TraceEvader: Making DeepFakes More Untraceable via Evading the Forgery Model Attribution

Mengjie Wu[1], Jingui Ma[1], Run Wang[1]*, Sidan Zhang[1], Ziyou Liang[1], Boheng Li[1], Chenhao Lin[2], Liming Fang[3], Lina Wang[1,4]

[1] Key Laboratory of Aerospace Information Security and Trusted Computing, Ministry of Education, School of Cyber Science and Engineering, Wuhan University, China
[2]Xi'an Jiaotong University, China
[3] Nanjing University of Aeronautics and Astronautics, China
[4] Zhengzhou Xinda Institute of Advanced Technology
{mengjiewu,majingui102,wangrun,sidanzhang,ziyouliang,bohengli,lnwang}@whu.edu.cn,
{linchenhao}@xjtu.edu.cn, {fangliming}@nuaa.edu.cn

Abstract

In recent few years, DeepFakes are posing serve threats and concerns to both individuals and celebrities, as realistic DeepFakes facilitate the spread of disinformation. Model attribution techniques aim at attributing the adopted forgery models of DeepFakes for provenance purposes and providing explainable results to DeepFake forensics. However, the existing model attribution techniques rely on the trace left in the DeepFake creation, which can become futile if such traces were disrupted. Motivated by our observation that certain traces served for model attribution appeared in both the high-frequency and low-frequency domains and play a divergent role in model attribution. In this work, for the first time, we propose a novel **training-free** evasion attack, TraceEvader, in the most practical **non-box** setting. Specifically, TraceEvader injects a **universal** imitated traces learned from wild DeepFakes into the high-frequency component and introduces adversarial blur into the domain of the low-frequency component, where the added distortion confuses the extraction of certain traces for model attribution. The comprehensive evaluation on **4** state-of-the-art (SOTA) model attribution techniques and fake images generated by **8** generative models including generative adversarial networks (GANs) and diffusion models (DMs) demonstrates the effectiveness of our method. Overall, our TraceEvader achieves the highest average attack success rate of **79%** and is robust against image transformations and dedicated denoising techniques as well where the average attack success rate is still around **75%**. Our TraceEvader confirms the limitations of current model attribution techniques and calls the attention of DeepFake researchers and practitioners for more robust-purpose model attribution techniques.

Introduction

With the rapid development of generative models, such as GANs (Karras et al. 2020) and DMs (Dhariwal and Nichol 2021; Liu et al. 2022), DeepFakes are becoming real threats to humans, which could synthesize realistic high-quality audios, images, and videos (Dolhansky et al. 2020; Juefei-Xu et al. 2022). Such realistic DeepFakes could be leveraged for malicious purposes, like producing disinformation, creating fake pornography, and releasing fake official statements (Leong 2023; Moneywatch 2023), *etc.*. Therefore, preventing the abuse and spread of malicious DeepFake has become an urgent need (Wang et al. 2021, 2022).

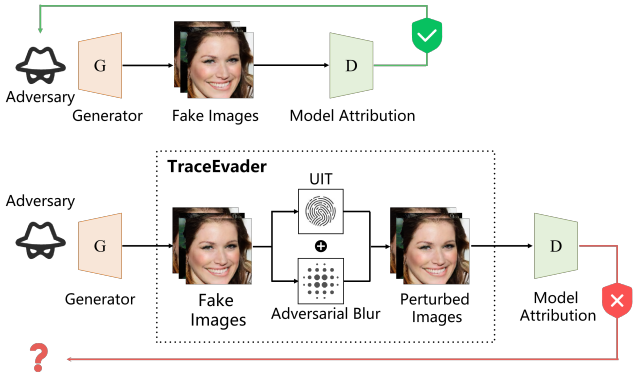


Figure 1: An overview of our framework. In the top panel, an adversary creates DeepFakes (*i.e.*, fake image) and releases them to the social network, then the defender discriminates the source of such DeepFakes and attributes them to a specific forgery model successfully for DeepFake forensics purposes. In the bottom panel, the adversary injects our crafted invisible adversarial perturbations before releasing them and evades the model attribution.

In the current research of DeepFake forensics, the DeepFake detection (Wang et al. 2020b; Zhao et al. 2020) tells us whether the sample is real or fake, while the DeepFake attribution[1] aims at investigating which forgery model is employed for creating such DeepFakes, further providing explainable results for DeepFake detection. In this paper, we explore an interesting question, *whether the existing DeepFake attribution techniques are robust enough to serve for DeepFake forensics and the potentials in deploying in real*

*Corresponding author. E-mail: wangrun@whu.edu.cn

[1]The DeepFake attribution has the same meaning as forgery model attribution. In this paper, we use them alternately.

scenarios.

Thus far, the DeepFake attribution could be classified as model-architecture attribution (Yang et al. 2022) and model-instance attribution (Yu, Davis, and Fritz 2019) based on whether the training settings are considered. Specifically, the model-architecture attribution infers the adopted model architecture merely (*e.g.*, CycleGAN (Zhu et al. 2017), StyleGAN (Karras et al. 2020)), while the model-instance attribution struggles to identify the model architecture equipped with specific training settings, like training dataset, initial seeds, *etc*.. Recent studies have shown that both the model-architecture and model-instance attribution rely on discriminating the traces left in the DeepFake creation (Yang et al. 2022; Yu, Davis, and Fritz 2019). In this work, we seek a practical trace disruption method to evade the two different types of model attribution techniques effectively in a black-box setting. Figure 1 presents the framework of our work, where our proposed evasion attack aims at injecting invisible adversarial perturbations to fool the attribution techniques.

The prior studies (Wesselkamp et al. 2022; Jia et al. 2022) are mostly working on exposing the vulnerability of DeepFake detectors by disturbing the clues for discriminating between real and fake. Actually, such clues should generalize diverse forgery models, especially unknown forgery models. However, the traces for model attribution are model-relevant, where the stable traces appeared in each model architecture and distinct traces vary among model instances. Thus, the evasion attack dedicated to DeepFake detectors is not appropriate for evading model attribution as their relied traces are vastly different.

The straightforward idea to evade the DeepFake attribution techniques in a black-box setting could be generating adversarial perturbations via accessing a surrogate model or querying the victim model by sending large amounts of samples (Chakraborty et al. 2018; Carlini and Farid 2020). However, both of them suffer poor transferability in tackling unknown models (Chakraborty et al. 2018; Hussain et al. 2021). Thus, we investigate the possibility of employing universal adversarial perturbations (Moosavi-Dezfooli et al. 2017; Mopuri et al. 2018) to disrupt the traces in black-box settings and further evade the target model attribution techniques. First of all, we need to explore a crucial question, *what makes DeepFake traceable?* Our preliminary experimental results illustrate that the forgery model leaves traces in both the high- and low-frequency components of an image. The high-frequency component (HFC) contains a globally consistent trace relevant to the model architecture, while the trace in the low-frequency component (LFC) is concentrated in specific regions and correlated with the model's weights. Both of them are widely served as an asset for model attribution. More detailed analysis on the exploration of traces in HFC and LFC refers to Section .

Motivated by the above observation that the traits of the trace in HFC and LFC are vastly different, in this paper, we introduce the adversarial distortions to the HFC and LFC, respectively. Specifically, for the HFC, inspired by the common ambiguity attack, we craft a universal imitated trace (UIT) learned from a set of fake images created by popular generative models to bring confusion to the DeepFake attribution techniques when extracting traces for model attribution. Inspired by prior works (Hou et al. 2023), blur is an effective natural degradation to mitigate spatial and frequency differences of DeepFake and real images. We employ Gaussian blurring mean shift to eliminate subtle traces varies in the LFC domain. Notably, our proposed evasion attack is **training-free** without involving any optimization of the tedious gradient computation, **non-box** manner working on black-box settings without obtaining any knowledge of the target DeepFake attribution techniques, **forgery model-agnostic** applicable to the popular generative models (*e.g.*, GANs, DMs), and **high-transferability** capabilities in tackling diverse DeepFake attribution techniques, in terms of model-architecture and model-instance attribution.

Our in-depth evaluation on **4** SOTA DeepFake attribution techniques and fake images generated by **6** GANs and **2** DMs demonstrate that our TraceEvader successfully fools the two types of DeepFake attribution techniques with an average ASR more than **79%** and survives the image transformations and denoising techniques well. This indicates that the current DeepFake attribution techniques highly rely on the stable traces between model architecture and distinct traces among model instances. Our study presents a new challenge for future DeepFake attribution techniques served for multimedia forensics, which needs to look for more advanced traces introduced in DeepFake creation.

Our main contributions are summarized as follows:

- To the best of our knowledge, this is the very first attempt to reveal the vulnerability of the existing DeepFake attribution techniques which calls for more robust model attribution techniques for better forgery model provenance and DeepFake forensics.
- We conduct an in-depth analysis of the model traces introduced in DeepFake creation which are employed for model attribution and discover the role of the stable traces in HFC for discriminating model architecture and distinct traces in LFC for certifying model instances.
- We propose a novel evasion attack, TraceEvader, to prevent the exact traces extraction for model attribution by injecting crafted universal imitated traces into the HFC and introducing Gaussian blur mean shift to eliminate the traces in the LFC, respectively.
- Experiments conducted on 4 state-of-the-art DeepFake attribution techniques demonstrate the effectiveness in evading the existing model attribution techniques and applicability to the GAN-based and DM-based forgery models and robustness in surviving the common sample transformations and intentional powerful denoising techniques.

Understanding Model Attribution

In this section, we explore three crucial questions, *what makes the model traceable, where are the traces, and what kind of features do they exhibit?* This in-depth understanding of model attribution will help us generate dedicated adversarial perturbations to disrupt the traces for DeepFake attribution. Here, we take GAN as an example to answer the three questions.

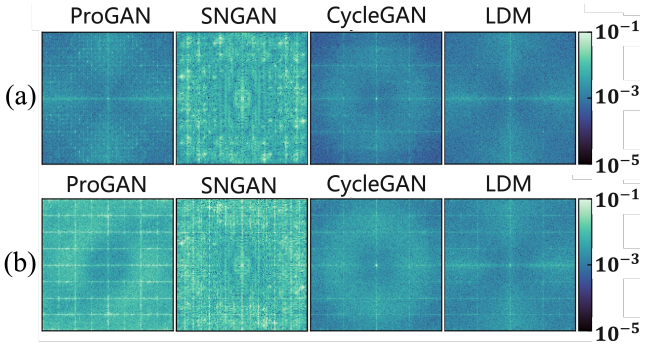


Figure 2: Spectrum analysis for popular generative models. We visualize the mean DFT spectra of high-pass filtered (a) generated raw fake images and (b) images added with our generated adversarial perturbations.

What Makes DeepFakes Traceable?

Studies have shown that GANs have stable fingerprints for discriminating whether two GANs (*e.g.*, ProGAN, StyleGAN) have the same architecture and distinct fingerprints that could be leveraged for distinguishing diverse GAN instances trained with specific training settings.

Stable fingerprints within model-architectures. The model-architecture attribution refers to attributing fake images to the specific generator architecture. It drives us to dig deeper into traces left by the structure of the model itself during the generative process. In generative models, the generator is equipped with a set of convolution filtering, upsampling, and non-linear components for learning a specific data distribution and accomplishing the task of image synthesis together. Previous studies reveal that specific upsampling strategies, a widely adopted component in generators, lead to periodic or chessboard artifacts in high-frequency components (Dzanic, Shah, and Witherden 2019; Durall, Keuper, and Keuper 2020). This regular artifact in GANs is leveraged for model attribution where the discrepancy in the spectrum is demonstrated effective for discriminating different GAN architectures (*e.g.* DNA-Det (Yang et al. 2022); DCT (Frank et al. 2020); Reverse (Asnani et al. 2021)). Figure 2 illustrates that the spectral artifacts among generators within images.

Distinct fingerprints among model-instances. Recent work, AttNet(Yu, Davis, and Fritz 2019), claims that model traces are characterized by model weights for parameters of each model instance converging to distinct values after training. Due to the unstable training phenomenon of GAN, two GAN instances with the same architecture trained under respective settings could result in individual GANs where their synthesized outputs are not exactly the same as the output image quality diverse. This suggests the existence of subtle spatial traces unique to the model instance even though they share the same model architecture.

Model Fingerprints in HFC and LFC

As discussed above, the same model architecture shares stable fingerprints, while the different model instances have distinct model fingerprints. Here, we explore where are these fingerprints, their role in model attribution, and further, we exploit the trait of these model fingerprints to conduct effective adversarial disruption. To answer the above question , we conduct several preliminary experiments by introducing degradation in the HFC and visualizing the regions of existing model attribution methods focused on. In our preliminary experiments, we consider four popular model attribution techniques, including model-architecture attribution 1) DNA-Det (Yang et al. 2022); 2) DCT (Frank et al. 2020); 3) Reverse (Asnani et al. 2021) and model-instance attribution 4) AttNet (Yu, Davis, and Fritz 2019). Further details on related works refer to the technical appendix.

Degradation introduced in HFC. Existing studies (Wang et al. 2020a) claim that HFC plays a crucial role in classification of DNNs. Inspired by this insight, we investigate the role of HFC and LFC in the classification results of model attribution methods. Here, we decompose the information of fake images into HFC and LFC via Fast Fourier Transform (FFT) proposed in a prior study (Wang et al. 2020a). Specifically, we simply drop the HFC and reconstruct an image $\mathbf{x}_l$ via inverse FFT with only the LFC rather than a whole component normally adopted. We have the following equations:

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

where $\mathcal{F}$ represents the FFT, $\mathbf{x}$ is the original input, $\mathcal{F}^{-1}$ indicates the inverse FFT for input reconstruction, $f(\cdot; r)$ denotes a threshold function that separates the HFC and LFC from $\mathcal{F}(\mathbf{x})$ with a specified hyperparameter, radius r, where a larger r means a wider frequency band is preserved.

As illustrated in Figure 3(a), when the r is large, the average accuracy of $\mathbf{x}_l$ on four popular attribution models is close to 100%. If we continue to decrease r, three attribution techniques experience a significant decline in performance, except AttNet. This indicates that traces in both HFC and LFC can be useful for model attribution. Next, we explore the role of traces in HFC and LFC and their characteristics.

Traces visualization. Figure 3(b) visualizes the regions of two model attribution techniques (*i.e.*, DNA-Det, AttNet) focused on. The AttNet mostly relies on the traces of LFC and tends to focus on local regions rich in semantic information like eyes and mouth. In contrast, model attribution methods mostly rely on HFC, like DNA-Det paying attention to more evenly distributed over the entire image. Thus, we have high confidence to believe that the traces in LFC are semantic-related and the HFC contains more low-level patterns. A recent study (Yang et al. 2022) also claims that GAN architecture is likely to leave fingerprints that are globally consistent across the entire image, while weight traces vary in different regions.

In summary, our preliminary experimental results prove that 1) both the HFC and LFC of the images provide useful traces for model attribution; 2) traces in HFC focus on low-level patterns better serve the model-architecture attribution, while the traces in LFC focus on the semantic-level pattern well serve the model-instance attribution; 3) the model-

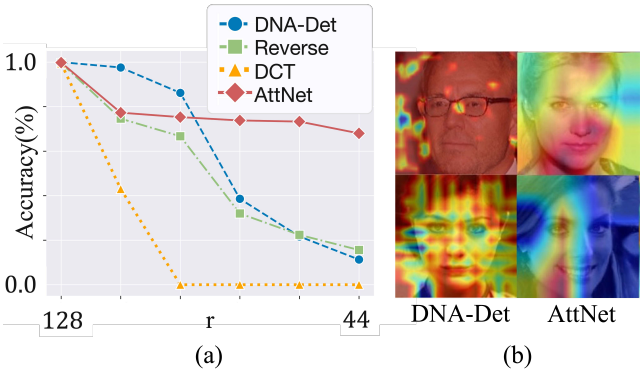


Figure 3: (a) The performance of four model attribution techniques in dealing with HFC removal. A small r indicates more distortions introduced in HFC. (b) Grad-cam of DNA-Det(left) method relying on HFC and AttNet(right) relying on LFC for attribution.

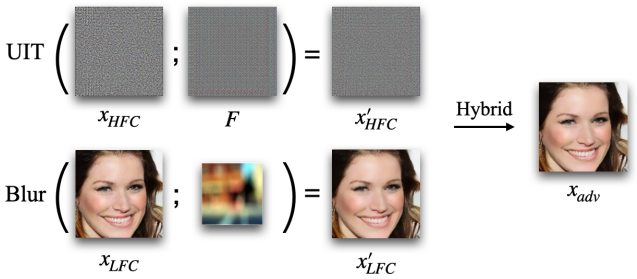


Figure 4: Adversarial perturbations crafted by our TraceEvader. We divide the HFC and LFC of the image x to add adversarial perturbations separately. For x_{HFC}, we apply the UIT attack by injecting a universal imitated trace F; For x_{LFC}, we apply the adversarial blur attack. Then we combine two components utilizing hybrid image transformation, where no visible artifacts exhibited.

architecture traces exhibit global consistency among full images, while the weight traces are concentrated on semantic-related local areas. These three findings tell us that the adversarial perturbations should disrupt the traces in both the HFC and LFC as the unknown attribution methods will adopt them possibly.

Methodology

Threat Model

Defender's Goals and Capabilities. The goal of the defender is to track the adopted forgery model and further provides explainable results to DeepFake forensics. The defender (1) trains the model attribution model with available fake images for attributing known forgery models, (2) performs simple input transformations or utilizes dedicated denoising methods with the help of collecting a small number of clean-perturbed pairs to remove the potential added adversarial perturbations intentionally.

Adversary's Goals and Capabilities. The attacker creates various realistic DeepFakes and improves the possibilities of surviving the potential DeepFake forensic methods, such as passive detection, forgery model attribution, *etc*.. Especially, the attacker is the owner of the model and has full knowledge of training dataset. To survive various potential model attribution techniques, the adversary may remove the traces involving the model modification or synthesized outputs manipulation to fool model attribution techniques finally.

Perturbating Model Traces

Our findings in Section indicate that the model fingerprints in the HFC are global consistency, while the model fingerprints in the LFC focus on local areas. Thus, it is non-trivial to generate unified perturbations to disrupt both the stable model fingerprints in HFC and distinct fingerprints in LFC as they are vastly different. In this paper, we craft the universal imitated trace (UIT) added to the HFC to disrupt the model fingerprints dedicated to the model-architecture attribution and introduce adversarial mean-shift blur to the LFC to prevent the extraction of model fingerprints for model-instance attribution, respectively. Figure 4 presents the two crafted adversarial perturbations for evading model attribution techniques.

Adversarial Imitated Traces. Motivated by the philosophy of ambiguity attack (Fan, Ng, and Chan 2019) in ownership verification where the adversary embeds a similar watermark into the victim model to claim his ownership. Then, the ownership of the target model is in doubt as the existence of their respective vouch. In this paper, we aim at crafting imitated traces to bring confusion to model attribution techniques like the ambiguous watermark in ownership verification in DNN models.

Residuals are shown to contain rich traces information related to generative models and extra random noises (Marra et al. 2019; Yu, Davis, and Fritz 2019). In this paper, we are inspired by previous works (Ulyanov, Vedaldi, and Lempitsky 2018; Wang et al. 2020c) that CNN has a tendency to learn universal and structured features in natural images before fitting to disordered random noise and are naturally embedded with trace prior (Sinitsa and Fried 2023) as its convolution and sampling operations are widely adopted in generator architectures. Intuitively, we can learn the universal and periodic trace pattern as UIT from the residuals by utilizing this CNN inductive bias. Specifically, let $\mathcal{X} \in \mathbb{R}^{3\times h\times w}$ represent image space and $\phi : \mathbb{R}^{c\times h\times w} \to \mathbb{R}^{3\times h\times w}$ demonstrate the generation process, which maps latent space to image space utilizing a CNN encoder-decoder network (Sinitsa and Fried 2023) ϕ. Given a mini-batch of n images $\boldsymbol{X} = [\mathbf{x}_1^{c_1}, ..., \mathbf{x}_n^{c_n}]$ where $c_i \in \{0, 1\}$ refers to generated images and real images, residuals $\boldsymbol{R}$ are extracted by a highpass filter (Zhang et al. 2017) f_{HP}, denoted as $\boldsymbol{R} = f_{HP}(\boldsymbol{X})$. Our goal is to generate the universal trace $\boldsymbol{F} \in \mathbb{R}^{3\times h\times w}$ exhibits a strong correlation with the traces in residuals of the generated images while showing less correlation with the clean residuals of real images. We formulate our task as follows:

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

where the the θ^* is optimized from parameters θ and random latent vector $\mathbf{z}$ follows uniform distribution.

In order to measure the correlation between the image residual and the $\boldsymbol{F}$, we adopt Pearson Correlation Coefficient as the correlation metric $\rho(\cdot)$, formulated as follows:

$$\rho(\boldsymbol{F}, \boldsymbol{R}) = \frac{E\left[\left(\boldsymbol{F} - \mu_F\right)\left(\boldsymbol{R} - \mu_R\right)\right]}{\sigma_F \sigma_R} \tag{3}$$

where σ and μ represent the standard deviations and mean values, respectively, $E(\cdot)$ calculates the mathematic expectation.

We design a contrastive loss function that forces residuals from different classes to have greater differentiation in the correlations with $\boldsymbol{F}$, promoting similarity in the correlations between $\boldsymbol{F}$ and residuals from the same class. The Euclidean distance metric is utilized in this work to quantify the correlation distance $\mathcal{D}_{ij}$ between two residual samples:

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

where $c_i \in \{0, 1\}$ refers to fake images and real images. When $c_i = c_j$, $\mathcal{D}_{ij}$ is denoted as $\mathcal{D}_{pos}$. Likewise, when $c_i \neq c_j$, $\mathcal{D}_{ij}$ is denoted as D_{neg}. The contrastive loss L_{con} is summarised below:

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

where m is a predefined margin parameter.

Adversarial Mean-shift Blur. As a self-crafted noise pattern, the former UIT has limited impact on the low frequency and semantics aspect of images. Due to its global consistency, UIT disturbs the image in a uniform manner, which shows less effectiveness in altering the overall pixel distribution and removing important local traces. Moreover, the interested region of the trace exploited by the victim model attribution techniques is unknown and varies among generative models. Thus, it is impossible to craft universal traces to bring confusion for the traces in LFC. In this paper, we employ the most straightforward idea by introducing blur to disrupt the model traces in LFC. Gaussian blurring mean shift (GBMS) is an effective edge-preserving filter that can smooth the low-frequency component and eliminate the imperfection or distortion artifacts of synthesis images (Hou et al. 2023).

GBMS, denoted as $G_\sigma(\cdot)$, calculates the positional density of each pixel in the spatial domain based on its color information and aggregates pixels with similar colors, resulting in a change in pixel distribution and a smoothing effect. To be specific, we first define a search window W for each pixel p_i in the image $\mathbf{x}$ (we set the window size as 3 in this work). Then we estimate the density of each pixel $p_j \in W$ by Gaussian kernel g_σ expressed as:

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

where σ determines the width of g and $\|\cdot\|$ indicates the Euclidean distance. The GBMS updates the position of the current pixel p_i with the mean shift vector dp:

$$\mathrm{d}p = \frac{\sum_{p_j \in W}\left(p_j - p_i\right) g_\sigma\left(p_i, p_j\right)}{\sum_{p_j \in W} g_\sigma\left(p_i, p_j\right)} \tag{7}$$

$$p_i' = p_i + \mathrm{d}p \tag{8}$$

We repeat this process until each pixel reaches its convergence.

Hybrid Image Transformation

After preparing the adversarial perturbations for the HFC and LFC, respectively, we add the adversarial perturbations in the created fake images with hybrid image transformation in a non-box manner. Formally, we separate the HFC and LFC of the image and perform attacks on both components. Then, these two perturbed components are combined to form the final adversarial hybrid image.

Specifically, we inject the crafted UIT into HFC residuals in order to produce ambiguous traces that are difficult to distinguish. This process can be formulated as:

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

where λ is a weight factor to control the intensity of perturbations. Next, we apply the GBMS function defined in Section to blur the low-frequency parts of the image:

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

Finally, we synthesize two-step manipulation to generate our adversarial hybrid image $\mathbf{x}_{adv}$ for evading model attribution techniques:

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

Experiments

Experiments Setup

Our experiments are conducted across two types of model attribution (*e.g.*, model-architecture attribution, model-instance attribution) on highly diverse data from 8 forgery models (*e.g.*, GANs, DMs). In experiments, we employ two kinds of adversarial attacks as the baselines, the one is the transfer-based attack (*i.e.*, BIM (Kurakin, Goodfellow, and Bengio 2016), MI-FGSM (Dong et al. 2018)). The other one is the non-box attack that served for evading the DeepFake detectors (*i.e.*, peak attack (Wesselkamp et al. 2022) and FakePolisher(Huang et al. 2020)). More details w.r.t. the evaluated model attribution techniques, evaluation metrics, employed baselines, and implementation details of our TraceEvader are available at the technical appendix.

To comprehensively evaluate our method, we perform effectiveness evaluation to explore whether the functionality has been compromised when we apply TraceEvader, robustness against common image transformation and intentional image denoising, and comparison with four baselines. Additionally, we also conduct ablation studies and extensive experiments to evaluate the effectiveness of our proposed method in evading GAN instance attribution and popular DeepFake detectors. The ablation studies and extensive experimental results refer to the technical appendix.

Effectiveness Evaluation

In this section, we evaluate the effectiveness of our TraceEvader against the model-architecture and model-instance techniques in the black-box setting and compare them with

four baselines. For the two transfer-based baselines (*i.e.*, BIM, MI-FGSM), adversarial perturbations are crafted by attacking DNA-Det in a white-box setting and hoping that it can be transferred to other model attribution techniques. For qualitative analysis, we visualize the spectrum of adversarial examples crafted by TraceEvader in Figure 2 (refer to the bottom row), where adversarial examples exhibit distinct frequency patterns from the original ones.

Evading Model-architecture Attribution. Experiments are conducted on DNA-Det (Yang et al. 2022), Reverse(Asnani et al. 2021) and DCT (Frank et al. 2020) where all of them focus on the traces in HFC. Experimental results in Table 1 illustrate that TraceEvader gives an average ASR **83.6%**. This indicates that all the model attribution techniques failed to serve the DeepFake forensic purposes. Our TraceEvader work perfectly in evading DNA-Det with an average ASR **97.95%** and Reverse with average ASR **90.6%**, outperforming all baselines. Our method yields relatively low ASR against DCT in some cases. Taking SNGAN as an example, we believe that this is because the added perturbation is not strong enough to cover up the strong peaks throughout the spectrum exhibited by SNGAN, referring to Figure 2. Therefore, we raise the value of λ to 0.02, and the attack success rate increases to **80.5%** with a PSNR value of 35.5 and an SSIM value of 0.9977, where the perturbed images do not show any visible artifacts.

Evading Model-instance Attribution. Experiments are conducted on AttNet (Yu, Davis, and Fritz 2019) where AttNet relies on traces in LFC. The huge gap between features exploited by AttNet and the other three methods leads to poor transfer performance of BIM and MI-FGSM. For non-box peak attack methods, they can only manipulate high frequencies of images, limiting their ability to attack AttNet. Our TraceEvader outperforms BIM, MI-FGSM and peak in all cases and obtains better average average attack success rate and image quality than FakePolisher. After retraining AttNet, the effectiveness of our attacks is not as significant as on the pre-trained model, yet this does not hinder us from achieving the highest average attack success rate up to **50.9%** against AttNet.

Image Quality Assessment. To demonstrate the imperceptibility of TraceEvader, we conduct both qualitative and quantitative analyses of perturbated examples. As illustrated in Figure 5, TraceEvader generates high quality samples with invisible perturbation. In contrast, samples with perturbations generated by MI-FGSM exhibit more noticeable distortions and images reconstructed by FakePolisher display a discernible level of blurriness. More samples can be found in the technical appendix. We present a detailed quality assessment with PSNR which is shown in the last columns of Table 1.

In summary, our TraceEvader achieves the highest average attack success rate of **79.07%** against both types of attribution, while the best average attack success rate among the four SOTA model attribution techniques is up to **97.95%**. Furthermore, our method had a top-2 attack success rate in more than 84% of cases. Furthermore, MI-FGSM and FakePolisher introduce visible noise, while our method maintain a relatively good image quality.

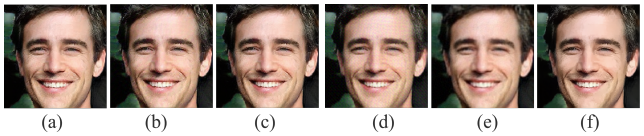


Figure 5: Visualization of the crafted adversarial examples. The visualizations include the raw fake images (a) and the perturbated images generated with BIM (b), peak (c), MI-FGSM (d), FakePolisher (e) and our proposed TraceEvader (f). Samples generated by MI-FGSM and FakePolisher exhibit visible artifacts when zoom in them.

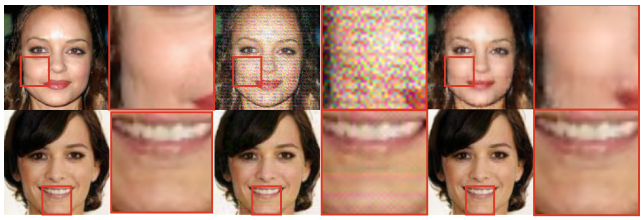


Figure 6: Visualization of the (a) original fake images (b) fake images added with crafted adversarial perturbations and (c) reconstructed images with denoising technique. To better illustrate the effectiveness of the employed denoising technique, the intensity of perturbations in the top row is larger than in the bottom row.

Robustness Evaluation

In a more real-world scenario, we consider a common phenomenon that the DeepFakes injected with our added perturbations will be corrupted via the popular image transformations (*e.g.*, compression, noise) and a strict assumption that the defender knows that the adversary may inject adversarial perturbations into the inputs to mislead the DeepFake attribution. In such circumstances, the defender will employ a powerful image-denoising method to defend the confusion of injected perturbations. Thus, in our experiments, we also investigate the robustness evaluation when the added perturbations are corrupted or maliciously denoised.

Surviving the intentional image denoising. To prove the strong capabilities of our proposed method in surviving the various attacks, we assume the defender collects a certain number of clean images and corresponding images injected with perturbations generated by our TraceEvader. In this paper, the popular Denoising Autoencoder (Vincent et al. 2008; Chiang et al. 2019; Zhang et al. 2020; Lee et al. 2021) is adopted as the backbone to remove noises. We train this denoise model with the supervision of image pairs and test it on new adversarial samples. As illustrated in Figure 6, the denoising model indeed eliminates the noticeable fingerprint artifacts. However, the result in Table 2, reconstructed adversarial examples still maintain high ASR, demonstrate that our attack survive from intentional image denoising well.

Surviving the common image transformations. We consider two common image transformations (*e.g.*, Gaussian

GMs	Methods	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
	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**	98.2	33.2	45.1*	31.9
	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
	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
	TraceEvader	95.2*	**99.3**	90.6*	**88.3**	36.3		**100.0**	**100.0**	61.8	22.7*	38.4

Table 1: The performance of TraceEvader in evading the four model attribution techniques measured by ASR (%). PSNR and SSIM are employed for measuring the image quality after adding adversarial perturbations. The symbol - denotes a white-box attack. We mark the top-2 ASR by bold and *, respectively. For the first two rows (*e.g.*, ProGAN, MMDGAN, SNGAN, and CramerGAN), we attack the pre-trained models provided by the four model attribution techniques. For the last two rows (*e.g.*, CycleGAN, StyleGNA2, PNDM, LDM), we train the model by ourselves for conducting the evasion attack.

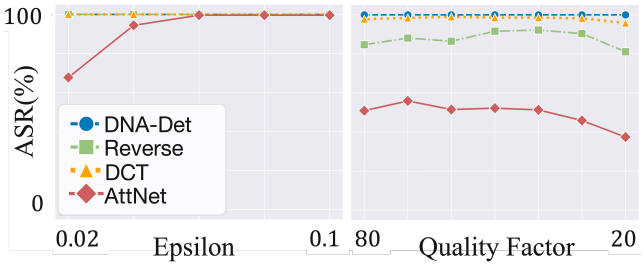
	DNA-Det	Reverse	DCT	AttNet
Adv	73.4	65.4	100	50.1
Recovered	75	65.9	75	84

Table 2: Comparison of ASR(%) between original adversarial samples and carefully denoised ones.

noises, and JPEG compression) to explore whether the transformation brings any degradation to our adversarial samples. Figure 7(a) shows that adding noise will not degrade our attack performance, because the traceable fingerprint has been destroyed already. In fact, noises will introduce even greater perturbations to images, which can further increase the attack success rate. Figure 7(b) shows that the attack performance will drop slightly only after extremely heavy compression. Experimental results illustrate that our method can survive common image degradation well.

Conclusion

In this paper, we investigate and introduce a non-box and training-free evasion attack TraceEvader against the popular model attribution techniques. To the best of our knowledge, this is the very first work to reveal the vulnerability of model attribution techniques which are vulnerable to imitated adversarial perturbations. Moreover, our studies observe that model traces appeared in both HFC and LFC where the traces in HFC are architecture-relevant and the traces in LFC are instance-relevant. More powerful defense mechanisms for DeepFake forensics should be proposed.



(a) Gaussian Noise (b) JPEG Compression

Figure 7: Performance in surviving the image transformation under different intensities.

Technical Appendix

In the technical appendix, we present the related works, the details of experimental setting, ablation studies, extensive experiments results. Additionally, the limitation and social impacts of our proposed TraceEvader are also refer to the technical appendix[2].

Acknowledgements

This research was supported in part by the National Key Research and Development Program of China under No.2021YFB3100700, the National Natural Science

[2]http://wangrun.github.io/supp/TraceEvader_AAAI2024.pdf

Foundation of China (NSFC) under Grants No. 62202340, 62372334, the CCF-NSFOCUS 'Kunpeng' Research Fund under No. CCF-NSFOCUS 2023005, the Open Foundation of Henan Key Laboratory of Cyberspace Situation Awareness under No. HNTS2022004, Wuhan Knowledge Innovation Program under No. 2022010801020127, the Fundamental Research Funds for the Central Universities under No. 2042023kf0121, the Natural Science Foundation of Hubei Province under No. 2021CFB089.

References

Asnani, V.; Yin, X.; Hassner, T.; and Liu, X. 2021. Reverse Engineering of Generative Models: Inferring Model Hyperparameters from Generated Images. arXiv:2106.07873.

Carlini, N.; and Farid, H. 2020. Evading Deepfake-Image Detectors With White- and Black-Box Attacks. In *Proceedings of the IEEE/CVF Computer Vision and Pattern Recognition (CVPR) Workshops*.

Chakraborty, A.; Alam, M.; Dey, V.; Chattopadhyay, A.; and Mukhopadhyay, D. 2018. Adversarial attacks and defences: A survey. *arXiv preprint arXiv:1810.00069*.

Chiang, H.-T.; Hsieh, Y.-Y.; Fu, S.-W.; Hung, K.-H.; Tsao, Y.; and Chien, S.-Y. 2019. Noise reduction in ECG signals using fully convolutional denoising autoencoders. *Ieee Access*, 7: 60806–60813.

Dhariwal, P.; and Nichol, A. 2021. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, 34: 8780–8794.

Dolhansky, B.; Bitton, J.; Pflaum, B.; Lu, J.; Howes, R.; Wang, M.; and Ferrer, C. C. 2020. The deepfake detection challenge (dfdc) dataset. *arXiv preprint arXiv:2006.07397*.

Dong, Y.; Liao, F.; Pang, T.; Su, H.; Zhu, J.; Hu, X.; and Li, J. 2018. Boosting Adversarial Attacks with Momentum. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Durall, R.; Keuper, M.; and Keuper, J. 2020. Watch Your Up-Convolution: CNN Based Generative Deep Neural Networks Are Failing to Reproduce Spectral Distributions. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Dzanic, T.; Shah, K.; and Witherden, F. 2019. Fourier Spectrum Discrepancies in Deep Network Generated Images.

Fan, L.; Ng, K. W.; and Chan, C. S. 2019. Rethinking deep neural network ownership verification: Embedding passports to defeat ambiguity attacks. *Advances in neural information processing systems*, 32.

Frank, J.; Eisenhofer, T.; Schönherr, L.; Fischer, A.; Kolossa, D.; and Holz, T. 2020. Leveraging Frequency Analysis for Deep Fake Image Recognition. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20. JMLR.org.

Hou, Y.; Guo, Q.; Huang, Y.; Xie, X.; Ma, L.; and Zhao, J. 2023. Evading DeepFake Detectors via Adversarial Statistical Consistency. *arXiv preprint arXiv:2304.11670*.

Huang, Y.; Juefei-Xu, F.; Wang, R.; Guo, Q.; Ma, L.; Xie, X.; Li, J.; Miao, W.; Liu, Y.; and Pu, G. 2020. FakePolisher: Making DeepFakes More Detection-Evasive by Shallow Reconstruction. In *Proceedings of the 28th ACM International Multimedia*.

Hussain, S.; Neekhara, P.; Jere, M.; Koushanfar, F.; and McAuley, J. 2021. Adversarial Deepfakes: Evaluating Vulnerability of Deepfake Detectors to Adversarial Examples. In *Proceedings of the IEEE/CVF Winter Applications of Computer Vision (WACV)*, 3348–3357.

Jia, S.; Ma, C.; Yao, T.; Yin, B.; Ding, S.; and Yang, X. 2022. Exploring frequency adversarial attacks for face forgery detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4103–4112.

Juefei-Xu, F.; Wang, R.; Huang, Y.; Guo, Q.; Ma, L.; and Liu, Y. 2022. Countering malicious deepfakes: Survey, battleground, and horizon. *International Journal of Computer Vision*, 130(7): 1678–1734.

Karras, T.; Laine, S.; Aittala, M.; Hellsten, J.; Lehtinen, J.; and Aila, T. 2020. Analyzing and Improving the Image Quality of StyleGAN. In *Proc. CVPR*.

Kurakin, A.; Goodfellow, I.; and Bengio, S. 2016. Adversarial machine learning at scale. *arXiv preprint arXiv:1611.01236*.

Lee, W.-H.; Ozger, M.; Challita, U.; and Sung, K. W. 2021. Noise learning-based denoising autoencoder. *IEEE Communications Letters*, 25(9): 2983–2987.

Leong, D. 2023. Deepfakes and Disinformation Pose a Growing Threat in Asia.

Liu, L.; Ren, Y.; Lin, Z.; and Zhao, Z. 2022. Pseudo numerical methods for diffusion models on manifolds. *arXiv preprint arXiv:2202.09778*.

Marra, F.; Gragnaniello, D.; Verdoliva, L.; and Poggi, G. 2019. Do GANs leave artificial fingerprints. In *2019 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*.

Moneywatch. 2023. Deepfake pornography could be a growing problem as AI editing programs become more sophisticated_2023. *CBS News*.

Moosavi-Dezfooli, S.-M.; Fawzi, A.; Fawzi, O.; and Frossard, P. 2017. Universal adversarial perturbations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1765–1773.

Mopuri, K. R.; Ojha, U.; Garg, U.; and Babu, R. V. 2018. Nag: Network for adversary generation. In *Proceedings of the IEEE Computer Vision and Pattern Recognition*, 742–751.

Sinitsa, S.; and Fried, O. 2023. Deep Image Fingerprint: Towards Low Budget Synthetic Image Detection and Model Lineage Analysis. arXiv:2303.10762.

Ulyanov, D.; Vedaldi, A.; and Lempitsky, V. 2018. Deep image prior. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 9446–9454.

Vincent, P.; Larochelle, H.; Bengio, Y.; and Manzagol, P.-A. 2008. Extracting and Composing Robust Features with Denoising Autoencoders. In *Proceedings of the 25th International Conference on Machine Learning*, ICML '08,

1096–1103. New York, NY, USA: Association for Computing Machinery. ISBN 9781605582054.

Wang, H.; Wu, X.; Huang, Z.; and Xing, E. P. 2020a. High-frequency component helps explain the generalization of convolutional neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 8684–8694.

Wang, R.; Huang, Z.; Chen, Z.; Liu, L.; Chen, J.; and Wang, L. 2022. Anti-Forgery: Towards a Stealthy and Robust DeepFake Disruption Attack via Adversarial Perceptual-aware Perturbations. *arXiv preprint arXiv:2206.00477*.

Wang, R.; Juefei-Xu, F.; Luo, M.; Liu, Y.; and Wang, L. 2021. FakeTagger: Robust Safeguards against DeepFake Dissemination via Provenance Tracking. In *Proceedings of the 29th ACM International Conference on Multimedia*, MM '21, 3546–3555. New York, NY, USA: Association for Computing Machinery. ISBN 9781450386517.

Wang, R.; Juefei-Xu, F.; Ma, L.; Xie, X.; Huang, Y.; Wang, J.; and Liu, Y. 2020b. FakeSpotter: A Simple yet Robust Baseline for Spotting AI-Synthesized Fake Faces. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*.

Wang, S.-Y.; Wang, O.; Zhang, R.; Owens, A.; and Efros, A. A. 2020c. CNN-generated images are surprisingly easy to spot...for now. In *CVPR*.

Wesselkamp, V.; Rieck, K.; Arp, D.; and Quiring, E. 2022. Misleading Deep-Fake Detection with GAN Fingerprints. In *2022 IEEE Security and Privacy Workshops (SPW)*, 59–65.

Yang, T.; Huang, Z.; Cao, J.; Li, L.; and Li, X. 2022. Deepfake Network Architecture Attribution. In *Proceedings of the 36th AAAI Conference on Artificial Intelligence (AAAI 2022)*.

Yu, N.; Davis, L.; and Fritz, M. 2019. Attributing Fake Images to GANs: Learning and Analyzing GAN Fingerprints. In *IEEE International Conference on Computer Vision (ICCV)*.

Zhang, C.; Zhou, L.; Zhao, Y.; Zhu, S.; Liu, F.; and He, Y. 2020. Noise reduction in the spectral domain of hyperspectral images using denoising autoencoder methods. *Chemometrics and Intelligent Laboratory Systems*, 203: 104063.

Zhang, K.; Zuo, W.; Chen, Y.; Meng, D.; and Zhang, L. 2017. Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising. *IEEE transactions on image processing*, 26(7): 3142–3155.

Zhao, Y.; Ge, W.; Li, W.; Wang, R.; Zhao, L.; and Ming, J. 2020. Capturing the persistence of facial expression features for deepfake video detection. In *Information and Communications Security: 21st International Conference, ICICS 2019, Beijing, China, December 15–17, 2019, Revised Selected Papers 21*, 630–645. Springer.

Zhu, J.-Y.; Park, T.; Isola, P.; and Efros, A. A. 2017. Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networkss. In *Computer Vision (ICCV), 2017 IEEE International Conference on computer vision*.