Yuzhe Fu

No.5, Yiheyuan Road, Haidian District
Beijing, P.R.China, 10087

Tel: (+86) 18573907771
Email: fuyz@stu.pku.edu.cn

EDUCATION

Peking University − Beijing, China *September 2021 – June 2024*
Master in Microelectronics and Solid-State Electronics

- GPA: 3.59/4.0
- Supervisor: Prof. Hailong Jiao

University of California, Berkeley − Berkeley, USA *August 2019 – December 2019*
Global Access Program

- GPA: 3.94/4.0

Southern University of Science and Technology − Shenzhen, China *September 2017 – June 2021*
Bachelor of Microelectronics Science and Engineering

- GPA: 3.88/4.0, 1st in Postgraduate Recommendation Evaluation
- Supervisor: Prof. Fengwei An

RESEARCH INTERESTS

- Algorithm-hardware co-design for neural network accelerator
- Energy-efficient and configurable artificial intelligence accelerator design

PROJECT EXPERIENCE

SoftAct: High-Precision Softmax Processing Hardware Unit for Transformer Network with Integrated Nonlinear Activation Function Support
Project Leader *July 2022 – Present*

- Developed an enhanced softmax function with penalty, optimizing precision in hardware.
- Introduced sparsity detection for softmax to fully skip redundant operations
- Designed a specialized architecture to implement the proposed techniques and non-linear functions
- SoftAct is synthesized in TSMC 28 nm and simulated based on MobileViT data, attaining peer-leading precision and hardware efficiency
- Wrote a journal paper submitted to ***IEEE Transactions on Circuits and Systems I: Regular Papers (TCAS-I)***

Nebula: An Energy-Efficient 3D Point Cloud-Based Neural Network Accelerator
Main Contributor & Module Leader *February 2022 – Present*

- Responsible for software work (model training, quantization and pruning, algorithm verification, etc.)
- Proposed a novel FPS and its hardware unit, ensuring precision while mitigating complexity
- Implemented data-sensitive analysis and adaptive filter pruning to significantly compress network
- Co-designed a coarse-to-fine-grained data-reuse dataflow scheme, leveraging spatial-temporal locality, block-wise pipelined delay-aggregation, and MLP fusion
- Nebula is fabricated in TSMC HPC 28 nm (In-progress). The simulated effective energy efficiency exceeds 50 TOPS/W
- Co-wrote a conference paper submitted to ***IEEE/ACM International Conference On Computer Aided Design (ICCAD)***

Sagitta: An Energy-Efficient Sparse 3D-CNN Accelerator for Real-Time 3D Understanding
Contributor *December 2021 – December 2022*

- Responsible for pruning and extraction of network data for analysis
- Leveraged locality and small differential value dropout to increase the sparsity of activations
- Co-revised a journal paper submitted to ***IEEE Internet of Things Journal (IoT-J)*** *(Minor revision)*

SGM Accelerator: A 4.29nJ/pixel Stereo-depth Coprocessor with Pixel-level Pipeline and Region-Optimized Semi-Global Matching for IoT Application

Project Leader *February 2020 – June 2021*

- Proposed a region-optimized stereo matching strategy improving the speed of traditional semi-global matching algorithm by 5 times while ensuring the accuracy
- Proposed a four-layer parallel pipeline hardware architecture and implemented it on FPGA platform which can extract depth information in real-time at 156MHz and 508fps under VGA resolution

PATENT

- **Y. Fu**, F. An, et al. (co-inventor), CN Patent 112070821A, Low-power-consumption stereo matching system and method for acquiring depth information, 2020.

PUBLICATION

- **Y. Fu**, C. Zhou, T. Huang, S. Qiu, Y. He, and H. Jiao, "SoftAct: High-Precision Softmax Processing Hardware Unit for Transformer Network with Integrated Nonlinear Activation Function Support," IEEE Transactions on Circuits and Systems I: Regular Papers (TCAS-I), 2023, under review.
- C. Zhou, **Y. Fu**, M. Liu, S. Qiu, G. Li, Y. He, and H. Jiao, "An Energy-Efficient 3D Point Cloud Neural Network Accelerator with Efficient Filter Pruning, MLP Fusion, and Dual-Stream Sampling," *IEEE/ACM International Conference On Computer Aided Design (ICCAD)*, 2023, under review.
- C. Zhou, M. Liu, S. Qiu, X. Cao, **Y. Fu**, Y. He, and H. Jiao, "Sagitta: An Energy-Efficient Sparse 3D-CNN Accelerator for Real-Time 3D Understanding," *IEEE Internet of Things Journal (IoT-J)*, 2023, under review (minor revision).
- P. Dong, Z. Chen, Z. Li, **Y. Fu**, L. Chen, F. An, "A 4.29nJ/pixel Stereo-depth Coprocessor with Pixel-level Pipeline and Region-Optimized Semi-Global Matching for IoT Application," IEEE Transactions on Circuits and Systems I: Regular Papers (TCAS-I), 2021.

TAPE OUT

- An energy-efficient pipelined and configurable 3D point cloud-based neural network accelerator is being designed in TSMC 28-nm HPC technology with an area of 2.0 mm×1.5 mm and is expected to be taped out in July 2023

HONORS & AWARDS

- Excellent Graduate Award, *Southern University of Science and Technology*, 2021
- **China National Scholarship**, *Ministry of Education of the PRC*, 2020
- First Class Scholarship, *Southern University of Science and Technology*, 2019
- First Class Scholarship, *Southern University of Science and Technology*, 2018
- Third Class Freshman Scholarship, *Southern University of Science and Technology*, 2017

SKILLS

- English: TOFEL 104
- Proficient in digital integrated circuit (IC) front-end development and logic synthesis, FPGA development and board-level verification, and neural network model compression
- Familiarity with the following tools: Cadence (Genus and NCSim), Vivado; Verilog HDL, PyTorch, Intel Distiller (Model Compression)
- Knowledgeable in Python, JAVA, MATLAB, Shell, Makefile