Perspective: Uniform switching of artificial synapses for large-scale neuromorphic arrays

APL Materials

Perspective: Uniform switching of artificial synapses for large-scale neuromorphic arrays

Cite as: APL Mater. **6**, 120901 (2018); https://doi.org/10.1063/1.5049137
Submitted: 19 July 2018 . Accepted: 19 September 2018 . Published Online: 05 December 2018

Scott H. Tan, Peng Lin, Hanwool Yeon, Shinhyun Choi, Yongmo Park, and Jeehwan Kim

Perspective: Uniform switching of artificial synapses for large-scale neuromorphic arrays

Scott H. Tan,[1,2] Peng Lin,[1,2] Hanwool Yeon,[1,2] Shinhyun Choi,[1,2] Yongmo Park,[1,2] and Jeehwan Kim[1,2,3,a]

[1]*Department of Mechanical Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139-4307, USA*

[2]*Research Laboratory of Electronics, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139-4307, USA*

[3]*Department of Materials Science and Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139-4307, USA*

(Received 19 July 2018; accepted 19 September 2018; published online 5 December 2018)

Resistive random-access memories are promising analog synaptic devices for efficient bio-inspired neuromorphic computing arrays. Here we first describe working principles for phase-change random-access memory, oxide random-access memory, and conductive-bridging random-access memory for artificial synapses. These devices could allow for dense and efficient storage of analog synapse connections between CMOS neuron circuits. We also discuss challenges and opportunities for analog synaptic devices toward the goal of realizing passive neuromorphic computing arrays. Finally, we focus on reducing spatial and temporal variations, which is critical to experimentally realize powerful and efficient neuromorphic computing systems. https://doi.org/10.1063/1.5049137

I. INTRODUCTION

Recently, artificial intelligence (AI) has allowed for significant technological advancements in image classification,[1–3] speech recognition,[4–6] strategic gaming,[7] and decision-making.[8–12] However, artificial neural networks (ANNs) require large amounts of computing power, especially for deep learning.[13] Because of this, there is a significant demand for more efficient hardware to accelerate training ANNs and improve classification accuracies for recognition tasks. For example, wider artificial neural network layers can be better at handling more complex datasets.[14] Increasing the number of computing cores has been the primary method to handle larger artificial neural networks. Hence, Graphics Processing Units (GPUs) have become the machine of choice for most AI. GPUs handle many operations in parallel by processing data using a centra1ized control of parallel arithmetic logic units (ALUs) that fetch data from a memory hierarchy.[15,16] Compared to GPUs, application-specific integrated circuit (ASIC) accelerators[17–23] and field-programmable gate array (FPGA) accelerators[24–27] demonstrated computing efficiency improvements. However, the performance of these complementary metal-oxide semiconductor (CMOS) based systems is limited by the large footprint of synaptic cells and frequent access to external memory.[28] This has engendered motivation to explore bio-inspired neuromorphic systems.[29–31] Figure 1 illustrates the trade-off between technologically mature CMOS-based hardware and ideal neuromorphic systems (Table I).

Contrasting electronic hardware, the human brain is efficient wetware consisting of roughly 80×10^9 neuronal cells and 150×10^{12} synapses.[32] A working hypothesis suggests that learning, in part, occurs when synapse connection strengths are re-programmed to process electrical signals differently than it did previously.[33] Memory and other cognitive processes emerge from encoded

[a]Author to whom correspondence should be addressed: jeehwan@mit.edu

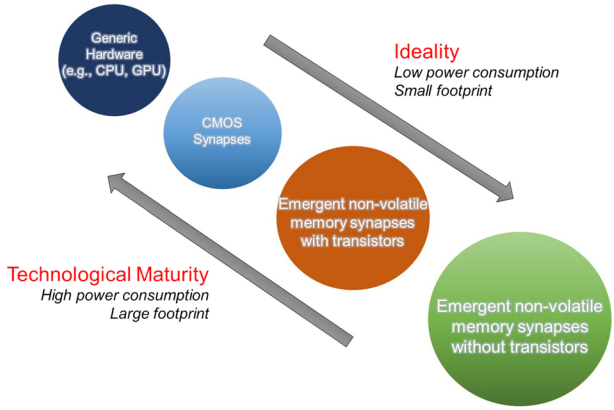

FIG. 1. The trade-off between ideality and technological maturity for bio-inspired neuromorphic systems. Passive arrays with emergent non-volatile memory synapses are ideal computing elements for artificial neural networks. Transistors can be used at each synapse to facilitate array control. CMOS synapses exhibit better robustness than emergent non-volatile memory synapses. Commercial technologies consume excessive power and have large footprint compared to neuromorphic computing architectures.

information in molecules, atoms, and ions within the cellular network.[34] While there is still much to learn about neurophysiological phenomena, discoveries in neuroscience[35–40] illuminate a path toward developing neuromorphic computing for more efficient AI.

In the late 1980s, Mead first described neuromorphic computing as a concept involving large integrated electronic analog systems that mimic biological neural networks.[41] Neuromorphic computing has since evolved into two categories: (1) bio-*plausible*: focused on mimicking dynamics of biological synapses and (2) bio-*inspired*: focused on developing electronic systems (loosely) inspired by the brain for more efficient artificial neural networks. In the first category, research has focused on experimentally demonstrating biologically observed phenomena, such as spike-timing-dependent plasticity (STDP), with artificial synapses.[42–44] However, how to efficiently utilize local updating rules with spikes remains an important question to tackle.[45] By contrast, bio-*inspired* neuromorphic computing focuses on implementing artificial neural network hardware built for learning algorithms that are well-defined mathematically.

This perspective focuses on recent progress toward achieving bio-inspired neuromorphic computing with analog arrays of artificial synapses,[46,47] as illustrated in Figs. 2(a) and 2(b). Like biological

TABLE I. Percent variation of potentiation voltage thresholds for two-terminal emergent non-volatile memory synapses with various strategies for reducing spatial and temporal variations.

Material	Strategy for uniformity	Temporal variation (%)	Spatial variation (%)	References
a-Si	Scaling	...	3	108
SiO_2	Nanocones	32	34	122
ZrO_x/HfO_x	Scaling	...	3	116
NiO	Stack modification	11	...	119
HfO_x	Stack modification	...	10	120
ZrO_2	Implanting Ti ions	35	...	128
CuC	Annealing	~30	...	129
GST	Cu_xO interface	13	...	118
TiO_2	Embedded Pt	9	...	131
ZnO	Nanorods	6	...	126
TiO_2	Ru nanorods	~30	...	130
ZrO_2	Cu nanocrystals	~50	...	127
SiO_x	Nanostructures	11	~30	125
TiO_x	Stack modification	...	16	86
ZnO w/Ti	Stack modification	~10	~25	117
SiGe	Dislocation confinement	3	5	75

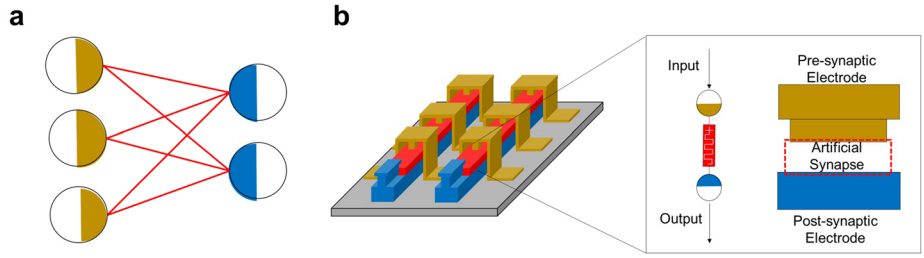

FIG. 2. A 2 × 3 fully connected neural network. (a) Synapses in an artificial neural network (lines) retain information about the connection strengths between neurons (circles). Each synaptic weight is trained or programmed to perform computations. (b) A 2 × 3 systolic array consists of pre-synaptic electrodes (gold) and post-synaptic electrodes (blue) with weighted associations determined by artificial synapse conductance values. Arrays can be used to compute weighted summations. Other components of the neural network, such as non-linear activations, can be designed in peripheral circuitry.

synapses, synaptic devices can reconfigure constituent ions and atoms to change the material conductivity between two electrode terminals. The conductance values, G, are analog weights used for passive weighted summations in crossbar arrays, as illustrated in Fig. 3(a). As voltage bias, V_i, are applied to M input rows, output currents, I_j, for all N output columns are weighted summations given by

$$I_j = \sum_{i=1}^{M} V_i G_{ij} \quad for\, j = 1, 2, \ldots, N,$$

where G_{ij} are synapse conductance values, as illustrated in Fig. 3(b). Passive arrays can execute weighted summations for all columns in parallel according to Ohm's law and Kirchhoff's current law. Each array implements vector-matrix multiplications given by

$$\bar{V} \cdot \boldsymbol{G}_{M\times N} = \begin{bmatrix} V_1 \\ V_2 \\ . \\ . \\ . \\ V_M \end{bmatrix}^T \begin{bmatrix} G_{11} & G_{12} & \cdots & G_{1N} \\ G_{21} & G_{22} & & G_{2N} \\ . & . & & . \\ . & . & & . \\ . & . & & . \\ G_{M1} & G_{M2} & \cdots & G_{MN} \end{bmatrix} = \begin{bmatrix} I_1 \\ I_2 \\ . \\ . \\ . \\ I_N \end{bmatrix}^T = \bar{I}.$$

Increasing and decreasing conductance values of synaptic devices are referred to as potentiation and depression, respectively. These operations are executed by applying electrical signals that exceed

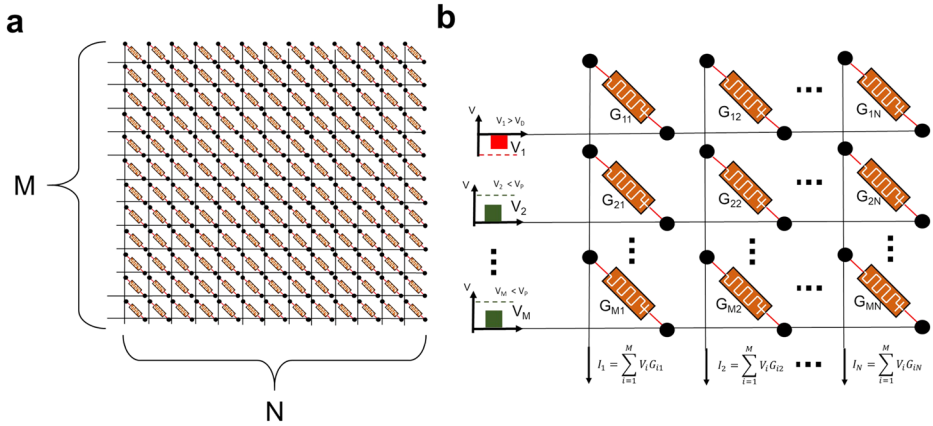

FIG. 3. Computing weighted summations with synapse arrays. (a) Synapses in an M by N crossbar array. Larger values of M and N could increase classification accuracy for more complex datasets. (b) Weighted summations using voltage inputs and synapse conductance values are computed at each output column. Sub-threshold voltage amplitudes are low enough that conductance values are not modified. In other words, synapses do not change conductance states as the weighted summations are read.

activation thresholds, V_P and V_D, respectively. After potentiation, the pre-synaptic input is more strongly associated with the post-synaptic output, i.e., the artificial synapse conductance increases, as illustrated in Fig. 4(a). After depression, an artificial synapse becomes more resistive, as illustrated in Fig. 4(b). Electrical signals weaker than threshold voltages do not frequently elicit the conductance change. This allows for negligible perturbance to conductance values while weighted summations are computed.

During online learning, artificial neural network training involves potentiation and depression of connection strengths between ensembles of neurons to minimize the global error of a training dataset. Out of the many existing learning algorithms,[33,48–51] a stochastic gradient descent using backpropagation is the most widely used method to update synaptic weights.[52–55] During training, the partial derivative of an error, δ_i, for each array output, i, is calculated by using CMOS neuron circuits, and conductance, $G_{ij,old}$, can be updated to $G_{ij,new}$ by the delta rule,[52]

$$G_{ij,new} = G_{ij,old} + \Delta G,$$

$$\Delta G = f(\eta, x_j, \delta_i),$$

where ΔG is the amount of conductance change, η is the learning rate parameter, and x_j is the activity at input j. The function $f(\eta, x_j, \delta_i)$ is a circuit implementation to physically manifest the synaptic weight change (computed by using CMOS neuron circuits) to reconfigure the material structure within an artificial synapse. For example, voltage pulses exceeding threshold values can be applied to potentiate or depress artificial synapses according to Eq. (3), as illustrated in Figs. 4(a) and 4(b). Protection voltages are commonly used to prevent unselected devices from unintentional conductivity changes. Similar schemes could be accomplished with circuitry comprised of integrators and comparators.[56–65]

Electrical CMOS circuits based on flash memory have been demonstrated for synaptic devices.[66,67] However, flash technologies are not ideal for online learning due to slow write speed, low endurance, and the requirement to erase an entire block at the same time. In addition to flash memory, three-terminal emergent non-volatile memory, such as organic transistors,[68] spin transfer torque magnetic random-access memory (STT-MRAM),[69] and ferroelectric devices[70] also show promise for multi-level storage of synaptic weights. These types of devices are in early stages of development.

Resistive random-access memories (RRAMs) are two-terminal emergent non-volatile memories that are suitable for artificial synapses since they are re-programmable and have analog

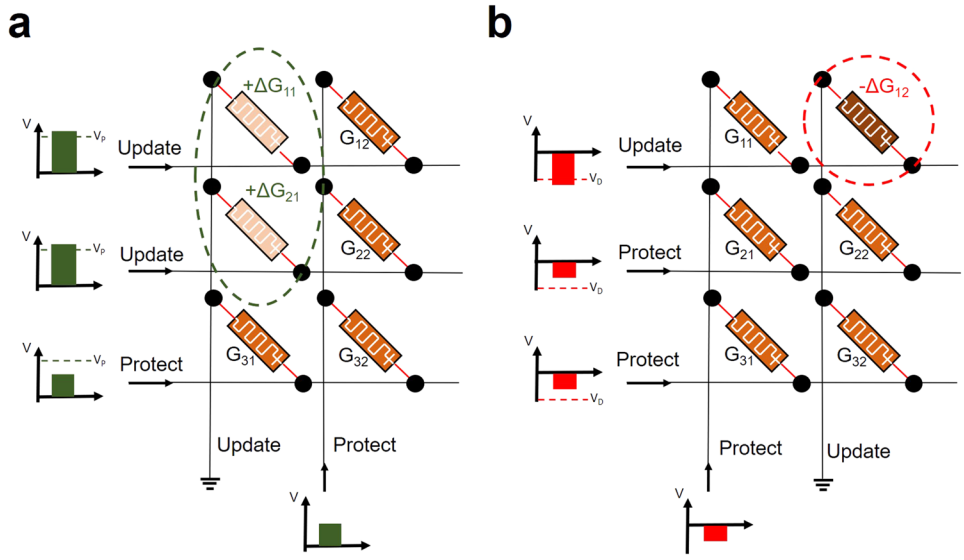

FIG. 4. Conductance values are updated during training or programming by applying supra-threshold voltage pulses. (a) Potentiation can occur when voltages applied across synapses exceed potentiation thresholds, V_P. (b) Depression can occur when voltages applied across the synapse exceed depression thresholds, V_D. Protection voltages bias could be used to suppress the weight change in other synapses.

conductance states. Device performance characteristics such as uniform switching, symmetric and linear potentiation/depression, many multi-level states, low operational power, fast switching speed, and good scaling are simultaneously desired for large-scale neuromorphic arrays. Three main types of RRAM are known as phase-change random-access memory (PCRAM), oxide random-access memory (OxRAM), and conductive bridging random-access memory (CBRAM). Many of these synaptic devices have demonstrated ideal characteristics for bio-inspired neuromorphic computing. To start, high signal contrast along with multi-level state accessibility is necessary to store precise synaptic weight values. HfO_x-based CBRAM has very large ON/OFF ratio, indicating that good signal contrast could be possible with these synaptic devices.[43] Also, Mo/TiO_x/TiN-based OxRAM remains stable at as many as 64 analog conductance levels.[71] For inference engines and other applications that rely on long-term memory, retention of conductance states is crucially important. Also, good cycling endurance is desired to allow online training of synapses.[45] Excellent endurance and retention have been observed in Ta_2O_{5-x}/TaO_{2-x} bilayer OxRAM. These devices exhibit switching for $>10^{12}$ cycles with over ten years of expected retention at room temperature.[72] Other metrics are the switching speed, scaling limitations, and linearity or symmetry of potentiation and depression. Sub-nanosecond switching speed is observed in Ta-based OxRAM.[73] Scaling as small as ~2 nm device size has been demonstrated using TiOx/HfO_2-based OxRAM.[74] Linear analog potentiation and depression has been demonstrated for a-Si-based and crystalline-SiGe-based CBRAM.[42,75]

Many review articles have surveyed RRAM devices.[45,76–83] Chen and Lin provided a review with focus on resistance variation for non-volatile memory or programmable logic applications.[84] In addition to resistance variation, switching threshold variation is also important since it correlates with repeatability and controllability of filament geometries and material configurations underlying the conductance change phenomena. This perspective focuses on switching threshold variations of artificial synapses for bio-inspired neuromorphic computing.

Section II of this article will introduce basic working principles for PCRAM, OxRAM, and CBRAM artificial synapses. After their operation mechanisms are discussed, Sec. III will highlight key challenges for these devices with a focus on strategies to reduce temporal and spatial variations of activation thresholds. By achieving better reliability and control of synapse performance, large-scale bio-inspired neuromorphic systems can potentially attain unprecedented computing efficiencies for advanced AI.

II. A BRIEF INTRODUCTION TO PCRAM, OxRAM, AND CBRAM

Two-terminal emergent non-volatile memories are promising synapses for analog arrays since they have a lithographic footprint size of only $4F^2$ in systolic crossbars and can take advantage of multi-level states and 3D architectures for further enhanced memory density.[85–88] Crossbar architectures have full-connectivity for passive weighted summations, which frequently repeat during artificial neural network operation. In this section, we will discuss the working principles of PCRAM, OxRAM, and CBRAM artificial synapses.[42,43,62,68,72,89,90]

A. PCRAM

Phase change random-access memory (PCRAM), also known as phase change memory (PCM), is a relatively mature RRAM technology.[91] As illustrated in Figs. 5(a) and 5(b), these devices are based on reversible crystallization and amorphization of chalcogenide materials. Potentiation and depression in PCRAM occur by altering the crystalline-to-amorphous volume ratio. The crystalline phase typically has dense long-range-ordered structure and high conductivity, whereas the amorphous phase is short-range ordered and porous with low conductivity.[64,65] Generally, resistive heater elements are used as bottom electrodes to localize Joule heating that induces material phase changes. Crystallization near the vicinity of the bottom interface depends on the temperature and the previous material structure. When the temperature of chalcogenide near the heater element rises above the glass transition temperature (T_g) while remaining below melting temperature (T_m), the volume ratio of crystalline-to-amorphous material increases, resulting in potentiation, as shown in Fig. 5(c).[64] During potentiation, atomic density increases, which increases the conductance state of the PCRAM

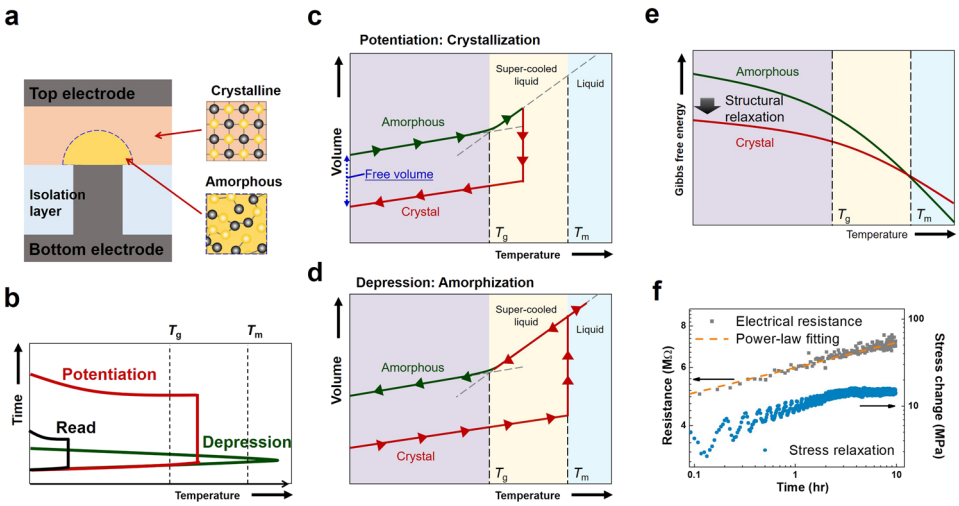

FIG. 5. PCRAM operation. (a) Schematics of vertical PCRAM and (b) temperature-time changes for potentiation, depression, and read. (c) Illustration of the volume changes and (d) Gibbs free energy of phase-change materials with respect to crystallization/amorphization. (e) Resistance drift and stress relaxation in amorphous $Ge_2Sb_2Te_5$ films during isothermal annealing at 80 °C. Reprinted with permission from Cho *et al.*, Electrochem. Solid-State Lett. **15**, H81–H83 (2012). Copyright 2012 The Electrochemical Society.

artificial synapse.[92] Depression occurs when the chalcogenide is rapidly quenched from a molten state to result in a larger volume fraction of amorphous material, as shown in Fig. 5(d).

Compared to other two-terminal emergent non-volatile memories, PCRAM is the most mature technology. Products based on PCRAM have been in the market since 2009, whereas OxRAM and CBRAM are still in earlier stages of development.[93] Modern PCRAM is scalable (device area smaller than 100 nm^2) and possesses a large conductance range (analog ON/OFF ratio of ~2000) due to large resistance contrast between amorphous and crystalline phases.[93]

However, despite the maturity of PCRAM, there remain several opportunities to improve its analog synapse performance. At low-conductance state, the amorphous structure is the dominant phase. PCRAM synaptic devices are inherently unstable because of thermodynamic instability of this amorphous phase. An amorphous phase has a larger specific volume than the crystalline phase, as illustrated in Fig. 5(e).[94,95] To form a more thermodynamically stable structure, structural rearrangement occurs in the amorphous phase even at much below crystalline temperature, known as structural relaxation (SR). During SR, electrical conductivity of the amorphous phase decreases continuously,[95] as shown in Fig. 5(f). Decaying conductivity could be caused by lower dangling bond density or release of a residual stress.

Although gradual potentiation is possible by inducing the spontaneous nucleation and growth of the crystalline region,[93,96] gradual depression is difficult since amorphization involves rapid quenching, which generates drastic changes in conductivity, as shown in Figs. 6(a) and 6(b). Additional circuit components could help to compensate for the drastic conductance change during depression.[97] However, temporal (cycle-to-cycle) variations are still a significant issue for PCRAM since Joule heating accelerates electromigration, which changes the chalcogenide stoichiometry by inducing phase segregation and void formation.[98–100]

B. OxRAM

Oxide random-access memory (OxRAM), also known as valence change memory (VCM), is another class of two-terminal synapses that is suitable for neuromorphic computing. Analog conductance states are accessible in many binary metal-oxide materials, such as HfO_x, TaO_x, TiO_x, AlO_x, ZrO_x, SiO_x, WO_x, ZnO_x, and many more.[72,101–104] In contrast to PCRAM, OxRAM relies on formation and rupture of localized conduction channels that result from the movement of oxygen vacancies.[72] After years of research, OxRAM has demonstrated promising properties such as

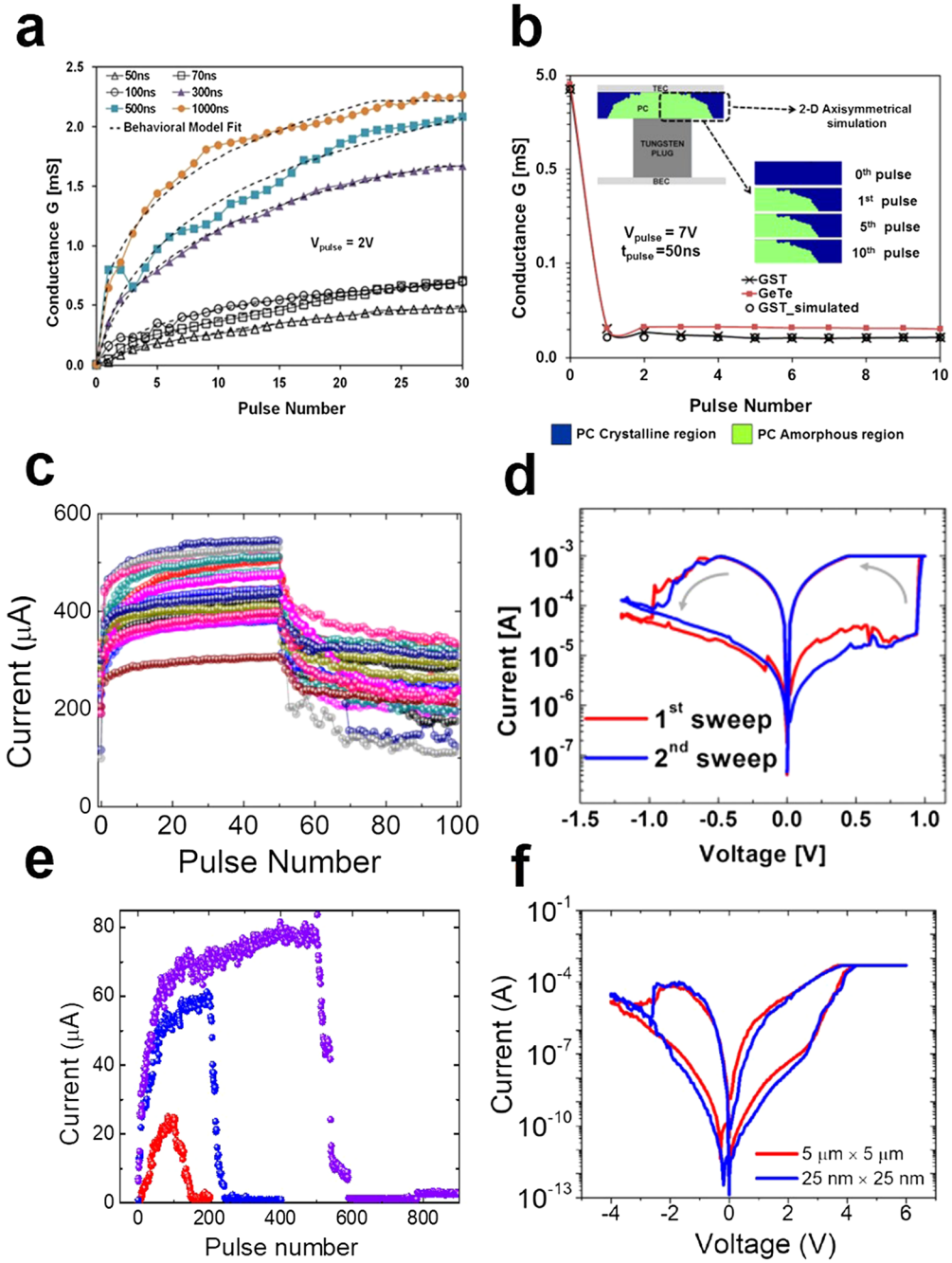

FIG. 6. Electrical characterization of PCRAM, oxide-RRAM, and CBRAM synapses. (a) Potentiation (30 pulses with an amplitude of 2 V and various pulse durations) and (b) depression (10 pulses with an amplitude of 7 V and a pulse duration of 50 ns) observed for GST-based PCRAM. Conductance states are read at −0.5 V for 1 ms after each potentiation and depression pulse. Reprinted with permission from Suri *et al.*, J. Appl. Phys. **112**, 054904 (2012). Copyright 2012 AIP Publishing LLC.[146] (c) Potentiation (50 pulses with an amplitude of 1.1 V and a pulse duration of 3 μs) followed by depression (50 pulses with an amplitude of −1.4 V and a pulse duration of 30 μs) for 18 Ta_2O_5-based oxide-RRAM synapses. Conductance states are read at 0.3 V for 1 ms after each potentiation and depression pulse. (d) Current-voltage (I-V) sweeps for Ta_2O_5-based oxide-RRAM with 100 μA current compliance. Reprinted with permission from Choi *et al.*, Nano Lett. **17**, 3113–3118 (2017). Copyright 2017 American Chemical Society.[147] (e) Potentiation (100/200/500 pulses with an amplitude of 5 V and a pulse duration of 5 μs) followed by depression (100/200/500 pulses with amplitude of −3 V for 5 μs) for epitaxial SiGe-based CBRAM. Conductance states are read at 2 V for 1 ms after each potentiation and depression pulse. (f) I-V sweeps for SiGe-based CBRAM for 5 μm × 5 μm and 25 nm × 25 nm synapses with 500 μA current compliance. Reprinted with permission from Choi *et al.*, Nat. Mater. **17**, 335 (2018). Copyright 2018 Springer Nature.

record-high endurance (>10^{12} cycles),[72] long data retention (>10 years at 85 °C),[105] good scalability (<10 nm),[106] fast switching speed (<1 ns),[73] and back-end-of-line (BEOL)/3D compatibility.[85–88]

Although demonstrations show promise, there are still challenges for OxRAM. First, temporal and spatial variations are unavoidable due to the stochasticity of conductive filament formation and rupture. Non-uniformity complicates the design of CMOS neuron circuits, decreases the efficiency of programming device conductance states, and results in reduced classification accuracies.[45] Furthermore, consistent conductance change step-size is hard to achieve in OxRAM since conductance has an exponential dependence on the conductive filament size and geometry.[107] As a result, OxRAM has been challenging to implement without additional CMOS at each synapse. Figure 6(c) shows spatial variation of potentiation and depression measured from Ta_2O_5-based OxRAM. Examples of current-voltage (I-V) cycles for Ta_2O_5-based OxRAM are shown in Fig. 6(d).

C. CBRAM

Conductive bridging random-access memory (CBRAM), also known as electrochemical metallization memory (ECM) or programmable metallization cell (PMC) memory, is also based on the formation and rupture of conduction channels. In contrast to OxRAM with oxygen vacancy conductive filaments, conduction channels in CBRAM are composed of metal, usually originating from one of the electrodes. These synaptic devices are also referred to as electrochemical metallization cells. CBRAM consists of an active metal moving through an amorphous solid electrolyte[42,43,108–111] or single-crystalline epitaxial film with dislocations.[75] During potentiation, a positive voltage is applied to the active electrode to oxidize metal into ions and electrons. Ions drift through the resistive medium and are reduced within the conduction channel, increasing the terminal-to-terminal conductivity. During depression, a negative potential is applied to the active electrode, and the metal conduction channel destabilizes and ruptures. Metal clusters have been observed forming and dissolving within a-Si and SiO_2.[109,112] Because CBRAM utilizes metal filaments rather than oxygen vacancies, the conductance range can be much larger than that of OxRAM.[43,77,112,113] Figure 6(e) shows potentiation and depression of epitaxial SiGe-based CBRAM. I-V cycles for epitaxial SiGe-based CBRAM are shown in Fig. 6(f).

In CBRAM, mid-level conductance states can initially tend to decay and stabilize at lower values due to metal clustering.[43] Stabilizing filaments by confining metal in dislocations could improve retention.[75] However, like PCRAM and OxRAM, CBRAM also suffers from large variations due to stochasticity associated with conduction channel formation and rupture. Achieving better uniformity for two-terminal emergent non-volatile memories is a critical challenge. Section III of this perspective will focus on strategies to minimize variations for PCRAM, OxRAM, and CBRAM artificial synapses.

III. STRATEGIES TO MINIMIZE VARIATIONS

Although stand-alone synapses show promise, array demonstrations have been limited to small-scale systems or require transistors to regulate each synapse individually. A major bottleneck limiting large-scale passive arrays is temporal (cycle-to-cycle) and spatial (device-to-device) variations, which result due to stochasticity of filament formation and rupture.[84,108,109,112,114] The more the conductance change amount per pulse fluctuates from synapse-to-synapse and in response to repetitive pulsing, the more difficult it becomes to implement conductance updating schemes. Furthermore, simulations accounting for variation predict that non-uniformity degrades classification accuracy.[45] As a general guideline, temporal variation above 2% will likely cause severe degradation of learning accuracy.[45] Arrays are fairly tolerant to spatial variations, but high yield (~100%) is essential since defective devices could permit excessive leakage currents.[45]

Potentiation voltage threshold is a useful metric for assessing uniformity since it is closely related to the conduction channel geometry. Consistent spatial (device-to-device) and temporal (cycle-to-cycle) response to identical voltage signals could allow for fast and low-power bio-inspired neuromorphic systems.[31,58] Conventionally, the percentage of variation is reported as the coefficient

of variation (standard deviation over the mean). In this perspective, we evaluate potentiation threshold voltage variations determined from current-voltage measurements. However, it is worth noting that thresholds for pulsing conditions slightly differ from DC measurements. Varying pulse duration alters the probability that a synapse will potentiate or depress in response.[75] An alternative metric to characterize uniformity is conductance state variations.[84] In this discussion, we evaluate synaptic devices variations by comparison of potentiation thresholds measured in DC current-voltage sweeps.

Consistent resistive switching fundamentally relies on controlling the evolution of the conductance channel. However, phase transitions in PCRAM and filament evolution in OxRAM and CBRAM are inherently stochastic processes.[115] While randomness cannot be entirely eliminated, stochasticity can be minimized by scaling, modifying device structures, embedding nanoparticles, and pre-defining filament geometries.

A. Scaling

Numerous candidate filament pathways exist under a large device area. Because of this, reducing the electrode area can effectively reduce stochasticity. Kim *et al.* reported a spatial variation of only 3% for 50 nm-node a-Si-based CBRAM, shown in Fig. 7(a).[108] Also, Lee *et al.* reported shrinking ZrO_x/HfO_x OxRAM to 50 nm reduces temporal variation to 3%.[116] Scaling down PCRAM device size can allow crystallization and amorphization across the entire volume to improve temporal variation.[93] However, this limits synapses to binary conductance values. Furthermore, as PCRAM device size decreases, potentiation thresholds are more sensitive to small differences in crystal grains and stoichiometry.[100]

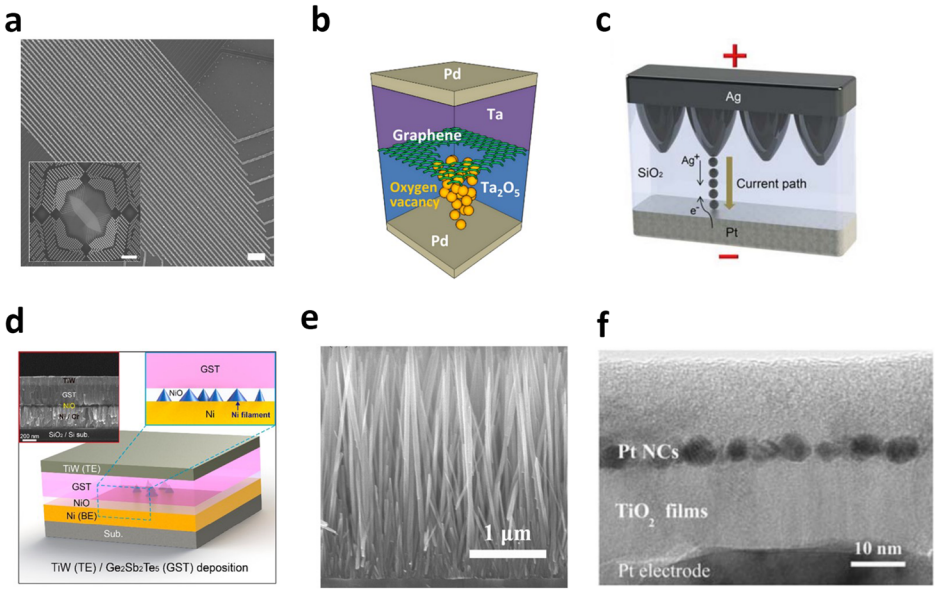

FIG. 7. Strategies for improving uniformity in emergent non-volatile memories. (a) SEM image of a crossbar 50 nm half pitch. Scaling down devices to nanometer length scales can enhance spatial uniformity of a-Si-based CBRAM. Scale bar: 500 nm. Inset: SEM of the 16 × 16 crossbar. Scale bar: 5 μm. Reprinted with permission from Kim *et al.*, Nano Lett. **12**, 389–395 (2012). Copyright 2012 American Chemical Society. (b) Schematic of a graphene nanopore directing oxygen vacancy filament formation in a Ta_2O_5-based oxide-RRAM. Reprinted with permission from Lee *et al.*, Appl. Phys. Lett. **100**, 142106 (2012). Copyright 2012 AIP Publishing LLC. (c) Schematic showing nanocones used for Ag filament localization in SiO_2-based CBRAM. Reprinted with permission from You *et al.*, ACS Nano **10**, 9478–9488 (2016). Copyright 2016 American Chemical Society. (d) Schematic showing nanopyramids used for phase transition localization in GST-based PCRAM. Inset: Cross section of the GST-based device. Reprinted with permission from You *et al.*, ACS Nano **9**, 6587–6594 (2015). Copyright 2015 American Chemical Society.[148] (e) SEM image of ZnO nanorods. Conductive filaments can be localized to the surface of nanorods, improving temporal uniformity. Reprinted with permission from Chang *et al.*, Appl. Phys. Lett. **96**, 242109 (2010). Copyright 2010 AIP Publishing LLC. (f) SEM image showing Pt nanocrystals embedded in TiO_2-based OxRAM for improved temporal uniformity. Reprinted with permission from Chang *et al.*, Appl. Phys. Lett. **95**, 042104 (2009). Copyright 2009 AIP Publishing LLC.

B. Structure modification

Modifying the device structure is a useful strategy to stabilize the formation and rupture of a conduction channel. For example, uniformity is improved for ZnO-based OxRAM by adding a TiO_x layer, resulting in about 25% spatial variation and 10% temporal variation.[117] In stacked crossbars, layered $TiN/TiO_{2-x}/Al_2O_3$ OxRAM between Pt electrodes has only 16% spatial variation.[86] Including a Ge-Sb-Te interface layer between Al and Cu_xO helps to isolate OxRAM filament formation for as low as 13% temporal variation.[118] Adding a layer of IrO_2 could help to stabilize oxygen migrations in NiO-based OxRAM, allowing for only 11% temporal variation.[119] Finally, embedding Al layers in HfO_x CBRAM can reduce spatial variation to 10%.[120]

Nanostructures have been employed to direct conductive pathways for improved uniformity. For example, nanopores in graphene could reduce variation in $Ta/G/Ta_2O_5$ OxRAM by acting as a diffusion barrier, as shown in Fig. 7(b).[121] Nanocones, as shown in Fig. 7(c), are capable of localizing Ag filaments in SiO_2 to result in 34% spatial variation and 32% temporal variation.[122] Uniformity improvement using pyramids has also been observed for Al_2O_3-based CBRAM,[123,124] and GST-based PCRAM, as shown in Fig. 7(d). Self-assembled SiO_x nanostructures in the inert electrode can help localize oxide-RRAM conductive filaments for about 30% spatial variation and 11% temporal variation.[125] Also, vertically aligned ZnO nanorods, as shown in Fig. 7(e), can also enhance uniformity by confining oxide-RRAM conductive filaments to nanorod surfaces for as low as 6% temporal variation.[126]

C. Embedding nanoparticles

Embedding nanoparticles is another effective method to reduce variation since nanoparticles can help to guide the formation of conductive filaments. Liu showed that adding nanocrystals to ZrO_2-based OxRAM results in 50% temporal variation.[127] Alternatively, Ti ions can be implanted into

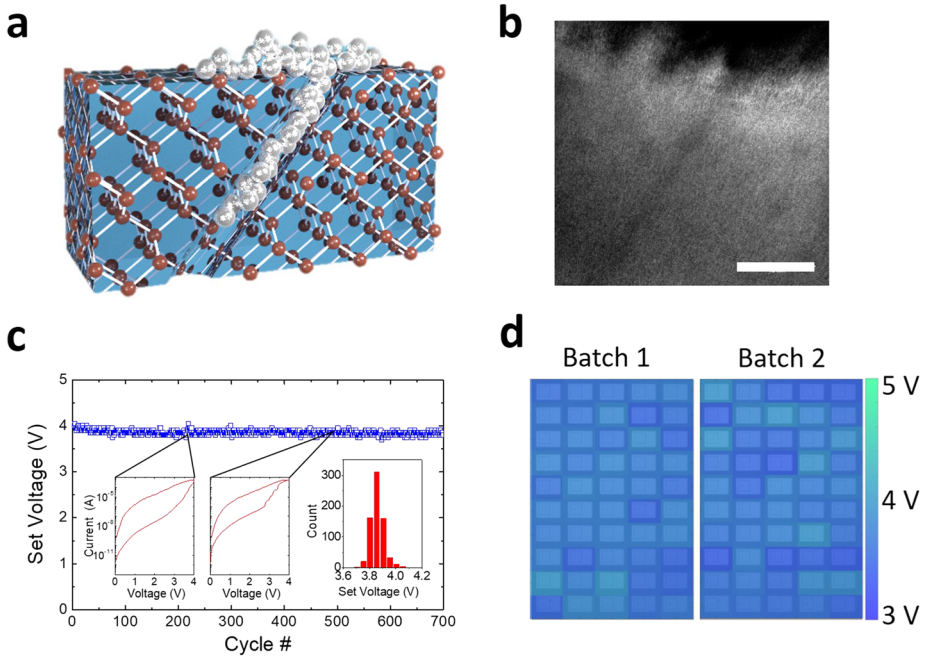

FIG. 8. Metal confinement through dislocations in single-crystalline films. (a) Illustration of metal atoms confined in a crystalline lattice to form a conduction channel. (b) TEM of an Ag filament confined in SiGe. Scale bar: 12 nm. (c) Temporal (cycle-to-cycle) evolution of potentiation threshold voltage, measured as the voltage when current exceeds 300 μA. Temporal variation is as low as 1%. (d) Spatial (device-to-device) color map of potentiation threshold voltages overlaid on optical images of the synaptic devices. Spatial variation is only 4.9% for 100 artificial synapses fabricated in two separate processing batches. Reprinted with permission from Choi *et al.*, Nat. Mater. **17**, 335 (2018). Copyright 2018 Springer Nature.

ZrO_2-based oxide-RRAM to allow for 35% temporal variation.[128] Lee *et al.* showed that annealing to agglomerate conductive filaments in CuC-based CBRAM allows for about 30% temporal variation.[129] In TiO_2-based OxRAM, embedding Ru nanodots allows for about 30% variation,[130] whereas embedding Pt nanocrystals, as shown in Fig. 7(f), isolates conductive filament formation, resulting in only 9% temporal variation.[131] Finally, in SiO_2-based OxRAM, embedding Pt allows for up to 1000 uniform cycles.[132]

D. Defining channels in single crystals

Pre-defining conduction channels using threading dislocations in a single-crystalline material is an effective technique for controlling spatial and temporal conductive filament dynamics. Potentiation and depression have been demonstrated using single-crystalline STO-based OxRAM,[133] as well as for crystalline-SiGe-based CBRAM.[75] Although ions are typically less mobile in defect-free single-crystals compared to amorphous materials, threading dislocations allow conduction channels to form with as low as 1% temporal variation and only 4.9% spatial variation,[75] as shown in Fig. 8. Under pulsing conditions, the potentiation threshold voltage shows 4.8% spatial variation and 3.9% temporal variation, while the depression threshold voltage shows 5.3% spatial variation and 4.8% temporal variation. These devices also demonstrate large conductance range, good endurance, and long retention time at high conductance states.

IV. CONCLUDING REMARKS

Bio-inspired neuromorphic computing arrays with PCRAM, OxRAM, and CBRAM artificial synapses have demonstrated pattern classification. To create robust arrays with non-uniform emergent non-volatile memories, transistors can help to compensate for variability and other device non-idealities. For example, with one transistor at each synaptic device, HfO_2/Al_2O_3-based OxRAM in a 128×8 array demonstrated facial recognition.[134] A 500×661 PCRAM array with one transistor at each artificial synapse has also demonstrated pattern classification.[62] Adding additional CMOS circuits to each artificial synapse can further enhance computing capabilities. For example, two PCRAMs, two transistors, and one capacitor for each artificial synapses can execute the training and image classification using MNIST, MNIST-back-rand, CIFAR-10, and CIFAR-100 datasets.[97]

Passive arrays without transistors regulating each synapse promise an ultimate reduction in footprint size and power consumption. Transistor-free Al_2O_3/TiO_2-based OxRAM demonstrated classification of 3×3-black/white pixel images using a passive 12×12 array.[57] Al_2O_3/TiO_{2-x}-based OxRAM in two 20×20 passive arrays demonstrated one-hidden layer perceptron classification.[135] WO_x-based OxRAM in a 32×32 crossbar array demonstrated offline learning for image analysis.[136]

Although these demonstrations are exciting, device variations impose limitations to classification accuracy with more complex databases. For example, a color image in the CIFAR-10 and CIFAR-100 datasets consists of $32 \times 32 \times 3$ inputs, which are too large for passive arrays to currently handle. Large arrays require complicated CMOS neuron control circuits to deal with variations. Robust and reliable artificial synapses could greatly simplify these peripheral access circuits to realize larger neuromorphic systems.

High-performance synaptic devices with minimal variations are critical to demonstrate passive arrays of analog artificial synapses. Spatial and temporal uniformity can be improved by shrinking device size, modifying device structure, embedding nanoparticles, or defining channels in single-crystals. Although some device non-idealities are tolerable, better control of conductance change is essential to realize wider arrays for powerful and efficient neuromorphic computing.

Several issues also influence device behavior and contribute to device properties. For example, the nanobattery effect caused by inhomogeneous ion distribution in the switching medium results in stored charges.[137,138] At metal/insulator interfaces, native oxides degrade reliability of devices.[139] Moisture changes alter device properties since water molecules can be incorporated into defect sites.[140,141] Nano-scale variations in local (short-range) material structure and density can cause significant switching variations as well.[113,142] Furthermore, the interplay between cation/anion mobility and oxidation/reduction reaction rates governs growth direction and shape of conductive filaments.[109]

Based on measured synaptic device characteristics, there is also significant opportunity to improve circuitry and algorithms to map calculated weight values to artificial synapses. Circuit implementation of Eq. (3) requires thoughtful design to efficiently apply voltage signals, read weighted summations, calculate and backpropagate partial derivatives of error, and modify artificial synapse conductance values. Mapping algorithms could also help to tolerate device non-idealities and variations. Highly integrated CMOS neuron circuits will be important to expand AI system capacities for more complicated tasks.

Accurate and detailed modeling of the dynamic processes that give rise to conductance change could help guide both device engineering and algorithm development. Advanced microscopic imaging, such as *in situ* transmission electron microscopy (TEM), may help reveal new strategies for controlling potentiation and depression.

In addition to improving emergent non-volatile memories, bio-inspired neuromorphic systems also could benefit from high-performance selector devices.[143–145] Selectors introduce non-linearity to suppress leakage currents at sub-threshold voltages. Two-terminal architectures allow for vertical integration with emergent non-volatile memories.

In summary, PCRAM, OxRAM, and CBRAM are promising emergent non-volatile memories for analog neuromorphic arrays. Additional transistors or CMOS circuitry can compensate for spatial and temporal non-uniformity and allow pattern classification with comparable accuracies to generic hardware. Passive analog artificial synapse arrays are capable of dense and efficient bio-inspired neuromorphic computing. Large-scale passive arrays could be achieved by reducing variations to better control conductance change without requiring transistors. In conclusion, improving uniformity of artificial synapses can allow for efficient and powerful hardware specialized for AI.

[1] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proceedings of the 25th International Conference on Neural Information Processing Systems* (NIPS Proceedings, 2012), Vol. 1, pp. 1097–1105, available at http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.

[2] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in Proceedings of the International Conference on Learning Representations; e-print arXiv:1409.1556 (2015).

[3] C. Szegedy *et al.*, "Going deeper with convolutions," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (IEEE, 2015).

[4] A. Graves, A. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing* (IEEE, 2013). pp. 6645–6649.

[5] C. Weng, D. Yu, S. Watanabe, and B.-H. F. Juang, "Recurrent deep neural networks for robust speech recognition," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (IEEE, 2014), pp. 5532–5536.

[6] G. Hinton *et al.*, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," IEEE Signal Process. Mag. **29**, 82–97 (2012).

[7] V. Mnih *et al.*, "Human-level control through deep reinforcement learning," Nature **518**, 529–533 (2015).

[8] X. Guo, S. Singh, H. Lee, R. L. Lewis, and X. Wang, "Deep learning for real-time Atari game play using offline Monte-Carlo tree search planning," in *NIPS Proceedings* (NIPS Proceedings, 2014), pp. 3338–3346.

[9] T. Schaul, J. Quan, I. Antonoglou, and D. Silver, "Prioritized experience replay," presented at International Conference on Learning Representations (ICLR), San Diego, CA; e-print arXiv:1511.05952 (2015), available at https://arxiv.org/abs/1511.05952.

[10] B. C. Stadie, S. Levine, and P. Abbeel, "Incentivizing exploration in reinforcement learning with deep predictive models" (2015).

[11] D. Silver *et al.*, "Mastering the game of Go without human knowledge," Nature **550**, 354–359 (2017).

[12] N. Brown and T. Sandholm, "Superhuman AI for heads-up no-limit poker: Libratus beats top professionals," Science **359**, 418 (2017).

[13] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," Nature **521**, 436–444 (2015).

[14] S. Zagoruyko and N. Komodakis, "DiracNets: Training very deep neural networks without skip-connections" (2017).

[15] V. Sze, Y.-H. Chen, T.-J. Yang, and J. Emer, "Efficient processing of deep neural networks: A tutorial and survey," Proc. IEEE **105**, 2295–2329 (2017).

[16] See http://www.nvidia.com/object/accelerate-inference.html for Deep Learning Inference Accelerators | NVIDIA Tesla | NVIDIA; accessed 17 January 2018.

[17] D. A. Hammerstrom, "VLSI architecture for high-performance, low-cost, on-chip learning," in *1990 IJCNN International Joint Conference on Neural Networks* (IEEE, 1990), Vol. 2, pp. 537–544.

[18] U. Ramacher *et al.*, *VLSI Design of Neural Networks* (Springer US, 1991), pp. 271–310.

[19] T. Chen *et al.*, "DianNao: A small-footprint high-throughput accelerator for ubiquitous machine-learning," ACM SIGPLAN Not. **49**, 269–284 (2014).

[20] Z. Du *et al.*, "ShiDianNao," in *Proceedings of the 42nd Annual International Symposium on Computer Architecture—ISCA'15* (ACM Press, 2015), Vol. 43, pp. 92–104.

[21] Y.-H. Chen, T. Krishna, J. S. Emer, and V. Sze, "Eyeriss: An energy-efficient reconfigurable accelerator for deep convolutional neural networks," IEEE J. Solid-State Circuits **52**, 127–138 (2017).
[22] N. P. Jouppi *et al.*, "In-datacenter performance analysis of a tensor processing unit," in *Proceedings of the 44th Annual International Symposium on Computer Architecture, ISCA '17* (ACM, 2017), pp. 1–12.
[23] Y. Chen *et al.*, "DaDianNao: A machine-learning supercomputer," in *2014 47th Annual IEEE/ACM International Symposium on Microarchitecture* (IEEE, 2014), pp. 609–622.
[24] S. Li *et al.*, "FPGA acceleration of recurrent neural network based language model," in *2015 IEEE 23rd Annual International Symposium on Field-Programmable Custom Computing Machines* (IEEE, 2015), pp. 111–118.
[25] C. Farabet, C. Poulet, J. Y. Han, and Y. LeCun, "CNP: An FPGA-based processor for convolutional networks," in *2009 International Conference on Field Programmable Logic and Applications* (IEEE, 2009), pp. 32–37.
[26] C. Zhang *et al.*, "Optimizing FPGA-based accelerator design for deep convolutional neural networks," in *Proceedings of the 2015 ACM/SIGDA International Symposium on Field-Programmable Gate Arrays—FPGA'15* (ACM Press, 2015), pp. 161–170.
[27] S. Chakradhar *et al.*, "A dynamically configurable coprocessor for convolutional neural networks," in *Proceedings of the 37th Annual International Symposium on Computer Architecture—ISCA'10* (ACM Press, 2010), Vol. 38, p. 247.
[28] N. P. Jouppi *et al.*, "In-datacenter performance analysis of a tensor processing unit," in *Proceedings of the 44th Annual International Symposium on Computer Architecture—ISCA'17* (ACM Press, 2017), Vol. 45, pp. 1–12.
[29] X. Liu *et al.*, "Harmonica: A framework of heterogeneous computing systems with memristor-based neuromorphic computing accelerators," IEEE Trans. Circuits Syst. I **63**, 617–628 (2016).
[30] X. Liu *et al.*, "RENO," in *Proceedings of the 52nd Annual Design Automation Conference on—DAC'15* (ACM Press, 2015), pp. 1–6.
[31] T. Gokmen, O. M. Onen, and W. Haensch, "Training deep convolutional neural networks with resistive cross-point devices," Front. Neurosci. **11**, 538 (2017).
[32] F. A. C. Azevedo *et al.*, "Equal numbers of neuronal and nonneuronal cells make the human brain an isometrically scaled-up primate brain," J. Comp. Neurol. **513**, 532–541 (2009).
[33] D. O. Hebb, *The Organization of Behavior : A Neuropsychological Theory* (Lawrence Erlbaum Associates, 2002).
[34] R. Yuste, "From the neuron doctrine to neural networks," Nat. Rev. Neurosci. **16**, 487–497 (2015).
[35] C. M. Armstrong, "Time course of TEA(+)-induced anomalous rectification in squid giant axons," J. Gen. Physiol. **50**, 491–503 (1966).
[36] C. M. Armstrong and L. Binstock, "Anomolous rectification in the squid giant axon injected with tetraethylammonium chloride," J. Gen. Physiol. **48**, 859–872 (1965).
[37] S. A. George, D. N. Mastronarde, and M. W. Dubin, "Prior activity influences the velocity of impulses in frog and cat optic nerve fibers," Brain Res. **304**, 121–126 (1984).
[38] A. L. Hodgkin and A. F. Huxley, "A quantitative description of membrane current and its application to conduction and excitation in nerve," J. Physiol. **117**, 500–544 (1952).
[39] I. Tasaki and A. S. Hagiwar, "Demonstration of two stable potential states in the squid giant axon under tetraethylammonium chloride," J. Gen. Physiol. **40**, 859–885 (1957).
[40] P. F. Baker, A. L. Hodgkin, and T. I. Shaw, "Replacement of the protoplasm of a giant nerve fibre with artificial solutions," Nature **190**, 885–887 (1961).
[41] C. Mead, "Neuromorphic electronic systems," Proc. IEEE **78**, 1629–1636 (1990).
[42] S. H. Jo *et al.*, "Nanoscale memristor device as synapse in neuromorphic systems," Nano Lett. **10**, 1297–1301 (2010).
[43] Z. Wang *et al.*, "Memristors with diffusive dynamics as synaptic emulators for neuromorphic computing," Nat. Mater. **16**, 101 (2016).
[44] S. Kim *et al.*, "NVM neuromorphic core with 64k-cell (256-by-256) phase change memory synaptic array with on-chip neuron circuits for continuous *in situ* learning," in *2015 IEEE International Electron Devices Meeting (IEDM)* (IEEE, 2015), pp. 17.1.1–17.1.4.
[45] S. Yu, "Neuro-inspired computing with emerging nonvolatile memorys," Proc. IEEE **106**, 260–285 (2018).
[46] K. Steinbuch, "Die lernmatrix," Kybernetik **1**, 36–45 (1961).
[47] C. Lehmann, M. Viredaz, and F. Blayo, "A generic systolic array building block for neural networks with on-chip learning," IEEE Trans. Neural Networks **4**, 400–407 (1993).
[48] F. Rosenblatt, "The perceptron: A probabilistic model for information storage and organization in the brain," Psychol. Rev. **65**, 386–408 (1958).
[49] E. L. Bienenstock, L. N. Cooper, and P. W. Munro, "Theory for the development of neuron selectivity: Orientation specificity and binocular interaction in visual cortex," J. Neurosci. **2**, 32–48 (1982).
[50] G. E. Hinton, P. Dayan, B. J. Frey, and R. M. Neal, "The 'wake-sleep' algorithm for unsupervised neural networks," Science **268**, 1158–1161 (1995).
[51] G. E. Hinton, "Training products of experts by minimizing contrastive divergence," Neural Comput. **14**, 1771–1800 (2002).
[52] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," Nature **323**, 533–536 (1986).
[53] S. Linnainmaa, "Taylor expansion of the accumulated rounding error," BIT **16**, 146–160 (1976).
[54] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," Proc. IEEE **86**, 2278–2324 (1998).
[55] Y. LeCun *et al.*, "Backpropagation applied to handwritten zip code recognition," Neural Comput. **1**, 541–551 (1989).
[56] G. W. Burr *et al.*, "Large-scale neural networks implemented with non-volatile memory as the synaptic weight element: Comparative performance analysis (accuracy, speed, and power)," in *2015 IEEE International Electron Devices Meeting (IEDM)* (IEEE, 2015), pp. 4.4.1–4.4.4.
[57] M. Prezioso *et al.*, "Training and operation of an integrated neuromorphic network based on metal-oxide memristors," Nature **521**, 61–64 (2015).

[58] S. Yu *et al.*, "Scaling-up resistive synaptic arrays for neuro-inspired architecture: Challenges and prospect," in *2015 IEEE International Electron Devices Meeting (IEDM)* (IEEE, 2015), pp. 17.3.1–17.3.4.
[59] B. Li, Y. Wang, Y. Wang, Y. Chen, and H. Yang, "Training itself: Mixed-signal training acceleration for memristor-based neural network," in *2014 19th Asia and South Pacific Design Automation Conference (ASP-DAC)* (IEEE, 2014), pp. 361–366.
[60] Z. Xu *et al.*, "Parallel programming of resistive cross-point array for synaptic plasticity," Proc. Comput. Sci. **41**, 126–133 (2014).
[61] J. Schmidhuber, "Deep learning in neural networks: An overview," Neural Networks **61**, 85 (2014).
[62] G. W. Burr *et al.*, "Experimental demonstration and tolerancing of a large-scale neural network (165,000 synapses), using phase-change memory as the synaptic weight element," in *2014 IEEE International Electron Devices Meeting* (IEEE, 2014), pp. 29.5.1–29.5.4.
[63] J. Seo *et al.*, "On-chip sparse learning acceleration with CMOS and resistive synaptic devices," IEEE Trans. Nanotechnol. **14**, 969–979 (2015).
[64] D. Soudry, D. Di Castro, A. Gal, A. Kolodny, and S. Kvatinsky, "Memristor-based multilayer neural networks with online gradient descent training," IEEE Trans. Neural Networks Learn. Syst. **26**, 2408–2421 (2015).
[65] P.-Y. Chen, L. Gao, and S. Yu, "Design of resistive synaptic array for implementing on-chip sparse learning," IEEE Trans. Multi-Scale Comput. Syst. **2**, 257–264 (2016).
[66] X. Guo *et al.*, "Fast, energy-efficient, robust, and reproducible mixed-signal neuromorphic classifier based on embedded NOR flash memory technology," in *2017 IEEE International Electron Devices Meeting (IEDM)* (IEEE, 2017), pp. 6.5.1–6.5.4.
[67] X. Guo *et al.*, "Temperature-insensitive analog vector-by-matrix multiplier based on 55 nm NOR flash memory cells," in *2017 IEEE Custom Integrated Circuits Conference (CICC)* (IEEE, 2017), pp. 1–4.
[68] Y. van de Burgt *et al.*, "A non-volatile organic electrochemical device as a low-voltage artificial synapse for neuromorphic computing," Nat. Mater. **16**, 414 (2017).
[69] J.-G. Zhu, "Magnetoresistive random access memory: The path to competitiveness and scalability," Proc. IEEE **96**, 1786–1798 (2008).
[70] Y. Kaneko, Y. Nishitani, and M. Ueda, "Ferroelectric artificial synapses for recognition of a multishaded image," IEEE Trans. Electron Devices **61**, 2827–2833 (2014).
[71] J. Park *et al.*, "TiOx-based RRAM synapse with 64-levels of conductance and symmetric conductance change by adopting a hybrid pulse scheme for neuromorphic computing," IEEE Electron Device Lett. **37**, 1559–1562 (2016).
[72] M. J. Lee *et al.*, "A fast, high-endurance and scalable non-volatile memory device made from asymmetric Ta2O(5-x)/TaO(2-x) bilayer structures," Nat. Mater. **10**, 625–630 (2011).
[73] A. C. Torrezan, J. P. Strachan, G. Medeiros-Ribeiro, and R. S. Williams, "Sub-nanosecond switching of a tantalum oxide memristor," Nanotechnology **22**, 485203 (2011).
[74] S. Pi *et al.*, "Memristor crossbar arrays with 6-nm half-pitch and 2-nm critical dimension," Nature Nanotechnology (published online).
[75] S. Choi *et al.*, "SiGe epitaxial memory for neuromorphic computing with reproducible high performance based on engineered dislocations," Nat. Mater. **17**, 335 (2018).
[76] I. Valov, R. Waser, J. R. Jameson, and M. N. Kozicki, "Electrochemical metallization memories—Fundamentals, applications, prospects," Nanotechnology **22**, 254003 (2011).
[77] D. Ielmini and R. Waser, "Resistive switching : From fundamentals of nanoionic redox processes to memristive device applications" (2016).
[78] H. S. P. Wong *et al.*, "Metal-oxide RRAM," Proc. IEEE **100**, 1951 (2012).
[79] L. Zhu, J. Zhou, Z. Guo, and Z. Sun, "An overview of materials issues in resistive random access memory," J. Materiomics **1**, 285 (2015).
[80] A. Prakash, D. Jana, and S. Maikap, "TaOx-based resistive switching memories: Prospective and challenges," Nanoscale Res. Lett. **8**, 418 (2013).
[81] D. S. Jeong and C. S. Hwang, "Nonvolatile memory materials for neuromorphic intelligent machines," Adv. Mater. **30**, 1704729 (2018).
[82] D. S. Jeong *et al.*, "Emerging memories: Resistive switching mechanisms and current status," Rep. Prog. Phys. **75**, 076502 (2012).
[83] S. T. Han, Y. Zhou, and V. A. Roy, "Towards the development of flexible non-volatile memories," Adv. Mater. **25**, 5425 (2013).
[84] A. Chen and M. R. Lin, "Variability of resistive switching memories and its impact on crossbar array performance," in *Proceedings of IEEE International Reliability Physics Symposium* (IEEE, 2011), pp. 843–846.
[85] B. Chakrabarti *et al.*, "A multiply-add engine with monolithically integrated 3D memristor crossbar/CMOS hybrid circuit," Sci. Rep. **7**, 42429 (2017).
[86] G. C. Adam *et al.*, "3-D memristor crossbars for analog and neuromorphic computing applications," IEEE Trans. Electron Devices **64**, 312–318 (2017).
[87] Z. Li, P.-Y. Chen, H. Xu, and S. Yu, "Design of ternary neural network with 3-D vertical RRAM array," IEEE Trans. Electron Devices **64**, 2721–2727 (2017).
[88] H. Li *et al.*, "Four-layer 3D vertical RRAM integrated with FinFET as a versatile computing unit for brain-inspired cognitive information processing," in *2016 IEEE Symposium on VLSI Technology* (IEEE, 2016), pp. 1–2.
[89] S. Kim *et al.*, "Experimental demonstration of a second-order memristor and its ability to biorealistically implement synaptic plasticity," Nano Lett. **15**, 2203–2211 (2015).
[90] T. Gokmen and Y. Vlasov, "Acceleration of deep neural network training with resistive cross-point devices: Design considerations," Front. Neurosci. **10**, 333 (2016).
[91] S. R. Ovshinsky, "Reversible electrical switching phenomena in disordered structures," Phys. Rev. Lett. **21**, 1450–1453 (1968).

[92] S. Menzel, U. Böttger, M. Wimmer, and M. Salinga, "Physics of the switching kinetics in resistive memories," Adv. Funct. Mater. **25**, 6306–6325 (2015).
[93] S. G. Sarwat, "Materials science and engineering of phase change random access memory," Mater. Sci. Technol. **33**, 1890–1906 (2017).
[94] S. R. Elliot, *Physics of Amorphous Materials* (John Wiley & Sons, Inc., 1986).
[95] J.-Y. Cho, T.-Y. Yang, Y.-J. Park, and Y.-C. Joo, "Study on the resistance drift in amorphous Ge2Sb2Te5 according to defect annihilation and stress relaxation," Electrochem. Solid-State Lett. **15**, H81–H83 (2012).
[96] G. W. Burr *et al.*, "Neuromorphic computing using non-volatile memory," Adv. Phys. X **2**, 89–124 (2017).
[97] S. Ambrogio *et al.*, "Equivalent-accuracy accelerated neural-network training using analogue memory," Nature **558**, 60–67 (2018).
[98] T.-Y. Yang, I.-M. Park, B.-J. Kim, and Y.-C. Joo, "Atomic migration in molten and crystalline Ge2Sb2Te5 under high electric field," Appl. Phys. Lett. **95**, 032104 (2009).
[99] K. Baek *et al.*, "Microstructure-dependent DC set switching behaviors of Ge–Sb–Te-based phase-change random access memory devices accessed by *in situ* TEM," NPG Asia Mater. **7**, e194 (2015).
[100] G. W. Burr *et al.*, "Phase change memory technology," J. Vac. Sci. Technol., B: Nanotechnol. Microelectron.: Mater., Process., Meas., Phenom. **28**, 223–262 (2010).
[101] K. Shibuya, R. Dittmann, S. Mi, and R. Waser, "Impact of defect distribution on resistive switching characteristics of Sr2TiO4 thin films," Adv. Mater. **22**, 411–414 (2010).
[102] G. S. Park *et al.*, "*In situ* observation of filamentary conducting channels in an asymmetric Ta2O5-x/TaO2-x bilayer structure," Nat. Commun. **4**, 2382 (2013).
[103] S. Yu, Y. Wu, R. Jeyasingh, D. Kuzum, and H.-S. P. Wong, "An electronic synapse device based on metal oxide resistive switching memory for neuromorphic computation," IEEE Trans. Electron Devices **58**, 2729–2737 (2011).
[104] D. B. Strukov, G. S. Snider, D. R. Stewart, and R. S. Williams, "The missing memristor found," Nature **453**, 80–83 (2008).
[105] H. Jiang *et al.*, "Sub-10 nm Ta channel responsible for superior performance of a HfO2 memristor," Sci. Rep. **6**, 28525 (2016).
[106] S. Pi, P. Lin, and Q. Xia, "Cross point arrays of 8 nm × 8 nm memristive devices fabricated with nanoimprint lithography," J. Vac. Sci. Technol., B: Nanotechnol. Microelectron.: Mater., Process., Meas., Phenom. **31**, 06FA02 (2013).
[107] S. Kim, S. Choi, and W. Lu, "Comprehensive physical model of dynamic resistive switching in an oxide memristor," ACS Nano **8**, 2369–2376 (2014).
[108] K. H. Kim *et al.*, "A functional hybrid memristor crossbar-array/CMOS system for data storage and neuromorphic applications," Nano Lett. **12**, 389–395 (2012).
[109] Y. Yang *et al.*, "Electrochemical dynamics of nanoscale metallic inclusions in dielectrics," Nat. Commun. **5**, 377–383 (2014).
[110] S. H. Jo and W. Lu, "CMOS compatible nanoscale nonvolatile resistance switching memory," Nano Lett. **8**, 392–397 (2008).
[111] R. Waser, R. Dittmann, C. Staikov, and K. Szot, "Redox-based resistive switching memories nanoionic mechanisms, prospects, and challenges," Adv. Mater. **21**, 2632–2663 (2009).
[112] Y. Yang *et al.*, "Observation of conducting filament growth in nanoscale resistive memories," Nat. Commun. **3**, 732 (2012).
[113] J. J. Yang, D. B. Strukov, and D. R. Stewart, "Memristive devices for computing," Nat. Nanotechnol. **8**, 13–24 (2013).
[114] K. Krishnan, T. Tsuruoka, C. Mannequin, and M. Aono, "Mechanism for conducting filament growth in self-assembled polymer thin films for redox-based atomic switches," Adv. Mater. **28**, 640–648 (2016).
[115] Y. Yang and W. Lu, "Nanoscale resistive switching devices: Mechanisms and modeling," Nanoscale **5**, 10076–10092 (2013).
[116] J. Lee *et al.*, "Diode-less nano-scale ZrOx/HfOx RRAM device with excellent switching uniformity and reliability for high-density cross-point memory applications," in *2010 International Electron Devices Meeting* (IEEE, 2010), pp. 19.5.1–19.5.4.
[117] J. H. Park, D. S. Jeon, and T. G. Kim, "Improved uniformity in the switching characteristics of ZnO-based memristors using Ti sub-oxide layers," J. Phys. D.: Appl. Phys. **50**, 015104 (2017).
[118] H. Lv, H. Wan, and T. Tang, "Improvement of resistive switching uniformity by introducing a thin GST interface layer," IEEE Electron Device Lett. **31**, 978–980 (2010).
[119] D. C. Kim *et al.*, "Improvement of resistive memory switching in NiO using IrO2," Appl. Phys. Lett. **88**, 232106 (2006).
[120] S. Yu *et al.*, "Improved uniformity of resistive switching behaviors in HfO[sub 2] thin films with embedded Al layers," Electrochem. Solid-State Lett. **13**, H36 (2010).
[121] J. Lee, C. Du, K. Sun, E. Kioupakis, and W. D. Lu, "Tuning ionic transport in memristive devices by graphene with engineered nanopores," ACS Nano **10**, 3571–3579 (2016).
[122] B. K. You *et al.*, "Reliable memristive switching memory devices enabled by densely packed silver nanocone arrays as electric-field concentrators," ACS Nano **10**, 9478–9488 (2016).
[123] K.-Y. Shin *et al.*, "Controllable formation of nanofilaments in resistive memories via tip-enhanced electric fields," Adv. Electron. Mater. **2**, 1600233 (2016).
[124] Y.-C. Huang *et al.*, "High-performance programmable metallization cell memory with the pyramid-structured electrode," IEEE Electron Device Lett. **34**, 1244–1246 (2013).
[125] B. K. You *et al.*, "Reliable control of filament formation in resistive memories by self-assembled nanoinsulators derived from a block copolymer," ACS Nano **8**, 9492–9502 (2014).
[126] W. Y. Chang, C. A. Lin, J. H. He, and T. B. Wu, "Resistive switching behaviors of ZnO nanorod layers," Appl. Phys. Lett. **96**, 242109 (2010).
[127] Q. Liu *et al.*, "Controllable growth of nanoscale conductive filaments in solid-electrolyte-based ReRAM by using a metal nanocrystal covered bottom electrode," ACS Nano **4**, 6162–6168 (2010).
[128] Q. Liu *et al.*, "Improvement of resistive switching properties in ZrO2-based ReRAM with implanted Ti ions," IEEE Electron Device Lett. **30**, 1335–1337 (2009).

[129] W. Lee *et al.*, "Improved switching uniformity in resistive random access memory containing metal-doped electrolyte due to thermally agglomerated metallic filaments," Appl. Phys. Lett. **100**, 142106 (2012).
[130] J. H. Yoon *et al.*, "Highly improved uniformity in the resistive switching parameters of TiO2 thin films by inserting Ru nanodots," Adv. Mater. **25**, 1987–1992 (2013).
[131] W.-Y. Chang *et al.*, "Improvement of resistive switching characteristics in TiO2 thin films with embedded Pt nanocrystals," Appl. Phys. Lett. **95**, 042104 (2009).
[132] B. J. Choi *et al.*, "Electrical performance and scalability of Pt dispersed SiO_2 nanometallic resistance switch," Nano Lett. **13**, 3213–3217 (2013).
[133] K. Szot, W. Speier, G. Bihlmayer, and R. Waser, "Switching the electrical resistance of individual dislocations in single-crystalline SrTiO3," Nat. Mater. **5**, 312–320 (2006).
[134] P. Yao *et al.*, "Face classification using electronic synapses," Nat. Commun. **8**, 15199 (2017).
[135] F. M. Bayat *et al.*, "Implementation of multilayer perceptron network with highly uniform passive memristive crossbar circuits," Nat. Commun. **9**, 2331 (2018).
[136] P. M. Sheridan *et al.*, "Sparse coding with memristor networks," Nat. Nanotechnol. **12**, 784–789 (2017).
[137] I. Valov *et al.*, "Nanobatteries in redox-based resistive switches require extension of memristor theory," Nat. Commun. **4**, 1771 (2013).
[138] S. Tappertzhofen, E. Linn, U. Bottger, R. Waser, and I. Valov, "Nanobattery effect in RRAMs-Implications on device stability and endurance," IEEE Electron Device Lett. **35**, 208 (2014).
[139] D. Y. Cho, M. Luebben, S. Wiefels, K. S. Lee, and I. Valov, "Interfacial metal-oxide interactions in resistive switching memories," ACS Appl. Mater. Interfaces **9**, 19287 (2017).
[140] T. Tsuruoka *et al.*, "Effects of moisture on the switching characteristics of oxide-based, gapless-type atomic switches," Adv. Funct. Mater. **22**, 70 (2012).
[141] F. Messerschmitt, M. Kubicek, and J. L. M. Rupp, "How does moisture affect the physical property of memristance for anionic-electronic resistive switching memories?," Adv. Funct. Mater. **25**, 5117 (2015).
[142] S. Choi, Y. Yang, and W. Lu, "Random telegraph noise and resistance switching analysis of oxide based resistive memory," Nanoscale **6**, 400 (2014).
[143] J. Liang and H.-S. P. Wong, "Cross-point memory array without cell selectors—Device characteristics and data storage pattern dependencies," IEEE Trans. Electron Devices **57**, 2531–2538 (2010).
[144] J. Zhou, K. H. Kim, and W. Lu, "Crossbar RRAM arrays: Selector device requirements during read operation," IEEE Trans. Electron Devices **61**, 1369–1376 (2014).
[145] S. Kim, J. Zhou, and W. D. Lu, "Crossbar RRAM arrays: Selector device requirements during write operation," IEEE Trans. Electron Devices **61**, 2820–2826 (2014).
[146] M. Suri *et al.*, "Physical aspects of low power synapses based on phase change memory devices," J. Appl. Phys. **112**, 054904 (2012).
[147] S. Choi, J. H. Shin, J. Lee, P. Sheridan, and W. D. Lu, "Experimental demonstration of feature extraction and dimensionality reduction using memristor networks," Nano Lett. **17**, 3113–3118 (2017).
[148] B. K. You, M. Byun, S. Kim, and K. J. Lee, "Self-structured conductive filament nanoheater for chalcogenide phase transition," ACS Nano **9**, 6587–6594 (2015).