Co-integration of single transistor neurons and synapses by nanoscale CMOS fabrication for highly scalable neuromorphic hardware

Joon-Kyu Han
Korea Advanced Institute of Science and Technology
Jungyeop Oh
Korea Advanced Institute of Science and Technology
Gyeong-Jun Yun
Korea Advanced Institute of Science and Technology
Dongeun Yoo
National Nanofab Center (NNFC)
Myung-Su Kim
Korea Advanced Institute of Science and Technology
Ji-Man Yu
Korea Advanced Institute of Science and Technology
Sung-Yool Choi
Korea Advanced Institute of Science and Technology
Yang-Kyu Choi (ykchoi@ee.kaist.ac.kr)
Korea Advanced Institute of Science and Technology https://orcid.org/0000-0001-5480-7027

Article

Keywords: artificial neural network (ANN), complementary metal-oxide-semiconductor (CMOS), co-integration, excitatory neuron, inhibitory neuron, MOSFET, neuromorphic system, neuron device, silicon-oxide-nitride-oxide-silicon (SONOS), single transistor latch (STL), single transistor neuron, spiking neural network (SNN), synapse device

Posted Date: January 6th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-120802/v1

Abstract

Co-integration of multi-state single transistor neurons and synapses was demonstrated for highly scalable neuromorphic hardware, using nanoscale complementary metal-oxide-semiconductor (CMOS) fabrication. The neurons and synapses were integrated on the same plane with the same process because they have the same structure of a metal-oxide-semiconductor field-effect transistor (MOSFET) with different functions such as homotype. By virtue of 100% CMOS compatibility, it was also realized to co-integrate the neurons and synapses with additional CMOS circuits, such as a current mirror and inverter. Such co-integration can enhance packing density, reduce chip cost, and simplify fabrication procedures. Neuronal inhibition and tunability of the firing threshold voltage were demonstrated for an energy efficient and reliable neural network. The multi-state single transistor neuron with low peak power consumption of 120 nW that can control neuronal inhibition and the firing threshold voltage was achieved. Spatio-temporal neuronal functionalities are demonstrated with analyses of a fabricated neuromorphic module, which is composed of a single transistor neuron and a set of single transistor synapse. Image processing for letter pattern recognition and face image recognition is performed using a hardware-based circuit simulation and a software-based neuromorphic simulation, respectively.

Introduction

Although software-based artificial neural networks (ANNs) have led to breakthroughs in a variety of intelligent tasks, they inevitably have inherent delays and significant energy consumption because the hardware structure to support the ANNs is still based on the von Neumann architecture.[1–3] To overcome these limitations, hardware-based ANNs, known as brain-inspired neuromorphic systems, have been intensively studied.[4–6] The human brain consists of neurons for the computational function and synapses for the memory function, as shown in **Fig. 1a**. There are about 10^{11} neurons and 10^{15} synapses, and thus it is important to implement neurons and synapses with high density in order to mimic the brain in hardware.[7,8]

Neurons are mainly composed of CMOS-based circuits, while synapses primarily comprise memristors.[9–15] However, circuit-based neurons are problematic for high packing density with low-cost because they are composed of a capacitor, integrator, and comparator including many transistors.[16,17] Furthermore, simultaneous integration of circuit-based neurons and memristor synapses in a single chip is challenging because they use different materials and fabrications. Even worse, they should be linked to each other *via* specific interconnections owing to their inherent heterotypic structures. Such interconnections hence impose constraints on increasing packing density and simplifying process complexity. Also, extra energy consumption cannot be avoided at the interface between the neurons and the synapses.

Meanwhile, few works to co-integrate memristor based artificial neurons and synapses in a single crossbar array for a fully memristive neural network have been reported.[18–20] It should be noted that the neuromorphic hardware heavily relies on extra interface and control circuitry that collect, process, and

transport data, as well as neurons and synapses.[21–23] Therefore, the neuromorphic hardware should contain additional CMOS circuits to support processing units, peripheral interfaces, memory, clocking circuits, and input/output for a complete application. However, memristors cannot be co-integrated with CMOS circuits on the same plane because they use different materials and fabrications. Although memristor arrays could be directly integrated on the top of the CMOS circuits, line resistance effects were observed and further scaling of the memristor arrays was difficult because the memristor arrays should be located between lower interconnections and upper interconnections or on top of the multi-layered metal interconnections.[14] These problems will be exacerbated as the metal interconnections of CMOS circuits become more complex and the number of the metal layers increases. Furthermore, commercialization of memristive devices is hindered by the immaturity of the fabrication process, which limits large scale integration with high yield.[24] From another point of view, neuronal inhibition and tunability of firing threshold voltage are important for an energy efficient and reliable neural network.[25–27] However, the memristor neurons cannot self-function for control of neuronal inhibition and firing threshold voltage due to the lack of controllability.

In this work, highly scalable neuromorphic hardware was implemented by simultaneously integrating multi-state single transistor neurons and synapses on the same plane. Both devices have the same homotypic MOSFET structure. In detail, the MOSFET for a neuron and a synapse encloses a charge trap layer in gate dielectrics with the same manner as a commercial flash memory based on a SONOS structure that comprise a gate poly-crystalline Si (S), blocking SiO_2 (O), charge trap Si_3N_4 (N), tunneling SiO_2 (O), and channel single-crystalline Si (S). Due to this CMOS compatibility, they were fabricated and integrated on the same plane using the standard Si CMOS fabrication. It is possible to co-integrate single transistor neurons and synapses with CMOS circuits for processing units, peripheral interfaces, memory, clocking circuits, and input/output at the same time and thus co-integration of the entire neuromorphic system is available. Therefore, a highly scalable neural network can be implemented in a single chip, which can enhance packing density, reduce chip cost, and simplify fabrication procedures. Unit neuron and a set of synapses were fabricated and directly interconnected. And their connection properties were analyzed. The abovementioned charge trap Si_3N_4 in the MOSFET can allow multi-states. The multi-states according to trapped charges control the excitatory/inhibitory function or changes the firing threshold voltage ($V_{T,firing}$) in the neuron, while they regulate synaptic weight in the synapse. These homotypic neurons and synapses were directly connected to realize spatio-temporal neural computations. At the same time, CMOS circuits such as a current mirror and inverter, which are key elements for analog and digital circuits, were fabricated on the same plane to show the feasibility of co-integration of the interface and control circuits. In addition to real device fabrication, image recognition was successfully implemented with the aid of experimental based simulations.

Results And Discussion

Unit device characteristics of neuron and synapse

N-channel single transistor neuron and synapse have the same SONOS structure, as shown in **Fig. 1b**. The intercalated charge trap nitride (Si_3N_4) in the multi-layered gate dielectrics allows multi-states according to the amount of trapped charges. They can perform two functions: (i) enable excitatory/inhibitory function or tuning the $V_{T,firing}$ in the neuron and (ii) control weight update in the synapse. Like the homotype, the neuron and the synapse have the same structure but operate differently, as shown in **Fig. 1c**. For neuron operation, input current (I_{in}) collected from the pre-synapses is applied to a n^+ drain (or source) electrode, and output voltage (V_{out}) is produced from the same n^+ drain (or source) electrode. For synapse operation, the voltage transferred from the pre-neuron (V_{in}) is applied to the gate electrode of the synapse, and output current (I_{out}) is flown from the n^+ source (or drain) electrode. These neurons and synapses were fabricated on an 8-inch wafer by using the same standard Si CMOS process, and were connected to each other through metallization for a monolithically integrated neuromorphic system, as shown in **Fig. 1d**. The fabrication details are described in **Supplementary Information 1**.

As mentioned earlier, the excitatory/inhibitory state of the neuron is determined by electron trapping in the nitride of the SONOS structure. An inhibitory function that disables the firing of the neuron is necessary, because it can improve the energy efficiency of the neuromorphic system by selectively firing a specific neuron. Hence it can realize effective learning and inference through the winner-takes-all (WTA) mechanism.[28-30] As shown in **Fig. 2a**, unless the electrons are trapped in the nitride, the neuron is at a low-resistance state (LRS). Thus, current flows through the channel when the I_{in} is applied. As a consequence, charges are not integrated and a leaky integrate-and-fire (LIF) function is inhibited. Otherwise, the neuron is at a high-resistance state (HRS) when trapped electrons in the nitride raise a potential barrier between n^+ source and p-type channel referred to as a p-n built-in potential. Accordingly, charges are integrated until the firing. For the neuron operation, the gate of the neuron transistor is a kind of a pseudo-gate, unlike a conventional actual-gate. It is used not for the LIF operation but for charge trapping. For electron trapping in the nitride, a positive voltage pulse is applied to the pseudo-gate. Afterwards, it is sustained in a floating state for the neuron operation. Due to non-volatility of the trapped charges even without gate biasing, energy consumption is much smaller compared to our previous study, which required additional and continuous gate voltage control.[31,32]

Fig. 2b shows output characteristics of the fabricated n-channel single transistor neuron, which is represented by the drain current versus drain voltage (I_D-V_D). Its gate length (L_G) and channel width (W_{CH}) are 880 nm and 280 nm, respectively. Before the electron trapping, I_D flows regardless of V_D. After the electron trapping with gate voltage (V_G) of 12 V and pulse time of 100 ms, the I_D does not flow at a low V_D. However, a large amount of I_D abruptly flows beyond a critical V_D; this is called latch-up voltage (V_{latch}). This is known as a phenomenon of single transistor latch (STL) and serves as a threshold switch.[33,34]

Fig. 2c shows the V_{out} versus time when a constant I_{in} was applied to the drain electrode of the single transistor neuron, before and after the electron trapping. The V_{out} was measured at the same drain

electrode. Before the electron trapping, the applied I_{in} directly flowed through the channel toward the source, and charge accumulation (integration) was not allowed. As a result, the inhibitory function was enabled, unlike the two-terminal based memristor neuron. After the electron trapping, the applied I_{in} did not flow out toward the source and charges accumulated in a parasitic capacitor (C_{par}). According to this integration process, V_D equivalent to V_{out} was increased prior to the $V_{T,firing}$. Simultaneously, iterative impact ionization was induced by the increased V_D, and holes accumulated in the body. When the V_{out} reaches V_{latch}, which is the same as the $V_{T,firing}$, the accumulated charges in C_{par} are suddenly discharged by STL. This is a firing process. Therefore, spiking of the neuron was mimicked. **Fig. S2** shows the energy band diagram during the LIF operation, which was extracted by a TCAD device simulation. Note that at the moment of the firing, the energy barrier between the n^+ source and p-type body is lowered enough to allow the integrated charges to escape toward the source. The measured spiking frequency (f) was increased as the I_{in} was increased. In addition, leaky characteristics appeared, as described in **Supplementary Information 3**. The single transistor neuron shows typical LIF operation.

In addition to the control of the excitatory/inhibitory state, the $V_{T,firing}$ was tunable by controlling the trapped charge density in the nitride. This tunable property of the $V_{T,firing}$ is important to implement a reliable neuromorphic system.[25-27] If the conductivity of the synapse is unsuitably low or high owing to process-induced variability and endurance problems, the targeted number of firings cannot be achieved. To suppress this instability, a tunable $V_{T,firing}$ is required. As shown in **Supplementary Information 4**, the $V_{T,firing}$ was increased as the number of pulses that can control the trapped charge density was increased. This tendency is because fewer electrons were injected over the built-in potential of the n^+ source and the p-type body due to the greater amount of trapped charges, and thereby the V_{latch} was increased. In summary, the demonstrated multi-state single transistor neuron harnesses both controllability of the excitatory/inhibitory and tunability of the $V_{T,firing}$.

The f of the LIF neuron can be modeled as follows:

$$f = \frac{1}{\int_0^{V_{T,firing}} \left(\frac{C_{par}}{I_{in} - \frac{V_{out}}{R_{off}}} \right) dV_{out}},$$

where R_{off} is the resistance at HRS during the integration. As the $V_{T,firing}$ decreases, the f increases because the firing occurs at the lower voltage. It should be noted that the $V_{T,firing}$, which corresponds to the V_{latch} in **Fig. 2b**, is determined by various parameters such as L_G (**Supplementary Fig. 5**), body doping concentration, and energy band gap.[34,35] As the I_{in} increases, charging speed becomes faster, the f tends to be increased. Besides the $V_{T,firing}$ and I_{in}, the C_{par} plays an important role in controlling the f. From the above equation, the f is increased as the C_{par} is reduced because it takes shorter time to charge the

smaller parasitic capacitor (**Supplementary Information 6**). Accordingly, energy consumption per spike (E/spike) is also decreased as the C_{par} is reduced. Power consumption was compared between the single transistor neuron and the memristor neuron. The peak power consumption was extracted from the multiplication of peak current and peak V_{out} (**Supplementary Information 7**). It was found that the single transistor neuron consumed 120 nW, which was 10 to 10^4-fold smaller than the consumption of the memristor neuron, owing to a small cross-sectional channel area for current flowing due to high scalability of the nano-CMOS fabrication. On the other hand, it is noteworthy that the single transistor neuron has a bidirectional characteristic, in which the spiking operation is possible in both the drain input/output (I/O) and source I/O (**Supplementary Information 8**). This bidirectional characteristic can provide more degrees of freedom in designing a neuromorphic system. Thus, we employed both methods to construct a neuromorphic system.

Since the synapse device has the same SONOS structure as the neuron, the weight of the synapse can be adjusted by controlling the trapped charge density in the nitride. For example, if the electrons are trapped by applying a positive bias to the gate, the threshold voltage (V_T) is shifted rightward and the channel conductance is decreased at the same read voltage, as depicted in **Fig. 2d**. This is a kind of depression. Otherwise, V_T is shifted leftward and the channel conductance is increased at the same read voltage. This is a kind of potentiation. **Fig. 2e** shows transfer characteristics of the fabricated n-channel single transistor synapse, which is represented by the drain current versus gate voltage (I_D-V_G). Its L_G and W_{CH} are 1880 nm and 180 nm, respectively. V_T was adjusted by the applied gate voltage that controls the trapped charge density. The potentiation-depression (P-D) curve in **Fig. 2f** shows the conductance change (weight update) according to the number of applied pulses with an identical amplitude and duty cycle. Both V_G and V_D for the reading operation were set as 1 V. The V_G for potentiation and depression was set as -11 V with a pulse width of 100 ms and 11 V with a pulse width of 10 ms, respectively. As a result, 32 levels (5 bits) of conductance states were secured.

Co-integration of neuron and synapse

If a neuron and a synapse are homotypic, they can be integrated on the same plane at the same time with the same fabrication. Thereafter they can be connected by metal interconnections. This co-integration is demonstrated for two layers in a neural network. One is a pre-layer composed of a pre-synaptic neuron and a transmitted synapse. The other is a post-layer comprising a transmitting synapse and a post-synaptic neuron. **Figs. 3a-c** show the co-integrated pre-synaptic neuron and transmitted synapse as the pre-layer. Referring to the circuit schematic of **Fig. 3a**, a constant input current ($I_{in,neuron}$) is applied to the drain electrode of the neuron, and the drain is connected to a gate of the synapse to apply the output voltage from the pre-synaptic neuron ($V_{out,pre\text{-}neuron}$). Note that this configuration employs the abovementioned drain I/O scheme. Therefore, when spiking of the neuron occurs, the corresponding drain current (I_D) flows through the channel of the synapse. Its magnitude is modulated by the synaptic weight.

In the case of a three-terminal synapse such as a MOSFET, input resistance to the gate is huge. On the contrary, in the case of a two-terminal synapse such as a memristor, the input resistance is too small to suppress the loading effect where a neuronal output is influenced by the resistance of the synapse when it is directly connected to the neuron.[36,37] This is a great advantage for large-scale co-integration of pre-synaptic neurons and transmitted synapses, as explained in **Supplementary Information 9**. **Fig. 3b** shows the fabricated pre-synaptic neuron and transmitted synapse interconnected through metallization. As shown in **Fig. 3c**, the spike-shaped output current of the transmitted synapse ($I_{out,syn}$) was increased according to the $V_{out,pre\text{-}neuron}$ of the excitatory pre-synaptic neuron in order of weight: $w_1<w_2<w_3$. It should be noted that the f of the $I_{out,syn}$ was determined by the $I_{in,neuron}$. The spiking was inhibited when it was connected to the inhibitory pre-synaptic neuron, as shown in **Supplementary Information 9**. Note that stable inference operation is allowed unless the tunneling oxide thickness of the SONOS-based synapse is reduced (**Supplementary Information 10**). This is because the synaptic weight would not be changed by $V_{out,pre\text{-}neuron}$, which is small compared to the voltage of potentiation/depression. **Figs. 3d-f** show the co-integrated post-layer composed of the transmitting synapse and the post-synaptic neuron. As shown in the circuit schematic of **Fig. 3d**, a constant gate voltage ($V_{in,syn}$) is applied to the transmitting synapse, and the drain of the synapse is connected to the source of the post-synaptic neuron. $I_{out,syn}$ is thus applied to the post-synaptic neuron. The output voltage is measured at the source of the post-synaptic neuron. In other words, it adopts the source I/O scheme. If the $I_{out,syn}$ is applied from the source of the transmitting synapse to the drain of the post-synaptic neuron (drain I/O scheme), the source voltage of the transmitting synapse is floated. This is the reason why the post-synaptic neuron is selected to have the source I/O scheme. **Fig. 3e** shows the fabricated transmitting synapse and post-synaptic neuron interconnected through metallization. As shown in **Fig. 3f**, the f of the output voltage from the excitatory post-synaptic neuron ($V_{out,post\text{-}neuron}$) is increased according to the increment of $I_{out,syn}$ from the transmitting synapse in order of weight: $w_1<w_2<w_3$.

Another way to connect the transmitting synapse and the post-synaptic neuron is suggested in **Supplementary Information 11**, where a current mirror is used. The current mirror is composed of two NMOSFETs and two PMOSFETs. In this case, the drain I/O scheme of the post-synaptic neuron is available by reversing the direction of the $I_{out,syn}$. This configuration is also attractive to modulate the $I_{out,syn}$ over a wide range by changing the channel width of the current mirror. In addition to the current mirror that can be used for analog circuitry, an inverter composed of an NMOSFET and a PMOSFET, which is a fundamental block to construct digital logic circuitry that controls the neural network for collecting, processing, and transporting data, was also fabricated on the same plane with co-integration of the neuron and synapse at the same time (**Supplementary Information 12**).

Gain modulation and coincidence detection

Using the co-integrated neurons and synapses, spatio-temporal neural computations such as gain modulation and coincidence detection were carried out. In biology, gain modulation is observed in many cortical areas and is thought to play an important role in maintaining stability.[38-41] Herein gain modulation was realized by co-integration of two transmitting synapses and one post-synaptic neuron, as shown in the circuit diagram of **Fig. 4a**. Two types of pre-synaptic inputs are applied to the gate electrodes of two synapses. A driving input ($V_{G,S1}$) enables the post-synaptic neuron to fire and a modulatory input ($V_{G,S2}$) tunes the effectiveness of the driving input, as illustrated in **Fig. 4b**. As shown in **Fig. 4c**, the *f* of the post-synaptic neuron was modulated by the $V_{G,S2}$ for the fixed $V_{G,S1}$. This is because the I_{in} applied to the post-synaptic neuron was increased as the $V_{G,S2}$ was increased. **Fig. 4d** shows the secondary data that the *f* was increased as the $V_{G,S2}$ was increased at various $V_{G,S1}$.

Coincidence detection is another important neural computation that encodes information by detecting the occurrence of temporally close but spatially distributed input signals. It has been found that coincidence detection is significant for highly efficient information processing in auditory and visual systems.[42-45] By the co-integration of neuron and synapses, coincidence detection is also possible. When two inputs were applied at the same time, the *f* was increased because the I_{in} applied to the post-synaptic neuron was increased, as illustrated in **Fig. 4b**. Accordingly, it is possible to determine whether two inputs are simultaneously applied. **Fig. 4e** shows the corresponding data. When the two input signals applied at the same time, the *f* of the neuron was larger than the other cases of the two signals that were not synchronized. In addition, when two input signals overlapped for a certain period of time, the *f* of the neuron increased only in the overlap region.

Letter recognition with hardware circuit simulation

The neuromorphic system is commonly used to recognize images such as letters, numbers, objects, and faces. Pattern recognition of a letter was demonstrated with the aid of SPICE circuit simulations that were based on the measured neuron-synapse characteristics. As a simple model, the neuron is composed of a threshold switch and a parasitic capacitor connected in parallel. As a result, the simulated electrical properties are similar to the measured characteristics from the fabricated neuron, as shown in **Supplementary Fig. 3**. The synapse was implemented with a three-terminal MOSFET, and the weight of the synapse was controlled by adjusting the V_T. We implemented two types of neural networks: a classifier based on a single-layer perceptron (SLP) and an auto-encoder based on a multi-layer perceptron (MLP). First, a neural network for the classifier was constructed to distinguish the letters ‘n’, ‘v’, and ‘z’, which was composed of 3×3 black-and-white pixels (**Fig. 5a**). It was composed of 9 input layers labeled with ‘i_1’ to ‘i_9’, which correspond to each pixel and 3 output layers labeled with ‘O_n’, ‘O_v’ and ‘O_z’ that are corresponding to each letter (**Fig. 5b**). The circuit diagram for the classifier is shown in **Supplementary Fig. 13**. Note that the output neurons were connected to each other to enable the lateral inhibition. According to the output voltage of the output neurons, each letter was identified. First spiking occurred in

the first neuron for the input of ‘n’, the second neuron was for the input of ‘v’, and the third neuron was for the input of ‘z’. It should be noted that the multi-state properties of the single transistor neuron play an important role in recognizing a pattern. First, it was confirmed that the unwanted spiking was inhibited by the inhibitory neurons prior to reaching the $V_{T,firing}$, which can enhance the energy efficiency of the neural network. Second, it was verified that the pattern was well recognized by appropriately tuning the $V_{T,firing}$, even if the synaptic weight was changed abnormally. This feature can enhance the reliability of the neural network **(Supporting Information 14)**.

In order to improve the recognition rate of an image, an auto-encoder is commonly used.[46] The auto-encoder can remove the effect of noisy input and reconstruct the image by encoding the image and decoding it again. As shown in **Fig. 5c**, we implemented the auto-encoder by use of the MLP network with one middle layer. The input layer and the output layer were composed of 9 neurons, and each layer represented each pixel. After encoding three letters in the first perception, the information of each pixel was newly decoded in the second perception. A circuit diagram for the auto-encoder is shown in **Supplementary Fig. 15**. It should be noted that the inhibitory function of the single transistor neuron allowed the auto-encoder operation. In more detail, the middle neurons were connected to each other to enable lateral inhibition, and hence the noisy signal could be removed. Receiving the signal from the middle neurons, some output neurons expressed spiking while others were inhibited. The excited output neuron was decoded as a black pixel, while the inhibited output neuron was decoded as a white pixel, as shown in **Fig. 5c**. As a result, noisy input images became clearer *via* the image reconstruction by the auto-encoder.

Face recognition with software simulation

Using the hardware-based circuit simulation, off-chip learning that is applicable to inference operation with fixed weights of the synapses was implemented. On the other hand, on-chip learning is also possible by using additional circuits. With the aid of a MATLAB software simulation, a network capable of face recognition through on-chip learning was explored. A fully connected two-layer spiking neural network (SNN) consisting of 32×32 input neurons, 20 neurons in a middle layer, and three output neurons was designed, as shown in **Fig. 6a**. The measured neuron-synapse characteristics were reflected to the simulation based on the circuit diagram of **Fig. 6b**. From the Yale Face Database, nine training images composed of 32×32 pixels were selected (**Fig. 6c**).[47] After clustering from an unsupervised crossbar, the classification was evaluated by a supervised crossbar. Neuronal output was converted through a waveform generator to make a proper pulse shape (**Supplementary Fig. 16a**). In addition, synaptic weight updates, which depend on the time difference between the pre-synaptic pulse (V_{pre}) and post-synaptic pulse (V_{post}) according to a learning rule of spike timing dependent plasticity (STDP), were made.[26,30] It should be emphasized that such circuits for waveform generation can be co-integrated on the same plane with neurons and synapses by standard CMOS fabrications. Also, it is noteworthy that lateral inhibition of the output neurons was enabled for efficient learning and inference by the WTA. After the

training, the conductance of the synapses was determined, as shown in the visual map diagram of the synapse array (**Fig. 6d**). The recognition rate was evaluated with a test set containing 24 images of three people. As a result, a recognition rate of 95.8 % was achieved for 'after training with the lateral inhibition' and that of below 60 % was observed for 'after training without the lateral inhibition,' as shown in **Fig. 6e**. Unless the lateral inhibition was applied, a high level recognition was not performed because the global weight updates were performed *via* the firing of all engaged neurons. In addition, even though the conductance of the synapses were abnormally changed by process-induced variability or endurance problems, the recognition failure was prevented by the $V_{T,firing}$ modulation (**Fig. S16d**). These results prove that the reliable neural network can be implemented by the multi-state single transistor neuron.

Conclusion

Completely CMOS-based neuromorphic hardware with high scalability was fabricated by the co-integration of single transistor-based neurons and synapses that are homotypic. The charge trapping layer intercalated in the neurons and synapses allows multi-states. They were used to control the excitatory/inhibitory function and to modulate the firing threshold voltage for the neurons. They were also utilized to determine the weight for the synapses. The single transistor neuron consumed peak power of 120 nW, which was 10 to 10^3-fold reduced relative to the consumption of the memristor neuron (**Supplementary Information 17**). Because the neuron and the synapse have exactly the same structure, they were simultaneously integrated on the same plane at the same time with the same fabrications. This feature permits improvement of packing density, reduction of chip cost, and simplification of the fabrication procedures. In addition, it is possible to co-integrate with additional CMOS circuits for processing units, peripheral interfaces, memory, clocking circuits, and input/output due to the same *in-situ* CMOS fabrications.

Methods

Fabrication: Neurons and synapses with the same SONOS structure, which had a tunneling oxide (SiO_2) of 3 nm, a charge trap nitride (Si_3N_4) of 6 nm, and a blocking oxide (SiO_2) of 8 nm, were fabricated. They were interconnected through metallization (Ti/TiN/Al) using a standard Si CMOS process. See **Supplementary Information 1** for details of the fabrication process.

Electrical characterization: Electrical characteristics of the co-integrated neurons and synapses were measured using a B1500 semiconductor parameter analyzer (Agilent Technologies). *I-V* characteristics of the neuron and synapse were measured by voltage source current measurement (VSCM) mode, and the spiking characteristic of the neuron was measured by current source voltage measurement (CSVM) mode. A semiconductor pulse generator unit (SPGU) was used to control the excitatory/inhibitory state and the firing threshold voltage of the neuron, as well as the weight of the synapse. The leaky characteristic of the neuron was measured using a Keithley 6221 current pulse source (Keithley). The source current was measured using a TDS 744A oscilloscope (Tektronix).

TEM analysis: TEM images were taken using FE-STEM (HD-2300A) by Hitachi High-Technologies Corporation.

SEM analysis: SEM images were taken using CD-SEM (S-9260A) by Hitachi High-Technologies Corporation.

Device simulation: Device simulations for the analysis of the neuron characteristics were performed using a TCAD Sentaurus simulator (Synopsys). All the device parameters were set as the closest values obtained from the SEM and TEM images.

Hardware-based circuit simulation: Circuit simulations for the letter pattern recognition were performed using LTspice software (Analog Devices). Neurons were modeled with a capacitor and a threshold switch, wherein the parasitic capacitance (C_{par}) and firing threshold voltage ($V_{T,firing}$) were extracted from the measured spiking characteristics of the neuron. Synapses were modeled with a three-terminal MOSFET, in which the device parameters were set as the closest values obtained from the SEM and TEM images. The weight of the synapses was controlled by changing the threshold voltage (V_T) of the MOSFET.

Software-based simulation: Software simulations for the face image recognition were performed using MATLAB. Spiking characteristics of the neurons and P-D characteristics of the synapses were reflected in the simulation.

Declarations

Acknowledgements

This work was supported by National Research Foundation (NRF) of Korea, under Grant 2018R1A2A3075302, 2019M3F3A1A03079603 and 2017R1A2B3007806, in part by the IC Design Education Center (EDA Tool and MPW).

Author Contributions

J.-K. Han and Y.-K. Choi conceived the idea and designed the experiments. J.-K. Han and D. Yoo fabricated the devices. J.-K. Han, M.-S. Kim, and J.-M. Yu performed the electrical measurements and data analysis. J.-K. Han performed the circuit simulation. J. Oh performed the software simulation. G.-J. Yun performed the device simulation. J.-K. Han wrote the manuscript. S.-Y. Choi and Y.-K. Choi supervised the research. All authors discussed the results and commented on the manuscript.

Data availability

The data that support the findings of this study are available from the corresponding author upon reasonable request.

Code availability

The codes used for plotting the deep neural network simulation data are available from the corresponding author on reasonable request.

References

1. Silver, D. et al. Mastering the game of Go without human knowledge. *Nature* **550**, 354 (2017).
2. Hadsell, R. et al. Learning long-range vision for autonomous off-road driving. *J. Field Robot.* **26**, 120–144 (2009).
3. Russakovsky, R. et al. ImageNet large scale visual recognition challenge. *Int. J. Comput. Vis.* **115**, 211–252 (2015).
4. Abbott, L. & Regehr, W. Synaptic computation. *Nature* **431**, 796–803 (2004).
5. Merolla, P. A. et al. A million spiking-neuron integrated circuit with a scalable communication network and interface. *Science* **345**, 668–673 (2014).
6. Qiao, N. et al. A reconfigurable on-line learning spiking neuromorphic processor comprising 256 neurons and 128 k synapses. *Frontiers Neurosci.* **9**, 1–17 (2015).
7. Burr, G. W. et al. Neuromorphic computing using non-volatile memory. *Adv. Phys.* **2**, 89–124 (2016).
8. Marković, D., Mizrahi, A., Querlioz, D. et al. Physics for neuromorphic computing. *Nat Rev Phys.* **2**, 499–510 (2020).
9. Indiveri, G. et al. Neuromorphic silicon neuron circuits. *Frontiers Neurosci.* **5**, 1–23 (2011).
10. Chu, M. et al. Neuromorphic hardware system for visual pattern recognition with memristor array and CMOS neuron. *IEEE Transactions on Industrial Electronics* **62**, 2410–2419 (2014).
11. Ebong, I. E. et al. CMOS and memristor-based neural network design for position detection. *Proceedings of the IEEE* **100**, 2050–2060 (2011).
12. Prezioso, M. et al. Training and operation of an integrated neuromorphic network based on metal-oxide memristors. *Nature* **521**, 61–64 (2015).
13. Sheridan, P. M. et al. Sparse coding with memristor networks. *Nat. Nanotechnol.* **12**, 784–789 (2017).
14. Cai, F. et al. A fully integrated reprogrammable memristor-CMOS system for efficient multiply-accumulate operations. *Nat. Electron.* **2**, 290–299 (2019)
15. An, H. et al. Monolithic 3D neuromorphic computing system with hybrid CMOS and memristor-based synapses and neurons. *Integration* **65**, 273–281 (2019).
16. Tuma, T. et al. Stochastic phase-change neurons. *Nature Nanotechnol.* **11**, 693–699 (2016).
17. Han, J. & Meyyappan, M. Leaky integrate-and-fire biristor neuron. *IEEE Electron Device Lett.* **39**, 1457–1460 (2018).
18. Wang, Z. et al. Fully memristive neural networks for pattern classification with unsupervised learning. *Nat. Electron* **1**, 137–145 (2018).
19. Woo, J., Wang, P. & Yu, S. Integrated crossbar array with resistive synapses and oscillation neurons. *IEEE Electron Device Lett.* **40**, 1313–1316 (2019).

20. Duan, Q. et al. Spiking neurons with spatiotemporal dynamics and gain modulation for monolithically integrated memristive neural networks, *Nat. Commun* **11**, 1–13 (2020).
21. Kim, K.-H. et al. A functional hybrid memristor crossbar-array/CMOS system for data storage and neuromorphic applications. *Nano Lett.* **12**, 389–395 (2012).
22. Indiveri, G. et al. Integration of nanoscale memristor synapses in neuromorphic computing architectures. *Nanotechnology* **24**, 384010 (2013).
23. Bayat, F. M. et al. Implementation of multilayer perceptron network with highly uniform passive memristive crossbar circuits. *Nat. Commun.* **9**, 2331 (2018).
24. Adam, G. C., Khiat, A. & Prodromakis, T. Challenges hindering memristive neuromorphic hardware from going mainstream. *Nat. Commun.* **9**, 5267 (2018).
25. Woo, S. Y. et al. Implementation of homeostasis functionality in neuron circuit using double-gate device for spiking neural network. *Solid-State Electron.* **165**, 107741 (2020).
26. Querlioz, D., Bichler, O., Dollfus, P. & Gamrat, C. Immunity to device variations in a spiking neural network with memristive nanodevices. *IEEE Trans. Nanotechnol.* **12**, 288–295 (2013).
27. Bartolozzi, C., Nikolayeva, O. & Indiveri, G. Implementing homeostatic plasticity in VLSI networks of spiking neurons. *Proc. 15th IEEE Int. Conf. Electron. Circuits Syst. ICECS*, 682–685 (2008).
28. Wang, Z. et al. Experimental demonstration of ferroelectric spiking neurons for unsupervised clustering. *IEEE Int. Electron Devices Meeting (IEDM)*, 13.3.1–13.3.4 (2018).
29. Luo, J. et al. Capacitor-less stochastic leaky-FeFET neuron of both excitatory and inhibitory connections for SNN with reduced hardware cost. *IEEE Int. Electron Devices Meeting (IEDM)*, 6.4.1–6.4.4 (2019).
30. Kim, S., Choi, B., Lim, M., Yoon, J., Lee, J., Kim, H. D. & Choi, S. J. Pattern recognition using carbon nanotube synaptic transistors with an adjustable weight update protocol. *ACS Nano* **11**, 2814–2822 (2017).
31. Han, J.-K. et al. Mimicry of excitatory and inhibitory artificial neuron with leaky integrate-and-fire function by a single MOSFET. *IEEE Electron Device Lett.* **41**, 208–211 (2020).
32. Han, J.-K. et al. Independently accessed double-gate for excitatory-inhibitory function and tunable firing threshold voltage. *IEEE Electron Device Lett.* **41**, 1157–1160. (2020).
33. Chen, C.–D. et al. Single-transistor latch in SOI MOSFETs. *IEEE Electron Device Lett.* **9**, 636–638 (1988).
34. Han, J. & Meyyappan, M. Trigger and self-latch mechanisms of n-p-n bistable resistor. *IEEE Electron Device Lett.* **35**, 387–389 (2014).
35. Moon, J.-B. et al. A bandgap-engineered silicon-germanium biristor for low-voltage operation. *IEEE Trans. Electron Devices* **61**, 2–7 (2014).
36. Han, J.-K. et al. One biristor-two transistor (1B2T) neuron with reduced output voltage and pulsewidth for energy-efficient neuromorphic hardware, *IEEE Trans. Electron Devices*, in press.

37. Suri, M. Advances in neuromorphic hardware exploiting emerging nanoscale devices. *Springer: New York.* (2017).
38. Futatsubashi, G., Sasada, S., Tazoe, T., & Komiyama, T. Gain modulation of the middle latency cutaneous reflex in patients with chronic joint instability after ankle sprain. *Clinical Neurophysiology* **124**, 1406–1413 (2013).
39. Chance, F. S., Abbott, L. F., & Reyes, A. D. Gain modulation from background synaptic input. *Neuron* **35**, 773–782 (2002).
40. Silver, R. A. Neuronal arithmetic. *Nat. Rev. Neurosci.* 1**1**, 474–489 (2010).
41. Wang, X. et al. Perisaccadic receptive field expansion in the lateral intraparietal area. *Neuron* **90**, 400–409 (2016).
42. Agmon-Snir, H., Carr, C. E. & Rinzel, J. The role of dendrites in auditory coincidence detection. *Nature* **393**, 268–272 (1998).
43. Joris, P. X., Smith, P. H. & Yin, T. Coincidence detection in the auditory system: 50 years after jeffress. *Neuron* **21**, 1235–1238 (1998).
44. Carr, C. E., & Konishi, M. Axonal delay lines for time measurement in the owl's brainstem. *Proceedings of the National Academy of Sciences* **85**, 8311–8315 (1988).
45. Engel, A. K., Fries, P. & Singer, W. Dynamic predictions: oscillations and synchrony in top-down processing. *Nat. Rev. Neurosci.* **2**, 704–716 (2001).
46. Mennel, L. et al. Ultrafast machine vision with 2D material neural network image sensors. *Nature* **579**, 62–66 (2020).
47. Belhumeur, P. N.; Hespanha, J. P.; Kriegman, D. J. Eigenfaces vs. fisherfaces: recognition using class specific linear projection. *IEEE Trans. Pattern Anal. Mach. Intell.* **19**, 711–720 (1997).

Figures

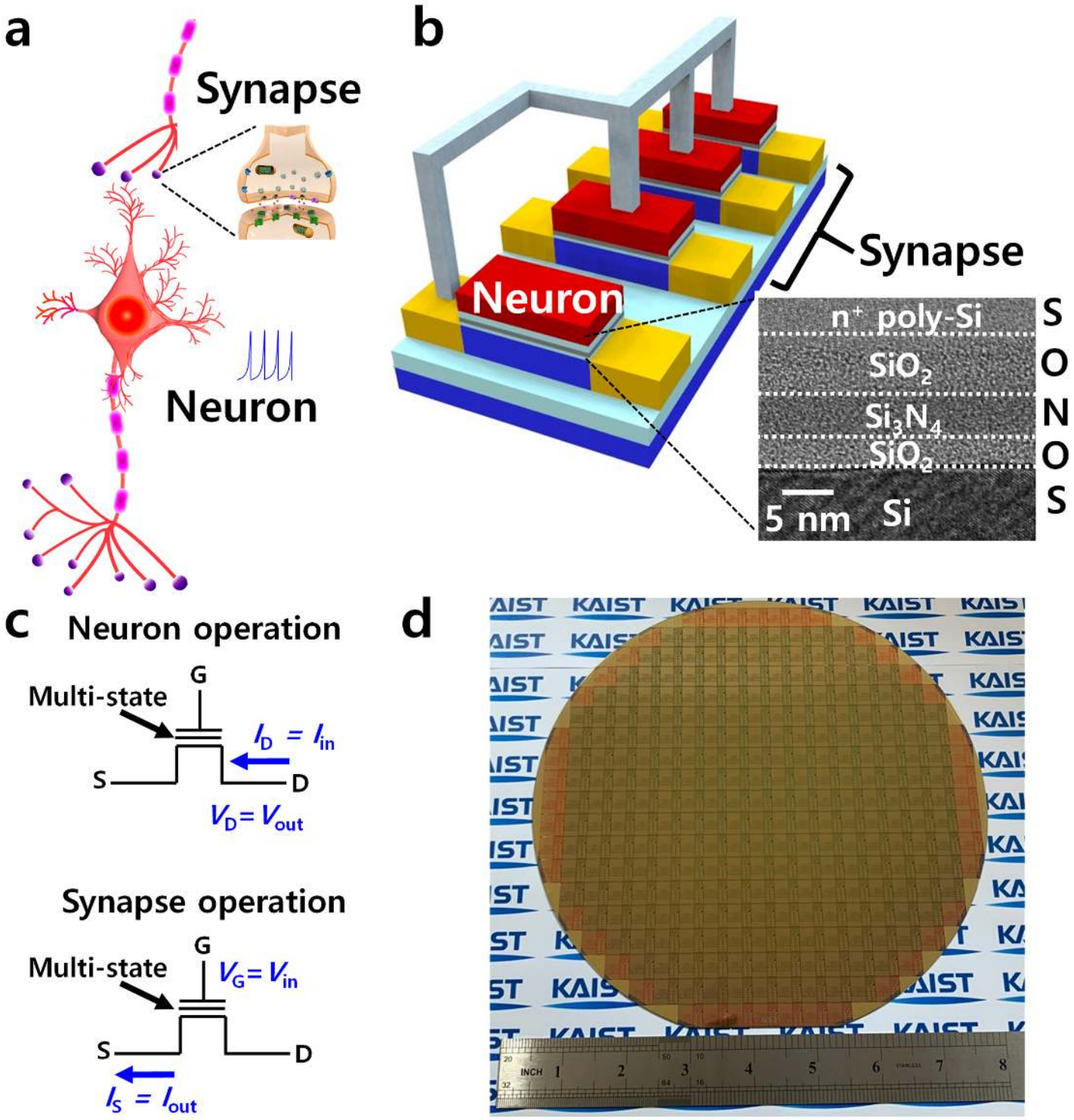

Figure 1

Concept of co-integrated single transistor neurons and synapses. a, Schematic of biological neuron and synapse. 1011 neurons and 1015 synapses are densely interconnected in human brain. b, Schematic of co-integrated single transistor neurons and synapses. They have exactly the same SONOS structure, which includes a charge trap layer (Si3N4) in the gate dielectrics as shown in the cross-sectional TEM image. They are fabricated with the same fabrications and connected through metallization. c, Operation

scheme of the neuron and synapse. The input and output of the neuron are current and voltage, respectively, while those of the synapse are voltage and current. d, Fabricated 8 inch wafer in which single transistor neurons, synapses, and additional CMOS circuits were co-integrated. It was fabricated with 100 % standard Si CMOS fabrications.

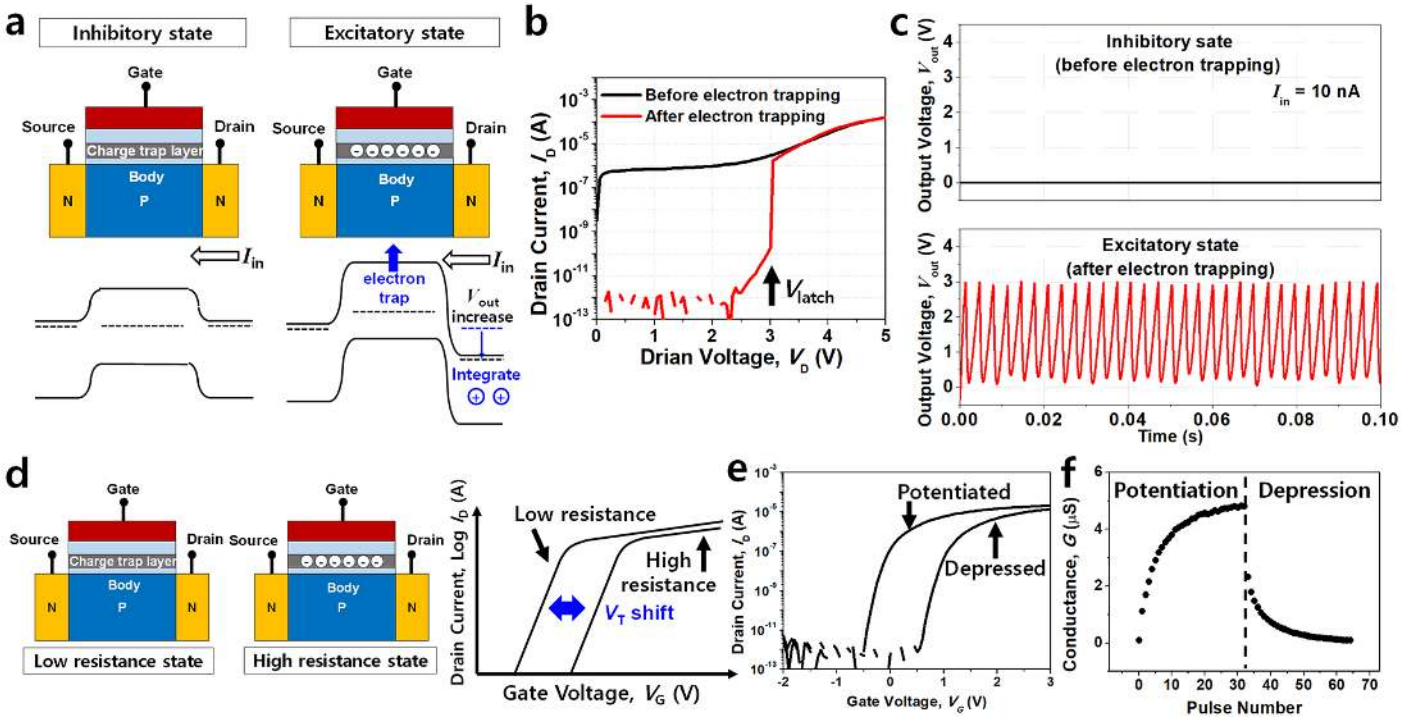

Figure 2

Unit device characteristics of single transistor neuron and synapse. a, Operation principle of the single transistor neuron. The excitatory and inhibitory state of the neuron are determined by electron trapping in the nitride. b, Output characteristic (ID-VD) of the fabricated single transistor neuron. The single transistor latch (STL) phenomenon that allows threshold switching near Vlatch was observed only after electron trapping (excitatory). c, Spiking characteristics of the fabricated single transistor neuron. The neuronal spiking by LIF operation was excited after electron trapping, while it was inhibited before electron trapping. d, Operation principle of the single transistor synapse. The weight of the synapse can be adjusted by controlling the trapped charge density in the nitride. e, Transfer characteristic (ID-VG) of the fabricated single transistor synapse after potentiation and depression. Threshold voltage (VT) was shifted leftward after potentiation and rightward after depression. f, Potentiation-depression (P-D) characteristic of the fabricated single transistor synapse. 32 levels of the conductance state were secured (5 bits).

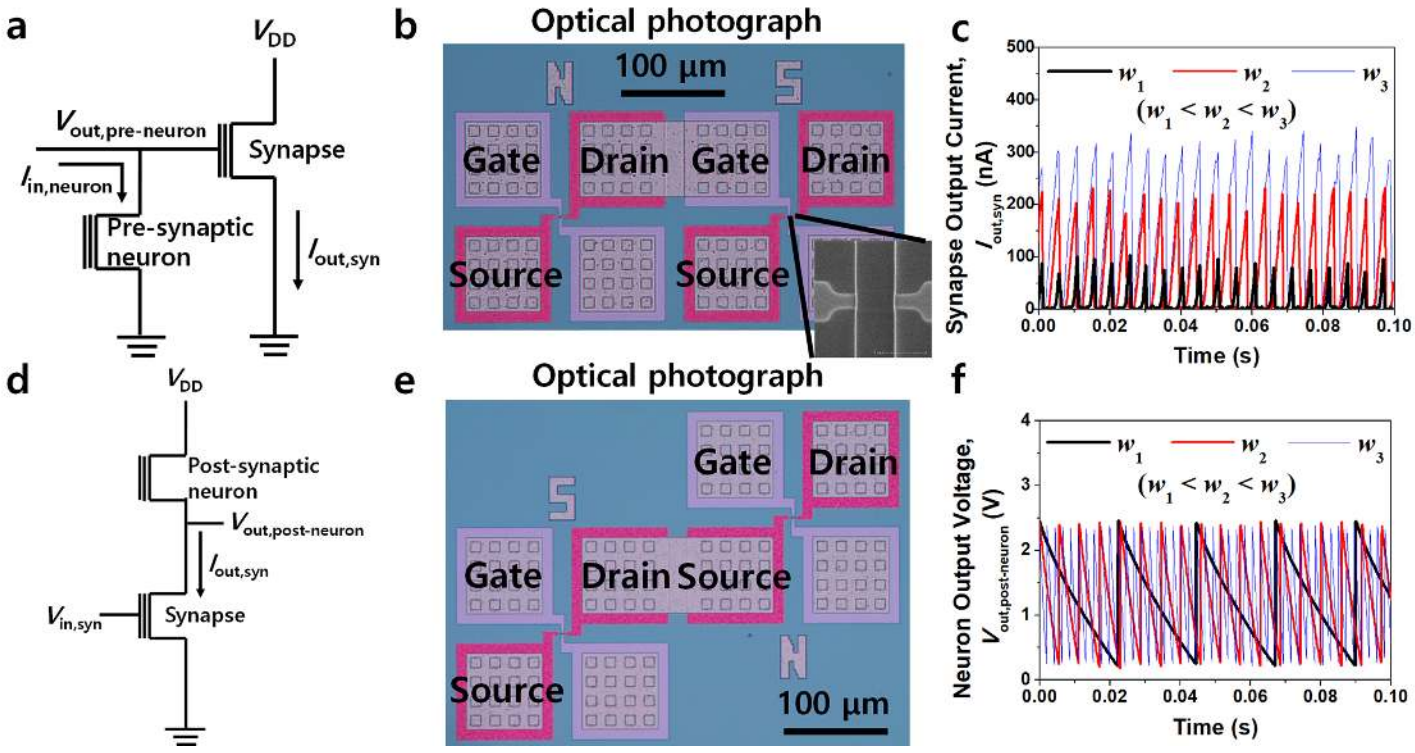

Figure 3

Co-integrated single transistor neuron and synapse. a, Circuit diagram of pre-synaptic neuron and transmitted synapse connection in the pre-layer of neural network. The output voltage of the pre-synaptic neuron (Vout,pre-neuron) is transmitted to the gate of the synapse. b, Fabricated pre-synaptic neuron and transmitted synapse interconnected through metallization. c, Measured synapse output current (Iout,syn) as a function of synaptic weight. The level of Iout,syn became higher when the synaptic weight was larger. d, Circuit diagram of transmitting synapse and post-synaptic neuron in the post-layer of neural network. The current of the transmitting synapse is applied to the source of the post-synaptic neuron. e, Fabricated transmitting synapse and post-synaptic neuron interconnected through metallization. f, Measured neuron output voltage (Vout,post-neuron) as a function of synaptic weight. The spiking frequency (f) of Vout,post-neuron became higher when the synaptic weight was larger.

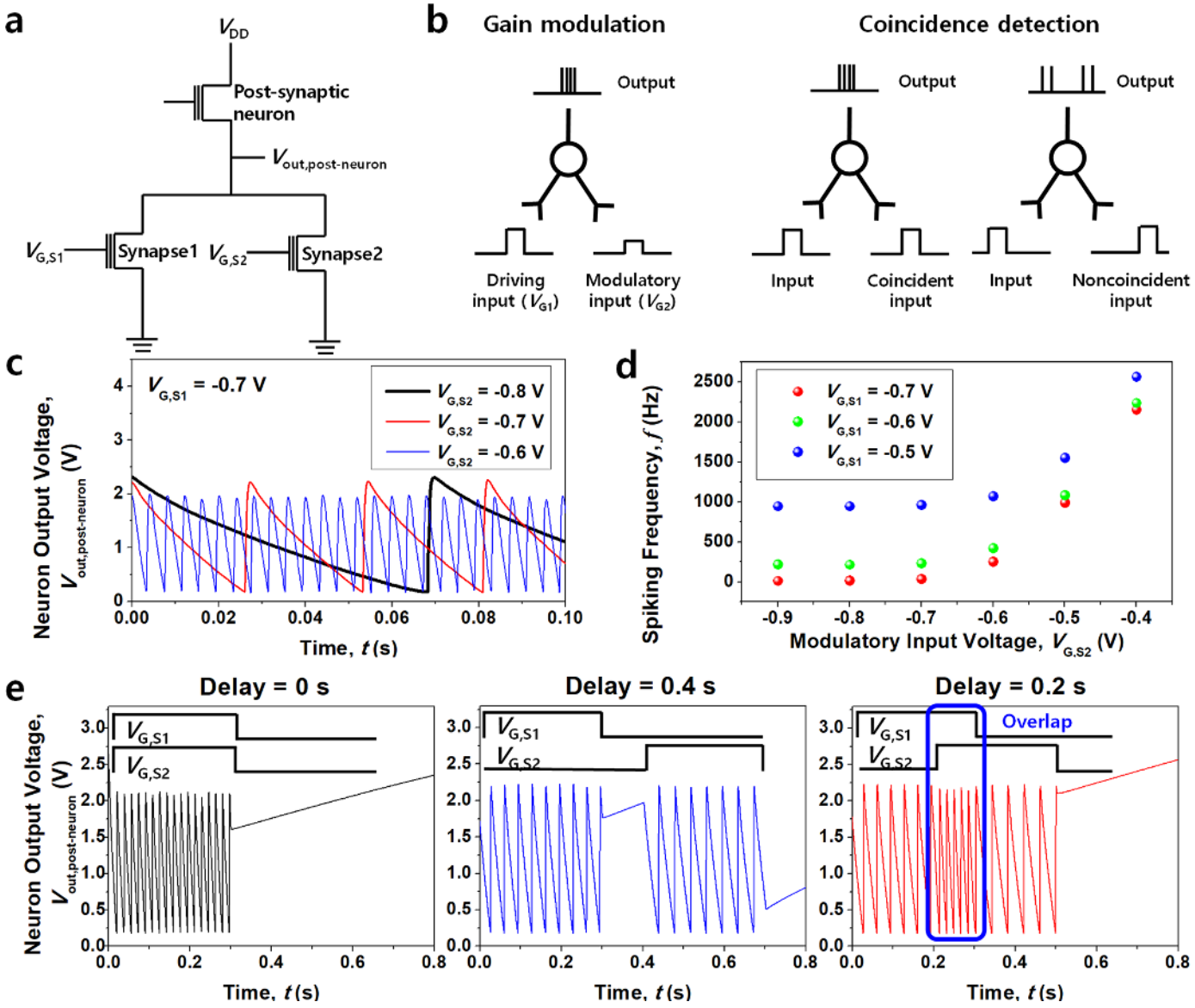

Figure 4

Gain modulation and coincidence detection by co-integrated single transistor neuron and synapses. a, Circuit diagram of connected two transmitting synapses and one post-synaptic neuron for gain modulation and coincidence detection. b, Schematic diagram of gain modulation and coincidence detection. Neuronal output can be determined by the modulatory input as well as the driving input, and the coincidence of the two inputs can be detected from the neuronal output. c, Spiking characteristics of the post-synaptic neuron depending on the modulatory input voltage (VG,S2) when the driving input voltage (VG,S1) was fixed. The spiking frequency (f) was increased as the VG,S2 was increased because the input current to the post-synaptic neuron was increased. d, f as a function of the VG,S2 at various VG,S1. e, Spiking characteristics of the post-synaptic neuron depending on the delay between the two signals. f was larger when two signals became more synchronized. When two signals overlapped for a certain period of time, the f was increased only in the overlap region.

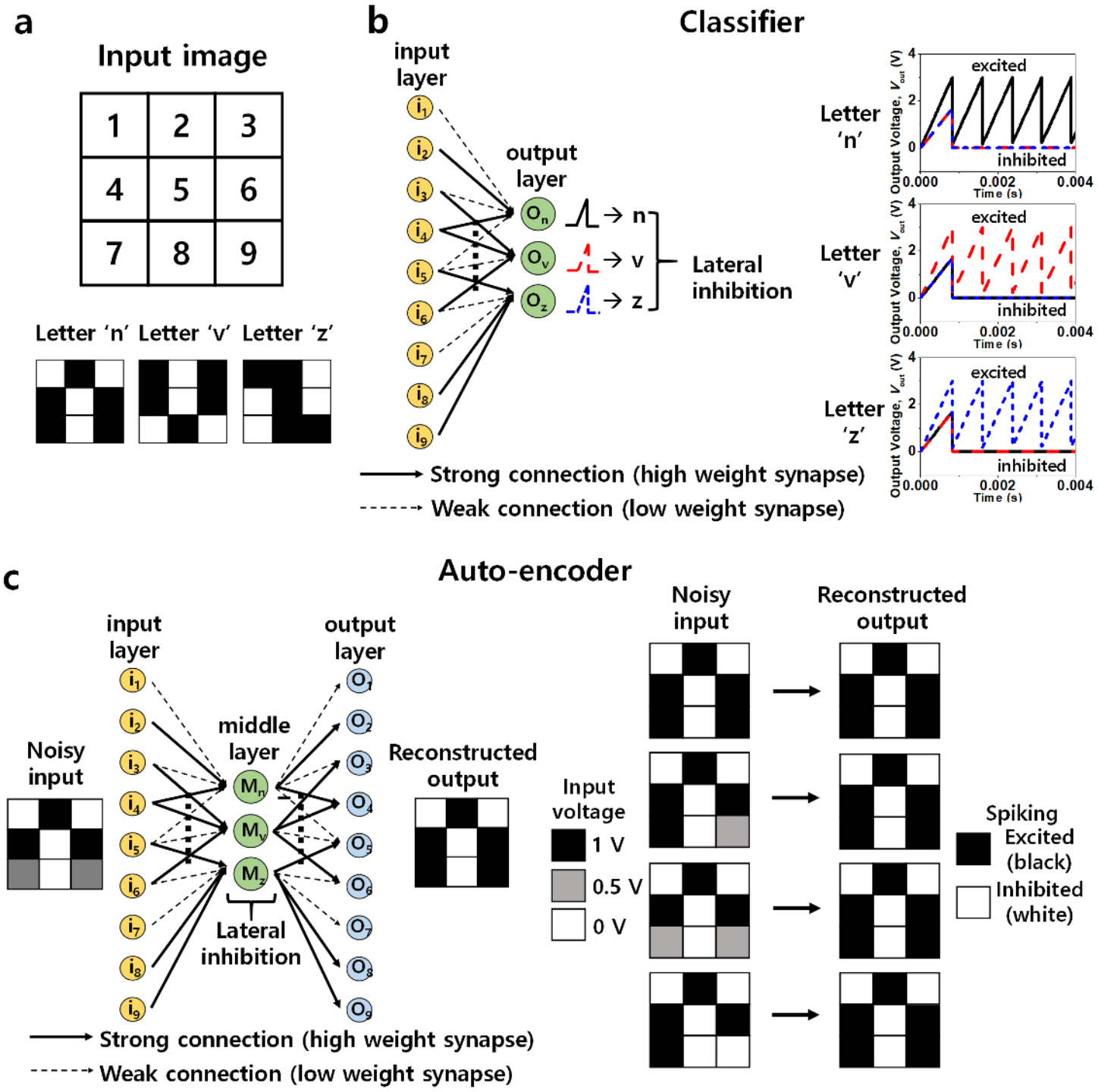

Figure 5

Letter recognition with hardware-based circuit simulation by reflecting the measured characteristics of single transistor neuron and synapse. a, Input image of the 3×3 pixel letter pattern. b, Single-layer perceptron (SLP) network for a classifier and classification results. Each input layer represents each pixel, and each output layer represents each letter. Classification determined by which neuron expressed spiking first was performed. All other neurons except the first spiked output neuron were laterally inhibited. c, Multi-layer perceptron (MLP) network for an auto-encoder and its encoding/decoding results. Each input layer represents each pixel of noisy input, and each output layer represents each pixel of

reconstructed output by the auto-encoder. The output neuron in which spiking was excited could be newly decoded as a black pixel, and the output neuron in which spiking was inhibited could be newly decoded as a white pixel to reconstruct a clearer image from a blurred noisy pattern.

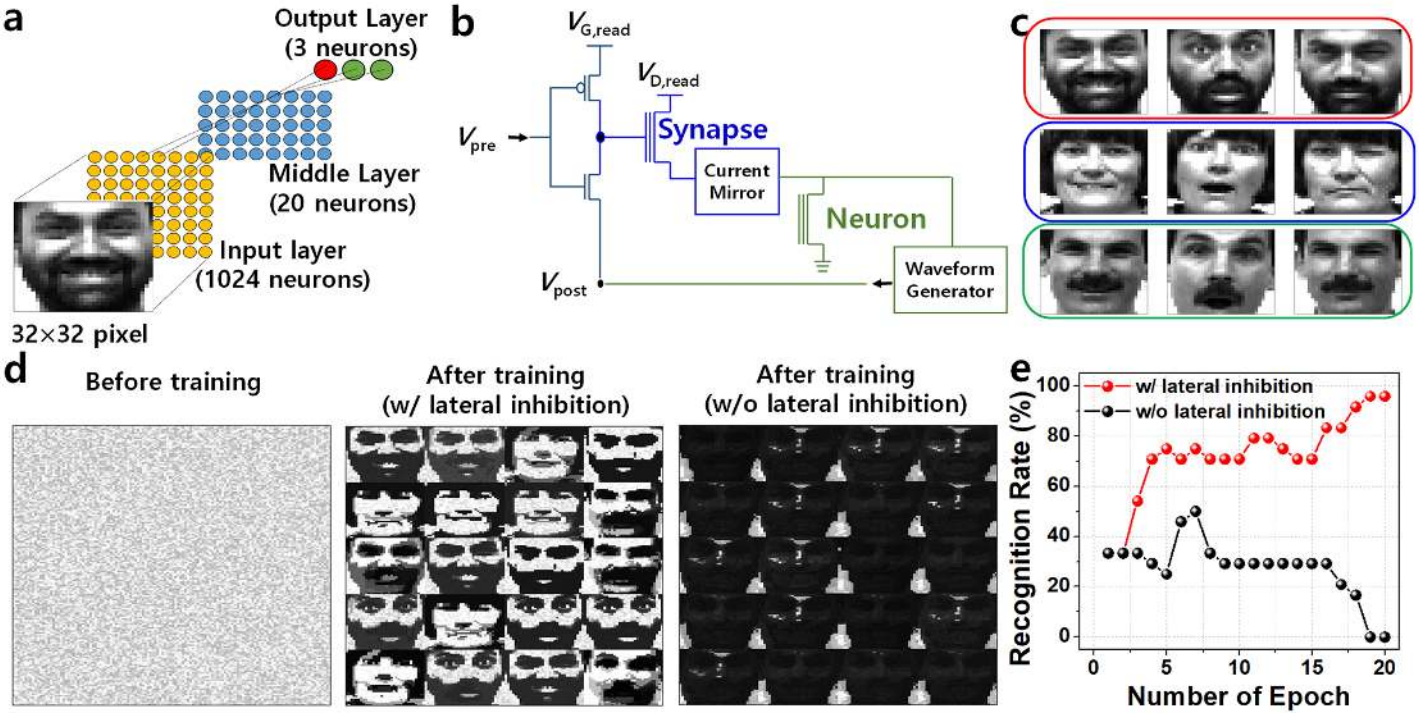

Figure 6

Face recognition with software-based simulation by reflecting the measured characteristics of single transistor neuron and synapse. a, Spiking neural network (SNN) for face recognition. The input layer is composed of 1024 neurons that represent each pixel, the middle layer is composed of 20 neurons, and the output layer is composed of three neurons that represent each person's face. b, Simplified circuit diagram to represent the connection of neuron-synapse. Neuronal output is converted through a waveform generator to make a proper pulse shape applied to the synapse for STDP learning. c, Nine training images of three people. d, Visual map of the synapse array to represent the conductance of the synapses, 'before training', 'after training with lateral inhibition', and 'after training without lateral inhibition'. e, Comparison of recognition rate depending on the number of training epochs between 'after training with lateral inhibition' and 'after training without lateral inhibition'. Higher recognition rate is achieved with the inhibitory function of the neurons.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryInformation.docx