Frontiers of Optoelectronics (2022) 15:15
https://doi.org/10.1007/s12200-022-00009-4

Frontiers of Optoelectronics

RESEARCH ARTICLE

A small microring array that performs large complex-valued matrix-vector multiplication

Junwei Cheng[1] · Yuhe Zhao[1] · Wenkai Zhang[1] · Hailong Zhou[1,2,3] · Dongmei Huang[3,4] · Qing Zhu[5] · Yuhao Guo[5] · Bo Xu[5] · Jianji Dong[1] · Xinliang Zhang[1]

Received: 26 September 2021 / Accepted: 2 November 2021

Abstract
As an important computing operation, photonic matrix–vector multiplication is widely used in photonic neutral networks and signal processing. However, conventional incoherent matrix–vector multiplication focuses on real-valued operations, which cannot work well in complex-valued neural networks and discrete Fourier transform. In this paper, we propose a systematic solution to extend the matrix computation of microring arrays from the real-valued field to the complex-valued field, and from small-scale (i.e., 4×4) to large-scale matrix computation (i.e., 16×16). Combining matrix decomposition and matrix partition, our photonic complex matrix–vector multiplier chip can support arbitrary large-scale and complex-valued matrix computation. We further demonstrate Walsh-Hardmard transform, discrete cosine transform, discrete Fourier transform, and image convolutional processing. Our scheme provides a path towards breaking the limits of complex-valued computing accelerator in conventional incoherent optical architecture. More importantly, our results reveal that an integrated photonic platform is of huge potential for large-scale, complex-valued, artificial intelligence computing and signal processing.

Keywords Photonic matrix–vector multiplication · Complex-valued computing · Microring array · Signal/image processing

1 Introduction

With the rapid advancement of technology in recent decades, there is a growing demand for large-capacity, high-speed computing over traditional computing. This is especially seen in the field of convolutional processing, a computationally intensive operation in electronics that occupies over 80% of the total processing time for image processing [1–3]. Optical computing has the ability of parallel processing with wavelength division multiplexing (WDM) due to its intrinsic high speed and low power consumption, thus has been proposed as a promising candidate for mass data processing [4]. Matrix multiplication is the kernel and most common operation in artificial intelligence (AI). It is widely used in artificial neutral networks (ANNs), which have been universally applied in signal processing, imaging recognition, voice recognition, real-time video analysis, and autonomous driving [5, 6]. The optical neural networks (ONNs) can improve the computation speed by several orders of magnitude. For example, a photonic convolutional accelerator comprised of soliton microcombs could carry out up to 10 trillion operations per second [7]. In addition, phase-change material (PCM) has been employed in non-volatile memory storage in optical computing to reduce the energy consumption of optical-electrical conversion during weight data refreshing [8–11]. Recently, an integrated photonic hardware accelerator has successfully executed 10^{12} multiply-accumulate operations per second by combining phase-change-material memory and soliton microcombs [9].

Junwei Cheng and Yuhe Zhao contributed equally to this work.

✉ Jianji Dong
jjdong@mail.hust.edu.cn

1 Wuhan National Laboratory for Optoelectronics, School of Optical and Electronic Information, Huazhong University of Science and Technology, Wuhan 430074, China

2 Photonics Research Centre, Department of Electronic and Information Engineering, The Hong Kong Polytechnic University, Hong Kong 999077, China

3 The Hong Kong Polytechnic University Shenzhen Research Institute, Shenzhen 518057, China

4 Photonics Research Centre, Department of Electrical Engineering, The Hong Kong Polytechnic University, Hong Kong 999077, China

5 Institute of Strategic Research, Huawei Technologies, Shenzhen 518129, China

Published online: 28 April 2022

A copious amount of research has been conducted in optical matrix computing using spatial light modulators [12, 13], electro-optic modulations [14–16], direct driven LED arrays [17], acousto-optic Bragg cells [18–20], and photorefractive medias [21–23]. Although spatial light modulators and other spatial elements are easily programmable, these methods are in general bulky, complex, and power-consuming. With the advancement of integrated photonics technology and hardware implementation of nanophotonic processors, integrated photonic platforms have shown huge potential for high-performance computing. At present, most existing neural networks are based solely on real-valued algorithms, but complex-valued algorithms may provide a significant advantage when performing tasks, such as the symmetry or XOR problem [24]. A great deal of research on integrated optical computing networks has been done using a cascaded Mach Zehnder interferometer (MZI) mesh [25–28]. MZI meshes have been widely used in linear optical circuits [25, 29], quantum information processing [30], universal multiport interferometers [27], optical modes descramblers [31, 32], and polarization processors [33]. For the linear section of optical neutral networks, impressive works, such as vowel recognition, have been demonstrated [34]. This method allows for good reconfigurability and independent control of both the amplitude and phase. However, the loading of the transmission matrix relies on iterative algorithms, which are quite slow and unsuitable for flexible matrix computations. Moreover, MZIs require a larger power consumption than resonant devices, such as microring resonators (MRRs), which are compact (several micron radius), more energy-efficient, highly integrated, and easily scalable [35, 36]. MRRs are resonant devices and the transmission coefficients are wavelength-sensitive. Parallel incoherent matrix computing can be achieved by controlling the resonant states of MRRs, which is commonly used in optical tensor computing and ONNs [11, 37]. The problem of MRR arrays is that the computation is incoherent, which means MRR arrays can only perform amplitude modulation without phase information. Thus, MMR arrays can only compute non-negative or real numbers assisted by differential detection. In addition, ultra-large-scale MRRs are difficult to implement because of the heavy thermal crosstalk and electronic circuits packaging. Hence, it is believed that MRRs cannot be implemented in a large-scale matrix multiplication to compute complex numbers.

In this paper, we present a systematic solution to extend the matrix computation of MRR arrays from the real-valued field to the complex-valued field, and from small scale (i.e., 4×4) to large scale matrix computation (i.e., 16×16). We experimentally demonstrate typical matrix–vector multiplication (MVM) applications of MRR arrays in Walsh Hardmard transform (WHT), discrete cosine transform (DCT), discrete Fourier transform (DFT), and image convolutional processing. These applications have significantly expanded the fields of optical computation based on MRR arrays. Our work shows huge potential for high-speed and universal matrix computations, such as applications in photonic accelerators and optical artificial intelligence.

2 Principle

The structure of the proposed on-chip MRR array (i.e., photonic complex-MVM core) is schematically illustrated in Fig. 1. The on-chip photonic complex-MVM core consists of a tunable silicon MRR array that includes 16 add-drop MRRs arranged in 4 rows and 4 columns. The entire architecture is based on wavelength-division multiplexing (WDM) and on-chip reconfigurable MRR array. The MRR array forms a complete network of a 4×4 transmission matrix, whose configuration can be realized by tuning the heater of each MRR.

Without consideration of the transmission loss, every add-drop MRR in each row of the array decides the through transmittance coefficient of $1-a_{ij}$ and drop transmittance coefficient of a_{ij}, respectively [38]. Then, the difference of these two ports is given by

$$\boldsymbol{O}=\boldsymbol{X}\boldsymbol{I}=\begin{bmatrix} 1-2a_{11} & 1-2a_{12} & 1-2a_{13} & 1-2a_{14}\\ 1-2a_{21} & 1-2a_{22} & 1-2a_{23} & 1-2a_{24}\\ 1-2a_{31} & 1-2a_{32} & 1-2a_{33} & 1-2a_{34}\\ 1-2a_{41} & 1-2a_{42} & 1-2a_{43} & 1-2a_{44} \end{bmatrix}\begin{bmatrix} i_1\\ i_2\\ i_3\\ i_4 \end{bmatrix}, \tag{1}$$

where the 4×1 vector $\boldsymbol{O}=\left[o_1,o_2,o_3,o_4\right]^{\mathrm{T}}$ represents the output vector, 4×1 vector $\boldsymbol{I}=\left[i_1,i_2,i_3,i_4\right]^{\mathrm{T}}$ represents the input vector, and 4×4 matrix $\boldsymbol{X}$ stands for the transmission matrix. When the transmission loss is ignored, the drop port coefficient a_{ij} falls in the range of $[0,1]$ and the corresponding coefficient in the transmission matrix, defined by $1-2a_{ij}$, falls in the range of $[-1,1]$. Thus, in the MVM operation, the input vector of $\boldsymbol{I}$ is non-negative, while the transmission matrix of $\boldsymbol{X}$ and the output vector of $\boldsymbol{O}$ can cover the real number field.

Figure 1 also shows the working principle to extend the matrix computation of the MRR array from the real-valued field to the complex-valued field, and from small-scale (i.e., 4×4) to large-scale matrix computation. Combining matrix decomposition and matrix partition, our photonic complex-MVM chip can support arbitrary large-scale and complex-valued matrix computation.

Without loss of generality, the MVM consists of an 8×1 complex input matrix of $\boldsymbol{I}$, 8×8 complex transmission matrix of $\boldsymbol{X}$, and output matrix of $\boldsymbol{O}$. To process a large amount of MVM, the size of the matrices is reduced

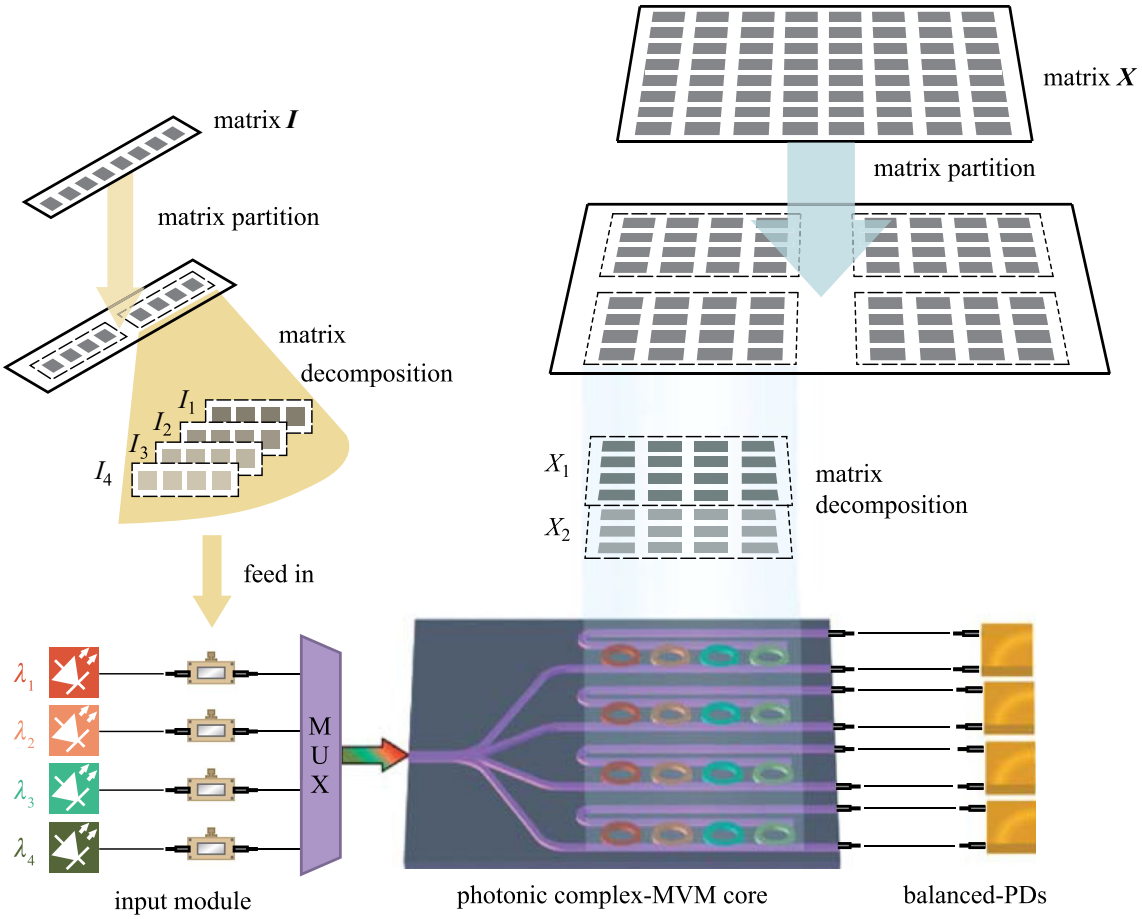

Fig. 1 Working principle of complex-valued MVM. The entire architecture containing the input module, photonic complex-MVM core, and balanced-PDs. To realize the complex-valued MVM, the input matrix $\boldsymbol{I}$ and transmission matrix $\boldsymbol{X}$ are partitioned, decomposed, and subsequently fed into the input module and MVM core, respectively. The different colors correspond to different light wavelengths

through matrix partition. Matrix $\boldsymbol{I}$ can be broken into two 4×1 matrices, while matrix $\boldsymbol{X}$ can be divided into four 4×4 matrices. To process complex MVM in full complex number field, matrix $\boldsymbol{I}$ is divided into $\boldsymbol{I}_1$, $\boldsymbol{I}_2$, $\boldsymbol{I}_3$, $\boldsymbol{I}_4$, defined as the positive real, positive imaginary, negative real, and negative imaginary parts of matrix $\boldsymbol{I}$, respectively. Matrix $\boldsymbol{X}$ is also divided into $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$, representing the real and imaginary parts of $\boldsymbol{X}$. The elements of the input submatrix, $\boldsymbol{I}_n = \left[i_1, i_2, i_3, i_4\right]_n^{\mathrm{T}} (n = 1, 2, 3, 4)$, are loaded onto the beams with different wavelengths of λ_1, λ_2, λ_3, and λ_4 by optical intensity modulators (IMs). After mixing by a wavelength multiplexer (MUX), the input is equally divided into four branches, each of which consists of four independent MRRs aligned to resonate the λ_1, λ_2, λ_3, and λ_4 wavelengths, respectively. Matrix $\boldsymbol{X}_n (n = 1, 2)$ is loaded onto the photonic complex MVM core with the 4×4 MRR array, where the coefficients are determined by the voltages applied to each MRR. The output matrix of $\boldsymbol{O}$ is detected by balanced photodetectors (PDs).

If the input vectors of $\boldsymbol{I}_1$, $\boldsymbol{I}_2$,…, $\boldsymbol{I}_m$ are loaded in series, the input vector can be expanded into a $n\times m$ matrix where $\boldsymbol{I} = [\boldsymbol{I}_1, \boldsymbol{I}_2, \ldots, \boldsymbol{I}_m]$. Similarly, the corresponding output powers of $\boldsymbol{O}_1, \boldsymbol{O}_2, \ldots, \boldsymbol{O}_m$ should be measured in series so that the output $m\times n$ matrix can be written as $\boldsymbol{O} = [\boldsymbol{O}_1, \boldsymbol{O}_2, \ldots, \boldsymbol{O}_m]$. Hence, the MVM can be expanded into matrix–matrix multiplication denoted by the following equation:

$$\left[\boldsymbol{O}_1, \boldsymbol{O}_2, \ldots, \boldsymbol{O}_m\right] = \boldsymbol{X}\left[\boldsymbol{I}_1, \boldsymbol{I}_2, \ldots, \boldsymbol{I}_m\right]. \tag{2}$$

3 Results

3.1 Fabrication and experimental setup

The proposed device was fabricated on a silicon-on-insulator (SOI) platform. A 725 μm SOI wafer with 220 nm of top silicon and 2 μm of buried oxide (BOX) was used. The layout is transferred onto photoresist using electron beam

lithography (EBL) and the top silicon is etched by inductively coupled plasma (ICP). The grating coupler is shallowly etched by 70 nm, while the silicon waveguide is fully etched by 220 nm. Between the waveguide and metal electrodes, 1 μm of silicon dioxide was deposited using plasma enhanced chemical vapor deposition (PECVD). The metal for the heaters and pads was deposited by electron beam evaporator (EBE). The heaters were made of 150 nm thick and 1 μm wide Ti. The electrical wires and pads were made of 20/250 μm thick Ti/Au.

The microscope image of the fabricated chip is illustrated in Fig. 2a. The input signal is injected through a grating coupler on the left and subsequently divided into four identical branches with a 4×4 MRR array. There are eight output gratings, representing the bus through waveguides and bus drop waveguides for each row of MRRs. The eight output gratings are placed in equal distances of 127 μm, the exact distance of the fiber array (FA) coupler. Figure 2b shows the packaged chip, where the metal pads are connected to the printed circuit board (PCB) by wire-bonding and the PCB is controlled by a custom 120-channel voltage source via a flexible flat cable. The input optical grating is coupled to an optical fiber that is vertically glued to the SOI chip. The output optical gratings on the chip are coupled to an optical FA that is attached to the PCB and equally distributed in 127 μm spacing V-groove, so that vertical output light from the chip is reflected 45° by the FA.

The experimental setup is shown in Fig. 2c. A continuous-wave (CW) laser was used as the stable optical source for the IMs. The electrical input data was encoded by a programmable voltage source and used as the driving signal that was temporally fed into the IMs. Since the output of the modulator is polarization-dependent, PCs were placed before and after the IMs to control the polarizations. A dense wavelength division multiplexing component (DWDM) was employed to combine the four wavelengths into a bus waveguide coupled to the packaged SOI chip. The optical powermeter is capable of both detecting and displaying the power values of the optical signals, which allowed us to obtain and record the results directly.

To verify the MVM function, IMs were used to configure the input vector $\boldsymbol{I}$, while the transmission matrix $\boldsymbol{X}$ was loaded by tuning the voltages applied on the MRR array. The output power values were then obtained from the balanced PDs. After calibration and normalization, the output vector $\boldsymbol{O}$ was obtained. When the input is the identity matrix, the

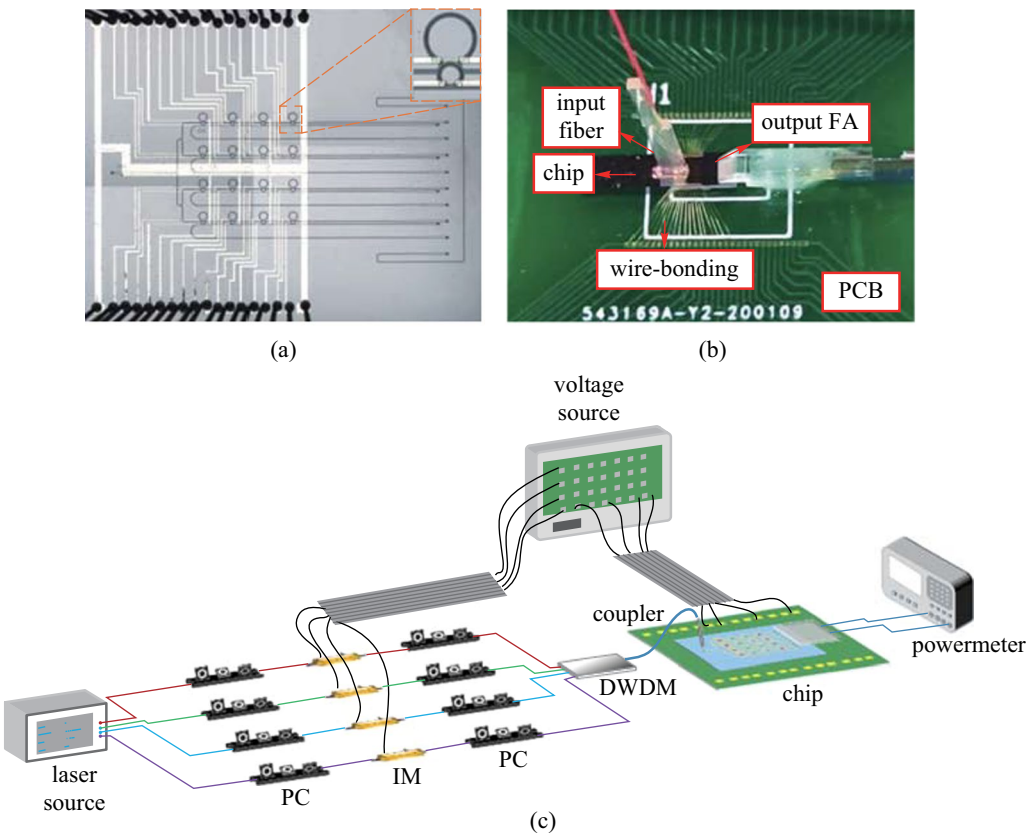

Fig. 2 **a** Microscope image of the on-chip 4×4 MRR array, inset: zoomed in microscope image of the tunable MZI-MRR. **b** Image of the packaged chip. FA fiber array, PCB printed circuit board. **c** Experimental setup of the matrix arithmetic processor. PC polarization controller, IM intensity modulator, DWDM dense wavelength division multiplexing

output matrix $\boldsymbol{O}$ is equivalent to the transmission matrix $\boldsymbol{X}$, allowing the transmission matrix $\boldsymbol{X}$ to be directly read at the output ports. In practical situations, the variation ranges of through transmittance coefficient and drop transmittance coefficient are different due to MRR loss. In this case, the coefficients will require recalibration for actual optical matrix computation (See Appendix A).

To statistically describe the performance of this multiplier, over 500 sets of input vector data and matrix, $\boldsymbol{X}$, were configured to the IMs and MRR array, respectively. Experimental results showed that the majority of the absolute values of the errors fall within the range of $0-0.1$, which suggets rather accurate computing. See Appendix B for more details.

3.2 Matrix–vector multiplication extending to the full real number field

Since the input vector $\boldsymbol{I}$ was determined by the optical powers modulated by the IMs, the elements must be non-negative. Although the transmission matrix $\boldsymbol{X}$ and output vector $\boldsymbol{O}$ can only cover the real number field, our proposed scheme allows for the conversion of the input elements into negative values, extending the MVM to the full real number field.

Figure 3 illustrates the proposed scheme. First, the input vector (real numbers) was divided into $\boldsymbol{I}_+$, containing all the positive elements and zeros, and $\boldsymbol{I}_-$, containing all the absolute values of the negative elements. The relationship between $\boldsymbol{I}_+$, and $\boldsymbol{I}_-$ are given by

$$\begin{cases} \boldsymbol{I}_+ = \frac{|\boldsymbol{I}|+\boldsymbol{I}}{2}, \\ \boldsymbol{I}_- = \frac{|\boldsymbol{I}|-\boldsymbol{I}}{2}, \\ I = \boldsymbol{I}_+ - \boldsymbol{I}_-. \end{cases} \tag{3}$$

The resulting two non-negative vectors, $\boldsymbol{I}_+$ and $\boldsymbol{I}_-$, are subsequently used in place of the origin input vector. The transmission matrix $\boldsymbol{X}$ was then loaded and the input vectors were configured as $\boldsymbol{I}_+$ and $\boldsymbol{I}_-$, respectively, to obtain the two output vectors, $\boldsymbol{P}$ and $\boldsymbol{Q}$. The targeted output matrix $\boldsymbol{O}$ was obtained following subtraction operation. The relationships between $\boldsymbol{P}$, $\boldsymbol{Q}$, and $\boldsymbol{O}$ are expressed below

$$\begin{cases} P = X\boldsymbol{I}_+, \\ Q = X\boldsymbol{I}_-, \\ O = P - Q. \end{cases} \tag{4}$$

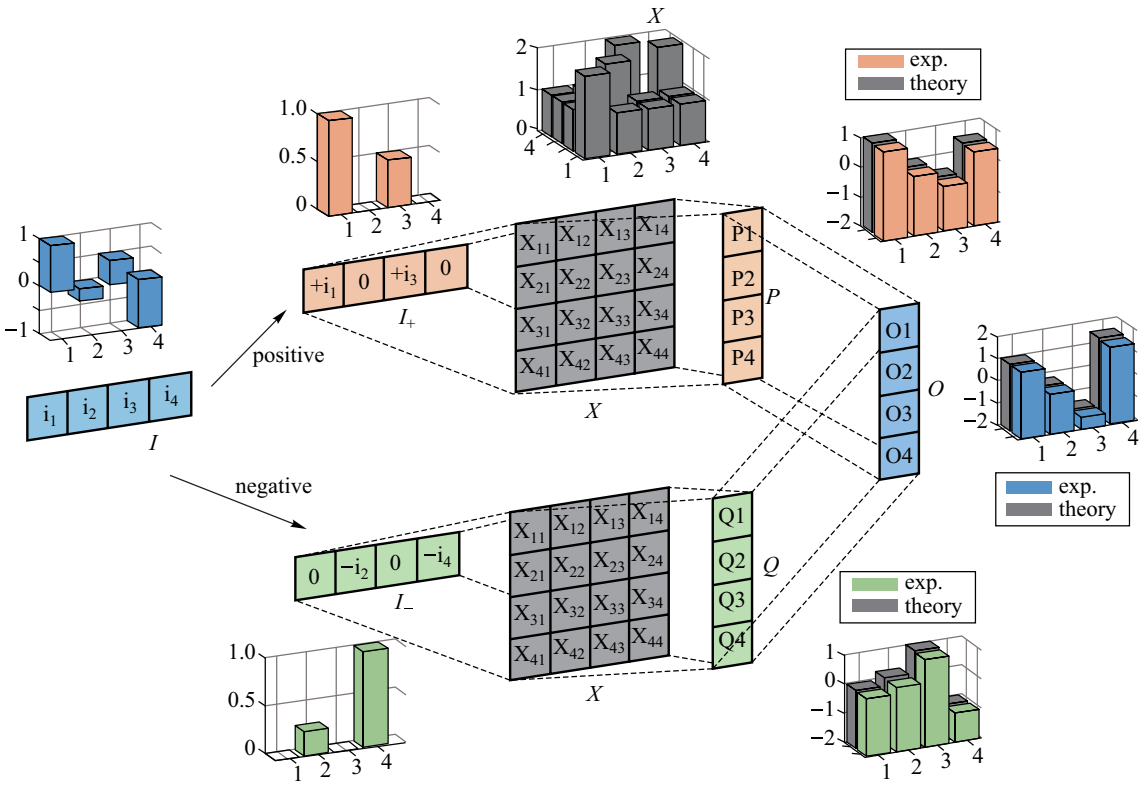

Fig. 3 Matrix computation extending to the full real number field. The 4×1 block array represents the input or output vectors and the 4×4 block array represents the transmission matrix. The bar graph shows the results from one operation, where the inputs or experimental outputs are represented by the colored bars and the theoretical outputs or transmission matrix are represented by the gray bars

Using the method described above, we were able to successfully split a real-valued optical MVM operation into two non-negative optical MVMs and one subtraction in the electrical domain. Figure 3 shows an experimental example of a real-valued MVM. The theoretical and experimental results are shown in three-dimensional bar graphs next to the corresponding matrices or vectors.

3.3 Matrix–vector multiplication extending to the full complex number field

To further extend our matrix computation into the complex number field, the input vector $\boldsymbol{I}$ and transmission matrix $\boldsymbol{X}$ were both separated into a real part and imaginary part. The output vector can be expressed as

$$\boldsymbol{O} = \boldsymbol{XI} = (\mathrm{real}(\boldsymbol{X}) + \mathrm{i} * \mathrm{imag}(\boldsymbol{X}))(\mathrm{real}(\boldsymbol{I}) + \mathrm{i} * \mathrm{imag}(\boldsymbol{I})), \tag{5}$$

where i is the square root of minus one, real($\boldsymbol{M}$) represents the real part of matrix $\boldsymbol{M}$, and imag($\boldsymbol{M}$) represents the imaginary part of matrix $\boldsymbol{M}$ (here, $\boldsymbol{M}$ can be $\boldsymbol{X}$, $\boldsymbol{I}$ or $\boldsymbol{O}$).

The matrix multiplication can then be divided into

$$\begin{cases} \mathrm{real}(\boldsymbol{O}) = \mathrm{real}(\boldsymbol{X})\mathrm{real}(\boldsymbol{I}) - \mathrm{imag}(\boldsymbol{X})\mathrm{imag}(\boldsymbol{I}), \\ \mathrm{imag}(\boldsymbol{O}) = \mathrm{real}(\boldsymbol{X})\mathrm{imag}(\boldsymbol{I}) + \mathrm{imag}(\boldsymbol{X})\mathrm{real}(\boldsymbol{I}). \end{cases} \tag{6}$$

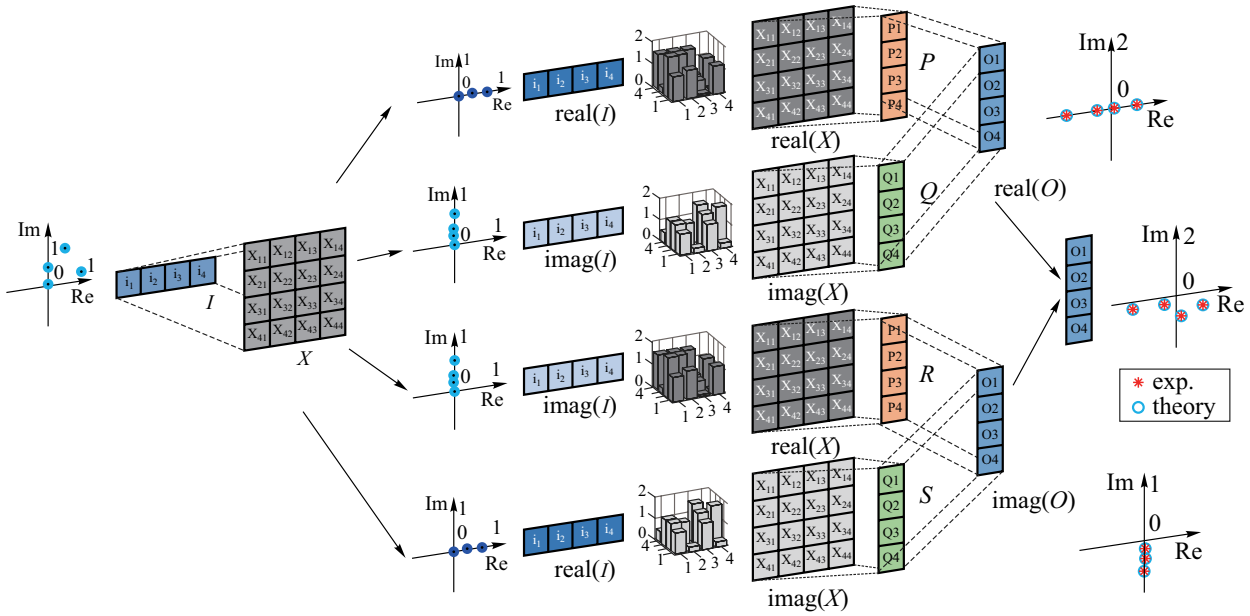

Fig. 4 **a** Matrix computation extending to the full complex number field. The 4×1 block array represents the input or output vectors and the 4×4 block array represents the transmission matrix. The bar graph shows the theoretical transmission matrix. The coordinate figure shows the complex inputs or experimental complex outputs of the operation. **b** Theoretically expected and experimental results for the complex-valued matrix multiplication

As seen in Fig. 4a, the complex-valued matrix multiplication was divided into four operations of optical MVMs, specifically real($\boldsymbol{X}$)real($\boldsymbol{I}$), imag($\boldsymbol{X}$)imag($\boldsymbol{I}$), real($\boldsymbol{X}$)imag($\boldsymbol{I}$), and imag($\boldsymbol{X}$)real($\boldsymbol{I}$), as well as two operations of electrical addition or subtraction operations. Figure 4a also shows an experimental demonstration of complex MVM. The two-dimensional coordinate diagrams in blue dots represent the corresponding input vectors or output vectors, and the three-dimensional gray bar graphs represent the transmission matrix. The experimental results are consistent with the theoretical results. In addition, the experimental results presented in Fig. 4b of the output of complex-valued matrix multiplication are also consistent with the predicted results.

3.4 Matrix–vector multiplication extending to higher dimensions

Considering the fact that partition of matrix can enlarge the matrix dimension, we were able to implement a high dimensional MVM with low dimensional MRR array via matrix partition. Figure 5 illustrates the basic principle of matrix partition. The input and output data are 8×1 vectors and the transmission matrix of $\boldsymbol{X}$ is an 8×8 matrix. To execute the 8×8 matrix computation using our 4×4 processor, the input and output vectors have to be split into two 4×1 vectors. Meanwhile, the transmission matrix is broken into four 4×4 matrices. Therefore, the equation can be written as

$$\boldsymbol{O} = \begin{pmatrix} o_1 \\ o_2 \end{pmatrix} = \begin{pmatrix} \boldsymbol{X}_{11} & \boldsymbol{X}_{12} \\ \boldsymbol{X}_{21} & \boldsymbol{X}_{22} \end{pmatrix} \begin{pmatrix} I_1 \\ I_2 \end{pmatrix} = \begin{pmatrix} X_{11}I_1 + X_{12}I_2 \\ X_{21}I_1 + X_{22}I_2 \end{pmatrix}. \tag{7}$$

Therefore, the partition of matrix can be realized by four rounds of optical MVMs and two rounds of electrical additions. Figure 5 shows an experimental demonstration of a partition of MVM, where the theoretical or experimental results are given in the three-dimensional bar graphs. It can be also seen from Fig. 5 that the experimental results are in agreement with the theoretical predictions.

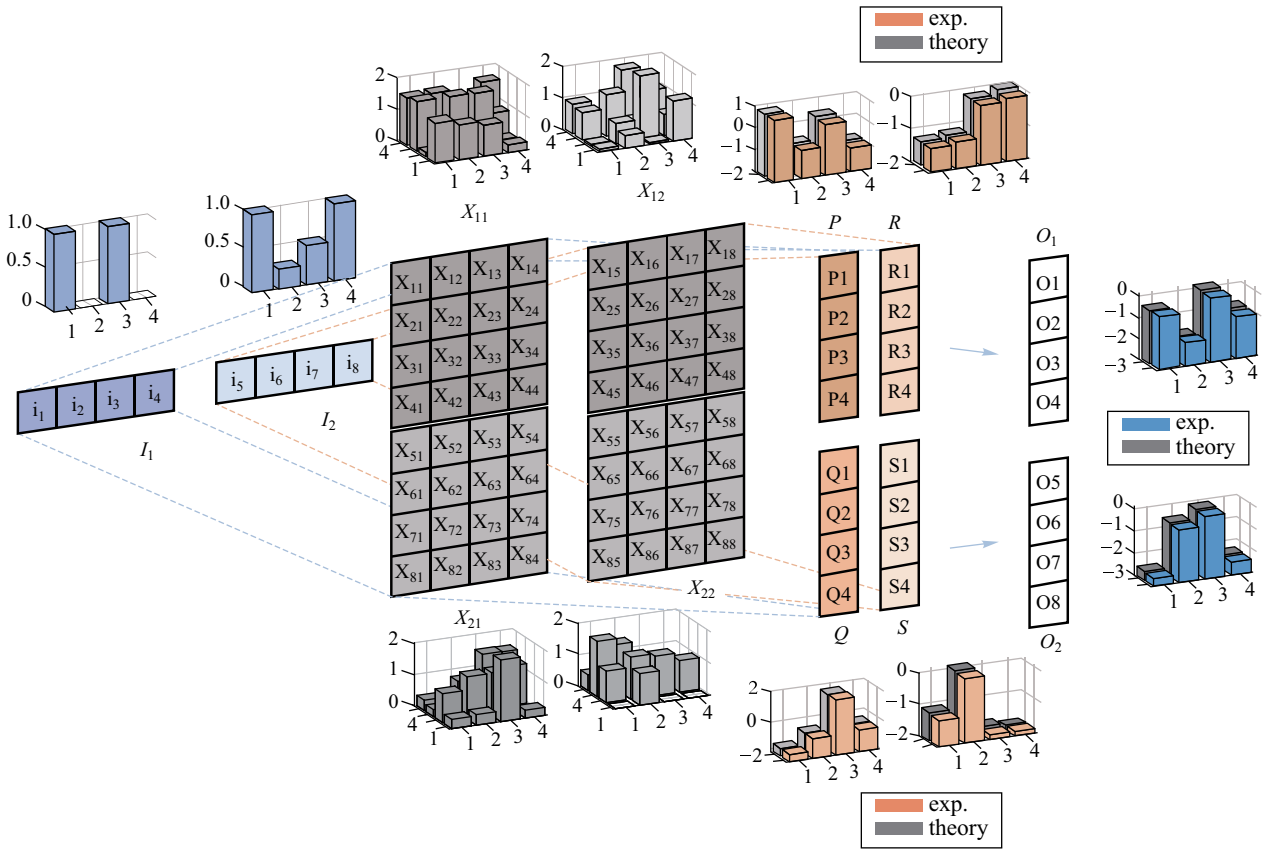

Fig. 5 Example of the partition of an 8×8 MVM. The 4×1 block array represents the input or output vectors, and the 4×4 block array represents the transmission matrix. The bar graph shows the results from one operation, where the inputs or experimental outputs are represented by colored bars and the theoretical outputs are represented by gray bars

3.5 Applications in signal transformation and image processing

Modern signal and image processing are two fields where algorithms based on large complex MVMs are widely utilized. This paper demonstrates three typical signal transformations, specifically, discrete WHT, DCT, and DFT [39]. WHT is orthogonal transformation that is widely used in imaging and code division multiple access [40]. The Hadamard matrix elements are equal to 1 or − 1, so that there are only addition and subtraction operations in the calculation, making it much simpler than DFT and DCT. Energy concentration is a characteristic of WHT, meaning the more uniform the numbers in the original data are, the more concentrated the transformed data are on the side. This property makes WHT advantageous for image compression [41]. Figure 6a shows the input signal and Fig. 6e shows the transformed signals after our matrix size to 16 × 16 was extended. One can see that WHT can compress information in the low frequency region if the input signal has a uniform amplitude distribution, thus the high frequency region can be ignored since it has a very low amplitude. DCT plays an important role in signal processing, signal modulation, and demodulation [42]. A periodic sequence was input into a 16 × 16 network and the output matrix was calculated, as shown in Fig. 6b and f. The first half of the former sequence was loaded into an 8 × 8 network as the input, depicted in Fig. 6c. The resulting output vector is quite similar to that presented in Fig. 6f and g. These results reflect the symmetry of DCT and provide supporting evidence that our system can correctly perform DCT. In addition, DFT can convert a signal sampling in time domain into frequency domain, one of the most frequently used operations in signal transformation [43]. Here, we used an input signal in the form of a square wave. Since DFT is a complex transformation, the amplitude of the output sequence is shown in form of its absolute value, which is shaped in a sinc function, as shown in Fig. 6d and h. The results show that not only can DFT be performed by our system, the calculation errors are also very small.

Image convolution is of paramount importance to convolutional neural networks and image processing, which can be performed in optical domain to achieve convolutional acceleration. To experimentally verify image convolution with our MVM, we choose the logo of Wuhan National Laboratory for Optoelectronics (WNLO) as an example, as well as seven different 3 × 3 sized kernels. The kernels are designed to perform different image processing functions or highlight different edges of the original image. The pixel values of the input image are loaded into the IMs by the electrical waveform and the on-chip MRR array is loaded by the transmission matrix representing the kernel. Figure 7 shows the

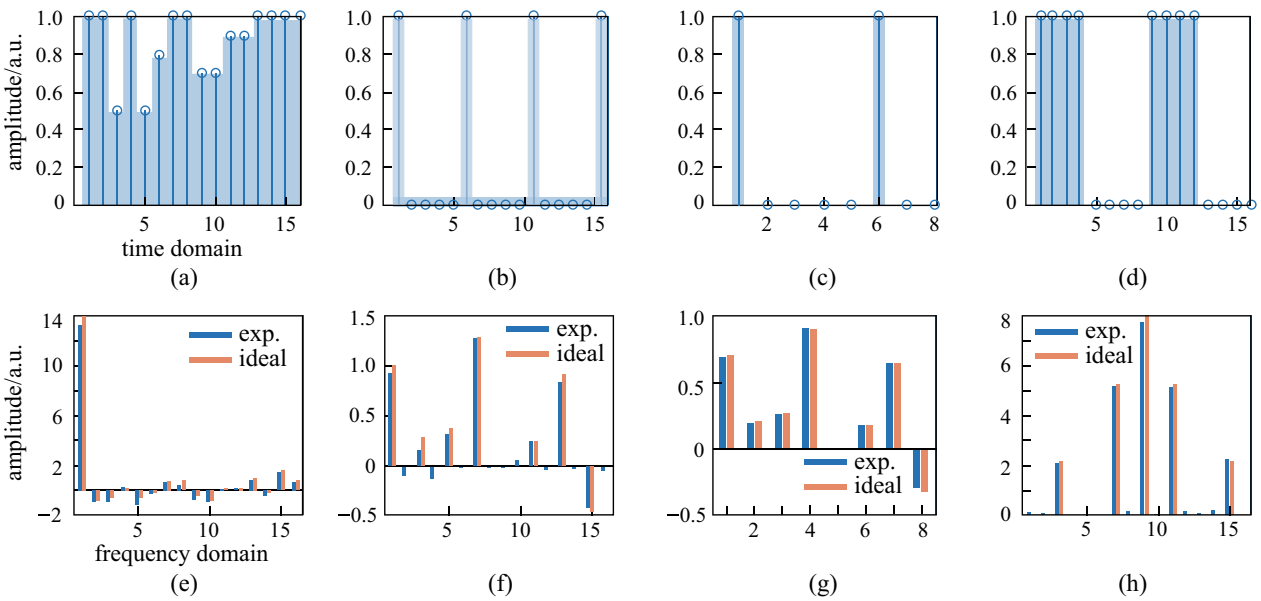

Fig. 6 Input and output signal sequence of three signal transformation. The input sequence of **a** WHT, **b** and **c**, DCT, and **d** DFT. **e** Experimental results (blue bars) and ideal (red bars) output of WHT. **f** and **g** Experimental results (blue bars) and ideal (red bars) output of DCT. **h** Experimental results (blue bars) and ideal (red bars) output of DFT

experimental results, including the recovered feature maps and corresponding transmission matrices of the kernels.

Compared with the original image, the edge features of the processed image are clearly visible in Fig. 7e–h, demonstrating the effectiveness of the optical convolution operation. The kernels in Fig. 7b–d correctly performed different image processing functions, including blur, motion blur, and sharpen. The kernels in Fig. 7e–h highlighted the edges of the original image in different directions. Using the theoretical results as reference, we determined that the calculation errors of the optical convolution operation was mainly concentrated on the bright part (i.e., high pixel value area) of the image, which indicates that these errors are largely caused by thermal crosstalk, rather than noise. Real-time calibration algorithms and external temperature control devices are implemented for system stability.

4 Discussion and future perspective

The experimental results of both signal and image processing clearly demonstrate that our proposed system is able to extend matrix computation to (1) real numbers, (2) full complex numbers, (3) higher processing dimensions, and (4) convolution. Thus, the processor can serve as a universal matrix arithmetic processor for complex tasks in various application scenarios.

However, the processor can be further improved in several ways. For example, the computational efficiency can be multiplied by making full use of parallel computation or by increasing the number of input wavelengths. Note that the transmission spectrum of MRR is repeated with a period of about 6 nm, which represents the free spectral range (FSR) of MRR. Therefore, multiple sets of input vectors with an interval equal to FSR can be operated simultaneously, as shown in Fig. 8. Suppose that there are m sets of different input vectors and the wavelengths of the input matrices are set as $(\lambda_1, \lambda_2, \lambda_3, \lambda_4) + p\mathrm{FSR}$, where $p = 0, 1, \ldots, m-1$. To obtain the output data, the output powers of each row are divided by the wavelength-division multiplexer and separately detected by m sets of corresponding balanced PDs. In this process, the state of the transmission matrix is fixed (i.e., the state of MRR array is fixed), while the m sets of input and output vectors are independently paralleled. This means that m sets of MVMs can be executed simultaneously, demonstrating the possibility of parallel optical computation. Secondly, full integration is crucial to improve the competitiveness of optical computing compared to electrical matrix processing. As shown in Fig. 8, an optical comb is integrated into the chip, providing a series of comb lines that are

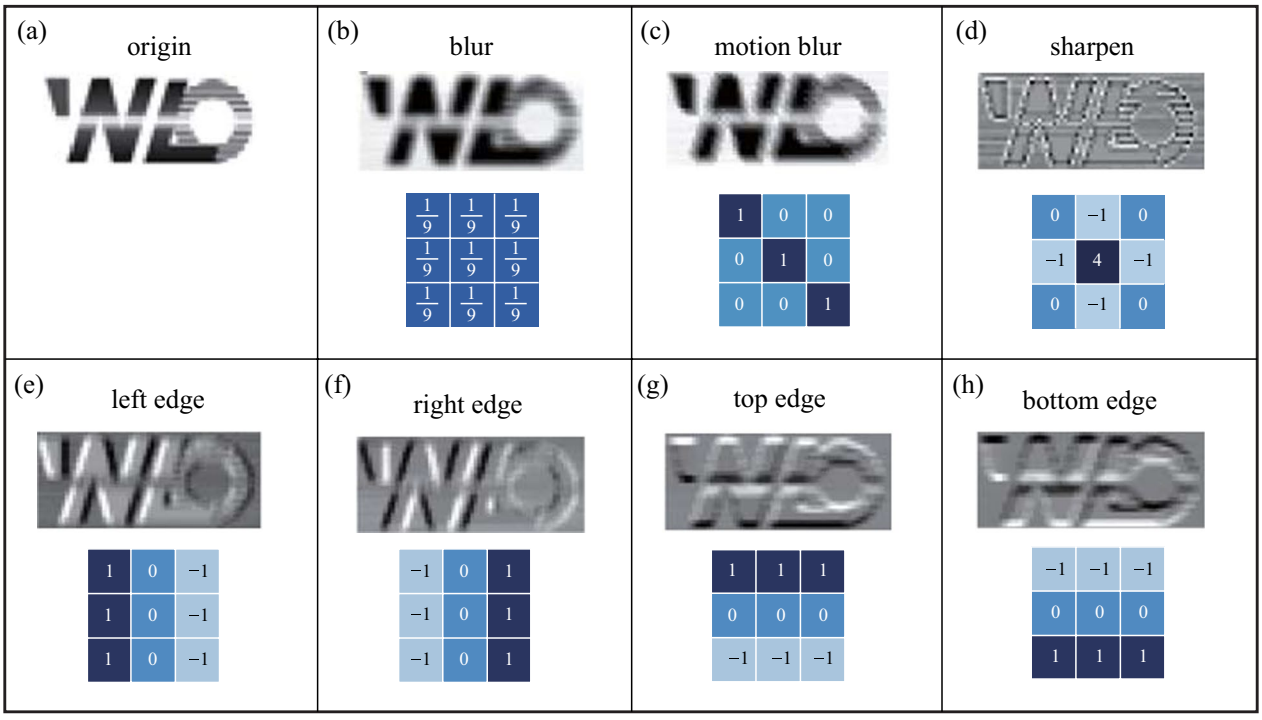

Fig. 7 Experimental results of image convolution. **a** Original image, which is the logo of Wuhan National Laboratory for Optoelectronics (WNLO). **b–h** Convolved images and the corresponding transmission matrices of 3×3 kernels. The kernels are designed to perform different image processing functions or highlight different edges of the original image

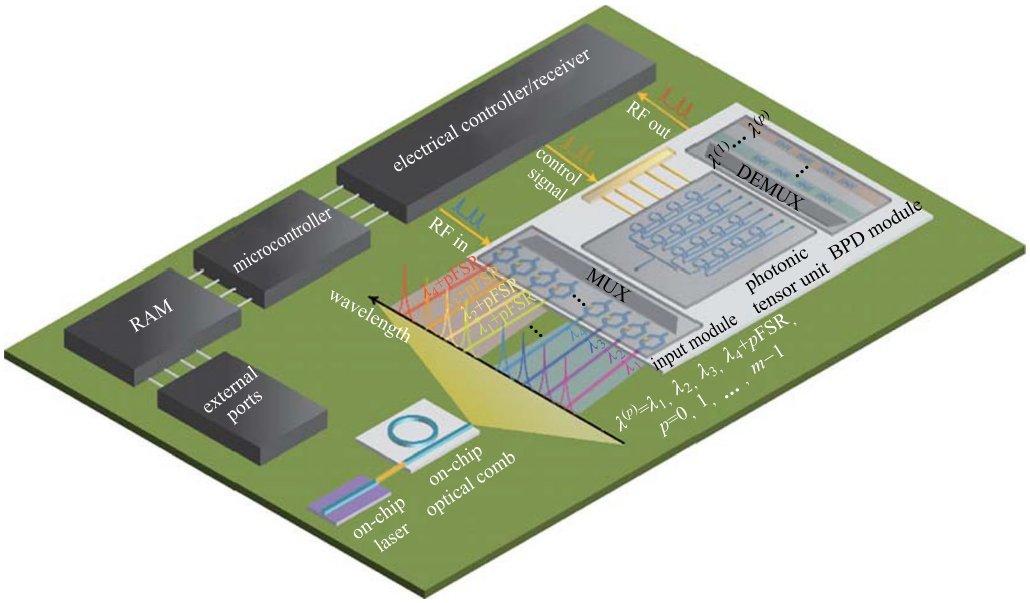

Fig. 8 Highly integrated on-chip scheme for optical parallel computation. There are m sets of different input vectors provided by multiwavelength light source (e.g., on-chip optical comb). Input signals are modulated in different wavelengths by the IMs, then multiplexed as the input of MRR array via wavelength division multiplexers (MUXs). The output powers of each row in MRR array are divided by the wavelength division demultiplexers (DEMUXs) and separately detected by m sets of corresponding photodiodes. Each set of wavelengths is used for one input vector. The electrical controller/receiver are driven by microcontroller equipped with RAM and external ports

modulated by IMs of the input module. With this, the experimental setup is greatly simplified. The thermally tuned MRRs can be replaced by electrically tuned ones, which might improve the response rate by several orders of magnitude. As for electrical control, the electrical controller/receiver, together with microcontroller, random access memory (RAM), and external ports are applied to improve system response rate.

5 Conclusion

In conclusion, we have demonstrated a small MRR array that performs large complex MVM. Through matrix decomposition and partition, we have also optimized the photonic complex-MVM core so that it can perform larger complex MVM and extended its matrix computation to (1) real number, (2) complex number, and (3) higher processing dimensions. We have fabricated the integrated photonic complex-MVM core on an SOI platform, which is compact and compatible with CMOS technology. With a small MRR array, the 4×4 matrix computation system can be scaled up to 8×8, 16×16, or even larger operation dimensions in complex field with traditional incoherent computing. The processor was then applied for WHT, DCT and DFT signal transformations. Image processing with 7 types of convolutional kernels is also experimentally demonstrated. Our proposed system shows adequate performance in various applications. The processing capacity of this matrix–vector multiplier can be further enhanced by enabling parallel WDM computation and full integration with on-chip laser sources and electrical microcontrollers in the future.

Appendix

A. Calibration of MRR array

Since the MRR is a resonant device, the transmittance of the through and drop ports depends on the difference between the laser and resonance wavelength of the MRR. Therefore, the four laser wavelengths need to be calibrated at the resonance peak of the corresponding MRR prior to experimentation. Figure 9a shows the state in which the laser wavelength is not aligned with the resonant peak of the MRR. In this case, the transmission coefficient of the MRR is $x_{ij} = 1$. As shown in Fig. 9b, the voltage values of the four MRRs were changed so that the four laser wavelengths coincide with the resonant peak of the MRRs, where the transmission coefficients were all $x_{ij} =$

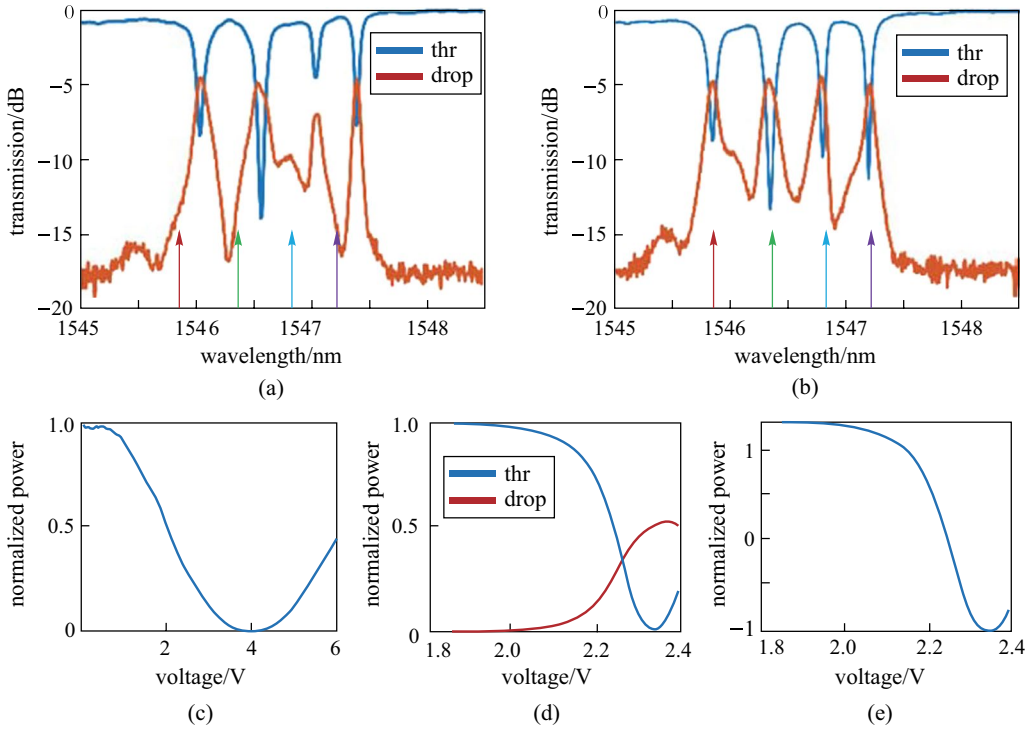

Fig. 9 Calibration of MRR array. **a** and **b** are the spectra before and after laser wavelength calibration respectively, and the four laser wavelengths are represented by four colors (red, green, blue, purple) respectively. **c** Voltage-input relational table. **d** Normalized optical power of the through port (red line) and drop port (blue line) of a certain MRR. **e** Normalized optical power curve after difference calculation

1. The calibration of the ring array is between these two states. Figure 9c shows the normalized power detected at the through port when the MRR is fixed in the all-pass state (i.e., the transmission coefficient of the MRR is 1) and the voltage applied on the IM is changed. The voltage-input relational table was obtained by choosing a fixed step length of 20 mV, applying 300 V steps to the IM, and measuring the corresponding output power. When a particular input needs to be loaded, the computer applies table look-up and loads the corresponding voltage into the IM. Similarly, the table look-up method is used in MRR calibration. First, the corresponding input is set at the maximum value of 1 and the voltages are selected according to a fixed step size between $x_{ij} = -1$ and $x_{ij} = 1$. Then, the voltages are applied to the MRR array and the output powers of MRR are measured. The voltage-transmission relational table was obtained and shown in Fig. 9d and e. When a particular transfer coefficient need to be loaded, the computer looks up the nearest value in the table using the look-up table method and loads the corresponding voltage onto the MRR electrodes.

B Experimental verification of matrix–vector multiplier

Figure 10a presents a sample experimental transmission function of $\boldsymbol{X}$, Fig. 10b lists the corresponding theoretical results, and Fig. 10c summarizes the vector data results. Each data point represents a dot product of one of the row vectors of $\boldsymbol{X}$ and the input vector. The blue line represents the experimental results and the red line represents the deviation of each experimental point. The error statistics are calculated and shown in Fig. 10d, where most of the absolute values of the errors fall within the range of 0–0.1.

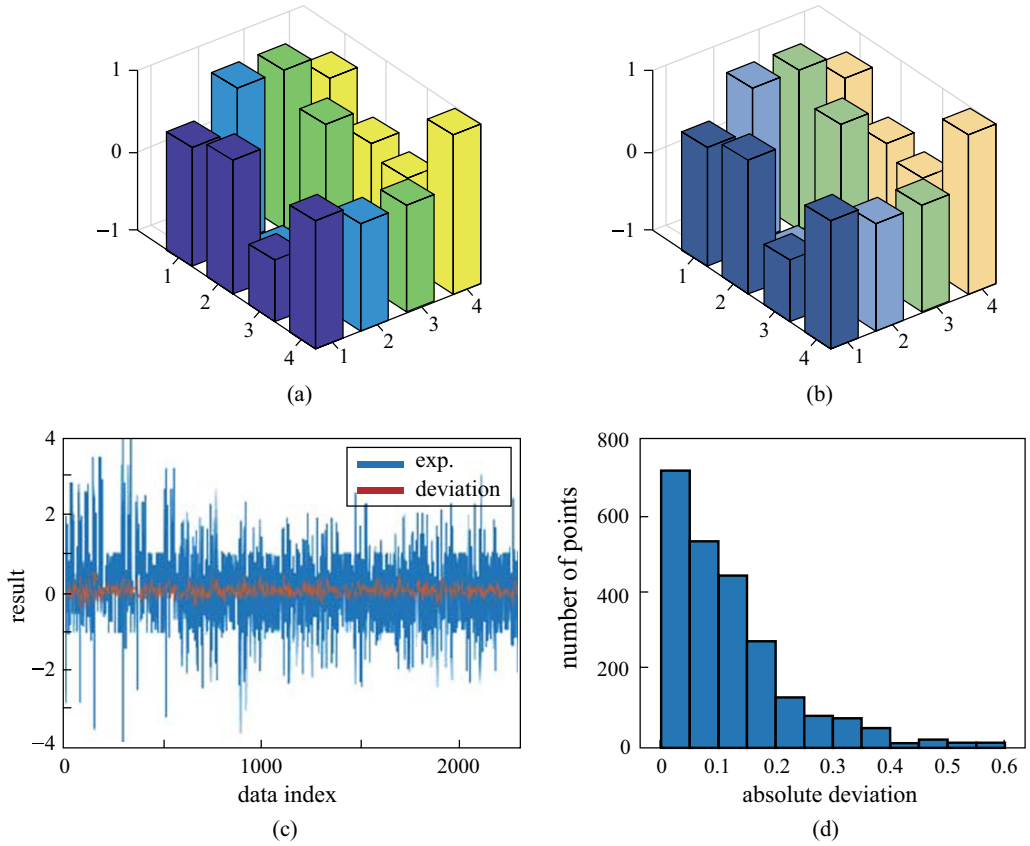

Fig. 10 Experimental results of matrix–vector multiplication. **a** Sample experimental transmission function of $\boldsymbol{X}$ and **b** its corresponding theoretical results. **c** Result vector data and **d** error distribution

Acknowledgements This work was partially supported by the National Key Research and Development Project of China (No. 2018YFB2201901), the National Natural Science Foundation of China (Grant Nos. 61805090 and 62075075), Shenzhen Science and Technology Innovation Commission (No. SGDX2019081623060558), and Research Grants Council of Hong Kong SAR (No. PolyU152241/18E).

Authors' contributions The authors read and approved the final manuscript.

Declarations

Competing interests The authors declare that they have no competing interests.

References

1. Shi, W., Caballero, J., Huszar, F., Totz, J., Aitken, A.P., Bishop, R., Rueckert, D., Wang, Z.: Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. Proc. CVPR 1874–1883 (2016)
2. Li, X., Zhang, G., Huang, H.H., Wang, Z., Zheng, W.: Performance analysis of GPU-based convolutional neural networks. Proc. ICPP 67–76 (2016)
3. Li, H., Lin, Z., Shen, X., Brandt, J., Hua, G.: A convolutional neural network cascade for face detection. Proc. CVPR 5325–5334 (2015)
4. Kitayama, K.I., Notomi, M., Naruse, M., Inoue, K., Kawakami, S., Uchida, A.: Novel frontier of photonics for data processing—photonic accelerator. APL Photonics **4**(9), 090901 (2019)
5. Krizhevsky, A., Sutskever, I., Hinton, G.E.: ImageNet classification with deep convolutional neural networks. Commun. ACM **60**(6), 84–90 (2017)
6. LeCun, Y., Bengio, Y., Hinton, G.: Deep learning. Nature **521**(7553), 436–444 (2015)
7. Xu, X., Tan, M., Corcoran, B., Wu, J., Boes, A., Nguyen, T.G., Chu, S.T., Little, B.E., Hicks, D.G., Morandotti, R., Mitchell, A., Moss, D.J.: 11 TOPS photonic convolutional accelerator for optical neural networks. Nature **589**(7840), 44–51 (2021)
8. Wu, C., Yu, H., Lee, S., Peng, R., Takeuchi, I., Li, M.: Programmable phase-change metasurfaces on waveguides for multimode photonic convolutional neural network. Nat. Commun. **12**(1), 96 (2021)

9. Feldmann, J., Youngblood, N., Karpov, M., Gehring, H., Li, X., Stappers, M., Le Gallo, M., Fu, X., Lukashchuk, A., Raja, A.S., Liu, J., Wright, C.D., Sebastian, A., Kippenberg, T.J., Pernice, W.H.P., Bhaskaran, H.: Parallel convolutional processing using an integrated photonic tensor core. Nature **589**(7840), 52–58 (2021)
10. Ríos, C., Youngblood, N., Cheng, Z., Le Gallo, M., Pernice, W.H.P., Wright, C.D., Sebastian, A., Bhaskaran, H.: In-memory computing on a photonic platform. Sci. Adv. **5**(2), 5759 (2019)
11. Feldmann, J., Youngblood, N., Wright, C.D., Bhaskaran, H., Pernice, W.H.P.: All-optical spiking neurosynaptic networks with self-learning capabilities. Nature **569**(7755), 208–214 (2019)
12. Lin, X., Rivenson, Y., Yardimci, N.T., Veli, M., Luo, Y., Jarrahi, M., Ozcan, A.: All-optical machine learning using diffractive deep neural networks. Science **361**(6406), 1004–1008 (2018)
13. Zhou, T., Lin, X., Wu, J., Chen, Y., Xie, H., Li, Y., Fan, J., Wu, H., Fang, L., Dai, Q.: Large-scale neuromorphic optoelectronic computing with a reconfigurable diffractive processing unit. Nat. Photonics **15**(5), 367–373 (2021)
14. Zhu, W., Zhang, L., Lu, Y., Zhou, P., Yang, L.: Design and experimental verification for optical module of optical vector-matrix multiplier. Appl. Opt. **52**(18), 4412–4418 (2013)
15. Habiby, S.F., Collins Jr, S.A.: Implementation of a fast digital optical matrix-vector multiplier using a holographic look-up table and residue arithmetic. Appl. Opt. **26**(21), 4639–4652 (1987)
16. Bocker, R.P., Clayton, S.R., Bromley, K.: Electrooptical matrix multiplication using the twos complement arithmetic for improved accuracy. Appl. Opt. **22**(13), 2019 (1983)
17. 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), 1–3 (1978)
18. Hong, J., Yeh, P.: Photorefractive parallel matrix-matrix multiplier. Opt. Lett. **16**(17), 1343–1345 (1991)
19. Cartwright, S.: New optical matrix-vector multiplier. Appl. Opt. **23**(11), 1683–1684 (1984)
20. Athale, R.A., Collins, W.C.: Optical matrix-matrix multiplier based on outer product decomposition. Appl. Opt. **21**(12), 2089–2090 (1982)
21. Mukhopadhyay, S., Das, D.N., Das, P.P., Ghosh, P.: Implementation of all-optical digital matrix multiplication scheme with nonlinear material. Opt. Eng. (Redondo Beach, Calif.) **40**(9), 1998–2002 (2001)
22. Liu, B., Liu, L.R., Shao, L., Chen, H.Q.: Matrix–vector multiplication in a photorefractive crystal. Opt. Commun. **146**(1–6), 34–38 (1998)
23. Gu, C., Campbell, S., Yeh, P.: Matrix–matrix multiplication by using grating degeneracy in photorefractive media. Opt. Lett. **18**(2), 146–148 (1993)
24. Nitta, T.: Orthogonality of decision boundaries in complex-valued neural networks. Neural Comput. **16**(1), 73–97 (2004)
25. Zhou, H., Zhao, Y., Xu, G., Wang, X., Tan, Z., Dong, J., Zhang, X.: Chip-scale optical matrix computation for pagerank algorithm. IEEE J. Sel. Top. Quantum Electron. **26**(2), 1–10 (2020)
26. Bogaerts, W., Pérez, D., Capmany, J., Miller, D.A.B., Poon, J., Englund, D., Morichetti, F., Melloni, A.: Programmable photonic circuits. Nature **586**(7828), 207–216 (2020)
27. Clements, W.R., Humphreys, P.C., Metcalf, B.J., Kolthammer, W.S., Walsmley, I.A.: Optimal design for universal multiport interferometers. Optica **3**(12), 1460–1465 (2016)
28. Miller, D.A.B.: Self-configuring universal linear optical component. Photonics Res. **1**(1), 1–15 (2013)
29. Mennea, P.L., Clements, W.R., Smith, D.H., Gates, J.C., Metcalf, B.J., Bannerman, R.H.S., Burgwal, R., Renema, J.J., Kolthammer, W.S., Walmsley, I.A., Smith, P.G.R.: Modular linear optical circuits. Optica **5**(9), 1087–1090 (2018)
30. Carolan, J., Harrold, C., Sparrow, C., Martín-López, E., Russell, N.J., Silverstone, J.W., Shadbolt, P.J., Matsuda, N., Oguma, M., Itoh, M., Marshall, G.D., Thompson, M.G., Matthews, J.C.F., Hashimoto, T., O'Brien, J.L., Laing, A.: Universal linear optics. Science **349**(6249), 711–716 (2015)
31. Zhou, H., Zhao, Y., Wang, X., Gao, D., Dong, J., Zhang, X.: Self-configuring and reconfigurable silicon photonic signal processor. ACS Photonics **7**(3), 792–799 (2020)
32. Annoni, A., Guglielmi, E., Carminati, M., Ferrari, G., Sampietro, M., Miller, D.A.B., Melloni, A., Morichetti, F.: Unscrambling light-automatically undoing strong mixing between modes. Light Sci Appl. **6**(12), e17110 (2017)
33. Zhou, H., Zhao, Y., Wei, Y., Li, F., Dong, J., Zhang, X.: All-in-one silicon photonic polarization processor. Nanophotonics **8**(12), 2257–2267 (2019)
34. Shen, Y., Harris, N.C., Skirlo, S., Prabhu, M., Baehr-Jones, T., Hochberg, M., Sun, X., Zhao, S., Larochelle, H., Englund, D., Soljačić, M.: Deep learning with coherent nanophotonic circuits. Nat. Photonics **11**(7), 441–446 (2017)
35. Tait, A.N., de Lima, T.F., Zhou, E., Wu, A.X., Nahmias, M.A., Shastri, B.J., Prucnal, P.R.: Neuromorphic photonic networks using silicon photonic weight banks. Sci. Rep. **7**(1), 7430 (2017)
36. Yang, L., Zhang, L., Ji, R.: On-chip optical matrix-vector multiplier. Optics and Photonics for Information Processing Vii (2013)
37. Miscuglio, M., Sorger, V.J.: Photonic tensor cores for machine learning. Appl. Phys. Rev. **7**(3), 031404 (2020)
38. Zhao, Y., Wang, X., Gao, D., Dong, J., Zhang, X.: On-chip programmable pulse processor employing cascaded MZI-MRR structure. Front. Optoelectron. **12**(2), 148–156 (2019)
39. Roy, A.B., Dey, D., Mohanty, B., Banerjee, D.: Comparison of FFT, DCT, DWT, WHT compression techniques on electrocardiogram and photoplethysmography signals. IJCA Special Issue on International Conference on Computing, Communication and Sensor Network CCSN, 2012. 6–11
40. Rahardja, S., Ser, W., Lin, Z.N.: UCHT-based complex sequences for asynchronous CDMA system. IEEE Trans. Commun. **51**(4), 618–626 (2003)
41. Andrushia, A.D., Thangarjan, R.: Saliency-based image compression using walsh–hadamard transform (WHT), pp. 21–42. Springer, Biologically rationalized computing techniques for image processing applications (2018)
42. Strang, G.: The discrete cosine transform. SIAM Rev. **41**(1), 135–147 (1999)
43. Oppenheim A.V., Schafer, R. W., Buck, J. R.: Discrete-TimeSignal Processing. Norwood: Pearson Education India (1999)

Junwei Cheng is currently a Ph.D. candidate in Huazhong University of Science and Technology (HUST), China. He received his B.Eng. degree from HUST in 2019. His current research interests include silicon photonics and photonic neuromorphic computing.

Yuhe Zhao received her Ph.D. degree in Optical Engineering from Huazhong University of Science and Technology, China in 2021. Her research interests are optoelectronic devices and integration, microwave photonics and arbitrary waveform generation.

Wenkai Zhang received his Bachelor's degree from Huazhong University of Science and Technology (HUST), China in 2021. Then he joined Wuhan National Laboratory for Optoelectronics at HUST as a Ph.D. candidate. His research interests include photonic integrated circuits and neuromorphic photonics.

Hailong Zhou received his B.S. degree from Huazhong University of Science and Technology (HUST), China in 2012/06. From 2012/09 – 2017/06, he studied in Wuhan National Laboratory for Optoelectronics at HUST as a doctoral candidate and received his Ph.D. degree in 2017/06. Currently, he is a post-doctor in Wuhan National Laboratory for Optoelectronics at HUST, China. His research interests are silicon photonics and photonic accelerators for artificial intelligence.

Dongmei Huang received her B.S. degree in 2014 from Huazhong University of Science and Technology, China, obtained her M.S. degree in 2017 from Chongqing University, China, and obtained her Ph.D. degree in 2020 from the Hong Kong Polytechnic University, China. She is currently a Research Assistant Professor at Photonics Research Centre of The Hong Kong Polytechnic University. Her research interests include wavelength swept lasers and its applications in optical coherence tomography and optical sensing systems, nonlinear microresonators.

Qing Zhu Senior Engineer from Huawei Technologies Co., LTD., who received her Ph.D. degree in Condensed Matter Physics from Institute of Physics, University of the Chinese Academy of Sciences, China, in 2020, and her bachelor's degree in Physics from Jilin University, China, in 2014. After finishing her PhD, she has been working in Institute of Strategic Research of Huawei Technologies Co., LTD. Her recent research interests include optical computing, silicon modulator and integrated photonics.

Yuhao Guo received the Ph.D. degree from Tianjin University, China, in 2020. Since 2021, he has been a senior engineer with Institute of Strategic Research in Huawei technologies Co. LTD. His research interests include integrated nanophotonics, chip-scale optical interconnects, metasurface, and optical computing. He has authored or co-authored 27 peer-reviewed articles and he has 3 patents issued.

Bo Xu Senior Engineer from Huawei Technologies Co., LTD., who received her Ph.D. degree in Optics from Shanghai Institute of Optics and Fine Mechanics, University of the Chinese Academy of Sciences, China, in 2020, and her bachelor's degree in Optoelectronic Technology and Science from Nankai University, Tianjin, China, in 2015. After finishing her Ph.D., she has been worked in Institute of Strategic Research of Huawei Technologies Co., LTD. Her recent research interests include photonic chip and optical computing.

Jianji Dong is Professor of Wuhan National Laboratory for Optoelectronics (WNLO), Huazhong University of Science and Technology (HUST), China. He received his Ph.D. degree in Optical Engineering from HUST in 2008. After that, he worked as postdoc at Cambridge University, UK till 2010. From March 2010, he returned to HUST and was promoted as a full professor in 2013. His research interests include integrated microwave photonics, silicon photonics, and photonic computing. He has published more than 100 Journal papers, including Nature Communications, Light science and applications, Physical Review Letters, etc. He has some special contribution to energy-efficient graphene silicon microheater, programmable temporal cloak, and complex spectrum analyzer of orbital angular momentum mode. He was honored First award of Natural Science of Hubei Province. He is the editorial member of Scientific Reports, associate editor of IET Optoelectronics, and executive editor-in-chief of Frontiers of Optoelectronics. He is an IEEE Senior Member and OSA member.

Xinliang Zhang received his Ph.D. degree in Physical Electronics from Huazhong University of Science and Technology (HUST), China in 2001. He is currently with Wuhan National Laboratory for Optoelectronics and School of Optical and Electronic Information, HUST, as a Professor. He is the author or coauthor of more than 300 journal and conference papers. His current research interests include InP-based and Si-based devices and integration for optical network, high-performance computing and ultrafast optical measurements. In 2016, he was elected as OSA Fellow.