王添毅

13514685743（微信同号） zjuwty@outlook.com 博士

基本信息

研究方向：对抗攻击与防御、目标检测、大模型安全 **求职意向：*****

个人主页：https://hill-wu-1998.github.io/ **Github：**Hill-Wu-1998

教育背景

浙江大学（985 工程） **控制科学与工程（博士研究生）** **2023.9-2026.7（预计毕业）**

- 隶属于孙优贤院士课题组，导师为百人计划研究员王聪老师，合作导师包括陈积明教授、程鹏教授和舒元超教授等

浙江大学（985 工程） **电子信息（硕士研究生）** **2021.9-2023.7（转博）**

郑州大学（211 工程） **自动化（本科）** **2016.9-2020.7**

科研成果

[CVPR 25] Learning Robust and Hardware-Adaptive Object Detectors against Latency Attacks for Edge Devices

CCF-A 类会议，一作，已录用

[AAAI 25] Fed-DFA: Federated Distillation for Heterogeneous Model Fusion through the Adversarial Lens

CCF-A 类会议，三作，已录用

[ICDCS 25] Pipelining Multi-DNN Inference on Heterogeneous Mobile Processors under Co-Execution Slowdown

CCF-B 类会议，三作，已录用

[ICCV 25] Exploiting the Hungarian Matching Loss in Detection Transformers for Fun and Profit

CCF-A 类会议，一作，在审

[KDD 25] Learning Pairwise Federated Distillation Online via Bandits with Hidden Context for Heterogeneous Model Fusion

CCF-A 类会议，二作，在审

[INFOCOM 24] Reverse Engineering Industrial Protocols Driven By Control Fields **CCF-A 类会议，五作，已录用**

[IEEE TSE] Better Pay Attention Whilst Fuzzing **CCF-A 类期刊，六作，已录用**

漏洞：CNVD 证书 17 项，均为第一贡献人，其中高危 6 项，中危 11 项

科研经历

2025.3-2025.4 **大模型安全攻防研究**

- 研究大模型中的提示注入攻击（prompt injection attack）。
- 研究推理类大模型中的越狱攻击（jailbreak attack），能否根据思维链的推理过程设计更有效的越狱攻击。
- 能否通过对抗训练、对齐等技术提升模型鲁棒性，增加大模型输出可信度、避免输出敏感问题等。

2024.1-2025.3 **跨架构目标检测模型鲁棒性分析与优化技术**

- 分析 CNN-based YOLO 模型在延迟攻击下的鲁棒性，发现被攻击样本中的对象区域存在**天然鲁棒性**，依此设计注意力区域加入对抗训练算法；使用 Nvidia Nsight 相关工具进行性能分析，发现不同设备在延迟攻击下存在**性能瓶颈迁移现象**，利用此现象提出**硬件自适应对抗训练算法**维持延迟攻击下的推理实时性（Jetson Orin NX 推理从 13 FPS 恢复到 43 FPS），同时维持模型精确度（相较于原始模型 mAP 提升 28.1%-58.8%）。
- 针对 DETR 系列模型存在的**独特攻击面**设计一种新的对抗攻击方法，基于攻击中发现的特殊现象调整攻击使得**攻击算法与攻击目标对齐**；**复现**针对 Attention 机制进行对抗攻击的工作，并结合我们的攻击发现一些有意思的现象。
- 科研产出：CVPR 一篇，ICCV 一篇在投

2023.9-2024.12 **面向分布式端侧系统的异构模型融合优化技术**

- 通过对抗攻击方法探测异构模型决策边界，利用 PGD 攻击对最近边界点的**动态性进行建模**；优化损失函数使得**蒸馏过程关注接近决策边界的样本**，支持卷积与 Transformer 模型混合分布式架构，将异构模型的分布式训练精度提升 0.5%-3.5%。
- 提出了一种**多任务混合模型并行推理机制**，该机制采用动态规划和负载均衡双层优化策略，旨在减少处理器间内存带宽竞争开销，实现模型的流水线并行推理。在包含 ARM CPU、OpenCL GPU、华为 DaVinci NPU 等异构多核处理器的并行任务环境中进行测试，对于麒麟 990、高通骁龙等多架构 SoC，该机制可将推理速度提升 2-8 倍，**显著提升了多任务在端侧的混合推理效率**。
- 科研产出：AAAI 一篇，ICDCS 一篇

自我评价

善于快速理解和分析问题，非常擅长复现领域内相关文章。同时具备良好的团队协作精神与极佳的沟通能力。