LASER & PHOTONICS REVIEWS
www.lpr-journal.org

Direct Optical Convolution Computing Based on Arrayed Waveguide Grating Router

Jialin Cheng, Chong Li, Jun Dai, Yayan Chu, Xinxiang Niu, Xiaowen Dong, and Jian-Jun He**

Optical convolution computing is gaining traction owing to its inherent parallelism, multi-dimensional processing, and energy efficiency. To handle input dimensions of N, conventional implementations necessitate N^2 optical elements, such as Mach–Zehnder interferometers or micro-ring resonators, to process multiply-accumulate (MAC) operations, limiting scalability and resulting in elevated power consumption. Here, a direct convolution computing method based on wavelength routing, utilizing the unique sliding property of an arrayed waveguide grating router (AWGR) to perform the sliding window operation of the convolution in the wavelength–space domains is proposed. With two input vectors directly loaded onto two modulator arrays, the convolution result is instantaneously produced at a photodetector array. The entire convolution computation is executed within a single clock cycle without the need for preprocessing or decomposition into elementary MAC operations. The number of active elements is minimal, only needed for input/output. The proposed optical convolution unit has striking advantages of high scalability, high speed, and processing simplicity compared to those based on optical matrix-vector multipliers. In the first experimental demonstration, a remarkable classification accuracy of up to 98.2% in handwritten digit recognition tasks using a LeNet-5 neural network is achieved.

1. Introduction

With the rapid development of artificial intelligence (AI) technology based on deep learning, deep neural networks have played an increasingly important role in many fields and applications.[1] Deep convolutional neural networks (CNNs), due to their incorporation of spatial invariance through convolutional operations and parameter sharing,[2] have demonstrated significantly superior performance in certain tasks compared to conventional fully connected networks.[3,4] This superiority is particularly pronounced in tasks related to vision, such as image recognition,[5–8] object detection,[9–12] semantic segmentation,[13–16] and image generation.[17–19] However, with the rapid increase in the depth and breadth of neural networks, the computational demands for training deep neural networks have escalated swiftly. As Moore's Law wanes, traditional electronic architectures become increasingly constrained.[20,21] Consequently, the quest for computing architectures that are faster and more energy-efficient, tailored for deep neural networks, becomes increasingly paramount. Optical methods show great potential for the next wave of neural network accelerators due to their advantages of ultra-wide bandwidth, low power consumption, and inherent parallelism, making them compelling candidates for accelerating deep learning hardware.[21,22] The previously proposed optical neural networks primarily involve matrix-vector multipliers (MVM),[23–26] leveraging light diffraction,[21,22,27–32] light interference,[33–36] light scattering,[37] and wavelength division multiplexing (WDM).[38–44] In these prior exploratory efforts, most of the work has predominantly focused on constructing fully connected neural networks through optical means. Existing optical convolutional neural networks predominantly follow two technological approaches. The first approach involves decomposing the overall convolution operation into numerous multiply-accumulate (MAC) operations within local kernel windows,[45–47] similar to Graphics Processing Units (GPUs). The photonic convolution is implemented as matrix-vector multiplication through preprocessing or decomposition into elementary MAC operations, using optical elements such as Mach–Zehnder interferometer (MZI),[21,23–25] micro-ring resonators (MRR),[22,42,43,48] time–wavelength interleaving,[41] frequency convolution,[49] phase-change materials,[38] and acousto-optical modulators array.[50] The number of optical elements increases quadratically as the dimensions of the input vectors increase ($O(N^2)$). In addition, phase controls are required between the optical elements, further increasing the number of active controls and

J. Cheng, J. Dai, J.-J. He
State Key Laboratory of Extreme Photonics and Instrumentation
Centre for Integrated Optoelectronics
College of Optical Science and Engineering
Zhejiang University
Hangzhou 310027, China
E-mail: jjhe@zju.edu.cn
C. Li, Y. Chu, X. Niu, X. Dong
Huawei Technologies Co., Ltd.
Bantian, Longgang, Shenzhen, Guangdong 518000, China
E-mail: xiaowen.dong@huawei.com
The ORCID identification number(s) for the author(s) of this article can be found under https://doi.org/10.1002/lpor.202301221

DOI: 10.1002/lpor.202301221

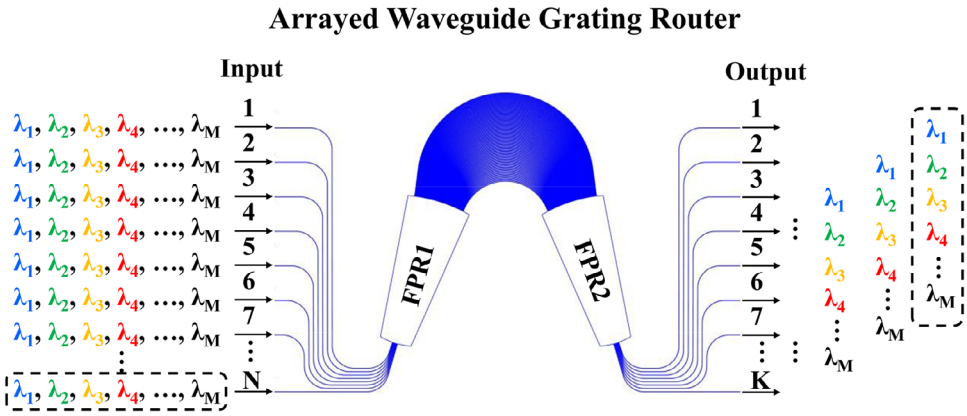

Figure 1. Schematic diagram of a N × K AWGR wavelength router. A distinctive feature of the AWGR is its "slide property," wherein the transition of an optical signal to an adjacent input port corresponds to an equivalent shift in the corresponding output port by the same number of channels.

the complexity of the computing algorithm. This increases the system complexity, power consumption, as well as the chip size. Multiwavelength comb sources in combination with WDM demultiplexers have been used to increase the processing parallelism in the MVM-based optical convolution methods.[38–40] The second approach converts the convolution operation into a multiply operation in the Fourier domain using the convolution theorem, with two Fourier transform processes.[30,51–53] It can reduce the required number of modulators to a linear relationship (O(N)) with the dimension of the input vector, as opposed to a quadratic relationship.[20,22] However, one of the input vectors cannot be directly loaded. Instead, its Fourier transform needs to be precalculated in the electrical domain and loaded on to the modulator array in the Fourier plane. The two on-chip Fourier transform components with the required coherent phase control complicates the calibration process and limits the scalability of the chip.[30] Therefore, most of current efforts face challenges in achieving both high efficiency and high scalability to support large network sizes.[54]

To address the above issues, we propose a direct optical convolution unit (OCU) based on arrayed waveguide grating router (AWGR).[55] The common arrayed waveguide grating (AWG) is a 1 × N device, and is widely used in telecom systems for multiplexing N wavelength signals from N transmitters into a single output fiber, or for demultiplexing N wavelength signals from an input fiber to N receivers. The AWGR is a N × N device that can be used for wavelength-routing and distributed optical switching in WDM systems.[56–60] Leveraging the unique sliding property in the wavelength–port relationship of the AWGR, which matches perfectly the sliding window operation of the convolution computations, our proposed OCU performs the convolution computation directly in a single step, without the need for decomposition into elementary MAC operations as in conventional MVM-based methods. It has striking advantages including high scalability, high speed and processing simplicity. The convolution is performed in the wavelength–space domain so that the entire convolution computing is executed in a single clock cycle. As soon as the two input vectors are loaded onto two modulator arrays, the convolution result is produced at the output photodetector array instantaneously at the speed of light propagation. The OCU maintains a linear relationship between the number of modulators and the dimension of the input vectors, thus exhibiting high scalability. Besides, the completely passive AWGR in combination with a minimal number of active elements required only for input/output make the system highly energy efficient. No selective summation operation is needed in the electronic or optical domain. The computational efficiency and scalability are thus much higher than those of optical matrix-vector multipliers based on cascaded MZI or MRR. In the proof-of-concept experiment, utilizing the proposed optical convolution unit we constructed a LeNet-5[3] network to perform the standard handwritten digit classification task.[61] With an output precision of four-bit, we achieved a recognition accuracy of 98.2%, which is comparable to electronic computers. This result demonstrates the superiority of our proposed OCU paradigm and its potential to address the current efficiency and scalability issues for achieving large-scale neural networks.

2. Principle

The basic structure of the N × N AWGR is shown in **Figure 1**, including N input waveguides, an input star coupler (i.e., the first free propagation region, FPR1), a grating composed of an array of waveguides with equal length differences, an output star coupler (FPR2), and *N* output waveguides. When the optical signal is transmitted from an input waveguide to the input FPR1, the light is diverged due to diffraction and then coupled to the arrayed waveguides. After propagating though the arrayed waveguide grating, the light from each arrayed waveguide is diffracted in the output FPR2, and then focused onto a specific position on the imaging surface according to the wavelength, due to the interference effect. Consequently, the light of different wavelengths is coupled into different output waveguides.

The AWGR represents a pivotal device for the efficient routing of optical signals at varying wavelengths from a solitary input port to multiple output ports, as depicted in Figure 1, where *M* represents the number of wavelengths, *N* represents the number of input ports of the AWGR and *K* is the number of output ports of the AWGR. For executing full convolution computation, these numbers satisfy the relationship $K = M + N - 1$. A distinctive feature of the AWGR is its "slide property", wherein the transition of the multi-wavelength input signals from one input port

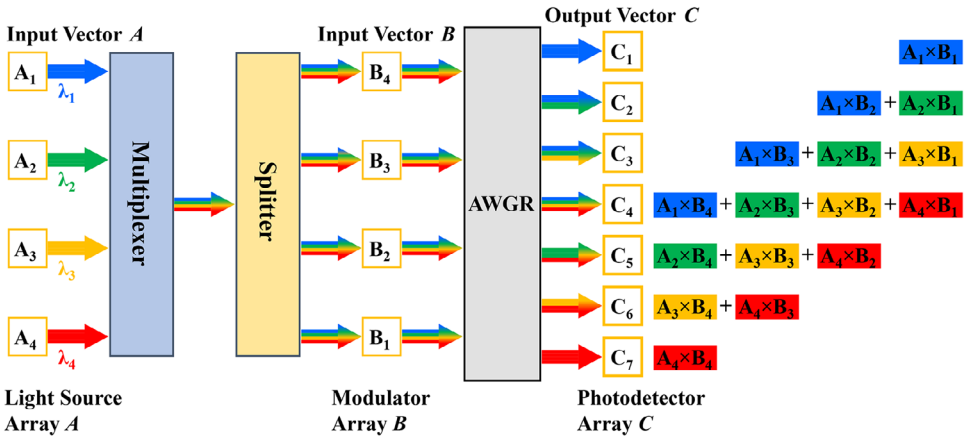

Figure 2. Optical Convolution Unit (OCU). Vector ***A*** is encoded into the intensities of optical signals of various wavelengths from Light Source Array A and vector ***B*** is loaded onto modulator array B, placed at the input ports of the AWGR. The optical vector–vector convolution results, denoted as ***C***, are produced instantaneously across the AWGR output ports and can be converted to digital electrical signal within one clock cycle.

to an adjacent input port results in an equivalent shift in the corresponding output ports by the same number of channels. This property closely mirrors the concept of sliding window operation in vector–vector convolution computations. Building upon this pivotal sliding property of the AWGR for wavelength routing, we propose a novel OCU for efficiently executing the convolution computing in the optical domain. As shown in **Figure 2**, within the OCU framework, vector ***A*** is encoded into intensity signals of different wavelengths through a directly modulated or externally modulated light source array denoted as A. Simultaneously, vector ***B*** is loaded onto another modulator array denoted as B. All the signals of different wavelengths of vector ***A*** are combined and then sent to each modulator (designated as B_j) of the modulator array B through a power splitter. Each modulator multiplies the vector element B_j to its input signals of different wavelengths carrying vector ***A***, and then send them to the corresponding input port of the AWGR. As shown by the dashed frames in Figure 1, the AWGR demultiplexes the signals of different wavelengths from an input port to different output ports with the above-mentioned space invariant sliding property.

Simultaneously, as illustrated by the color-coded product terms at the right-hand side of Figure 2, it also multiplexes signals of different wavelengths from different input ports to each output port according to its wavelength routing property, to be detected by a photodetector, which performs the summation operation when converting the optical signals to an electrical signal. The results of the entire vector–vector convolution, labeled as vector ***C***, are thus obtained at the output of the photodetector array. In the context of full convolution between two vectors with N elements, the output vector ***C*** attains a length of $2N-1$. It is noteworthy that the optical power detected at each output port of the wavelength router is directly proportional to the convolution of vectors ***A*** and ***B***

$$C_k = A \otimes B = \sum_{i+j=k-1} A_i B_j, i = 1, 2, \ldots, M, j = 1, 2, \ldots, N \quad (1)$$

which represents the sum of the dot product between two input vectors with different displacements corresponding to different sliding positions. The distinctive routing property of the AWGR enables the implementation of a sliding window operation for convolution. This inherent characteristic establishes a linear correlation between the number of modulators and the dimension of the computation vector, as opposed to a quadratic relationship (N^2), thus rendering our OCU remarkably scalable.

Figure 3a provides a detailed experimental demo setup of the proposed OCU paradigm, which comprises a laser array, electro–optic modulator (EOM) array A, MUX (Wavelength-Division Multiplexer), splitter, EOM array B, AWGR, a photodetector (PD) array integrated with a transimpedance amplifier (TIA) circuit, and a field-programmable gate array (FPGA) driver board equipped with digital-to-analog converter (DAC) and analog-to-digital converter (ADC). The AWGR used in the experiment is fabricated on silica-on-silicon platform. It features a configuration with 32 input and 32 output ports, enabling it to execute vector convolution calculations with a maximum length of 16 for both vectors. Operating at a central wavelength of 1550 nm, the 32 × 32 AWGR characterizes a channel spacing of 100 GHz (0.8 nm). The AWGR module used in the current experiment includes a built-in heater to ensure the wavelength stability of its channels, which can potentially be eliminated by using an athermal design to reduce the power consumption. A photograph of the AWGR chip is shown in Figure 3b. It has a footprint of 25 × 10 mm^2 with curved profile. Figure 3c shows the measured transmission spectra of the AWGR for input port #17. The 32 × 32 AWGR is designed with cyclic characteristics, with a free spectral range (FSR) equal to 3200 GHz (25.6 nm). The insertion loss for the central channel is about 3.5 dB, with a channel non-uniformity of 2.5 dB. Notably, the observed crosstalk is below −31 dB.

The laser array employed in experiment produces four distinct wavelength sat 1548.52, 1547.72, 1546.92, and 1546.12 nm, aligned with four consecutive wavelength channels of the 32 × 32 AWGR. The lasers are coupled to lithium niobate ($LiNbO_3$) EOM array ***A*** through polarization maintaining fibers, then combined by an AWG multiplexer (MUX). The aggregated multi-wavelength signals are then sent to each of the second EOM array ***B*** coupled to the input ports of the AWGR. The $LiNbO_3$ intensity modulators used in both EOM arrays are iXblue MX-LN-20, characterized by a specified maximum modulation speed of

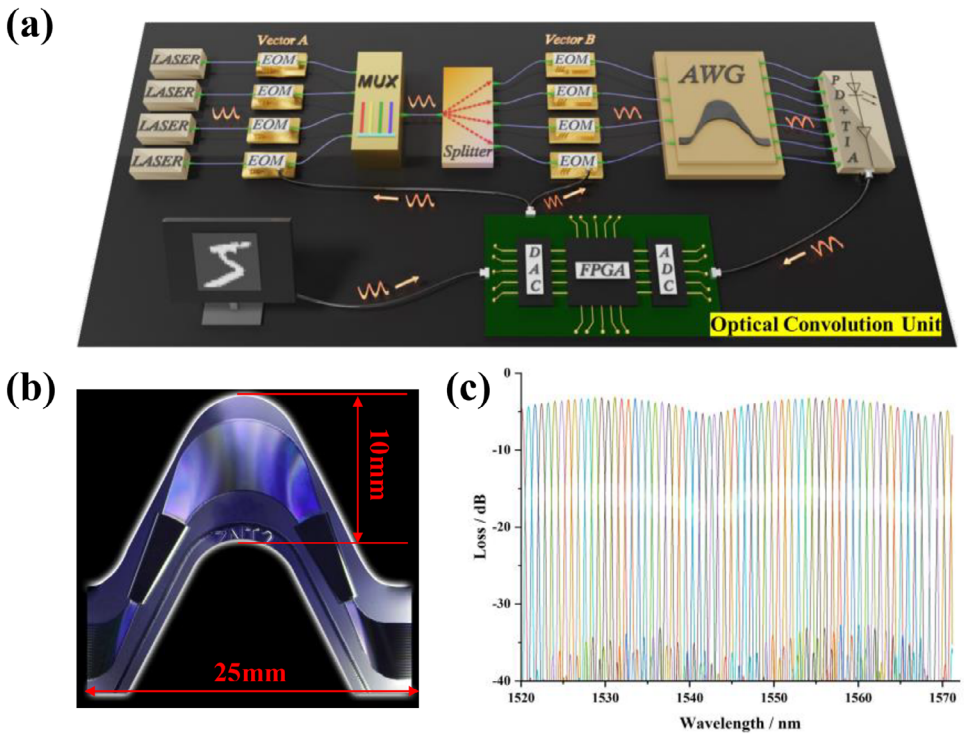

Figure 3. a) OCU experimental setup composed of a laser array, electro–optic modulator (EOM) array A, MUX (wavelength-division multiplexer), Splitter, EOM array B, 32 × 32 AWGR, PD (photodiode) and transimpedance amplifier (TIA) array, digital-to-analog converter (DAC), analog-to-digital converter (ADC), field programmable gate array (FPGA), PC. b) Microscopic image of the fabricated 32 × 32 AWGR with a footprint of 25 × 10 mm^2. c) Measured transmission spectra from one input port to 32 output ports.

20 GHz, a half-wave voltage of 7 V, and an insert loss of 4 dB. At the output of the AWGR, optical power-to-voltage conversion is executed using photodetectors (PDs) equipped with integrated TIA circuits—specifically, KY-PRM-500M-I-FC, featuring a transimpedance gain of 5000 mV μW^{-1} and a 3-dB bandwidth of 500 MHz. The input vector data and the convolution kernels are loaded onto the modulators though the DAC under FPGA control. Based on the AWGR property described above, once the data vectors are loaded onto the two modulator arrays, the vector–vector convolution result is obtained at once at the PD array and is convert to digital electrical signal within one system clock cycle by an ADC, without the need for preprocessing or decomposition into elementary MAC operations. It is worth mentioning that all the optical devices can be potentially integrated on a chip, similar to wavelength transmitter–router developed for optical switching.[60] Additionally, the input vector dimensions can be easily scaled to 32 with a 32 × 64 AWGR.

For high-speed data input/output, instead of using expensive instruments such as arbitrary waveform generator and oscilloscope with very limited channel counts,[33] we designed and fabricated a more practical FPGA-controlled drive circuit to efficiently execute vector data input/output, accommodating sequences with a maximum length of 32. This drive circuit incorporates a 64-channel DAC, enabling push–pull drive for MZI modulators, resulting in a high dynamic extinction ratio. The drive circuit also provides differential output voltages spanning from 0.3 to 2.1 V, maintaining an output voltage accuracy of up to 16 bits. A 64-channel ADC is integrated into the drive circuit, offering a precision of 10 bits with the capacity to sample voltages up to 1 V. Operating at a clock speed of 100 MHz, the entire drive circuit empowers the system with a peak computational capacity of up to $32 \times 32 \times 2 \times 100\ \mathrm{M} = 0.2$ TMACs (tera multiply-accumulate operations). By employing the state-of-the-art DAC/ADC at clock rates in the order of 10 GHz, a computing capacity beyond 20 TMACs can be accommodated.

As a proof-of-concept demonstration for the OCU proposed in this work, we constructed a 4 × 4 convolution operator based on the wavelength routing to implement an optical–electronic hybrid CNN for Modified National Institute of Standards and Technology (MNIST) dataset recognition.[61] The CNN usually consist of convolutional layers (which perform the convolutional operations), pooling layers (which reduce the size of the convoluted matrices, that is, the feature maps), nonlinear layers (which provide nonlinear activations such as rectified linear unit, or ReLU for short), and fully connected layers (which performs the task of classification based on the features extracted through the previous layers). The convolutional layers enable CNNs to possess translation-invariant characteristics so that they can identify and extract patterns and features from data irrespective of variations in position, orientation, or scale, which is crucial for object recognition. The convolutions are computationally demanding so an optical accelerator is highly desirable. The CNN employed in the OCU experiment has been implemented in accordance with the architectural framework of LeNet-5,[3] a very efficient CNN for handwritten character recognition. **Figure 4** illustrates the architecture of the neutral network,

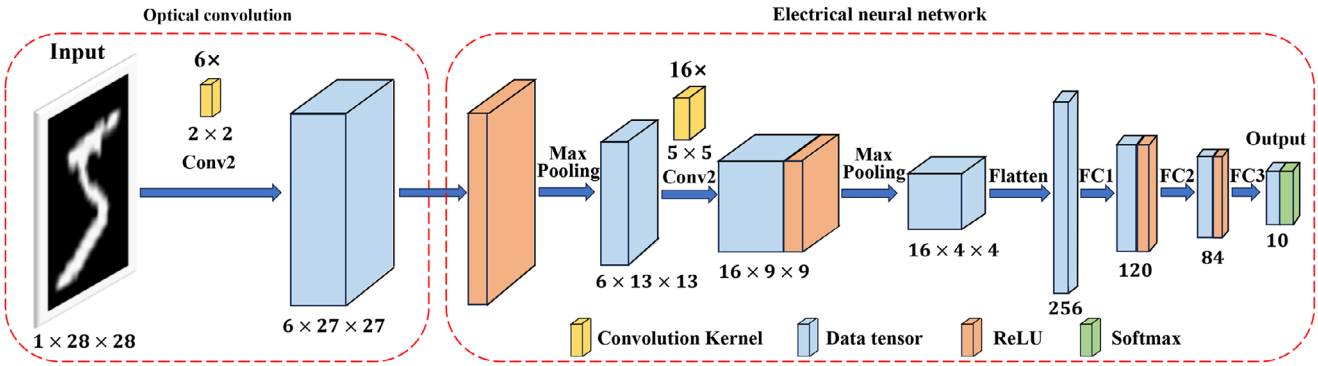

Figure 4. Framework of the optical–electronic hybrid convolutional neural network for MNIST dataset in accordance with LeNet-5.

Table 1. Neural network structure.

Layer	Type	Kernel size	Stride	No. of filters
Conv 1	**Conv2d**	**2 × 2**	**1**	**6**
Pool 1	Maxpool2d	2 × 2	2	
Conv 2	Conv2d	5 × 5	1	16
Pool 2	Maxpool2d	2 × 2	2	
FC1	FC/Linear			120
FC2	FC/Linear			84
FC3	FC/Linear			10

Note that the optical functionality is currently limited to only the first CONV layer, while the subsequent computations are executed in the electrical domain. In the future, it is possible to execute multiple layers of convolutions in the optical domain by looping back the data after pooling and nonlinear activation of each layer in the electrical domain to the input of the OCU or by cascading multiple OCU units with nonlinear optical elements for executing nonlinear activations such as ReLU operations.[26,62]

comprising two convolutional layers and three fully connected layers. The input image size for the first layer is 28 × 28. The first layer is performed by the AWGR OCU in the optical domain, featuring six convolution kernels with a size of 2 × 2. The resulting output tensor size is 6 × 27 × 27. The precise network structure is detailed in **Table 1**.

3. Results

3.1. MNIST Dataset Recognition

The MNIST dataset are split into the training (60 000 images) and testing sets (10 000 images). Our model is trained on the entire training set. To maintain compatibility with the four-bit precision of the system, the input images utilized during network training undergo quantization to match the same precision. As depicted in **Figure 5a**, after 15 training iterations, the inference accuracy stabilizes at ≈99.0%, closely approximating the theoretical accuracy limit of the neural network.

The six convolution kernels of the Conv 1 layer in the trained neural network are displayed in Figure 5b (top), along with the mathematical convolution result (middle) and the experimental optical convolution result (bottom) for an example handwriting image. Notably, the optical domain convolution results exhibit slight deviations from the mathematical outcomes. And the computation accuracy can be calculated by the deviations of the experimental optical convolution results from the mathematical convolution results. It is important to note that due to limitations in the signal-to-noise ratio of the system, the experimental convolution result achieves a four-bit accuracy. Nevertheless, the overall accuracy of the vector–vector convolution calculation reaches an impressive 96.3%. It is noteworthy that the discrepancies between the mathematical and experimental results predominantly fall within ±1, as elucidated in Figure 5c.

Subsequently, the optical convolution result is fed into the posterior network for further processing, ultimately yielding recognition results at the output layer. A subset of 512 images from the original MNIST test dataset is selected and the handwritten digit recognition experiments are conducted following the aforementioned procedure. Subsequent to obtaining the computational results from the optical domain convolution layer, the ensuing image processing is conducted directly in the electrical domain, maintaining the parameters of the remaining network layers unaltered. A recognition accuracy of 92.8% is achieved on this randomly selected set of 512 images. The confusion matrix is depicted in Figure 5d, where out of a total of 512 images, 475 were precisely inferred and correctly identified. Note that this performance falls slightly short compared to an electronically trained LeNet-5 network (99.0%). The observed reduction in accuracy during network inference is primarily attributed to the noise within the preliminary demo system and the constrained precision four-bit of convolution calculations within optical domain.

3.2. Fine-Tuning Training

To enhance the performance of the neural network, a fine-tuning is performed in the training process. The feature map of the 512 images is divided into re-training and testing dataset in a ratio of 400:112. Out of these, 400 images are designated for re-training within the electrical domain, emphasizing the adjustment of parameters in the remaining network layers to mitigate the influence of noise on convolution calculation accuracy. During the fine-tuning process, the parameters of the Conv 1 layer were kept unchanged, allowing only the parameters of the subsequent layers to undergo fine-tuning retraining. After this fine-tuning

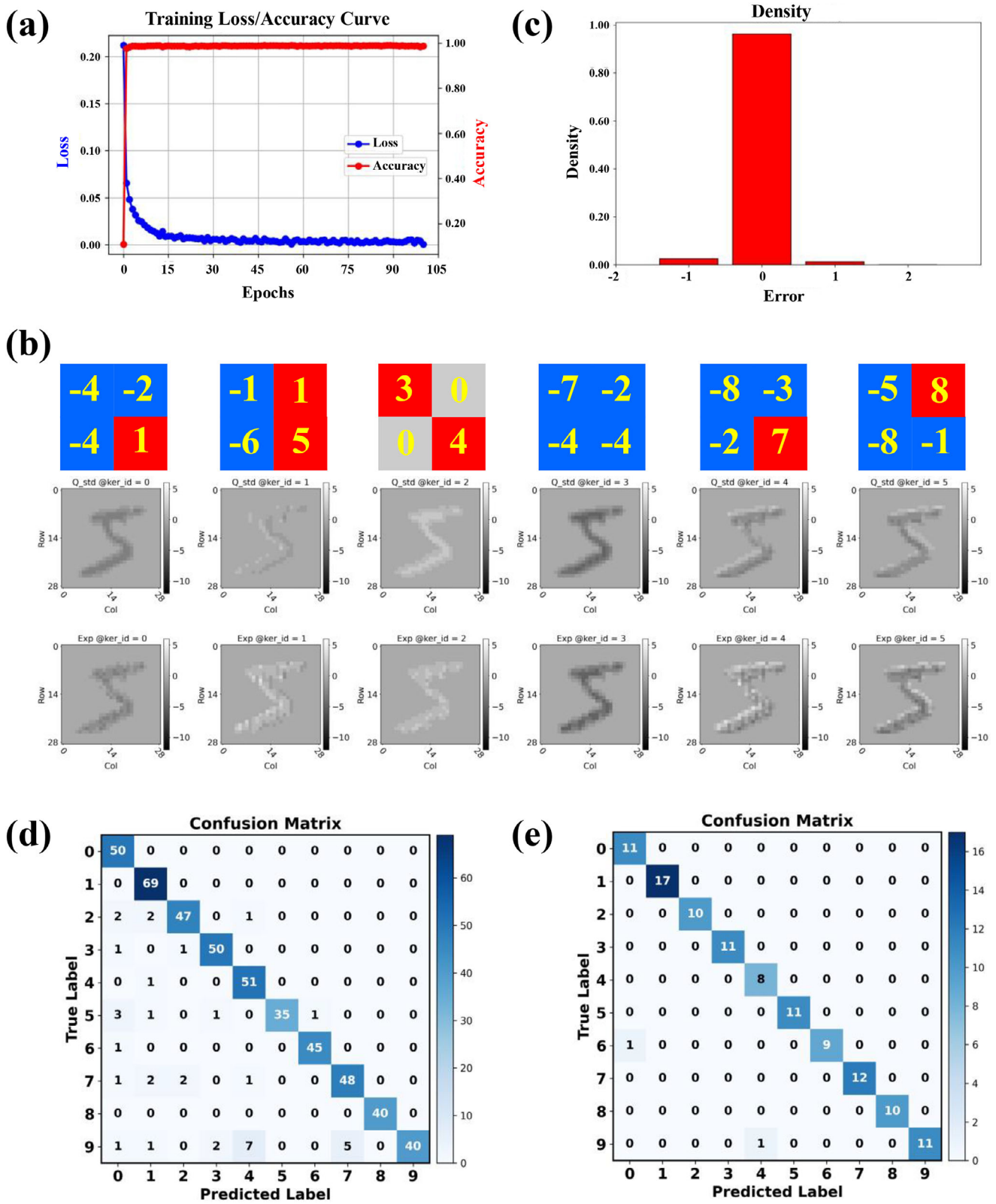

Figure 5. a) The simulation accuracy of test dataset and network inference loss during 100 epochs of training. b) Six convolution kernels (top), math results (middle), and experimental results (bottom). c) Error density chart showing the vector–vector convolution calculation accuracy of the 4 × 4 OCU based on AWGR is 96.3%. d) Confusion matrix corresponding to 92.8% classification inference accuracy for a directly trained network. e) Improved confusion matrix with additional fine-tuning in the training process, showing 98.2% inference accuracy for the MNIST dataset.

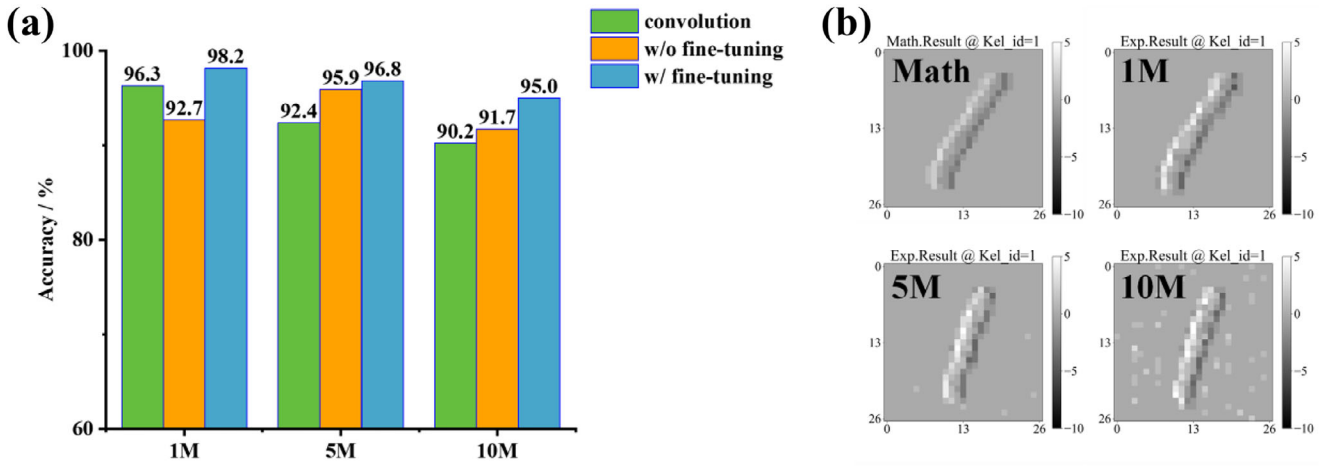

Figure 6. a) Measured accuracy of convolution calculations, and recognition accuracies for trainings with and without fine-tuning, under different data transferring speed. b) Comparison of mathematical image of convolution results under 1, 5, and 10 MHz.

process, the network achieved a recognition accuracy of 98.2% on the testing dataset (112 images). The final confusion matrix is presented in Figure 5e. This level of performance is very close to the inference performance achieved on electronic computers, indicating that the proposed optical convolution paradigm possesses computational capabilities that are on par with those of electronic computers (99.0%).

Due to the constraint of available lasers and modulators, we used 2 × 2 kernel size in our first experimental demonstration of a 4 × 4 OCU based on discrete components. The 2 × 2 convolution kernel size suffices to implement relatively simple Lenet-5 neural network, to showcase the feasibility of our proposed OCU based on AWGR wavelength routing. We are currently working on an integrated Photonic Integrated Circuit (PIC) version of the OCU with much larger vector dimensions so that larger kernel sizes can be implemented. This will allow more complex network architectures such as ResNet to be demonstrated with larger datasets such as Fashion MNIST.

3.3. Accuracy versus System Noise under Varying Data Speed

The entire 4 × 4 OCU system operates on a clock frequency of 100 MHz. To achieve noise suppression, we initially represent one single user data point with a sequence spanning 100 system clock cycles (equivalent to 1 MHz data transferring speed), and average the 100 cycles at the ADC output. This meticulous approach has allowed us to achieve a network inference accuracy of 98.2% on the MNIST dataset. To study the impact of noises under different data transferring rates, we use a representation scheme equivalent to 5 MHz (where 20 clock cycles represent one data point) and 10 MHz (where ten clock cycles represent one data point) for transmitting user data. In our comprehensive evaluation, we compared the accuracy of the vector–vector convolution calculations, and the recognition accuracies with and without fine-tuning in the training process under varying data transmission rates.

As illustrated in **Figure** 6a, the network inference accuracies without the fine-tuning training are 92.7% (1 MHz), 95.9% (5 MHz), and 91.7% (10 MHz), respectively. At a frequency of 1 MHz, the recognition accuracy stood at 98.2% after fine-tuning training. As the data transmission speed is increased to 5 MHz, we observe a marginal decrease in accuracy, resulting in a still commendable 96.8%. When operating at 10 MHz, there is a notable decline in accuracy to 95.0%. This observed trend aligns with our anticipated outcome, as heightened system noise levels invariably lead to diminished accuracy, as shown in Figure 6b. It is important to note that the corresponding vector–vector convolution calculation accuracy exhibited a similar diminishing trend, from 96.3% (1 MHz) to 92.4% (5 MHz) and 90.2% (10 MHz).

The accuracy deviations between different data transferring rates reflect the limited bit precision due to system noises, which are caused by numerous factors, including the electrical and optical noises, and instability of some optical devices such as the temperature drift and mechanical disturbances, and polarization sensitivity of the EOMs. By averaging multiple samples with increased ADC sampling frequency, the accuracy of the convolution calculation can be significantly improved. The noises can also be mitigated with high-degree photonic integration[63–65] and improved drive circuit, which aligns with the objective of our on-going effort. Advanced photonic–electronic co-packaging technologies[66] can also reduce high-frequency parasitic crosstalk noises and minimize signal losses at elevated frequencies. We believe that it is possible to improve the current four-bit accuracy to eight-bit, thereby meeting the demand of a more extensive array of AI applications.

3.4. Prospect of Photonic Integration

It is worth noting that while the current experimental system is based on an AWGR chip with many standalone components, all the optical components have the potential to be integrated onto a single silicon photonic chip with hybrid integrated light sources. Apart from light sources, diverse optical devices including multiplexers, power splitters, modulators, AWGR, and germanium (Ge) photodetectors can be integrated on silicon-on-insulator (SOI) platform. The integration of lasers on-chip

can be achieved through heterogeneous integration,[63] hybrid integration solutions,[64] or flip-chip bonding.[65] Such photonic integration with advanced optical–electronic co-packaging technologies offers significant advantages in terms of reductions in both chip size and power consumption, in addition to much increased system scalability and stability. For example, a SOI-based 32 × 32 AWGR from ref. [67] has a footprint of only about 1.1 mm × 2.35 mm. Furthermore, the AWGR can be replaced by a more compact echelle diffraction grating router, and an intra-cavity Etched Diffraction Grating (EDG) can be used to make a self-aligned multi-wavelength laser instead of the combination of the laser array of different wavelengths and the wavelength multiplexer,[68] to further reduce the chip size and device count. A cascaded MZI structure with the same number of input/output ports requires 1984 MZI's and has a chip size of 24 mm × 18 mm for achieving 32 × 32 optical switch.[69] The chip size for constructing a 32 × 32 matrix-vector multiplier would be even much bigger considering the phase shifters needed at each MZI input/output port.[21] Similarly, for convolution calculation based on MRR, it is observed that a total of 1024 rings are necessary to construct a 32 × 32 matrix-vector multiplier.[48] Such a large number of devices would consume a significant amount of power since each of them needs to be thermally tuned. Take the more efficient MRR approach as an example. The tuning efficiency of a typical SOI-based MRR is 27.53 mW/FSR.[70] Considering the average tuning of half FSR, the power consumption for thermal tuning of a single MRR amounts to 13.77 mW. For a 32 × 32 MVM, the total power consumption for the thermal tuning is assessed at 14.1 W. In comparison, the tuning efficiency of a SOI-based AWGR is typically measured at 7.5 nm W^{-1}.[71] The required maximum power consumption is 0.43 W for tuning a channel. It is evident that the thermal tuning power consumption of the AWGR is substantially lower than that of the MRR array. Furthermore, upon integrating a laser array on-chip, a more efficient strategy is to thermally tune the wavelengths of the lasers to align with the AWGR channels, resulting in a substantial reduction in the total power consumption. We are currently implementing a 32 × 64 AWGR based OCU on the SOI platform. The total chip size is only 3.6 mm × 2.8 mm, including a 32 × 64 AWGR, a 1 × 32 splitter, 32 MZIs and 64 PDs/grating couplers for data input/output, etc. More detailed comparisons of the size, power consumption and other metrics will be highlighted in the subsequent work with experimental results.

In the wavelength routing based OCU proposed in this article, the speed of convolution computing is primarily constrained by the clock frequencies of the DAC and ADC circuits. Elevating the clock frequency through the entire driver system would yield an exceptionally high computing power. By employing a 128 × 128 OCU with 10 GHz DAC/ADC, for example, a computing power of 327.68 TMACs can be attained. This computing capacity surpasses Nvidia GPU A100, but only requires a 130 nm silicon photonic fab instead of 7 nm Integrated Circuit (IC) technology. To sum up, our proposed optical convolution computing paradigm offers several key advantages compared to previously reported schemes based on matrix-vector multipliers:

(a) Processing simplicity: Two input vectors are directly loaded onto two modulator arrays, one for application data and the other for kernel or weight matrix, no pre-processing is needed at the input side, either electronically or optically. At the output side, each vector element of the convolution result is obtained directly at the photodetector array, without the need for regrouping or selective summation operation in the electronic or optical domain.

(b) High speed: The vector–vector convolution computation is executed within a single clock cycle without the need for preprocessing or decomposition into elementary MAC operations as in conventional MVM-based methods. The sliding operation is performed in the wavelength–space domains instead of the time-space domains. The multiply and accumulate operations corresponding to different sliding window positions are executed simultaneously at different modulators and photodetectors at the input and output ports of the AWGR, respectively. This high degree of parallelism significantly boosts the operational speed of the system, making it suitable for real-time applications.

(c) High scalability: The AWGR executes the convolution operation in both space and wavelength domains. The number of elements in the two input vectors is mapped to the number of input wavelengths and the number of input ports on the AWGR, respectively. This 2D parallel processing feature results in $2N$ as opposed to N^2 scalability. Both the number of AWGR ports and the number of modulators/detectors scale linearly with the vector dimension N.

(d) Low device count: For 32 × 32 vector convolution, for example, only a single AWGR is required, as opposed to thousands of MZIs or MRRs with even more control electrodes in the case of vector-matrix multipliers. This reduction in device count can lead to more compact chip size and high computing power density.

(e) Low power consumption: The AWGR is completely passive, while the number of active elements such as MZIs and PDs is minimal, which are only needed for data input/output. The energy efficiency is thus much higher than those of optical matrix-vector multipliers based on cascaded MZIs or MRRs. Although an integrated heater is usually required for the AWGR to ensure wavelength stability, in the case of integrated OCU chip, only a single temperature controller is needed for the entire chip. This is in contrast to MZI/MRR based optical convolution architectures, where each of the numerous MZI/MRR devices requires a heater for fine tuning and stabilization, in additional to a common temperature controller to stabilize the system including other components such as lasers and grating couplers.

(f) High reconfigurability: Since the kernel or weight matrix is loaded onto a high-speed modulator array, the system allows for reconfigurability at high speed as needed.

(g) High accuracy: Computational efficiency of the OCU and the algorithm allowed us to achieve a high inference accuracy of 98.2%, even with an ADC output accuracy of four-bit constrained by system noises. Reducing the system noises can further improve the accuracy of the convolution operation and enhance the data throughput.

(h) Low requirements on fabrication technology: A silicon photonic OCU accelerator based on the wavelength routing can potentially reach similar computing capacity and computing

power density to state-of-the-art GPUs, but only needs 130 nm processing node, as opposed to 7 nm or below.

4. Conclusion

In this work, we have proposed a direct optical convolution computing architecture based on wavelength routing, which bring forth a range of significant advantages including high scalability, high speed, processing simplicity, minimized device counts, and high efficiency. Building upon this novel architecture, image recognition based on a 4 × 4 optical convolution unit has been demonstrated, for the first time, with a DAC/ADC based high speed electronic driver developed in-house. Owing to multi-dimensional parallelism, the convolution operation can be obtained within one single system clock cycle, without decomposition into numerous MAC operations and execution in time sequence. The convolution results have been directly subjected to inference on MNIST dataset by a trained neural network implemented in accordance with the framework of Lenet-5, achieving an accuracy rate of 92.8%. Through meticulous fine-tuning training, it has been shown that the inference accuracy can be further elevated to an exceptional 98.2%. It is the first time that the AWGR is demonstrated for direct optical convolution computing with potentially much superior characteristics compared to other optical computing systems reported in the literature. Our proposed optical convolution computing paradigm exhibits promising potential for large-scale photonic integration. These developments will lay the foundation for the next generation ultra-high–speed artificial intelligence platforms.

Acknowledgements

J.C. and C.L. contributed equally to this work. This work was supported by the National Key Research and Development Program of China (2021YFB2801700). The authors thank Huimin Yan for her help in preparing some of the figures, and Zhiqiang Yun for his help in experimental setup. The authors at ZJU acknowledge the support by Huawei through ZJU-Huawei Center for Innovation on Optical Computing.

Conflict of Interest

The authors declare no conflict of interest.

Data Availability Statement

The data that support the findings of this study are available from the corresponding author upon reasonable request.

Keywords

arrayed waveguide grating router, handwritten digit recognition, LeNet-5 neural network, optical convolution computing

Received: November 21, 2023
Revised: March 1, 2024
Published online:

[1] Y. LeCun, Y. Bengio, G. Hinton, *Nature* **2015**, *521*, 436.
[2] Y. Bengio, I. Goodfellow, A. Courville, in *Deep Learning*, Vol. *1*, MIT press, Cambridge, MA **2017**.
[3] Y. LeCun, L. Bottou, Y. Bengio, P. Haffner, *Proc. IEEE* **1998**, *86*, 2278.
[4] N. Aloysius, M. Geetha, presented at Int. Conference on Communication and Signal Processing (ICCSP), Chennai, India, April **2017**.
[5] A. Krizhevsky, I. Sutskever, G. E. Hinton, presented at Advances in Neural Information Processing Systems (NIPS), Nevada, USA, December **2012**.
[6] K. Simonyan, A. Zisserman, arXiv:1409.1556, **2015**.
[7] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, presented at IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), Massachusetts, USA, October **2015**.
[8] K. He, X. Zhang, S. Ren, J. Sun, presented at IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), Nevada, USA, June **2016**.
[9] R. Girshick, J. Donahue, T. Darrell, J. Malik, *IEEE Trans. Pattern Anal. Mach. Intell.* **2015**, *38*, 142.
[10] R. Girshick, presented at IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), Massachusetts, USA, October**2015**.
[11] S. Ren, K. He, R. Girshick, J. Sun, presented at Advances in Neural Information Processing Systems (NIPS), Montréal, Canada, December **2015**.
[12] J. Redmon, S. Divvala, R. Girshick, A. Farhadi, presented at IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), Nevada, USA, June **2016**.
[13] L.-C. Chen, M. Collins, Y. Zhu, G. Papandreou, B. Zoph, F. Schroff, H. Adam, J. Shlens, presented at Advances in Neural Information Processing Systems (NIPS), Montréal, Canada, December **2018**.
[14] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, H. Adam, presented at European Conf. on Computer Vision (ECCV), Munich, Germany, September **2018**.
[15] K. He, G. Gkioxari, P. Dollár, R. Girshick, presented at IEEE Int. Conf. on Computer Vision (ICCV), Venice, Italy, October **2017**.
[16] H. Zhao, J. Shi, X. Qi, X. Wang, J. Jia, presented at IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), Hawaii, USA, July **2017**.
[17] S.-Y. Wang, O. Wang, R. Zhang, A. Owens, A. A. Efros, presented at IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR), Washington, USA, June **2020**.
[18] A. Creswell, T. White, V. Dumoulin, K. Arulkumaran, B. Sengupta, A. A. Bharath, *IEEE Signal Process. Mag.* **2018**, *35*, 53.
[19] J. Ho, A. Jain, P. Abbeel, presented at Advances in Neural Information Processing Systems (NIPS), Jiangsu, China, December **2020**.
[20] H. J. Caulfield, J. Kinser, S. K. Rogers, *Proc. IEEE* **1989**, *77*, 1573.
[21] Y. Shen, N. C. Harris, S. Skirlo, M. Prabhu, T. Baehr-Jones, M. Hochberg, X. Sun, S. Zhao, H. Larochelle, D. Englund, M. Soljačić, *Nat. Photonics* **2017**, *11*, 441.
[22] V. Bangari, B. A. Marquez, H. Miller, A. N. Tait, M. A. Nahmias, T. F. De Lima, H.-T. Peng, P. R. Prucnal, B. J. Shastri, *IEEE J. Sel. Top. Quantum Electron.* **2020**, *26*, 7701213
[23] X. Wang, P. Xie, B. Chen, X. Zhang, *Nano-Micro Lett.* **2022**, *14*, 221.
[24] H. Zhou, Y. Zhao, X. Wang, D. Gao, J. Dong, X. Zhang, *ACS Photonics* **2020**, *7*, 792.
[25] H. Zhou, Y. Zhao, G. Xu, X. Wang, Z. Tan, J. Dong, X. Zhang, *IEEE J. Sel. Top. Quantum Electron.* **2019**, *26*, 8300910.
[26] Z. Chen, A. Sludds, R. Davis III, I. Christen, L. Bernstein, L. Ateshian, T. Heuser, N. Heermeier, J. A. Lott, S. Reitzenstein, R. Hamerly, D. Englund, *Nat. Photonics* **2023**, *17*, 723.
[27] X. Lin, Y. Rivenson, N. T. Yardimci, M. Veli, Y. Luo, M. Jarrahi, A. Ozcan, *Science* **2018**, *361*, 1004.
[28] Y. Zuo, B. Li, Y. Zhao, Y. Jiang, Y.-C. Chen, P. Chen, G.-B. Jo, J. Liu, S. Du, *Optica* **2019**, *6*, 1132.

[29] Y. Luo, D. Mengu, N. T. Yardimci, Y. Rivenson, M. Veli, M. Jarrahi, A. Ozcan, *Light: Sci. Appl.* **2019**, *8*, 112.
[30] H. H. Zhu, J. Zou, H. Zhang, Y. Z. Shi, S. B. Luo, N. Wang, H. Cai, L. X. Wan, B. Wang, X. D. Jiang, J. Thompson, X. S. Luo, X. H. Zhou, L. M. Xiao, W. Huang, L. Patrick, M. Gu, L. C. Kwek, A. Q. Liu, *Nat. Commun.* **2022**, *13*, 1044.
[31] T. Fu, Y. Zang, Y. Huang, Z. Du, H. Huang, C. Hu, M. Chen, S. Yang, H. Chen, *Nat. Commun.* **2023**, *14*, 70.
[32] Y. Chen, M. Nazhamaiti, H. Xu, Y. Meng, T. Zhou, G. Li, J. Fan, Q. Wei, J. Wu, F. Qiao, L. Fang, Q. Hai, *Nature* **2023**, *623*, 48.
[33] X. Meng, G. Zhang, N. Shi, G. Li, J. Azaña, J. Capmany, J. Yao, Y. Shen, W. Li, N. Zhu, M. Li, *Nat. Commun.* **2023**, *14*, 3000.
[34] H. Zhang, M. Gu, X. D. Jiang, J. Thompson, H. Cai, S. Paesani, R. Santagati, A. Laing, Y. Zhang, M. H. Yung, Y. Z. Shi, F. K. Muhammad, G. Q. Lo, X. S. Luo, B. Dong, D. L. Kwong, L. C. Kwek, A. Q. Liu, *Nat. Commun.* **2021**, *12*, 457.
[35] T. W. Hughes, M. Minkov, Y. Shi, S. Fan, *Optica* **2018**, *5*, 864.
[36] J. Cheng, W. Zhang, W. Gu, H. Zhou, J. Dong, X. Zhang, *ACS Photonics* **2023**, *10*, 2173.
[37] Y. Qu, H. Zhu, Y. Shen, J. Zhang, C. Tao, P. Ghosh, M. Qiu, *Sci. Bull.* **2020**, *65*, 1177.
[38] 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, H. Bhaskaran, *Nature* **2021**, *589*, 52.
[39] B. Shi, N. Calabretta, R. Stabile, *IEEE J. Sel. Top. Quantum Electron.* **2020**, *26*, 7701111.
[40] B. Shi, N. Calabretta, R. Stabile, *IEEE J. Sel. Top. Quantum Electron.* **2023**, *29*, 7400310.
[41] 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, D. J. Moss, *Nature* **2021**, *589*, 44.
[42] S. Ohno, R. Tang, K. Toprasertpong, S. Takagi, M. Takenaka, *ACS Photonics* **2022**, *9*, 2614.
[43] J. Cheng, Y. Zhao, W. Zhang, H. Zhou, D. Huang, Q. Zhu, Y. Guo, B. Xu, J. Dong, X. Zhang, *Front Optoelectron* **2022**, *15*, 15.
[44] B. Dong, S. Aggarwal, W. Zhou, U. E. Ali, N. Farmakidis, J. S. Lee, Y. He, X. Li, D.-L. Kwong, C. D. Wright, W. H. P. Prenice, H. Bhaskaran, *Nat. Photonics* **2023**, *17*, 1080.
[45] W. Shi, Z. Huang, H. Huang, C. Hu, M. Chen, S. Yang, H. Chen, *Light: Sci. Appl.* **2022**, *11*, 121.
[46] M. J. Filipovich, Z. Guo, M. Al-Qadasi, B. A. Marquez, H. D. Morison, V. J. Sorger, P. R. Prucnal, S. Shekhar, B. J. Shastri, *Optica* **2022**, *9*, 1323.
[47] C. Wu, H. Yu, S. Lee, R. Peng, I. Takeuchi, M. Li, *Nat. Commun.* **2021**, *12*, 96.
[48] B. Bai, Q. Yang, H. Shu, L. Chang, F. Yang, B. Shen, Z. Tao, J. Wang, S. Xu, W. Xie, W. Zou, W. Hu, J. E. Bowers, X. Wang, *Nat. Commun.* **2023**, *14*, 66.
[49] L. Fan, Z. Zhao, K. Wang, A. Dutt, J. Wang, S. Buddhiraju, C. C. Wojcik, S. Fan, *Phys. Rev. Appl.* **2022**, *18*, 034088.
[50] S. Xu, J. Wang, R. Wang, J. Chen, W. Zou, *Opt. Express* **2019**, *27*, 19778.
[51] M. Ahmed, Y. Al-Hadeethi, A. Bakry, H. Dalir, V. J. Sorger, *Nanophotonics* **2020**, *9*, 4097.
[52] J.-J. He, J. Guo, X. Dong, *U.S. 17/582,164*, **2022**.
[53] L. Fan, X. Long, J. Dai, C. Li, X. Dong, J.-J. He, *Appl. Opt.* **2023**, *62*, 1337.
[54] C. Li, X. Zhang, J. Li, T. Fang, X. Dong, *PhotoniX* **2021**, *2*, 20.
[55] Y. Chu, X. Dong, U.S. 18/362 200 **2023**
[56] C. Dragone, *IEEE Photonics Technol. Lett.* **1991**, *3*, 812.
[57] S. B. Yoo, *J. Lightwave Technol.* **2021**, *40*, 2214.
[58] J. Guo, Z. Fan, S. Zhou, B. Liu, J. Meng, J. Zhu, Y. Li, Q. Li, J. Zhao, J.-J. He, *Opt. Lett.* **2022**, *47*, 2762.
[59] Y. Chu, Q. Chen, Z. Fan, J.-J. He, *IEEE Photonics Technol. Lett.* **2019**, *31*, 943.
[60] Z. Fan, J. Guo, S. Zhang, J. Zhu, J. Meng, Q. Li, Y. Li, J. Zhao, J.-J. He, *IEEE Photonics J.* **2022**, *14*, 6642508.
[61] L. Deng, *IEEE Signal Process. Mag.* **2012**, *29*, 141.
[62] Y. Li, Y. Yuan, presented at Advances in Neural Information Processing Systems (NIPS), California, USA, December **2017**.
[63] C. Xiang, J. Liu, J. Guo, L. Chang, R. N. Wang, W. Weng, J. Peters, W. Xie, Z. Zhang, J. Riemensberger, J. Selvidge, T. J. Kippenberg, J. E. Bowers, *Science* **2021**, *373*, 99.
[64] C. Xiang, W. Jin, O. Terra, B. Dong, H. Wang, L. Wu, J. Guo, T. J. Morin, E. Hughes, J. Peters, Q.-X. Ji, A. Feshali, M. Paniccia, K. J. Vahala, J. E. Bowers, *Nature* **2023**, *620*, 78.
[65] T. Matsumoto, T. Kurahashi, R. Konoike, K. Suzuki, K. Tanizawa, A. Uetake, K. Takabayashi, K. Ikeda, H. Kawashima, S. Akiyama, S. Akiyama, *J. Lightwave Technol.* **2018**, *37*, 307.
[66] C. Minkenberg, R. Krishnaswamy, A. Zilkie, D. Nelson, *IET. Optoelectron.* **2021**, *15*, 77.
[67] J. Zou, L. Li, Y. Zhuang, C. Wang, M. Zhang, Z. Le, X. Wang, G. Cai, S. Feng, J.-J. He, *J. Lightwave Technol.* **2023**, *41*, 226.
[68] J.-J. He, B. Liu, Y. Chu, X. Dong, *China 202210054701.8*, **2023**.
[69] W. Gao, X. Li, L. Lu, J. Chen, L. Zhou, presented at Optical Fiber Communication Conf. (OFC), California, USA, March **2022**.
[70] P. Pintus, M. Hofbauer, C. L. Manganelli, M. Fournier, S. Gundavarapu, O. Lemonnier, F. Gambini, L. Adelmini, C. Meinhart, C. Kopp, F. Testa, H. Zimmermann, C. J. Oton, *Laser Photonics Rev.* **2019**, *13*, 1800275.
[71] S. Tondini, C. Castellan, M. Mancinelli, C. Kopp, L. Pavesi, *J. Lightwave Technol.* **2017**, *35*, 5134.