

# FALCON: A Fast and Low-Power Current-Mode Near-Sensor-Computing Architecture for Real-Time Edge Visual Processing

Liang Zhang<sup>1,2</sup>, Jing Kou<sup>1,2</sup>, Jinyao Mi<sup>1,2</sup>, Yang Liu<sup>1,2</sup>, Junda Zhao<sup>1,2</sup>, Junzhan Liu<sup>1,2\*</sup> and Wang Kang<sup>1,2\*</sup>

<sup>1</sup>State Key Laboratory of Spintronics, Hangzhou International Innovation Institute, Beihang University, Hangzhou 311115, China

<sup>2</sup>School of Integrated Circuit Science and Engineering, Beihang University, Beijing 100191, China

Email: liujunzhan@buaa.edu.cn, wkang@buaa.edu.cn

**Abstract**—Near-sensor computing (NSC) has emerged as a promising paradigm for edge visual processing and data compression, to mitigate data transmission and computing overheads at IoT nodes. However, existing NSC still suffers from limited precision, reduced frame rate and low energy efficiency under complex DNN tasks due to inefficient analog memory, exponential computation overheads and considerable ADC burden. This paper introduces FALCON, a novel current-mode (CM) NSC architecture featuring in-current-register-processing (ICRP) unit and two-step multiply-and-accumulate (TS-MAC) for high-precision and low-latency feature extraction. Additionally, a reconfigurable ADC with embedded ReLU and pooling functionality is employed to improve ADC overhead and compression ratio. Implemented under a 55nm CIS process, FALCON achieves 12.92 TOPS/W with 7-bit weight precision and supports a frame rate of 3096 fps under 8 filters, with an iFOM of 10.1 pJ/pix-fps.

**Index Terms**—near-sensor-computing, CMOS image sensor, convolutional neural network, analog-to-digital converter

## I. INTRODUCTION

With the rapid expansion of the Internet of Things (IoT), enabling fast and energy-efficient AI inference on resource-constrained edge devices has become increasingly challenging. In conventional edge-vision systems, massive raw sensor data are digitized and transferred to downstream processors for computation. The resulting heavy reliance on analog-to-digital conversion and frequent sensor-processor data movement incurs substantial power and latency overhead, constituting a primary bottleneck for edge-vision applications [1–4].

To break this bottleneck, researchers are turning to near-sensor computing (NSC) [5–9], which shifts part of the processing into the analog domain before ADC. Raw image data are compressed into compact feature maps by leveraging energy-efficient analog computation, greatly reducing data transfers and conversion overheads. NSC achieves orders-of-magnitude improvements in energy-efficiency and latency, while maintaining acceptable accuracy for low-to-medium-complexity tasks [10–16].

However, existing NSCs still face challenges such as limited weight precision, low frame rate and reduced compression rate under complex DNN workloads. In this work, we present

This work was supported by the Beijing Nova Program (20250484807), Fundamental Research Funds for the Central Universities (GW2025-08), Zhejiang Provincial Key R&D Program (2025C01071), Research Funding of Hangzhou International Innovation Institute of Beihang University (2024KQ157).

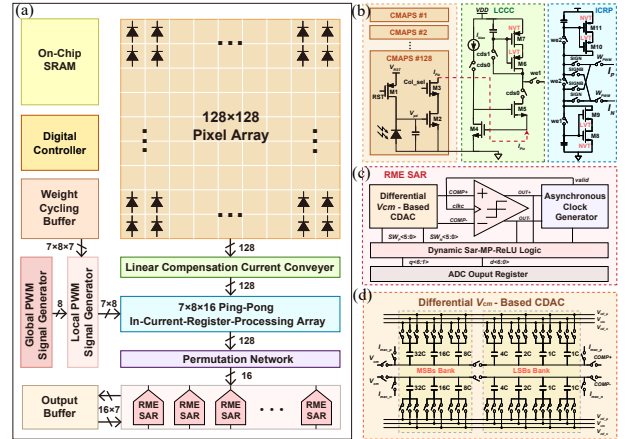


Fig. 1: (a) Overall architecture of FALCON, (b) current-mode CDS and ICRP unit, (c) architecture of the RME-SAR ADC, (d) CDAC reused for computing and A/D conversion.

FALCON, a fast and low-power CM NSC architecture for edge visual processing. Key innovations include: a linear and variation-tolerant current-sensing circuit; a differential ICRP unit and a two-step multiply-accumulate (MAC) strategy for fast and energy-efficient 7-b weighted computation; a reconfigurable ADC with embedded ReLU and max pooling (MP) for enhancing data compression ratio and power reduction.

## II. OPERATION STRATEGY AND OVERALL ARCHITECTURE

Fig. 1(a) illustrates the overview of FALCON with two operating modes, which consist of a 128x128 current-mode APS (CMAPS) pixel array, 128 linear compensation current conveyors (LCCC) for correlated double sampling (CDS), a 7x8x16 ping-pong ICRP (PP-ICRP) array, a permutation network, 16 ReLU-MP-embedded SAR ADCs (RME-SAR), a weight cycling buffer (WCB), a global PWM signal generator (GPG), 7x8 local PWM generators (LPG), and a digital controller. The imaging mode produces 7b 128x128 images, which are necessary to compare the near-sensor convolution, subject to analog non-idealities, with an ideal software baseline. The feature extraction mode executes 7b weighted convolution of the image. Programmable parameters include a maximum filter size of 5x5 with stride S=1/2, and the number of filters ranging from 1 to 64. The generated feature maps can be further configured for ReLU and MP during readout.

### III. CIRCUIT IMPLEMENTATION AND DESIGN

#### A. Linear and variation-tolerant sensing, CDS and storage

Fig. 1(b) illustrates the current-mode sensing datapath. The LCCC double-samples the selected CMAPS pixel and stores the CDS currents in the ICRP unit. The CMAPS pixel adopts a 3T structure, where the readout transistor  $M_2$  operates in the linear region to provide an approximately constant V-I gain ( $g_m$ ) from  $V_{pd}$  to  $I_{pix}$ . To reduce nonlinearity, the LCCC integrates a regulated-cascode topology that stabilizes  $V_{DS}$  of  $M_2$ . To mitigate write/read errors in the current mirror, we employ a current-programmed dynamic-cascode sampling scheme [17], which is less sensitive to threshold-voltage variation than voltage-programmed approaches. The CDS proceeds in three phases. During reset, the reset current is sampled and stored by  $M_6$  and  $M_7$ . After exposure, the NMOS branch of the ICRP samples the difference between the reset and exposure currents, thereby suppressing FPN.

#### B. High-Performance In-Situ Computing

1) **PWM Signal Generation:** The GPG comprises four differential flip-flops in a negative-feedback loop to generate eight multi-phase clocks. These clocks are combined with simple logic and retimed by dual-edge-triggered flip-flops to produce global PWM signals. The programmed filter weights in the WCB are delivered to LPGs, where a weight decoder and multiplexer select and generate the local PWM signals.

2) **Two-Step Current-PWM MAC:** As shown in Fig. 1(b), each ICRP unit includes a current register and read-select switches. During MAC, the stored current is multiplied by the PWM magnitude  $W_{pwm}$ , while the sign bit controls the polarity of the differential output current. In a direct implementation, charge and latency scale exponentially with weight precision. To improve PPA at higher precision, we propose TS-MAC by reusing the CDAC (Fig. 1(d)) for bit partitioning and accumulation, enabled by an added array-separation switch. In Step 1, the 3 LSBs are computed by the ICRPs. Then, the MSB bank is reset while the LSB bank holds its charge, realizing a charge-domain divide-by-8 for MSB/LSB mapping. In Step 2, the 3 MSBs are computed and accumulated. This reduces the energy/latency scaling with  $n$ -bit weights from  $2^n$  to  $2^{\frac{n}{2}+1}$ .

#### C. Reconfigurable SAR ADC with Embedded ReLU and MP

The proposed RME-SAR, as shown in Fig. 1(c), integrates a CDAC, a two-stage dynamic comparator, an asynchronous clock generator, a dynamic SAR/MP/ReLU logic unit, and output registers. In the first approximation cycle, if the comparator output is 0, the clock and subsequent SAR cycles are disabled and the output register is set to zero, realizing ReLU. To support max pooling (MP), an extra comparison cycle is inserted before second approximation cycles, where the stored code drives CDAC switching for voltage subtraction. If the comparator output is 1, the conversion continues and the output is updated as the new maximum; otherwise, the conversion is aborted, retaining the previous maximum. As a result, most conversions are gated by ReLU and MP, reducing ADC energy and output data traffic without additional latency.

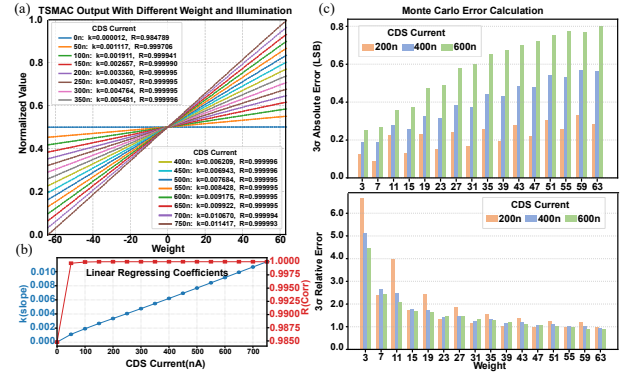


Fig. 2: (a) Transfer function of TS-MAC with 7b weight and 16 illumination level, (b) linearity analysis of computation results, (c) Evaluation of absolute and relative error under 200 point monte carlo evaluation and 3 illumination level.

### IV. SIMULATION RESULTS AND SYSTEM PERFORMANCE

#### A. Functional Verification

We conducted transient simulations of the full signal chain to validate the computational performance of the proposed NSC. The photocurrent was swept from 0 to 15 nA with a 1.6  $\mu$ s exposure time, yielding an effective CDS current range of 0–800 nA. A  $5 \times 5$  filter was enabled with TS-MAC, and the full 7-bit weight range was swept to evaluate linearity. Fig. 2(a) plots the measured outputs, while Fig. 2(b) summarizes the linear-regression parameters. Except for a minor deviation below 100 nA, the average  $R^2$  reaches 0.99997 with consistent regression slopes, indicating excellent linearity. Fig. 2(c) shows the low absolute and relative error under 200 points monte carlo simulation. These results verify the effectiveness of the proposed linearity and robustness compensation techniques in FALCON, leading to reduced computation error.

#### B. System-Level Performance Analysis

FALCON achieves 12.92 TOPS/W for 7-bit-weight convolutions, placing it among the most energy-efficient designs reported at this precision. With a 512.4  $\mu$ W total power, the GMPG, WCG, LPG, and digital buffers consume 43.3%. The TS-MAC and imaging front-end contribute 22.7% and 20.4%, respectively, while the RME-ADC accounts for 13.6%. The TS-MAC latency is 120 ns, whereas row-level CDS requires 2  $\mu$ s. With eight filters and stride two, the pipeline is balanced and reaches a peak frame rate of 3096 fps, corresponding to an iFOM of 10.1 pJ/pix-fps.

### V. CONCLUSION

In this paper, we present FALCON, a fast and low-power current-mode NSC architecture. By combining a low-error processing strategy, energy-efficient in-situ computing circuits, and a multifunctional readout, FALCON achieves **12.92 TOPS/W** energy efficiency and a peak frame rate of **3096 FPS**. Current limitations include limited operator support; future work will extend FALCON with additional functionalities under an event-driven architecture.

## REFERENCES

- [1] Ryoji Eki et al. “9.6 A 1/2.3 inch 12.3 Mpixel with on-chip 4.97 TOPS/W CNN processor back-illuminated stacked CMOS image sensor”. In: *2021 IEEE International Solid-State Circuits Conference (ISSCC)*. Vol. 64. IEEE, 2021, pp. 154–156.
- [2] Kea-Tiong Tang et al. “Considerations of integrating computing-in-memory and processing-in-sensor into convolutional neural network accelerators for low-power edge devices”. In: *2019 Symposium on VLSI Circuits*. IEEE, 2019, T166–T167.
- [3] Sun-II Hwang et al. “A 2.7-M pixels 64-mW CMOS image sensor with multicolumn-parallel noise-shaping SAR ADCs”. In: *IEEE Transactions on Electron Devices* 65.3 (2018), pp. 1119–1126.
- [4] Daniel García Moreno et al. “A cluster of FPAA’s to recognize images using neural networks”. In: *IEEE Transactions on Circuits and Systems II: Express Briefs* 68.11 (2021), pp. 3391–3395.
- [5] Feichi Zhou and Yang Chai. “Near-sensor and in-sensor computing”. In: *Nature Electronics* 3.11 (2020), pp. 664–671.
- [6] Tzu-Hsiang Hsu et al. “AI edge devices using computing-in-memory and processing-in-sensor: From system to device”. In: *2019 IEEE International Electron Devices Meeting (IEDM)*. IEEE, 2019, pp. 22–5.
- [7] Jialong Liu et al. “TFT-based near-sensor in-memory computing: circuits and architecture perspectives of large-area eDRAM and ROM CiM chips”. In: *IEEE Transactions on Circuits and Systems I: Regular Papers* 71.2 (2023), pp. 620–633.
- [8] Han Xu et al. “Macsen: A processing-in-sensor architecture integrating mac operations into image sensor for ultra-low-power bnn-based intelligent visual perception”. In: *IEEE Transactions on Circuits and Systems II: Express Briefs* 68.2 (2020), pp. 627–631.
- [9] Hyunsoo Song et al. “A 120 Frames/s CMOS Image Sensor With 8.19 TOPS/W Computing-In-Pixel for Energy-Efficient Low-Latency Face Detection”. In: *IEEE Journal of Solid-State Circuits* (2025).
- [10] Martin Lefebvre and David Bol. “MANTIS: A Mixed-Signal Near-Sensor Convolutional Imager SoC Using Charge-Domain 4b-Weighted 5-to-84-TOPS/W MAC Operations for Feature Extraction and Region-of-Interest Detection”. In: *IEEE Journal of Solid-State Circuits* (2024).
- [11] Zhe Chen et al. “Processing near sensor architecture in mixed-signal domain with CMOS image sensor of convolutional-kernel-readout method”. In: *IEEE Transactions on Circuits and Systems I: Regular Papers* 67.2 (2019), pp. 389–400.
- [12] Tianrui Ma et al. “Leca: In-sensor learned compressive acquisition for efficient machine vision on the edge”. In: *Proceedings of the 50th Annual International Symposium on Computer Architecture*. 2023, pp. 1–14.
- [13] Tzu-Hsiang Hsu et al. “A 0.8 V intelligent vision sensor with tiny convolutional neural network and programmable weights using mixed-mode processing-in-sensor technique for image classification”. In: *IEEE Journal of Solid-State Circuits* 58.11 (2023), pp. 3266–3274.
- [14] Tzu-Hsiang Hsu et al. “A 0.5-V real-time computational CMOS image sensor with programmable kernel for feature extraction”. In: *IEEE Journal of Solid-State Circuits* 56.5 (2020), pp. 1588–1596.
- [15] Min-Yang Chiu et al. “A multimode vision sensor with temporal contrast pixel and column-parallel local binary pattern extraction for dynamic depth sensing using stereo vision”. In: *IEEE Journal of Solid-State Circuits* 58.10 (2023), pp. 2767–2777.
- [16] Junzhan Liu et al. “A 1000FPS@ 360,000 pixels mixed-signal sensing with computing macro featuring analog compression and maximum parallelism for objective detection tasks”. In: *Sensors and Actuators A: Physical* 379 (2024), p. 115951.
- [17] Jiahao Song et al. “A 4-bit calibration-free computing-in-memory macro with 3T1C current-programmed dynamic-cascode multi-level-cell eDRAM”. In: *IEEE Journal of Solid-State Circuits* 59.3 (2023), pp. 842–854.