CURRICULUM VITAE

Sean Xuefeng Du
Assistant Professor
College of Computing and Data Science
Nanyang Technological University - Singapore

Address: 50 Nanyang Ave, Block N 3, Singapore 639798
Website: Homepage, Google Scholar (Citation $\geq$ 3,300 as of Feb. 2026)
Email: xuefeng.du@ntu.edu.sg, **Phone**: +65-89657205

Academic Qualifications

Ph.D. in Computer Sciences Jan. 2021 - May 2025
University of Wisconsin–Madison, Madison, WI

- Ph.D. research in reliable machine learning.
- Advisor: Prof. Sharon Y. Li.
- Dissertation: Foundations of Unknown-aware Machine Learning
- Thesis Committee Members: Profs. Jerry Zhu, Robert D. Nowak, Yong Jae Lee.

B. Eng. in Electrical Engineering Sept. 2016 – Jun. 2020
Xi'an Jiaotong University, Xi'an, China

- Overall GPA: 91.60/100(3.83/4.0), Rank: 1st/170.
- Honors: National Scholarships (2017, 2018).

Summary of Working Experience

Assistant Professor Sept. 2025 - Present
College of Computing and Data Science, Nanyang Technological University

Academic Honours and Awards

Year	Academic Honour / Award
2026	Recipient, AAAI-26 New Faculty Highlights
2025	Recipient, CS Ivanisevic Award in UW-Madison ($3,000, awarding to one student in the department for research excellence)
2024	Recipient, Rising Stars in Data Science (**worldwide 30 recipients**)
2023	Recipient, Jane Street Graduate Research Fellowship ($50,000 grant; **6 selected out of 600**)
2022	Recipient, NeurIPS Scholar Award
2021	Recipient, CS departmental research fellowship, UW-Madison
2018, 2017	Recipient, National Scholarship, Ministry of Education in China

RESEARCH SUMMARY

Key Areas of Research

Reliable machine learning, foundation model reliability, and related applications, specifically I work on algorithm and theory design for:

- *Algorithmic reliability*: designing adaptive and interpretable learning principles that help ML models detect and generalize on out-of-distribution (OOD) samples, such as semantic/covariate shifts, adversarial and noisy samples.
- *Language model reliability*: understanding the blindspots of Large Language Models (LLMs) and Multimodal LLMs through uncertainty estimation, such as model hallucination detection and mitigation, malicious prompt prevention, etc.
- *Interdisciplinary research*: Biometrics, AI-aided cryo-microscopy/protein structural analysis.

Research Awards / Recognition

Year	Research Award / Recognition
2025	Research on LLM safety is reported on Wisconsin Engineer Magazine
2024	NeurIPS Spotlight Presentation on paper *HaloScope: Harnessing Unlabeled LLM Generations for Hallucination Detection*.
2022	Research on out-of-distribution object detection is interviewed by Yannic Kilcher (over 313K subscribers) on Youtube: Youtube video, Author interview
2022	CVPR Oral Presentation on paper *Unknown-Aware Object Detection: Learning What You Don't Know from Videos in the Wild*
2019	BMVC Spotlight Presentation on paper *Open-set Recognition of Unseen Macromolecules in Cellular Electron Cryo-Tomograms by Soft Large Margin Centralized Cosine Loss*

Invited Presentations

- Talk at AAAI-26 new faculty highlights program, 1/2026
  Teach AI What It Doesn't Know
- Talk at Huawei Singapore Research Center, 11/2025
  Teach AI What It Doesn't Know
- Talk at University of Queensland, 6/2025
  Teach AI What It Doesn't Know
- Talk at MBZUAI, 5/2025
  Teach AI What It Doesn't Know
- Talk at Microsoft Research, 4/2025
  Teach AI What It Doesn't Know
- Talk at Pennsylvania State University, 3/2025
  Teach AI What It Doesn't Know
- Talk at University of Texas at Arlington, 3/2025
  Teach AI What It Doesn't Know
- Talk at Oregon State University, 2/2025
  Teach AI What It Doesn't Know
- Talk at UC Riverside, Statistics and CS departments, 2/2025 and 3/2025
  Teach AI What It Doesn't Know
- Talk at Rochester Institute of Technology, 2/2025
  Teach AI What It Doesn't Know
- Talk at Johns Hopkins University, 2/2025
  Teach AI What It Doesn't Know
- Talk at OpenAI, 1/2025
  Teach AI What It Doesn't Know
- Talk at NTU CCDS, 1/2025
  Teach AI What It Doesn't Know
- Talk at HKUST CSE, 12/2024
  Teaching AI What They Don't Know: Foundations and Algorithms for Reliable ML in the Open World
- Talk at MLOPT Seminar, UW-Madison, 12/2024
  HaloScope: Harnessing Unlabeled LLM Generations for Hallucination Detection
- Talk at UCSD, 11/2024
  HaloScope: Harnessing Unlabeled LLM Generations for Hallucination Detection
- Talk at Microsoft, 08/2024
  Defending Multimodal LLMs from Malicious Attacks
- Talk at Jane Street, 04/2023
  Unknown-aware Object Detection
- Talk at AI talks, 02/2023
  Unknown-aware Object Detection
- Talk at MLOPT Seminar, UW-Madison, 05/2022
  Unknown-aware Object Detection: Learning from Videos in the Wild

- Talk at Google Research, 01/2022
  Unknown-aware Object Detection: Learning from Videos in the Wild
- Talk at Adobe Research, 12/2021
  Unknown-aware Object Detection: Learning from Videos in the Wild
- Talk at Microsoft Research, 12/2021
  Unknown-aware Object Detection: Learning from Videos in the Wild

Research Funding

Internal Grants

Role	Year	Project Title	Amount (S$)	Source of Grant
PI	2025–2028	*Reliable Foundation Models: Theory and Algorithms for Detecting and Mitigating Hallucinations in LLMs and Multimodal Systems*	150,000	NTU Start-up grant
PI	2025–2028	*From Output to Process: Calibrated Confidence Estimation for Large AI Models*	100,000	MOE AcRF Tier 1 Seed

External Grants

Role	Year	Project Title	Amount (S$)	Source of Grant
PI	2025–2026	*Reliable Foundation Models*	6,450	Google Cloud Research Credits
PI	2026	*Detecting hallucinations in foundation models: Theory and algorithms*	7,400	NSCC Young Investigator Seed Project

Citation Summary

Database	Citation Count		H-index
	without self-citations	with self-citations	
Scopus	-	-	-
Web of Science (SCI)	-	-	-
Google Scholar	-	3428	26

Publications

See full list in my google scholar page, I publish under the name “Xuefeng Du”.
◇ indicates Equal Contribution.

Conference Papers

27. Limited Preference Data? Learning Better Reward Model with Latent Space Synthesis
    Leitian Tao, **Xuefeng Du**, Sharon Li.
    NeurIPS 2025.

26. How to Steer LLM Latents for Hallucination Detection?
    Seongheon Park, **Xuefeng Du**, Min-Hsuan Yeh, Haobo Wang, Yixuan Li.
    ICML 2025.

25. Understanding Multimodal LLMs Under Distribution Shifts: An Information-Theoretic Approach
    Changdae Oh, Zhen Fang, Shawn Im, **Xuefeng Du**, Yixuan Li.
    ICML 2025.

24. Challenges and Future Directions of Data-Centric AI Alignment
    Min-Hsuan Yeh, Jeffrey Wang, **Xuefeng Du**, Seongheon Park, Leitian Tao, Shawn Im, Yixuan Li.
    ICML 2025.

23. HaloScope: Harnessing Unlabeled LLM Generations for Hallucination Detection

Xuefeng Du, Chaowei Xiao, Yixuan Li.
NeurIPS 2024, **Spotlight**.

22. When and How does In-distribution Label Help Out-of-distribution Detection?
Xuefeng Du, Yiyou Sun, Yixuan Li.
International Conference on Machine Learning (ICML) 2024.

21. How does Unlabeled Data Provably Help Out-of-distribution Detection?
Xuefeng Du◇, Zhen Fang◇, Ilias Diakonikolas, Yixuan Li.
International Conference on Learning Representations (ICLR) 2024.

20. Feed Two Birds with One Scone: Exploiting Wild Data for Both Out-of-Distribution Generalization and Detection
Haoyue Bai, Gregory Canal, **Xuefeng Du**, Jeongyeol Kwon, Robert D Nowak, Yixuan Li.
International Conference on Machine Learning (ICML) 2023.

19. Dream the Impossible: Outlier Imagination with Diffusion Models
Xuefeng Du, Yiyou Sun, Jerry Zhu, Yixuan Li.
Advances in Neural Information Processing Systems (NeurIPS) 2023.

18. Non-parametric Outlier Synthesis
Leitian Tao, **Xuefeng Du**, Jerry Zhu, Yixuan Li.
International Conference on Learning Representations (ICLR) 2023.

17. SIREN: Shaping Representations for Detecting Out-of-distribution Objects
Xuefeng Du, Gabriel Gozum, Yifei Ming, Yixuan Li.
Advances in Neural Information Processing Systems (NeurIPS) 2022.

16. OpenOOD: Benchmarking Generalized Out-of-Distribution Detection
Jingkang Yang, Pengyun Wang, Dejian Zou, Zitang Zhou, Kunyuan Ding, Wenxuan Peng, Haoqi Wang, Guangyao Chen, Bo Li, Yiyou Sun, **Xuefeng Du**, Kaiyang Zhou, Wayne Zhang, Dan Hendrycks, Yixuan Li, Ziwei Liu.
Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track, 2022.

15. Unknown-Aware Object Detection: Learning What You Don't Know from Videos in the Wild
Xuefeng Du, Xin Wang, Gabriel Gozum, Yixuan Li.
IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) 2022, **Oral**.

14. Performance-Aware Mutual Knowledge Distillation for Improving Neural Architecture Search
Pengtao Xie, **Xuefeng Du**.
IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) 2022.

13. VOS: Learning What You Don't Know by Virtual Outliers Synthesis
Xuefeng Du, Eric Wang, Mu Cai, Yixuan Li.
International Conference on Learning Representations (ICLR) 2022.

12. Learning Diverse-Structured Networks for Adversarial Robustness
Xuefeng Du, Jingfeng Zhang, Bo Han, Tongliang Liu, Yu Rong, Gang Niu, Junzhou Huang, Masashi Sugiyama.
International Conference on Machine Learning (ICML) 2021.

11. How to save your annotation cost for Panoptic Segmentation?
Xuefeng Du, Chenhan Jiang, Hang Xu, Gengwei Zhang, Zhenguo Li.
The AAAI Conference on Artificial Intelligence (AAAI) 2021.

10. Node Classification on Graphs with Few-Shot Novel Labels via Meta Transformed Network Embedding
Lin Lan, Pinghui Wang, **Xuefeng Du**, Kaikai Song, Jing Tao, Xiaohong Guan.
Advances in Neural Information Processing Systems (NeurIPS) 2020.

9. Cross-domain palmprint recognition based on transfer convolutional autoencoder
Huikai Shao, Dexing Zhong, **Xuefeng Du**.
IEEE International Conference on Image Processing (ICIP) 2019.

8. Continual palmprint recognition without forgetting
   Xuefeng Du, Dexing Zhong, Huikai Shao.
   IEEE International Conference on Image Processing (ICIP) 2019.

7. Building an active palmprint recognition system
   Xuefeng Du, Dexing Zhong, Huikai Shao.
   IEEE International Conference on Image Processing (ICIP) 2019.

6. Low-shot palmprint recognition based on meta-siamese network
   Xuefeng Du, Dexing Zhong, Pengna Li.
   IEEE International Conference on Multimedia and Expo (ICME) 2019.

5. Classification in Cryo-Electron Tomograms
   Ilja Gubins, Gijs van der Schot, Remco C Veltkamp, Friedrich Förster, **Xuefeng Du**, Xiangrui Zeng, Zhenxi Zhu, Lufan Chang, Min Xu, Emmanuel Moebel, Antonio Martinez-Sanchez, Charles Kervrann, Tuan M Lai, Xusi Han, Genki Terashi, Daisuke Kihara, Benjamin A Himes, Xiaohua Wan, Jingrong Zhang, Shan Gao, Yu Hao, Zhilong Lv, Xiaohua Wan, Zhidong Yang, Zijun Ding, Xuefeng Cui, Fa Zhang.
   Eurographics 2019.

4. Open-set Recognition of Unseen Macromolecules in Cellular Electron Cryo-Tomograms by Soft Large Margin Centralized Cosine Loss
   Xuefeng Du, Xiangrui Zeng, Bo Zhou, Alex Singh, Min Xu.
   British Machine Vision Conference (BMVC) 2019, **Spotlight**.

3. Semi-supervised Macromolecule Structural Classification in Cellular Electron Cryo-Tomograms using 3D Autoencoding Classifier
   Siyuan Liu, **Xuefeng Du**, Rong Xi, Fuya Xu, Xiangrui Zeng, Bo Zhou, Min Xu.
   British Machine Vision Conference (BMVC) 2019.

2. Palm vein recognition with deep hashing network
   Dexing Zhong, Shuming Liu, Wenting Wang, **Xuefeng Du**.
   Pattern Recognition and Computer Vision 2019.

1. Palmprint recognition using siamese network
   Dexing Zhong, Yuan Yang, **Xuefeng Du**.
   13th Conference of Biometric Recognition (CCBR) 2018.

Journal Papers

13. Robust Palmprint Recognition via Multi-stage Noisy Label Selection and Correction
    Huikai Shao, Siyu Shi, **Xuefeng Du**, Dan Zeng, Dexing Zhong.
    IEEE Transactions on Image Processing, 2025.

12. Out-of-Distribution Learning with Human Feedback
    Haoyue Bai, **Xuefeng Du**, Katie Rainey, Shibin Parameswaran, Yixuan Li.
    Transactions on Machine Learning Research (TMLR), 2025.

11. Exploring Transition States of Protein Conformational Changes via Out-of-Distribution Detection in the Hyperspherical Latent Space
    Bojun Liu, Jordan G Boysen, Ilona Christy Unarta, **Xuefeng Du**, Yixuan Li, Xuhui Huang.
    Nature Communications, 2024.

10. OpenOOD v1.5: Enhanced Benchmark for Out-of-Distribution Detection
    Jingyang Zhang, Jingkang Yang, Pengyun Wang, Haoqi Wang, Yueqian Lin, Haoran Zhang, Yiyou Sun, **Xuefeng Du**, Kaiyang Zhou, Wayne Zhang, Yixuan Li, Ziwei Liu, Yiran Chen, Hai Li.
    Journal of Data-centric Machine Learning Research, 2023.

9. Noise-robust Graph Learning by Estimating and Leveraging Pairwise Interactions
   Xuefeng Du, Tian Bian, Yu Rong, Bo Han, Tongliang Liu, Tingyang Xu, Wenbing Huang, Yixuan Li, Junzhou Huang.
   Transactions on Machine Learning Research (TMLR), 2023.

8. Active learning to classify macromolecular structures in situ for less supervision in cryoelectron tomography
Xuefeng Du, Haohan Wang, Zhenxi Zhu, Xiangrui Zeng, Yi-Wei Chang, Jing Zhang, Eric Xing, Min Xu.
Bioinformatics, 2021.

7. A deep biometric hash learning framework for three advanced hand-based biometrics
Huikai Shao, Dexing Zhong, **Xuefeng Du**, 2021.
IET Biometrics.

6. Few-shot learning for palmprint recognition via meta-siamese network
Huikai Shao, Dexing Zhong, **Xuefeng Du**, Shaoyi Du, Raymond NJ Veldhuis.
IEEE Transactions on Instrumentation and Measurement, 2021.

5. Deep distillation hashing for unconstrained palmprint recognition
Huikai Shao, Dexing Zhong, **Xuefeng Du**.
IEEE Transactions on Instrumentation and Measurement, 2021.

4. Cross-domain palmprint recognition via regularized adversarial domain adaptive hashing
Xuefeng Du, Dexing Zhong, Huikai Shao.
IEEE Transactions on Circuits and Systems for Video Technology, 2020.

3. Effective deep ensemble hashing for open-set palmprint recognition
Huikai Shao, Dexing Zhong, **Xuefeng Du**.
Journal of Electronic Imaging, 2020.

2. A hand-based multi-biometrics via deep hashing network and biometric graph matching
Dexing Zhong, Huikai Shao, **Xuefeng Du**.
IEEE Transactions on Information Forensics and Security, 2019.

1. Decade progress of palmprint recognition: A brief survey
Dexing Zhong, **Xuefeng Du**, Kuncai Zhong, 2019.
Neurocomputing.

Others

2. Safety-Aware Fine-Tuning of Large Language Models
Hyeong Kyu Choi, **Xuefeng Du**, Yixuan Li.
Advances in Neural Information Processing Systems 2024, Safe Generative AI Workshop.

1. Efficient Deep Palmprint Recognition via Distilled Hashing Coding
Huikai Shao, Dexing Zhong, **Xuefeng Du**.
IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops 2019.

Working Papers

10. VAUQ: Vision-Aware Uncertainty Quantification for LVLM Self-Evaluation
Seongheon Park, Changdae Oh, Hyeong Kyu Choi, **Xuefeng Du**, Sharon Li.
arXiv preprint arXiv:2602.21054.

9. Towards Reducible Uncertainty Modeling for Reliable Large Language Model Agents
Changdae Oh, Seongheon Park, To Eun Kim, Jiatong Li, Wendi Li, Samuel Yeh, **Xuefeng Du**, Hamed Hassani, Paul Bogdan, Dawn Song, Sharon Li.
arXiv preprint arXiv:2602.05073.

8. Harnessing Reasoning Trajectories for Hallucination Detection via Answer-agreement Representation Shaping
Jianxiong Zhang, Bing Guo, Yuming Jiang, Haobo Wang, Bo An, **Xuefeng Du**.
arXiv preprint arXiv:2601.17467.

7. Unlocking the Pre-Trained Model as a Dual-Alignment Calibrator for Post-Trained LLMs
Beier Luo, Cheng Wang, Hongxin Wei, Sharon Li, **Xuefeng Du**.
arXiv preprint arXiv:2601.04277.

6. VLMGuard: Defending VLMs against Malicious Prompts via Unlabeled Data
   Xuefeng Du, Reshmi Ghosh, Robert Sim, Ahmed Salem, Vitor Carvalho, Emily Lawton, Yixuan Li, Jack W. Stokes.
   arXiv preprint arXiv:2410.00296.

5. The Ghanaian NLP Landscape: A First Look
   Sheriff Issaka, Zhaoyi Zhang, Mihir Heda, Keyi Wang, Yinka Ajibola, Ryan DeMar, **Xuefeng Du**
   arXiv preprint arXiv:2405.06818.

4. Learning by Passing Tests, with Application to Neural Architecture Search
   Xuefeng Du, Pengtao Xie
   arXiv preprint arXiv:2011.15102.

3. Skillearn: Machine Learning Inspired by Humans' Learning Skills
   Pengtao Xie, **Xuefeng Du**, Hao Ban
   arXiv preprint arXiv:2012.04863.

2. Small-Group Learning, with Application to Neural Architecture Search
   Xuefeng Du, Pengtao Xie
   arXiv preprint arXiv:2012.12502.

1. Towards efficient unconstrained palmprint recognition via deep distillation hashing
   Huikai Shao, Dexing Zhong, **Xuefeng Du**
   arXiv preprint arXiv:2004.03303.

Innovation

Patents Granted

- A kind of cross-platform palm grain identification method
  Dexing Zhong, Huikai Shao, **Xuefeng Du**, Runzhao Yao, CN109800677A.
- A kind of vehicle early warning vehicle intelligent system based on multisensor
  Dexing Zhong, Huikai Shao, **Xuefeng Du**, CN108492624A.

Patents Filed

- Guarding Multimodal Artificial Intelligence Systems From Malicious Prompt Attacks (pending, with Microsoft)
  Xuefeng Du, Reshmi Ghosh, Vitor Carvalho, Robert Sim, Emily Lawton, Jay Stokes, Lukas Wutschitz, Reshmi Ghosh, Ahmed Salem

TEACHING SUMMARY

Key Courses Taught (Last 3 years)

Course Code	*Course Title*	*Academic Year*	*Course Level*	*Type*	*Semester*
SC5005	NLP, LLM and Applications	AY25 – AY26	UG	Lecture & Tutorial	2
SC3000	AI	AY25 – AY26	UG	Tutorial	2

Academic Supervision and Mentoring

PhD students

No.	*PhD student*	*Period*	*Role*	*Thesis*	*Current Status*
Current					
2	Thanh Do Nguyen	2026-	Sole Supervisor	Reliability in Multi-Agent Systems	-
1	Junlin Fang	2026-	Sole Supervisor	Hallucination Detection in Large Reasoning Models	-

Masters students (By Coursework) & Undergraduate Students

Current

7. Wenyu Chen (Msc in Modeling and Simulation, SPMS, Independent study, Fall 2025-)

6. Yeqing Tang (Undergraduate student, CCDS, Independent study, Fall 2025-)

5. Kuo Eugene (Undergraduate FYP, CCDS, Spring 2026-)
   Role: Supervisor
   Project title: Detecting malicious prompts for multimodal language models

4. Wang Min-Jen (Undergraduate FYP, CCDS, Spring 2026-)
   Role: Supervisor
   Project title: Hallucination detection for LLMs via uncertainty estimation

3. Zhiyuan Wang (MSCS, CCDS, Thesis, Spring 2026-)
   Role: Supervisor
   Project title: Detecting malicious prompts for multimodal language models

2. Paliwal Siddharth (MSAI, CCDS, Thesis, Spring 2026-)
   Role: Supervisor
   Project title: Hallucination Detection for LLMs via Uncertainty Estimation

1. Sourabh Sooraj (MSAI, CCDS, Thesis, Spring 2026-)
   Role: Supervisor
   Project title: Hallucination detection for LLM agents

Past

1. Shunzhang Chen (MSAI, CCDS, Independent study, Fall 2025)
   Current status: Research intern at Singapore Huawei Research Center

Visiting Students
Current

2. Jianxiong Zhang (Visiting PhD student, Sichuan University, Fall 2025-Spring 2026)

1. Dazhou Liu (Undergraduate GEM Trailblazer Exchange, SUSTech, Spring 2026)

Confirmation Presentation Committees Served On

2. Fan Wu (PhD student, NTU CCDS, Spring 2026)
   Exam topic: Multi-Level Knowledge Alignment in Distributed Foundation Models
   Advisor: Lim Wei Yang Bryan

1. Yuzhou Cao (PhD student, NTU CCDS, Fall 2025)
   Exam topic: Convex, Smooth, and Consistent Surrogate Losses with Linear Surrogate Regret for General Prediction Tasks
   Advisor: Bo An

Thesis Committees Served On

5. Zonglin Yang (PhD 2025, NTU CCDS)
   Thesis topic: From Inductive Reasoning to Automated Scientific Discovery: Foundations and Methods with Large Language Models
   Role: Internal Examiner
   Advisor: Erik Cambria

4. Xiaobei Yan (PhD 2026, NTU CCDS)
   Thesis topic: Deep Learning Under The Microscope: Hardware-Oriented Attacks And Robust Defenses
   Role: Internal Examiner
   Advisor: Tianwei Zhang

3. Yutong Wu (PhD 2026, NTU CCDS)
   Thesis topic: Mutual Information as a Unified Lens for Deep Learning Security
   Role: Internal Examiner
   Advisor: Tianwei Zhang

2. Teng Yao Long (PhD student, NTU CCDS)
   Role: Thesis Advisory Committee
   Advisor: Bo An

1. Heyi Zhang (PhD student, NTU CCDS)
   Role: Thesis Advisory Committee
   Advisor: Tianwei Zhang

Mentees before tenure-track

9. Seongheon Park, PhD student at UW-Madison, now PhD student at UW-Madison.
8. Sheriff Issaka, undergraduate student at UW-Madison, 2024, now PhD student at UCLA.
7. Leitian Tao, undergraduate student at WHU, 2022, now PhD student at UW-Madison.
6. Tian Bian, PhD student at CUHK, 2021, now researcher at Alibaba DAMO Academy.
5. Gabriel Gozum, undergraduate student at UW-Madison, 2021, now perception engineer at Applied Intuition.
4. Eric Wang, undergraduate student at UW-Madison, 2021, now research assistant at UCSD.
3. Zhenxi Zhu, undergraduate student at BUPT, 2019, now PhD student at Nanjing University.
2. Alex Singh, undergraduate student at CMU, 2019, now AI Engineer at Tesla.
1. Pengna Li, undergraduate student at XJTU, 2019, now PhD student at XJTU.

Other Teaching Experiences

Exam Moderator for

- NTU, SC 4000 / CZ4041 Machine Learning (Fall 2025)

Teaching Assistant for

- CS762 (graduate course), UW-Madison: Advanced Deep Learning, Fall 2022.
  –Designing lectures on topics in trustworthy machine learning, grading students' written papers, holding office hours, leading discussions and answering questions.
- CS540 (undergraduate course), UW-Madison: Intro to AI, Spring 2021.
  –Designing lectures on topics in unsupervised learning, deep learning, and reinforcement learning. Preparing all the homework related to the topic of deep learning, and design final exams, guiding students on final projects, holding office hours, and delivering discussion sections.

SERVICE SUMMARY

College and School

- "Learning with AI" evaluation panel, NTU CCDS, 2025.
- UM-Madison, Admission Committee, Master for Data Science Program, 2024.

Academic Community

Area Chair for
ICLR 2026, ICML 2026

Session Chair for
AAAI 2026, *Interpretability and Internal Representations in Deep Models*

Reviewer for

- NeurIPS, ICML, ICLR 2021-2025.
- CVPR 2022-2024, ECCV 2022, ICCV 2023.
- WACV 2023, AAAI 2020, IJCAI 2021, MICCAI 2020.
- TMLR, Nature Communications, IJCV, IEEE TCSVT, TMM, TIP, TIFS.

Workshop Organization

- Co-organizer, Agentic AI in the Wild: From Hallucinations to Reliable Autonomy workshop at ICLR 2026.
- Co-organizer, Toward Trustworthy Vision-language Models in the Wild: Theory, Algorithm and Application workshop at ICMR 2026.
- Co-organizer, Socially Responsible and Trustworthy Foundation Models (ResponsibleFM) workshop at NeurIPS 2025.
- Co-organizer, Special session of Responsible Foundation Models in the Wild at IJCNN 2025.

ADDITIONAL EXPERIENCES

Research Intern June 2024 – Sept. 2024
Microsoft Research, Redmond, WA, USA
- Hosts: Dr. Robert Sim, Jay Stokes and Reshmi Ghosh.
- Research on *Malicious prompt detection for vision language models*

Student Researcher June 2022 – Sept. 2022
Google Research, Sunnyvale, CA, USA
- Hosts: Dr. Zizhao Zhang, Ting Chen and Han Zhang.
- Research on *Open-vocabulary object detection with language models*

Research Intern Mar. 2021 – June 2021
Tencent AI Lab, Shenzhen, China
- Supervised by Dr. Yu Rong and Junzhou Huang.
- Research on *Robust graph neural networks against noisy labels*.

Research Assistant Oct. 2020 – Jan. 2021
Department of Computer Science, Hong Kong Baptist University, Hong Kong
- Supervised by Dr. Bo Han.
- Research on *Effective network architecture for adversarial robustness*.

Research Engineer Intern Aug. 2020 – Sept. 2020
AI Lab, Bytedance Inc., Beijing, China
- Supervised by Dr. Changhu Wang.
- Research on *Fine-grained image classification*.

Research Intern Dec. 2019 – July 2020
AI Theory Group, Noah's Ark Lab, Shenzhen, China
- Supervised by Dr. Hang Xu and Chenhan Jiang.
- Research on *Hybrid supervised panoptic segmentation*.

Student Intern July 2018 – Feb. 2020
Department of Computational Biology, CMU, Pittsburgh, PA, USA
- Supervised by Dr. Min Xu and Dr. Haohan Wang.
- Research on *Deep learning for cellular electron cryo-tomography analysis*.

Research Assistant Nov. 2018 – Jun. 2020
Intelligent Networks and Network Security Lab, XJTU, Xi'an, China
- Supervised by Dr. Pinghui Wang.
- Research on *Few-shot node classification with meta learning*.

Research Assistant Jun. 2017 – Apr. 2019
Institute of Automatic Control, XJTU, Xi'an, China
- Supervised by Dr. Dexing Zhong.
- Research on *Machine learning for hand-based biometrics*.