Title

Femtojoule optical nonlinearity for deep learning with incoherent illumination.

Journal
Science Advances, 11(5)

Authors
Feng, Qixin
Uzundal, Can
Guo, Ruihan
et al.

Publication Date
2025-01-31

DOI
10.1126/sciadv.ads4224

APPLIED SCIENCES AND ENGINEERING

Femtojoule optical nonlinearity for deep learning with incoherent illumination

Qixin Feng[1,2]†, Can B. Uzundal[2,3]†, Ruihan Guo[2,4], Collin Sanborn[2,5], Ruishi Qi[1,2], Jingxu Xie[1,6], Jianing Zhang[1,6], Junqiao Wu[2,4], Feng Wang[1,2,7]*

Optical neural networks (ONNs) are a promising computational alternative for deep learning due to their inherent massive parallelism for linear operations. However, the development of energy-efficient and highly parallel optical nonlinearities, a critical component in ONNs, remains an outstanding challenge. Here, we introduce a nonlinear optical microdevice array (NOMA) compatible with incoherent illumination by integrating the liquid crystal cell with silicon photodiodes at the single-pixel level. We fabricate NOMA with more than half a million pixels, each functioning as an optical analog of the rectified linear unit at ultralow switching energy down to 100 femtojoules per pixel. With NOMA, we demonstrate an optical multilayer neural network. Our work holds promise for large-scale and low-power deep ONNs, computer vision, and real-time optical image processing.

INTRODUCTION

As all-purpose digital computation, particularly for artificial intelligence and deep neural networks, reaches an energy bottleneck, alternative physics–based computational architectures are attracting increasing attention (*1*–*13*). Among these, optical neural networks (ONNs) are a promising alternative due to their high parallelism, energy efficiency, and minimal latency (*14*–*17*). Highly parallel linear operations, such as matrix multiplications (*18*–*25*) and convolutions (*26*–*28*), can readily be implemented using nearly dissipationless linear optical transformations in ONNs. On this front, ONNs offer substantial energy savings per linear operation compared to cutting-edge all-digital counterparts (*24*, *29*–*32*). However, achieving an efficient optical nonlinearity poses inherent challenges (*33*, *34*), leading ONNs to often rely on hybrid systems that incorporate electronic nonlinear activations. These hybrid ONNs require preamplifiers and analog-to-digital converters to process weak optical signals that increase latency and power spent per operation (*15*). To realize deep ONNs with low energy consumption, the development of a sub-picojoule optical nonlinearity is crucial.

Recently, a “receiverless” approach has been proposed for energy-efficient optical modulation, which obviates the need for power-hungry electronics by in situ integration of a photodiode (PD) with an electro-optical modulator (EOM) (*32*, *35*). In this configuration, a portion of the input light generates photocarriers, which directly charge (or discharge) the EOM, thereby modulating the remaining part as the output. This process facilitates light self-modulation, with energy consumption that scales with the capacitance of the PD and the EOM. Following this approach, optical nonlinear operation with switching energy on the order of femtojoule per activation has been demonstrated in integrated photonics circuits by integrating femto-farad capacitance PDs and EOMs such as InGaAsP photonic crystals or micro-ring resonators. (*23*, *36*). However, integrated photonic devices face scalability challenges and lack compatibility with incoherent light, strongly restricting their use in large-scale ONNs in ambient light scenarios.

A free-space counterpart of the receiverless optical nonlinearity has the potential to address the scalability concerns by harnessing the immense parallel computing capabilities afforded by free-space light propagation. Further, free-space ONNs have compelling applications in object detection and sensing where conventional neural networks are routinely used to run inference on digitized camera images. In such applications, free-space ONNs could remove the need for the digitization step and run inference directly on the ambient light (*25*). Previously, liquid crystal (LC) light valves (LCLVs) have been developed for controlling a read beam with a write beam by placing a photosensitive film next to an LC EOM layer, with a dielectric mirror separating the two (*37*–*39*). With LCLVs, the sigmoid-like nonlinear dependence of the read beam intensity on the write beam intensity has been demonstrated and applied to the early research on ONNs (*40*, *41*). More recently, self-modulation of light has been realized by resistively coupling the LC layer to two-dimensional (2D) material phototransistor arrays, but the energy consumption is well above picojoule per operation (*42*). To the best of our knowledge, a femtojoule–rectified linear unit (ReLU) for self-activation of the input patterns—the predominant nonlinear function in contemporary deep neural networks—has never been realized.

In this study, we present an energy-efficient and highly parallel nonlinear optical microdevice array (NOMA) for free-space optical computation. Each pixel of the device contains a silicon (Si) PD capacitively coupled to an LC cell, allowing for nonlinear activations at the femtojoule scale. By leveraging the mature fabrication processes for Si-based integrated circuits and LC display technologies, our design readily enables the fabrication of devices with millions of pixels. Through the characterization of NOMA, we present an optical ReLU nonlinearity operating on an incoherent optical beam. Further, we demonstrate the practical applications of this optical ReLU in two optical processing tasks: real-time image contrast enhancement and nonlinear activation within a multilayer ONN (ML-ONN).

[1]Department of Physics, University of California, Berkeley, Berkeley, CA 94720, USA. [2]Materials Sciences Division, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA. [3]Department of Chemistry, University of California, Berkeley, Berkeley, CA 94720, USA. [4]Department of Materials Science and Engineering, University of California, Berkeley, Berkeley, CA 94720, USA. [5]Graduate Group in Applied Science and Technology, University of California, Berkeley, Berkeley, CA 94720, USA. [6]School of Physical Science and Technology, ShanghaiTech University, Pudong District, Shanghai 201210, China. [7]Kavli Energy NanoScience Institute, University of California, Berkeley, Berkeley, CA 94720, USA.
*Corresponding author. Email: fengwang76@berkeley.edu
†These authors contributed equally to this work.

RESULTS

Device structure and working principle

We fabricated an NOMA comprising 750 × 700 pixels, each with a dimension of 20 μm by 20 μm (Fig. 1A). The Materials and Methods of the supplementary text describe the detailed structure and the fabrication process for the entire device. Figure 1B schematically illustrates the structure of the NOMA. Within each pixel, there is a Si PD connected to an Al mirror while the LC fills the gap formed between the Si substrate and the indium tin oxide (ITO)–coated glass. The Si substrate is grounded, while a source voltage V_s is applied to the ITO electrode. An additional n-doped region beneath the Al mirror serves as a global electrode for a control voltage V_c. Both V_s and V_c are nonnegative to ensure a reverse-biased PD.

For optimal optical modulation, we use a vertically aligned nematic LC with a large contrast ratio (*43*, *44*). We characterize the optical intensity modulation of the LC cell using cross-polarized (CP) reflectance as a function of V_{LC} (Fig. 1C). In this configuration, the LC cell acts as a tunable half-wave plate placed between two crossed polarizers.

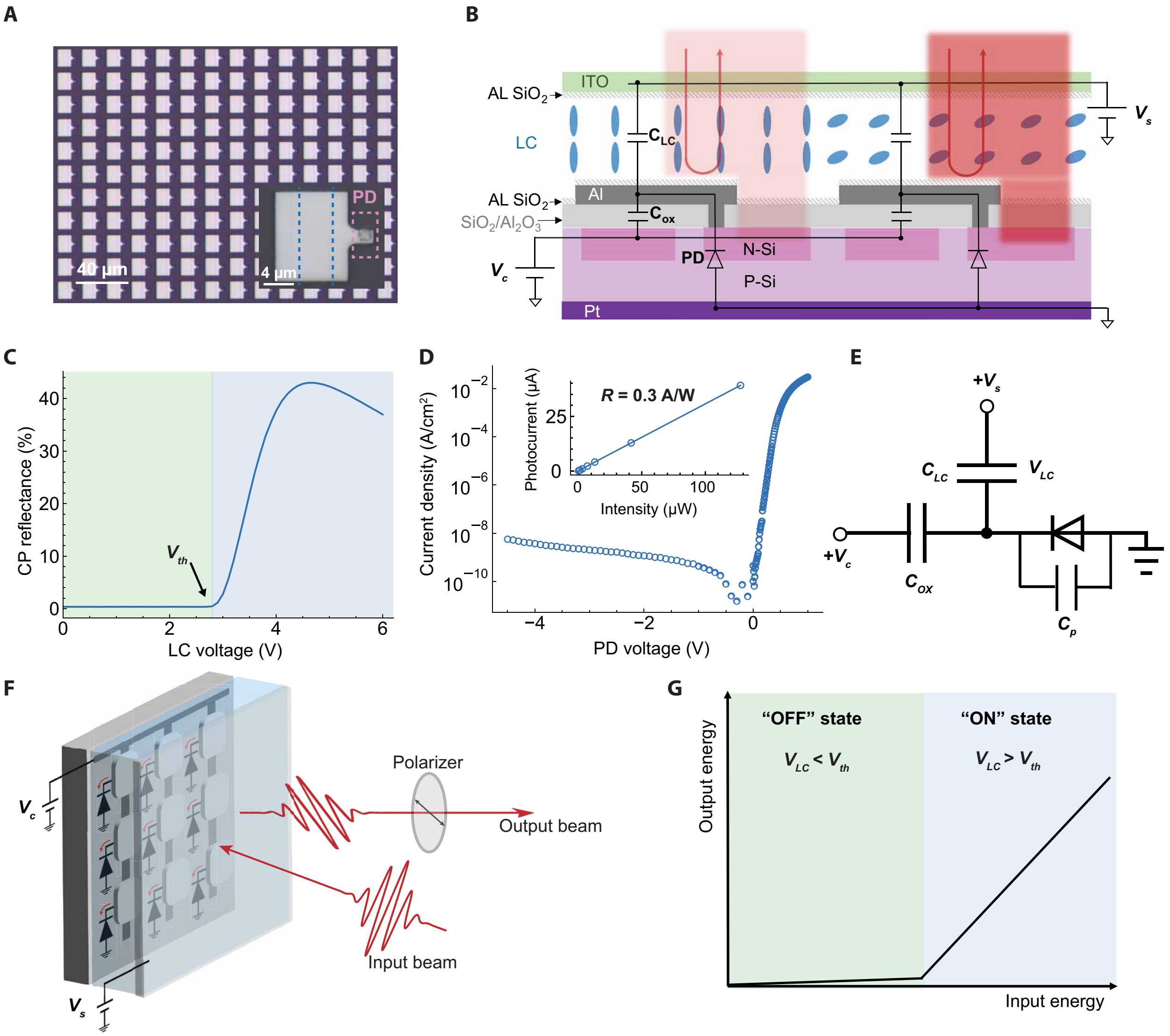

Fig. 1. Device structure and working principle. (**A**) Microscope image of NOMA. The inset shows a single pixel, depicting the metallic region that is the Al mirror. The n-doped region outlining the PD is indicated by pink dashed lines, and the additional n-doped region underneath the Al mirror is outlined by the blue dashed lines. (**B**) Cross section of NOMA pixels. The device consists of an LC layer, represented by blue rods between an ITO-covered glass and a Si substrate with PDs. Both substrates are coated with a SiO$_2$ for LC alignment (AL). Upon illumination, PD charges the LC cell, leading to rotation of the LC molecules and modulation of reflected light through LC birefringence. (**C**) Characterization of the LC birefringence using voltage-dependent CP reflectance of the LC cell with $\lambda = 680$ nm, showing a threshold voltage (V_{th}) at 2.8 V. (**D**) Current-voltage relation of the Si PD under dark conditions. The measured PD is the one formed between the auxiliary n-doped region and substrate. The insert shows the relation between photocurrent and light intensity, indicating a responsivity (*R*) of 0.3 A/W. (**E**) Equivalent circuit for a NOMA pixel. C_{LC} denotes the capacitance between the Al mirror and the ITO electrode. C_{ox} denotes the capacitance between the Al mirror and the n-doped region beneath it. C_p represents the junction capacitance. (**F**) Configuration of NOMA for light self-modulation. The input beam is a linearly polarized pulse, and the output beam is the CP reflected pulse. (**G**) ReLU-like input-output relationship. For incident light with small pulse energy, the NOMA remains in the OFF state and the transmitted light energy through the polarizer (output energy) is suppressed. For incident light with large pulse energy, the NOMA is switched to the ON state, characterized by a high CP reflectance.

At low bias, the LC cell appears dark as LC molecules are aligned parallel to the propagation direction of the incident beam and no change in the polarization state of the incident beam occurs. At a threshold bias (V_{th}) of 2.8 V, the LC molecules start to tilt, causing the incident beam to attain ellipticity and the LC cell appears brighter. The measured value closely approximates the theoretical value of 2.2 V predicted by the Freedericksz transition theory (*45*, *46*), considering the mismatch in electrode work functions (0.4 V, between Al and ITO). As the bias is increased above V_{th}, a clear maximum in reflectance occurs where the polarization of the incident beam is completely rotated to the perpendicular polarization. The LC cell exhibits a contrast ratio of 120, providing a broad optical modulation range.

As the Si PD provides in situ optical-to-electrical feedback, low dark current and high optical responsivity are crucial for energy-efficient nonlinear operations. By using an Al_2O_3 passivation layer (*47*), we achieve a low-junction dark current of 10 nA/cm^2 (Fig. 1D). Given that the PD area in each pixel is approximately 100 μm^2, the dark current per pixel is around 10 fA. We measure the responsivity of our PD as 0.3 A/W ($\lambda = 680$ nm), indicating efficient collection of photocarriers.

To elucidate the dynamics of the optical ReLU nonlinearity, we introduce a simple circuit model depicted in Fig. 1E, which consists of three main elements: (i) an LC capacitor (C_{LC}), which forms between the Al mirror and the ITO electrode; (ii) a Si PD, which can be described by an ideal diode in parallel with a junction capacitor C_p; (iii) an oxide dielectric capacitor (C_{ox}), which forms between the Al mirror and the additional n-doped region. We list the estimated capacitance values of each in table S1. Upon illumination, part of the incident light is reflected by the Al mirror, while the rest is absorbed by the Si PD. The photocarriers generated in the Si PD accumulate on the Al mirror, altering the voltage across the LC cell (V_{LC}). Consequently, the orientation of the LC molecules changes, which then modulates the reflected light through the LC's birefringence (Fig. 1F). Under dark or weak illumination, the V_{LC} remains below V_{th}, so the NOMA remains in the "OFF" state with a low CP reflectance. Thus, the energy of the CP-reflected light (output energy) is suppressed. On the other hand, for incident light with a high pulse energy, the LC capacitor is charged above V_{th}, so the NOMA is switched to the "ON" state, characterized by a high CP reflectance. In this state, the output energy exhibits a linear dependence on the input energy. This characteristic behavior mimics the ReLU function, where the output remains zero for low input values and increases linearly with higher inputs (Fig. 1G).

Optical switching dynamics

We periodically operate NOMA between an active and erase phase, as illustrated in Fig. 2A. During the active phase, the LC capacitor first charges to an initial voltage V_i, which is determined by the capacitance divider: $V_i = \frac{C_p + C_{ox}}{C_{tot}} V_s - \frac{C_{ox}}{C_{tot}} V_c$, where $C_{tot} = C_{LC} + C_{ox} + C_p$.

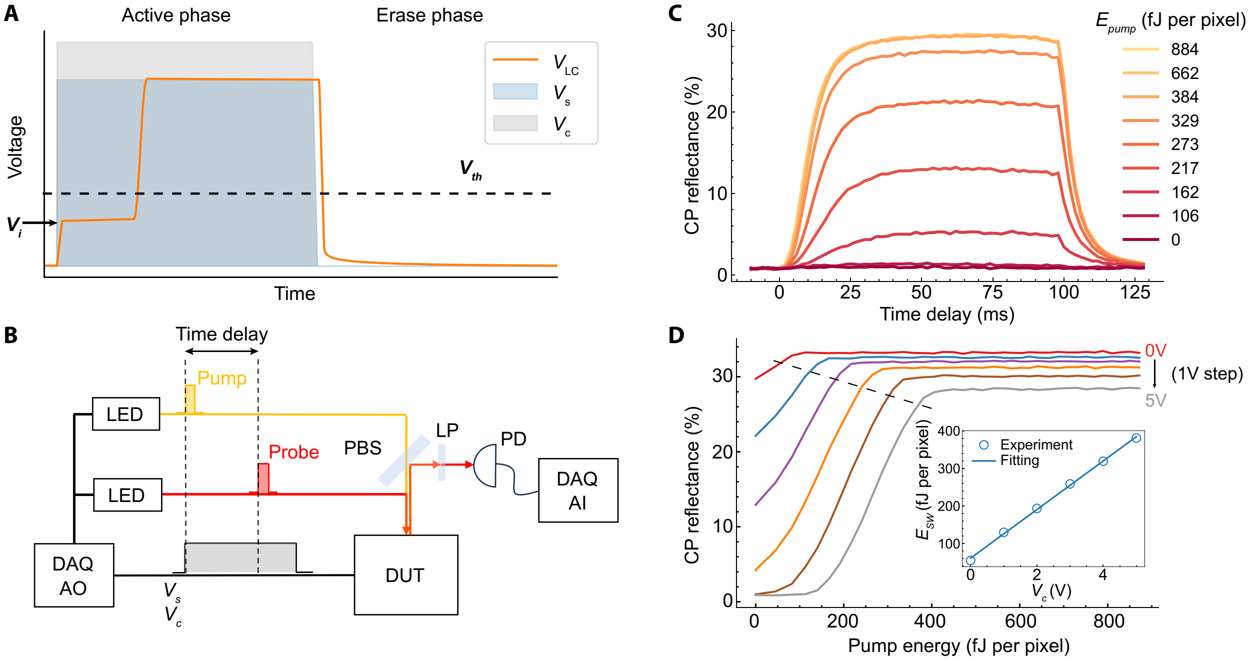

Fig. 2. Optical switching dynamics. (**A**) Waveform of V_s, V_c, and the resulting LC capacitor voltage V_{LC}. Both V_s and V_c are synchronized square waves, characterized by an active and erase phase. In the active phase, an intense optical pulse can charge the V_{LC} from the below-threshold initial value V_i to above-threshold value V_s, resulting in NOMA switching from the OFF to the ON state. In the erase phase, the LC capacitor is discharged, setting it to the default OFF state. (**B**) Illustration of the pump-probe experiment (see fig. S3A for an optical layout), where pump ($\lambda = 630$ nm) and probe pulses ($\lambda = 680$ *nm*) from LEDs are directed onto NOMA. Using analog outputs (AO) of a data acquisition card (DAQ), the light pulses are synchronized with the applied bias such that the pump arrives at the rising edge of the bias, while the probe pulse is time-delayed. Dynamics in the CP reflectance of NOMA is measured using a polarized beam splitter (PBS) with a long-pass filter (LP) in a wide-field microscope geometry. The total intensity of the reflected light is measured using a PD and digitized using DAQ analog inputs (AI). (**C**) The dynamics of CP reflectance of NOMA at increasing pump energies (lighter shades) E_{pump} ($V_s = 4$ V and $V_c = 4.5$ V) largely follow the outline of the voltage pulse with a rise and fall time limited by the LC. (**D**) E_{pump}-dependent CP reflectance at different V_c (V_s fixed at 4 V). The insert illustrates the expected linear relationship between switching energy E_{sw} and V_c, showing a minimum E_{sw} at 60 fJ/pixel.

This charging process happens on a timescale of $\tau_1 = R_s C_{tot}$, where R_s is the series resistance of bulk silicon and the contact. Given that R_s is approximately $\sim 10^6$ ohm and C_{tot} is around 20 fF, τ_1 occurs on the timescale of microseconds. Under dark condition, the LC capacitor slowly charges to V_s where the charging time (τ_2) is set by the dark current I_R of the Si PD. Given that I_R is in the order of 10 fA / pixel, τ_2 is in the order of seconds. In our experiments, we apply a square wave of V_s and V_c with a period T of milliseconds, which is determined by the LC's response time. Considering that $\tau_1 \ll T / 2 \ll \tau_2$, the LC voltage remains relatively constant at V_i under dark condition during the $T / 2$ ms active phase. Under light illumination, the photocarriers generated in the PD lead to a rapid charging of the LC capacitor to V_s. We maintain V_s above V_{th} and adjust V_c so that V_i remains below V_{th}, ensuring that the LC is in the OFF state without light but can transition to the ON state under sufficient light illumination. In the erase phase, we keep V_s below V_{th} and set V_c to 0 V, dissipating the accumulated charges in the LC capacitor and reverting the device to the default OFF state.

The optical switching energy E_{sw} is determined by the amount of photocarriers (Q_{ph}) needed to fully charge the LC and oxide capacitors, which can be estimated from our circuit model as

$$E_{sw} = \frac{Q_{ph}}{\alpha} = \frac{C_{LC}\Delta V_s + C_{ox}\Delta V_c}{\alpha} \tag{1}$$

where α is the optical-to-electrical coefficient and ΔV_s and ΔV_c, respectively, denote the changes of V_s and V_c between the active and erase phases. Considering the responsivity of silicon PD (0.3 A/W), fill factor of Al electrode, and transmission loss through the ITO layer, α is approximately 0.1 C/J. Given that both C_{ox} and C_{LC} are in the range of femtofarad, Q_{ph} is estimated to be in the tens of femtocoulombs. Thus, the optical switching energy for each pixel is calculated to be in the hundreds of femtojoule range. Regarding the electronic energy consumed during the switching process, it is essentially the work done by the voltage sources, which can be calculated as $E_{electronic} = C_{LC}\Delta V_s^2 + C_{ox}\Delta V_c^2$. It indicates that the electric switching energy is on the same scale of the optical switching energy.

To investigate the dynamics of the optical switching process, we carried out pump-probe experiments capable of optically probing the device dynamics at the characteristic time scale of the LC molecules (Fig. 2B). We use short pulses (2-ms duration) from colored light-emitting diodes (LEDs) for both pump and probe lights. The pump light is synchronized with the electrical signal's rising edge, which charges the LC cell and initiates the optical switching. The probe light measures the CP reflectance of the device as a function of time delay between the pump and probe. To ensure that the observed dynamics are only from the pump-induced changes, we use a probe pulse energy (80 fJ per pixel) below the threshold energy of NOMA and use probe pulse durations (2 ms) much shorter than the LC response time (~10 ms). We map the CP reflectance as a function of V_s, V_c, and pump energy (E_{pump}). Using these results, we quantify the capacitances in our circuit model and identify the optimal conditions for efficient optical nonlinearity.

Figure 2C shows time traces of CP reflectance at increasing pump energies at $V_s = 4$V and $V_c = 4.5$V. For subthreshold pump energies (i.e., 0 and 106 fJ per pixel), CP reflectance traces show a negligible increase after pumping, indicating that the device remains in the OFF state. In contrast, at a higher pump energy, the CP reflectance increases notably after pumping and reaches a plateau. With increasing pump energy, the plateau value rises from 0.8% to a maximum of 29%, showing a large modulation range of 35 between the OFF and ON states.

We further characterize the optical switching by measuring the CP reflectance at a fixed probe time delay at 60 ms, at which point the CP reflectance reaches its plateau. Figure 2D displays the measured CP reflectance as a function of E_{pump} at different V_c. The derived switching energy E_{sw} (defined as the pump energy at which the CP reflectance reaches 95% of its saturation level) shows a linear dependence on V_c, consistent with our circuit model (Eq. 1). From the slope and intersection of the linear fit, we derive $C_{ox} = 6.5$ fF per pixel and $C_{LC} = 5.8$ fF per pixel, which are close to their estimated values (table S1).

Femtojoule optical ReLU for image contrast enhancement

We demonstrate the optical ReLU function using a single but longer LED light pulse (50 ms) synchronized with the rising edge of the electrical pulses. The voltages are set at $V_s = 4$V and $V_c = 5.5$V. The distinct signature of the optical nonlinearity is depicted in Fig. 3A, where the relationship between CP reflected light energy (output energy) and incident light energy (input energy) resembles an ReLU function with a switching energy of 280 fJ per pixel. At low input energies, the reflectivity of NOMA is marginal at around ~1%. At higher input energies, the reflectivity increases to 24% and saturates. The saturated reflectivity is limited by the fill factor of the Al electrode and the optical losses at the interfaces of the device stack. Analogous to the response to short pulses, we can manipulate the switching energy value of the ReLU function by adjusting V_c (fig. S4). This additional tunability is useful in applications where the optical input varies dynamically, enabling the device to maintain optimal performance across various scenarios.

To further investigate the response at the individual pixel level, we capture wide-field images of NOMA and track the pixel-by-pixel dynamics. We segment the wide-field images into a regular grid, each grid containing only one pixel of NOMA (fig. S5A). Figure 3B shows the statistics from 10,201 NOMA pixels. The ReLU function has a normal distribution of the switching energy with a mean value of 280 fJ and an SD of 18 fJ. In applications such as ONNs, some degree of heterogeneity in the response of individual pixels is acceptable and even leveraged through hardware-aware fine-tuning of the models (*48*, *49*). Yet, achieving uniform and consistent response across many pixels is generally desirable as cascading errors can potentially impede computation, especially for deep neural networks.

To demonstrate the ReLU functionality and large-scale uniformity, we performed a contrast enhancement task of a binary grayscale image through interaction with more than 15,000 NOMA pixels. As our baseline, we capture an image reflected from NOMA under the linear response (Fig. 3C). We ensure the linear response of the device by setting $V_s = -4$V to forward bias the Si PD, which maintains the LC cell in the ON state. To capture the contrast-enhanced image (Fig. 3D), we set $V_s = 4$ V and $V_c = 5.5$ V to ensure that NOMA is under an ReLU response. Compared with the dark regions of the reference image, the contrast enhanced image shows a darker background while the bright regions retain their average brightness. Quantitatively, the image processed with nonlinearity exhibits a contrast four times greater than that of the reference image (Fig. 3, E and F). The ability of NOMA to selectively amplify the contrast of specific image regions showcases its potential in applications such as real-time image processing and optical edge computing.

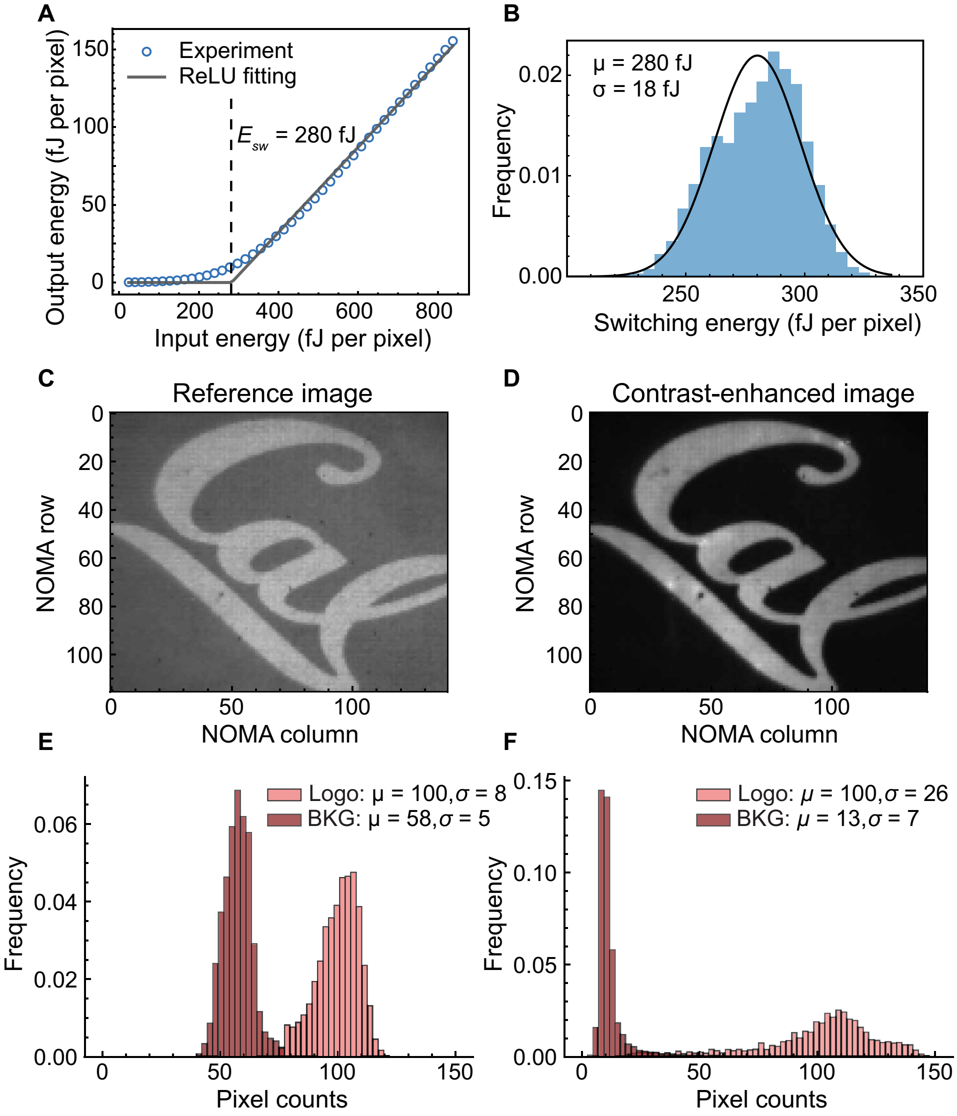

Fig. 3. Demonstration of the optical ReLU and image contrast enhancement. (**A**) Average output energy versus input energy across more than 10,000 NOMA pixels, showing a nonlinear dependence, which is fitted with an ReLU function (gray line) with a switching energy E_{sw} of 280 fJ. The histogram of E_{sw} (**B**) shows a normal distribution with a mean (μ) of 280 fJ and an SD (σ) of 18 fJ. The reference (**C**) and contrast-enhanced (**D**) image reflected from the NOMA, which is operated under a linear response and the ReLU response, respectively. The input image is a "*Cal*" logo, which consists of two regions: the logo region, chosen to be bright, and the background, chosen to be dark. Investing the two regions separately in terms of their pixel values reveals histograms with a mean (μ) and SD (σ) value for the reference image (**E**) and contrast-enhanced image (**F**), demonstrating four-time improvement in the image contrast.

In these cases, the nonlinear layer prunes or maintains connections between successive layers in the network. In spirit, this contrast enhancement task can be thought of as one such application where we already demonstrate more than 10,000 nonlinear connections with potential to expand into deep neural networks with more than one hidden layer.

ML-ONN with ReLU activations

We highlight the role of the ReLU nonlinearity by demonstrating an ML-ONN. The implemented ML-ONN consists of two fully connected linear layers linked by the NOMA, serving as the nonlinear activation layer. We leverage the ML-ONN to tackle two distinct binary classification tasks characterized by nonlinear decision boundaries. Figure 4A illustrates the ground truth for one such boundary, separating the 2D space defined by the input vector $\boldsymbol{x} = (x_1, x_2)$ into red and blue regions by a circular decision boundary. Our network is configured with two inputs, four hidden neurons, and two output neurons with the goal of learning these nonlinear decision boundaries. The ML-ONN maps the input vector $\boldsymbol{x}$ to an output vector $\boldsymbol{y} = (y_1, y_2)$ through two transformation matrices and one ReLU nonlinear activation. We determine the class of the input point (red or blue) by comparing the magnitudes of y_1 and y_2 or, more precisely, calculating the posterior probabilities with the SoftMax function.

In our optical implementation, we encode $\boldsymbol{x}$ into light intensity and the weight matrices [$W^{(1)}$ and $W^{(2)}$] into the reflectivity of

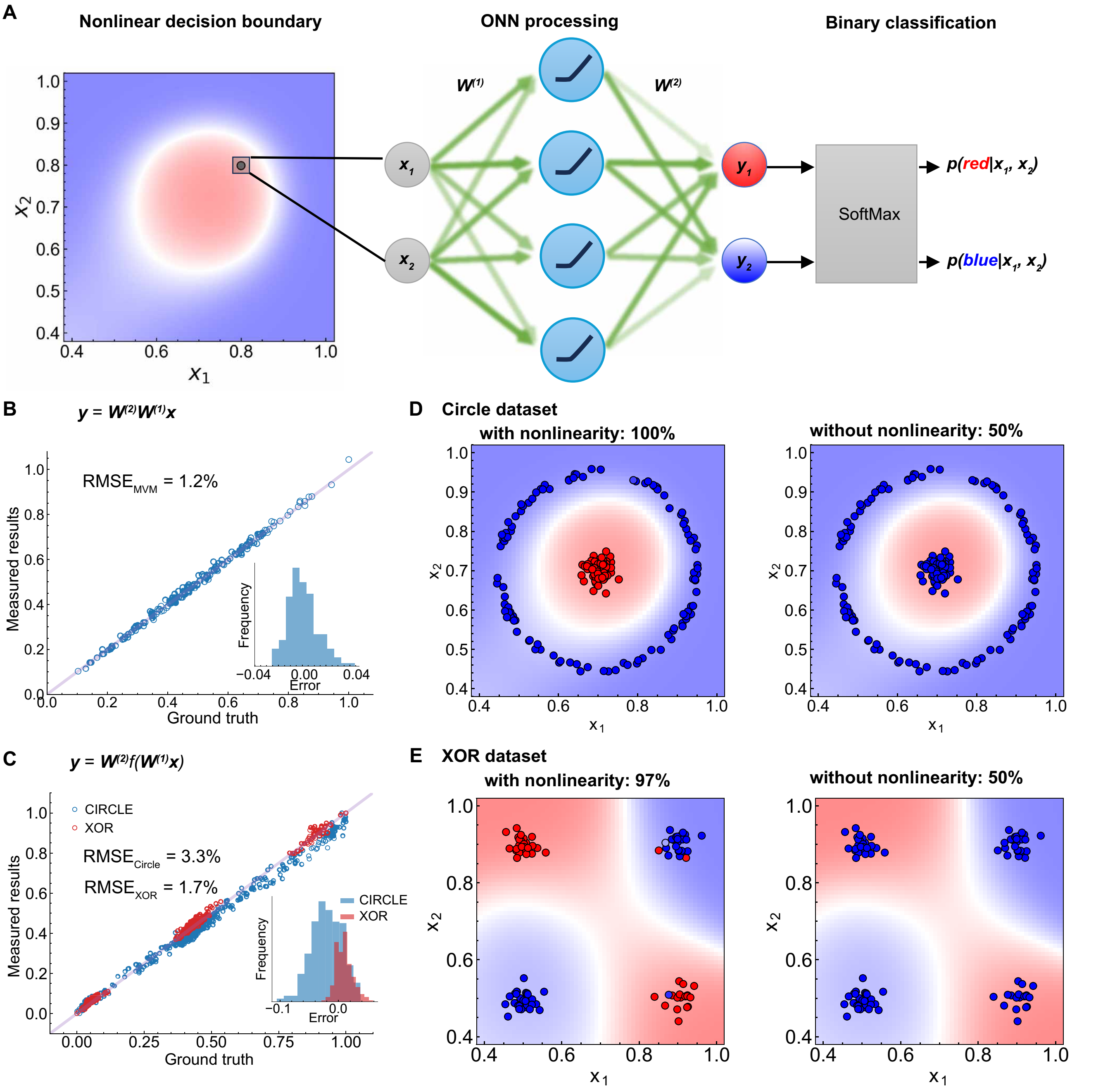

Fig. 4. ML-ONN for binary classification. (**A**) Schematic illustration of the ML-ONN used for the binary classification task. The input to the ML-ONN consists of coordinates (x_1, x_2) of a point in 2D space, separated into two classes by a nonlinear boundary. The implemented ML-ONN is a two-layer fully connected neural network with optical nonlinear activation using the NOMA with ReLU response shown in Fig. 3A. (**B**) Performance of the linear operation of the ML-ONN shown as a scatter plot of measured outputs against ground truth of 80 random MVM. In an ideal implementation, the measured output of the MVM and the ground truth fall on a line with slope of 1, marked by the purple line in the figure. The inset panel is an error histogram illustrating the scatter around this ideality line. The histogram is characterized by a root mean squared error (RMSE) of 1.2%. (**C**) Scatter plots characterizing the RMSE of the full ML-ONN, including the nonlinear activation between MVMs, along with the error histogram (inset). RMSE rates for two classification tasks on two datasets (Circle and XOR) are characterized as 3.3% and 1.7% respectively. The distribution and the underlying decision boundaries of the Circle dataset and the XOR dataset are shown in (**D**) and (**E**), respectively. In both cases, the classification accuracy is >97% when the NOMA is operated under an ReLU response, while the accuracy is 50% (random chance) when it is operated under linear response conditions.

spatial light modulators (SLMs). We use an optical fan out to implement matrix vector multiplications (MVMs) (*25*). The Supplementary Materials contain a detailed description of the optical layout (fig. S7) as well as the test and validation of the implemented optical hardware in supplementary section S3.

We first quantify the operation precision of the linear and nonlinear layers in the ML-ONN. We evaluate the linear operation capabilities of our ML-ONN by performing random MVMs on the SLMs. Figure 4B shows the measured light intensities after the multiplication operation versus the theoretical values. On the basis of this comparison, we determine a root mean square uncertainty of 1.2% in our optical MVM implementation. This error rate indicates an effective calculation precision of 6.2 bits for our linear operations.

We next performed an optical inference experiment with the complete ML-ONN depicted in Fig. 4A across two binary classification datasets, namely, Circle and XOR. Figure 4C shows the measured light intensity versus the theoretical values. The comparison shows that the root mean square uncertainty increases to 1.7 ~ 3.3% when we include a nonlinear activation through NOMA, resulting in an effective calculation precision of 5 bits. Figure 4 (D and E) illustrates the inference results of the Circle and XOR classifications, respectively, where the shaded regions show the underlying decision boundaries. For XOR and Circle datasets, the test accuracy stands at 97 and 100%, respectively. In contrast, the inference accuracy without nonlinearity is only 50% for both datasets. The poor accuracy in the absence of nonlinearity is expected as the network without optical nonlinearity essentially functions as a linear regression model, incapable of capturing the inherent nonlinear decision boundary of these datasets.

DISCUSSIONS

In this study, we implemented an NOMA for an energy-efficient optical nonlinearity by integrating Si PD and LC EOM at a single-pixel level. The NOMA achieves an optical ReLU function with switching energy down to 100 fJ across more than half a million pixels. We further demonstrated NOMA's energy efficiency, uniform nonlinear response, and compatibility with incoherent light through an image contrast enhancement task and highlighted the optical ReLU function in a binary classification task for deep ML-ONNs. In contrast with the state-of-the-art analog optoelectronic neural networks (*17*), NOMA eliminates the need for shuttling signals back and forth between optical and electrical domains, which should enable neural networks with more than one hidden layer in an energy-efficient manner. Further improvements to the switching energy can be achieved by decreasing EOM and PD capacitance, which is ultimately limited by the circuit Johnson noise. For instance, by reducing the pixel pitch from the current 20 to 3 μm (comparable to the state-of-art LC on silicon technology) and using a smaller-capacitance P-I-N junction as the PD, the capacitance of a single pixel can be as low as 100 aF, enabling optical modulation at sub-femtojoule switching energies.

We report switching times on the order of milliseconds for our energy-efficient nonlinearity. At these switching times, an NOMA-based ONN can be used as a drop in energy-efficient replacement for digital neural networks in applications where the inference task is frame rate limited. Such situations arise in a broad range of image recognition tasks, including applications in autonomous vehicles and facial recognition. Furthermore, an NOMA-based optical nonlinearity could be used in image compression as an efficient optical encoder layer that alleviates bandwidth challenges associated with large images. More excitingly, an energy-efficient optical nonlinearity, such as NOMA, could enable the development of optical neuromorphic computation platforms that mimic biological functions, such as visual perception.

The NOMA initially addressed a fundamental challenge of nonlinearity within the all-optical neuromorphic computing framework, which generally requires high energy efficiency, scalability, and broadband compatibility. We believe that our approach could catalyze the development of large-scale deep ONNs for intelligent edge computing and sensing in the future.

MATERIALS AND METHODS

Figure S1A is the photograph image of a completed NOMA with 750 × 700 pixels. Figure S1B illustrates the design of a 20 μm–by–20 μm NOMA pixel. Each pixel contains a rectangular Al mirror that occupies approximately $A_{\mathrm{Al}} = 240\ \mu\mathrm{m}^2$ or 60% of the total pixel area. The Al mirror is connected to a Si PD through a 4-$\mu\mathrm{m}^2$ contact via. The area of the Si PD is $A_{\mathrm{PD}} = 100\ \mu\mathrm{m}^2$. Beneath the Al mirror, a 6-μm-wide additional n-doped region extends across the column of pixels and connects to a common Al electrode at the edge of the device. The device's vertical structure, displayed in Fig. 1B, consists of an ITO glass and a Si backplane, forming an LC cell with an approximate thickness of $d = 3$ μm. Given the LC's refractive index anisotropy ($\Delta n = n_e - n_o$) of about 0.1, the retroreflected light's maximum path difference (ΔL) between the ordinary (o) and extraordinary (e) light is roughly 0.6 μm ($\Delta L = 2d\Delta n$), corresponding to a 1.8π phase retardation for a wavelength of 670 nm. This phase shift is sufficient for a full-range intensity modulation, which typically requires a phase modulation between 0 and π.

We used conventional planar fabrication techniques for the silicon substrate. To define the Si PD, we doped a 6-inch (15.24-cm) p-type Si wafer (ρ = 10 to 20 ohm·cm, Silicon Valley Microelectronics, USA) using phosphorus thermal diffusion under a $POCl_3$ atmosphere at 840°C. We next established a dielectric stack on the Si substrate through the deposition of 15 nm Al_2O_3 at 250°C by atomic layer deposition and 530 nm SiO_2 at 350°C by plasma-enhanced chemical vapor deposition. Notably, the Al_2O_3/Si interface hosts a substantial built-in charge density, which effectively passivates minor carrier recombination at the surface, reducing the surface leakage current of the Si PD (*50*). Following the formation of the dielectric stack, the wafer underwent a patterning and etching process to create contact vias for Al mirrors. This was followed by 200-nm Al sputtering and patterning. Then, we established an ohmic contact to the p-type Si using 100-nm Pt that was sputtered on the back side of the wafer (*51*). We used the ITO glass (MSE Supplies LLC) with a sheet resistance of around 30 to 60 ohm/sq. We coated both the silicon and ITO substrates with 40-nm SiO_2 via oblique e-beam evaporation, which served as the alignment layer for the LC (*52*). We bonded the Si and ITO pieces together using an ultraviolet (UV) curing adhesive (OG142, Fiber Optic Center, USA). We controlled the cell gap using microspheres (Micropearl SP-203, Sekisui Chemical Co. Ltd., Japan) with a diameter of 3 μm, which were placed along the periphery of the chip within the UV adhesive. We filled the sealed cell through a small fill port at the edge of the bonded chips with nematic LC (LC-VAST14, INSTEC, USA) using the capillary effect to uniformly form the LC layer. Last, we sealed the fill port using the same UV adhesive, completing the assembly process.

Supplementary Materials

This PDF file includes:
Supplementary Text S1 to S3
Figs. S1 to S9
Tables S1 and S2

REFERENCES AND NOTES

1. 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. http://ieeexplore.ieee.org/document/6757323/.
2. AI hardware has an energy problem. *Nat. Electron.* **6**, 463 (2023).
3. V. Sze, Y.-H. Chen, T.-J. Yang, J. S. Emer, Efficient processing of deep neural networks: A tutorial and survey. *Proc. IEEE* **105**, 2295–2329 (2017).
4. D. Marković, A. Mizrahi, D. Querlioz, J. Grollier, Physics for neuromorphic computing. *Nat. Rev. Phys.* **2**, 499–510 (2020).
5. H. Jaeger, B. Noheda, W. G. van der Wiel, Toward a formal theory for computing machines made out of whatever physics offers. *Nat. Commun.* **14**, 4911 (2023).
6. B. W. Blankenship, R. Li, R. Guo, N. Zhao, J. Shin, R. Yang, S. H. Ko, J. Wu, Y. Rho, C. Grigoropoulos, Photothermally activated artificial neuromorphic synapses. *Nano Lett.* **23**, 9020–9025 (2023).
7. E. Tseng, G. Kuo, S.-H. Baek, N. Matsuda, A. Maimone, F. Schiffers, P. Chakravarthula, Q. Fu, W. Heidrich, D. Lanman, F. Heide, Neural étendue expander for ultra-wide-angle high-fidelity holographic display. *Nat. Commun.* **15**, 2907 (2024).
8. 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).
9. S. Colburn, Y. Chu, E. Shilzerman, A. Majumdar, Optical frontend for a convolutional neural network. *Appl. Optics* **58**, 3179–3186 (2019).
10. A. Ryou, J. Whitehead, M. Zhelyeznyakov, P. Anderson, C. Keskin, M. Bajcsy, A. Majumdar, Free-space optical neural network based on thermal atomic nonlinearity. *Photonics Res.* **9**, B128–B134 (2021).
11. H. Li, P. Fu, Z. Zhou, W. Sun, Y. Li, J. Wu, Q. Dai, Performing calculus with epsilon-near-zero metamaterials. *Sci. Adv.* **8**, eabq6198 (2022).
12. N. Morrison, S. Pan, E. Y. Ma, Physics-agnostic inverse design using transfer matrices. *APL Mach. Learn.* **2**, 016115 (2024).
13. S. K. Vadlamani, T. P. Xiao, E. Yablonovitch, Physics successfully implements Lagrange multiplier optimization. *Proc. Natl. Acad. Sci. U.S.A.* **117**, 26639–26650 (2020).
14. C. Denz, *Optical Neural Networks* (Springer Science & Business Media, 2013).
15. 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).
16. P. L. McMahon, The physics of optical computing. *Nat. Rev. Phys.* **5**, 717–734 (2023).
17. 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).
18. D. A. B. Miller, Self-configuring universal linear optical component. *Photonics Res.* **1**, 1–15 (2013).
19. 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).
20. 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).
21. 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).
22. 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).
23. S. Bandyopadhyay, A. Sludds, S. Krastanov, R. Hamerly, N. Harris, D. Bunandar, M. Streshinsky, M. Hochberg, D. Englund, Single-chip photonic deep neural network with forward-only training. *Nat. Photonics* **18**, 1335–1343 (2024).
24. Z. Chen, A. Sludds, R. Davis, I. Christen, L. Bernstein, L. Ateshian, T. Heuser, N. Heermeier, J. A. Lott, S. Reitzenstein, R. Hamerly, D. Englund, Deep learning with coherent VCSEL neural networks. *Nat. Photonics* **17**, 723–730 (2023).
25. 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).
26. 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).
27. M. Miscuglio, Z. Hu, S. Li, J. K. George, R. Capanna, H. Dalir, P. M. Bardet, P. Gupta, V. J. Sorger, Massively parallel amplitude-only Fourier neural network. *Optica* **7**, 1812–1819 (2020).
28. 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).
29. L. Bernstein, A. Sludds, C. Panuski, S. Trajtenberg-Mills, R. Hamerly, D. Englund, Single-shot optical neural network. *Sci. Adv.* **9**, eadg7904 (2023).
30. T. Wang, S.-Y. Ma, L. G. Wright, T. Onodera, B. C. Richard, P. L. McMahon, An optical neural network using less than 1 photon per multiplication. *Nat. Commun.* **13**, 123 (2022).
31. S.-Y. Ma, T. Wang, J. Laydevant, L. G. Wright, P. L. McMahon, Quantum-noise-limited optical neural networks operating at a few quanta per activation. arXiv:2307.15712 [physics.optics] (2023).
32. D. A. B. Miller, Attojoule optoelectronics for low-energy information processing and communications. *J. Lightwave Technol.* **35**, 346–396 (2017).
33. R. W. Boyd, A. L. Gaeta, E. Giese, Nonlinear optics in *Springer Handbook of Atomic, Molecular, and Optical Physics*, G. W. F. Drake, Ed. (Springer International Publishing, 2023; https://doi.org/10.1007/978-3-030-73893-8_76) *Springer Handbooks*, pp. 1097–1110.
34. J. B. Khurgin, Nonlinear optics from the viewpoint of interaction time. *Nat. Photonics* **17**, 545–551 (2023).
35. I. A. D. Williamson, T. W. Hughes, M. Minkov, B. Bartlett, S. Pai, S. Fan, Reprogrammable electro-optic nonlinear activation functions for optical neural networks. *IEEE J. Select. Top. Quantum Electron.* **26**, 1–12 (2020).
36. K. Nozaki, S. Matsuo, T. Fujii, K. Takeda, A. Shinya, E. Kuramochi, M. Notomi, Femtofarad optoelectronic integration demonstrating energy-saving signal conversion and nonlinear functions. *Nat. Photonics* **13**, 454–459 (2019).
37. U. Efron, J. Grinberg, P. O. Braatz, M. J. Little, P. G. Reif, R. N. Schwartz, The silicon liquid-crystal light valve. *J. Appl. Phys.* **57**, 1356–1368 (1985).
38. J. A. Neff, R. A. Athale, S. H. Lee, Two-dimensional spatial light modulators: A tutorial. *Proc. IEEE* **78**, 826–855 (1990).
39. P. K. Shrestha, Y. T. Chun, D. Chu, A high-resolution optically addressed spatial light modulator based on ZnO nanoparticles. *Light Sci. Appl.* **4**, e259 (2015).
40. N. Collings, W. Xue, Liquid-crystal light valves as thresholding elements in neural networks: Basic device requirements. *Appl. Optics* **33**, 2829–2833 (1994).
41. I. F. Saxena, Adaptive multilayer optical neural network with optical thresholding. *Opt. Eng.* **34**, 2435–2440 (1995).
42. D. Zhang, D. Xu, Y. Li, Y. Luo, J. Hu, J. Zhou, Y. Zhang, B. Zhou, P. Wang, X. Li, B. Bai, H. Ren, L. Wang, A. Zhang, M. Jarrahi, Y. Huang, A. Ozcan, X. Duan, Broadband nonlinear modulation of incoherent light using a transparent optoelectronic neuron array. *Nat. Commun.* **15**, 2433 (2024).
43. E. Calisto, M. G. Clerc, V. Zambra, Magnetic field-induced vortex triplet and vortex lattice in a liquid crystal cell. *Phys. Rev. Res.* **2**, 042026 (2020).
44. R. Basu, L. J. Atwood, Homeotropic liquid crystal device employing vertically aligned carbon nanotube arrays as the alignment agent. *Phys. Rev. E* **102**, 022701 (2020).
45. H. Wang, T. X. Wu, X. Zhu, S.-T. Wu, Correlations between liquid crystal director reorientation and optical response time of a homeotropic cell. *J. Appl. Phys.* **95**, 5502–5508 (2004).
46. D. Andrienko, Introduction to liquid crystals. *J. Mol. Liq.* **267**, 520–541 (2018).
47. B. Hoex, S. B. S. Heil, E. Langereis, M. C. M. van de Sanden, W. M. M. Kessels, Ultralow surface recombination of c-Si substrates passivated by plasma-assisted atomic layer deposited Al2O3. *Appl. Phys. Lett.* **89**, 042112 (2006).
48. B. Chakraborty, S. Mukhopadhyay, Heterogeneous recurrent spiking neural network for spatio-temporal classification. *Front. Neurosci.* **17**, 994517 (2023).
49. N. Perez-Nieves, V. C. H. Leung, P. L. Dragotti, D. F. M. Goodman, Neural heterogeneity promotes robust learning. *Nat. Commun.* **12**, 5791 (2021).
50. R. Hezel, K. Jaeger, Low-temperature surface passivation of silicon for solar cells. *J. Electrochem. Soc.* **136**, 518–523 (1989).
51. H. C. Card, Aluminum—Silicon Schottky barriers and ohmic contacts in integrated circuits. *IEEE Trans. Electron Devices* **23**, 538–544 (1976).
52. A. Marino, E. Santamato, N. Bennis, X. Quintana, J. M. Otón, V. Tkachenko, G. Abbate, Ellipsometric study of vertically aligned nematic liquid crystals. *Appl. Phys. Lett.* **94**, 013508 (2009).

Acknowledgments

Funding: The liquid crystal fabrication, simulation, and optical measurements in this work are funded by the US Department of Energy, Office of Science, Basic Energy Sciences, Materials Sciences and Engineering Division under contract no. DE-AC02-05-CH11231 within the van der Waals heterostructure program KCFW16. The silicon wafer fabrication is funded by the National Science Foundation Award 2311205 (silicon wafer fabrication). **Author contributions:** F.W. conceived and supervised the research. Q.F. fabricated the devices with the help from R.G., C.S., R.Q., J.X., J.Z., and J.W. Q.F. and C.B.U. performed all the measurements with the assistance from R.G. and R.Q. C.B.U. and Q.F. wrote the code to control the ML-AONN system. Q.F., C.B.U., and F.W. analyzed the data. Q.F., C.B.U., and F.W. prepared the manuscript with input from all authors. All authors discussed the results. **Competing interests:** F.W., Q.F., and C.B.U. are inventors on patent application (no. 63/743,688) submitted by LBNL and UC Berkeley, which covers the nonlinear optical microdevice arrays discussed in this work. The other 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 12 August 2024
Accepted 2 January 2025
Published 31 January 2025
10.1126/sciadv.ads4224