All-Spin Artificial Neural Network Based on Spin–Orbit Torque-Induced Magnetization Switching

Zhen Cao, Shuai Zhang, Jincheng Hou, Wei Duan, and Long You, *Senior Member, IEEE*

***Abstract*—A reliable design of all spin artificial neural networks based on spin–orbit torque (SOT) devices has been proposed and demonstrated in W/CoFeB/MgO heterostructures. In our scheme, a single device acts as a neuron with an rectified linear unit (ReLU) activation function. Besides, the synaptic function is also realized using the devices made of the same film structure as that used in neural devices, but with different film thicknesses. Furthermore, system-level simulations are performed to classify the MNIST database by exploiting SOT neurons and SOT synapses characteristics. The high recognition rate (91.01%) confirms the feasibility of our scheme.**

***Index Terms*—All-spin neural network, rectified linear unit (ReLU), spin–orbit torque (SOT), synapse.**

I. Introduction

COMPUTERS based on the von-Neumann architecture consume substantial energy shuttling information between memory units and central processing unit [1]. In contrast, brains store information locally where it is processed [2]. Neuromorphic computing which uses brain-inspired principles can perform cognitive tasks, such as recognition and reasoning, more efficiently due to its massive parallelism and energy efficiency [3]. Spintronic devices stand out from the emerging devices due to their ultrafast voltage operation and high energy efficiency [4]. Meanwhile, spintronic devices possess key features required for artificial synapses and neurons, such as nonvolatility and nonlinearity [5].

Recently, spintronic devices emulating the functions of synapses and neurons have been proposed [6], [7], [8]. It is worth noting that most reported spintronic neurons perform the sigmoidal activation function. However, due to the saturation characteristic of the sigmoid function, the vanishing gradient problem often occurs in deep neural networks, while the rectified linear unit (ReLU) function can overcome the vanishing gradient problem because exponential terms are eliminated on its back propagation [9]. However, in the existing experimental implementation of the ReLU activation function scheme [10], [11], the neural components need to be applied with additional bias currents and microwave signals, which increases the complexity and area of the circuit [12].

In this work, an spin–orbit torque (SOT) neuron with ReLU activation function is proposed and experimentally demonstrated in a single device relying on SOT-induced magnetization switching. Besides, synaptic function is achieved. To demonstrate the feasibility of our scheme, we constructed a three-layer perceptron with experimentally measured characteristics of SOT synapses and SOT neurons, and performed training on the written digit dataset from the Modified National Institute of Standards and Technology (MNIST) database and achieve high recognition rates.

II. Device Fabrication and Characterization

Our SOT synapse device is composed of the following layers: W (5 nm)/CoFeB (1.1 nm)/MgO (2 nm)/Ta (2 nm) and is patterned into $30 \times 150\ \mu\text{m}^2$ Hall bar structures by standard photolithography and ion-milling techniques, as shown in Fig. 1(a).

To quantitatively monitor the magnetization, anomalous Hall resistance (R_{AH}) measurements are performed to evaluate the perpendicular component of the magnetization when R_{AH} versus out-of-plane field H_z is measured [13]. The loops of R_{AH} as a function of H_z show hysteresis with sharp switching, indicating a strong perpendicular magnetic anisotropy of the devices, as shown in Figs. 1(b). Then, we investigate the response of R_{AH} to the in-plane current (I_x) under a small in-plane field, $H_x = +100$ Oe. The results are shown in Figs. 1(c). Such bipolar switching behaviors are the signature of SOT generation at the W/CoFeB interface due to the spin

Manuscript received 7 July 2023; revised 26 September 2023; accepted 19 October 2023. Date of publication 3 November 2023; date of current version 28 November 2023. This work was supported in part by the National Key Research and Development Program of China under Grant 2020AAA0109005; in part by the National Natural Science Foundation of China (NSFC) under Grant 62074063, Grant 61821003, and Grant 61674062; in part by the Wuhan National High Magnetic Field Center through the Interdisciplinary Program under Grant WHMFC202119; in part by the Huazhong University of Science and Technology, Shenzhen Virtual University Park, under Grant 2021Szvup091; and in part by the Shenzhen Science and Technology Program under Grant JCYJ20220818103410022. The review of this article was arranged by Editor P. Narayanan. *(Corresponding author: Long You.)*

Zhen Cao, Shuai Zhang, Jincheng Hou, and Wei Duan are with the School of Integrated Circuits, Huazhong University of Science and Technology, Wuhan 430074, China.

Long You is with the School of Integrated Circuits, Huazhong University of Science and Technology, Wuhan 430074, China, and also with the Shenzhen Huazhong University of Science and Technology Research Institute, Shenzhen 518000, China (e-mail: lyou@hust.edu.cn).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TED.2023.3327031.

Digital Object Identifier 10.1109/TED.2023.3327031

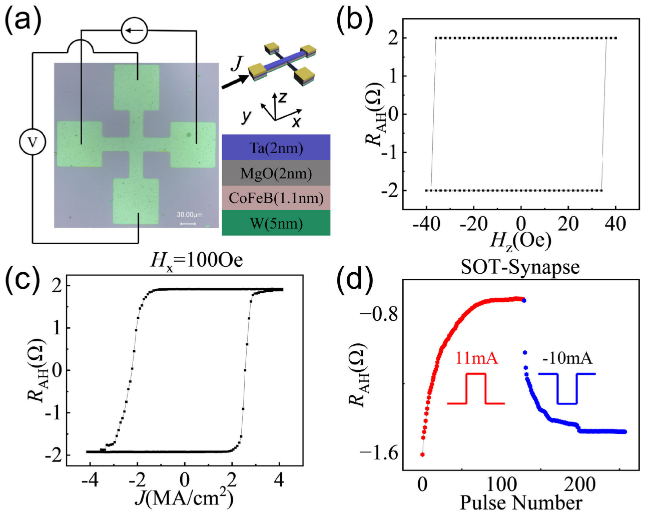

Fig. 1. (a) SOT synapse device structure with measurement setup. (b) R_{AH} versus applied out-of-plane magnetic field (H_z), the coercivity field is 37 Oe. (c) Current-induced magnetization switching under a small constant magnetic field $H_x = +100$ Oe. (d) R_{AH} response to sequential current pulses with a duration of 1 ms, the pulse amplitudes are +11 and −10 mA for positive and negative values, respectively.

Hall effect and at the CoFeB/MgO interface due to the Rashba effect [14], [15].

III. Results and Discussion

The synapses in the brain, located at junctions between neurons, take the role of memorizing and learning. The synaptic weight is changed in an analog manner in the process of learning [5]. Similarly, R_{AH} of SOT synapses represents the weight of synapses which can be modulated in an analog and nonvolatile manner by consecutive current pulses [16], [17].

To achieve synaptic function, a domain wall (DW) is driven back and forth in a continuous manner in the CoFeB layer by applying in-plane current pulses along the W layer. Hence, the magnetization and consequently R_{AH} are modulated in an analog manner [18], [19]. We first saturate the magnetization pointing up (+z-direction). Then, consecutive current pulses with a width of 1 ms and magnitude of −11 mA are applied to the device, and the quasi-continuous R_{AH} gradually decreases. Next, we apply consecutive 8-mA current pulses with a width of 1 ms to the device, which achieves a continuous increase of R_{AH}, as shown in Fig. 1(d). Through current pulses regulation, the R_{AH} states of the SOT synapse can reach 100. Therefore, the R_{AH} modulation induced by consecutive pulses can be utilized to imitate synaptic behaviors.

The linearity of the curve between the R_{AH} and the number of programming pulses is desired to be linear and symmetric, which can perfectly model the weight change in the algorithm. However, R_{AH} of the realistic synaptic devices typically does not vary linearly with the number of applied pulses. Asymmetry refers to the similarity between the weight-increasing trajectory (synaptic potentiation, P) and the weight-decreasing trajectory (synaptic depression, D). As shown in Fig. 2(a), R_{AH} of the SOT synapse does not vary linearly with the number of applied pulses. To quantitatively analyze its nonlinearity and asymmetry, we adopt the extracted nonlinearity behavioral model in [20], and R_{AH} change with the number of pulses is described with the following equations:

$$R_P = B\left(1 - e^{\left(-\frac{P}{A}\right)}\right) + R_{\min}$$
$$R_D = B\left(1 - e^{\left(-\frac{P-P_{\max}}{A}\right)}\right) + R_{\max}$$
$$B = (R_{\max} - R_{\min})/\left(1 - e^{\frac{-P_{\max}}{A}}\right)$$

where $R_{\max}$, $R_{\min}$, and $P_{\max}$ represent the maximum R_{AH}, minimum R_{AH}, and maximum pulse number to tune the device between $R_{\min}$ and $R_{\max}$. The parameter A controls the nonlinearity of conductance tuning behavior, and B is simply a fitting function. $R_{\min}$, $R_{\max}$, $P_{\max}$, A, and B may be different in R_P and R_D. According to the results in [21], High nonlinearity is tolerable when P/D has the same polarity; otherwise, the accuracy degrades dramatically with the increasing asymmetric nonlinearity.

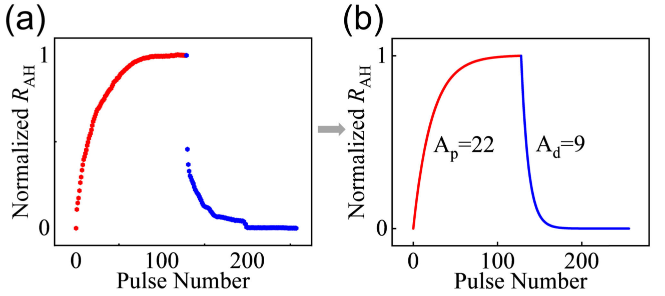

Fig. 2. (a) Normalized R_{AH} behavior of SOT synapse. (b) R_{AH} behavior of SOT synapse after fitting using the nonlinear model.

By fitting in the above way, the parameter A of our synaptic potentiation and depression is 22 and 9, respectively, As shown in Fig. 2(b). Incorporating nonlinearity and symmetry into neural network simulations, the accuracy of neural networks is very low (11.35%). Designing the programming scheme that updates the weight smartly is possible to improve the nonlinearity. For example, the duration of programming pulses can be adjusted in a way that a shorter pulse is applied at the beginning stages, while gradually wider pulses are applied at subsequent stages [22].

Synaptic variations may significantly deteriorate system performance. Basically, there are two types of variation: a random distribution of conductance when the same operating voltages are applied to one device from cycle-to-cycle (C2C), and the conductance variation from device-to-device (D2D). C2C variation is investigated by introducing a Gaussian distribution when the weight is updated, and the influence of D2D variation is treated by bringing errors to each weight. Fig. 3(a) shows the weight increase and decrease of the same device repeated for ten cycles, and the C2C variation of synaptic potentiation and depression is 0.065 and 0.029, respectively. Similarly, Fig. 3(b) shows the results of the weight increase and decrease operation of the five devices, the D2D variation of synaptic potentiation and depression is 0.13 and 0.48, respectively. As shown in Fig. 3(c) and (d), the accuracy decreases as the C2C variation and D2D variation increases. Compared with C2C variation, the network has a better tolerance to D2D variation.

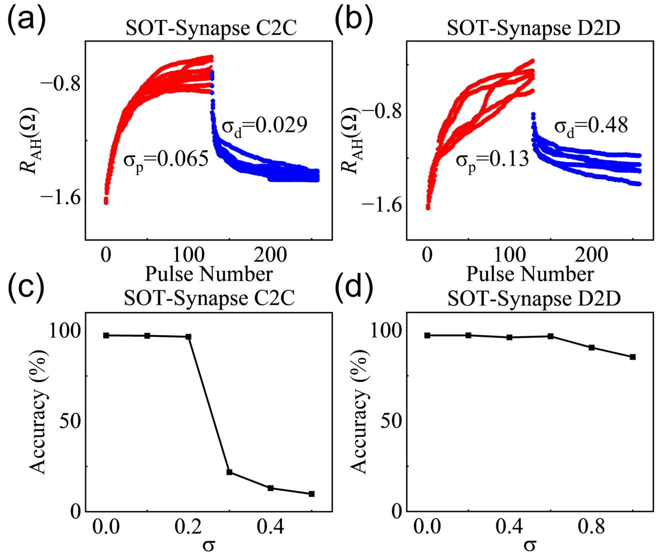

Fig. 3. (a) R_{AH} behavior of SOT synapse during six cycles of 128 conductance states with C2C variation. (b) R_{AH} behaviors for three devices with D2D variation. (c) Recognition accuracy with different C2C variations. (d) Recognition accuracy with different D2D variations.

The role of neurons in the brain is to process information. The neurons fire once the membrane potential reaches a threshold, which enlightens the idea of activation function for neuromorphic computing. Similarly, the SOT neuron also acts as an activation function. The response of R_{AH} to the input current can mimic the ReLU activation function.

Similar to the device fabrication and characterization method of synaptic device, a stack consisting of W (5 nm)/CoFeB (1.2 nm)/MgO (2 nm)/Ta (3 nm) is patterned into 50 × 200 μm^2 Hall bar structures to construct SOT neuron device, as shown in Fig. 4(a). R_{AH}–H_z loop and R_{AH}–I loop of SOT neuron device are shown in Fig. 4(b) and (c), respectively. When the amplitude of the current is not large enough, the magnetization does not switch and the R_{AH}–I curve is flat. With the gradual increase of the current amplitude, the magnetization begins to switch, and R_{AH} shows a good linearity as a function of the current, as shown in the red part of the curve in Fig. 4(c), which is very similar to the ReLU function. The different R_{AH}–I response compared with SOT synapse may be due to different switching modes caused by different thicknesses of the film. Then, an SOT ReLU function is constructed by mapping the selected data (the red dots) in Fig. 4(c) onto a ReLU function. The resulting SOT ReLU function is plotted in Fig. 4(d), which differs only slightly from an ideal ReLU function. Therefore, the nonlinear relationship between R_{AH} and current can be utilized to emulate neuronic behaviors.

Similar to the method used to study synaptic variations, we performed ten measurements of the same device and a single measurement of five different devices, as shown in Fig. 5(a) and (b), respectively. The C2C variation of neuron is 0.014, and the D2D variation of a neuron is 0.18. As shown in Fig. 5(c) and (d), the accuracy decreases as the C2C and D2D increases. Compared with C2C, the network has a better tolerance to D2D.

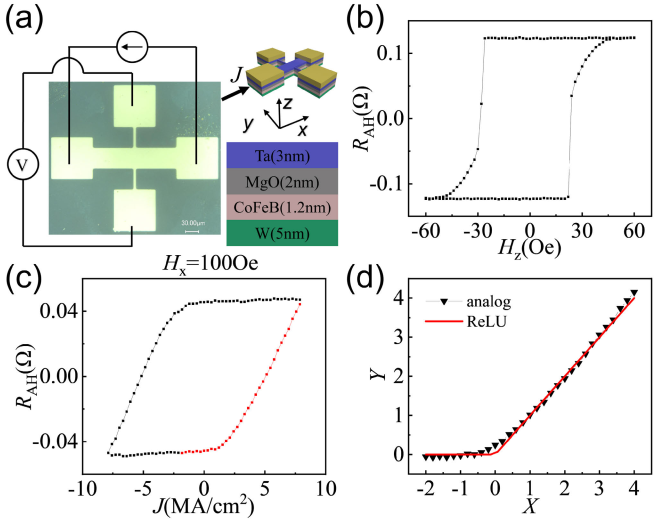

Fig. 4. (a) SOT neuron device structure with measurement setup. (b) RAH versus applied out-of-plane magnetic field (H_z), the coercivity field is around 30 Oe. (c) Current-induced magnetization switching under a small constant magnetic field H_x = +100 Oe. Data points marked by red in the shaded region are selected as the basis of the analog ReLU function. (d) Exemplary plots of the SOT ReLU function and ideal ReLU function.

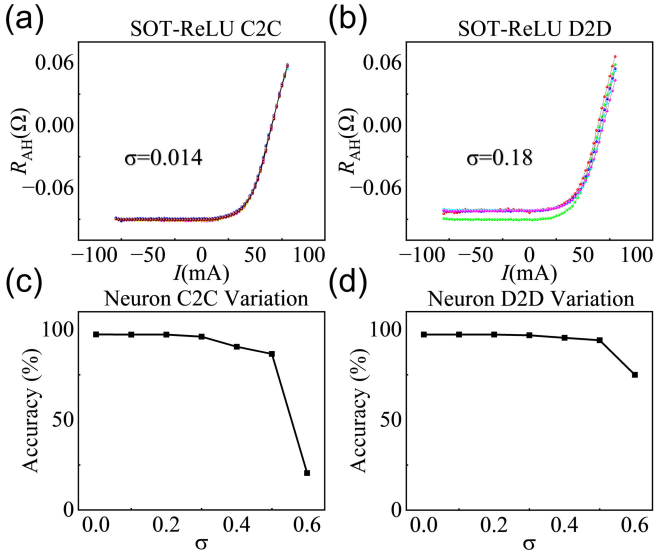

Fig. 5. (a) R_{AH} behavior of SOT neuron during ten cycles with C2C variation. (b) R_{AH} behaviors of SOT neurons for five devices with D2D variation. (c) Recognition accuracy with different C2C variations SOT neuron. (d) Recognition accuracy with different D2D variations SOT neuron.

In the neural network implemented by digital circuits, multistage registers are needed because of the operation involving multiple cycles. With our nonvolatile ReLU activation function, registers are not required.

In our case, the device size of the SOT neuron is 50 × 200 μm^2, the resistance of our device is 250 Ω, as measured in our experiments, and the average amplitude of input current is 0.2 mA. According to [23], considering that the DW location can be displaced and sensed over a minimum distance of 20 nm, we can assume that the size of the device can be reduced to 50 × 800 nm^2 (4-bit discretization), and then, the resistance and the average amplitude of input current of the

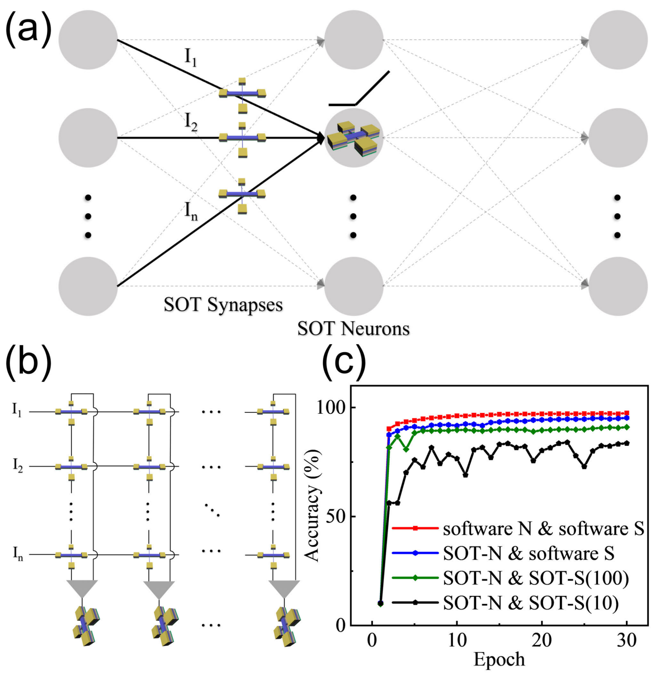

Fig. 6. (a) Three-layer perceptron used for training. (b) SOT synapse array between the input layer and hidden layer and the SOT neuron array of the hidden layer. (c) Pattern recognition accuracy as a function of training iteration. *S* means synapse, and *N* means neuron.

scaled device is 1000 Ω and 0.2 μA, respectively. Assuming that the DW moves at 80 m/s, the input time and reset time can be reduced to 10 ns. Besides, we can use the field-free SOT switching method mentioned in [15] to eliminate the power consumption of the external magnetic field required to achieve the deterministic SOT switching. Therefore, the energy consumption (I^2Rt) of the scaled SOT neuron is around 0.1 fJ. In contrast, through circuit simulation, we can get the area and power consumption of 100-MHz, 4-bit digital CMOS neuron based on 45-nm CMOS technology is 4.7 μm^2 and 2.622 fJ, which indicates that the scaled SOT neurons have great advantages over CMOS counterpart in area and power consumption.

It should be noted that the anomalous Hall effect (AHE) voltage signal is too small to drive the next-level circuit. Therefore, in circuit design, the operational amplifier with high amplification can be used to enhance the AHE voltage signal of the device. In addition, a V–I converter can be used to convert AHE voltage into current to drive the next-level circuit. In conceptual demonstration, we adopt the Hall bar structure, while in practical application, the magnetic tunnel junction (MTJ) structure is widely used to obtain a larger signal.

By exploiting measured characteristics of SOT synapses and SOT neurons, we analyze the training on a three-layer perceptron as shown in Fig. 6(a). The network has 784, 100, and 10 neurons in the input layer, hidden layer, and output layer, respectively [24], [25], [26].

Fig. 6(b) illustrates the SOT synapse array between the input layer and the hidden layer. At each cross point, R_{AH} of the SOT synapse is locally stored as a synaptic weight. In each column, the Hall voltage detection terminals are connected in series, and the summation can be obtained according to Kirchhoff's voltage law. The current signals depending on the training data from the input layer are modulated by the corresponding SOT synapse. These synaptic voltages are summed and converted to a current signal, which is achieved by a V–I converter. Then, the resultant synaptic currents are input to the SOT neuron located at the end of each vertical line.

In order to validate its feasibility, we performed four types of training, using different types of synapses and hidden layer neurons. The first type is based on neurons with software ReLU function and software synapses. The second type is based on SOT neurons and software synapses. The third type is based on SOT neurons and SOT synapses with $100R_{\mathrm{AH}}$ states. The fourth type is based on SOT neurons and SOT synapses with ten R_{AH} states. All four types of output layers employ neurons with softmax activation functions. Using 60 000 training examples and 10 000 testing examples, the recognition rates of the four types are 97.53%, 95.28%, 91.01%, and 83.60%, respectively, as illustrated in Fig. 6(c). It can be seen from the results that SOT synapses with ten R_{AH} states are sufficient to support the training of artificial neural network (ANN).

IV. Conclusion

In this work, at the device level, we experimentally demonstrate the functions of artificial neurons and synapses based on SOT-induced magnetization manipulation in W/CoFeB/MgO heterostructures with two different film thicknesses. In the SOT neuron device, the R_{AH}–I curve is used to construct the ReLU activation function, while in the SOT synapse device, multiple R_{AH} state modulation is realized which can imitate synaptic behaviors. At the system level, an all-spin ANN is constructed based on our proposed SOT neuron and SOT synapse. The high recognition rates proved the feasibility of the scheme. Such all-spin ANN can potentially pave the way for the integration of energy-efficient and high-density building blocks for deep-learning neural systems.

References

[1] J. von Neumann, "First draft of a report on the EDVAC," *IEEE Ann. Hist. Comput.*, vol. 15, no. 4, pp. 27–75, 1993, doi: 10.1109/85.238389.

[2] S. Fukami and H. Ohno, "Perspective: Spintronic synapse for artificial neural network," *J. Appl. Phys.*, vol. 124, no. 15, Oct. 2018, Art. no. 151904, doi: 10.1063/1.5042317.

[3] A. Kurenkov, S. Fukami, and H. Ohno, "Neuromorphic computing with antiferromagnetic spintronics," *J. Appl. Phys.*, vol. 128, no. 1, Jul. 2020, Art. no. 010902, doi: 10.1063/5.0009482.

[4] Q. Shao et al., "Roadmap of spin-orbit torques," *IEEE Trans. Magn.*, vol. 57, no. 7, pp. 1–39, Jul. 2021, doi: 10.1109/TMAG.2021.3078583.

[5] J. Grollier, D. Querlioz, K. Y. Camsari, K. Everschor-Sitte, S. Fukami, and M. D. Stiles, "Neuromorphic spintronics," *Nature Electron.*, vol. 3, no. 7, pp. 360–370, Mar. 2020, doi: 10.1038/s41928-019-0360-9.

[6] J. Zhou et al., "Spin-orbit torque-induced domain nucleation for neuromorphic computing," *Adv. Mater.*, vol. 33, no. 36, Sep. 2021, Art. no. 2103672, doi: 10.1002/adma.202103672.

[7] S. Zhang et al., "A spin-orbit-torque memristive device," *Adv. Electron. Mater.*, vol. 5, no. 4, Apr. 2019, Art. no. 1800782, doi: 10.1002/aelm.201800782.

[8] A. W. Stephan and S. J. Koester, "Spin Hall MTJ devices for advanced neuromorphic functions," *IEEE Trans. Electron Devices*, vol. 67, no. 2, pp. 487–492, Feb. 2020, doi: 10.1109/TED.2019.2959732.

[9] M. M. Lau and K. H. Lim, "Review of adaptive activation function in deep neural network," in *Proc. IEEE-EMBS Conf. Biomed. Eng. Sci. (IECBES)*, Sarawak, Malaysia, Dec. 2018, pp. 686–690, doi: 10.1109/IECBES.2018.8626714.

[10] E. Raimondo et al., "Reliability of neural networks based on spintronic neurons," *IEEE Magn. Lett.*, vol. 12, pp. 1–5, 2021, doi: 10.1109/LMAG.2021.3100317.

[11] A. W. Stephan and S. J. Koester, "Convolutional neural networks utilizing multifunctional spin-Hall MTJ neurons," 2019, *arXiv:1905.03812*.

[12] J. Cai et al., "Sparse neuromorphic computing based on spin-torque diodes," *Appl. Phys. Lett.*, vol. 114, no. 19, May 2019, Art. no. 192402, doi: 10.1063/1.5090566.

[13] X. Wang, Y. Chen, H. Xi, H. Li, and D. Dimitrov, "Spintronic memristor through spin-torque-induced magnetization motion," *IEEE Electron Device Lett.*, vol. 30, no. 3, pp. 294–297, Mar. 2009, doi: 10.1109/LED.2008.2012270.

[14] L. Liu, C.-F. Pai, Y. Li, H. W. Tseng, D. C. Ralph, and R. A. Buhrman, "Spin-torque switching with the giant spin Hall effect of tantalum," *Science*, vol. 336, no. 6081, pp. 555–558, May 2012, doi: 10.1126/science.1218197.

[15] I. M. Miron et al., "Perpendicular switching of a single ferromagnetic layer induced by in-plane current injection," *Nature*, vol. 476, no. 7359, pp. 189–193, Aug. 2011, doi: 10.1038/nature10309.

[16] Q. Zhang et al., "Perpendicular magnetization switching driven by spin-orbit torque for artificial synapses in epitaxial Pt-based multilayers," *Adv. Electron. Mater.*, vol. 8, no. 12, Dec. 2022, Art. no. 2200845, doi: 10.1002/aelm.202200845.

[17] W. A. Borders et al., "Analogue spin-orbit torque device for artificial-neural-network-based associative memory operation," *Appl. Phys. Exp.*, vol. 10, no. 1, Jan. 2017, Art. no. 013007, doi: 10.7567/APEX.10.013007.

[18] T. Shibata et al., "Linear and symmetric conductance response of magnetic domain wall type spin-memristor for analog neuromorphic computing," *Appl. Phys. Exp.*, vol. 13, no. 4, Apr. 2020, Art. no. 043004, doi: 10.35848/1882-0786/ab7e07.

[19] W. A. Borders, S. Fukami, and H. Ohno, "Characterization of spin-orbit torque-controlled synapse device for artificial neural network applications," *Jpn. J. Appl. Phys.*, vol. 57, no. 10, Oct. 2018, Art. no. 1002B2, doi: 10.7567/JJAP.57.1002B2.

[20] X. Sun and S. Yu, "Impact of non-ideal characteristics of resistive synaptic devices on implementing convolutional neural networks," *IEEE J. Emerg. Sel. Topics Circuits Syst.*, vol. 9, no. 3, pp. 570–579, Sep. 2019, doi: 10.1109/JETCAS.2019.2933148.

[21] P.-Y. Chen, X. Peng, and S. Yu, "NeuroSim: A circuit-level macro model for benchmarking neuro-inspired architectures in online learning," *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst.*, vol. 37, no. 12, pp. 3067–3080, Dec. 2018, doi: 10.1109/TCAD.2018.2789723.

[22] P.-Y. Chen et al., "Mitigating effects of non-ideal synaptic device characteristics for on-chip learning," in *Proc. IEEE/ACM Int. Conf. Comput.-Aided Design (ICCAD)*, Austin, TX, USA, Nov. 2015, pp. 194–199, doi: 10.1109/ICCAD.2015.7372570.

[23] A. Sengupta, B. Han, and K. Roy, "Toward a spintronic deep learning spiking neural processor," in *Proc. IEEE Biomed. Circuits Syst. Conf. (BioCAS)*, Shanghai, China, Oct. 2016, pp. 544–547, doi: 10.1109/BioCAS.2016.7833852.

[24] Z. Wang et al., "Fully memristive neural networks for pattern classification with unsupervised learning," *Nature Electron.*, vol. 1, no. 2, pp. 137–145, Feb. 2018, doi: 10.1038/s41928-018-0023-2.

[25] A. Kurenkov, S. DuttaGupta, C. Zhang, S. Fukami, Y. Horio, and H. Ohno, "Artificial neuron and synapse realized in an antiferromagnet/ferromagnet heterostructure using dynamics of spin-orbit torque switching," *Adv. Mater.*, vol. 31, no. 23, Jun. 2019, Art. no. 1900636, doi: 10.1002/adma.201900636.

[26] R. Li et al., "In-memory mathematical operations with spin-orbit torque devices," *Adv. Sci.*, vol. 9, no. 25, Sep. 2022, Art. no. 2202478, doi: 10.1002/advs.202202478.