

Jaehyeong Sim

+82-2-3277-6547 | jh.sim@ewha.ac.kr | <https://acpl.ewha.ac.kr/>

 [LinkedIn](#) |  [Google Scholar](#) | Updated 2026-10-12

Seoul, 03760, Republic of Korea

Summary

I am an accomplished researcher and academic with extensive expertise in hardware-software co-design, neural processing units (NPUs), digital systems, and efficient deep learning. I currently serve as an Associate Professor at the Department of Computer Science and Engineering, Ewha Womans University, where he leads the AI Computing Platform Laboratory, supervising 8 graduate students. My professional experience includes practical work at Samsung Advanced Institute of Technology, contributing to industrial NPU hardware architecture and performance analysis. I have spearheaded diverse projects, such as on-device learning for embedded systems and automated FPGA-based AI semiconductor design frameworks. I hold numerous patents, have published prolifically in premier journals and conferences, and received the Best Paper Award at IEEE ICCD. With my interdisciplinary expertise and a track record of bridging hardware and software for AI systems, I believe that I am well-positioned to excel in roles driving innovation at the intersection of deep learning and hardware design.

Experience

- **Ewha Womans University** *September 2026 - Present*
Seoul, Republic of Korea
Associate Professor (Tenure Track)
 - Department of Computer Science and Engineering, School of Artificial Intelligence
- **Ewha Womans University** *April 2021 - August 2026*
Seoul, Republic of Korea
Assistant Professor (Tenure Track)
 - Department of Computer Science and Engineering, School of Artificial Intelligence
 - Department of Artificial Intelligence (Former Adjunct)
 - HOKMA College of General Education (Former Adjunct)
- **Samsung Advanced Institute of Technology (SAIT)** *March 2020 - March 2021*
Suwon, Republic of Korea
Staff Researcher
 - Developed datacenter NPU hardware architecture.
 - Developed variable-precision processing elements for NPU.
 - Conducted performance and power analysis for NPU and its PEs.
- **Korea Advanced Institute of Science and Technology** *March 2019 - February 2020*
Daejeon, Republic of Korea
Postdoctoral Researcher
 - Actively participated in a hardware-software co-design framework to accelerate CNN models for a super-resolution task.
 - Actively participated in developing customized 4T embedded DRAM cell array architecture for accelerating binary neural networks.
 - Actively participated in developing processing in-memory architecture for accelerating LSTM based on SOT-MRAM.

Education

- **Korea Advanced Institute of Science and Technology (KAIST)** *March 2014 - February 2019*
Daejeon, Republic of Korea
PhD in Electrical Engineering
 - Advisor: Prof. Lee-Sup Kim, IEEE Fellow
 - Thesis: Energy Efficient Processors and In-DRAM Processing Framework for Deep Convolutional Neural Network
 - Have experience of 6 silicon tapeouts.
- **Korea Advanced Institute of Science and Technology (KAIST)** *February 2012 - February 2014*
Daejeon, Republic of Korea
Master in Electrical Engineering
 - Advisor: Prof. Lee-Sup Kim, IEEE Fellow
 - Thesis: Timing Error Masking by Exploiting Operand Value Locality in SIMD Architecture
- **Korea Advanced Institute of Science and Technology (KAIST)** *February 2007 - February 2012*
Daejeon, Republic of Korea
Bachelor in Electrical Engineering
 - Cum Laude

Research Projects

- **Basic Research Laboratory, National Research Foundation of Korea (NRF)** *June 2025 - May 2028*
Energy-Efficient, General-Purpose Multi-Modal AI with Heterogeneous Computing Accelerators Co-Investigator
 - Developing an NPU architecture for multi-modal AI with multi-tenancy.
- **LG Electronics** *March 2025 - December 2025*
Transformer-based On-Device Segmentation for Next Generation Robotic Vacuum Cleaner Principal Investigator
 - Improving on-device efficiency of SAM2 model on NVIDIA Jetson Orin Nano platform.
- **LG Electronics** *January 2024 - December 2024*
On-Device Learning of an Object Detection Model for Robotic Vacuum Cleaners Principal Investigator
 - Improved detection accuracy of YOLO-X model in a personalized in-door environment using knowledge distillation, few-shot learning, and domain adaptation.
 - Converted the proposed model into TFLite, and implemented C code to execute it on the commercial NPU developed by LG Electronics.
 - Implemented a few-shot learning and domain adaptation methods in C to execute them on the SoC embedded in the commercial robotic vacuum cleaner.
- **First Research in Life Time, National Research Foundation of Korea (NRF)** *March 2023 - February 2026*
End-to-End Design Automation Framework for FPGA-Based AI Semiconductor Principal Investigator
 - Developed a hardware-aware neural architecture search (NAS) framework for deep learning acceleration in a FPGA.
 - Developed an automatic architecture search framework for a FPGA-based deep learning accelerator.
- **AI Semiconductor Processing SW Research Center, IITP** *July 2022 - December 2029*
Hardware-Aware Neural Architecture Search and Accelerator Design Co-Investigator
 - Developed a hardware-aware neural architecture search framework for an NPU.
 - Developed an energy-efficient accelerator for YOLOX inference.
 - Developed a bit-serial architecture exploiting weight bit sparsity for efficient deep learning acceleration.
- **Artificial Intelligence Innovation Hub, IITP** *September 2021 - December 2025*
Self-evolving Multi-task Intelligence Co-Investigator
 - Developed a hardware-aware automatic operation search framework for a squash function for capsule networks.
 - Developed a novel token merging method to accelerate vision transformer models.

Publications

International Conference

- [C.28] Eunyoung Yoo, Eunseo Kim, Jaehyeong Sim. **LCM-KD: Teacher-Compatible Channel Merging for Recoverability-Oriented Compression and Distillation.** In *Asian Conference on Computer Vision (ACCV)*, 2026. (to appear).
- [C.27] Soeun Choi, Jieui Kang, Eunyoung Yoo, Yeonhee Kim, Jaehyeong Sim. **CSP-KD: Clustering-Based Sentence Pruning for Multi-Teacher Knowledge Distillation.** In *International Conference on Communications, Computing, Cybersecurity, and Informatics (CCCI)*, 2026. (to appear).
- [C.26] Jaeyoung Choi, Jaehyeong Sim. **RLE-Aware Weight Loader: A Lightweight Decompression Front-End for Fast LLM Cold Starts.** In *International SoC Design Conference (ISOCC)*, 2026. (to appear).
- [C.25] Eunjin Lee, Jiho Lee, Hayoung Lim, Jaehyeong Sim. **MAGNETO: A Genetic Algorithm-Based Power-Aware Mapping Optimization Framework for Mobile NPUs.** In *International Conference on Communications, Computing, Cybersecurity, and Informatics (CCCI)*, 2025.
- [C.24] Soeun Choi, Yejin Lee, Juhee Kim, Minji Kim, Jaehyeong Sim. **DS-CAE: a Dual-Stream Cross-Attentive Autoencoder for Robust and Cluster-Aware Retrieval-Augmented Generation.** In *International Conference on Communications, Computing, Cybersecurity, and Informatics (CCCI)*, 2025.
- [C.23] Eunjin Lee, Eunseo Kim, Eunyoung Yoo, Jaehyeong Sim. **GATHER: A Gated-Attention Accelerator for Efficient LLM Inference.** In *International SoC Design Conference (ISOCC)*, 2025.
- [C.22] Soeun Choi, Jaehyeong Sim. **LoRA-PIM: In-Memory Delta-Weight Injection for Multi-Adapter LLM Serving.** In *International SoC Design Conference (ISOCC)*, 2025.
- [C.21] Eunyoung Yoo, Jaehyeong Sim. **ViT-Slim: A Genetic Algorithm-based NAS Framework for Efficient Vision Transformer Design.** In *IEEE Conference on Artificial Intelligence (CAI)*, 2025.
- [C.20] Chaeyun Kim, Eunseo Kim, Yeonhee Kim, Jaehyeong Sim, Jonkil Kim. **Enhancing Gender Prediction Model Performance through Automatic Individual Entity Extraction and Class Balance.** In *IEEE International Conference on Big Data and Smart Computing (BigComp)*, 2025.

- [C.19] Jieui Kang, Subeun Lee, Eunseo Kim, Soeun Choi, **Jaehyeong Sim**. **OCW: Enhancing Few-Shot Learning with Optimized Class-Weighting Methods**. In *International Conference on Communications, Computing, Cybersecurity, and Informatics (CCCI)*, 2024.
- [C.18] Jieui Kang, Sooyoung Kwon, Hyojin Kim, **Jaehyeong Sim**. **AutoCaps-Zero: Searching for Hardware-Efficient Squash Function in Capsule Networks**. In *International Conference on Communications, Computing, Cybersecurity, and Informatics (CCCI)*, 2024.
- [C.17] Jiho Lee, Jieui Kang, Eunjin Lee, Yejin Lee, **Jaehyeong Sim**. **AlphaAccelerator: An Automatic Neural FPGA Accelerator Design Framework Based on GNNs**. In *International SoC Design Conference (ISOCC)*, 2024.
- [C.16] Eunseo Kim, Subeun Lee, Chaeyun Kim, Hayoung Lim, Jimin Nam, **Jaehyeong Sim**. **BS2: Bit-Serial Architecture Exploiting Weight Bit Sparsity for Efficient Deep Learning Acceleration**. In *International SoC Design Conference (ISOCC)*, 2024.
- [C.15] Kyungmi Kim, Soeun Choi, Eunkyeol Hong, Yoonseo Jang, **Jaehyeong Sim**. **An Energy-Efficient Hardware Accelerator for On-Device Inference of YOLOX**. In *International SoC Design Conference (ISOCC)*, 2024.
- [C.14] Hayoung Lim, Yeseo Jang, Juyeon Kim, **Jaehyeong Sim**. **TD-NAAS: Template-Based Differentiable Neural Architecture Accelerator Search**. In *International SoC Design Conference (ISOCC)*, 2023.
- [C.13] Jieui Kang, **Jaehyeong Sim**, Hyokyung Bahn. **Optimization of the Modified Gaussian Filter for Mobile GPU Usage in Game Workloads**. In *International Conference on Communications, Computing, Cybersecurity, and Informatics (CCCI)*, 2023.
- [C.12] Seungkyu Choi, **Jaehyeong Sim**, Myeonggu Kang, Yeongjae Choi, Hyeonuk Kim, Lee-Sup Kim. **A 47.4 uJ/epoch Trainable Deep Convolutional Neural Network Accelerator for In-Situ Personalization on Smart Devices**. In *IEEE Asian Solid-State Circuits Conference (A-SSCC)*, 2019.
- [C.11] Hyein Shin, **Jaehyeong Sim**, Daewoong Lee, Lee-Sup Kim. **A PVT-Robust Customized 4T Embedded DRAM Cell Array for Accelerating Binary Neural Networks**. In *IEEE/ACM International Conference on Computer-Aided Design (ICCAD)*, 2019.
- [C.10] Kyeonghan Kim, Hyein Shin, **Jaehyeong Sim**, Myeonggu Kang, Lee-Sup Kim. **An Energy-Efficient Processing-in-Memory Architecture for Long Short Term Memory in Spin Orbit Torque MRAM**. In *IEEE/ACM International Conference on Computer-Aided Design (ICCAD)*, 2019.
- [C.9] Youngbeom Jung, Yeongjae Choi, **Jaehyeong Sim**, Lee-Sup Kim. **eSRCNN: A Framework for Optimizing Super-Resolution Tasks on Diverse Embedded CNN Accelerators**. In *IEEE/ACM International Conference on Computer-Aided Design (ICCAD)*, 2019.
- [C.8] Hyeonuk Kim, **Jaehyeong Sim**, Yeongjae Choi, Lee-Sup Kim. **NAND-Net: Minimizing Computational Complexity of In-Memory Processing for Binary Neural Networks**. In *International Symposium on High-Performance Computer Architecture (HPCA)*, 2019.
- [C.7] **Jaehyeong Sim**, Hoseok Seol, Lee-Sup Kim. **NID: Processing Binary Convolutional Neural Network in Commodity DRAM**. In *IEEE/ACM International Conference on Computer-Aided Design (ICCAD)*, 2018.
- [C.6] Seungkyu Choi, **Jaehyeong Sim**, Myeonggu Kang, Lee-Sup Kim. **TrainWare: A Memory Optimized Weight Update Architecture for On-Device Convolutional Neural Network Training**. In *IEEE International Symposium on Low-Power Electronics and Design (ISLPED)*, 2018.
- [C.5] Myung-Hoon Choi, Seungkyu Choi, **Jaehyeong Sim**, Lee-Sup Kim. **SENIN: An Energy-Efficient Sparse Neuromorphic System with On-Chip Learning**. In *IEEE International Symposium on Low-Power Electronics and Design (ISLPED)*, 2017.
- [C.4] Hyeonuk Kim, **Jaehyeong Sim**, Yeongjae Choi, Lee-Sup Kim. **A Kernel Decomposition Architecture for Binary-Weight Convolutional Neural Networks**. In *IEEE/ACM Design Automation Conference (DAC)*, 2017.
- [C.3] **Jaehyeong Sim**, Jun-Seok Park, Minhye Kim, Dongmyung Bae, Yeongjae Choi, Lee-Sup Kim. **A 1.42 TOPS/W Deep Convolutional Neural Network Recognition Processor for Intelligent IoT Systems**. In *IEEE International Solid-State Circuits Conference (ISSCC)*, 2016.
- [C.2] **Jaehyeong Sim**, Jun-Seok Park, Seungwook Paek, Lee-Sup Kim. **Timing Error Masking by Exploiting Operand Value Locality in SIMD Architecture**. In *IEEE International Conference on Computer Design (ICCD)*, 2014.
- [C.1] Seungwook Paek, Seok-Hwan Moon, Wongyu Shin, **Jaehyeong Sim**, Lee-Sup Kim. **PowerField: A Transient Temperature-to-Power Technique Based on Markov Random Field Theory**. In *IEEE/ACM Design Automation Conference (DAC)*, 2012.

- [J.14] Soeun Choi, Jaehyeong Sim. Genetic Fragmentation Gradient Descent Failure Constrained Scheduling for GPU Sharing Clusters. *Journal of Supercomputing*, 2026. (to appear).
- [J.13] Eunjoung Yoo, Jieui Kang, Soeun Choi, Yeonhee Kim, Jaehyeong Sim. Token-Based Task-Aware Knowledge Distillation for Encoder Adaptation. *IEEE Access*, Vol. 14, pp. 99526 - 99540, 2026.
- [J.12] Jieui Kang, Jaeyoung Choi, Wonhui Roh, Jaehyeong Sim. QubitCache: Quantum-Inspired Probabilistic Attention Preservation for KV-Cache Compression. *IEEE Access*, Vol. 14, pp. 57983 - 57996, 2026.
- [J.11] Jieui Kang, Eunjoung Yoo, Soeun Choi, Yeonhui Kim, Jaehyeong Sim. SHARP: Structured Hierarchical Attention Rank Projection for Efficient Language Model Distillation. *IEEE Access*, Vol. 14, pp. 56679 - 56693, 2026.
- [J.10] Jieui Kang, Hyungon Ryu, Jaehyeong Sim. PRISM-Med: Parameter-efficient Robust Interdomain Specialty Model for Medical Language Tasks. *IEEE Access*, Vol. 13, pp. 4957-4965, 2025.
- [J.9] Jieui Kang, Soeun Choi, Eunjin Lee, Jaehyeong Sim. SpDRAM: Efficient In-DRAM Acceleration of Sparse Matrix-Vector Multiplication. *IEEE Access*, Vol. 12, pp. 176009-176021, 2024.
- [J.8] Jieui Kang, Jihye Park, Soeun Choi, Jaehyeong Sim. Q-LATte: An Efficient and Versatile LSTM Model for Quantized Attention-Based Time Series Forecasting in Building Energy Applications. *IEEE Access*, Vol. 12, pp. 69325-69341, 2024.
- [J.7] Myeonggu Kang, Hyeonuk Kim, Hyein Shin, Jaehyeong Sim, Kyeonghan Kim, Lee-Sup Kim. S-FLASH: A NAND Flash-based Deep Neural Network Accelerator Exploiting Bit-Level Sparsity. *IEEE Transactions on Computers*, Vol. 71, Issue 6, pp. 1291-1304, 2022.
- [J.6] Yeongjae Choi, Jaehyeong Sim, Lee-Sup Kim. CREMON: Cryptography Embedded on the Convolutional Neural Network Accelerator. *IEEE Transactions on Circuits and Systems II - Express Briefs*, Vol. 67, Issue 12, pp. 3337-3341, 2020.
- [J.5] Seungkyu Choi, Jaehyeong Sim, Myeonggu Kang, Yeongjae Choi, Hyeonuk Kim, Lee-Sup Kim. An Energy-Efficient Deep Convolutional Neural Network Training Accelerator for In Situ Personalization on Smart Devices. *IEEE Journal of Solid-State Circuits*, Vol. 55, Issue 10, pp. 2691-2702, 2020.
- [J.4] Jaehyeong Sim, Somin Lee, Lee-Sup Kim. An Energy-Efficient Deep Convolutional Neural Network Inference Processor with Enhanced Output Stationary Dataflow in 65-nm CMOS. *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, Vol. 28, Issue 1, pp. 87-100, 2019.
- [J.3] Yeongjae Choi, Dongmyung Bae, Jaehyeong Sim, Seungkyu Choi, Minhye Kim, Lee-Sup Kim. Energy-Efficient Design of Processing Element for Convolutional Neural Network. *IEEE Transactions on Circuits and Systems II - Express Briefs*, Vol. 64, Issue 11, pp. 1332-1336, 2017.
- [J.2] Taeho Lee, Yong-Hun Kim, Jaehyeong Sim, Jun-Seok Park, Lee-Sup Kim. A 5-Gb/s 2.67-mW/Gb/s Digital Clock and Data Recovery with Hybrid Dithering Using a Time-Dithered Delta-Sigma Modulator. *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, Vol. 24, Issue 4, pp. 1450-1459, 2015.
- [J.1] Seungwook Paek, Wongyu Shin, Jaehyeong Sim, Lee-Sup Kim. PowerField: A Probabilistic Approach for Temperature-to-Power Conversion Based on Markov Random Field Theory. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, Vol. 32, Issue 10, pp. 1509-1519, 2013.

Domestic Conference

- [DC.7] 박나담, 이나경, 이주원, 심재형. ProgressiveServe: 서버리스 LLM 콜드 스타트 완화를 위한 점진적 모델 로딩 및 복구 기법. In 한국소프트웨어종합학술대회 (KSC), 2025.
- [DC.6] 김민서, 김수현, 하지연, 심재형. 메모리 용량 제약 하에서 하드웨어 최적화 트랜스포머 설계를 위한 HPO-NAS 통합 프레임워크. In 대한전자공학회 하계학술대회 (IEIE-Summer), 2025.
- [DC.5] 류이정, 박지원, 소예림, 최종원, 심재형. T-FLIP: 어텐션 가중치 기반 지식 증류를 통한 안면 위조 방지 모델 경량화. In 대한전자공학회 추계학술대회 (IEIE-Autumn), 2024.
- [DC.4] 임하영, 김경미, 장예서, 김주연, 심재형. QTNAAS: 템플릿 기반 양자화된 신경망 구조 및 가속기 탐색 프레임워크. In 대한전자공학회 추계학술대회 (IEIE-Autumn), 2023.
- [DC.3] 권수영, 권민서, 김효진, 심재형. ToMato: Token Merging을 이용한 Vision Transformer 가속화. In 대한전자공학회 추계학술대회 (IEIE-Autumn), 2023.
- [DC.2] 강지의, 심재형, 반효경. 게임 워크로드에 최적화된 모바일 GPU 설계방안 연구. In 한국소프트웨어종합학술대회 (KCC), 2022.

[DC.1] 김정민, 박지민, 이로운, 조서원, 심재형. 딥러닝 기반의 MBTI 성격유형 분류 연구. In 한국통신학회 하계종합학술발표회 (KICS), 2022.

Patents

Granted Patents

- [P.12] 계층적 주의 랭크 투영에 기반한 언어모델 지식 증류 장치 및 방법. Korean Intellectual Property Office (KIPO), Patent No. 10-3022935, 2026.
- [P.11] 토큰 병합을 이용한 비전 트랜스포머 장치 및 방법. Korean Intellectual Property Office (KIPO), Patent No. 10-3022917, 2026.
- [P.10] 가중치 매트릭스를 이용한 메모리 연산 처리 장치 및 방법. Korean Intellectual Property Office (KIPO), Patent No. 10-3013413, 2026.
- [P.9] 비트 직렬 연산 처리 장치 및 방법. Korean Intellectual Property Office (KIPO), Patent No. 10-2940812, 2026.
- [P.8] 그래프 신경망을 이용하여 하드웨어 구조의 설계를 가속하는 하드웨어 구조 설계 장치 및 하드웨어 구조 설계 방법. Korean Intellectual Property Office (KIPO), Patent No. 10-2897328, 2025.
- [P.7] 인공지능 기반의 스마트 윈도우 제어 시스템 및 제어 방법. Korean Intellectual Property Office (KIPO), Patent No. 10-2853870, 2025.
- [P.6] 가우시안 플러스 필터에 기반하는 이미지 처리장치 및 그 방법. Korean Intellectual Property Office (KIPO), Patent No. 10-2820700, 2025.
- [P.5] **Accelerator, method of operating an accelerator, and electronic device including an accelerator.** United States, Patent No. US12130756, 2024.
- [P.4] **Method and apparatus with deep learning operations with adder tree structure.** USPTO, Patent No. US12423057, 2024.
- [P.3] **Computing device and method for allocating resources using cost matrix.** USPTO, Patent No. US12175299, 2024.
- [P.2] **Method and apparatus for performing convolution operation in neural network.** Korean Intellectual Property Office (KIPO), Patent No. KR10-2452951, 2022.
- [P.1] **Neural network method and apparatus.** United States, Patent No. US10699160, 2020.

Patent Applications

- [A.23] 주파수 특성 기반 언어 모델의 컨텍스트 정보 압축 장치 및 방법. Korean Intellectual Property Office (KIPO), Application No. 10-2026-0123437, 2026.
- [A.22] ETA 기반 의사결정을 이용한 분산 학습 시스템의 장애 대응 장치 및 방법. Korean Intellectual Property Office (KIPO), Application No. 10-2026-0107448, 2026.
- [A.21] 멀티코어 신경망 처리 장치의 멀티모달 추론 장치 및 방법. Korean Intellectual Property Office (KIPO), Application No. 10-2026-0107362, 2026.
- [A.20] 비전 언어 모델의 다방향 관계 기반 지식 증류 장치 및 방법. Korean Intellectual Property Office (KIPO), Application No. 10-2026-0103636, 2026.
- [A.19] GPU 공유 클러스터에서 실패 제약을 고려한 경량 스케줄링을 위한 정책 학습 및 스케줄링 장치 및 방법. Korean Intellectual Property Office (KIPO), Application No. 10-2026-0074507, 2026.
- [A.18] 채널 라우팅 및 분기 연산을 이용한 트랜스포머 모델의 추론 장치 및 방법. Korean Intellectual Property Office (KIPO), Application No. 10-2026-0071793, 2026.
- [A.17] 서버리스 기반의 대규모 언어 모델 서빙 시스템 및 방법. Korean Intellectual Property Office (KIPO), Application No. 10-2025-0194750, 2025.
- [A.16] 대규모 언어모델의 키-값 캐시를 압축하기 위한 확률적 어텐션 보존 기반의 캐시 압축 시스템 및 방법. Korean Intellectual Property Office (KIPO), Application No. 10-2025-0179129, 2025.
- [A.15] 사용자 맞춤형 공간 제어가 가능한 인공지능 기반 실내 인테리어 변환 시스템 및 방법. Korean Intellectual Property Office (KIPO), Application No. 10-2025-0168914, 2025.
- [A.14] 계층적 주의 랭크 투영에 기반한 언어모델 지식 증류 장치 및 방법. PCT, Application No. PCT/KR2025/017383, 2025.
- [A.13] 검색 증강 지식 응답 생성 장치 및 방법. Korean Intellectual Property Office (KIPO), Application No. 10-2025-0159362, 2025.

- [A.12] 딥러닝 기반 이미지 처리 장치 및 방법. Korean Intellectual Property Office (KIPO), Application No. 10-2025-0105299, 2025.
- [A.11] 클러스터링 기반 문장 가지치기를 활용한 문장 기반 지식 증류 장치 및 동작 방법. Korean Intellectual Property Office (KIPO), Application No. 10-2025-0099179, 2025.
- [A.10] 동적 토큰 선택 및 동적 토큰 통합에 기반하여 태스크 인지 기반 지식 증류를 수행하는 태스크 인지 기반 지식 증류 장치 및 방법. Korean Intellectual Property Office (KIPO), Application No. 10-2025-0069129, 2025.
- [A.9] 토큰 병합을 이용한 비전 트랜스포머 장치 및 방법. PCT, Application No. PCT/KR2024/018690, 2024.
- [A.8] 도메인 적응형 언어모델 처리 장치 및 방법. Korean Intellectual Property Office (KIPO), Application No. 10-2024-0094108, 2024.
- [A.7] 정확도 정보 및 유사도 정보를 이용한 양자화 인공지능 학습 처리 장치 및 방법. PCT, Application No. PCT/KR2024/006433, 2024.
- [A.6] 템플릿에 기반하는 신경 구조 탐색장치 및 그 방법. PCT, Application No. PCT/KR2024/005651, 2024.
- [A.5] 캡슐 네트워크의 스쿼시 함수 탐색장치 및 그 방법. PCT, Application No. PCT/KR2024/003896, 2024.
- [A.4] 최적화된 클래스 가중치를 이용한 인공지능 학습 처리 장치 및 방법. Korean Intellectual Property Office (KIPO), Application No. 10-2024-0031351, 2024.
- [A.3] 정확도 정보 및 유사도 정보를 이용한 양자화 인공지능 학습 처리 장치 및 방법. Korean Intellectual Property Office (KIPO), Application No. 10-2023-0194206, 2023.
- [A.2] 템플릿에 기반하는 신경 구조 탐색장치 및 그 방법. Korean Intellectual Property Office (KIPO), Application No. 10-2023-0178909, 2023.
- [A.1] 캡슐 네트워크의 스쿼시 함수 탐색장치 및 그 방법. Korean Intellectual Property Office (KIPO), Application No. 10-2023-0121855, 2023.

Skills

-
- **Programming Languages:** C, C++, Python, Java, Lua, Node.js, CUDA, Verilog HDL, System Verilog, SystemC, MATLAB, MySQL
 - **Machine Learning Framework:** Caffe2, TensorFlow, Keras, PyTorch
 - **Inference Framework:** TensorRT, TFLite, vLLM
 - **Hardware Design Tools:** NC-Verilog, Synopsys Design Compiler, Synopsys Fusion Compiler, Synopsys PrimeTime-PX, Synopsys IC Compiler, Synopsys Astro, Simvision, Verdi, Cadence Virtuoso, Cadence Allegro, OrCAD, Xilinx Vivado, Mentor Catapult HLS
 - **Hardware Simulation/Profiling Tools:** GPGPU-Sim, SCALE-Sim, MAESTRO, CACTI, DRAMSim2, HotSpot, NVIDIA Nsight Systems
 - **Cloud Technologies:** Amazon Web Service (EC2, Lambda, S3, DynamoDB, RDS, SageMaker)
 - **Specialized Area:** Digital System Design, NPU, Parallel Computing, Performance and Power Modeling, Deep Learning, Hardware-Software Co-Design, AutoML, Processing-in-Memory, Model Compression, DL Inference Optimization
 - **AI Services:** Claude Code, ChatGPT, Claude, Gemini, Obsidian, Antigravity, Cursor, Codex, Suno

Honors and Awards

-
- | | |
|---|---|
| <ul style="list-style-type: none"> • Teaching Excellence Award
<i>Ewha Womans University</i> • Rising Star
<i>Electronic & Information Research Information Center (EIRIC)</i> • Outstanding Paper Poster Award
<i>Samsung-KAIST Industry-Academia Cooperation</i> • IDEC Congress Chip Design Contest 2019 Best Design Award
<i>IC Design Education Center (IDEC)</i> • IDEC Congress Chip Design Contest 2019 Best Poster Award
<i>IC Design Education Center (IDEC)</i> • Best Paper Award
<i>IEEE 32nd International Conference on Computer Design (ICCD)</i> | <div>2025</div> <div>2023</div> <div>2019</div> <div>2019</div> <div>2019</div> <div>2014</div> |
|---|---|
- 

Academic Services

• IEEE Transactions on Circuit and Systems I: Regular Papers (TCAS-I) <i>External Reviewer</i>	2022 - 2025
• IEEE Transactions on Circuit and Systems II: Express Briefs (TCAS-II) <i>External Reviewer</i>	2019 - 2025
• IEEE Transactions on Very Large Scale Integration Systems (TVLSI) <i>External Reviewer</i>	2020 - Present
• IEEE Transactions on Computers (TC) <i>External Reviewer</i>	2024 - 2025
• IEEE Access <i>External Reviewer</i>	2023 - Present
• IEEE Transactions on Knowledge and Data Engineering (TKDE) <i>External Reviewer</i>	2025
• IEEE Transactions on Artificial Intelligence (TAI) <i>External Reviewer</i>	2025
• IEEE Transactions on Neural Network and Learning Systems (TNNLS) <i>External Reviewer</i>	2025
• IEEE Transactions on Reliability (TR) <i>External Reviewer</i>	2024
• IEEE Journal on Emerging and Selected Topics in Circuits and Systems (JETCAS) <i>External Reviewer</i>	2023
• IEEE Open Journal of the Solid-State Circuits Society (OJ-SSCS) <i>External Reviewer</i>	2022
• IEEE Transactions on Signal Processing (TSP) <i>External Reviewer</i>	2021
• Scientific Reports (Springer) <i>External Reviewer</i>	2024 - Present
• Signal, Image and Video Processing (Springer) <i>External Reviewer</i>	2024 - 2025
• Journal of Real-Time Image Processing (Springer) <i>External Reviewer</i>	2024 - 2025
• Analog Integrated Circuits and Signal Processing (Springer) <i>External Reviewer</i>	2025
• Journal of Nondestructive Evaluation (Springer) <i>External Reviewer</i>	2024
• The Journal of Supercomputing (Springer) <i>External Reviewer</i>	2024
• Nature (Springer) <i>External Reviewer</i>	2018
• IEEE International Symposium on Circuits and Systems (ISCAS) <i>External Reviewer</i>	2021
• ACM KDD <i>External Reviewer</i>	2026 - 2027

Additional Information

Languages: Korean (Native), English (Intermediate High)