面向 DeepFake 伪造模型溯源的逃避攻击

吴梦洁[1], 于佳艺[1], 汪　润[1], 叶　茜[1], 张钰洋[1], 蔺琛皓[2], 方黎明[3], 王丽娜[1]

[1]武汉大学国家网络安全学院 空天信息安全与可信计算教育部重点实验室 武汉 中国 430072

[2]西安交通大学 西安 中国 710049

[3]南京航空航天大学 南京 中国 210016

摘要 近年来, 深度伪造技术(DeepFake)的泛滥引起了公众和知名人士的极大警觉。这些高度逼真的伪造图像以及视频可能大规模传播虚假信息, 对声誉造成伤害, 甚至可能引发社会动荡。为了应对生成的伪造图像及视频, DeepFake 取证领域的研究得到了广泛关注。在当前的 DeepFake 取证研究中, DeepFake 检测技术负责判断给定样本真实与否, 而 DeepFake 溯源技术则旨在追溯生成该类 Deepakes 的伪造模型类型, 为 DeepFake 检测提供更具解释性的结果。具体而言, DeepFake 溯源可以分为模型-架构溯源和模型-实例溯源两类, 其中模型-架构溯源仅推断使用的具体模型架构, 而模型-实例溯源则试图识别具有特定训练设置的模型实例。而无论模型-架构溯源还是模型-实例溯源方法, 都依赖于识别 DeepFake 生成过程中留下的特定痕迹, 精明的攻击者可以破坏或篡改这些痕迹, 从而使得溯源技术失效。本文观察到, 用于模型溯源的特定痕迹同时存在于高频分量和低频分量中, 并在溯源过程中起着不同的作用。基于此, 本文首次提出一种无训练的逃避攻击方法——TraceEvader, 并在最符合现实环境的无盒场景下进行了测试。具体来说, TraceEvader 将从原始 DeepFakes 中学习到的通用模仿痕迹注入到高频分量中, 并在低频分量中引入对抗性模糊, 以混淆某些痕迹的提取过程, 从而逃避模型溯源。本文对 4 种最先进的模型溯源技术进行了实验, 评估其在 8 种生成模型(包括生成对抗网络(Generative Adversarial Networks, GAN)和扩散模型(Diffusion Models, DM))生成的伪造图像上的表现。结果表明, TraceEvader 实现了 79%的最高平均攻击成功率, 并且在面对图像转换和专业去噪技术时依然表现出了良好的鲁棒性, 平均攻击成功率保持在 75%左右。TraceEvader 证实了当前模型溯源技术的局限性, 并提醒 DeepFakes 研究人员和从业者探索更强大的模型溯源技术。

关键词 深度伪造; 深度伪造溯源; 对抗攻击; 伪造人脸

中图法分类号 TP391 DOI 号 10.19363/J.cnki.cn10-1380/tn.2026.03.05

Evading Attacks for DeepFake Fake Model Traceability

WU Mengjie[1], YU Jiayi[1], WANG Run[1], YE Xi[1], ZHANG Yuyang[1], LIN Chenhao[2], FANG Liming[3], WANG Lina[1]

[1] Key Laboratory of Aerospace Information Security and Trusted Computing, Ministry of Education, School of Cyber Science and Engineering, Wuhan University, Wuhan 430072, China

[2] Xi'an Jiaotong University, Xi'an 710049, China

[3] Nanjing University of Aeronautics and Astronautics, Nanjing 210016, China

Abstract In recent years, the proliferation of DeepFakes has caused great alarm among the public and prominent figures. These highly realistic fake images and videos can spread disinformation on a large scale, cause reputational harm, and may even trigger social unrest. In order to deal with the generated fake images and videos, the research in the field of DeepFake forensics has been widely concerned. In the current DeepFake forensics research, DeepFake detection technology is responsible for judging whether a given sample is true or not, while DeepFake traceability technology aims to trace the type of counterfeit model that generates such Deepakes, so as to provide more explanatory results for DeepFake detection. Specifically, DeepFake traceability can be divided into model-schema traceability and model-instance traceability, where model-schema traceability only inferences the specific model schema used, while model-instance traceability attempts to identify model instances with specific training Settings. Both model-architecture and model-instance traceability methods rely on identifying specific traces left by the generation of deepfakes that savvy attackers can destroy or tamper with, rendering the traceability techniques ineffective. It is observed that specific traces used for model traceability exist in both high-frequency and low-frequency components and play different roles in the traceability process. Based on this, this paper proposes an untrained attack evading method—TraceEvader for the first time, and tests it in the most practical non-box

通讯作者: 汪润, 博士, 副教授, Email: wangrun@whu.edu.cn。

本课题得到国家重点研发计划青年科学家项目(No. 2021YFB3100700)、国家自然科学基金项目(No. 62202340、No. 62372334)、河南省网络空间态势感知重点实验室开放课题基金重点项目(No. HNTS2022004)、武汉市知识创新计划项目(No. 2022010801020127)、中央高校基本科研业务费专项(No. 2042023kf0121)、CCF-绿盟科技“鲲鹏”科研基金(No. CCF-NSFOCUS 2023005)资助。本文的早期会议版本发表在 AAAI 2024 (https://doi.org/10.1609/aaai.v38i18.29973)

收稿日期: 2024-08-22; 修改日期: 2024-10-29; 定稿日期: 2026-01-26

setting. Specifically, TraceEvader injects generic imitation traces learned from the original DeepFakes into the high-frequency component and introduces adversarial ambiguity into the low-frequency component to obfuscate the extraction process of certain traces, thereby evading model traceability. In this paper, we experiment with four state-of-the-art model traceability techniques and evaluate their performance in eight generative models, including Generative Adversarial Networks (GANs) and Diffusion Models (DMs) generate representations on forged images. The results show that TraceEvader achieves the highest average attack success rate of 79%, and still shows good robustness in the face of image conversion and professional denoising techniques, and the average attack success rate remains around 75%. TraceEvader confirms the limitations of current model traceability techniques and reminds DeepFakes researchers and practitioners to explore more powerful model traceability techniques.

Key words DeepFake; DeepFake attribution; adversarial attack; forged face

1 引言

随着生成模型(Generative Models, GM)如 GAN[1]和 DM[2-3]的迅速发展, DeepFakes 已经成为严峻的挑战, 生成模型可以合成高度逼真的音频、图像和视频[4-5]。这些 DeepFake 可以被恶意用于制造虚假信息、生成色情内容, 甚至发布虚假官方声明[6]等。因此, 防止恶意 DeepFake 的滥用和传播已成为当务之急[7-8]。

在当前的 DeepFake 取证研究中, DeepFake 检测技术[9-10]负责判断给定样本真实与否, 而 DeepFake 溯源技术则旨在追溯生成该类 Deepake 的伪造模型类型, 为 DeepFake 检测提供更具解释性的结果。本文探讨了一个关键问题: 现有的 DeepFake 溯源技术是否能够胜任 DeepFake 取证工作, 并且是否具备在实际应用场景中部署的潜力。

截至目前, DeepFake 溯源可以分为模型-架构溯源[10]和模型-实例溯源[11]两类。具体而言, 模型-架构溯源仅推断使用的具体模型架构(例如, CycleGAN[12], StyleGAN[1]), 而模型-实例溯源则试图识别具有特定训练设置的模型实例, 如特定训练数据集和初始种子等条件下训练的模型。最近的研究表明, 无论模型-架构溯源还是模型-实例溯源, 都依赖于识别 DeepFake 生成过程中留下的特定痕迹[10-11]。因此, 本文探索了一种实用的痕迹破坏方法, 旨在有效地逃避黑盒设置下这两种不同类型的模型溯源技术。图 1 展示了本文的工作框架: 上半部分中, 攻击者创建并发布 DeepFake, 防御者则追踪其来源并归因于特定伪造模型用于取证。下半部分中, 攻击者通过在发布前注入难以察觉的对抗性扰动, 成功逃避溯源技术。这种逃避攻击方法旨在通过注入不可见的对抗性扰动来欺骗溯源技术。

先前的研究[13-14]大多集中于通过干扰真伪鉴别边界的关键痕迹特征来揭示 DeepFake 检测器的脆弱性, 这些痕迹特征应当能够泛化多种伪造模型, 特别是未知的伪造模型。然而, 模型溯源的痕迹与具体模型相关, 每种模型架构都有其稳定的痕迹, 并且显著的痕迹因模型实例的不同而有所变化。因此, 专为 DeepFake 检测器设计的逃避攻击并不适用于模型溯源, 因为它们依赖的痕迹大相径庭。

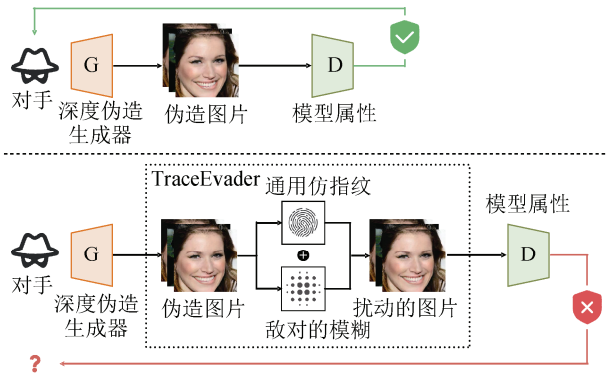


图 1 工作框架概述

Figure 1 An overview of our framework

在黑盒设置中逃避 DeepFake 溯源技术的一个主要方法是通过访问代理模型或发送大量样本查询受害者模型来生成对抗性扰动[15-16]。然而, 这两种方法在处理未知模型时都表现出较差的迁移性能[15, 17]。因此, 本文探索采用通用对抗性扰动[18-19]破坏黑盒设置中的痕迹, 并以此来逃避目标模型溯源技术。首先, 需要探索一个关键问题: 是什么让 DeepFake 具有可溯源性?

初步实验结果表明, 伪造模型在图像的高频分量和低频分量中都留下了痕迹。高频分量(High-Frequency Component, HFC)包含与模型架构相关的全局一致痕迹, 而低频分量(Low-Frequency Component, LFC)中的痕迹集中在特定区域并与模型的权重相关。这两种痕迹被广泛用于模型溯源。

根据上述观察, HFC 和 LFC 中的痕迹特征差异显著, 本文分别对 HFC 和 LFC 引入了对抗性扰动。具体而言, 对于 HFC, 受常见混淆攻击的启发, 本文从常见生成模型创建的一组伪造图像中学习到了一种通用模仿痕迹(Universal Imitated Trace, UIT), 用于在提取模型溯源痕迹时混淆 DeepFake 溯源技术。

受先前工作[20]的影响, 模糊是一种有效的图像劣化方式, 可以减轻 DeepFake 与真实图像之间的空间和频率差异。本文使用高斯模糊均值偏移来消除 LFC 域中细微的痕迹变化。值得注意的是, 本文提出的逃避攻击是**无训练的**, 不需要进行烦琐的梯度计算优化; 在**无盒**设置下, 无需获得目标 DeepFake 溯源技术的任何具体知识; 适用于常见的生成模型(如 GAN 和 DM); 对伪造模型的具体架构和实例溯源具有很强的**迁移能力**。

本文对 4 种最先进的 DeepFake 溯源技术, 6 种 GAN 模型和 2 种 DM 模型生成的伪造图像进行了深入评估。结果显示, 本文提出的 TraceEvader 成功地欺骗了两种类型的 DeepFake 溯源技术, 平均攻击成功率(Attack Success Rate, ASR)超过了 79%, 并且在面对图像转换和去噪技术时依然表现出色。这表明当前的 DeepFake 溯源技术在很大程度上依赖于模型架构之间的稳定痕迹和模型实例之间的局部语义痕迹。本文的研究为未来用于 DeepFake 取证的模型溯源技术提出了新的挑战, 需要寻找更先进的痕迹检测方法来识别 DeepFake 生成过程中引入的痕迹。

本文主要贡献总结如下:

1) 据我们所知, 这是第一次尝试揭示现有 DeepFake 溯源技术的脆弱性, 强调了为了更好地追溯伪造模型的来源和完成深度伪造取证, 需要开发更稳健的模型溯源技术。

2) 深入分析了在 DeepFake 生成过程中引入的用于模型溯源的模型痕迹, 发现了 HFC 中的稳定痕迹可用于区分模型架构, 而 LFC 中的局部语义痕迹则可用于认证模型实例。

3) 提出了一种新的逃避攻击 TraceEvader, 通过在 HFC 中注入伪造的通用模仿痕迹和引入高斯模糊平均位移来消除 LFC 中的痕迹, 以防止模型溯源中准确痕迹的提取。

4) 在对 4 种先进的 DeepFake 溯源技术进行的实验中, TraceEvader 能够有效躲避现有的模型溯源技术, 适用于基于 GAN 和基于 DM 的伪造模型, 并且能够应对常见的图像变换和针对性的去噪技术。

2 模型溯源

本节将探讨三个关键问题: 模型可被溯源依赖于什么样的痕迹, 这些痕迹存在于哪里? 以及这些痕迹具有什么样的特征? 深入理解这些问题有助于生成专门的对抗性扰动, 以破坏用于 DeepFake 模型溯源的痕迹。本文以 GAN 为例来回答这三个问题。

2.1 DeepFake 的溯源机制

研究表明, GAN 具有稳定的特征, 可用于区分两个 GAN(例如, ProGAN 和 StyleGAN)是否具有相同的架构, 不同的特征可以用来区分经过特定设置训练的不同 GAN 实例。

2.1.1 模型架构特征

模型-架构溯源指的是将伪造图像溯源到特定的模型结构。这驱使我们深入挖掘模型本身结构在生成过程中留下的痕迹。在生成模型中, 生成器使用一组卷积滤波、上采样和非线性组件来学习特定的数据分布, 从而完成图像合成任务。先前的研究表明, 特定的上采样策略(生成器中广泛采用的组件)可能导致高频分量中的周期性或棋盘状伪影[21-22]。这些规则伪影在 GAN 中被用于模型溯源, 其频谱差异已被证明能有效区分不同的 GAN 架构(例如 DNA-Det[10]、DCT[23]、Reverse[24])。图 2 展示了生成器之间频谱伪影的示例。

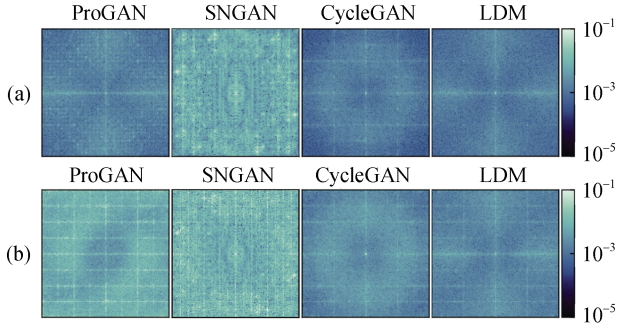


图 2 常见生成模型的频谱分析(a)生成的原始假图像和(b)添加了 TraceEvader 生成的对抗性扰动的图像的平均 DFT 谱

Figure 2 Spectrum analysis for popular generative models. (a) generated raw fake images and (b) images added with our generated adversarial perturbations

2.1.2 模型实例特征

近期的研究(例如 AttNet[11])声称, 模型痕迹的特征是每个模型实例在训练后参数收敛到不同值的模型权重。由于 GAN 训练的不稳定性, 即使是具有相同架构的两个 GAN 实例在各自的训练设置下也可能产生不完全相同的合成输出, 表现出参差不齐的图像质量。这表明, 尽管它们共享相同的模型体系结构, 但仍存在模型实例特有的微妙空间痕迹。

2.2 HFC 和 LFC 模型指纹

如前所述, 相同的模型体系结构具有共享的稳定指纹, 而不同的模型实例则具备不同的模型指纹。本节探讨了这些指纹的具体位置, 它们在模型溯源中的作用, 并进一步利用这些模型指纹的特征进行

有效的对抗性干扰。

为了回答上述问题, 通过在 HFC 中引入劣化, 并将现有模型溯源方法关注的区域进行了可视化, 并进行了几个初步实验。实验考虑了四种常见的模型溯源技术, 包括模型-架构溯源方法 DNA-Det[10]、DCT[23]、Reverse[24], 以及模型-实例溯源方法 AttNet[11]。

2.2.1 HFC 分离实验

根据现有研究的观点[25], HFC 在深度神经网络分类中具有重要作用。受此启发, 本文研究了 HFC 和 LFC 在模型溯源方法中的分类结果。利用之前研究[25]中提出的快速傅里叶变换(Fast Fourier Transform, FFT)将伪造图像的信息分解为 HFC 和 LFC。具体来说, 通过放弃 HFC, 仅使用逆 FFT 重建图像 $\boldsymbol{x}_l$, 不包括通常采用的整个组件。我们有如下等式:

$$\begin{aligned}\mathbf{LFC},\mathbf{HFC} &= f\left(\mathcal{F}(\boldsymbol{x});r\right)\\ \boldsymbol{x}_l &= \mathcal{F}^{-1}(\mathbf{LFC})\end{aligned} \tag{1}$$

其中, $\mathcal{F}$ 表示快速傅里叶变换, $\boldsymbol{x}$ 为原始输入, $\mathcal{F}^{-1}$ 表示输入重构的逆 FFT, $f(\cdot;r)$ 表示一个阈值函数, 它将 HFC 和 LFC 从 $\mathcal{F}(\boldsymbol{x})$ 中分离出来, 具有指定的超参数半径 r, 其中 r 越大意味着保留的频带越宽。

如图 3(a)所示, 当 r 较大时, 四种常见的溯源模型 $\boldsymbol{x}_l$ 的平均准确率接近 100%。进一步降低 r 时, 除了 AttNet 之外, 其他三种溯源技术的性能显著下降。这表明 HFC 和 LFC 中的痕迹对模型溯源都非常重要。接下来, 进一步探讨 HFC 和 LFC 中痕迹的作用及其特征。

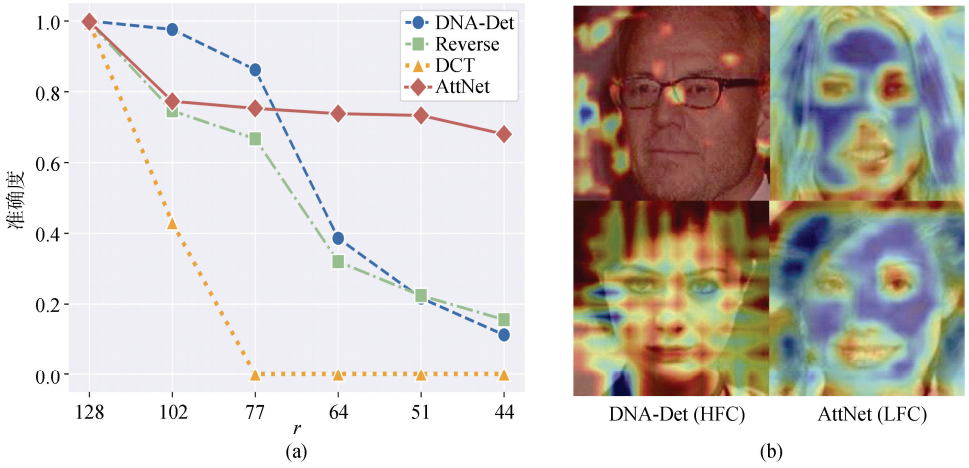


图 3 (a) 四种模型溯源技术在处理 HFC 去除中的表现(b)基于 HFC 的 DNA-Det(左)方法和基于 LFC 的 AttNet(右)方法的分级图

Figure 3 (a) The performance of four model attribution techniques in dealing with HFC removal (b) Grad-cam of DNADet(left) method relying on HFC and AttNet(right) relying on LFC for attribution

2.2.2 注意力可视化

图 3(b)展示了两种模型溯源技术(即 DNA-Det 和 AttNet)关注区域的可视化结果。AttNet 主要依赖于 LFC 中的痕迹, 倾向于关注具有丰富语义信息的局部区域, 如眼睛和嘴巴。相比之下, 像 DNA-Det 这样的模型溯源方法更依赖于 HFC, 注意力在整个图像上的均匀分布。因此, 本文有理由相信 LFC 中的痕迹与语义相关, 而 HFC 则包含更多底层模式。最近的研究[10]也指出, GAN 架构可能会在整个图像中留下全局一致的痕迹, 而不同区域的权重痕迹则有所不同。

综上所述, 本文的初步实验结果证明:

1) 图像的 HFC 和 LFC 都为模型溯源提供了有用的痕迹;

2) HFC 中关注底层模式的痕迹更适合于模型-架构属性, 而 LFC 中关注语义层模式的痕迹更适合于模型-实例属性;

3) 模型-架构痕迹在整个图像中呈现出全局一致性, 权重痕迹则集中在语义相关的局部区域。

这些发现表明, 对抗性扰动应该能够同时破坏 HFC 和 LFC 中的痕迹, 因为未知的溯源方法可能会同时采用它们。

3 技术方法

3.1 威胁模型

3.1.1 防御者的目标和能力

防御者的目标是追踪采用的伪造模型, 并为

DeepFake 取证提供解释性证据。防御者的主要策略包括:

1) 利用现有的伪造图像来训练模型溯源模型, 用于追踪已知的伪造模型;

2) 执行简单的图像变换或利用专门的去噪方法, 通过收集少量干净和攻击样本对来有意识地去除可能添加的对抗性扰动。

3.1.2 攻击者的目标和能力

攻击者的目标是生成各种逼真的 DeepFake, 并提高在应对可能的 DeepFake 取证方法(如被动检测、伪造模型溯源等)的能力。特别是, 攻击者通常是模型的创建者, 对训练数据集有充分了解。为了避免被各种潜在的模型溯源技术追踪, 攻击者可能会删除涉及模型修改或合成输出的痕迹, 从而成功欺骗模型溯源技术。

3.2 扰动模型痕迹

本文的研究结果表明, HFC 中的模型指纹具有全局一致性, 而 LFC 中的模型指纹则集中于局部区域, 这两者之间存在显著的差异。因此, 设计统一的扰动来破坏 HFC 中的稳定模型指纹和 LFC 中的显著指纹至关重要。本文提出了两种精心设计的对抗性扰动, 分别添加到图像的 HFC 和 LFC 中。HFC 中的通用模仿痕迹(UIT)来破坏模型-架构溯源的模型指纹, 并在 LFC 中引入均值偏移模糊来防止提取模型-实例溯源的模型指纹。图 4 给出了逃避模型溯源技术的两个精心制作的对抗性扰动, 首先将图像 x 的 HFC 和 LFC 分离, 分别添加对抗性扰动。对于 x_{HFC}, 通过注入通用模仿痕迹 F 来实现 UIT 攻击; 对于 x_{LFC}, 采用对抗性模糊攻击。然后利用混合图像变换将两个分量组合在一起, 而不产生明显的伪影。

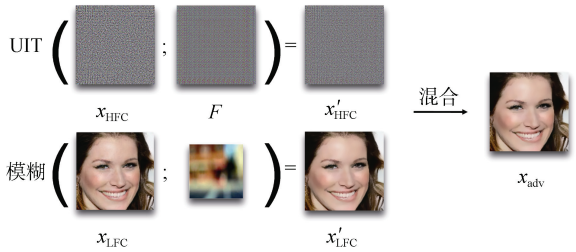


图 4 TraceEvader 制造的对抗性扰动

Figure 4 Adversarial perturbations crafted by our TraceEvader

3.2.1 对抗性模仿痕迹

受到混淆攻击的启发[26], 在所有权验证中, 攻击者将类似的水印嵌入到受害者模型中以声明其所有权, 这样, 目标模型的所有权就变得模糊不清, 因为各自的存在都可以作为证据。本文旨在制作模仿痕迹, 以混淆模型溯源技术, 如 DNN 模型中的所有权验证中的模糊水印。

残差显示了与生成模型和额外随机噪声相关的丰富指纹信息[27]。受先前工作的启发[28-29], 本文认为 CNN 在拟合无序随机噪声之前倾向于学习自然图像中的通用和结构化特征, 并且由于其卷积和采样操作在生成器架构中广泛应用, 自然地嵌入了指纹先验信息[30]。直观地说, 通过这种CNN的归纳偏置, 可以从残差中学习到通用和周期性的指纹模式作为 UIT。具体而言, 给定图像空间 $\mathcal{X}\in\mathbb{R}^{3\times h\times w}$, 用 $\boldsymbol{\phi}:\mathbb{R}^{c\times h\times w}\to\mathbb{R}^{3\times h\times w}$ 表示生成过程, 该过程利用 CNN 编码器-解码器网络[30] $\boldsymbol{\phi}$ 将潜在空间映射到图像空间。给定 n 个图像的小批量 $\boldsymbol{X}=\left[\boldsymbol{x}_1^{c_1},\ldots,\boldsymbol{x}_n^{c_n}\right]$, 其中 $c_i\in\{0,1\}$ 表示生成图像和真实图像, 残差 $\boldsymbol{R}$ 通过高通滤波器[31] f_{HP} 提取, 记为 $\boldsymbol{R}=f_{\mathrm{HP}}(\boldsymbol{X})$。本文的目标是生成通用痕迹 $\boldsymbol{F}\in\mathbb{R}^{3\times h\times w}$, 使其与生成图像残差中的痕迹有较强的相关性, 而与真实图像的干净残差相关性较弱。

$$\theta^*=\underset{\theta}{\operatorname{argmin}}\,L_{\mathrm{con}}\left(\boldsymbol{\phi}(\boldsymbol{z};\theta);\boldsymbol{R}\right)\quad \boldsymbol{F}=\boldsymbol{\phi}\left(\boldsymbol{z};\theta^*\right) \tag{2}$$

其中, θ^* 由参数 θ 优化而成, 随机潜向量 $\boldsymbol{z}$ 服从均匀分布。

为了衡量图像残差与 $\boldsymbol{F}$ 的相关性, 本文采用 Pearson 相关系数作为相关度量 $\rho(\cdot)$, 其公式为

$$\rho(\boldsymbol{F},\boldsymbol{R})=\frac{E\left[(\boldsymbol{F}-\mu_{\mathrm{F}})(\boldsymbol{R}-\mu_{\mathrm{R}})\right]}{\sigma_{\mathrm{F}}\sigma_{\mathrm{R}}} \tag{3}$$

其中, σ 和 μ 分别代表标准差和平均值, $E(\cdot)$ 计算数学期望。

本文设计了一个对比损失函数, 以促进来自不同类别的残差在与 $\boldsymbol{F}$ 的相关性中存在更大的差异, 从而提高 $\boldsymbol{F}$ 与来自同一类别的残差之间的相关性。本文利用欧几里得距离来量化两个残差样本之间的相关距离 $\mathcal{D}_{ij}$:

$$\mathcal{D}_{ij}=\left\|\rho\left(\boldsymbol{F},\boldsymbol{R}_i^{c_i}\right)-\rho\left(\boldsymbol{F},\boldsymbol{R}_j^{c_j}\right)\right\| \tag{4}$$

其中, $c_i\in\{0,1\}$ 为伪造图像和真实图像。当 $c_i=c_j$ 时, $\mathcal{D}_{ij}$ 表示为 $\mathcal{D}_{\mathrm{pos}}$; 当 $c_i\neq c_j$ 时, $\mathcal{D}_{ij}$ 表示为 $\mathcal{D}_{\mathrm{neg}}$。对比损失 L_{con} 总结如下:

$$L_{\mathrm{con}}=\frac{1}{2}D_{\mathrm{pos}}^2-\frac{1}{2}\left(\max\left\{0,m-\mathcal{D}_{\mathrm{neg}}\right\}\right)^2 \tag{5}$$

其中, m 是预定义的边距参数。

3.2.2 对抗性均值偏移模糊

作为一种自定义的噪声模式, 通用模仿痕迹(UIT)对图像的低频和语义方面影响有限。由于其全局一致性, UIT 以均匀的方式干扰图像, 这在改变整体像素分布和去除重要局部痕迹方面的效果较差。此外, 目标模型溯源技术所利用的感兴趣区域痕迹是未知的, 并且在不同的生成模型中有所变化。因此, 很难制作通用痕迹来混淆 LFC 中的痕迹。本文采用一种直接的方法, 即引入模糊来破坏 LFC 中的模型痕迹。高斯模糊平均移位(Gaussian Blurring Mean Shift, GBMS)是一种有效的边缘保持滤波器, 可以平滑低频分量, 消除合成图像的缺陷或畸变伪影[20]。

GBMS 记作 $G_\sigma(\cdot)$, 根据像素的颜色信息计算每个像素在空间域中的位置密度, 并将颜色相似的像素聚集在一起, 从而改变像素分布以达到平滑效果。具体来说, 首先为图像 x 中的每个像素 p_i 定义一个搜索窗口 W (本文工作中将窗口大小设置为 3)。然后用高斯核函数 $\boldsymbol{g}_\sigma$ 估计每个像素 $p_j \in W$ 的密度, 表示为

$$\boldsymbol{g}_\sigma\left(p_i, p_j\right)=\exp\left(-\frac{1}{2}\left(\left\|p_i-p_j\right\| / \sigma\right)^2\right) \tag{6}$$

其中, σ 表示 $\boldsymbol{g}$ 的宽度, $\|\cdot\|$ 表示欧几里得距离。GBMS 使用平均位移向量 dp 更新当前像素 p_i 的位置:

$$dp=\frac{\sum_{p_j \in W}\left(p_j-p_i\right)\boldsymbol{g}_\sigma\left(p_i, p_j\right)}{\sum_{p_j \in W}\boldsymbol{g}_\sigma\left(p_i, p_j\right)} \tag{7}$$

$$p_i'=p_i+dp \tag{8}$$

重复上述过程, 直到每个像素达到收敛。这种对抗性平均偏移模糊方法旨在破坏 LFC 中的模型痕迹, 提供了一种有效的方法来干扰模型-实例溯源技术的痕迹提取过程。

3.3 混合图像变换

在为 HFC 和 LFC 分别准备了对抗性扰动之后, 以混合图像变换将这些对抗性扰动添加到伪造图像中。具体操作是分离图像的 HFC 和 LFC, 对这两个分量施加攻击, 然后将这两个扰动分量组合成最终的对抗性混合图像。

首先, 将 UIT 注入到 HFC 的残差中, 以产生难以区分的模糊痕迹。这个过程可以表述为

$$x_{\mathbf{HFC}}=\lambda \cdot\left[f_{\mathrm{HP}}(x)+\boldsymbol{F}\right] \tag{9}$$

其中, λ 是控制整体扰动强度的权重因子。

其次, 使用 GBMS 函数对图像的低频部分进行模糊处理:

$$x_{\mathbf{LFC}}=G_\sigma\left(x-f_{\mathbf{HP}}(x)\right) \tag{10}$$

最后, 生成对抗性混合图像 x_{adv}, 用于逃避模型溯源技术:

$$\begin{aligned} x_{\mathrm{adv}} &= x_{\mathrm{HFC}}+x_{\mathrm{LFC}} \\ &= G_\sigma\left(x-f_{\mathrm{HP}}(x)\right)+\lambda \cdot\left[f_{\mathrm{HP}}(x)+\boldsymbol{F}\right] \end{aligned} \tag{11}$$

4 实验

4.1 实验设置

在研究中, 为了全面评估提出的 TraceEvader 方法的有效性和鲁棒性, 进行了多个实验。首先是有效性实验, 评估在应用 TraceEvader 后是否会影响其功能表现。接着是对常见图像变换和有意图像去噪情况下的鲁棒性实验。此外, 还与四种不同的基线方法进行了对比评估。本文的实验涵盖了两种类型的模型溯源技术(模型-架构溯源和模型-实例溯源), 针对来自 8 种深度伪造模型(例如 GAN 和 DM)的高度多样化数据进行详尽的探究。实验采用两种对抗性攻击作为基线, 一种是基于迁移的攻击, 包括 BIM[32] 和 MI-FGSM[33]; 另一种是无盒攻击, 如 Vera22-peak[13], TN-Net[34]和 FakePolisher[35]。

4.1.1 模型溯源技术

实验涵盖模型-架构溯源和模型-实例溯源两种类型的模型溯源技术, 评估了 TraceEvader 在四种先进的模型溯源技术上的效果。具体来说, 针对模型-架构溯源方面, 本文评估了 DNA-Det[10]、Reverse[24]和 DCT[23]; 针对模型-实例溯源方面, 评估了 AttNet[11]。为了准确评估 TraceEvader 的性能, 对来自 8 个深度伪造模型的数据进行了微调, 以确保在模型溯源方面表现出最佳性能。表 1 展示了对这四种模型溯源技术在 8 种深度伪造模型中的评估结果。

4.1.2 数据集

本文针对不同模型进行攻击的数据集来自 DNA-Det[10]以及 DiffDetect[1]中的测试集。

4.1.3 评价指标

本文使用攻击成功率(ASR)和图像质量指标峰值信噪比(PSNR)来评估 TraceEvader 的有效性。PSNR 被广泛用于衡量添加对抗性扰动后的图像质量, 较高的数值表示更好的图像质量。

4.1.4 基线方法

本文将 TraceEvader 与两种常用的对抗性攻击进行比较, 一种是基于迁移的攻击(即 BIM[32] 和 MIFGSM[36]), 另一种是无盒攻击包括 FakePolisher[35], TN-Net[34]和 Vera22-peak[13]。具体来说, Vera22-peak

表 1 四种模型溯源方法在 6 个 GAN 和 2 个 DM 溯源中的表现

Table 1 The performance of four model attribution methods in attributing the six GANs and two DMs

模型类型	数据集来源	模型&数据集	DNA-Det ↑	Reverse ↑	DCT ↑	AttNet ↑
GAN	DNA-Det[10]	ProGAN[36]	0.987	0.8487	0.469	0.81
		MMDGAN[37]	1.0	1.0	0.996	1.0
		SNGAN[38]	0.94	0.997	0.994	0.99
		CramerGAN[39]	1.0	0.995	0.985	1.0
		CycleGAN[12]	1.0	0.423	0.924	0.575
		StyleGAN2[1]	0.998	0.779	0.811	0.97
DMs	DiffDetect[1]	PNDM[3]	1.0	0.574	0.953	0.302
		LDM[40]	1.0	0.248	0.633	0.978

通过移除频域中异常周期峰值的频率系数阈值来进行, 而 FakePolisher 则通过重建图像以去除痕迹。TN-Net 设计了包含一个生成器和多个判别器的痕迹移除网络, 旨在移除了深度伪造的空间异常、频谱差异和噪声指纹。

4.1.5 实现细节

对于基线实验, 在表 2 中指定了实现 Vera22-peak 的最佳阈值, 这些阈值是根据文献[14]确定的, 以获得最佳性能。为了实现 BIM 和 MI-FGSM, 设定 $\epsilon = 8/255$, 最大迭代 $T = 40$, 在 MI-FGSM 中采用 $\mu = 1.0$, 本文直接使用了 Fakepolisher 作者提供的模型权重。

在 UIT 生成过程中, 本文采用基于图 5 所示的 U-Net 架构的网络 ϕ。具体的架构细节如表 3 所示。编码器由四个编码卷积模块构成, 解码器包含四个解码卷积模块和最后的卷积层。

表 2 每个 GM 在峰值攻击中的最优阈值

Table 2 The optimal threshold in the peak attack for each GM

GMs	ProGAN	MMDGAN	SNGAN	CramerGAN	CycleGAN	StyleGAN2	PNDM	LDM
阈值	0.65	0.25	0.4	0.2	0.7	0.5	0.95	0.6

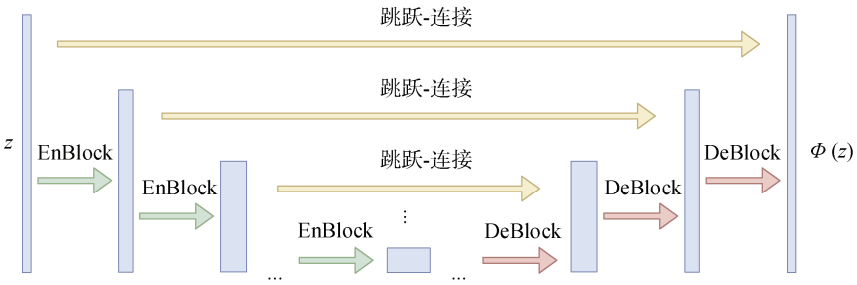


图 5 用于学习单元的网络体系结构

Figure 5 The architecture of the network employed for learning UITs

表 4 详细列出了所有模块的结构。卷积层的核大小表示为[高度,宽度, 步幅]。c_i 和 c_o 代表输入和输出通道, 其中的"+"表示输入包括跳跃连接。生成器的输入是一个 16 通道的潜码 z。输出列描述输出形状, Conv 表示卷积层, BN 表示批归一化, LReLU 表示 Leaky ReLU 激活函数, SC 表示跳跃连接。h 和 w 是输入形状。f_{HP} 是来自先前研究[41-42]的预训练 DnCNN, 其中权重在生成模拟痕迹期间被冻结。预定义的边距 m 设置为 0.01。学习率设为 5×10^{-4}, 并使用 Adam 优化器进行优化。

在本文的实验中, UIT 是从一个包含 1000 张伪造图像和 1000 张真实图像的训练数据集中学习得到的。在注入 UIT 的过程中, 权重因子 λ 设置为 0.01, 在高斯平均位移中, 权重因子 σ 设置为 8。经过视觉检查, 35dB 的 PSNR 值定为伪造痕迹不可见的下限, 本文设置上述参数值以确保对抗性样本的 PSNR 值大于 35dB, 保证攻击的隐蔽性。

4.2 有效性实验

本节将攻击黑盒设置中的模型-架构和模型-实例溯源技术评估 TraceEvader 的有效性, 并将其与四

表 3 编码器和解码器的网络结构

Table 3 Network architecture of the encoder and decoder

结构	层	c_i	c_o
编码器	z	16	16
	EnBlock	16	32
	EnBlock	32	64
	EnBlock	64	128
	EnBlock	128	256
解码器	DeBlock	256+256	128
	DeBlock	128+128	64
	DeBlock	64+64	32
	DeBlock	32+32	32
	最后一层	32	3

表 4 编码块, 解码块和最后一层的架构

Table 4 EnBlock, DeBlock and Final_Layer architecture

结构	层	内核	输出
EnBlock	Conv	[3,3,1]	$c_o\times h\times w$
	BN,LReLU	-	$c_o\times h\times w$
	Conv	[3,3,1]	$c_o\times h\times w$
	BN,LReLU	-	$c_o\times h\times w$
	Maxpool	[2,2,2]	$c_o\times h/2\times w/2$
	SC	-	$2c_i\times h\times w$
DeBlock	Conv	[3,3,1]	$c_o\times h\times w$
	BN,LReLU	-	$c_o\times h\times w$
	Conv	[3,3,1]	$c_o\times h\times w$
	BN,LReLU	-	$c_o\times h\times w$
	DeConv	[2,2,2]	$c_o\times 2h\times 2w$
	BN	-	$c_o\times 2h\times 2w$
最后一层	Conv	[1,1,1]	$c_o\times h\times w$
	TanH	-	$c_o\times h\times w$

个基线进行比较。对于两个基于迁移的基线方法(即 BIM 和 MI-FGSM), 通过在白盒设置中攻击 DNA-Det 生成对抗扰动并迁移到其他模型溯源技术上。为了更直观地展示 TraceEvader 生成的对抗性示例的特点, 本文在图 2 中展示了 TraceEvader 生成的对抗性示例的频谱可视化, 其中对抗性样本显示出与原始样本不同的频域模式。

4.2.1 逃避模型-架构溯源

本文针对 DNA-Det[10]、Reverse[24]和 DCT[23] 进行了实验, 这些技术均专注于 HFC 中的痕迹。实验结果见表 1, 显示 TraceEvader 的平均 ASR 为 83.6%。这表明所有的模型溯源技术都未能有效满足 DeepFake 取证目的。TraceEvader 在规避 DNA-Det 时表现出色, 平均 ASR 达到了 97.95%。Reverse 的平均 ASR 为 90.6%, 明显高于所有基线方法。本文的方法在某些情况下对 DCT 的攻击成功率相对较低。例如, 对于 SNGAN, 发现在默认设置下, 添加的对抗性扰动不足以掩盖 SNGAN 生成图像中出现的强烈峰值。为了解决这个问题, 本文调整了扰动强度 λ 值至 0.02, 攻击成功率显著提高到了 80.5%。在此设置下, PSNR 值为 35.5, 生成的图像仍能保持很高的视觉质量, 没有明显的伪影。

4.2.2 逃避模型-实例溯源

为了进一步验证 TraceEvader 的有效性, 本文在 AttNet 上进行对抗性实验[11]。AttNet 依赖于 LFC 中的痕迹进行深度伪造溯源。由于 AttNet 利用的特性与其他三种技术有显著差异, 这导致了 BIM 和 MI-FGSM 方法在迁移性方面表现不佳。对于无盒 Vera22-peak 方法, 它们只能操作图像的高频分量, 限制了其攻击 AttNet 的能力。TraceEvader 在所有情况下都优于 BIM、MI-FGSM、Vera22-peak 和 TR-Net, 并获得比 FakePolisher 更好的平均攻击成功率和图像质量。尽管在重新训练 AttNet 后, 本文的攻击效果与预训练模型相比不如前者显著, 但仍然实现了对 AttNet 的最高平均攻击成功率, 达到了 50.9%。

4.2.3 图像质量评估

为了证明 TraceEvader 的不可感知性, 本文对扰动样本进行了定性和定量分析。如图 6 所示, TraceEvader 生成了具有不可见扰动的高质量样本。相比之下, 由 MI-FGSM 生成的扰动样本显示出更明显的失真, 而 FakePolisher 重建的图像则呈现出可识别的模糊。本文详细列出了 PSNR 质量评估结果, 见表 5 的最后一列。在添加对抗性扰动后, 使用 PSNR 来衡量图像质量。符号“-”表示白盒攻击。前 2 名 ASR 分别用加粗和“*”标记。对于前 4 行(例如 ProGAN, MMDGAN, SNGAN 和 CramerGAN), 使用四种模型溯源技术提供的预训练模型进行攻击。对于后 4 行

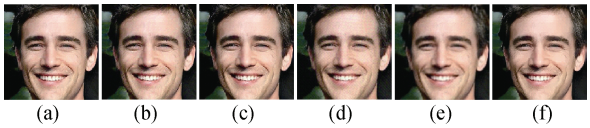


图 6 精心制作的对抗性样本的可视化。可视化包括由 BIM (a)、Vera22-peak (b)、TR-Net (c)、MIFGSM (d)、FakePolisher (e)和 TraceEvader (f)生成的扰动图像

Figure 6 Visualization of the crafted adversarial examples.The visualizations include the perturbated images generated with BIM (a), Vera22-peak (b), TR-Net (c), MIFGSM (d), FakePolisher (e) and our proposed TraceEvader (f)

表 5 TraceEvader 对四种模型溯源技术的逃避性能以 ASR(%)衡量

Table 5 The performance of TraceEvader in evading the four model attribution techniques measured by ASR (%)

GMs	方法	DNA-Det	Reverse	DCT	AttNet	PSNR	GMs	DNA-Det	Reverse	DCT	AttNet	PSNR
ProGAN	BIM	-	**96.0**	46.3	3.0	41.1	MMDGAN	-	0	78.1	0	40.5
	MI-FGSM	-	85.0	54.0	10.2	30.5		-	1.7	**98.0**	11.6	30.5
	Vera22-peak	**97.4**	51.5	81.8	0	50.0		0	1.7	0	0	50.9
	FakePolisher	0	70.2	84.0	30.8	31.8		**100.0**	**98.3**	24.7	39.5*	32.4
	TR-Net	88.7	48.7	96.1*	**60.1**	35.8		20.0	3.8	52.4	6.8	36.3
	TraceEvader	96.9*	92.3*	**99.6**	59.0*	36.6		98.7*	78.8*	96.4*	82.7	36.3
SNGAN	BIM	-	61.9	64.6	0	41.5	CramerGAN	-	17.2	0	0	40.2
	MI-FGSM	-	83.8	**100.0**	3.1	30.5		-	99.8*	69.0*	10.9	30.6
	Vera22-peak	48.2	9.7	75.2*	1.0	41		0	10.1	2.2	0	50.5
	FakePolisher	**100.0**	94.6*	72.9	48.4*	32.5		**100.0**	98.2	33.2	45.1	31.9
	TR-Net	66.2	4.4	49.8	8.7	36.7		0	93.8	30.4	45.7*	35.8
	TraceEvader	**100.0**	**98.5**	65.8	**71.0**	36.3		97.5*	**100.0**	**93.4**	**48.2**	37.5
CycleGAN	BIM	-	23.7	3.7	2.2	38.0	StyleGAN2	-	42.5	**100.0**	0	37.32
	MI-FGSM	-	34.1	80.0*	2.3	30.6		-	99.8*	**100.0**	0	30.4
	Vera22-peak	94.8	79.5*	8.3	2.1	50.6		94.4	47.8	12.0	1.9	51.1
	FakePolisher	**100.0**	**88.1**	**100.0**	**45.2**	27.7		**100.0**	14.0	**100.0**	46.0	27.6
	TR-Net	70.2	54.5	30.1	12.3	35.5		98.7	81.9	36.2	4.2	36.7
	TraceEvader	96.2*	55.9	46.9	17.4*	36.9		99.1*	**100.0**	60.1	17.8*	38.2
PNDM	BIM	-	98.0	73.3	15.3	38.2	LDM	-	**100.0**	3.6	9.4	38.2
	MI-FGSM	-	94.6	**100.0**	61.5*	30.4		-	99.7	**100.0**	10.1	30.5
	Vera22-peak	72.0	99.0*	4.5	0	51.2		**100.0**	60.0	20.8	0.3	51.1
	FakePolisher	**100.0**	52.9	0	28.1	28.5		**100.0**	0	**100.0**	**89.9**	28.7
	TR-Net	94.4	83.5	75.2	23.2	36.2		99.2	59.2	70.1	11.7	35.9
	TraceEvader	95.2*	**99.3**	90.6*	**88.3**	36.3		**100.0**	**100.0**	61.8	22.7*	38.4

(例如 CycleGAN, StyleGNA2, PNDM 和 LDM), 由于原论文溯源方法没有在上述 4 类模型生成的伪造样本上训练, 本文对训练数据集进行扩充, 采用提供的代码进行训练, 在重新训练的模型进行逃避攻击, 溯源性能可见表 1。

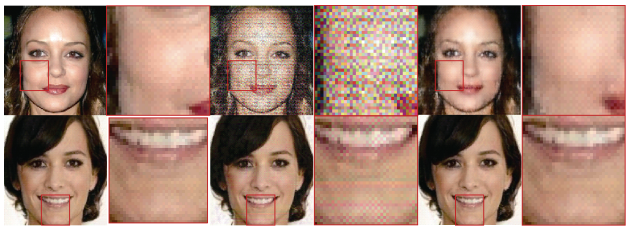
综上所述, TraceEvader 在两种溯源技术下的平均攻击成功率最高, 可达到 79.07%, 而在四种最先进模型溯源技术中, 平均攻击成功率最高, 可达到 97.95%。此外, 本文方法在超过 84%的情况下具有前 2 名的攻击成功率。相比之下, MI-FGSM 和 FakePolisher 引入了可见噪声, 而本文的方法则保持了相对较好的图像质量。

4.3 鲁棒性实验

在处理真实场景时, 本文考虑了一些现实场景: 通过 DeepFakes 注入的添加扰动会受到常见图像变换(如压缩和噪声)的影响而被破坏。同时, 本文严格假设防御者已经知晓攻击者可能在输入中注入对抗性扰动, 防御者会采用图像去噪方法来消除这些被注入的扰动影响。因此, 在实验中, 本文还研究了当添加扰动被破坏或故意去噪时的鲁棒性评估。

4.3.1 针对图像去噪的鲁棒性实验

为了验证本文提出的方法在各种攻击下的强大性能, 假设防御者收集了一定数量的干净图像及其相应由 TraceEvader 生成的注入扰动的图像对。本文采用了常见的去噪自动编码器[43-45]作为主要的去噪工具。该模型是在成对的图像监督下进行训练的, 并在新的对抗性样本上进行测试。如图 7 所示, 为了更



(a) 原始 (b) Adv (c) 重建

图 7 可视化(a)原始假图像(b)添加了精心制作的对抗性扰动的假图像和(c)使用去噪技术重建的图像

Figure 7 Visualization of the (a) original fake images (b) fake images added with crafted adversarial perturbations and (c) reconstructed images with denoising technique

好地展示所使用的去噪技术的有效性, 顶部行显示的是扰动程度显著的样本, 底部为扰动强度不显著的样本。实验结果显示, 去噪模型成功消除了明显的指纹伪影。然而, 表 6 的结果显示, 重建的对抗样本仍然保持较高的 ASR, 表明本文的攻击对恶意的图像去噪有较强的抵抗能力。

表 6　原始对抗样本与仔细去噪样本的 ASR(%)比较

Table 6　Comparison of ASR(%) between original adversarial samples and carefully denoised ones

	DNA-Det	Reverse	DCT	AttNet
Adv	73.4	65.4	100	50.1
Recovered	75	65.9	75	84

4.3.2　针对常见图像变换的鲁棒性实验

本文考虑了两种常见的图像变换(例如高斯噪声和 JPEG 压缩), 以探索这些变换是否会给 TraceEvader 生成的对抗性样本造成退化。图 8(a)显示, 添加噪声并未削弱攻击性能, 因为可溯源的指纹痕迹已经被破坏。实际上, 噪声会给图像引入更大的扰动, 从而进一步提高攻击成功率。从图 8(b)可以看出, 仅在大幅压缩后, 攻击性能才略有下降。实验结果表明, 本文的方法能够较好地抵御常见的图像退化。

4.4　消融实验

4.4.1　唯一对抗性扰动的影响

本节探讨了每种攻击模式(例如 HFC 的 UIT, LFC 的对抗性模糊)对攻击有效性的影响。如表 7 所示, 在攻击 DNA-Det 时, 仅使用 UIT 通常就可以达到与本文提出方法相近的高攻击成功率。而攻击 AttNet 时, 仅使用对抗性模糊能够产生类似于建议攻击的性能。可以得出结论, 单一模式的 UIT 攻击对于使用高频分量中痕迹特征的模型溯源技术时有效, 但在逃避使用低频分量中痕迹特征的模型溯源技术时其效果有限, 而单一对抗性模糊则呈现相反结果。因此, 它们各自的效果有所局限。通过混合对抗攻击, 能够成功提升攻击性能, 这表明本文方法融合了各个设计的优势, 并将其兼容地集成于一体。

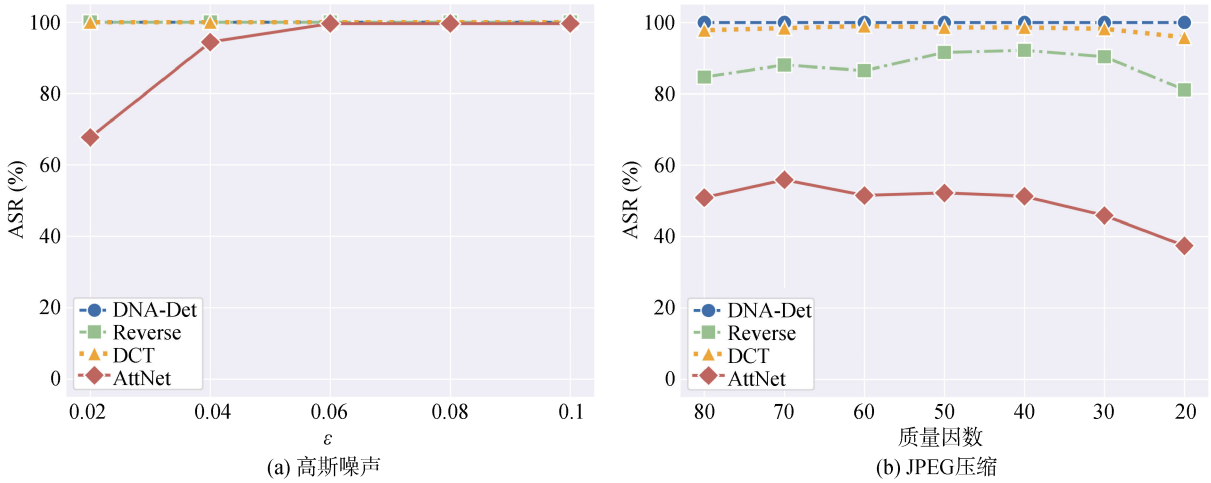


图 8　在不同强度下图像变换的存活性能

Figure 8　Performance in surviving the image transformation under different intensities

4.4.2　权重因子 λ 对 HFC 对抗性扰动的影响

本文讨论了通过权重因子 λ 控制的对抗性 UIT 强度对最终模型的溯源规避攻击效果的影响。具体来说, 将 λ 从 0.001 调整到 0.01, 并研究了伪造模型 ProGAN 的攻击成功率。如图 9(a)所示, AttNet 的攻击成功率保持稳定, 而其他三种方法的攻击成功率随着扰动强度的增加而增加。这种现象可以归因于 AttNet 利用的权重痕迹主要分布在图像的低频分量中, 因此其性能受高频分量扰动的影响较小。相比之下, 依赖高频分量的其他方法的攻击成功率会随着 λ 值的增加而提高。

4.4.3　高斯核宽度 σ 对 LFC 对抗性扰动的影响

本节研究了模糊程度对攻击成功率的影响。通过将高斯核宽度 σ 从 1 调整到 9, 如图 9(b)所示, 模糊操作的程度对实例级溯源方法 AttNet 有显著影响。然而, 对于架构级溯源方法, 攻击成功率略有提高。

4.5　对比实验

为了验证本文提出的模型效果, 本节进行了对比实验, 探讨模糊模块是否可以替代。对比实验采用了两种基于 ProGAN 的攻击方式: UIT&模糊和 UIT&噪声, 并在 DNA-Det、Reverse、DCT、AttNet 和 PSNR 这五种模型溯源技术上进行了测试。作为常见的图

表 7　基于四种 GAN 的四种模型溯源技术的单 UIT 和对抗模糊性能

Table 7　Performance of sole UIT and adversarial blur on four model attribution techniques with four GANs

GMs	攻击模式	DNA-Det	Reverse	DCT	AttNet
ProGAN[36]	UIT	96.9	86	100	4.2
	blur	40.7	12.5	90.0	59.0
	UIT & blur	96.9	92.3	99.6	59.0
MMDGAN[37]	UIT	98.5	81.5	88.2	8.0
	blur	0	0	25.8	88.2
	UIT & blur	98.7	78.8	96.4	82.7
SNGAN[38]	UIT	100	97.3	56.2	3.9
	blur	36.2	2	49.6	72
	UIT & blur	100.0	98.5	65.8	71.0
CramerGAN[39]	UIT	95.9	99.8	100	4.9
	blur	0	0	0	51.0
	UIT & blur	97.5	100.0	93.4	48.2

像变换操作，本文在原有基线基础上添加了方差为 0.02 的噪声。如表 8 所示，在 AttNet 攻击模式下，添加模糊的攻击成功率显著高于添加噪声的攻击模式。同时，PSNR 数值表明，采用模糊攻击模式时，图像的质量更高。这可能是因为相较于噪声攻击，模糊攻击在图像视觉上的影响更为自然且更难以被检测到。模糊处理有效地模糊了图像的边缘和细节，使得 DeepFake 检测模型难以从图像的局部特征中分辨出篡改痕迹，从而降低了检测性能。相反，噪声攻击通常会在图像中引入明显的颗粒和失真，这些异常特征更容易被深度学习模型识别并标记为潜在的伪造痕迹。

4.6　实例级溯源模型逃避攻击

为了深入探讨模型-实例溯源的脆弱性，我们训练了 10 个具有不同初始化种子的 ProGAN 实例，标记为 seed#i，并使用 AttNet 进行溯源。图 10 和图 11 展示了在应用 TraceEvader 攻击前后模型-实例溯源的不同之处。显然， TraceEvader 能够有效逃避模型-实例溯源的检测。平均检测成功率下降了 61.5%，最糟糕的情况下下降了 90.8%。图 12 显示了对抗性扰动对图像频域的影响。实验结果进一步证明本文的攻击能够利用模型-实例溯源方法中存在的脆弱性。

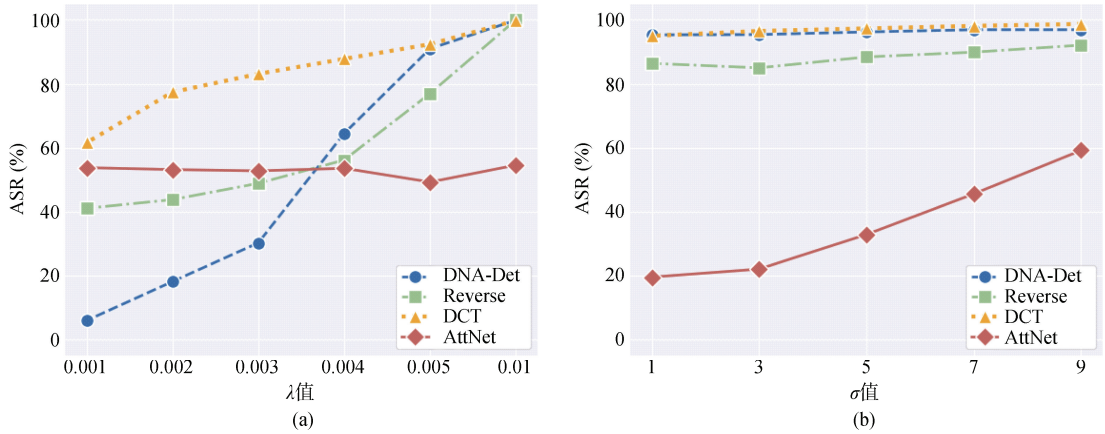


图 9　不同权重因子 λ 和 σ 对四种溯源方法的 ASR(%)

Figure 9　The ASR (%) of different weight factors λ and σ against four attribution techniques

表 8　基于 ProGAN 的五种模型溯源技术的单 UIT 和对抗模糊性能

Table 8　Performance of sole UIT and adversarial blur on four model attribution techniques with four GANs

GMs	攻击模式	DNA-Det	Reverse	DCT	AttNet	PSNR
ProGAN	UIT&模糊	96.9	92.3	99.6	59.0	36.6
	UIT&噪声	96.9	91.7	100	12.3	34.1

4.7　DeepFake 检测逃避攻击

本节探讨 TraceEvader 是否具备逃避常见 DeepFake 检测器的潜力。表 9 显示了 TraceEvader 在逃避基于空域、基于频域和基于指纹的 DeepFake 检测方法(如 DCT、CNNDetection[9]和 AttNet)方面的性能。实验结果表明，在四种不同的伪造方法中，TraceEvader 对三种 DeepFake 检测方法的平均攻击成功率超过 43.3%，达到了与其他专门为 DeepFake 检测设计的无盒方法相当的性能，甚至在某些情况下表现更加优越。

5　结论

本文针对当前常见的模型溯源技术进行了研究，并引入了一种无盒无训练的逃避攻击方法—

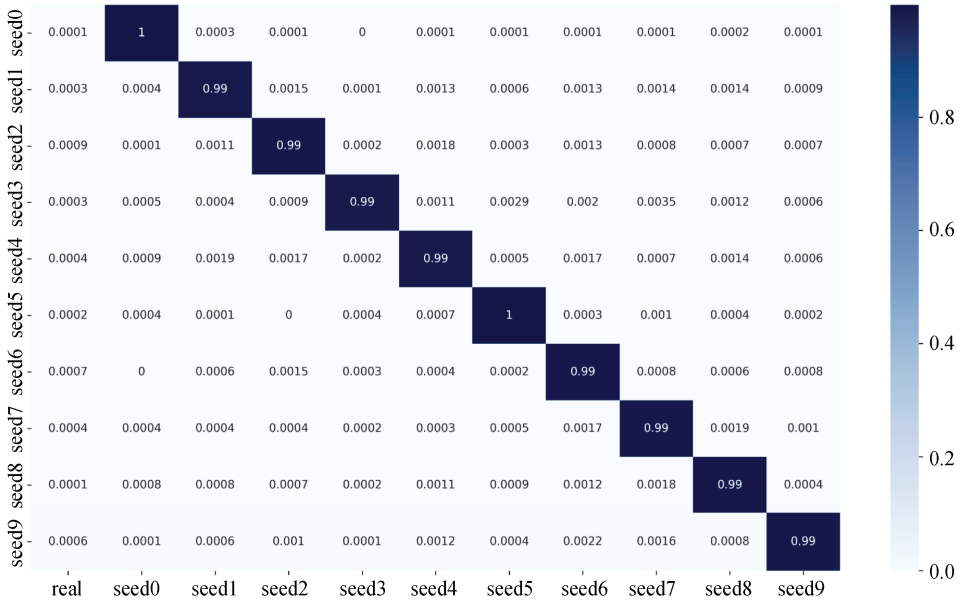


图 10　溯源交叉种子 ProGAN 实例的混淆矩阵

Figure 10　Confusion matrice on attributing cross-seed ProGAN instances

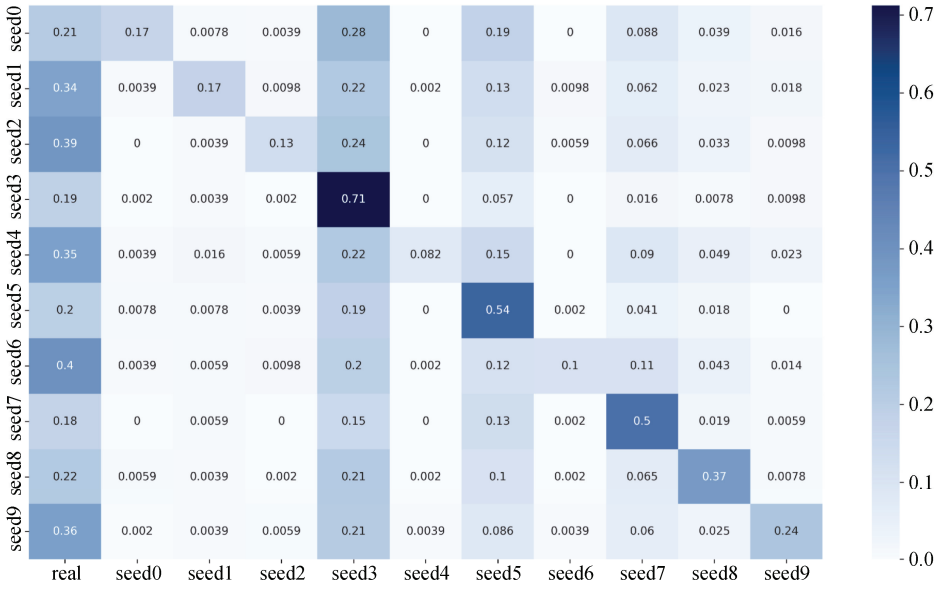


图 11　溯源添加了对抗性扰动样本的混淆矩阵

Figure 11　Confusion matrice on attributing the samples with our added adversarial perturbations

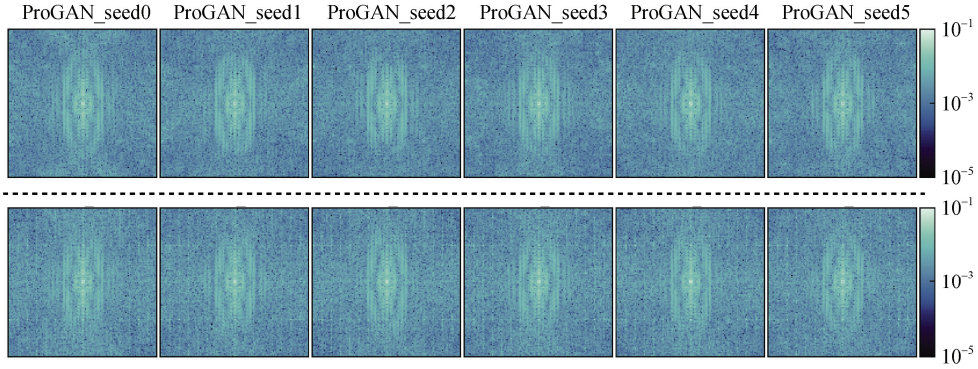


图 12　不同 ProGAN 实例的频谱分析

Figure 12　Spectrum analysis for different ProGAN instances

表 9　四种对抗攻击在躲避 DeepFake 检测中的表现

Table 9　The performance of four adversarial attack in evading DeepFake detection

GM&检测器	ProGAN			MMDGAN		
方法	DCT	CNNDetection	AttNet	DCT	CNNDetection	AttNet
Vera22-peak	82.2	44.9	0	0.6	24.5	0.2
FakePolisher	92.4	47.5	25.7	22.0	30.1	38.2
TraceEvader	97.0	49.0	48.8	41.2	33.0	20.9
GM&Detector	SNGAN			CramerGAN		
方法	DCT	CNNDetection	AttNet	DCT	CNNDetection	AttNet
Vera22-peak	72.8	62.4	0.2	0.2	20.7	0
FakePolisher	55.4	60.5	46.2	26.4	30.3	43.4
TraceEvader	56.1	61.9	30.4	82.3	55.4	4.5

—TraceEvader。这是首个揭示模型溯源技术脆弱性的研究, 该技术容易受到混淆对抗性扰动影响。此外, 本文的研究还发现, 在 HFC 和 LFC 中都存在生成模型留下的痕迹, 其中 HFC 中的痕迹与模型架构相关, 而 LFC 中的痕迹则与具体实例相关。综上所述, 本文希望促进更加强大和全面的 DeepFake 取证技术的发展, 以提高对深度伪造内容的检测能力和安全性。

参考文献

[1] Karras T, Laine S, Aittala M, et al. Analyzing and Improving the Image Quality of StyleGAN[C]. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020: 8107-8116.

[2] Dhariwal P, Nichol A. Diffusion Models Beat GANs on Image Synthesis[C]. *The 35th International Conference on Neural Information Processing Systems*, 2021: 8780-8794.

[3] Liu L P, Ren Y, Lin Z J, et al. Pseudo Numerical Methods for Diffusion Models on Manifolds[EB/OL]. 2022: arXiv: 2202.09778. https://arxiv.org/abs/2202.09778.

[4] Dolhansky B, Bitton J, Pflaum B, et al. The DeepFake Detection Challenge (DFDC) Dataset[EB/OL]. 2020: arXiv: 2006.07397. https://arxiv.org/abs/2006.07397.

[5] Juefei-Xu F, Wang R, Huang Y H, et al. Countering Malicious DeepFakes: Survey, Battleground, and Horizon[J]. *International Journal of Computer Vision*, 2022, 130(7): 1678-1734.

[6] Leong D. Deepfakes and Disinformation Pose a Growing Threat in Asia[J]. *The Diplomat*, 2023: 11.

[7] Wang R, Juefei-Xu F, Luo M, et al. FakeTagger: Robust Safeguards Against DeepFake Dissemination *via* Provenance Tracking[C]. *The 29th ACM International Conference on Multimedia*, 2021: 3546-3555.

[8] Wang R, Huang Z H, Chen Z K, et al. Anti-Forgery: Towards a Stealthy and Robust DeepFake Disruption Attack *via* Adversarial Perceptual-Aware Perturbations[EB/OL]. 2022: arXiv: 2206.00477. https://arxiv.org/abs/2206.00477.

[9] Zhao Y R, Ge W F, Li W X, et al. Capturing the Persistence of Facial Expression Features for Deepfake Video Detection[M]. Information and Communications Security. Cham: Springer International Publishing, 2020: 630-645.

[10] Yang T Y, Huang Z Y, Cao J, et al. Deepfake Network Architecture Attribution[J]. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2022, 36(4): 4662-4670.

[11] Yu N, Davis L, Fritz M. Attributing Fake Images to GANs: Learning and Analyzing GAN Fingerprints[C]. *2019 IEEE/CVF International Conference on Computer Vision*, 2019: 7555-7565.

[12] Zhu J Y, Park T, Isola P, et al. Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks[C]. *2017 IEEE International Conference on Computer Vision*, 2017: 2242-2251.

[13] Wesselkamp V, Rieck K, Arp D, et al. Misleading Deep-Fake Detection with GAN Fingerprints[C]. *2022 IEEE Security and Privacy Workshops*, 2022: 59-65.

[14] Jia S, Ma C, Yao T P, et al. Exploring Frequency Adversarial Attacks for Face Forgery Detection[C]. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022: 4093-4102.

[15] Chakraborty A, Alam M, Dey V, et al. Adversarial Attacks and Defences: A Survey[EB/OL]. 2018: arXiv: 1810.00069. https://arxiv.org/abs/1810.00069.

[16] Carlini N, Farid H. Evading Deepfake-Image Detectors with White- and Black-Box Attacks[C]. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2020: 2804-2813.

[17] Hussain S, Neekhara P, Jere M, et al. Adversarial Deepfakes: Evaluating Vulnerability of Deepfake Detectors to Adversarial Examples[C]. *2021 IEEE Winter Conference on Applications of Computer Vision*, 2021: 3347-3356.

[18] Moosavi-Dezfooli S M, Fawzi A, Fawzi O, et al. Universal Adversarial Perturbations[C]. *2017 IEEE Conference on Computer Vision and Pattern Recognition*, 2017: 86-94.

[19] Mopuri K R, Ojha U, Garg U, et al. NAG: Network for Adversary Generation[C]. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018: 742-751.

[20] Hou Y, Guo Q, Huang Y H, et al. Evading DeepFake Detectors *via* Adversarial Statistical Consistency[C]. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023: 12271-12280.

[21] Dzanic T, Shah K, Witherden F. Fourier spectrum discrepancies in deep network generated images[J]. *Advances in neural information*

processing systems, 2020, 33: 3022-3032.

[22] Durall R, Keuper M, Keuper J. Watch Your Up-Convolution: CNN Based Generative Deep Neural Networks Are Failing to Reproduce Spectral Distributions[C]. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020: 7887-7896.

[23] Frank J, Eisenhofer T, Schönherr L, et al. Leveraging Frequency Analysis for Deep Fake Image Recognition[C]. *The 37th International Conference on Machine Learning*, 2020: 3247-3258.

[24] Asnani V, Yin X, Hassner T, et al. Reverse Engineering of Generative Models: Inferring Model Hyperparameters from Generated Images[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023, 45(12): 15477-15493.

[25] Wang H H, Wu X D, Huang Z Y, et al. High-Frequency Component Helps Explain the Generalization of Convolutional Neural Networks[C]. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020: 8681-8691.

[26] Fan L X, Ng K W, Chan C S. Rethinking Deep Neural Network Ownership Verification: Embedding Passports to Defeat Ambiguity Attacks[C]. *The 33rd International Conference on Neural Information Processing Systems*, 2019: 4714-4723.

[27] Marra F, Gragnaniello D, Verdoliva L, et al. Do GANs Leave Artificial Fingerprints? [C]. *2019 IEEE Conference on Multimedia Information Processing and Retrieval*, 2019: 506-511.

[28] Lempitsky V, Vedaldi A, Ulyanov D. Deep Image Prior[C]. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018: 9446-9454.

[29] Wang S Y, Wang O, Zhang R, et al. CNN-Generated Images Are Surprisingly Easy to Spot… for now[C]. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020: 8692-8701.

[30] Sinitsa S, Fried O. Deep Image Fingerprint: Towards Low Budget Synthetic Image Detection and Model Lineage Analysis[C]. *2024 IEEE/CVF Winter Conference on Applications of Computer Vision*, 2024: 4055-4064.

[31] Zhang K, Zuo W M, Chen Y J, et al. Beyond a Gaussian Denoiser: Residual Learning of Deep CNN for Image Denoising[J]. *IEEE Transactions on Image Processing*, 2017, 26(7): 3142-3155.

[32] Kurakin A, Goodfellow I, Bengio S. Adversarial Machine Learning at Scale[EB/OL]. 2016: arXiv: 1611.01236. https://arxiv.org/abs/1611.01236.

[33] Dong Y P, Liao F Z, Pang T Y, et al. Boosting Adversarial Attacks with Momentum[C]. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018: 9185-9193.

[34] Liu C, Chen H J, Zhu T Q, et al. Making DeepFakes More Spurious: Evading Deep Face Forgery Detection *via* Trace Removal Attack[J]. *IEEE Transactions on Dependable and Secure Computing*, 2023, 20(6): 5182-5196.

[35] Huang Y H, Juefei-Xu F, Wang R, et al. FakePolisher: Making DeepFakes More Detection-Evasive by Shallow Reconstruction[C]. *The 28th ACM International Conference on Multimedia*, 2020: 1217-1226.

[36] Karras T, Aila T, Laine S, et al. Progressive Growing of GANs for Improved Quality, Stability, and Variation[EB/OL]. 2017: arXiv: 1710.10196. https://arxiv.org/abs/1710.10196.

[37] Wang W, Sun Y, Halgamuge S. Improving MMD-GAN Training with Repulsive Loss Function[EB/OL]. 2018: arXiv: 1812.09916. https://arxiv.org/abs/1812.09916.

[38] Miyato T, Kataoka T, Koyama M, et al. Spectral Normalization for Generative Adversarial Networks[EB/OL]. 2018: arXiv: 1802.05957. https://arxiv.org/abs/1802.05957.

[39] Bellemare M G, Danihelka I, Dabney W, et al. The Cramer Distance as a Solution to Biased Wasserstein Gradients[EB/OL]. 2017: arXiv: 1705.10743. https://arxiv.org/abs/1705.10743.

[40] Rombach R, Blattmann A, Lorenz D, et al. High-Resolution Image Synthesis with Latent Diffusion Models[C]. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022: 10674-10685.

[41] Chen P Y, Zhang H, Sharma Y, et al. ZOO: Zeroth Order Optimization Based Black-Box Attacks to Deep Neural Networks without Training Substitute Models[C]. *The 10th ACM Workshop on Artificial Intelligence and Security*, 2017: 15-26.

[42] Vincent P, Larochelle H, Bengio Y, et al. Extracting and Composing Robust Features with Denoising Autoencoders[C]. *The 25th International Conference on Machine Learning*, 2008: 1096-1103.

[43] Chiang H T, Hsieh Y Y, Fu S W, et al. Noise Reduction in ECG Signals Using Fully Convolutional Denoising Autoencoders[J]. *IEEE Access*, 2019, 7: 60806-60813.

[44] Zhang C, Zhou L, Zhao Y Y, et al. Noise Reduction in the Spectral Domain of Hyperspectral Images Using Denoising Autoencoder Methods[J]. *Chemometrics and Intelligent Laboratory Systems*, 2020, 203: 104063.

[45] Lee W H, Ozger M, Challita U, et al. Noise Learning-Based Denoising Autoencoder[J]. *IEEE Communications Letters*, 2021, 25(9): 2983-2987.

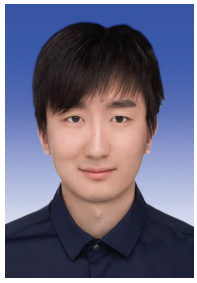
吴梦洁 于 2022 年在武汉大学信息安全专业获得工学学士学位。现在武汉大学网络空间安全专业攻读硕士学位。研究领域为人工智能安全。Email：mengjiewu@whu.edu.cn

于佳艺 于 2024 年在西安交通大学软件工程专业获得学士学位。现在武汉大学电子信息专业攻读硕士学位。研究领域为人工智能安全、深度伪造。Email：2206123961@stu.xjtu.edu.cn

汪润 于 2018 年在武汉大学信息安全专业获得博士学位。现任武汉大学国家网络安全学院副教授，博士生导师。研究领域人工智能安全。研究兴趣包括：人工智能安全、软件与系统安全。Email：wangrun@whu.edu.cn

叶茜 于 2023 年在南京航空航天大学获得硕士学位。现在武汉大学网络空间安全专业攻读博士学位。研究领域为视觉隐私保护、深度伪造。Email：rmde_f@nuaa.edu.cn

张钰洋 于 2023 年在武汉大学网络空间安全专业获得工学学士学位。现在武汉大学网络空间安全专业攻读硕士学位。研究领域为人工智能安全与隐私。Email：yuwanz@whu.edu.cn



蔺琛皓 于 2018 年在香港理工大学电子计算学专业获得博士学位。现任西安交通大学电子与信息学部网络空间安全学院教授，博士生导师。研究领域为人工智能安全：对抗性机器学习、AIGC 攻防、大模型安全、AI 测试、公平性、可解释性等；智能身份安全：人机智能系统身份安全与认证。Email：linchenhao@xjtu.edu.cn

方黎明 现任南京航空航天大学深圳研究院副院长、智能与安全实验室副主任、密码学与应用安全实验室主任、高安全系统的软件开发与验证技术工信部重点实验室学术带头人，博士生导师。研究领域为工智能安全。研究兴趣包括：密码学、隐私计算、人工智能安全、区块链。Email：fangliming@nuaa.edu.cn

王丽娜 于 2001 年在东北大学获得博士学位。现任武汉大学二级教授，博士生导师，空天信息安全与可信计算教育部重点实验室主任。研究领域为软件与系统安全、隐写分析和人工智能安全。Email：lnwang@whu.edu.cn