Andi Zhang

https://andi.ac az381@cantab.ac.uk (+44) 7481814001

</> Summary

I am currently a Postdoctoral Research Associate at the University of Manchester. I obtained my PhD from the University of Cambridge, where I proposed the **Probabilistic Adversarial Attack**, a principled probabilistic perspective to adversarial attacks. After my PhD, I developed the **Concept-based Adversarial Attack**, which leverages the probabilistic perspective of adversarial attacks to extend traditional single-image adversarial attacks to broader semantic concepts or entities, expanding the applicability of adversarial methods.

Education

Trinity College, University of Cambridge

PhD in Computer Science Oct.2019 - Dec.2024

- Thesis Title: **Anomalous Inputs in Deep Learning: a Probabilistic Perspective**
- Supervisor: Dr. Damon Wischik
- Keywords: Probabilistic Generative Models, Large Language Models, Out-of-distribution Detection, Adversarial Attack

MASt in Mathematics (Part III of the Mathematical Tripos) Oct.2018 - Jun.2019

- Relevant modules: Statistical Learning in Practise, Convex Optimization, Astrostatistics, Bayesian Modelling and Computation, Modern Statistical Methods

MPhil in Advanced Computer Science Oct.2017 - Jun.2018

- Relevant modules: Machine Learning for Natural Language Processing, Algebraic Path, Deep Learning for Natural Language Processing, Advanced Functional Programming, Probabilistic Machine Learning

School of Computer Science, University of Manchester

BSc (Hons) in Computer Science and Mathematics Sep.2014 - Jun.2017

- Relevant modules: Machine Learning, Quantum Computing, Compilers, Convex Optimization, Probability, Statistical Methods, Functional Analysis, Topology, Number Theory

Experience

University of Manchester — Manchester, United Kingdom

Postdoc Research Associate in Machine Learning | Supervisor: Prof. Samuel Kaski Nov.2024 -

- Supported by UKRI Generative AI Hub.
- Concentrating on user modelling and AI safety. Proposed concept-based adversarial attack.

MediaTek Research — Cambridge, United Kingdom

Deep Learning Intern | Host: Dr. Alberto Bernacchia Aug.2024 - Nov.2024

- Participated to a project in accelerating the sampling of diffusion models by proposing a new second-order estimation of the posterior.

University College London — London, United Kingdom

Visitor | Hosts: Dr. David Barber and Dr. Mingtian Zhang July.2023 - Oct.2023

- Worked on a new image generative model that based on Diffusion models.
- Worked on Amortized LoRA, a more principle and probabilistic way to generate LoRA parameters.

Kirin AI Group, Huawei — Cambridge, United Kingdom

Research Intern (Part-time) | Host: Mr. Arnau Raventos Nov.2021 - Oct.2022

- Conducted a comprehensive evaluation of various traditional and AI-based frame interpolation methods, providing insights for the product department to leverage in their decision-making processes.
- Contributed to an AI global illumination project, successfully implemented various samplers within the Mitsuba 3 framework.

Noah's Ark Lab, Huawei — Shenzhen / Beijing, China

Research Intern | Hosts: Dr. Yitong Sun and Dr. Shifeng Zhang Sep.2020 - Sep.2021

- Co-developed the local autoregressive model, which significantly advanced the state-of-the-art in Out-Of-Distribution (OOD) detection and lossless compression. The empirical evidence supporting this enhancement was recognized and published in NeurIPS 2021.
- Contributed to a computer vision project through researching deep learning uncertainty for active learning applications. The outcomes of this work were recognized and published in BMVC 2021.
- Contributed to an indoor positioning project leveraging WiFi signal strength. I specifically conducted research on Deep Gaussian Processes, leading our team to enhance indoor positioning accuracy. The successful results of this work were published in IPIN 2021.

University of Cambridge — Cambridge, United Kingdom

Course Supervisor (Teaching Assistant) — Oct.2019 - Dec.2024

- Supervised IA Algorithms (2020 Lent, 2023 Lent), IA Machine Learning in Real World (2020 Lent), IA Introduction to Probability (2020 Easter, 2021 Easter), IB Further Java (2020 Michaelmas), IB Data Science (2022 Michaelmas, 2023 Michaelmas), IB Further Graphics (2022 Michaelmas) and II Machine Learning and Bayesian Inference (2023 Lent)

MRC Biostatistics Unit, University of Cambridge — Cambridge, United Kingdom

Research Assistant | Hosts: Dr. Sofia Villar and Dr. David Robertson — Jul.2019 - Sep.2019

- Developed a simulation system to tackle the multi-armed bandit problem (MABP) in clinical trial settings and pioneered a unique test statistic for binary response MABP situations. Although the time constraints prevented this work from being formalized into a publication, it provided substantial practical and theoretical insight.

Tencent (WeChat AI) — Beijing, China

R & D Intern | Host: Dr. Jie Zhou — Jul.2018 - Sep.2018

- Engaged in comprehensive exploration and research in deep reinforcement learning and evolution strategies. While this research did not culminate in a tangible output, the experience gained in these cutting-edge fields has proven invaluable.

Amazon (Alexa) — Cambridge, United Kingdom

Applied Scientist Intern | Host: Dr. Emilio Monti — Jun.2017 - Sep.2017

- Employed the Seq2Seq model, the predecessor of the Transformer, with the Attention mechanism for semantic parsing tasks. Semantic parsing involves translating natural language into a structured internal language used for database or knowledge base querying.

Robotics Institute, Carnegie Mellon University — Pittsburgh, United States

Student Intern | Host: Dr. Stelian Coros — Jun.2016 - Sep.2016

- Contributed to the implementation of a general stance controller for 3D-printable robotic creatures. The controller's adaptability enables it to be used with any human-designed multi-foot robots, hence its designation as 'general'.

Publications

Concept-based Adversarial Attack: A Probabilistic Perspective — Arxiv Preprint
Andi Zhang, Xuan Ding, Steven McDonagh, Samuel Kaski
https://arxiv.org/abs/2507.02965

Compositional Attribute Imbalance in Vision Datasets — AAAI 2026
Yanbiao Ma, Jiayi Chen, Wei Dai, Dong Zhao, Zeyu Zhang, Yuting Yang, Bowei Liu, Jiaxuan Zhao, **Andi Zhang**
In the 40th Annual AAAI Conference on Artificial Intelligence (AAAI)

Noise Diffusion for Enhancing Semantic Faithfulness in Text-to-Image Synthesis — CVPR 2025
Boming Miao, Chunxiao Li, Xiaoxiao Wang, **Andi Zhang**, Rui Sun, ZiZhe Wang, Yao Zhu
In the IEEE / CVF Computer Vision and Pattern Recognition Conference (CVPR)

Improving Probabilistic Diffusion Models With Optimal Diagonal Covariance Matching — ICLR 2025 Oral
Zijing Ou[*], Mingtian Zhang[*], **Andi Zhang**, Tim Z. Xiao, Yingzhen Li, David Barber
In the 13th International Conference on Learning Representations

Your Finetuned Large Language Model is Already a Powerful OOD Detector — AISTATS 2025
Andi Zhang, Tim Z. Xiao, Weiyang Liu, Robert Bamler, Damon Wischik
In Proceedings of the 28th International Conference on Artificial Intelligence and Statistics

Constructing Semantics-Aware Adversarial Examples with a Probabilistic Perspective — NeurIPS 2024
Andi Zhang, Mingtian Zhang, Damon Wischik
In the 38th Annual Conference on Neural Information Processing Systems, 2024

SR-OOD: Out-of-Distribution Detection via Sample Repairing — Arxiv Preprint
Rui Sun[*], **Andi Zhang**[*], Haiming Zhang[*], Yao Zhu, Ruimao Zhang, Zhen Li
https://arxiv.org/abs/2305.18228

[*]Equal contribution.

Falsehoods that ML researchers believe about OOD detection **NeurIPS 2022 Workshop**
Andi Zhang, Damon Wischik
In ML Safety Workshop, 36th Conference on Neural Information Processing Systems

Out-of-Distribution Detection with Class Ratio Estimation **NeurIPS 2022 Workshop**
Mingtian Zhang*, **Andi Zhang***, Tim Z. Xiao, Yitong Sun, Steven McDonagh
In ML Safety Workshop, 36th Conference on Neural Information Processing Systems

On the Out-of-distribution Generalization of Probabilistic Image Modelling **NeurIPS 2021**
Mingtian Zhang*, **Andi Zhang***, Steven McDonagh
In Neural Information Processing Systems annual meeting, 2021

Measuring Uncertainty in Signal Fingerprinting with Gaussian Processes Going Deep **IPIN 2021**
Ran Guan, **Andi Zhang**, Mengchao Li, Yongliang Wang
In International Conference on Indoor Positioning and Indoor Navigation, 2021

Towards dynamic and scalable active learning with neural architecture adaption for object detection **BMVC 2021**
Fuhui Tang, Dafeng Wei, Chenhan Jiang, Hang Xu, **Andi Zhang**, Wei Zhang, Hongtao Lu, Chunjing Xu
In British Machine Vision Conference, 2021

Solipsistic Reinforcement Learning **ICLR 2021 Workshop**
Mingtian Zhang*, Peter Noel Hayes*, Tim Z. Xiao, **Andi Zhang**, David Barber
In Self-Supervision for Reinforcement Learning Workshop - International Conference on Learning Representations, 2021

Honors & Awards

Top Reviewer (10%) of AISTATS 2023 Feb.2023

The Williams/Kilburn Medal for outstanding final year student Jul.2017
School of Computer Science, University of Manchester
The medal is awarded on an occasional basis to students of exceptional distinction whose performance throughout their undergraduate course has been of outstanding merit.

University awards for outstanding academic achievement Jul.2017
University of Manchester
The university will consider the top 0.5% of undergraduates (nominated by each school) for this award.

Academic Service

I have served as a reviewer for the following conferences:

- NeurIPS (2024, 2025)
- ICML (2025)
- ICLR (2025, 2026)
- AISTATS (2023, 2024, 2025, 2026)
- CVPR (2025, 2026)
- ICCV (2025)
- ACM MM (2025)
- AAAI (2026)