Attacking On *Disrupting-Deepfakes*: Methodology Transferability And Perturbation Removal

Ting-Chih Chen
Department of Computer Science
Virginia Tech
tingchih@vt.edu

Xiao Guo
Department of Computer Science
Virginia Tech
kevinguo2003@vt.edu

Abstract

In recent years, with the advancement and maturity of technology, deepfakes can not only modify specific attributes but also transfer image styles. Face modification systems can now fully make a fake human to disrupt people's decisions. The main research topic in this area is to improve the performance of deepfakes **(Attacker)** and to protect images from deepfakes' attacks **(Defender)**. The protection method is to add perturbation in the images. The defender can efficiently disrupt facial manipulation systems. In this paper, we mainly focus on how to break down the **Defender**. We propose two attack surfaces: transferability and perturbation removal. We prove the perturbation is only for corresponding generative adversarial networks (GANs), and neural network models can efficiently remove perturbation from images.

Keywords: security, adversarial attacks, GANs, privacy, deepfakes

1 Introduction

Generative adversarial networks (GANs) already have impressive performance on image translation [1, 2, 3], face manipulation [4], and photograph editing [5] topics. Image translation is taking images from one domain and transforming them, so they have the style of images from another domain. Researchers already achieve transforming specific attributes and styles between images. Face manipulation is also a viral topic aiming to modify the age, gender, hair color, and other attributes of the person. People can apply face manipulation to affect the public's decision while no one can determine whether these images and videos are true. Another photograph editing technique is image reconstruction. This technology can not only reconstruct the faces with specific features but also modify the environment and background from photographs, such as removing rain and snow. Combining the above research, GANs already have great results that make humans unable to distinguish which one is true or fake easily.

Due to the widespread availability of these deepfakes, malicious actors can modify images of a person without their content. This action seriously affects personal privacy and reputation issues. There have been existing methods to protect the attacks from deepfakes. The most impressive example is LinkedIn. LinkedIn adds some unnoticeable interference in the source images. This can avoid malefactors from applying people's profile images. The other famous method is the fast gradient sign method (FGSM) [6]. FGSM uses the gradients of the neural network to create an adversarial example. The method uses the gradients of the loss with respect to the input image to create a new image that maximizes the loss. The advantages are: (1) the perturbations can create especially based on input images, and (2) people can not obtain or reproduce the perturbations easily because this is based on neural networks and training data. The current main research direction

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

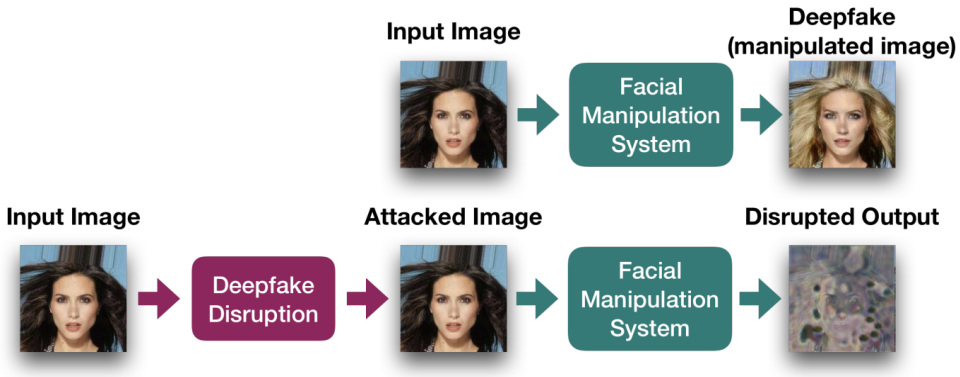

Figure 1: The overview of deepfakes disrupting.

on protecting images is to add the perturbations in the source images. This approach can disrupt deepfakes' results without letting people know the neural network backbone.

In this project, we mainly focus on attacking Disrupting Deepfakes: Adversarial Attacks Against Conditional Image Translation Networks and Facial Manipulation Systems [7]. The pipeline of their scheme is shown in Figure 1. One drawback of this paper is that the authors assume their single-model-based perturbation can successfully disrupt any deepfakes models. In addition, they believe the existing perturbations can not be fully removed using other neural networks. Given the above discussion, we adopt three GAN models [8, 9, 10, 11] to test the perturbation transferability. For the second drawback of the target paper, we accordingly apply auto-encoder and image restoration model [12] to efficiently remove perturbations.

2 Related Work

Deepfakes have been viral for years. Ever since the first paper that automated lips movement based on speech text was published in 1997 [13], researchers and engineers have expanded the usage of deepfakes to other fields like Face Reenactment [14] and Face Replacement [15]. The rise of these models has raised the immediate need for defense strategies. In general, the defense can be divided into two fields: detection and prevention. Detection focus on judging whether the sampled image is from generation methods. A recent paper called has utilized a neural network to achieve high detection accuracy [16]. Another approach to detecting fake images is using Watermarking. The defender can inject artificial fingerprinting to the source image so that the deepfakes images generated will heritage these marks [17].

However, this detection-based defense strategy can't stop attackers from generating images. Thus, more defenders start to explore ways to prevent deepfakes models from generating meaningful content [18]. In our paper, we especially targetting on Perturbation-based defense methods [7, 19]. The authors try to inject invisible noise, or perturbation, into the source image. The result from generating with perturbed image will be severely distorted and unrecognizable to human eyes, achieving the defending goals. However, recent attacks have proposed ways to remove this kind of perturbation. A paper called MagDR uses a complex processing pipeline that integrates both detector and denoiser to achieve near-perfect restoration result [20].

3 Methodology

In this section, we mainly introduce our methods to solve perturbation transferability and perturbation removal. We describe the testing models for perturbation transferability in section 3.1 and the models used for perturbation removal in section 3.2. In the transferability task, we apply few-shot face

translation, GHOST, and StyleGAN-NADA to test the perturbation from starGan. In the perturbation removal task, we use auto-encoder and Restormer to check whether these two models can remove the perturbation or not.

3.1 Transferability

Few-shot face translation is a GAN model with a generator and a discriminator. The generator is an encoder-decoder-based network. Our approach is to extract the latent embedding information from a pre-trained face recognition model with SPAD [11] and AdaIN [10]. When we use AdaIN in the generator, we can obtain the global appearance information, the content of images, and the location of facial features. For example, we can obtain the locations of eyes with AdaIN. After we have the facial features, we fed the facial features into SPAD to generate the semantic images. At last, we use these semantic images to do deepfakes with target images and input the deepfakes images into the discriminator to assist the generator's performance.

GHOST [8] is a GAN model with two encoders, one generator, and one discriminator. This model can generate deepfakes images by mixing the features of source images and attributes of the target images. The first encoder, the identity encoder, is a pre-trained ArcFace model that extracts the features from the source images. The second encoder, the attribute encoder, is a model with a U-Net architecture that extracts attributes from the target images. U-Net is a fully convolutional network. We think U-Net can yield more precise segmentations. The generator, AAD generator, is a model to mix attributes and identities. We use GHOST to generate the deepfakes from the perturbation images. After generating deepfakes images, we fed these images into the discriminator to decide whether these images are fake or real.

StyleGAN-NADA [9] is a CLIP-Guided Domain Adaptation of Image Generators. It converts a pre-trained generator to new domains using only a textual prompt and no training data. For example, one can input several face images and enter "sketch" as a text prompt. Then the model will produce corresponding face images in sketch style without any additional inputs. The magic behind lies in its special architecture. From a high level, StyleGAN-NADA combines the generation ability of StyleGAN with the semantic knowledge of CLIP. It uses the required text description to find the editing direction of latent code in the latent space of pre-trained GAN. During training, there are two phases. Before starting, the model will freeze all the network layers and prevents them from changing any weights. In the first phase, the model selects the k most relevant layers by using Global Clip Loss and observes the layers where its latent code changes the most. Once we determined the most different codes, the training can enter phase II. In this phase, we unfreeze the weights of selected layers and then optimize these layers using Directional CLIP loss.

In our project, we are going to combine target images as input and text prompts to achieve a face-swapping task. To prove our statement regarding transferability is correct, we are hoping that the face-swapping result from the perturbed image and non-protected image shall look the same.

3.2 Perturbation Removal

Auto-encoder: As our target is to remove the added perturbation in the image, the first methodology we think about is the denoising auto-encoder. A denoising auto-encoder is a type of auto-encoder that is trained to remove noise from input data. We hope that the auto-encoder would learn to extract the important features from the images and ignore the noise. To get a more satisfying mapping from the perturbed image to the unprotected image, we train our own denoising auto-encoder model with the perturbed images generated from the original paper's code and the unmodified photo from the CelebA dataset.

However, it's widely known that the denoising auto-encoder is going to produce blurred output as not every detail will be preserved during the down-sampling process. We propose a follow-up **Super Resolution Autoencoder** that will take the blurry predictions from the denoising auto-encoder as the

Figure 2: The few-shot face translation results. The first row is the source images from the CelebA. The second row is deepfakes from source images. The last row is deepfakes from perturbation images.

input and try to produce a more high-quality image. Similarly, we train our model with the perturbed image and a degraded version of the same set. With the combination of two auto-encoders, we expect the result to be both high-quality and free from perturbation.

Restormer [12]: Restormer is an encoder-decoder Transformer for multi-scale local-global representation learning on high-resolution images without disintegrating them into local windows, thereby exploiting distant image context. In the Restormer architecture, it has a multi-Dconv head transposed attention to process local and non-local pixel interactions. This attention layer has enough power to process high-resolution images. In addition, it has a gated-Dconv feed-forward network to control feature transformation. This feed-forward network can decide which information should be propagated into the next transformer block. In the Restormer paper, they have already achieved state-of-the-art performance in image restoration, such as deblurring, deraining, gaussian denoising, and real denoising. We use Restormer to test whether this model can process the invisible perturbation or not.

4 Experiments

In this section, we mainly show our transferability results in section 4.1 and perturbation removal results in section 4.2. Also, we discuss the image similarity evaluation, restoration quality, and removal effectiveness in this section. For the experiment environment, we mainly used Google Colab with a Nvidia T4 graphic card and 24G RAM. The unprotected and perturbed images are all of size 256x256 generated by the original paper's code [7]. We use 500 real images and 500 perturbed images for the experiment. The deepfakes results used by evaluation are also generated by the exact same model that our target paper uses with its highest level of protection setting: Spread-spectrum.

4.1 Transferability

Few-shot face translation results are shown in Figure 2. In these results, we can observe the perturbation generated based on starGAN cannot affect this deepfakes model. Although some examples have different results, most of the deepfake results from perturbed and unprotected source images look the same. One example of the difference is that in the second column, we observe the left eye is obviously different here. One explanation is that the perturbation cannot affect SPAD and AdaIN normalization layers. However, the discriminator's performance is not good to discriminate the images. Since the semantic images are mostly the same, we can generate the same deepfakes.

GHOST results are shown in Figure 3. We think the results are better than the few-shot face translation. The reason is we use two encoders to extract the image features rather than two

Figure 3: The GHOST results. The first row is the source images from the CelebA. The second row is deepfakes from source images. The last row is deepfakes from perturbation images.

Figure 4: The StyleGAN-NADA results. The first and fourth columns are the source images from the CelebA. The second and fifth columns are deepfakes results from source images. The third and last columns are deepfakes from perturbation images. We can see that there is almost no difference between generating perturbed or unprotected images.

normalization layers. In addition, the perturbations cannot affect the encoders to extract the facial features. The result of GHOST model perfectly shows that perturbations added by the author of our target paper does not transfer to other models.

StyleGAN-NADA results are shown in Figure 4. In this experiment, we choose the pretrained model on FFHQ dataset [21]. While swapping the target image(Obama etc.) faces to the victims' faces, we also specify to use sketchy style, as StyleGAN-NADA requires a domain to adapt to. The training iteration was set to 150 and the encoder choice is *e4e*, a novel encoder specifically designed to allow for the subsequent editing of inverted real images. Still, except for some background differences due to noise, the perturbed images have the same face-swapping performance compared to the unprotected images on this model. However, we do notice that the output of the model has a certain level of randomness as the same input from one single image may result in two slightly different images. This is shown numerically by StyleGAN-NADA-Baseline in numerical evaluation.

Table 1: The numerical results from our transferability experiments. We compare the deepfakes results using each model from (1) the perturbed image and (2) the unprotected original image.

Models	MSE	SSIM	PSNR
GHOST	**0.73**	**0.99**	**50.40**
Few-shot Face Translation	57.68	0.95	32.30
StyleGAN-NADA	247.15	0.78	25.26
StyleGAN-NADA-Baseline	226.91	0.80	26.61

Table 2: The numerical result from our perturbation-removal task. The AE_Output denotes the restored image using the auto-encoder while the perturbed images are generated by the original paper's code. Similarly, we compare the result of the restored images using Restormer with the perturbed image in the second row, denoted by Restormer_Output.

Source image pairs	MSE	SSIM	PSNR
AE_Output & Perturbed	267.63	0.68	24.16
Restormer_output & Perturbed	8.46	0.97	39.16

Numerical Evaluation
We show the results in Table 1. We show all three models that the generated deepfakes results from two sources have very high similarity, indicating that the perturbations trained on one model don't transfer to others. Note that StyleGAN-NADA's performance suffers from randomness. We tested it with the exact same image twice for our baseline, it shows the result still presents a numerically-noticeable difference. This explains the relatively poor performance of StyleGAN-NADA.

4.2 Perturbation Removal

Auto-encoder: We used a denoising auto-encoder of 8 Conv layers, Average pooling, and 720k parameters. For the super-resolution, the model includes 10 Conv layers with a mix of upsampling and maximum pooling, resulting in a total of 1.1 million parameters. We train both models for 50 epochs and a batch size of 32. The running time is within minutes. However, due to the limitation of the auto-encoder's structure and implementation difficulties, the result images are still blurred while a small amount of perturbation still remains. We show the graphical in Figure 5.

Restormer: Restormer work mainly focuses on visible perturbation. However, in the results, we can see Restormer removes the most of perturbation. Although some results have disruptions, these disruptions do not affect people to recognize the attribute changes.

Evaluation
We show the results in Table 2 and Table 3. We show that: (1). While suffering from blurring and low image quality, our auto-encoder model does remove most of the perturbation by making the deepfakes results no longer completely distorted. (2) Our Restormer model offers high-quality restoration images with most of the perturbation removed. Also, from Figure 5, we could see that the generated deepfakes image with Restromer's output is very close to the unprotected version. This finding suggests that the perturbation added by the defenders could be removed by image restoration or denoising models. Thus, we conclude the perturbation-based defense proposed by our target paper [7] is broken.

5 Conclusion

In this paper, we mainly explore how to break down the **Defender**. We propose a transferability test and a perturbation removal method. We explore whether the perturbation from another deepfakes model can affect the unknown deepfakes models and the feasibility of using neural networks to remove perturbation. According to our experiments, we have three observations here. Firstly, the perturbation can only affect the corresponding deepfakes model on which they were generated. Secondly, the U-Net backbone network has a great performance in generating deepfakes images

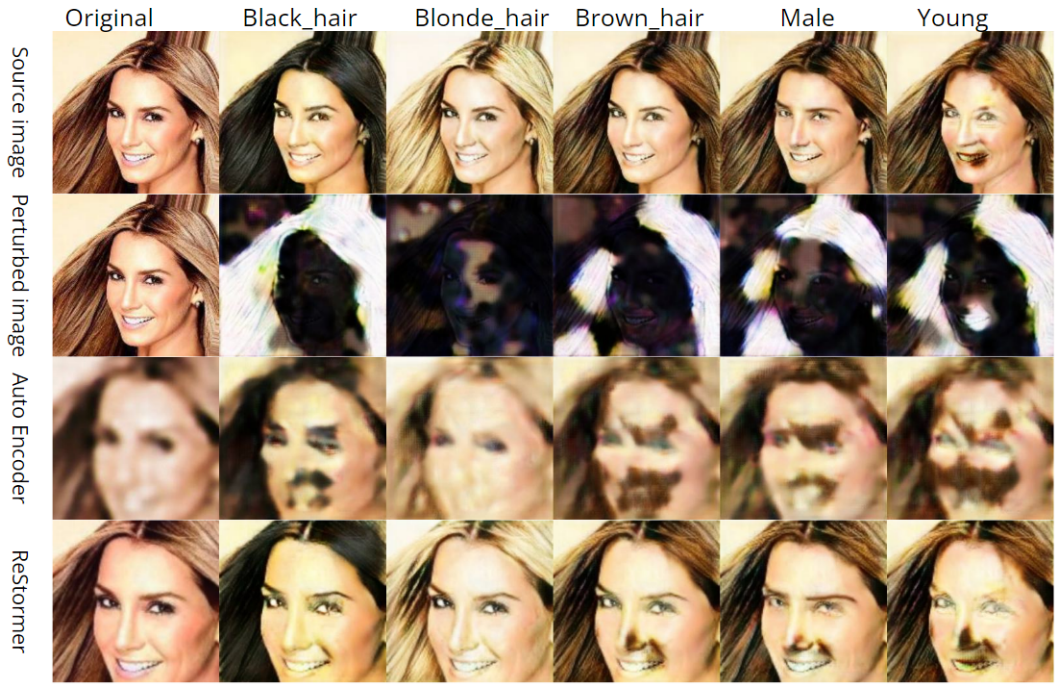

Figure 5: Results of our perturbation-removal task. The first column represents the original image before deepfakes, the rest five columns indicates the five attribute modified by StarGAN. The first row is generated with the image without any perturbation. The second row is perturbed based on StarGAN. The third row is the denoised image from our auto-encoder model. The bottom row is the image from our Restormer model. It's clear that the image was processed with either auto-encoder or Reformer

Table 3: Numerical result of deepfakes performance comparison between restored and unprotected image. AE_Deepfake denotes the deepfakes image generated using the auto-encoder-restored image. Similarly for Restormer_Deepfake, it takes Restormer-generated images into the deepfakes model pipeline. Unprotected here means generating deepfakes with the original image from CelebA dataset without any perturbation.

Deepfakes results Pair	MSE	SSIM	PSNR
AE_Deepfake & Unprotected	5661.78	0.33	11.17
Restormer_Deepfake & Unprotected	5830.58	0.32	11.04

despite of noise or perturbations. Note it's all shown that U-Net also handles the task of denoising images extraordinarily [22]. In addition, the invisible perturbation can also be removed while the original paper only announces support for visible perturbation. Most importantly, because we're simply using Restormer as a per-trained model without any adversarial training, this proves how easily the perturbation added by our target paper can be removed. Without the need for training data, the attacker can break the defense without knowing the details of the defense model. It also suggests the attack can initiate the perturbation removal process in a zero-shot manner because the attackers don't need to query any image pairs from the defense scheme.

Limitation and Future Work In our experiment, we used a small dataset and auto-encoder model due to hardware limitations. A larger sample set with a more complex auto-encoder model shall produce a more pleasant result. However, the auto-encoder method does have its limitations like blurring, we suggest exploring new methodologies like diffusion models based on U-Net [23].

6 Statement Of Work

1. Ting-Chih Chen
    - Few-shot face translation
    - GHOST
    - Auto-encoder
    - Restormer
2. Xiao Guo
    - StyleGAN-NADA
    - Auto-encoder
    - Restormer
    - Evaluation

References

[1] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. Image-to-image translation with conditional adversarial networks, 2016. URL https://arxiv.org/abs/1611.07004.

[2] 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. 2017. doi: 10.48550/ARXIV.1711.09020. URL https://arxiv.org/abs/1711.09020.

[3] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks, 2017. URL https://arxiv.org/abs/1703.10593.

[4] Ruben Tolosana, Ruben Vera-Rodriguez, Julian Fierrez, Aythami Morales, and Javier Ortega-Garcia. Deepfakes and beyond: A survey of face manipulation and fake detection. 2020. doi: 10.48550/ARXIV.2001.00179. URL https://arxiv.org/abs/2001.00179.

[5] Guim Perarnau, Joost van de Weijer, Bogdan Raducanu, and Jose M. Álvarez. Invertible conditional gans for image editing, 2016. URL https://arxiv.org/abs/1611.06355.

[6] Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial machine learning at scale, 2016. URL https://arxiv.org/abs/1611.01236.

[7] Nataniel Ruiz, Sarah Adel Bargal, and Stan Sclaroff. Disrupting deepfakes: Adversarial attacks against conditional image translation networks and facial manipulation systems, 2020. URL https://arxiv.org/abs/2003.01279.

[8] Alexander Groshev, Anastasia Maltseva, Daniil Chesakov, Andrey Kuznetsov, and Denis Dimitrov. Ghost—a new face swap approach for image and video domains. *IEEE Access*, 10: 83452–83462, 2022. doi: 10.1109/ACCESS.2022.3196668.

[9] Rinon Gal, Or Patashnik, Haggai Maron, Gal Chechik, and Daniel Cohen-Or. Stylegan-nada: Clip-guided domain adaptation of image generators, 2021. URL https://arxiv.org/abs/2108.00946.

[10] Ming-Yu Liu, Xun Huang, Arun Mallya, Tero Karras, Timo Aila, Jaakko Lehtinen, and Jan Kautz. Few-shot unsupervised image-to-image translation. 2019. doi: 10.48550/ARXIV.1905.01723. URL https://arxiv.org/abs/1905.01723.

[11] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. 2019. doi: 10.48550/ARXIV.1903.07291. URL https://arxiv.org/abs/1903.07291.

[12] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Restormer: Efficient transformer for high-resolution image restoration, 2021. URL https://arxiv.org/abs/2111.09881.

[13] Christoph Bregler, Michele Covell, and Malcolm Slaney. Video rewrite: Driving visual speech with audio. In *Proceedings of the 24th annual conference on Computer graphics and interactive techniques*, pages 353–360, 1997.

[14] Justus Thies, Michael Zollhofer, Marc Stamminger, Christian Theobalt, and Matthias Nießner. Face2face: Real-time face capture and reenactment of rgb videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2387–2395, 2016.

[15] Lingzhi Li, Jianmin Bao, Hao Yang, Dong Chen, and Fang Wen. Faceshifter: Towards high fidelity and occlusion aware face swapping, 2019. URL `https://arxiv.org/abs/1912.13457`.

[16] David Güera and Edward J. Delp. Deepfake video detection using recurrent neural networks. In *2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 1–6, 2018. doi: 10.1109/AVSS.2018.8639163.

[17] Ning Yu, Vladislav Skripniuk, Sahar Abdelnabi, and Mario Fritz. Artificial fingerprinting for generative models: Rooting deepfake attribution in training data, 2020. URL `https://arxiv.org/abs/2007.08457`.

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

[19] Paarth Neekhara, Brian Dolhansky, Joanna Bitton, and Cristian Canton Ferrer. Adversarial threats to deepfake detection: A practical perspective. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 923–932, 2021.

[20] Zhikai Chen, Lingxi Xie, Shanmin Pang, Yong He, and Bo Zhang. Magdr: Mask-guided detection and reconstruction for defending deepfakes, 2021. URL `https://arxiv.org/abs/2103.14211`.

[21] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks, 2018. URL `https://arxiv.org/abs/1812.04948`.

[22] Chi-Mao Fan, Tsung-Jung Liu, and Kuan-Hsien Liu. Sunet: Swin transformer unet for image denoising, 2022. URL `https://arxiv.org/abs/2202.14009`.

[23] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.