Towards Retentive Proactive Defense Against DeepFakes

Tao Jiang[1], Hongyi Yu[2], Wenjuan Meng[3], and Peihan Qi[4(✉)]

[1] School of Cyber Engineering, Xidian University, Xi'an, China
taojiang@xidian.edu.cn
[2] Guangzhou Institute of Technology, Xidian University, Xi'an, China
hongyiyu@stu.xidian.edu.cn
[3] College of Information Engineering, Northwest A&F University, Yangling, Shaanxi, China
wjmeng@nwsuaf.edu.cn
[4] State Key Laboratory of Integrated Service Networks, Xidian University, Xi'an, China
phqi@xidian.edu.cn

Abstract. In recent years, with the development of artificial intelligence, many facial manipulation methods based on deep neural networks have been developed, known as DeepFakes. Unfortunately, DeepFakes are always maliciously used, and if the spread of DeepFakes cannot be controlled in a timely manner, it will pose a certain threat to both society and individuals. Researchers have studied the detection of DeepFakes, but this type of detection belongs to post-evidence collection and still has a certain degree of negative impact. Therefore, we propose a retentive and proactive defense method to protect DeepFakes before malicious operations. The main idea is to train a perturbation generator end-to-end, and introduce the perturbation generated by the perturbation generator into the image to make it adversarial and immune to DeepFakes. White-box experiments on a typical DeepFake manipulation method (facial attribute editing) demonstrate the effectiveness of our proposed method, and a comparison with an adversarial attack PGD proves the superiority of our method in terms of similarity and inference efficiency.

Keywords: DeepFake · Retentive · Proactive defense · Adversarial attack · Perturbation

1 Introduction

With the booming field of artificial intelligence, advanced image and video synthesis techniques have started to emerge. Various generative adversarial networks [1], that can generate more realistic images and videos, have already produced a great impact on artificial intelligence and related fields. DeepFake [2–4] is a technique for synthesizing images and videos, which can edit the facial attributes of

J. Liu et al. (Eds.): TridentCom 2023, LNICST 523, pp. 139–153, 2024.
https://doi.org/10.1007/978-3-031-51399-2_8

the target, as well as can modify the expression of the target samples. DeepFakes can also replace the face of the object with the face of the target as a way to generate some fake but realistic images and videos. Some people used DeepFakes to generate pornographic videos of celebrities for posting on the Internet, which has made DeepFakes notorious on the Internet. Although DeepFakes have an important role in many fields, effective defenses should be developed to prevent the misuse of DeepFakes.

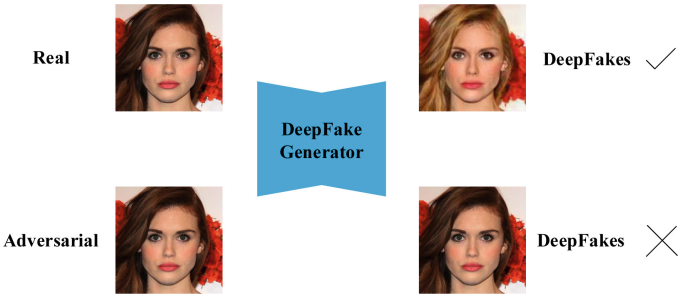

Fig. 1. An illustration of retentive proactive defense against a DeepFake generator. After applying an imperceptible perturbation on the image, the output of StarGAN [2] is similar to the input, which assure the defense method is immune to the function of the DeepFake.

Researchers' defensive countermeasures against DeepFakes are mainly divided into passive and proactive defenses. Passive defense focuses on the detection of DeepFakes [5–7]. This type of technique far exceeds other methods in the overall research on the defense of DeepFakes, which can detect the generated false images and videos. Specifically, DeepFake detectors classify a given image or video as true or false based on the characteristics of the face in it. Although all existing detectors have high accuracy, they can only mitigate the negative impact of false information already spread on the Internet due to being only a passive defense with post-event forensics, and cannot completely eliminate it. Moreover, the detectors usually use deep neural networks, which are very vulnerable in the face of adversarial attacks [8–10]. Therefore, some researchers have started to study how to proactively defend against DeepFakes [11–13], i.e., to proactively defend against DeepFakes before they produce a series of negative effects. They focus on adding invisible perturbations to the images, so that the output of the perturbed images input into the DeepFake model has obvious distortion, so as to achieve the purpose of proactive defense. Although these studies have been effective in disrupting DeepFakes, the added adversarial perturbations are easily detected and eliminated by noise detectors [14,15]. Moreover, most of the previous studies involve disruptions that distort the output of DeepFakes. Such disruptions hit DeepFakes to some extent, but sometimes people on the Internet do not care if things are true or false, which still has some negative

impact. Therefore, it is still necessary to investigate effective methods that still have sufficient defense capability while ensuring no degradation in image quality.

In this paper, as illustrated in Fig. 1, we delve into the disruption of DeepFakes functionality and make our endeavor to assure the outputs of DeepFakes as similar as their inputs. More precisely, we train a perturbation generator through an end-to-end pipeline to obtain an invisible perturbation, then add the perturbation to the image to obtain an adversarial image similar to the original image that can disrupt the function of DeepFakes. Finally, in the training and inference phases, the adversarial image with the perturbation added has no change visible to the naked eye after the DeepFakes output, achieving the purpose of breaking the function of DeepFakes. In summary, the adversarial image with added perturbations must satisfy two objectives: (i) visually consistent with the original image (ii) no visual change between the input and the output of DeepFakes. To evaluate the superiority of our approach, we conducted experiments mainly on the facial attribute editing StarGAN and compared it with the powerful adversarial attack PGD. The experimental results show that our proposed method can better enable the addition of perturbed images with the ability to corrupt DeepFakes functionality.

Our main contributions are summarized as follows:

1. We propose a retentive framework for proactive defense against DeepFakes that can be adapted to different DeepFakes.
2. In this framework we train a perturbation generator that can destroy DeepFakes, the perturbation generator can quickly and efficiently image the corresponding perturbation. The original image with the addition of this invisible perturbation has immunity to DeepFakes by making the DeepFakes output preserve the visual effect of their inputs.
3. We demonstrate the superiority of our approach in terms of similarity and inference efficiency by comparing it with adversarial attack PGD under different evaluation metrics and different datasets.

2 Related Work

2.1 DeepFake Generation

As generative adversarial networks have made significant progress in recent years, DeepFakes have evolved from crude to exquisite. DeepFakes that are indistinguishable from the human eye pose a huge threat to people's security and privacy. Currently, there are four main types of DeepFakes, namely the entire synthesis, facial attribute editing, face swapping, and facial expression swapping. Since the entire synthesis does not involve a specific person, we mainly consider the other three types of threats.

The entire synthesis is to generate people that do not exist in the world, such as PGGAN [16], and StyleGAN [3]. Facial attribute editing is the use of GANs such as StarGAN [2] for a particular attribute of the face, e.g., hair color, or age. Face swapping is the most notorious method that replaces the source image face

with the target face, FaceSwap and DeepFaceLab [17] are two commonly used tools. Facial expression swapping can be done by replacing the source image facial expression with the target facial expression using tools like Face2Face [18].

2.2 DeepFake Defense

Passive Defense. Initially researchers only considered distinguishing the fake works generated by DeepFakes, so they invented a series of DeepFake detectors, all of which have good accuracy rates. But DeepFake detection is a passive defense of post-event forensics, which can only mitigate the negative effects caused by DeepFakes. Current research is mainly based on DeepFakes detection of artificial artifacts in the spatial [5] and frequency domains [6], and there are also some other methods such as biosignals [7]. These methods commonly use deep neural networks, and some researchers have identified the threat of adversarial perturbations to DeepFakes detectors. There are several other issues [19] that researchers need to address.

Proactive Defense. DeepFakes passive defense can only stop things after they happen, and it is not effective to mitigate the impact of DeepFakes. To solve this problem, various proactive defense methods [11–13,20–26] have been proposed to defend against DeepFakes.

Some consider inserting predefined marks into a synthesized face and then using these marks to determine whether the image or video has been manipulated by DeepFakes. Wang et al. [20] inserted label information into the protected image. Images subject to DeepFakes manipulation can also be obtained with the label information inserted at the beginning. Sun et al. [21] designed two types of traces, sustainable traces and erasable traces. A model trained using images with both types of traces added generates face images with only sustainable traces. In this way, it is possible to determine whether an image or video has been manipulated by DeepFakes.

Others considered adding adversarial perturbation to the image, and the image with added perturbation is not visually different from the original image. Finally, this image with added invisible perturbation can destroy the function of DeepFakes, and this method can effectively mitigate the impact of DeepFakes. Ruiz et al. [11] proposed an adversarial attack in the gray box case, and then Huang et al. [12] proposed a method that can achieve the goal by alternating training strategies and using some task-specific strategies to enhance the defense performance. Wang et al. [22] considered that distorted images need to be judged not only by the naked eye, but also by DeepFake detectors. Huang et al. [23] consider that the generated perturbations are usually specific to a particular image and a particular generative model, and thus propose a two-level perturbation fusion strategy as a way to generate cross-model generic adversarial watermarks. Wang et al. [24] consider that perturbations are generally not robust, so a method to generate perceptual-aware perturbations incessantly was proposed to improve the robustness of perturbations.

However, the goal of the attack that most researchers have considered is to distort the output of DeepFakes. This does allow people to tell if an image is real or fake, but people on the Internet these days don't really care if things are real or fake a lot of the time. Unless the distortion is so severe that you can't see a face, it can have a negative impact. Therefore, we tend to attack the goal of making the output of DeepFakes visually unchanged from the input. Yeh et al. [25] proposed an invalid attack using adversarial attack PGD to minimize the distance between the adversarial output and the original input. They [26] later proposed to solve this problem in the black box case. He et al. [13] proposed a method to find neighbors in the latent space that are similar to the original image but can reach the disabled DeepFakes target. This method slightly alters the appearance of the face image although it has better visual quality less likely to be detected by noise detectors. In this paper, we present an end-to-end approach to training a perturbation generator. Images can achieve the goal of disabling DeepFakes after adding the perturbations generated by the perturbation generator.

3 Method

In this section, we first describe the adversarial attack on DeepFakes. Then we introduce our proactive defense framework and its loss function and optimization algorithms.

3.1 Adversarial Attacks Against DeepFake

Proactive defense against DeepFakes by adding invisible perturbations to the original image is a more important direction for research than using those DeepFake detectors that have been extensively studied for passive defense. An adversarial attack against DeepFakes can disrupt the functionality of DeepFakes. There are two ways of this destruction, which are distortion attack and invalid attack. The distortion attack is to make the output of DeepFakes distorted, so that visually it can be judged that it is not a real image. The invalid attack is one that causes DeepFakes to lose the ability to manipulate the image, making it visually impossible to tell the gap between the input and the output of the DeepFakes. Formally, we denote x as the original image and x_{adv} as the adversarial image, where $x_{adv} = x + \eta$ and η is a visually invisible perturbation with a common norm constraint. The corresponding outputs $G(x)$ and $G(x_{adv})$ can be obtained by feeding the original image x and the adversarial image x_{adv} into the DeepFake generator $G(\cdot)$. We let t be the target of the attack and the objective function can be written as

$$\mathcal{L}(G(x+\eta), t), \quad s.t. \quad ||\eta||_2 < \epsilon, \tag{1}$$

where $\mathcal{L}$ is a distance function normally using the L_0, L_2 or L_∞ norms. If t is set to the original output $G(x)$, maximizing this function is a distortion attack, and we can obtain adversarial images x_{adv} that distort the output of the DeepFakes. If t is set to the original image x, minimizing this function is an invalid attack, and

we can obtain adversarial images x_{adv} that invalidates DeepFakes. Minimizing this function can also be considered as target image generation if t is set to the desired target.

The optimal perturbation η of the original image can be efficiently optimized by an adversarial attack on Eq. 1., e.g., Iterative Fast Gradient Sign Method (IFGSM) [8] or Projected Gradient Descent (PGD) [9] . Though the optimal η can be effectively solved through the iterations of IFGSM or PGD, it could be time-consuming to deal with the large-scale image dataset. For each image, we have to optimize it individually and iteratively to obtain the corresponding best perturbation η, which is a waste of resources. Training a perturbation generator that is effective for DeepFakes not only saves time, but may even yield better performance.

3.2 Perturbation Generator

As shown in Fig. 2, the model pipeline of this method consists of a perturbation generator and a DeepFake generator. Next, we will introduce them one by one and provide an overall optimization framework. Compared to generating the corresponding perturbation for each image separately, we tend to learn a perturbation generator PG to generate perturbation $P(x)$ for image x. We add perturbation $P(x)$ to image x to obtain adversarial image x_{adv}. Therefore, the generated image of the adversarial image x_{adv} under the given DeepFake generator G can be written as $G(x_{adv})$. Therefore, the objective function of disrupting DeepFake can be rewritten as

$$x_{adv} = x + P(x) \tag{2}$$

$$\min_{P} \mathcal{L}(x_{adv}, G(x_{adv})), s.t. \quad ||P(x)||_2 < \epsilon, \tag{3}$$

where ϵ is our chosen range of the perturbation.

Magnitude Loss. Due to the use of images with added invisible perturbations to replace the original image to combat DeepFakes, the original image and the image with added perturbations should be visually as similar as possible. Therefore, we propose the first loss to constrain the size of the perturbation generated by the perturbation generator PG. $\mathcal{L}_{mag}$ is composed of two loss functions, $\mathcal{L}_{pix}$ is the pix-level loss with L_2 norm of x and x_{adv}, and $\mathcal{L}_{feat}$ is the perceptual loss of both,

$$\mathcal{L}_{pix} = \|x - x_{adv}\|_2 \tag{4}$$

$$\mathcal{L}_{feat} = \|F(x) - F(x_{adv})\|_2 \tag{5}$$

$$\mathcal{L}_{mag} = \mathcal{L}_{pix} + \lambda_{feat}\mathcal{L}_{feat} \tag{6}$$

where λ_{feat} is hyper-parameters to balance different loss items. $F(\cdot)$ represents a feature extraction model for acquiring perceptual loss.

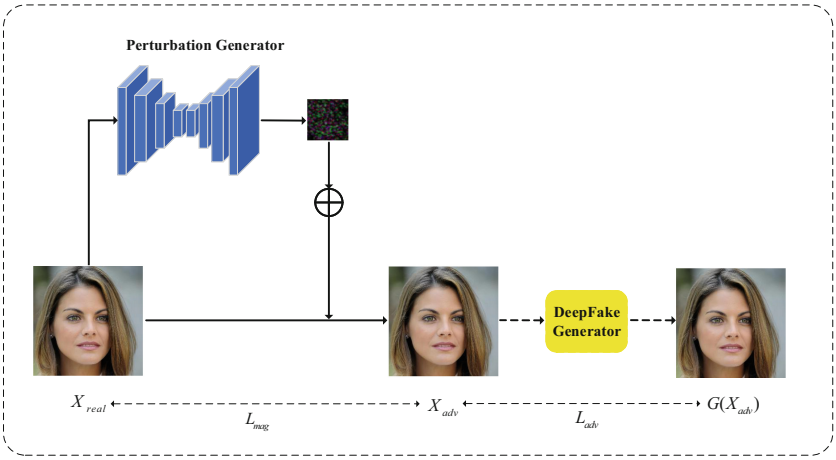

Fig. 2. The training phase of the overall pipeline of the perturbation generator that combats DeepFakes. The source image x is fed into the perturbation generator PG to obtain a perturbation $P(x)$ that is invisible to the naked eye. This perturbation $P(x)$ is then added to the corresponding source image x to obtain the adversarial image x_{adv}. Finally, this adversarial image x_{adv} is input to the DeepFake generator G to obtain the corresponding output $G(x_{adv})$. The corresponding perturbation generator PG is obtained by training under the constraints of the loss functions.

Adversarial Loss. The purpose of training the Perturbation Generator PG is to destroy the function of the DeepFake generative model, i.e., to make the input and output of the DeepFakes similar. Therefore, we introduce the adversarial loss to constrain the l_2 distance between the adversarial image x_{adv} and its output,

$$\mathcal{L}_{adv} = \|x_{adv} - G(x_{adv})\|_2 \tag{7}$$

By incorporating the above loss functions, we thus achieve the resulting objective function,

$$\mathcal{L} = \lambda_{mag}\mathcal{L}_{mag} + \lambda_{adv}\mathcal{L}_{adv} \tag{8}$$

where λ_{mag} and λ_{adv} are hyper parameters to balance different loss items. The DeepFake models will be selected from the pre-trained SOTA models. Their parameters will not be changed during the training, which means that the whole end-to-end training will only update the parameters of the perturbation generator PG. By optimizing Eq. 8, we can learn the appropriate perturbation generator parameters from the training set. In the inference phase, we input the source image into the perturbation generator PG to obtain the corresponding optimal perturbation, and add this invisible perturbation to the source image x to obtain the corresponding adversarial image x_{adv}, which can destroy functionality of the DeepFakes, i.e., DeepFakes are unable to manipulate x_{adv}.

4 Experiments

In this paper, we have experimented with attribute editing in DeepFakes, a classic facial manipulation task, using one of the representative works, StarGAN [2]. To demonstrate the effectiveness as well as the superiority of the method, we first demonstrate that this proactive defense framework can effectively disrupt the functionality of DeepFakes. At the same time we ensure that the processed image is visually indistinguishable from the source image. Then, we compare it with the adversarial attack PGD in a white-box environment, and we can clearly see the superiority of our method from the experiment.

4.1 Architectures and Datasets

Similar perturbation generators have been covered in previous work [12,22,27]. We refer to the network architectures used in previous work. The network architecture of the perturbation generator PG is chosen to be U-Net [28], which is a classical network architecture that consists of an encoder that extracts the abstract features of an image and a decoder that fuses the various features. Its U-shaped network structure densely fuses the shallow features with the deeper features, and our experiments use the UNet-128 network.

We use the pre-trained StarGAN model for attribute editing operations. The dataset uses face images from CelebA, which has 202,599 face images with a resolution of 178×218. We first crop the center of the image to 178×178 to increase the proportion of the face in the image, and then resize it to 128×128. Finally, 200,600 of these images were used in the training phase and the remaining 1,999 images were used in the test phase. In the training phase, the batchsize is set to 16, and the learning rate of the perturbation generator is set to 0.0001. The attribute domain is set to five attributes ('Black Hair', 'Blond Hair', 'Brown Hair', 'Male', 'Young'). The network is updated for 200,000 iterations. We use PGD for adversarial attacks, simultaneously attacking five attribute domains of one image at a time, with a step size of 0.0001 and an iteration count of 100. In the inference phase, we additionally selected the public face datasets LFW [29] and FFHQ [30] for evaluation. We select 1999 images from each of these two datasets. Since the image size of LFW is 250×250, we first crop the center of the image to 200×200 to increase the proportion of the face in the image, and then resize it to 128×128.

4.2 Evaluation Metrics

In order to comprehensively evaluate the effectiveness of the attack, we have employed the L_2 norm distance, Structural Similarity Index (SSIM), Peak Signal-to-Noise Ratio (PSNR), perceptual loss [31], and Learned Perceptual Image Patch Similarity (LPIPS) [32] to evaluate the similarity between the images with the addition of the invisible perturbation and their outputs under the generative model, respectively. The similarity between the images and their outputs

of the DeepFakes is evaluated. The network used for feature extraction by perceptual loss is VGG16 and the network used for feature extraction by LPIPS is AlexNet. We propose the Defense Success Rate (DSR) to better evaluate the effectiveness of the attack. If the pixel-level loss $L_2 \leq 0.05$ between the input and the output of the DeepFakes, we can consider that it successfully destroys the functionality of the DeepFakes. So we define the DSR as the percentage of adversarial images that successfully destroy the functionality of the DeepFakes in the total test images. In order to show the effectiveness of the defense more intuitively, we use Local Binary Pattern (LBP) to describe the local features of an image. LBP features are widely used in face recognition.

4.3 Attack Performance Evaluation

In order to demonstrate the effectiveness of the proposed proactive defense, we conducted experiments in a white-box manner, i.e., we know the network architecture of the DeepFakes, the domain information, and various information about the internals.

For the facial attribute editing task, our approach ensures better disruption of DeepFakes while the image with added invisible perturbations has better similarity to the original image. As shown in Fig. 3, we can visualize the effect of our method on the five attribute domains that StarGAN chooses to manipulate. The first column is the input of the DeepFake, and the following five columns are StarGAN's manipulation of the five attribute domains ('Black Hair', 'Blond Hair', 'Brown Hair', 'Male', 'Young') of the input image. The first line is the original result with no defense, and the second line is the result after using our method.

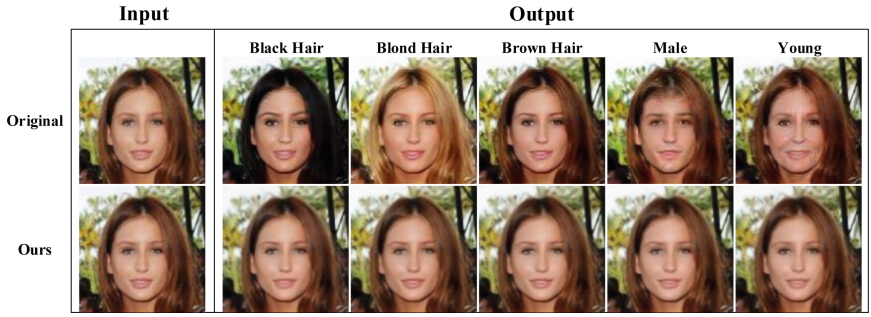

Fig. 3. Visual examples of defending against StarGAN [2].

4.4 Comparison with Other Methods

Similarity. We compared our method with adversarial attack PGD at $\epsilon = 0.05$ under StarGAN on different datasets. In Table 1, it can be clearly seen that the

adversarial image x_{adv} obtained by our method is more similar to the original image than PGD. Thus it is visually more difficult to detect the existence of the perturbation.

Table 1. Similarity comparison between the original image and the image with perturbation.($\epsilon = 0.05$)

DataSets	Methods	$L_2 \downarrow$	PSNR↑	SSIM↑	perceptual↓	LPIPS↓
CelebA	PGD	0.0008	36.9832	0.9360	239.0099	0.0115
	Ours	**0.0002**	**42.6844**	**0.9887**	**11.6263**	**0.0012**
LFW	PGD	0.0008	36.9324	0.9381	240.8695	0.0130
	Ours	**0.0002**	**43.0926**	**0.9909**	**11.7893**	**0.0015**
FFHQ	PGD	0.0008	36.9449	0.9473	225.5404	0.0070
	Ours	**0.0004**	**40.6340**	**0.9876**	**18.1097**	**0.0012**

In Table 2, we compare the similarity between the input and output of StarGAN. We find that both our method and PGD can effectively combat DeepFakes. However, our method performs better in terms of similarity, which indicates that our method has better defense.

Table 2. Similarity comparison between the Input and the Output of the StarGAN. ($\epsilon = 0.05$)

Datasets	Methods	$L_2 \downarrow$	PSNR↑	SSIM↑	perceptual↓	LPIPS↓	ASR↑
CelebA	No defense	0.0435	21.5832	0.8195	2994.8579	0.0869	-
	PGD	0.0039	30.2694	0.8654	2014.5713	0.0581	100%
	Ours	**0.0018**	**34.0527**	**0.9594**	**941.7181**	**0.0306**	100%
LFW	No defense	0.0414	21.2920	0.8106	3298.6406	0.0998	-
	PGD	0.0040	30.1451	0.8673	2068.6641	0.0578	100%
	Ours	**0.0018**	**33.8998**	**0.9622**	**908.7238**	**0.0274**	100%
FFHQ	No defense	0.0477	20.6184	0.7889	4039.8401	0.0988	-
	PGD	0.0059	28.5643	0.8556	2565.0344	0.0747	100%
	Ours	**0.0038**	**30.8490**	**0.9386**	**1670.6807**	**0.0542**	100%

In Table 3, we compare the similarity between the original image and the adversarial output of StarGAN. We can see that our method has higher similarity than PGD. From this we can know that our method outperforms PGD in terms of similarity regardless of whether the image with the added invisible perturbation is attacked by DeepFake or not.

Table 3. Similarity comparison between the original image and the output of image with perturbation on StarGAN. ($\epsilon = 0.05$)

Datasets	Methods	$L_2 \downarrow$	PSNR↑	SSIM↑	perceptual↓	LPIPS↓
CelebA	PGD	0.0032	31.1604	0.9052	1977.8158	0.0481
	Ours	**0.0024**	**32.8416**	**0.9446**	**1016.6274**	**0.0355**
LFW	PGD	0.0033	31.0122	0.9064	1999.1094	0.0486
	Ours	**0.0023**	**32.7748**	**0.9514**	**970.7018**	**0.0328**
FFHQ	PGD	0.0052	29.1619	0.8903	2593.4036	0.0665
	Ours	**0.0050**	**29.5974**	**0.9178**	**1821.7161**	**0.0615**

As shown in Fig. 4, we can see the superiority of our method more intuitively at the pixel-level and at the LBP-level. At the pixel-level we can see that both our method and PGD are effective enough. However, at the LBP-level we can see that the gap for our method compared to PGD is smaller.

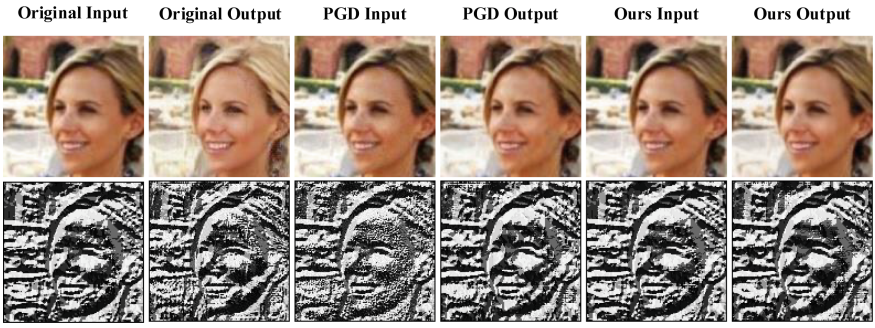

Fig. 4. Visual comparison of the adversarial attack PGD with our method in the pixel-level (top) and the LBP-level (bottom) on the defense of StarGAN [2].

Inference Efficiency. We also make a comparison in inference efficiency, where we randomly select 100 images for corresponding perturbation generation using our method and PGD. It is not meaningful to observe the specific inference time since different computer performance leads to different time. We choose to compute the multiples of the difference in time they take. The final test on StarGAN shows that PGD takes more than ten times as long as our method, which demonstrates the superiority of our method in inference efficiency.

4.5 Ablation Study

In our training process, we use three loss functions. To demonstrate their necessity, we provided ablation evaluation on three different training sessions, each

time using only two of the loss functions, and compare them to our full model. As the ablation evaluation shown in Table 4, eliminating $\mathcal{L}_{pix}$ is inferior to our full model in all parameters. Eliminating $\mathcal{L}_{feat}$ gives a slight advantage in the l_2 distance between the adversarial image x_{adv} and the output of the adversarial image x_{advout}, but falls short of our full model on the other two figures. Eliminating $\mathcal{L}_{adv}$ causes the l_2 distance between the original image x and the adversarial image x_{adv} to reach an extremely low case, but has no effect at all on defense. In summary, the choice of complete model on the loss function is a balance between the size of the perturbation and the effect on defense.

Table 4. Ablation study to remove losses used in our training. Removing any of the losses degrades the performance of the entire active defense framework compared to our proposed method. The data in the table respectively represent the l_2 distance between the original image x and the adversarial image x_{adv}, the adversarial image x_{adv} and the output of the adversarial image x_{advout}, and the original image x and the output of the adversarial image x_{advout}. ($\epsilon = 0.05$)

Loss removed	StarGAN ($L_2 \downarrow$)		
	$l_2(x, x_{adv})$	$l_2(x_{adv}, x_{advout})$	$l_2(x, x_{advout})$
Magnitude loss ($\mathcal{L}_{pix}$)	0.0003	0.0020	0.0028
Magnitude loss ($\mathcal{L}_{feat}$)	0.0006	**0.0015**	0.0028
Adversarial loss ($\mathcal{L}_{adv}$)	1×10^{-7}	0.0436	0.0436
None (ours)	0.0002	0.0018	**0.0024**

5 Conclusion

In this paper, we propose a new method for proactive defense against DeepFakes with image visual retentivity. By training a perturbation generator to obtain perturbations and then adding invisible perturbations to images to combat DeepFakes, our work can help people's photos to be immune to DeepFakes by generating visually similar face images. Experiments on the facial attribute editing, StarGAN, validate the effectiveness of the approach. Compared to the perturbations obtained through adversarial attacks such as PGD, our method obtains smaller perturbations, acquires the perturbations faster, and is more effective against DeepFakes. Therefore, our method can generate face images with stronger protection and more similar to the original image. Currently, the proposed method works only on white-box setups, which does not allow us to obtain effective defenses in realistic black-box situations. Although it works well on single-model counterpart defenses, it does not support cross-models defense. We will consider the defense with black-box and cross-model setup in our future work. Although it is still challenging to design a general, effective, interpretable, and robust proactive defense method against DeepFake, we hope that our method can give new ideas for securing the Internet.

Acknowledgement. This research was funded by National Natural Science Foundation of China No. 62171334, Fundamental Research Funds for the Central Universities No. ZYTS23162 and Scientific Research Foundation of Northwest A&F University No. Z1090121092.

References

1. Goodfellow, I., et al.: Generative adversarial nets. In: Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N., Weinberger, K. (eds.) Advances in Neural Information Processing Systems, vol. 27. Curran Associates, Inc. (2014)
2. Choi, Y., Choi, M., Kim, M., Ha, J.W., Kim, S., Choo, J.: StarGAN: unified generative adversarial networks for multi-domain image-to-image translation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2018
3. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 2019
4. Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., Aila, T.: Analyzing and improving the image quality of StyleGAN. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 2020
5. Li, L., et al.: Face X-ray for more general face forgery detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 2020
6. Luo, Y., Zhang, Y., Yan, J., Liu, W.: Generalizing face forgery detection with high-frequency features. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 16317–16326, June 2021
7. Zhou, Y., Lim, S.N.: Joint audio-visual deepfake detection. In: 2021 IEEE/CVF International Conference on Computer Vision (ICCV), pp. 14780–14789 (2021). https://doi.org/10.1109/ICCV48922.2021.01453
8. Goodfellow, I.J., Shlens, J., Szegedy, C.: Explaining and harnessing adversarial examples (2015)
9. Madry, A., Makelov, A., Schmidt, L., Tsipras, D., Vladu, A.: Towards deep learning models resistant to adversarial attacks (2019)
10. Carlini, N., Wagner, D.: Towards evaluating the robustness of neural networks. In: 2017 IEEE Symposium on Security and Privacy (SP), pp. 39–57 (2017). https://doi.org/10.1109/SP.2017.49
11. Ruiz, N., Bargal, S.A., Sclaroff, S.: Disrupting deepfakes: adversarial attacks against conditional image translation networks and facial manipulation systems (2020)
12. Huang, Q., Zhang, J., Zhou, W., Zhang, W., Yu, N.: Initiative defense against facial manipulation. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 35, no. 2, pp. 1619–1627 (2021). https://doi.org/10.1609/aaai.v35i2.16254
13. He, Z., Wang, W., Guan, W., Dong, J., Tan, T.: Defeating deepfakes via adversarial visual reconstruction. In: Proceedings of the 30th ACM International Conference on Multimedia, MM 2022, pp. 2464–2472. Association for Computing Machinery, New York, NY, USA (2022). https://doi.org/10.1145/3503161.3547923
14. Li, S., et al.: Connecting the dots: detecting adversarial perturbations using context inconsistency. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020. LNCS, vol. 12368, pp. 396–413. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58592-1_24

15. Agarwal, A., Singh, R., Vatsa, M., Ratha, N.: Image transformation-based defense against adversarial perturbation on deep learning models. IEEE Trans. Dependable Secure Comput. **18**(5), 2106–2121 (2021). https://doi.org/10.1109/TDSC.2020.3027183
16. Karras, T., Aila, T., Laine, S., Lehtinen, J.: Progressive growing of GANs for improved quality, stability, and variation. arXiv preprint arXiv:1710.10196 (2017)
17. Perov, I., et al.: Deepfacelab: integrated, flexible and extensible face-swapping framework. arXiv preprint arXiv:2005.05535 (2020)
18. Thies, J., Zollhöfer, M., Stamminger, M., Theobalt, C., Nießner, M.: Face2Face: real-time face capture and reenactment of RGB videos. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2387–2395 (2016). https://doi.org/10.1109/CVPR.2016.262
19. Juefei-Xu, F., Wang, R., Huang, Y., Guo, Q., Ma, L., Liu, Y.: Countering malicious DeepFakes: survey, battleground, and horizon. Int. J. Comput. Vis. **130**(7), 1678–1734 (2022). https://doi.org/10.1007/s11263-022-01606-8
20. Wang, R., Juefei-Xu, F., Luo, M., Liu, Y., Wang, L.: FakeTagger: robust safeguards against deepfake dissemination via provenance tracking. In: Proceedings of the 29th ACM International Conference on Multimedia, MM 2021, pp. 3546–3555. Association for Computing Machinery, New York, NY, USA (2021). https://doi.org/10.1145/3474085.3475518
21. Sun, P., Qi, H., Li, Y., Lyu, S.: FakeTracer: proactively defending against face-swap DeepFakes via implanting traces in training. arXiv preprint arXiv:2307.14593 (2023)
22. Wang, X., Huang, J., Ma, S., Nepal, S., Xu, C.: DeepFake disrupter: the detector of DeepFake is my friend. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 14900–14909 (2022). https://doi.org/10.1109/CVPR52688.2022.01450
23. Huang, H., et al.: CMUA-watermark: a cross-model universal adversarial watermark for combating deepfakes. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 36, no. 1, pp. 989–997 (2022). https://doi.org/10.1609/aaai.v36i1.19982
24. Wang, R., Huang, Z., Chen, Z., Liu, L., Chen, J., Wang, L.: Anti-forgery: towards a stealthy and robust deepfake disruption attack via adversarial perceptual-aware perturbations. In: Raedt, L.D. (ed.) Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22, main Track, pp. 761–767, July 2022. https://doi.org/10.24963/ijcai.2022/107
25. Yeh, C.Y., Chen, H.W., Tsai, S.L., Wang, S.D.: Disrupting image-translation-based DeepFake algorithms with adversarial attacks. In: 2020 IEEE Winter Applications of Computer Vision Workshops (WACVW), pp. 53–62 (2020). https://doi.org/10.1109/WACVW50321.2020.9096939
26. Yeh, C.Y., Chen, H.W., Shuai, H.H., Yang, D.N., Chen, M.S.: Attack as the best defense: nullifying image-to-image translation GANs via limit-aware adversarial attack. In: 2021 IEEE/CVF International Conference on Computer Vision (ICCV), pp. 16168–16177 (2021). https://doi.org/10.1109/ICCV48922.2021.01588
27. Xiao, C., Li, B., Yan Zhu, J., He, W., Liu, M., Song, D.: Generating adversarial examples with adversarial networks. In: Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18, pp. 3905–3911, July 2018. https://doi.org/10.24963/ijcai.2018/543
28. Ronneberger, O., Fischer, P., Brox, T.: U-Net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F.

(eds.) MICCAI 2015. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28
29. Huang, G.B., Mattar, M., Berg, T., Learned-Miller, E.: Labeled faces in the wild: a database for studying face recognition in unconstrained environments. In: Workshop on Faces in 'Real-Life' Images: Detection, Alignment, and Recognition (2008)
30. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4401–4410 (2019)
31. Johnson, J., Alahi, A., Fei-Fei, L.: Perceptual losses for real-time style transfer and super-resolution. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9906, pp. 694–711. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46475-6_43
32. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2018