An in-memory computing architecture based on a duplex 2D material structure for *in-situ* machine learning

Hongkai Ning[1]†, Zhihao Yu[1,2]†*, Qingtian Zhang[3]†, Hengdi Wen[1]†, Bin Gao[3]†*, Yun Mao[1], Yuankun Li[3], Ying Zhou[3], Yue Zhou[4], Jiewei Chen[4], Lei Liu[1], Wenfeng Wang[1], Taotao Li[1], Yating Li[1], Wanqing Meng[1], Weisheng Li[1], Yun Li[1], Hao Qiu[1], Yi Shi[1], Yang Chai[4], Huaqiang Wu[3]* & Xinran Wang[1,5,6]*

[1]*National Laboratory of Solid State Microstructures, School of Electronic Science and Engineering and Collaborative Innovation Center of Advanced Microstructures, Nanjing University; Nanjing 210093, China*

[2]*School of Integrated Circuit Science and Engineering, Nanjing University of Posts and Telecommunications; Nanjing 210023, China*

[3]*School of Integrated Circuits, Tsinghua University; Beijing 100084, China*

[4]*Department of Applied Physics, The Hong Kong Polytechnic University; Hung Hom, Kowloon, Hong Kong, China*

[5]*School of Integrated Circuits, Nanjing University; Suzhou 215163, China.*

[6]*Suzhou Laboratory, Suzhou, China.*

*Corresponding author. Email: zhihao@nju.edu.cn, gaob1@tsinghua.edu.cn, wuhq@tsinghua.edu.cn, xrwang@nju.edu.cn

† These authors contributed equally to this work

Abstract

The growing computational demand in artificial intelligence (AI) calls for hardware solutions that are capable of in-situ machine learning, where both training and

inference are performed by edge computation. This not only requires extremely energy-efficient architecture (such as in-memory computing, IMC) but also memory hardware with tunable properties to simultaneously meet the demand for training and inference. Here, we report a duplex device structure based on ferroelectric field-effect transistor (FeFET) and atomically thin MoS_2 channel and realize a universal IMC architecture for in-situ learning. By exploiting the tunability of ferroelectric energy landscape, the duplex building block demonstrates overall excellent performance in endurance ($>10^{13}$), retention (>10 years), speed (4.8 ns) and energy consumption (22.7 fJ/(bit·μm^2)). We implemented a hardware neural network using arrays of two-transistor-one-duplex-FeFET (2T1D) cells and achieved 99.86% accuracy in non-linear localization task with in-situ trained weights. Simulations show that the proposed device architecture could achieve the same level of performance as graphics processing unit under notably improved energy efficiency. Our device core can be combined with silicon circuitry through three-dimensional heterogeneous integration to give a hardware solution toward general edge intelligence (EI).

Introduction

Modern AI relies on central cloud to process data generated at edge devices. Such cloud-edge separated model is not energy efficient due to the von Neumann architecture underpinning digital computing systems as well as the data communications [1-3]. There is strong motivation to develop *in-situ* machine learning hardware with training-and-inference-in-one (TIIO) architecture (Fig. 1a), which is the ultimate goal of EI. TIIO offers the benefit of data security, real-time processing and bandwidth, but it requires extremely high energy and area efficiency due to limited resources at edge. For example, typical edge training scenarios involve over 10^{12} MAC operations per second under milliwatt power, which far exceed the capability of existing hardware technologies [4]. Recently, IMC based on non-volatile memories (NVMs) emerges as a promising solution for EI [4,5,7-14]. However, using a single NVM technology to perform simultaneous training and inference has been challenging [5-7]. This is because training and inference take different aspects of memory properties [5]. In particular, training involves abundant data so it requires good endurance, speed and energy efficiency. On the other hand, inference relies on pre-stored cell weights so retention is critical. In both scenarios, analog capability is desirable to improve the accuracy and energy efficiency of neural networks [5]. Unfortunately, most NVMs lack large tunability in memory properties, preventing a universal IMC architecture that simultaneously satisfies the requirements for training and inference.

Ferroelectrics was proposed as NVM in 1950s and recently became technologically promising after the discovery of ferroelectricity in binary fluorite oxides (HfO_2 and ZrO_2) down to the thickness limit [15-19]. As the basic device building block for IMC, FeFET has been demonstrated on various channel materials, delivering some of the most promising characteristics for edge computing [20-28]. Among the channel materials,

2D semiconductors (such as transition-metal dichalcogenides) are especially appealing because: 1) they have atomic thickness and therefore low power consumption through leakage at scaled device dimension [29-31]; 2) the reduced screening allows the reduction of gate voltage and expands design margin for analog computing [32,33]; 3) they are back-end-of-line (BEOL) compatible with complementary metal-oxide-semiconductor (CMOS) and can be integrated with peripheral circuitry [34], although some challenges of material, device, and integration need to be addressed [35,36]; 4) they offer a variety of sensory properties to facilitate the fusion of sensor with computing [23,37,38].

Here, we combined FeFET with monolayer MoS_2 channel and devised a duplex device structure for *in-situ* machine learning. The duplex structure comprised of a split-gate FeFET with different ferroelectric (FE)/dielectric (DE) capacitance ratio (C_{FE}/C_{DE}) optimized for training and inference, respectively. The duplex structure exhibited excellent endurance (>10^{13}), retention (>10 years), speed (4.8 ns) and energy consumption (22.7 fJ/(bit·μm^2)) simultaneously to meet the requirement for edge training and inference. Multi-layer neural network was implemented with array of 2T1D cells and achieved 99.86% accuracy in non-linear localization using *in-situ* trained weights and all-analog computing. Our results suggest that combining 2D materials with ferroelectrics is a promising hardware solution for EI.

The duplex FeFET device structure

We exploited the tunability of FeFET by engineering the FE energy landscape in the metal–ferroelectric–metal–insulator–semiconductor (MFMIS) device structure. Fig. 1b shows the schematic illustration of the duplex device structure consisting of two split gates with different C_{FE} sharing the same MoS_2 channel. The metal layer between FE and DE acts as floating gate in memory operation. The potential drop across the FE is expressed as:

$$V_{FE} = V_g \times \frac{C_{DE}}{C_{DE}+C_{FE}}, \quad (1)$$

where V_g, C_{FE} and C_{DE} represent the gate voltage and the capacitance of FE and DE layer, respectively. The Gibbs free energy of the FE-DE system is expressed as $G_{FE}t_{FE} + G_{DE}t_{DE}$ where $G_{FE}(G_{DE})$ and $t_{FE}(t_{DE})$ are the free energy and thickness of the FE (DE) layer, respectively. By changing the C_{FE}/C_{DE}, the FeFET can evolve from "FE-like" to "DE-like" as a result of the evolving FE energy landscape, leading to continuously tunable memory characteristics (Extended Data Fig. 1). Specifically, when operating on the gate with small (large) C_{FE}, the duplex FeFET is "FE-like" ("DE-like"), which is more suitable for inference (training). In the extreme case of infinite C_{FE}/C_{DE} (without FE), the device is "pure DE" and can serve as selector transistor in a cross-bar array.

Fig. 1c displays the optical micrograph of a 2T1D duplex cell, where the two split gates of the duplex FeFET are connected to training- (T-) and inference- (I-) selectors through vertical vias. Fig. 1d illustrates the programing sequence during the *in-situ* machine learning process. The T- and I- word line, which is the gate voltage of the corresponding selector, is used to select the T-type and I-type synapse during training and inference, respectively. During *in-situ* training, multiple weight-tuning pulses are applied on T-type synapses through the bit line. After the network has been trained, the weights are transferred to I-type synapses through the same bit line, which are stored there and used for inference. The V_{in}, which is the drain voltage of the FeFET, acted as both weight read for backpropagation during training and data voltage input for feed forward during inference.

Device performance of duplex FeFET

All the devices in this work used chemical-vapor deposited monolayer MoS_2 as

channel[39] and local backgate structure consisting of two dielectrics layers (16 nm $Hf_xZr_{1-x}O_2$ FE and 12 nm HfO_2 DE, corresponding permittivity of 18 and 19 respectively) and three metal layers (backgate, floating gate, and source/drain) (Fig. 2a, see Methods for details of fabrication). All detailed geometric device parameters can be found in Supplementary Table 1. The temperature of the entire MoS_2 transfer and device fabrication process was kept below 450 ℃. We first studied the memory properties of the FeFET as a function of C_{FE}/C_{DE}. To this end, we fabricated a test structure with FE/DE area ratio A_{FE}/A_{DE} ranging from 0.007 to 2.667. Fig. 2b plots the double-sweep I_{ds}-V_g characteristics of the FeFET as a function of A_{FE}/A_{DE}. As expected, the memory window progressively narrowed with A_{FE}/A_{DE} due to the increasing DE contribution in the gate stack (Extended Data Fig. 1). In the pure DE case (by shorting the floating gate and backgate), the device returned to transistor behavior with negligible hysteresis (Fig. 2b, black line).

The retention and endurance characteristics of the FeFET were summarized in Figs. 2c, 2d and Extended Data Figs. 2, 3. In contrast to the binary memory in logic circuits, multi-bit data retention is desirable for inference using IMC. We performed accelerated retention test of 16 states in an I-type FeFET (A_{FE}/A_{DE} = 0.053) under 85 ºC before (Fig. 2c) and after endurance cycling (Extended Data Fig. 3h). Both fresh device and device undergone 10^5 endurance cycles, the conductance of the states was well separated and did not show obvious degradation up to 10^3 s. More remarkably, even under 125 ºC accelerated test [40], we could still extrapolate 10-year retention in the I-type FeFET with on/off ratio of 10^6 (Extended Data Fig. 3f). To evaluate endurance, we

continuously applied programing and erasing voltage pulses with a period of 20 ns and measured the transfer curve to extract I_{on} and I_{off} at intervals of several cycles [41] (Fig. 2d inset, see Methods for more details). Fig. 2d and Supplementary Fig. 1 show the endurance of a T-type FeFET ($A_{FE}/A_{DE} = 0.67$). The devices survived 10^{13} cycles without breaking, and the on/off ratio of 10^5 which was adequate for memory operations. We measured the endurance for a range of A_{FE}/A_{DE} and observed trade-off behavior with retention, which was consistent with the transition from "FE-like" to "DE-like" behavior (Extended Data Figs. 1 and 2a). Nevertheless, the endurance exceeded the requirement for edge training (10^9) and even cloud training (10^{12}) in a wide range of A_{FE}/A_{DE} [7,42], providing a large design space for different applications. To experimentally evaluate the scaling potential of device metrics, we fabricated and measured scaled devices with different channel length. We found that even for channel length down to 85 nm, the strong A_{FE}/A_{DE} dependence, the high retention of DE-like devices and the high endurance of FE-like devices maintained a high consistency with long-channel devices (Supplementary Note 5 and Extended Data Fig. 2).

We further performed benchmark with existing memory technologies, including Flash, RRAM, PCRAM, MRAM, FTJ and FeRAM (Fig. 2f, Supplementary Tables 2 and 3). As a building block for IMC, our duplex FeFET structure simultaneously demonstrated good endurance and retention characteristics. It is worth noting that the degenerated endurance of Hf-based FeFET originates from numerous factors, such as high coercive field for saturation polarization, imprint induced by interfacial traps or defects, uncompensated charge by MFIS structure, etc. Compared with MFIS structure,

the MFMIS releases interfacial voltage stress and reduces the trap and defect generation[20] while embracing symmetrical electrodes and compensated charge. Moreover, benefit from the strong gate dependence of atomic MoS_2, the reduced V_{FE} with the unsaturated polarization can still achieve the multi-bit storage required for training (more flattened *E-P* relationship, see details in Extended Data Fig. 1), thereby effectively improving endurance. As a result, our devices improved the endurance over existing Si- and MoS_2-based FeFETs by 10^2 and 10^8, respectively.

Memory speed and energy consumption was also critical for training with massive data. We characterized the switching speed and read speed by ultrafast pulse measurements and read-after-write measurements (Extended Data Fig. 4). As shown in Fig. 2e, the FeFET could be reliably programed and erased by 4.8 ns electrical pulses (limited by our experimental setup) with good retention and on/off ratio. The FE polarization can be effectively read with minimal delay of 20 ns after programmed, and there is almost no visible shift in both of high and low threshold voltages, which demonstrates very leading read speed (Supplementary Table 4). The switching speed was one of the fastest in FeFET and already met the International Roadmap for Devices and Systems (IRDS) target for NVM [43]. We also calculated the switching energy of 3.4 pJ (or 22.7 fJ/μm^2) from the transient response (Extended Data Fig. 5), which was also among the lowest in NVM (Supplementary Table 2, 5). More importantly, Hf-based ferroelectric has been successfully integrated with advanced processes such as Fin-FET[44] and FDSOI[45], and the memory window has also been reduced to 1.5 V or even lower, which demonstrates the great advantages of ferroelectrics in future advanced

manufacturing integrated circuit applications.

We further assessed the analog storage capability. Extended Data Fig. 6a-c shows the 7-bit (128-state) output characteristics and the corresponding potentiation/depression process of a T-type FeFET (A_{FE}/A_{DE} = 0.43, see Methods for details of measurement). The good linearity of output curves allows all-analog computing (as demonstrated later in the neural network), which is more energy efficient than binary encoding (Supplementary Note 2). The reliable multi-level performance is attributed to the dangling bond-free interface of MoS_2 which could potentially overcome the trap-induced performance degradation in Si-based FeFET[33]. Overall, our duplex FeFET demonstrated excellent memory performance to meet the *in-situ* learning requirements on device level.

Hardware implementation of *in-situ* learning

To demonstrate the potential of the duplex FeFET architecture in *in-situ* learning, we built an artificial neural network (ANN) (Fig. 3a) containing three neuron layers (input, hidden and output) and solved the localization problem in 2D space (Fig. 3b), which is higher-order classification problem that cannot be implemented by single-layer or binary network (Supplementary Note 3). The neural network was physically implemented by an 8×3 array of 2T1D TIIO cells (Fig. 1c). Two 7-bit cells were combined together to realize positive and negative weights to imitate excitation and inhibition in biology. Therefore, the size of L1 synapse (connecting input and hidden layer) and L2 synapse (connecting hidden and output layer) are 2×4 and 4×1, respectively, with 8-bit precision. Within each cell, the T-type and I-type synapse shared

the MoS_2 channel with A_{FE}/A_{DE} of 0.43 and 0.053, respectively. In this pseudo-cross-bar array, training and inference functions were performed as the voltage sequence operation described in Fig. 1d. The datasets were imported as a voltage sequence to V_{in} without any encoding (Extended Data Fig. 6d). The weight was stored by FE polarization and translated as the channel conductance of FeFET and the output I_{ds} was summed over each column according to Kirchhoff's Law. Owing to the linear I_{ds}-V_{ds} curve and long data retention, high-fidelity analog output waveforms were achieved. The measurement setup, software and interfaces were customized to facilitate the hardware test flow (see Methods, Extended Data Fig. 7).

The *in-situ* learning process was divided into three steps, namely on-chip training, weight transfer, and on-chip inference. Fig. 3c shows a typical training process using T-type synapse, where the accuracy and loss gradually converge with distinguished boundaries of the dataset as the epoch progresses. The 2D heatmaps represent the pristine input data and the classification results after the 6^{th}, 12^{th}, and 17^{th} training epoch. After the 17^{th} epoch, the accuracy of both training and test reached 100%, while the cost dropped to 0.067 and 0.083, respectively. The evolution of the localization boundary during the training process was displayed in Supplementary Video 1. The histogram distribution of the synapse weights before and after training are shown in Fig. 3d, suggesting that the weights were changed effectively by the backpropagation algorithm. The robustness and reliability of the classification results were further verified by computer simulations using the same architecture and learning scheme (Extended Data Fig. 8).

After training, the weights were transferred to I-type synapses in the TIIO cell for subsequent inference (see Methods). Subsequently, we performed classification of additional 10,000 data points as shown in Fig. 3e. Thanks to the excellent retention of I-type synapse, the output maintains high accuracy of 99.86% (14 mis-classified points out of 10,000). The histogram shows that most data points are distributed around 0.9 ("inside") or 0 ("outside") away from the boundary (0.5), indicating high fidelity of the inference results.

Simulation of large-scale artificial neural network

Autonomous robotic vision is an important application for *in-situ* learning. Biological systems typically adopt binocular vision, which rely on disparity of optical path difference entering the left and right eyes to render the real-time 3D space. Monocular depth estimation, on the other hand, is attractive for computer vision due to the reduced hardware volume and computation resources[46] (Fig. 4a). However, monocular depth estimation is like seeing 3D space when one eye, which requires repeated data training to adapt the foreshortening effects and therefore extremely high energy efficiency.

A widely adopted approach for monocular depth estimation is the encoder-decoder architecture (Fig. 4b). The encoder part uses massive pre-trained weights through transfer learning[47] but minimal weight update, which requires long data retention (corresponding to I-type synapse). On the contrary, the decoder part focused on feature extraction from training with abundant data (corresponding to T-type synapse). Here, a 15-block U-Net[48] with 178 layers was simulated using the duplex architecture, where

I-type (T-ype) synapses were used in the 9-block encoder (6-block decoder) with all the device parameters derived from experiments (see Methods and Extended Data Fig. 9). Two variation models were constructed for training and inference to ensure the simulation reliability (See Methods). The simulated chip consisted of 128 × 128 2T1D cells with peripheral analog-to-digital converter (ADC), sample-and-hold circuits multiplexer, controller, and driver (Fig. 4c inset). Fig. 4d and Extended Data Fig. 10e show several street scenes in autonomous driving. Our duplex TIIO chip successfully identified all the features and captured their relative depth with comparable convergence rate as GPU (Fig. 4c). The recognition accuracy (sigma 3 level of threshold) and RMSE (Root Mean Square Error) reached 96.85% and 6.31%, respectively (see Extended Data Fig. 10a,10c). Compared to GPU, the convolution circuit of duplex TIIO exhibits better energy efficiency while maintaining the equal computing accuracy. We designed rigorous scaling rules based on ITRS reports and cell layout with appraised parasitic parameter to perform energy efficiency projection for our TIIO cell at advanced 22 nm node (Supplementary Note 6). For training (inference) process, the pre- and post- simulation of projected cell energy efficiency is 2110 (111.86) TOPS/W and 1151 (111.86) TOPS/W. We noticed that the reduced energy efficiency in the post-simulation is induced by the larger operating voltage of the bit line, which also leads to further drop in energy efficiency as the array scale increases. Therefore, reducing the thickness of HZO and realizing the integration of more advanced technology node will be crucial to improving chip-level energy efficiency. Thanks to the BEOL advantages of 2D materials and ferroelectric HZO, the neuromorphic

computing cores can be the monolithically integrated with other necessary functional blocks of pooling, activation, routing and buffering in the future, and further improve overall energy efficiency.

Conclusions

In this work, we have shown large tunability of memory metrics by device architecture design, which is lacking for most non-volatile memory technologies. We demonstrated an IMC architecture that can complete in-situ machine learning, using a unitary device technology. By integrating split FE capacitors with complementary characteristics in the same memory cell, the proposed duplex architecture solves the problem of conflicting memory requirements for training and inference, which has long plagued EI applications. It not only simplifies the hardware fabrication process, but also merges the training and inference process in one memory building block. Such compact design can improve parallel computation and thus deliver higher energy efficiency. Based on 22 nm technology node, our architecture shows a post-simulation projected energy efficiency for training of 1151 TOPS/W, using the single TIIO cell. It is, however, worth noting that the projection here is somewhat overestimated because the contribution of the necessary peripheral circuitry is not included. Compared with previous work that focused on training and inference, we use the non-volatile multi-bit characteristics for both learning and inference on a single device, and demonstrate 2D localization task on a small-scale hardware circuit, which maintains high area efficiency and energy efficiency for IMC applications. Our design also embraces transfer learning which is widely applied in image processing, natural language processing and emotion recognition, thus will likely become a key component in lifelong learning applications.

Acknowledgements.

This work is supported by the National Key R&D Program of China (grant no. 2022YFB4400100 (X. W.), 2021YFA0715600 (H. Q.), 2021YFA1202903 (W. L.)), the National Natural Science Foundation of China (grant no. T2221003 (X. W.), 61927808 (X. W.), 61734003 (X. W.), 61851401 (X. W.), 91964202 (Z. Y.), 62204124 (Z. Y.), 51861145202 (X. W.)), the Leading-edge Technology Program of Jiangsu Natural Science Foundation (grant no. BK20202005 (X. W.)), the Strategic Priority Research Program of Chinese Academy of Sciences (grant no. XDB30000000 (X. W.)), the Research Grant Council of Hong Kong (no. 15205619 (Y. C.)), Key Laboratory of Advanced Photonic and Electronic Materials, Collaborative Innovation Center of Solid-State Lighting and Energy-Saving Electronics, and the Fundamental Research Funds for the Central Universities, China. In addition, we thank the NJU Micro-fabrication and Integration Center for support during device fabrication and measurement, and thank the Beijing Advanced Innovation Center for Integrated Circuits for support on device modeling and simulations.

Author Contributions.

Z. Y. and X. W. conceived and supervised the project. H. N. performed the fabrication of device and TIIO array with assistance from Z. Y., H. Wen., W. M., W. L., Yating L. and Y. L. H. N. did electrical measurements with assistance with Z. Y. and H. Wen. Y. Mao and H. Qiu performed projections and simulations of 22nm-node FeFET. L. L., W. W. and T. L. performed MoS_2 growth. Q. Z., Yuankun. L., Ying Z., Yue Z., J. C. contributed to simulations of monocular depth estimation, with the guidance from B. G., Y. C. and H. Wu. H. N., Z. Y., X. W co-wrote the manuscript with input from other authors. All the authors contributed to discussions.

Competing interests. The authors declare no competing financial interest.

Fig. 1. *In-situ* machine learning with TIIO cell. a, Inference and training process in machine learning. During inference, the weights are saved in a synapse array, where massive multiply-and-accumulate (MAC) are done in parallel. During training, weights in synapses are updated frequently. The proposed TIIO cell can integrate inference (I-) type and training (T-) type synapse in the same memory building block to realize *in-situ* learning. **b**, Schematics of duplex 2D material CIM device. **c**, Optical microscope image and programming sequence of a TIIO cell comprised of 2T1D. Besides the duplex FeFET core, two selector transistors (T- and I-) are involved to form a pseudo-crossbar structure. Scale bar, 20 μm.

Fig. 2. The duplex FeFET device performance. **a**, Schematic drawing of the test structure with different A_{FE}/A_{DE} sharing the same MoS_2 channel. **b**, Transfer characteristics of FeFET with different A_{FE}/A_{DE}, revealing large tunability of memory window. **c**, 16-level (chosen from 128 states) data retention of an I-type FeFET (A_{FE}/A_{DE} = 0.053) under 85ºC accelerated test. **d**, The endurance of a T-type FeFET (A_{FE}/A_{DE} = 0.67). Inset shows the pulse sequence during test. **e**, Switching of FeFET under 4.8 ns programming and erasing pulses. **f**, Benchmark of endurance and retention with other memory technologies. The three horizontal lines mark the endurance requirement for cloud training ($>10^{12}$), edge training ($>10^{9}$), and storage ($>10^{5}$). STP: short-term plasticity; LTP: long-term plasticity. The references for the data in **f** are summarized in Supplementary Tables 2, 3.

Fig. 3. *In-situ* machine learning with TIIO ANN. a, Left, microscopic image of chip layout with TIIO ANNs and test structures. Scale bar, 1 mm. Right, one TIIO ANN with pseudo-crossbar structure containing two synapse layers (L1 and L2), 8 bit lines, 8 hidden nodes, 6 word lines, 2 input lines and 1 output line. Scale bar, 100 μm. **b**, Scene illustration of the 2D localization task. This non-linear classification requires neural network with at least 2 synapse layers. The target of this ANN was classifying location data as "inside (1)" or "outside (0)" with a high accuracy. **c-e**, Training (**c, d**) and inference (**e**) with the TIIO ANN. **c**, Cost and accuracy as a function of training epoch (blue stands for training data and yellow stands for test data). The training finished at the 17^{th} epoch with 100% accuracy. Classification heatmaps of the initial and 6^{th}, 12^{th} and 17^{th} epoch are plotted. Data points with white (210 points) and black border (90 points) stand for training and test data, respectively. **d**, The distribution of weights and bias parameters before and after training. **e**, The inference result of 10,000 data points using in-situ trained weights. 99.86% accuracy was achieved. The dash line at 0.5 draws out the threshold of classification, where the outputs ≥ 0.5 were classified as "inside (1)", and the outputs <0.5 were classified as "outside (0)".

Fig. 4. Simulation of large-scale TIIO ANN. a, The scene illustration of monocular depth estimation in autonomous driving. **b**, The employed neural network with encoder-decoder architecture. **c**, Test loss as a function of epoch simulated on GPU (gray) with 8-bit precision and 128×128 TIIO ANN (yellow). The yellow shaded region stands for standard error from 5 independent runs. And the center line with yellow symbols stands for the mean values of these 5 runs. Inset, schematic chip architecture used in this simulation. **d**, A representative scene of depth estimation containing 4 cars and 5 poles. The TIIO correctly distinguishes all the features with sharp edges.

References:

1. Hutson, M. Has artificial intelligence become alchemy? *Science*. **360**, 478–478 (2018).

2. Christensen, D. V. et al. 2022 roadmap on neuromorphic computing and engineering. *Neuromorph. Comput. Eng.* **2**, 022501 (2022).

3. Mehonic, A. & Kenyon, A. J. Brain-inspired computing needs a master plan. *Nature*. **604**, 255–260 (2022).

4. Salahuddin, S., Ni, K. & Datta, S. The era of hyper-scaling in electronics. *Nat. Electron.* **1**, 442–450 (2018).

5. Kendall, J. D. & Kumar, S. The building blocks of a brain-inspired computer. *Appl. Phys. Rev.* **7**, 011305 (2020).

6. Ambrogio, S. et al. Equivalent-accuracy accelerated neural-network training using analogue memory. *Nature*. **558**, 60–67 (2018).

7. Yu, S. Neuro-inspired computing with emerging nonvolatile memory. *Proc. IEEE*. **106**, 260–285 (2018).

8. Zhou, Z. et al. Edge intelligence: paving the last mile of artificial intelligence with edge computing. *Proc. IEEE*. **107**, 1738–1762 (2019).

9. Keshavarzi, A., Ni, K., Hoek, W. V. D., Datta, S. & Raychowdhury, A. FerroElectronics for edge intelligence. *IEEE Micro*. **40**, 33–48 (2020).

10. Yao, P. et al. Fully hardware-implemented memristor convolutional neural network. *Nature*. **577**, 641–646 (2020).

11. Demasius, K.-U., Kirschen, A. & Parkin, S. Energy-efficient memcapacitor devices for neuromorphic computing. *Nat. Electron.* **4**, 748–756 (2021).

12. Chen, W. et al. CMOS-integrated memristive non-volatile computing-in-memory for AI edge processors. *Nat. Electron.* **2**, 420–428 (2019).

13. Cheng, C. et al. In-memory computing with emerging nonvolatile memory devices. *Sci. China Inf. Sci.* **64**, 221402 (2021).

14. Li, C. et al. Analogue signal and image processing with large memristor crossbars. *Nat. Electron.* **1**, 52–59 (2018).

15. Müller, J. et al. Ferroelectricity in simple binary ZrO_2 and HfO_2. *Nano Lett.* **12**, 8, 4318–4323 (2012).

16. Böscke, T. S., Müller, J., Bräuhaus, D., Schröder, U. & Böttger, U. Ferroelectricity in hafnium oxide thin films. *Appl. Phys. Lett.* **99**, 102903 (2011).

17. Cheema, S. S. et al. Enhanced ferroelectricity in ultrathin films grown directly on silicon. *Nature*. **580**, 478–482 (2020).

18. Cheema, S. S. et al. Emergent ferroelectricity in subnanometer binary oxide films on silicon. *Science*. **376**, 648-652 (2022).

19. Gao, Z. et al. Identification of ferroelectricity in a capacitor with ultra-thin (1.5-nm) $Hf_{0.5}Zr_{0.5}O_2$ film. *IEEE Electron Device Lett.* **42**, 1303-1306 (2021).

20. Khan, A. I., Keshavarzi, A. & Datta, S. The future of ferroelectric field-effect transistor technology. *Nat. Electron.* **3**, 588–597 (2020).

21. Schroeder, U., Park, M. H., Mikolajick, T. & Hwang, C. S. The fundamentals and applications of ferroelectric HfO_2. *Nat. Rev. Mater.*, 1–17 (2022).

22. Jerry, M. et al. Ferroelectric FET analog synapse for acceleration of deep neural network training. In *2017 IEEE International Electron Devices Meeting (IEDM)* 6.2.1-6.2.4. (IEEE, 2017).

23. Ni, K. et al. SoC logic compatible multi-bit FeMFET weight cell for neuromorphic applications. In *2018 IEEE International Electron Devices Meeting (IEDM)* 13.2.1-13.2.4. (IEEE, 2018).

24. Sun, X., Wang, P., Ni, K., Datta, S. & Yu, S. Exploiting hybrid precision for training and inference: a 2T-1FeFET based analog synaptic weight cell. In *2018 IEEE International Electron Devices Meeting (IEDM)* 3.1.1-3.1.4. (IEEE, 2018).

25. Tong, L. et al. 2D materials-based homogeneous transistor-memory architecture for neuromorphic hardware. *Science*. **373**, 1353–1358 (2021).

26. Zhang, W. et al. Neuro-inspired computing chips. *Nat. Electron.* **3**, 371–382 (2020).

27. Luo, Q. et al. A highly CMOS compatible hafnia-based ferroelectric diode. *Nat. Commun.* **11**, 1391 (2020).

28. Radisavljevic, B., Radenovic, A., Brivio, J. et al. Single-layer MoS2 transistors. Nature Nanotech **6**, 147–150 (2011).

29. Akinwande, D. et al. Graphene and two-dimensional materials for silicon technology. *Nature*. **573**, 507–518 (2019).

30. Liu, C. et al. Two-dimensional materials for next-generation computing technologies. *Nat. Nanotechnol.* **15**, 545–557 (2020).

31. Marega, M. et al. Logic-in-memory based on an atomically thin semiconductor. *Nature* **587,** 72–77 (2020).

32. Chung, Y.-Y. et al. High-accuracy deep neural networks using a contralateral-gated analog synapse composed of ultrathin MoS_2 nFET and nonvolatile charge-trap memory. *IEEE Electron Device Lett.*, 1–1 (2020).

33. Chen, L., Pam, M. E., Li, S. & Ang, K. -W. Ferroelectric memory based on two-dimensional materials for neuromorphic computing. *Neuromorph. Comput. Eng.* **2**, 022001 (2022).

34. Meng, W. et al. Three-dimensional monolithic micro-LED display driven by atomically thin transistor matrix. *Nat. Nanotechnol.* **16**, 1231–1236 (2021).

35. Schram, T., Sutar, S., Radu, I. & Asselberghs, I. Challenges of wafer-scale integration of 2D semiconductors for high-performance transistor circuits. *Adv. Mater.* **2109796** (2022).

36. Waltl, M. et al. Perspective of 2D integrated electronic circuits: scientific pipe dream or disruptive technology? *Adv. Mater.* **2201082** (2022).

37. Chai, Y. In-sensor computing for machine vision. *Nature*. **579**, 32–33 (2020).

38. Mennel, L. et al. Ultrafast machine vision with 2D material neural network image sensors. *Nature*. **579**, 62–66 (2020).

39. Li, T. et al. Epitaxial growth of wafer-scale molybdenum disulfide semiconductor single crystals on sapphire. *Nat. Nanotechnol*. **16**, 1201–1207 (2021).

40. Müller, J. et al., Ferroelectric hafnium oxide: A CMOS-compatible and highly scalable approach to future ferroelectric memories. In *2013 IEEE International Electron Devices Meeting (IEDM)* 10.8.1-10.8.4 (IEEE, 2013).

41. N. Gong & T. -P. Ma, A Study of Endurance Issues in HfO2-Based Ferroelectric Field Effect Transistors: Charge Trapping and Trap Generation. *IEEE Electron Device Lett.* **39**, 1, 15-18 (2018).

42. Y. Liu et al., 4.7 A 65nm ReRAM-enabled nonvolatile processor with 6× reduction in restore time and 4× higher clock frequency using adaptive data retention and self-write-termination nonvolatile logic. In *2016 IEEE International Solid-State Circuits Conference (ISSCC)* 84-86 (IEEE, 2016).

43. *International Roadmap for Devices and Systems (IRDS™) 2021 Edition* (IEEE, 2021); *https://irds.ieee.org/editions/2021* .

44. Krivokapic, Z. et al. 14nm Ferroelectric FinFET technology with steep subthreshold slope for ultra low power applications. In *2017 IEEE International Electron Devices Meeting (IEDM)* 15.1.1-15.1.4 (IEEE, 2017).

45. Dünkel, S. et al. A FeFET based super-low-power ultra-fast embedded NVM technology for 22nm FDSOI and beyond. In *2017 IEEE International Electron Devices Meeting (IEDM)* 19.7.1-19.7.4 (IEEE, 2017).

46. Zhao, C., Sun, Q., Zhang, C., Tang, Y. & Qian, F. Monocular depth estimation based on deep learning: An overview. *Sci. China Technol. Sci.* **63**, 1612–1627 (2020).

47. Alhashim, I. & Wonka, P. High quality monocular depth estimation via transfer learning. Preprint at https://arxiv.org/abs/1812.11941 (2018).

48. Ronneberger, O., Fischer, P. & Brox, T. U-Net: convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI* 234-241 (Springer, 2015).

Methods:

The fabrication of the duplex FeFET/ TIIO Array.

On the p-type silicon substrate with 275 nm SiO_2, back gate (M1) was defined by electron beam lithography (EBL), 3 nm Ti/9 nm Pt were deposited by electron beam evaporator (EBE). 16 nm $H_{0.5}Z_{0.5}O_2$ (HZO) film was deposited at 200℃ by atomic layer deposition (ALD) using precursors TDMA-Hf and TDMA-Zr, while water as the oxygen source. Next, floating gate (FG) was defined by EBL, and about 14nm Pt were evaporated using EBE. With FG metal covered, crystallizing of ferroelectric HZO was realized by rapid thermal annealing (RTA) at 450℃ in N_2 atmosphere for 30s. 12 nm HfO_2 film was deposited at 150℃ by ALD using precursor TEMA-Hf, while O_2 plasma as the oxygen source.

There are slight differences for fabrication of TIIO array. During the substrate fabrication, the input line and output line were made with M1 metal. The training/inference word line was made with M2 metal. The bit line and hidden nodes were made with M3 metal. Interconnection via was defined by EBL and etched using BCl_3/Ar by GSE C200 Series Plasma Etcher, with Pt as etch stop after M1, M2 and dielectric. There were 4 different via types in a TIIO array: M1-M3, The input line in connection with drain of the FeFET; M1-M3, The bottom metal in connection with source of selector FET; M1, For probing the input line and output line; M2, For probing the training/inference word line. The size and distribution of pads were specially designed for customized probe cards.

Single-crystalline monolayer MoS_2 films were grown on custom-designed C/A-plane sapphire wafers in a home-made CVD furnace. Assisted by 35nm flat Au films, the MoS_2 film was transferred to target substrate by PMMA and PDMS. PDMS/PMMA/Au stack was laminated on fresh new MoS_2/sapphire. Next, MoS_2 was dry-delaminated from the sapphire and transferred onto substrates with pre-patterned gate layout in glovebox. Then, the unnecessary Au/MoS_2 (defined by EBL) was removed by sequential Au etching (using Transense TFA) and MoS_2 etching (using SF_6 and O_2 plasma in reactive ion etcher (RIE)). The last step of EBL defined the Source/Drain pattern, and M3 metal of 10nm Ti/35nm Pd/10 nm Ti were deposited by EBE. Finally, self-aligned etch was performed to open channel via S/D metal mask using Transense TFA. An annealing (200℃) was performed to remove adsorbates and improve contact with base pressure ~10^{-6} Pa in vacuum atmosphere.

Electrical measurement of duplex FeFET

We developed a home-made system for the various *in-situ* measurements of FeFET and TIIO array. The system contained Keithley 4200 semiconductor characterization system (SCS) with 4 SMUs for DC test and 4225- remote pulse and switch module (RPM) for pulse test, a National Instruments (NI)-PXIe 2532B matrix switch (with 8×64 terminal block), PXIe-5433 arbitrary waveform generator (AWG) and Keysight MSOX6004A oscilloscope.

The transfer and output characteristics were measured by SMUs with pre-amplifier, which enable a current resolution of 0.1 fA. As for data retention, the FeFET was programmed to ON state or erased to OFF state, then a DC sampling test ran for thousands of seconds. In addition, we measured retention at temperatures of 85^{o}C and 125^{o}C in a vacuum atmosphere in the Lake Shore CRX-VF probe station. We extrapolated the high-temperature 10-year retention by linear fitting.

In the multi-state test, potentiation and degression were realized by positive and negative pulses, generated by 4225-RPMs. While applying pulses at gate, the drain and source of FET were both grounded. Once the pulse finished, we ground the gate and applied V_{ds} to read the conductance of FeFET. For a shorter pulse width in the speed test, we switched to NI PXIe-5433 AWG, which can generate pulses with amplitudes up to 10 V and pulse widths as small as 4.8 ns. A Keysight MSOX6004A oscilloscope was used here to collect real-time pulse amplitude and width. The mode of AWG was set to user-defined waveform, list output, and immediate trigged. The duration was set carefully to make sure that only one pulse generated for every output. In endurance test, based on the AWG, we change the duration to output a sequence of identical pulses, with different cycles number of 1, 10, 1E3, 1E4, …, 1E13. At the very end of one sequence, we check the transfer curve of FeFET to monitor the performance degradation. Considering the time spent was very large for 1E12, 1E13 cycling, we just measured I_{on} and I_{off} for discrete cycles rather than every cycle.

Hardware *in-situ* machine learning on TIIO array

In the array measurement, we modified the vacuum probe station to meet the special requirements. The NI PXIe 2532B matrix switch, in the 8×64 terminal set-up,

helped connect the SMUs/PMUs test sources with the device under test (DUT). We loaded two customized probe cards (12 pins for A, 15 pins for B) on the original arms in the probe station. These probe cards were electrically connected with flexible flat cables (FFC), adapters, cable hub (48-line feed through), and in the end, the test instruments outside. All the test were performed in the vacuum environment.

For training process, we added one PC here in connected with Keithley 4200 SCS for running program codes, which defined the initial parameters and hyperparameters, flow of ANN training, interfaces for software-hardware interaction, and related data processing. On the level of hardware, two selectors share one drain in one 2T1D cell, thus the operation mode depends on which word line (T- or I-type) accesses the duplex FeFET. A typical process is mode transferring from training to inference, which means resetting the T-type capacitor, switching to I-word line, and programing I-type capacitor to a well-trained weight. More details about the algorithm can be found in the Supplementary Note 4.

Device modeling and hardware evolution

Based on the measuring results of duplex FeFET devices (Supplementary Fig. 6,7), we constructed two variation models for the inference and training, respectively. Without loss of generality, random variables sampled from the Gaussian distribution with zero mean and σ^2 variance are used to simulate the inference and training variation. A linear variation model is used in this work.

$$W_{w/\ noise} = W_{w/o\ noise} + W_{w/o\ noise} \times Noise_{weights}, \quad (2)$$

$$Noise_{weights} \sim N\left(0, \sigma^2_{weights}\right). \quad (3)$$

The standard deviation $\sigma_{\mathrm{weights}}$ is 0.056 μS (3μm L_{ch}) and 0.040μS (scaled device, 85nm L_{ch}).

Similar to the inference variation, a similar linear model is used to simulate the training variation:

$$V_{update\ w/\ noise} = V_{update\ w/o\ noise} + V_{update\ w/o\ noise} \times Noise_{update}, \quad (4)$$

$$Noise_{update} \sim N\left(0, \sigma^2_{update}\right). \quad (5)$$

The standard deviation σ_{update} is 0.043 μS (3μm L_{ch}) and 0.017μS (scaled device, 85nm L_{ch}).

Dataset and neural network structure in monocular depth estimation

We evaluated our devices on a monocular pixel-level depth prediction task based on a subset of the KITTI dataset. The data in KITTI dataset is captured by driving around in rural areas and on highways in the mid-size city of Karlsruhe. The dataset comprises stereo and optical flow image pairs, stereo visual odometry sequences, and object annotations captured scenarios[49]. In this work, we tried to predict the depth of each pixel in the raw RGB images from a monocular camera. We randomly selected 2,802 images for training and 608 images for the test.

We simulated a transfer learning algorithm to demonstrate the superiority of TIIO architecture in both inference and training. The neural network adopts the U-Net structure, which consists of the encoder and decoder [47,48]. The encoder is realized by a 169-layer DenseNet [50] with four dense blocks and four transition blocks. The decoder is realized by a convolutional layer and five upsampling blocks. Each upsampling block contains a bilinear upsampling layer and two convolutional layers with Leaky-ReLU activations. The four dense blocks in the encoder are connected to the first four upsampling blocks, respectively. The whole network configuration is shown in Supplementary Table 9. The encoder is pretrained on ImageNet classification task [50,51]. While the decoder is randomly initialized using a uniform model and trained for this depth prediction task with the encoder together.

Training details in monocular depth estimation

The loss function with L1-norm loss and structural similarity (SSIM) loss[52] is used:

$$Loss = \lambda L_1(y, \hat{y}) + L_{SSIM}(y, \hat{y}), \quad (6)$$

where y indicates predicted image and $\hat{y}$ indicates ground truth. The pixel-wise L1-norm loss is defined as:

$$L_1(y, \hat{y}) = \frac{1}{n}\sum_p^n |y_p - \hat{y}_p|. \quad (7)$$

The SSIM loss is defined as:

$$L_{SSIM}(y, \hat{y}) = \frac{1-SSIM(y,\hat{y})}{2}. \quad (8)$$

λ is set to 0.1 in this work.

To update the weights according to the gradients, a series of identical pulses are applied on the duplex FeFET devices and the without-verify strategy is used in this simulation. When the gradient is less than a quarter of the average change of one pulse, the devices will not be changed.

The other parameter setting of the training are listed in Supplementary Table 10.

Evaluation of predicted depth

We evaluated the accuracy of predicted depth with different tolerant level ($\delta_1, \delta_2, \delta_3$), the absolute relative depth error (*abs* Rel.), the root mean square error of depth (RMS), and the Log Mean Absolute Error (*log* MAE) [53] of our duplex FeFET *in-situ* training algorithm. The predicted depth of a pixel is considered correct with tolerant level δ depending on whether the relative error between the predicted depth and the ground truth is smaller than δ.

$$max\left(\frac{depth_{pred}}{depth_{gt}}, \frac{depth_{gt}}{depth_{pred}}\right) < \delta. \quad (9)$$

The tolerant level used in this work is 1.25, 1.25^2, and 1.25^3. The other evaluation indicators are calculated as follows.

$$abs\ \text{Rel.} = \frac{1}{n}\sum \frac{|y_{pred}-y_{gt}|}{y_{gt}}, \quad (10)$$

$$\text{RMS} = \sqrt{\frac{1}{n}|y_{pred} - y_{gt}|^2}, \quad (11)$$

$$log\ \text{MAE} = \frac{1}{n}\sum |log(y_{pred}) - log(y_{gt})|. \quad (12)$$

The comparisons between GPU and TIIO are shown in Extended Data Fig 10.

Data availability: Source data are provided with this paper.

Code availability: The codes used to build the interfaces (0~3) in the demonstrations in Fig. 3, and used for the simulations in Extended Data Fig. 8 are available from the corresponding author upon reasonable request.

Methods-only references:

49. Geiger, A., Lenz, P. & Urtasun, R. Are we ready for autonomous driving? The KITTI vision benchmark suite. In *2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* 3354-3361 (IEEE, 2012).

50. Huang, G., Liu, Z., Maaten, L. V. D. & Weinberger, K. Q. Densely connected convolutional networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* 2261-2269 (IEEE, 2017).

51. Deng, J. et al. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* 248-255 (IEEE, 2009).

52. Wang, Z., Bovik, A. C., Sheikh, H. R. & Simoncelli, E. P. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing* **13**, 600-612 (2004).

53. Eigen, D., Puhrsch, C. & Fergus, R. Depth Map Prediction from a Single Image using a Multi-Scale Deep Network. In *the 28th Conference on Neural Information Processing Systems (NIPS)* (NIPS, 2014).