Anti-Forgery: Towards a Stealthy and Robust DeepFake Disruption Attack via Adversarial Perceptual-aware Perturbations

Run Wang[1,2*] , **Ziheng Huang**[1,2] , **Zhikai Chen**[3] , **Li Liu**[4] , **Jing Chen**[1,2] , **Lina Wang**[1,2,5]

[1]School of Cyber Science and Engineering, Wuhan University, China

[2]Key Laboratory of Aerospace Information Security and Trusted Computing, Ministry of Education, China

[3]Tencent Zhuque Lab

[4]Fudan Development Institute, Fudan University, China

[5]Zhengzhou Xinda Institute of Advanced Technology

Abstract

DeepFake is becoming a real risk to society and brings potential threats to both individual privacy and political security due to the DeepFaked multimedia are realistic and convincing. However, the popular DeepFake passive detection is an ex-post forensics countermeasure and failed in blocking the disinformation spreading in advance. To address this limitation, researchers study the proactive defense techniques by adding adversarial noises into the source data to disrupt the DeepFake manipulation. However, the existing studies on proactive DeepFake defense via injecting adversarial noises are not robust, which could be easily bypassed by employing simple image reconstruction revealed in a recent study MagDR.

In this paper, we investigate the vulnerability of the existing forgery techniques and propose a novel *anti-forgery* technique that helps users protect the shared facial images from attackers who are capable of applying the popular forgery techniques. Our proposed method generates perceptual-aware perturbations in an incessant manner which is vastly different from the prior studies by adding adversarial noises that is sparse. Experimental results reveal that our perceptual-aware perturbations are robust to diverse image transformations, especially the competitive evasion technique, MagDR via image reconstruction. Our findings potentially open up a new research direction towards thorough understanding and investigation of perceptual-aware adversarial attack for protecting facial images against DeepFakes in a proactive and robust manner. Code is available at https://github.com/AbstractTeen/AntiForgery.

1 Introduction

In recent years, we have witnessed the remarkable development of GAN in image synthesis and fine-grained image ma-

*Run Wang's Email: wangrun@whu.edu.cn

nipulation. Attackers could leverage GAN to generate realistic and natural synthetic images, audios, and videos (also known as DeepFake), which poses potential security and privacy concerns to individuals [Juefei-Xu *et al.*, 2021]. In this AI era, we are living in a world where we cannot believe our eyes and ears anymore. Thus, effective countermeasures should be developed for fighting against DeepFakes.

To defend against DeepFakes, various countermeasures are developed in both passive and proactive manner. However, both of them are still in their early stage and not prepared for tackling this emerging severe threat. The passive DeepFake detection merely determines the real or fake which is an ex-post DeepFake defense manner [Hu *et al.*, 2021]. More importantly, the existing DeepFake detection techniques can not well tackle the DeepFakes created by unknown synthetic techniques [Wang *et al.*, 2021a]. According to a result of the DeepFake detection challenge (DFDC) held by Facebook, the winner team can give a detection accuracy less than 70%. Thus, some researchers are working on developing proactive DeepFake defending techniques by adding adversarial noises into the source images to disrupt the DeepFake creation [Yang *et al.*, 2020; Segalis, 2020]. Specifically, they hope that the created DeepFakes with added perturbations exhibit visually noticeable artifacts and provide signals for detectors. The DeepFake disruption is the most promising countermeasures for fighting against DeepFakes in a proactive manner, which shows potential for tackling DeepFakes in the wild [Chen *et al.*, 2021]. However, existing DeepFake disruption studies by adding adversarial noises suffer challenge in tackling input transformations, which limits their practicality applications.

Existing studies mostly borrow the idea of prior adversarial attack (*e.g.*, FGSM, PGD) via gradient-based or optimization-based strategies to generate imperceptible adversarial noises. However, a series of studies reveal that such adversarial noises could be easily removed or destroyed. A recent study, MagDR [Chen *et al.*, 2021], illustrated that a simple input reconstruction could destroy the added adversarial noises for disrupting DeepFakes.

The ultimate goal of proactive DeepFake defense is to create low-quality DeepFakes when applying various techniques for forgery purpose. The created low-quality DeepFakes will exhibit noticeable artifacts and could be easily
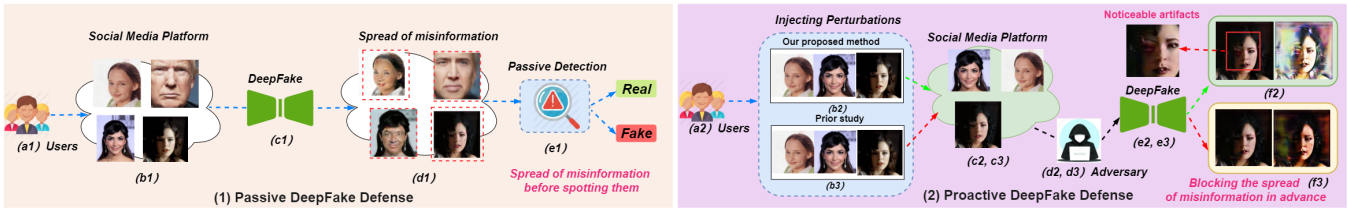
Figure 1: An overview of the difference between passive and proactive DeepFake defense. The passive DeepFake detection failed in blocking the spread of misinformation before the fake is confirmed, while the proactive DeepFake defense provides a restricted spread of misinformation as the noticeable artifacts exhibited in the created DeepFakes. In contrast to the prior study working on RGB color space, our proposed anti-forgery method operating on the Lab color space could survive the input transformation well with noticeable artifacts to block the further spread of misinformation.

spotted even with simple DeepFake detectors. Thus, a desired proactive DeepFake defense technique should satisfy the following three properties, *robust* to input transformations (*e.g.*, reconstruction), visually *natural* perturbations to human eyes, working on the *black-box* settings without obtaining any knowledge of the forgery model.

In this paper, we investigate the vulnerability of GANs in image synthesis and seek a kind of robust adversarial perturbation which could survive drastic input transformation. Specifically, we propose a novel anti-forgery approach by adding perceptual-aware perturbations into the source data to enforce the created DeepFakes exhibiting noticeable artifacts and detectable with simple DeepFake detectors. Our perceptual-aware perturbations operate on the *Lab* color space in the de-correlated a and b channel in an incessant manner. Unlike the prior studies working on the *RGB* color space which introduces perceivable distortions incurred by unnatural colors even with small variations [Huang *et al.*, 2021], our proposed method works on a natural-color range of semantic categories which generates perceptually natural facial images. More importantly, our generated perceptual-aware perturbation is robust against input transformation which shows potential to be deployed in the real world. Figure 1 illustrates the comparison with the existing common DeepFake disruption techniques by adding adversarial noises on the RGB color space.

Extensive experiments conducted on the three types of DeepFake demonstrate the effectiveness of our proposed method and robustness against the input transformation, like reconstruction, compression. Additionally, some ablation studies are conducted for illustrating that our added perturbations could provide a clear signal for DeepFake detection. Our main contributions are summarized as follows:

- We introduce a novel anti-forgery method by adding adversarial perceptual-aware perturbations against forgery in a proactive manner by transforming facial images in an incessant and natural way, as opposed to adding meaningless and sparse noises in prior studies which are not robust to drastic input transformations, like reconstruction.

- We employ a simple yet effective method to generate the perceptual-aware perturbations by operating on the *Lab* color space in the decorrelated a and b channel that generates a visually natural face which could be further leveraged disrupting the DeepFake creation.

- For the first time, we conduct experiments on three types of DeepFakes to demonstrate the effectiveness of our devised perceptual-aware perturbations in creating DeepFakes with noticeable artifacts and robustness in surviving the input reconstruction compared with prior studies.

- Our research findings hint a new research direction towards perceptual-aware adversarial attack by investigating natural and robust perturbations against various transformations. We hope that more studies working on exploiting the vulnerabilities of GAN in image synthesis to develop DeepFake proactive countermeasures which could fight against DeepFakes in the wild.

2 Related Work

2.1 DeepFake Creation

In the past years, GAN and its variants have achieved tremendous success in image synthesis and manipulation. DeepFake leverages the success of GANs to generate forgery facial images or videos, which pose potential threats to individual privacy and political security. We are living in a world where we cannot believe our eyes anymore. In general, *entire synthesis*, *attribute editing*, *identity swap*, and *face reenactment* are four common types of DeepFakes.

The entire synthesis produces non-existent images in the world with random vectors to the network, such as PGGAN, StyleGAN. The attribute editing is a fine-grained face manipulation by modifying the simple face attributes (*e.g.*, hair color, bald) and complex face attributes (*e.g.*, gender, age) with popular GANs, like STGAN [Liu *et al.*, 2019], StarGAN [Choi *et al.*, 2018]. The identity swap is known as the popular face swap by swapping the face between target and source facial image. FaceSwap[1] and DeepFaceLab [Perov *et al.*, 2020] are two popular available tools for identity swap. Similar to identity swap, face reenactment also known as expression swap involves facial expression swap between target and source facial image by applying popular tools, such as Face2Face [Thies *et al.*, 2016].

Among the four types of DeepFake, the entire synthesis produces non-existent facial images without involving any source image manipulation, thus it will not bring any privacy or security concerns to individuals. In this paper, we mainly

---

[1]https://github.com/Oldpan/Faceswap-Deepfake-Pytorch

focus on the other three types DeepFakes (*e.g.*, attribute editing, identity swap, and face reenactment), which pose security and privacy concerns to individuals.

2.2 DeepFake Defense

DeepFake Detection. Determining if the facial image is real or fake by observing the subtle differences between real and fake images is a straightforward idea for defending Deep-Fakes. However, DeepFake detection is an ex-post forensic manner and suffers various known challenges, such as generalization to tackle unknown synthetic techniques, robust against quality degradations, evasion to adversarial attacks, which are obstacles for its practical usage to be deployed in the wild [Juefei-Xu *et al.*, 2021].

Studies have shown that the subtle differences between real and fake could be revealed in the spatial [Wang *et al.*, 2020; Wang *et al.*, 2021b] and frequency domain [Qian *et al.*, 2020], however, they are susceptible to adversarial attack via various image-level and frequency-level manipulations [Carlini and Farid, 2020; Huang *et al.*, 2020]. Another line work investigate the biological signals which are hard to replicate [Hu *et al.*, 2021]. Unfortunately, these passive DeepFake detection techniques cannot be applied for blocking the wide spreading of DeepFakes in advance before causing damage impacts.

DeepFake Disruption. To address the aforementioned limitations of DeepFake detection, DeepFake disruption aims at disrupting the images being DeepFaked proactively by adding adversarial perturbations into the source image to produce damaged images. Recently, massive studies have shown that Deep Neural Networks (DNN) are vulnerable to adversarial attack (*e.g.*, FGSM, PGD) by adding imperceptible noises to craft so-called adversarial examples. Thus, some researchers explore whether the DeepFakes are vulnerable to adversarial examples as well.

Yang *et al.* leveraged differentiable random image transformations to yield adversarial faces with noticeable artifacts [Yang *et al.*, 2020]. Ruiz *et al.* presented a spread-spectrum adversarial attacks against conditional image translation network in a grey-box scenario [Ruiz *et al.*, 2020]. Yeh *et al.* proposed two types of adversarial attack against image translation GANs with designed adversarial loss function by gradient optimization to output blurred and distorted output [Yeh *et al.*, 2020]. Huang *et al.* [Huang *et al.*, 2021] proposed an initiative defense framework against facial manipulation by injecting venom into the target manipulation model in the black-box settings. This work adds invisible perturbations by training a surrogate model, however, they merely claim its effectiveness in attribute editing and face reenactment since these two types of manipulation model share a similar pipeline. Unfortunately, it will be hardly applied in defending identity swap which is more challenge to find an appropriate surrogate model to generate perturbations.

Unfortunately, the existing DeepFake disruption techniques employ a naive adversarial attack to generate invisible perturbations which could be easily detected and corrupted by input reconstruction. A recent study MagDR [Chen *et al.*, 2021] also demonstrates this and reveals that the Deep-Fake disruption techniques by adopting C&W and PGD are all failed in producing damaged images. Specifically, MagDR

employs image restoration techniques to remove the added perturbations and obtains the desired form further.

In this paper, we challenge the effectiveness of MagDR in dealing with our perceptual-aware perturbations which are generated in an incessant and natural manner. We aim at developing stealthy and robust imperceptible adversarial perturbations which could effectively survive MagDR and provide strong protection to our shared faces in social media.

3 Problem Statement

In this paper, we describe the DeepFake defense from both passive DeepFake detection and proactive DeepFake defense in Figure 1. In the real world, a user uploads his/ her personal facial images to social media, like Twitter to share with friends or anyone on the Internet. Unfortunately, an attacker could easily pick the victim's photos for malicious manipulation with GAN to create fake pornography for individuals to raise privacy concerns or to generate a fake official statement for celebrities to cause panic. Our method adds adversarial perceptual-aware perturbations into personal images before uploading to social media without introducing any visual artifacts. The facial image with our applied perturbations could prevent the GAN-based manipulations well by generating damaged facial images with noticeable artifacts. The key idea here is that our perturbed facial images should be robust enough to survive input transformations, like reconstruction before the DeepFake process. Finally, the proactive Deep-Fake defense is achieved for creating damaged images and exhibiting clear fake signals for detectors.

Here, we elaborate the details regarding Figure 1. In the left panel, we present the pipeline of passive DeepFake detection. A user (Figure 1-a1) first uploads personal facial images to the social media platform (Figure 1-b1). Then, attackers pick the shared photos to create DeepFake (Figure 1-c1) (*e.g.*, change the color of hair, the intensity of facial expressions) which causes the spread of misinformation in the platform (Figure 1-d1). Finally, the DeepFake detector (Figure 1-e1) determines whether the suspicious one is real or fake in an ex-post forensics manner to further block the misinformation spreading in social media platforms if fake is confirmed. In the right panel, we present the pipeline of proactive Deep-Fake defense by comparing our proposed method with the prior DeepFake disruption approach by employing popular adversarial noises.

In contrast to the passive DeepFake detection in the left panel, a user first (Figure 1-a2) injects perturbations into facial images before uploading them to social media platforms (Figure 1-c2, c3). Our proposed method injects perceptual-aware perturbations by operating the Lab color space in an incessant manner (Figure 1-b2), in comparing with prior studies inject imperceptible adversarial noises by perturbing the RGB color space in a sparse manner (Figure 1-b3). However, in a real battleground scenario, an adversary (Figure 1-d2, d3) tries to remove or corrupt the added perturbations by employing input transformations (*e.g.*, reconstruction) before creating DeepFakes. In creating DeepFakes (Figure 1-e2, e3), our proposed method adding perceptual-aware perturbations could survive the adversarial attack with input transforma-

tions and exhibit noticeable artifacts in the manipulated facial images (Figure 1-f2) and blocked immediately without further spreading to cause any panic and privacy concerns for individuals. However, the prior studies are not robustly output realistic and natural DeepFakes which cause the spreading of misinformation (Figure 1-f3).

In summary, in such a real battleground scenario, both the passive DeepFake detection and prior proactive Deep-Fake disruption methods failed in blocking the misinformation spreading before resulting severe social impacts.

4 Methodology

4.1 Insight

DeepFake disruption is a promising countermeasure for defending DeepFakes proactively [Chen et al., 2021]. However, the existing studies suffer two issues, 1) they are not robust to the input transformation which is demonstrated in MagDR, 2) the effectiveness on all types of DeepFakes is unclear [Huang et al., 2021], which is an obstacle for its practical usage. Thus, the community is not prepared for tackling this emerging threat.

A straightforward idea for defending GAN-based Deep-Fake proactively is to explore the vulnerability of GAN in image synthesis and manipulation. To this end, adversarial examples are employed for attacking GAN by adding imperceptible perturbations into the input to introduce distortions on the output. Generally speaking, the created output with perturbations should satisfy the following three properties when defending DeepFakes in the wild.

- Visually natural to human eyes. The uploaded facial images to the social media platform are mainly used to share exciting moments to friends at the most of the time, thus the visually quality degradation is not allowed.

- Robust to input transformation (e.g., reconstruction). In the real world, adversaries will corrupt the added perturbations intentionally to break the defense mechanism.

- Generalizing well to a wide types of DeepFakes. A facial image spreading in the social media platform would suffer various forgery which all cause security and privacy concerns to individuals.

Inspired by the advantages of Lab color space that are perceptual uniform and more satisfy the humans' to the outside world, we operate the a and b channel of Lab color space to generate natural color perturbations. More importantly, our perturbations are generated in an incessant manner, unlike the perturbations generated on RGB color space which are sparse and introduce distortions even small variations existed. We illustrate the power of Lab color space in color representation in Section 4.3. In the following subsections, we introduce the framework of our proposed anti-forgery method to generate adversarial perceptual-aware perturbations to disrupt the DeepFake creation.

4.2 Adversarial Attack to GAN

The foundation of DeepFake creation is GAN which consists of two deep learning networks, the Generator G and Discriminator D. In the GAN training, the Generator G tries to generate samples indistinguishable to the real, while the Discriminator D learns to differentiate synthesized sample $G(z)$ from noises z and real samples. For example, a typical image-to-image translation network CycleGAN could be employed for identity swap[2]. Specifically, CycleGAN learns to build a mapping $G : x \rightarrow y$ and an inverse mapping $F : y \rightarrow x$ between two image domain x and y.

Similar to adversarial examples, the imperceptible perturbations used to disrupt DeepFakes are produced by introducing minimal distortion to preserve the natural effects of human eye vision. Let x be the input source image, θ is the imperceptible perturbation, $\tilde{x}$ is the crafted input with adversarial perturbations.

$$\tilde{x} = x + \theta \tag{1}$$

A generator G receives the two input x and $\tilde{x}$ and outputs $G(x)$ and $G(\tilde{x})$ further. Ideally, $G(x)$ and $G(\tilde{x})$ are totally different where $G(\tilde{x})$ exhibits noticeable artifacts to human eyes. More specifically, the objective function can be formulated as follows by maximize the distortion between $G(x)$ and $G(\tilde{x})$.

$$\max_{\theta} \mathcal{L}(G(x + \theta), r), \text{ subject to } \|\theta\|_{\infty} \leq \epsilon \tag{2}$$

where ϵ is the maximum magnitude of the perturbation, $\mathcal{L}$ is a distance function for measuring the similarity, r is a ground-truth, $r = G(x)$.

4.3 Lab Color Space

The Lab color space is designed to conform with human's sense of color which is perceptual uniform. A light channel L and two color channels a and b consist the Lab color space, where the L channel ranges from black (0) to white (100) representing the light, a channel ranges from green (-128) to red ($+127$), b channel ranges from blue (-128) to yellow ($+127$). Unlike RGB color space, we can change the lightness by simply modifying the value of L channel without involving any changes to the color channel a and b.

Due to the power of Lab color space in representing color space that is perceptual uniform and the wide value range in representing color, in this paper, our adversarial perturbations are generated by operating the Lab color space to add perturbation into the channel a and b.

4.4 Proposed Anti-forgery Method

Specifically, Algorithm 1 describes the details of our proposed method. Our goal for the generated perturbations is natural to human eyes without introducing any quality degradation and resistant enough to potential input transformation attacks. First, we convert the input image from the RGB color space to the Lab color space for adding perceptual uniform perturbations into the channel a and b. Then, the perturbations is updated by attacking the surrogate model M via optimization-based strategy adopted in C&W.

To improve the transferability in tackling multiple facial attributes, we select different facial attribute label c in each iteration. The objective function can be formulated as follows.

$$\text{minimize } \mathcal{L}(M(x_{adv}, c), o) \tag{3}$$

[2]https://github.com/shaoanlu/faceswap-GAN

Algorithm 1: Our proposed adversarial perceptual-aware perturbations.

Input : Input image x, Surrogate model M, Label $c \in C$, Iteration K, Objective o, Learning rate τ.

Output: Adversarial sample x_{adv} with perturbation.

1   initialization θ_a, θ_b
2   **for** $i \in \{1...K\}$ **do**
3      $l, a, b \leftarrow rgb2lab(x)$
4   ▷ **Add perturbations for both channel a and b.**
5      $a' \leftarrow a + \theta_a$
6      $b' \leftarrow b + \theta_b$
7   ▷ **Convert into RGB color space.**
8      $x_{adv} \leftarrow lab2rgb(l, a', b')$
9   ▷ **Update the perturbations of channel a and b.**
10     $\theta_a \leftarrow \theta_a - \tau \cdot \nabla_{\theta_a} \mathcal{L}(M(x_{adv}, c), o)$
11     $\theta_a \leftarrow clip(\theta_a, -\epsilon, \epsilon)$
12     $\theta_b \leftarrow \theta_b - \tau \cdot \nabla_{\theta_b} \mathcal{L}(M(x_{adv}, c), o)$
13     $\theta_b \leftarrow clip(\theta_b, -\epsilon, \epsilon)$
14   **return** x_{adv}

---

where $\mathcal{L}$ could be L_1 or L_2 norm, o could be 0, 1, or even Gaussian noises. When o is a regular translation image, the objective can be formulated as follows.

$$\text{minimize} - \mathcal{L}(M(x_{adv}, c), o) \qquad (4)$$

4.5 Comparison with Popular Adversarial Noises

A straightforward idea for attacking GAN to disrupt the DeepFake creation would be employing the gradient-based and optimization-based strategies to generate imperceptible perturbations. However, such restricted perturbations operating on the RGB color space suffer sparse issues which could be easily detected and corrupted by input transformations [Meng and Chen, 2017]. Specifically, the three channels in RGB color space have strong correlations, indicating that natural perturbations require the modification of these three channels simultaneously. Thus, the perturbations generated by operating the RGB color space are not robust against real world input transformations.

In summary, our perceptual-aware adversarial perturbations by operating the Lab color space have the following main strengths. 1) Robust to input reconstruction. Our perturbations operated on Lab color space which is perceptual uniform, as opposite to the prior studies operating on the RGB space which is sparse. The perceptual uniform perturbations operated on the Lab color space is more likely to share the same distribution as the unperturbed input. 2) Generic to a wide types of fogery. There is a wide range of color values for operating on the Lab color space. We will always explore the best perturbations for disruption effectively.

5 Experiments

The experimental setup and more experimental results on the performance across diverse GANs, exploring other color spaces, ablation studies, visualizations, and the description of broader impact *etc.* , see the Appendix here[3].

---

[3]http://wangrun.github.io/supp/Anti_forgery_IJCAI2022.pdf

5.1 Effectiveness Evaluation

We report the experimental results of effectiveness evaluation on all three types of DeepFakes involving the manipulation of source images. Since a large portion of GANs in image synthesis are fine-grained attribute editing, we employ three different GANs (*e.g.*, StarGAN, AttGAN, and Fader Network) which could represent the SOTA performance of GANs in attribute editing. Experimental results show that our proposed method by operating the Lab color space could disrupt the SOTA GANs for DeepFake creation with competitive performance. The visualization of our proposed method in disrupting DeepFaked images refer to the supplemental material.

Table 1 summarizes the effectiveness evaluation and comparison with two baselines. For the three attributes editing GANs, our method achieved competitive results with the two baselines measured by four different metrics and significantly outperforms them in AttGAN. In the metrics of ASR by giving a distortion restriction, our method outperforms the two baselines in the three GANs for attributes editing.

To the best of our knowledge, this is the first work to present a comprehensive evaluation on the three types of DeepFakes, including attribute editing, identity swap, and face reenactment. Experimental results in Table 1 demonstrated the effectiveness of our proposed method and illustrated that our proposed method achieved competitive results in comparing with the two baselines.

5.2 Robustness Analysis

In this section, we report the evaluation results of our method against common input transformations, including JPEG compression, Gaussian Blur, reconstruction, *etc.* . Specifically, in the real world scenarios, the input with added perturbations would be spread in the social media platform and suffers various real degradations (*e.g.*, compression, blur, *etc.*). A recent study MagDR revealed that the prior DeepFake disruption studies failed in tackling input reconstruction where the added imperceptibly could be easily removed or corrupted by simple reconstruction. Thus, we aim to answer one question that *whether our proposed method is robust against the common input transformations, especially the input reconstruction studied in MagDR.*

Performance on common input transformations. Table 2 summarizes the performance of our method in tackling various input transformations. Experimental results show that our method shows promising performance in the adversary settings and demonstrated its robustness in surviving various input transformations. Specifically, our method significantly outperforms the two competitive baselines in all four adversary settings. In Table 2, the second row *Defense* means a clean setting without any input transformations; the third row *JPEG compression* indicates that the input is compressed before leveraged for creating DeepFakes; the fourth row *JPEG Compression ($\epsilon = 0.1$)* denotes that the added perturbations is restricted to a certain range larger than the default $\epsilon = 0.05$; the fifth row *Gaussian Blur* represents that Gaussian blur is employed to input; the last row *Blur (Data augmentation)* explores whether the resistant against blur can be improved via data augmentation. Experimental results show that our

DeepFake Model	PGD				C&W				Our method			
	L_2 ↑	PSNR ↓	SSIM↓	ASR↑	L_2 ↑	PSNR↓	SSIM↓	ASR↑	L_2 ↑	PSNR↓	SSIM↓	ASR↑
StarGAN [Choi et al., 2018]	1.192	5.386	0.297	**100.0**	**1.409**	**4.781**	0.285	**100.0**	1.302	5.292	**0.251**	100.0
AttGAN [He et al., 2019]	0.091	16.885	0.701	80.50	0.095	16.419	0.679	84.10	**0.103**	**15.975**	**0.634**	**85.30**
Fader Network [Lample et al., 2017]	0.195	13.663	0.374	89.70	**0.242**	**10.932**	0.365	91.40	0.229	11.047	**0.320**	**93.10**
Identity swap	0.035	15.976	0.468	5.40	**0.031**	16.710	**0.513**	3.90	0.029	**17.153**	0.530	4.20
Face reenactment [Tripathy et al., 2020]	-	18.336	0.688	-	**-**	16.621	**0.627**	**-**	-	**17.824**	0.640	-

Table 1: Performance of our method in disrupting the three DeepFake types with a comparison with two competitive baselines. The manipulated attributes for the four GANs are Black hair, Blond hair, Brown hair, male, and young. For the adopted four metrics, the large value of L_2 and ASR indicates the large distortion is introduced while the small value of PSNR and SSIM means the large corruption is introduced. For the face reenactment, the ASR is applicable since the frames are extracted from the video, thus we leave it blank.

Defense	PGD				C&W				Our method			
	L_2 ↑	PSNR ↓	SSIM↓	ASR↑	L_2 ↑	PSNR↓	SSIM↓	ASR↑	L_2 ↑	PSNR↓	SSIM↓	ASR↑
No defense	1.192	5.386	0.297	**100.0**	1.409	**4.781**	0.285	**100.0**	1.302	5.292	**0.251**	100.0
JPEG compression	0.038	17.236	0.682	41.30	0.043	16.774	0.611	45.20	**0.058**	12.722	**0.433**	**54.10**
JPEG Compression ($\epsilon = 0.1$)	0.568	9.430	0.484	83.50	0.501	11.355	0.552	81.20	**0.722**	7.459	**0.370**	**90.70**
Gaussian Blur ($\delta = 3$)	0.007	28.086	0.944	0.0	0.004	29.335	0.913	0.0	**0.012**	20.252	**0.865**	**5.80**
Blur (Data Augmentation)	0.096	26.871	0.806	77.90	0.114	21.335	0.713	81.30	**0.302**	18.252	**0.631**	**88.50**

Table 2: Performance of our method on StarGAN in adversary settings with a comparison with two competitive baselines. The inputs are transformed by blurring, input compression.

Metric	PGD	Our Method	PGD (MagDR)	Our Method (MagDR)
SSIM (I)	0.914	**0.953**	0.914	**0.906**
PSNR (I)	35.313	**35.618**	35.356	**30.699**
MSE (I)	29.836	**27.889**	29.546	**88.320**
SSIM (O)	0.604	**0.522**	0.892	**0.819**
PSNR (O)	15.029	**13.107**	30.838	**20.828**
MSE (O)	4117.143	**6795.289**	85.429	**1083.934**

Table 3: Performance of our method in bypassing input reconstruction. SSIM (I) indicates the input with added perturbations, while SSIM (O) means the output of SSIM (I) by employing DeepFake manipulations. PGD (MagDR) and our method (MagDR) represent the sample is reconstructed by prior study MagDR.

method outperforms the two baselines measured by four different metrics on four input transformations.

Performance on MagDR [Chen et al., 2021]. Beyond the aforementioned popular input transformations, a recent study MagDR revealed that existing DeepFake disruption techniques are not robust to input reconstruction. Specifically, the added imperceptible perturbations could be corrupted or even removed by employing simple reconstruction and further output a high-quality DeepFake without noticeable artifacts. In our experiments, we also explore whether our method could resist the threat of input reconstruction.

Specifically, Table 3 summarizes the experimental results and compares them with one baseline. We reproduce MagDR and ensure the implementation details are correct by checking with the authors of MagDR. For the input with added perturbations, our method exhibits fewer perturbations than the baseline but leaves more perturbations in the input after employing MagDR which is our desired result as it would be lead to corrupted DeepFakes. For the output by employing MagDR, the value of the three metrics indicates the effectiveness in creating DeepFakes with noticeable artifacts. We can find that the MagDR failed in corrupting our added perceptual-aware perturbations over the three metrics by operating the Lab color space in comparison with the baseline.

In summary, a comprehensive robustness evaluation against both the common input transformations and input reconstruction demonstrated that our proposed method shows strong capabilities in resisting common input transformations and outperforms the two baselines measured by four different metrics in tackling the three common input transformations. Our proposed method shows its potential to be deployed in the real-world scenarios.

5.3 Discussion

Here, we discuss the limitations of our proposed method by operating on the Lab color space to generate perceptual-aware perturbations for DeepFake disruption and suggest provisional countermeasures against them. Our generated perturbations are added into the whole input which suffers the potential threat that the adversary crops the background to destroy the perturbations directly. In our future work, we explore more robust technique to evade such attack by setting masks to enforce the perturbations are added into the internal region of face.

6 Conclusion and Future Research Direction

We propose the first adversarial perceptual-aware perturbations operating on the Lab color space for anti-fogery and performed an extensive evaluation of our method on three types of DeepFakes. Experimental results demonstrate its effectiveness in disrupting DeepFakes by exposing noticeable artifacts to human eyes while preserving the visually naturalness of the source image. More importantly, our method exhibits its robustness against input reconstruction which significantly outperforms prior studies. In addition, the created DeepFakes by adding our perceptual-aware perturbations provide strong signals for DeepFake detection. Our findings of perceptual-aware adversarial perturbations present a new insight for defending DeepFakes proactively by investigating more natural and robust adversarial perturbations like facial background [Xiao et al., 2020], etc. . Thus, in our future work, we will explore more physical scenes to serve as guards for protecting images being DeepFaked.

Acknowledgments

References

[Carlini and Farid, 2020] Nicholas Carlini and Hany Farid. Evading deepfake-image detectors with white-and black-box attacks. In *CVPR Workshops*, pages 658–659, 2020.

[Chen *et al.*, 2021] Zhikai Chen, Lingxi Xie, Shanmin Pang, Yong He, and Bo Zhang. Magdr: Mask-guided detection and reconstruction for defending deepfakes. In *CVPR*, pages 9014–9023, 2021.

[Choi *et al.*, 2018] Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *CVPR*, pages 8789–8797, 2018.

[He *et al.*, 2019] Zhenliang He, Wangmeng Zuo, Meina Kan, Shiguang Shan, and Xilin Chen. Attgan: Facial attribute editing by only changing what you want. *TIP*, 28(11):5464–5478, 2019.

[Hu *et al.*, 2021] Shu Hu, Yuezun Li, and Siwei Lyu. Exposing gan-generated faces using inconsistent corneal specular highlights. In *ICASSP*, pages 2500–2504. IEEE, 2021.

[Huang *et al.*, 2020] Yihao Huang, Felix Juefei-Xu, Run Wang, Qing Guo, Lei Ma, Xiaofei Xie, Jianwen Li, Weikai Miao, Yang Liu, and Geguang Pu. Fakepolisher: Making deepfakes more detection-evasive by shallow reconstruction. In *ACM Multimedia*, pages 1217–1226, 2020.

[Huang *et al.*, 2021] Qidong Huang, Jie Zhang, Wenbo Zhou, Weiming Zhang, and Nenghai Yu. Initiative defense against facial manipulation. In *AAAI*, volume 35, pages 1619–1627, 2021.

[Juefei-Xu *et al.*, 2021] Felix Juefei-Xu, Run Wang, Yihao Huang, Qing Guo, Lei Ma, and Yang Liu. Countering malicious deepfakes: Survey, battleground, and horizon. *arXiv preprint arXiv:2103.00218*, 2021.

[Lample *et al.*, 2017] Guillaume Lample, Neil Zeghidour, Nicolas Usunier, Antoine Bordes, Ludovic Denoyer, and Marc'Aurelio Ranzato. Fader networks: Manipulating images by sliding attributes. *arXiv preprint arXiv:1706.00409*, 2017.

[Liu *et al.*, 2019] Ming Liu, Yukang Ding, Min Xia, Xiao Liu, Errui Ding, Wangmeng Zuo, and Shilei Wen. Stgan: A unified selective transfer network for arbitrary image attribute editing. In *CVPR*, pages 3673–3682, 2019.

[Meng and Chen, 2017] Dongyu Meng and Hao Chen. Magnet: a two-pronged defense against adversarial examples. In *CCS*, pages 135–147, 2017.

[Perov *et al.*, 2020] Ivan Perov, Daiheng Gao, Nikolay Chervoniy, Kunlin Liu, Sugasa Marangonda, Chris Umé, Mr Dpfks, Carl Shift Facenheim, Luis RP, Jian Jiang, et al. Deepfacelab: A simple, flexible and extensible face swapping framework. *arXiv preprint arXiv:2005.05535*, 2020.

[Qian *et al.*, 2020] Yuyang Qian, Guojun Yin, Lu Sheng, Zixuan Chen, and Jing Shao. Thinking in frequency: Face forgery detection by mining frequency-aware clues. In *ECCV*, pages 86–103. Springer, 2020.

[Ruiz *et al.*, 2020] Nataniel Ruiz, Sarah Adel Bargal, and Stan Sclaroff. Disrupting deepfakes: Adversarial attacks against conditional image translation networks and facial manipulation systems. In *ECCV*, pages 236–251. Springer, 2020.

[Segalis, 2020] Eran Segalis. Disrupting deepfakes with an adversarial attack that survives training. *arXiv e-prints*, pages arXiv–2006, 2020.

[Thies *et al.*, 2016] Justus Thies, Michael Zollhofer, Marc Stamminger, Christian Theobalt, and Matthias Nießner. Face2face: Real-time face capture and reenactment of rgb videos. In *CVPR*, pages 2387–2395, 2016.

[Tripathy *et al.*, 2020] Soumya Tripathy, Juho Kannala, and Esa Rahtu. Icface: Interpretable and controllable face reenactment using gans. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3385–3394, 2020.

[Wang *et al.*, 2020] Sheng-Yu Wang, Oliver Wang, Richard Zhang, Andrew Owens, and Alexei A Efros. Cnn-generated images are surprisingly easy to spot... for now. In *CVPR*, pages 8695–8704, 2020.

[Wang *et al.*, 2021a] Run Wang, Felix Juefei-Xu, Meng Luo, Yang Liu, and Lina Wang. Faketagger: Robust safeguards against deepfake dissemination via provenance tracking. In *ACM Multimedia*, pages 3546–3555, 2021.

[Wang *et al.*, 2021b] Run Wang, Felix Juefei-Xu, Lei Ma, Xiaofei Xie, Yihao Huang, Jian Wang, and Yang Liu. Fakespotter: a simple yet robust baseline for spotting ai-synthesized fake faces. In *IJCAI*, pages 3444–3451, 2021.

[Xiao *et al.*, 2020] Kai Yuanqing Xiao, Logan Engstrom, Andrew Ilyas, and Aleksander Madry. Noise or signal: The role of image backgrounds in object recognition. In *ICLR*, 2020.

[Yang *et al.*, 2020] Chaofei Yang, Lei Ding, Yiran Chen, and Hai Li. Defending against gan-based deepfake attacks via transformation-aware adversarial faces. *arXiv preprint arXiv:2006.07421*, 2020.

[Yeh *et al.*, 2020] Chin-Yuan Yeh, Hsi-Wen Chen, Shang-Lun Tsai, and Sheng-De Wang. Disrupting image-translation-based deepfake algorithms with adversarial attacks. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision Workshops*, pages 53–62, 2020.