Haoran You

Personal Website Google Scholar LinkedIn
haoran.you@gatech.edu +1 (281) 236-1978

EDUCATION

- **Georgia Institute of Technology** — Atlanta, Georgia
  Ph.D. Candidate in CS. Advisor: Prof. Yingyan (Celine) Lin — Dec. 2022 - Present
- **Rice University** — Houston, Texas
  M.S. Degree & Ph.D. Student in ECE. Advisor: Prof. Yingyan (Celine) Lin — Sep. 2019 - Dec. 2022
- **Huazhong University of Science and Technology** — Wuhan, China
  B.S. Degree in Electronic and Information Engineering. Advanced Class of 2015 — Sep. 2015 - Jul. 2019

WORKING EXPERIENCE

- **Adobe Research & Photoshop** — May. 2024 - Dec. 2024
  Research Intern on Efficient GenAI
  Mentors: Sohrab Amirghodsi, Yuqian Zhou, Zhe Lin, Yan Kang, Connelly Barnes, Eli Shechtman
- **CREATE-X & VentureLab** — May. 2023 - Aug. 2023
  Startup Founder. Idea: EyeCoD [IEEE Micro's Top Pick of 2023]
  Co-founders: Yang (Katie) Zhao, Yingyan (Celine) Lin.
- **Meta Reality Lab (Team Now: Meta GenAI)** — May. 2022 - Dec. 2022
  Research Intern on Efficient ViTs. Publication: Caslting-ViT [CVPR'23]
  Mentors: Peizhao Zhang, Yunyang Xiong, Xiaoliang Dai, Bichen Wu, Peter Vajda
- **Baidu USA** — May. 2021 - Apr. 2022
  Research Intern on Efficient ML. Mentors: Baopu Li
  Publications: ShiftAddNAS [ICML'22], SuperTickets [ECCV'22]

RESEARCH STATEMENT

- **Hierarchical algorithm optimization, particularly for Transformers, towards efficient and ubiquitous AI system across both edge (e.g., AR/VR) and cloud (e.g., GenAI)**, addressing
  (1) operator-level efficiency through techniques including ShiftAddNet [1,3,7,12];-
  (2) module-level scalability with approaches including Linearized-Transformer [2,4];
  (3) model-level compactness via methods including drawing Early-Bird Tickets [6,9,11,13]; and
  (4) algorithm-hardware co-design as integrated systems exemplified by ViTCoD [5] and EyeCoD [8].

PUBLICATIONS

[1] **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", in *38th Annual Conference on Neural Information Processing Systems* (**NeurIPS 2024**). [Paper], [Code].

[2] **H. You**, Y. Fu, Z. Wang, A. Yazdanbakhsh, Y. Lin. “When Linear Attention Meets Autoregressive Decoding: Towards More Effective and Efficient Linearized Large Language Models”, in *41th International Conference on Machine Learning* (**ICML 2024**). [Paper], [Code], [Poster].

[3] **H. You***, H. Shi*, Y. Guo*, Y. Lin. “ShiftAddViT: Mixture of Multiplication Primitives Towards Efficient Vision Transformer”, in *37th Annual Conference on Neural Information Processing Systems* (**NeurIPS 2023**). [Paper], [Code], [Slide], [Poster], [Talk].

[4] **H. You***, Y. Xiong*, X. Dai, B. Wu, P. Zhang, H. Fan, P. Vajda, Y. Lin. “Castling-ViT: Compressing Attention via Switching Towards Linear-Angular Attention at ViT Inference”. in *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (**CVPR 2023**). [Paper], [Code], [Slide], [Poster], [Talk].

[5] **H. You**, Z. Sun, H. Shi, Z. Yu, Y. Zhao, Y. Zhang, C. Li, B. Li, Y. Lin. “ViTCoD: Vision Transformer Acceleration via Dedicated Algorithm and Accelerator Co-Design”, in *29th IEEE International Symposium on High-Performance Computer Architecture* (**HPCA 2023**). [Paper], [Code], [Slide], [Poster], [Talk@GT], [Talk@HPCA]. **Selected as the Meta Faculty Research Award of 2022!**

[6] **H. You**, B. Li, Z. Sun, X. Ouyang, Y. Lin. “SuperTickets: Drawing Task-Agnostic Lottery Tickets from Supernets via Jointly Searching and Pruning”, in *European Conference on Computer Vision* (**ECCV 2022**). [Paper], [Code], [Slide], [Poster], [Talk].

[7] **H. You**, B. Li, H. Shi, Y. Lin. “ShiftAddNAS: Hardware-Inspired Search for More Accurate and Efficient Neural Networks”, in *39th International Conference on Machine Learning* (**ICML 2022**). [Paper], [Code], [Slide], [Talk].

[8] **H. You***, Y. Zhao*, Z. Yu*, C. Wan*, Y. Fu, C. Li, S. Zhang, S. Wu, J. Yuan, Y. Zhang, Vivek B., Ashok V., Z. Li, Y. Lin. “EyeCoD: Eye Tracking System Acceleration via FlatCam-Based Algorithm and Accelerator Co-Design”, in *49th Annual International Symposium on Computer Architecture* (**ISCA 2022**). [Paper], [Slide], [Poster], [Talk], [TopPick’23]. **Selected as the IEEE Micro’s Top Pick of 2023!**

[9] **H. You**, Z. Lu, Z. Zhou, Y. Fu, Y. Lin. “Early-Bird GCNs: Graph-Network Co-Optimization Towards More Efficient GCN Training and Inference via Drawing Early-Bird Lottery Tickets”. in *36th AAAI Conference on Artificial Intelligence* (**AAAI 2022**). [Paper], [Code], [Slide], [Poster], [Talk].

[10] **H. You**, T. Geng, Y. Zhang, A. Li, Y. Lin. “GCoD: Graph Convolutional Network Accelerationvia Dedicated Algorithm and Accelerator Co-Design”, in *28th IEEE International Symposium on High-Performance Computer Architecture* (**HPCA 2022**). [Paper], [Code], [Slide], [Talk].

[11] **H. You***, R. Balestriero*, Z. Lu, Y. Kou, Y. Lin, R.G. Baraniuk. “Max-Affine Spline Insights Into Deep Network Pruning”, in *Transactions on Machine Learning Research* (**TMLR 2022**). [Paper], [Code].

[12] **H. You**, X. Chen, Y. Zhang, C. Li, S. Li, Z. Liu, Z. Wang, Y. Lin. “ShiftAddNet: A Hardware-Inspired Deep Network”, in *34th Annual Conference on Neural Information Processing Systems* (**NeurIPS 2020**). [Paper], [Code], [Slide], [Poster], [Talk@RICE], [Talk@NeurIPS].

[13] **H. You**, C. Li, P. Xu, Y. Fu, Y. Wang, X. Chen, R.G. Baraniuk, Z. Wang, Y. Lin. “Drawing Early-Bird Tickets: Towards More Efficient Training of Deep Networks”, in *8th International Conference on Learning Representations* (**ICLR 2020**). [Paper], [Code], [Slide], [OpenReview], [Talk]. **Selected as the Spotlight Oral Paper (4%)!**

[14] **H. You**, Y. Cheng, T. Cheng, C. Li, P. Zhou. “Bayesian Cycle-Consistent Generative Adversarial Networks via Marginalizing Latent Sampling”, in *IEEE Transactions on Neural Networks and Learning Systems* (**TNNLS 2020**). [Paper], [Code].

[15] Z. Yu, Z. Wang, Y. Li, **H. You**, R. Gao, X. Zhou, S. Bommu, Y. Zhao, Y. Lin. “EDGE-LLM: Enabling

Efficient LLM Adaptation on Edge Devices via Layerwise Unified Compression and Adaptive Layer Tuning and Voting", in *61th Design Automation Conference* (**DAC 2024**).

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

[17] Y. Fu, Z. Ye, J. Yuan, S. Zhang, S. Li, **H. You**, Y. Lin. "Gen-NeRF: Efficient and Generalizable Neural Radiance Fields via Algorithm-Hardware Co-Design", in *50th International Symposium on Computer Architecture* (**ISCA 2023**).

[18] S. Li, C. Li, W. Zhu, B. Yu, Y. Zhao, C. Wan, **H. You**, H. Shi, Y. Lin. "Instant-3D: Instant Neural Radiance Fields Training Towards Real-Time AR/VR 3D Reconstruction", in *50th International Symposium on Computer Architecture* (**ISCA 2023**).

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

[20] X. Chen, Y. Zhao, Y. Wang, P. Xu, **H. You**, C. Li, Y. Fu, Y. Lin, Z. Wang. "SmartDeal: Re-Modeling Deep Network Weights for Efficient Inference and Training", in *IEEE Transactions on Neural Networks and Learning Systems* (**TNNLS 2021**).

[21] Y. Zhao, Z. Li, Y. Fu, Y. Zhang, C. Li, C. Wan, **H. You**, S. Wu, X. Ouyang, V. Boominathan, A. Veeraraghavan, Y. Lin. "i-FlatCam: A 253 FPS, 91.49 µJ/Frame Ultra-Compact Intelligent Lensless Camera Eye Tracking System", in *IEEE Symposium on VLSI Technology & Circuits* (**VLSI 2022**). **Won First Place in University Best Demonstration at DAC 2022!**

[22] Z. Yu, Y. Fu, S. Wu, M. Li, **H. You**, Y. Lin. "LDP: Learnable Dynamic Precision for Efficient Deep Neural Network Training and Inference", in *tinyML Research Symposium* (**TinyML 2022**).

[23] T. Geng, C. Wu, Y. Zhang, C. Tan, C. Xie, **H. You**, M. Herbordt, Y. Lin, A. Li. "I-GCN: A GCN Accelerator with Runtime Locality Enhancement through Islandization", in *54th IEEE/ACM International Symposium on Microarchitecture* (**MICRO 2021**).

[24] C. Li, Z. Yu, Y. Fu, Y. Zhang, Y. Zhao, **H. You**, Q. Yu, Y. Wang, C. Hao, Y. Lin. "HW-NAS-Bench: Hardware-Aware Neural Architecture Search Benchmark", in *9th International Conference on Learning Representations* (**ICLR 2021**). **Selected as the Spotlight Oral Paper (4%)!**

[25] Y. Zhang, **H. You**, Y. Fu, T. Geng, A. Li, Y. Lin. "G-CoS: GNN-Accelerator Co-Search Towards Both Better Accuracy and Efficiency", in *International Conference on Computer-Aided Design* (**ICCAD 2021**).

[26] Y. Zhang, Y. Fu, W. Jiang, C. Li, **H. You**, M. Li, V. Chandra, Y. Lin. "DIAN: differentiable accelerator-network co-search towards maximal DNN efficiency", in *ACM/IEEE International Symposium on Low Power Electronics and Design* (**ISLPED 2021**).

[27] Y. Fu, **H. You**, Y. Zhao, Y. Wang, C. Li, K. Gopalakrishnan, Z. Wang, Y. Lin. "FracTrain: Fractionally Squeezing Bits Both Temporally and Spatially for Efficient DNN Training", in *34th Annual Conference on Neural Information Processing Systems* (**NeurIPS 2020**).

[28] C. Li, T. Chen, **H. You**, Z. Wang, Y. Lin. "HALO: Hardware-Aware Learning to Optimize", in *European Conference on Computer Vision* (**ECCV 2020**).

[29] Y. Zhao, X. Chen, Y. Wang, C. Li, **H. You**, Y. Fu, Y. Xie, Z. Wang, Y. Lin. "SmartExchange: Trading Higher-cost Memory Storage/Access for Lower-cost Computation", in *47th International Symposium on Computer Architecture* (**ISCA 2020**).

SELECTED AWARDS

- Haoran received the NeurIPS 2023 Scholar Award ! Dec. 2023
- Haoran won First Place in ACM Student Research Competition (SRC) at ICCAD ! Nov. 2023
- ICCAD 2023 Student Travel Grant ! Oct. 2023
- Haoran was selected as one of the ML and Systems Rising Stars of 2023 ! May. 2023
- Haoran won the Best Poster Award at the SCS Poster Competition of GaTech ! Apr. 2023
- Haoran was selected by GT to receive the Outstanding Graduate Research Assistant Award ! Mar. 2023
- EyeCoD (ISCA'22) was selected for being the IEEE Micro's Top Pick of 2023 ! Jan. 2023
- HPCA 2023 Student Travel Grant ! 2022 - 2023
- Haoran got PhD Growth Annual Rock Star Award in EIC Lab ! 2022 - 2023
- ViTCoD (HPCA'23) was selected for Meta Faculty Research Award of 2022 ! Nov. 2022
- Our i-FlatCam won First Place in University Demo Best Demonstration at DAC Conference ! Jul. 2022
- ISCA 2022 Student Trave Grant ! Jun. 2022
- Samsung Scholarship ! 2017 - 2018
- International Interdisciplinary Contest in Modeling (ICM), Meritorious Winner ! Apr. 2018
- Selected as *Outstanding Undergraduates in Term of Academic Performance*, (HUST 1%) ! Dec. 2017
- The Eighth Chinese Mathematics Competition (CMC), First Price (China 5%) ! Nov. 2016

INVITED TALKS

- Invited Speaker at Georgia Tech CS 4803/8803 Efficient ML Course, 2024. [Link]
- Invited Speaker at Georgia Tech ECE 8803 SW/HW Co-Design Course, 2024. [Link]
- Invited Speaker at Google, 2023. [Link]
- Invited Speaker at University of Rochester IntelliArch Lab, 2023. [Link]
- Invited Speaker at ISCA AutoDL Workshop, 2023. [Link]
- Invited Speaker at UT Austin Energy-Aware Computing Group, 2023. [Link]
- Invited Speaker at the Tutorial in Asian Conference on Computer Vision, 2022. [Link]
- Invited Speaker at the Rice University Ken Kennedy Institute Data Science Conference, 2021. [Link]

MENTORING EXPERIENCE

- Zhanyi Sun (B.S. Student), Published two 2nd author HPCA'23/ECCV'22 papers, Next move: MS@CMU
- Zhihan Lu (B.S. Student), Published two 2nd author AAAI'22/TMLR papers, Next move: MS@CMU
- Shang Wu (M.S. Student), Published co-authored ISCA'22/TMLR papers, Next move: PhD@Northwestern
- Xu Ouyang (M.S. Student), Published one co-authored ECCV'22 paper, Next move: PhD@Virginia
- Zijian Zhou (B.S. Student), Published one co-authored AAAI'22 paper, Next move: MS@Rice
- Yichao Fu (M.S. Student), Published one 2nd authored ICML'24 paper, Next move: PhD@UCSD
- Yi Sun (B.S. Student), Submitted a co-authored paper, Next move: MS@CMU
- Yutong Kou (B.S. Student), Published a co-authored TMLR paper, Next move: PhD@CAS
- Yipin Guo (M.S. Student), Published one co-1st author NeurIPS'23 paper, Current position: MS@ZJU
- Wei Zhou (B.S. Student), Submitted a co-authored paper, Next move: MS@GaTech

PROPOSAL EXPERIENCE

- Georgia Tech's Tech Ready Grant from the Office of Technology Licensing. [News]
  - Leading PI; Writing draft; Key preliminary results: EyeCoD [8]

TEACHING EXPERIENCE

- TA at CS 4803/8803: Efficient Machine Learning, Georgia Tech Spring 2024
- TA at ELEC 220: Fundamentals of Computer Engineering, Rice University Fall 2021
- TA at ELEC 220: Fundamentals of Computer Engineering, Rice University Spring 2021
- TA at ELEC 220: Fundamentals of Computer Engineering, Rice University Fall 2020

CONFERENCE PRESENTATION

- The 41th International Conference on Machine Learning (ICML), 2024. [Link]
- The Second CoCoSys Annual Review, 2024. [Link]
- The 37th Conference on Neural Information Processing Systems (NeurIPS), 2023. [Link]
- Semiconductor Research Corporation (SRC) TECHCON, 2023. [Link]
- The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023. [Link]
- The 50th International Symposium on Computer Architecture (ISCA), 2023. [Link]
- The 29th International Symposium on High-Performance Computer Architecture (HPCA), 2023. [Link]
- European Conference on Computer Vision (ECCV), 2022. [Link]
- The 49th International Symposium on Computer Architecture (ISCA), 2022. [Link]
- The 39th International Conference on Machine Learning (ICML), 2022. [Link]
- The 28th International Symposium on High-Performance Computer Architecture (HPCA), 2022. [Link]
- The 36th AAAI Conference on Artificial Intelligence (AAAI), 2022. [Link]
- The 34th Conference on Neural Information Processing Systems (NeurIPS), 2020. [Link]
- The 8th International Conference on Learning Representations (ICLR), 2020. [Link]

SERVICES

- **Conference Reviewer**: ICML, NeurIPS, ICLR, MLSys, CVPR, ICCV, ECCV, AAAI
- **Journal Reviewer**: TPAMI, TNNLS, TCAS-II, JSTC, IMAVIS
- **Technical Program Committee**: DATE
- **Volunteer & Website Developer**: ICCAD HALO Workshop, EECS Rising Stars Workshop of 2023