Electronic Supplementary Information

An Electro-Photo-Sensitive Synaptic Transistor for Edge Neuromorphic Visual System

Nian Duan, [1] Yi Li, [1]* Hsiao-Cheng Chiang, [2] Jia Chen, [1] Wen-Qian Pan, [1] Ya-Xiong Zhou, [1] Yu-Chieh Chien, [2] Yu-Hui He, [1] Kan-Hao Xue, [1] Gang Liu,[3] Ting-Chang Chang,[2]* and Xiang-Shui Miao[1]*

[1] Wuhan National Research Center for Optoelectronics, School of Optical and Electronic Information, Huazhong University of Science and Technology, Wuhan 430074, China

[2] Department of Physics, National Sun Yat-Sen University, Kaohsiung 80424, Taiwan.

[3] School of Chemistry and Chemical Engineering, Shanghai Jiao Tong University, Shanghai 200240, China.

Corresponding Author

* E-mail: liyi@hust.edu.cn, tcchang3708@gmail.com, miaoxs@hust.edu.cn

1. Electrical properties of the IGZO synaptic transistors

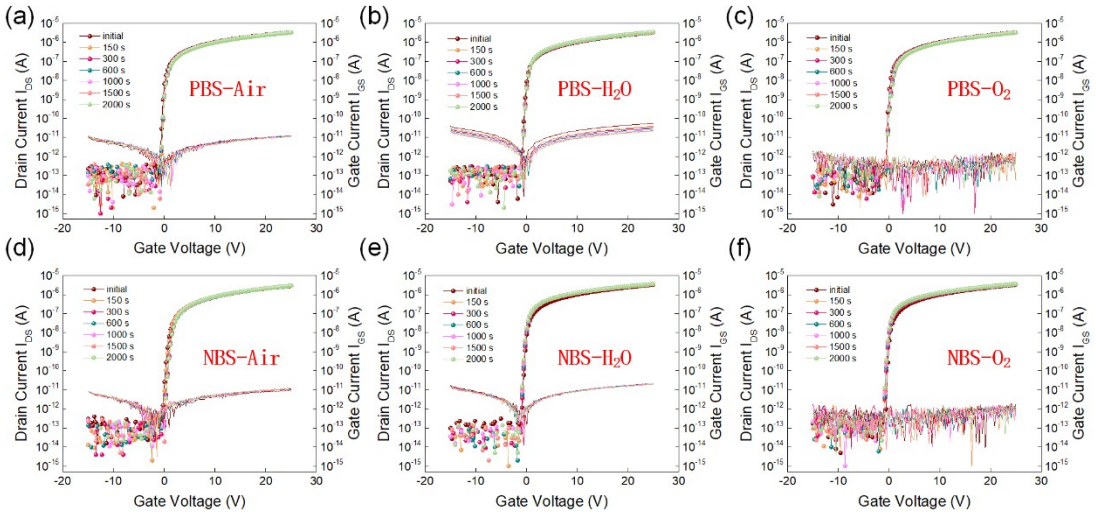

Fig. S1 The transfer characteristics curves of the bottom gate under the negative and positive stress voltages show trivial and no harmful degradation in different ambient, including the humidity environment (H_2O), oxygen atmosphere (O_2) and air environment (air). The results indicate the high reliability of the device in various environments. The high reliability and stability which is one of the factors that determine whether a device can be applied to a neural network indicates the potential of our device for integration of the neural system. In addition, the stability of the device under the negative bias and illumination stress (NBIS) is one of the key indicators of the of the thin film transistor. In our previous studies, the threshold voltages of the IGZO TFT devices will shift to the negative direction after NBIS. Methods to suppress this phenomenon have been investigated too. Optimizing the device structure (such as fringe filed structure) or applying N_2O plasma treatment to both gate insulator and active layer can play a positive role on improving the stability of the device during NBIS.[1-3]

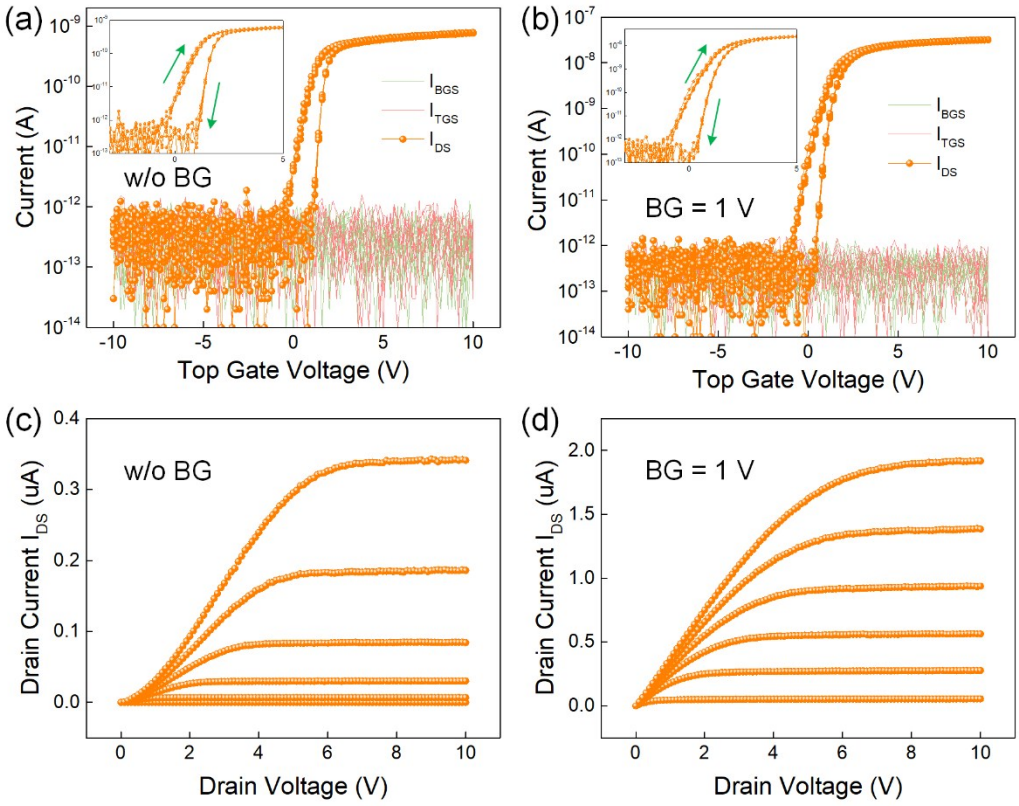

Fig. S2 The electrical properties measured on the top gate of the IGZO thin film transistor. When the device works as a synapse, the presynaptic signal is input from the top gate while the voltage of the bottom gate keeps at V_{sw} which is about 1 V. Hence, the transfer curves and output curves of the top gate was measured at the cases with and without the bottom gate voltage (BG = 1 V) respectively. The transfer curves show obvious clockwise hysteresis which confirms the mechanism of the electron trapping. (c) and (d) shows the output characteristics of the device measured on the top gate. The top gate voltage sweeps in a range of 0 to 10 V with a step of 2 V.

2. Channel current during the positive pulse.

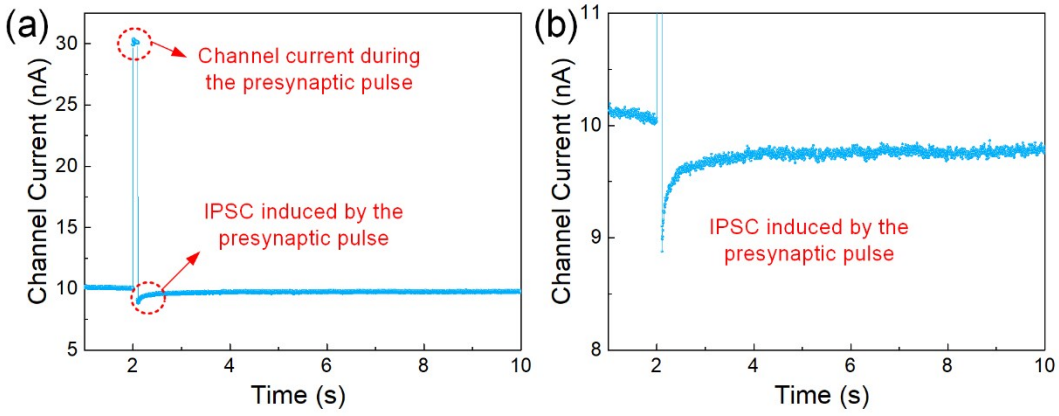

Fig. S3 Channel current during the positive pulse. As (a) shows, the channel current during the presynaptic pulse (5 V, 100 ms) increased and reached about 30 nA. This phenomenon is temporary. After the pulse is finished, the channel current dropped instantly, and the inhibitory postsynaptic current (IPSC) is generated. Compared with the current value during the pulse, the IPSC is relatively small and it will be less obvious if the peak of the current during the pulse is included in the figure. So in our experiment, the current during the pulse is not recorded (Fig. 3(a)).

3. Transition from short-term plasticity (STP) to long-term plasticity (LTP)

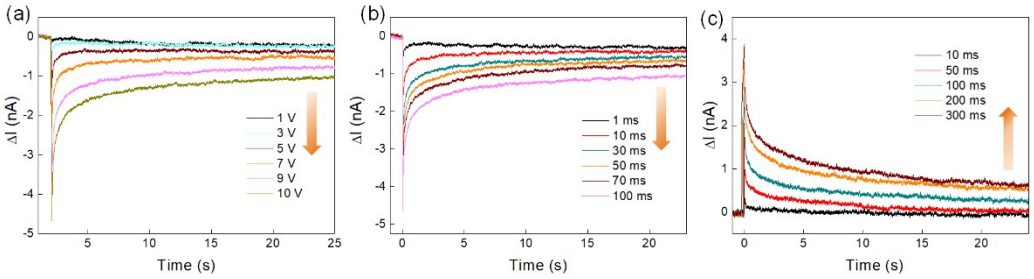

Fig. S4 The IPSCs triggered by the presynaptic spikes became larger as the amplitude (a) and width (b) of the applied voltage increased. The residual IPSCs gradually increased which indicate a trend of transition from STP to LTP. In the same way, when the width of

light pulse increased from 10 ms to 300 ms (c), the residual EPSCs increased, indicating a trend of transition from STP to LTP.

4. **Mathematical model of electron trapping to explain the mechanism of IPSC**

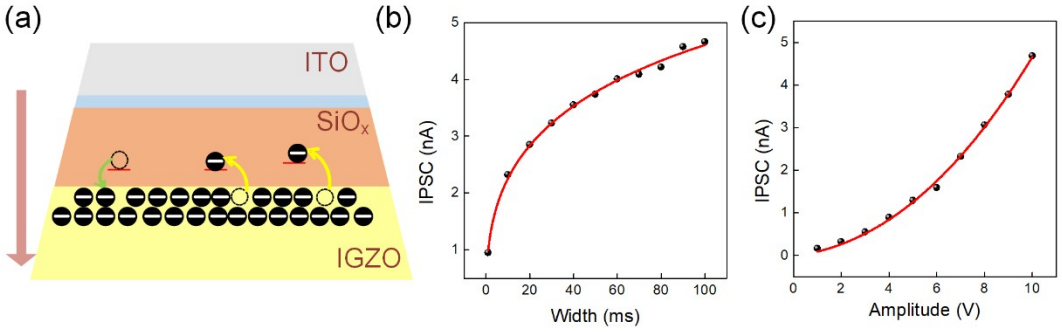

Fig. S5 Mechanism of the generation of IPSC. (a) Schematic of the movement of electrons under the positive electric pulse applied on the top gate. (b) The relationship between IPSC and pulse width fitted with the mathematical model (Equation (3)). (c) The relationship between IPSC and pulse amplitude fitted with the mathematical model (Equation (5)).

The V_{th} shift attributed to charge trapping can be modeled with the following function:

$$\Delta V_{th} = V_0 \left(1 - e^{-\left(\frac{t}{\tau}\right)^{\beta}} \right) \tag{1}$$

Where $V_0 = V_{pre} - V_{th}$ is ΔV_{th} at the infinite time, of which V_{th} is the initial threshold voltage, β is the stretched-exponential exponent, τ is the characteristic trapping time of carriers and t is the stress time, namely the presynaptic pulse width.[4-6]

In our experiments, the IGZO synaptic transistor worked in the linear region with V_{DS} of 0.1 V, and the channel current can be modeled by a simple linear relationship with gate voltage:

$$I_{DS} = \frac{\mu_n C_{ox} V_D W}{L} \left(V_{pre} - V_{th} \right) \tag{2}$$

When the device received the presynaptic pulse, the shift of V_{th} led to a change in the I_{DS}, described as below:

$$\Delta I_{DS} = \mathrm{A}(V_{pre} - V_{th})(1 - e^{-\left(\frac{t}{\tau}\right)^{\beta}}) \quad (3)$$

In which $\mathrm{A} = \frac{\mu_n C_{ox} V_D W}{L}$ is a constant. When the IPSC was modulated by controlling the pulse width t with a fixed amplitude, the relationship between ΔI_{DS} and t accords with the stretched-exponential function (Equation (3)) derived above. The fitting result of the experimental data is in consistent with theoretical model, as shown in Fig. S5b. In this case, the fitting parameter is calculated to be $\tau = 6169$ and $\beta = 0.34$.

The relationship between ΔI_{DS} and presynaptic pulse amplitude V_{pre} can also be modeled. Since the magnitude of presynaptic pulse width t is hundreds of milliseconds, which is far less than τ, by using $e^{-x} \approx 1 - x$ when $x = 1$ and $\tau = aV_{pre}{}^{-\alpha}$ [7], the equation (3) can be simplified as below.[4]

$$\Delta I_{DS} = \mathrm{A}(V_{pre} - V_{th})(\frac{t}{a})^{\beta} V_{pre}{}^{\alpha\beta} \quad (4)$$

As t and other parameters are independent of V_{pre}, ΔI_{DS} can be described as a function of gate bias V_{pre}:

$$\Delta I_{DS} = \mathrm{C}V_{pre}{}^{n+1} - \mathrm{D}V_{pre}{}^{n} \quad (5)$$

Where C, D and n are fitting parameters.

As shown in Fig. S5c, the experimental data was fitted with the function above, and the R-square of this fitting can reach 99.8%. In the fitting, C = 0.015, D = -0.08, and n = 1.31. As an inference of the fitting result, the IPSCs induced by electrical pulses originate from the electron trapping and de-trapping in the top gate insulator.

5. **Convolutional Neural Network (CNN)**

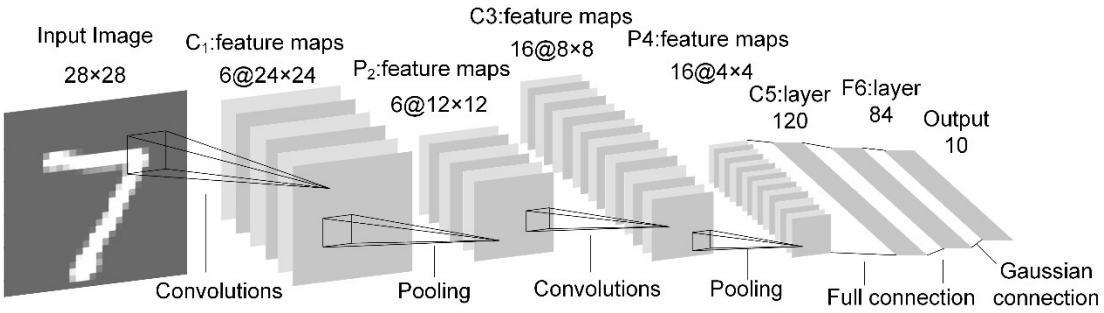

Fig. S6 The LeNet-5 CNN architecture, consisted of two convolutional layers, two average pooling layers and three fully connected layers with ReLU activation function uses the gradient-based algorithm in the learning process. The number of input is 784 since the input image from the MNIST digit handwritten dataset is divided into 28 × 28 grayscale pixels. The convolution operation is a matrix multiplication between the previous input layer and a convolution kernel. In this CNN architecture, 6 different 5 × 5 convolution kernels were used in the first convolution operation to get 6 different 24 × 24 feature maps while the second convolution layer uses 16 convolution kernels to get 16 different 8 × 8 feature maps. The pooling operations alternates with the convolution operations. The number of the 2 × 2 filters used in the first and the second pooling operation is 6 and 16 respectively, and the stride is set to 2. After the operation of second pooling layer, the output is reshaped into a one-dimensional array and then transferred to the fully connected layers. The three-layer fully connected layer contains 120 neurons, 84 neurons and 10 neurons respectively. The output results of the 10 neurons correspond to the learning results.

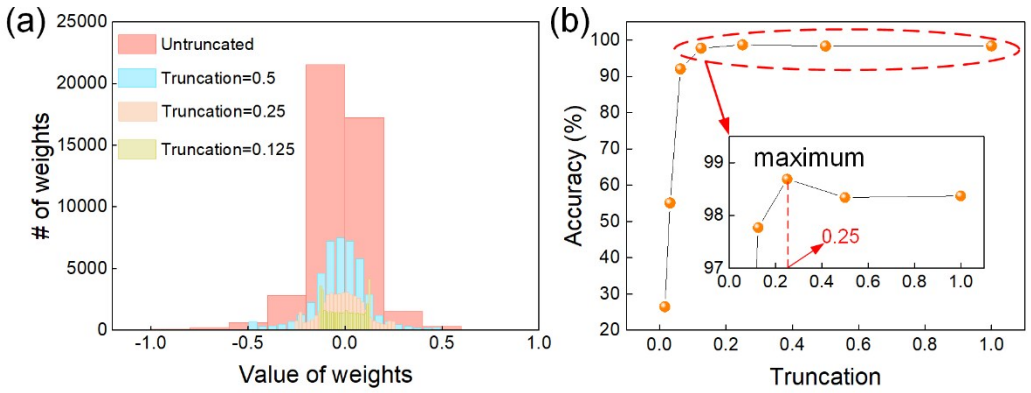

Fig. S7 Since the weight range in actual device is limited, the weights in the neural network

should also be in a range.[8] Determination of the weight truncation range used in the simulation is necessary. (a) Weight distribution in different normalized truncated ranges. (b) The relationship between learning accuracy and the weight truncation. The highest recognition accuracy (98.69%) was achieved when the weight truncation is 0.25. Therefore, a truncation range of [-0.25, 0.25] is adopted in our following simulation.

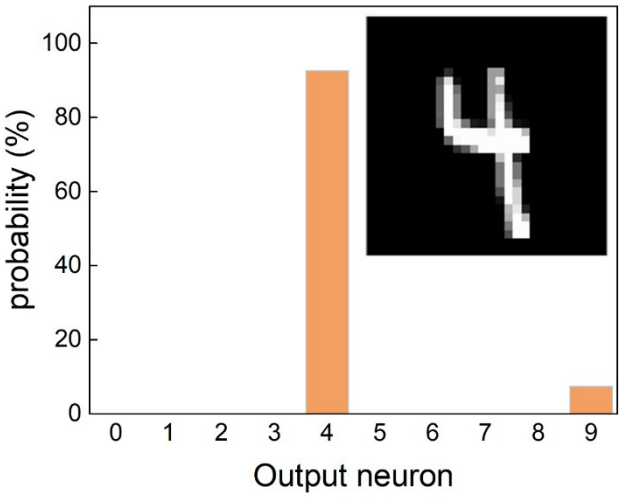

Fig. S8 In the simulation based on standard MNIST dataset, a correctly classified digit from the MNIST test set was chosen randomly and the corresponding probability of each output neuron was shown. The probability distribution map indicates the learning result is correct.

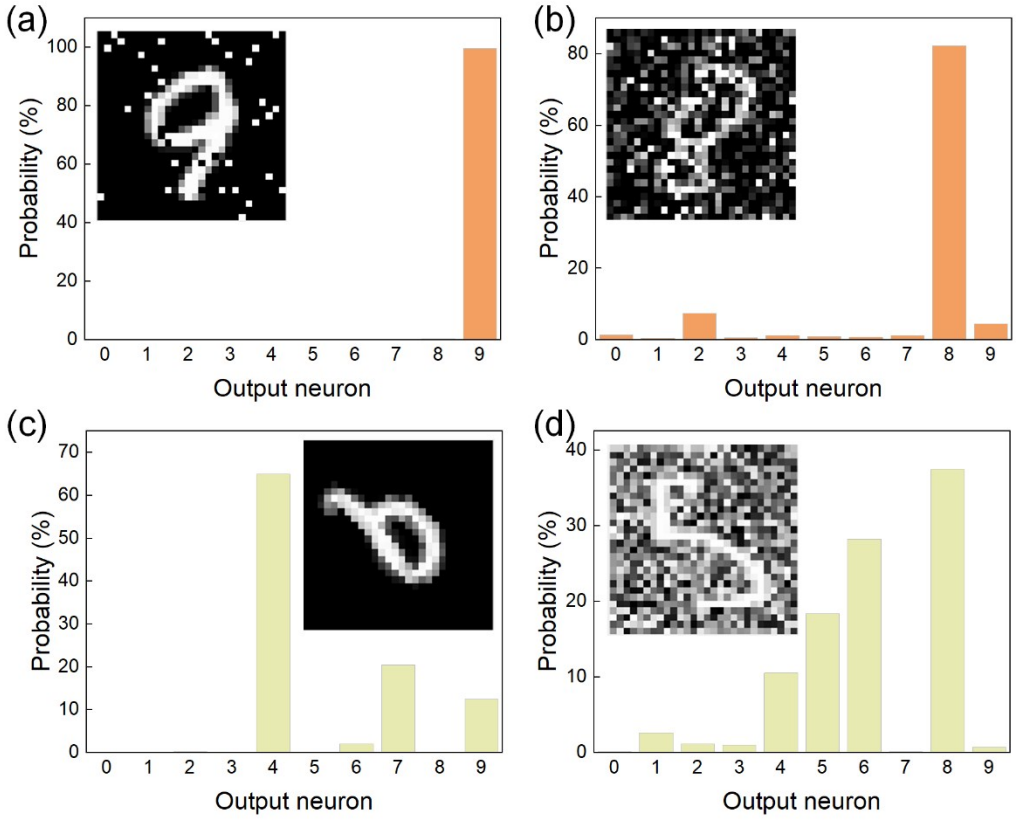

Fig. S9 The probability distribution map of randomly selected digits, when the trained network is subjected to different noisy test datasets: (a) 10% salt & pepper noise, (b) 10% Gaussian noise, (c) MNIST-rot, and (d) MNIST-back-rand datasets. In the cases of (a) and (b), the learning results are correct, while that in (c) and (d) are incorrect.

6. **Calculation of Power Consumption**

In our synaptic transistors, the total energy consumption per spike is defined by the sum of channel resistive energy (E_r), capacitive energy (E_c) and operation pulse energy (E_o). When our device is working as a synapse in the neural network simulation, the parameters of the electric and optical spike are (8 V, 100 ms) and (0.5 W cm^{-2}, 100 ms) respectively.

When the electric pulse (8 V, 100 ms) is applied on the top gate, the E_r is calculated by $V_{DS} \times I_{DS} \times t$ and E_o is calculated by $V_{GS} \times I_{GS} \times t$. The I_{DS} during the electric pulse (8

V, 100ms) is measured to be 36 nA with a read voltage V_{DS} = 0.1 V as Figure S10 shows. And the I_{GS} through the top gate is measured to be less than 1 pA as shown in Fig. S2. In this case, the E_r and E_o are calculated to be 360 pJ and 0.8 pJ, respectively. the E_o is much less than the E_r.

The capacitive energy E_c can be calculated by

$$E_c = \frac{Q_c^2}{2C_{eff}} = \frac{(I_{GS}t)^2}{2C_{eff}} \quad (6)$$

where Q_c is the amount of charge stored by single pulse. I_{GS} and t are the current flowing through top gate-source electrodes and the pulse width, respectively. Further, C_{eff} means the effective capacitance.[9] Since I_{GS} is too small (less than 10^{-12} A), The E_c is calculated to be at a magnitude of fJ, which is too small and can be neglected. So the sum energy consumption under the electric pulse is 360.8 pJ.

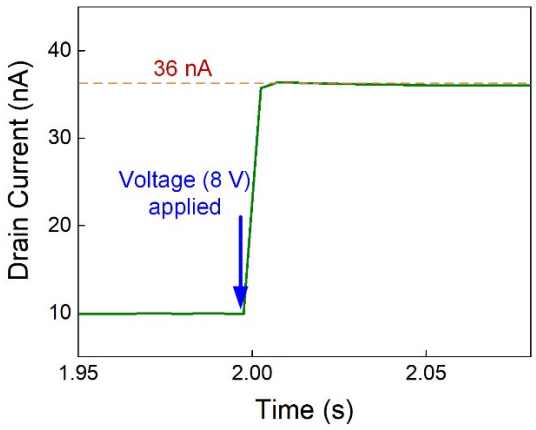

Fig. S10 The channel current during the electric pulse (8 V, 100 ms).

When the device is operated by the light pulse (0.5 W cm^{-2}, 100 ms), the maximum channel current is 12.3 nA (Fig. 4b). Then the E_r is calculated to be 123 pJ. The E_c of light pulse is the same with the that of electric pulse, which is too small and can be neglected. The calculation method of E_o under light pulse is different, and it related to the area of our device. Since the area of our device is 4×10^{-5} cm^2, the E_o=0.5 W $cm^{-2}\times4\times10^{-5}$ $cm^2\times0.1$ s=2 uJ. Hence, the sum energy consumption under the light pulse is about 2 uJ.

7. **Logic operations with the photoelectric synapse**

Apart from neuromorphic computing functions, Boolean logic operations which can perform some specific functions such as judgement function could also be realized based on the synaptic devices. Here, the functional complete Boolean logic implemented in the IGZO synaptic devices is demonstrated. Fig. S11 shows the logic operation scheme and the truth table of *p* RIMP *q* and *p* NIMP *q* logic operations. {RIMP, NIMP} is a functionally complete logic set. Similar to the synaptic inhabitation and excitation operations, the voltage and light pulses were used as input and current signals were output to realize the logic operations.

To be specific, the implementation of a logic function requires two steps, defined as the initialization, light and electric inputs, respectively. To implement RIMP (NIMP) logic operation, we initialize the device state by applying a light (electric) initialization pulse at first. Then the logic input *p* and *q* with the forms of light and electric pulse, respectively, are simultaneously applied to the device to perform RIMP (NIMP) logic. Meanwhile, the channel current response signal represents the logic output. We define a light (electric) pulse input as logic 1 and no light (electric) signal as logic 0, while high output current response is logic 1 and low output current response is logic 0.

Fig. S11b and S11d show the experimental results of the RIMP and NIMP logic operations. The width and amplitudes of the light and electric pulses were designed so that a SET process will only take place when the light and electric pulses are temporally paired, due to the tunable switching kinetics of the devices. Here, the parameter of the initialization light and electric pulses are (0.5 W cm^{-2}, 200 ms) and (8 V, 100 ms), respectively. The logic 1 inputs of *p* and *q* are 0.5 W cm^{-2}-400 ms light pulse and 10 V-200 ms electric pulse, respectively, while logic 0 means no signal is input. As a result, the logic output “0” (“1”) will only obtain upon the combined logic inputs of “0, 1” (“1, 0”), which has effectively implemented the *p* RIMP *q* (*p* NIMP *q*) function in two logic steps. Thus, our synaptic devices have simultaneously achieved the implementation of synaptic plasticity and

functionally Boolean logic operation. Such multi-functionality has not been reported in previous studies, making our synaptic devices particularly intriguing for optoelectronic computing applications.

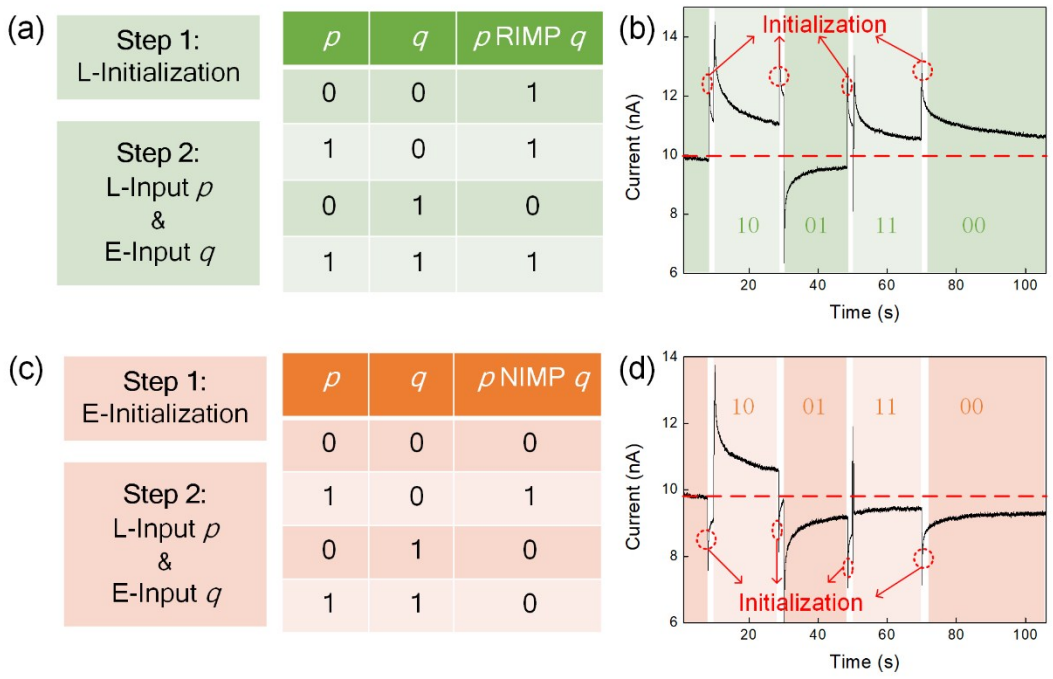

Fig. S11 Realization of proof-of-concept functionally complete Boolean logic. (a) The truth table of RIMP logic. The light input (0.5 W cm^{-2}, 200 ms) was applied as the initialization signal while the light pulse (0.5 W cm^{-2}, 400 ms) and electric pulse (10 V, 200 ms) are applied as the logic input *p* and *q*. (b) Experimental demonstrations of the RIMP operation. (c) The truth table of NIMP logic. The electric input (8 V, 100 ms) was applied as the initialization signal and the logic inputs *p* and *q* are the same as logic RIMP. (d) Experimental demonstration of the NIMP operation.

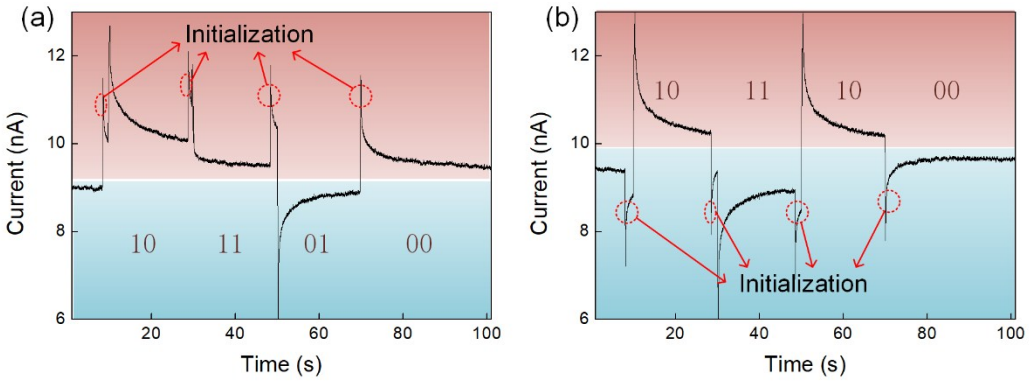

Fig. S12 Realization of the RIMP and NIMP logic operation with the inputs in different order. (a) Logic operation *p* RIMP *q* when the input signal pair (*p*, *q*) followed the sequence 10-11-01-00. The outputs match the RIMP logic correctly. (b) The demonstration of *p* NIMP *q* when the input signal pairs (*p*, *q*) followed the sequence 10-11-10-00.

References

1 Y.-C. Chen, T.-C. Chang, H.-W. Li, T.-Y. Hsieh, T.-C. Chen, C.-P. Wu, C.-H. Chou, W.-Ch. Chung, J.-F. Chang and Y.-H. Tai, *Appl. Phys. Lett.*, 2012, **101**, 223502.

2 T.-Y. Hsieh, T.-C. Chang, T.-C. Chen, M.-Y. Tsai, W.-H. Lu, S.-C. Chen, F.-Y. Jian and C.-S. Lin, *Thin Solid Films*, 2011, **520**, 1427-1431.

3 W.-C. Su, T.-C. Chang, P.-Y. Liao, Y.-J. Chen, B.-W. Chen, T.-Y. Hsieh, C.-I Yang, Y.-Y. Huang, H.-M. Chang, S.-C. Chiang, K.-C. Chang and T.-M. Tsai, *Appl. Phys. Lett.*, 2017, **110**, 103502.

4 F. R. Libsch and J. Kanicki, *Appl. Phys. Lett.*, 1993, **62**, 1286.

5 J. S. Park, W. J. Maeng, H. S. Kim and J. S. Park, *Thin Solid Films*, 2012, **520**, 1679-1693.

6 T.-C. Chen, T.-C. Chang, C.-T. Tsai, T.-Y. Hsieh, S.-C. Chen, C.-S. Lin, M.-C. Hung, C.-H. Tu, J.-J. Chang and P.-L. Chen, *Appl. Phys. Lett.*, 2010, **97**, 112104.

7 K. K. Ryu, I. Nausieda, D. Da He, A. I. Akinwande, V. Bulovic and C. G. Sodini, *IEEE Trans. Electron Devices*, 2010, **57**, 1003-1008.

8 S. Agarwal, S. J. Plimpton, D. R. Hughart, A. H. Hsia, I. Richter, J. A Cox, C. D. James and M. J. Marinella, *IEEE Int. Jt. Conf. Neural Networks*, 2016, 929-938.

9 J. Sun, S. Oh, Y. Choi, S. Seo, M. J. Oh, M. Lee, W. B. Lee, P. J. Yoo, J. H. Cho and J.-H. Park, *Adv. Funct. Mater.*, 2018, **28**, 1804397.