GENG YUAN

Assistant Professor, School of Computing, University of Georgia, Athens, Georgia, USA
geng.yuan@uga.edu, Google Scholar, ORCID: 0000-0001-9844-992X

EDUCATION

Ph.D. Candidate in Computer Engineering *08/2017 - 08/2023*
Northeastern University, Boston, MA,
Department of Electrical and Computer Engineering

M.S. in Electrical & Computer Engineering *08/2014 - 05/2016*
Syracuse University, Syracuse, NY,
Department of Electrical Engineering and Computer Science

B.E. in Electric Information Engineering *09/2010 - 06/2014*
Beijing University of Technology, China
College of Electric Information and Control Engineering

PROFESSIONAL EXPERIENCE

School of Computing, University of Georgia *08/2023 - present*
Assistant Professor (Tenure-Track)

Northeastern University *08/2017 - 08/2023*
Graduate Research Assistant

Snap Inc. *05/2022 - 08/2022*
Research Internship

Mellanox Technologies, Inc. | **NVIDIA** *07/2019 - 12/2019*
Engineer Internship

AREA OF INTEREST

- **Energy-efficient deep learning and artificial intelligence systems design**
- **AI for data analytics and advanced scientific research**
- **Hardware-software co-design for DNN architectures and accelerators**
- **Deep learning systems with emerging technologies**

AWARDS

09/2025 ACM/IEEE International Conference on Computer-Aided Design (ICCAD 2025):
- **Best Paper Nomination**

01/2023 The 11th International Conference on Learning Representations (ICLR 2023):
- **Spotlight Paper Award**

12/2021 The 35th Conference on Neural Information Processing Systems (NeurIPS 2021):
- **Spotlight Paper Award**

05/2021 Hardware Aware Efficient Training workshop in ICLR (HAET 2021):
- **Best Paper Award**

03/2021 The 24th Design, Automation and Test in Europe Conference (DATE, 2021):
➢ **Best Paper Nomination**

08/2020 International Symposium on Low Power Electronics and Design (ISLPED 2020) - Design Contest:
➢ **Design Contest 1st Place Winner**

06/2019 Design Automation Conference - System Design Contest (DAC-SDC 2019):
➢ **Special Service Recognition Award**

03/2018 The 19th International Symposium on Quality Electronic Design (ISQED 2018):
➢ **Best Paper Nomination**

INSTRUCTION AND MENTORSHIP

➢ **Instructor:**

Advanced Topics in Efficient Deep Learning (CSCI 8000)
Computer Architecture and Organization (CSCI 4720)
Special Topics in Deep Learning (CSCI 4900/6900)

➢ **Advisor:**

PhD - Chence Yang ***(08/2025 - present)***
PhD - Qingchan Zhu ***(08/2025 - present)***
PhD - Ningxi Cheng ***(08/2025 - present)***
PhD - Qitao Tan ***(08/2024 - present)***
PhD - Ci Zhang ***(08/2024 - present)***
MS project - Ayush Ummadi ***(08/2024 - present)***
MS project - Murali Krishna Dalavai Kumar ***(01/2024 - present)***
Research internship - Rui Xia (MS at University of Pennsylvania) ***(08/2024 - present)***
Research internship - Huidong Ji (PhD at Fudan University) ***(08/2024 - present)***
Research internship - John T Settles* (MS from UGA) ***(01/2024 - 08/2024)***
Research internship - Thong Minh Nguyen* (BS from Eastern Oregon University) ***(01/2024 - 05/2024)***

**** indicates the student decided to pursue PhD after the research internship.***

PROFESSIONAL ACTIVITIES

➢ **Guest Editor:**

2024 MDPI Symmetry - Symmetry and Asymmetry in Embedded Systems
2023 MDPI Electronics - Scalable Deep Learning System: Principle and Practice

➢ **Session Chair:**

[25'GLSVLSI] Great Lakes Symposium on VLSI.
[24'ICCAD] ACM/IEEE International Conference on Computer-Aided Design.

➢ **Proposal Review:**
2024 DoE - Advancements in Artificial Intelligence for Science.

➢ **Technical Program Committee:**
[25'ISQED] International Symposium on Quality Electronic Design.
[24'GLSVLSI] Great Lakes Symposium on VLSI.
[24'ISLPED] ACM/IEEE International Symposium on Low Power Electronics and Design.
[24'ISQED] The 25th International Symposium on Quality Electronic Design.

➢ **Reviewer:**
[25'CVPR] IEEE/CVF Conference on Computer Vision and Pattern Recognition.
[25'ICLR] The International Conference on Learning Representations.
[25'AAAI] The 39th Annual AAAI Conference on Artificial Intelligence.

[24'NeurIPS] Annual Conference on Neural Information Processing Systems.
[24'ICCP] International Conference on Parallel Processing.
[24'NEWCAS] 22nd IEEE International NEWCAS Conference.
[24'ICML] International Conference on Machine Learning.
[24'CVPR] Conference on Computer Vision and Pattern Recognition.
[23'TOSN] ACM Transactions on Sensor Networks.
[23'NeurIPS] Conference on Neural Information Processing Systems.
[23'TCAS-I] IEEE Transactions on Circuits and Systems I.
[23'MWCAS] IEEE International Midwest Symposium on Circuits and Systems.
[23'JOE] The Journal of Engineering.
[23'CCOS] Connection Science.
[23'ICCV] IEEE/CVF International Conference on Computer Vision.
[23'ISVLSI] IEEE Computer Society Annual Symposium on VLSI.
[23'AICAS] Artificial Intelligence Circuits and Systems.
[23'NEWCAS] 21st International NEWCAS conference.
[23'IJCAI] The International Joint Conference on Artificial Intelligence.
[23'ICML] The International Conference on Machine Learning.
[23'ICLR-SNN] ICLR 2023 Workshop SNN.
[23'TNNLS] IEEE Transactions on Neural Networks and Learning Systems.
[23'NEUCOM] NeuroComputing Journal.
[23'CVPR] IEEE/CVF Computer Vision and Pattern Recognition Conference.
[22'JOE] The Journal of Engineering.
[22'TPAMI] IEEE Transactions on Pattern Analysis and Machine Intelligence.
[22'VLSIJ] Integration, the VLSI Journal.
[22'JETCAS] IEEE Journal on Emerging and Selected Topics in Circuits and Systems.
[22'ICCD] IEEE International Conference on Computer Design.
[22'NeurIPS] Conference on Neural Information Processing Systems.
[22'ECCV] European Conference on Computer Vision.
[22'TCAS-II] IEEE Transactions on Circuits and Systems II.
[22'ISVLSI] IEEE Computer Society Annual Symposium on VLSI.
[22'T-CAD] IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems.
[22'NEWCAS] 20th International NEWCAS conference.
[22'TNNLS] IEEE Transactions on Neural Networks and Learning Systems.
[22'AICAS] Artificial Intelligence Circuits and Systems.
[22'GLSVLSI] Great Lakes Symposium on VLSI.
[22'ICML] The Thirty-ninth International Conference on Machine Learning.
[22'CVPR] IEEE/CVF Computer Vision and Pattern Recognition Conference.
[22'NEUCOM] NeuroComputing Journal.
[21'TECS] ACM Transactions on Embedded Computing Systems.
[21'ASAP] IEEE International Conference on Application-Specific Systems, Architectures and Processors.
[21'NEWCAS] 19th International NEWCAS conference.
[20'ISVLSI] IEEE Computer Society Annual Symposium on VLSI.
[20'NEWCAS] 18th International NEWCAS conference.
[20'FCCM] 28th IEEE International Symposium on Field-Programmable Custom Computing Machines.
[19'FPT] International Conference on Field-Programmable Technology.

FULL PUBLICATION LIST (* EQUAL CONTRIBUTION)

1. **[25'ICCAD] [Best Paper Nomination] Qitao Tan, Sung-En Chang, Rui Xia, Huidong Ji, Chence Yang, Ci Zhang, Jun Liu, Zheng Zhan, Zhenman Fang, Zhuo Zou, Yanzhi Wang, Jin Lu, Geng Yuan, "Perturbation-efficient zeroth-order optimization for hardware-friendly on-device training", in Proceedings of 2025**

ACM/IEEE International Conference on Computer-Aided Design (ICCAD).

2. **[25'ASP-DAC] Huidong Ji, Sheng Li, Yue Cao, Chen Ding, Jiawei Xu, Qitao Tan, Jun Liu, Ao Li, Xulong Tang, Lirong Zheng, Geng Yuan, Zhuo Zou, "A computation and energy efficient hardware architecture for ssl acceleration", in Proceedings of the 30th Asia and South Pacific Design Automation Conference.**

3. **[25'ICS] Zhengang Li, Hongwu Peng, Xuan Shen, Masoud Zabihi, Xi Xie, Geng Yuan, Yanzhi Wang, Olivia Chen, Caiwen Ding, "Graph Convolutional Network Acceleration Using Adiabatic Superconductor Josephson Devices", in Proceedings of the 39th ACM International Conference on Supercomputing.**

4. **[25'GLSVLSI] Ci Zhang, Chence Yang, Qitao Tan, Jun Liu, Ao Li, Yanzhi Wang, Jin Lu, Jinhui Wang, Geng Yuan, "Towards memory-efficient and sustainable machine unlearning on edge using zeroth-order optimizer", in Proceedings of the 2025 Great Lakes Symposium on VLSI.**

5. **[25'HIR Journal] Huayu Li, Zhengxiao He, Xiwen Chen, Ci Zhang, Stuart F Quan, William DS Killgore, Shu-Fen Wung, Chen X Chen, Geng Yuan, Jin Lu, Ao Li, "Smarter Together: Combining Large Language Models and Small Models for Physiological Signals Visual Inspection", in Proceedings of Journal of Healthcare Informatics Research.**

6. **[25'IJCAI] Changdi Yang, Zheng Zhan, Ci Zhang, Yifan Gong, Yize Li, Zichong Meng, Jun Liu, Xuan Shen, Hao Tang, Geng Yuan, Pu Zhao, Xue Lin, Yanzhi Wang, "FairSMOE: Mitigating Multi-Attribute Fairness Problem with Sparse Mixture-of-Experts", in Proceedings of 2025 International Joint Conferences on Artificial Intelligence.**

7. **[25'ICLR] Sheng Li, Qitao Tan, Yue Dai, Zhenglun Kong, Tianyu Wang, Jun Liu, Ao Li, Ninghao Liu, Yufei Ding, Xulong Tang, Geng Yuan, "Mutual Effort for Efficiency: A Similarity-based Token Pruning for Vision Transformers in Self-Supervised Learning", in Proceedings of International Conference on Learning Representations.**

8. **[25'AAAI] Jun Liu, Zhenglun Kong, Pu Zhao, Changdi Yang, Hao Tang, Geng Yuan, Wei Niu, Zhang Wenbin, Xue Lin, Dong Huang, Yanzhi Wang, "Toward Adaptive Large Language Models Structured Pruning via Hybrid-grained Weight Importance Assessment", in Proceedings of Annual AAAI Conference on Artificial Intelligence.**

9. **[25'ICASSP] Jun Liu, Zhenglun Kong, Peiyan Dong, Xuan Shen, Pu Zhao, Hao Tang, Geng Yuan, Wei Niu, and *et.al.*, "RoRA: Efficient Fine-Tuning of LLM with Reliability Optimization for Rank Adaptation", in Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing.**

10. **[24'TACO] Danlin Jia, Geng Yuan, Yiming Xie, Xue Lin, Ningfang Mi, "A Data-Loader Tunable Knob to Shorten GPU Idleness for Distributed Deep Learning", in Proceedings of ACM Transactions on Architecture and Code Optimization.**

11. **[24'ISNCC] Yiming Xie, Pinrui Yu, Geng Yuan, Xue Lin, Ningfang Mi, "Adaptive Homogeneity-Based Client Selection Policy for Federated Learning", in Proceedings of International Symposium on Networks, Computers and Communications.**

12. **[24'ICLR] Sheng Li, Chao Wu, Ao Li, Yanzhi Wang, Xulong Tang, Geng Yuan, "Waxing-and-Waning: a Generic Similarity-based Framework for Efficient Self-Supervised Learning", in Proceedings of the 12th International Conference on Learning Representations.**

13. **[24'JBHI Journal] Huayu Li, Ana S Carreon-Rascon, Xiwen Chen, Geng Yuan, Ao Li, "MTS-LOF: Medical Time-Series Representation Learning via Occlusion-Invariant Features", in Proceedings of IEEE Journal of Biomedical and Health Informatics.**

14. **[24'Healthcare Journal] Ao Li, Huayu Li, Geng Yuan, "Continual Learning with Deep Neural Networks in Physiological Signal Data: A Survey", in Proceedings of MDPI Healthcare.**

15. **[24'CPAL] Haoyu Ma, Chengming Zhang, Xiaolong Ma, Geng Yuan, and *et.al.*, "HRBP: Hardware-friendly Regrouping towards Block-based Pruning for Sparse CNN Training", in Proceedings of Conference on Parsimony and Learning.**

16. **[24'TCAD Journal] Jun Liu, Zhenglun Kong, Pu Zhao, Weihao Zeng, Hao Tang, Xuan Shen, Changdi Yang, Wenbin Zhang, Geng Yuan, Wei Niu, Xue Lin, Yanzhi Wang, "TSLA: A Task-Specific Learning Adaptation for Semantic Segmentation on Autonomous Vehicles Platform", in Proceedings of IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems.**

17. **[24'NeurIPS] Zheng Zhan, Yushu Wu, Yifan Gong, Zichong Meng, Zhenglun Kong, Changdi Yang, Geng Yuan, Pu Zhao, Wei Niu, Yanzhi Wang, "Fast and Memory-Efficient Video Diffusion Using Streamlined Inference", in Proceedings of The Thirty-Eighth Annual Conference on Neural Information Processing Systems.**

18. **[24'DATE] Yanyue Xie, Peiyan Dong, Geng Yuan, Zhengang Li, Masoud Zabihi, and *et.al.*, "SuperFlow: A Fully-Customized RTL-to-GDS Design Automation Flow for Adiabatic Quantum-Flux-Parametron Superconducting Circuits", in Proceedings of 2024 Design, Automation & Test in Europe Conference & Exhibition (DATE).**

19. **[24'DATE] Yushu Wu, Chao Wu, Geng Yuan, Yanyu Li, Weichao Guo, and *et.al.*, "DACO: Pursuing Ultra-low Power Consumption via DNN-Adaptive CPU-GPU CO-optimization on Mobile Devices", in Proceedings of 2024 Design, Automation & Test in Europe Conference & Exhibition (DATE).**

20. **[24'ICCAD] Chao Wu, Yifan Gong, Liangkai Liu, Mengquan Li, Yushu Wu, Xuan Shen, Zhimin Li, Geng Yuan, Weisong Shi, Yanzhi Wang, "AyE-Edge: Automated Deployment Space Search Empowering Accuracy yet Efficient Real-Time Object Detection on the Edge", in Proceedings of 2024 ACM/IEEE International Conference on Computer-Aided Design (ICCAD).**

21. **[23'NeurIPS] Peiyan Dong, Lei Lu, Chao Wu, Cheng Lyu, Geng Yuan, Hao Tang, Yanzhi Wang, "PackQViT: Faster Sub-8-bit Vision Transformers via Full and Packed Quantization on the Mobile", in Proceedings of Advances in Neural Information Processing Systems.**

22. **[23'NeurIPS] Peiyan Dong, Zhenglun Kong, Xin Meng, Pinrui Yu, Yifan Gong, Geng Yuan, Hao Tang, Yanzhi Wang, "HotBEV: Hardware-oriented Transformer-based Multi-View 3D Detector for BEV Perception", in Proceedings of Advances in Neural Information Processing Systems.**

23. **[23'MICRO] Zhengang Li, Geng Yuan, Tomoharu Yamauchi, Zabihi Masoud, Yanyue Xie, Peiyan Dong, Xulong Tang, Nobuyuki Yoshikawa, Devesh Tiwari, Yanzhi Wang, Olivia Chen, "SupeRBNN: Randomized Binary Neural Network Using Adiabatic Superconductor Josephson Devices", in Proceedings of the 56th Annual IEEE/ACM International Symposium on Microarchitecture.**

24. **[23'ICCAD] Yushu Wu, Yifan Gong, Zheng Zhan, Geng Yuan, Yanyu Li, Qi Wang, Chao Wu, Yanzhi Wang, "MOC: Multi-Objective Mobile CPU-GPU Co-Optimization for Power-Efficient DNN Inference", in Proceedings of IEEE/ACM International Conference on Computer Aided Design (ICCAD, 2023).**

25. **[23'AMC-SME] Jun Liu, Chao Wu, Geng Yuan, Wei Niu, Wenbin Zhang, Houbing Herbert Song, "A Scalable Real-time Semantic Segmentation Network for Autonomous Driving", in Proceedings of Advanced Multimedia Computing for Smart Manufacturing and Engineering.**

26. **[23'ICLR] [Spotlight] Sheng Li*, Geng Yuan*, Yue Dai, Youtao Zhang, Yanzhi Wang, Xulong Tang, "SmartFRZ: An Efficient Training Framework using Attention-Based Layer Freezing", in Proceedings of the 11th International Conference on Learning Representations (ICLR, 2023).**

27. **[23'ICLR] Sizhe Chen, Geng Yuan, Xinwen Cheng, Yifan Gong, Minghai Qin, Yanzhi Wang, Xiaolin Huang, "Self-Ensemble Protection: Training Checkpoints Are Good Data Protectors", in Proceedings of the 11th International Conference on Learning Representations (ICLR, 2023).**

28. **[23'DATE] Sung-En Chang*, Geng Yuan*, Alec Lu, Mengshu Sun, Yanyu Li, Xiaolong Ma, Zhengang Li, Yanyue Xie, Minghai Qin, Xue Lin, Zhenman Fang and Yanzhi Wang, "ESRU: Extremely Low-Bit and Hardware-Efficient Stochastic Rounding Unit Design for 8-Bit DNN Training", in Proceedings of the 26th Design, Automation & Test in Europe Conference & Exhibition (DATE, 2023).**

29. **[23'AAAI] Zhenglun Kong, Haoyu Ma, Geng Yuan, Mengshu Sun, Yanyue Xie, and *et.al.*, "Peeling the Onion: Hierarchical Reduction of Data Redundancy for Efficient Vision Transformer Training", submitted and under review in the 37th AAAI Conference on Artificial Intelligence (AAAI, 2023).**

30. **[23'AAAI] Yanyu Li, Changdi Yang, Pu Zhao, Geng Yuan, Wei Niu, and *et.al.*, "Towards Real-Time Segmentation on the Edge", submitted and under review in the 37th AAAI Conference on Artificial Intelligence (AAAI, 2023).**

31. **[22'NeurIPS] Geng Yuan*, Yanyu Li*, Sheng Li, Zhenglun Kong, Sergey Tulyakov, Xulong Tang, Yanzhi Wang, Jian Ren, "Layer Freezing & Data Sieving: Missing Pieces of a Generic Framework for Sparse Training", in Proceedings of the 36th Conference on Neural Information Processing Systems (NeurIPS, 2022).**

32. **[22'NeurIPS] Yanyu Li*, Geng Yuan*, Yang Wen, Ju Hu, Georgios Evangelidis, Sergey Tulyakov, Yanzhi Wang, Jian Ren, "EfficientFormer: Vision Transformers at MobileNet Speed", in Proceedings of the 36th Conference on Neural Information Processing Systems (NeurIPS, 2022).**

33. **[22'NeurIPS] Zifeng Wang, Zheng Zhan, Yifan Gong, Geng Yuan, Wei Niu, Tong Jian, Bin Ren, Stratis Ioannidis, Yanzhi Wang, Jennifer Dy, "SparCL: Sparse Continual Learning on the Edge", in Proceedings of the 36th Conference on Neural Information Processing Systems (NeurIPS, 2022).**

34. **[22'ECCV] Geng Yuan, Sung-En Chang, Qing Jin, Alec Lu, Yanyu Li, Yushu Wu, Zhenglun Kong, Yanyue Xie, Peiyan Dong, Minghai Qin, Xiaolong Ma, Xulong Tang, Zhenman Fang, Yanzhi Wang "You Already Have It: A Generator-Free Low-Precision DNN Training Framework using Stochastic Rounding", in Proceedings of the European Conference on Computer Vision (ECCV, 2022).**

35. **[22'TECS Journal] Geng Yuan, Peiyan Dong, Mengshu Sun, Wei Niu, Zhengang Li, Yuxuan Cai, Yanyu Li, Jun Liu, Weiwen Jiang, Xue Lin, Bin Ren, Xulong Tang, Yanzhi Wang, "Mobile or FPGA? A Comprehensive Evaluation on Energy Efficiency and a Unified Optimization Framework", in ACM Transactions on Embedded Computing Systems (TECS, 2022).**

36. **[22'TODAES Journal] Yifan Gong, Geng Yuan, Zheng Zhan, Wei Niu, Zhengang Li, Pu Zhao, Yuxuan Cai, Sijia Liu, Bin Ren, Xue Lin, Xulong Tang, Yanzhi Wang, "Automatic Mapping of the Best-Suited DNN Pruning Schemes for Real-Time Mobile Acceleration", in ACM Transactions on Design Automation of Electronic Systems (TODAES, 2022).**

37. **[22'DATE] Siyue Wang*, Geng Yuan*, Xiaolong Ma, Yanyu Li, Xue Lin, Bhavya Kailkhura, "Fault-tolerant deep neural networks for processing-in-memory based autonomous edge systems", in Proceedings of the 25th Design, Automation & Test in Europe Conference & Exhibition (DATE, 2022).**

38. **[22'DAC] Sung-En Chang*, Geng Yuan*, Alec Lu*, Mengshu Sun, Yanyu Li, Xiaolong Ma, Zhengang Li, Yanyue Xie, Minghai Qin, Xue Lin, Zhenman Fang, Yanzhi Wang, "Hardware-efficient stochastic rounding unit design for DNN training: Late Breaking Results", in Proceedings of the 59th ACM/IEEE Design Automation Conference (DAC, 2022).**

39. **[22'ISQED] Xiaolong Ma*, Geng Yuan*, Zhengang Li*, Yifan Gong, Tianyun Zhang, Wei Niu, Zheng Zhan, Pu Zhao, Ning Liu, Jian Tang, Xue Lin, Bin Ren, Yanzhi Wang, "BLCR: Towards Real-time DNN Execution with Block-based Reweighted Pruning", in Proceedings of the 23rd International Symposium on Quality Electronic Design (ISQED, 2022).**

40. **[22'WWW] Bingyao Li*, Qi Xue*, Geng Yuan*, Sheng Li, Xiaolong Ma, Yanzhi Wang, Xulong Tang, "Optimizing Data Layout for Training Deep Neural Networks", in Companion Proceedings of the Web Conference (WWW, 2022).**

41. **[22'CLOUD] Danlin Jia, Geng Yuan, Xue Lin, Ningfang Mi, "A Data-Loader Tunable Knob to Shorten GPU Idleness for Distributed Deep Learning", in Proceedings of the IEEE 15th International Conference on Cloud Computing (CLOUD, 2022).**

42. **[21'ISCA] Geng Yuan*, Payman Behnam*, Zhengang Li, Ali Shafiei, Sheng Lin, Xiaolong Ma, Hang Liu, Xuehai Qian, Mahdi Nazm Bojnordi, Yanzhi Wang, Caiwen Ding, "FORMS: Fine-grained Polarized ReRAM-based In-situ Computation for Mixed-Signal DNN Accelerator", in Proceedings of the 48th International Symposium on Computer Architecture (ISCA, 2021).**

43. **[21'NeurIPS] [Spotlight (Top 3%)] Geng Yuan*, Xiaolong Ma*, Wei Niu, Zhengang Li, Zhenglun Kong, Ning Liu, Yifan Gong, Zheng Zhan, Chaoyang He, Qing Jin, and *et.al.*, "MEST: Accurate and Fast Memory-Economic Sparse Training Framework on the Edge", in Proceedings of the 35th Conference on Neural Information Processing Systems (NeurIPS, 2021).**

44. **[21'NeurIPS] Xiaolong Ma*, Geng Yuan*, Xuan Shen, Tianlong Chen, Xuxi Chen, Xiaohan Chen, Ning Liu, Minghai Qin, Sijia Liu, Zhangyang Wang, Yanzhi Wang, "Sanity Checks for Lottery Tickets: Does Your Winning Ticket Really Win the Jackpot?", in Proceedings of the 35th Conference on Neural Information Processing Systems (NeurIPS, 2021).**

45. **[21'ICML] Ning Liu*, Geng Yuan*, Zhengping Che, Xuan Shen, Xiaolong Ma, Qing Jin, Jian Ren, Jian Tang, Sijia Liu, Yanzhi Wang, "Lottery Ticket Preserves Weight Correlation: Is It Desirable or Not?", in Proceedings of the 38th International Conference on Machine Learning (ICML, 2021).**

46. **[21'DATE] [Best Paper Nomination] Geng Yuan, Payman Behnam, Yuxuan Cai, Ali Shafiee, Jingyan Fu, Zhiheng Liao, Zhengang Li, Xiaolong Ma, Jieren Deng, Jinhui Wang, Mahdi Bojnordi, Yanzhi Wang, Caiwen Ding, "TinyADC: Peripheral Circuit-aware Weight Pruning Framework for Mixed-signal DNN Accelerators", in Proceedings of the 24th Design, Automation and Test in Europe Conference (DATE, 2021).**

47. **[21'CVPR] [Oral Paper (Top 5%)] Zhengang Li*, Geng Yuan*, Wei Niu, Pu Zhao, Yanyu Li, Yuxuan Cai, Xuan Shen, Zheng Zhan, Zhenglun Kong, Qing Jin, Zhiyu Chen, Sijia Liu, Kaiyuan Yang, Bin Ren, Yanzhi Wang, Xue Lin, "NPAS: A Compiler-aware Framework of Unified Network Pruning and Architecture Search for Beyond Real-Time Mobile Acceleration", in Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR, 2021).**

48. **[21'AAAI] Yuxuan Cai*, Geng Yuan*, Hongjia Li, Wei Niu, Yanyu Li, Xulong Tang, Bin Ren, Yanzhi Wang, "A Compression-Compilation Co-Design Framework Towards Real-Time Object Detection on Mobile Devices", in Proceedings of the AAAI Conference on Artificial Intelligence (AAAI, 2021).**

49. **[21'AAAI] Yuxuan Cai*, Hongjia Li*, Geng Yuan*, Wei Niu, Yanyu Li, Xulong Tang, Bin Ren, Yanzhi Wang, "YOLObile: Real-time object detection on mobile devices via compression-compilation co-design", in Proceedings of the AAAI Conference on Artificial Intelligence (AAAI, 2021).**

50. **[21'ISQED] Geng Yuan, Zhiheng Liao, Xiaolong Ma, Yuxuan Cai, Zhenglun Kong, Xuan Shen, Jingyan Fu, Zhengang Li, Chengming Zhang, Hongwu Peng, Ning Liu, Ao Ren, Jinhui Wang, Yanzhi Wang, "Improving DNN Fault Tolerance using Weight Pruning and Differential Crossbar Mapping for ReRAM-based Edge AI", in Proceedings of the 22th International Symposium on Quality Electronic Design (ISQED, 2021).**

51. **[21'ICS] Chengming Zhang, Geng Yuan, Wei Niu, Jiannan Tian, Sian Jin, Donglin Zhuang, Zhe Jiang, Yanzhi Wang, Bin Ren, Shuaiwen Leon Song, Dingwen Tao, "ClickTrain: efficient and accurate end-to-end deep learning training via fine-grained architecture-preserving pruning", in Proceedings of the ACM International Conference on Supercomputing (ICS, 2021).**

52. **[21'DAC] Pu Zhao, Geng Yuan, Yuxuan Cai, Wei Niu, Qi Liu, Wujie Wen, Bin Ren, Yanzhi Wang, Xue Lin, "Neural Pruning Search for Real-Time Object Detection of Autonomous Vehicles", in Proceedings of the ACM/IEEE 58th Design Automation Conference (DAC, 2021).**

53. **[21'ICLR-HAET] [Best Paper Award] Xiaolong Ma*, Zhengang Li*, Geng Yuan*, Wei Niu, Bin Ren, Yanzhi Wang, Xue Lin, "Memory-Bounded Sparse Training on the Edge", (ICLR 2021 workshop of Hardware-Aware Efficient Training of Deep Learning Models).**

54. **[21'ICCV] Zheng Zhan, Yifan Gong, Pu Zhao, Geng Yuan, Wei Niu, Yushu Wu, and *et.al.*, "Achieving on-Mobile Real-Time Super-Resolution with Neural Architecture and Pruning Search", in Proceedings of the international conference on computer vision (ICCV, 2021).**

55. **[21'RTAS] Geng Yuan, Peiyan Dong, Mengshu Sun, Wei Niu, Zhengang Li, and *et.al.*, "Work in progress: Mobile or FPGA? A comprehensive evaluation on energy efficiency and a unified optimization framework", in Proceedings of the IEEE 27th Real-Time and Embedded Technology and Applications Symposium (RTAS, 2021).**

56. **[21'RTAS] Pu Zhao, Wei Niu, Geng Yuan, Yuxuan Cai, Hsin-Hsuan Sung, and *et.al.*, "Industry paper: Towards real-time 3D object detection for autonomous vehicles with pruning search", in Proceedings of the IEEE 27th Real-Time and Embedded Technology and Applications Symposium (RTAS, 2021).**

57. **[21'ASP-DAC] Hongjia Li, Geng Yuan, Wei Niu, Yuxuan Cai, Mengshu Sun, Zhengang Li, Bin Ren, Xue Lin, Yanzhi Wang, "Real-time mobile acceleration of DNNs: From computer vision to medical applications", in Proceedings of the 26th Asia and South Pacific Design Automation Conference (ASP-DAC, 2021).**

58. **[21'TNNLS Journal] Xiaolong Ma, Sheng Lin, Shaokai Ye, Zhezhi He, Linfeng Zhang, Geng Yuan, and *et.al.*, "Non-Structured DNN Weight Pruning–Is It Beneficial in Any Platform?", IEEE Transactions on Neural Networks and Learning Systems (TNNLS, 2021).**

59. **[21'IJCAI demo] Xuan Shen*, Geng Yuan*, Wei Niu, Xiaolong Ma, Jiexiong Guan, Zhengang Li, Bin Ren, Yanzhi Wang, "Towards Fast and Accurate Multi-Person Pose Estimation on Mobile Devices", in Proceedings of the 30th International Joint Conference on Artificial Intelligence (IJCAI-21) Demonstrations Track.**

60. **[21'ICCAD] Xu, Weizheng, Ashutosh Pattnaik, Geng Yuan, Yanzhi Wang, Youtao Zhang, and Xulong Tang, "ScaleDNN: Data Movement Aware DNN Training on Multi-GPU", in Proceedings of the 40th International Conference on Computer-Aided Design (ICCAD, 2021).**

61. **[20'ISLPED] [Design Contest 1st Place Winner] Geng Yuan, Wei Niu, Pu Zhao, Xue Lin, Bin Ren, and Yanzhi Wang, "CoCoPIE: A Framework of Compression-Compilation Co-design Towards Ultra-high Energy Efficiency and Real-Time DNN Inference on Mobile Devices", IEEE/ACM International Symposium on Low Power Electronics and Design (ISLPED, 2020).**

62. **[20'ICCAD] Gongye, Cheng, Hongjia Li, Xiang Zhang, Majid Sabbagh, Geng Yuan, Xue Lin, Thomas Wahl, and Yunsi Fei, "New passive and active attacks on deep neural networks in medical applications", in Proceedings of the 39th International Conference on Computer-Aided Design (ICCAD, 2020).**

63. **[20'ASP-DAC] Xiaolong Ma, Geng Yuan, Sheng Lin, Caiwen Ding, Fuxun Yu, Tao Liu, Wujie Wen, Xiang Chen, Yanzhi Wang, "Tiny but Accurate: A Pruned, Quantized and Optimized Framework of an Ultra Efficient DNN Device", in 25th Asia and South Pacific Design Automation Conference (ASP-DAC, 2020).**

64. **[19'ISLPED] Geng Yuan, Xiaolong Ma, Caiwen Ding, Sheng Lin, Tianyun Zhang, Zeinab S. Jalali, Yilong Zhao, Li Jiang, Sucheta Soundarajan, Yanzhi Wang, "An Ultra-Efficient Memristor-Based DNN Framework with Structured Pruning and Quantization Using ADMM", in IEEE/ACM International Symposium on Low Power Electronics and Design (ISLPED, 2019).**

65. **[19'NANOARCH] Xiaolong Ma, Geng Yuan, Sheng Lin, Zhengang Li, Yanzhi Wang, "ResNet Can Be Pruned 60x: Introducing Network Purification and Unused Path Removal (P-RM) after Weight Pruning", in 15th IEEE / ACM International Symposium on Nanoscale Architectures (NANOARCH, 2019).**

66. **[18'AAAI] Yanzhi Wang, Caiwen Ding, Zhe Li, Geng Yuan, Siyu Liao, Xiaolong Ma, Bo Yuan, Xuehai Qian, Jian Tang, Qinru Qiu, Xue Lin. "Towards ultra-high performance and energy efficiency of deep learning systems: an algorithm-hardware co-optimization framework", in AAAI Conference on Artificial Intelligence (AAAI, 2018).**

67. **[18'ISQED] [Best Paper Nomination] Xiaolong Ma, Yipeng Zhang, Geng Yuan, Ao Ren, Zhe Li, Jie Han, Jingtong Hu, Yanzhi Wang. "An Area and Energy Efficient Design of Domain-Wall Memory-Based Deep Convolutional Neural Networks using Stochastic Computing", in International Symposium on Quality Electronic Design (ISQED, 2018).**

68. **[18'GLSVLSI] Caiwen Ding, Ao Ren, Geng Yuan, Xiaolong Ma, Jiayu Li, Ning Liu, Bo Yuan, Yanzhi Wang. "Structured Weight Matrices-Based Hardware Accelerators in Deep Neural Networks: FPGAs and ASICs" in Proceedings of the 2018 on Great Lakes Symposium on VLSI (GLSVLSI, 2018).**

69. **[17'MICRO] Caiwen Ding, Siyu Liao, Yanzhi Wang, Zhe Li, Ning Liu, Youwei Zhuo, Chao Wang, Xuehai Qian, Yu Bai, Geng Yuan, and *et.al.*, "CirCNN: accelerating and compressing deep neural networks using block-circulant weight matrices", in Proceedings of the International Symposium on Microarchitecture (MICRO, 2017).**

70. **[17'MWSCAS] Geng Yuan, Caiwen Ding, Ruizhe Cai, Xiaolong Ma, Ziyi Zhao, Ao Ren, Bo Yuan, Yanzhi Wang. "Memristor crossbar-based ultra-efficient next-generation baseband processors", in IEEE 60th International Midwest Symposium on Circuits and Systems (MWSCAS, 2017).**