

Wei Niu

Assistant Professor, School of Computing, University of Georgia | wniu@uga.edu | [Google Scholar](#) | [Homepage](#)

RESEARCH INTERESTS

- Real-time Machine Learning
- AGI on Embedded Devices
- Parallel Computing
- Compiler Optimizations

WORK EXPERIENCE

University of Georgia

Assistant Professor, School of Computing

Athens, GA

Aug. 2023 — Present

Bytedance

Software Engineer

Beijing, China

Jul. 2016 — Jul. 2018

EDUCATION EXPERIENCE

William & Mary

Doctor of Philosophy in Computer Science, advised by [Dr. Bin Ren](#)

Williamsburg, VA

Aug. 2018 — Jul. 2023

Beihang University

Bachelors of Science in Software Engineering

Beijing, China

Aug. 2012 — Aug. 2016

EXTERNAL GRANTS

- G1 NSF (#2428108): “SHF Core: Memory Hierarchy Optimizations Meet Transformers (MITTEN)” Oct. 2024 – Sep. 2027
‣ **Leading PI** (Total \$600,000)
- G2 NSF (#2403090): “Collaborative Research: OAC Core: CropDL - Scheduling and Checkpoint/Restart Support for Deep Learning Applications on HPC Clusters” Oct. 2024 – Sep. 2027
‣ **PI** (Total \$600,000)

PUBLICATIONS

Peer-reviewed Conference Publications (* means equal contribution, my Ph.D. advisees are highlighted)

- C1 [NeurIPS'24] **Zhihao Shu***, Xiaowei Yu*, Zihao Wu, Wenqi Jia, Yinchun Shi, Miao Yin, Tianming Liu, Dajiang Zhu, **Wei Niu**, “Real-time Core-Periphery Guided ViT with Smart Data Layout Selection on Mobile Devices”, *The Thirty-Eighth Annual Conference on Neural Information Processing Systems, 2024*
- C2 [NeurIPS'24] 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”, *The Thirty-Eighth Annual Conference on Neural Information Processing Systems, 2024*
- C3 [NeurIPS'24] Zheng Zhan, Zhenglun Kong, Yifan Gong, Yushu Wu, Zichong Meng, **Hangyu Zheng**, Xuan Shen, Stratis Ioannidis, **Wei Niu**, Pu Zhao, Yanzhi Wang, “Exploring Token Pruning in Vision State Space Models”, *The Thirty-Eighth Annual Conference on Neural Information Processing Systems, 2024*
- C4 [Healthcare'24] Jiexiong Guan, Junjie Wang, **Wei Niu**, Zhen Peng, Shuangquan Wang, Zhenming Liu, Gang Zhou, Bin Ren, “Towards Recognizing Food Types for Unseen Subjects”, *ACM Transactions on Computing for Healthcare, 2024*
- C5 [ASPLOS'24] **Wei Niu**, Gagan Agrawal, Bin Ren, “SoD²: Statically Optimizing Dynamic Deep Neural Network Execution”, *The 29th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, 2024*

- C6 [ASPLOS'24] **Wei Niu**, Md Musfiqur Rahman Sanim, **Zhihao Shu**, Jiexiong Guan, Xipeng Shen, Miao Yin, Gagan Agrawal, Bin Ren, "SmartMem: Layout Transformation Elimination and Adaptation for Efficient DNN Execution on Mobile", *29th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, 2024*
- C7 [ICLR'24] Gen Li, Lu Yin, Jie Ji, **Wei Niu**, Minghai Qin, Bin Ren, Linke Guo, Shiwei Liu, Xiaolong Ma, "NeurRev: Train Better Sparse Neural Network Practically via Neuron Revitalization", *The Twelfth International Conference on Learning Representations*
- C8 [ECCV'24] Gen Li, **Zhihao Shu**, Jie Ji, Minghai Qin, Fatemeh Afghah, **Wei Niu**, Xiaolong Ma, "Data Overfitting for On-Device Super-Resolution with Dynamic Algorithm and Compiler Co-Design", *The European Conference on Computer Vision*
- C9 [AMC-SME'23] Jun Liu, Chao Wu, Geng Yuan, **Wei Niu**, Wenbin Zhang, Houbing Herbert Song, "A Scalable Real-time Semantic Segmentation Network for Autonomous Driving", *Advanced Multimedia Computing for Smart Manufacturing and Engineering*
- C10 [USENIX ATC'23] Hsin-Hsuan Sung, Jiexiong Guan, **Wei Niu**, Jou-An Chen, Bin Ren, Xipeng Shen, "Decentralized Application-Level Adaptive Scheduling for Multi-Instance DNNs on Open Mobile Devices", *2023 USENIX Annual Technical Conference*
- C11 [CVPR'23] Gen Li, Jie Ji, Minghai Qin, **Wei Niu**, Bin Ren, Fatemeh Afghah, Linke Guo, Xiaolong Ma, "Towards High-Quality and Efficient Video Super-Resolution via Spatial-Temporal Data Overfitting", *IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023*
- C12 [CVPR'23] Changdi Yang, Pu Zhao, Yanyu Li, **Wei Niu**, Jiexiong Guan, Hao Tang, Minghai Qin, Bin Ren, Xue Lin, Yanzhi Wang, "Pruning Parameterization with Bi-level Optimization for Efficient Semantic Segmentation on the Edge", *IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023*
- C13 [AAAI'23] Yanyu Li, Changdi Yang, Pu Zhao, Geng Yuan, **Wei Niu**, Jiexiong Guan, Hao Tang, Minghai Qin, Qing Jin, Bin Ren, Xue Lin, Yanzhi Wang, "Towards Real-Time Segmentation on the Edge", *Thirty-Seventh AAAI Conference on Artificial Intelligence*
- C14 [NeurIPS'22] 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", *The 36th Conference on Neural Information Processing Systems, 2022*
- C15 [MICRO'22] **Wei Niu**, Jiexiong Guan, Xipeng Shen, Yanzhi Wang, Gagan Agrawal, Bin Ren, "GCD²: A Globally Optimizing Compiler for Mapping DNNs to Mobile DSPs", *The 55th IEEE/ACM International Symposium on Microarchitecture, 2022*
- C16 [ECCV'22] Zhenglun Kong, Peiyan Dong, Xiaolong Ma, Xin Meng, **Wei Niu**, Mengshu Sun, Xuan Shen, Geng Yuan, Bin Ren, Minghai Qin, Hao Tang, Yanzhi Wang, "SPViT: Enabling Faster Vision Transformers via Soft Token Pruning", *Proceeding of European Conference on Computer Vision, 2022*
- C17 [ECCV'22] Yushu Wu, Yifan Gong, Pu Zhao, Yanyu Li, Zheng Zhan, **Wei Niu**, Hao Tang, Minghai Qin, Bin Ren, Yanzhi Wang, "Compiler-Aware Neural Architecture Search for On-Mobile Real-time Super-Resolution", *Proceeding of European Conference on Computer Vision, 2022*
- C18 [ISQED'22] Xiaolong Ma, Geng Yuan, Zhengang Li, Yifan Gong, Tianyun Zhan, **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", *23rd International Symposium on Quality Electronic Design, 2022*
- C19 [CVPR'21] 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", *IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021 (oral paper: top 5%)*
- C20 [AAAI'21] **Wei Niu***, Mengshu Sun*, Zhengang Li*, Jou-An Chen, Jiexiong Guan, Xipeng Shen, Yanzhi Wang, Xue Lin, Bin Ren, "Achieving Real-Time Execution of 3D Convolutional Neural Networks on Mobile Devices", *The 35th AAAI Conference on Artificial Intelligence, February 2021*

- C21 [**NeurIPS'21**] Geng Yuan, Xiaolong Ma, **Wei Niu**, Zhengang Li, Zhenglun Kong, Ning Liu, Yifan Gong, Zheng Zhan, Chaoyang He, Qing Jin, Siyue Wang, Minghai Qin, Bin Ren, Yanzhi Wang, Sijia Liu, Xue Lin, "MEST: Accurate and Fast Memory-Economic Sparse Training Framework on the Edge", *The 35th Conference on Neural Information Processing Systems, 2021*
- C22 [**PLDI'21**] **Wei Niu**, Jiexiong Guan, Yanzhi Wang, Gagan Agrawal, Bin Ren, "DNNFusion: Accelerating Deep Neural Networks Execution with Advanced Operator Fusion", *42nd ACM SIGPLAN Conference on Programming Language Design and Implementation, 2021*
- C23 [**HiPC'21**] Qihan Wang, **Wei Niu**, Li Chen, Ruoming Jin, Bin Ren, "HEALS: A Parallel eALS Recommendation System on CPU/GPU Heterogeneous Platforms", *IEEE International Conference on High Performance Computing, Data, Analytics, 2021*
- C24 [**ICCV'21**] Zheng Zhan, Yifan Gong, Pu Zhao, Geng Yuan, **Wei Niu**, Yushu Wu, Tianyun Zhang, Malith Jayaweera, David Kaeli, Bin Ren, Xue Lin, Yanzhi Wang, "Achieving on-Mobile Real-Time Super-Resolution with Neural Architecture and Pruning Search", *International Conference on Computer Vision, 2021*
- C25 [**ICS'21**] 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", *The 35th ACM International Conference on Supercomputing, 2021*
- C26 [**DAC'21**] 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", *The 58th Annual Design Automation Conference, 2021*
- C27 [**ASP-DAC'21**] Hongjia Li, Geng Yuan, **Wei Niu**, Yuxuan Cai, Mengshu Sun, Zhengang Li, Bin Ren, Xue Lin, and Yanzhi Wang, "Real-Time Mobile Acceleration of DNNs: From Computer Vision to Medical Applications", *Proceeding of Asia and South Pacific Design Automation Conference, 2021*
- C28 [**CACM'21**] Hui Guan, Shaoshan Liu, Xiaolong Ma, **Wei Niu**, Bin Ren, Xipeng Shen, Yanzhi Wang, Pu Zhao (in alphabet order), "CoCoPIE: Making Mobile AI Sweet as PIE - Compression-Compilation Co-Design Goes a Long Way", *Communications of the ACM (CACM), 2021 (flagship journal of ACM, featured with a video report)*
- C29 [**AAAI'21**] Yuxuan Cai, Hongjia Li, Geng Yuan, **Wei Niu**, Yanyu Li, Xulong Tang, Bin Ren, Yanzhi Wang, "YOLOmobile: Real-Time Object Detection on Mobile Devices via Compression-Compilation Co-Design", *The 35th AAAI Conference on Artificial Intelligence, February 2021*
- C30 [**GLSVLSI'20 Special Session Paper**] Yifan Gong, Zheng Zhan, Zhengang Li, **Wei Niu**, Xiaolong Ma, Wenhao Wang, Bin Ren, Caiwen Ding, Xue Lin, Xiaolin Xu, Yanzhi Wang, "A Privacy-Preserving-Oriented DNN Pruning and Mobile Acceleration Framework", *Proceedings of 2020 on Great Lakes Symposium on VLSI, Sep. 2020*
- C31 [**ECCV'20**] Xiaolong Ma*, **Wei Niu***, Tianyun Zhang, Sijia Liu, Sheng Lin, Hongjia Li, Wujie Wen, Xiang Chen, Jian Tang, Kaisheng Ma, Bin Ren, Yanzhi Wang, "An Image Enhancing Pattern-based Sparsity for Real-time Inference on Mobile Devices", *16th European Conference on Computer Vision, August 2020*
- C32 [**DAC'20**] Peiyan Dong, Siyue Wang, **Wei Niu**, Chengming Zhang, Sheng Lin, Zhengang Li, Yifan Gong, Bin Ren, Xue Lin, and Dingwen Tao, "RTMobile: Beyond Real-Time Mobile Acceleration of RNNs for Speech Recognition", *The 57th Annual Design Automation Conference, July 2020*
- C33 [**AAAI'20**] Xiaolong Ma, Fu-Ming Guo, **Wei Niu**, Xue Lin, Jian Tang, Kaisheng Ma, Bin Ren, Yanzhi Wang, "PCONV: The Missing but Desirable Sparsity in DNN Weight Pruning for Real-time Execution on Mobile Devices", *The 34th AAAI Conference on Artificial Intelligence, February 2020*
- C34 [**ASPLOS'20**] **Wei Niu**, Xiaolong Ma, Sheng Lin, Shihao Wang, Xuehai Qian, Xue Lin, Yanzhi Wang, Bin Ren, "PatDNN: Achieving Real-Time DNN Execution on Mobile Devices with Pattern-based Weight Pruning", *The 25th International Conference on Architectural Support for Programming Languages and Operating Systems, 2020*

Peer-reviewed Journal Publications

- J1 [**CSUR'22**] Jou-an Chen, **Wei Niu**, Bin Ren, Yanzhi Wang, Xipeng Shen, "Survey: Exploiting Data Redundancy for Optimization of Deep Learning", *ACM Computing Surveys(CSUR), 2022*

- J2 [TDAES'22] 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", *ACM Transactions on Design Automation of Electronic Systems (TDAES)*, 2022
- J3 [TECS'22] Geng Yuan, 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", *ACM Transactions on Embedded Computing Systems (TECS)*, 2022
- J4 [TPAMP'21] **Wei Niu**^{*}, Zhengang Li^{*}, Xiaolong Ma, Peiyang Dong, Gang Zhou, Xuehai Qian, Xue Lin, Yanzhi Wang, Bin Ren, "GRIM: A General, Real-Time Deep Learning Inference Framework for Mobile Devices based on Fine-Grained Structured Weight Sparsity", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021

TEACHING

- CSCI 4130/6130: CUDA C Programming on GPUs for High Performance Computing Instructor
 - 24'Fall, 23'Fall
- CSCI 8000: Advanced Topics in Machine Learning Systems Instructor
 - 24'Spring
- CSCI 434: Network Systems and Design Teaching Assistant
 - 20'Spring
- CSCI 304: Computer Organization Teaching Assistant
 - 19'Fall, 18'Fall
- CSCI 312: Principles of Programming Languages Teaching Assistant
 - 19'Spring

PROFESSIONAL ACTIVITIES

Program Committee

- [MLSys] The Eighth Annual Conference on Machine Learning and Systems 2025
- [PPOPP] ACM SIGPLAN Symposium on Principles and Practice of Parallel Programming 2025
- [WHPC@SC] ACM/IEEE Supercomputing Conference WHPC 2025
- [HiPC] IEEE International Conference on High Performance Computing, Data, and Analytics 2024
- [ICPP] International Conference on Parallel Processing 2024
- [ICCD] IEEE International Conference on Computer Design 2024
- [IPDPS] The 38th IEEE International Parallel & Distributed Processing Symposium 2024
- [ICCCN] International Conference on Computing and Communication Networks 2024
- [MASS] IEEE International Conference on Mobile Ad-Hoc and Smart Systems 2024
- [HiPC] IEEE International Conference on High Performance Computing, Data, and Analytics 2019

Reviewer

- [AISTATS] The 28th International Conference on Artificial Intelligence and Statistics 2025
- [ICLR] The International Conference on Learning Representations 2025
- [AAAI] The Annual AAAI Conference on Artificial Intelligence 2025
- [NeurIPS] The 28th Conference on Neural Information Processing Systems 2024
- [ECCV] European Conference on Computer Vision 2024
- [TCAS] IEEE Transactions on Circuits and Systems II: Express Briefs 2024
- [TACO] Transactions on Architecture and Code Optimization 2023
- [MWSCAS] The 2023 IEEE 66th International Midwest Symposium on Circuits and Systems 2023
- [NeurIPS] Thirty-seventh Conference on Neural Information Processing Systems 2023
- [ICCV] International Conference on Computer Vision 2023
- [IMWUT] Proceedings of the ACM on Interactive, Wearable and Ubiquitous Technologies 2022
- [ICCD] The 40th IEEE International Conference on Computer Design 2022
- [ECCV] European Conference on Computer Vision 2022

- [IJCAI] 31st International Joint Conference on Artificial Intelligence 2022
- [CVPR] IEEE/CVF Conference on Computer Vision and Pattern Recognition 2022
- [AAAI] 36th AAAI Conference on Artificial Intelligence 2022
- [IPDPS] Distributed Processing Symposium 2021
- [TC] IEEE Transactions on Computers 2021
- [TPDS] IEEE Transactions on Parallel and Distributed Systems 2020
- [Bench] BenchCouncil International Symposium on Benchmarking, Measuring and Optimizing 2020
- [TC] IEEE Transactions on Computers 2020
- [NPC] Annual IFIP International Conference on Network and Parallel Computing 2019
- [SOFC] The 3rd ACSIC Symposium on Frontiers in Computing 2019
- [BIGCOMM] International Conference on Big Data Computing and Communications 2019

AWARDS

- Stephen K.Park Graduate Research Award (highest research award at CS@William & Mary) 2021
- Best Paper Award at ICLR Workshop on Hardware Aware Efficient Training 2021
- Featured Cover Article in Communications of the ACM 2021
- First Place of Design Contest in International Symposium on Low Power Electronics and Design 2020
- Travel Grant: MICRO'22, William & Mary SA Conference Fund'22, ASPLOS'20

INVITED TALKS

- “Real-time DNN Execution on Mobile Devices with Compiler Optimizations” 2024
 - Invited Talk at University of Texas at Arlington, September, 2024
- “SmartMem: Layout Transformation Elimination and Adaptation for Efficient DNN Execution on Mobile” 2024
 - ASPLOS, San Diego, CA, April, 2024
- “SoD²: Statically Optimizing Dynamic Deep Neural Network” 2024
 - ASPLOS, San Diego, CA, April, 2024
- “Achieving Real-Time Execution of Extremely Deep Neural Networks on Mobile Devices” 2024
 - Samsung, San Francisco, CA, April, 2024
- “Real-time Machine Learning Systems with Compiler Optimizations” 2023
 - Clemson University, Guest Lecture, January, 2023
- “GCD²: A Globally Optimizing Compiler for Mapping DNNs to Mobile DSPs” 2022
 - MICRO, Chicago, Illinois, October, 2022
- “DNNFusion: Accelerating Deep Neural Networks Execution with Advanced Operator Fusion” 2020
 - UMASS, Guest Lecture, Virtual, October, 2020
- “PatDNN: Achieving Real-Time DNN Execution on Mobile Devices with Pattern-based Weight Pruning” 2020
 - Google Brain, Virtual, November, 2020

PATENTS

- Method for accelerating deep neural networks execution with advanced operator fusion [US17/842,748]
- BPDNN: A general, real-time DNN execution framework on mobile devices with block-based column-row pruning [No.: 62/976,577]
- RTMobile: A mobile acceleration framework of RNNs for beyond real-time speech recognition [No.: 62/965,275]
- PatDNN: Achieving real-time DNN execution on mobile devices with pattern-based weight pruning [No.: 62/976,595]
- DNNFusion: Accelerating Deep Neural Networks Execution with Advanced Operator Fusion [US-2022-0413862-A1]