Making DeepFakes More Spurious: Evading Deep Face Forgery Detection via Trace Removal Attack

# Chi Liu; Huajie Chen; Tianqing Zhu et al.

## 2023

The experimental results show that the proposed attack can significantly compromise the detection accuracy of six state-of-the-art DeepFake detectors while causing only a negligible loss in visual quality to the original DeepFake samples.

# Abstract

DeepFakes are raising significant social concerns. Although various DeepFake detectors have been developed as forensic countermeasures, these detectors are still vulnerable to attacks. Recently, a few attacks, principally adversarial attacks, have succeeded in cloaking DeepFake images to evade detection. However, these attacks have typical detector-specific designs, which require prior knowledge about the detector, leading to poor transferability. Moreover, these attacks only consider simple security scenarios. Less is known about how effective they are in high-level scenarios where either the detector’s defensive capability or the attacker’s knowledge varies. In this paper, we aim to solve the above challenges with presenting a novel attack pattern for DeepFake anti-forensics, namely, the trace removal attack. Instead of investigating the detector side, this trace removal attack looks into the original DeepFake creation pipeline, attempting to remove all detectable natural DeepFake traces to render the fake images more “authentic”. This detector-agnostic design benefits the attack to be effective against arbitrary or even unknown detectors. To implement this attack, first, we perform an in-depth DeepFake trace discovery, which identifies three discernible traces: spatial anomalies, spectral disparities, and noise fingerprints. Then a trace removal network (TR-Net) is proposed based on an adversarial learning framework involving one generator and multiple discriminators. Each discriminator is responsible for one individual trace representation to avoid cross-trace interference. These multiple discriminators are arranged in parallel, which prompts the generator to remove various traces simultaneously. To evaluate the efficacy of the attack, we crafted heterogeneous security scenarios where the detectors were embedded with different levels of defense and the attackers’ background knowledge of data varies. The experimental results show that the proposed attack can significantly compromise the detection accuracy of six state-of-the-art DeepFake detectors while causing only a negligible loss in visual quality to the original DeepFake samples.

# Study subjects

**66000 semantically-closest pairs**

Thus, for a thorough evaluation, we created All-in-One-DF, a new DeepFake dataset based on previous datasets. **The All-in-One-DF dataset consists of 66, 000 semantically-closest pairs of real and fake images (i.e., 132, 000 images in total) from four sources: (1) CelebA: A large-scale dataset containing more than 200k real face images**. The images are cropped and aligned to the size of 128 ∗ 128 ∗ 3 with the face in the centre

# Findings

All detectors here reached a satisfactory level of accuracy at over 90.00%, except for the NF detector

After the trace removal attack, the classification accuracy of all detectors had decreased markedly, and the average accuracy of the six had dropped from 90.89% to 25.66%

(d) PSNR and SSIM in the out-of-distribution DeepFake case group (where ProGAN is not included in the training set), the Xception’s detection accuracy increases from 36.12% to 40.21% and the F3-Net’s detection accuracy increases from 20.14% to 30.90% (Figure 14.c)

# Scholarcy Highlights

* Along with the recent progress in automated digital face manipulation techniques based on deep learning, deep face forgeries, known as DeepFakes, are raising serious social concerns for information security [1]
* We focused on an anti-forensics attack against DeepFake detectors
* We presented a novel detector-agnostic attack, called a trace removal attack, that is capable of refining DeepFake images by removing all possible DeepFake traces via an one-versus-multiple adversarial learning network
* The refined DeepFake images are closer to the real images and can bypass arbitrary and even unknown detectors
* We assessed the efficacy of the trace removal attack against a wide range of state-of-theart detectors in heterogeneous high-level security scenarios where the detectors were embedded with various defensive strategies and the attacker’s knowledge of data was limited
* The proposed trace removal attack achieves the highest attack effectiveness while introducing minimal visual quality loss compared with contemporary adversarial and reconstruction-based attacks

# Scholarcy Summary

## Introduction

Along with the recent progress in automated digital face manipulation techniques based on deep learning, deep face forgeries, known as DeepFakes, are raising serious social concerns for information security [1].

Even in a universal black-box attack scenario, information from surrogate detectors is always needed to imitate the behavior of the target detector

These detector-specific designs lead to poor transferability and a lack of stability across different detectors or unknown detectors [14], [15].

Other attacks emerging in this field generally require reconstructing the DeepFake samples to modify the distribution of features-of-interest of the target detector to evade detection [10]–[13]

This is a detector-specific design, which means that these attacks are less transferable to detectors interested in different forgery features.

These attacks only pay attention to a single type of feature.

Their efficacy may deteriorate significantly against advanced detectors that operate on hybrid features

## Objectives

We aim to solve the above challenges with presenting a novel attack pattern for DeepFake anti-forensics, namely, the trace removal attack

## Findings

All detectors here reached a satisfactory level of accuracy at over 90.00%, except for the NF detector.

(d) PSNR and SSIM in the out-of-distribution DeepFake case group, the Xception’s detection accuracy increases from 36.12% to 40.21% and the F3-Net’s detection accuracy increases from 20.14% to 30.90% (Figure 14.c)

## Discussion

The above model trace analysis identifies three typical model traces throughout the complete DeepFake pipeline

The interplay of these traces distinguishes DeepFake images from real ones.

A solid and universal trace removal attack is desired to eliminate all these possible traces

In this way, the distribution of DeepFake images can be much closer to that of the real ones in the trace feature space, by which the modified DeepFake images can evade arbitrary detectors.

Since our knowledge of DeepFake traces is derived from the fundamental pipeline shared by shared by the different methods used to generate DeepFakes, the trace removal attack to be presented is applicable to all these DeepFake types.

Assume the target victim model is an arbitrary DeepFake detector C, which is a machine learning classifier that distinguishes trace features between real and DeepFake images.

The attacking goal of fraudulence is well satisfied by the proposed trace removal attack

## Conclusion

We focused on an anti-forensics attack against DeepFake detectors.

We presented a novel detector-agnostic attack, called a trace removal attack, that is capable of refining DeepFake images by removing all possible DeepFake traces via an one-versus-multiple adversarial learning network.

The refined DeepFake images are closer to the real images and can bypass arbitrary and even unknown detectors.

We assessed the efficacy of the trace removal attack against a wide range of state-of-theart detectors in heterogeneous high-level security scenarios where the detectors were embedded with various defensive strategies and the attacker’s knowledge of data was limited.

The proposed trace removal attack achieves the highest attack effectiveness while introducing minimal visual quality loss compared with contemporary adversarial and reconstruction-based attacks.

We will focus on developing more robust forensics countermeasures against trace removal attacks

# Builds on previous work

Notably, since the subtle spatial anomalies may be imperceptible to humans but can be captured by machines, we demonstrate their existence and spatial distributions in the RGB color space with the spatial attention map (SAM) of a toy Xception detector. **Grad-CAM [51] is used to calculate the SAMs regarding different DeepFake types** (Details of these DeepFake types and the Xception detector is introduced in Section 5)

(2) Face synthesis: We employed ProGAN, one of the most popular unconditional GANs to synthesize nonexisting face images. **We utilize th**e pre-trained1 ProGAN instance [42] to generate 22, 000 fake images

(3) Facial attribute editing: We selected STGAN, a stateof-the-art GAN for facial attribute editing for this use. **We randomly** sampled 22, 000 real images from the remaining CelebA dataset and applied the official pre-trained1 STGAN instance [20] to modify the attributes on these real samples, 1

# Contributions

CONCLUSION AND FUTURE WORKIn this paper, we focused on an anti-forensics attack against DeepFake detectors. We presented a novel detector-agnostic attack, called a trace removal attack, that is capable of refining DeepFake images by removing all possible DeepFake traces via an one-versus-multiple adversarial learning network. The refined DeepFake images are closer to the real images and can therefore bypass arbitrary and even unknown detectors. We assessed the efficacy of the trace removal attack against a wide range of state-of-theart detectors in heterogeneous high-level security scenarios where the detectors were embedded with various defensive strategies and the attacker’s knowledge of data was limited. Our findings reveal that, the proposed trace removal attack achieves the highest attack effectiveness while introducing minimal visual quality loss compared with contemporary adversarial and reconstruction-based attacks. In the future, we will focus on developing more robust forensics countermeasures against trace removal attacks.