ACS Photonics

pubs.acs.org/journal/apchd5

Article

A Compact Butterfly-Style Silicon Photonic−Electronic Neural Chip for Hardware-Efficient Deep Learning

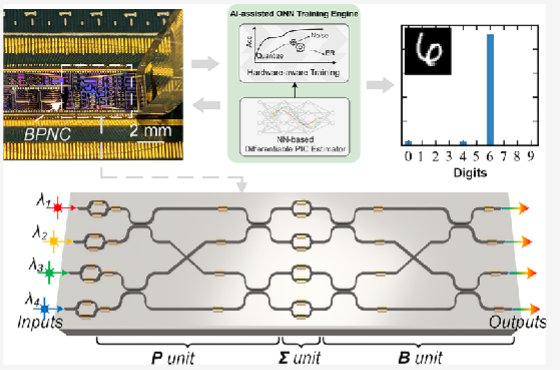
Chenghao Feng,[#] Jiaqi Gu,[#] Hanqing Zhu, Zhoufeng Ying, Zheng Zhao, David Z. Pan,* and Ray T. Chen*

Cite This: ACS Photonics 2022, 9, 3906−3916

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: The optical neural network (ONN) is a promising hardware platform for next-generation neurocomputing due to its high parallelism, low latency, and low energy consumption. Previous ONN architectures are mainly designed for general matrix multiplication (GEMM), leading to unnecessarily large area cost and high control complexity. Here, we move beyond classical GEMM-based ONNs and propose an optical subspace neural network (OSNN) architecture, which trades the universality of weight representation for lower optical component usage, area cost, and energy consumption. We devise a butterfly-style photonic−electronic neural chip to implement our OSNN with up to 7× fewer trainable optical components compared to GEMM-based ONNs. Additionally, a hardware-aware training framework is provided to minimize the required device programming precision, lessen the chip area, and boost the noise robustness. We experimentally demonstrate the utility of our neural chip in practical image recognition tasks, showing that a measured accuracy of 94.16% can be achieved in handwritten digit recognition tasks with 3 bit weight programming precision.



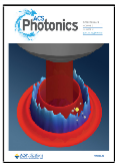
KEYWORDS: *optical subspace neural network, butterfly-style photonic−electronic neural chip, hardware-aware training, hardware efficiency, deep learning*

1. INTRODUCTION

Deep neural networks (DNNs) have demonstrated superior performance in various intelligence tasks, such as image recognition, decision making, and language translation.[1−4] Hardware accelerators capable of performing high-speed, energy-efficient, and parallel multiply−accumulate (MAC) operations are in high demand with the rapidly escalating DNN model size and data volume. However, electronic digital hardware accelerators, including but not limited to graphical processing units (GPUs), field-programmable gate arrays (FPGAs),[5] and other digital application-specific integrated circuits (ASICs),[6] are inevitably limited by millisecond-level latency, high energy consumption, excessive heat, and high interconnect cost.[7,8] In contrast, analog neuromorphic computing represents a paradigm shift in efficient DNN acceleration, significantly increasing parallelism and energy efficiency.[9,10]

The optical neural network (ONN) is a promising analog artificial intelligence (AI) accelerator that features low latency, wide bandwidth, and high parallelism of light.[11−15] Earlier work has presented a range of high-performance integrated photonic neural networks that implement multilayer perceptrons (MLPs)[11,16,17] or convolutional neural networks (CNNs).[18,19] The fundamental matrix−vector multiplication (MVM) unit is realized using Mach−Zehnder interferometer (MZI) arrays or microring-resonator (MRR) arrays. By tuning the phase shifters in MZIs or the transmission of MRRs, these photonic systems are designed to implement universal linear operations or general matrix multiplication (GEMM) with a relatively high requirement in device control precision. Recent studies show that the construction of DNNs can move beyond conventional GEMM with restricted matrix parameter space, e.g., low-rank NNs[20−22] and structured NNs,[23−25] which shows not only considerable hardware efficiency improvement but also comparable representability to classical GEMM-based NNs. We refer to such NN architectures as subspace neural networks. The success of such a design concept can be reproduced in ONNs by trading the universality of weight representation for higher hardware efficiency. Several structured ONNs have been proposed to reduce the number of optical components, e.g., the fast-Fourier-transform-based

Received: July 31, 2022
Published: November 30, 2022



https://doi.org/10.1021/acsphotonics.2c01188
ACS Photonics 2022, 9, 3906−3916

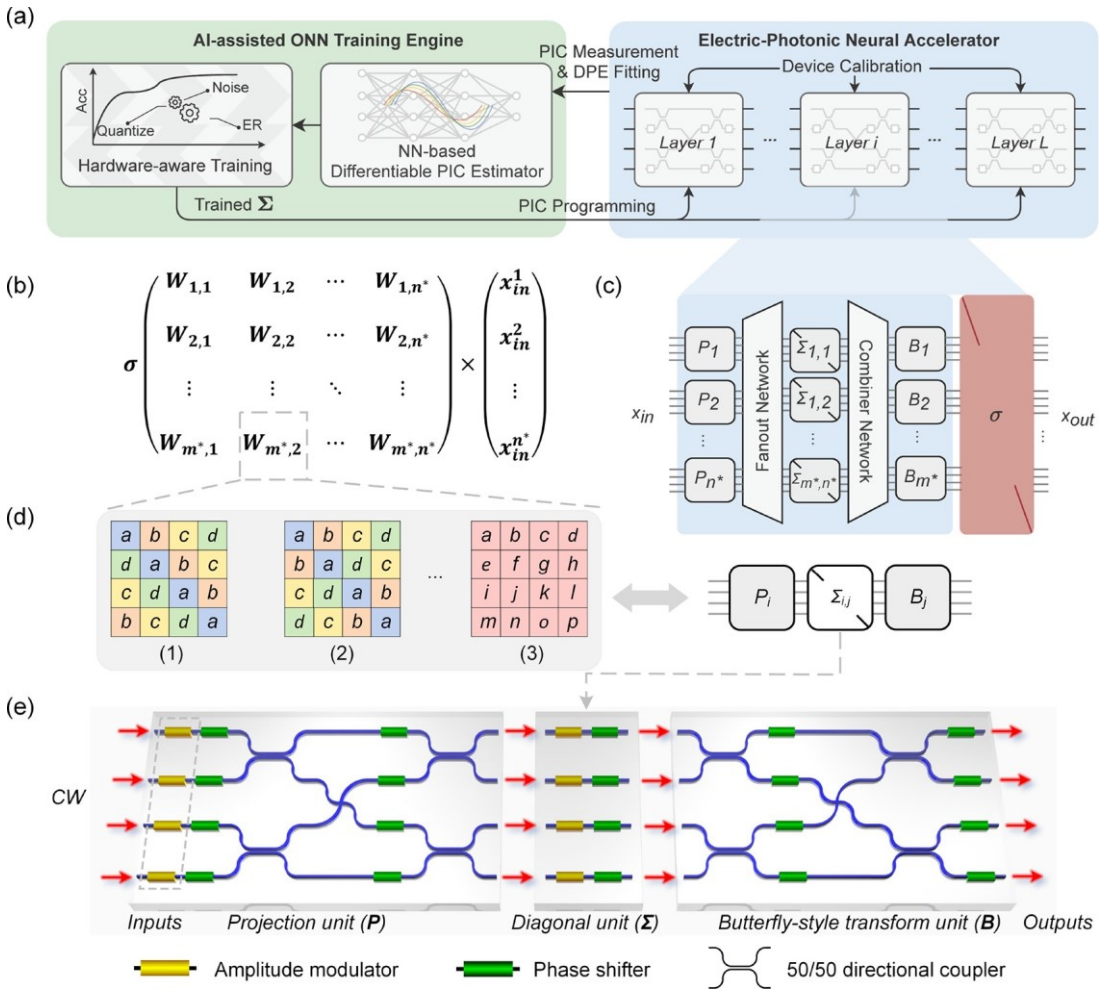


Figure 1. General architecture of the optical subspace neural network (OSNN). The hardware-aware training framework is shown in (a). The mathematical representation of one layer in our OSNN is shown in (b), where an $m \times n$ matrix is partitioned to $\frac{mn}{k^2}$ blocks. The hardware representation of one layer is shown in (c), which consists of $n^* = \frac{n}{k}$ projection units ($\boldsymbol{P}$ units), $m^* = \frac{m}{k}$ butterfly-style transform units ($\boldsymbol{B}$ units), and $m^* \times n^*$ diagonal matrix units ($\boldsymbol{\Sigma}$ units). The fanout network and the combiner network are used to distribute and combine the optical signals to different optical paths. By setting phase shifters in $\boldsymbol{B}$ units and $\boldsymbol{P}$ units, several popular structured ONNs based on (1) fast Fourier transform and inverse fast Fourier transform, (2) Hadamard transform, or (3) other implementable transforms are shown in (d). (e) Schematic of 4×4 $\boldsymbol{B}$, $\boldsymbol{P}$, and $\boldsymbol{\Sigma}$ units, respectively, which are the building blocks of a four-point OSNN. To train the OSNN, we first perform an on-chip device calibration on our OSNN. The measurement data are learned and modeled in our hardware-aware training flow. Moreover, other factors such as the precision of controlling signals, limited extinction ratio (ER) of input modulators, and other noises are also considered in our training engine to improve the accuracy and robustness of our OSNN.

(FFT-based) ONN.[25−27] In this work, we further explore this subspace NN design concept in the optical domain and experimentally demonstrate a novel butterfly-style photonic−electronic neural chip (BPNC) with superior hardware efficiency and compactness.

Additionally, noise-tolerant ONN training currently still lacks an efficient, scalable, and physically evaluated solution. As is the case with other analog computing platforms, ONNs will inevitably encounter performance degradation or even malfunction due to non-ideal factors, e.g., process variations,[28,29] limited control precision,[30,31] and dynamic noises.[32] Recently, on-chip training has become an appealing trend toward noise-resilient ONNs. Numerous on-chip training algorithms have been proposed to directly optimize optical devices with in situ noise handling.[33−37] However, prior ONN on-chip training protocols suffer from algorithmic inefficiency and require costly hardware overhead, e.g., phase detection,[33] high-resolution optical component control,[34] or per-device field monitoring.[33,37] Therefore, applying them to practical ONN training is still technically challenging.

In this work, we propose an OSNN for next-generation hardware-efficient deep learning. Our proposed OSNN partitions each layer's weight matrix into smaller $k \times k$ (k = 4, 8) submatrices with restricted parameter space. Our architecture can achieve photonic neural computing with 7× fewer trainable optical components compared to MZI-based ONN architectures designed for general MVMs,[11] resulting in a 3.3× smaller device footprint and 5.5× lower latency. The number of trainable optical components can be further reduced by ∼70% using structured circuit pruning[38] with negligible (<0.2%) task performance loss. Moreover, an efficient and scalable hardware-aware training framework is experimentally deployed to enable ONN training with high noise robustness and a low control precision requirement. Our

https://doi.org/10.1021/acsphotonics.2c01188
ACS Photonics 2022, 9, 3906−3916

OSNN is then experimentally demonstrated on a 4×4 BPNC and evaluated by a MNIST handwritten digit classification task[39] with a measured accuracy of 94.16%. Our performance analysis reveals that our OSNN can achieve a computational density of ~225 tera (10^{12}) operations per second/mm^2 (TOPS/mm^2) and energy efficiency of ~9.5 TOPS/W using compact optical devices, e.g., microdisk-based active devices.[40] Our proposed OSNN architecture and hardware-aware training framework provide a synergistic solution that unleashes the power of optics from a novel co-design perspective and pushes the limits of next-generation efficient AI.

2. OPTICAL SUBSPACE NEURAL NETWORK

The proposed OSNN and its training framework are depicted in Figure 1a. The mathematical representation of one layer in a typical DNN with n inputs and m outputs is shown in Figure 1(b), along with its hardware implementation shown in Figure 1(c). Here we partition the $m \times n$ weight matrix $\boldsymbol{W}$ into $m^* \times n^*$ submatrices $\{\boldsymbol{W}_{i,j} \in \mathbb{C}^{k\times k}\}_{i\in[n^*],j\in[m^*]}$. The input vector $\boldsymbol{x}_{\text{in}}$ is encoded as the amplitude of the optical signals and will also be partitioned into n^* segments $\boldsymbol{x}_{\text{in}} = (\boldsymbol{x}_{\text{in}}^1, \boldsymbol{x}_{\text{in}}^2, ..., \boldsymbol{x}_{\text{in}}^{n^*})$. Thus, the MVM operation can be expressed using the block matrix multiplication formula as follows,

$$\boldsymbol{x}'_{\text{out}} = \boldsymbol{W}\boldsymbol{x}_{\text{in}} = \begin{bmatrix} \sum_{j=1}^{n^*} \boldsymbol{W}_{1,j}\boldsymbol{x}_{\text{in}}^j \\ \sum_{j=1}^{n^*} \boldsymbol{W}_{2,j}\boldsymbol{x}_{\text{in}}^j \\ \vdots \\ \sum_{j=1}^{n^*} \boldsymbol{W}_{m^*,j}\boldsymbol{x}_{\text{in}}^j \end{bmatrix} \tag{1}$$

Each submatrix $\boldsymbol{W}_{i,j}$ can be decomposed as $\boldsymbol{W}_{i,j} = \boldsymbol{B}\boldsymbol{\Sigma}_{i,j}\boldsymbol{P}$, where $\boldsymbol{B}$ and $\boldsymbol{P}$ are both $k \times k$ unitary matrices shown in Figure 1d, and $\boldsymbol{\Sigma}_{i,j}$ is a $k \times k$ diagonal matrix. Here, we use two butterfly-style programmable photonic integrated circuits (PICs), namely, a butterfly-style transform unit ($\boldsymbol{B}$ unit) and a projection unit ($\boldsymbol{P}$ unit), to implement the unitary matrices $\boldsymbol{B}$ and $\boldsymbol{P}$ using phase shifters, directional couplers, and waveguide crossings, while the diagonal matrix unit ($\boldsymbol{\Sigma}$ unit) is composed with a column of modulators. Figure 1e shows the photonic circuit structure of these matrix units when $k = 4$. Compared to previous work with similar butterfly-style circuit topology, which uses MZI modulators as building blocks,[41,42] our BPNC uses more basic components as building blocks such as directional couplers and phase shifters. As a result, we save ~50% number of optical components compared to previous work.

One of the key advantages of our OSNN is that the chip footprint and the total number of trainable optical devices are considerably smaller than previous GEMM-based MZI-ONN. Specifically, while an $k \times k$ MZI array consumes (k^2) MZIs, our $\boldsymbol{B}$ and $\boldsymbol{P}$ units only use $(k\log_2 k)$ couplers and phase shifters. Moreover, instead of having all devices to be trainable, only the $\boldsymbol{\Sigma}_{i,j}$ units need to be trained. The $\boldsymbol{B}$ and $\boldsymbol{P}$ units will not be modified throughout the training, and mapping processes after their desired states are accomplished by tuning the phase shifters in them. As a result, the total number of trainable optical devices is $\frac{mn}{k}$ in an n-input, m-output layer, significantly reducing the weight loading cost and reprogramming complexity.

Based on the statistical evaluation,[43] our butterfly-style $\boldsymbol{B\Sigma P}$ block demonstrates good flexibility and matrix expressivity by only using $1/k$ total trainable components compared to MZI arrays (details in the Supporting Information). The $\boldsymbol{B}$ and $\boldsymbol{P}$ units in OSNN can flexibly support a wide range of unitary transforms. For instance, when $k = 4$, our butterfly unit $\boldsymbol{B}$ itself can express 80.2% arbitrary unitary matrices and the $\boldsymbol{B\Sigma P}$ block can realize 64.4% fidelity in expressing general matrices (see Figure S3). As depicted in Figure 1d, several commonly used structured matrices can be realized by configuring the phase shifters in $\boldsymbol{B}$ and $\boldsymbol{P}$ units. For example, a block-circulant matrix can be realized (Figure 1d1) when the $\boldsymbol{P}$ unit performs optical FFT, while the $\boldsymbol{B}$ unit performs optical inverse FFT (IFFT).[23] Furthermore, our $\boldsymbol{B}$ and $\boldsymbol{P}$ units can realize Hadamard transformation (HT), which is a popular choice to construct efficient DNNs.[44] Detailed proves can be found in the Supporting Information. Figure 1d2 shows the matrix pattern when both $\boldsymbol{P}$ and $\boldsymbol{B}$ units implement HT. The superior versatility and expressivity of our $\boldsymbol{B\Sigma P}$ units guarantee that our OSNN can have enough learning capability.

Another essential property of our OSNN is that different $\boldsymbol{\Sigma}_{i,j}$ units can share the $\boldsymbol{B}$ and $\boldsymbol{P}$ units, leading to significant chip area reduction. Since all $\boldsymbol{W}_{i,j}$s are constructed by the same $\boldsymbol{B}$, $\boldsymbol{P}$ transforms, $\boldsymbol{B}$ and $\boldsymbol{P}$ units can be reused in the optical domain. Here, we rewrite eq 1 with matrix multiplication's distributive and associative properties:

$$\boldsymbol{x}'_{out} = \boldsymbol{W}\boldsymbol{x}_{in} = \begin{pmatrix} \sum_{j=1}^{n^*} \boldsymbol{B}\boldsymbol{\Sigma}_{1,j}\boldsymbol{P}\boldsymbol{x}_j \\ \sum_{j=1}^{n^*} \boldsymbol{B}\boldsymbol{\Sigma}_{2,j}\boldsymbol{P}\boldsymbol{x}_j \\ \vdots \\ \sum_{j=1}^{n^*} \boldsymbol{B}\boldsymbol{\Sigma}_{m^*,j}\boldsymbol{P}\boldsymbol{x}_j \end{pmatrix} = \begin{pmatrix} \boldsymbol{B}\sum_{j=1}^{n^*} \boldsymbol{\Sigma}_{1,j}\aleph_j \\ \boldsymbol{B}\sum_{j=1}^{n^*} \boldsymbol{\Sigma}_{2,j}\aleph_j \\ \vdots \\ \boldsymbol{B}\sum_{j=1}^{n^*} \boldsymbol{\Sigma}_{m^*,j}\aleph_j \end{pmatrix}, \tag{2}$$

where $\aleph_j = \boldsymbol{P}\boldsymbol{x}_j$. By sharing the unitary matrix units, one can implement an MVM operation of size $m \times n$ with only $m/k\boldsymbol{P}$ units and $n/k\boldsymbol{B}$ units, dramatically reducing the footprint of the OSNN compared to previous FFT-based ONN architecture, which requires $\frac{mn}{k}\boldsymbol{P}$ and $\boldsymbol{B}$ units.[25] The mechanism of the optical architecture shown in Figure 1b can then be stated as follows: First, the input optical signal $\boldsymbol{x}_{in}$ is partitioned into n^* segments $\boldsymbol{x}_j$s and then propagate through the $\boldsymbol{P}$ units to generate $\aleph_j$s, which will then be distributed via a fanout network to $m^* \times n^*$ diagonal matrix units $\boldsymbol{\Sigma}_{i,j}$s. After propagating through the $\boldsymbol{\Sigma}_{i,j}$ units, the signals will be combined and fed into each $\boldsymbol{B}$ unit with a combiner network to obtain the MVM result. Finally, as with all typical NN architectures, a nonlinear activation unit (σ unit) is required to generate the output of the layer $\boldsymbol{x}_{\text{out}}$, which has been realized by all-optical non-linear devices or optoelectronic circuits in previous work.[45,46] In this work, we assume the activation functions are realized electronically. We omit details on this and do not consider its effects on system performance in our discussion hereinafter.

3. HARDWARE-AWARE TRAINING FRAMEWORK

After manufacturing, fabrication variances in optical components and dynamic noises will introduce uncertainty into the ONN. Moreover, the numerical resolution of implementable weight matrices is limited by the precision of electrical signals used to program the ONN. Consequently, the task performance will deteriorate. To remedy this robustness issue, we build a multistage hardware-aware training framework.

The general procedures of the training framework are summarized in Figure 1a, and more details are provided in the

https://doi.org/10.1021/acsphotonics.2c01188
ACS Photonics 2022, 9, 3906−3916

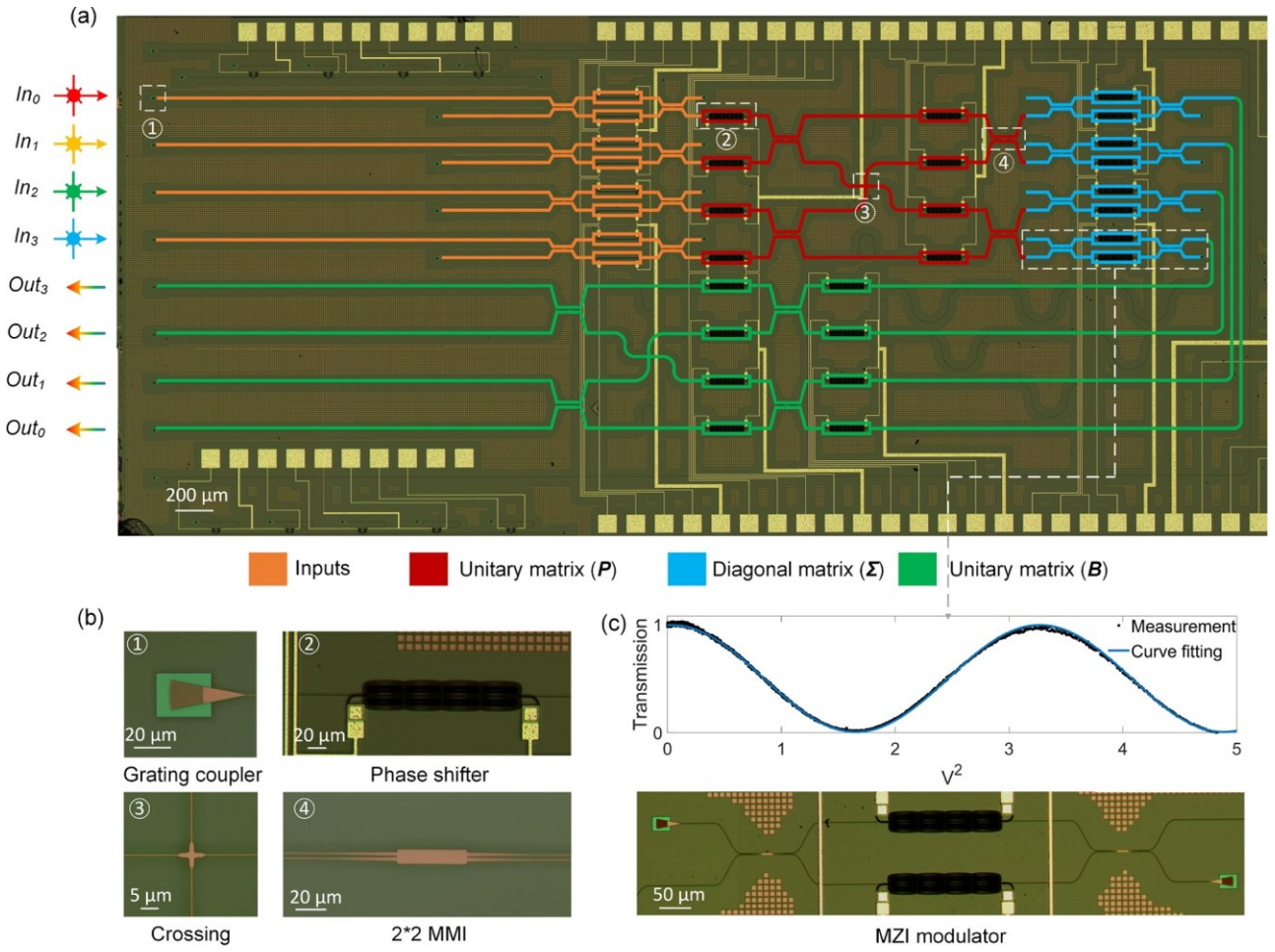


Figure 2. Schematic of the butterfly-style silicon photonic−electronic neural chip. The micrograph of the neural chip is shown in (a). The input optical beams with different wavelengths are shown in different colors. The necessary optical components are highlighted in (b). (c) shows the schematic and the normalized transmission curve of an MZI attenuator in the diagonal matrix unit (**Σ** unit). Only the attenuators in **Σ** are programmed in training.

Supporting Information. First, the OSNN is calibrated on-chip to measure the performance of tunable components and prepare the state of the photonic chip to approach the desired transfer matrix. In reality, the actual transfer matrix of the photonic neural chip deviates from the designed one due to performance variations of the optical components, e.g., the unbalanced splitting ratio of directional couplers. Second, to model the non-ideal behavior and predict the response of the real optical neural chip, we develop an NN-based differentiable PIC estimator (DPE) using measurement data and AI algorithms. Our DPE explicitly models the behavior of the real physical chip during forward and gradient backpropagation that enables gradient-based physical-variation-aware optimization. The third step involves determining the DNN parameters and mapping them to the electrical control signals using a hardware-aware training and parameter mapping process. The DPE is used to efficiently emulate the real chip response to enable variation-aware gradient backpropagation. Quantization-aware training with dynamic noise injection techniques is used to improve the noise tolerance with limited device control resolution. Thus, our OSNN can achieve the performance target despite control precision restrictions and other non-idealities. More details of our hardware-aware training framework are shown in the Supporting Information. Compared to prior on-chip training protocols based on derivative-free optimization algorithms[11,35,36,47] and gradient approximation using ideal simulation models,[48−50] our AI-assisted ONN learning shows considerably higher scalability and effectiveness in robust optical neural chip training.

4. EXPERIMENT

In this work, we experimentally demonstrate the practicality of the OSNN on the silicon photonics platform using a butterfly-style photonic−electronic neural chip (BPNC) capable of implementing 4×4 $B\Sigma P$ blocks in our OSNN. The layout of the chip was drawn and verified using Synopsys OptoDesigner, while the chip was fabricated by the Advanced Micro Foundry (AMF). The schematic of the BPNC is shown in Figure 2a, while the close-ups of its components, such as phase shifters, 50−50 directional couplers, and crossings, are depicted in Figure 2b. The unitary matrix units B/P are marked in red/green in Figure 2a. The active phase shifters in these regions support enough flexibility to realize different unitary transforms, but note that they are not optimized as parameters during ONN training. The diagonal matrix unit (**Σ** unit) is built using an array of MZI attenuators for magnitude and phase control.[51] One MZI attenuator and its transmission are shown in Figure 2c.

The schematic of the testing setup is shown in Figure 3. Continuous-wave (CW) light of different wavelengths is

https://doi.org/10.1021/acsphotonics.2c01188
ACS Photonics 2022, 9, 3906−3916

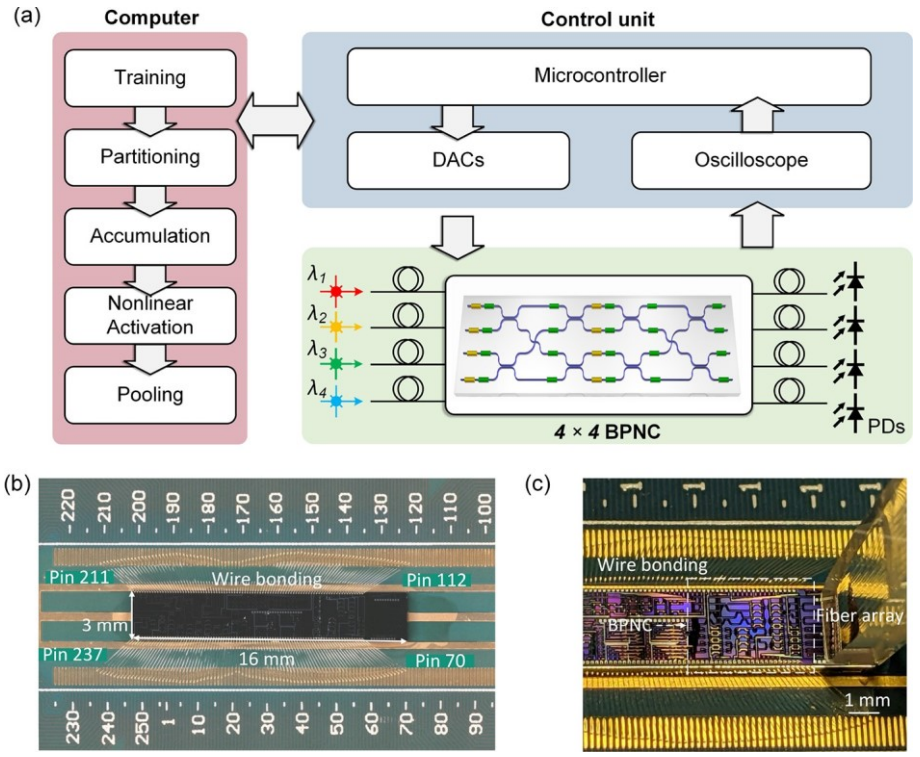


Figure 3. Experimental setup of OSNN. (a) Schematic of our OSNN test flow. The entire MVM is first partitioned into multiple 4×4 blocks, and each block is implemented optically on a butterfly-style photonic−electronic neural chip (BPNC). The wavelengths of input signals are λ_1 = 1548.0 nm, λ_2 = 1549.0 nm, λ_3 = 1550.0 nm, and λ_4 = 1551.0 nm. (b) shows the wire-bonded photonic chip and its starting/ending electrical pin numbers, while (c) is the photography of the chip testing setup. The parameters and the input signals are programmed by a multichannel digital-to-analog converter (DAC), while the output signals are read by the oscilloscope. Both the oscilloscope and the DAC are controlled by a microcontroller. The MVM results are provided to the computer for data processing in order to train and deploy the DNN.

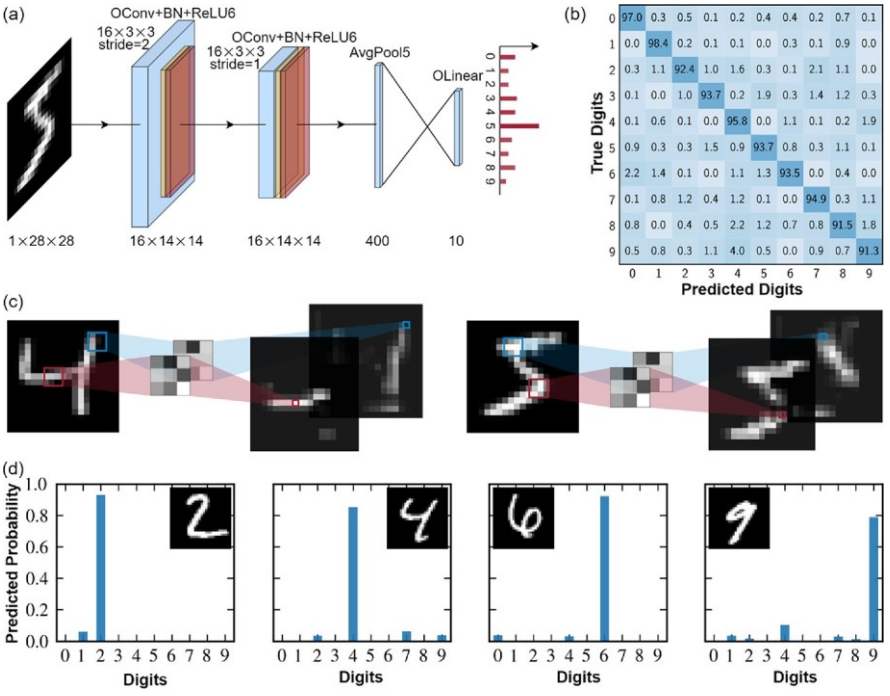


Figure 4. Experimental data of digit recognition with the OSNN. (a) Structure of the CNN; the convolution is realized by OSNN with the im2col approach. The first convolutional layer has one input channel and 16 output channels with a stride of 2. The subsequent convolutional layer has 16 input/output channels with a stride of 1, and the size of the convolutional kernel is 3 × 3. After adaptive average pooling, we have 5 × 5 × 16 = 400 hidden features, followed by a linear classifier with 10 outputs. (b) Confusion matrix of the trained OSNN on MNIST, showing a measured accuracy of 94.16%. (c) Experimental results of convolving two input images with convolution kernels of size 3 × 3 in our OSNN. (d) Predicted probability distribution of our OSNN on four selected test digits in the MNIST dataset.

https://doi.org/10.1021/acsphotonics.2c01188
ACS Photonics 2022, 9, 3906−3916

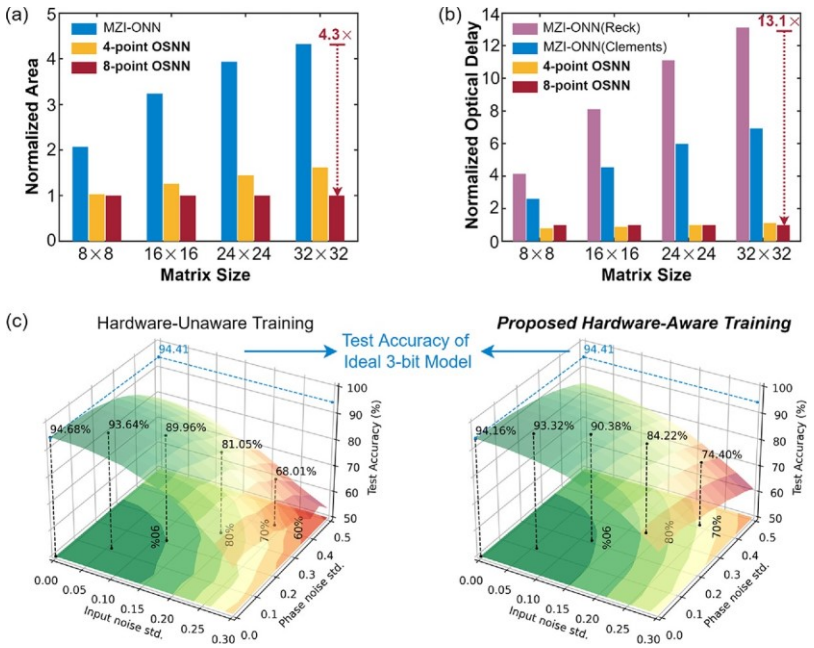


Figure 5. Performance analysis of the OSNN. (a) Normalized area comparison and (b) normalized optical delay comparison between OSNN and MZI-based ONN[11] when implementing weight matrices of different sizes. The comparison is evaluated using PDK components from AMF. (c) Robustness comparison between the hardware-unaware training and our proposed hardware-aware training, which includes dynamic noise injection techniques. With noise-awareness, our proposed hardware-aware training flow can effectively boost the noise tolerance of OSNN against various nonideal factors, such as input noises and weight-encoding noises. As a reference, we can achieve 94.41% test accuracy with 3 bit weight programming precision using the ideal transfer matrix model of the BPNC but will suffer complete malfunction (~10%) when mapped onto the real chip due to the huge discrepancy between the simulation model and the manufactured chip.

coupled in different input grating couplers separately. There are three reasons to use multiwavelength inputs in our BPNC: First, we can use compact resonator-based modulators as input modulators and avoid additional hardware costs for phase control, which requires high-speed phase shifters. Second, phase detection at the outputs can be avoided using multiwavelength inputs.[49] Third, multiwavelength inputs eliminate the phase fluctuations of optical signals in off-chip fibers and improve the robustness of OSNN to input phase noises, which have been reported in other work.[52] More details about the weight matrix of the BPNC when we use multiwavelength inputs are provided in the Supporting Information. Using multiwavelength inputs, our BPNC can express arbitrary all-positive or all-negative 4 × 4 matrices with a surprisingly high fidelity of 92.2% (See Figure S3 and the Supporting Information). The input modulators and phase shifters of the BPNC are programmed by a high-precision multichannel digital-to-analog converter (DAC). Off-chip photodetector arrays will collect the output signals, which will subsequently be read using oscilloscopes or analog-to-digital converters (ADCs). A microcontroller is used to write electrical signals to the DAC and read the output signals in this work. The measurement data are processed by computers to train and implement the DNN model. It should be noted that current fabrication and packaging technologies enable the integration of electrical circuits, photodetectors, and the laser on a single chip[53] with potentially much higher compactness, shorter interconnect paths, and higher efficiency. The experimental setup is described in detail in the Supporting Information.

Here, we experimentally implement the multistage hardware-aware training flow on our BPNC. In the calibration stage, the performance of modulators and phase shifters in the BPNC are first calibrated individually, such that we can precisely control the state of active devices, especially the input modulators and the $\mathbf{\Sigma}$ matrix (the calibration results are detailed in the Supporting Information). The second stage is to learn desired device configurations via ONN training. We first program the BPNC with representative input signals and phase shifter control voltages and collect the corresponding outputs. The above measured input−output pairs are used to train our differentiable PIC estimator for accurate and efficient chip response modeling. Then, we embed our DPE into our ONN training procedure to effectively enable hardware-aware training. Quantization-aware training and dynamic noise injection techniques are used during training to adapt the ONN model to limited phase-shifter control resolutions and boost the PIC robustness to dynamic system noises.

In this work, we construct a CNN with our BPNC and benchmark its performance on a handwritten digit classification dataset MNIST.[39] We use MVM operations to implement CNNs with a widely-applied tensor unrolling method (im2col),[54] as detailed in the Supporting Information. Figure 4a illustrates the network structure. Here, large-size tensor operations are partitioned into 4 × 4 blocks and mapped onto our BPNC. When the voltage control resolution is set to 3 bits (eight attenuation levels for each MZI attenuator in the $\mathbf{\Sigma}$ unit), the inference accuracy of the CNN reaches 94.16% in our experimental demonstration, comparable to the simulated value of 94.59%. The confusion matrix depicting the prediction results is shown in Figure 4b. Figure 4c visualizes the tested output images after being convolved by learned kernels. Figure 4d shows the tested probability distribution of different hand-written digits. More testing results are included in the Supporting Information, where we evaluate the accuracy of OSNN with different control voltage

https://doi.org/10.1021/acsphotonics.2c01188
ACS Photonics 2022, 9, 3906−3916

ranges and control resolutions. They will also be discussed in the following section.

5. DISCUSSION

5.1. Footprint. Our OSNN outperforms SVD-based MZI ONN architectures[11] in the number of trainable devices and the footprint. Rather than deploying area-costly MZI arrays, we use basic optical components such as directional couplers, phase shifters, and crossings to construct the unitary matrix units B and P. The second reason that leads to our superior compactness is that many Σ units share the B and P units, which reduces the chip area for implementing unitary transforms. Shown in Figure 5a, when the matrix size is 32 × 32, our 8-point OSNN consumes ~3.6× fewer phase shifters and ~ 4.8× fewer directional couplers, leading to ~3.3× footprint reduction compared to SVD-based MZI ONN architectures[11] with the same matrix size and optical component selection. The footprints of different ONN architectures are estimated by summing the areas of their constituent optical components provided by the same foundry (AMF). See the detailed evaluation of the chip area in the Supporting Information.

The chip area or hardware cost of OSNN can be further optimized with structured circuit pruning strategies. In an n-input, m-output layer, the $\frac{mn}{k^2}$ diagonal matrix units can be treated as $\frac{mn}{k^2}$ parameter groups. When all of the transmission coefficients in one Σ unit are zeros, this unit or parameter group is unnecessary and can be omitted in OSNN designs. When training the DNN, penalty terms encouraging higher sparsity can be added to the training objective, allowing for the elimination of unneeded Σ units while minimizing task performance degradation. Our simulation results indicate that more than 70% of neural connections in our OSNN can be pruned with negligible (<0.2%) accuracy loss when implementing image recognition tasks such as MNIST[39] or Fashion-MNIST.[55] (Results are provided in the Supporting Information). On these datasets, our pruned OSNN can save around 70% of the trainable optical components, resulting in ~52% chip area reduction compared to unpruned OSNN.

5.2. Computational Speed and Energy Efficiency. Our OSNN utilizes light to implement MVM operations, which outperforms electronic counterparts in both speed and energy efficiency. Taking into account the delay contributed by high-speed modulators (10 ps),[40,56] photodetectors (10 ps),[57] ADCs (100 ps),[58] and the optical path (43.8 ps), the total delay required to implement a 32×32 MVM can reach ~164 ps, which corresponds to an operating frequency of around 6 GHz. Using the same component library,[59] the propagation delay of the optical path in our OSNN is 5.5× less than that of an MZI-based ONN,[11] as depicted in Figure 5b. The computational speed of OSNN is now constrained by optical-to-electrical (OE) or electrical-to-optical (EO) conversion, but it can be increased further by using all-optical devices as non-linear activation functions[60] (see the Supporting Information).

The total power consumption of OSNN for MVM operations is composed of the power to drive the laser/modulators/photodetectors, the power to set the weight matrix, and the power to drive the ADCs. Numerous energy-efficient active optical components have been developed in recent years. For instance, the silicon microdisk modulator achieves approximately 1 fJ per bit.[40] Maintaining the weight matrix takes less than 2.5 mW per phase shifter in our AMF-manufactured neural chip, which can be decreased to zero by setting weights with phase change materials or nano-opto-electromechanical devices.[61,62] Concerning the power consumption of ADCs, despite the availability of high-speed ADCs, the power consumption of ADCs is significantly higher than that of other components. For example, an 8 bit, 40 GSPS ADC consumes 200 mW per channel, while an 8 bit, 10 GSPS ADC consumes 39 mW per channel.[58] In addition, the number of trainable devices in our k-point OSNN is only $O\left(\frac{mn}{k}\right)$, which saves energy for storing and reconfiguring weights. In comparison to ONN architectures designed for general MVM, where the number of programmable devices is around $O(mn)$[16] or $O(\max(m^2, n^2))$,[11] the memory cost of storing and accessing the weight matrix and the energy required to reconfigure corresponding active devices are also reduced by k times. This feature of OSNN will bring considerable energy efficiency improvement when weights need to be reconfigured frequently in large-scale DNNs, where weight loading takes nontrivial hardware cost even with weight-stationary dataflow.[63]

5.3. Resolution Analysis. Our OSNN is capable of achieving a high accuracy under low bit control of optical components. Prior ONN architectures designed for general MVMs require high-precision control of optical devices for parameter mapping to maintain accuracy.[11] Otherwise, we may encounter severe task performance degradation because of large mapping errors,[64] which will quickly accumulate as the size of weight matrices or the number of layers increases. Given that the control precision of some energy-efficient photonic tensor cores is only 4 or 5 bits,[30] it is necessary to reduce the resolution requirement of ONN architectures and enhance the tolerance of quantization errors. In this study, quantization-aware training is applied to our OSNN to adopt the limited voltage control precision and mitigate the accuracy loss. In experiments, we have shown that ~94% accuracy can be achieved for digit recognition when the precision of the DACs for controlling the phase shifters is around 3 bits (see the Supporting Information). What is more, low-resolution device control can also lessen the energy cost for weight storage, access, and reconfiguration.[65]

5.4. Robustness. The robustness of our OSNN is guaranteed by our hardware-aware training framework. Our AI-assisted DPE provides accurate variation modeling of static noises, e.g., process variations, device calibration errors, thermal cross-talk, and non-ideal extinction ratio of modulators. Moreover, our noise-injection training algorithm further considers the impacts of dynamic noises, e.g., thermal noises from the laser source and photodetection noises. The robustness of our architecture is evaluated by varying the signal-to-noise-ratio (SNR) of the inputs and the phase drifts of phase shifters in MZI attenuators, and our analysis results are shown in Figure 5c. Thanks to our noise injection techniques, our OSNN maintains greater than 90% average inference accuracy even when the standard deviation of input noise and phase drifts reach 0.1 and 0.2, respectively.

Additionally, the robustness of our OSNN can be enhanced with more reliable optical components and more reliable control circuits.[66] For bandwidth-driven and robustness-driven OSNN design, one can directly select broadband MZI with low temperature sensitivity as a robust variable optical attenuator (VOA). Less robust but more compact or energy-

https://doi.org/10.1021/acsphotonics.2c01188
ACS Photonics 2022, 9, 3906−3916

efficient components can also be employed, e.g., ultracompact MRR modulators with on-chip feedback controls and PCM-based modulators with advanced high-endurance materials.

5.5. Scaling and Outlook. The performance metrics of the OSNN can be further improved in several directions. First, our OSNN is compatible with the majority of the device-level enhancement techniques. For example, by using smaller directional couplers,[67] crossings,[68] and VOAs,[69] the chip area of the OSNN can be optimized, resulting in a competitive computing density of >200 TOPS/mm^2 and energy efficiency of ∼9.5 TOPS/W (see the Supporting Information). Second, massive multiplexing techniques can substantially boost the throughput of our architecture. Because all of the optical components in our architecture can be broadband devices, wavelength-division multiplexing (WDM) techniques can be applied to our architecture: If k-wavelength input signals propagate through the chip simultaneously to implement the MVM in parallel, the throughput and the computing density can then be improved by $(k - 1)$ times over a single-wavelength OSNN. Furthermore, more circuit structures and optical components can be investigated to construct our OSNN. Notably, the BPNC is not the only option to implement $B\Sigma P$. For example, recent work demonstrates that multiport n-to-n directional couplers, multimode interference (MMI) couplers, and diffractive cells can be utilized to build unitary matrices,[70,71] which can achieve smaller footprint but less matriX representativity compared to our proposed BPNC. They can also be used to build the B and P units to reduce the chip area. Finally, faster or more energy efficient EO/OE conversion techniques are demanded to improve the computational speed and energy efficiency for data movement between electrons and photons, which currently restricts the performance of optical computing platforms.

6. CONCLUSIONS

We present a hardware-efficient optical subspace neural network (OSNN) architecture with experimental demonstrations on a silicon photonic programmable butterfly-style photonic−electronic neural chip (BPNC). By exploring optical neurocomputing beyond conventional GEMMs with restricted weight representability, our OSNN consumes up to 7× fewer trainable optical components than prior MZI-based ONN architectures designed for GEMMs. This advantage can be further increased to ∼23× using structured circuit pruning strategies with negligible accuracy loss. Our proposed hardware-aware training framework efficiently models the behavior of the OSNN to help reduce control precision requirements, enhance noise robustness, and fully exploit the expressivity in the subspace. The performance of OSNN can be further improved with smaller optical components as well as faster and more efficient EO/OE conversion techniques. Our OSNN pushes the limits of scalability and the robustness of ONNs and creates a new design paradigm for next-generation high-performance AI accelerators with improved hardware efficiency.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acsphotonics.2c01188.

Our supporting information includes the following contents: EXpressivity evaluation and phase configuration discussion of B/P units; Methods to scale B/P units; Details of our proposed hardware-aware training framework; Weight matriX of BPNC with multi-wavelength inputs; Testing setup; On-chip calibration results; The approach to implementing 2D convolution with MVM operations; EXperimental results of resolution analysis; Evaluation of footprint, computational speed, propagation loss, power consumption, computing density, and scaling approaches of BPNC (PDF)

AUTHOR INFORMATION

Corresponding Authors

David Z. Pan − *Department of Electrical and Computer Engineering, The University of Texas at Austin, Austin, Texas 78705, United States*; Email: dpan@ece.utexas.edu

Ray T. Chen − *Microelectronics Research Center, The University of Texas at Austin, Austin, Texas 78758, UnitedStates; Department of Electrical and Computer Engineering,The University of Texas at Austin, Austin, Texas 78705, United States; Omega Optics, Inc., Austin, Texas 78757, United States*; Email: chenrt@austin.utexas.edu

Authors

Chenghao Feng − *Microelectronics Research Center, The University of Texas at Austin, Austin, Texas 78758, United States; Department of Electrical and Computer Engineering, The University of Texas at Austin, Austin, Texas 78705, United States;* orcid.org/0000-0002-0751-7681

Jiaqi Gu − *Department of Electrical and Computer Engineering, The University of Texas at Austin, Austin, Texas78705, United States*

Hanqing Zhu − *Department of Electrical and Computer Engineering, The University of Texas at Austin, Austin, Texas 78705, United States*

Zhoufeng Ying − *Microelectronics Research Center, The University of Texas at Austin, Austin, Texas 78758, United States; Alpine Optoelectronics, Fremont, California 94538, United States*

Zheng Zhao − *Department of Electrical and Computer Engineering, The University of Texas at Austin, Austin, Texas 78705, United States; Synopsys Inc., Mountain View, California 94043, United States*

Complete contact information is available at: https://pubs.acs.org/10.1021/acsphotonics.2c01188

Author Contributions

#C.F. and J.G. contributed equally to this work.

Funding

Multidisciplinary University Research Initiative (MURI) program (FA 9550-17-1-0071) and NSF Award #1932753.

Notes

The authors declare no competing financial interest.

The data and codes that support the findings of this study are available from the corresponding author upon reasonable request. A reference implementation of OSNN and hardware-aware training approach is available at our open-source photonic AI library TorchONN (https://github.com/JeremieMelo/pytorch-onn).

ACKNOWLEDGMENTS

The authors acknowledge support from the Multidisciplinary University Research Initiative (MURI) program through the

Air Force Office of Scientific Research (AFOSR), monitored by Dr. Gernot S. Pomrenke. The photonic-electronic neural chip reported herein is also targeted at the AI and machine learning applications for the NSF Fourier Transform Spectroscopy program (Award #1932753).

REFERENCES

(1) Lecun, Y.; Bengio, Y.; Hinton, G. Deep Learning. *Nature* 2015, *521*, 436−444.

(2) Krizhevsky, A.; Sutskever, I.; Hinton, G. E. ImageNet Classification with Deep Convolutional Neural Networks. *Adv. Neural Inf. Process. Syst.* 2012, *25*, 1097−1105.

(3) Silver, D.; Huang, A.; Maddison, C. J.; Guez, A.; Sifre, L.; van den Driessche, G.; Schrittwieser, J.; Antonoglou, I.; Panneershelvam, V.; Lanctot, M.; Dieleman, S.; Grewe, D.; Nham, J.; Kalchbrenner, N.; Sutskever, I.; Lillicrap, T.; Leach, M.; Kavukcuoglu, K.; Graepel, T.; Hassabis, D. Mastering the Game of Go with Deep Neural Networks and Tree Search. *Nature* 2016, *529*, 484−489.

(4) Morioka, T.; Iwata, T.; Hori, T.; Kobayashi, T.Multiscale Recurrent Neural Network Based Language Model. In *Interspeech*; 2010; Vol. *2*, pp. 1045−1048. DOI: 10.21437/interspeech.2015-512.

(5) Nurvitadhi, E.; Venkatesh, G.; Sim, J.; Marr, D.; Huang, R.; Gee, J.; Ong, H.; Liew, Y. T.; Srivatsan, K.; Moss, D.; Subhaschandra, S.; Boudoukh, G.Can FPGAs Beat GPUs in Accelerating Next-Generation Deep Neural Networks?*Proceedings of the 2017 ACM/SIGDA International Symposium on Field-Programmable Gate Arrays*; FPGA2017. DOI: 10.1145/3020078.

(6) Chen, Y.-H.; Krishna, T.; Emer, J. S.; Sze, V. Eyeriss: An Energy-Efficient Reconfigurable Accelerator for Deep Convolutional Neural Networks. *IEEE J. Solid-State Circuits* 2017, *52*, 127−138.

(7) Dennard, R. H.; Gaensslen, F. H.; Yu, H.-N.; Rideout, V. L.; Bassous, E.; Leblanc, A. R. Design Of Ion-Implanted MOSFET's with Very Small Physical Dimensions. *P. IEEE* 1999, *87*, 668−678.

(8) Waldrop, M. M. More Than Moore. *Nature* 2016, *530*, 144−147.

(9) Li, C.; Hu, M.; Li, Y.; Jiang, H.; Ge, N.; Montgomery, E.; Zhang, J.; Song, W.; Dávila, N.; Graves, C. E.; Li, Z.; Strachan, J. P.; Lin, P.; Wang, Z.; Barnell, M.; Wu, Q.; Williams, R. S.; Yang, J. J.; Xia, Q. Analogue Signal and Image Processing with Large Memristor Crossbars. *Nature Electronics* 2018, *1*, 52−59.

(10) Yao, P.; Wu, H.; Gao, B.; Tang, J.; Zhang, Q.; Zhang, W.; Yang, J. J.; Qian, H. Fully Hardware-Implemented Memristor Convolutional Neural Network. *Nature* 2020, *577*, 641.

(11) Shen, Y.; Harris, N. C.; Skirlo, S.; Prabhu, M.; Baehr-Jones, T.; Hochberg, M.; Sun, X.; Zhao, S.; Larochelle, H.; Englund, D.; Soljacic, M. Deep Learning with Coherent Nanophotonic Circuits. *Nat. Photonics* 2017, *11*, 441−446.

(12) Lin, X.; Rivenson, Y.; Yardimci, N. T.; Veli, M.; Luo, Y.; Jarrahi, M.; Ozcan, A. All-Optical Machine Learning Using Diffractive Deep Neural Networks. *Science* 2018, *361*, 1004−1008.

(13) Shastri, B. J.; Tait, A. N.; de Lima, T. F.; Pernice, W. H. P.; Bhaskaran, H.; Wright, C. D.; Prucnal, P. R. Photonics for Artificial Intelligence and Neuromorphic Computing. *Nat. Photonics* 2021, *15*, 102−114.

(14) Ying, Z.; Feng, C.; Zhao, Z.; Dhar, S.; Dalir, H.; Gu, J.; Cheng, Y.; Soref, R.; Pan, D. Z.; Chen, R. T. Electronic-Photonic Arithmetic Logic Unit for High-Speed Computing. *Nat. Commun.* 2020, *11*, 2154.

(15) Feng, C.; Ying, Z.; Zhao, Z.; Gu, J.; Pan, D. Z.; Chen, R. T. Toward High-Speed and Energy-Efficient Computing: A WDM-Based Scalable On-Chip Silicon Integrated Optical Comparator. *Laser Photonics Rev.* 2021, *15*, No. 2000275.

(16) Tait, A. N.; De Lima, T. F.; Zhou, E.; Wu, A. X.; Nahmias, M. A.; Shastri, B. J.; Prucnal, P. R. Neuromorphic Photonic Networks Using Silicon Photonic Weight Banks. *Sci. Rep.* 2017, *7*, 7430.

(17) Huang, C.; Fujisawa, S.; de Lima, T. F.; Tait, A. N.; Blow, E. C.; Tian, Y.; Bilodeau, S.; Jha, A.; Yaman, F.; Peng, H. T.; Batshon, H. G.; Shastri, B. J.; Inada, Y.; Wang, T.; Prucnal, P. R. A Silicon Photonic−Electronic Neural Network for Fibre Nonlinearity Compensation. *Nat. Electron.* 2021, *4*, 837−844.

(18) Xu, X.; Tan, M.; Corcoran, B.; Wu, J.; Boes, A.; Nguyen, T. G.; Chu, S. T.; Little, B. E.; Hicks, D. G.; Morandotti, R.; Mitchell, A.; Moss, D. J. 11 TOPS Photonic Convolutional Accelerator for Optical Neural Networks. *Nature* 2021, *589*, 44−51.

(19) Feldmann, J.; Youngblood, N.; Karpov, M.; Gehring, H.; Li, X.; Stappers, M.; Le Gallo, M.; Fu, X.; Lukashchuk, A.; Raja, A. S.; Liu, J.; Wright, C. D.; Sebastian, A.; Kippenberg, T. J.; Pernice, W. H. P.; Bhaskaran, H. Parallel Convolutional Processing Using an Integrated Photonic Tensor Core. *Nature* 2021, *589*, 52−58.

(20) Denton, E. L.; Zaremba, W.; Bruna, J.; LeCun, Y.; Fergus, R. Exploiting Linear Structure within Convolutional Networks for Efficient Evaluation. *Adv. Neural Inf. Process. Syst.* 2014, 1269−1277.

(21) Lebedev, V.; Ganin, Y.; Rakhuba, M.; Oseledets, I.; Lempitsky, V.Speeding-up Convolutional Neural Networks Using Fine-Tuned CP-Decomposition. In *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings; International Conference on Learning Representations*; ICLR, 2015.

(22) Tai, C.; Xiao, T.; Zhang, Y.; Wang, X.; Weinan, E.Convolutional Neural Networks with Low-Rank Regularization. *4th International Conference on Learning Representations, ICLR 2016 - Conference Track Proceedings*; ICLR2015.

(23) Ding, C.; Liao, S.; Wang, Y.; Li, Z.; Liu, N.; Zhuo, Y.; Wang, C.; Qian, X.; Bai, Y.; Yuan, G.; Ma, X.; Zhang, Y.; Tang, J.; Qiu, Q.; Lin, X.; Yuan, B.CIRCNN: Accelerating and Compressing Deep Neural Networks Using Block-Circulant Weight Matrices. In *Proceedings of the Annual International Symposium on Microarchitecture*; MICRO; ACM: New York, NY, USA, 2017; Vol. Part *F1312*, pp. 395−408. DOI: 10.1145/3123939.3124552.

(24) Li, Z.; Wang, S.; Ding, C.; Qiu, Q.; Wang, Y.; Liang, Y.Efficient Recurrent Neural Networks Using Structured Matrices in FPGAS. *6th International Conference on Learning Representations, ICLR 2018 - Workshop Track Proceedings*; ICLR2018, 1−4.

(25) Gu, J.; Zhao, Z.; Feng, C.; Liu, M.; Chen, R. T.; Pan, D. Z.Towards Area-Efficient Optical Neural Networks: An FFT-Based Architecture. In *2020 25th Asia and South Pacific Design Automation Conference (ASP-DAC)*; IEEE, 2020; Vol. 2020-Janua, pp. 476−481. DOI: 10.1109/ASP-DAC47756.2020.9045156.

(26) Gu, J.; Zhao, Z.; Feng, C.; Ying, Z.; Liu, M.; Chen, R. T.; Pan, D. Z. Toward Hardware-Efficient Optical Neural Networks: Beyond FFT Architecture via Joint Learnability. *IEEE Trans. Comput.-Aided Des. Integr. Circuits Syst.* 2021, *40*, 1796−1809.

(27) Miscuglio, M.; Hu, Z.; Li, S.; George, J. K.; Capanna, R.; Dalir, H.; Bardet, P. M.; Gupta, P.; Sorger, V. J. Massively Parallel Amplitude-Only Fourier Neural Network. *Optica* 2020, *7*, 1812.

(28) Nikdast, M.; Nicolescu, G.; Trajkovic, J.; Liboiron-Ladouceur, O. Modeling Fabrication Non-Uniformity in Chip-Scale Silicon Photonic Interconnects. In *2016 Design, Automation Test in Europe Conference Exhibition (DATE)*; IEEE2016; pp. 115−120.

(29) Zhu, Y.; Zhang, G. L.; Li, B.; Yin, X.; Zhuo, C.; Gu, H.; Ho, T.-Y.; Schlichtmann, U.Countering Variations and Thermal Effects for Accurate Optical Neural Networks. In *2020 IEEE/ACM International Conference On Computer Aided Design (ICCAD)*; IEEE2020; pp. 1−7.

(30) Miscuglio, M.; Sorger, V. J. Photonic Tensor Cores for Machine Learning. *Appl. Phys. Rev.* 2020, *7*, No. 031404.

(31) Cheng, Q.; Kwon, J.; Glick, M.; Bahadori, M.; Carloni, L. P.; Bergman, K. Silicon Photonics Codesign for Deep Learning. *Proc. IEEE* 2020, *108*, 1261−1282.

(32) Banerjee, S.; Nikdast, M.; Chakrabarty, K.Modeling Silicon-Photonic Neural Networks under Uncertainties. *Proceedings -Design, Automation and Test in Europe, DATE*; IEEE2021, 2021-Febru, 98−101. DOI: 10.23919/DATE51398.2021.9474000.

(33) Hughes, T. W.; Minkov, M.; Shi, Y.; Fan, S. Training of Photonic Neural Networks through in Situ Backpropagation and Gradient Measurement. *Optica* 2018, *5*, 864.

(34) Gu, J.; Zhao, Z.; Feng, C.; Li, W.; Chen, R. T.; Pan, D. Z.FLOPS: EFficient On-Chip Learning for Optical Neural Networks Through Stochastic Zeroth-Order Optimization. In *2020 57th ACM/*

IEEE Design Automation Conference (DAC); IEEE, 2020; Vol. 2020-July, pp. 1−6. DOI: 10.1109/DAC18072.2020.9218593.
(35) Gu, J.; Feng, C.; Li, W.; Chen, R. T.; Pan, D. Z. Efficient On-Chip Learning for Optical Neural Networks Through Power-Aware Sparse Zeroth-Order Optimization. In *Proceedings of the AAAI Conference on Artificial Intelligence*; 2020; Vol. *35*, pp. 7583−7591. DOI: 10.1109/DAC18072.2020.9218593.
(36) Zhang, T.; Wang, J.; Dan, Y.; Lanqiu, Y.; Dai, J.; Han, X.; Sun, X.; Xu, K. Efficient Training and Design of Photonic Neural Network through Neuroevolution. *Opt. Express* 2019, *27*, 37150.
(37) Pai, S.; Williamson, I. A. D.; Hughes, T. W.; Minkov, M.; Solgaard, O.; Fan, S.; Miller, D. A. B. Parallel Programming of an Arbitrary Feedforward Photonic Network. *IEEE J. Sel. Top. Quantum Electron.* 2020, *26*, 1.
(38) Wang, Y.; Wen, W.; Liu, B.; Chiarulli, D.; Li, H. H. Group Scissor: Scaling Neuromorphic Computing Design to Large Neural Networks. *Proceedings - Design Automation Conference*; IEEE2017, *Part 128280*. DOI: 10.1145/3061639.3062256.
(39) LeCun, Y.*The MNIST Database of Handwritten Digits*. http://yann.lecun.com/exdb/mnist/1998.
(40) Timurdogan, E.; Sorace-Agaskar, C. M.; Sun, J.; Shah Hosseini, E.; Biberman, A.; Watts, M. R. An Ultralow Power Athermal Silicon Modulator. *Nat. Commun.* 2014, *5*, 4008.
(41) Jing, L.; Shen, Y.; Dubcek, T.; Peurifoy, J.; Skirlo, S.; Lecun, Y.; Tegmark, M.; Soljacic, M. S. Tunable Efficient Unitary Neural Networks (EUNN) and Their Application to RNNs. In *Proceedings of the 34th International Conference on Machine Learning*; PMLR2017. DOI: 10.5555/3305381.
(42) Flamini, F.; Spagnolo, N.; Viggianiello, N.; Crespi, A.; Osellame, R.; Sciarrino, F. Benchmarking Integrated Linear-Optical Architectures for Quantum Information Processing. *Sci. Rep.* 2017, *7*, 1−10.
(43) Tian, Y.; Zhao, Y.; Liu, S.; Li, Q.; Wang, W.; Feng, J.; Guo, J. Scalable and Compact Photonic Neural Chip with Low Learning-Capability-Loss. *NANO* 2022, *11*, 329−344.
(44) Zhao, R.; Hu, Y.; Dotzel, J.; De Sa, C.; Zhang, Z.Building Efficient Deep Neural Networks With Unitary Group Convolutions. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE, 2019, pp. 11295−11304. DOI: 10.1109/CVPR.2019.01156.
(45) Tait, A. N.; Ferreira De Lima, T.; Nahmias, M. A.; Miller, H. B.; Peng, H.-T.; Shastri, B. J.; Prucnal, P. R. Silicon Photonic Modulator Neuron. *Phys. Rev. Appl.* 2019, *11*, No. 064043.
(46) Williamson, I. A. D.; Hughes, T. W.; Minkov, M.; Bartlett, B.; Pai, S.; Fan, S. Reprogrammable Electro-Optic Nonlinear Activation Functions for Optical Neural Networks. *IEEE J. Sel. Top. Quantum Electron.* 2020, *26*, 1.
(47) Zhou, H.; Zhao, Y.; Xu, G.; Wang, X.; Tan, Z.; Dong, J.; Zhang, X. Chip-Scale Optical MatriX Computation for PageRank Algorithm. *IEEE J. Sel. Top. Quantum Electron.* 2020, *26*, 1−10.
(48) Wright, L. G.; Onodera, T.; Stein, M. M.; Wang, T.; Schachter, D. T.; Hu, Z.; McMahon, P. L. Deep Physical Neural Networks Trained with Backpropagation. *Nature* 2022, *601*, 549−555.
(49) Zhou, T.; Lin, X.; Wu, J.; Chen, Y.; Xie, H.; Li, Y.; Fan, J.; Wu, H.; Fang, L.; Dai, Q. Large-Scale Neuromorphic Optoelectronic Computing with a Reconfigurable Diffractive Processing Unit. *Nat. Photonics* 2021, *15*, 367−373.
(50) Spall, J.; Guo, X.; Guo, X.; Guo, X.; Lvovsky, A. I.; Lvovsky, A. I.; Lvovsky, A. I. Hybrid Training of Optical Neural Networks. *Optica* 2022, *9*, 803−811.
(51) Miller, D. A. B. Analyzing and Generating Multimode Optical Fields Using Self-Configuring Networks. *Optica* 2020, *7*, 794.
(52) Zhang, H.; Thompson, J.; Gu, M.; Jiang, X. D.; Cai, H.; Liu, P. Y.; Shi, Y.; Zhang, Y.; Karim, M. F.; Lo, G. Q.; Luo, X.; Dong, B.; Kwek, L. C.; Liu, A. Q. Efficient On-Chip Training of Optical Neural Networks Using Genetic Algorithm. *ACS Photonics* 2021, 1662.
(53) Atabaki, A. H.; Moazeni, S.; Pavanello, F.; Gevorgyan, H.; Notaros, J.; Alloatti, L.; Wade, M. T.; Sun, C.; Kruger, S. A.; Meng, H.; Al Qubaisi, K.; Wang, I.; Zhang, B.; Khilo, A.; Baiocco, C. V.; Popovic, ́M. A.; Stojanovic, ́V. M.; Ram, R. J. Integrating Photonics with Silicon Nanoelectronics for the next Generation of Systems on a Chip. *Nature* 2018, *556*, 349−354.
(54) Chetlur, S.; Woolley, C.; Vandermersch, P.; Cohen, J.; Tran, J.; Catanzaro, B.; ShelhamerCudnn: Efficient Primitives for Deep Learning. *arXiv preprint arXiv:1410.0759*2014.
(55) Xiao, H.; Rasul, K.; Vollgraf, R.Fashion-MNIST: A Novel Image Dataset for Benchmarking Machine Learning Algorithms. *Arxiv*2017.
(56) Timurdogan, E.; Su, Z.; Shiue, R.-J.; Poulton, C. V.; Byrd, M. J.; Xin, S.; Watts, M. R.APSUNY Process Design Kit (PDKv3.0): O, C and L Band Silicon Photonics Component Libraries on 300mm Wafers. *2019 Optical Fiber Communications Conference and Exhibition (OFC)*; Optical Society of America2019, 1−3. DOI: 10.1364/ofc.2019.tu2a.1.
(57) Vivien, L.; Polzer, A.; Marris-Morini, D.; Osmond, J.; Hartmann, J. M.; Crozat, P.; Cassan, E.; Kopp, C.; Zimmermann, H.; Fédéli, J. M. Zero-Bias 40Gbit/s Germanium Waveguide Photodetector on Silicon. *Opt. Express* 2012, *20*, 1096.
(58) *ADC (Analog-to-Digital converters) − Alphacore*. https://www.alphacoreinc.com/adc-analog-to-digital-converters/ (accessed 2021-08-25).
(59) Siew, S. Y.; Li, B.; Gao, F.; Zheng, H. Y.; Zhang, W.; Guo, P.; Xie, S. W.; Song, A.; Dong, B.; Luo, L. W.; Li, C.; Luo, X.; Lo, G. Q. Review of Silicon Photonics Technology and Platform Development. *J. Lightwave Technol.* 2021, *39*, 4374−4389.
(60) Bao, Q.; Zhang, H.; Ni, Z.; Wang, Y.; Polavarapu, L.; Shen, Z.; Xu, Q. H.; Tang, D.; Loh, K. P. Monolayer Graphene as a Saturable Absorber in a Mode-Locked Laser. *Nano Res.* 2011, *4*, 297−307.
(61) Wuttig, M.; Bhaskaran, H.; Taubner, T. Phase-Change Materials for Non-Volatile Photonic Applications. *Nat. Photonics* 2017, *11*, 465−476.
(62) Midolo, L.; Schliesser, A.; Fiore, A. Nano-Opto-Electro-Mechanical Systems. *Nat. Nanotechnol.* 2018, *13*, 11−18.
(63) Chen, Y. H.; Emer, J.; Sze, V.Eyeriss: A Spatial Architecture for Energy-Efficient Dataflow for Convolutional Neural Networks. *Proceedings - 2016 43rd International Symposium on Computer Architecture, ISCA 2016*2016, 367−379. DOI: 10.1109/ISCA.2016.40.
(64) Gu, J.; Zhao, Z.; Feng, C.; Zhu, H.; Chen, R. T.; Pan, D. Z.ROQ: A Noise-Aware Quantization Scheme Towards Robust Optical Neural Networks with Low-Bit Controls. *Proceedings of the 2020 Design, Automation and Test in Europe Conference and Exhibition, DATE 2020*; IEEE2020, No. 1, 1586−1589. DOI: 10.23919/date48585.2020.9116521.
(65) Han, S.; Liu, X.; Mao, H.; Pu, J.; Pedram, A.; Horowitz, M. A.; Dally, W. J.EIE: Efficient Inference Engine on Compressed Deep Neural Network. *2016 ACM/IEEE 43rd Annual International Symposium on Computer Architecture (ISCA)*; IEEE2016, 243−254. DOI: 10.1109/ISCA.2016.30.
(66) Wade, M.; Anderson, E.; Ardalan, S.; Bhargava, P.; Buchbinder, S.; Davenport, M. L.; Fini, J.; Lu, H.; Li, C.; Meade, R.; Ramamurthy, C.; Rust, M.; Sedgwick, F.; Stojanovic, V.; Van Orden, D.; Zhang, C.; Sun, C.; Shumarayev, S. Y.; OKeeffe, C.; Hoang, T. T.; Kehlet, D.; Mahajan, R. V.; Guzy, M. T.; Chan, A.; Tran, T. TeraPHY: A Chiplet Technology for Low-Power, High-Bandwidth In-Package Optical I/O. *IEEE Micro* 2020, *40*, 63−71.
(67) Ye, C.; Dai, D. Ultra-Compact Broadband 2 × 2 3 DB Power Splitter Using a Subwavelength-Grating-Assisted Asymmetric Directional Coupler. *J. Lightwave Technol.* 2020, *38*, 2370−2375.
(68) Han, H.-L.; Li, H.; Zhang, X.-P.; Liu, A.; Lin, T.-Y.; Chen, Z.; Lv, H.-B.; Lu, M.-H.; Liu, X.-P.; Chen, Y.-F. High Performance Ultra-Compact SOI Waveguide Crossing. *Opt. Express* 2018, *26*, 25602.
(69) Haffner, C.; Joerg, A.; Doderer, M.; Mayor, F.; Chelladurai, D.; Fedoryshyn, Y.; Roman, C. I.; Mazur, M.; Burla, M.; Lezec, H. J.; Aksyuk, V. A.; Leuthold, J. Nano−Opto-Electro-Mechanical Switches Operated at CMOS-Level Voltages. *Science* 2019, *366*, 860−864.

(70) Tanomura, R.; Tang, R.; Ghosh, S.; Tanemura, T.; Nakano, Y. Robust Integrated Optical Unitary Converter Using Multiport Directional Couplers. *J. Lightwave Technol.* 2020, *38*, 60−66.
(71) Zhu, H. H.; Zou, J.; Zhang, H.; Shi, Y. Z.; Luo, S. B.; Wang, N.; Cai, H.; Wan, L. X.; Wang, B.; Jiang, X. D.; Thompson, J.; Luo, X. S.; Zhou, X. H.; Xiao, L. M.; Huang, W.; Patrick, L.; Gu, M.; Kwek, L. C.; Liu, A. Q. Space-Efficient Optical Computing with an Integrated Chip Diffractive Neural Network. *Nat. Commun.* 2022, *13*, 1−9.

https://doi.org/10.1021/acsphotonics.2c01188
ACS Photonics 2022, 9, 3906−3916