Yuzhe Fu

No.5, Yiheyuan Road, Haidian District Tel: (+86) 18573907771
Beijing, P.R.China, 10087 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*
Semester Exchange, 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
- Energy-efficient and configurable artificial intelligence accelerator design

PUBLICATIONS

- **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," *plan to submit to TCAS-I in August*, 2023.
- 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.（Accepted)
- 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.

PATENT

- **Y. Fu**, F. An, et al. (co-inventor), CN Patent, Low-power-consumption stereo matching system and method for acquiring depth information, 2020, CN112070821A / WO2022021912A1.

HONORS & AWARDS

- Excellent Graduate Award, *Southern University of Science and Technology*, 2021
- Best Presentation Award, IEEE CASS Shanghai and Shenzhen Joint Workshop, 2021
- **China National Scholarship**, *Ministry of Education of the PRC*, 2020
- First Class Scholarship **(Top 5%)**, *Southern University of Science and Technology*, 2019
- First Class Scholarship **(Top 5%)**, *Southern University of Science and Technology*, 2018

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 and optimized precision in hardware.
- Introduced a stage-wise sparsity detecting method and designed a reconfigurable architecture.
- Implemented in TSMC HPC 28 nm and benchmarked with MobileViT, attaining peer-leading precision and improved hardware efficiency by 7.5×.

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 accelerating unit, ensuring precision while mitigating complexity by 14.22×.
- Co-designed a coarse-to-fine-grained data-reuse dataflow scheme, leveraging spatial-temporal locality, block-wise pipelined delay-aggregation, and MLP fusion.

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.

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× 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.

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 taped out in July 2023.
- A 4.5 TOPS/W sparse 3D-CNN accelerator for real-time 3D understanding was fabricated in UMC 55-nm low-power CMOS technology with an area of 4.2 mm×3.6 mm in August 2020.

SKILLS

- Proficient in digital integrated circuit (IC) front-end development (RTL implementation and logic synthesis), FPGA development, and neural network model compression.
- Familiar tools: Cadence (Genus and NCSim), Vivado; PyTorch, Intel Distiller (Model Compression).
- Knowledgeable languages: Verilog HDL, Python, JAVA, MATLAB, Shell, Makefile.
- For additional information, please visit my website: https://yuzhe-fu.github.io