Nature Communications

https://doi.org/10.1038/s41467-026-71907-w

Article in Press

Temporal-Spatial Fusion Vision Hardware Enables Streamlined In-Sensor Computing for Dynamic Scenes

Received: 19 June 2025

Accepted: 1 April 2026

Cite this article as: Wu, Y., Deng, W., Liu, R. *et al.* Temporal-Spatial Fusion Vision Hardware Enables Streamlined In-Sensor Computing for Dynamic Scenes. *Nat Commun* (2026). https://doi.org/10.1038/s41467-026-71907-w

Yi Wu, Wenjie Deng, Ruihao Liu, Chutian Xiao, Jianmiao Guo, Chaoyi Zhu, Qinqi Ren, Zehao Li, Yushan Wu, Kexin Li, Xueliang Ma, Xiaoting Wang, Zhangyang Xu, Zikang Zhao, Zhijie Chen, Yang Chai & Yongzhe Zhang

Temporal-Spatial Fusion Vision Hardware Enables Streamlined In-Sensor Computing for Dynamic Scenes

Yi Wu[1,2,3,†], Wenjie Deng[2, †, *], Ruihao Liu[2], Chutian Xiao[2], Jianmiao Guo[3], Chaoyi Zhu[3], Qinqi Ren[3], Zehao Li[2], Yushan Wu[2], Kexin Li[2], Xueliang Ma[2], Xiaoting Wang[2], Zhangyang Xu[2], Zikang Zhao[2], Zhijie Chen[2,*], Yang Chai[3,*] and Yongzhe Zhang[1,2,*]

[1] State Key Laboratory of Materials Low-Carbon Recycling, College of Materials Science and Engineering, Beijing University of Technology, Beijing, 100124, China

[2] Key Laboratory of Optoelectronics Technology of Education Ministry of China, School of Information Science and Technology, Beijing University of Technology, Beijing 100124, China.

[3] Department of Applied Physics, The Hong Kong Polytechnic University, Kowloon, Hong Kong, China

[†] These authors contributed equally: Yi Wu, Wenjie Deng.

*Corresponding author. E-mail: yzzhang@bjut.edu.cn; ychai@polyu.edu.hk; wjdeng@bjut.edu.cn; chenzj@bjut.edu.cn

Keywords: in-sensor computing, photodetector, temporal-spatial fusion, artificial vision system, machine vision

Abstract

Time and space constitute fundamental dimensions of physical reality, making their integrated processing crucial for advanced vision perception systems. Current visual information processing faces dual limitations: von Neumann architecture-induced data transfer bottlenecks and spatial feature processing often disregard temporal dynamics, while temporal analyzers oversimplify spatial complexity. Here we propose an artificial vision hardware enabling intrinsic temporal-spatial fusion through voltage-tunable temporal differentiation with microsecond-scale resolution and photoresponse-weighted spatial compression via pixel binning. The architecture achieves millisecond-level latency from sensing to decision in autonomous driving scenarios through in-sensor spatiotemporal fusion, eliminating external computing dependencies. Experimental validation demonstrates 95 % recognition accuracy in human actions database while the operation counts required is only 1/10 of conventional convolutional processing. This work facilitates physical-level spatiotemporal fusion through the co-optimization of photodetector arrays and weighted control circuits, which could fundamentally reshape machine vision architectures with potential extensions to real-time decision systems.

Introduction

Temporal-spatial integration constitutes the fundamental framework of physical reality, wherein temporal progression and spatial architecture form an inseparable continuum governing real-world dynamics. Biological visual systems achieve sophisticated environmental perception through intrinsic spatiotemporal fusion mechanisms.[1,2] This bioinspired framework finds technological resonance in artificial perception systems, where spatiotemporal integration becomes a computational imperative rather than merely biological phenomenon. Recent breakthroughs in autonomous driving technologies, exemplified by the BEVFormer (Bird's Eye View Transformer) algorithm[3], demonstrate the transformative potential of artificial spatiotemporal fusion through temporal and spatial attention mechanisms. This approach generates compact BEV feature maps that simultaneously encode multidimensional spatiotemporal information, thereby enabling computationally efficient target detection critical for vision-based

autonomous systems like Tesla's. Current vision systems, reliant on the von Neumann architecture and backend algorithmic processing for spatiotemporal fusion perception, suffer from excessive energy consumption and latency due to the transmission and processing of massive raw, unprocessed data. Hence, co-innovation must begin at the sensory hardware level, bridging spatiotemporal fusion principles with novel detector architectures.

The emerging in-sensor computing paradigm presents a promising solution through in-situ computation within photodetectors, effectively bypassing redundant data conversion and transmission.[4] However, current in-sensor computing vision strategies only process information from either the temporal or spatial dimension, failing to capture integrated spatiotemporal dynamics. The processing of spatial information is primarily inspired by the structure of receptive fields in the human eye, similar to the mechanisms of artificial neural networks, where the core operation involves the weighted summation of light intensity. One approach is to make the responsivity of the device itself adjustable.[5,6] By assembling multiple devices with tunable photoresponse into an array, optical information can be processed. Alternatively, photodetectors can be connected with devices that have adjustable weights.[7,8] However, this functionality only allows for the spatial weighted summation of light intensity, without considering the temporal distribution characteristics of the light intensity. For the extraction of temporal information features, most methods rely on the device's ability to store photogenerated carriers through trap or floating gate structures, effectively integrating light intensity over time to process event information.

In recent years, reservoir computing based on this principle has demonstrated exceptional advantages in handling temporal information.[9–11] However, the accumulated (reservoir computing-processed) images still require additional computational units to process spatial characteristics. Thus, it is evident that in practical implementations, designers face a dichotomy: either perform frame-by-frame spatial analysis followed by temporal sequencing, or integrate multiple temporal frames before extracting spatial features. Both approaches introduce inherent latency and computational overhead, particularly problematic for high-speed object detection. This

fundamental trade-off underscores the need for true co-processing of spatiotemporal information at the sensory level. Although recent computational event-driven vision sensors demonstrate motion recognition capability, their dependence on the fact that each pixel requires two devices to be integrated in order to realize the function of information processing fundamentally conflicts with miniaturization requirements.[12]

In this work, we present a temporal-spatial fusion artificial vision hardware (TSF-AVH) achieving native temporal-spatial co-processing through two mechanisms: (1) electrically tunable temporal differentiation with microsecond-scale resolution; (2) pixel binning executing spatial compression through photoresponse-weighted summation (**Figure 1)**. The temporal information processing based on differentiation exhibits higher temporal resolution compared to conventional integration-based. Integration-based methods necessitate the collection of light intensity over a short period, resulting in delays when recognizing high-speed moving objects. By employing voltage adjustments to modulate the responsivity of photodetectors and concurrently binning pixels with varying light responses through specific components, fundamental recognition can be directly accomplished at the detector end. This temporal-spatial fusion operation facilitates in-sensor feature extraction, thereby reducing backend computation compared to traditional systems. To experimentally verify the functionality of our vision hardware, a spatial light modulator (SLM) is employed to provide rich spatiotemporal visual elements. For autonomous driving, the system directly recognizes changes in signal lights, achieving a decision delay benefit of 1 ms from the detection end to the control end, which satisfies the time delay requirements for autonomous driving. In the complex scenario of human action recognition, this system directly implements feature extraction by temporal and spatial fusion, requiring only 10 % of the operation counts of convolutional neural networks to achieve a recognition accuracy rate exceeding 95 %. This system is anticipated to advance the development of the next generation of artificial vision systems.

Results

Design and characteristics of the devices

Figure 2a delineates the hierarchical architecture of the TSF-AVH. The device stack integrates five functional layers (top to bottom): (i) Au top electrodes arrays (70 nm, e-beam evaporated) interfacing with weighting circuits, (ii) graphene electrodes (bilayer, CVD-grown) offering both conductivity and light transmission, (iii) SiO_2 dielectric spacer (300 nm) for DC signal isolation, (iv) intrinsic Si absorption layer (500 μm), and (v) Au bottom electrodes connected to transimpedance amplifiers (TIA, gain: 10^7 V/A). The material characterization is shown in **Supplementary Figure 1**. The morphology of the TSF-AVH is as shown in **Figure 2b**. The temporal processing capability originates from the device's differential photoresponse mechanism. As demonstrated in **Figure 2c**, under rectangular optical pulses (λ=605 nm, rise time <1 μs), the detector generates bipolar current spikes, precisely matching the temporal derivative of optical power (dP/dt).[13] The magnified view of the time - differential pulse signal in **Figures 2d-e** shows that the bipolar current spikes duration is less than 1 millisecond. Time-resolved measurements using a high-speed oscilloscope (20 GHz bandwidth) revealed sub-millisecond transient responses, as quantitatively demonstrated in **Figures 2f** and **g**. The response time (τ), defined as the interval between 10 % and 90 % of peak signal amplitude, was measured at 100 μs. This corresponds to a -3dB bandwidth of ~3.5 kHz, calculated as the reciprocal of the response time. This temporal resolution corresponds to equivalent frame rates exceeding 3500 fps (frames per second) in conventional imaging systems, demonstrating capability for high-speed motion tracking applications.

Electrically tunable differential response, a key enabler for neuromorphic computing, is quantified in **Figure 2h**. Sweeping top electrode from -1 V to +5 V modulates photoresponse amplitude at fixed illumination. This characteristic distinguishes our device from standard silicon photodiode arrays (as shown in the **Supplementary Figure 2**), which lack obvious voltage-tunable photoresponse. The mechanism of electrical tunable time-differential photoresponse is shown in **Supplementary Note 1**.[14,15] **Figure 2i** systematically maps the voltage-dependent output signals

across three decades of optical power (10-1000 μW), demonstrating stable operation over the full dynamic range. The inherent built-in potential, arising from Fermi-level mismatch between Au and graphene, induces asymmetric voltage modulation characteristics about 0 V, requiring threshold voltages to overcome interfacial band bending.[16] The photoresponsivity was calculated from the measured voltage signal through a transimpedance amplifier (TIA) and oscilloscope. The max photoresponsivity reaches the order of mA/W (typically $R=49\ mA\cdot W^{-1}$ at $P_{light}=10$ μW, $V_{mod}= 6$ V), depending on the incident light intensity and applied modulation voltage. Although slightly lower, the photoresponsivity remains within a comparable range to that of state-of-the-art commercially available silicon PIN photodiodes (S1223).

The scalability from single-pixel devices to a 6×6 array necessitated systematic uniformity evaluation. As depicted in **Figure 3a**, uniformity estimation was performed using a collimated light spot (λ=685 nm, spot diameter 200 μm) to sequentially illuminate individual pixels under four gate biases (-1 V, 1 V, 3 V, 5 V). **Figure 3b** shows the highly consistent photoresponse characteristics for all 36 pixels. Statistical analysis of response amplitude distributions **(Figure 3c**) reveals voltage-dependent separation while maintaining intra-group uniformity.[10] The observed minor non-uniformities prove inconsequential for neuromorphic computing implementations, as artificial neural networks inherently tolerate parameter variations through distributed information encoding. Information is encoded holistically in the entirety of network weights through distributed representations, as opposed to localized exact storage in traditional computing. Neural architectures process signals via population coding rather than precise analog computations[17], where collective responses from multiple neurons (analogous to array pixels) compensate for individual deviations. The spatial consistency satisfies the tolerance threshold for neuromorphic array operations, confirming feasibility for large-area image processing applications. Furthermore, the noise characteristics and the quantitative definition of computing accuracy is explored in **Supplementary Note 2**. Theoretically, this analog in-sensor spatiotemporal scheme may include stably set 32767 distinguishable weight states with an effective bit resolution of approximately 15

bits at the output.

Mechanism for Spatial Information Processing

TSF-AVH enables spatial information processing through bio-inspired spatially distributed optoelectronic weighting, as schematically illustrated in **Figure 4a**. A biological neuron processes information by receiving multiple input signals through synaptic structures, integrating them, and transmitting the result through dendrites to the next neuron, thereby completing the information processing.[18,19] This process can be simplified as a weighted summation operation, where multiple inputs are multiplied by weights and then summed to produce an output signal.[20] The weighted summation operation is the central computational operation in fully connected artificial neural networks, and artificial neural networks based on this operation can be trained using gradient descent algorithm. The fully connected artificial neural network achieves the classification of the image by weighting and summing the grey values of the pixels at different locations of the pattern as indicated in the left panel of **Figure 4b**. A pattern undergoes weighted summation operations with four different weight matrices, resulting in four corresponding output values.[21] Classification is achieved by comparing the magnitudes of these four outputs. For this TSF-AVH array, voltage can be used as a weight to control the photoresponse of each pixel, with light intensity serving as the input signal. By connecting multiple photodetectors in parallel and using Kirchhoff's current law to output the total current, a neuron-like information processing capability is achieved. On the device, the pattern is projected onto four positions on the array: upper left, upper right, lower left, and lower right as indicated in the right panel of **Figure 4b**. The pixels at each position can be voltage-modulated to control their responsivity, and by connecting the pixels in parallel, a weighted summation of the optical signals similar to a full connected neural network is achieved. During the training of the artificial neural network, one-hot encoding is used for four different letters. The classification result is determined by comparing the final output voltage values. The accuracy of the classification is dependent on the values of the weight matrices, which are obtained through extensive data-driven learning and training. When an unknown image is projected onto

the visual system, the channel number corresponding to the maximum output represents the recognized result. Since the position of each pixel determines the spatial distribution of the collected light intensity, spatial optical information processing is accomplished.

To control the weight of each pixel, this work designed a weight control circuit centered around a multi-channel Digital-to-Analog Converter (DAC), using an FPGA to control the voltage output at each DAC terminal, thereby adjusting the weight of each photodetector, as shown in **Figure 4c**. The DAC chip is the core of the circuit, featuring multiple functional pins: SCK (Serial Clock) for synchronizing serial data transmission; SDI (Serial Data Input) for data input; and LD/CS (Load/Chip Select) for selecting the chip and controlling data loading. The weight matrix (ω (1,1) …ω (6,6)) is input via a touchpad, and the FPGA controls the DAC to apply the corresponding voltage (V (1,1) … V (6,6)) to each pixel. The long-term stability of the bias voltages supplied by the weighting control circuit is demonstrated in the **Supplementary Figure 6**. The currents from all detectors are aggregated through an amplifier to an oscilloscope for testing. The physical diagram of the weight control system is shown in **Figure 4d**. The complete equipment setup is shown in **Supplementary Figure 7**. The optical information input end consists of a laser paired with a reflective spatial light modulator, capable of projecting any spatial distribution pattern onto the photodetector array. The schematic of the laser and spatial light modulator connection is shown in **Figure 4e**. **Figure 4f** displays four different letter patterns projected onto the photodetector array, with noise signals superimposed on the original letters to increase pattern diversity.

Figures 4g-h show the weight distribution of the neural network before and after training and the corresponding voltage values applied to the artificial visual system. During network training, experimentally characterized photoresponse curves were incorporated into the forward model, while weight ranges were constrained to match the achievable voltage-tuning range of the hardware. This enables direct mapping of trained algorithmic parameters to executable hardware configurations, ensuring compatibility between computational models and physical

implementation. The detail of training of artificial neural networks is shown in **Supplementary Note 3**. **Figure 4i** illustrates the change curves of accuracy and loss function during the training cycles of the neural network, achieving 100 % accuracy after 20 cycles. **Figure 4j** displays the output values of each channel when different patterns are projected onto the device. When the input is C, i.e., the first column, Output1 produces the highest pulse signal, so the neural network decision result is C, matching the output pattern. When the input is L, Output2 is the highest, so the neural network decision result is L. **Figure 4k** presents the data of the highest pulse amplitude for the four channels under different pattern projections. Experimental results show that by simply setting a certain threshold, the incident pattern can be directly classified, completely independent of external computational support. Thus, the system completes the processing of spatial information. The subsequent work will integrate the temporal characteristics of differential response with the spatial characteristics of weighted summation to realize an artificial visual system based on spatiotemporal fusion.

Spatiotemporally fused information processing

With the increasing adoption of vision-based solutions in autonomous driving, photodetectors capable of processing visual information are expected to play a significant role in this field.[22] By integrating voltage-tunable temporal differentiation at the pixel level with neural-inspired spatial Information compression at the array level, our architecture achieves hardware-level spatiotemporal fusion (More details are shown in **Supplementary Note 1**).

Figure 5a illustrates a typical autonomous driving scenario where a vehicle is waiting for a traffic light change. The traffic light has three different directions: "go straight," "turn left," and "turn right." Initially, the traffic light can be in one of three states and may change to one of the other two states at any given moment. As shown in **Figure 5b**, there are six different event states in total. These correspond to six different steering wheel decision states: "rotate the steering wheel 90 degrees counterclockwise from the go-straight state," "rotate 90 degrees clockwise from the go-

straight state," "rotate 90 degrees clockwise from the left-turn state," "rotate 180 degrees clockwise from the left-turn state," "rotate 90 degrees counterclockwise from the right-turn state," and "rotate 180 degrees counterclockwise from the right-turn state." In order to complete the decision making for this scenario, the orientation of the traffic light, i.e. the spatial information of the light intensity, and the change of the traffic light, i.e. the temporal information of the light intensity, need to be analyzed. Traditional vision solutions involve recognizing arrows pointing in three different directions and analyzing the results before and after a change to command the steering wheel accordingly. This approach is sensitive to delays, which are critical in the high-speed information processing required for autonomous driving. The separation of perception and computation necessitates data conversion and transfer, leading to delays. In contrast, temporal differential sensing can quickly detect changes in light intensity and use appropriate weights to recognize changes in image states. By employing a similar artificial neural network training process as described in the previous section, the optimal weight distribution for this scenario can be obtained. The motion scene is then projected onto a photodetector array with six different weight distributions. By comparing the signal magnitudes from six different output ports, steering wheel decisions can be made in real-time without the need for data conversion and transfer. As shown in **Figure 5c**, when the traffic light changes from "turn right" to "go straight," the fifth channel shows a higher signal value, leading to the decision to "rotate the steering wheel 90 degrees counterclockwise from the right-turn state." In practical experiments, due to the limitations of array size, simultaneous output from six channels is not feasible. Therefore, a time-division multiplexing method is used, where the same scene is projected onto the detector array under different weight voltage controls to obtain six channel outputs.

Based on the above, it is evident that compared to traditional vision solutions, the TSF-AVH developed in this study is inherently well-suited for adapting to changes in light intensity. By adjusting the weight voltages, it can directly distinguish events related to varying light intensities.[23] This approach overcomes the delays associated with traditional frame-based methods and avoids

the need for data conversion and transmission inherent in conventional von Neumann architectures.[24] It enables direct decision-making from the detector to the control end, with the device's response speed equating to the decision speed. According to previous characterizations of the photodetector's response speed, the pulse photodetector responds in less than 100 μs, which meets the millisecond-level decision-making requirements of autonomous driving.[25]

Using voltage as a weight to simulate the behavior of artificial neural networks allows for rapid target recognition. However, the photodetection process is unidirectional, unlike the bidirectional processing of electrical signals, which can be repeated. As a result, neural networks based on detector responsivity control can only implement single-layer networks. This characteristic limit the chip's ability to recognize more complex scenarios. In contrast to fully connected neural networks, convolutional neural networks (CNNs) are also an efficient type of artificial neural network for image recognition. CNNs are characterized by their ability to extract image features through convolution operations with kernels and reduce data size while preserving features through pooling layers. The convolution operation involves multiplying and summing the grayscale values of an image with the corresponding values in the convolution kernel.[26,27] Suppose the pixel values of an image region are represented by matrix

$$\begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} \quad (1)$$

, and the convolution kernel by matrix

$$\begin{bmatrix} k_{11} & k_{12} & k_{13} \\ k_{21} & k_{22} & k_{23} \\ k_{31} & k_{32} & k_{33} \end{bmatrix} \quad (2)$$

; then the convolved pixel value is

$$b = a_{11}k_{11} + a_{12}k_{12} + a_{13}k_{13} + a_{21}k_{21} + a_{22}k_{22} + a_{23}k_{23} + a_{31}k_{31} + a_{32}k_{32} + a_{33}k_{33} \quad (3)$$

. This convolution operation is similar to the weighted summation described in the previous section. Therefore, matrix **a** can be considered as a light intensity matrix, and matrix **k** as a detector

matrix controlled by voltage. Measuring the output electrical signal yields the result of feature extraction of the light intensity signal under the convolution kernel. By leveraging the time-differential photoresponse characteristics of the detectors, direct feature extraction and recognition of moving targets can be achieved. **Figure 5d** illustrates two different feature extraction methods for Weizmann human actions database.[28] The TSF-AVH can directly recognize moving scenes. In contrast, traditional detectors require algorithms to first extract moving targets, such as using frame difference methods to perform differential operations on image sequences, followed by feature extraction using convolution kernels. This traditional method inevitably involves data transfer and handling. The TSF-AVH does not rely on additional computational units, achieving both spatial feature extraction of images and temporal motion feature extraction of moving targets. By combining these two aspects of feature extraction, the system retains motion characteristics while reducing data size, significantly enhancing the efficiency of subsequent recognition by fully connected neural networks. **Figure 5e** compares the training epochs and accuracy rates during the recognition training process of the artificial vision system based on traditional detectors and spatiotemporal fusion. It is evident that images captured by TSF-AVH significantly enhance subsequent recognition efficiency. The improvement in accuracy can be attributed to two main factors. First, the static background and dynamic targets exhibit different temporal characteristics, which are extracted by the photodetector's inherent differential photoresponse.[29] Second, by adjusting the responsivity of the photodetector through voltage control, the spatial features of moving targets are further processed. This enables the photodetector to perform temporal-spatial fusion of visual information, allowing the subsequent recognition network to distinguish these features with minimal computation. **Figure 5f** compares several common vision strategies using accuracy and computation as metrics. The experimental results demonstrate that this approach maintains high recognition accuracy up to 95 % while requiring very low computational operation counts, which is about 10 % compared with that by CNNs. For specific methods of computational efficiency evaluation, please see the **Supplementary Note 4**. In addition, the potential comparison

against conventional vision systems on key metrics such as accuracy, latency, and energy efficiency is evaluated in the **Supplementary Table 1**. Due to its robust structural design (More detail see **Supplementary Note 5**), this architecture supports scaling to larger arrays for more complex functions. However, some extreme flickering environments can still lead to erroneous output results (**Supplementary Figure 10**), necessitating further optimization in algorithm design. Furthermore, established circuit solutions like active-matrix addressing provide manageable pathways for control expansion in scaled arrays, ensuring the architecture's scalability without compromising performance. Furthermore, scaling up the detector array further can substantially lessen reliance on external computation, enabling a more efficient artificial vision system. Additionally, TSF-AVH exhibits excellent stability as shown in **Supplementary Note 6**. Experimental results demonstrate a 10 % decay over 15 months, with no significant reduction in light response after ten minutes of cyclic light stimulation. It maintains good stability even under high-temperature and high-humidity conditions.

Discussion

This study demonstrates a temporal-spatial fusion artificial vision system with in-sensor computing capabilities to overcome the von Neumann bottleneck in visual data processing. By integrating a 6 × 6 differential photodetector arrays with bias-tunable photoresponsivity and a multi-channel DAC/FPGA control system, we enable direct temporal-spatial vision feature processing through optical intensity differentiation and voltage-programmable convolutional operations. The proposed temporal-spatial fusion vision processor demonstrates exceptional performance in dynamic feature extraction, achieving sensor-to-control decision latency ≤1 ms in autonomous driving scenarios. For complex human action recognition tasks, the system attains >95 % accuracy with merely 10 % computational overhead compared to conventional CNN. This work has accomplished board-level functional verification based on the array chip. By

leveraging the current mature integrated circuit fabrication process, we are poised to achieve an artificial vision system that matches the pixel scale of contemporary CMOS image sensors while incorporating temporal-spatial fusion capabilities for visual information processing. These advancements validate the device's capability in efficient temporal-spatial information processing, combining in-sensor optical differentiation with hardware-accelerated convolutional operations, establishing a new paradigm for real-time edge computing in intelligent perception systems.

The TSF-AVH architecture demonstrates broader importance through its CMOS-compatible, scalable design that transcends its initial autonomous driving application. Its in-sensor computing paradigm fundamentally addresses the data bottleneck of conventional systems by transforming pixel scaling into parallel analog computing resource expansion, maintaining ultra-low latency and power efficiency even at increased resolutions. While the current 3.5 kHz bandwidth and predefined weight configuration present certain constraints for extreme high-speed scenarios and dynamic class adaptation, the 90 dB dynamic range and neuromorphic population coding of the system provide inherent robustness in noisy environments. Future development paths, including 3D integration, adaptive voltage regulation, and memristor-based weight updates, are expected to improve low-light performance and enable real-time learning. A comparative analysis of TSF-AVH against recent hardware technologies is listed in Supplementary Table 2. These characteristics make TSF-AVH a foundational technology for various applications such as industrial automation, robot vision, and extended reality systems, and its hardware-level spatiotemporal processing capabilities provide a new reference for efficient and rapid-response machine perception.

Methods

Fabrication of the devices

The intrinsic silicon substrates and CVD graphene were purchased from Nanjing MKNANO Tech.

Co., Ltd. (www.mukenano.com). The silicon wafers were sequentially cleaned in acetone, isopropyl alcohol, and deionized water. A wet etching process was employed to transfer CVD-grown graphene onto silicon substrates. The graphene patterns were defined through photolithography followed by oxygen plasma etching. Ti/Au contact layers (10 nm/70 nm) were deposited using UV lithography (SUSS MJB4 mask aligner) and electron beam evaporation (HHV FL400 system). The final device structure was completed using a lift-off process. The electrode-patterned silicon chip was electrically connected to a custom PCB daughterboard through a custom-built silver wire bonding system. The daughterboard was linked to a mainboard integrating a 16-bit DAC（LTC2688) and support electronics via board-to-board connectors. Dynamic voltage patterning was executed by an FPGA (Xilinx Zynq-7010), commanding the DAC to achieve pixel-specific responsivity weighting through bias modulation.

Characterization of the devices

The material characterizations of Raman were carried out by a confocal Raman spectroscopy system (WITec Alpha 300R) with 532 nm laser excitation (100 μW, room temperature). Fs-Pro semiconductor parameter instrument was used to measure the electrical properties of the devices. The electrical characterizations were performed in ambient air with a transimpedance amplifier and oscilloscope.

Data Availability

The data that support the findings of this study are presented in the paper and the Supplementary Information. Source data are provided with this paper.

Code availability

The codes that support the findings of this study are available from the corresponding authors on request.

References

1. Van Essen, D. C. & Gallant, J. L. Neural mechanisms of form and motion processing in the primate visual system. *Neuron* **13**, 1–10 (1994).

2. Chen, G. & Gong, P. A spatiotemporal mechanism of visual attention: Superdiffusive motion and theta oscillations of neural population activity patterns. *Science Advances* **8**, eabl4995 (2022).

3. Li, Z. *et al.* BEVFormer: Learning Bird's-Eye-View Representation from Multi-Camera Images via Spatiotemporal Transformers. Preprint at https://doi.org/10.48550/arXiv.2203.17270 (2022).

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

5. He, Z. *et al.* Perovskite retinomorphic image sensor for embodied intelligent vision. *Science Advances* **11**, eads2834 (2025).

6. Yang, Y. *et al.* In-sensor dynamic computing for intelligent machine vision. *Nat Electron* **7**, 225–233 (2024).

7. Dang, B. *et al.* Reconfigurable in-sensor processing based on a multi-phototransistor–one-memristor array. *Nat Electron* **7**, 991–1003 (2024).

8. Li, F. *et al.* An artificial visual neuron with multiplexed rate and time-to-first-spike coding. *Nat Commun* **15**, 3689 (2024).

9. Wu, X. *et al.* Ultralow-power optoelectronic synaptic transistors based on polyzwitterion dielectrics for in-sensor reservoir computing. *Science Advances* **10**, eadn4524 (2024).

10. Huang, H. *et al.* Fully integrated multi-mode optoelectronic memristor array for diversified in-sensor computing. *Nat. Nanotechnol.* **20**, 93–103 (2025).

11. Gao, H. *et al.* Bio-inspired mid-infrared neuromorphic transistors for dynamic trajectory perception using PdSe2/pentacene heterostructure. *Nat Commun* **16**, 5241 (2025).

12. Zhou, Y. *et al.* Computational event-driven vision sensors for in-sensor spiking neural networks. *Nat Electron* **6**, 870–878 (2023).

13. Reissig, L., Dalgleish, S. & Awaga, K. A differential photodetector: Detecting light modulations using transient photocurrents. *AIP Advances* **6**, 015306 (2016).

14. Herrera, C. T. & Labram, J. G. Quantifying the performance of perovskite retinomorphic sensors. *J. Phys. D: Appl. Phys.* **54**, 475110 (2021).

15. Al Mahfuz, M. M., Islam, R. & Ko, D.-K. Artificial Amacrine Retinal Circuits. *ACS Appl. Mater. Interfaces* **16**, 46454–46460 (2024).

16. Kumar, M., Park, H. & Seo, H. A Single-Pixel Event Photoactive Device for Real-Time, In-Sensor Spatiotemporal Optical Information Processing. *Advanced Materials* **37**, 2406607 (2024).

17. Yamamoto, H. *et al.* Modular architecture facilitates noise-driven control of synchrony in neuronal networks. *Science Advances* **9**, eade1755 (2023).

18. Sinha, M. & Narayanan, R. Active Dendrites and Local Field Potentials: Biophysical Mechanisms and Computational Explorations. *Neuroscience* **489**, 111–142 (2022).

19. Yi, G., Wang, J., Wei, X. & Deng, B. Action potential initiation in a two-compartment model of pyramidal neuron mediated by dendritic Ca2+ spike. *Sci Rep* **7**, 45684 (2017).

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

21. Wang, C.-Y. *et al.* Gate-tunable van der Waals heterostructure for reconfigurable neural network vision sensor. *Sci. Adv.* **6**, eaba6173 (2020).

22. Yao, S. *et al.* Radar-Camera Fusion for Object Detection and Semantic Segmentation in Autonomous Driving: A Comprehensive Review. *IEEE Transactions on Intelligent Vehicles* **9**, 2094–2128 (2024).

23. Wu, Y. *et al.* CMOS-compatible retinomorphic Si photodetector for motion detection. *Sci. China Inf. Sci.* **66**, 162401 (2023).

24. Chen, G. *et al.* Event-Based Neuromorphic Vision for Autonomous Driving: A Paradigm Shift for Bio-Inspired Visual Sensing and Perception. *IEEE Signal Processing Magazine* **37**, 34–49 (2020).

25. Liu, L. *et al.* Computing Systems for Autonomous Driving: State of the Art and Challenges. *IEEE Internet of Things Journal* **8**, 6469–6486 (2021).

26. Kim, M.-K., Kim, I.-J. & Lee, J.-S. CMOS-compatible compute-in-memory accelerators based on integrated ferroelectric synaptic arrays for convolution neural networks. *Science Advances* **8**, eabm8537 (2022).

27. Wu, C. *et al.* Programmable phase-change metasurfaces on waveguides for multimode photonic convolutional neural network. *Nat Commun* **12**, 96 (2021).

28. Gorelick, L., Blank, M., Shechtman, E., Irani, M. & Basri, R. Actions as Space-Time Shapes. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **29**, 2247–2253 (2007).

29. Wu, Y. *et al.* A Spiking Artificial Vision Architecture Based on Fully Emulating the Human Vision. *Advanced Materials* **36**, 2312094 (2024).

Acknowledgements

This work was supported by the National Key Research and Development Project of China (2023YFB2806701 W.D.), the National Natural Science Foundation of China under Grant (U23A20357 Y.Z., 62334001 Y.Z., 62305013 W.D. and 62574019 W.D.), and the China National Postdoctoral Program for Innovative Talents (No. BX20230033 W.D.).

Author contributions

Y.Z., W.D., Y.W. and Y.C. conceived the concept and designed the experiments. W.D., Y.C. and Y.Z. supervised the project. Y.W. fabricated the devices. Z.C., Y.W., R.L., C.X., Z.L. and Y.S.W. design weight control circuits. Y.W., J.G., C.Z., Q.R., X.M., Z.X, and Z.Z. performed the optoelectronic measurements. Y.W., D.W., K.L., X.W., Z.C., Y.C., and Y.Z. analyzed the data. Y.W. and W.D. wrote the paper. All the authors discussed the results and implications and reviewed the paper.

Competing interests

The authors declare no competing interests.

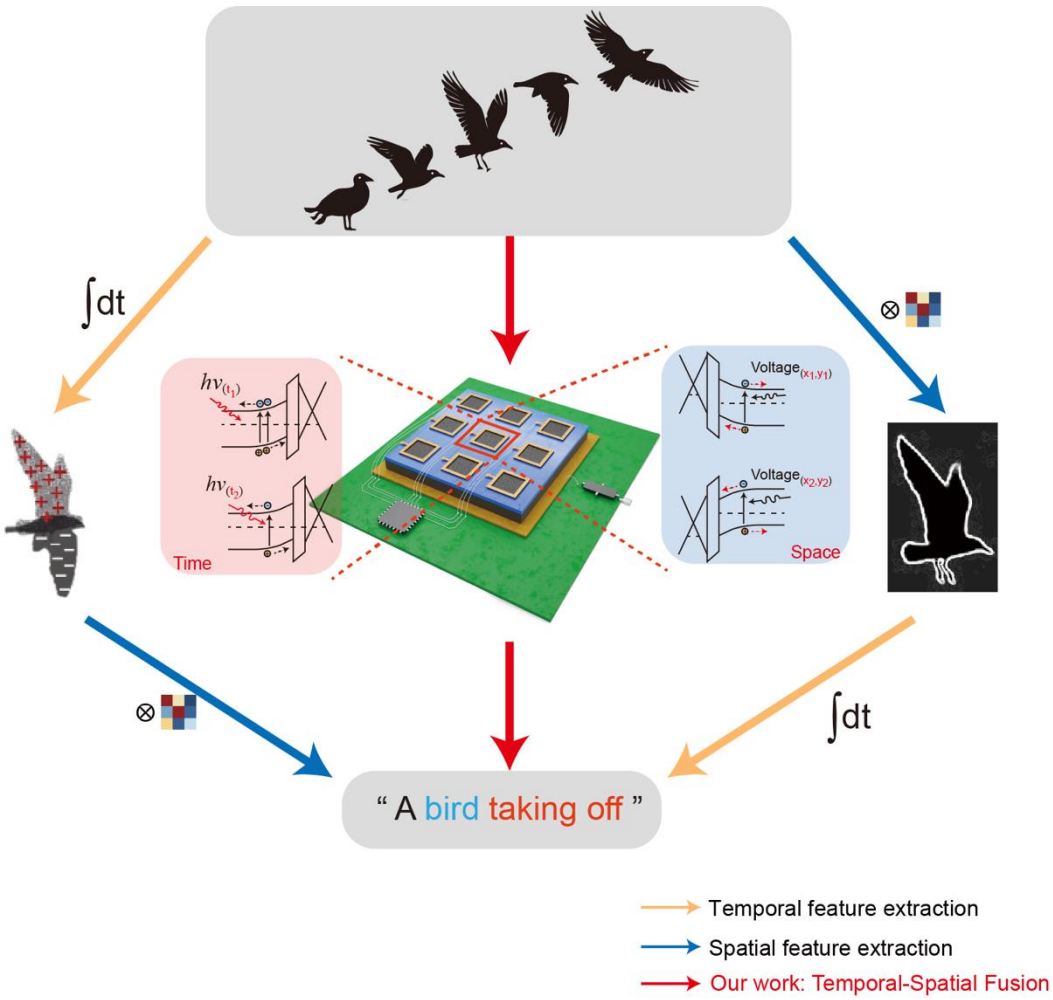

Figure 1. Schematic diagram of temporal-spatial fusion artificial vision system for vision information processing. In the same pixel, temporal information processing is achieved by photodetectors sensitive only to changes in light intensity over time., while spatial information processing is realized by voltage-modulated photoresponse, thereby accomplishing temporal-spatial fusion visual information processing. Traditional methods, instead, require the sequential processing of temporal and spatial information.

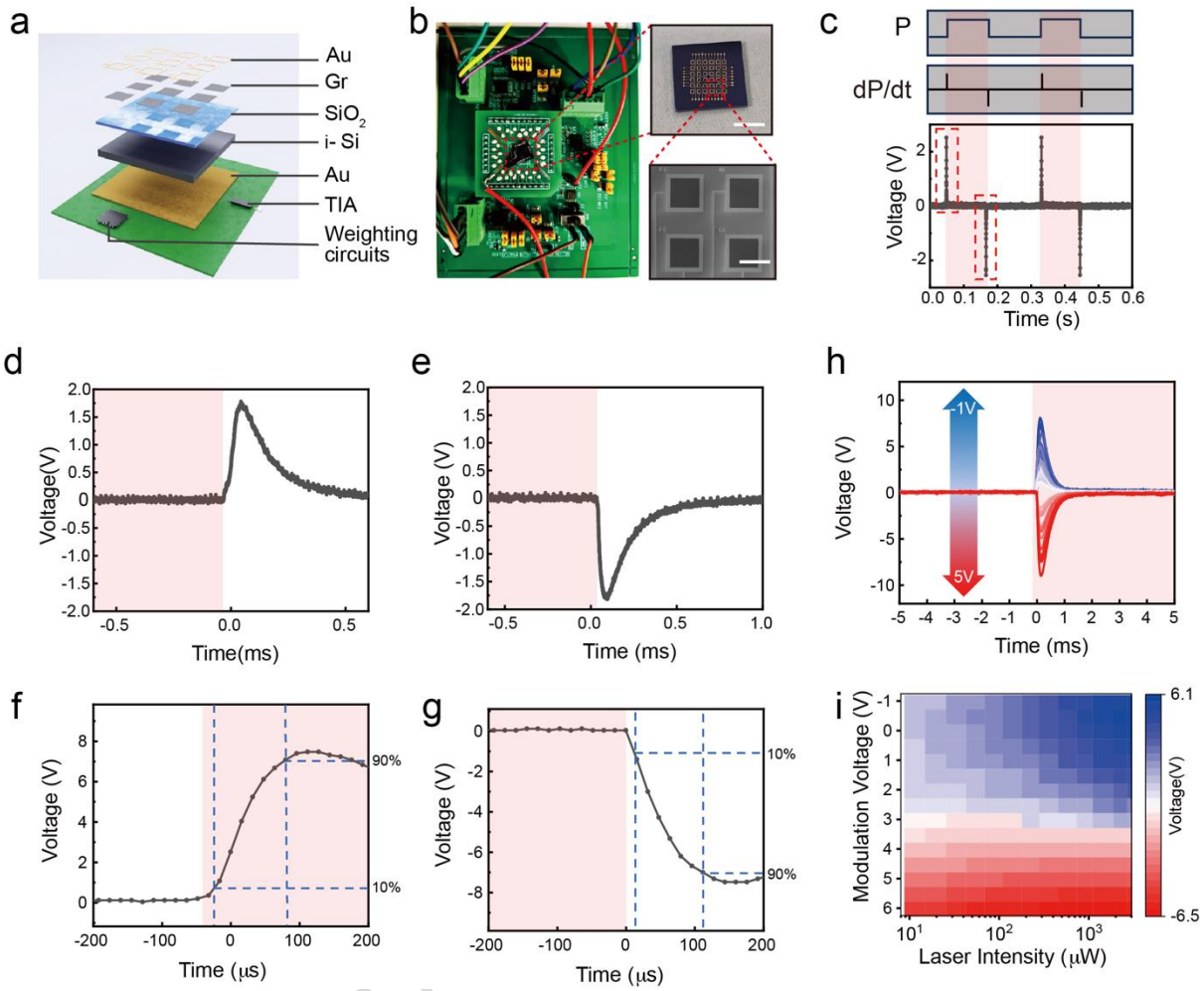

Figure 2. The temporal information processing characteristics of TSF-AVH. a) Schematic of TSF-AVH. b) Photograph of the assembled TSF-AVH, and magnified image of the pixels array. Scale bar: 5mm (top-right), 500μm (bottom-right). c) The time differential photoresponse characteristics of TSF-AVH. The photoresponse of the TSF-AVH is matched to the differential of the light intensity against time, so the processing of the temporal features of the light intensity can be directly implemented based on this property. d-e) Magnified view of the time - differential pulse signal. f-g) Time - resolved measurement characterizes the response speed of the TSF-AVH. h) Voltage-tunable differential photoresponse. i) Mapping of the voltage-dependent responsivity across three decades of optical power.

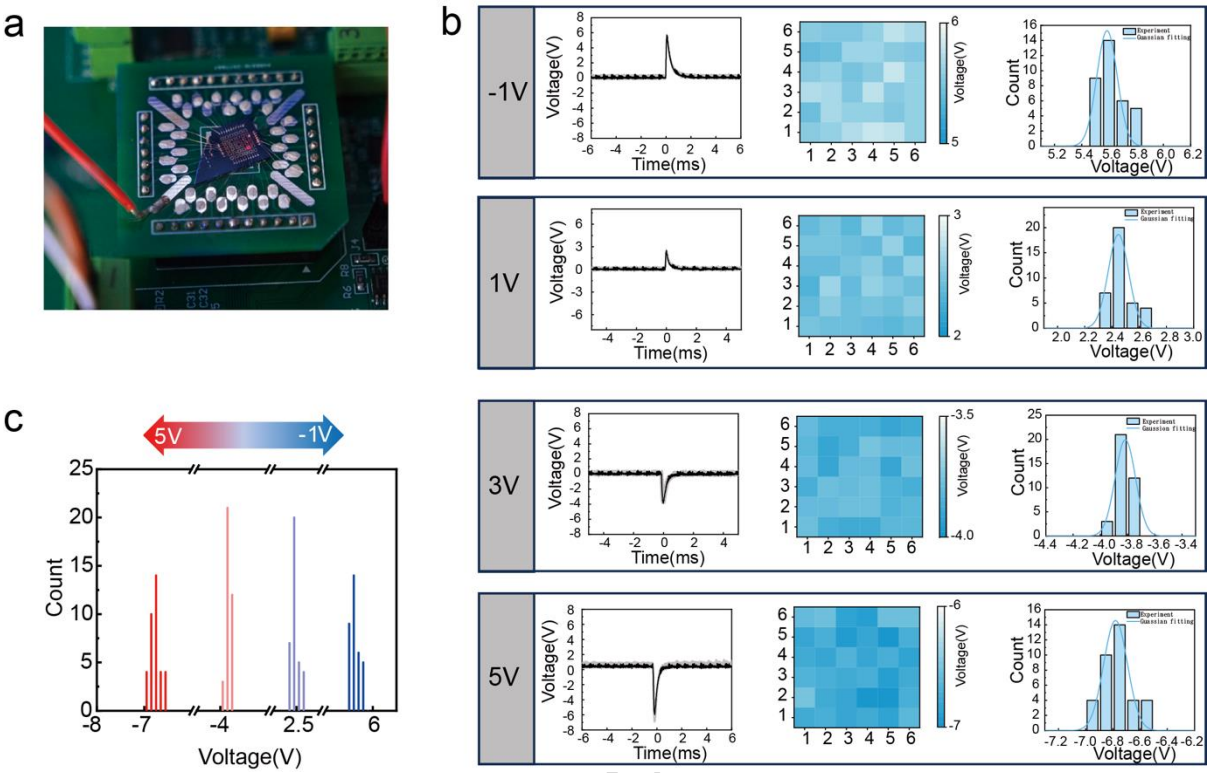

Figure 3. The variability of the TSF-AVH array. a) Image of the laser spot illuminating on one of the pixels of the TSF-AVH. b) Photoresponse characteristics for all 36 pixels. c) The statistical analysis graph of the photoresponse amplitude distribution indicates a voltage-dependent separation, with the intra-group consistency retained.

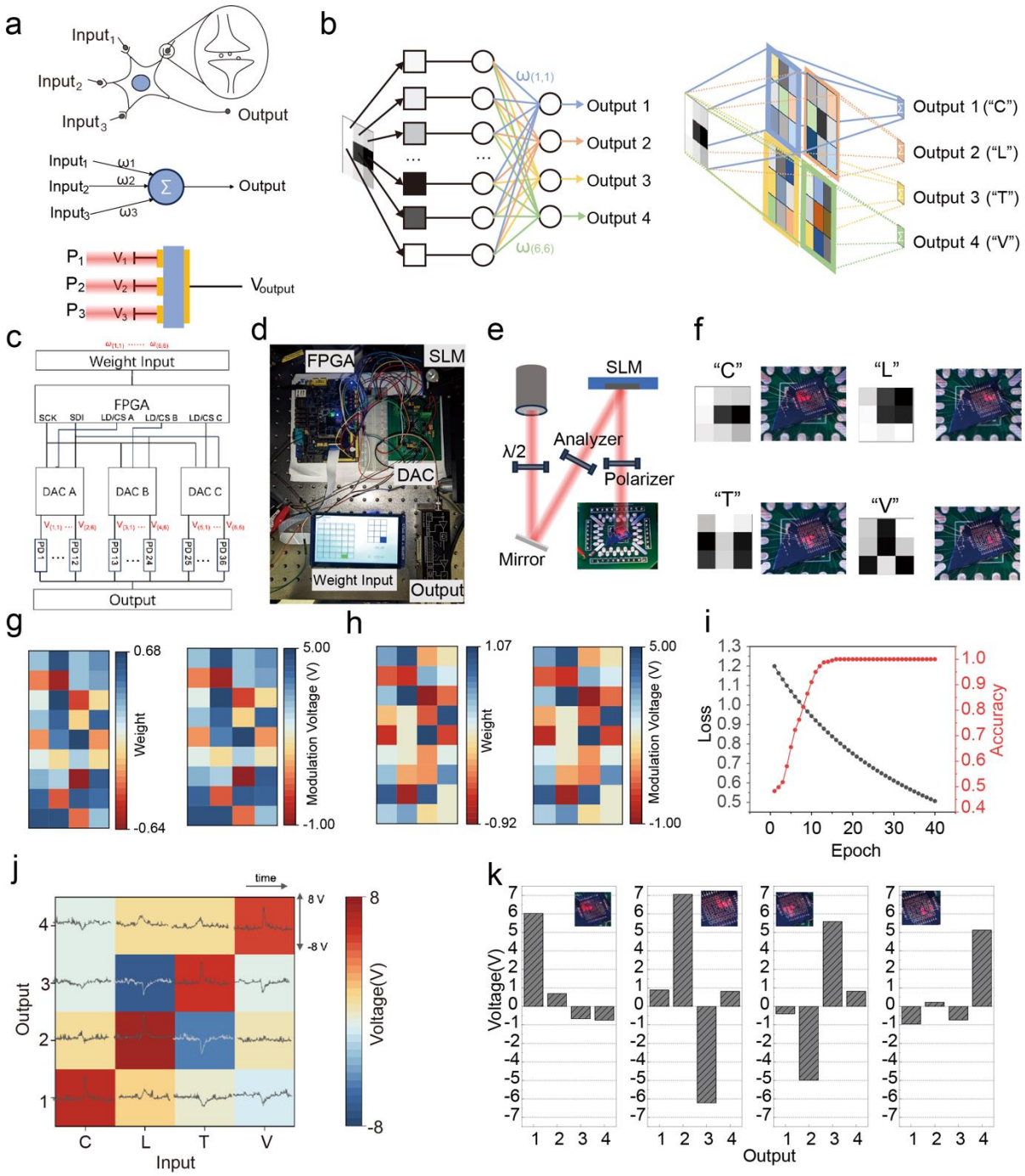

Figure 4. The spatial information processing characteristics of TSF-AVH. a) Schematic diagram of the analogy among biological neurons, mathematical models and device designs. b) Schematic diagram of the mechanism of spatial information processing. c) Schematic diagram of the weight control circuit. d) The image of the weight control circuit. e) Schematic diagram of the spatial light modulator. f) Image of four different letter patterns projected onto the photodetector

array. g-h) the weight distribution of the neural network before and after training and the corresponding voltage. i) Accuracy and loss during the training cycles. j-k) The output values of each channel when different patterns are projected onto the device.

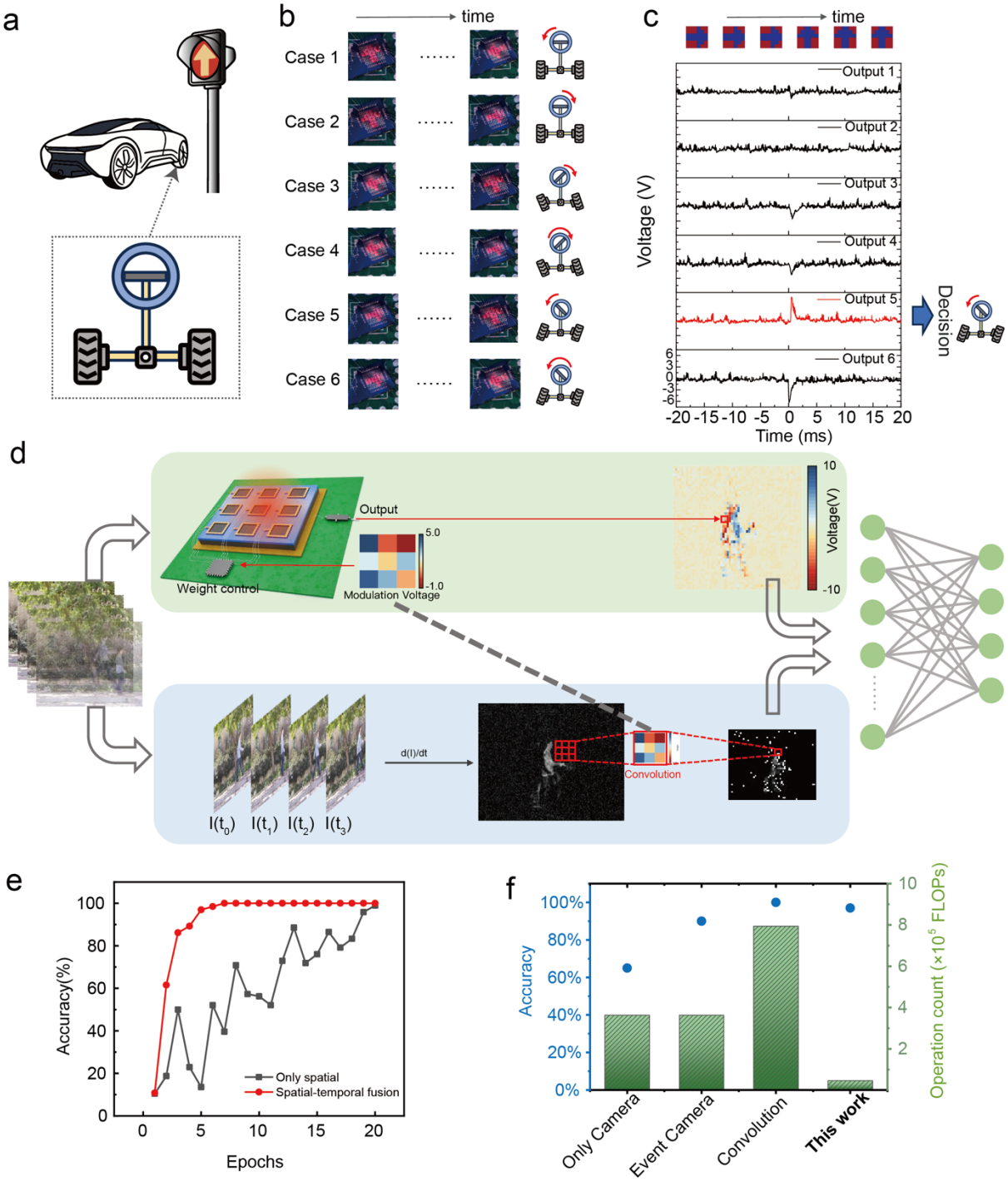

Figure 5. Temporal-spatial feature fusion for dynamic scene perception. (a) In a typical autonomous driving scenario, the vehicle passes through a signalized intersection and steers the steering wheel according to the indicator lights, (b) The pattern actually illuminated on the device and the corresponding steering wheel operation. c) For a certain scene, the photoelectric signals

output by 6 channels, by comparing the magnitude of the peak, the decision result can be obtained. All six output panels share the same y-axis label. d) Comparison of feature extraction methods between the Spatiotemporal Fusion Vision System and the Traditional Detector Vision System Strategies. e) Comparison of training epochs and accuracy rates during the recognition training process between the camera only and the spatiotemporal - fusion - based artificial vision system. f) Comparison of recognition accuracy and computing power among several different visual strategies.

Editor's summary:

Current artificial vision systems suffer from high energy consumption and latency due to the von Neumann bottleneck and separated spatiotemporal processing. Wu et al. propose a vision system achieving native spatiotemporal co-processing with millisecond latency and 95% action recognition accuracy at low computational cost.

Peer review information: *Nature Communications* thanks Hyeok Kim, Chengkuo Lee and Haotong Wei for their contribution to the peer review of this work. A peer review file is available.