nature photonics

Article

https://doi.org/10.1038/s41566-023-01170-8

Image sensing with multilayer nonlinear optical neural networks

Received: 29 August 2022

Accepted: 3 February 2023

Published online: 23 March 2023

Check for updates

Tianyu Wang[1,4] ✉, Mandar M. Sohoni[1,4] ✉, Logan G. Wright[1,2] ✉, Martin M. Stein[1], Shi-Yuan Ma[1], Tatsuhiro Onodera[1,2], Maxwell G. Anderson[1] & Peter L. McMahon[1,3] ✉

Optical imaging is commonly used for both scientific and technological applications across industry and academia. In image sensing, a measurement, such as of an object's position or contour, is performed by computational analysis of a digitized image. An emerging image-sensing paradigm relies on optical systems that—instead of performing imaging—act as encoders that optically compress images into low-dimensional spaces by extracting salient features; however, the performance of these encoders is typically limited by their linearity. Here we report a nonlinear, multilayer optical neural network (ONN) encoder for image sensing based on a commercial image intensifier as an optical-to-optical nonlinear activation function. This nonlinear ONN outperforms similarly sized linear optical encoders across several representative tasks, including machine-vision benchmarks, flow-cytometry image classification and identification of objects in a three-dimensionally printed real scene. For machine-vision tasks, especially those featuring incoherent broadband illumination, our concept allows for a considerable reduction in the requirement of camera resolution and electronic post-processing complexity. In general, image pre-processing with ONNs should enable image-sensing applications that operate accurately with fewer pixels, fewer photons, higher throughput and lower latency.

Optical images are widely used to capture and convey information about the state or dynamics of physical systems, in both fundamental science and technology. They are used to guide autonomous machines, to assess manufacturing processes, and to inform medical diagnoses and procedures. In such applications, an optical system such as a microscope forms an image of a subject on a camera, which converts the photonic, analogue image into an electronic, digital image. Digital images are typically many megabytes; however, for most applications, nearly all of this information is redundant or irrelevant. There are three main reasons: (1) natural images contain sparse information and are therefore compressible[1–3]; (2) most applications involve images of subjects with additional underlying commonalities beyond sparsity; and, finally (3), most information in an image is irrelevant to the image's end use. Here we refer to machine-vision applications for which factor (3) is applicable as image sensing—only a specific subset of information from each image is sought for these applications, as demonstrated in Fig. 1a.

The information inefficiency of conventional imaging has inspired machine-vision paradigms in which optics are designed not as conventional imaging systems, but instead as optical encoders—computational pre-processors that extract relevant information from an image[1,4–9]. Techniques include end-to-end optimization[5,8–17],

[1]School of Applied and Engineering Physics, Cornell University, Ithaca, NY, USA. [2]Physics & Informatics Laboratories, NTT Research Inc., Sunnyvale, CA, USA. [3]Kavli Institute at Cornell for Nanoscale Science, Cornell University, Ithaca, NY, USA. [4]These authors contributed equally: Tianyu Wang, Mandar M. Sohoni. ✉e-mail: tw329@cornell.edu; mms477@cornell.edu; lgw32@cornell.edu; pmcmahon@cornell.edu

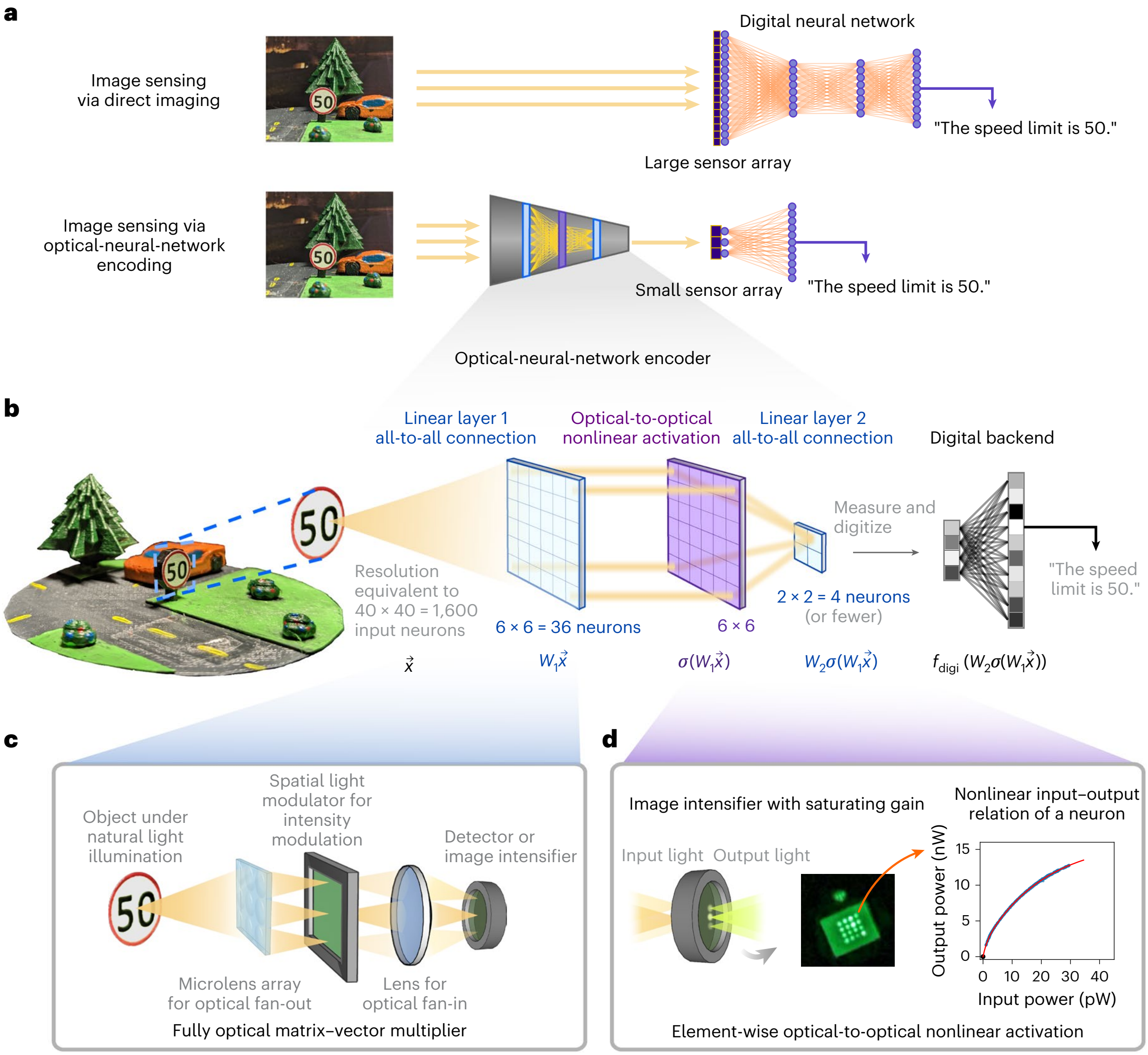


Fig. 1 | A multilayer optical-neural-network encoder as a frontend for image sensing. a, Image sensing via direct imaging versus optical encoding. In conventional image sensing, an image is collected by a camera and then processed, often using a neural network to extract a small piece of relevant information such as the text of a sign. Rather than faithfully reproducing the full image of a scene onto the sensor array, an ONN encoder instead pre-processes the image, compressing and extracting only the image information necessary for its end use, allowing a much smaller (fewer pixel) sensor array. As with more widely studied ONN inference accelerators, such a system can improve the speed or energy efficiency of neural-network-based machine-vision and image-based sensors. However, an important distinction is that an ONN image sensor takes a natural image as input—a pattern of incoherent photons scattered from a real object—and can improve sensor performance in ways that extend beyond latency and power consumption, such as effective resolution or sensitivity. **b**, The neural-network diagram and corresponding mathematical operations of the ONN encoder used in this study. The ONN encoder consists of interleaved linear and nonlinear layers before the compressed signal is captured by a small photodetector array. $\vec{x}$, input image; W_i, weight matrix of fully connected layer i; σ, optical-to-optical nonlinear activation function; f_{digi}, digital backend function. **c**, The schematic of the fully optical matrix–vector multiplier used for constructing both linear layers in **b**. **d**, The schematic of the optical-to-optical nonlinear activation layer realized with a saturating image intensifier. The inset plot shows that the output light intensity of a single spatial mode begins to saturate as the input light intensity increases, resembling the sigmoid activation function.

compressed sensing and single-pixel imaging[1,3,18,19], coded apertures[17,20,21] and related approaches for computational lensless imaging[22]. Related trends include the broader fields of smart cameras[23], in- and near-sensor computing[7,24,25], variational quantum sensors[26] and machine-learning-enabled smart sensors[27].

Optical encoders improve machine-vision systems by reducing the number of photodetectors. Many performance metrics such as frame rate and photon efficiency are directly bottlenecked by the number of pixels in the camera, including the energy and time costs of transducing images from the optical to digital electronic domain, of transporting them from the sensor to the post-processor, and performing high-dimensional digital post-processing[2,23]. Consequently, using a camera with C-fold fewer pixels typically leads to a C-fold improvement in the achievable frame rate, in the total number of photons required for each detector to achieve a high signal-to-noise ratio, and in the total system power and cost. Although high compression ratios ($C \gg 10$) are routinely achieved with electronic deep neural networks (DNNs), the computational capacity of simple optical encoders (such as single random or optimized masks) is rarely sufficient to realize such high compression.

Fortunately, much richer optical processing is possible with optical neural networks (ONNs)[6,28]—optoelectronic systems that perform mathematical operations involved in typical DNN inference calculations with optics. Optical neural networks are thus ideal for enabling a new class of image-sensing devices called ONN sensors[4–8,29], where an ONN pre-processes data from, and in, the analogue optical domain before its conversion into digital electronic signals. Unlike the complementary application of ONNs for accelerating deep-learning calculations on digital-domain data, the goal of ONN sensors is not only to accelerate calculations by replacing electronic operations with optical ones, but to also improve sensing performance, both by allowing faster, lower latency image processing, as well as by performing optical-domain computations that might be impractical or even impossible to perform after converting the photonic signal into a digital electronic one.

However, most optical image encoders experimentally demonstrated so far have involved only linear optical operations or, equivalently, a single-layer neural network[5,6,8,14,23]. Nonlinearity is essential for deep networks and high-performance image processing: multilayer, nonlinear networks are exponentially (in the number of neurons) more efficient than single-layer neural networks at approximating practically relevant functions[30]. There have been several promising proposals and proof-of-concept demonstrations for incorporating optoelectronic nonlinearity to enable multilayer ONNs[31–35]; for example, an integrated photonic system for all-optical, low-latency classification of images under laser illumination[36]. However, most machine-vision settings would require ONNs to process high-spatial-resolution images obtained with conventional or natural illumination, that is, patterns of broadband, incoherent light scattered from real, three-dimensional physical objects and scenes.

Here we demonstrate an optical neural network image sensor that uses optoelectronic optical-to-optical nonlinear activation (OONA) to perform multilayer ONN pre-processing for a variety of image-sensing applications. Our multilayer, nonlinear ONN pre-processor takes natural images (that is, patterns of incoherent photons scattered from real objects) as input, and conditionally compresses the image data into a low-dimensional latent feature space in a single shot, achieving compression ratios of up to 800:1. This allows image sensing to be performed with much simpler cameras (for example, a few pixels rather than millions of pixels), and vastly reduced digital post-processing and associated latency. At high compression ratios, our device consistently outperforms conventional image sensing and linear optical pre-processing on experiments based on standard machine-vision datasets, on flow-cytometry image classification, and for real scene object detection and measurement. The OONA used in our experiments is based on a commercial image intensifier typically used, for example, in night-vision goggles or low-light scientific imaging. Broadly, our findings support the use of multilayer ONNs with nonlinear activations as optical-domain pre-processors for sensors. Given the numerous ONN platforms[36–38] being developed, we expect that a variety of deep ONN sensors are possible; these future sensors may detect information encoded in light's spatial, spectral, and/or temporal degrees of freedom.

Results

ONN-based image sensors with optical-to-optical nonlinearity

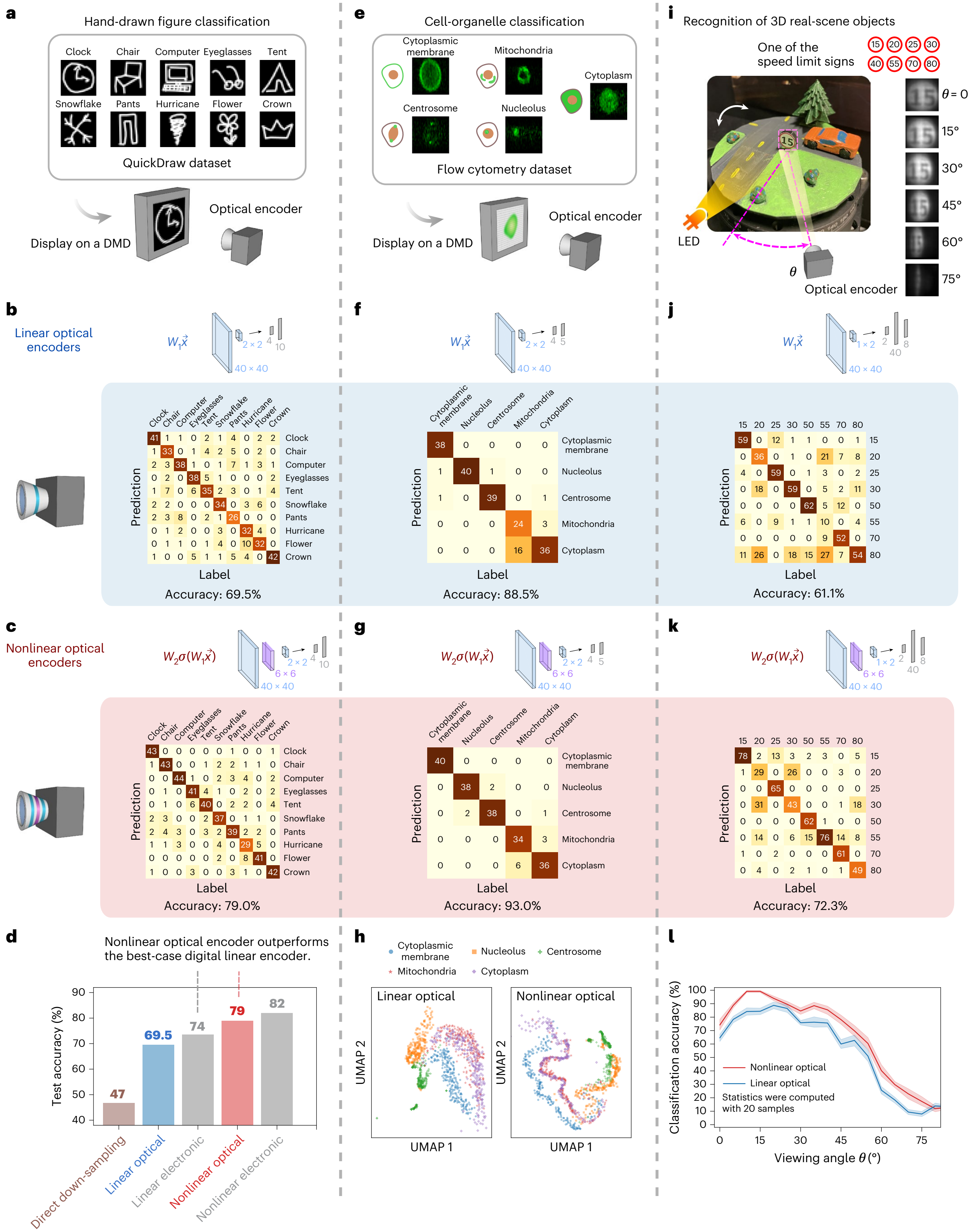
Our experimental ONN image sensor consists of two fully connected optical linear layers with an element-wise OONA layer in-between them (Fig. 1b). The linear layers (matrix–vector multiplications) in our ONN are implemented using a technique designed to facilitate broadband, incoherent light as direct inputs. Optical fully connected matrix–vector multiplications are performed using a method similar to past works[38,39] (see Methods and Supplementary Note 2). Natural input images are first fanned out (multiple spatially distinct copies of the input images are created) by an array of microlenses. Multiplication is then achieved by attenuating the copies of the input image in proportion to the components of the weight matrix, which can be typically implemented with a spatial light modulator (SLM) for intensity modulation. Finally, the summation of each output vector element is realized by focusing the attenuated light components using a lens (Fig. 1c). To realize the OONA operations applied to each element of this output vector, light is focused onto a commercial image intensifier tube. Incident light generates free electrons from a photocathode, which are locally amplified by a microchannel plate (MCP) and then produce new, amplified bright spots as they strike a phosphor screen[40]. The local saturation of the MCP's amplification leads to a saturating nonlinear response that is qualitatively similar to the positive half of the sigmoid function (Fig. 1d and Supplementary Fig. 8). Although the OONA is optoelectronic rather than all-optical, its local, in-place realization preserves the spatial parallelism of the ONN, and avoids the time and energy costs required for read-out/in when the nonlinear activation is computed on a separate electronic processor[29,38,39]. To implement the second layer of the ONN, the light produced by the intensifier is processed by a second copy of the optical matrix–vector multiplier depicted in Fig. 1c. The output from this layer (a four-dimensional vector) was detected by a camera (see Methods), but in principle can be captured by an array of four photodetectors.

Nonlinear encoders are more efficient than linear encoders

We first performed several image-classification tasks to evaluate the performance of the multilayer, nonlinear ONN encoder (Fig. 2). As a benchmark, we trained classifiers for ten pre-selected classes of the Quick, Draw! (QuickDraw) image dataset[41]. By placing a beamsplitter before the intensifier OONA, we could reconfigure the ONN image sensor for direct imaging (by setting the SLM of the first linear layer to be transparent) and for single-layer linear encoding (by applying linear layer weights to the first SLM for intensity modulation). Input images (28 × 28 pixels) were binarized and displayed on a digital micromirror display (DMD), which was placed in front of the image sensor working in nonlinear multilayer, linear single-layer or direct imaging mode. For a direct comparison, the vector dimension at the optical electronic bottleneck in each sensor is the same—a 2 × 2 array or four-dimensional latent space, which represents a 196:1 image compression ratio (in the direct imaging mode, the images were directly down-sampled by binning the four quadrants of the image into four pixels). The multilayer, nonlinear ONN encoder achieved better classification accuracy than the linear ONN encoder and direct downsampling of images (Fig. 2b–d). To ensure that the accuracy advantage of the nonlinear, multilayer ONN encoder over linear encoders is consistent for any possible linear encoder with the same bottleneck dimension, we also trained all-digital (with real number weights and biases) single-layer linear encoders for the same task, without image downsampling (Fig. 2d). Despite the

Fig. 2 | Comparison between linear and nonlinear ONN encoders on diverse image classification tasks. a, Classification of hand-drawn figures from ten different classes in the QuickDraw dataset. **b**,**c**, The results of QuickDraw[41] classification with a linear (**b**) or nonlinear (**c**) ONN encoder as the frontend. The neural-network architecture with corresponding mathematical operations is placed above the confusion matrix it produces (blue slabs, linear optical neurons; purple slabs, nonlinear activations; grey bars, digital neurons). **d**, Comparison of the accuracy derived from classifiers equipped with different frontends. In all cases, the encoder's output dimension (number of pixels) is 4. **e**, Classification of HeLa cells labelled for different organelles from a dataset acquired from flow-cytometry experiments[42]. **f**,**g**, The results of cell-organelle classification with a linear (**f**) or a nonlinear (**g**) ONN encoder as the frontend. **h**, Visualization of the compressed cell-organelle data with density uniform manifold approximation and projection (DensMAP)[57]. **i**, Recognition of three-dimensional objects: each of eight 3D-printed speed-limit signs are viewed from different perspectives by an ONN encoder, which classifies the speed-limit number on the sign. **j**,**k**, The results of classifying speed limits with a linear (**j**) or a nonlinear (**k**) ONN encoder as the frontend. **l**, Classification accuracy as a function of the viewing angle, θ. The shaded area denotes 1 s.d. from the mean for repeated classification tests.

constraint of non-negative weights and the non-analytical form of our OONA, the experimental, multilayer nonlinear ONN encoder's performance (79% test accuracy) surpassed that of linear encoders, beating both the optical (69.5%) and optimized digital (74%) single-layer encoders. Compared with an ideal digital multilayer encoder with real-valued weights and biases and a sigmoid nonlinear activation function, the



b

Prediction \ Label	Clock	Chair	Computer	Eyeglasses	Tent	Snowflake	Pants	Hurricane	Flower	Crown
Clock	41	1	1	0	2	1	4	0	2	2
Chair	1	33	0	1	4	2	5	0	2	0
Computer	2	3	38	1	0	1	7	1	3	1
Eyeglasses	0	2	0	38	5	1	0	0	0	2
Tent	1	7	0	6	35	2	3	0	1	4
Snowflake	2	2	0	0	0	34	0	3	6	0
Pants	2	3	8	0	2	1	26	0	0	0
Hurricane	0	1	2	0	0	3	0	32	4	0
Flower	0	1	1	0	1	4	0	10	32	0
Crown	1	0	0	5	1	1	5	4	0	42

Accuracy: 69.5%

f

Prediction \ Label	Cytoplasmic membrane	Nucleolus	Centrosome	Mitochondria	Cytoplasm
Cytoplasmic membrane	38	0	0	0	0
Nucleolus	1	40	1	0	0
Centrosome	1	0	39	0	1
Mitochondria	0	0	0	24	3
Cytoplasm	0	0	0	16	36

Accuracy: 88.5%

j

Prediction \ Label	15	20	25	30	50	55	70	80
15	59	0	12	1	1	0	0	1
20	0	36	0	1	0	21	7	8
25	4	0	59	0	1	3	0	2
30	0	18	0	59	0	5	2	11
50	0	0	0	0	62	5	12	0
55	6	0	9	1	1	10	0	4
70	0	0	0	0	0	9	52	0
80	11	26	0	18	15	27	7	54

Accuracy: 61.1%

c

Prediction \ Label	Clock	Chair	Computer	Eyeglasses	Tent	Snowflake	Pants	Hurricane	Flower	Crown
Clock	43	0	0	0	0	0	1	0	0	1
Chair	1	43	0	0	1	2	2	1	1	0
Computer	0	0	44	1	0	2	3	4	0	2
Eyeglasses	0	1	0	41	4	1	0	2	0	2
Tent	0	1	0	6	40	0	2	2	0	4
Snowflake	2	3	0	0	2	37	0	1	1	0
Pants	2	4	3	0	3	2	39	2	2	0
Hurricane	1	1	3	0	0	4	0	29	5	0
Flower	0	0	0	0	0	2	0	8	41	0
Crown	1	0	0	3	0	0	3	1	0	42

Accuracy: 79.0%

g

Prediction \ Label	Cytoplasmic membrane	Nucleolus	Centrosome	Mitochondria	Cytoplasm
Cytoplasmic membrane	40	0	0	0	0
Nucleolus	0	38	2	0	0
Centrosome	0	2	38	0	1
Mitochondria	0	0	0	34	3
Cytoplasm	0	0	0	6	36

Accuracy: 93.0%

k

Prediction \ Label	15	20	25	30	50	55	70	80
15	78	2	13	3	2	3	0	5
20	1	29	0	26	0	0	3	0
25	0	0	65	0	0	0	0	0
30	0	31	0	43	0	0	1	18
50	0	0	0	0	62	1	0	0
55	0	14	0	6	15	76	14	8
70	1	0	2	0	0	0	61	0
80	0	4	0	2	1	0	1	49

Accuracy: 72.3%

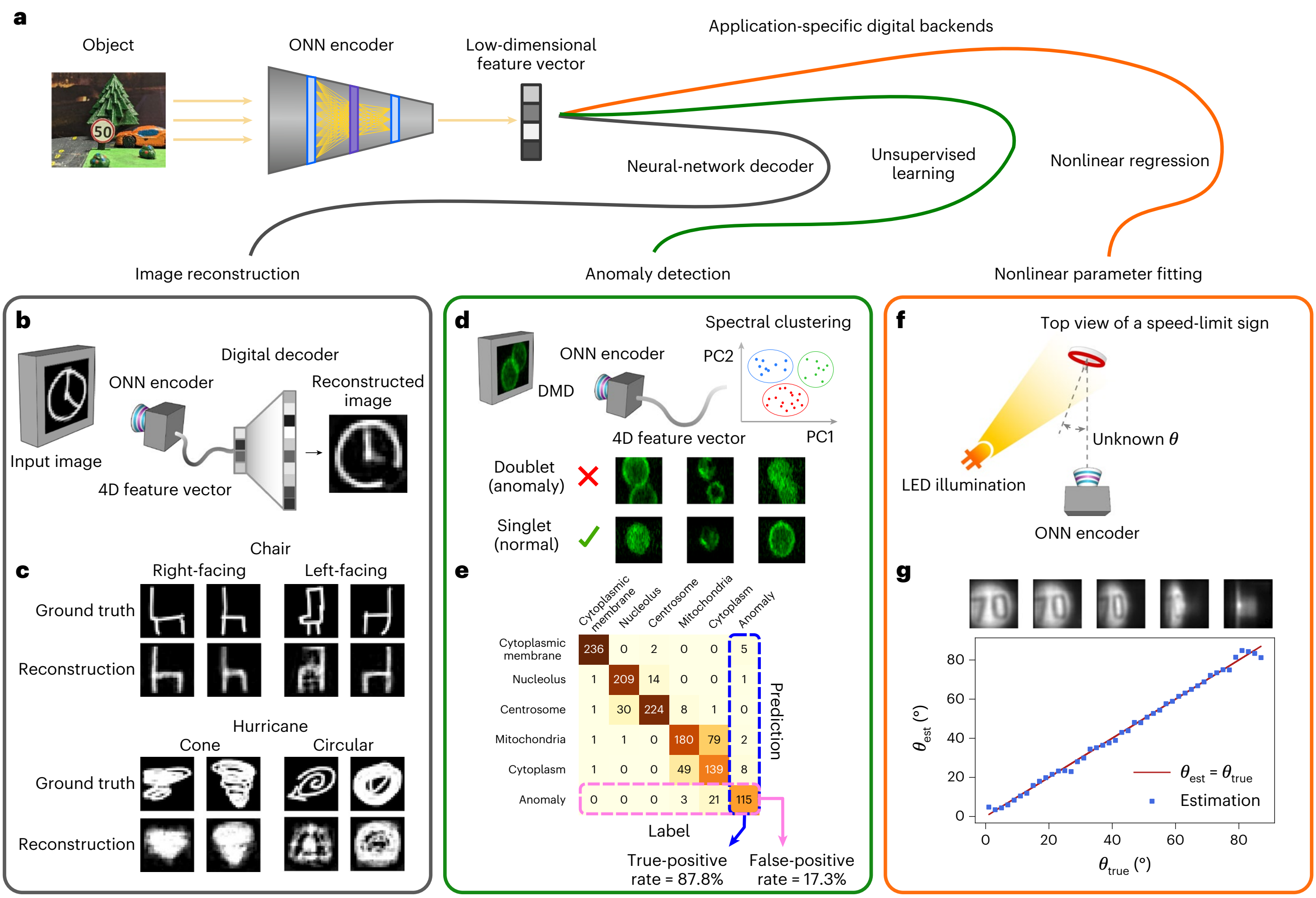


Fig. 3 | Nonlinear ONN encoders trained for classification can be reused for diverse image-sensing tasks by training only new digital backends. a, New image-sensing tasks can be performed by using the feature vectors produced by the nonlinear ONN encoders trained for classification as input to new digital backends. **b**, Images from the QuickDraw dataset were reconstructed by training a new digital decoder to reconstruct rather than classify images from the feature vectors. The encoder is exactly the same ONN, including weights, as in Fig. 2c,d. **c**, Although the encoder was only trained to preserve class information, randomly selected reconstructed images show that feature information, such as the direction or shape of the chairs and hurricanes, is preserved. **d**,**e**, By performing unsupervised clustering (see Methods) on the feature vector produced by the cell-organelle-classifier ONN frontend from Fig. 2g (**d**), we can accurately detect anomalous doublet images that were not part of the encoder's training set (**e**). **f**, We trained a new digital backend to, rather than classify the content of a speed-limit sign, use the speed-sign classifier's feature vector to infer the viewing angle of the sign. **g**, The speed-limit images above the viewing angle inference plot refer to the ground-truth images at a few different viewing angles. θ_{est}, estimated viewing angle; θ_{true}, true viewing angle.

experimental nonlinear ONN encoder has a slightly lower test accuracy (79% versus 82%). For single-shot processing of incoherent light, ONNs are restricted to non-negative weights. Although this is not a severe limitation for the tasks considered here (see Supplementary Note 16), it may need to be addressed in future work.

To explore the potential of ONN image sensors for a more practically important application, we next tested our image sensors on the task of classifying fluorescent images of cell organelles acquired in a flow-cytometry device[42]. Image-based flow cytometry is an emerging technique in which cells travel through a fluidic channel and are sorted, ideally one-by-one, on the basis of their, for example, fluorescence and/or phase images[42–44]. To process statistically useful collections of cells, so as to detect, for instance, extremely rare cancerous cells, it is essential to minimize the latency of each sorting decision, maintaining a high throughput of, for example, 100,000 cells per second[42–44]. In our experiments we displayed binarized images from the dataset in ref. [42] on the DMD and performed classification with each ONN encoder, as in the QuickDraw experiments (Fig. 2e). When each cell image was compressed to a four-dimensional feature vector, the multilayer, nonlinear ONN encoder exhibited a better classification accuracy for the five considered classes than that of the linear ONN encoder (93% versus 88.5% test accuracy, Fig. 2f,g; higher local density within clusters, Fig. 2h).

Although it is helpful in improving both the accuracy and flexibility of our ONN sensors, the small digital post-processing layer we employ in these networks is not a necessity and can be eliminated if applications require a particularly short latency. For the same flow-cytometry task considered above, we show that all-optical classification (that is, classification without an electronic digital backend) is also possible (see Supplementary Note 13). Even in this case, nonlinear optical classifiers outperform linear optical classifiers.

The two tasks considered so far are effectively experimental simulations of image-sensing tasks; real image-sensing tasks involve directly processing photons arriving from real three-dimensional objects. To test this setting, we applied the image sensors to the task of classifying traffic signs in a real-model scene, the three-dimensionally printed intersection shown in Fig. 2i. Due to the limited field-of-view of the particular microlens array used in this experiment, the input images to the image sensors (insets of Fig. 2g) primarily contain only the speed limit sign being classified. The nonlinear, multilayer ONN encoder results

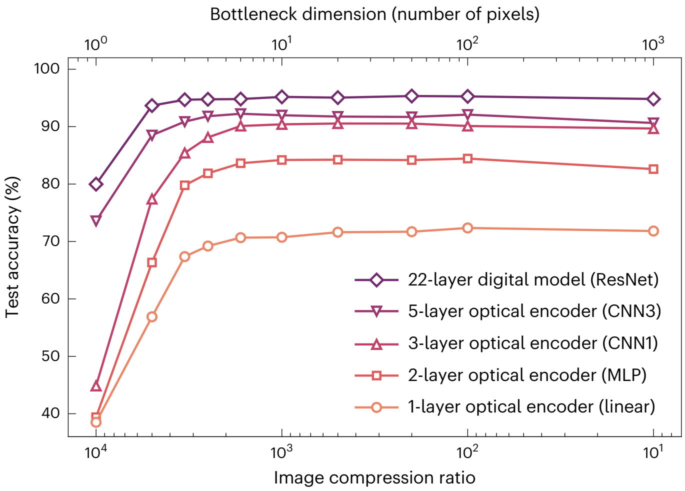


Fig. 4 | Simulations of performance scaling with deeper nonlinear optical neural network encoders for ten-class cell-organelle classification. Classification accuracy as a function of image compression ratio (or bottleneck feature vector dimension) for all of the models. Linear, single-layer (linear) ONN encoder; multilayer perceptron (MLP), a nonlinear encoder with two fully connected layers; CNN1, a three-layer nonlinear ONN encoder with a convolutional layer followed by two fully connected layers; CNN3, a five-layer nonlinear ONN encoder with three convolutional and two fully connected layers. The ResNet-based model is a state-of-the-art digital model shown here as an estimate of the upper bound on performance at each compression ratio. Deeper models generally produce higher accuracy, especially at higher compression ratios.

in better identification of the speed limit than the linear ONN encoder across a range of viewing angles from 0° to 80° (Fig. 2j–l).

A versatile optical frontend for assorted vision tasks

By training new digital post-processers only, the same optical encoders trained for classification in the previous section can be reused for a variety of other image-sensing tasks. If suitably trained (see Methods), encoders can produce robust representations of high-dimensional images in the low-dimensional latent space, which preserve far more information than the bare minimum required for classification. For example, although the QuickDraw-classification encoder (Fig. 2a–c) was trained only to facilitate classification, the feature space evidently preserves more complex attributes of the original images beyond just the figure's class. When a digital decoder is trained to reconstruct QuickDraw images from the classification encoder's features (Fig. 3b,c), it produces reconstructions that—although often lacking specific details, such as the position of the clock's hands—capture coarse intra-class details such as the orientation or shape of chairs and hurricanes. Although nonlinear encoders generally enable improved image reconstruction performance[45], this is not necessarily the case for all datasets or models. In the case of the QuickDraw dataset considered here, we find only a marginal benefit from nonlinear encoding (Supplementary Fig. 25).

As another example of the versatility of optical image encoding, using the same multilayer ONN encoder previously trained for traffic-sign classification (Fig. 2i–l), we trained a new digital backend to predict the angle at which a traffic sign was viewed (Fig. 3f,g). The resulting predictions are very accurate, although the performance is reduced if the network is required to predict the viewing angle for all of the speed-limit classes, rather than just one at a time (Supplementary Fig. 27).

Finally, in many image-sensing applications, initial device training will not be able to account for edge cases that may be encountered in deployment. To test the capacity for detecting anomalies not previously observed (and on which the optical encoder was not trained), we introduced anomalous images of doublet cell clusters to the ONN image sensor (Fig. 3d). To detect these anomalies, we applied spectral clustering to the normalized four-dimensional feature vectors produced by the ONN encoder previously trained for cell-organelle classification (see Methods). By identifying the six most prominent clusters as the five trained classes, plus one last class corresponding to anomalous images, we were able to adapt the digital decoder to reliably identify anomalous images in the test set (Fig. 3e). These results show that the nonlinear ONN encoder does not overfit to the initial training dataset, but instead preserves important data structure beyond the initially chosen classes, while still compressing the original images to a low-dimensional space.

Deeper ONN image sensors for more complex tasks

The results presented in Figs. 2 and 3 illustrate that a two-layer nonlinear ONN pre-processor enables consistently better image-sensing performance across a wide range of tasks than conventional imaging with direct downsampling or linear ONN pre-processing. Nonetheless, an ONN encoder with two fully connected layers is merely a first step. A key motivation for using an OONA is that it will facilitate even deeper ONN encoders. To explore what may soon be possible with deeper, nonlinear ONN encoders, we performed realistic simulations of four different optical pre-processors (see Fig. 4), performing an extended (ten-class) version of the organelle classification task considered in Figs. 2 and 3 (Supplementary Fig. 28). This dataset—which is more challenging than the five-class cell-organelle classification demonstrated in earlier experiments—allowed us to study the performance of more complicated ONN encoders. Our simulations (see Methods for details) consider physical noise, and involve strictly non-negative weights, which is a critical constraint for ONNs operating on incoherent light, such as fluorescence.

Figure 4 shows how the classification accuracy of the different ONN pre-processors varies as the compression ratio is changed. The compression ratio is changed by modifying the number of output neurons in the final optical layer, which determines the number of pixels or photodetectors required on the photosensor. As a reference for achievable performance, we also performed the task with a fully digital classifier based on a ResNet model (an 18-layer pretrained ResNet plus four additional adapting layers)[46]. All networks, including the all-digital reference, have the same single-layer digital decoder architecture.

The key result in Fig. 4 is that deeper ONNs with multiple nonlinear layers lead to progressively better classification performance across a wide range of compression ratios. The benefit of pre-processor depth becomes especially evident at very high compression ratios: for a compression ratio of 10^4 (bottleneck dimension 1), the five-layer pre-processor (CNN3) achieves nearly double the accuracy of shallower networks.

Discussion

We demonstrated a nonlinear ONN system that can—in a single shot, and without relying on a separate digital electronic processor to implement the nonlinearity—perform a variety of nonlinear image processing tasks on natural images, that is, on patterns of incoherent photons scattered from a real object. By performing image compression in the optical domain, ONN image sensors can fundamentally bypass the optoelectronic bandwidth limit of high-resolution cameras, allowing for faster, more sensitive and more efficient machine-vision systems. Our results show that the nonlinear processing capacity of ONNs enables image sensors to outperform image sensors based either on direct downsampling of conventional images or purely linear optical pre-processing. We also see that the performance advantages of nonlinear ONN encoders scale favourably with additional layers of ONN pre-processing. Such nonlinear optical encoders extend the paradigm of end-to-end image system optimization[5,8,10–17]. Given the numerous promising optical[47–50] and optoelectronic nonlinearities proposed for ONNs, we are optimistic about future prospects for compact, scalable multilayer ONN image sensors (see Supplementary Note 15).

An important benefit of ONN sensors is their potential for dramatically reduced latency in high-speed control scenarios such as manufacturing robotics, human–machine interfaces, active flow and plasma stabilization, and defence applications. For ONN-based image sensors, capturing and parallel processing of the largest number of spatial modes motivates the use of free-space systems, an approach employed by this work. Nonetheless, implementing more than a few layers in this format will eventually encounter trade-offs with respect to system size (see Supplementary Note 15 and Supplementary Table 9). The output of free-space layers are still in the optical domain, however, so a promising solution is to route the compressed optical feature vector directly from a free-space layer to integrated-photonics neural networks for further optical processing, rather than to digital electronics[6,28,36,51,52]. The resulting all-optical intelligent sensors could entirely bypass electronic bottlenecks on speed, sensitivity and resolution, and could one day operate with multigigahertz bandwidth, gigapixel effective spatial resolution and subnanosecond-scale latency.

Finally, beyond allowing information encoded in many spatial modes to be transmitted into just a few pixels, ONN image sensors are also exciting because they may be sensitive to other optical information that is traditionally lost in photodetection, such as hyperspectral[53] and vectorial (for example, ray direction) information, both of which can drastically change what information can be extracted from a scene.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41566-023-01170-8.

References

1. Duarte, M. F. et al. Single-pixel imaging via compressive sampling. *IEEE Signal Process. Mag.* **25**, 83–91 (2008).
2. Sterling, P. *Principles of Neural Design* (MIT Press, 2015).
3. Gibson, G. M., Johnson, S. D. & Padgett, M. J. Single-pixel imaging 12 years on: a review. *Opt. Express* **28**, 28190–28208 (2020).
4. Chang, J., Sitzmann, V., Dun, X., Heidrich, W. & Wetzstein, G. Hybrid optical-electronic convolutional neural networks with optimized diffractive optics for image classification. *Sci. Rep.* **8**, 1–10 (2018).
5. Martel, J. N. P., Mueller, L. K., Carey, S. J., Dudek, P. & Wetzstein, G. Neural sensors: learning pixel exposures for HDR imaging and video compressive sensing with programmable sensors. *IEEE Trans. Pattern Anal. Mach. Intell.* **42**, 1642–1653 (2020).
6. Wetzstein, G. et al. Inference in artificial intelligence with deep optics and photonics. *Nature* **588**, 39–47 (2020).
7. Mennel, L. et al. Ultrafast machine vision with 2D material neural network image sensors. *Nature* **579**, 62–66 (2020).
8. Pad, P. et al. Efficient neural vision systems based on convolutional image acquisition. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition* 12285–12294 (IEEE, 2020).
9. Zheng, H. et al. Meta-optic accelerators for object classifiers. *Sci. Adv.* **8**, eabo6410 (2022).
10. Matic, R. M. & Goodman, J. W. Comparison of optical predetection processing and postdetection linear processing for partially coherent image estimation. *J. Opt. Soc. Am. A* **6**, 213–228 (1989).
11. Kubala, K., Dowski, E. & Cathey, W. T. Reducing complexity in computational imaging systems. *Opt. Express* **11**, 2102–2108 (2003).
12. Stork, D. G. & Robinson, M. D. Theoretical foundations for joint digital-optical analysis of electro-optical imaging systems. *Appl. Opt.* **47**, B64–B75 (2008).
13. Sitzmann, V. et al. End-to-end optimization of optics and image processing for achromatic extended depth of field and super-resolution imaging. *ACM Trans. Graph.* **37**, 1–13 (2018).
14. Colburn, S., Chu, Y., Shilzerman, E. & Majumdar, A. Optical frontend for a convolutional neural network. *Appl. Optics* **58**, 3179–3186 (2019).
15. Kim, K., Konda, P. C., Cooke, C. L., Appel, R. & Horstmeyer, R. Multi-element microscope optimization by a learned sensing network with composite physical layers. *Opt. Lett.* **45**, 5684–5687 (2020).
16. Markley, E., Liu, F. L., Kellman, M., Antipa, N. & Waller, L. Physics-based learned diffuser for single-shot 3D imaging. In *NeurIPS 2021 Workshop on Deep Learning and Inverse Problems* (NeurIPS, 2021).
17. Vargas, E., Martel, J. N. P., Wetzstein, G. & Arguello, H. Time-multiplexed coded aperture imaging: learned coded aperture and pixel exposures for compressive imaging systems. In *Proc. IEEE/CVF International Conference on Computer Vision* 2692–2702 (IEEE, 2021).
18. Liutkus, A. et al. Imaging with nature: compressive imaging using a multiply scattering medium. *Sci. Rep.* **4**, 1–7 (2014).
19. Li, J. et al. Spectrally encoded single-pixel machine vision using diffractive networks. *Sci. Adv.* **7**, eabd7690 (2021).
20. Asif, M. S., Ayremlou, A., Sankaranarayanan, A., Veeraraghavan, A. & Baraniuk, R. G. Flatcam: thin, lensless cameras using coded aperture and computation. *IEEE Trans. Comput. Imaging* **3**, 384–397 (2016).
21. Baek, S.-H. et al. Single-shot hyperspectral-depth imaging with learned diffractive optics. In *Proc. IEEE/CVF International Conference on Computer Vision* 2651–2660 (IEEE, 2021).
22. Sinha, A., Lee, J., Li, S. & Barbastathis, G. Lensless computational imaging through deep learning. *Optica* **4**, 1117–1125 (2017).
23. Brady, D. J. et al. Smart cameras. Preprint at https://arxiv.org/abs/2002.04705 (2020).
24. Burt, P. J. Smart sensing within a pyramid vision machine. *Proc. IEEE* **76**, 1006–1015 (1988).
25. Zhou, F. & Chai, Y. Near-sensor and in-sensor computing. *Nat. Electron.* **3**, 664–671 (2020).
26. Marciniak, C. D. et al. Optimal metrology with programmable quantum sensors. *Nature* **603**, 604–609 (2022).
27. Ballard, Z., Brown, C., Madni, A. M. & Ozcan, A. Machine learning and computation-enabled intelligent sensor design. *Nat. Mach. Intell.* **3**, 556–565 (2021).
28. Shastri, B. J. et al. Photonics for artificial intelligence and neuromorphic computing. *Nat. Photon.* **15**, 102–114 (2021).
29. Li, H.-Y. S., Qiao, Y. & Psaltis, D. Optical network for real-time face recognition. *Appl. Opt.* **32**, 5026–5035 (1993).
30. Lin, H. W., Tegmark, M. & Rolnick, D. Why does deep and cheap learning work so well? *J. Stat. Phys.* **168**, 1223–1247 (2017).
31. Wagner, K. & Psaltis, D. Multilayer optical learning networks. *Appl. Opt.* **26**, 5061–5076 (1987).
32. Zuo, Y. et al. All-optical neural network with nonlinear activation functions. *Optica* **6**, 1132–1137 (2019).
33. Fard, M. M. P. et al. Experimental realization of arbitrary activation functions for optical neural networks. *Opt. Express* **28**, 12138–12148 (2020).
34. Ryou, A. et al. Free-space optical neural network based on thermal atomic nonlinearity. *Photon. Res.* **9**, B128–B134 (2021).
35. Feldmann, J., Youngblood, N., Wright, C. D., Bhaskaran, H. & Pernice, W. H. All-optical spiking neurosynaptic networks with self-learning capabilities. *Nature* **569**, 208–214 (2019).

36. Ashtiani, F., Geers, A. J. & Aflatouni, F. An on-chip photonic deep neural network for image classification. *Nature* **607**, 501–506 (2022).
37. Zhou, T. et al. Large-scale neuromorphic optoelectronic computing with a reconfigurable diffractive processing unit. *Nat. Photon.* **15**, 367–373 (2021).
38. Bernstein, L. et al. Single-shot optical neural network. Preprint at https://arxiv.org/abs/2205.09103 (2022).
39. Wang, T. et al. An optical neural network using less than 1 photon per multiplication. *Nat. Commun.* **13**, 1–8 (2022).
40. Zemel, J. N. *Sensors V 6—Optical Sensors—A Comprehensive Survey* (John Wiley & Sons, 1991).
41. Jongejan, J., Rowley, H., Kawashima, T., Kim, J. & Fox-Gieg, N. *The Quick, Draw! AI Experiment* https://quickdraw.withgoogle.com/ (2016).
42. Schraivogel, D. et al. High-speed fluorescence image–enabled cell sorting. *Science* **375**, 315–320 (2022).
43. Li, Y. et al. Deep cytometry: deep learning with real-time inference in cell sorting and flow cytometry. *Sci. Rep.* **9**, 1–12 (2019).
44. Lee, K. C. M., Guck, J., Goda, K. & Tsia, K. K. Toward deep biophysical cytometry: prospects and challenges. *Trends Biotechnol.* **39**, 1249–1262 (2021).
45. Hinton, G. E. & Salakhutdinov, R. R. Reducing the dimensionality of data with neural networks. *Science* **313**, 504–507 (2006).
46. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition* 770–778 (IEEE, 2016).
47. Li, G. H.et al. All-optical ultrafast ReLU function for energy-efficient nanophotonic deep learning. *Nanophotonics* https://doi.org/10.1515/nanoph-2022-0137 (2022).
48. Guo, Q. et al. Femtojoule femtosecond all-optical switching in lithium niobate nanophotonics. *Nat. Photon.* **16**, 625–631 (2022).
49. Nahmias, M. A., Shastri, B. J., Tait, A. N. & Prucnal, P. R. A leaky integrate-and-fire laser neuron for ultrafast cognitive computing. *IEEE J. Sel. Top. Quantum Electron.* **19**, 1–12 (2013).
50. Mirek, R. et al. Neural networks based on ultrafast time-delayed effects in exciton polaritons. *Phys. Rev. Appl.* **17**, 054037 (2022).
51. Bandyopadhyay, S. et al. Single chip photonic deep neural network with accelerated training. Preprint at https://arxiv.org/abs/2208.01623 (2022).
52. Xu, X. et al. 11 TOPS photonic convolutional accelerator for optical neural networks. *Nature* **589**, 44–51 (2021).
53. Makarenko, M. et al. Real-time hyperspectral imaging in hardware via trained metasurface encoders. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition* 12692–12702 (IEEE, 2022); https://doi.org/10.1109/CVPR52688.2022.01236
57. Narayan, A., Berger, B. & Cho, H. Assessing single-cell transcriptomic variability through density-preserving data visualization. *Nat. Biotechnol.* **39**, 765–774 (2021).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Methods

Multilayer optical-neural-network image pre-processor

The ONN pre-processor (Supplementary Fig. 1a) comprises an optical matrix–vector multiplier unit, an OONA unit, a second optical matrix–vector multiplier, and a camera. Light is detected in the compressed, low-dimensional latent space on the camera, and is subsequently digitally post-processed (see Supplementary Note 1 for more details).

The optical matrix–vector multiplier treats an image with N pixels as an N-dimensional vector and multiplies it with a user-specified matrix. To implement the matrix–vector product between an N by N' matrix, W and an N-dimensional input vector, we take the following steps, which are also illustrated graphically in Supplementary Fig. 3a. First, the input image (vector) is fanned-out to create N' identical copies. This is realized by using a microlens array (MLA) to form N' identical images on regions of an SLM. Second, each optically fanned-out copy of the image covers N pixels on the SLM, and the intensity of each image was modulated in an element-wise fashion according to a different column of matrix W. Finally, after the intensity modulation by the SLM, which implements the weight multiplication, the intensity-modulated image copies are optically fanned-in by forming a demagnified image of the N' copies onto an image intensifier or a camera. Provided that the size of the focused image of each attenuated copy is smaller than the resolution of the image intensifier (or the size of the camera superpixel), the photoelectrons generated by each optical copy are pooled to achieve the summation step of the matrix–vector multiplication for each row, producing the N'-dimensional output vector.

For the optical matrix–vector multiplier that implements the first fully connected layer, the square MLA array has a pitch of 1.1 ± 0.001 mm and a focal length of 128.8 mm (APO-Q-P1100-F105, OKO Optics). This first MLA contains 26 × 26 = 676 lenslets altogether, but we used only 6 × 6 = 36 of them to create 36 optical copies of the input image, limited by several practical constraints detailed in Supplementary Note 2. For the second fully connected layer, the MLA has a rectangular pitch of 4 mm × 3 mm and a focal length of 38.10 mm (no. 63–230, Edmund Optics). The weights of each layer are stored as pixel values on a liquid crystal display (LCD, Sony LCX029, with LCX017 controllers by Bild- und Lichtsysteme GmbH). The LCDs were operated as transmissive intensity modulation SLMs by placing two polarizers—oriented at +45° and −45° relative to the pixel grid of the LCD—before and after the LCD panel. The transmission was calibrated as a function of the LCD pixel value. The calibration procedure for the LCD-based matrix–vector multipliers is described in Supplementary Note 2. Under white-light illumination, the extinction ratio of the LCD pixels was measured to be at least 400, and the LCD can provide 256 discrete modulation levels.

The optical fan-in for the first layer was implemented by demagnifying optical fan-out copies after they were modulated by an LCD. The demagnification factor of ×30 was implemented by a 4f imaging system composed of a singlet lens (LA1484-A-ML, Thorlabs; f = 300 mm) and an objective lens (MY20X-804, 20x, Mitutoyo; f = 10 mm). The optical fan-in of the second layer was performed using a zoom lens (Zoom 7000, Navitar) and imaged onto a camera (Prime 95B Scientific CMOS Camera, Teledyne Photometrics). The pixels values were summed digitally after read-out, but could equivalently be summed in an analogue fashion by binning camera pixels or by using larger pixels/photodetectors.

Among the two optical fully connected layers, the optical transmission of the first layer is critical to the sensitivity of the ONN image sensor because it directly operates on the limited amount of light from the physical environment. The optical transmission of our first matrix–vector multiplier was measured to be 2.9%. Through detailed analysis (Supplementary Note 5), we estimated that the pure optical transmission (that is, without adding any additional attenuation by setting all of the pixels on the LCD to the maximum transmission) can be improved to close to 50% with customized optical elements. The 50% hard limit is due to the inevitable loss of half of the power when incoherent light goes through a polarizer. Even higher (at least 90%) transmission should be possible for devices that can assume coherent illumination, or that perform incoherent spatial light modulation without polarizers.

The OONA after the first matrix–vector multiplication was realized with a commercial image intensifier tube (MCP125/Q/S20/P46/GL, Photek). The image intensifier provides large input-output gains (around 800 in our work), a crucial feature for multilayer networks and low-light operation. A more subtle feature of the image intensifier's OONA is that it resets the number of spatial optical modes: even though the number of modes incident to the photocathodes is equal to the number of weights in the weight matrix NN', the number of distinct output beams is only equal to the output vector size N'.

The device, and its local nonlinearity, operates as follows. In the image intensifier, light is collected on a photocathode, which produces photoelectrons in proportion to the local input light intensity. These photoelectrons are then locally amplified with a microchannel plate (MCP). The amplified photoelectrons in each channel then excite photons on a phosphor screen, producing the light input to the next layer. The saturation of this input-output response results in the nonlinearity used in our ONN encoders. The image intensifier used in our experiments is from Photek, and includes a S20 photocathode, one-stage MCP and P46 phosphor. We find that the nonlinearity of intensifier varies slightly from channel to channel, so we calibrated the input-output response for all 36 illuminated regions separately (Supplementary Fig. 8), fitting them each to a curve of the form $y = a(1 - e^{-bx}) + c(1 - e^{-dx})$, where a, b, c, d are fit parameters for each region. The intensifier's response time was measured to be approximately 20 µs (Supplementary Fig. 7).

For most experiments in this work, the ONN device and architecture are similar: the input is a 1,600 (40 × 40 pixels) image, and the first fully connected layer consists of a 1,600 × 36 weight matrix, whereas the second fully connected layer—after the optical-to-optical nonlinearity—usually comprised a 36 × 4 weight matrix, except for the traffic-sign classification task, which used a 36 × 2 weight matrix. The convention for matrix size used throughout this paper is: the first dimension is the length of input vector or the number of neurons in the input layer, and the second dimension is the output vector dimension or the number of neurons in the output layer. The effective input image size equals the number of LCD pixels each optically fanned-out copy of the input image covers on the first LCD, which is used as a transmissive SLM for element-wise multiplication.

To monitor the light at intermediate locations in the ONN pre-processor, and to enable us to perform experiments with direct imaging and single-layer ONN pre-processing, we included a beamsplitter (BP245B1, Thorlabs) after the first LCD, and another (BS013, Thorlabs) immediately after the image intensifier. Each beamsplitter directs part of the light to a monitoring camera, which enabled us to observe several intermediate steps of computation. The full experimental set-up is depicted in Supplementary Fig. 2.

QuickDraw image classification

We chose the QuickDraw dataset[4,41] to benchmark the performance of the encoders as it: (1) is much harder than the MNIST dataset; and (2) can be binarized and displayed on a DMD without substantial loss of image information. Ten classes (clock, chair, computer, eyeglasses, tent, snowflake, pants, hurricane, flower, crown) were chosen arbitrarily (by hand, but with no deliberate rationale other than to ensure the classes were not too similar) from the available 250+ classes. Inappropriate images, or images that were obviously not of the intended class, were removed by hand. The first 300 images remaining for each class were used for the training set (total size 3,000) with a random train–validation split of 250:50, whereas the next 50 were used for testing (total size 500). This dataset is included with all other data for this manuscript at https://doi.org/10.5281/zenodo.6888985.

For experiments, the QuickDraw images were resized to 100 × 100 pixels, binarized and then displayed on a DMD (V650L Vialux GmbH). The DMD was illuminated by a white-light source (MNWHL4–4900 K, Thorlabs).

To train the ONN's weights, we needed to first measure the input images that were seen by the ONN device. This was necessary as the optically fanned-out images that formed on the LCD differed slightly from the digital image loaded onto the DMD, due to the imaging resolution limit and aberrations of the MLA. To measure these, we displayed each QuickDraw image on the DMD—leaving all of the LCD pixels at their highest transmission—and then inserted a pellicle beamsplitter (BP245B1, Thorlabs) after the LCD to reflect part of light to a monitoring camera (see Supplementary Fig. 2 for details). An image of the LCD panel was formed on the camera so that each fanned-out copy of the input image could be captured by the monitoring camera as the effective ground truth of the input images. These ground-truth images were used for training the weights of the ONN pre-processor on a computer (Supplementary Note 6) and for checking the accuracy of optical matrix–vector multipliers (Supplementary Figs. 5 and 6). Each ground-truth image of the fanned-out copies was resized to 40 × 40 = 1,600 pixels, corresponding to the 40 × 40 LCD pixels used as the weights for each image.

For the QuickDraw image classification task shown in Fig. 2a, the multilayer ONN encoder consisted of a matrix–vector multiplication, with a weight matrix size of 1,600 × 36, the 36 optical-to-optical nonlinear activations, and a final matrix–vector multiplication with a weight matrix size of 36 × 4 (Supplementary Fig. 11). The digital decoder consisted of a single matrix–vector multiplication with a weight matrix size of 4 × 10. The linear ONN pre-processor involved just a single optical matrix–vector multiplication with a weight matrix size of 1,600 × 4, followed by a 4 × 10 digital decoder. For direct imaging, the 40 × 40 ground-truth images were resized to 2 × 2 images by averaging the pixel values, and sent to a digital decoder comprising a 4 × 10 weight matrix. The linear digital neural network shown in Fig. 2d consists of a linear layer with a 1,600 × 4 weight matrix, followed by another linear layer with a 4 × 10 weight matrix. There is no nonlinear activation function between the two linear layers, and both have real-valued weights and bias terms. The nonlinear digital neural network shown in Fig. 2d has a linear layer with a 1,600 × 36 weight matrix, followed by element-wise nonlinear activations (sigmoid), followed by another linear layer with a 36 × 4 weight matrix, and finally a linear layer with a 4 × 10 weight matrix. There is no nonlinear activation between the 36 × 4 linear layer and 4 × 10 linear layer. All layers have real-valued weights and bias terms.

Optical-neural-network training

Training of the ONN layers was achieved primarily by creating an accurate model (digital twin) of the optical layers, and training the model's parameters in silico, including the digital post-processing layer(s). The digital model treated each optical fully connected layer as matrix–vector multiplication, and included the 36 individually calibrated nonlinear curves for the image intensifier activation functions. As our optical matrix–vector multiplier was engineered to perform matrix–vector multiplication, our digital models are composed of mathematical operations like those in regular digital neural networks, but do not require simulation of any physical process such as optical diffraction. To improve the robustness of the model and allow it to be accurately implemented experimentally despite the imperfection of this calibration, we made use of three key techniques: an accurate calibrated digital model as described above, data augmentation for modelling physical noise and errors, and a layer-by-layer fine-tuning with experimentally collected data.

We performed data augmentation on training data with random image misalignments and convolutions, which were intended to mimic realistic optical aberrations and misalignments. This included translations (±5% of the image size in each direction) and mismatched zoom factor (±4% image scale). To manage the computational cost of this augmentation, we found that it was sufficient to only apply these augmentations to the input layer. During each forward pass, we also added random noise to the input of each layer of the ONN that is equivalent to about 2% of the input values (more details in Supplementary Note 6).

We first trained models entirely digitally. We used a stochastic gradient optimizer (AdamW[54]) for training. The training parameters such as learning rate vary from task to task and are included in training code deposited in GitHub or Zenodo. Generally, each model was trained for multiple times with each training parameter randomly generated within a range. The parameters were fine-tuned from trial to trial by using the package Optuna[55], until the best training result was achieved (for example, the highest validation accuracy without obvious overfitting).

After this digital training step, we fine-tuned the trained models using a layer-by-layer training scheme that incorporated data collected from the experimental device. We first uploaded the weights for the first optical layer obtained by training the digital model, and collected the nonlinear activations for each training image after the image intensifier using the monitoring imaging systems (see Supplementary Note 3). Using the images after the image intensifier as the input, we then retrained the second optical layer. We then uploaded the obtained weights for the second optical layer, and for each image in the training set collected the output from this second layer experimentally, which was used to finally retrain the last digital linear layer. Only after this layer-by-layer fine-tuning did we perform experimental testing with the test dataset.

To ensure that all of the weights of the optical layers in the trained ONNs are non-negative, we clamped each element of the weight matrix, setting negative weights to zero after each parameter update during our training of the ONNs. One can think of this as applying a rectified linear unit to the weight matrix. This clamping slows down training and is prone to instabilities if large learning rates are used, due to vanishing gradients (once clamped, the gradient for that element is 0). We worked around this by training with smaller and decaying learning rates, but more epochs. We also applied techniques such as hyperparameter searches to improve training results (Supplementary Note 6), which was effective for all of the models used in the experiments we ran.

Flow-cytometry image classification

We performed an experimental benchmark of image-based cell-organelle classification using a procedure mostly similar to the QuickDraw benchmarks, including the experimental collection of input ground-truth images, and the training procedures. Images from ref. [42] (S-BSST644, available from https://www.ebi.ac.uk/biostudies/) were filtered into five classes based on the organelles (nucleolus, cytoplasm, centrosomes, cell mask, mitochondria) and the first 200 valid images per class were selected by hand for training (1,000 images in total), with a random train–validation split of 160:40, and the next 40 valid images per class were used for testing. Our selection criterion was to discard invalid images that involve multiple or no cells. Incidentally, images with multiple cells were added back later for the anomaly detection benchmark shown in Fig. 3. As with the QuickDraw images, these images were binarized and displayed on the DMD with a 100 × 100 resolution (in terms of DMD pixels), illuminated by the white-light source.

Real-scene image classification

For classification of objects in a real scene, we 3D-printed a small scene consisting of a road intersection centred around a traffic-sign holder, in which different speed-limit signs could be placed. We used a zoom lens (Zoom 7000, Navitar) to image the speed-limit sign onto the input of the ONN image processor (Supplementary Fig. 1). The demagnification of this lens was chosen so that the image of the sign relayed in front of the ONN encoder approximately spanned about 1 mm × 1 mm, which was the same physical size of the images displayed on the DMD. The scene was illuminated by two green LED lights (M530L4-C1, Thorlabs) from different angles for more uniform illumination.

To train the ONN weights for the real scene tasks, we collected ground-truth input images using a procedure similar to the classification tasks performed with the DMD input. These images were collected for each angle (0 to 88° in 1° increments), for each of eight classes (15, 20, 25, 30, 40, 55, 70 and 80 speed limits). Every fourth angle collected was used in the validation set, so the total dataset included 536 images for training and 176 images for validation.

All other aspects of the training and network design are similar to the previous tasks, with only two exceptions: (1) the digital backend consisted of two layers instead of one layer; (2) the compressed dimension was 2, rather than 4. As with other tasks, this compression ratio was selected as the highest compression ratio for which the nonlinear, multilayer ONN was still able to perform the task with a reasonable accuracy.

Additional image-sensing tasks based on different digital backends

Image reconstruction with autoencoders. We reconstructed QuickDraw images (as shown in Fig. 3b,c) with a digital decoder in the following way. Starting with the four-dimensional feature vectors produced by the ONN encoder previously trained for classification (Fig. 2), we trained a new digital decoder neural network that would produce an image whose structural similarity index was minimized relative to the ground-truth training dataset images. The decoder neural network was chosen to be a multilayer perceptron, with batch normalization layers before each sigmoid activation function. The number and widths of the hidden layers were found by random neural architecture search, which produced a best-performing network with three hidden layers, where the final output dimension corresponds to the reconstructed 28 × 28 image. We found that larger (that is, more powerful) decoders were unable to produce better reconstructions, suggesting that the four-dimensional bottleneck is the limit on reconstruction accuracy here. The reconstructed images shown in Fig. 3c are randomly chosen test images in the hurricane and chair classes. All reconstructed images in the test set are shown in Supplementary Figs. 15–24. See Supplementary Note 10 for more details.

Anomaly detection with unsupervised learning. We performed the anomaly detection shown in Fig. 3d,e as follows. First, we created a dataset consisting of 418 anomaly images by including images containing at least two cells from all of the five original classes. These were previously excluded from training dataset for the cell-organelle classifier shown in Fig. 2e–h. Next, we displayed all images in this new, anomaly dataset on the DMD and, with the ONN encoder's weights kept identical to those originally obtained for cell-organelle classification, collected the four-dimensional feature vector for each image. Principal component analysis on these feature vectors (Supplementary Fig. 26) shows that the anomalous images are distinct from the previously trained classes, occupying a part of the latent space that was previously not accessed by any of the trained classes. As a result, we were able to successfully perform spectral clustering on these feature vectors. This procedure involves computing the nearest-neighbour distances of the vectors to compute an affinity matrix, whose eigenvectors correspond to localized clusters. The largest five clusters (largest eigenvalues) of this matrix were found to correspond to each of the previously trained classes, while the sixth cluster was found to correspond to anomalous images. After assigning the most probable class label to each cluster for the maximum overall likelihood (Supplementary Fig. 26), we computed the confusion matrix of classifying the original 5 classes plus the new anomaly class by comparing to the ground-truth labels (Fig. 3e). The true positive rate was calculated as the percentage of anomalous images classified as anomalous images, and the false positive rate was calculated as the percentage of normal images in the total number of images classified as anomalies.

Nonlinear parameter fitting. We performed the estimation of speed-sign viewing angle as follows. Using the two-dimensional feature vectors produced by the ONN encoder trained for speed-limit classification in Fig. 3, we trained a new digital decoder neural network to predict sign viewing angle. The dataset split between training and validation here was that every even angle was used in the train set and every odd angle was used in the validation set, except we only considered one class (that is, one speed limit) at a time (in other words, the sign angle estimation decoder only works for a given speed-limit sign, rather than for an arbitrary sign). A multilayer perceptron with dimensions 2 → 50 → 100 → 1 was found to perform well when trained with an L1 loss function, that is, $|\theta_{\text{predicted}} - \theta_{\text{true}}|$. The angle prediction can be performed for all, rather than just one, speed-limit class at a time, albeit with reduced performance. The results are shown in Supplementary Fig. 27.

Simulation of deeper optical neural networks for ten-class cell-organelle classification

To explore the possible performance and applications of future, scaled-up nonlinear ONN encoders, we performed realistic physical simulations of optical neural networks based on our experiments, for a more challenging task: ten-class cell-organelle classification for image-based cytometry.

The dataset used for this task was adapted from ref. [42] (S-BSST644; available from https://www.ebi.ac.uk/biostudies/) in the following way. First, we selected ten of the twelve provided classes (the other two, Golgi and Control, had too few images and no fluorescent channel respectively). Unlike our dataset preparation for the five-class version of this task performed experimentally, here we retained all of the images, including those with multiple or no cells.

The five networks considered are as follows. The first ONN pre-processor is a wide (100 × 100 = 10,000-dimensional input vector), linear single-layer ONN (Linear). The second is a two-layer multilayer perceptron with 10,000-dimensional input, and a 200-dimensional hidden layer. Besides both the input and hidden dimension being much larger, this network is similar to the two-layer fully connected ONN we realized experimentally. The third and fourth models extend this network deeper, adding one, for CNN1, or three, for CNN3, optical convolutional layers. Multichannel optical convolutional layers of this kind have been realized before with 4*f* systems (for example, refs. [4,14]), which are in many regards simpler and more amenable to compact implementation than fully connected optical layers. These CNNs also include a shifted ReLU activation (that is, trained batch normalization layers followed by ReLU), which could be realized with a slight modification of the image intensifier electronics, or by the threshold-linear behaviour of optically controlled VCSEL[56] or LED arrays. We have primarily assumed pooling operations are AvgPool, which are straightforwardly implemented with optical summation. The MaxPool operation used once in CNN3 is more challenging but could plausibly be realized effectively by using a broad-area semiconductor laser or placing a master limit on the energy available to a VCSEL or LED array, such that the first unit to rise above threshold would suppress activity in others. These ONN designs are ultimately speculative; In general, we anticipate that practically realizing more powerful ONN encoders will require jointly designing compact, low-cost ONN hardware components and developing optics-friendly DNN architectures, rather than simply directly adapting existing digital DNN architectures.

The decoders used for all networks are identical linear layers with dimensions $N \to 10$, where N is the bottleneck dimension. Note that the compression ratio is taken to be 100 × 100 = 10,000, the original image resolution, divided by N.

All-optical networks were simulated with emulated physical noise on the forward pass (as described in detail in Supplementary Note 14), and weights were constrained to be non-negative (because we assumed incoherent light). We note that it is possible that intermediate layers could be realized with coherent light (and therefore with real-valued weights) even if the input light is strictly incoherent, that is, by using arrays of VCSELs[56]. We find that, while non-negative weights can generally be trained for various tasks, the performance of these networks is generally inferior to what is possible with real (that is, both positive and negative) weights. Consequently, our results here are roughly a lower bound with respect to the performance of coherent-light-based ONN encoders.

As a reference for achievable classification accuracy on this cell-organelle classification task (that is, a practical upper bound, in part due to the presence of anomalous multi- or no-cell images), we trained a purely digital classifier based on a ResNet-18 (ref. [46]) which was pretrained on ImageNet. This network includes four additional layers to adapt input images to the ResNet core, and to produce the final classification output. All weights of this network were fine-tuned by training with the training set.

Data availability

The demonstration data for data gathering, as well as training data for the all-optical/digital neural networks, are available at https://github.com/mcmahon-lab/Image-sensing-with-multilayer-nonlinear-optical-neural-networks.

Code availability

All of the data generated, and code used, in this work are available at https://doi.org/10.5281/zenodo.6888985.

References

54. Loshchilov, I. & Hutter, F. Decoupled weight decay regularization. In *7th International Conference on Learning Representations* (ICLR, 2019).
55. Akiba, T., Sano, S., Yanase, T., Ohta, T. & Koyama, M. Optuna: a next-generation hyperparameter optimization framework. In *Proc. 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 2623–2631 (ACM, 2019).
56. Heuser, T. et al. Developing a photonic hardware platform for brain-inspired computing based on 5×5 VCSEL arrays. *J. Phys. Photon.* **2**, 44002 (2020).

Acknowledgements

We wish to thank NTT Research for their financial and technical support (to T.W., L.G.W., S.-Y.M., T.O. and P.L.M.). Portions of this work were supported by the National Science Foundation (award no. CCF-1918549 to T.W., M.M. Stein and P.L.M.), a Kavli Institute at Cornell instrumentation grant (to T.W. and P.L.M.), and a David and Lucile Packard Foundation Fellowship (to P.L.M.). P.L.M. acknowledges membership of the CIFAR Quantum Information Science Program as an Azrieli Global Scholar. T.W. acknowledges the partial support from Schmidt Futures via an Eric and Wendy Schmidt AI in Science Postdoctoral Fellowship to Cornell University. We acknowledge helpful discussions with A. Senanian, B. Malia, F. Presutti, V. Kremenetski, S. Prabhu, A. Barth, R. Oliver, and D. Schraivogel. We also acknowledge S. Sohoni for help with figure design.

Author contributions

T.W., L.G.W., M.M. Sohoni and P.L.M. conceived the project and designed the experiments. M.M. Sohoni and T.W. built and performed the experiments on the nonlinear and linear ONN encoders, and analysed the data. T.W. performed the extended cell-organelle simulations. M.M. Stein performed the neural architecture search for QuickDraw reconstruction. S-Y.M. and T.O. aided in simulations of deep optical encoders. M.G.A. assisted with 3D-scene modelling. L.G.W., T.W., M.M. Sohoni and P.L.M. wrote the manuscript. P.L.M. and L.G.W. supervised the project.

Competing interests

T.W., M.M. Sohoni, L.G.W. and P.L.M. are listed as inventors on a US provisional patent application (serial no. 63/392,042) on nonlinear optical neural network pre-processors for imaging and image sensing. The other authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41566-023-01170-8.

Correspondence and requests for materials should be addressed to Tianyu Wang, Mandar M. Sohoni, Logan G. Wright or Peter L. McMahon.

Peer review information *Nature Photonics* thanks Jacques Carolan and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Reprints and permissions information is available at www.nature.com/reprints.