

Compressed meta-optical encoder for image classification

Anna Wirth-Singh^{a,*†} Jinlin Xiang^{b,†} Minhó Choi^{b,†} Johannes E. Fröch^{a,b} Luocheng Huang^b Shane Colburn^b Eli Shlizerman^{b,c} and Arka Majumdar^{a,b,*}

^aUniversity of Washington, Department of Physics, Seattle, Washington, United States

^bUniversity of Washington, Department of Electrical and Computer Engineering, Seattle, Washington, United States

^cUniversity of Washington, Department of Applied Mathematics, Seattle, Washington, United States

Abstract. Optical and hybrid convolutional neural networks (CNNs) recently have become of increasing interest to achieve low-latency, low-power image classification, and computer-vision tasks. However, implementing optical nonlinearity is challenging, and omitting the nonlinear layers in a standard CNN comes with a significant reduction in accuracy. We use knowledge distillation to compress modified AlexNet to a single linear convolutional layer and an electronic backend (two fully connected layers). We obtain comparable performance with a purely electronic CNN with five convolutional layers and three fully connected layers. We implement the convolution optically via engineering the point spread function of an inverse-designed meta-optic. Using this hybrid approach, we estimate a reduction in multiply-accumulate operations from 17M in a conventional electronic modified AlexNet to only 86 K in the hybrid compressed network enabled by the optical front end. This constitutes over 2 orders of magnitude of reduction in latency and power consumption. Furthermore, we experimentally demonstrate that the classification accuracy of the system exceeds 93% on the MNIST dataset of handwritten digits.

Keywords: neural network; meta-optics; image classification; knowledge distillation; optical computing.

Received Nov. 13, 2024; accepted for publication Jan. 14, 2025; published online Feb. 25, 2025.

© The Authors. Published by SPIE and CLP under a Creative Commons Attribution 4.0 International License. Distribution or reproduction of this work in whole or in part requires full attribution of the original publication, including its DOI.

[DOI: [10.1117/1.APN.4.2.026009](https://doi.org/10.1117/1.APN.4.2.026009)]

1 Introduction

Convolutional neural networks (CNNs) represent a significant milestone in image classification, recognition, and tracking.¹ CNNs, for example, AlexNet, are composed of several convolutional layers that adaptively learn spatial representations from input images. Although powerful, the convolution operation is computationally expensive, leading to high latency and power consumption. In fact, it has been estimated that ~80% of the total runtime of CNNs is used in performing convolution operations.² Reducing this latency and, as a result, power consumption has become an active area of research, with multiple works proposing free-space optical systems as a solution.^{3–8} Beyond reducing latency and power consumption, optical information processing offers advantages including high bandwidth,

spatial parallelism, and low-loss transmission, which have led to a surge of interest in the field.³

For decades, it has been known that a $4f$ system can be used to perform convolutions optically by placing an appropriate filter at the Fourier plane of the lens.^{4,9–11} This was demonstrated in 2018⁴ using a diffractive optical element as the filtering element and traditional refractive lenses composing the $4f$ system. Spatial light modulators⁶ and digital micromirror devices¹² can also be used as the filtering element. However, one drawback of the Fourier-based $4f$ approach is that it requires three elements (two lenses and a spatial filter), resulting in a bulky optical system with a greater propensity for misalignments than single-element optical systems. Such misalignments from each optical convolutional layer cannot be ignored even when weights are trained with noisy inputs (more details in the [Supplementary Material](#)). In addition, the filtering optics must be contained within a compact area at the focal plane in the $4f$ system, which limits parallel processing ability unless creative measures are taken, such as utilizing naturally present diffraction orders¹² or lenslet arrays.⁵

*Address all correspondence to Anna Wirth-Singh, annaw77@uw.edu; Arka Majumdar, arka@uw.edu

†These authors contributed equally to this work.

Advantageously, the convolution operation can also be performed using free-space optics and requires only a single element. The resultant image produced by any optics is the input convolved with the point spread function (PSF) of the optics.^{13–15} Therefore, by engineering optics to produce a particular PSF, convolution can be performed optically simply via passing light through the optics. Further, by passing the input through several of these optics in parallel, multiple convolution operations can be performed simultaneously at the speed of light.^{5,16} This approach leverages the inherent parallelism of light enabling the passive processing of a vast amount of data without increasing computation time.^{3,5,12} This unique optical capability circumvents scalability issues when handling high-resolution images in traditional electronic-based CNN systems.

However, a challenge to all optically implemented CNN approaches is that nonlinear layers are interspersed with the linear layers. For example, AlexNet consists of five convolutional layers followed by three fully connected layers.¹ Specifically, each convolutional layer in the architecture utilizes the rectified linear unit (ReLU) as its nonlinear activation function, followed by the local response normalization and max pooling layer, ensuring an effective mechanism for spatial hierarchy extraction. Therefore, nonlinearity is consistently applied across all layers, serving as a foundational element of the network design to enhance its learning capability. Without the nonlinear activations (e.g., ReLU), the classification accuracy of the CNN drops by $\sim 20\%$.¹⁷ The nonlinear layers cannot be implemented using simple lens-like optics; to implement them optically, some physical nonlinearity must be introduced, for instance, using an atomic vapor cell^{6,18} or image intensifier.¹⁹ Hybrid approaches involving repeated transduction of the signal to perform linear operations in optics and nonlinear operations in electronics provide little benefit due to large latency and power consumption in signal transduction.^{5,11,19} Implementing only one of many required convolution operations does not provide much benefit in terms of speed and latency. Alternatively, there have been recent breakthroughs in using end-to-end designs for physical and hybrid network designs that perform image classification or other tasks without explicitly using convolution.^{3,20–23} Such an approach can effectively implement multiple linear layers in one optical front end. Although novel, these end-to-end neural networks are computationally expensive to train and are applicable only to the physical system for which they were specifically designed. In another approach, all-optical classifiers composed of several layers of diffractive optics have achieved reasonable classification accuracy using coherent illumination at terahertz^{24,25} and near-infrared wavelengths.²⁶ Further, a metasurface-based on-chip diffractive neural network has also been demonstrated.²⁷ However, these all-optical approaches are limited to implementing only linear operations.

In this work, we experimentally demonstrate a hybrid optical–electronic CNN consisting of a single optical convolution layer with an electronic single fully connected layer to achieve similar accuracy as AlexNet on handwritten digit classification tasks. To overcome the limitation of the absence of optical nonlinearity, we apply knowledge distillation (KD) to remove the nonlinear layers and compress multiple layers into a single linear layer.¹⁷ KD (more details in Sec. 4) circumvents the need for nonlinearity without a significant reduction in the performance by transferring knowledge from a larger, pretrained network (the “teacher” network) to a more compact network (the “student” network). Here, we use a modified AlexNet, denoted

AlexNet-Mod, as the teacher network and a single convolutional layer coupled with a single fully connected layer as the student network. We use this architecture to demonstrate a hybrid meta-optical platform, wherein an optical front end based on a single meta-optic performs the linear convolution operation, followed by an electronic backend that contains a linear calibration layer and a fully connected layer. In such a way, the most computationally expensive operation is performed optically to leverage the benefits of optical computing, namely, high spatial bandwidth and low power consumption. The use of a single meta-optic layer drastically simplifies the experimental setup and provides a compact geometry.

The optical front end of this network is realized using inverse-designed meta-optics. The meta-optics are arrays of subwavelength scatterers that act as phase masks, imparting spatially coded phase shifts to incident light. Here, we design meta-optics to perform the desired convolutional steps of the CNN by engineering the PSF. We fabricate and experimentally validate the performance of the designed optics using incoherent green-light illumination from a light-emitting diode (LED), centered at 525 nm. Further, we experimentally demonstrate the classification accuracy of the entire hybrid CNN on the MNIST data set. The hybrid CNN is described in Fig. 1, where we compare the architectures of multilayer electronic CNNs, compressed electronic CNNs (a linear single-layer CNN), and our hybrid system. The number of multiply-accumulate (MAC) operations in the entire network is reduced by over 2 orders of magnitude by compressing multiple convolutional layers into a single layer and implementing them optically. The classification accuracy of the compressed hybrid CNN is reduced by only 5% from AlexNet-Mod (98% accuracy) to achieve 93% classification accuracy on the MNIST data set.

2 Results

In each convolutional layer of a CNN, an optimized kernel is convolved with the input to generate a feature map, which is then passed to the next layer. Using the KD approach, we optimize eight convolutional kernels (each 6 pixel \times 6 pixel in size) for the MNIST data set of handwritten digits. The selected number and size of the kernels are based on previous experimental results (details shown in the [Supplementary Material](#)).²⁸ As described in Sec. 2.1, we design the optics to implement these optimized convolutional kernels and combine them with an electronic backend for image classification.

2.1 Compressing Multiple Convolutional Layers Using KD

To select the ideal network architecture for the linear optical–electronic hybrid system, underfitting and overfitting issues must be avoided. Smaller networks face underfitting concerns, particularly in optical settings that limit the system to just eight kernels and remove nonlinear functions.²⁹ This is in stark contrast to AlexNet, which uses over 300 kernels.¹ However, complex models are prone to overfitting, and practical issues such as fabrication noise and misalignments introduce further challenges for optically implementing a large number of kernels.

Therefore, to obtain a balanced network that could be implemented optically, we use KD to compress an AlexNet-Mod as a base model. AlexNet-Mod consists of five convolutional layers and three fully connected layers (eight total layers) that we compress to the desired structure of one convolutional layer and one

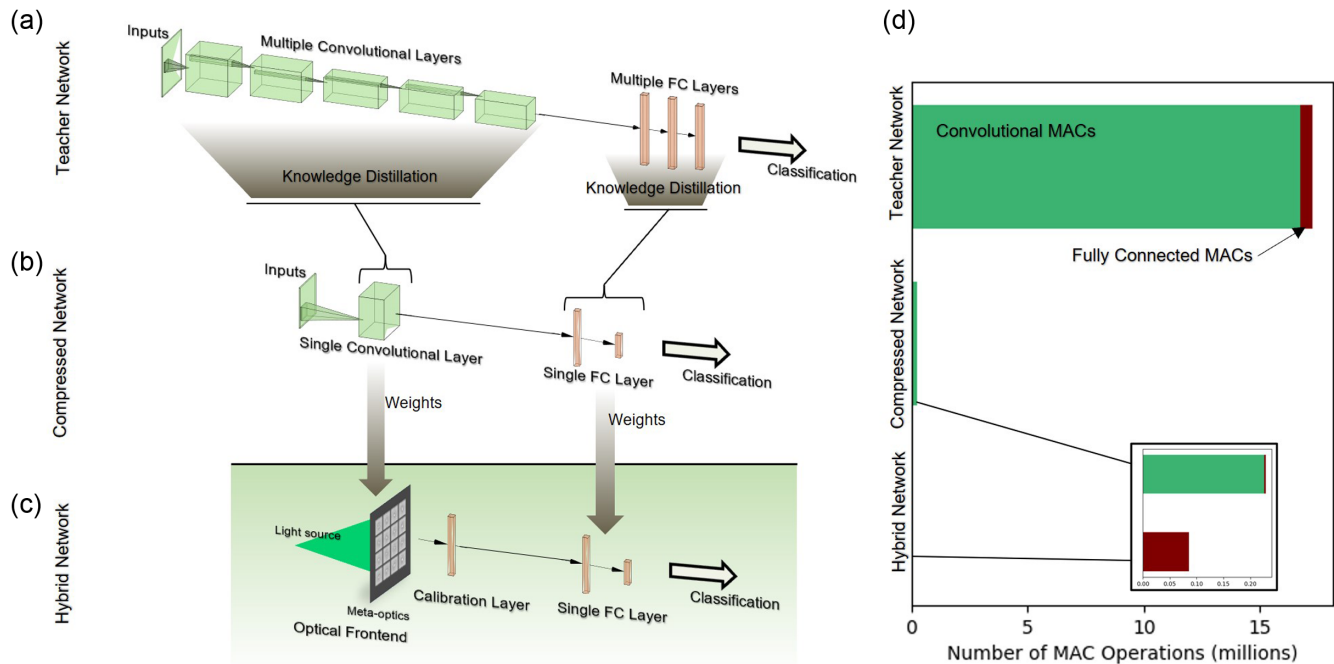


Fig. 1 Schematic of CNNs for image classification tasks. (a) All-electronic multi-layered CNN. (b) All-electronic compressed CNN. (c) Hybrid CNN that combines an optical meta-optical front end and electronic backend. (d) Number of MAC operations of each network configuration, with convolutional MACs in green and fully connected MACs in brown.

fully connected layer (two layers). The KD approach assumes that the teacher network is already trained and performs the desired task with high accuracy; in this case, we use AlexNet-Mod as the teacher network, which achieves $98.9\% \pm 0.33\%$ on training and $98.4\% \pm 0.32\%$ on testing classification accuracy MNIST data sets over repeated trials. In addition, AlexNet-Mod employs nonlinear activation functions (ReLU) to optimize performance; these are circumvented by KD for a result that is compatible with our optical setting. In the compressed network, we limit the number of kernels in the compressed convolutional layer to eight, and each kernel is $6\text{ pixel} \times 6\text{ pixel}$ in size. After training, the compressed electronic network achieves an approximate classification accuracy of 96% on both training and testing data sets.

2.2 Optical Convolution Using Meta-Optics

The optical component of this network is implemented using inverse-designed meta-optics. We used the Gerchberg–Saxton (GS) phase-retrieval method³⁰ to design phase masks that correspond to the optimized convolutional kernels. Specifically, each suboptic is designed to have a PSF that resembles the convolutional kernel. As electronic convolutional kernels include both positive and negative values, we separated each kernel into positive and negative parts and designed meta-optics for each. The positive and negative images were computationally subtracted afterward to produce the net convolution. Two-dimensional phase maps of an example set of suboptics are shown in Fig. 2(c). The phase maps were implemented by silicon nitride pillars that are 750 nm tall but with varied widths corresponding to their relative phase delay. Scanning electron microscope (SEM) images of the fabricated optics are also shown for comparison, highlighting the fabrication quality.

All 16 optics (corresponding to 8 convolutional kernels) were fabricated on a single substrate (more details in Sec. 4), allowing convolution from 16 different kernels in a single image capture. A photograph of the fabricated meta-optic is shown in Fig. 2(b). Each kernel meta-optic is $470\ \mu\text{m} \times 470\ \mu\text{m}$, with a center-to-center distance of $705\ \mu\text{m}$. The optics are arranged in two rows of eight optics for a total footprint of $5.6\ \text{mm} \times 1.4\ \text{mm}$.

To verify the performance of the optics, we measured the PSF of each suboptic using a single-mode fiber as a point light source, as shown in Fig. 2(a). The PSFs from all 16 suboptics are simultaneously captured by the two-dimensional CMOS camera (more details in Sec. 4). Accordingly, the convolved images from all 16 suboptics are also captured at the same time when we replaced the single-mode fiber with the display, as described in Fig. 2(a). Exemplary electronic convolutional kernels, which represent the ground truth PSFs for the optically implemented kernels, are shown in Fig. 2(d). We also present the simulated PSFs from the meta-optics using angular spectrum propagation.^{31,32} The experimentally measured PSFs shown in Fig. 2(d) match the simulated PSFs well, confirming fabrication accuracy. However, due to the constraints on physically realizable PSFs, there are notable differences between the ground-truth PSFs and experimentally measured PSFs. To correct for these differences, as well as slight noise and misalignments in the optical system, we introduce a calibration layer to the computational backend, further discussed in Sec. 4.2.

2.3 Hybrid Network Classification Results

To address optical noise and misalignment affecting image classification performance, an additional calibration layer is introduced to adjust optical representations for compatibility

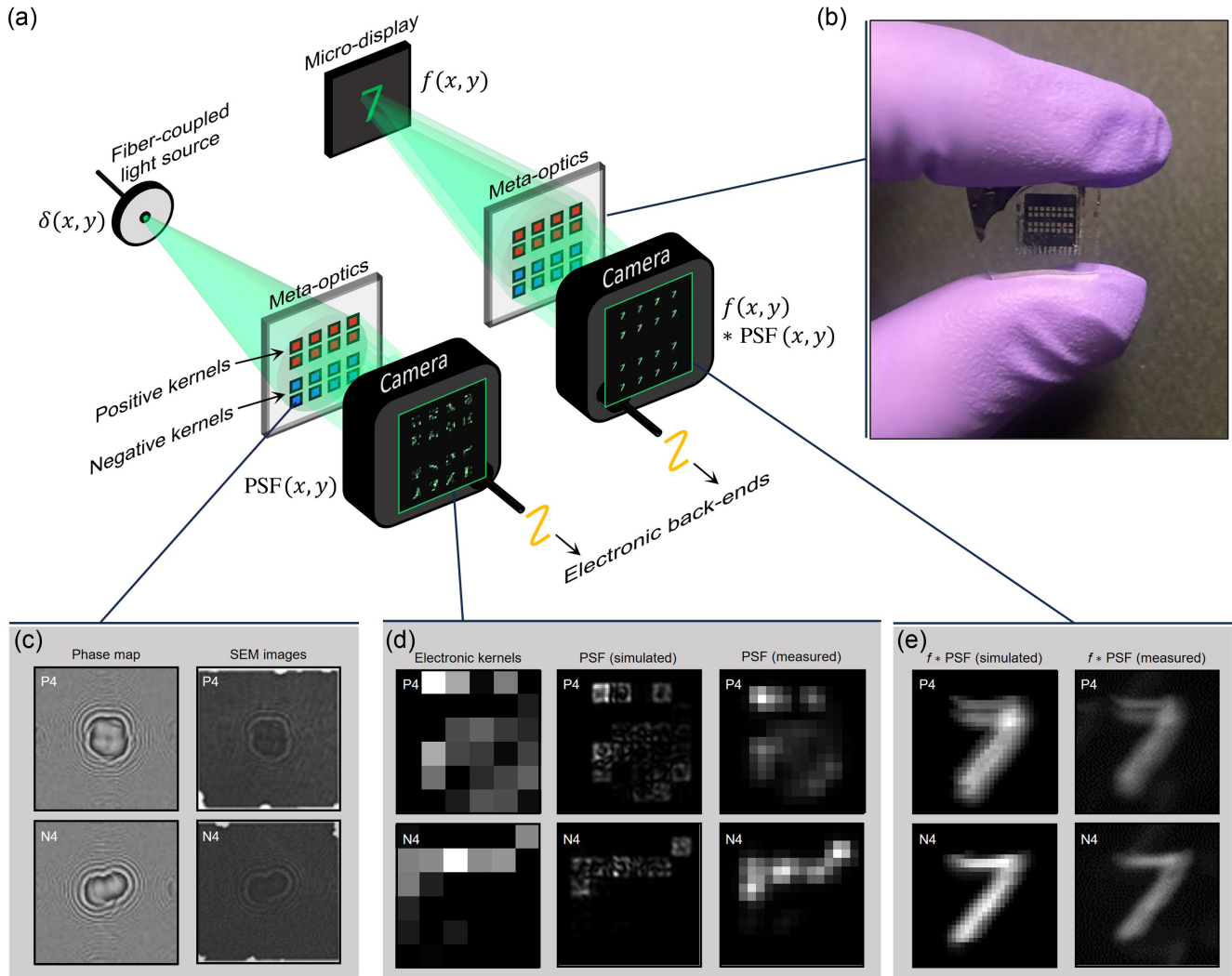


Fig. 2 Schematic of the optical system. (a) PSF measurement setup using a monochromatic point light source (left) and optical convolution measurements using a micro-LED display (right). (b) Photograph of the fabricated meta-optics. The meta-optic contains 16 different suboptics, spatially distributed in a single layer, operating in parallel for classification tasks. (c) Phase maps and SEM images of exemplary suboptics corresponding to the positive and negative parts of a particular convolutional kernel. (d) Positive and negative parts of an example convolutional kernel (left), the corresponding PSF simulation (middle), and experiment (right). (e) Simulated electronic output (left) and optical experiment (right) convolved output for the example kernel, for the case of an input “7” from MNIST.

with the computational backend. Specifically, this calibration layer is a single fully connected neural network layer with an input dimension of 288 and an output dimension of 288. We fine-tune it with only 10% of the training data set and ensure that the computational backend does not need to be retrained. Therefore, the electronic backend of the hybrid network consists of the calibration layer followed by the original compressed electronic backend.

We compare three CNN architectures (AlexNet-Mod, the compressed electronic network, and the hybrid optical–electronic network) in Table 1. The MAC for the convolutional layer depends on the image size (H, W), kernel size (k^2), number of kernels (c_{in}), and input channels (#kernels), and the MACs are calculated as $c_{in}HWk^2\#\text{kernels}$. The MAC for a fully connected layer depends on the input size (m) and output size

(n), and the MACs are calculated as mn . The AlexNet-Mod achieves a classification accuracy exceeding 98% on both training and testing data sets. The number of MAC operations of this network is 17 million with 8-bit precision (details are shown in the [Supplementary Material](#), Sec. S1). The compressed electronic CNN achieves greater than 96% accuracy; this 2% decline reflects the inherent challenges of compressing multiple layers into a single layer. Primarily due to the compression of the convolution layers, the number of MAC operations is reduced to 228,672. The hybrid network, which integrates the optical convolution layer with the calibration layer and single fully connected layer electronic backend, experimentally achieves a classification accuracy of 93.9% ($\pm 0.25\%$) and 93.4% ($\pm 0.22\%$) on the training and testing data sets, respectively, and requires only 85,824 MAC operations, which is 0.5% and 37% of that

Table 1 Classification results.

Network architecture	Train (%)	Test (%)	MAC operations
AlexNet-Mod	98.9 ± 0.33	98.4 ± 0.32	17,323,520
Compressed electronic CNN (without KD)	84.2 ± 0.47	82.1 ± 0.69	228,672
Compressed electronic CNN (KD)	97.2 ± 0.35	96.2 ± 0.29	228,672
Hybrid CNN (KD)	93.9 ± 0.25	93.4 ± 0.22	85,824

required for AlexNet-Mod and the compressed electronic networks, respectively.

Figure 3 illustrates the confusion matrices for three neural network configurations. Each matrix visually represents the model's tested performance across different classes (in this case, digits labeled 0 through 9), with the true labels on the rows and the predicted labels on the columns. The multilayer electronic CNN, AlexNet-Mod, displays high values along the diagonal,

exceeding 98.1% accuracy in each class. The compressed electronic CNN, while having a slight decline in diagonal values, still demonstrates robust classification accuracy with a minimum of 94.3%. The hybrid network exhibits a more diverse range of values along the diagonal, with some classes exhibiting lower predictive accuracy compared with the compressed electronic network. We attribute this slight decline to noise in the optical experiment, which may be due to optics fabrication,

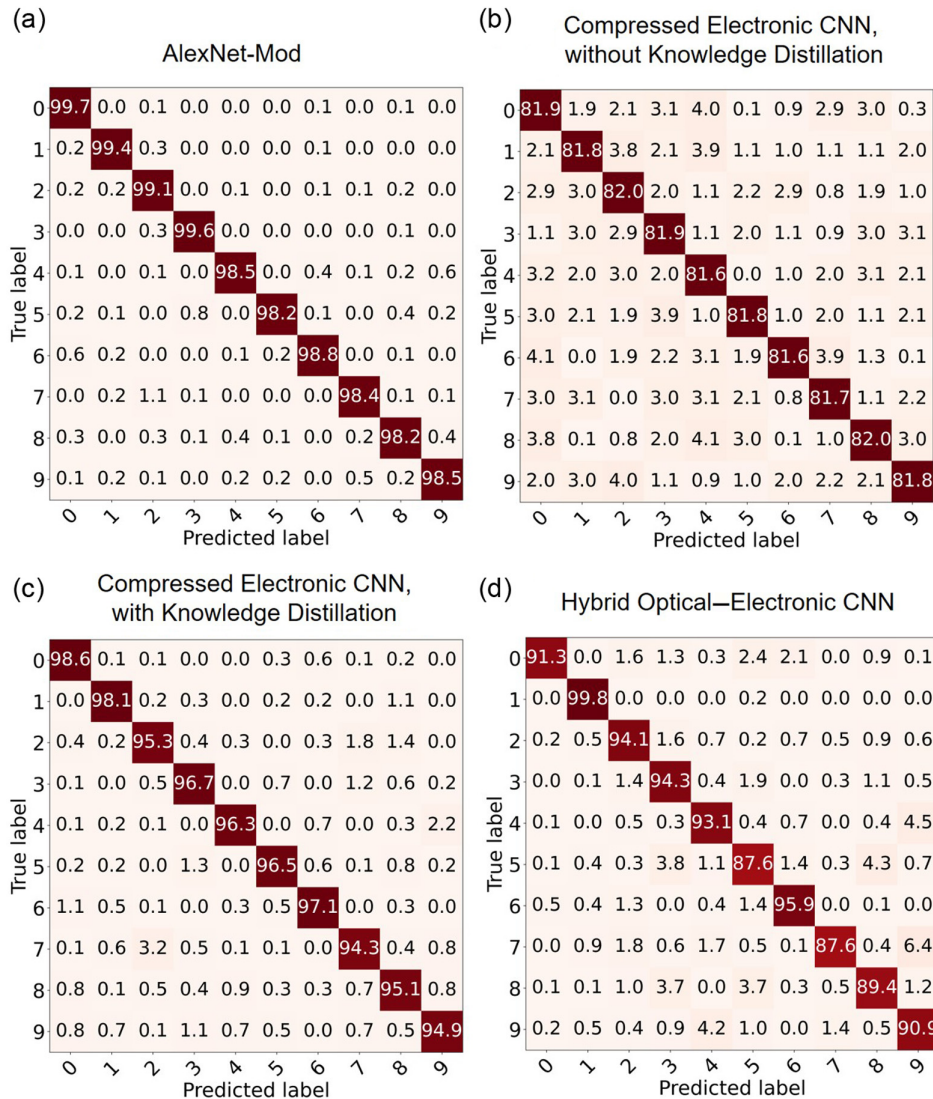


Fig. 3 Confusion matrices for different network architectures. (a) Classification results for AlexNet-Mod (multiple-layer electronic CNN). (b) Classification results for the all-electronic CNN compressed without using KD. (c) Classification results for the all-electronic CNN compressed with KD. (d) Classification results for the hybrid optical-electronic CNN.

Table 2 Network ablation study.

Configuration	Train accuracy	Test accuracy
Backend only	89%	87%
Calibration + backend	84%	80%
Optics + calibration + backend	94% (+5%/+10%)	93% (+6%/+13%)

camera sensor noise, and optical noise due to vibrations that cannot be fully compensated for by the calibration layer. Despite these factors, the hybrid network still performs reasonably well with the network, correctly predicting each class with a minimum of 87.6% accuracy. This indicates that the hybrid network maintains a reasonable level of accuracy against these noises and discrepancies.

3 Discussion

3.1 Ablation Study and Principal Component Analysis

To understand the contribution of each component of the hybrid optoelectronic CNN, we perform an ablation study and principal component analysis (PCA). In the ablation study, we evaluate the classification accuracy when using only the electronic backend structure for classification. We summarize the ablation study results in Table 2, and the numbers in brackets represent the accuracy gain compared with “backend only” and “calibration + backend,” respectively. For a fair comparison, the backend layer here is the same structure as used in the hybrid network but is re-optimized for the best performance in the absence of any convolutional front end. Specifically, we compare the performance of a backend layer only (a single fully connected layer), the calibration and backend layers together (two fully connected layers), and the entire hybrid network. For a single fully connected layer alone, an accuracy of 89% was attained, which is less than that of the hybrid network; this highlights the utility of the optical front end. Further, for two fully connected layers (representing the calibration and backend layer) without any convolutional front end, the accuracy is reduced to 84%. This can be attributed to the fact that two layers are fully connected and, in the absence of nonlinear activation functions, tend to converge toward a “saddle point” in the optimization landscape.³³

We hypothesize that the calibration layer only remaps the optical representations and does not improve the performance of the electronic backend alone. To examine this, we further analyze the effect of the calibration layer and the overall performance of the hybrid network as compared with an all-electronic network using PCA. PCA is a statistical method widely used in various fields, especially in analyzing the quality of neural networks,³⁴ to project original, high-dimensional data into a new, simpler coordinate system for explicit interpretation. Specifically, PCA computes the eigenvalue decomposition of the covariance matrix or the singular value decomposition of input data to determine the principal direction (known as “principal components”). Principal component 1 denotes the axis of maximum variance, encapsulating the most substantial relationships among variables, whereas principal component 2, orthogonal to principal component 1, captures the second most significant variance direction. Notably, the first two principal components typically contain the most crucial information. In this study, we compress the output dimensions from electronic and optical convolutional layers into two principal components, respectively, to compare the classification efficacy of each approach; this is observed by comparing the clusters observed in PCA visualizations.³⁵

In Fig. 4, we use PCA to show that the raw experimental data do identify the fundamental components necessary for classification, but that the calibration function is necessary to shift optical representations to a form that is compatible with the pre-designed electronic backend. As shown in Figs. 4(a) and 4(c), we observe that both the all-electronic CNN outputs and the uncalibrated hybrid network experimental results can effectively distinguish among different classes due to the clustering behavior of specific classes, e.g., light blue (number 6) and navy blue (number 7). This clustering behavior indicates that despite any observable shifts in the PCA plot, the fundamental capacity of the hybrid network to classify data remains comparable to that of all-electronic networks. However, due to the differences between the optical experimental output and the expected input to the electronic backend, directly using the original backend network results in a notable drop in accuracy, down to 16.3%. The calibration layer is designed to recalibrate the outputs from the optical convolution layer back to the original outputs, thereby enabling the use of the original backend without retraining. As shown in Figs. 4(b) and 4(c), the calibrated experimental result exhibits very similar clustering behavior to the all-electronic

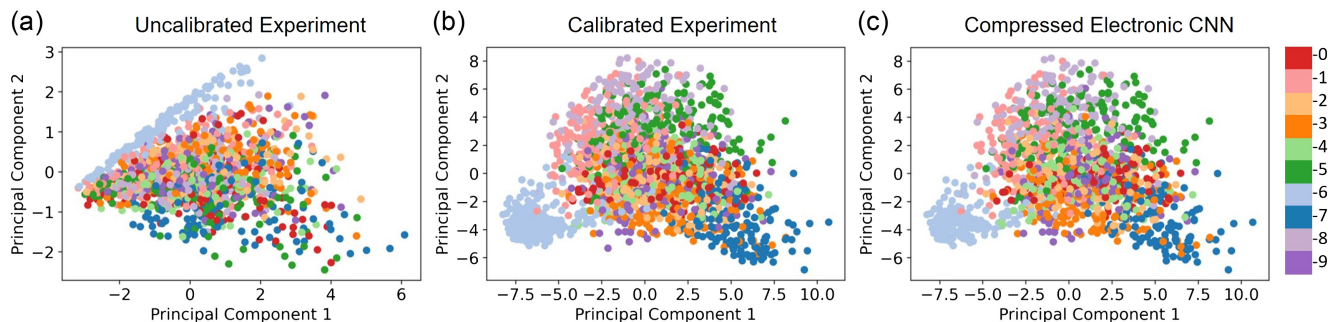


Fig. 4 PCA of the hybrid CNN. (a) PCA of the uncalibrated experimental hybrid CNN classification data. (b) PCA of the calibrated experimental data, which has been remapped and exhibits clustering behavior similar to that of the compressed electronic CNN data. (c) PCA of the compressed electronic CNN data.

network, further demonstrating that the hybrid network classification is comparable to that of the all-electronic network.

3.2 PSF-Engineered Meta-Optics

We emphasize two advantages of our PSF engineering method for performing optical convolution. First, we highlight the simplicity of the optical system, as this method requires only a display, a single layer of optic, and a camera, making it compact and simple to execute. Incoherent illumination is used, so this approach can be applied to real-world image classification scenarios. Second, we highlight the ease of integration with the optimized electronic system. One of the challenges faced by optical computing is that electronic computing is already extremely powerful, having had decades of research and development into algorithms and hardware.³ In our approach, the optics are designed to implement the electronically optimized kernels. These kernels may be modified to reflect improvements in electronic CNN models and architectures, and the optics can accordingly be adapted to implement convolutions with arbitrary kernel matrices.

Furthermore, the GS algorithm³⁰ used to design the optics is a well-established technique. The GS algorithm is an iterative phase-retrieval algorithm to determine the phase (in the optic plane) that produces an intensity pattern in another desired plane (the focal plane). In other words, the meta-optics are phase-only holograms producing the desired PSFs as their images. Due to the fact that only amplitude, and not phase, contributes to the intensity pattern, the iteratively designed phase masks are not unique. More details on the implementation of the GS algorithm are available in the [Supplementary Material](#). In the [Supplementary Material](#), we also discuss an alternative design method based on automatic differentiation, which also produces viable optics, but we found the GS-designed optics to produce slightly brighter, clearer images.

There are, however, two major limitations of the PSF engineering approach. One limitation is that there is no guarantee that the desired PSF is physically realizable. That is, a single-phase mask that satisfies the desired amplitude constraints may not exist. However, by introducing an electronic calibration layer, the resultant PSF does not need to be perfect to effectively classify the data. Alternatively, to ensure physically realizable PSFs, one could adopt an end-to-end optimization scheme wherein the phase mask is simultaneously optimized with a backend. However, training this large phase mask (on the order of 10^5 to 10^6 unit cells per kernel optic) is prohibitively costly, and the design space is potentially too large to attain convergence. In contrast, separately training the electronic convolutional kernels and the optics are both reasonable steps, which, as demonstrated, are effective when combined.

A second limitation of the described approach is that we assume the PSF is spatially invariant. In reality, light from different spatial locations on the imaging object intersects the meta-optic at various angles of incidence, resulting in a different PSF for the off-axis rays. In contrast, we design the optics assuming a normal incident illumination. To mitigate this discrepancy, we ensure the incoming angles of incidence are relatively small by placing the display far away from the optics (90 mm) relative to the focal length (2.4 mm). Therefore, for a displayed image size of $8 \text{ mm} \times 8 \text{ mm}$, we ensure that the maximum deviation from normal incidence is 3.6 deg, and therefore, the assumption of spatial invariance is reasonable for our system. However, for

a large field-of-view imaging system, we may need to explicitly model the spatially varying PSF. We note that Wei et al.¹⁶ used reparameterization techniques to design spatially varying kernel optics and reported higher classification accuracy using this method (73.8%) versus designing optics with the assumption of spatial invariance (71.6%) on the CIFAR-10 data set. In another variation of a PSF engineering technique, Zheng et al.²³ engineered the PSF of polarization-sensitive meta-optics to provide an array of focal spots that produce images of intensities relative to the kernel weights. However, we note that this approach requires a more complex optical system and incurs transmission losses under ambient light due to polarization sensitivity.

3.3 Outlook—Computational Effectiveness of the Hybrid CNN

The number of MAC operations required of a network serves as a metric of computational complexity that is independent of the employed hardware technology. In a modern digital system, one MAC operation consumes $\sim 1 \text{ pJ}$,^{19,22} so the hybrid network is expected to reduce the required power to classify an input from $17 \text{ }\mu\text{J}$ to 85 nJ based on the reduction in MAC operations. The latency of such a classification task is also expected to decrease proportionately. For input that is already in the optical domain, the power and latency required to capture an image and convert it to digital input is the same regardless of the network. Specifically, the handwritten digits of MNIST were captured using a standard camera, and the optical front end of our network uses a standard camera sensor, with meta-optics replacing the refractive camera lens. The network's performance could be further enhanced by adding additional calibration layers, but incorporating more fully connected layers might risk gradient vanishing or explosion.³⁶ We also find that a larger number of kernels could increase overall accuracy, but this would require increasing the number of suboptics in the meta-optic layer and thereby the overall footprint of the meta-optical layer. This footprint is ultimately limited by the camera sensor size.

The benefit of optically implementing the convolutional step becomes more significant as the number of input pixels is increased. The electronic computational complexity of each convolutional layer is determined by the height and width of the input images (H, W), as well as the size of the kernels (k), and is $\mathcal{O}(HWk^2)$.¹⁷ However, the computational complexity decreases to $\mathcal{O}(1)$ in the optical convolutional layer. For example, when the MNIST data set's typical image size of $28 \text{ pixel} \times 28 \text{ pixel}$ is increased to $100 \text{ pixel} \times 100 \text{ pixel}$, the electronic computational complexity of each convolutional layer is expected to increase 12.76 times (assuming the kernel size remains unchanged), but an optically implemented convolution would not incur any increase in computation time. Therefore, as the resolution of real-world images continues to increase, hybrid networks such as the one described offer a promising solution to scaling problems incurred by all-electronic networks.

In summary, we demonstrate single-layer optical convolution with an electronic backend to achieve similar accuracy as AlexNet-Mod on MNIST handwritten digit classification, with a 99.5% reduction in computational complexity. To circumvent the nonlinearity of AlexNet-Mod, we use KD to compress the CNN into linear layers, which are then implemented in a hybrid format. As a further innovation, we implement the convolution optically via engineering the PSF of meta-optics, which results

in a more compact and resilient optical front end than the commonly used $4f$ lens system and does not require coherent illumination or polarization control. This hybrid approach integrates seamlessly with existing CNN architectures, utilizing simple optical design and requiring no retraining of the electronic backend to classify the data in the experiment. This approach is also suitable for scaling to higher-resolution data sets; unlike in all-electronic networks where the convolution time scales with the number of input pixels, for optical convolution, the processing time is independent of the resolution of the data set. This work serves as a baseline for other optical and hybrid neural networks for higher bandwidth as well as lower power and latency in increasingly prevalent CNN applications.

4 Materials and Methods

4.1 Transfer Knowledge to Linear Networks

The KD algorithm is designed to compress neural networks. KD accomplishes this by transferring knowledge from a larger, pre-trained network (referred to as the “teacher model”) to a more compact network (referred to as the “student model”). In our implementation, we use AlexNet-Mod as the teacher network and a linear electronic network as the student network. The student network only comprises a single CNN coupled with a single fully connected layer. Straightforwardly training this linear network tends to easily converge to suboptimal saddle points; however, with the KD approach, the student network converges faster and obtains better performance than it would achieve without the KD training.

The KD algorithm optimizes the linear student network by combining two types of losses: temperature loss and student loss. Similar to other conventional losses in image classification, KD uses the training labels as “hard labels” and computes the probabilities distribution vector p^{hl} , where each element p_i^{hl} in the vector corresponds to the probability of the current input belonging to the class i . The softmax function is used to compute probabilities

$$p_i^{hl} = \exp(z_i) / \sum \exp(z_i), \quad (1)$$

where z_i is the student logits after the last fully connected layer.

Then, KD treats the prediction of the teacher model as “soft labels” that inform the student model because training with hard labels only is a sensitive process, especially for compact linear networks. KD includes a softening parameter, T , named the distillation temperature for teacher probabilities. Therefore, for each input, at the same time of computing p^{hl} with the student model, KD computes the soft probabilities vector, p^{sl} , with the teacher model according to

$$p_i^{sl} = \exp(y_i/T) / \sum \exp(y_i/T). \quad (2)$$

Therefore, the total loss is then calculated as a weighted summation of the two losses,

$$\mathcal{L}(x, \Phi) = \alpha \mathcal{L}_C(y, p^{hl}) + (1 - \alpha) \mathcal{L}_k[(p^{sl,t}; T = \tau), (p^{sl,s}; T = \tau)], \quad (3)$$

where x corresponds to the input, y is the training data, Φ is the student model weight, \mathcal{L}_C is the cross-entropy loss function,

\mathcal{L}_k is the Kullback–Leibler divergence loss function,³⁷ p^{hl} corresponds to the student model hard predictions, $p^{sl,s}$ corresponds to the student predictions under given teacher model probabilities $p^{sl,t}$, and α is the weighting parameter.

4.2 Calibrating Optical Experimental Results with Limited Data

The calibration function is designed to minimize the discrepancy between the ideal model and the previously trained backend. This addition addresses a variety of differences that may occur between the optical and electronic counterparts, including scaling, translation, rotation, and optical noise. With the addition of the calibration layer, the weights of the fully connected layer are preserved, and the optical front end can be integrated into the existing network framework without retraining the backend or fine-tuning the optical alignment.

Specifically, the backend includes a calibration layer and the original backend layer used in the “compressed CNN,” which are both fully connected layers. The input of the original backend layer is 288, and the output is 10, where 288 is the flattened image size after the optical convolutional layer. The input and output of the calibration layer are both 288 to align with the optical front end and the original backend. We summarized the network structure details in [Supplementary Material S1](#). The loss function used in this process is defined as

$$\mathcal{L} = \min[f_{\text{calibrate}}(\text{ON}), \text{EN}], \quad (4)$$

where ON is the network with the optical experimental results, and EN is the all-electronic network. This approach aims to refine the experimental output to align more closely with the predesigned electronic network. To prevent overfitting, we strategically limit our training to only 10% of the available data, ensuring that our model remains efficient.^{38,39} The calibration layer addresses diverse types of noise encountered in the optical system, thus ensuring a more robust hybrid network.

To obtain accuracy, the model generates scores for each input image. These scores are the output of the last layer (size 10 by 1). We use the following equation to convert the scores to accuracy:

$$\text{Accuracy} = \frac{1}{N} \sum_{i=1}^N [\arg \max(\text{scores}_i) == \text{label}_i], \quad (5)$$

where N is the total number of samples. We convert these scores into predicted class indices by selecting the class with the highest score for each data point. Next, we check whether the predicted class for each data point matches the true class. Finally, we sum the number of matches to determine how many predictions were correct.

4.3 Meta-Optics Design

The meta-optics are designed with our experimental setup in mind. Due to the sensitivity of our camera (GT-1930C) and available light sources, we design the optics specifically for 525 nm illumination. For all electromagnetic simulations, we use a simulation grid size of 586 nm to be both comparable to the wavelength of the light and evenly divisible by the size of the camera pixels (5.86 μm per pixel). Each suboptic is square, 800 \times 800 simulation pixels in size, which provides a

compact footprint and reasonable computation time. Although it is not necessary to propagate the electric fields on a subwavelength grid, it is necessary to design the meta-optic scatterers with subwavelength periodicity. Therefore, we divide each meta-optic pixel into a 2×2 block of square meta-optic scatterers, each with a period of 293 nm. With 750-nm SiN ($n = 2.06$) pillars, we select a set that provides a 0 to 2π phase shift at the desired wavelength. The scatterer unit cells were simulated using S4 RCWA.⁴⁰ More details on the meta-optic design are available in the [Supplementary Material](#).

4.4 Meta-Optics Fabrication

The convolutional meta-optics are fabricated on a silicon nitride layer on a quartz substrate. We first deposited silicon nitride on a double-side polished quartz wafer using plasma-enhanced chemical vapor deposition (Oxford; Plasma Lab 100). Then, we patterned on a positive-tone resist (ZEP-520A) using e-beam lithography (JEOL; JBX6300FS). We used alumina as a hard mask for etching the silicon nitride layer, so we deposited the alumina using e-beam evaporator (CHA; SEC-600) and did liftoff with 1-methyl-2-pyrrolidinone. We etched the silicon nitride layer with a plasma etcher (Oxford; PlasmaLab 100, ICP-180) using fluorine-based gases. To minimize the stray light of the meta-optics, we blocked the light except for the 16-kernel suboptics by putting apertures around each one using photolithography (Heidelberg; DWL66+) and metal deposition followed by a liftoff process.

4.5 Optical Measurements

Two-dimensional PSFs of 16 different optical kernels are measured simultaneously with a simple measurement setup. A single-mode optical fiber-coupled light source acts as a point source, and the meta-optics are placed 92 mm apart from the source. As we put the meta-optics on a three-axis linear stage and a kinetic mount with two rotation-adjusting knobs, both the position and angle of the meta-optics can be well defined with respect to the designed setup. Then, we put a high-resolution color camera (GT-1930C with $5.86 \mu\text{m}$ per pixel resolution) 2.4 mm from the meta-optics to collect the PSFs of the kernels. We simply replaced the light source from the single-mode optical fiber to a microdisplay presenting the MNIST data set of handwritten digits to get the convolved images. The camera could capture all 16 convolved images from different kernels at the same time, and we used Python code to automatically collect convolved images from 10,000 numbers of the MNIST data set.

Disclosures

A. M. is a co-founder of Tunoptix, which aims to commercialize meta-optics technology.

Code and Data Availability

The code and data generated and/or analyzed are available from the corresponding authors upon reasonable request.

Author Contributions

A. W. S., J. X., and M. C. contributed equally to this work. E. S. and A. M. conceived the idea and provided funding. S. C. was involved in designing the preliminary convolutional meta-optics and KD approaches. J. X. and E. S. developed the KD approach.

J. X. trained the neural network and analyzed the experimental data. A. W. S. developed the PSF engineering approach and designed the meta-optics. M. C. fabricated the meta-optics and conducted the optical experiments. L. H. wrote the script to automate the experiment. J. F. advised on the fabrication and experiment. J. F., E. S., and A. M. provided project management and edited the paper.

Acknowledgments

This work was supported by the National Science Foundation (Grant Nos. NSF-ECCS-2127235 and EFRI-BRAID-2223495). Part of this work was conducted at the Washington Nanofabrication Facility/Molecular Analysis Facility, a National Nanotechnology Coordinated Infrastructure (NNCI) site at the University of Washington with partial support from the National Science Foundation (Grant Nos. NNCI-1542101 and NNCI-2025489).

References

1. A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Adv. Neural Inf. Process. Syst.* **25** (2012).
2. X. Li et al., "Performance analysis of GPU-based convolutional neural networks," in *45th Int. Conf. Parallel Process. (ICPP)*, pp. 67–76 (2016).
3. P. L. McMahon, "The physics of optical computing," *Nat. Rev. Phys.* **5**, 717–734 (2023).
4. J. Chang et al., "Hybrid optical-electronic convolutional neural networks with optimized diffractive optics for image classification," *Sci. Rep.* **8**, 12324 (2018).
5. S. Colburn et al., "Optical frontend for a convolutional neural network," *Appl. Opt.* **58**(12), 3179–3186 (2019).
6. M. Yang et al., "Optical convolutional neural network with atomic nonlinearity," *Opt. Express* **31**(10), 16451–16459 (2023).
7. A. M. McNeil et al., "Fundamentals and recent developments of free-space optical neural networks," *J. Appl. Phys.* **136**(3), 030701 (2024).
8. H. Chen et al., "Diffractive deep neural networks: theories, optimization, and applications," *Appl. Phys. Rev.* **11**(2), 021332 (2024).
9. L. Cutrona et al., "Optical data processing and filtering systems," *IRE Trans. Inf. Theory* **6**(3), 386–400 (1960).
10. J. N. Mait, G. W. Euliss, and R. A. Athale, "Computational imaging," *Adv. Opt. Photonics* **10**, 409–483 (2018).
11. C. M. V. Burgos et al., "Design framework for metasurface optics-based convolutional neural networks," *Appl. Opt.* **60**, 4356–4365 (2021).
12. Z. Hu et al., "High-throughput multichannel parallelized diffraction convolutional neural network accelerator," *Laser Photonics Rev.* **16**, 2200213 (2022).
13. W. Lohmann and W. T. Rhodes, "Two-pupil synthesis of optical transfer functions," *Appl. Opt.* **17**(7), 1141–1151 (1978).
14. J. N. Mait and W. T. Rhodes, "Two-pupil synthesis of optical transfer functions: 2: pupil function relationships," *Appl. Opt.* **25**(12), 2003–2007 (1986).
15. J. N. Mait, "Pupil-function design for complex incoherent spatial filtering," *J. Opt. Soc. Am. A* **4**(7), 1185–1193 (1987).
16. K. Wei et al., "Spatially varying nanophotonic neural networks," *Sci. Adv.* **10**(45) (2024).
17. J. Xiang et al., "Knowledge distillation circumvents nonlinearity for optical convolutional neural networks," *Appl. Opt.* **61**(9), 2173–2183 (2022).
18. A. Ryou et al., "Free-space optical neural network based on thermal atomic nonlinearity," *Photon. Res.* **9**, B128–B134 (2021).
19. T. Wang et al., "Image sensing with multilayer nonlinear optical neural networks," *Nat. Photonics* **17**(5), 408–415 (2023).

20. H. Zheng et al., “Meta-optic accelerators for object classifiers,” *Sci. Adv.* **8**(30), eabo6410 (2022).
21. T. Zhou et al., “Large-scale neuromorphic optoelectronic computing with a reconfigurable diffractive processing unit,” *Nat. Photonics* **15**, 367–373 (2021).
22. L. Huang et al., “Photonic advantage of optical encoders,” *Nanophotonics* **13**(7), 1191–1196 (2023).
23. H. Zheng et al., “Multichannel meta-imagers for accelerating machine vision,” *Nat. Nanotechnol.* **19**, 471–478 (2024).
24. X. Lin et al., “All-optical machine learning using diffractive deep neural networks,” *Science* **361**(6406), 1004–1008 (2018).
25. D. Mengu et al., “Analysis of diffractive optical neural networks and their integration with electronic neural networks,” *IEEE J. Sel. Top. Quantum Electron.* **26**(1), 3700114 (2020).
26. C. He et al., “Pluggable multitask diffractive neural networks based on cascaded metasurfaces,” *Opto-Electron. Adv.* **7**, 230005 (2024).
27. X. Luo et al., “Metasurface-enabled on-chip multiplexed diffractive neural networks in the visible,” *Light Sci. Appl.* **11**, 158 (2022).
28. S. Öztürk et al., “Convolution kernel size effect on convolutional neural network in histopathological image processing applications,” in *Int. Symp. Fund. Elect. Eng. (ISFEE)*, IEEE, pp. 1–5 (2018).
29. H. Cai et al., “Network augmentation for tiny deep learning,” in *Int. Conf. Learn. Represent.* (2022).
30. R. W. Gerchberg, “A practical algorithm for the determination of phase from image and diffraction plane pictures,” *Optik* **35**, 237–246 (1972).
31. J. W. Goodman, *Introduction to Fourier Optics*, Roberts & Company, Englewood, Colorado (2005).
32. K. Matsushima and T. Shimobaba, “Band-limited angular spectrum method for numerical simulation of free-space propagation in far and near fields,” *Opt. Express* **17**(22), 19662–19673 (2009).
33. A. Jacot et al., “Saddle-to-saddle dynamics in deep linear networks: small initialization training, symmetry, and sparsity,” <https://doi.org/10.48550/arXiv.2106.15933> (2021).
34. J. Ma and Y. Yuan, “Dimension reduction of image deep feature using PCA,” *J. Visual Commun. Image Represent.* **63**, 102578 (2019).
35. G. Ivosev, L. Burton, and R. Bonner, “Dimensionality reduction and visualization in principal component analysis,” *Anal. Chem.* **80**(13), 4933–4944 (2008).
36. S. Hochreiter, “The vanishing gradient problem during learning recurrent neural nets and problem solutions,” *Int. J. Uncertainty Fuzziness Knowledge Based Syst.* **6**(2), 107–116 (1998).
37. G. Seo et al., “KL-divergence-based region proposal network for object detection,” in *IEEE Int. Conf. Image Process. (ICIP)*, IEEE, pp. 2001–2005 (2020).
38. J. Xiang and E. Shlizerman, “TKIL: tangent kernel optimization for class balanced incremental learning,” in *Proc. IEEE/CVF Int. Conf. Comput. Vision*, pp. 3529–3539 (2023).
39. Y. Zheng et al., “BI-MAML: balanced incremental approach for meta learning,” <https://doi.org/10.48550/arXiv.2006.07412> (2020).
40. V. Liu and S. Fan, “S4: a free electromagnetic solver for layered periodic structures,” *Comput. Phys. Commun.* **183**, 2233–2244 (2012).
41. J. Chang et al., “Reducing MAC operation in convolutional neural network with sign prediction,” in *Int. Conf. Inf. Commun. Technol. Convergence (ICTC)*, IEEE, pp. 177–182 (2018).
42. S. B. Damelin and W. Miller, *The Mathematics of Signal Processing*, Cambridge University, Cambridge (2011).
43. B. Dube, “Phase retrieval II: iterative transform algorithms,” <https://www.retrorefractions.com/blog/phase-retrieval-02/> (2023).
44. B. Dube, “PRAISE,” GitHub Repository, <https://github.com/brandondube/praise/tree/master> (2019).

Biographies of the authors are not available.