

REVIEW ARTICLE

Open Access

Optical neural networks: progress and challenges

Tingzhao Fu^{1,2,3}, Jianfa Zhang^{1,2,3}, Run Sun^{4,5}, Yuyao Huang^{4,5}, Wei Xu^{1,2,3}, Sigang Yang^{4,5}, Zhihong Zhu^{1,2,3} and Hongwei Chen^{4,5}✉

Abstract

Artificial intelligence has prevailed in all trades and professions due to the assistance of big data resources, advanced algorithms, and high-performance electronic hardware. However, conventional computing hardware is inefficient at implementing complex tasks, in large part because the memory and processor in its computing architecture are separated, performing insufficiently in computing speed and energy consumption. In recent years, optical neural networks (ONNs) have made a range of research progress in optical computing due to advantages such as sub-nanosecond latency, low heat dissipation, and high parallelism. ONNs are in prospect to provide support regarding computing speed and energy consumption for the further development of artificial intelligence with a novel computing paradigm. Herein, we first introduce the design method and principle of ONNs based on various optical elements. Then, we successively review the non-integrated ONNs consisting of volume optical components and the integrated ONNs composed of on-chip components. Finally, we summarize and discuss the computational density, nonlinearity, scalability, and practical applications of ONNs, and comment on the challenges and perspectives of the ONNs in the future development trends.

Introduction

Artificial intelligence (AI) has penetrated diverse fields in society and obtained achievements beyond humanity. Neural networks play a crucial role as the core technology supporting the development of AI. In the 1940s, McCulloch and Pitts introduced the working principle of neural networks and the structure of neurons^{1,2}. In 1949, Hebb systematically elucidated the theory of neuropsychology³. During the 1950s, critical issues regarding the development of AI were discussed and raised⁴, which promoted the crucial technologies in neural networks and the establishment of the discipline of AI. In 1958, Rosenblatt systematically introduced the mathematical model and working principle of perceptron⁵, laying a significant foundation for the further development of neural networks. In 1986, following early pioneering work, Rumelhart, Hinton, and Williams proposed the error back-propagation algorithm to train multi-layer perceptrons⁶.

In 1990, LeCun et al. introduced a convolutional neural network based on the back-propagation algorithm and demonstrated its application performance in handwritten digit recognition⁷. In 2012, Krizhevsky et al. proposed deep convolutional neural networks⁸, which further improved the inference abilities of neural networks and made the application fields of AI more extensive^{9–13}.

In the meantime, the rapid development of semiconductor process technology has created a host of advanced electronic computing hardware^{14–20} with better performance than CPUs in the past decade. The advanced computing hardware predominantly ensures the computing efficiency requirements of neural networks during the iteration process, thereby promoting the rapid development and application of AI in multitudinous fields. However, the machining accuracy of semiconductor manufacturing technology has approached 3nm²¹, and the size has drawn near the physical limit of the transistor. Transistors with such sizes are highly susceptible to quantum tunneling and thermal effects, making them difficult to work well. Accordingly, improving the machining accuracy of semiconductor processes to acquire higher computing power will be unsustainable. The development of neural network technology relies on

Correspondence: Hongwei Chen (chenhw@tsinghua.edu.cn)

¹College of Advanced Interdisciplinary Studies, National University of Defense Technology, Changsha, China

²Hunan Provincial Key Laboratory of Novel Nano-Optoelectronic Information Materials and Devices, National University of Defense Technology, Changsha, China

Full list of author information is available at the end of the article

© The Author(s) 2024



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



massive data, advanced algorithms, high computing power, and contemporary social demands. In the future, neural networks may face bottlenecks in meeting computing power requirements during training or inference processes. Therefore, exploring a new computational paradigm is of great significance.

Extensive matrix operations in the training or inferring stage of neural networks can be equivalent to the propagation process of light. Due to the advantages of low latency, low power consumption, large bandwidth, and parallel signal processing of light, the matrix operations process can be executed by modulating the optical feature quantities (amplitude, phase, polarization, angular momentum, etc.) during the light propagation process. Thus, further designing optical systems to achieve the inference function of ONNs is more advantageous than its electronic counterparts. As early as the 1960s and 1970s, preliminary basic research regarding ONNs was carried out, e.g., optical signal detection²² and optical signal transmission in optical systems²³. Afterward, in 1985, Farhat et al.²⁴ proposed an optical implementation method for the Hopfield model. In 1987, Fisher et al.²⁵ presented a method for implementing optical networks with variable adaptive learning capability. In 1989, Caulfield et al.²⁶ systematically introduced ONNs and pointed out that ONNs emerged from the mutual infiltration of traditional optical information processing and neural network knowledge systems. Caulfield et al.²⁶ consider ONNs would be superior to electronic neural networks in certain situations. In 1990, Psaltis et al.²⁷ introduced a method for implementing nonlinear functionality in ONNs using photoreactive crystals. In 1994, Reck et al.²⁸ used optical devices such as beam splitters and phase shifters to achieve the operation function of unitary matrices in a cascaded manner. In 2016, Clements et al.²⁹ proposed a matrix factorization method based on the Mach–Zehnder interferometer (MZI) cascade approach. Tait et al.³⁰ achieved wavelength filtering and power allocation for micro-ring resonators (MRRs). These research works have laid a solid and crucial theoretical foundation for the subsequent development of ONNs. Figure 1 shows the timeline of partial emblematic works in the progression of ONNs.

ONNs have made hosts of research progress^{31–35} through the continuous exploration and efforts of predecessors. This article will provide a comprehensive introduction to the evolution of ONNs in recent years. Firstly, we briefly introduce the principle of ONNs, including the structure of ONNs, the model of optical neurons, and the implementation methods of matrix multiplication function. Secondly, we systematically survey the research of ONNs in recent years from two modules (divided into seven aspects, Fig. 2), including

non-integrated ONNs based on volume optical elements and integrated ONNs composed of on-chip optical components. Finally, we summarize and analyze the current advantages and challenges of ONNs regarding computational density, nonlinearity, scalability, practical applications, etc. We will also provide an outlook and discussion on the future development trends of ONNs.

Principle of ONNs

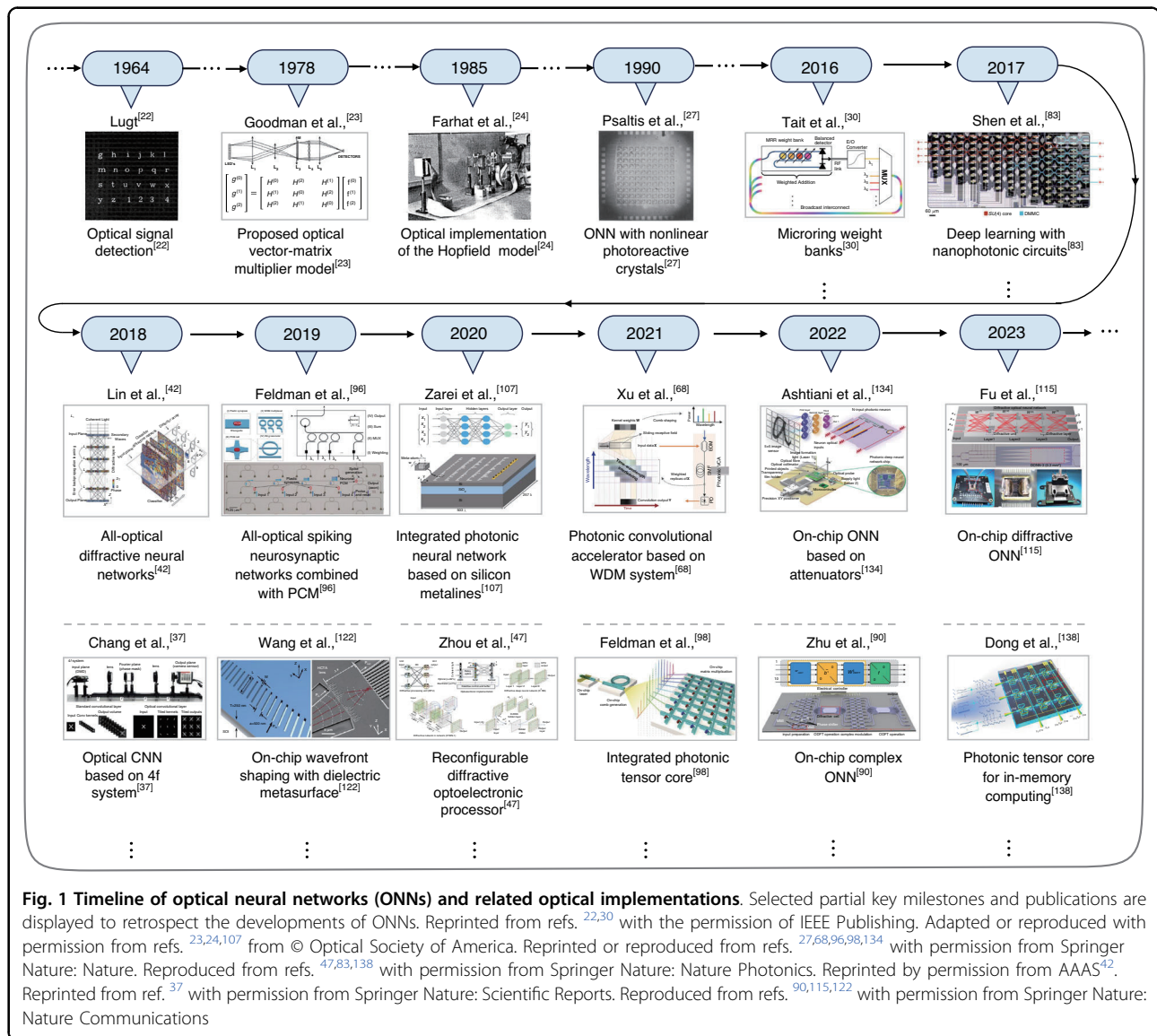
Biological neuron model and artificial neural network

Neural networks are mathematical models established by imitating the human brain's nervous system. The functions effectuated by artificial neurons mimic those of dendrites, cell nucleus, axons, and synapses in biological neurons. Figure 3a, b show the abstract structure of a biological neuron and its logical inferring model. In Fig. 3b, w_n ($n = 1, 2, \dots$) is referred to as a weight, which is a parameter that controls the importance of the input signal. b is bias, which is a parameter that adjusts the ease of activation of neurons. Furthermore, artificial neurons can construct a neural network (multi-layer perceptron) by combining various weight connections, as depicted in Fig. 3c.

Implementation of optical matrix operations

As shown in Fig. 3c, the neural network consists of many interconnected neurons, and the connections between these neurons are termed as weights. The neurons on the first hidden layer interact with the input signals through the first weight matrix (w_{ij}^1), and then such neurons transmit the calculated results as a new signal to the neurons on the next hidden layer. Each neuron in the entire neural network repeats the above process until the input signals reach the output end. Optical systems can achieve the function of matrix multiplications, and the implementation methods among different optical systems are diverse, including MZI mesh (Fig. 4a), MRR weight banks (Fig. 4b), and other optical components (Fig. 4c). It is worth noting that the weight matrix values implemented by the optical system mentioned above can correspond one-to-one with the weight matrix values in Fig. 3c, which can be obtained through training by optimization algorithms. However, for ONNs constructed based on diffractive components, namely diffractive optical neural networks (DONNs), which are shown in Figs. 4d, e, the weights (w_{ij}^n , $n = 1, 2, \dots$) between the adjacent layers, such as the input layer, hidden layers, and the output layer, are fixed due to the connection of neurons through light diffraction. The training parameters of hidden layers (such as T_{ij} in Fig. 4f) are usually the transmission coefficients of sub-wavelength structures on the diffractive layers. The weight matrix (T_{ij}) that DONN needs to train is different from the weight matrix values (w_{ij}^n) of the ONNs based on the





MZI mesh, MRR weight banks, and other optical components.

Non-integrated ONNs based on volume elements ONNs based on volume optical 4f system

The optical 4f system is a typical optical transfer system consisting of two lenses with a focal length of f , which can achieve a Fourier transform of the input optical field. After the input signal is transformed from the time domain to the frequency domain, the frequency spectrum of the input signal can be modified or modulated conveniently in the Fourier plane to obtain the corresponding output response. The work of using a 4f system to perform matrix multiplication has been validated³⁶. The frequency domain information at the Fourier plane can be designed and modulated to achieve the inference function of ONNs^{37–41}.

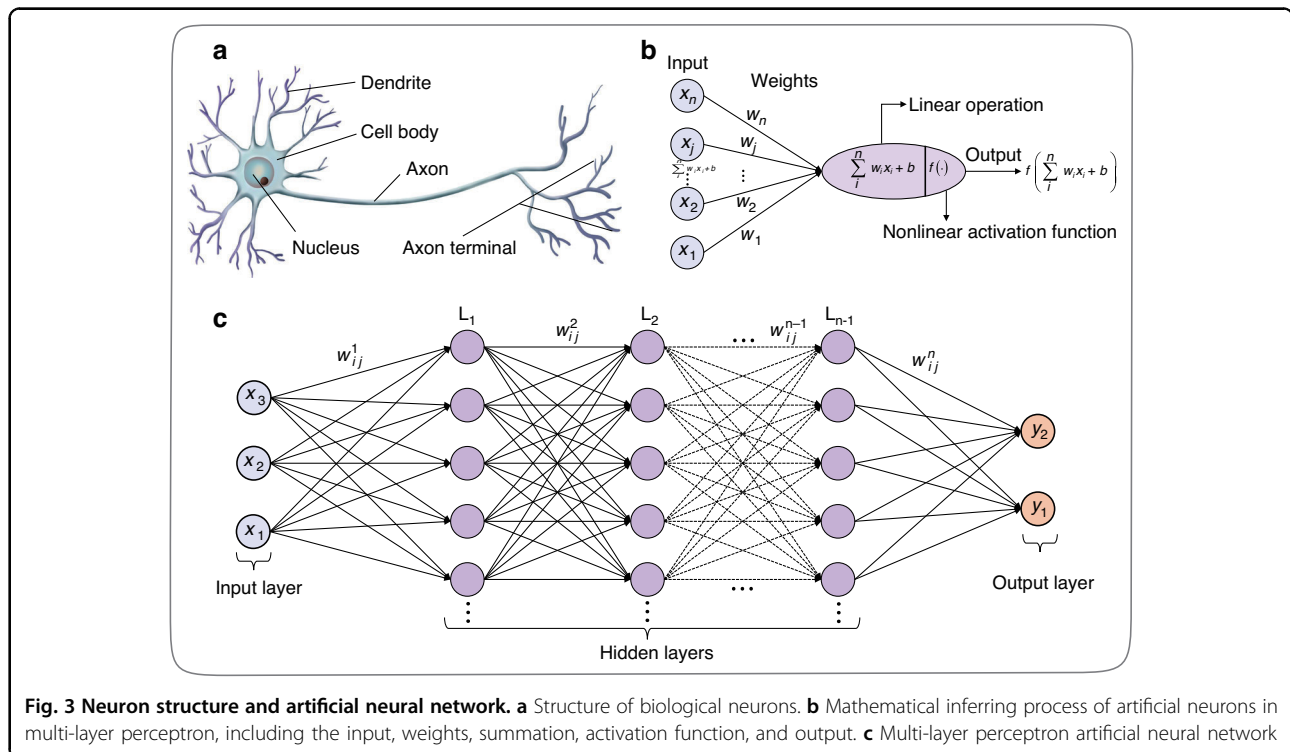
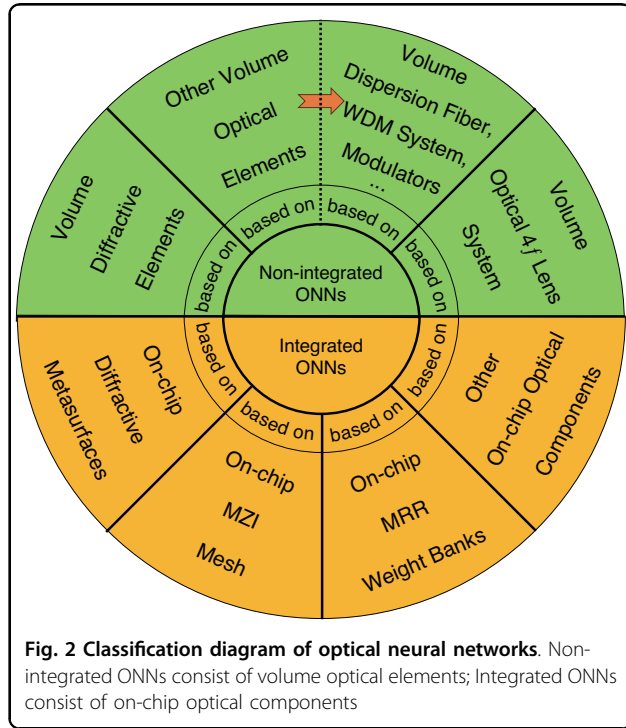
In 2018, Chang et al.³⁷ proposed a scheme for devising optical convolutional neural networks (OCNNs) based on the 4f system (Fig. 5a), which involves placing a phase mask in the Fourier plane of the 4f system to achieve the function of convolutional kernels. Although the convolutional kernels cannot support parameter reconfiguration, the OCNN brings lower training costs, lower computation, and better predictive performance to the optoelectronic hybrid systems. In this work, the nonlinear function is fulfilled in the electrical domain. In July 2019, Yan et al.³⁸ placed a diffractive depth neural network (D²NN) in the Fourier plane of a 4f system (Fig. 5b) through numerical simulation, which significantly improved the classification accuracy and robustness of the optical system by adding an optical nonlinear activation function (optical characteristic parameters of ferroelectric

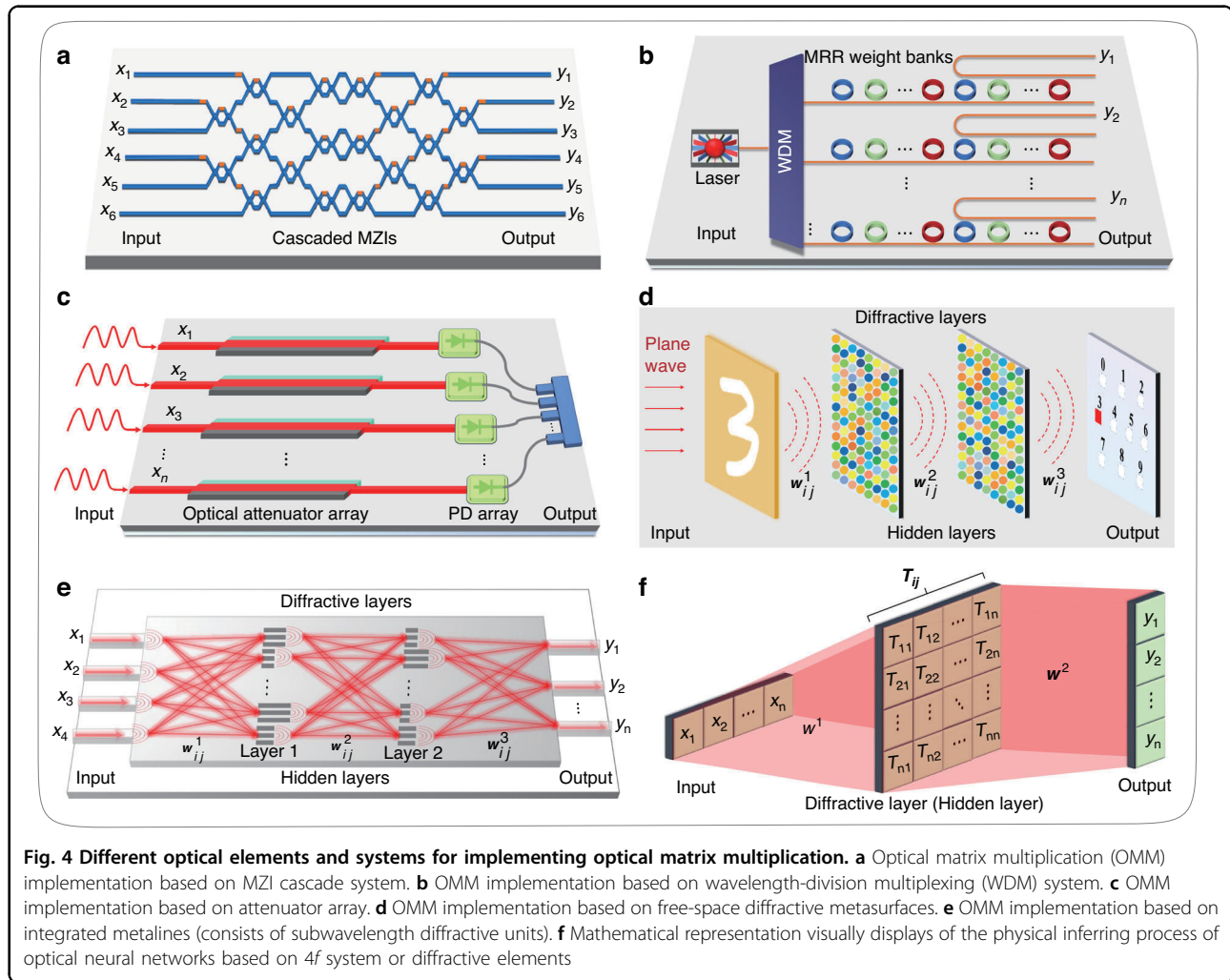


thin films) to the D²NN. Yan et al.³⁸ compared and analyzed the performance of the optical system after training with/without nonlinear activation layers of ferroelectric thin films introduced during the training process, as well

as the number of nonlinear activation layers introduced. This work may provide theoretical guidance for designing and fabricating nonlinear devices in the subsequent 4f system.

In August 2019, Zou et al.³⁹ used a spatial light modulator (SLM) and Fourier lens to complete the input signal loading and weight matrix construction, and the introduced SLM made the matrix operation process programmable in real-time. In addition, Zou et al.³⁹ introduced nonlinear optical media in the matrix operation process, endowing the operation process of ONNs with nonlinearity. The specific physical implementation method is shown in Fig. 6a. On this basis, the research group used two SLMs and a 4f system group to complete a two-layer all-optical neural network with nonlinear and reconfigurable functions, as shown in Fig. 6b. This work enhances the functionality (implemented nonlinearity and reconfigurability) of the ONN based on the 4f system and endows it with more powerful logical inferring abilities. In 2020, Miscuglio et al.⁴⁰ proposed a massively parallel amplitude-only Fourier ONN (Figs. 6c, d). Specifically, the kilohertz-fast reprogrammable high-resolution digital micromirror devices (DMDs) were introduced in the 4f system to implement input signal loading and Fourier plane signal modulation. The modulation rate of DMD is at least 2 orders of magnitude faster than that of SLM with the same pixel resolution, making the ONN system faster and more efficient.





ONNs based on discrete optical diffractive elements

The trainable parameters (modulation units) of the ONNs built based on the 4f system are used to be placed in the Fourier plane^{37–40}, which limits the expansion of trainable parameters and the number of hidden layers. However, free-space DONNs constructed by diffractive elements^{42–59} can overcome these limitations well. In 2018, Lin et al.⁴² proposed an all-optical deep learning framework, in which neural networks are physically formed by multiple layers of diffractive surfaces (Fig. 7a). The diffractive surfaces are composed of diffractive units, each of which is termed as a hidden layer in DONNs, and each diffractive unit on the diffractive surface is defined as a neuron. Meanwhile, the function of the diffractive unit is to change the phase difference after the light passes, and the specific phase difference values can be pre-trained on the computer through intelligent algorithms such as forward propagation, gradient descent, error back-propagation, etc. Among them, the forward propagation process (wave analysis) is demonstrated by the Rayleigh-

Sommerfeld diffraction equation⁴². After obtaining all phase difference values, the diffractive surfaces can be fabricated using 3D printing technology to obtain the physical implementation device of DONNs (Fig. 7c). The relationship between the thickness of diffractive units and phase difference values is linear, and the specific value of the thickness of each diffractive unit can be calculated by the formula: $h = \lambda\phi / (2\pi\Delta n)$, where Δn is the refractive index difference between the 3D printing material and air, ϕ is the phase difference value (trainable parameter), λ is the wavelength of light propagation in free-space. Finally, Lin et al.⁴² validated the classification tasks of the MNIST and Fashion MNIST datasets by encoding the input signal into the amplitude and phase of light, respectively. This work proposed a novel scheme for the study of ONNs. In 2 years, Qian et al.⁴⁴ designed a DONN (Fig. 7d) for logic gate operations (AND, OR, NOT, etc.) based on Rayleigh-Sommerfeld diffraction. The hidden layers function of the DONN is implemented by the Huygens metasurfaces fabricated through mechanical processing.



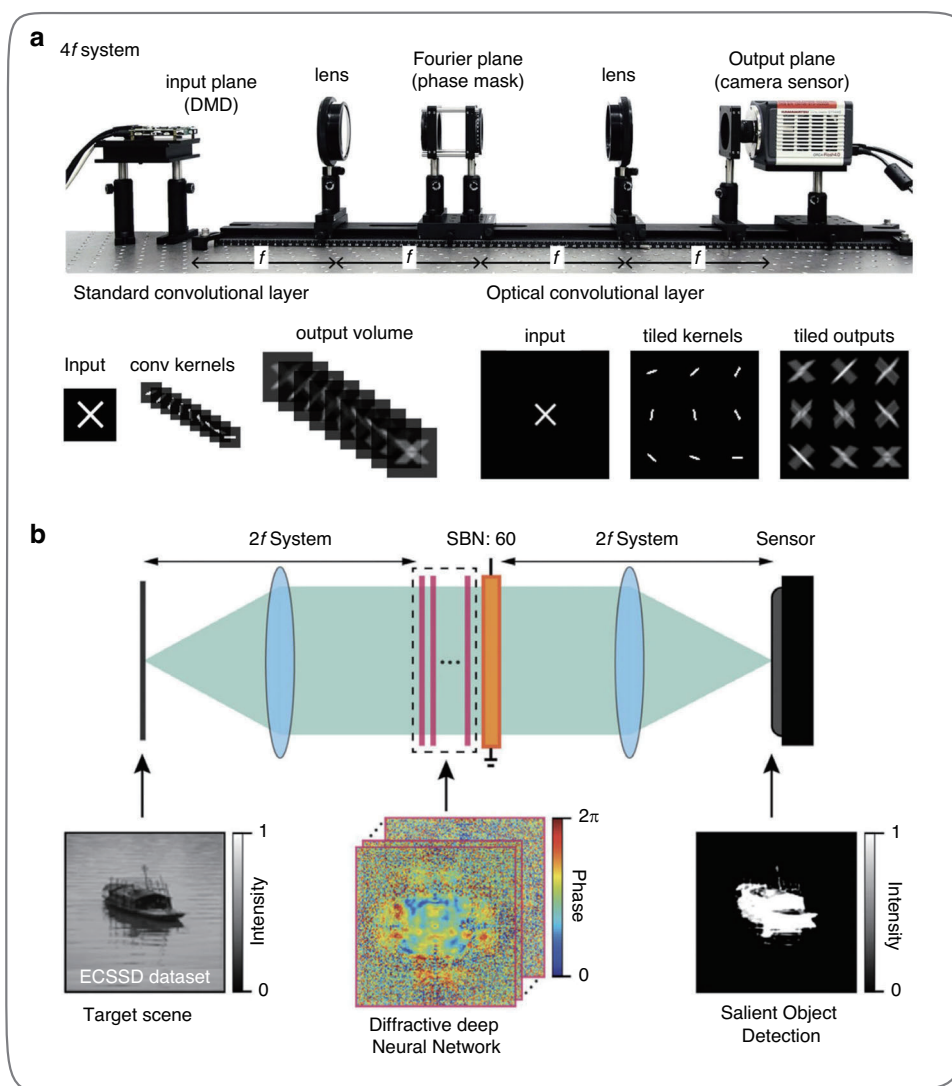


Fig. 5 ONNs implementation based on 4f system. **a** Optoelectronic hybrid convolutional neural network with phase mask placed in the Fourier plane of the 4f system³⁷. **b** All-optical neural network with diffractive depth neural network placed in the Fourier plane of the 4f system³⁸. **a** Reprinted from ref.³⁷ with permission from Springer Nature: Scientific Reports. **b** Reprinted with permission from ref.³⁸ from © The American Physical Society

Noteworthiness, the current DONNs^{42,44} does not have reconfigurable functionality, and nonlinearity is not introduced in the network except for the output layer (optical intensity detection⁶⁰). In addition, the fabricating errors generated during the machining process might accumulate with the depth increases of ONNs, resulting in inevitable systematic errors. The difficulties mentioned above extensively limit the model complexity and experimental performance of existing DONN processors. Thus, in 2021, Zhou et al.⁴⁷ proposed a reconfigurable diffractive processing unit (DPU), as shown in Fig. 8a, based on which an optoelectronic fusion computing architecture can be constructed, and this architecture can support different neural networks and achieve high model

complexity with millions of neurons. The DPU consists of an input layer, an information processing layer, and an output layer. The input data of the DPU input layer is optically encoded by DMD, and the physical process of the information processing layer is implemented by SLM and optical diffraction. The optical field summation and nonlinear activation function of the output layer are achieved by the photoelectric effect on each pixel of the complementary metal-oxide-semiconductor (CMOS) sensor (Fig. 8b). In this work, Zhou et al.⁴⁷ designed a feedforward deep ONN (Fig. 8c) and a diffractive recurrent ONN (Fig. 8d) using DPU as the basic units. In 2022, Liu et al.⁵³ designed a programmable DONN (Fig. 8e) based on information metasurfaces⁶¹. The meta-structure

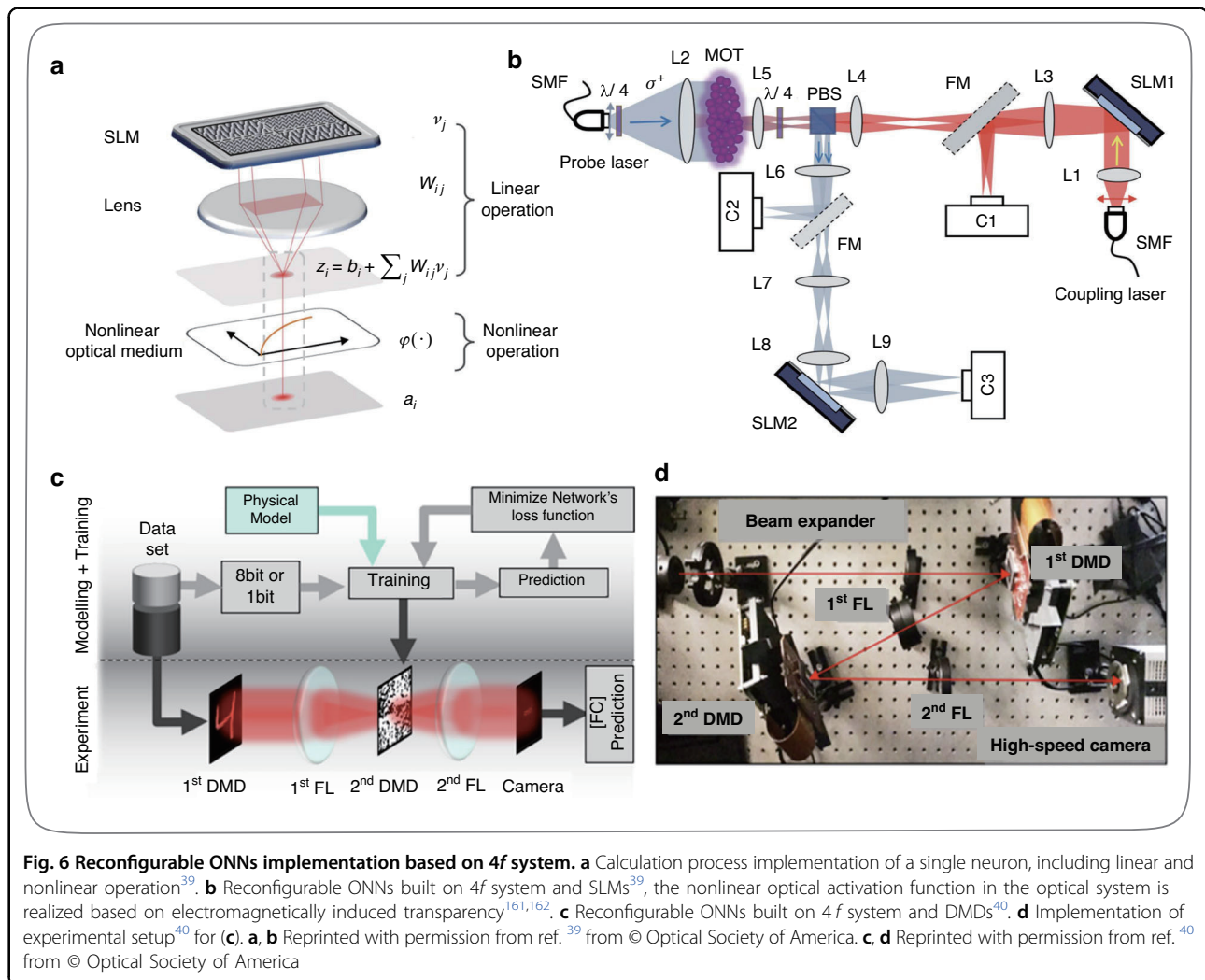


Fig. 6 Reconfigurable ONNs implementation based on 4f system. **a** Calculation process implementation of a single neuron, including linear and nonlinear operation³⁹. **b** Reconfigurable ONNs built on 4f system and SLMs³⁹, the nonlinear optical activation function in the optical system is realized based on electromagnetically induced transparency^{161,162}. **c** Reconfigurable ONNs built on 4f system and DMDs⁴⁰. **d** Implementation of experimental setup⁴⁰ for (c). **a, b** Reprinted with permission from ref. ³⁹ from © Optical Society of America. **c, d** Reprinted with permission from ref. ⁴⁰ from © Optical Society of America

(neuron) arrays on such a metasurface (hidden layer) can be uniformly routed and controlled through the field-programmable gate arrays (FPGA), thereby achieving real-time control of the amplitude and phase coefficients of each neuron. In other words, the values of each neuron on the DONN hidden layers can be flexibly set. This work is expected to promote the application of DONN in the microwave fields, such as remote control, wireless communication, signal enhancement, medical imaging, and the Internet of Things. It is worth mentioning that the introduction of programmable information metasurfaces enhances the programmability of DONNs, and embedding optical power amplification devices on each diffractive unit is conducive to further expanding the depth of DONNs.

ONNs based on other volume optical components

Other volume optical components refer to the ordinary single-mode fibers, dispersion fibers,

modulators, attenuators, filters, Fabry-Perot laser with saturable absorber, polarizing beam splitter, wavelength-division multiplexing (WDM) system, etc., which can flexibly construct ONNs in different combinations^{62–76}. In 2012, Duport et al.⁶² achieved the experimental demonstration of all-optical reservoir calculation for the first time based on single-mode fibers, semiconductor optical amplifiers (SOA), tunable optical attenuators, delay loops, bandpass filters, etc. (Fig. 9a). Meanwhile, the nonlinearity is provided by the saturation of the optical gain in the SOA. In 2021, Stelzer et al.⁷¹ constructed the computational function of a single neuron using a nonlinear device and multiple time-delay feedback loops and implemented a deep neural network of any size by continuously iteratively reconstructing the parameters of the single neuron. The network's connection weights were implemented by adjusting the feedback modulation signal and delay within the loop, as shown in Fig. 9b.



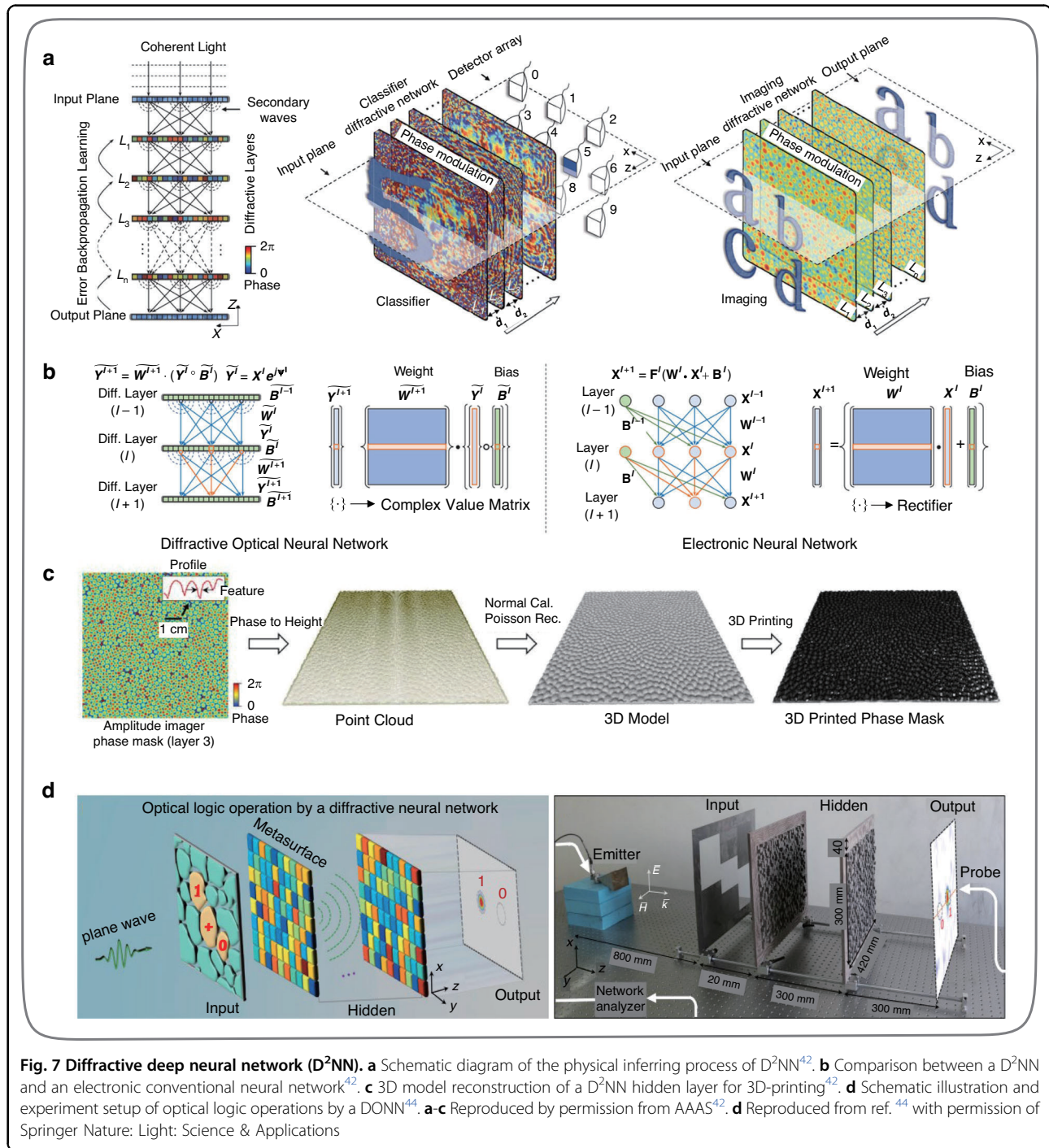


Fig. 7 Diffractive deep neural network (D²NN). **a** Schematic diagram of the physical inferring process of D²NN⁴². **b** Comparison between a D²NN and an electronic conventional neural network⁴². **c** 3D model reconstruction of a D²NN hidden layer for 3D-printing⁴². **d** Schematic illustration and experiment setup of optical logic operations by a DONN⁴⁴. **a–c** Reproduced by permission from AAAS⁴². **d** Reproduced from ref. ⁴⁴ with permission of Springer Nature: Light: Science & Applications

In 2021, Xiang et al.⁶⁹ designed photonic neurons and synapses based on the vertical-cavity surface-emitting laser with an embedding saturable absorber (VCSEL-SA) and the vertical-cavity semiconductor optical amplifiers (VCSOA) respectively, and based on which they constructed an all-optical spiking neural network (SNN), as shown in Fig. 9c. This work built up a new framework for classification tasks through supervised learning manner

and developed a self-consistent unified neuron-synapse-learning model, providing a hardware-friendly method for implementing SNN in the optical domain. In the same year, Xu et al.⁶⁸ designed a universal optical vector convolution accelerator adopting the time-wavelength interleaving method (Fig. 9d), and then implemented the function of the optical convolution accelerator on hardware based on an optical frequency comb and WDM

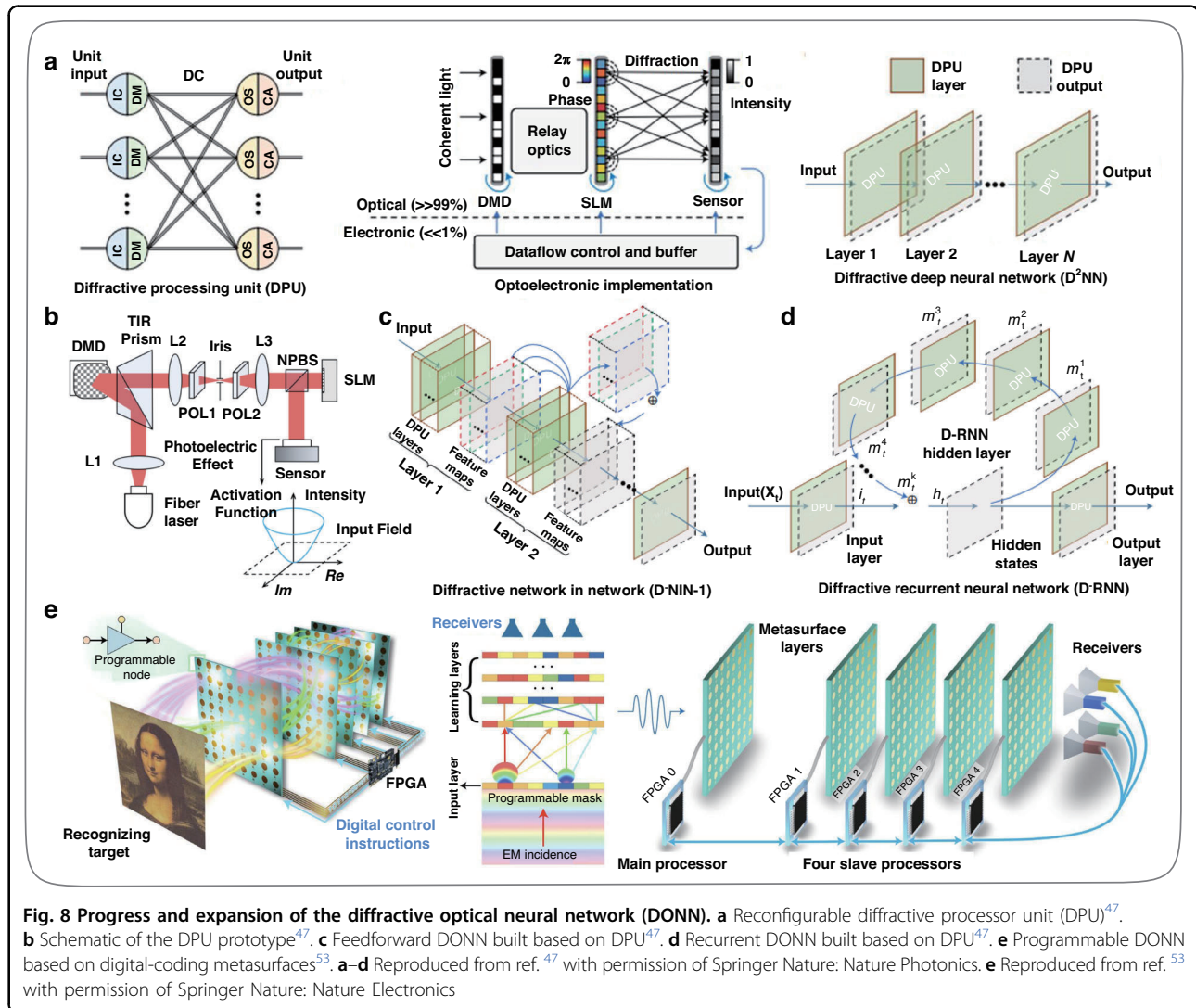


Fig. 8 Progress and expansion of the diffractive optical neural network (DONN). **a** Reconfigurable diffractive processor unit (DPU)⁴⁷. **b** Schematic of the DPU prototype⁴⁷. **c** Feedforward DONN built based on DPU⁴⁷. **d** Recurrent DONN built based on DPU⁴⁷. **e** Programmable DONN based on digital-coding metasurfaces⁵³. **a–d** Reproduced from ref. ⁴⁷ with permission of Springer Nature: Nature Photonics. **e** Reproduced from ref. ⁵³ with permission of Springer Nature: Nature Electronics

system, reaching a computing speed at more than 10 TOPS (trillions of operations per second). The input of the optical convolution accelerator is completed through an on-chip optical frequency comb, while its output, weight matrix operation process, and convolution kernel allocation are completed through a WDM system. Furthermore, Xu et al.⁶⁸ designed an optical convolutional neural network for recognizing handwritten digital images, with an experimental recognition accuracy of 88%, which is very close to 90% of the numerical calculation results obtained through optimization training. The method proposed by Xu et al.⁶⁸ can be used for training more complex networks, which is promising to be applied to complex scenes such as autonomous vehicles and real-time video recognition. In 2023, Xiang et al.⁷³ developed a photonic spiking neuron chip based on an integrated Fabry-Perot laser with a saturable absorber (FP-SA), which provides an indispensable foundational module for

constructing photonic spiking neural network (PSNN) hardware (Fig. 9e). Additionally, they proposed time-multiplexed temporal spike encoding to achieve a functional PSNN that far exceeds hardware integration scale limitations, paving the way for multi-layer PSNN with nonlinearity that can handle complex tasks. Besides designing spiking neurons based on FP-SA, many other methods can also achieve the functionality of spiking neurons^{77–81}, and further construct PSNNs.

Integrated ONNs based on on-chip optical components

With the assistance of advanced semiconductor process technologies, miniaturization proves to be a significant trend in the development of ONNs. Compared with ONNs constructed with volume optical elements, on-chip integrated ONNs have advantages such as high computational density, portability, and stability, which may play a crucial role in the birth of new computing machines



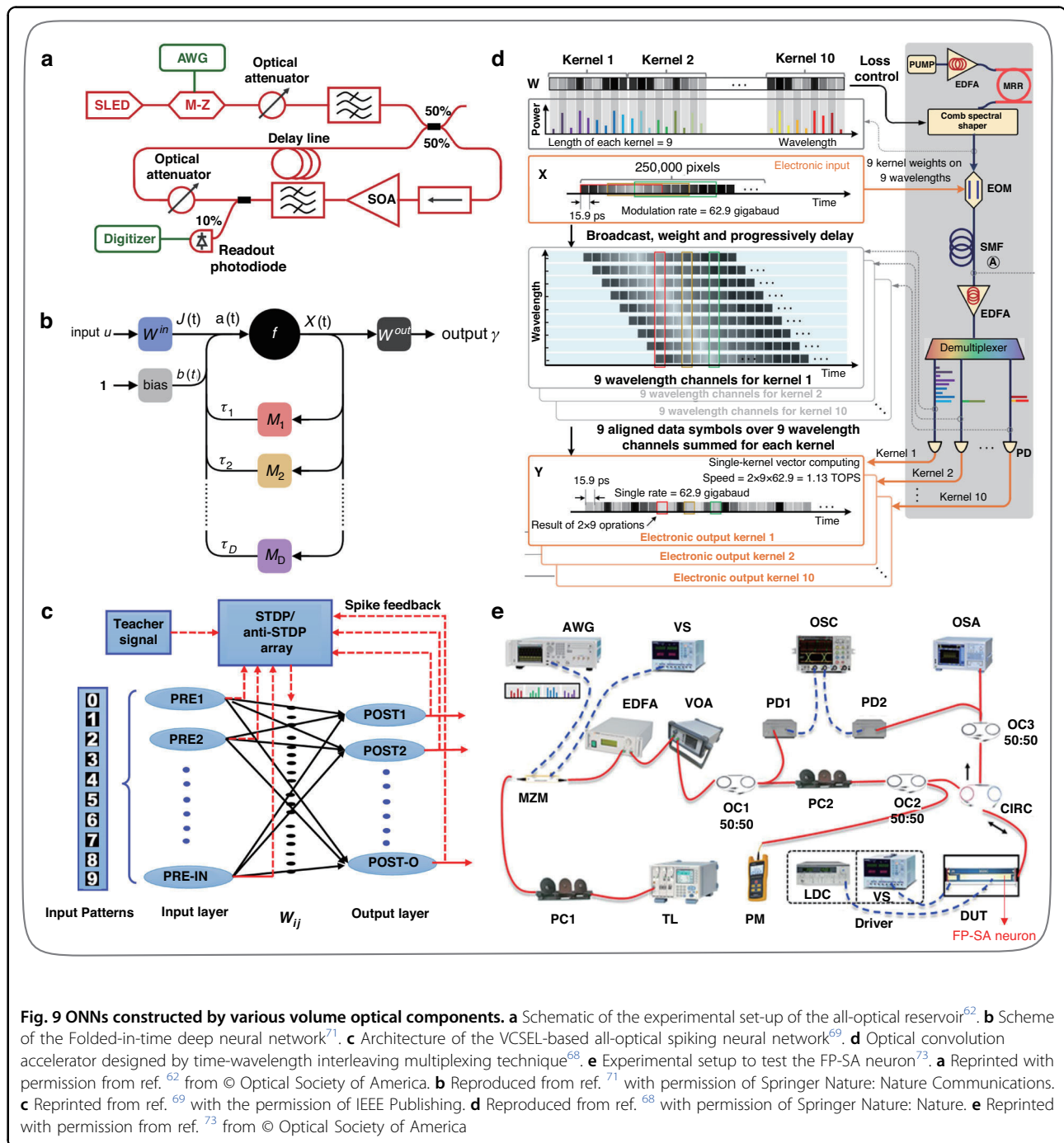


Fig. 9 ONNs constructed by various volume optical components. **a** Schematic of the experimental set-up of the all-optical reservoir⁶². **b** Scheme of the Folded-in-time deep neural network⁷¹. **c** Architecture of the VCSEL-based all-optical spiking neural network⁶⁹. **d** Optical convolution accelerator designed by time-wavelength interleaving multiplexing technique⁶⁸. **e** Experimental setup to test the FP-SA neuron⁷³. **a** Reprinted with permission from ref. ⁶² from © Optical Society of America. **b** Reproduced from ref. ⁷¹ with permission of Springer Nature: Nature Communications. **c** Reprinted from ref. ⁶⁹ with the permission of IEEE Publishing. **d** Reproduced from ref. ⁶⁸ with permission of Springer Nature: Nature. **e** Reprinted with permission from ref. ⁷³ from © Optical Society of America

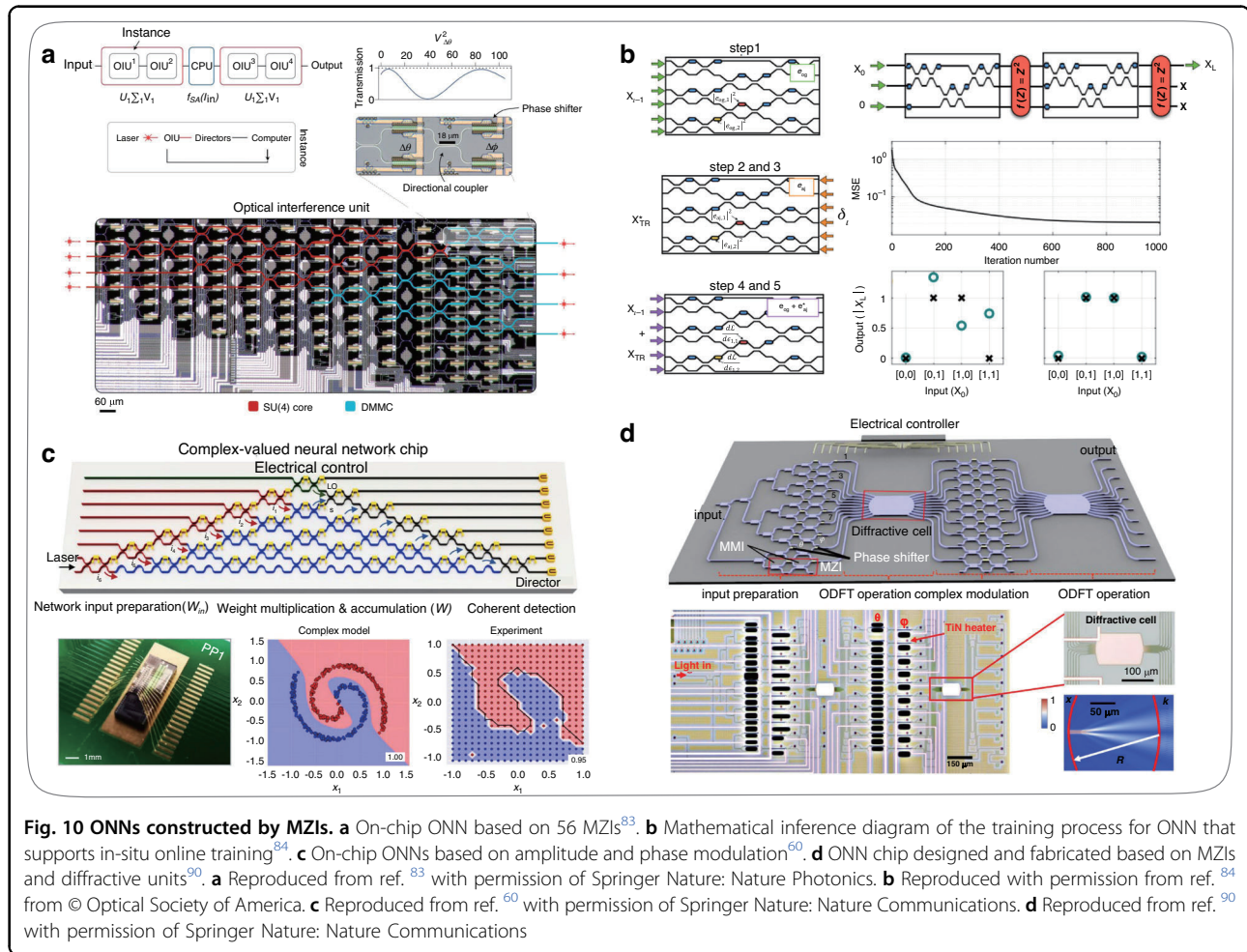
(e.g., optical computers). This section will introduce the research on integrated ONNs based on various on-chip optical components.

ONNs based on on-chip MZI mesh

The propagation of light is a natural process of matrix operation. In 1994, Reck et al.²⁸ conducted experimental verification of unitary matrix operation using traditional bulk optical components (beam splitters and phase

shifters). On this basis, Clements et al.²⁹ proposed an optimized matrix factorization method in 2016, which makes optical components more efficient in achieving matrix factorization. The same year, Ribeiro et al.⁸² implemented a reconfigurable 4×4 -dimensional matrix by cascading on-chip MZIs. From then on, many on-chip ONNs^{60,83–92} based on MZI mesh have emerged.

In 2017, Shen et al.⁸³ designed an ONN with a matrix dimension of 4×4 by cascading 56 MZIs and completed



the fabrication of an ONN chip on a silicon-based substrate, as shown in Fig. 10a. Theoretically, any matrix can be decomposed into one diagonal matrix and two unitary matrices by using the singular value decomposition method. The optical attenuators can implement any diagonal matrix function, and the beam splitters and phase shifters can achieve any unitary matrix function. Thus, the training weight matrices of ONNs can be physically implemented one-to-one via integrated optical elements. In this work, the parameters of on-chip ONN are obtained through pre-training on a computer, and the optical transmission characteristic curve of a saturated absorber is adopted as the nonlinear function during the ONN training process. Afterward, Shen et al.⁸³ completed the 4×4 -dimension matrix operation by reconstructing on-chip MZIs step by step in the ONN inferring stage, which achieved a blind test accuracy of 76.7% in the vowel recognition dataset in experiments. This work leads to a new method for studying on-chip ONNs. In 2018, Hughes et al.⁸⁴ proposed an in-situ training method for on-chip ONNs designed by MZI mesh. Firstly, the optical power of the output port of the on-chip ONNs is accurately

measured. Then, the adjoint variable method is employed to realize the gradient derivation and error back-propagation during the on-chip ONNs in-situ training process (Fig. 10b). This method allows the structural parameters of on-chip ONNs to be directly trained by optical hardware, thus overcoming the error problem caused by the chip fabricating process, which is a crucial advancement in the training method of on-chip ONNs based on MZI mesh, from offline training to in-situ online training. In addition, this work also has a reference for the research of optoelectronic fusion ONNs, such as how to solve the problems in the process of repeated signal conversion between optical chips and electronic hardware.

In 2021, Zhang et al.⁶⁰ pointed out that most research on ONNs still only uses traditional real-value frameworks designed for digital computers. Therefore, the team utilized both the phase and amplitude of light during the training process of ONNs, resulting in higher degrees of freedom for the optimized structural parameters of ONNs during the training process (doubling the adjustable variables). Consequently, the trained



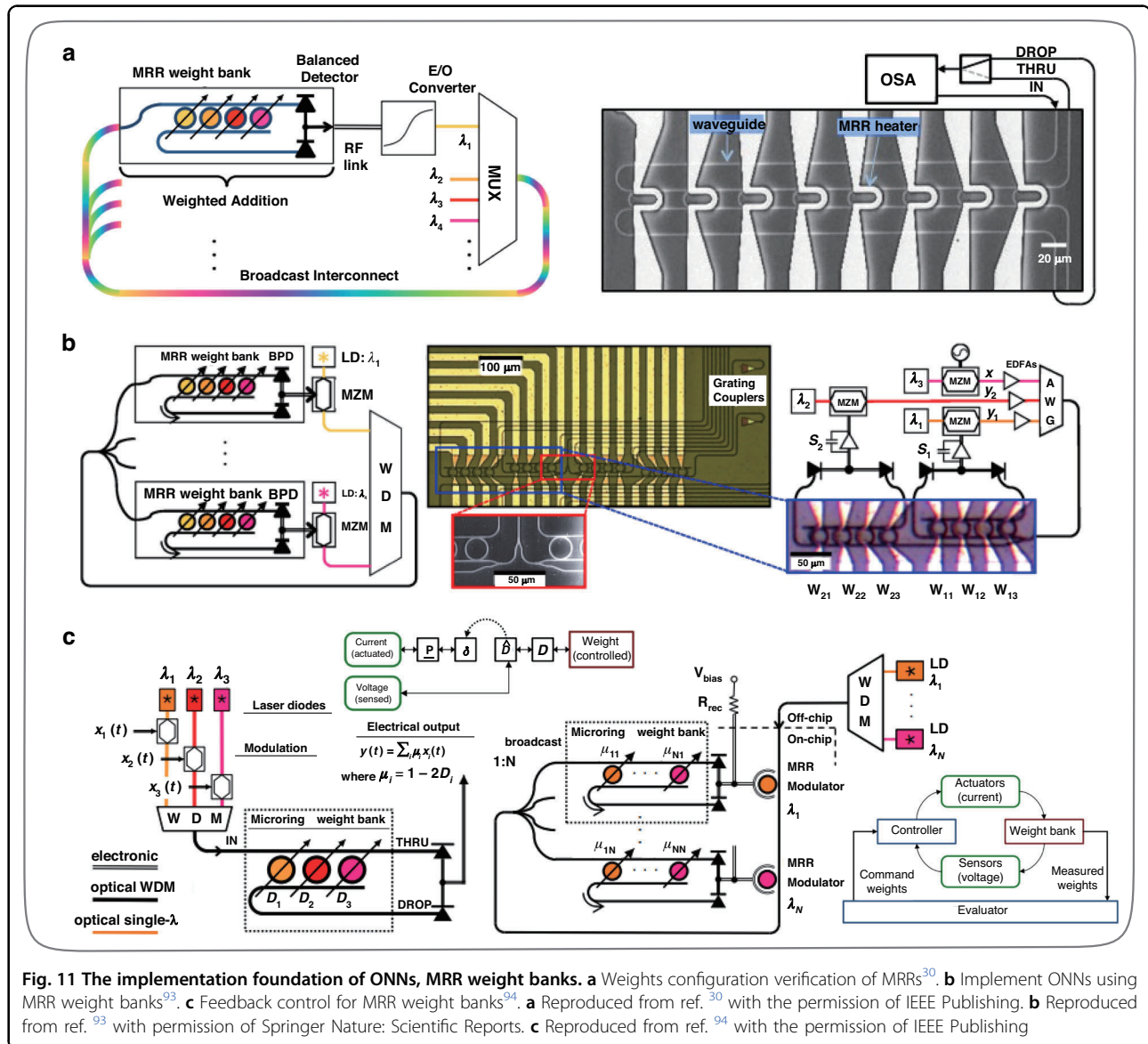


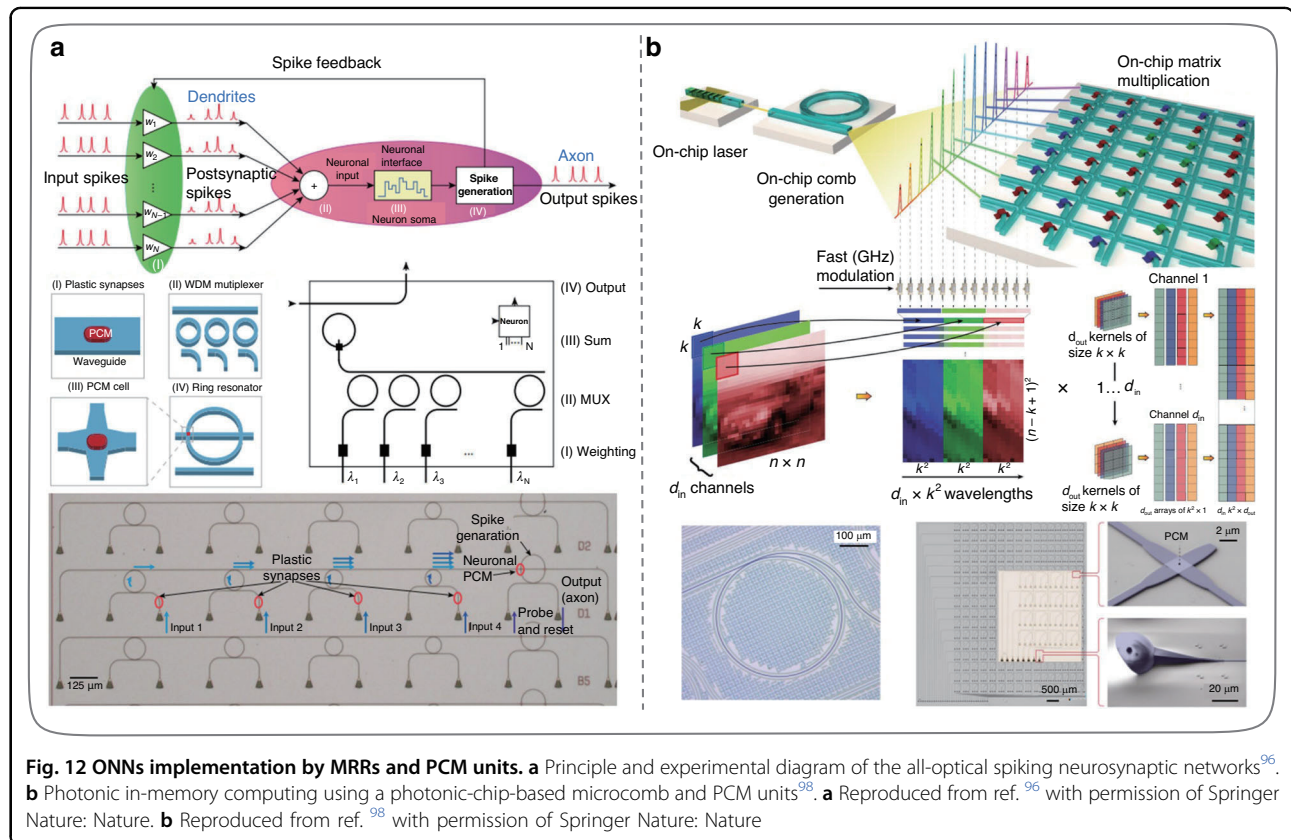
Fig. 11 The implementation foundation of ONNs, MRR weight banks. **a** Weights configuration verification of MRRs³⁰. **b** Implement ONNs using MRR weight banks⁹³. **c** Feedback control for MRR weight banks⁹⁴. **a** Reproduced from ref. ³⁰ with the permission of IEEE Publishing. **b** Reproduced from ref. ⁹³ with permission of Springer Nature: Scientific Reports. **c** Reproduced from ref. ⁹⁴ with the permission of IEEE Publishing

ONNs (Fig. 10c) exhibit superior logical inferring abilities. The ONNs have completed tasks such as logic gate operation, IRIS dataset category prediction, nonlinear data (circle and spiral) classification, and MNIST dataset handwritten digit recognition. The performance of ONNs obtained by this design method is better, with classification accuracy and convergence speed of the loss function compared to other ONNs designed based on MZI mesh at the same matrix size, prominently improving the computational speed and energy efficiency of ONNs. The following year, Zhu et al.⁷⁸, who belonged to the same research group as Zhang et al.⁶⁰, went further in their existing study by introducing integrated diffractive elements that can implement the Fourier transform and inverse transform (Fig. 10d), thereby improving the matrix dimensions resolved by

ONNs during the operation process and reducing their computational energy consumption. Compared with the previous work⁶⁰, this new design scheme outperformed ONNs solely based on MZIs topology cascading regarding integration and energy consumption in classification experiments on the same IRIS and MNIST datasets.

ONN based on on-chip MRR weight banks

MRR has a filtering function and can regulate the optical power of different wavelengths. In 2016, Tait et al.³⁰ conducted a detailed study on the MRR weight banks (Fig. 11a), including the principle of MRR, mutual channel crosstalk, and its design methods. They predict that the MRR weight banks may unlock brand new domains of computing based on silicon photonics. After,



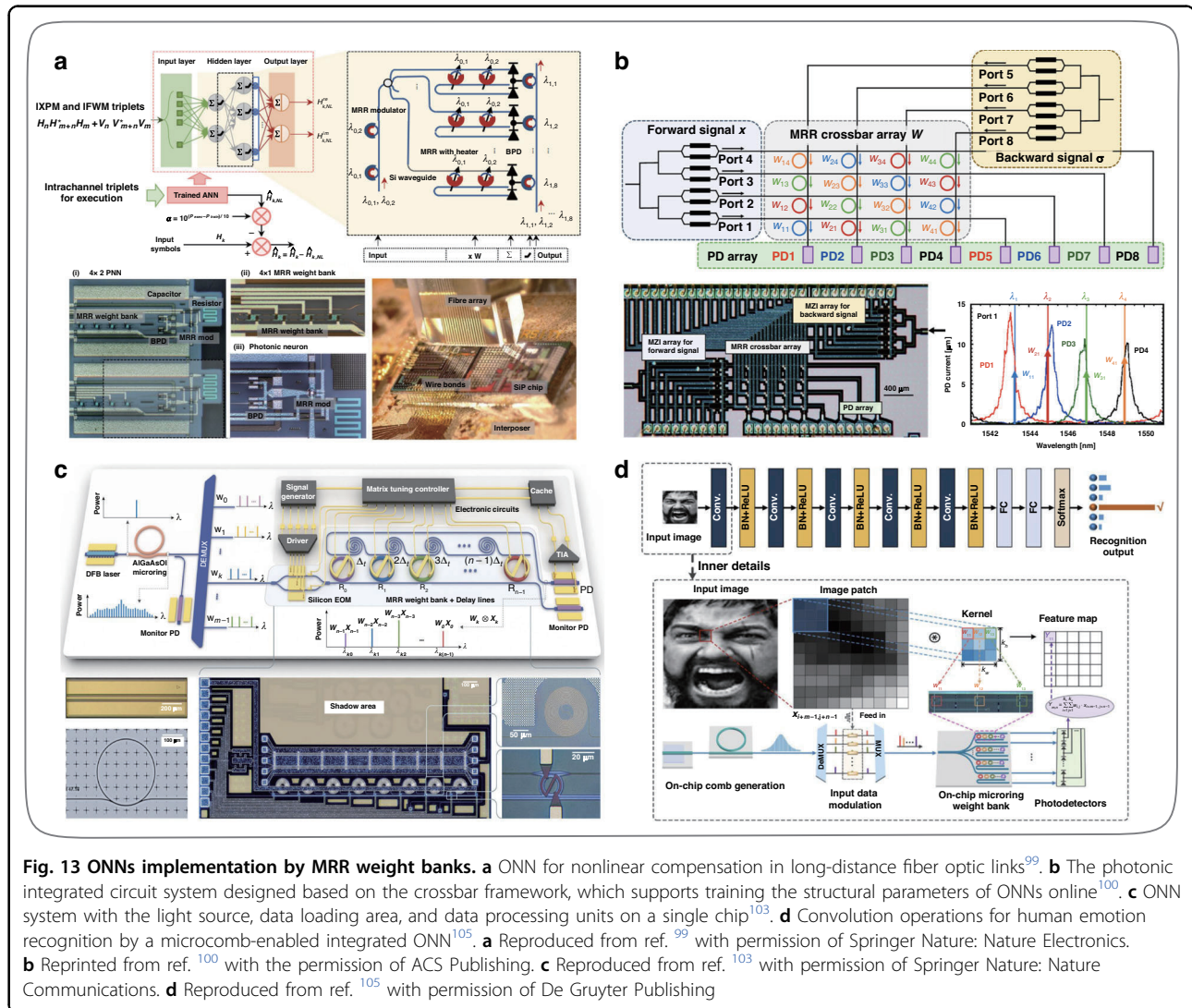
in 2017, Tait et al.⁹³ used the MRR weight banks to configure the connection weights of ONNs (Fig. 11b), proving the mathematical isomorphism between silicon photonic circuits and continuous neural network models. In addition, the team derived and analyzed the power consumption of modulator-class neurons. Further, in 2018, Tait et al.⁹⁴ investigated the fabrication process and thermal sensitivity challenges of MRRs, rendering a feedback weight control method (Fig. 11c) to overcome the weight control problem in reconfigurable photonic networks. Research labors by Tait et al.^{30,93,94} have laid a significant foundation for the subsequent developments of ONNs based on MRR weight banks.

Following the research of predecessors, new research on ONNs^{95–105} based on MRR weight banks has emerged one after another. In 2019, Feldmann et al.⁹⁶ designed a spiking all-optical synaptic system (Fig. 12a) based on MRRs and phase change material (PCM), which avoids the physical separation of memory and processor in conventional computing architectures. Therefore, the inference process of the spiking ONN designed by Feldmann et al.⁹⁶ is more analogous to the brain. The specific working principle of the ONNs implemented by this method is to use PCM units to weight the input pulses, and then couple the corresponding wavelength of light into a single-mode waveguide through MRRs for power

summation. When the accumulated optical power in the single-mode waveguide exceeds a certain threshold, the PCM unit on the last MRR will switch the crystal state and generate an output pulse, completing a calculation on the optical domain. Meanwhile, the optical nonlinearity in the inferring process of the ONNs is physically implemented through the PCM unit on the last MRR. Two years later, Feldmann et al.⁹⁸ further designed and fabricated an integrated photonic processor (Fig. 12b) on the silicon nitride platform, which parallelly implemented the processing function of traditional convolutional kernels in an all-optical manner on the chip. The operating speed of the photonic processor can reach trillions of operations per second. PCM has the characteristic of nonvolatility. The nonlinearity and reconfigurability of the on-chip ONNs can be realized by introducing PCM units on integrated chips. Feldmann et al.^{96,98} provide a valuable reference for the subsequent developments of all-optical ONNs.

In 2021, Huang et al.⁹⁹ designed and fabricated an on-chip ONN based on the MRR weight banks on the silicon-on-insulator (SOI) platform (Fig. 13a). The function of this ONN is to assist electronic hardware systems in completing nonlinear compensation in submarine fiber optic links. As the ONN system can process optical signals in the analog domain, it enormously reduces the high

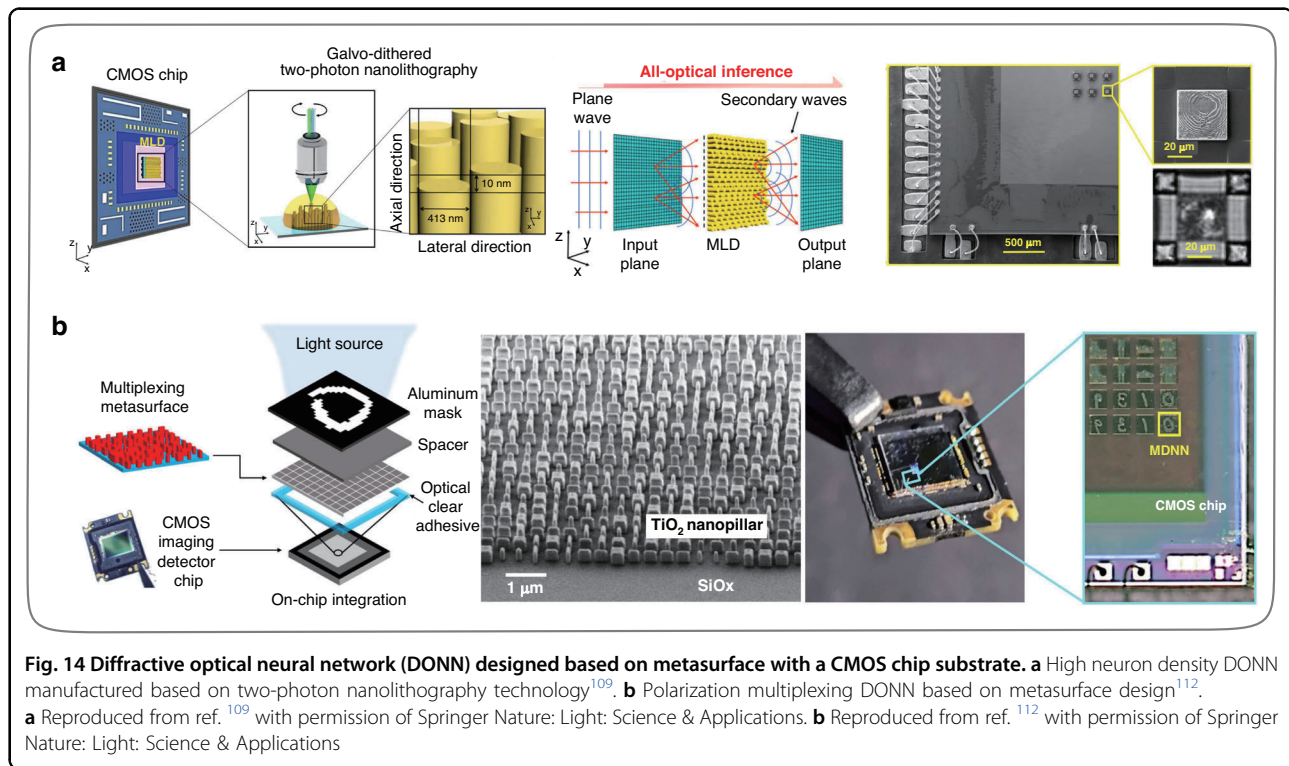




complexity and high-speed requirements of conventional digital signal processing circuits in handling nonlinear digital signal processing circuits in handling nonlinear compensation in submarine long-distance fiber optic communication links. In 2022, Ohno et al.¹⁰⁰ pointed out that most of the Si programmable photonic integrated circuits (PICs) proposed for ONNs suffer the issues of low scalability and incomplete all-optical training frameworks. Therefore, Ohno et al.¹⁰⁰ proposed a crossbar array framework based on the MRR weight banks and designed a PIC based on this framework for parameters online training of on-chip ONNs. The programmable PIC (Fig. 13b) does without complex algorithms such as singular value decomposition during matrix-vector multiplication, supporting a gradient back-propagation algorithm for online training of on-chip ONNs, which is beneficial for integrated system error calibration. In January 2023, Bai et al.¹⁰³ designed an on-chip ONN system (Fig. 13c) using integrated optical frequency combs, MRR

weight banks, and silicon-based spiral waveguide delay lines. The multiple wavelength sources, data loading areas, and data processing centers are fully integrated into a single chip. The convolution function is realized by adopting the time-wavelength stretching method, and the computational density can reach about 1.04 TOPS per square millimeter. In October 2023, Cheng et al.¹⁰⁵ designed a microcomb-enabled integrated ONN based on microcomb and MRR weight banks (Fig. 13d). It can perform tensor convolution operations and complete intelligent tasks for human emotion recognition (6 types of human emotions) at low power consumption and light speed, with a blind testing accuracy of 78.5%. This work is a meaningful attempt by ONNs in practical applications, with a potential throughput of up to 51.2 TOPS, providing a reference for the next generation of computing hardware in intensively computational artificial intelligence applications.





ONN based on on-chip diffractive metasurfaces

The clever design of metasurfaces can theoretically achieve arbitrary control of the wavefront of reflected/refracted beams¹⁰⁶. Based on this, research on integrated DONNs^{107–120} has also been carried out.

To reduce the size of computing units and further improve the integration and system stability of the free-space DONNs. In 2021, Goi et al.¹⁰⁹ selected a near-infrared wavelength (785 nm) and fabricated an integrated DONN with high neuron density by two-photon nanolithography using the complementary metal-oxide semiconductor (COMS) chip as the substrate (Fig. 14a). The research team fabricated a multilayer diffraction metasurface on a CMOS chip consisting of an array of subwavelength cylindrical structures in steps of 10 nm in the height direction of the cylinder. This DONN is highly integrated and can fabricate about 500 million neurons per square centimeter. In 2022, Luo et al.¹¹² selected the visible wavelength (532 nm) to fabricate an integrated DONN on a CMOS chip substrate that can support performing multi-channel sensing and multi-tasks in a visible light environment (Fig. 14b). The metasurface (hidden layer) in the integrated DONN consists of a subwavelength nanopillar array with a fixed height of 600 nm and a central spacing between adjacent nanopillars (the nanopillar array period) of 400 nm, which modulates the phase of propagated light by varying the length and width of each nanopillar. In this

work, a multi-channel classifier framework was constructed by implementing a polarization multiplexing scheme using subwavelength nanostructures, which further demonstrates that the comprehensive design of optical feature quantities can endow ONNs architectures with the ability to handle multiple tasks (the various classification tasks of MNIST and Fashion-MNIST were demonstrated). In addition, the number of neurons on this CMOS chip that could be integrated per square millimeter was about 6.25×10^{16} .

Additionally, the on-chip one-dimensional metasurfaces (metalines) composed of subwavelength diffraction units can also achieve the wavefront shaping of light in slab waveguides^{121,122}. The on-chip DONNs constructed from diffraction metalines on slab waveguides have better stability, portability, and scalability. In 2020, Zarei et al.¹⁰⁷ designed an on-chip DONN composed of diffraction metalines by using subwavelength rectangular slots on the SOI platform and validated the performance of the DONNs (Fig. 15a) through simulation calculations. The computing speed of such an on-chip DONN is about 1.2×10^{16} multiply-accumulate operations per second. In 2021, Fu et al.¹⁰⁸ developed an on-chip integrated two-dimensional spatial electromagnetic propagation model and a weight mapping model, which can complete parameter training via computers for on-chip DONN and ensure the pre-trained parameters accurately map onto the physical



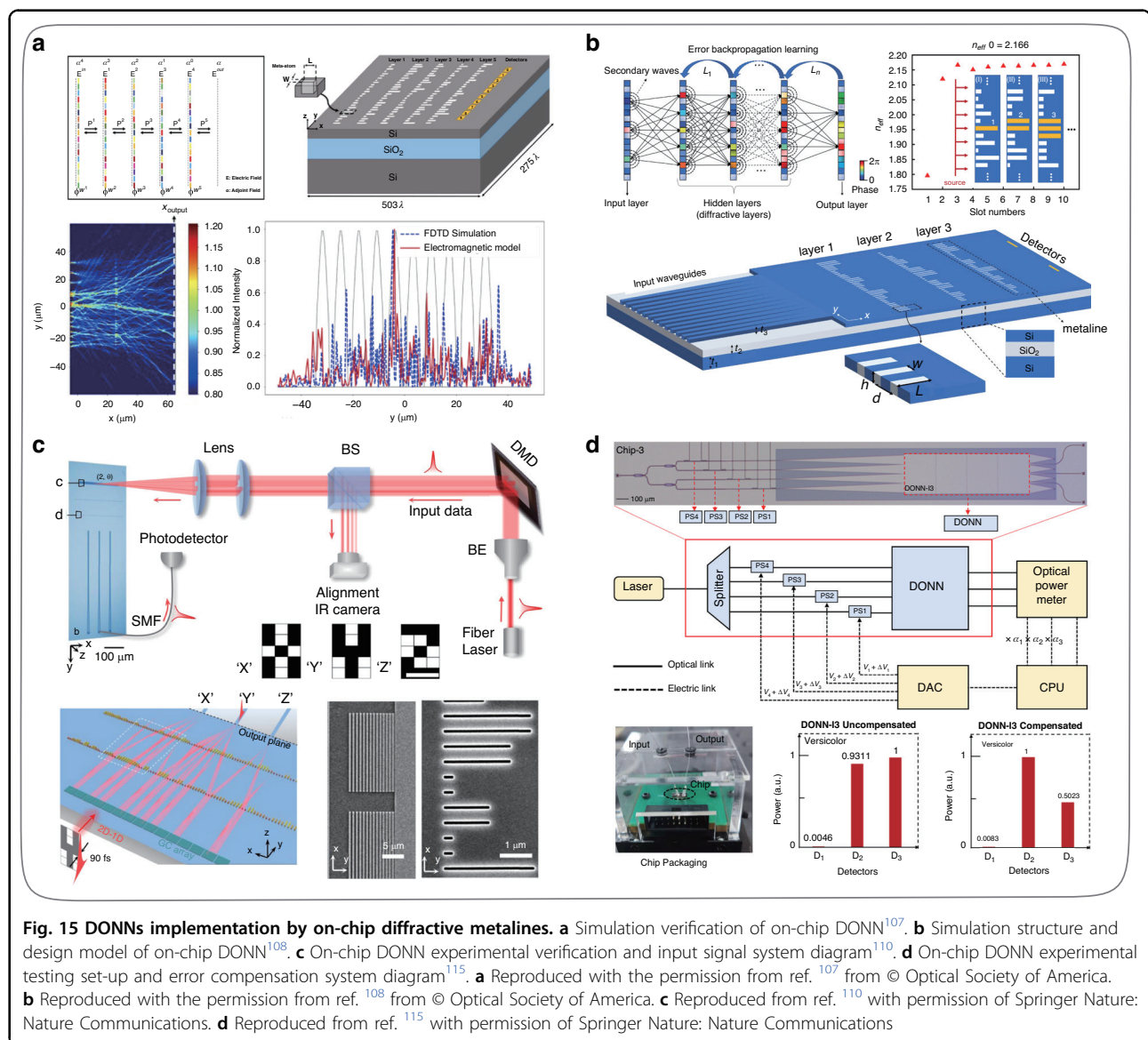


Fig. 15 DONNs implementation by on-chip diffractive metalines. **a** Simulation verification of on-chip DONN¹⁰⁷. **b** Simulation structure and design model of on-chip DONN¹⁰⁸. **c** On-chip DONN experimental verification and input signal system diagram¹¹⁰. **d** On-chip DONN experimental testing set-up and error compensation system diagram¹¹⁵. **a** Reproduced with the permission from ref.¹⁰⁷ from © Optical Society of America. **b** Reproduced with the permission from ref.¹⁰⁸ from © Optical Society of America. **c** Reproduced from ref.¹¹⁰ with permission of Springer Nature: Nature Communications. **d** Reproduced from ref.¹¹⁵ with permission of Springer Nature: Nature Communications

devices (Fig. 15b). Two years later, Fu et al.¹¹⁵ fabricated on-chip DONNs (Fig. 15d) based on the previous theoretical exploration and considered the practical issues that were difficult to care for during the simulation part. Meanwhile, to effectively reduce the system errors caused by chip fabricating and packaging processes, an in-situ training scheme based on the particle swarm algorithm for error compensation was proposed as well. The effectiveness of this scheme remarkably ensured the performance of the DONN chips and improved the robustness of the experimental testing system. In 2022, Wang et al.¹¹⁰ further advanced their previous work¹²² by designing and fabricating an on-chip DONN (Fig. 15c) based on subwavelength diffractive units. In this work, to reduce the mutual interference between adjacent subwavelength diffractive units, Wang et al.¹¹⁰

combined two identical subwavelength diffractive units as a new computing cell in the hidden layer of the on-chip DONNs. In addition, the input signals are loaded onto the DONN chip through a DMD and lens system, which indicates that the input dimension of the on-chip DONN can be unrestricted by the number of waveguides, providing a feasible solution for the problem of limited input dimension of on-chip ONNs.

ONN based on other on-chip optical components

In addition to the preceding introduced integrated optical devices^{83,90,96,103,109,115}, there are a variety of on-chip optical components that can construct ONNs^{123–140}, such as single-mode waveguides, multi-mode interferometers, phase shifters, attenuators, detectors, three-dimensional integrated waveguides, etc.



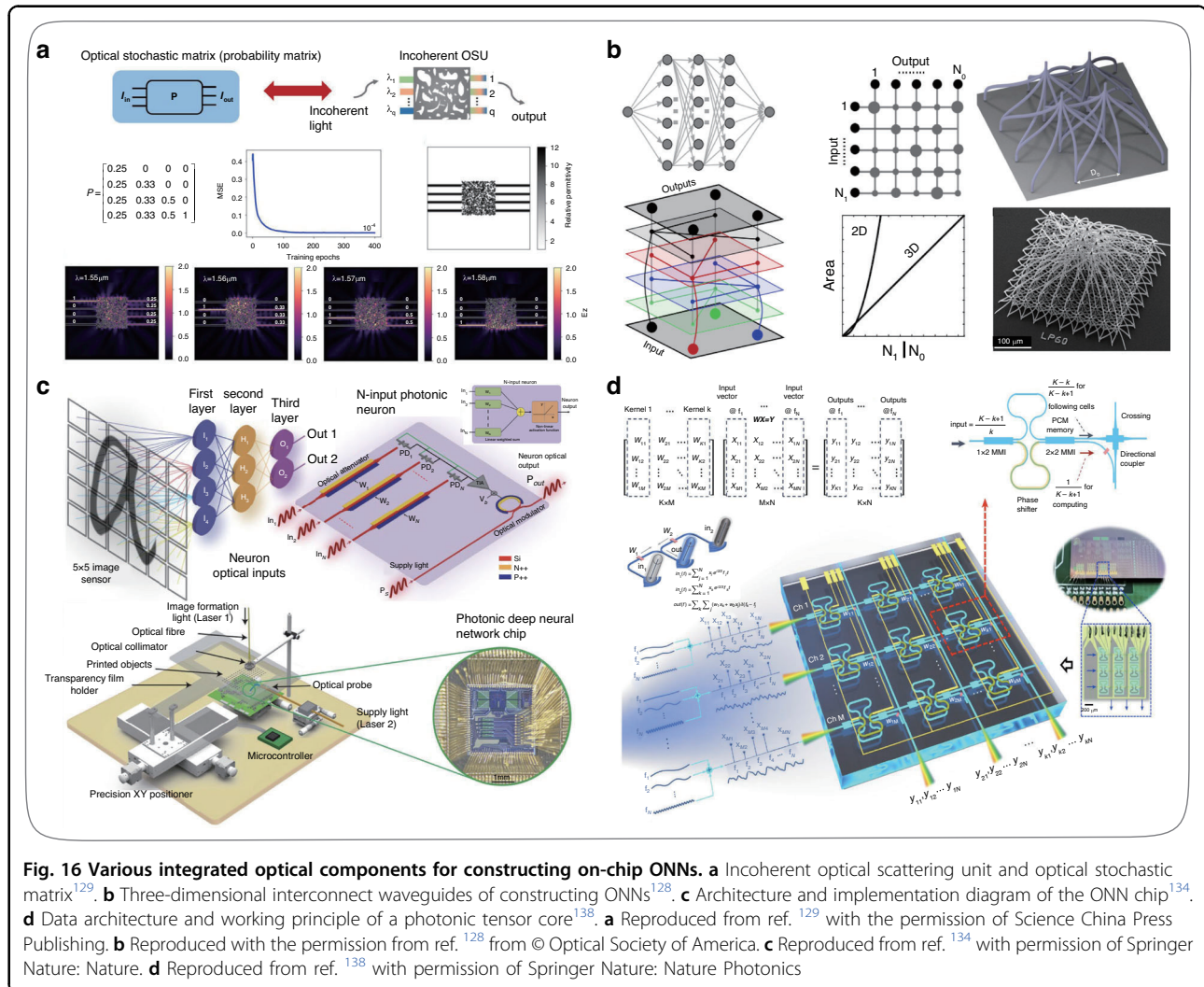


Fig. 16 Various integrated optical components for constructing on-chip ONNs. **a** Incoherent optical scattering unit and optical stochastic matrix¹²⁹. **b** Three-dimensional interconnect waveguides of constructing ONNs¹²⁸. **c** Architecture and implementation diagram of the ONN chip¹³⁴. **d** Data architecture and working principle of a photonic tensor core¹³⁸. **a** Reproduced from ref. ¹²⁹ with the permission of Science China Press Publishing. **b** Reproduced with the permission from ref. ¹²⁸ from © Optical Society of America. **c** Reproduced from ref. ¹³⁴ with permission of Springer Nature: Nature. **d** Reproduced from ref. ¹³⁸ with permission of Springer Nature: Nature Photonics

In April 2020, Qu et al.¹²⁹ proposed an optical stochastic architecture based on optical scattering units (Fig. 16a). They adopt the inverse design method to optimize the parameters of the optical scattering units, thereby obtaining on-chip ONNs for deep learning tasks with fast speed, low power consumption, and small footprint. The team designed an on-chip ONN for the MNIST dataset and achieved a prediction accuracy of 97.1% on the blind test set. Theoretically, the inverse design approach can implement any function of the optical scattering units. The advantage of inverse design is that it can achieve target optimization results without knowing the analytical process of parameter acquisition. However, this method has a fatal limitation, such as the optimization process of parameters is time-consuming and demands massive computing power when solving complex tasks. In June 2020, Moughames et al.¹²⁸ utilized two-photon polymer printing technology to fabricate on-chip three-dimensional integrated low-loss

photonic waveguide arrays. The waveguides interconnect structure corresponds to large-scale vector-matrix products (Fig. 16b) and is the core of neural network computing. In this work, the diameter of the photonic waveguide is 1.2 μm and the spacing between adjacent waveguides is 20 μm . Consequently, there is about 2200 neurons can be integrated per cubic millimeter through the two-photon polymer printing technology. This scheme provides a novel direction for the development of on-chip ONNs.

In 2022, Ashtiani et al.¹³⁴ proposed and fabricated a silicon-based integrated system using on-chip photonic devices such as attenuators and detectors, which can perform the inference function of ONNs in steps (Fig. 16c). Specifically, the integrated system can only perform one calculation process for a single neuron at one time, which demands modulating the coefficients of the attenuators multiple times to complete weight allocation and matrix operation. Although the computing capacity



of the on-chip ONNs is low, it has advantages such as simple structure, reconfigurability, and nonlinearity. In 2023, Dong et al.¹³⁸ proposed a method for three-dimensional data processing, which introduced radio-frequency modulation of photonic signals to increase parallelism, thereby adding the input dimensions of the on-chip ONNs based on spatially distributed nonvolatile memory and wavelength multiplexing. The optical system constructed by the photonic tensor core (Fig. 16d) attaches parallelism of 100 degrees, achieving two orders of magnitude higher than using only spatial and wavelength interleaving methods. This work provides a significant inspiration for solving the input dimension limitation problem of on-chip ONNs. Subsequent research regarding on-chip ONNs can consider simultaneously mapping data features onto multiple feature quantities of light, thereby increasing the dimensionality of input data in the limited physical input channels.

Discussion

ONNs have been developing for decades since the 1960s. This review summarizes seven different optical devices designed for ONNs under the two major modules of non-integrated ONNs and integrated ONNs. In this chapter, we score these ONNs regarding integration level, computing capacity, stability/portability, universality, reconfigurability, nonlinearity, and scalability. Then, we analyze and summarize the corresponding performances of the various ONNs and envisage their future development trends and challenges.

Non-integrated ONNs are mainly designed based on $4f$ systems^{37,39,40}, diffractive elements^{42,44,45,53}, and other bulk optical components^{62,68,71,73}. All types of the ONNs can achieve reconfigurable functions in terms of technology. Meanwhile, ONNs based on diffractive elements can achieve large computing capacity^{42,47}, but it is composed of discrete components, which may cause unavoidable errors in alignment calibration between discrete components. The error will accumulate with the increase of ONN layers, and the accumulated errors will cause adverse impacts on the performance of ONNs. In addition, the ONNs consist of other bulk optical components can already implement nonlinear functions in the optical domain^{39,141}, but those ONNs are tough to realize large-scale implementation, which limits their scalability and makes them difficult to achieve universality. In summary, non-integrated ONNs are more suitable for specialized applications, such as specific holographic imaging^{142,143}, pre-sensing optical calculation^{52,109,112,144}, etc.

For integrated ONNs, thanks to the CMOS process technology, the problem of alignment errors between discrete components have been efficiently solved, and the large-scale and low-cost processing conditions make the scalability and versatility of the integrated ONNs

realistic, these advantages effectively make up for the shortcomings of non-integrated ONNs. Among them, the integration level of the ONNs designed based on MZI mesh^{60,83,90}, MRR weight banks^{98,100,103}, and other components^{129,134,138} is relatively high. However, since these ONNs require constant energy supplies during the working process, which leads to the limitation of large-scale expansion of the computing units by taking into account the adjacent modulator's thermal crosstalk, thus the computing capacity of these ONNs always keeps a low level. By contrast, the integrated ONNs based on diffractive metasurfaces^{110,115} have a large computing capacity because of the sub-wavelength computing units. However, due to the small size of the computing units (e.g., the size of the diffractive computing unit proposed by Fu et al.¹¹⁵ is about $0.5\mu\text{m}\times 2\mu\text{m}$), it is tough to be precisely modulated. Currently, although integrated ONNs have corresponding implementation technologies in terms of reconfigurability^{83,90,103,134,138}, nonlinearity^{96,145}, and high computing capacity^{110,115}, it is still very difficult to simultaneously achieve these advantages on the same ONN, and further exploration is still needed.

Computational density and computing capacity of ONNs

The computational density in this paper refers to the number of operations that can be completed per square millimeter/centimeter per second. The computing capacity means the maximum matrix dimension that ONNs can handle at one time. Thus, the computational density and computing capacity here are different, in other words, when the computing capacity is large, the computational density may not be high (e.g., the ONN proposed by Lin et al.⁴²). In fact, the computational density is significantly related to the integration level of the neurons (computing units) of the ONNs, and the higher the integration level of the computing units of ONNs, the higher the computational density of the ONNs will be^{109,110,112,115}. Therefore, the computational density of the integrated ONNs is often higher than that of the non-integrated ONNs. However, in terms of computing capacity, even if the integration level of integrated ONNs is higher than that of non-integrated ONNs, the integrated ONNs may not necessarily have greater computing capacity than non-integrated ONNs.

In addition, regarding the computing capacity of ONNs constructed based on the $4f$ system³⁷, we consider that it is related to the integration level of the mask (computing/modulation units are always fabricated on the mask) placed on the Fourier plane. The higher the integration level of the mask on the Fourier plane, the higher the computing capacity of the ONNs.

Finally, in the design of ONNs, it is unnecessary to pursue computational density perversely. It is better to



Table 1 Performance comparison of the non-integrated/integrated ONNs (I)

Types	Performance Basic components	Integration level	Computing capacity	Stability/ Portability	Universality
Non-integrated ONNs	Volume 4f system	--- ●●●●	★★★ ●●●●	★ ●●●	--- ●●●●
	Volume diffractive elements	--- ●●●●	★★★★★ ●	★ ●●●●	★★ ●●●
	Other volume elements	--- ●●●●●	★★ ●●●	★ ●●●	★★ ●●●
Integrated ONNs	On-chip MZI mesh	★★★ ●●●	★ ●●●●	★★★★★ ●	★★ ●●
	On-chip MRR weight banks	★★★★ ●●●	★★ ●●●●	★★★★★ ●	★★ ●●
	On-chip diffractive metasurfaces	★★★★★ ●	★★★★★ ●	★★★★★ ●	--- ●●●●●
	Other on-chip optical elements	★★ ●●●●	★ ●●●●●	★★★★★ ●	★★ ●●

design the ONNs by determining the actual demands, some ONNs with appropriate computing capacity but not high computational density, might also serve the oriented tasks well.

Nonlinearity and reconfigurability of ONNs

Nowadays, there are cases where both non-integrated ONNs^{39,62,71} and integrated ONNs^{96,98,116,146} can achieve nonlinearity. However, most ONNs that can achieve nonlinearity are generally relatively small in scale. Larger scale ONNs, such as those with larger depths, must consider the optical power attenuation introduced by nonlinear layers.

Regarding reconfigurability, non-integrated ONNs are usually built through optical devices such as SLMs and DMDs^{39,40,47}. However, it is tough to achieve optical nonlinearity on these devices. In the future, metasurfaces may be able to simultaneously address issues such as reconfigurability, nonlinearity, and the insertion loss of ONNs. For example, the ONNs designed based on information metasurfaces by Liu et al.⁵³ can conveniently achieve the programmable function of each diffractive unit (neuron), and it is promising to further achieve nonlinearity and power amplification function by introducing stable nonlinear amplifiers into artificial neurons⁵³.

Integrated ONNs are used to be constructed through optical components such as on-chip MZIs, MRRs, thermal/electro-optical phase shifters, and attenuators^{83,96,133,134}. The best performance for integrated

ONNs is to achieve both nonlinearity and reconfigurability simultaneously. In fact, the adoption of non-volatile PCM units^{147–150} can endow ONNs with both reconfigurable and nonlinear functions during the inferring process^{96,98}. However, the introduction of new materials for PCM still faces the challenge of insertion loss. Thus, it is imperative to develop new devices that can achieve both reconfigurable and nonlinear functions at low power consumption. Fortunately, Zhong et al.¹⁴⁵ designed a low power and reconfigurable phase-relevant on-chip activation function device based on Graphene/Silicon heterojunction, which offers a new insight into the on-chip ONNs. The activation function device proposed by Zhong et al.¹⁴⁵ may be applied to various on-chip ONNs with different architectures in the future.

Scalability of ONNs

The scalability of non-integrated ONNs is not as good as that of integrated ONNs because of their large size and composition of discrete devices. The scalability of integrated ONNs can be discussed from the following aspects.

First, limitations on the parallel input dimension of signals. In fact, except for integrated ONNs^{109,112} designed based on CMOS chip substrate that are not limited by input dimensions, other integrated ONNs are extremely low (e.g., the input waveguide of less than 10) in the input dimensions due to the limitation of the number of inputting on-chip waveguides^{83,90,115,119}.

Second, the scalability of computing units in a single integrated ONN. The on-chip ONNs constructed based on MZI mesh or MRR weight banks require additional energy supply during the operation of the computing units, and many computing units face difficulties in synchronous modulation under high-speed situations. These challenges greatly hinder the large-scale expansion of such on-chip ONNs. Distinctively, due to the computing unit of on-chip DONN consisting of subwavelength diffractive structures^{110,115}, its scalability is superior to other on-chip ONNs. The large-scale expansion of on-chip DONN computing units will not cause a significant increase in computing energy consumption, which is advantageous to expand in scale. However, the computing units of on-chip DONN are difficult to achieve reconfigurable and nonlinear functions.

Third, the scalability of cascading on-chip ONNs. At present, all on-chip ONNs have not achieved good scalability in the cascading method. If we hope to further improve the scalability between different on-chip ONNs, electronic circuits may be essential to assist in implementation. Besides, ensuring the supplementation or regeneration⁵³ of optical power during the light propagation in cascaded ONNs is crucial. Optoelectronic hybrid ONNs⁴⁷ may be an effective way to achieve cascade scalability between various on-chip ONNs. However, the energy consumption caused by the photoelectric conversion interface and the impact of conversion speed on the overall efficiency of the optoelectronic hybrid ONN system will become undeniable. Notably, scalabilities between the on-chip ONNs require the support of nonlinear functionality, otherwise, the performance improvement of the scalable ONNs will be compromised. Thus, there are still existing many challenges in the scalable process of ONNs.

Energy efficiency of ONNs

The calculation process of ONNs is completed during the propagation of light, thus their energy efficiency during operation can be designed to be excellent. Here we provide quantitative comparisons of the energy efficiency between different ONNs and compare it with the energy efficiency of the advanced computing hardware, as shown in Table 3. It is not difficult to find that the energy efficiency of some ONNs^{83,134} is better than that of existing advanced computing hardware^{151,152}, and the advantage in energy efficiency will be extremely prominent¹⁵³ after optimizing the on-chip ONN's architecture. For more performance summaries of ONNs, including energy efficiency and other performance aspects please refer to more relevant reference works^{32,47,110,115,153–156}.

Applications of ONNs

Currently, the application of ONNs is not as widespread as its electronic counterparts, most research works still

focus on handling simple datasets, and the practical application scenarios of ONNs are rare. For all that, scientific researchers are striving to combine the research study of ONNs with real-world applications. For example, Huang et al.⁹⁹ applied on-chip integrated ONNs to nonlinear compensation in submarine fiber optic communication links. Sludds et al.¹⁵⁷ developed the Netcast, an edge-computing architecture based on photonic deep learning, to complete the inferring process on edge devices. The dedicated ONNs in real-world scenarios have indeed achieved positive results, bringing benefits to practical application systems regarding computing speed, energy consumption, etc.

In addition, the industry has also invested in research in the field of ONNs or optical computing. For example, Lightmatter has successively released a series of products such as Enviser and Passage¹⁵⁸. The company has comprehensively considered software and hardware collaboration and energy consumption to better leverage the inherent advantages of light in the optical computing process. Lightelligence has released the photonic arithmetic computing engine (PACE), which integrates more than 10,000 discrete photonic devices in PACE's photonic chip and has a system clock of 1 GHz¹⁵⁹.

Admittedly, ONNs are difficult to independently complete inference tasks without the assistance of electronic computing hardware, and the application of using ONNs to thoroughly replace the electronic counterpart is still far away. The universality performance of all ONNs as shown in Table 2 is relatively low, which indicates that the application fields of ONNs are not broad. Here, we introduce an optoelectronic hybrid framework in the application of ONNs. In the framework, ONNs are the optical accelerators as part of the optoelectronic hybrid system, as shown in Fig. 17.

Among the optoelectronic hybrid ONNs system framework, the main task of ONNs is to efficiently process a host number of linear matrix operations at the speed of light and undertake the main computational work in optoelectronic hybrid systems. Meanwhile, the responsibility of electronic auxiliary hardware is to achieve parameter reconstruction of ONNs and handle nonlinear operations, data storage, and flow control that are difficult to implement by ONNs. By combining the advantages of electronic hardware and ONNs, the performance of an optoelectronic hybrid system will be superior to traditional electronic methods in terms of energy consumption, computing capacity, computing speed, and so forth^{83,115,134,153}. It is worth noting that when processing massive amounts of data, there will be a host number of routing operations and optoelectronic (electro-optic) conversion operations in the calculation process of optoelectronic hybrid systems. Therefore, new optoelectronic communication protocols and optoelectronic



Table 2 Performance comparison of the non-integrated/integrated ONNs (II)

Types	Performance	Reconfigurability	Nonlinearity	Scalability
	Basic components			
Non-integrated ONNs	Volume 4 <i>f</i> system	★★ ●	★ ●●●●	★ ●●●●●
	Volume diffractive elements	★★★ ●	★ ●●●●	★★ ●●●●
	Other volume elements	★★★★ ●	★★ ●●	★ ●●●●
Integrated ONNs	On-chip MZI mesh	★★★★ ●	★★ ●●●●	★★ ●●●●
	On-chip MRR weight banks	★★★★ ●	★★ ●●●●	★★ ●●●●
	On-chip diffractive metasurfaces	★ ●●●●	★ ●●●●●	★★★ ●●
	Other on-chip optical elements	★★★★ ●	★★★ ●●	★★ ●

In Table 1 and Table 2, the number of “★” represents the degree to which the corresponding performance of ONNs has been achieved presently. The more “★”, the better the performance (the best performance is five “★”), and vice versa. The number of “●” represents the degree of difficulty that ONNs may achieve the corresponding excellent performance in the future. The more “●”, the more difficult it is to achieve the corresponding performance, and vice versa. Meanwhile, five “●” indicates that the corresponding performance is the most difficult to implement. In Table 1, the integration level refers to the number of computing units of ONNs that can integrate per unit area. Computing capacity refers to the maximum matrix dimension that ONNs built on different optical components can handle at one time

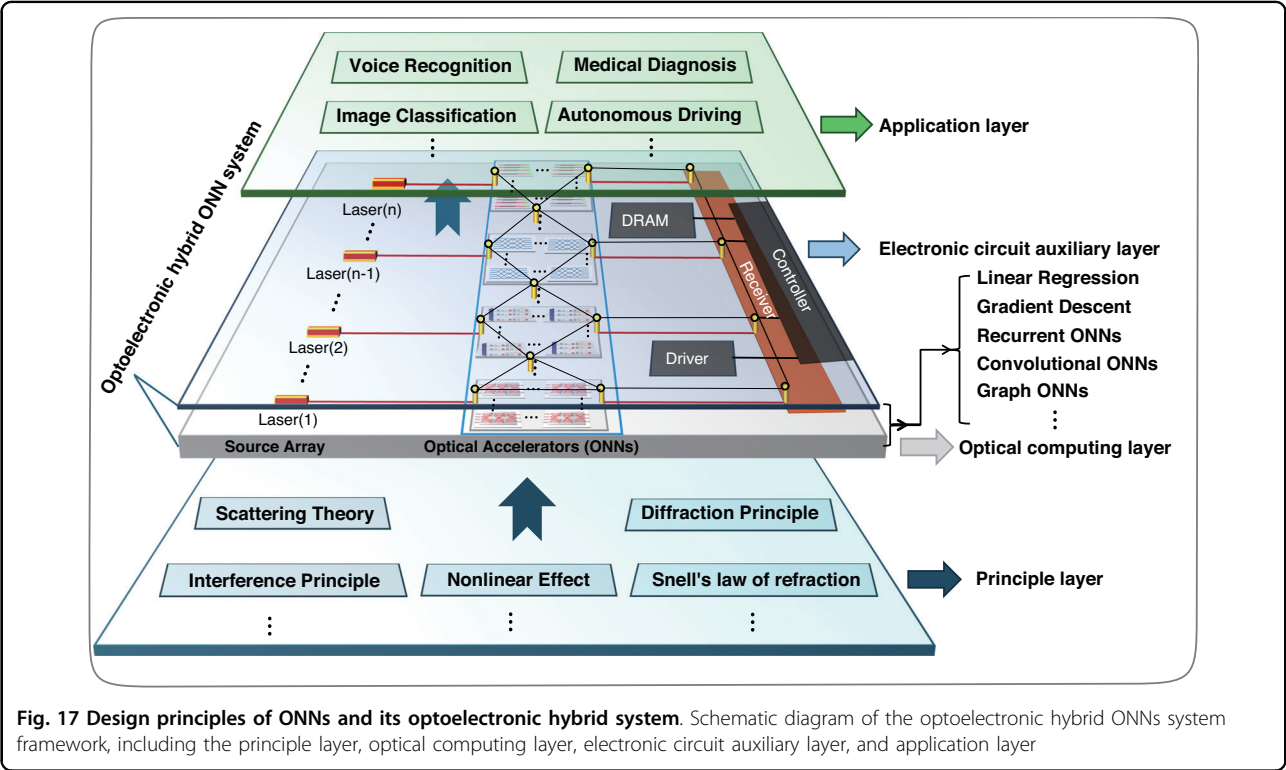
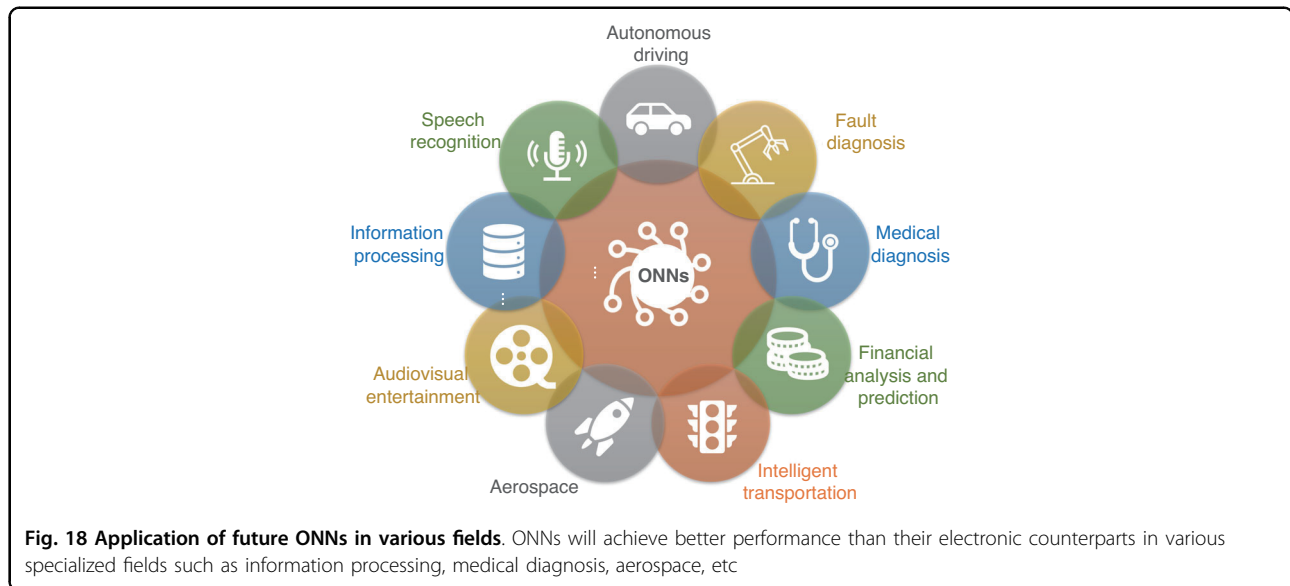


Fig. 17 Design principles of ONNs and its optoelectronic hybrid system. Schematic diagram of the optoelectronic hybrid ONNs system framework, including the principle layer, optical computing layer, electronic circuit auxiliary layer, and application layer

Table 3 Energy efficiency comparison of partial different optical computing architecture and electronic hardware

Works	Technology	Theoretical energy efficiency (TOPS/W)
Ref. ⁴⁷	MPLC with a reconfigurable DPU	0.71
Ref. ⁸³	Coherent MZI mesh●	6.67
Ref. ⁹⁸	Photonic in-memory computing based on WDM system and PCM units	0.50
Ref. ¹³⁴	Based on integrated attenuators	2.90
Ref. ¹⁵³	Based on integrated subwavelength diffractive units and MZI mesh	160.82
Ref. ¹⁵¹	HUAWEI Ascend 910	1.65
Ref. ¹⁵²	NVIDIA Tesla T4	1.85
Ref. ¹⁶⁰	Google TPU	4.65

In Table 3, the abbreviations are *MPLC* multiple plane light conversion, *DPU* diffractive processing unit, *TPU* tensor processing unit, and *PCM* phase change material. ●Indicates that the modulation and detection rates are set at 10 GHz during calculation



(electro-optic) conversion efficiency as well need to be further optimized.

In the future, it may take a considerable period to continuously optimize the architecture or hybrid framework of the ONNs system to obtain better performances, so that they can achieve outstanding results in certain dedicated fields (Fig. 18) compared to their electronic counterparts. During this period, there may be challenges in comprehensively considering the construction of the ONNs application ecosystem, including software, hardware, protocols, optical algorithms, industry standards, manufacturing technology, and other aspects.

Conclusion

In this review, we first introduce the development history of ONNs, the mathematical model of optical artificial

neurons, and the distinctions between various optical components to achieve optical matrix operations. Secondly, we systematically retrospect the development of ONNs based on seven diverse optical components of non-integrated ONNs and integrated ONNs and introduce the typical research works in detail. Then, we score and analyze the performances of different types of ONNs in the discussion part, including computational density, computing capacity, reconfigurability, nonlinearity, and scalability. Finally, we discuss the challenges that the various ONNs may encounter and envisage their applications and development trends in the future.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (NSFC) (62135009, 12274462).



Author details

¹College of Advanced Interdisciplinary Studies, National University of Defense Technology, Changsha, China. ²Hunan Provincial Key Laboratory of Novel Nano-Optoelectronic Information Materials and Devices, National University of Defense Technology, Changsha, China. ³Nanhu Laser Laboratory, National University of Defense Technology, Changsha, China. ⁴Department of Electronic Engineering, Tsinghua University, Beijing, China. ⁵Beijing National Research Center for Information Science and Technology (BNRist), Beijing, China

Author contributions

T.Z.F. and H.W.C. proposed the framework of this review. T.Z.F. prepared the manuscript. T.Z.F., J.F.Z., R.S., Y.Y.H., W.X., S.G.Y., Z.H.Z. and H.W.C. involved in the discussion. All authors have approved submission.

Conflict of interest

The authors declare no competing interests.

Received: 10 January 2024 Revised: 29 July 2024 Accepted: 18 August 2024
Published online: 20 September 2024

References

- McCulloch, W. & Pitts, W. A logical calculus of the ideas immanent in nervous activity. *Bull. Math. Biophys.* **5**, 115–133 (1943).
- McCulloch, W. & Pitts, W. The statistical organization of nervous activity. *Biometrics* **4**, 91–99 (1948).
- Hebb, D. O. *The Organization of Behavior* (Wiley, 1949).
- McCarthy, J. et al. A proposal for the Dartmouth summer research project on artificial intelligence: August 31, 1955. *AI Mag.* **27**, 12–14 (2006).
- Rosenblatt, F. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychol. Rev.* **65**, 386–408 (1958).
- Rumelhart, D. E. et al. Learning representations by back-propagating errors. *Nature* **323**, 533–536 (1986).
- Le Cun, Y. et al. Handwritten digit recognition with a back-propagation network. In *Proc. 2nd International Conference on Neural Information Processing Systems* 396–404 (MIT Press, 1989).
- Krizhevsky, A. et al. ImageNet classification with deep convolutional neural networks. In *Proc. 25th International Conference on Neural Information Processing Systems* 1097–1105 (Curran Associates Inc, 2012).
- Liu, P. R. et al. Application of artificial intelligence in medicine: an overview. *Curr. Med. Sci.* **41**, 1105–1115 (2021).
- Zhao, S., Blaabjerg, F. & Wang, H. An overview of artificial intelligence applications for power electronics. *IEEE Trans. Power Electron.* **36**, 4633–4658 (2021).
- Lawal, A. I. & Kwon, S. Application of artificial intelligence to rock mechanics: an overview. *J. Rock. Mech. Geotech. Eng.* **13**, 248–266 (2021).
- Assunção, G. et al. An overview of emotion in artificial intelligence. *IEEE Trans. Artif. Intell.* **3**, 867–886 (2022).
- Hochegger, B. et al. Artificial intelligence for cardiothoracic imaging: overview of current and emerging applications. *Semin. Roentgenol.* **58**, 184–195 (2023).
- Misra, J. & Saha, I. Artificial neural networks in hardware: a survey of two decades of progress. *Neurocomputing* **74**, 239–255 (2010).
- Poon, C. S. & Zhou, K. Neuromorphic silicon neurons and large-scale neural networks: challenges and opportunities. *Front. Neurosci.* **5**, 108 (2011).
- Silver, D. et al. Mastering the game of Go with deep neural networks and tree search. *Nature* **529**, 484–489 (2016).
- Esser, S. K. et al. Convolutional networks for fast, energy-efficient neuromorphic computing. *Proc. Natl Acad. Sci. USA* **113**, 11441–11446 (2016).
- Shafiee, A. et al. ISAAC: a convolutional neural network accelerator with in-situ analog arithmetic in crossbars. In *Proc. 2016 ACM/IEEE 43rd Annual International Symposium on Computer Architecture (ISCA)* 14–26 (IEEE, 2016).
- Graves, A. et al. Hybrid computing using a neural network with dynamic external memory. *Nature* **538**, 471–476 (2016).
- Chen, Y. H. et al. Eyeriss: an energy-efficient reconfigurable accelerator for deep convolutional neural networks. *IEEE J. Solid-State Circuits* **52**, 127–138 (2017).
- Liu, Y. Q. et al. Effective scaling of blockchain beyond consensus innovations and Moore's law: challenges and opportunities. *IEEE Syst. J.* **16**, 1424–1435 (2022).
- Lugt, A. V. Signal detection by complex spatial filtering. *IEEE Trans. Inf. Theory* **10**, 139–145 (1964).
- Goodman, J. W., Dias, A. R. & Woody, L. M. Fully parallel, high-speed incoherent optical method for performing discrete Fourier transforms. *Opt. Lett.* **2**, 1–3 (1978).
- Farhat, N. H. et al. Optical implementation of the Hopfield model. *Appl. Opt.* **24**, 1469–1475 (1985).
- Fisher, A. D. et al. Optical implementations of associative networks with versatile adaptive learning capabilities. *Appl. Opt.* **26**, 5039–5054 (1987).
- Caulfield, H. J., Kinser, J. & Rogers, S. K. Optical neural networks. *Proc. IEEE* **77**, 1573–1583 (1989).
- Psaltis, D. et al. Holography in artificial neural networks. *Nature* **343**, 325–330 (1990).
- Reck, M. et al. Experimental realization of any discrete unitary operator. *Phys. Rev. Lett.* **73**, 58–61 (1994).
- Clements, W. R. et al. Optimal design for universal multiport interferometers. *Optica* **3**, 1460–1465 (2016).
- Tait, A. N. et al. Microring weight banks. *IEEE J. Sel. Top. Quantum Electron.* **22**, 312–325 (2016).
- Liu, J. et al. Research progress in optical neural networks: theory, applications and developments. *Photonix* **2**, 5 (2021).
- Zhou, H. L. et al. Photonic matrix multiplication lights up photonic accelerator and beyond. *Light Sci. Appl.* **11**, 30 (2022).
- Huang, C. R. et al. Prospects and applications of photonic neural networks. *Adv. Phys. X* **7**, 1981155 (2022).
- Sheng, H. Y. Review of integrated diffractive deep neural networks. *Highlights Sci. Eng. Technol.* **24**, 264–278 (2022).
- Bai, Y. P. et al. Photonic multiplexing techniques for neuromorphic computing. *Nanophotonics* **12**, 795–817 (2023).
- Chen, Y. S. 4f-type optical system for matrix multiplication. *Opt. Eng.* **32**, 77–79 (1993).
- Chang, J. L. et al. Hybrid optical-electronic convolutional neural networks with optimized diffractive optics for image classification. *Sci. Rep.* **8**, 12324 (2018).
- Yan, T. et al. Fourier-space diffractive deep neural network. *Phys. Rev. Lett.* **123**, 023901 (2019).
- Zuo, Y. et al. All-optical neural network with nonlinear activation functions. *Optica* **6**, 1132–1137 (2019).
- Miscuglio, M. et al. Massively parallel amplitude-only Fourier neural network. *Optica* **7**, 1812–1819 (2020).
- Momeni, A. et al. Backpropagation-free training of deep physical neural networks. *Science* **382**, 1297–1303 (2023).
- Lin, X. et al. All-optical machine learning using diffractive deep neural networks. *Science* **361**, 1004–1008 (2018).
- Luo, Y. et al. Design of task-specific optical systems using broadband diffractive neural networks. *Light Sci. Appl.* **8**, 112 (2019).
- Qian, C. et al. Performing optical logic operations by a diffractive neural network. *Light Sci. Appl.* **9**, 59 (2020).
- Zhou, T. K. et al. In situ optical backpropagation training of diffractive optical neural networks. *Photonics Res.* **8**, 940–953 (2020).
- Wu, Z. C. et al. Neuromorphic metasurface. *Photonics Res.* **8**, 46–50 (2020).
- Zhou, T. K. et al. Large-scale neuromorphic optoelectronic computing with a reconfigurable diffractive processing unit. *Nat. Photonics* **15**, 367–373 (2021).
- Fang, T. et al. Classification accuracy improvement of the optical diffractive deep neural network by employing a knowledge distillation and stochastic gradient descent β -Lasso joint training framework. *Opt. Express* **29**, 44264–44274 (2021).
- Wang, P. P. et al. Orbital angular momentum mode logical operation using optical diffractive neural network. *Photonics Res.* **9**, 2116–2124 (2021).
- Li, Y. J. et al. Real-time multi-task diffractive deep neural networks via hardware-software co-design. *Sci. Rep.* **11**, 11013 (2021).
- Sun, Y. C. et al. Modeling and simulation of all-optical diffractive neural network based on nonlinear optical materials. *Opt. Lett.* **47**, 126–129 (2022).
- Shi, W. X. et al. LOEN: Lensless opto-electronic neural network empowered machine vision. *Light Sci. Appl.* **11**, 121 (2022).
- Liu, C. et al. A programmable diffractive deep neural network based on a digital-coding metasurface array. *Nat. Electron.* **5**, 113–122 (2022).



54. Li, J. X. et al. Polarization multiplexed diffractive computing: all-optical implementation of a group of linear transformations through a polarization-encoded diffractive network. *Light Sci. Appl.* **11**, 153 (2022).
55. Rahman, M. S. S. et al. Universal linear intensity transformations using spatially incoherent diffractive processors. *Light Sci. Appl.* **12**, 195 (2023).
56. Li, J. X. et al. Rapid sensing of hidden objects and defects using a single-pixel diffractive terahertz sensor. *Nat. Commun.* **14**, 6791 (2023).
57. Li, J. X. et al. Massively parallel universal linear transformations using a wavelength-multiplexed diffractive optical network. *Adv. Photonics* **5**, 016003 (2023).
58. Zheng, Z. Y. et al. Dual adaptive training of photonic neural networks. *Nat. Mach. Intell.* **5**, 1119–1129 (2023).
59. Duan, Z. Y., Chen, H. & Lin, X. Optical multi-task learning using multi-wavelength diffractive deep neural networks. *Nanophotonics* **12**, 893–903 (2023).
60. Zhang, H. et al. An optical neural chip for implementing complex-valued neural network. *Nat. Commun.* **12**, 457 (2021).
61. Cui, T. J., Liu, S. & Zhang, L. Information metamaterials and metasurfaces. *J. Mater. Chem. C* **5**, 3644–3668 (2017).
62. Duport, F. et al. All-optical reservoir computing. *Opt. Express* **20**, 22783–22795 (2012).
63. Paquot, Y. et al. Optoelectronic reservoir computing. *Sci. Rep.* **2**, 287 (2012).
64. Bueno, J. et al. Reinforcement learning in a large-scale photonic recurrent neural network. *Optica* **5**, 756–760 (2018).
65. Cheng, T. Y. et al. Optical neural networks based on optical fiber-communication system. *Neurocomputing* **364**, 239–244 (2019).
66. Huang, Y. Y. et al. Programmable matrix operation with reconfigurable time-wavelength plane manipulation and dispersed time delay. *Opt. Express* **27**, 20456–20467 (2019).
67. Zang, Y. B. et al. Electro-optical neural networks based on time-stretch method. *IEEE J. Sel. Top. Quantum Electron.* **26**, 7701410 (2020).
68. Xu, X. Y. et al. 11 TOPS photonic convolutional accelerator for optical neural networks. *Nature* **589**, 44–51 (2021).
69. Xiang, S. Y. et al. Computing primitive of fully VCSEL-based all-optical spiking neural network for supervised learning and pattern classification. *IEEE Trans. Neural Netw. Learn. Syst.* **32**, 2494–2505 (2021).
70. Zhang, L. H. et al. Optical machine learning using time-lens deep neural networks. *Photonics* **8**, 78 (2021).
71. Stelzer, F. et al. Deep neural networks using a single neuron: folded-in-time architecture using feedback-modulated delay loops. *Nat. Commun.* **12**, 5164 (2021).
72. Gu, B. L. et al. Enhanced prediction performance of a time-delay reservoir computing system based on a VCSEL by dual-training method. *Opt. Express* **30**, 30779–30790 (2022).
73. Xiang, S. Y. et al. Hardware-algorithm collaborative computing with photonic spiking neuron chip based on an integrated Fabry–Perot laser with a saturable absorber. *Optica* **10**, 162–171 (2023).
74. Xiang, S. Y. et al. Photonic integrated neuro-synaptic core for convolutional spiking neural network. *Opto Electron. Adv.* **6**, 230140 (2023).
75. Shi, Y. C. et al. Photonic integrated spiking neuron chip based on a self-pulsating DFB laser with a saturable absorber. *Photonics Res.* **11**, 1382–1389 (2023).
76. Guo, X. X. et al. Photonic implementation of the input and reservoir layers for a reservoir computing system based on a single VCSEL with two Mach-Zehnder modulators. *Opt. Express* **32**, 17452–17463 (2024).
77. Nahmias, M. A. et al. A leaky integrate-and-fire laser neuron for ultrafast cognitive computing. *IEEE J. Sel. Top. Quantum Electron.* **19**, 1800212 (2013).
78. Chakraborty, I. et al. Toward fast neural computing using all-photonic phase change spiking neurons. *Sci. Rep.* **8**, 12980 (2018).
79. Peng, H. T. et al. Temporal information processing with an integrated laser neuron. *IEEE J. Sel. Top. Quantum Electron.* **26**, 5100209 (2020).
80. Xiang, J. L. et al. All-optical silicon microring spiking neuron. *Photonics Res.* **10**, 939–946 (2022).
81. Jha, A. et al. Photonic spiking neural networks and graphene-on-silicon spiking neurons. *J. Lightwave Technol.* **40**, 2901–2914 (2022).
82. Ribeiro, A. et al. Demonstration of a 4×4 -port universal linear circuit. *Optica* **3**, 1348–1357 (2016).
83. Shen, Y. C. et al. Deep learning with coherent nanophotonic circuits. *Nat. Photonics* **11**, 441–446 (2017).
84. Hughes, T. W. et al. Training of photonic neural networks through in situ backpropagation and gradient measurement. *Optica* **5**, 864–871 (2018).
85. Harris, N. C. et al. Linear programmable nanophotonic processors. *Optica* **5**, 1623–1631 (2018).
86. Ong, J. R. et al. Photonic convolutional neural networks using integrated diffractive optics. *IEEE J. Sel. Top. Quantum Electron.* **26**, 7702108 (2020).
87. Shokraneh, F., Geoffroy-Gagnon, S. & Liboiron-Ladouceur, O. The diamond mesh, a phase-error-and loss-tolerant field-programmable MZI-based optical processor for optical neural networks. *Opt. Express* **28**, 23495–23508 (2020).
88. Bogaerts, W. et al. Programmable photonic circuits. *Nature* **586**, 207–216 (2020).
89. Bandyopadhyay, S., Hamerly, R. & Englund, D. Hardware error correction for programmable photonics. *Optica* **8**, 1247–1255 (2021).
90. Zhu, H. H. et al. Space-efficient optical computing with an integrated chip diffractive neural network. *Nat. Commun.* **13**, 1044 (2022).
91. Wang, X. Y. et al. Chip-based high-dimensional optical neural network. *Nano-Micro Lett.* **14**, 221 (2022).
92. Pai, S. et al. Experimentally realized in situ backpropagation for deep learning in photonic neural networks. *Science* **380**, 398–404 (2023).
93. Tait, A. N. et al. Neuromorphic photonic networks using silicon photonic weight banks. *Sci. Rep.* **7**, 7430 (2017).
94. Tait, A. N. et al. Feedback control for microring weight banks. *Opt. Express* **26**, 26422–26443 (2018).
95. Tait, A. N. et al. Silicon photonic modulator neuron. *Phys. Rev. Appl.* **11**, 064043 (2019).
96. Feldmann, J. et al. All-optical spiking neurosynaptic networks with self-learning capabilities. *Nature* **569**, 208–214 (2019).
97. Ohno, S. et al. Si microring resonator crossbar arrays for deep learning accelerator. *Jpn. J. Appl. Phys.* **59**, SGGE04 (2020).
98. Feldmann, J. et al. Parallel convolutional processing using an integrated photonic tensor core. *Nature* **589**, 52–58 (2021).
99. Huang, C. R. et al. A silicon photonic-electronic neural network for fibre nonlinearity compensation. *Nat. Electron.* **4**, 837–844 (2021).
100. Ohno, S. et al. Si microring resonator crossbar array for on-chip inference and training of the optical neural network. *ACS Photonics* **9**, 2614–2622 (2022).
101. Xu, S. F. et al. High-order tensor flow processing using integrated photonic circuits. *Nat. Commun.* **13**, 7970 (2022).
102. Zhang, W. P. et al. Silicon microring synapses enable photonic deep learning beyond 9-bit precision. *Optica* **9**, 579–584 (2022).
103. Bai, B. W. et al. Microcomb-based integrated photonic processing unit. *Nat. Commun.* **14**, 66 (2023).
104. Yin, R. Y. et al. Integrated WDM-compatible optical mode division multiplexing neural network accelerator. *Optica* **10**, 1709–1718 (2023).
105. Cheng, J. W. et al. Human emotion recognition with a microcomb-enabled integrated optical neural network. *Nanophotonics* **12**, 3883–3894 (2023).
106. Yu, N. F. et al. Light propagation with phase discontinuities: generalized laws of reflection and refraction. *Science* **334**, 333–337 (2011).
107. Zarei, S., Marzban, M. R. & Khavasi, A. Integrated photonic neural network based on silicon metalines. *Opt. Express* **28**, 36668–36684 (2020).
108. Fu, T. Z. et al. On-chip photonic diffractive optical neural network based on a spatial domain electromagnetic propagation model. *Opt. Express* **29**, 31924–31940 (2021).
109. Goi, E. et al. Nanoprinted high-neuron-density optical linear perceptrons performing near-infrared inference on a CMOS chip. *Light Sci. Appl.* **10**, 40 (2021).
110. Wang, Z. et al. Integrated photonic metasystem for image classifications at telecommunication wavelength. *Nat. Commun.* **13**, 2131 (2022).
111. Yan, T. et al. All-optical graph representation learning using integrated diffractive photonic computing units. *Sci. Adv.* **8**, eabn7630 (2022).
112. Luo, X. H. et al. Metasurface-enabled on-chip multiplexed diffractive neural networks in the visible. *Light Sci. Appl.* **11**, 158 (2022).
113. Zarei, S. & Khavasi, A. Realization of optical logic gates using on-chip diffractive optical neural networks. *Sci. Rep.* **12**, 15747 (2022).
114. Huang, Y. Y. et al. Sophisticated deep learning with on-chip optical diffractive tensor processing. *Photonics Res.* **11**, 1125–1138 (2023).
115. Fu, T. Z. et al. Photonic machine learning with on-chip diffractive optics. *Nat. Commun.* **14**, 70 (2023).
116. Poordashtban, O., Marzban, M. R. & Khavasi, A. Integrated photonic convolutional neural network based on silicon metalines. *IEEE Access* **11**, 61728–61737 (2023).
117. Fu, T. Z. et al. Integrated diffractive optical neural network with space-time interleaving. *Chin. Opt. Lett.* **21**, 091301 (2023).



118. Liu, W. C. et al. C-DONN: compact diffractive optical neural network with deep learning regression. *Opt. Express* **31**, 22127–22143 (2023).
119. Fu, T. Z. et al. Miniature on-chip diffractive optical neural network design. In *Proc. Conference on Lasers and Electro-Optics (CLEO)* 1–2 (IEEE, 2023).
120. Sun, R. et al. Multimode diffractive optical neural network. *Adv. Photonics Nexus* **3**, 026007 (2024).
121. Zhang, J. J. et al. Ultrashort and efficient adiabatic waveguide taper based on thin flat focusing lenses. *Opt. Express* **25**, 19894–19903 (2017).
122. Wang, Z. et al. On-chip wavefront shaping with dielectric metasurface. *Nat. Commun.* **10**, 3547 (2019).
123. Vandoorne, K. et al. Experimental demonstration of reservoir computing on a silicon photonics chip. *Nat. Commun.* **5**, 3541 (2014).
124. Perez, D. et al. Silicon photonics rectangular universal interferometer. *Laser Photonics Rev.* **11**, 1700219 (2017).
125. Khoram, E. et al. Nanophotonic media for artificial neural inference. *Photonics Res.* **7**, 823–827 (2019).
126. Hughes, T. W. et al. Wave physics as an analog recurrent neural network. *Sci. Adv.* **5**, eaay6946 (2019).
127. Zhang, T. et al. Efficient training and design of photonic neural network through neuroevolution. *Opt. Express* **27**, 37150–37163 (2019).
128. Moughames, J. et al. Three-dimensional waveguide interconnects for scalable integration of photonic neural networks. *Optica* **7**, 640–646 (2020).
129. Qu, Y. R. et al. Inverse design of an integrated-nanophotonics optical neural network. *Sci. Bull.* **65**, 1177–1183 (2020).
130. Zhao, X. M. et al. On-chip reconfigurable optical neural networks. at URL: <https://www.researchsquare.com/article/rs-155560/v1> (2021).
131. Wu, C. M. et al. Programmable phase-change metasurfaces on waveguides for multimode photonic convolutional neural network. *Nat. Commun.* **12**, 96 (2021).
132. Sunada, S. & Uchida, A. Photonic neural field on a silicon chip: large-scale, high-speed neuro-inspired computing and sensing. *Optica* **8**, 1388–1396 (2021).
133. Cheng, J. W. et al. Photonic emulator for inverse design. *ACS Photonics* **10**, 2173–2181 (2023).
134. Ashtiani, F., Geers, A. J. & Aflatouni, F. An on-chip photonic deep neural network for image classification. *Nature* **606**, 501–506 (2022).
135. Liao, K. et al. Matrix eigenvalue solver based on reconfigurable photonic neural network. *Nanophotonics* **11**, 4089–4099 (2022).
136. Ling, Q. et al. On-chip optical matrix-vector multiplier based on mode division multiplexing. *Chip* **2**, 100061 (2023).
137. Wu, T. W. et al. Lithography-free reconfigurable integrated photonic processor. *Nat. Photonics* **17**, 710–716 (2023).
138. Dong, B. W. et al. Higher-dimensional processing using a photonic tensor core with continuous-time data. *Nat. Photonics* **17**, 1080–1088 (2023).
139. Meng, X. Y. et al. Compact optical convolution processing unit based on multimode interference. *Nat. Commun.* **14**, 3000 (2023).
140. Giamougiannis, G. et al. Neuromorphic silicon photonics with 50 GHz tiled matrix multiplication for deep-learning applications. *Adv. Photonics* **5**, 016004 (2023).
141. Zheng, D. Z. et al. Experimental demonstration of coherent photonic neural computing based on a Fabry-Perot laser with a saturable absorber. *Photonics Res.* **11**, 65–71 (2023).
142. Rivenson, Y. et al. Phase recovery and holographic image reconstruction using deep learning in neural networks. *Light Sci. Appl.* **7**, 17141 (2018).
143. Shi, L. et al. Towards real-time photorealistic 3D holography with deep neural networks. *Nature* **591**, 234–239 (2021).
144. Zheng, H. Y. et al. Multichannel meta-imagers for accelerating machine vision. *Nat. Nanotechnol.* **19**, 471–478 (2024).
145. Zhong, C. Y. et al. Graphene/silicon heterojunction for reconfigurable phase-relevant activation function in coherent optical neural networks. *Nat. Commun.* **14**, 6939 (2023).
146. Shi, Y. et al. Nonlinear germanium-silicon photodiode for activation and monitoring in photonic neuromorphic networks. *Nat. Commun.* **13**, 6048 (2022).
147. Cheng, Z. G. et al. On-chip photonic synapse. *Sci. Adv.* **3**, e1700160 (2017).
148. Wright, C. D., Bhaskaran, H. & Pernice, W. H. P. Integrated phase-change photonic devices and systems. *MRS Bull.* **44**, 721–727 (2019).
149. Nisar, M. S. et al. On-chip integrated photonic devices based on phase change materials. *Photonics* **8**, 205 (2021).
150. Delaney, M. et al. Nonvolatile programmable silicon photonics using an ultralow-loss Sb₂Se₃ phase change material. *Sci. Adv.* **7**, eabg3500 (2021).
151. HUAWEI. Ascend 910 at URL: https://www.actfor.net/products/intelligent-computing/atlas/huawei-ai/ai-chips/Ascend_910/features (2024).
152. NVIDIA. T4 tensor core datasheet at URL: <https://www.nvidia.com/en-us/data-center/tesla-t4/> (2024).
153. Xu, Z. H. et al. Large-scale photonic chiplet Taichi empowers 160-TOPS/W artificial general intelligence. *Science* **384**, 202–209 (2024).
154. Ferdous, W. et al. New advancements, challenges and opportunities of multi-storey modular buildings—a state-of-the-art review. *Eng. Struct.* **183**, 883–893 (2019).
155. Cheng, J. W., Zhou, H. L. & Dong, J. J. Photonic matrix computing: from fundamentals to applications. *Nanomaterials* **11**, 1683 (2021).
156. Chen, Y. T. et al. All-analog photoelectronic chip for high-speed vision tasks. *Nature* **623**, 48–57 (2023).
157. Sluuds, A. et al. Delocalized photonic deep learning on the internet's edge. *Science* **378**, 270–276 (2022).
158. Lightmatter. Enviser and Passage at URL: <https://lightmatter.co> (2023).
159. Lightelligence. Pace at URL: <https://www.lightelligence.ai/index.php> (2023).
160. Jouppi, N. P. et al. In-datacenter performance analysis of a tensor processing unit. In *Proc. ACM/IEEE 44th Annual International Symposium on Computer Architecture (ISCA)* 1–12 (IEEE, 2017).
161. Harris, S. E. Electromagnetically induced transparency. *Phys. Today* **50**, 36–42 (1997).
162. Fleischhaue, M., Imamoglu, A. & Marangos, J. P. Electromagnetically induced transparency: optics in coherent media. *Rev. Mod. Phys.* **77**, 633–673 (2006).

