



ZHENGYUAN JIANG

📍 2080 Duke University Road, Durham, NC 27708 📞 (1)-984-312-9065
HomePage |  GitHub |  LinkedIn | Google Scholar | ✉️ zhengyuan.jiang@duke.edu

EDUCATION

Duke University (Advised by Prof. Neil Gong)
Ph.D. Student, Electrical and Computer Engineering

North Carolina, USA
2022.9 - 2027 (Expected)

University of Science and Technology of China
B.Eng., Information Science and Technology with Honors (top 5%)

Hefei, P.R. China
2018.9 - 2022.7

RESEARCH EXPERIENCE

Evading Watermark-based AI-generated Image Detection [Code]

January 2023 - May 2023

Advisor: Prof. Neil Gong, Duke University

- Proposed WEvade, the state-of-the-art image watermark removal attack, which can add small, human-imperceptible perturbations to AI-generated images to evade watermark-based detectors
- Extended adversarial examples to watermarking and were the first to introduce the double-tailed detector
- Theoretically analyzed the evasion rates of WEvade in both white-box and black-box settings via rigorous derivation

Watermark-based Detection and Attribution of AI-Generated Content

May 2023 - November 2023

Advisor: Prof. Neil Gong, Duke University

- Conducted the first systematic study on watermark-based, user-level attribution of AI-generated content
- Formally quantified the behavior of watermarking, based on which we provide a theoretical analysis of detection and attribution performance
- Building on our theoretical insights, we formulated watermark selection as an optimization problem and developed an efficient approximate solution

Certifiably Robust Image Watermark [Code]

November 2023 - April 2024

Advisor: Prof. Neil Gong, Duke University

- Proposed the image watermarks with certified robustness guarantees against removal and forgery attacks
- Extended randomized smoothing, a popular technique for constructing certifiably robust classifiers and regression models, to the task of watermarking
- Designed multi-class, multi-label, and regression smoothing to build certifiably robust image watermarks.
- Achieved certified robustness and also better empirical robustness

AudioMarkBench: Benchmarking Robustness of Audio Watermarking [Code]

January 2024 - June 2024

Collaborator: Dr. Lun Wang, Google DeepMind

- Conducted the first systematic and comprehensive benchmark for assessing the robustness of audio watermarking
- Evaluated robustness against both watermark removal and forgery attacks

AI-generated Image Detection: Passive or Watermark [Code]

May 2024 - November 2024

Collaborator: Dr. Amir Sadovnik, Oak Ridge National Lab

- Proposed ImageDetectBench, the first benchmark designed to systematically compare the effectiveness, robustness, and efficiency of passive and watermark-based AI-generated image detectors
- Incorporated four diverse datasets, eleven types of perturbations, and nine detectors
- Presented key findings and offered recommendations for AI-generated image detection

SafeText: Safe Text-to-image Models via Aligning the Text Encoder

April 2024 - October 2024

Advisor: Prof. Neil Gong, Duke University

- Proposed SafeText, a novel alignment method for text-to-image models.
- Fine-tuned the text encoder of a text-to-image model to preserve image utility to the greatest extent.
- Demonstrated that SafeText outperforms existing alignment methods for text-to-image models, achieving state-of-the-art performance across three prompt datasets with different models.

Jailbreaking Safeguarded Text-to-Image Models via Large Language Models

May 2024 - November 2024

Collaborator: Prof. Yinzhi Cao, Johns Hopkins University

- Proposed PromptTune, a query-free jailbreak attack to bypass guardrails of a safeguarded text-to-image model.
- Utilized SFT and DPO to fine-tune a large language model to generate adversarial prompts.
- Demonstrated that three variants of our PromptTune outperform current attacks.

PUBLICATIONS

Pengfei Zhang, Zhengyuan Jiang, Yixuan Wang, Yu Li. CLMB: deep contrastive learning for robust metagenomic binning. International Conference on Research in Computational Molecular Biology (RECOMB), 2022. [Paper]

Zhengyuan Jiang, Jinghui Zhang, Neil Gong. Evading Watermark based Detection of AI-Generated Content. ACM Conference on Computer and Communications Security (CCS), 2023. [Paper]

Zhengyuan Jiang, Minghong Fang, Neil Gong. IPCert: Provably Robust Intellectual Property Protection for Machine Learning. IEEE/CVF International Conference on Computer Vision (ICCV) Workshop, 2023. [Paper]

Zhengyuan Jiang, Moyang Guo, Yuepeng Hu, Neil Gong. Certifiably Robust Image Watermark. European Conference on Computer Vision (ECCV), 2024. [Paper]

Hongbin Liu, Moyang Guo, Zhengyuan Jiang, Lun Wang, Neil Gong. AudioMarkBench: Benchmarking Robustness of Audio Watermarking. NeurIPS Datasets and Benchmarks Track, 2024. [Paper]

Yuepeng Hu, Zhengyuan Jiang, Moyang Guo, Neil Gong. A Transfer Attack to Image Watermarks. International Conference on Learning Representations (ICLR), 2025. [Paper]

Yuepeng Hu, Zhengyuan Jiang, Neil Gong. SafeText: Safe Text-to-image Models via Aligning the Text Encoder. Under Submission, 2024.

Zhengyuan Jiang, Yuepeng Hu, Yuchen Yang, Yinzhi Cao, Neil Gong. Jailbreaking Safeguarded Text-to-Image Models via Large Language Models. Under Submission, 2024.

TECHNICAL SKILLS

Programming	Python (Advanced), C, MATLAB, HTML
Frameworks	Pytorch, Tensorflow, Scikit-Learn, Matplotlib
Software&Tools	Git, PyCharm, VSCode, MATLAB
Soft Skills	Academic Writing & Speaking, Teamwork, Critical Thinking

REWARDS

USTC Undergraduate Honorary Rank Candidate	2021
Huawei Scholarship	2021
ZengHua Scholarship (top 2% at USTC)	2020
CASC Scholarship	2020
Talent Student Scholarship (top 5% at USTC)	2019

ADDITIONAL INFORMATION

Research Interests: AI Security, GenAI Security, Diffusion Model, MLLM, Robustness, ect.

Program Committee Service: ICLR 2025, ICML 2025, ACM Multimedia 2023 & 2024.

Other Interests: Photography, Swimming, Badminton, Table Tennis, Video Game.