PHOTONICS Research

Optical multi-imaging–casting accelerator for fully parallel universal convolution computing

Guoqing Ma,[1,2] Junjie Yu,[1,2,3] Rongwei Zhu,[1,2] and Changhe Zhou[1,2,*]

[1]Laboratory of Information Optics and Optoelectronic Technology, Shanghai Institute of Optics and Fine Mechanics, Chinese Academy of Sciences, Shanghai 201800, China
[2]Center of Materials Science and Optoelectronics Engineering, University of Chinese Academy of Sciences, Beijing 100049, China
[3]e-mail: Junjiey@siom.ac.cn
*Corresponding author: chazhou@mail.shcnc.ac.cn

Received 8 August 2022; revised 1 December 2022; accepted 20 December 2022; posted 23 December 2022 (Doc. ID 472741); published 1 February 2023

Recently, optical computing has emerged as a potential solution to computationally heavy convolution, aiming at accelerating various large science and engineering tasks. Based on optical multi-imaging–casting architecture, we propose a paradigm for a universal optical convolutional accelerator with truly massive parallelism and high precision. A two-dimensional Dammann grating is the key element for generating multiple displaced images of the kernel, which is the core process for kernel sliding on the convolved matrix in optical convolutional architecture. Our experimental results indicate that the computing accuracy is typically about 8 bits, and this accuracy could be improved further if high-contrast modulators are used. Moreover, a hybrid analog–digital coding method is demonstrated to improve computing accuracy. Additionally, a convolutional neural network for the standard MNIST dataset is demonstrated, with recognition accuracy for inference reaching 97.3%. Since this architecture could function under incoherent light illumination, this scheme will provide opportunities for handling white-light images directly from lenses without photoelectric conversion, in addition to convolutional accelerators.

https://doi.org/10.1364/PRJ.472741

1. INTRODUCTION

A convolutional neural network (CNN), as "convolutional" implies, involves extensive convolution operations among neighboring layers, followed by batch normalization and nonlinear activation for the expected performance [1–3]. Remarkably, these massive linear matrix multiply–accumulate (MAC) operations account for more than 80% of the total number of deep neural network (DNN) calculations [4]. However, the convolution operation, which is unsuitable for modern advanced electric serial processors, is becoming the biggest burden for high-performance computing tasks, particularly for artificial intelligence (AI) algorithms. Furthermore, as the scale of the matrix increases, so does the computational overhead of convolution operations. It has been demonstrated that the amount of computing power required to train state-of-the-art DNNs doubles every 3.5 months [5], far exceeding that of traditional electrical integrated circuits (EICs) following Moore's law. Although parallel electrical coprocessors such as graphics processing units (GPUs) and tensor processing units (TPUs) can accelerate the convolution calculation, it is still difficult to handle millions of MAC operations in a fully parallel manner for DNNs practically [6,7]. In contrast, it has been proven that many MAC operations can be executed concurrently during a single pass of light, and this may be the prime motivation for the recent interest in optical computing [8,9]. Photonic solutions for computing have been investigated for at least 70 years [10,11]. However, compared with fast-growing EICs, the development of optical computing gradually slowed in the late 2000s [12], owing to a lack of application-driven motivation and adequate optical computing architectures.

Recently, due to the remarkable achievements in AI, there has been renewed interest in attempting to improve computing power, energy efficiency, and processing speed by exploiting photonic or hybrid optical–electric processors rather than their electronic counterparts [13–15]. Two mainstream optical computing architectures have been rapidly developed. The first is based on a planar waveguide on a two-dimensional (2D) substrate [16–18], whereas the second is realized by multiple cascading diffractive optical elements (DOEs) in three-dimensional (3D) space [19,20]. However, planar architecture, which includes Mach–Zehnder interferometers [16], microring resonators [21,22], waveguide modulators [23], and acousto-optical modulators [24], does not fully use the 3D interconnectivity of optics, whereas 3D architecture requires full manipulation of the electromagnetic field with high precision,

and fabricating large-sized and high-precision subwavelength DOEs in 3D space will still be difficult [19,20].

Despite predictions that photonic processors could be at least 10,000 times faster than state-of-the-art EICs [13,14], the past schemes have not realized fully parallel convolution computing compared with their electronic counterparts, particularly when high precision is required. Here, we propose a new paradigm for a universal convolutional accelerator with full parallelism and adequate precision based on optical multi-imaging–casting architecture (OMica), capable of calculating arbitrarily encoded hybrid analog–digital matrix convolutions. The architecture can be viewed as the starting point for a new roadmap for optical computing, with the potential for building fully massively parallelized optical convolutional accelerators to overcome the intrinsic computing power shortage and unsatisfactory energy efficiency of EICs. Furthermore, the incoherent illumination implies the possibility of handling white-light images directly from lenses without traditional photoelectric conversion, promising to fully exploit the benefits of AI algorithms or accelerate other practical applications where rapid big data processing is desired.

2. PRINCIPLE OF OMICA

A. Optical Multi-Imaging–Casting Architecture

The OMica architecture, as depicted in Fig. 1, employs an incident-modulated light (matrix A) and a spatial light modulator (SLM) (matrix B), as well as a confocal $4f$ system with a diffractive beam splitter (BS), and another focusing system with a photodetector (matrix C). The planes of matrices A and B, the confocal plane of the $4f$ system, and the plane of the detector are all in a conjugated object–image relationship with each other. When a BS, such as a Dammann grating (DG) [25–27], is placed behind the plane of matrix A, the two pairs of imaging–casting relationships mentioned above still hold. When the DG is inserted, the optical signal carrying the information of matrix A is duplicated into multiple diffraction orders, with excellent uniformity due to the properties of DG. The different diffraction orders inherently have different angular spectral components (θ_1 and θ_2). However, they all carry the same information as matrix A, as shown in Fig. 1(c). This implies that the multiplexing of matrix A is achieved over the spatial pattern. When we pass a pinhole through one of the diffraction orders in the confocal plane, the image corresponding to that diffraction order can be seen clearly on the plane of matrix B through lens L_2 (as shown in Appendix A and Fig. 9). Because these diffraction orders have different diffraction angles (θ_1 and θ_2), the images of the diffraction orders on the plane of matrix B are displaced when we sequentially pass each diffraction order through the pinhole. Thus, as shown in Fig. 1(c), all images are aligned by adjusting the distance d between the DG and matrix A, according to a paraxial relation:

$$s = l\frac{f_1}{f_2}\tan\theta, \tag{1}$$

where s is a convolutional stride, f_1 and f_2 are focal lengths of L_1 and L_2, respectively, and θ is the angle difference between any two adjacent diffraction orders. According to

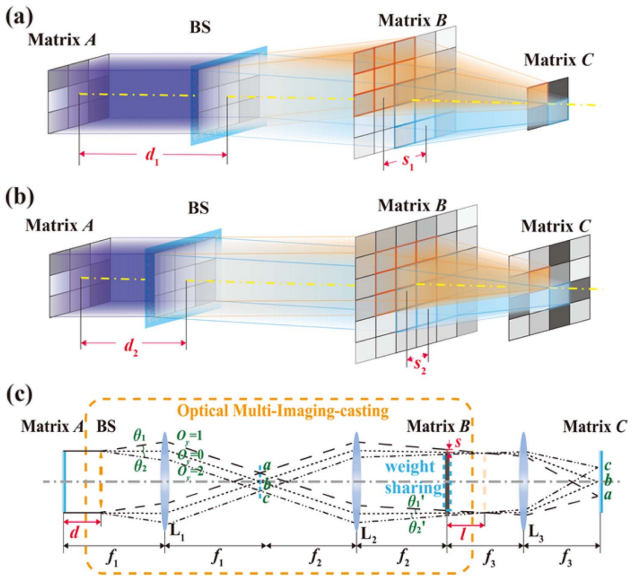

Fig. 1. Schematic of the optical multi-imaging–casting architecture: optical parallel convolution process with different convolutional strides s_1 (a) and s_2 (b); (c) optical architecture principle of OMica, where the beam splitter (BS) is a diffractive beam splitter; O_y is diffraction order in the y direction (indicated by different line types), and θ is the angle difference between any two adjacent diffraction orders in object space (θ_1 and θ_2 are diffraction angles of $O_y = 1$ and $O_y = 2$ diffraction orders, respectively); θ' is the angle difference in image space (θ_1' and θ_2' are diffraction angles of $O_y = 1$ and $O_y = 2$ diffraction orders, respectively); d is the distance between matrix A and BS, and l is the distance between matrix B and the image of BS. a, b, and c are spot arrays corresponding to different diffraction orders diffracted from a BS. The imaging–casting system is composed of L_1 and L_2, with focal lengths f_1 and f_2. L_3 is a focusing lens with focal length f_3. s is the lateral shifts of the image of diffraction orders of DG on the SLM_2 plane corresponding to the convolutional stride, and this convolutional stride could be tunable by changing the distance d [s_1 and s_2 correspond to different convolutional strides shown in (a) and (b)].

the grating equation, $\theta_m = \arcsin(m\lambda/\Lambda)$, $\theta_{m+1} - \theta_{m-1} \approx \theta$, and $\tan\theta \approx \sin\theta$, Eq. (1) can be re-written as

$$s = d\frac{f_2}{f_1}\frac{\lambda}{\Lambda}, \tag{2}$$

where θ_m is the diffraction angle of the mth order of a DG, Λ is the grating period, and λ is wavelength. Therefore, s can also be adjusted to adapt to different convolutional strides by changing d [Figs. 1(a) and 1(b)].

Because of the conjugation relationship and different angles, the images of all diffraction orders are superimposed on the matrix B plane with naturally shifted displacements when the pinhole is removed. This means that the SLM can modulate these shifted images simultaneously. That is, all multiplications of multiple images of matrix A and matrix B can be implemented in parallel. These multiplications are then summed through L_3 and separated from each other in the C plane due to the angular spectrum differences. Therefore, the convolution of the two matrices can be performed in parallel after the light passes through the system once. This process is a perfect

optical implementation of mathematical convolution, i.e., $C = A \otimes B$, where "$\otimes$" is the convolution operator. Owing to the object–image conjugate configuration, the OMica proposed here avoids the size trade-off of elements in the matrix between spatial and frequency domains in the $4f$ optical convolutional system [28,29], allowing massive parallelism with sufficiently high accuracy to be realized. Moreover, because of the object–image conjugate configuration, the OMica can work under both coherent and incoherent light illumination. Thus, this optical hardware allows it to handle white-light images directly from lenses without traditional photoelectric conversion if achromatic lenses are used as the projection system.

B. Negative Matrix Coding Method

In our proof-of-concept implementation, a homemade 2D 28×20 DG (see details in Appendix B) was inserted into a $4f$ system. Two amplitude-only SLMs (8-bit grayscale) are located on the object and image planes of the $4f$ system, where the two convolution matrices are loaded sequentially. In the experiment, light intensity was used as the information carrier, and the two SLMs were used to load the information of matrix B and matrix A into the incident uniform light beam. Therefore, in principle, only nonnegative matrices can be loaded and calculated based on this hardware. To address this limitation, a negative matrix encoding method for hybrid analog–digital optical convolution computing was developed. In a hybrid analog–digital framework, a grayscale matrix with negative elements can be easily decomposed into one larger-scale or several same-size negabinary digit (NBD) matrices in spatial or temporal sequences, respectively [30,31]. In other words, each decimal element in the original matrix can be converted into NBD representation as follows:

$$(a)_{10} = \sum_{i=0}^{\lceil N/k \rceil} c_i(-2^k)^i, \tag{3}$$

where $\{c_{\lceil N/k \rceil}, c_{\lceil N/k \rceil - 1}, \ldots, c_0\}$ NBD is called c_i bits, with $c_i \in [0, 2^k - 1]$; N is maximum bits of NBD, k is an integer, and the operator "$\lceil \cdot \rceil$" indicates rounding the number to the nearest integer greater than it. Following this decomposition, a grayscale matrix with negative elements is transformed into a larger matrix spatially or several same-sized matrices in temporal series represented by $\lceil N/k \rceil$ nonnegative bits, allowing these matrices to be loaded directly on the SLMs. The principle of this encoding method is depicted schematically in Fig. 2. Notably, there is a trade-off between computing precision and computing power, which can be adjusted by varying parameter k. A small k indicates that high precision with low computing power will be achieved, whereas a large k indicates high computing power with relatively low precision. Therefore, this encoding method can improve computing precision to the same extent compared with pure-analog optical convolution computing [30,31].

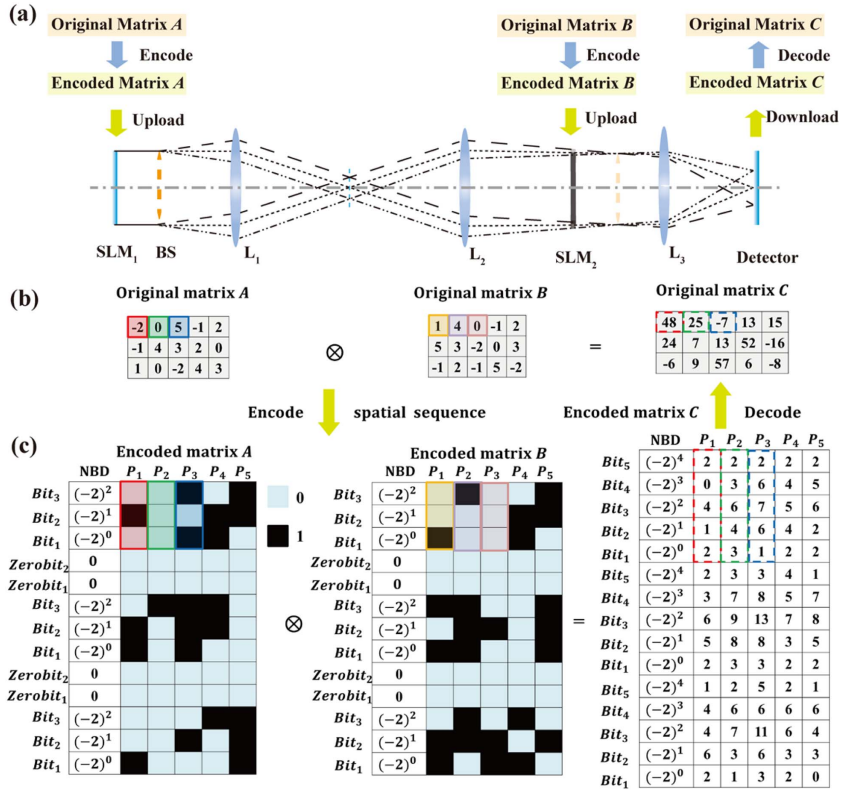

Fig. 2. Procedure of converting the original grayscale matrix with negative elements into encoded matrices of NBD. (a) The encoding matrices are loaded into the OMica system to compute the convolution, with the experimental encoded convolutional result decoded into the original matrix. (b) Original grayscale matrices A and B, and original convolutional results matrix C. (c) Larger encoded matrices A and B in spatial sequence and the same size encoded convolutional results matrix C.

Here, as an example, under the condition of $k = 1$, the encoding process of a grayscale matrix with negative elements ranging from −2 to 5 is demonstrated step by step. As shown in Figs. 2(b) and 2(c), the grayscale number for each element of the original matrix to be encoded is expressed in multiple NBDs after encoding. For example, the first element in the original matrix A is written as $-2 = 0 \times (-2)^2 + 1 \times (-2)^1 + 0 \times (-2)^0$. Therefore, the elements of the matrix are arranged in rows after encoding, denoted as P_1, P_2, and P_3. Each element in the column direction is encoded with three NBDs, denoted as Bit_3, Bit_2, and Bit_1, as shown in Fig. 2(c). Thus, the first element, −2, is expressed as {010} in the first column of the encoded matrix, that is, $c_2 = 0$, $c_1 = 1$, and $c_0 = 0$. Subsequently, the converted matrices are loaded onto the SLMs in spatial sequence for computing [Fig. 2(c)]. Notably, to avoid aliasing in a spatial sequence, some zero elements should be inserted into the encoded matrix between two adjacent rows or columns of the original high-bit matrix, where the number of zero elements is $\lceil N/k \rceil - 1$. Here, the physical pixels of the SLMs will not be fully used because of the redundant zero elements. The computational advantage can be realized only by increasing the matrix scale, but doing so will significantly slow down the system's refresh rate because the convolution must be performed among all bits of either matrix A or B. Therefore, when the OMica is used for computing acceleration, a compromise should be struck between high computing power and high computing precision by choosing an appropriate parameter k.

3. EXPERIMENTAL RESULTS

A. Hybrid Analog–Digital Matrix Convolution

As an example, the hybrid analog–digital optical convolution of two randomly generated 2-bit grayscale 3×10 matrices, A_1 and B_1, with elements in the range of 0 to 3, and two negabinary 3-bit grayscale 2×10 matrices, A_2 and B_2, with negative

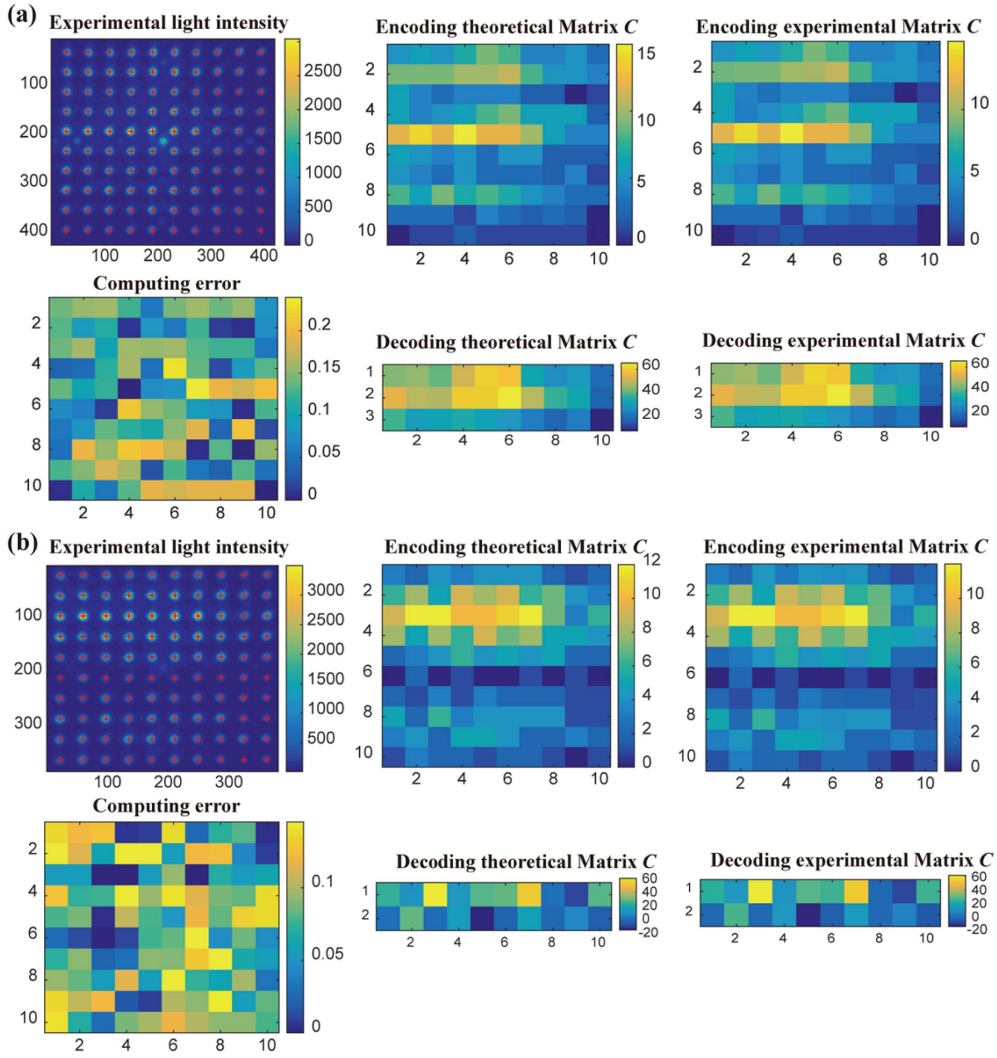

Fig. 3. Experimental results of hybrid analog–digital matrix convolution for two groups of matrices based on spatial sequence encoding. The subfigures from left to right are the light intensity distribution of the spot array denoting the convolution, theoretical convolutional values, experimental convolutional results, error map between theoretical and experimental results, and decoded convolutional results, respectively, in (a) matrices A_1 and B_1 and (b) matrices A_2 and B_2. The red cross marks the centroid positions of each spot.

elements in the range of −2 to 5, is demonstrated, and the convolutional results are shown in Fig. 3. In each box, the light intensity distributions of the spot arrays on the detected plane, denoting the raw results of convolution, are shown in the first subfigure of the first row. The theoretical results obtained by an electric computer (full precision, 64 bit) are illuminated in the second subfigure, and the experimental results before decoding are shown in the third subfigure. The absolute error map is shown in the first subfigure of the second row, which is defined as follows:

$$\mathrm{AE} = |C_{\mathrm{theo}} - C_{\mathrm{exp}}|, \tag{4}$$

where C_{theo} and C_{exp} are the theoretical and experimental convolutional results, respectively. "$|\cdot|$" denotes the absolute operation. Additionally, the theoretical and experimental results of the convolution after decoding are shown in the second and third subfigures in the second row, respectively. It is demonstrated that the overall trend of the experimental and theoretical results of the convolution is consistent.

Figures 3(a) and 3(b) show the results of the convolution of two matrices, A_1, A_2 and B_1, B_2, respectively. The mean values of the absolute errors AE are 0.114 and 0.08, and it is seen that the maximum values are approximately 0.239 and 0.145, respectively, before decoding, indicating that the optical convolutional architecture achieves high precision. It should be noted that the former has a higher mean error before decoding than the latter, owing to increased cross talk caused by relatively large convolutional elements. Moreover, the two encoded matrices in spatial coding methods are filled with zero elements to avoid aliasing, which further reduces the cross talk and final error. Because the maximum absolute errors for the two cases are all less than 0.5, the correct convolutional results, with 100% accuracy, can still be obtained after digitalization. Thus, the experimental light intensity distribution of the two cases precisely reflects the values of the convolutional results.

B. High-Accuracy Matrix Convolution

As an example, the high-accuracy optical convolution of randomly generated 8-bit grayscale 10×10 matrices A_3 and B_3 and 20×20 matrices A_4 and B_4 with elements in the range of 0 to 255 is demonstrated. Figure 4 compares the experimental results of the optical convolution of matrices A_3, A_4 and matrices B_3, B_4 with the theoretical results. In each box, the light intensity distributions of the spot arrays on the detected plane, denoting the raw results of convolution, are shown in the first subfigure of all columns. The theoretical results obtained using an electric computer (full precision, 64 bits) are highlighted in the second subfigure, and the experimental results are shown in the third subfigure. The relative error is defined as follows:

$$\mathrm{RE} = |C_{\mathrm{exp}} - C_{\mathrm{theo}}|/(|C_{\mathrm{max}} - C_{\mathrm{min}}|/256), \tag{5}$$

where C_{exp} represents experimental convolution, C_{theo} represents theoretical convolution, C_{max} represents the maximum value of theoretical convolution, and C_{min} is the minimum value of theoretical convolution. Furthermore, "$|\cdot|$" denotes the absolute operation. This relative error indicates that the precision of 8 bits will be obtained if its value is less than one.

Figures 4(a) and 4(b) show the results of the convolution of matrices A_3, A_4 and matrices B_3, B_4, respectively. It is demonstrated that the overall trend of the experimental and theoretical results of convolution is very consistent. After further assessment, the mean values of the relative error RE are 0.424 and 0.39, and the maximum values are 2.258 and 1.293, respectively. Also, from these error maps, one can see that the relative errors for most of points [98.06% and 97.25% in Figs. 4(a) and 4(b), respectively] are less than one, indicating that the computing accuracy is very close to 8 bits, which is high enough for most AI inference tasks and, at least, some training tasks. Additionally, other examples of the experimental results of larger-scale matrices were also demonstrated in the appendix (see Appendix C).

4. OPTICAL CNN INFERENCE TASKS BASED ON MNIST

With its ability to accelerate universal convolutional computation, this OMica could find applications in a variety of fields where dense convolutions are involved, such as simulation of optical imaging, multi-input multi-output systems, and training and inference of a CNN. As an example, we demonstrate the inference tasks of recognition of handwritten digits based on the OMica using the above-mentioned negative matrix coding method and hybrid analog–digital matrix convolution (see details of CNN in Appendix D). Here, a binary neural network (BNN) [32] is implemented as an example to test the robustness and accuracy of the proposed optical hardware. For a BNN, the input signal is a nonnegative binary (0 or 1) image, and the kernel is a signed binary matrix (−1 or +1) [33]. Each kernel of the BNN trained in advance is encoded into two identical-sized nonnegative matrices, one of which is a low-bit (positive) matrix and the other a high-bit (negative) matrix, as shown in Fig. 5(a). Intuitively, it seems that two convolution operations should be executed in the temporal sequence. Remarkably, 10 original kernels need to be divided into 10 high-bit sub-kernels and the same low-bit sub-kernel because the low-bit sub-kernels are the same. Furthermore, the first high-bit sub-kernel and low-bit sub-kernel are the same with unity transmittance. Thus, the total number of convolutional kernels after encoding is still 10, implying that no additional computational overhead incurs. Figure 5(b) shows the inference process of the CNN based on encoding low- and high-bit kernels. The 10 encoded kernels are sequentially loaded onto the SLM located at the input plane of matrix A, and the binary input images with a scale of 28×28 are sequentially loaded onto the SLM at the input plane of matrix B. When light passes through the two SLMs in sequence and is then focused and separated by the focusing lens, the detector on the focal plane captures the spot array denoting the convolutional results. Finally, the original convolutional results are obtained by decoding the corresponding low- and high-bit convolutions. By adding the results of the positive and negative convolutions and multiplying them by the weight −2, the final convolutional results can be obtained.

Figure 5(c) shows the absolute error AE map between the theoretical and experimental results of an input image of a handwritten digit 7 convolved by the first kernel. Compared

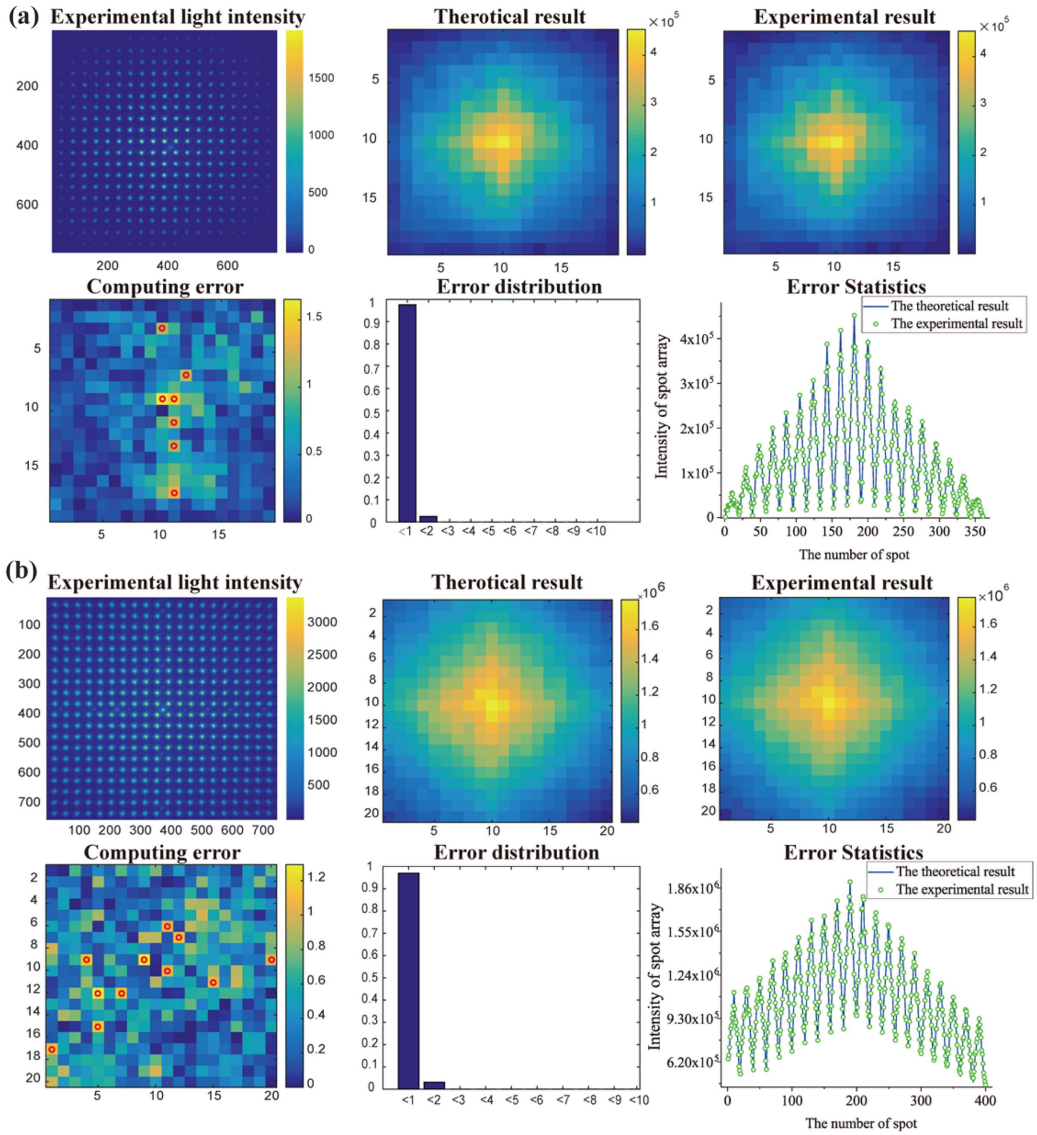

Fig. 4. Experimental results of high-accuracy convolution for two groups of grayscale matrices. (a), (b) Randomly generated 8-bit grayscale 10×10 matrices A_3 and B_3, 8-bit grayscale 20×20 matrices A_4 and B_4, respectively. The subfigures from left to right show the light intensity distribution of the spot array denoting the convolution, theoretical convolutional values, experimental convolutional results, error map between theoretical and experimental results (the red circle indicates the computing accuracy at that point is less than 8 bit), and histogram of the error distribution, respectively. The comparison of experimental convolutional results expands into one-dimensional (1D) vectors and theoretical convolutional results.

with the matrices in Fig. 4, the size of a standard input image of handwritten digits is 28×28, whereas the size of the convolutional kernel is nearly the same, and the average value of the absolute errors is 0.405. This implies that it is possible to calculate the optical convolution of larger-scale matrices using OMica with high precision. The following pooling layer, nonlinear operations, and full connections are executed by a classical electrical computer.

To validate the reliability and robustness of the system, we performed blind testing for the first 1000 sets of MNIST images with serial numbers ranging from 1 to 1000. As shown in Figs. 5(d) and 5(e), the experimental results indicate that the optical convolutional accelerator achieved blind-testing accuracy of up to 97.3%, whereas electrical computers achieved recognition accuracy of 96.7% for the same test dataset. This may be due to the computing error of the optical convolution carrying characteristics of the input images, thus further strengthening the feature extraction ability. It can be seen that the error maps for different handwritten digits are highly correlated with the input image, as shown in Fig. 5(c) (see Appendix E). By optimizing the kernel weights of the optical convolutional system, direct training of the optical CNN is expected to yield better results than those of an electronic computer. Based on this, the architecture can be effectively used as a hardware accelerator with large computing power in various DNNs.

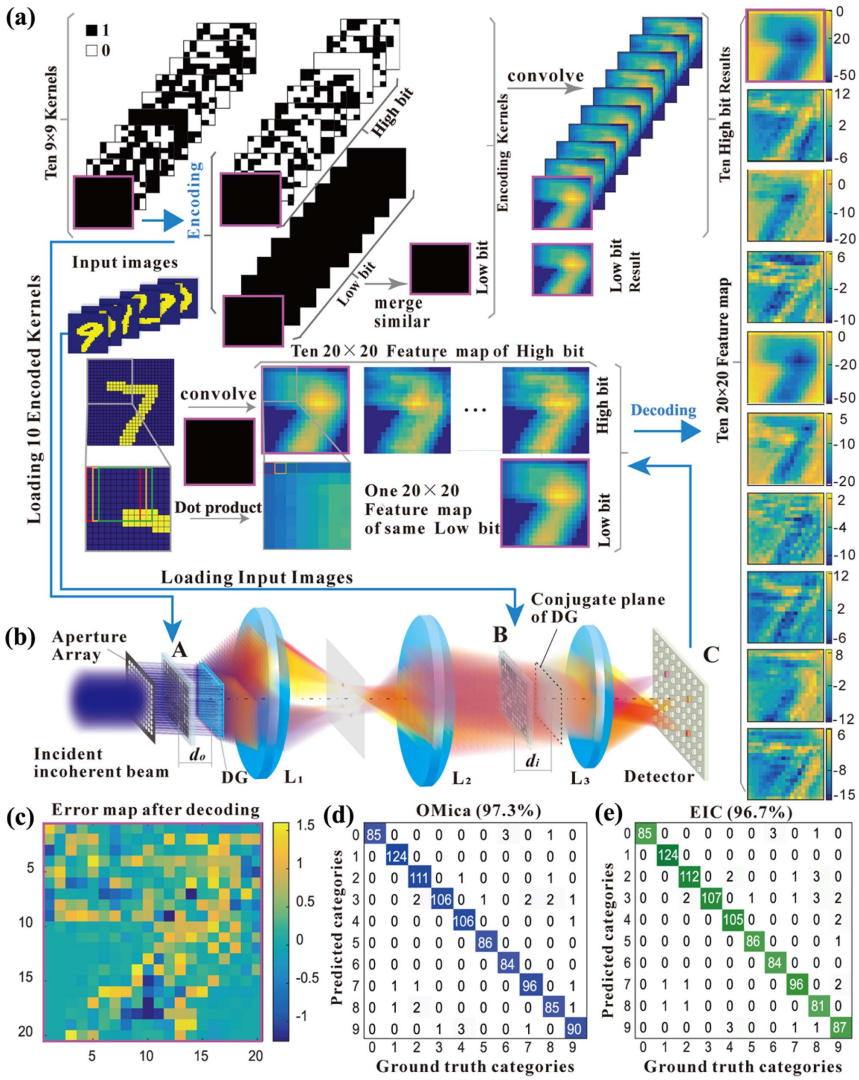

Fig. 5. Inference process for the convolutional neural network performed by OMica based on the MNIST dataset. (a) Execution of convolution operation by encoding each original convolutional kernel into high-bit and low-bit kernels; (b) schematic of the optical convolutional architecture performing CNN inference; (c) absolute error AE map comparing theoretical and experimental results of the convolution of a handwritten digit 7 as an input; confusion matrix of blind-testing 1000 images from the MNIST dataset when matrix convolutions are executed by the optical hardware (d) and by pure electric hardware (e). The purple box marks the first convolutional kernel to realize the whole process of encoding, convolution, and decoding.

5. DISCUSSION

A. Computing Power Scalability

As shown in Fig. 1, even when the suitable distance d between matrix A and the BS is adjusted to match the convolutional stride s, each diffraction order of the BS involved in the convolution is still imaged to the plane of matrix B. Therefore, it is possible to greatly reduce the physical size of the matrix elements. Given these conditions, the peak computing power of the optical convolutional architecture will reach 10 peta (10^{15}) operations per second (POPS) [34], which is even faster than the state-of-the-art GPU, such as TITAN RTX (Nvidia) [35], if a modulator with a higher refresh rate (typically 10 kHz) is used, such as a digital mirror device (DMD) or a specially designed micro–electro–mechanical system. Furthermore, if other multiplexing methods, such as polarization, wavelength, and spatial mode, are used, then speeds at least 10 to 10^2 times faster than this estimation can be achieved [36,37]. Therefore, based on the OMica, the computing power for convolution may, in the near future, be superior, or at least comparable, to that of the most powerful supercomputer (peak performance of the top system, Frontier [38] with Linpack Performance 1102 POPS), with larger-scale and higher-updating-frequency devices.

B. Energy Efficiency Ratio

Additionally, the power consumption of the optical convolutional system is significantly lower than that of an electronic processor with the same computing power, even for such a bulk optical system at present. This fully accounts for the operating power consumption of the optoelectronic device and assumes that the total power consumption of the entire optical convolution computing system, including the light source, two

modulators, and the detector, is less than 100 W. Of course, the power consumption of 100 W is meaningless for the MNIST dataset. However, as the matrix size increases, along with the aperture size and DG splitting ratio, etc., the increase in computing power is proportional to N^4, whereas the increase in the power consumption of this system is insignificant. Therefore, as computing power continues to grow, the energy efficiency ratio of this architecture will significantly outperform that of existing electronic computing systems. Furthermore, if a more sensitive detection device, such as a multiphoton counter, is used, power consumption will be drastically reduced [39]. In contrast, a powerful supercomputer is energy hungry, with power consumption typically reaching 10^4 to 10^5 kW (Frontier's power is 21,100 kW). Evidently, the optical convolutional architecture will consume far less power than supercomputers, whereas its computing power for a specific task (convolution) could be at least comparable to that of Frontier, the top supercomputer this year.

C. Potential Applications

To the best of our knowledge, the OMica is the only optical parallel acceleration solution that can produce both high-precision convolutional computers and AI hardware accelerators with high recognition accuracy. Additionally, if an appropriate distance d [Figs. 1(a) and 1(b)] is chosen, this OMica architecture could realize not only convolutional layers but also pooling layers and fully connected layers (all layers are linear convolution calculations). For AI algorithms, it has been demonstrated that very high accuracy is not required [40] and that neural networks can operate effectively with both low-accuracy and fixed-point operations. Inference models function nearly as well with 4–8 bits of precision and are trained with nearly 8–16 bits of precision per computation [41]. Our results indicate that computing accuracy is close to 8 bits, which is sufficiently accurate for most AI inference applications. Moreover, if high-contrast modulators, such as DMDs, are used, computing accuracy could be improved even further, and the results obtained from this optical accelerator would be sufficient for training most AI models. Furthermore, when training the neural network directly in this optical convolutional system, the physical characteristics of the system itself are also trained, such as alignment errors and cross talk, which are expected to further improve the performance of the aforementioned neural network.

Presently, only one kernel A and one input feature map B are loaded onto these two SLMs. It is also possible to load multiple kernels on the first SLM, allowing for parallel convolutions among multiple kernels and multiple input channel feature maps by filling an appropriate number of zero elements between any two adjacent kernels. By swapping the positions of feature map B and kernel A, a CNN can be built, and the key is to make full use of pixels to increase computing power. Also, it is worth noting that considering the actual hardware scale, it is often necessary to split and reorganize the input feature map to further improve the hardware utilization, that is, to load different matrix combinations to the SLMs to execute the convolution process.

Although these task-specific devices are not yet available, the current CMOS technology, in principle, is adequate for developing high-quality devices, such as SLMs and detectors, for optical computing. This work presents a promising method for building optical convolutional processors to overcome the intrinsic shortage of computing power and unsatisfactory energy efficiency in traditional electrical processors. Furthermore, the experimental results validate the benefits of optical convolutional systems for various application scenarios, including computationally intensive tasks and neuromorphic computing.

6. CONCLUSION

An optical convolutional accelerator for fully parallel universal convolution computing was proposed, and a negative matrix coding scheme with sufficiently high precision was demonstrated. In principle, a suitable encoding scheme and the OMica can be used to efficiently calculate the convolution of an arbitrary bit matrix with massive parallelism and sufficient accuracy. Moreover, convolution is universal, and the computing results obtained may be easily transferred to any other computing platform. Our proof-of-concept experimental results proved the feasibility of the optical convolution of 20 × 20 matrices with an accuracy of about 8 bits. Furthermore, a BNN for handwritten digit recognition tasks on the standard MNIST dataset was constructed, and the inference process was demonstrated based on this optical hardware. The results indicated that the blind test recognition accuracy can reach 97.3%, which is comparable with that predicted by pure electrical networks. These proof-of-concept experimental results indicated that the OMica could be used for massive parallelism, high-precision, and high-efficiency AI accelerators, and this computing paradigm has potential applicability in the construction of task-specific cloud computing centers or other AI computing centers. By developing high-speed SLMs with higher contrast, optimizing a specially proposed projection imaging system, and setting up a dedicated dot array lighting source, it is possible to build a photonic coprocessor with higher computing power and lower energy consumption than state-of-the-art supercomputers, such as Frontier, based on the OMica. Additionally, the characteristics of the imaging system itself suggest that the computing power of the system can be exponentially increased by cascading multiple $4f$ systems and employing extra multiplexing degrees of freedom. Thus, a hybrid optical–electrical computer center or data center could be directly constructed. Furthermore, because the optical hardware could work under incoherent white-light illumination if an achromatic lens projection system is used, the OMica architecture allows it to handle white-light images directly from lenses without traditional photoelectric conversion.

In summary, the OMica is expected to be used in self-driving vehicles [42], machine vision [43], and other fields that require high computing power for real-time or quasi-real-time data processing. This opens the door to increasing the computing power and energy efficiency of convolution by using high-performance devices, such as larger-scale modulators with higher updating frequencies and detectors or detector arrays with wider dynamic ranges and higher sampling frequencies, which would be superior to the most powerful supercomputers, in the near future.

APPENDIX A: EXPERIMENTAL SETUP AND METHODS

Figure 6 shows a proof-of-concept experimental system based on the OMica. Figure 7 shows photographs of the experimental setup. Two large-scale matrices, A and B, are assumed to be two convolved matrices and are loaded onto two modulators, SLM_1 and SLM_2, respectively. The convolutional matrix C is detected by $sCMOS_1$. The DG was removed before alignment, and the monitoring camera was placed in the focal plane of L_9. During the alignment process, some specially designed patterns shown in Fig. 8 are used. Subsequently, the DG is inserted before SLM_1, and the distance d_0 between the DG and SLM_1 should be adjusted carefully to make the lateral shift of the image correspond to normalized convolutional stride size $s = 1$ [Eq. (1)].

We can then confirm this alignment by placing a single tunable iris on the focal plane of L_6 to allow only one order to pass through the iris aperture at a time. Noticeably, when the iris is moved, the window slides along the moving direction, allowing different diffraction orders to pass through the aperture in sequence. Here, as an example, one rectangular array of 8×8 square blocks [Fig. 9(a)] and one rectangular array of eight circular blocks [Fig. 9(b)] are loaded onto SLM_2 and SLM_1, respectively. As shown in Fig. 9(c), matrix A markedly slides on matrix B. In this case, the iris moves from left to right, and the diffraction orders of $(+1^{st}, +1^{st})$, $(-1^{st}, +1^{st})$, $(-3^{rd}, +1^{st})$, $(-5^{th}, +1^{st})$, $(-7^{th}, +1^{st})$, $(-9^{th}, +1^{st})$, $(-11^{th}, +1^{st})$, and $(-13^{th}, +1^{st})$ are sequentially passed through the iris aperture. Because of the cut-off effect of the square aperture located on the conjugating plane of SLM_2, we can observe that eight columns of the rectangular array of circular blocks are changed to one. Alignment between the two matrices is achieved if all circular blocks loaded on SLM_1 are aligned with the square blocks loaded on SLM_2.

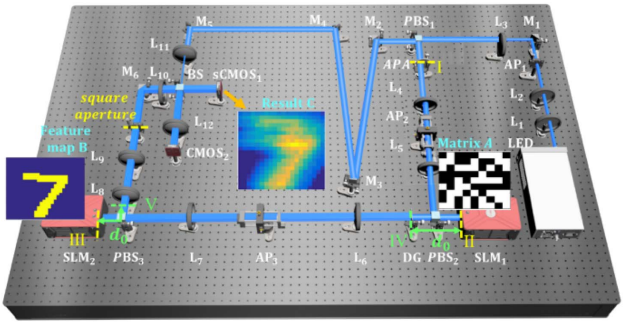

Fig. 6. Schematic of the optical convolution experimental system using the DG. LED, light-emitting diode with wavelength $\lambda = 450$ nm; M_{1-6}, reflective aluminum mirrors; $AP_{1,2,3}$, aperture pinholes; L_{1-5}, convergent lenses; L_6, L_7, L_{10}, Fourier transform lenses; $PBS_{1,2,3}$, cube polarization beam splitters; SLM_1, SLM_2, reflected liquid crystal SLMs; APA, aperture array; DG, Dammann grating; BS, non-polarizing beam splitter; $sCMOS_1$, scientific complementary metal–oxide–semiconductor camera for detection; $CMOS_2$, CMOS camera for monitoring. I, II, III, and the plane of the square aperture are one group of object–image conjugate planes. IV and V are other groups of object–image conjugate planes. Plane V is the image plane of the DG. d_0 is the characteristic distance corresponding to $s = 1$, which can be adjusted to match the physical size of the matrix unit of matrix B to the different stride size.

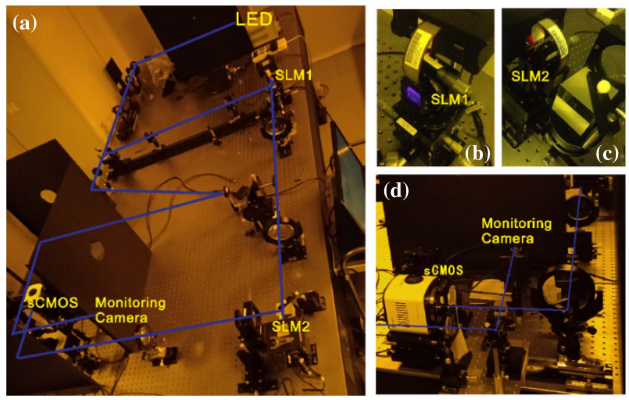

Fig. 7. Photographs of the experiment system of OMica. (a) Entire optical system; (b) SLM mounted on a 4D manual stage for loading kernel A, (c) SLM mounted on a 4D manual stage for loading matrix B, and (d) enlarged part of the $sCMOS_1$ detector and monitoring $CMOS_2$ camera.

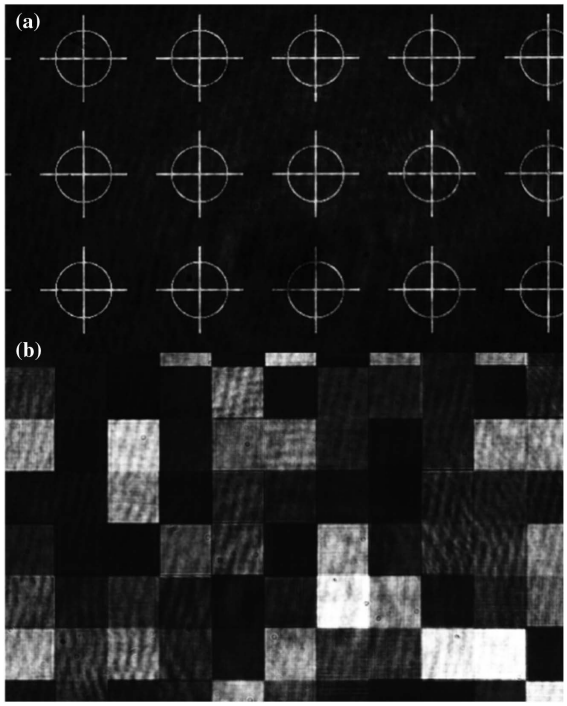

Fig. 8. Typical patterns loaded onto two SLMs for alignment. (a) Alignment pattern and (b) square array pattern.

APPENDIX B: DESIGN AND MANUFACTURING OF DAMMANN GRATING

Here, a simulated annealing algorithm is used to optimize the structure of DGs. The normalized energy distributions of 1×20 and 1×28 DGs with diffraction orders for ideal π phase retardation are shown in Fig. 10. Under ideal conditions, the efficiencies of 1×20 and 1×28 1D gratings were 81.93% and 82.38%, respectively, and the energy uniformity was less than 1%. The structure of a 2D DG can be easily obtained after the orthogonal superposition of two crossing 1D gratings.

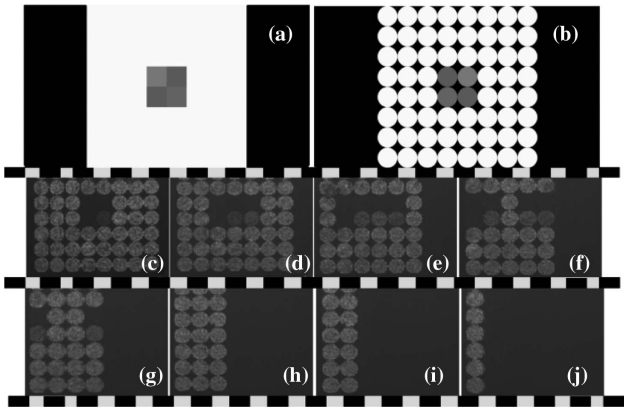

Fig. 9. Experimental results for demonstration of kernel sliding. (a), (b) Images loaded onto two SLMs. (c)–(j) Images captured by the monitoring $CMOS_2$ camera as the iris moves from left to right, allowing only one diffraction order to pass through its aperture in sequence.

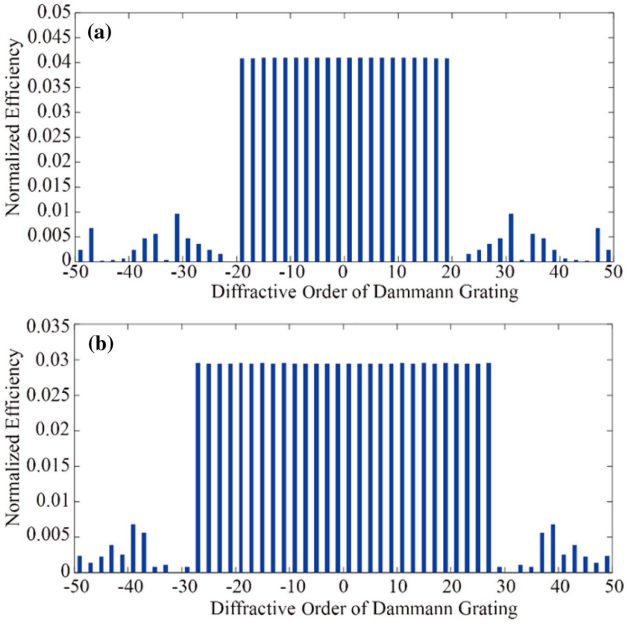

Fig. 10. 1 × 20 (a) and 1 × 28 (b) DG beam splitting order normalized energy distribution.

Figure 11 depicts the intensity distributions and diffractive angles versus diffraction orders for a 20 × 28 2D DG. It can be seen that the normalized energy corresponding to each diffraction order is approximately 0.12% of the incident light energy (without considering the interface reflection and other losses). Furthermore, the angle distribution versus diffraction order shows that the maximum diffraction angle is approximately 3.78°. This implies that the paraxial condition is approximately maintained in this situation. In our experiment, the grating period was designed to be 225 μm in both directions, and the feature size was approximately 2.47 μm. Fused silica was chosen as the substrate, and the grating sample was fabricated using lithography and ion etching. The diffractive pattern in the far field is captured in the focal plane of a Fourier lens, and the intensity map is shown in Fig. 11(c). The intensity distribution for this spot array was estimated by calculating the accumulation of the intensity around the centroid for each spot. The results indicate that the uniformity of the fabricated 20 × 28 DG was approximately 5%. This level of uniformity had almost no negative effect on the final computed results after calibration.

APPENDIX C: CONVOLUTIONAL RESULTS FOR TWO 8-BIT GRAYSCALE 180 × 224 LARGE MARTRICES

In principle, the OMica can achieve high computing power due to its true parallel processing capabilities. Furthermore, the convolution of two 180 × 224 matrices was also demonstrated in the analog framework. The theoretical and experimental results, as well as the experimental detection of the light distribution of the convolution, are shown in Figs. 12(a)–12(c). The relative errors defined above are shown in Fig. 12(e). The mean errors for the five groups of data computed using OMica hardware were 10.87, 10.93, 11.12, 11.17, and 11.48, respectively. This low precision was mainly caused by the alignment error. This alignment error could be significantly reduced using piezo actuators with resolutions in the nanometer range. Under this condition, a matrix scale of ∼200 × 200 indicates that the peak computing power reaches 3.2×10^9 MAC operations when light passes through the system once.

APPENDIX D: CONFIGURATION OF THE CNN

The configuration of the CNN model used in our experiment for demonstration of the handwritten digit recognition based on the MNIST dataset is shown in Fig. 13. It can be seen that this CNN network contains five layers: convolutional layer, pooling layer, nonlinear activation layer, and two fully connected layers. To achieve a higher recognition rate while avoiding overfitting, we set the learning rate to 0.05 and the training batch size to 50. The number of epochs was set to four to avoid overfitting. The activation function for the first layer was the rectified linear unit (ReLU) function, and 10 9×9 convolutional kernels with binary element values of −1 or +1 were used. Owing to its simple derivative formation, the training speed of the ReLU function is faster than that of the sigmoid and tanh functions when the kernel weights are trained based on the backpropagation algorithm. Because the derivative is not zero, it can effectively address the vanishing gradient problem and further reduce overfitting. The average pooling method was selected for the pooling layer because all the information in the feature map can be obtained on average without losing too much information. Because the image is processed through binarization in advance, the foreground and background information in the feature map maintains a high resolution after average pooling. The first fully connected layer had 200 nodes, and the activation function was chosen as the ReLU function. The last fully connected layer had 10 nodes, and the activation function was selected as the sigmoid function. Because the sigmoid function is used in the final layer for classification tasks,

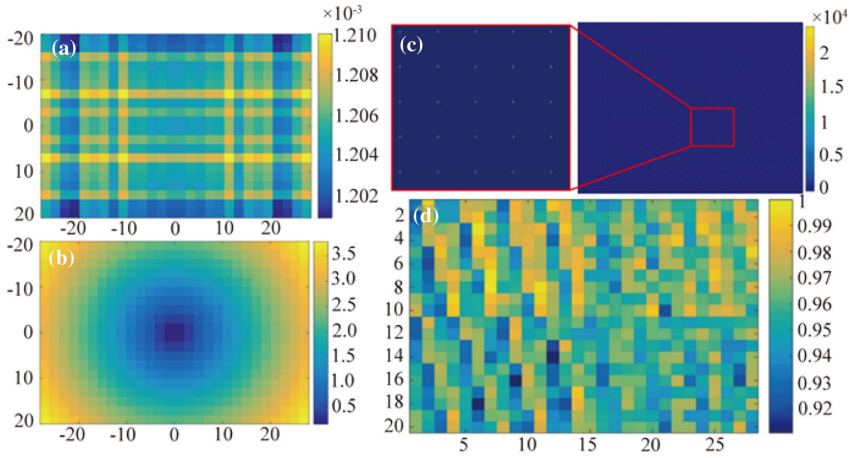

Fig. 11. Intensity and angle distribution of 20 × 28 2D DG. (a) Simulation result of intensity distribution versus different orders; (b) simulation result of diffraction angle versus diffraction order; (c) intensity map of the spot array captured in the experiment (the cross represents the centroid); (d) experimental results of normalized intensity distribution versus diffraction order.

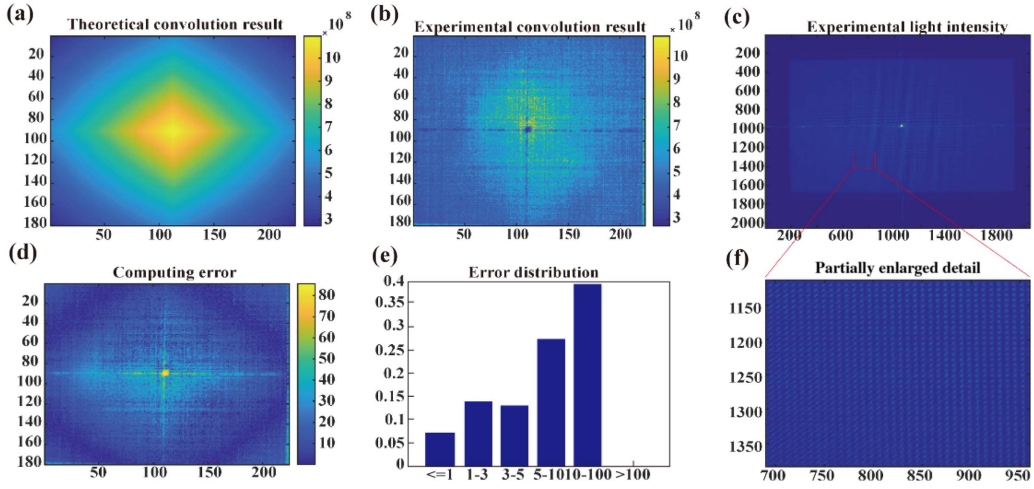

Fig. 12. Experimental convolutional results for 180 × 224 matrices. (a)–(c) Theoretical convolutional results, experimental convolutional results, and experimental detection light distribution, respectively; (d) partially enlarged view of the experimental light spot on (c); (e) error distribution; (f) proportion of experimental light intensity distribution.

we chose the cross-entropy loss function to avoid the vanishing gradient problem. For the 60,000 training set, the total training time of the CNN network was approximately 3 min (Intel Core i7-4790 CPU at 3.60 GHz), and the recognition accuracy on the 1000-sample test set was 96.7%. The recognition accuracy on the 10,000-sample test set was 96.3%.

Here, we divide the 60,000 training samples into two parts: training set and validation set. The learning curve is shown in Fig. 13. The difference between the true and predicted values is defined as the error, and the loss is the average of the errors of all samples. The loss function used in this network is the cross-entropy loss function, $H(p,q) = \sum_{i=1}^{n} p(x_i) \log \frac{1}{q(x_i)} = -\sum_{i=1}^{n} p(x_i) \log q(x_i)$, where $p(x)$ is the predicted probability, and $q(x)$ is the experimental probability. The loss of the training set decreases demonstrably as the number of training batches increases. Additionally, the decreasing trend of the losses of the validation and training sets is almost coincidental, indicating the good fitting ability of the model.

It can also be observed from Fig. 14 that the loss curve of the training set drops rapidly until it fluctuates slightly near a stable value. Additionally, the validation and training set losses are slightly different, indicating that the model has some generalization ability. The accuracy of the validation set gradually reaches a stable value as the number of samples increases, indicating that no overfitting effect occurs in this model.

APPENDIX E: INPUT-RELATED CROSS TALK

Figure 15 shows the distribution of relative errors between the experimental convolutional results and theoretical convolutional results for different digital inputs. These error maps are clear characteristic of the input numbers. This may be due to optical cross talk between different pixel channels. Optical cross talk is an important factor that limits the improvement of optical computing accuracy. However, for the AI algorithm, if training of the deep learning network model

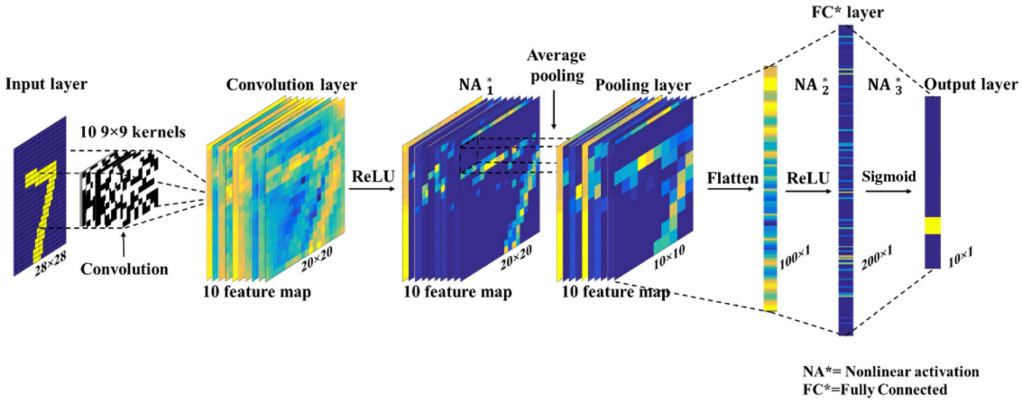

Fig. 13. Schematic of the CNN architecture.

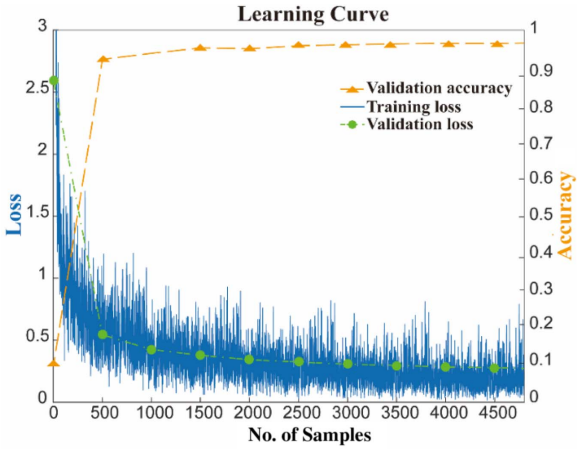

Fig. 14. Learning curve of the CNN.

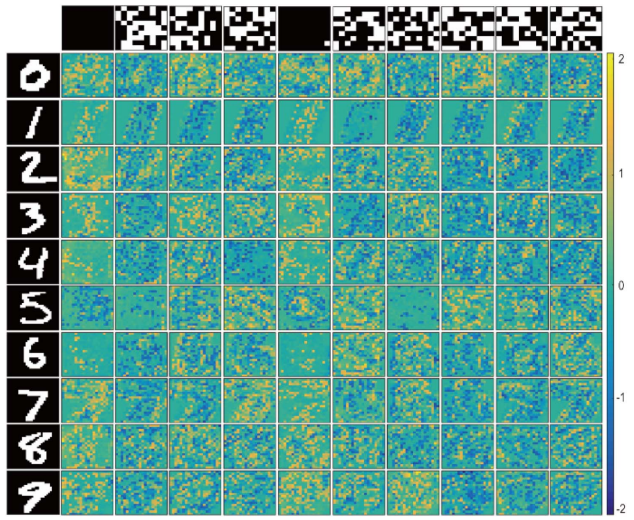

Fig. 15. Typical error maps between convolutional results obtained from the optical hardware and that of an electrical computer with the full precision of different input handwritten digits (from 0 to 9) for these 10 convolutional kernels after encoding.

is directly based on an optical computing system, then the optical cross talk may help improve the recognition accuracy of the system. This result has implications for developing optical AI accelerators with high recognition accuracy.

APPENDIX F: SUMMARY OF DIFFERENT OPTICAL CONVOLUTIONAL ARCHITECTURES

Table 1 shows a summary of various mainstream optical convolutional architectures (OIU, optical interference unit; MRs, microring resonators; OFC, optical frequency comb; PCM, phase change materials; D^2NN, diffractive DNN). It has been shown that precision of only about 4–5 bits is achieved for most photonic accelerators reported, although they work well for most artificial learning tasks after retraining with noise. However, it has been verified empirically that for most neural networks, inference models work nearly just as well with 4–8 bits of precision, while training with nearly 8–16 bits of precision per computation [41]. This is one important reason why most photonic accelerators have been used only for inference tasks. Besides artificial neural networks, the OMica provides the ability for accelerating universal convolution computation, and thus could find applications in many other fields, such as simulation of optical imaging, and multi-input multi-output systems.

Compared with the most popular scheme involving planar waveguides on a 2D substrate [17,18,22,24], the scheme of multiple cascading DOEs inherently takes full advantage of the 3D connection ability of optics. Thus, it can achieve higher computing power in a single computing step. Recently, Xu *et al.* realized a type of photonic convolutional accelerator based on optical frequency combs [17], whose computing power is as high as tera operations per second (TOPS). The use of optical frequency combs to realize multi-wavelength light sources is remarkable progress. However, the scalability of this architecture is still limited by the number of channels of the optical frequency combs. Mario *et al.* [29] proposed an optical system that performs fast updating of optical neural networks based on two amplitude-only DMDs, where one amplitude-only DMD is located at the Fourier transform plane of the other. Although the mapping relationship between the input images and the recognition digits can be successfully established using this method, the computing results are essentially not standard convolutions. Therefore, this method cannot be used for

Table 1. Summary of Different Optical Convolutional Architectures

Architecture	Principle	Pros/Cons	Computing Accuracy	References
OIUs and delay line	Matrix–vector multiplication	• High integration and high modulation speed. • Limited by the integration of integrated photonic devices, it is difficult to realize the parallel convolution process of multiple convolutional kernels.	~5 bits	[16,18,24]
MRs, OFC, and PCM	Matrix-vector multiplication	• High integration and high modulation speed. • The OFC can provide multi-wavelength light sources and timing modulation, and the system integration is higher. • Low power consumption using non-volatile PCM. • Complex electronic control and test configuration.	~5 bits	[17,22]
$4f$ filter	Multiplication in frequency domain equals convolution in spatial domain	• Object and spectrum are limited by the Fourier transform relationship. There is a trade-off between computing accuracy and computing size. • Configuration is very simple.	/	[28,29]
D^2NN	Diffraction	• High-precision 3D macro-nano structures are difficult to fabricate, and computational accuracy is limited. • High computing power.	~5 bits	[19,44]
Shadow casting	2D matrix–matrix multiplication	• Diffraction effect exists when matrix A is projected onto matrix B, and computing accuracy cannot be guaranteed. • Configuration is very simple.	/	[30,31,39]
OMica	2D matrix–matrix convolution and multiplication	• DG and object–image conjugation avoids diffraction effects by wavefront recombination. • DG is 2D DOE, and it is easy to manufacture. Computing power can be expanded easily by using large-scale DGs. • Can work under incoherent light illumination and directly handle optical images. • Computational accuracy is high.	~8 bits	This work

high-precision universal convolution computing. Moreover, it is difficult to align the two DMDs pixel by pixel. Because of the Fourier transform, the relationship between input and filter planes, realizing large-scale optical networks will be difficult. Recently, Zhou *et al.* [44,45] demonstrated a reconfigurable scheme for realizing 3D architecture with multiple cascading DOEs, using two programmable modulators and a DMD, as well as another pure-phase SLM, for amplitude and phase modulation, respectively. Because of the coherent working mode, micrometer-sized pixels, alignment error between the DMD and SLM, and alignment errors between different layers, achieving high computing precision is difficult. Therefore, recognition is drastically degraded without adaptive training. Although this scheme performs well after adaptive training, it cannot be used for universal convolution computing because of its low precision.

In contrast, because of the object–image conjugate relationship, a CMOS monitoring camera can be added to the conjugating plane of two SLMs, making it simple to align two SLMs with a monitor camera. Additionally, an incoherent light source could be used in this architecture to prevent sensitivity and speckle noise. More importantly, this configuration makes it possible to handle images directly from a lens under white-light illumination, which is very challenging for all mainstream architectures, to the best of our knowledge.

Therefore, the convolutional accelerator enabled by the OMica can be used to compute universal matrix convolution, and the results obtained by the hybrid optical–electrical hardware can be easily transferred to any other computing platform, including photonic, hybrid optical–electrical, and traditional electric processors or coprocessors. Because of its universality, this architecture can be used for building task-specific cloud computing centers, or some other AI accelerating centers, as well as the present bulk optical system. In the future, with the advancement of nonlinear optical elements, a scheme based on the OMica could also be integrated into pure photonic accelerators by combining planar waveguides [46,47], metasurfaces [48–50], and advanced modulator arrays, etc.

Funding. Chinese Academy of Sciences (QYZDJ-SSW-JSC014); Science and Technology Commission of Shanghai Municipality (19DZ2291102, 19JC1415400, 20ZR1464700). Service Platform of Shanghai Precision Optical Manufacture and Test.

Acknowledgment. The authors appreciate the critical discussion on this concept with Guowei Li and also his assistance in the experiment.

Disclosures. G. Ma, J. Yu, and C. Zhou have filed a patent on the proposed OMica architecture (patent number 2021107423134).

Data Availability. Data underlying the results presented in this paper are not publicly available at this time but may be obtained from the authors upon reasonable request.

REFERENCES

1. Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," Proc. IEEE **86**, 2278–2324 (1998).

2. A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," Adv. Neural Inf. Process. Syst. **25**, 1097–1105 (2013).
3. Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," Nature **521**, 436–444 (2015).
4. J. Cong and B. Xiao, "Minimizing computation in convolutional neural networks," in *International Conference on Artificial Neural Networks* (2014), pp. 281–290.
5. T. F. De Lima, H.-T. Peng, A. N. Tait, M. A. Nahmias, H. B. Miller, B. J. Shastri, and P. R. Prucnal, "Machine learning with neuromorphic photonics," J. Lightwave Technol. **37**, 1515–1534 (2019).
6. Y. Ito, R. Matsumiya, and T. Endo, "OOC-cuDNN: accommodating convolutional neural networks over GPU memory capacity," in *IEEE International Conference on Big Data* (2017), pp. 183–192.
7. K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE Conference on Computer Vision and Pattern Recognition* (2016), pp. 770–778.
8. G. Wetzstein, A. Ozcan, S. Gigan, S. Fan, D. Englund, M. Soljačić, C. Denz, D. A. B. Miller, and D. Psaltis, "Inference in artificial intelligence with deep optics and photonics," Nature **588**, 39–47 (2020).
9. B. J. Shastri, A. N. Tait, T. F. de Lima, W. H. P. Pernice, H. Bhaskaran, C. D. Wright, and P. R. Prucnal, "Photonics for artificial intelligence and neuromorphic computing," Nat. Photonics **15**, 102–114 (2021).
10. P. Ambs, "Optical computing: a 60-year adventure," Adv. Opt. Photon. **2010**, 1–15 (2010).
11. A. Maréchal and P. Croce, "Un filtre de fréquences spatiales pour l'amélioration du contraste des images optiques," C. R. Acad. Sci. **237** (1953).
12. L. De Marinis, M. Cococcioni, P. Castoldi, and N. Andriolli, "Photonic neural networks: a survey," IEEE Access **7**, 175827 (2019).
13. P. R. Prucnal and B. J. Shastri, *Neuromorphic Photonics* (CRC Press, 2017).
14. F. Thomas, B. J. Shastri, A. N. Tait, M. A. Nahmias, and P. R. Prucnal, "Progress in neuromorphic photonics," Nanophotonics **6**, 577–599 (2017).
15. Q. Zhang, H. Yu, M. Barbiero, B. Wang, and M. Gu, "Artificial neural networks enabled by nanophotonics," Light Sci. Appl. **8**, 42 (2019).
16. Y. Shen, N. C. Harris, S. Skirlo, M. Prabhu, T. Baehr-Jones, M. Hochberg, X. Sun, S. Zhao, H. Larochelle, D. Englund, and M. Soljačić, "Deep learning with coherent nanophotonic circuits," Nat. Photonics **11**, 441–446 (2017).
17. X. Xu, M. Tan, B. Corcoran, J. Wu, A. Boes, T. G. Nguyen, S. T. Chu, B. E. Little, D. G. Hicks, R. Morandotti, A. Mitchell, and D. J. Moss, "11 TOPS photonic convolutional accelerator for optical neural networks," Nature **589**, 44–51 (2021).
18. H. Bagherian, S. Skirlo, Y. Shen, H. Meng, V. Ceperic, and M. Soljacic, "On-chip optical convolutional neural networks," arXiv, arXiv:1808.03303v2 (2018).
19. X. Lin, Y. Rivenson, N. T. Yardimci, M. Veli, Y. Luo, M. Jarrahi, and A. Ozcan, "All-optical machine learning using diffractive deep neural networks," Science **361**, 1004–1008 (2018).
20. A. Silva, F. Monticone, G. Castaldi, V. Galdi, A. Alù, and N. Engheta, "Performing mathematical operations with metamaterials," Science **343**, 160–163 (2014).
21. J. Feldmann, N. Youngblood, C. D. Wright, H. Bhaskaran, and W. H. P. Pernice, "All-optical spiking neurosynaptic networks with self-learning capabilities," Nature **569**, 208–214 (2019).
22. J. Feldmann, N. Youngblood, M. Karpov, H. Gehring, X. Li, M. Stappers, M. Le Gallo, X. Fu, A. Lukashchuk, A. S. Raja, J. Liu, C. D. Wright, A. Sebastian, T. J. Kippenberg, W. H. P. Pernice, and H. Bhaskaran, "Parallel convolutional processing using an integrated photonic tensor core," Nature **589**, 52–58 (2021).
23. F. Ashtiani, A. J. Geers, and F. Aflatouni, "An on-chip photonic deep neural network for image classification," Nature **606**, 501–506 (2022).
24. S. Xu, J. Wang, R. Wang, J. Chen, and W. Zou, "High-accuracy optical convolution unit architecture for convolutional neural networks by cascaded acousto-optical modulator arrays," Opt. Express **27**, 19778–19787 (2019).
25. H. Dammann and E. Klotz, "Coherent optical generation and inspection of two-dimensional periodic structures," Opt. Acta **24**, 505–515 (1977).
26. C. Zhou and L. Liu, "Numerical study of Dammann array illuminators," Appl. Opt. **34**, 5961–5969 (1995).
27. J. Yu, C. Zhou, W. Jia, W. Cao, S. Wang, J. Ma, and H. Cao, "Three-dimensional Dammann array," Appl. Opt. **51**, 1619–1630 (2012).
28. J. Chang, V. Sitzmann, X. Dun, W. Heidrich, and G. Wetzstein, "Hybrid optical-electronic convolutional neural networks with optimized diffractive optics for image classification," Sci. Rep. **8**, 12324 (2018).
29. M. Miscuglio, Z. Hu, S. Li, J. K. Grorge, R. Capanna, H. Dalir, P. M. Bardet, P. Gupta, and V. J. Sorger, "Massively parallel amplitude-only Fourier neural network," Optica **7**, 1812–1819 (2020).
30. C. Zhou, L. Liu, and Z. Wang, "Binary-encoded vector–matrix multiplication architecture," Opt. Lett. **17**, 1800–1802 (1992).
31. L. Liu, G. Li, and Y. Yin, "Optical complex matrix–vector multiplication with negative binary inner products," Opt. Lett. **19**, 1759–1761 (1994).
32. H. Qin, R. Gong, X. Liu, X. Bai, J. Song, and N. Sebe, "Binary neural networks: a survey," Pattern Recogn. **105**, 107281 (2020).
33. M. Courbariaux, I. Hubara, D. Soudry, R. El-Yaniv, and Y. Bengio, "Binarized neural networks training neural networks with weights and activations constrained to +1 or –1," arXiv, arXiv:1602.02830 (2016).
34. C. Zhou, J. Yu, G. Li, and G. Ma, "Roadmap of optical computing," Proc. SPIE **11898**, 118981B (2021).
35. https://www.nvidia.com/en-us/deep-learning-ai/products/titan-rtx/.
36. P. Minzioni, C. Lacava, T. Tanabe, J. Dong, X. Hu, G. Csaba, W. Porod, G. Singh, A. E. Willner, A. Almaiman, V. Torres-Company, J. Schröder, A. C. Peacock, M. J. Strain, F. Parmigiani, G. Contestabile, D. Marpaung, Z. Liu, J. E. Bowers, L. Chang, S. Fabbri, M. R. Vázquez, V. Bharadwaj, S. M. Eaton, P. Lodahl, X. Zhang, B. J. Eggleton, W. J. Munro, K. Nemoto, O. Morin, J. Laurat, and J. Nunn, "Roadmap on all-optical processing," J. Opt. **21**, 063001 (2019).
37. J. Wang, J.-Y. Yang, I. M. Fazal, N. Ahmed, Y. Yan, H. Huang, Y. Ren, Y. Yue, S. Dolinar, M. Tur, and A. E. Willner, "Terabit free-space data transmission employing orbital angular momentum multiplexing," Nat. Photonics **6**, 488–496 (2012).
38. https://www.top500.org/system/180047/.
39. T. Wang, S.-Y. Ma, L. G. Wright, T. Onodera, B. C. Richard, and P. L. McMahon, "An optical neural network using less than 1 photon per multiplication," Nat. Commun. **13**, 123 (2022).
40. S. Gupta, A. Agrawal, K. Gopalakrishnan, and P. Narayanan, "Deep learning with limited numerical precision," arXiv, arXiv:1502.02551 (2015).
41. M. A. Nahmias, T. F. de Lima, A. N. Tait, H.-T. Peng, B. J. Shastri, and P. P. Prucnal, "Photonic multiply-accumulate operations for neural networks," IEEE J. Quantum Electron. **26**, 7701518 (2020).
42. J. Han, A. Jentzen, and E. Weinan, "Solving high-dimensional partial differential equations using deep learning," Proc. Natl. Acad. Sci. USA **115**, 8505–8510 (2018).
43. L. Mennel, J. Symonowicz, S. Wachter, D. K. Polyushkin, A. J. Molina-Mendoza, and T. Mueller, "Ultrafast machine vision with 2D material neural network image sensors," Nature **579**, 62–66 (2020).
44. T. Zhou, X. Lin, J. Wu, Y. Chen, H. Xie, Y. Li, J. Fan, H. Wu, L. Fang, and Q. Dai, "Large-scale neuromorphic optoelectronic computing with a reconfigurable diffractive processing unit," Nat. Photonics **15**, 367–373 (2021).
45. T. Zhou, L. Fang, T. Yan, J. Wu, Y. Li, J. Fan, H. Wu, X. Lin, and Q. Dai, "*In situ* optical backpropagation training of diffractive optical neural networks," Photon. Res. **8**, 940–953 (2020).
46. M. Gruber, "Multichip module with planar-integrated free-space optical vector-matrix-type interconnects," Appl. Opt. **43**, 463–470 (2004).
47. G. Mínguez-Vega, M. Gruber, J. Jahns, and J. Lancis, "Achromatic optical Fourier transformer with planar-integrated free-space optics," Appl. Opt. **44**, 229–235 (2005).
48. Y. Zhang, C. Fowler, J. Liang, B. Azhar, M. Y. Shalaginov, S. Deckoff-Jones, S. An, J. B. Chou, C. M. Roberts, V. Liberman, M. Kang, C. Ríos, K. A. Richardson, C. Rivero-Baleine, T. Gu, H. Zhang, and J. Hu, "Electrically reconfigurable non-volatile metasurface using low-loss optical phase-change material," Nat. Nanotechnol. **3**, 661–666 (2021).
49. Z. Wu, M. Zhou, E. Khoram, B. Liu, and Z. Yu, "Neuromorphic metasurface," Photon. Res. **8**, 46–50 (2020).
50. H. Kwon, D. Sounas, A. Cordaro, A. Polman, and A. Alù, "Nonlocal metasurfaces for optical signal processing," Phys. Rev. Lett. **121**, 173004 (2018).