A 2.54µJ/Frame Event-Driven Vision Sensor with Oculomotor-Adaptive Workload Allocation and Reconfigurable In-Pixel Feature Extraction for Eye-Tracking in AR/VR Applications

Jiakun Zheng, Fengshi Tian, Xihao Guan, Kwang-Ting (Tim) Cheng, Chi-Ying Tsui

Hong Kong University of Science and Technology (HKUST), Hong Kong SAR
AI Chip Center for Emerging Smart Systems (ACCESS), Hong Kong SAR

***Abstract*—This paper presents an event-driven vision sensor for eye-tracking task in AR/VR human-computer interaction (HCI) applications, with an oculomotor-adaptive scheme to match computation intensity with HCI needs and balance power-accuracy trade-off. The sensor dynamically distinguishes rapid saccades and stable fixations. During saccades, where ballistic shifts occur and coarse trajectory tracking is sufficient for HCI, dense 2D frames are projected into sparse 1D feature vectors, allowing a lightweight on-sensor neural network for coarse tracking. Upon detecting stable fixations that require precise gaze estimation, the system switches to an off-sensor high-precision mode, with in-pixel convolutional feature extraction to compress data before streaming out, achieving an 8x data reduction to mitigate the I/O workload. Furthermore, the sensor features an in-pixel motion-detection mode that gates all redundant processing for static scenes. Fabricated in 40nm CMOS, the prototype demonstrates the efficacy of dynamic workload switching, with a power of 52.8µW for always-on in-pixel motion-detection and an energy efficiency of 2.54µJ/frame for coarse tracking using on-sensor processor.**

Keywords*— *Always-on sensing, in-sensor computing, event-driven sensor, energy-proportionality architecture, eye-tracking

I. Introduction

The evolution of wearable AR/VR devices has driven eye-tracking an emerging method for natural human-computer interaction (HCI). However, deploying high-precision eye-tracking algorithms on-edge remains challenging due to the high energy cost of continuously capturing, transferring, and processing high-frame-rate eye images under the strict thermal and battery constraints of smart glasses [1].

Conventional frame-based sensor–processor architecture suffers from a severe data movement bottleneck, where energy consumption is dominated by the continuous transmission of pixel data from the camera to the processor [2]. Furthermore, continuous sampling the eye during periods of inactivity consumes unnecessary system resources.

To alleviate this, event-based sensors have been proposed to achieve data sparsity by capturing only intensity changes [1, 2]. While efficient for tracking fast movements, these event-driven approaches struggle during slow ocular movements and static fixations, as the sparse pixel-level deltas become indistinguishable from noise, leading to a significant degradation in localization accuracy [3].

Eye movements involve rapid saccades (ballistic shifts between points) and stable fixations (focus on specific targets) [4]. During saccades, the human visual system undergoes saccadic suppression, meaning the system only requires coarse trajectory tracking for predictive rendering. Conversely, during fixations, the user is actively interacting with the interface, necessitating high-precision gaze estimation [4]. However, neither frame-based nor event-based designs make use of this physiological characteristic.

This work was partially supported by ACCESS – AI Chip Center for Emerging Smart Systems, sponsored by InnoHK funding, Hong Kong SAR.

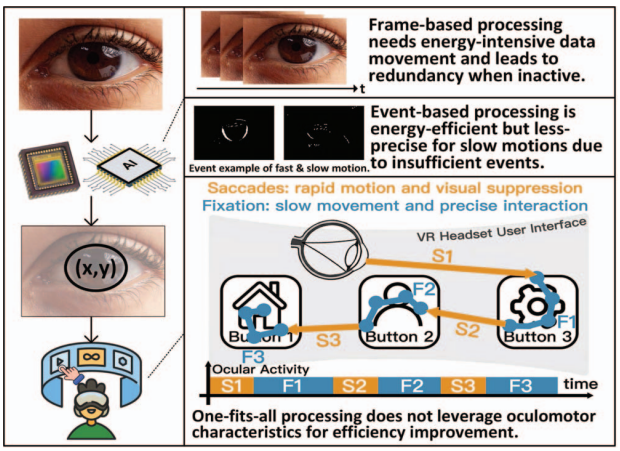


Fig. 1(a). Current solutions and challenges of eye-tracking systems.

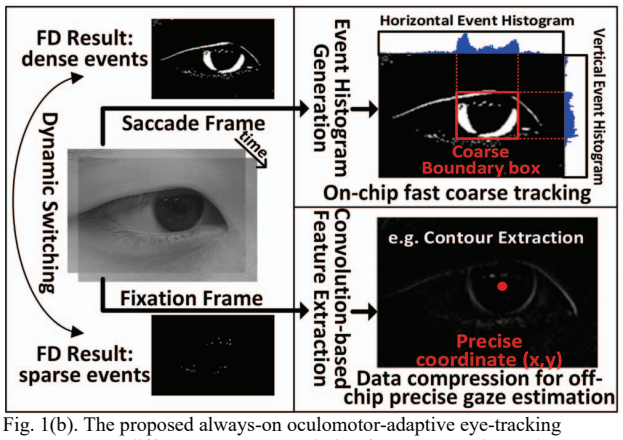


Fig. 1(b). The proposed always-on oculomotor-adaptive eye-tracking system. Frame-difference (FD), convolution feature extraction and histogram generation are all performed in the sensor pixel array.

To address these challenges, we propose an intelligent vision sensor specifically optimized for ultra-low-power, always-on eye tracking. Our design features:

1. A hardware-efficient CMOS imager circuit that achieves in-pixel frame-difference (FD) for motion detection and convolution-based feature extraction for compressing data before off-chip transmission, with minimal modifications to a standard 3-transistor (3-T) active pixel sensor (APS), incurring negligible area overhead.

2. An event-triggered oculomotor-adaptive system architecture, which leverages an in-pixel motion-event detection circuit for dynamic transition between (a) a low-power on-sensor event-based coarse tracking mode during saccades and (b) a frame-based off-sensor high-precision mode with in-pixel feature-extraction for data compression during fixations.

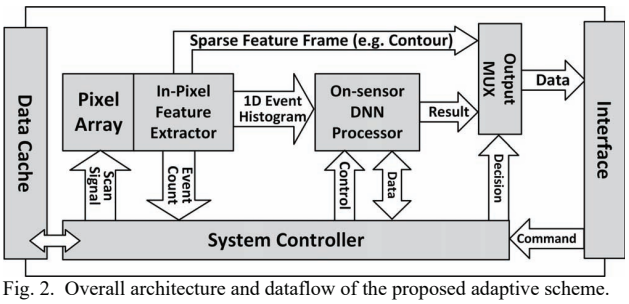


Fig. 2. Overall architecture and dataflow of the proposed adaptive scheme.

II. System Architecture and Dataflow

In AR/VR HCI, eye movements primarily alternate between rapid saccades and stable fixations. As explained in Fig. 1, the proposed design implements a dynamic workload allocation strategy to exploit these distinct physiological phases.

During high-velocity saccades, the in-pixel FD circuit inherently exploits spatio-temporal correlation by capturing pixel-level intensity deltas between consecutive frames. Given that saccades are transient and typically do not involve precise interaction, high-precision localization is computationally redundant. Thus, when high FD event density is detected, the chip directly utilizes the FD result to do eye-tracking. As discussed in [5], the FD result can be used to efficiently locate the region of interest (ROI), by calculating the event numbers in the frame horizontally and vertically, which equivalently projects the 2D frame into two 1D event histogram vectors. By compressing the spatial data into 1D, the computational complexity is drastically reduced, allowing a simple Multi-Layer Perceptron (MLP) network running on the on-sensor neural network processor to determine coarse pupil locations. This dimensionality reduction ensures trajectory continuity with minimal energy consumption, bypassing the need for high-precision, power-hungry 2D processing during rapid motion.

However, the efficacy of FD-based sensing is inherently tied to the frame rate and degrades significantly during slow ocular movements. Crucially, these slow movements correspond to fixations, where high-precision localization is mandatory for accurate HCI. To bridge this gap, when the event count falls below a predefined threshold, the system switches to a frame-based high-precision mode using an off-sensor convolution neural network (CNN). To circumvent the energy-hungry transmission of full raw frames to the off-chip processor, the sensor leverages its in-pixel convolution circuit to perform feature extraction before streaming out frames. By applying Sobel-like operators and thresholding to extract salient contour information (e.g., pupil-iris boundaries) directly at the sensing frontier, the raw data is compressed into a binary sparse feature map, leading to an 8x data reduction compared to an 8-bit full precision readout, with only minor information loss.

When the scene remains static (no event generated), meaning an inactive ocular state, fixation mode is applied but only the last settled frame is processed, after which only the FD circuit is working as a trigger to detect the next motion event, suppressing redundant processing for inactive frames.

III. Circuit Design and Implementation

As shown in Fig. 3, the proposed sensor chip consists of an array of reconfigurable pixel groups (RPGs), column/row-parallel multi-functional readout units (MFRUs), decoders and drivers, and a digital deep neural network (DNN) processor with digital computing-in-memory (DCIM) macros and controllers. A block of on-chip 1-Mb SRAM is used as DNN weight memory. The pixel circuit is based on an APS structure, with additional capacitors as memory for in-pixel storage and computation. An RPG employs a binning design that makes the adjacent 4 pixels share two capacitors and two source followers (SFs), supporting both low-resolution feature extraction and full-resolution readout while reducing area overhead, resulting in 3.5-T per pixel and a fill-factor (FF) of 85% with 7.9-um pixel pitch. An MFRU consists of two binary-weighted 9-bit switched-capacitor network (SCN) and a sense-amplifier (SA), which is a typical SAR ADC structure. Each 9-bit SCN is designed to be three C-2C-4C segments, which is not only for reducing area but also for multiple weighted samplings that equivalently multiplies the input signals with weights, as explained in Figs. 4-7. The MFRU can be configured in different modes to achieve not only SAR ADC but also multiplication and summation through weighted sampling and charge-redistribution. The column/row-parallel MFRUs with corresponding decoders and drivers wrap at the four sides of the RPG array to allow parallel feature extraction from vertical and horizontal directions. With this RPG and MRFU circuit reconfigurability, the sensor can work in the following modes.

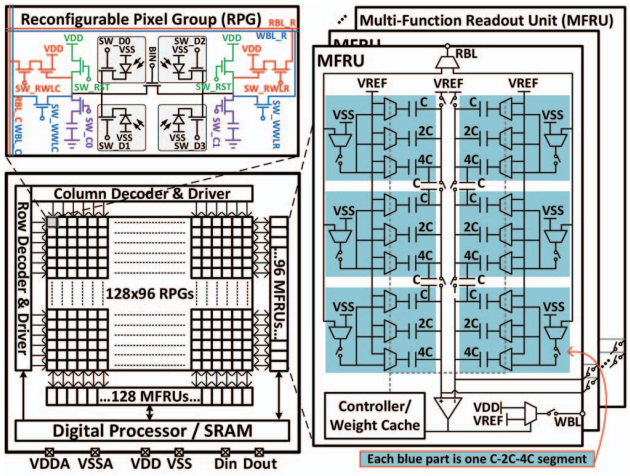


Fig. 3. Block diagram of the proposed sensor, and the circuit design of the pixel array and the readout unit.

A. Programmable Convolution

Fig. 4 shows the working flow of a convolution-based feature extraction. Like traditional 3T-APS, the RPG uses photodiodes to convert light intensity to voltage signals, which are buffered by SFs and sampled to the MFRUs. The capacitors in the RPGs can store the exposure results of the pixels. The SCNs in the MFRUs are configured to sample the exposure and reset voltages of different rows/columns with different weights, which is equivalently a multiplication between the analog input signal from RPG and a 4-bit weight (signed-magnitude). The weight magnitudes are programed by multiplexing different numbers of capacitors in each C-2C-4C segment. As later conversion reads out the differential voltage of the left side or right side of the 9-bit SCNs, depending on the sign-selection mux, reset and exposure signals are sampled to the left side or right side of the 9-bit SCNs, resulting in a $V_{rst} - V_{exp}$ or $V_{exp} - V_{rst}$, effectively

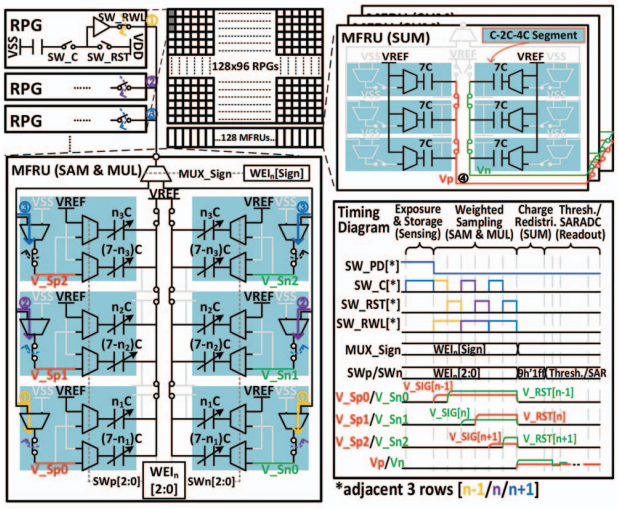


Fig. 4. The MFRU is configured to do sampling (SAM) with multiplication (MUL), and charge-redistribution-based summation (SUM), followed by a thresholding readout (1-bit SAR ADC). Timing diagram shows an example MAC process of one 3x3 patch, with kernel size=3 and stride=1.

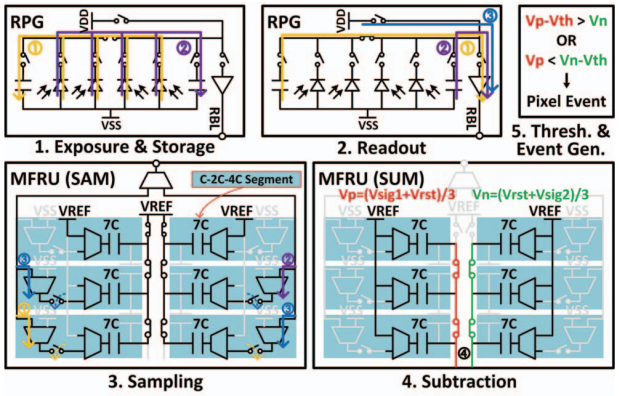


Fig. 5. Always-on motion detection using frame-difference for one row. Sampling process is like MAC with weights set to +1 and -1, followed by a thresholding process. This process is repeated across all rows/columns to generate a binary event frame.

programs the signs of weights. A multiplication-and-accumulation (MAC) is done through this multiplication followed by a charge redistribution across the capacitor segments within and with adjacent MFRUs, which averages out the voltages and is equivalently a summation. After MAC operations, the SCNs can perform SAR ADC conversion by switching the capacitors, with up to 9-bit resolution. A thresholding operation can be achieved by applying a 1-bit SAR ADC. The timing diagram in Fig. 4 illustrates that n times of sampling is required to achieve the kernel sliding, where n is the kernel size. Nonetheless, this process is pipelined with rolling-shutter exposure time, incurring no latency overhead compared to a normal readout.

B. Always-on Motion-Event Detection

Fig. 6 explains the configuration of motion detection mode. The two in-pixel capacitors are used as analog memory cells to store the photovoltages of two consecutive frames, which are sampled into the MFRUs like the MAC operation with weights programed to be +1 and -1 (equivalently a subtraction between two frames), followed by a thresholding to output a binary event, indicating whether there is a significant intensity change between two frames in that pixel or not. This process is repeated for all rows, generating an event frame for later event density calculation.

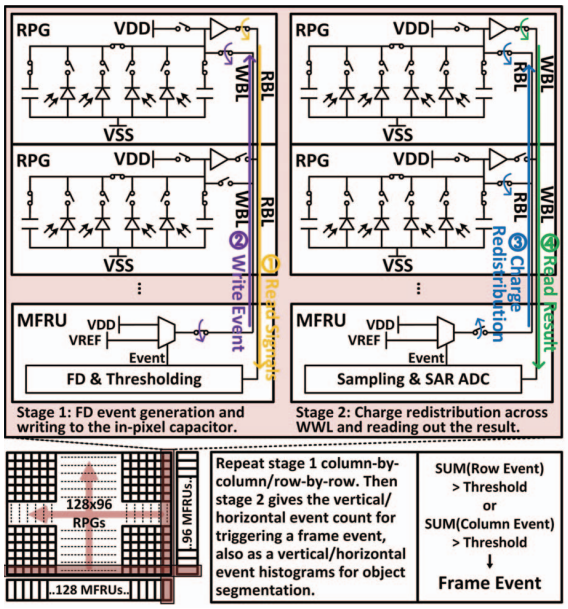


Fig. 6. RPG and MFRU work as row/column event counters for triggering dynamic switch between saccade and fixation modes, as well as generating event histograms for estimating the positions of moving objects.

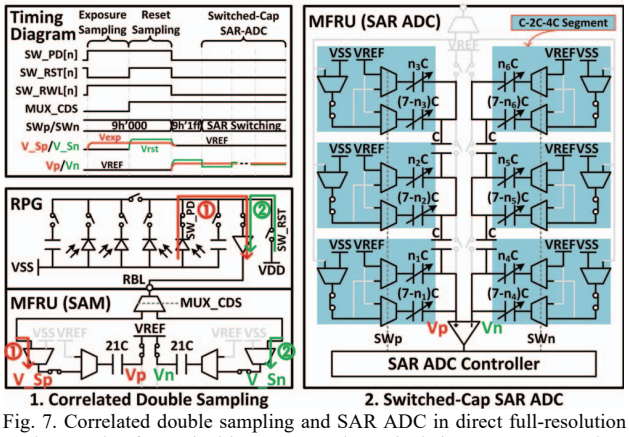


Fig. 7. Correlated double sampling and SAR ADC in direct full-resolution readout mode of one pixel in one row. The 4 pixels in an RPG are readout sequentially, and the process is repeated for all rows, resulting in a full-resolution (256x192) frame.

C. Event Count and Horizontal/Vertical Event Histogram

As shown in Fig. 6, after each pixel subtraction, the events generated by MFRUs are written back into the pixel capacitor through the write-bit-line (WBL), followed by a charge redistribution on the WBL, which averages out the voltage on all the in-pixel capacitors connected on the line. The eventual voltage on the WBL is proportional to the number of events triggered on that line of pixels. Reading out this event count row/column-wise gives the horizontal and vertical event-count histograms. This histogram feature is then feed into the on-sensor Multi-layer Perception (MLP) processor for energy-efficient coarse tracking.

D. Full-Resolution Readout

The RPG and MFRU can also read out the raw frame of the pixel array like a typical 3T-APS with SAR ADC. Fig. 7 shows the timing diagram and circuit configuration of such full-frame readout. In this mode, the SCN and SA work as a SAR ADC for sampling, holding and conversion. The conversion reads out the differential voltage, resulting in a correlated double sampling (CDS) to suppress offset and reset noise [6]. CDS also applies in other modes mentioned above.

E. *On-sensor DNN Inference*

The sensor also features a DCIM-based DNN processor to process the sparse 1D histogram feature vector with a lightweight MLP network to perform the coarse tracking. This near-sensor processing eliminates the need for off-chip data movement and thus is energy-efficient.

IV. Experiment Setup and Measurement Results

1. Network Training and Performance Simulation

To ensure the neural networks are optimized for the sensor's outputs, two specialized neural networks are trained using the Labelled Pupils in the Wild (LPW) dataset [7]. To emulate the in-pixel feature extraction circuits of the sensor output (histogram and contour features), behavioral simulation modules are integrated into the network front-ends. Specifically, the on-sensor MLP network is trained on 1D histogram vectors derived from simulated FD operations, while convolution with Sobel-filter is applied to mimic the in-pixel contour extraction before the off-sensor CNN.

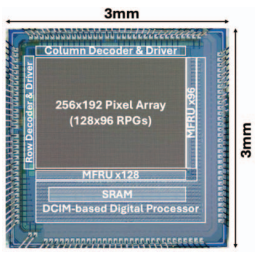
These behavioral simulations utilize temporal and spatial gradients rather than raw pixel values, reducing dependence on absolute light intensity and thus mitigates performance degradation caused by sensor parameter mismatches between the original dataset's acquisition hardware and the fabricated prototype. Training result shows an average pupil-center error of 12.53 pixels for saccades and 4.65 pixels for fixations with a frame size of 128x96. The fixation mode error is nearly on par with the 4.52-pixel error of raw-frame software baseline, showing that the proposed data compression strategy significantly reduces I/O bandwidth and energy consumption without significantly compromising the performance.

TABLE I. SUMMARY OF THE TRAINED MODELS

Model	On-sensor event-based coarse	Off-sensor feature-based high-accuracy	Software baseline
Structure	3-layer MLP	MobileNetV3	MobileNetV3
Parameter #	250K	1.08M	1.08M
MAC #	250K	14.52M	14.52M
Input Feature & Dimension	1D event histogram vector	2D binary contour feature map	2D grayscale full frame
	[224, 1]@8bit	[128, 96]@1bit	[128, 96]@8bit
Average Err.	12.53 pixels	4.65 pixels	4.52 pixels

2. Chip Measurement Result

The proposed sensor chip is fabricated in TSMC 40-nm CMOS process. Measurement result shows a power consumption of 52.8µW for always-on motion detection and an energy efficiency of 2.54µJ/frame for on-sensor coarse tracking. As summarized in Table II, this work maintains superior energy-proportionality through its oculomotor-adaptive architecture, achieving a favorable energy efficiency compared to prior works.



Process	**40nm**
Area	**9mm²**
Analog Voltage	**1.1 V**
Digital Voltage	**0.8 V**
Frequency	**50 MHz**
Power	**52.8uW @30FPS FD**
On-Chip SRAM	**1 Mb**

Fig. 8. Die photo and specs of the chip.

V. CONCLUSION

This paper presents an event-driven oculomotor-adaptive vision sensor for eye-tracking tasks. By dynamically switching between an always-on motion detection mode, an on-sensor efficient saccade tracking, and an off-sensor high-precision fixation mode with data compression, the architecture and circuit design optimize the energy-accuracy trade-off, providing an energy-proportional solution for AR/VR HCI applications.

TABLE II. COMPARISON WITH PRIOR WORKS

Reference	VLSI 2015 [8]	VLSI 2022 [9]	TCAS-I 2025 [1]	ASP-DAC 2026 [2]	This Work
System	Frame-based + Ellipse fitting	Frame-based + CNN	Event-based + CNN	Event-based + CNN	**Oculomotor-adaptive Feature Extraction + MLP/CNN**
Sensor Integration	Yes	Yes	No	No	**Yes**
Processing Method	In-sensor	Near-sensor	Near-sensor	Near-sensor	**In-pixel**
Technology	65 nm	28 nm	40 nm	12 nm	**40 nm**
Voltage	1.2-3.3 V	0.51-0.88 V	0.99 V	0.8 V	**1.1V (A); 0.8V (D)**
Clock	50 MHz	90-370 MHz	200 MHz	400 MHz	**50 MHz**
Dataset	/	OpenEDS	EVBEYE	3ET+	**LPW**
Resolution	320 x 240	640 x 400	120 x 220 96° x 64°	80 x 60	**128 x 96**
Average Error	1 px / 0.5°	3.16°	18.5 px / 0.91°	2.45 px	**12.53 px (coarse) 4.65 px (high-acc.)**
Frame Rate	30 FPS	253 FPS	1252 FPS	2000 FPS	**30 FPS (Limited by PD sensitivity)^**
Energy Efficiency	333.33* uJ/Frame	91.49 µJ/frame	12.7 µJ/frame	18.9 µJ/frame	**2.54 µJ/frame (on-chip coarse result)**
					10.19 µJ/frame (off-chip high-acc.)#

***Derived from frame rate and power consumption.**
#Simulated on DCIM-based DNN processor in 40-nm process.
^Different from processor-only design, integrated sensor pipelines sensing with processing, resulting in limited FPS due to long exposure time required by process-dependent photodiode.

VI. REFERENCES

[1] S. Tan et al., "Toward Efficient Eye Tracking in AR/VR Devices: A Near-Eye DVS-Based Processor for Real-Time Gaze Estimation," in IEEE Transactions on Circuits and Systems I: Regular Papers, vol. 72, no. 10, pp. 5748-5760, Oct. 2025, doi: 10.1109/TCSI.2025.3553497.

[2] T. Han, A. Li, Q. Chen and C. Gao, "JaneEye: A 12-nm 2K-FPS 18.9-µJ/Frame Event-based Eye Tracking Accelerator," 2026 31st Asia and South Pacific Design Automation Conference (ASP-DAC), Lantau, Hong Kong, 2026, pp. 170-176, doi: 10.1109/ASP-DAC66049.2026.11420815.

[3] Q. Chen, Z. Wang, S. -C. Liu and C. Gao, "3ET: Efficient Event-based Eye Tracking using a Change-Based ConvLSTM Network," 2023 IEEE Biomedical Circuits and Systems Conference (BioCAS), Toronto, ON, Canada, 2023, pp. 1-5, doi: 10.1109/BioCAS58349.2023.10389062.

[4] R. Krueger, S. Koch and T. Ertl, "Saccadelenses: interactive exploratory filtering of eye tracking trajectories," 2016 IEEE Second Workshop on Eye Tracking and Visualization (ETVIS), Baltimore, MD, USA, 2016, pp. 31-34, doi: 10.1109/ETVIS.2016.7851162.

[5] J. Acharya et al., "EBBIOT: A Low-complexity Tracking Algorithm for Surveillance in IoVT using Stationary Neuromorphic Vision Sensors," 2019 32nd IEEE International System-on-Chip Conference (SOCC), Singapore, 2019, pp. 318-323, doi: 10.1109/SOCC46988.2019.1570553690.

[6] T. Ma et al., "Systematic Methodology of Modeling and Design Space Exploration for CMOS Image Sensors," in IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, vol. 45, no. 2, pp. 1047-1060, Feb. 2026, doi: 10.1109/TCAD.2025.3585753.

[7] M. Tonsen, X. Zhang, Y. Sugano, and A. Bulling, "Labeled pupils in the wild: A dataset for studying pupil detection in unconstrained environments," arXiv, Nov. 18, 2015, arXiv:1511.05768.

[8] K. Bong, I. Hong, G. Kim and H. -J. Yoo, "A 0.5-degree error 10mW CMOS image sensor-based gaze estimation processor with logarithmic processing," 2015 Symposium on VLSI Circuits (VLSI Circuits), Kyoto, Japan, 2015, pp. C46-C47, doi: 10.1109/VLSIC.2015.7231321.

[9] Y. Zhao et al., "i-FlatCam: A 253 FPS, 91.49 µJ/Frame Ultra-Compact Intelligent Lensless Camera for Real-Time and Efficient Eye Tracking in VR/AR," 2022 IEEE Symposium on VLSI Technology and Circuits (VLSI Technology and Circuits), Honolulu, HI, USA, 2022, pp. 108-109, doi: 10.1109/VLSITechnologyandCir46769.2022.