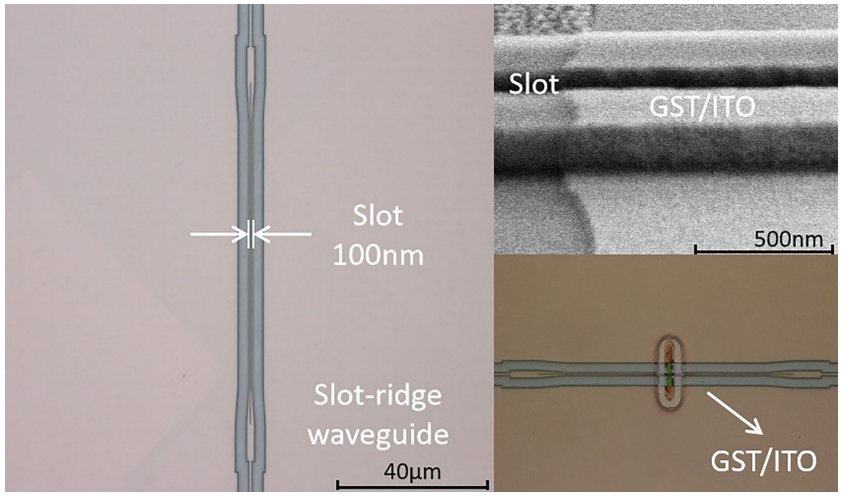
IEEE Photonics Journal
An IEEE Photonics Society Publication

Open Access

On-Chip Photonic Synapses Based on Slot-Ridge Waveguides With PCMs For In-Memory Computing

Volume 13, Number 2, April 2021

Huan Zhang
Beiju Huang
Zanyun Zhang
Chuantong Cheng
Zan Zhang
Hengjie Zhang
Run Chen
Hongda Chen



DOI: 10.1109/JPHOT.2021.3066500

On-Chip Photonic Synapses Based on Slot-Ridge Waveguides With PCMs For In-Memory Computing

Huan Zhang [1,2] Beiju Huang [1,2,3] Zanyun Zhang [4] Chuantong Cheng,[1,2] Zan Zhang,[5] Hengjie Zhang,[1,2] Run Chen,[1,2] and Hongda Chen[1,2,3]

[1]State Key Laboratory on Integrated Optoelelctronics, Institute of Semiconductors, Chinese Academy of Sciences, Beijing 100083, China
[2]College of Materials Science and Opto-Electronic Technology, University of Chinese Academy of Sciences, Beijing 100049, China
[3]Beijing Key Laboratory of Inorganic Stretchable and Flexible Information Technology, Beijing 100083, China
[4]School of Electrical and Electronic Engineering, Tiangong University, Tianjin 300387, China
[5]School of Electronic and Control Engineering, Chang'an University, Xi'an 710064, China

DOI:10.1109/JPHOT.2021.3066500

Manuscript received February 11, 2021; revised March 6, 2021; accepted March 13, 2021. Date of publication March 17, 2021; date of current version April 8, 2021. This work was supported in part by the National Key R&D Program of China under Grant 2018YFA0209000, in part by the Natural Science Foundation of China under Grants 61675191, 61974099, and 61634006, and in part by the Youth Innovation Promotion Association of Chinese Academy of Sciences under Grant 2018146. Corresponding author: Beiju Huang (e-mail: bjhuang@semi.ac.cn).

Abstract: Because of the von Neumann bottleneck, neuromorphic networks aimed at in-memory computing, such as brains, are extensively studied. As artificial synapses are essential in neuromorphic networks, a photonic synapse based on slot-ridge waveguides with nonvolatile phase-change materials (PCMs) was proposed and demonstrated in an SOI platform with standard complementary metal-oxide-semiconductor (CMOS) process for a larger weight dynamic range. The change of the optical transmission spectrum of our photonic synapses was about 3.5dB higher than that of primitive synapses, which meant large weight dynamic range and more weight values. A 90.7% recognition accuracy based on our photonic synapses, which was 2.6% higher than that of primitive synapses, was realized for the MNIST handwritten digits recognition task performed by a three-layer perceptron. Besides, because of the nonvolatile nature of PCMs, the weights achieved by our photonic synapses can be stored in situ ensuring a lower consumption in in-memory computing. This framework can potentially achieve a more efficient in-memory computing neuromorphic network in silicon photonics.

Index Terms: Photonic synapses, PCMs, in-memory computing, neuromorphic networks, silicon photonics.

1. Introduction

Usually, traditional computer systems suffer from the bottleneck of the von Neumann architecture [1], which considerably limits the computing efficiency and increases the power consumption due to the separating of the computing and memory units. To achieve efficient computing and low power consumption, the distance between the computing and memory units must be shortened. To reach

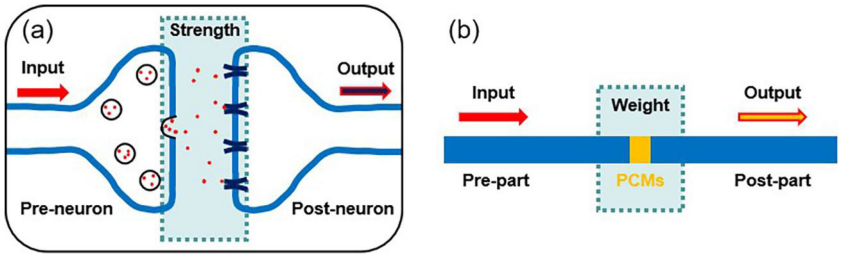


Fig. 1. (a) The schematic diagram of a biological synapse. Pre-neuron and post-neuron are connected by synapses. (b) The schematic diagram of a photonic artificial synapse based on the waveguide with PCMs.

this goal, neuromorphic computing networks aimed at in-memory computing, like human brains, are attracting increasing attention [2]. With approximately 10^{11} neurons and 10^{15} synapses [3] handling and storing information between neurons, the human brains are so efficient that they use only 20W of power to perform 10^{20} MAC/s (multiply accumulate per second). Therefore, artificial synapses are essential to realize efficient in-memory computing neuromorphic networks. Several electronic synapses, such as memory [4]–[7], memristors [8]–[10], and the field-effect transistors [11]–[13] have already been developed. However, they severely suffer from low bandwidth and interconnect bottleneck. Photonic synapses as potential alternatives are becoming popular owing to their high speed, low crosstalk, scalability, and high interconnect bandwidth [14].

Previously, there have been many demonstrations of photonic synapses for in-memory computing [14]–[20] based on ridge waveguides with nonvolatile phase-change materials (PCMs) directly put on the top of the waveguides to modulate optical transmission by transforming their phase states between amorphous and crystalline states. The working principle of this structure is based on the weak interaction between the waveguide mode and PCMs, which seriously hinders the optical modulation efficiency. In this paper, we proposed and demonstrated a photonic synapse design based on slot-ridge waveguides, which enhanced the interaction and the change of the interaction between the optical mode and PCMs. Employing this architecture, we synthesized and demonstrated photonic synapses with a larger weight dynamic range, which was possibly capable of achieving more multilevel weights. In addition, the weights achieved by our photonic synapses can be stored in situ because of the nonvolatile PCMs ensuring a lower consumption in in-memory computing. Therefore, the framework has the potential to make photonic synapses more weight values and lower power consumption, and it can realize neuromorphic networks with more efficient in-memory computing in silicon photonics.

2. Device Concept

In human brains, synapses are numerous, and they control the communication between neurons [14]. As Fig. 1(a) shows, the connection strength between the pre-neuron and post-neuron depends on the synapse with a spike signal coefficient, the connection coefficient. Further, the connection strength varies with the pre-neuron and post-neuron representing as multilevel weights. Accordingly, in Fig. 1(b) a photonic artificial synapse based on the waveguide with PCMs also attaches a weight coefficient to the input optical signal by realizing the different phase states of the PCMs. Therefore, the connection coefficient and the weight coefficient, which is achieved by the interaction between the optical mode and PCMs, are found to be analogous when we compare photonic synapses with biological synapses.

For photonic synapses, the conventional architecture of positioning PCMs directly on the top of ridge waveguides considerably suffers from the weak interaction between the evanescent optical field and the PCMs. To enhance the interaction, the slot structure was utilized and introduced to a ridge waveguide structure, where the optical field was confined and enhanced around the

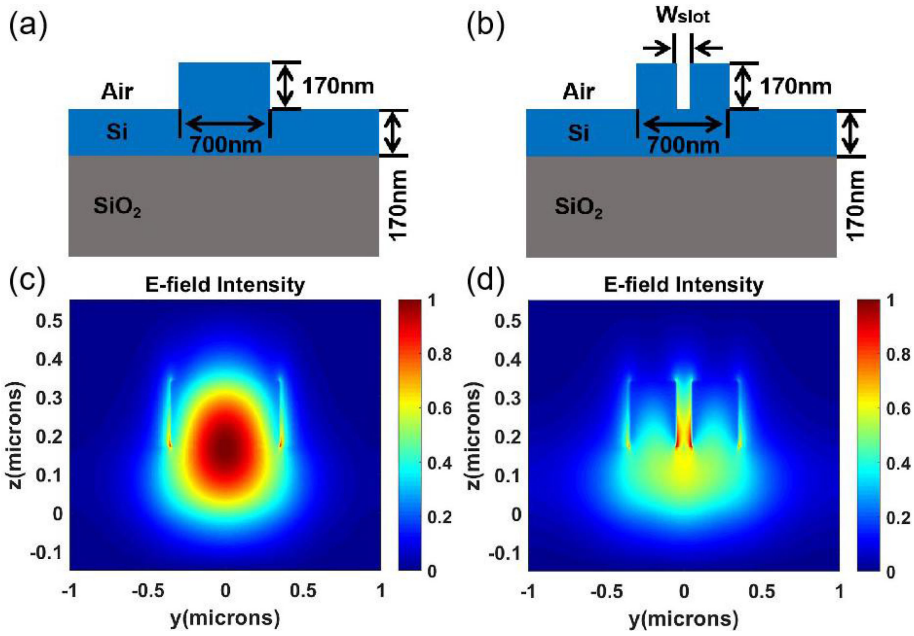


Fig. 2. The cross-sectional view of (a) ridge waveguides and (b) slot-ridge waveguides. The E-field intensity of the TE optical mode profiles of (c) ridge waveguides and (d) slot-ridge waveguides.

slot because of the continuity of the normal component of electric flux density [21]. It needs to be emphasized that the slot waveguides are not exploited to be photonic synapses as they are larger loss and more difficult to be fabricated by processes because of their large depth-to-width ratio. The cross-sectional view of ridge and slot-ridge waveguides are shown in Fig. 2(a) and (b), respectively, where the gray and blue parts refer to 2 μm-thick silicon dioxide and 340 nm-thick silicon, respectively. Fig. 2(c) and (d) show the E-field intensity of the TE optical mode profiles of ridge and slot-ridge waveguides at the wavelength of 1550 nm simulated using Lumerical MODE Solutions. Clearly, the optical mode of slot-ridge waveguides was largely confined in the slot, where a type of nonvolatile PCMs, chalcogenide $Ge_2Sb_2Te_5$ (GST), was filled. It means that the optical mode of slot-ridge waveguides exhibits higher overlap efficiency with GST compared to that of the ridge waveguides. More details of the optical mode profiles were shown in Fig. 3. The orange part represents GST, whose thickness was 50 nm and the sidewall is about half of the thickness according to our experience of the fabrication process. The complex refractive indices for amorphous GST (a_GST) and crystalline GST (c_GST) are 4.6+0.18i and 7.2+1.9i using spectroscopic ellipsometry [22]. The E-field intensity of the optical mode profiles of ridge and slot-ridge waveguides are shown in Fig. 3(c)–(f). The effective complex refractive indices (n+ik) of ridge waveguides with a_GST and c_GST were 3.15+0.03i and 4.18+1.53i while those of slot-ridge waveguides with a_GST and c_GST were 3.03+0.02i and 3.80+1.63i. The difference of the imaginary part Δk refers to the change of the optical transmission in the waveguides, which were 1.50 and 1.61 for ridge waveguides and slot-ridge waveguides, respectively. This means that the change of the optical transmission of slot-ridge waveguides was larger than that of ridge waveguides. Besides, the power confinement factor Γ_t is defined as [18]:

$$\Gamma_t = \frac{\iint_{\Omega t} E^* \times H dx dy}{\iint_{\Omega T} E^* \times H dx dy} \tag{1}$$

where Ωt refers to the GST region. ΩT represents the entire cross-section of the waveguides. The change of Γ_t between c_GST and a_GST are 36.81% and 36.57% for slot-ridge waveguides and ridge waveguides, respectively. It means that the change of the normalized optical power in

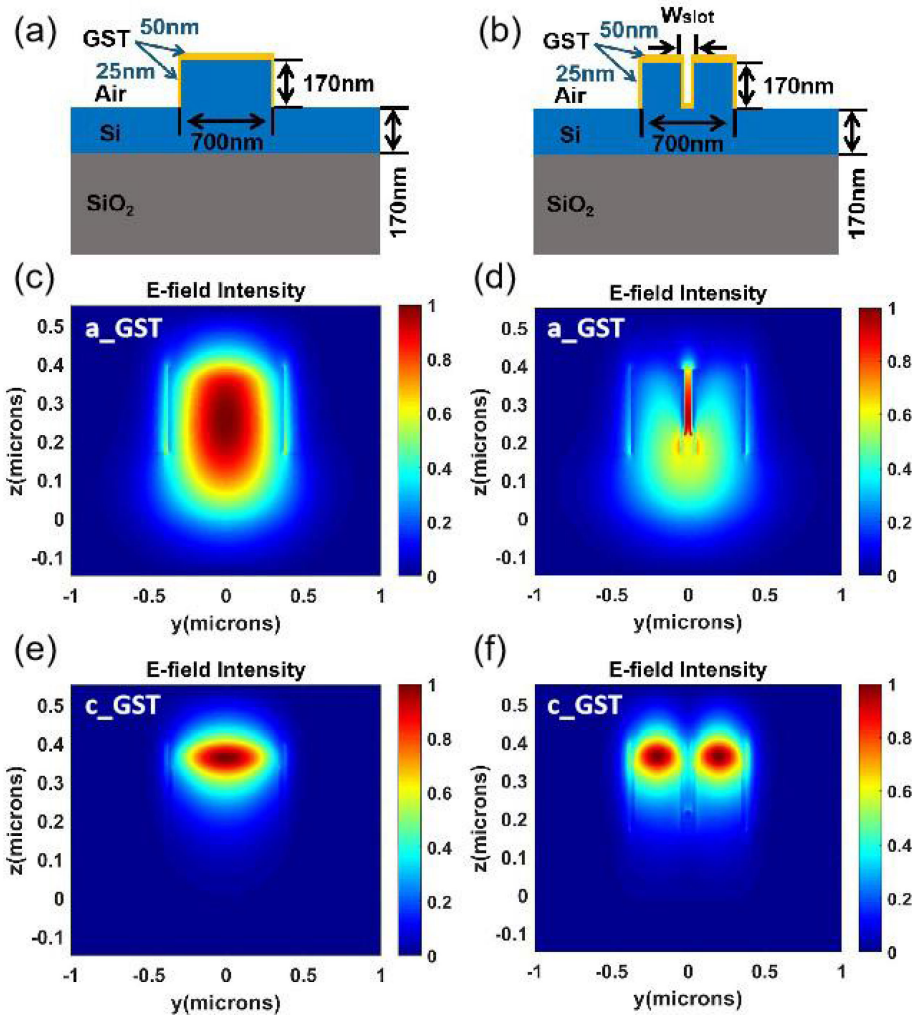


Fig. 3. (a) The cross-sectional view of (a) ridge waveguides with GST and (b) slot-ridge waveguides with GST. The E-field intensity of the optical mode profiles of (c) ridge waveguides with a_GST and (d) slot-ridge waveguides with a_GST. The E-field intensity of the optical mode profiles of (e) ridge waveguides with c_GST and (f) slot-ridge waveguides with c_GST.

the GST region is larger for slot-ridge waveguides with the phase transition between c_GST and a_GST.

3. Design and Optimization

For our slot-ridge waveguide structure, the optical transmission T was monitored to analyze and quantify the enhancement effect. The change in the optical transmission ΔT was obtained using 3D FDTD simulation method to compare slot-ridge waveguides with ridge waveguides. Both amorphous and crystalline phase states of the GST were employed to attach different weight coefficients to the input optical signal. In the simulation, the main parameters of slot-ridge waveguides, such as the length of GST Lgst and the width of the slot Wslot were designed and optimized. For the optimization of Lgst, the thickness of GST and Wslot were set to be 50 nm and 100 nm, respectively. As shown in Fig. 4(a), T of both ridge and slot-ridge waveguides decreases as Lgst increases. Clearly, T of the slot-ridge waveguide is lower than that of the ridge waveguide because of the

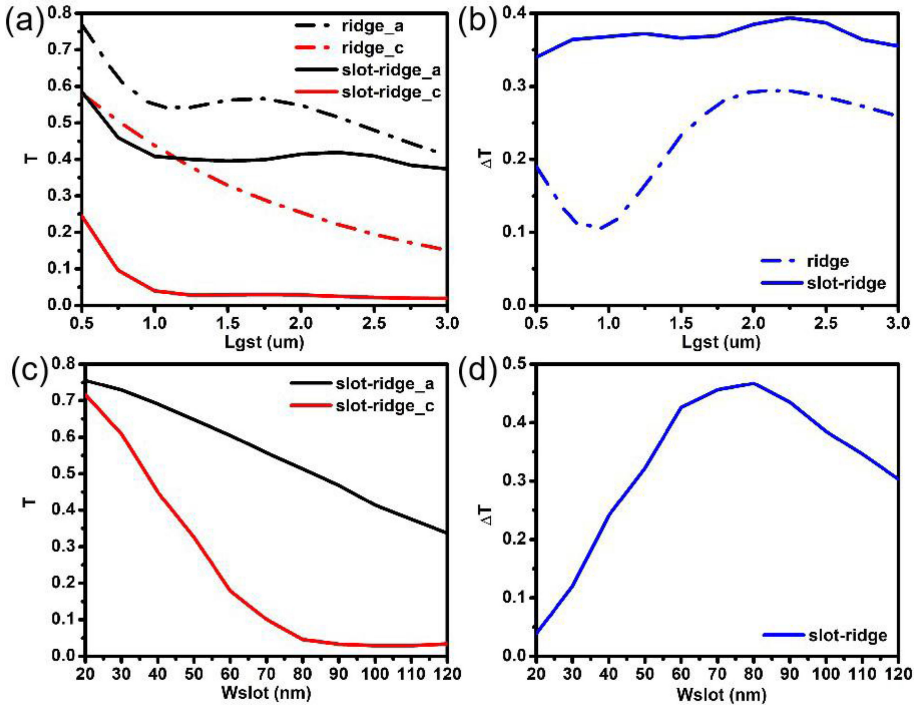


Fig. 4. (a) The optical transmission and (b) the optical transmission change of ridge and slot-ridge waveguides with different length of a_GST and c_GST, where the thickness of GST and Wslot are 50 nm and 100 nm, respectively. (c)The optical transmission and (d) the optical transmission change of slot-ridge waveguides with different width of the slot, where the thickness of GST and Lgst were 50 nm and 2 μm, respectively.

stronger interaction of the optical mode with GST and the coupling loss of the mode converter about 0.22dB by simulation. The mode converter was designed with a symmetric multimode interference (MMI) structure, which is based on the similarity of the optical field distributions between its two-fold image and the slot waveguide [23]. The width and length of MMI are 2 μm and 4.7 μm while the length of the slot taper is 10.5 μm, which is used to connect MMI and the slot waveguide for more efficient coupling. There was a dip in the black short dash dotted-line in Fig. 4(a), which maybe resulted from the non-negligible scattering loss in comparison with the total absorption while the length of GST is shorter than 1.25 μm [24]. Also, ΔT between different phase states of GST was larger for the slot-ridge waveguide device shown in Fig. 4(b). Specially, for a GST length of 1 μm, ΔT of slot-ridge waveguides was 2.53 times larger than that of ridge waveguides. This implies that photonic synapses utilizing slot-ridge waveguides have a larger weight dynamic range, which results in more multilevel weights and lower dynamic power consumption to achieve a precise weight. In addition, because of the nonvolatile nature of GST, the weights can be stored in situ by employing our photonic synapses, which refers to almost no static power consumption, thus ensuring in-memory computing with lower power consumption.

For further investigation on the advantages of slot-ridge waveguides, T and ΔT as a function of the slot width Wslot are analyzed and the results were shown in Fig. 4(c) and (d). The thickness of GST and Lgst were 50 nm and 2 μm, respectively. T of slot-ridge waveguides with c_GST decreases until Wslot reaches 80 nm, above which the optical transmission remains almost unchanged while that with a_GST decreases steadily with increasing Wslot. Therefore, ΔT initially increases reaches a maximum of 0.467 for a Wslot value of 80 nm, and finally decreases, implying that the difference of the overlap between the optical mode and different phases of GST is possibly the best with such a slot size.

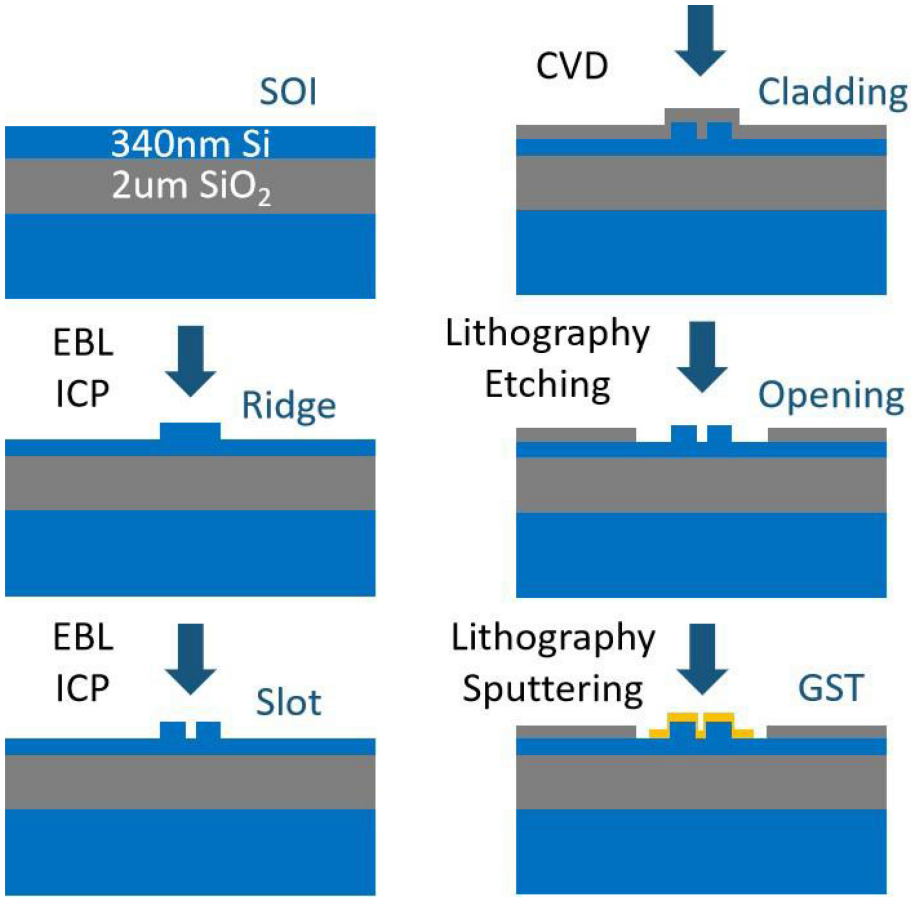


Fig. 5. The fabrication process flow of the slot-ridge waveguides with GST.

4. Results and Discussion

4.1 Device Fabrication

Based on the parameter optimization, the slot-ridge waveguides with GST were realized and demonstrated in experiments. The fabrication process flow was illustrated in Fig. 5. The device was fabricated on an SOI wafer with 340 nm-thick top silicon layer and 2 μm-thick buried silicon oxide layer. Firstly, the ridge waveguide was patterned by electron beam lithography (EBL) process and then etched by inductively coupled plasma (ICP) process. The 100 nm-width slot structure was shared with the same process. After these processes, 700 nm-thick SiO_2 cladding layer was deposited by plasma enhanced chemical vapor deposition (PECVD). Then a 5 μm-width sputtering window was opened by lithography and wet-etching with buffered oxide etch (BOE) for 3 minutes. Finally, 20 nm GST was sputtered on the top of the slot-ridge waveguide and followed by 10 nm indium tin oxide (ITO), which prevented the oxidation of the 20 nm GST. Following these processes, the fabrication was finally realized after the liftoff process. The micrograph image of the whole slot-ridge waveguides is shown in Fig. 6(a), and the scanning electron microscope (SEM) images of the mode converters, the slot-ridge waveguide, and the grating coupler are shown in Fig. 6(b)–(d). The width of the slot is about 100 nm while the width of the ridge is about 700 nm, which are good agreement with the design parameters. The SEM images of slot-ridge and ridge waveguides covered with GST are shown in Fig. 7(a) and (b) while those of the sputtering window and waveguides with GST could be found in Fig. 7(c) and (d). The width of the sputtering window and GST were approximately 5 μm and 2 μm, respectively.

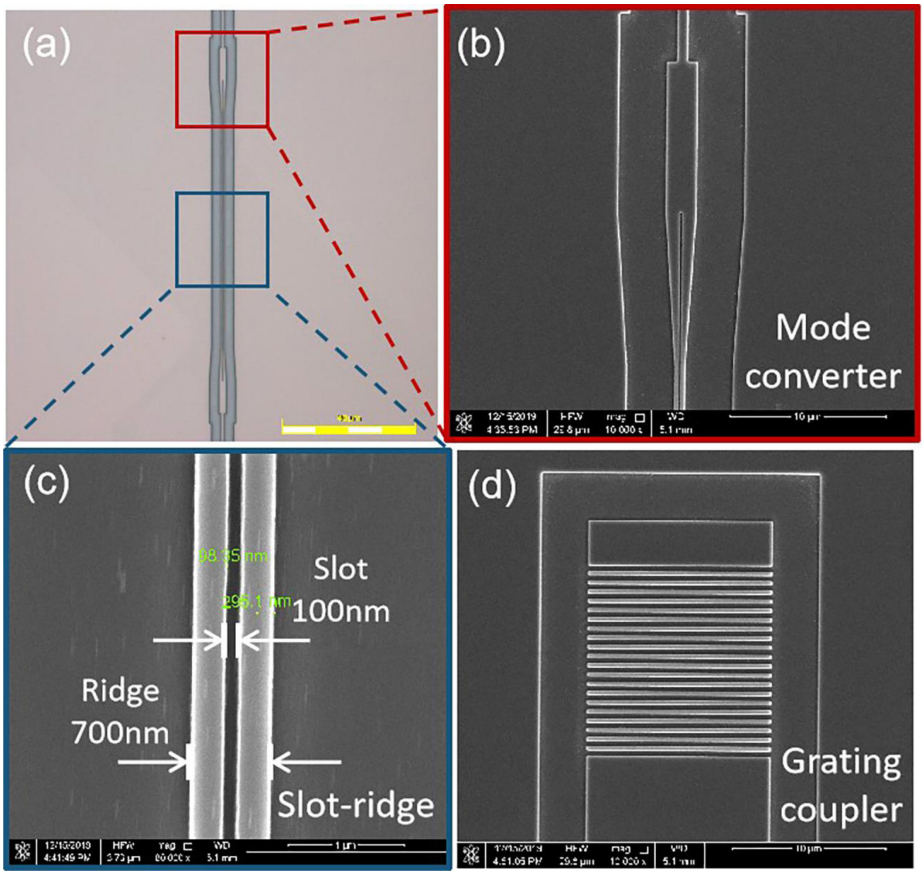


Fig. 6. (a) The micrograph image of the whole slot-ridge waveguides. (b-d) The SEM images of the mode converters, the slot-ridge waveguides, and the grating coupler.

4.2 Experimental Results

For the phase transition of GST, it could be achieved by thermal, electrical, and optical pulses [25]–[26]. Here the phase transition was triggered by heating samples on a hot plate at 300 °C for 4 minutes and then naturally cooling. From Fig. 7(e) and (f), it could be seen that the color of GST/ITO changed from green to brown with the phase state changed from amorphous to crystalline state. The optical transmission spectra of the slot-ridge and ridge waveguides with a_GST and c_GST were shown in Fig. 8(a), which were normalized by the source power. The length of the slot was approximately 200 μm, 100 μm, and 50 μm for three test devices named for slot_1, slot_2, and slot_3, respectively. The optical transmission spectra of the slot-ridge waveguides were lower than those of the ridge waveguides because of the stronger interaction between the optical mode and GST, the extra coupling loss of the mode converter, and larger propagation loss resulting from the roughness of the sidewalls. The change of optical transmission spectra of waveguides with a_GST and c_GST, were got in Fig. 8(b), which showed that the change of optical transmission spectrum of the slot-ridge waveguides was about 3.5dB higher than that of the ridge waveguides on average, which meant that photonic synapses based on the slot-ridge waveguide had the potential to achieve large weight dynamic range and more weight values. Such an attempt is left for our future work. It was noted that the change of optical transmission spectrum of slot_2 was a little higher than those of the others, which maybe resulted from the error of the manufacture and test process.

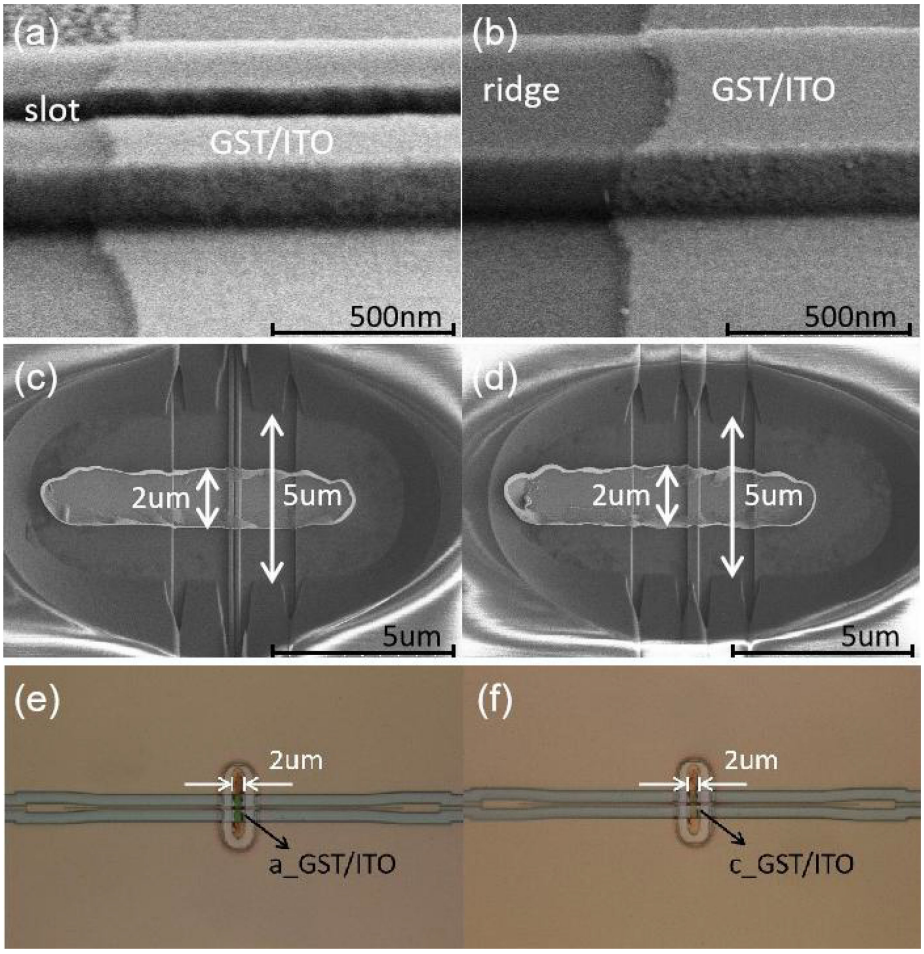


Fig. 7. The SEM images of (a) slot-ridge waveguides covered with GST, (b) ridge waveguides covered with GST, (c) the sputtering window and slot-ridge waveguides with GST, and (d) the sputtering window and ridge waveguides with GST. The micrographs of slot-ridge waveguide with a_GST and c_GST (e) before the phase transition, and (f) after the phase transition.

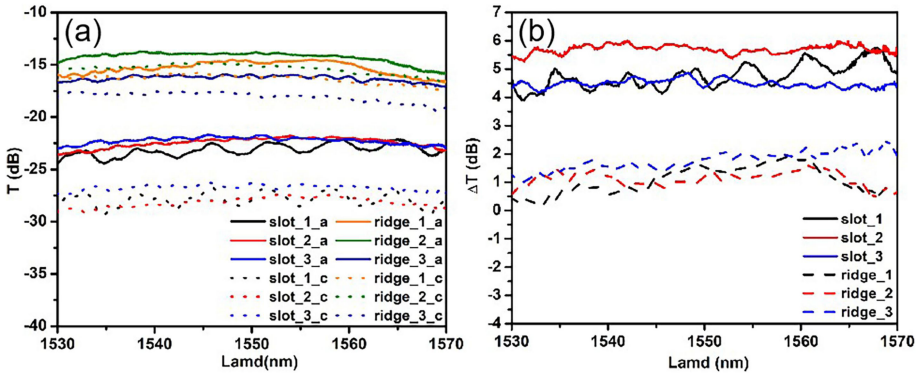


Fig. 8. (a) The optical transmission spectra of the slot-ridge and ridge waveguides with a_GST and c_GST. (b) The change of optical transmission spectra of the slot-ridge and ridge waveguides.

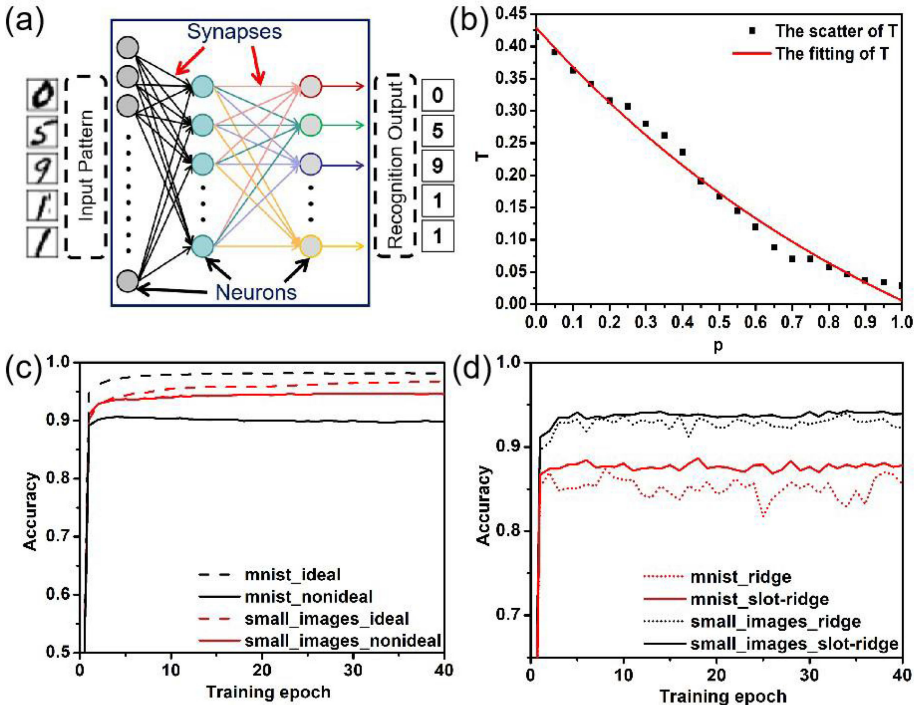


Fig. 9. (a) The schematic diagram of our neural network, which is a three-layer perceptron. (b) The relationship between the optical transmission and the crystallization level p. (c) The accuracies of two datasets small_images and mnist using devices with ideal characteristic and devices based on our photonic synapses with nonideal characteristic. (d) The accuracies of the recognition task of small_images and mnist with devices based on the simulation of ridge and slot-ridge waveguides.

4.3 Recognition Task

To determine the performance of our photonic synapses after the parameter optimization, we performed a recognition task using our photonic synapses based on slot-ridge waveguides with the simulated properties. First, we introduced a crystallization level parameter, which led to multilevel weights needed in the recognition task. The crystallization level resulted from the mixed state of amorphous and crystallized molecules with different proportions during crystallization [27]. Theoretically, the effective permittivity $\varepsilon_{\mathrm{eff}}$(p) is approximately defined by the effective-medium theory:

$$\frac{\varepsilon_{eff}(p)-1}{\varepsilon_{eff}(p)+2} = p \times \frac{\varepsilon_c - 1}{\varepsilon_c + 2} + (1-p) \times \frac{\varepsilon_a - 1}{\varepsilon_a + 2} \tag{2}$$

where p is the crystallization level, and ε_{a} and ε_{c} are the permittivities for a_GST and c_GST. The relationship between the complex refractive indices and the permittivities of GST is governed by the equation, $\varepsilon(\lambda) = (n + ik)^2$, where the complex refractive indices of a_GST and c_GST are 4.6+0.18i and 7.2+1.9i, respectively. Further, $\varepsilon_{\mathrm{eff}}$(p) is indirectly calculated by using (2). The optical transmission at different crystallization levels obtained in Lumerical FDTD is plotted in Fig. 9(b). The black scatter is the simulation data, whereas the red solid curve is the fitting data given by the function:

$$y = y_0 + A_1 \exp(-x/t_1) \tag{3}$$

where $-1/t_1$ is the nonlinearity factor estimated to be 0.8576 by the fitting data. It is an analogous parameter compared with that of the memristor conductance [28], which indicates the nonideal characteristic of the device when used to apply weights.

A handwritten digit recognition task was performed by a three-layer perceptron shown in Fig. 9(a) with gradient descent, which is based on the backpropagation (BP) algorithm [29]. The optimized parameters, weights and bias, are obtained by calculating gradient to make the cost function minimum and the accuracy maximum possibly. In the optimization, photonic synapses play the part of applying weights, which refer to the connect strength between the neurons. A crossbar simulator CrossSim is used to perform the recognition task, which is an API to model realistic device properties and variability [30]. For the recognition task, there are two datasets: an 8×8 version of MNIST, small_images [31], and the standard MNIST handwritten digits, mnist [32]. Therefore, the sizes of the perceptron networks are 64×36×10 and 784×300×10 for small_images and mnist, which means that the number of the neurons in the hidden layers are 36 for small_images and 300 for mnist, respectively. Additionally, the sigmoid function is chosen to be the neuron activation function. The weights of all three layers are mapped to the photonic synapses, where the multiplication matrix operations of the input patterns and their corresponding specific weights are performed. For both two datasets, the number of training epochs and learning rates are 40 and 0.1, respectively, and the weight levels are 127. The recognition accuracies of the neural network are based on the test dataset after the network was trained on the training dataset. The accuracies of the two datasets using our photonic synapses based on slot-ridge waveguides are shown in Fig. 9(c). The dotted lines are the theoretical baseline accuracies, which refer to the devices with ideal characteristics. The solid lines represent the devices based on our proposed synapses with nonideal characteristics. For our synapses, the accuracies of small_images and mnist are 94.7% and 90.7%, which are 2.2% and 8.4% lower than those of devices with ideal characteristics. That is because the weight deviations resulting from the nonlinear characteristics during the weight mapping.

Furthermore, to compare our photonic synapses based on slot-ridge waveguides with those primitive ones based on ridge waveguides, write noises related to the weight dynamic range are added to the simulation for qualitative analysis as the mapped weights are normalized with BP algorithm during the gradient descent. The write noise is modeled after a Gaussian distribution with a standard deviation σ as discussed in [33]:

$$W = W_0 + \Delta W + N(\sigma) \tag{4}$$

$$\sigma = \sqrt{\Delta W \times W_{range}} \times \frac{W_0}{W_{range}} \times \sigma_{MN} \tag{5}$$

where W and ΔW are the target photonic weight and the intended photonic weight change, while W0 and Wrange represent the initial photonic weight state and the photonic weight range. Further, σ_{MN} is a dimensionless standard deviation with a typical value of 0.1. For $\mathrm{W_{range}}$, the maximum weight is set to be 1, and the minimum weight is given by $\mathrm{T_{min}/T_{max}}$. $\mathrm{T_{min}}$ and $\mathrm{T_{max}}$ represent the minimum and maximum optical transmission, which are easily estimated from the simulation results in Fig. 4(a). Then the minimum weight of our synapses and primitive synapses are 0.07 and 0.46, respectively. The solid and dotted lines representing the results of small_images and mnist recognition tasks by using the slot-ridge and ridge waveguides are shown in Fig. 9(d). Compared with those of the synapses based on ridge waveguides, the accuracies of our photonic synapses in small-images and mnist recognition are improved by 1% and 2.6% on average, which implies that our photonic synapses have better performance in applying weights to the network in the recognition task. Also, compared with Fig. 9(c), there are more fluctuations in the curves with write noises in Fig. 9(d) due to the fluctuations in the write noises themselves.

To make sure the enhance effect of the fabricated slot-ridge waveguides, the same recognition task was performed based on the same nonlinearity factor of the previous simulation. $\mathrm{T_{max}}$ and $\mathrm{T_{min}}$ could be got from the experimental results in Fig. 8(a), and the minimum weight of our synapses and primitive synapses were calculated be 0.61 and 0.85. The other parameters of the neural network were the same as the previous recognition task. The results of small_images and mnist recognition tasks are shown in Fig. 10. Obviously, the accuracies of recognition tasks both small-images and mnist using our synapses are better than those of primitive synapses. For the mnist recognition

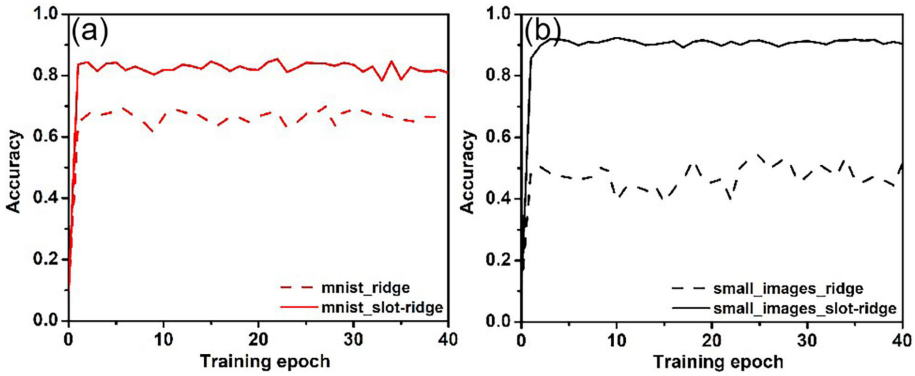


Fig. 10. The accuracies of the recognition task of (a) mnist and (b) small_images with fabricated devices based on the ridge and slot-ridge waveguides.

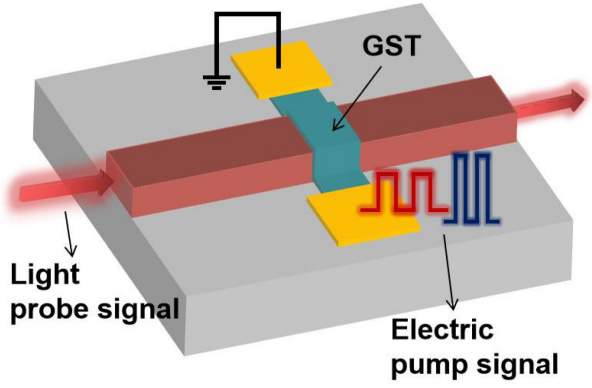


Fig. 11. The schematic diagram of the system with electric pulses triggering the phase change of the GST material.

task, the accuracy of our photonic synapses is 15.6% higher than that of primitive synapses based on ridge waveguides.

In the above-mentioned recognition task, a simple perceptron network is chosen to verify our photonic synapses. However, it needs to be emphasized that the architectures of the neural networks are universality and flexible enough, such as convolutional and recurrent neural networks [34]–[35], to tackle more tough recognition tasks. In addition, there are different neural algorithms to optimize the recognition accuracy, such as genetic algorithms [36]. In other words, the crossbar simulator can be considered as an algorithm accelerator, and several different neural architectures and algorithms will be demonstrated in our future work. Moreover, the pump signal triggering the phase change and the probe signal modulated by PCMs in recent photonic synapses with PCMs are both light pulses [4], [15], [17], [20]. However, in this protocol there are some disadvantages, such as complex measurement systems, inconvenient test operation, and unstable transition conditions. To solve these problems, electric pulses could be used as the pump signal to trigger the phase transition of PCMs in our next work. A schematic diagram of this method is shown in Fig. 11. The phase change process is thermally induced by Joule heating during a current pulse [26].To switch c_GST to a_GST, a short electrical pulse with high power, RESET pulse, will be used to increase the temperature to above the melting temperature of GST while the transition of a_GST to c_GST is triggered by a long one with lower power, SET pulse, to make the GST material above its crystallization. Intermediate phase states are achieved by specific electrical pulses with specific width and power, which are between those of the SET and RESET pulses. The phase

states are modulated by specific electric pump pulses, whereas a light probe pulse is weighted by the interaction between the optical mode and PCMs. The measurement setup is simpler, and the transition conditions are more controllable compared to those of the light pump pulses' setup current used. Besides, the fabrication of the device is compatible with the traditional CMOS process, which indicates a potential larger-scale integration. In brief, photonic synapses have the potential to be the highlight components in building neuromorphic networks for efficient in-memory computing in silicon photonics.

5. Conclusion

In this paper, we presented and demonstrated the design of a photonic synapse based on slot-ridge waveguides with nonvolatile PCMs. Compared with photonic synapses based on ridge waveguides with PCMs, the change of the interaction between the optical mode and PCMs is enhanced by using the slot waveguide structure in our proposed photonic synapses. The optical transmission change of our synapses was observed to be higher than that of primitive ones by the simulation calculation. It implies that larger weight dynamic range could be achieved by employing our proposed photonic synapses. The device was experimentally demonstrated on an SOI platform using the standard CMOS process. The change of optical transmission spectrum of the slot-ridge waveguides was about 3.5dB higher than that of the ridge waveguides, which meant that photonic synapses based on the slot-ridge waveguides could have larger weight dynamic range and more weight values. In addition, a recognition task was performed using our photonic synapses, and for the MNIST handwritten digits a 90.7% recognition accuracy was achieved, which was 2.6% higher than that of synapses based on ridge waveguides. Meanwhile, lower power in-memory computing with our photonic synapses will be achieved because of the nonvolatile nature of the PCMs. This work has the potential to synthesize photonic synapses with more weight levels and lower power consumption and realize neuromorphic networks with higher efficient in-memory computing in silicon photonics in future.

References

[1] J. Neumann, "First draft of a report on the EDVAC," *IEEE Ann. Hist. Comput.*, vol. 15, no. 4, pp. 27–75, Oct. 1993.
[2] A. Sebastian, M. Le Gallo, R. Khaddam-Aljameh, and E. Eleftheriou, "Memory devices and applications for in-memory computing," *Nat. Nanotechnol.*, vol. 15, no. 7, pp. 529–544, 2020.
[3] J. Hasler and B. Marr, "Finding a roadmap to achieve large neuromorphic hardware systems," *Front. Neurosci.*, vol. 7, no. 118, pp. 1–29, 2013.
[4] Z. Cheng, C. Rios, W. H. P. Pernice, C. D. Wright, and H. Bhaskaran, "On-chip photonic synapse," *Sci. Adv.*, vol. 3, no. 9, pp. 1–6, 2017.
[5] I. Boybat *et al.*, "Neuromorphic computing with multi-memristive synapses," *Nat. Commun.*, vol. 9, no. 2514, pp. 1–12, 2018.
[6] S. Lim *et al.*, "Improved synapse device with MLC and conductance linearity using quantized conduction for neuromorphic systems," *IEEE Electron. Device Lett.*, vol. 39, no. 2, pp. 312–315, Feb. 2018.
[7] S. Oh, Z. Huang, Y. Shi, and D. Kuzum, "The impact of resistance drift of phase change memory (PCM) synaptic devices on artificial neural network performance," *IEEE Electron. Device Lett.*, vol. 40, no. 8, pp. 1325–1328, Aug. 2019.
[8] S. R. Nandakumar, M. Le Gallo, I. Boybat, B. Rajendran, A. Sebastian, and E. Eleftheriou, "A phase-change memory model for neuromorphic computing," *J. Appl. Phys.*, vol. 124, no. 15, p. 152135, 2018.
[9] P. Yao *et al.*, "Face classification using electronic synapses," *Nat. Commun.*, vol. 8, no. 15199, pp. 1–8, 2017.
[10] C. Li *et al.*, "Efficient and self-adaptive in-situ learning in multilayer memristor neural networks," *Nat. Commun.*, vol. 9, no. 2385, pp. 1–8, 2018.
[11] F. Cai *et al.*, "A fully integrated reprogrammable memristor–CMOS system for efficient multiply–accumulate operations," *Nat. Electron.*, vol. 2, no. 7, pp. 290–299, 2019.
[12] J. Shi, S. D. Ha, Y. Zhou, F. Schoofs, and S. Ramanathan, "A correlated nickelate synaptic transistor," *Nat. Commun.*, vol. 4, no. 1, pp. 1–9, 2013.
[13] G. Agnus *et al.*, "Two-terminal carbon nanotube programmable devices for adaptive architectures," *Adv. Mater.*, vol. 22, no. 6, pp. 702–706, 2010.
[14] C. L. Chen, K. Kim, Q. Truong, A. Shen, Z. Li, and Y. Chen, "A spiking neuron circuit based on a carbon nanotube transistor," *Nanotechnology*, vol. 23, no. 27, 2012, Art. no. 275202.
[15] J. Feldmann, N. Youngblood, C. D. Wright, H. Bhaskaran, and W. H. P. Pernice, "All-optical spiking neurosynaptic networks with self-learning capabilities," *Nature*, vol. 569, no. 7755, pp. 208–214, 2019.

[16] H. Zhang *et al.*, "Miniature multilevel optical memristive switch using phase change material," *ACS Photon.*, vol. 6, no. 9, pp. 2205–2212, 2019.
[17] J. Feldmann *et al.*, "Calculating with light using a chip-scale all-optical abacus," *Nat. Commun.*, vol. 8, no. 1256, pp. 1–8, 2017.
[18] H. Zhang *et al.*, "Ultracompact Si-GST hybrid waveguides for nonvolatile light wave manipulation," *IEEE Photon. J*, vol. 10, no. 1, pp. 1–10, Feb. 2018.
[19] X. Li *et al.*, "Fast and reliable storage using a 5 bit, nonvolatile photonic memory cell," *Optica*, vol. 6, no. 1, pp. 1–6, 2018.
[20] C. Rios *et al.*, "In-memory computing on a photonic platform," *Sci. Adv.*, vol. 5, no. 2, pp. 1–9, 2019.
[21] V. R. Almeida, Q. Xu, C. A. Barrios, and M. Lipson, "Guiding and confining light in void nanostructure," *Opt. Lett.*, vol. 29, no. 11, pp. 1209–1211, 2004.
[22] S. Y. Kim, S. J. Kim, H. Seo, and M. R. Kim, "Variation of the complex refractive indices with Sb-addition in Ge-Sb-Te alloy and their wavelength dependence," *Opt. Data Storage*, vol. 3401, no. 98, pp. 112–115, 1998.
[23] Q. Deng, L. Liu, X. Li, and Z. Zhou, "Strip-slot waveguide mode converter based on symmetric multimode interference," *Opt. Lett.*, vol. 39, no. 19, pp. 5665–5668, 2014.
[24] C. Rios *et al.*, "Controlled switching of phase-change materials by evanescent-field coupling in integrated photonics " *Opt. Mater. Exp.*, vol. 8, no. 9, pp. 2455–2470, 2018.
[25] M. Wuttig and N. Yamada, "Phase-change materials for rewriteable data storage," *Nat. Mater.*, vol. 6, no. 11, pp. 824–832, 2007.
[26] M. H. Lankhorst, B. W. Ketelaars, and R. A. Wolters, "Low-cost and nanoscale non-volatile memory concept for future silicon chips," *Nat. Mater.*, vol. 4, no. 4, pp. 347–352, 2005.
[27] Y. Chen *et al.*, "Engineering the phase front of light with phase-change material based planar lenses," *Sci. Rep.*, vol. 5, no. 1, pp. 1–7, 2015.
[28] J. Chen *et al.*, "LiSiOx-based analog memristive synapse for neuromorphic computing," *IEEE Electron Device Lett.*, vol. 40, no. 4, pp. 542–545, Apr. 2019.
[29] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *Nature*, vol. 323, no. 6088, pp. 533–536, 1986.
[30] S. Agarwal *et al.*, "CrossSim: Crossbar simulator," Sandia National Laboratories, 2017. [Online]. Available: https://cross-sim.sandia.gov/
[31] D. Dua and C. Graff, "UCI machine learning repository," 2019. [Online]. Available: http://archive.ics.uci.edu/ml
[32] Y. LeCun, C. Cortes, and C. J. Burges, "The MNIST database of handwritten digits," 1998. [Online]. Available: http://yann.lecun.com/exdb/mnist
[33] S. Agarwal *et al.*, "Resistive memory device requirements for a neural algorithm accelerator," in *Proc. 2016 Int. Joint Conf. Neural Netw.*, 2016, pp. 929–938.
[34] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," in *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.
[35] J. L. Elman, "Finding structure in time," *Cogn. Sci.*, vol. 14, no. 2, pp. 179–211, 1990.
[36] D. Goldberg, *Genetic Algorithms in Search Optimization and Machine Learning*, Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc., 1989, pp. 1–25.