Zhongzhi Yu

Personal Website: https://yuzz1020.github.io/ Email : zyu401@gatech.edu
Google Scholar: https://scholar.google.com/citations?user=KjvcaBQAAAAJ&hl=en Mobile : +1-832-709-3055
Linkedin: https://www.linkedin.com/in/zhongzhi-yu

Education

- **Georgia Institute of Technology** Atlanta, GA
  Ph.D. in Computer Science *Jan. 2023 - Now*
- **Rice University** Houston, TX
  Ph.D. in Electrical and Computer Engineering *Aug 2020 - Dec 2022*
  Transferred to Georgia Institute of Technology with my advisor.
- **Columbia University** New York, NY
  Master of Science in Electrical Engineering *Aug 2017 - May 2019*
- **Zhejiang University** Zhejiang, China
  Bachelor of Engineering in Opto-electronic Information Science and Engineering with honor *Sep 2013 - June 2017*

Experiences

- **NVIDIA Research** Austin, TX
  Advisor: Mark Ren *May 2024 - Aug 2024*
  Research on developing LLM agent systems to facilitate LLMs in generating complex hardware code from unstructured and lengthy instructions.
- **MIT-IBM Watson AI Lab** Cambridge, MA
  Advisor: Yang Zhang *May 2022 - Aug 2022*
  Research on developing modular models to equip existing ASR systems with multilingual scalability and low-resource adaptation ability.

Publications

- **Yu, Zhongzhi**, Zheng Wang, Zhenyang Chen, Chaojian Li, Hyewon Suh, Yonggan Fu, Dachuan Shi, Hongxu Yin, Jan Kautz, Pavlo Molchanov, and Yingyan (Celine) Lin. "*ZoomVLM: A Tuning-Free Framework for Efficient Video Understanding via Adaptive Zooming in Vision-Language Models*". *Under Review*
- **Yu, Zhongzhi**, Zheng Wang, Yonggan Fu, Huihong Shi, Khalid Shaikh, and Yingyan (Celine) Lin. "*Unveiling and Harnessing Hidden Attention Sinks: Enhancing Large Language Models without Training through Attention Calibration*" In International Conference on Machine Learning (ICML 2024)
- **Yu, Zhongzhi**, Zheng Wang, Xiaoya Zhou, Yuhan Li, Ruijie Gao, Sreenidhi Reddy Bommu, Yang (Katie) Zhao, and Yingyan (Celine) Lin. "*EDGE-LLM: Enabling Efficient Large Language Model Adaptation on Edge Devices via Unified Compression and Adaptive Layer Voting.*" In the 2024 61th ACM/IEEE Design Automation Conference (DAC 2024).
- Yongan Zhang, **Zhongzhi Yu**, Yonggan Fu, Cheng Wan, Yingyan (Celine) Lin, "*MG-Verilog: Multi-grained Dataset Towards Enhanced LLM-assisted Verilog Generation.*" In the First IEEE International Workshop on LLM-Aided Design (LAD 2024, Best Paper Award).
- **Yu, Zhongzhi**, Yang Zhang, Kaizhi Qian, Cheng Wan, Yonggan Fu, Yongan Zhang, and Yingyan (Celine) Lin. "*Master-ASR: Achieving Multilingual Scalability and Low-Resource Adaptation in ASR with Modular Learning.*" In International Conference on Machine Learning (ICML 2023).
- **Yu, Zhongzhi**, Shang Wu, Yonggan Fu, Shunyao Zhang, and Yingyan (Celine) Lin. "*Hint-Aug: Drawing Hints from Foundation Vision Transformers Towards Boosted Few-Shot Parameter-Efficient Tuning.*" In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2023).
- Fu, Yonggan*, Yongan Zhang*, **Zhongzhi Yu***, Sixu Li, Zhifan Ye, Chaojian Li, Cheng Wan, and Yingyan (Celine) Lin. "*GPT4AIGChip: Towards Next-Generation AI Accelerator Design Automation via Large Language Models.*" In 2023 IEEE/ACM International Conference On Computer-Aided Design (ICCAD 2023).
- **Yu, Zhongzhi**, Yonggan Fu, Jiayi Yuan, Haoran You, and Yingyan (Celine) Lin. "*NetBooster: Empowering Tiny Deep Learning By Standing on the Shoulders of Deep Giants.*" In the 2023 60th ACM/IEEE Design Automation Conference (DAC 2023).
- **Yu, Zhongzhi**, Yonggan Fu, Sicheng Li, Chaojian Li, and Yingyan Lin. "*MIA-Former: Efficient and Robust Vision Transformers via Multi-Grained Input-Adaptation.*" In Proceedings of the AAAI Conference on Artificial Intelligence (AAAI 2022).

- **Yu, Zhongzhi**, Yonggan Fu, Shang Wu, Mengquan Li, Haoran You, and Yingyan Lin. "*LDP: Learnable Dynamic Precision for Efficient Deep Neural Network Training and Inference.*" TinyML Research Symposium 2022.
- Wang, Zheng, Boxiao Jin, Yuming Chang, **Zhongzhi Yu**, Minjia Zhang, "*Model Tells You Where to Merge: Adaptive KV Cache Merging for LLMs on Long-Context Tasks*", *Under Review*
- You, Haoran*, Cheng Wan*, Yang Zhao*, **Zhongzhi Yu***, Yonggan Fu, Jiayi Yuan, Shang Wu, Shunyao Zhang, Yongan Zhang, Chaojian Li, Vivek Boominathan, Ashok Veeraraghavan, Ziyun Li, and Yingyan (Celine) Lin. "*EyeCoD: Eye Tracking System Acceleration via Flatcam-based Algorithm and Accelerator Co-design.*" In Proceedings of the 49th Annual International Symposium on Computer Architecture (ISCA 2022).
- **Yu, Zhongzhi**, and Yemin Shi. "*Centralized Space Learning for Open-set Computer-aided Diagnosis.*" In Scientific Reports (2023), 13(1), 1630.
- **Yu, Zhongzhi**, and Yemin Shi. "*Kernel Quantization for Efficient Network Compression.*" IEEE Access 10 (2022): 4063-4071.
- Zhang, Yongan, Yonggan Fu, **Zhongzhi Yu**, Kevin Zhao, Cheng Wan, Chaojian Li, and Yingyan (Celine) Lin. "*Invited: Data4AIGChip: An Automated Data Generation and Validation Flow for LLM-assisted Hardware Design.*" In the 2024 61st ACM/IEEE Design Automation Conference (DAC 2024).
- Fu, Yonggan, Zhifan Ye, **Zhongzhi Yu**, and Yingyan (Celine) Lin. "*S6-DAMON: Unlocking Structured Sparsity in Self-Supervised Speech Models via Data-Model Co-Compression.*" *Under Review.*
- Li, Chaojian, **Zhongzhi Yu**, Yonggan Fu, Yongan Zhang, Yang Zhao, Haoran You, Qixuan Yu, Yue Wang, and Yingyan Lin. "*HW-NAS-Bench: Hardware-Aware Neural Architecture Search Benchmark.*" In the 9th International Conference on Learning Representations 2021 (ICLR 2021).
- Fu, Yonggan, Yang Zhang, Kaizhi Qian, Zhifan Ye, **Zhongzhi Yu**, Cheng-I Lai, and Yingyan Lin. "*Losses Can Be Blessings: Routing Self-Supervised Speech Representations Towards Efficient Multilingual and Multitask Speech Processing.*" In Thirty-sixth Conference on Neural Information Processing Systems (NeurIPS 2022).
- You, Haoran, Zhanyi Sun, Huihong Shi, **Zhongzhi Yu**, Yang Zhao, Yongan Zhang, Chaojian Li, Baopu Li, and Yingyan Lin. "*Vitcod: Vision Transformer Acceleration via Dedicated Algorithm and Accelerator Co-design.*" In 2023 IEEE International Symposium on High-Performance Computer Architecture (HPCA 2023). IEEE, 2023.
- Fu, Yonggan, **Zhongzhi Yu**, Yongan Zhang, Yifan Jiang, Chaojian Li, Yongyuan Liang, Mingchao Jiang, Zhangyang Wang, and Yingyan Lin. "*InstantNet: Automated Generation and Deployment of Instantaneously Switchable-Precision Networks.*" In 2021 58th ACM/IEEE Design Automation Conference (DAC 2021), pp. 757-762. IEEE, 2021.
- Fu, Yonggan, Yongan Zhang, Chaojian Li, **Zhongzhi Yu**, and Yingyan Lin. "*A3C-S: Automated Agent Accelerator Co-Search Towards Efficient Deep Reinforcement Learning.*" In 2021 58th ACM/IEEE Design Automation Conference (DAC 2021), pp. 13-18. IEEE, 2021.
- Li, Mengquan, **Zhongzhi Yu**, Yongan Zhang, Yonggan Fu, and Yingyan Lin. "*O-HAS: Optical Hardware Accelerator Search for Boosting Both Acceleration Performance and Development Speed.*" In 2021 IEEE/ACM International Conference On Computer Aided Design (ICCAD 2021), pp. 1-9. IEEE, 2021.
- Fu, Yonggan, **Zhongzhi Yu**, Yongan Zhang, and Yingyan Lin. "*Auto-agent-distiller: Towards efficient deep reinforcement learning agents via neural architecture search.*" arXiv preprint arXiv:2012.13091 (2020).
- Zhao, Guangyuan, Mohammad M. Kabir, Kimani C. Toussaint, Cuifang Kuang, Cheng Zheng, **Zhongzhi Yu**, and Xu Liu. "*Saturated Absorption Competition Microscopy.*" Optica 4, no. 6 (2017): 633-636.
- **Yu, Zhongzhi**, Shaocong Liu, Dazhao Zhu, Cuifang Kuang, and Xu Liu. "*Parallel Detecting Super-resolution Microscopy Using Correlation Based Image Restoration.*" Optics Communications 404 (2017): 139-146.

Awards and Services

- Best paper award in the First IEEE International Workshop on LLM-Aided Design (LAD'24)
- Second place in University Best Demonstration at DAC 2023
- Served as reviewer for NeurIPS, CVPR, ICML, ICLR, ECCV, Transactions on Computers, AAAI, and AICAS