Shunchang Liu

Rte Cantonale, 1015 Lausanne, Switzerland
shunchangliu@outlook.com | (+86)18810220881 | Google Scholar

RESEARCH INTERESTS

AI Safety & Privacy, Adversarial Robustness, Alignment, Generative AI Copyright Protection

EDUCATION

EPF Lausanne - ETH Zurich (joint-degree program), Lausanne & Zurich, Switzerland — Sep 2024 | Present
Master of Computer Science: Cybersecurity — GPA: 5.70 / 6.00

Beihang University, Beijing, China — Sep 2021 | Oct 2023
Joint Study of Master and Doctoral Degree in Computer Science (Quit) — GPA: 3.83 / 4.00

Beihang University, Beijing, China — Sep 2017 | Jun 2021
Bachelor of Engineering: Automation — GPA: 3.87 / 4.00

WORK EXPERIENCES

Concordia AI (Social Enterprise focused on AI Safety and Governance) — Beijing, China
Intern, Mentor: Brian Tse — Apr 2024 | Aug 2024

- Built the database for China's AI safety evaluation methods and benchmarks, with results published in the "China AI Security Panorama Report (Spring 2024 Edition)"
- Provided third-party services to businesses, such as testing the model robustness through jailbreaks and assisting businesses in developing responsible scaling policies

PUBLICATIONS

First author / Corresponding author (†)

- **Shunchang Liu**, Zhuan Shi, Lingjuan Lyu, Yaochu Jin, Boi Faltings. CopyJudge: Automated Copyright Infringement Identification and Mitigation in Text-to-Image Diffusion Models, *The ACM International Conference on Multimedia (MM)*, 2025 [paper]
- Tony Ma, Songze Li, Yisong Xiao, **Shunchang Liu†**. Boosting Cross-task Transferability of Adversarial Patches with Visual Relations, *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshop*, 2023 [paper]
- **Shunchang Liu**, Jiakai Wang, Aishan Liu, Yingwei Li, Yijie Gao, Xianglong Liu, Dacheng Tao. Harnessing Perceptual Adversarial Patches for Crowd Counting, *The ACM Conference on Computer and Communications Security (CCS)*, 2022 [paper] [code]

Co-author

- Yisong Xiao, Tianyuan Zhang, **Shunchang Liu**, Haotong Qin. Benchmarking the Robustness of Quantized Models, *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshop*, 2023 [paper]
- Aishan Liu, Huiyuan Xie, Xianglong Liu, Zixin Yin, **Shunchang Liu**. Revisiting audio visual scene-aware dialog, *Neurocomputing*, 2022 [paper]
- Jiakai Wang, Aishan Liu, Zixin Yin, **Shunchang Liu**, Shiyu Tang, Xianglong Liu. Dual Attention Suppression Attack: Generate Adversarial Camouflage in Physical World, *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021 (**Oral**) [paper] [code]

RESEARCH EXPERIENCES

Reward Hacking Mitigation in RLHF via Sparse Autoencoder — ETH Zurich
Semester Project, Supervisor: Prof. Andreas Krause and Dr. Xin Chen — Mar 2025 | Present

- Used sparse autoencoder (SAE) to construct an interpretable preference model and employing causal intervention on SAE features to mitigate reward hacking during RLHF (in progress)

Copyright Infringement Detection and Mitigation for Diffusion Models — EPFL
Semester Project, Supervisor: Prof. Boi Faltings and Dr. Zhuan Shi — Sep 2024 | Jan 2025

- Proposed CopyJudge, an automated framework that uses large vision-language models to detect copyright infringement by comparing AI-generated and copyrighted images using abstraction-filtration-comparison and multi-model debate, and reducing risk through prompt tuning and non-infringing noise exploration in latent space.

- Published one [paper] on ACM MM

Adversarial Examiner for Human-Object Interaction Johns Hopkins University
Summer Intern, Supervisor: Prof. Alan L. Yuille and Prof. Tianmin Shu July 2023 | Oct 2023

- Generated out-of-distribution 3D human-object interaction (HOI) samples automatically by manipulating latent parameters for full-body interactions with multilayer perceptron, which provided an efficient method to test the generalization performance of current HOI reconstruction models

Cross-task Adversarial Patch Generation Beihang University
Research Assistant, Supervisor: Dr. Jiakai Wang Jan 2023 | Jun 2023

- Proposed a novel visual relation-based adversarial patch generation method combing object misclassification and predicate-based relation elimination, which improved black-box adversarial transferability across diverse visual reasoning tasks such as image captioning and visual question answering
- Published one [paper] on IEEE CVPR Workshop

Adversarial Patch Generation for Crowd Counting Beihang University
Research Assistant, Supervisor: Prof. Aishan Liu and Dr. Yingwei Li (JHU) Jun 2021 | Aug 2022

- Proposed a novel perceptual adversarial patch generation framework that exploited model-inherent perceptual properties, e.g., scale and position perceptions, of crowd counting models, which led to the SOTA transferable attack
- Employed adversarial training with our patches to improve models' cross-dataset generalization and robustness towards complex backgrounds, which showcased empirical evidence of its beneficial impact on vanilla models' performance
- Published one [paper] on ACM CCS

Bias Exploration in Audio Visual Scene-Aware Dialog Task Beihang University
Research Assistant, Supervisor: Prof. Aishan Liu and Dr. Huiyuan Xie (Cambridge) Sep 2020 | Dec 2020

- Revisited the audio-visual scene-aware dialog task unveiled biases in models, datasets, and evaluation metrics, which notably exposed the overreliance of current models on textual data and inadequate utilization of visual information
- Published one [paper] on Neurocomputing

Adversarial Camouflage Generation in Physical World Beihang University
Research Assistant, Supervisor: Dr. Jiakai Wang July 2020 | Nov 2020

- Proposed a novel dual attention suppression attack to generate visually natural adversarial vehicle camouflages by evading both model-shared attention and human-specific attention, which achieved state-of-the-art black-box attacking performance in both digital and physical world towards classification and detection tasks
- Published one [paper] on IEEE CVPR

PROJECTS

Trustworthiness Evaluation on LLMs University of Illinois Urbana-Champaign
Autumn Intern, Supervisor: Prof. Bo Li Oct 2023 | Dec 2023

- Contributed to a [leaderboard] for open large language models that considered various trustworthiness perspectives, which contributed to a comprehensive evaluation of large-scale models' performance across aspects such as toxicity, bias, robustness, privacy, ethics, and fairness

AWARDS

- Outstanding Graduate of Beihang University **(Top 20%)** Jun 2021
- CVPR Security AI Challenge: No-limit Adversarial Attacks on ImageNet **(11/1559)** Apr 2021
- First Prize of Discipline Competition Scholarship, BUAA Dec 2019
- First Prize of Study Excellence Scholarship, BUAA Dec 2019
- Outstanding Student of Beihang University **(Top 5%)** Jun 2019
- Second Prize of Group A of Non-physics in the 35th National Physics Competition for College Students Dec 2018

SERVICES

- **Program Committee / Reviewer:** Pattern Recognition, Frontiers of Computer Science, IEEE TCSVT, IEEE T-ITS, and Workshops (ICML MoFA 2025, CVPR AdvCV 2023, etc.)

SKILLS

- **Programming:** C, Python, R, etc.
- **Software:** Matlab, Pytorch, TensorFlow, MindSpore, etc.
- **Language:** Chinese, English (IELTS Academic overall 7.0, listening 6.5, reading 8.0, speaking 6.0, writing 7.0)