Jie Zhang

Curriculum Vitae

*CFAR and IHPC, A*STAR*
Singapore
(+65) 87102696
zhang_jie@cfar.a-ster.edu.sg
Personal Homepage

Work Experience

08/2024 – **Research Scientist & Innovation Lead**, *A*STAR Centre for Frontier AI Research (CFAR)*, Singapore, work with Dr. Qing Guo and Prof. Ivor Tsang.

03/2023 – 07/2024 **Research Fellow**, *Nanyang Technological University*, Singapore, work with Prof. Tianwei Zhang and Prof. Yang Liu.

07/2022 – 02/2023 **Postdoc**, *University of Waterloo*, Canada, remote work with Prof. Florian Kerschbaum.

Education

09/2017 – 06/2022 **PhD of Cyber Science and Technology**, *University of Science and Technology of China*, Hefei, China.

09/2013 – 06/2017 **Bachelor of Electrical Engineering and Automation**, *China University of Geosciences*, Beijing, China.

Research Interests

Trustworthy AI & GenAI

Vulnerability Evaluation [TIP 2022], [AAAI 2023], [MM 2023], [AAAI 2024], [AAAI 2024], [AAAI 2024], [CCS 2024], [NeurIPS 2024], [Information Fusion 2024], [USENIX Security 2025], [NAACL 2025], [USENIX Security 2025], [TMM 2025], [CVPR 2025]

Proactive Safeguard [AAAI 2021], [MM 2023], [IJCAI 2024], [ICML 2024], [MM 2024], [NDSS 2025], [AAAI 2025], [ICASSP 2025]

Post-hoc Forensic [AAAI 2020], [NeurIPS 2020], [MM 2020], [TPAMI 2021], [AAAI 2022], [TAI 2023], [Springer Book], [AAAI 2023], [AAAI 2023], [TKDE 2023], [TPAMI 2024], [NDSS 2024], [ICML 2024], [ECCV 2024], [S&P 2025], [TIFS 2025], [ICLR 2025]

Others

Affective Computing [MM 2024], [CVPR 2025]

AI for Science [arXiv 2023]

Selected Recent Works

- ⋆ Guoyin Wang, Shengyu Zhang, Tianyu Zhan, Zhouzhou Shen, Jiwei Li, Xueyu Hu, Xiaofei Sun, Fei Wu, Gelei Deng, **Jie Zhang**, Runyi Hu, Tianwei Zhang, Xiaoya Li, Shuhe Wang, Eduard Hovy, Unlocking the Mysteries of OpenAI o1: A Survey of the Reasoning Abilities of Large Language Models, arXiv, 2025
- ⋆ Shuhe Wang, Shengyu Zhang, **Jie Zhang**, Runyi Hu, Xiaoya Li, Tianwei Zhang, Jiwei Li, Fei Wu, Guoyin Wang, Eduard Hovy, Reinforcement Learning Enhanced LLMs: A Survey, arXiv, 2025
- ⋆ Zhongyi Zhang, **Jie Zhang**, Wenbo Zhou, Xinghui Zhou, Qing Guo, Weiming Zhang, Tianwei Zhang, and Nenghai Yu, FACETRACER: Unveiling Target Identities from Swapped Face Images and Videos, *submitted to TPAMI, Major Revision*, *Corresponding Author.*
- ⋆ Geilei Deng, Haoran Ou, **Jie Zhang**, Tianwei Zhang, and Yang Liu, OEDIPUS: LLM-enchanced Reasoning CAPTCHA Solver, *submitted to ACM CCS 2025*, *Corresponding Author.*
- ⋆ Guanlin Li, Shuai Yang, **Jie Zhang**, and Tianwei Zhang, PRIME: Protect Your Videos From Malicious Editing, *submitted to ICML 2025*.
- ⋆ Weitao Feng, Jiyan He, **Jie Zhang**, Tianwei Zhang, Wenbo Zhou, Weiming Zhang, and Nenghai Yu, Catch You Everything Everywhere: Guarding Textual Inversion via Concept Watermarking, *submitted to TPAMI*, *Corresponding Author.*
- ⋆ Meng Tong, Kejiang Chen, Yuang Qi, **Jie Zhang**, Tianwei Zhang, Weiming Zhang, Nenghai Yu, Privinfer: Privacy-preserving inference for black-box large language model, *Major revision by TDSC*.
- ⋆ Jiyan He, Weitao Feng, Yaosen Min, Jingwei Yi, Kunsheng Tang, Shuai Li, **Jie Zhang**, Kejiang Chen, Wenbo Zhou, Xing Xie, Weiming Zhang, Nenghai Yu, Shuxin Zheng,Control Risk for Potential Misuse of Artificial Intelligence in Science, *submitted to Nature Machine Intelligence*.
- ⋆ Xi Yang, **Jie Zhang**, Chang Liu, Han Fang, Zehua Ma, Kejiang Chen, Weiming Zhang, Nenghai Yu, Synthesizing Glyph Vectors for Practical Information Hiding in Documents, *submitted to TDSC, 2024, Corresponding Author.*
- ⋆ Zhe Lei, **Jie Zhang**, Jingtao Li, Weiming Zhang, and Nenghai Yu, Aparecium: Revealing Secrets from Physical Photographs, *submitted to ICME*, *Corresponding Author.*

Publications (Google Scholar)

Vulnerability Evaluation

- ⋆ **Jie Zhang**, Dongdong Chen, Jing Liao, Qidong Huang, Hua Gang, Weiming Zhang, Nenghai Yu, Poison Ink: Robust and Invisible Backdoor Attack, *IEEE Transactions on Image Processing (TIP), 2022*.
- ⋆ Kunsheng Tang, Wenbo Zhou, **Jie Zhang**†, Aishan Liu, Gelei Deng, Shuai Li, Peigui Qi, Weiming Zhang, Tianwei Zhang, Nenghai Yu, GenderCARE: A Comprehensive Framework for Assessing and Reducing Gender Bias in Large Language Models, *The ACM Conference on Computer and Communications Security (CCS) 2024*, †*Corresponding Author.*

- ⋆ Shuai Li, **Jie Zhang**†, Yuang Qi, Kejiang Chen, Tianwei Zhang,Weiming Zhang, and Nenghai Yu, Clean Image May be Dangerous: Data Poisoning Attacks Against Deep Hashing, *IEEE Transactions on Multimedia (MM), 2025*, †*Corresponding Author.*
- ⋆ Yue Cao, Yun Xing, **Jie Zhang**, Di Lin, Tianwei Zhang, Ivor Tsang, Yang Liu, Qing Guo, SceneTAP: Scene-Coherent Typographic Adversarial Planner against Vision-Language Models in Real-World Environments, *The IEEE / CVF Computer Vision and Pattern Recognition Conference (CVPR), 2025*.
- ⋆ Haolin Wu, Chang Liu, Jing Chen, Ruiying Du, Kun He, Yu Zhang, Cong Wu, Tianwei Zhang, Qing Guo, **Jie Zhang**, Yueqiang Cheng, and Weiming Zhang, When Translators Refuse to Translate: A Novel Attack to Speech Translation Systems, *USENIX Security, 2025*.
- ⋆ Meng Tong, Kejiang Chen, Xiaojian Yuan, Jiayang Liu, Weiming Zhang, Nenghai Yu, **Jie Zhang**, Yueqiang Cheng, and Weiming Zhang, On the Vulnerability of Text Sanitization, *NAACL, 2025*.
- ⋆ Haoxiang Tian, Xingshuo Han, Guoquan Wu, An Guo, Yuan Zhou, **Jie Zhang**, Shuo Li, Jun Wei, Tianwei Zhang, An LLM-empowered Adaptive Evolutionary Algorithm For Multi-Component Deep Learning Systems, *AAAI (Oral), 2025*.
- ⋆ Linqing Hu, Junqi Zhang, **Jie Zhang**, Shaoyin Cheng, Yuyi Wang, Weiming Zhang, Nenghai Yu, Security Analysis and Adaptive False Data Injection against MultiSensor Fusion Localization for Autonomous Driving, *Information Fusion, 2024*.
- ⋆ Junqi Zhang, Shaoyin Cheng, Linqing Hu, **Jie Zhang**, Chenyu Shi, Xingshuo Han, Tianwei Zhang, Yueqiang Cheng, and Weiming Zhang, The Ghost Navigator: Revisiting the Hidden Vulnerability of Localization in Autonomous Driving, *USENIX Security, 2025*.
- ⋆ Guanlin Li, Kangjie Chen, Shudong Zhang, **Jie Zhang**, and Tianwei Zhang , ART: Automatic Red-teaming for Text-to-Image Models to Protect Benign Users, *NeurIPS, 2024*.
- ⋆ Yihao Huang, Felix Juefei-Xu, Qing Guo, **Jie Zhang**, Yutong Wu, Ming Hu, Tianlin Li, Geguang Pu, Yang Liu, Zero-Day Backdoor Attack against Text-to-Image Diffusion Models via Personalization, *AAAI Conference on Artificial Intelligence (AAAI), 2024.*
- ⋆ Xiaojian Yuan, Kejiang Chen, Wen Huang, **Jie Zhang**, Weiming Zhang, Nenghai Yu, Data-Free Hard-Label Robustness Stealing Attack, *AAAI Conference on Artificial Intelligence (AAAI), 2024.*
- ⋆ Yi Xie, **Jie Zhang**, Shiqian Zhao, Tianwei Zhang, Xiaofeng Chen, SAME: Sample Reconstruction Against Model Extraction Attacks, *AAAI Conference on Artificial Intelligence (AAAI), 2024.*
- ⋆ Yanru He, Kejiang Chen, Guoqiang Chen, Zehua Ma, Kui Zhang, **Jie Zhang**, Huanyu Bian, Han Fang, Weiming Zhang, Nenghai Yu, ProTegO: Protect Text Content against OCR Extraction Attack, *ACM MM, 2023*.
- ⋆ Kui Zhang, Hang Zhou, **Jie Zhang**, Qidong Huang, Weiming Zhang, and Nenghai Yu. Ada3Diff: Defending against 3D Adversarial Point Clouds via Adaptive Diffusion, *ACM MM, 2023*.
- ⋆ Xiaojian Yuan, Kejiang Chen, **Jie Zhang**, Weiming Zhang, and Nenghai Yu, Pseudo Label-Guided Model Inversion Attack via Conditional Generative Adversarial Network, *AAAI Conference on Artificial Intelligence (AAAI), 2023*.

Proactive Safeguard

- ⋆ Zhiling Zhang, **Jie Zhang**†, Kui Zhang, Wenbo Zhou, Weiming Zhang, Nenghai Yu, Segue: Side-information Guided Generative Unlearnable Examples for Facial Privacy Protection in Real World, *ICASSP 2025*, †*Corresponding Author.*
- ⋆ Yutong Wu, **Jie Zhang**, Florian Kerschbaum, and Tianwei Zhang, THEMIS: Regulating Textual Inversion for Personalized Concept Censorship, *the Network and Distributed System Security Symposium (NDSS), 2025*,*Corresponding Author.*
- ⋆ Yanghao Su, **Jie Zhang**†, Ting Xu, Tianwei Zhang, Weiming Zhang, Nenghai Yu, Model X-ray: Backdoor Detection for MLaaS via Decision Boundary, *ACM MM 24*, †*Corresponding Author.*
- ⋆ Qidong Huang*, **Jie Zhang***, Wenbo Zhou, Weiming Zhang, Nenghai Yu, Initiative Defense against Facial Manipulation, *AAAI Conference on Artificial Intelligence (AAAI), 2021*, ** Equal Contribution.*
- ⋆ Kui Zhang, Hang Zhou, **Jie Zhang**, Wenbo Zhou, Weiming Zhang, Nenghai Yu,Transferable Facial Privacy Protection against Blind Face Restoration via Domain-Consistent Adversarial Obfuscation, *ICML 24*.
- ⋆ Hanlin Gu, Gongxi Zhu, **Jie Zhang**, Yuxing Han, Lixin Fan, Qiang Yang, Unlearning during Learning: An Streamlined Federated Machine Unlearning Method, *IJCAI 24*.

Post-hoc Forensic

- ⋆ **Jie Zhang**, Dongdong Chen, Jing Liao, Zehua Ma, Han Fang, Weiming Zhang, Hua Gang, Nenghai Yu, Robust Model Watermarking for Image Processing Networks via Structure Consistency, *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), 2024.*
- ⋆ **Jie Zhang**, Dongdong Chen, Jing Liao, Weiming Zhang, Nenghai Yu, "Digital Watermarking for Machine Learning Models - Chapter 6: Protecting Image Processing Networks via Model Watermarking", *Springer book, 2023.*
- ⋆ **Jie Zhang**, Dongdong Chen, Jing Liao, Weiming Zhang, Hua Gang, Huamin Feng, Nenghai Yu, Deep Model Intellectual Property Protection via Deep Watermarking, *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), 2021.*
- ⋆ **Jie Zhang**, Dongdong Chen, Jing Liao, Weiming Zhang, Hua Gang, Nenghai Yu, Passport-aware Normalization for Deep Model Protection, *Advances in Neural Information Processing Systems (NeurIPS), 2020.*
- ⋆ **Jie Zhang**, Dongdong Chen, Jing Liao, Han Fang, Weiming Zhang, Wenbo Zhou, Hao Cui, Nenghai Yu, Model Watermarking for Image Processing Networks, *AAAI Conference on Artificial Intelligence (AAAI), 2020.*
- ⋆ Runyi Hu, **Jie Zhang**†, Yiming Li, Jiwei Li, Qing Guo, Han Qiu, and Tianwei Zhang, VideoShield: Regulating Diffusion-based Video Generation Models via Watermarking, *ICLR 2025*, †*Corresponding Author.*

- ⋆ Boheng Li, Yanhao Wei, Yankai Fu, Zhenting Wang, Yiming Li, **Jie Zhang**†, Rui Wang, and Tianwei Zhang, Towards Reliable Verification of Unauthorized Data Usage in Personalized Text-to-Image Diffusion Models, *S&P 2025*, †*Corresponding Author.*
- ⋆ Runyi Hu, **Jie Zhang**†, Ting Xu, Jiwei Li, Tianwei Zhang, Robust-Wide: Robust Watermarking against Instruction-driven Image Editing, *ECCV 24*, †*Corresponding Author.*
- ⋆ Weitao Feng, Wenbo Zhou, Jiyan He, **Jie Zhang**†, Tianyi Wei, Guanlin Li, Tianwei Zhang, Weiming Zhang, and Nenghai Yu, AquaLoRA: Toward White-box Protection for Customized Stable Diffusion Models via Watermark LoRA, *ICML 24*, †*Corresponding Author.*
- ⋆ Chang Liu, **Jie Zhang**†, Tianwei Zhang, Xi Yang, Weiming Zhang, and Nenghai Yu, Detecting Voice Cloning Attacks via Timbre Watermarking, *the Network and Distributed System Security Symposium (NDSS), 2024.*,†*Corresponding Author.*
- ⋆ Haozhe Chen, **Jie Zhang**†, Kejiang Chen, Weiming Zhang, Nenghai Yu, Model Access Control Based on Hidden Adversarial Examples for Automatic Speech Recognition, *IEEE Transactions on Artificial Intelligence, 2023,* †*Corresponding Author.*
- ⋆ Xi Yang*, **Jie Zhang***, Han Fang, Zehua Ma, Chang Liu, Weiming Zhang, and Nenghai Yu, AutoStegaFont: Synthesizing Vector Fonts for Hiding Information in Documents, *AAAI Conference on Artificial Intelligence (AAAI), 2023*, ** Equal Contribution.*
- ⋆ Chang Liu*, **Jie Zhang***, Han Fang, Zehua Ma, Weiming Zhang, and Nenghai Yu, DeAR: A Deep-learning-based Audio Re-cording Resilient Watermarking, *AAAI Conference on Artificial Intelligence (AAAI), 2023*, ** Equal Contribution.*
- ⋆ Xi Yang*, **Jie Zhang***, Kejiang Chen, Weiming Zhang, Zehua Ma, Feng Wang, Nenghai Yu, Tracing Text Provenance via Context-aware Lexical Substitution, *AAAI Conference on Artificial Intelligence (AAAI), 2022*, ** Equal Contribution.*
- ⋆ Shuai Li, Kejiang Chen, Kunsheng Tang, Wen Huang, **Jie Zhang**, Weiming Zhang, Nenghai Yu. Turning Your Strength into Watermark: Watermarking Large Language Model via Knowledge Injection, *TIFS*, 2025
- ⋆ Zhiwen Ren, Han Fang, **Jie Zhang**, Zehua Ma, Ronghao Lin, Weiming Zhang, Nenghai Yu, A Robust Database Watermarking Scheme That Preserves Statistical Characteristics, *IEEE Transactions on Knowledge and Data Engineering (TKDE), 2023*.
- ⋆ Haozhe Chen, Hang Zhou, **Jie Zhang**, Dongdong Chen, Weiming Zhang, Kejiang Chen, Nenghai Yu, Perceptual Hashing of Deep Convolutional Neural Networks for Model Copy Detection, *ACM Transactions on Multimedia Computing Communications and Applications (TOMM), 2022*.
- ⋆ Kunlin Liu, Dongdong Chen, Jing Liao, Weiming Zhang, Hang Zhou, **Jie Zhang**, Wenbo Zhou, Nenghai Yu, JPEG Robust Invertible Grayscale, *IEEE Transactions on Visualization and Computer Graphics (TVCG), 2021.*
- ⋆ Xiquan Guan, Weiming Zhang, Huaming Feng, Hang Zhou, **Jie Zhang**, Nenghai Yu, Reversible Watermarking in Deep Convolutional Neural Networks for Integrity Authentication, *Proceedings of the 28th ACM International Conference on Multimedia (ACM MM), 2020.*

- ⋆ Han Fang, Dongdong Chen, Qidong Huang, **Jie Zhang**, Weiming Zhang, Nenghai Yu, Deep Template-based Watermarking, *IEEE Transactions on Circuits and Systems for Video Technology (TCSVT), 2020.*

Others

- ⋆ Ruiqi Wang, Jinyang Huang, **Jie Zhang**†, Xin Liu, Xiang Zhang, Zhi Liu, Peng Zhao, Sigui Chen, and Xiao Sun, FacialPulse: An Efficient RNN-based Depression Detection via Temporal Facial Landmarks, *ACM MM 24*, *Oral (3.97%), †Corresponding Author.*
- ⋆ Wenbo Zhou, Dongdong Chen, Jing Liao, **Jie Zhang**, Kejiang Chen, Weiming Zhang, Nenghai Yu, Attribute-Aware Head Swapping Guided by 3d Modeling, *EEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2024*.

Collaboration

- ⋆ Nils Lukas, Assistant Professor at MBZUAI, Sep 2023 – present
- ⋆ Florian Kerschbaum, Professor in the David R. Cheriton School of Computer Science at the University of Waterloo, Sep 2022 – present
- ⋆ Dongdong Chen, Senior Researcher at Microsoft Research, Sep 2019 – present
- ⋆ Jing Liao, Assistant Professor with the Department of Computer Science, City University of Hong Kong (CityU), Sep 2019 – present
- ⋆ Gang Hua, Vice President and Chief Scientist of Wormpex AI Research, Feb 2020 – present

Services

- ⋆ Reviewer for ICML, ICLR, NeurIPS, AAAI, IJCAI, CVPR, ICCV, ECCV, ACM CCS, ACM MM, etc.
- ⋆ Reviewer for TPAMI, IJCV, TIP, TIFS, TDSC, TMM, TCSVT, SPL, etc.

Awards & Honors

2024 Distinguished Artifact Award, CCS, 2024

2021 National Scholarship for Doctoral Students, China

2020 Cyberspace Science Scholarship (funded by Academician Xiaomo Wang), China

Grants

01/2020–12/2021 **Research on Intellectual Property (IP) Protection for Deep Models**, *leader*, the Fundamental Research Funds for the Central Universities, No. WK5290000001.

11/2019–10/2021 **Research on the Mechanism of Attack and Defense for Deep Models**, *student leader*, the Exploration Fund Project of University of Science and Technology of China under Grant, No. YD3480002001.

1/2021–12/2024 **Research on Basic Theory and Key Technology of Attack and Defense Analysis for Deep Models**, *student leader*, the Natural Science Foundation of China under Grant, No. U20B2047.

Projects

06/2019–06/2020 **Research on Intellectual Property (IP) Protection for Medical Image Processing Models**, *leader*, with Pvmed.

09/2020–02/2021 **Research on Intellectual Property (IP) Protection for Products Data**, *student leader*, with JD.COM.

10/2021–10/2022 **Research on Security Assessment of Automatic Driving Models**, *student leader*, with NIO.

Interests

Sports, Hiking, Traveling, Reading