









# Space-efficient optical computing with an integrated chip diffractive neural network

H. H. Zhu<sup>1</sup>, J. Zou<sup>1</sup>, H. Zhang<sup>1</sup>, Y. Z. Shi<sup>2</sup>, S. B. Luo<sup>1</sup>, N. Wang<sup>3</sup>, H. Cai<sup>3</sup>, L. X. Wan<sup>1</sup>, B. Wang<sup>1</sup>, X. D. Jiang<sup>1</sup> <sup>✉</sup>, J. Thompson<sup>4</sup> , X. S. Luo<sup>5</sup> , X. H. Zhou<sup>6</sup> <sup>✉</sup>, L. M. Xiao<sup>7</sup> <sup>✉</sup>, W. Huang<sup>8</sup>, L. Patrick<sup>5</sup>, M. Gu<sup>9</sup> <sup>✉</sup>, L. C. Kwek<sup>1,4</sup> <sup>✉</sup> & A. Q. Liu<sup>1</sup> <sup>✉</sup>

Large-scale, highly integrated and low-power-consuming hardware is becoming progressively more important for realizing optical neural networks (ONNs) capable of advanced optical computing. Traditional experimental implementations need  $N^2$  units such as Mach-Zehnder interferometers (MZIs) for an input dimension  $N$  to realize typical computing operations (convolutions and matrix multiplication), resulting in limited scalability and consuming excessive power. Here, we propose the integrated diffractive optical network for implementing parallel Fourier transforms, convolution operations and application-specific optical computing using two ultracompact diffractive cells (Fourier transform operation) and only  $N$  MZIs. The footprint and energy consumption scales linearly with the input data dimension, instead of the quadratic scaling in the traditional ONN framework. A ~10-fold reduction in both footprint and energy consumption, as well as equal high accuracy with previous MZI-based ONNs was experimentally achieved for computations performed on the *MNIST* and *Fashion-MNIST* datasets. The integrated diffractive optical network (IDNN) chip demonstrates a promising avenue towards scalable and low-power-consumption optical computational chips for optical-artificial-intelligence.

<sup>1</sup>Quantum Science and Engineering Centre (QSec), Nanyang Technological University, Singapore 639798, Singapore. <sup>2</sup>National Key Laboratory of Science and Technology on Micro/Nano Fabrication, Department of Micro/Nano Electronics, Shanghai Jiao Tong University, Shanghai 200240, China. <sup>3</sup>Institute of Microelectronics, A\*STAR (Agency for Science, Technology and Research), Singapore 138634, Singapore. <sup>4</sup>Centre for Quantum Technologies, National University of Singapore, Singapore 117543, Singapore. <sup>5</sup>Advanced Micro Foundry, 11 Science Park Road, 117685 Singapore, Singapore. <sup>6</sup>State Key Joint Laboratory of ESPC, Center for Sensor Technology of Environment and Health, School of Environment, Tsinghua University, Beijing 100084, China. <sup>7</sup>Shanghai Engineering Research Center of Ultra-Precision Optical Manufacturing, School of Information Science and Technology, Fudan University, Shanghai 200433, China. <sup>8</sup>Suzhou Institute of Nano-Tech and Nano-Bionics (SINANO), Chinese Academy of Sciences (CAS), Suzhou 215123, China. <sup>9</sup>Quantum Hub, School of Physical and Mathematical Science, Nanyang Technological University, 50 Nanyang Ave, 639798 Singapore, Singapore. ✉email: [exdjiang@ntu.edu.sg](mailto:exdjiang@ntu.edu.sg); [xhzhou@mail.tsinghua.edu.cn](mailto:xhzhou@mail.tsinghua.edu.cn); [liminxiao@fudan.edu.cn](mailto:liminxiao@fudan.edu.cn); [gumile@ntu.edu.sg](mailto:gumile@ntu.edu.sg); [cqtklc@nus.edu.sg](mailto:cqtklc@nus.edu.sg); [eaqliu@ntu.edu.sg](mailto:eaqliu@ntu.edu.sg)

Optical neural networks (ONNs) that exploit photonic hardware acceleration to compute complex matrix-vector multiplication, have the advantages of ultra-high bandwidth, high calculation speed, and high parallelism over electronic counterparts<sup>1,2</sup>. With the rapidly increasing complexity in data manipulation techniques and dataset size, highly integrated and scalable ONN hardware with ultracompact size and the reduced energy consumption is strongly desired to conquer the resource bottleneck in artificial neural networks. Due to the advantages of ultracompact size, high-density integration, and power-efficiency silicon photonic integrated circuits (PICs) are emerging as a promising candidate for establishing large and compact computing units predominantly required in optical-artificial-intelligence computers<sup>3–5</sup>. Moreover, the fabrication of silicon PICs is compatible with CMOS technology, thus current infrastructure can support mass production at a low cost<sup>6</sup>. Typical silicon PIC architectures to realize chip-scale ONNs<sup>7–15</sup> use cascades of multiple Mach–Zehnder interferometers (MZIs)<sup>7–13</sup> and microring resonator-based wavelength division multiplexing technology<sup>14,15</sup>. Though PICs have shown great potential in integrated optics, the space utilization (directional couplers and phase modulators)<sup>16</sup>, energy consumption (heaters to control phase), and the complex control circuits for reducing fluctuation of the resonance wavelength<sup>17,18</sup> restrict the development of a large-scale programmable photonic neural network. Additionally, phase-change materials such as  $\text{Ge}_2\text{Sb}_2\text{Te}_5$ <sup>5</sup>, inverse-designed metastructures<sup>19</sup>, and nanoscale neural medium<sup>20</sup> combined with CMOS-compatible integrated photonic platform, can further minimize the size of integrated photonic circuits. However, the number of units for these ONN architectures scales quadratically with the input data dimension<sup>5,7,14</sup> for both convolution operations and matrix multiplication, leading to quadratic scaling in the footprint and energy consumption.

Meanwhile, Fourier transforms and convolutions, which are the fundamental building blocks of ONN architectures, have been realized using various approaches including the spatial light modulators<sup>21–24</sup>, micro-lens array<sup>25–28</sup>, and holographic elements<sup>29–33</sup>. Fourier transforms and convolutions as well as applications in classification tasks have been proposed utilizing the optical diffraction masks in free space<sup>26,31</sup>. However, these free-space approaches require heavy auxiliary or modulation equipment which is space-consuming<sup>34,35</sup> and exceedingly challenging to program in real-time<sup>31,36,37</sup>, hindering their feasibilities towards the large-scale implementation and commercialization of photonic neural networks.

Here, we demonstrate the first scalable integrated diffractive neural network (IDNN) chip using silicon PICs, which is capable of performing the parallel Fourier transform and convolution operations. Due to the utilization of on-chip compact diffractive cells (slab waveguides), both the footprint and power consumption of the proposed architecture is reduced from quadratic scaling in the input data dimensions required for MZI-based ONN architectures<sup>7,10,13</sup> to linear scaling for the IDNN. This reduction in the resource scaling from quadratic to linear will have a profound impact on the realization of large-scale silicon-photonics computing circuits with current fabrication technologies<sup>38–40</sup>.

The parallel Fourier transform and convolution operations in our IDNN chip can be applied to classification tasks with Iris flower, Handwriting digit, and Fashion product datasets. The cases of 1D sequence and 2D digit image recognition demonstrate the enhanced performance of the convolution operations. Two typical cases of Iris flower classification with a one-layer neural network, and Handwriting digit and Fashion product classification with a two-layer neural network as well as the comparison with the fully connected neural network, are conducted to further

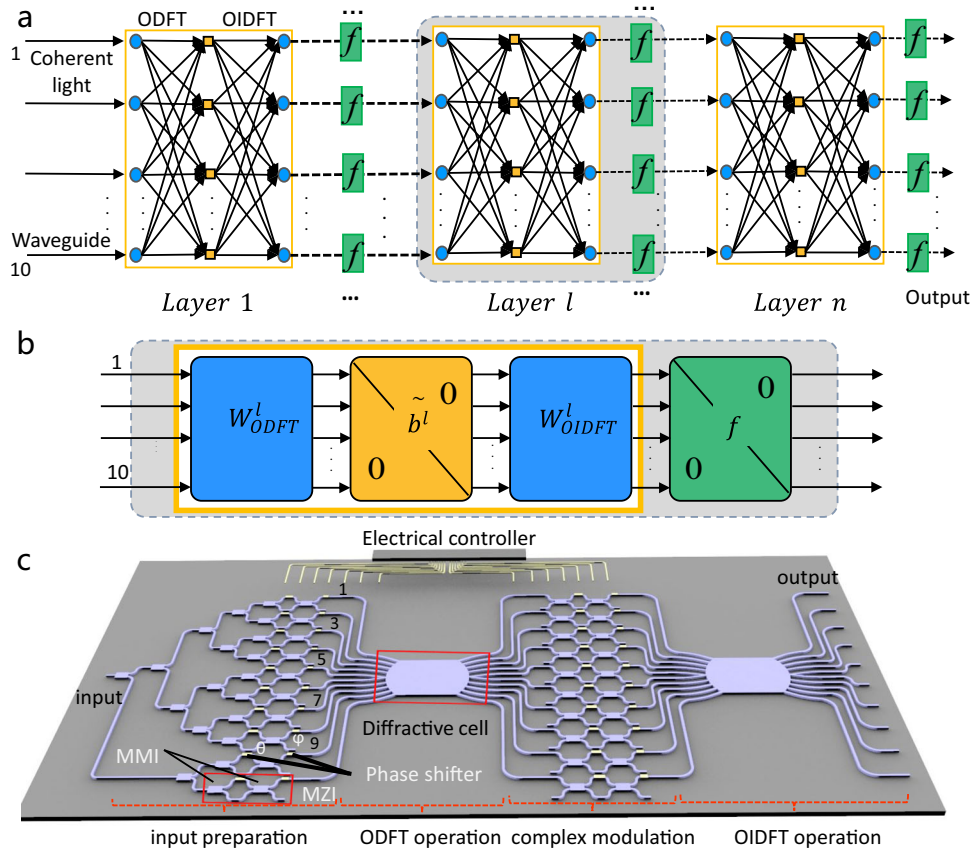
characterize the classification performance of our IDNN chip. For the *Iris* dataset classification, we obtain a chip testing accuracy of 98.3% in the one-layer IDNN using a complex modulation. For Handwriting and Fashion recognition tasks using 500 different sample inputs, the testing accuracies of 89.3 and 81.3% are obtained, respectively, which are equivalent to those obtained in a fully connected neural network. Compared to the previous design with an area of 5 mm<sup>2</sup> and power of 3–7 W<sup>7,10,13</sup>, our IDNN chip achieves the reduced hardware footprint (0.53 mm<sup>2</sup>) and low power consumption (17.5 mW), manifesting the advantage of the Fourier-based design as a scalable and power-efficient solution for data-heavy artificial intelligence applications.

## Results

**Design and fabrication.** A typical electronic one-layer neural network consists of an input layer, an output layer, connections between them (weight matrix), and elementwise nonlinearity (activation function). In a departure from the electronic neural network, our introduced IDNN framework implements a convolution transformation physically in the optical field using PICs. The convolution transformation is special matrix multiplication, and the complex-valued matrix elements are circulant<sup>41</sup>. The theoretical framework of the IDNN with multi-layers is shown in Fig. 1a. The preprocessed signals are encoded via modulating the amplitude and phase (two degrees of freedom) of coherent light and then input to the multi-layer IDNNs. In each layer as depicted in Fig. 1b, the information propagates through matrix multiplications followed by a nonlinear activation function. The linear part of each layer is composed of two ultracompact diffractive cells to implement optical discrete Fourier transform (ODFT) operation ( $W_{ODFT}^l$ ) and optical inverse discrete Fourier transform (OIDFT) operation ( $W_{OIDFT}^l$ ), and a complex-valued transmission modulation region in the Fourier domain behind the ODFT operation to achieve the Hadamard product. The nonlinear activation function  $f$  originates from the intensity detection of the complex outputs. We can hereby achieve a single-layer neural network. This IDNN can also be multi-layered by cascading multiple chips or recycling one single chip. The recycling process can be realized by configuring the phase shifters through a computer-controlled digital-to-analog converter.

Figure 1c shows the holistic diagram of the proposed silicon PICs-based IDNN chip, which consists of four major sections, e.g., input preparation, ODFT operation, complex modulation, and OIDFT operation. A continuous-wave laser (wavelength 1550 nm) is incoupled into the device through a grating coupler to realize the input source. The basic computing unit of a modulated MZI cell comprises multimode interferometers (MMI)-based beam splitters and two thermo-optic phase shifters in the “input preparation” section. The internal phase shifters between two beam splitters (indicated as  $\theta$ ) and external beam splitters (indicated as  $\phi$ ) are used to control the amplitude and phase of the input signals, respectively. All phase shifters are thermally tuned by the TiN heaters that are packaged to a printed circuit board (Fig. S5a). Then, they can be manipulated by an external electric circuit controller.

Subsequently, the ten input signals are fed into the diffractive cell, which is composed of a slab waveguide to implement the Fourier transform in the “ODFT operation” section (details will be shown in Fig. 2). After the operation, an array of MZIs is utilized to modulate the amplitude and phase of the ten signals after the ODFT operation. The modulated signals are then introduced to another diffractive cell to implement the OIDFT. After that, the ten output intensity signals are detected by an array of photodetectors. In addition, a thermoelectric controller is implemented as the substrate beneath the chip to control and



**Fig. 1 Optical integrated diffractive neural networks (IDNNs).** **a** The multi-layer neural networks. One layer contains three main parts: optical discrete Fourier transform (ODFT) operation, amplitude/phase modulation, and optical inverse discrete Fourier transform (OIDFT) operation. A nonlinear activation function is added between two layers. **b** IDNN operates on complex-valued inputs using coherent light. There are two matrices based on diffractive cells and a Hadamard product operation raised by phase and amplitude modulation behind the ODFT operation. **c** Schematics of the experimental device. The device includes four functional parts: (1) input signal preparation; (2) implementing ODFT operation; (3) modulating amplitude/phase in the Fourier domain; (4) implementing OIDFT operation.

stabilize the temperature (details in Supplementary Note 2). The holistic 3.2 mm × 2 mm chip monolithically integrates 10 modes, 20 MZIs, 2 slab waveguides, and 40 thermo-optic phase shifters.

Figure 2a, b shows the optical micrographs of the whole chip and the diffractive cell, respectively. The Fourier transform, as a core operation of our network, can be achieved using a slab waveguide-based ultracompact diffraction cell, whose input waveguides are coded by different phase shifters incorporated in the “input preparation” section. The normalized electric field after the ODFT operation  $[E_i(k)]$  (Supplementary Note 3) can be expressed as

$$E_i(k) = \frac{1}{j\sqrt{\lambda_0 R/n_s}} \int E_0(x) \exp(-j2\pi f_k x) dx \quad (1)$$

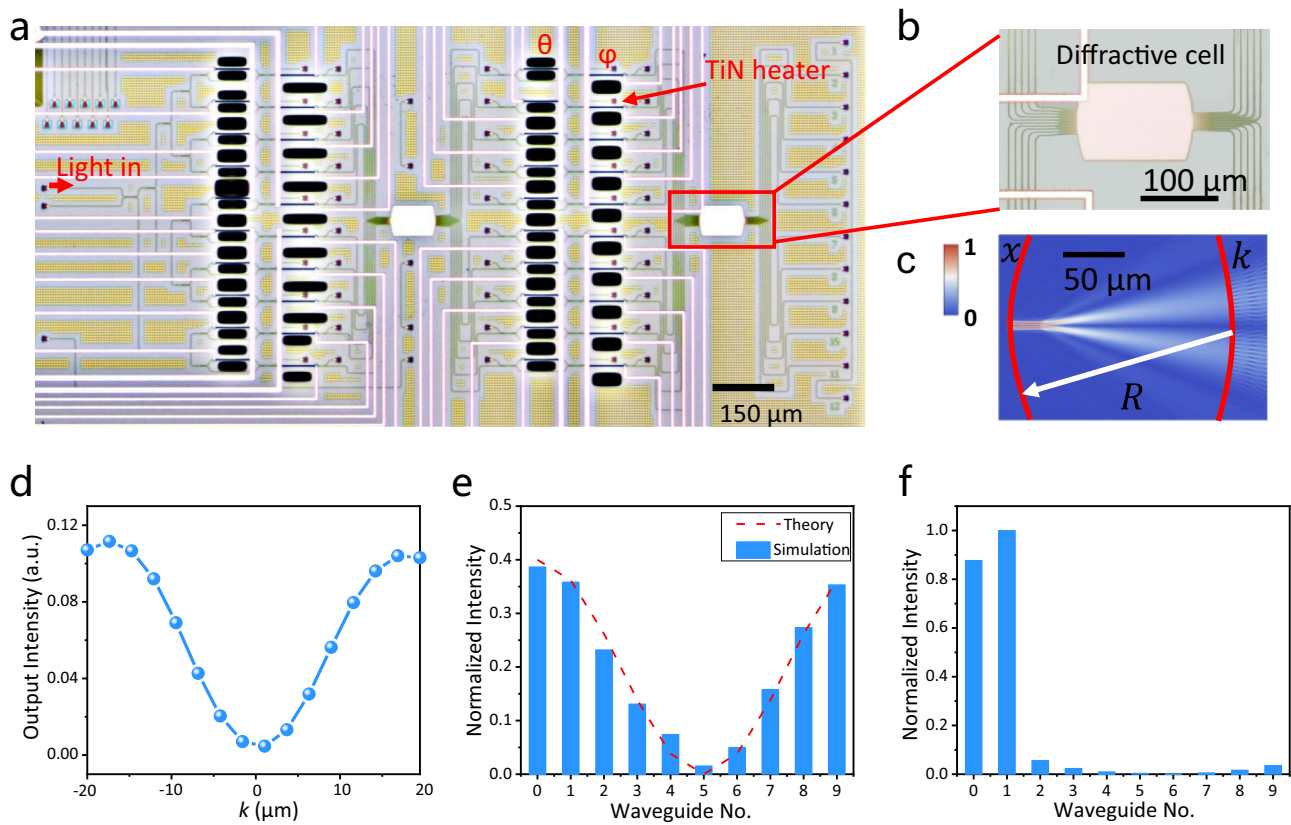
where  $E_0$  is the input electric field,  $f_k = kn_s/(\lambda_0 R)$  is the angle spectrum of the incident light field,  $\lambda_0$  is the wavelength of light in vacuum,  $R$  is the arc radius of the diffraction cell as shown in Fig. 2c,  $n_s$  is the effective refractive index of the waveguide, and  $k$  is the axis of output coordinate plane.

The simulated electric field distribution in the diffraction region and the output intensity distribution is shown in Fig. 2c, d, respectively. The input light is centered into two central waveguides along the propagation direction and the outputs are the electric field profile after the ODFT operation. The detailed analysis for light propagating in the entire chip is shown in Supplementary Note 3. The measured normalized intensities of ODFT operation in the diffractive cell and the retrieved signal

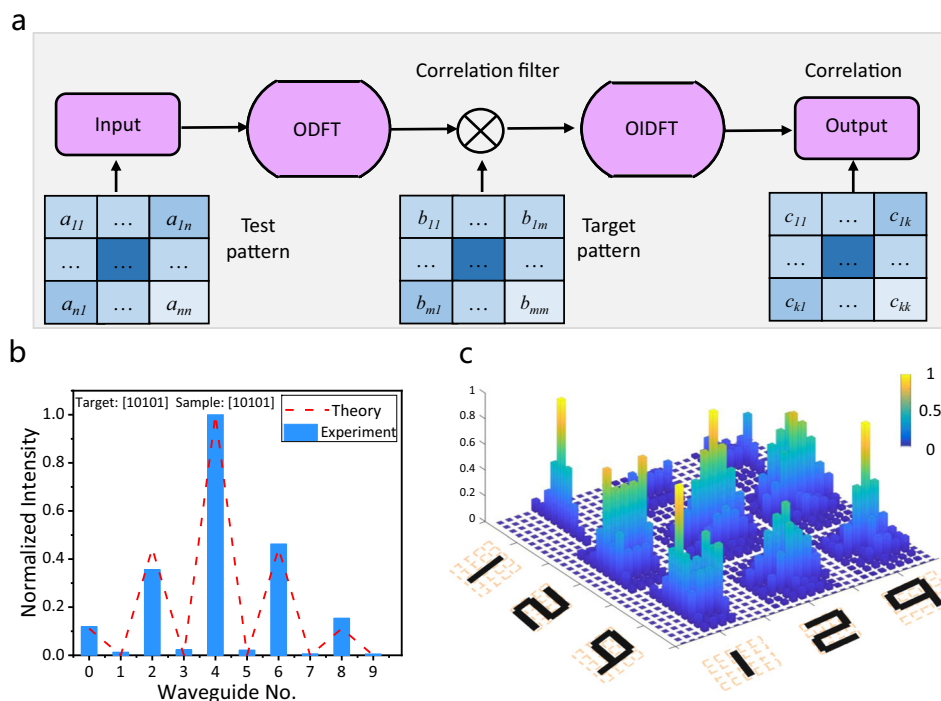
from all output channels after OIDFT operation are shown in Fig. 2e, f, respectively. The input signal can be well retrieved after ODFT and OIDFT operations are applied with no intervening phase modulation as seen from Fig. 2f (more experimental results can be found in Fig. S7).

**Image recognition with correlation algorithm.** To demonstrate the performance of our IDNN chip and validate the design for realizing the convolution matrix, we first use the chip as a convolution operator to implement image recognition using the correlation algorithm. The calibration for the initial amplitude and phase of the signal, the realization of ODFT and OIDFT operations, and the implementation of weight matrices are shown in Supplementary Note 4. The detailed mathematical elucidation of correlation algorithms can be found in Supplementary Note 5.

The 1D and 2D convolution operations have various applications in machine learning, such as convolution layers in convolutional neural networks<sup>41–43</sup> and correlation recognition of human faces from still images or video frames<sup>44,45</sup>. Here, we demonstrate the cases of 1D sequence and 2D image recognitions using a classical correlation algorithm<sup>44</sup>, which demonstrates the capability of the convolution operation realized by our IDNN chip. The image recognition process is shown in Fig. 3a. The correlation calculation of the input test image ( $n \times n$ ) and target image ( $m \times m$ ) is conducted using the Hadamard product between the Fourier transform of the input test image and the Fourier transform of a target image. Then the height of that peak



**Fig. 2 Optical microscope photos and some simulation results.** **a** The optical micrograph of the whole chip. **b** The optical micrograph of the diffractive cell. **c** Simulated electric field distribution in the slab waveguide region. **d** Simulated output amplitude distribution along the output waveguide plane. **e** Simulated normalized intensity of ten channels after ODFT operation. **f** Simulated normalized retrieved intensity information of ten channels.



**Fig. 3 Image recognition.** **a** Schematics of the image recognition process. The input test image dimension is  $n \times n$ , target image dimension is  $m \times m$  and the output dimension is  $k = n + m - 1$ . **b** Experimental normalized intensity from the output waveguides when the sample sequence [10101] is the same as the target sequence [10101]. **c** Experimental correlation results for digit image recognition. There are three-digit images (1, 2, 9), and each digit image is composed of  $5 \times 5$  matrix.

(relative to the background) of the output result is used to determine whether the test image matches the target image.

The experimental recognized sequence ([10101]) is the same as the target sequence as shown in Fig. 3b (More results in Fig. S8). Furthermore, we extend the 1D sequence to 2D image recognition of digits ( $5 \times 5$  matrices) with the same correlation algorithm. The correlation signals for nine different digit-image combinations are shown in Fig. 3c (details in Supplemental Note 6). Strong correlation peaks are observed along the diagonal line of the measured cross-correlation matrix, reflecting the intuition that two images have maximum correlation and similarity when they are the same. Our approach to convolutional processing provides an effective method for accelerating computation via reducing computational complexity, as compared with traditional methods<sup>43,41</sup>. The traditional electronic single convolution operation involves  $(n - m + 1)^2$  matrix-vector multiplication (MVM) operations with  $n \times n$  and  $m \times m$  dimension matrices<sup>43,41</sup>. Similarly, the state-of-the-art research works for photonic convolutional accelerators<sup>5,35</sup> also perform convolution operation by MVM operations, but in combination with parallel operations using WDM technology to accelerate the computing. Here, we realize an optical Fourier transform-based convolution. Each 2D convolution with  $n \times n$  and  $m \times m$  dimension matrices needs  $(n - m + 1) \times m$  1D convolution operations<sup>46</sup>. Since a 1D convolution is converted into an MVM operation in our design, our approach only needs  $(n - m + 1) \times m$  MVM operations to compute a 2D convolution. This design can thus accelerate the convolution operations to achieve improved scaling over traditional electronic counterparts<sup>43,41</sup>. More discussions on the computational convolution approach can be found in Supplementary Note 6. It can further be extended to circulant convolution operation<sup>45</sup> and employed in typical classification tasks<sup>47–49</sup>.

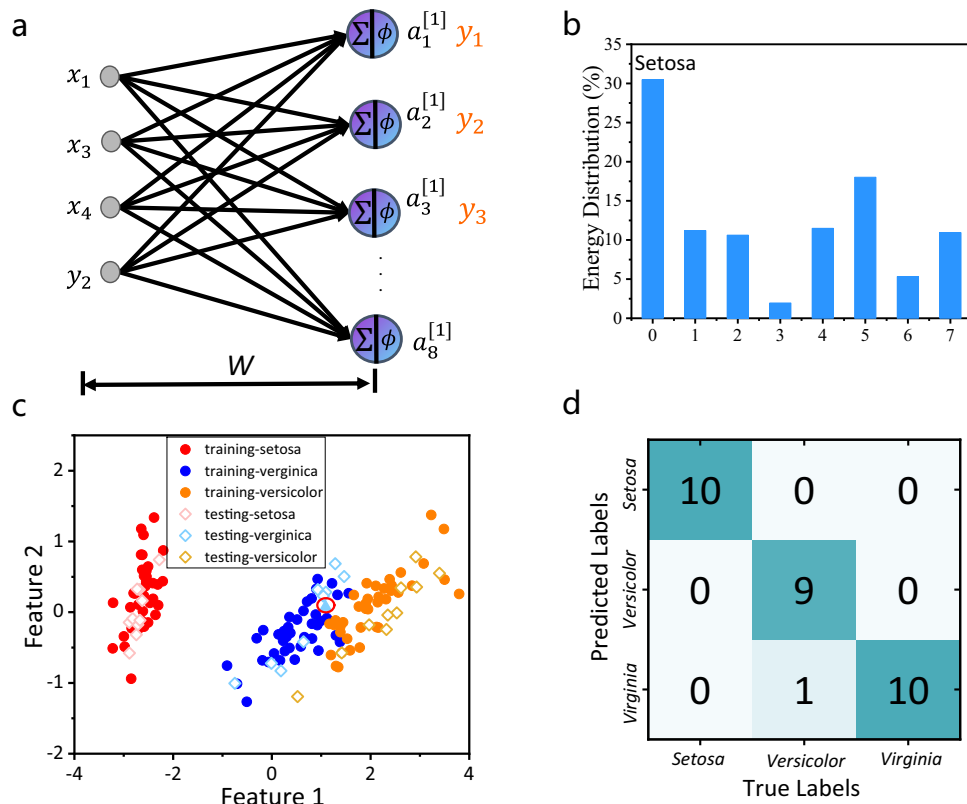
**Iris flower classifier.** We demonstrate the performance of our network in classifying the Iris flower dataset. Here we have four input parameters (the lengths and widths of the sepals and petals of a candidate flower). The task is to determine which of the three possible subspecies the flower belongs to (*setosa*, *versicolor*, and *virginica*). We implement this classification task using a one-layer neural network with eight neurons as shown in Fig. 4a. The input and output are connected by a complex, circulant weight matrix ( $W$ ) with eight trained complex-valued parameters. Input signals are encoded by modulating the amplitude of the input field of the IDNN. Amplitude and phase modulators in the Fourier domain are trained to map inputs into three output waveguides, which are detected by external power sensors. The highest intensity measurement outcome is used to indicate the subspecies.

The entire dataset with 150 instances is split into the training set and testing set according to a ratio of 8:2. The weights are trained only on the training set. In the training process, the amplitude and phase of each channel in the Fourier domain are learnable parameters, representing a complex-valued modulation on the neural network. Trained with the error back-propagation algorithm<sup>16</sup>, the numerical convergence of the accuracy and loss versus epoch number are depicted in Fig. S10. The training accuracy of 98.0% is achieved in the complex modulation, while the training accuracy is 97.3% in the conventional one-layer fully connected neural network (Fig. S10).

After training, the performance of the IDNN chip with complex modulation is evaluated on the Iris flower classification testing dataset, see Fig. 4. The experimentally obtained energy distribution from the eight output waveguides is shown in Fig. 4b for one testing data point as an example. The channel “0” has the highest output intensity, indicating that this flower is classified as “Setosa”. More experimental results can be found in Fig. S10. The

intensity distribution from the output detectors reveals that the IDNN can achieve a maximum signal at the corresponding detector channel identified with candidate flower species. About 150 samples, including the training set and testing set, are extracted into two features and visualized as a scatter plot in Fig. 4c. A total of 30 testing samples are experimentally evaluated using the IDNN chip with only one wrong classification as marked by the red circle in Fig. 4c. The one-layer IDNN chip can achieve a high classification accuracy of 96.7% over 30 testing images in the experiment (Fig. 4d), which is the same as the testing result (96.7%) of the conventional one-layer fully connected neural network, manifesting that our IDNN has comparable accuracy with a fully connected network. However, our design has fewer components (i.e., the number of MZIs is reduced from 16 to 8) and more efficient space overhead compared with the fully connected architectures.

**Handwritten digit and Fashion product classifier.** More complicated datasets, *MNIST* (handwrite digits images), and *Fashion-MNIST* (clothing images) datasets are used to further validate the functionality of our IDNN chip. The two datasets are both split into the training (60,000 images) and testing sets (10,000 images). Our model is trained on the entire training set, and 500 instances are uniformly and randomly drawn from the testing set to validate the trained model on-chip<sup>35</sup>. Figure 5a shows that the network is composed of two layers, a hidden layer  $W$  with  $16 \times 10$  trained complex parameters and an output layer  $W^{out}$  with ten trained complex parameters. The outputs of the output layer are the recognition results (The computing details are shown in Supplementary Note 8). The numerical testing accuracy and loss vary with the epoch number as shown in Fig. 5b with a classification accuracy of 92.5% (see Fig. S11). Besides, the simulated energy distribution from the intensity detection results of ten outputs is shown in Fig. S11f to display the classification performance, indicating that our numerical simulations predict the channel with the highest energy will generally correspond to the correct handwritten digit. We experimentally tested 500 images and the confusion matrix (Fig. 5c) shows an accuracy of 91.4% in the generated predictions, in contrast to 92.6% in the numerical simulation. The output energy distributions from 10 waveguides for the classification of the digit “2” are depicted in Fig. 5d (More experimental results are shown in Fig. S12). These results show that we have successfully implemented the classification task on the integrated diffractive optical computing platform. Our experimental testing accuracy is slightly lower (91.4%) than the numerically predicted value obtained from simulations (92.6%), which is attributed to the errors from the diffractive region, heater calibration, and other experimental factors (Supplementary Note 10). Although our method does not achieve as high accuracy as previous reservoir computing demonstrations<sup>50,51</sup>, their internal dynamics of reservoirs are uncontrollable with fixed nodes, which leads to the stringent requirement of obtaining many parameters with high accuracy. Consequently, they are difficult to be implemented on chip. In contrast, our approach requires fewer parameters in controllable weight matrices that significantly reduce the footprint and energy consumption. Meanwhile, the input data compression process in our method also loses some information and lowers the numerical classification accuracy. The numerical blind testing accuracy (92.5%) with 10,000 samples can be further improved after enlarging the size of the network with more input information. It is worth noting that the classification accuracy obtained with our approach is comparable to the fully connected network (Fig. 5b), but our method consumes less energy and occupies less footprint, representing the major advantages of the design.



**Fig. 4** Iris flower classification using the diffractive neural network. **a** Network consists of one layer. The outputs of the layer are the recognition results after the intensity detection.  $x_1$  is sepal length,  $x_2$  is sepal width,  $x_3$  is petal length, and  $x_4$  is petal width.  $y_1$  is *Setosa*,  $y_2$  is *Versicolor*, and  $y_3$  is *Verginica*. **b** Experimental output intensity distribution of the IDNN for a class of flowers as “*Setosa*” is demonstrated. **c** Training and testing results of classification. One classification error appears on the testing part. **d** Confusion matrix in our experiment using 30 different flowers.

We further use the IDNN to perform the classification of the *Fashion-MNIST* dataset. In general, we can encode the preprocessed signals by using the amplitude and phase channels of coherent light. In the case of the *MNIST* dataset classifier, the output of the first hidden layer is encoded to the amplitude channels of the output layer. To verify that the network also performs well when input signals are encoded into the phase channel, the signal from the input layer of the fashion product images is encoded in the phase channel in our experiment. The numerical testing results for the classification of fashion products are depicted in Fig. S13. The corresponding accuracy for 500 testing images with one hidden layer is 82.0% in simulation versus 80.2% in an experiment (Fig. 5f). The experimental energy distributions from ten waveguide outputs using one image of a “pullover”, is shown in Fig. 5g and the experimental results from other images are depicted in Fig. S14. We implement additional experiments using the phase channel encoding input signal on the *MNIST* dataset and using the amplitude channel encoding input signal on the *Fashion-MNIST* dataset. The confusion matrix for 500 images (Fig. S15a) shows an experimental accuracy of 89.4%, in contrast to 92.6% for the numerical simulation results on the *MNIST* dataset. The results for the *Fashion-MNIST* dataset are 81.4% according to the numerical simulations versus 80.4% for the experiment (Fig. S15b). We note that the accuracy error between the experimental and numerical results has no distinct difference between using the amplitude channel and the phase channel. However, for phase encoding, the intensity of the input signal would not be attenuated, which leads to a higher output intensity in our experiment.

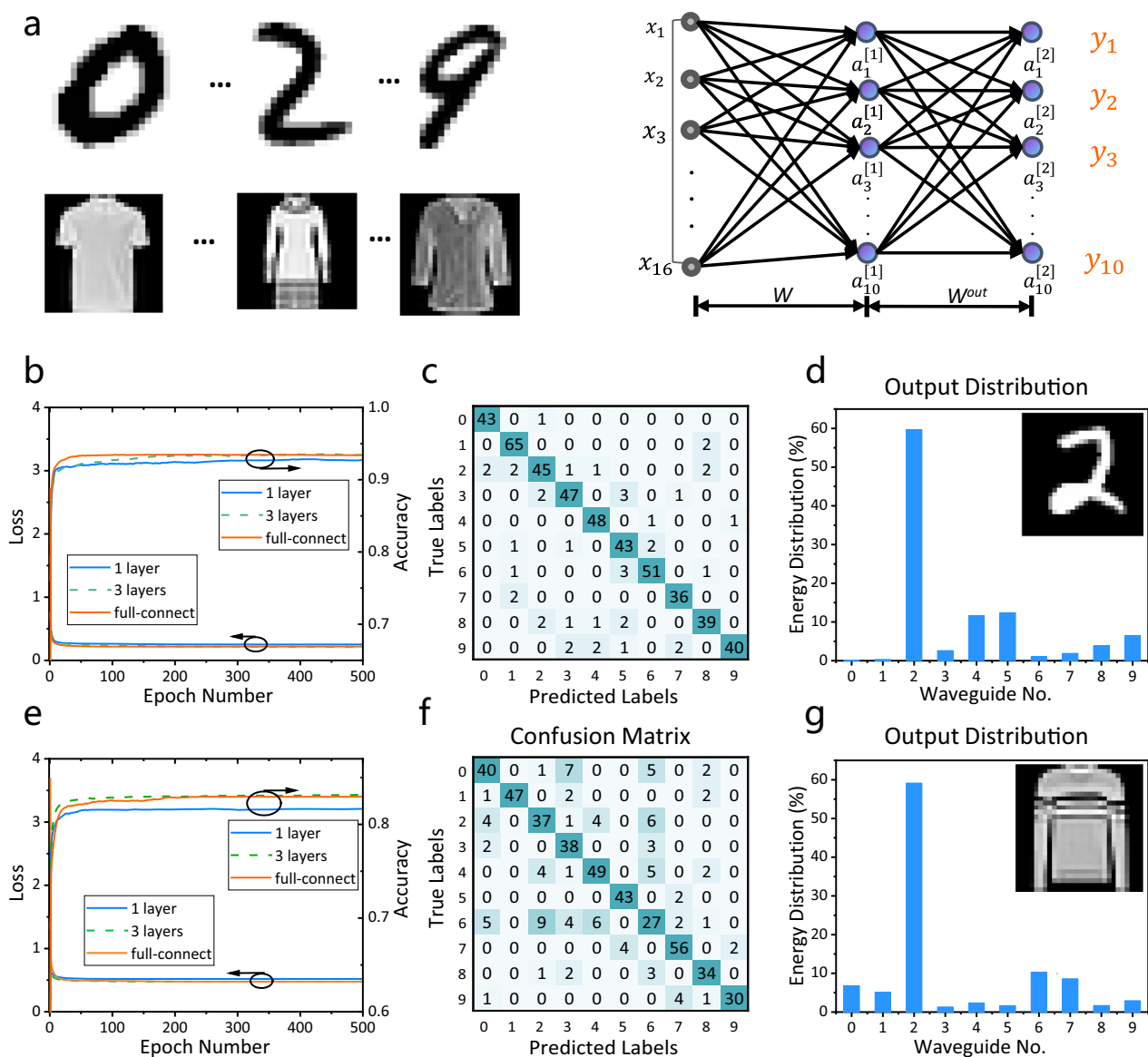
We compare our IDNN architecture with a fully connected neural network with the training results of the *MNIST* dataset and *Fashion-MNIST* dataset as shown in Fig. 5b, e, respectively. The

simulation model is built in Pytorch with a learning rate of 0.0001, a training period of 500 iterations, and a batch size of 100. Our complex circulant matrix-based neural network with one hidden layer achieves a numerical testing accuracy of 92.5%, which is comparable to that obtained in both a fully connected neural network with one hidden layer (93.4%) and a circulant matrix-based neural network with three-layers (93.5%) for the *MNIST* dataset. For the *Fashion-MNIST* dataset, both the simulated testing accuracies of the complex circulant matrix-based neural network with one hidden layer and three layers are higher than 81% (81.7 and 83.2%, respectively), which is comparable with 83.0% for fully connected based algorithm. For the *MNIST* and *Fashion-MNIST* datasets, the classification accuracy of our IDNN chip with less trainable parameters can be comparable with the traditional fully connected network.

Our IDNN chip can achieve comparable testing accuracy with classical fully connected neural networks in datasets *MNIST* and *Fashion-MNIST*. Some key metrics evaluating the physical network are listed by the comparison with the traditional fully connected MZI-based<sup>7,10,13</sup> architectures for the optical realization, as shown in Table 1. For the *MNIST* and *Fashion-MNIST* datasets, a  $10 \times 10$  matrix on-chip is achieved with only 0.53 mm<sup>2</sup> area and 17.5 mW power consumption to maintain the phase of heaters, while the traditional MZI-based ONN architectures need an area around 5 mm<sup>2</sup> and requisite energy consumption of 3–7 W. (More details are in Supplementary Notes 1).

## Discussion

We propose and experimentally demonstrate the reconfigurable IDNN chip, leveraging diffractive optics to realize an efficient



**Fig. 5 Handwriting and fashion recognition using the IDNN.** **a** The network consists of a hidden layer  $W$  and an output layer  $W^{out}$ . The outputs of the output layer are recognition results. The input layer is calculated by the traditional computer and the complex output results of the input layer are converted into amplitude and phase information as the input of the chip. **b** The numerical testing results of accuracy and loss versus epoch number for the *MNIST* dataset. **c** The confusion matrix for our experimental results, using 500 different handwritten digits. **d** The output intensity distribution of the IDNN for a handwritten input of “2” is demonstrated. **e** The numerical testing results of accuracy and loss versus epoch number for the *MNIST-Fashion* dataset. **f** The confusion matrix in the experiment. **g** As an example, the output intensity distribution of the IDNN for a fashion product input of “pullover” is demonstrated.

**Table 1 The performance of our proposed IDNN framework.**

Type	The no. modulators	Matrix dimension	Area ( $10 \times 10$ neurons)	Power consumption ( $10 \times 10$ neurons)
MZI-VMM <sup>10</sup>	16	$4 \times 4$	$5.2 \text{ mm}^2$	-----
MZI-VMM <sup>7</sup>	9	$4 \times 4$	$6 \text{ mm}^2$	3 W
MZI-VMM <sup>13</sup>	15	$4 \times 4$	$4.8 \text{ mm}^2$	7 W
Our work	10	$10 \times 10$	$0.53 \text{ mm}^2$	17.5 mW

matrix-vector multiplication. Previous MZI-based integrated photonic approaches for computing have been predominantly restricted by large footprints (around  $0.5 \text{ mm}^2$  per MZI unit<sup>23,26,29</sup>) and the excessive energy consumption to tune the phase, e.g., one needs  $\sim 50 \text{ mW}$  to tune each thermo-optic heater ( $2\pi$  phase shift). Our approach can effectively decrease footprints and energy

consumption (see Table 1) by reducing the numbers of MZI from quadratic scaling with input data size to linear. The key strategy of the scaling reduction is using integrated ultracompact diffractive cells (slab waveguides) to replace MZI units and realize Fourier transform operations. The slab waveguide is a passive device with a small size ( $0.15 \text{ mm}^2$ ) and does not require electric tuning.

Our IDNN chip shows an equivalent classification capability while cutting down the number of optical basic components (i.e., MZI modulators) to 10% of original on *MNIST* and *Fashion-MNIST* classification tasks and a half of original on the *Iris* datasets, compared with the previous MZI-based ONN architectures. We achieved a  $10 \times 10$  matrix on-chip with 10 MZIs and two diffractive cells, which has an area of  $0.53 \text{ mm}^2$ , while 100 MZIs are required with an area around  $5 \text{ mm}^2$  for a fully connected network. Meanwhile, the power consumption used to maintain the phase of modulators, which is also the dominant source of energy consumption, decreases from 3–7 W (previous techniques) to 17.5 mW (see Fig. S1). Using the compact design method, the matrix size can easily be scaled up to  $64 \times 64$  with acceptable loss (see Supplementary Note 1). Considering the current fabrication level of the integrated photonic chip, reduction in resource scaling from quadratic to linear and the relevant footprint and energy consumption reduction are profoundly meaningful towards the goal of a large-scale programmable photonic neural network and the achievement of photonic AI computing. It is worth mentioning that the IDNN focuses on implementing the convolution matrix with the advantages of scalability and low power consumption, instead of an arbitrary matrix-vector multiplication modulation. Since a convolution matrix contains fewer free parameters than those in an arbitrary matrix, the current demonstrated one-layer network may not be an optimal solution for complicated classification tasks. This issue can be mitigated by implementing multi-layers or multiple convolutions in each layer to obtain a stronger expression capability. The speed of IDNN is also limited by the electrical equipment and can be accelerated when the modulators and detectors are high-speed programmed.

Our integrated chip is adaptable to the Fourier-based convolution acceleration algorithm and multiple tasks, which has great potentials for many applications in compact and scalable application-specific optical computing, future optical-artificial-intelligence computers, and quantum information processing, such as image analysis, object classification, and live video processing in autonomous driving<sup>5,42</sup>. The footprint and energy consumption scale linearly with the input data dimension, which will hugely reduce the resources used for future sophisticated and data-heavy computing, and facilitate the prospects of a power-efficient, ultracompact, and large-scale integrated optical computing chip.

## Methods

**Fabrication.** The whole optical diffractive neural network is fabricated on the silicon-on-insulator (SOI) platform with a 220-nm thick silicon top layer and a 2- $\mu\text{m}$  thick buried oxide. Subsequently, a thin layer of titanium nitride (TiN) is deposited as the resistive layer for heaters. A thin aluminum film is patterned as the electrical connection to the heaters and photodetectors. Isolation trenches are created by etching the  $\text{SiO}_2$  top cladding and Si substrate.

**ODFT operation.** The critical part of the IDNN chip is the ODFT/OIDFT operations composed of the diffractive regions with a suitable phase difference from different waveguides before the diffractive components. The phase shift,  $\phi_{n-k}$ , between input  $k$  and output  $n$  can be designed to satisfy the relation  $\phi_{n-k} = \frac{2\pi}{N}(k - \frac{N-1}{2})(n - \frac{N-1}{2})$  by setting positions of input and output waveguides. In addition, the phase offset  $\varphi_k = \frac{\pi}{N}(N-1)k$  is additionally added to the input  $k$  using length adjustment or a phase shift. Then the phase difference  $\Delta\phi_n$  between two paths to the output  $n$  originating from the inputs,  $k+1$  and  $k$  is derived as  $\Delta\phi_n = (\phi_{n-(k+1)} + \varphi_{k+1}) - (\phi_{n-k} + \varphi_k) = \frac{2\pi n}{N}$ . This allows the ODFT computation to be realized.

**Circulant matrix.** For the linear part of the IDNN, a complex circulant matrix multiplication is achieved utilizing ODFT and OIDFT operations. Assuming each circulant matrix  $B$  is defined by a vector  $b$ , which is the first-row vector of  $B$ . Based on the theories of circulant convolution<sup>52–54</sup> and diffractive optics, the matrix-vector multiply can be efficiently performed using Fourier transform, which is expressed as  $y^{j+1} = W_{\text{ODFT}}^l [(W_{\text{ODFT}}^l \circ y^j) \circ b^j] = F^{-1}[F(y^j) \circ F(b^j)]$ .  $W_{\text{ODFT}}^l$  is the

matrix for ODFT,  $W_{\text{OIDFT}}^l$  is the matrix for OIDFT,  $b^j$  is a vector of the circulant matrix  $B$  of the  $l$ -layer neural network,  $y^j$  is the input vector of the  $l$ -layer neural network.  $F(\circ)$  represents an  $n$ -point ODFT operation,  $F^{-1}(\circ)$  is the inverse of  $F(\circ)$  (OIDFT operation),  $\circ$  represents the Hadamard product and  $b^j$  is the Fourier transform of  $b^j$ .

**Numerical simulation.** The training process of Iris flower, Handwritten digit, and Fashion product datasets was conducted in Pytorch (Google Inc.), a package for Python. The training process of Iris flower, Handwritten digit, and Fashion product datasets was conducted in Pytorch (Google Inc.), a package of Python. The learned complex parameter ( $w$ ) is expressed as  $w = a + ib = Ae^{i\alpha}$  where  $A$  is encoded into the internal phase shifters ( $\theta$ ) with  $\theta = \arcsin(A^2)$  and  $\alpha$  is encoded into the external phase shifters ( $\phi$ ) with  $\phi = \alpha - \theta/2$ . The ODFT and OIDFT operations are implemented using the angular spectrum method. A three-dimensional finite-difference-time-domain method is used to simulate the optical field distribution and transmission spectra of the diffractive cells.

## Data availability

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

Received: 5 May 2021; Accepted: 20 January 2022;

Published online: 24 February 2022

## References

- Kitayama, K. et al. Novel frontier of photonics for data processing—Photonic accelerator. *APL Photonics* **4**, 090901 (2019).
- Caulfield, H. J., Kinser, J. & Rogers, S. K. Optical neural networks. *Proc. IEEE* **77**, 1573–1583 (1989).
- O'Brien, J. L., Furusawa, A. & Vuckovic, J. Photonic quantum technologies. *Nat. Photon* **3**, 687–695 (2009).
- Kues, M. et al. On-chip generation of high-dimensional entangled quantum states and their coherent control. *Nature* **546**, 622–626 (2017).
- Feldmann, J. & Youngblood, N. et al. Parallel convolutional processing using an integrated photonic tensor core. *Nature* **589**, 52–58 (2021).
- Dong, P., Chen, Y. K., Duan, G. H. & Neilson, D. T. Silicon photonic devices and integrated circuits. *Nanophotonics* **3**, 215–228 (2014).
- Shen, Yichen et al. Deep learning with coherent nanophotonic circuits. *Nat. Photonics* **11**, 441–446 (2017).
- Reck, M. et al. Experimental realization of any discrete unitary operator. *Phys. Rev. Lett.* **73**, 58–61 (1994).
- Harris, N. C. et al. Linear programmable nanophotonic processors. *Optica* **5**, 1623–1631 (2018).
- Ribeiro, A. et al. Demonstration of a 4x4-port universal linear circuit. *Optica* **3**, 1348–1357 (2016).
- Bogaerts, W. et al. Programmable photonic circuits. *Nature* **586**, 207–216 (2020).
- Clements, W. R. et al. Optimal design for universal multiport interferometers. *Optica* **3**, 1460–1465 (2016).
- Zhang, H. et al. An optical neural chip for implementing complex-valued neural network. *Nat. Commun.* **12**, 1–11 (2021).
- Tait, A. N. et al. Neuromorphic silicon photonic networks. *Sci. Rep.* **7**, 7430 (2016).
- Feldmann, J. et al. All-optical spiking neurosynaptic networks with selflearning capabilities. *Nature* **569**, 208–214 (2019).
- Wetzstein, G. et al. Inference in artificial intelligence with deep optics and photonics. *Nature* **588**, 39–47 (2020).
- Zhang, Y. et al. Temperature sensor with enhanced sensitivity based on silicon Mach-Zehnder interferometer with waveguide group index engineering. *Opt. Express* **26**, 26057–26064 (2018).
- Carolan, J. et al. Scalable feedback control of single photon sources for photonic quantum technologies. *Optica* **6**, 335–340 (2019).
- Estakhri, N. M., Edwards, B. & Engheta, N. Inverse-designed metastructures that solve equations. *Science* **363**, 1333–1338 (2019).
- Khoram, E. et al. Nanophotonic media for artificial neural inference. *Photonics Res.* **7**, 823–827 (2019).
- Armitage, D. et al. High-speed spatial light modulator. *IEEE J. Quant. Electron* **21**, 1241–1248 (1985).
- Li, S.-Q. et al. Phase-only transmissive spatial light modulator based on tunable dielectric metasurface. *Science* **364**, 1087–1090 (2019).
- Kagalwala, K. H., Di Giuseppe, G., Abouraddy, A. F. & Saleh, B. E. A. Single-photon three-qubit quantum logic using spatial light modulators. *Nat. Commun.* **8**, 739 (2017).



24. Zuo, Y. et al. All-optical neural network with nonlinear activation functions. *Optica* **6**, 1132–1137 (2019).
25. Lu, T., Wu, S., Xu, X. & Francis, T. Two-dimensional programmable optical neural network. *Appl. Opt.* **28**, 4908–4913 (1989).
26. Chang, J. L. et al. Hybrid optical-electronic convolutional neural networks with optimized diffractive optics for image classification. *Sci. Rep.* **8**, 12324 (2018).
27. Yuan, W. et al. Fabrication of microlens array and its application: a review. *Chin. J. Mech. Eng.* **31**, 1–9 (2018).
28. Luo, Y. et al. Design of task-specific optical systems using broadband diffractive neural networks. *Light Sci. Appl.* **8**, 112 (2019).
29. Psaltis, D., Brady, D., Gu, X. G. & Lin, S. Holography in artificial neural networks. *Nature* **343**, 325–330 (1990).
30. Yeh, S. L. et al. Optical implementation of the Hopfield neural network with matrix gratings. *Appl. Opt.* **43**, 858–865 (2004).
31. Lin, X. et al. All-optical machine learning using diffractive deep neural networks. *Science* **361**, 1004–1008 (2018).
32. Mengu, D. et al. Analysis of diffractive optical neural networks and their integration with electronic neural networks. *IEEE J. Sel. Top. Quant. Electron* **26**, 1–14 (2020).
33. Fan, F. et al. Holographic element-based effective perspective image segmentation and mosaicking holographic stereogram printing. *Appl. Sci.* **9**, 920 (2019).
34. Hamerly, R., Bernstein, L., Sludds, A., Soljačić, M. & Englund, D. Large-scale optical neural networks based on photoelectric multiplication. *Phys. Rev. X* **9**, 021032 (2019).
35. Xingyuan, X. et al. 11 TOPS photonic convolutional accelerator for optical neural networks. *Nature* **589**, 44–51 (2021).
36. Silva, A. et al. Performing mathematical operations with metamaterials. *Science* **343**, 160–163 (2014).
37. Colburn, S. et al. Optical frontend for a convolutional neural network. *Appl. Opt.* **58**, 3179–3186 (2019).
38. Wang, J. et al. Multidimensional quantum entanglement with large-scale integrated optics. *Science* **360**, 285–291 (2018).
39. Sun, J., Timurdogan, E., Yaacobi, A., Hosseini, E. S. & Watts, M. R. Large-scale nanophotonic phased array. *Nature* **493**, 195–199 (2013).
40. Harris, N. C. et al. Quantum transport simulations in a programmable nanophotonic processor. *Nat. Photon* **11**, 447–452 (2017).
41. Shafiee, A. et al. ISAAC: a convolutional neural network accelerator with in-situ analog arithmetic in crossbars. In *Proc. 2016 43rd Annual International Symposium on Computer Architecture (ISCA)* (IEEE, 2016).
42. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. In *2016 Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)* (IEEE, 2016).
43. Gao, L., Chen, P. Y. & Yu, S. Demonstration of convolution kernel operation on resistive cross-point array. *IEEE Electron Device Lett.* **37**, 870–873 (2016).
44. Qiu, X., Zhang, D., Zhang, W. & Chen, L. Structured-pump-enabled quantum pattern recognition. *Phys. Rev. Lett.* **122**, 123901 (2019).
45. Kumar, B. et al. Correlation pattern recognition for face recognition. *Proc. IEEE* **94**, 1963 (2006).
46. Gray, R. M. Toeplitz and circulant matrices: a review. *now publishers inc.* **2**, 155–239 (2006).
47. Suthaharan, S. *Machine Learning Models and Algorithms for Big Data Classification* (Springer, 2016).
48. Nilsback, M.-E. & Andrew Z. A visual vocabulary for flower classification. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)* 2 (IEEE, 2006).
49. Krizhevsky, A., Ilya, S. & Geoffrey, E. H. Imagenet classification with deep convolutional neural networks. *Adv. neural Inf. Process. Syst.* **25**, 1097–1105 (2012).
50. Hermans, M., Soriano, M., Dambre, J., Bienstman, P. & Fischer, I. Photonic delay systems as machine learning implementations. *J. Machine Learn Res.* **16**, 2081–2097 (2015).
51. Antonik, P., Nicolas, M. & Damien, R. Large-scale spatiotemporal photonic reservoir computer for image classification. *IEEE J. Sel. Top. Quant. Electron* **26.1**, 1–12 (2019).
52. Ding, C.W., et al. Structured weight matrices-based hardware accelerators in deep neural networks: Fpgas and asics. In *Proceedings of the 2018 on Great Lakes Symposium on VLSI* (Association for Computing Machinery, 2018).
53. Liang, Z., et al. Theoretical properties for neural networks with weight matrices of low displacement rank. In *Proc. 34th International Conference on Machine Learning* 4082–4090 (JMLR.org, 2017).
54. Jiaqi, G. et al. Towards area-efficient optical neural networks: an FFT-based architecture. In *2020 25th Asia and South Pacific Design Automation Conference (ASP-DAC)* (IEEE, 2020).

## Acknowledgements

This work was supported by the Singapore National Research Foundation under the Competitive Research Program (NRFCRP13-2014-01) and the Singapore Ministry of Education (MOE) Tier 3 grant (MOE2017-T3-1-001).

## Author contributions

H.H.Z., L.C.K., M.G., and A.Q.L. jointly conceived the idea. H.H.Z., J.Z., H.Z., S.B.L., X.D.J., and A.Q.L. performed the numerical simulations and theoretical analysis. H.H.Z., J.Z., and L.X.W. did the experiments. H.H.Z., B.W., N.W., H.C., Y.Z.S., X.S.L., L.P., X.D.J., L.C.K., M.G., and A.Q.L. involved in the discussion and data analysis. H.H.Z., Y.Z.S., J.T., M.G., and A.Q.L. prepared the manuscript. X.H.Z., L.M.X., and W.H. revised the manuscript. L.C.K., M.G., and A.Q.L. supervised and coordinated all the work. All authors commented on the manuscript.

## Competing interests

The authors declare no competing interests.

## Additional information

**Supplementary information** The online version contains supplementary material available at <https://doi.org/10.1038/s41467-022-28702-0>.

**Correspondence** and requests for materials should be addressed to X. D. Jiang, X. H. Zhou, L. M. Xiao, M. Gu, L. C. Kwek or A. Q. Liu.

**Peer review information** *Nature Communications* thanks Dong-Ling Deng, Miguel Soriano and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

**Reprints and permission information** is available at <http://www.nature.com/reprints>

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



**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2022