# Xuefeng Du

Department of Computer Sciences
University of Wisconsin–Madison
1210 W Dayton St, Madison, WI 53706
Homepage, +1-608-720-4664
xfdu@cs.wisc.edu, xuefengdu1@gmail.com

## Research Interests

Reliable machine learning, with a special focus on developing empirical algorithms and theoretical understandings for learning & inference with out-of-distribution data; More recently on LLM hallucination and harmful prompt detection.

## Education Background

*Ph.D. candidate in Computer Sciences* Jan. 2021-Present
University of Wisconsin–Madison, Madison, WI

- Ph.D. research in open-world machine learning.
- Advisor: Prof. Sharon Yixuan Li.

*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.
- Major GPA: 93.04/100(3.91/4.0), Rank: 3rd/170.

## Research Experience

*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*
- Work on detecting malicious prompts for VLMs.
- Design a new general framework for representation-based detection leveraging unlabeled user prompts.

*Student Researcher* June 2022 – Sept. 2022
Google Research, Google Inc., Sunnyvale, CA, USA

- Hosts: Dr. Zizhao Zhang, Ting Chen and Han Zhang.
- Research on *open-vocabulary object detection with language models*
- Work on exploiting language models on open-vocabulary object detection.
- Design a new general framework for prompt-based object detection transformers.

*Research Intern* Mar. 2021 – June 2021
Machine Learning Center, Tencent AI Lab, Shenzhen, China

- Supervised by Dr. Yu Rong and Junzhou Huang.
- Research on *pairwise interactions for robust graph neural networks against noisy labels* .
- Work on exploiting pairwise interactions in graph neural network's training to combat noisy labels for semi-supervised node classification.
- Submit one paper on semi-supervised node classification against noisy labels to TMLR and get accepted.

*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.*
- Work on learning a diverse network architecture to improve adversarial robustness.
- Submit one paper to ICML and get accepted.

*Research Engineer Intern* Aug. 2020 -Sept. 2020

AI Lab, Bytedance Inc., Beijing, China

- Supervised by Dr. Changhu Wang.
- Research on *Fine-grained image classification.*
- Exploit pretraining techniques for fine-grained image classification.
- Design effective fine-grained image classification models for deployment on the product line.

*Research Intern* Dec. 2019 -July 2020

AI Theory Group, Noah's Ark Lab, Huawei Inc., Shenzhen, China

- Supervised by Dr. Hang Xu and Chenhan Jiang.
- Research on *Hybrid Supervised Panoptic Segmentation.*
- Propose a new Hybrid Supervised Panoptic Segmentation (HSPS) framework to significantly reduce the annotation cost for the most complex panoptic segmentation task. HSPS fully utilizes various kinds of weak supervision in the dataset, i.e., image labels, boxes and also semantic coherence between the thing and stuff branch, which achieves a 41.7% Panoptic Quality via 30% fully annotated data.
- Submit one paper to AAAI and get accepted.

*Student Intern* April 2019 – Nov. 2019

Department of Machine Learning, Carnegie Mellon University, Pittsburgh, PA, USA

- Supervised by Dr. Haohan Wang.
- Research on *Robust Machine Learning on Adversarial Attacks.*
- Propose two simple and effective intuitions to improve adversarial training. Apply One-Vs-All (OVA) models to improve adversarial training, which naturally allows the perturbation bound to be different for different classes.
- Propose a conditional adversarial training method that gradually improves the perturbation bound until no perturbed adversarial examples are considered valid.

*Student Intern* July 2018 – Feb. 2020

Department of Computational Biology, CMU, Pittsburgh, PA, USA

- Supervised by Dr. Min Xu.
- Research on *Deep Learning for Cellular Electron Cryo-Tomography (CECT).*
- Work on domain-specific problems in CECT, such as open-set novel macromolecule detection and recognition and label-efficient sub-tomogram classification.
- Submit four papers and get accepted.

*Research Assistant* Nov. 2018 – Jun. 2020

Intelligent Networks and Network Security Lab, XJTU, Xi'an, China

- Supervised by Dr. Pinghui Wang.
- Research on *Network Embedding and Meta Learning*.
- Propose a novel setting in network embedding: Node classification with Few-shot Novel labels.
- Integrate Meta-Learning flavored few-shot learning with classic network embedding techniques, such as DeepWalk and LINE to jointly learn the structure and classification information in graphs.
- Help submit one paper to NeurIPS and get accepted.

*Research Assistant* Jun. 2017 - Apr. 2019
Institute of Automatic Control, XJTU, Xi'an, China

- Supervised by Dr. Dexing Zhong.
- Research on *Machine Learning Powered Hand-based Biometrics*.
- Explore deep-learning based methods in hand-based biometrics, such as continual learning, adversarial domain adaptation for domain-specific problems.
- Help submit four papers.

## Achievements

*Awards*

- Jane Street Graduate Research fellowship, 2023.
- NeurIPS Scholar Award, 2022.
- CS departmental fellowship, UW-Madison, 2021.
- National Scholarship (2x), Ministry of Education, 2017, 2018.

*Talks*

- Talk at Adobe, 12/2021
- Talk at Microsoft, 12/2021
- Talk at Google, 01/2022
- Talk at MLOPT Seminar, UW-Madison, 05/2022
- Talk at AI talks, 02/2023
- Talk at Jane Street, 04/2023

## Papers and preprints on ML

See also my google scholar page.

20. Haoyue Bai, **Xuefeng Du**, Katie Rainey, Shibin Parameswaran, Yixuan Li , "Out-of-Distribution Learning with Human Feedback", arXiv preprint arXiv:2408.07772.
19. Sheriff Issaka, Zhaoyi Zhang, Mihir Heda, Keyi Wang, Yinka Ajibola, Ryan DeMar, **Xuefeng Du**, "The Ghanaian NLP Landscape: A First Look", arXiv preprint arXiv:2405.06818.
18. **Xuefeng Du**, Yiyou Sun, Yixuan Li, When and How does In-distribution Label Help Out-of-distribution Detection?, ICML 2024.
17. **Xuefeng Du**, Zhen Fang, Ilias Diakonikolas, Yixuan Li, How does Unlabeled Data Provably Help Out-of-distribution Detection, ICLR 2024.
16. **Xuefeng Du**, Yiyou Sun, Jerry Zhu, Yixuan Li, Dream the Impossible: Outlier Imagination with Diffusion Models, NeurIPS 2023.
15. Haoyue Bai, Gregory Canal, **Xuefeng Du**, Jeongyeol Kwon, Robert D Nowak, Yixuan Li, Feed Two Birds with One Scone: Exploiting Wild Data for Both Out-of-Distribution Generalization and Detection, ICML 2023.

14. Leitian Tao, **Xuefeng Du**, Jerry Zhu, Yixuan Li, Non-parametric Outlier Synthesis, ICLR 2023.

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

12. 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, OpenOOD v1.5: Enhanced Benchmark for Out-of-Distribution Detection, arXiv preprint arXix: 2306.09301

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

10. 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, OpenOOD: Benchmarking Generalized Out-of-Distribution Detection, Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track, 2022.

9. **Xuefeng Du**, Xin Wang, Gabriel Gozum, Yixuan Li, "Unknown-Aware Object Detection: Learning What You Don't Know from Videos in the Wild", CVPR 2022, **oral**.

8. Pengtao Xie, **Xuefeng Du**, "Performance-Aware Mutual Knowledge Distillation for Improving Neural Architecture Search", CVPR 2022.

7. **Xuefeng Du**, Eric Wang, Mu Cai, Yixuan Li , "VOS: Learning What You Don't Know by Virtual Outliers Synthesis", ICLR 2022.

6. **Xuefeng Du**, Jingfeng Zhang, Bo Han, Tongliang Liu, Yu Rong, Gang Niu, Junzhou Huang, Masashi Sugiyama , "Learning Diverse-Structured Networks for Adversarial Robustness", ICML 2021.

5. **Xuefeng Du**, Chenhan Jiang, Hang Xu, Gengwei Zhang, Zhenguo Li, "How to save your annotation cost for Panoptic Segmentation?", in AAAI 2021.

4. Lin Lan, Pinghui Wang, **Xuefeng Du**, Kaikai Song, Jing Tao, Xiaohong Guan, "Node Classification on Graphs with Few-Shot Novel Labels via Meta Transformed Network Embedding", in NeurIPS 2020.

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

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

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

### Papers in Submission

2. HaloScope: Harnessing Unlabeled LLM Generations for Hallucination Detection

1. VLMGuard: Safeguarding VLMs against Malicious Prompts with Unlabeled Data

## Papers on Bioinformatics

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

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

7. **Xuefeng Du**, Dexing Zhong, Huikai Shao, “Cross-domain palmprint recognition based on adversarial domain adaptative hashing”, in IEEE Transactions on Circuits and Systems for Video Technology, 2020.

6. Huikai Shao, Dexing Zhong and **Xuefeng Du**, “Efficient Deep Palmprint Recognition via Distilled Hashing Coding”, in CVPR Workshops 2019.

5. Dexing Zhong, Huikai Shao and **Xuefeng Du**, “A Hand-based Multi-biometrics via Deep Hashing Network and Biometric Graph Matching”, in IEEE Transactions on Information Forensics and Security. (TIFS), vol.14, issue.12, pp. 3140 - 3150. (IF 6.211)

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

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

2. 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, “Classification in Cryo-Electron Tomograms”, in Eurographics 2019.

1. Dexing Zhong, **Xuefeng Du**, and Kuncai Zhong, ”Decade progress of palmprint recognition: a brief survey”, Neurocomputing, 2018, vol. 328, pp.16-28. (IF 4.072)

## Service

*Reviewer*

- ICML, NeurIPS, ICLR
- CVPR, ECCV, ICCV
- IJCAI, IJCV, TCSVT, TIP, MICCAI

## Teaching

*Teaching Assistant for*

- CS540, UW-Madison: Intro to AI, Spring 2021.
- CS762, UW-Madison: Advanced Deep Learning, Fall 2022.

## Additional Information

*Language skills*

- Native speakers of Mandarin with fluent English speaking capability (CET4: 633, CET6: 627, TOEFL: 106 (S22), GRE: 160+166+4.5).

*Programming skills*

- Proficient with Python, TensorFlow, PyTorch. Familiar with C, Matlab, C++.