手机/微信：(+86) 15397310001
邮箱：zhaoyuchen20@fudan.edu.cn
地址：上海市杨浦区邯郸路 220 号

陈兆宇

Google Scholar
Home Page

教育背景

山东大学 **计算机科学与技术（本科）** 2016 年 9 月到 2020 年 7 月

荣誉奖项：CET-4，CET-6，一等奖学金（前 5%），山东省优秀毕业生 **GPA：**87.93/100.0（5/101）

复旦大学 **计算机应用技术（直博生）** 2020 年 9 月至今

研究方向：AI 安全、多模态大模型、AIGC、计算机视觉 **GPA：**3.63/4.0（7/50）

专业技能：常用 C++和 Python，了解 Pytorch 框架；熟悉 AI 安全，包括对抗样本、模型窃取、版权保护、大模型安全评估等；熟悉 BERT、GPT、CLIP、LLaVA、Stable Diffusion 等大模型，了解 LoRA、Adapter、ControlNet 等微调方法。

荣誉奖项：华泰证券科技奖学金（前 1%），一等奖奖学金（前 5%），一等优秀博士候选人奖学金（前 3%）

论文专利：发表 CVPR、ICCV、AAAI、TIFS 等 CCF A 类会议/期刊和 SCI 一区期刊 22 篇，其中一作和共一论文 8 篇，并申请国家发明专利 5 项；担任 TPAMI、CVPR、NeurIPS、TNNLS 等 19 个顶会和顶刊的审稿人，谷歌学术引用 627；

实习经历

深圳市腾讯计算机系统有限公司 CSIG-优图实验室-盘古研究中心（基础研究实习生） 2021 年 4 月到 2024 年 4 月

- 业务：为活体检测和人脸篡改检测等人脸安全业务提供数字域和物理域的对抗样本，评估业务模型的对抗鲁棒性；
- 科研：为业务提供对抗攻防的技术预研，调研和总结人脸安全的相关技术，对物理攻击和黑盒攻击进行落地实践；

维沃移动通信有限公司（vivo） 影像算法研究部-质量增强算法中心（助理算法工程师） 2024 年 5 月至今

- 基于多模态大模型，实现面向摄影场景的 Agent，使其可以根据用户反馈和拍摄场景给出拍摄参数、模式等建议；

科研经历

基于生成模型的内容生成：利用最新的生成模型来实现图像编辑、图像水印和图像去噪等功能。

- *Content-based Unrestricted Adversarial Attack* 以第一作者发表于 NeurIPS 2023, 论文借助 BLIP v2 构造文本, 基于 Stable Diffusion 将图像映射到隐变量空间，用跳接梯度来对抗优化图像，实现轻微编辑，引入 ControlNet 后扩展至 TPAMI 在投；
- *VideoPure: Diffusion-based Adversarial Purification for Video Recognition* 以共同通讯作者在投于 IEEE TIFS 期刊，该工作基于 ModelScopeT2V，实现了时序 DDIM 反演和时空优化来重建视频和去除对抗噪声，借助投票机制提高去噪性能；
- *Cmua-watermark: A Cross-model Universal Adversarial Watermark for Combating Deepfakes* 以第三作者发表于 AAAI 2022，基于对抗噪声和多目标优化，首次提出了一个跨模型通用的图像水印，能够保护人脸不受 deepfake 模型的篡改；

基于基础视觉模型的鲁棒性研究：面向预训练视觉模型和多模态大模型进行对抗鲁棒性的评估。

- *Improving Adversarial Transferability of VLP Models through Collaborative Multimodal Interaction* 以共同第一作者获得 CVPR 2024 Workshop: 视觉基础模型黑盒攻击第一名，基于模态交互和对抗文本在 GPT-4V、Qwen-VL 等大模型上测试。
- *Towards Practical Certifiable Patch Defense with Vision Transformer* 以第一作者发表于 CVPR 2022，该工作首次将 ViT 引入可信防御，针对随机平滑的可信图块防御机制，提出多阶段渐进式重建预训练来提升可信鲁棒性（~26%↑）；
- *Boosting the Transferability of Adversarial Attacks with GMI* 以共同第一作者发表于 ESWA 期刊（中科院一区）；

无数据场景下的知识蒸馏：探究隐私安全场景下，如何更好地提升轻量化模型的正常性能和对抗鲁棒性。

- *Sampling to Distill: Knowledge Transfer from Open-World Data* 以第二作者发表于 ACMMM 2024；
- *Out of Thin Air: Exploring Data-free Adversarial Robustness Distillation* 以共同第一作者发表于 AAAI 2024；

面向视频模型的对抗鲁棒性分析：在针对视频任务的特点，在白盒和黑盒设置下提出了对应的对抗鲁棒性评估方法。

- *Efficient Decision-based Black-box Patch Attacks on Video Recognition* 以共同第一作者发表于 ICCV 2023；
- *Towards Decision-based Sparse Attacks on Video Recognition* 以共同第一作者发表于 ACMMM 2023（Oral）；
- *Exploring the Adversarial Robustness of VOS via One-shot Adversarial Attacks* 以共同通讯作者发表于 ACMMM 2023；

面向深度视觉模型的对抗图块：从形状和位置的角度研究对抗图块的攻击性能，并分析不同视觉架构的鲁棒性。

- *Shape Matters: Deformable Patch Attack* 以第一作者发表于 ECCV 2022，基于几何关系提出了可微形变的优化策略；
- *Query-Efficient Decision-based Black-Box Patch Attack* 以第一作者发表于 IEEE TIFS 期刊（CCF A，中科院一区）；

竞赛经历

CVPR24 Workshop：视觉基础模型的黑盒攻击第一名 CVPR21 Workshop：防御模型的白盒对抗攻击（9/1681）

CVPR21 Workshop：ImageNet 无限制对抗攻击（10/1599） 2019 年 ACM ICPC 南昌邀请赛铜牌