A Local Perturbation Generation Method for GAN-generated Face Anti-forensics

Haitao Zhang, Beijing Chen, Jinwei Wang, Guoying Zhao, *Fellow, IEEE*

***Abstract*—Although the current generative adversarial networks (GAN)-generated face forensic detectors based on deep neural networks (DNNs) have achieved considerable performance, they are vulnerable to adversarial attacks. In this paper, an effective local perturbation generation method is proposed to expose the vulnerability of state-of-the-art forensic detectors. The main idea is to mine the fake faces' areas of common concern in multiple-detectors' decision-making, then generate local anti-forensic perturbations by GANs in these areas to enhance the visual quality and transferability of anti-forensic faces. Meanwhile, in order to improve the anti-forensic effect, a double-mask (soft mask and hard mask) strategy and a three-part loss (the GAN training loss, the adversarial loss consisting of ensemble classification loss and ensemble feature loss, and the regularization loss) are designed for the training of the generator. Experiments conducted on fake faces generated by StyleGAN demonstrate the proposed method's advantage over the state-of-the-art methods in terms of anti-forensic success rate, imperceptibility, and transferability. The source code is available at https://github.com/imagecbj/A-Local-Perturbation-Generation-Method-for-GAN-generated-Face-Anti-forensics.**

***Index Terms*—local perturbation, generated face, anti-forensics, generative adversarial network, double mask.**

I. Introduction

IMAGE synthesis technological advances have increasingly enabled the generation of forged faces visually realistic. The emergence of generative adversarial networks (GANs) [1] in particular makes it hard for human eyes to tell a real face from a fake one. As human faces have been widely used in facial recognition and biometric authentication, the broad dissemination of fake faces may cause ethical, social and security issues.

Therefore, researchers have proposed various forensic methods [2-10] to detect the authenticity of face images by distinguishing between fake faces and real faces. Although the considerable performance has been obtained by the current forensic detectors on multiple benchmark datasets, these detectors are vulnerable to adversarial attacks. Sophisticated malicious forgers may try to exploit these loopholes to generate faces that can bypass the detectors while maintaining high visual quality of face images. These detected and certified images may do more harm as they spread. Therefore, some researchers have investigated counter-measures [11-13] against forensics, called anti-forensics, to expose the vulnerability of the current forensic detectors. The study of anti-forensics could enable researchers to develop advanced forensic methods against forgery technologies. In addition, it can help prevent anonymous faces from being detected while ensuring the visual quality in the field of image anonymity based on face synthesis [14-16].

Regarding anti-forensics, in recent years, it has been shown that forensic detectors based on deep neural networks (DNNs) are vulnerable to adversarial perturbations [17-21]. By adding carefully designed and imperceptible anti-forensic noise to the fake face image, the forensic detectors are rendered ineffective. However, the existing attack-based methods [17-21] use the method of perturbing all pixels, and there are many redundant and meaningless perturbations. In addition, the transferability of these methods is still insufficient.

In order to further obtain a good trade-off between the visual quality and transferability of anti-forensic faces, we propose an effective local perturbation generation method. The predicted local areas are meaningful regions for the decision-making of the forensic detector. For the generation of anti-forensic perturbations, the GAN architecture is exploited due to its fast generation speed. Meanwhile, owing to the effectiveness of the ensemble model [22-23] and latent feature [24-26] to improve the ability of adversarial attacks, these means are combined to improve the anti-forensic effect.

In this paper, our contributions are as follows:

● We propose an invisible local region perturbation method to fool many state-of-the-art GAN-generated face forensic CNNs. A good trade-off between the visual quality and transferability of anti-forensic images is obtained by mining the areas of common concern of multiple detectors and adding adversarial noises in these areas.

● We reveal that the forensic detectors have areas of common concern when making decisions. Then, we utilize multiple detectors to predict the soft-masks and hard-masks of perturbed regions and design a double mask guided local perturbation generation method.

● In order to train an effective anti-forensic perturbation generator, in addition to the GAN training loss, the adversarial loss consisting of ensemble classification loss and ensemble feature loss, and the regularization loss are designed to guide the optimization and ensure anti-forensic ability and visual quality.

This work was supported in part by the National Natural Science Foundation of China under Grants 62072251, and 62072250; in part by the Priority Academic Program Development of Jiangsu Higher Education Institutions (PAPD) fund.

Haitao Zhang, Beijing Chen (corresponding author), and Jinwei Wang are with the Engineering Research Center of Digital Forensics, Ministry of Education, Nanjing University of Information Science and Technology, Nanjing 210044, China, also with School of Computer Science, Nanjing University of Information Science and Technology, Nanjing 210044, China, and also with Jiangsu Collaborative Innovation Center of Atmospheric Environment and Equipment Technology (CICAEET), Nanjing University of Information Science and Technology, Nanjing 210044, China (e-mail: yzzhanghaitao@163.com, nbutimage@126.com, wjwei_2004@163.com).

Guoying Zhao is with the Center for Machine Vision and Signal Analysis, University of Oulu, Oulu FI-90014, Finland (e-mail: guoying.zhao@oulu.fi).

II. RELATED WORKS

In this section, we first briefly discuss the existing technologies for generating fake face. Then, the existing anti-forensic methods are systematically summarized. The method proposed in this paper is based on local adversarial perturbation to realize anti-forensics. Therefore, finally, we introduce the relevant local adversarial attack methods in other image classification tasks.

A. Fake Face Generation and Forensics

Fake Face Generation. The GAN framework was first proposed by Goodfellow *et al.* in 2014 [1]. It can be described as an adversarial training between a generator and discriminator. Specific to fake face generation, Karras *et al.* proposed a network structure, called PGGAN [27], using a progressive training strategy to generate high-resolution images from low-resolution images. Later, inspired by image style transfer, they designed a new generator architecture, StyleGAN [28], by decoupling and separating high-level semantic attributes of images through unsupervised learning. Over time, GAN-generated faces have come ever closer to natural images and have been able to ‘mix the spurious with the genuine’.

Fake Face Forensics. Traditional image forensic methods are usually based on specific traces left by a certain forgery method. However, their effectiveness is easily affected by forgery methods and changes in data distribution. Recently, deep learning-based approaches have proven capable of handling more sophisticated forgery methods. Yamagishi *et al.* [5] proposed a network called MesoNet that has fewer layers and focused on the mesoscopic features of images to detect tampering. Wang *et al.* [6] used the residual network ResNet50 as a detection network and achieved good generalization under various GAN-generated face datasets. Liu *et al.* [7] added the Gram matrix module to ResNet18 to enhance its ability to extract global texture features. Rossler *et al.* [8] showed that the Xception network model outperforms other models on the task of fake face classification. Chen *et al.* [9] integrated of luminance and chrominance component and showed its effectiveness for GAN-generated face detection. Wang *et al.* [10] proposed an attention-based data augmentation framework to improve the ability of face forgery detection. In practice, many advanced network architectures in computer vision also achieve good performance, such as EfficientNet [29], DenseNet [30], and others.

B. Image Anti-forensics

In its early stages, research on image anti-forensics mainly focused on hiding tampering artifacts such as JPEG compression [31] and median filtering [13]. With the development of face forgery and generation models, some achievements have been made in anti-forensic researches on image forgery, especially fake face generation. In the anti-forensic process, detectors with knowledge such as architecture and parameters are called white-box detectors, while zero-knowledge detectors are called black-box detectors. For an anti-forensics task to be successful, the perturbed face must have a good visual quality and strong transferability. Transferability means that anti-forensic images generated by white-box detectors can also fool black-box detectors. At present, current research on anti-forensics seeks to improve the transferability and visual quality of anti-forensic images. From the perspective of algorithmic thinking, they can be divided into two categories: reconstruction and attack. Table I shows a summary of relevant studies on image anti-forensics.

From a reconstruction perspective, Nguyen *et al.* [32] proposed a method to fool detectors by transforming computer-generated images into new images with encoding features of natural images. Peng *et al.* [33] proposed a new generative adversarial network architecture, CGR-GAN, to resolve the problem of insufficient color, lack of texture details, and light changes of the work [32]. Peng *et al.* [34] subsequently proposed the BDC-GAN structure to realize bidirectional conversion between natural and computer-generated images. Neves *et al.* [11] tried to remove “GAN fingerprints” belonging to high-frequency signals using an autoencoder acting as a nonlinear low-pass filter. Zhao *et al.* [12] achieved good anti-forensic performance by attacking ensemble detectors with a GAN structure. Huang *et al.* [35] reduced artifact patterns of GAN-generated images based on dictionary learning.

From an attack point of view, adding carefully crafted perturbations to GAN-generated face images can render forensic detectors ineffective. Wang *et al.* [17] applied the fast gradient sign method (FGSM) [36] and the momentum iterative gradient sign method (MI-FGSM) [23] attacks to face images. Since they found that most of the perturbations in the YCbCr color space were concentrated in the Y channel, they assigned more perturbations to the Cb and Cr channels to improve the visual quality of anti-forensic images. Goebel *et al.* [18] utilized an optimization-based attack to incorporate the co-occurrence matrices taken from real images into GAN-generated images. Li *et al.* [19] performed an iterative attack called project gradient descent (PGD) [37] on the input latent vector and noise of the trained StyleGAN model to directly generate anti-forensic images. Ding *et al.* [20] proposed a new GAN structure with multiple generators and discriminators to improve visual quality of the anti-forensic images. Hussain *et al.* [21] utilized input transformation and PGD attack to generate adversarial deepfakes that are robust to JPEG compression.

The reconstruction-based methods are generally based on learning natural image features or removing forged traces. However, the visual effect and anti-forensic performance of the reconstructed images still require improved. The attack-based methods are realized by adding adversarial perturbations to the clean fake faces. Such methods retain the image details of faces to a certain extent, and the perturbations have the characteristic of transferability. Therefore, attack-based methods can better ensure anti-forensic performance. However, all of these methods use the method of perturbing all pixels, and there are many redundant and meaningless perturbations. In addition, the transferability these attack-based methods [17-21] is still insufficient. Therefore, the proposed method utilizes local perturbation method to further obtain a good trade-off between

TABLE I
THE SUMMARY OF RELEVANT RESEARCHES ON IMAGE ANTI-FORENSICS

Category	Methods	Description	Advantages	Disadvantages
Reconstruction	Nguyen et al. [32]	Use two autoencoders and a transformer net to transform the computer-generated images into new images with encoding features of natural images.	Transformed images look natural.	Nearly 50% of the transformed images can still be detected.
	Peng et al. [33]	Use style transfer to regenerate the computer-generated images.	Anti-forensic images possess the style of natural images and most of them can deceive detectors.	Anti-forensic ability on computer-generated faces with black-skin is poor.
	Peng et al. [34]	Use GAN technology to learn the bidirectional conversion between computer-generated images and natural images.	Transformed images achieve good anti-forensic performance.	Visual quality of transformed computer-generated images is affected by natural images.
	Neves et al. [11]	Use an autoencoder as a nonlinear low-pass filter to remove "GAN fingerprints" belonging to high-frequency signals.	It achieves a certain anti-forensic effect through a simple network.	Anti-forensic ability is general on many detectors.
	Huang et al. [35]	Use shallow reconstruction method based on dictionary learning to reduce up-sampling artifact patterns.	Artificial traces of reconstructed faces are reduced and such faces have a good anti-forensic ability.	Reconstructed faces look slightly blurry.
	Zhao et al. [12]	Use GAN technology and multiple white-box detectors to falsify forensic traces associated with real images.	Anti-forensic images can learn forensic traces and achieve good anti-forensic ability.	Anti-forensic transferability on some other detectors is insufficient.
Adversarial Attack	Wang et al. [17]	Use PGD attack method in YCbCr color space to allocate more perturbations to Cb and Cr channels instead of Y channel.	It reduces the perturbations of Y channel and improve the visual quality.	Anti-forensic transferability on some other detectors is insufficient.
	Goebel et al. [18]	Use optimization-based attack to impart the co-occurrence matrices taken from real images into GAN-generated images.	Anti-forensic images achieve good anti-forensic effect on the detectors based on co-occurrence features.	Anti-forensic images only aim at co-occurrence based detectors.
	Li et al. [19]	Use the trained StyleGAN model to perturb the input latent vector and noise.	Searching on StyleGAN's manifold make anti-forensic faces have a good visual quality.	Trained StyleGAN model is required and the transferability needs to be improved.
	Ding et al. [20]	Use multiple generators and discriminators to enhance image visual quality and implement face swap anti-forensics.	It enhances the visual quality by using multiple generators.	The framework of network is complicated.
	Hussain et al. [21]	Use PGD attack method and input transformation to generate robust adversarial deepfakes.	Anti-forensic videos and images are robust against JPEG compression.	The detectors in zero-knowledge scenarios are not considered.

visual quality and transferability of anti-forensic images.

C. *Local Adversarial Attacks*

Several local adversarial attack methods have been proposed for general image classification tasks. According to the location of the perturbations, in [38], local attack methods have been divided into those that perturb several pixels and those that perturb a region. The former perturbs pixels that are usually nonadjacent, while the latter modifies the values of pixels in a continuous region. The proposed method belongs to the latter. The following describes these two types of local attack methods in detail.

Attack methods that perturb several pixels are usually realized by constraining the number of perturbed pixels. The classical method JSMA [39] attempted to perturb a small number of pixels and it first utilized the norm constraint L_0. Su *et al.* [40] tried to attack DNNs by perturbing only one pixel. Croce *et al.* [41] restricted the perturbed pixels to be located in regions with rich colors in addition to minimizing the L_0 distance. Although these methods reduce the number of perturbed pixels, there may be two problems. Firstly, since the number of perturbed pixels is limited, the adversarial perturbation is sparse; as a result, the adversarial ability is weak, especially for high-resolution images. Secondly, the methods only constrain a few perturbed pixels, which may cause excessive modifications of these pixels that are easily detected by human eyes.

According to the visibility of adversarial perturbation, the attacks that perturb a region of pixels can be further divided into visible adversarial patch attacks and invisible local-region perturbation attacks:

Visible adversarial patch attack. This attack type mainly uses patches to cover regions of images. Brown *et al.* [42] optimized the given patch by their designed objective functions to create a general, robust and targeted visible adversarial image patch. Karmon *et al.* [43] introduced some latent properties and proposed the LaVAN attack to generate local adversarial noises that do not cover the main object pixels.

Invisible local-region perturbation attack. This type of attack mainly utilizes a mask to determine local areas and adds imperceptible perturbations in these areas. Qian *et al.* [38] proposed the CFR attack using the interpretability of neural networks and an optimization-based attack. Xiang *et al.* [44] utilized model interpretability and a gradient-based attack to generate an initial adversarial example. Then, they generated the final example through gradient estimation and random search.

Perturbation invisibility is an important requirement in the task of anti-forensics. In addition, in order to avoid the

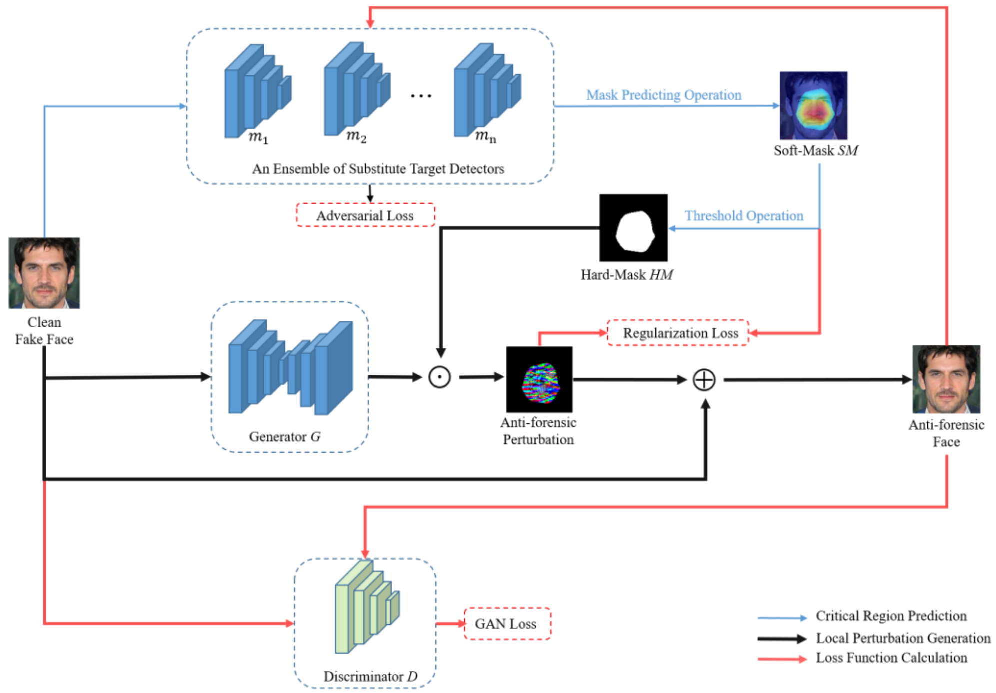

Fig. 1. Overall framework of our anti-forensic GAN

problems of perceptible perturbation and general adversarial ability caused by attack methods that perturb several pixels, we adopt the invisible local-region perturbation attack method. This method can reduce the perturbation while ensuring attack ability, thus achieving a good trade-off.

III. Proposed Method

In this section, we introduce our anti-forensic method in detail. In subsection III.A, the definition of anti-forensic problem is described first. Then, the overall framework is presented in subsection III.B. In the final three subsections, the specific methods adopted and loss functions for training are elaborated.

A. *Problem Definition of Anti-Forensics*

The goal of this paper is to fool GAN-generated face forensic detectors by adding well-designed perturbations to GAN-generated faces. The forensic detector predicts an input face x as：

$$y = \underset{k\in\{0,1\}}{\operatorname{argmax}} f_k(x), \tag{1}$$

where $f_k(x)$ is the formulation of an initial detector whose output is the probability for the category k, y is the predicted label, $y = 0$ means that the label is 'fake' and $y = 1$ denotes the 'real' label.

In the anti-forensic scenario, the goal for the perturbed GAN-generated fake face is to be predicted as 'real' with the label 1. The targeted attack process can be represented as follows:

$$\underset{k\in\{0,1\}}{\operatorname{argmax}} f_k(x_{fake} + \delta) = 1, \tag{2}$$

where x_{fake} is the original GAN-generated face and δ is the added adversarial perturbation. The reason why we do not adopt an untargeted attack is that the detector may incorrectly predict a few images and we should take these images into account.

B. *Overall Framework*

We are the first to introduce the local attacks in the task of anti-forensics. Unlike the existing works [17-21] using the global adversarial attacks that perturb each pixel of images, our work using local attacks only changes the values of some pixels. We adopt such local perturbations for two main reasons. Firstly, forensic detectors have areas of common concern in decision-making, and these areas' pixels are more useful than others for detector prediction. Therefore, perturbing such sensitive areas is enough to achieve strong anti-forensic effects. Secondly, local perturbations make the most of unmodified areas, which can improve the visual effect. In general, the proposed method is based on GANs to generate imperceptible and transferable anti-forensic adversarial examples. The overall framework of our anti-forensic GAN is based on AdvGAN [45], a classic adversarial perturbation generation framework. On the basis of AdvGAN, we design a double-mask guided local attack anti-forensic framework and three-part loss to ensure strong anti-forensic ability.

1) Double-Mask Guided Local Attack Framework

As shown in Fig. 1, our anti-forensic framework consists of two parts: critical region prediction and local-perturbation generation. The former predicts the local areas of the face images. Such areas play a key role in the decision-making of the ensemble of substitute target detectors. The area prediction is mainly realized by the proposed double-mask (soft mask and hard mask) strategy. The hard mask is used to restrict the perturbation area, while soft mask is utilized to constrain the perturbation degree of each pixel by regularization loss (please find more details in subsection III.E). The ensemble is a

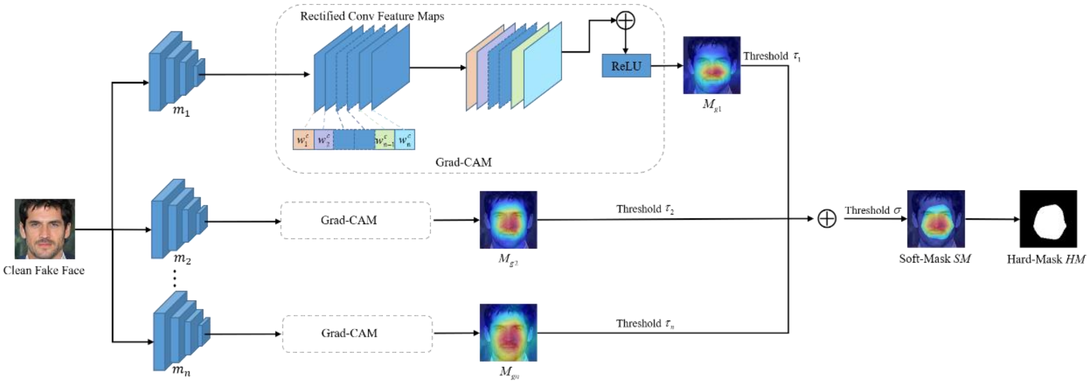

Fig. 2. Critical region prediction

collection of well-trained white-box detectors. The latter (local-perturbation generation) makes the generated perturbations possess the anti-forensic capabilities through adversarial training of generator *G* and discriminator *D*.

Specifically, the process of generating anti-forensic faces is described as follows. Firstly, a fake face is input into the ensemble of substitute target detectors to obtain a double-mask through the mask predicting and thresholding operations. Secondly, the face is input into *G* to generate the preliminary perturbation. Then, this perturbation is combined with the hard mask to retain the perturbations of critical areas and discard those of other regions. Finally, the obtained local perturbations are added to the input face to get the final anti-forensic face.

2) Three-part Loss

In the anti-forensic task, the three-part loss is designed for training *G* to make the perturbed faces achieve a strong anti-forensic ability and a good visual quality. The three-part loss is composed of GAN training loss, adversarial loss, and regularization loss. For the adversarial training of *G* and *D*, the GAN training loss is adopted to make perturbed faces indistinguishable from the input faces. For the anti-forensic ability of perturbed faces, the adversarial loss (consisting of ensemble classification loss and ensemble feature loss) is designed to give the generated local perturbations anti-forensic capabilities and excellent generalizability. Extra regularization loss is proposed to constrain the magnitude of perturbation and ensure visual quality. More details can be found in subsection III.E.

C. *Critical Region Prediction*

As described in subsection III.B, for a GAN-generated image, it is necessary to identify local areas and add adversarial perturbation to them to deceive detectors. These local areas must be critical regions for detectors' decision-making. Considering that we must generate anti-forensic faces with strong transferability and good visual quality, the critical regions are obtained from the target detectors themselves. We look for the areas of common concern in their decision-making to designate as perturbed regions. On the one hand, such local perturbations leave most areas of the original face image unchanged, reducing the magnitude of the perturbation and ensuring visual quality. On the other hand, the local regions are predicted from multiple target detectors and are considered to be common sensitive areas of these detectors. These areas are likely to contribute more than others to model decisions. To a certain extent, such areas can intuitively guarantee strong anti-forensic ability and transferability.

For the convenience of calculation, the shape of the critical region obtained can be represented by a binary matrix *M*:

$$M(i,j) = \begin{cases} 1 & x_{fake}(i,j) \in r, \\ 0 & \text{otherwise}, \end{cases} \tag{3}$$

where $x_{fake}(i,j)$ denotes the pixel value at coordinates (i,j) of x_{fake} and r represents the critical regions predicted. Thus, r can be transferred to $x_{fake} \odot M$, where $\odot$ represents the Hadamard product. Correspondingly, the local perturbation can be expressed as $\delta \odot M$. The whole anti-forensic process can be described as follows:

$$\underset{k\in\{0,1\}}{\operatorname{argmax}} f_k(x_{fake} + G(x_{fake}) \odot M) = 1. \tag{4}$$

In order to find the critical regions *r*, gradient-weighted class activation mapping (Grad-CAM) is selected as our basic mask predicting method. The Grad-CAM proposed by Selvaraju *et al.* [46] tried to produce a visual explanation for CNN-based models. Using such a visual attention technique helps us quickly predict the sensitive areas of the input face for model decisions. In addition, to make better use of the predicted key areas, we calculate a double mask instead of a single binary matrix *M*. The double mask includes a soft mask and a hard mask. The former not only retains the information of critical areas, but also preserves the contribution of each pixel in those areas. The latter corresponds to the binary matrix *M* in (3), which only retains the information of the areas. The advantage of the double mask is that we can better guide the generation of local perturbations, utilizing the relative importance of different pixels. The entire critical region predicting process using Grad-CAM technology (i.e., the double-mask predicting operation) is shown in Fig. 2.

As shown in Fig.2, the GAN-generated face x_{fake} is input into each forensic CNN in the ensemble. Then the feature maps output from the last convolutional layer of each detector can be taken out. Here, the Grad-CAM based operations (i.e., gradient computations, global average pooling, a weighted combination and rectified linear unit (ReLU)) are utilized to form an attention mask M_s of the same size as the input. The value of each element in M_s represents the importance of the coordinate pixels of the GAN-generated face. According to the magnitude of the value, an appropriate threshold can be set to obtain the

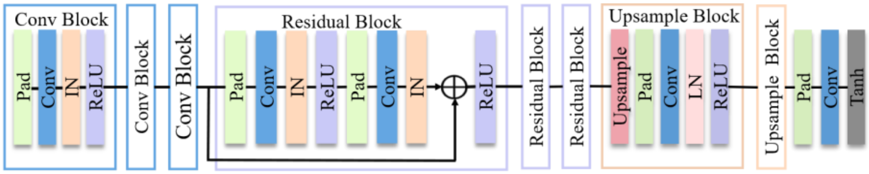

Fig. 3. Architecture of the adopted generator *G*.

soft mask of the sensitive areas for a single CNN. Then, the hard mask can be obtained by setting the soft mask value in the sensitive area to 1.

Mathematically, firstly, the soft mask can be computed as follows:

$$SM_s(i,j) = \begin{cases} M_s(i,j) & M_s(i,j) \geq \tau, \\ 0 & \text{otherwise}, \end{cases} \quad (5)$$

where SM_s is the soft mask obtained for the s-th single detector and τ is a threshold.

Then, the binary matrix called a hard mask can be calculated as follows:

$$HM_s(i,j) = \begin{cases} 1 & SM_s(i,j) > 0, \\ 0 & \text{otherwise}, \end{cases} \quad (6)$$

where HM_s is the shape of the critical region in (3).

Certainly, the aim of critical region prediction is to obtain the common critical areas of multiple detectors. Therefore, an ensemble of detectors rather than just one detector is used to determine the mask. Thus, a weighted combination operation is finally adopted to obtain the common areas of *n* detectors in the ensemble as follows:

$$SM(i,j) = \begin{cases} \sum_{s=1}^{n} \alpha_s SM_s\,(i,j) & \sum_{s=1}^{n} \alpha_s SM_s\,(i,j) \geq \sigma, \\ 0 & \text{otherwise}, \end{cases} \quad (7)$$

where n is the number of detectors in the ensemble, σ is a threshold, α_s are the weights satisfying $\sum_{s=1}^{n} \alpha_s = 1$ and $\alpha_s = 1/n, \forall s \in \{1,2,\dots,n\}$.

After that, the final hard mask of the critical region can also be obtained by:

$$HM(i,j) = \begin{cases} 1 & SM(i,j) > 0, \\ 0 & \text{otherwise}, \end{cases} \quad (8)$$

D. Local Perturbation Generation

Once the critical regions are obtained through the mask predicting operation, the local perturbation is generated by the well-trained generator. The training of the generator *G* is an iterative process with *D* and *G* alternately performing gradient descent on mini-batches.

The generator *G* is utilized to map the input GAN-generated face to the adversarial perturbation manifold to achieve effective anti-forensic perturbations. The discriminator *D* distinguishes the input faces from the perturbed faces, and the ensemble of substitute target detectors gives the perturbations generated by *G* anti-forensic capabilities. The architectures of *G* and *D* are shown as follows.

Generator *G*. Our generator adopts a structure similar to that of StarGAN [47]. The architecture is shown in Fig. 3. Since the dimension of the perturbation is the same as that of the input image, the generator adopts an encoder-decoder architecture, which is commonly used. Specifically, the encoder adopts a typical convolutional neural network architecture (mainly including three convolution blocks and three residual blocks) to down-sample the input image. Each convolution block's convolution layer is followed by an instance normalization and a ReLU activation. For three blocks, the numbers of convolution kernels are 64, 128, and 256, respectively; the kernel sizes are 7×7, 4×4, and 4×4, respectively; the strides are 2, 1, and 1, respectively. Subsequently, three residual blocks are adopted to further encode the features to a latent space representation. With the help of skip connections in the residual block, the efficiency of network convergence is improved. Among the residual blocks, each convolutional layer contains 256 kernels with a size of 3×3. In order to obtain perturbed images with higher visual quality, our decoder uses the resize-convolution method for up-sampling rather than deconvolution as the deconvolution operation is prone to suffer from checkerboard artifacts. The decoder is composed of two up-sample blocks and one convolution layer. The numbers of convolution kernels in the decoder are 128, 64, and 3, while the kernel sizes are 5×5, 5×5, and 7×7, respectively. Finally, we apply a Tanh function to obtain the adversarial perturbation, which has the same dimensions as the input face.

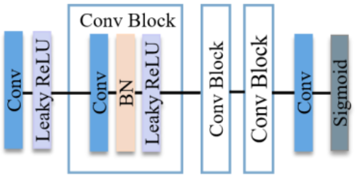

Fig. 4. Architecture of the adopted discriminator *D*.

Discriminator *D*. For the GAN training, a discriminator with a simple structure can perform well. Accordingly, our discriminator adopts a relatively simple structure based on DCGAN [48]. Its architecture is shown in Fig. 4. It contains three convolutional blocks and each block is composed of a convolutional layer followed by a batch normalization layer and a leaky ReLU function. We set the kernel size to 4×4 and the stride to 2 for the convolutional layers in all the blocks. The final 512-dimensional features are fed into a convolutional layer followed by a sigmoid function, to generate a one-dimensional output.

E. Training Loss Function

For the training of the whole anti-forensic GAN, the training loss function consists of the objective function of *G* and the function of *D* trained alternately. As mentioned in Section II.B, a well-designed perturbed face must achieve two objectives: a strong anti-forensic ability and a good visual effect.

Therefore, a three-part loss is designed to jointly guide the learning process of *G* according to the task requirements. It is composed of GAN training loss L_{GAN_G}, adversarial loss L_{adv}, and regularization loss L_{reg} , as follows:

$$L_G = L_{GAN\_G} + L_{adv} + L_{reg}. \quad (9)$$

For the optimization of *D*, the GAN training loss L_{GAN_D} is calculated to implement adversarial training with *G*, as follows:

$$L_D = L_{GAN\_D}. \tag{10}$$

Next, each of these loss functions is introduced in detail.

GAN Training Loss L_{GAN_G} and L_{GAN_D}. In order to ensure that the perturbed images are similar to the original images, this loss function is implemented to train the generator and discriminator. In the training process, the generator G requires to ensure that the images with local perturbations are as photo-realistic as possible so that the discriminator considers the generated images are the same distribution as the input images. On the contrary, the discriminator D tries to correctly distinguish the adversarial faces from the original faces. To increase the stability of the training process, the proposed method adopts the least square loss [49] instead of the commonly used cross-entropy loss. The function can be defined as follows:

$$L_{GAN\_D} = \mathbb{E}_{x\sim P_{data}(x)}[(D(x)-1)^2 + ((D(x+HM\odot G(x)))^2), \tag{11}$$

$$L_{GAN\_G} = \mathbb{E}_{x\sim p_{data}(x)}[(D(x+HM\odot G(x))-1)^2], \tag{12}$$

where $p_{data}(x)$ is the distribution of the input GAN-generated faces and *HM* is the binary matrix mentioned in (3) and (8). Here, *HM* mainly serves to discard the perturbations outside the critical areas.

Adversarial Loss L_{adv}. The goal of the adversarial loss is to give the generated local perturbations anti-forensic capabilities and excellent generalizability. Specifically, we hope that the perturbed image can fool the ensemble of substitute target detectors for training. The main reason for adopting a set of detectors instead of a single detector is that an ensemble attack can effectively enhance the transferability of adversarial examples. Liu *et al.* [22] pointed out that the decision boundaries of different models were similar, indicating that the ensemble method can improve the transferability of adversarial examples and allow them more easily attack black box models. Dong *et al.* [23] also improved the attack ability of adversarial examples via fooling multiple trained white-box source detectors in parallel. Therefore, the idea of integrated detectors is also adopted to allow the generated anti-forensic images to fool more detectors.

The adversarial loss consists of two parts: ensemble classification loss and ensemble feature loss. A basic integrated loss strategy is introduced by the following:

$$L_{adv} = \lambda \mathbb{E}_{x\sim p_{data}(x)} \sum_{s=1}^{n} \beta^{(s)} L_c^{(s)}(x) + \mu \mathbb{E}_{x\sim p_{data}(x)} \sum_{s=1}^{n} \beta^{(s)} L_f^{(s)}(x), \tag{13}$$

where $L_c^{(s)}$and $L_f^{(s)}$are the classification loss and feature loss of the s-th detector in the ensemble of substitute target detectors, respectively, λ and μ are two hyperparameters and $\beta^{(s)}$ are the ensemble weights, i.e., $\sum_{s=1}^{n} \beta^{(s)} = 1$ and $\beta^{(s)} = 1/n, \forall s \in \{1,2,\dots,n\}$.

The aim of the classification loss is to make the white-box detectors classify the perturbed GAN-generated face images, including images originally predicted as 'real', into the 'real' class. It utilizes the output space of each substitute detector, that is, the sigmoid cross-entropy between the anti-forensic image output and the real class:

$$L_c = -\sum_{b=1}^{2} t_b\, log(f(HM\odot G(x))_b), \tag{14}$$

where $f(\cdot)$ is the substitute detector and t_b is the b-th entry of target sigmoid vector with 1 for the 'real' class and 0 for the 'fake' class.

The feature loss forces to enlarge the distance between the anti-forensic image and the input image. It exploits the feature space of each substitute detector. Some works have shown that latent features can be utilized to improve the attack ability of adversarial examples. For example, Singh *et al.* [24] pointed out that latent features were more susceptible to adversarial perturbation, and Yu *et al.* [26] showed that the latent features in a specific robust model were very vulnerable to adversarial attacks. Thus, in this paper, the output space and feature space are combined to make the detectors classify the perturbed face images into the 'real' class as much as possible and boost the transferability of anti-forensic images. Concretely, the reciprocal of KL divergence is exploited as the feature space loss:

$$L_f = \frac{1}{\mathrm{KL}(\mathbb{D}(x),\, \mathbb{D}(HM\odot G(x))}$$

$$= C/\sum_{c=1}^{C}(\mathbb{D}(HM\odot G(x),c)\cdot \log\left(\frac{\mathbb{D}(HM\odot G(x),c)}{\mathbb{D}(x,c)+\epsilon}+\epsilon\right)), \tag{15}$$

where $\mathbb{D}(\cdot)$ is the output feature map of a certain hidden layer, C is the number of channels, and ϵ is a smoothing term that avoids the division by zero.

Regularization Loss L_{reg}. The main purpose of regularization loss is to constrain the magnitude of perturbation, thus ensuring visual quality and avoiding the overfitting of the ensemble of substitute target detectors. As mentioned in subsection III.B, we predict a double mask to guide the generation of local perturbations. The soft-mask *SM* preserves the contribution of each pixel in the perturbed region. When *HM* is used to restrict the perturbation area, *SM* is utilized here to constrain the perturbation degree of each pixel in more detail. Concretely, we aim to add the greater perturbation in the more important pixels; the greater the value of the *SM*, the less the constraint on the magnitude of the pixel's perturbation. In addition, a perception loss is employed to improve the naturalness of the anti-forensic faces and reduce the overfitting of the above feature loss. The regularization loss is defined as follows:

$$L_{reg} = \mathbb{E}_{x\sim p_{data}(x)}\, max\left(0, \left\|(1-SM)\odot\left(HM\odot G(x)\right)\right\|_2 - \varsigma\right) + \mathbb{E}_{x\sim p_{data}(x)} L_{perc}, \tag{16}$$

where ς is a hyperparameter to control the amount of perturbation, and L_{perc} is a commonly used VGG perception loss [50] in style transfer tasks.

After the iterative optimization of the above objective functions, a well-trained G combined with the mask predicting operation can quickly generate local perturbations and then add perturbations to the GAN-generated faces to achieve anti-forensic effects.

IV. Experimental results and analysis

In this section, the experiments results are presented. In subsection IV.A, the experimental setting is introduced first. Then, we analyze the results of the proposed method in subsection IV.B. Finally, other experiments are discussed in subsection IV.C, including ablation study.

TABLE II
BASELINE FORENSIC ACCURACIES OF THE TRAINED DETECTORS

Detectors	Denotes	Accuracy
ResNet-50 [6]	m_1	1.0000
Xception [8]	m_2	0.9999
EfficientNet-b0 [29]	m_3	0.9993
DenseNet-121 [30]	m_4	0.9996
MesoInception [5]	m_5	0.9981
AlexNet [52]	m_6	0.9980
Discriminator [48]	m_7	0.9980
GramNet [7]	m_8	1.0000
RFM [10]	m_9	0.9565

A. *Basic Setting*

Datasets. The experiments in this paper adopt CelebA [51] as the initial dataset, which contains 202,599 natural face images of 10,177 people. For the anti-forensics task, we create a GAN-generated face dataset based on the CelebA dataset. The created dataset is composed of 200,000 photo-realistic facial images generated by StyleGAN, which is the most commonly used tool to generate high-quality face images. The dataset is split into 190,000 images for anti-forensic GAN training and 10,000 images for testing. For all detection models used in the experiments, the natural image dataset CelebA and our created GAN-generated face dataset are uniformly adopted for training. For training convenience, all the images in the dataset are background-removed and resized to 128×128.

Detectors. Before training and evaluating the anti-forensic attack, we first train the forensic detectors adopted in the paper. We select nine effective detectors [5-8, 10, 32-33, 48, 52] for computer vision and multimedia forensics tasks, and all the detectors are trained under the same conditions. The training details are as follows: the datasets are the same as those mentioned above, the target function is cross-entropy loss, the learning rate is set to 0.0001 with a weight decay of 0.0004, the batch size is 64 and AdamW is selected as the optimizer. For the convenience of expression, we denote the detectors as m_i, $i = 1,2,\ldots,8$. The specific detectors and their detection accuracies are shown in Table II. We can find that all the selected detectors achieve accuracies close to 1.0 for the task of recognizing GAN-generated faces.

Evaluation Metrics. To evaluate the performance of our method, the above forensic detectors are used to classify anti-forensic images and calculate the anti-forensic success rate (*ASR*), defined as follows:

$$ASR = \frac{ASF\_NUM}{F\_NUM}, \tag{17}$$

where *ASF_NUM* is the number of perturbed faces classified as ‘real’ and *F_NUM* is the total number of GAN-generated fake faces.

Moreover, the mean of structural similarity (SSIM), seak signal-to-soise ratio (PSNR), learned perceptual image patch similarity (LPIPS) of all images in the testing dataset are adopted to evaluate the visual quality of anti-forensic images.

Given the clean face *U* and the anti-forensic face *V*, the SSIM can be defined as follows:

$$\mathrm{SSIM}(U,V) = \frac{(2\mu_U\mu_V+C_1)(2\sigma_{UV}+C_2)}{(\mu_U^2\mu_V^2+C_1)(\sigma_U^2\sigma_V^2+C_2)}, \tag{18}$$

where μ_U and μ_V are the means, σ_U and σ_V are the standard deviations, σ_{UV} is the cross-covariance, C_1 = 0.0001 and C_2=0.0009 are two constants for avoiding a null denominator.

The PSNR is formulated as:

$$\mathrm{PSNR}(U,V) = 10\,log_{10}(\frac{M^2}{MSE(U,V)}), \tag{19}$$

where *M* is the maximum pixel value, *MSE* is Mean Squared Error defined as:

$$MSE = \sum_{m=1}^{p}(U_m - V_m)^2, \tag{20}$$

where *p* is the number of pixels.

The LPIPS is a popular CNN-based image quality assessment metric for semantic similarity measurement.

Finally, we take the computational time required to generate an anti-forensic image in the inference stage as the index of anti-forensic efficiency.

Experimental Environment. In this paper, all the anti-forensic experiments are implemented via PyTorch. The computational complexity of our method is 15.09GFlops and the number of parameters is 56.34M. We run our network and other experiments on 24GB GeForce RTX 3090, 3.80GHz i7-10700KF CPU, and 32GB RAM.

Model Selection. For fairness, the ensemble of white-box substitute detectors used is uniformly the ensemble of m_1, m_2, and m_3 in each experiment. And the remaining six detectors are black-box detectors to test the performance of the methods. The three white-box detectors are chosen for the reason that they are the commonly used and perform very well in GAN-generated face detection.

Hyperparameter Setting. The specific details of our method are described as follows. The combination thresholds σ in (7) and τ in (5) are dominant hyperparameters. Among that, σ is set as 0.3 and τ for three detectors, m_1, m_2, and m_3, are set as 0.4, 0.25 and 0.5, respectively. The hyper-parameters ς in (16), λ and μ in (13) are 1, 10 and 1, respectively. Adam is selected as the optimizer and the learning rate is set to 0.0001. The batch size is set to 64.

Regarding the hyperparameter setting of the compared methods, the experimental settings are carried out according to their original literature. Five methods are developed as baseline in this paper: FGSM attack [36] under the $\mathrm{L_{inf}}$ constraint, PGD attack [37] under the $\mathrm{L_{inf}}$ norm constraint, AdvGAN [45] attack, and two anti-forensic methods [12, 17]. For the FGSM and PGD attacks, the perturbation bound ε is set as 6.0 for the trade-off between the *ASR* and visual quality. For the anti-forensic method of Wang *et al.*, perturbation thresholds ε_Y, ε_{Cb}, and ε_{Cr} for different channels of YCbCr color space are set as 1.5, 3, and 3, respectively. Normal GAN training is conducted for AdvGAN and the method of Zhao *et al*.

B. *Anti-forensic Attack Experiments*

In the experiments of our anti-forensic method, *G* and *D* are trained alternately in each iteration. The proposed anti-forensic generator is trained for approximately 20,000 iterations in total. The generator training loss gradually tends to be smooth after iterating nearly seven epochs with 2969 iterations per epoch, in Fig. 5.

1) Experiments in Zero-Defense Scenario

TABLE III
ANTI-FORENSICS PERFORMANCE FOR THE WHITE-BOX SUBSTITUTE DETECTORS AND THE ZERO-KNOWLEDGE BLACK-BOX DETECTORS IN TERMS OF *ASR*.

Methods	White-box detectors			Black-box detectors						Avg. *ASR*
	m_1	m_2	m_3	m_4	m_5	m_6	m_7	m_8	m_9	
FGSM [36]	0.6289	0.9575	**1.0000**	0.7613	0.9985	**0.2725**	0.6030	0.8238	0.2223	0.6964
PGD [37]	0.9828	0.9826	**1.0000**	0.9720	0.9997	0.2521	0.5614	0.9836	0.9253	0.8511
AdvGAN[45]	0.9989	**1.0000**	**1.0000**	0.9535	0.9999	0.0412	0.2205	0.9957	0.4511	0.7401
Wang's [17]	0.9999	0.9999	0.9999	**0.9998**	0.9999	0.0950	0.3260	**0.9999**	0.9694	0.8211
Zhao's [12]	**1.0000**	**1.0000**	**1.0000**	0.9996	**1.0000**	0.0232	0.2113	0.9916	**0.9766**	0.8003
Prop.	0.9999	**1.0000**	0.9990	0.9923	0.9409	0.2115	**0.9904**	0.9684	0.8910	**0.8882**

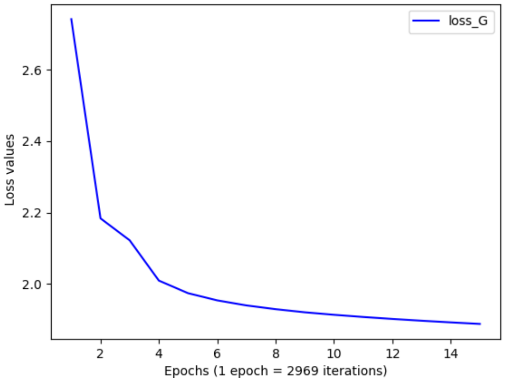

Fig. 5. Curve of the generator training loss L_G.

Here, the anti-forensic faces encounter no defense means. In this scenario, the performances and discussion of our method and comparison methods are reported in white-box and black-box settings.

For the substitute detectors used in the training process, the anti-forensic performance is shown in Table III. Most methods, including the proposed method, achieve an *ASR* close to 1.0 for fooling three white-box detectors. This performance illustrates the effectiveness of the proposed method for anti-forensic tasks of known detectors. In addition, the proposed method outperforms the other methods regarding to visual quality. As shown in Table IV, it achieves a mean SSIM of 0.9849, a mean PSNR of 44.5047, and a mean LPIPS of 0.0045. These results show that the perturbed faces produced by our proposed generator maintain high visual quality. In Table IV, we also present the computational time of each method for generating an anti-forensic face with a resolution of 128×128 in the same operating environment. Although the proposed method is not the fastest of the evaluated methods, it is still fast.

For the zero-knowledge black-box detectors, the results are shown in Table III. Although the *ASR* of the proposed method on a single detector may be slightly lower than the optimal result, the method has the best average *ASR*, reaching 0.8441. The proposed method has better transferability than other methods. The possible reason is that different detectors share areas of common concern for a face image, which is consistent with the ideas of tasks in the classification scenario [38, 53]. We also find that the *ASR*s of all methods for m_6 are lower than those on other detectors. The same phenomenon is shown for m_7 in addition to our method. The reason is speculated that the detectors are different in network modules and depth so that they have the sophisticated decision landscapes. Detectors m_6 and m_7 are more lightweight in architectures and shallower in depth. The anti-forensic perturbation generated from the detectors (m_1, m_2 and m_3) with relatively complex structure may have poor transferability on detectors (m_6 and m_7) with relatively simple structure. This is consistent with the experimental conclusion in work [54]. The work [54] divided the models into three categories: huge CNNs, lightweight CNNs, and non-CNNs. Its experimental analysis shows that the adversarial transferability of different model families is quite different; in most cases, the adversarial samples obtained by models in one category have poor transferability to models in another category. These above results suggest that the proposed method has posed a sufficiently strong threat to both white-box detectors and black-box detectors, though the *ASR* for m_6 is not ideal. In order to further analyze the effects of different anti-forensic methods, the attention saliency maps of faces before and after perturbation are shown in Figs. 6 and 7.

TABLE IV
VISUAL QUALITY MEASURES AND COMPUTATIONAL TIME IN THE INFERENCE STAGE OF DIFFERENT ANTI-FORENSIC METHODS

Methods	Mean SSIM	mean PSNR	mean LPIPS	Computational Time (s)
FGSM [36]	0.8582	38.6579	0.0090	0.0543
PGD [37]	0.9360	42.5955	**0.0044**	0.4399
AdvGAN[45]	0.9432	38.0376	0.0127	**0.0127**
Wang's [17]	0.9435	43.0803	0.0046	0.9148
Zhao's [12]	0.9467	34.5811	0.0405	0.0107
Prop.	**0.9849**	**44.5047**	0.0045	0.0949

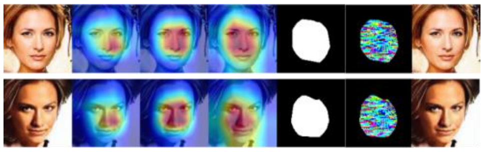

Fig. 6. Examples of attention saliency maps generated by the ensemble of substitute target detectors, *HM* based on them and the corresponding perturbations. The first column are the original faces, the next three columns are the maps obtained by m_i, $i = 1,2,3$, followed by the *HM*, perturbations and the anti-forensic faces.

In Fig. 6, it can be seen that different detectors have areas of common interest when making decisions. These areas are mainly located in the face areas, which is in line with human intuition. Meanwhile, the faces before and after perturbation are visually indistinguishable. The *HM* obtained by the ensemble of white-box detectors not only retains the areas of common concern, but also retains the important areas predicted by each single detector. The perturbed regions are the local areas obtained strictly according to *HM*.

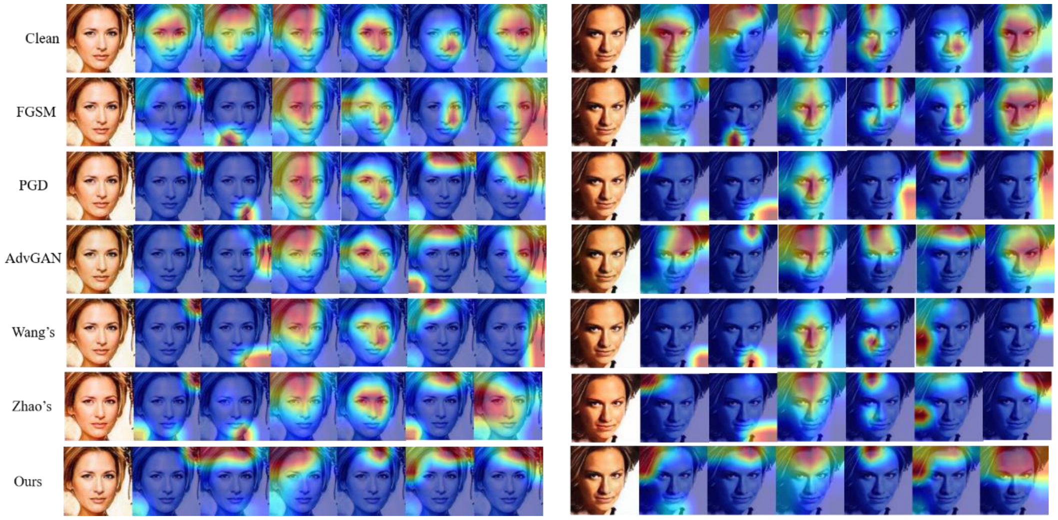

Fig. 7. Examples of attention saliency maps of images before and after being perturbed. For the same face, the first column are the original faces and anti-forensic faces and the rest of columns from left to right are the corresponding attention saliency maps of six black-box detectors m_i, $i = 4, \dots, 9$.

As shown in Fig. 7, the perturbations generated by the proposed method may greatly change the area originally focused on by the detector and achieve the overall best performance. The possible reason is that the proposed method utilizes the fake faces' areas of common concern in multiple-detector decision-making to add local adversarial perturbations. The pixels in these areas of common concern are more useful for detector prediction than other pixels. The results are consistent with those of Wang *et al.* [55], who use feature importance to guide adversarial examples to disrupt important object-aware features for optimization; as a result, the attacked model could not capture important areas but focus on trivial areas instead. This change in the areas successfully improves the transferability. In addition, for detectors, such as m_6 and m_7, the anti-forensic faces obtained by some compared methods do not change the critical area. To a certain extent, this shows that the transferability of related methods on these detectors is poor. Likewise, although the area of concern of m_6 has been changed by the proposed perturbations, the change is not complete. This may be a possible reason for the inadequate anti-forensic effect on m_6 shown in Table III.

2) Experiments in Defense Scenario

In practice, generated anti-forensic faces may encounter defense solutions against anti-forensic attacks. In such scenario, the added anti-forensic perturbations may fail. Therefore, in this subsection, we specifically discuss the anti-forensic performances of the proposed method under defense operations.

(a) Defenses that modify input faces

In this type of defense, the parameters of the trained detectors (i.e., m_i, $i = 1, 2, \dots, 9$) are not changed. Common image filtering operations have been shown to be effective defense methods that can remove the adversarial noises in [56]. In addition, adding additional noise may also destroy anti-forensic perturbations. Thus, Gaussian filtering, median filtering, bilateral filtering, and Gaussian noise are respectively applied to the anti-forensic faces. The kernel sizes of all the filters are 3×3 and the standard deviation of Gaussian noise is set to 3. The results are shown in Table V. It can be seen from Table V that the common filtering and noise operations fail to defend against our anti-forensic perturbations. On the contrary, anti-forensics are even more successful when these operations are applied. For example, the *ASR* of anti-forensic faces increases by 0.0266 after Gaussian filtering. In essence, image filtering does not improve the cognitive ability of neural networks. However, it can usually reduce the classification accuracy of clean samples due to the filtering of image details. Moreover, when the intensity of adversarial perturbation on the clean image is small, simple filtering is difficult to distinguish image details and adversarial noise. In this paper, the intensity of local perturbation is constrained, and the perturbation is subtle and invisible. Therefore, the common filtering and noise operations reduce detectors' accuracy and promote the success of anti-forensics.

Image compression [57] and input reconstruction [58] have also been proved to be effective defense means to eliminate the impact of adversarial perturbations. In this paper, we consider JPEG compression with the compression factor of 90. Concerning image reconstruction, two methods are considered: shallow reconstruction [35] and MagNet [58]. The shallow reconstruction used in Fakepolisher [35] originally utilizes dictionary learning to reconstruct the image and eliminate traces of up-sampling. MagNet [58] is an adversarial defense framework independent of the adversarial attack method. It tries to train an adversarial example detector and reformer through clean samples and reconstruct the perturbed examples closest to the clean samples. The anti-forensic effects against image compression and input reconstruction are shown in Table VI. The results show that: (1) Image compression has a small impact on our method with the average *ASR* decreased from 0.8882 to 0.7959, but it cannot make the local perturbation generated by the proposed method lose the anti-forensic performance; (2) Anti-forensic performance is enhanced after shallow reconstruction with the average *ASR* increased from 0.8882 to 0.9581. The possible reason is that shallow reconstruction fails to fully destroy the anti-forensic perturbations but reduces the up-sampling artifact patterns, which some detectors may rely upon as forensic features; (3)

TABLE V
ANTI-FORENSICS PERFORMANCE AFTER FILTERING OR ADDING NOISE IN TERMS OF *ASR*.

Filter/Noise	m_1	m_2	m_3	m_4	m_5	m_6	m_7	m_8	m_9	Avg.
None	0.9999	1.0	0.9990	0.9923	0.9409	0.2115	0.9904	0.9684	0.8910	0.8882
Gaussian Filter	1.0	1.0	1.0	0.9991	0.9999	0.2413	0.9996	0.9998	0.9932	0.9148
Median Filter	1.0	1.0	1.0	0.9996	1.0	0.2548	0.9957	0.9998	0.9971	0.9163
Bilateral Filter	1.0	1.0	0.9999	0.9991	0.9992	0.1079	0.9972	0.9992	0.9863	0.8988
Gaussian Noise	0.9999	1.0	1.0	0.9959	0.9611	0.9095	0.9989	0.9740	0.9397	0.9754

TABLE VI
ANTI-FORENSICS PERFORMANCE AGAINST IMAGE COMPRESSION AND INPUT RECONSTRUCTION DEFENSES IN TERMS OF *ASR*

Defense	m_1	m_2	m_3	m_4	m_5	m_6	m_7	m_8	m_9	Avg.
JPEG Compression	0.9996	0.9793	0.7018	0.9998	0.4313	0.3680	0.6975	1.0	0.9857	0.7959
Shallow Reconstruction [35]	1.0	1.0	1.0	0.9998	1.0	0.6486	0.9743	1.0	1.0	0.9581
MagNet [58]	1.0	0.9999	0.5564	0.9993	1.0	0.0625	0.5296	1.0	0.9999	0.7942
None	0.9999	1.0	0.9990	0.9923	0.9409	0.2115	0.9904	0.9684	0.8910	0.8882

TABLEVII
ANTI-FORENSICS PERFORMANCE AGAINST DETECTORS RETRAINED BY ANTI-FORENSIC FACES GENERATED BY PGD ATTACK IN TERMS OF *ASR* WHEN THE WHITE-BOX DETECTORS USED IN THE ANTI-FORENSIC PROCESS ARE KNOWN

METHOD	m_1	m_2	m_3	m_4	m_5	m_6	m_7	m_8	m_9	Avg.
FGSM [36]	0.0096	0.0000	0.0149	0.0080	0.0000	0.0000	0.0072	0.0017	0.0013	0.0047
PGD [37]	0.0236	0.0161	0.0071	0.0080	0.0001	0.0000	0.0042	0.0008	0.0038	0.0071
ADVGAN [45]	0.0962	0.0005	0.0167	0.0285	0.0001	0.0000	0.0258	0.0322	0.0031	0.0225
WANG'S [17]	0.0233	0.0813	0.0039	0.0151	**0.0043**	0.0000	0.1013	0.0003	0.0080	0.0264
ZHAO'S [12]	**0.8075**	0.0176	0.3715	**0.7617**	0.0001	0.0034	0.1012	0.8020	0.2360	0.3446
PROP.	0.6932	**0.6757**	**0.5320**	0.6722	0.0011	**0.2079**	**0.8523**	**0.8040**	**0.2692**	**0.5230**

TABLE VIII
ANTI-FORENSICS PERFORMANCE AGAINST DETECTORS RETRAINED BY ANTI-FORENSIC FACES GENERATED BY DIFFERENT METHODS IN TERMS OF *ASR* WHEN THE WHOLE ANTI-FORENSIC ATTACK MECHANISM IS KNOWN

Method	m_1	m_2	m_3	m_4	m_5	m_6	m_7	m_8	m_9	Avg.
FGSM [36]	0.0	0.0001	0.0009	0.0005	0.0001	0.0006	0.0002	0.0034	0.0	0.0006
PGD [37]	**0.0043**	**0.0158**	0.0162	0.0043	0.0001	0.0009	**0.0068**	0.0036	0.0006	0.0058
AdvGAN [45]	0.0004	0.0002	0.0051	0.0020	0.0011	0.0024	0.0019	0.0052	0.0	0.0012
Wang's [17]	0.0017	0.0026	0.0285	0.0023	0.0039	**0.0045**	0.0031	0.0113	**0.0049**	0.0070
Zhao's [12]	0.0	0.0005	0.0004	0.0	0.0001	0.0019	0.0012	0.0002	0.0	0.0005
Prop.	0.0032	0.0001	**0.0335**	**0.0055**	**0.0099**	0.0029	0.0	**0.0159**	0.0001	**0.0078**

TABLE IX
ANTI-FORENSICS PERFORMANCE OF SECONDARY ATTACK AGAINST RETRAINED DETECTORS RETRAINED BY ANTI-FORENSIC FACES GENERATED BY DIFFERENT METHODS IN TERMS OF *ASR*

Method	m_1	m_2	m_3	m_4	m_5	m_6	m_7	m_8	m_9	Avg.
FGSM [36]	0.0006	0.0005	0.1596	0.0008	0.0001	0.0001	0.0008	0.0123	0.0	0.0194
PGD [37]	0.0912	0.9523	0.9946	0.0059	0.0001	0.0008	0.0084	0.0084	0.0109	0.2303
AdvGAN [45]	0.9988	0.9998	0.9993	0.2635	0.0408	0.0093	0.0609	0.9616	0.1272	0.4957
Wang's [17]	**1.0**	**1.0**	**1.0**	0.012	0.0204	0.0040	0.0013	**0.9830**	0.0829	0.4560
Zhao's [12]	0.9936	0.9966	0.9954	0.4038	0.1309	0.1323	0.2196	0.9570	0.2128	0.5602
Prop.	0.9763	0.9947	0.9745	**0.5132**	**0.6023**	**0.4187**	**0.2521**	0.1863	**0.2258**	**0.5715**

The average *ASR* is decreased about 0.09 under the MagNet input reconstruction defense, but it can still maintain its anti-forensic effect for most detectors.

(b) Defenses that modify detectors

In addition to the above defenses that modify input faces, there are also defense means of changing detector parameters. If our attack mechanism is partially or completely known, the defender may retrain the detectors using data augmentation to enhance the detectors robustness against adversarial attack [59]. Specifically, the defender may know the white-box detectors used in our anti-forensic process or even the whole anti-forensic method. Therefore, we conduct experiments on these two types of retraining defenses.

If the white-box detectors used in our anti-forensic process are known, the defender may try to use a known attack method to generate adversarial examples, then use these adversarial examples to retrain the detectors. In the test, we choose PGD attack as the known attack method, as it is the most commonly used method. Table VII presents the corresponding anti-forensic performance. It can be found that our anti-forensic faces can still maintain a certain anti-forensic ability against the robust detectors retrained by PGD-attacked faces.

If the whole anti-forensic attack mechanism is known, the defender can directly add the anti-forensic faces generated by the attack to the detector training process through data augmentation. Therefore, in this test, we replace the original clean fake faces in the training process of the nine detectors with anti-forensic faces at a proportion of 50%. Table VIII shows that all the compared methods are ineffective under the retraining defense. This is a normal phenomenon because using adversarial examples for data augmentation is an important way to reduce model overfitting and improve model robustness, including resisting anti-forensics. Although the retrained detectors have strong defensive performance against anti-forensic faces, they are still vulnerable to a secondary attack by the corresponding attack method. In other words, in the real-world context, the white-box detectors are available; thus, we can replace the original white-box detectors trained by using clean faces with the above-mentioned retrained robust detectors. Then, our attack method can be carried out again. Table IX shows the results of the secondary attack on the retrained model. The proposed method still has good anti-forensic performance for the retrained white-box detectors. This is mainly because the retrained white-box detectors directly participate in the training of our perturbation generator. In addition, our method still has stronger transferability to the retrained black-box detectors than other attack methods.

C. Other Experiments

1) Ensemble Strategy

For the ensemble strategy of substitute target detectors, there are four ensemble strategies: ensemble in loss, ensemble in logits, ensemble in predictions [22], and an alternative method [19]. Ensemble in loss is the above-mentioned basic method in which the loss function of each substitute white-box forensic detector is calculated separately and then aggregated. Ensemble in logits means fusing the output logits to compute the final loss. Ensemble in predictions calculates the loss directly via the output probabilities. In the alternative scenario, the detector in the ensemble is utilized alternatively in each step to carry out a gradient-descent updating process. The performances of the four strategies are shown in Table X. All the strategies achieve good anti-forensic results for white-box detectors, but the effects differ for zero-knowledge detectors. Although some strategies have achieved the highest *ASR* for a single black-box detector, in general, the basic loss ensemble strategy has the best overall anti-forensic transferability.

2) Perturbations in Some Specific Areas

In this paper, the perturbed areas (the predicted masks) are fully determined by the ensemble of substitute target detectors. In order to further explore the influence of the perturbed areas on the effect of anti-forensics, we carry out some additional experiments on perturbing specific areas, such as the eyes, nose, and mouth. Specifically, we use the Dlib tool for face key-point detection to obtain the masks of specific areas. These areas are used to directly replace the hard masks in our framework shown in Fig. 1. The anti-forensic performance is shown in Table XI. In order to illustrate the visual quality after adding perturbation, we also present some examples of specific areas and their corresponding perturbed faces in Fig. 8. The following can be observed from Table XI and Fig. 8: (1) When the perturbation area is small, such as the nose and mouth areas, the anti-forensic performance is generally low for both white-box and black-box detectors. The reason is that the small perturbation area weakens the adversarial ability and makes the perturbation intensity in the small area strong. As a consequence, the perturbation is easily detected by human eyes; (2) When simultaneously perturbing the eyes, nose and mouth, the anti-forensic perturbations are invisible and achieve good anti-forensic performance for the white-box detector. However, their transferability to the black-box detectors is not as good as the proposed method. This shows that in addition to specific areas, other areas of the face are also meaningful for the forensic detectors.

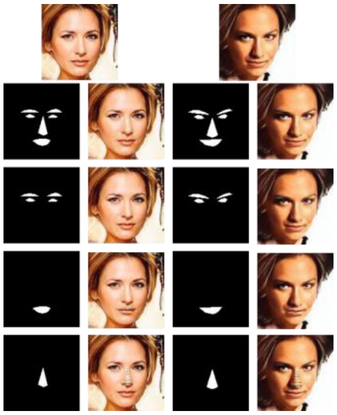

Fig. 8. Examples of specific masks and corresponding perturbed faces. The first line are the original clean faces. For the same face, the first column are the specific masks and the second column are the perturbed faces.

3) Ablation study

The architecture of our anti-forensic GAN is based on the baseline classical adversarial sample generation method, AdvGAN. Compared with AdvGAN, the mask predicting operation is proposed to generate perturbation in critical local regions, and the ensemble strategy and hidden layer information are combined to improve the anti-forensic effect. In this subsection, we mainly analyze the effectiveness of relevant strategies, including the mask predicting operation, the ensemble and ensemble feature loss L_f . Moreover, the additional role of the VGG perceptual loss L_{perc} in reducing overfitting is also presented.

The results of ablation experiments on adopting an ensemble of substitute detectors are shown in Table XII. For each single detector, it can be found that its performance of anti-forensics is average in the black-box scenario that is, the generated perturbation is less transferable. This demonstrates the effectiveness of the ensemble.

TABLE X
PERFORMANCE COMPARISON OF FOUR ENSEMBLE STRATEGIES FOR BOTH THE WHITE-BOX AND BLACK-BOX DETECTORS

Strategies	White-box detectors			Black-box detectors						Avg. *ASR*	mean SSIM	mean PSNR	mean LPIPS
	m_1	m_2	m_3	m_4	m_5	m_6	m_7	m_8	m_9				
Loss	0.9999	**1.0000**	0.9990	0.9923	**0.9409**	0.2115	**0.9904**	0.9684	**0.8910**	**0.8882**	0.9849	44.5047	0.0045
Logits	**1.0000**	**1.0000**	0.9269	0.8752	0.3522	**0.2717**	0.5321	0.6735	0.8272	0.7176	0.9836	44.6694	0.0044
Predictions	0.9997	**1.0000**	0.9988	**0.9930**	0.8543	0.0714	0.3419	**0.9778**	0.8875	0.7916	**0.9866**	**45.3070**	0.0037
Alternative	**1.0000**	**1.0000**	**1.0000**	0.9882	0.8359	0.0022	0.0067	0.9376	0.2499	0.6689	0.9791	43.2220	**0.0026**

TABLE XI
ANTI-FORENSICS PERFORMANCE OF SPECIFIC MASKS FOR BOTH THE WHITE-BOX AND BLACK-BOX DETECTORS IN TERMS OF *ASR*

Specific Mask	White-box detectors			Black-box detectors						Avg.
	m_1	m_2	m_3	m_4	m_5	m_6	m_7	m_8	m_9	
Eyes	0.8995	0.9985	0.1848	0.7342	0.9880	0.0095	0.1418	0.1940	0.1041	0.4727
Nose	0.4817	0.9989	0.0112	0.5244	0.8817	0.1755	0.8133	0.1785	0.0509	0.4573
Mouth	0.4565	0.9983	0.0437	0.0280	0.9148	0.1574	0.2236	0.0296	0.0431	0.3217
All	0.9894	0.9990	0.9483	0.9919	**0.9885**	0.0118	0.0164	0.4325	0.1198	0.6108
Prop.	**0.9999**	**1.0000**	**0.9990**	**0.9923**	0.9409	**0.2115**	**0.9904**	**0.9684**	**0.8910**	**0.8882**

TABLE XII
PERFORMANCE COMPARISON OF ADOPTING AN ENSEMBLE OF SUBSTITUTE DETECTORS OR A SINGLE ONE. THE FIRST COLUMN REPRESENTS THE WHITE-BOX DETECTOR USED IN TRAINING PROCESS

Detectors	m_1	m_2	m_3	m_4	m_5	m_6	m_7	m_8	m_9	Avg.
m_1	**0.9999**	0.2096	0.0577	0.7062	0.0655	0.0009	0.0439	0.7149	0.7638	0.3958
m_2	0.4789	0.9967	0.0361	0.2268	0.1953	0.1282	0.1973	0.2257	0.6721	0.3507
m_3	0.3514	0.0429	0.9987	0.2596	0.3499	0.0040	0.0762	0.1817	0.6185	0.3204
m_1+m_2+m_3 (Prop.)	**0.9999**	**1.0000**	**0.9990**	**0.9923**	**0.9409**	**0.2115**	**0.9904**	**0.9684**	**0.8910**	**0.8882**

TABLE XIII
ABLATION STUDY OF THE MASK PREDICTING (MP) OPERATION, FEATURE SPACE LOSS AND VGG PERCEPTUAL LOSS ABLATION EXPERIMENTS

Ablation	White-box detectors			Black-box detectors						Avg.
	m_1	m_2	m_3	m_4	m_5	m_6	m_7	m_8	m_9	
Without MP	**1.0000**	**1.0000**	**1.0000**	0.6420	0.8275	0.0018	0.0047	**0.9992**	**0.9346**	0.7122
Without L_f	0.9998	**1.0000**	0.9998	**0.9986**	0.7944	0.0379	0.9856	0.9886	0.7802	0.8428
Without L_{perc}	**1.0000**	**1.0000**	**1.0000**	0.9981	0.9380	0.1081	0.7948	0.9520	0.7930	0.8427
Prop.	0.9999	**1.0000**	0.9990	0.9923	**0.9409**	**0.2115**	**0.9904**	0.9684	0.8910	**0.8882**

TABLE XIV
ANTI-FORENSICS PERFORMANCE ON PGGAN AND WGAN_GP DATASETS

Datasets	White-box detectors			Black-box detectors						Avg. *ASR*	mean SSIM	mean PSNR	mean LPIPS
	m_1	m_2	m_3	m_4	m_5	m_6	m_7	m_8	m_9				
PGGAN	0.9960	0.9978	0.9618	0.9756	0.8920	0.3674	0.4651	0.9458	0.8113	**0.8236**	0.9859	42.8203	0.0038
WGAN_GP	0.9920	0.9915	0.9501	0.4986	0.4457	0.1224	0.6109	0.7581	0.8144	**0.6971**	0.9873	45.9328	0.0028
StyleGAN (Prop.)	0.9999	1.0000	0.9990	0.9923	0.9409	0.2115	0.9904	0.9684	0.8910	**0.8882**	0.9849	44.5047	0.0045

The corresponding results of the ablation experiments on the mask predicting (MP) operation and two losses are shown in Table XIII. The results without MP operation show that removing this operation leads to a drop in the *ASR*s for the black-box detectors. This also shows that different detectors have areas of common concern and demonstrates the effectiveness of our proposed MP operation on anti-forensic transferability. The results without L_f and L_{perc} show that these two loss functions contribute to the effect of anti-forensic perturbations. The reason may be that the hidden layer features are helpful for the transferability of adversarial perturbations and the VGG perceptual loss essentially utilizes the intermediate layer features of the model pretrained on the ImageNet dataset, reducing overfitting.

4) Experiments with Other Datasets

All the above experiments are carried out on StyleGAN-generated faces. Here, we further verify the anti-forensic effectiveness of the proposed method on two other face generation models, i.e., PGGAN [27] and WGAN_GP [60]. As shown in Table XIV, the proposed method also performs well with the faces generated by PGGAN and WGAN_GP, achieving an average *ASR* of 0.8236 for the PGGAN-generated faces and 0.6971 for the WGAN_GP-generated faces. Moreover, it also achieves a good trade-off between anti-forensic ability and visual quality. However, compared to the results for StyleGAN, the average *ASR*s for both PGGAN and WGAN_GP are lower. The possible reason is that the detectors have a strong forensic ability for GAN-generated faces of poor quality, and the quality of WGAN_GP and PGGAN-generated face is inferior to that of StyleGAN-generated face [61].

V. Conclusion

In this paper, an effective local anti-forensic method is proposed. The proposed method utilizes the fake faces' areas of common concern in multiple-detector decision-making to add local adversarial perturbations. Through the utilization of ensemble model and latent features, good anti-forensic performance is achieved for the white-box and zero-knowledge scenarios. The experimental results show that the anti-forensic method exposes the vulnerability of state-of-the-art forensic detectors.

However, the transferability of our method to some black-box detectors requires further enhanced. In addition, possible defense solutions, especially methods that modify detectors by retraining with adversarial examples, may affect the adversarial ability of the existing anti-forensic methods including the proposed method. Therefore, in the future, we want to focus on improving the transferability and robustness of perturbation against possible defense solutions.

References

[1] Goodfellow I, Pouget-Abadie J, Mirza M, *et al.*, "Generative adversarial nets," in Proceedings of the 28th Conference on Neural Information Processing Systems (NeurIPS 2014), pp. 2672-2680, 2014.

[2] Yang J, Xiao S, Li A, *et al.*, "MSTA-Net: forgery detection by generating manipulation trace based on multi-scale self-texture attention," IEEE Transactions on Circuits and Systems for Video Technology, doi: 10.1109/TCSVT.2021.3133859, 2021.

[3] Chen B, Ju X, Xiao B, *et al.*, "Locally GAN-generated face detection based on an improved Xception," Information Sciences, vol. 572, pp. 16-28, 2021.

[4] Yang Q, Yu D, Zhang Z, *et al.*, "Spatiotemporal trident networks: detection and localization of object removal tampering in video passive forensics," IEEE Transactions on Circuits and Systems for Video Technology, vol. 31, no. 10, pp. 4131-4144, 2021.

[5] Afchar D, Nozick V, Yamagishi J, *et al.*, "Mesonet: a compact facial video forgery detection network," in 2018 IEEE International Workshop on Information Forensics and Security (WIFS). IEEE, pp. 1-7, 2018.

[6] Wang S Y, Wang O, Zhang R, *et al.*, "CNN-generated images are surprisingly easy to spot... for now," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 8695-8704, 2020.

[7] Liu Z, Qi X, Torr P H S, "Global texture enhancement for fake face detection in the wild," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 8060-8069, 2020.

[8] Rossler A, Cozzolino D, Verdoliva L, *et al.*, "Faceforensics++: Learning to detect manipulated facial images," in Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 1-11, 2019.

[9] Chen B, Liu X, Zheng Y, *et al.*, "A Robust GAN-Generated Face Detection Method Based on Dual-Color Spaces and an Improved Xception," IEEE Transactions on Circuits and Systems for Video Technology, vol. 32, no. 6, pp. 3527-3538, 2022.

[10] Wang C, Deng W, "Representative forgery mining for fake face detection," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 14923-14932, 2021.

[11] Neves J C, Tolosana R, Vera-Rodriguez R, *et al.*, "Ganprintr: Improved fakes and evaluation of the state of the art in face manipulation detection," IEEE Journal of Selected Topics in Signal Processing, vol.14, no.5, pp. 1038-1048, 2020.

[12] Zhao X, Stamm M C, "Making GAN-generated images difficult to spot: a new attack against synthetic image detectors," arXiv preprint arXiv:2104.12069, 2021.

[13] Xie H, Ni J, Shi Y -Q, "Dual-domain generative adversarial network for digital image operation anti-forensics," IEEE Transactions on Circuits and Systems for Video Technology, vol. 32, no. 3, pp. 1701-1706, 2022.

[14] Maximov M, Elezi I, Leal-Taixé L, "Ciagan: Conditional identity anonymization generative adversarial networks," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 5447-5456, 2020.

[15] Kuang Z, Liu H, Yu J, *et al.*, "Effective De-identification Generative Adversarial Network for Face Anonymization," in Proceedings of the 29th ACM International Conference on Multimedia, pp. 3182-3191, 2021.

[16] Zhu B, Fang H, Sui Y, *et al.*, "Deepfakes for medical video de-identification: Privacy protection and diagnostic information preservation," in Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, pp. 414-420, 2020.

[17] Wang Y, Ding X, Yang Y, et al., "Perception matters: Exploring imperceptible and transferable anti-forensics for GAN-generated fake face imagery detection," Pattern Recognition Letters, vol.146, pp. 15-22, 2021.

[18] Goebel M, Manjunath B S, "Adversarial attacks on co-occurrence features for GAN detection," arXiv preprint arXiv:2009.07456, 2020.

[19] Li D, Wang W, Fan H, *et al.*, "Exploring adversarial fake images on face manifold," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5789-5798, 2021.

[20] Ding F, Zhu G, Li Y, *et al.*, "Anti-Forensics for face swapping videos via adversarial training," IEEE Transactions on Multimedia, 2021. doi: 10.1109/TMM.2021.3098422.

[21] Hussain S, Neekhara P, Jere M, *et al.*, "Adversarial deepfakes: Evaluating vulnerability of deepfake detectors to adversarial examples," in Proceedings of the IEEE/CVF winter conference on applications of computer vision, pp. 3348-3357, 2021.

[22] Liu Y, Chen X, Liu C, *et al.*, "Delving into transferable adversarial examples and black-box attacks," arXiv preprint arXiv:1611.02770, 2016.

[23] Dong Y, Liao F, Pang T, *et al.*, "Boosting adversarial attacks with momentum," in Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 9185-9193, 2018.

[24] Singh M, Sinha A, Kumari N, *et al.*, "Harnessing the vulnerability of latent layers in adversarially trained models," arXiv preprint arXiv:1905.05186, 2019.

[25] Che Z, Borji A, Zhai G, *et al.*, "SMGEA: A new ensemble adversarial attack powered by long-term gradient memories. IEEE Transactions on Neural Networks and Learning Systems, vol. 33, no. 3, pp. 1051-1065, 2022.

[26] Yu Y, Gao X, Xu C Z, "LAFEAT: Piercing through adversarial defenses with latent features," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5735-5745, 2021.

[27] Karras T, Aila T, Laine S, et al., "Progressive growing of gans for improved quality, stability, and variation," arXiv preprint arXiv:1710.10196, 2017.

[28] Karras T, Laine S, Aila T, "A style-based generator architecture for generative adversarial networks," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4401-4410, 2019.

[29] Tan M, Le Q, "EfficientNet: Rethinking model scaling for convolutional neural networks," in International Conference on Machine Learning. PMLR, pp. 6105-6114, 2019.

[30] Iandola F, Moskewicz M, Karayev S, *et al.*, "Densenet: Implementing efficient convnet descriptor pyramids," arXiv preprint arXiv:1404.1869, 2014.

[31] Luo Y, Zi H, Zhang Q, *et al.*, "Anti-forensics of jpeg compression using generative adversarial networks," in 2018 26th European Signal Processing Conference (EUSIPCO). IEEE, pp. 952-956, 2018.

[32] Nguyen H H, Tieu N D T, Nguyen-Son H Q, *et al.*, "Transformation on computer-generated facial image to avoid detection by spoofing detector," in 2018 IEEE International Conference on Multimedia and Expo (ICME). IEEE, pp. 1-6, 2018.

[33] Peng F, Yin L P, Zhang L B, *et al.*, "CGR-GAN: CG Facial Image Regeneration for Antiforensics Based on Generative Adversarial Network," IEEE Transactions on Multimedia, vol. 22, no. 10, pp. 2511-2525, 2019.

[34] Peng F, Yin L, Long M, "BDC-GAN: Bidirectional conversion between computer-generated and natural facial images for anti-forensics," IEEE Transactions on Circuits and Systems for Video Technology, 2022. doi:

10.1109/TCSVT.2022.3177238.

[35] Huang Y, Juefei-Xu F, Wang R, *et al.,* "Fakepolisher: Making deepfakes more detection-evasive by shallow reconstruction," in Proceedings of the 28th ACM international conference on multimedia, pp. 1217-1226, 2020.

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

[37] Madry A, Makelov A, Schmidt L, *et al.*, "Towards deep learning models resistant to adversarial attacks," arXiv preprint arXiv:1706.06083, 2017.

[38] Qian Y, Wang J, Wang B, et al, "Visually Imperceptible Adversarial Patch Attacks on Digital Images," arXiv preprint arXiv:2012.00909, 2020.

[39] Papernot N, McDaniel P, Jha S, *et al.*, "The limitations of deep learning in adversarial settings," in 2016 IEEE European symposium on security and privacy (EuroS&P), pp. 372-387, 2016.

[40] Su J, Vargas D V, Sakurai K, "One pixel attack for fooling deep neural networks," IEEE Transactions on Evolutionary Computation, vol. 23, no. 5, pp. 828-841, 2019.

[41] Croce F, Hein M, "Sparse and imperceivable adversarial attacks," in Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 4724-4732, 2019.

[42] Brown T B, Mané D, Roy A, *et al.*, "Adversarial patch," arXiv preprint arXiv:1712.09665, 2017.

[43] Karmon D, Zoran D, Goldberg Y, "Lavan: Localized and visible adversarial noise," in International Conference on Machine Learning. PMLR, pp. 2507-2515, 2018.

[44] Xiang T, Liu H, Guo S, *et al.*, "Local black-box adversarial attacks: A query efficient approach," arXiv preprint arXiv:2101.01032, 2021.

[45] Xiao C, Li B, Zhu J Y, *et al.*, "Generating adversarial examples with adversarial networks," arXiv preprint arXiv:1801.02610, 2018.

[46] Selvaraju R R, Cogswell M, Das A, *et al*., "Grad-cam: Visual explanations from deep networks via gradient-based localization," in Proceedings of the IEEE international conference on computer vision, pp. 618-626, 2017.

[47] Choi Y, Choi M, Kim M, *et al.*, "Stargan: Unified generative adversarial networks for multi-domain image-to-image translation," in Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 8789-8797, 2018.

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

[49] Mao X, Li Q, Xie H, *et al.*, "Least squares generative adversarial networks," in Proceedings of the IEEE international conference on computer vision. pp. 2794-2802, 2017.

[50] Gatys L A, Ecker A S, Bethge M, "Image style transfer using convolutional neural networks," in Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2414-2423, 2016.

[51] Liu Z, Luo P, Wang X, *et al.*, "Deep learning face attributes in the wild," in Proceedings of the IEEE International Conference on Computer Vision, pp. 3730-3738, 2015.

[52] Krizhevsky A, Sutskever I, Hinton G E, "Imagenet classification with deep convolutional neural networks," Advances in neural information processing systems, vol.25, pp. 1097-1105, 2012.

[53] Wu W, Su Y, Chen X, *et al.*, "Boosting the transferability of adversarial samples via attention," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1161-1170, 2020.

[54] Tang S, Gong R, Wang Y, *et al.*, "Robustart: Benchmarking robustness on architecture design and training techniques," arXiv preprint arXiv:2109.05211, 2021.

[55] Wang Z, Guo H, Zhang Z, *et al.*, "Feature importance-aware transferable adversarial attacks," in Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 7639-7648, 2021.

[56] Osadchy M, Hernandez-Castro J, Gibson S, *et al.*, "No bot expects the DeepCAPTCHA! Introducing immutable adversarial examples, with applications to CAPTCHA generation," IEEE Transactions on Information Forensics and Security, vol. 12, no. 11, pp. 2640-2653, 2017.

[57] Guo C, Rana M, Cisse M, *et al.*, "Countering adversarial images using input transformations," arXiv preprint arXiv:1711.00117, 2017.

[58] Meng D, Chen H, "Magnet: a two-pronged defense against adversarial examples," in Proceedings of the 2017 ACM SIGSAC conference on computer and communications security, pp. 135-147, 2017.

[59] Sharif M, Bauer L, Reiter M K, "On the Suitability of Lp-Norms for Creating and Preventing Adversarial Examples," 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 1686-16868, 2018.

[60] Gulrajani I, Ahmed F, Arjovsky M, *et al.*, "Improved training of wasserstein gans," Advances in neural information processing systems, pp. 30, 2017.

[61] Wang Z W, She Q, Ward T E, "Generative Adversarial Networks in Computer Vision: A Survey and Taxonomy," ACM Computing Surveys, vol. 54, no. 37, pp. 1-38, 2021.

Haitao Zhang received the B.S. degree in Engineering in 2020 from Nanjing University of Information Science & Technology, Nanjing, China. He is currently pursuing the M.S. degree in Computer Science from Nanjing University of Information Science & Technology, Nanjing, China. His research interests include image processing, and image forensics.

Beijing Chen received the Ph.D. degree in Computer Science in 2011 from Southeast University, Nanjing, China. Now he is a Professor in the School of Computer, Nanjing University of Information Science & Technology, China. His research interests include color image processing, image forensics, image watermarking, and pattern recognition. He serves as an Editorial Board Member of the Journal of Mathematical Imaging and Vision.

Jinwei Wang received the Ph.D. degree in information security from the Nanjing University of Science and Technology in 2007. He is currently a Professor with the Nanjing University of Information Science and Technology. He has published over 50 papers. His research interests include multimedia copyright protection, multimedia forensics, multimedia encryption, and data authentication.

Guoying Zhao received the Ph.D. degree in Computer Science from the Chinese Academy of Sciences, Beijing, China, in 2005. She is currently a full Professor with Center for Machine Vision and Signal Analysis, University of Oulu, Finland. She has authored or coauthored more than 240 articles in journals and conferences. Her articles have currently over 15,600 citations in Google scholar (H-index 57). She is a Fellow of the IEEE and the IAPR. She has served as the area chairs for several conferences. She is an Associate Editor of Pattern Recognition, IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, and Image and Vision Computing journal.