
Education Background

- Aug. 2021 - **Ph.D. Candidate**, Department of Computer Science, School of Computing, Present National University of Singapore, Singapore.
- Supervisor: Prof. Mohan Kankanhalli.
 - Research field: safety in machine learning [1, 2, 3, 4, 5, 6, 7, 8].
- Sep. 2017 - **Undergraduate**, Taishan College, Shandong University, Jinan, China.
- Jun. 2021
- Advisor: Prof. Lizhen Cui.
 - Research field: adversarial machine learning [9, 10, 11].

Publications

- [1] Zihao Luo*, Xilie Xu*, Feng Liu, Yun Sing Koh, Di Wang, and Jingfeng Zhang. Privacy-preserving low-rank adaptation for latent diffusion models. 2025.
- [2] Keyi Kong*, Xilie Xu*, Di Wang, Jingfeng Zhang, and Mohan Kankanhalli. Perplexity-aware correction for robust alignment with noisy preferences. In *The 38th Neural Information Processing Systems Annual Conference (NeurIPS)*, 2024.
- [3] Xilie Xu, Jingfeng Zhang, and Mohan Kankanhalli. Autolora: A parameter-free automated robust fine-tuning framework. In *The Twelfth International Conference on Learning Representations (ICLR)*, 2024.
- [4] Xilie Xu*, Keyi Kong*, Ning Liu, Lizhen Cui, Di Wang, Jingfeng Zhang, and Mohan Kankanhalli. An LLM can fool itself: A prompt-based adversarial attack. In *The Twelfth International Conference on Learning Representations (ICLR)*, 2024.
- [5] Jingfeng Zhang and Xilie Xu. Towards robust foundation models: Adversarial contrastive learning. In *The Third Blogpost Track at ICLR 2024*, 2024.
- [6] Xilie Xu*, Jingfeng Zhang*, Feng Liu, Masashi Sugiyama, and Mohan Kankanhalli. Efficient adversarial contrastive learning via robustness-aware coreset selection. In *The 37th Neural Information Processing Systems Annual Conference (NeurIPS)*, 2023.
- [7] Xilie Xu*, Jingfeng Zhang*, Feng Liu, Masashi Sugiyama, and Mohan Kankanhalli. Enhancing adversarial contrastive learning via adversarial invariant regularization. In *The 37th Neural Information Processing Systems Annual Conference (NeurIPS)*, 2023.

⁰An asterisk (*) beside authors' names indicates equal contributions.

- [8] Xilie Xu*, Jingfeng Zhang*, Feng Liu, Masashi Sugiyama, and Mohan Kankanhalli. Adversarial attack and defense for non-parametric two-sample tests. In *The 39th International Conference on Machine Learning (ICML)*, 2022.
- [9] Jingfeng Zhang*, Xilie Xu*, Bo Han, Tongliang Liu, Gang Niu, Lizhen Cui, and Masashi Sugiyama. Noilin: Improving adversarial training and correcting stereotype of noisy labels. *Transactions on Machine Learning Research*, 2022.
- [10] Chen Chen*, Jingfeng Zhang*, Xilie Xu, Lingjuan Lyu, Chaochao Chen, Tianlei Hu, and Gang Chen. Decision boundary-aware data augmentation for adversarial training. *IEEE Transactions on Dependable and Secure Computing*, 2022.
- [11] Jingfeng Zhang*, Xilie Xu*, Bo Han, Gang Niu, Lizhen Cui, Masashi Sugiyama, and Mohan Kankanhalli. Attacks which do not kill training make adversarial learning stronger. In *The 37th International Conference on Machine Learning (ICML)*, 2020.

Professional Service

- 2021-Present **Conference reviewer at NeurIPS’[21-24], ICML’[22-24], ICLR’[22-25].**
- 2022-Present **Journal reviewer at TAI, TMLR, IPL.**
- Apr. 2024 **DAAD AInet Fellow**, Germany.
- Apr. 2023 **Create the website of N-CRiPT Technical Workshop 2023 [link] and present a poster at the workshop**, National University of Singapore.
- Oct. 2022 **Member of the executive group of TrustML Young Scientist Seminars**, RIEKN-AIP, Tokyo.
- Apr. 2022 **Student reviewing member of the Master of Computing admission**, School of Computing, National University of Singapore.

Award&Grant

- Jul. 2024 **Championship**, ICML 2024 TiFA Workshop MLLM Attack Challenge.
- Mar. 2024 **Postgraduate Student Travel Grant**, School of Computing, National University of Singapore.
- Dec. 2023 **Postgraduate Student Travel Grant**, School of Computing, National University of Singapore.
- Oct. 2023 **NeurIPS 2023 Scholar Award**, NeurIPS 2023.
- Aug. 2022 **Postgraduate Student Travel Grant**, School of Computing, National University of Singapore.
- Aug. 2022 **Research Achievement Award**, School of Computing, National University of Singapore.
- Jun. 2022 **ICML 2022 Participation Grant**, ICML 2022.
- Oct. 2021 **Outstanding Reviewer Award**, NeurIPS 2021.
- Jun. 2021 **Outstanding Undergraduate Thesis Award**, Shandong University.
- Sep. 2020 **Specialty Scholarship (Research Innovation Award)**, *First Prize*, Shandong University.

- Sep. 2018 - **Outstanding Student Scholarship**, Shandong University.
Sep. 2020
Oct. 2018 **First Prize at the 10th Mathematics Competition of Chinese College Student**, Chinese Mathematical Society.

Internship

- Jun. 2024 - **Quantitative Research Intern**, Baiont Quant, Nanjing, China.
Present
Jun. 2021 - **Research Intern**, Department of Ant Group-CRO Line-Security and Risk
Jul. 2021 Management, Ant Z Space, Hangzhou, China.
 - Mentor: Dr. Lingjuan Lyu.
 - Research topic: adversarial machine learning and privacy.
 - Result: Proposed an innovative patent to protect model intellectual property and data privacy.

Teaching

- Jan. 2024 - **Teaching assistant for CS5342 Multimedia Computing and Applications**, School of Computing, National University of Singapore.
May 2024
Jan. 2024 - **Teaching assistant for CS3244 Machine Learning**, School of Computing,
May 2024 National University of Singapore.
Aug. 2023 - **Teaching assistant for CS3244 Machine Learning**, School of Computing,
Dec. 2023 National University of Singapore.
Jan. 2022 - **Teaching assistant for CS5242 Deep Learning and Neural Networks**,
May. 2022 School of Computing, National University of Singapore.