From Retina to Silicon: The IRIS Framework for Bio-inspired Visual Intelligence

Subhradip Chakraborty*
University of Wisconsin - Madison
Madison, Wisconsin, USA
chakrabort42@wisc.edu

Shay Snyder*
George Mason University
Fairfax, Virginia, USA
ssnyde9@gmu.edu

Maryam Parsa
George Mason University
Fairfax, Virginia, USA
mparsa@gmu.edu

Gregory Schwartz
Northwestern University
Evanston, Illinois, USA
mparsa@gmu.edu

Akhilesh Jaiswal
University of Wisconsin - Madison
Madison, Wisconsin, USA
akhilesh.jaiswal@wisc.edu

Abstract

Traditional frame-based vision systems are ill-suited for extreme-edge intelligence due to high bandwidth, latency, and power overheads. The IRIS framework overcomes these limits through a unified neuroscience, software, and hardware co-design. Inspired by mammalian retinal circuitry, IRIS embeds multiple key visual functions: object motion sensitivity, looming detection, and motion prediction, directly into the image sensor. Using spatio-temporal filtering and predictive coding, it enables efficient mixed-signal CMOS implementations. Leveraging 3D-integration scheme, IRIS delivers real-time, event-driven feature extraction at lower latency, power, and complexity, enabling sensor-level bio-inspired visual intelligence. Our results highlight that IRIS maintains 98% feature fidelity while integrating three individual visual functions and reducing energy consumption by 2.41x.

Keywords: Retina inspired sensor, Motion prediction, Object motion sensitivity, Looming detection, Neuromorphic sensor, 3D integration.

ACM Reference Format:
Subhradip Chakraborty, Shay Snyder, Maryam Parsa, Gregory Schwartz, and Akhilesh Jaiswal. 2026. From Retina to Silicon: The IRIS Framework for Bio-inspired Visual Intelligence. In *Proceedings of On-Sensor Vision Workshop (OSV '26).* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 Introduction

The retina performs visual processing at the earliest stage of perception, transforming photoreceptor inputs into structured representations of looming threats, motion prediction, and other features [8]. By integrating perception and processing within the same substrate, the retina achieves unmatched levels of efficiency [1]. In contrast, traditional frame-based vision pipelines separate sensing and computation, relying on power-intensive sensors, high-bandwidth data transfer, and computationally expensive post-processing [13]. Recent neuromorphic sensors, such as Dynamic Vision Sensors (DVS), were developed to mimic the retina's asynchronous sampling [4]. However, even these lack embedded feature computations resulting in greater data redundancy and heavier downstream processing. DVS sensors are an important step toward realizing retina-inspired vision in cameras, yet their functionality is limited to the outer retina - specifically the first retinal synapse, while excluding the complex computations performed by inner retinal layers. Furthermore, biological retina extract diverse decision-critical spikes through parallel retinal circuits, which is in sharp contrast to DVS cameras that focus only on the change-detection aspect of bipolar cells. The lack of real-time extraction of context-appropriate, decision-critical features in current DVS cameras calls for modernizing the symbiotic relationship between retinal neuroscience, software, and hardware to enable next generation of retina-inspired cameras with wide application to computer vision, specifically for high-speed, bandwidth constrained autonomous maneuvering.

To address these limitations, we introduce Integrated Retinal Functionality in Image Sensors (IRIS), a neuroscience-software-hardware co-design framework that embeds retinal computations directly into image sensors. IRIS translates three retinal features (and can be scaled to include more features) into algorithmically tractable models and corresponding CMOS circuit implementations. The following sections outline the biological foundations, hardware implementations, and algorithmic models of each retinal computation. We show that IRIS preserves 98% feature fidelity while integrating three distinct visual functions and achieving a 2.41x reduction in energy consumption.

*Both authors contributed equally to this research.

OSV '26, Denver, CO, USA
2026. ACM ISBN 978-x-xxxx-xxxx-x/YYYY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

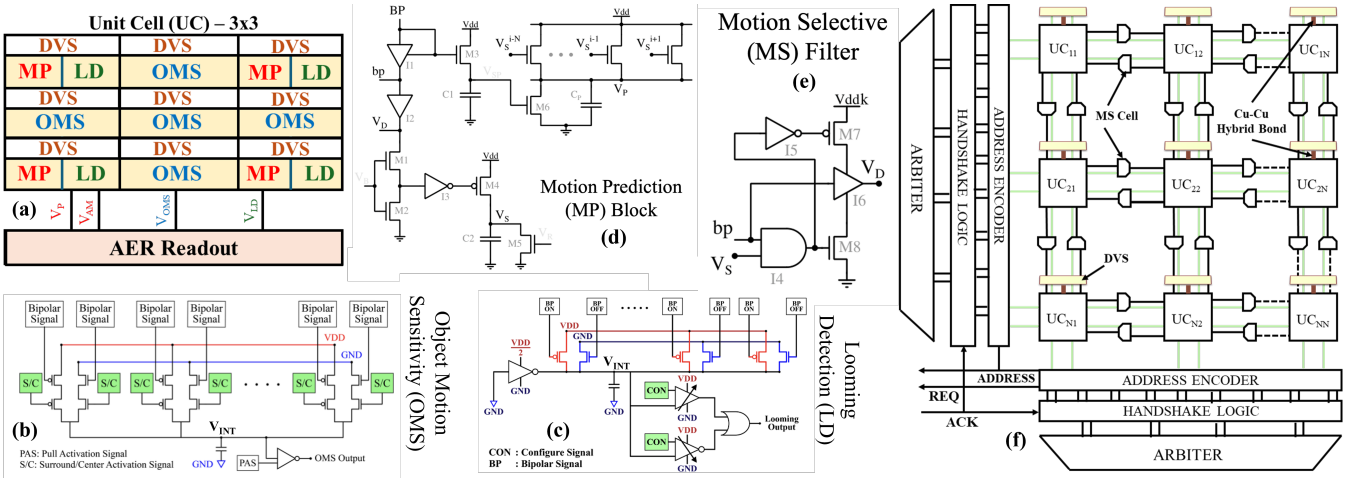

Figure 1. (a) Proposed unit cell (UC) integrating all retinal features (highlighted in yellow and implemented on a separate chip through 3D integration) together with a DVS pixel, interconnected via Cu–Cu hybrid bonding. Circuit schematics of (b) OMS, (c) LD, and (d–e) MP are also shown. (f) Two-dimensional array-level implementation of the UC employing an address-event representation (AER) interface.

2 Research Methods

Hardware Implementation: We implement circuit-level designs in the GF22 FDX technology to realize Object Motion Sensitivity (OMS) [14], Looming Detection (LD) [11], and Motion Prediction (MP) [2] functionalities. Each of these features is connected to the DVS pixel through 3D Cu–Cu hybrid bonding [5, 9, 12]. To enable the IRIS camera architecture, we propose a unit cell (UC), as shown in Figure. 1(a), which allows these features to be shared among multiple DVS pixels in an interleaved manner, rather than equipping each DVS pixel with all three features individually. This shared UC approach significantly reduces both power consumption, area overhead and enables selectivity of features.

OMS distinguishes object motion from global motion caused by self-movement (ego-motion). This mechanism relies on concentric center-surround receptive fields (RFs) where the RF center excites retinal ganglion cell (RGC) output. Amacrine cells (ACs) integrate contrast signals from the RF surround to inhibit RGC activity. The OMS algorithm models this interaction through two convolutional filters. Surround responses are subtracted from center responses before passing through a nonlinear activation [3]. The hardware maps OMS onto mixed-signal circuits that emulate bipolar cells (BCs), ACs and RGCs as shown in Figure. 1(b). The center and surround RFs are compared, suppressing global motion while emitting event spikes only for object motion [10]. OMS demonstrates how biologically grounded center–surround suppression can be implemented efficiently in hardware to isolate object motion with reduced parameter counts and computational complexity and consumes as little as 4.3 pJ per spike.

LD detects approaching objects, signaling potential collisions or threats. This arises from interactions between BCs, ACs, and RGCs, where an imbalance between ON/OFF BC activations produces transient RGC responses as objects expand within the RF [8]. The LD models this using spatial convolution filters and temporal differencing between ON/OFF activations. Their activations are subtracted and passed through a non-linear activation to generate binary responses for potential threats [11]. The hardware implementation uses paired PMOS/NMOS transistors to emulate ON/OFF bipolar behavior within RFs as shown in Figure. 1(c). Tunable thresholds and reconfigurable RFs enable adaptation across lighting and motion conditions with energy costs as low as 0.36 pJ per spike. LD demonstrates how biologically grounded excitatory–inhibitory dynamics can be realized in software and silicon for low-power, event-driven threat detection.

MP anticipates the trajectory of moving objects with a combination of biphasic filtering and non-linear activations. In biology, predictive responses arise from temporally offset excitatory and inhibitory pathways that allow RGCs to respond in advance of a moving object's true position [8]. This maintains a history of object positions while projecting their position forward in time. MP produces a predictive activation map aligned with the object's expected position. The hardware uses BCs that generate ON/OFF signals in response to luminance changes. These signals drive a MP circuit as shown in Figure. 1(d) that integrates past and current spikes to anticipate object positions and amplify motion-correlated responses using motion selective (MS) filter as shown in Figure. 1(e) [2]. Validated on real-world stimuli, the MP hardware demonstrates its ability to perform MP with low energy consumption, achieving just 15.8 pJ per prediction. These results highlight MP's potential in predictive computation for decision-making applications, including robotics and autonomous navigation.

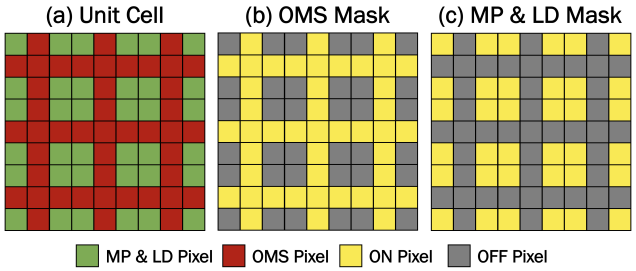

Figure 2. Simulating a 3x3 unit cell array with per-pixel masking. (a) The pixel-wise allocation of each pixel to the downstream feature. (b) A binary mask to filter events for OMS. (c) A binary mask to filter events for MP and LD.

The UC approach, as shown in Figure. 1(f), is further extended into a two-dimensional grid that integrates all the proposed features. This organization enables object motion sensitivity, looming detection, and motion prediction in a real-world neuromorphic camera that communicates using the address-event representation (AER) protocol.

Software Implementation: To algorithmically evaluate the proposed unit-cell, we emulate this behavior through event-level binary masking. Rather than modifying the underlying feature extraction algorithms, this approach isolates the effect of unit-cell information sharing. Events are selectively routed to downstream feature pipelines, closely mirroring the hardware-level multiplexing performed by the unit cell.

Events generated by the DVS pixel array are filtered using binary masks that assign each pixel's events to a single downstream retinal feature. These masks operate directly on the event stream, transmitting or blocking events on a per-pixel basis before feature computation. The layout of the masks are illustrated in Figure 2. This preserves the original temporal characteristics of the event stream while enforcing the unit-cell's constraints.

The masked events are processed using the original OMS, MP, and LD algorithms, as described in their respective works [2, 3, 10, 11]. This ensures that any observed differences in feature quality arise solely from the unit-cell sharing mechanism rather than algorithmic changes. By decoupling unit-cell masking from the downstream feature computations, we enable a fair comparison between the original implementations and the proposed unit-cell. The resulting features can therefore be quantitatively and qualitatively evaluated to assess the impact of unit-cell sharing on feature fidelity.

3 Results & Discussion

Hardware Implementation: The hardware results presented in Figure 3 are obtained using simulations in GF 22 nm FDX technology. Figure 3(a) shows the OMS input for an object moving relative to global motion, where the center and surrounding pixels generate spikes over time. When the accumulated signal reaches a predefined threshold, the OMS output node fires, as shown in Figure 3(e). Figures 3(b) and 3(c) illustrate the LD inputs for objects moving from left to right, right to left, and for an approaching (looming) object. The LD output spikes only when the object is approaching; in the other cases, the internal node voltage remains below the threshold, as shown in Figure 3(f). Figure 3(d) shows the MP input spikes for an object approaching from left to right. The predictive node fires based on the incoming object motion, as shown in Figure 3(h). If the prediction matches the actual object motion, an amplified spike is produced, as shown in Figure 3(g). The latency of the proposed retinal circuits is dominated by the DVS spike generation, which takes approximately 1 µs. With the proposed unit-cell approach, the power consumption is significantly reduced, since each unit cell integrates all features instead of implementing all features within every pixel. As a result, the overall energy consumption is reduced by a factor of 2.41.

Table 1. Quantitatively comparing the unit cells performance across each feature with respect to SSIM and energy savings.

	OMS	MP	LD
SSIM	0.98	0.97	0.99
Normalized Energy Savings	1.8	2.25	2.63

Software Implementation: To evaluate the resulting feature fidelity of the IRIS unit-cell, we leverage a series of video sequences from [7] depicting a ball approaching the observer and another with the ball translating horizontally. We refer to these sequences as sequence approaching S_A and sequence translation S_T. These sequences mimic the evolutionary purpose of OMS, LD, and MP where S_A enables the perception and prediction of object motion, whereas S_T enables the perception of looming threats. These pre-recorded video sequences were recorded at 60Hz with a resolution of 400×400. The DVS frames are generated through sequential frame differencing with a 10% threshold.

The quality of generated features is evaluated using the Structural Similarity Index Measure (SSIM) [6]. SSIM is a commonly used metric for assessing perceptual similarity in visual signals [6]. Because unit-cell multiplexing reduces the effective event rate observed by downstream features, the activation thresholds for OMS and LD are scaled linearly with respect to the event rate. This approach mirrors the adaptive firing thresholds used in spiking neural networks under variable input conditions and ensures comparable feature activation behavior [15]. As shown in Table 1, the unit-cell achieves consistently high feature fidelity across all three features with a mean SSIM of 0.98 across the S_A and S_T video sequences. These results indicate that performing unit-level OMS, MP, and LD computations introduces minimal distortion relative to the original per-feature implementations.

Figure 4 highlights the unit-cell's ability to capture OMS and MP on the S_T sequence. In both cases, the generated

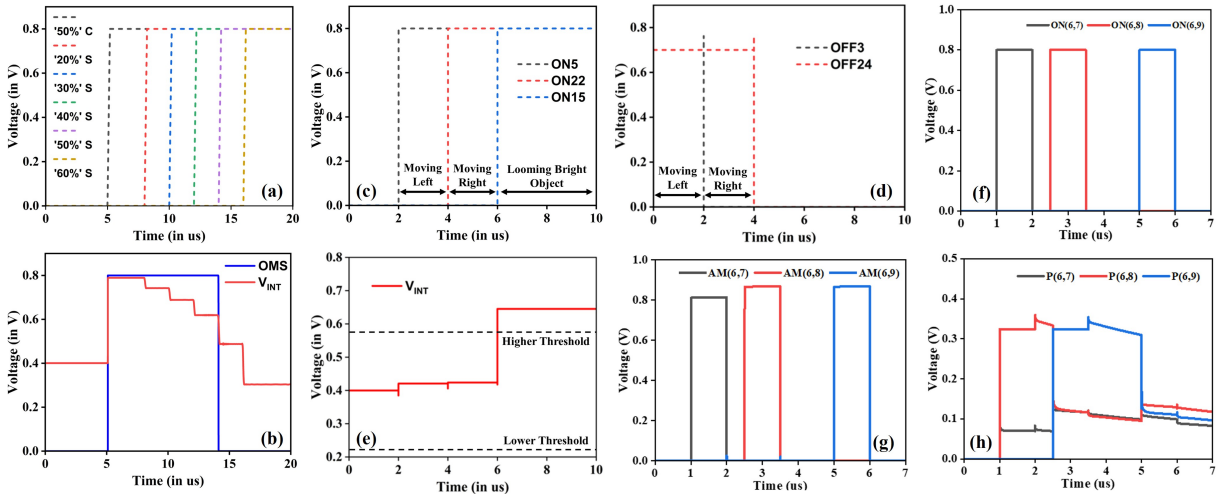

Figure 3. (a) Input spike trains generated by the DVS pixel for the OMS under different activation levels of the center (C) and surround (S) receptive fields, along with (b) the corresponding OMS output response. (c–d) Input spike patterns generated for leftward, rightward, and looming motion, and (e) the resulting output spike, which is produced exclusively during looming motion. (f) Input spikes corresponding to an object moving from left to right, and (g–h) MP output responses demonstrating the generation of predictive and amplified spikes in the direction of motion.

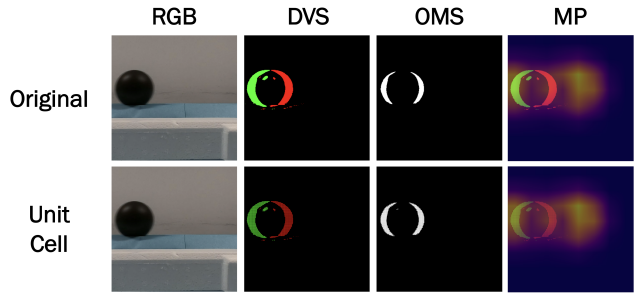

Figure 4. Evaluating the unit cell's OMS and MP performance against the original cells. The RGB video frames are from the translating sequence.

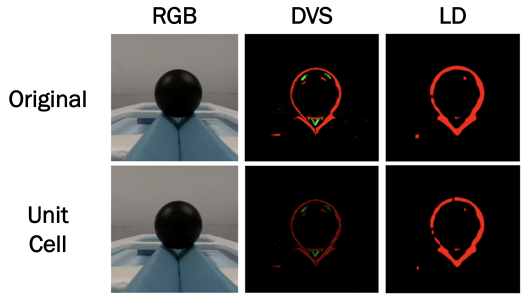

Figure 5. Evaluating the unit cell's LD performance against the original cells.

features closely match those produced by the original implementations. This preserves the spatial structure and temporal behavior of the motion responses. The key takeaway here is that the unit-cell is able to accurately generate OMS and MP features despite the shared hardware resources.

Figure 5 shows the unit-cell's LD performance on the S_A sequence. The original implementation highlights spikes corresponding to approaching threats, and the unit-cell output very closely mimics this behavior. The near-perfect SSIM score indicates that both spatial information and temporal activation patterns are well preserved.

Together, these results quantitatively and qualitatively demonstrate that the IRIS unit-cell generates high-fidelity retinal features while significantly reducing power consumption. When combined with circuit-level measurements, the unit-cell design achieves a 2.41x reduction in energy consumption without sacrificing feature quality. This reinforces the effectiveness of IRIS as a low-power, sensor-level approach to bio-inspired visual intelligence.

4 Conclusion

IRIS leverages neuroscience insights to combine algorithm and hardware design into a unified framework for biologically-inspired visual intelligence. By translating retinal computations into CMOS hardware designs and enabling on-chip spatio-temporal signaling through 3D hybrid-bonding with mixed signal compute circuits, IRIS bridges the gap between biological vision and silicon. This neuroscience-to-hardware co-design not only reduces the bandwidth, latency, and energy bottlenecks of conventional pipelines but also opens new opportunities for sensor-level adaptive and event-driven perception. Across multiple visual features, IRIS preserves 98% feature fidelity while integrating three distinct visual functions, achieving a 2.41x reduction in energy consumption and validating the benefits of neuroscience-inspired co-design.

References

[1] Karl Leif Bates. 2021. *Living Retina Achieves Sensitivity and Efficiency Engineers Can Only Dream About.*

[2] Subhradip Chakraborty, Shay Snyder, Md Abdullah-Al Kaiser, Maryam Parsa, Gregory Schwartz, and Akhilesh R Jaiswal. 2025. A retina-inspired pathway to real-time motion prediction inside image sensors for extreme-edge intelligence. *Neuromorphic Computing and Engineering* 5, 3 (2025), 034005.

[3] Victoria Clerico, Shay Snyder, Arya Lohia, Md Abdullah-Al Kaiser, Gregory Schwartz, Akhilesh Jaiswal, and Maryam Parsa. 2025. Retina-inspired object motion segmentation for event-cameras. In *2025 Neuro Inspired Computational Elements (NICE)*. IEEE, 1–6.

[4] Guillermo Gallego, Tobi Delbrück, Garrick Orchard, Chiara Bartolozzi, Brian Taba, Andrea Censi, Stefan Leutenegger, Andrew J Davison, Jörg Conradt, Kostas Daniilidis, et al. 2020. Event-based vision: A survey. *IEEE transactions on pattern analysis and machine intelligence* 44, 1 (2020), 154–180.

[5] Y Kagawa and Hayato Iwamoto. 2019. 3D integration technologies for the stacked CMOS image sensors. In *2019 International 3D Systems Integration Conference (3DIC)*. IEEE, 1–4.

[6] Jim Nilsson and Tomas Akenine-Möller. 2020. Understanding SSIM. arXiv:2006.13846 [eess.IV] https://arxiv.org/abs/2006.13846

[7] Ziyan Qin, Qinbing Fu, and Jigen Peng. 2024. A computationally efficient and robust looming perception model based on dynamic neural field. *Neural Networks* 179 (2024), 106502. doi:10.1016/j.neunet.2024.106502

[8] Greg Schwartz. 2021. *Retinal computation.* Academic Press.

[9] Kan Shimizu, Takumi Kamibayashi, Kenichi Saito, Nobutatsu Araki, Ryoichi Nakamura, Wataru Otsuka, Yoshihisa Kagawa, and Hayato Iwamoto. 2025. Development of a Novel WoWoW Process for 1/1.3-Inch 50 Megapixel Three-Wafer-Stacked CMOS Image Sensor with DNN Circuits. In *2025 IEEE 75th Electronic Components and Technology Conference (ECTC)*. IEEE, 559–564.

[10] Jason Sinaga, Victoria Clerico, Md Abdullah-AI Kaiser, Shay Snyder, Arya Lohia, Gregory Schwartz, Maryam Parsa, and Akhilesh Jaiswal. 2024. Hardware-algorithm re-engineering of retinal circuit for intelligent object motion segmentation. In *2024 International Conference on Neuromorphic Systems (ICONS)*. IEEE, 256–263.

[11] Jason Sinaga, Shay Snyder, Md Abdullah-Al Kaiser, Dan Jinoy, Gregory Schwartz, Maryam Parsa, and Akhilesh Jaiswal. 2025. Reconfigurable Retina-Inspired Looming Detection. In *2025 IEEE 33rd Annual International Symposium on Field-Programmable Custom Computing Machines (FCCM)*. IEEE, 01–07.

[12] Pascal Vivet, Gilles Sicard, Laurent Millet, Stephane Chevobbe, Karim Ben Chehida, Luis Angel Cubero, Monte Alegre, Maxence Bouvier, Alexandre Valentian, Maria Lepecq, et al. 2019. Advanced 3d technologies and architectures for 3d smart image sensors. In *2019 Design, Automation & Test in Europe Conference & Exhibition (DATE)*. IEEE, 674–679.

[13] Yiwen Xu, Tariq M Khan, Yang Song, and Erik Meijering. 2025. Edge deep learning in computer vision and medical diagnostics: a comprehensive survey. *Artificial Intelligence Review* 58, 3 (2025), 93.

[14] Zihan Yin, Md Abdullah-Al Kaiser, Lamine Ousmane Camara, Mark Camarena, Maryam Parsa, Ajey Jacob, Gregory Schwartz, and Akhilesh Jaiswal. 2023. Iris: Integrated retinal functionality in image sensors. *Frontiers in Neuroscience* 17 (2023), 1241691.

[15] Zihao Zhang, Yu Zhang, Daniel O'Boy, and Miguel Martínez-García. 2025. Adaptive threshold in Leaky-Integrated-and-Fire function for audio-based industrial diagnosis. *Journal of Industrial Information Integration* 48 (2025), 100944. doi:10.1016/j.jii.2025.100944