Received 26 August 2021; accepted 3 September 2021. Date of publication 7 September 2021; date of current version 13 December 2021.
The review of this article was arranged by Editor C. Yang.

Digital Object Identifier 10.1109/JEDS.2021.3110877

Nonlinear Weight Quantification for Mitigating Stress Induced Disturb Effect on Multilevel RRAM-Based Neural Network Accelerator

LINDONG WU, ZHIZHEN YU, YABO QIN, QINGYU CHEN, YIMAO CAI (Member, IEEE), AND RU HUANG (Fellow, IEEE)

Institute of Microelectronics, Peking University, Beijing 100871, China

CORRESPONDING AUTHOR: Y. CAI (e-mail: caiyimao@pku.edu.cn)

This work was supported in part by the National Key Research and Development Project of China under Grant 2018YFB1107701 and Grant 2019YFB2205401; in part by the National Natural Science Foundation of China under Grant 61834001, Grant 62025401, Grant 61904003, and Grant 61421005; in part by the 111 Project under Grant B18001; and in part by the Beijing Academy of Artificial Intelligence (BAAI).

ABSTRACT The RRAM-based array is one of the most promising core functional primitives to accelerate the inference process of neural networks. However, the stress-induced disturbance can cause a significant accuracy drop during inference process where input vectors with different voltage levels are fed to the device. This kind of disturb can hardly be avoided by optimizing the fabrication process. Here, we investigate this phenomenon based on TaO_x-based devices with different electrodes. The results indicate that the stress-induced disturb mainly appears in the intermediate resistance states when the voltage is applied on the device. The simulation result by COMSOL reveals the relationship between read disturb and electric field. Therefore, we propose a nonlinear weight quantification method to mitigate read disturb effect on inference accuracy by reducing the number of devices in intermediate resistance states. The simulation results based on the fully-connected networks for MNIST recognition indicate that stress disturb phenomenon can be well suppressed by nonlinear weight quantification compared with the conventional linear quantification method, which will advance the application of the RRAM-based accelerator.

INDEX TERMS Read disturb, multilevel RRAM, nonlinear weight quantification, neural network accelerator.

I. INTRODUCTION

Nowadays, resistive random access memory (RRAM) is considered as one of the most promising candidates for the construction of neuromorphic computing system to break the bottleneck of Von-Neumann architecture [1]–[8]. Especially, the TaO_x-based RRAM exhibits great electrical characteristic, including excellent endurance, high uniformity, and good linearity, and thus has drawn lots of research attention [9]–[12]. Moreover, the multilevel RRAM-based array has been adopted to accelerate inference process of deep neural networks (DNN) and behaves well in various tasks, such as image classification and object detection [13]–[15].

However, it is inevitable that read disturb phenomenon reflected in the resistance drift will happen during inference process when a read voltage is applied on the device [16]–[18]. Read disturb will induce inference accuracy drop. The RRAM-based array serves as the neural network accelerator by speeding up the matrix vector multiplication [19]–[20]. The devices are used to store the weights. Therefore, the resistance drift of the devices will have an influence on the stored weights. As a result, the summated current along the column will be affected. What is worse, it cannot be solved by optimizing the fabrication process, which will greatly restrict the application of RRAM-based accelerator. To handle this dilemma, [16] proposed a bipolar read scheme during inference process. The simulation results indicate that this scheme can enhance the resilience against the read disturb. However, this method will obviously increase the complexity of the peripheral circuit, which is unfavorable for the application.

In this paper, we extend our previous work presented in [21], where we investigate the read disturb effect on resistance drift of multilevel RRAM with a structure of TiN/TaO_x/Pt and proposed a nonlinear weight quantification (NWQ) method to reduce the number of devices in intermediate-resistance states (IRS) to mitigate the read disturb effect on inference accuracy of DNN. However, to speed up the resistance drift phenomenon, high read voltages with the width of 2 s (V_rs) are applied on the devices when they are in different resistance states, which is a little unconvincing because the read voltage adopted during inference process is often small (e.g., 0.1 V or 0.2 V). Therefore, additional measurement experiments are carried out to explore the read disturb phenomenon in this paper. In these experiments, small V_rs (0.1 V, 0.2 V and 0.3 V) with the width of 100 s are applied on the devices of TiN/TaO_x/Pt. Moreover, to demonstrate the universality of read disturb, we further investigate an alternative device with the structure of TiN/TaO_x/TaN. The measurement results indicate that read disturb has a greater impact on devices in IRSs, which is consistent with [21]. In order to better reveal this mechanism, a simulation based on COMSOL is carried out in this paper. The results show that the electric field related to gap distance between fractured filaments plays an important role in the appearance of read disturb. Finally, fully-connected neural networks with different layers are constructed to demonstrate the effect of NWQ method on mitigating read disturb phenomenon. The simulation results indicate that NWQ can effectively reduce the number of devices in IRSs. Compared with traditional linear weight quantification (LWQ) method, a higher accuracy can be achieved by adopting the proposed NWQ mapping strategy when read disturb is considered, which will greatly broaden the application of RRAM-based neural network accelerator.

II. EXPERIMENTS

To investigate the read disturb on multilevel RRAM, two kinds of devices were fabricated, including TiN/TaO_x/Pt and TiN/TaO_x/TaN. For the TiN/TaO_x/Pt device, Pt bottom electrode (BE) was deposited on silicon substrate and a 40-nm TaO_x film was deposited by reactive sputtering. Finally, TiN top electrode (TE) was deposited and patterned. The electrical measurement was conducted with an Agilent B1500A semiconductor analyzer and the voltage was always applied on the TE with the BE grounded. For the TiN/TaO_x/TaN device, the BE (TaN) was deposited on silicon substrate firstly. The TaO_x film of 30 nm was deposited by reactive sputtering. Finally, the TE (TiN) was deposited and patterned. The electrical measurement was conducted when the voltage was always applied on the BE with the TE grounded.

III. RESULTS AND DISCUSSION

Fig. 1(a) shows the typical I-V curve of TiN/TaO_x/Pt device. The set process can be observed when the positive voltage reaches 0.6 V with a compliance current (I_{cc}) of 500 μA. The

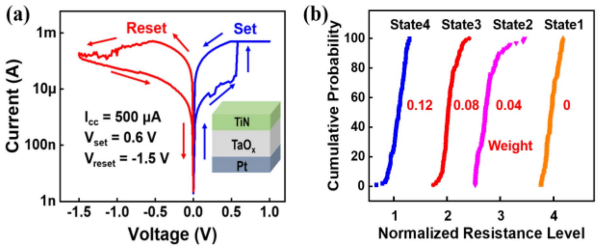

FIGURE 1. (a) The typical I-V curve of the device of TiN/TaO_x/Pt. (b) The cumulative probability distribution of four states obtained by varying reset voltages.

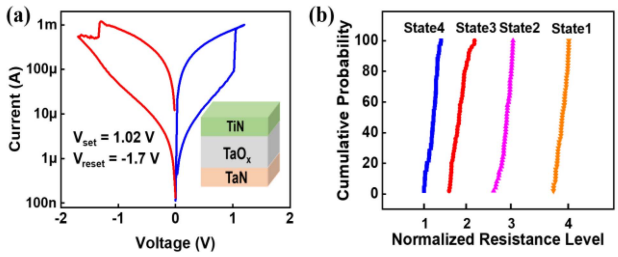

FIGURE 2. (a) The typical I-V curve of the device of TiN/TaO_x/TaN. (b) The cumulative probability distribution of four states obtained by varying reset voltages.

reset process can be accomplished by applying a negative voltage of −1.5 V. The transition between high resistance state (HRS) and low resistance state (LRS) is associated with formation and fracture of the filament. Moreover, four states can be obtained by varying V_{reset} (−1 V, −1.5 V and −2 V for State 3, State 2 and State 1 respectively) while the set voltage (V_{set}) and I_{cc} were kept unchanged, as indicated in Fig. 1(b). The same operation is carried out on the device of TiN/TaO_x/TaN. Fig. 2(a) shows the I-V curve of the device. As can be seen, V_{set} is about 1.02 V and V_{reset} is −1.7 V. The I_{cc} is not needed. By varying V_{reset}s, four states can be achieved as shown in Fig. 2(b). The resistances of the device in State 1, 2, 3 and 4 are about 300 kΩ, 100 kΩ, 50 kΩ and 33 kΩ respectively.

Read disturb will emerge when the multilevel RRAM-based inference accelerator works for a long time, which will affect inference accuracy of the network. Therefore, a comprehensive research is made on read disturb phenomenon based on these two kinds of devices. In these experiments, small V_rs (0.1 V, 0.2 V and 0.3 V) with the width of 100 s are applied on the devices of TiN/TaO_x/Pt and TiN/TaO_x/TaN in different resistance states. The results of resistance drift of four states of TiN/TaO_x/Pt device are shown in Fig. 3. Fig. 3(a) and Fig. 3(d) indicate that read disturb has little effect on the device in HRS or LRS. When the device is in IRS, read disturb has a significant effect on resistance drift, as shown in Fig. 3(b) and Fig. 3(c). As can be seen from Fig. 3(b), the resistance of the device in State 2 will decrease when V_r of 0.1 V is applied on the device. What is worse, the resistance will shift from level 3 to level 2 when V_r of 0.2 V is applied on the device for

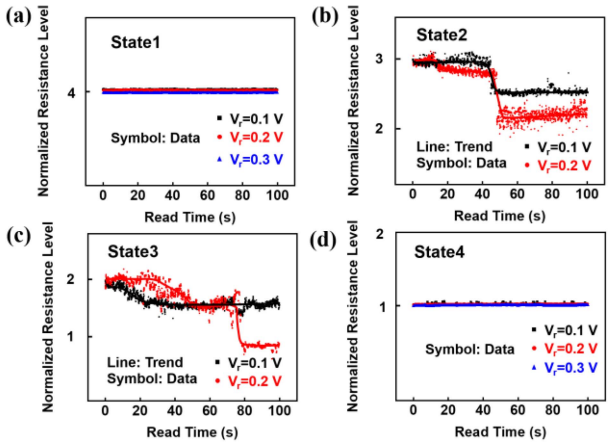

FIGURE 3. Read disturb effect on TiN/TaO_x/Pt devices in State 1 (a), State 2 (b), State 3 (c), and State 4 (d) when different V_rs are applied on the devices for 100 s respectively. Read disturb phenomenon has a significant effect on the devices in State 2 and State 3.

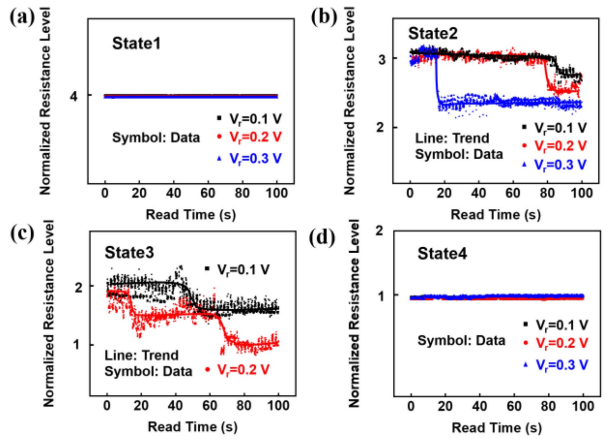

FIGURE 4. Read disturb effect on TiN/TaO_x/TaN devices in State 1 (a), State 2 (b), State 3 (c), and State 4 (d) when different V_rs are applied on the devices for 100 s respectively. Read disturb also has a significant effect on the devices in State 2 and State 3.

50 s. Fig. 3(c) indicates that the resistance of the device in State 3 will change more when the V_r of 0.2 V with the width of 100 s is applied on the device. Similar phenomena happen on the devices of TiN/TaO_x/TaN, as shown in Fig. 4. Read disturb has little effect on the devices in HRS or LRS. However, Fig. 4(b) shows that when V_r of 0.1 V is applied on TiN/TaO_x/TaN device in State 2, the resistance of the device will decrease. When V_r of 0.2 V is applied on the device in State 3, the resistance of the device will shift to level 1, as indicated in Fig. 4(c).

To better reveal the mechanism of read disturb, a simulation by COMSOL is carried out, as shown in Fig. 5. The simulation is constructed based on TiN/TaO_x/Pt device. As can be seen from Fig. 5(a), the fractured filaments can serve as virtual electrodes. Therefore, the electrical potential on the top filament is nearly equal to the voltage applied on TE and the potential on the bottom potential is 0, equal to BE. Here the voltage applied on the device is set as 1 V. Therefore,

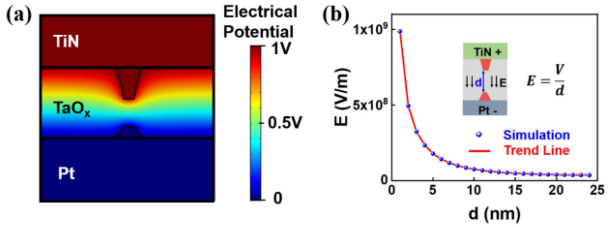

FIGURE 5. (a) The simulation result of TiN/TaO_x/Pt device by COMSOL. The electrical potential of top filament is nearly equal to the voltage applied on the device (1 V). (b) The relationship between the electrical field (*E*) and distance (*d*).

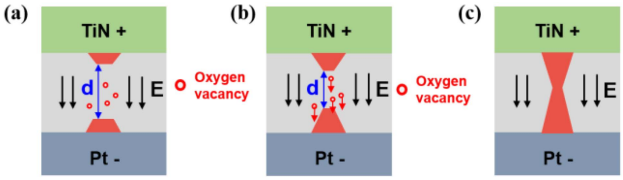

FIGURE 6. The schematic mechanism of read disturb phenomenon on the TiN/TaO_x/Pt device in HRS (a), IRS (b) and LRS (c).

the electric field (E) between fractured filaments can be calculated using (1):

$$E = \frac{V}{d} \tag{1}$$

where V is the applied voltage (1 V) on the device and d is the distance between the fractured filaments. As shown in Fig. 5(b), E decreases with the increase of d. Read disturb is mainly associated with the driving of residual oxygen vacancy by electrical field between fractured filaments. Therefore, when the device is in HRS, d is relatively large so that E is small and cannot drive the oxygen vacancy to move, as shown in Fig. 6(a). However, when the device is in IRS, the distance becomes smaller so that the electric filed is bigger and thus can drive oxygen vacancy to move towards BE to refill the filament, leading to resistance drift (Fig. 6(b)). When the device is in LRS, read disturb has little effect on it because a stable filament is generated, as indicated in Fig. 6(c).

We designed two fully-connected neural networks (FCNNs) with different layers (three and four layers) for MNIST recognition task to investigate the impacts of read disturb. The schematic structure of four-layer FCNN is shown in Fig. 7(a), in which there are 784 input nodes and 10 output nodes. The first hidden layer consists of 100 nodes and the second hidden layer has 50 nodes. For the three-layer FCNN, there are 784 input nodes and 10 output nodes. The hidden layer consists of 100 nodes. Fig. 7(b) indicates that weight values in a well-trained four-layer FCNN are basically in the interval between -0.12 and 0.12. Therefore, we quantify the positive weight into four values, which are 0, 0.04, 0.08, and 0.12. During training process of the network based on LWQ method as shown in Fig. 8(a), the weight values greater than 0.12 are normalized to 0.12, while the weight values between 0.08 and 0.12 are normalized to 0.08.

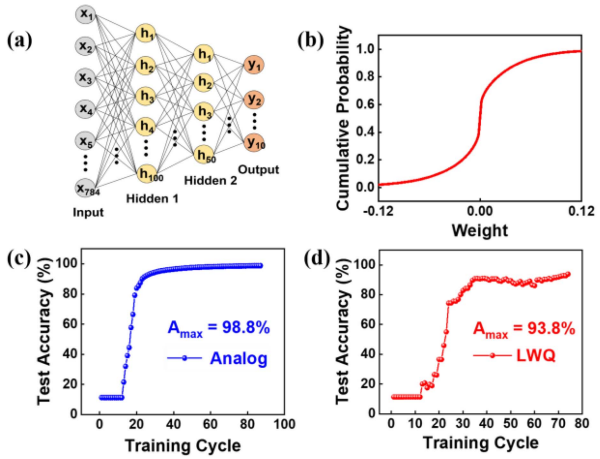

FIGURE 7. **(a) The schematic structure of four-layer fully-connected neural network. (b) The cumulative probability distribution of weights. (c-d) Inference accuracy curves against training cycles of the neural networks based on analog weights and linear weight quantification (LWQ) method.**

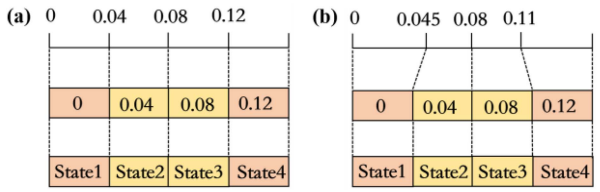

FIGURE 8. **The schematic of linear weight quantification (LWQ) method (a) and nonlinear weight quantification (NWQ) method (b).**

The weight values between 0.04 and 0.08 are normalized to 0.04. The weight values smaller than 0.04 are normalized to 0. The same method is employed on negative weight values. The inference accuracy curves against training cycles of the networks based on analog weights and LWQ method are shown in Fig. 7(c) and Fig. 7(d), respectively. The maximum accuracy (A_{max}) of LWQ-based network is 93.8%, which is tolerable compared with the analog-weight-based network.

The impact of resistance drift caused by read disturb on inference accuracy is investigated, as shown in Fig. 9. The accuracy drops significantly if large resistance drift appears. Resistance drift makes devices transform from IRS to LRS, which induces the increase of current along the column. Therefore, we propose the NWQ method to mitigate the effect on inference accuracy. The schematic of nonlinear quantification method is shown in Fig. 8(b). Unlike LWQ method, during training process, the weight values greater than 0.11 are normalized to 0.12, while weight values between 0.08 and 0.11 are normalized to 0.08. The weight values between 0.045 and 0.08 are normalized to 0.04. Weight values smaller than 0.045 are normalized to 0. The inference accuracy based on NWQ method is shown in Fig. 9(a). The A_{max} of NWQ-based network is 93.4%, which is similar to that of LWQ-based network. Moreover, the number of devices in IRSs is greatly

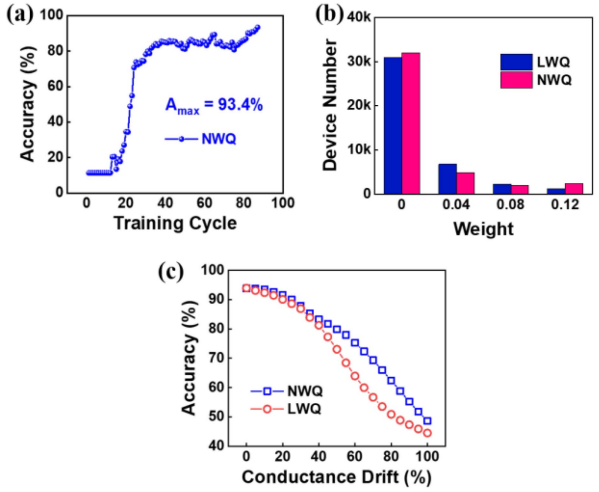

FIGURE 9. **(a) The inference accuracy against training cycles of the NWQ-based network. (b) The histogram of number of devices storing quantized weights of LWQ-based network and NWQ-based network. (c) The inference accuracy of NWQ-based network and LWQ-based network when read disturb is considered.**

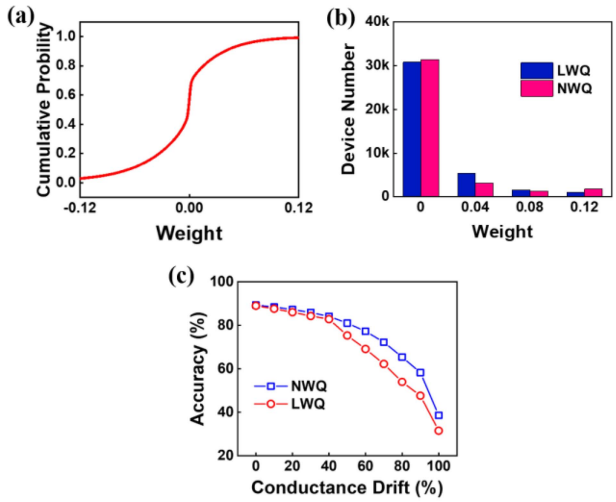

FIGURE 10. **(a) The cumulative probability distribution of weights of three-layer neural network. (b) The histogram of number of devices storing quantized weights of LWQ-based network and NWQ-based network. (c) The inference accuracy of NWQ-based network and LWQ-based network when read disturb is considered.**

reduced compared with LWQ-based network, as shown in Fig. 9(b). Finally, read disturb effect on NWQ-based and LWQ-based FCNNs is studied, as shown in Fig. 9(c). With the number of devices in IRSs suffering from read disturb increasing, the inference accuracy of NWQ-based FCNN and LWQ-based FCNN both decreases. However, the inference accuracy of NWQ-based network is nearly always higher than that of LWQ-based network, due to the decrease of devices in IRSs of LWQ-based network. The simulation results of three-layer FCNN also demonstrate that. As can be seen from Fig. 10(a), the weight values in the well-trained network are also basically in the interval between -0.12 and 0.12. Therefore, the same quantification

method shown in Fig. 8 is adopted. Fig. 10(b) indicates that the number of devices in IRSs of NWQ-based network is greatly decreased compared with that of LWQ-based network. Moreover, as can be seen from Fig. 10(c), the inference accuracy of NWQ-based FCNN is always higher than that of LWQ-based FCNN when read disturb is considered. These results demonstrate that NWQ-based network possesses better resilience than LWQ-based network and NWQ method can effectively mitigate the read disturb effect on multilevel RRAM-based neural network accelerator, which will advance the application of RRAM-based accelerator.

IV. CONCLUSION

In this paper, we investigate the read disturb effect on multilevel RRAM-based neural network accelerator. Two kinds of devices are comprehensively studied. The measurement results indicate that read disturb can induce the resistance drift when the device is in IRS. The simulation result based on COMSOL reveals that the electrical field plays a vital role in the appearance of read disturb phenomenon. When the device is in IRS, the electrical field is so big that the oxygen vacancy can be driven towards BE to refill the filament. Therefore, we propose an NWQ method to reduce the number of devices in IRSs to mitigate read disturb effect on the DNNs. The simulation results indicate that the influence of read disturb can be effectively mitigated by the NWQ method compared with traditional LWQ method, which will give an important guidance to enhance the stability of multilevel RRAM-based accelerator.

REFERENCES

[1] S. H. Jo, T. Chang, I. Ebong, B. B. Bhadviya, P. Mazumder, and W. D. Lu, "Nanoscale memristor device as synapse in neuromorphic systems," *Nano Lett.*, vol. 10, no. 4, pp. 1297–1301, Apr. 2010, doi: 10.1021/nl904092h.

[2] Z. Zhang *et al.*, "Memory materials and devices: From concept to application," *InfoMat*, vol. 2, no. 2, pp. 261–290, 2020. [Online]. Available: https://doi.org/10.1002/inf2.12077

[3] Q. Xia and J. J. Yang, "Memristive crossbar arrays for brain-inspired computing," *Nat. Mater.*, vol. 18, no. 4, pp. 309–323, Apr. 2019, doi: 10.1038/s41563-019-0291-x.

[4] L. Wu *et al.*, "Emulation of biphasic plasticity in retinal electrical synapses for light-adaptive pattern pre-processing," *Nanoscale*, vol. 13, no. 6, pp. 3483–3492, Feb. 2021, doi: 10.1039/d0nr08012h.

[5] J. Tang *et al.*, "Bridging biological and artificial neural networks with emerging neuromorphic devices: Fundamentals, progress, and challenges," *Adv. Mater.*, vol. 31, no. 49, Dec. 2019, Art. no. 1902761, doi: 10.1002/adma.201902761.

[6] Z. Wang *et al.*, "Self-activation neural network based on self-selective memory device with rectified multilevel states," *IEEE Trans. Electron Devices*, vol. 67, no. 10, pp. 4166–4171, Oct. 2020, doi: 10.1109/TED.2020.3014566.

[7] Y. Fang *et al.*, "Investigation of NbO_x-based volatile switching device with self-rectifying characteristics," *Sci. China Inf. Sci.*, vol. 62, no. 12, 2019, Art. no. 229401, doi: 10.1007/s11432-019-9894-0.

[8] R. Han, P. Huang, Y. Zhao, X. Cui, X. Liu, and J. Kang, "Efficient evaluation model including interconnect resistance effect for large scale RRAM crossbar array matrix computing," *Sci. China Inf. Sci.*, vol. 62, no. 2, pp. 1–11, 2019. [Online]. Available: https://doi.org/10.1007/s11432-018-9555-8

[9] P.-Y. Jung, D. Panda, S. Chandrasekaran, S. Rajasekaran, and T.-Y. Tseng, "Enhanced switching properties in TaO_X memristors using diffusion limiting layer for synaptic learning," *IEEE J. Electron Devices Soc.*, vol. 8, pp. 110–115, 2020, doi: 10.1109/JEDS.2020.2966799.

[10] M. Yu *et al.*, "Encapsulation layer design and scalability in encapsulated vertical 3D RRAM," *Nanotechnology*, vol. 27, no. 20, 2016, Art. no. 205202. [Online]. Available: https://doi.org/10.1088/0957-4484/27/20/205202

[11] Z. Wang *et al.*, "Localized metal doping effect on switching behaviors of TaO_X-based RRAM device," in *Proc. IEEE Non-Volatile Memory Technol. Symp. (NVMTS)*, Pittsburgh, PA, USA, 2016, pp. 1–3, doi: 10.1109/NVMTS.2016.7781516.

[12] M.-J. Lee *et al.*, "A fast, high-endurance and scalable non-volatile memory device made from asymmetric $Ta2O5_{-x}/TaO_{2-x}$ bilayer structures" *Nat. Mater.*, vol. 10, no. 8, pp. 625–630, 2011. [Online]. Available: https://doi.org/10.1038/nmat3070

[13] P. Yao *et al.*, "Face classification using electronic synapses," *Nat. Commun.*, vol. 8, May 2017, Art. no. 15199, doi: 10.1038/ncomms15199.

[14] Z. Wang *et al.*, "In situ training of feed-forward and recurrent convolutional memristor networks," *Nat. Mach. Intell.*, vol. 1, no. 9, pp. 434–442, 2019, doi: 10.1038/s42256-019-0089-1.

[15] C. Li *et al.*, "Efficient and self-adaptive in-situ learning in multilayer memristor neural networks," *Nat. Commun.*, vol. 9, p. 2385, Jun. 2018, doi: 10.1038/s41467-018-04484-2.

[16] W. Shim, Y. Luo, J.-S. Seo, and S. Yu, "Investigation of read disturb and bipolar read scheme on multilevel RRAM-based deep learning inference engine," *IEEE Trans. Electron Devices*, vol. 67, no. 6, pp. 2318–2323, Jun. 2020, doi: 10.1109/TED.2020.2985013.

[17] M. Zhao, B. Gao, J. Tang, H. Qian, and H. Wu, "Reliability of analog resistive switching memory for neuromorphic computing," *Appl. Phys. Rev.*, vol. 7, no. 1, 2020, Art. no. 011301. [Online]. Available: https://doi.org/10.1063/1.5124915

[18] Y. Cai *et al.*, "Technology-array-algorithm co-optimization of RRAM for storage and neuromorphic computing: Device non-idealities and thermal cross-talk," in *Proc. IEEE Int. Electron Device Meeting (IEDM)*, San Francisco, CA, USA, 2020, pp. 1–4, doi: 10.1109/IEDM13553.2020.9371968.

[19] 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, Jul. 2019, doi: 10.1038/s41928-019-0270-x.

[20] D. Ielmini and H.-S. P. Wang, "In-memory computing with resistive switching devices," *Nat. Electron.*, vol. 1, no. 6, pp. 333–343, Jun. 2018, doi: 10.1038/s41928-018-0092-2.

[21] L. Wu *et al.*, "Nonlinear weight quantification for mitigating read disturb effect on multilevel RRAM-based neural network," in *Proc. IEEE Electron Devices Technol. Manu. Conf.*, Chengdu, China, 2021, pp. 1–3.