Experience

**NVIDIA Research**
- **Staff Research Scientist, Learning and Perception Research (LPR)**  
  **Apr. 2024 - Now**

**NVIDIA Research**
- **Senior Research Scientist, Learning and Perception Research (LPR)**  
  **May 2022 - Mar. 2024**

**NVIDIA Research**
- **Research Scientist, Learning and Perception Research (LPR)**  
  **May 2020 - Apr. 2022**

**NVIDIA Research**
- **Research Intern, Learning and Perception Research (LPR)**  
  **May 2019 - Nov. 2019**

**Alibaba U.S.**
- **Research Intern, Machine Learning Team**  
  **May 2018 - Nov. 2018**

Education

**Princeton University**
- Ph.D. in Electrical Computer Engineering, advised by Prof. Niraj K. Jha  
  **New Jersey, USA**  
  **2015 - 2020**
- Research focus: Efficient and Secure Deep Learning

**Nanyang Technological University**
- B.Eng in Electronic & Electronics Engineering (GPA 3.9/4.0, dean's lister all four years)  
  **Singapore, SG**  
  **2011 - 2015**
- Minor in Business (GPA 4.0/4.0)

**University of California, Berkeley**
- Undergraduate summer exchange  
  **California, USA**  
  **2012**

**University of Cambridge**
- High school elite exchange program  
  **Cambridge, UK**  
  **2007**

Selected Awards

- 36 Kr Top 100 Global Outstanding Chinese Awards  
  **2022**
- Forbes Top 60 Elite Chinese North America  
  **2021**
- Princeton ECE Best Dissertation Award Finalist (Top-3 in department)  
  **2020**
- Princeton Yan Huo *94 Fellowship (Top-3 in department)  
  **2019**
- Princeton Natural Science and Foundation Fellowship  
  **2015-2017**
- Gold Medal - Defense Science and Technology  
  **2015**
- Gold Medal - Thomas Asia Pacific Holdings  
  **2015**
- Department Dean's Lister Award  
  **2011-2015**
- Nanyang Best Industrial Orientation Award  
  **2014**
- Nanyang Presidential Scholar with Highest Distinction  
  **2012-2015**

Conference Publications

(*: equal contribution; †: advised intern)

49. Gongfan Fang†, **Hongxu Yin**, Saurav Muralidharan, Greg Heinrich, Jeff Pool, Jan Kautz, Pavlo Molchanov, Xinshao Wang  
   *MaskLLM: Learnable semi-structured sparsity for large language models*  
   Advances in Neural Information Processing Systems (NeurIPS), 2024  
   *(Spotlight Paper)*

48. An-Chieh Cheng†, **Hongxu Yin**, Yang Fu, Qiushan Guo, Ruihan Yang, Jan Kautz, Xiaolong Wang, Sifei Liu  
   *SpatialRGPT: Grounded spatial reasoning in vision language model*  
   Advances in Neural Information Processing Systems (NeurIPS), 2024
47. Ji Lin†, Hongxu Yin*, Wei Ping, Yao Lu, Pavlo Molchanov, Andrew Tao, Huizi Mao, Jan Kautz, Mohammad Shoeybi, Song Han
   VILA: On pre-training for visual language models
   Conference on Computer Vision and Pattern Recognition (CVPR), 2024

46. Qiushan Guo†, Shalini De Mello*, Hongxu Yin*, Wonmin Byeon, Ka Chun Cheung, Yizhou Yu, Ping Luo, Sifei Liu
   RegionGPT: Towards region understanding vision language model
   Conference on Computer Vision and Pattern Recognition (CVPR), 2024

45. Shih-Yang Liu‡, Chien-Yi Wang, Hongxu Yin, Pavlo Molchanov, Yu-Chiang Frank Wang, Kwang-Ting Cheng, Min-Hung Chen
   DoRA: Weight-decomposed low-rank adaptation
   International Conference on Machine Learning (ICML), 2024
   (Oral Presentation - top 1.5% paper)

44. Jingwen Sun†, Ziyue Xu, Hongxu Yin, Dong Yang, Daguang Xu, Yiran Chen, Holger R. Roth
   FedBPT: Efficient federated black-box prompt tuning for large language models
   International Conference on Machine Learning (ICML), 2024
   AAAI Symposium, 2024 (Best Paper Award)

43. Ruisi Cai‡, Saurav Muralidharan, Greg Henrich, Hongxu Yin, Zhangyang Wang, Jan Kaut, Pavlo Molchanov
   FlexTron: Many-in-One flexible large language models
   International Conference on Machine Learning (ICML), 2024
   (Oral Presentation - top 1.5% paper)

42. De-an Huang, Shijia Liao, Subhashree Radhakrishnan, Hongxu Yin, Pavlo Molchanov, Zhiding Yu, Jan Kautz
   LIITA: Language instructed temporal-localization assistant
   European Conference on Computer Vision (ECCV), 2024

41. Anna Bair†, Hongxu Yin, Maying Shen, Pavlo Molchanov, Jose M. Alvarez
   Adaptive sharpness-aware pruning for robust sparse networks
   International Conference on Learning Representations (ICLR), 2024

40. Ali Hatamizadeh, Greg Heinrich, Hongxu Yin, Andrew Tao, Jose M. Alvarez, Jan Kautz, Pavlo Molchanov
   FasterViT: Fast vision transformers with hierarchical attention
   International Conference on Machine Learning (ICML), 2024

39. Xinlong Sun†, Maying Shen, Hongxu Yin, Lei Mao, Pavlo Molchanov, Jose M Alvarez
   Advancing weight and channel sparsification with enhanced saliency
   Winter Conference on Applications of Computer Vision (WACV), 2024

38. Jiaming Song, Qinsheng Zhang, Hongxu Yin, Morteza Mardani, Ming-yu Liu, Jan Kautz, Yongxin Chen, Arash Vahdat
   Loss-guided diffusion models for Plug-and-Play controllable generation
   International Conference on Machine Learning (ICML), 2023

37. Ali Hatamizadeh, Hongxu Yin, Jan Kautz, Pavlo Molchanov
   Global context vision transformer
   International Conference on Machine Learning (ICML), 2023

36. Divyam Madaan‡, Hongxu Yin, Wonmin Byeon, Jan Kautz, Pavlo Molchanov
   Heterogeneous continual learning
   Conference on Computer Vision and Pattern Recognition (CVPR), 2023
   (Highlight Paper - top 2.5% paper)

35. Huanrui Yang†, Hongxu Yin, Pavlo Molchanov, Hai Li, Jan Kautz
   NViT: Vision transformer compression and parameter redistribution
   Conference on Computer Vision and Pattern Recognition (CVPR), 2023

34. Paul Micaelli†, Pavlo Molchanov, Arash Vahdat, Hongxu Yin, Jan Kautz
   Recurrence without recurrence: stable video landmark detection with deep equilibrium models
   Conference on Computer Vision and Pattern Recognition (CVPR), 2023

33. Xin Dong†, Hongxu Yin, Jose Alvarez, Jan Kautz, Pavlo Molchanov
   Privacy vulnerability of split computing to data-free model inversion attacks
   British Machine Vision Conference (BMVC), 2022
32. Maying Shen*, Hongxu Yin*, Pavlo Molchanov, Lei Mao, Jianna Liu, Jose Alvarez
   *Structural pruning via latency-saliency Knapsack*
   Advances in Neural Information Processing Systems (NeurIPS), 2022

31. Hongxu Yin, Arash Vahdat, Jose Alvarez, Arun Mallya, Jan Kautz, Pavlo Molchanov
   A-ViT: Adaptive tokens for efficient vision transformer
   Conference on Computer Vision and Pattern Recognition (CVPR), 2022
   (Oral Presentation)

30. Ali Hatamizadeh*, Hongxu Yin*, Holger Roth, Wenqi Li, Jan Kautz, Daguang Xu, Pavlo Molchanov
   GradViT: Gradient inversion of vision transformers
   Conference on Computer Vision and Pattern Recognition (CVPR), 2022

29. Maying Shen, Pavlo Molchanov, Hongxu Yin, Jose Alvarez
   When to prune? A policy towards early structural pruning
   Conference on Computer Vision and Pattern Recognition (CVPR), 2022

28. Pavlo Molchanov*, Jimmy Hall*, Hongxu Yin*, Jan Kautz, Nicolo Fusi, Arash Vahdat
   HANT: Hardware-aware network transformation
   European Conference on Computer Vision (ECCV), 2022

27. Hongxu Yin, Arun Mallya, Arash Vahdat, Jose Alvarez, Jan Kautz, Pavlo Molchanov
   See through gradients: Image batch recovery via GradInversion
   Conference on Computer Vision and Pattern Recognition (CVPR), 2021

26. Yerlan Idelbayev†, Pavlo Molchanov, Maying Shen, Hongxu Yin, M. C. Perpinan, Jose Alvarez
   Optimal quantization using scaled codebook
   Conference on Computer Vision and Pattern Recognition (CVPR), 2021

25. Akshay Chawla†, Hongxu Yin, Pavlo Molchanov, Jose Alvarez
   Data-free knowledge distillation for object detection
   Winter Conference on Applications of Computer Vision (WACV), 2021

24. Hongxu Yin, Arun Mallya, Arash Vahdat, Jose Alvarez, Jan Kautz, Pavlo Molchanov
   Dreaming to distill: Data-free knowledge transfer via DeepInversion
   Conference on Computer Vision and Pattern Recognition (CVPR), 2020
   (Oral Presentation)

23. Wenhan Xia, Hongxu Yin, Niraj K. Jha
   Efficient synthesis of compact deep neural networks
   IEEE Design Automation Conference (DAC), 2020

22. Xiaoliang Dai, Peizhao Zhang, Bichen Wu, Hongxu Yin, Fei Sun, Yanghan Wang, Marat Dukhan, Yunqing Hu,
   Yiming Wu, Yangqing Jia, Peter Vajda, Matt Uyttendaele, Niraj K. Jha
   ChamNet: Towards efficient network design through platform-aware model adaptation
   Conference on Computer Vision and Pattern Recognition (CVPR), 2019

21. Ozge Akmandor, Hongxu Yin, and Niraj K. Jha
   Simultaneously ensuring smartness, security, and energy efficiency in Internet-of-Things sensors
   IEEE Custom Integrated Circuits Conference (CICC), 2017

20. Hongxu Yin, Bah Hwee Gwee, Zhiping Lin, Kumar Anil, Galul R. Sirajudeen, and Choo M. S. See
   Novel real-time system design for floating-point sub-Nyquist multi-coset signal blind reconstruction
   IEEE Int. Symp. on Circuits and Systems (ISCAS), 2015
   (Oral Presentation)

Journal Publications

19. Ali Hatamizadeh, Hongxu Yin, Pavlo Molchanov, Andriy Myronenko, Wenqi Li, Prerna Dogra, Andrew Feng,
   Mona G Flores, Jan Kautz, Daguang Xu, Holger R. Roth
   Do gradient inversion attacks make federated learning unsafe?
   IEEE Transactions on Medical Imaging, 2023

18. Shayan Hassantabar, Joe Zhang, Hongxu Yin, Niraj K. Jha
   MHDeep: Mental health disorder detection system based on body-area and deep neural networks
   ACM Transactions on Embedded Computing Systems, 2022
17. Hongxu Yin, Guoyang Chen, Yingmin Li, Shuai Che, Weifeng Zhang, and Niraj K. Jha
Hardware-guided symbiotic training for compact, accurate, yet execution-efficient LSTMs
IEEE Trans. Emerging Topics in Computing, 2021

16. Wenhan Xia, Hongxu Yin, Xiaoliang Dai, Niraj K. Jha
Fully dynamic inference with deep neural networks
IEEE Trans. Emerging Topics in Computing, 2021

15. Xiaoliang Dai*, Hongxu Yin*, and Niraj K. Jha
Grow and prune compact, fast, and accurate LSTMs
IEEE Trans. Computers, 2020

14. Hongxu Yin, Bilal Mukadam, Xiaoliang Dai, and Niraj K. Jha
DiabDeep: Pervasive diabetes diagnosis based on wearable medical sensors and efficient neural networks
IEEE Trans. Emerging Topics in Computing, 2020

13. Xiaoliang Dai, Hongxu Yin, and Niraj K. Jha
Incremental learning using a grow-and-prune paradigm with efficient neural networks
IEEE Trans. Computers, 2020

12. Xiaoliang Dai, Hongxu Yin, and Niraj K. Jha
NeST: A neural network synthesis tool based on a grow-and-prune paradigm
IEEE Trans. Computers, 2019

11. Hongxu Yin, Zeyu Wang, and Niraj K. Jha
A hierarchical inference model for Internet-of-Things
IEEE Trans. Multi-scale Computing Systems, 2018

10. Hongxu Yin and Niraj K. Jha
A health decision support system for disease diagnosis based on wearable medical sensors and machine learning ensembles
IEEE Trans. Multi-scale Computing Systems, 2017

9. Ozge Akmandor, Hongxu Yin and Niraj K. Jha
Smart, secure, yet energy-efficient, Internet-of-Things sensors
IEEE Trans. Multi-scale Computing Systems, 2017

Book Chapter

8. Hongxu Yin, Ozge Akmandor, Arsalan Mosenia, and Niraj K. Jha
Smart healthcare
Foundations and Trends, 2017

Preprint (publicly available & under review)

7. Hanrong Ye†, De-An Huang, Yao Lu, Zhiding Yu, Wei Ping, Andrew Tao, Jan Kautz, Song Han, Dan Xu, Pavlo Molchanov, Hongxu Yin
X-VILA: Cross-modality alignment for large language model
preprint, 2024

6. Yunhao Fang†*, Ligeng Zhu*, Yao Lu, Yan Wang, Pavlo Molchanov, Jang Hyun Cho, Marco Pavone, Song Han, Hongxu Yin
VILA*: VILA augmented VILA
preprint, 2024

5. Yecheng Wu, Zhuoyang Zhang, Junyu Chen, Haotian Tang, Dacheng Li, Yunhao Fang, Ligeng Zhu, Enze Xie, Hongxu Yin, Li Yi, Song Han, Yao Lu
VILA-U: a unified foundation model integrating visual understanding and generation
preprint, 2024

4. Fuzhao Xue†*, Yukang Chen†*, Dacheng Li†*, Qinghao Hu†*, Ligeng Zhu, Xiuyu Li, Yunhao Fang, Haotian Tang, Shang Yang, Zhijian Liu, Ethan He, Hongxu Yin, Pavlo Molchanov, Jan Kautz, Linxi Fan, Yuke Zhu, Yao Lu, Song Han
LongVILA: Scaling long-context visual language models for long videos
preprint, 2024

3. Min Shi, Fuxiao Liu, Shihao Wang, Shijia Liao, Subhashree Radhakrishnan, De-An Huang, Hongxu Yin, Karan Sapra, Yaser Yacoob, Humphrey Shi, Bryan Catanzaro, Andrew Tao, Jan Kautz, Zhiding Yu, Guilin Liu
EAGLE: Exploring the design space for multimodal LLMs with mixture of encoders
preprint, 2024

2. Yazhou Xing†, Amrita Mazumdar, Anjul Patney, Chao Liu, Hongxu Yin, Qifeng Chen, Jan Kautz, Iuri Frosio
Online overexposed pixels hallucination in videos with adaptive reference frame selection
preprint, 2023

1. Zhen Dong†, Hongxu Yin, Arash Vahdat, Jan Kautz, Pavlo Molchanov
Efficient transformation of architectures through hardware-aware nonlinear optimization
preprint, 2022

Workshop & Tutorial Organizer

- Efficient Deep Learning for Foundation Models Workshop
  ECCV 2024

- Efficient Computer Vision Workshop
  CVPR 2024

- Full-Stack, GPU-based Acceleration of Deep Learning Tutorial
  CVPR 2024

- Data-efficient Learning for Large Model Tutorial
  ICCV 2023

- Full-Stack, GPU-based Acceleration of Deep Learning Tutorial
  CVPR 2023

- Transformers for Vision Workshop
  CVPR 2022

Invited Keynote & Talk (till Dec. 2022)

- Efficient Deep Learning
  Invited Panelist, Open Compute Project (OCP) Global Summit
  Oct. 2022

- Towards Efficient and Secure Deep Learning
  Invited Keynote, Design & Automation Conference (DAC’60)
  Jul. 2022

- Towards Efficient and Secure Deep Nets
  University of British Columbia ECE Department
  May 2022

- Inverting Deep Nets
  Princeton University, Department of Computer Science research groups
  Aug. 2021

- See through Gradients
  Europe ML meeting
  Apr. 2021

- Dreaming to Distill
  Synced AI (largest AI media in Asia)
  Jul. 2020

- Dreaming to Distill
  Facebook AR/VR
  Jun. 2020

- Making Neural Networks Efficient
  Alibaba Cloud / Platform AI group
  Feb. 2020

- Efficient Neural Networks
  NVIDIA Research, Facebook Research
  Dec. 2019

- Efficient Neural Networks
  Baidu Research, ByteDance A.I. Lab US
  Dec. 2019

- Efficient Neural Networks
  Alibaba A.I. Research, Kwai Lab
  Nov. 2019

- Applied Machine Learning: From Theory to Practice
  Invited Keynote, IEEE Circuits and Systems Society (Singapore Chapter)
  Feb. 2018

- A Health Decision Support System for Disease Diagnosis
  New Jersey Tech Council
  Jun. 2016
| Patent Number | Title                                                                 | Year |
|---------------|----------------------------------------------------------------------|------|
| 25            | Pruning Neural Networks                                              | 2022 |
| 24            | Neural Network Training Technique                                     | 2022 |
| 23            | Techniques to Identify Data used to Train One or More Neural Networks | 2022 |
| 22            | Pruning Vision Transformers under Latency Budget and a Method to Distribute Parameters across Layers | 2022 |
| 21            | GradViT: Gradient Inversion of Vision Transformers                    | 2022 |
| 20            | Adaptive Token Depth Adjustment Algorithm for Networks with Transformer Blocks | 2022 |
| 19            | Global Context Model for Transformer Neural Networks                  | 2022 |
| 18            | Towards Understanding the Risks of Gradient Inversion in Federated Learning | 2022 |
| 17            | When to Prune? A Policy for Early Structural Pruning                  | 2021 |
| 16            | See Through Gradients: Image Batch Recovery via GradInversion         | 2021 |
| 15            | Network similarity metric as a Pruning Indicator                      | 2021 |
| 14            | Zero-shot Model Inversion for Data-free Distillation                  | 2021 |
| 13            | MHDeep: Mental Health Disorder Detection System based on Body-Area and Deep Neural Networks | 2019 |
| 12            | Optimal MSE Quantization with Fixed Codebook and Rescaling           | 2020 |
| 11            | Dreaming Data for Continual Learning                                 | 2020 |
| 10            | Data-Free Knowledge Distillation for Object Detection                 | 2020 |
| 9             | Hardware-aware Latency Neural Network Pruning                         | 2020 |
| 8             | Image Generation for Data Free Pruning                               | 2019 |
| 7             | Hardware-guided Symbiotic Training for Compact, Accurate, yet Execution-efficient LSTMs | 2019 |
| 6             | Incremental Learning using a Grow-and-prune Paradigm with Efficient Neural Networks | 2019 |
| 5             | DiabDeep: Persasive Diabetes Diagnosis based on Wearable Medical Sensors and Efficient Neural Networks | 2019 |
| 4             | Smart, Secure, yet Energy-efficient Internet-of-Things Sensors        | 2019 |
| 3             | NeST: A Neural Network Synthesis Tool based on a Grow-and-prune Paradigm | 2018 |
| 2             | Grow and Prune Compact, Fast, yet Accurate LSTMs                      | 2018 |
1. A Hierarchical Health Decision support System based on Wearable Medical Sensors and Machine Learning Ensembles
Princeton University

Academic Services

Teaching Assistant - Princeton University
ELE 364, Machine Learning for Predictive Data Analytics
ELE464, Embedded Computing

Conference Reviewer & Committee
Computer Vision and Pattern Recognition (CVPR)
Conference on Neural Information Processing Systems (NeurIPS)
International Conference on Learning Representations (ICLR)
International Conference on Machine Learning (ICML)
International Conference on Computer Vision (ICCV)
European Conference on Computer Vision (ECCV)
British Machine Vision Conference (BMVC)
Winter Conference on Applications of Computer Vision (WACV)
AAAI Conference on Artificial Intelligence (AAAI)
Design Automation Conference (DAC)
High-Performance Computer Architecture (HPCA)

Journal Reviewer & Committee
IEEE Transactions on Pattern Analysis and Machine Intelligence
IEEE Transactions on Neural Networks and Learning Systems
International Journal of Computer Vision
IEEE Journal of Biomedical and Health Informatics
IEEE Journal of Selected Topics in Signal Processing
IEEE Sensors Journal
IEEE Consumer Electronics Magazine
International Journal on Artificial Intelligence Tools
International Journal of Systems Architecture
International Journal of Healthcare Technology and Management
International Journal of Electronic Imaging

Mentorship

NVIDIA Research Mentees
Baifeng Shi, University of California, Berkeley
Hanrong Ye, Hong Kong University of Science and Technology
Ji Lin, Massachusetts Institute of Technology
Huanrui Yang, Duke University
Zhen Dong, University of California, Berkeley
Xin Dong, Harvard University
Annamarie Bair, Carnegie Mellon University
Divyam Madaan, New York University
Paul Micaelli, University of Edingburgh
Yerlan Idelbayev, University of California, Merced
Vu Nguyen, Stony Brooks University
Akshay Chawla, Carnegie Mellon University

Princeton Senior Thesis Mentees
Joe Zhang, now Ph.D. at Stanford
Hari Santhanam, now Ph.D. at University of Pennsylvania
Frederick Hertan, now at SIG Trading
Kyle Johnson, now at Princeton University
Bilal Mukadam, now at Microsoft
Chloe Song, now at Astra Inc.