INTELLIGENT VISION DEVICES AND SYSTEMS

JIANG BIYI

PhD

The Hong Kong Polytechnic University

2026

The Hong Kong Polytechnic University

Department of Applied Physics

Intelligent Vision Devices and Systems

JIANG Biyi

A thesis submitted in partial fulfilment of the requirements for the degree of Doctor of Philosophy

August 2025

CERTIFICATE OF ORIGINALITY

I hereby declare that this thesis is my own work and that, to the best of my knowledge and belief, it reproduces no material previously published or written, nor material that has been accepted for the award of any other degree or diploma, except where due acknowledgement has been made in the text.

________________________________(Signed)

JIANG Biyi ____________(Name of student)

Abstract

Intelligent vision systems that can sense, understand, and process the visual scenes in real time hold great promise for versatile and complex vision scenarios. Conventional vision system with Von Neumann architecture cannot process these scenarios efficiently. Emerging intelligent device-based systems hold promise, but they still suffer from limited reconfigurability and advanced processing capabilities at device level, as well as the lack of co-designed circuits and algorithms at the system level.

To solve the limited reconfigurability, we employ a multi-paradigm device with four sensory computing modes, including 1) photo-spiking neuron (PN) and 2) photo-synaptic (PS) modes with spiking and non-spiking in-sensor low-/mid- level processing capabilities for neuromorphic imaging (NI) functions, and 3) electrical-synaptic (ES) and 4) electrical-neuron (EN) modes with spiking and non-spiking in-memory high-level processing capabilities for artificial intelligence (AI) functions. With the co-designed mode control circuit and reconfigurable dataflow architecture, we achieve an intelligent vision system with hierarchical reconfigurability from device to system level. It is capable of on-demand allocating resources between NI and AI functionalities towards high-quality smart imaging and high-accuracy recognition tasks, and spiking/non-spiking mode transitions towards frameless dynamic and frame-based static scenarios. Exceptional power efficiencies of up to 52.6 TOPS/W for NI-centric computing and 75.5 TOPS/W for AI-centric computing are achieved, which are up to two orders of magnitude larger than the state-of-the-art intelligent vision system.

To address the limited processing capability in complex multi-target detection scenarios, systems capable of efficiently perceiving multidimensional optical information are highly desirable. Inspired by the multilevel biological visual filters, we introduce a multidimensional and multilevel MoS_2/HZO ferroelectric field-effect transistor (FeFET) device. It exhibits enhanced wavelength selectivity for low-level colour clustering and mid-level multidimensional decoupling due to the polarization-enhanced interfacial barrier height, as well as polarization-enhanced multi-photoresponsivity states for high-level processing. With the co-designed hierarchical visual filter bank algorithm (HVBF), a multidimensional hierarchical intelligent vision system is constructed, capable of performing low-level colour clustering, mid-level segmentation, and high-level recognition within a single MoS_2/HZO device array. It achieves 93.8% accuracy in identifying overlapping digits in noisy scenarios, demonstrating high robustness and precision.

For scenarios requiring three-dimensional (3D) understanding, we demonstrate a bipolar MoS_2/In_2Se_3 device that generates photocurrent with different polarities for near-infrared (NIR) and red, green, and blue (RGB) light under the same biasing condition based on the visible light-induced MoS_2/Au interfacial barrier-lowering mechanism. Combining with near-infrared marker-based depth perception technology, the MoS_2/In_2Se_3 array-based RGB-D camera can efficiently sense and decouple two-dimensional (2D) RGB and 3D depth images simultaneously. An intelligent visual system with 3D scene understanding capability is realized, including the RGB-D camera for sensing and decoupling, and two neural networks (the convolutional neural network (CNN) and the long short-term memory (LSTM) network) for processing the 2D and 3D images, respectively. In the augmented reality (AR) task involving two hands interacting with one object, the system recognizes static objects and dynamic gestures, and places virtual annotations near their real-world counterparts with 91.8% accuracy.

Publications

1. Biyi Jiang[#], Jiayi Xu[#], Liang Ran[#], Yida Li[*], Longyang Lin[*] and Feichi Zhou[*], " An On-Demand Neuromorphic Vision System Enabled by A Multi-Paradigm Neuromorphic Device and Hierarchical Reconfigurability designed from Device to System Level", Advanced Science, under minor revision.
2. Jiayi Xu[#], Biyi Jiang[#], Weizhen Wang[#], Zhifeng Guo, Junsen Gao, Xinyan Hu, Jingkai Qin, Liang Ran, Longyang Lin, Songhua Cai[*], Yida Li[*], and Feichi Zhou[*], "High-Order Dynamics in An Ultra-Adaptive Neuromorphic Vision Device ", Nature nanotechnology, 2025, 20, 1419–1430. (co-first author)
3. Biyi Jiang et al., " Bio-Inspired Multilevel Visual Filter with A MoS_2-HZO Ferroelectric Field-Effect Transistor for Multidimensional and Multi-Object Processing", under review.
4. Biyi Jiang et al., " RGB-NIR Decoupled Bipolar MoS_2/In_2Se_3 Device based RGB-D Camera for Augmented Reality", to be submitted.
5. Jiang Bi-Yi, Zhou Fei-Chi[*], Chai Yang[*]. Application of neuromorphic resistive random access memory in image processing. Acta Phys. Sin., 2022, 71(14): 148504
6. Huang, PY.[#], Jiang, BY.[#], Chen, HJ.[#] et al., "Neuro-inspired optical sensor array for high-accuracy static image recognition and dynamic trace extraction", Nature Communications, 2023, 14, 6736. (co-first author)
7. Zhang, HJ.[#], Jiang, BY.[#] et. al., "A Self-Rectifying Synaptic Memristor Array with Ultrahigh Weight Potentiation Linearity for a Self-Organizing-Map Neural Network", Nano Letters, 2023, 23, 3107-311. (co-first author)

Acknowledgments

It is my great honour to study at PolyU and pursue my Ph.D. in the Department of Applied Physics. I have learnt a lot here. From a beginner who knows very little about the field, to being able to do my own research.

For these, I would like to express my sincere gratitude to my PolyU Chief Supervisor, Prof. CHAI Yang. His guidance has been invaluable to me. I am also very grateful to my other supervisor, Dr. ZHOU Feichi. She has always been very patient in helping me and answering my questions. Especially in my first two years, since I knew very little about the device physics, I often asked some questions that now seem basic, even a little stupid. Despite this, Dr. ZHOU Feichi patiently answered my questions every time, and finally helped me get through the beginner period and get me on the right track of research. Finally, I would like to thank my mother and father for their companionship and support, as well as for myself.

Table of Contents

Chapter 1 Introduction: Demands for intelligent vision systems and devices

Demands for intelligent vision systems, which are defined as systems that can not only acquire data from vision scenarios but also interpret and process it efficiently, are increasing. The reason comes not only from the complexity of the real-world vision scenarios themselves, but also from that many modern technologies require smarter processing. Many factors could influence the real-world visual scenes, such as lighting conditions, depth variations, overlapping objects, and background noise. For example, low-light conditions significantly degrade imaging quality[1]. Noisy and cluttered background also makes it difficult to recognize the target object accurately[2]. On the other hand, with the continuous development of technologies, the demands from modern application scenarios such as autonomous driving, intelligent robotics, augmented reality (AR), and smart security are increasing. For instance, in autonomous driving, systems must accurately recognize vehicles, pedestrians, and traffic signs in complex traffic environments[3]. Vision security surveillance systems must identify and track dynamic targets in real-time and with high robustness[4]. Augmented reality systems must precisely align virtual category labels with the physical objects in the real scene to ensure an immersive user experience[5]. Conventional complementary metal-oxide-semiconductor (CMOS) vision systems are inefficient in realizing intelligent vision systems due to the rigid sensing-computing architecture and bio-inspired functionalities lacking device[6-11]. The abundant data transfer and conversion between sensors, computing units and memory lead to high power consumption and time latency. The lack of devices with efficient biologically emulated processing capability also decreases the area and power efficiency, since multiple transistors, resistors, and capacitors are required for implementing a single function.

Neuromorphic vision device-based intelligent vision systems process visual information efficiently in a biologically emulated way. However, their reconfigurability is limited. They are usually limited to only static image preprocessing (such as contrast enhancement, noise reduction, and visual adaptation) based on frame-based processing, or only dynamic scenarios preprocessing (such as motion extraction) based on event-based processing[12-16]. Besides, current devices are usually capable of sensing only single-dimensional information from optical signals (such as only intensity or only wavelength), lacking the advanced processing capability to process rich and complex visual information. At the system level, they usually lack effective co-design among circuits, algorithms, and system architecture for fully utilizing the characteristics of the device, resulting in limited processing accuracy and efficiency. Therefore, to achieve truly intelligent vision systems, we need innovations at both device level and system level.

To address the limited reconfigurability, we present a reconfigurable intelligent vision system with both high flexibility and power efficiency. In this system, a unique multi-paradigm neuromorphic vision device with versatile behaviours is utilized. We first show a device-algorithm co-design making use of device's behaviours, achieving four sensory computing modes, including 1) PN and 2) PS modes with spiking and non-spiking in-sensor low-/mid-level processing capabilities for NI functions, and 3) ES and 4) EN modes with spiking and non-spiking in-memory high-level processing capabilities for AI functions. Further, we co-design mode control circuits that enable flexible switching among device modes and support multiple array architectures, achieving array-level reconfigurability. Finally, with the co-designed reconfigurable dataflow architecture, a highly reconfigurable neuromorphic vision system achieving full spiking/non-spiking NI/AI reconfigurability across device, cell, array and system levels is realized. The system supports flexible NI-centric to AI-centric architecture reconfigurations in both spiking and non-spiking modes and, correspondingly, NI/AI flexible algorithm deployment depending on real-time scenario demands. NI-centric architectures and algorithms are specifically designed for high-resolution, high-quality smart imaging, while AI-centric configurations are designed for versatile, high-accuracy and efficient motion and static recognition tasks. The flexibility of reconfiguring to different NI/AI sizes in our system allows maximal resource allocation within a unified array, thus enabling high accuracy, power and area efficiency (up to 52.6 TOPS/W for NI-centric computing and 75.5 TOPS/W for AI-centric computing) compared with other vision systems.

We also propose two intelligent vision systems aiming for two complex and demanding tasks. To overcome the lack of advanced processing capability when handling multi-target detection tasks, we take inspiration from the biological visual system, which efficiently and robustly processing multidimensional visual information following a hierarchical pathway of multiple levels of visual filters. A multidimensional and hierarchical intelligent vision system based on a single MoS_2/HZO FeFET array is demonstrated in this work, fully emulating the low-level visual filter in retina for low-level colour clustering, as well as mid- and high-level visual filters for targets segmentation and recognition. The novel multidimensional and multilevel MoS_2/HZO FeFET-based visual filter device exhibits enhanced wavelength selectivity, which is enabled by the polarization-modulated interfacial barrier mechanism, allowing for the low-level colour clustering and mid-level multidimensional decoupling. It also shows non-volatile polarization-enhanced multi-photoresponsivity states characteristics for high-level processing. The photoresponsivity of the MoS_2/HZO FeFET device can be tuned by input intensity, wavelength, and the stored polarization states. A multidimensional and intelligent vision system is built by employing a single MoS_2/HZO FeFET array. Combined with the co-designed novel HVBF algorithm, the system can perform low-level colour clustering, mid-level image segmentation and RGB noise filtering, and high-level recognition, achieving high accuracy in two complex and noisy scenarios including recognizing coloured digits (93.3%) and identifying two overlapping digits in a noisy background (93.8%). The implemented intelligent vision system shows promising potential in the multi-target scenarios with high accuracy, noise-resilience and high integration level.

Additionally, to solve the problem of limited 3D scene understanding capability involving depth and visual appearance perception, we design an intelligent vision system integrating an RGB-D camera based on MoS_2/In_2Se_3 heterostructure device. The device features a bipolar photoresponse to RGB and NIR light input, effectively decoupling the two signals. Under positive bias, the device produces a positive photoresponse to NIR light pulse input, while RGB light results in negative photoresponse owing to the visible light-induced interfacial barrier-lowering effect at the source side of the device. When combined with a near-infrared marker-based depth perception technology, the MoS_2/In_2Se_3 array-based RGB-D camera can acquire and decouple a 2D RGB image and a 3D depth image at the same time, effectively obtaining both visual appearance and 3D location data of the input scene. An intelligent visual system with 3D scene understanding capability is realized, including the RGB-D camera at the front end for sensing and decoupling the depth and RGB information, and two neural networks at the back end for processing the acquired information. A SqueezeNet CNN is used to extract the spatial features from visual appearance information for static object recognition. The long short-term memory (LSTM) network is used to analyse the temporal features of 3D location information for dynamic gesture recognition. In the augmented reality (AR) task for the two-hand-interacting-with-one-object scenarios, the system achieves both accurate schematic alignment and spatial alignment between virtual augmentations and their real-world counterparts. It can recognize both the static objects and dynamic gestures, and place the corresponding virtual annotations near their real-world counterparts based on the acquired 3D location data with 91.8% accuracy.

The main content of this thesis is as follows. Chapter 1 introduces the motivation behind the projects, focusing on intelligent vision systems and devices. Chapters 2-4 introduce the three main projects. Chapter 5 summarizes the projects and discusses potential directions for future research.

Chapter 2 Reconfigurable intelligent vision system based on multi-paradigm Pd/CuO_x/ITO device

2.1. Introduction

An intelligent vision system is designed with two things in mind: ultra-flexibility (reconfigurability) for different scenario adaptation and ultra-high power and area efficiency for an intended application. However, the current vision system falls short since it is not possible to achieve all concurrently. This is especially critical for edge computing with hardware and power resource limitations.

Conventional intelligent vision systems are limited by specialized processing paradigms, such as spiking-based processing for event-based motion scenes[17,18] or non-spiking based processing methods for frame-based static scenes[19], which lead to reduced adaptability and computational inefficiency in dynamic and unpredictable real-world environments. While current state-of-the-art reconfigurable CMOS intelligent vision systems have recently made some progress in realizing spiking/non-spiking reconfigurability and algorithmic deployment (at the chip and block level) in front-end imaging hardware or back-end AI processors[6-9,20], their performance are still limited. This is because they are inefficient from both device and architectural perspective. Specifically, at the device cell level, they largely rely on devices that are heterogeneously integrated and lack biological functionality, and lack multi-paradigm reconfigurability and implementable device-level algorithms. Furthermore, at the architecture level, rigid sensing computing architectures hinder the flexible deployment of algorithms in various scenarios with specific requirements (e.g., high-resolution smart imaging scenarios requiring neuromorphic imaging (NI) and high-precision decision-making scenarios requiring artificial intelligence (AI)). Ideal AI vision systems are expected to alleviate these problems by co-designing novel hardware algorithms at two different levels. The first level is to assess its higher-order behaviour from the base device cell level, and the second level is to implement it at the system level for a variety of different scenarios and applications.

Guidance is therefore taken from a biological visual system that is highly adaptive to ever-changing scenarios and performs reliable visual sensing and processing with great on-demand intelligence, flexibility and efficiency. Visual neurons in the retina and cortex support multiple levels of sensing and processing, spanning low, mid and high levels through spiking and non-spiking signalling patterns. In addition, due to its rich elemental dynamics, inherent structural homogeneity and large connectivity, the human visual system is able to operate in a task-dependent manner through 'cognitive' architectures[21-25], which can rearrange its connections in an ultra-flexible manner for optimal efficiency based on the computational demands of real-time tasks. While the state-of-the-art emerging reconfigurable neuromorphic vision system designs have unique biological-emulated processing capabilities, current designs suffer from insufficient multi-paradigm reconfiguration capabilities at the device level, thus limiting their ability to seamlessly integrate multi-level processing in both spiking and non-spiking modes. In addition, they are unable to support full device-level spiking

/non-spiking/NI/AI algorithm deployment, ultimately leading to reduced power efficiency and area efficiency. In addition, reconfigurable circuits and efficient algorithmic designs still fall short of fully supporting hierarchical reconfigurability from device to system level for different scenarios, and these aspects remain to be addressed[12,13,26-32].

To address these challenges, we propose a NI/AI reconfigurable intelligent vision system (**Fig. 2.1**a). The system employs a unique multi-paradigm neuromorphic vision device and a highly reconfigurable sensing-computing architecture, and co-designed reconfigurable circuits and visual algorithms (**Fig. 2.1**b and c). We first demonstrate a device-algorithm co-design that takes advantage of the multi-dynamics of the device to implement four sensing computation modes, including 1) photo-spiking neuron (PN) and 2) photo-synaptic (PS) modes with spiking and non-spiking in-sensor low-/mid-level processing capabilities for NI functions, and 3) electrical-synaptic (ES) and 4) electrical-neuron (EN) modes with spiking and non-spiking in-memory high-level processing capabilities for AI functions. In addition, we demonstrate a highly reconfigurable neuromorphic vision system with full spiking/non-spiking NI/AI reconfigurability across device, cell, array, and system levels (**Fig. 2.1**d). The system supports flexible NI-centric to AI-centric architecture reconfigurations in both spiking and non-spiking modes, and accordingly supports flexible deployment of NI/AI algorithms according to real-time scene requirements. The NI-centric architecture and algorithms are designed for high-resolution, high-quality intelligent imaging with a unified device cell array structure, and all device cells are assigned to either PN (spiking) or PS (non-spiking) modes for frameless motion and frame-based static scenes, respectively (**Fig. 2.1**e). In contrast, the AI-centric configurations allocate the device cells with a smaller NI proportion in PN or PS mode and a greater proportion in ES and EN modes for spiking and non-spiking AI, respectively, favouring versatile, high-accuracy and efficient motion and static recognition tasks. Our system can be flexibly reconfigured to different NI/AI sizes, thus allowing for maximum resource allocation within a unified array, achieving high accuracy, power and area efficiency compared to other vision systems (NI vision systems, AI vision systems and non-reconfigurable hybrid NI/AI vision systems) (**Fig. 2.1**e).

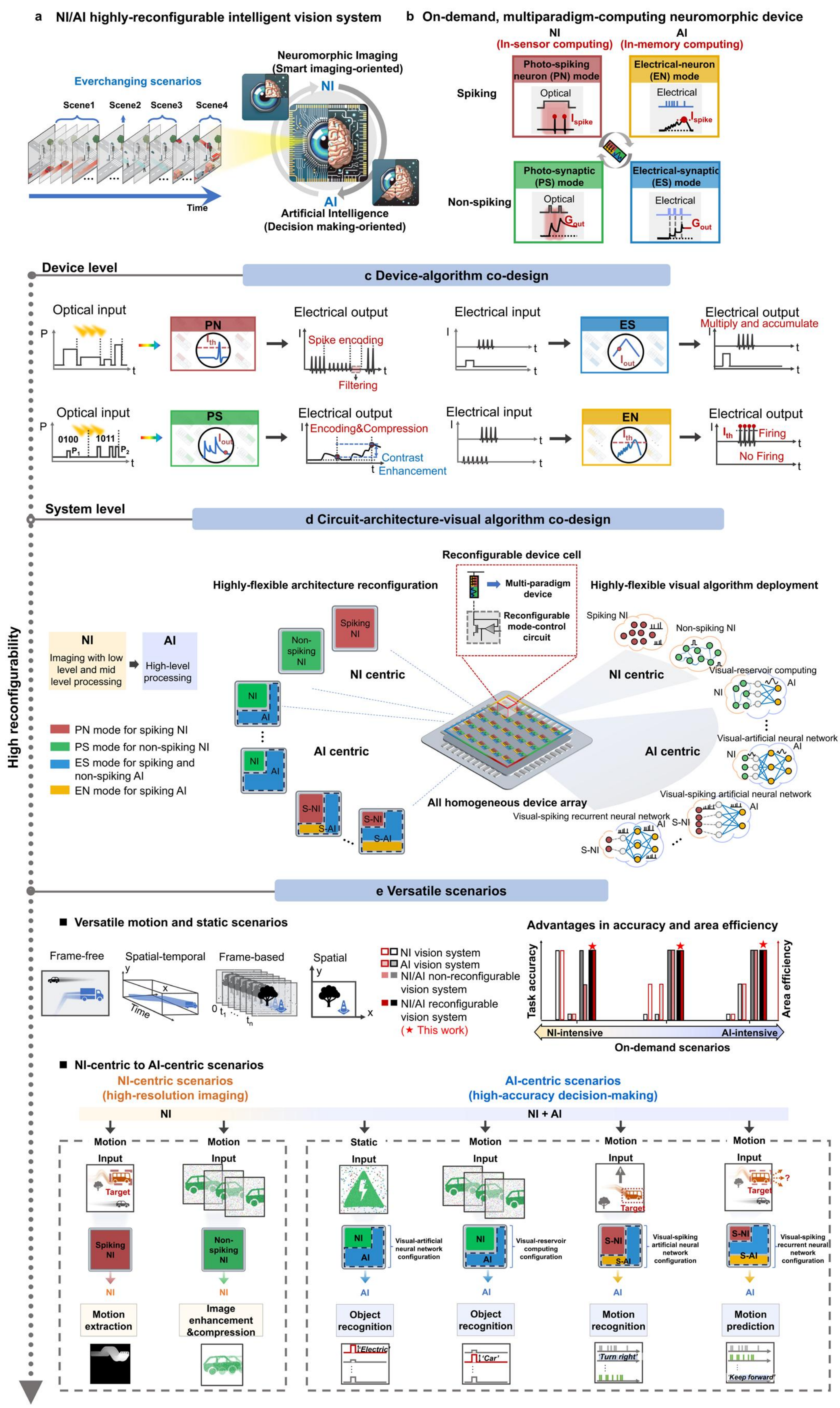

a NI/AI highly-reconfigurable intelligent vision system
Everchanging scenarios
Scene1 Scene2 Scene3 Scene4
Time
Neuromorphic Imaging (Smart imaging-oriented)
NI
AI
Artificial Intelligence (Decision making-oriented)
b On-demand, multiparadigm-computing neuromorphic device
NI (In-sensor computing)
AI (In-memory computing)
Spiking
Non-spiking
Photo-spiking neuron (PN) mode
Electrical-neuron (EN) mode
Photo-synaptic (PS) mode
Electrical-synaptic (ES) mode
Optical
Electrical
Device level
c Device-algorithm co-design
Optical input
Electrical output
Spike encoding
Filtering
Electrical input
Multiply and accumulate
Encoding&Compression
Contrast Enhancement
Firing
No Firing
System level
d Circuit-architecture-visual algorithm co-design
Reconfigurable device cell
Multi-paradigm device
Reconfigurable mode-control circuit
Highly-flexible architecture reconfiguration
Highly-flexible visual algorithm deployment
NI Imaging with low level and mid level processing
AI High-level processing
PN mode for spiking NI
PS mode for non-spiking NI
ES mode for spiking and non-spiking AI
EN mode for spiking AI
NI centric
AI centric
All homogeneous device array
Spiking NI
Non-spiking NI
Visual-reservoir computing
Visual-artificial neural network
Visual-spiking artificial neural network
Visual-spiking recurrent neural network
High reconfigurability
e Versatile scenarios
Versatile motion and static scenarios
Frame-free
Spatial-temporal
Frame-based
Spatial
NI vision system
AI vision system
NI/AI non-reconfigurable vision system
NI/AI reconfigurable vision system
(★ This work)
Advantages in accuracy and area efficiency
Task accuracy
Area efficiency
NI-intensive
AI-intensive
On-demand scenarios
NI-centric to AI-centric scenarios
NI-centric scenarios (high-resolution imaging)
AI-centric scenarios (high-accuracy decision-making)
NI
NI + AI
Motion
Static
Input
Target
Visual-artificial neural network configuration
Visual-reservoir computing configuration
Visual-spiking artificial neural network configuration
Visual-spiking recurrent neural network configuration
Motion extraction
Image enhancement &compression
Object recognition
Motion recognition
Motion prediction
'Electric'
'Car'
Turn right
Keep forward

Fig. 2.1 Highly-reconfigurable intelligent vision system with hierarchical reconfigurability across device to system levels. (a) A highly flexible NI/AI reconfigurable intelligent vision system enabled by (b) a multi-paradigm neuromorphic vision device with co-existence of four visual processing modes, mimicking the elemental functionalities in the retina (NI) and cortex (AI). (c) Device-algorithm co-design of achieving four visual processing modes including: PN mode with in-sensor spike encoding and filtering capabilities, PS mode with in-sensor data compression and enhancement capabilities, ES mode with in-memory multiplication capability, and EN mode with in-memory threshold detection capability. (d) Highly-flexible architectural reconfigurability across device, array and system levels, supporting various NI-centric to AI-centric configurations and algorithm deployments with different NI/AI sizes. (e) The system supports a broad spectrum of intelligent vision applications, spanning from spatiotemporally frame-free motion to spatially frame-based static scenarios and NI-centric (high-resolution and high-quality smart imaging with pre-processing) to AI-centric (high-efficiency and high-accuracy recognition and prediction) scenarios. Configurations including NI-centric spiking and non-spiking configurations, and AI-centric configurations such as visual-spiking recurrent neural network, visual-spiking artificial neural network, visual-artificial neural network, and visual-reservoir computing, and versatile scenarios including motion extraction, image enhancement, image compression, object recognition, motion recognition, and motion direction prediction are illustrated here. As compared to other vision systems (NI vision system, AI vision system, or hybrid NI/AI vision system), our designed system excels whether in terms of task accuracy or power and area efficiency.

2.2. Results

2.2.1. Reconfigurable system hardware prototype

We experimentally demonstrate a reconfigurable system hardware prototype, consisting of a multi-paradigm device array and reconfigurable circuitry. **Fig. 2.2** schematically illustrates the reconfigurable prototype architecture based on a reconfigurable cell array, with each cell integrating a multi-paradigm device and a reconfigurable mode-control circuit. We fabricated the multi-paradigm device with a stack structure of Pd/CuO_x/ITO (**Fig. 2.3**a) and a 12 × 12 device array (**Fig. 2.3**b). Each device in the array exhibits device-level reconfigurability (PS, PN, ES, EN modes). Further, a reconfigurable system prototype is demonstrated as shown in **Fig. 2.3**c. The reconfigurable mode-control circuitry is implemented with peripheral modules including transimpedance amplifiers (TIAs), analog-to-digital converters (ADCs), and a field-programmable gate array (FPGA), enabling dynamic switching between different device modes and array architectures. The hardware prototype design supports hierarchical reconfigurability across device, cell and array levels, which are experimentally demonstrated in the following sections.

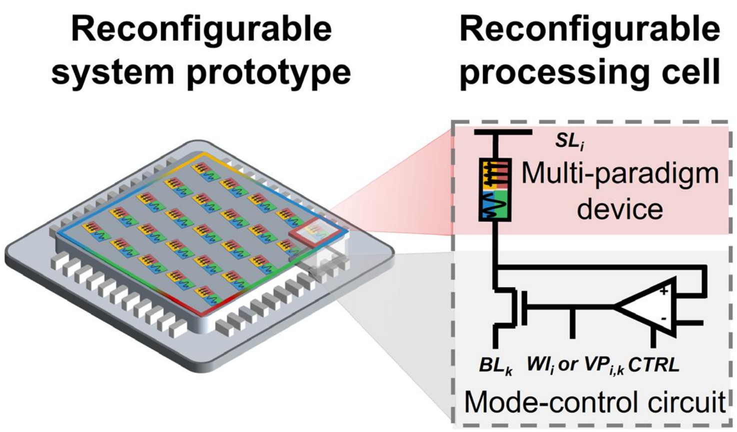


Fig. 2.2 3D-stacked homogeneous array architecture. The multi-paradigm device array is stacked on top of CMOS implemented mode-control circuits array. Each reconfigurable

processing cell consists of a multi-paradigm device stacked over a mode-control circuit.

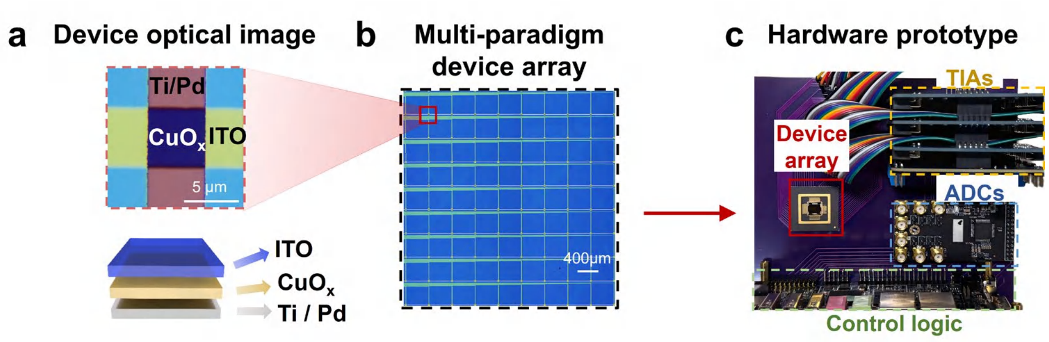


Fig. 2.3 Reconfigurable system hardware prototype. (a) Optical image of the multi-paradigm $Pd/CuO_x/ITO$ device. (b) Optical image of a 12 × 12 multi-paradigm device array. (c) Photo of the board-level reconfigurable system prototype.

2.2.2. Multi-paradigm device

2.2.2.1. Device architecture and computation modes

The two-terminal multi-paradigm device consists of a $Pd/CuO_x/ITO$ stack. The bottom electrode was first patterned using optical lithography, followed by 35 nm-thick Pd deposited via electron-beam evaporation and a lift-off process. Thereafter, the sandwich layer was patterned using optical lithography, followed by 40 nm CuO_x layer deposited via RF sputtering using a CuO target and a lift-off process. Finally, the top electrode was formed by optical lithography patterning, deposition of 40 nm ITO via DC sputtering, followed by lift-off process. The same process steps were used for the device array fabrication except that the bottom electrodes of the same column of devices were connected and the top electrodes of each device were connected individually.

Four different behaviours corresponding to different sensory computing modes can be accessed in different voltage ranges, including PN mode for V_{bias}= 0 V, PS mode for -0.1 V < V_{bias} < -0.04 V, ES mode for 0.1 V < V_{bias} < 0.5 V, and EN mode for V_{bias} >

1.3 V as shown in the basic *I-V* characteristics (**Fig. 2.4**). All sensing computation modes operate at low power consumption, resulting in an efficient intelligent vision system.

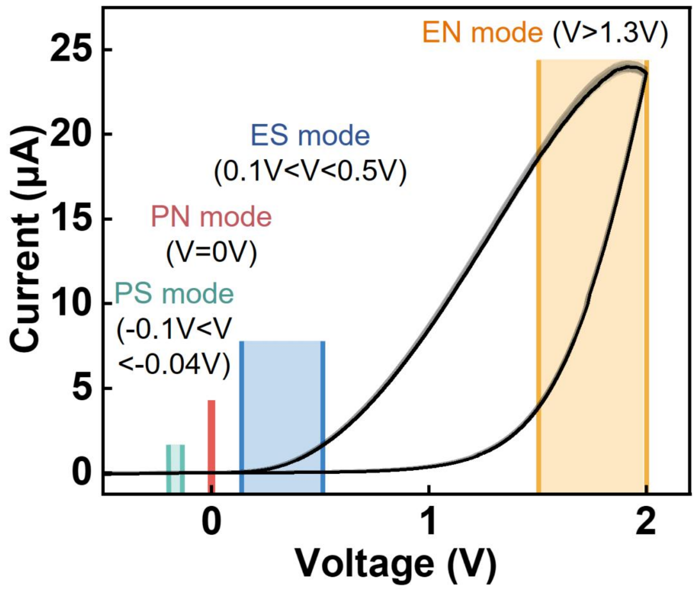


Fig. 2.4 Basic *I-V* characteristics of the device with a bidirectional DC sweeping voltage from 0 V to 2 V to 0 V to -2 V to 0 V with 50 cycles. Four modalities are accessed under different voltage regimes: PN mode at 0V, PS mode at -0.1 V to -0.04 V, ES mode at the range of 0.1-0.5 V and EN mode at a higher voltage range of >1.3V.

The estimated cost per device of multi-paradigm device is shown in **Table 2.1**, which is comparable to that of a conventional CMOS pixel[33,34].

Table 2.1 Estimated cost per device for fabricating Pd/CuO_x/ITO device

Layer	Estimated fabrication cost (USD)	CMOS pixel cost (USD)
Pd	9.125×10^{-7}	
CuO_x	1.093×10^{-6}	
ITO	1.093×10^{-6}	
Total cost	3.099×10^{-6}	10^{-6} - 10^{-5}

* The cost conversion is based on an exchange rate of 1 USD = 7.05 RMB.

2.2.2.2. Device mechanisms

The photoresponsive LIF (leaky integrate and fire) behaviour arises from the dynamic trapping and de-trapping of electrons at the ITO/CuO_x interface (**Fig. 2.5**a). Upon illumination, photoexcited electrons are generated in CuO_x, some of which are collected by the ITO electrode, producing a baseline photocurrent. At the same time, the remaining electrons occupy trap states at the ITO/CuO_x interface. As these interfacial trap states become populated, the barrier width and effective barrier height (the electron barrier seen from the CuO_x side to the ITO side) at the interface gradually decrease, corresponding to the integration phase. Once the barrier narrows sufficiently and reaches a tunnelling threshold, the trapped electrons are detrapped and released by the illumination and tunnel through the ITO/CuO_x interfacial barrier, resulting in a sharp spiking current (the firing event). Afterwards, the barrier at the ITO/CuO_x interface naturally returns to its original width, completing a self-recovery cycle.

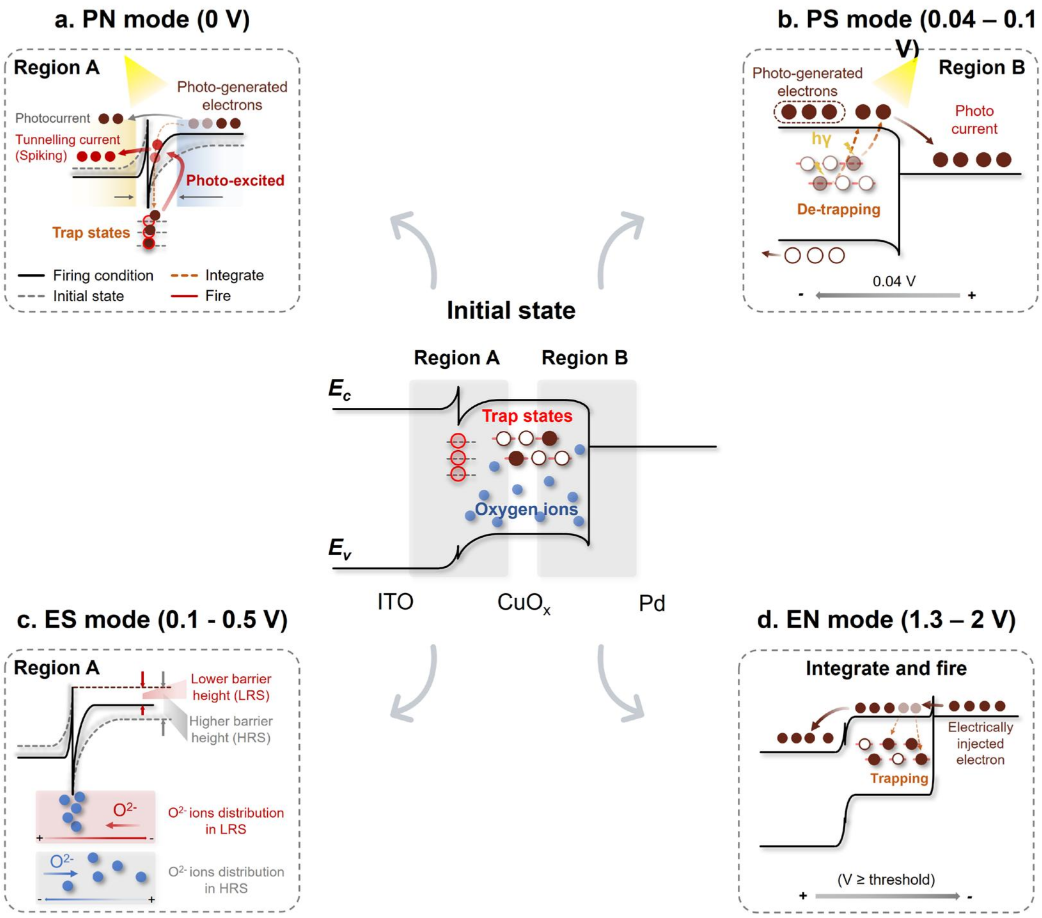


Fig. 2.5 Device mechanisms. (a) PN mode photoresponsive LIF mechanism (0 V bias). The integration and firing processes are indicated. (b) PS mode photoresponse mechanism under the light illumination (-0.1 to -0.04 V bias). (c) ES mode switching mechanisms, indicating the transition from high resistance state (HRS) to low resistance state (LRS) (0.1 to 0.5 V bias). (d) EN mode electrical LIF behaviour mechanism, indicating the integration, firing, and recovery process (1.3 - 2 V bias).

The observed photosynaptic response originates from the capture and release dynamics of photo-generated electrons within the trap states of CuO_x (**Fig. 2.5**b). When exposed to light, electron–hole pairs are generated in the CuO_x film. These charge carriers are driven apart by a weak applied electric field, directing electrons toward the Pd electrode and holes toward the ITO electrode, thereby enhancing the photocurrent.

Simultaneously, electrons within the trap states of the CuO_x are also excited and contribute to the photocurrent. As illumination continues, the current increase begins to slow down due to saturation of both photogeneration and de-trapping processes. Once the light is turned off, the free carriers rapidly recombine, while some excess electrons become re-trapped in the CuO_x trap states, resulting in a non-linear decay of the current over time, characteristic of short-term synaptic plasticity.

The analogue-type resistive switching in ES mode originates from the change of the interfacial energy barrier, which is regulated by the movement of oxygen ions within the CuO_x layer (**Fig. 2.5**c). In the initial state, the random distribution of oxygen ions results in moderate resistance. The set and reset operations are described in the following two stages, where 1) Set to low resistance state (LRS): when a small positive voltage is applied to the ITO electrode, oxygen ions migrate towards and into the ITO electrode. This results in a net hole doping effect, causing the Fermi level in the CuO_x layer to shift downwards, reducing the effective electron barrier height; 2) Reset to high resistance state (HRS): when a negative voltage is applied, the oxygen ions move back into the CuO_x layer eliminating the net hole doping effect, thus raising the effective electron barrier height.

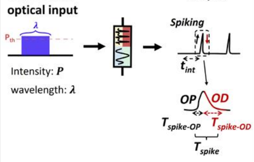
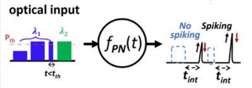
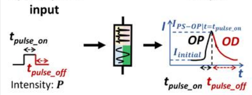
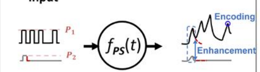
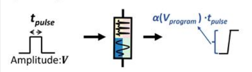
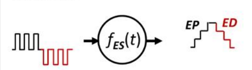
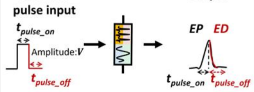
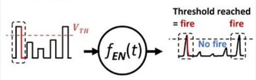
The LIF behaviour in the EN mode arises from the modulation of the interfacial energy barrier and the dynamic trapping of electrons within the CuO_x film (**Fig. 2.5**d). Initially, the device resides in a high resistance state (HRS). When a sequence of positive voltage pulses is applied to the ITO electrode, the barrier for electron injection from the Pd side is effectively lowered, allowing electrons to migrate into the CuO_x

layer and drift toward the ITO electrode. During this process, a portion of the injected electrons are captured in trap states within CuO_x, producing a net doping effect that enhances the CuO_x film conductivity, corresponding to the leaky integration process. Upon cessation of the voltage pulses, the device enters a recovery phase, during which the effective barrier height increases, causing an initial steep drop in current, followed by a slower decay as trapped charges are gradually released. Eventually, the system reverts to its original high-resistance state, thus completing the LIF cycle.

2.2.2.3. Device behaviours and co-designed algorithms for Each Operating Mode

We present the experimental measurements of the basic behaviours in the four sensory computing modes. We then model the behaviour of the device and present the co-designed device algorithms for each sensory computing mode (**Table 2.2**), for use in the subsequent system-level algorithms for dedicated functions. Descriptions of each sensory computing mode are presented in the subsections.

Table 2.2 Summary of the four sensory computing modes basic behaviours, together with the corresponding implementable algorithms and applications.

Operating mode			Model equations	Applications
Photo-spiking neuron	Continuous optical input / Output 	Basic Behaviour	*Basic PN mode behaviour for one LIF period (0-t$_0$):* $F_{PN}(P,\lambda,I,t) = \begin{cases} I = I_{baseline}, & \text{Condition 1} \\ I_{PN-OP} = f_{PN-OP}(P,\lambda,t) = I_{baseline} + \beta(P,\lambda)\cdot(-e^{-\alpha(P,\lambda)\cdot t}+1), & \text{Condition 2} \\ I_{PN-OD} = f_{PN-OD}(P,\lambda,t-t_{spike-OP}) = I_{PN-OP \vert t=T_{spike-OP}} \cdot e^{-\gamma(t-T_{spike-OP})}, & \text{Condition 3} \\ I = I_{baseline}, & \text{Condition 4} \end{cases}$ $Condition\ 1: t = 0\ to\ t = T_{integration}$ $Condition\ 2: t = T_{integration}\ to\ t = T_{integration} + T_{spike}$ $Condition\ 3: t = T_{integration}\ to\ t = T_{integration} + T_{spike}$ $Condition\ 4: t = T_{integration} + T_{spike}\ to\ t = t_0$ $\gamma \propto I;$	• In-sensor motion extraction and filtering
	Continuous optical input / Output 	Algorithm	*Device algorithm for spike encoding and filtering (0-T):* $n = \frac{T}{T_{spike} + T_{integration}}$ $f_{encoding\,\&\,filtering}(P,t,\lambda) = U_{fire}(P,t)\cdot\sum_{i=1}^{n} F_{PN}(P,\lambda,I,t)$ $U_{fire}(P,t) = \begin{cases} 1, & if\ P > P_{th}\ and\ t > t_{th} \\ 0, & Otherwise \end{cases}$	
Photo-synaptic	Optical pulse input / Output 	Basic Behaviour	$Basic\ PS\ mode\ behavior\ for\ one\ optical\ pulse\ period\ (0-t_0):$ $F_{PS}(P,t) = \begin{cases} I_{PS-OP} = f_{PS-OP}(P,t) = I_{initial} + \alpha(I,P)\cdot t, & t = 0(t_{pulse\_on})\ to\ t = t_{pulse\_off} \\ I_{PS-OD} = f_{PS-OD}(P,t) = A_1(P,t)e^{-\gamma_1(t-t_{pulse\_on})} + A_2(P,t)e^{-\gamma_2(t-t_{pulse\_on})}, & t = t_{pulse\_off}\ to\ t_0 \end{cases}$ $\gamma_1,\gamma_2 \propto I$	• Image compression • Image enhancement
	Optical pulse input / Output 	Algorithm	$Device\ algorithm\ for\ data\ compression\ (0-T):$ $f_{data-compression}(P,n_c) = \sum_{i=1}^{n_c} F_{PS}(P_i,t_i)$ $For\ i^{th}\ optical\ pulse, one\ pulse\ period\ (0-t_i):$ $if\ input\ pulse\ intensity = \text{"1"},$ $F_{PS}(P,t) = \begin{cases} I_{initial} + \alpha(I,P)\cdot t, t = 0\ (t_{pulse\_on})\ to\ t = t_{pulse\_off} \\ A_1(P,t)e^{-\gamma_1(t-t_{pulse\_on})} + A_2(P,t)e^{-\gamma_2(t-t_{pulse\_on})}, t = t_{pulse\_off}\ to\ t = t_i \end{cases}$ $if\ input\ pulse\ intensity = 0, F_{PS}(P,t) = A_1(P,t)e^{-\gamma_1\cdot t} + A_2(P,t)e^{-\gamma_2\cdot t}$ $Device\ algorithm\ for\ optical\ contrast\ enhancement\ (0-T):$ $f_{optical-contrast-enhancement}(P,t) = \sum_{i=1}^{n} A_1(P_i,t_i)e^{-\gamma_1(T-\sum_{i=1}^{n} t_i)} + A_2(P_i,t_i)e^{-\gamma_2(T-\sum_{i=1}^{n} t_i)}$	
Electrical-synaptic	Electrical pulse input / Output 	Basic Behaviour	$Basic\ ES\ mode\ behavior\ for\ one\ electrical\ pulse\ period\ (0-t):$ $F_{ES} = \begin{cases} G_{ES-EP} = f_{ES-EP}(V,t) = G_{initial} + s(V)\cdot t_{pulse}, & V \geq 0 \\ G_{ES-ED} = f_{ES-ED}(V,t) = G_{initial} - s(V)\cdot t_{pulse}, & V < 0 \end{cases}$	• Multiply and accumulate (MAC) function
	Electrical pulse input / Output 	Algorithm	$Device\ algorithm\ for\ multiplication\ (0-T):$ $G_{ES-EP}(V_{program},n) = f_{ES-EP}(V_{program},n) = G_{initial} + \sum_{i=1}^{n} s(V_{program})\cdot t_{pulse}$ $G_{ES-ED}(V_{program},n) = f_{ES-ED}(V_{program},n) = G_{initial} - \sum_{i=1}^{n} s(V_{program})\cdot t_{pulse}$ $f_{multiply}(V_{input}, i = n) = V_{input}\cdot G_{ES-EP/ED}(V_{i=n-1}, t_{i=n-1})$	
Electrical-neuron	Electrical pulse input / Output 	Basic Behaviour	$Basic\ EN\ mode\ behavior\ for\ i^{th}\ electrical\ pulse, for\ one\ pulse\ duration\ (0-t_i):$ $F_{EN}(V_i,t_i) = \begin{cases} I_{EN-EP} = f_{EN-EP}(V_i,t_i) = \beta(V,i)\cdot(-e^{-\alpha(V,i)\cdot t}+1), t = 0\ (pulse\_on) to\ t = t_{pulse\_off} \\ I_{EN-ED} = f_{EN-ED}(V_i,t_i) = I_{EN-EP}(V_i,t_i)\cdot e^{-\gamma t_{pulse\_off}}, t = t_{pulse\_off}\ to\ t_i \end{cases}$ $\gamma \to \infty$	• Comparing function
	Electrical pulse input / Output 	Algorithm	$Device\ algorithm\ for\ comparing\ function\ (0-T):$ $f_{comparing}(V) = \begin{cases} 1, & V \geq V_{th}\ and\ F_{EN}(V,t) \geq I_{th} \\ 0, & otherwise \end{cases}$	

2.2.2.3.1. PN mode

At a voltage bias of 0 V, the device operates in photo-spiking neuron (PN) mode with photoresponsive LIF spiking characteristics that can directly convert and encode the continuous light into current spikes, as shown in **Fig. 2.6**a. An increase in light intensity from 220 mW/cm^2 to 300 mW/cm^2 results in a larger spike amplitude (**Fig. 2.6**a upper panel). **Fig. 2.6**a (lower panel) shows the photoresponsive LIF spiking under different wavelengths (450 nm, 520nm and 638nm) under the light intensity of 300 mW/cm^2, exhibiting wavelength-dependent spiking amplitudes. The spiking current amplitude and baseline photocurrent shows a positive relation with light power from 20 mW/mm^2 to 220 mW/mm^2 (**Fig. 2.7**). It can be noted that the PN mode shows light intensity- and illumination duration-based threshold spiking features, where the spiking occurs only when the intensity and illumination duration exceeds their thresholds., i.e. spikes can be initiated only when light intensity > 20 mW/mm^2. These two features allow the device to be used for in-sensor motion extraction and filtering. In addition, the device exhibits LIF spiking behaviours across the visible spectrum (450nm, 520nm, and 638 nm) under different light intensities ranging from 20 mW/mm^2 to 220 mW/mm^2.

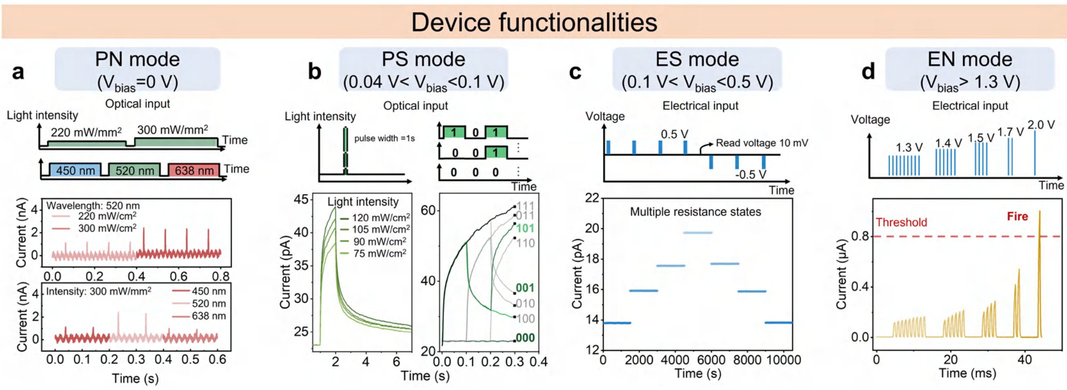


Fig. 2.6 Device behaviours. (a) PN mode device behaviour. Upper panel: photoresponsive LIF spiking characteristics in PN mode under light intensities of 220 mW/cm^2 and 300 mW/cm^2. Lower panel: photoresponsive LIF spiking characteristics of the device in PN mode under 450 nm, 520 nm and 638 nm light illumination with intensity of 300 mW/cm^2. (b) PS mode device

behaviour. Left panel: Device response under 520 nm optical pulse with varying optical pulse intensities ranging from 75 - 120 mW/cm^2. The pulse width is 1s. Right panel: Device response under eight different 3-bit light pulse stimuli patterns denoted by '000' to '111' for data compression (pulse intensities: 0 mW/cm^2 for '0' and 1200 mW/cm^2 for '1', pulse width = 100 ms, V_{read} = -40 mV, 520 nm wavelength). (c) ES mode device behaviour. LTP and LTD behaviours under 0.5 V and -0.5 V electrical pulses with a pulse width of 50 ms, respectively. (d) EN mode device behaviour. LIF characteristics in EN mode under electrical pulse stimuli with varying amplitudes and pulse numbers (from 1.3 V for 10 pulses to 2.0 V for 1 pulse), with a pulse width of 0.5 ms and a pulse interval of 0.5 ms.

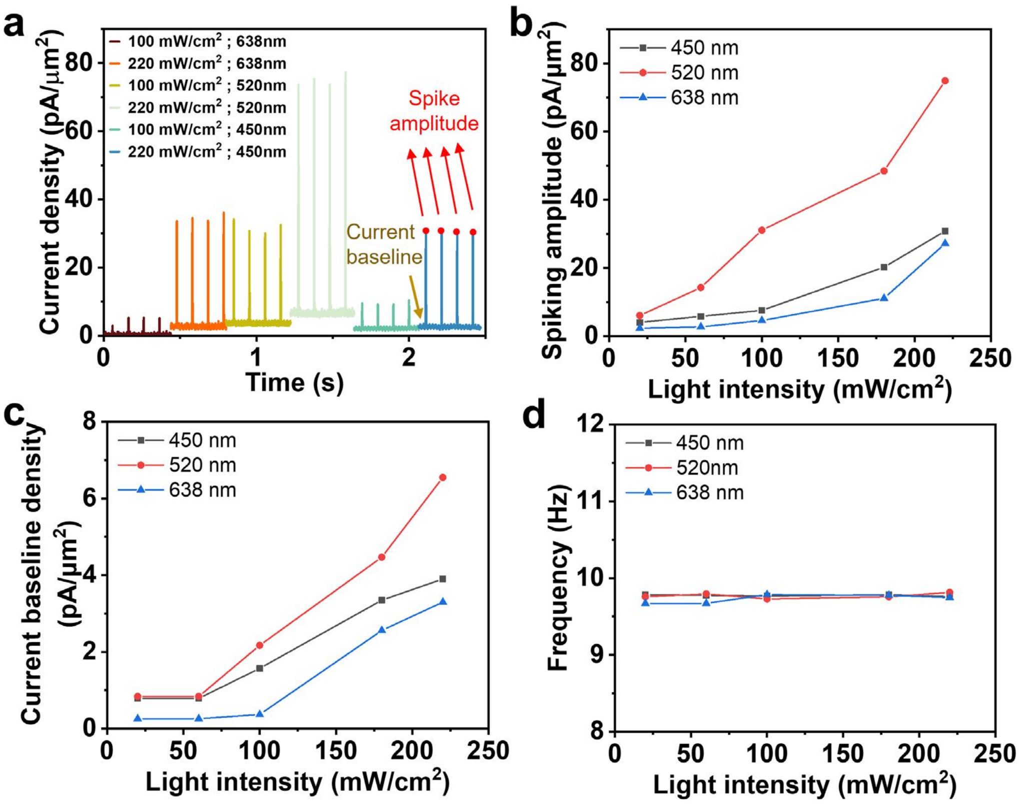


Fig. 2.7 Photoresponsive spiking characteristics under (a) 450 nm, 520 nm, and 638 nm light, each tested at different light intensities (100 mW/cm^2 and 220 mW/cm^2). Summary plots of (b) spiking amplitude and (c) current baseline density and (d) frequency as a function of light

intensity.

We conducted further experiments on the PN mode at temperatures of 240 K **(Fig. 2.8**a), 270 K (**Fig. 2.8**b), 300 K (**Fig. 2.8**c), 330 K (**Fig. 2.8**d) and 350 K (**Fig. 2.8**e). When the temperature is lower than 273 K, no photo-spiking behaviour is observed, attributed to the lower free carrier concentration and thermal energy of the carrier under the low temperature. At high temperatures (>350 K), the scattering process begins to dominate, which could contribute to noise, masking out the possible dynamics. Therefore, no photo-spiking was observed in these two cases.

At 273 K to 330 K, the device functions normally with photoresponsive LIF spiking behaviours. While the spiking amplitudes and integration time nearly remain unchanged within the temperature range (**Fig. 2.8**f and g), we see a more obvious reduction of spiking width of the device from 5.18 ± 0.5 ms to 4.03 ± 0.12 ms as the temperature goes up from 273 K to 330 K (**Fig. 2.8**h). This observation is consistent with proposed photo-spiking mechanism, as shown in **Fig. 2.5**a. The photo-spiking mechanism of the device can be understood by the charging/discharging of the electrons at the ITO/CuO_x interfacial trap states. In this process, the spiking current amplitude is mainly determined by the number of accumulated electrons at the interface. The integration period is determined by the duration required to attain the threshold barrier width necessary for Fowler-Nordheim (FN) tunnelling initiation, which is influenced by the interfacial trap site density and electron injection rate. As the electron trapping at the interface undergoes no significant change within the small temperature range, the spiking current amplitude and integration time exhibit minimal variation. On the other hand, the spiking width is defined by the temporal interval spanning from the initiation of FN tunnelling to the complete discharge of all trapped electrons. At a higher temperature, the discharging and tunnelling process is further aided by the thermionic emission of high-energy carriers, resulting in a quicker discharge time and shortened spiking period, which agrees well with observations.

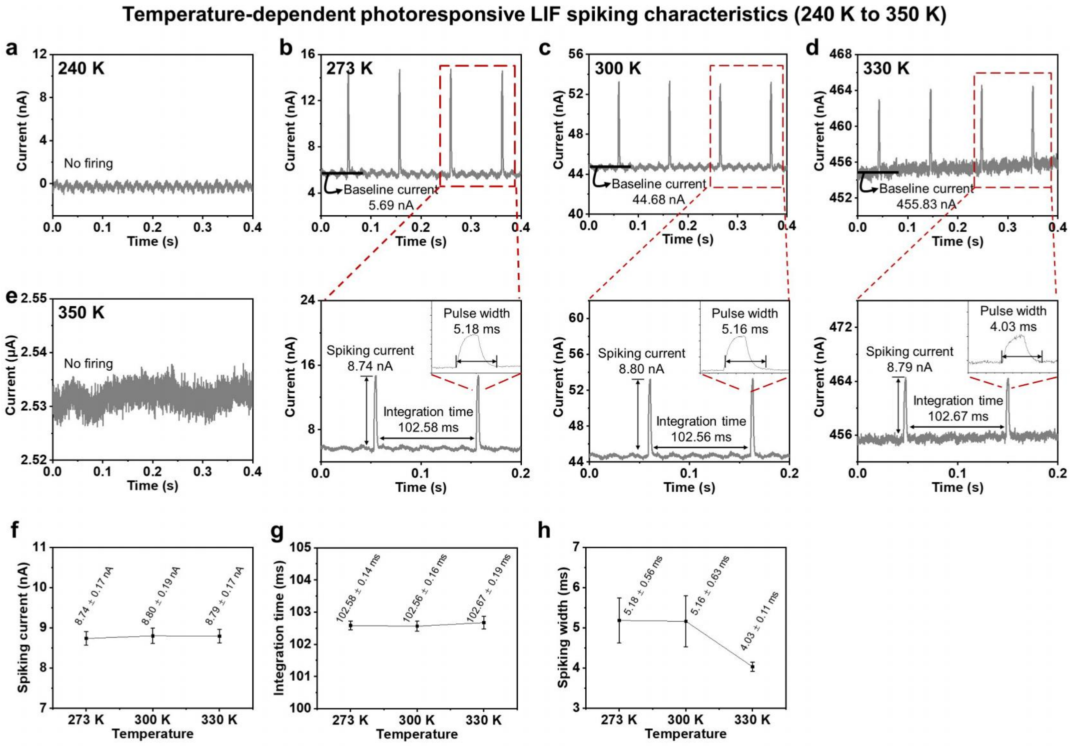


Fig. 2.8 Photoresponsive LIF spiking characteristics under different temperatures (a) 240 K, (b) 273 K, (c) 300 K, (d) 330 K and (e) 350K. The device features an area of 25 μm×25 μm. (f) Plot of spiking amplitude as a function of temperature. (g) Plot of integration time as a function of temperature. (h) Plot of spiking width as a function of temperature.

The basic behaviour of the device for one LIF spiking period and device algorithm describing the in-sensor spike encoding and filtering capabilities are described by the following functions (indicated in **Table 2.2**):

Basic PN mode behaviour for one LIF period (0-t_0):

$$F_{PN}(P,\lambda,I,t)=\begin{cases} I=I_{baseline}, & Condition\ 1\\ I_{PN-OP}=f_{PN-OP}(P,\lambda,\ t)=I_{baseline}+\beta(P,\ \lambda)\cdot\left(-e^{-\alpha(P,\lambda)\cdot t}+1\right), & Condition\ 2\\ I_{PN-OD}=f_{PN-OD}\left(P,\lambda,\ t-t_{spike-OP}\right)=I_{PN-OP\,|\,t=T_{spike-OP}}\cdot e^{-\gamma\left(t-T_{spike-OP}\right)}, & Condition\ 3\\ I=I_{baseline}, & Condition\ 4\end{cases} \tag{1}$$

$Condition\ 1: t=0\ to\ t=T_{integration}$

$Condition\ 2: t=T_{integration}\ to\ t=T_{integration}+T_{spike}$

$Condition\ 3: t=T_{integration}\ to\ t=T_{integration}+T_{spike}$

$Condition\ 4: t=T_{integration}+T_{spike}\ to\ t=t_0$

$$\gamma\propto I;$$

, where *spike-OP* and *spike-OD* are the potentiation section and depression section in one spike respectively, P is the applied light power, λ is the light wavelength, $I_{baseline}$ is baseline current, $T_{spike\text{-}OP}$ is the time required for the potentiation section, $T_{spike\text{-}OD}$ is the time required for the depression section, $T_{integration}$ is the time required before the firing occurs, T_{spike} is the time duration of each spike, β is the spike amplitude, α is the rise factor, and γ is the decay factor.

Building upon the basic behaviour of the PN mode, the device algorithm is designed correspondingly to perform spike encoding and filtering of the input optical data as follows.

Device algorithm for spike encoding and filtering (0-T):

$$n=\frac{T}{T_{spike}+T_{integration}}$$

$$f_{encoding\,\&\,filtering}(P,t,\lambda)=U_{fire}(P,t)\cdot\sum_{i=1}^{n}F_{PN}(P,\lambda,I,t) \tag{2}$$

$$U_{fire}(P,t)=\begin{cases}1, & if\ P>P_{th}\ and\ t>t_{th}\\ 0, & Otherwise\end{cases} \tag{3}$$

, where n is the number of spikes within period T, $U_{fire}(P,t)$ is the firing condition with respect to the light intensity and illumination time, P_{th} is the threshold light intensity for firing, t_{th} is the threshold integration time for firing. Based on the spike encoding and filtering device algorithms designed above, the device can encode the light information into spike amplitude and filter the light signal with below-threshold light intensity and integration time.

2.2.2.3.2. PS mode

The device operates in photo-synaptic (PS) mode under the bias range of -0.1 V to -0.04 V. In this mode, when subjected to light pulses stimulation, the device current initially undergoes continuous potentiation, followed by nonlinear relaxation after stimulus removal, as shown in **Fig. 2.6**b. The PS mode exhibits nonlinear synaptic plasticity-dependent on optical pulse intensity, pulse number, and pulse interval, as shown in **Fig. 2.9-2.11**. **Fig. 2.9** demonstrates that increasing the light pulse number from single to multiple enhances the plasticity, as indicated by the increased relaxation time and current amplitude. **Fig. 2.10** demonstrates that the light stimuli with higher frequency generate enhanced plasticity with longer relaxation time. In addition, the device exhibits photo-synaptic behaviours across the visible spectrum under 450 nm, 520 nm, and 638 nm. **Fig. 2.11** further investigates the pulse frequency-dependent plasticity under low input intensity (7.5 mW/cm^2). Under the same input intensity, the maximum photocurrent amplitude and relaxation time increase with increasing input intensity.

These properties enable in-sensor optical data enhancement (**Fig. 2.6**b (left panel)) and compression (**Fig. 2.6**b (right panel)). **Fig. 2.6**b (left panel) shows the current responses under varying light intensities (75–120 mW/cm^2, pulse width = 1 s, pulse number = 1), suggesting that higher intensities induce larger current amplitudes with slow decay rates, while lower intensities produce rapidly decaying currents. This temporal amplification of output differences enhances contrast between distinct light intensities. **Fig. 2.6**b (right panel) illustrates data compression achieved through the pulse temporal pattern-dependent nonlinear synaptic plasticity. This enables encoding of 3-bit light pulse sequences (“000”– “111”, where “1” represents a 100 ms light pulse at 1200 mW/cm^2 and “0” denotes 100 ms light pulse at 0 mW/cm^2) into eight distinct 1-bit current magnitudes, achieving data compression with a compression ratio of 3.

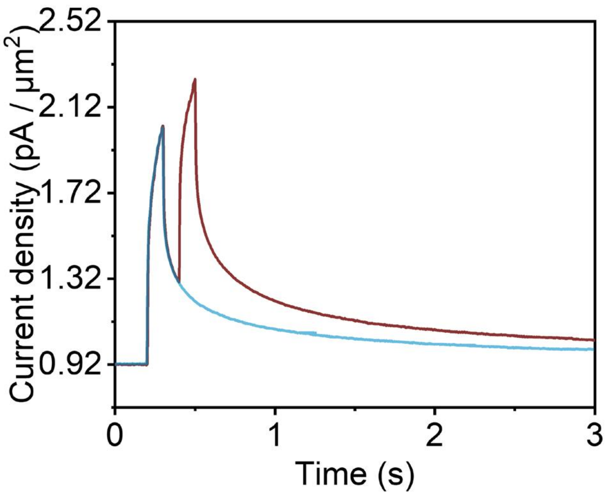


Fig. 2.9 Basic current response of multi-paradigm device in PS mode under single and multiple optical light pulses. The light intensity used is 1200 mW/cm^2, and the pulse width is 100 ms. The wavelength is 520 nm. A read voltage of -40mV is applied during the measurement.

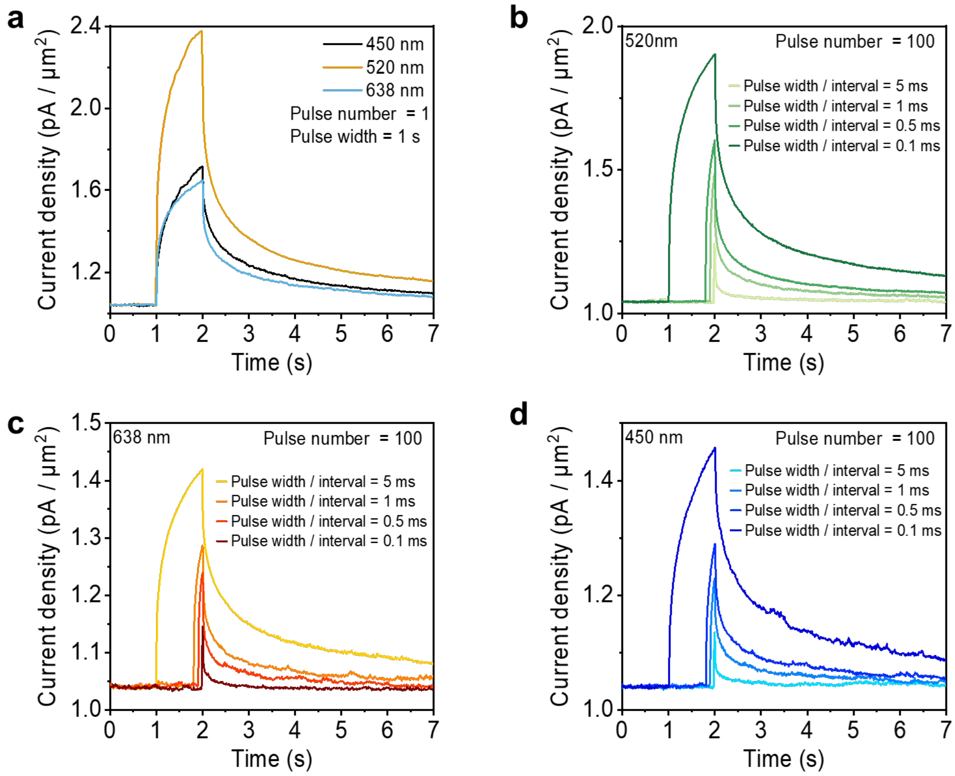


Fig. 2.10 Variation of the photoresponse of the multi-paradigm device in PS mode with light wavelength, pulse width and pulse interval. (a) Current response to light pulses of different wavelengths (450 nm, 520 nm, and 638 nm) under the same conditions of the same pulse width (1 s) and light intensity (120 mW/cm^2). Current response measured with different pulse widths and pulse intervals under the same 120 mW/cm^2 light intensity for (b) 520 nm, (c) 638 nm and (d) 450 nm light, respectively. All measurements were performed at a read voltage of -40mV.

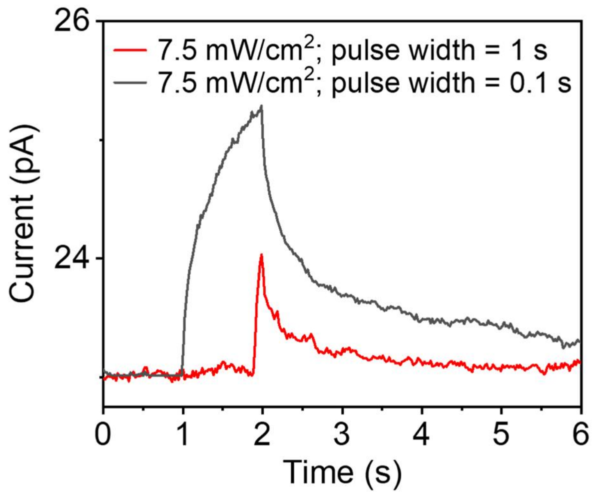


Fig. 2.11 Frequency-dependent optical response of the device in PS mode at 520 nm illumination, 7.5 mW/cm^2 light intensity. The photoresponse under 1 Hz and 10 Hz (50% duty cycle) light pulses is demonstrated. A read voltage of -40mV is applied.

The basic behaviour of the device in the PS mode within one optical pulse period and device algorithms describing data enhancement and compression functions are described by the following functions (indicated in **Table 2.2**):

$Basic\ PS\ mode\ behavior\ for\ one\ optical\ pulse\ period\ (0 - t_0)$:

$$F_{PS}(P,t) = \begin{cases} I_{PS-OP} = f_{PS-OP}(P,t) = I_{initial} + \alpha(I,P)\cdot t, & t = 0(t_{pulse\_on})to\ t = t_{pulse\_off} \\ I_{PS-OD} = f_{PS-OD}(P,t) = A_1(P,t)e^{-\gamma_1(t-t_{pulse\_on})} + A_2(P,t)e^{-\gamma_2(t-t_{pulse\_on})}, & t = t_{pulse\_off}\ to\ t_0 \end{cases} \quad (4)$$

, where *OP* is the optical potentiation process under light stimulation, *OD* is the optical depression process after stimulus removal, $I_{initial}$ is the instantaneous current before the start of the process, t_{pulse_on} is the time where the light is on, t_{pulse_off} is the time where the light is off, α is the rise factor, γ_1 and γ_2 are the decay factors corresponding to the two exponential decay processes, and A_1 and A_2 are the scale factors for the respective exponential decay components. We also extracted the decay constant γ_2 by fitting experimental photocurrent curves under varying light intensities using the double-exponential function. As shown in **Fig. 2.12**, γ_2 exhibits clear intensity-dependent behaviour, with higher illumination levels resulting in slower decay rates.

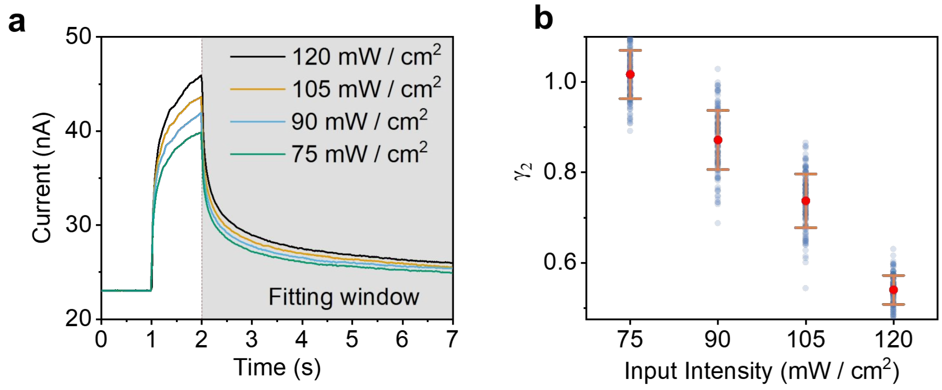


Fig. 2.12 Relationship between the decay constant γ_2 and input light intensity (a) Photoresponse in PS mode under different illumination intensities. The wavelength and pulse width are 520 nm and 1 s, respectively. The read voltage is -40 mV. The shaded area indicates the fitting window used for extracting decay constants. (b) The decay constant extracted by fitting the photocurrent response with a double exponential decay model under varying illumination intensities. Each blue dot represents the γ_2 value of one device in the 12 × 12 array . Red dots indicate the average γ_2 for each intensity level, and the orange error bars represent the standard deviation.

Building upon the basic behaviour of the PS mode, the device algorithm is designed correspondingly to perform contrast enhancement and data compression of the input optical data as follows:

$Device\ algorithm\ for\ optical\ contrast\ enhancement\ (0-T)$:

$$f_{optical-contrast-enhancement}(P,t)=\sum_{i=1}^{n}A_1(P_i,t_i)e^{-\gamma_1\left(T-\sum_{i=1}^{n}t_i\right)}+A_2(P_i,t_i)e^{-\gamma_2\left(T-\sum_{i=1}^{n}t_i\right)}\quad(5)$$

$Device\ algorithm\ for\ data\ compression\ (0-T)$:

$$f_{data-compression}(P,n_c)=\sum_{i=1}^{n_c}F_{PS}(P_i,t_i)\quad(6)$$

$For\ i^{th}\ optical\ pulse, one\ pulse\ period\ (0 - t_i)$:

$if\ input\ pulse\ intensity = "1"$,

$$F_{PS}(P,t) = \begin{cases} I_{initial} + \alpha(I,P) \cdot t, & t = 0\ (t_{pulse\_on})\ to\ t = t_{pulse\_off} \\ A_1(P,t)e^{-\gamma_1(t-t_{pulse\_on})} + A_2(P,t)e^{-\gamma_2(t-t_{pulse\_on})}, & t = t_{pulse\_off}\ to\ t = t_i \end{cases} \quad (7)$$

$if\ input\ pulse\ intensity = "0"$,

$$F_{PS}(P,t) = A_1(P,t)e^{-\gamma_1 \cdot t} + A_2(P,t)e^{-\gamma_2 \cdot t} \quad (8)$$

, where n is the number of training cycles, n_c is the compression ratio, and the rest of the variables as described above.

2.2.2.3.3. ES mode

The device functions as an electrical-synaptic (ES) mode with a low programming voltage of 0.1 V to 0.5 V, demonstrating analog and non-volatile electrical resistive switching behaviours. The basic behaviour of the device in this mode results in long-term potentiation (LTP) and long-term depression (LTD) characteristics under application of constant positive (0.5V, 50 ms) and negative (-0.5V, 50 ms) voltage pulses, respectively, as illustrated in **Fig. 2.6**c. Multiple analog resistance states remain stable and reliably maintainable. **Fig. 2.13** further shows the cyclic LTP/LTD behaviours with excellent linearity.

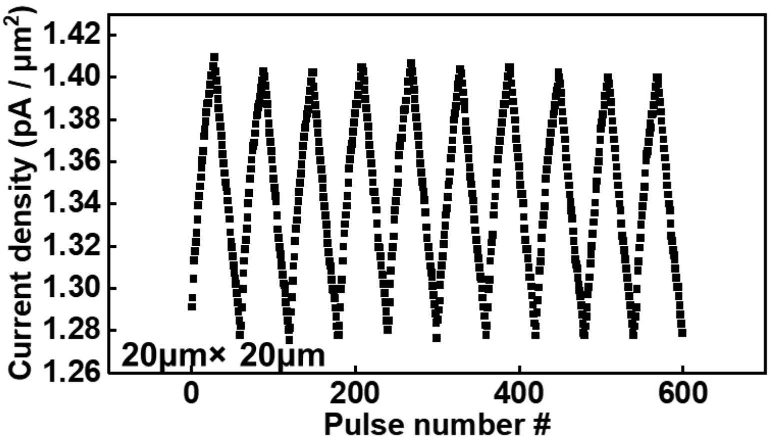


Fig. 2.13 Cyclic LTP and LTD of device in ES mode. The device is programmed by 30 positive pulses (+ 0.5 V/100 ms) and 30 negative pulses (- 0.5 V/100 ms) for 10 cycles with a read voltage of 0.01 V.

The retention of the device in the ES mode is measured over 10 hours without any observable degradation in the low resistance state and well extrapolated to 10 years, shown in **Fig. 2.14**a. Also, the retention of multiple resistance states over 3000 s is shown in **Fig. 2.14**b, demonstrating the non-volatile multilevel resistance states (8 clearly separated states) programmed by voltage pulse amplitude and duration of 0.5 V and 50 ms, respectively, with a read voltage of 10 mV. To assess the uniformity of the retention in ES mode across the device array, we randomly selected 10 devices from a 4-inch wafer with thirty 100×100 arrays and performed retention tests. As shown in **Fig. 2.14**c, the devices all exhibit consistent low resistance state retention over 3000 s, with small variation across the large-area array.

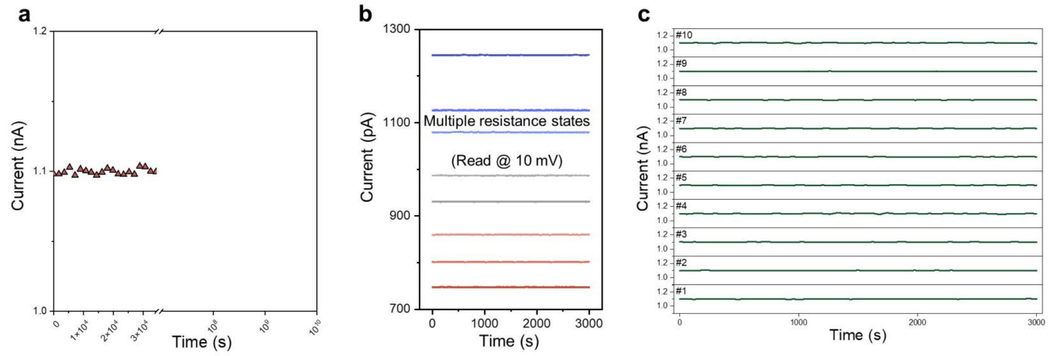


Fig. 2.14 Retention behaviour of ES mode device. (a) Retention of ES mode device with read voltage of 10 mV over 10 hours with device area of 25 μm × 25 μm. (b) Retention of non-volatile multilevel resistance states (8 clearly separated states) programmed by voltage pulse amplitude and duration of 0.5 V and 50 ms, respectively. (c) Retention of 10 randomly selected devices with device area of 25 μm × 25 μm in ES mode with a read voltage of 10 mV.

The endurance of ES mode device for 10000 cycles is tested, validating the basic endurance reliability (**Fig. 2.15**).

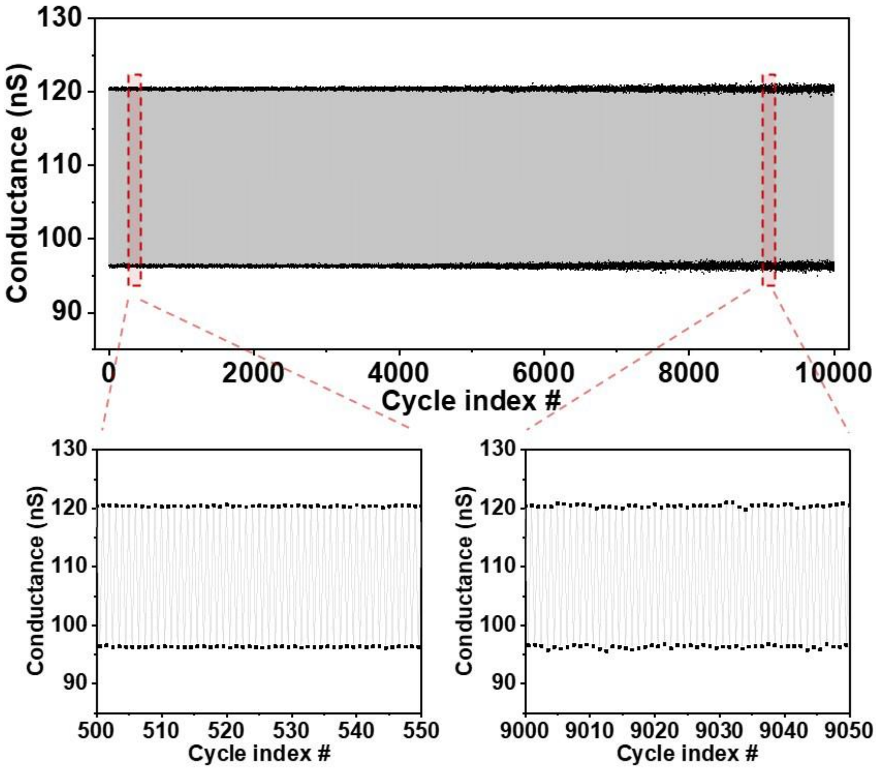


Fig. 2.15 Endurance of ES mode device for 10000 cycles. The device in the ES mode is programmed by the voltage sweeping of 0 - ±0.5 V and read at a voltage of 10 mV. The device area is 25 μm × 25 μm.

The basic behaviour for one electrical pulse period is described by the following functions:

$Basic\ ES\ mode\ behavior\ for\ one\ electrical\ pulse\ period\ (0-t)$:

$$F_{ES} = \begin{cases} G_{ES-EP} = f_{ES-EP}(V,t) = G_{initial} + s(V)\cdot t_{pulse}, & V \geq 0 \\ G_{ES-ED} = f_{ES-ED}(V,t) = G_{initial} - s(V)\cdot t_{pulse}, & V < 0 \end{cases} \tag{9}$$

, where G is the conductance of the device, EP is the potentiation process, ED is the depression process, $G_{initial}$ is the initial conductance of the device, t_{pulse} is the time duration when the electrical pulse is on, and s is the scaling factor for the increase and decrease in conductance.

Building upon the basic behaviour of the device, multiplication operation algorithm can be implemented. Functions that describe the multiplication algorithms are represented by the following functions (indicated in **Table 2.2**):

$Device\ algorithm\ for\ multiplication\ (0 - T)$:

$$G_{ES-EP}(V_{program}, n) = f_{ES-EP}(V_{program}, n) = G_{initial} + \sum_{i=1}^{n} s(V_{program}) \cdot t_{pulse} \quad (10)$$

$$G_{ES-ED}(V_{program}, n) = f_{ES-ED}(V_{program}, n) = G_{initial} - \sum_{i=1}^{n} s(V_{program}) \cdot t_{pulse} \quad (11)$$

$$f_{multiply}(V_{input},\ i = n) = V_{input} \cdot G_{ES-EP/ED}(V_{i=n-1}, t_{i=n-1}) \quad (12)$$

, where $V_{program}$ is the programming voltage, n is the number of pulses applied to the device, and the rest of the variables as described above. The device-level multiplication algorithm ultimately supports multiply-and-accumulate (MAC) operations in cross-point cell array configurations for practical applications.

2.2.2.3.4. EN mode

The device operates in electrical-neuron (EN) mode at a larger voltage range of >1.3 V, exhibiting LIF behaviour under continuous electrical pulse stimuli. **Fig. 2.6**d presents the current responses to electrical pulses with varying amplitudes and numbers: 1.3 V (10 pulses), 1.4 V (7 pulses), 1.5 V (4 pulses), 1.7 V (2 pulses) and 2.0 V (1 pulse), using 0.5 ms pulse width/interval. The firing occurs only when the pulse amplitude exceeds the threshold levels ($V \geq 2.0$ V). **Fig. 2.16** reveals frequency-dependent firing characteristics under constant amplitude, successful firing requires a minimum 2.0 ms pulse width. These results confirm that firing only initiates when input stimulus strength exceeds critical thresholds. The pulse amplitude-, pulse frequency-, and pulse number-dependent LIF can be used for mimicking neurons for selecting above-threshold signals and facilitating the decision-making process in spiking-based neural networks.

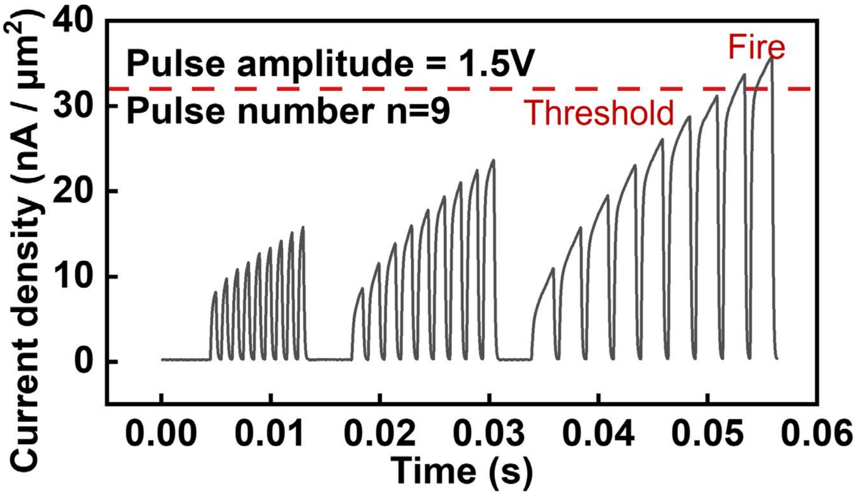


Fig. 2.16 EN mode device frequency-dependent firing characteristic. The device's current response is shown for varying electrical pulse widths and intervals (500 µs/500 µs, 1000 µs/500 µs, and 2000 µs/500 µs) with a pulse number of 9.

The basic EN mode behaviour for one pulse duration can be described by the following functions:

$Basic\ EN\ mode\ behavior\ for\ i^{th}\ electrical\ pulse, for\ one\ \ pulse\ duration\ (0 - t_i)$:

$$F_{EN}(V_i, t_i) = \begin{cases} I_{EN-EP} = f_{EN-EP}(V_i, t_i) = \beta(V, i) \cdot \left(-e^{-\alpha(V,i)\cdot t} + 1\right), & t = 0\ (pulse\_on)\ to\ t = t_{pulse\_off} \\ I_{EN-ED} = f_{EN-ED}(V_i, t_i) = I_{EN-EP}(V_i, t_i) \cdot e^{-\gamma t_{pulse\_off}}, & t = t_{pulse\_off}\ to\ t_i \end{cases} \quad (13)$$

$$\gamma \to \infty$$

, where *EP* is the potentiation process, *ED* is the depression process, α is the rise factor, γ is the decay factor, t_{pulse_on} is the duration where the electrical pulse in on, t_{pulse_off} is the duration where the electrical pulse is off, β is the amplitude of the device current at i^{th} step.

Building upon the basic behaviour of the device, a comparing function model is as described. The model takes into account the pulse amplitude-based firing conditions.

$Device\ algorithm\ for\ comparing\ function\ (0 - T)$:

$$f_{comparing}(V) = \begin{cases} 1, & V \geq V_{th} \ and \ F_{EN}(V,t) \geq I_{th} \\ 0, & otherwise \end{cases} \quad (14)$$

, where V_{th} is the threshold voltage and I_{th} the required threshold current for firing.

2.2.2.4. Energy consumption of sensory computing modes of device

All sensory computing modes are operated with low energy consumption, as shown in **Table 2.3**. enabling the highly efficient intelligent vision system.

Table 2.3 Energy and power consumption of the device

Mode	Metric	Calculation
PN mode	Device power consumption	P = 0 W (zero voltage bias)
	Device energy consumption	E = 0 J (zero voltage bias)
PS mode	Device power consumption	$P = E / t$ = 152.000 fJ / 100.000 ms = 1.520 pW
	Device energy consumption	$E \approx \lvert V \rvert \times (1/2 \times (I_{Dark}+I_{Photo}) \times t)$ = 0.04 V × (0.5 × (23.000 + 53.000) pA) × 100.000 ms = 152.000 fJ
ES mode	Device power consumption	$P = V \times I$ = 0.01 × 20.000 pA = 200.000 fW
	Device energy consumption	$E = V \times I \times t$ = 0.01 × 20.000 pA × 20.000 ms = 4.000 fJ
EN mode	Device power consumption	$P = E / t$ = 157.500 pJ / 500.000 µs = 315.000 nW
	Device energy consumption	$E \approx V \times (1/2 \times (I_{Base}+I_{Firing}) \times t)$ = 1.5 V × (0.5 × (20.000 + 400.000) nA) × 500.000 µs) = 157.500 pJ

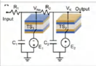
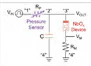
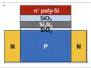
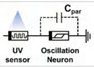
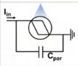
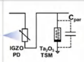
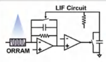
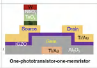
2.2.2.5. Benchmarking with other neuromorphic devices

The benchmarking results for energy efficiency and performance across various neuromorphic devices are presented in **Table 2.4**[28,30,32,35] and **Table 2.5**[30,36-42]. The comparison includes both the comparison of our multi-paradigm device with state-of-the-art reconfigurable devices (**Table 2.4**) and the comparison of our device in PN mode with other electrical and photoresponsive spiking neurons (**Table 2.5**).

Table 2.4 Comparisons of the state-of-the-art emerging reconfigurable neuromorphic devices with our multi-paradigm device

		[35]	[28]	[30]	[32]	**This work**
Configuration		$MoTe_2$/P(VDF -TrFE) transistor	V/VO_x/$HfWO_x$/ Pt + capacitor	Phototransistor + Ta/TaO_x/NbO_x /W	TiN/ZnO/ TiO_x/Pd	Pd/CuO_x/ITO
		1T	1R+1C	1T + 1R	1R	1R (single device)
Reconfigurability (dynamics)		PS+ES mode	Only ES and EN modes	Only PS and PN modes	Only PS and ES modes	**PN+PS+ES+ EN modes**
Energy per operation	**PN mode**	Not support	Not support	-	Not support	**0 pJ**
	PS mode	0 pJ	Not support	-	2 nJ	**152 fJ**
	ES mode	-	60 fJ	Not support	2 nJ	**4 fJ**
	EN mode	Not support	1 nJ	Not support	Not support	**157. 50 pJ**
CMOS-Back End of Line (BEOL) compatible		Yes	Yes	Yes	Yes	**Yes**
Photoresponse		340 - 1310 nm	No	450-800 nm	405-650 nm	**375 nm – 1064 nm**

Table 2.5 Comparisons of the electrical and photoresponsive spiking neurons with our multi-paradigm device in PN mode[30,36-42]

	[36]	[37]	[38]	[39]	[40]	[41]	[42]	[30]	This work
Neuron type	Electrical neuron			Photoresponsive neuron					
Structure	Mott memristor + Resistor + Capacitor	Mott memristor + Resistor + Capacitor	Silicon-oxide-nitride-oxide-silicon based MOSFET	UV sensor + Memristor + Capacitor	InGaAs BJT + Capacitor	Photoristor + Memristor + Capacitor	Optoelectronic RRAM + LIF circuit	Phototransistor + Memristor +capacitor	**Multi-paradigm device in PN mode**
Integration level									**Single device (with extra three modes)**
Photoresponse	No			UV light (254 nm ~ 365 nm)	Visible light (520 nm) ~IR light (1550 nm)	UV light (360 nm) ~Visible light (532 nm)	Visible light (445 nm ~ 623 nm)	Visible light (450 - 800 nm)	**UV light (375 nm) ~ IR light (1064 nm)**
Reconfigurability	No		Yes	No			Yes		**Yes**
Capacitors needed	Yes		No	Yes					**No**
Operation voltage	1.45 V	1.34 V	3 V	1.5 V	1.22-1.34 V	0.2 V	40 V	1.5 V	**0 V**
Read voltage	1.45 V	1.34 V	3 V	3 V	1.22-1.34 V	0.2 V	0.1 V	1.5 V	**0 V**
Energy per spike	-	-	0.7 pJ	-	270 pJ	30 pJ	-	-	**0 J**
Pixel core area	Memristor (5×5 μm²) + Capacitor	Memristor (5×5 μm²) + Capacitor	$L_G \times W_{CH}$ (880 × 280 nm)	-	Biristor (100 ×100 μm²) + Capacitor	70×70 μm²	ORRAM (~20×9 μm²) + LIF circuit	Phototransistor (10×20 μm²) + Memristor (5×5 μm²)	**5 × 5 μm²**
Preparation temperature	Room Temperature	120 ℃	1000 ℃	Room Temperature	350 ℃	Room Temperature	-	Room Temperature	**Room Temperature**

2.2.2.6. Device uniformity and stability

The long-term stability tests for four months are conducted to verify the time stability of the device over time. Specifically, we characterized the key functionalities (PN, PS, ES, and EN modes) of the same device when as prepared, after two months, and after four months, as shown in **Fig. 2.17**. The photo-spiking in PN mode, photocurrent response in PS mode, LTP and LTD in ES mode, as well as current response in EN mode, all exhibit no obvious degradation in device performance, demonstrating good time stability across all operational modes.

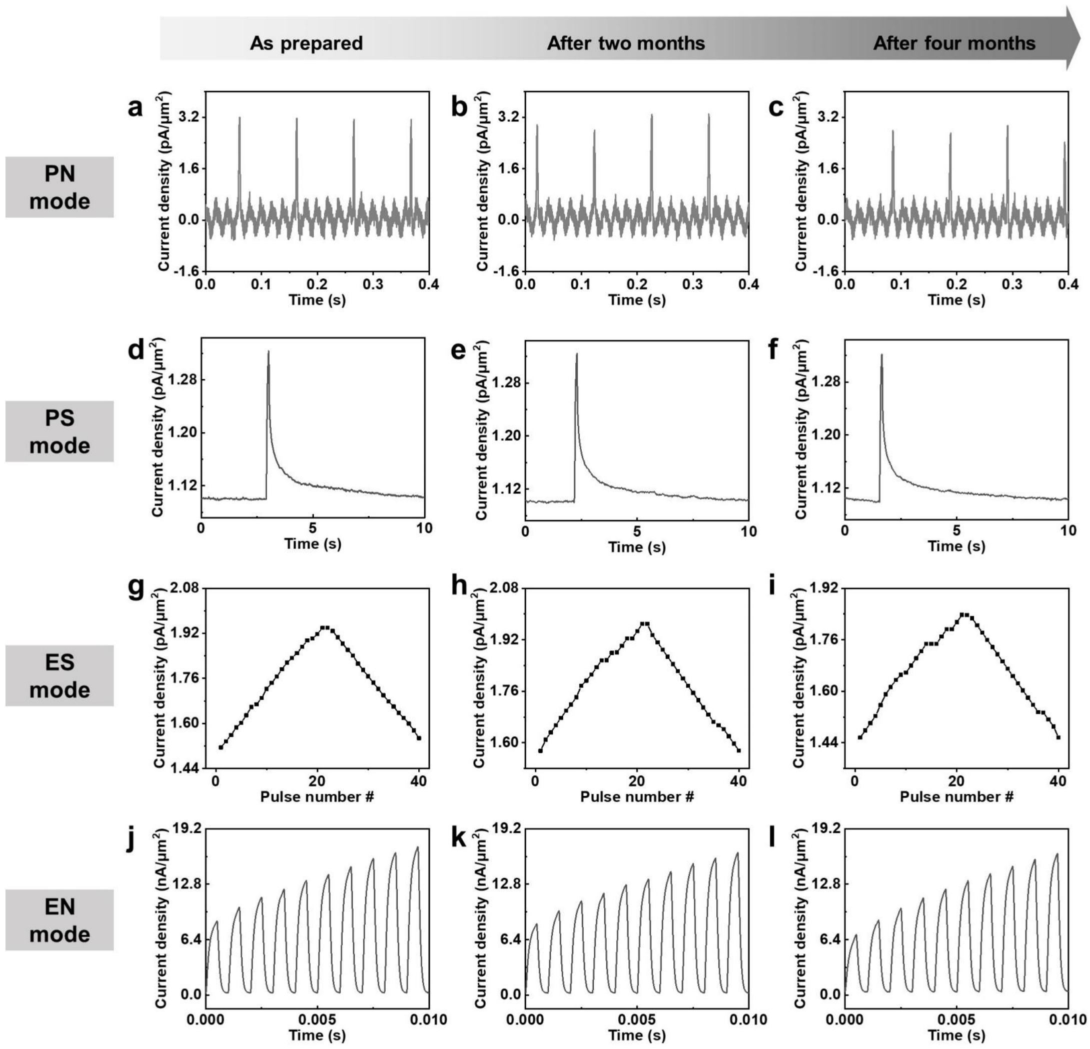


Fig. 2.17 Time stability of the multiparadigm device at different operating modes over four months. (a-c) PN mode behaviours under 638 nm, 20 mW/cm^2 light illumination, (d-f) PS mode triggered by 100 light pulses (638 nm, 120 mW/cm^2, pulse width of 1 ms and interval of 1 ms) and a -40 mV read voltage, (g-i) LTP and LTD behaviours of ES mode, programmed by ±0.5 V voltage pulses (pulse width of 50 ms), and (j-l) EN mode behaviours triggered by 10 voltage pulses (1.5 V, pulse width of 0.5 ms and interval of 0.5 ms).

To further examine the device-to-device variability over large areas, we fabricated the multi-paradigm devices on a 4-inch wafer, as shown **Fig. 2.18**a. 90 devices were randomly selected at different spots (indicated in white boxes) for device-to-device variation tests. The device-to-device variations operating in the respective four modes are provided as shown in **Fig. 2.18**b-e, showing the small device-to-device variance. These results indicate the scalability of multi-paradigm device array.

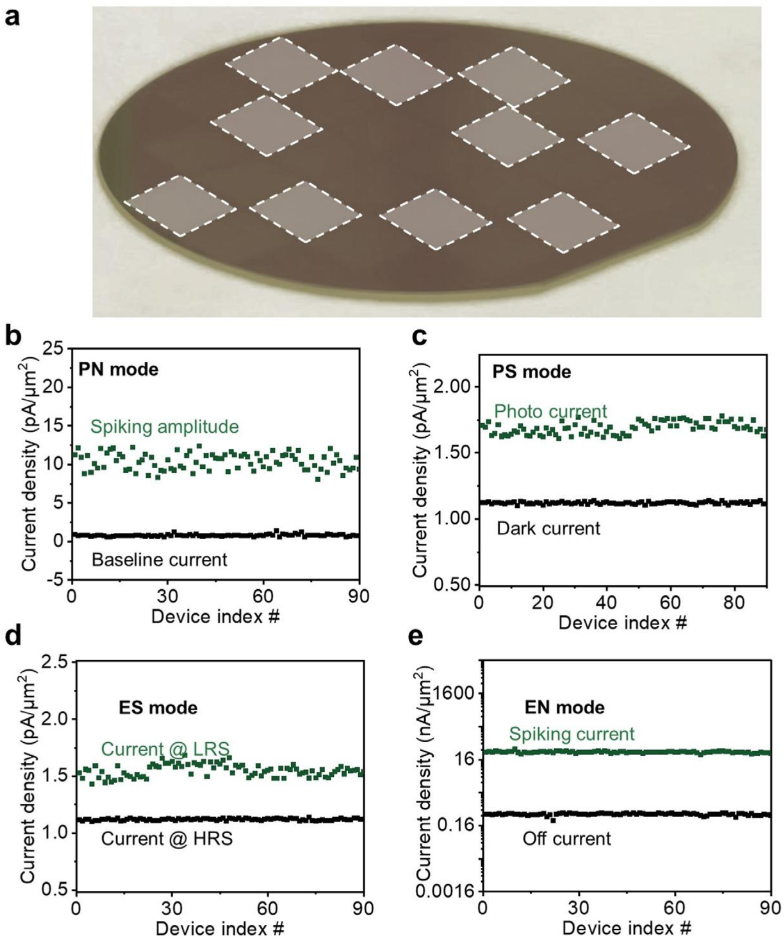


Fig. 2.18 Device-to-device variability over large areas. (a) Photo of fabricated multi-paradigm devices on a 4-inch wafer, with a total of 90 randomly selected devices with size of 25 μm×25 μm at different spots (indicated in white boxes) for device-to-device variation tests. (b) Device-to-device variation operating in PN mode under 375 nm light. (c) Distribution of Photocurrent and dark current of devices in PS mode, triggered by 450 nm light pulse with pulse width of 50 ms), with a read voltage of -40 mV. (d) Distribution of HRS and LRS when the device operates in ES mode. The HRS and LRS are programmed by 0 - ±0.5 V voltage

sweeping, respectively. (e) Distribution of spiking currents and OFF currents of the devices operating in EN mode. The spiking currents are triggered by 10 voltage pulses of 1.5 V with a pulse width of 0.5 ms and an interval of 0.5 ms.

2.2.3. Reconfigurable mode-control circuit and cell array

To achieve the full reconfigurability from the device to the system level, reconfigurable control circuits that can fully support device mode switching, cell array-architecture reconfigurability, and system-level reconfigurability are designed and demonstrated.

To emphasize the feasibility of our methodology at a hardware level, the system is validated in two stages. Firstly, we demonstrate a board-level hardware prototype (**Fig. 2.3**), by integrating a 12 × 12 multi-paradigm device array and designed reconfigurable mode-control circuits, verifying the reconfigurability across the cell to the array level. Secondly, a large-scale system with system-level reconfigurability is designed for complete hardware system verification, using a commercial 180 nm CMOS process design kit (PDK) and verified through electronic design automation (EDA) tools simulation (presented in the **Section 2.2.4.3**).

Fig. 2.19 schematically illustrates a reconfigurable cell, integrating a multi-paradigm device with a reconfigurable mode-control circuit consisting of an NMOS transistor and a low-power leakage-based amplifier. This circuit provides tuneable device biases by operating as a voltage clamp via negative feedback loops, maintaining mode-specific biases: 0 V (PN), -0.1 to -0.04 V (PS), 0.1 to 0.5 V(ES), and >1.3V (EN). During voltage clamping, simultaneous device readout is achieved with the mode-control circuit functioning as a transimpedance amplifier (TIA) for current-to-voltage conversion.

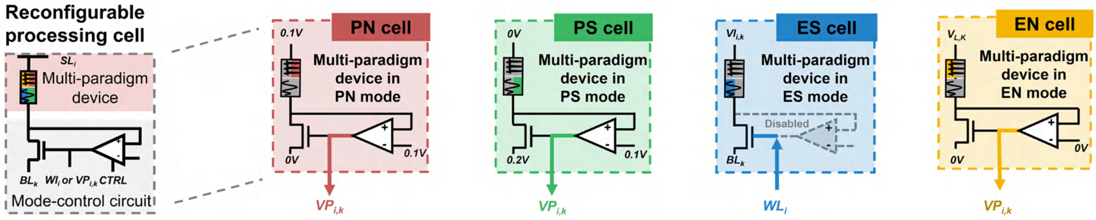


Fig. 2.19 Reconfigurable cell including a multi-paradigm device and mode-control circuit.

Additionally, the circuit supports different cell array architecture reconfigurations including 1R cell array (PN/PS/EN modes) (**Fig. 2.20**a) and 1T1R cross-point cell array (ES mode) configurations (**Fig. 2.20**b). When configured to 1T1R configuration, the amplifier is disabled to form a 1T1R cell unit for cell in ES mode, and the entire array can form a 1T1R structure for MAC operations. This dedicated circuit design maintains a unified architecture, which not only enables individual control of all four sensory modes per device but also supports flexible array configurations and thus the further system-level visual algorithms. The board-level hardware prototype in **Fig. 2.3** is used for functional verification for this reconfigurable cell array.

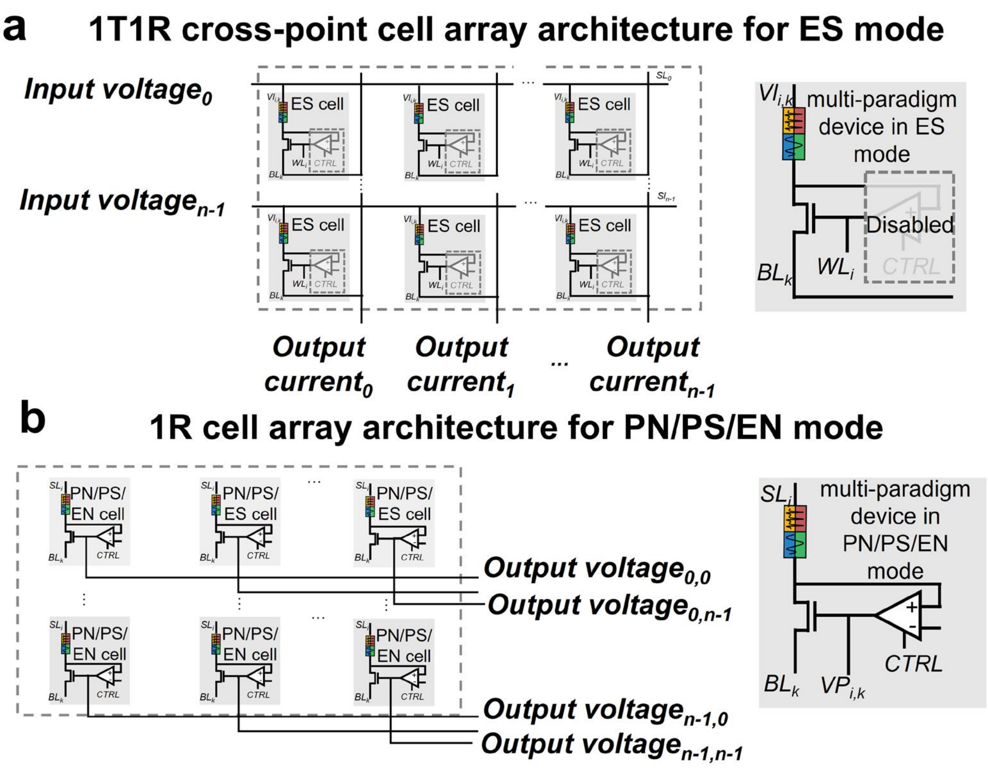


Fig. 2.20 Array reconfigurability: (a) 1R cell array architecture for PN/PS/EN modes and (b) 1T1R cross-point cell array architecture for ES mode.

The cell design's full access to four sensory modes and array architecture reconfigurability is experimentally validated by presenting outputs from both single reconfigurable cells and a 12 × 12 cell array across operational modes, completely readout by the prototype platform. In PN mode (**Fig. 2.21**a), continuous 520 nm and 450 nm illumination at 300 mW/cm^2 generates optical LIF spiking. The 12 × 12 cell array's spiking output under 520 nm illumination further verifies 1R cell array functionality with good uniformity (**Fig. 2.21**b). **Fig. 2.21**c shows the behaviour of the cell in PS mode, demonstrating pulse intensity-dependent nonlinear synaptic plasticity and encoding capability. The 12 × 12 cell array in PS mode (**Fig. 2.21**d) responds uniformly to the "001" pattern, generating 144 matching voltage amplitudes under a

1R architecture. ES mode cell operation exhibits non-volatile resistive switching, with three distinct resistance states programmed via positive and negative 0.5 V pulses (**Fig. 2.21**e). The multiplication results at high-resistance state (HRS) and low-resistance state (LRS) when a 0.015 V voltage pulse is applied are also shown in **Fig. 2.21**e. **Fig. 2.21**f demonstrates the individual cell readouts at HRS and LRS, and MAC output in a 12 × 12 ES mode cell array. **Fig. 2.21**g shows amplitude-dependent LIF behaviour in EN mode cell under electrical pulses (1.3-2.0 V), with firing activation exclusively at 2.0 V. Corresponding array outputs (**Fig. 2.21**h) show effective firing across 144 device cells when stimulated at 2.0 V, exceeding the firing threshold uniformly. The detailed metrics of each mode of a multi-paradigm device and a cell are presented in **Table 2.6**. And the calculation details are included in **Table 2.7**.

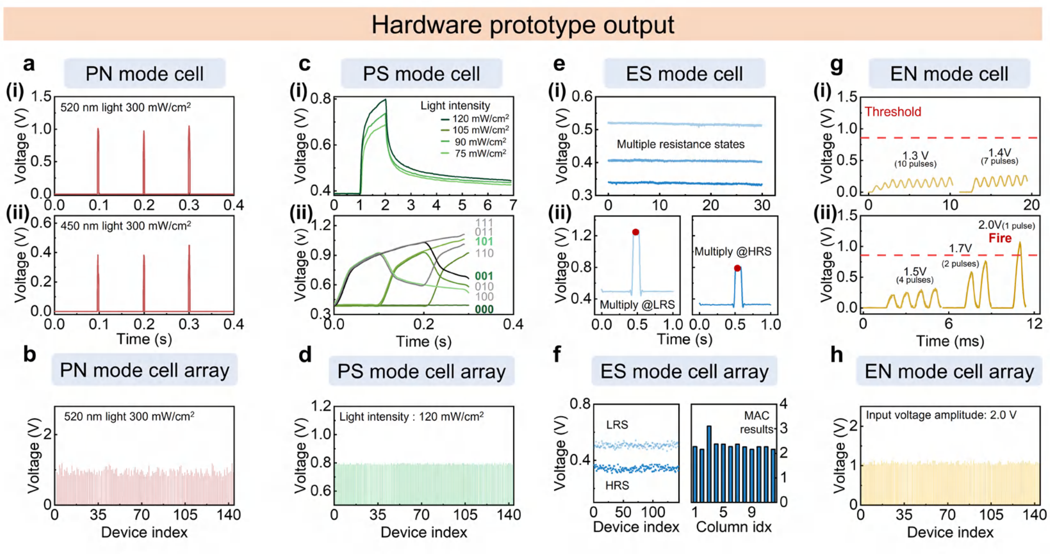


Fig. 2.21 Device cell and array output. (a) PN mode cell output under i. 520 nm and ii. 450 nm illumination with light intensity of 300 mW/cm^2. (b) Spiking results of the 12 × 12 cell array reconfigured to PN mode under 520 nm illumination with light intensity of 300 mW/cm^2. (c) PS mode cell output under i. single 520 nm, 1s light pulse with varying illumination

intensities ranging from 75-120 mW/cm^2, and ii. eight different input optical temporal patterns ("000" - "111"). The "0" represent optical input with 1s pulse width, 520 nm wavelength, and intensity of 0 mW/cm^2, while the "1" represents optical input with 1s pulse width, 520 nm wavelength, and intensity of 0 mW/cm^2. The read voltage is -40 mV. (d) Photoresponse output of the cell array reconfigured to PS mode under 520 nm, 1s light pulse with intensity of 120 mW/cm^2. (e) ES mode cell output showing i. different resistance states under a constant read voltage of 0.01 V and ii. Multiplication results at HRS and LRS when a 0.015 V voltage pulse is applied. (f) Readout results at HRS and LRS, and MAC results of the 12 × 12 ES mode cell array. (g) EN mode cell response under varying pulse amplitudes and numbers i. 1.3 V (10 pulses), 1.4 V (7 pulses), and ii. 1.5 V (4 pulses), 1.7 V (2 pulses), 2.0 V (1 pulse), with pulse width and interval of 0.5 ms. (h) Output firing results of the 12×12 EN mode cell array under a 2 V, 0.5 ms electrical pulse.

Table 2.6 Detailed metrics of each mode of a multi-paradigm device and a cell in the vision system

		PN mode	PS mode	ES mode	EN mode
Per device	Programming voltage	0 V	-0.04 V	0.1V	>1.3 V
	Read voltage	0 V	-0.04 V	10 mV	>1.3 V
	Energy consumption	0 J	152 fJ	4 fJ	157.5 pJ
	Power consumption	0 W	1.52 pW	0.2 pW	315 nW
Per cell*	Power consumption*	1.1 pW	2.62 pW	0.2 pW	315 nW

* The leakage-based amplifier in mode-control circuit consumes only 1.1 pW for the targeted 200 Hz bandwidth. Obtained from simulation using Cadence Virtuoso with a commercial

180 nm CMOS PDK.

Table 2.7 Calculation details of energy and power of the device and cell

Mode	Metric	Calculation
PN mode	Device power consumption	P = 0 W (zero voltage bias)
	Device energy consumption	E = 0 J (zero voltage bias)
	Circuit power consumption	P = 1.1 pW
	Cell power consumption	0 W + 1.1 pW = 1.1 pW
PS mode	Device power consumption	$P = E / t$ = 152.000 fJ / 100.000 ms = 1.520 pW
	Device energy consumption	$E \approx \lvert V \rvert \times (1/2 \times (I_{Dark}+I_{Photo}) \times t)$ = 0.04 V × (0.5 × (23.000 + 53.000) pA) × 100.000 ms = 152.000 fJ
	Circuit power consumption	P = 1.1 pW
	Cell power consumption	1.520 pW + 1.1pW = 2.620 pW
ES mode	Device power consumption	$P = V \times I$ = 0.01 × 20.000 pA = 200.000 fW
	Device energy consumption	$E = V \times I \times t$ = 0.01 × 20.000 pA × 20.000 ms = 4.000 fJ
	Circuit power consumption	P = 0 pW
	Cell power consumption	200.000 fW
EN mode	Device power consumption	$P = E / t$ = 157.500 pJ / 500.000 μs = 315.000 nW
	Device energy consumption	$E \approx V \times (1/2 \times (I_{Base}+I_{Firing}) \times t)$ = 1.5 V × (0.5 × (20.000 + 400.000) nA) × 500.000 μs) = 157.500 pJ
	Circuit power consumption	P = 0 pW
	Cell power consumption	315.000 nW

Further details on the cycle-to-cycle and cell-to-cell performance variations are provided in **Fig. 2.22** and **Fig. 2.23**, respectively. This verified device-to-array reconfigurability enables deployment of diverse system-level NI/AI algorithms combining spiking and non-spiking paradigms on the prototype platform.

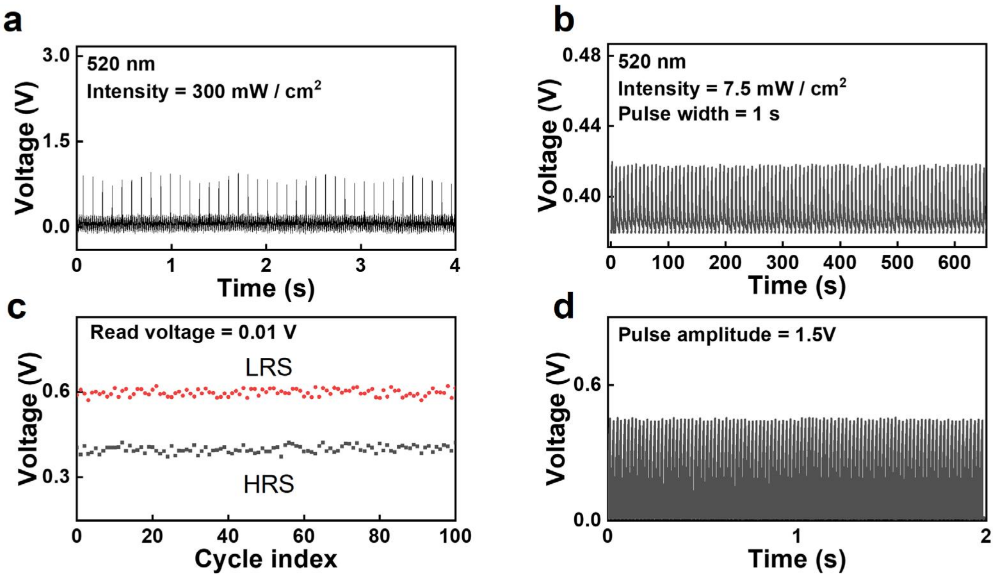


Fig. 2.22 Cycle-to-cycle performance of the multi-paradigm cell. The devices in the array are read by the board-level dynamic control circuit in our prototype, outputting voltage signals. (a) Cyclic optical LIF spiking of PN mode under 520 nm illumination (300 mW/cm^2). (b) Cyclic photoresponse of PS mode cell triggered by 450 nm light pulses (7.5 mW/cm^2, pulse width of 1 s), with a read voltage of -40 mV. (c) Cyclic HRS and LRS behaviour of ES mode cell, programmed by 0 to ±0.5 V voltage sweeps and read at 0.01 V. (d) Cyclic voltage response of EN mode cell under 10 voltage pulses (1.5 V, 0.5 ms width and interval) for each cycle.

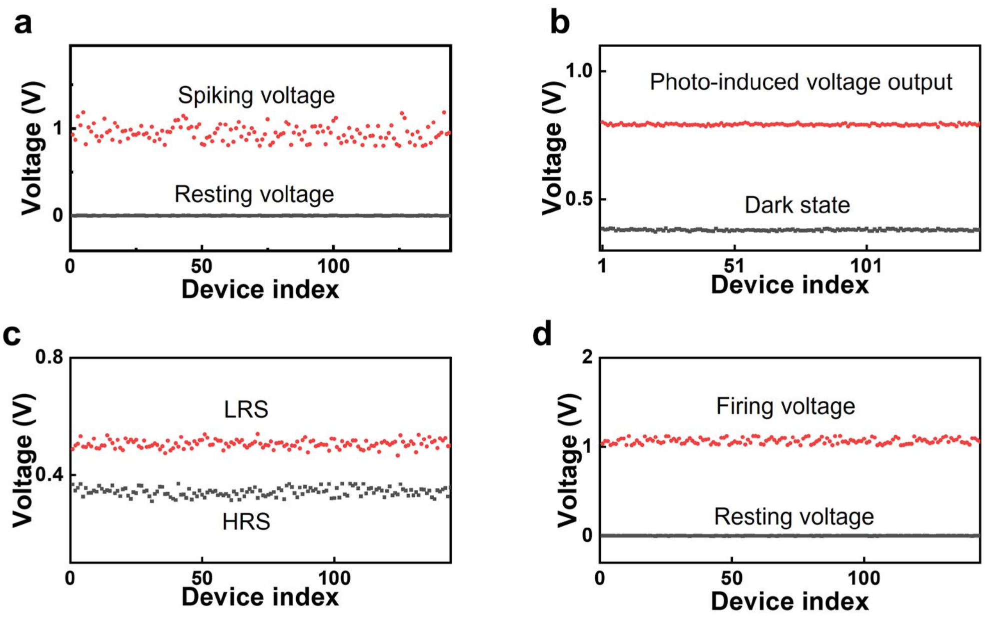


Fig. 2.23 Device-to-device variation of four modes in a 12 × 12 multi-paradigm device array measured by our reconfigurable hardware prototype. (a) Spiking voltage distribution of cell operating in PN mode under 520 nm, 300 mW/cm^2 illumination. (b) Distribution of output voltages in PS mode under a 520 nm light pulse with intensity of 7.5 mW/cm^2, and pulse width of 1 s for each cycle. (c) Distribution of readout voltages at HRS and LRS states in ES mode, programmed by 0 to ± 0.5 V voltage sweeps. (d) Distribution of firing voltages in EN mode, triggered by a 2.0 V, 0.5 ms pulse.

2.2.4. On-demand Reconfigurable Intelligent Vision System

Exploiting the behaviours in the reconfigurable cell array prototype, an on-demand reconfigurable intelligent vision system is designed and demonstrated. Firstly, system-level visual algorithms are designed and simulated based on the experimentally characterized behaviour of the hardware prototype, enabling the system to adapt dynamically across various vision scenarios. **Section 2.2.4.1** shows the NI-centric architecture configurations and co-designed algorithms towards high-resolution smart imaging scenarios. In these types of configurations, the cell array with in-sensor computing capabilities can perform high-resolution imaging and preprocessing in real time. **Section 2.2.4.2** demonstrates the AI-centric configurations and co-designed algorithms, including spiking and non-spiking visual-neural networks (visual-spiking recurrent neural network, visual-spiking artificial neural network, visual-artificial neural network, visual-reservoir computing), for accurate recognition scenarios. The AI-centric configuration is designed with an NI/AI hierarchical architecture, combining the traits of NI implemented with in-sensor pre-processing at the front end, and AI implemented with in-memory processing at the back end. These configurations and algorithms are implemented based on the hardware prototype platform (**Fig. 2.3**), supported by large-scale simulation results. Therefore, the visual neural networks are sized to match what we can currently measure and validated at this stage.

Then, the system-level architecture with reconfigurable signal paths and corresponding peripheral circuits that can fully support the system-level algorithms is further designed for full hardware implementation verification, as shown in **Section 2.2.4.3**.

2.2.4.1. NI-centric configurations and visual algorithms

NI-centric configurations, operating in in-sensor computing mode, can be categorized into spiking or non-spiking for event-based and frame-based scenarios. In the spiking NI configuration (**Fig. 2.24**a), all cells operate in PN mode, enabling executing event-based colour-mixed motion extraction and filtering with high resolution. Utilizing the device algorithm describing the wavelength-dependent LIF spiking characteristics with intensity and integration time-based thresholds, the event-based colour-mixed motion extraction and filtering is illustrated (**Fig. 2.24**b). The PN mode cell array can intelligently capture, extract and encode the continuous motion information (L, from 0 to T period) of the colored targeted vehicles (red and blue cars), outputting high-resolution voltage spike maps (U, from 0 to T period) with continuous trace and detailed object contour while filtering out the noisy vehicles (green motorcycle) and the static background. These motion capturing, extraction and encoding processes are defined by the algorithm as:

$$\begin{aligned} U &= (U_1 \cdots U_{13456})_{t_0-t_{24}} \\ &= F_{motion-extraction}\left((L_1 \cdots L_{13456})_{t_0-t_{24}}\right) \\ &= \left(\{f_{spike\ encoding\ \&\ filtering}(P_m,\ \lambda_m)\}_{m=1}^{13456}\right)_{t_0-t_{24}} \end{aligned} \tag{15}$$

, with each device complied with the device-level spike encoding and filtering algorithm $f_{spike\ encoding\ \&\ filtering} = U_{fire}(P,t) \cdot \sum_{i=1}^{n} F_{PN}(P,\lambda,I,t)$ (equation 2).

Owing to the light-intensity-threshold and integration-time-threshold-based firing condition $U_{fire}(P,t) = \begin{cases} 1, & if\ P > P_{th}\ and\ t > t_{th} \\ 0, & Otherwise \end{cases}$ in the PN mode device (equation 3), the targeted moving vehicles only with both above-threshold input intensity ($P_1, P_2 > P_{th}$) and integration time ($t_1, t_2 > t_{th}$) can be extracted by the device, generating spikes. Conversely, static dark background with the below-threshold input intensity ($P_4 < P_{th} | t_4 > t_{th}$) and noisy motorcycle with a shorter integration time ($t_3 < t_{th} | P_3 > P_{th}$) result in no spikes and can be effectively filtered out. The motion information of detected targeted moving cars with different colors is consequently extracted and encoded into spikes with wavelength-dependent amplitudes based on the

PN mode device algorithm $f_{spike\ encoding\ \&\ filtering}$ (equation 2), where the spike amplitude is a function of the input wavelength λ. The resulting spike maps over time are illustrated in **Fig. 2.24**b. The system can accurately real-time extract and encode motion information, faithfully representing the spatial features and temporal dynamics of the target objects.

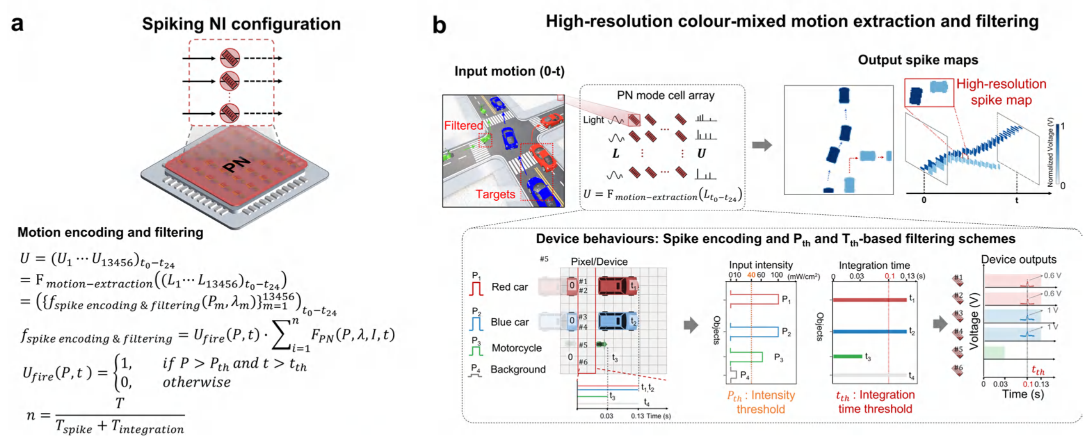


Fig. 2.24 Spiking NI configuration and visual algorithms for high-quality and high-resolution neuromorphic imaging with low- and mid-level processing capabilities. (a) Spiking NI configuration with all reconfigurable cells in the array configured into PN mode and the corresponding motion encoding and filtering algorithms. (b) Scenarios and output spike map results of event-based colour-mixed motion encoding and filtering.

In contrast, the non-spiking NI configuration operates in PS mode (**Fig. 2.25**a), enabling high-resolution frame-based imaging and preprocessing (**Fig. 2.25**b). This configuration enables two visual algorithms: image contrast enhancement ($F_{image-enhancement}$) and image compression ($F_{image-compression}$). For image enhancement (**Fig. 2.25**b(i)), the high-resolution traffic sign image L is processed by the cell array, resulting in enhanced body/background and body/noise contrasts. The image enhancement algorithm is represented as:

$$U = F_{image-enhancement}(L) = \{f_{optical-contrast-enhancement}(P_m, t = t_{sample})\}_{m=1}^{13456} \quad (16)$$

This algorithm represents the processing of each pixel P_m in the image L at the sampling time, t_{sample}. By deploying the device-level optical-contrast-enhancement algorithm: $f_{optical-contrast-enhancement}(P, t)$ (equation 5), the cell array behaves as generating voltage with varying decay rates at different light intensities. The PS mode device generates a slower-decaying current for the body pixels than that for the noise pixels. This difference in decay rates enhances the contrast between the voltage signals from the body pixels and noise pixels at the sampling time. Meanwhile, it generates a rapidly-decaying current for the background pixels, resulting in no current at the sampling time, therefore enhancing the contrast between the body and background pixels and improving the quality of the traffic sign image. **Fig. 2.25**b(ii) shows the vehicle image compression demonstration with a threefold data volume reduction enabled by the PS mode cell array. The image compression process is represented by the following algorithm as:

$$U = F_{image-compression}(L) = \{f_{data-compression}(P_m, n_c = 3)\}_{m=0}^{36963} \quad (17)$$

where each device complied with the device-level data-compression algorithm $f_{data-compression}(P, n_c) = \sum_{i=3m+1}^{n_c+3m+1} F_{PS}(P_i, t_i)$ (as detailed in equations 6, 7 and 8), and m and n_c denotes the device index and the compression ratio, respectively. Specifically, during the image compression implementation, each device receives and processes a 3-bit light information from three adjacent frames of the grayscale vehicle image stream, and compresses it into a 1-bit conductance state, denoted as

$f_{data-compression}(P_m, n_c = 3)$ for a device indexed by m. Therefore, the cell array can encode 8 different 3-bit optical streams ranging from '000' to '111' into 8 distinct 1-bit conductance states, thereby compressing the input images.

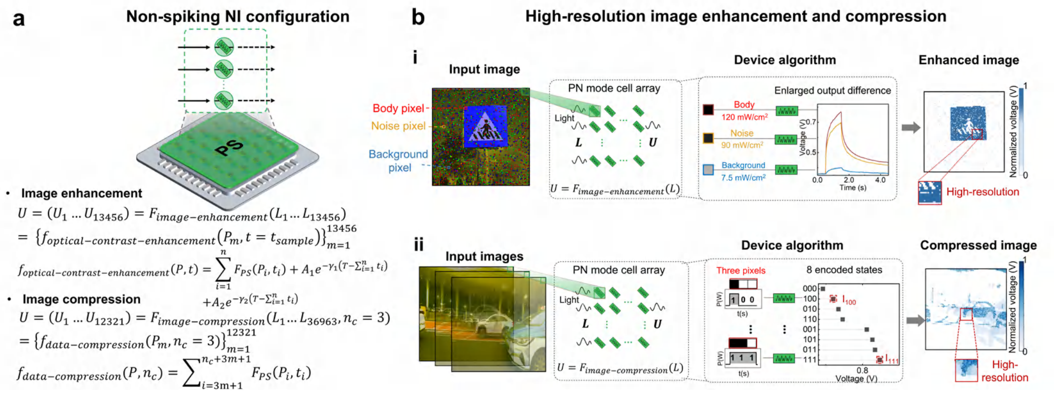


Fig. 2.25 Non-spiking NI configuration and visual algorithms for high-quality and high-resolution neuromorphic imaging with low-level processing capabilities. (a) Non-spiking NI configuration with all cells configured into PS mode and the corresponding image enhancement and compression algorithms. (b) i. Image enhancement results, and ii. image compression results.

The frame rate and latency metrics for each NI-centric configuration are provided in **Table 2.8**.

Table 2.8 Performance metrics of NI-based processing functionalities

Array configuration	NI-based processing	Frame Rate	Latency
Spiking mode	Motion extraction and filtering	10 fps	0.1 s
Non-spiking mode	Image enhancement	Up to 80 fps	~ 12.5 ms
	Image compression	80 fps	~ 12.5 ms

2.2.4.2. AI-centric configurations and visual algorithms

AI-centric configurations can be categorized into mixtures of AI- and NI-array of different proportions, combining in-sensor and in-memory computing. These configurations feature NI-based imaging with low-/mid-level processing, followed by AI-based high-level processing. Compared with NI-centric configuration, a smaller portion of device cells is allocated to PN/PS modes for NI with lower imaging resolution, while a larger portion is allocated to ES/EN modes for AI recognition. Visual neural network algorithms are designed for the AI-centric configurations. The NI proportion serving as the visual neural network's input sensory neurons utilizes the device's inherent behaviours to sense and extract the most essential features of the image input required for the subsequent AI processing, allowing dramatically reduced AI network size with enhanced efficiency and accuracy.

2.2.4.2.1. Visual-spiking artificial neural network configuration

Fig. 2.26a shows the spiking AI-centric visual-spiking artificial neural network configuration for event-based colour-mixed motion trajectory extraction, filtering, and recognition. The cell array is reconfigured into a PN mode cell array as the input sensory spiking neurons for the spiking NI, an ES mode cell array and an EN mode cell array as the synapses and output neurons for AI. More details about the network structure and hardware implementation can be found in **Fig. 2.27**, with example scenarios in **Fig. 2.28**.

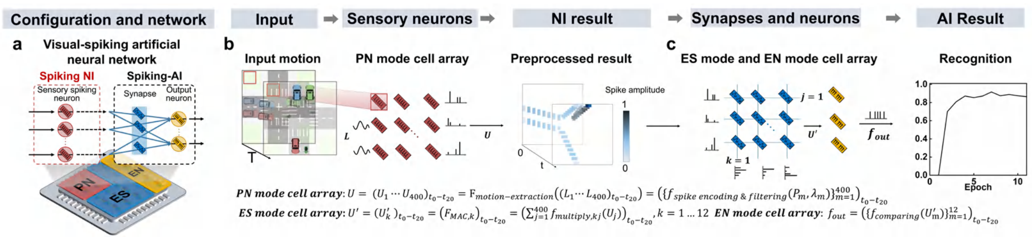


Fig. 2.26 AI-centric visual-spiking artificial neural network configuration. (a) In this configuration, the cell array is configured into a PN mode cell array as the input sensory spiking

neurons for the spiking NI, an ES mode cell array and an EN mode cell array as the synapses and output neurons for the spiking AI. (b) NI-based processing: illustration of the event-based colour-mixed motion trajectory extraction and filtering enabled by the PN mode cell array, outputting spike maps with filtered and extracted motion trajectories (0-*T*). (c) AI-based processing: recognition accuracy of the configuration on 12 classes of mixed-colour vehicle motion images for the target cars.

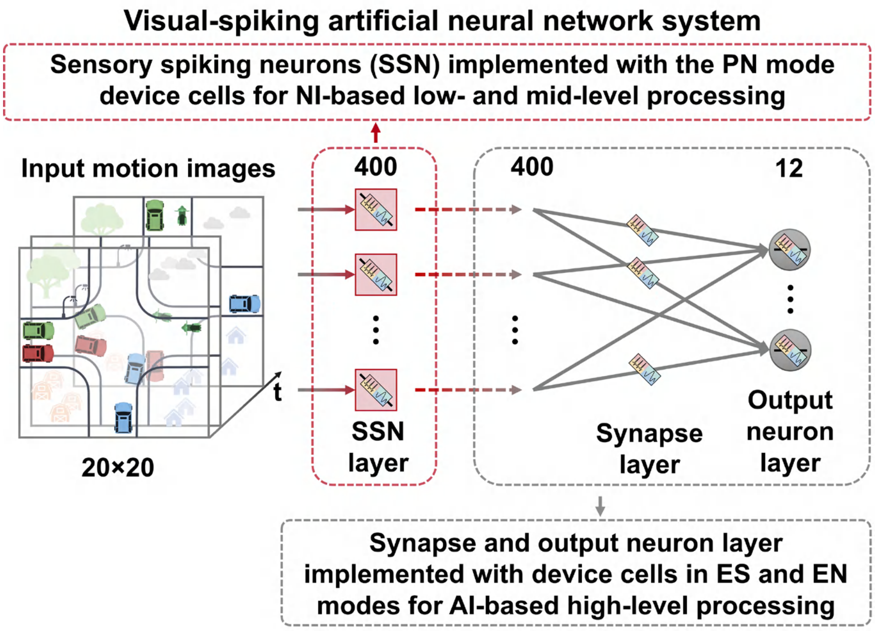


Fig. 2.27 Network structure and hardware implementation of the visual-spiking artificial neural network system. The visual-spiking artificial neural network has 400 input sensory spiking neurons corresponding to the 20 × 20 PN mode device cell array, and 12 output neurons realized by the EN mode device cell array. The synapses that connect the input and output layers are implemented by the ES mode cell array.

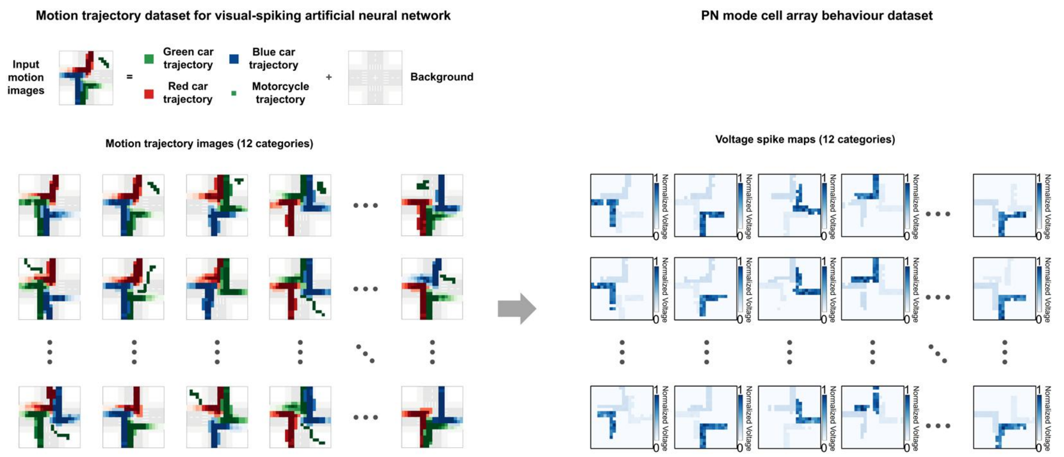


Fig. 2.28 Visual-spiking artificial neural network dataset. The motion trajectory dataset consists of 12 different categories of colour-mixed motion images, each comprising 20 consecutive motion steps. The size of each motion image is 20 × 20 pixels. Similarly, the PN mode cell array behaviour dataset consists of 12 different categories of spike map patterns, each with the same resolution of 20 × 20 pixels, containing 20 sequential spikes. 480 of these spike maps were used for training the AI section in this configuration, and 120 spike maps were used for testing the AI section of this configuration.

As shown in **Fig. 2.26**b, the PN mode cell array first extracts and encodes the continuous colour-mixed motion information $(L, 0-T)$ of target vehicles (cars) while filtering out the noisy motorcycle and static background based on the schemes depicted in **Fig. 2.24**b. The output is a voltage spike map $U = (U_1 \cdots U_{400})_{t_0-t_{20}}$ with extracted motion trajectories. This process is represented as:

$$U = (U_1 \cdots U_{400})_{t_0-t_{20}} = F_{motion-extraction}\big((L_1 \cdots L_{400})_{t_0-t_{20}}\big) \tag{18}$$

, where the motion extraction algorithm $F_{motion-extraction}$ is defined as:

$$F_{motion-extraction}\big((L_1 \cdots L_{400})_{t_0-t_{20}}\big) = \big(\{f_{spike\ encoding\ \&\ filtering}(P_m,\ \lambda_m)\}_{m=1}^{400}\big)_{t_0-t_{20}} \tag{19}$$

in which each PN mode device performs a device-level spike encoding and filtering algorithm, as detailed in equation 2.

These voltage spike maps from the PN mode cell array are subsequently input to the ES mode cell array for MAC operations (**Fig. 2.29**). The MAC algorithm is represented as:

$$U' = (U'_k)_{t_0-t_{20}} = \left(F_{MAC,k}\right)_{t_0-t_{20}} = \left(\sum_{j=1}^{400} f_{multiply,kj}\left(U_j\right)\right)_{t_0-\, t_{20}} \tag{20}$$

, where each ES mode device performs a device-level multiplication algorithm $f_{multiply}$, as detailed in equation 12. k is the column index, and j is the row index.

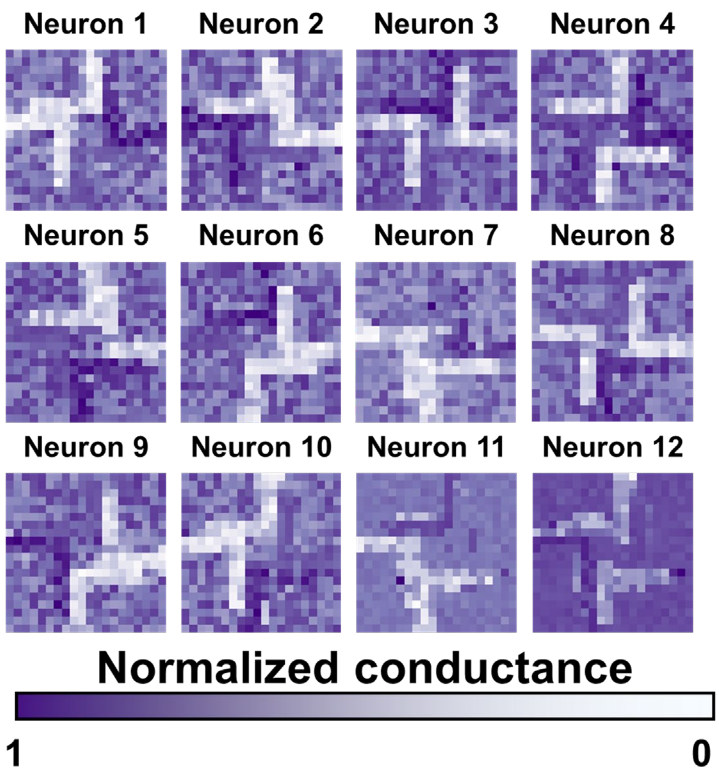


Fig. 2.29 Weight maps of visual-spiking artificial neural network synapse layer.

The output voltage spikes are then processed by the EN mode cell array for selecting the above-threshold signals and final decision-making (**Fig. 2.26**c). The final output voltage spike trains are denoted as

$$f_{out} = \left(\{f_{comparing}(U'_m)\}_{m=1}^{12}\right)_{t_0-t_{20}} \tag{21}$$

, where each device complied with the device-level algorithm $f_{comparing}$ (equation 14). The spike rates of the EN mode cell array $\frac{f_{out}}{T}$ indicate the recognition result (**Fig. 2.30**). The trained system achieves a recognition accuracy of 91.7% across 12 mixed-colour vehicle motion classes.

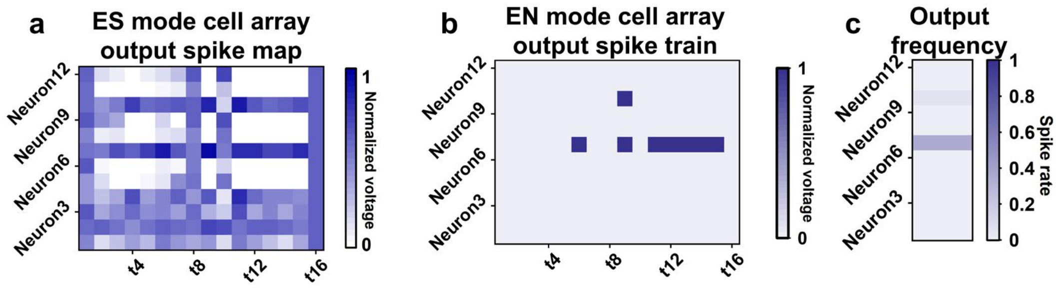


Fig. 2.30 Spike maps from ES and EN mode cell array. (a) The output voltage spike maps from the ES mode cell array and (b) the output spike trains from the EN mode cell array and the corresponding (c) output frequency, generated using an image from the test dataset labelled as class 5.

2.2.4.2.2. Visual-spiking recurrent neural network configuration

Fig. 2.31a shows the spiking AI-centric (visual-spiking recurrent neural network) configuration for event-based target vehicle (car) motion trajectory extraction and prediction. This configuration reconfigures the cell array into a PN mode cell array for spiking NI, and ES and EN mode cell arrays for spiking AI. The PN mode cell array serves as the sensory spiking neurons. The ES mode cell array serves as synapses, including recurrent forwards-connection (RFC) synapses, recurrent backwards-connection (RBC) synapses and fully connected (FC) synapses, while the EN mode cell array serves as recurrent neurons and output neurons. This configuration processes ten continuous motion steps and predicts future motion directions at the 11th step. Details of the network structure are shown in **Fig. 2.32**, and examples of scenarios in **Fig. 2.33**.

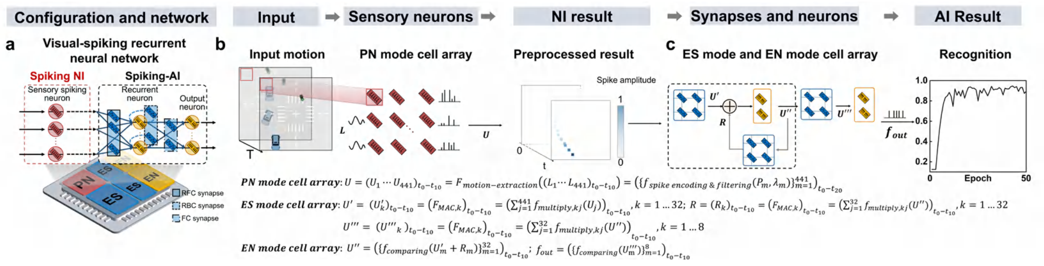


Fig. 2.31 AI-centric visual-spiking recurrent neural network configuration. (a) In this configuration, the cell array is configured into a PN mode cell array as sensory spiking neurons for the spiking NI, an ES mode cell array and an EN mode cell array for the spiking AI. (b) NI-based processing: illustration of the event-based target vehicle motion trajectory extraction and static dark background filtering, outputting spike maps with extracted and filtered motion trajectory (0-*T*). (c) AI-based processing: the prediction accuracy of the trained configuration on the future motion direction.

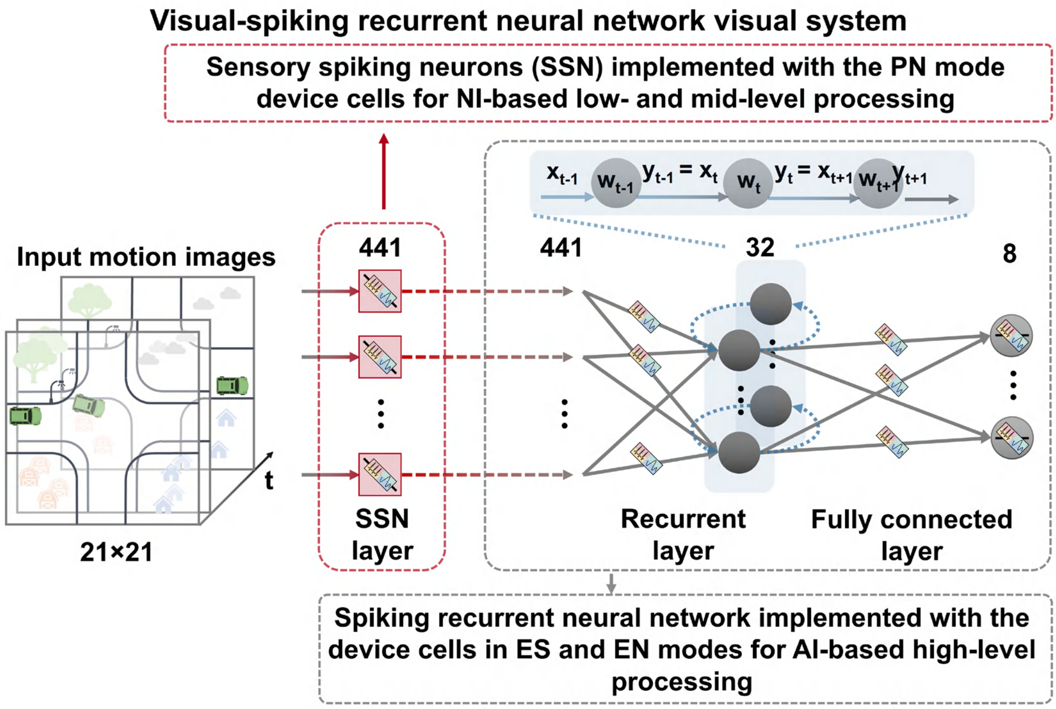


Fig. 2.32 Network structure of visual-spiking recurrent neural network. The input sensory spiking neuron layer of visual-spiking recurrent neural network configuration comprises 441

sensory spiking neurons implemented by the PN mode cell array. The recurrent layer is composed of 32 recurrent neurons, 441 × 32 forward recurrent synapses, and 32 × 32 backward recurrent synapses. The recurrent neurons and recurrent synapses are implemented by the EN mode cell array and ES mode cell array, respectively. The final fully connected layer consists of 8 output neurons implemented with the EN mode cell array and 256 (32 × 8) synapses implemented by the ES mode cell array.

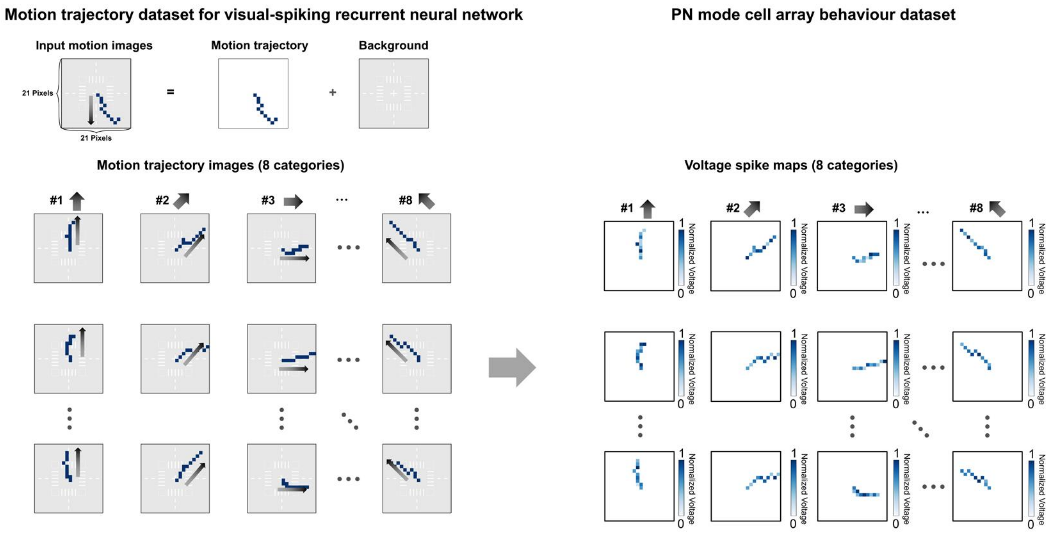


Fig. 2.33 Visual-spiking recurrent neural network datasets. The motion trajectory dataset contains 1,640 distinct motion images (t_0 to t_{10}) with 8 possible moving directions. Each motion image (21 × 21) contains 10 sequential vehicle motion steps. Similarly, the simulated PN mode cell array behaviour dataset consists of 8 possible spike maps (21 × 21) containing 10 sequential spikes. 1,312 of these spike maps were used for training the AI section in this configuration, and the rest of the 328 spike maps were used for testing the AI section of this configuration.

As shown in **Fig. 2.31**b, the PN mode cell array encodes the motion data from input motion image L into voltage spike signals and real-time filters the static background (**Fig. 2.24**b), outputting voltage spike maps $\left(U = (U_1 \cdots U_{441})_{t_0 - t_{10}}\right)$ with extracted motion trajectory. The motion extraction algorithm in this configuration is represented as:

$$F_{motion-extraction}\left((L_1 \cdots L_{441})_{t_0-t_{10}}\right) = \left(\{f_{spike\ encoding\ \&\ filtering}(P_m,\ \lambda_m)\}_{m=1}^{441}\right)_{t_0-t_{20}} \quad (22)$$

The resulting voltage spike maps U are then fed to the ES mode cell arrays for MAC (**Fig. 2.34**) and the EN mode cell arrays for final prediction (**Fig. 2.31**c). The MAC algorithm for each ES mode cell array column (index with k) is represented as:

$$\text{RFC synapse:}\left(F_{MAC,k}\right)_{t_0-t_{10}} = \left(\sum_{j=1}^{441} f_{multiply,kj}(U_j)\right)_{t_0-t_{10}} \quad (23)$$

$$\text{RBC synapse: }\left(F_{MAC,k}\right)_{t_0-t_{10}} = \left(\sum_{j=1}^{32} f_{multiply,kj}(U_j)\right)_{t_0-t_{10}} \quad (24)$$

$$\text{FC synapse:}\left(F_{MAC,k}\right)_{t_0-t_{10}} = \left(\sum_{j=1}^{8} f_{multiply,kj}(U_j)\right)_{t_0-t_{10}} \quad (25)$$

, where each ES mode device performs a device-level multiplication algorithm $f_{multiply}$, as detailed in equation 12.

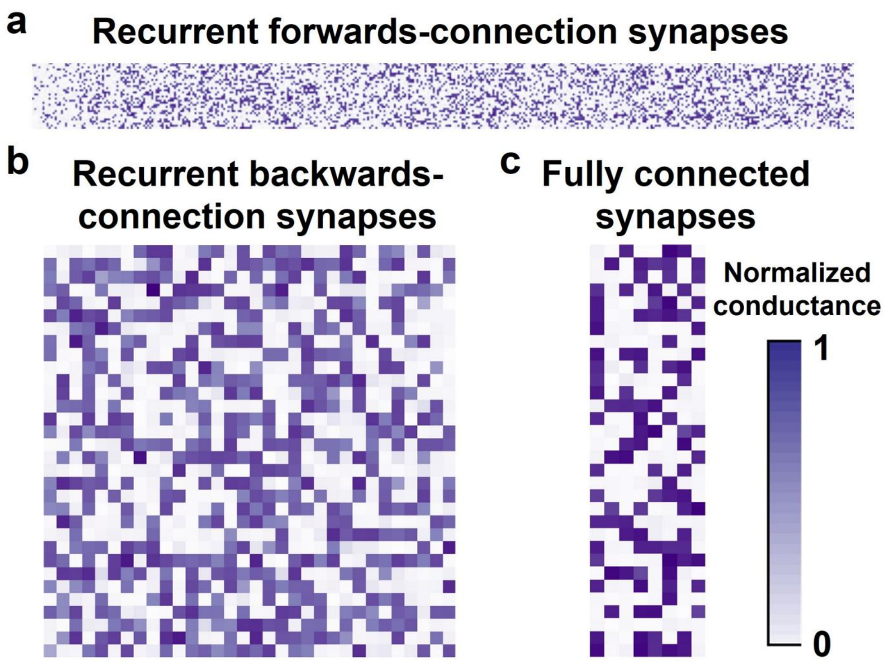


Fig. 2.34 Weight maps of visual-spiking recurrent neural network. (a) recurrent forwards-

connection synapses, (b) recurrent backwards-connection synapses and (c) fully connected synapses.

The predicted motion direction of the 11th step is determined by the spike rates of the EN mode cell array. (details in **Fig. 2.35**). This configuration achieves a prediction accuracy of 92.5%.

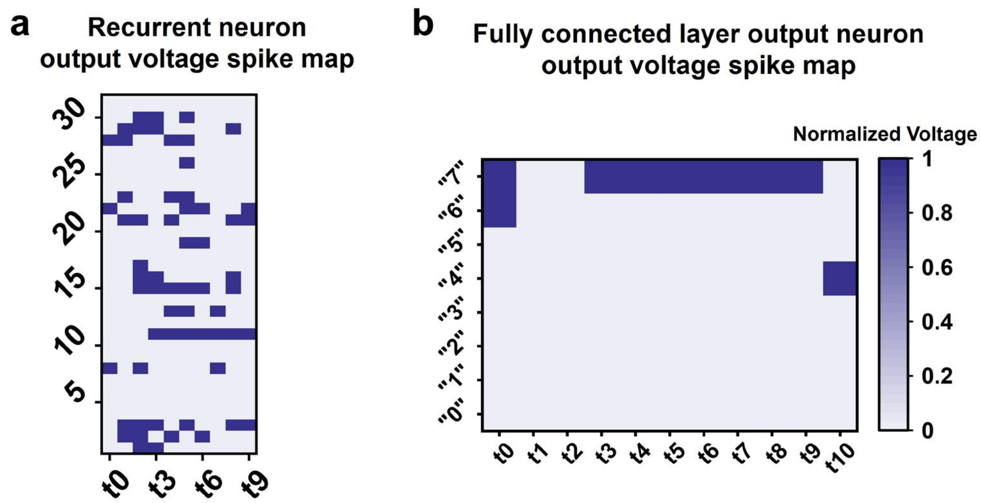


Fig. 2.35 Spike maps of visual-spiking recurrent neural network configuration. (a) The output voltage spike maps of the recurrent neuron and (b) the output voltage spike map of the fully connected layer output neuron generated using an image from the test dataset labelled as class 7.

2.2.4.2.3. Visual-artificial neural network configuration

Fig. 2.36a demonstrates the non-spiking visual-artificial neural network configuration for image enhancement and classification, wherein the cell array is configured into a PS mode cell array as the input sensory spiking neurons for the non-spiking NI, and an ES mode cell array as the synapses for AI. The network architecture is detailed in **Fig. 2.37**, with scenario examples provided in **Fig. 2.38**.

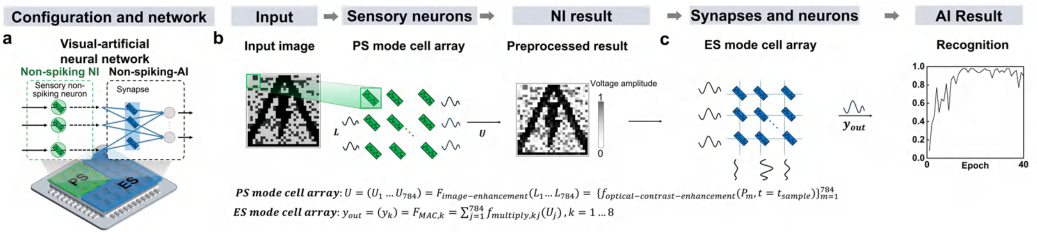


Fig. 2.36 AI-centric visual-artificial neural network configuration. (a) In this configuration, the cell array is configured into a PS mode cell array as the input sensory non-spiking neurons for the non-spiking NI and an ES mode cell array as the synapses for the AI. (b) NI-based processing: illustration of the image enhancement enabled by PS mode cell array, outputting traffic sign image with enhanced body/background and body/noise contrasts. (c) AI-based processing: classification accuracy of the trained visual-artificial neural network configuration on eight categories of noisy traffic sign images.

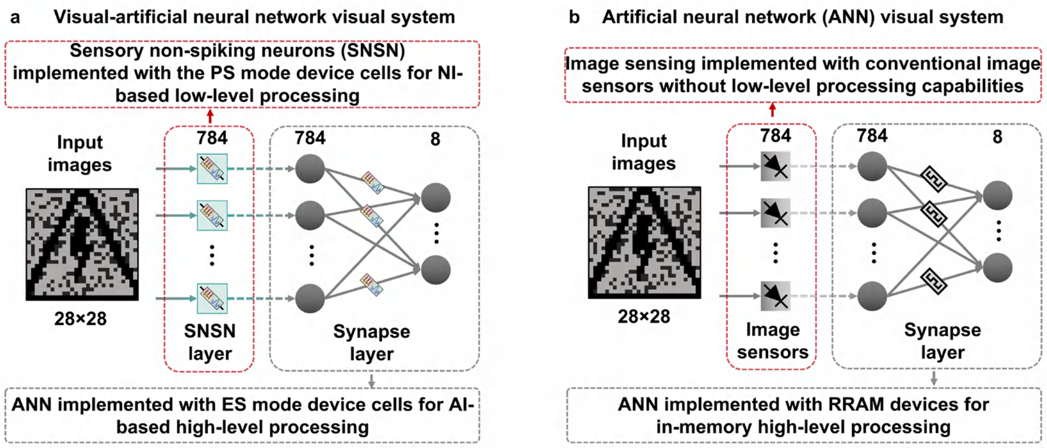


Fig. 2.37 Network structure and implementations of visual-artificial neural network. The visual-artificial neural network consists of an input sensory non-spiking neuron layer (784 neurons) and a synapse layer with 784 × 8 connections. Two different implementations of the visual-artificial neural network-based vision system and a traditional artificial neural network (ANN) vision system are demonstrated: one with multi-paradigm devices and one without. (a) Visual-artificial neural network implemented with PS mode device cells for the sensory non-spiking neuron layer and ES mode device cells for the synapse layer. (b) Traditional ANN

implemented with conventional image sensors for the input layer and electrical resistive random-access memory (RRAM) devices for the synapse layer.

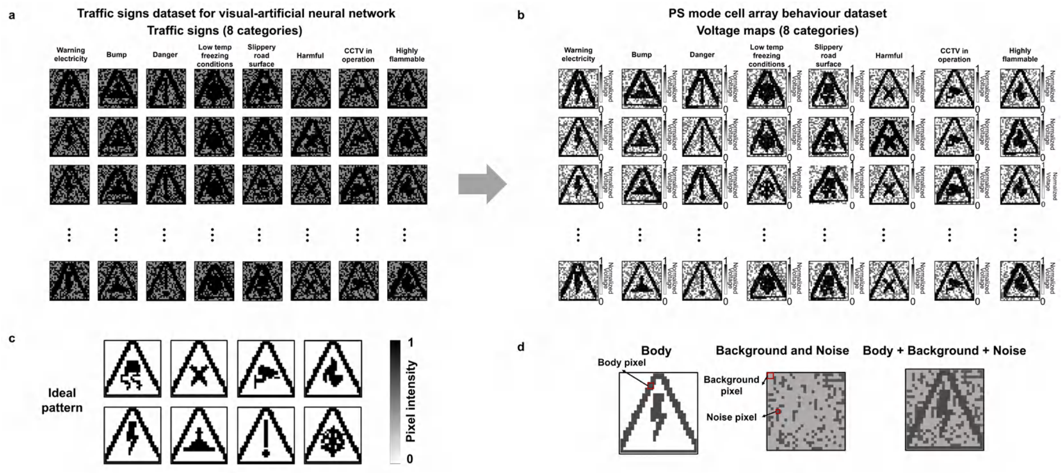


Fig. 2.38 Visual-artificial neural network datasets. (a) The traffic sign image dataset includes 8 categories of 28 × 28 noisy traffic sign images (warning electricity, bump, danger, low temperature/freezing conditions, slippery road surface, harmful, CCTV in operation, and highly flammable). (b) The simulated PS mode cell array behaviour dataset consists of 8 possible voltage map patterns, each with a resolution of 28 × 28 pixels. In the dataset, 192 samples are used for training, 39 for validation, and 48 for testing. (c) Ideal patterns for the 8 traffic sign image categories. (d) Example of a noisy traffic sign image.

As shown in **Fig. 2.36**b, the PS mode cell array first in situ performs the contrast enhancement of the noisy traffic sign images L according to the processing schemes in **Fig. 2.25**b(i). This processing scheme, denoted as $F_{image-enhancement}(L)$, enhances the contrast between the body/background and body/noise, resulting in the enhanced contrasts between body/background and body/noise and a clarified image in the voltage map format $U = (U_1 \ldots U_{784})$. The image enhancement algorithm is expressed as:

$$F_{image-enhancement}(L_1 \ldots L_{784}) = f_{optical-contrast-enhancement}\left(P_m,\ t = t_{sample}\right)_{m=1}^{784} \quad (26)$$

, in which each PS mode device performs a device-level optical contrast enhancement algorithm.

The voltage map U from the PS mode cell array is then fed into the ES mode cell array for MAC operations (**Fig. 2.39**), outputting voltage signals y_{out} with different amplitudes (**Fig. 2.36**c). The highest amplitude indicates the final classified category. The MAC algorithm for each ES mode cell array column (indexed by k) in this configuration is represented as:

$$F_{MAC,k} = \sum_{j=1}^{784} f_{multiply,kj}(U_j) \tag{27}$$

where $f_{multiply}$ denotes the device-level multiplication operation, as defined in equation 12.

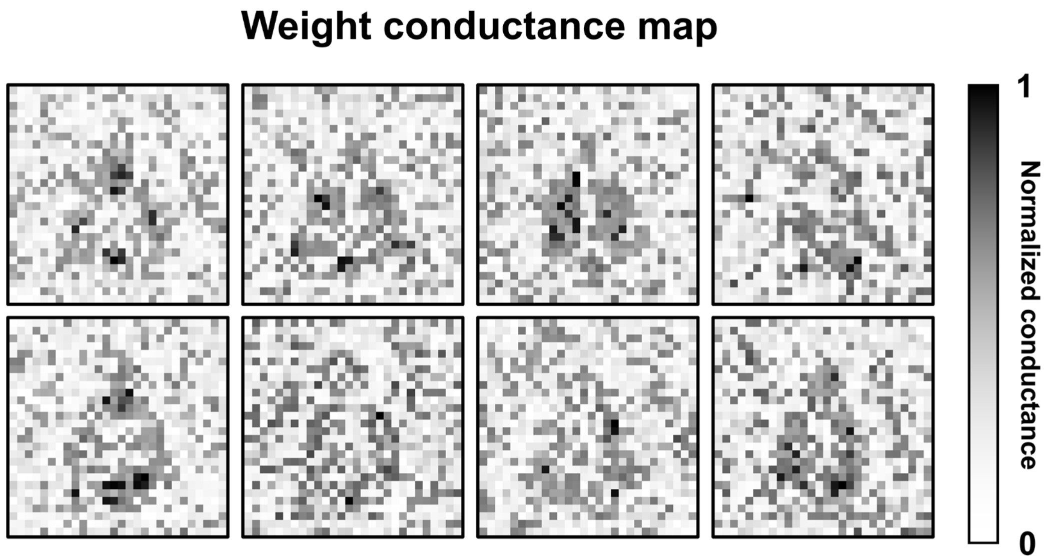


Fig. 2.39 Weight maps of visual-artificial neural network synapse layer.

The trained visual-artificial neural network correctly classifies eight categories of noisy traffic sign images with a high accuracy of 97 %. This represents an 18 % improvement in classification accuracy compared with the system without non-spiking

NI-based low-level processing (**Fig. 2.40**).

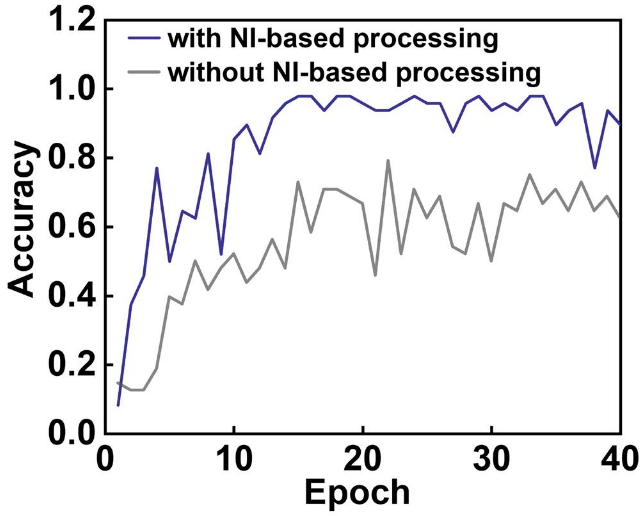


Fig. 2.40 Comparison of classification accuracy of visual-artificial neural network-based system with and without NI-based low-level processing. The system with NI-based processing achieves higher classification accuracy of 97% when recognizing noisy traffic sign images, representing a > 10 % improvement compared to the system without non-spiking NI-based low-level processing functions.

2.2.4.2.4. Visual-reservoir computing configuration

Fig. 2.41 demonstrates the non-spiking visual-reservoir computing configuration for image compression and classification, wherein the cell array is reconfigured into a PS mode cell array as the sensory non-spiking neurons for the non-spiking NI, and an ES mode cell array as the synapses for the reservoir computing based AI. Details of the network structure are shown in **Fig. 2.42**. The PS mode cell array reduces the data volume of dynamic traffic images L (27×27×3) (**Fig. 2.43**) based on the processing schemes depicted in **Fig. 2.25**b(ii), outputting the compressed image in the form of voltage map. The image compression algorithm is expressed as:

$$F_{image-compression}\left((L_1 \dots L_{2187}), n_c = 3\right) = \left\{f_{data-compression}(P_m, n_c = 3)\right\}_{m=1}^{729} \quad (28)$$

, where each PS mode device performs a device-level data compression algorithm. n_c denotes the compression ratio.

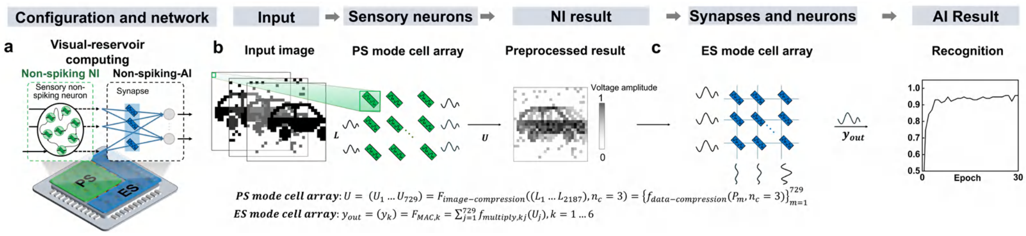


Fig. 2.41 AI-centric visual-reservoir computing configuration. (a) In this configuration, the cell array is configured into a PS mode cell array as the sensory non-spiking neurons for the non-spiking NI, and an ES mode cell array as the synapses for the AI. (b) NI-based processing: illustration of the image compression enabled by PS mode cell array, reducing the output image size to one-third. (c) AI-based processing: classification accuracy of the trained visual-reservoir computing configuration on six categories of noisy traffic images.

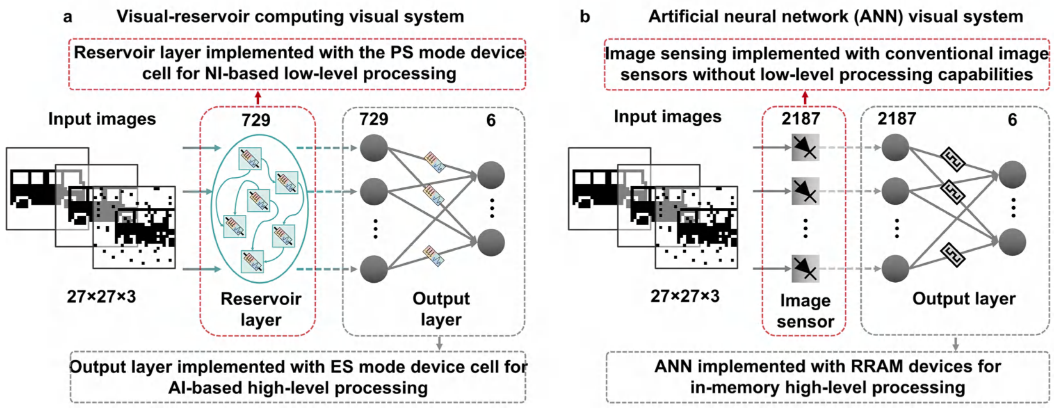


Fig. 2.42 Comparison of the network structures between (a) Visual-reservoir computing based vision system with NI-based processing fully implemented with multi-paradigm devices and (b) Visual-reservoir computing vision system without NI-based processing implemented with conventional image sensors and electrical RRAMs.

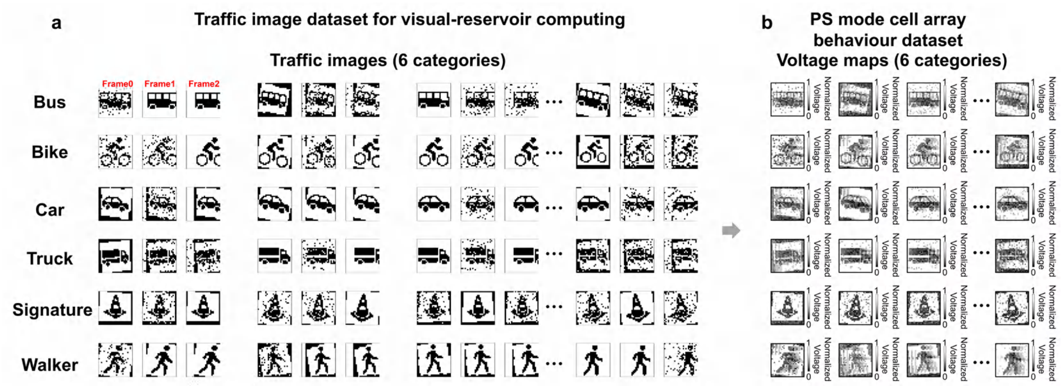


Fig. 2.43 The traffic image dataset and the simulated PS mode behaviour dataset for visual-reservoir computing. (a) The traffic image dataset includes six categories (bus, bike, car, truck, sign, walker). Image noise includes perspective distortion (applied with 50% probability), random rotation (angle ∈ [–15°, 15°], applied with 50% probability), and salt-and-pepper noise (50% probability, amount = 0.2). (b) The corresponding simulated voltage output of the PS mode cell array was generated based on device behaviour measured under data compression operations, which is further used as the PS mode array behaviour dataset to train the AI section in this configuration. Device-level variations, including cycle-to-cycle and cell-to-cell fluctuations, were modelled as Gaussian noise and incorporated during simulation to reflect more realistic PS mode cell array responses.

The voltage maps U from the PS mode cell array are fed into the ES mode cell array for MAC (**Fig. 2.44**), generating output voltage signals y_{out} (**Fig. 2.41**c). The MAC algorithm for each ES mode cell array column (indexed by k) in this configuration is represented as:

$$F_{MAC,k} = \sum_{j=1}^{729} f_{multiply,kj}(U_j) \tag{29}$$

where $f_{multiply}$ denotes the device-level multiplication operation, as defined in equation 12.

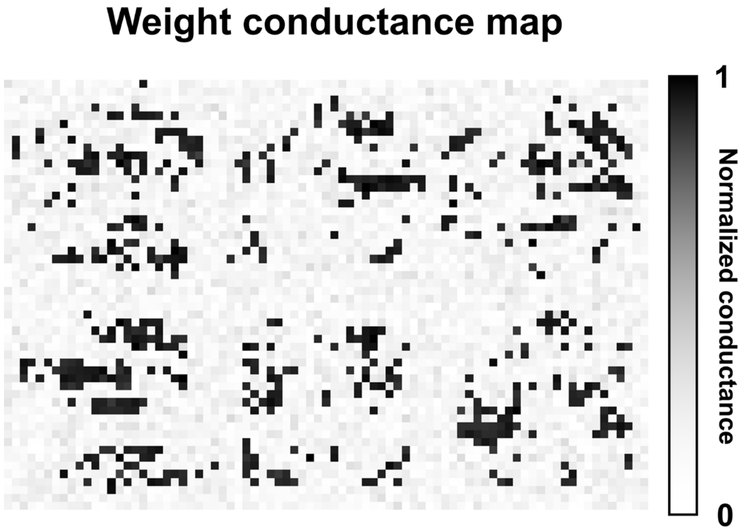


Fig. 2.44 Weight map of visual-reservoir computing output layer. Cell variability was incorporated during simulation by adding Gaussian noise to the conductance values in the visual-reservoir computing output layer, with the noise variance extracted from experimental measurements of the ES mode cell array, reflecting both cell-to-cell and cycle-to-cycle noise.

After training, the visual-reservoir computing configuration classifies all six categories of noisy traffic images with an accuracy of 95.8 %. Compared to systems without NI-based processing, the NI-based image compression in the non-spiking visual-reservoir computing configuration can reduce the data volume at the front end, not only leading to a further reduced network size and reduced computational resources but also enhancing the network training and classification efficiencies (**Fig. 2.42**).

2.2.4.3. Reconfigurable system dataflow and architecture

Fig. 2.45 shows the full hardware design at the system level that can support all the system-level algorithms and versatile scenarios described in **Section 2.2.4.1** and **Section 2.2.4.2**. All the circuits in this vision system are designed using a commercial 180nm CMOS Process Design Kit (PDK) and verified through state-of-the-art EDA tools simulations.

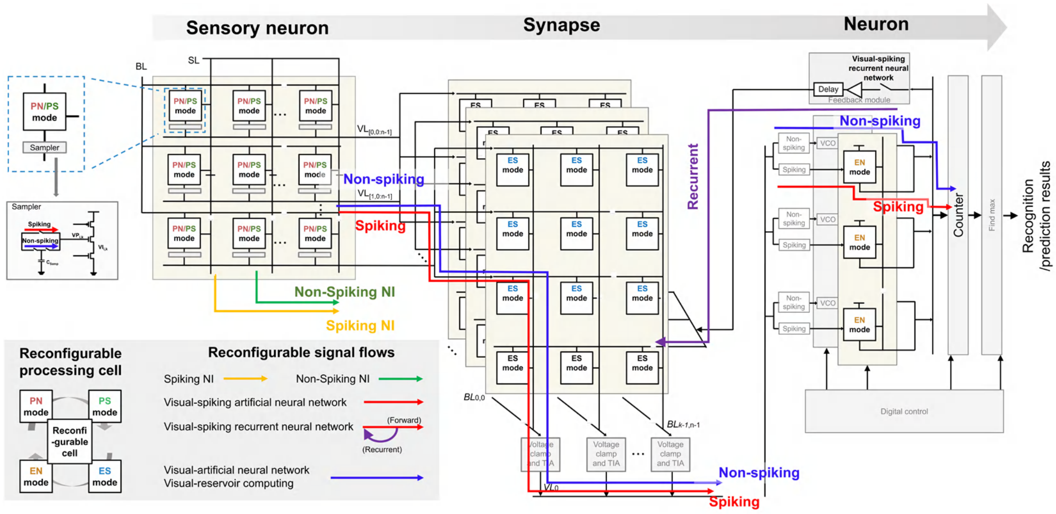


Fig. 2.45 Reconfigurable vision system with system-level reconfigurability. The reconfigurable dataflow architecture supports spiking/non-spiking/NI/AI processing and all six configurations.

The system's dataflow initiates with the reconfigurable spiking/non-spiking NI processing path. As shown in **Fig. 2.45**, the PN or PS cells serve as sensory neurons for smart imaging and low- and mid-level processing. Each sensory neuron is constructed by a PN or PS mode device cell, a bypassable capacitive sampler and a driver. The bypass capacitive sampler supports reconfiguring the dataflow path for spiking and non-spiking NI. Specifically, the non-spiking PS mode cell output is sampled at a configurable time, while the spiking PN mode cell output is relayed directly. The NI configuration results can then be read from the driver as sensory neuron outputs.

When the system is reconfigured for AI-centric processing, the sensory neuron outputs are then fed into the synapse module consisting of a 1T1R-structured ES mode cell array for efficient MAC operations. For the visual-spiking recurrent neural network cases, additional recurrent paths are facilitated between the synapse and neuron modules with a configurable delay. Finally, the synapse outputs are passed to the neuron module for making a decision, which includes a dual-path spike generator and a unified decision-making block for supporting both spiking and non-spiking signals efficiently. The dual-path spike generator outputs spike trains using an EN cell during spiking processing or converts the non-spiking voltage signals to spike trains using a voltage-controlled oscillator (VCO) during the non-spiking cases. A decision is thus made by counting the number of output spikes and finding the maximum one using the unified decision-making block. **Fig. 2.46** shows the hardware implementation results of system operating modes (visual-spiking artificial neural network mode, visual-spiking recurrent neural network mode, visual-artificial neural network mode and visual-reservoir computing mode), demonstrating the expected behaviours of the system.

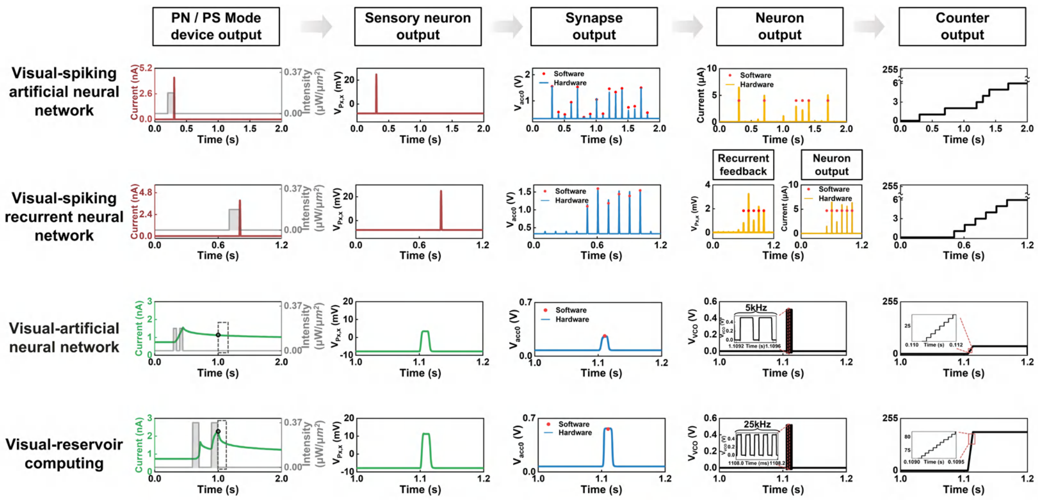


Fig. 2.46 Hardware simulation results of visual-spiking artificial neural network, visual-spiking recurrent neural network, visual-artificial neural network, and visual-reservoir computing modes.

The flexible reconfiguration between NI- and AI-centric configurations allows our system to be not only suitable for edge vision scenarios, but also for scenarios that demand AI workloads involving millions of parameters. In the NI-centric configurations, it handles high-resolution image sensing and preprocessing. In the AI-centric configurations, it can perform some preliminary high-level tasks, like object recognition. These integrated processing capabilities help reduce the amount of data that needs to be sent to the processor at the data center, in addition with enhancing the data quality, improving the overall system efficiency and accuracy.

2.2.4.4. Benchmarking with other reconfigurable vision systems

In NI-based processing, our system can achieve 9.1-52.6 TOPS/W energy efficiency, which is 86-107 times higher than the currently reported state-of-the-art. In AI-based processing, our system shows an energy efficiency of 2.3-2.7 TOPS/W for spiking neural networks and 76.0-76.5 TOPS/W for non-spiking neural networks. Moreover, our system can uniquely perform NI-AI combined processing (V-NNs), demonstrating 2.3-75.5 TOPS/W energy efficiency, which are 28-922 times better than the reported state-of-the-art. The area efficiency is 2.29×10^4 to 7.53×10^5 OPS/F^2, which is 2986-3900 times better than current state-of-the-art reported neuromorphic vision systems[43-48]. **Table 2.9** presents the comparisons of the representative reconfigurable vision systems with our work. The evaluation of the power consumption and energy efficiency of the system across various configurations in **Table 2.9** is detailed in **Table 2.10-2.15**.

Table 2.9 Comparisons of the representative reconfigurable vision systems with our work

43-48

		This work *	High-Speed 3D Vision Chip [43]	Spiking Vision Chip [44]	4.9MP Programmable Sensor [45]	SCAMP-7 [46]	SCAMP-5 [47]	High-Speed In-Focal-Plane Chip [48]
Architecture		In-sensor and In-memory computing	Near-sensor computing	Near-sensor and Near-memory computing	Near sensor computing			
Processing modes		• Spiking NI • non-spiking NI; • Spiking AI • non-spiking AI	Non-spiking NI	• Spiking NI • Spiking AI	Non-spiking NI			
Reconfigurable level	Reconfiguration granularity	Device level	Circuit level					
	Reconfigurable processing	Comprehensive spiking/ non-spiking/ NI/ AI reconfigurable	Only NI-reconfigurable	Only NI- or AI-reconfigurable (NI-AI non-reconfigurable)	Only NI-reconfigurable			
NI-based processing	Power consumption per pixel	Spiking NI: 1.1 pW Non-spiking NI: 1.52 pW	Non-spiking NI 181.83 nW	Not reported	Non-spiking NI 65.1 nW	Not reported	Non-spiking NI 18.77 μW	Non-spiking NI 0.916 μW
	Energy efficiency	Spiking NI: 9.1 TOPS/W Non-spiking NI: 52.6 TOPS/W	Non-spiking NI 0.61 TOPS/W		Not reported		Non-spiking NI 535 GOPS/W	Non-spiking NI 85 GOPS/W
AI-based processing	Power consumption	V-SANN/V-SRNN: 0.09 – 0.23 μW V-ANN/V-RC: 18.4- 26.3 nW **	Not supported		Not supported			
	Energy efficiency	V-SANN/V-SRNN: 2.3 – 2.7 TOPS/W V-ANN/V-RC: 76-76.5TOPS/W ***						
Hybrid NI/AI processing	Power consumption	V-SANN/V-SRNN: 0.09 – 0.23 μW V-ANN/V-RC: 19.5- 27.4 nW ****		875 mW				
	Energy efficiency	V-SANN/V-SRNN: 2.3-2.7 TOPS/W V-ANN/V-RC: 74.5-75.5 TOPS/W *****		81.9 GOPS/W				
Area efficiency		$2.29\times10^4 – 7.53\times10^5$ OPS/F^2	1.93×10^2 OPS/F^2	24 OPS/F^2	Not reported	Not reported	1.73×10^2 OPS/F^2	7.68 OPS/F^2

* The power consumption is calculated using 5 μm × 5 μm multi-paradigm devices. The power consumption can be further reduced when the device is scaled down. The image size for spiking NI is 116 × 116, while for non-spiking NI it is 111 × 111. The image size for V-SANN (visual-spiking artificial neural network configuration) is 20 × 20, for V-SRNN (visual-spiking recurrent neural network configuration) is 21 × 21, for V-ANN (visual-artificial neural network configuration) is 28 × 28, and for V-RC (visual-reservoir computing configuration) is 27 × 27. The V-SANN and V-SRNN are estimated at a frequency of 10Hz, while the V-ANN and V-RC are estimated at 80Hz. The power consumption of the peripheral circuits is obtained using Cadence Virtuoso @ 180nm.

** For AI-based processing, the power consumption is 85.2 nW for V-SANN, 228 nW for V-SRNN, 26.3 nW for V-ANN and 18.4 nW for V-RC.

*** For AI-based processing, the energy efficiency is 2.3 TOPS W^{-1} @ 10 Hz for V-SANN, 2.7 TOPS W^{-1} @ 10 Hz for V-SRNN, 76.5 TOPS W^{-1} @ 80 Hz for V-ANN and 76.0 TOPS W^{-1} @ 80 Hz for V-RC.

**** For hybrid NI/AI processing, the power consumption is 85.2 nW for V-SANN, 228 nW for V-SRNN, 27.4 nW for V-ANN, and 19.5 nW for V-RC.

***** For hybrid NI/AI processing, the energy efficiency is 2.3 TOPS W^{-1} @ 10 Hz for V-SANN, 2.7 TOPS W^{-1} @ 10 Hz for V-SRNN and 75.5 TOPS W^{-1} @ 80 Hz for V-ANN and 74.5 TOPS W^{-1} @ 80 Hz for V-RC.

Table 2.10 Calculation Table for Spiking NI configuration (10Hz; input image size 116 × 116)

Metrics	Calculation details
Operations per second	$N_{OP,\,PN}$ = 13456 × 10 Hz = 134560 OPS
Power	P_{Total} = 1.1 pW × 13456 = 14.8 nW
Power Efficiency	Efficiency = $N_{OP,\,PN}$ / P_{Total} = 134560OPS / 1.480 nW = 9.09 TOPS/W

Table 2.11 Calculation Table for Non-spiking NI configuration (80Hz; input image size 111 × 111)

Metrics	Calculation details
Operations per second	$N_{OP,\,PS}$ = 12321 × 80 Hz = 985680 OPS
Power	P_{Total} = 1.52 pW × 12321 = 18.73 nW
Power Efficiency	Efficiency = $N_{OP,\,PS}$ / P_{Total} = 985680 / 18.73 nW = 52.63 TOPS/W

Table 2.12 Calculation Table for visual-spiking artificial neural network configuration (10 Hz; input image size 20 × 20)

Metrics	**Calculation details**		
Operations per second	NI-based processing	PN mode cell array	$N_{OP, PN}$ = 400 × 1 OP × 10 Hz = 4000 OPS
	AI-based processing	ES mode cell array	$N_{OP, ES}$ = (4800 × 2 × 2) OP × 10 Hz = 192000 OPS
		EN mode cell array	$N_{OP, EN}$ = 12 × 1 OP = 12 OPS
		Digital circuits	$N_{OP, Digital\ circuits}$ = 24 × 1 OP × 10 Hz = 240 OPS
	Total		$N_{OP, Total} = N_{OP, PN} + N_{OP, ES} + N_{OP, EN} + N_{OP, Digital\ circuits}$ = 196252 OPS
Power	NI-based processing	PN mode cell array	P_{PN} = 1.10 × 10^{-12} W × 400 =0.44 nW
	AI-based processing	ES mode cell array	P_{ES} = (4800 × 2/10 × 19 pW) + (12 × 291 pW) = 0.217 nW
		EN mode cell array	P_{EN} = 0.2 × 315 nW + 12 × 1.1 pW = 63 nW
		Digital circuits	24.72 pW
	Total		$P_{Total} = P_{PN} + P_{ES} + P_{EN} + P_{digital\_circuits}$ = 85.2 nW
Power efficiency	NI-based processing		$N_{OP, PN} / P_{PN}$ = 4000 OPS / 0.44 nW = 9.09 TOPS/W
	AI-based processing		$(N_{OP, ES} + N_{OP, EN} + N_{OP,Digital\ circuits}) / (P_{ES} + P_{EN} + P_{digital\_circuits})$ = 192252 OPS / 84.8 nW = 2.27 TOPS/W
	Hybrid NI/AI computing		$N_{OP, Total} / P_{Total}$ = 196252 OPS / 85.2 nW = 2.30 TOPS/W

Table 2.13 Calculation Table for visual-spiking recurrent neural network configuration (10 Hz; input image size 21 × 21)

Metrics	Calculation details		
Operations per second	NI-based processing	PN mode cell array	$N_{OP,\ PN}$ = 441 × 1 OP × 10 Hz = 4410 OPS
	AI-based processing	ES mode cell array	$N_{OP,\ ES}$ = (14112 + 1024 + 256) × 2 × 2 OP × 10 Hz = 615680 OPS
		EN mode cell array	$N_{OP,\ EN}$ = (32 + 8) × 1 OP = 40 OPS
		Digital circuits	$N_{OP,\ Digital\ circuits}$ = 16 × 1 OP × 10 Hz = 160 OPS
	Total		$N_{OP,\ Total} = N_{OP,\ PN} + N_{OP,\ ES} + N_{OP,\ EN} + N_{OP,\ Digital\ circuits}$ = 620290 OPS
Power	NI-based processing	PN mode cell array	P_{PN} = 1.10 × 10^{-12} W × 441 = 485 pW
	AI-based processing	ES mode cell array	$P_{ES} = P_{ES\_RFC} + P_{ES\_RBC} + P_{PES\_FC}$ = (14112 × 2/10 × 19 pW) + (1024 × 2/10 × 19 pW) + (32 × 291 pW) + (256 × 2/10 × 19 pW) + (8 × 291 pW) = 70.1 nW
		EN mode cell array	P_{EN} = 32 × 0.125 × 0.1 × 315 nW + 32 × 1.1 pW + 8 × 0.125 × 0.1 × 315 nW + 8 × 1.1 pW = 158 nW
		Digital circuits	16.48 pW
	Total		$P_{Total} = P_{PN} + P_{ES} + P_{EN} + P_{digital\_circuits}$ = 228 nW
Power efficiency	NI-based processing		$N_{OP,\ PN}$ / P_{PN} = 4410 OPS / 485 pW = 9.09 TOPS/W
	AI-based processing		($N_{OP,\ ES} + N_{OP,\ EN} + P_{digital\_circuits}$) / ($P_{ES} + P_{EN} + P_{digital\_circuits}$) = 615880 OPS / 228 nW = 2.70 TOPS/W
	Hybrid NI/AI computing		$N_{OP,\ Total}$ / P_{Total} = 620290 OPS / 228 nW = 2.72 TOPS/W

Table 2.14 Calculation Table for visual-artificial neural network configuration (80Hz; input image size 28 × 28)

<table>
<tr><th>Metrics</th><th colspan="3">Calculation details</th></tr>
<tr><td rowspan="4">Operations per second</td><td>NI-based processing</td><td>PS mode cell array</td><td>$N_{OP, PS}$ = 784 × 1 OP × 80 Hz = 62720 OPS</td></tr>
<tr><td rowspan="2">AI-based processing</td><td>ES mode cell array</td><td>$N_{OP, ES}$ = (6272 × 2 × 2) OP × 80 Hz = 2007040 OPS</td></tr>
<tr><td>Digital circuits</td><td>$N_{OP, Digital\ circuits}$ = 24 × 1 OP × 80 Hz = 1920 OPS</td></tr>
<tr><td colspan="2">Total</td><td>$N_{OP, Total} = N_{OP, PS} + N_{OP, ES} + N_{OP, Digital\ circuits}$ = 2071680 OPS</td></tr>
<tr><td rowspan="4">Power</td><td>NI-based processing</td><td>PS mode cell array</td><td>$P_{PS} = 1.52 \times 10^{-12}$ W × 784 = 1.19 nW</td></tr>
<tr><td rowspan="2">AI-based processing</td><td>ES mode cell array</td><td>P_{ES} = (6272 × 2/10 × 19 pW) + (8 × 291 pW) = 26.2 nW</td></tr>
<tr><td>Digital circuits</td><td>92.48 pW</td></tr>
<tr><td colspan="2">Total</td><td>$P_{Total} = P_{PS} + P_{ES} + P_{digital\_circuits}$ = 27.4 nW</td></tr>
<tr><td rowspan="3">Power efficiency</td><td colspan="2">NI-based processing</td><td>$N_{OP, PS} / P_{PS}$= 62720 OPS / 1.19 nW = 52.63 TOPS/W</td></tr>
<tr><td colspan="2">AI-based processing</td><td>$(N_{OP, ES} + N_{OP, Digital\ circuits}) / (P_{ES} + P_{digital\_circuits})$ = 2008960 OPS / 26.3 nW = 76.52 TOPS/W</td></tr>
<tr><td colspan="2">Hybrid NI/AI computing</td><td>$N_{OP, Total} / P_{Total}$ = 2071680 OPS / 27.4 nW = 75.48 TOPS/W</td></tr>
</table>

Table 2.15 Calculation Table for visual-reservoir computing configuration (80Hz; input image size 27 × 27)

Metrics	Calculation details		
Operations per second	NI-based processing	PS mode cell array	$N_{OP,\ PS}$ = 729 × 1 OP × 80 Hz = 58320 OPS
	AI-based processing	ES mode cell array	$N_{OP,\ ES}$ = (4374 × 2 × 2) OP × 80 Hz = 1399680 OPS
		Digital circuits	$N_{OP,\ Digital\ circuits}$ = 18 × 1 OP × 80 Hz = 1440 OPS
	Total		$N_{OP,\ Total} = N_{OP,\ PS} + N_{OP,\ ES} + N_{OP,\ Digital\ circuits}$ = 1459440 OPS
Power	NI-based processing	PS mode cell array	P_{PS} = 1.52 pW × 729 = 1.11 nW
	AI-based processing	ES mode cell array	P_{ES} = (4374 × 2/10 × 19 pW) + (6 × 291 pW) = 18.4 nW
		Digital circuits	68.16 pW
	Total		$P_{Total} = P_{PS} + P_{ES} + P_{digital\_circuits}$ = 19.5 nW
Power efficiency	NI-based processing		$N_{OP,\ PS} / P_{PS}$ = 58320 OPS / 1.11 nW = 52.63 TOPS/W
	AI-based processing		$(N_{OP,\ ES} + N_{OP,\ Digital\ circuits}) / (P_{ES} + P_{digital\_circuits})$ = 1401120 OPS / 18.4 nW = 76.0 TOPS/W
	Hybrid NI/AI computing		$N_{OP,\ Total} / P_{Total}$ = 1459440 OPS / 19.5 nW = 74.68 TOPS/W

For this work, area efficiency was estimated as: $\text{Area Efficiency (OPS/F}^2) = \frac{\text{Opeartions per second}}{\text{Feature size (F}^2)}$. Feature size was obtained by scaling system area to the adopted technology node (180 nm). The **Table 2.16** shows the minimum and maximum values for area efficiency based on the system configurations.

Table 2.16 Calculation Table for area efficiency.

Metric	Min Value	Max Value
Operations per Second per Watt (OPS/W)	2.3×10^{12}	7.55×10^{13}
Total Area (um^2)	3.2495×10^{6}	
Scaled Area (F^2)	1.0029×10^{8}	
Area Efficiency (OPS/F^2) = Operations Per second (OPS) / Feature size (F^2)	2.2933×10^{4}	7.5279×10^{5}

Chapter 3 Multidimensional intelligent vision system based on MoS_2/HZO FeFET device

3.1. Introduction

Human visual perception and recognition systems excel in their robustness and efficiency under complex and dynamic scenarios, able to handle multidimensional visual input attributes (e.g. intensity, colour, shape, contour features, concept, category features) and challenging conditions (e.g. occlusion, clutter and corruption). Although advancements have been made in machine vision systems, conventional systems are vulnerable and inefficient in processing multidimensional data with random image corruption, adversarial perturbation and distribution shifts[49,50]. Conventional machine vision systems typically employ a separate sensing-memory-processing Von Neumann architecture, relying on bulky, multiple colour filters and sensors, followed by complex processing, including multiscale-multichannel feature extraction and reconstruction, neural networks-based segmentation and identification[51-54], resulting in high latency, large chip size, and bandwidth bottlenecks. Furthermore, they suffer from adaptability challenges to distribution shifts and varying noise[55,56] in dynamically changing visual environments.

Biological visual systems excel in efficiently and robustly processing multidimensional visual information in an intellectual way, following a hierarchical pathway composed of multiple levels of visual filters[57]. The raw multidimensional visual information is first captured and filtered to multidimensional pixel features, then to lower-dimensional surface features, and finally to the one-dimensional semantic features[58-60]. Visual processing starts from the low-level filtering in the retina, where photoreceptors convert multidimensional light data (e.g. brightness and colour[61]) into distinguishable electrical signals. As these signals move through the brain, they are further simplified: mid-level areas (e.g. V4) extract lower-dimensional features like texture and contour, while high-level regions like the inferotemporal cortex (IT) refine the data into one-dimensional semantic representations for object recognition. Such hierarchical refinement allows the brain to manage complex visual environments efficiently, filtering out noise and separating clutters to better focus on relevant patterns[62-64]. Hardware-wise, a straightforward neuromorphic approach is to design and integrate emerging device-emulated multiple visual filters that fully mimic these hierarchical multidimensional processing pathways. However, severe limitations still exist in current reported devices. They are constrained to single-dimensional sensing at the pixel level, and have limited multidimensional light information resolving capabilities. In addition, they are not able to conduct mid-level processing (e.g. segmentation of the clusters in input) for multidimensional data, a function critical for distinguishing objects within a scene and enabling multi-object recognition. Moreover, while constructing a phototransistor-based array-based photoresponsivity matrix with subpixels-constituted pixel core architecture can enable high-level processing (e.g. recognition)[65], they face severe challenges in scalability, noise, and interconnection complexity when locally integrating multiple subpixels and globally integrating pixel cores. Consequently, significant gaps remain in the device functionalities and efficient and hierarchical multidimensional processing architectures in neuromorphic vision systems, calling for serious efforts to address them.

In this work, we address the above adequacies by reporting a multifunctional visual filter based on MoS_2/HZO FeFET devices, which can fully emulate the comprehensive low- to mid- to high-level filtering for multidimensional visual data within a single array, using the co-designed hierarchical visual filter bank algorithm (HVBF). Importantly, both low-level visual filtering (pixel-level filtering) and mid-level filtering (surface-level filtering) can be directly realized by the device behaviours. The low-level filtering enables pixel-level colour clustering based on colour clustering capability, while mid-level processing decouples the intensity and wavelength dimensions by leveraging the device's multidimensional decoupling capability. The mid-level filtering can reliably decouple the colour and intensity dimension across a wide wavelength range (three colours, RGB) and a wide dynamic intensity range (0.007 mW/mm^2 to 0.700 mW/mm^2, two orders of magnitude), facilitating effective image segmentation based on colour and intensity. Thereafter, the high-level filtering (semantic-level filtering) is then achieved based on the device's capabilities in non-volatile ferroelectric-based multi-level storage and polarization-enhanced multi-photoresponsivity states, and together with the co-designed HVBF algorithm, enabling the array to compute the similarity between input patterns with stored templates for recognition and filter out irrelevant data based on response strength. We then demonstrate the high robustness and accuracy in two complex, noisy multidimensional scenarios, achieving high accuracy rates of 93.3% and 93.8% in recognizing noisy coloured digits and identifying overlapping digits in a noisy background.

3.2. Results

3.2.1. Multidimensional MoS_2/HZO FeFET device

3.2.1.1. Device structure and characterization

The device structure and optical image are shown in **Fig. 3.1**. The grazing-incidence X-ray diffraction (GIXRD) pattern of the HZO layer is shown in **Fig. 3.2**, confirming the crystallization of the ferroelectric o-phase in HZO. The ferroelectric switching behaviour of the HZO films is further proven by the piezo-response force microscopy (PFM) technique, showing functional switching (**Fig. 3.3**). Following the formation of the dielectric stack, a monolayer of MoS_2 was transferred onto the HfO_2/HZO stack to serve as the channel material. The Atomic Force Microscopy (AFM), Raman spectroscopy, Photoluminescence (PL), and Second Harmonic Generation (SHG) characterizations are performed to verify the monolayer thickness of the MoS_2 (**Fig. 3.4**a-e). **Fig. 3.4**f shows that the absorption spectrum of MoS_2 cuts off near 700 nm. A final S/D contact using Au is deposited to complete the device fabrication.

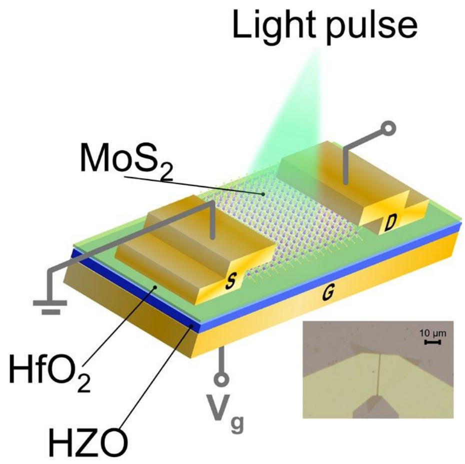


Fig. 3.1 Schematic illustration of the device structure and the optical image of the MoS_2/HZO device.

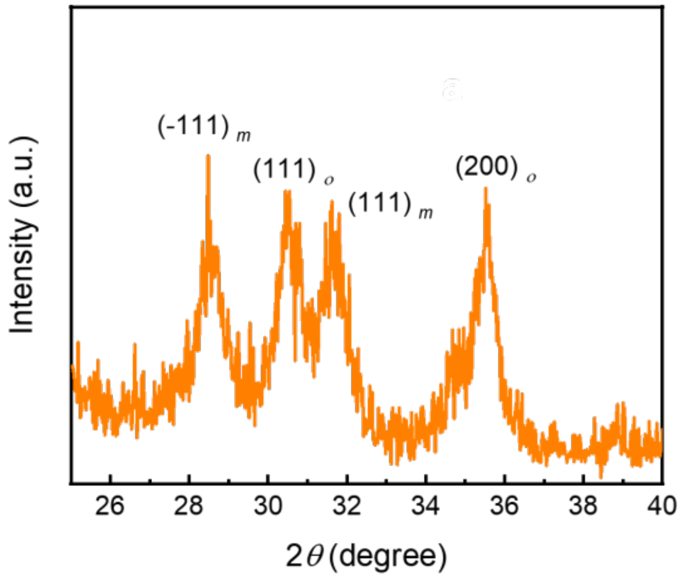


Fig. 3.2 GI-XRD (Grazing Incidence X-Ray Diffraction) characterization of HZO thin film. The diffraction patterns exhibit clear peaks at 30.6° and 35.5°, corresponding to the (111) and (200) crystal planes of the ferroelectric orthorhombic structural phase ($Pca2_1$, o phase), respectively. Additional peaks at ~28.5° and ~31.5° are assigned to the (−111) and (111) reflections of the monoclinic phase ($P2_1/c$).

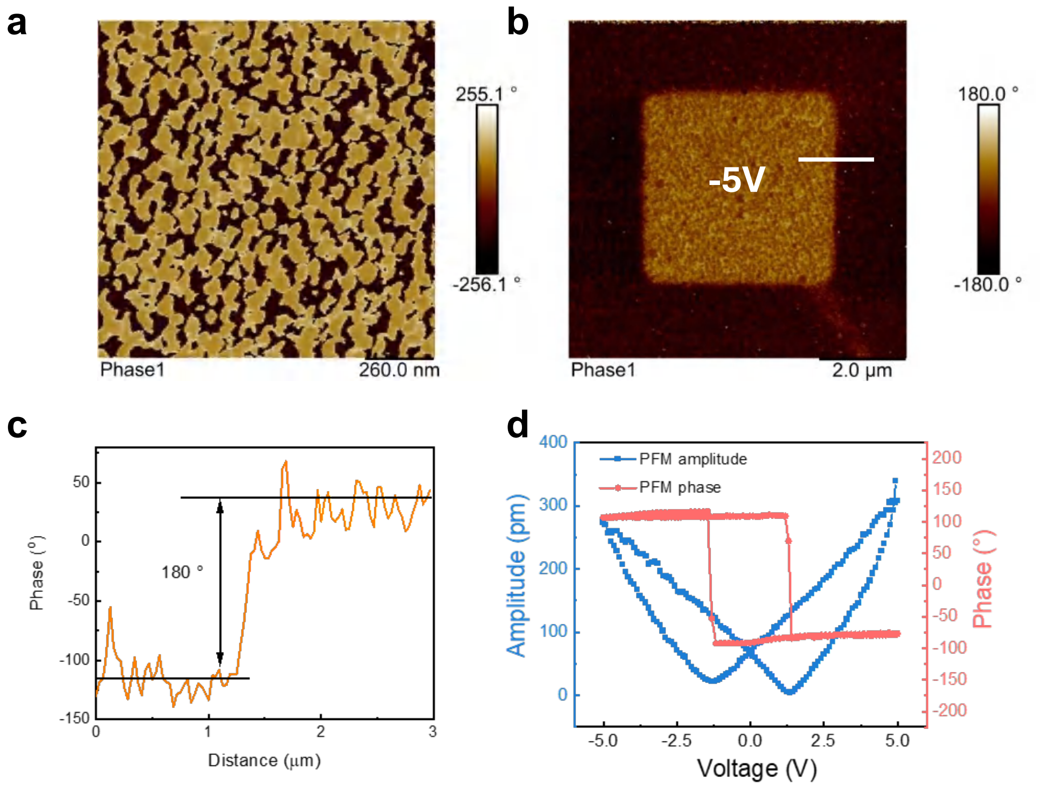


Fig. 3.3 PFM characterization of HZO thin film. (a) PFM phase image of HZO film in initial state. (b) PFM phase image of HZO film after applying a ±5V programming voltage at the pin tip. (c) PFM phase profile of the white line in b. (d) Phase and amplitude switching spectroscopy loops for HZO thin film, demonstrating ferroelectric hysteresis.

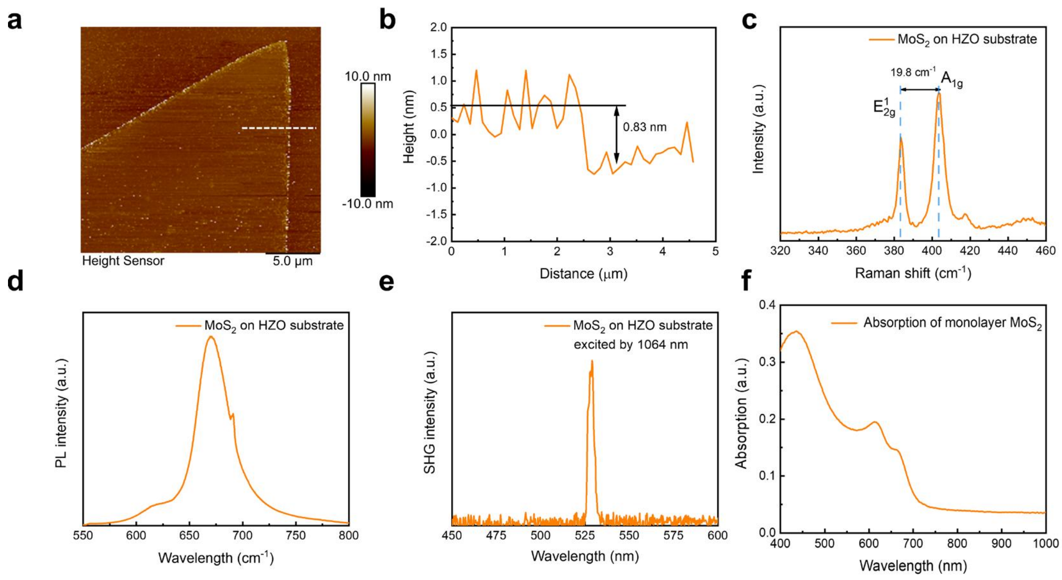


Fig. 3.4 Characterization of Monolayer MoS_2 material. (a) AFM image of the MoS_2 channel

material. (b) The thickness profile over distance of the dotted line in (a), displaying a monolayer thickness of approximately 0.83 nm. (c) Raman spectroscopy analysis of MoS_2 channel material using 532 nm laser excitation. The observed wavenumber difference between the A_{1g} and E_{2g} Raman modes of MoS_2 is consistent with a monolayer MoS_2, as reported[66]. (d) Photoluminescence (PL) of MoS_2 channel material. The photoluminescence (PL) spectra show two peaks at approximately 620 nm and 670 nm, corresponding to the B_1 and A_1 excitons of monolayer MoS_2, respectively, following the reported values in literature[67]. (e) Second harmonic generation (SHG) of MoS_2 channel material. Under 1064 nm excitation, the MoS_2 monolayer channel material presents a sharp peak at ~532 nm, consistent with its expected second harmonic generation. (f) Monolayer MoS_2 absorption spectrum. The absorption spectrum shows a gradual decrease from 405 nm and 532 eventually cutting off around 700 nm.

The transfer curve of the MoS_2/HZO FeFET device in a dark environment exhibits a typical counterclockwise hysteresis behaviour under a gate voltage (V_g) sweep from -4 V to 4 V with an applied V_{ds} of 0.1 V, indicating ferroelectric polarization switching within the HZO layer (**Fig. 3.5**a).

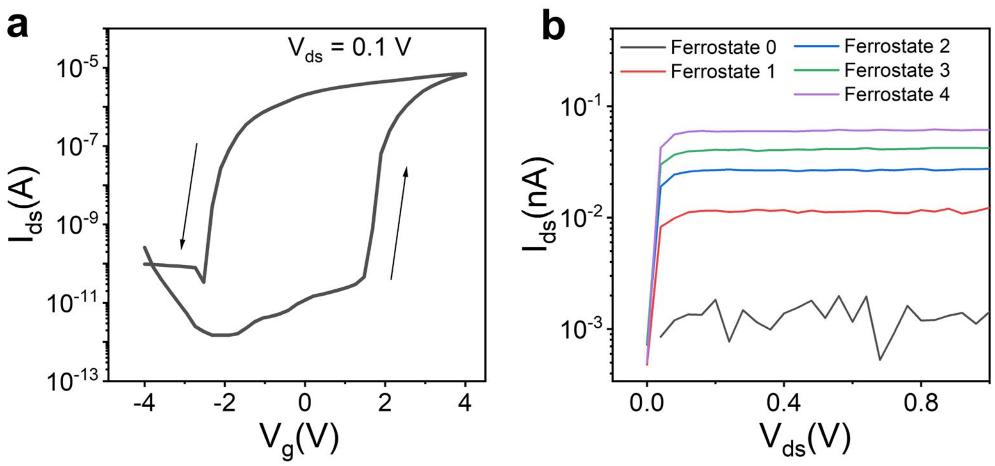


Fig. 3.5 Electric characterization of MoS_2/HZO device. (a) Transfer curve (I_{ds}-V_g) of the

MoS_2/HZO device under -4 V to 4 V gate voltage sweep. (b) Output characteristics (I_{ds}-V_{ds}) of the MoS_2/HZO device measured in the dark condition.

The corresponding output curves (I_{ds}-V_{ds}) presented in **Fig. 3.5**b are measured without any gate voltage applied but with the HZO layer programmed to four different ferroelectric states in advance. They are labelled as ferrostate1, ferrostate2, ferrostate3, and ferrostate4, programmed by the application of four different gate voltages of -0.5 V, -0.8 V, -1.2 V, and -1.5 V, respectively.

The mechanism of multilevel conductance states observed under the different gate voltages can be attributed to the modulation of ferroelectric polarization on the MoS_2/Au interfacial barrier height between the source and drain Au metal electrode and the channel[68-70]. After applying a gate voltage of 2 V, the device is modulated to high resistance state, in which the upward-oriented ferroelectric domains in HZO dominate. The negative interfacial charge at the HfO_2/HZO interface will increase the MoS_2/Au interfacial barrier height, which suppresses the electron collection at the drain end. As a result, the channel conductance state is lower. As the device transitions from the ferroelectric1 to ferroelectric4 (the gate voltage changes from -0.5 V to -1.5 V), the proportion of downward-oriented ferroelectric domains in HZO increases, reducing the barrier height and promoting the electron collection at the drain end, leading to an increased channel conductance state.

Fig. 3.6 shows the output curves under illumination wavelengths (405 nm, 532 nm, and 785 nm, corresponding to blue, green, and red light, respectively) and two illumination intensities (0.007 mW/mm^2 and 0.700 mW/mm^2) under ferrostate1 to ferrostate4, respectively. Under light illumination, the device exhibits increasing I_{ds} with a shorter wavelength and higher light intensity, exhibiting clearly distinguishable states without any overlap under different combinations of wavelength and intensities.

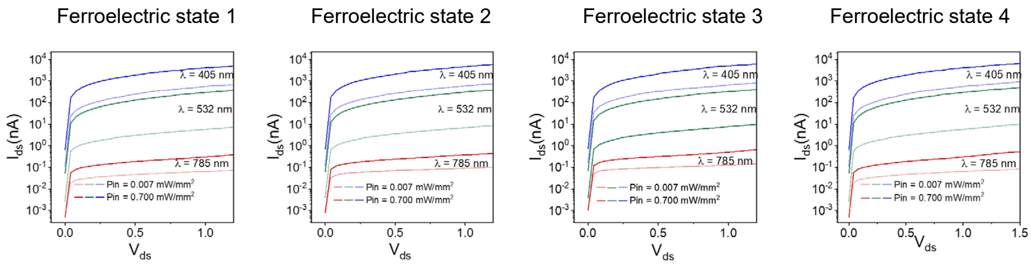


Fig. 3.6 Output curves (I_{ds}-V_{ds}) of the device in four ferrostates under two illumination intensities (0.007 mW/mm^2, and 0.700 mW/mm^2) and three wavelengths (405 nm, 532 nm, and 785 nm).

The estimated cost per device is shown in **Table 3.1**. The cost is higher than that of a CMOS pixel, which is typically on the order of 10^{-6} to 10^{-5} USD per pixel[33,34], due to the growth and integration processes of the monolayer MoS_2 channel material.

Table 3.1 Estimated cost per device for fabricating MoS_2/HZO FeFET device

Layer	Estimated fabrication cost (USD)	CMOS pixel cost (USD)
Au (Gate electrode)	2.891×10^{-5}	
HZO (Gate dielectric)	1.685×10^{-5}	
MoS_2 (Channel)	4.863×10^{-3}	
Au (Source and drain electrode)	2.891×10^{-5}	
Total cost	4.938×10^{-3}	10^{-6} - 10^{-5}

* The cost conversion is based on an exchange rate of 1 USD = 7.05 RMB.

3.2.1.2. Low-level colour clustering behaviour

The device's low-level colour clustering capability is that output photocurrent corresponds uniquely to each illumination wavelength (405 nm, 532 nm, and 785 nm) across a wide range of input intensities (0.007 mW/mm^2 to 0.700mW/mm^2) (**Fig. 3.7**).

The output does not overlap with each other when the illuminated wavelength changes. As a result, the device is able to cluster the optical inputs into three current ranges based on the maximum photocurrent in the following manner: 36–667 nA for 405 nm (blue), 1.6–29 nA for 532 nm (green), and 0.04–0.4 nA for 785 nm (red). This clear separation across the entire intensity range demonstrates wavelength selectivity and enables reliable clustering of illumination colours. This behaviour is observed across multiple ferroelectric states.

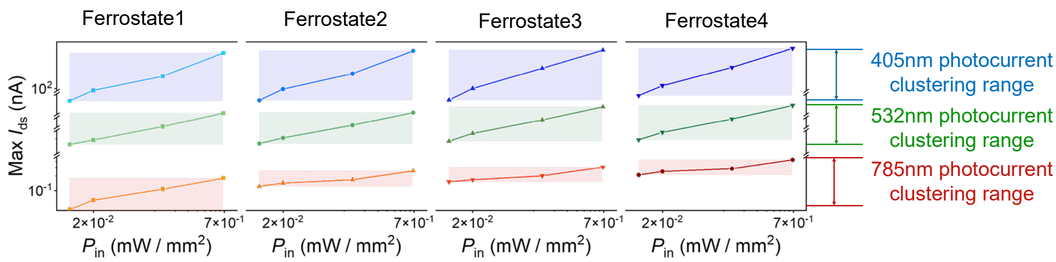


Fig. 3.7 The relationship between the maximum output photocurrent and input optical power at four distinct ferroelectric states for the three wavelengths (405 nm, 532 nm, and 785 nm), demonstrating colour clustering capability with the photocurrent range corresponding to each wavelength does not overlap with each other.

3.2.1.3. Mid-level multidimensional decoupling behaviour

As shown in **Fig. 3.8,** the device processes nine optical input patterns defined by combinations of three wavelengths (red, green, blue) and three intensity levels (P1 is 0.007 mW/mm^2, P2 is 0.119 mW/mm^2, P3 is 0.700 mW/mm^2), ranging from (R, P1) to (B, P3). Each of these combinations generates a unique and non-overlapping photocurrent response. For example, under ferrostate1, the device outputs 0.1 nA, 5 nA, and 36 nA for the input conditions (R, P2), (G, P2), and (B, P1), respectively. This one-to-one mapping between the optical input combinations and the resulting photocurrent ensures that both wavelength and intensity can be uniquely identified based on the output current. This capability is consistent across multiple ferroelectric states.

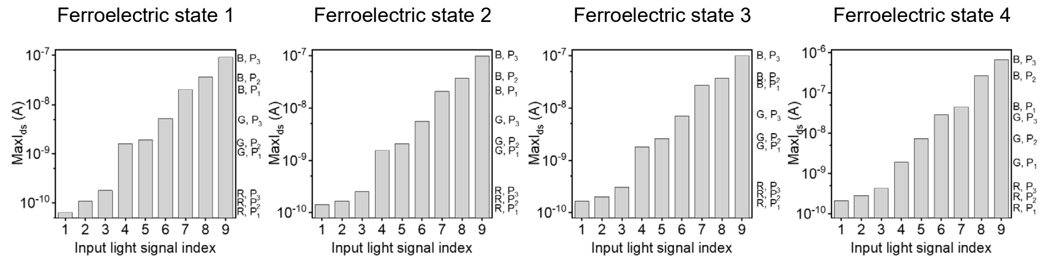


Fig. 3.8 Decoupling of intensity and wavelength features in input light pulse (Pulse width = 1 s) when memorizing different ferrostates. In all ferrostates, the MoS_2/HZO device is able to decouple the nine (wavelength, intensity) patterns of the input pulse from photocurrent level.

The mechanism of multidimensional decoupling and colour clustering capabilities of MoS_2/HZO FeFET device can be mainly attributed to the enhanced wavelength selectivity, enabled by controlling the interfacial barrier through ferroelectric modulation, as shown in **Fig. 3.9**. The wavelength selectivity of the MoS_2/HZO device is characterized by its ability to discriminate among blue (405 nm), green (532 nm), and red (785 nm) light across a range of light intensities. This mechanism is driven by two key factors: (1) the wavelength-dependent absorption coefficient of MoS_2 channel material and (2) the high MoS_2/Au interfacial barrier height modulated by HZO polarization, which influences the collection efficiency of photogenerated carriers at drain.

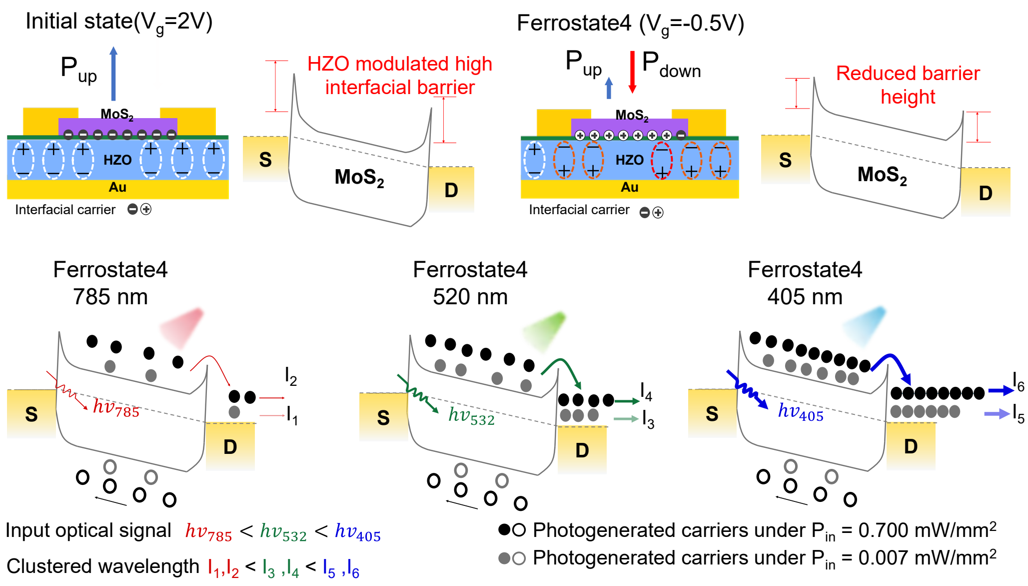


Fig. 3.9 Working mechanism of the enhanced wavelength selectivity.

The monolayer MoS_2 material exhibits a wavelength-dependent absorption profile. As shown in **Fig. 3.4**f, the absorption spectrum of MoS_2 decreases gradually from 405 nm (blue) and 532 nm (green) before cutting off near 700 nm (red). Shorter wavelengths, with higher absorption coefficient, resulting in more efficient generation of electron-hole pairs. This leads to a larger number of photogenerated carriers in the MoS_2 layer for shorter wavelengths compared to longer wavelengths, which corresponds to a larger photocurrent level and thereby contributes to the MoS_2/HZO device's wavelength selectivity.

The second factor enhancing wavelength selectivity is the high interfacial barrier height. The interfacial barrier is modulated by the ferroelectric polarization of the HZO layer for MoS_2/HZO device (**Fig. 3.9**) or by applying a gate voltage in MoS_2/SiO_2 devices. An increase in the interfacial barrier height affects the efficiency with which photogenerated carriers are transported across the MoS_2/Au interface, particularly for carriers generated by longer wavelengths. Carriers generated by longer wavelengths, which correspond to lower energy experience a larger reduction in carrier transportation efficiency across the barrier when the interfacial barrier height increases, therefore reducing the photocurrent.

To further prove this effect, we measured the MoS2/SiO2 device with SiO2 as gate dielectric as comparison. This device does not show wavelength selectivity behaviour as the output photocurrent corresponds to each illumination wavelength (405 nm, 532 nm, and 785 nm) across the same range of input intensities (0.007 mW/mm^2 to 0.700mW/mm^2) overlaps (**Fig. 3.10**).

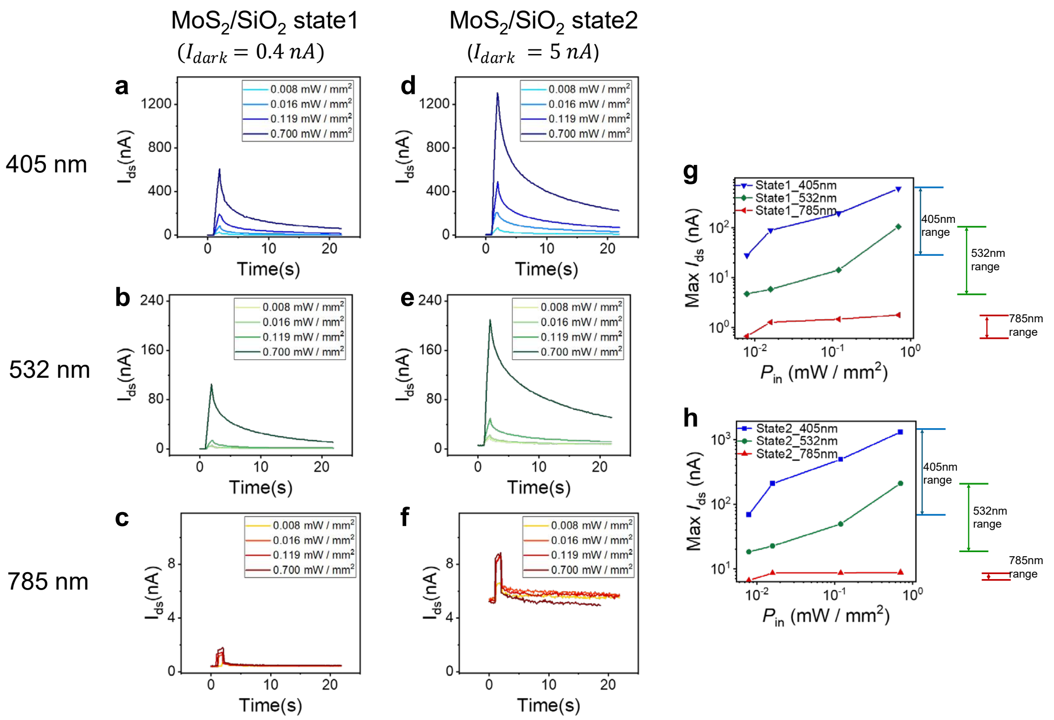


Fig. 3.10 Photoresponse behaviour of the MoS$_2$/SiO$_2$ device under two gate voltage conditions: V_g= -10 V (State1) and V_g = 0 V (State2). For State1, I_{ds} over time is shown under illumination with (a) 405 nm, (b) 532 nm, and (c) 785 nm wavelengths. For State2, I_{ds} is also recorded under (d) 405 nm, (e) 532 nm, and (f) 785 nm. Light intensity ranges from 0.007 to 0.700 mW/cm². The maximum I_{ds} over input intensity for all three wavelengths is presented for

(g) State1 and (h) State2, demonstrating no wavelength selectivity. The photocurrent responses for 405 nm, 532 nm, and 785 nm overlap in both states.

The MoS_2/HZO device, exhibiting a higher on-off ratio (**Fig. 3.11**), allows for better control over the interfacial barrier height and can reach lower channel conductance states (0.01 to 0.05 nA from ferrostate1 to ferrostate4) than the MoS_2/SiO_2 device (0.4 nA in state1, 5 nA in state2), corresponding to higher interfacial barrier heights.

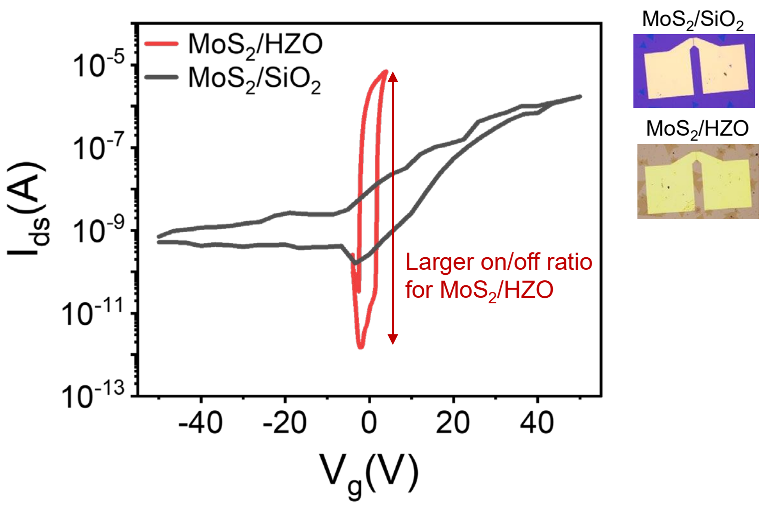


Fig. 3.11 Comparison of I_{ds}-V_g characteristic for MoS_2/HZO device and MoS_2/HZO device. The MoS_2/HZO device reaches higher on-off ratio than MoS_2/SiO_2 device. This is due to the use of HZO, as a high-k gate dielectric, offers better gate control and lower leakage current over the channel compared to SiO_2 used in MoS_2/SiO_2 device[71-73].

Fig. 3.12 illustrates the impact of interfacial barrier height on photoresponsivity. As the dark channel conductance decreases from MoS_2/SiO_2 device state 2 (5 nA) to MoS_2/HZO device ferrostate1 (0.01 nA), corresponding to an increasing barrier height, the responsivity to blue light shows a small change (approximately 4 times) between the highest and lowest MoS_2 channel conductance state. This indicates that shorter wavelengths are less affected by barrier height changes. In contrast, the responsivity differences for green and red light are more significant, especially for red light, where the detectivity drops by approximately 70 times as the barrier height increases from lowest to highest. This demonstrates that longer wavelengths are more significantly affected by barrier height changes.

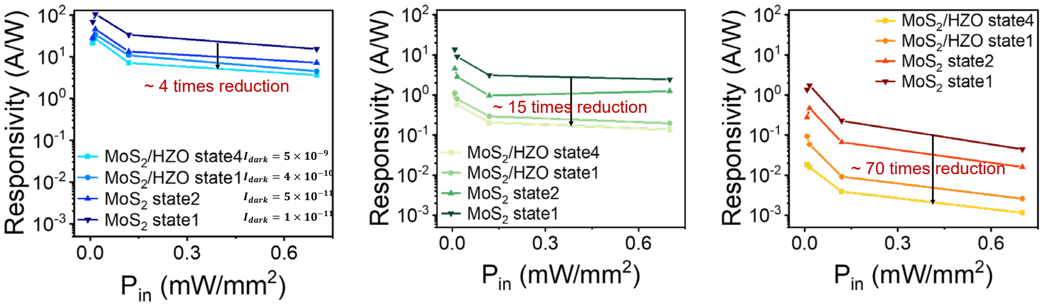


Fig. 3.12 Responsivity comparison of MoS_2/HZO device and MoS_2/SiO_2 device under different channel conductance states across three wavelengths: blue, green, and red. Each plot features data from two devices: MoS_2/SiO_2 with higher dark current levels (0.4 nA at state1, 5 nA at state2), and MoS_2/HZO with lower dark current levels (0.01 nA at ferrostate1, 0.05 nA at ferrostate4), representing four distinct channel conductance states.

In summary, the combined effect of the wavelength-dependent absorption coefficient and the modulation of interfacial barrier height enhances the photocurrent contrast between wavelengths, enabling wavelength selectivity. The absorption coefficient of MoS_2, along with the difference in photon energy across different incident wavelengths, creates an initial difference in carrier generation between shorter and longer wavelengths. More importantly. the high interfacial barrier height modulated by HZO further enhances the difference in carrier collection. As a result, the MoS_2/HZO device demonstrates enhanced wavelength selectivity, while MoS_2/SiO_2 device lacks this capability.

3.2.1.4. High-level non-volatile polarization enhanced multi-photoresponsivity states behaviour

Fig. 3.13a illustrates the non-volatile and multilevel ferroelectric states (state1 to state4) during the long-term potentiation (LTP) and long-term depression (LTD) processes, with programming gate voltages ranging from -0.5 V to -1.5 V and 0.5 V to 1.5 V.

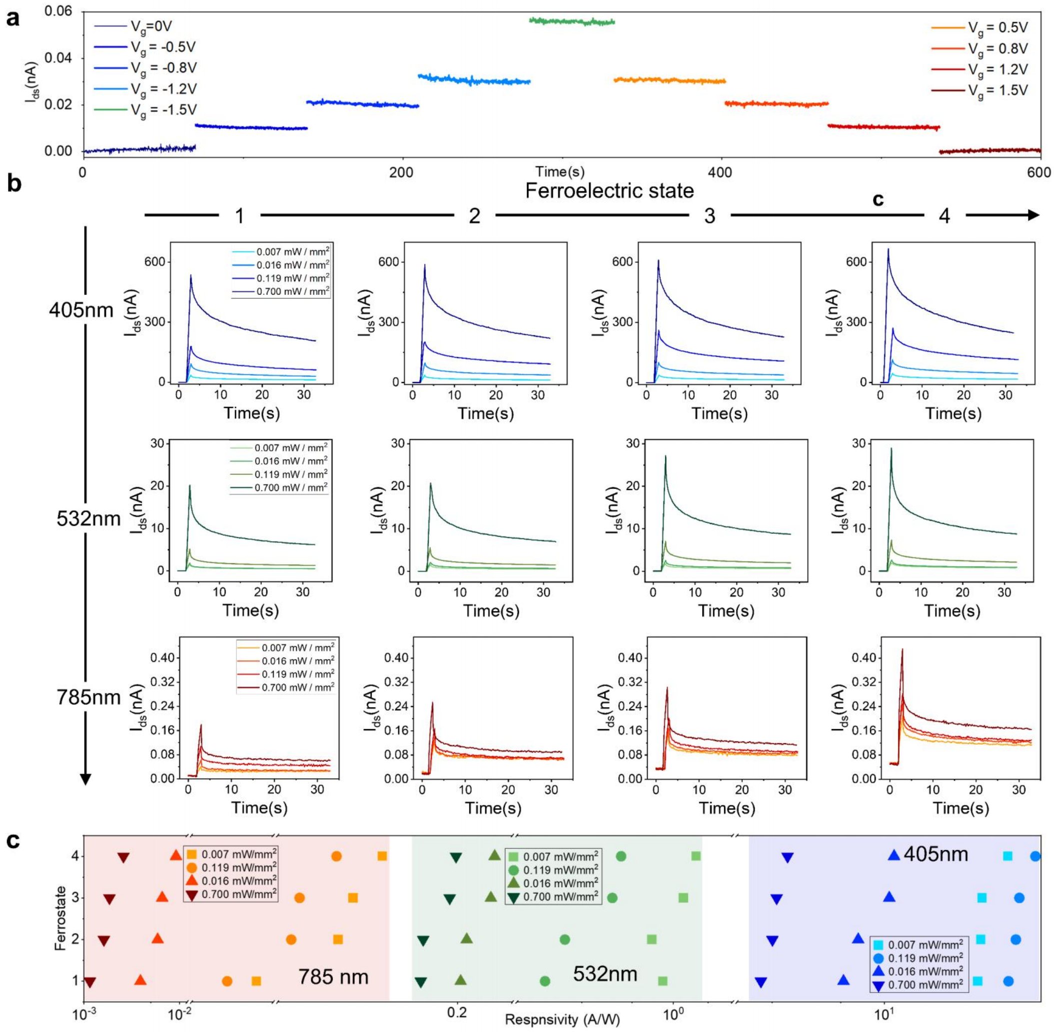


Fig. 3.13 High-level non-volatile polarization enhanced multi-photoresponsivity states behaviour of MoS_2/HZO FeFET device (a) Long-term potentiation and long-term depression

(LTP/LTD) characteristics. (b) The I_{ds}-T characteristics of the output photocurrent generated by light pulses with varying intensities, ranging from 0.007 mW/mm^2 to 0.700mW/mm^2, and wavelengths of 405 nm, 532 nm, and 785 nm (corresponding to each row, respectively) across different ferroelectric polarization states (ferrostate1 to ferrostate4, represented in the 1st to 4th columns). (c) 48 distinct photoresponsivity states, each depends on wavelength, intensity, and memorized ferrostate.

Fig. 3.13b presents the I_{ds}-T characteristics of photocurrent generated by the light pulses with varying intensities (from 0.007 mW/mm^2 to 0.700mW/mm^2) and wavelengths (405, 532, and 785nm in each row, respectively) at the four different ferroelectric states (ferrostate1 to ferrostate4 from 1st to 4th column). The I_{ds}-T curves demonstrate that the maximum photocurrent amplitude is strongly dependent on the ferroelectric state of the HZO layer. Under identical optical input conditions, the maximum photocurrent amplitude increases progressively from ferrostate1 to ferrostate4, indicating that the ferroelectric polarization effectively enhances photocarrier separation and transport, resulting in an increase in photoresponsivity. In addition to the ferroelectric state, the photocurrent amplitude is also influenced by the input light wavelength and intensity. At a fixed ferroelectric state and wavelength, higher light intensity results in a higher maximum photocurrent amplitude, while at a fixed ferroelectric state and intensity, shorter wavelengths lead to a higher maximum photocurrent amplitude due to their higher photon energy, corresponding to higher responsivity.

Fig. 3.13c presents 48 distinct photoresponsivity states, each tuned by combinations of wavelengths (405, 532, and 785 nm), intensities (0.007 mW/mm^2, 0.016 mW/mm^2, 0.119 mW/mm^2, 0.700 mW/mm^2), and ferroelectric state (ferrostate1 to ferrostate4). The device exhibits the highest photoresponsivity at ferrostate4 and the lowest at ferrostate1. Moreover, the photoresponsivity decreases with increasing wavelength, with 405 nm producing the strongest photoresponsivity and 785 nm the lowest. The output photocurrent due to the input light intensity and wavelength clearly depends on the ferroelectric state of the HZO. At a fixed wavelength and ferroelectric state, higher light intensity results in a higher photocurrent, while at a fixed ferroelectric state and intensity, shorter wavelengths lead to a higher photocurrent. Moreover, the photocurrent increases as the ferroelectric state increases under the same lighting conditions.

Fig. 3.14 and **Fig. 3.15** further investigate the polarization-modulated photoresponsivity behaviour under different pulse widths and pulse numbers.

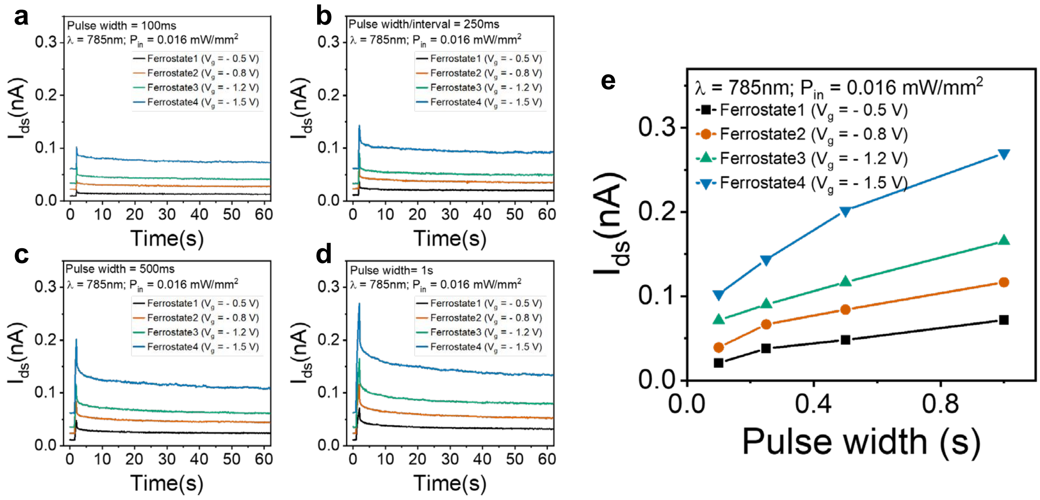


Fig. 3.14 Polarization-modulated photoresponsivity under varying pulse widths with 785 nm light illumination (pulse number = 1, intensity = 0.016 mW/mm^2). I_{ds}-T characteristics of the MoS_2/HZO device in response to light pulses of different widths: (a) 0.1 s, (b) 0.25 s, (c)

0.5 s, and (d) 1 s. (e) The maximum source-drain current I_{ds} as a function of input light pulse width, demonstrating the dependence of the photocurrent response on the duration of the light pulse.

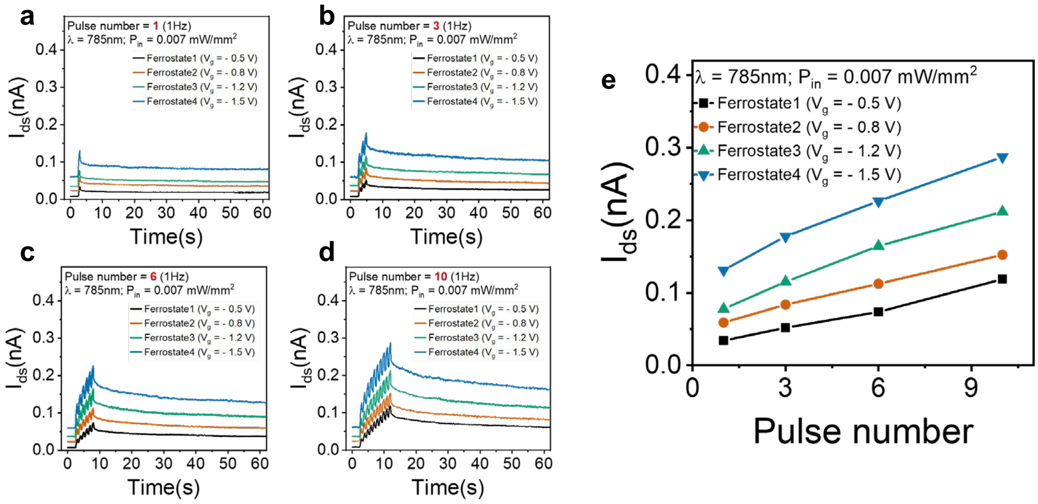


Fig. 3.15 Polarization-modulated photoresponsivity under varying numbers of light pulses at 785 nm illumination (pulse width = 0.5 s, pulse interval = 0.5 s, intensity = 0.007 mW/mm^2). I_{ds}-T characteristics of the MoS_2/HZO device in response to 785 nm light pulses of varying numbers: (a) 1, (b) 3, (c) 6, and (d) 10 pulses. (e) The maximum source-drain current I_{ds} as a function of input light pulse number.

The photoresponsivity of MoS_2/HZO FeFET device can be widely tuned by adjusting ferroelectric polarization states, which mainly affects the MoS_2/Au interfacial barrier height (**Fig. 3.16**). As the ferroelectric state in the device is changed from state1 to state4 by applying more negative gate voltages, the proportion of downward ferroelectric domains in the HZO increases, which reduces the MoS_2/Au interfacial electron barrier height at the channel/drain electrode and channel/source electrode interface. The lower barrier height facilitates more effective separation of electron-hole pairs, therefore improving photoresponsivity and resulting in a larger photocurrent amplitude.

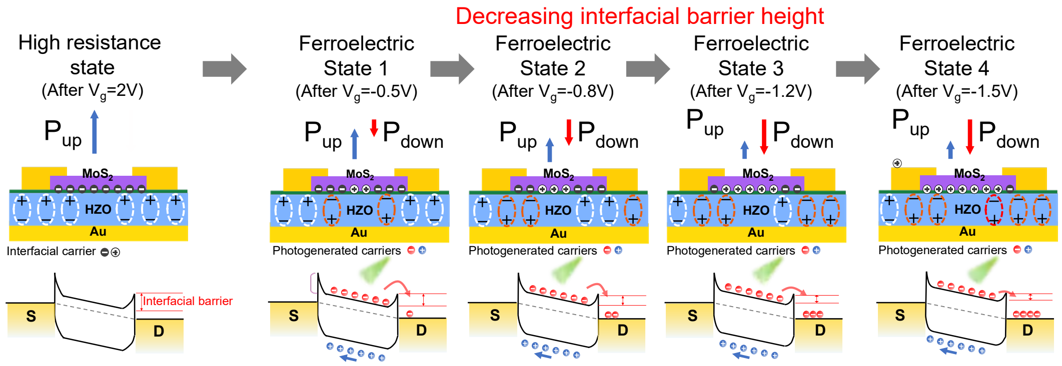


Fig. 3.16 The mechanism of non-volatile polarization enhanced multi-photoresponsivity states behaviour.

3.2.2. Multidimensional and multilevel processing based on MoS_2/HZO FeFET array-based vision system

An efficient multidimensional visual processing system is constructed with a single FeFET array integrated with a co-designed HVBF algorithm (**Fig. 3.17**). The HVBF algorithm leverages the device's inherent computational capabilities for the low- and mid-level processing, while the high-level processing is achieved through a combination of device behaviour and a correlational filtering algorithm, progressively refining data to enhance both accuracy and versatility in visual processing.

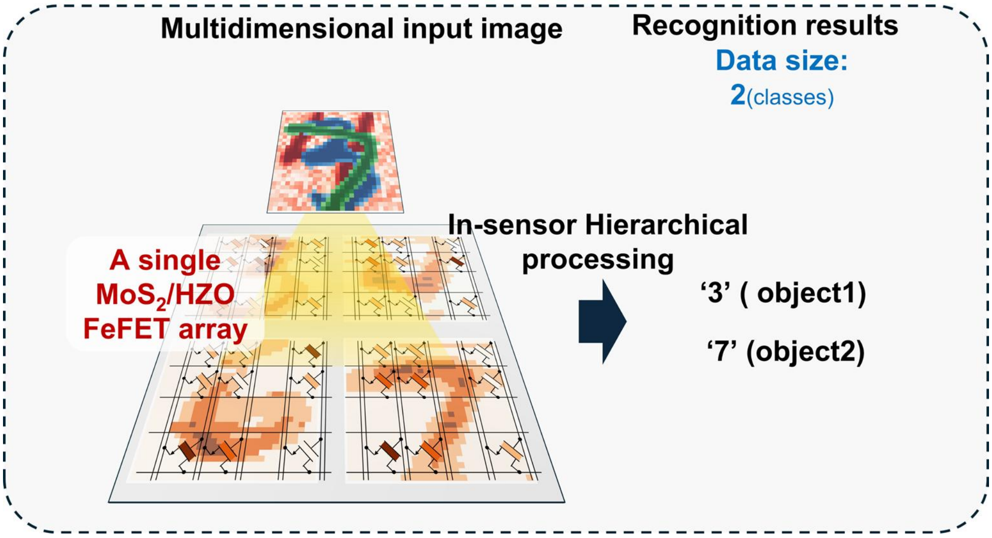


Fig. 3.17 Schematic illustration of the multidimensional and hierarchical system consists of a single MoS_2/HZO FeFET array and a co-designed HVBF algorithm.

At the low-level, the system employs its colour clustering capability to distinguish the colour of each pixel within the input multidimensional input image (**Fig. 3.18**). The colour of each pixel is identified based on the corresponding output photocurrent range. Specifically, the pixel is classified as blue, green, and red when the output photocurrent is in the range of 36-667 nA, 1.6-29 nA, and 0.04-0.4 nA, respectively. The decoupled colour information is represented as a colour map ⟨λ⟩.

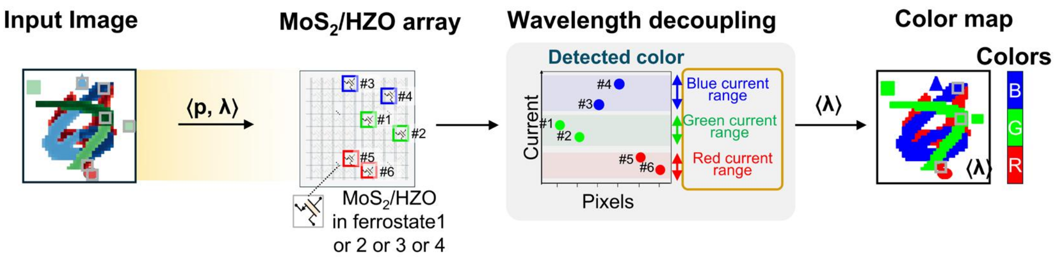


Fig. 3.18 Low-level colour clustering: The MoS_2/HZO array discriminates the colour of each pixel based on its colour clustering capability.

At the mid-level, the system performs surface-level filtering, segmenting the two overlapping digits by decoupling intensity and colour information based on the FeFET device's inherent multidimensional decoupling capability. The input of this stage is the decoupled colour map from the low-level stage and the input multidimensional image. The purpose of this stage is to accurately separate the spatially overlapping digits, which share a common region but differ in spectral characteristics. The multidimensional decoupling capability of the device enables a one-to-one photocurrent-intensity relationship for each wavelength. Since the wavelength information of each pixel has already been obtained from the low-level filtering, this relationship further allows for accurate determination of optical intensity for each pixel. This decoupling results in an intensity map ⟨p⟩ and a colour map ⟨λ⟩. Based on these maps, the image is segmented into two overlapping digits and RGB noise (**Fig. 3.19**). The RGB noise is then filtered.

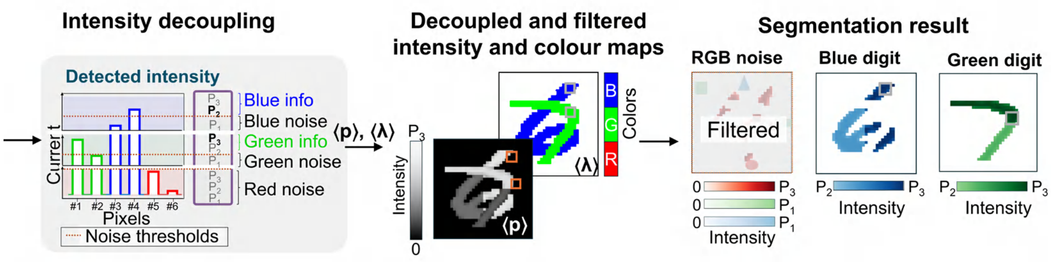


Fig. 3.19 Mid-level image segmentation: MoS_2/HZO array decouples the wavelength and intensity at each pixel using its multidimensional decoupling capability, enabling segmentation into RGB noise, blue digits, and green digits. The RGB noise is subsequently filtered out.

At the high-level, the system performs schematic-level filtering, simultaneously recognizes the two segmented objects by comparing their similarity to class feature maps memorized in ferroelectric states. The input of this stage is the segmented two images from mid-level processing. This high-level filtering processing involves two key stages: filtering template learning and similarity-based recognition (**Fig. 3.20**).

In the filtering template learning stage, feature maps for digit classes ('3', '4', '6', and '7') are first memorized in the ferroelectric states in four sections of the MoS_2/HZO array, by utilizing its non-volatile polarization enhanced multi-photoresponsivity states characteristic. These feature maps, extracted from the MNIST dataset, are initially generated by averaging training images within each class. They are then refined by removing noise and shared features to highlight the unique characteristics of each digit class. Once refined, the feature maps are quantized into four discrete grayscale levels (0, 1, 2, and 3) and stored in four separate sections of the FeFET array as high-level filtering templates. The corresponding ferroelectric states encode these values, where state1, state2, state3, and state4 represent grayscale levels 0, 1, 2, and 3, respectively.

Once the feature maps are stored, each array section serves as a visual filter. The system recognizes the input digits by computing the similarity between the segmented input and each class template via summed photocurrent output as:

$$I_{array} = \sum_{pixel} I_{pixel} = \sum_{pixel} R_{pixel} P_{pixel} = \vec{R} \cdot \overrightarrow{P_{in}}$$

In this equation, I_{pixel} represents the photocurrent of each FeFET device, R_{pixel} is the photoresponsivity modulated by the ferroelectric state, and P_{pixel} denote the optical power input at each pixel. This process is equivalent to performing a dot product operation between the class template $\vec{R}$ and the input object $\overrightarrow{P_{in}}$, where the highest photocurrent indicating the strongest correlation[74] and therefore the recognized class.

As a result, the two images from the mid-level stage are further reduced to a one-dimensional semantic vector of size 2, each element representing the recognized class of a segmented digit. This completes the system's hierarchical filtering processing, from a noisy N × H × W input to a compact and interpretable 2-class output.

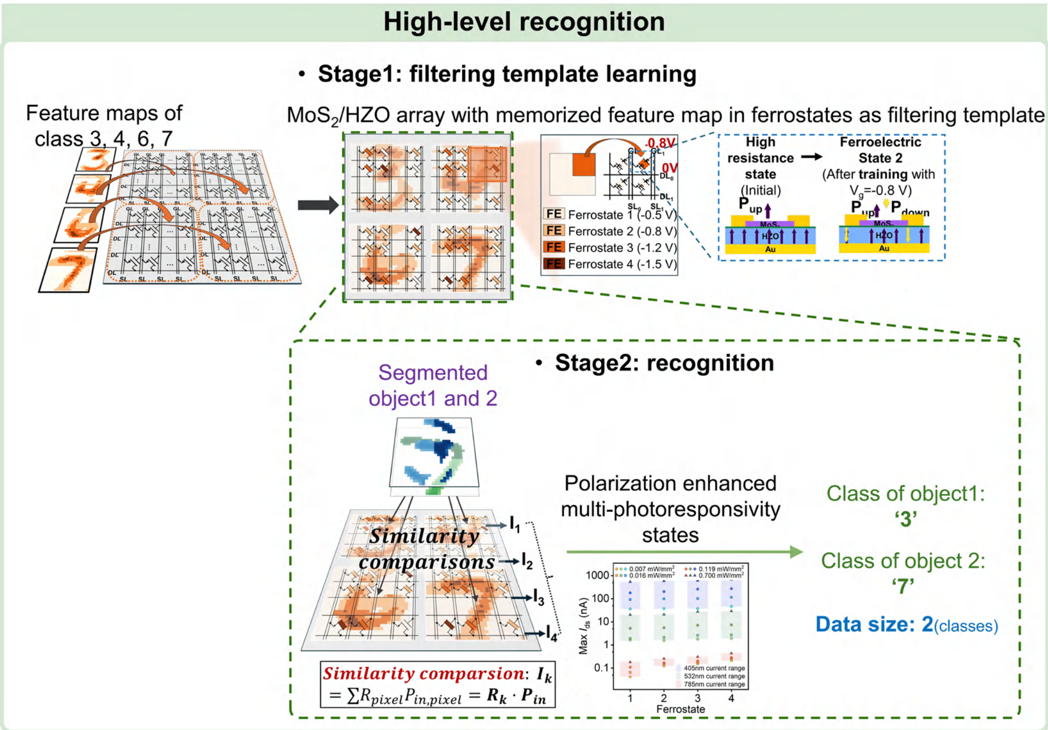


Fig. 3.20 High-level schematic feature filtering. This process consists of two stages. In the first stage, the MoS_2/HZO array memorizes four filtering templates corresponding to digit classes ('3', '4', '6', and '7') in ferrostates. In the second stage, the MoS_2/HZO array performs similarity comparison between the segmented image and the memorized filtering templates using its non-volatile polarization-enhanced multi-photoresponsivity states capability. This enables the recognition of both overlapping digits, resulting in a one-dimensional semantic representation with a final data size of 2 (two recognized classes).

To demonstrate the effectiveness of the MoS_2/HZO FeFET array-based multidimensional and multilevel vision system in processing noisy, multidimensional, and complex visual information, we evaluate its performance on two challenging tasks: recognizing noisy, coloured handwritten digits and identifying overlapping digits within a noisy, mixed-coloured image. The system achieves high accuracy in both tasks by leveraging its multidimensional sensing and decoupling capability, which allows it to extract sufficient information from the scene, as well as its hierarchical filtering capability, which progressively refines visual data to enhance robustness and precision.

Fig. 3.21 illustrates the task of recognizing noisy, coloured handwritten digits, where variations in both colour and intensity add complexity to accurate recognition. The MoS_2/HZO visual filter array addresses this challenge through a multilevel filtering process. The input consists of a single digit (selected from 3, 4, 6, or 7) in one of three colours (blue, green, or red), forming 12 distinct input classes. The MoS_2/HZO visual filter array can effectively capture colour information from the input by utilizing its low-level colour clustering capability, outputting distinct current maps for digits with different colours. This enables accurate low-level colour recognition while retaining intensity information for subsequent high-level processing. The digit class is determined based on the maximum summed output currents using the similarity-based recognition method.

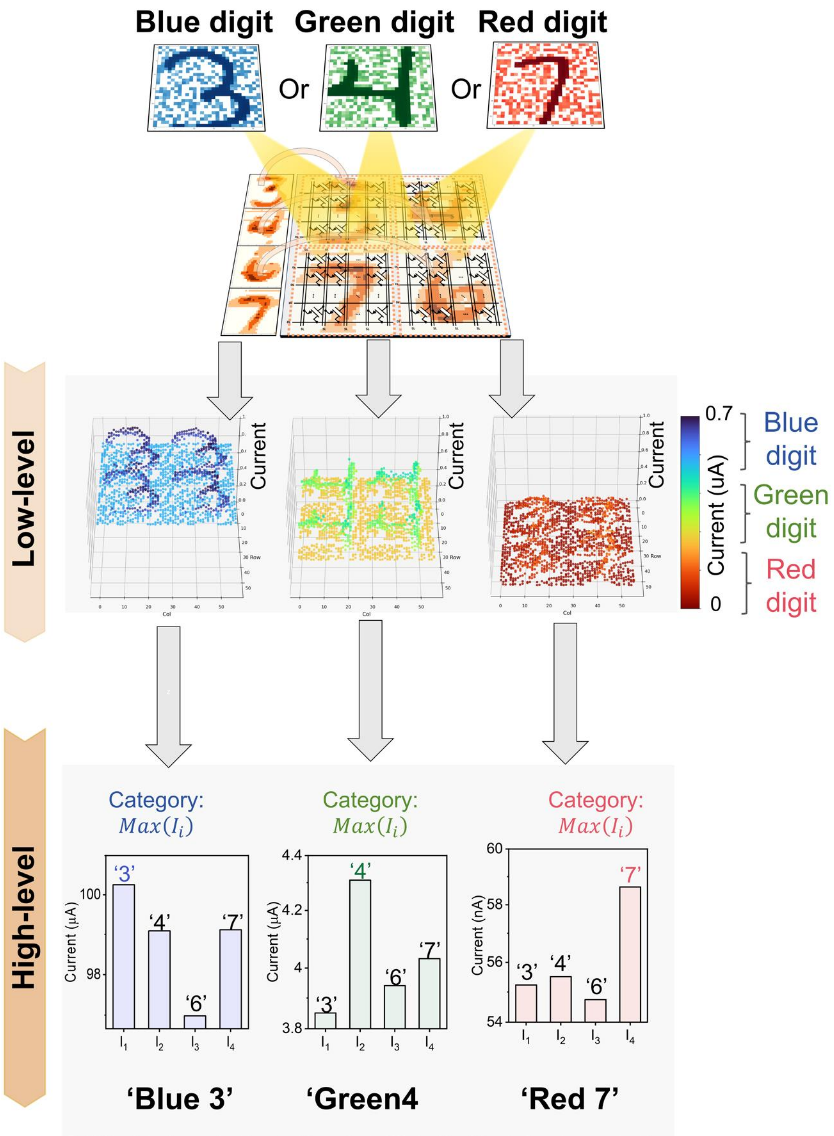


Fig. 3.21 Workflow for recognition of noisy, coloured handwritten digits, including low-level colour clustering and high-level recognition. The low-level output is shown as a 3D current map, where each point corresponds to a pixel's current value, with colour and z-axis height indicating the current level. In high-level processing, the current maps from four distinct array sections are summed separately, and the highest current value determines the recognition result.

The ColoredMNIST dataset used for evaluation is shown in **Fig. 3.22**a. The MoS_2/HZO visual filter array achieves a high accuracy of 93.3% (**Fig. 3.22**b). In comparison, a traditional single-dimensional (capturing only grayscale information) and single-level (only high-level processing) system based on a colour-blind CMOS sensor and CNN achieves only 33.3% accuracy (**Fig. 3.23**). This 60% performance gap highlights the MoS_2/HZO visual filter array's effectiveness in accurate in-sensor recognition and multidimensional visual processing.

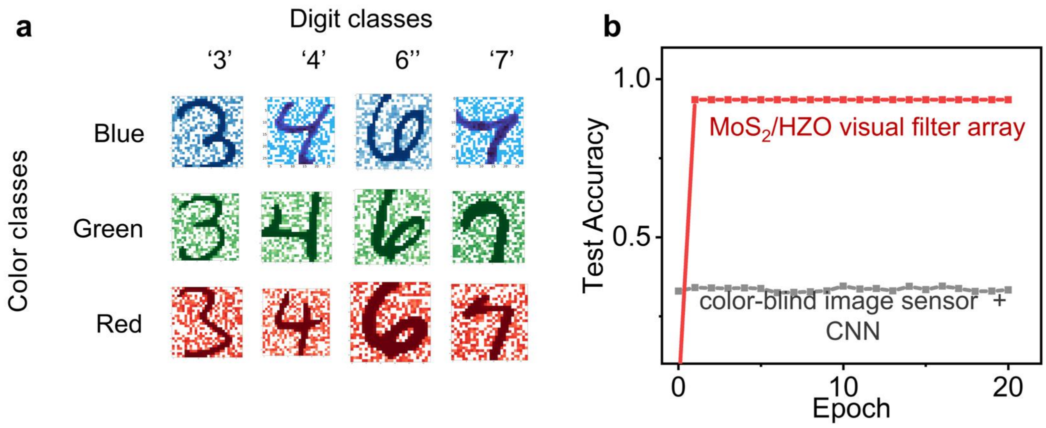


Fig. 3.22 Input dataset and output accuracy in the task of recognizing coloured digits (a) Custom ColoredMNIST dataset. (b) Accuracy comparison between the multidimensional, multilevel MoS_2/HZO visual filter array (93.3%) and a conventional single-dimensional, single-level system (33.3%) on the ColoredMNIST dataset.

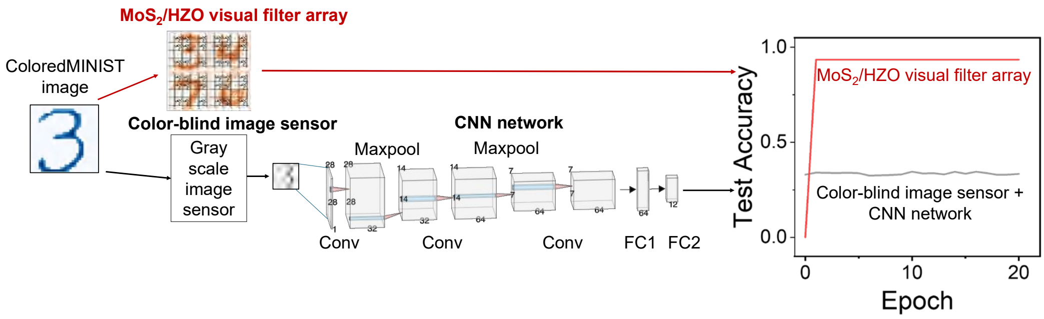


Fig. 3.23 Workflow of the MoS_2/HZO visual filter array-based multidimensional vision system and the traditional vision system. The MoS_2/HZO visual filter array is a multidimensional, multilevel system capable of performing both low-level colour clustering and high-level recognition. When an image is input into the array, it accurately recognizes 12

classes of images, achieving a recognition accuracy of 93.3%. In contrast, the traditional vision system relies on a colour-blind CMOS sensor and a deep CNN with 8 layers. Due to its single-dimensional sensing and single-level processing capability, the traditional system achieves a much lower accuracy of 33%.

Fig. 3.24 presents a more complex challenge of recognizing overlapping digits in a noisy, mixed-colour image, where the system must not only filter out noise interference but also accurately separate the two overlapping objects and correctly recognize each digit class. The MoS_2/HZO visual filter array effectively handles this task through its multidimensional and hierarchical low-, mid-, and high-level filtering capability. The input image consists of four components: a main digit in green, an overlapping digit in blue, a noisy red digit, and a noisy red background. At the low-level, the MoS_2/HZO visual filter array performs noise filtering by leveraging the colour clustering capability of the MoS_2/HZO device to identify all red pixels. This effectively removes both the red noisy digit and the red background, ensuring that only relevant information is retained for further processing. At the mid-level, the array performs surface-level filtering, segmenting the two digits by utilizing its multidimensional decoupling capability. This results in two distinct current maps, each corresponding to one of the overlapping digits, ensuring clear object separation. These segmented digits are then processed at the high-level, where the similarity-based recognition method is applied to determine the class of each digit.

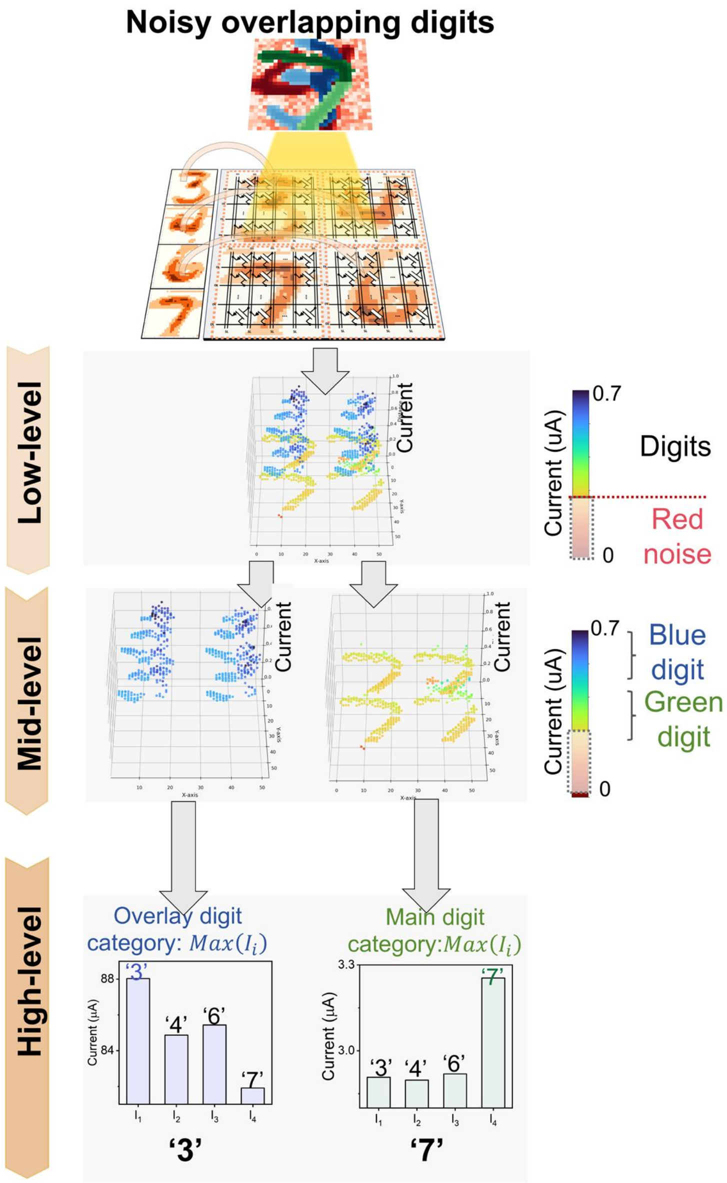


Fig. 3.24 Workflow for recognition of two overlapping digits in a noisy, mixed-colour image, including low-level noise filtering, mid-level image segmentation, and high-level recognition. The low-level output is shown as a 3D current map, where each point represents the current value of a pixel, with colour and z-axis height indicating the current level. At the mid-level, the array's multidimensional decoupling capability segments the current map into two separate 3D maps corresponding to the green main digit and the blue overlapping digit. In the high-level processing stage, the summed current maps for each digit produce two sets of output currents, with the highest current level in each set indicating the recognition result.

On the custom MultiMNIST dataset (**Fig. 3.25**a), the MoS_2/HZO visual filter array achieves an accuracy of 93.8% (**Fig. 3.25**b**)**, which is 10% higher than that of a traditional single-dimensional, single-level system based on a colour-blind CMOS sensor and CNN (**Fig. 3.26**). This accuracy improvement demonstrates the MoS_2/HZO vision system's ability to handle complex, mixed-dimensional recognition tasks, where conventional approaches might struggle with the object occlusion and spectral noise interference.

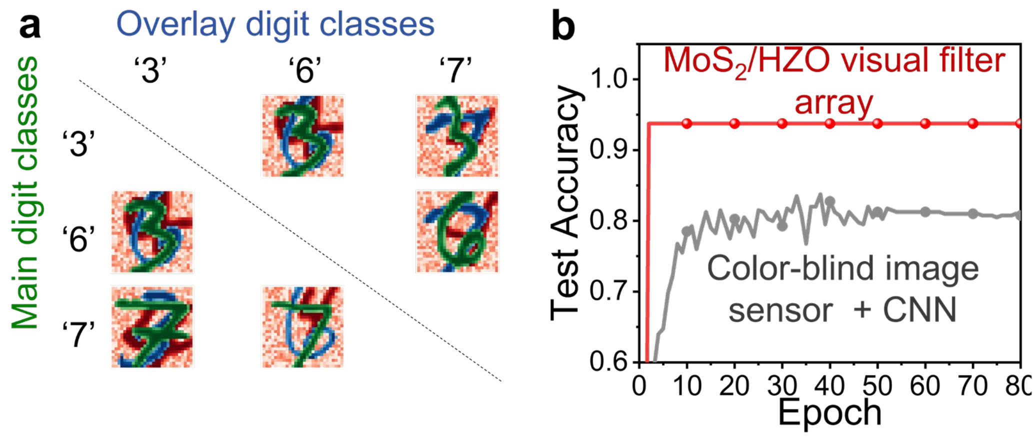


Fig. 3.25 Input dataset and output accuracy in the task of recognizing overlapping digits. (a) Custom MultiMNIST dataset. (b) Accuracy comparison between the multidimensional, multilevel MoS_2/HZO visual filter array (93.8%) and a conventional single-dimensional, single-level system (83.8%) on the MultiMNIST dataset.

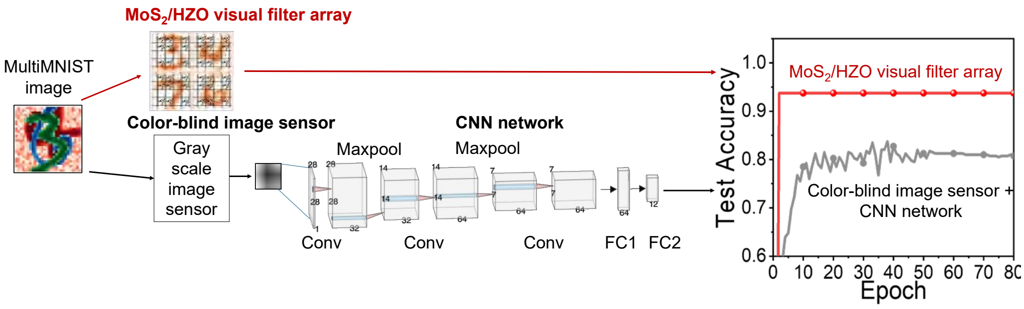


Fig. 3.26 Comparison of the MoS_2/HZO visual filter array with a traditional single-

dimensional, single-level vision system for recognizing overlapping digits in a noisy, mixed-colour image from the MultiMNIST dataset. The mixed-dimensional and hierarchical MoS_2/HZO visual filter array achieves 93.8% accuracy, outperforming the conventional system, which relies on a colour-blind CMOS sensor and a deep convolutional neural network (CNN) and achieves only 83.0% accuracy.

Chapter 4 RGB-Depth aware intelligent vision system based on RGB-NIR decoupled bipolar MoS_2/In_2Se_3 device

4.1. Introduction

Augmented reality (AR) refers to the technology that overlays digital information onto the real-world environment in real time, which helps the user to better understand and connect to their surroundings [75]. For example, AR can display virtual texts near the corresponding object or action to annotate their names and types. This augmentation helps users better understand and interact with the real world. As a result, AR has found applications in diverse areas such as navigation, industrial management, design, and education, where it enhances user experience, improves the decision-making speed, and increases learning quality.

The quality of an AR system depends on the alignment accuracy between virtual augmentations and their real-world counterparts[76]. This alignment includes two aspects: the schematic alignment and spatial alignment. First, the virtual information must correctly reflect the category or function[75] of the associated object (such as labelling a product during manufacturing[77], displaying a landmark name during navigation[78], or identifying the type of a medical instrument in a surgical setting[79]). This task is particularly challenging in real-world vision scenarios where static and dynamic objects both exist. To achieve the accurate recognition, the system should first capture high-quality two-dimensional (2D) Red, Green, and Blue (RGB) images that preserve the necessary visual appearance for static object recognition. In addition, the system should also extract the 3D position information for dynamic objects and gesture recognition. Infrared (IR) depth sensing methods, such as structured light[80], time-of-flight (ToF)[81], or near-infrared (NIR) marker[82,83], are commonly adopted for obtaining 3D position information since they are robust to variations in visible lighting conditions, therefore enabling accurate depth sensing even in low-light environments[84,85]. Besides, to achieve the spatial alignment accuracy, the 3D location information is also critical in determining the real-time location of the dynamic gesture to enable precise spatial positioning of the virtual arguments. As a result, Red, Green, and Blue-Depth (RGB-D) cameras are widely adopted in AR systems for capturing both 2D RGB images and 3D position data simultaneously[86,87].

However, conventional CMOS RGB-D cameras face certain limitations that hinder their performance in AR applications. These CMOS RGB-D camera typically relies on two separate imaging units, including one RGB camera for capturing 2D RGB images for visual appearance in the visible spectrum, and one NIR camera for acquiring 3D depth images for 3D positional information[88-90]. Since the two cameras are physically separated and usually have different intrinsic parameters (such as focal length, focus distance, and resolution), accurate calibration is required to align the captured RGB and depth images. This alignment process typically relies on classical camera calibration techniques[91] or deep learning algorithms[92], which can be complex, time-consuming, and inaccurate. Such dual-camera-based RGB-D camera introduces two major drawbacks for AR system. First, the use of two sensing units increases power and area consumption, which brings problems in resource-constrained edge scenarios such as AR navigation on mobile devices. Second, the misalignment between RGB and depth images due to imperfect calibration can result in the same object appearing at different positions in images captured by different cameras. This difference between object visual appearance and 3D location data reduces the schematic accuracy and the spatial position accuracy of virtual argumentation, since the virtual arguments might be placed on the wrong objects, therefore degrading the accuracy of the AR system. For instance, incorrect placement of virtual directional arrows in AR navigation can mislead users, and in industrial assembly, inaccurate labelling of part names can lead to assembly errors.

In this work, we report an optoelectronic RGB-NIR decoupled bipolar MoS_2/In_2Se_3 device for constructing an RGB-D camera based on a single MoS_2/In_2Se_3 array, tailored for AR applications. This device features an asymmetric structure that allows RGB-photogenerated carriers to flow to the source electrode through the visible-light-induced Schottky barrier-lowering effect, resulting in a negative photoresponse, while NIR-photogenerated carriers flow to the drain electrode following the direction of applied positive bias, generating a positive photoresponse. This bipolar photoresponse (positive for NIR and negative for RGB) effectively decouples RGB and NIR optical signals. Consequently, the MoS_2/In_2Se_3 device array-based RGB-D camera can simultaneously sense and decouple both RGB light and NIR light with a single device array. This enables simultaneous acquisition of 2D visual appearance and 3D location data, which are output in the form of negative current and positive current maps, respectively. Additionally, since the 2D RGB image and 3D location are obtained by the same imaging unit, they are spatially aligned on x- and y-axis, eliminating the need for complex alignment procedures required in traditional dual-camera systems. The resulting RGB-D camera is integrated with two neural networks to construct an RGB-Depth aware intelligent vision system. A SqueezeNet convolutional neural network (CNN) is used to process the 2D visual appearance information, and a long short-term memory (LSTM) network is used to analyse the 3D location information. This system can be used for AR applications during interactive events where two hands are interacting with one object. The system can recognize the hand gesture and the static object with 91.8% accuracy, and accurately place the corresponding category labels at their respective 3D locations, ensuring both schematic and positional accuracy.

4.2. Results

4.2.1. MoS_2/In_2Se_3 heterostructure device with decoupled bipolar responses to RGB and NIR light

4.2.1.1. Device structure and characterization

The device architecture of the designed RGB-NIR decoupled bipolar MoS_2/In_2Se_3 device with decoupled response for RGB and NIR light is illustrated in **Fig. 4.1**a. This device features an asymmetric structure where the source electrode is in contact with the MoS_2 and the drain electrode is in contact with the In_2Se_3. **Fig. 4.1**b presents the Raman spectrum of the In_2Se_3 exhibiting prominent peaks at ~88, 102, 179, and 196 cm^{-1}. These peaks correspond to the vibrational modes characteristic of the α-phase In_2Se_3, consistent with previously reported values[93-95]. **Fig. 4.1**c presents the Raman spectrum of the MoS_2. The observed frequency difference between the E_{2g}^{1} and A_{1g} modes is approximately 22 cm^{-1}, consistent with the characteristic value for bilayer MoS_2, as reported in previous studies[96,97]. **Fig. 4.1**d presents the atomic force microscopy (AFM) image of the MoS_2/In_2Se_3 heterostructure device, with the inset displaying the corresponding height profile measured along the orange dotted line. **Fig. 4.1**e and f present the KPFM image of the MoS_2/In_2Se_3 heterostructure device and the corresponding surface potential profile along the white dotted line. A surface potential difference of approximately 0.125 eV is observed between the MoS_2 and the source Au electrode, while a larger difference of approximately 0.25 eV appears between the In_2Se_3 and the drain Au electrode.

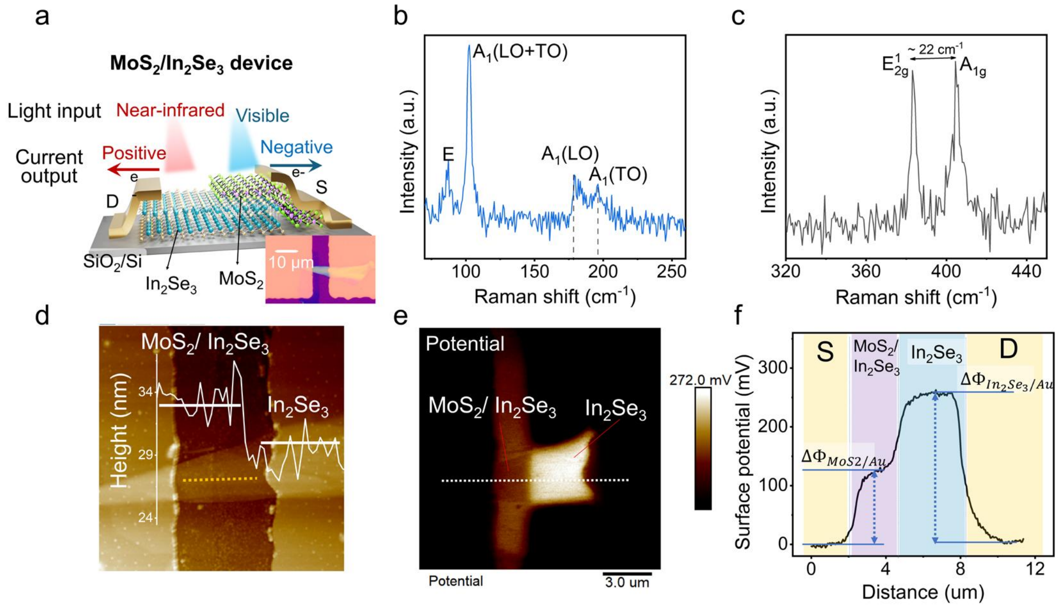


Fig. 4.1 Device structures and basic characterizations. (a) Schematic illustration of the device structure and the optical image of the MoS_2/In_2Se_3 heterostructure device. (b) Raman spectra of α-In_2Se_3, demonstrating the four E, A_1(LO+TO) , A_1(LO) and A_1(TO) phonon modes. (c) Raman spectrum of the MoS_2 flake showing a ~22 cm^{-1} separation between the E_{2g}^{1} and A_{1g} modes, consistent with a bilayer thickness. (d) AFM image of the MoS_2/In_2Se_3 device, with the

inset showing the height profile along the orange dotted line. (e) KPFM image of the MoS_2/In_2Se_3 device. (f) KPFM measurement along the white dotted line in (e), showing the surface potential difference across the source electrode, the MoS_2/In_2Se_3 heterostructure, In_2Se_3 and the drain electrode.

The estimated cost per device is shown in **Table 4.1**. The cost is higher than that of a CMOS pixel, which is typically on the order of 10^{-6} to 10^{-5} USD per pixel[33,34]. The higher cost is mainly due to the low material utilization efficiency in the process of manually aligning and transferring two distinct 2D materials.

Table 4.1 Estimated cost per device for fabricating MoS_2/In_2Se_3 heterostructure device

Layer	**Estimated fabrication cost (USD)**	**CMOS pixel cost (USD)**
In_2Se_3	2.472×10^{-7}	
MoS_2	2.553×10^{-1}	
Au	8.326×10^{-4}	
Total cost	2.562×10^{-1}	10^{-6} - 10^{-5}

* The cost conversion is based on an exchange rate of 1 USD = 7.05 RMB.

Fig. 4.2 depicts the *I-V* characteristics of the device under 450nm illumination conditions. The blue lines depict the output characteristics under 450 nm illumination at different light intensities (0 mW/mm^2 to 6.0 mW/mm^2). The output current increases with the increasing illumination intensity. Under 450 nm illumination, the device shows a positive open circuit voltage and negative short circuit current, indicating a photovoltaic behaviour with negative photo-induced electric field pointing from source to drain.

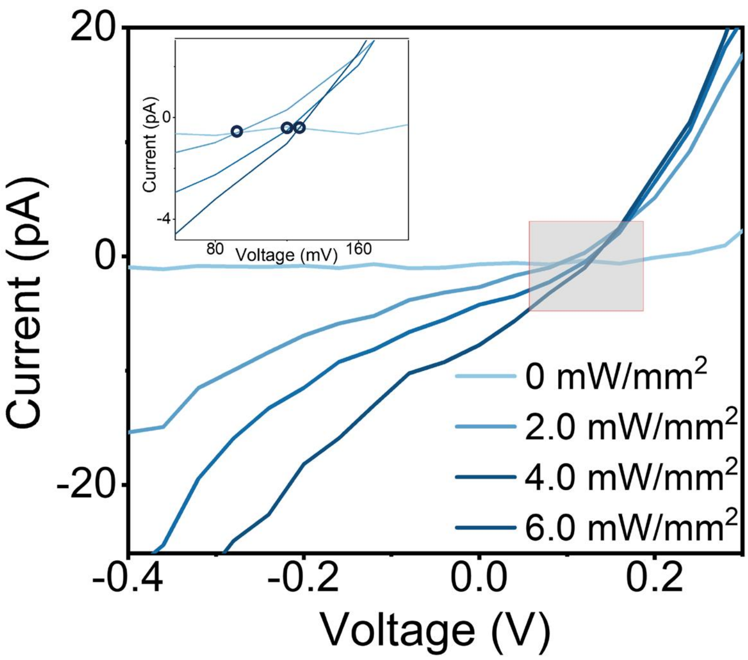


Fig. 4.2 The *I-V* characteristic under 450 nm illumination at different illumination intensities ranging from 0 mW/mm^2 (dark) to 6.0 mW/mm^2.

Fig. 4.3 compares the *I–V* characteristics of the device under 450, 520, 638, and 808 nm illumination at the same intensity (5.97 mW/mm^2). The device demonstrates larger open circuit voltage under visible light illumination (450, 520, 638 nm) than under NIR light illumination (808 nm). Under small forward external bias (such as 0.05 V), the photocurrent is positive for 808 nm but remains negative for visible wavelengths, demonstrating a visible-light-induced photocurrent polarity switching effect. We attribute this effect to the significant lowering of the Schottky barrier at the drain side under visible illumination, caused by the trapping of photoexcited holes at the MoS_2/Au interface, which generates the negative photo-induced electric field. Only when the forward bias is sufficiently large can the drain-side barrier be lowered enough to reverse the photocurrent polarity under visible illumination.

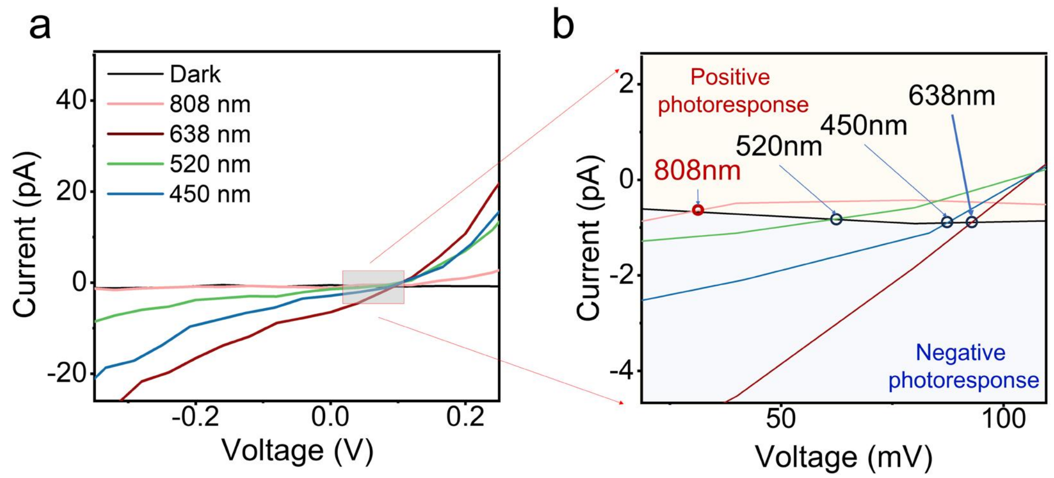


Fig. 4.3 *I-V* characteristics of the device under visible light and near-infrared light illumination. (a) *I-V* characteristics under 450, 520, 638, and 808 nm illumination at the same intensity (5.97 mW/mm^2). (b) Enlarged view of (a), demonstrating the photocurrent of 450, 520, 638 and 808 nm under bias voltage from 0.02 V to 0.1 V.

4.2.1.2. Mechanisms for the RGB-NIR decoupled bipolar photoresponse

The RGB-NIR bipolar photoresponse behaviour can also be seen from the photocurrent mapping measurements under different illumination wavelengths. **Fig. 4.4**a shows the optical image of the MoS_2/In_2Se_3 heterostructure device, with the regions outlined by the black dashed line corresponding to the material areas. **Fig. 4.4**b and **Fig. 4.4**c present the photocurrent mapping result under 808 nm and 638 nm, respectively (V_{bias} = 0.05 V).

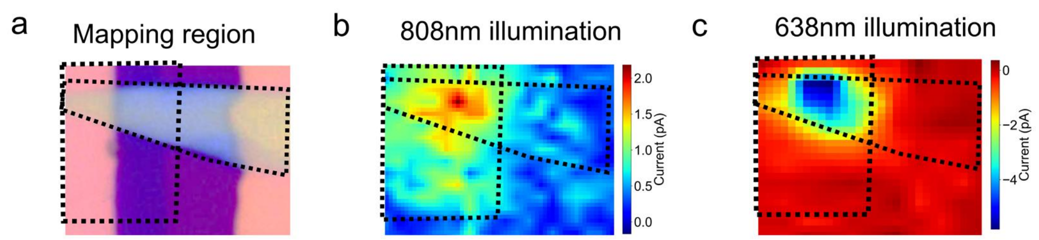


Fig. 4.4 Photocurrent mapping results. (a) The optical image of the MoS_2/In_2Se_3 heterostructure device. The material regions are outlined by the black dashed. Scanning

photocurrent mapping result under (b) 808 nm and (c) 638 nm laser illumination, respectively. The bias voltage is 0.05 V. The diameter of the laser spot is 1μm, and the intensity of the laser is 0.64 $\mu W/\mu m^2$.

As shown in the photocurrent mapping result, the devices exhibits clear positive current signals under 808 nm illumination and negative current signals under 638 nm, consistent with the wavelength-dependent bipolar photoresponse observed in *I-V* measurements (**Fig. 4.3**). Notably, under 638 nm illumination, the strongest photocurrent signal is located near the MoS_2/Au interface (source side). This observation supports our proposed mechanism that photocurrent polarity switching under visible light illumination is driven by visible-light-induced reduction of the Schottky barrier at the MoS_2/Au interface (source side), which alters the photogenerated electron flow direction and induces a reversed photocurrent, as illustrated in **Fig. 4.5**.

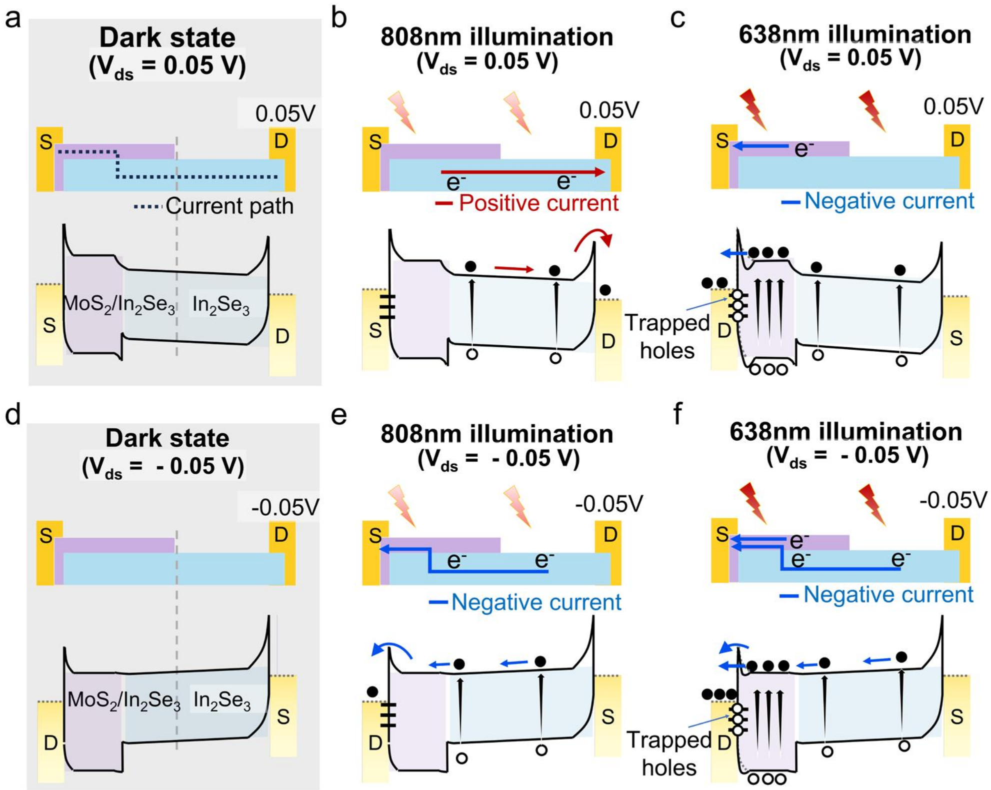


Fig. 4.5 The band diagram of the device. Band diagram under 0.05 V bias in (a) dark condition, (b) 808 nm illumination condition, and (c) 638 nm illumination condition. Band diagram under -0.05 V bias in (d) dark condition, (e) 808 nm illumination condition, and (f) 638 nm illumination condition.

As shown in **Fig. 4.5**a-c, under positive bias, the Schottky barrier at the In_2Se_3/Au interface (drain side) is lower than that of the MoS_2/Au interface (source side). Under 808 nm illumination, the photon energy is smaller than the bandgap (E_g) of MoS_2(~1.6 eV)[98], leading to only a limited number of photogenerated carriers in MoS_2 through defect-assisted absorption[99]. The resulting hole density is insufficient to fill sulfur vacancy states at the MoS_2/Au interface to appreciably reduce the Schottky barrier height[100,101]. In contrast, the photon energy exceeds the E_g of In_2Se_3(~1.4 eV)[102], enabling efficient carrier generation within the In_2Se_3. The photogenerated electrons in

In_2Se_3 experience a reduced effective barrier height at the drain side under positive bias, giving rise to a positive photocurrent. Under visible light illumination (e.g., 638 nm), the photon energy exceeds the E_g of MoS_2, leading to a significantly higher number of photogenerated carriers. A substantial number of photogenerated holes become trapped at sulfur vacancy states near the MoS_2/Au interface. This interfacial hole accumulation lowers the Schottky barrier height at the source side. As a result, the photogenerated electrons in MoS_2 encounter a lower effective barrier at the source than at the drain, resulting in a negative photocurrent. The photogenerated electrons in In_2Se_3 encounter a similar barrier height at the drain and source side, resulting in negligible photocurrent contribution.

The device behaviour under negative bias (-0.05 V) is also analysed, where the photocurrent is negative for all wavelengths. As shown in **Fig. 4.5**d, the applied negative bias voltage lowers the Schottky barrier at the MoS_2/Au interface. Under 808 nm illumination (**Fig. 4.5**e), the photogenerated electrons in In_2Se_3 experience a lower effective barrier height at the source side, giving rise to a negative photocurrent. Under 638 nm illumination (**Fig. 4.5**f), the photocurrent is also negative, and its magnitude is larger than under 0.05 V. This arise from that the effective barrier height at the drain side is reduced by both the applied field and the visible light-induced hole trapping, therefore, the photogenerated electrons in MoS_2 and In_2Se_3 both experience a lower barrier at the source side, resulting in a negative photocurrent with larger magnitude.

4.2.1.3. Optoelectronic characterization of the MoS_2/In_2Se_3 heterostructure device

The device photoresponse under different bias conditions is measured (**Fig. 4.6**).

For all wavelengths, the magnitude of the output photocurrent increases with increasing input intensity ranging from 10 mW/mm^2 to 80 mW/mm^2. Under 0.05 V bias voltage, the device exhibits the RGB-NIR bipolar photoresponse behaviour (**Fig. 4.6**a). Under -0.05 V bias, the photocurrent of the device under both visible light illumination and NIR light illumination conditions is negative (**Fig. 4.6**b). The photocurrent magnitude increases with increasing input intensity. These results are consistent with the proposed mechanism. At zero bias, 808 nm illumination yields negligible photoresponse, while visible light generates negative photocurrent (**Fig. 4.6**c).

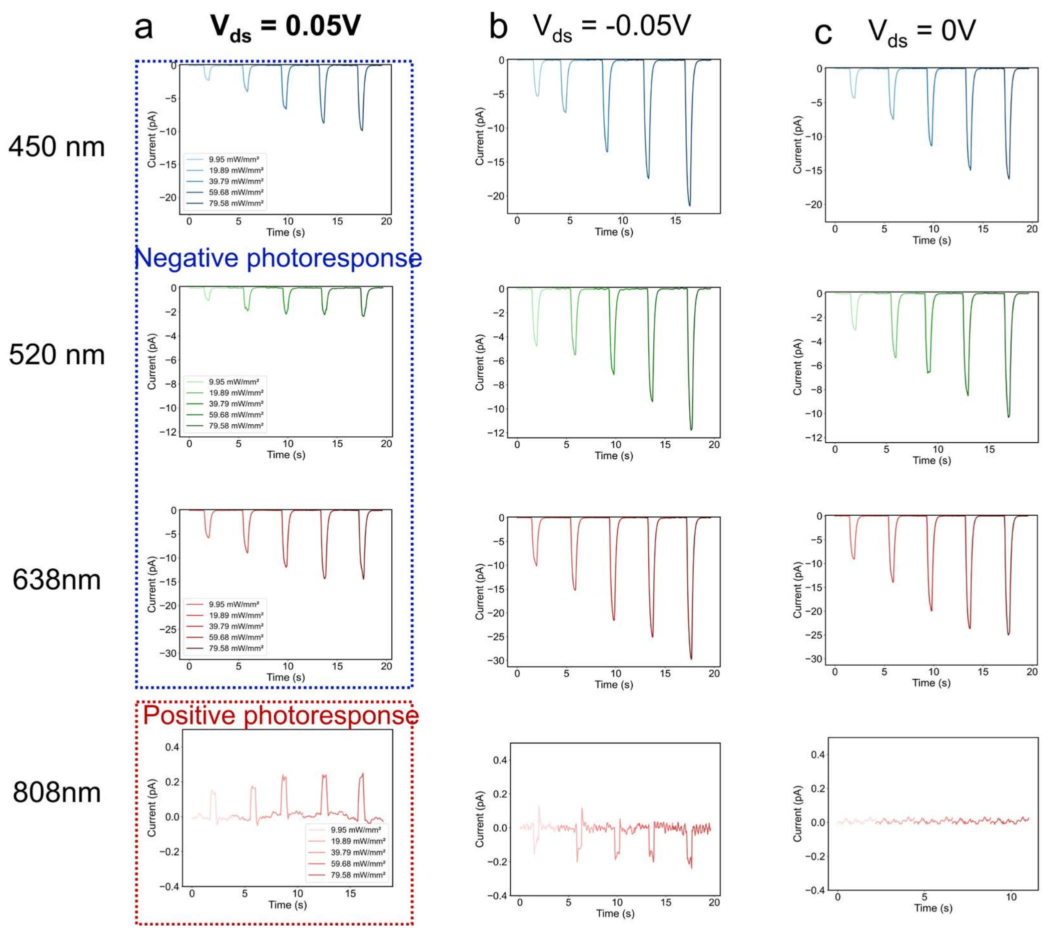


Fig. 4.6 Photocurrent of the device under 450, 520, 638, and 808 nm illumination conditions(intensity) at bias voltage of (a) 0.05 V, (b) -0.05 V, and (c) 0 V. The optical pulse frequency is 1Hz, with intensity ranging from 10 to 80 mW/mm^2.

4.2.2. Bipolar MoS_2/In_2Se_3 device array-based RGB-D camera

4.2.2.1. Bipolar MoS_2/In_2Se_3 device array-based RGB-D camera for 3D hand gesture reconstruction

In this section, we conducted a 3D hand gesture reconstruction task. In this demonstration, a simulated 28×28 MoS_2/In_2Se_3 device array is used as an RGB-D camera to simultaneously capture both the 2D RGB image for visual appearance and 3D position data of the hand. Thanks to the RGB-NIR bipolar photoresponse of the device, the simulated RGB-D camera is capable of simultaneously sensing and decoupling visible and NIR components in a single detection. Moreover, because both signals are acquired by the same array, the resulting RGB and depth information are spatially aligned at the x- and y-axis, eliminating the need for external calibration or complex alignment processing (**Fig. 4.7**).

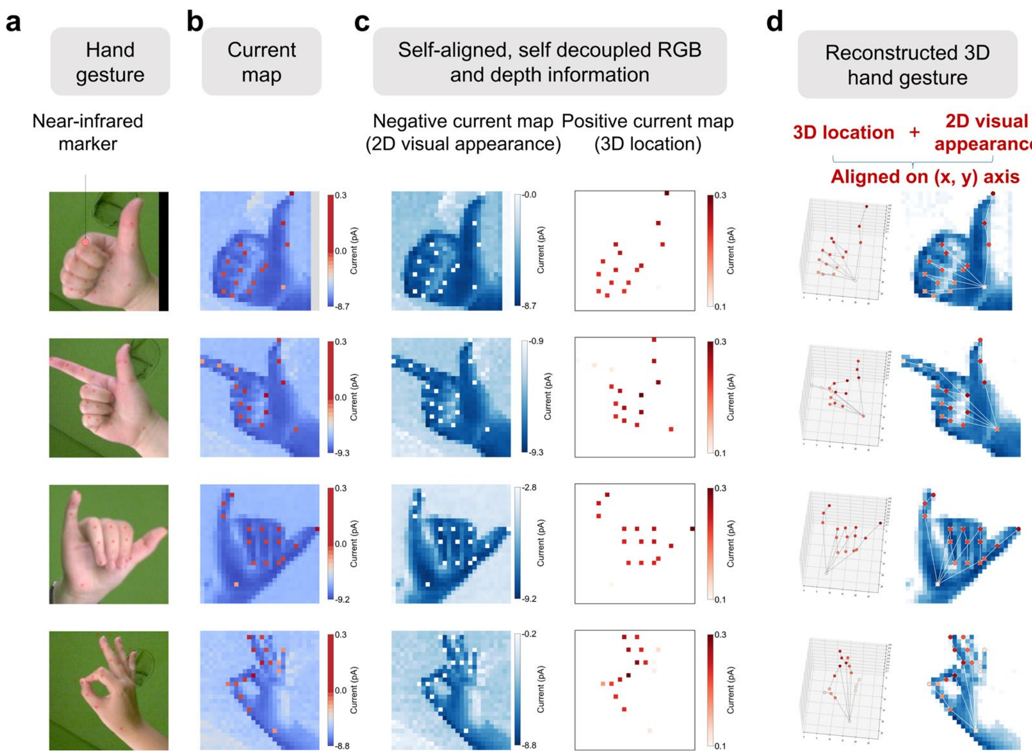


Fig. 4.7 MoS_2/In_2Se_3 array-based RGB-D camera for 3D hand gesture reconstruction. (a)

Input hand gestures including four representative gestures: “good”, “eight”, “six”, and “OK”. (b) The output current map from the RGB-D camera. (c) Positive and negative current maps decoupled from output current map, representing the 2D appearance and the 3D location information, respectively. (d) Reconstructed 3D hand gesture. The visual appearance and 3D location information are spatially aligned.

The input consists of a hand performing four different gestures, including “good”, “eight”, “six”, and “OK” (**Fig. 4.7**a). To support accurate 3D location detection, NIR markers were placed on key parts of the hand, such as fingertips and knuckles. These markers emit NIR signals, allowing their positions to be detected by the RGB-D camera alongside the visible hand image.

During image capture, each pixel in the array receives either visible light or NIR light. Devices illuminated by visible light generate a negative photocurrent, while those illuminated by NIR light from the markers generate a positive photocurrent. As a result, the total current output of the array (**Fig. 4.7**b) naturally contains both signals and can be easily separated into two distinct maps: a negative photocurrent map representing the grayscale image of the hand, and a positive photocurrent map indicating the positions of the NIR markers (**Fig. 4.7**c).

The negative current map is used to reconstruct the 2D visual appearance of the hand. By mapping photocurrent levels to optical intensity, we obtain grayscale images that clearly show the hand gesture’s contour and shape. The positive current map shows the 3D positions of the NIR markers. The x and y coordinates are taken directly from the locations of the positive photocurrents in the current map. To get the z-axis (depth), we first convert the photocurrent value into light intensity, then use a linear relationship between NIR intensity and distance to estimate the distance between the marker and the camera. In this way, we can determine the full 3D coordinates of each marker.

Importantly, because both the RGB and depth information are derived from the same current map output from a single camera, they are inherently aligned. This alignment avoids the need for additional alignment processing that is typically required in dual-camera systems, which often introduce additional energy consumption and system complexity.

With the 3D marker positions and the 2D hand images, we reconstructed full 3D representations of the hand gestures (**Fig. 4.7**d). These results demonstrate the capability of our MoS_2/In_2Se_3 array-based RGB-D camera on 3D hand gesture reconstruction task.

4.2.2.2. RGB-Depth aware intelligent vision system for AR

By integrating the RGB-D camera with back-end neural network models, an RGB-Depth aware intelligent vision system for AR applications can be constructed. The MoS_2/In_2Se_3 array-based RGB-D camera can provide both 2D RGB images for visual appearance and spatially aligned 3D position data. The 2D RGB images allow the neural networks at the back end to accurately recognize the objects, while 3D position enables dynamic gesture recognition through back-end neural networks and allows the category labels to be accurately overlayed on physical scene.

The system workflow is shown in the **Fig. 4.8**. The input scene involves two hands interacting with one object. Specifically, there are three categories of objects (“lotion”, "milk", and "chip") and two types of gestures ("grab" and "place"), resulting in six different input scenarios. NIR markers are placed on key locations of hands, enabling the MoS_2/In_2Se_3 array-based RGB-D camera to obtain the real-time 3D location information. The purpose is to simultaneously recognize both the hand gesture and object type, and accurately overlay the corresponding category labels at the appropriate positions.

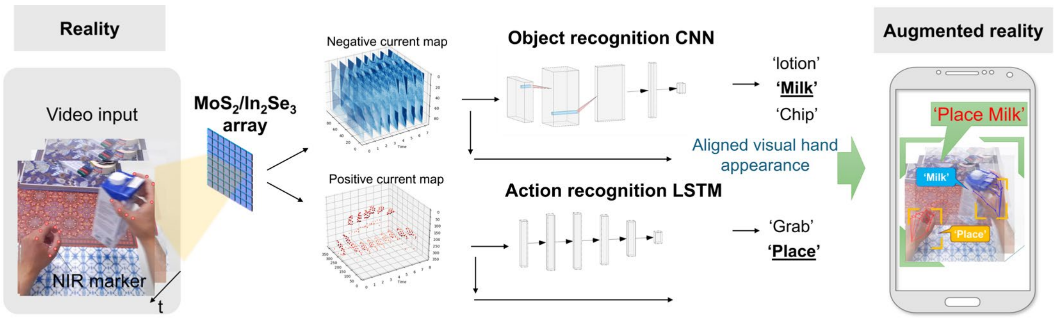


Fig. 4.8 Workflow of the RGB-Depth aware intelligent vision system for AR based on MoS_2/In_2Se_3 array-based RGB-D camera and neural networks.

During processing, the scene is first input to the MoS_2/In_2Se_3 array, outputting current maps that encode 2D visual appearance information and 3D location information in negative currents and positive currents, respectively (**Fig. 4.9**a). To recognize the object type, the negative current maps generated by the array, which contains 2D visual appearance information, are input into a SqueezeNet CNN[103] to extract the semantic features of the object. Meanwhile, the 3D location data extracted from the positive current map is input into an LSTM network for gesture recognition, since the LSTM model excels at capturing temporal relationships in frame sequences[104]. The use of 3D position information enhances the robustness of gesture recognition, as it provides motion patterns that remain consistent across varying lighting conditions and background noisy objects. The accuracy for correctly recognizing both hand gestures and object categories is 91.8% (**Fig. 4.9**b). Besides, with the accurate and spatially aligned 3D position data, the system is able to overlay the recognized gesture labels onto the hand region, while the object labels are overlayed onto the position between the hands, which is expected to be the location of the object.

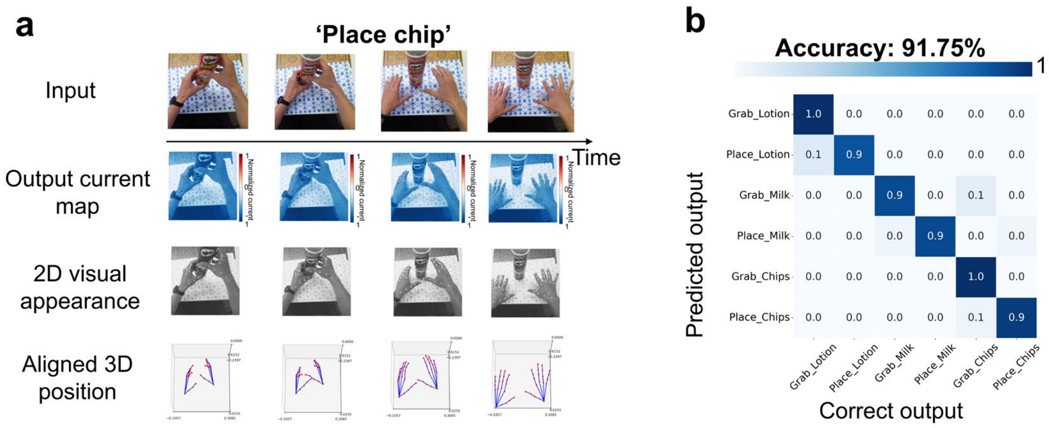


Fig. 4.9 Outputs of the AR system. (a) Simulated output from the MoS_2/In_2Se_3 array and the extracted 2D visual appearance information and aligned 3D location data. The first row shows four frames from an input scene "place chip". The second row presents the corresponding output current maps from the array. The negative current values encode the 2D visual appearance information and the positive current values encode the 3D location information. The third and fourth rows display the extracted 2D visual appearance and 3D hand position, respectively. (b) Accuracy of the system in correctly recognizing both hand gestures and object categories.

Chapter 5 Conclusions and outlooks

In this thesis, we present three intelligent vision devices and construct three intelligent vision systems, each targeting for solving a specific challenge in realizing a truly intelligent system.

Targeting the lack of reconfigurability in vision devices and systems, we present a bio-inspired, intelligent vision system that achieves both high reconfigurability and high efficiency. The system integrates a multi-paradigm ITO/CuO_x/Pd device array with a reconfigurable mode-control circuit array and reconfigurable dataflow architecture, achieving full spiking/non-spiking/NI/AI reconfigurability across device to system-level. The intelligent vision system enables dynamic resource allocation between NI and AI functionalities towards high-quality smart imaging and high-accuracy recognition tasks, and spiking/non-spiking mode transitions towards frameless dynamic and frame-based static scenarios. The system reaches up to 52.6 TOPS/W for NI-centric computing and 75.5 TOPS/W for AI-centric computing, outperforming state-of-the-art intelligent vision system by up to two orders of magnitude. These results suggest potential opportunities for advancing future neuromorphic vision systems.

Looking ahead, further enhancements could expand the system’s capabilities. First, engineering richer device dynamics (e.g., adaptive and nonlinear responses) can enable advanced operations like normalization and complex activation functions, supporting more complex models and broader visual scenarios. Second, regarding the device reliability and robustness, achieving endurance beyond 10^6 cycles is an important step towards the practical deployment. Several strategies could be adopted to further improve device performance. During fabrication, a heat dissipation layer can be introduced for better thermal management. During device programming, the write, verify, and adaptive programming strategy can be adopted. This would allow us to program based on the current state of the device and only program when needed, avoid overstressing the device. In addition, the behaviour of the device across the full commercial temperature range, from -40 °C to +85 °C is also important. Fully characterizing these behaviours will help us better understand and optimize the system's performance for commercial use. Third, in terms of scalability, further scaling the array up would allow the supporting of larger computing workload. The main barrier that lies in scaling our multi-paradigm device to megapixel arrays is large-scale parallel readout. Therefore, the development of 3D BEOL integration is required. This technology will allow for each pixel to be directly connected to CMOS control circuits for parallel operation and readout, enabling further scaling to megapixel arrays.

To address the problem of lack of advanced processing power in complex visual scenarios, especially in multi-object detection tasks, we also demonstrate a multidimensional intelligent vision system with advanced multidimensional information processing capabilities, inspired by the multilevel visual filters in retina and cortex. This system is based on the multidimensional MoS_2/HZO device, which exhibits enhanced wavelength selectivity between blue, green, and red wavelengths based on the non-volatile polarization modulated MoS_2/Au interfacial barrier mechanism. This results in the low-level colour clustering and mid-level multidimensional decoupling capability of the device. The device also demonstrates a non-volatile polarization enhanced multi-photoresponsivity states behaviour for high-level processing. Based on these characteristics and the co-designed HVBF algorithm, the MoS_2/HZO array-based intelligent vision system can perform hierarchical multidimensional information processing within a single array. In a noisy multi-target scene, it can effectively filter the RGB noise, segment the two overlapping digits, and recognize them accurately with 93.8% recognition rate.

Looking ahead, we need to further optimize the fabrication process for large-scale megapixel MoS_2/HZO arrays. A larger array can process more complex visual scenarios that require higher resolution to capture detailed information, such as identifying small objects or differentiating objects by their textures. This can be done by advancement from two aspects. First, the growth of MoS_2 thin film and HZO with high uniformity and quality is important. For MoS_2, this can be achieved by further optimization of the CVD system to improve mass transport uniformity and ensure stable nucleation. For HZO, improvements in deposition and annealing processes help reduce defects and oxygen vacancies, therefore improving both the ferroelectric properties and material uniformity. Second, the development of the 3D BEOL technology is required to solve the interconnection complexity problem caused by the requirement of three signal lines per device in the MoS_2/HZO FeFET array. In parallel, to reduce the cost associated with exotic materials and complex fabrication processes, future efforts will explore the incorporation of large-scale, low-cost oxide-based materials that implement similar multidimensional device concepts for achieving a highly functional and cost-effective system.

For 3D scene understanding tasks involving both depth perception and RGB information perception, we propose an intelligent bipolar MoS_2/In_2Se_3 vision device. Based on the visible light-induced barrier-lowering mechanism, the device exhibits negative photoresponse to RGB light and positive photoresponse to NIR light under the same positive biasing condition. This characteristic results in the simultaneous sensing and decoupling of RGB light and NIR light. As a result, the bipolar MoS_2/In_2Se_3 device can perceive both visual appearance information, which is obtained from the RGB light, and 3D depth information, which is obtained by receiving signals from NIR markers. An intelligent vision system is constructed. The bipolar MoS_2/In_2Se_3 array constructed RGB-D camera at the front end enables accurate and efficient capture of both 2D RGB images and 3D depth images. Capturing both 2D RGB images and 3D depth images using a single array ensures the spatial alignment between these two images, eliminating complex and inefficient alignment procedures required in traditional dual-camera systems. At the back end, a SqueezeNet CNN and an LSTM network extract 2D visual appearance feature from RGB images for static object recognition and temporal patterns of 3D location data for dynamic gesture recognition, respectively. This system ensures precise alignment between virtual annotation and real-world objects for immersive AR experience with 91.8% accuracy.

In the future, wafer-scale fabrication technology for uniform thin film MoS_2 and multilayer In_2Se_3 should be investigated for a megapixel RGB-D camera with higher resolution for more complex input scenes. In parallel, replicating the demonstrated device behaviour in a large-scalable, low-cost oxide-based device is also a promising route toward larger array for practical deployment. This can be realized by appropriate oxide material selection, controlled doping, and engineering of bulk and interface dynamics. Beyond device development, the edge readout circuit should be co-designed with the device array to ensure maximum accuracy and energy efficiency on perceiving visual appearance and 3D location data. For example, the transimpedance amplifier's working range should match the device output current range and bandwidth to ensure linear and precise current readout with minimum power consumption.

To further increase the intelligence in vision devices and systems, future development can be first oriented around device performance. Improving the device response speed allows the resulting intelligent vision system to support rapidly changing scenarios. Additionally, enhancing the device dynamic range helps the system handle extreme lighting conditions, including low-light environments and mixed lighting in the same scene. Enhancing device endurance and extending the operational temperature range are also essential for reliable, real-world deployment. At the array level, electrical inspection can be employed to identify pixels suffering from short circuits or large leakage current, while novel deep-learning-based optical inspection will enable the automatic detection of contamination, fabrication damage, and other abnormal patterns. After detecting the defected pixels, mild defects can be compensated algorithmically, while severely defective pixels can be compensated through interpolation from neighbouring pixels. At the circuit level, co-design of circuits with the device array is required for optimized area and energy efficiency. At the system level, 3D BEOL CMOS integration technology is crucial for supporting efficient, high-throughput control and readout, which enables large-scale intelligent vision systems capable of addressing increasingly complex visual tasks.

Reference

1 Shi, Y. *et al.* in *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).* 3015-3024.

2 Saleh, K., Szénási, S. & Vámossy, Z. in *2021 IEEE 19th World Symposium on Applied Machine Intelligence and Informatics (SAMI).* 000477-000484.

3 Wang, K., Zhou, T., Li, X. & Ren, F. Performance and Challenges of 3D Object Detection Methods in Complex Scenes for Autonomous Driving. *IEEE Transactions on Intelligent Vehicles* **8**, 1699-1716 (2023). https://doi.org/10.1109/TIV.2022.3213796

4 Alamri, F. Comprehensive study on object detection for security and surveillance: A concise review. *Multimedia Tools and Applications* (2025). https://doi.org/10.1007/s11042-025-20801-6

5 Gao, Y.-F., Wang, H.-Y. & Bian, X.-N. (IEEE).

6 Dudek, P. *et al.* Sensor-level computer vision with pixel processor arrays for agile robots. *Science Robotics* **7**, eabl7755 (2022). https://doi.org/doi:10.1126/scirobotics.abl7755

7 Pei, J. *et al.* Towards artificial general intelligence with hybrid Tianjic chip architecture. *Nature* **572**, 106-111 (2019). https://doi.org/10.1038/s41586-019-1424-8

8 Wilson, G. & Premson, Y. FPGA Implementation of Hardware Efficient Algorithm for Image Contrast Enhancement Using Xilinx System Generator. *Procedia Technology* **24**, 1141-1148 (2016). https://doi.org/10.1016/j.protcy.2016.05.067

9 Merolla, P. A. *et al.* A million spiking-neuron integrated circuit with a scalable communication network and interface. *Science* **345**, 668-673 (2014).

10 Zhang, W. *et al.* Edge learning using a fully integrated neuro-inspired memristor chip. *Science* **381**, 1205-1211 (2023). https://doi.org/doi:10.1126/science.ade3483

11 Zhou, F. & Chai, Y. Near-sensor and in-sensor computing. *Nature Electronics* **3**, 664-671 (2020). https://doi.org/10.1038/s41928-020-00501-9

12 Wang, T. *et al.* Reconfigurable neuromorphic memristor network for ultralow-power smart textile electronics. *Nat Commun* **13**, 7432 (2022). https://doi.org/10.1038/s41467-022-35160-1

13 Huang, X. *et al.* 2D MoS2-based reconfigurable analog hardware. *Nature Communications* **16**, 101 (2025). https://doi.org/10.1038/s41467-024-55395-4

14 Huang, P.-Y. *et al.* Neuro-inspired optical sensor array for high-accuracy static image recognition and dynamic trace extraction. *Nature Communications* **14**, 6736 (2023). https://doi.org/10.1038/s41467-023-42488-9

15 Zhou, G. *et al.* Full hardware implementation of neuromorphic visual system based on multimodal optoelectronic resistive memory arrays for versatile image processing. *Nature Communications* **14**, 8489 (2023). https://doi.org/10.1038/s41467-023-43944-2

16 Zhou, Y. *et al.* Computational event-driven vision sensors for in-sensor spiking neural networks. *Nature Electronics* **6**, 870-878 (2023). https://doi.org/10.1038/s41928-023-01055-2

17 Davies, M. *et al.* Loihi: A Neuromorphic Manycore Processor with On-Chip Learning. *IEEE Micro* **38**, 82-99 (2018). https://doi.org/10.1109/MM.2018.112130359

18 Richter, O. *et al.* DYNAP-SE2: a scalable multi-core dynamic neuromorphic asynchronous spiking neural network processor. *Neuromorphic Computing and Engineering* **4**, 014003 (2024).

19 Rico, A. *et al.* AMD XDNA NPU in Ryzen AI Processors. *IEEE Micro* **44**, 73-82 (2024).

https://doi.org/10.1109/MM.2024.3423692
20 Yang, Z. *et al.* A vision chip with complementary pathways for open-world sensing. *Nature* **629**, 1027-1033 (2024). https://doi.org/10.1038/s41586-024-07358-4
21 Felleman, D. J. & Van Essen, D. C. Distributed hierarchical processing in the primate cerebral cortex. *Cerebral cortex (New York, NY: 1991)* **1**, 1-47 (1991).
22 Hauptman, M., Liu, Y.-F. & Bedny, M. Built to Adapt: Mechanisms of Cognitive Flexibility in the Human Brain. *Annual Review of Developmental Psychology* **6**, 133-162 (2024). https://doi.org/https://doi.org/10.1146/annurev-devpsych-120621-042108
23 Cole, Michael W., Bassett, Danielle S., Power, Jonathan D., Braver, Todd S. & Petersen, Steven E. Intrinsic and Task-Evoked Network Architectures of the Human Brain. *Neuron* **83**, 238-251 (2014). https://doi.org/https://doi.org/10.1016/j.neuron.2014.05.014
24 Gonzalez-Castillo, J. & Bandettini, P. A. Task-based dynamic functional connectivity: Recent findings and open questions. *NeuroImage* **180**, 526-533 (2018). https://doi.org/https://doi.org/10.1016/j.neuroimage.2017.08.006
25 Pinto, L. *et al.* Task-Dependent Changes in the Large-Scale Dynamics and Necessity of Cortical Regions. *Neuron* **104**, 810-824 e819 (2019). https://doi.org/10.1016/j.neuron.2019.08.025
26 Li, Z. *et al.* A reconfigurable heterostructure transistor array for monocular 3D parallax reconstruction. *Nature Electronics* **8**, 46-55 (2025). https://doi.org/10.1038/s41928-024-01261-6
27 Liang, X., Luo, Y., Pei, Y., Wang, M. & Liu, C. Multimode transistors and neural networks based on ion-dynamic capacitance. *Nature Electronics* **5**, 859-869 (2022). https://doi.org/10.1038/s41928-022-00876-x
28 Fu, Y. *et al.* Reconfigurable Synaptic and Neuronal Functions in a V/VOx/HfWOx/Pt Memristor for Nonpolar Spiking Convolutional Neural Network. *Advanced Functional Materials* **32**, 2111996 (2022). https://doi.org/https://doi.org/10.1002/adfm.202111996
29 Gao, J. *et al.* Reconfigurable neuromorphic functions in antiferroelectric transistors through coupled polarization switching and charge trapping dynamics. *Nature Communications* **16**, 4368 (2025). https://doi.org/10.1038/s41467-025-59603-7
30 Dang, B. *et al.* Reconfigurable in-sensor processing based on a multi-phototransistor–one-memristor array. *Nature Electronics* **7**, 991-1003 (2024). https://doi.org/10.1038/s41928-024-01280-3
31 Xu, J. *et al.* in *2024 IEEE International Conference on IC Design and Technology (ICICDT).* 1-3.
32 Huang, H. *et al.* Fully integrated multi-mode optoelectronic memristor array for diversified in-sensor computing. *Nature Nanotechnology* **20**, 93-103 (2025). https://doi.org/10.1038/s41565-024-01794-z
33 Fossum, E. *et al.* Development of CMOS Active Pixel Image Sensors for Low Cost Commercial Applications. (1994).
34 Eric, R. F. in *75th Anniversary of the Transistor* 281-291 (IEEE, 2023).
35 Wu, G. *et al.* Ferroelectric-defined reconfigurable homojunctions for in-memory sensing and computing. *Nature Materials* **22**, 1499-1506 (2023). https://doi.org/10.1038/s41563-023-01676-0
36 Yang, Y. *et al.* Firing feature-driven neural circuits with scalable memristive neurons for

robotic obstacle avoidance. *Nature Communications* **15**, 4318 (2024).

37 Zhu, J. *et al.* A heterogeneously integrated spiking neuron array for multimode-fused perception and object classification. *Advanced Materials* **34**, 2200481 (2022).

38 Han, J.-K. *et al.* Cointegration of single-transistor neurons and synapses by nanoscale CMOS fabrication for highly scalable neuromorphic hardware. *Science Advances* **7**, eabg8836 (2021). https://doi.org/doi:10.1126/sciadv.abg8836

39 Wu, Q. *et al.* Spike Encoding with Optic Sensory Neurons Enable a Pulse Coupled Neural Network for Ultraviolet Image Segmentation. *Nano Letters* **20**, 8015-8023 (2020). https://doi.org/10.1021/acs.nanolett.0c02892

40 Han, J.-K. *et al.* in *2021 IEEE International Electron Devices Meeting (IEDM).* 1-4 (IEEE).

41 Wang, X. *et al.* Vertically integrated spiking cone photoreceptor arrays for color perception. *Nat Commun* **14**, 3444 (2023). https://doi.org/10.1038/s41467-023-39143-8

42 John, R. A. *et al.* Optogenetics inspired transition metal dichalcogenide neuristors for in-memory deep recurrent neural networks. *Nature Communications* **11**, 3211 (2020). https://doi.org/10.1038/s41467-020-16985-0

43 Yamazaki, T. *et al.* in *2017 IEEE International Solid-State Circuits Conference (ISSCC).* 82-83.

44 Yang, X. *et al.* A Bio-Inspired Spiking Vision Chip Based on SPAD Imaging and Direct Spike Computing for Versatile Edge Vision. *IEEE Journal of Solid-State Circuits* **59**, 1883-1898 (2024). https://doi.org/10.1109/JSSC.2023.3340018

45 Murakami, H. *et al.* in *2022 IEEE International Solid-State Circuits Conference (ISSCC).* 104-106.

46 Bose, L., Dudek, P., Carey, S. J. & Chen, J. in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW).* 3995-3996.

47 Carey, S. J., Lopich, A., Barr, D. R. W., Wang, B. & Dudek, P. in *2013 Symposium on VLSI Circuits.* C182-C183.

48 Millet, L. *et al.* A 5500-frames/s 85-GOPS/W 3-D Stacked BSI Vision Chip Based on Parallel In-Focal-Plane Acquisition and Processing. *IEEE Journal of Solid-State Circuits* **54**, 1096-1105 (2019). https://doi.org/10.1109/JSSC.2018.2886325

49 Dziugaite, G. K., Roy, D. M. & Ghahramani, Z. Training generative neural networks via maximum mean discrepancy optimization. *arXiv preprint arXiv:1505.03906* (2015).

50 Ribeiro, M. T., Singh, S. & Guestrin, C. in *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining.* 1135-1144.

51 Kimmel, R. Demosaicing: image reconstruction from color CCD samples. *IEEE Transactions on Image Processing* **8**, 1221-1228 (1999). https://doi.org/10.1109/83.784434

52 Li, Y., Bi, Y., Zhang, W., Ren, J. & Chen, J. M2GF: Multi-Scale and Multi-Directional Gabor Filters for Image Edge Detection. *Applied Sciences* **13**, 9409 (2023).

53 Ataky, S., Saqui, D., Matos, J., Britto, A. & Koerich, A. Multiscale Analysis for Improving Texture Classification. *Applied Sciences* **13**, 1291 (2023). https://doi.org/10.3390/app13031291

54 Pal, S. K., Pramanik, A., Maiti, J. & Mitra, P. Deep learning in multi-object detection and tracking: state of the art. *Applied Intelligence* **51**, 6400-6429 (2021). https://doi.org/10.1007/s10489-021-02293-7

55 Iwashita, A. S. & Papa, J. P. An Overview on Concept Drift Learning. *IEEE Access* **7**, 1532-

1547 (2019). https://doi.org/10.1109/ACCESS.2018.2886026

56 Lu, J. *et al.* Learning under Concept Drift: A Review. *IEEE Transactions on Knowledge and Data Engineering* **31**, 2346-2363 (2019). https://doi.org/10.1109/TKDE.2018.2876857

57 Lee, D. K., Itti, L., Koch, C. & Braun, J. Attention activates winner-take-all competition among visual filters. *Nat Neurosci* **2**, 375-381 (1999). https://doi.org/10.1038/7286

58 Riesenhuber, M. & Poggio, T. Hierarchical models of object recognition in cortex. *Nature Neuroscience* **2**, 1019-1025 (1999). https://doi.org/10.1038/14819

59 Peirce, J. W. Understanding mid-level representations in visual processing. *Journal of Vision* **15**, 5-5 (2015). https://doi.org/10.1167/15.7.5

60 Sowden, P. T. & Schyns, P. G. Channel surfing in the visual brain. *Trends in Cognitive Sciences* **10**, 538-545 (2006). https://doi.org/https://doi.org/10.1016/j.tics.2006.10.007

61 Carandini, M. *et al.* Do We Know What the Early Visual System Does? *The Journal of Neuroscience* **25**, 10577-10597 (2005). https://doi.org/10.1523/jneurosci.3726-05.2005

62 Groen, II, Silson, E. H. & Baker, C. I. Contributions of low- and high-level properties to neural processing of visual scenes in the human brain. *Philos Trans R Soc Lond B Biol Sci* **372** (2017). https://doi.org/10.1098/rstb.2016.0102

63 Medathati, N. V. K., Neumann, H., Masson, G. S. & Kornprobst, P. Bio-inspired computer vision: Towards a synergistic approach of artificial and biological vision. *Computer Vision and Image Understanding* **150**, 1-30 (2016). https://doi.org/https://doi.org/10.1016/j.cviu.2016.04.009

64 Thorpe, S., Fize, D. & Marlot, C. Speed of Processing in the Human Visual System. *Nature* **381**, 520-522 (1996). https://doi.org/10.1038/381520a0

65 Mennel, L. *et al.* Ultrafast machine vision with 2D material neural network image sensors. *Nature* **579**, 62-66 (2020). https://doi.org/10.1038/s41586-020-2038-x

66 Migliato Marega, G. *et al.* Logic-in-memory based on an atomically thin semiconductor. *Nature* **587**, 72-77 (2020). https://doi.org/10.1038/s41586-020-2861-0

67 Yu, Y. *et al.* Controlled Scalable Synthesis of Uniform, High-Quality Monolayer and Few-layer MoS2 Films. *Scientific Reports* **3**, 1866 (2013). https://doi.org/10.1038/srep01866

68 Lopez-Sanchez, O., Lembke, D., Kayci, M., Radenovic, A. & Kis, A. Ultrasensitive photodetectors based on monolayer MoS2. *Nature Nanotechnology* **8**, 497-501 (2013). https://doi.org/10.1038/nnano.2013.100

69 Jiang, Y. *et al.* Asymmetric Ferroelectric-Gated Two-Dimensional Transistor Integrating Self-Rectifying Photoelectric Memory and Artificial Synapse. *ACS Nano* **16**, 11218-11226 (2022). https://doi.org/10.1021/acsnano.2c04271

70 Zhang, Q. *et al.* Tunable Multi-Bit Nonvolatile Memory Based on Ferroelectric Field-Effect Transistors. *Advanced Electronic Materials* **8**, 2101189 (2022). https://doi.org/https://doi.org/10.1002/aelm.202101189

71 Mahlouji, R. *et al.* Influence of High-κ Dielectrics Integration on ALD-Based MoS2 Field-Effect Transistor Performance. *ACS Applied Nano Materials* **7**, 18786-18800 (2024). https://doi.org/10.1021/acsanm.4c02214

72 Kumar, J., Birla, S. & Agarwal, G. A review on effect of various high-k dielectric materials on the performance of FinFET device. *Materials Today: Proceedings* **79**, 297-302 (2023). https://doi.org/https://doi.org/10.1016/j.matpr.2022.11.204

73 Yim, K. *et al.* Novel high-κ dielectrics for next-generation electronic devices screened by

automated ab initio calculations. *NPG Asia Materials* **7**, e190-e190 (2015). https://doi.org/10.1038/am.2015.57

74 Cohen, E. & Lewis, D. D. Approximating Matrix Multiplication for Pattern Recognition Tasks. *Journal of Algorithms* **30**, 211-252 (1999). https://doi.org/https://doi.org/10.1006/jagm.1998.0989

75 Arena, F., Collotta, M., Pau, G. & Termine, F. An Overview of Augmented Reality. *Computers* **11**, 28 (2022).

76 Abdallah, M., Sawalhi, G., Mazhar, A., AlRifaee, M. & Salah, M. Factors Influencing the Quality of Augmented Reality Applications. *Procedia Computer Science* **251**, 150-156 (2024). https://doi.org/https://doi.org/10.1016/j.procs.2024.11.095

77 Kim, M., Choi, S. H., Park, K.-B. & Lee, J. Y. A Hybrid Approach to Industrial Augmented Reality Using Deep Learning-Based Facility Segmentation and Depth Prediction. *Sensors* **21**, 307 (2021).

78 Kumar, D., Iyer, S., Raja, E., Kumar, R. & Kafle, V. P. Improving Pedestrian Navigation in Urban Environment Using Augmented Reality and Landmark Recognition. *IEEE Communications Standards Magazine* **8**, 20-26 (2024). https://doi.org/10.1109/MCOMSTD.0003.2300017

79 von Atzigen, M. *et al.* Marker-free surgical navigation of rod bending using a stereo neural network and augmented reality in spinal fusion. *Medical Image Analysis* **77**, 102365 (2022). https://doi.org/https://doi.org/10.1016/j.media.2022.102365

80 Kim, G. *et al.* Metasurface-driven full-space structured light for three-dimensional imaging. *Nature Communications* **13**, 5920 (2022). https://doi.org/10.1038/s41467-022-32117-2

81 Ruget, A. *et al.* Pixels2Pose: Super-resolution time-of-flight imaging for 3D pose estimation. *Science Advances* **8**, eade0123 (2022). https://doi.org/doi:10.1126/sciadv.ade0123

82 Kolahi, A. *et al.* Design of a marker-based human motion tracking system. *Biomedical Signal Processing and Control* **2**, 59-67 (2007). https://doi.org/https://doi.org/10.1016/j.bspc.2007.02.001

83 Wang, T., Liu, Y. & Wang, Y. in *2008 International Conference on Computer Science and Software Engineering.* 816-819.

84 Zhangu, W., Jun, Z., Chuanguang, D., Xin, G. & Kai, Y. Traffic vehicle cognition in severe weather based on radar and infrared thermal camera fusion. *Measurement Science and Technology* **32**, 095111 (2021). https://doi.org/10.1088/1361-6501/abfe34

85 Qin, C. *et al.* in *2020 IEEE International Conference on Robotics and Automation (ICRA).* 8899-8906.

86 Chidsin, W., Gu, Y. & Goncharenko, I. AR-Based Navigation Using RGB-D Camera and Hybrid Map. *Sustainability* **13**, 5585 (2021).

87 Pan, Y., Chen, C., Li, D., Zhao, Z. & Hong, J. Augmented reality-based robot teleoperation system using RGB-D imaging and attitude teaching device. *Robotics and Computer-Integrated Manufacturing* **71**, 102167 (2021). https://doi.org/https://doi.org/10.1016/j.rcim.2021.102167

88 Wang, S.-A., Lu, M.-H., Lee, A.-T., Chen, C.-Y. & Lee, L.-W. Performance of Microsoft Azure Kinect DK as a tool for estimating human body segment lengths. *Scientific Reports* **14**, 15811 (2024). https://doi.org/10.1038/s41598-024-66798-0

89 Kurillo, G., Hemingway, E., Cheng, M.-L. & Cheng, L. Evaluating the Accuracy of the Azure

Kinect and Kinect v2. *Sensors* **22**, 2469 (2022).

90 Albert, J. A. *et al.* Evaluation of the Pose Tracking Performance of the Azure Kinect and Kinect v2 for Gait Analysis in Comparison with a Gold Standard: A Pilot Study. *Sensors* **20**, 5104 (2020).

91 Zhang, C., Huang, T. & Shen, Y. A Review of RGB-D Camera Calibration Methods. *Journal of Geodesy and Geoinformation Science* **4**, 11-33 (2021). https://doi.org/10.11947/j.JGGS.2021.0402

92 Qiu, D., Pang, J., Sun, W. & Yang, C. in *Proceedings of the ieee/cvf international conference on computer vision.* 9994-10003.

93 Dutta, D., Mukherjee, S., Uzhansky, M. & Koren, E. Cross-field optoelectronic modulation via inter-coupled ferroelectricity in 2D In2Se3. *npj 2D Materials and Applications* **5** (2021). https://doi.org/10.1038/s41699-021-00261-w

94 Mu, Y. *et al.* Homo-type α-In₂Se₃/PdSe₂ Ferroelectric van der Waals Heterojunction Photodetectors with High-performance and Broadband. *Advanced Functional Materials* **34** (2024). https://doi.org/10.1002/adfm.202315543

95 Wang, L. *et al.* A Scanning Microwave Impedance Microscopy Study of α-In₂Se₃ Ferroelectric Semiconductor. *Advanced Functional Materials* **34** (2024). https://doi.org/10.1002/adfm.202316583

96 Lee, C. *et al.* Anomalous Lattice Vibrations of Single- and Few-Layer MoS₂. *ACS Nano* **4**, 2695-2700 (2010). https://doi.org/10.1021/nn1003937

97 Li, H. *et al.* Optical Identification of Single- and Few-Layer MoS₂ Sheets. *Small* **8**, 682-686 (2012). https://doi.org/10.1002/smll.201101958

98 Zhong, H. *et al.* Interfacial Properties of Monolayer and Bilayer MoS2 Contacts with Metals: Beyond the Energy Band Calculations. *Scientific Reports* **6**, 21786 (2016). https://doi.org/10.1038/srep21786

99 Xie, Y. *et al.* Ultrabroadband MoS₂ Photodetector with Spectral Response from 445 to 2717 nm. *Advanced Materials* **29**, 1605972 (2017). https://doi.org/10.1002/adma.201605972

100 Abnavi, A., Ahmadi, R., Ghanbari, H., Akinwande, D. & Adachi, M. M. Switchable Photovoltaic Effect Induced by Light Intensity. *ACS Nano* **18**, 34147-34157 (2024). https://doi.org/10.1021/acsnano.4c10392

101 Zhou, C. *et al.* Self-Powered and Reconfigurable Double-Terminal MoS2 Photodetector for Image Recognition. *Nano Letters* **25**, 3515-3523 (2025). https://doi.org/10.1021/acs.nanolett.4c06084

102 Lu, Z. *et al.* 2D Materials Based on Main Group Element Compounds: Phases, Synthesis, Characterization, and Applications. *Advanced Functional Materials* **30**, 2001127 (2020). https://doi.org/https://doi.org/10.1002/adfm.202001127

103 Iandola, F. *et al.* SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <0.5MB model size. (2016). https://doi.org/10.48550/arXiv.1602.07360

104 Athira, K. A. & Divya Udayan, J. Temporal Fusion of Time-Distributed VGG-16 and LSTM for Precise Action Recognition in Video Sequences. *Procedia Computer Science* **233**, 892-901 (2024). https://doi.org/https://doi.org/10.1016/j.procs.2024.03.278