



# Sukhyun Han

**Ph.D Student      Sungkyunkwan University**  
kavin1010@g.skku.edu      &      +82) 10-6709-2535  
Suwon, Republic of Korea  
 sukhyun-han1010       hsh-notes

## Research Interests

---

Domain Specific Accelerator (NPU) Architecture

- Deep Learning Accelerator Architecture
- Hardware-Software Co-Design Methodology

AI Acceleration

- Optimization for DL Accelerators
- LLM model Performance Engineering

## Education

---

### Ph.D Course

Sungkyunkwan university, Suwon, South Korea (Aug. 2025 ~ Present)  
Department of Electrical and Computer Engineering

### Master's Degree

Sungkyunkwan university, Suwon, South Korea (Mar. 2024~ Aug. 2025)  
Department of Semiconductor Convergence Engineering

### Bachelor's Degree

Sungkyunkwan university, Suwon, South Korea (Mar. 2019 ~ Mar. 2024)  
Department of Electronic and Electrical Engineering

## Skills and Techniques

---

Programming Languages

- Digital system design with Verilog HDL
- Software programming with C, C++, Python, CUDA on Linux system

EDA Tool

- Synopsys Design Compiler
- Vivado FPGA Suite
- Cadence Virtuoso Suite

## Honors and Awards

---

### 2022 AI Semiconductor design Competition (4th place prize)

- Design General HW Accelerator based on Row-stationary Systolic Array.
- Our team aimed to design and verify accelerator architectures that optimize dataflow and maximize throughput by row-stationary SA.

### 2022 Deep learning Hardware Design Competition (Popularity Award)

- Design tiny-YOLOv3 HW Accelerator with Verilog HDL.
- Our team aimed to design and verify accelerator architectures that maximize throughput by minimizing off-chip memory access and leveraging maximum board resources. My work included implementation of Fused-Conv layer which is integrated with Pooling layer in CNN, reducing the access of On-chip SPM.

## Major Publications

---

### **Avalanche: Optimizing Cache Utilization via Matrix Reordering for Sparse Matrix Multiplication Accelerator**

- 2025 International Symposium on Computer Architecture (ISCA 2025)
- Gwangeun Byeon, Seongwook Kim, Hyungjin Kim, **Sukhyun Han**, Jinkwon Kim, Prashant Nair, Taewook Kang, Seokin Hong
- This paper addresses cache contention in sparse matrix multiplication accelerators with tiled outer product dataflow. The proposed architecture maximizes on-chip cache utilization by evicting early non-important data to external memory. My major contribution to this paperwork is to give advice about some details of the paper writing.

### **Zebra: Leveraging Diagonal Attention Pattern for Vision Transformer Accelerator**

- 2025 Design, Automation & Test in Europe Conference & Exhibition (DATE 2025)
- **Sukhyun Han**, Seongwook Kim, Gwangeun Byeon, Jihun Yoon, Seokin Hong
- This paper proposes a novel accelerator for ViT. This paper reveals that the self-attention map of ViT models usually show the diagonal pattern, and it is possible to apply pruning with respect to the pattern. To leverage these sparse diagonal patterns, this paper proposes novel hardware architecture that only compute necessary vector operations in attention mechanism. I am the first author of this paper.

## Projects and Research Experiences

---

### **Research Fellow**

Aug. 2023 - Present

Computer Architecture and Systems Lab – COMPASS LAB SKKU

- Profiling and Performance Analysis for Transformer models.
- Teaching Assistant for Semiconductor System Engineering courses.
- Maintain and Manage GPU Computing server.

### **Comprehensive HW/SW Solution for Edge Self Supervised Learning** COMPASS Lab

Sep. 2023 - Present

- Co-working with Hanwha Systems Inc.
- Evaluate the performance of Mobile AP embedded with NPU for ISAR-based object detection and classification task. My work included implementation of efficient ISAR-based algorithms on Mobile AP and design the control application.

### **NPU Technology for Accelerating Ladar-based Object Detection** COMPASS Lab

Sep. 2023 - Feb. 2025

- Design NPU Microarchitecture including ISA and the programming model integrated RISC-V Instruction Set Simulator.
- My work included managing setup and documentation of Spike Simulator SDK and contributing to the development of RISC-V based extended ISA for edge NPU specialized for Self-Supervised Learning.

### **Samsung Computer Engineering Challenge 2023**

Aug. 2023 - Nov. 2023

- Optimize LLaMav2 model Inference without accuracy drop.
- My work included implementing multi-GPU parallelization and advanced dataset sorting with dynamic padding reduction to minimize redundant computations.

### **Robust BGR and LDO Design against PVT Variations**

Jun. 2022 - Aug. 2022

CIRCUITS Lab in SKKU (Undergraduate Research Program)

- Design and evaluate AMP, BGR, and LDO circuits meeting specifications using 180nm TSMC process.
- My work included Improving and analyzing performance for PVT variation robustness by utilizing Thick PMos and Cascode PMos devices.

### **Implementation of Neural Network Computation using Low-Power CIM Circuit**

Mar. 2023 - Jul. 2023

Undergraduate Thesis

- Undergraduate Poster session at the 2023 International Technical Conference on Circuits/Systems,

Computers, and Communications (ITC-CSCC).

- My work included Implementation an energy-efficient Computing-in-Memory circuit based on 8T SRAM by incorporating a Layer Recycling technique.

**Real Time Face Recognition using MTCNN and FaceNet with TensorRT**

Sep. 2021 - Dec. 2021

Undergraduate Course Project

- Optimize MTCNN and FaceNet algorithms on Jetson Nano using CUDA for GPU/CPU co-processing.
- My work included real-time multi-face recognition with K-NN algorithm and enabled new face registration functionality, enhancing accuracy.