AI相关

深度伪造技术

背景

深度伪造(Deepfake)通常指的是使用深度学习进行涉及人脸和人声的多媒体伪造技术，如果被恶意滥用会给社会带来灾难。深度伪造不仅限于面部的替换，还有修改面部特征、修改表情、唇形同步、姿势变换、完整脸生成、篡改音频到视频以及文本到视频等方式。人类面部在社会、政治、经济等方面的敏感性，使得深度伪造技术威胁着社会和个人的安全。

深度伪造算法

自解码器

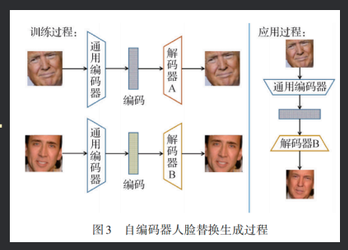

图3　自编码器人脸替换生成过程

生成对抗网络(GAN)

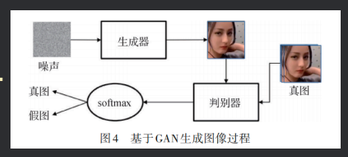

图4　基于GAN生成图像过程

扩散模型

检测技术

图6　深度伪造检测技术分类方法及类别

挑战

检测可行路径

一些想法

攻击 — 针对AI智能体的对抗性深度伪造攻击 — 能欺骗人类和AI

防御 — 面向智能体环境的深度伪造动态防御系统

- 框架
  - 多维特征提取层：利用频域解耦技术提取动态特征，同时结合空间域的物理逻辑（如光影、纹理异常）和时序逻辑（如帧间不一致性）进行初步研判
  - 多模态对齐校验层：通过多模态对比学习，校验图像、音频与文本语义之间的逻辑一致性
  - 动态反馈闭环：根据检测到的伪造概率动态调整智能体的置信度阈值，甚至触发"主动验证"行为
- 可能的实现路径
  - 频域解耦的动态特征提取 — 不直接处理RGB像素，而是通过离散余弦变换（DCT）等手段将图像/视频信号分解为不同频段
  - 多模态逻辑一致性校验
    - 视听同步检测
    - 物理规律校验
    - 语义注入防御
  - 动态反馈与主动验证机制 — 置信度自适应：系统根据检测结果实时计算一个“伪造概率”。当概率达到中间值时，触发Agent的主动验证行为
- 可能的创新点 — 动作-感知一致性校验：“环境反馈验证"。例如，若Agent执行了开灯指令，但视觉感知中的光影变化不符合物理模型，则判定输入环境为伪造