APPLIED SCIENCES AND ENGINEERING

Spatially varying nanophotonic neural networks

Kaixuan Wei[1]†, Xiao Li[1]†, Johannes Froech[2]†, Praneeth Chakravarthula[1], James Whitehead[2], Ethan Tseng[1], Arka Majumdar[2], Felix Heide[1]*

The explosive growth in computation and energy cost of artificial intelligence has spurred interest in alternative computing modalities to conventional electronic processors. Photonic processors, which use photons instead of electrons, promise optical neural networks with ultralow latency and power consumption. However, existing optical neural networks, limited by their designs, have not achieved the recognition accuracy of modern electronic neural networks. In this work, we bridge this gap by embedding parallelized optical computation into flat camera optics that perform neural network computations during capture, before recording on the sensor. We leverage large kernels and propose a spatially varying convolutional network learned through a low-dimensional reparameterization. We instantiate this network inside the camera lens with a nanophotonic array with angle-dependent responses. Combined with a lightweight electronic back-end of about 2K parameters, our reconfigurable nanophotonic neural network achieves 72.76% accuracy on CIFAR-10, surpassing AlexNet (72.64%), and advancing optical neural networks into the deep learning era.

INTRODUCTION

Increasing demands for high-performance artificial intelligence (AI) in the last decade have levied immense pressure on computing architectures across domains, including robotics, transportation, personal devices, medical imaging and scientific imaging. Although electronic microprocessors have undergone drastic evolution over the past 50 years (*1*), providing us with general-purpose central processing units and custom accelerator platforms (e.g., graphical processing unit and Digital Signal Processor (DSP) ASICs), this growth rate is far outpaced by the explosive growth of AI models. Specifically, the Moore's law delivers a doubling in transistor counts every 2 years (*2*), whereas deep neural networks (DNNs) (*3*), arguably the most influential algorithms in AI, have doubled in size every 6 months (*4*). However, the end of voltage scaling has made the power consumption, and not the number of transistors, the principal factor limiting further improvements in computing performance (*5*). Overcoming this limitation and radically reducing compute latency and power consumption could drive unprecedented applications from low-power edge computation in the camera, potentially enabling computation in thin eyeglasses or microrobots and reducing power consumption in data centers used for training of neural network architectures.

Optical computing has been proposed as a potential avenue to alleviate several inherent limitations of digital electronics, e.g., compute speed, heat dissipation, and power, and could potentially boost computational throughput, processing speed, and energy efficiency by orders of magnitude (*6*–*10*). Such optical computers leverage several advantages of photonics to achieve high throughput, low latency, and low power consumption (*11*). These performance improvements are achieved by sacrificing reconfigurability. Thus, although general-purpose optical computing has yet to be practically realized due to obstacles such as larger physical footprints and inefficient optical switches (*12*, *13*), several notable advances have already been made toward optical/photonic processors tailored specifically for AI (*14*, *15*). Representative examples include optical computers that perform widely used signal processing operators (*16*–*22*), e.g., spatial/temporal differentiation, integration, and convolution with performance far beyond those of contemporary electronic processors. Most notably, optical neural networks (ONNs) (*6*, *23*–*38*) can perform AI inference tasks such as image recognition when implemented as fully optical or hybrid opto-electronical computers.

Existing ONNs can be broadly classified into two categories based on either integrated photonics (*24*–*30*) [e.g., Mach-Zehnder interferometers (*23*, *26*), phase change materials (*24*), microring resonators (*29*), multimode fibers (*30*)] for physically realizing multiply-adds floating point operations (FLOPs), or with free-space optics (*6*, *31*–*37*) that implement convolutional layers with light propagation through diffractive elements [e.g., 3D-printed surfaces (*6*), 4F optical correlators (*37*), optical masks (*35*), and metasurfaces (*36*)]. The design of these ONN architectures has been fundamentally restricted by the underlying network design, including the challenge of scaling to large numbers of neurons (within integrated photonic circuits) and the lack of scalable energy-efficient nonlinear optical operators. As a result, even the most successful ensemble ONNs (*31*) that use dozens of ONNs in parallel, have only achieved LeNet (*39*)–level accuracy on image classification, which was achieved by their electronic counterparts over 30 years ago. Moreover, most high-performance ONNs can only operate under coherent illumination, prohibiting the integration into the camera optics under natural lighting conditions. Although hybrid opto-electronic networks (*35*, *36*, *40*) working on incoherent light do exist, most of them do not yield favorable results as their optical front-end is designed for small-kernel spatially uniform convolutional layers, which this work finds does not fully exploit the design space available for optical convolution.

In this work, we report a novel nanophotonic neural network that lifts the aforementioned limitations, allowing us to close the gap to the first modern DNN architectures (*41*) with optical compute in a flat form factor of only 4 mm length, akin to performing computation on the sensor cover glass, in lieu of the bulky compound 4-f system–based Fourier filter setup (*40*). We leverage the ability of a lens system to perform large-kernel spatially varying (LKSV) convolutions

[1]Department of Computer Science, Princeton University, Princeton, NJ, USA. [2]Department of Electrical and Computer Engineering, University of Washington, Seattle, WA, USA.
*Corresponding author. Email: fheide@princeton.edu
†These authors contributed equally to this work.

tailored specifically for image recognition and semantic segmentation. These operations are performed during the capture before the sensor makes a measurement. We learn large kernels via low-dimensional reparameterization techniques, which circumvent spurious local extremum caused by direct optimization. To physically realize the ONN, we develop a differentiable spatially varying inverse design framework that solves for metasurfaces (*42*–*46*) that can produce the desired angle-dependent responses under spatially incoherent illumination. Because of the compact footprint and complementary metal-oxide semiconductor (CMOS) sensor compatibility, the resulting optical system is not only a photonic accelerator but also an ultracompact computational camera that directly operates on the ambient light from the environment before the analog to digital conversion. We find that this approach facilitates generalization and transfer learning to other tasks, such as semantic segmentation, reaching performance comparable to AlexNet (*41*) in 1000-category ImageNet (*47*) classification and PASCAL VOC (*48*) semantic segmentation.

Recent work (*49*) concurrent to ours reported a novel metasurface doublet that implements a multichannel optical convolution via angular and polarization multiplexing under spatially incoherent illuminance, and extensions (*50*, *51*) leverage large convolutional kernels for image classification and semantic segmentation. While this work shares advantages with ours, such as multichannel operation, high performance, and the use of incoherent light, our method uses a single metasurface and relies on LKSV convolution instead of uniform convolutions increasing the parameter space by an order of magnitude.

Hence, by on-chip integration of the flat-optics front-end (>99% FLOPs) with an extremely lightweight electronic back-end (<1% FLOPs), we achieve higher classification performance than modern fully electronic classifiers [73.80% in simulation and 72.76% in experiment, compared to 72.64% by AlexNet (*41*) on CIFAR-10 (*52*) test set] while simultaneously reducing the number of electronic parameters by four orders of magnitude, thus bringing ONNs into the modern deep learning era.

RESULTS

LKSV parameterization

The working principle and optoelectronic implementation of the proposed spatially varying nanophotonic neural network (SVN^3) are illustrated in Fig. 1A. The SVN^3 is an optoelectronic neuromorphic computer that comprises a metalens array nanophotonic front-end and a lightweight electronic back-end (embedded in a low-cost microcontroller unit) for image classification or semantic segmentation. The metalens array front-end consists of 50 metalens elements that are made of 390-nm pitch nano-antennas and are optimized for incoherent light in a band around 525 nm. The wavefront modulation induced by each metalens can be represented by the optical convolution of the incident field and the point spread functions (PSFs) of the individual device. Therefore, the nanophotonic front-end performs parallel multichannel convolutions, at the speed of light, without any power consumption. We also refer to texts S1 and S3 for additional details on the physical forward model and the neural network design, respectively.

Unlike existing ONNs (*35*–*37*, *53*) that engineer the optical response to mimic a convolutional layer that consists of spatially invariant small-sized kernels, the SVN^3 uses large-sized angularly varying PSFs (Fig. 1B) as the convolution kernels to construct a LKSV convolutional layer.

Such an LKSV convolutional layer is not used in conventional DNNs due to immense computation costs and challenges in training. Nevertheless, we demonstrate that with low-dimensional reparameterization techniques, namely, large kernel factorization, and low-rank spatially varying reparameterization, this computing layer can be effectively learned in silicon, circumventing spurious local minima that can arise from naïve overparameterization (text S3).

We reparameterize a large (15×15) convolutional kernel into a stack of (seven) small 3×3 kernels, which are convolved sequentially to the large kernel (Fig. 1C). The spatially varying structure is reparameterized through a spatially variant weighted linear combination of a (large) kernel basis, which resembles the low-rank approximation of a general spatially varying kernel. Hence, we construct a three-layer convolutional neural network (CNN) composed of an LKSV convolutional stem, a depth-wise separable convolutional layer, and a fully connected classification head, for CIFAR-10 image classification. This CNN is trained in silicon by minimizing the standard cross-entropy loss with tailored regularizations (an isotropic total variation regularization and a specialized spectrum regularization) on the spatially varying kernels (text S4). Validated by the spatial combining weights and the Fourier spectrum profiles of learned kernels in fig. S2, these regularizations enforce smooth transitions of spatially varying kernels (Fig. 1E) and penalize high-pass and ill-conditioned kernels, which are challenging to implement in an optical system.

After in-silicon training, our LKSV design performs favorably compared to the conventional small-kernel spatially invariant counterpart by a sizable margin, lifting from the LeNet-level accuracy (65.45%) to the AlexNet-level accuracy (73.80%); see also Fig. 1D and tables S1 and S2.

The high computational cost of LKSV convolution in silicon can be entirely eliminated by designing a passive optical system with metalenses whose PSFs are inverse designed to mimic the designated target kernels. While the target kernels may contain both positive and negative values, optical PSFs contain only non-negative values. Thus, to generate each target kernel, we use a pair of metalenses and we take the subtraction of their image features postconvolution to achieve positive and negative values (*54*–*57*).

To optically realize a 25-channel LKSV convolutional layer, we instantiate an on-chip metalens array that consists of 50 metalenses with the device layout shown in Figs. 1A and 2A. To engineer spatially varying PSFs, we simulate the optical system and use a differentiable spatially varying inverse design framework to compute the phase profiles of the metalenses via stochastic gradient–based optimization. The angularly varying PSFs are optimized by minimizing the mean square error loss with respect to the target electronic kernels and using an energy regularization to maximize the localized energy in the region of interest on the sensor plane. By using energy regularization, we improve the light efficiency of the designed metalenses from 39.37 to 93.88% without affecting the PSF accuracy and make the ONNs more robust to unwanted scattering light and other noise in real-world measurement (text S5).

Experimental validation

The inverse design–optimized metalens array was fabricated on a single chip in a silicon nitride on quartz film (text S6). We used a

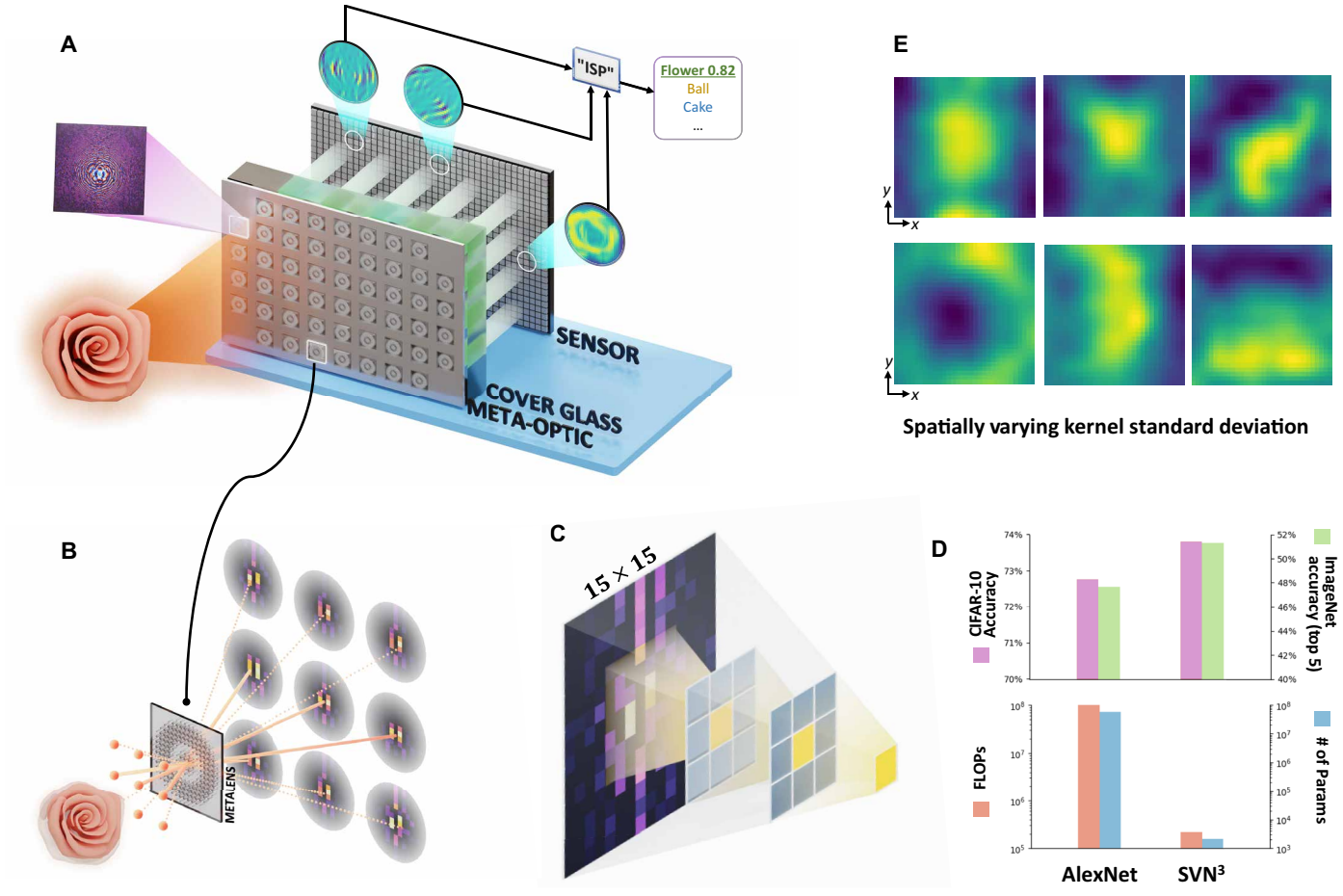

Fig. 1. Spatially varying nanophotonic neural networks. (**A**) Illustration of the proposed opto-electronic network, which comprises a nanophotonic array front-end that optically encodes the scene into multichannel image features and a lightweight electronic back-end that performs the final prediction, in a programmable manner, for image classification or semantic segmentation. (**B**) Each metalens is designed for specific learned large and angularly varying PSFs that comprise the feature kernels of the early network layers, which vary over the sensor. These kernels are learned electronically using a spatially varying reparameterization. (**C**) Large kernels of size 15 × 15 (for digital 32 × 32 image classification) are reparameterized by factorizing them into a cascade of smaller ones. (**D**) Assessment of purely electronic AlexNet (*41*) compared to SVN^3: classification accuracies on CIFAR-10 and ImageNet datasets (top barplot), digital multiply-adds floating point operations (FLOPs), and digital parameters (bottom barplot) for CIFAR-10 image recognition. The proposed method outperforms a network with multiple orders of magnitude more electronic parameters with multiple orders of magnitude fewer FLOPs, see table S2 for details. (**E**) Illustration of kernel SD that varies smoothly across space.

nanopatterning approach using electron beam lithography (EBL) to define the outline of the design in a resist, deposited a hard mask, and subsequently transferred the pattern into the underlying silicon nitride using reactive ion etching. To exclude transmission of light through nonpatterned sections, we further deposited a metal aperture around the ONN metalens kernels.

The close-up of the resulting metalens array camera and a metalens array device before mounting are shown in Fig. 2A. The PSFs (over 3 × 3 varied sampling incident angles) of three randomly selected kernels are illustrated in Fig. 2C, which illustrates the spatially varying features of the designed optical kernels. To experimentally realize the optical system and measure the image features of the metalenses, we devise the setup shown in Fig. 2B. The green channel of a smartphone organic light-emitting diode (OLED) display, which is placed at the designed object distance, is used as the incoherent light source, and a large-area CMOS sensor is placed at the focal plane of the metalens array device. When the dataset images are displayed on the display, the sensor captures the corresponding image features of all the metalens elements in a single shot.

The captured positive, negative, and real-valued features through subtraction closely resemble the electronic ground truth from both qualitative and quantitative comparisons, which verifies the effectiveness of the implemented inverse design framework (Figs. 2D and 3A). Interested readers are also referred to movie S1 for prototype demonstration of SVN^3 for dynamic content.

To extensively assess the performance of our opto-electronic neural network SVN^3, we captured the entire grayscale CIFAR-10 dataset, including 50,000 training images and 10,000 test images, with the setup described above and shown in Fig. 2B. The image features in each frame are equally spaced in a regular 6 × 9 array with the four corners being traditional hyperbolic metalenses used for device alignment (fig. S6). After cropping the image features of all the metalenses and computing the real-valued target features

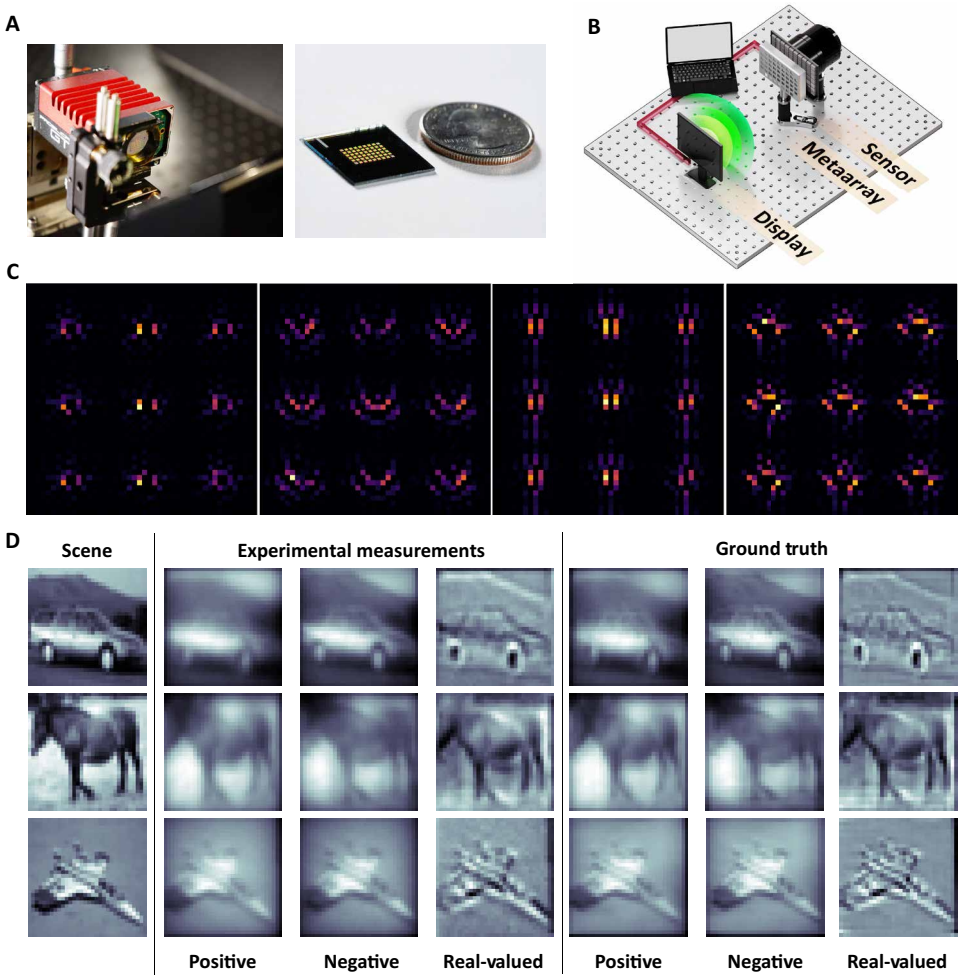

Fig. 2. Experimental validation of SVN^3. (**A**) Flat camera prototype (left) and a metalens array device before mounting (right). (**B**) Illustration of the experimental setup, consisting of an OLED display placed at the designated object distance, metalens array, and CMOS sensor. Note that no additional optics are used. Camera and display are synchronized for data capture. (**C**) Spatially varying PSF visualization on a 3 × 3 sampling grid of incident angles. Here, we show four representative kernels. (**D**) Side-by-side comparison of the experimental measurements that match the corresponding ground truth feature channels. "Real-valued" denotes the target feature channel, the negative image feature subtracted from positive image features postconvolution.

through paired subtraction, the resulting multichannel optical features are fed into the pretrained lightweight electronic back-end to obtain the final predictions. We finetune the electronic back-end using the cross-entropy loss on the experimentally captured CIFAR-10 training dataset. The finetuning procedure is identical to the prior in-silicon training of the target electronic neural network, except no extra regularization losses are applied (text S4). SVN^3 reaches to 72.76% on the CIFAR-10 test dataset, which is comparable to 73.80% of the corresponding electronic model. Similar observations are also drawn in the confusion matrices in Fig. 3B, which reveals the similar recognition behavior of the SVN^3 in real experiment and simulation. Figure 4 reports predictions on random samples from CIFAR-10 testset. The method consistently assigns a high probability to the true class (top 2). These experimental results collectively validate the effectiveness of SVN^3 in classifying common objects, extending beyond the realm of handwritten digit recognition investigated in existing work. Furthermore, we emphasize that almost all computations (>99% of FLOPs) of SVN^3 are executed on the optical side with zero energy consumption (table S2). This AlexNet-level classification accuracy is thus achieved with an ultralow power device.

Versatile reconfigurable computational camera

Our approach is generic, which we validate by instantiating SVN^3 for other datasets and tasks. Next, we describe such an instance for ImageNet classification with 1000 object categories. ImageNet is the first large-scale image classification dataset with 1.28 million labeled training data, serving as a major driver to advance modern AI. To the best of our knowledge, no existing ONN has reported results on 1000-class ImageNet classification so far. To tackle this challenging 1000-class recognition task, we use an enlarged electronic back-end with four depth-wise separable convolutional layers and one fully connected classifier. We inverse design and fabricate an on-chip metasurface array to optically encode features for 64 × 64 low-resolution ImageNet classification. Akin to the CIFAR-10 experiment, the entire training and validation datasets of ImageNet

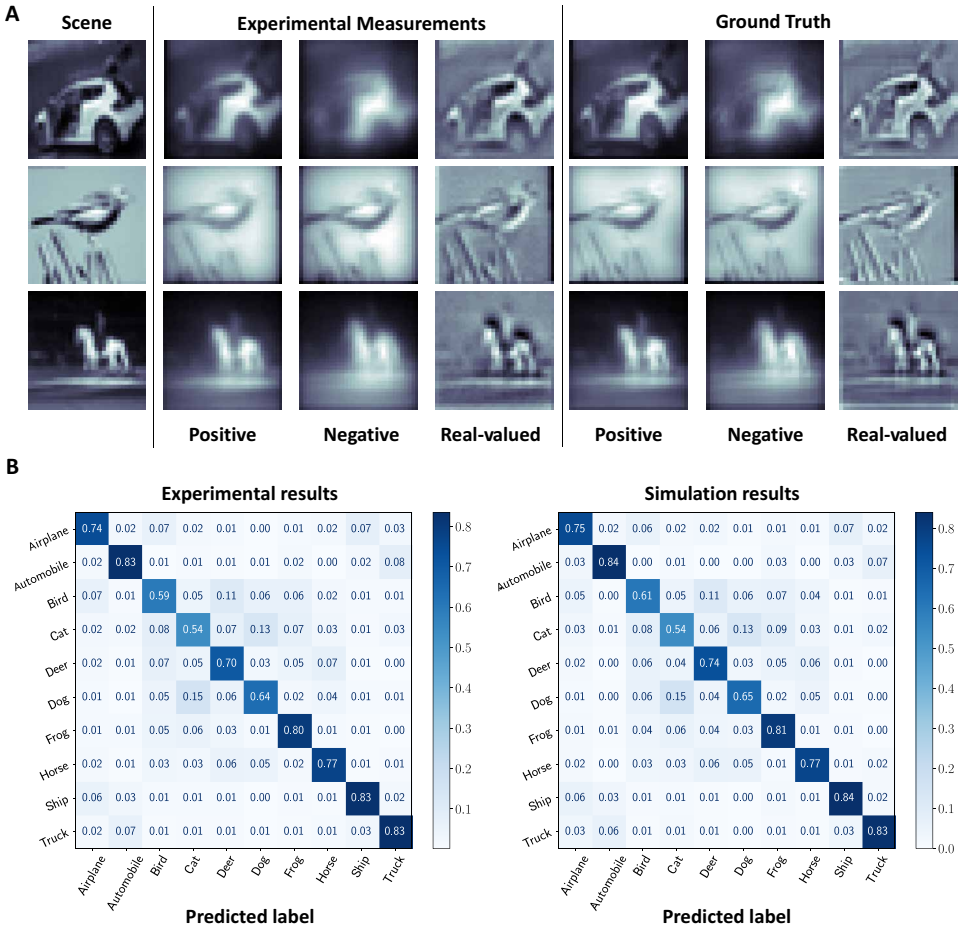

Fig. 3. Experimental measurements of a fabricated chip of a design for CIFAR-10 image classification. (**A**) Qualitative assessment of the experimental measurements compared with the ground truth feature channels. Real-valued again denotes the target feature channels via subtracting the negative from the positive image features postconvolution. (**B**) The confusion matrices of the experimental and simulation results on the CIFAR-10 test dataset validate the effectiveness of the method.

are encoded into optical features by the imaging system for finetuning and evaluation. The experimentally captured features consistently align with their electronic ground truth (Fig. 5A), validating the scalability and effectiveness of SVN^3 to process large-sized image features. After finetuning the electronic back-end on the ImageNet training set, SVN^3 achieves 48.64% top 5 classification accuracy in ImageNet validation set, outperforming AlexNet (47.60%) by 1.03%. Note that the SVN^3 for 64 × 64 ImageNet classification has 1.67 million digital multiply-accumulate operations (FLOPs), which is only 0.9% of AlexNet (180.26 million).

Although the optical front-end (encoder) in SVN^3 is not programmable after being fabricated, we demonstrate that SVN^3 can serve as a reconfigurable versatile computational camera with a universal optical encoder. By adjusting the electronic back-end (decoder) using transfer learning, SVN^3 is capable of performing diverse vision tasks beyond the initially designed task. Using the same physical setup for ImageNet classification, we conduct image recognition experiments on the CIFAR-100 (*52*), Flowers-102 (*58*), Food-101 (*59*), and Pet-37 (*60*) datasets. For all of these datasets, we achieve comparable or better performance than the (finetuned) AlexNet (Fig. 5B), consistently validating the flexibility of our hybrid opto-electronic system without adapting the optical front-end. We also validate this capability for other computer vision tasks, e.g., semantic segmentation in PASCAL VOC (*48*) dataset, where our hybrid network is competitive to the AlexNet-based segmentation network as validated in Fig. 5C. Our SVN^3 achieves a pixel accuracy of 65.73% compared with 66.34% of AlexNet-based segmentation on the PASCAL VOC test set.

DISCUSSION

In this work, we investigate a novel nanophotonic neural network that lifts the limitations of existing ONNs, propelling them to performance parity with the first modern digital neural network, AlexNet. To this end, we embed computation in the camera lens, performed during the image capture, and we exploit the spatially varying nature of large optical aberrations. Specifically, we propose a LKSV CNN, learned via low-dimensional reparameterization techniques, and physically realizing it via a meta-optical system. The proposed method shifts almost all computations (99.64%) from electronic processors into the optical domain, while allowing for an ultrathin optical stack of only 4 mm, similar to performing computation on the sensor cover glass. We find that this approach achieves an image classification accuracy of (top 1)

Fig. 4. Experimental (top 2) classification (probability) results on random samples from CIFAR-10 test set. Green- and orange-colored labels under the images denote the correct and incorrect predictions, respectively. The method accurately predicts the correct class or a visually similar class. See figs. S15 and S16 for additional examples.

72.76% on CIFAR-10 and (top 5) 48.64% on (1000-class) ImageNet, shrinking the gap between photonic and electronic AI, while ensuring generalization to diverse vision tasks without needing to fabricate new optics. Possible multi-aperture extensions of our work in the future may allow for high-resolution, multi-channel optical computing, and enable future photonic AI to bridge this gap.

MATERIALS AND METHODS

Design and optimization

We used PyTorch to design and evaluate our spatially varying nanophotonic neural network. See texts S3 to S5 for details on the architectural design, in-silicon training, and differentiable inverse design of SVN[3].

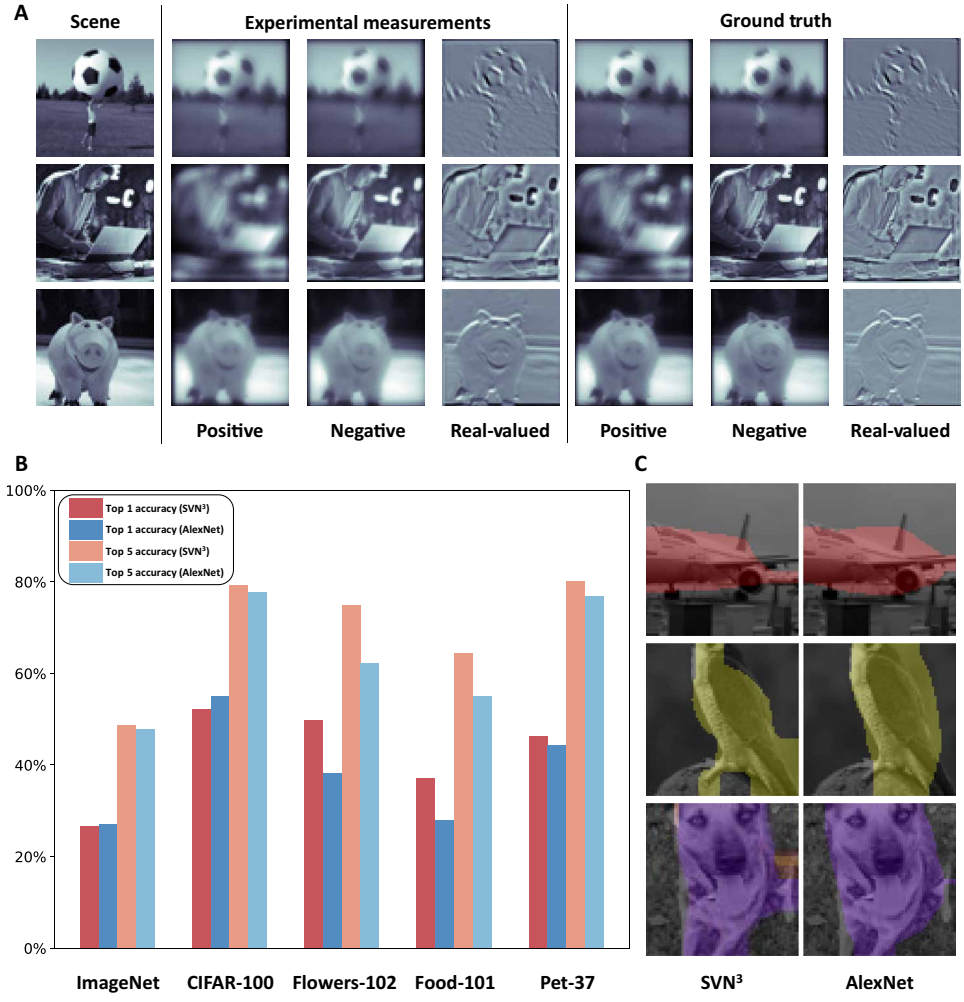

Fig. 5. Validation of SVN³ as a versatile camera for diverse vision tasks. (**A**) Experimentally measured feature maps of SVN³ on the ImageNet dataset. (**B**) Recognition on ImageNet and other downstream datasets (CIFAR-100, Flowers-102, Food-101, and Pet-37) using the same optical front-end and the transfer-learned electronic decoder. (**C**) Transfer learning for semantic segmentation on PASCAL VOC dataset. SVN³ again achieves comparable or better performance than the AlexNet-based segmentation network (see fig. S17 for additional examples). These findings validate that the proposed camera, with a fixed optical encoder, can generalize to diverse tasks by adapting the electronic back-end.

Sample fabrication

We fabricated the meta-optic on top of a 500-μm-thick double-side polished fused silica wafer. First, a 800-nm film of silicon nitride was deposited via plasma-enhanced chemical vapor deposition (PECVD) in a SPTS DeltaX PECVD using silane and ammonia as the precursor for a growth at 350°C. After growth, the wafer is diced in pieces of 2 × 2 cm and cleaned in a sonicating bath of acetone, followed by a rinse in isopropyl alcohol (IPA). Then, the sample was shortly cleaned in a O_2 plasma using a barrel etcher at 100 W for ~15 s. After the cleaning step, we spin-coated the sample with ZEP 520A resist (~400 nm), followed by a layer of a discharging polymer (DisCharge H2O). The arrays of kernels were then written on single chips for the spatially varying and spatially invariant designs via electron beam lithography (EBL) using a JEOL-JBX6300FS with acceleration voltage of 100 kV and 8-nA beam current. After EBL, the sample was rinsed in IPA and developed in amyl acetate for 2 min and rinsed in IPA. To define a hard mask, we evaporated 65 nm of alumina using a laboratory-built e-beam evaporator and a Al_2O_3 evaporation source. The resist was then lift-off overnight in N-Methylpyrrolidone (NMP) at 110°C and the sample was further cleaned in a brief O_2 plasma etch to remove remaining organic residues. We then used inductively coupled reactive ion etching (Oxford Instruments, PlasmaLab100) with an etch chemistry based on fluorine to transfer the metasurface layout from the hard mask into the silicon nitride film to a thickness of ~750 nm, whereas the remaining 50 nm of PECVD ensures higher stability of the etched device layer. After fabrication of the device layer, we deposited a metal aperture layer surrounding the metasurfaces to exclude any stray light. These apertures were created through optical direct write lithography (Heidelberg-DWL66) and subsequent deposition of a 150-nm-thick metal film (Cr).

Experimental setup

We built two experimental setups to characterize the optical performance of metalens array samples, as described in detail in text S7: The first one is used to experimentally measure the PSFs of the metalens array samples. In this setup, a 520-nm pigtailed single-mode fiber laser is used to mimic a point light source, and a CMOS sensor is used as the detector to measure the intensity response of a metalens

array sample upon the incidence of a point light source positioned at the designed object distance. A microscope objective, together with a relay lens, is used to magnify the PSF measurement on the detector plane. The second setup is used to realize the designed optical system and to measure the image features, as shown in Fig. 2 (A and B). In this setup, the green channel of a smartphone OLED display that is placed at the designated object distance is used as the incoherent light source, and a large-area CMOS sensor is placed at the focal plane of the metalens array device. The smartphone and the sensor are controlled by a computer and synchronized such that when the dataset images are displayed on the smartphone sequentially, the sensor captures the corresponding image features of all the metalens elements in a single shot.

Supplementary Materials

The PDF file includes:

Supplementary Text
Tables S1 to S7
Figs. S1 to S23
Legend for movie S1
References

Other Supplementary Material for this manuscript includes the following:

Movie S1

REFERENCES AND NOTES

1. G. E. Moore, Cramming more components onto integrated circuits. *Proc. IEEE* **86**, 82–85 (1998).
2. M. M. Waldrop, The chips are down for Moore's law. *Nature* **530**, 144–147 (2016).
3. Y. LeCun, Y. Bengio, G. Hinton, Deep learning. *Nature* **521**, 436–444 (2015).
4. J. Sevilla, L. Heim, A. Ho, T. Besiroglu, M. Hobbhahn, P. Villalobos, Compute trends across three eras of machine learning. arXiv:2202.05924 (2022).
5. M. Horowitz, "1.1 computing's energy problem (and what we can do about it)," in *2014 IEEE International Solid-State Circuits Conference Digest of Technical Papers (ISSCC)* (IEEE, 2014), pp. 10–14.
6. X. Lin, Y. Rivenson, N. T. Yardimci, M. Veli, Y. Luo, M. Jarrahi, A. Ozcan, All-optical machine learning using diffractive deep neural networks. *Science* **361**, 1004–1008 (2018).
7. D. R. Solli, B. Jalali, Analog optical computing. *Nat. Photonics* **9**, 704–706 (2015).
8. H. J. Caulfield, S. Dolev, Why future supercomputing requires optics. *Nat. Photonics* **4**, 261–263 (2010).
9. D. A. Miller, Attojoule optoelectronics for low-energy information processing and communications. *J. Lightwave Technol.* **35**, 346–396 (2017).
10. T. Wang, M. M. Sohoni, L. G. Wright, M. M. Stein, S. Y. Ma, T. Onodera, M. G. Anderson, P. L. McMahon, Image sensing with multilayer nonlinear optical neural networks. *Nat. Photonics* **17**, 408–415 (2023).
11. P. L. McMahon, The physics of optical computing. *Nat. Rev. Phys.* **5**, 717–734 (2023).
12. D. A. B. Miller, Are optical transistors the logical next step? *Nature Photonics* **4**, 3–5 (2010).
13. R. S. Tucker, The role of optics in computing. *Nature Photonics* **4**, 405 (2010).
14. G. Wetzstein, A. Ozcan, S. Gigan, S. Fan, D. Englund, M. Soljačić, C. Denz, D. A. B. Miller, D. Psaltis, Inference in artificial intelligence with deep optics and photonics. *Nature* **588**, 39–47 (2020).
15. B. J. Shastri, A. N. Tait, T. Ferreira de Lima, W. H. P. Pernice, H. Bhaskaran, C. D. Wright, P. R. Prucnal, Photonics for artificial intelligence and neuromorphic computing. *Nat. Photonics* **15**, 102–114 (2021).
16. N. Mohammadi Estakhri, B. Edwards, N. Engheta, Inverse-designed metastructures that solve equations. *Science* **363**, 1333–1338 (2019).
17. X.-Y. Xu, X.-L. Huang, Z.-M. Li, J. Gao, Z.-Q. Jiao, Y. Wang, R.-J. Ren, H. P. Zhang, X.-M. Jin, A scalable photonic computer solving the subset sum problem. *Sci. Adv.* **6**, eaay5853 (2020).
18. W. Liu, M. Li, R. S. Guzzon, E. J. Norberg, J. S. Parker, M. Lu, L. A. Coldren, J. Yao, A fully reconfigurable photonic integrated signal processor. *Nat. Photonics* **10**, 190–195 (2016).
19. H. Kwon, D. Sounas, A. Cordaro, A. Polman, A. Alù, Nonlocal metasurfaces for optical signal processing. *Phys. Rev. Lett.* **121**, 173004 (2018).
20. A. Silva, F. Monticone, G. Castaldi, V. Galdi, A. Alù, N. Engheta, Performing mathematical operations with metamaterials. *Science* **343**, 160–163 (2014).
21. T. Zhu, Y. Zhou, Y. Lou, H. Ye, M. Qiu, Z. Ruan, S. Fan, Plasmonic computing of spatial differentiation. *Nat. Commun.* **8**, 15391 (2017).
22. M. Ferrera, Y. Park, L. Razzari, B. E. Little, S. T. Chu, R. Morandotti, D. J. Moss, J. Azaña, On-chip CMOS-compatible all-optical integrator. *Nat. Commun.* **1**, 29 (2010).
23. S. Pai, Z. Sun, T. W. Hughes, T. Park, B. Bartlett, I. A. D. Williamson, M. Minkov, M. Milanizadeh, N. Abebe, F. Morichetti, A. Melloni, S. Fan, O. Solgaard, D. A. B. Miller, Experimentally realized in situ backpropagation for deep learning in photonic neural networks. *Science* **380**, 398–404 (2023).
24. J. Feldmann, N. Youngblood, C. D. Wright, H. Bhaskaran, W. H. Pernice, All-optical spiking neurosynaptic networks with self-learning capabilities. *Nature* **569**, 208–214 (2019).
25. X. Xu, M. Tan, B. Corcoran, J. Wu, A. Boes, T. G. Nguyen, S. T. Chu, B. E. Little, D. G. Hicks, R. Morandotti, A. Mitchell, D. J. Moss, 11 TOPS photonic convolutional accelerator for optical neural networks. *Nature* **589**, 44–51 (2021).
26. Y. Shen, N. C. Harris, S. Skirlo, M. Prabhu, T. Baehr-Jones, M. Hochberg, X. Sun, S. Zhao, H. Larochelle, D. Englund, M. Soljačić, Deep learning with coherent nanophotonic circuits. *Nat. Photonics* **11**, 441–446 (2017).
27. J. Feldmann, N. Youngblood, M. Karpov, H. Gehring, X. Li, M. Stappers, M. le Gallo, X. Fu, A. Lukashchuk, A. S. Raja, J. Liu, C. D. Wright, A. Sebastian, T. J. Kippenberg, W. H. P. Pernice, H. Bhaskaran, Parallel convolutional processing using an integrated photonic tensor core. *Nature* **589**, 52–58 (2021).
28. F. Ashtiani, A. J. Geers, F. Aflatouni, An on-chip photonic deep neural network for image classification. *Nature* **606**, 501–506 (2022).
29. A. N. Tait, T. F. de Lima, E. Zhou, A. X. Wu, M. A. Nahmias, B. J. Shastri, P. R. Prucnal, Neuromorphic photonic networks using silicon photonic weight banks. *Sci. Rep.* **7**, 7430 (2017).
30. U. Teğin, M. Yıldırım, İ. Oğuz, C. Moser, D. Psaltis, Scalable optical learning operator. *Nat. Comput. Sci.* **1**, 542–549 (2021).
31. M. S. S. Rahman, J. Li, D. Mengu, Y. Rivenson, A. Ozcan, Ensemble learning of diffractive optical networks. *Light Sci. Appl.* **10**, 14 (2021).
32. X. Luo, Y. Hu, X. Ou, X. Li, J. Lai, N. Liu, X. Cheng, A. Pan, H. Duan, Metasurface-enabled on-chip multiplexed diffractive neural networks in the visible. *Light Sci. Appl.* **11**, 158 (2022).
33. R. Hamerly, L. Bernstein, A. Sludds, M. Soljačić, D. Englund, Large-scale optical neural networks based on photoelectric multiplication. *Phys. Rev. X* **9**, 021032 (2019).
34. T. Zhou, X. Lin, J. Wu, Y. Chen, H. Xie, Y. Li, J. Fan, H. Wu, L. Fang, Q. Dai, Large-scale neuromorphic optoelectronic computing with a reconfigurable diffractive processing unit. *Nat. Photonics* **15**, 367–373 (2021).
35. W. Shi, Z. Huang, H. Huang, C. Hu, M. Chen, S. Yang, H. Chen, LOEN: Lensless opto-electronic neural network empowered machine vision. *Light Sci. Appl.* **11**, 121 (2022).
36. H. Zheng, Q. Liu, Y. Zhou, I. I. Kravchenko, Y. Huo, J. Valentine, Meta-optic accelerators for object classifiers. *Sci. Adv.* **8**, eabo6410 (2022).
37. J. Chang, V. Sitzmann, X. Dun, W. Heidrich, G. Wetzstein, Hybrid optical-electronic convolutional neural networks with optimized diffractive optics for image classification. *Sci. Rep.* **8**, 12324 (2018).
38. Y. Chen, M. Nazhamaiti, H. Xu, Y. Meng, T. Zhou, G. Li, J. Fan, Q. Wei, J. Wu, F. Qiao, L. Fang, Q. Dai, All-analog photoelectronic chip for high-speed vision tasks. *Nature* **623**, 48–57 (2023).
39. Y. Le Cun, B. Boser, J. S. Denker, R. E. Howard, W. Habbard, L. D. Jackel, D. Henderson, Handwritten digit recognition with a back-propagation network. *Adv. Neural Inf. Process. Syst.* **2**, 396–404 (1989).
40. S. Colburn, Y. Chu, E. Shilzerman, A. Majumdar, Optical frontend for a convolutional neural network. *Appl. Optics* **58**, 3179–3186 (2019).
41. A. Krizhevsky, I. Sutskever, G. E. Hinton, Imagenet classification with deep convolutional neural networks. *Adv. Neural Inf. Process. Syst.* **60**, 84–90 (2012).
42. N. Yu, P. Genevet, M. A. Kats, F. Aieta, J. P. Tetienne, F. Capasso, Z. Gaburro, Light propagation with phase discontinuities: Generalized laws of reflection and refraction. *Science* **334**, 333–337 (2011).
43. A. V. Kildishev, A. Boltasseva, V. M. Shalaev, Planar photonics with metasurfaces. *Science* **339**, 1232009 (2013).
44. M. Khorasaninejad, F. Capasso, Metalenses: Versatile multifunctional photonic components. *Science* **358**, eaam8100 (2017).
45. A. H. Dorrah, F. Capasso, Tunable structured light with flat optics. *Science* **376**, eabi6860 (2022).
46. E. Tseng, S. Colburn, J. Whitehead, L. Huang, S. H. Baek, A. Majumdar, F. Heide, Neural nano-optics for high-quality thin lens imaging. *Nat. Commun.* **12**, 6493 (2021).
47. J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition* (IEEE, 2009), pp. 248–255.
48. M. Everingham, L. Van Gool, C. K. Williams, J. Winn, A. Zisserman, The pascal visual object classes (VOC) challenge. *Int. J. Comput. Vis.* **88**, 303–338 (2010).

49. H. Zheng, Q. Liu, I. I. Kravchenko, X. Zhang, Y. Huo, J. G. Valentine, Multichannel meta-imagers for accelerating machine vision. *Nat. Nanotechnol.* **19**, 471–478 (2024).
50. Q. Liu, H. Zheng, B. T. Swartz, H. Lee, Z. Asad, I. Kravchenko, J. G. Valentine, Y. Huo, Digital modeling on large kernel metamaterial neural network. *J. Imaging Sci. Technol.* **67**, 1–11 (2023).
51. Q. Liu, B. T. Swartz, I. Kravchenko, J. G. Valentine, Y. Huo, ExtremeMETA: High-speed lightweight image segmentation model by remodeling multi-channel metamaterial imagers. arXiv:2405.17568 (2024).
52. A. Krizhevsky, G. Hinton, "Learning multiple layers of features from tiny images" (Tech. Rep. 0, University of Toronto, 2009).
53. W. Fu, D. Zhao, Z. Li, S. Liu, C. Tian, K. Huang, Ultracompact meta-imagers for arbitrary all-optical convolution. *Light Sci. Appl.* **11**, 62 (2022).
54. A. Lohmann, W. T. Rhodes, Two-pupil synthesis of optical transfer functions. *Appl. Optics* **17**, 1141–1151 (1978).
55. J. N. Mait, W. T. Rhodes, Two-pupil synthesis of optical transfer functions: 2: Pupil function relationships. *Appl. Optics* **25**, 2003–2007 (1986).
56. J. N. Mait, Pupil-function design for bipolar incoherent spatial filtering. *J. Opt. Soc. Am. A* **3**, 1826–1832 (1986).
57. J. N. Mait, Pupil-function design for complex incoherent spatial filtering. *J. Opt. Soc. Am. A* **4**, 1185–1193 (1987).
58. M.-E. Nilsback, A. Zisserman, "Automated flower classification over a large number of classes," in *2008 Sixth Indian Conference on Computer Vision, Graphics & Image Processing* (IEEE, 2008), pp. 722–729.
59. L. Bossard, M. Guillaumin, L. Van Gool, "Food-101–mining discriminative components with random forests," in *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6–12, 2014, Proceedings, Part VI 13* (Springer, 2014), pp. 446–461.
60. O. M. Parkhi, A. Vedaldi, A. Zisserman, C. Jawahar, "Cats and dogs," in *2012 IEEE Conference on Computer Vision and Pattern Recognition* (IEEE, 2012), pp. 3498–3505.
61. J. W. Goodman, *Introduction to Fourier Optics* (Roberts & Co. Publishers, 2005).
62. J. Li, D. Mengu, Y. Luo, Y. Rivenson, A. Ozcan, Class-specific differential detection in diffractive optical neural networks improves inference accuracy. *Adv. Photonics* **1**, 046001 (2019).
63. J. Sasián, *Introduction to Aberrations in Optical Imaging Systems* (Cambridge Univ. Press, 2013).
64. A. Arbabi, A. Faraon, Advances in optical metalenses. *Nat. Photonics* **17**, 16–25 (2023).
65. K. Shastri, F. Monticone, Nonlocal flat optics. *Nat. Photonics* **17**, 36–47 (2023).
66. F. Chollet, Xception: "Deep learning with depthwise separable convolutions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (IEEE, 2017), pp. 1251–1258.
67. A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, H. Adam, Mobilenets: Efficient convolutional neural networks for mobile vision applications. arXiv:1704.04861 (2017).
68. I. Goodfellow, Y. Bengio, A. Courville, *Deep Learning* (MIT Press, 2016).
69. K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556 (2014).
70. K. He, X. Zhang, S. Ren, J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (IEEE, 2016), pp. 770–778.
71. Y. LeCun, Generalization and network design strategies. *Connectionism in Perspective* (Elsevier, 1989), pp. 143–155.
72. S. Bartunov, A. Santoro, B. A. Richards, L. Marris, G. E. Hinton, T. P. Lillicrap, "Assessing the scalability of biologically-motivated deep learning algorithms and architectures." *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, Red Hook, NY, USA, 2018 (Curran Associates Inc., 2018), pp. 9390–9400.
73. D. L. Ruderman, W. Bialek, Statistics of natural images: Scaling in the woods. *Phys. Rev. Lett.* **73**, 814–817 (1994).
74. B. A. Olshausen, D. J. Field, Natural image statistics and efficient coding. *Network* **7**, 333–339 (1996).
75. E. P. Simoncelli, B. A. Olshausen, Natural image statistics and neural representation. *Annu. Rev. Neurosci.* **24**, 1193–1216 (2001).
76. X. Glorot, Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics* (JMLR Workshop and Conference Proceedings, 2010), pp. 249–256.
77. H. Li, Z. Xu, G. Taylor, C. Studer, T. Goldstein, Visualizing the loss landscape of neural nets. *Adv. Neural Inf. Process. Syst.* **31**, 6799–6810 (2018).
78. R. Sun, D. Li, S. Liang, T. Ding, R. Srikant, The global landscape of neural networks: An overview. *IEEE Signal Process. Mag.* **37**, 95–108 (2020).
79. S. B. Damelin, W. Miller Jr., *The Mathematics of Signal Processing, 48* (Cambridge Univ. Press) (2012).
80. U. Hasson, I. Levy, M. Behrmann, T. Hendler, R. Malach, Eccentricity bias as an organizing principle for human high-order object areas. *Neuron* **34**, 479–490 (2002).
81. M. J. Arcaro, S. A. McMains, B. D. Singer, S. Kastner, Retinotopic organization of human ventral visual cortex. *J. Neurosci.* **29**, 10638–10652 (2009).
82. R. Lafer-Sousa, B. R. Conway, Parallel, multi-stage processing of colors, faces and shapes in macaque inferior temporal cortex. *Nat. Neurosci.* **16**, 1870–1878 (2013).
83. Z. M. Saygin, D. E. Osher, E. S. Norton, D. A. Youssoufian, S. D. Beach, J. Feather, N. Gaab, J. D. E. Gabrieli, N. Kanwisher, Connectivity precedes function in the development of the visual word form area. *Nat. Neurosci.* **19**, 1250–1255 (2016).
84. I. Loshchilov, F. Hutter, Decoupled weight decay regularization. arXiv:1711.05101 (2017).
85. K. Matsushima, T. Shimobaba, Band-limited angular spectrum method for numerical simulation of free-space propagation in far and near fields. *Opt. Express* **17**, 19662–19673 (2009).
86. K. Matsushima, Shifted angular spectrum method for off-axis numerical propagation. *Opt. Express* **18**, 18453–18463 (2010).
87. D. P. Kingma, J. Ba, Adam: A method for stochastic optimization. arXiv:1412.6980 (2014).
88. V. Liu, S. Fan, S^4: A free electromagnetic solver for layered periodic structures. *Comput. Phys. Commun.* **183**, 2233–2244 (2012).

Acknowledgments

Funding: This work was supported by an NSF CAREER Award (2047359), a Packard Foundation Fellowship, a Sloan Research Fellowship, a Disney Research Award, a Project X Innovation Award, a Bosch Research Award, an Amazon Science Research Award, a Google PhD Fellowship, DARPA (contract no. DARPAW31P4Q21C0043), and an NSF grant (NSF-2127235). 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 NSF via Awards NNCI-1542101 and NNCI-2025489. **Author contributions:** K.W. and F.H. designed and analyzed the spatially varying ONN. K.W. and X.L. performed the experiments. K.W. and F.H. led the manuscript writing. J.F., J.W., and A.M. fabricated the optical devices. J.F., E.T., X.L., and A.M. assisted in design and analysis, experiments, and manuscript writing. P.C. also assisted in writing the manuscript. F.H. supervised the project. **Competing interests:** The authors declare that they have no competing interests. **Data and materials availability:** All data needed to evaluate the conclusions in the paper are present in the paper and/or the Supplementary Materials.

Submitted 4 March 2024
Accepted 1 October 2024
Published 8 November 2024
10.1126/sciadv.adp0391