Research Article

ADVANCED PHOTONICS

Waveguide-array optical neural network on thin-film lithium niobate platform

Xueyun Li,[a] Zhizhang Wang,[a,*] Jitao Ji,[a] Zihan Gao,[a] Shengjie Wu,[a] Jiacheng Sun,[a] Zhiyuan Lin,[a] Jian Li,[a] Zhilin Ye,[b] Yuexin Chang,[b] Xian Long,[a] Shanshan Hu,[a] Kelan Chen,[a] Shining Zhu,[a] and Tao Li[a,*]

[a]Nanjing University, College of Engineering and Applied Sciences, School of Physics, National Laboratory of Solid State Microstructures, Key Laboratory of Intelligent Optical Sensing and Manipulations, Jiangsu Key Laboratory of Artificial Functional Materials, Nanjing, China

[b]NanZhi Institute Advanced Optoelectronic Integration, Nanjing, China

Abstract. Photonic parallel computing inspires important approaches to artificial intelligence, promising substantial gains in processing speed, energy efficiency, and latency. Nevertheless, the integration density and robustness of most optical computing architectures continue to pose significant challenges. Here, we report a waveguide-array optical neural network based on thin-film lithium niobate (TFLN). By virtue of the compact size of the waveguide arrays, this architecture yields a computational density of 9.48×10^4 TOPS/mm^2. Moreover, the high-speed and energy-efficient characteristics of the TFLN platform enable a computing speed of 135.6 TOPS with an energy efficiency of 6.28 TOPS/W. Our experiments on the iris flower and handwritten digit recognition demonstrate accuracies of 100% and 89.7%, respectively. The device achieves over 90% accuracy across a 75-nm operating bandwidth in the iris flower recognition task, indicating the wavelength insensitivity and superior robustness of the waveguide array. As a compact paradigm for optical neural networks, we demonstrate the potential of waveguide arrays in performing optical computation, thereby offering solutions for the advancement of edge computing, optoelectronic integration, and intelligent automation.

Keywords: waveguide-array optical neural network; thin-film lithium niobate; high speed; high energy efficiency; high computational density.

Received Jan. 28, 2026; revised manuscript received Aug. 25, 2026; accepted for publication Sep. 1, 2026; published online Oct. 6, 2026.

© The Authors. Published by SPIE and CLP under a Creative Commons Attribution 4.0 International License. Distribution or reproduction of this work in whole or in part requires full attribution of the original publication, including its DOI.

[DOI: 10.1117/1.AP.8.5.056013]

1 Introduction

Artificial intelligence has promoted a number of advancements in a multitude of industries, including autonomous driving,[1] computer vision,[2] preference analysis,[3] and large language models,[4] while also generating substantial demand for data and computational resources.[5,6] Traditional electronic computing architectures, including central processing units, graphics processing units, and application-specific integrated circuits, face severe challenges from high energy consumption and latency in processing large-scale deep learning tasks.[7,8]

To overcome these limitations, optical neural networks (ONNs) are recognized as a crucial direction for next-generation high-performance computing,[9–18] leveraging their inherent advantages of ultra-high energy efficiency, ultra-high bandwidth, ultra-low latency, and parallel processing at the speed of light. Research into on-chip ONNs primarily follows two main technical routes: one is based on interferometer arrays such as Mach–Zehnder interferometers (MZIs)[19–23] and micro-ring resonators (MRRs)[24–29] to realize linear matrix multipliers; the other leverages static diffractive optical elements, referred to as on-chip diffractive ONNs (DONNs).[30–34] However, existing on-chip photonic computing solutions still face significant challenges. MZI/MRR-based schemes, although programmable, suffer from a large physical footprint, limiting integration density, and their active modulation mechanisms (e.g., the thermo-optic effect, plasma dispersion effect, or material phase transition) are often power-hungry and sensitive to environmental fluctuations. Conversely, on-chip DONNs utilize planar diffractive elements

*Address all correspondence to Zhizhang Wang, zhizhangwang@nju.edu.cn; Tao Li, taoli@nju.edu.cn

to perform computing functions, aiming for high computational density and energy efficiency. Nevertheless, the precise diffraction structures of DONNs demand extremely high fabrication accuracy, and this leads to poor fabrication error tolerance as well as high sensitivity to operating wavelengths, which hinders the application in wavelength division multiplexing and the further scaling of the computational complexity.

To address the critical challenges of speed, power consumption, integration density, and robustness concurrently, a compelling pathway is to leverage advanced material platforms and novel photonic architectures. The thin-film lithium niobate (TFLN) platform, renowned for its strong electro-optic effect, provides a superior foundation for high-speed and low-power optical modulation, directly mitigating the speed and power consumption limitations of traditional silicon-based or silicon nitride-based devices.[35–40] Meanwhile, waveguide arrays stand out as a mature and versatile photonic structure. They offer remarkable design flexibility, stable fabrication processes, and a compact form factor. Although waveguide arrays have demonstrated free-form design potential in areas such as topological photonics,[41–43] their application in ONNs has historically been limited to simple, regular layouts,[35,44–46] underutilizing their full potential for complex computational design.

In this work, we propose a waveguide-array optical neural network (WONN) based on the TFLN platform. Building upon this material advantage, we innovatively adopt the waveguide array structure as the core computing unit as shown in Fig. 1 to improve the integration density and robustness compared with the existing ONN structures. Leveraging the compact nature of the waveguide array, we achieve computational units with dimensions of 14.3 μm × 100 μm (for the iris flower classification task) and 24.8 μm × 300 μm [for the Modified National Institute of Standards and Technology (MNIST) classification task], which can realize a computational density of 9.48×10^4 TOPS/mm^2. Meanwhile, our device achieves over 90% accuracy across a 75-nm operating bandwidth in the iris classification task, demonstrating the wavelength insensitivity and superior robustness of the waveguide array. Furthermore, by exploiting the superior electro-optic effect of the TFLN platform, it achieves a theoretically estimated high computational speed of 135.6 TOPS and an energy efficiency of 6.28 TOPS/W. This approach exhibits an exceptional performance as shown in Fig. 1(c) and introduces a novel paradigm for the design of neural network computational units, which is not only adaptable across various material platforms but also enables additional design flexibility in optical computing.

2 Materials and Methods

Figure 1(a) illustrates the architecture of our WONN device. The chip integrates multimode interference (MMI) splitters in multiple stages, electro-optic phase shifters, waveguide arrays, and input/output grating couplers. The input task is first preprocessed and converted into voltage signals, which are applied to the electrodes of the phase shifters to modulate the optical phase. The modulated light then propagates through the waveguide array, where the matrix-vector multiplication (MVM) is carried out, and the computational outcome is subsequently retrieved as an optical intensity output. The dense arrangement of the waveguides introduces coupling interactions among them, which governs the propagation of the light field. By precisely controlling the strength of each coupling coefficient, the transmission matrix T of the waveguide array can be deliberately designed [Figs. 1(a) and 1(b)]. Implementing the transmission matrix T for tasks of varying complexity presents a trade-off: although more waveguides are necessary for complex operations, an overabundance wastes space, which strikes a balance in selecting the waveguide number. Accordingly, a 10-waveguide array (representing a 10×10 weight matrix) and a 20-waveguide array (representing a 20×20 weight matrix) are used for the iris task and the MNIST task, respectively. Meanwhile, when configuring the input and output ports of the array, a well-considered selection strategy is adopted to achieve optimal task performance. The specific waveguides selected thereby establish the matrices $M_{n\times p}$ and $N_{q\times n}$ (p, q, and n are the number of neural units in input layer, output layer, and hidden layer, respectively), which physically implement the input-to-hidden and hidden-to-output layer weight matrices (see Note S1 in the Supplementary Material).

Generally, the coupling interaction can be described by the coupled mode theory (CMT).[47–49] For the waveguide array system in Figs. 2(a) and 2(b), assuming the coupling among waveguides is $c_1, c_2, \ldots, c_{n-1}$, the system Hamiltonian can be written as

$$H = \begin{pmatrix} \beta_1 & c_1 & & & \\ c_1 & \beta_2 & c_2 & & \\ & c_2 & \beta_3 & \ddots & \\ & & \ddots & \ddots & c_{n-1} \\ & & & c_{n-1} & \beta_n \end{pmatrix}, \tag{1}$$

where $\beta_1, \beta_2, \ldots, \beta_n$ are propagation constants in each waveguide. In this work, all the propagation constants are the same for the identical structural features of the waveguides. According to the CMT, the propagation of the light field in the waveguide array can be described in the similar mathematic form to the Schrödinger equation

$$-i\frac{\mathrm{d}}{\mathrm{d}z}A(z) = HA(z), \tag{2}$$

where i is the imaginary unit, z is the coordinate along the direction of the light field evolution, and $A(z) = [A_1(z) \quad A_2(z) \quad \ldots \quad A_n(z)]$ is the light field distribution at coordinate z. Solving this differential equation, we can get $A(z) = A(0)e^{i\int_0^z H\mathrm{d}z}$. Because there is no modulation on the coupling coefficients and propagation constants along the evolution in our device, the equation can be simplified and the transmission matrix is derived as $T = e^{iHz}$. A limitation arises because the CMT requires the tight-binding approximation (nearest-neighbor coupling only) and the constant-mode profile assumption (inter-waveguide coupling does not perturb the modal field distribution). In the practical realization with TFLN rib waveguides, the optical mode is not perfectly confined, introducing significant higher-order coupling, which causes the CMT-calculated field distribution to deviate from the actual one. A more accurate method for simulating light propagation in waveguide arrays is the eigenmode expansion (EME) technique. This approach begins by calculating the eigenmodes of the entire waveguide array. The input field is then decomposed into these eigenmodes, and the propagation of each individual mode is computed. The output field distribution is finally obtained by the coherent superposition of all these evolved eigenmodes at the output. Although this method is computationally more

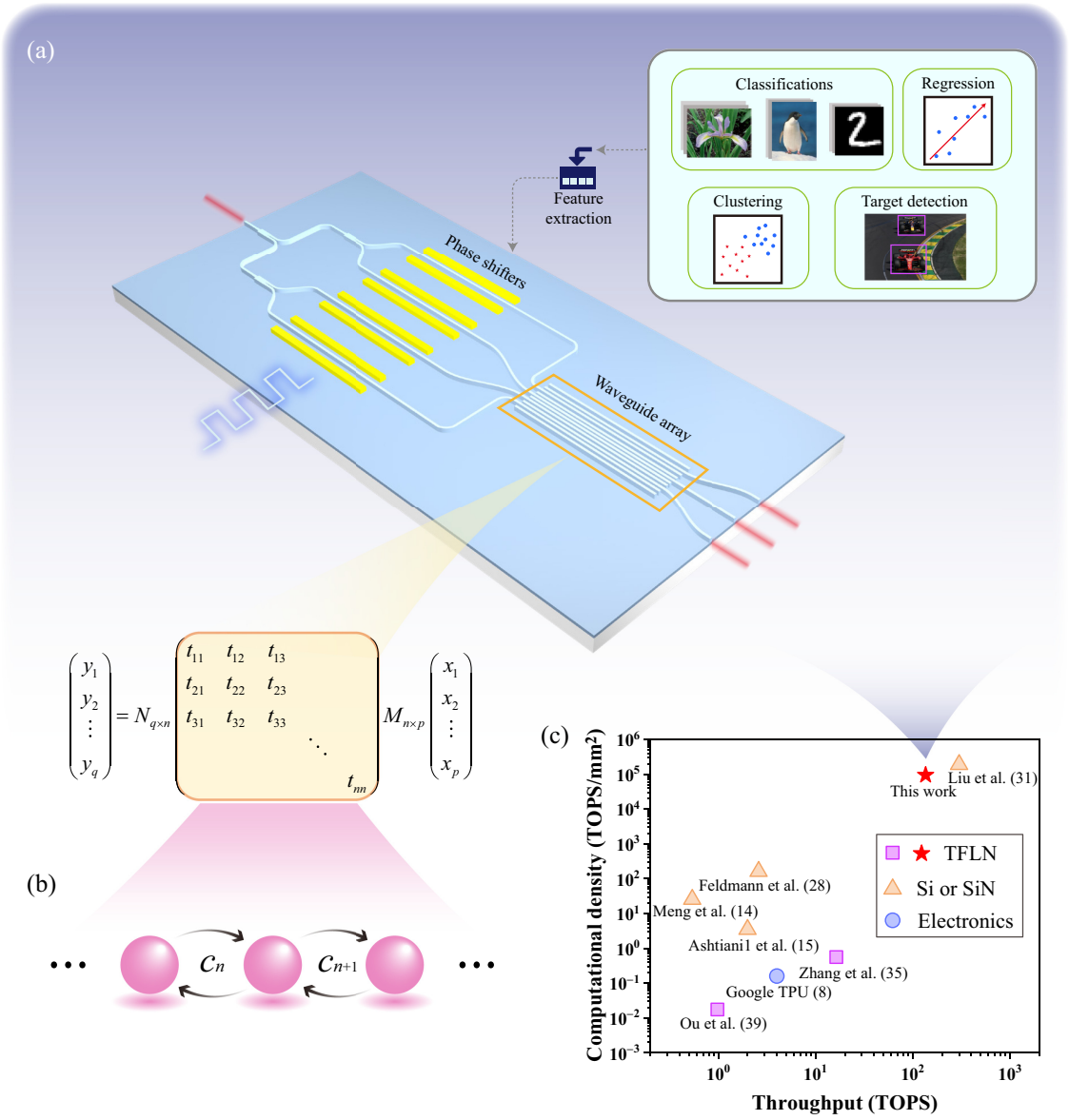


Fig. 1 Schematic illustration of waveguide-array optical neural network. (a) Schematic diagram of the proposed device, which primarily consists of MMI splitters, electro-optic phase shifters, and waveguide arrays. The waveguide arrays are elaborately designed to execute the MVM operations in the network, which is carried out within an extremely low latency when the input signals converted from the specific tasks are fed into the device. The mathematical formulation of the waveguide array in the neural network is presented in the lower-left corner, where the transmission matrix T highlighted in the orange box is obtained by precisely controlling each inter-waveguide coupling according to the CMT as shown in panel (b). (c) Performance of the proposed device compared with that of other on-chip optical neural network works on the different material platforms in terms of the throughput and the computational density.

intensive than CMT, it yields results that are in excellent agreement with rigorous numerical simulations. To achieve both speed and precision during the training process, in the design of our device, we introduce an efficient training optimization method. As shown in Fig. 2(b), the WONN is first trained using a CMT model to achieve a coarse approximation of the coupling coefficients, which are subsequently refined via an EME model to yield the final values (see Notes S2 and S3, Fig. S3 in the Supplementary Material).

Based on the training results, the chip is processed using a complementary metal-oxide-semiconductor (CMOS)-compatible manufacturing process. Then, the fabricated chip is wire-bonded to a printed circuit board and the photograph of the packaged chip is exhibited in Fig. 2(c). Figure 2(d) presents the microscope picture of the overall view of the device used in the iris flower classification task, and Figs. 2(e) and 2(f) show the scanning electron microscope (SEM) images of the waveguide array and an enlarged view of the input part. The

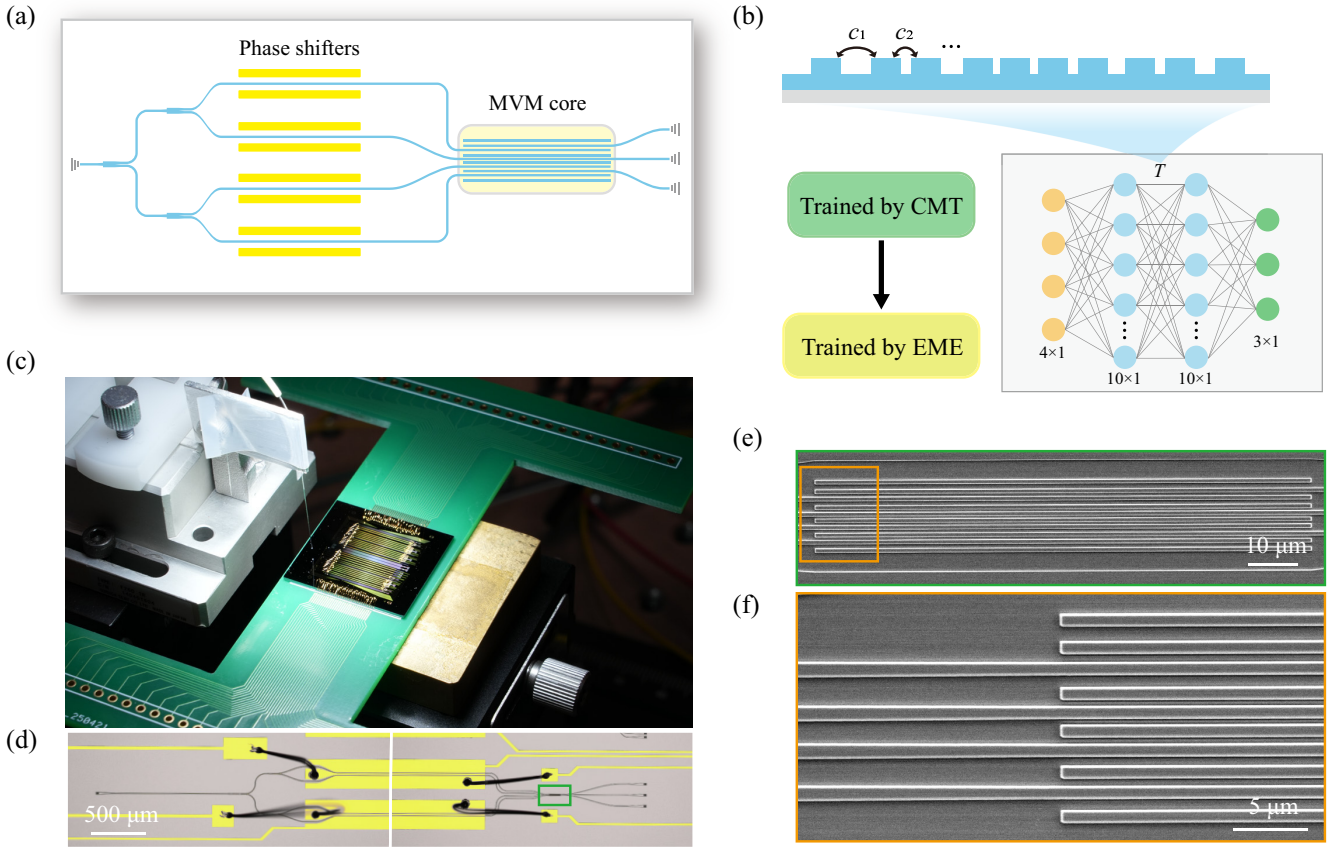


Fig. 2 Design principle and the prototype of the proposed device. (a) Modeling diagram of the WONN. (b) Cross-profile of the waveguide array and the designing principle. A two-step training optimization method is utilized to achieve the proper coupling coefficients powered by CMT method and EME method. (c) Photograph of the packaged chip. (d) Microscope image of the device used in the iris flower classification task. (e) SEM image of the overall view of the waveguide array highlighted by the green box. Within the orange box is an (f) enlarged view of the input part.

waveguides feature a width of 800 nm, a ridge height of 300 nm with a 300-nm slab waveguide below, a propagation length of 100 μm, and gaps ranging from 0.48 to 1.09 μm. The structural parameters of the phase shifters are based on the previous works.[50,51] The MZI modulator of the same design features a half-wave voltage of 3.37 V by the pull–push configuration and an extinction ratio of 26 dB. The phase shifters used in the specific tasks are also characterized by calculating the mean square error (MSE) function between the output intensity distribution and the reference, indicating a voltage of ∼13 V for a full 2π phase shift by the single-arm configuration (see Note S4 and Fig. S6 in the Supplementary Material).

3 Results

3.1 Iris Flower Classifier

In the iris flower classification task, the complete dataset of 150 samples was randomly divided into training and test sets with a 4:1 split, i.e, 120 samples for training and 30 for testing. Figure 3(a) illustrates the progression of the loss function and accuracy throughout the two-stage training process of the neural network. A slight increase in loss and a corresponding dip in accuracy are observed at the transition from the first to the second training stage, consistent with our previous discussion. After the training, our device achieves 100% recognition accuracy on the test set in the simulation. To validate the generalizability of our proposed architecture, we also design a WONN for the Palmer Penguins dataset, an alternative to the iris flower dataset, achieving a theoretical test accuracy of 98.6%.

After the chip processing and packaging, the device undergoes initial characterization of a pre-operational calibration and rectification procedure owing to processing imperfections. There are two influences brought from fabrication errors. First, morphological imperfections within the waveguide array result in a deviation of the inter-waveguide coupling coefficients from their designated values, thereby inducing errors in the transmission matrix. Second, an initial phase error is imparted to the light field upon its passage through the phase shifters before it is coupled into the waveguide array. We begin by analyzing the influence of the manufacturing imperfections in the waveguide array. Figures 3(b) and 3(c) show the simulated variations in the loss function and task accuracy with respect to the evolution length error and the waveguide width error, respectively. It can be observed that within a length error range of approximately −7 to +8 μm, the WONN maintains a recognition accuracy above 90%, and across the entire considered width error range from −100 to +100 nm, the recognition accuracy remains

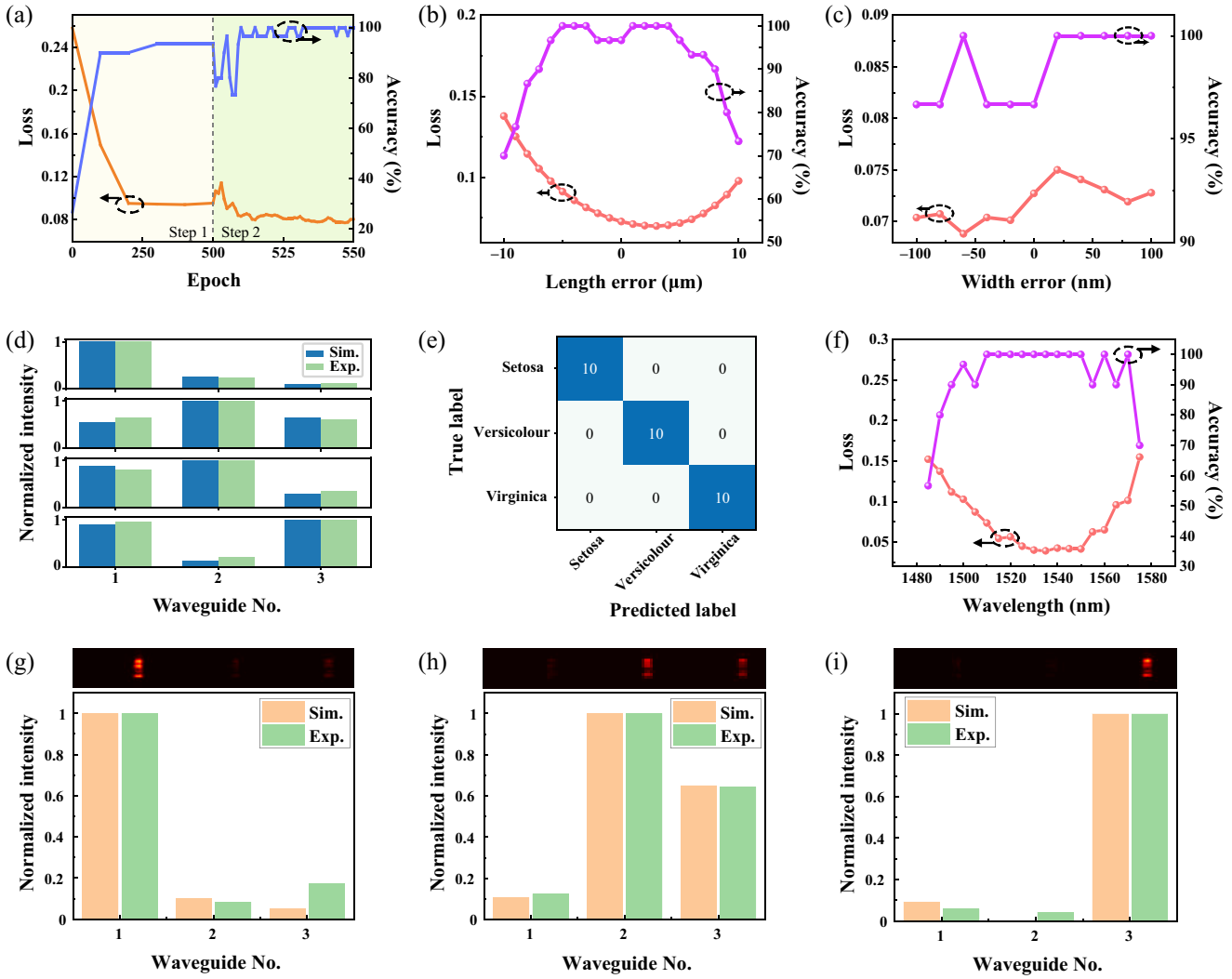


Fig. 3 Analysis of fabrication errors and experimental results of iris flower classification task. (a) Loss function and the test accuracy versus the training epochs in the two-step training optimization method. (b) and (c) Variation in the loss function and accuracy versus different evolution length errors and waveguide width errors, respectively. (d) Comparison between the output intensity distribution measured after exciting each input port of an isolated waveguide array and the simulation results with a propagation length of 105 μm. (e) Confusion matrix of the classifying results on the test dataset with an accuracy of 100%. (f) Classification accuracy of the test dataset as a function of wavelength, showing an operating bandwidth of 75 nm from 1495 to 1570 nm with a recognition accuracy in excess of 90%. (g) to (i) Simulated outputs and the experimental outputs of the designed waveguide array when three different species of iris flower are fed into the WONN at the operating wavelength of 1550 nm. The illustration above each figure presents the picture captured in the experiment, where the brightest port indicates the recognition result.

above 95%, demonstrating an excellent tolerance to fabrication errors. We also characterize the performance of an isolated waveguide array and compare the experimental results with theoretical simulations. By successively stimulating each input port, the intensity outputs are subsequently recorded and shown from the top to the bottom in Fig. 3(d). The experimental results align with the theoretical simulations at an evolution length of 105 μm, which indicates that the associated error in the evolution process is around 5 μm and well within the permissible range. For the compensation of the initial phase error, a method involving training-based *in situ* compensation is proposed, which entails retraining the model directly on the erroneous chip. The optimization variables are defined as the compensatory voltages for each electrode, and the MSE is set as the objective function, which is minimized through iterative gradient descent optimization (see Note S5 and Fig. S7 in the Supplementary Material).

Figure 3(e) shows the confusion matrix of the identification results of our device on the test dataset and the accuracy reaches 100%, and the accuracy on the whole samples in the dataset is 96%, demonstrating the advantages of accuracy and reliability in matrix operations. The wavelength response of the WONN is tested in Fig. 3(f) and the chip's classification accuracy on the test dataset across different wavelengths is measured. The results show that within the range of 1495 to 1570 nm, the

recognition accuracy remains above 90%, indicating the waveguide array's insensitivity to wavelength and confirming its superior robustness. Figures 3(g)–3(i) present a comparison between the simulated and experimentally measured outputs of the waveguide array for three different species of iris flowers at wavelength of 1550 nm, revealing excellent agreement. The illustration above each figure shows the measured intensity distribution of the output grating, in which the brightest port indicates the recognition result and the waveguide numbers 1, 2, and 3 represent three species of Setosa, Versicolour, and Virginica, respectively.

3.2 Handwritten Digit Classifier

The MNIST handwritten digit images classifying is a more complicated task. Utilizing the method above, a new WONN is designed accordingly, and the architecture is shown in Fig. 4(a). The MNIST dataset is divided into a training dataset of 60,000 images and a test dataset of 10,000 images, and each image consists of 28 × 28 grayscale pixels. Prior to being fed into the device, every image is first flattened into a 784 × 1 vector, and then the vector is processed by a fully connected layer, which compresses it to a 10 × 1 vector to match the number of inputs on the device [Fig. 4(b)]. In this task, the number of waveguides in the array is set to 20, and the evolution length is 300 μm to suit the task's complexity. The waveguides, with a width of 600 nm and gaps ranging from 0.48 to 1.17 μm, give a total width of 24.8 μm. This configuration exhibits a recognition accuracy of 91% on the test dataset in the simulation. We also perform a comparison on a digital chip, where the compression operation is kept unchanged, and the WONN chip is replaced with a complex-valued linear regression layer which is retrained. The resulting recognition accuracy reaches 87.9%, demonstrating the effectiveness of WONN in the recognition task. Figure 4(c) is the SEM image of the input section of the fabricated waveguide array. When evaluated on 300

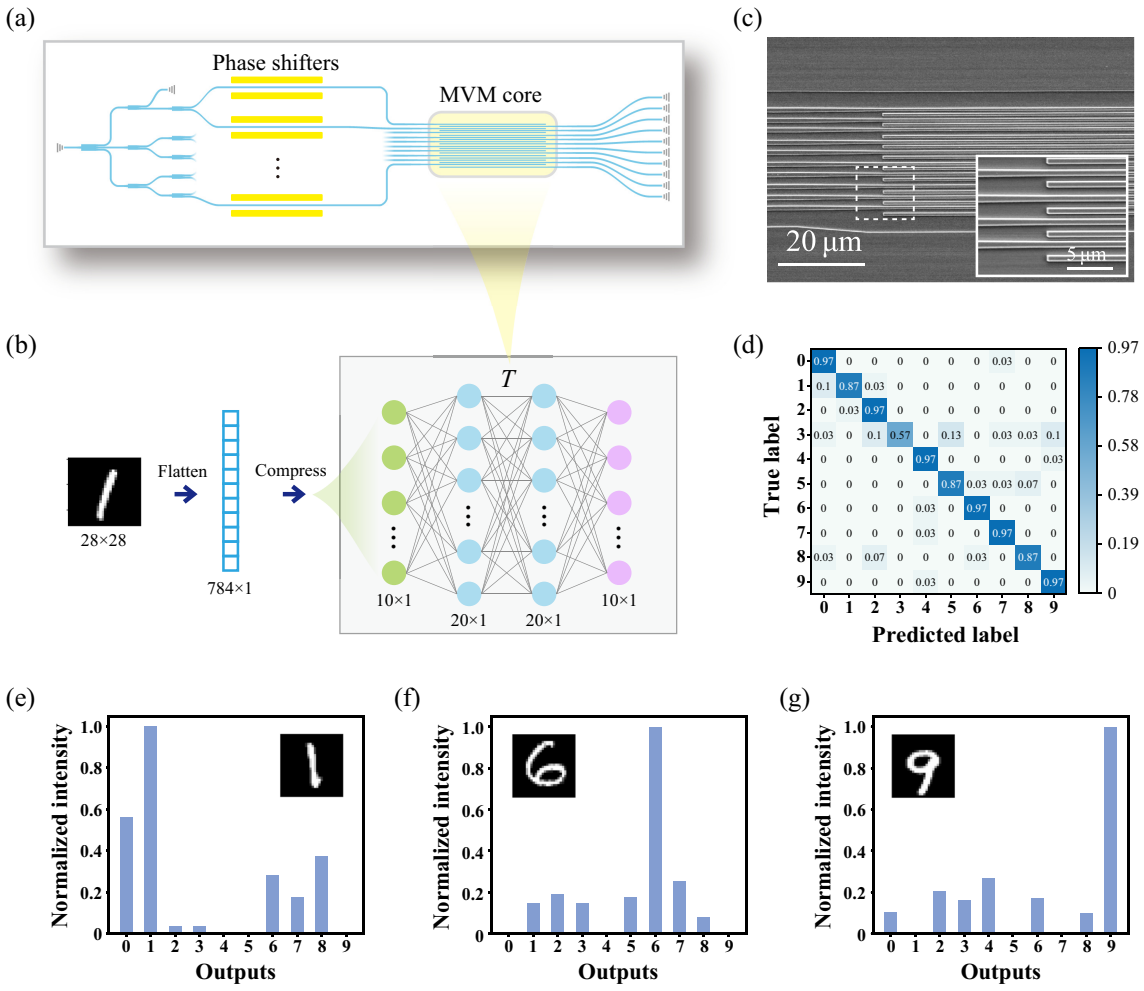


Fig. 4 Schematic illustration and experimental results of handwriting digit classification task. (a) Schematic diagram of the device for the MNIST identification task, where the number of waveguides in the array increases to 20. (b) Framework of the neural network. The features of the handwriting digit images are extracted by operations of flattening and compressing and then processed by the proposed device. (c) SEM image of the input section of the waveguide array and the enlarged view shown in the insert. (d) Confusion matrix of the experimental classification results with an accuracy of 89.7%. (e)–(g) Classification results for digits "1," "6," and "9," which are determined by the output exhibiting the maximum intensity.

Table 1 Performance comparison between our proposed WONN and other on-chip ONN architectures.

Ref.	Platform	Basic units	Throughput (TOPS)	Energy efficiency (TOPS/W)	Computational density (TOPS/mm^2)	Operating bandwidth (nm)
Google TPU[8]	Electronics	N/A	4	2	0.16	N/A
Shen et al.[21]	Si	MZI	6.4@100 GHz[d]	N/A	N/A	N/A
Feldmann et al.[28]	Si/SiN	PCM[a]	2.6@18 GHz	0.5	162	100
Ashtiani et al.[15]	Si	Attenuator + PD[b] + MRR	2@21.7 GHz	2.9	3.5	N/A
Fu et al.[30]	Si	Diffraction units	1.38×10^4@100 GHz[d]	5.082	4.6×10^4	N/A
Liu et al.[31]	Si	Diffraction units	300@10 GHz[d]	1.1	1.9×10^5	N/A
Meng et al.[14]	SiN	MMI	0.54@16.6 GHz	0.41	25.48	2.4
Ou et al.[39]	TFLN	MZI	0.98@10 GHz	40	0.0175	6
Zhang et al.[35]	TFLN	MDCs[c] + phase shifters	16.38@128 GHz	4.55	0.56	4.81
This work	TFLN	Waveguide arrays	135.6@100 GHz[d]	6.28	9.48×10^4	75

[a]PCM, phase-change material.
[b]PD, photodiode.
[c]MDCs, multiport directional couplers.
[d]The operating speed is a theoretically estimated data.

randomly selected samples from the test dataset, the system achieves a recognition accuracy of 89.7% under the multi-level compensation scheme for this increased system complexity. The confusion matrix is shown in Fig. 4(d), and the classification results for the digits "1," "6," and "9" after error compensation are presented in Figs. 4(e)–4(g), where the output with the highest intensity represents the final recognition result (see Note S5 in the Supplementary Material).

4 Discussion and Conclusion

This work demonstrates a matrix-vector multiplication scheme utilizing waveguide arrays for optical neural networks, which capitalizes on the inherent compactness, fabrication stability, and robustness of waveguide arrays, while synergistically combining them with the high-speed and low-power consumption advantages of TFLN. The resulting chip exhibits exceptional performance in computational density, computational speed, power efficiency, and operating bandwidth. Figure 1(c) and Table 1 present the key performance metrics of the WONN and the comparison with other representative works in the same field. Taking the iris flower classification task as an example, the waveguide array dimensions are merely 14.3 μm × 100 μm. Operating at the experimental speed of 10 kHz, the computational speed of this chip, based on the referenced calculation method,[13] is 1.36×10^7 OPS. This result is limited by our measurement system and does not achieve the optimal performance of the TFLN modulators. Assuming the switching operation at 100 GHz which is widely adopted in optical computation and can be achieved by TFLN modulators,[21,30,52] the computational speed can be calculated to be 135.6 TOPS, corresponding to an extremely high computational density of 9.48×10^4 TOPS/mm^2. These metrics for speed and density rank among the highest in the field of on-chip optical computing and establish a leading position for this work among implementations on the TFLN platform, as quantified in Fig. 1(c), where the work in Ref. 31 ranks at the top for its sophisticated design and compact size utilizing the diffraction units. Regarding the chip's power consumption, the dominant portion is attributed to the capacitive load of the phase shifters. Simulations indicate a phase shifter capacitance of 10 pF, which corresponds to a calculated energy efficiency of 6.28 TOPS/W (see Note S6 in the Supplementary Material). Furthermore, the device exhibits an operating bandwidth of 75 nm, underscoring the wavelength tolerance inherent to the waveguide array architecture. The future integration of wavelength division multiplexing to leverage the wavelength dimension in computation enables the expansion of channels and the increase in information capacity, further unlocking the computational potential of optical neural networks based on waveguide arrays.

In summary, we propose, for the first time, the use of waveguide arrays to perform matrix-vector multiplication and to construct an optical neural network. By flexibly controlling the inter-waveguide coupling coefficients to configure the desired transmission matrix, we achieved 100% recognition accuracy on the iris dataset and 89.7% on the MNIST dataset in the experiment, demonstrating reliable performance in optical computing. Moreover, it surpasses other TFLN platforms in computational speed, density, and operating wavelength bandwidth, securing a leading position in the field of on-chip optical neural networks, and proposing a pioneering architectural paradigm for high-speed optical computing cores. Prospective integration with dynamically programmable ONNs to harness their synergistic potential heralds a promising pathway toward scalable, high-performance optical computing.

Disclosures

The authors declare no competing interests.

Code and Data Availability

The data underlying the results presented in this paper may be obtained from the authors upon reasonable request.

Acknowledgments

The authors acknowledge the financial support from the National Natural Science Foundation of China (Grant Nos.

62288101, 62325504, 12204233, and 12174186), the Key Research and Development Program of Jiangsu Province (Grant No. BE2023083), the Natural Science Foundation of Jiangsu Province (Grant No. BK20251223), and the Fundamental Research Funds for the Central Universities (Grant No. KG202513).

References

1. E. Yurtsever et al., "A survey of autonomous driving: common practices and emerging technologies," *IEEE Access* **8**, 58443–58469 (2020).
2. X. Zhao et al., "A review of convolutional neural networks in computer vision," *Artif. Intell. Rev.* **57**(4), 99 (2024).
3. Z. Su et al., "A multigrained preference analysis method for product iterative design incorporating AI-generated review detection," *Sci. Rep.* **15**(1), 2528 (2025).
4. A. Radford et al., "Improving language understanding by generative pre-training," https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf (2018).
5. X. Zhu et al., "swPredictor: a data-driven performance model for distributed data parallelism training on large-scale HPC clusters," *Perform. Eval.* **170**, 102530 (2025).
6. T. Sun et al., "Optimizing resource allocation for geographically-distributed inference by large language models," *Perform. Eval.* **170**, 102527 (2025).
7. T. Fu et al., "Optical neural networks: progress and challenges," *Light Sci. Appl.* **13**(1), 263 (2024).
8. Y. Sun and A. M. Kist, "Deep learning on edge TPUs," arXiv: 2108.13732v2 (2022).
9. X. Lin et al., "All-optical machine learning using diffractive deep neural networks," *Science* **361**(6406), 1004–1008 (2018).
10. T. Yan et al., "Fourier-space diffractive deep neural network," *Phys. Rev. Lett.* **123**(2), 023901 (2019).
11. M. Miscuglio et al., "Massively parallel amplitude-only Fourier neural network," *Optica* **7**(12), 1812–1819 (2020).
12. K. P. Kalinin et al., "Analog optical computer for AI inference and combinatorial optimization," *Nature* **645**(8080), 354–361 (2025).
13. T. Zhou et al., "Large-scale neuromorphic optoelectronic computing with a reconfigurable diffractive processing unit," *Nat. Photonics* **15**(5), 367–373 (2021).
14. X. Meng et al., "Compact optical convolution processing unit based on multimode interference," *Nat. Commun.* **14**(1), 3000 (2023).
15. F. Ashtiani, A. J. Geers, and F. Aflatouni, "An on-chip photonic deep neural network for image classification," *Nature* **606**(7914), 501–506 (2022).
16. J. Spall, X. Guo, and A. I. Lvovsky, "Training neural networks with end-to-end optical backpropagation," *Adv. Photonics* **7**(1), 016004 (2025).
17. F. Yu et al., "Time-wavelength multiplexed photonic neural network accelerator for distributed acoustic sensing systems," *Adv. Photonics* **7**(2), 026008 (2025).
18. W. Shi et al., "Review of nonlinear activation functions in optical neural networks," *Adv. Photonics* **7**(6), 064004 (2025).
19. H. Zhang et al., "An optical neural chip for implementing complex-valued neural network," *Nat. Commun.* **12**(1), 457 (2021).
20. H. Zhu et al., "Space-efficient optical computing with an integrated chip diffractive neural network," *Nat. Commun.* **13**(1), 1044 (2022).
21. Y. Shen et al., "Deep learning with coherent nanophotonic circuits," *Nat. Photonics* **11**(7), 441–446 (2017).
22. S. Xu et al., "Optical coherent dot-product chip for sophisticated deep learning regression," *Light Sci. Appl.* **10**(1), 221 (2021).
23. Z. Xu et al., "Large-scale photonic chiplet Taichi empowers 160-TOPS/W artificial general intelligence," *Science* **384**(6692), 202–209 (2024).
24. S. Xu et al., "High-order tensor flow processing using integrated photonic circuits," *Nat. Commun.* **13**(1), 7970 (2022).
25. B. Bai et al., "Microcomb-based integrated photonic processing unit," *Nat. Commun.* **14**(1), 66 (2023).
26. R. Yin et al., "Integrated WDM-compatible optical mode division multiplexing neural network accelerator," *Optica* **10**(12), 1709–1718 (2023).
27. S. Jiang et al., "MRR-assisted MZI crossbar array for energy-efficient optical parallel computing," *Laser Photonics Rev.* **19**(24), e01035 (2025).
28. J. Feldmann et al., "Parallel convolutional processing using an integrated photonic tensor core," *Nature* **589**(7840), 52–58 (2021).
29. J. Feldmann et al., "All-optical spiking neurosynaptic networks with self-learning capabilities," *Nature* **569**(7755), 208–214 (2019).
30. T. Fu et al., "Photonic machine learning with on-chip diffractive optics," *Nat. Commun.* **14**(1), 70 (2023).
31. W. Liu et al., "Ultra-compact multi-task processor based on in-memory optical computing," *Light Sci. Appl.* **14**(1), 134 (2025).
32. Z. Wang et al., "Integrated photonic metasystem for image classifications at telecommunication wavelength," *Nat. Commun.* **13**(1), 2131 (2022).
33. H. Yang et al., "Near-energy-free photonic Fourier transformation for convolution operation acceleration," *Adv. Photonics* **7**(5), 056007 (2025).
34. J. Zhang et al., "Highly integrated all-optical nonlinear deep neural network for multi-thread processing," *Adv. Photonics* **7**(4), 046003 (2025).
35. X. Zhang et al., "Unlocking high-speed and energy-efficiency: integrated convolution processing on thin-film lithium niobate," *Laser Photonics Rev.* **19**(10), 2401583 (2025).
36. Y. Zheng et al., "Photonic neural network fabricated on thin film lithium niobate for high-fidelity and power-efficient matrix computation," *Laser Photonics Rev.* **18**(10), 2400565 (2024).
37. Y. Hu et al., "Integrated lithium niobate photonic computing circuit based on efficient and high-speed electro-optic conversion," *Nat. Commun.* **16**(1), 8178 (2025).
38. Z. Lin et al., "120 GOPS photonic tensor core in thin-film lithium niobate for inference and *in situ* training," *Nat. Commun.* **15**(1), 9081 (2024).
39. S. Ou et al., "Hypermultiplexed integrated photonics-based optical tensor processor," *Sci. Adv.* **11**(23), eadu0228 (2025).
40. F. Labbé et al., "Thin-film lithium niobate quantum photonics: review and perspectives," *Adv. Photonics* **7**(4), 044002 (2025).
41. W. Song et al., "Artificial gauge fields in photonics," *Nat. Rev. Phys.* **7**(11), 606–620 (2025).
42. Z. Lin et al., "Observation of topological transition in Floquet non-Hermitian skin effects in silicon photonics," *Phys. Rev. Lett.* **133**(7), 073803 (2024).
43. S. Wu et al., "Approaching the adiabatic infimum of topological pumps on thin-film lithium niobate waveguides," *Nat. Commun.* **15**(1), 9805 (2024).
44. R. Tang et al., "Ten-port unitary optical processor on a silicon photonic chip," *ACS Photonics* **8**(7), 2074–2080 (2021).
45. M. Markowitz, K. Zelaya, and M.-A. Miri, "Learning arbitrary complex matrices by interlacing amplitude and phase masks with fixed unitary operations," *Phys. Rev. A* **110**(3), 033501 (2024).
46. R. Tanomura et al., "Scalable and robust photonic integrated unitary converter based on multiplane light conversion," *Phys. Rev. Appl.* **17**(2), 024071 (2022).
47. A. Yariv, "Coupled-mode theory for guided-wave optics," *IEEE J. Quantum Electron.* **9**(9), 919–933 (1973).
48. W.-P. Huang, "Coupled-mode theory for optical waveguides: an overview," *J. Opt. Soc. Amer. A* **11**(3), 963–983 (1994).
49. Q. Cheng et al., "Asymmetric topological pumping in nonparaxial photonics," *Nat. Commun.* **13**(1), 249 (2022).
50. Z. Wang et al., "Fast-speed and low-power-consumption optical phased array based on lithium niobate waveguides," *Nanophotonics* **13**(13), 2429–2436 (2024).

51. J. Ji et al., "Dynamic holographic display with addressable on-chip metasurface network based on lithium niobate photonics," *Light Sci. Appl.* **14**(1), 332 (2025).
52. C. Wang et al., "Integrated lithium niobate electro-optic modulators operating at CMOS-compatible voltages," *Nature* **562**(7725), 101–104 (2018).
53. Y. Chen et al., "All-analog photoelectronic chip for high-speed vision tasks," *Nature* **623**(7985), 48–57 (2023).
54. S. Pai et al., "Experimentally realized *in situ* backpropagation for deep learning in photonic neural networks," *Science* **380**(6643), 398–403 (2023).

Xueyun Li is a PhD candidate working under supervision of Tao Li at the College of Engineering and Applied Sciences, Nanjing University. He received his bachelor's degree from Nanjing University. He is currently working on nanophotonic integrations and optical computation.

Zhizhang Wang is an assistant researcher at the College of Engineering and Applied Sciences, Nanjing University. He received his PhD from Nanjing University. He is currently focusing on nanophotonic integrations and LNOI devices.

Tao Li is a professor at the College of Engineering and Applied Sciences, Nanjing University. He received his PhD from Nanjing University in 2005. His research interest includes optical metamaterials, topological photonics, and on-chip photonic integrations.

Biographies of the other authors are not available.