Yijun Yang, Ph.D. Candidate

yjyang@cse.cuhk.edu.hk yangyj16@tsinghua.org.cn
Yijun's Homepage Github
Google Scholar Research Gate
+86 18613370368 +852 59862685

Short Bio

- I am a final-year Ph.D. candidate at the CUhk REliable Laboratory (CURE), part of the Department of Computer Science and Engineering at The Chinese University of Hong Kong, under the guidance of Prof. Qiang Xu. Our laboratory primarily concentrates on tasks related to AI security and robustness. Prior to this, I earned my M.Phil in Electrical Engineering, with a focus on hardware security chip design, from Tsinghua University in 2019. Currently, my research interests are concentrated in the realms of AI Safety. This includes working on designing robust LLM and diffusion models and developing adversarial attacks on diffusion-based Text-to-Image models.

Education

- **Ph.D., The Chinese University of Hong Kong**, Hong Kong S.A.R.
  Department of Computer Science and Engineering. GPA: 3.7/4.0
- **M.Phil., Tsinghua University**, Beijing, China.
  Department of Integrated Circuit Engineering. GPA: 3.7/4.0, Ranking: 3/45
- **B.Eng, Central South University**, Changsha, China.
  Department of Automation. GPA: 3.7/4.0

Research intern

- July 2023 - Nov. 2023, AIGC group, Wenge·YaYi Large Model, Beijing, China
- Mar. 2022 - June 2023, Foundation Model, Megvii, Beijing, China
- Mar. 2020 - June 2020, 2012 Lab, Huawei, Shenzhen, China

Selected Research Publications

1. **Yijun Yang**, Ruiyuan Gao, Xiaosen Wang, Tsung-Yi Ho, Nan Xu, & Qiang Xu. (2023). Mma-diffusion: Multimodal attack on diffusion models. *Submit to IEEE / CVF Computer Vision and Pattern Recognition Conference (**CVPR 2024** under review)*. Retrieved from https://arxiv.org/abs/2311.17516
2. **Yijun Yang**, Xiangyu Wen, Ruiyuan Gao, Xiangyu Zhang, & Qiang Xu. (2023). Defending object detectors against adversarial hiding attacks with semantic input validation. *On going*.
3. Zhiyuan He*, **Yijun Yang***, Pin-yu Chen, Qiang Xu, & Tsung-Yi Ho. (2023). Be your own neighborhood: Detecting adversarial example by the neighborhood relations built on self-supervised learning. *Submit to International Conference on Learning Representations (**ICLR 2024** under review) * co-first author.*
4. **Yijun Yang**, Ruiyuan Gao, & Qiang Xu. (2022). Out-of-distribution detection with semantic mismatch under masking. *European Conference on Computer Vision (**ECCV 2022**)*. Retrieved from https://arxiv.org/abs/2208.00446
5. **Yijun Yang**, Ruiyuan Gao, Yu Li, Qiuxia Lai, & Qiang Xu. (2022). What you see is not what the network infers: Detecting adversarial examples based on semantic contradiction. *Network and Distributed Systems Security (**NDSS 2022**)*.
6. **Yijun Yang**, Ruiyuan Gao, Yu Li, Qiuxia Lai, & Qiang Xu. (2021). Mixdefense: A defense-in-depth framework for adversarial example detection. *The International Symposium on Computer Architecture (**ISCA 2021**) Workshop*. Retrieved from https://sites.google.com/usc.edu/spsl/home

Services

- I am invated as a reviewer of academic conferences: **ICASSP**, **NeurIPS**, **ICLR**, **CVPR**

Selected Award and Honors

- **International Algorithm Case Competition 2023** - Adversarial Defence Competition, 2_{nd} place.
- **Full Postgraduate Studentship**, The Chinese University of Hong Kong.
- **Outstanding Master Graduate**, Tsinghua University (Top 2%).
- **Outstanding Thesis Award**, Tsinghua University (Top 3%).
- **Scholarship for Advancement in Academic Work**, Tsinghua University (Top 5%).
- **Scholarship for Advancement in Academic Work**, Tsinghua University (Top 5%).
- **Outstanding Bachelor Graduate**, Central South University (Top 3%).
- **Outstanding Thesis Award**, Central South University (Top 5%).