Opto-Electronic Advances

CN 51-1781/TN ISSN (Print) 2096-4579 ISSN (Online) 2097-3993

Hardware-aware lightweight photonic spiking neural network for pattern classification

Shuiying Xiang*, Yahui Zhang, Shangxuan Shi, Haowen Zhao, Dianzhuang Zheng, Xingxing Guo, Yanan Han, Ye Tian, Liyue Zhang, Yuechun Shi and Yue Hao

Citation: Xiang SY, Zhang YH, Shi SX et al. Hardware-aware lightweight photonic spiking neural network for pattern classification. *Opto-Electron Adv* **9**, 250319 (2026).

https://doi.org/10.29026/oea.2026.250319

Received: 19 November 2025; Accepted: 16 March 2026; Published online: 23 April 2026

OEA | Opto-Electronic Advances

Article

2026, Vol. , No.

Open Access

Hardware-aware lightweight photonic spiking neural network for pattern classification

Shuiying Xiang[1]*, Yahui Zhang[1], Shangxuan Shi[1], Haowen Zhao[1], Dianzhuang Zheng[1], Xingxing Guo[1], Yanan Han[1], Ye Tian[1], Liyue Zhang[2], Yuechun Shi[3] and Yue Hao[1]

Abstract: There exists a significant scale gap between photonic neural network integrated chips and neural networks, which hinders the deployment and application of photonic neural network. Here, we propose hardware-aware lightweight spiking neural networks (SNNs) architecture tailored to our photonic neuromorphic chips, and conduct hardware-software collaborative computing for solving pattern classification tasks. We employed a simplified Mach-Zehnder interferometer (MZI) mesh for performing linear computation, and 16-channel distributed feedback lasers with saturable absorber (DFB-SA) array for performing nonlinear spike activation. Both photonic neuromorphic chips based on the MZI mesh and DFB-SA array were designed, optimized and fabricated. Furthermore, we propose a lightweight SNN with discrete cosine transform to reduce input dimension and match the input/output ports number of the photonic neuromorphic chips. We demonstrated an end-to-end inference of an entire layer of the lightweight photonic SNN. The hardware-software collaborative inference accuracy is 90% and 80.5% for MNIST and Fashion-MNIST datasets, respectively. The energy efficiency is 1.39 TOPS/W for the MZI mesh, and is 987.65 GOPS/W for the DFB-SA array. The lightweight architecture and experimental demonstration address the challenge of scale mismatch between the photonic chip and SNN, paving the way for the hardware deployment of photonic SNNs.

Keywords: photonic neuromorphic; photonic spiking neural network; pattern classification

DOI: 10.29026/oea.2026.250319 | CSTR: 32247.14.oea.2026.250319

Citation: Xiang SY, Zhang YH, Shi SX et al. Hardware-aware lightweight photonic spiking neural network for pattern classification. *Opto-Electron Adv*, 250319 (2026).

1 Introduction

With the rapid development of artificial intelligence (AI) represented by ChatGPT and Deepseek, the demand for computing power has shown explosive growth. Traditional electronic computing chips based on the von Neumann architecture are facing severe challenges of the "memory wall" and "power wall", and their computing power growth is gradually approaching physical limits. Photonic neural networks (PNNs) stand out with their ultra-high speed, ultra-low power consumption, and inherent large-scale parallel processing capabilities, becoming one of the most promising disruptive computing paradigms in the post-Moore era[1–12].

In recent years, PNNs based on Mach-Zehnder interferometers (MZI)[13–23], micro-ring resonator[24–27], phase change material crossbars[28–30], and spiking laser neurons[31–37] have been studied extensively. Specifically, programmable MZI mesh-based PNN chips have been developed rapidly since 2017, with their matrix scale gradually advancing from the initial 4×4[13,14] to the sophisticated 128×128[23]. For diverse photonic computing paradigms, reported PNNs implementations cover photonic multi-layer perceptrons[38–39], convolutional neural networks[40–43], tensor processing[44–48], and diffractive neural networks[49–55].

Among numerous neural network models, spiking neural networks (SNNs) have attracted significant attention due to their adoption of the efficient and sparse processing mechanisms of the biological brain. Combining the ultra-high speed and low-power advantages of photonic technology

Received: 19 November 2025

Accepted: 16 March 2026

Published online: 23 April 2026

[1]State Key Laboratory of Integrated Service Networks, State Key Discipline Laboratory of Wide Bandgap Semiconductor Technology, Xidian University, Xi'an 710071, China; [2]Key Laboratory of Photonic-Electronic Integration and Communication-Sensing Convergence (Ministry of Education), Southwest Jiaotong University, Chengdu 611756, China; [3]Yongjiang laboratory, No. 1792 Cihai South Road, Ningbo 315202, China.
*Correspondence: SY Xiang, E-mail: syxiang@xidian.edu.cn

with the biologically inspired efficiency of SNNs has spawned the emerging interdisciplinary field of photonic spiking neural networks (PSNNs). PSNNs are expected to achieve remarkable computing speeds while keeping energy consumption at an extremely low level, providing an ideal hardware solution for scenarios such as edge computing and real-time intelligent processing. In addition, these photonic SNN systems exhibit outstanding noise immunity while easing the rigorous requirements for high-precision analog-to-digital (AD) and digital-to-analog (DA) conversions during the input of spike signals and retrieval of spike output data. Consequently, PSNN emerges as a hardware-friendly architecture. Recently, PSNNs have attracted numerous attention[28,32–37,56–60]. However, there is a huge gap between the scale of actual photonic neuromorphic chips and the scale of SNNs required to complete specific tasks. This mismatch in scale between "software" (neural network models) and "hardware" (neuromorphic chips) seriously hinders the transition of PSNNs from theoretical models to practical applications.

To address this core contradiction, the key lies in model compression of PSNNs. That is, on the premise of maintaining their core functions and performance, the complexity and scale of the model are significantly reduced to enable efficient deployment on resource-constrained photonic chips. Traditional model compression methods, such as pruning and quantization, have been applied in electronic SNNs, but they often fail to fully consider the unique physical characteristics and hardware constraints of photonic computing. Therefore, there is an urgent need for a new lightweight architecture that is highly compatible with the characteristics of photonic hardware.

We propose a lightweight PSNN architecture based on the discrete cosine transform (DCT)[61], and deploy it on a photonic neuromorphic chip for hardware-software collaborative computing, aiming to fundamentally bridge the gap between algorithm scale and hardware chip size. We introduce DCT into the design of PSNNs to reduce the dimensionality and eliminate the redundancy of the input data, retaining only the most informative frequency components as input to the SNN. This fundamentally reduces the dimensionality of the processed data and the architectural scale required for subsequent networks, thereby enabling the deployment of the entire SNN on a photonic chip. We consider a fully connected network architecture in order to achieve single-time-step SNN training and complete the classification tasks for the MNIST and Fashion-MNIST datasets. The hardware-software co-inference based on the photonic neuromorphic chip and the lightweight PSNN is further carried out. This work not only offers an innovative technical solution to address the scale bottleneck of PSNNs, but also marks a key step in promoting the transition of photonic neuromorphic computing from the laboratory to practical applications such as edge computation and embodied intelligence.

2 Experimental setup and method

2.1 Lightweight architecture of PSNN with DCT

To match the limited fan-in size of the photonic neuromorphic computing chip, we propose a frequency-domain pruning method by applying DCT to the input image of the SNN. As shown in Fig. 1(a), after DCT, the frequency spectrum of the input image is obtained. We select the low-frequency components and feed them into the SNN for further processing, which significantly reduces the input size of the SNN. The frequency-domain components are flattened and then fed into the first linear layer of the SNN. Following the linear layer, the weighted summation signals are conveyed to the leaky integrate-and-fire (LIF) neuron for nonlinear spike activation. Different hidden layers can be considered for different tasks.

The DCT is a mathematical transformation that converts spatial domain data into the frequency domain, capturing the frequency characteristics of the input data. In image processing, two-dimensional (2D) DCT is applied to transform a spatial-domain image (e.g., a 28×28 MNIST image) into its frequency spectrum, where the coefficients represent the energy distribution at different frequency components. Mathematically, the 2D DCT is defined as:

$$F(u,v)=\sum_{x=0}^{N-1}\sum_{y=0}^{N-1} I(x,y)\cos\left[\frac{\pi(2x+1)u}{2N}\right]\cos\left[\frac{\pi(2y+1)v}{2N}\right],$$

where $I(x,y)$ is the input image pixel value, and $F(u,v)$ represents the frequency coefficients after transformation. u,v are the indices of the frequency components, and N is the image dimension.

The 2D DCT concentrates low-frequency components in the top-left corner of the transformed spectrum, while the high-frequency components spread towards the bottom-right corner. In our scheme, the upper triangular region of the 2D DCT spectrum is extracted, retaining only the most significant low-frequency features. This reduces the input data dimension while preserving essential information, thereby simplifying the computational complexity for the subsequent neural network processing.

2.2 Photonic synapse chip based on simplified MZI mesh

Note, there is currently no single material system that can simultaneously satisfy the high-performance requirements of photonic SNNs for large-scale and low-loss linear photonic computation and on-chip efficient nonlinear photonic operation. Our design leverages the complementary strengths of silicon-photonics-based and InP-based materials to ensure the performance, manufacturability and integration potential.

As shown in Fig. 1(b), we deployed the linear layer on a photonic synapse array chip to perform photonic linear

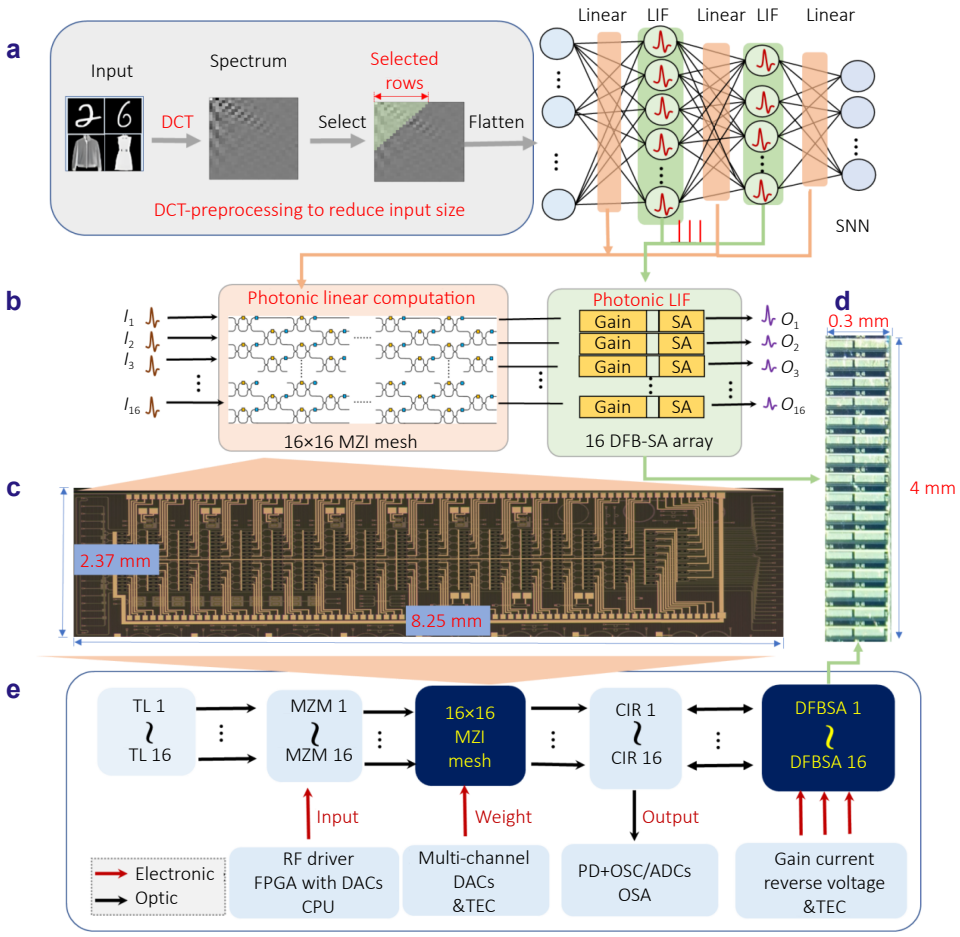

Fig. 1 | (a) The overall architecture of the frequency-domain SNN with DCT. (b) Schematic diagram of photonic linear computation and photonic LIF neuron. (c) Microscopic image of the fabricated MZI mesh chip[60]. (d) Microscopic image of the fabricated DFB-SA array chip[60]. (e) The testing experimental setup[60]. TL: tunable laser; MZM: Mach-Zehnder modulator; RF: radio frequency; FPGA: field programmable gate array; DAC: digital-analogy converter; CPU: central processing unit; MZI: Mach-Zehnder interferometers; TEC: temperature controller; CIR: optical circulator, PD: photodetector; OSC: oscilloscope; OSA: optical spectrum analyzer; DFB-SA: distributed feedback lasers with saturable absorber.

computation. For this purpose, we designed and fabricated a chip based on a simplified MZI mesh on a silicon photonic platform to execute incoherent optical matrix-vector multiplication (MVM). The structure with a 16 × 16 MZI mesh is designed, where each MZI incorporates only a single phase shifter on one of its inner arms. For representing a 16 × 16 weight matrix, the design utilizes 17 × (17−1)/2 + 16 = 152 phase shifters. The average loaded voltage for each phase shifter is approximately 3.1 V and the resistance is about 800 ohms. Thus, the average power consumption for each phase shifter is about 0.012 W, and the total power consumption is about 1.8 W. Thus, this structure achieves significant area, transmission loss and power consumption reduction. The chip features a length of 8.25 mm, a width of 2.37 mm, and a corresponding area of 19.55 mm^2, with its microscopic image presented in Fig. 1(c). When light is injected through a single port, the chip demonstrates an approximate insertion loss of 13 decibels (dB).

2.3 Photonic nonlinear spiking neuron chip based on DFB-SA

As shown in Fig. 1(b), the LIF layer was deployed on 16-channel distributed feedback lasers with saturable absorber (DFB-SA) laser array, fabricated on an III-V platform to enable element-wise nonlinear spike activation. As presented in Fig. 1(d), the array chip measures 4 mm in length and 0.3 mm in width, with a total area of 1.2 mm^2. Each single-channel DFB-SA laser within the array is structured with two sections, namely a gain region and a saturable absorber (SA) region. The gain region is forward-biased by a current source (I_G), while the SA region is reverse-biased by a voltage source (V_{SA}). The epitaxial wafer structure for single-channel DFB-SA laser unit is presented in Fig. 2(a). The optimized multi-quantum well (MQW) structure comprises 7 well layers, and the thickness of the upper and lower confining layers is approximately 70 nm. Furthermore, anti-reflection (AR) coatings and high-reflection (HR) coatings

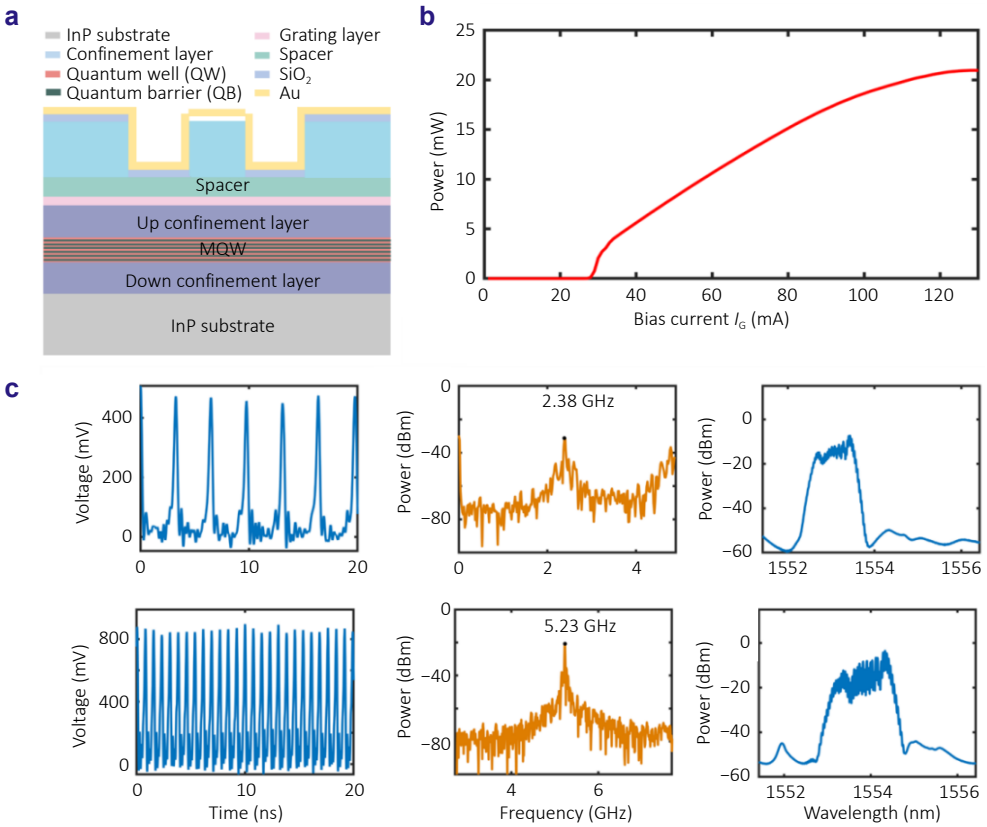

Fig. 2 | The epitaxial wafer structure and properties of DFB-SA laser. (a) The schematic of epitaxial wafer structure of DFB-SA laser. (b) Power current curve of DFB-SA laser. (c) The self-pulsation, frequency spectra of self-pulsation state and optical spectra of DFB-SA laser, for I_G =40.6 mA, V_{SA} = -0.26 V (upper row), and I_G =98.8 mA, V_{SA} = -1.88V (lower row).

are deposited on the two laser facets to boost the light emission power, while the SA region is placed close to the HR side. As shown in Fig. 2(b), the measured lasing threshold is about 28 mA, and is almost the same for all the 16 DFB-SA lasers. The self-pulsation outputs, the corresponding frequency spectra and optical spectra for different I_G and V_{SA} are displayed in Fig. 2(c). The maximum self-pulsation frequency is measured to be 5.23 GHz. When biasing slightly below the self-pulsation state, the nonlinear neuron-like response can be achieved[28], similar to our previous demonstration of single DFB-SA laser. The optical spectrum of the DFB-SA laser broadens when the device operates in the self-pulsation state.

2.4 Experimental setup for testing the photonic neuromorphic computing system

The experimental setup of photonic neuromorphic computing system for deploying frequency-domain SNN is illustrated in Fig. 1(e). The MZI mesh chip and DFB-SA laser array chips are packaged separately, and the two chips are interconnected via optical fiber coupling. To ensure stable and scalable operations, we calibrate the wavelength, optical power, modulation point, optical polarization, and latency of the entire optical link before each experiment. Optical carriers with different wavelengths are generated by multi-channel tunable lasers (TLs). The input signals for the photonic neuromorphic computing system were generated by a Xilinx FPGA platform, ZCU216 evaluation board, which features a Zynq UltraScale+ RFSoC 49DR chip and 16-channel high-speed AD/DA converters. The FPGA was controlled by a digital computer. The input signals were modulated into optical carriers by the Mach-Zehnder modulators (MZMs). The modulated optical signal were then injected into the MZI mesh. The weights of the frequency-domain SNN were deployed on the MZI mesh, which was controlled through a multi-channel voltage source and a TEC for thermal stabilization. To perform the nonlinear computation, the 16-channel output from the MZI mesh was fed into a 16-channel DFB-SA laser array. The output from the 16-channel DFB-SA laser array was routed through an array of three-port optical circulators. The resulting optical signals were then characterized using an optical spectrum analyzer and converted into electrical signals by a photodetector (Agilent/HP 11982A) for subsequent acquisition with an oscilloscope (Keysight DSOZ592A). To mitigate the influence of temperature fluctuations on device performance, the operating temperature of the MZI mesh

chip and DFB-SA laser array chip is actively stabilized at 25 °C by the TEC.

3 Results

3.1 Software-hardware collaborative training-inference framework

To tackle the challenges of SNN training difficulty and accuracy degradation in optical computing, we propose an end-to-end three-stage software-hardware-software collaborative training-inference framework tailored for frequency-domain photonic SNNs. Three training phases are incorporated into this framework, namely software pre-training phase, local photonic hardware in-situ training phase, and hardware-aware software fine-tuning phase. The complete workflow of the hardware-software collaborative training and inference process is illustrated in Fig. 3.

SNN pre-training. As shown in Fig. 3(a), the process begins with training an ANN, which is then converted to a SNN. Next, direct training of the SNN is performed using backpropagation based on surrogate gradients. Notably, a temporal pruning method is adopted to achieve low latency inference. This method treats the threshold and leakage of SNN as trainable parameters, initiating training with an SNN configured for T time steps and progressively reducing the time steps in each iteration, ultimately yielding an SNN with a single time step (T=1). In our implementation, the time steps of SNN are gradually compressed from 5 to 3, and finally to 1. Note, for T=1, low-latency inference and stable hardware spike activation outputs can be achieved. We designate the software-trained weights to be mapped onto the hardware MZI chip as $\boldsymbol{W}^{\mathrm{S}}$. The complete hybrid training algorithm, incorporating DCT, ANN-to-SNN conversion, direct training with surrogate gradient and temporal pruning, is summarized in Algorithm 1.

Algorithms 1. Single time step SNN training with DCT and temporal pruning

Input: Input 28×28 MNIST image I, number of selected rows in spectrum diagram N_r, trained the SNN model with N time steps (T_N), the time steps reduction step size b, number of epochs to train e.

DCT pre-processing:

Apply DCT to I to obtain a 28×28 spectrum diagram F.

Extract the upper triangle region from the spectrum diagram using N_r:

for each pixel $F\ (i,j)$:

Include pixel if $i+j \le N_r-1$

// The size of input data for the network is $\dfrac{N_r \times (N_r+1)}{2}$

Training with temporal pruning:

Train an ANN, new SNN initialized with trained parameters of T_N, reduced latency $T_r = N - b$

while $T_r \ge 1$ **do**

// Training phase

for epoch 1: e **do**

// Train network with T_r timesteps

end for

// Initialize another iso-architecture SNN with parameters of above trained network

// Temporal pruning

$T_r = T_r - b$

end while

Local photonic hardware in-situ training. To represent the target weight matrix $\boldsymbol{W}^{\mathrm{S}}$, the MZI mesh chip is configured using the stochastic parallel gradient descent (SPGD) algorithm based on local in-situ training[18]. This configuration strategy is due to the implicit nature of the optical matrix characterized by the MZI mesh. Specifically, there is no direct correspondence between one or more phase shifter heater voltages and a single element of the matrix. However, the weights trained on the hardware MZI chip may not perfectly match $\boldsymbol{W}^{\mathrm{S}}$, as they are affected by noise, fabrication imperfections, and the limited precision of the multi-channel DACs that drive the phase shifters. The actual weight matrix represented by the trained MZI mesh is denoted as $\boldsymbol{W}^{\mathrm{H}}$. We first consider a lightweight frequency-domain SNN with a size of 45×16×16×10 to classify the MNIST dataset. Specifically, after applying the DCT, we select only 45 low-frequency components for classification with frequency-domain SNN. The network comprises one input layer with 45 neurons, two hidden layers with 16 neurons each, and one output layer with 10 neurons. Such a small input layer size matches well with the fan-in number of an MZI mesh chip that can be fabricated using currently available silicon photonics process. In our experimental demonstration, the entire hidden layer with a size of 16×16 is deployed to the fabricated MZI mesh chip and the DFB-SA array chip.

The experiment setup for configuring MZI mesh chip is depicted in Fig. 3(b). Continuous wave (CW) input is generated by a 16-channel TL source, while a 16-channel optical switch (OS) controls the injection light into the input ports of the MZI chip. Multi-channel voltage sources, controlled by a host computer, are used to adjust the weighting of the MZI. The output power of the MZI chip is detected by a 16-channel optical power meter (PM) and then is feedback to the host computer. For the iterative update of MZI mesh weights, the SPGD algorithm is implemented to optimize the driving voltage of each phase shifter within the MZI mesh. The optimized voltages are then loaded onto the corresponding phase shifters by the host computer through regulating the multi-channel voltage sources via an Ethernet link.

The progress of the in-situ training for the MZI chip is illustrated in Fig. 3(e), which primarily consists of five key steps, system initialization, weight matrix characterization, performance evaluation, gradient estimation via perturbation, and iterative optimization.

1) System initialization. A set of random voltages $\boldsymbol{U_i}$ is applied to the MZI phase shifters to initialize the MZI mesh chip.

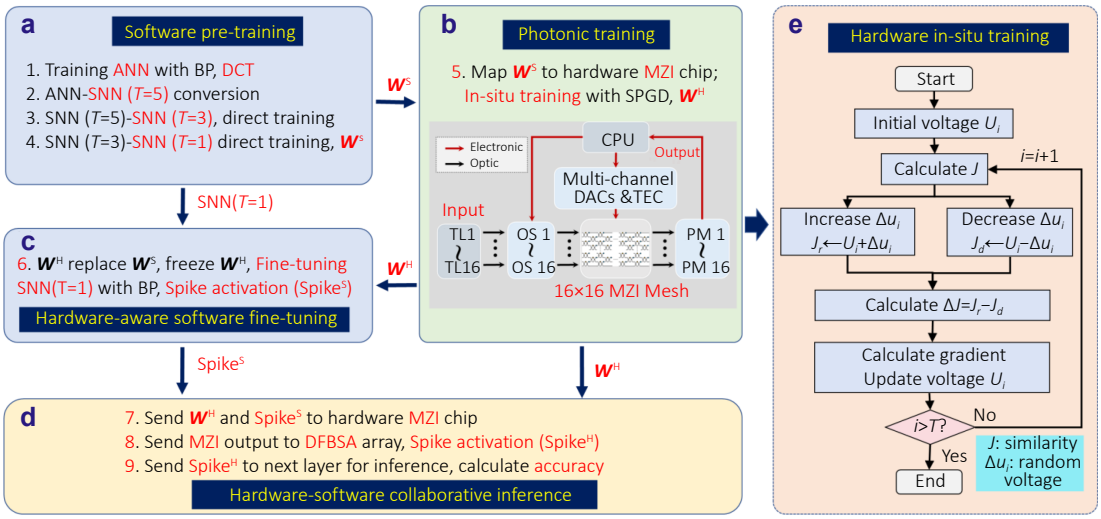

Fig. 3 | Software-hardware-software collaborative training-inference framework for frequency-domain photonics SNNs.

2) Weight matrix characterization. To characterize the weight matrix, CW light is first injected into a single input port. The optical power measured from the 16 output ports then forms one column of the matrix. This process is repeated by sequentially injecting light into each of the 16 input ports, ultimately constructing the complete 16×16 hardware weight matrix $\boldsymbol{W}^{\mathrm{H1}}$.

3) Performance evaluation. The similarity J between the characterized hardware weight matrix W^{H1} and the target matrix W^{S} is quantified using cosine similarity[18].

4) Gradient estimation. The gradient is estimated by applying positive and negative voltage random perturbations ($U_i+\Delta u_i$ and $U_i-\Delta u_i$), remeasuring the similarity metric ($\boldsymbol{J}_{\mathrm{r}}$ and $\boldsymbol{J}_{\mathrm{d}}$), and computing the difference $\Delta \boldsymbol{J} = \boldsymbol{J}_{\mathrm{r}} - \boldsymbol{J}_{\mathrm{d}}$. The phase shifter voltages are then updated by calculating the gradient based on $\Delta \boldsymbol{J}$ and Δu_i.

5) Iterative optimization. The phase shifter voltages are updated based on the calculated gradients, and steps 2–5 are repeated until the hardware weight matrix converges to the target.

Hardware-aware software fine-tuning stage. As shown in Fig. 3(c), the software-trained weights $\boldsymbol{W}^{\mathrm{S}}$ are substituted with their hardware-calibrated counterparts $\boldsymbol{W}^{\mathrm{H}}$. During the subsequent fine-tuning stage, the weights corresponding to the MZI layer ($\boldsymbol{W}^{\mathrm{H}}$) are fixed, while the remaining weights in the network are adjusted. This hardware-aware fine-tuning strategy ensures that the weights of the MZI layer are accurately represented by the physical chip, effectively compensating for inherent manufacturing variations and hardware noise.

Hardware-software collaborative inference. As illustrated in Fig. 3(d), following the hardware-aware software fine-tuning stage, the converged weights $\boldsymbol{W}^{\mathrm{H}}$ and input spike signals are sent to the MZI chip for optical processing. The output of MZI is then fed into a DFB-SA array to generate activated spikes ($\mathrm{Spike}^{\mathrm{H}}$), which are subsequently transmitted to the next software-based layer for further inference, with the aim of calculating accuracy.

3.2 MNIST and Fashion-MNIST classification using photonic neuromorphic chips

To evaluate the performance, we first consider the lightweight frequency-domain SNN with a size of 45×16×16×10 to classify the MNIST dataset. The pre-training and hardware-aware fine-tuning results are presented in Fig. 4(a). Here, the evolution of the training and testing accuracy and loss, as well as the confusion matrix are shown. We can find that the accuracy is 95.66% for ANN, 94.4% for SNN with T=5, 94.04% for SNN with T=3, 91.23% for SNN with T=1, and reaches 90.17% after performing hardware-aware fine-tuning. The hardware in-situ training results are presented in Fig. 4(b). The accuracy of the hardware in-situ training of the MZI mesh is converged to 0.896 for the MNIST dataset. There is a mismatch between simulation and experimental matrices. That is primarily arises from non-ideal effects inherent to the physical chip, including thermal effects and fabrication-induced inhomogeneities, as well as experimental noise and alignment fluctuations. Nevertheless, the mismatch exerts a negligible influence on the computational process. Moreover, the convergence behavior of the SPGD-based in-situ training under measurement noise and drift is demonstrated through independent training runs with different random weight initializations and optical carrier wavelengths. It can be found that, even with only a single time step and with the input size of network reduced to only 45 (compared to 784 for conventional scheme), the accuracy maintains a relatively high value. Thus, this lightweight frequency-domain SNN with DCT closely matches the size of available photonic SNN

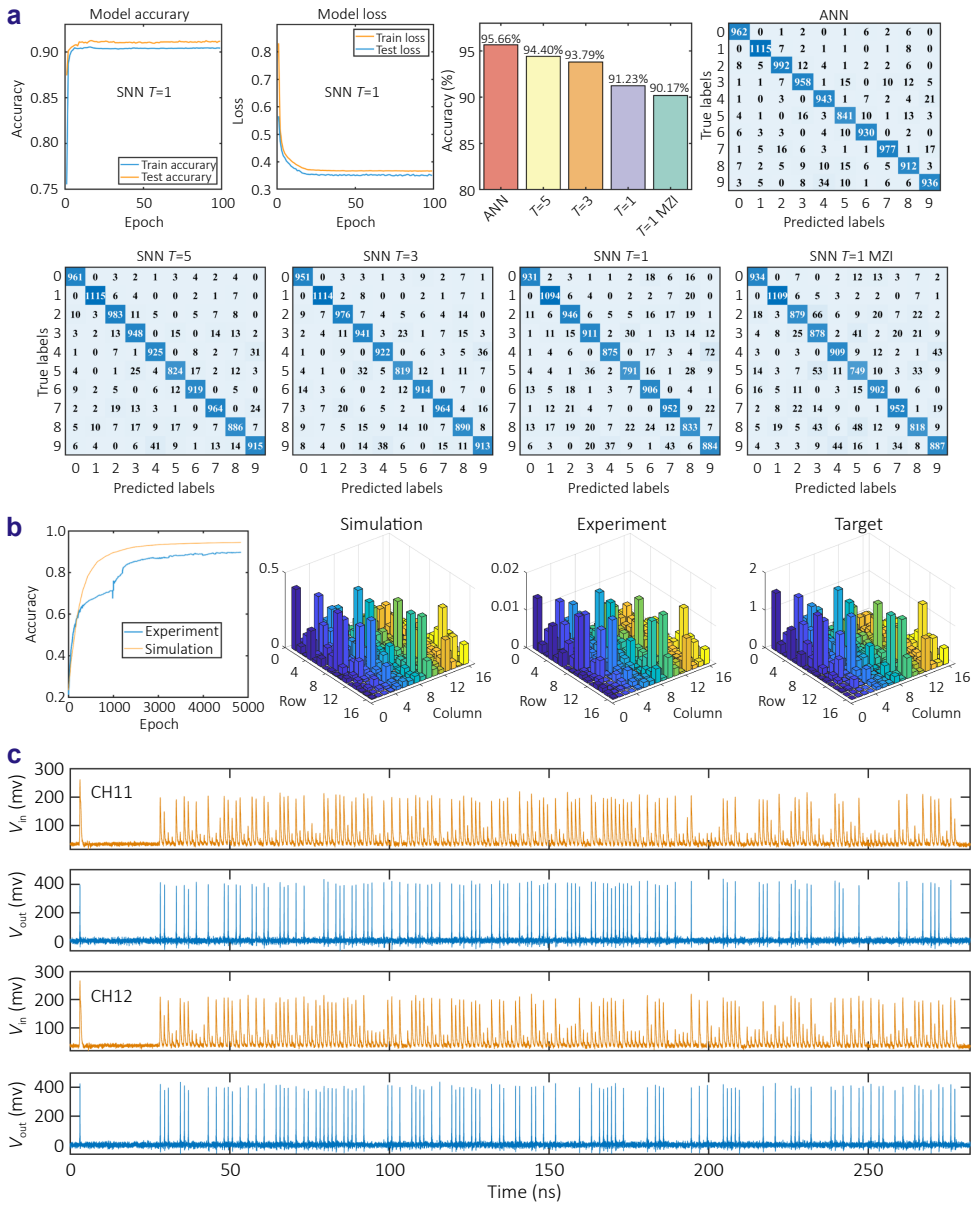

Fig. 4 | The results for the training and inference stage for MNIST datasets. (a) Accuracy and loss as functions of epoch, the test accuracy and confusion matrices for ANN training, SNN with *T*=5, SNN with *T*=3, SNN with *T*=1, and hardware-aware fine-tuning SNN with *T*=1. (b) Local in-situ training results including experimental and simulated accuracy as functions of epoch. (c) The experimentally measured linear computation and nonlinear spike activation results of two representative channels.

chips, effectively bridging the gap between the network size and the integrated photonic chip scale.

By deploying the weights to the MZI mesh chip, we measured the hardware linear computation and nonlinear spike activation results for 200 randomly selected test images for hardware and software inference. The experimentally measured results are presented in Fig. 4(c). Here, the temporal results of two channels are displayed for simplicity. Clearly, the MZI mesh chip can perform linear MVM function, and the DFB-SA laser array can accurately perform the nonlinear spike activation. Based on the neuron-like response of the DFB-SA laser, input pulses with relatively small amplitudes cannot trigger spike responses, and the nonlinear outputs become much sparser than the linear outputs.

The confusion matrix for the hardware chip inference for the MNIST dataset is presented in Fig. 5(a), and the hardware testing accuracy is 90%, which is slightly lower than 90.17% that achieved in the hardware-aware software fine-tuning stage, indicating that the proposed frequency-domain SNN is robust to fabrication errors and system noise.

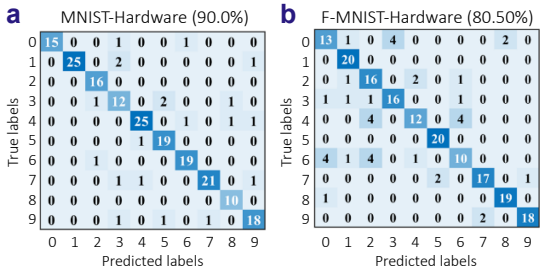

Fig. 5 | The experimentally measured testing accuracy for (a) MNIST and (b) Fashion-MNIST dataset.

We also use the same frequency-domain SNN network structure and photonic neuromorphic computing chips to classify the Fashion-MNIST. The results of training accuracy and loss, confusion matrices, as well as the photonic hardware in-situ training results can be found in Fig. 6(a–b). It can be seen that, the hardware in-situ training accuracy of the MZI mesh is converged to 0.836 for the Fashion-MNIST dataset. We find that the accuracy is 85.57% for ANN, 84.45% for SNN with T=5, 84.35% for SNN with T=3, 81.23% for SNN with T=1, and 80.66% for hardware-aware

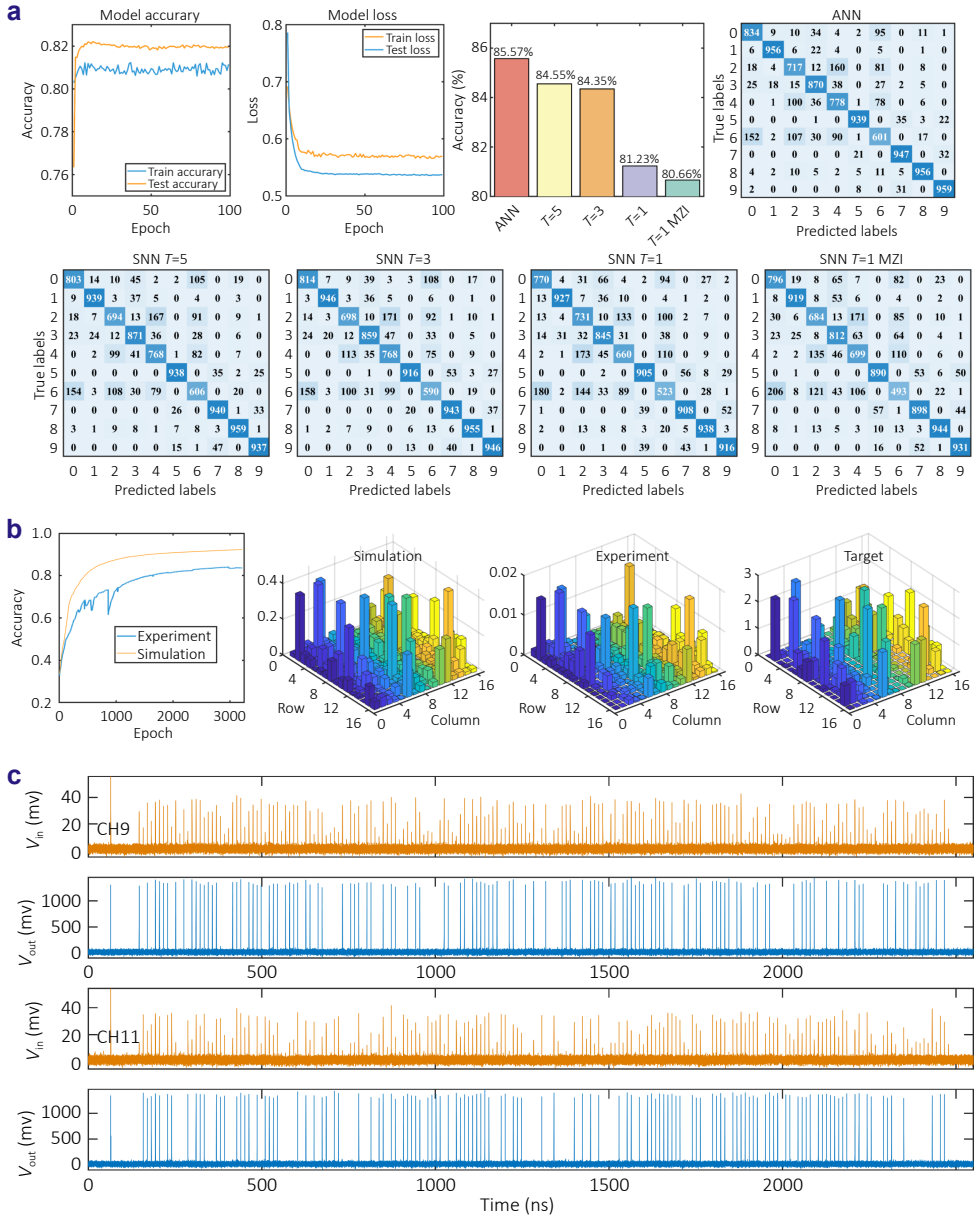

Fig. 6 | The results for the training and inference stage for Fashion-MNIST dataset. (a) Accuracy and loss as functions of epoch, the test accuracy and confusion matrices for ANN training, SNN with T=5, SNN with T=3, SNN with T=1, and hardware-aware fine-tuning SNN with T=1. (b) Local in-situ training results including experimental and simulated accuracy as functions of epoch. (c) The experimentally measured linear computation and nonlinear spike activation results of two representative channels.

fine-tuning SNN with T=1. The hardware linear and nonlinear computation temporal results for Fashion-MNIST can be found in Fig. 6(c). Correspondingly, as shown in Fig. 5(b), the hardware testing accuracy is 80.50% for the Fashion-MNIST dataset, and is slightly lower than 80.66% for software inference.

We further consider numerical simulations to explore different network sizes. The numerical results for frequency-domain SNNs with size of 15×16×16×10 and 120×64×64×10 are presented for both datasets in Fig. 7. For MNIST dataset, the accuracy for SNN with T=1 is 81.57% for the network size of 15×16×16×10, and is 96.68% for the network size of 120×64×64×10. For the Fashion-MNIST, the accuracy for SNN with T=1 is 76.79% for the network size of 15×16×16×10, and is 86.56% for the network size of 120×64×64×10. The numerical findings validate the feasibility of enhancing the inference accuracy with larger-scale SNNs.

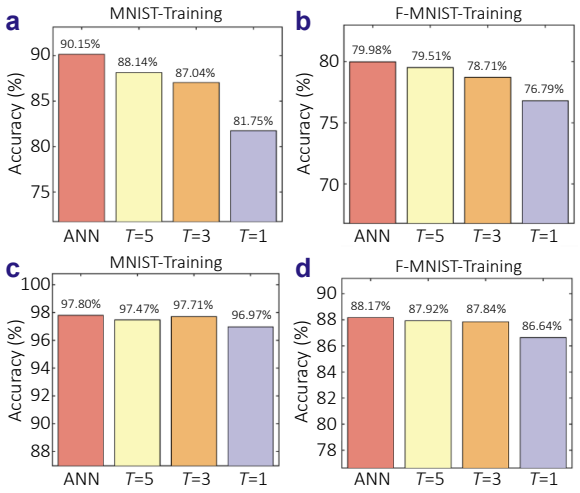

Fig. 7 | SNN pre-training results for (a) MNIST and (b) Fashion-MNIST dataset with a size of 15×16×16×10. SNN pre-training results for (c) MNIST and (d) Fashion-MNIST dataset with a size of 120×64×64×10.

4 Discussion and conclusion

Metrics. For PSNNs encompassing both linear and nonlinear computing, we evaluated the key performance metrics of the MZI mesh chip and the DFB-SA laser array chip. Here, the maximum achievable refresh rate of the entire system is determined by the refractory period. Note, as the DFB-SA laser array operates at a maximum self-pulsation frequency of approximately 5 GHz, which defines the upper limit of the refractory period. Thus, the operating signal rate of the MZI is accordingly set to 5 GHz. Under this operational parameter setting, the computing rate of the MZI chip is calculated as 2×16×16×5 GHz, yielding a value of 2.5 TOPS. Leveraging the metrics of the area and total power consumption of chip, we calculated its energy efficiency at 1.39 TOPS/W and computing density at 0.13 TOPS/mm^2. As for the 16-channel DFB-SA laser array chip, its throughput is calculated as 640 GOPS, considering that each input undergoes 8 equivalent operations when emulating the LIF neuron, with the formulation being 16×8×5 GHz. With a total energy consumption of approximately 0.648 W, the chip achieves an energy efficiency of 987.65 GOPS/W and a computing density of 533.33 GOPS/mm^2. The end-to-end latency for implementing a full layer of the SNN, including both photonic MVM and photonic LIF modules, is calculated as 320 ps. Table 1 presents a comparative analysis of our chips against state-of-the-art PNN chips including optical nonlinear computation and electronic neuromorphic chips. To the best of our knowledge, among previously reported PNN chips capable of supporting optical nonlinear computing, our design is superior to counterparts in terms of latency and the number of trainable parameters.

Scalability. Notably, the photonic spiking neurons based on the DFB-SA laser array exhibit excellent scalability. They can be readily expanded to 150 channels while maintaining high wavelength precision[66]. For the MZI mesh architecture, recent work has already demonstrated the feasibility of large-scale integration with size of 128×128[23]. Thus, we expect that the PSNN may be scale-up expansion to a 128-channel with manageable loss levels. The power consumption of a 128-channel DFB-SA laser array is estimated as 7.78 W. By similar structure optimization as reported in ref.[23], the estimated linear propagation loss of 128-channel MZI chip is about 4.6 dB, and the corresponding power consumption is 3.2 W[23]. In addition, silicon dioxide (SiO_2) isolation trenches can be introduced into the array during chip design to mitigate the intra-chip thermal crosstalk. Furthermore, advanced heterogeneous integration technologies offer the potential to achieve even tighter integration of photonic synapse chips and photonic spiking neuron chips, laying the groundwork for higher-performance PSNN systems.

All-optical DCT potential. DCT can compress image into low-frequency-dominated frequency-domain features, which match the input ports of PSNN chip after dimensionality reduction and thus lower the difficulty of hardware implementation. It is worth noting, however, that DCT also introduces additional power consumption and latency in the preprocessing stage. Fortunately, DCT can be inherently implemented in diffractive neural networks relying on optical passive devices[67]. Looking ahead, integration with this all-optical DCT paradigm could empower all-optical pattern classification to enable pattern preprocessing and recognition without the need for EO/OE conversions. This integration would leverage inherent optical signal processing capabilities of the system, achieving remarkable low latency and superior energy efficiency.

Conclusion. In summary, we present a programmable photonic neuromorphic computing chip equipped with 272

Table 1 | Comparison with state-of-the-art photonic neural network chips and electronic neuromorphic chips.

Metrics	Computing density	Energy efficiency	Latency	Trainable parameters	Computing accuracy (%) (numerical/experimental results)	Optical nonlinear computation	For SNN
Our work	**0.13 TOPS/mm²**	**1.39 TOPS/W**	**320 ps**	**272**	**91.23/90 (MNIST) 81.23/80.5 (Fashion-MNIST)**	**√**	**√**
Ashtiani et al.[15]	1.75 TOPS/mm²	2.9 TOPS/W	570 ps	67	93.8 (hand letters)	√	X
Bandyopadhyay et al.[17]	0.02 TOPS/mm²	0.013 TOPS/W	410 ps	132	92.5 (vowel classification)	√	X
Feldmann et al.[28]	N/A	N/A	N/A	64	N/A	√	√
Fang et al.[19]	0.23 TOPS/mm²	N/A	N/A	N/A	95.5/94.5 (MNIST)	√	√
Dong et al.[20]	N/A	121.7 pJ/OP	4.1 ns	N/A	96 (fashion images) 94 (handwritten digits)	√	X
Taichi[62]	878.90TMACS/mm²	160.82TOPS/W	~5 ns	160	94.96/91.89 (Omniglot)	√	X
Dong et al.[63]	N/A	29fJ/OP(AWG)	1.33 ns	20	98.3 (spectral classification) 83.75 (MNIST)	√	X
Tianjic[64]	84.08 GOPS/mm²	1278 GOPS/W	~84.84 ms	N/A	96.48 (NMNIST)	N/A	√
NVIDIA H100[65]	N/A	0.15 TOPS/W	N/A	N/A	N/A	N/A	X

trainable model parameters, which supports the full deployment of a SNN layer. A core strength of this chip lies in its capability to execute both linear and nonlinear spike computations directly within the optical domain, eliminating the need for frequent optical-to-electrical conversion and thereby laying the foundation for high-efficiency computing.

To address the inherent challenges of SNN training and enhance the accuracy and robustness of photonic SNNs, we further developed a software-hardware collaborative training-inference framework. By introducing DCT for input data dimension reduction, we realized a high-performance lightweight frequency-domain SNN operating at a single time step (T=1). Hardware inference validated its effectiveness, achieving 90% accuracy on the MNIST dataset and 80.5% accuracy on the Fashion-MNIST dataset. The estimated energy efficiency is 1.39 TOPS/W for the MZI mesh and 987.65 GOPS/W for the DFB-SA spiking neuron array. Our work presents a lightweight SNN architecture with DCT that tailed for the fabricated photonic neuromorphic chip. Compared with the traditional photonic neural networks based on continuous-value activation, the PSNN employs binary activation values as inputs and outputs of the photonic chip. Thus, the lightweight PSNN architecture with DCT only requires low-bit AD/DA converters for input and output signals, and also exhibits superior noise robustness, which paves the way for the hardware deployment of photonic SNNs in fields of lightweight visual perception task, edge computation and embodied intelligence.

References

1. Wetzstein G, Ozcan A, Gigan S et al. Inference in artificial intelligence with deep optics and photonics. *Nature* **588**, 39–47 (2020).
2. Shastri BJ, Tait AN, Ferreira De Lima T et al. Photonics for artificial intelligence and neuromorphic computing. *Nat Photonics* **15**, 102–114 (2021).
3. Liao K, Chen Y, Yu ZC et al. All-optical computing based on convolutional neural networks. *Opto-Electron Adv* **4**, 200060 (2021).
4. Zhou HL, Dong JJ, Cheng JW et al. Photonic matrix multiplication lights up photonic accelerator and beyond. *Light Sci Appl* **11**, 30 (2022).
5. Huang CR, Sorger VJ, Miscuglio M et al. Prospects and applications of photonic neural networks. *Adv Phys X* **7**, 1981155 (2022).
6. Li CH, Du W, Huang YX et al. Photonic synapses with ultralow energy consumption for artificial visual perception and brain storage. *Opto-Electron Adv* **5**, 210069 (2022).
7. Zhao AK, Jiang N, Peng JF et al. Parallel generation of low-correlation wideband complex chaotic signals using CW laser and external-cavity laser with self-phase-modulated injection. *Opto-Electron Adv* **5**, 200026 (2022).
8. Fu TZ, Zhang JF, Sun R et al. Optical neural networks: progress and challenges. *Light Sci Appl* **13**, 263 (2024).
9. Zhou CD, Huang Y, Yang YG et al. Streamlined photonic reservoir computer with augmented memory capabilities. *Opto-Electron Adv* **8**, 240135 (2025).
10. Wang YZ, Chen MJ, Yao CH et al. Asymmetrical estimator for training encapsulated deep photonic neural networks. *Nat Commun* **16**, 2143 (2025).
11. Yan T, Guo YC, Zhou TK et al. A complete photonic integrated neuron for nonlinear all-optical computing. *Nat Comput Sci* **5**, 1202–1213 (2025).
12. Yang T, Zhang LY, Li SS et al. Spatiotemporal multiplexed photonic reservoir computing: parallel prediction for the high-dimensional dynamics of complex semiconductor laser network. *Opto-Electron Adv* **8**, 250159 (2025).
13. Shen YC, Harris NC, Skirlo S et al. Deep learning with coherent nanophotonic circuits. *Nat Photonics* **11**, 441–446 (2017).
14. Tian Y, Zhao Y, Liu SP et al. Scalable and compact photonic neural chip with low learning-capability-loss. *Nanophotonics* **11**, 329–344

(2022).
15. Ashtiani F, Geers AJ, Aflatouni F. An on-chip photonic deep neural network for image classification. *Nature* **606**, 501–506 (2022).
16. Peng B, Hua SY, Su Z et al. A 64×64 integrated photonic accelerator. In *2022 IEEE Photonics Conference (IPC)* 1–2 (IEEE, 2022). http://doi.org/10.1109/IPC53466.2022.9975501.
17. Bandyopadhyay S, Sludds A, Krastanov S et al. Single-chip photonic deep neural network with forward-only training. *Nat Photonics* **18**, 1335–1343 (2024).
18. Wan YJ, Liu XD, Wu GZ et al. Efficient stochastic parallel gradient descent training for on-chip optical processor. *Opto-Electron Adv* **7**, 230182 (2024).
19. Xue ZW, Zhou TK, Xu ZH et al. Fully forward mode training for optical neural networks. *Nature* **632**, 280–286 (2024).
20. Wu B, Huang CR, Zhang JL et al. Scaling up for end-to-end on-chip photonic neural network inference. *Light Sci Appl* **14**, 328 (2025).
21. Ouyang JY, Liu SP, Yang ZY et al. 16-channel photonic solver for optimization problems on a silicon chip. *Chip* **4**, 100117 (2025).
22. Hua SY, Divita E, Yu SS et al. An integrated large-scale photonic accelerator with ultralow latency. *Nature* **640**, 361–367 (2025).
23. Ahmed SR, Baghdadi R, Bernadskiy M et al. Universal photonic artificial intelligence acceleration. *Nature* **640**, 368–374 (2025).
24. Tait AN, De Lima TF, Zhou E et al. Neuromorphic photonic networks using silicon photonic weight banks. *Sci Rep* **7**, 7430 (2017).
25. Ohno S, Tang R, Toprasertpong K et al. Si microring resonator crossbar array for on-chip inference and training of the optical neural network. *ACS Photonics* **9**, 2614–2622 (2022).
26. Xu TJ, Zhang WP, Zhang JW et al. Control-free and efficient integrated photonic neural networks via hardware-aware training and pruning. *Optica* **11**, 1039–1049 (2024).
27. Jiang S, Liu XY, Wu B et al. MRR-assisted MZI crossbar array for energy-efficient optical parallel computing. *Laser Photonics Rev* **19**, e01035 (2025).
28. Feldmann J, Youngblood N, Wright CD et al. All-optical spiking neurosynaptic networks with self-learning capabilities. *Nature* **569**, 208–214 (2019).
29. Zhou W, Dong BW, Farmakidis N et al. In-memory photonic dot-product engine with electrically programmable weight banks. *Nat Commun* **14**, 2887 (2023).
30. Wei ML, Xu K, Tang B et al. Monolithic back-end-of-line integration of phase change materials into foundry-manufactured silicon photonics. *Nat Commun* **15**, 2786 (2024).
31. Nahmias MA, Shastri BJ, Tait AN et al. A leaky integrate-and-fire laser neuron for ultrafast cognitive computing. *IEEE J Sel Top Quantum Electron* **19**, 1800212 (2013).
32. Prucnal PR, Shastri BJ, de Lima TF et al. Recent progress in semiconductor excitable lasers for photonic spike processing. *Adv Opt Photonics* **8**, 228–299 (2016).
33. Peng HT, Angelatos G, de Lima TF et al. Temporal information processing with an integrated laser neuron. *IEEE J Sel Top Quantum Electron* **26**, 5100209 (2020).
34. Guo XH, Xiang JL, Zhang YJ et al. Integrated neuromorphic photonics: synapses, neurons, and neural networks. *Adv Photonics Res* **2**, 2000212 (2021).
35. Xiang SY, Ren ZX, Song ZW et al. Computing primitive of fully VCSEL-based all-optical spiking neural network for supervised learning and pattern classification. *IEEE Trans Neural Netw Learn Syst* **32**, 2494–2505 (2021).
36. Zhang YN, Xiang SY, Yu CY et al. Photonic neuromorphic pattern recognition with a spiking DFB-SA laser subject to incoherent optical injection. *Laser Photonics Rev* **19**, 2400482 (2025).
37. Guo XX, Song ZW, Xiang SY et al. Four-channel full-function photonic spiking neural network chips for gene analysis. *Laser Photonics Rev* **19**, e00864 (2025).
38. Fang MYS, Manipatruni S, Wierzynski C et al. Design of optical neural networks with component imprecisions. *Opt Express* **27**, 14009–14029 (2019).
39. Dai F, Chen YW, Huang ZY et al. Comparing the performance of multi-layer perceptron training on electrical and optical network-on-chips. *J Supercomput* **79**, 10725–10746 (2023).
40. Feldmann J et al. Parallel convolutional processing using an integrated photonic tensor core. *Nature* **589**, 52–58 (2021).
41. Xu XY et al. 11 TOPS photonic convolutional accelerator for optical neural networks. *Nature* **589**, 44–51 (2021).
42. Meng XY et al. Compact optical convolution processing unit based on multimode interference. *Nat Commun* **14**, 3000 (2023).
43. Bai BW et al. Microcomb-based integrated photonic processing unit. *Nat Commun* **14**, 66 (2023).
44. Xu SF et al. Optical coherent dot-product chip for sophisticated deep learning regression. *Light Sci Appl* **10**, 221 (2021).
45. Dong BW et al. Higher-dimensional processing using a photonic tensor core with continuous-time data. *Nat Photonics* **17**, 1080–1088 (2023).
46. Xu SF, Wang J, Yi SC et al. High-order tensor flow processing using integrated photonic circuits. *Nat Commun* **13**, 7970 (2022).
47. Lin ZJ, Shastri BJ, Yu SX et al. 120 GOPS photonic tensor core in thin-film lithium niobate for inference and in situ training. *Nat Commun* **15**, 9081 (2024).
48. Ning SP, Zhu HQ, Feng CH et al. Hardware-efficient photonic tensor core: accelerating deep neural networks with structured compression. *Optica* **12**, 1079–1089 (2025).
49. Lin X, Rivenson Y, Yardimci NT et al. All-optical machine learning using diffractive deep neural networks. *Science* **361**, 1004–1008 (2018).
50. Zhu HH, Zou J, Zhang H et al. Space-efficient optical computing with an integrated chip diffractive neural network. *Nat Commun* **13**, 1044 (2022).
51. Fu TZ, Zang YB, Huang YY et al. Photonic machine learning with on-chip diffractive optics. *Nat Commun* **14**, 70 (2023).
52. Zheng ZY, Duan ZY, Chen H et al. Dual adaptive training of photonic neural networks. *Nat Mach Intell* **5**, 1119–1129 (2023).
53. Cheng JW, Huang CR, Zhang JL et al. Multimodal deep learning using on-chip diffractive optics with in situ training capability. *Nat Commun* **15**, 6189 (2024).
54. Chen HJ, Lou SZ, Wang Q et al. Diffractive deep neural networks: theories, optimization, and applications. *App Phys Rev* **11**, 021332 (2024).
55. Zhang YF, Liu XB, Yang CG et al. Direct tensor processing with coherent light. *Nat Photonics* **20**, 102–108 (2026).
56. Jha A, Huang CR, Peng HT et al. Photonic spiking neural networks and graphene-on-silicon spiking neurons. *J Lightwave Technol* **40**, 2901–2914 (2022).
57. Xiang SY, Shi YC, Guo XX et al. Hardware-algorithm collaborative computing with photonic spiking neuron chip based on an integrated Fabry–Perot laser with a saturable absorber. *Optica* **10**, 162–171 (2023).
58. Xiang SY, Shi YC, Zhang YH et al. Photonic integrated neurosynaptic core for convolutional spiking neural network. *Opto-Electron Adv* **6**, 230140 (2023).
59. Xiang SY, Han YN, Gao S et al. Semiconductor lasers for photonic neuromorphic computing and photonic spiking neural networks: a perspective. *APL Photonics* **9**, 070903 (2024).
60. Xiang SY, Chen YH, Zhao HW et al. Nonlinear photonic neuromorphic chips for spiking reinforcement learning. *Optica* **13**, 457–468 (2026).
61. Britanak V, Yip PC, Rao KR. *Discrete Cosine and Sine Transforms* (Amsterdam: Academic, 2007).
62. Xu ZH, Zhou TK, Ma MZ et al. Large-scale photonic chiplet Taichi empowers 160-TOPS/W artificial general intelligence. *Science* **384**, 202–209 (2024).
63. Zhang SJ, Jiang XY, Wu B et al. Photonic edge intelligence chip for

multi-modal sensing, inference and learning. *Nat Commun* **16**, 10136 (2025).
64. Pei J, Deng L, Song S et al. Towards artificial general intelligence with hybrid Tianjic chip architecture. *Nature* **572**, 106–111 (2019).
65. NVIDIA. NVIDIA H100 Tensor Core GPU Datasheet. (2024). [Online]. Available: https://resources.nvidia.com/en-us-tensor-core/nvidia-tensor-core-gpu-datasheet.
66. Zhang Y, Sun ZX, Xiao RL et al. Experimental demonstration of 150-channel REC-DFB laser array based on asymmetric multiple-quantum-well. *IEEE J Quantum Electron* **61**, 2200307 (2025).
67. Ren H, Feng YX, Zhou S et al. All-optical DCT encoding and information compression based on diffraction neural network. *ACS Photonics* **12**, 1196–1211 (2025).

Acknowledgements

We would like to thank Prof. Guoqi Li, Lei Deng, Jian Wang, and Jianji Dong for their helpful discussions. We also thank Prof. Jianji Dong and Hailong Zhou for fabricating the MZI mesh chip. We are also grateful for financial supports from the National Natural Science Foundation of China (No.62535015, 62575231); National Key Research and Development Program of China (2021YFB2801900, 2021YFB2801901, 2021YFB2801902, 2021YFB2801903, 2021YFB2801904); The Fundamental Research Funds for the Centrale Universities (QTZX23041).

Author contributions

S. Y. Xiang and Y. H. Zhang are co-first authors of the article. S. Y. Xiang proposed the idea, S. Y. Xiang, Y. H. Zhang prepared the manuscript, S. X. Shi, H. W. Zhao, D. Z. Zheng performed system experiments, Y. H. Zhang, X. X. Guo tested the chips, Y. N. Han, Y. Tian, L. Y. Zhang developed the algorithms, Y. C. Shi fabricated the samples, S. Y. Xiang, Y. Hao supervised the overall projects. All the authors analyzed and discussed the results.

Competing interests

The authors declare no competing financial interests.

Data availability

The example dataset used in this paper are publicly available and can be accessed at https://yann.lecun.com/exdb/mnist/ for MNIST dataset, and https://github.com/zalandoresearch/fashion-mnist for fashion-MNIST dataset. Other data that support the findings of this study are available from the corresponding author upon reasonable request.