Disrupting Deepfakes via Union-Saliency Adversarial Attack

Guisheng Zhang, Mingliang Gao, Qilei Li, *Graduate Student Member, IEEE*, Wenzhe Zhai, Guofeng Zou, and Gwanggil Jeon, *Senior Member, IEEE*

***Abstract*—With the rapid development of electronic payment technologies, facial recognition-based payment systems have become increasingly popular and indispensable. However, the majority of facial recognition payment systems are vulnerable to being manipulated by facial deepfake technology, and it would be a serious threat to personal property and privacy. In order to effectively defend deepfake models on the premise of minimizing alterations to the original image, we propose a union-saliency attack model which is a well-trained deepfake model while maintaining plausible detail of the original face images. To this aim, we derive a union mask mechanism to accurately determine facial region as a prior in guiding the subsequent perturbations, with the objective of minimizing the information loss on input images. Additionally, we propose a novel structural similarity loss and a noise generator to minimize detail degradation. Experiments prove that the proposed method can interfere with deepfake models effectively and minimize the distortion of the original image simultaneously.**

***Index Terms*—Deepfake, generative adversarial network, facial recognition payment, model attack.**

I. Introduction

FACIAL recognition system is a critical component of digital payment. Although this technology brings a lot of conveniences, it can be maliciously abused with the deepfake model, which is capable of generating real-time fake face images to fool face payment systems with unauthorized transactions [1]. To solve this issue, various countermeasures have been developed to combat deepfakes. These methods can be categorized into two modes, namely deepfake detection and deepfake disruption. Deepfake detection is to detect whether the image is real or fake. In contrast, deepfake disruption aims to prevent the creation of convincing synthetic images. The schematic diagrams of the two modes are shown in Fig. 1.

Some attempts were made to explore deepfake detection methods [2], [3], [4]. Hsu et al. [2] proposed a two-streamed network that facilitates the extraction of discriminative synthetic features at intermediate and advanced levels by consolidating cross-layer representations. Zhao et al. [3] designed a pair-wise self-consistency learning to detect fake images based on source feature inconsistency. Yang et al. [4] proposed a Masked Relation Learning model, which decreases the redundancy to learn informative relational features. Specifically, a spatio-temporal attention module is exploited to learn the attention features of multiple facial regions. Li et al. [5] introduced an Artifacts-Disentangled Adversarial Learning framework, which aims to achieve accurate deepfake detection through the disentanglement of artifacts from irrelevant information. To address an issue of discrepancies in quality between test faces and training faces, Wang et al. [6] employed an innovative technique termed the Localization Invariance Siamese Network (LiSiam) for the purpose of deepfake detection. The primary focus of LiSiam is to ensure localization invariance and enhance the ability of the model to cope with diverse image degradation. These deepfake detection methods have made significant progress in detecting fake images. However, these methods work as an ex-post approach [7] and cannot prevent the generation of fake images ahead of time. More critically, a common issue with current deepfake detection algorithms is their insufficient generalization capacity [8], [9], [10].

In contrast, deepfake disruption, aimed at preventing the generation of fake images proactively. The process of disrupting deepfakes typically involves the introduction of imperceptible noise into the images [11], [12], [13]. Although it can disrupt the deepfake models, it unavoidably results in the loss of image details. Besides, existing methods often introduce noise across the entire image without considering whether those regions correspond to facial areas. Although there exist some methods [14] that perturb only the face region, they are mostly driven by a pre-trained facial region detector, which is prone to be inaccurate caused by the variations between the training and testing distribution.

In this work, we propose a union-saliency attack model, which provides an effective and practical solution to counter deepfake models while keeping image modifications to a minimum. It effectively addresses the issue of fraudulent utilization of face images generated by deepfake technology

Manuscript received 15 May 2023; revised 1 August 2023 and 26 September 2023; accepted 23 November 2023. Date of publication 28 November 2023; date of current version 26 April 2024. This work was supported in part by the National Natural Science Foundation of Shandong Province under Grant ZR2022MF307. *(Corresponding authors: Mingliang Gao; Gwanggil Jeon.)*

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

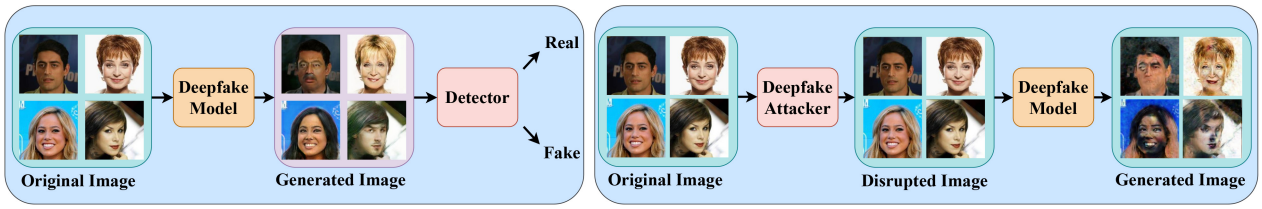
Qilei Li is with the School of Electronic Engineering and Computer Science, Queen Mary University of London, E1 4NS London, U.K. (e-mail: qilei.li@outlook.com).

Gwanggil Jeon is with the School of Electrical and Electronic Engineering, Shandong University of Technology, Zibo 255000, China, and also with the Department of Embedded Systems Engineering, Incheon National University, Incheon 22012, South Korea (e-mail: ggjeon@gmail.com).

Digital Object Identifier 10.1109/TCE.2023.3337207

1558-4127 © 2023 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See https://www.ieee.org/publications/rights/index.html for more information.

Authorized licensed use limited to: Queen Mary University of London. Downloaded on March 06,2025 at 11:54:15 UTC from IEEE Xplore. Restrictions apply.



(a) Deepfake Detection (b) Deepfake Disruption

Fig. 1. Diagrams of deepfake detection and deepfake disruption. (a) Deepfake detection. Using a detector determines whether an image was synthesized by a deepfake model. (b) Deepfake disruption. Adding perturbations to input images disrupts deepfake models from producing realistic images. The images generated by deepfake models have obvious artifacts and can be easily distinguished visually.

in the context of electronic payments. Specifically, given a face image, we first generate a preliminary salient mask to distinguish and extract the face region. As a complementary, a manipulation mask is introduced to guide addition perturbation. It provides details of the facial area where the deepfake model modifies. Therefore, a union mask is capable of inducing the noise to be added to the facial region. Furthermore, we introduce a noise generator with a structural similarity loss function to further minimize detail degradation. In sum, the contributions of the work are three-fold:

- We propose a union-saliency attack framework to focus on the face region. The proposed framework can minimize the distortion of perturbed images and disrupt deepfake systems.
- We employ a union mask, which consists of a salient mask and a manipulation mask, to provide a more precious face region. Furthermore, we adopt a noise generator to optimize a small initial perturbation, and to reduce the loss of details of the original image with disturbing the deepfake model.
- We introduce a structural similarity loss to focus on the difference between the original image and the disrupted image. This is beneficial to reduce the detail loss of the disrupted image.

The rest paper is structured as follows: Section II presents the related work. Section III illustrates the proposed method in detail. Section IV analyses the experimental results. The paper is concluded in Section V.

II. Related Work

A. Generative Adversarial Network

The original GAN model was put forward by Goodfellow et al. [15]. It has excited a sensation in the domain of deep learning and expanded the application areas of deep learning. There has been a significant increase in the development and proposal of efficient GAN architectures that are capable of learning the various variations of individual faces, including differences in hair color, age, expression, and gender. Zhu et al. [16] proposed an image-to-image translation model named CycleGAN. Compared to traditional paired image-to-image translation methods, the CycleGAN model does not require training on paired images. However, for training across multiple domains and datasets, CycleGAN has shown unsatisfactory results. Choi et al. [17] further addressed this problem by introducing the StarGAN model in 2018.

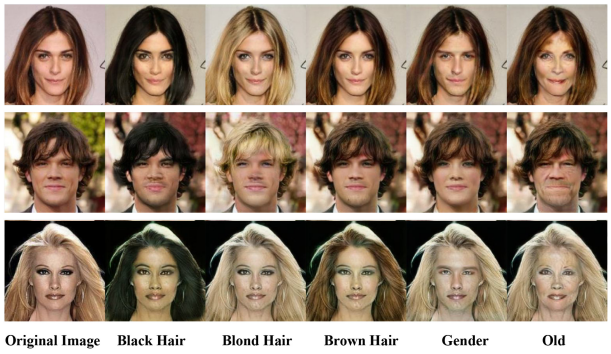


Fig. 2. Generated fake faces with different facial attributes.

Pumarola et al. [18] employed GANimation for continuous domain facial expression synthesis. The GANimation employs attention mechanisms to enhance the robustness of the network in background and lighting variations. Another representative work termed StyleGAN3 was proposed by Karras et al. [19]. In comparison with StyleGAN [20] and StyleGAN2 [21], StyleGAN3 enhances the capture of perceptual information utilizing reversible multi-layer perceptron and microstructure discriminator.

In this work, we utilize a StarGAN model that has been trained to create face images with alterations to diverse attributes. Furthermore, to seek a kind of adversarial perturbation that could disrupt the fully trained StarGAN model, we propose an anti-forgery approach termed union-saliency attacker.

B. Deepfake Creation

Although GAN offers numerous benefits, they have also been used to create inappropriate adult content and spread misinformation, which pose a threat to personal privacy and have negative impacts on politics. Deepfake creation is to utilize deep learning models to synthesize facial images. It is employed to replace one individual facial features with another individual facial features or change face attributes. Some generated fake faces with different facial attributes, *e.g.*, changing hair, color changing gender, age, are depicted in Fig. 2.

Recently, with the development of GAN, numerous GAN-based deepfake models [22], [23], [24], [25] have been designed to generate forgery facial images, that are difficult to be distinguished visually. He et al. [26] implemented AttGAN

Authorized licensed use limited to: Queen Mary University of London. Downloaded on March 06,2025 at 11:54:15 UTC from IEEE Xplore. Restrictions apply.

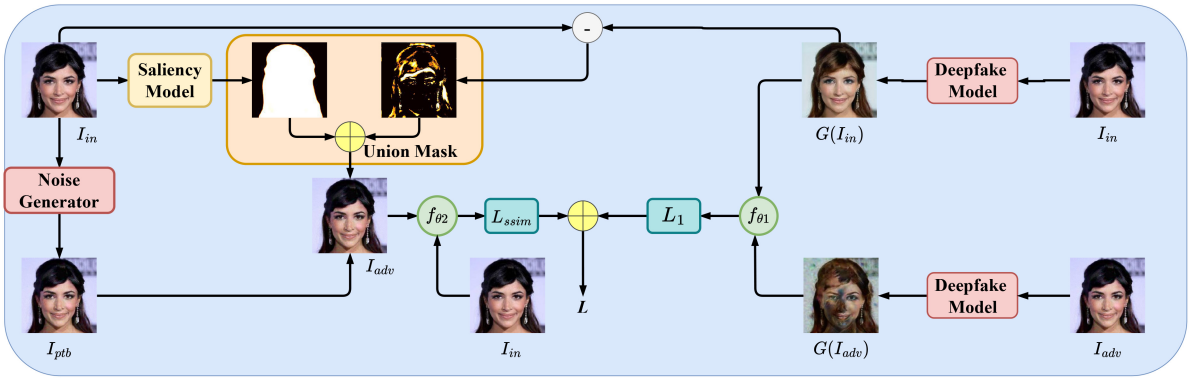


Fig. 3. Illustration of the union-saliency attack for defending deepfake. By perturbing the input images in the salient area, particularly the face region, the deepfake system is effectively disrupted. As a result, the generated images become less realistic and display noticeable artifacts that can be easily detected visually.

with an attribute classification constraint. The attribute classification constraint is utilized to edit the generated images while preserving the qualities of the original image. Liu et al. [27] proposed STGAN to highlight specific facial features while maintaining the other areas intact. Gao et al. [28] reported a high-fidelity arbitrary face editing (HifaFace) that can make accurate face editing while keeping rich details of undesired attribute areas.

C. Deepfake Disruption

The objective of deepfake disruption is to prevent the creation of deepfakes proactively. In recent years, some studies [29], [30], [31] have illustrated that adding imperceptible noise to the original sample can deceive Deep Neural Networks (DNNs). Goodfellow et al. [32] proposed a Fast Gradient Sign Method (FGSM) to generate adversarial samples by computing the gradient of the loss function with respect to the input data. The follow-up work I-FGSM [33] performs multiple gradient updates to generate stronger perturbations. These studies proved that the existence of such adversarial examples causes the vulnerability of the DNN. Hence, researchers attempted to determine the potential vulnerability of deepfakes by interrupting them with adversarial examples. Athalye et al. [34] proposed the expectation-over-transformation (EoT) method to generate adversarial examples that can withstand preprocessing transformations. Yang et al. [35] designed a defense technology to protect users from GAN-based deepfake attacks. Ruiz et al. [12] reported spread-spectrum adversarial attacks, that can bypass blurring defense mechanisms in a gray-box scenario. The spread-spectrum adversarial attack method exhibits high transferability across different types and magnitudes of blur. Huang et al. [13] introduced an initiative defense model to safeguard facial data against manipulation. The presence of the embedded poison perturbations substantially impairs the performance of face forgery models at the inference and training stages.

In this work, we design a union-saliency attack framework to generate adversarial examples. Unlike the aforementioned methods that manipulate the entire image, the proposed model can minimize adding disturbance to the background and preserve more facial details.

D. Saliency Detection

The task of saliency detection involves the identification and localization of the most visually conspicuous area, termed as "salient region". In order to enhance the accuracy of salient object detection, numerous works [36], [37], [38], [39], [40], [41] have improved the saliency detection module. Wu et al. [42] introduced a Sample Adaptive View Transformer module which includes three different transformation branches. This configuration facilitates the acquisition of diverse features corresponding to distinct perspectives. Cong et al. [43] designed a weakly-supervised model by employing hybrid labels. These models enable the network to focus on the foreground region and disregard the background, improving its ability to perform the task of salient object detection.

In this work, we adopt a pre-trained TRACER [44] model to generate a salient mask. This model can produce a weak annotation for the facial region of interest and highlight the most important and relevant main portrait area. As a complementary to the saliency map, an assistive manipulation mask is simultaneously introduced to guide the perturbation. The manipulation mask is regarded as a region that the deepfake model altered to the original image. In this paper, the coupling of the salient mask and the manipulation mask is termed a union mask.

III. Method

A. Overview

The primary aim of the union-saliency attack proposed in this work is to efficiently impair deepfake systems by subtly perturbing the facial region in the original image. The overall framework is shown in Fig. 3. The attack aims to minimize the image information loss while achieving maximum effectiveness. This framework is designed in a four-step process:

(1) Creating union masks (Section III-B).

(2) Perturbing the facial image in a union-saliency method while ensuring imperceptibility (Section III-C).

Authorized licensed use limited to: Queen Mary University of London. Downloaded on March 06,2025 at 11:54:15 UTC from IEEE Xplore. Restrictions apply.

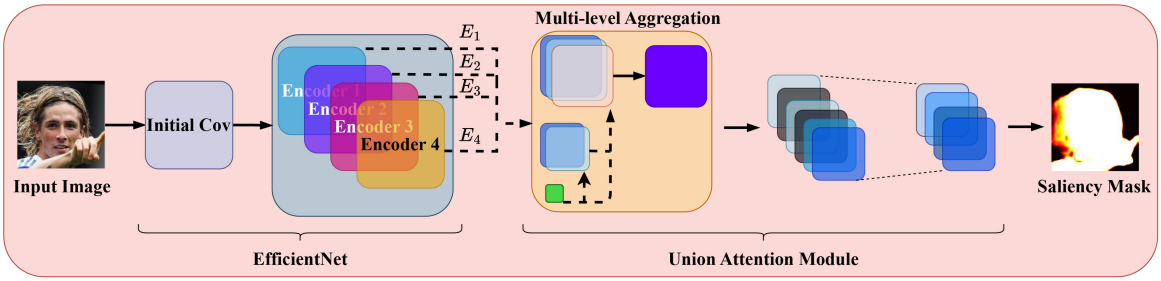


Fig. 4. The architecture of TRACER.

(3) Adding the structural similarity loss function to the original loss function so as to reduce the distortion of the input image caused by the perturbation (Section III-D).

(4) The face images were fed forward into the deepfake model (Section III-E), with the expectation that it would be readily distinguishable visually.

B. Union Mask

Deepfake models typically manipulate the face through techniques such as face swapping or attribute manipulation. In order to effectively thwart deepfake systems and minimize image distortion resulting from the introduction of perturbations, we adopt a pre-trained model to accurately isolate the facial area from the background elements that are not relevant to the face. However, due to the inability to accurately identify portrait areas with only model-generated saliency masks, a manipulation mask is introduced to guide addition perturbation with a saliency mask. The union mask is created through the combination of the saliency mask and the manipulation mask.

The saliency mask is generated by the well-trained TRACER [44] model. The architecture of TRACER is shown in Fig. 4. Specifically, the TRACER model consists of an efficient backbone encoder and attention-guided salient object-tracking modules. The TRACER model employs EfficientNet as the backbone encoder to extract multi-scale features and uses the masked edge attention module to improve memory efficiency. The union and object attention modules are integrated within the decoder to effectively combine multi-level features. The outputs of encoders are incorporated into these modules. The Union Attention Module (UAM) effectively combines multi-level features and captures significant contextual information from both channel and spatial representations. The UAM can enhance the detection performance. Furthermore, the object attention module enhances object detection and edge extraction by leveraging refined channels and spatial representations. Therefore, it exhibits an increased level of precision in generating saliency masks.

Denoting an original image as I_{in}, and a pre-trained TRACER model is defined as $f_{\text{sd}}(\cdot)$, the generated saliency mask is defined as,

$$M_{\text{sm}} = f_{\text{sd}}(n(I_{\text{in}}, \mu, \sigma), \theta_s), \tag{1}$$

where $n(\cdot)$ refers to the process of normalization. μ and σ are the mean and variance of the raw input image, respectively. θ_s in the TRACER model is a fixed value that remains unchanged during the inference process of the model.

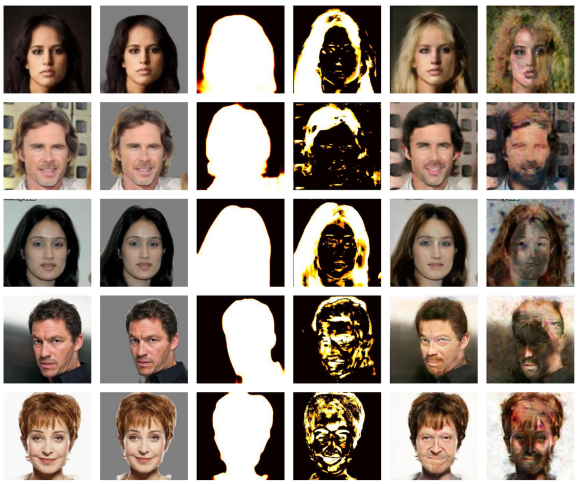


Fig. 5. Results of the intermediate process. The images from left to right are original images, saliency masks, foreground face region of images, manipulation masks, deepfake-generated images from the original images and final disrupted images by the proposed model.

The manipulation mask provides a precise facial region to guide perturbation introduction. The manipulation mask is formulated as,

$$M_{\text{ma}} = \begin{cases} 0, & \text{if } \|G(I_{\text{in}}) - I_{\text{in}}\| < 0.5, \\ \|G(I_{\text{in}}) - I_{\text{in}}\|, & \text{otherwise}, \end{cases} \tag{2}$$

where $G(I_{\text{in}})$ represents the image generated by the GAN model from the original input image I_{in}. The $G(\cdot)$ represents the GAN model. In contrast to utilizing post-processing techniques for the conversion into a binary map, regions with minimal changes in the image are assigned a value of zero, and the regions are determined by $\|G(I_{\text{in}}) - I_{\text{in}}\|$.

The union mask is a linear combination of the manipulation mask and the saliency mask. It is defined as,

$$M_{\text{un}} = \alpha \cdot M_{\text{sm}} + \beta \cdot M_{\text{ma}}, \tag{3}$$

where α and β are the weights of the saliency mask and manipulation mask, respectively. The saliency mask M_{sm} and the manipulation mask M_{ma} are shown in Fig. 5. Compared with the saliency mask, the union mask can better highlight the foreground of the image. With the union mask M_{un}, the foreground face region is obtained as,

$$I_{\text{fface}} = I_{\text{in}}(x, y) \times M_{\text{un}}(x, y), \tag{4}$$

The incorporation of the manipulation mask M_{ma} with the saliency mask M_{sm} enables the preservation of fine facial

Authorized licensed use limited to: Queen Mary University of London. Downloaded on March 06,2025 at 11:54:15 UTC from IEEE Xplore. Restrictions apply.

details. Applying the union mask to the foreground region ensures that subsequent perturbations have no effect on the background. Therefore, this model can reduce the loss of image details.

C. Union-Saliency Image Perturbation

While contemporary deepfake models exhibit enhanced performance relative to conventional methods, they remain susceptible to adversarial attacks. To disrupt a pre-trained deepfake model, a simple yet efficient solution is to adopt Section II-C assumption by disrupting the original image. The perturbed images can be generated as,

$$I_{\text{ptb}} = I_{\text{in}} + \tau, \tag{5}$$

where I_{ptb} is the perturbed image. τ is an imperceptible perturbation that can be uniform noise, Gaussian noises, or salt-and-pepper noises. However, the imperceptible disruption is randomly generated and may not be the smallest perturbation that obstructs a deepfake model. To solve this issue, a noise generator is employed to iterate the random perturbation. The objective function can be formulated as,

$$min\mathcal{L}(G(I_{\text{in}} + \tau_{\text{i}}, c)), \tag{6}$$

where i represents the number of iterations for the perturbation, and c is the facial attribute label. The τ_{i} is obtained as,

$$\tau_{\text{i}} = \tau_{\text{i-1}} - \eta \cdot \nabla_{\tau_{\text{i-1}}} \mathcal{L}(G(I_{\text{in}} + \tau_{\text{i-1}}, c)), \tag{7}$$

where η is the learning rate, and it is set to 1e-4. The final disturbance generated by the noise generator is defined as $\tilde{\tau}$. For a deepfake system, when the perturbed image I_{ptb} is input, it has the capability to output a generated image O_{ptb} which will have noticeable artifacts. This proves that adding perturbations can disturb the deepfake system.

The balance between disrupting the deepfake system and preserving image details is crucial. To mitigate the loss of original face image details and disturb the deepfake system, we present a novel union-saliency approach. Given the union mask M_{un} generated by Eq. (4), the perturbation is accomplished by redefining Eq. (7) as,

$$\tilde{\tau}_{\text{ptb}} = M_{\text{un}} \times \tilde{\tau}, \quad I_{\text{adv}} = I_{\text{in}} + \tilde{\tau}_{\text{ptb}}, \tag{8}$$

where the parameter I_{adv} represents the perturbation factor, which is added to the original image to achieve the desired level of disruption. I_{adv} is the perturbed image with a union mask.

D. Loss Function

Given a deepfake system $G(\cdot)$, the attack has two objectives, each on a separate manifold.

- The disrupted image I_{adv} is envisioned to have minimal distortion while effectively disrupting the deepfake system, *i.e.*, the introduced perturbation is supposed to be invisible, which is expressed as,

$$\min_{\tau_{\text{ptb}}} \mathcal{L}\big(I_{\text{in}} + \tau_{\text{ptb}}\big), \quad \text{s.t.} \|\tau_{\text{ptb}}\|_\infty \leq \epsilon. \tag{9}$$

- The deepfake-generated images $G(I_{\text{in}})$ from the original images and the deepfake-generated disrupted image $G(I_{\text{adv}})$ is projected to be highly unnatural contrasting with $G(I_{\text{in}})$, *i.e.*, the introduced perturbation effectively induces degradation in the generated image, which is expressed as,

$$\max_{\tau_{\text{ptb}}} \mathcal{L}\big(G(I_{\text{in}}), G(I_{\text{in}} + \epsilon_{\text{ptb}})\big), \quad \text{s.t.} \|\tau_{\text{ptb}})\|_\infty \leq \epsilon. \tag{10}$$

To mitigate distortions arising from the introduction of perturbations to images, we introduce a structural similarity (SSIM) loss [45] based on the original loss function. The original loss is formulated as,

$$\mathcal{L}_1(G(I_{\text{adv}}), G(I_{\text{in}})) = \frac{\sum_{i=1}^{n} (G(I_{\text{adv}} - G(I_{\text{in}}))^2}{n}. \tag{11}$$

The structural similarity loss is frequently utilized as an image quality metric to calculate the likeness between two images. It is commonly employed as a loss function for image processing. The SSIM loss is formulated as

$$\begin{aligned} SSIM(I_{in}, I_{adv}) &= \frac{\big(2\mu_{I_{in}}\mu_{I_{adv}} + \lambda_1\big)\big(2\sigma_{I_{in}I_{adv}} + \lambda_2\big)}{\Big(\mu_{I_{in}}^2 + \mu_{I_{adv}}^2 + \lambda_1\Big)\Big(\sigma_{I_{in}}^2 + \sigma_{I_{adv}}^2 + \lambda_2\Big)}, \\ \mathcal{L}_{\text{ssim}}(I_{\text{adv}}, I_{\text{in}}) &= 1 - SSIM(I_{in}, I_{adv}), \end{aligned} \tag{12}$$

where μ and σ denote the mean and variance. The values of λ_1 and λ_2 are set as 1e-4 and 9e-4, respectively. Unlike the original loss, the SSIM loss focuses on the relationship between I_{in} and I_{adv} in luminance, contrast, and structure. The overall loss function is formulated as,

$$\mathcal{L} = \mathcal{L}_1(G(I_{\text{adv}}), G(I_{\text{in}})) + \gamma * \mathcal{L}_{\text{ssim}}(I_{\text{adv}}, I_{\text{in}}), \tag{13}$$

where the parameter γ is a weight of the SSIM loss.

E. Deepfake Attack Model

The aim of a deepfake attack model is to trick a deepfake system $G(\cdot)$ that has been trained with parameters θ_g to generate a visually authentic facial image $g(I_{\text{in}})$ using the original image I_{in}. To this aim, the model should be fooled by a disrupted image I_{adv}, which is known as an adversarial attack. The Fast Gradient Signed Method (FGSM) [32] is a well-established technique to attack neural networks. It can modify the input through the addition of a small noise in the direction of the loss gradient with respect to the input data. It is formulated as,

$$I_{\text{adv}} = I_{\text{in}} + m \,\text{sign}\big(\nabla_x \mathcal{L}(\theta_g, x, I_{\text{in}})\big), \tag{14}$$

where $\nabla_x \mathcal{L}$ is the loss function and m denotes the size of the FGSM step. The I-FGSM [33] further improves the FGSM algorithm by iteratively applying small perturbations to the input in the direction of the gradient of the loss. It is mathematically denoted as,

$$\begin{aligned} I_{\text{adv}}^0 &= I_{\text{in}}, \\ I_{\text{adv}}^t &= I_{\text{adv}}^{t-1} + a \,\text{sign}\Big(\nabla_x \mathcal{L}(\theta_g, I_{\text{adv}}^{t-1}, y)\Big), \end{aligned} \tag{15}$$

where a is the step size of the I-FGSM.

The perturbed image is inevitably distorted after adding perturbations to the original image. To address this issue, we propose a union-saliency attack. The core idea of this

Authorized licensed use limited to: Queen Mary University of London. Downloaded on March 06,2025 at 11:54:15 UTC from IEEE Xplore. Restrictions apply.

TABLE I
COMPARATIVE RESULTS IN L^1 ERROR AND L^2 ERROR. THE BEST RESULTS ARE HIGHLIGHTED IN **BOLD**

Attacker	Tiny (500)		Small (1000)		Middle (2000)		Large (4000)		Average	
	$L^1 \downarrow$	$L^2 \downarrow$	$L^1 \downarrow$	$L^2 \downarrow$	$L^1 \downarrow$	$L^2 \downarrow$	$L^1 \downarrow$	$L^2 \downarrow$	$L^1 \downarrow$	$L^2 \downarrow$
I-FGSM [33]	0.019	0.087	0.019	0.087	0.019	0.087	0.019	0.087	0.019	0.087
EoT-Blur [34]	0.020	0.089	0.020	0.089	0.020	0.089	0.020	0.089	0.020	0.089
Spread-Spectrum [12]	0.017	0.071	0.017	0.071	0.017	0.071	0.017	0.071	0.017	0.071
Saliency-Aware [14]	0.016	0.068	0.016	0.068	0.016	0.068	0.016	0.068	0.016	0.068
Union-Saliency	**0.015**	**0.058**	**0.015**	**0.058**	**0.015**	**0.058**	**0.015**	**0.058**	**0.015**	**0.058**

Algorithm 1 Pseudocode of the Union-Saliency Adversarial Attack Model

Input: Deepfake generator $G(\cdot)$, Pre-trained saliency detection model $f_{\text{sd}}(\cdot)$, Input face image I_{in}, Disruption magnitude a.
Output: Disrupted face image I_{adv}, generated image from the disrupted counterpart $g(I_{\text{adv}})$.

Using saliency detection to generate saliency mask M_{sm} as Eq. (1).
Generating the union mask M_{un} as Eq. (3).
Using noise generator and the union mask M_{un} to generate perturbation $\tilde{\tau}_{\text{ptb}}$ as Eq. (8).
Applying perturbation $\tilde{\tau}_{\text{ptb}}$ on the face region to get the residual component by I-FGSM algorithm by Eq. (16).
Feeding the union-saliency perturbed face image I_{adv} into the deepfake generator to get the translated image $g(I_{\text{adv}})$.

method is to constrain the disturbance in the salient face area through the union mask. Meanwhile, the SSIM loss function is employed to constrain the differences between the original image and its counterpart with added noise. This process is formulated as,

$$\begin{aligned} I^0_{\text{adv}} &= I_{\text{in}} + \tilde{\tau}_{\text{ptb}}, \\ I^{t+1}_{\text{adv}} &= I^t_{\text{adv}} + a\ \text{sign}\big(\nabla_x \mathcal{L}(\theta_g, I^t_{\text{adv}}, y)\big) \times M_{\text{un}}, \end{aligned} \tag{16}$$

where the $\mathcal{L}(\cdot)$ is formulated as Eq. (13), and "$\times$" denotes the matrix multiplication operation.

Following the recent work [12], a Gaussian smoothing filter is employed to blur the input image at each iteration so as to bootstrap the attacker. The pseudocode of the proposed union-saliency adversarial attack model is summarized in Algorithm 1.

IV. EXPERIMENTS AND ANALYSIS

A. Implementation Details

In the union-saliency attack framework, the TRACER [44] is employed to generate the saliency mask. The kernel size is set to 11 and σ is set to 1. A Gaussian blur filter is employed as a pre-processing step for the image. The magnitude of perturbation in Eq. (5) is set to 0.05. The step size a in Eq. (16) is set to 0.01 [12], and the loss function for the union-saliency framework in Eq. (13) is a combination of MSE loss and SSIM loss. We empirically set the parameter α to 1.0, β to 0.1 and γ to 1.0. The framework is implemented in PyTorch [46] framework with an NVIDIA GeForce GTX 3090Ti GPU.

B. Datasets

The experiments are performed on the CelebFaces Attributes (CelebA) dataset [47]. It is an openly available dataset consisting of 202,599 face images provided by 10,177 celebrities. We conduct a comprehensive evaluation of the proposed methods by sampling images at various scales, including a small scale (500 images), a medium scale (1,000 images), an intermediate scale (2,000 images), and a large scale (4,000 images).

C. Evaluation Metrics

To validate the effectiveness of the proposed union-saliency attack framework, we conduct an evaluation objectively and subjectively. For objective evaluation, four evaluation metrics, *i.e.,* L^1 error, L^2 error, peak signal-to-noise ratio (PSNR) and structural similarity index measure (SSIM), are adopted to measure the difference between the original face images and the disrupted image. For subjective evaluation, we visually depicted the generated images produced by the deepfake system using both I_{in} and I_{adv} as inputs and visualize the generated image by the deepfake framework from the pair of these two inputs, respectively.

D. Comparison With State-of-the-Art Methods

In order to validate the effectiveness of the proposed union-saliency attack framework, we compared it with the state-of-the-art (SOTA) competitors under the same configuration [12], [14], [33], [34].

For objective evaluation, the results of L^1 and L^2 errors are shown in Table I. Compared with the second-best Saliency-Aware [14], the proposed method reduces L^1 and L^2 errors by 6.2% and 14.7%, respectively. Moreover, to further evaluate the similarity between the original and the disturbed images comprehensively, the results of the PSNR and the SSIM are shown in Table II. The results prove that the proposed framework outperforms all other competitors in terms of both the PSNR and the SSIM. Specifically, compared with the second-best method, the proposed framework improves the PSNR and SSIM by 2.1% and 0.9%, respectively. This signifies that the proposed method merely requires less noise injection into the original image to achieve effective disruption of the deepfake mode. Table I and Table II prove that the proposed attack framework outperforms other competitors, on the same set of benchmark settings with four evaluation metrics. While both the proposed method and the second-best Saliency-Aware model [14] integrate saliency masks, the union-saliency attack framework goes a step further by introducing the union masks. This strategic enhancement leads to a more comprehensive representation of facial features in the framework. Therefore, the proposed attack method

Authorized licensed use limited to: Queen Mary University of London. Downloaded on March 06,2025 at 11:54:15 UTC from IEEE Xplore. Restrictions apply.

TABLE II
COMPARATIVE RESULTS IN *PSNR* AND *SSIM*. THE BEST RESULTS ARE HIGHLIGHTED IN **BOLD**

Attacker	Tiny (500)		Small (1000)		Middle (2000)		Large (4000)		Average	
	$PSNR\uparrow$	$SSIM\uparrow$	$PSNR\uparrow$	$SSIM\uparrow$	$PSNR\uparrow$	$SSIM\uparrow$	$PSNR\uparrow$	$SSIM\uparrow$	$PSNR\uparrow$	$SSIM\uparrow$
I-FGSM [33]	33.614	0.942	33.616	0.943	33.616	0.943	33.616	0.943	33.616	0.943
EoT-Blur [34]	33.536	0.937	33.538	0.938	33.536	0.938	33.536	0.938	33.536	0.938
Spread-Spectrum [12]	34.521	0.954	34.520	0.955	34.520	0.954	34.520	0.954	34.520	0.954
Saliency-Aware [14]	34.706	0.954	34.709	0.955	34.709	0.955	34.709	0.955	34.708	0.955
Union-Saliency	**35.427**	**0.964**	**35.430**	**0.964**	**38.446**	**0.964**	**35.446**	**0.964**	**35.437**	**0.964**

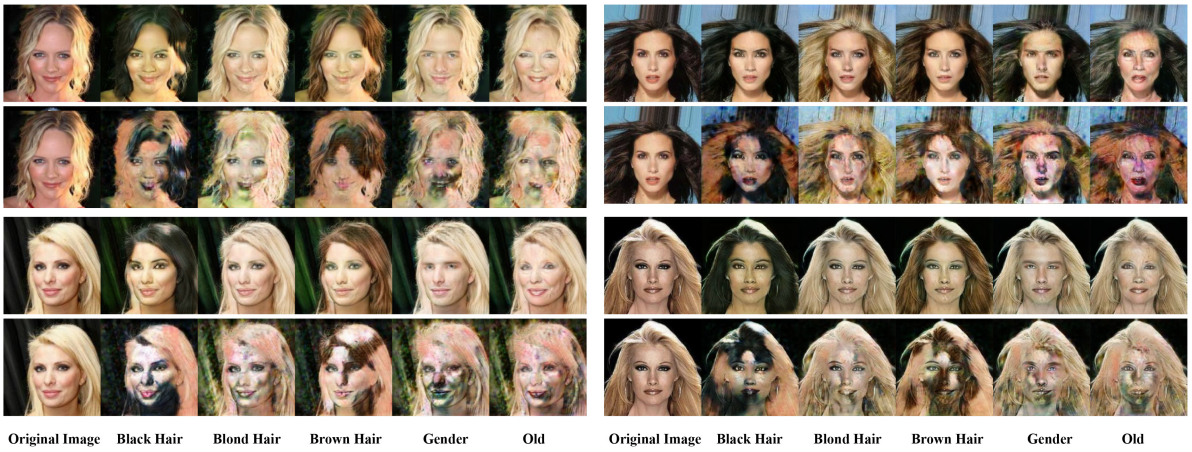


Fig. 6. Visualized results with different facial attributions.

TABLE III
ABLATION STUDIES ON THE KEY COMPONENTS AND THE PROPOSED LOSS FUNCTION. THE BEST RESULTS ARE HIGHLIGHTED IN **BOLD**

Methods	$L^1\downarrow$	$L^2\downarrow$	$PSNR\uparrow$	$SSIM\uparrow$
baseline	0.0169	0.0714	34.5212	0.9541
baseline+Saliency mask	0.0167	0.0697	34.6044	0.9544
baseline+Union mask	0.0167	0.0703	34.5748	0.9531
baseline+Union mask+Noise generator	0.0167	0.0702	34.5756	0.9532
baseline+Saliency mask+Noise generator+L_{ssim}	0.0149	0.0584	35.3852	0.9635
baseline+Union mask+Noise generator+L_{ssim}	**0.0147**	**0.0580**	**35.4274**	**0.9636**

is able to further reduce the distortion of the image. In consequence, the solid foundation provided by these results is unquestionably essential for subsequent deepfake attack methodologies.

The visualized results are depicted in Fig. 6. The original images and disrupted images are processed by the widely adopted StarGAN model [17] to perform operations on different attributes. The subjective results demonstrate that the image generated by feeding the disrupted image into StarGAN has noticeable artifacts, and these images can be easily distinguished visibly. Additionally, the concentration of artifacts in the analyzed images is primarily localized in the facial region, with a notably lower occurrence of artifacts in non-facial images. This finding proves that the proposed framework is capable of effectively disrupting the deepfake model.

E. Ablation Studies

To validate the efficacy of the key components and the adopted loss function, the ablation experiments are conducted. Comparative results are listed in Table III. The corresponding counterparts are depicted as follows.

- "baseline" is Spread-Spectrum attack [12]. The scores of L^1 error, L^2error, *PSNR* and *SSIM* are 0.0169, 0.0714, 34.5212 and 0.9541, respectively.
- "baseline+Saliency mask" denotes the baseline model with a saliency mask which is generated by TRACER module [44]. Compared to the baseline model, adding the saliency mask can reduce L^1 and L^2 errors by 1.2 % and 2.4%, respectively.
- "baseline+Union mask" indicates the baseline model with a union mask, which generated by adding a saliency feature mask and a manipulation mask. It shows that adopting the union mask yields better performance than the baseline.
- "baseline+Union mask+Noise generator" represents the baseline by adding a union mask and noise generator. It shows that the performance outperforms the baseline by adding a single union mask in L^2 error, PSNR and SSIM.
- "baseline+Saliency mask+Noise generator+L_{ssim}" indicates the baseline by adding a saliency mask, noise generator, and the SSIM loss. The scores of L1 error, L2 error, PSNR and SSIM are 0.0149, 0.0584, 35.3852, and 0.9636, respectively.

Authorized licensed use limited to: Queen Mary University of London. Downloaded on March 06,2025 at 11:54:15 UTC from IEEE Xplore. Restrictions apply.

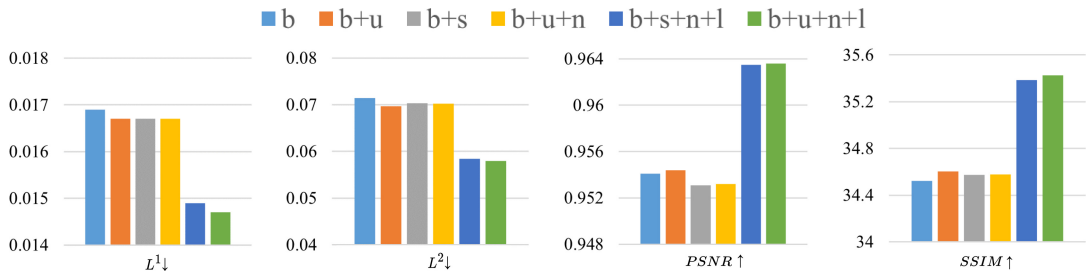


Fig. 7. Ablation study on different components. "b": "baseline", "b+s": "baseline+Saliency mask", "b+u": "baseline+Union mask", "b+u+n": "baseline+Union mask+Noise generator", "b+s+n+l": "baseline+Saliency mask+Noise generator+L_{ssim}", "b+u+n+l": "baseline+Union mask+Noise generator+L_{ssim}".

- "baseline+Union mask+Noise generator+L_{ssim}" denotes the baseline by adding a union mask, noise generator, and the SSIM loss, which is the proposed framework. The scores of L^1 error, L^2 error, PSNR and SSIM are 0.0147, 0.0580, 35.4274, and 0.9636, respectively. The results indicate that the performance of the proposed framework surpasses that of other combination models in terms of L^1 error, L^2 error, PSNR, and SSIM metrics.

Table III proves that the individual components within the proposed framework interact synergistically. The synergistic interplay among these modules contributes to a substantial enhancement in network performance. The performance comparison of the ablation studies is shown in Fig. 7.

V. Conclusion

In this paper, we propose a union-saliency attack framework to provide an effective and practical solution to counter deepfake models while keeping image modifications to a minimum. It employs a union mask to enhance the focus on the facial features of individuals while adding perturbations to these specific details rather than the irrelevant background regions. This strategy introduces minimal changes to the original image, with making the attack remain effective. Experiments were conducted to showcase the effectiveness of the proposed method, revealing its superiority over SOTA competitors. We consider the proposed model to have the potential to safeguard the privacy of individuals and address ethical concerns by mitigating the harm caused by deepfake technology.

References

[1] R. Tolosana, R. Vera-Rodriguez, J. Fierrez, A. Morales, and J. Ortega-Garcia, "Deepfakes and beyond: A survey of face manipulation and fake detection," *Inf. Fusion*, vol. 64, pp. 131–148, Dec. 2020.

[2] C.-C. Hsu, Y.-X. Zhuang, and C.-Y. Lee, "Deep fake image detection based on pairwise learning," *Appl. Sci.*, vol. 10, no. 1, p. 370, Jan. 2020.

[3] 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.*, 2021, pp. 15003–15013.

[4] Z. Yang, J. Liang, Y. Xu, X.-Y. Zhang, and R. He, "Masked relation learning for deepfake detection," *IEEE Trans. Inf. Forensics Security*, vol. 18, pp. 1696–1708, 2023.

[5] X. Li, R. Ni, P. Yang, Z. Fu, and Y. Zhao, "Artifacts-disentangled adversarial learning for deepfake detection," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 4, pp. 1658–1670, Apr. 2023.

[6] J. Wang, Y. Sun, and J. Tang, "LiSiam: Localization invariance Siamese network for deepfake detection," *IEEE Trans. Inf. Forensics Security*, vol. 17, pp. 2425–2436, 2022.

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

[8] 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.*, 2022, pp. 1–13.

[9] R. Wang, F. Juefei-Xu, M. Luo, Y. Liu, and L. Wang, "Faketagger: Robust safeguards against deepfake dissemination via provenance tracking," in *Proc. 29th ACM Int. Conf. Multimedia*, 2021, pp. 3546–3555.

[10] L. Jiang, W. Wu, C. Qian, and C. C. Loy, "DeepFakes detection: The DeeperForensics dataset and challenge," in *Handbook of Digital Face Manipulation Detection*. Cham, Switzerland, Springer, 2022, pp. 303–329.

[11] 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*, 2020, pp. 53–62.

[12] N. Ruiz, S. A. Bargal, and S. Sclaroff, "Disrupting deepFakes: Adversarial attacks against conditional image translation networks and facial manipulation systems," in *Proc. ECCV Workshops*, 2020, pp. 236–251.

[13] Q. Huang, J. Zhang, W. Zhou, W. Zhang, and N. Yu, "Initiative defense against facial manipulation," in *Proc. AAAI Conf. Artif. Intell.*, vol. 35, 2021, pp. 1619–1627.

[14] Q. Li, M. Gao, G. Zhang, and W. Zhai, "Defending Deepfakes by saliency-aware attack," *IEEE Trans. Comput. Social Syst.*, early access, May 8, 2023, doi: 10.1109/TCSS.2023.3271121.

[15] I. Goodfellow et al., "Generative adversarial nets," in *Proc. NeurIPS*, vol. 1, 2014, pp. 1–8.

[16] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. ICCV*, 2017, pp. 2242–2251.

[17] 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*, 2018, pp. 1–14.

[18] 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.

[19] T. Karras et al., "Alias-free generative adversarial networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, pp. 852–863, 2021.

[20] 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.*, 2019, pp. 4401–4410.

[21] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, "Analyzing and improving the image quality of stylegan," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 8110–8119.

[22] D. Bau et al., "Semantic photo manipulation with a generative image prior," *ACM Trans. Graph.*, vol. 38, no. 4, pp. 1–11, Jul. 2019.

[23] Z. Liu, X. Qi, and P. H. Torr, "Global texture enhancement for fake face detection in the wild," in *Proc. IEEE/CVF Conf. Comput. Vis. pattern Recognit.*, 2020, pp. 8060–8069.

[24] A. Khodabakhsh, R. Ramachandra, and C. Busch, "Subjective evaluation of media consumer vulnerability to fake audiovisual content," in *Proc. 11th Int. Conf. Qual. Multimedia Exp. (QoMEX)*, 2019, pp. 1–6.

[25] Y. Shen, J. Gu, X. Tang, and B. Zhou, "Interpreting the latent space of GANs for semantic face editing," in *Proc. IEEE/CVF Conf. Comput. Vis. pattern Recognit.*, 2020, pp. 9243–9252.

Authorized licensed use limited to: Queen Mary University of London. Downloaded on March 06,2025 at 11:54:15 UTC from IEEE Xplore. Restrictions apply.

[26] 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.
[27] M. Liu et al., "STGAN: A unified selective transfer network for arbitrary image attribute editing," in *Proc. IEEE/CVF Conf. Comput. Vis. pattern Recognit.*, 2019, pp. 3673–3682.
[28] Y. Gao et al., "High-fidelity and arbitrary face editing," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 16115–16124.
[29] N. Carlini and D. Wagner, "Towards evaluating the robustness of neural networks," in *Proc. IEEE Symp. Security Privacy (SP)*, 2017, pp. 39–57.
[30] A. Nguyen, J. Yosinski, and J. Clune, "Deep neural networks are easily fooled: High confidence predictions for unrecognizable images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 427–436.
[31] 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. Security Privacy (EuroS P)*, 2016, pp. 372–387.
[32] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," 2014, *arXiv:1412.6572*.
[33] A. Kurakin, I. J. Goodfellow, and S. Bengio, "Adversarial examples in the physical world," in *Artificial Intelligence Safety and Security*. Boca Raton, FL, USA: Chapman and Hall/CRC, 2018, pp. 99–112.
[34] A. Athalye, L. Engstrom, A. Ilyas, and K. Kwok, "Synthesizing robust adversarial examples," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 284–293.
[35] C. Yang, L. Ding, Y. Chen, and H. Li, "Defending against GAN-based deepfake attacks via transformation-aware adversarial faces," in *Proc. Int. Jt. Conf. Neural Netw. (IJCNN)*, 2021, pp. 1–8.
[36] P. Zhang, D. Wang, H. Lu, H. Wang, and X. Ruan, "Amulet: Aggregating multi-level convolutional features for salient object detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 202–211.
[37] X. Qin, Z. Zhang, C. Huang, C. Gao, M. Dehghan, and M. Jagersand, "BasNet: Boundary-aware salient object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 7471–7481.
[38] J. Wei, S. Wang, Z. Wu, C. Su, Q. Huang, and Q. Tian, "Label decoupling framework for salient object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. pattern Recognit.*, 2020, pp. 13025–13034.
[39] Y. Pang, Y. Li, J. Shen, and L. Shao, "Towards bridging semantic gap to improve semantic segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 4230–4239.
[40] J. Wei, S. Wang, and Q. Huang, "F^3Net: Fusion, feedback and focus for salient object detection," in *Proc. AAAI Conf. Artif. Intell.*, vol. 34, 2020, pp. 12321–12328.
[41] Y. Pang, X. Zhao, L. Zhang, and H. Lu, "Multi-scale interactive network for salient object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 9413–9422.
[42] J. Wu, C. Xia, T. Yu, and J. Li, "View-aware salient object detection for 360° omnidirectional image," *IEEE Trans. Multimedia*, vol. 25, pp. 6471–6484, Sep. 2022.
[43] R. Cong et al., "A weakly supervised learning framework for salient object detection via hybrid labels," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 2, pp. 534–548, Feb. 2023.
[44] M. S. Lee, W. Shin, and S. W. Han, "TRACER: Extreme attention guided salient object tracing network (student abstract)," in *Proc. AAAI Conf. Artif. Intell.*, vol. 36, 2022, pp. 12993–12994.
[45] 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.
[46] A. Paszke et al., "Automatic differentiation in PyTorch," in *Proc. NIPS-W*, 2017, pp. 1–4.
[47] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *Proc. ICCV*, 2015, pp. 3730–3738.

Guisheng Zhang is pursuing the M.S. degree with the School of Electrical and Electronic Engineering, Shandong University of Technology, Zibo, China. His research interests include deepfake detection and deep learning.

Mingliang Gao received the Ph.D. degree in communication and information systems from Sichuan University. He is currently an Associate Professor and the Vice Dean with the Shandong University of Technology. He was a Visiting Lecturer with the University of British Columbia from 2018 to 2019. He has published over 150 journal/conference papers in IEEE, Springer, Elsevier, and Wiley. His research interests include computer vision, machine learning, and intelligent optimal control.

Qilei Li (Graduate Student Member, IEEE) received the M.S. degree from Sichuan University in 2020. He is currently pursuing the Ph.D. degree in computer science from the Queen Mary University of London, supervised by Prof. S. Gong. His research interests include computer vision and deep learning, particularly focusing on person ReID, video/image enhancement.

Wenzhe Zhai received the M.S. degree from the School of Electrical and Electronic Engineering, Shandong University of Technology, Zibo, China, in 2023. He is currently pursuing the Doctoral degree with Harbin Engineering University. His research interests include smart city systems, information fusion, crowd analysis, and deep learning.

Guofeng Zou received the Ph.D. degree in pattern recognition and intelligent systems from the College of Automation, Harbin Engineering University, Harbin, in 2013. He is currently working as an Associate Professor with the College of Electrical and Electronic Engineering, Shandong University of Technology, Zibo, China. He has published more than 50 papers in major journals and conferences. His current research interests include pattern recognition, digital image processing and analysis, and machine learning.

Gwanggil Jeon (Senior Member, IEEE) received the Ph.D. degree from the Department of Electronics and Computer Engineering, Hanyang University, Seoul, South Korea, in 2008. From September 2009 to August 2011, he was with the School of Information Technology and Engineering, University of Ottawa, Ottawa, ON, Canada, as a Postdoctoral Fellow. He serves as a Full Professor with the Shandong University of Technology, Zibo, China, and Incheon National University, Incheon, South Korea. His research interests include computer vision, machine learning, and Internet of Things. He is a recipient of the IEEE Chester Sall Award in 2007, the *ETRI Journal* Paper Award in 2008, and the Industry-Academic Merit Award by the Ministry of SMEs and Startups of Korea Minister in 2020.

Authorized licensed use limited to: Queen Mary University of London. Downloaded on March 06,2025 at 11:54:15 UTC from IEEE Xplore. Restrictions apply.