Global-Gate Controlled One-Transistor One-Digital-Memristor Structure for Low-Bit Neural Network

Mingqiang Huang, Guangchao Zhao, Xingli Wang, Wei Zhang, Philippe Coquet, Beng Kang Tay, *Senior Member, IEEE*, Gaokuo Zhong, and Jiangyu Li

***Abstract*— Memristor based neuromorphic computing system has recently attracted enormous attention due to its fast and energy-efficient matrix vector multiplication, thus providing a novel approach to implement neural networks for artificial intelligence. However, the widely studied analogue memristors exhibit major flaws in terms of high conductance variation and nonlinear/asymmetric characteristics. In this work, we develop global gate controlled one transistor one digital memristor (1T1DM) architecture as the basic binary electronic synapse. Inspired by the current research highlights about low-bit networks, we further implement low-bit neuromorphic computing onto our 1T1DM systems by simulation. Compared to the classical analogue type of memristor with one transistor one analogue memristor (1T1R) structure, our 1T1DM network is light-weighted, highly robust and can work well on challenging visual tasks. Besides, benefiting from the global gated device structure, the on-state conductance of the digital memristors in the network can be simultaneously modulated by the controlling gate, offering possibility to tune the power consumption and operation speed while will not increase the circuit complexity.**

***Index Terms*— Memristor, low-bit neural network, low power.**

I. Introduction

ANALOGUE memristor with multi-level conductance has been considered as a promising and novel computation element for energy-efficient neuromorphic computing due to its innate properties of nonvolatility, reconfigurability and conductance updating characteristics [1]–[4]. However, there exists major flaw in the hardware implementation due to the high conductance variation. For instance, as demonstrated in reference [1], the convolutional neural network (CNN) theoretically shows more than 98% recognition accuracy on Modified National Institute of Standards and Technology (MNIST) test in software. Once the weight parameters had been transferred to the memristor based hardware networks, the practical accuracy was degraded to be only 79.76% due to conductance variation.

Recently, it has been demonstrated that the low-bit (1-2 bit) network is also able to achieve similar performance compared to the full precision (32-bit precision) network, wherein it costs much less storage space and computing resource. Particularly, 1-bit neural network has attracted enormous attention because of its extreme energy-efficiency and light-weighted architecture [5]–[10]. Several previous works [11]–[15] have shown that the digital memristor with bistable conductance states (usually known as a nonvolatile memory) seems to be promising for the 1-bit neural networks, in which SRAM (Static Random-Access Memory) and 1S1R (1-selector 1-memristor) are usually used [16]–[22]. However, SRAM consumes large device and circuit area, while the 1S1R structure has sacrificed its reconfigurability with fixed on-state and off-state conductance [12], [13], [18].

In this work, we demonstrated global gate controlled 1T1DM device as the binary electronic synapse for the neuromorphic computing tasks. Compared to the classical three-terminal 1T1R device, it has largely simplified the circuit complexity, and is also likely to suppress the conductance variation issue in circuit design, thus providing higher performance robustness. Furthermore, the 1T1DM device can be globally controlled by a back gate, providing much more reconfigurable features (such as its tunable maximum conductance or power consumption) while not increase the circuit complexity.

II. 1T1DM Device Structure

The key point for 1T1DM device is to connect one bistable digital type memristor in-series with one transistor together. Fig. 1a shows the cross section schematic view of the device demonstrated in this work. The fabrication process can be described as following: Few-layer black phosphorus (BP) flakes were firstly mechanically exfoliated onto the HfO_2/Si

Manuscript received October 12, 2020; revised October 31, 2020; accepted November 4, 2020. Date of publication November 10, 2020; date of current version December 24, 2020. This work was supported in part by the Guangdong Basic and Applied Basic Research Foundation under Grant 2019A1515111142; in part by the Shenzhen Science and Technology Innovation Committee under Grant KQTD20170810160424889 and Grant JCYJ20170818163902553; in part by the National Natural Science Foundation of China under Grant 51902337; and in part by the Ministry of Education, Singapore, under Grant MOE2019 T1-001-113. The review of this letter was arranged by Editor B. Govoreanu. *(Mingqiang Huang and Guangchao Zhao contributed equally to this work.) (Corresponding author: Mingqiang Huang.)*

Mingqiang Huang, Wei Zhang, Gaokuo Zhong, and Jiangyu Li are with the Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen 518055, China (e-mail: mq.huang2@siat.ac.cn).

Guangchao Zhao, Xingli Wang, Philippe Coquet, and Beng Kang Tay are with CNRS-International-NTU-THALES-Research-Alliance, Nanyang Technological University, Singapore 639798.

Color versions of one or more of the figures in this letter are available online at https://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/LED.2020.3037203

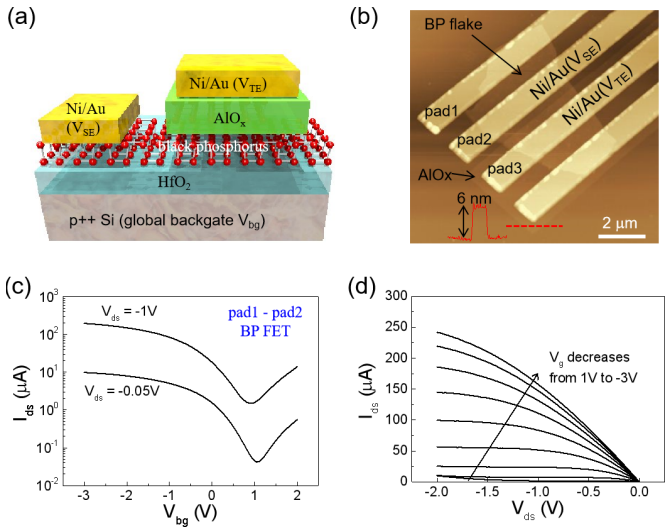

Fig. 1. (a) Schematic view of the 1T1DM device. (b) AFM morphology of the device. Pad 1 and pad 2 are directly contacted with BP flake, pad 3 and pad 4 are on top of AlO_x. (c) I_{ds}-V_{bg} curve of the individual BP FET (pad 1 – pad 2). (d) I_{ds}-V_{ds} curve of the BP FET.

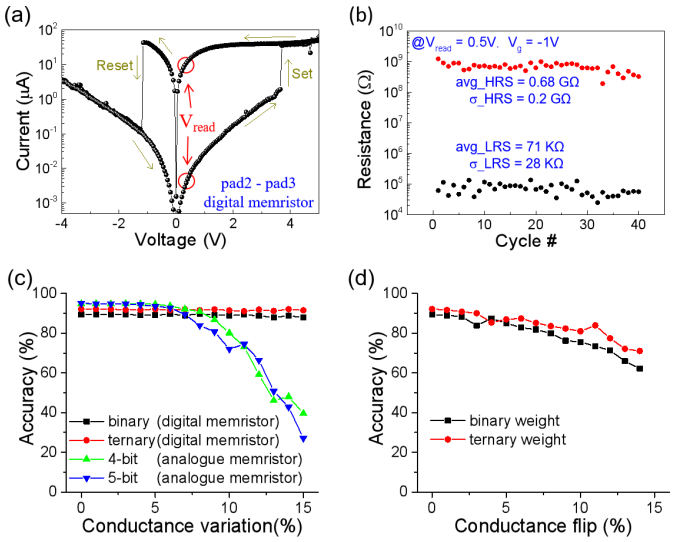

Fig. 2. (a) and (b) Measured HRS-LRS transition of the memristor. (c) Robustness simulation for the binary, ternary, 4-bits and 5-bits networks with conductance variation. (d) Robustness simulation of BWN and TWN with conductance flip (failure in Set/Reset process).

substrate, in which the 30nm-thick HfO_2 layer was deposited by atomic layer deposition (ALD) at 300 °C on silicon wafer. Then e-beam lithography was used to define the AlO_x region and 2 nm Al seeding layer was deposited by e-beam evaporation sequentially. Next, thin Al film would be spontaneously oxidized to AlO_x (5 nm – 6 nm) during the fabrication process. Finally, the Ni/Au side electrode and top electrode were defined and deposited using standard nano-fabrication process.

Fig. 1b shows AFM morphology image of the device, in which the left two metal pads are directly contacted with BP flake, and the right two pads are on top of the AlO_x. In our design, pad1 - pad2 forms the typical FET, pad2 - pad3 forms the memristor, pad2 - pad3 - backgate constitutes the developed 1T1DM device. Fig. 1c and 1d plots the I_{ds}-V_{bg} and I_{ds}-V_{ds} curves of the individual BP transistors (pad1 - pad2), respectively. The moderate on/off ratio (about 3 order) and high output current as well as the linear Ohmic Contact indicates the good quality of semiconducting BP [23], [24].

III. Robust Digital Memristor Network

Pad2 - pad3 forms a typical AlO_x memristor. Similar to other oxide based memristors, the AlO_x device here also needs an electrical forming process [25], [26], and the forming voltage of our device was about |VSE-VTE| = 3V. After the electrical forming, standard DC current-voltage curve was measured and plotted in fig. 2a, where a clear resistive switching phenomenon between high resistance state (HRS) and low resistance state (LRS) can be seen. The switching mechanism is mainly attributed to the oxygen vacancies in Al_2O_3 layer, which is controlled by varying the applied voltage [25].

Statistically, we have fabricated tens of devices with the same structure, and 73% of them show similar and good performance. Generally, the set voltage of the device ranges from 2V to 5V, and the reset voltage ranges from −1V to −2V. The HRS to LRS ratio can reach as large as 104 at a small read voltage of 0.5V. Such bistable property offers great potential for low bit neural network, in which the weights are represented by the memristor conductance [13]–[15].

A fair comparison between 1T1R and 1T1DM base network has been then conducted by using a two-layer (784-100-10) network for the MNIST recognition test. All the weight parameters were firstly trained in a standard full-precision neural network in software, and the recognition accuracy was about 95%. Then the fix-point quantization method was performed on the weight parameters.

For the 1-bit network, the weights were binarized (threshold is 0) to be either 1 or -1, corresponding to the LRS and HRS of our digital memristor.

For the ternary network, we need twice the number of 1T1DM arrays in a differential circuit (represented as G_{+ij} and G_{-ij}) to map the conductance to −1, 0 and 1, which was also commonly used in the analogue memristor networks [2]–[4], [27], [28]. Then the parameters were ternarized using hard threshold method (threshold was chosen to be ±0.17 by our calculation).

For the analogue memristor based multi-bit neural network, linear segmentation method was used to get the 4-bit (16-states) and the 5-bit (32-states) parameters matrix [4]. After the segmentation, the recognition accuracy was 89.34%, 92.14, 94.87% and 95.09% for binary (1-bit), ternary (2-bits), 4-bits and 5-bits network, respectively.

Next, different levels of Gaussian noise, namely the conductance variation data (corresponding to the standard deviation in fig.2b), was induced into the network to represent the variation of weight parameters. Fig. 2c plots the simulation results at different variations. It is clear to see that with the increasing of conductance variation, the accuracy drops. However, the analogue memristor networks had been largely degraded once the variation is large than 10%. Meanwhile, the 1T1DM based binary weighted network (BWN) and the ternary weighted network (TWN) show excellent robustness.

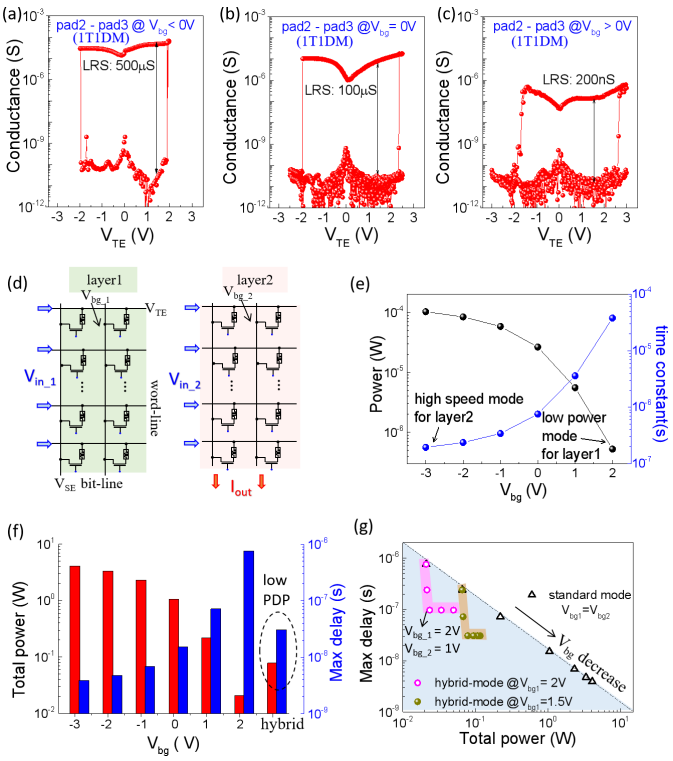

Fig. 3. (a-c) Tunable on/off ratio of the 1T1DM device at different gate voltages. (d) Hybrid mode (with different V_{bg} at different layers). (e) Power consumption and simulated time delay of individual 1T1DM device at different backgate voltages V_{bg}. (f) Power and delay of the system at different V_{bg}. (g) Hybrid operation mode for low power and high speed.

And further simulation has shown that the binary or ternary network can still reserve a high accuracy about 80% when the variation is up to 40%.

Furthermore, the performance of BWN and TWN are evaluated for even severer conditions when a certain number of memristors are reversed to their opposite conductance state (conductance flip: LRS↔HRS) due to the failure in set or reset process. For BWN, the weights were changed as 1→-1, or -1→1. For TWN, in consideration of the differential circuits mapping method, the weight were changed as 0→ ±1, 1→0, or -1→0. As shown in fig. 2d, if 3% of the total memristors failed in Set/Reset process, the accuracy can almost remain the same. Once 10% were failed, the accuracy can still maintain about 75% and 80% for BWN and TWN, respectively. Increasing the write/erase voltage meanwhile reducing the read voltage can be an efficient method to avoid such conductance flip [11].

IV. Tunable Power Consumption and Speed

Now we will focus on the unique and tunable DC properties of our device. The operating current of 1T1DM flowing from the side electrode V_{SE} to the top electrode V_{TE} will be modulated by both the FET (V_{bg}) and the memristor (V_{TE}). Under certain V_{bg}, the device shows bistable states (LRS and HRS) dominated by the memristor. Once the V_{bg} is changed, the device conductance will be correspondingly tuned with different compliance current constrained by the BP transistor.

Fig. 3a, 3b and 3c present the typical device behavior under different gate voltages V_{bg}. When V_{bg} is negative (fig. 3a), the FET is in its "on" state, then the 1T1DM structure will show a intrinsic LRS/HRS conductance ratio about 5×10^6. When V_{bg} is positive (fig. 3c), the FET is in "off" state, then the current will be largely suppressed, and the LRS/HRS conductance ratio will also be decreased to about 10^4. Due to the moderate on/off ratio (2-3 order) of the BP FET [23], the device can always show a high LRS/HRS ratio at any backgate voltage modulation, thus providing excellent selectivity (defined as the conductance ratio at V_d and $V_d/2$) and offering great potential in neural network application.

Note that the neuromorphic application shows hierarchy behavior in data flowing from input layer to the output layer, thus offering an opportunity to design different circuit blocks with different operation modes (fig. 3d).

Firstly, we simulated the power consumption and time constant of the individual 1T1DM device at different V_{bg} by using the classical circuits delay $\tau = RC$ model (the capacitance is estimated to be $C = 20$ pF as an empirical value) [29], [30]. As shown in figure 3e, with the increasing of V_{bg}, the device power consumption drops, but the time constant (delay) increases because of the lower current.

Then we extended the simulation onto our binary network. If all the connected synapses, namely the 1T1DM devices, were in high-performance mode with $V_{bg} = -3$ V, the max transient total power is about 4.07 W and the calculated delay is 3.9 ns. If they were in the low-power mode with $V_{bg} = 2$ V, the power consumption can be decreased to only 21 mW, meanwhile, the system delay increases to 760 ns (fig. 3f).

Since the first layer holds much more synapses than that of the second layer, we can develop a hybrid mode to make different layers operate at different V_{bg} to get the best power delay product (PDP). As shown in fig. 3g, the black triangles indicate a clear trade-off relationship between delay and total power at a standard mode with a fixed V_{bg}. The purple dots represent an alternative hybrid mode with $V_{bg\_layer2} = 2$ V and V_{bg_layer1} varies from -1 to 2V. The elbow point ($V_{bg_layer2} = 2$V and $V_{bg\_layer1} = 1$ V) shows the best power delay product about $2.27*10^{-9}$J, which is 7.03 times better than the fixed V_{bg} mode. Similarly, the brown dots represent V_{bg_layer2} is 1.5 V and V_{bg_layer1} varies from -2 to 1.5 V. The elbow point ($V_{bg\_layer2} = 1.5$ V and $V_{bg\_layer1} = 0$ V) shows a power delay product about $2.41*10^{-9}$J, which shows an improvement of 6.62 times.

V. Conclusion

In this work, we proposed a global gate controlled 1T1DM synaptic device for robust low-bit neuromorphic computing. The networks exhibit high performance, high robustness, and tunable power consumption. Compared to the classical three-terminal analogue 1T1R structure, our 1T1DM has greatly simplified the network circuit and avoided the conductance variation issue. Compared to the two-terminal 1S1R for a digital memristor, the 1T1DM provides much more reconfigurable features such as the hybrid mode to get better power delay product, meanwhile, it will not increase the circuit complexity. This work opens the possibility for digital memristor based re-configurable computation system and can be extended in ubiquitous applications in the future.

REFERENCES

[1] M. Prezioso, F. Merrikh-Bayat, B. D. Hoskins, G. C. Adam, K. K. Likharev, and D. B. Strukov, "Training and operation of an integrated neuromorphic network based on metal-oxide memristors," *Nature*, vol. 521, no. 7550, pp. 61–64, May 2015, doi: 10.1038/nature14441.

[2] P. Yao, H. Wu, B. Gao, S. B. Eryilmaz, X. Huang, W. Zhang, Q. Zhang, N. Deng, L. Shi, H.-S.-P. Wong, and H. Qian, "Face classification using electronic synapses," *Nature Commun.*, vol. 8, no. 1, pp. 1–8, Aug. 2017, doi: 10.1038/ncomms15199.

[3] C. Li, D. Belkin, Y. Li, P. Yan, M. Hu, N. Ge, H. Jiang, E. Montgomery, P. Lin, Z. Wang, W. Song, J. P. Strachan, M. Barnell, Q. Wu, R. S. Williams, J. J. Yang, and Q. Xia, "Efficient and self-adaptive *in-situ* learning in multilayer memristor neural networks," *Nature Commun.*, vol. 9, no. 1, pp. 1–8, Dec. 2018, doi: 10.1038/s41467-018-04484-2.

[4] P. Yao, H. Wu, B. Gao, J. Tang, Q. Zhang, W. Zhang, and H. Qian, "Fully hardware-implemented memristor convolutional neural network," *Nature*, vol. 577, no. 7792, pp. 641–646, 2020, doi: 10.1038/.s41586-020-1942-4.

[5] I. Hubara, M. Courbariaux, D. Soudry, R. El-Yaniv, and Y. Bengio, "Binarized neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 4107–4115.

[6] F. Li, B. Zhang, and B. Liu, "Ternary weight networks," 2016, *arXiv:1605.04711*. [Online]. Available: http://arxiv.org/abs/1605.04711

[7] C. Leng, H. Li, S. Zhu, and R. Jin, "Extremely low bit neural network: Squeeze the last bit out with ADMM," 2017, *arXiv:1707.09870*. [Online]. Available: http://arxiv.org/abs/1707.09870

[8] P. Wang, Q. Hu, Y. Zhang, C. Zhang, Y. Liu, and J. Cheng, "Two-step quantization for low-bit neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4376–4384.

[9] M. Rastegari, V. Ordonez, J. Redmon, and A. Farhadi, "XNOR-net: ImageNet classification using binary convolutional neural networks," in *Computer Vision—ECCV* (Lecture Notes in Computer Science), vol. 9908, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds. Cham, Switzerland: Springer, 2016. [Online]. Available: doi: 10.1007/978-3-319-46493-0_32.

[10] H.-Y. Chen, S. Brivio, C.-C. Chang, J. Frascaroli, T.-H. Hou, B. Hudec, M. Liu, H. Lv, G. Molas, J. Sohn, S. Spiga, V. M. Teja, E. Vianello, and H.-S.-P. Wong, "Resistive random access memory (RRAM) technology: From material, device, selector, 3D integration to bottom-up fabrication," *J. Electroceram.*, vol. 39, nos. 1–4, pp. 21–38, Dec. 2017, doi: 10.1007/s10832-017-0095-9.

[11] B. Gao, J. Kang, Z. Zhou, Z. Chen, P. Huang, L. Liu, and X. Liu, "Metal oxide resistive random access memory based synaptic devices for brain-inspired computing," *Jpn. J. Appl. Phys.*, vol. 55, no. 4S, Apr. 2016, Art. no. 04EA06.

[12] S. Truong, S.-J. Ham, and K.-S. Min, "Neuromorphic crossbar circuit with nanoscale filamentary-switching binary memristors for speech recognition," *Nanosc. Res. Lett.*, vol. 9, no. 1, p. 629, 2014, doi: 10.1186/1556-276X-9-629.

[13] T. Tang, L. Xia, B. Li, Y. Wang, and H. Yang, "Binary convolutional neural network on RRAM," in *Proc. 22nd Asia South Pacific Design Autom. Conf. (ASP-DAC)*, Jan. 2017, pp. 782–787, doi: 10.1109/ASPDAC.2017.7858419.

[14] S. Yin, X. Sun, S. Yu, and J.-S. Seo, "High-throughput in-memory computing for binary deep neural networks with monolithically integrated RRAM and 90nm CMOS," 2019, *arXiv:1909.07514*. [Online]. Available: http://arxiv.org/abs/1909.07514

[15] X. Sun, S. Yin, X. Peng, R. Liu, J.-S. Seo, and S. Yu, "XNOR-RRAM: A scalable and parallel resistive synaptic architecture for binary neural networks," in *Proc. Design, Autom. Test Eur. Conf. Exhib. (DATE)*, Mar. 2018, pp. 1423–1428, doi: 10.23919/DATE.2018.8342235.

[16] X. Sun, X. Peng, P.-Y. Chen, R. Liu, J.-S. Seo, and S. Yu, "Fully parallel RRAM synaptic array for implementing binary neural network with (+1, –1) weights and (+1, 0) neurons," in *Proc. 23rd Asia South Pacific Design Autom. Conf. (ASP-DAC)*, Jan. 2018, pp. 574–579, doi: 10.1109/ASPDAC.2018.8297384.

[17] X. Sun, R. Liu, X. Peng, and S. Yu, "Computing-in-Memory with SRAM and RRAM for binary neural networks," in *Proc. 14th IEEE Int. Conf. Solid-State Integr. Circuit Technol. (ICSICT)*, Oct. 2018, pp. 1–4, doi: 10.1109/ICSICT.2018.8565811.

[18] A. Ascoli, R. Tetzlaff, and L. O. Chua, "The first ever real bistable Memristors—Part II: Design and analysis of a local fading memory system," *IEEE Trans. Circuits Syst. II, Exp. Briefs*, vol. 63, no. 12, pp. 1096–1100, Dec. 2016, doi: 10.1109/TCSII.2016.2613560.

[19] S. Fukami, C. Zhang, S. DuttaGupta, A. Kurenkov, and H. Ohno, "Magnetization switching by spin–orbit torque in an antiferromagnet–ferromagnet bilayer system," *Nature Mater.*, vol. 15, no. 5, pp. 535–541, May 2016, doi: 10.1038/nmat4566.

[20] S. G. Hu, Y. Liu, Z. Liu, T. P. Chen, J. J. Wang, Q. Yu, L. J. Deng, Y. Yin, and S. Hosaka, "Associative memory realized by a reconfigurable memristive hopfield neural network," *Nature Commun.*, vol. 6, no. 1, pp. 1–8, Nov. 2015, doi: 10.1038/ncomms8522.

[21] X. Hong, D. J. Loy, P. A. Dananjaya, F. Tan, C. Ng, and W. Lew, "Oxide-based RRAM materials for neuromorphic computing," *J. Mater. Sci.*, vol. 53, no. 12, pp. 8720–8746, Jun. 2018, doi: 10.1007/s10853-018-2134-6.

[22] D. Garbin, O. Bichler, E. Vianello, Q. Rafhay, C. Gamrat, L. Perniola, G. Ghibaudo, and B. DeSalvo, "Variability-tolerant convolutional neural network for pattern recognition applications based on OxRAM synapses," in *IEDM Tech. Dig.*, Dec. 2014, pp. 4–28, doi: 10.1109/IEDM.2014.7047126.

[23] L. Li, Y. Yu, G. J. Ye, Q. Ge, X. Ou, H. Wu, D. Feng, X. H. Chen, and Y. Zhang, "Black phosphorus field-effect transistors," *Nature Nanotechnol.*, vol. 9, no. 5, pp. 372–377, May 2014, doi: 10.1038/nnano.2014.35.

[24] M. Huang, S. Li, Z. Zhang, X. Xiong, X. Li, and Y. Q. Wu, "Multifunctional high-performance van der Waals heterostructures," *Nature Nanotech.*, vol. 12, no. 12, pp. 1148–1154, Oct. 2017, doi: 10.1038/nnano.2017.208.

[25] W. Sun, B. Gao, M. Chi, Q. Xia, J. J. Yang, H. Qian, and H. Wu, "Understanding memristive switching via *in situ* characterization and device modeling," *Nature Commun.*, vol. 10, no. 1, pp. 1–13, Dec. 2019, doi: 10.1038/s41467-019-11411-6.

[26] X. D. Huang, Y. Li, H. Y. Li, Y. F. Lu, K. H. Xue, and X. S. Miao, "Enhancement of DC/AC resistive switching performance in AlOx memristor by two-technique bilayer approach," *Appl. Phys. Lett.*, vol. 116, no. 17, Apr. 2020, Art. no. 173504, doi: 10.1063/5.0006850.

[27] Y. Zhang, X. Wang, and E. G. Friedman, "Memristor-based circuit design for multilayer neural networks," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 65, no. 2, pp. 677–686, Feb. 2018, doi: 10.1109/TCSI.2017.2729787.

[28] G. Zhong, M. Zi, C. Ren, Q. Xiao, M. Tang, L. Wei, F. An, S. Xie, J. Wang, X. Zhong, M. Huang, and J. Li, "Flexible electronic synapse enabled by ferroelectric field effect transistor for robust neuromorphic computing," *Appl. Phys. Lett.*, vol. 117, no. 9, Aug. 2020, Art. no. 092903, doi: 10.1063/5.0013638.

[29] M. Nazmus Sakib, R. Hassan, S. N. Biswas, and S. R. Das, "Memristor-based high-speed memory cell with stable successive read operation," *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst.*, vol. 37, no. 5, pp. 1037–1049, May 2018, doi: 10.1109/TCAD.2017.2729464.

[30] C. A. Desoer, *Basic Circuit Theory*. New York, NY, USA: McGraw-Hill, 2010.