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

1 Experimental Setup

Dataset. All our experiments are conducted on a popular face dataset CelebFaces Attributes (CelebA) [Liu *et al.*, 2015] which is widely employed in the recent DeepFake studies [Mirsky and Lee, 2021; Juefei-Xu *et al.*, 2021]. The facial images in CelebA are employed for creating fake faces (*e.g.*, , attribute editing, face reenactment, identity swap) via various GANs. CelebA contains more than 200K facial images with 40 attributes annotation for each face. All the facial images are cropped to 256×256.

Model Architectures. Our proposed method is evaluated on three types of DeepFake that involve source images manipulation. For the attribute editing, StarGAN [Choi *et al.*, 2018], AttGAN [He *et al.*, 2019], and Fader Network [Lample *et al.*, 2017] are employed for fine-grained facial attribute editing. For face reenactment, we employ the public available tool Icface [Tripathy *et al.*, 2020] to swap facial expressions. For identity swap, we adopt the popular DeepFake tool faceswap [FaceSwap, 2021] to swap faces freely.

Baselines. In evaluation, we compare our work with prior studies by employing gradient-based PGD [Madry *et al.*, 2017] and optimization-based strategy C&W [Carlini and Wagner, 2017] to generate imperceptible perturbations for a comprehensive comparison.

Evaluation Metrics. To evaluate the visual quality of manipulated facial images by injecting our proposed perceptual-aware perturbations, we adopt three different metrics, the average MSE, PSNR, and SSIM, for measuring the similarity between the original image and the disrupted fake image. Furthermore, we employ attack success rate (ASR) to report the successfully disrupted facial images, where the distortion measured by $L_2 \geq 0.05$.

Implementation. In our comparison experiments, the iteation for the PGD is 10, the optimizer for C&W and our method is Adam, the learning rate is 10^{-4}, the iteration is 500, the ϵ set to 0.05.

Table 1: Performance of our method in tackling a diverse GANs. The performance is evaluated by employing ASR. Our indicates our proposed method.

GAN	**StarGAN**			**AttGAN**			**Fader Network**		
	PGD	**C&W**	**Our**	**PGD**	**C&W**	**Our**	**PGD**	**C&W**	**Our**
StarGAN	-	-	-	7.1	11.5	**13.6**	9.7	**16.8**	15.3
AttGAN	26.3	**37.1**	35.4	-	-	-	18.4	**21.5**	19.6
Fader Network	**16.2**	20.7	18.9	5.3	**7.8**	7.0	-	-	-

Table 2: Performance of our method operating on the Lab color space on StarGAN in an adversary settings with a comparison with other three color space. The input are tranformed by input compression and blurring.

Defense	**RGB**	**Lab**	**HSV**	**CMYK**
JPEG Compression	0.038	**0.058**	0.053	0.041
Gaussian Blur (σ=1)	0.031	**0.049**	0.040	0.038
Gaussian Blur (σ=2)	0.016	**0.025**	0.019	0.013
Gaussian Blur (σ=3)	0.007	**0.012**	0.008	0.007

2 Performance across Diverse GANs

In this section, we explore whether our generated perturbations on one model are effective on other GAN models as well. Table 1 summarizes the experimental results of our proposed method in tackling diverse GANs in black-box settings. Specifically, our method outperforms PGD except one case where the perturbations are generated from Fader Network and apply to StarGAN. However, the other baseline C&W achieved the best performance in almost all the cases. It should be noted that the ASR value of our method is similar to C&W, thus our method also has a good transferability across GANs. Thus, it would be interesting to explore perturbations with strong transferability across diverse GANs, especially to combine C&W for achieving both high transferability and robustness against input transformation attack, which is our future work.

3 Exploring other Color Spaces

To better illustrate the advances by operating on the Lab color space, we investigate the other three popular color space, namely RGB, HSV, and CMYK to explore their performance in disrupting DeepFakes and their stealthiness in evading detection. Experimental results in Table 2 show that our proposed method operating on the Lab color space outperforms the other three baselines measure by L_2 in resisting the two types of input transformation (*e.g.*, compression and Gaussian blur), which exposes more visually artifacts.

To further explore the stealthiness of the perturbations generated by operating the Lab color space, we employ an adversarial noise detector by using local intrinsic dimensionality for detection [Ma *et al.*, 2018]. Experimental results in Table 3 shows that our proposed method is more stealthy than

Table 3: Performance of our method operating on the Lab color space in evading noise detection comparison with three baselines from three different color space.

GAN	**RGB**	**Lab**	**HSV**	**CMYK**
StarGAN	89.51	**85.43**	88.52	88.73
AttGAN	91.53	84.60	**82.53**	89.37
Fader Network	92.31	87.33	87.15	**86.46**

the other three baselines with lower AUC score.

4 Ablation Study

In our ablation study, we explore the trend our added perturbations ϵ and the degree of damaged DeepFakes in comparison with two baselines. Figure 1 shows us the manipulated images with StarGAN by adding the perturbation ϵ from 0.01 to 0.1 with a default interaction 500. Experimental results shown that a large added perturbation leads to large distortion in the created DeepFakes, however large perturbations also introduce unnatural artifacts into the source data which will break our "natural" property requirement. Thus, we set the maximum perturbation to 0.05 in our whole experiments to ensure a fair comparison with the two baselines.

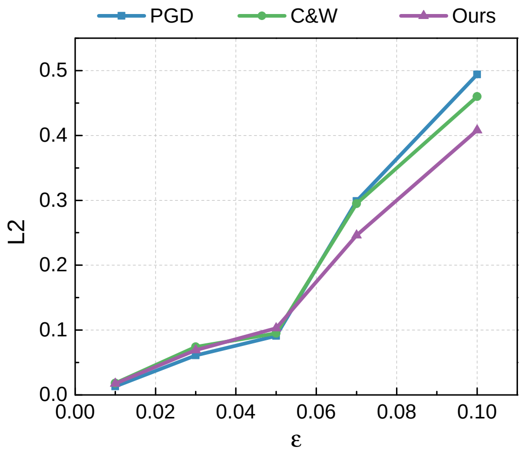

Figure 1: The trend of the magnitude of perturbations ϵ and the intensity of damaged DeepFake.

On the one hand, our generated perturbations for disrupting DeepFakes by exposing noticeable artifacts to avoid users believe the misinformation caused by such DeepFakes, on the other side, we hope that the disrupted DeepFakes could provide strong fake signals for the simple classifier. Table 4 summaries the performance of a simple and popular DeepFake detector [Wang *et al.*, 2020] in spotting unknown DeepFakes. Wang *et al.* 's [Wang *et al.*, 2020] study achieved merely 70.69% in spotting images manipulated by StarGAN while the source image is clean. Wang *et al.* give an accuracy more than 87% in spotting our disrupted images by adding our perceptual-aware perturbations into the source image where the magnitude of the perturbation is only 0.01. Thus, our proposed method show promising results in providing clear fake textual signals for DeepFake detectors in tacking unknown DeepFakes.

Table 4: The performance of Wang *et al.* in spotting unknown DeepFakes.

Clean	iter =10 ($\epsilon = 0.01$)	iter =100 ($\epsilon = 0.01$)	iter =100 ($\epsilon = 0.03$)	iter =300 ($\epsilon = 0.03$)
70.69%	75.52%	**87.81%**	85.54%	80.70%

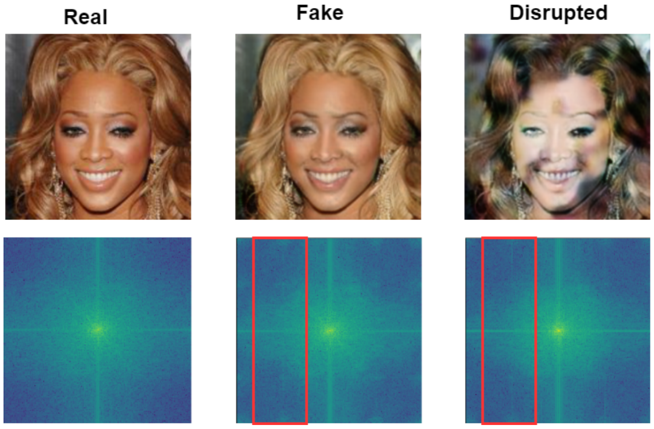

Figure 2: In the first row, the images in turn are a real image from CelebA, a fake image produced by StarGAN on the clean real image, a disrupted image produced by StarGAN on a image added our perceptual-aware perturbations. In the second row, the images are the spectrum corresponding to the images above.

Additionally, we also present Figure 2 to visualize whether our proposed method could enhance the fake textual. In Figure 2, the red rectangle highlights that our created DeepFakes from the images with our added perceptual-aware perturbations exhibit obvious fake textual in spectrum than the fake image manipulated on clean image.

5 Visualization

Figure 3 presents the visualization of our proposed method in disrupting DeepFaked images.

6 Broader Impact

With the rapid development of GAN in image synthesis and the popularity of social media in sharing the exciting moments with personal photos, DeepFake is becoming a real threat to individuals and celebrities as the potential of creating fake pornography and releasing fake statements. The community seeks various countermeasures to fight DeepFakes in both passive and proactive defense manner, unfortunately both of them are still in its fancy and not prepared for tackling this emerging threat. The passive DeepFake detection is an ex-post forensics method while the existing studies for proactive DeepFake disruption are not robust to input reconstruction.

To address the challenges in DeepFake defense, our work is the first attempt to identify and showcase that adversarial perturbations by operating on the Lab color space is not only feasible, but also leads to a robust protection of facial images without introducing visually artifacts. In a large sense, this work can and will provide new thinking into how to better design robust anti-forgery techniques for defending DeepFakes in the wild in order to mitigate the security and privacy concerns caused by the spreading of DeepFakes.

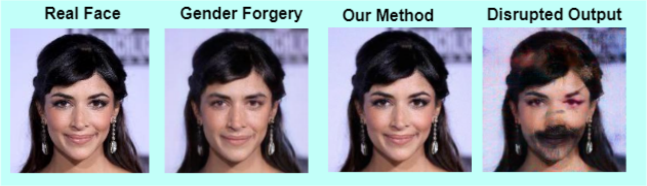

Figure 3: Visualization of our proposed anti-fogery method. The first column is the real face and the second column indicates the forgery face by manipulating gender. The third column indicates the face by adding our perceptual-aware perturbations with high visually quality and the last column represents the disrupted output by manipulating the faces with our added perturbations.

References

[Carlini and Wagner, 2017] Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *2017 ieee symposium on security and privacy (sp)*, pages 39–57. IEEE, 2017.

[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 *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8789–8797, 2018.

[FaceSwap, 2021] FaceSwap. FaceSwap, 2021. https://github.com/Oldpan/Faceswap-Deepfake-Pytorch.

[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. *IEEE transactions on image processing*, 28(11):5464–5478, 2019.

[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.*, 2015] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision*, pages 3730–3738, 2015.

[Ma *et al.*, 2018] Xingjun Ma, Bo Li, Yisen Wang, Sarah M Erfani, Sudanthi Wijewickrema, Grant Schoenebeck, Dawn Song, Michael E Houle, and James Bailey. Characterizing adversarial subspaces using local intrinsic dimensionality. *arXiv preprint arXiv:1801.02613*, 2018.

[Madry *et al.*, 2017] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

[Mirsky and Lee, 2021] Yisroel Mirsky and Wenke Lee. The creation and detection of deepfakes: A survey. *ACM Computing Surveys (CSUR)*, 54(1):1–41, 2021.

[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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8695–8704, 2020.