

# ZHANHAO HU

Personal Homepage: <https://whothu.github.io/>  
Soda Hall, University of California, Berkeley, CA, 94720  
+1 341-333-8522 ◇ zhanhaohu.cs@gmail.com

## EDUCATION

---

**Tsinghua University, Beijing** 2017 - 2023

Ph.D., Computer Science and Technology

Advisor: Bo Zhang and Xiaolin Hu

Dissertation: The Practicality of Physical Adversarial Examples for Deep Learning Models.

**Tsinghua University, Beijing** 2013 - 2017

B.S., Mathematics and Physics

Advisor: Xiaolin Hu

Dissertation: STDP-based Learning for Spiking Neural Networks

## RESEARCH EXPERIENCE

---

**University of California, Berkeley** January 2024 - Present

*Postdoctoral Researcher* *Berkeley, California*

- I'm affiliated with the Institute for Data Science (BIDS) and the Electrical Engineering and Computer Science (EECS) department, and I am advised by Prof. David Wagner. I focus on safety/security issues in Large Language Models (LLMs), including red-teaming, detecting, and defending against existing threats to these models.

**Tsinghua University** July 2023 - December 2023

*Researcher* *Beijing, China*

- I worked with Prof. Xiaolin Hu at the Tsinghua Laboratory of Brain and Intelligence (THBI). My research primarily focused on robustness/security issues in Computer Vision (CV) models, such as adversarial examples and privacy.

## PUBLICATIONS

---

(Sorted by time; \* for equal contribution)

### *Under review & Preprint*

1. Dennis Jacob, Emad Alghamdi\*, **Zhanhao Hu\***, Basel Alomair, and David Wagner (2025). Better Privilege Separation for Agents by Restricting Data Types (arxiv)
2. Xiaopei Zhu, Guanning Zeng, **Zhanhao Hu**, Jun Zhu, and Xiaolin Hu (2025). Multimodal Physical Adversarial Clothing Evades Visible-Thermal Detectors with Non-Overlapping RGB-T Pattern (under review)

3. Qiongxiu Li, Lixia Luo, Agnese Gini, Changlong Ji, **Zhanhao Hu**, Xiao Li, Chengfang Fang, Jie Shi, Xiaolin Hu (2024). Perfect Gradient Inversion in Federated Learning: A New Paradigm from the Hidden Subset Sum Problem (arxiv)

## **Published**

4. **Zhanhao Hu\***, Xiao Huang\*, Patrick Mendoza, Emad A. Alghamdi, Basel Alomair, Raluca Ada Popa, and David Wagner (2025). GradShield: Alignment Preserving Finetuning (Accepted by ICLR 2026)
5. Jesson Wang\*, **Zhanhao Hu\***, and David Wagner (2025). JULI: Jailbreak Large Language Models by Self-Introspection (Accepted by ICLR 2026)
6. Julien Piet, Xiao Huang, Dennis Jacob, Annabella Chow, Maha Alrashed, Geng Zhao, **Zhanhao Hu**, Chawin Sitawarin, Basel Alomair, and David Wagner (2025). Jailbreaksovertime: Detecting jailbreak attacks under distribution shift (AISeC)
7. Dennis Jacob, Hend Alzahrani, **Zhanhao Hu**, Basel Alomair, and David Wagner (2025). PromptShield: Deployable Detection for Prompt Injection Attacks (CODASPY)
8. Xiaopei Zhu, Siyuan Huang, **Zhanhao Hu**, Jianmin Li, Jun Zhu, Xiaolin Hu (2025). Physical Adversarial Examples for Person Detectors in Thermal Images Based on 3D Modeling. (TPAMI)
9. **Zhanhao Hu**, Julien Piet, Geng Zhao, Jiantao Jiao, and David Wagner (2024). Toxicity Detection for Free. Advances in Neural Information Processing Systems (Neurips **Spotlight**)
10. Zhi cheng, **Zhanhao Hu**, Yugu Liu, jianmin Li, Hang Su, Xiaolin Hu (2024). Full-Distance Evasion of Pedestrian Detectors inthe Physical World. Advances in Neural Information Processing Systems (Neurips)
11. Xiao Li, Wei Zhang, Yining Liu, **Zhanhao Hu**, Bo Zhang, Xiaolin Hu (2024). Language-Driven Anchors for Zero-Shot Adversarial Robustness. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).
12. Xiao Li, Qiongxiu Li, **Zhanhao Hu**, Xiaolin Hu (2024). On the Privacy Effect of Data Enhancement via the Lens of Memorization. IEEE Transactions on Information Forensics and Security (IEEE TIFS).
13. Xiaopei Zhu, Yuqiu Liu, **Zhanhao Hu**, Jianmin Li, and Xiaolin Hu (2024) Infrared Adversarial Car Stickers ()
14. Xiaopei Zhu, **Zhanhao Hu**, Siyuan Huang, Jianmin Li, Xiaolin Hu (2023). Hiding from Infrared Detectors in Real World with Adversarial Clothes. Applied Intelligence.
15. **Zhanhao Hu**, Wenda Chu, Xiaopei Zhu, Hui Zhang, Bo Zhang, Xiaolin Hu (2023). Physically Realizable Natural-Looking Clothing Textures Evade Person Detectors via 3D Modeling. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).

16. Tong Wang, Xiaohui Kuang, Qianjin Du, **Zhanhao Hu**, Huan Deng, Gang Zhao. (2023). Driving into Danger: Adversarial Patch Attack on End-To-End Autonomous Driving Systems Using Deep Learning. In 2023 IEEE Symposium on Computers and Communications (ISCC)
  17. **Zhanhao Hu**, Jun Zhu, Bo Zhang, Xiaolin Hu (2022). Amplification trojan network: Attack deep neural networks by amplifying their inherent weakness. Neurocomputing, 505, 142-153.
  18. **Zhanhao Hu**, Siyuan Huang, Xiaopei Zhu, Fuchun Sun, Bo Zhang, Xiaolin Hu (2022). Adversarial texture for fooling person detectors in the physical world. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR **Oral**).
  19. Xiaopei Zhu, **Zhanhao Hu**, Siyuan Huang, Jianmin Li, Xiaolin Hu (2022). Infrared invisible clothing: Hiding from infrared detectors at multiple angles in real world. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR **Oral**).
  20. **Zhanhao Hu**, Tao Wang, Xiaolin Hu (2017). An stdp-based supervised learning algorithm for spiking neural networks. In Neural Information Processing: 24th International Conference (ICONIP).

#### **TEACHING EXPERIENCE**

## Teaching Assistant

- Neural and Cognitive Computation (No.80240642), Tsinghua University *2019 Autumn*
  - Neural and Cognitive Computation (No.80240642), Tsinghua University *2018 Autumn*

## **PROFESSIONAL SERVICE**

Reviewer

- Journal Reviewer: TIP, TPAMI, TNNLS
  - Conference Reviewer: Neurips, ICML, CVPR, ICLR, ICCV, ECCV, AAAI, ICIST, ICACI, ICICIP, ISNN