Beyond the depth representation limit of holographic displays," Nature Communications **13**, 6012 (2022).

[108] J. Kim, J.-Y. Kim, J. Kim, Y. Hyeong, B. Neseli, J.-B. You, J. Shim, J. Shin, H.-H. Park, and H. Kurt, "Inverse design of nanophotonic devices enabled by optimization algorithms and deep learning: recent achievements and future prospects," Nanophotonics (2025).

[109] H. Ren, W. Shao, Y. Li, F. Salim, and M. Gu, "Three-dimensional vectorial holography based on machine learning inverse design," Science Advances **6**, eaaz4261 (2020).

[110] A. R. Hall, "Beyond the fringe: diffraction as seen by Grimaldi, Fabri, Hooke and Newton," Notes and Records of the Royal Society of London **44**, 13 (1990).

[111] D. Li and D. Pacifici, "Strong amplitude and phase modulation of optical spatial coherence with surface plasmon polaritons," Science Advances **3**, e1700133 (2017).

[112] N. A. Rubin, A. Zaidi, A. H. Dorrah, Z. Shi, and F. Capasso, "Jones matrix holography with metasurfaces," Science Advances **7**, eabg7488 (2021).

[113] S. N. Khonina, N. L. Kazanskiy, S. V. Karpeev, and M. A. Butt, "Bessel beam: Significance and applications—A progressive review," Micromachines **11**, 997 (2020).

[114] F. Yang, A. Kadis, R. Mouthaan, B. Wetherfield, A. Kaczorowski, and T. D. Wilkinson, "Perceptually motivated loss functions for computer generated holographic displays," Scientific Reports **12**, 7709 (2022).

[115] A. Mustafa, A. Mikhailiuk, D. A. Iliescu, V. Babbar, and R. K. Mantiuk, "Training a task-specific image reconstruction loss," in Proceedings of the IEEE/CVF winter conference on applications of computer vision (2022), p. 2319.

[116] J.-H. Lee, J. Kim, J.-H. Kim, and D.-J. Kim, "Generating High-Resolution SAR Images with Loss Function Customization," in IGARSS 2023-2023 IEEE International Geoscience and Remote Sensing Symposium (2023), p. 5218.

[117] D. P. Kingma, "Adam: A method for stochastic optimization," arXiv:1412.6980.

[118] A. Song, S. N. Murty Kottapalli, R. Goyal, B. Scholkopf, and P. Fischer, "Low-power scalable multilayer optoelectronic neural networks enabled with incoherent light," Nature Communications **15**, 10692 (2024).

[119] A. Song, S. N. M. Kottapalli, and P. Fischer, "Multi-channel free space optical convolutions," in EPJ Web of Conferences, Vol. 287 (2023), p. 13020.

[120] B. Cottier, R. Rahman, L. Fattorini, N. Maslej, and D. Owen, "The rising costs of training frontier AI models," arXiv:2405.21015.

[121] J.-Q. Yang, Y. Zhou, and S.-T. Han, "Functional applications of future data storage devices," Advanced Electronic Materials **7**, 2001181 (2021).

[122] D. Patterson and P. Ranganathan, TPUs improved carbon-efficiency of AI workloads by 3x, en-US, Feb. 2025.

[123] SherryX, Inside Maia 100: Revolutionizing AI Workloads with Microsoft's Custom AI Accelerator, en-US, Sept. 2024.

[124] Co-designing hardware and software for a sustainable future of AI, en-US, May 2024.

[125] K. Ueyoshi, I. A. Papistas, P. Houshmand, G. M. Sarda, V. Jain, M. Shi, Q. Zheng, S. Giraldo, P. Vrancx, J. Doevenspeck, et al., "DIANA: An end-to-end energy-efficient digital and ANAlog hybrid neural network SoC," in 2022 IEEE International Solid-State Circuits Conference (ISSCC), Vol. 65 (IEEE, 2022), p. 1.

[126] W. Wan, R. C. Kubendran, C. J. S. Schaefer, S. B. Eryilmaz, W. Zhang, D. Wu, S. R. Deiss, P. Raina, H. Qian, B. Gao, et al., "Edge AI without Compromise: Efficient, Versatile and Accurate Neurocomputing in Resistive Random-Access Memory," arXiv:2108.07879.

[127] S. Jain, H. Tsai, C.-T. Chen, R. Muralidhar, I. Boybat, M. M. Frank, S. Woźniak, M. Stanisavljevic, P. Adusumilli, P. Narayanan, et al., "A heterogeneous and programmable compute-in-memory accelerator architecture for analog-ai using dense 2-d mesh," IEEE Transactions on Very Large Scale Integration (VLSI) Systems **31**, 114 (2022).

[128] V. Chandra, Y. Chen, and S. Yoo, "Introduction to the special section on energy-efficient AI chips," ACM Transactions on Design Automation of Electronic Systems (TODAES) **27**, 1 (2022).

[129] I. Chakraborty, "Toward Energy-Efficient Machine Learning: Algorithms and Analog Compute-In-memory Hardware," PhD thesis (Purdue University, 2021).

[130] S. J. Kim, I. H. Im, J. H. Baek, S. Choi, S. H. Park, D. E. Lee, J. Y. Kim, S. Y. Kim, N.-G. Park, D. Lee, et al., "Linearly programmable two-dimensional halide perovskite memristor arrays for neuromorphic computing," Nature Nanotechnology **20**, 83 (2025).

[131] G. Indiveri, B. Linares-Barranco, T. J. Hamilton, A. v. Schaik, R. Etienne-Cummings, T. Delbruck, S.-C. Liu, P. Dudek, P. Häfliger, S. Renaud, et al., "Neuromorphic silicon neuron circuits," Frontiers in Neuroscience **5**, 73 (2011).

[132] C. S. Thakur, J. L. Molin, G. Cauwenberghs, G. Indiveri, K. Kumar, N. Qiao, J. Schemmel, R. Wang, E. Chicca, J. Olson Hasler, et al., "Large-scale neuromorphic spiking array processors: A quest to mimic the brain," Frontiers in Neuroscience **12**, 891 (2018).

[133] R. K. Malviya, R. R. Danda, K. K. Maguluri, and B. V. Kumar, "Neuromorphic computing: Advancing energy-efficient ai systems through brain-inspired architectures," Nanotechnology Perceptions **14**, 1548 (2024).

[134] J. W. Goodman, "Fan-in and fan-out with optical interconnections," Optica Acta: International Journal of Optics **32**, 1489 (1985).

[135] N. H. Farhat, D. Psaltis, A. Prata, and E. Paek, "Optical implementation of the Hopfield model," Applied Optics **24**, 1469 (1985).

[136] P. E. Keller and A. F. Gmitro, "Computer-generated holograms for optical neural networks: on-axis versus off-axis geometry," Applied Optics **32**, 1304 (1993).

[137] C. X.-G. Gu, Optical neural networks using volume holograms (California Institute of Technology, 1990).

[138] Y. Owechko and B. H. Soffer, "Optical neural networks based on liquid-crystal light valves and photorefractive crystals," in Liquid-Crystal Devices and Materials, Vol. 1455 (1991), p. 136.

[139] S. PHOTOREFRACTIVE, "OPTICAL NEURAL NETWORKS," Contract **144**, 89 (1992).

[140] P. E. Keller and A. F. Gmitro, "Design and analysis of fixed planar holographic interconnects for optical neural networks," Applied Optics **31**, 5517 (1992).

[141] D. R. Collins, J. B. Sampsell, L. J. Hornbeck, J. M. Florence, P. A. Penz, and M. T. Gately, "Deformable mirror device spatial light modulators and their applicability to optical neural networks," Applied Optics **28**, 4900 (1989).

[142] A. Bergeron, H. H. Arsenault, E. Eustache, and D. Gingras, "Optoelectronic thresholding module for winner-take-all operations in optical neural networks," Applied Optics **33**, 1463 (1994).

[143] D. J. Brady, K. Hsu, and D. Psaltis, "Learning in Optical Neural Networks," Optical Computing (1991).

[144] Y. Qiao and D. Psaltis, "Local learning algorithm for optical neural networks," Applied Optics **31**, 3285 (1992).

[145] C. Denz, Optical neural networks (Springer Science & Business Media, 2013).

[146] N. C. Thompson, K. Greenewald, K. Lee, G. F. Manso, et al., "The computational limits of deep learning," arXiv:2007.05558.

[147] T. Haigh, "Between the Booms: AI in Winter," Communications of the ACM **67**, 18 (2024).

[148] R. S. Tucker, "The role of optics in computing," Nature Photonics **4**, 405 (2010).

[149] D. A. Miller, "Are optical transistors the logical next step?" Nature Photonics **4**, 3 (2010).

[150] B. J. Shastri, A. N. Tait, T. Ferreira de Lima, W. H. Pernice, H. Bhaskaran, C. D. Wright, and P. R. Prucnal, "Photonics for artificial intelligence and neuromorphic computing," Nature Photonics **15**, 102 (2021).

[151] Y. Shen, N. C. Harris, S. Skirlo, M. Prabhu, T. Baehr-Jones, M. Hochberg, X. Sun, S. Zhao, H. Larochelle, D. Englund, et al., "Deep learning with coherent nanophotonic circuits," Nature Photonics **11**, 441 (2017).

[152] H. Zhang, M. Gu, X. Jiang, J. Thompson, H. Cai, S. Paesani, R. Santagati, A. Laing, Y. Zhang, M.-H. Yung, et al., "An optical neural chip for implementing complex-valued neural network," Nature Communications **12**, 457 (2021).

[153] T. W. Hughes, M. Minkov, Y. Shi, and S. Fan, "Training of photonic neural networks through in situ backpropagation and gradient measurement," Optica **5**, 864 (2018).

[154] M. Reck, A. Zeilinger, H. J. Bernstein, and P. Bertani, "Experimental realization of any discrete unitary operator," Physical Review Letters **73**, 58 (1994).

[155] W. R. Clements, P. C. Humphreys, B. J. Metcalf, W. S. Kolthammer, and I. A. Walmsley, "Optimal design for universal multiport interferometers," Optica **3**, 1460 (2016).

[156] A. N. Tait, T. F. De Lima, E. Zhou, A. X. Wu, M. A. Nahmias, B. J. Shastri, and P. R. Prucnal, "Neuromorphic photonic networks using silicon photonic weight banks," Scientific Reports **7**, 7430 (2017).

[157] B. Bai, Q. Yang, H. Shu, L. Chang, F. Yang, B. Shen, Z. Tao, J. Wang, S. Xu, W. Xie, et al., "Microcomb-based integrated photonic processing unit," Nature Communications **14**, 66 (2023).

[158] J. Feldmann, N. Youngblood, M. Karpov, H. Gehring, X. Li, M. Stappers, M. Le Gallo, X. Fu, A. Lukashchuk, A. S. Raja, et al., "Parallel convolutional processing using an integrated photonic tensor core," Nature **589**, 52 (2021).

[159] J. Feldmann, N. Youngblood, C. D. Wright, H. Bhaskaran, and W. H. Pernice, "All-optical spiking neurosynaptic networks with self-learning capabilities," Nature **569**, 208 (2019).

[160] S. Zarei, M.-r. Marzban, and A. Khavasi, "Integrated photonic neural network based on silicon metalines," Optics Express **28**, 36668 (2020).

[161] T. Fu, Y. Zang, Y. Huang, Z. Du, H. Huang, C. Hu, M. Chen, S. Yang, and H. Chen, "Photonic machine learning with on-chip diffractive optics," Nature Communications **14**, 70 (2023).

[162] E. Goi, X. Chen, Q. Zhang, B. P. Cumming, S. Schoenhardt, H. Luan, and M. Gu, "Nanoprinted high-neuron-density optical linear perceptrons performing near-infrared inference on a CMOS chip," Light: Science & Applications **10**, 40 (2021).

[163] X. Luo, Y. Hu, X. Ou, X. Li, J. Lai, N. Liu, X. Cheng, A. Pan, and H. Duan, "Metasurface-enabled on-chip multiplexed diffractive neural networks in the visible," Light: Science & Applications **11**, 158 (2022).

[164] Z. Wang, L. Chang, F. Wang, T. Li, and T. Gu, "Integrated photonic metasystem for image classifications at telecommunication wavelength," Nature Communications **13**, 2131 (2022).

[165] Y. Qu, H. Zhu, Y. Shen, J. Zhang, C. Tao, P. Ghosh, and M. Qiu, "Inverse design of an integrated-nanophotonics optical neural network," Science Bulletin **65**, 1177 (2020).

[166] J. Moughames, X. Porte, M. Thiel, G. Ulliac, M. Jacquot, L. Larger, M. Kadic, and D. Brunner, "Three dimensional waveguide-interconnects for scalable integration of photonic neural networks," arXiv:1912.08203.

[167] F. Ashtiani, A. J. Geers, and F. Aflatouni, "An on-chip photonic deep neural network for image classification," Nature **606**, 501 (2022).

[168] Z. Xu, T. Zhou, M. Ma, C. Deng, Q. Dai, and L. Fang, "Large-scale photonic chiplet Taichi empowers 160-TOPS/W artificial general intelligence," Science **384**, 202 (2024).

[169] A. Montes McNeil, Y. Li, A. Zhang, M. Moebius, and Y. Liu, "Fundamentals and recent developments of free-space optical neural networks," Journal of Applied Physics **136** (2024).

[170] J. Hu, D. Mengu, D. C. Tzarouchis, B. Edwards, N. Engheta, and A. Ozcan, "Diffractive optical computing in free space," Nature Communications **15**, 1525 (2024).

[171] G. Wetzstein, A. Ozcan, S. Gigan, S. Fan, D. Englund, M. Soljačić, C. Denz, D. A. Miller, and D. Psaltis, "Inference in artificial intelligence with deep optics and photonics," Nature **588**, 39 (2020).

[172] H. Chen, J. Feng, M. Jiang, Y. Wang, J. Lin, J. Tan, and P. Jin, "Diffractive deep neural networks at visible wavelengths," Engineering **7**, 1483 (2021).

[173] D. Mengu and A. Ozcan, "All-optical phase recovery: diffractive computing for quantitative phase imaging," Advanced Optical Materials **10**, 2200281 (2022).

[174] Y. Luo, Y. Zhao, J. Li, E. Çetintaş, Y. Rivenson, M. Jarrahi, and A. Ozcan, "Computational imaging without a computer: seeing through random diffusers at the speed of light," elight **2**, 4 (2022).

[175] T. Yan, J. Wu, T. Zhou, H. Xie, F. Xu, J. Fan, L. Fang, X. Lin, and Q. Dai, "Fourier-space diffractive deep neural network," Physical Review Letters **123**, 023901 (2019).

[176] S. Jahani and Z. Jacob, "All-dielectric metamaterials," Nature Nanotechnology **11**, 23 (2016).

[177] N. Yu, P. Genevet, M. A. Kats, F. Aieta, J.-P. Tetienne, F. Capasso, and Z. Gaburro, "Light propagation with phase discontinuities: generalized laws of reflection and refraction," Science **334**, 333–337 (2011).

[178] A. V. Kildishev, A. Boltasseva, and V. M. Shalaev, "Planar photonics with metasurfaces," Science **339**, 1232009 (2013).

[179] B. Xiong, L. Deng, R. Peng, and Y. Liu, "Controlling the degrees of freedom in metasurface designs for multi-functional optical devices," Nanoscale Advances **1**, 3786 (2019).

[180] A. Arbabi, Y. Horie, M. Bagheri, and A. Faraon, "Dielectric metasurfaces for complete control of phase and polarization with subwavelength spatial resolution and high transmission," Nature Nanotechnology **10**, 937 (2015).

[181] Y. Wang, A. Yu, Y. Cheng, and J. Qi, "Matrix diffractive deep neural networks merging polarization into meta-devices," Laser & Photonics Reviews **18**, 2300903 (2024).

[182] B. Xiong, Y. Liu, Y. Xu, L. Deng, C.-W. Chen, J.-N. Wang, R. Peng, Y. Lai, Y. Liu, and M. Wang, "Breaking the limitation of polarization multiplexing in optical metasurfaces with engineered noise," Science **379**, 294 (2023).

[183] M. Khorasaninejad, W. T. Chen, R. C. Devlin, J. Oh, A. Y. Zhu, and F. Capasso, "Metalenses at visible wavelengths: Diffraction-limited focusing and subwavelength resolution imaging," Science **352**, 1190–1194 (2016).

[184] J. Li, T. Gan, B. Bai, Y. Luo, M. Jarrahi, and A. Ozcan, "Massively parallel universal linear transformations using a wavelength-multiplexed diffractive optical network," Advanced Photonics **5**, 016003–016003 (2023).

[185] J. Li, Y.-C. Hung, O. Kulce, D. Mengu, and A. Ozcan, "Polarization multiplexed diffractive computing: all-optical implementation of a group of linear transformations through a polarization-encoded diffractive network," Light: Science & Applications **11**, 153 (2022).

[186] W. Ma, Y. Xu, B. Xiong, L. Deng, R.-W. Peng, M. Wang, and Y. Liu, "Pushing the limits of functionality-multiplexing capability in metasurface design based on statistical machine learning," Advanced Materials **34**, 2110022 (2022).

[187] S. Turtaev, I. T. Leite, K. J. Mitchell, M. J. Padgett, D. B. Phillips, and T. Čižmár, "Comparison of nematic liquid-crystal and DMD based spatial light modulation in complex photonics," Optics Express **25**, 29874 (2017).

[188] M. Miscuglio, Z. Hu, S. Li, J. K. George, R. Capanna, H. Dalir, P. M. Bardet, P. Gupta, and V. J. Sorger, "Massively parallel amplitude-only Fourier neural network," Optica **7**, 1812–1819 (2020).

[189] Y. Zuo, B. Li, Y. Zhao, Y. Jiang, Y.-C. Chen, P. Chen, G.-B. Jo, J. Liu, and S. Du, "All-optical neural network with nonlinear activation functions," Optica **6**, 1132 (2019).

[190] T. Wang, M. M. Sohoni, L. G. Wright, M. M. Stein, S.-Y. Ma, T. Onodera, M. G. Anderson, and P. L. McMahon, "Image sensing with multilayer nonlinear optical neural networks," Nature Photonics **17**, 408 (2023).

[191] J. Chang, V. Sitzmann, X. Dun, W. Heidrich, and G. Wetzstein, "Hybrid optical-electronic convolutional neural networks with optimized diffractive optics for image classification," Scientific Reports **8**, 1 (2018).

[192] D. Mengu, Y. Zhao, N. T. Yardimci, Y. Rivenson, M. Jarrahi, and A. Ozcan, "Misalignment resilient diffractive optical networks," Nanophotonics **9**, 4207–4219 (2020).

[193] G. Qu, G. Cai, X. Sha, Q. Chen, J. Cheng, Y. Zhang, J. Han, Q. Song, and S. Xiao, "All-dielectric metasurface empowered optical-electronic hybrid neural networks," Laser & Photonics Reviews **16**, 2100732 (2022).

[194] T. Zhou, X. Lin, J. Wu, Y. Chen, H. Xie, Y. Li, J. Fan, H. Wu, L. Fang, and Q. Dai, "Large-scale neuromorphic optoelectronic computing with a reconfigurable diffractive processing unit," Nature Photonics **15**, 367 (2021).

[195] H. Zheng, Q. Liu, Y. Zhou, I. I. Kravchenko, Y. Huo, and J. Valentine, "Meta-optic accelerators for object classifiers," Science Advances **8**, eabo6410 (2022).

[196] R. Pratap, V. Agarwal, and R. Singh, "Review of various available spice simulators," in 2014 International Conference on Power, Control and Embedded Systems (ICPCES) (IEEE, 2014), p. 1.

[197] X. Yang, Y. Lin, T. Wu, Z. Yan, Z. Chen, H.-C. Kuo, and R. Zhang, "An overview on the principle of inkjet printing technique and its application in micro-display for augmented/virtual realities," Opto-Electronic Advances **5**, 210123 (2022).

[198] W. Chau, S. Wong, and S. Wan, "A critical analysis of dithering algorithms for image processing," in IEEE TENCON'90: 1990 IEEE Region 10 Conference on Computer and Communication Systems. Conference Proceedings (IEEE, 1990), p. 309.

[199] Stabilize Your Transimpedance Amplifier, en, 2025.

[200] B. Brown, IMPLEMENTATION AND APPLICATIONS OF CURRENT SOURCES AND CURRENT RECEIVERS, en, 1990.

[201] M. Zamora, Precision current sources and sinks using voltage references, en, 2020.

[202] Transimpedance Amplifier Noise Considerations, en, 2018.

[203] G. Di Patrizio Stanchieri, A. De Marcellis, G. Battisti, M. Faccio, E. Palange, and U. Guler, "A 1.8 V low-power low-noise high tunable gain TIA for CMOS integrated optoelectronic biomedical applications," Electronics **11**, 1271 (2022).

[204] F. Khateb, F. Kacar, N. Khatib, and D. Kubanek, "High-precision differential-input buffered and external transconductance amplifier for low-voltage low-power applications," Circuits, Systems, and Signal Processing **32**, 453 (2013).

[205] Nvidia dgx platform, Nvidia, 2024.

[206] Nvidia ada gpu architecture, Nvidia, 2023.

[207] Tpu v4, Google Cloud, 2024.

[208] N. Jouppi and D. Patterson, Google's cloud tpu v4 provides exaflops-scale ml with industry-leading efficiency, Google Cloud Blog, 2023.

[209] J. Feldmann, N. Youngblood, M. Karpov, H. Gehring, X. Li, M. Stappers, M. Le Gallo, X. Fu, A. Lukashchuk, A. S. Raja, et al., "Parallel convolutional processing using an integrated photonic tensor core," Nature **589**, 52 (2021).

[210] X. Xu, M. Tan, B. Corcoran, J. Wu, A. Boes, T. G. Nguyen, S. T. Chu, B. E. Little, D. G. Hicks, R. Morandotti, et al., "11 TOPS photonic convolutional accelerator for optical neural networks," Nature **589**, 44 (2021).

[211] S. Hasegawa and Y. Hayasaki, "Holographic vector wave femtosecond laser processing," International Journal of Optomechatronics **8**, 73 (2014).

[212] I. Kim, G. Yoon, J. Jang, P. Genevet, K. T. Nam, and J. Rho, "Outfitting next generation displays with optical metasurfaces," ACS Photonics **5**, 3876 (2018).

[213] S. Residori, U. Bortolozzo, and J. Huignard, "Liquid crystal light valves as optically addressed liquid crystal spatial light modulators: optical wave mixing and sensing applications," Liquid Crystals Reviews **6**, 1 (2018).

[214] R. Rice, W. Li, and G. Moddel, "High-speed optically-addressed spatial light modulator for optical computing," Optical Computing, MF1 (1989).

[215] J. Huignard, "Spatial light modulators and their applications," Journal of Optics **18**, 181 (1987).

[216] G. Moddel, K. Johnson, W. Li, R. Rice, L. Pagano-Stauffer, and M. Handschy, "High-speed binary optically addressed spatial light modulator," Applied Physics Letters **55**, 537 (1989).

[217] R. Xu, P. Lv, F. Xu, and Y. Shi, "A survey of approaches for implementing optical neural networks," Optics & Laser Technology **136**, 106787 (2021).

[218] S. Colburn, Y. Chu, E. Shilzerman, and A. Majumdar, "Optical frontend for a convolutional neural network," Applied Optics **58**, 3179 (2019).

[219] H. Kato, D. Beker, M. Morariu, T. Ando, T. Matsuoka, W. Kehl, and A. Gaidon, "Differentiable rendering: A survey," arXiv:2006.12057.

[220] C. Wang, N. Chen, and W. Heidrich, "Lens design optimization by back-propagation," in International Optical Design Conference (Optica Publishing Group, 2021), 120781O.

[221] Y. Nie, J. Zhang, R. Su, and H. Ottevaere, "Freeform optical system design with differentiable three-dimensional ray tracing and unsupervised learning," Optics Express **31**, 7450 (2023).

[222] C. Wang, "Differentiable ray tracing for optical design and modeling," FM3A–2 (2023).

[223] Z. Fang, R. Chen, J. E. Froech, Q. A. Tanguy, A. I. Khan, X. Wu, V. Tara, A. Manna, D. Sharp, C. Munley, et al., "Nonvolatile phase-only transmissive spatial light modulator with electrical addressability of individual pixels," ACS Nano **18**, 11245 (2024).

[224] J.-A. Pan, Z. Rong, Y. Wang, H. Cho, I. Coropceanu, H. Wu, and D. V. Talapin, "Direct optical lithography of colloidal metal oxide nanomaterials for diffractive optical elements with 2pi phase control," Journal of the American Chemical Society **143**, 2372–2383 (2021).

[225] R. Gounella, G. M. Ferreira, M. L. Amorim, J. N. Soares Jr, and J. P. Carmo, "A review of optical sensors in CMOS," Electronics **13**, 691 (2024).

[226] T. Francois and D. Christophe, "Challenges and Solutions for the Fabrication of CMOS-driven Microled Displays," in Proceedings of the International Display Workshops (International Display Workshop, 2022), p. 871.

[227] J. Park, D.-M. Geum, W. Baek, J. Shieh, and S. Kim, "Monolithic 3D sequential integration realizing 1600-PPI red micro-LED display on Si CMOS driver IC," in 2022 IEEE Symposium on VLSI Technology and Circuits (VLSI Technology and Circuits) (IEEE, 2022), pp. 383–384.

[228] J. Park, D.-M. Geum, W. Baek, J. Shieh, and S. Kim, "Monolithic 3D sequential integration realizing 1600-PPI red micro-LED display on Si CMOS driver IC," in Symposium on VLSI Technology and Circuits (VLSI Technology and Circuits) (IEEE, 2022), pp. 383–384.

[229] G. Brown, M. Ravindran, R. Burns, and N. Miller, "How to deploy AI software to self driving cars," in Proceedings of the International Workshop on OpenCL (2019), p. 1.

LIST OF FIGURES

LIST OF TABLES