Minxuan Zhou

Tenure-track Assistant Professor · Computer Science

Illinois Institute of Technology

mzhou26@iit.edu | minxuanz.weebly.com/

Education

University of California San Diego — *La Jolla, CA, USA*

PhD in Computer Science — *2018.9 - 2023.9*

- Thesis: Software-Hardware Co-design for Processing In-Memory Accelerators
- Committee: Tajana Rosing (Advisor), Farinaz Kushanfar, Steven Swanson, Dean Tullsen, Jishen Zhao

University of California San Diego — *La Jolla, CA, USA*

MS in Computer Science — *2015.9 - 2017.6*

- Research: Efficient Temperature Management for 3D-stacked DRAM
- Advisor: Dean Tullsen

Beihang University — *Beijing, China*

BS in Computer Science and Technology — *2011.9 - 2015.6*

- Thesis: Efficient Checkpoint Infrastructure in Micro-kernel Operating System
- Advisor: Yuebin Bai
- Outstanding undergraduate thesis in School of Computer Science and Engineering

Publications

Conferences

Minxuan Zhou, Yujin Nam, Xuan Wang, Youhak Lee, Chris Wilkerson, Raghavan Kumar, Sachin Taneja, Sanu Mathew, Rosario Cammarota, and Tajana Rosing, "UFC: A Unified Accelerator for Fully Homomorphic Encryption", 57th IEEE/ACM International Symposium on Microarchitecture, 2024 (accepted)

Chien-Yi Yang, **Minxuan Zhou**, Flavio Ponzina, Suraj Sathya Prakash, Raid Ayoub, Pietro Mercati, Mahesh Subedar, and Tajana Rosing, "Multi-Objective Software-Hardware Co-Optimization for HD-PIM via Noise-Aware Bayesian Optimization", ACM/IEEE International Conference on Computer-Aided Design, 2024 (accepted)

Eunji Kwon, **Minxuan Zhou**, Weihong Xu, Tajana Rosing and Seokhyeong Kang, "RL-PTQ: RL-based Mixed Precision Quantization for Hybrid Vision Transformers", Design Automation Conference (DAC), 2024

Jaeyoung Kang, You Hak Lee, **Minxuan Zhou**, Weihong Xu and Tajana Rosing, "HygHD: Hyperdimensional Hypergraph Learning", Design, Automation, and Test in Europe (DATE), 2024

Yue Pan, **Minxuan Zhou**, Chonghan Lee, Zheyu Li, Rishika Kushwah, Vijaykrishnan Narayanan, and Tajana Rosing, "PRIMATE: Processing in Memory Acceleration for Dynamic Token-pruning Transformers", 29th Asia and South Pacific Design Automation Conference (ASP-DAC), 2024

Yujin Nam, **Minxuan Zhou**, Saransh Gupta, Gabrielle De Micheli, Rosario Cammarota, Chris Wilkerson, Daniele Micciancio, and Tajana Rosing, "Efficient Machine Learning on Encrypted Data using Hyperdimensional Computing", IEEE/ACM International Symposium on Low Power Electronics and Design (ISLPED), 2023

Minxuan Zhou*, Xuan Wang*, and Tajana Rosing, "OverlaPIM: Overlap Optimization for Processing In-Memory Neural Network Acceleration", Design, Automation and Test in Europe Conference (DATE), 2023

Jaeyoung Kang, **Minxuan Zhou**, Abhinav Bhansali, Weihong Xu, Anthony Thomas and Tajana Rosing, "RelHD: A Lightweight Graph-based Learning with Hyperdimensional Computing", The 40th IEEE International Conference on Computer Design (ICCD), 2022

Minxuan Zhou*, Weihong Xu*, Jaeyoung Kang, and Tajana Rosing, "TransPIM: A Memory-based Acceleration via Software-Hardware Co-Design for Transformers", The 28th IEEE International Symposium on High-Performance Computer Architecture (HPCA), 2022

Yizhou Wei, **Minxuan Zhou**, Sihang Liu, Korakit Seemakhupt, Tajana Rosing and Samira Khan. "PIMProf: An Automated Program Profiler for Processing-in-Memory Offloading Decisions", Design, Automation and Test in Europe Conference

(DATE), 2022

Yeseong Kim, Mohsen Imani, Saransh Gupta, **Minxuan Zhou**, and Tajana Rosing. *Massively Parallel Big Data Classification on a Programmable Processing In-Memory Architecture.*, IEEE/ACM International Conference On Computer Aided Design (ICCAD), 2021

Minxuan Zhou*, Lingxi Wu*, Muzhou Li, Niema Moshiri, Kevin Skadron, and Tajana Rosing, "Ultra Efficient Acceleration for De Novo Genome Assembly via Near-Memory Computing", International Conference on Parallel Architectures and Compilation Techniques (PACT), 2021

Minxuan Zhou, Guoyang Chen, Mohsen Imani, Saransh Gupta, Weifeng Zhang, and Tajana Rosing, "PIM-DL: Boosting DNN Inference on Digital Processing In-Memory Architectures via Data Layout Optimizations", International Conference on Parallel Architectures and Compilation Techniques (PACT), 2021

Minxuan Zhou, Yunhui Guo, Weihong Xu, Bin Li, Kevin Eliceiri, and Tajana Rosing, "MAT: Processing In-Memory Acceleration for Long-Sequence Attention", Design Automation Conference (DAC), 2021

Xiao Liu, **Minxuan Zhou**, Rachata Ausavarungnirun, Sean Eilert, Ameen Akel, Tajana Rosing, Vijaykrishnan Narayanan, Jishen Zhao, "FPRA: A Fine-grained Parallel RRAM Architecture", IEEE/ACM International Symposium on Low Power Electronics and Design (ISLPED), 2021

Minxuan Zhou, Muzhou Li, Mohsen Imani, and Tajana Rosing, "HyGraph: Accelerating Graph Processing with Hybrid Memory-centric Computing", Design, Automation and Test in Europe Conference (DATE), 2021

Minxuan Zhou, Mohsen Imani, Yeseong Kim, Saransh Gupta, and Tajana Rosing, "DPSim: A Full-stack Simulation Infrastructure for Digital Processing In-Memory Architecture", 26th Asia and South Pacific Design Automation Conference (ASP-DAC), 2021

Mohsen Imani, Saikishan Pampana, Saransh Gupta, **Minxuan Zhou**, Yeseong Kim, and Tajana Rosing. *Dual: Acceleration of clustering algorithms using digital-based processing in-memory.*, 53rd Annual IEEE/ACM International Symposium on Microarchitecture (MICRO), 2020

Xiao Liu, **Minxuan Zhou**, Tajana Rosing, and Jishen Zhao. 2019. HR3AM: A Heat Resilient Design for RRAM-based Neuromorphic Computing. ACM/IEEE International Symposium on Low Power Electronics and Design (ISLPED), 2019

Mohsen Imani, Saransh Gupta, Yeseong Kim, **Minxuan Zhou**, and Tajana Rosing. DigitalPIM: Digital-based Processing In-Memory for Big Data Acceleration. ACM Proceedings of the 2019 on Great Lakes Symposium on VLSI

Minxuan Zhou, Mohsen Imani, Saransh Gupta, and Tajana Rosing, "Thermal-Aware Design and Management for Search-based In-Memory Acceleration", Design Automation Conference (DAC), 2019.

Minxuan Zhou, Mohsen Imani, Saransh Gupta, Yeseong Kim, and Tajana Rosing, "GRAM: Graph Processing in a ReRAM-based Computational Memory", 24th Asia and South Pacific Design Automation Conference (ASP-DAC), 2019

Minxuan Zhou, Mohsen Imani, Saransh Gupta, and Tajana Rosing, "GAS: A Heterogeneous Memory Acceleration for Graph Processing", IEEE/ACM International Symposium on Low Power Electronics and Design (ISLPED), 2018.

Journals

Xuan Wang*, **Minxuan Zhou* (co-first author)**,and Tajana Rosing. "Fast-OverlaPIM: A Fast Overlap-driven Mapping Framework for Processing In-Memory Neural Network Acceleration". IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems (TCAD), 2024

Lingxi Wu*, **Minxuan Zhou* (co-first author)**, Weihong Xu, Ashish Venkat, Tajana Rosing, and Kevin Skadron, "Abakus: Accelerating k-mer Counting With Storage Technology", ACM Transactions on Architecture and Code Optimization (TACO), 2024

Minxuan Zhou, Andreas Prodromou, Rui Wang, Hailong Yang, Depei Qian, Dean Tullsen. "Temperature-Aware DRAM Cache Management -Relaxing Thermal Constraints in 3D Systems". IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems (TCAD), 2019

Cheng, Kun, Yuebin Bai, Yongwang Zhao, Yao Ma, Duo Lu, Yuanfeng Peng, and **Minxuan Zhou**. "HV 2 M: A novel approach to boost inter-VM network performance for Xen-based HVMs." Journal of Systems and Software 114 (2016): 54-68.

Under Review / Preprint

Minxuan Zhou, Yujin Nam, Pranav Gangwar, Weihong Xu, Arpan Dutta, Kartikeyan Subramanyam, Chris Wilkerson, Rosario Cammarota, Saransh Gupta, and Tajana Rosing, "FHEmem: A Processing In-Memory Accelerator for Fully Homomorphic Encryption", arXiv:2311.16293, 2023

Weihong Xu, Junwei Chen, Po-Kai Hsu, Jaeyoung Kang, **Minxuan Zhou**, Sumukh Pinge, Shimeng Yu, and Tajana Rosing, "Proxima: Near-storage Acceleration for Graph-based Approximate Nearest Neighbor Search in 3D NAND", arXiv:2312.04257, 2023

Non-public Conferences

Minxuan Zhou, Yujin Nam, Pranav Gangwar, Weihong Xu, Arpan Dutta, Chris Wilkerson, Rosario Cammarota, Saransh Gupta and Tajana Rosing, "FHEmem: A Processing In-Memory Accelerator for Fully Homomorphic Encryption", SRC TECHCON, 2023

Minxuan Zhou, Yujin Nam, Pranav Gangwar, Weihong Xu, Arpan Dutta, Chris Wilkerson, Rosario Cammarota, Saransh Gupta and Tajana Rosing, "HEM: Accelerating Fully Homomorphic Encryption In and Near Memory", DARPA GOMACTech, 2023

Minxuan Zhou, Muzhou Li, Mohsen Imani, and Tajana Rosing, "Accelerating Graph Processing with Hybrid Memory-centric Computing", SRC TECHCON, 2020

Minxuan Zhou, Mohsen Imani, Saransh Gupta, and Tajana Rosing, "Thermal-Aware Design and Management for Search-based In-Memory Acceleration", SRC TECHCON, 2019

Minxuan Zhou, Mohsen Imani, Saransh Gupta, and Tajana Rosing, "GP3: Graph Processing in a Parallel Processing-in-Memory Architecture", SRC TECHCON, 2018

Presentations

Invited Talks

2021.11. *TransPIM: A Processing In-Memory Accelerator for Transformers*. Invited talk: SRC CRISP Annual Review (Student Research Deep Dive), Charlottesville, VA (virtual)

2021.11. *Ultra-efficient De Novo Assembly using Near-data Processing*. Invited talk: SRC CRISP Annual Review (Student Research Deep Dive), Charlottesville, VA (virtual)

PHD Forum

2023.07. *Software-hardware co-design for Processing In-memory Accelerator*. Design Automation Conference PhD Forumn, San Francisco, CA.

Conference Talks

2023.09. *FHEmem: A Processing In-Memory Accelerator for Fully Homomorphic Encryption.* SRC TECHCON, Austin, TX, USA

2022.04. *TransPIM: A Memory-based Acceleration via Software-Hardware Co-Design for Transformers.* The 28th IEEE International Symposium on High-Performance Computer Architecture (HPCA'2022), virtual

2021.12. *MAT: Processing In-Memory Acceleration for Long-Sequence Attention.* Design Automation Conference (DAC), San Francisco, California, USA

2021.09. *PIM-DL: Boosting DNN Inference on Digital Processing In-Memory Architectures via Data Layout Optimizations.* International Conference on Parallel Architectures and Compilation Techniques (PACT), virtual

2021.02. *HyGraph: Accelerating Graph Processing with Hybrid Memory-centric Computing.* Design, Automation and Test in Europe Conference (DATE), virtual

2021.01. *DPSim: A Full-stack Simulation Infrastructure for Digital Processing In-Memory Architecture.* 26th Asia and South Pacific Design Automation Conference (ASP-DAC), virtual

2020.09. *Accelerating Graph Processing with Hybrid Memory-centric Computing.* SRC TECHCON, virtual

2019.09. *Thermal-Aware Design and Management for Search-based In-Memory Acceleration.* SRC TECHCON, Austin, TX, USA

2019.06. *Thermal-Aware Design and Management for Search-based In-Memory Acceleration.* Design Automation Conference (DAC), Las Vegas, NV, USA

2019.01. *GRAM: Graph Processing in a ReRAM-based Computational Memory.* 24th Asia and South Pacific Design Automation Conference (ASP-DAC), Tokyo, Japan

2018.09. *GP3: Graph Processing in a Parallel Processing-in-Memory Architecture.* SRC TECHCON, Austin, TX, USA

2018.07. *GAS: A Heterogeneous Memory Acceleration for Graph Processing.*, International Symposium on Low Power Electronics and Design (ISLPED), Bellevue, Washington, USA

Work Experience

Illinois Institute of Technology *Chicago, IL*

Tenure-track Assistant Professor *2024.8 -*

- Research: Computer architecture, system

UC San Diego *La Jolla, CA*

Postdoctoral Scholar *2023.9 - 2024.8*

Graduate Student Researcher *2018.9 - 2023.9*

- Research: software-hardware co-optimization for processing in-memory architecture for emerging computer applications
- Advisor: Tajana Rosing

Intel Corporation *Santa Clara, CA (virtual)*

Security Hardware Research Intern (full-time) *2022.6 - 2022.9*

Security Hardware Research Intern (part-time) *2022.9 - 2023.7*

- Research: Architecture and compiler optimization for fully-homophobic encryption accelerator
- Mentors: Chris Wilkerson, Rosario Cammarota, Sanu Mathew
- 1 paper publication, 1 paper submission, 2 US Patents, 1 chip tapout

Apple Inc. *Cupertino, CA (virtual)*

Machine Learning Research Intern *2021.6 - 2021.9*

- Research: Compiler optimization for Apple Neural Engine
- Mentor: Cecile Foret

Meta *Menlo Park, CA (virtual)*

PhD Software Engineering Intern *2020.6 - 2020.9*

- Research: Efficient Multi-GPU training of large-scale machine learning models
- Mentor: Yuchen Hao

Alibaba Group US. *Sunnyvale, CA*

Research Intern *2019.6 - 2019.9*

- Research: Compiler-level data layout optimization for processing in-memory accelerators
- Mentor: Weifeng Zhang
- 1 paper publication, 2 US Patents

Participated Proposals and Grants

2023	**2 SEED funds in JUMP2.0-PRISM**, Semiconductor Research Corporation	*$ 200k*
2023	**JUMP2.0-PRISM**, Semiconductor Research Corporation	*$ 50.5M*
2023	**Travel grant for DAC60 PhD Forum**, Association for Computing Machinery	*$ 500*
2022-2023	**DPRIVE subcontract**, DARPA	*$ 12.3M*
2020	**Travel grant for DAC57 Young Fellow**, Association for Computing Machinery	*$ 500*
2019	**Brain-Inspired Hyperdimensional Computing for IoT Applications**, NSF#1911095	*$ 500k*
2019	**SEED fund in JUMP-CRISP**, Semiconductor Research Corporation	*$ 100k*
2018-2021	**GRC IoT Reliability**, Semiconductor Research Corporation	*$ 240k*
2018	**Gift for thermal and power optimization in smartphones**, A major smartphone vendor	*$ 100k*

Teaching Experience

Fall23 **CSE193, Introduction to CS Research**, Research Mentor *UC San Diego*
Winter20 **CSE237A, Introduction to Embedded Computing**, Teaching Assistant *UC San Diego*

Mentoring

2023 - . Peter Wang. Undergraduate.

2023 - . Karen Yan. Undergraduate.

2023 - . Ishika Agrawal, Warren Trinh, Vivian Liu, Shirley Bian. Undergraduate
UCSD CSE-ERSP for addressing the underrepresentation of minority students

2022 - . Aatash Pestonjamasp. Undergraduate

2022 - . Arjun Sampath. Undergraduate. Qualcomm

2022 - . Kartikeyan Subramanyam. Undergraduate. Co-authored 1 publication

2022 - . Junwei Chen. Undergraduate. Co-authored 1 submission

2020 - . Xuan Wang. Undergraduate
UCSD CSE-ERSP for addressing the underrepresentation of minority students
Co-authored 1 publication and 1 submission. UCSD PhD

2023 - . Enzo Han. Master

2021 - 2022. Arpan Dutta. Master. Co-authored 1 publication. NVIDIA

2022. Monil Shah. Master. Samsung Research

2022. Abhinav Bhansali. Master. Co-authored 1 publication. Samsung Semiconductors

2020 - 2021. Muzhou Li. Master. Co-authored 2 publications. LinkedIn

2023 - . Haein Choi. PhD

2023 - . Jangseon Park. PhD

2023 - . Chien-Yi Yang. PhD. Co-authored 1 submission

2023 - . Youhak Lee. PhD. Co-authored 1 publication

2022 - . Yue Pan. PhD. Co-authored 1 publication

2021 - . Yujin Nam. PhD. Co-authored 2 publications and 1 submission

2021 - 2022. Pranav Gangwar. PhD. Co-authored 1 publication

2020 - . Weihong Xu. PhD. Co-authored 3 publications and 1 submission

2019 - 2023. Jaeyoung Kang. PhD. Co-authored 3 publications. Apple

Outreach & Professional Development

Paper Review

IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems (TCAD)
IEEE Transactions on Computers (TC)
Applied Soft Computing Journal (ASOC)
MDPI Sensors
MDPI Electronics
MDPI Applied Sciences