Ultra-Compact and NonVolatile Nanophotonic Neural Networks

Huan Yuan, Zhicheng Wang, Zheng Peng, Jiagui Wu, and Junbo Yang**

A nanophotonic neural network (N-PNN) architecture is proposed with compact nanophotonic scattering units and a hybrid structure of silicon and nonvolatile arrayed Sb_2Se_3. This PNN can execute deep neural networks (DNN) classification and identification tasks with a broad operation bandwidth and very compact footprint. The reconstruction of the convolutional kernel core is realized by digitally switching the phase state of the Sb_2Se_3 array. Based on a three-dimensional finite-difference time-domain analysis, the core unit received only 4.92 × 2.34 μm^2 footprint. The convolution kernel unit weights are reconfigured with high-accuracy (7-bit) image processing and recognition in the wavelength C-band (1530–1570 nm). Furthermore, various deep-learning tasks (speech, digital patterns, and clothing patterns) are investigated. The accuracy of the classification and recognition efficiency reached almost the same level as that of a 64-bit computer. The size of the N-PNN is almost two orders of magnitude smaller than that of classic Mach–Zehnder interferometer meshes. It is conducive for scalability, high-radix DNN, and optoelectronic fusion of photonic integrated circuits and electronic integrated circuits.

1. Introduction

Machine learning, particularly deep neural networks[1,2] has attracted considerable attention for its efficient information processing, such as image classification[3–8] speech recognition,[9–11] and decision-making[12] In addition, neural networks are very useful in scientific fields such as biomedical science[13–15] autonomous driving[16] genetics[17,18] and photon chip design[19–22] At present, the computing platform of an artificial neural network is usually an electronic neural network (ENN) solution, which faces problems such as limited speed and high energy consumption[23] Compared with ENNs, photonic neural networks (PNNs) show dramatically high-speed and low-energy consumption features[5] making them a potential next-generation computing platform.

Various photonic architectures have been proposed for PNNs. For example, Feldmann et al. developed a micro-ring-based optical spiking neurosynaptic network[24] Shen et al. demonstrated coherent PNNs using a Mach–Zehnder interferometer (MZI) array[11] Tian et al. demonstrated a photonic matrix architecture[25] using the real part of a nonuniversal N × N MZI mesh to effectively reduce the array numbers. MZI arrays have gradually become a classical PNN scheme. Lin et al. reported a diffraction PNN architecture utilizing light diffraction during 3D-printed screens[26] with instantaneous convolution processing and almost zero power consumption. However. It is desirable to develop scalable and high-radix (e.g., 1024 × 1024) neural networks. Recently, Lightmatter's Mars device integrated 64 × 64 MZI meshes with 150 mm^2 size[27] The predicted size for the 1024 × 1024 MZI meshes was approximately 384 cm^2[28] Size is a significant obstacle. Additionally, most MZI meshes are volatile. The representation weight matrix requires a series of controllers (e.g., thermal electrodes), which consume a significant amount of energy.

Therefore, nonvolatile and high-density nanophotonics are extremely attractive. Optical phase change materials (O-PCMs) offer tempting solutions for nonvolatility, and nanophotonics designs can receive very compact components. Adibi et al. systematically summarized the properties of O-PCMs, and their marvelous applications[29] In O-PCMs, the crystalline and amorphous states provide large phase and amplitude modulation on the micro/nanoscale[30] The specific nonvolatility of O-PCMs offers significantly low energy consumption. Previously, O-PCMs have been successfully applied in programmable optical switching[30–34] optical storage[35–37] and reconfigurable elements[38–42] etc. However, classic O-PCMs, such as the Ge_x-Sb_y-Te_z alloy and VO_2 exhibit some optical loss in different states, which is not conducive to the ultra-low loss requirement.

H. Yuan, J. Wu
School of Physical Science and Technology
Southwest University
Chongqing 400715, China
E-mail: mgh@swu.edu.cn

H. Yuan, Z. Wang, Z. Peng, J. Yang
Center of Material Science
National University of Defense Technology
Changsha 410073, China
E-mail: yangjunbo@nudt.edu.cn

H. Yuan
College of Electronics and Information Engineering
Sichuan University
Chengdu 610065, China

Z. Wang, Z. Peng
College of Artificial Intelligence
Southwest University
Chongqing 400715, China

The ORCID identification number(s) for the author(s) of this article can be found under https://doi.org/10.1002/adom.202300215

DOI: 10.1002/adom.202300215

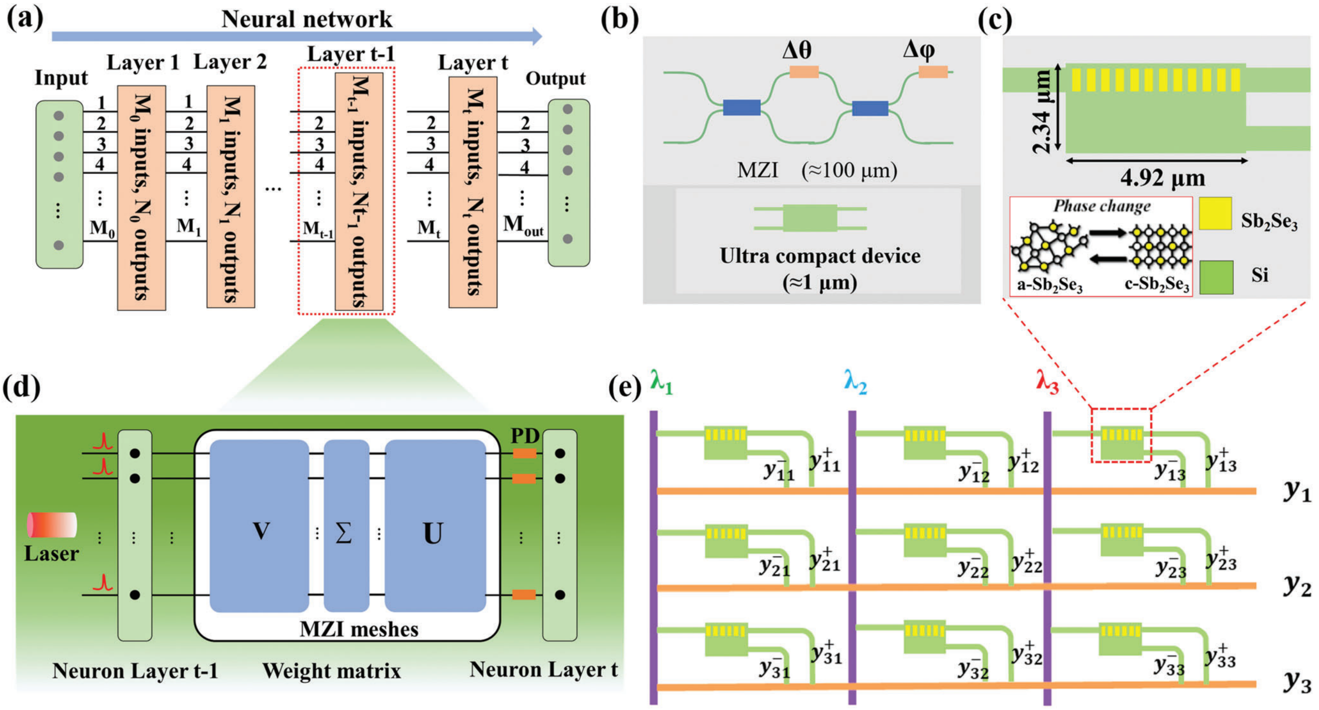


Figure 1. Schematic diagram of N-PNNs. a) neural network framework comprising an input layer, multiple hidden layers, and one output layer. b) Structure diagram of a classic Mach–Zehnder interferometer (MZI) and ultra-compact device. c) Detailed structure of ultra-compact nanophotonic units. d) Schematic representation of conventional MZI-based optical matrix multiplication from layer n-1 to layer n. e) Schematic diagram of 3 × 3 matrix vector multiplication of ultra-compact nanophotonic PNN architecture.

Additionally, VO_2 is a pseudo-phase transition material that is not very nonvolatile. Very recently, Sb_2Se_3 has been reported as a novel nonvolatile choice with almost zero loss in different states[43,44] making it very attractive for ultralow-loss nanophotonics, and the high saturation, high resolution, and high-efficiency metapixels in metasurfaces with Sb_2Se_3[45] As the cyclability, fortunately, some good advances have been reported. For example, Raoux et al. achieved 10^{11} cycles by switching PCM states[42] Zheng et al. demonstrated that $Ge_2Sb_2Te_5$ can reversibly trigger over 1000 large-area phase transitions with almost zero additional loss[41] and for the new family PCM Sb_2Se_3, Delaney et al. demonstrated a stable switching endurance of over 4000 cycles[43]

Compared with ENNs, PNNs have broader bandwidth and faster speed, but still face challenges such as low-density integration. Very recently, Shi et al. proposed new on-chip neurons using nonlinear germanium silicon photodiodes, whose nonlinear parts are concentrated in a compact size ≈4.3 × 8 μm^2[46] Bai et al. proposed an microcomb-driven chip-based photonic processing unit, and show a preeminent photonic-core compute density of over 1 trillion of operations per second per square millimeter (TOPS mm^{-2})[47] Furthermore, high-density photonic integrated circuits (PICs) are essential when considering tens of millions of possible photonic components in a monolithic chip. Digital nanophotonics is an effective method for realizing high-density PICs[48] and inverse-designed digital nanophotonics could have an ultra-small footprint, breaking the limitations of traditional periodic metamaterials and allowing arbitrary topological structures[49–51]

Here, we propose a nanophotonic neural network (N-PNN) architecture that combines nonvolatile O-PCM (Sb_2Se_3) and the inverse design of digital nanophotonics. It integrates ultra-compact nanophotonic units, whose power distribution between waveguides could be controlled with 128 distinguishable levels of power contrast. The core unit reaches a 4.92 × 2.34 μm^2 footprint. Next, a fully programmable convolution kernel array is constructed to realize the deep learning tasks of image recognition, digital recognition, and speech recognition, aiding in the regulation of state distribution in the O-PCM array. Comparatively, our N-PNNs are ultra-compact and nonvolatile. It may offer new nanophotonic solutions for large-scale photonic processing and over conventional photonic designs on chip size, insertion loss, and power consumption.

2. Results and Discussion

Figure 1 shows the N-PNN architecture for deep learning tasks. Figure 1a,d presents the framework of typical neural networks, comprising three parts: the input layer, middle hidden layer, and output layer. Each layer comprises a group of neurons, and each neuron is connected to other neurons in the next layer. These interconnected neurons propagate the input signals in the form of matrix multiplication. In each iteration, by inputting training data into the neural network, forward propagation is performed to output the training results, and then, backpropagation is used to optimize the weight matrix.

As shown in Figure 1b, the core of a PNN is typically constructed using an MZI grid. The neuronal signals of each layer

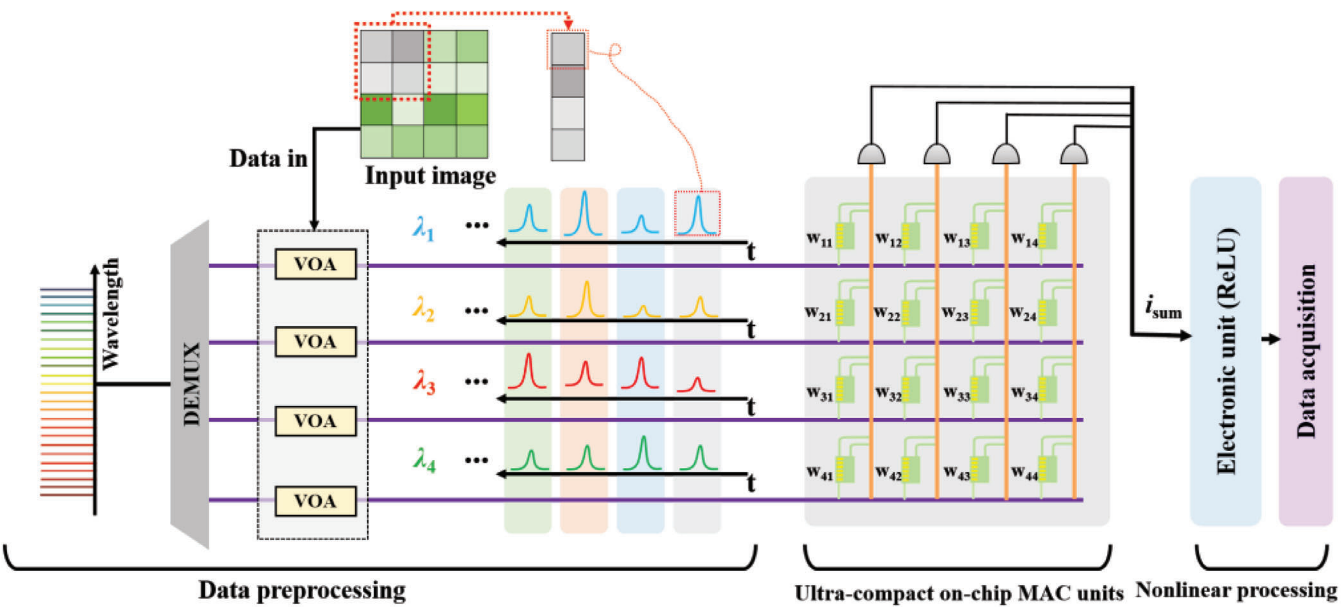


Figure 2. The holistic diagram of the proposed N-PNNs chip based on nanophotonic units. There are three major sections: data preprocessing sections, multiply-accumulate (MAC) section, and nonlinear processing section. The input vector are multiplied by the nonvolatile phase-change memory unit matrix after passing through the VOA and summing after optical detection along the parallel photodetectors.

are converted into optical signals, and matrix multiplication is realized by thermal electrode modulation (typical illustration is shown in **Figure** 2d in[11]). Comparatively, an ultra-compact nanophotonic unit has a very small footprint, as shown in Figure 1b. In our design, the nanophotonic unit, comprising a hybrid structure of silicon and arrayed Sb_2Se_3(as shown in Figure 1c), has a footprint of only approximately 2.34 × 4.92 µm^2. The size exhibits a significant reduction of two orders of magnitude. A 3 × 3 matrix-vector multiplication of the designed compact neural network is shown in Figure 1e. The input vector is converted into different input wavelengths. Each weight of the matrix is represented by an ultra-compact nanophotonic unit, and the output vector can be expressed as $y_i = y_{ij}^{+} - y_{ij}^{-}$. Figure 2 shows the holistic diagram of the proposed N-PNNs chip based on nanophotonic units, which consists of three major sections: data preprocessing sections, multiply-accumulate (MAC) section, and nonlinear processing section. Multi-wavelength laser source injects into de-multiplexer (DEMUX). Pixel blocks of image data are encoded as optical pulses and transmitted to MAC units by variable optical attenuator (VOA), and then input into the ultra-compact on-chip MAC units, where the nanophotonic structures be assembled into integrated photonic networks for parallel convolutional processing of matrix-vector multiplication. Next, the outputs of MAC units are detected by photodetectors, and input the nonlinear processing section, where the nonlinear activation function is realized with an electronic unit.

Figure 3a depicts the detailed structure of the nanophotonic unit device. This unit, exhibiting low loss and compactness, can be used as the weight unit of the optical convolution kernel in the N-PNN. The device consists of an input waveguide, two output waveguides, and a silicon photonic waveguide coupling region consisting of an Sb_2Se_3 array and etched hole array embedded in silicon. The device is divided into upper and lower coupling regions. The upper coupling region is composed of 30 Sb_2Se_3 arrays with a length of 90 nm and a width of 620 nm, and two adjacent cells are separated by 70 nm. A 70 nm interval makes the central separation between Sb_2Se_3 cells goes to 160 nm. It is over 100 nm and useful to avoid the possible thermal crosstalk between adjacent PCM cells[52,53] The lower coupling region is discretized into hundreds of 120 × 120 nm^2 pixels, each one being square shaped with a central circular hole with a radius of 45 nm and a hole depth of 220 nm. This structure with irregular patterns provides a substantial design degree of freedom that allows the input light to undergo complex scattering behavior within an ultra-compact region. When the device is in the OFF state, the TE$_0$ transverse electric mode light enters the coupling region through the left input channel and outputs through the right upper output channel. When the device is *ON*, the TE$_0$ transverse mode light is outputted through the lower output channel. Moreover, the Sb_2Se_3 array of the device is programmable, which means that we can dynamically switch the power output ratio across the device.

To realize logic functions, we optimized the device using a silicon-Sb_2Se_3 hybrid digital element structure. At 1550 nm wavelength, the complex refractive index of amorphous Sb_2Se_3 (a-Sb_2Se_3) is 3.285 + 0.000i, whereas that of crystalline Sb_2Se_3 (c-Sb_2Se_3) increases to 4.050 + 0.000i[43] The phase transition process can be triggered electrically or optically. When a phase transformation occurs, its optical properties change significantly, and a logic function can be realized by switching between different states. Compared with the well-known GST and GSST O-PCMs[30] c-Sb_2Se_3 and a-Sb_2Se_3 absorptions are both 0. Comparatively, the refractive indexes of GST and GSST are as follows: amorphous-GST: 4.600 + 0.120i, crystalline-GST: 7.450 + 1.490i, amorphous-GSST: 3.390 + 0.00018i, and crystalline-GSST: 5.140 + 0.420i. Therefore, the optical absorption of GST and GSST is relatively large in the crystalline state. Additionally, the

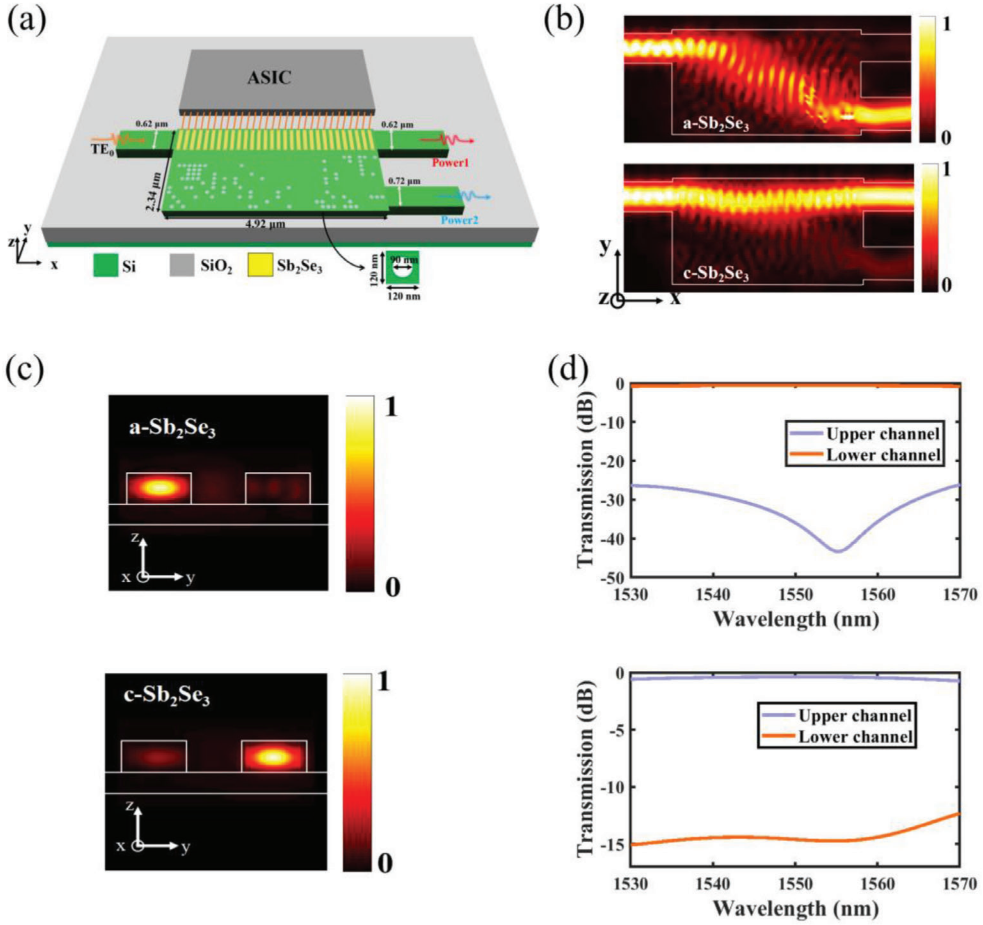


Figure 3. a) Schematic of the 1 × 2 programmable nanophotonic units. b) The energy density at 1550 nm in both crystalline and amorphous conditions was simulated using the FDTD. We plotted the electromagnetic energy density $U = \epsilon|E|^2 + \mu|H|^2$. c) The TE$_0$ mode energy intensity of the cross-sections of the two output channels. d) The spectral transmission curves of nanophotonic units in the upper and lower channels are simulated.

refractive index of a-Sb$_2$Se$_3$ is very close to that of Si; therefore, the transmission of encoded regulated light depends mainly on c-Sb$_2$Se$_3$.

As shown in Figure 3, the upper input and upper output waveguides are both 620 nm wide, and the lower output waveguide is 720 nm wide. In this study, all devices are used to build N-PNN kernels, which is 2.34 × 4.92 µm^2. In this design, different pixel states determine the refractive index distribution in the core region. Compared with the conventional periodic photonic bandgap structure, this artificial uneven distribution allows for more flexible refractive-index changes at the subwavelength scale. The waveguide supports the TE$_0$ transverse mode. In addition, our proposed devices are compatible with casting processes for standard silicon photonic devices. The three-dimensional finite-difference time-domain (3D FDTD) method is used for the simulation and transmission calculation, as shown in Figure 3b, which shows the electromagnetic energy densities corresponding to a-Sb$_2$Se$_3$ and c-Sb$_2$Se$_3$ at λ = 1550 nm. Figure 3c shows the energy intensity of the TE$_0$ modulus of the output channel cross-section. Figure 3d shows the transmission curves of the two output channels in the wavelength range −1530–1570 nm. At 1550 nm, the insertion losses (ILs) of the device are −0.545 dB (a-Sb$_2$Se$_3$: lower channel) and −0.378 dB (c-Sb$_2$Se$_3$: upper channel), and the crosstalk values are −35.482 dB (a-Sb$_2$Se$_3$: upper channel) and −14.610 dB (c-Sb$_2$Se$_3$: lower channel). From the amorphous state to the crystalline state, the transmission of TE$_0$ increases from −0.545 dB to −0.378 dB, and the transmission is close and efficient. The ILs and crosstalk between the different output channels remain well above the 40 nm bandwidth range. This design method effectively utilizes the refractive index variation and extremely low extinction coefficient of Sb$_2$Se$_3$ during the phase transformation to promote light transmission to different output ports. This ultra-compact nanophotonics structure of the N-PNN system can provide significantly high photonic-core compute density and tensor processing units. Taking into account the additional footprint area required for MAC operation, we estimated the total footprint of each unit, still with good compactness. For the proposed 3 × 3 matrix, we obtain the compute density of the structure assuming a normal optics pulse rate of 10 GHz, and compare it with some related work (See Note S6, Supporting Information).

To realize a programmable optical convolutional neural network, we determine the power contrast τ by measuring the power of the upper and lower output channels and calculating their

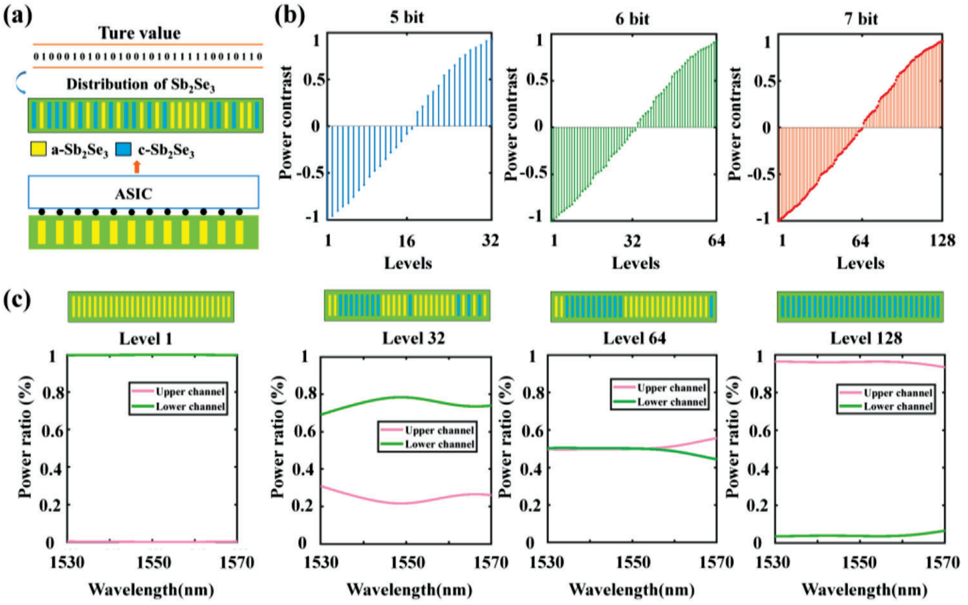


Figure 4. a) Sb_2Se_3 distribution corresponding to the true value. b) programmable power beam splitter with different levels, 32, 64, and 128, of power contrast τ, corresponding to 5-, 6-, and 7-bit programming resolution, respectively. c) Different levels of power contrast are shown at 7-bit resolution. Level 1, 32, 64, 128 shows the distribution of the Sb_2Se_3 array and the gradient process of the power distribution.

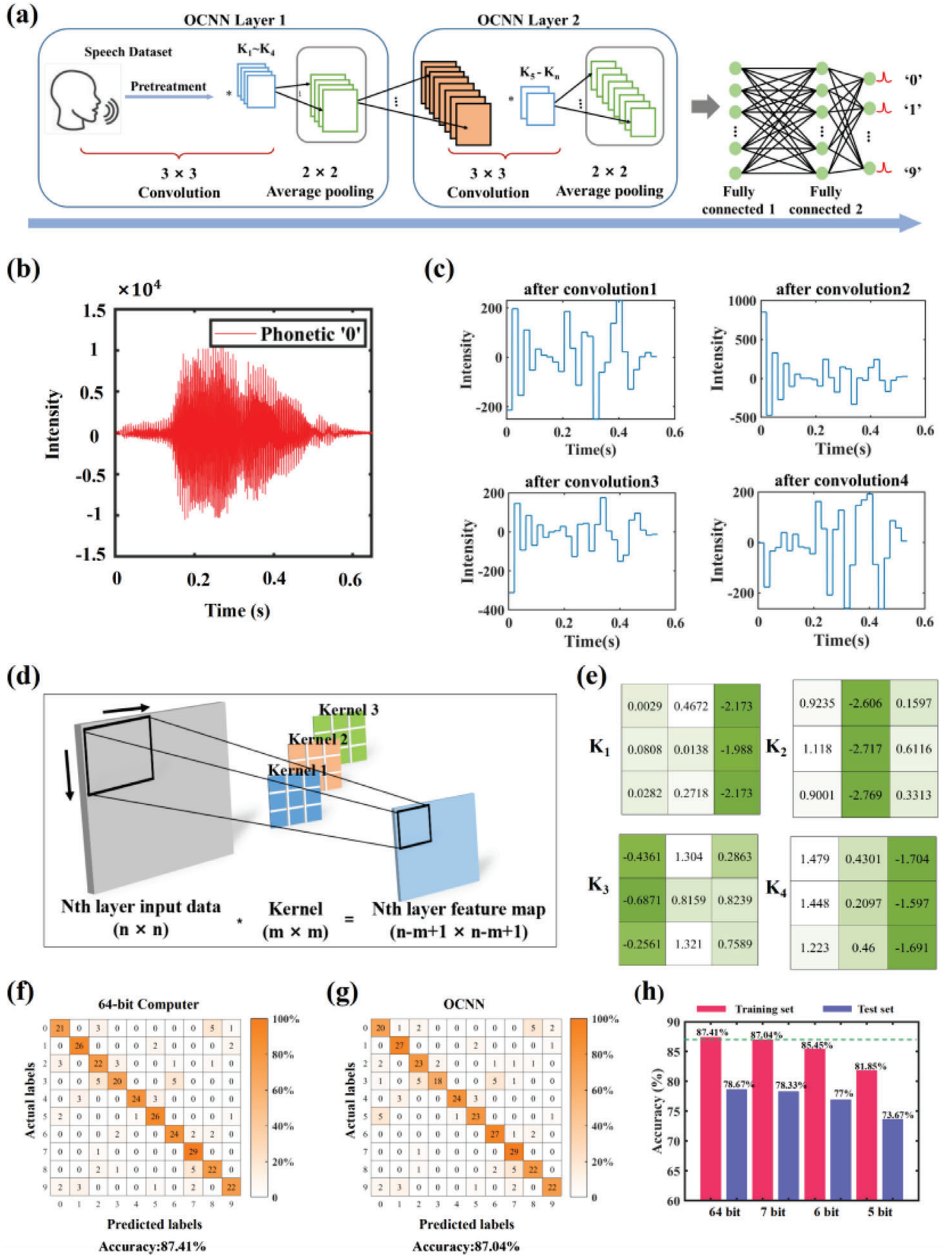
difference. Here, we define the power contrast τ as

$$\tau = \frac{T_{\text{upper}} - T_{\text{lower}}}{T_{\text{upper}} + T_{\text{lower}}} \tag{1}$$

where T_{upper} and T_{lower} are the transmission efficiencies of the upper and lower output channels at specified wavelengths, respectively, and τ ranges from −1 to 1. As shown in **Figure 4**a, the logic circuit of the ASIC can be connected to the distribution of the target light field through a specific truth table (the optical pulse or the electric pulse can digitally achieve the O-PCM state switching, as shown in Figure 3a). The ASIC function mainly controls the electrical heating of the Sb_2Se_3 unit in a program-controlled manner. To program the Sb_2Se_3 transition from amorphous (1) to crystalline (0) (or crystalline (0) to amorphous (1)), the control pulse sequence of the ASIC is used to program the array of phase-transition materials in the device. When the input TE_0 mode light passes through the device, matrix multiplication can be performed by calculating the power contrast ratio τ: $y = x{\cdot}w$. Figure 4b shows the multistage programmability of the power beam splitter, with an ideal power contrast range of −1–1. At 1550 nm, the device reaches a distinguishable level of 128 (7 bits), ranging from −0.9995 to +0.9274, indicating the low loss and high precision of the device, which is crucial for the construction and training of large-scale neural networks. In Figure 4b, we show the discrimination levels under different resolutions, which are discussed later in the neural-network recognition task (**Figure 5**h). Based on the network architecture shown in Figure 1e, we considered the phase state of each Sb_2Se_3 cell of the device as a learnable variable and obtained the final phase distribution through multiple iterations according to the corresponding optimization target. Figure 4c shows the power-split ratio corresponding to different Sb_2Se_3 array distributions. When Sb_2Se_3 is in an amorphous state, the power beam splitter can effectively transmit TE_0 mode light from the lower output channel, with a power ratio of 99.95% in the broadband range. When Sb_2Se_3 switches to the crystal state, the power beam splitter effectively transmits the TE_0 mode light from the upper output channel, with a power ratio of 92.74% in the broadband range. Compared with the conventional MZI-PNN with the inability of memory weight without a power supply, the nonvolatile N-PNNs are very conductive to processing in memory [54,55]

3. N-PNNs for Deep Learning Recognition

Our N-PNNs architecture is based on the above nonvolatile storage computing capability and high-precision programmable optical convolutional multiplication computing unit and adopts several classical datasets (such as MNIST and Fashion-MNIST) to perform classification and recognition tasks. The network architecture consists of two alternating 3 × 3 convolution layers, 2 × 2 maximum pooling layers, and two fully connected layers. An open-source machine learning framework in Python was used to build the network, and the network was trained using the stochastic gradient descent method to obtain the desired output. The nonlinear activation function was ReLU, the loss function was the mean square error, and the optimizer was the Adam optimizer. We use the stochastic gradient descent method on a traditional computer and the standard back propagation algorithm to train the matrix parameters used in the PNN. The closest value is searched in the simulated weight parameter with 7-bit resolution (corresponding to the distribution of different Sb_2Se_3 arrays). The optimization process of Sb_2Se_3 arrays is detailed in Note S4, Supporting Information. After the framework of the N-PNNs is determined, the convolutional kernel



K_1

0.0029	0.4672	-2.173
0.0808	0.0138	-1.988
0.0282	0.2718	-2.173

K_2

0.9235	-2.606	0.1597
1.118	-2.717	0.6116
0.9001	-2.769	0.3313

K_3

-0.4361	1.304	0.2863
-0.6871	0.8159	0.8239
-0.2561	1.321	0.7589

K_4

1.479	0.4301	-1.704
1.448	0.2097	-1.597
1.223	0.46	-1.691

64-bit Computer

	0	1	2	3	4	5	6	7	8	9
0	21	0	3	0	0	0	0	0	5	1
1	0	26	0	0	0	2	0	0	0	2
2	3	0	22	3	0	0	1	0	1	0
3	0	0	5	20	0	0	5	0	0	0
4	0	3	0	0	24	3	0	0	0	0
5	2	0	0	0	1	26	0	0	0	1
6	0	0	0	2	0	0	24	2	2	0
7	0	0	1	0	0	0	0	29	0	0
8	0	0	2	1	0	0	0	5	22	0
9	2	3	0	0	0	1	0	2	0	22

OCNN

	0	1	2	3	4	5	6	7	8	9
0	20	1	2	0	0	0	0	0	5	2
1	0	27	0	0	0	2	0	0	0	1
2	2	0	23	2	0	0	1	1	1	0
3	1	0	5	18	0	0	5	1	0	0
4	0	3	0	0	24	3	0	0	0	0
5	5	0	0	0	1	23	0	0	0	1
6	0	0	0	0	0	0	27	1	2	0
7	0	0	1	0	0	0	0	29	0	0
8	0	0	1	0	0	0	2	5	22	0
9	2	3	0	0	0	1	0	2	0	22

Figure 5. Construction process of N-PNN architecture and the operation process of speech sequence recognition. a) The N-PNN consists of two sets of kernel computing layers composed of convolutional layers and pooling layers, and a fully connected layer. The final output determines the accuracy of the input image. b,c) The time series of the speech "0" signals and the speech sequence signal are generated by passing through the first convolution layer with the four convolution kernels k_1-k_4. d) Optical convolution schematic for image processing, convolution of convolution kernels. e) Convolutional kernel matrix k_1-k_4 obtained after the completion of N-PNN training. f,g) The identification results of the 64-bit computer and the N-PNN calculation show high close accuracy. h) Identification accuracy of four convolutional neural networks with different resolutions.

matrix can be programmed and reprogrammed in the system, and then, a number of complex target recognition tasks can be completed.

The three types of datasets contain different information, and the basic content of the dataset is as follows.

(a) The Handwritten-MNIST dataset contains 10 handwritten digits, ranging from 0 to 9. It is divided into a training set and a test set, with 60 000 and 10 000 images, respectively.

(b) The Fashion-MNIST dataset contains 10 clothing images. It is divided into a training set and a test set, with 60 000 and 10 000 images, respectively.

(c) The speech dataset contains 10 speech sequences ranging from 0 to 9. It is divided into a training set and a test set, with 2700 and 300 voice data, respectively.

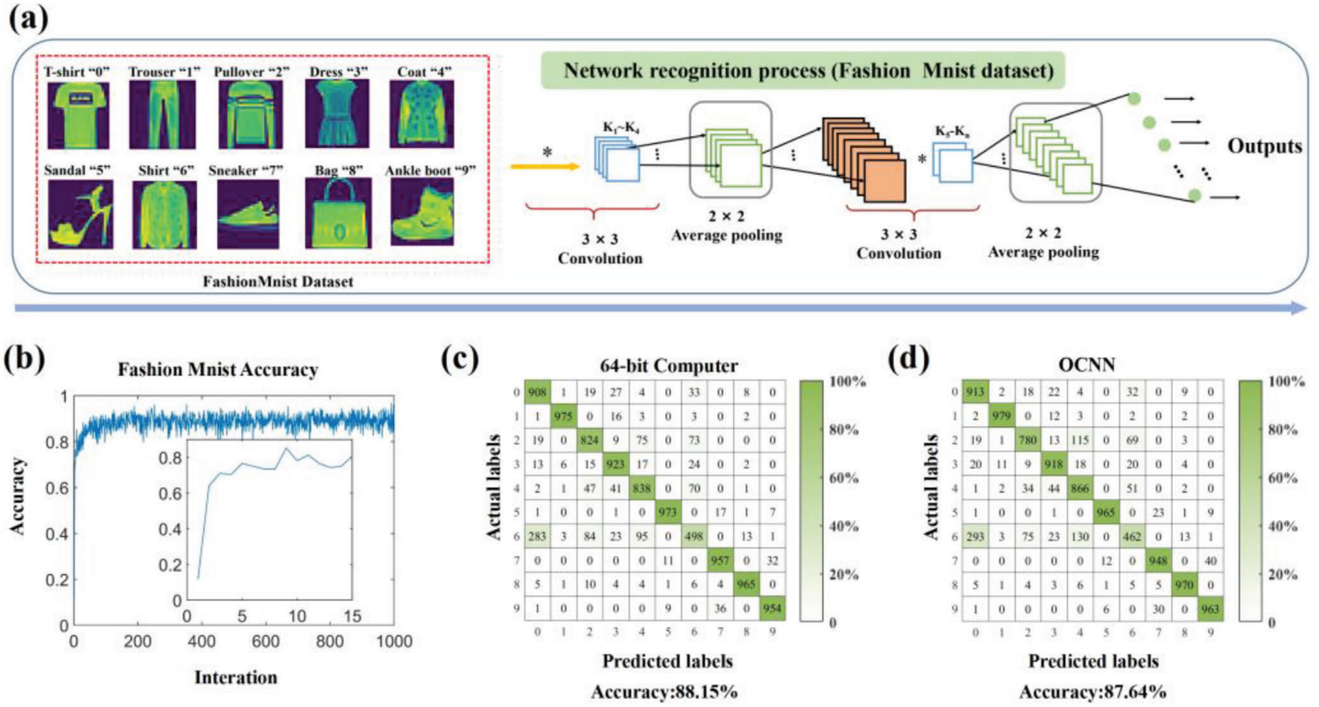


Figure 6. Construction process of the N-PNN architecture and the operation process of Fashion-MNIST dataset recognition. a) N-PNNs consists of two sets of kernel computing layers composed of convolutional layers and pooling layers, and a fully connected layer. The final output determines the accuracy of the input image. b) Identification accuracy varied with the number of training iterations. c,d) The identification results of the 64-bit computer and the N-PNNs.

Optical convolution operations can be converted into kernel matrix multiplications. Furthermore, the multiplication operation using an optical signal with a compact device can also be applied in the average pooling and fully connected layers to achieve complete N-PNNs. Figure 5a shows the architecture and identification results of N-PNNs (Subsequent N-PNN structure for the digital and garment images is consistent with it). By preprocessing the speech sequence, an optical signal can be transmitted in a photonic circuit. Taking a total of ten speech data from 0 to 9 as an example, different types of speech sequences output different encoded information after passing through the N-PNN structure. The input signal of the speech sequence "0" is shown in Figure 5b, and Figure 5c shows the intermediate signal processed by four different convolution cores in the first convolutional layer. Figure 5d illustrates the working principle of N-PNN data processing, where the gray matrix is the $n \times n$ input data, producing $(n - m + 1)^2$ feature data information after $m \times m$ convolution kernel processing. The input black box corresponds to the feeling field of the pixels in the feature data. Figure 5e shows the numerical values of the different convolution kernels in the first layer. There are distinct numerical segmentations between different columns in the kernel matrix; thus, it exhibits good signal processing power. The traditional CNN of a 64-bit computer trains and identifies the same speech sequence signal on the same basic network architecture. As shown in Figure 5f,g, the proposed N-PNN architecture can achieve a classification accuracy of 87.04%, which is almost consistent with a 64-bit computer (87.41%). Figure 5h shows the recognition accuracy corresponding to the network structure with different resolutions. Compared with the relatively different recognition efficiency of the 5-bit, 6-bit, and 64-bit computers, the N-PNN recognition results with 7-bit resolution can be very close to the recognition efficiency of the electronic computer.

Arbitrary weights can be achieved by programming the regulated phase distribution of the Sb_2Se_3 arrays. Therefore, our designed compact photonic matrix units are digitally programmable. Using the proposed N-PNN architecture, we analyzed another task, the identification of the Fashion-MNIST datasets. We compared a 64-bit computer-based CNN and N-PNN. **Figure 6**a shows the basic framework of the neural network, where we present the original images of ten different categories of ornaments. Figure 6b shows that the accuracy of the network training on a 64-bit computer changes with iteration. Although the training process has some fluctuations, it is still stable at 80%–90% efficiency. The classification test results are shown in Figure 6c,d. For the Fashion-MNIST dataset, the proposed N-PNN structure can achieve 87.64% classification accuracy, which is close to 88.15% for the 64-bit computer.

Similarly, we show that the most common MNIST datasets are used in neural network identification to validate the efficiency of N-PPNs. Figure 7a shows the network architecture of N-PNNs, with a total of ten (0–9) digital pictures. The corresponding output signal was obtained after passing through the trained N-PNN. **Figure 7**b shows that the network training accuracy on a 64-bit computer varies with iteration. Because the digital image data have a relatively simple pattern shape and strong differences, the convergence speed is very fast, and its recognition accuracy can be almost close to 100%. The classification test results are

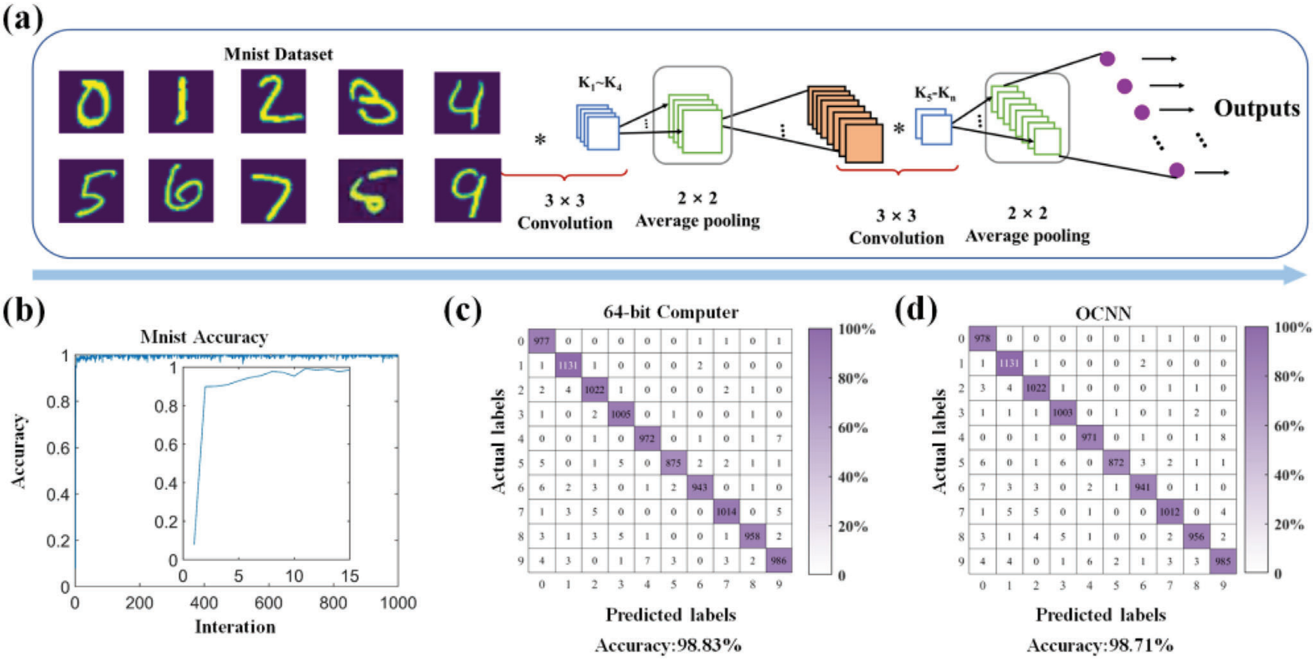


Figure 7. Construction process of the N-PNN architecture and the operation process of MNIST dataset recognition. a) The N-PNN consists of two sets of kernel computing layers comprising convolutional layers and pooling layers, and a fully connected layer. The final output determines the accuracy of the input image. b) The identification accuracy varied with the number of training iterations. c,d) The identification results of the 64-bit computer and the N-PNN calculation.

Table 1. Structural parameters and properties of photonic neural networks.

Refs.	[11]	[24]	[25]	This work
Core unit size	>100 µm	>20 µm	>100 µm	2.34 × 4.92 µm^2
Containing O-PCM	No	Yes ($Ge_2Sb_2Te_5$)	No	Yes (Sb_2Se_3)
Structure type	MZI	Ring resonator	MZI	Ultra-compact nanophotonics
Nonvolatile	No	Yes	No	Yes
Reconfigurable	No	Yes	No	Yes
Multitarget recognition	No	No	No	Yes
Identify the number of types	Vowels (4 Types)	Digit (4 Types)	Digit (10 Types)	Digit (10 Types)/ Decoration (10 Types)/ Speech (10 Types)

shown in Figure 7c,d. For the MNIST dataset, the proposed N-PNN structure can achieve 98.71% classification accuracy, which is close to 98.83% for a 64-bit computer.

Furthermore, we compare the results with those of other recently reported PNN. As shown in **Table 1**, the proposed N-PNN has advantages in terms of the network size and reconstruction properties. For instance, most reported networks were not nonvolatile[11,25] Compared to other O-PCMs (e.g., GST and GSST)[24] Sb_2Se_3 has a smaller optical loss and is therefore conducive for larger-scale neural networks[43] Moreover, the footprint of N-PNNs is much smaller. Compared with the results of[11,25] the size is reduced by approximately two orders of magnitude. Furthermore, our N-PNN is capable of multirecognition tasks. In the tasks of very large-scale image recognition, it has the combined advantages of fast speed, low power consumption, nonvolatility, and ultra-small size.

The ultra-compact photonic matrix unit is suitably extended to large-scale arrays because of its small size and low loss. **Figure 8** shows an example of the potential construction of larger N-PNNs using these core units. As shown in Figure 8, we present an architecture for extended matrix-vector multiplication in large-scale PNNs. Figure 8a shows the numerical calculation model corresponding to the components of the first N-PNN convolution layer. Figure 8b illustrates the conversion of the numerical computation portion of a convolutional layer into an actual photonic network. The photonic network consists of a series of photonic matrix interconnections. The input vector $x = [x_1, x_2, ..., x_M]$ is transformed through different input wavelengths $[\lambda_1, \lambda_2, ..., \lambda_M]$, and the corresponding output vector $y = [y_1, y_2, ..., y_N]^T$ is obtained by the weight matrix composed of M × N horizontal and vertical compact devices (the value of each input signal x_i is represented by a different input power). Large-scale

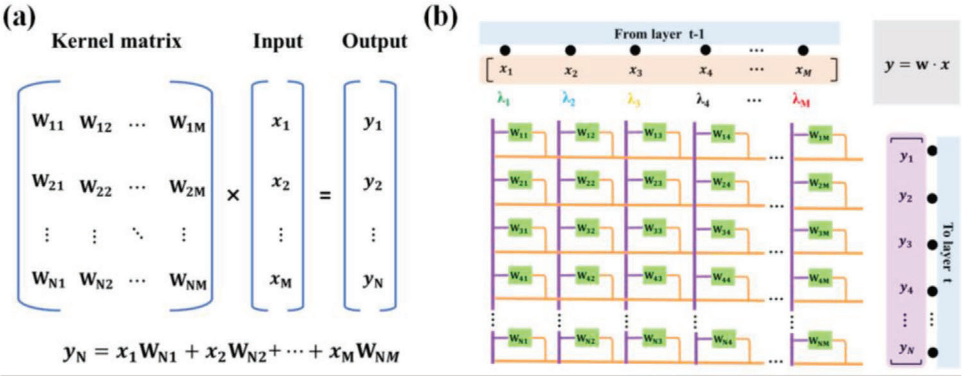

(a)
Kernel matrix
Input
Output
W_{11} W_{12} ⋯ W_{1M}
W_{21} W_{22} ⋯ W_{2M}
W_{N1} W_{N2} ⋯ W_{NM}
x_1 x_2 x_M
y_1 y_2 y_N
×
=
$y_N = x_1W_{N1} + x_2W_{N2} + \cdots + x_MW_{NM}$
(b)
From layer t-1
x_1 x_2 x_3 x_4 ⋯ x_M
λ_1 λ_2 λ_3 λ_4 ⋯ λ_M
$y = w \cdot x$
y_1 y_2 y_3 y_4 y_N
To layer t

Figure 8. a) Schematic diagram of the extended matrix vector multiplication. b) Schematic representation of large photon arrays of ultra-compact programmable devices used to perform multiplication operations. Horizontal and vertical waveguides subdivide the network into M × N units.

N-PNNs constructed with nanophotonic units can greatly avoid the defects caused by the large unit footprint when considering tens of millions of components in a single monolithic PIC.

4. Conclusions

We propose an architecture of N-PNNs based on ultra-compact nanophotonic units, which combine a nonvolatile O-PCM (Sb_2Se_3) and the inverse design of digital nanophotonics. The power distribution between waveguides can be controlled with 128 distinguishable power contrast levels. The core unit reaches only 4.92 × 2.34 μm^2 footprint. Next, the task of high-precision image processing and recognition in the C-band (1530–1570 nm) is realized by reconstructing the convolutional kernel weights through such N-PNNs. These results show that the recognition efficiency of our designed N-PNNs is almost the same as that of 64-bit computers in multiple target recognition tasks. It may offer a candidate for next-generation PNNs for scalability and high-radix DNN, featuring large-scale, nonvolatile, ultra-low power consumption, and ultra-compact size.

Supporting Information

Supporting Information is available from the Wiley Online Library or from the author.

Acknowledgements

H.Y., Z.W., and Z.P. contributed equally to this work. This work was funded by the Innovation Research 2035 Pilot Plan of Southwest University (SWU-XDPY22012); Chongqing Science Funds for Distinguished Young Scientists (cstc2021jcyj-jqX0027); the National Natural Science Foundation of China (60907003, 61805278, 12272407, 62275269, 62275271); National Key R&D Program of China (2022YFF0706005); Innovation Support Program for Overseas Students in Chongqing (cx2021008); China Postdoctoral Science Foundation (2018M633704); Foundation of NUDT (JC13-02-13, ZK17-03-01); Hunan Provincial Natural Science Foundation of China (13JJ3001); and the Program for New Century Excellent Talents in University (NCET-12-0142).

Conflict of Interest

The authors declare no conflict of interest.

Data Availability Statement

The data that support the findings of this study are available from the corresponding author upon reasonable request.

Keywords

digital nanophotonics, phase change material, photonic neural networks

Received: January 28, 2023
Revised: April 3, 2023
Published online: May 14, 2023

[1] Y. Lecun, Y. Bengio, G. Hinton, *Nature* **2015**, *521*, 436.
[2] J. Schmidhuber, *Neural Networks* **2015**, *61*, 85.
[3] A. Krizhevsky, I. Sutskever, G. E. Hinton, *Commun. ACM* **2017**, *60*, 84.
[4] K. H. Jin, M. T. Mccann, E. Froustey, M. Unser, *IEEE Trans. Image Process* **2017**, *26*, 4509.
[5] 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, *Nature* **2021**, *589*, 44.
[6] C. Wu, H. Yu, S. Lee, R. Peng, I. Takeuchi, M. Li, *Nat. Commun.* **2021**, *12*, 96.
[7] T. Wang, S. Y. Ma, L. G. Wright, T. Onodera, B. C. Richard, P. L. McMahon, *Nat. Commun.* **2022**, *13*, 123.
[8] H. H. Zhu, J. Zou, H. Zhang, Y. Shi, S. Luo, N. Wang, H. Cai, L. Wan, B. Wang, X. Jiang, *Nat. Commun.* **2022**, *13*, 1044.
[9] G. Hinton, L.i Deng, D. Yu, G. Dahl, A.-R. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. Sainath, B. Kingsbury, *IEEE Signal Process Mag* **2012**, *29*, 82.
[10] A. Graves, A. Mohamed, G. Hinton, in *Speech Recognition with Deep Recurrent Neural Networks[C]//2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, IEEE, Vancouver, BC, Canada **2013**, pp. 6645–6649.
[11] Y. Shen, N. C. Harris, S. Skirlo, M. Prabhu, T. Baehr-Jones, M. Hochberg, X. Sun, S. Zhao, H. Larochelle, D. Englund, M. Soljačić, *Nat. Photonics* **2017**, *11*, 441.
[12] D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. Van Den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, S. Dieleman, D. Grewe, J. Nham, N. Kalchbrenner, I. Sutskever, T. Lillicrap, M. Leach, K. Kavukcuoglu, T. Graepel, D. Hassabis, *Nature* **2016**, *529*, 484.

[13] P. Baldi, *Annu. Rev. Biomed. Data Sci.* **2018**, *1*, 181.
[14] A. Esteva, K. Chou, S. Yeung, N. Naik, A. Madani, A. Mottaghi, Y. Liu, E. Topol, J. Dean, R. Socher, *NPJ Digit Med.* **2021**, *4*, 5.
[15] T. Saba, *Microsc Res. Tech.* **2021**, *84*, 1272.
[16] V. Rausch, A. Hansen, E. Solowjow, C. Liu, E. Kreuzer, J. K. Hedrick, in *Learning a Deep Neural Net Policy for End-to-End Control of Autonomous Vehicles[C]//2017 American Control Conference (ACC)*, IEEE, Seattle, WA, USA **2017**, pp. 4914–4919.
[17] M. K. K. Leung, H. Y. Xiong, L. J. Lee, B. J. Frey, *Bioinformatics* **2014**, *30*, i121.
[18] H. Y. Xiong, B. Alipanahi, L. J. Lee, H. Bretschneider, D. Merico, R. K. C. Yuen, Y. Hua, S. Gueroussov, H. S. Najafabadi, T. R. Hughes, Q. Morris, Y. Barash, A. R. Krainer, N. Jojic, S. W. Scherer, B. J. Blencowe, B. J. Frey, *Science* **2015**, *347*, 1254806.
[19] J. Peurifoy, Y. Shen, L.i Jing, Y.i Yang, F. Cano-Renteria, B. G. Delacy, J. D. Joannopoulos, M. Tegmark, M. Soljačić, *Sci. Adv.* **2018**, *4*, eaar4206.
[20] W. Ma, F. Cheng, Y. Liu, *ACS Nano* **2018**, *12*, 6326.
[21] I. Malkiel, M. Mrejen, A. Nagler, I. Malkiel, M. Mrejen, A. Nagler, U. Arieli, L. Wolf, and H. Suchowski, *Light: Sci. Appl.* **2018**, *7*, 60.
[22] Z. Liu, D. Zhu, S. P. Rodrigues, K.-T. Lee, W. Cai, *Nano Lett.* **2018**, *18*, 6570.
[23] N. Jones, *Nature* **2018**, *561*, 163.
[24] J. Feldmann, N. Youngblood, C. D. Wright, H. Bhaskaran, W. H. P. Pernice, *Nature* **2019**, *569*, 208.
[25] Y.e Tian, Y. Zhao, S. Liu, Q. Li, W. Wang, J. Feng, J. Guo, *Nanophotonics* **2022**, *11*, 329.
[26] X. Lin, Y. Rivenson, N. T. Yardimci, M. Veli, Y.i Luo, M. Jarrahi, A. Ozcan, *Science* **2018**, *361*, 1004.
[27] C. Ramey, in *Silicon Photonics for Artificial Intelligence Acceleration: Hotchips 32[C]//2020 IEEE Hot Chips 32 Symposium (HCS)*, IEEE, Palo Alto, CA, USA **2020**, pp. 1–26.
[28] X. Xiao, S. J. B. Yoo, *Scalable and Compact 3D Tensorized Photonic Neural Networks[C]//2021 Optical Fiber Communications Conference and Exhibition (OFC)*, IEEE, San Francisco, CA, USA **2021**, pp. 1–3.
[29] S. Abdollahramezani, O. Hemmatyar, H. Taghinejad, A. Krasnok, Y. Kiarashinejad, M. Zandehshahvar, A. Alù, A. Adibi, *Nanophotonics* **2020**, *9*, 1189.
[30] Q. Zhang, Y. Zhang, J. Li, R. Soref, T. Gu, J. Hu, *Opt. Lett.* **2018**, *43*, 94.
[31] C. Zhang, M. Zhang, Y. Xie, Y. Shi, R. Kumar, R. R. Panepucci, D. Dai, *Photonics Res.* **2020**, *8*, 1171.
[32] P. Xu, J. Zheng, J. K. Doylend, A. Majumdar, *ACS Photonics* **2019**, *6*, 553.
[33] D. Tanaka, Y. Shoji, M. Kuwahara, X. Wang, K. Kintaka, H. Kawashima, T. Toyosaki, Y. Ikuma, H. Tsuda, *Opt. Express* **2012**, *20*, 10283.
[34] Z. Zhang, J. Yang, W. Bai, Y. Han, X. He, J. Huang, D. Chen, S. Xu, W. Xie, *Appl. Opt.* **2019**, *58*, 7392.
[35] C. Ríos, M. Stegmaier, P. Hosseini, D.i Wang, T. Scherer, C. D. Wright, H. Bhaskaran, W. H. P. Pernice, *Nat. Photonics* **2015**, *9*, 725.
[36] C. Rios, P. Hosseini, C. D. Wright, H. Bhaskaran, W. H. P. Pernice, *Adv. Mater.* **2014**, *26*, 1372.
[37] B. Gholipour, J. Zhang, K. F. Macdonald, D. W. Hewak, N. I. Zheludev, *Adv. Mater.* **2013**, *25*, 3050.
[38] Q. Wang, E. T. F. Rogers, B. Gholipour, C.-M. Wang, G. Yuan, J. Teng, N. I. Zheludev, *Nat. Photonics* **2016**, *10*, 60.
[39] X. Yin, T. Steinle, L. Huang, T. Taubner, M. Wuttig, T. Zentgraf, H. Giessen, *Light: Sci. Appl.* **2017**, *6*, e17016.
[40] J. Zheng, Z. Fang, C. Wu, S. Zhu, P. Xu, J. K. Doylend, S. Deshmukh, E. Pop, S. Dunham, M.o Li, A. Majumdar, *Adv. Mater.* **2020**, *32*, 2001218.
[41] Y. Zhang, J. B. Chou, J. Li, H. Li, Q. Du, A. Yadav, S.i Zhou, M. Y. Shalaginov, Z. Fang, H. Zhong, C. Roberts, P. Robinson, B. Bohlin, C. Ríos, H. Lin, M. Kang, T. Gu, J. Warner, V. Liberman, K. Richardson, J. Hu, *Nat. Commun.* **2019**, *10*, 4279.
[42] S. Raoux, F. Xiong, M. Wuttig, E. Pop, *MRS Bull.* **2014**, *39*, 703.
[43] M. Delaney, I. Zeimpekis, D. Lawson, D. W. Hewak, O. L. Muskens, *Adv. Funct. Mater.* **2020**, *30*, 2002447.
[44] M. Delaney, I. Zeimpekis, H. Du, X. Yan, M. Banakar, D. J. Thomson, D. W. Hewak, O. L. Muskens, *Sci. Adv.* **2021**, *7*, eabg3500.
[45] O. Hemmatyar, S. Abdollahramezani, I. Zeimpekis, S. Lepeshov, A. Krasnok, A. I. Khan, K. M. Neilson, C. Teichrib, T. Brown, E. Pop, *arXiv. 2107*, 12159.
[46] Y. Shi, J. Ren, G. Chen, W. Liu, C. Jin, X. Guo, Y.u Yu, X. Zhang, *Nat. Commun.* **2022**, *13*, 6048.
[47] B. Bai, Q. Yang, H. Shu, L. Chang, F. Yang, B. Shen, Z. Tao, J. Wang, S. Xu, W. Xie, W. Zou, W. Hu, J. E. Bowers, X. Wang, *Nat. Commun.* **2023**, *14*, 66.
[48] J. Huang, H. Ma, D. Chen, H. Yuan, J. Zhang, Z. Li, J. Han, J. Wu, J. Yang, *Nanophotonics* **2021**, *10*, 1011.
[49] K. Wang, X. Ren, W. Chang, L. Lu, D. Liu, M. Zhang, *Photonics Res.* **2020**, *8*, 528.
[50] J. Huang, J. Yang, D. Chen, W. Bai, J. Han, Z. Zhang, J. Zhang, X. He, Y. Han, L. Liang, *Nanophotonics* **2020**, *9*, 159.
[51] H. Ma, J. Huang, K. Zhang, J. Yang, *Opt. Express* **2020**, *28*, 17010.
[52] Y. Wang, P. Landreman, D. Schoen, K. Okabe, A. Marshall, U. Celano, H.-S. P. Wong, J. Park, M. L. Brongersma, *Nat. Nanotechnol.* **2021**, *16*, 667.
[53] A. Pirovano, A. L. Lacaita, A. Benvenuti, F. Pellizzer, S. Hudgens, R. Bez, *2003 IEEE International Electron Devices Meeting (IEDM)*, IEEE, Washington, DC, USA **2003**.
[54] W. Zhang, R. Mazzarello, M. Wuttig, E. Ma, *Nat. Rev. Mater.* **2019**, *4*, 150.
[55] B. J. Shastri, A. N. Tait, T. Ferreira De Lima, W. H. P. Pernice, H. Bhaskaran, C. D. Wright, P. R. Prucnal, *Nat. Photonics* **2021**, *15*, 102.