PHOTONICS Research

Sophisticated deep learning with on-chip optical diffractive tensor processing

Yuyao Huang, Tingzhao Fu, Honghao Huang, Sigang Yang, and Hongwei Chen*

Beijing National Research Center for Information Science and Technology, Department of Electronic Engineering, Tsinghua University, Beijing 100084, China
**Corresponding author: chenhw@tsinghua.edu.cn*

Received 3 January 2023; revised 11 April 2023; accepted 17 April 2023; posted 17 April 2023 (Doc. ID 484662); published 1 June 2023

Ever-growing deep-learning technologies are making revolutionary changes for modern life. However, conventional computing architectures are designed to process sequential and digital programs but are burdened with performing massive parallel and adaptive deep-learning applications. Photonic integrated circuits provide an efficient approach to mitigate bandwidth limitations and the power-wall brought on by its electronic counterparts, showing great potential in ultrafast and energy-free high-performance computation. Here, we propose an optical computing architecture enabled by on-chip diffraction to implement convolutional acceleration, termed "optical convolution unit" (OCU). We demonstrate that any real-valued convolution kernels can be exploited by the OCU with a prominent computational throughput boosting via the concept of structral reparameterization. With the OCU as the fundamental unit, we build an optical convolutional neural network (oCNN) to implement two popular deep learning tasks: classification and regression. For classification, Fashion Modified National Institute of Standards and Technology (Fashion-MNIST) and Canadian Institute for Advanced Research (CIFAR-4) data sets are tested with accuracies of 91.63% and 86.25%, respectively. For regression, we build an optical denoising convolutional neural network to handle Gaussian noise in gray-scale images with noise level $\sigma = 10$, 15, and 20, resulting in clean images with an average peak signal-to-noise ratio (PSNR) of 31.70, 29.39, and 27.72 dB, respectively. The proposed OCU presents remarkable performance of low energy consumption and high information density due to its fully passive nature and compact footprint, providing a parallel while lightweight solution for future compute-in-memory architecture to handle high dimensional tensors in deep learning.

https://doi.org/10.1364/PRJ.484662

1. INTRODUCTION

Convolutional neural networks (CNNs) [1–3] power enormous applications in the artificial intelligence (AI) world, including computer vision [4–6], self-driving cars [7–9], natural language processing [10–12], medical science [13–15], etc. Inspired by biological behaviors of visual cortex systems, CNNs have brought remarkable breakthroughs in manipulating high-dimensional tensor such as images, videos, and speech, enabling efficient processing with more precise information extractions but much fewer network parameters, compared with the classical feed-forward one. However, advanced CNN algorithms have rigorous requirements on computing platforms, which are responsible for massive data throughputs and computations, which have triggered the flourishing development of high-performance computing hardware such as the central processing unit [16], graphics processing unit [17], tensor processing unit (TPU) [18], and field-programmable gate array [19]. Nonetheless, today's electronic computing architectures are facing physical bottlenecks in processing distribution and parallel tensor operations, e.g., bandwidth limitation, high-power consumption, and the fading of Moore's law, causing serious computation force mismatches between AI and the underlying hardware frameworks.

Important progress has been made to further improve the capabilities of future computing hardware. In recent years, the optical neural network (ONN) [20–27] has received growing attention with its extraordinary performance in facilitating complex neuromorphic computations. The intrinsic parallelism nature of optics enables more than 10 THz interconnection bandwidth [28], and the analog fashion of photonics system [29] decouplings demonstrates the urgent need for high-performance memory in conventional electronic architectures and therefore prevents energy wasting and time latency from continuous AD/DA conversion and arithmetic logic unit (ALU)-memory communication, thus boosting computational speed and reducing power consumption essentially.

To date, numerous ONNs have been proposed to apply various neuromorphic computations such as optical inference networks based on Mach–Zehnder interferometer (MZI) mesh [30–32], photonics spiking neural networks based on an

wavelength division multiplexing (WDM) protocol and ring modulator array [33–35], photonics tensor core based on phase change materials [36,37], and optical accelerator based on time-wavelength interleaving [38–40]. For higher computation capabilities of ONNs, diffractive optical neural networks [41–44] have been proposed to provide millions of trainable connections and neurons optically by means of light diffraction. To further improve network density, the integrated fashion of diffractive optical neural networks based on an optical discrete Fourier transform [45], multimode interference [46], and metasurface technologies [47–49] has been studied. However, these on-chip diffraction approaches are limited by power consumption and input dimensions, making them difficult to scale up for adapting massive parallel high-order tensor computations. Here, we take one step forward to address this issue by building an optical convolution unit (OCU) with on-chip optical diffraction and cascaded 1D metalines on a standard silicon on insulator (SOI) platform. We demonstrate that any real-valued convolution kernels can be exploited by an OCU with a prominent computation power. Furthermore, with the OCU as the basic building block, we build an optical convolutional neural network (oCNN) to perform classification and regression tasks. For classification tasks, Fashion-MNIST and CIFAR-4 data sets are tested with accuracies of 91.63% and 86.25%, respectively. For regression tasks, we build an optical denoising convolutional neural network (oDnCNN) to handle Gaussian noise in gray-scale images with noise level $\sigma = 10$, 15, and 20, resulting in clean images with average peak signal-to-noise ratio (PSNR) of 31.70, 29.39, and 27.72 dB. The proposed OCU and oCNN are fully passive in processing massive tensor data and compatible for ultrahigh bandwidth interfaces (for both electronic and optical), being capable of integrating with electronic processors to reaggregate computational resources and power penalties.

2. PRINCIPLE

Figure 1(a) presents the operation principle of 2D convolution. Here, a fixed kernel $\mathbf{K}$ with size of $H \times H$ slides over the image $\mathbf{I}$ with size of $N \times N$ by stride of S and does weighted addition with the image patches that are covered by the kernel, resulting an extracted feature map $\mathbf{O}$ with size of $G \times G$, where $G = \lfloor (N - H)/S + 1 \rfloor$ (in this case, we ignore the padding process of convolution). This process can be expressed in Eq. (1), where $\mathbf{O}[i, j]$ represents a pixel of the feature map, and m and n are related to the stride S. Based on this, one can simplify the operation as multiplications between an $H^2 \times 1$ vector $\hat{\mathbf{K}}$ reshaped by the kernel and a $G^2 \times H^2$ matrix $\hat{\mathbf{I}}$ composed by image patches, as shown in Eq. (2):

$$\mathbf{O}[i,j] = \sum_{m=-\infty}^{\infty} \sum_{n=-\infty}^{\infty} \mathbf{I}[i-m, j-n] \cdot \mathbf{K}[m,n], \tag{1}$$

$$\hat{\mathbf{O}} = \hat{\mathbf{K}} \cdot \hat{\mathbf{I}} = \begin{bmatrix} k_1 \\ k_2 \\ \vdots \\ k_{H^2} \end{bmatrix}^T \cdot \begin{bmatrix} i_{11} & i_{12} & \cdots & i_{1,G^2} \\ i_{21} & i_{22} & \cdots & i_{2,G^2} \\ \vdots & \vdots & & \vdots \\ i_{H^2,1} & i_{G^2,2} & \cdots & i_{H^2,G^2} \end{bmatrix}, \tag{2}$$

where $\hat{\mathbf{I}}_m = [i_{m,1}, i_{m,2}, \ldots, i_{m,H^2}]$, and $m = (1, 2, \ldots, G^2)$ is a corresponding image patch covered by a sliding kernel, and $\hat{\mathbf{O}}$ denotes the flattened feature vector. Consequently, the fundamental idea of optical 2D convolution is to manipulate multiple vector-vector multiplications optically and keep their

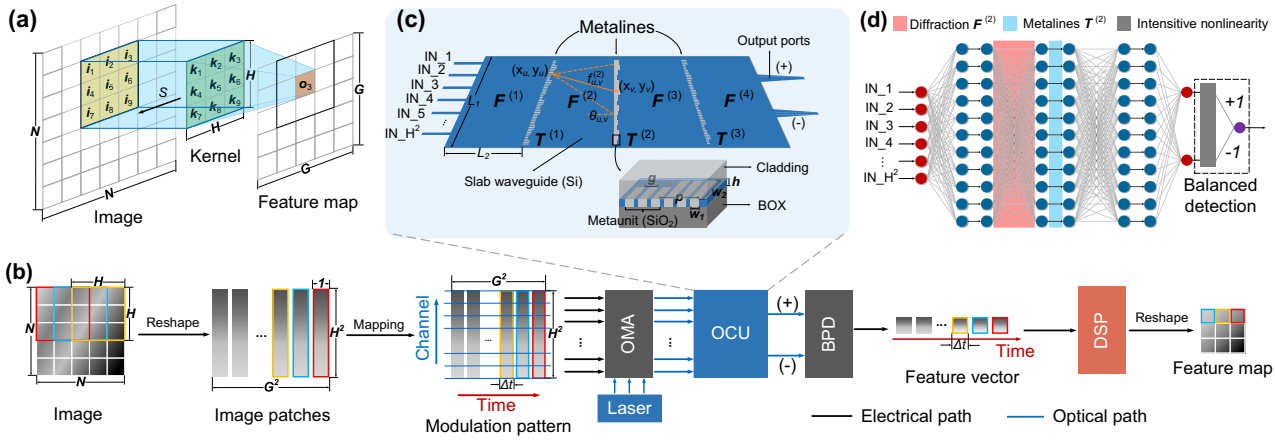

Fig. 1. Principle of optical image convolution based on OCU. (a) Operation principle of 2D convolution. A fixed kernel with size of $H \times H$ slides over the image with size of $N \times N$ by stride of S and does weighted addition with the image patches that are covered by the kernel, resulting in an extracted feature map with size of $G \times G$, where $G = \lfloor (N - H)/S + 1 \rfloor$. (b) Optical image convolution architecture with OCU. An image is first flattened into patches according to the kernel size and sliding stride and then mapped into a modulation pattern confined with time and channel number, which modulates a coherent laser via a modulation array. The modulated light is sent to OCU to perform optical convolution, whose positive and negative results are subtracted by a balanced photodetector and reshaped by a DSP to form a new feature map. OMA, optical modulator array; BPD, balanced photodetector; DSP, digital signal processor. (c) Details of OCU. H^2 waveguides are used to send a laser signal into a silicon slab waveguide with size of $L_1 \times L_2$, and layers of metaline are exploited successively with layer gap of L_2, which are composed by well-arranged metaunits. Three identical silica slots with sizes of $w_1 \times w_2 \times h$ are used to compose one metaunit with gap of g, and the period of metaunits is p. The phase modulation is implemented by varying w_2. The transfer function of the diffraction in slab waveguide and phase modulation of metalines are denoted as $\mathbf{F}$ and $\mathbf{T}$. (d) The feedforward neural network abstracted from the OCU model. Red and blue boxes denote diffractions and phase modulations of metalines; gray box represents intensitive nonlinear activation of complex-valued neural networks introduced by photodetection.

products in series for reshaping to a new map. Here, we use on-chip optical diffraction to implement this process, as described in Fig. 1(b). The input image with size of $N \times N$ is first reshaped into flattened patches according to the kernel size H and sliding stride S, which turns the image into a $G^2 \times H^2$ matrix $\hat{\mathbf{I}}$. Then, $\hat{\mathbf{I}}$ is mapped into a plane of space channels and time, in which each row of $\hat{\mathbf{I}}$ is varied temporarily with period of G^2 and each column of $\hat{\mathbf{I}}$ (namely, pixels of a flattened image patch) is distributed in H^2 channels. A coherent laser signal is split into H^2 paths and then modulated individually by the time-encoded and channel-distributed image patches, in either amplitude or phase. In this way, one time slot with duration of Δt contains one image patch with H^2 pixels in corresponding channels, and G^2 of these time slots can fully express image patch matrix $\hat{\mathbf{I}}$. Then, the coded light is sent to the proposed OCU to perform matrix multiplications as Eq. (2) shows, and the corresponding positive and negative results are detected by a balanced photodetector (BPD) to do subtractions between the two. The balanced detection scheme assures the OCU operates in a real-valued field. The detected information is varied temporarily with symbol duration of Δt and then reshaped into a new feature map by a digital signal processor (DSP). The principle is also applicable for images and kernels that have nonsquare shapes.

The details of OCU are given in Fig. 1(c). Here, H^2 silicon strip waveguides are exploited for receiving signals simultaneously from modulation channels, which diffract and interfere with each other in a silicon slab waveguide with size of $L_1 \times L_2$ before it encounters well-designed 1D metalines. The 1D metaline is a subwavelength grating consisting of silica slots with each slot having a size of $w_1 \times w_2 \times h$, which is illustrated in the inset of Fig. 1(c). Furthermore, we use three identical slots with slot gap of g to constitute a metaunit with period of p to ensure a constant effective refractive index of the 1D metaline when it meets light from different angles, as demonstrated in our previous work [48,49]. The incoming signal is phase-modulated from 0 to 2π by changing the length of each metaunit w_2 but with w_1 and h fixed. Accordingly, the corresponding length $w_{2,v}^{(l)}$ of the vth metaunit in the lth metaline can be ascertained from the introduced phase delay $\Delta\phi_v^{(l)}$ by Eq. (3), where n_1 and n_2 are the effective refractive index of the slab and slots, respectively. After layers of propagation, the interfered light is sent to two ports, which output positive and negative part of computing results

$$w_{2,v}^{(l)} = \frac{\lambda}{2\pi(n_1 - n_2)} \cdot \Delta\phi_v^{(l)}, \tag{3}$$

$$f_{u,v}^{(l)} = \frac{1}{j\lambda} \cdot \left(\frac{1 + \cos\theta_{u,v}}{2r_{u,v}}\right) \cdot \exp\left(j\frac{2\pi r_{u,v} n_1}{\lambda}\right) \cdot \eta \exp(j\Delta\psi). \tag{4}$$

For more precise analysis, the diffraction in the slab waveguide between two metalines with U and V metaunits, respectively, is characterized by a $U \times V$ matrix $\mathbf{F}^{(l)}$ based on the Huygens–Fresnel principle under restricted propagation conditions, whose element $f_{u,v}^{(l)}$, as shown in Eq. (4), is the diffractive connection between the uth metaunit located at (x_u, y_u) of the $(l-1)$th metaline, and the vth metaunit locates at (x_v, y_v) in the lth metaline, $\cos\theta_{u,v} = (x_u - x_v)/r_{u,v}$, $r_{u,v} = \sqrt{(x_u - x_v)^2 + (y_u - y_v)^2}$ denotes the distance between the two metaunits, λ is working wavelength, j is the imaginary unit, and η and $\Delta\psi$ are the amplitude and phase coefficients, respectively. As for each metaline, the introduced phase modulation is modeled by a $V \times V$ diagonal matrix $\mathbf{T}^{(l)}$, as expressed in Eq. (5):

$$\mathbf{T}^{(l)} = \begin{bmatrix} \exp(j\Delta\phi_1^{(l)}) & 0 & \cdots & 0 \\ 0 & \exp(j\Delta\phi_2^{(l)}) & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & \exp(j\Delta\phi_V^{(l)}) \end{bmatrix}. \tag{5}$$

To prove the accuracy of the proposed model in Eqs. (4) and (5), we evaluate the optical field of an OCU with finite-different time-domain (FDTD) method, as shown in Fig. 2(a). Three metalines with 10 metaunits for each are configured based on a standard SOI platform, the size of slab waveguide between the metalines is 40 μm × 15 μm, the width and gap of slots are set to be 200 and 500 nm, the period of metaunit is 1.5 μm, and a laser source is split to nine waveguides with working wavelength of 1550 nm. We monitor the amplitude and phase response of the diffracted optical field at Position A of Fig. 2(a), which agree well with the proposed analytical model in Eq. (4), as shown in Figs. 2(b) and 2(c). Phase modulation of the metaline is also validated by monitoring the optical phase response at Position B in Fig. 2(d), with the incident light

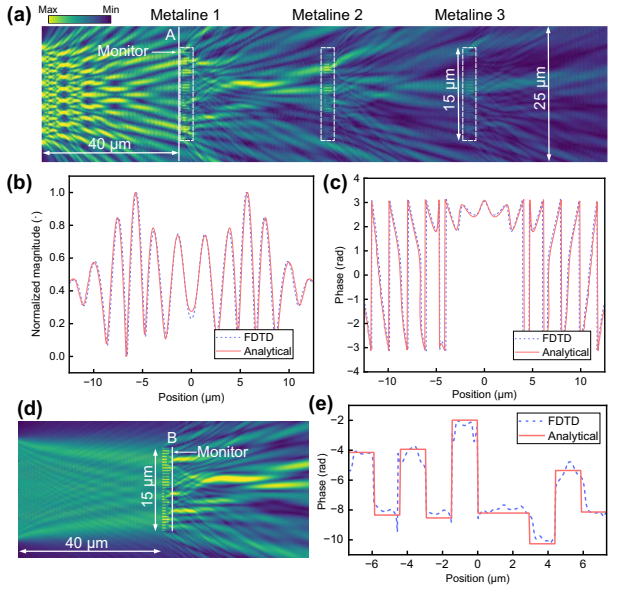

Fig. 2. (a) Optical field of the OCU evaluated by FDTD method. A monitor is set at Position A to receive the optical field of the incident light. (b) Magnitude and (c) phase response of the optical field at Position A (red solid curve) match well with the analytical model (purple dash curve) in Eq. (5). (d) Optical field of the metaline with incident light of a plane wave. A monitor is set behind the metaline at Position B to obtain its phase response. (e) The analytical model (purple dash curve) of Eq. (6) fits well with the FDTD calculation (red solid curve).

of a plane wave. Figure 2(e) shows an ideal match between the FDTD calculation and the analytical model in Eq. (5).

Consequently, we conclude the OCU model in Eq. (6), where M is the layer number of OCU and $\mathbf{R}_{\mathrm{OCU}}$ is the response of the OCU when the input is a reshaped image patch matrix $\hat{\boldsymbol{I}}$. Besides, the numbers of metaunits in the M metaline layers are all designed to be V, which leads to $\mathbf{F}^{(l+1)}$ and $\mathbf{T}^{(l)}$ $(l = 1, 2, \ldots, M\text{-}2)$ are matrices with size of $V \times V$. Specifically, $\mathbf{F}^{(1)}$ is a $V \times H^2$ matrix since H^2 waveguides are exploited, and $\mathbf{F}^{(M)}$ is a $2 \times V$ matrix since we only focus on the signals at two output ports:

$$\mathbf{R}_{\mathrm{OCU}} = \left\{\mathbf{F}^{(M)}\mathbf{T}^{(M-1)}\left[\prod_{l=1}^{M-2}(\mathbf{F}^{(l+1)}\mathbf{T}^{(l)})\right]\mathbf{F}^{(1)}\right\}\cdot\hat{\mathbf{I}}. \quad (6)$$

Therefore, $\mathbf{R}_{\mathrm{OCU}}$ is a $2 \times G^2$ matrix with column of $\mathbf{R}_1$ and $\mathbf{R}_2$, which are $1 \times G^2$ vectors, and the corresponding response of balanced detection is described in Eq. (7) accordingly, where $\odot$ denotes a Hadamard product, and κ is a amplitude coefficient introduced by the photodetection

$$\mathbf{R}_{\mathrm{BPD}} = \kappa\{\|\mathbf{R}_1 \odot \mathbf{R}_1^*\| - \|\mathbf{R}_2 \odot \mathbf{R}_2^*\|\}. \quad (7)$$

Furthermore, the OCU and balanced detection model in Eqs. (6) and (7) can be abstracted as a feedforward neural network, as illustrated in Fig. 1(d), where the dense connections denote diffractions, and the single connections are phase modulations introduced by metalines. The BPD's square-law detection performs as a nonlinear activation in the network since the phase-involved computing makes the network complex-valued [50–52].

Note that it is rarely possible to build a one-to-one mapping between the metaunit lengths and kernel value directly, because the phase modulation of metalines introduces complex-valued computations while the kernels are usually real-valued. However, the feedforward linear neural network nature of OCU model facilitates another approach to implement 2D convolution optically. Structural reparameterization [53–55] (SRP) is a networking algorithm in deep learning, in which the original network structure can be substituted equivalently with another one to obtain same outputs, as illustrated in Fig. 3. Here, we leverage this concept to create a regression between the diffractive feedforward neural network and 2D convolution. In other words, we train the network to learn how to perform 2D convolution instead of mapping the kernel value directly into metaunit lengths. More details are shown in the following sections.

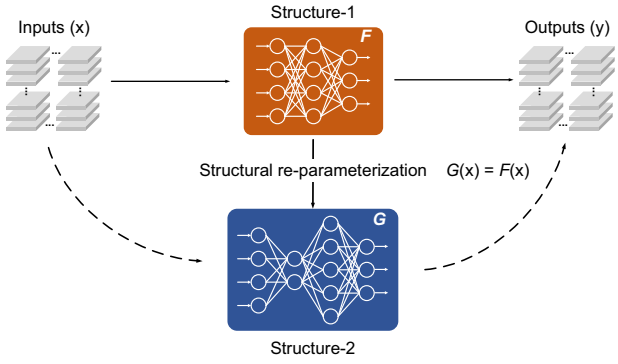

Fig. 3. Concept of structural reparameterization in deep learning. Network Structure 1 has a transfer function of F, which can be substituted equivalently by Network Structure 2, whose transfer function is G. Accordingly, both structures have the same outputs y under the same inputs of x.

3. RESULTS

In this section, we evaluate the performance of OCU in different aspects of deep learning. In Subsection 3.A, we present the basic idea of 2D optical convolution with the concept of SRP; we also demonstrate that the proposed OCU is capable of representing arbitrary real-valued $H \times H$ convolution kernel (in our following demos, we take $H = 3$) and therefore implementing a basic image convolution optically. In Subsection 3.B, we use the OCU as a fundamental unit to build an oCNN, with which classification and regression applications of deep learning are carried out with remarkable performance.

A. Optical Convolution Functioning

As aforementioned, an OCU cannot be mapped from a real-valued kernel directly since the phase modulation of metalines makes the OCU model a complex-valued feedforward neural network. Therefore, we need to train the OCU to "behave" as a real-valued convolution model with the SRP method, which is referred to as the training phase of OCU, as illustrated in Fig. 4(a). We utilize a random pattern as the training set to make a convolution with a real-valued kernel, and the corresponding result is reshaped as a training label $\hat{\mathbf{R}}$. Then, we apply the training set on the OCU model to obtain a feature vector $\mathbf{R}_{\mathrm{BPD}}$ and calculate a mean square error loss $\mathbb{J}$ with the collected label. Through the iteration of a backward propagation algorithm in our model, all the trainable parameters are updated to minimize loss, and the OCU is evolved to the targeting real-valued kernel, as shown in Eqs. (8) and (9), where $\Delta\Phi$ is metaline-introduced phase; it is also the trainable parameter of the OCU. Accordingly, images can be convolved with the well-trained OCU; we term this process as an "inference phase," as presented in Fig. 4(b):

$$\mathbb{J} = \frac{1}{2}\cdot\sum_{i=1}^{G^2}\|\mathbf{R}_{\mathrm{BPD}}(\Delta\Phi)[i] - \hat{\mathbf{R}}[i]\|^2, \quad (8)$$

$$\Delta\Phi^* = \arg\min_{\Delta\Phi}\ \mathbb{J}(\Delta\Phi). \quad (9)$$

For proof-of-concept, a 128×128 random pattern (the OCU's performance receives almost no improvement with a random pattern that is larger than 128×128) and eight unique real-valued 3×3 convolution kernels are exploited to generate training labels, and a 256×256 gray-scale image is utilized to test the OCUs' performance, as shown in Fig. 5. In this case, we use three layers of metalines in OCU with $L_1 = 75$ μm and $L_2 = 300$ μm; each metaline consists of 50 metaunits with $w_1 = 200$ nm, $g = 500$ nm, and $p = 1.5$ μm; further, the number of input waveguides is set to be nine according to the size of the utilized real-valued kernel. The training and inference process of OCU are conducted with TensorFlow2.4.1 framework. From Fig. 5, we can see good matches between the

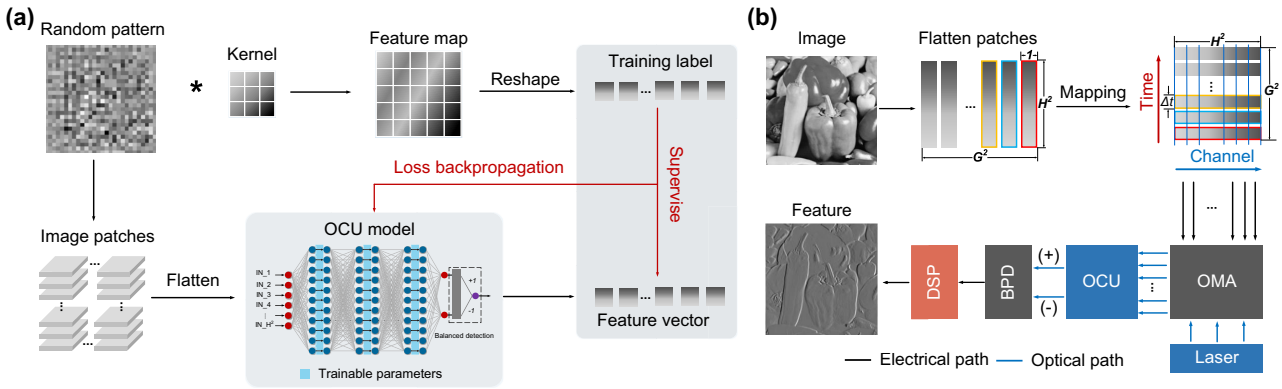

Fig. 4. Training and the inference phase of an OCU to perform real-valued optical convolution with the idea of SRP in deep learning. (a) A 128 × 128 random pattern is utilized to generate a training pair for OCU model. The output feature vector of OCU model is supervised by the training label with the input of flattened image patches decomposed from the random pattern. (b) A 256 × 256 gray-scale image is reshaped to flattened patches and sent to the well-trained OCU to perform a real-valued convolution. OMA, modulator array; BPD, balanced photodetector; DSP, digital signal processor.

ground truths generated by real-valued kernels and the outputs generated by OCUs with high peak signal-to-noise ratios (PSNRs); moreover, the average PSNR between the two can be calculated as 36.58 dB, indicating that the OCU can respond as a real-valued convolution kernel with remarkable performance.

B. Optical Convolutional Neural Network

With OCU as the basic unit for feature extraction, more sophisticated architectures can be carried out efficiently to interpret the hidden mysteries in higher dimensional tensors. In this section, we build an optical convolutional neural network (oCNN) to implement tasks in two important topics of deep learning: classification and regression.

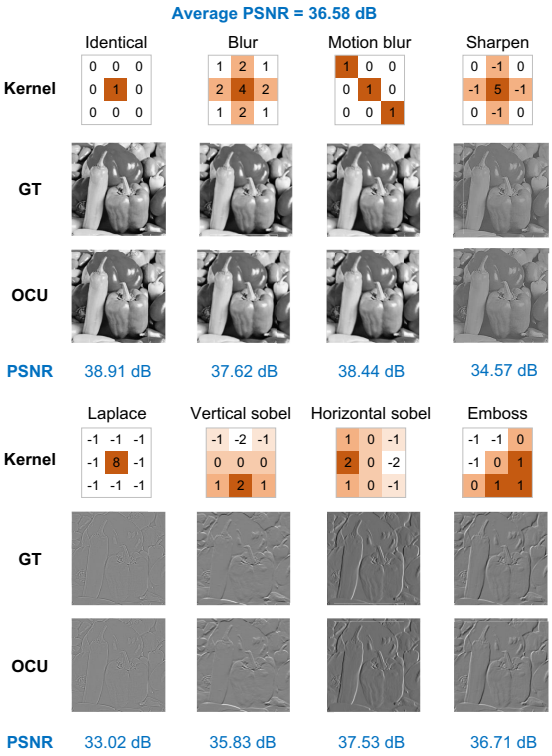

Fig. 5. Well-agreed convolution results between the ground truths and the outputs of OCUs with eight unique real-valued convolution kernels, with average PSNR of 36.58 dB. GT, ground truth.

1. Image Classification

Figure 6 shows the basic architecture of oCNN for image classifications. Images with size of $N \times N \times C$ are first flattened into C groups of patches and concatenated as a data batch with size of $G^2 \times C \cdot H^2$ according to the kernel size H; then, they loaded to a modulator array with total $C \cdot H^2$ modulators in parallel. Here, C denotes the image channel number, and N, G, and H are already defined in the principle section. The modulated data batch is copied q times and split to q optical convolution kernels (OCKs) by means of optical routing. Each OCK consists of C OCUs corresponding to C data batch channels, and the nth channel of the data batch is convolved by the nth OCU in each OCK, where $n = 1, 2, \ldots, C$. Balanced photodetection is utilized after each OCU to give a subfeature map FM_{mn} with size of $G \times G$, where $m = 1, 2, \ldots, q$, and all C subfeature maps in a OCK are summed up to generate a final feature map FM_m. For convenience, we term this process as an "optical convolution layer" (OCL), as denoted inside the blue dashed box of Fig. 5. After OCL, the feature maps are further downsampled by the pooling layer to form more abstracted information. Multiple OCLs and pooling layers can be exploited to establish deeper networks when the distribution of tensors (herein this case, images) is more complicated. At last, the extracted output tensors are flattened and sent to a small but fully connected (FC) neural network to play the final classifications.

We demonstrate the oCNN classification architecture on gray-scale image data set Fashion-MNIST and colored image data set CIFAR-4, which are selected from the widely used CIFAR-10 with much more complex data distribution. We visualize the two data sets with the t-distributed stochastic neighbor embedding method in a 2D plane, as shown in Fig. 7(d). For Fashion-MNIST, we use four OCKs to compose an OCL for feature extraction and three cascaded FC layers to give the final classification, assisted with the loss of cross

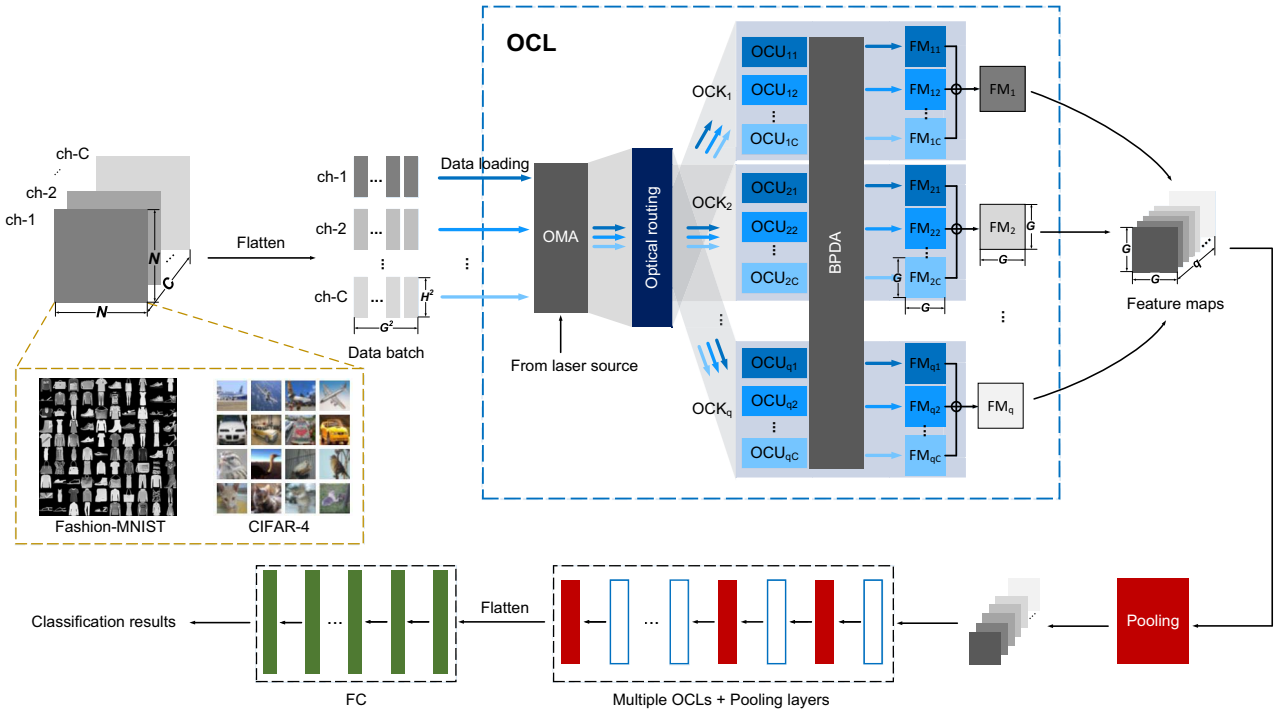

Fig. 6. Architecture of oCNN for image classification. Images with size of $N \times N \times C$ are first flattened into C groups of patches and concatenated as a data batch with size of $G^2 \times C \cdot H^2$, according to the kernel size H, and then loaded to a modulator array with total $C \cdot H^2$ modulators in parallel. The modulated signal is split to q OCKs by an optical router, each of which contains C OCUs to generate C subfeature maps; then, all the subfeature maps of each OCK are summed up to form a final feature map. We refer to this process as an OCL, denoted by the blue dashed box. After OCL, the feature maps are further downsampled by a pooling layer, and multiple OCLs and pooling layers can be utilized to build deeper networks to manipulate more complicated tasks. A small-scale fully connected layer is used to give the final classification results. OMA, optical modulator array; OCK, optical convolution kernel; BPDA, balanced photodetector array; FM, feature map; FC, fully connected layer.

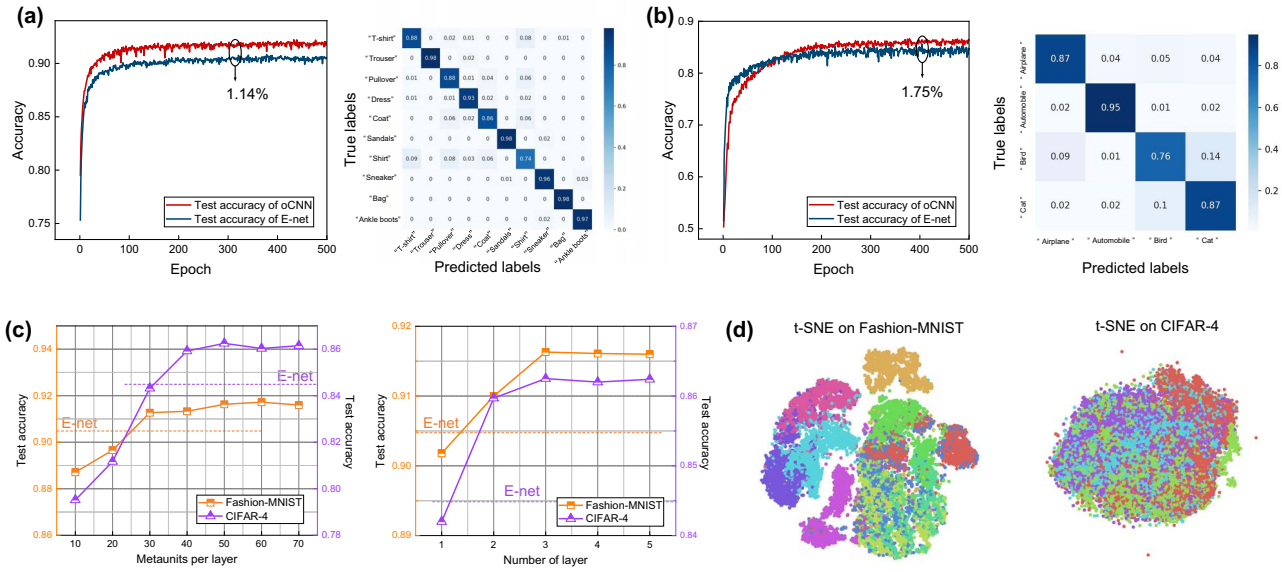

Fig. 7. Classification results of oCNNs for (a) fashion-MNIST and (b) CIFAR-4 data sets. Accuracies of 91.63% and 86.25% are obtained with oCNNs for the corresponding two data sets, which outperform their electrical counterparts with 1.14% and 1.75% respectively. (c) Classification performance evaluations on both data sets with respect to two main physical parameters of OCU: the number of metaunit per layer and the number of the exploited metaline layer. (d) 2D visualizations of the two applied data sets with t-distributed stochastic neighbor embedding (t-SNE) method.

entropy. Here, in this case, each OCK only has one OCU since gray-scale images have only one channel, and each OCU performs as a 3×3 convolution. We use 60,000 samples as the training set and 10,000 samples as the test set; after 500 epochs of iterations, the loss of the training and test sets is converged, and the accuracy of test set is stable at 91.63%, as given in Fig. 7(a) attached with a confusion matrix. For CIFAR-4 data set, a similar method is leveraged: an OCL with 16 OCKs is carried out with each OCK consisting of three OCUs, according to RGB channels of the image, and then three FC layers are applied after the OCL. Further, each OCU also performs as a 3×3 convolution. Here, 20,000 samples are used as the training set and another 4000 samples as the test set; the iteration epoch is set as 500. We also use cross entropy as the loss function. After iterations of training, the classification accuracy is stable at 86.25%, as shown in Fig. 7(b), and the corresponding confusion matrix is also presented. The OCU's parameter we use here is the same as the settings in Subsection 3.A. Furthermore, we also evaluate the performances of electrical neural networks (denoted as E-net) with the same architecture as optical ones in both two data sets; the results show that the proposed oCNN outperforms E-net with accuracy boosts of 1.14% for Fashion-MNIST and 1.75% for CIFAR-4.

We also evaluate the classification performance of the oCNN, with respect to two main physical parameters of the OCU: the number of metaunit per layer and the number of the exploited metaline layer, as shown in Fig. 7(c). In the left of Fig. 7(c), three metaline layers are used with the number of metaunit per layer varied from 10 to 70; the result shows that increasing the metaunit numbers gives accuracy improvements for both data sets; however, the task for CIFAR-4 has a more significant boost of 6.73% than the Fashion-MNIST of 2.92% since the former has a more complex data structure than the latter; therefore, it is more sensitive to model complexity. In the right of Fig. 7(c), 50 metaunits are used for each metaline layer, and the result indicates that increasing the layer amount of the metaline also gives a positive response on test accuracy for both data sets, with accuracy improvements of 1.45% and 1.05%, respectively. To conclude, the oCNN can further improve its performance by increasing the metaunit density of the OCU, and adding more metaunits per layer is a more efficient way than adding more layers of the metaline to achieve this goal.

2. Image Denoising

Image denoising is a classical and crucial technology that has been widely applied for high-performance machine vision [56,57]. The goal is to recover a clean image $\mathbf{X}$ from a noisy one $\mathbf{Y}$, and the model can be written as $\mathbf{Y} = \mathbf{X} + \mathbf{N}$, where in general $\mathbf{N}$ is assumed to be an additive Gaussian noise. Here, we refer to the famous feed-forward denoising convolutional neural network (DnCNN) [58] to build its optical fashion, termed as "optical denoising convolutional neural network" (oDnCNN), to demonstrate the feasibility of the proposed OCU in deep-learning regression.

Figure 8(a) shows the basic architecture of oDnCNN, which includes three different parts (as follows). (i) Input layer: OCL with q_1 OCKs is utilized (details are presented in Fig. 6). Each OCK consists of C_{in} OCUs, which perform $3 \times 3 \times C_{\text{in}}$ 2D convolutions, where $C_{\text{in}} = 1$ for gray-scale images and $C_{\text{in}} = 3$ for colored images. Then, ReLUs are utilized for nonlinear activation. (ii) Middle layer: OCL with q_2 OCKs is exploited; for the first middle layer q_1, OCUs are used in each OCK; for the rest of the middle layers, the number is q_2. ReLUs are also used as nonlinearity, and batch normalization is added between OCL and ReLU. (iii) Output layer: only one OCL with one OCK is leveraged, which has q_2 OCUs.

With this architecture, basic Gaussian denoising with known noise level σ is performed. We follow Ref. [59] to

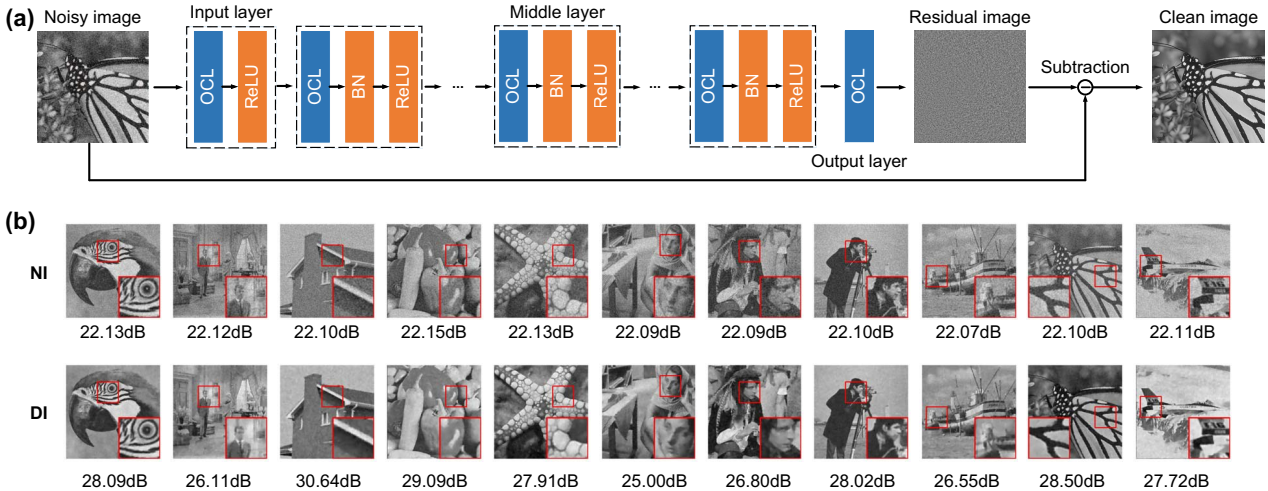

Fig. 8. (a) Architecture of the proposed oDnCNN. A Gaussian noisy image with a known noise level is first flattened and modulated into a lightwave and then sent to oDnCNN, which is composed of three parts: input layer with OCL and ReLU; middle layer with an extra batch normalization between the two; and output layer with only an OCL. After the oDnCNN, a residual image is obtained, which is the extracted noise. By subtracting the noisy image with the extracted residual one, the clean image can be aquired. OCL, optical convolution layer; ReLU, rectified linear unit; BN, batch normalization. (b) The denoised result of Set12 data set leveraged by the proposed oDnCNN with noise level $\sigma = 20$, giving much clearer textures and edges as the details show in red boxes. In this case, the average PSNR of the denoised images is 27.02 dB, compared with 22.10 dB of the noisy ones. NI, noisy images; DI, denoised images.

Table 1. Performance Comparisons of the Proposed oDnCNN and E-net in Average PSNR, with Noise Level $\sigma = 10$, 15, and 20

Noise Level	Noisy (dB)	oDnCNN (dB)	E-net (dB)
$\sigma = 10$	28.13	31.70	30.90
$\sigma = 15$	24.61	29.39	29.53
$\sigma = 20$	22.10	27.72	27.74

use 400 gray-scale images with size of 180 × 180 to train the oDnCNN and crop them into 128 × 1600 patches with patch size of 40 × 40. For test images, we use the classic Set12 data set, which contains 12 gray images with size of 256 × 256. Three noise levels, i.e., $\sigma = 10$, 15, and 20, are considerd to train the oDnCNN and are also applied to the test images. The oDnCNN we apply for this demonstration includes one input layer, one middle layer, and one output layer, among which eight OCKs are exploited for the input and middle layer, respectively, and one OCK for the output layer. Similar physical parameters of OCUs in the OCKs are set as the ones in Subsection 3.A; the only difference is that we use only two layers of metaline in this denoising demo. Figure 8(b) shows the denoising results under test images with noise level $\sigma = 20$. We evaluate the average PSNR for each image before and after the oDnCNN's denoising as 22.10 and 27.72 dB, posing a 5.62 dB improvement of image quality; further, details in red boxes show that clearer textures and edges are obtained at the oDnCNN's output. More demonstrations are carried out for noise level $\sigma = 10$ and 15, and the performances of E-net are also evaluated, as presented in Table 1. The results reveal that the oDnCNN provides 3.57 and 4.78 dB improvements of image quality for $\sigma = 10$ and $\sigma = 15$, which is comparable with the E-net's performance. Our demonstrations are limited by the computation power of the utilized server, and the overall performance can be further improved by increasing the metaunit density of the OCUs.

4. DISCUSSION

A. Computation Throughput and Power Consumption

The operation number of a 2D convolution composes the production part and accumulation part, which can be addressed by the kernel size H as shown in the first equation in Eq. (10). Consequently, for a convolution kernel in CNN, the operation number (OPs) can be further scaled by input channel C, as shown in the second equation in Eq. (10). Here, O_{conv} and O_{kernel} denote operation numbers of a 2D convolution and a convolution kernel in a CNN:

$$\begin{aligned} O_{\text{conv}} &= 2 \cdot H^2 - 1 \quad \text{OPs}, \\ O_{\text{kernel}} &= C \cdot O_{\text{conv}} \quad \text{OPs}. \end{aligned} \tag{10}$$

Consequently, the computation speed of an OCU can be calculated by the operation number O_{conv} and modulation speed r of OMA; the speed of an OCK with C input channels can be also acquired, by evaluating the number of operations per second (OPS), referred to as computation throughput. The calculations are presented in Eq. (11), where S_{ocu} and S_{ock} represent the computation throughput of OCU and OCK, respectively:

$$\begin{aligned} S_{\text{ocu}} &= O_{\text{conv}} \cdot r \quad \text{OPS}, \\ S_{\text{ock}} &= C \cdot S_{\text{ocu}} \quad \text{OPS}. \end{aligned} \tag{11}$$

From Eq. (11), we can see that the computation throughput of the OCU or OCK is largely dependent on modulation speed of OMA. Meanwhile, a high-speed integrated modulator has received considerable interest, in terms of new device structure or new materials, and the relative industries are also going to be mature [60]. Assuming that the modulation speed is 100 GBaud per modulator, for an OCU performing 3 × 3 optical convolutions, the computation throughput can be calculated as $(2 \times 3 \times 3 - 1) \times 100 = 1.7$ TOPS. For instance, in the demonstration in the last section, 16 OCKs are utilized to classify the CIFAR-4 data set, which contains three channels for each image; therefore, the total computation throughput of the OCL can be addressed as $3 \times 1.7 \times 16 = 81.6$ TOPS.

Because the calculations of OCU are all passive, its power consumption mainly comes from the data loading and photodetection process. Schemes of a photonics modulator with small driving voltage [61–63] have been proposed recently to provide low power consumption; further, integrated photodetectors [64,65] are also investigated with negligible energy consumed. Therefore, the total power of an OCU with equivalent kernel size of H can be calculated as Eq. (12), where E_{ocu}, E_{dri}, and E_{dec} are the energy consumptions of OCU, data driving, and detection, respectively; E_b is the utilized modulator's energy consumption; P_d is the power of the photodetector; B denotes the symbol or pixel number; and D is the symbol precision. Assuming that a 100 GBaud modulator and a balanced photodetecor with energy and power consumption of 100 fJ/bit and 100 mW are used for a 4K image with more than 8 million pixels and 8-bit depth for each, the total energy consumed by a 3 × 3 optical convolution can be calculated as $(3 \times 3) \times (8 \times 10^6 \times 8 \times 3 \times 100 \times 10^{-15}) + 0.1 \times [8 \times 10^{-6}/(100 \times 10^9)] = 1.808 \times 10^{-4}$ J.

$$\begin{aligned} E_{\text{mod}} &= H^2 \cdot (B \cdot D \cdot C \cdot E_b), \\ E_{\text{det}} &= P_d \cdot (B/r), \\ E_{\text{ocu}} &= E_{\text{mod}} + E_{\text{det}}. \end{aligned} \tag{12}$$

B. Data Preprocessing

As with most electronic integrated circuits, the photonic integrated circuits we rely on in our manuscript are basically 2D circuit planes; therefore, as with its electronic counterparts, our OCU can only read 2D information by preprocessing it to 1D-shape data. The preprocessing method we use in our OCU is generalized matrix multiplication (GeMM) [66], which is widely adopted by the state-of-the-art electronic integrated computing architectures such as Tensor Core of NVIDIA Ampere [67], Huawei Davinci [68], Google TPU [69], and the cross-bar array of a memristor [70]. The idea of GeMM is to transform high-order tensor convolutions into 2D matrix multiplications (referred to as "im2col operation") so that computation can be performed in parallel. The benefit from using GeMM is that it makes the storage of data in memory more regular and closer to computational cores to further shorten

the computation latency and reduce the impact of a bandwidth wall. This benefit also facilitates optical tensor computing architecture for loading data from memory with higher efficiency.

However, GeMM reshapes and duplicates tensor data continually; it also increases memory usage and additional access drastically. The consequence of this drawback is that more memory spaces are required to store high-order tensors and accordingly increase electronic system complexity. Even so, the relative computation performance of OCU is still unaffected compared with electronic ones since GeMM is applied on both architectures. In the long run, however, optimization for GeMM is crucial. Considerable studies have been done in GeMM optimizations [71–73]; some photonic solutions [36,74,75] are also carried out by transferring data manipulation from memory to photonic devices.

C. Scalability of OCU

In this work, the proposed OCU is a prototype with the simplest form, i.e., one OCU represents one kernel. Based on this, performing N optical convolutions in parallel requires N prototypes, as shown in Fig. 9(a), where each kernel size is assumed as $H \times H$. This prototype OCU can be scaled up intrinsically by simply diffracting light from the final metalines layer to the OCU's output facet, since the optical diffraction broadcasts information densely and produces a spatially-varied electrical field at the OCU's output facet; in addition, this field can be further multiplexed to train and perform multiple convolutions in parallel at different spatial position of the output facet, as shown in Figs. 9(b) and 9(c). In this case, we refer to the OCU as spatial-multiplexed. Assume N convolutions are expected to perform in one space-multiplexed OCU, with each kernel having size of $H \times H$; then, the fitting target of this OCU is a matrix $\mathbf{K}$ with size of $H^2 \times N$, whose column $\mathbf{K}_i$ is a $1 \times H^2$ kernel vector, where $i = 1, 2, \ldots, N$. Note that, for space-multiplexed OCU, the number of metaunits and metalines layers may increase

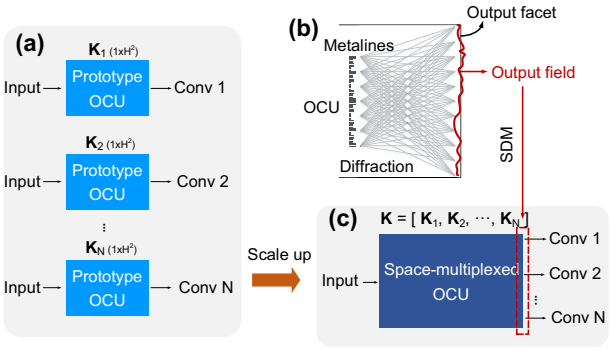

Fig. 9. Method of OCU's scaling for computing multiple convolutions. (a) Computing with prototype OCU. N convolutions require N prototype OCUs, each of which represents one kernel. (b) Principle of scaling up the prototype. Optical diffraction in silicon slab waveguide provides spatially-varied field at the output facet of OCU, which enables the possibility for multiplexing multiple convolution operations spatially. (c) Computing with space-multiplexed OCU. N convolutions can be calculated by just one space-multiplexed OCU with its fitting target of a kernel matrix $\mathbf{K} = [\mathbf{K}_1, \mathbf{K}_2, \ldots, \mathbf{K}_N]$, where $\mathbf{K}_i$ is a $1 \times H^2$ vector with $i = 1, 2, \ldots, N$. SDM, space division multiplexing.

with the number of the spatially performed convolutions, because OCU's representative space evolves from a single kernel (vector) to multiple kernels (matrix), which makes the training of space-multiplexed OCU difficult. Besides, physical limitations such as insertion loss, crosstalk, and signal-to-noise ratio must also be considered carefully with the scaling. Therefore, the extent of OCU's spatial scaling requires a thorough evaluation to find a trade-off between the benefits and losses.

D. Superiority of OCU

In recent years, significant efforts have been made in exploring high-performance computing with integrated photonics; we also find, however, that bottlenecks impede the practical applications of these optical computing schemes, the leading obstacle of which is the issue of scalability brought by the utilized photonic devices and techniques, which were supposed to power the positive development of optical computing. Coherent interference and WDM technique are the two most used approaches in the optical computing world. Even though both techniques are highly reconfigurable and programmable, their scalabilities remain limited. The Mach–Zehnder interferometer (MZI) enabled coherent interference scheme [13,32,76] is based on singular value decomposition, which facilitates the implementation of the matrix operation naturally and has a bulky size compared with other silicon photonic processors. In Ref. [13], 56 thermo-optical MZIs are used to implement a 4×4 matrix multiplication with the areas of each MZI of around 174 μm × 66 μm; in Ref. [32], a 6×6 complex-valued matrix multiplication is realized with a chip size of 0.53 mm × 1.46 mm. In addition, these sizes would be larger if high-speed modulation is further applied. As for the WDM-based optical computing scheme, microring resonators (MRRs) are often used for wavelength manipulation [35,36,74,77], which have a much greater compact footprint and lower power consumption than MZIs. Nonetheless, MRRs are sensitive to environment variations such as temperature and humidity fluctuation, which shift the MRRs' resonance wavelength drastically. Therefore, feedback control circuits and algorithms are intensively applied to stabilize the MRRs' resonance wavelength, especially in the case of high-speed modulation, causing significant downsides of system complexity and power consumption. For other WDM schemes such as time-wavelength interleaving [40], a multiwavelength source brings considerable energy consumption, a dispersive medium required platform from other materials such as silicon nitride, and on-chip dense WDM MUXers and DeMUXers, which face challenges from crosstalk and bandwidth steering. Further, synchronization of multiple-weight bank modulators remains tricky to address.

In contrast, diffraction-based OCUs have two benefits in accelerating computations optically. (1) Natural parallel computing architecture. Optical diffraction enables dense connections between two adjacent layers and produces fruitful information at the output facet by broadcasting inputs spatially, laying a solid foundation for large-scale parallel computations. Most importantly, these connections are built up simultaneously and passively, with simple physical implementations and speed of lightwave. (2) More powerful representation ability. This benefit comes from the compact footprint of a

Table 2. Comparison of State-of-the-Art Integrated Photonic Computing Hardware[a]

Works	Footprint (mm^2)[b]	Matrix Dimension	Operation Density (OPs/mm^2)	Power Efficiency (TOPS/W)
MZI mesh [13]	0.68	4 × 4	28/0.68 = 41.17	—
MZI mesh [32]	0.77	6 × 6	66/0.77 = 85.71	—
Cascaded MZI [76]	9.33	5 × 5 (convolution)	49/9.33 = 5.25	—
MRRs [35]	0.38	4 × 2	12/0.38 = 31.58	—
WDM + PCM [36]	6.07	9 × 4	63/6.07 = 10.37	0.4
MRRs + delay lines [74]	0.81	3 × 3 (convolution)	17/0.81 = 20.98	—
MRRs + TWI [77]	1.31	2 × 2 (convolution)	6/1.31 = 4.58	1.52×10^{-3}
Diffractive cell [45]	2.36	10 × 10	190/2.36 = 80.51	0.11
This work	0.088	3 × 3 (convolution)	17/0.088 = 193.18	0.37[c]

[a]PCM, phase change material; TWI, time-wavelength interleaving.
[b]Total area of photonic chip is considered.
[c]10 Gbit/s modulators with power of 51 mW for each [78] and receivers with power of 2.97 mW [79] are used for the estimation.

photonic neuron facilitated by 1D metalines, which are subwavelength gratings with each unit of hundreds of nanometers in width and up to 2 to 3 μm in length. The compact size of a photonic neuron creates a larger parameter space than other integrated approaches since the number of trainable photonic neurons is greater in the same area, making the mathematical model of an OCU more powerful to represent diverse and multiple linear transformations with the support of structural reparameterization. Based on the above two benefits, we believe the OCU has greater performance in scaling up computations, and we list more detailed comparisons with representative works in terms of footprint, computation density, and power consumption in Table 2. Notably, since different architectures have distinct modulation schemes for input and weight loading, we only evaluate computation density that is normalized by modulation speed, termed "operation density." The definition of operation density and power efficiency is given by Eqs. (13) and (14).

$$\text{Operation density} = \frac{\text{Operation number (OPs)}}{\text{Total area of photonic chip (mm}^2\text{)}}, \tag{13}$$

$$\text{Power efficiency} = \frac{\text{Computation throughput (TOPS)}}{\text{Power consumption (W)}}. \tag{14}$$

E. Networking with Optical Tensor Core

Today's cutting-edge AI systems are facing a double test of computation forces and energy cost in performing data-intensive applications [80]; models like the ResNet50 [81] and VGG16 [82] are power-hungry in processing high-dimensional tensors such as images and videos, molecular structures, time-serial signals, and languages, especially when the semiconductor fabrication process approaches its limitations [83]. Edge computing [84–86], a distributive computing paradigm, is proposed to process data near its source to mitigate the bandwidth wall and further improve computation efficiency, which requires computing hardware and has low run-time energy consumption and short computing latency. Compute-in-memory (CIM) [87–89] has received considerable attention in recent years since it avoids long time latency in data movement and reduces intermediate computations, thus showing potential as an AI edge processor. However, reloading large-scale weights repeatedly from DRAM to local memory also weakens energy efficiency significantly. Notably, the proposed OCU can be regarded as a natural CIM architecture because the computations are performed with the optical flow connecting the inputs and outputs with the speed of light; more importantly, its weights are fixed at the metaunits; therefore, the data loading process is eliminated.

Consequently, from a higher perspective, we consider a general networking method with multiple OCUs and optoelectrical interfaces, by leveraging the idea of network rebranching [90], to build an optoelectrical CIM architecture, as shown in Fig. 10. The idea of rebranching is to decompose the model mathematically into two parts (i.e., trunk and branch); by fixing the major parameters in the trunk and altering the minor ones in the branch, the network can be programmed with low energy consumption. The trunk part, which is responsible for major computations of the model, has fixed weights provided optically, referred to as the optical tensor core (OTC). The laser bank is exploited as the information carrier and routed by optical I/O to multiple optical tensor units (OTUs), and tensor data in the DRAM are loaded into OTUs by high-speed drivers. The OTUs contain a modulator array, OCUs, and a balanced PD array, and manipulate tensor convolutions passively; the calculation results are read out by DSPs. The branch part, which is a programmable lightweight electrical network, is responsible for reconfiguring the whole model with negligible computations. With this structure, big models can be performed with the speed of the TOPS level, but almost no power is consumed, and time latency is also shortened since fewer weights are reloaded from the DRAM. This scheme is promising for future photonics AI edge computing.

Technologies for the implementation of OTC are quite mature these days. The DFB laser array [91,92] can be applied as the laser bank, which has been widely used in commercial optical communication systems; further, an on-chip optical frequency comb [93–95] can provide an even more compact and efficient source supply with the Kerr effect in a silicon-nitride waveguide. Integrated optical routing schemes have been proposed recently based on the MZI network [96–98], ring modulators [99–102], and MEMS [103–105], with low insertion loss and flexible topology. Integrated modulators with ultrahigh bandwidth and low power consumption are also investigated intensively based on MZ [106] and ring [63] structures, with

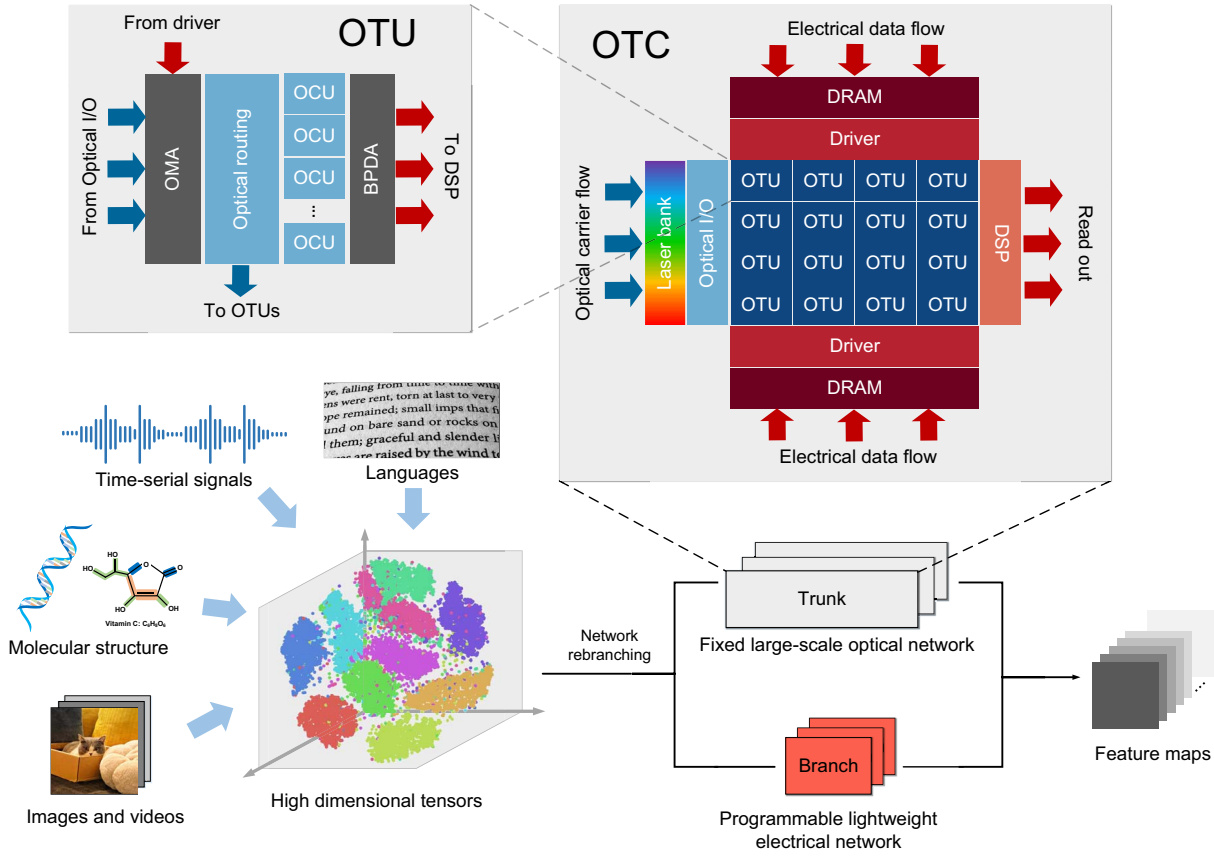

Fig. 10. Highly efficient optical deep learning framework with network rebranching and optical tensor core. Deep learning models are decomposed mathematically into two parts: trunk and branch, which carry the major and minor calculations of the model, respectively. The trunk part is computed by an optical tensor core with fixed weights, and the branch part is performed by a lightweight electrical network to reconfigure the model. OTC, optical tensor core; OTU, optical tensor unit.

diverse material platforms, including SOI [60], lithium niobate [107], and indium phosphide [108]. High-speed photodetectors with high sensitivity, low noise, and low dark current based on either silicon or III-V materials have also been studied and massively produced for optical communications [109] and microwave photonics [110] industries. The existing commercial silicon photonics foundries [111,112] are capable of fabricating metasurfaces with a minimum linewidth smaller than 180 nm via universal semiconductor techniques, showing potential for future pipeline-based production of the proposed OCU.

5. CONCLUSION

In this work, we propose an optical convolution architecture, OCU, with light diffraction on a 1D metasurface to process large-scale tensor information. We demonstrate that our scheme is capable of performing any real-valued 2D convolution by using the concept of structural reparameterization. We then apply the OCU as a computation unit to build a convolutional neural network optically, implementing classification and regression tasks with extraordinary performances. The proposed scheme shows advantages in either computation speed or power consumption, posing a novel networking methodology of large-scale but lightweight deep-learning hardware frameworks.

Funding. National Natural Science Foundation of China (62135009); Beijing Municipal Science and Technology Commission (Z221100005322010).

Disclosures. The authors declare no conflicts of interest.

Data Availability. The data that support the plots within this paper and other findings of this study are available from the corresponding author 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. Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," Nature **521**, 436–444 (2015).
3. J. Gu, Z. Wang, J. Kuen, L. Ma, A. Shahroudy, B. Shuai, T. Liu, X. Wang, G. Wang, J. Cai, and T. Chen, "Recent advances in convolutional neural networks," Pattern Recogn. **77**, 354–377 (2018).
4. A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," Commun. ACM **60**, 84–90 (2017).
5. A. Voulodimos, N. Doulamis, A. Doulamis, and E. Protopapadakis, "Deep learning for computer vision: a brief review," Comput. Intell. Neurosci. **2018**, 7068349 (2018).

6. D. A. Forsyth and J. Ponce, *Computer Vision: A Modern Approach* (Prentice Hall, 2002).
7. M. Bojarski, D. Del Testa, D. Dworakowski, B. Firner, B. Flepp, P. Goyal, L. D. Jackel, M. Monfort, U. Muller, J. Zhang, X. Zhang, J. Zhao, and K. Zieba, "End to end learning for self-driving cars," arXiv, arXiv:1604.07316 (2016).
8. J. Levinson, J. Askeland, J. Becker, J. Dolson, D. Held, S. Kammel, J. Z. Kolter, D. Langer, O. Pink, V. Pratt, M. Sokolsky, G. Stanek, D. Stavens, A. Teichman, M. Werling, and S. Thrun, "Towards fully autonomous driving: systems and algorithms," in *IEEE Intelligent Vehicles Symposium (IV)* (2011), pp. 163–168.
9. S. Grigorescu, B. Trasnea, T. Cocias, and G. Macesanu, "A survey of deep learning techniques for autonomous driving," J. Field Robot. **37**, 362–386 (2020).
10. J. Hirschberg and C. D. Manning, "Advances in natural language processing," Science **349**, 261–266 (2015).
11. P. M. Nadkarni, L. Ohno-Machado, and W. W. Chapman, "Natural language processing: an introduction," J. Am. Med. Inf. Assoc. **18**, 544–551 (2011).
12. K. Chowdhary, "Natural language processing," in *Fundamentals of Artificial Intelligence* (2020), pp. 603–649.
13. D. Shen, G. Wu, and H.-I. Suk, "Deep learning in medical image analysis," Annu. Rev. Biomed. Eng. **19**, 221 (2017).
14. E. Gawehn, J. A. Hiss, and G. Schneider, "Deep learning in drug discovery," Mol. Inf. **35**, 3–14 (2016).
15. C. Angermueller, T. Pärnamaa, L. Parts, and O. Stegle, "Deep learning for computational biology," Mol. Syst. Biol. **12**, 878 (2016).
16. J. L. Hennessy and D. A. Patterson, *Computer Architecture: A Quantitative Approach* (Elsevier, 2011).
17. D. Kirk, "NVIDIA CUDA software and GPU parallel computing architecture," in *6th International Symposium on Memory Management* (2007), pp. 103–104.
18. N. P. Jouppi *et al.*, "In-datacenter performance analysis of a tensor processing unit," in *Proceedings of the 44th Annual International Symposium on Computer Architecture* (2017), pp. 1–12.
19. C. Zhang, P. Li, G. Sun, Y. Guan, B. Xiao, and J. Cong, "Optimizing FPGA-based accelerator design for deep convolutional neural networks," in *Proceedings of the 2015 ACM/SIGDA International Symposium on Field-Programmable Gate Arrays* (2015), pp. 161–170.
20. H. J. Caulfield and S. Dolev, "Why future supercomputing requires optics," Nat. Photonics **4**, 261–263 (2010).
21. D. A. Miller, "The role of optics in computing," Nat. Photonics **4**, 406 (2010).
22. J. Touch, A.-H. Badawy, and V. J. Sorger, "Optical computing," Nanophotonics **6**, 503–505 (2017).
23. G. Carleo, I. Cirac, K. Cranmer, L. Daudet, M. Schuld, N. Tishby, L. Vogt-Maranto, and L. Zdeborová, "Machine learning and the physical sciences," Rev. Mod. Phys. **91**, 045002 (2019).
24. B. J. Shastri, A. N. Tait, T. Ferreira de Lima, W. H. Pernice, H. Bhaskaran, C. D. Wright, and P. R. Prucnal, "Photonics for artificial intelligence and neuromorphic computing," Nat. Photonics **15**, 102–114 (2021).
25. W. Bogaerts, D. Pérez, J. Capmany, D. A. Miller, J. Poon, D. Englund, F. Morichetti, and A. Melloni, "Programmable photonic circuits," Nature **586**, 207–216 (2020).
26. D. Marković, A. Mizrahi, D. Querlioz, and J. Grollier, "Physics for neuromorphic computing," Nat. Rev. Phys. **2**, 499–510 (2020).
27. H. Zhou, J. Dong, J. Cheng, W. Dong, C. Huang, Y. Shen, Q. Zhang, M. Gu, C. Qian, H. Chen, Z. Ruan, and X. Zhang, "Photonic matrix multiplication lights up photonic accelerator and beyond," Light Sci. Appl. **11**, 30 (2022).
28. Q. Cheng, M. Bahadori, M. Glick, S. Rumley, and K. Bergman, "Recent advances in optical technologies for data centers: a review," Optica **5**, 1354–1370 (2018).
29. J. Yao, "Microwave photonics," J. Lightwave Technol. **27**, 314–335 (2009).
30. 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).
31. G. Wetzstein, A. Ozcan, S. Gigan, S. Fan, D. Englund, M. Soljačić, C. Denz, D. A. Miller, and D. Psaltis, "Inference in artificial intelligence with deep optics and photonics," Nature **588**, 39–47 (2020).
32. H. Zhang, M. Gu, X. Jiang, J. Thompson, H. Cai, S. Paesani, R. Santagati, A. Laing, Y. Zhang, M. Yung, Y. Z. Shi, F. K. Muhammad, G. Q. Lo, X. S. Luo, B. Dong, D. L. Kwong, L. C. Kwek, and A. Q. Liu, "An optical neural chip for implementing complex-valued neural network," Nat. Commun. **12**, 457 (2021).
33. A. N. Tait, T. F. De Lima, E. Zhou, A. X. Wu, M. A. Nahmias, B. J. Shastri, and P. R. Prucnal, "Neuromorphic photonic networks using silicon photonic weight banks," Sci. Rep. **7**, 7430 (2017).
34. A. N. Tait, M. A. Nahmias, B. J. Shastri, and P. R. Prucnal, "Broadcast and weight: an integrated network for scalable photonic spike processing," J. Lightwave Technol. **32**, 4029–4041 (2014).
35. C. Huang, S. Fujisawa, T. F. de Lima, A. N. Tait, E. C. Blow, Y. Tian, S. Bilodeau, A. Jha, F. Yaman, H.-T. Peng, H. G. Batshon, B. J. Shastri, Y. Inada, T. Wang, and P. R. Prucnal, "A silicon photonic–electronic neural network for fibre nonlinearity compensation," Nat. Electron. **4**, 837–844 (2021).
36. 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).
37. C. Wu, H. Yu, S. Lee, R. Peng, I. Takeuchi, and M. Li, "Programmable phase-change metasurfaces on waveguides for multimode photonic convolutional neural network," Nat. Commun. **12**, 96 (2021).
38. Y. Huang, W. Zhang, F. Yang, J. Du, and Z. He, "Programmable matrix operation with reconfigurable time-wavelength plane manipulation and dispersed time delay," Opt. Express **27**, 20456–20467 (2019).
39. X. Xu, M. Tan, B. Corcoran, J. Wu, T. G. Nguyen, A. Boes, S. T. Chu, B. E. Little, R. Morandotti, A. Mitchell, D. G. Hicks, and D. J. Moss, "Photonic perceptron based on a Kerr microcomb for high-speed, scalable, optical neural networks," Laser Photon. Rev. **14**, 2000070 (2020).
40. 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).
41. 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).
42. 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).
43. Z. Xu, X. Yuan, T. Zhou, and L. Fang, "A multichannel optical computing architecture for advanced machine vision," Light Sci. Appl. **11**, 255 (2022).
44. T. Yan, R. Yang, Z. Zheng, X. Lin, H. Xiong, and Q. Dai, "All-optical graph representation learning using integrated diffractive photonic computing units," Sci. Adv. **8**, eabn7630 (2022).
45. H. Zhu, J. Zou, H. Zhang, Y. Shi, S. Luo, N. Wang, H. Cai, L. Wan, B. Wang, X. Jiang, J. Thompson, X. S. Luo, X. H. Zhou, L. M. Xiao, W. Huang, L. Patrick, M. Gu, L. C. Kwek, and A. Q. Liu, "Space-efficient optical computing with an integrated chip diffractive neural network," Nat. Commun. **13**, 1044 (2022).
46. X. Zhao, H. Lv, C. Chen, S. Tang, X. Liu, and Q. Qi, "On-chip reconfigurable optical neural networks," 2021, https://assets.researchsquare.com/files/rs-155560/v1_stamped.pdf.
47. Z. Wang, L. Chang, F. Wang, T. Li, and T. Gu, "Integrated photonic metasystem for image classifications at telecommunication wavelength," Nat. Commun. **13**, 2131 (2022).
48. T. Fu, Y. Zang, H. Huang, Z. Du, C. Hu, M. Chen, S. Yang, and H. Chen, "On-chip photonic diffractive optical neural network based on a spatial domain electromagnetic propagation model," Opt. Express **29**, 31924–31940 (2021).

49. T. Fu, Y. Zang, Y. Huang, Z. Du, H. Huang, C. Hu, M. Chen, S. Yang, and H. Chen, "Photonic machine learning with on-chip diffractive optics," Nat. Commun. **14**, 70 (2023).
50. A. Hirose, *Complex-Valued Neural Networks: Theories and Applications* (World Scientific, 2003), Vol. **5**.
51. N. Özdemir, B. B. İskender, and N. Y. Özgür, "Complex valued neural network with Möbius activation function," Commun. Nonlinear Sci. Numer. Simulation **16**, 4698–4703 (2011).
52. S. Scardapane, S. Van Vaerenbergh, A. Hussain, and A. Uncini, "Complex-valued neural networks with nonparametric activation functions," IEEE Trans. Emerg. Top. Comput. Intell. **4**, 140–150 (2018).
53. X. Ding, X. Zhang, J. Han, and G. Ding, "Diverse branch block: building a convolution as an inception-like unit," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2021), pp. 10886–10895.
54. X. Ding, X. Zhang, N. Ma, J. Han, G. Ding, and J. Sun, "Repvgg: making vgg-style convnets great again," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2021), pp. 13733–13742.
55. X. Ding, T. Hao, J. Tan, J. Liu, J. Han, Y. Guo, and G. Ding, "Resrep: lossless CNN pruning via decoupling remembering and forgetting," in *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2021), pp. 4510–4520.
56. B. Goyal, A. Dogra, S. Agrawal, B. S. Sohi, and A. Sharma, "Image denoising review: from classical to state-of-the-art approaches," Inf. Fusion **55**, 220–244 (2020).
57. C. Tian, L. Fei, W. Zheng, Y. Xu, W. Zuo, and C.-W. Lin, "Deep learning on image denoising: an overview," Neural Netw. **131**, 251–275 (2020).
58. K. Zhang, W. Zuo, Y. Chen, D. Meng, and L. Zhang, "Beyond a Gaussian denoiser: residual learning of deep cnn for image denoising," IEEE Trans. Image Process. **26**, 3142–3155 (2017).
59. Y. Chen and T. Pock, "Trainable nonlinear reaction diffusion: a flexible framework for fast and effective image restoration," IEEE Trans. Pattern Anal. Mach. Intell. **39**, 1256–1272 (2016).
60. A. Rahim, A. Hermans, B. Wohlfeil, D. Petousi, B. Kuyken, D. Van Thourhout, and R. G. Baets, "Taking silicon photonics modulators to a higher performance level: state-of-the-art and a review of new technologies," Adv. Photon. **3**, 024003 (2021).
61. A. Samani, M. Chagnon, D. Patel, V. Veerasubramanian, S. Ghosh, M. Osman, Q. Zhong, and D. V. Plant, "A low-voltage 35-GHz silicon photonic modulator-enabled 112-Gb/s transmission system," IEEE Photon. J. **7**, 7901413 (2015).
62. T. Baehr-Jones, R. Ding, Y. Liu, A. Ayazi, T. Pinguet, N. C. Harris, M. Streshinsky, P. Lee, Y. Zhang, A. E.-J. Lim, T.-Y. Liow, S. H.-G. Teo, G.-Q. Lo, and M. Hochberg, "Ultralow drive voltage silicon traveling-wave modulator," Opt. Express **20**, 12014–12020 (2012).
63. M. Sakib, P. Liao, C. Ma, R. Kumar, D. Huang, G.-L. Su, X. Wu, S. Fathololoumi, and H. Rong, "A high-speed micro-ring modulator for next generation energy-efficient optical networks beyond 100 Gbaud," in *CLEO: Science and Innovations* (2021), paper SF1C–3.
64. C. Liu, J. Guo, L. Yu, J. Li, M. Zhang, H. Li, Y. Shi, and D. Dai, "Silicon/2D-material photodetectors: from near-infrared to mid-infrared," Light Sci. Appl. **10**, 1 (2021).
65. Y.-Q. Bie, G. Grosso, M. Heuck, M. M. Furchi, Y. Cao, J. Zheng, D. Bunandar, E. Navarro-Moratalla, L. Zhou, D. K. Efetov, T. Taniguchi, K. Watanabe, J. Kong, D. Englund, and P. Jarillo-Herrero, "A $MoTe_2$-based light-emitting diode and photodetector for silicon photonic integrated circuits," Nat. Nanotechnol. **12**, 1124–1129 (2017).
66. S. Chetlur, C. Woolley, P. Vandermersch, J. Cohen, J. Tran, B. Catanzaro, and E. Shelhamer, "CUDNN: efficient primitives for deep learning," arXiv, arXiv:1410.0759 (2014).
67. J. Choquette, W. Gandhi, O. Giroux, N. Stam, and R. Krashinsky, "NVIDIA A100 tensor core GPU: performance and innovation," IEEE Micro **41**, 29–35 (2021).
68. H. Liao, J. Tu, J. Xia, and X. Zhou, "Davinci: a scalable architecture for neural network computing," in *Hot Chips Symposium* (2019), pp. 1–44.
69. N. Jouppi, C. Young, N. Patil, and D. Patterson, "Motivation for and evaluation of the first tensor processing unit," IEEE Micro **38**, 10–19 (2018).
70. P. Yao, H. Wu, B. Gao, J. Tang, Q. Zhang, W. Zhang, J. J. Yang, and H. Qian, "Fully hardware-implemented memristor convolutional neural network," Nature **577**, 641–646 (2020).
71. Y. Li, J. Dongarra, and S. Tomov, "A note on auto-tuning GEMM for GPUs," in *Computational Science–ICCS 2009: 9th International Conference* (2009), pp. 884–892.
72. R. Nath, S. Tomov, and J. Dongarra, "An improved magma gemm for fermi graphics processing units," Int. J. High Performance Comput. Appl. **24**, 511–515 (2010).
73. D. Yan, W. Wang, and X. Chu, "Demystifying tensor cores to optimize half-precision matrix multiply," in *IEEE International Parallel and Distributed Processing Symposium (IPDPS)* (2020), pp. 634–643.
74. S. Xu, J. Wang, S. Yi, and W. Zou, "High-order tensor flow processing using integrated photonic circuits," Nat. Commun. **13**, 7970 (2022).
75. V. Bangari, B. A. Marquez, H. Miller, A. N. Tait, M. A. Nahmias, T. F. De Lima, H.-T. Peng, P. R. Prucnal, and B. J. Shastri, "Digital electronics and analog photonics for convolutional neural networks (DEAP-CNNS)," IEEE J. Sel. Top. Quantum Electron. **26**, 7701213 (2019).
76. S. Xu, J. Wang, H. Shu, Z. Zhang, S. Yi, B. Bai, X. Wang, J. Liu, and W. Zou, "Optical coherent dot-product chip for sophisticated deep learning regression," Light Sci. Appl. **10**, 221 (2021).
77. 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, and X. Wang, "Microcomb-based integrated photonic processing unit," Nat. Commun. **14**, 66 (2023).
78. W. M. Green, M. J. Rooks, L. Sekaric, and Y. A. Vlasov, "Ultra-compact, low RF power, 10 Gb/s silicon Mach-Zehnder modulator," Opt. Express **15**, 17106–17113 (2007).
79. C. Sun, M. T. Wade, Y. Lee, J. S. Orcutt, L. Alloatti, M. S. Georgas, A. S. Waterman, J. M. Shainline, R. R. Avizienis, S. Lin, B. R. Moss, R. Kumar, F. Pavanello, A. H. Atabaki, H. M. Cook, A. J. Ou, J. C. Leu, Y.-H. Chen, K. Asanović, R. J. Ram, M. A. Popović, and V. M. Stojanović, "Single-chip microprocessor that communicates directly using light," Nature **528**, 534–538 (2015).
80. Y. LeCun, "1.1 deep learning hardware: past, present, and future," in *IEEE International Solid-State Circuits Conference (ISSCC)* (2019), pp. 12–19.
81. K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2016), pp. 770–778.
82. K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," arXiv, arXiv:1409.1556 (2014).
83. R. R. Schaller, "Moore's law: past, present and future," IEEE Spectrum **34**, 52–59 (1997).
84. W. Shi, J. Cao, Q. Zhang, Y. Li, and L. Xu, "Edge computing: vision and challenges," IEEE Internet Things J. **3**, 637–646 (2016).
85. Y. Mao, C. You, J. Zhang, K. Huang, and K. B. Letaief, "A survey on mobile edge computing: the communication perspective," Commun. Surveys Tuts. **19**, 2322–2358 (2017).
86. M. Satyanarayanan, "The emergence of edge computing," Computer **50**, 30–39 (2017).
87. D. Ielmini and H.-S. P. Wong, "In-memory computing with resistive switching devices," Nat. Electron. **1**, 333–343 (2018).
88. A. Sebastian, M. Le Gallo, R. Khaddam-Aljameh, and E. Eleftheriou, "Memory devices and applications for in-memory computing," Nat. Nanotechnol. **15**, 529–544 (2020).
89. N. Verma, H. Jia, H. Valavi, Y. Tang, M. Ozatay, L.-Y. Chen, B. Zhang, and P. Deaville, "In-memory computing: advances and prospects," IEEE Solid-State Circuits Mag. **11**, 43–55 (2019).
90. Y. Chen, G. Yin, Z. Tan, M. Lee, Z. Yang, Y. Liu, H. Yang, K. Ma, and X. Li, "Yoloc: deploy large-scale neural network by rom-based computing-in-memory using residual branch on a chip," arXiv, arXiv:2206.00379 (2022).
91. B. Mukherjee, "WDM optical communication networks: progress and challenges," IEEE J. Sel. Areas Commun. **18**, 1810–1824 (2000).
92. B. B. Buckley, S. T. Fryslie, K. Guinn, G. Morrison, A. Gazman, Y. Shen, K. Bergman, M. L. Mashanovitch, and L. A. Johansson, "WDM source based on high-power, efficient 1280-nm DFB lasers for terabit interconnect technologies," IEEE Photon. Technol. Lett. **30**, 1929–1932 (2018).

93. L. Chang, S. Liu, and J. E. Bowers, "Integrated optical frequency comb technologies," Nat. Photonics **16**, 95–108 (2022).
94. T. J. Kippenberg, R. Holzwarth, and S. A. Diddams, "Microresonator-based optical frequency combs," Science **332**, 555–559 (2011).
95. Y. K. Chembo, "Kerr optical frequency combs: theory, applications and perspectives," Nanophotonics **5**, 214–230 (2016).
96. Y. Shoji, K. Kintaka, S. Suda, H. Kawashima, T. Hasama, and H. Ishikawa, "Low-crosstalk 2 × 2 thermo-optic switch with silicon wire waveguides," Opt. Express **18**, 9071–9075 (2010).
97. L. Lu, S. Zhao, L. Zhou, D. Li, Z. Li, M. Wang, X. Li, and J. Chen, "16 × 16 non-blocking silicon optical switch based on electro-optic Mach-Zehnder interferometers," Opt. Express **24**, 9295–9307 (2016).
98. L. Qiao, W. Tang, and T. Chu, "32 × 32 silicon electro-optic switch with built-in monitors and balanced-status units," Sci. Rep. **7**, 42306 (2017).
99. P. Dong, S. F. Preble, and M. Lipson, "All-optical compact silicon comb switch," Opt. Express **15**, 9600–9605 (2007).
100. Y. H. Wen, O. Kuzucu, T. Hou, M. Lipson, and A. L. Gaeta, "All-optical switching of a single resonance in silicon ring resonators," Opt. Lett. **36**, 1413–1415 (2011).
101. B. G. Lee, A. Biberman, P. Dong, M. Lipson, and K. Bergman, "All-optical comb switch for multiwavelength message routing in silicon photonic networks," IEEE Photon. Technol. Lett. **20**, 767–769 (2008).
102. N. Sherwood-Droz, H. Wang, L. Chen, B. G. Lee, A. Biberman, K. Bergman, and M. Lipson, "Optical 4 × 4 hitless silicon router for optical networks-on-chip (NOC)," Opt. Express **16**, 15915–15922 (2008).
103. S. Han, T. J. Seok, K. Yu, N. Quack, R. S. Muller, and M. C. Wu, "Large-scale polarization-insensitive silicon photonic MEMS switches," J. Lightwave Technol. **36**, 1824–1830 (2018).
104. K. Kwon, T. J. Seok, J. Henriksson, J. Luo, L. Ochikubo, J. Jacobs, R. S. Muller, and M. C. Wu, "128 × 128 silicon photonic MEMS switch with scalable row/column addressing," in *CLEO: Science and Innovations* (2018), paper SF1A–4.
105. H. Y. Hwang, J. S. Lee, T. J. Seok, A. Forencich, H. R. Grant, D. Knutson, N. Quack, S. Han, R. S. Muller, G. C. Papen, M. C. Wu, and P. O'Brien, "Flip chip packaging of digital silicon photonics MEMS switch for cloud computing and data centre," IEEE Photon. J. **9**, 2900210 (2017).
106. L. Liao, D. Samara-Rubio, M. Morse, A. Liu, D. Hodge, D. Rubin, U. D. Keil, and T. Franck, "High speed silicon Mach-Zehnder modulator," Opt. Express **13**, 3129–3135 (2005).
107. C. Wang, M. Zhang, X. Chen, M. Bertrand, A. Shams-Ansari, S. Chandrasekhar, P. Winzer, and M. Lončar, "Integrated lithium niobate electro-optic modulators operating at CMOS-compatible voltages," Nature **562**, 101–104 (2018).
108. Y. Ogiso, J. Ozaki, Y. Ueda, H. Wakita, M. Nagatani, H. Yamazaki, M. Nakamura, T. Kobayashi, S. Kanazawa, Y. Hashizume, H. Tanobe, N. Nunoya, M. Ida, Y. Miyamoto, and M. Ishikawa, "80-GHz bandwidth and 1.5-V V_π InP-based IQ modulator," J. Lightwave Technol. **38**, 249–255 (2019).
109. Z. Zhao, J. Liu, Y. Liu, and N. Zhu, "High-speed photodetectors in optical communication system," J. Semicond. **38**, 121001 (2017).
110. S. Malyshev and A. Chizh, "State of the art high-speed photodetectors for microwave photonics application," in *15th International Conference on Microwaves, Radar and Wireless Communications* (2004), pp. 765–775.
111. A. E.-J. Lim, J. Song, Q. Fang, C. Li, X. Tu, N. Duan, K. K. Chen, R. P.-C. Tern, and T.-Y. Liow, "Review of silicon photonics foundry efforts," IEEE J. Sel. Top. Quantum Electron. **20**, 405–416 (2013).
112. S. Y. Siew, B. Li, F. Gao, H. Y. Zheng, W. Zhang, P. Guo, S. W. Xie, A. Song, B. Dong, L. W. Luo, C. Li, X. Luo, and G.-Q. Lo, "Review of silicon photonics technology and platform development," J. Lightwave Technol. **39**, 4374–4389 (2021).