Zhongliang Guo

Blue texts refer to clickable links

guo_zhongliang@hotmail.com| Homepage| Google Scholar

Technology Stack

Areas of Expertise:	AI Robustness, AI Safety, Trustworthy AI, Computer Vision, LLMs
Programming Language:	Python, JAVA, SQL, C#, JavaScript, LaTeX, HTML5
Libraries & Frameworks:	PyTorch, Diffusers, OpenCV, NumPy, Pandas, Matplotlib, Django
Tools & Technologies:	Linux, Shell, Vim, Slurm, Docker, Git

Education

PhD ***Computer Science***, University of St Andrews, *Full scholarship with stipend*, Supervisor: Oggie and Lei	2022 - 2025
MSc ***Artificial Intelligence*** with **Distinction**, University of St Andrews, Nominated on 2021/2 *Deans' List*	2021 - 2022
BSc ***Forensic Science***, NWUPL, GPA: 88.4/100 (ranked 1/55), Awarded 2021 Outstanding UG Dissertation	2017 - 2021

Honor & Grant

	Date
• 2021 Outstanding Undergraduate Dissertation	Jun 2021
• 2021/2022 Dean's List at University of St Andrews	Sep 2022
• 2022 - 2026 full PhD scholarships with stipend	Oct 2022
• ECAI 2024 Conference Travel Grant from EurAI (22/547)	Oct 2024
• CVPR 2025 Highlight Paper (235/2360)	Jun 2025
• Leading Guest Editor for Pattern Recognition	Oct 2025

Research Experience

1. Adversarial Attack for Social Good

- **Principal Investigator**. Explore the benign use of adversarial attack in terms of computer vision.
- Propose an **adversarial pre-processing method** to protect artwork from unauthorized neural style transfer, allowing safeguarding unique style against popular transfer techniques, coming with a color-centric **Image Quality Assessment**.
- Propose a near black-box attack method against **Latent Diffusion Models**, achieving SOTA performance at 4× faster than existing approaches, reducing the VRAM occupation by 60%.
- Propose a Diffusion-based facial privacy protection method using adversarial techniques.

2. Adversarial Attack for AI Robustness

- **Principal Investigator**. Explore the **vulnerability** of existing machine learning models and potential **defenses**.
- Expose the **illusory robustness** in SOTA signature verification models, proposing a False Positive attack to address the unbalanced performance of existing attack methods.
- Propose an attack framework against **multi-modal diffusion models**, utilizing distilled backbones and optimized noise predictors to generate high-fidelity adversarial examples with superior **transferability** and **robustness** against defenses.
- Propose a **one-step** diffusion-based adversarial purification method using controlled purification and noise distillation, speed up 100× while maintain 76% robustness.

3. LLMs and its Robustness

- Propose, implement, and deploy a dual-retrieval **RAG**, to improve the Q&A performance of **LLMs** in the industry. Propose a **multi-agent evaluation protocol**, come with a **new data generation paradigm** for industrial scenarios.
- Propose the feature alignment enhancement paradigm and a new **backdoor attack** method for **LLMs**. The proposed method significantly improved the backdoor attack success rate while maintaining the model's conventional task reasoning performance, revealing the undiscovered weaknesses of large language models.
- Propose a new **data synthesis** method for backdoor attacks on Chinese LLMs, which significantly improves the concealment of backdoor attacks. The proposed new method achieves SOTA performance on various models and various baselines.

4. Object Counting (Supervised/Semi-Supervised/Unsupervised)

- Use **density graph estimation** network architecture to effectively improve the accuracy and robustness of target counting in complex scenarios.
- For scenarios where labeling data is limited, develop a new **semi-supervised learning** method, using only 40% labeling data to achieve the accuracy comparable to full labeling.
- Reveal that existing zero-shot methods is insensitive to text prompts, and the widely-used dataset has labelling bias. Leveraging the T2I Diffusion Model, achieve **text-guided**, **zero-shot** object counting.
- Solve the problem of aberration between the existing natural image and thermal image crowd counting dataset. Use the **unsupervised modal alignment** based on visual prompts to achieve high-precision counting **without natural images**.

5. **Content Safety and High-Risk Scenario Assessment for Large Language Models**
   - Benchmarking the content safety for LLMs in public health-sensitive information provision.
   - Constructed a dataset containing 2,160 QA pairs covering **safety boundary check**, **quantitative information accuracy**, and **risk level reasoning**.
   - Exposed failure risks of existing LLM safety mechanisms in high-risk contexts.
   - Demonstrated that **excessive safety** constraints may lead to **refusal of service**, compromising model availability.

Industry Work Experience

1. **AIGC Algorithm Engineer (Intern)**, DreamFace AI Video Generation Team — Sep 2025 - OCt 2025
   - **AI Agent System Development**. Design end-to-end AI agent for **automated long biographical video generation**.
   - Develop timeline-based storyboard generation pipeline using **multi-agent** coordination for long-form video creation.
   - Engineer smooth video transitions using first-frame and last-frame **constrained generation techniques**.
   - Implement quality assessment framework with **Visual Quality Assessment** and **vLLM** evaluation metrics.
   - Design **iterative refinement workflow** with agent-based quality control for continuous improvement.

Academic Work Experience

1. **Research Collaboration with City University of Hong Kong** — Apr 2024 - Now
   - **Technical Mentor**. Serve as a technical mentor for Prof. Chun Pong Lau's lab.
   - Provide academic guidance to **3 PhD students** on adversarial attack/defense and diffusion-based generation.
   - Participate in research ideation sessions, helping to conceptualize and validate experimental approaches.
   - Contribute to 2 papers in **CVPR 2025**, 1 paper in **Pattern Recognition**, many papers in writing.
2. **LLMs Research Fellow**, University of St Andrews, funded by **Tapoly** — Jan 2025 - Mar 2025
   - **Principal Researcher**. An AI **agent** based on LLMs with unique knowledge of insurance industry.
   - Design, implement and deploy AI agent for the **insurance industry** to automatically solve customer needs, such as policy inquiries and intelligent claims settlement.
   - Propose a more **economical** AI agent implementation and deployment framework, making the response of AI agents more accurate **without fine-tuning or training**.
   - Propose a performance evaluation framework for **multi-agent** to align with the requirement of data sensitivity, and data insufficiency in the insurance industry.
3. **Radar Algorithm Research Fellow**, University of St Andrews, funded by **MathWorks** — Dec 2023 - Nov 2024
   - **Principal Researcher**. Machine Learning based **drone** and **bird radar detection** using micro-Doppler radar signature.
   - Design and implement physical models to simulate **avian** and **drone dynamics**.
   - Conduct field experiments to collect various radar frequency data of birds and drones.
   - Process the signal data into corresponding micro-Doppler signatures and categorizing, labeling a new dataset.
   - Use **physic-driven data transformation** reduces the redundancy and complexity of radar signal, making it machine learning-friendly with **data compression** rates up to **96%**.
   - Develop multiple usage neural network for bird-drone-clutter-noise classification and moving object tracking.
4. **Teaching Assistant** (Covers UG level and PGT level) — Sep 2023 - May 2025
   - Modules include CS1002 OOP, CS3105 AI, and ID5059 KDD.
   - Topic covers Programming Languages, Machine Learning, Artificial Intelligence, Deep Learning, and Statistics.
   - Demonstrate lab session, tutorial, lecture, and mark coursework.
   - Designed a seminar about AI robustness with replacing slides to an interactive webpage, delivering to audiences unfamiliar with the topic. [link]

Academic Service

- **Leading Guest Editor** for Trustworthy AI Special Issue of *Pattern Recognition*
- Reviewer for *Pattern Recognition*, *Information Sciences*, *IEEE T-IFS*, *ICLR*, *CVPR*, *NeurIPS*, *AAAI*
- Volunteer for *ECAI 2024*

Selected Publication & Patent

*equal contribution †corresponding author

- **Zhongliang Guo**, Chun Tong Lei, Lei Fang, Shuai Zhao, Yifei Qian, Jingyu Lin, Zeyu Wang, Cunjian Chen, Ognjen Arandjelović, and Chun Pong Lau. A Grey-box Attack against Latent Diffusion Model-based Image Editing by Posterior Collapse. *IEEE Transactions on Information Forensics and Security* (**IEEE TIFS**), 2025.
- **Zhongliang Guo**, Yifei Qian, Shuai Zhao, Junhao Dong, Yanli Li, Ognjen Arandjelović, Fang Lei, and Chun Pong Lau. Artwork Protection Against Unauthorized Neural Style Transfer and Aesthetic Color Distance Metric. ***Pattern Recognition***, 2025.
- **Zhongliang Guo**†, Yifei Qian, Kaixuan Wang, Weiye Li, Ziheng Guo, Yuheng Wang, Yanli Li, Ognjen Arandjelović, and Lei Fang. Artwork Protection Against Neural Style Transfer Using Locally Adaptive Adversarial Color Attack. In *The 27th European Conference on Artificial Intelligence* (**ECAI 2024 Oral**), 2024.
- **Zhongliang Guo**†, Weiye Li, Yifei Qian, Ognjen Arandjelovic, and Lei Fang. A White-Box False Positive Adversarial Attack Method on Contrastive Loss-Based Offline Handwritten Signature Verification Models. In *The 27th International Conference on Artificial Intelligence and Statistics* (**AISTATS 2024**), 2024.
- **Zhongliang Guo**†, Ognjen Arandjelović, David Reid, Yaxiong Lei, and Jochen Büttner. A Siamese Transformer Network for Zero-Shot Ancient Coin Classification. ***Journal of Imaging***, 2023.
- **Zhongliang Guo**, Dian Jia, Zhaokai Wang, and Yongqi Zhou. A Method of Video Recognition Network of Face Tampering Based on Deep Learning, **A.U. Patent** 2019101186A4, Oct. 2019.
- **Zhongliang Guo**, Yifei Qian, Yanli Li, Weiye Li, Chun Tong Lei, Shuai Zhao, Lei Fang, Ognjen Arandjelović, and Chun Pong Lau. Beyond Vulnerabilities: A Survey of Adversarial Attacks as Both Threats and Defenses in Computer Vision Systems. ***Under review*** *of ACM Computing Surveys*, 2025.
- **Zhongliang Guo**, Siyuan Huang, and Chun Pong Lau. MMAD-Purify: A Precision-Optimized Framework for Efficient and Scalable Multi-Modal Attacks. ***Under review*** *of Pattern Recognition*, 2025.
- Yifei Qian*, **Zhongliang Guo***, Bowen Deng, Chun Tong Lei, Shuai Zhao, Chung Pong Lau, Xiaopeng Hong, and Michael P Pound. T2ICount: Enhancing Cross-modal Understanding for Zero-Shot Counting. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (**CVPR 2025 Highlight**), 2025.
- Chun Tong Lei, Hon Ming Yam, **Zhongliang Guo**, Yifei Qian, and Chun Pong Lau. Instant Adversarial Purification with Adversarial Consistency Distillation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (**CVPR 2025**), 2025.
- Yanli Li, **Zhongliang Guo**, Nan Yang, Huaming Chen, Dong Yuan, and Weiping Ding. Threats and Defenses in the Federated Learning Life Cycle: A Comprehensive Survey and Challenges. *IEEE Transactions on Neural Networks and Learning Systems* (**IEEE TNNLS**), 2025.
- Yifei Qian, Xiaopeng Hong, **Zhongliang Guo**, Ognjen Arandjelović, and Carl R Donovan. Semi-Supervised Crowd Counting with Masked Modeling: Facilitating Holistic Understanding of Crowd Scenes. *IEEE Transactions on Circuits and Systems for Video Technology* (**IEEE TCSVT**), 2024.
- Yifei Qian, Liangfei Zhang, **Zhongliang Guo**, Xiaopeng Hong, and Ognjen Arandjelović. Perspective-assisted Prototype-based Learning for Semi-supervised Crowd Counting. ***Pattern Recognition***, 2025.
- Shuai Zhao, Meihuizi Jia, **Zhongliang Guo**, Leilei Gan, Xiaoyu Xu, Xiaobao Wu, Jie Fu, Feng Yichao, Fengjun Pan, and Anh Tuan Luu. A Survey of Recent Backdoor Attacks and Defenses in Large Language Models. *Transactions on Machine Learning Research* (**TMLR Survey Certificate**), 2025.
- Jiang Liu, Chun Pong Lau, **Zhongliang Guo**, Yuxiang Guo, Zhaoyang Wang, and Rama Chellappa. DiffProtect: Generative Adversarial Examples Using Diffusion Models for Facial Privacy Protection. ***Pattern Recognition***, 2025.
- Man Hu, Yatao Yang, Deng Pan, **Zhongliang Guo**, Luwei Xiao, Deyu Lin, and Shuai Zhao. Syntactic Paraphrase-based Synthetic Data Generation for Backdoor Attacks Against Chinese Language Models. ***Information Fusion***, 2025.
- Yuqi Li, Yanli Li, Kai Zhang, Fuyuan Zhang, Chuanguang Yang, **Zhongliang Guo**, Weiping Ding, and Tingwen Huang. Achieving Fair Medical Image Segmentation in Foundation Models with Adversarial Visual Prompt Tuning. ***Information Sciences***, 2025.
- Shuai Zhao, Yulin Zhang, Luwei Xiao, Xinyi Wu, Yanhao Jia, **Zhongliang Guo**, Xiaobao Wu, Cong-Duy Nguyen, Guoming Zhang, and Anh Tuan Luu. Affective-ROPTester: Capability and Bias Analysis of LLMs in Predicting Retinopathy of Prematurity. *IEEE Transactions on Affective Computing* (**IEEE TAC**), 2025.