MagDR: Mask-guided Detection and Reconstruction for Defending Deepfakes

Zhikai Chen
Xi'an Jiaotong University
zhikai_chen@outlook.com

Lingxi Xie
Huawei Inc.
198808xc@gmail.com

Shanmin Pang✉
Xi'an Jiaotong University
pangsm@xjtu.edu.cn

Yong He
Xi'an Jiaotong University
hy0275@stu.xjtu.edu.cn

Bo Zhang
Tencent Blade Team
cradminzhang@tencent.com

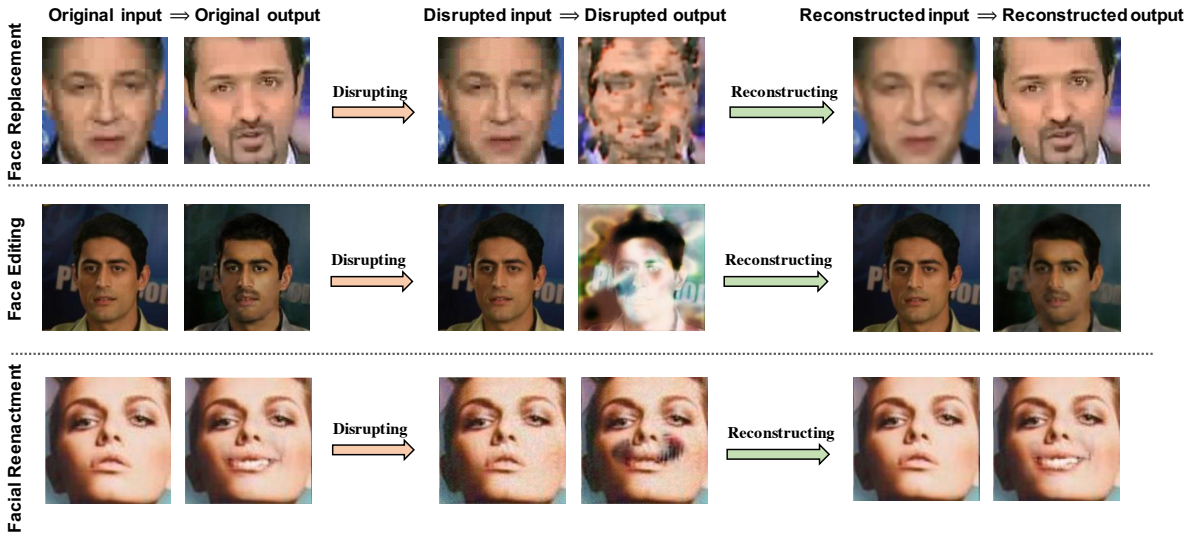

Figure 1: MagDR defends deepfakes from adversarial attacks in three main tasks, face replacement (**top**), face editing (**middle**), and facial reenactment (**bottom**). In each group, we show the original generation effect (**left**), how adversarial perturbations damage the generation (**middle**), and how the proposed defender recovers the desired output (**right**).

Abstract

[1]*Deepfakes raised serious concerns on the authenticity of visual contents. Prior works revealed the possibility to disrupt deepfakes by adding adversarial perturbations to the source data, but we argue that the threat has not been eliminated yet. This paper presents MagDR, a mask-guided detection and reconstruction pipeline for defending deepfakes from adversarial attacks. MagDR starts with a detection module that defines a few criteria to judge the abnormality of the output of deepfakes, and then uses it to guide a learnable reconstruction procedure. Adaptive masks are extracted to capture the change in local facial regions. In experiments, MagDR defends three main tasks of deepfakes, and the learned reconstruction pipeline transfers across input data, showing promising performance in defending both black-box and white-box attacks.*

1. Introduction

Deepfakes originally appeared as a neutral technology that can synthesize images with the human face replaced by another identity. While the technique benefits the community in the scenarios of *e.g.* creating new characters or decorate them with vivid facial expressions, it gradually be-

[1]This work was supported by NSFC under Grant 61972312 and by the Key Research and Development Program of Shaanxi under Grant 2020GY-002.

comes infamous for the unethical applications (*e.g.* swap fake of celebrities into pornographic videos, or generate a fraud video that delivers fake and malicious messages). To avoid negative impacts to the public, researchers started to develop algorithms to detect the images and videos that have been contaminated by deepfakes [36, 56]. However, the follow-up research [2, 10, 31] quickly realized that these detectors are easily fooled by adversarial perturbations. Another way to confront deepfakes is to add adversarial perturbations to the source image so that the output is severely damaged [38, 53]. This was believed to be more robust than the deepfakes detectors.

However, in this paper, we reveal the feasibility of defending the adversarial attacks to deepfakes. We propose a framework named **mask-guided detection and reconstruction** (MagDR). It starts with defining a few criteria (*e.g.*, SSIM, PSNR, *etc.*) that are sensitive to the abnormality of the outputs. Then, a mask-guided detector is trained to judge, from the output image, whether the input image has been contaminated. If yes, a reconstruction algorithm follows to eliminate the damage of the adversarial perturbations and recover the desired output.

A highlight of our approach is that we maintain a number of masks and use them to provide auxiliary information in the detection and reconstruction procedures. The masks can be learned from an individual training process, and each of them corresponds to a specific part of the human face. Guided by the masks, the detector can be partitioned into two components which detect distortion and inconsistency, both of which indicate the regions that are likely to be contaminated. To reconstruct the desired output, we design a pipeline containing several modules and equip them with a changeable execution order and adjustable parameters. Then, we perform an adaptive optimization that suppresses all the pre-defined criteria and produce the recovered output.

We evaluate our approach on two popular datasets, namely, FaceForensics++ [37] and CelebA [23]. We correspond three image-to-image translation methods, CycleGAN [57], StarGAN [6], and GANimation [35], to the three main functions of deepfakes, face editing, facial reenactment, and face replacement, respectively. We investigate two settings, one is the oblivious attack in which the attackers transfer the perturbations computed on original deepfakes models to the defender, and the other is the adaptive attack in which the parameters of the defenders are known to the attacker. Experiments show that deepfakes are vulnerable in both scenarios, but MagDR is able to eliminate the impact of the attack in most cases. Typical examples are shown in Figure 1. MagDR also shows advantages in extensive experiments against state-of-the-art adversarial attackers [4, 25] and defenders [24, 27, 26, 38, 39, 30, 9, 48]. Interestingly, MagDR is able to transfer across different scenarios, demonstrating its ability in both black-box and white-box attacks, and implying that adversarial perturbations are detectable by some common rules.

The contributions of this paper are as follows:

- We reveal that the threats of deepfakes have not yet been eliminated by adding adversarial perturbations to the input image or videos.
- We find that the corruption to image-to-image translation can influence either a part of the image or the entire image. The proposed mask-guided design follows this property and achieves satisfying performance.
- We propose a heuristic, hierarchical reconstruction module for each conditional attribute patch. We adjust it through a progressive approach, which can largely reduce the computational costs. Therefore, we verify that different regions are complementary in recovering the detailed textures, and the layer-by-layer architecture with a proper execution order can enhance the performance of defense.

2. Related Work

Deepfakes. Deepfakes have gained a lot of concern for it can generate fake images, video, voice, *etc*. Those generated products can highly mislead the judgment of humans. [28, 44] survey deepfakes, which divide deepfakes on facial image or video into four main regions: Face Synthesis, Face Editing, Facial Reenactment, and Face Replacement. Face synthesis can generate entire non-existent face images [16, 32, 8] that usually use GAN based methods *e.g.*, ProGAN [15] or StyleGAN [16], *etc*. Face Editing means some attributes of the face can be added, removed, or changed. Those attributes can be the hair, age, clothes, ethnicity, gender [11], *etc*. And the methods often related with GAN with attributes *e.g.*, StarGan [6], attGAN [13] and STGAN [22]. Facial Reenactment modifying the facial expression of the person can be achieved by [22, 43, 42]. Finally, face replacement [33, 19] is an operation to swap the face of the source image to the target image by considering the face size, pose, and skin color *etc*. In this paper, we mainly focus on defending face editing, facial reenactment, and face replacement. This is because that these deepfakes alter source images, while face synthesis does not need any input images.

Adversarial Attack and Defense. Researchers designed a lot of attacking algorithms to add imperceptible perturbations onto well-trained neural networks so that the prediction is dramatically destroyed. Successful scenarios include image classification [17, 29, 25], object detection and semantic segmentation [49], image captioning [51], video classification [5] *etc*. Among the first to introduce adversarial examples against deep neural networks was [41]. After that, Goodfellow *et al.* [12] used the sign

of the gradient to propose a fast attack method called Fast Gradient Sign Method (FGSM). FGSM seeks the direction that can maximize the classification errors to update each pixel. In [25], an iterative method called Projection Gradient Descent (PGD) was proposed. PGD makes the perturbations project back to the ϵ-ball which center is the original data when perturbations over the ϵ-ball. There have been a lot of adversarial defense strategies [55] (Adversarial Detecting[26, 24, 27], Input Reconstruction[26, 50, 21], Adversarial (Re)training[17, 45], *etc*).

Attacking and Defending Deepfakes. A lot of deepfake detectors are proposed to detect fake images in Face Synthesis [40], Face Editing [36, 56], Facial Reenactment[46] and Face Replacement [20]. Despite popularity, recent researches [2, 10, 31] found those deepfake detectors are easily to be misled via adversarial perturbations. Another method to confront deepfakes was proposed by [38, 53]. They found that similar to other traditional computer vision systems, deepfakes are also vulnerable to adversarial examples. Through adding adversarial perturbations to source images, the output can be corrupted, highly influencing the effectiveness of deepfakes. While in this paper, we demonstrate that there is a method to resolve the newly proposed disrupting methods.

3. Methodology

3.1. An Overview of Deepfakes and Disrupted Images Generation

CycleGAN [57] uses two sets of GANs, in which two Generators transform the images from both domains, *i.e.*, $\mathbf{G_x} : \mathbf{x} \rightarrow \mathbf{y}$ and $\mathbf{G_y} : \mathbf{y} \rightarrow \mathbf{x}$, and two Discriminator $\mathbf{D_x}$ and $\mathbf{D_y}$ learn to distinguish between $\mathbf{x}$ and $\mathbf{G_Y}(\mathbf{y})$ as well as between $\mathbf{y}$ and $\mathbf{G_X}(\mathbf{x})$. An enhanced approach named StarGAN [6], which performs image-to-image translations for multiple domains using only a single model. It is a conditional attribute transfer network trained by attribute classification loss and cycle consistency loss. Another work called GANimation [35] reenacts face and uses an expression prediction loss to penalizes $\mathbf{G}$ for realistic expressions. In this model, the output of deepfakes can be simplified as $\mathbf{G}(\mathbf{x}, \mathbf{c})$, where $\mathbf{c}$ is the target class that defines the specific condition where we want to modify in different deepfakes.

Generating disrupted images, which adds some imperceptible perturbation in the source image, is similar to generating adversarial examples. Recently, [38, 53] utilized iterative gradient-based methods (*e.g.* I-FGSM [17]) to generate disrupting images. In this work, we use both the iterative gradient-based method PGD [25] and the optimize-based strategy C&W [4] to generate disrupted images for comprehensive attack settings.

Let $\hat{\mathbf{x}}$ be a generated disrupted input image, *i.e.*, $\hat{\mathbf{x}} = \mathbf{x} + \delta$, where $\mathbf{x}$ is the input image and δ is a human-imperceptible perturbation. As such, the disrupted output can be formulated as $\mathbf{G}(\hat{\mathbf{x}}, \mathbf{c})$, and the objective function for perturbation generation is:

$$\max_{\delta} \boldsymbol{\mathcal{L}}(\mathbf{G}(\hat{\mathbf{x}}, \mathbf{c}), \mathbf{r}), \quad \text{subject to } ||\delta||_\infty \leq \epsilon, \tag{1}$$

where ϵ is the maximum magnitude of the perturbation and $\boldsymbol{\mathcal{L}}$ is the loss function to define the difference between the inputs. If we pick $\mathbf{r}$ to be the ground-truth output, *i.e.*, $\mathbf{r} = \mathbf{G}(\mathbf{x}, \mathbf{c})$, we will get the *ideal* disruption which aims to maximize the distortion of the output.

While sometimes we may need to get some specific altered output image. Accordingly, we need to minimize the distance $\boldsymbol{\mathcal{L}}$ between $\mathbf{G}(\hat{\mathbf{x}}, \mathbf{c})$ and $\mathbf{r_t}$, where $\mathbf{r_t}$ can represent any image we want it to be. This is known as the targeted attack and its formulation can be organized as:

$$\min_{\delta} \boldsymbol{\mathcal{L}}(\mathbf{G}(\hat{\mathbf{x}}, \mathbf{c}), \mathbf{r_t}), \quad \text{subject to } ||\delta||_\infty \leq \epsilon, \tag{2}$$

In addition, to disrupt a network in many defense scenarios, we perform a modified $\boldsymbol{\mathcal{L}}$ calculation method that corresponds to *adaptive attacks* [3, 1, 14]:

$$\max_{\delta} \sum_{k=1}^{K} \boldsymbol{\mathcal{L}}(\mathbf{f}_k(\mathbf{G}(\mathbf{x} + \delta, \mathbf{c}), \mathbf{r})), \quad \text{subject to } ||\delta||_\infty \leq \epsilon, \tag{3}$$

where $\mathbf{f}_k$ is a defense pre-processing operation, and we have K different defense methods with different magnitudes and types.

3.2. The Framework for Defending Deepfakes

The proposed framework is named MagDR, standing for mask-guided detection and reconstruction. As shown in Figure 2 (a), it contains two major components, a detector and a reconstructor, both of which are guided by a set of pre-defined criteria computed on adaptive masks. The overall idea is to sense the presence of adversarial attacks from the output image (which is often significantly perturbed), and perform an adjustable algorithm to suppress all the criteria to an acceptable value, after which we believe that the output has been reconstructed.

Before entering the elaboration of technical details, we point out that the implementation of each module can be freely changed under the designed framework. That being said, we look forward to future research that improves the performance of MagDR by using more accurate criteria as well as stronger detectors and reconstructors.

3.2.1 Predefined Module for MagDR

Distance Metrics Definition. We define $\mathbf{D}$ as the whole set of distance estimation functions. Each $\mathbf{D}_i(\mathbf{x}_i, \mathbf{x}_j)$ rep-

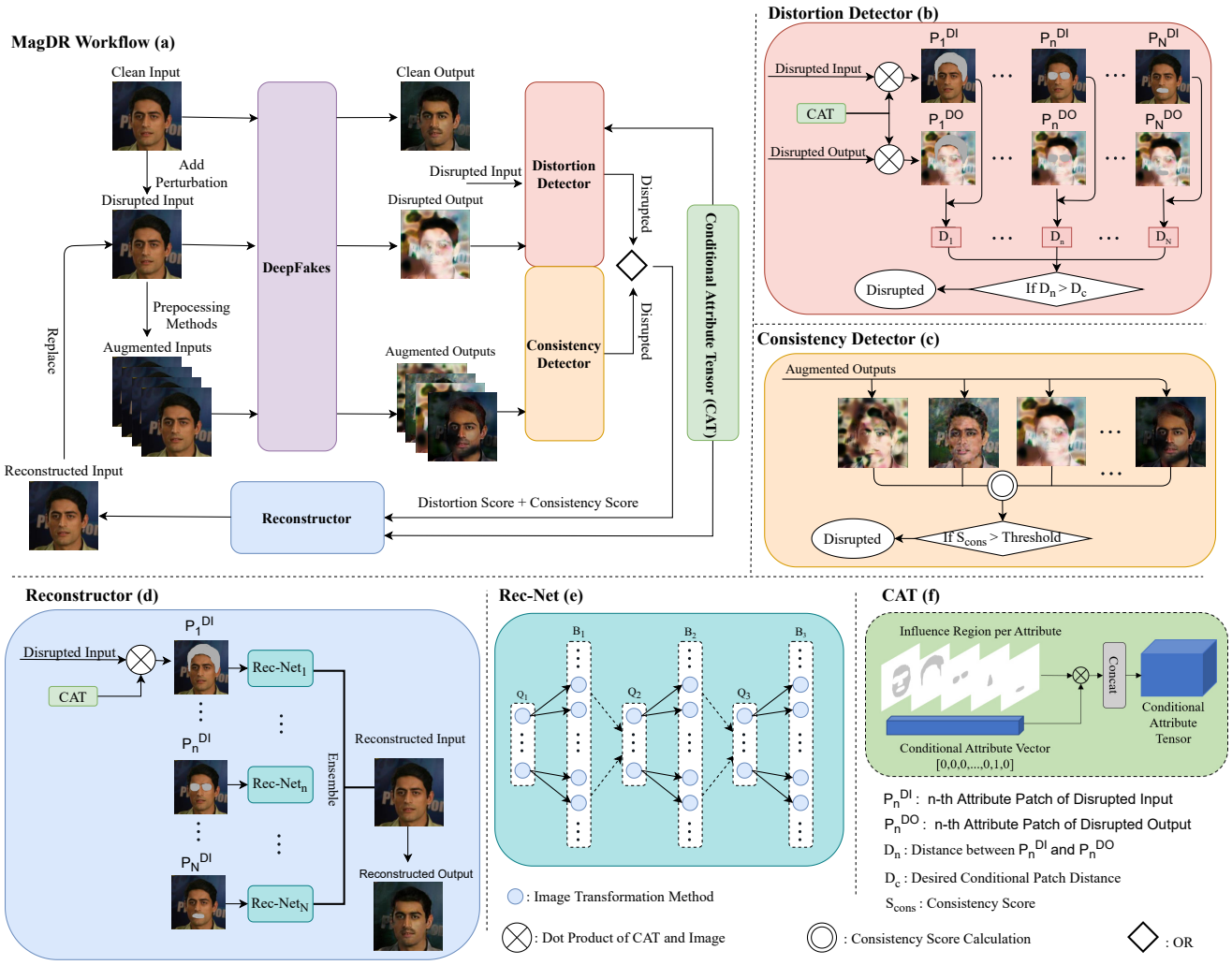

Figure 2: **The MagDR Framework.** It is a unified framework suitable for various deepfake models, *e.g.*, StarGAN, and GANimation. Input with adversarial perturbations is first fed into detectors. If it is considered disrupted, MagDR reconstructs it through reconstructor, replace the disrupted input with reconstructed one, then do the processes again. The detector consists of two sub-detectors, and the core of the reconstructor is Rec-Nets. We use the pre-trained conditional attribute tensor to help detector and reconstructor for detailed image information, where the grey region of masks needs to be preserved when doing calculations.

resents a distance estimation function which calculates the difference of input image pair. In particular, $\mathbf{D}_i$ can be computed in a number of different ways. In our settings, we compute the distances between $\mathbf{\Phi}_l(\mathbf{G}(\hat{\mathbf{x}}, \mathbf{c}), \mathbf{r}))$ and $\mathbf{\Phi}_l(\mathbf{G}(\mathbf{x}, \mathbf{c}), \mathbf{r}))$ on the l-th layer using L_p, SSIM and PSNR, where $\mathbf{\Phi}_l$ is the mapping from an image to its internal DNN representation at layer l. Besides, we calculate the cosine similarity at layer $L-1$, where L denotes the number of layers of the network.

Conditional Mask Tensor Generation. For enhancing the detection and reconstruction, we bring 19-class soft facial region masks in our MagDR from a pre-trained face parser same as MagGAN [47] introduced. The face parser is a modified BiseNet [54] trained on the CelebAMask-HQ dataset [18] [2]. Then we select N attribute region masks based on our deepfake tasks. For each attribute a_i, we define its *influence regions* represented by two probability masks $M_i \in [0,1]^{H \times W}$. Then, we concatenate these attribute mask regions into a conditional mask tensor, where the mask can be denoted as a probability map $\mathbf{M} \in [0,1]^{N \times H \times W}$ of the N facial parts, satisfying $\sum_{i=1}^{N} \mathbf{M}_{i,h,w} = \mathbf{1}_{h,w}$. The module of conditional mask tensor generation is shown in Figure 2 (f).

[2] https://github.com/zllrunning/face-parsing.PyTorch

3.2.2 The Detector

As Figure 2 (a) shown, to detect different patterns of corruption, our detector contains two sub-detectors: a distortion detector and a consistency detector. We combine these two sub-detectors together as a function $d : \mathbb{S} \rightarrow \{0, 1\}$ to decide whether the input is an adversarial sample or not.

Distortion Detector. As Figure 1 shows, deepfakes normally modify the conditional attribute region of the input. Thus, it is weird for the output that there is abnormality out of the conditional attribute region. Motivated by this observation, we use the changes of attribute regions measured by the aforementioned conditional attribute region mask and distance metrics to develop a distortion detector, as shown in Figure 2 (b).

$$\bar{\mathbf{d}}_i = \frac{\mathbf{D}(\mathbf{M}_i \circ x, \mathbf{M}_i \circ \mathbf{G}(\mathbf{x}, \mathbf{c}))}{S_i} \tag{4}$$

where $\bar{\mathbf{d}}_i$ is the distance metrics vector with the conditional attribute region, and S_i is the number of pixels of the attribute region mask.

In practice, we cannot determine whether the image is polluted or not by directly using the distance of the target attribute region larger than a threshold. This is because that the threshold is affiliated with specific deepfake tasks and images. To address the problem, we propose to use the distance metrics of the target attribute region as the benchmark to compare with other regions.

$$\mathbf{V}_i = \begin{cases} 1 & \text{if } \max\{\bar{\mathbf{d}}_i - \mathbf{d}_c\} \geq 0, \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

where $\mathbf{d}_c$ is the distance of the target attribute region M_c. Based on the difference of conditional regions, we can conclude the input is disrupted or not. In particular, if the distance metrics of the calculated region is larger than the benchmark, we will assume this region is corrupted and set the flag as 1. Otherwise, it is considered as clean and set as 0. Thus, we naturally use the number of disrupted patches to decide whether an image is disrupted or not.

In addition, we can use the vector $\bar{\mathbf{d}}_i$ to define distortion score that measures the distortion magnitude of the disrupted output. Specifically, it is formulated as:

$$\mathbf{S}_{\text{dist}} = \sum_{i \neq c} \mathbf{w} \circ \text{sigmoid}(\bar{\mathbf{d}}_i) \tag{6}$$

where $\mathbf{w}$ is the regularization vector of the distance metric $\bar{\mathbf{d}}_i$.

Consistency Detector. The distortion detector is effective in detecting disrupted images when there are a lot of corruptions out of the conditional attribute region. However, it becomes less effective if corruption is caused in the whole image. To overcome this problem, we consider the vulnerability of adversarial perturbations as a breakthrough to detect disrupted images. We use many image processing methods to obtain a set of augmented images. These augmented images are regarded as the protected input of deepfakes. Correspondingly, there are a lot of different representations contained in the output. With them, as shown in Figure 2 (c), we calculate the consistency score $\mathbf{S}_{\text{cons}}$:

$$\begin{aligned} \mathbf{d}^k &= \mathbf{D}(\mathbf{G}(\hat{\mathbf{x}}, \mathbf{c}), \mathbf{G}(\mathbf{f}_k(\hat{\mathbf{x}}), \mathbf{c})) \\ \boldsymbol{\mu} &= \mathbf{E}_{\mathbf{f}_k|\mathbf{F}}[\mathbf{d}^k] \\ \boldsymbol{\sigma}^2 &= \mathbf{E}_{\mathbf{f}_k|\mathbf{F}}[(\mathbf{d}^k - \boldsymbol{\mu})^2] \\ \mathbf{S}_{\text{cons}} &= \boldsymbol{\sigma} \end{aligned} \tag{7}$$

Here, K is the number of image pre-processing methods. If $\mathbf{S}_{\text{cons}}$ is larger than the predefined threshold, we treat the input image as the adversarial example which may influence the whole image region, and vice versa.

3.2.3 The Reconstructor

The reconstructor aims to recover the correct output of deepfakes by reducing the effect of added perturbations. Formally, it is modelled as minimizing the distance between original output and reconstructed output:

$$\min \mathbf{D}(\mathbf{G}(\mathbf{x}, \mathbf{c}), \mathbf{G}(\mathbf{T}(\hat{\mathbf{x}}), \mathbf{c})) \tag{8}$$

where $\mathbf{T}(\cdot)$ is an image transformation function. Ideally, $\mathbf{T}(\cdot)$ should be model-agnostic, sophisticated and non-differentiable, making it harder for the adversary to circumvent the transformed model by back-propagating the distance metrics through it.

As shown in Figure 2 (d), our proposed approach uses the image restoration technique to purify disrupted images. It has two components, which together form an effective pipeline that is difficult to bypass. First, we apply the conditional region mask to help us obtain specific facial patches. Second, we use a multi-stage module Rec-Net shown in Figure 2 (e), to enhance the image quality and simultaneously remove adversarial perturbations. Rec-Net is the core component of the Reconstructor, and its algorithmic description is outlined in Algorithm. 1.

The final criteria score $\mathbf{S}_{\text{final}}$ uses the distortion score $\mathbf{S}_{\text{dist}}$ in Eq. 6 and the consistency score $\mathbf{S}_{\text{cons}}$ in Eq. 7 to judge the abnormality of the output:

$$\mathbf{S}_{\text{final}} = \lambda \mathbf{S}_{\text{dist}} + \mathbf{S}_{\text{cons}} \tag{9}$$

where λ is the hyper-parameter to balance two different detectors.

3.3. Advantages of Proposed Method

Our proposed method offers a number of advantages. First and most important, it is agnostic to attack algorithms

Algorithm 1: Rec-Net Training for a Patch

/* Pre-training Procedure */
Input: Training set $\mathbf{X}$, Transformation methods f_k^p for reconstruction in layer $\mathbf{B}_i$, $\mathbf{S}_{\text{final}}$ refer to Eq. 9
Output: Layer $\mathbf{Q}_i$

K = the number of method categories in $\mathbf{B}_i$.
$\mathbf{P}$ = the number of parameters for each category in f_k.
Evaluate each methods f_k^p in training set $\mathbf{X}$ by $\mathbf{S}_{\text{final}}$.
Find the optimal parameters $\mathbf{q}$ for each category methods f_k.
Insert f_k^q into layer $\mathbf{Q}_i$.
K = the number of methods in $\mathbf{Q}_i$.

/* Obtain the i-th transformation method for rec-net $\mathbf{R}$ */
Input: Input image patch x, Layer $\mathbf{Q}_i$, Layer $\mathbf{B}_i$, $\mathbf{S}_{\text{final}}$
Output: Processing methods sequence of rec-net $\mathbf{R}$

Evaluate each methods f_k^q in layer $\mathbf{Q}_i$ for patch x.
Select the top-3 method categories $\mathbb{C}_j$ based on the $\mathbf{S}_{\text{final}}$.
Use the top-3 categories as the index, obtain the subset layer $\mathbf{B}'_{\mathbf{i}}$.
Select the top-1 scored method $f^b_{\mathbb{C}_j}$ in the subset as the optimal transformation method in layer $\mathbf{B}_{\mathbf{i}}$.
Insert $f^b_{\mathbb{C}_j}$ into $\mathbf{R}_{\mathbf{i}}$.

and attacked models. Second, as it leverages the vulnerability of perturbations, it thus can achieve strong detection efficiency of those perturbations with little altering. Third, it takes strong adaptive defense ability with different attack degrees. Fourth, unlike many recently proposed techniques, which degrade critical image information as part of their defense, our proposed method preserves image quality while simultaneously providing a strong defense. Last, due to its modular nature, the proposed approach can be used as a universal module in existing deepfake models.

4. Experiments

4.1. Datasets and Model Architecture

Datasets. We mainly use two datasets in our experiments: FaceForensics++ [37] and CelebA [23]. FaceForensics++ contains 1000 original video sequences. And all of them have been manipulated by Deepfakes, Face2Face, FaceSwap and NeuralTextures. The data is collected from 977 youtube videos and all videos contain a trackable mostly frontal face without occlusions. CelebFaces Attributes Dataset (CelebA) contains 200K celebrity images that cover large pose variations and background clutter. CelebA has large diversities, large quantities, and rich annotations, including 10,177 number of identities, 202,599 number of face images, and 5 landmark locations, 40 binary attributes annotations per image. The size of each image is cropped to 128×128.

Model Architectures. We use the CycleGAN, StarGAN, and GANimation image translation architectures to demonstrate our framework on different scenarios mentioned above. For CycleGAN, we use FaceForensics++ dataset to train a face to face model with 200 epochs. For StarGAN and GANimation, We use the open-source implementation [3] refered in Nataniel *et al.* [38] and fine-tuned on the CelebA dataset.

4.2. Attack Settings and Evaluation Metircs

We mainly use C&W and PGD methods to craft adversarial examples. They are different attack methods in which one is gradient-based and another is optimization-based, which can prove our attack setting is comprehensive. We also use the different hyper-parameters, as well as the loss function to control the distortion in the different deepfake models. For CycleGAN attacking, we use the target attack settings in which the target label is the input image. And perform the untarget attack in StarGAN and GANimation models that could make more distortion of the output. These objective functions can refer to Eq. 1 and Eq. 2.

Specifically, all of those images are under the attack success situation and the magnitude of perturbation is constrained in the ϵ norm ball. And we adopt the *recall* rate, the *precision* rate and the *F1* score to quantify the detection performance. All experiments are run on the same set of images and against the same attacks for a fair comparison.

4.3. Detection Performance

We compare our proposed detector with a number of state-of-the-art adversarial examples detectors. These include training a model to distinguish the difference of adversarial examples and normal samples [24, 27, 7], calculating the reconstruction errors to detect adversarial examples [26], and training a network to do binary classification on disrupted and clean outputs (OTD).

4.3.1 Detecting Defense-unaware Attacks

We test the performance of three deepfakes under the defense-unaware attack, where the attackers generate the disrupted images in the models without any defense modules. We calculate the precision, accuracy, and F1 score of detectors on disrupted images w.r.t. the perturbation ϵ. Table. 1 shows the performance of detectors under different attack methods and different deepfake scenarios. From Table. 1, we can see that our method performs the best in all cases. Under the same attack settings, those methods proposed to detect adversarial examples in the classification tasks are at a low performance of detection. And we also perform the **ablation study** to demonstrate the effectiveness of the consistency detector (CD) and distortion detector (DD). The results show consistency detector performs well in the face replacement. Because the corruption in the

[3] https://github.com/nataniellruiz/disrupting-deepfakes

Table 1: Comparison of detection performance across different attack methods and deepfakes. Note that all of these perturbations are generated under the defense-unaware situation.

Defense Models	Attack Methods	Face Replacement			Face Editing			Facial Reenactment		
		Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1
Lu *et al.* [24]	C&W	0.48	0.50	0.43	0.62	0.57	0.52	0.56	0.60	0.56
	PGD	0.45	0.53	0.46	0.66	0.65	0.64	0.50	0.55	0.50
Metzen *et al.* [27]	C&W	0.52	0.52	0.51	0.48	0.46	0.40	0.48	0.51	0.44
	PGD	0.46	0.44	0.43	0.50	0.55	0.49	0.48	0.47	0.40
Meng *et al.* [26]	C&W	0.51	0.51	0.50	0.53	0.53	0.52	0.49	0.50	0.46
	PGD	0.56	0.56	0.56	0.90	0.89	0.89	0.56	0.55	0.55
OTD	C&W	0.28	0.53	0.36	0.20	0.33	0.25	0.60	0.57	0.46
	PGD	0.27	0.52	0.35	0.52	0.51	0.50	0.45	0.49	0.36
NNIF *et al.* [7]	C&W	0.82	0.66	0.73	0.83	0.85	0.84	0.85	0.54	0.66
	PGD	0.78	0.62	0.69	0.83	0.79	0.81	0.80	0.52	0.63
DD (ours)	C&W	0.84	1.00	0.91	0.96	0.96	0.96	0.92	0.92	0.92
	PGD	0.76	1.00	0.87	0.99	0.99	0.99	0.94	0.94	0.94
CD (ours)	C&W	0.87	0.99	0.92	0.96	0.96	0.96	0.76	0.86	0.81
	PGD	0.95	0.94	0.94	0.98	0.99	0.99	0.97	0.78	0.87
MagDR (CD+DD)	C&W	0.96	1.00	**0.98**	0.96	0.96	**0.96**	0.92	0.92	**0.92**
	PGD	0.95	1.00	**0.97**	1.00	1.00	**1.00**	0.97	0.94	**0.95**

situation is huge and influenced in the whole image, the distortion detector can not get a proper benchmark for comparison. And the distortion detector is good at detecting partial corruption. So it performs well in the facial reenactment. Finally, MagDR combines the advantages of the two components to obtain superior detection performance.

4.3.2 Detecting Defense-aware Attacks

For a complete analysis, we investigate the detection performance under the defense-aware attack, which is also called adaptive attack [1]. It is the most difficult defense scenarios because the adversary knows the technique details of the detection methods. When launching an attack, the attacker can leverage the knowledge to fool the detector by generating specific perturbations. The adaptive-attack methods can refer to Eq. 3, which ensures it can reduce the performance of detectors through more iterations. While a good detection method should increase the attack cost which means attackers should continuing to increase the iterations for the desired attack. As Figure 3 shows, with the iteration increase, the F1-score going down under the adaptive attack settings. And the detectors which trained on some datasets shows their poor performance and high vulnerability. While our method can greatly keep the stability of highest detection performance under more aggressive attacks.

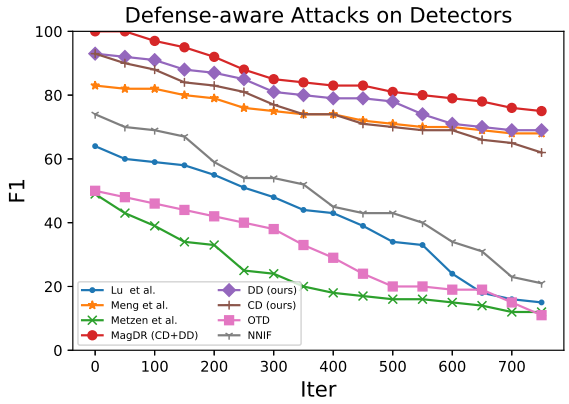

Figure 3: Illustration the F1-score under defense-aware attacks. Iter denotes the iteration number of searching the adversarial perturbations.

4.4. Reconstruction Performance

We compare our proposed reconstructor with number of recently introduced state-of-the-art image transformation based defense schemes in the literature. These include JPEG Compression [9], Adversarial training + Blur [38], Auto-encoder reformer [26], Random Noise [48], Super-Resolution [30], Me-net [52], Pixel Deflection(PD) [34].

As Table. 2 shows, we evaluate reconstructors under two dimensions: input-pair and output-pair. We expect the reconstructors should alter less in the input-pair while keeping high similarity in the output-pair. The results show the randomized method [48, 34] does not have any reconstruction ability in deepfake tasks, even make things worse. The Auto-encoder-based reconstructor [52, 26] has a huge difference between the desired images. And our method can perform superior in both two evaluated dimensions.

Figure 4 shows the effect of all of the compared defense methods on a disrupted image. The perturbations applied to

Table 2: Performance comparison with state-of-the-art reconstruction mechanisms. (I) is the difference between original input and reconstructed input. (O) is the difference between original outputs and reconstructed outputs.

Metrics	Xie [48]	Prakash [34]	Meng [26]	Yang [52]	Nataniel [38]	Gintare [9]	Mustafa [30]	MagDR
MSE (I)	25.59	25.56	27.66	21.17	24.59	17.69	15.53	**11.40**
SSIM (I)	0.72	0.72	0.76	0.81	0.75	0.82	0.88	**0.89**
PSNR (I)	32.90	32.91	30.12	32.44	33.25	36.11	38.60	**39.92**
Feature Similarity (I)	0.72	0.73	0.67	0.73	0.69	0.72	0.75	**0.77**
MSE (O)	257.19	242.34	34.98	27.47	43.64	35.25	30.60	**23.93**
SSIM (O)	0.09	0.08	0.68	0.75	0.69	0.82	0.86	**0.88**
PSNR (O)	12.86	13.37	28.08	30.17	28.26	30.12	31.35	**33.48**
Feature Similarity (O)	0.09	0.14	0.66	0.74	0.66	0.71	0.75	**0.76**

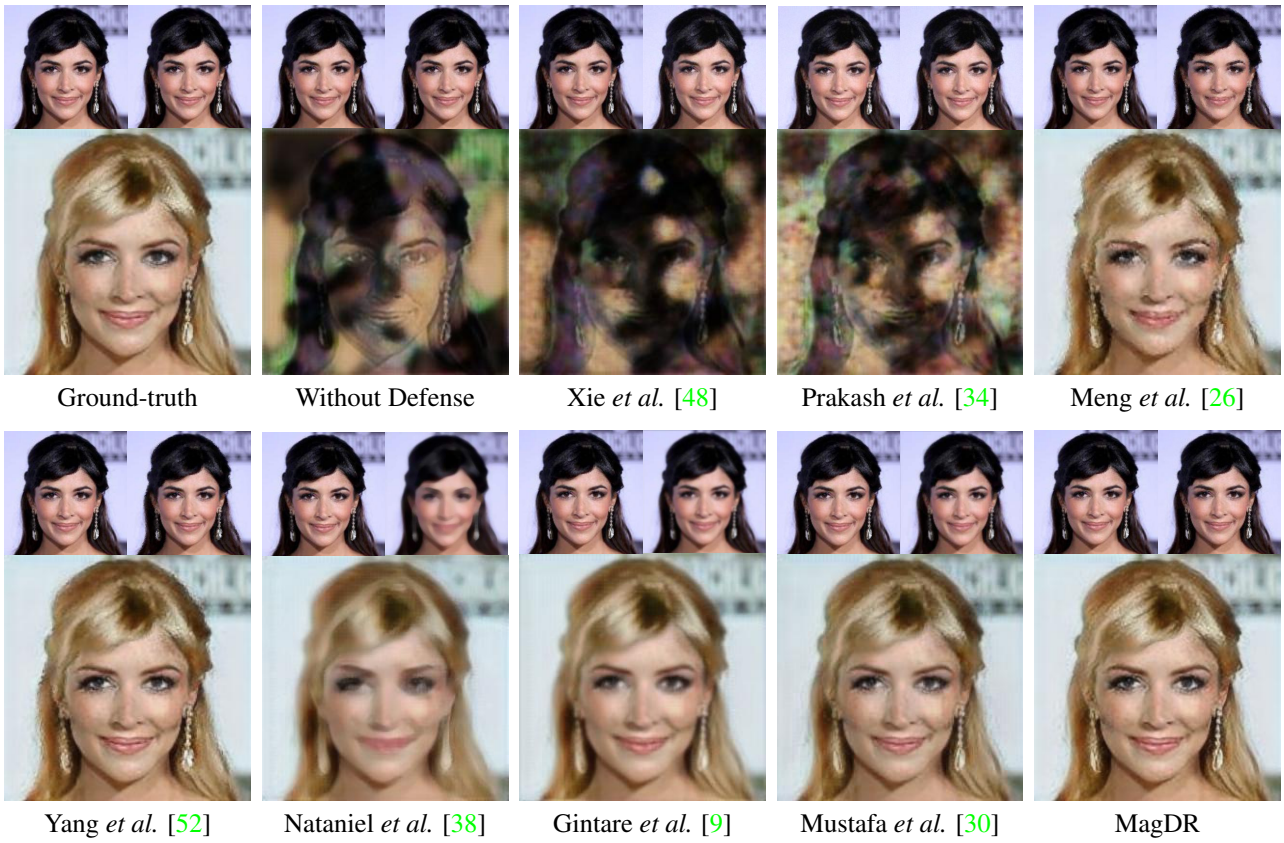

Figure 4: Visual comparison of the deepfakes outputs of different reconstructed inputs by defense methods. In each case, (top-left) is the disrupted input, (top-right) is the reconstructed input, and (bottom) is the output obtained from the reconstructed input. The output images are enlarged two times for better visualization.

samples are the same. And the reconstruction performance is quite equally with which is reflected in Table. 2.

5. Conclusions

This paper presents a two-step framework named MagDR (mask-guided detection and reconstruction) to defend deepfakes from adversarial attacks. The core idea is to compute a few unsupervised criteria that are sensitive to the adversarial perturbations on the output image. Then, an iterative process involving detection and reconstruction is performed, recovering the output to the desired form.

Beyond the promising results, our work delivers a message to the community that image-to-image translation algorithms seem easier to protect themselves from adversarial attacks because the attacks often generate meaningless patterns on the output image (rather than semantic predictions), making themselves easy to be detected. We expect the attacks to become stronger when they realize this weakness and generate more 'natural' perturbations, and it may raise new challenges to defend such 'smarter' attackers.

References

[1] Nicholas Carlini, Anish Athalye, Nicolas Papernot, Wieland Brendel, Jonas Rauber, Dimitris Tsipras, Ian Goodfellow, Aleksander Madry, and Alexey Kurakin. On evaluating adversarial robustness. *arXiv preprint arXiv:1902.06705*, 2019. 3, 7

[2] Nicholas Carlini and Hany Farid. Evading deepfake-image detectors with white- and black-box attacks, 2020. 2, 3

[3] Nicholas Carlini and David Wagner. Adversarial examples are not easily detected: Bypassing ten detection methods. In *ACM Workshop on Artificial Intelligence and Security (AISec)*, 2017. 3

[4] 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. 2, 3

[5] Zhikai Chen, Lingxi Xie, Shanmin Pang, Yong He, and Qi Tian. Appending adversarial frames for universal video attack, 2019. 2

[6] Y. Choi, M. Choi, M. Kim, J.W. Ha, S. Kim, and J. Choo. StarGAN: Unified Generative Adversarial Networks for Multi-Domain Image-to-Image Translation. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018. 2, 3

[7] Gilad Cohen, Guillermo Sapiro, and Raja Giryes. Detecting adversarial samples using influence functions and nearest neighbors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 6, 7

[8] H. Dang, F. Liu, J. Stehouwer, X. Liu, and A. Jain. On the Detection of Digital Face Manipulation. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020. 2

[9] Gintare Karolina Dziugaite, Zoubin Ghahramani, and Daniel M. Roy. A study of the effect of jpg compression on adversarial images, 2016. 2, 7, 8

[10] Apurva Gandhi and Shomik Jain. Adversarial perturbations fool deepfake detectors, 2020. 2, 3

[11] E. Gonzalez-Sosa, J. Fierrez, R. Vera-Rodriguez, and F. Alonso-Fernandez. Facial Soft Biometrics for Recognition in the Wild: Recent Works, Annotation and COTS Evaluation. *IEEE Transactions on Information Forensics and Security*, 13(8):2001–2014, 2018. 2

[12] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv:1412.6572*, 2014. 2

[13] Z. He, W. Zuo, M. Kan, S. Shan, and X. Chen. AttGAN: Facial Attribute Editing by Only Changing What You Want. *IEEE Transactions on Image Processing*, 28(11):5464–5478, 2019. 2

[14] Cormac Herley and Paul C Van Oorschot. Sok: Science, security and the elusive goal of security as a scientific pursuit. In *2017 IEEE Symposium on Security and Privacy (S&P)*, pages 99–120. IEEE, 2017. 3

[15] T. Karras, T. Aila, S. Laine, and J. Lehtinen. Progressive Growing of GANs for Improved Quality, Stability, and Variation. In *Proc. International Conference on Learning Representations*, 2018. 2

[16] T. Karras, S. Laine, and T. Aila. A Style-Based Generator Architecture for Generative Adversarial Networks. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019. 2

[17] Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial examples in the physical world. *arXiv:1607.02533*, 2016. 2, 3

[18] Cheng-Han Lee, Ziwei Liu, Lingyun Wu, and Ping Luo. Maskgan: towards diverse and interactive facial image manipulation. *arXiv preprint arXiv:1907.11922*, 2019. 4

[19] Lingzhi Li, Jianmin Bao, Hao Yang, Dong Chen, and Fang Wen. Faceshifter: Towards high fidelity and occlusion aware face swapping. *arXiv preprint arXiv:1912.13457*, 2019. 2

[20] Lingzhi Li, Jianmin Bao, Ting Zhang, Hao Yang, Dong Chen, Fang Wen, and Baining Guo. Face x-ray for more general face forgery detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5001–5010, 2020. 3

[21] Bin Liang, Hongcheng Li, Miaoqiang Su, Xirong Li, Wenchang Shi, and XiaoFeng Wang. Detecting adversarial image examples in deep neural networks with adaptive noise reduction. *IEEE Transactions on Dependable and Secure Computing*, page 1–1, 2019. 3

[22] M. Liu, Y. Ding, M. Xia, X. Liu, E. Ding, W. Zuo, and S. Wen. STGAN: A Unified Selective Transfer Network for Arbitrary Image Attribute Editing. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019. 2

[23] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015. 2, 6

[24] Jiajun Lu, Theerasit Issaranon, and David Forsyth. Safetynet: Detecting and rejecting adversarial examples robustly. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 446–454, 2017. 2, 3, 6, 7

[25] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv:1706.06083*, 2017. 2, 3

[26] Dongyu Meng and Hao Chen. Magnet: a two-pronged defense against adversarial examples. In *Proceedings of the 2017 ACM SIGSAC conference on computer and communications security*, pages 135–147, 2017. 2, 3, 6, 7, 8

[27] Jan Hendrik Metzen, Tim Genewein, Volker Fischer, and Bastian Bischoff. On detecting adversarial perturbations. *arXiv preprint arXiv:1702.04267*, 2017. 2, 3, 6, 7

[28] Yisroel Mirsky and Wenke Lee. The creation and detection of deepfakes: A survey, 2020. 2

[29] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. Deepfool: A simple and accurate method to fool deep neural networks. In *CVPR*, 2016. 2

[30] Aamir Mustafa, Salman H. Khan, Munawar Hayat, Jianbing Shen, and Ling Shao. Image super-resolution as a defense against adversarial attacks. *IEEE Transactions on Image Processing*, 29:1711–1724, 2020. 2, 7, 8

[31] Paarth Neekhara, Shehzeen Hussain, Malhar Jere, Farinaz Koushanfar, and Julian McAuley. Adversarial deepfakes: Evaluating vulnerability of deepfake detectors to adversarial examples, 2020. 2, 3

[32] J.C. Neves, R. Tolosana, R. Vera-Rodriguez, V. Lopes, H. Proença, and J. Fierrez. GANprintR: Improved Fakes and Evaluation of the State-of-the-Art in Face Manipulation Detection. *IEEE Journal of Selected Topics in Signal Processing*, 2020. 2

[33] Yuval Nirkin, Yosi Keller, and Tal Hassner. Fsgan: Subject agnostic face swapping and reenactment. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 7184–7193, 2019. 2

[34] Aaditya Prakash, Nick Moran, Solomon Garber, Antonella DiLillo, and James Storer. Deflecting adversarial attacks with pixel deflection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8571–8580, 2018. 7, 8

[35] Albert Pumarola, Antonio Agudo, Aleix M Martinez, Alberto Sanfeliu, and Francesc Moreno-Noguer. Ganimation: Anatomically-aware facial animation from a single image. In *Proceedings of the European conference on computer vision (ECCV)*, pages 818–833, 2018. 2, 3

[36] C. Rathgeb, A. Botaljov, F. Stockhardt, S. Isadskiy, L. Debiasi, A. Uhl, and C. Busch. PRNU-based Detection of Facial Retouching. *IET Biometrics*, 2020. 2, 3

[37] Andreas Rossler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. Faceforensics++: Learning to detect manipulated facial images. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1–11, 2019. 2, 6

[38] Nataniel Ruiz, Sarah Adel Bargal, and Stan Sclaroff. Disrupting deepfakes: Adversarial attacks against conditional image translation networks and facial manipulation systems. 2020. 2, 3, 6, 7, 8

[39] Pouya Samangouei, Maya Kabkab, and Rama Chellappa. Defense-gan: Protecting classifiers against adversarial attacks using generative models, 2018. 2

[40] Joel Stehouwer, Hao Dang, Feng Liu, Xiaoming Liu, and Anil Jain. On the detection of digital face manipulation. *arXiv preprint arXiv:1910.01717*, 2019. 3

[41] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv:1312.6199*, 2013. 2

[42] J. Thies, M. Zollhöfer, and M. Nießner. Deferred Neural Rendering: Image Synthesis using Neural Textures. *ACM Transactions on Graphics*, 38(66):1–12, 2019. 2

[43] J. Thies, M. Zollhofer, M. Stamminger, C. Theobalt, and M. Nießner. Face2face: Real-Time Face Capture and Reenactment of RGB Videos. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2016. 2

[44] 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. 2

[45] Florian Tramèr, Alexey Kurakin, Nicolas Papernot, Ian Goodfellow, Dan Boneh, and Patrick McDaniel. Ensemble adversarial training: Attacks and defenses. *arXiv preprint arXiv:1705.07204*, 2017. 3

[46] 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 Conference on Computer Vision and Pattern Recognition*, volume 7, 2020. 3

[47] Yi Wei, Zhe Gan, Wenbo Li, Siwei Lyu, Ming-Ching Chang, Lei Zhang, Jianfeng Gao, and Pengchuan Zhang. Maggan: High-resolution face attribute editing with mask-guided generative adversarial network. In *Proceedings of the Asian Conference on Computer Vision (ACCV)*, November 2020. 4

[48] Cihang Xie, Jianyu Wang, Zhishuai Zhang, Zhou Ren, and Alan Yuille. Mitigating adversarial effects through randomization. In *International Conference on Learning Representations*, 2018. 2, 7, 8

[49] Cihang Xie, Jianyu Wang, Zhishuai Zhang, Yuyin Zhou, Lingxi Xie, and Alan Yuille. Adversarial examples for semantic segmentation and object detection. In *ICCV*, 2017. 2

[50] Weilin Xu, David Evans, and Yanjun Qi. Feature squeezing: Detecting adversarial examples in deep neural networks. *Proceedings 2018 Network and Distributed System Security Symposium*, 2018. 3

[51] Yan Xu, Baoyuan Wu, Fumin Shen, Yanbo Fan, Yong Zhang, Heng Tao Shen, and Wei Liu. Exact adversarial attack to image captioning via structured output learning with latent variables. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4135–4144, 2019. 2

[52] Yuzhe Yang, Guo Zhang, Dina Katabi, and Zhi Xu. Me-net: Towards effective adversarial robustness with matrix estimation. *arXiv preprint arXiv:1905.11971*, 2019. 7, 8

[53] Chin-Yuan Yeh, Hsi-Wen Chen, Shang-Lun Tsai, and Sheng-De Wang. Disrupting image-translation-based deepfake algorithms with adversarial attacks. In *The IEEE Winter Conference on Applications of Computer Vision (WACV) Workshops*, March 2020. 2, 3

[54] Changqian Yu, Jingbo Wang, Chao Peng, Changxin Gao, Gang Yu, and Nong Sang. Bisenet: Bilateral segmentation network for real-time semantic segmentation. In *ECCV*, 2018. 4

[55] XiaoYong Yuan, Pan He, Qile Zhu, and Xiaolin Li. Adversarial examples: Attacks and defenses for deep learning. *IEEE Transactions on Neural Networks and Learning Systems*, PP:1–20, 01 2019. 3

[56] X. Zhang, S. Karaman, and S.F. Chang. Detecting and Simulating Artifacts in GAN Fake Images. In *Proc. IEEE International Workshop on Information Forensics and Security*, 2019. 2, 3

[57] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017. 2, 3