Zhifan Ye

346-542-6369 | zye327@gatech.edu | google scholar

Education

- **Georgia Institute of Technology** — Atlanta, GA, USA
  Ph.D. Student, Computer Science, Advisor: Prof. Yingyan Lin — *Aug. 2023 – now*
- **Rice University** — Houston, TX, USA
  Ph.D. Student, Electrical and Computer Engineering, Advisor: Prof. Yingyan Lin — *Aug. 2022 – Aug. 2023*
- **University of Science and Technology of China** — Hefei, Anhui, China
  B.E., School of the Gifted Young, Computer Science (Ranking: 3/251) — *Aug. 2018 – July 2022*

Research Interest

- Efficient deep neural network (DNN) training and inference algorithms
- 3D Computer Vision and Computer Graphics
- Efficient DNN accelerators on the edge

Publications

1. Y. Fu, **Z. Ye**, S. Zhang, J. Yuan, Z. Yu, Y. Lin, "S6-DAMON: Bridging Self-Supervised Speech Models and Real-Time Speech Recognition", **under review**, *International Conference on Learning Representations* (**ICLR**), 2024.

2. **Z. Ye**, C. Li, Y. Fu, H. You, S. Li, H. Qu, S. Zhang, Y. Lin, "MarryRecon: Marry Radiance Fields and Meshes Towards Efficient 3D Reconstruction and Rendering", **under review**, *Annual AAAI Conference on Artificial Intelligence* (**AAAI**), 2024.

3. Y. Fu*, Y. Zhang*, Z. Yu*, S. Li, **Z. Ye**, C. Li, C. Wan, Y. Lin, "GPT4AIGChip: Towards Next-Generation AI Accelerator Design Automation via Large Language Models", *International Conference on Computer-Aided Design* (**ICCAD**), 2024. (*: co-first author)

4. Y. Fu*, **Z. Ye***, J. Yuan, S. Zhang, S. Li, H. You, Y. Lin, "Gen-NeRF: Efficient and Generalizable Neural Radiance Fields via Algorithm-Hardware Co-Design", *International Symposium on Computer Architecture* (**ISCA**), 2023. (*: co-first author)

5. Y. Fu, Y. Zhang, K. Qian, **Z. Ye**, Z. Yu, C. Lai, Y. Lin, "Losses Can Be Blessings: Routing Self-Supervised Speech Representations Towards Efficient Multilingual and Multitask Speech Processing", *Advances in Neural Information Processing Systems* (**NeurIPS**), 2022.

6. J. Dass, S. Wu, H. Shi, C. Li, **Z. Ye**, Z. Wang, Y. Lin, "ViTALiTy: Unifying Low-rank and Sparse Approximation for Vision Transformer Acceleration with a Linear Taylor Attention", *IEEE International Symposium on High-Performance Computer Architecture* (**HPCA**), 2023.

7. X. Tian, **Z. Ye**, A. Lu, L. Guo, Y. Chi, Z. Fang, "SASA: A Scalable and Automatic Stencil Acceleration Framework for Optimized Hybrid Spatial and Temporal Parallelism on HBM-based FPGAs, *ACM Transactions on Reconfigurable Technology and Systems* (**TRETS**), 2022.

Experiences

- **Research Scientist Intern, Meta Reality Labs** — Manager: Prof. Yuecheng Li
  Efficient Full-Body Codec Avatar Decoder — *May 2023 – Aug 2023*
- **Research Intern, Rice University** — Advisor: Prof. Yingyan Lin
  Efficient Deep Learning Systems — *Sept 2021 – July 2022*
- **Research Intern, Simon Fraser University** — Advisor: Prof. Zhenman Fang
  Accelerating Big Data Algorithms on Datacenter FPGAs — *Jan 2021 – Aug 2021*
- **Research Intern, University of Science and Technology of China** — Advisor: Prof. Chao Wang
  FPGA-based DNN Accelerators — *Sept 2020 – Jan 2021*

Skills

Programming

- Programming Languages: C/C++, Python, Bash
- ML Frameworks: Pytorch, DGL, NCNN
- FPGA Programming: HLS, Verilog, Vivado, Vitis

Language

- **GRE General Test** Verbal 164 (94%), Quantitative 169 (93%), Analytical Writing 3.5 (37%)
- **TOEFL iBT** Reading 30, Listening 30, Speaking 24, Writing 26

Awards

- HPCA Travel Grant — March. 2023
- Honors Graduate of University of Science and Technology of China (5%) — Jan. 2022
- Huawei Scholarship (2%) — Oct. 2021
- Chinese Academy of Science SINANO Scholarship (2%) — Oct. 2020
- School of the Gifted Young SiHua Scholarship (5%) — Oct. 2019
- Gold Scholarship of University of Science and Technology of China (10%) — Oct. 2019 & Oct. 2020
- First Prize in High School Physics Olympiad in Anhui Province — Oct. 2017

Other Research Projects

- **Hardware Accelerators for 3D Gaussians** Aug 2023 – now @ Georgia Tech
  project leader
  1. Supported both efficient training and real-time rendering for 3D Gaussians.
  2. Targeted at both static scene reconstruction and video sequence reconstruction.
  3. Supported AR (Augmented Reality) applications such as scene composition.

- **Gaussian Splatting for Efficient Codec Avatar** May 2023 – now @ Georgia Tech & Meta
  project leader
  1. Introduced gaussian splatting to codec avatar pipeline.s
  2. Achieved favourable rendering quality and efficiency trade-off for on-device deployment.
  3. Proposed end to end training pipeline for learning anistropic 3D Gaussians of a human avatar.

- **Benchmarking NeRF Algorithms** April 2023 – now @ Georgia Tech
  project leader
  1. Established an unified framework for representative NeRF algorithms.
  2. Extensive benchmarking to understand the inherent bias and performance differences resulting from different neural scene representations.

- **Efficient Distributed Training Algorithms for GNNs** Aug 2022 – now @ Georgia Tech
  project co-leader
  1. Proposed a node-sampling strategy and a novel data-compression technique to reduce data movement.
  2. Supported both GraphSage and GAT.
  3. Cut down communication volume between servers by $> 90\%$, with less than 1% accuracy drop.

Teaching

- **CS 3220: Processor Design** (*Teaching Assistant*) 2023 Fall
  1. Designed coding labs for implementing a RISC-V processor with branch prediction in RTL.
  2. Hosted weekly office hours and take charge of Piazza discussions.