李潇

(+86) 188-1096-3683 | xiaoli.cst@gmail.com | lixiaothu.github.io | Google Scholar

教育背景

清华大学

在读博士生　计算机科学与技术 *2020 年 9 月 - 至今*

- 导师: 胡晓林教授、张钹院士
- TSAIL组 (张钹院士及朱军教授带领)
- 研究兴趣: 我的研究目标旨在建立可信的人工智能系统，希望人工智能模型能够在各类极端情况下工作良好，能达到接近或超越人类智能的水平。基于这个目标，我探索了包括对抗机器学习、表征学习、大脑和认知功能启发的学习、可扩展多模态学习、以及扩散生成模型在内的主题。我最近的研究兴趣集中在提高基础模型的可信性，包括大型文本到图像和文本到视频的扩散生成模型以及大型语言模型。

清华大学

工学学士　计算机科学与技术 *2016 年 9 月 - 2020 年 6 月*

- 辅修**统计学** (清华统计学研究中心)

实习经历

华为 2012 实验室

可信机器学习算法实习生 *2023 年 7 月-至今*

- 指导: 时杰博士
- 研究课题: 文本到图像扩散生成模型，大型语言模型，通用目标检测算法等的可信属性
- 工作内容: 围绕华为公司的盘古系列大模型，终端相机的目标检测算法等企业智能业务和计算机视觉领域前沿开展相关算法的预研工作，包括对文生图扩散模型和大型语言模型的可信属性等问题进行研究和洞察前瞻，发表高水平论文并申请专利
- 产出: 一篇关于利用扩散模型防御对抗样本 (ADBM) 的一作在投论文，两篇分别关于大语言模型和目标检测算法安全性的拟投稿一作论文，两项在审专利
- 两次受邀华为 2012 实验室内部繁星论坛讲座:
  1) *Insights into Security Risks of the Diffusion (Text-to-Image) Generative Models*
  2) *Adversarial Suffix Prompt Attacks and Defenses on Large Language Models*

清华大学

研究生助教 *2021 年 8 月-2023 年 2 月*

- 工作内容: 协助老师讲授 pytorch 等深度学习编程基础，设计符合课程要求的作业，批改答疑等
- 深度学习训练营, 主讲: 胡晓林老师, 2021 年 8 月 - 2023 年 2 月
- 神经与认知计算 (THU-80240642), 主讲: 胡晓林老师, 2022 秋季学期
- 深度学习导论 (THU-00240332), 主讲: 胡晓林老师, 2021 秋季学期

Momenta自动驾驶

计算机视觉算法实习生 *2018 年 8 月-2018 年 12 月*

- 指导: 李翔教授
- 研究课题: 人眼视线识别、智慧报警系统、多任务学习、持续学习等
- 工作内容: 围绕自动驾驶场景辅助司机防止疲劳驾驶等业务需求进行相关计算机视觉算法的研发和迭代，主要包括视频监控场景下的司机视线方向识别等问题，结合多任务学习、持续学习、注意力机制等前沿技术持续提升识别精度和识别效率
- 产出: 最终我的方案在公司内部举办的挑战赛中获得第一名，并后续被集成落地

论文发表

更多论文见Google Scholar.

Preprint & Under review

- **Xiao Li**, Wenxuan Sun, Huanran Chen, Qiongxiu Li, Yining Liu, Yingzhe He, Jie Shi, Xiaolin Hu. *Adversarial Diffusion Bridge Model for Reliable Adversarial Purification*. *Submitted to ICML 2024*.
- **Xiao Li**, Yining Liu, Na Dong, Sitian Qin, Xiaolin Hu. *PartImageNet++ Dataset: Scaling up Part-based Models for Robust Recognition*. *Submitted to ECCV 2024*.
- **Xiao Li**, Hang Chen, Xiaolin Hu. *On the Importance of Backbone to the Adversarial Robustness of Object Detectors*. *Under review & Preprint*. arXiv:2305.17438. *Submitted to IEEE TIFS 2024*.
- Hang Chen, Chufeng Tang, **Xiao Li**, Lesi Wei, Xiaolin Hu. *Improving Neuron Segmentation in Electron Microscopy with Affinity-Guided Queries*. *Submitted to ECCV 2024*.
- Qiongxiu Li, Lixia Luo, Agnese Gini, Zhanhao Hu, **Xiao Li**, Chengfang Fang, Xiaolin Hu, Jie Shi. *On the Hardness of Input Reconstruction Attack via Gradient Sharing in Federated Learning: A Cryptographic View*. *Under review*.
- Wei Zhang, Zhanhao Hu, **Xiao Li**, Xiaopei Zhu, Xiaolin Hu. *Adversarial Patch Defenses Give a False Sense of Security for Physical Defense: Circumventing Defenses with a Single Set of Clothes*. *Under review*.

Published & Accepted

- **Xiao Li**, Wei Zhang, Yining Liu, Zhanhao Hu, Xiaolin Hu. *Language-Driven Anchors for Zero-Shot Adversarial Robustness*. **CVPR 2024 (CCF A)**.
- **Xiao Li**, Qiongxiu Li, Zhanhao Hu, Xiaolin Hu. *On the Privacy Effect of Data Enhancement via the Lens of Memorization*. **IEEE TIFS 2024 (CCF A)**.
- **Xiao Li**, Ziqi Wang, Bo Zhang, Fuchun Sun, Xiaolin Hu. *Recognizing Object by Components with Human Prior Knowledge Enhances Adversarial Robustness of Deep Neural Networks*. **IEEE TPAMI 2023 (CCF A)**.
- Hang Chen, **Xiao Li**, Zefan Wang, Xiaolin Hu. *Robust logo detection in e-commerce images by data augmentation*. **ACM MM Workshop**.
- Xiaolin Hu, Chufeng Tang, Hang Chen, **Xiao Li**, Jianmin Li, Zhaoxiang Zhang. *Improving Image Segmentation with Boundary Patch Refinement*. **IJCV 2022 (CCF A)**.
- Chufeng Tang, Hang Chen, **Xiao Li**, Jianmin Li, Zhaoxiang Zhang, Xiaolin Hu. *Look closer to segment better: Boundary patch refinement for instance segmentation*. **CVPR 2021 (CCF A)**.
- Xiaopei Zhu, **Xiao Li**, Jianmin Li, *et al*. *Fooling thermal infrared pedestrian detectors in real world using small bulbs*. **AAAI 2021 (CCF A)**.

荣誉奖励

- 2023 年　综合优秀奖学金 (博士)，　清华大学
- 2020 年　鲁棒商标检测挑战赛，　第五名/36489 队，　ACM MM & 阿里巴巴
- 2018 年　深度学习视线跟踪挑战赛，　第一名/23 队，　Momenta
- 2018 年　科技创新优秀奖 (本科)，　清华大学
- 2018 年　学业优秀奖学金 (本科)，　清华大学
- 2017 年　学业优秀奖学金 (本科)，　清华大学

学术服务

- 会议审稿人: AAAI 2023, CVPR 2023, NeurIPS 2023, AAAI 2024, CVPR 2024, ECCV 2024 等
- 期刊审稿人: IEEE TPAMI, IEEE TIFS, IEEE TIP 等

编程技能

编程语言　Python / C&C++ / Java / Matlab / R / Nodejs 等

开源框架　PyTorch / mmdetection / detectron2 / diffusers 等

开发平台　Macos / Linux / Windows 等

智能工具　ChatGPT / Claude 3 / Github Copilot 等