IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 26 December 2023, accepted 7 January 2024, date of publication 15 January 2024, date of current version 25 January 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3353785

RESEARCH ARTICLE

Coexistence of Deepfake Defenses: Addressing the Poisoning Challenge

JAEWOO PARK, LEO HYUN PARK, (Graduate Student Member, IEEE), HONG EUN AHN, AND TAEKYOUNG KWON, (Member, IEEE)
Graduate School of Information, Yonsei University, Seoul 03722, South Korea

Corresponding author: Taekyoung Kwon (taekyoung@yonsei.ac.kr)

This work was supported in part by the Institute of Information and Communications Technology Planning and Evaluation (IITP) Grant funded by the Korea Government through the Ministry of Science and Information and Communications Technology (MSIT), South Korea, (Advanced and Proactive AI Platform Research and Development Against Malicious Deepfakes), under Grant RS-2023-00230337; and in part by MSIT under the Information Technology Research Center (ITRC) Support Program supervised by IITP under Grant IITP-2023-2020-0-01602.

ABSTRACT As Generative Adversarial Networks advance, deepfakes have become increasingly realistic, thereby escalating societal, economic, and political threats. In confronting these heightened risks, the research community has identified two promising defensive strategies: proactive deepfake disruption and reactive deepfake detection. Typically, proactive and reactive defenses coexist, each addressing the shortcomings of the other. However, this paper brings to the fore a critical yet overlooked issue associated with the simultaneous deployment of these deepfake countermeasures. Genuine images gathered from the Internet, already imbued with disrupting perturbations, can lead to data poisoning in the training datasets of deepfake detection models, thereby severely affecting detection accuracy. We propose an improved training framework to address this problem in deepfake detection models. Our approach involves purifying the disrupting perturbations in disruptive images using a backward process of the denoising diffusion probabilistic model (DDPM). Images purified using our DDPM-based technique closely mimic the original, unperturbed images, thereby enabling the successful generation of deepfake images for training purposes. Moreover, our purification process outperforms DiffPure, a prominent adversarial purification method, in terms of speed. While conventional defensive techniques struggle to preserve detection accuracy in the face of a poisoned training dataset, our framework markedly reduces this accuracy drop, thus achieving superior performance across a range of detection models. Our experiments demonstrate that deepfake detection models trained using our framework exhibit an increase in detection accuracy ranging from 11.24%p to 45.72%p when compared to models trained with the DiffPure method. Our implementation is available at https://github.com/seclab-yonsei/Anti-disrupt.

INDEX TERMS Deepfake, deepfake detection, deepfake disruption, data poisoning, adversarial purification.

I. INTRODUCTION

Generative Adversarial Networks (GANs) [1] have revolutionized the field of synthetic face generation. With rapid progress in GAN-based technologies [2], [3], [4], the emergence of highly convincing deepfake imagery poses a formidable challenge, often imperceptible to the naked eye. Such advancements have substantial implications, especially when misused for nefarious purposes in cybercrimes, with the potential to inflict severe political, social, and economic harm. In response, research communities have intensified to develop robust defense mechanisms against deepfake exploitation.

Defense strategies against deepfakes typically fall into two categories: proactive and reactive methods. Proactive defenses aim to preemptively thwart the creation or spread of deepfakes, while reactive defenses focus on identifying and mitigating the effects post-creation. These two paradigms

The associate editor coordinating the review of this manuscript and approving it for publication was S. K. Hafizul Islam.

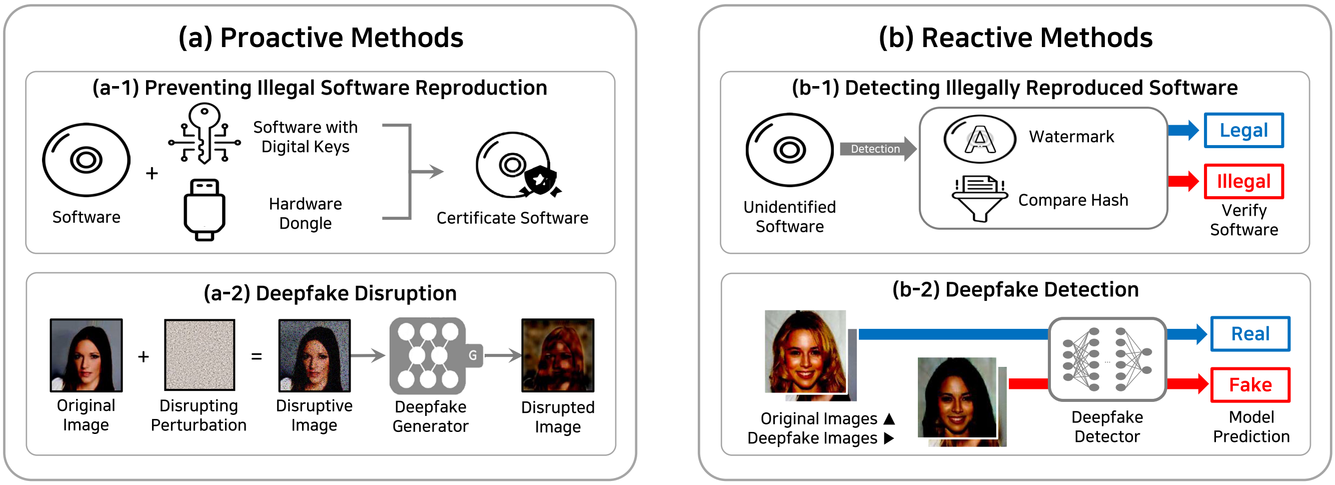


FIGURE 1. **Defense techniques are categorized as either proactive or reactive based on their timing of intervention. Proactive methods are designed to prevent issues before they arise, while reactive methods address issues post-occurrence. This categorization also applies to deepfake defense strategies.**

generally operate in tandem, each fortifying the other, to form a comprehensive security infrastructure. Their synergistic application is crucial for a robust defense against the multifaceted threats posed by deepfake technology.

In this context, two countermeasures are currently under investigation: proactive deepfake disruption and deepfake detection. Deepfake disruption [5] aims to obstruct the generation of deepfakes by introducing disruptive perturbations to the authentic image. This method, inspired by existing adversarial attacks [6], [7], [8], utilizes a gradient of loss as a perturbation that maximizes the distortion of the deepfake generator output. Conversely, deepfake detection [9], [10] is a technology that employs deep learning models to discern whether an input image is authentic or fabricated. Techniques for deepfake detection include spatial-based detection [11], which seeks visual artifacts; frequency-based detection [12], which identifies features in the frequency domain; and biological signal-based detection [13], which leverages biological signals.

The two defense techniques have distinct roles and naturally coexist within a deepfake environment. However, this prompts a fundamental question: *Are there any complications arising from the simultaneous adoption of both countermeasures?* Due to the divergent objectives of disruption and detection, we have found that disruption can inadvertently contribute to the issue of data poisoning, leading to a substantial decrease in detection model accuracy. Disruption techniques aim to ensure that images uploaded by users onto the Internet are safeguarded from use as deepfake images. In contrast, detection techniques necessitate the generation of a training dataset from deepfake images downloaded from the Internet. Therefore, the training dataset for detectors can potentially be contaminated through the inclusion of disrupted images. While these issues may arise in real-world applications, they have not been previously addressed. To tackle this problem, it's crucial to ensure the effective operation of the deepfake generator. Adversarial defense techniques can be implemented to stabilize the model's output. However, existing solutions fail to generate fake images that are advantageous for the detection model. Adversarial training [5] helps form robust model parameters but can inadvertently cause model deformations. On the other hand, adversarial purification effectively eliminates perturbations in the input. However, the leading method called DiffPure [14] compromises the semantics of the original input by introducing unnecessary noise.

In this paper, we introduce a detector training framework designed to address the problem we first raised above. For this purpose, We employ a diffusion model [15] to cleanse the perturbations from the distorted real images, without altering the parameters of the deepfake generator responsible for producing the fake training images. Our approach uses a diffusion model similar to DiffPure [14] but we do not conduct a forward process that introduces noise to the input. Instead, we perform solely a reverse process to remove perturbations, rendering purification more effective and efficient. We evaluated the performance of our framework in comparison to DiffPure [14] and adversarially trained StarGAN (StarGAN AT) [5]. The results of our purification process yield L_2 distances comparable to those of existing purification methods. Furthermore, the distribution of deepfake images produced by our method aligns more closely with the original deepfakes compared to existing methods.

The contributions of this paper are as follows:

- We revisit deepfake disruption and detection as countermeasures from a conventional security perspective. Our first concern is that the coexistence of two countermeasures, *disruption* and *detection*, against deepfakes may inevitably cause their conflict at the training phase of detection models(§III). To highlight this concern, we empirically demonstrate that the accuracy of detection models drops significantly when their training datasets are poisoned by disrupted images.

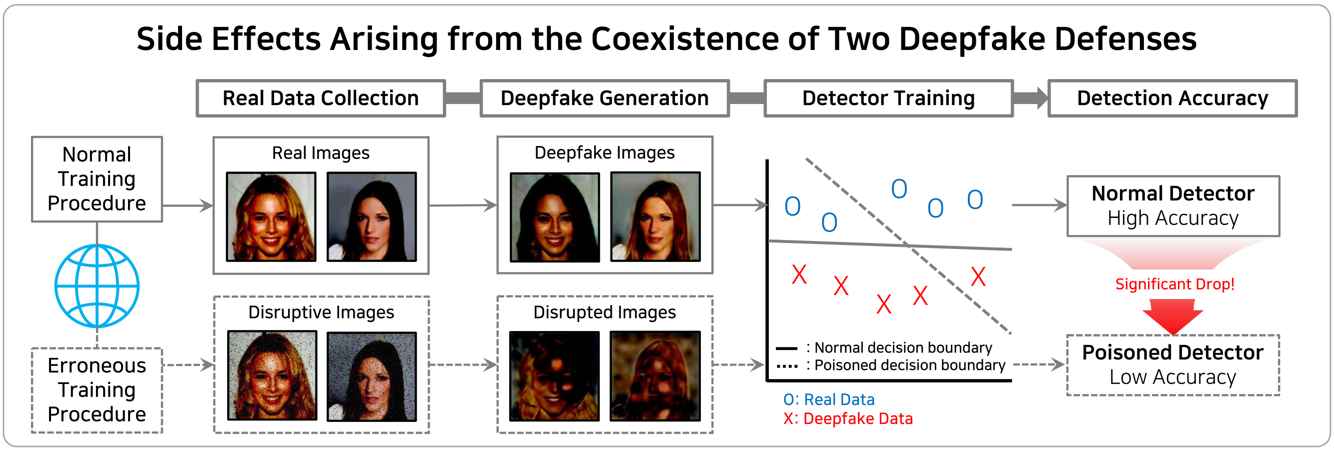


FIGURE 2. Background and motivation of this work. The coexistence of two defense mechanisms can lead to data poisoning in the training dataset of detectors. If disruptive images are collected from the Internet and used to train the detector, the detection accuracy will be significantly reduced.

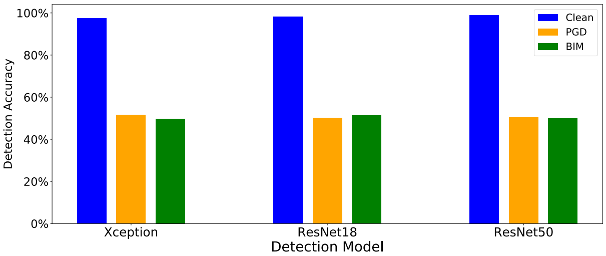


FIGURE 3. Dectection accuracy. Deepfake detection models are trained on a clean dataset and a poisoned dataset, respectively. PGD and BIM are adversarial attack methods used for disruption. From a defender's perspective, the percentage of poisoned data is set to 100%, assuming the worst-case scenario.

- We introduce a robust training method that involves the purification of disruptive perturbations to mitigate data poisoning in deepfake detection models (§IV). By capitalizing on the denoising method DDPM and bypassing the step of adding random noise, our approach decreases both the distortion and execution time of the generated outputs.
- Through comparative analysis with DiffPure and adversarially trained StarGAN, we demonstrate that our framework is successful in generating real images that result in deepfake images, visually similar to standard fake images. This similarity is comparable to that achieved by existing methods (§V-B). Furthermore, our method significantly outperforms existing methods in terms of detection accuracy when the training dataset is poisoned (§V-C).

II. BACKGROUND

In this section, we summarize traditional security issues, deepfake issues, and their respective defenses to view the coexistence of deepfake defense methods that are not covered from a traditional security perspective.

A. SECURITY CONCERNS IN THE PHYSICAL WORLD

1) TRADITIONAL ISSUES

As technology develops, various risks arise in various fields. Examples include software piracy, network intrusion, and image piracy.

a: SOFTWARE PIRACY

Software piracy is the use or distribution of software without permission or payment of fair value. Software piracy infringes on copyright and causes economic losses for software developers. In the long run, this can lead to a decrease in the quality of the software.

b: NETWORK INTRUSION

Network intrusion means gaining access to an unauthorized network or server. Allowing unauthorized access to a network can result in the theft or corruption of data on the network.

c: IMAGE PIRACY

Image piracy is the use of an image without permission from the copyright holder. Image piracy, similar to software piracy, infringes on the creator's copyright and causes economic losses. Creators suffer from reduced creativity and, in the long run, lower-quality images.

2) RECENT ISSUE: DEEPFAKE

There arise many side effects that exploit the technology of generative AI. The most representative problem is deepfake. Deepfake is a technology that modifies the face or features of a person in an image or video. The improvement of generative AI makes it difficult for the human eye to distinguish the difference between an AI-generated image and a normal image. By exploiting this, attackers can create deepfake images by blending the faces of acquaintances into sensitive images, or create deepfake images of politicians to cause political damage. In May 2023, a deepfake video of an explosion at the Pentagon went viral, causing stock markets to drop.[1] A video of the president of warring Ukraine declaring his surrender went viral as well.[2] Similar to traditional security issues, deepfakes undermine the reliability of information and can instigate legal and social issues. Thus, it is imperative to adopt a traditional security perspective that encompasses both proactive and reactive approaches to adequately defend against the threats posed by deepfakes.

B. PROACTIVE DEFENSE METHOD

The proactive defense method aims to prevent issues before they occur, as shown in Fig.1-a.

1) TRADITIONAL PROACTIVE DEFENSE

A variety of proactive defense methods such as software piracy prevention, firewalls, and watermarking have been used to prevent the traditional security concerns described in II-A1.

a: SOFTWARE PIRACY PREVENTION

Software piracy prevention strategies are designed to combat illegal copying by employing obfuscation techniques or ensuring that only authorized users have access to the software. Typically, these approaches incorporate the use of license keys [16] or hardware dongles [17], and of late, blockchain technology [18], which has been applied across various fields.

b: FIREWALL

A firewall [19] functions as a security barrier that selectively blocks communication from specific IPs or ports to prevent unauthorized network access. It operates on predefined security rules to filter out potential attackers by restricting entry through unauthorized ports.

c: WATERMARKING

Watermarking [20] operates as a prevention against unauthorized use by embedding a unique mark, such as a company logo or owner's mark, into an image. However, this method carries the risk that watermarks, if merely added, can be removed, permitting the unauthorized use of protected content.

2) PROACTIVE DEEPFAKE DEFENSE: **DEEPFAKE DISRUPTION**

Similar to traditional proactive defense methods, deepfake disruption prevents potential issues associated with deepfake techniques, which are described in II-A2. Deepfake disruption introduces a perturbation to the real image, as depicted in Fig. 1-(a-2), prompting the deepfake generator to produce an image significantly different from the typical deepfake image. Although there are various methods to create perturbations, we focus on the approach that leverages adversarial attacks [5]. Adversarial attacks generate perturbations using the gradient of a deep learning model to deviate the output from the correct result. In this paper, we employ BIM [7] and PGD [8] attacks to generate disrupting perturbations, thereby interrupting deepfake generation. Further details on adversarial attacks can be found in §VII-B.

C. REACTIVE DEFENSE METHOD

The reactive defense methods aim to address issues after they have occurred, as depicted in Fig.1-b.

1) TRADITIONAL REACTIVE DEFENSE

A variety of methods are utilized to address the issues delineated in II-A1. Examples of traditional reactive defense methods include software piracy detection, network intrusion detection, and adding digital identifiers.

a: SOFTWARE PIRACY DETECTION

Software piracy detection is a method that verifies the legitimacy of the software in use by checking for any unauthorized modifications or copies through techniques like software integrity checking.

b: NETWORK INTRUSION DETECTION

Network intrusion detection systems examine network communications within a device to determine their legitimacy and identify potential malicious intent. A notable example is Kitsune [21], which measures anomalies in network traffic for use in network intrusion detection systems(NIDS).

c: DIGITAL IDENTIFIER

Digital Identifiers are similar to adversarial attack techniques; they embed data within images using noise patterns imperceptible to the human eye. This embedded data can then trace the source of an image's leak or distribution and identify the responsible parties.

2) REACTIVE DEEPFAKE DEFENSE: **DEEPFAKE DETECTION**

Deepfake detection, similar to traditional reactive defense methods, addresses issues arising from deepfake technology, which is detailed in II-A2. Deepfake detection is a technique that primarily uses DNN models to ascertain whether an input image is a deepfake, as depicted in Fig. 1-(b-2). The training dataset for DNN models used in deepfake detection requires a substantial collection of real data, along with deepfakes generated from this collected data. In this paper, we select Xception [22] and ResNet [23], which are widely employed for deepfake detection, as our target models. Both models [11], [24] exploit unnatural artifacts present in the image to detect deepfakes.

[1] https://edition.cnn.com/2023/05/22/tech/twitter-fake-image-pentagon-explosion/index.html

[2] https://www.foxnews.com/world/zelenskyy-deepfake-quality-manipulate-opinions

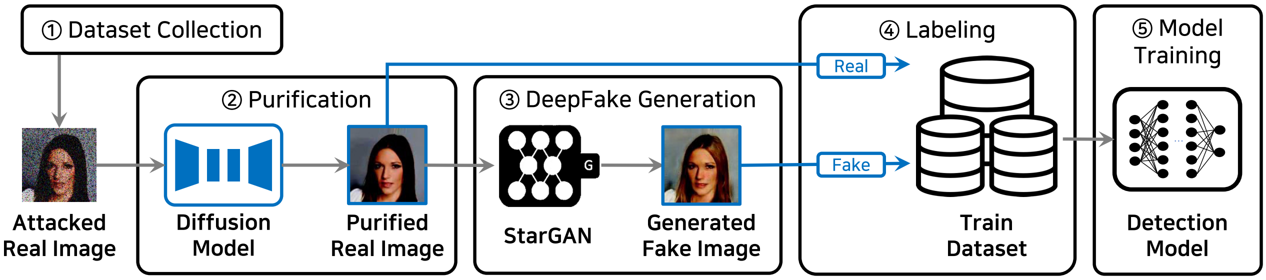


FIGURE 4. Overview of our detection model training framework. The disruptive image with disruption perturbation is purified through the diffusion model and put into the deepfake generator to generate a deepfake image. Generated deepfake images are used to train the deepfake detection model.

III. PROBLEM DEFINITION: COEXISTENCE OF DEEPFAKE DEFENSE METHODS

Deepfake defense methods, such as deepfake disruption and deepfake detection, are designed to operate at different stages within a deepfake ecosystem. Diverse methods in a traditional security ecosystem currently operate at different stages. They not only coexist but also synergize, thereby reinforcing the overall defense. However, the efficacy of this paradigm is uncertain when applied to the deepfake ecosystem. Unlike the harmonious coexistence of strategies in traditional cases, the coexistence of deepfake disruption and detection techniques introduces a unique issue, as depicted in Fig. 2: the issue of 'training data poisoning'. This phenomenon ultimately reduces detection accuracy.

Users who employ disruption may upload their perturbed images x' to the Internet to prevent their images from being exploited for deepfake generation. However, if these disruptive images x' are inadvertently collected for deepfake detector D training data while sourcing real data x from the Internet, the quality of the training data for detectors suffers. In essence, the training dataset becomes inundated with disruptive images, as opposed to standard deepfake images. As a result, vital information about deepfake images fails to reach the detectors, resulting in a significant decrease in their accuracy.

Fig. 3 illustrates the detection accuracy drop observed in our small experiments. Note that deepfake detectors perform quite accurately when their training data is comprised of clean real images only, but the detection accuracy severely drops almost by half when training data is constructed with disruptive real images in the worst-case scenario. From the defender's perspective, we set the percentage of poisoned data to 100% for the rest of this paper, assuming the worst case scenario.

IV. ROBUST TRAINING FRAMEWORK FOR DEEPFAKE DETECTION AGAINST DATA POISONING

A. OVERVIEW

Our framework proceeds in a sequence of five steps as shown in Fig. 4: ① data collection, ② purification, ③ deepfake generation, ④ dataset labeling, and ⑤ model training. Our approach has two differences from the traditional training of deepfake detection models in Fig. 2 (c). First, we newly deploy the purification step to address disruptive real images. Second, we discard the disruptive images and only consider the purified images from them for the later steps.

In the ① data collection step, we collect real images, regardless of whether the images are disruptive. Instead, all collected images are then fed into the diffusion model during the ② purification step. The diffusion model outputs new real images from the disruptive images where perturbations are removed. We adopt DDPM [15] as the diffusion model structure. In the ③ deepfake generation step, we feed the output images of the diffusion model into the generative model instead of the collected real images to get the normal fake images. The generative model produces results that are almost similar to the output of the normal real images. We label the purified real images as real and the generated fake images as fake during the ④ dataset labeling step. Finally, in the ⑤ training step, we train the detector using the labeled data.

B. DIFFUSION PURIFICATION

Basic idea of DDPM. In the training phase, DDPM takes an input image x_0, transforms it into a completely noisy image x_T through a forward process, and then restores it back to the original image x_0 through a reverse process. The forward process is denoted as $q\left(\mathrm{x}_t \mid \mathrm{x}_{t-1}\right)$. The image x_t is generated by combining the preceding image x_{t-1} with noise $\mathbf{I}$ in a ratio of $1-\beta_t$ to β_t. This process creates a noisy image by introducing noise up to the targeted timestep T, as dictated by t. Here, t is an integer between 0 and T that represents the level of noise in the image.

$$q\left(\mathrm{x}_t \mid \mathrm{x}_{t-1}\right) := \mathcal{N}\left(\mathrm{x}_t; \sqrt{1-\beta_t}\mathrm{x}_{t-1}, \beta_t\mathbf{I}\right) \tag{1}$$

The reverse process, denoted as $p_\theta\left(\mathrm{x}_{t-1} \mid \mathrm{x}_t\right)$, essentially reverses the forward process to produce an image x_0' from the noisy image x_T. An image x_{t-1} at timestep $t-1$ is generated using the mean, $\mu\theta\left(\mathrm{x}_t, t\right)$, and variance, $\Sigma_\theta\left(\mathrm{x}_t, t\right)$, derived from the image x_t from the previous step.

$$p_\theta\left(\mathrm{x}_{t-1} \mid \mathrm{x}_t\right) := \mathcal{N}\left(\mathrm{x}_{t-1}; \mu_\theta\left(\mathrm{x}_t, t\right), \Sigma_\theta\left(\mathrm{x}_t, t\right)\right) \tag{2}$$

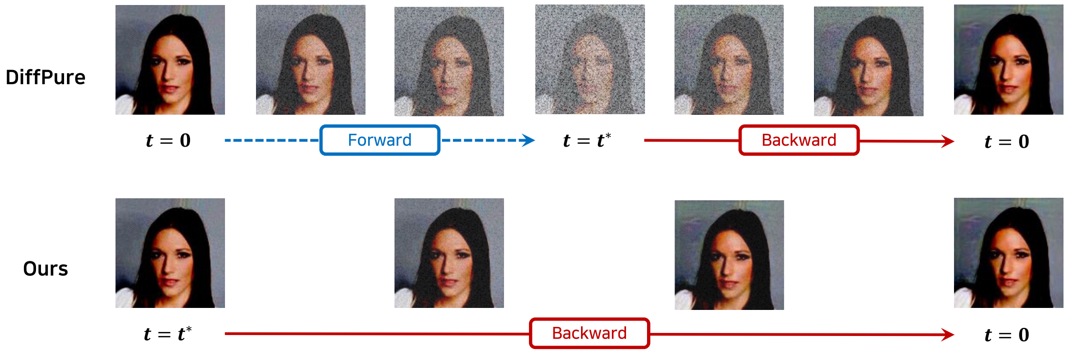


FIGURE 5. The difference between our method and DiffPure [14]. DiffPure adds noise to the image in a forward process up to timestep t^* and then removes it in a reverse process. Ours removes the noise using only the reverse process from timestep t^*.

1) OUR PURIFICATION STRATEGY.

In our purification method, we only employ the reverse process R of DDPM. Given a clean real image x, let x' denote the corresponding disruptive image. We input x' into DDPM, specifying the timestep $t = t^*$ as the starting point of purification, so $x_{t^*} = x'$. In essence, we are assuming that x' is a real image with added little noise, not the complete noise.

$$x_{t-1} = R_{t-1}(x_t) \tag{3}$$

Our reverse process continues until the timestep reaches $t = 0$ with $x_0 = \hat{x}$ where $\hat{x}$ represents the purified image.

$$\hat{x} = \sum_{t=t^*}^{0} R_t(x') \tag{4}$$

2) TECHNICAL DIFFERENCE FROM DIFFPURE

Our purification step is designed by referring to DiffPure. DiffPure also inputs x' into the diffusion model, but it employs both forward and reverse processes, starting from $t = t_0$ where $x_0 = x'$. The objective of DiffPure is to retain only the generalized knowledge of the input image, enabling the classification model to correctly classify the image. To accomplish this, DiffPure adds noise to the image during the forward process from $t = 0$ to $t = t^*$, thereby gradually eliminating the local structures of adversarial examples. However, our goal differs in that we do not generalize the image, but restore it precisely to its original form. According to the work of [15], as more time steps are taken in the forward process and the magnitude of the noise increases, the image restored by the reverse process becomes more different from the original image. This suggests that the more the forward process repeats, the more detail in the original image is destroyed. Therefore, we preserve more detail in the original image than DiffPure by removing the forward process.

C. TIMESTEP FOR PURIFICATION

DiffPure argues that a substantial amount of noise is needed to remove the local structure, leading them to set the timestep t^* to a relatively large value of 300. However, a larger timestep not only introduces more disruption to the image in both the forward and reverse processes, but it also prolongs the process due to an increased number of iterations. As the magnitude of the disruption perturbation is nearly imperceptible to the human eye, there's no need for a large t^*. Consequently, we choose a smaller timestep value of $t^* = 10$ to enhance efficiency and resilience. In this scenario, the DDPM's timestep t_T is set to 1000.

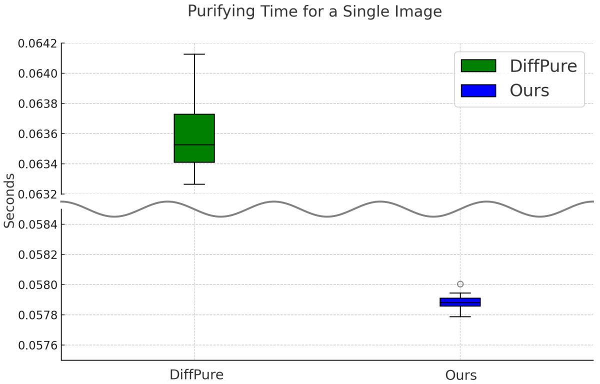


FIGURE 6. Time is taken to purify a single image in Diffpure and Ours. Ours is 0.00572 seconds faster than DiffPure to process one image, achieving a 9.88% average time savings. DiffPure's fastest time is 8.31% slower than Ours' slowest time. We used Group B images in the CelebA dataset.

V. EVALUATION

To evaluate the performance of the proposed framework, we formulated two research questions and conducted experiments to answer them. **RQ1** investigates whether normal deepfake images can be generated from disruptive images through purification (§V-B). **RQ2** verifies whether a training dataset constructed by our purification can uphold the accuracy of the deepfake detector (§V-C).

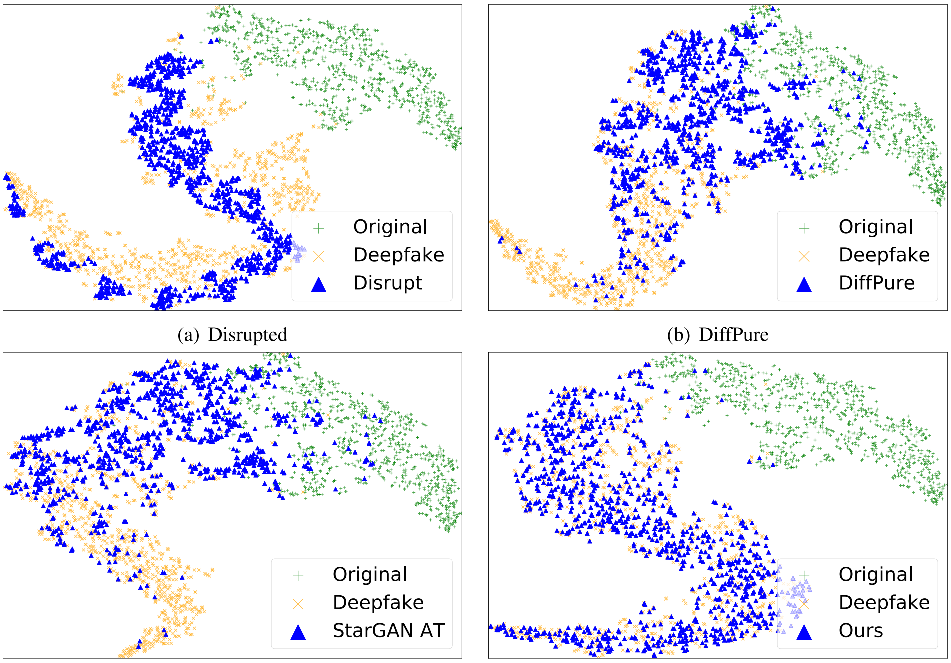
A. EXPERIMENTAL SETTINGS

1) DEEPFAKE DATASET AND DETECTION MODEL

We divide the entire CelebA [26] dataset into three groups. Group A occupies the first 60% of the dataset and is assumed to be the original image dataset initially owned by the defender. All defense methods against disruption are trained with Group A images. Group B occupies the following 30%

TABLE 1. Purification performance of defense methods against the disruption. We measured the distance between the StarGAN output from the disruptive image and the normal deepfake image in the input (L_1 and L_2) and feature spaces (FID). Following the previous work [5], we consider the disruption fails when L_2 distance is greater than or equal to 0.05.

Defense Model	PGD			BIM		
	FID	L_1	L_2	FID	L_1	L_2
No Defense	151.6	0.418	0.267	141.62	0.430	0.278
MagNet Reformer [25]	181.8	0.265	0.118	177.7	0.265	0.118
StarGAN AT [5]	33.5	0.080	0.012	34.1	0.081	0.012
DiffPure [14]	30.6	0.072	0.010	31.56	0.073	0.010
Ours	38.3	0.083	0.013	32.3	0.075	0.010



(a) Disrupted

(b) DiffPure

(c) StarGAN AT

(d) Ours

FIGURE 7. The distribution of defended deepfake images (in blue) against PGD-based disruption. Image features were extracted from the ResNet18 model for t-SNE visualization. The total number of samples in all subfigures is equal, and the data ratio of real images (Original), deepfake images of original real images (Deepfake), and deepfake images of disrupting real images is equal to 1:1:1 (i.e., about 33.3%). It can be seen that the deepfake images generated by 'Ours' in Figure 7-(d) are most similar in distribution to the original deepfake images when compared to the deepfake images generated directly from the disrupting images (Figure 7-(a)) or the images generated by the existing defense techniques 'DiffPure' and 'StarGAN AT' (Figure 7-(b) and Figure 7-(c), respectively).

of the dataset and is used to train the deepfake detection model. We assume that data poisoning based on disruption occurs for this group. Group C occupies the last 10% of the dataset and is used to evaluate the accuracy of detectors. The images in all datasets are cropped to 178 × 178 and subsequently resized to 128 × 128. Our baseline deepfake detection models are the Xception [22] model used in FaceForensics++ [11], and the ResNet18 and ResNet50 [23] models used in the disruption perturbation paper [9].

2) PARAMETERS FOR DISRUPTION

The disrupting perturbation is generated to make the deepfake image get closer to a black image, targeting StarGAN [2]. We assume the grey-box disrupter knows the structure and parameters of the StarGAN, but not those of the defense model. We use a 10-step BIM and a 10-step PGD with $\epsilon = 0.05$ with step size 0.01 for disruption.

3) COMPARISON TARGET MODEL

Our comparison targets are MagNet reformer [25], adversarially trained StarGAN with PGD (StarGAN AT) [5], and DiffPure [14], a SOTA purification technique using DDPM. All defense methods are trained with the Group A dataset.

4) ENVIRONMENTS

We performed all experiments on a single machine Ubuntu 20.04 environment with two NVIDIA RTX 4090 (24GB) GPUs.

TABLE 2. Detection accuracy of detectors trained with the poisoned Group B dataset. Normal real and deepfake images in the Group C dataset are used to measure the accuracy.

Defense Model	Xception		ResNet18		ResNet50	
	PGD	BIM	PGD	BIM	PGD	BIM
No Attack	97.48		98.19		99.04	
No Defense	51.56	49.80	50.15	51.31	50.40	50.05
MagNet Reformer [25]	50.05	50.05	50.05	50.10	54.13	49.95
StarGAN AT [5]	50.05	50.05	49.55	50.00	49.95	50.05
DiffPure [14]	53.53	58.97	53.43	53.83	73.39	85.28
Ours	**83.37**	**95.58**	**89.37**	**99.55**	**90.88**	**96.52**

B. PURIFICATION ABILITY FOR DISRUPTIVE IMAGES

From Table 1, we can see that our L_2 distance is lower than the disruption threshold ($L_2 \geq 0.05$), indicating the successful generation of deepfake images. For the PGD, our performance is better than the MagNet reformer and similar to DiffPure and StarGAN AT. Moreover, our method results in the lowest L_2 distance alongside DiffPure for the BIM. Fig. 7 illustrates the distribution of deepfake images generated by each defense method against PGD-based disruption. We found our method is more effective in terms of the distribution of defended images. Our distribution in Fig. 7 (d) closely resembles the distribution of normal deepfake images. This result is better than DiffPure (Fig. 7 (b)) and StarGAN AT (Fig. 7 (c)) whose distributions are located between the distributions of normal real and deepfake images. We also found our method is faster than DiffPure because of our reverse process alone and fewer timesteps. DiffPure took 0.0636 seconds on average to purify one image, while ours took 0.0572 seconds on average, which is a 9.88% reduction in defense time over DiffPure.

C. ACCURACY OF DETECTION MODELS UNDER THE POISONED DATASET

Table 2 presents the accuracy of defense methods for a deepfake detection model trained by a poisoned dataset with disruptive images. The deepfake detector trained using our approach exhibits the highest detection accuracy among defense methods, regardless of the model structure and disruption method. The MagNet reformer and StarGAN AT still demonstrate severely decreased accuracy across all model structures and disruption methods. DiffPure demonstrates satisfactory accuracy on ResNet50, but it exhibits underwhelming accuracy on Xception and ResNet18 models. MagNet reformer and StarGAN AT exhibited an average drop in accuracy of 48.77%p and 48.87%p respectively from the accuracy without disruption (“No Attack”). DiffPure showed a slight improvement with a drop of 43.32%p. Our method achieved an average accuracy drop of 7.83%p, and notably, the accuracy against BIM in ResNet18 was even higher than the “No Attack” scenario.

VI. DISCUSSION

A. TARGETED DEEPFAKE GENERATION AND DETECTION MODELS.

In this paper, we focused on StarGAN as the target deepfake generation model, thereby excluding other deepfake generation techniques such as faceswap. DDPM used in our purification is not trained for a specific disruption method or target model. Therefore, our method is agnostic to disruption methods or deepfake generation methods. Nevertheless, experimental verification is necessary to confirm our capability. Furthermore, in this paper, the experiments were conducted specifically on spatial-based deepfake detection techniques using a DNN model. Additional experiments are required for other detection methods, including frequency-based and biological-based ones. We leave the extension of our method as future work.

B. POISONING RATE IN TRAINING DATASET

In this paper, we assumed a worst-case scenario, setting the poisoning rate of the training dataset to 100%. Achieving this level of data poisoning is challenging, often limited by the attacker’s knowledge of the targeted deepfake generator and the attacker’s capabilities. Nevertheless, from a defensive perspective, it is essential to consider such scenarios to strengthen defense mechanisms. Moreover, as with the traditional issue, the deepfake issue will become more applicable as the defense techniques improve. It is not impossible to have a 100% poisoning rate if defense techniques are applied to all devices with generation suppression, just like firewalls are applied to all computers today. Therefore, in this paper, we conducted experiments under the fully poisoned training data where the attacker acquires the knowledge of the deepfake generation model.

VII. RELATED WORK

A. DEEPFAKE GENERATION

Various deepfake generation techniques exist for face generation [3], face conversion [4], [118], attribute manipulation [2], [119], and expression conversion [120]. Among them, StarGAN [2], a prominent technique for modifying attributes such as gender and age in faces, has been extensively used

TABLE 3. Deepfake disruption papers: Dataset, Target Model, and Evaluation Metric. In the GitHub row, 'O' means Official GitHub link, and 'U' means Unofficial GitHub link.

	Paper	Shamshad, et al. [27]	Van, et al. [28]	Zhong, et al. [29]	Li, et al. [30]	Zhai, et al. [31]	Sun, et al. [32]	Tang, et al. [33]	Shim, et al. [34]	Ruiz, et al. [35]	Salman, et al. [36]	Dong, et al. [37]	Li, et al. [38]	Huang, et al. [39]	He, et al. [40]	Guan, et al. [41]	Yeh, et al. [42]	Dong, et al. [43]	Ruiz, et al. [5]	Sun, et al. [44]	Ruiz, et al. [45]
	Year	23	23	23	23	23	23	23	23	23	23	23	23	22	22	22	21	21	20	20	20
	Github	O	O								O		O	O			O		O		
Dataset	CelebA [26]			✓	✓	✓		✓		✓		✓	✓	✓			✓		✓		✓
	CelebA-HQ [46]	✓	✓	✓					✓						✓						
	CelebAMask-HQ [41]															✓					
	Celeb-DF [47]						✓													✓	
	LFW [48]	✓						✓						✓							
	LFWA [48]					✓															
	FF++ [11]							✓													
	Stable Diffusion [49]										✓										
	FFHQ [3]												✓								
	LADN [50]	✓																			
	Face Scrub [51]																	✓			
	VGG Face2 [52]		✓																		
	Horse2Zebra [53]																		✓		
	Photograph2Monet [53]																		✓		
	CityScapes [54]																		✓		
Target Model	StarGAN [2]			✓	✓	✓		✓		✓		✓		✓	✓	✓			✓		✓
	StarGAN2 [55]			✓																	
	AttGAN [56]			✓		✓		✓				✓		✓		✓					
	GANimation [57]									✓					✓	✓			✓		✓
	Celeb-DF [47]						✓													✓	
	StyleGAN1 [3]												✓								
	StyleGAN2 [58]	✓		✓									✓								
	StyleGAN3 [59]			✓																	
	AG-GAN [60]					✓		✓						✓							
	HiSD [61]					✓		✓						✓							
	CycleGAN [53]			✓													✓		✓		
	StyleSwin [62]			✓																	
	DDGAN [63]			✓																	
	CUT [64]			✓																	
	SimSwap [65]								✓						✓	✓					
	StyleCLIP [66]								✓												
	Diffusion AE [67]								✓												
	Icface [68]								✓												
	FaceSwap [69]											✓									
	DCGAN [70]												✓								
	WGAN [71]												✓								
	FaceShifter [72]															✓					
	Self-AE																	✓			
	Stable Diffusion [49]		✓								✓										
	pix2pixHD [73]																		✓		
Evaluation Metric	l-norm			✓	✓			✓	✓	✓			✓	✓	✓	✓			✓		✓
	SSIM [74]					✓	✓	✓				✓	✓					✓		✓	
	FSIM [75]										✓	✓						✓			
	PSNR	✓				✓		✓			✓		✓								
	FID [76]	✓		✓		✓					✓			✓							
	ASR (DSR)							✓	✓	✓							✓		✓		✓
	Bit Acc			✓			✓														
	Query Count									✓							✓				✓
	BRISQUE [8]		✓									✓			✓			✓			
	Detection Acc						✓														
	TFHC							✓						✓							
	ACS							✓						✓							
	PR [77]										✓										
	VIFP [78]										✓										
	Matching Rate												✓								
	MCD														✓						
	Nima [79]														✓						
	DBCNN [80]														✓						
	TV [81]														✓						
	ID Sim (PSR, ISM)	✓	✓												✓						
	Perceptual Loss															✓					
	NME [82]														✓					✓	
	DFDR																				
	SER-FQA [83]		✓																		
	Identity Loss [84]								✓												
	LPIPS [85]								✓												
	FDFR		✓																		

TABLE 4. Deepfake detection papers: train dataset, test dataset, and evaluation metric. In the GitHub row, 'O' means the official GitHub link, and 'U' means the unofficial GitHub link. 'S', 'F', and 'B' in the Method column stand for Spatial, Frequency, and Biological detection methods, respectively.

	Paper	Wang et al. [86]	Dong et al. [87]	Ricker et al. [88]	Chen et al. [89]	Zhuang et al. [90]	Chen et al. [91]	Dong et al. [92]	Shiohara et al. [93]	Cao et al. [94]	Zheng et al. [95]	Zhao et al. [96]	Haliassos et al. [97]	Chandrasegaran et al. [98]	Luo et al. [99]	Li et al. [100]	Zhao et al. [101]	Cozzolino et al. [102]	Li et al. [103]
	Year	23	23	22	22	22	22	22	22	22	21	21	21	21	21	21	21	21	20
	Method	S	S	F	S	S	S	S	S	S	S	S	B	F	F	F	S	S	S
	Github	O	O	O	O	O	O	O	O	O	O	U	O	O	O	U	O	O	U)
	Pretrain	✓	✓	✓		✓			✓				✓	✓			✓		✓
Train Dataset	FF++ [11]	✓	✓		✓	✓	✓		✓	✓	✓	✓	✓		✓	✓	✓		✓
	SBI [93]		✓																
	MS-Celeb [104]							✓											
	CDF-V2 [47]											✓	✓						
	DFDC-P [105]																		
	DFDC [106]											✓	✓						
	DF												✓						
	CelebA [26]													✓					
	CelebA-HQ [46]													✓					
	LSUN-Bedroom [107]			✓										✓					
	VoxCeleb2 [108]																	✓	
Test Dataset	FF++-HQ [11]		✓			✓	✓	✓	✓	✓	✓	✓	✓		✓	✓	✓		✓
	FF++-LQ [11]						✓										✓		
	CDF-V1 [47]	✓	✓			✓	✓	✓	✓	✓	✓	✓	✓		✓		✓	✓	
	CDF-V2 [47]					✓		✓				✓							
	DFD [109]	✓			✓	✓		✓	✓						✓				✓
	FSH [110]										✓		✓						
	DeeperForensics [111]	✓			✓		✓	✓			✓	✓	✓		✓				
	DFDC [106]		✓		✓		✓		✓		✓		✓		✓				✓
	DFDC-P [105]					✓			✓			✓						✓	
	FFIW [112]								✓										
	Wild [113]									✓									
	Celeb-DF [47]																		✓
	CelebA [26]													✓					
	CelebA-HQ [46]													✓					
	ImageNet [114]			✓															
	FFHQ [3]			✓															
	LSUN-Cat [107]			✓															
	LSUN-Horse [107]			✓															
	LSUN-Bedroom [107]			✓										✓					
	VoxCeleb2 [108]																	✓	
Evaluation Metric	ACC				✓	✓				✓				✓		✓	✓	✓	
	AUC	✓	✓	✓	✓	✓	✓	✓	✓	✓		✓	✓		✓	✓	✓	✓	✓
	Time Consumptions				✓														
	EER				✓					✓									✓
	Feature Visualization								✓	✓	✓					✓			
	Forgery Localizaiton					✓		✓	✓	✓	✓				✓		✓		
	A.P.																		✓
	LogLoss									✓							✓		
	pAUC [115]															✓			
	MSL [116]																	✓	
	Triplet [117]																	✓	

for evaluating deepfake detection [9], [10], [121], [122] and disruption methods [5], [9], [121], [123]. Therefore, in this paper, we have chosen StarGAN as the target generator to assess the issues arising when disruption and detection coexist.

B. DEEPFAKE DISRUPTION

The early works on deepfake disruption [124] utilized adversarial attacks [6], [7], [8] as their core strategy. Ruiz et al. [5] added a disrupting perturbation to the real image along with the gradient of the generative model toward a distorted deepfake image. Some other works utilize perturbation generators which are also generative models to disrupt deepfake generators [9], [125]. While previous deepfake disruption defenses were primarily conducted in white-box environments [5], [40], recent research has shifted towards more realistic scenarios akin to grey-box and black-box settings. Research is being made to achieve disruption even when the model's structure is unknown, and the model's output is either partially known [35], [42], [45] or completely unknown [37]. To do this, they made sure that disruption works well in situations where the number of queries is limited. Another emerging trend in recent research emphasizes transitivity, to ensure that disruption is not confined to a specific type of deepfake generation [41]. As deepfake generation models have traditionally been based on GANs, the advent of diffusion models has prompted research into extending disruption techniques to these as well [28], [34], [36]. Concurrently, there is ongoing research on nullification [42] and warning watermarking [31], which aim to do more than just suppress deepfakes; they seek to prevent any alteration of the image in its entirety. To synthesize the research landscape more clearly, we present Table 3, which details the datasets, target models, and evaluation metrics used in each study.

C. DEEPFAKE DETECTION

Deepfake detection methods can be broadly categorized into three types: spatial-based, frequency-based, and biological-signal-based. Spatial-based detection identifies visual artifacts in deepfake images [11], [122]. Frequency-based detection uncovers artifacts in deepfake images within the frequency domain [12], [121]. Biological- signal-based detection analyzes natural biological signals exclusive to real faces [13], [126].

Historically, deepfake detection models have demonstrated high detection rates within the same datasets they were trained on (intra-datasets), but their performance significantly drops when applied to new, unseen datasets (cross-datasets). To address this, recent studies [89], [90] have focused on enhancing the generalization capabilities of detection models to maintain high detection rates across various datasets. The detection models were initially designed to identify deepfakes generated by Generative Adversarial Networks (GANs). With the emergence of diffusion models [15], a novel approach to deepfake creation [127], the scope of research has broadened to include the detection of deepfakes originating from these sophisticated diffusion techniques [88]. For a comprehensive summary of the datasets used for training and testing, as well as the evaluation metrics employed in each study, refer to Table.4

The coexistence of deepfake detection and disruption is likely, as they operate at different stages of the deepfake ecosystem. Wang et al. raised concerns that disrupted images could potentially fool detection models [9]. While their work explored the interplay between detection and disruption, our focus is on the unintended consequences that disruption may have on the training of detection models.

D. ADVERSARIAL PURIFICATION

Generative models are frequently employed for adversarial purification. MagNet [25], for instance, uses an autoencoder to learn a manifold of normal images. Images situated far from the learned manifold are rejected, and those close to the manifold are purified. DiffPure harnesses a diffusion model for adversarial purification due to its robust performance in image generation and noise reduction [14]. We also utilized DDPM [15], a standard diffusion model. However, unlike previous approaches, we omitted the forward process and used a smaller timestep.

VIII. CONCLUSION

We revisited two countermeasures, deepfake disruption, and detection, from the perspective of traditional security issues. We found that, due to the coexistence of the two countermeasures, deepfake disruption may cause data poisoning problems for deepfake detection models during their training phase. To solve this problem, we propose a robust training framework for deepfake detection models. Our framework is tailored to address the training data poisoning problem in deepfake detection models, the first of its kind in academia. Using the DDPM [15] denoising model, it minimizes image distortion and cuts defense time by eliminating extra noise introduced by DiffPure [14]. Our approach generates deepfake images that better resemble normal ones compared to those created by DiffPure or StarGAN Adversarial Training [5]. Moreover, our method achieves a 7.75% defense time reduction compared to DiffPure, and when applied to training deepfake detection models, it outperforms StarGAN AT and DiffPure in detection accuracy on Xception [22], ResNet18 [23], and ResNet50 [23] under PGD [8] and BIM [7] disruption attacks.

ACKNOWLEDGMENT

An earlier version of this paper was presented in part at the Pre-Proceeding of the 24th World Conference on Information Security Applications (WISA 2023) [DOI: 10.1007/978-981-99-8024-6_14].

REFERENCES

[1] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," 2014, *arXiv:1406.2661*.

[2] 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 Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8789–8797.

[3] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 4396–4405.

[4] Y. Nirkin, Y. Keller, and T. Hassner, "FSGAN: Subject agnostic face swapping and reenactment," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 7183–7192.

[5] N. Ruiz, S. A. Bargal, and S. Sclaroff, "Disrupting Deepfakes: Adversarial attacks against conditional image translation networks and facial manipulation systems," in *Proc. Eur. Conf. Comput. Vis.*, Aug. 2020, pp. 236–251.

[6] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," 2014, *arXiv:1412.6572*.

[7] K. Lee, J. Kim, S. Chong, and J. Shin, "Making stochastic neural networks from deterministic ones," School Elect. Eng., Korea Adv. Inst. Sci. Technol. (KAIST), Republic of Korea, Tech. Rep., 2017.

[8] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," 2017, *arXiv:1706.06083*.

[9] X. Wang, J. Huang, S. Ma, S. Nepal, and C. Xu, "DeepFake Disrupter: The detector of Deepfake is my friend," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 14900–14909.

[10] B. Liu, F. Yang, X. Bi, B. Xiao, W. Li, and X. Gao, "Detecting generated images by real images," in *Proc. Eur. Conf. Comput. Vis.*, Cham, Switzerland: Springer, Oct. 2022, pp. 95–110.

[11] A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Niessner, "FaceForensics++: Learning to detect manipulated facial images," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 1–11.

[12] X. Zhang, S. Karaman, and S.-F. Chang, "Detecting and simulating artifacts in GAN fake images," in *Proc. IEEE Int. Workshop Inf. Forensics Secur. (WIFS)*, Dec. 2019, pp. 1–6.

[13] U. A. Ciftci, I. Demir, and L. Yin, "FakeCatcher: Detection of synthetic portrait videos using biological signals," *IEEE Trans. Pattern Anal. Mach. Intell.*, p. 1, Jul. 2020.

[14] W. Nie, B. Guo, Y. Huang, C. Xiao, A. Vahdat, and A. Anandkumar, "Diffusion models for adversarial purification," in *Proc. Int. Conf. Mach. Learn.*, Jun. 2022, pp. 16805–16827.

[15] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, Eds. Red Hook, NY, USA: Curran Associates, 2020, pp. 6840–6851. [Online]. Available: https://proceedings.neurips.cc/paper/2020/file/4c5bcfec8584af0d967f1ab10179ca4b-Paper.pdf

[16] M. Peyravian, A. Roginsky, and N. Zunic, "Methods for preventing unauthorized software distribution," *Comput. Secur.*, vol. 22, no. 4, pp. 316–321, May 2003.

[17] M. J. Morgan and D. J. Ruskell, "Software piracy—The problems," *Ind. Manag. Data Syst.*, vol. 87, nos. 3–4, pp. 8–12, Mar. 1987.

[18] J. Herbert and A. Litchfield, "A novel method for decentralised peer-to-peer software license validation using cryptocurrency blockchain technology," in *Proc. 38th Australas. Comput. Sci. Conf. (ACSC)*, vol. 27, Jan. 2015, p. 30.

[19] S. Ioannidis, A. D. Keromytis, S. M. Bellovin, and J. M. Smith, "Implementing a distributed firewall," in *Proc. 7th ACM Conf. Comput. Commun. Secur.*, Nov. 2000, pp. 190–199.

[20] V. M. Potdar, S. Han, and E. Chang, "A survey of digital image watermarking techniques," in *Proc. INDIN 05. 3rd IEEE Int. Conf. Ind. Informat.*, Aug. 2005, pp. 709–716.

[21] Y. Mirsky, T. Doitshman, Y. Elovici, and A. Shabtai, "Kitsune: An ensemble of autoencoders for online network intrusion detection," 2018, *arXiv:1802.09089*.

[22] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1800–1807.

[23] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[24] Y. Li and S. Lyu, "Exposing DeepFake videos by detecting face warping artifacts," 2018, *arXiv:1811.00656*.

[25] D. Meng and H. Chen, "MagNet: A two-pronged defense against adversarial examples," in *Proc. ACM SIGSAC Conf. Comput. Commun. Secur.*, Oct. 2017, pp. 135–147.

[26] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 3730–3738.

[27] F. Shamshad, M. Naseer, and K. Nandakumar, "CLIP2Protect: Protecting facial privacy using text-guided makeup via adversarial latent search," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 20595–20605.

[28] T. Van Le, H. Phung, T. H. Nguyen, Q. Dao, N. N. Tran, and A. Tran, "Anti-DreamBooth: Protecting users from personalized text-to-image synthesis," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2023, pp. 2116–2127.

[29] H. Zhong, J. Chang, Z. Yang, T. Wu, P. C. M. Arachchige, C. Pathmabandu, and M. Xue, "Copyright protection and accountability of generative AI: Attack, watermarking and attribution," in *Proc. Companion ACM Web Conf.*, Apr. 2023, pp. 94–98.

[30] Q. Li, M. Gao, G. Zhang, and W. Zhai, "Defending DeepFakes by saliency-aware attack," *IEEE Trans. Computat. Social Syst.*, pp. 1–8, May 2023.

[31] R. Zhai, R. Ni, Y. Chen, Y. Yu, and Y. Zhao, "Defending fake via warning: Universal proactive defense against face manipulation," *IEEE Signal Process. Lett.*, vol. 30, pp. 1072–1076, 2023.

[32] P. Sun, H. Qi, Y. Li, and S. Lyu, "FakeTracer: Proactively defending against face-swap DeepFakes via implanting traces in training," 2023, *arXiv:2307.14593*.

[33] L. Tang, D. Ye, Z. Lu, Y. Zhang, S. Hu, Y. Xu, and C. Chen, "Feature extraction matters more: Universal deepfake disruption through attacking ensemble feature extractors," 2023, *arXiv:2303.00200*.

[34] J. Shim and H. Yoon, "LEAT: Towards robust Deepfake disruption in real-world scenarios via latent ensemble attack," 2023, *arXiv:2307.01520*.

[35] N. Ruiz, S. A. Bargal, C. Xie, and S. Sclaroff, "Practical disruption of image translation DeepFake networks," in *Proc. Conf. Artif. Intell. (AAAI)*, vol. 37, Jun. 2023, pp. 14478–14486.

[36] H. Salman, A. Khaddaj, G. Leclerc, A. Ilyas, and A. Madry, "Raising the cost of malicious AI-powered image editing," 2023, *arXiv:2302.06588*.

[37] J. Dong, Y. Wang, J. Lai, and X. Xie, "Restricted black-box adversarial attack against Deepfake face swapping," *IEEE Trans. Inf. Forensics Security*, vol. 18, pp. 2596–2608, 2023.

[38] Z. Li, N. Yu, A. Salem, M. Backes, M. Fritz, and Y. Zhang, "UnGANable: Defending against GAN-based face manipulation," in *Proc. 32nd USENIX Secur. Symp. (USENIX)*, Aug. 2023, pp. 7213–7230.

[39] H. Huang, Y. Wang, Z. Chen, Y. Zhang, Y. Li, Z. Tang, W. Chu, J. Chen, W. Lin, and K. K. Ma, "CMUA-watermark: A cross-model universal adversarial watermark for combating Deepfakes," in *Proc. AAAI Conf. Artif. Intell.*, vol. 36, no. 1, 2022, pp. 989–997.

[40] Z. He, W. Wang, W. Guan, J. Dong, and T. Tan, "Defeating Deepfakes via adversarial visual reconstruction," in *Proc. 30th ACM Int. Conf. Multimedia*, Oct. 2022, pp. 2464–2472.

[41] W. Guan, Z. He, W. Wang, J. Dong, and B. Peng, "Defending against Deepfakes with ensemble adversarial perturbation," in *Proc. 26th Int. Conf. Pattern Recognit. (ICPR)*, Aug. 2022, pp. 1952–1958.

[42] C.-Y. Yeh, H.-W. Chen, H.-H. Shuai, D.-N. Yang, and M.-S. Chen, "Attack as the best defense: Nullifying image-to-image translation GANs via limit-aware adversarial attack," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 16168–16177.

[43] J. Dong and X. Xie, "Visually maintained image disturbance against Deepfake face swapping," in *Proc. IEEE Int. Conf. Multimedia Expo. (ICME)*, Jul. 2021, pp. 1–6.

[44] P. Sun, Y. Li, H. Qi, and S. Lyu, "Landmark breaker: Obstructing Deepfake by disturbing landmark extraction," in *Proc. IEEE Int. Workshop Inf. Forensics Secur. (WIFS)*, Dec. 2020, pp. 1–6.

[45] N. Ruiz, S. A. Bargal, and S. Sclaroff, "Protecting against image translation Deepfakes by leaking universal perturbations from black-box neural networks," 2020, *arXiv:2006.06493*.

[46] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive growing of GANs for improved quality, stability, and variation," 2017, *arXiv:1710.10196*.

[47] Y. Li, X. Yang, P. Sun, H. Qi, and S. Lyu, "Celeb-DF: A large-scale challenging dataset for Deepfake forensics," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 3204–3213.

[48] L. J. Karam and T. Zhu, "Quality labeled faces in the wild (QLFW): A database for studying face recognition in real-world environments," in *Proc. Human Vis. Electron. Imag. XX*, Mar. 2015, pp. 87–96.

[49] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 10674–10685.

[50] Q. Gu, G. Wang, M. T. Chiu, Y.-W. Tai, and C.-K. Tang, "LADN: Local adversarial disentangling network for facial makeup and de-makeup," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 10480–10489.

[51] H.-W. Ng and S. Winkler, "A data-driven approach to cleaning large face datasets," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2014, pp. 343–347.

[52] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman, "VGGFace2: A dataset for recognising faces across pose and age," in *Proc. 13th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, May 2018, pp. 67–74.

[53] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2242–2251.

[54] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 3213–3223.

[55] Y. Choi, Y. Uh, J. Yoo, and J.-W. Ha, "StarGAN V2: Diverse image synthesis for multiple domains," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 8185–8194.

[56] Z. He, W. Zuo, M. Kan, S. Shan, and X. Chen, "AttGAN: Facial attribute editing by only changing what you want," *IEEE Trans. Image Process.*, vol. 28, no. 11, pp. 5464–5478, Nov. 2019.

[57] A. Pumarola, A. Agudo, A. M. Martinez, A. Sanfeliu, and F. Moreno-Noguer, "Ganimation: Anatomically-aware facial animation from a single image," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 818–833.

[58] Y. Viazovetskyi, V. Ivashkin, and E. Kashin, "Stylegan2 distillation for feed-forward image manipulation," in *Proc. Eur. Conf. Comput. Vis.*, Cham, Switzerland: Springer, Aug. 2020, pp. 170–186.

[59] T. Karras, M. Aittala, S. Laine, E. Härkönen, J. Hellsten, J. Lehtinen, and T. Aila, "Alias-free generative adversarial networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, Dec. 2021, pp. 852–863.

[60] H. Tang, D. Xu, N. Sebe, and Y. Yan, "Attention-guided generative adversarial networks for unsupervised image-to-image translation," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2019, pp. 1–8.

[61] X. Li, S. Zhang, J. Hu, L. Cao, X. Hong, X. Mao, F. Huang, Y. Wu, and R. Ji, "Image-to-image translation via hierarchical style disentanglement," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 8635–8644.

[62] B. Zhang, S. Gu, B. Zhang, J. Bao, D. Chen, F. Wen, Y. Wang, and B. Guo, "StyleSwin: Transformer-based GAN for high-resolution image generation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 11294–11304.

[63] Z. Xiao, K. Kreis, and A. Vahdat, "Tackling the generative learning trilemma with denoising diffusion GANs," 2021, *arXiv:2112.07804*.

[64] P. Taesung, A. A. Efros, R. Zhang, and J. Y. Zhu, "Contrastive learning for unpaired image-to-image translation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Aug. 2020, pp. 319–345.

[65] R. Chen, X. Chen, B. Ni, and Y. Ge, "SimSwap: An efficient framework for high fidelity face swapping," in *Proc. 28th ACM Int. Conf. Multimedia*, Oct. 2020, pp. 2003–2011.

[66] O. Patashnik, Z. Wu, E. Shechtman, D. Cohen-Or, and D. Lischinski, "StyleCLIP: Text-driven manipulation of StyleGAN imagery," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 2065–2074.

[67] K. Preechakul, N. Chatthee, S. Wizadwongsa, and S. Suwajanakorn, "Diffusion autoencoders: Toward a meaningful and decodable representation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 10619–10629.

[68] S. Tripathy, J. Kannala, and E. Rahtu, "ICface: Interpretable and controllable face reenactment using GANs," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2020, pp. 3374–3383.

[69] P. Korshunov and S. Marcel, "DeepFakes: A new threat to face recognition? Assessment and detection," 2018, *arXiv:1812.08685*.

[70] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," 2015, *arXiv:1511.06434*.

[71] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein generative adversarial networks," in *Proc. Int. Conf. Mach. Learn.*, Jul. 2017, pp. 214–223.

[72] L. Li, J. Bao, H. Yang, D. Chen, and F. Wen, "FaceShifter: Towards high fidelity and occlusion aware face swapping," 2019, *arXiv:1912.13457*.

[73] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, A. Tao, J. Kautz, and B. Catanzaro, "High-resolution image synthesis and semantic manipulation with conditional GANs," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8798–8807.

[74] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.

[75] L. Zhang, L. Zhang, X. Mou, and D. Zhang, "FSIM: A feature similarity index for image quality assessment," *IEEE Trans. Image Process.*, vol. 20, no. 8, pp. 2378–2386, Aug. 2011.

[76] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "Gans trained by a two time-scale update rule converge to a local nash equilibrium," in *Proc. Adv. Neural Inf. Process. Syst.*, Jun. 2017, pp. 25–34.

[77] M. S. M. Sajjadi, O. Bachem, M. Lucic, O. Bousquet, and S. Gelly, "Assessing generative models via precision and recall," in *Proc. Adv. Neural Inf. Process. Syst.*, May 2018, pp. 1–15.

[78] H. R. Sheikh and A. C. Bovik, "Image information and visual quality," *IEEE Trans. Image Process.*, vol. 15, no. 2, pp. 430–444, Feb. 2006.

[79] H. Talebi and P. Milanfar, "NIMA: Neural image assessment," *IEEE Trans. Image Process.*, vol. 27, no. 8, pp. 3998–4011, Aug. 2018.

[80] W. Zhang, K. Ma, J. Yan, D. Deng, and Z. Wang, "Blind image quality assessment using a deep bilinear convolutional neural network," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 1, pp. 36–47, Jan. 2020.

[81] L. I. Rudin, S. Osher, and E. Fatemi, "Nonlinear total variation based noise removal algorithms," *Phys. D, Nonlinear Phenomena*, vol. 60, nos. 1–4, pp. 259–268, Nov. 1992.

[82] A. Jourabloo and X. Liu, "Pose-invariant 3D face alignment," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 3694–3702.

[83] P. Terhörst, J. N. Kolf, N. Damer, F. Kirchbuchner, and A. Kuijper, "SER-FIQ: Unsupervised estimation of face image quality based on stochastic embedding robustness," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 5650–5659.

[84] E. Richardson, Y. Alaluf, O. Patashnik, Y. Nitzan, Y. Azar, S. Shapiro, and D. Cohen-Or, "Encoding in style: A styleGAN encoder for image-to-image translation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 2287–2296.

[85] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 586–595.

[86] Z. Wang, J. Bao, W. Zhou, W. Wang, and H. Li, "AltFreezing for more general video face forgery detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 4129–4138.

[87] S. Dong, J. Wang, R. Ji, J. Liang, H. Fan, and Z. Ge, "Implicit identity leakage: The stumbling block to improving Deepfake detection generalization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 3994–4004.

[88] J. Ricker, S. Damm, T. Holz, and A. Fischer, "Towards the detection of diffusion model Deepfakes," 2022, *arXiv:2210.14571*.

[89] L. Chen, Y. Zhang, Y. Song, J. Wang, and L. Liu, "OST: Improving generalization of Deepfake detection via one-shot test-time training," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 35, Dec. 2022, pp. 24597–24610.

[90] W. Zhuang, Q. Chu, Z. Tan, Q. Liu, H. Yuan, C. Miao, Z. Luo, and N. Yu, "UIA-ViT: Unsupervised inconsistency-aware method based on vision transformer for face forgery detection," in *Proc. Eur. Conf. Comput. Vis.*, Cham, Switzerland: Springer, Oct. 2022, pp. 391–407.

[91] L. Chen, Y. Zhang, Y. Song, L. Liu, and J. Wang, "Self-supervised learning of adversarial example: Towards good generalizations for Deepfake detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 18689–18698.

[92] X. Dong, J. Bao, D. Chen, T. Zhang, W. Zhang, N. Yu, D. Chen, F. Wen, and B. Guo, "Protecting celebrities from Deepfake with identity consistency transformer," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 9458–9468.

[93] K. Shiohara and T. Yamasaki, "Detecting Deepfakes with self-blended images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 18699–18708.

[94] J. Cao, C. Ma, T. Yao, S. Chen, S. Ding, and X. Yang, "End-to-end reconstruction-classification learning for face forgery detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 4103–4112.

[95] Y. Zheng, J. Bao, D. Chen, M. Zeng, and F. Wen, "Exploring temporal coherence for more general video face forgery detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 15024–15034.

[96] T. Zhao, X. Xu, M. Xu, H. Ding, Y. Xiong, and W. Xia, "Learning self-consistency for Deepfake detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 15003–15013.

[97] A. Haliassos, K. Vougioukas, S. Petridis, and M. Pantic, "Lips Don't lie: A generalisable and robust approach to face forgery detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 5037–5047.

[98] K. Chandrasegaran, N.-T. Tran, and N.-M. Cheung, "A closer look at Fourier spectrum discrepancies for CNN-generated images detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 7196–7205.

[99] Y. Luo, Y. Zhang, J. Yan, and W. Liu, "Generalizing face forgery detection with high-frequency features," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 16312–16321.

[100] J. Li, H. Xie, J. Li, Z. Wang, and Y. Zhang, "Frequency-aware discriminative feature learning supervised by single-center loss for face forgery detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 6454–6463.

[101] H. Zhao, T. Wei, W. Zhou, W. Zhang, D. Chen, and N. Yu, "Multi-attentional Deepfake detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 2185–2194.

[102] D. Cozzolino, A. Rössler, J. Thies, M. Nießner, and L. Verdoliva, "ID-reveal: Identity-aware Deepfake video detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 15088–15097.

[103] L. Li, J. Bao, T. Zhang, H. Yang, D. Chen, F. Wen, and B. Guo, "Face X-ray for more general face forgery detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 5000–5009.

[104] Y. Guo, L. Zhang, Y. Hu, X. He, and J. Gao, "MS-Celeb-1M: A dataset and benchmark for large-scale face recognition," in *Proc. Eur. Conf. Comput. Vis.*, Oct. 2016, pp. 87–102.

[105] B. Dolhansky, R. Howes, B. Pflaum, N. Baram, and C. C. Ferrer, "The Deepfake detection challenge (DFDC) preview dataset," 2019, *arXiv:1910.08854*.

[106] B. Dolhansky, J. Bitton, B. Pflaum, J. Lu, R. Howes, M. Wang, and C. Canton Ferrer, "The Deepfake detection challenge (DFDC) dataset," 2020, *arXiv:2006.07397*.

[107] F. Yu, A. Seff, Y. Zhang, S. Song, T. Funkhouser, and J. Xiao, "LSUN: Construction of a large-scale image dataset using deep learning with humans in the loop," 2015, *arXiv:1506.03365*.

[108] J. Son Chung, A. Nagrani, and A. Zisserman, "VoxCeleb2: Deep speaker recognition," 2018, *arXiv:1806.05622*.

[109] *Deep Fake Detection Dataset*. Accessed: Oct. 29, 2019. [Online]. Available: https://ai.googleblog.com/2019/09/contributing-data-to-deepfake-detection.html

[110] L. Li, J. Bao, H. Yang, D. Chen, and F. Wen, "Advancing high fidelity identity swapping for forgery detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 5073–5082.

[111] L. Jiang, R. Li, W. Wu, C. Qian, and C. C. Loy, "DeeperForensics-1.0: A large-scale dataset for real-world face forgery detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 2886–2895.

[112] T. Zhou, W. Wang, Z. Liang, and J. Shen, "Face forensics in the wild," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 5774–5784.

[113] B. Zi, M. Chang, J. Chen, X. Ma, and Y.-G. Jiang, "WildDeepfake: A challenging real-world dataset for Deepfake detection," in *Proc. 28th ACM Int. Conf. Multimedia*, Oct. 2020, pp. 2382–2390.

[114] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, Dec. 2015.

[115] D. K. McClish, "Analyzing a portion of the ROC curve," *Med. Decis. Making*, vol. 9, no. 3, pp. 190–195, Aug. 1989.

[116] X. Wang, X. Han, W. Huang, D. Dong, and M. R. Scott, "Multi-similarity loss with general pair weighting for deep metric learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 5017–5025.

[117] E. Hoffer and N. Ailon, "Deep metric learning using triplet network," in *Proc. Int. Workshop Similarity-Based Pattern Recognit.*, Nov. 2015, pp. 84–92.

[118] Y. Zhu, Q. Li, J. Wang, C. Xu, and Z. Sun, "One shot face swapping on megapixels," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 4832–4842.

[119] W. Xia, Y. Yang, J.-H. Xue, and B. Wu, "TediGAN: Text-guided diverse face image generation and manipulation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 2256–2265.

[120] G.-S. Hsu, C.-H. Tsai, and H.-Y. Wu, "Dual-generator face reenactment," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 632–640.

[121] V. Asnani, X. Yin, T. Hassner, S. Liu, and X. Liu, "Proactive image manipulation detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 15365–15374.

[122] N. Hulzebosch, S. Ibrahimi, and M. Worring, "Detecting CNN-generated facial images in real-world scenarios," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2020, pp. 2729–2738.

[123] R. Wang, Z. Huang, Z. Chen, L. Liu, J. Chen, and L. Wang, "Anti-forgery: Towards a stealthy and robust DeepFake disruption attack via adversarial perceptual-aware perturbations," 2022, *arXiv:2206.00477*.

[124] C.-Y. Yeh, H.-W. Chen, S.-L. Tsai, and S.-D. Wang, "Disrupting image-translation-based DeepFake algorithms with adversarial attacks," in *Proc. IEEE Winter Appl. Comput. Vis. Workshops (WACVW)*, Mar. 2020, pp. 53–62.

[125] Q. Huang, J. Zhang, W. Zhou, W. Zhang, and N. Yu, "Initiative defense against facial manipulation," in *Proc. AAAI Conf. Artif. Intell.*, vol. 35, no. 2, 2021, pp. 1619–1627.

[126] S. Hu, Y. Li, and S. Lyu, "Exposing GAN-generated faces using inconsistent corneal specular highlights," in *Proc. ICASSP - IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Jun. 2021, pp. 2500–2504.

[127] K. Kim, Y. Kim, S. Cho, J. Seo, J. Nam, K. Lee, S. Kim, and K. Lee, "DiffFace: Diffusion-based face swapping with facial guidance," 2022, *arXiv:2212.13344*.

JAEWOO PARK received the B.S. degree in electronics engineering from Kyungpook National University, Daegu, South Korea, in 2022. He is currently pursuing the M.S. degree in information security with Yonsei University, Seoul, South Korea. His research interests include adversarial robustness, generative adversarial networks, diffusion models, testing deep neural networks, and adversarial machine learning.

LEO HYUN PARK (Graduate Student Member, IEEE) received the B.S. degree in computer engineering from Kwangwoon University, Seoul, South Korea, in 2017. He is currently pursuing the Ph.D. degree in information security with Yonsei University, Seoul. His research interests include adversarial robustness, testing deep neural networks, adversarial machine learning, malware analysis, and usable security.

HONG EUN AHN received the B.S. degree in cyber security from Ewha Womans University, Seoul, South Korea, in 2023. She is currently pursuing the M.S. degree in information security with Yonsei University, Seoul. Her research interests include adversarial machine learning, computer vision, membership inference attack, and natural language model.

TAEKYOUNG KWON (Member, IEEE) received the B.S., M.S., and Ph.D. degrees in computer science from Yonsei University, Seoul, South Korea, in 1992, 1995, and 1999, respectively. He is currently a Professor of information security with Yonsei University, where he is also the Director of the Information Security Laboratory. From 1999 to 2000, he was a Postdoctoral Research Fellow with the University of California at Berkeley. From 2001 to 2013, he was a Professor of computer engineering with Sejong University, Seoul. His research interests include authentication, cryptographic protocols, system security, fuzzing, usable security, AI security, adversarial robustness, and adversarial machine learning. He is a member of ACM and Usenix. He is on the director board of the Korea Institute of Information Security and Cryptology (KIISC) and on the editorial committee of the Korean Institute of Information Scientists and Engineers (KIISE).