Research Article

ADVANCED PHOTONICS

Time-wavelength multiplexed photonic neural network accelerator for distributed acoustic sensing systems

Fuhao Yu,[a,b,†] Kangjian Di,[c,†] Wenjun Chen,[c] Sen Yan,[a,b] Yuanyuan Yao,[a,b] Silin Chen,[c] Xuping Zhang,[a,b] Yixin Zhang,[a,b,*] Ningmu Zou,[b,c,d,*] and Wei Jiang[a,b,*]

[a]Nanjing University, College of Engineering and Applied Sciences, Nanjing, China
[b]Nanjing University, Ministry of Education Key Laboratory of Intelligent Optical Sensing and Manipulation, Nanjing, China
[c]Nanjing University, School of Integrated Circuits, Suzhou, China
[d]Nanjing University, Interdisciplinary Research Center for Future Intelligent Chips (Chip-X), Suzhou, China

Abstract. Distributed acoustic sensors (DASs) can effectively monitor acoustic fields along sensing fibers with high sensitivity and high response speed. However, their data processing is limited by the performance of electronic signal processing, hindering real-time applications. The time-wavelength multiplexed photonic neural network accelerator (TWM-PNNA), which uses photons instead of electrons for operations, significantly enhances processing speed and energy efficiency. Therefore, we explore the feasibility of applying TWM-PNNA to DAS systems. We first discuss processing large DAS system data for compatibility with the TWM-PNNA system. We also investigate the effects of chirp on optical convolution in complex tasks and methods to mitigate its impact on classification accuracy. Furthermore, we propose a method for achieving an optical full connection and study the influence of pruning on the full connection to reduce the computational burden of the model. Experimental results indicate that decreasing the ratio of $\Delta\lambda_{\text{chirp}}/\Delta\lambda$ or choosing push–pull modulation can eliminate the impact of chirp on recognition accuracy. In addition, when the full connection parameter retention rate is no less than 60%, it can still maintain a classification accuracy of over 90%. TWM-PNNA provides an innovative computational framework for DAS systems, paving the way for the all-optical fusion of DAS systems with computational systems.

Keywords: optical computing; distributed acoustic sensors; time-wavelength multiplexing; optical full connection; chirp.

Received Nov. 4, 2024; revised manuscript received Feb. 3, 2025; accepted for publication Feb. 13, 2025; published online Mar. 17, 2025.

[DOI: 10.1117/1.AP.7.2.026008]

1 Introduction

Fiber-optic distributed acoustic sensors (DASs) are at the forefront of fiber sensing measurement technology. In addition to the characteristics of traditional distributed fiber-optic sensing systems, such as resistance to electromagnetic interference, good concealment, corrosion resistance, and long detection distances, DAS can also achieve real-time detection of dynamic strain with high sensitivity and rapid response. This technology shows significant advantages and application prospects in major infrastructures, including seismic wave monitoring,[1] oil and gas resource exploration,[2] submarine cable monitoring,[3] railway traffic operation monitoring,[4] and overhead transmission lines.[5]

However, environmental factors and benign human interference can lead to false alarms and missed detections, and the high rates of false alarms and missed detections have been a bottleneck limiting its performance in field applications. In recent years, with the continuous advancements in machine-learning and deep-learning technologies, several algorithms have emerged that can autonomously summarize patterns by analyzing large data sets and flexibly adjusting their parameters, achieving high-precision event recognition in a short time. Various event

*Address all correspondence to Yixin Zhang, zyixin@nju.edu.cn; Ningmu Zou, nzou@nju.edu.cn; Wei Jiang, weijiang@nju.edu.cn

†These authors contributed equally to this work.

recognition techniques for DAS systems have been proposed: morphological feature extraction methods have achieved over 90% accuracy in classifying three types of disturbances;[6] wavelet energy spectrum analysis combined with relevance vector machines has achieved an 88.6% recognition rate;[7] attention mechanism-based LSTM methods have achieved a 94.3% recognition rate among five types of disturbances;[8] convolutional neural networks (CNNs) combined with bandpass filtering and gray-scale conversion preprocessing have achieved an accuracy of 96.67%;[9] artificial neural networks (ANNs) can improve the linearity of strain noise and underground motion measurements;[10] and Markov transition fields and nonnegative matrix factorization methods can enhance the average recognition rate of fence events by over 13%.[11] In addition, transfer learning methods can achieve an accuracy of 95.56% with low computational costs, even with small sample training.[12]

Currently, the ANN employed in the aforementioned DAS systems primarily relies on central processing units. Although GPUs offer significant advantages in processing demodulated DAS data, they cannot overcome the limitations of electronic computing on DAS systems. The emergence of optical computing technology offers a potential solution to these challenges. It utilizes photons instead of electrons to perform computational operations, enabling processing speeds that far exceed those of traditional electronic computing. Furthermore, optical computing can achieve high parallelism through wavelength multiplexing and offer higher energy efficiency. Therefore, this article proposes the integration of optical computing systems and DAS systems and emphasizes the use of optical computing to achieve co-GPU functionality and process-demodulated DAS data, providing a new direction for postprocessing DAS system data.

Photonic neural networks can be categorized into three types: those based on on-chip coherent principles, those based on spatial optical structures, and those utilizing time-wavelength multiplexing (TWM) technology. Neural networks based on on-chip coherent principles mainly rely on topologically cascaded Mach–Zehnder interferometer (MZI) arrays, employing coherent light and matrix singular value decomposition to achieve an integrated photonic neural network architecture on a chip. With interference from factors such as detector noise and thermal cross talk, errors accumulate during the MZI cascading process, resulting in an accuracy of only 76.7% in voice recognition tasks involving four vowels.[13] In addition, a phase control scheme has been proposed to enhance this architecture, enabling the realization of complex photonic neural networks that can improve handwritten digit recognition accuracy to 90.5%.[14] Nonetheless, challenges related to fabrication errors[15] and uneven distribution of rotating rotors[16] hinder the large-scale integration of photonic neural networks based on on-chip coherent principles. Neural networks based on spatial optical structures can be implemented using various spatial optical devices. For example, a fully optical deep-learning framework that uses only optical diffraction and passive optical components for machine-learning tasks has been proposed and applied to handwritten recognition tasks.[17] For further integration, a novel on-chip diffractive optical neural network architecture based on a silicon-on-insulator platform has been proposed, featuring high integration and low power consumption, which allows for low-cost mass production.[18] Although this approach improves integration, classification accuracy and generalization to other tasks need further enhancement. Optical neural networks based on TWM technology offer a parallel optical neural network that does not require matrix decomposition and can directly correspond wavelengths to matrix elements, enabling large-scale scalability. For instance, a photonic convolution accelerator based on an optical frequency comb has been proposed, utilizing wavelength division multiplexing (WDM) technology and fiber dispersion delay to achieve large-scale parallel convolution, achieving over 10 trillions of operations per second (TOPS) of computational power while recognizing handwritten digits with an accuracy of 88%.[19] Furthermore, a fully integrated photonic processing unit containing a microcomb source, silicon electro-optic modulator, microring weight library, and optical delay line has been proposed for convolution operations, realizing image edge detection and handwritten digit recognition.[20] To enhance computational capability, a scheme combining TWM and multimode interference coupling has been proposed, supporting three 2×2 real-valued kernels and enabling parallel convolution operations.[21] In addition, TWM photonic neural networks have been used to process high-order tensors,[22,23] achieving promising results.

Advanced photonic neural networks have made significant progress in handling classification tasks, offering a promising alternative to traditional electronic neural networks. This paper proposes a time-wavelength multiplexed photonic neural network accelerator (TWM-PNNA) system, aimed at achieving optical convolution and optical fully connected computations while exploring the potential applications of TWM-PNNA in DAS systems. The system employs optical technologies to enhance computational efficiency, supporting effective data processing and feature extraction. As the data collected by DAS contain spatial and temporal dimensions and are large in volume, real-time processing is challenging; thus, the issue of adapting DAS data to the TWM-PNNA system needs to be addressed. In addition, modulation chirp can cause pulse broadening and intersymbol interference, necessitating solutions for quantifying and mitigating the effects of modulation chirp. Furthermore, the fully connected layers have numerous parameters and complex computations, making it essential to explore how to reduce the model size through pruning and achieve real-time, high-speed optical fully connected operations. This paper investigates the related research on these three issues.

2 Principles

2.1 Data Acquisition and Processing of a DAS System

DAS is a novel distributed optical fiber sensing technology using coherent detection on the basis of Φ-OTDR. The DAS we designed acquires the phase information of Rayleigh scattering light through spatial differential interferometry technology to achieve the reconstruction of external vibration signals.[24] We employ a DAS system based on heterodyne coherent detection to collect experimental data,[5,25–28] as shown in Fig. 1(a). The narrow linewidth laser (NLL) with a 3-kHz frequency width operating at 1550.12 nm was selected as the light source. The output of the NLL was split into two components, at 80% and 20% as the probe light and the local reference light by an optical coupler, respectively. The probe light is modulated by an acousto-optic modulator (AOM) into a probe pulse with a 150-MHz frequency shift. The probe pulse with a pulse width of 100 ns and repetition rate of 2 kHz was amplified in an erbium-doped fiber amplifier (EDFA), and then, it was injected into the sensing fiber through a circulator. The Rayleigh backscattering (RBS) light returning from the sensing fiber is mixed with the local reference light.

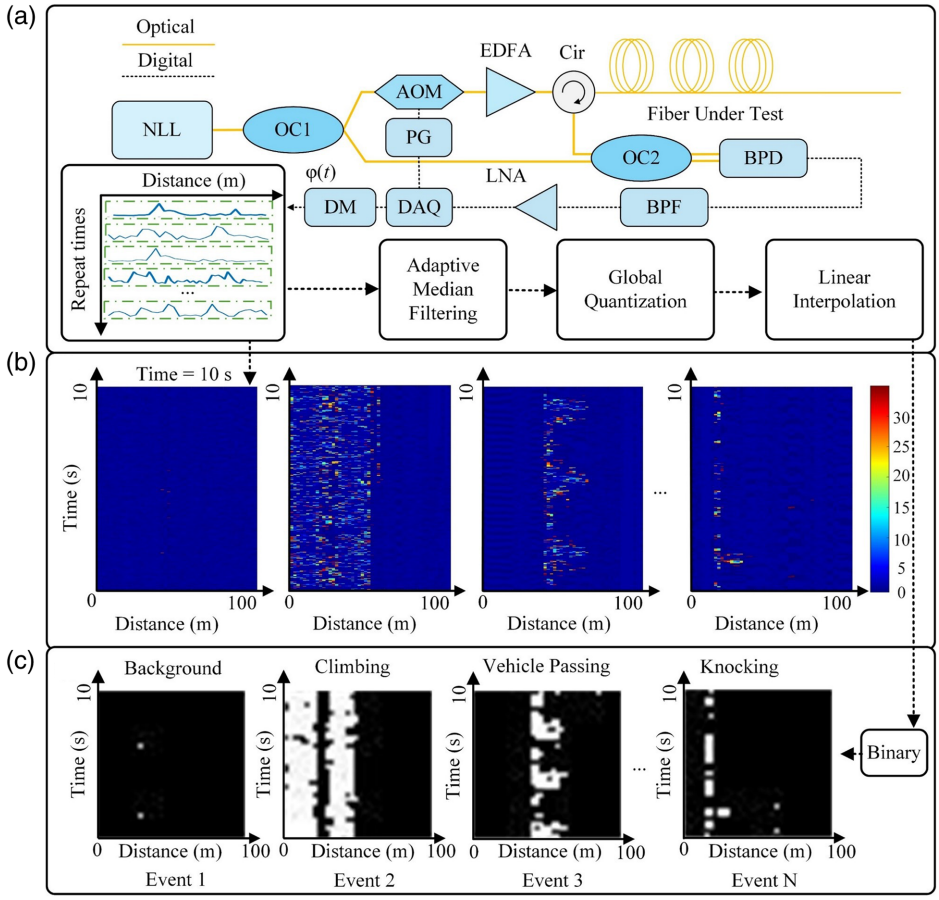

Fig. 1 (a) Setup of DAS system; (b) raw signals collected by the system; (c) spatiotemporal maps of various events after denoising and quantization. NLL, narrow linewidth laser; OC, optical coupler; AOM, acousto-optic modulator; DAQ, digital acquisition card; Cir, circulator; EDFA, erbium-doped fiber amplifier; BPD, balanced photodetector; PG, pulse generator; BPF, bandpass filter; LNA, low-noise amplifier; DM, demodulation module.

The mixed signal is detected by a balanced photodetector (BPD) with a 200-MHz bandwidth. The signals from coupler2 enter the photodetectors (PDs). Then, an 8-bit data acquisition (DAQ) card continuously sampled the output data with a 1-GHz sampling rate. The phase demodulation was completed by an IQ demodulation module (DM). When external vibrations induce fluctuations in the backward Rayleigh light power at specific points along the sensing fiber over time, the variations in the interference signal can be analyzed to facilitate the reconstruction of vibration information.[29] Finally, the two-dimensional (2D) time-space matrix can be obtained by accumulating the temporal responses along the spatial axis.

The raw signals collected by the system are shown in Fig. 1(b). During the 10-s sensing period, various disturbance events were recorded. Due to the influence of system and environmental noise, directly distinguishing different disturbance characteristics becomes challenging. Therefore, we employ an adaptive median filter to reduce noise interference in the sensing data and perform global quantization processing. As a result of the large amount of redundant information in the denoised data, which increases the burden of subsequent processing, we use linear interpolation to compress the data into a 36×36 image while preserving the original image features and distance information. After processing, different types of disturbances still exhibit their unique 2D spatiotemporal characteristics, including background noise, climbing, vehicle passage, and tapping events, as shown in Fig. 1(c). In addition, considering that the absolute value of phase-amplitude information below the threshold in Fig. 1(b) is all background information, the threshold is set to 7.35, and the compressed image is binarized, as shown in Fig. 1(c). After collecting a large amount of experimental data, we reconstructed the data set and randomly split it into training and testing sets at a ratio of 8:2, with detailed information provided in Table 1.

Table 1 Database construction for vibration events.

Event type	Train	Test	Total	Label
Event 1 (background)	826	206	1032	0
Event 2 (climbing)	676	170	846	1
Event 3 (vehicle passing)	712	178	890	2
Event 4 (knocking)	757	189	946	3
…	…	…	…	…
Event *N*	*A*	*B*	*A* + *B*	*N* − 1

2.2 Photonic Neural Network Accelerator

In CNN, convolution operations are performed on images to obtain feature maps. Each point in the feature map is the result of multiplying and summing the values of the convolution kernel with the corresponding pixel values at that position. The convolution kernel slides over the image with a specified stride, ultimately forming a feature map. As the convolution kernel remains unchanged during this sliding process, a single kernel can only extract one feature map. To obtain multiple feature maps, multiple convolution kernels are required. To facilitate the implementation of this process in optical systems, it is necessary to flatten the 2D image data into one-dimensional (1D) data. The flattening method and the principles of photonic convolution operations are illustrated in Fig. 2. First, l^2 groups of 1D data are combined to form an image matrix $\mathbf{X} = [X_{\mathrm{CON}11}, X_{\mathrm{CON}12}, \ldots, X_{\mathrm{CON}1l^2}]$. Simultaneously, the $l \times l$ convolution kernel is flattened into a vector $\mathbf{W} = [w_{11}, w_{12}, \ldots, w_{21}, \ldots, w_{ll}]$ containing l^2 elements and multiplied with the image matrix X to obtain the convolution results $[X_{\mathrm{CON}1}, X_{\mathrm{CON}2}, \ldots, X_{\mathrm{CON}K}]$. This process is akin to assigning weights to the image matrix. Subsequently, each resulting column vector is delayed by one symbol, and the corresponding values of the row vectors are summed to obtain 1D feature information. The above operations can be repeated for different convolution kernels. After passing through the pooling layer and nonlinear activation function in the electronic processing, the feature map can yield the underlying information and features of the image. Subsequently, all the information is consolidated in the fully connected layer. Moreover, the number of output classes in the output layer of the fully connected layer is set according to the data categories in the DAS database. Each neuron in the output layer is connected to all neurons in the previous layer with specific connection weights. Therefore, the fully connected input data will be multiplied by N different weight groups and added together, and the result with the highest probability output is the classification result, as shown in Fig. 2.

In the specific experiment, we used the setup shown in Fig. 3. n independent tunable lasers emit light at n wavelengths, spaced by $\Delta\lambda$, which constitute the $\sqrt{n} \times \sqrt{n}$ convolution kernels and are combined into a single beam after passing through a WDM. The effectiveness of WDM in enhancing data transmission, system capacity, and signal integrity has been extensively demonstrated in prior studies.[30,31] This beam then enters an optical switch (OSW), output from port A of the switch into a Mach–Zehnder modulator (MZM, Fujitsu, FTM7937EZ, bandwidth 30 GHz). A 36×36 pixels binarized image obtained from the DAS database is flattened into a 1D vector $\boldsymbol{X}$ and encoded into an arbitrary waveform generator (AWG, Tektronix, AWG7000B). The modulator working at the quadrature point encodes this into an optical time-domain signal, with each pixel represented by one bit of the modulated signal. The AWG transmits the signal at a certain baud rate. At this point, vector $\boldsymbol{X}$ is simultaneously modulated onto all n wavelengths. The output vector is obtained by detecting the signals in each time slot. A single-mode fiber (SMF) of a certain length provides progressive delays for each channel to match the baud rate of

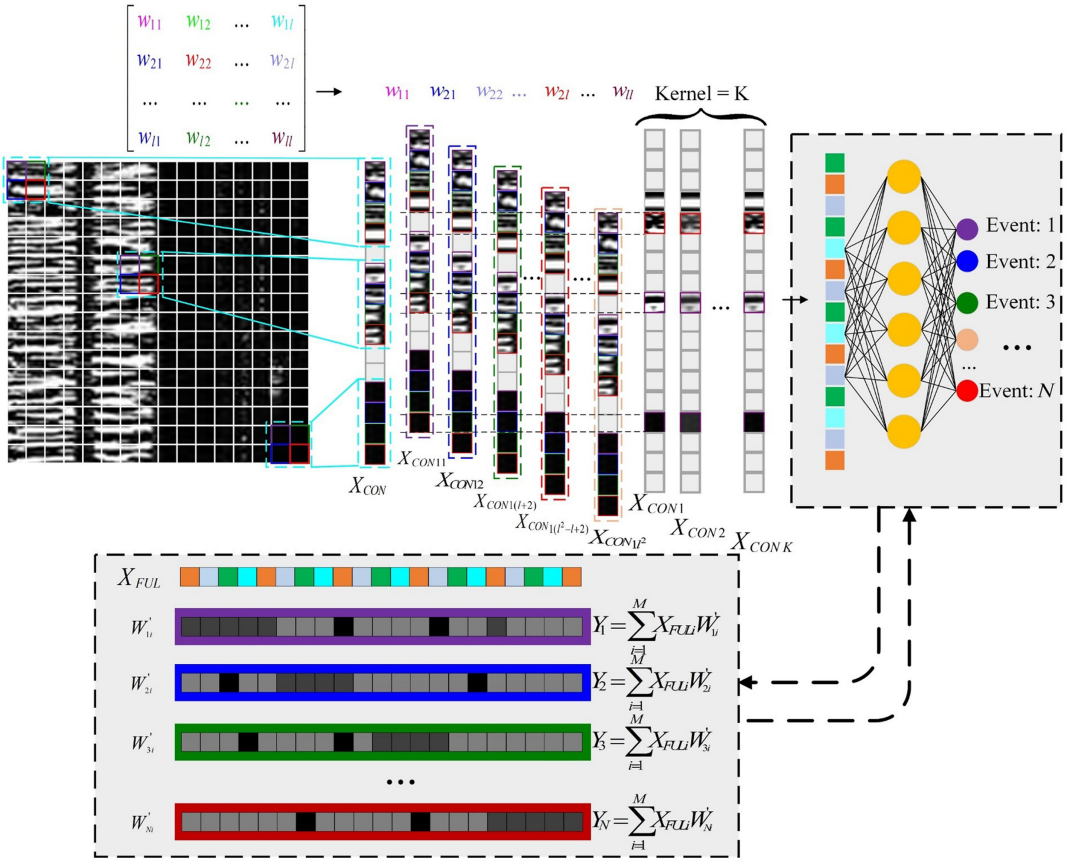

Fig. 2 Photonic neural network accelerator.

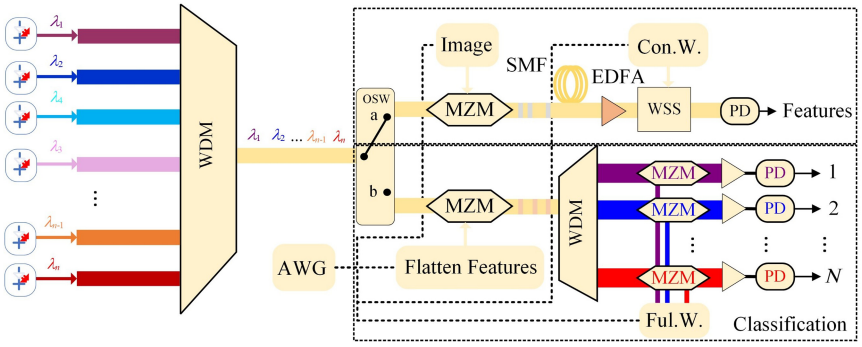

Fig. 3 Experimental setup of the TWM-PNNA. AWG, arbitrary waveform generator; MZM, Mach–Zehnder modulator; WSS, wavelength-selective switch; OSW, optical switch; WDM, wavelength-division multiplexer; PD, photodetector.

the data emitted by the AWG, ensuring that the signals on adjacent wavelength channels are time-shifted by the same number of symbol positions. In addition to SMF, as long as the delay time matches the AWG rate, a waveshaper can also be used for dispersion. Next, an EDFA is used to compensate for the insertion loss of the modulator and the loss of light of different wavelengths after passing through the SMF. The amplified wavelength channels are then shaped by a wavelength-selective switch (WSS, CoAdna, 50 GHz Nx1-1.2). WSS can be seen as a combination of WDM, Mux/Demux, and filters, providing flexible wavelength selection and routing functions while controlling the attenuation of each wavelength channel. This attenuation represents the weight information applied to each wavelength channel, allowing the weight value W_i to be assigned to wavelength λ_i. Subsequently, the output light enters a high-speed PD (Finisar, XPDV21x0, bandwidth 40 GHz), which aggregates the total optical power at each wavelength. In addition, the WSS can realize different convolution kernels by reconfiguring the routing and attenuation of different wavelength channels. Finally, the electrical output waveform after the photonic convolution will be sampled and digitized by a high-speed oscilloscope (OSC, Tek, DPO75902SX, bandwidth 70 GHz) to obtain the feature map. The optical switch is then adjusted so that the laser light emitted from the lasers enters through port B of the switch. The flattened feature map is also loaded into the optical path through the AWG and MZM. The output light from the modulator passes through the WDM, filtering the laser light so that each output channel contains only one wavelength. Furthermore, each output channel of the WDM connects to an MZM to load different weight parameters, thereby achieving multiplication operations. The output light then enters the PD and is collected by the OSC, completing the addition operations required for the fully connected layer (for details of the experimental parameters, see Note S7 in the Supplementary Material).

3 Pretraining of TWM-PNNA

A complete TWM-PNNA architecture includes optical convolution layers and optical fully connected layers. Therefore, in our demonstration, we first used all data sets in Table 1 to complete the pretraining of the CNN, thus obtaining the necessary parameters for each layer of the neural network. In addition, as discrete tunable lasers were chosen as the input light source to reduce experimental costs, we used four independent lasers, resulting in a convolution kernel size of 2×2. To simplify the model complexity, we set the number of convolution layers to 1 and the number of convolution kernels to 10. The more convolution kernels there are, the more features can be extracted within the same layer,[32] allowing for the capture of various patterns, textures, and edge information from the input data. It is important to note that as the WSS can only attenuate optical intensity information, negative convolution kernels cannot be implemented. Therefore, during pretraining, negative convolution kernels need to be trimmed in advance to ensure they are nonnegative while still achieving the classification task. Furthermore, to prevent information loss during the transfer through the neural network, we added a layer of blank pixels around the image, known as padding. Subsequently, the stride of the convolution kernel was set to 1, and the 2D image data were flattened into 1D data according to the size of the convolution kernel.

Subsequently, we loaded the 1D data into the optical signal using the AWG. As the 1D data only contain two states, 0 and 1, the output baud rate of the AWG is the same as the bit rate. To ensure that the signals on adjacent wavelength channels are time-shifted by one bit, the relationship among the AWG bit rate v_{bit}, the wavelength interval $\Delta\lambda$, and the length L of the SMF can be expressed as

$$t_{\text{bit}} = \frac{1}{v_{\text{bit}}} = \Delta\lambda \cdot L \cdot D, \tag{1}$$

where D is the dispersion coefficient of the SMF, $\sim 17\ \text{ps km}^{-1}\,\text{nm}^{-1}$. According to Eq. (1), it can be seen that the TWM-PNNA is insensitive to wavelength. The delay is determined by the wavelength interval, not the wavelength itself, which in turn affects the matrix multiplication and subsequent addition operations.[33] When we set the wavelength interval $\Delta\lambda$ to 2 nm and the AWG baud rate to 10 Gbaud, the required fiber length to achieve a symbol shift is 2.9 km. Subsequently, we used the WSS to set different weight parameters for the 10 convolution kernels, as shown in Figs. 4(b1)–4(b10). Using climbing events as an example, we present the results obtained after performing optical convolution operations through the aforementioned system, as shown in Figs. 4(a1)–4(a10). From the figure, it can be observed that multiplying the image data with different convolution kernels yields different data features as reflected in the power of the waveforms. For instance, when the convolution kernel is 0, as shown in Fig. 4(b9), the waveform power is close to 0, indicating no features. In addition, when the convolution kernel coefficients are larger, as shown in

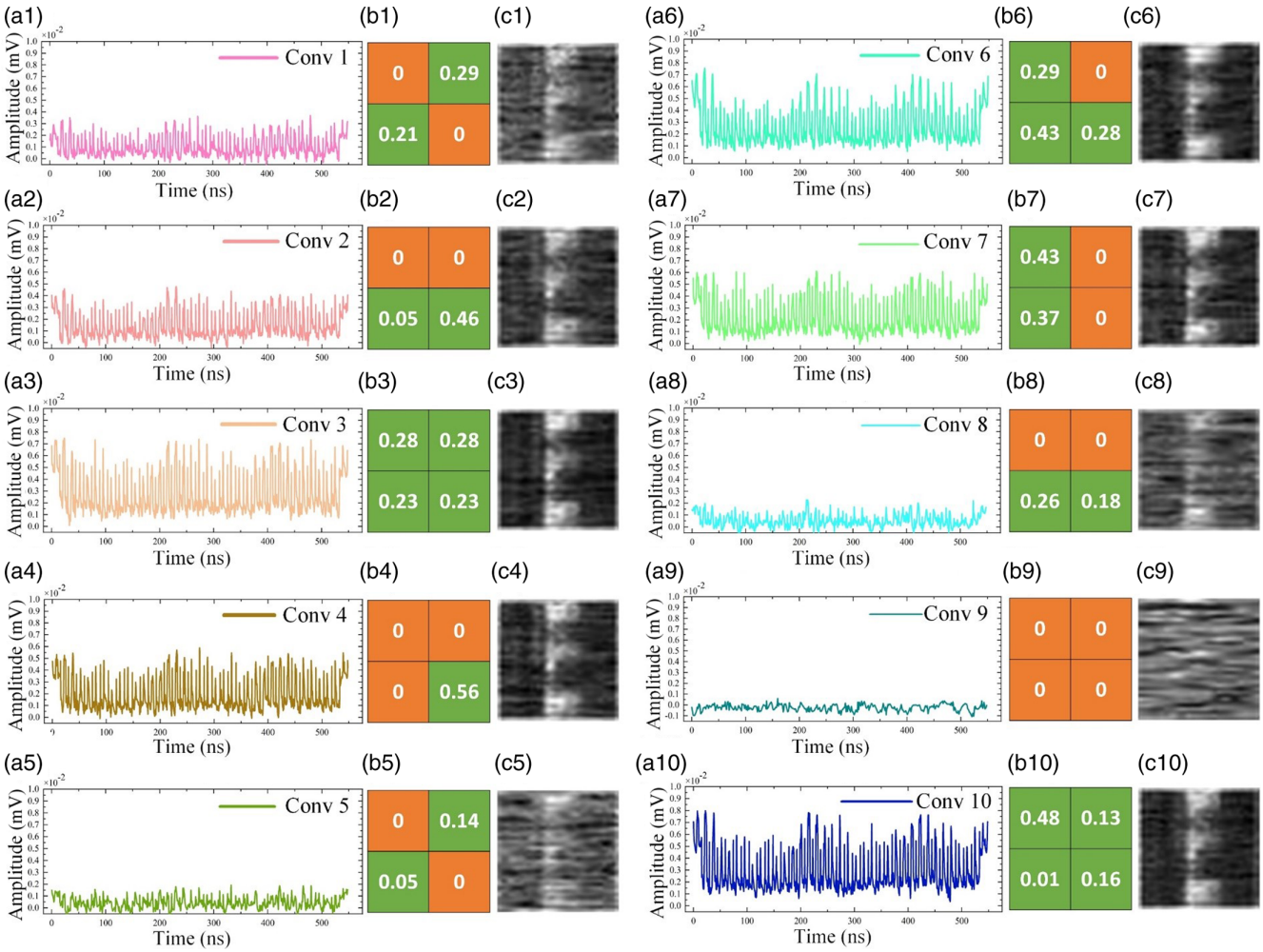

Fig. 4 Results of convolution operations in the TWM-PNNA. (a1)–(a10) 1D feature maps after optical convolution; (b1)–(b10) 10 different convolution kernels; (c1)–(c10) restored 2D feature maps.

Figs. 4(b6) and 4(b10), the power of the waveforms in the feature maps is greater and more pronounced. However, due to the presence of some noise in the system, increasing the number of convolution kernels not only allows for the extraction of more image features but also helps prevent the feature maps from those kernels with smaller coefficients from being drowned out by noise, thus affecting recognition accuracy. Next, we reverse the process of flattening the 2D image into 1D data to convert the 1D feature maps obtained after the optical convolution operation back into 2D feature maps, as shown in Figs. 4(c1)–4(c10). A detailed explanation of the results of photonics convolution for other events can be found in Note S1 in the Supplementary Material.

To independently verify the performance of the optical convolution accelerator in the TWM-PNNA, the subsequent pooling, nonlinear, and fully connected parts were implemented on a computer. The experimental results showed that the optical convolution accelerator achieved a classification accuracy of 95% on the test set, whereas the electronic CNN reached an accuracy of 98.3% on the test set. Owing to certain noise in the optical path, the accuracy of the optical convolution accelerator is slightly lower compared with the electronic CNN.

In addition, we experimentally verified the accuracy of the test and training sets for image sizes of 16×16, 32×32, 36×36, 64×64, and 128×128, as shown in Fig. 5(a). From the figure, it can be seen that when the image size is 16×16, the accuracy significantly decreases. This is due to excessive downsampling, which leads to the loss of image details and information, causing the low-resolution image to fail to accurately reflect the features of the original image. In addition, it can be seen from the table that the accuracies of the 32×32 image and the 128×128 image are relatively close, indicating that once the size of the image exceeds a certain range, the impact of image size on accuracy can be ignored (for details of the comparison of classification results, see Note S6 in the Supplementary Material). Furthermore, we studied the classification accuracy of binary and nonbinary images on the electrical neural network under different image sizes, as shown in Fig. 5(b). We can see that binarizing the DAS data image does not significantly reduce image accuracy, indicating that the information of the DAS data image is not lost after binarization (for details of the comparison of classification results and reliability and repeatability of results, see Notes S6 and S8 in the Supplementary Material).

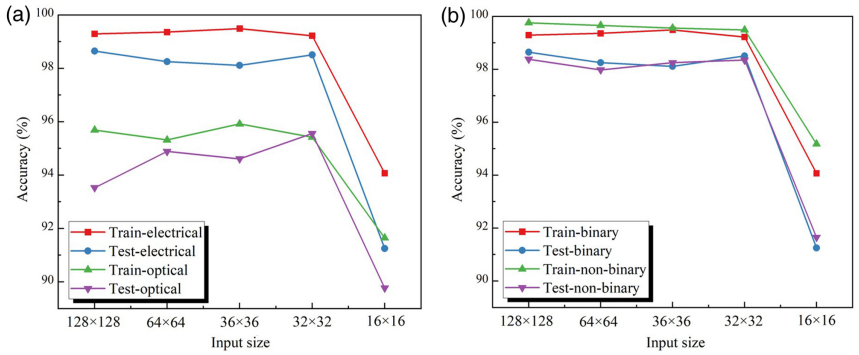

Fig. 5 Classification accuracy under different image sizes. (a) Electrical neural networks and optical neural networks in binary images; (b) electrical neural networks in binary images and nonbinary images.

4 Results and Analysis of the Impact of Chirp on TWM-PNNA

From the previous descriptions, it is evident that the wavelength spacing among different channels of the laser has a significant impact on the TWM optical computing system. However, as this system uses an MZM to achieve intensity modulation, its output expression during single-drive conditions can be represented as

$$|E_{\text{out}}|^2 = |E_{\text{in}}|^2 \cdot \frac{1}{2}\left(1 + \cos\left(\Phi_0 + \frac{\pi V_{\text{pp}}}{V_\pi}\right)\right), \tag{2}$$

where E_{out} is the output electric field intensity of the modulator, E_{in} is the input electric field intensity of the modulator, Φ_0 is the initial phase difference between the two arms of the MZM, V_{pp} is the driving voltage of the modulator, and V_π is the half-wave voltage of the modulator. From Eq. (2), it can be seen that as the output intensity of the modulator changes, its phase also changes. As $\Delta\omega = \mathrm{d}\varphi/\mathrm{d}t$, where $\Delta\omega$ is the frequency variation in the output signal, phase changes will produce modulation chirp. In addition, $\Delta\omega = (-2\pi c/\lambda^2)\Delta\lambda_{\text{chirp}}$, where c is the speed of light in vacuum, λ is the operating wavelength, and $\Delta\lambda_{\text{chirp}}$ is the wavelength variation caused by the chirp. Furthermore, $\Delta\lambda_{\text{chirp}}$ will cause pulse broadening, and the pulse broadening time can be expressed as $\Delta t_{\text{chirp}} = D \cdot L \cdot \Delta\lambda_{\text{chirp}}$. Moreover, when the two arms of the MZM operate in a push–pull state, meaning a relative RF signal $v_1'(t) = -v_2'(t)$ is applied between the two modulation arms, the output optical field can be expressed as

$$E_{\text{out}}(t) = E_{\text{in}}(t) \cdot \cos\left(\frac{\pi}{2V_\pi}(2v_{\text{in}}(t) + V_{\text{bias}})\right)\exp\left(j\frac{\pi(V_1+V_2)}{2V_\pi}\right), \tag{3}$$

where $v_{\text{in}}(t) = 2v_1'(t)$. V_1 and V_2 are the DC bias voltages for the two arms, and $V_{\text{bias}} = V_1 - V_2$ is the total DC bias signal. In this case, the output signal of the modulator retains a constant phase shift term, which does not change with variations in the RF signal, thus helping to reduce modulation chirp. In the actual experiment, the signal output from the AWG passes through an amplifier (SHF 804b 65 GHz) before entering the MZM. For single-drive modulation, the driving voltage is $V_{\text{pp}} = 3$ V; for push–pull modulation, the driving voltage is $V_{\text{pp1}} = -V_{\text{pp2}} = 1.5$ V (for details of noise control of push–pull modulation and single-drive modulation, see Note S3 in the Supplementary Material).

Based on the experimental results in Fig. 5, we set the image size to 36×36. To ensure that the data-partitioning method is consistent with common partitioning methods in the literature, such as Refs. 11, 34, and 35, we randomly selected 50 images from the training set of each category for training and 30 images from the testing set of each category for testing. In the TWM-PNNA system, for a single wavelength channel, the MZM can load electrical signals with intensities $(x_1, x_2, \ldots, x_{10})$ onto the optical domain. In the absence of modulation chirp, each bit sequence remains independent, as shown in Fig. 6(a). However, due to the interaction between modulation chirp and fiber dispersion, pulse broadening occurs.[36,37] When optical signals carrying different intensity information pass through the delayed optical fiber, interference arises between adjacent bits, resulting in intersymbol interference (ISI) within the channel, as illustrated in Fig. 6(b). This phenomenon degrades signal transmission quality and increases the bit error rate (BER). Furthermore, when the PD sums the power of signals from different

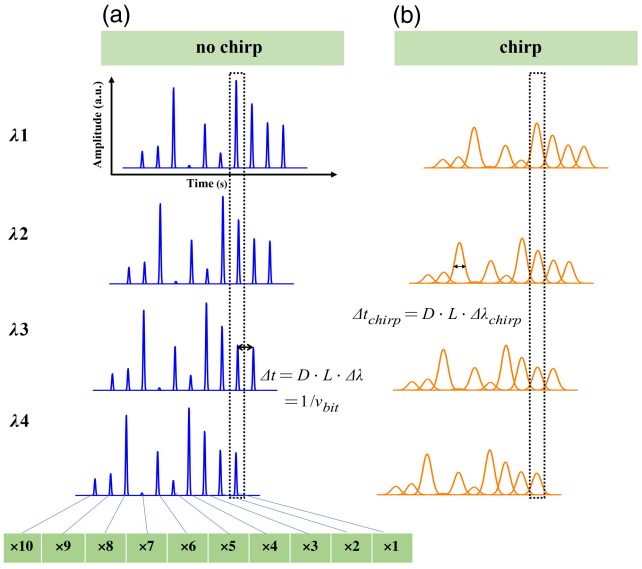

Fig. 6 Comparison of chirp effects. (a) Without chirp. (b) With chirp.

wavelengths, modulation chirp amplifies signal distortion, exacerbating its negative impact on the computational accuracy of convolution kernels in the TWM-PNNA system, as indicated by the black dashed box in Fig. 6. In addition to pulse broadening, modulation chirp also causes spectral broadening of the signal, significantly increasing the probability of spectral overlap between adjacent channels in WDM systems.[38] This spectral overlap can increase interchannel cross talk, reducing the signal-to-noise ratio. These combined effects significantly impact the computation accuracy and BER performance of the system, as reflected in the experimental results.

Furthermore, according to Eq. (1), although the time interval Δt between each bit within each wavelength channel is directly determined by the AWG rate, there exists a certain quantitative relationship with the wavelength interval $\Delta\lambda$. There is also a quantitative relationship between the pulse broadening Δt_{chirp} and the wavelength change $\Delta\lambda_{\text{chirp}}$ caused by chirp. As long as the optical fiber length matches the AWG rate, we can obtain this relationship $\Delta t_{\text{chirp}}/\Delta t = \Delta\lambda_{\text{chirp}}/\Delta\lambda$, allowing us to ignore the effect of fiber length. Therefore, we only need to calculate $\Delta\lambda_{\text{chirp}}/\Delta\lambda$ to measure the impact of modulation chirp on the TWM-PNNA.

To further investigate the impact of modulation chirp on the TWM-PNNA, we reduced the wavelength spacing. Given that the minimum wavelength spacing of the used WSS is 0.4 nm, we set the wavelength spacing of the lasers to $\Delta\lambda = 0.4$ nm to match the WSS. At the same time, we adjusted the AWG rate to 20 Gbps. In this case, the interval between each bit is 50 ps. In addition, according to Eq. (1), we chose the length of SMF to be 7.35 km to ensure signal matching across different wavelength channels. Moreover, the corresponding calculations indicate that the signal bandwidth is ~14 GHz, corresponding to $\Delta\lambda_{\text{chirp}} = 0.11$ nm. Thus, $\Delta\lambda_{\text{chirp}}/\Delta\lambda = 0.275$. To verify the effect of modulation chirp on the experimental results, the subsequent pooling, nonlinear, and fully connected parts were implemented on a computer, consistent with previous experiments.

The experimental results indicate that under these conditions, when the modulator operates in single-drive modulation, the classification accuracy is only 77%, as shown in the confusion matrix in Fig. 7(e). By contrast, when the modulator operates in push–pull conditions, the classification accuracy significantly improves to 90.8%, as shown in Fig. 7(d). When the modulation rate is 20 Gbps and the wavelength spacing is 2 nm, the corresponding $\Delta\lambda_{\text{chirp}}/\Delta\lambda$ is 0.055, with the confusion matrices for both push–pull and single-drive modulation shown in Figs. 7(b) and 7(c). From the figures, we can see that both modulation methods achieve high accuracy in classification tasks, with a

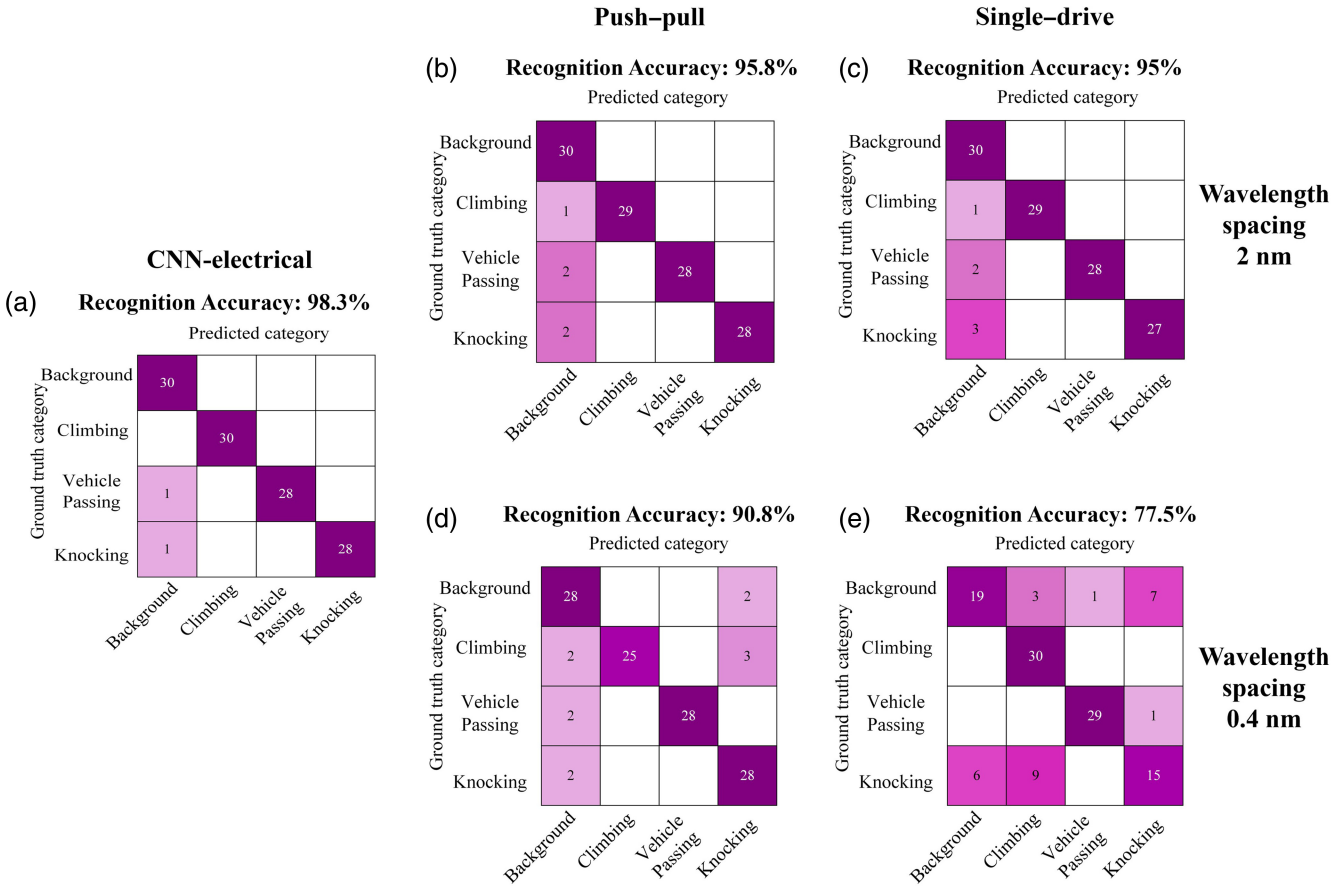

Fig. 7 Confusion matrix for the DAS classification task with image size of 36 × 36. (a) Electrical CNN confusion matrix; (b)–(e) photonic convolution kernel confusion matrices; (b) push–pull modulation with wavelength spacing of 2 nm; (c) single-drive modulation with wavelength spacing of 2 nm; (d) push–pull modulation with wavelength spacing of 0.4 nm; (e) single-drive modulation with wavelength spacing of 0.4 nm.

difference of less than 1%. Keeping the modulation rate constant, we selected different wavelength spacings and conducted classification experiments under both modulation methods, with results shown in Fig. 8. Regardless of the modulation method used, as the wavelength spacing increases, the classification accuracy shows a rising trend. Moreover, the classification accuracy of push–pull modulation is significantly better than that of single-drive modulation at different wavelength spacings, indicating that modulation chirp reduces the accuracy of TWM-PNNA; this effect is more pronounced with smaller wavelength spacings. In addition, we found that when $\Delta\lambda = 1.2$ nm, which corresponds to $\Delta\lambda_{\text{chirp}}/\Delta\lambda \approx 0.1$, the classification accuracy for single-drive modulation is 90%, as indicated by the dashed box in Fig. 8. Although modulation chirp has some impact on classification, the results remain acceptable. Therefore, we can set $\Delta\lambda_{\text{chirp}}/\Delta\lambda = 0.1$ as the chirp threshold for this system. Above this threshold, it is preferable to choose push–pull modulation, whereas below this threshold, single-drive modulation can still be considered.

Optical signals are affected by dispersion when propagating in optical fibers, especially chirped signals, which can exacerbate pulse broadening and cause signal distortion. Studying chirping can help optimize the phase and spectral characteristics of optical signals, reduce the impact of dispersion, and thus improve signal quality. In the TWM-PNNA system, the interaction between wavelengths may lead to signal degradation, especially when the chirp is not optimized. Effective control of chirping can improve the bandwidth utilization of optical signals, providing the possibility for optical computing systems to support more wavelength multiplexing channels, thereby achieving higher computational efficiency and communication rates. This aspect has not been adequately analyzed in existing work.

Traditional photonic neural network tasks mainly use the MNIST standard data set, where images are small in size (28×28) with distinct features. As a result, smaller convolution kernels (2×2 or 3×3) are sufficient to accomplish recognition tasks. However, as photonic neural networks begin to tackle more general tasks, the image sizes in data sets increase, and the features become more complex, necessitating larger convolution kernels, such as 5×5 or 7×7. This presents certain challenges to the TWM-PNNA system. In this system, the larger the convolution kernel, the more lasers are required. As the devices used in this system mainly operate in the C-band (1530 to 1560 nm), the available wavelength range is limited. Therefore, it is necessary to expand the size of the convolution kernel by reducing the wavelength spacing $\Delta\lambda$. However, reducing $\Delta\lambda$ is affected by the modulation chirp $\Delta\lambda_{\text{chirp}}$, making it important to study the impact of modulation chirp on the TWM-PNNA system. In addition, chirp-free modulation requires applying two opposite RF signals to the MZM, which increases system complexity and cost. Thus, studying the relationship between $\Delta\lambda_{\text{chirp}}/\Delta\lambda$ under single-drive modulation can provide a novel perspective on how to increase the convolution kernel size without increasing system complexity and cost. This would establish a theoretical foundation for enabling the system to handle more general tasks. Furthermore, compared with the solution of using dispersion compensation fibers to reduce modulation chirp, this approach is more flexible and cost-effective.[39]

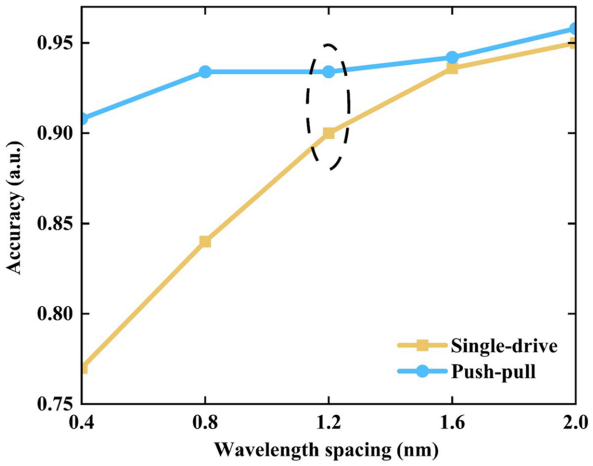

Fig. 8 Relationship between wavelength spacing and classification accuracy under different modulation schemes.

5 Pruning Experiment Results and Analysis of TWM-PNNA

Pruning is a commonly used model compression technique in neural networks that helps make the constructed network models more lightweight. Currently, pruning techniques have been shown to achieve a good balance between computational resources and model performance during the inference phase.[40] After pruning a network, the model's performance can sometimes even improve compared with the original, resulting in significant enhancements in memory usage and processing speed. Recently, this technique has been widely applied to various architectures, including CNNs,[41] recurrent neural networks,[42] transformer-based models,[43] and diffusion models.[44]

Pruning in deep-learning networks can be divided into three categories: pruning before training, pruning during training, and pruning after training (PAT). Currently, most optical neural networks are not capable of online training. The architecture of these optical neural networks is typically realized by loading pretrained neural network weight parameters to perform optical computation. Therefore, pruning techniques for optical neural networks are based on the principles of PAT, where pruning is applied to the trained weight parameters. In densely connected electrical neural networks, unstructured pruning can be treated as a constrained optimization problem, and its expression is as follows:[45]

$$\min_{\mathbf{w}} \mathcal{L}(\mathbf{w};\mathcal{D}) = \min_{\mathbf{w}} \frac{1}{N}\sum_{i=1}^{N} \ell(\mathbf{w};(x_i,y_i)), \quad \text{s.t. } \|\mathbf{w}\|_0 \leqslant k. \tag{4}$$

Here, $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{n}$ represents a data set, $\ell(\cdot)$ is the standard loss function, and $\mathbf{w}$ represents a series of weights in the neural network. $\|\cdot\|$ denotes the standard L_0-norm, and k is the total number of nonzero elements in the weight set $\mathbf{w}$. In this paper, the proposed TWM-PNNA architecture is suitable for unstructured pruning, where it is not necessary to directly set the weights to 0. Instead, the corresponding elements in the mask matrix $\mathbf{m}$ can be set to 0, and then, a dot product is performed. The result can be expressed as

$$\min_{\mathbf{w}} \mathcal{L}(\mathbf{w}\odot\mathbf{m};\mathcal{D}) = \min_{\mathbf{w},\mathbf{m}} \frac{1}{N}\sum_{i=1}^{N} \ell(\mathbf{w}\odot\mathbf{m};(x_i,y_i)), \quad \text{s.t. } \|\mathbf{m}\|_0 \leqslant k. \tag{5}$$

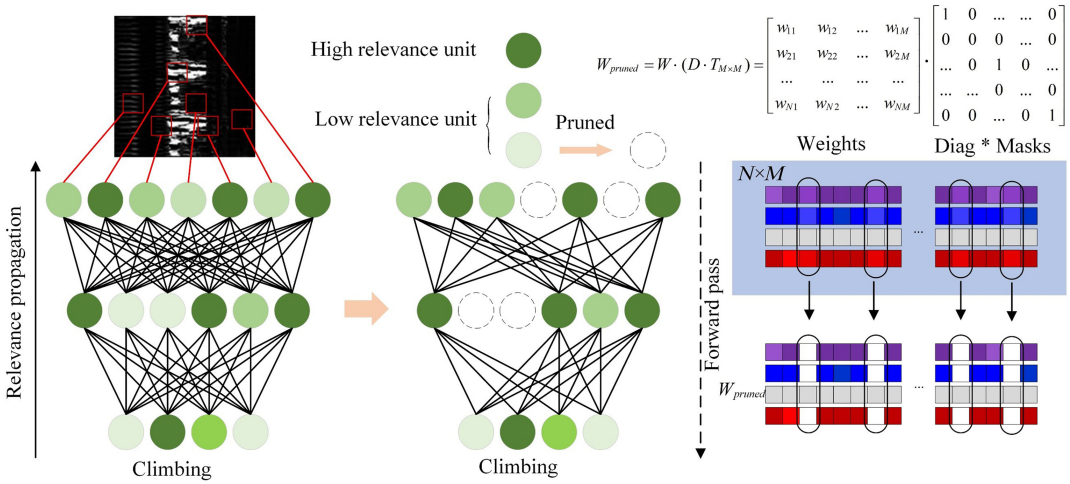

Fig. 9 Pruning scheme in the TWM-PNNA.

In the TWM-PNNA architecture, the network at weight initialization can be represented as $f(\mathbf{x};\mathbf{w}_0)$. After the network has completed training and achieved optimal performance, the weights of the network are updated, which can be expressed as $f(\mathbf{x};\mathbf{w}_T)$. Following PAT optimization, the pruned network is represented as $f(\mathbf{x};\mathbf{w}_T\odot\mathbf{m}')$, where $\mathbf{m}'$ is the corresponding mask matrix. This mask matrix $\mathbf{m}'$ preserves the performance of the trained model while also helping to compress the entire model.

In Fig. 9, we primarily focus on posttraining pruning. Neurons in the fully connected layer can be classified as strongly correlated and weakly correlated based on their relevance to the classification results. In fully connected layers, the weakly correlated neurons are more likely to be pruned. The strength of neurons can be determined through gradient descent. Recognizing sensor events is an M-class classification task, so the parameters of the fully connected layer form an $M \times N$ weight matrix. Therefore, each wavelength needs to load a different weight vector of size $1 \times N$.

For electrical fully connected layers, the pruning operation can be represented as the dot product of the weight matrix $\mathbf{w}_T$ and the mask matrix $\mathbf{m}'$, meaning that the zeros in each row of the weight matrix are randomly distributed. In optical pruning, the lightweight nature of the pruned network model must be considered, and the number of weights loaded per wavelength should be consistent. This means that the number of nonzero elements in each row of the weight matrix must be uniform and equal to the number of features X_{FC} input into the fully connected layer. In addition, the weight length loaded for each wavelength should be the same. Therefore, the pruned weight matrix for the fully connected layer can be expressed as

$$\mathbf{W}_{\text{pruned}} = \mathbf{W} \times (\mathbf{D} \times \mathbf{M}). \tag{6}$$

Here, $\mathbf{W}$, $\mathbf{D}$, and $\mathbf{M}$ represent the trained weight matrix, a diagonal matrix, and a mask matrix, respectively. After the pruning process, the pruned weight matrix is obtained, and the size of the model can be compressed by retaining only the nonzero elements.

We used the experimental setup shown in Fig. 3 and switched the optical switch to port B. The pooled signals were loaded into the optical path through the first modulator, and the WDM divided the optical signal into different optical paths, with each optical path corresponding to a neuron in the fully connected layer.[46] During the pretraining process, pruning techniques were applied to obtain parameters under different retention rates. These parameters were then loaded into the optical paths via the second modulator, where they were multiplied with the corresponding optical signals. At the same time, both modulators operate at the quadrature point. Finally, the multiplied signals were collected by a PD and accumulated to obtain the classification results of the fully connected layer. A detailed explanation of the principle of achieving fully connected operation through an MZM can be found in Note S2 in the Supplementary Material. WSS is limited by essentially fixed weights and port numbers in fully connected operations, making it difficult to meet dynamic and large-scale requirements. We propose using modulators to achieve fully connected computation. Compared with WSS, modulators have higher scalability and flexibility in fully connected computation and can dynamically adjust weights with fast speeds to meet complex network requirements. WSS is more suitable for convolution operations with fixed weights. At the same time, in the scheme of implementing full connection based on modulators, we also studied the impact of pruning on the fully connected results. Pruning optimizes resource allocation, improves computational efficiency, and reduces hardware requirements and system complexity by reducing redundant weight connections.

The experimental results are shown in Fig. 10. As seen in the figure, when the retention rate is between 0.6 and 1, the accuracy of the system decreases only slightly. However, when the retention rate drops below 0.6, there is a significant drop in accuracy. Therefore, a retention rate of 0.6 is a critical point for the system's accuracy, providing a basis for model simplification.

6 Discussion

In addition, the convolution operation speed in the TWM-PNNA system can be expressed as: 2 × (number of kernels) × (kernel size) × (baud rate) = 2 × 10 × 4 × 20 = 1.6 TOPS. In this work, a 2 × 2 kernel generates a convolution window with a vertical stride of 1, so the effective matrix computation speed is $1.6/1 = 1.6$ TOPS. According to the calculation method in Ref. 19, the computation speed can be further

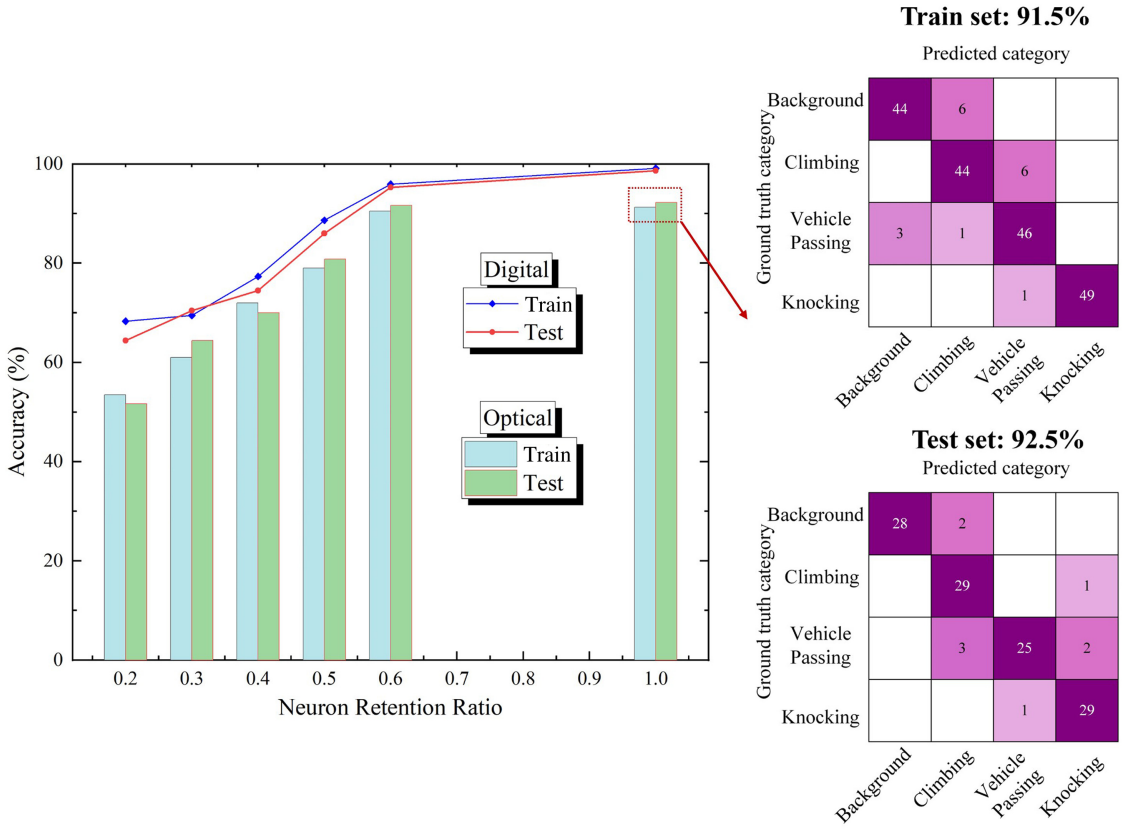

Fig. 10 Pruning results in TWM-PNNA.

improved by increasing the number of wavelengths and data rate. For example, when the number of wavelengths is 81 and the speed of AWG reaches 50 Gbaud, the speed of the TWM-PNNA system can reach 81 TOPS. Furthermore, besides speed, another important indicator of optical computing is power consumption. In our proof-of-concept configuration, power consumption mainly comes from tunable lasers, EDFA, modulator drivers,[47–50] WSS,[51] etc. However, the demonstration in this work relies on discrete devices; it is difficult to fully showcase the ultimate capabilities of our approach. Thanks to the development of hybrid and monolithic integration techniques,[52,53] as well as advancements in related technologies, the functions of the discrete devices used in the experiment can be implemented using integrated chips, which provides the possibility to measure the power consumption of TWM-PNNA systems. Therefore, we provide an estimation of the energy efficiency for a fully integrated optical TWM-PNNA (in line with the similar protocols in Refs. 19, 20, and 54). The excepted power consumption is ∼1.83 W, corresponding to an excepted energy efficiency of 0.87 TOPS/W for a fully integrated TWM-PNNA with a 20 Gbaud modulation rate and 2×2 kernel size. In addition, theoretically, when the processing speed reaches 81 TOPS, the power consumption is 3.85 W, achieving an energy efficiency of 21.02 TOPS/W. TWM-PNNA demonstrates certain advantages compared with existing electrical processors[55] and optical computing schemes.[54,56–58] In addition, we have conducted analyses on the system's latency and time-to-completion,[13,59] cost and physical size,[60–62] and reliability[63–65] (for detailed information on power consumption estimation, please refer to Note S4 in the Supplementary Material).

According to the principles of distributed acoustic sensing (DAS) systems, the information collected at each moment contains data from different spatial sensing points along the optical fiber. Therefore, it is necessary to store the data from the previous moment while collecting data from the next moment, which imposes certain requirements on storage duration.[66] However, existing optical storage technologies struggle to achieve storage times exceeding 10 s,[67] making it difficult to meet the needs of DAS systems. Consequently, without altering the current DAS system structure, the only option is to use DAQ systems to complete the data collection tasks. Furthermore, the data processing capability of the DAS system is limited by high-performance analog-to-digital conversion (ADC) and the speed of electronic signal processing. This bottleneck will result in the DAS system being able to support only complex algorithms at limited locations, preventing global coverage of these algorithms. To address these limitations, optical computing, including optical ADCs, presents a promising solution. Although it is true that the nonlinearity and noise in optical systems pose challenges for optical ADCs to simultaneously achieve high precision (high resolution) and high bandwidth in DAS systems, these limitations do not diminish the broader potential of optical computing. By leveraging its inherent advantages in parallelism and high energy efficiency, optical computing can be effectively utilized to process DAS data. Moreover, combining optical sensing with optical computing could offer significant benefits in size (e.g., potentially allowing for all-optical on-chip integration with DAS systems in the future), weight, and environmental adaptability, thus paving the way for intelligent monitoring and response in previously inaccessible or challenging environments.

The work of this article focuses on using TWM-PNNA as a co-GPU to process DAS data. By integrating optical computing into DAS systems, it is possible to complement or replace GPU-based processing, providing a novel and efficient approach to handling DAS data, and also proving that using optical computing to process DAS data is theoretically feasible. Future work can improve the DAS system and reduce its need for storage and high-performance ADCs. Meanwhile, in future work, optical demodulation methods, optical nonlinear calculations,[68] and *in situ* training methods[59,69] of optical neural networks will be developed to achieve DAS system data processing, advancing toward an all-optical DAS event recognition system (for details of future-oriented all-optical DAS event recognition system, see Note S5 in the Supplementary Material). In practical DAS applications, environmental factors such as temperature fluctuations, vibrations, fiber perturbations, and noise can impact performance. To address these, the TWM-PNNA can integrate adaptive mechanisms, including a temperature stabilization loop for laser and modulator stability, a vibration isolation platform to minimize mechanical disturbances, optical noise suppression via narrowband filters, and periodic self-calibration for long-term accuracy. These enhancements can enhance the robustness and reliability of the TWM-PNNA in dynamic and unpredictable environments.

7 Conclusion

In this paper, we introduce a photonic neural network accelerator into the real-time data processing of DAS systems for the first time, based on the TWM mechanism. By combining the high performance of fiber sensors with the high-speed data processing capabilities of optical computing, rapid processing of optical signals with low power consumption and high energy efficiency is enabled, providing accurate classification results. In addition, we investigate the impact of modulation chirp on the CNN accelerator. The experimental results show that the ratio of the wavelength shift caused by modulation chirp to the wavelength spacing between adjacent laser channels $\Delta\lambda_{\text{chirp}}/\Delta\lambda$ is an important metric for assessing the impact of chirp. When the ratio of $\Delta\lambda_{\text{chirp}}/\Delta\lambda$ is greater than 0.1, recognition accuracy is significantly affected by chirp. In this case, using a push–pull modulation method or reducing the ratio of $\Delta\lambda_{\text{chirp}}/\Delta\lambda$ can eliminate the impact of chirp on the accuracy, allowing the accuracy to surpass 90% and even approach the 98.3% achieved by electrical systems. This metric provides important guidance for flexibly selecting modulation methods and reducing system complexity. Furthermore, based on this architecture, we propose an optical computing architecture capable of fully connected layers and explore the impact of pruning on the performance of this fully connected architecture. The outcomes further support that retaining more than 60% of the fully connected parameters corresponding to the neurons can still maintain over 90% recognition accuracy, providing a basis for further reducing the model size.

Code and Data Availability

All data are available in the paper or the Supplementary Material. All related data, code, and materials in the main text or the supplementary information are available from the corresponding authors.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (NSFC) (Grant Nos. U2001601, 62175100, 62175103, and 61775094), the Equipping Pre-research Project (Grant No. 30601010104), the Fundamental Research Funds for the Central Universities (Grant Nos. 2024300447, 0213-14380211, and 0213-14380265), and the Jiangsu Innovation Teams and AI & AI for Science Project of Nanjing University.

References

1. N. J. Lindsey et al., "Illuminating seafloor faults and ocean dynamics with dark fiber distributed acoustic sensing," *Science* **366**(6469), 1103–1107 (2019).
2. J. Tejedor et al., "Real field deployment of a smart fiber-optic surveillance system for pipeline integrity threat detection: architectural issues and blind field test results," *J. Lightwave Technol.* **36**(4), 1052–1062 (2018).
3. A. Masoudi et al., "Subsea cable condition monitoring with distributed optical fiber vibration sensor," *J. Lightwave Technol.* **37**(4), 1352–1358 (2019).
4. Z. Li et al., "Fiber distributed acoustic sensing using convolutional long short-term memory network: a field test on high-speed railway intrusion detection," *Opt. Express* **28**(3), 2925–2938 (2020).
5. Z.-W. Ding et al., "Phi-OTDR based on-line monitoring of overhead power transmission line," *J. Lightwave Technol.* **39**(15), 5163–5169 (2021).
6. Q. Sun et al., "Recognition of a phase-sensitivity OTDR sensing system based on morphologic feature extraction," *Sensors* **15**(7), 15179–15197 (2015).
7. Y. Wang et al., "Pattern recognition using relevant vector machine in optical fiber vibration sensing system," *IEEE Access* **7**(1), 5886–5895 (2019).
8. X. Chen et al., "Disturbance pattern recognition based on an ALSTM in a long-distance φ-OTDR sensing system," *Microwave Opt. Technol. Lett.* **62**(1), 168–175 (2020).
9. Y. Shi et al., "An event recognition method for Φ-OTDR sensing system based on deep learning," *Sensors* **19**(15), 3421 (2019).
10. S. Liehr et al., "Real-time dynamic strain sensing in optical fibers using artificial neural networks," *Opt. Express* **27**(5), 7405–7425 (2019).
11. Z. Wei et al., "A representation-enhanced vibration signal imaging method based on MTF-NMF for Φ-OTDR recognition," *J. Lightwave Technol.* **42**(18), 6395–6401 (2024).
12. Y. Shi et al., "An easy access method for event recognition of φ-OTDR sensing system based on transfer learning," *J. Lightwave Technol.* **39**(13), 4548–4555 (2021).
13. Y. Shen et al., "Deep learning with coherent nanophotonic circuits," *Nat. Photonics* **11**(7), 441–446 (2017).
14. H. Zhang et al., "An optical neural chip for implementing complex-valued neural network," *Nat. Commun.* **12**(1), 457 (2021).
15. M. Y. S. Fang et al., "Design of optical neural networks with component imprecisions," *Opt. Express* **27**(10), 14009–14029 (2019).
16. S. Yu et al., "Heavy tails and pruning in programmable photonic circuits for universal unitaries," *Nat. Commun.* **14**(1), 1853 (2023).
17. X. Lin et al., "All-optical machine learning using diffractive deep neural networks," *Science* **361**(6406), 1004–1008 (2018).
18. T. Fu et al., "Photonic machine learning with on-chip diffractive optics," *Nat. Commun.* **14**(1), 70 (2023).
19. X. Xu et al., "11 TOPS photonic convolutional accelerator for optical neural networks," *Nature* **589**(7840), 44–51 (2021).
20. B. Bai et al., "Microcomb-based integrated photonic processing unit," *Nat. Commun.* **14**(1), 66 (2023).

21. X. Meng et al., "Compact optical convolution processing unit based on multimode interference," *Nat. Commun.* **14**(1), 3000 (2023).
22. S. Xu et al., "High-order tensor flow processing using integrated photonic circuits," *Nat. Commun.* **13**(1), 7970 (2022).
23. K. Tang et al., "Photonic tensor processing unit with single dataflow and programmable high-precision weighting control," *J. Lightwave Technol.* **42**(2), 659–669 (2023).
24. P. Healey, "Fading in heterodyne OTDR," *Electron. Lett.* **20**(1), 30–32 (1984).
25. R. Hong et al., "Distributed dynamic strain measurement with a direct detection scheme by using a three-step-phase-shifted double pulse in a UWFBG array," *Opt. Lett.* **48**(8), 2090–2093 (2023).
26. J. Shao et al., "Near-surface structure investigation using ambient noise in the water environment recorded by fiber-optic distributed acoustic sensing," *Remote Sens.* **15**(13), 3329 (2023).
27. M. Zabihi et al., "Continuous fading suppression method for Φ-OTDR systems using optimum tracking over multiple probe frequencies," *J. Lightwave Technol.* **37**(14), 3602–3610 (2019).
28. J. Shao et al., "Tracking moving ships using distributed acoustic sensing data," *IEEE Geosci. Remote Sens. Lett.* **22**(1), 7502605 (2025).
29. Y. Li et al., "A deep learning model enabled multi-event recognition for distributed optical fiber sensing," *Sci. China Inf. Sci.* **67**(3), 132404 (2024).
30. E. E. Elsayed et al., "Performance evaluation and enhancement of the modified OOK based IM/DD techniques for hybrid fiber/FSO communication over WDM-PON systems," *Opt. Quantum Electron.* **52**(9), 385 (2020).
31. E. E. Elsayed et al., "Investigations on wavelength-division multiplexed fibre/FSO PON system employing DPPM scheme," *Opt. Quantum Electron.* **54**(6), 358 (2022).
32. Y. Lecun et al., "Gradient-based learning applied to document recognition," *Proc. IEEE* **86**(11), 2278–2324 (1998).
33. X. Meng et al., "On-demand reconfigurable incoherent optical matrix operator for real-time video image display," *J. Lightwave Technol.* **41**(6), 1637–1648 (2023).
34. N. Yang et al., "Real-time classification for 0-OTDR vibration events in the case of small sample size datasets," *Opt. Fiber Technol.* **76**(1), 103217 (2023).
35. R. Yao et al., "Vibration event recognition using SST-based Φ-OTDR system," *Sensors* **23**(21), 8773 (2023).
36. F. Koyama et al., "Frequency chirping in external modulators," *J. Lightwave Technol.* **6**(1), 87–93 (1988).
37. X. Chi et al., "Joint intra and inter-channel nonlinear compensation scheme based on improved learned digital back propagation for WDM systems," *Opt. Express* **32**(4), 5095–5116 (2024).
38. E. E. Elsayed, "Atmospheric turbulence mitigation of MIMO-RF/FSO DWDM communication systems using advanced diversity multiplexing with hybrid N-SM/OMI M-ary spatial pulse-position modulation schemes," *Opt. Commun.* **562**, 130558 (2024).
39. Y. Y. Huang et al., "Programmable matrix operation with reconfigurable time-wavelength plane manipulation and dispersed time delay," *Opt. Express* **27**(15), 20456–20467 (2019).
40. H. Cheng et al., "A survey on deep neural network pruning: taxonomy, comparison, analysis, and recommendations," *IEEE Trans. Pattern Anal. Mach. Intell.* **46**, 10558–10578 (2024).
41. D. Ghimire et al., "A survey on efficient convolutional neural networks and hardware acceleration," *Electronics* **11**(6), 945 (2022).
42. G. Fang et al., "DepGraph: towards any structural pruning," in *IEEE/CVF Conf. Comput. Vision and Pattern Recognit. (CVPR)*, pp. 16091–16101 (2023).
43. E. Frantar et al., "SparseGPT: massive language models can be accurately pruned in one-shot," arXiv:2301.00774 (2023).
44. G. Fang et al., "Structural pruning for diffusion models," arXiv:2305.10924 (2023).
45. N. Lee et al., "SNIP: single-shot network pruning based on connection sensitivity," arXiv:1810.02340 (2019).
46. A. Krizhevsky et al., "ImageNet classification with deep convolutional neural networks," *Adv. Neural Inf. Process. Syst.* **25**(1), 1097–1105 (2012).
47. E. Temporiti et al., "A 56 Gb/s 300 mW silicon-photonics transmitter in 3D-integrated PIC25G and 55 nm BiCMOS technologies," in *63rd IEEE Int. Solid-State Circuits Conf. (ISSCC)*, pp. 404–405 (2016).
48. L. Kull et al., "A 24-to-72GS/s 8b time-interleaved SAR ADC with 2.0-to-3.3 pJ/conversion and >30 dB SNDR at Nyquist in 14 nm CMOS FinFET," in *65th IEEE Int. Solid-State Circuits Conf. (ISSCC)*, pp. 358–360 (2018).
49. M. A. Kossel et al., "An 8b DAC-based SST TX using metal gate resistors with 1.4 pJ/b efficiency at 112 Gb/s PAM-4 and 8-tap FFE in 7 nm CMOS," in *IEEE Int. Solid-State Circuits Conf. (ISSCC)*, pp. 130–132 (2021).
50. K. R. Lakshmikumar et al., "A process and temperature insensitive CMOS linear TIA for 100 Gb/s/λ PAM-4 optical links," *IEEE J. Solid-State Circuits* **54**(11), 3180–3190 (2019).
51. T. Li et al., "Nonvolatile switching in In_2Se_3-silicon microring resonators," in *Conf. Lasers and Electro-Opt. (CLEO)*, SM4B.5 (2021).
52. B. Stern et al., "Battery-operated integrated frequency comb generator," *Nature* **562**(7727), 401–405 (2018).
53. Thorlabs, https://www.thorlabs.com/newgrouppage9.cfm?objectgroup_id=3944.
54. J. Feldmann et al., "Parallel convolutional processing using an integrated photonic tensor core," *Nature* **591**(7849), E13 (2021).
55. Nvidia, https://www.nvidia.com/en-us/autonomous-machines/embedded-systems/jetson-nano/product-development.
56. B. Shi et al., "Deep neural network through an InP SOA-based photonic integrated cross-connect," *IEEE J. Sel. Top. Quantum Electron.* **26**(1), 7701111 (2020).
57. F. Ashtiani et al., "An on-chip photonic deep neural network for image classification," *Nature* **606**(7914), 501–506 (2022).
58. H. H. Zhu et al., "Space-efficient optical computing with an integrated chip diffractive neural network," *Nat. Commun.* **13**(1), 1044 (2022).
59. S. Pai et al., "Experimentally realized in situ backpropagation for deep learning in photonic neural networks," *Science* **380**(6643), 398–404 (2023).
60. P. Kaur et al., "Hybrid and heterogeneous photonic integration," *APL Photonics* **6**(6), 061102 (2021).
61. M. R. Billah et al., "Hybrid integration of silicon photonics circuits and InP lasers by photonic wire bonding," *Optica* **5**(7), 876–883 (2018).
62. J. Zhang et al., "Silicon photonics fiber-to-the-home transceiver array based on transfer-printing-based integration of III-V photodetectors," *Opt. Express* **25**(13), 14290–14299 (2017).
63. A. Yi et al., "Silicon carbide for integrated photonics," *Appl. Phys. Rev.* **9**(3), 031302 (2022).
64. S. Shekhar et al., "Roadmapping the next generation of silicon photonics," *Nat. Commun.* **15**(1), 751 (2024).
65. J. E. Johnson et al., "Performance and reliability of advanced CW lasers for silicon photonics applications," in *Opt. Fiber Commun. Conf. and Exhibit. (OFC)*, Tu2D.1 (2022).
66. H. Wu et al., "Simultaneous extraction of multi-scale structural features and the sequential information with an end-to-end mCNN-HMM combined model for fiber distributed acoustic sensor," *J. Lightwave Technol.* **39**(20), 6606–6616 (2021).
67. B. Jing et al., "Approaching scalable quantum memory with integrated atomic devices," *Appl. Phys. Rev.* **11**(3), 031304 (2024).
68. G. H. Y. Li et al., "All-optical ultrafast ReLU function for energy-efficient nanophotonic deep learning," *Nanophotonics* **12**(5), 847–855 (2023).
69. Z. Xue et al., "Fully forward mode training for optical neural networks," *Nature* **632**(8024), 280–286 (2024).

Fuhao Yu received a master's degree in 2022 from the Taiyuan University of Technology for research on distributed fiber optic sensing. He is currently pursuing his PhD in the College of Engineering and Applied Sciences, at Nanjing University, Nanjing, China. His research interests include silicon photonics, optical computing and modulators.

Kangjian Di is a PhD student in the School of Integrated Circuits at the Nanjing University. His research focuses on optical neural network and AI for optics.

Wenjun Chen is currently pursuing a master's degree in the School of Integrated Circuits at the Nanjing University. His research interests focus on the field of photonic chips and intelligent algorithms.

Sen Yan received his MS degree in control science and engineering from Qilu University of Technology (Shandong Academy of Sciences), Jinan, China. He is currently pursuing his PhD in the College of Engineering and Applied Sciences, Nanjing University.

Yuanyuan Yao is a doctoral candidate in optical engineering. Her primary research focuses on data analysis of distributed optical fiber sensing systems and the application of machine learning techniques in engineering.

Silin Chen is a PhD student at the School of Integrated Circuits, Nanjing University. His research interests include AI for chips and for remote sensing, image processing and deep learning.

Xuping Zhang is a professor with the College of Engineering and Applied Sciences, Nanjing University, Nanjing, China. She is the director of the Key Laboratory of Intelligent Optical Sensing and Manipulation, Ministry of Education of China, and the dean of the Institute of Optical Communication Engineering of Nanjing University since 2002. Her current research interest focuses on distributed optical fiber sensing and its applications on structure health monitoring

Yixin Zhang is a professor with the College of Engineering and Applied Sciences, Nanjing University, Nanjing, China. His research interests include fiber-sensor-based health monitoring technology and digital signal processing.

Ningmu Zou is an associate professor at the Nanjing University School of Integrated Circuits. From 2017 to 2023, he was a machine learning engineer at AMD in Texas and an adjunct professor at Texas State University. He specializes in applying AI to advanced semiconductor processes, photolithography, optical mask correction, yield improvement, and defect detection.

Wei Jiang (Senior Member, IEEE) received his BS degree in physics from Nanjing University, an MA degree in physics and a PhD in electrical engineering from the University of Texas at Austin. He is currently a professor in the College of Engineering and Applied Sciences at Nanjing University, Nanjing, China, and serves as an associate director of Optical Communications Systems & Network Engineering Research Center of Jiangsu Province.