

# Shaoyi Huang (she / her / hers)

---

## CONTACT INFORMATION

Web: <https://www.shaoyihuang.com/>  
E-mail: [shaoyi.huang@uconn.edu](mailto:shaoyi.huang@uconn.edu)  
Mobile: 484-747-0300

Google Scholar: [\[Link\]](#)  
Updated: May 21th, 2024

## RESEARCH INTERESTS

My research agenda is anchored in advancing **AI systems**, specifically focused on the development of energy-efficient, sustainable and privacy preserving AI systems. I am actively engaged in **sparsity & parallelism exploitation sparse training, emerging deep learning models inference acceleration, efficient privacy preserving machine learning**, and **machine learning in EDA**. My research goal is to bridge the research gap between algorithms and practical applications, especially the deployments on edge devices such as GPUs, FPGAs, ReRAM, superconducting technologies, and other emerging devices, as well as to improve the energy efficiency of electronic design automation (EDA) and chip design, reinforcing the sustainable development of computing infrastructure.

## EDUCATION

**University of Connecticut**  
Ph.D., Computer Engineering (exp.) July 2024  
Advisor: [Caiwen Ding](#)  
Co-major-advisor: [Omer Khan](#)

**University of Rochester**  
M.S.E., Electrical and Computer Engineering May 2018  
Advisor: [Gonzalo Mateos](#)

**Wuhan University of Technology**  
B.S., Electrical and Information Engineering May 2015

## EDUCATION OUTREACH

- The 6th Workshop for Women in Hardware and Systems Security (WISE 2023), student panelist (shared experiences with female researchers and K-12 students)
- NSF REU Site proposal (Topic: Trustable Embedded Systems Security), mentored 3 undergraduates
- USDA funded Agriculture and Food Research Initiative Award Project, mentored 1 undergraduate

## PUBLICATION SUMMARY

**Published: 20** papers, **11** first/co-first authored papers (**leader**) or the second author (**main contributor**).  
**Impact: 389** citations, h-index: 12, i10-index: 13 (as of May 21th, 2024).  
Published at HPCA, ASPLOS, SC, DAC, ICCAD, ACL, ICCV, NeurIPS, IJCAI, ICCD, ISQED, etc.

### • Sparsity & Parallelism Exploitation Sparse Training

— **[DAC 2023]** [Shaoyi Huang](#), Bowen Lei, Dongkuan Xu, et al, “Dynamic Sparse Training via Balancing the Exploration-Exploitation Trade-off”, Design Automation Conference, 2023

— **[DAC 2023]** [Shaoyi Huang](#), Haowen Fang, et al, “Neurogenesis Dynamics-inspired Spiking Neural Network Training Acceleration”, Design Automation Conference, 2023

— **[HPCA 2024]** Deniz Gurevin, Mohsin Shan, [Shaoyi Huang](#), et al, "PruneGNN: An Optimized Algorithm-Hardware Framework for Graph Neural Network Pruning", IEEE International Symposium on High-Performance Computer Architecture, 2024

- **Emerging Deep Learning Models Inference Acceleration**

- [ACL 2022] Shaoyi Huang, Dongkuan Xu, Ian En-Hsu Yen, et al, "Sparse Progressive Distillation: Resolving Overfitting under Pretrain-and-Finetune Paradigm", Annual Meeting of the Association for Computational Linguistics, 2022
- [SC 2021] Shaoyi Huang\*, Shiyang Chen\*, et al, "E.T.: Re-Thinking Self-Attention for Transformer Models on GPUs", International Conference for High Performance Computing, Networking, Storage and Analysis, 2021
- [DAC 2022, Publicity Paper] Shaoyi Huang\*, Hongwu Peng\*, et al, "A Length Adaptive Algorithm-Hardware Co-design of Transformer on FPGA Through Sparse Attention and Dynamic Pipelining", Design Automation Conference, 2022

- **Privacy Preserving Machine Learning**

- [ICCV 2023] Shaoyi Huang\*, Hongwu Peng\*, Tong Zhou\*, et al, "AutoReP: Automatic ReLU Replacement for Fast Private Network Inference", International Conference on Computer Vision, 2023

PARTICIPATING GRANTS	<b>Semiconductor Research Corporation Research Grant</b>	01/2023 - 12/2025
	<i>Co-writer and primary student representative</i> for Caiwen Ding's group	
	"Exploring Extreme Sparsity in Training and Inference for Graph Neural Networks to Achieve High Performance Scaling on Large Core Count Machines" - \$225,000	
	PI: Caiwen Ding, Omer Khan	
	<b>National Science Foundation (NSF)</b>	07/2023 - 06/2027
	<i>Primary student representative</i> for Caiwen Ding's group	
	"Collaborative Research: SaTC: CORE: Medium: Accelerating Privacy-Preserving Machine Learning as a Service: From Algorithm to Hardware" - \$399,775	
	PI: Caiwen Ding	
	<b>USDA-NIFA Agriculture and Food Research Initiative</b>	01/2022 - 12/2025
	<i>Primary Student representative</i> for Caiwen Ding's group	
	"Evaluating the Impact of Preferential Trade Agreements on Agricultural and Food Trade: New Insights from Natural Language Processing and Machine Learning" - \$650,000	
	PI: Sandro Steinbach, Caiwen Ding, Dongjin Song, Jeremy Jelliffe	
	<b>Eversource Energy Center Seed Grant</b>	09/2021 - 08/2023
	<i>Primary student representative</i> for Caiwen Ding's group	
	"Optigrid: Planning & Optimizing the Power Grid During the Low Carbon Transition in Connecticut" - \$69,000	
	PI: Caiwen Ding, Mikhail A Bragin	
	<b>Travelers Insurance Research Grant</b>	09/2021 - 02/2022
	<i>Primary student representative</i> for Caiwen Ding's group	
	"Change and Damage Detection from Aerial Images" - \$292,406	
	PI: Caiwen Ding, Jinbo Bi, Dongjin Song	

- PEER-REVIEWED PAPERS
- [1] Deniz Gurevin, Shaoyi Huang, Mohsin Shan, MD Amit Hasan, Caiwen Ding, Omer Khan, "PruneGNN: An Optimized Algorithm-Hardware Framework for Graph Neural Network Pruning", IEEE International Symposium on High-Performance Computer Architecture (**HPCA 2024**)  
Acceptance rate: 75/410=18.3%
  - [2] Shaoyi Huang\*, Hongwu Peng\*, Tong Zhou\*, Yukui Luo, Chenghong Wang, Zigeng Wang, Jiahui Zhao, Xi Xie, Ang Li, Tony Geng, Kaleel Mahmood, Wujie Wen, Xiaolin Xu, Caiwen Ding, "AutoReP: Automatic ReLU Replacement for Fast Private Network Inference", International Conference on Computer Vision (**ICCV 2023**)
  - [3] Shaoyi Huang, Bowen Lei, Dongkuan Xu, Hongwu Peng, Yue Sun, Mimi Xie, Caiwen Ding, "Dynamic Sparse Training via Balancing the Exploration-Exploitation Trade-off", Design Automation Conference (**DAC 2023**)  
Acceptance rate: 263/1156=22.7%
  - [4] Shaoyi Huang, Haowen Fang, Kaleel Mahmood, Bowen Lei, Nuo Xu, Bin Lei, Yue Sun, Dongkuan Xu, Wujie Wen and Caiwen Ding, "Neurogenesis Dynamics-inspired Spiking Neural Network Training Acceleration", Design Automation Conference (**DAC 2023**)  
Acceptance rate: 263/1156=22.7%
  - [5] Shaoyi Huang, Dongkuan Xu, Ian En-Hsu Yen, Yijue Wang, Sung-En Chang, Bingbing Li, Shiyang Chen, Mimi Xie, Sanguthevar Rajasekaran, Hang Liu, Caiwen Ding, "Sparse Progressive Distillation: Resolving Overfitting under Pretrain-and-Finetune Paradigm", Annual Meeting of the Association for Computational Linguistics (**ACL 2022**)  
Acceptance rate: 714/3350=21.3%
  - [6] Shaoyi Huang\*, Hongwu Peng\*, Shiyang Chen, Bingbing Li, Tong Geng, Ang Li, Weiwen Jiang, Wujie Wen, Jinbo Bi, Hang Liu and Caiwen Ding, "A Length Adaptive Algorithm-Hardware Co-design of Transformer on FPGA Through Sparse Attention and Dynamic Pipelining", Design Automation Conference (**DAC 2022, Publicity Paper**)
  - [7] Shaoyi Huang, Ning Liu, Yueying Liang, Hongwu Peng, Hongjia Li, Dongkuan Xu, Mimi Xie, Caiwen Ding, "An automatic and efficient bert pruning for edge ai systems", International Symposium on Quality Electronic Design (**ISQED, 2022**)
  - [8] Shaoyi Huang\*, Shiyang Chen\*, Santosh Pandey, Bingbing Li, Guang Gao, Long Zheng, Caiwen Ding, Hang Liu, "E.T.: Re-Thinking Self-Attention for Transformer Models on GPUs", International Conference for High Performance Computing, Networking, Storage and Analysis (**SC 2021**)  
Acceptance rate: 86/365=23.6%
  - [9] Shaoyi Huang, Shiyang Chen, Hongwu Peng, Daniel Manu, Zhenglun Kong, Geng Yuan, Lei Yang, Shusen Wang, Hang Liu, Caiwen Ding, "HMC-Tran: A Tensor-core Inspired Hierarchical Model Compression for Transformer-based DNNs on GPU", Great Lakes Symposium on VLSI (**GLSVLSI 2021**)
  - [10] Hongwu Peng, Shaoyi Huang, Tong Geng, Ang Li, Weiwen Jiang, Hang Liu, Shusen Wang, Caiwen Ding, "Accelerating transformer-based deep learning models on fpgas using column balanced block pruning", International Symposium on Quality Electronic Design (**ISQED 2021**)
  - [11] Daniel Manu, Shaoyi Huang, Caiwen Ding, Lei Yang, "Co-Exploration of Graph Neural Network and Network-on-Chip Design Using AutoML", Great Lakes Symposium on VLSI (**GLSVLSI 2021**)

- [12] Hongwu Peng, Xi Xie, Kaustubh Shivdikar, MD Amit Hasan, Jiahui Zhao, Shaoyi Huang, Omer Khan, David Kaeli, Caiwen Ding, "MaxK-GNN: Towards Theoretical Speed Limits for Accelerating Graph Neural Networks Training", Proceedings of the Twenty-Second International Conference on Architectural Support for Programming Languages and Operating Systems (**ASPLOS 2024**)
- [13] Hongwu Peng\*, Ran Ran\*, Yukui Luo, Jiahui Zhao, Shaoyi Huang, Kiran Thorat, Tong Geng Chenghong Wang, Xiaolin Xu, Wujie Wen, Caiwen Ding, "LinGCN: Structural Linearized Graph Convolutional Network for Homomorphically Encrypted Inference", Advances in Neural Information Processing Systems (**NeurIPS 2023**)
- [14] Xi Xie\*, Hongwu Peng\*, MD Amit Hasan, Shaoyi Huang, Jiahui Zhao, Haowen Fang, Wei Zhang, Tong Geng, Omer Khan, Caiwen Ding, "Accel-GCN: High-Performance GPU Accelerator Design for Graph Convolution Networks", IEEE/ACM International Conference On Computer Aided Design (**ICCAD 2023**)
- [15] Bingbing Li, Zigeng Wang, Shaoyi Huang, Mikhail Bragin, Ji Li, Caiwen Ding, "Towards Lossless Head Pruning through Automatic Peer Distillation for Large Language Models", International Joint Conference on Artificial Intelligence (**IJCAI 2023**)
- [16] Yijue Wang, Nuo Xu, Shaoyi Huang, Kaleel Mahmood, Dan Guo, Caiwen Ding, Wujie Wen, Sanguthevar Rajasekaran, "Analyzing and Defending against Membership Inference Attacks in Natural Language Processing Classification", IEEE International Conference on Big Data (**IEEE Big Data 2022**)
- [17] Yixuan Luo\*, Payman Behnam\*, Kiran Thorat, Zhuo Liu, Hongwu Peng, Shaoyi Huang, Shu Zhou, Omer Khan, Alexey Tumanov, Caiwen Ding, Tong Geng, "CoDG-ReRAM: An Algorithm-Hardware Co-design to Accelerate Semi-Structured GNNs on ReRAM", IEEE International Conference on Computer Design (**ICCD 2022**)
- [18] Hongwu Peng\*, Deniz Gurevin\*, Shaoyi Huang, Tong Geng, Weiwen Jiang, Omer Khan, Caiwen Ding, "Towards Sparsification of Graph Neural Networks", IEEE International Conference on Computer Design (**ICCD 2022**)
- [19] Panjie Qi, Edwin Hsing-Mean Sha, Qingfeng Zhuge, Hongwu Peng, Shaoyi Huang, Zhenglun Kong, Yuhong Song, Bingbing Li, "Accelerating Framework of Transformer by Hardware Design and Model Compression Co-Optimization", IEEE/ACM International Conference On Computer Aided Design (**ICCAD 2021**)
- [20] Panjie Qi, Yuhong Song, Hongwu Peng, Shaoyi Huang, Qingfeng Zhuge, Edwin Hsing-Mean Sha, "Accommodating transformer onto fpga: Coupling the balanced model compression and fpga-implementation optimization", Great Lakes Symposium on VLSI (**GLSVLSI 2021**)

## TALKS

### Invited Talks

- **Student panelist** - The 6th Workshop for Women in Hardware and Systems Security  
Oct 2023, California State University, Fullerton, Fullerton, CA
- **Towards Efficient Model Inference and Training**  
Sep 2023, University of Rochester, Rochester, NY
- **Towards Efficient Training and Inference Under Pretrain-and-Finetune Paradigm**  
Sep 2023, TechCon - Semiconductor Research Corporation (SRC), Austin, TX

- **Exploring Extreme Sparsity in Training and Inference for Graph Neural Networks to Achieve High Performance Scaling on Large Core Count Machines**  
May 2023, Semiconductor Research Corporation (SRC) AIHW & CADT Annual Review, IBM Research, San Jose, CA
- **Efficient Model Inference and Training**  
Apr 2023, Machine Learning and Natural Language Processing Community (MLNLP), Virtual

#### Conference Presentations

- **Dynamic Sparse Training via Balancing the Exploration-Exploitation Trade-off**  
DAC, July 2023, San Francisco, CA
- **Neurogenesis Dynamics-inspired Spiking Neural Network Training Acceleration**  
DAC, July 2023, San Francisco, CA
- **Sparse Progressive Distillation: Resolving Overfitting under Pretrain-and-Finetune Paradigm**  
ACL, Oct 2022

#### PROFESSIONAL Panelist

##### SERVICE

- The 6th Workshop for Women in Hardware and Systems Security (WISE)

#### Program Committee

- Annual Meeting of the Association for Computational Linguistics (ACL) 2024
- SIAM International Conference on Data Mining (SDM) 2024
- AAAI Conference on Artificial Intelligence (AAAI) 2024, 2023
- NeurIPS Datasets and Benchmarks 2023
- SIGKDD Conference on Knowledge Discovery and Data Mining (KDD) 2023
- The First Workshop on DL-Hardware Co-Design for AI Acceleration @AAAI 2023

#### Journal Reviewer

- Neurocomputing
- Pattern Recognition
- Neural Networks
- Journal of Systems Architecture
- Engineering Applications of Artificial Intelligence
- IEEE Transactions on Neural Networks and Learning Systems
- ACM Transactions on Design Automation of Electronic Systems

#### HONORS AND AWARDS

- Machine Learning and Systems Rising Stars (ML Commons) 2024
- Marion and Frederick Buckman Engineering Fellowship 2024
- NeurIPS Travel Grant from CACC 2023
- WISE Student Travel Award 2023
- GE Fellowship of Excellence 2023
- Predoctoral Prize for Research Excellence 2023
- DAC Publicity Paper Award 2022
- GE Fellowship for Excellence 2022
- Synchrony Fellowship 2022
- Predoctoral Prize for Research Excellence 2022
- Eversource Energy Graduate Fellowship 2022
- DAC Young Fellow 2021
- Cigna Graduate Fellowship 2021

PROFESSIONAL POSITIONS	<b>ByteDance</b> , Austin, TX			Summer 2022
	<ul style="list-style-type: none"><li>• Research Intern</li><li>• Project: Transformer-based Model Training Acceleration</li></ul>			
	<b>Moffett AI</b> , Los Altos, CA			Summer 2021
	<ul style="list-style-type: none"><li>• Research Intern</li><li>• Mentor: Ian En-Hsu Yen, Co-founder</li><li>• Project: Language Model Compression [ACL 2022]</li></ul>			
TEACHING EXPERIENCE	Guest Lecturer (University of Rochester)			
	<ul style="list-style-type: none"><li>• ECE 403-1: Advanced Computer Architecture for Machine Learning, Fall 2023</li></ul> Topic: Towards Efficient Model Inference and Training			
	Teaching Assistant (University of Connecticut)			
	<ul style="list-style-type: none"><li>• CSE 4502 &amp; 5717: BigData Analytics, Spring 2023</li><li>• CSE5819: Introduction to Machine Learning, Fall 2022</li></ul>			
	Teaching Assistant (University of Rochester)			
	<ul style="list-style-type: none"><li>• Microcontroller, Spring 2018</li><li>• Circuits &amp; Signals LAB, Fall 2017</li></ul>			
REFERENCES	<b>Dr. Caiwen Ding</b> Assistant Professor University of Connecticut caiwen.ding@uconn.edu			
	<b>Dr. Marcus Pan</b> Program Manager SRC marcus.pan@src.org		<b>Dr. Yanzhi Wang</b> Associate Professor Northeastern University yanz.wang@northeastern.edu	
	<b>Dr. Omer Khan</b> Professor University of Connecticut omer.khan@uconn.edu		<b>Dr. Ian En-Hsu Yen</b> Co-founder & Chief Scientist Moffett AI ian.yan@moffett.ai	
	<b>Dr. Dongkuan Xu</b> Assistant Professor NC state University dxu27@ncsu.edu		<b>Dr. Wujie Wen</b> Associate Professor NC State University wwen2@ncsu.edu	
	<b>Dr. Tony Geng</b> Assistant Professor University of Rochester tong.geng@rochester.edu		<b>Dr. Suining He</b> Assistant Professor University of Connecticut suining.he@uconn.edu	