A 40nm 3.58 mJ/Frame PCM-based POSIT/INT Mixed-Precision Digital/Analog CIM Chip for Edge Foveated Rendering

Bowen Wang[1#], Zelun Pan[1#], Zhe Zhan[1], Longhao Yan[1*], Jiukai Fang[1], Xile Wang[1], Yuqi Li[1], Minghui Yin[1], Xi Li[3], Zhitang Song[3*], Yaoyu Tao[1*], Yuchao Yang[1,2,4*]
[1]School of Integrated Circuits, Peking University, Beijing, China. [2]Shenzhen Graduate School, Peking University, Shenzhen, China. [3]Shanghai Institute of Microsystem and Information Technology, Chinese Academy of Sciences, Shanghai, China. [4]Center for Brain Inspired Intelligence, Chinese Institute for Brain Research (CIBR), Beijing, China.
[#]Equally contributed authors.
*Corresponding Author's Email: yanlonghao@pku.edu.cn, ztsong@mail.sim.ac.cn, taoyaoyutyy@pku.edu.cn, yuchaoyang@pku.edu.cn

***Abstract*—This work proposes a 40nm PCM-based chip with 3 key innovations to address the challenges of latency, reliability and energy-efficiency in edge real-time rendering: (1) a hardware-software co-designed mixed-precision foveated neural radiance field (MP-FovNeRF) system with analog/digital compute-in-memory (CIM) for low latency; (2) a PCM-oriented error correction codes (ECC) and a redundant ADC (R-ADC) for reliable computation; (3) a 2-level sparsity aware multiply-accumulate (MAC) method for high energy efficiency. Together, the chip achieves 2.38 FPS/mm^2 and 3.58 mJ/frame for edge real-time rendering. High area efficiency and energy efficiency are achieved with nearly no decrease in image quality.**

Keywords*—PCM, compute in memory, posit, mixed-precision, foveated rendering.

I. Introduction

Edge VR/AR devices impose stringent constraints on latency, energy efficiency, and reliability, making real-time on-device rendering particularly challenging. Neural Radiance Field (NeRF)[1] enables high-quality scene synthesis without explicit geometric modeling; however, its reliance high spatial resolution (large number of rays) leads to prohibitive latency. Despite the development of foveated NeRF (FovNeRF) effectively lower the resolution requirement[2], the intensive neural network inference using high precision numerical format still poses heavy burden on resource-constrained edge platforms.

Beyond computational complexity, rendering workloads exhibit strong sensitivity to computation errors. While PCM-based compute-in-memory (CIM) architectures provide high throughput by enabling massively parallel matrix operations and reducing data movement, they inherently suffer from device non-idealities and ADC settling errors, which can significantly degrade rendering quality.

Furthermore, rendering workloads dominate system energy consumption in edge VR/AR devices. Under stringent battery capacity constraints, achieving high energy efficiency is therefore essential, placing additional demands on the underlying computing platform.

To address the above challenges, we propose a 40nm PCM-based mixed-precision CIM (MP-CIM) chip with a hardware-software co-designed framework for efficient real-time edge rendering. As shown in Fig. 1, we introduce a mixed-precision FovNeRF (MP-FovNeRF) that exploits the non-uniform acuity of human vision by allocating high-precision, high-resolution inference (POSIT16 in DCIM) to the foveal region and low-precision, high-throughput computation (INT8 in ACIM) to the peripheral region, significantly reducing system latency. Second, we propose a reliability enhancement design, incorporating PCM-oriented CIM error correction code (ECC) and a redundant analog-to-digital-converter (R-ADC) to mitigate computation errors arising from PCM conductance drift and ADC settling, thereby preserving excellent image quality. Thirdly, a two-

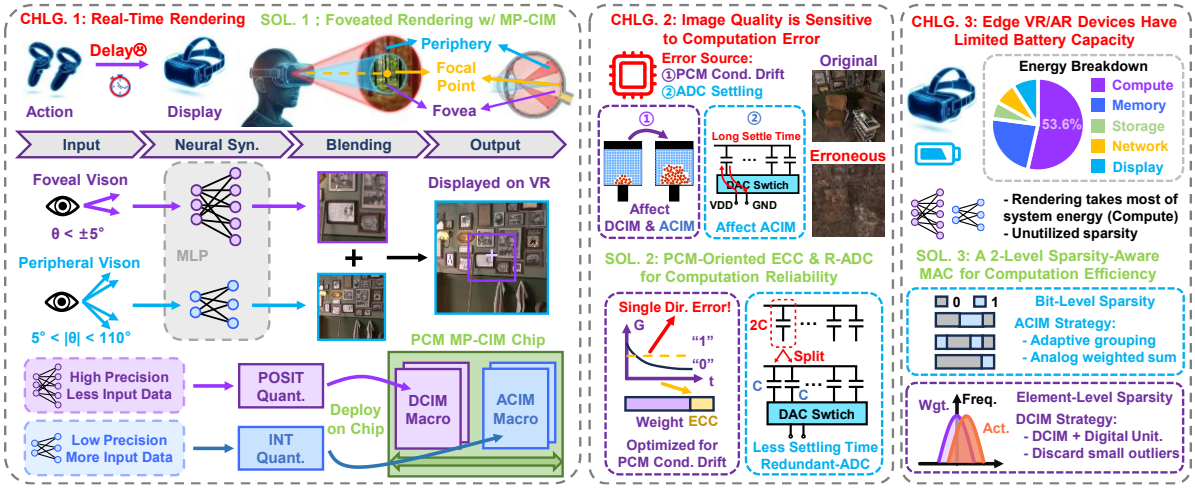

Figure. 1. Real-time rendering demands low computing latency. Additionally, high-resolution rendering quality are sensitive to computation errors and rendering on edge requires low energy consumption. So we propose a robust mixed-precision FovNeRF with MP-CIM that utilizes sparsity.

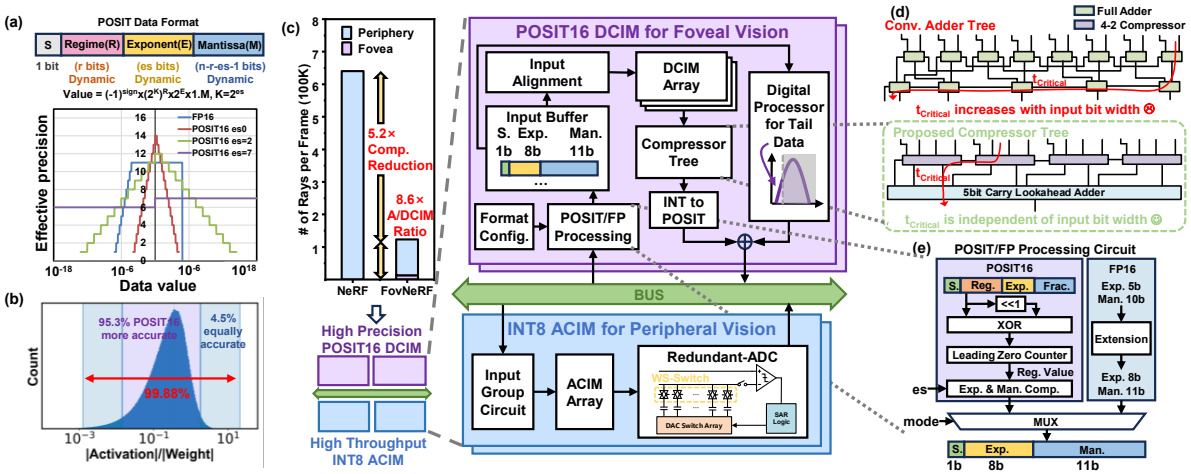

Figure. 2. (a) Comparison of effective precision of FP16/POSIT16. (b) Distribution of activations & weights in FovNeRF. (c) Architecture overview of our proposed PCM-based mixedprecision MP-CIM chip. (d) Proposed compressor tree. (e) Design of POSIT/FP processing circuit.

level sparsity-aware MAC strategy is proposed, jointly leveraging bit-level sparsity in ACIM and element-level sparsity in DCIM to minimize redundant computation latency and energy consumption.

II. Proposed PCM-based Mixed-Precision CIM Chip for Edge Foveated Rendering

A. *Hardware-software Co-designed MP-FovNeRF with Analog/Digital CIM for Low Latency*

Our proposed MP-FovNeRF partitions the perceptual field into foveal and peripheral regions based on gaze direction. The region within 5° of the gaze center is classified as the foveal region and is processed using DCIM with high spatial resolution (i.e., high ray density) under POSIT16 precision. As illustrated in Fig. 2(a), the POSIT16 format provides superior numerical precision around unity compared to FP16 [4]. We adopt POSIT16 with es=2 to maximize the fraction of weights and activations that benefit from higher precision. As shown in Fig. 2(b), this design enables 95.3% of computations to be performed with higher precision than FP16, thereby improving image quality in the foveal region.

The remaining perceptual area is defined as the peripheral region and is processed using ACIM with reduced ray density under INT8 precision. To balance the computational workload between regions, we carefully adjust the ray density by jointly considering the spatial coverage of both regions and the throughput differences between ACIM and DCIM. This co-design ensures matched latency across regions. As a result, MP-FovNeRF reduces total computation by 5.2× compared to standard NeRF, while preserving ray density—and thus visual fidelity—in the foveal region.

Fig. 2(c) shows the architecture of our proposed MP-CIM chip. The chip has 2 DCIM macros and 2 ACIM macros, containing a total of 256Kb PCM devices. In ACIM macros, the input group circuit is used for reducing MAC energy and latency with lossless compression, and leveraging bit-level sparsity. The R-ADC is used for mitigating ACIM computation errors. To support large numerical range for POSIT16, we design a two pathway MAC circuit in the DCIM macros, where most weights that have a compact distribution are pre-aligned, stored and processed in DCIM array, while the small number of small outliers are stored in a custom digital processor (Digit.) to avoid alignment loss in DCIM. The Digit. can also leverage element-level sparsity to skip computations associated with small values. The detail of POSIT/FP Processing Circuit is shown in Fig. 2(e). The input POSIT number is converted into a custom FP20 format to minimize precision loss during computation. The input format can also be FP16, preserving the versatility for other tasks. As shown in Fig. 2(d), we use 4-2 compressors to reduce the critical path latency, making it irrelevant to input bit width unlike conventional adder tree and enabling higher clock frequency, and thus higher rendering efficiency.

B. *PCM-oriented ECC and R-ADC for Reliable Computation*

PCM conductance drift deteriorates the stored weights and can cause errors in computations (Fig.3(a)). We observe that the error caused by drift is mostly from a single direction (1 to 0, as shown in Fig.3(b)) because low-resistance states(LRS)

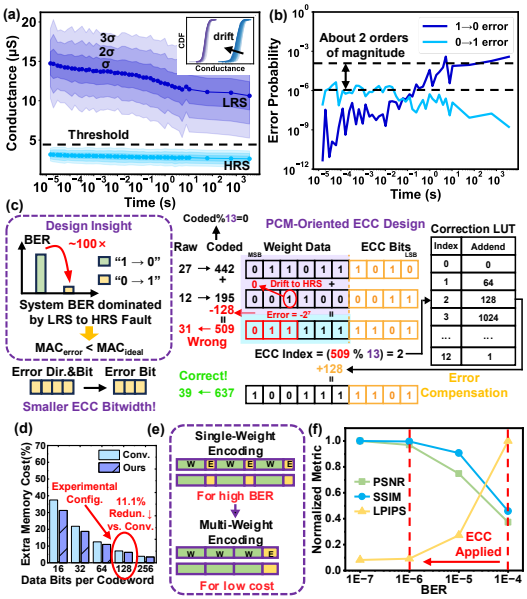

Figure. 3. Measured (a)PCM conductance with drift and (b)BER of different bit-flip directions. (c) Proposed PCM-oriented CIM ECC design for single direction faults due to PCM drift. (d)(e) Multi-weight encoding and cost of proposed ECC design. (f) The impact of BER on rendering performance.

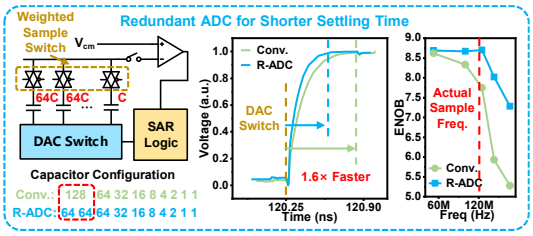

Figure. 4. Redundant ADC for reliable computation at higher frequency.

tend to drift across the threshold into high-resistance states(HRS). So we design a PCM-oriented CIM ECC with shorter code length considering the unidirectional error characteristics. As shown in Fig. 3(c), the unidirectional bit error characteristics result in a behavior where erroneous MAC results are always smaller than the ideal ones. Under this assumption, we can eliminate one redundancy bit that indicates the MAC error direction in each ECC code, thereby simplifying the residue code used in CIM-oriented ECC. Fig. 3(d) demonstrates a 11.1% reduction in redundancy over prior CIM-oriented ECC schemes[5], under our experimental configuration of 128 data bits per codeword. Since the bitwidth of our weights is 16 bits, we adopt the multi-weight encoding method where multiple weights are packed into each codeword, as shown in Fig. 3(e). The normalized metrics of rendering quality, before and after ECC is applied(BER=10^{-4} and 10^{-6}, respectively), are shown in Fig.3(f). We observe that PCM drift error has no effect on the final rendering quality after ECC is applied.

Fig. 4 shows our R-ADC design for shorter settling time. By breaking the largest capacitor in conventional SAR ADC into 2 equal capacitors, we achieve 1.6× faster settling time and the resul ing ENOB under 120MHz sample frequency shows obvious advantage. Thus, reliable computation is achieved at high throughput.

C. *2-level Sparsity Aware MAC Method for High Energy Efficiency*

As shown in Fig. 2(b), a large number of upper zeros exist when the activations and weights are quantized to INT8. And the long tail on the left indicates that there are some extremely small values. High energy efficiency can be achieved by exploiting the aforementioned bit-level and element-level

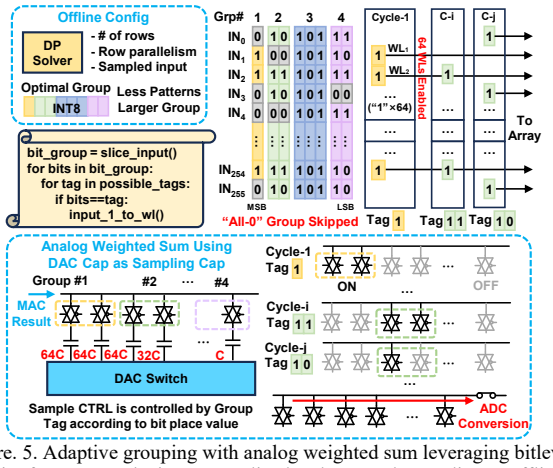

Figure. 5. Adaptive grouping with analog weighted sum leveraging bitlevel sparsity for ACIM. The inputs are sliced and grouped according to offline determined configuration to achieve better efficiency.

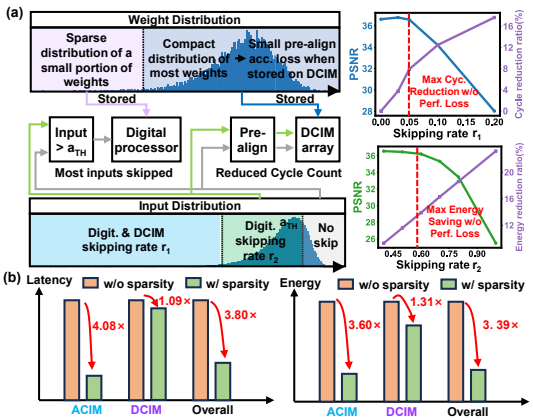

Figure. 6. (a) Input skipping strategy for DCIM leveraging element-level sparsity. (b) Overall performance of 2-level sparsity aware MAC is shown.

sparsity. So we design a 2-level sparsity-aware MAC for ACIM and DCIM respectively.

As shown in Fig 5, in ACIM we design an adaptive input grouping method leveraging bit-level sparsity. Input grouping method can group multiple bits and compute 64 same multi-bit patterns in one input cycles. Unlike conventional design[6] that has a fixed group size, our method dynamically adjusts to the optimal group size according to an offline configuration searched by a solver considering the input distribution. All-zero patterns can be skipped directly, and fewer distinct patterns among grouped multiple bits also reduce the input cycles. MAC energy and latency can be reduce owing to the concentrated distribution of operands and abundant high-order zeros. Along with our binary weighted sampling R-ADC, the weighted sum of the partial MAC results is computed in the analog domain[12]. The grouped bits are used to control the DAC sampling switch in the R-ADC, generating corresponding weighted partial-sum. Such design eliminates complicated digital shift-add logic, and thus largely increases energy efficiency.

As shown in Fig. 6(a), for DCIM, we adopt a skipping strategy leveraging element-level sparsity. Inputs of different magnitude are divided into 3 parts. The largest part is sent to both DCIM and Digit. and the middle part is sent only to DCIM (Digit. skip rate r2), while the smallest part is completely discarded (DCIM and Digit. skip rate r1). We investigate the effect of skipping rate r1 and r2 on rendered image (one is fixed at 0 while the other is parametrically swept), and we choose the skipping rate that can maximize cycle reduction and energy saving without compromising image quality (<1% accuracy loss).

Together, the 2-level sparsity-aware MAC method offers 3.80× reduction in MAC latency and 3.39× reduction in MAC energy, shown in Fig. 6(b).

III. Performance Evaluation

As shown in Fig. 7(a), our chip, two FPGAs and power supply circuits are integrated on a single test board, connected to a PC via a UART port. A rendering result of MP-FovNeRF at a resolution of (1660,1440) is shown in Fig. 7(b). It can be observed that the same object in foveal region exhibits higher rendering quality than that in peripheral region. Since the foveal region of human eyes provides 5–20× higher detail resolution than the peripheral region[11], rendering accuracy

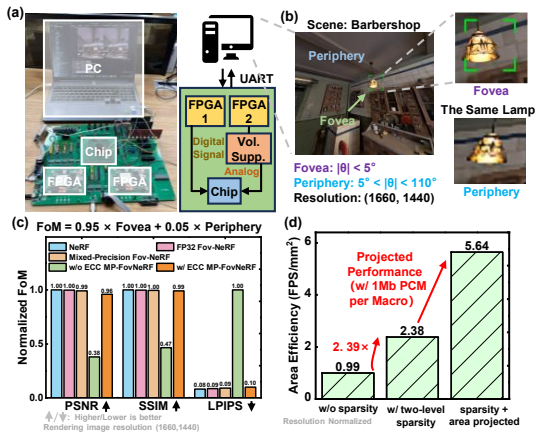

Figure. 7. (a) Chip test platform and (b) an example of MP-FovNeRF under high resolution. (c) FoM of rendered image quality comparison and (d) total area efficiency utilizing the sparsity-aware MAC is demonstrated.

in the foveal region is more critical than in the periphery. So we adopt an accuracy FoM to evaluate the human visual perception of rendering results in Fig. 7(c). The FoM of our mixed-precision hardware implementation with ECC is nearly equal to that of the full-precision software implementation. The area efficiency is improved by 2.39× after two-level sparsity is exploited, as shown in Fig.7(d). Constrained by the size of PCM array per macro, higher area efficiency can be achieved potentially.

As shown in Fig. 8, by exploiting the two-level sparsity, the chip has achieved 3.04x reduction in energy per frame compared with the baseline. The chip is fabricated in a 40nm process, and the die area is 0.831 mm^2 excluding the IO pads. The peak frequencies of ACIM and DCIM are 120MHz and 200MHz, respectively. Power and area breakdown of our proposed PCM chip is also shown.

The comparison between state-of-the-art(SOTA) works is shown in Table I. A hardware-software co-designed MP-FovNeRF is implemented, supporting POSIT16, FP16 and INT8 data formats. Our work achieves 2.38 FPS/mm^2 area efficiency. And the energy per frame is 3.58 mJ, which outperforms SOTA works.

IV. Conclusion

This work presents a 40nm PCM-based mixed-precision digital/analog CIM chip for edge foveated rendering. By

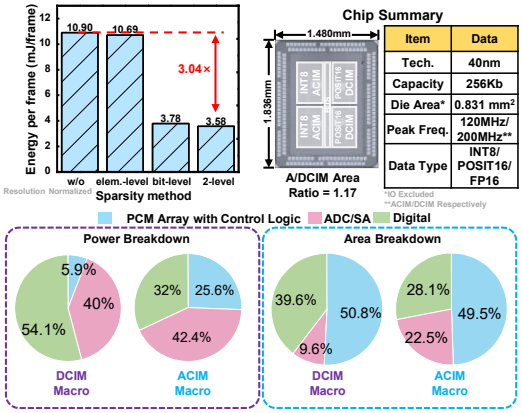

Figure. 8. System energy efficiency utilizing the sparsity-aware MAC, chip summary, power and area breakdown of our proposed PCM chip.

	NVIDIA A100	ISSCC' 23[7]	VLSI' 23[8]	ISSCC' 24[9]	ISSCC' 25[10]	This work
Technology	7nm	28nm	28nm	28nm	28nm	40nm
Die Area [mm²]	826	20.25	5.07	20.25	20.25	0.83
Model/ Method	Foveated NeRF	NeRF	NeRF	NeRF	Gaussian Splatting	MP-FovNeRF
Data Precision Supported	FP32	FP8/ FP16	INT8	FP16/FP8	FP16/FP8	POSIT16/ FP16,INT8
MAC Implementation	GPU PE	Digital PE	Digital CIM	Digital PE	Digital PE	DCIM + ACIM
Frequency [MHz]	1410	50-250	100-200	50-200	50-200	20-120 /20-200
Area Efficiency[1] [FPS/mm²]	0.00707	1.12-3.78	4.2	1.12	2.49	2.38[2] | 5.64[2][3]
Energy per frame [mJ/frame]	475200	5.36-8.17	4.2	7.95-9.94	8.55	3.58[2]

1) Area is normalized to 40-nm technology
2) Normalized from resolution (1600,1440) to (800,800) due to dataset discrepancies. Most existing works use datasets with resolutions equal to or below 800×800.
3) Projected to 1Mb per macro from 64Kb per macro

Figure. 9. Comparison with state-of-the-art chips for rendering

hardware-software co-designing MP-FovNeRF with POSIT16/INT8 mixed-precision computing, PCM-oriented ECC and R-ADC for reliability, and a 2-level sparsity-aware MAC scheme for efficiency, the proposed chip achieves low-latency, reliable, and energy-efficient rendering. Measurement results show 3.58 mJ/frame energy consumption with nearly no loss in image quality, demonstrating its strong potential for real-time edge VR/AR rendering.

Acknowledgment

This work was supported by Guangdong S&T Program (2026B0101070006), the National Key R&D Program of China (2023YFB4502200), the National Natural Science Foundation of China (92164302, 62404004, U25A20487), Guangdong Provincial Key Laboratory of In-Memory Computing Chips (2024B1212020002), Shenzhen Science and Technology Program (ZDCY20250901103401002, JCYJ20241202125907011), Beijing Natural Science Foundation (L234026, L257010) and the 111 Project (B18001). This work has been supported by the New Cornerstone Science Foundation and Financial Support for Outstanding scientific and technological innovation Talents Training Fund in Shenzhen.

References

[1] Mildenhall B, Srinivasan P P, Tancik M, et al. "Nerf: Representing scenes as neural radiance fields for view synthesis." *Communications of the ACM*, 2021.

[2] Deng, Nianchen, et al. "Fov-nerf: Foveated neural radiance fields for virtual reality." *IEEE Transactions on Visualization and Computer Graphics,* 2022.

[3] Leng, Yue, et al. "Energy-efficient video processing for virtual reality." *Proceedings of the 46th International Symposium on Computer Architecture*. 2019.

[4] Wang, Yang, et al. "34.1 A 28nm 83.23 TFLOPS/W POSIT-based compute-in-memory macro for high-accuracy AI applications." *ISSCC*, 2024.

[5] He, Zhen, et al. "ER-DCIM: Error-resilient digital CIM architecture with run-time MAC-cell error correction." *HPCA*, 2025.

[6] Wen, Tai-Hao, et al. "34.8 A 22nm 16Mb floating-point ReRAM compute-in-memory macro with 31.2 TFLOPS/W for AI edge devices." *ISSCC*, 2024.

[7] Han, Donghyeon, et al. "2.7 MetaVRain: A 133mW real-time hyper-realistic 3D-NeRF processor with 1D-2D hybrid-neural engines for metaverse on mobile devices." *ISSCC*, 2023.

[8] Jo, Wooyoung, et al. "NeRPIM: A 4.2 mJ/frame neural rendering processing-in-memory processor with space encoding block-wise mapping for mobile devices." *VLSI Technology and Circuits*, 2023.

[9] Park, Gwangtae, et al. "Space-Mate: A 303.5-mW Real-Time Sparse Mixture-of-Experts-Based NeRF-SLAM Processor for Mobile Spatial Computing." *IEEE Journal of Solid-State Circuits*, 2025.

[10] Song, Seokchan, et al. "IRIS: A 8.55 mJ/frame spatial computing SoC for interactable rendering and surface-aware modeling with 3D Gaussian splatting." *ISSCC*, 2025.

[11] Anstis, Stuart M. "Chart demonstrating variations in acuity with retinal position." *Vision research,* 1974.

[12] Z. Pan et al., "A 40nm Mixed-Precision PCM Chip with Fused Analog/Digital Compute-in-Memory and Adaptive Drift Compensation for Embodied AI Applications," *IEDM*, 2025.