HUIHONG SHI

Tel.: +86-17390940085 **Email**: shihh@smail.nju.edu.cn **Profile**: Google Scholar **Homepage**: Link

EDUCATION

Nanjing University Nanjing, China
Ph.D. student supervised by Prof. Zhongfeng Wang (**IEEE Fellow**), Electronic Science and Engineering (EE) Sep. 2020 - Present

Georgia Institute of Technology Atalanta, USA
Visiting student supervised by Prof. Yingyan Lin, Computer Science (CS) Dec. 2023 - Jun. 2024

Rice University Remote
Visiting student supervised by Prof. Yingyan Lin, Electrical and Computer Engineering (ECE) Mar. 2021 - Dec. 2022

Jilin University Changchun, China
Bachelor of Communication Engineering (GPA: 3.8/4.0) Sep. 2016 - Jul. 2020

RESEARCH INTERESTS

- **Efficient and Automated Machine Learning** (ICCAD, TCAS-I, ICML, TMLR, NeurIPS)
- **Algorithm and Hardware Accelerator Co-Design** (ICCAD, HPCA, TCAS-I, TVLSI, ISCAS, ISCA)

UNDERGOING PROJECTS

- **Diffusion Transformer (DiT) Quantization**
  We focus on post-training quantization (PTQ) for DiTs, investigating (1) the impact of different reconstruction methods on DiT quantization and (2) the varying quantization sensitivities across different layers within DiTs.
- **FPGA-Based Large Language Model (LLM) Acceleration**
  We aim to (1) accelerate global and window attention patterns during both the prefill and generation stages of LLMs, and (2) utilize quantization and KV caching strategies to reduce bandwidth requirements during the generation stage.

FINISHED PROJECTS

- **Neural Architecture Search and Acceleration for Hardware Inspired Multiplication-Reduced Networks** (ICCAD 2022, TCAS-I 2023/2024)
  We propose a Neural Architecture Search and Acceleration framework (NASA) to enable automated search and acceleration of multiplication-reduced models, aiming to marry the powerful performance of multiplication-based models and the hardware efficiency of multiplication-free models.
- **Accelerating Vision Transformer with Linear Attention and Dedicated Hardware Accelerator** (HPCA 2023)
  We first approximate the vanilla softmax with first-order Taylor attention for linear complexity and unify low-rank and sparse components to enhance accuracy. We further develop a dedicated accelerator that leverages the linearized workload to improve hardware efficiency.
- **Mixture of Multiplication Primitives Towards Efficient Vision Transformer** (NeurIPS 2023)
  We reparameterize pre-trained Vision Transformers (ViTs) with a mixture of multiplication primitives, such as bitwise shifts and additions, to achieve end-to-end inference speedups on GPUs without requiring training from scratch.
- **Post-Training Quantization and Acceleration of Vision Transformers** (TVLSI 2024 TCAS-I 2024)
  For Standard Vision Transformers (ViTs), we propose a dedicated quantization scheme with Power-of-Two (PoT) scaling factors to minimize re-quantization overhead and further develop a tailored accelerator to enhance throughput.
  In the later work, we further marry the hardware efficiency of both quantization and efficient Vision Transformer (ViT) architectures, which (1) eliminate the troublesome Softmax and (2) integrate linear attention.

PUBLICATIONS

- **H. Shi**, W Mao, Z Wang.
  LITNet: A Light-weight Image Transform Net for Image Style Transfer.
  International Joint Conference on Neural Networks, **IJCNN 2021**

- **H. Shi**, H. You, Y. Zhao, Z. Wang, Y. Lin.
  NASA: Neural Architecture Search and Acceleration for Hardware Inspired Hybrid Networks.
  International Conference on Computer-Aided Design, **ICCAD 2022**

- **H. Shi**, H. You, Z. Wang, Y. Lin.
  NASA+: Neural Architecture Search and Acceleration for Multiplication-Reduced Hybrid Networks.
  IEEE Transactions on Circuits and Systems I, **TCAS-I 2023**

- **H. Shi**, C Xin, W Mao, Z Wang.
  P^2-ViT: Power-of-Two Post-Training Quantization and Acceleration for Fully Quantized Vision Transformer.
  IEEE Transactions on Very Large Scale Integration Systems, **TVLSI 2024**

- **H Shi**, H Shao, W Mao, Z Wang
  Trio-ViT: Post-Training Quantization and Acceleration for Softmax-Free Efficient Vision Transformer.
  IEEE Transactions on Circuits and Systems I, **TCAS-I 2024**

- **H Shi***, Y Xu*, Y Wang, W Mao, Z Wang. **(*Co-First Authors)**
  NASA-F: FPGA-Oriented Search and Acceleration for Multiplication-Reduced Hybrid Networks.
  IEEE Transactions on Circuits and Systems I, **TCAS-I 2024**

- H You*, **H Shi***, Y Guo*, Y Lin **(*Co-First Authors)**
  ShiftAddViT: Mixture of Multiplication Primitives Towards Efficient Vision Transformer.
  Thirty-seventh Conference on Neural Information Processing Systems, **NeurIPS 2023**

- J Dass*, S Wu*, **H Shi***, C Li, Z Ye, Z Wang, Y Lin. **(*Co-First Authors)**
  ViTALiTy: Unifying Low-rank and Sparse Approximation for Vision Transformer Acceleration with a Linear Taylor Attention.
  International Symposium on High-Performance Computer Architecture, **HPCA 2023**

- Yang X*, **H Shi***, Z Wang **(*Co-First Authors)**
  NASH: Neural Architecture and Accelerator Search for Multiplication-Reduced Hybrid Models.
  IEEE Transactions on Circuits and Systems I, **TCAS-I 2024**

- H Shao, **H Shi**, W Mao, Z Wang
  An FPGA-Based Reconfigurable Accelerator for Convolution-Transformer Hybrid EfficientViT.
  IEEE International Symposium on Circuits and Systems 2024, **ISCAS 2024**

- Y Liang, **H Shi**, Z Wang
  M^2-ViT: Accelerating Convolution-Transformer Hybrid Vision Transformers with Double-Mixed Quantization.
  IEEE Transactions on Circuits and Systems II, **TCAS-II 2024 (Under Review)**

- W Mao, S Yang, **H Shi**, J Liu, Z Wang
  Intelligent Typography: Artistic Text Style Transfer for Complex Texture and Structure.
  IEEE Transactions on Multimedia 2022, **TMM 2022**

- H. You, B. Li, **H. Shi**, Y. Lin
  ShiftAddNAS: Hardware-Inspired Search for More Accurate and Efficient Neural Networks.
  International Conference on Machine Learning, **ICML 2022**

- H. You, Z. Sun, **H. Shi**, Z. Yu, Y. Zhao, Y. Zhang, C. Li, B. Li, Y. Lin.
  Selected as the Meta Faculty Research Award of 2022!

ViTCoD: Vision Transformer Acceleration via Dedicated Algorithm and Accelerator Co-Design.
International Symposium on High-Performance Computer Architecture, **HPCA 2023**

- M. She, W. Mao, **H. Shi**, Z Wang.
S^2R: Exploring a Double-Win Transformer-Based Framework for Ideal and Blind Super-Resolution.
International Conference on Artificial Neural Networks, **ICANN 2023**

- S Zhang, W Mao, **H Shi**, Z Wang,
A Computationally Efficient Neural Video Compression Accelerator Based on a Sparse CNN-Transformer Hybrid Network.
Design, Automation and Test in Europe Conference, **DATE 2024**

- Z Yu, Z Wang, Y Fu, **H Shi**, K Shaikh, Y Lin,
Unveiling and Harnessing Hidden Attention Sinks: Enhancing Large Language Models without Training through Attention Calibration.
International Conference on Machine Learning, **ICML 2024**

- H You, R Balestriero, Z Lu, Y Kou, **H Shi**, S Zhang, S Wu, Y Lin,
Max-Affine Spline Insights Into Deep Network Pruning.
Transactions on Machine Learning Research, **TMLR 2022**

- H You, Y Guo, Y Fu, W Zhou, **H Shi**, X Zhang, S Kundu, A Yazdanbakhsh, Y Lin,
ShiftAddLLM: Accelerating Pretrained LLMs via Post-Training Multiplication-Less Reparameterization.
Thirty-eight Conference on Neural Information Processing Systems, **NeurIPS 2024**

- S Li, C Li, W Zhu, C Wan, H You, **H Shi**, Y Lin,
Instant-3D: Instant Neural Radiance Field Training Towards On-Device AR/VR 3D Reconstruction.
International Symposium on Computer Architecture, **ISCA 2023**

SELECTED AWARDS

- The First-Class Academic Scholarship for Postgraduate Students at Nanjing University — Sep. 2024
- The First-Class Academic Scholarship for Postgraduate Students at Nanjing University — Sep. 2023
- The First-Class Academic Scholarship for Postgraduate Students at Nanjing University — Sep. 2022
- Excellence Scholarship for Postgraduate Students at Nanjing University — Sep. 2021
- The First-Class Academic Scholarship for Postgraduate Students at Nanjing University — Sep. 2021
- President's Special Scholarship for Doctoral Candidate of Nanjing University — Sep. 2020
- Post and Telecommunications Alumni Scholarship for Undergraduates of Jilin University — Oct. 2019
- The First-Class Academic Scholarship for Undergraduates of Jilin University — Sep. 2019
- The First-Class Academic Scholarship for Undergraduates of Jilin University — Sep. 2018
- National Scholarship Award for Undergraduates Issued by Ministry of Education of China — Sep. 2017

MENTORING EXPERIENCE

- **Yang Xu** (MS@Nanjing University), Working on two co-1st author journal papers
- **Xinyu Ding** (MS@Sun Yat-sen University), Guiding her undergraduate graduation project, which received the best graduation paper from Zhengzhou University in 2023
- **Yanbiao Liang** (MS@Nanjing University), Working on one journal paper
- **Xinyan Liu** (MS@Nanjing University), Working on one journal paper

REVIEW EXPERIENCE

- I am serving as a reviewer for ICLR 2025, NeurIPS 2025, TNNLS, and TCAS