

Sehyeon Oh

HARDWARE ACCELERATOR · QUANTIZATIONR · SYSTEM SOFTWARE

☎ (+82) 10-4159-7632 | ✉ somae0604@gmail.com | 🌐 osehyeon

Education

UST(Korea University of Science and Technology)

M.S. IN ARTIFICIAL INTELLIGENCE

- GPA: 4.44/4.5

Daejeon, S.Korea

Mar. 2024 - Feb. 2026 (Expected)

Dong-Eui University

B.S. IN ROBOT AND AUTOMATION ENGINEERING (MINOR: APPLIED SOFTWARE)

- GPA: 4.4/4.5

Busan, S.Korea

Mar. 2018 - Feb. 2024

Publications

Journal

Luthier: Bridging Auto-Tuning and Vendor Libraries for Efficient Deep Learning Inference [Paper]

ACM TRANSACTIONS ON EMBEDDED COMPUTING SYSTEMS(SCIE)

2025

- Yongin Kwon, JooHyounG Cha, **Sehyeon Oh**, Misun Yu, Jeman Park, Jemin Lee

Data Transmission Optimization for NPU System Scalability [Paper]

IEMEK JOURNAL OF EMBEDDED SYSTEMS AND APPLICATIONS(KCI)

2025

- Sehyeon Oh**, Yongin Kwon, Jemin Lee

Optimizing Real-Time Object Detection in a Multi-Neural Processing Unit System [Paper]

SENSORS (SCIE)

2025

- Sehyeon Oh**, Yongin Kwon, Jemin Lee

Conference

TriPlanNet: Triangle Path Planning Network for A Variable Truss Robot with Deep Learning [Paper]

THE 13TH INTERNATIONAL WORKSHOP ON ASSISTIVE COMPUTER VISION AND ROBOTICS, ICCV

2025

- Choonghan Lee, Leah Harris, **Sehyeon Oh**, JooHyounG Cha, Jemin Lee, Yongin Kwon, Andrew Jang-ho Bae

Work-in-Progress: I-FlashAttention: Fully Integer Fused Attention for Efficient Vision Transformers [Paper]

INTERNATIONAL CONFERENCE ON COMPILERS, ARCHITECTURES, AND SYNTHESIS FOR EMBEDDED SYSTEMS (CASES)

2025

- Sehyeon Oh**, Yongin Kwon, Jemin Lee

Optimizing Real-Time Object Detection in a Multi NPU System with Double Buffering and Queue-Based Processing [Paper]

INTERNATIONAL CONFERENCE ON ARTIFICIAL INTELLIGENCE COMPUTING AND SYSTEMS (AICOMPS)

2024

- Sehyeon Oh**, Yongin Kwon, Jemin Lee

PCIe-Based Multi-NPU Data Transmission Optimization [Paper]

IEMEK AUTUMN CONFERENCE

2024

- Sehyeon Oh**, Yongin Kwon, Jemin Lee

Projects

- | | | |
|------|---|----------|
| 2025 | Implementation of an integer arithmetic kernel based on Triton-lang | [Github] |
| 2024 | Optimized the object detection on the Neubla NPU for the ETRI Conference 2024 | [NEWS] |
| 2024 | C-based implementation of AI operators and models for NPU verification | [Github] |

Honors & Awards

- | | | |
|------|--|-------|
| 2025 | UST Research Paper Award , UST Research Paper Award 2025 | UST |
| 2025 | WoonDang Student Paper Award , WoonDang Academic Award Ceremony | KIPS |
| 2025 | Encuragement Award , ARX/RVX-Based AI Semiconductor Design Competition | ISE |
| 2024 | Distinguished Paper Award , International Conference on Artificial Intelligence Computing and Systems | KIPS |
| 2024 | Distinguished Paper Award , IEMEK Autumn Conference | IEMEK |