Ziyang Zhang

· ziyang.zhang@polimi.it · https://bigboyzzy.github.io

EDUCATION

Postdoctoral Fellow, Networked Embedded Software Lab Feb. 2025 - Jan. 2026
Politecnico di Milano Milan, Italy
Advisor: Prof. Luca Mottola

Ph.D., Computer Science and Technology Sep. 2020 - Sep. 2024
Harbin Institute of Technology Harbin, China
Advisor: Prof. Jie Liu (IEEE Fellow, ACM Distinguished Scientist)

M.Eng., Electronics and Communications Engineering Sep. 2018 - July. 2020
Nankai University (GPA 4.09/5.0) Tianjin, China
Advisor: Prof. Guiling Sun (Assistant dean of school of electronic information)

B.Eng., Electronic Information Science and Technology Sep. 2014 - July. 2018
Shandong University of Science and Technology (GPA 4.32/5.0) Qingdao, China

RESEARCH INTEREST

My main research interest is **edge computing**, **embedded intelligence**, with a focus on **high-performance and energy-efficent edge DNN inference**.

- **Edge Computing**: optimizing energy-efficient DNN inference via system (e.g., compiler, DVFS)-algorithm (e.g., early exit, reinforcement learning) co-design.
- **Machine Learning System**: designing hybrid scheduling frameworks for multi-tenant edge-cloud environments.

PUBLICATIONS

Conference Papers

[1] E4: Energy-Efficient DNN Inference for Edge Video Analytics Via DVFS and Early Exiting
Ziyang Zhang, Yang Zhao, Ming-Ching Chang, Changyao Lin, Jie Liu
AAAI, 1165-1173, 2025

[2] E3: Early Exiting with Explainable AI for Real-Time and Accurate DNN Inference in Edge-Cloud Systems
Changyao Lin, Zhenming Chen, Ziyang Zhang, Jie Liu
ACM SenSys, 1-13, 2025 (*Acceptance Rate: 46/245, 18.8%*)

[3] POS: An Operator Scheduling Framework for Multi-model Inference on Edge Intelligent Computing
Ziyang Zhang, Huan Li, Yang Zhao, Changyao Lin, Jie Liu
ACM/IEEE IPSN, 40-52, 2023 (*Acceptance Rate: 22/83, 26.5%*)

[4] Octopus: SLO-Aware Progressive Inference Serving via Deep Reinforcement Learning in Multi-Tenant Edge Cluster
Ziyang Zhang, Yang Zhao, Jie Liu
Springer ICSOC, 242-258, 2023 (*Acceptance Rate: 35/208, 16.8%*)

[5] Choosing Appropriate AI-enabled Edge Devices, Not the Costly Ones
Ziyang Zhang, Feng Li, Changyao Lin, Shihui Wen, Xiangyu Liu, Jie Liu
IEEE ICPADS, 201-208, 2021

[6] DVFO: Dynamic Voltage, Frequency and Offloading for Efficient AI on Edge Devices (Poster)
Ziyang Zhang, Yang Zhao, Jie Liu
ACM/IEEE IPSN, 304-305, 2023

[7] E4: Energy-Efficient Early-Exit DNN Inference Framework for Edge Video Analytics (Poster)
Ziyang Zhang, Yang Zhao, Jie Liu
ACM SenSys, 512-513, 2023

[8] ECSRL: A Learning-Based Scheduling Framework for AI Workloads in Heterogeneous Edge-Cloud Systems (Poster)

Changyao Lin, Huan Li, Ziyang Zhang, Jie Liu
ACM SenSys, 386-387, 2021

[9] Exploiting Operator-Level Concurrency Control to Guide Deployment for Real-Time Tasks in Edge AI Cluster (Poster)
Changyao Lin, Ziyang Zhang, Jie Liu
ACM SenSys, 1-2, 2025

Journal Papers

[1] DVFO: Learning-Based DVFS for Energy-Efficient Edge-Cloud Collaborative Inference
Ziyang Zhang, Yang Zhao, Huan Li, Changyao Lin, Jie Liu
IEEE Transactions on Mobile Computing, 9042-9059, 2024

[2] BCEdge: SLO-Aware DNN Inference Services with Adaptive Batch-Concurrent Scheduling on Edge Platforms
Ziyang Zhang, Yang Zhao, Huan Li, Jie Liu
IEEE Transactions on Network and Service Management, 4131-4145, 2024

[3] POS2: Sparse and Operator-Aware Hybrid Scheduling for Edge DNN Inference (Submitted)
Ziyang Zhang, Luca, Mottola, Jie Liu
IEEE Transactions on Mobile Computing, 1-17, 2025

[4] E3A: Energy-efficient Edge Analytics with Early Exit and Frequency Domain Distillation (Submitted)
Ziyang Zhang, Shaowei He, Shusheng Li, Yang Zhao, Jie Liu
IEEE Internet of Things Journal, 1-15, 2025

[5] TOP: Task-Based Operator Parallelism for Asynchronous Deep Learning Inference on GPU
Changyao Lin, Zhenming Chen, Ziyang Zhang, Jie Liu
IEEE Transactions on Parallel and Distributed Systems, 266-281, 2024

[6] Multi-Sensor Data Fusion Algorithm Based on Trust Degree and ImprovedGenetics
Ziyang Zhang, Guiling Sun, Bowen Zheng, Yangyang Li
Sensors, 19(9), 1-18, 2019

Professional Service

- IEEE Transactions on Computers (TC), reviewer
- IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), reviewer
- IEEE Transactions on Parallel and Distributed Systems (TPDS), reviewer
- IEEE Transactions on Network and Service Management (TNSM), reviewer
- IEEE Transactions on Vehicular Technology (TVT), reviewer
- IEEE Internet of Things Journal (IoTJ), reviewer
- Elsevier Internet of Things, reviewer
- ACM SenSys'24, AE Committee
- ACM SenSys/BuildSys Workshop on DATA' 23, TPC member

Skill

- Programming Language: C, C++, Python, CUDA
- Embedded Platforms: Arm-Linux, GPU, FPGA
- Deep Learning Frameworks: TensorFlow, PyTorch, TensorRT, TVM, ONNX
- DevOps Tools: Docker, Kubernetes (k8s)

Selected Awards and Honors

- **The First Prize**, MCM/ICM, 2017
- **National Scholarship**, Highest honor in China, 2019
- **Tencent Scholarship**, Tencent, 2024