

ZIYANG ZHANG

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

EDUCATION

Postdoctoral Fellow, Networked Embedded Software Lab <i>Politecnico di Milano</i> <i>Advisor: Prof. Luca Mottola</i>	Feb. 2025 - Present Milan, Italy
Ph.D., Computer Science and Technology <i>Harbin Institute of Technology</i> <i>Advisor: Prof. Jie Liu (IEEE Fellow, ACM Distinguished Scientist)</i>	Sep. 2020 - Sep. 2024 Harbin, China
M.Eng., Electronics and Communications Engineering <i>Nankai University (GPA 4.09/5.0)</i> <i>Advisor: Prof. Guiling Sun (Assistant dean of school of electronic information)</i>	Sep. 2018 - July. 2020 Tianjin, China
B.Eng., Electronic Information Science and Technology <i>Shandong University of Science and Technology (GPA 4.32/5.0)</i>	Sep. 2014 - July. 2018 Qingdao, China

RESEARCH INTEREST

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

- Edge Computing
- Machine Learning System
- Deep Reinforcement Learning

PUBLICATIONS

Conference Papers

- 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 (CCF-A), 2025
- 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 (CCF-B, 清华 A), 2025 (Acceptance Rate: 46/245, 18.8%)
- 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 (CCF-B, 清华 A), 2023 (Acceptance Rate: 22/83, 26.5%)
- Octopus: SLO-Aware Progressive Inference Serving via Deep Reinforcement Learning in Multi-Tenant Edge Cluster
Ziyang Zhang, Yang Zhao, Jie Liu
Springer ICSOC (CCF-B), 2023 (Acceptance Rate: 35/208, 16.8%)
- Choosing Appropriate AI-enabled Edge Devices, Not the Costly Ones
Ziyang Zhang, Feng Li, Changyao Lin, Shihui Wen, Xiangyu Liu, Jie Liu
IEEE ICPADS (CCF-C), 2021

Journal Papers

- 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 (CCF-A), 9042-9059, 2024
- 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
- 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 (CCF-A), 1-16, 2024

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

Poster Papers

- DVFO: Dynamic Voltage, Frequency and Offloading for Efficient AI on Edge Devices
Ziyang Zhang, Yang Zhao, Jie Liu
ACM/IEEE IPSN (CCF-B, 清华 A), 2023
- E4: Energy-Efficient Early-Exit DNN Inference Framework for Edge Video Analytics
Ziyang Zhang, Yang Zhao, Jie Liu
ACM SenSys (CCF-B, 清华 A), 2023
- ECSRL: A Learning-Based Scheduling Framework for AI Workloads in Heterogeneous Edge-Cloud Systems
Changyao Lin, Huan Li, Ziyang Zhang, Jie Liu
ACM SenSys (CCF-B, 清华 A), 2021

Under Review

- E3A: Energy-efficient Edge Analytics with Early Exit and Frequency Domain Distillation
Shaowei He, Ziyang Zhang, Shusheng Li, Yang Zhao
IEEE Internet of Things Journal, 2025

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

SELECTED AWARDS AND HONORS

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