Defending Deepfakes by Saliency-Aware Attack

Qilei Li, Mingliang Gao, Guisheng Zhang, and Wenzhe Zhai

***Abstract*— With the rapid development of deep learning, especially the generative adversarial network (GAN), face modification has been substantially advanced and enables the generated images to look more realistic. Given an image or a video frame of a person, such a system can create fake images, which manipulates the movement, expression, and even appearance, e.g., hair color, eye color, and age. Such a system is termed Deepfake, which has raised significant ethical issues, especially for celebrities. With the pretrained Deepfake models being widely available on the Internet, its negative applications, such as face manipulation and pornographic generation, have exposed the dark side of the Deepfake technology to the sociocyber world. In this article, we aim to defend a well-trained Deepfake model by manipulating the raw image with unperceived perturbation. To minimize the alterations to the original image while effectively fooling the Deepfake model, we propose to selectively perturb only the foreground person region and maintain the irrelevant background. This is based on the observation that the salient object in a person's image is always the foreground face region. Such a strategy introduces negligible alterations to the original image, which makes the attack remain effective. We experimentally demonstrate the superiority of the proposed attacking framework over the existing models and show our approach is ready to be applied for out-of-the-box development.**

***Index Terms*— Deepfake, face image, generative adversarial network (GAN), model attack, saliency detection.**

I. Introduction

GENERATIVE adversarial networks (GANs) have been rapidly developed in the last decade. By training in a zero-sum manner, GANs are capable of generating realistic images with random sampled noise. It has been widely adopted in artist creation, image superresolution, and multimodal information fusion. One of the most popular applications is face generation, which is termed Deepfake. It aims to manipulate a face image to create new counterparts, which reveal different poses, emotions, and even attributes, such as hair color, gender, or even race.

Although it allows creative generation of new nonexisting people, it inevitably brings negative dark-side influence to society. For example, with the wide availability of large-scale trained Deepfake models, anyone can simply deploy these models to modify face images for the own purpose without the consent of the belonging person. It has been known to all that some face images of the celebrities have been maliciously manipulated for explicit content, which brings them substantial harassment. Because of the arousal of such bad behaviors and negative concerns due to Deepfake, some companies, such as Facebook, have carried out the corresponding policy to avoid the uploading of the generated content from Deepfake.

To mitigate such an issue, one intuitive remedy is to derive a detector that is dedicated to distinguishing whether one image has been modified. There has been a wide range of explorations in this direction in recent years [1], [2]. However, given these models are mostly based on deep learning and there are inevitably domain shifts between the training data and the test data, the performance of these detection systems is not guaranteed.

Apart from detecting the image to distinguish manipulation, another way is to attack the trained generation model so as to make the generated images less realistic and be easily distinguished by human eyes. This is based on the observation that deep neural networks are sensitive to adversarial attacks, which has been widely validated in the classification models [3], [4], [5]. By changing a bit the pixel value of the raw image, owning to the deep structure of the network, the corresponding values on the feature maps are dramatically different; therefore, the generated output will not follow the pretrained patterns. Such attacks, based on the knowledge of the target model, can be categorized into three classes [6]. The most simple but unrealistic class is called the "white-box" attack. Under this condition, the attacker is expected to have an entire understanding of the model, such as the parameters and the architecture, so that it is dedicated to this kind of model. However, it cannot be generalized to other models under different scenarios. Oppositely, there is another kind of attack named "black-box" attack, which assumes that the attacker is blindly unaware of the model. This condition is more versatile but less effective as it ignores the data characteristics. The third category is the "gray-box" attack, in which the model and the parameters are expected to be known, whereas the defense, such as image preprocessing, remains unknown.

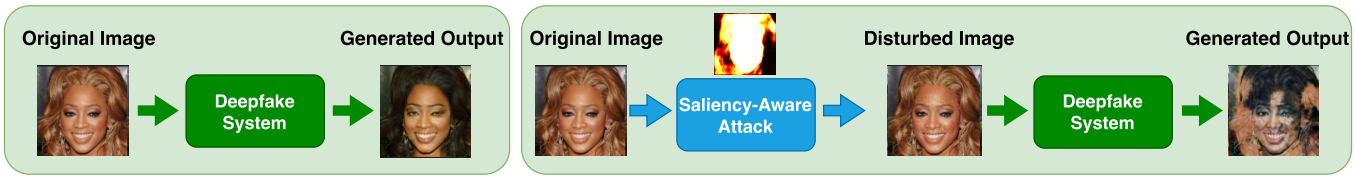
Given the gray-box condition is more realistic and can be easily satisfied, recent works [4], [7] mostly focus on this condition. The general idea for a Deepfake attack is to perform image perturbation that is invisible to human perception. Such image perturbation is expected to cause significant interference in the image translation network. The evaluation protocol for the attack is that the generated images significantly deteriorated and can be distinguished to be fake by human eyes. The general idea of a Deepfake attack is shown in Fig. 1. Some recent works [6], [8] designed different kinds of attacks, such as blur attacks, feature attack, and class condition attack. However, these attackers perform distribution on the whole raw image and inevitably harm the image details,

Manuscript received 10 January 2023; revised 26 March 2023; accepted 21 April 2023. Date of publication 8 May 2023; date of current version 2 August 2024. This work was supported in part by the National Natural Science Foundation of Shandong Province under Grant ZR2021QD041. *(Corresponding author: Mingliang Gao.)*

Qilei Li is with the School of Electrical and Electronic Engineering, Shandong University of Technology, Zibo 255000, China, and also with the School of Electronic Engineering and Computer Science, Queen Mary University of London, E1 4NS London, U.K. (e-mail: qilei.li@outlook.com).

Mingliang Gao, Guisheng Zhang, and Wenzhe Zhai are with the School of Electrical and Electronic Engineering, Shandong University of Technology, Zibo 255000, China (e-mail: mlgao@sdut.edu.cn; sdut_guisheng@163.com; wenzhezhai@163.com).

Digital Object Identifier 10.1109/TCSS.2023.3271121



(a) (b)

Fig. 1. (a) Deepfake model inference. By feedforwarding a face image into the Deepfake system, which includes a pretrained deep model, a realistic face image that manipulates the original input can be generated. (b) Defending Deepfake by the saliency-aware attack. By attacking the Deepfake system with the disturbed input image on the salient area (face region), the generated image is less realistic and can be easily distinguished by human eyes.

even if it is irrelevant to the person of interest, i.e., the background.

In this work, we solve the side effect caused by the unnecessary attack and propose the saliency-aware attack to minimize the information loss caused by image perturbation. Specifically, given a face image, we aim to separate the discriminative person region (i.e., face) out of the person's irrelevant region (i.e., background). Given the nonexistence of the fine-grained saliency label that indicates the face region, we propose to use weak labels provided by the saliency detection model that is well trained on large-scale datasets. This is based on the assumption that the face region is dominantly salient in a face region. The contemporary Deepfake models are trained to solely manipulate the facial attributes rather than the background, which also provides rich details to human perceptual systems for scenario understanding; therefore, there is no need to redundantly perturb background regions. With this observation in mind, we design the saliency-aware attack framework to adaptively disrupt the face images only in the salient region and to prevent information loss in the background region. By integrating the proposed attack framework into any existing Deepfake model, the perturbed input images lost fewer details, whereas the generated images are significantly noisy and so can be easily distinguished by human eyes.

In summary, the contributions are threefold.

1) We claim that the current Deepfake attackers can be further improved by focusing on the foreground face region in a targeted manner and can avoid the distribution of the face-irrelevant background region. Preserving the background can enhance the explainability of the disrupted image and facilitate human understanding.
2) We implement this regularization with the designed saliency-aware attack framework, which can separate the facial region from the irrelevant background region. Instead of exhaustively labeling images manually, the proposed framework utilizes weak labels from a pretrained saliency detection model to generate the mask for the face region.
3) We derive the saliency-aware attacker as a universally applicable framework, which can be adapted to any existing Deepfake models in a plug-and-play manner. Experimental results prove that it can maintain equivalent attack performance while minimizing the loss of image details, compared with the state-of-the-art (SOTA) works.

The rest of this article is structured as follows. In Section II, some recent works that are related to the proposed framework are listed. Section III details the proposed method. Section IV shows the experimental results as well as the ablation studies. In addition, this article is concluded in Section V.

II. Related Work

A. Generative Adversarial Networks

The advancement of GANs enables the generation of highly realistic face images, which are difficult to be distinguished by human eyes. The first GANs model was originally proposed in 2014 by Goodfellow et al. [9], which marked a landmark in generative models. GANs are a special type of neural network, in which two components, namely a generator and a discriminator, are trained simultaneously, with the former focusing on image generation and the latter centering on distinguishing the generated image. GANs produce reasonably satisfying output through a dynamic zero-sum game consisting of a generator and a discriminator. The GAN's greatest appeal lies in the universality of the kind of confrontation mechanism, which is believed to have made breakthroughs in the complex research areas of many target functions. Although it brings us a lot of conveniences, the GAN has been used to generate adult content and disinformation, which has a bad impact on politics and public privacy.

Recent attempts have significantly advanced realistic image generation using GANs. In 2017, Zhu et al. [10] introduced a new GAN model termed CycleGAN. Different from the traditional GANs that are composed of only a single generator and a discriminator, CycleGAN consists of two generators and two discriminators. The CycleGAN model solves the one-to-one problem, which is the translation from one domain to another. When there are many domain transformations, each domain transformation should retrain a model to solve the problem. In consequence, for training across multiple domains and multiple datasets, CycleGAN performs poorly. This issue was taken a step further in 2018 by Choi et al. [11] who put forward a StarGAN model. Different from the traditional GANs in which the image transformation is predefined in a deterministic manner, e.g., black-to-blond hair, StarGAN also considers the representative domain information as the input during the model training phase. In 2020, Choi et al. [12] introduced a StarGAN v2 model based on StarGAN. This model is mainly designed to solve the problem of converting the image of one domain into multiple images of the target domain and supporting multiple target domains. The work showed that the StarGAN v2 model can generate images with rich styles across multiple domains.

Another representative work termed GANimation [13] was proposed to generate face images with expressive facial expressions. In contrast to the previous GANs, GANimation designs the network in an attentive learning framework by focusing only on the relevant region in conveying the novel expression. Furthermore, Geng et al. [14] designed a fine-grained face manipulation model to synthesize face images of one person with diverse expressions.

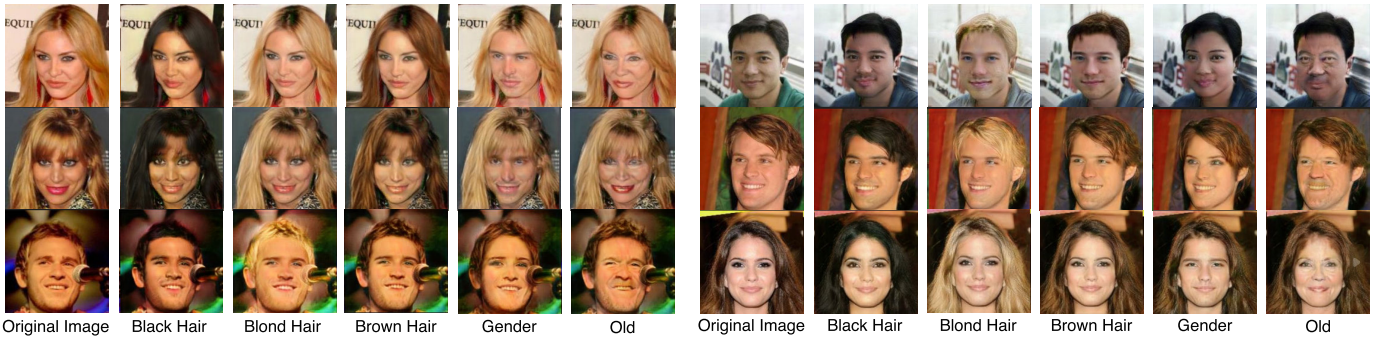


Fig. 2. Examples of face image manipulation by Deepfake model on StarGAN. The first column is the original input, and the following columns are the generated images by changing different facial attributes.

In this work, we use the large-scale trained StarGAN model to generate face images by modifying different attributes and apply the proposed saliency-aware attacker on the raw images to defend this well-trained StarGAN model.

B. Deepfake

Deepfake is a kind of artificial intelligence technology that uses the machine learning model of GANs [9] to alter face images. It was first created by a Reddit user in 2017 as a script used to generate face-swapped adult pornographic content [15]. The technology of making fake images or videos is also known as AI face swap in the industry. The core principle is to use algorithms, such as GANs or convolutional neural networks, to graft the face of the target object into the imitated object. Specifically, first, the target face features are extracted, and then, the target imitation object's face is replaced by the target object's face. Finally, the target images are resynthesized into fake images.

Deep learning technology can automate this process, and this approach enables the common encoder to discover and learn the similarities between two sets of input face images. This is relatively unchallenging for the encoder because faces usually have similar features, such as eye, ear, and mouth positions. Examples of face image manipulation by the Deepfake model are shown in Fig. 2.

As deep learning technology is gradually applied to Deepfakes, facial images synthesized by the Deepfake technology are becoming more and more realistic. It has been used positively for several applications [16], [17]. Deepfakes could boost photography, virtual reality, filmmaking, and the economy. Moreover, it can make people's life more convenient. For example, people can choose clothes online, try different styles by using Deepfake technology, and use it to generate photos of loved ones who have passed away. However, the malicious use of Deepfakes is on the rise. It is often used to produce false news or some adult content [18], [19], [20], [21], e.g., Deepfakes have the potential to spread untrue information about candidates and other political leaders. Even if this fake news turns out to be fake, it could still stick in voters' minds and change their perceptions of the characters in it.

The main uses of Deepfake technology are pornography, political subterfuges, extortion, and financial fraud, and contents produced by Deepfakes are difficult to distinguish for humans. Therefore, in this work, we used some adversarial attacks [6], [22], [23] to reduce the fidelity of the synthesized face images.

C. Deepfake Attack

A Deepfake attack aims to deceive neural network models, such as image forgery and target recognition, by adding noise that is unperceptive to the human eye so that the images or videos produced by these models can be recognized by the human eye.

There are plenty of works [5], [24], [25], [26] that conduct adversarial attacks in deep classification networks. Goodfellow et al. [4] explore a fast gradient sign method (FGSM), which significantly enhanced the utility of adversarial examples and provided an alternative regularization method to traditional regularization methods. The iterative FGSM (I-FGSM) is also a representative iterative attack, which was introduced by Kurakin et al. [7]. The key idea of I-FGSM is to iteratively perform gradient sign operations on the original images. This work demonstrates the feasibility of adversarial examples for machine learning systems in the real world. Besides, Madry et al. [23] studied the adversarial robustness of neural networks through the lens of robust optimization, and another stronger iterative attack, projected gradient descent (PGD), has been proposed by them in 2017. Based on the previous adversarial sample generation, Athalye et al. [22] made a further study on physical environment transformation to perturb raw images. Athalye et al. [22] also proposed the expectation-over-transformation (EoT) method as a further exploration, which inspired the real-world antagonist who proves the existence of a 3-D antagonist. Furthermore, Ruiz et al. [6] proposed a faster heuristic iterative spread-spectrum disruption for evading blur defenses.

Recent work by Yeh et al. [27] explored two types of adversarial attackers by designing a novel adversarial loss function to adversarially learn the blurred and distorted outputs jointly so as to optimize the network with gradient descent. Huang et al. [28] proposed a novel antiforgery model by disrupting the manipulation model in a black-box setting, which can incessantly generate perceptual-aware disruption.

In this work, unlike the aforementioned methods, which manipulate the whole image to attack the Deepfakes, we propose a saliency-aware attack framework, which selectively focuses only on the discriminative face region, so as to minimize the alterations to the raw image and preserve more details.

D. Saliency Detection

The core idea of saliency detection is to extract important information from the picture and enable the network to pay

Fig. 3. Few samples of face images. From top to bottom, first row: Face images. Second row: Saliency maps. Third row: Disrupted images. Fourth row: Generated images by Deepfake from disrupted images.

more attention to significant areas in the picture. In recent years, image saliency detection has made great progress, especially since the method based on deep learning has achieved superior performance.

According to a survey by Borji et al. [29], image saliency detection methods can be classified into bottom-up model [30], [31], [32] and top-down model [33], [34], [35]. The bottom-up approach is based directly on the primary information, such as contrast, whereas the top-down approach takes into account our analysis of goals and scenarios. Qu et al. [36] designed an RGBD saliency detection model based on the conditional random field under superpixels. They made full use of the advantages of different modal cues, effectively integrated them, improved the performance of the saliency detection model, and played a role in the field of visual-driven applications. First, Hanet et al. [37] exposed the cross-modal discrepancy in the RGBD data and proposed two cross-modal transfer learning strategies.

There are several works [38], [39] that have been done to capture the comprehensive and implicit attributes from the feature map. An anisotropic center-surround difference (ACSD) measure with 3-D spatial prior refinement was introduced by Ju et al. [38], which combines both depth-based background estimation and learning-based direction contrast weighting. Guo et al. [39] proposed a novel salient object detection method for RGBD images based on an evolution strategy. This model fully explored the potential of color cues and depth cues in the whole procedure of salient object detection saliency propagation.

It is known that the feature maps extracted in different layers encode heritable information captured at different scales under different levels of receptive files. These multiscale features contribute to the saliency detection model with diverse knowledge. One recent work PFAN [40] was proposed by Zhao et al. to attentively reweighting the feature maps at different hierarchies. In PFAN, the high- and low-level features are processed by the channel and spatial attention modules to highlight the significant representation while suppressing the insignificant ones, so as to boost the network in learning discriminative detection capacity.

In this work, to generate a weak label for the foreground face region, we employ PFAN [40] to generate a saliency map, which can naturally identify the dominant face region as the saliency region.

III. Defending Deepfake by Saliency-Aware Attack

In order to defend Deepfakes from generating realistic face images, one effective solution that has been extensively studied is to disrupt the input image with ideally imperceptible noise. This is based on the observation that deep learning-based models generally follow the independent identically distribution (IID) assumption, and once the test data have different distributions from the training data, the performance of a well-trained Deepfake model can drop significantly. However, disrupting the input face image will inevitably bring negative influence to image quality and will hinder human perception. The tradeoff between disruption and interpretability is merely considered by previous methods. In this work, we aim to effectively attack Deepfake models by perturbing the input face image while minimizing its degradation. This is realized by the saliency-aware attack, which adaptively disrupts only the foreground face region to fool the Deepfake generator.

A. Overview

The ultimate goal of the proposed saliency-aware attack is to defend Deepfake models by imperceptibly disrupting the salient face region on the input image, so as to achieve an effective attack with minimal image information loss. This framework is formulated in three steps: 1) detection of the salient face region in the input face image to get the face mask (Section III-B); 2) imperceptibly perturb the face image in a saliency-aware manner (Section III-C); and 3) feedforward the putter face image into the Deepfake model to get the generated image (Section III-D), which is expected and can be distinguished by human eyes. The overall framework of the proposed method is shown in Fig. 4.

B. Salient Face Region Detection

Given a face image as the input for the Deepfake system, the face region is attentively the target area for manipulation. In order to better disrupt the input image for a Deepfake attack with less information loss, it is nontrivial to accurately separate the face region of the face-irrelevant background. However, given the nonexistence of masks for the face region, it is highly expensive and even impossible to label manually for all the images. Therefore, considering that the face region is intuitively the salient part of a face image, this face mask generation task can naturally be transferred as a salience detection counterpart.

By tasking the recent advantage of saliency detection, which is to locate the dominant part of one natural image, the weak label can be easily obtained by simply feeding the original input image to a pretrained saliency detection model. Denoting an input image as I_{ori} and a trained saliency detection model as $f_{\text{sa}}(\cdot)$, the saliency mask is calculated as

$$M = f_{\text{sa}}(n(I_{\text{ori}}, \mu, \sigma), \theta_f) \tag{1}$$

where $n(\cdot)$ is the normalization operation to diminish the numerical scale variations of the raw input image based on

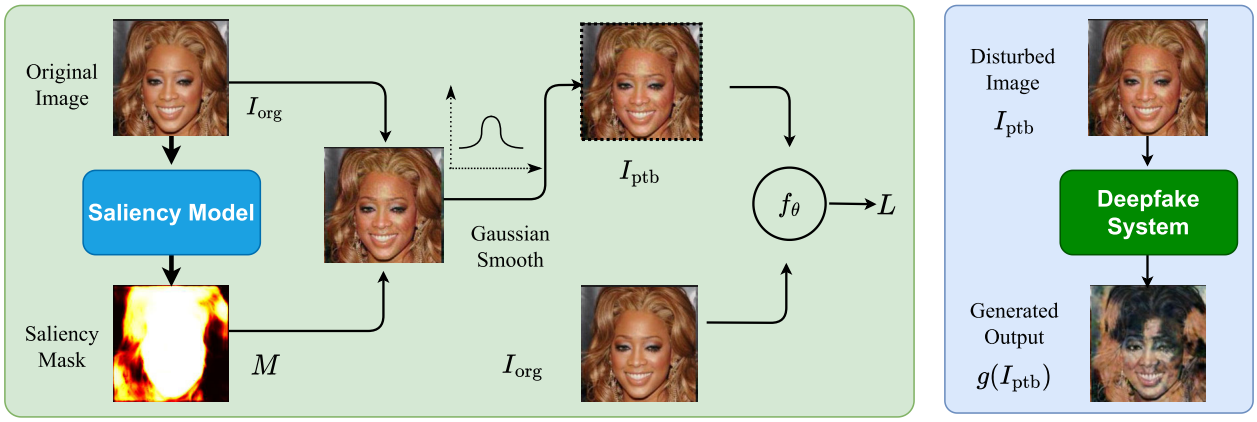


Fig. 4. (a) Deepfake model inference. By feedforwarding a face image into the Deepfake system, which includes a pretrained deep model, a realistic face image that manipulates the original input can be generated. (b) Defending Deepfake by the saliency-aware attack. By attacking the Deepfake system with the disturbed input image on the salient area (face region), the generated image is less realistic and can be easily distinguished by human eyes.

the mean μ and variance σ, and θ_f is the parameter of the saliency model, which is frozen during model inference.

The saliency map M can highlight the dominant face region, as shown in Fig. 3. Instead of applying postprocessing to transform it into a binary map, we keep its raw value to indicate a different level of saliency. This can mitigate the unavoidable false prediction in the weak label to some extent. With the saliency map M, the foreground face region is obtained as

$$I_{\text{face}} = I_{\text{org}}(x, y) \times M(x, y). \tag{2}$$

Remark: The saliency map M as the wake label to indicate the face region is obtained by utilizing the pretrained detection model, which is cheap and can be computed in real time. By applying the saliency map to separate the foreground region, the subsequent perturbation will not influence the background, hence leading to less perceptual detail loss.

C. Saliency-Aware Image Perturbation

Though contemporarily deep learning-based models have achieved significant performance improvement compared with traditional methods, they generally follow a simple and unrealistic assumption that the training data and the test data are collected in the same distribution, which is termed independent and identical distribution (IID). Once this assumption is no longer satisfied, the performance of a well-trained model can significantly drop. This observation is also applicable to the Deepfake model. In order to attack a pretrained Deepfake model, a straightforward yet effective solution is to break the IID assumption by altering the distribution of the input data, which is realized by disrupting the image as

$$I_{\text{ptb}} = I_{\text{org}} + \epsilon \tag{3}$$

where I_{ptb} is the perturbed face image, and ϵ is a human-imperceptible disruption that can be Gaussian noise or an adversarial noise. Given the perturbed image I_{ptb} as the input, a Deepfake model can output a generated image O_{ptb}, which is expected to be highly unnatural and, therefore, to bring to notice that the corresponding source image is untrustworthy. Although manually disrupting the input face images in such a way can break the IID assumption so as to attack the Deepfake model, it unavoidably sacrifices the image quality and lost facial details. There exists a balance between disrupting the face input and preserving the image detail. However, this trade-off is always omitted in the literature. We address the issue by disrupting the image in a saliency-aware manner. Given the saliency-aware map M generated by (2), the perturbation is achieved by reformulating (3) as

$$\epsilon_{\text{ptb}} = M \times \epsilon, \quad I_{\text{ptb}} = I_{\text{org}} + \epsilon_{\text{ptb}} \tag{4}$$

where ϵ_{ptb} is the disruption factor that is added to the original image to achieve perturbation. This saliency-aware perturbation is intuitively equivalent to applying the disruption upon the face region as

$$I_{\text{ptb}} = (I_{\text{face}} + \epsilon) + \tilde{I}_{\text{face}} \tag{5}$$

where $\tilde{I}_{\text{face}}$ is the nonsalient region, which is the complementary set to I_{face}.

Denoting the Deepfake model as $g(\cdot)$, the objectives of the attack are in the two manifolds.

1) The disrupted image I_{ptb} is expected to maintain maximum image information, i.e., the perturbation should be human-imperceptible, which is formulated as

$$\min_{\epsilon_{\text{ptb}}} L(I_{\text{org}} + \epsilon_{\text{ptb}}), \quad \text{s.t. } \|\epsilon_{\text{ptb}})\|_\infty \leq \tau. \tag{6}$$

2) The generated image from the original input $g(I_{\text{org}})$ and the disrupted image $g(I_{\text{ptb}})$ is expected to be significantly different, i.e., the perturbation successfully leads to the degradation of the generated image, which is formulated as

$$\max_{\epsilon_{\text{ptb}}} L(g(I_{\text{org}}), g(I_{\text{org}} + \epsilon_{\text{ptb}})), \quad \text{s.t. } \|\epsilon_{\text{ptb}})\|_\infty \leq \tau. \tag{7}$$

Remark: In order to attack the Deepfake model with minimal sacrifices of the image quality, we propose to distribute only the salient area. Although the saliency detection model is trained with natural image rather than purely facial images, it may unavoidably introduce errors due to labels being weakly labeled. To mitigate this issue, our solution is to use residual noise and add it to the original input.

D. Deepfake Model Attack

Given a trained Deepfake model $g(\cdot)$ parameterized by θ_g, which is able to generate a realistic face image $g(I_{\text{org}})$ with the source image I_{org}, our objective is to confuse it with a disrupted sample I_{ptb}, which is called an adversarial attack.

TABLE I

COMPARISON WITH THE SOTA METHODS. THE BEST RESULTS ARE SHOWN IN **RED**. ALL L^2 ERRORS ARE AMPLIFIED 100 TIMES TO HIGHLIGHT THE DIFFERENCE

Attacker	Tiny (100)		Small (500)		Middle (1000)		Large (2000)		Average	
	L^1	L^2	L^1	L^2	L^1	L^2	L^1	L^2	L^1	L^2
I-FGSM [7]	0.020	0.089	0.019	0.087	0.019	0.087	0.019	0.087	0.019	0.088
Spread-Spectrum [6]	0.017	0.071	0.017	0.071	0.017	0.071	0.017	0.071	0.017	0.071
EoT-Blur [6]	0.020	0.090	0.020	0.089	0.020	0.089	0.020	0.089	0.020	0.089
Saliency-Aware	**0.016**	**0.068**	**0.016**	**0.068**	**0.016**	**0.068**	**0.016**	**0.068**	**0.016**	**0.068**

Algorithm 1 Defending Deepfake by Saliency-Aware Attack

Input: Deepfake generator $g(\cdot)$, Pre-trained saliency detection model $f_{\text{sa}}(\cdot)$, Input face image I_{org}, Disruption magnitude m.

Output: Disrupted face image I_{ptb}, Generated image from the disrupted counterpart $g(I_{\text{ptb}})$.

Using saliency detection to generate saliency mask M as (1).

Applying noise on the face region to get the residual component by I-FGSM algorithm by (10).

Feeding the saliency-aware perturbed face image I_{ptb} into the deepfake generator to get the translated image $g(I_{\text{ptb}})$.

One of the most popular attackers is the fast gradient signed method (FGSM), which was proposed by Goodfellow et al. [4]. The principle of FGSM is to generate a new sample toward the direction of gradient descent so as to minimize the loss, which is formulated as

$$I_{\text{ptb}} = I_{\text{org}} + m * \text{sign}(\nabla_x J(\theta_g, x, I_{\text{org}})) \tag{8}$$

where $\nabla_x J$ is the cost function for training $g(\cdot)$, and the multiplier m controls the magnitude of the disruption. The FGSM algorithm is further improved by *iteratively* performing the perturbation and termed I-FGSM [7], which can be mathematically denoted as

$$\begin{aligned} I^0_{\text{ptb}} &= I_{\text{org}} \\ I^{t+1}_{\text{ptb}} &= I^t_{\text{ptb}} + m * \text{sign}\left(\nabla_x J\left(\theta_g, I^t_{\text{ptb}}, y\right)\right). \end{aligned} \tag{9}$$

These attack models achieved advanced performance in defending the trained generative models. However, they unavoidably degrade the image quality after perturbation.

To seek an appropriate tradeoff, we propose the saliency-aware attack. Specifically, instead of adversarially disrupting the whole image, we separate the salient face area and add noise to it only. This process is formulated as

$$\begin{aligned} I^0_{\text{ptb}} &= I_{\text{face}} \\ I^{t+1}_{\text{ptb}} &= I^t_{\text{ptb}} + m * \text{sign}\left(\nabla_x J\left(\theta_g, I^t_{\text{ptb}} \times M, y\right)\right) \times M. \end{aligned} \tag{10}$$

Inspired by the recent work [6], to bootstrap the attacker, the input image is first blurred by a Gaussian smoothing filter in each iteration.

Remark: By iteratively applying the gradient attack upon the input, we can get the final perturbed input I^t_{ptb}. Feeding it into the Deepfake model g, one fake image can be generated, but it is less realistic and can be obviously distinguished.

IV. EXPERIMENTAL RESULTS AND ANALYSIS

A. Implementation Details

Our saliency-aware attacker adopted the PFAN [40] to generate the saliency mask. We used a Gaussian blur filter for image preprocessing. The kernel size was set to 11, and σ was set to 1. The magnitude of disruption in (3) was 0.1, and the cost function for the I-FGCA method in (3) was MSE Loss. The framework was implemented by the PyTorch [41] framework, and all the experiments were performed on an NVIDIA A100 GPU.

B. Datasets

The CelebFaces Attributes Dataset (CelebA) [42] is an open-source dataset. It is a large-scale dataset of real face attributes. These face images contain abundant face poses with different backgrounds. Each image in CelebA had been labeled with five facial feature points and 40 attribute labels. Given the overwhelming amount of images (202, 599) in CelebFaces, it is extremely time-consuming to process all of them, and we sampled images in different scales to evaluate the methods, namely tiny (100 images), small (500), middle (1000), and large (2000).

C. Evaluation Protocols

We assessed the performance of the proposed saliency-aware attack framework from both quantitative and qualitative aspects. For quantitative evaluation, to measure the difference between the original face images and the disrupted image, we adopted L^1 and L^2 errors as the metrics. The L^1 and L^2 errors are defined as the 1-norm, 2-norm, and infinity norm of the difference between I_{org} and I_{ptb}, respectively. Lower errors denote less information loss. For qualitative evaluation, we visualized the generated image by the Deepfake framework from the I_{org} and I_{ptb}, respectively.

D. Comparison With SOTA

To validate the proposed saliency-aware attack framework, we compared it with the SOTA competitors under various settings [4], [6], [7]. The quantitative evaluation results are shown in Table I. It can be seen that the proposed attack framework exhibits superior performance against all other competitors, on a wide range of benchmark settings with different evaluation metrics. Our method can always achieve less degradation on the original image, which provides a solid foundation for the following attack.

We provided visualizations of the generated image, respectively, from the original and the disrupted images by changing the different attributes with the widely adopted StarGAN [11]. From Fig. 5, one can see that the proposed method can successfully defend the Deepfake by making the generated output being distinguishable with human eyes, whereas the disrupted images remain informative as the original ones.

Fig. 5. Visualization of the generated images from the original and disrupted faces images. It can be observed that the generated images from the original image are realistic, whereas the counterparts from the disrupted images are highly unrealistic and can be easily detected by human. Therefore, we demonstrate the effectiveness of the proposed attack framework.

The output from the disrupted images has obvious artifacts, therefore validating that our framework can reliably defend the Deepfake model.

V. Conclusion

A. Discussion

The proposed saliency-aware attack framework can prevent the Deepfake models from generating real-like face images. This can protect the privacy of the public and prevent some ethical issues. However, adding perturbations to the original face image will inevitably cause degradation to the original images, resulting in the loss of information and details. Besides, our framework is highly driven by the pretrained saliency detection model, which may not be reliable for accurately detecting the foreground region and cause the perturbation less effective.

B. Summary

In this article, we proposed the saliency-aware attack framework to defend a well-trained Deepfake model by manipulating the raw image with unperceived perturbation. It is achieved by selectively perturbing only the foreground person region and maintaining the irrelevant background to simultaneously fool the Deepfake model while minimizing the alterations to the original image effectively. Such a strategy introduces negligible alterations to the original image, which makes the attack remain effective. We performed extensive experiments to demonstrate the effectiveness of the proposed method and showed superiority over the SOTA methods.

REFERENCES

[1] R. Wang et al., "FakeSpotter: A simple yet robust baseline for spotting AI-synthesized fake faces," 2019, *arXiv:1909.06122*.

[2] S.-Y. Wang, O. Wang, R. Zhang, A. Owens, and A. A. Efros, "CNN-generated images are surprisingly easy to spot... for now," in *Proc. CVPR*, Jun. 2020, pp. 8695–8704.

[3] N. Carlini and D. Wagner, "Towards evaluating the robustness of neural networks," in *Proc. IEEE Symp. Secur. Privacy (SP)*, May 2017, pp. 39–57.

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

[5] N. Papernot, P. McDaniel, S. Jha, M. Fredrikson, Z. B. Celik, and A. Swami, "The limitations of deep learning in adversarial settings," in *Proc. IEEE Eur. Symp. Secur. Privacy (EuroSP)*, Mar. 2016, pp. 372–387.

[6] 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.* Glasgow, U.K.: Springer, 2020, pp. 236–251.

[7] A. Kurakin, I. J. Goodfellow, and S. Bengio, "Adversarial examples in the physical world," in *Artificial Intelligence Safety and Security*. Boca Raton, FL, USA: CRC Press, 2018, pp. 99–112.

[8] Q. Li, Y. Guo, and H. Chen, "Practical no-box adversarial attacks against DNNs," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 12849–12860.

[9] I. Goodfellow et al., "Generative adversarial nets," in *Proc. NeurIPS*, 2014, vol. 1, no. 2, pp. 1–9.

[10] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. ICCV*, Oct. 2017, pp. 2223–2232.

[11] 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. CVPR*, Jun. 2018, pp. 8789–8797.

[12] Y. Choi, Y. Uh, J. Yoo, and J.-W. Ha, "StarGAN v2: Diverse image synthesis for multiple domains," in *Proc. CVPR*, Jun. 2020, pp. 8188–8197.

[13] 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)*, 2018, pp. 818–833.

[14] Z. Geng, C. Cao, and S. Tulyakov, "3D guided fine-grained face manipulation," in *Proc. CVPR*, 2019, pp. 9821–9830.

[15] J. Kietzmann, L. W. Lee, I. P. McCarthy, and T. C. Kietzmann, "Deepfakes: Trick or treat?" *Bus. Horizons*, vol. 63, no. 2, pp. 135–146, Mar. 2020.

[16] T.-C. Wang, A. Mallya, and M.-Y. Liu, "One-shot free-view neural talking-head synthesis for video conferencing," in *Proc. CVPR*, Jun. 2021, pp. 10039–10049.

[17] J. H. Kietzmann, K. Hermkens, I. P. McCarthy, and B. S. Silvestre, "Social media? Get serious! Understanding the functional building blocks of social media," *Bus. Horizons*, vol. 54, no. 3, pp. 241–251, 2011.

[18] A. K. Jain, P. Flynn, and A. A. Ross, *Handbook of Biometrics*. Nashville, TN, USA: Springer, 2007.

[19] T. Mak and D. Temple-Raston. (2020). Where are the deepfakes in this presidential election. NPR. [Online]. Available: https://www.npr.org/2020/10/01/918223033/where-are-the-deepfakes-in-this-presidential-election

[20] D. K. Citron, *How Deepfakes Undermine Truth and Threaten Democracy*. TED, 2019.

[21] R. Cellan-Jones, "Deepfake videos double in nine months," BBC News, London, U.K., Oct. 2019.

[22] A. Athalye, L. Engstrom, A. Ilyas, and K. Kwok, "Synthesizing robust adversarial examples," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 284–293.

[23] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," in *Proc. ICLR*, 2018, pp. 1–28.

[24] S.-M. Moosavi-Dezfooli, A. Fawzi, O. Fawzi, and P. Frossard, "Universal adversarial perturbations," in *Proc. CVPR*, Jul. 2017, pp. 1765–1773.

[25] S.-M. Moosavi-Dezfooli, A. Fawzi, and P. Frossard, "DeepFool: A simple and accurate method to fool deep neural networks," in *Proc. CVPR*, Jun. 2016, pp. 2574–2582.

[26] A. Nguyen, J. Yosinski, and J. Clune, "Deep neural networks are easily fooled: High confidence predictions for unrecognizable images," in *Proc. CVPR*, 2015, pp. 427–436.

[27] 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/CVF Winter Conf. Appl. Comput. Vis. Workshops*, Mar. 2020, pp. 53–62.

[28] 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," in *Proc. IJCAI*, 2022, pp. 761–767.

[29] A. Borji, M.-M. Cheng, H. Jiang, and J. Li, "Salient object detection: A benchmark," *IEEE Trans. Image Process.*, vol. 24, no. 12, pp. 5706–5722, Jan. 2015.

[30] J. Shi, Q. Yan, L. Xu, and J. Jia, "Hierarchical image saliency detection on extended CSSD," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 4, pp. 717–729, Apr. 2016.

[31] J. Lei et al., "A universal framework for salient object detection," *IEEE Trans. Multimedia*, vol. 18, no. 9, pp. 1783–1795, Sep. 2016.

[32] H. Peng, B. Li, H. Ling, W. Hu, W. Xiong, and S. J. Maybank, "Salient object detection via structured matrix decomposition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 4, pp. 818–832, Apr. 2017.

[33] T. Chen, L. Lin, L. Liu, X. Luo, and X. Li, "DISC: Deep image saliency computing via progressive representation learning," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 27, no. 6, pp. 1135–1149, Jun. 2016.

[34] H. Jiang, J. Wang, Z. Yuan, Y. Wu, N. Zheng, and S. Li, "Salient object detection: A discriminative regional feature integration approach," in *Proc. CVPR*, Jun. 2013, pp. 2083–2090.

[35] G. Li and Y. Yu, "Deep contrast learning for salient object detection," in *Proc. CVPR*, Jun. 2016, pp. 478–487.

[36] L. Qu, S. He, J. Zhang, J. Tian, Y. Tang, and Q. Yang, "RGBD salient object detection via deep fusion," *IEEE Trans. Image Process.*, vol. 26, no. 5, pp. 2274–2285, May 2017.

[37] J. Han, H. Chen, N. Liu, C. Yan, and X. Li, "CNNs-based RGB-D saliency detection via cross-view transfer and multiview fusion," *IEEE Trans. Cybern.*, vol. 48, no. 11, pp. 3171–3183, Nov. 2017.

[38] R. Ju, Y. Liu, T. Ren, L. Ge, and G. Wu, "Depth-aware salient object detection using anisotropic center-surround difference," *Signal Process., Image Commun.*, vol. 38, pp. 115–126, Oct. 2015.

[39] J. Guo, T. Ren, and J. Bei, "Salient object detection for RGB-D image via saliency evolution," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jul. 2016, pp. 1–6.

[40] T. Zhao and X. Wu, "Pyramid feature attention network for saliency detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Long Beach, CA, USA, 2019, pp. 3080–3089, doi: 10.1109/CVPR.2019.00320.

[41] A. Paszke et al., "Automatic differentiation in PyTorch," in *Proc. NIPS-W*, 2017, pp. 1–4.

[42] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *Proc. ICCV*, Dec. 2015, pp. 3730–3738.