Swin-SR: A 0.044∼1.14 μJ/token Multi-Modal Signal Restoration Processor with Hierarchical Operation-in-Memory-Access Streaming

Fengshi Tian[1], Zhipeng Liao[2], Zilu Liu[3], Hui Wu[2], Jiakun Zheng[1], Ziyang Shen[2], Pingcheng Dong[1], Xihao Guan[1], Chaoming Fang[2], Jinbo Chen[2], Shiqi Zhao[4], Jie Yang[2], Mohamad Sawan[2], Chi-Ying Tsui[1], Kwang-Ting (Tim) Cheng[1]
[1]Hong Kong University of Science and Technology, Hong Kong SAR, China; [2] Westlake University, China; [3] Great Bay University, China; [4] Northeastern University, China

***Abstract*—This paper presents Swin-SR, the first unified Swin-Transformer (Swin-T) processor designed for multi-modal signal restoration applications, ranging from neural video codecs to brain–computer interfaces. To address the severe bottlenecks of diverse computation patterns and excessive data movement inherent in Swin-T models, Swin-SR introduces two co-designed architectural innovations: 1) A heterogeneous 3D-Cubic/2D-Array core (HP-Core) that enables operation-fused streaming, boosting hardware utilization via a novel "Transpose-in-Cubic" bidirectional dataflow and hierarchical streaming engines; and 2) A hierarchical Operation-in-Memory-Access (OIMA) paradigm. OIMA features pre-mask on-the-fly shifted-window (PMSW) processing to eliminate redundant attention computations, and a residual-aware update (RAU) mechanism leveraging dynamic quantization and zero-skipping delta updates to minimize memory access traffic. Fabricated in 40nm CMOS, the 9.00 mm^2 Swin-SR chip achieves 0.044–1.14 μJ/token energy efficiency and delivers 33.41–37.85 dB PSNR across multiple restoration tasks, providing up to 12.99× higher energy efficiency and 5.88× latency speedup compared to baseline implementations.**

***Index Terms*—Swin-Transformer, Heterogeneous Porcessor, Signal Restoration, Multi-Modal AI**

I. Introduction

High-fidelity signal restoration (SR) is a foundational enabler for emerging applications, including neural video codecs, medical imaging (MRI), and brain–computer interfaces (BCI). While prior CNN-based accelerators primarily target vision-only workloads with localized receptive fields [1]–[4], Swin-Transformer (Swin-T) models have emerged as a unified restoration backbone, delivering superior quality across diverse signal modalities by capturing long-range dependencies [5]–[8] (Fig. 1).

Despite their algorithmic superiority, deploying Swin-T models on energy-constrained edge devices remains highly challenging. The core components of a Swin-T block—Shifted-Window Multi-Head Self-Attention (SW-MHSA), Multi-Layer Perceptrons (MLP), and deep residual connections—introduce significant computational heterogeneity and memory movement overheads. Specifically, existing hardware accelerators face two system-level bottlenecks (Fig. 1): **C1) Low Execution Utilization:** The workload is a heterogeneous mixture of dense GEMM/VMM operations and feature-dependent element-wise operations (e.g., LayerNorm, Softmax). In conventional monolithic accelerators, this dispar-

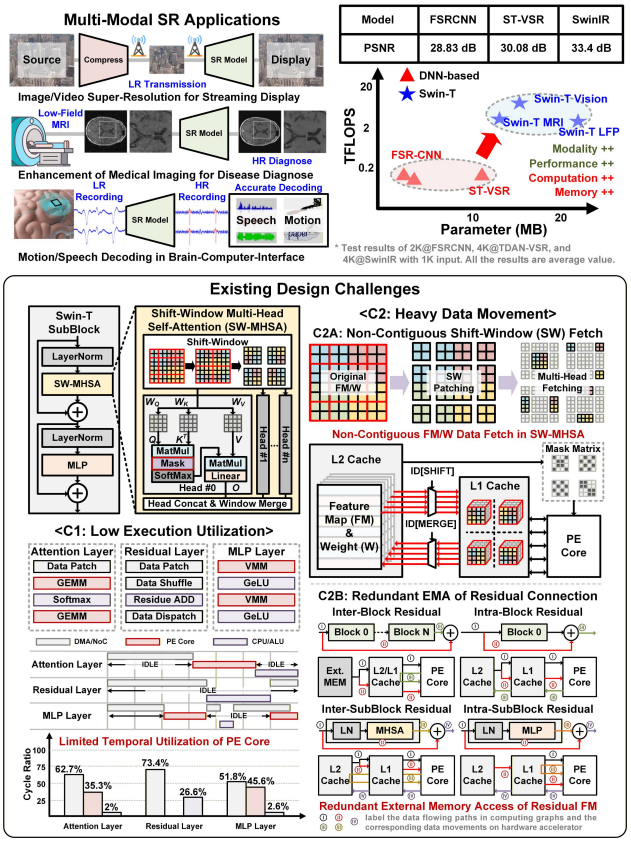


Fig. 1. Signal restoration models with Swin-T raise two kinds of challenges for Swin-T accelerator design.

This research was partially conducted by ACCESS – AI Chip Center for Emerging Smart Systems, supported by the InnoHK initiative of the Innovation and Technology Commission of the Hong Kong Special Administrative Region Government. This work was supported in part by the STI2030-Major Projects under Grant 2022ZD0208805, in part by the "Pioneer" and "Leading Goose" Research and Development Program of Zhejiang under Grant 2024C03002, and in part by the Key Project of Westlake Institute for Optoelectronics under Grant 2023GD004. The first two authors contributed equally. The corresponding author of this paper is Fengshi Tian.

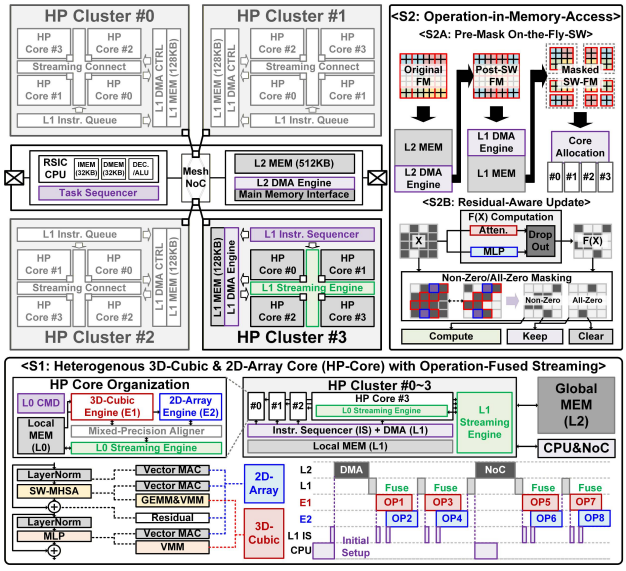


Fig. 2. Proposed Swin-SR processor with two key architecture features.

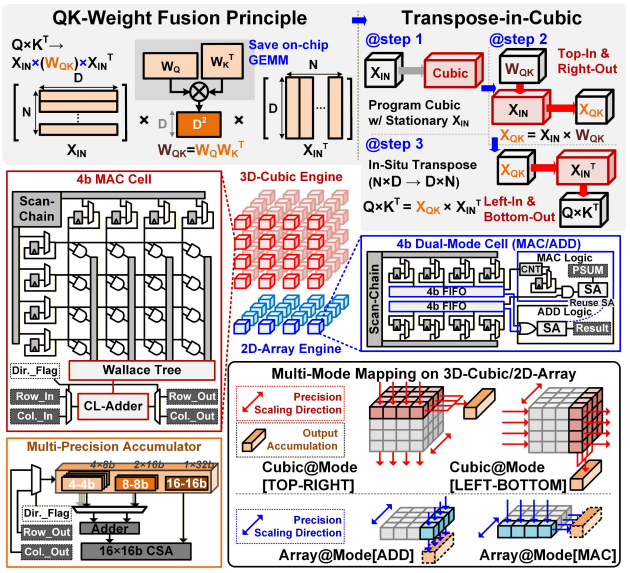


Fig. 3. Proposed circuit designs of 3D-Cubic/2D-Array engines and the transpose-in-cubic scheme.

ity causes severe pipeline stalls, leaving Processing Elements (PEs) idle for over 60% of cycles during data dispatch and residual phases. **C2) Heavy Data Movement:** This bottleneck manifests in two distinct forms. **C2A (Non-Contiguous Fetch):** The SW-MHSA mechanism requires non-contiguous memory access for window patching, and computes redundant attention scores on padded/masked tokens. **C2B (Redundant Residual EMA):** Deep intra- and inter-block residual connections trigger frequent read-modify-write operations to external memory. Consequently, EMA scales linearly with token count, dominating the overall system energy. To overcome these challenges, we propose **Swin-SR**, a highly efficient domain-specific processor. By co-optimizing the algorithm mapping, core architecture, and memory hierarchy, Swin-SR effectively dismantles the utilization and memory walls of Swin-Transformers.

II. Swin-SR Processor Architecture

To address the utilization and data-movement bottlenecks, Swin-SR is organized into four parallel clusters interconnected by a mesh Network-on-Chip (NoC). As shown in Fig. 2, a distributed instruction dispatch system and a two-level DMA hierarchy coordinate computation and dataflow across the chip.

A. Heterogeneous 3D-Cubic & 2D-Array Core (HP-Core)

To tackle the computation diversity (C1), we partition the workload into dense and sparse/element-wise operations, mapping them onto a heterogeneous HP-Core. Each HP-Core integrates a **3D-Cubic Engine (E1)** for dense GEMM/VMM and a **2D-Array Engine (E2)** for vector-wise operations. The rationale behind assembling the GEMM engine into a 3D cubic topology is to facilitate dynamic grouping. Organizing multiple 2D MAC slides along the depth dimension allows flexible combination of 4-bit cells. This maintains massive spatial parallelism while natively supporting multi-precision scaling (4b/8b/16b) without footprint underutilization. Fig. 3 details the circuit designs of the E1 and E2 engines. A major challenge in SW-MHSA is the generation of the attention matrix ($Q \times K^T$), which traditionally requires explicit, energy-intensive matrix transposition in SRAM. Leveraging a co-designed **QK-Weight Fusion** strategy, we pre-compute the weights ($W_{QK} = W_Q W_K^T$) off-chip. The 3D-Cubic engine then buffers the input feature map (X_{IN}) as stationary data. As shown in Steps 1 and 2, X_{IN} is multiplied by W_{QK} with partial sums accumulated along the *row* direction to produce the intermediate matrix X_{QK}. In Step 3, to compute $Q \times K^T = X_{QK} \times X_{IN}^T$, the intermediate X_{QK} is directly routed back to the input ports. Crucially, by switching the accumulation dataflow to the orthogonal *column* direction, the 3D-Cubic engine inherently satisfies the dimensional requirement of a transposed multiplication ($N \times D \rightarrow D \times N$). This "Transpose-in-Cubic" mechanism implicitly performs transposition, completely eliminating the explicit memory read/write overhead. The 3D-Cubic engine comprises 4-bit MAC cells equipped with scan-chains for stationary data loading and Wallace-tree multipliers. Complementing E1, the 2D-Array Engine consists of 4-bit dual-mode cells configurable for either parallel element-wise addition (for residuals) or vector-wise MAC operations.

B. Hierarchical Streaming Architecture

To sustain high spatial and temporal utilization, Swin-SR abandons traditional layer-by-layer execution in favor of **Operation-Fused Streaming (OFS)**. As illustrated in the timing diagram of Fig. 2 (S1), this mechanism drastically

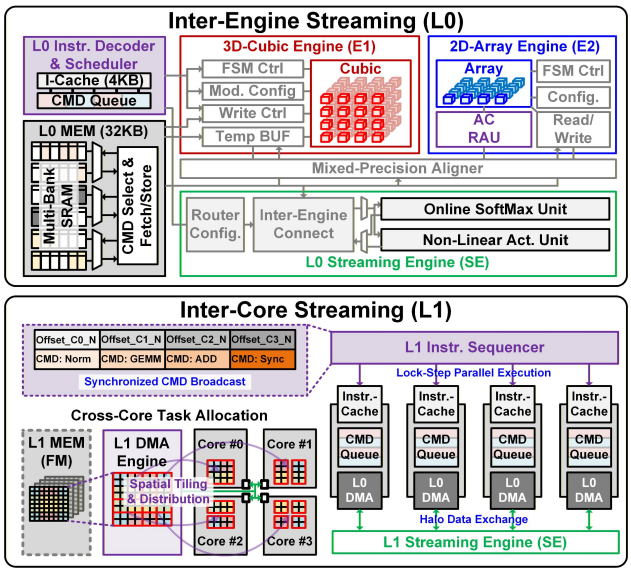


Fig. 4. Proposed hierarchical (L0 & L1) streaming and the hardware fabrics.

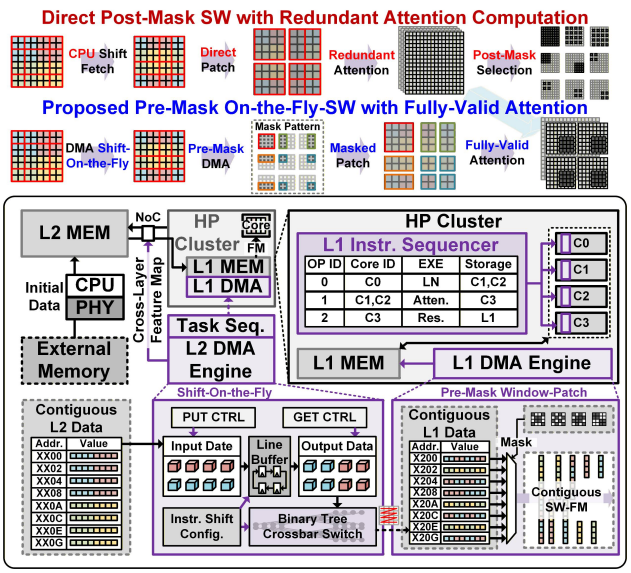


Fig. 5. Proposed pre-mask on-the-fly-shifted-window scheme and the operation-in-memory-access fabrics.

optimizes execution timing by overlapping dependent operations. The hierarchical streaming architecture operates at two levels (Fig. 4): **Inter-Engine Streaming (L0):** Instead of stalling the pipeline to write full-tensor outputs back to memory, intermediate results (e.g., LayerNorm outputs) are streamed directly to the next engine via static switchboxes and asynchronous FIFOs. While the 3D-Cubic engine executes dense GEMM (OP1), the 2D-Array engine concurrently processes the subsequent vector operations (OP2), effectively hiding memory latency and shrinking the end-to-end execution timeline. To prevent non-linear operations from bottlenecking the streaming pipeline, the L0 SE also integrates dedicated hardware units–Online SoftMax Unit and Non-Linear Activation Unit. **Inter-Core Streaming (L1):** At the cluster level, Swin-T partitions feature maps into spatial windows; the L1 SE allocates these patches across parallel cores. For the Shifted-Window (SW) mechanism, the L1 SE orchestrates DMA-driven request-packet communication over the L2 Mesh NoC to perform "Halo Data Exchange". This enables adjacent cores to directly share boundary tokens from their local L1 memories, bypassing the L2 MEM access entirely and perfectly aligning with Swin-T's spatial computing flow.

III. Operation-in-Memory-Access (OIMA)

To eliminate the severe C2 memory bottleneck, we propose an OIMA hierarchy (Fig. 5 and Fig. 6) that shifts complex data reorganization from instruction-driven fetches to configuration-driven memory transfers.

A. Pre-Mask On-the-Fly Shifted-Window (PMSW)

Conventional SW-MHSA execution requires the CPU to reorder data in memory to form shifted windows, wasting cycles and bandwidth. Swin-SR introduces a DMA-based **Shift-On-the-Fly** mechanism. Using reconfigurable line buffers and a crossbar switch within the L2 DMA, contiguous data is fetched and automatically reorganized into shifted windows during transit to the L1 MEM. Furthermore, to address the redundancy caused by useless token attention computation, a L1 **Pre-Mask Window-Patch** DMA applies validity masks *before* data enters the HP-Core. This ensures that the 3D-Cubic engine only processes fully-valid attention patches (Fig. 5).

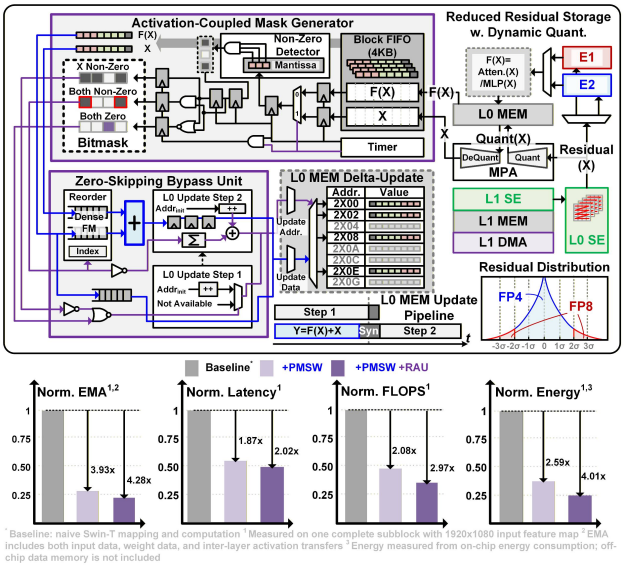
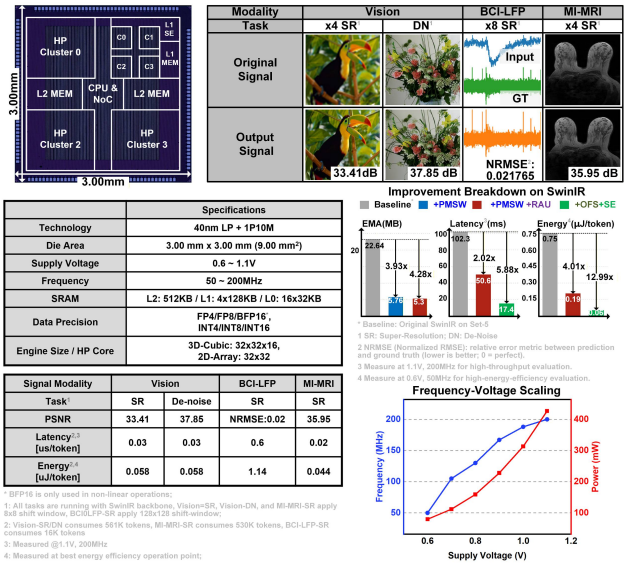


Fig. 6. Proposed residual-aware update circuit designs and the achieved improvements.



	Specifications
Technology	40nm LP + 1P10M
Die Area	3.00 mm x 3.00 mm (9.00 mm²)
Supply Voltage	0.6 ~ 1.1V
Frequency	50 ~ 200MHz
SRAM	L2: 512KB / L1: 4x128KB / L0: 16x32KB
Data Precision	FP4/FP8/BFP16*, INT4/INT8/INT16
Engine Size / HP Core	3D-Cubic: 32x32x16, 2D-Array: 32x32

Signal Modality	Vision		BCI-LFP	MI-MRI
Task[1]	SR	De-noise	SR	SR
PSNR	33.41	37.85	NRMSE:0.02	35.95
Latency[2,3] [us/token]	0.03	0.03	0.6	0.02
Energy[2,4] [uJ/token]	0.058	0.058	1.14	0.044

* BFP16 is only used in non-linear operations;
1: All tasks are running with SwinIR backbone, Vision=SR, Vision-DN, and MI-MRI-SR apply 8x8 shift window, BCI&LFP-SR apply 128x128 shift-window;
2: Vision-SR/DN consumes 561K tokens, MI-MRI-SR consumes 530K tokens, BCI-LFP-SR consumes 16K tokens
3: Measured @1.1V, 200MHz
4: Measured at best energy efficiency operation point;

Fig. 7. Chip die photo, voltage-frequency scaling curve, demonstration across four SR tasks, and acceleration analysis.

TABLE I
COMPARISON WITH STATE-OF-THE-ART WORKS

	JSSC 2023 [1]	ISSCC 2025 [2]	VLSI 2025 [3]	VLSI 2025 [4]	TCAS-I 2025 [5]	This Work
Function	CNN	CNN	CNN	CNN	Swin-T	Swin-T/CNN
Technology (nm)	65	40	16	28	FPGA-based	40
Core Engine Implementation	Monolithic	Monolithic	Monolithic	Monolithic	Monolithic	Heterogenous Cubic/Array
OP-Fusion Fabric	-	Inter-Core	Inter-Core	Inter-Core	-	Inter-Engine/Core
Supply Voltage (V)	1.0	0.55-0.95	0.6-1.1	0.7-1.1	-	0.6-1.1
Frequency (MHz)	200	10-200	100-400	50-250	170	50-200
Die Area (mm²)	10.00	6.8	4.00	20.25	-	9.00
Precision	INT8 + 10% FP8	INT8	INT8	INT8/16	INT8	INT4/8/16, FP4/8, BFP16
Power (mW)	98	25-344	127-1158	-	160-970	79.56[a]-426.7[b]
Throughput (TOPS or TFLOPS)	-	34	5.1	61.4-117.1	0.83	3.5[a,f]-37.6[a,e,g]
Energy Efficiency (TOPS/W or TFLOPS/W)	-	98.8-163.2	4.4-10	8.6-36.7	0.045	43.99[a,e,f]-427.5[a,e,g]
Application	SR	SR	SR / DN	SR / DN	Object Detection	Multi-Modal SR / DN
Signal Modality	Vision	Vision	Vision	Vision	Vision	Vision, MRI, BCI
Algorithm Model	FSR-CNN	-	U-Net	CNN	Swin-ViT	SwinIR[c]
Vision PSNR	2K: 26.1dB	4K: 27.8dB	4K: 31.68dB	4K: 32.0dB	-	4K: 33.41dB
Energy Consumption	0.92mJ/frame	-	-	10.8mJ/frame	-	3.25mJ/frame[a,c], 0.058uJ/token[a,c]
Frame Rate (FPS)	-	4K@210[d]	4K@30	4K@30	-	4K@57.47[b,c]
Latency	-	48ms/frame	-	-	-	17.4ms/frame[b,c]

a: @0.6V, 50MHz; b: @1.1V, 200MHz; c: SwinIR model with 1K input, x4 upscale d: inter-frame normalization result; e: Assume 0% sparsity, full utilization, 1MAC=2OPS; f: INT8 precision; g: FP4 precision.

B. Residual-Aware Update (RAU)

The deep residual connections in Swin-T ($Y = F(X) + X$) force a continuous, brute-force "read-modify-write" penalty: massive intermediate feature maps (X) must be repeatedly fetched from external memory, added to ($F(X)$) element-by-element, and the entire updated tensor (Y) written back, wasting immense bandwidth. Our RAU scheme optimizes this at three levels (Fig. 6): **Dynamic Quantization:** Leveraging the Laplacian distribution of residual features, a Multi-Precision Aligner (MPA) dynamically quantizes the residual X to 4-bit formats while maintaining the main branch $F(X)$ in 8-bit, halving the residual storage footprint. **Activation-Coupled Zero-Skipping:** Due to non-linear activations, $F(X)$ exhibits high sparsity. An Activation-Coupled Mask Generator generates a bitmask identifying non-zero indices. A Zero-Skipping Bypass Unit utilizes this mask to clock-gate *specific addition cells* within the 2D-Array engine. This skips the element-wise addition only for zero-valued tokens, without disabling or stalling the entire engine. **L0 MEM Delta-Update:** Instead of writing back the entire updated tensor Y, the delta-update pipeline only writes back values where $F(X) \neq 0$. This configuration-driven sparse update minimizes L0 write energy and reduces internal memory traffic drastically.

IV. Implementation Results

Fabricated in 40nm CMOS (Fig. 7), the 9.00 mm² Swin-SR chip operates from 50–200 MHz at 0.6–1.1 V. For vision tasks (×4 Super-Resolution, Denoising), Swin-SR achieves up to 37.85 dB PSNR. For BCI neuronal spike reconstruction (LFP ×8 SR) and Medical Imaging (MRI ×4 SR), the processor delivers high-fidelity outputs with ultra-low latency (down to 0.02 μs/token). At its optimal energy-efficiency point (0.6 V, 50 MHz), Swin-SR achieves a peak energy efficiency of **43.99 TOPS/W** (INT8). The architectural and memory-access innovations—specifically PMSW, RAU, and OPS—collectively deliver a **12.99× total energy reduction** and a **5.88× latency speedup** over a baseline monolithic implementation without these features. Table I compares Swin-SR with state-of-the-art works [1]–[5]. Swin-SR is the first chip to provide native, highly-optimized hardware support for multi-modal signal restoration. It outperforms recent designs by delivering significantly lower energy consumption while supporting high-resolution, multi-modal generative tasks.

References

[1] Z. Li et al. An efficient deep-learning-based super-resolution accelerating soc with heterogeneous accelerating and hierarchical cache. *IEEE Journal of Solid-State Circuits*, 58(3):614–623, 2023.

[2] Y. Lin et al. 2.8: A 210fps image signal processor for 4k ultra hd true video super resolution. In *2025 IEEE International Solid-State Circuits Conference (ISSCC)*, volume 68, pages 58–60, 2025.

[3] Y. Chen et al. Cattus: A 4k-uhd 30fps deep image processor for channel attention equipped u-net acceleration in 16nm finfet. In *2025 Symposium on VLSI Technology and Circuits (VLSI Technology and Circuits)*, pages 1–3, 2025.

[4] S. Kim et al. Nuvpu: A 4.8-9.6mj/frame progressive ntt-based unified video processor for stable video streaming and processing with neural video codec. In *2025 Symposium on VLSI Technology and Circuits (VLSI Technology and Circuits)*, pages 1–3, 2025.

[5] Q. Dong et al. An efficient window-based vision transformer accelerator via mixed-granularity sparsity. *IEEE Transactions on Circuits and Systems I: Regular Papers*, 72(9):4751–4764, 2025.

[6] J. Liang et al. Swinir: Image restoration using swin transformer. In *2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pages 1833–1844, 2021.

[7] N. Hong et al. Machine learning-based high-frequency neuronal spike reconstruction from low-frequency and low-sampling-rate recordings. *Nature Communications*, 15(635), 2024.

[8] P. Sousa et al. Swin transformer applied to breast mri super-resolution in a cross-cohort dataset. In *2025 IEEE 38th International Symposium on Computer-Based Medical Systems (CBMS)*, pages 1–6, 2025.