Research Article

Research on Multimodality Face Antispoofing Model Based on Adversarial Attacks

Junjie Mao,^{1,2,3} Bin Weng,^{1,2,3} Tianqiang Huang^{1,2,3} ,^{1,2,3} Feng Ye,^{1,2,3} and Liqing Huang^{1,2,3}

¹College of Mathematics and Informatics, Fujian Normal University, Fuzhou 350007, China

²Digital Fujian Institute of Big Data Security Technology, Fuzhou 350007, China

³Fujian Provincial Engineering Research Center of Big Data Analysis and Application, Fuzhou 350007, China

Correspondence should be addressed to Tianqiang Huang; fjhtq@fjnu.edu.cn

Received 18 April 2021; Accepted 22 July 2021; Published 10 August 2021

Academic Editor: Zhili Zhou

Copyright © 2021 Junjie Mao et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Face antispoofing detection aims to identify whether the user's face identity information is legal. Multimodality models generally have high accuracy. However, the existing works of face antispoofing detection have the problem of insufficient research on the safety of the model itself. Therefore, the purpose of this paper is to explore the vulnerability of existing face antispoofing models, especially multimodality models, when resisting various types of attacks. In this paper, we firstly study the resistance ability of multimodality models when they encounter white-box attacks and black-box attacks from the perspective of adversarial examples. Then, we propose a new method that combines mixed adversarial training and differentiable high-frequency suppression modules to effectively improve model safety. Experimental results show that the accuracy of the multimodality face antispoofing model is reduced from over 90% to about 10% when it is attacked by adversarial examples. But, after applying the proposed defence method, the model can still maintain more than 90% accuracy on original examples, and the accuracy of the model can reach more than 80% on attack examples.

1. Introduction

Facial recognition has been gradually integrated into our daily life nowadays. Its application in mobile payment, security monitoring, and other fields is becoming more and more extensive. For example, in the ticket gates of the Chinese railway station, face recognition ticket entry has replaced traditional manual ticket verification, which greatly speeds up the entry process. At the same time, Alipay and other software launched the function of face scan payment, and users do not need to carry mobile devices such as cell phones, as long as the face recognition system that detects the real face bound to the account can realize the face payment.

However, previous research has found that facial recognition systems are easily spoofed by various face presentation attacks (PAs) [1–3] (as shown in Figure 1). These attacks include print attacks, video replay attacks, and 3D mask attacks. Print attacks, as shown in Figure 1(b), are where the face image of a legitimate user is printed on paper to attack the facial recognition systems. Video replay attacks,

as shown in Figures 1(c) and 1(d), are to replay the face that includes blinking, head movement, and facial expressions on a digital device. 3D mask attacks, as shown in Figure 1(e), are where the 3D face synthesis technology is used to make face hoods or masks. In 2018, there was a case where a gang used software to create 3D avatars of relevant citizens' avatars, so as to pass Alipay's facial recognition authentication and register the personal information of the aforementioned users to obtain rewards for new users. Therefore, a safe face antispoofing detection model to distinguish whether there is a live face is research-wise significant and valuable.

In recent years, scholars' research on face antispoofing detection can be roughly categorized into methods based on hand-crafted features and methods based on deep neural networks. Methods based on hand-crafted features are mostly classified according to manually extracted spatial texture features [4–8] or temporal information between video frames [9–12], but the features extracted by such methods are single, and attackers can use low-cost methods to deal with the corresponding features of the image in a

FIGURE 1: Face presentation attacks. (a) Live face. (b) Print attack. (c) Video replay attack, eye-open state. (d) Video replay attack, eye-closed state. (e) 3D mask attack.

targeted manner to deceive the detector. Methods based on deep neural networks aim to design a unified feature extraction and classification model through network architectures such as CNN [13, 14] and LSTM [15]. The feature information extracted by the deep neural network is diverse and no longer depends on a single feature for judgment. Therefore, the existent deep learning-based methods can generally achieve better performance.

On the other hand, the face antispoofing detection methods can be divided into single-modality and multimodality models. The single-modality model refers to a single type of image input such as RGB or Depth images into the network, and CNN-LSTM [16], CNN-RNN [17], or other network structures are used to extract temporal or spatial information. However, in order to improve the generalization and robustness of the model, recent research has gradually tended to adopt multimodality models. To this end, in addition to capturing the human face through the conventional visible light camera, researchers also use depth cameras and infrared cameras to capture depth maps and infrared maps. The input stream of the network includes RGB, Depth, and IR images, which allows the model to increase the learning of the inherent differences between live and spoofing faces in terms of reflection properties, three-dimensional structure, and so forth. For example, Yu et al. [2] proposed a multimodality network structure based on differential convolution. This method effectively merged the image detail information in three input streams and shows excellent performance in detecting presentation attacks.

Although neural network-based face antispoofing models have achieved good performance in the face of presentation attacks, a small number of security researches on single-modality models [18–21] have shown that spoofing face will be incorrectly recognized by the model as the real face after adding small perturbations, which poses a new challenge to the security of the model. However, the above research is only conducted in the convolutional neural network with a relatively simple structure and did not involve multimodality detection models that have achieved excellent performance in recent years; meanwhile, the generation of perturbed images only involved RGB images and did not involve Depth and IR images. Therefore, this paper focuses on the safety of multimodality models in the face of adversarial attacks, and the contributions of this paper are summarized as follows:

- (1) We select the advanced models of single-modality and multimodality face antispoofing and verify the vulnerability of the models in the face of white-box and black-box attacks on RGB, Depth, and IR images.
- (2) We assess white-box and black-box attacks on a single-input stream in the multimodality model. Through the experiments, we find that the multimodality detection model is robust against single-modality attacks.
- (3) We explored the security of single-modality and multimodality models against different numbers of patch attacks.
- (4) We propose a new defence method to combine hybrid adversarial training and diffusible high-frequency suppression modules in single-modality and multimodality models to improve the security of models. Experimental results demonstrate that the security of the multimodality model is significantly higher than that of the single-modality model. At the same time, this is the first time that the research on the multimodality model's adversarial attacks has been proposed.

The rest of the paper is structured as follows. Section 2 reports a literature review of the latest related work. Section 3 reports all the details about the proposed method. Section 4 details the experiment. Finally, Section 5 concludes the paper.

2. Related Work

While deep learning is widely used in computer vision, there are also obvious security vulnerabilities in deep neural networks, that is, adding certain imperceptible perturbations to an image may lead the model to make wrong prediction. To counter this threat, adversarial attacks and defence methods are becoming a research hotspot.

2.1. Adversarial Attack Methods. Most of the existing attacks are divided into two categories: white-box attacks and black-box attacks. A white-box attack is one in which the attacker is able to know the algorithm and parameters of the model in advance, while a black-box attack does not get this

information. Ma et al. [18] proposed a white-box attack algorithm that considers the concentration property of the human eye visual cascade and concentrates the perturbations on a few dimensions to make the generated adversarial examples imperceptible. Goswami et al. [19] evaluated the robustness of face recognition models and demonstrated the impact on models through four imperceptible perturbations. Bousnina et al. [20] achieved high attack success rate by a black-box attack method in a CNN network based on transfer learning. Yang et al. [21] introduced a new Attentional Adversarial Attack Generation Network (A^3GN) as a way to generate the same adversarial examples as the original face images; unlike the traditional GAN, A^3GN uses the face recognition network for the third player to participate in the competition between the generator and the discriminator. Yang et al. [22] observed that ℓ_0 -norm has good sparsity but is hard to solve; for this reason, they proposed to use ℓ_q -norm to approximate ℓ_0 -norm and attack the face antispoofing task with the goal of minimizing the ℓ_q distance from the original image. Zhang et al. [23] shifted the attacks from digital domain to physical domain and verified the realistic threat of adversarial attacks in face antispoofing detection. However, there is a certain lack of research on whether the above algorithms can successfully attack multimodality face antispoofing models with excellent performance.

2.2. Defence Methods. Existing defences can be broadly classified into active defences and passive defences. Xie et al. [24] proposed adding a randomization layer before the classifier to transform the input image. Dhillon et al. [25] proposed Stochastic Activation Pruning (SAP) to the pre-trained networks to improve robustness. Liao et al. [26] used a denoising network in the form of U-Net to remove the perturbation noise in the adversarial examples. Chen et al. [27] observed by comparing heatmaps between clean examples and adversarial examples that the adversarial instances are able to deceive DNNs by scattering key regions of the images and blurring the object contours; for this reason, they proposed a defence method (referred to as RCA-SOC) that refocuses on key regions and reinforces object contours. All the above methods are studied from the perspective of passive defence, which has limited robustness and transferability. Goodfellow et al. [28] were the first to propose an adversarial training method that adds adversarial examples to the model training process from the perspective of active defence, which allows the model to learn the perturbation features directly. Shafahi et al. [29] improved the traditional adversarial training method by proposing recycling the gradient information computed when updating the model parameters to reduce the overhead. In the study of face

antispoofing detection, Ma et al. [30] proposed a multiregion convolutional neural network for the shortcoming that the gradient of the face representation attacks detector is concentrated in a local region and thus is vulnerable to adversarial example attacks, but the method was not explored in multimodality model. Although the research on defence methods has been rapidly developed in recent years, there are still few studies on defence against multimodality models.

3. Proposed Defence Method

The proposed work aims to achieve the purpose of deceiving the model to make its classification wrong by adding small perturbation values to the images. For this purpose, two white-box attacks and one black-box attack are used in this paper. At the same time, this paper combines the methods of active defence and passive defence; we propose a combination of mixed confrontation training and diffusible high-frequency suppression module to improve the safety of the model. The overall framework is shown in Figure 2. In Figure 2, the spoofed image is correctly recognized when fed into the original face antispoofing model, while the model is incorrectly predicted when the image is attacked. For this reason, this paper proposes a defence method so that the attacked image can also be correctly predicted. The red-box part in Figure 2 indicates the attack process of this paper, and the green-box part shows the defence method of this paper. We will explain each part of the framework in detail.

In the defensive research of adversarial examples, the defence method of adversarial training can enable the model to directly learn the perturbation characteristics through retraining. Therefore, it is widely applied in resisting adversarial attacks and has achieved excellent performance. Existing adversarial training methods used an original example and an adversarial example generated from the original one for collaborative training. Due to the differences in the disturbance characteristics formed by different attack methods, this type of methods may cause the defensive effect of confrontation training to be unable to defend against multiple attack types at the same time. Consequently, in order to better improve the security of the model in the face of multiple adversarial attacks, we proposed a hybrid adversarial training method to directly learn multiple disturbance features. The proposed method generates multiple different adversarial examples on the same original example. In detail, this paper uses a method of two white-box attacks and one black-box attack to generate adversarial examples and train them together with the original one. Therefore, the optimization objectives of the mixed adversarial training proposed in this paper are

$$\min_{\omega} \left\{ L(f_{\omega}(z), t) + \rho \max_{\|\delta_i\| \leq \epsilon} L(f_{\omega}(z), \alpha_1 f_{\omega}(z + \delta_1) + \alpha_2 f_{\omega}(z + \delta_2) + \alpha_3 f_{\omega}(z + \delta_3)) \right\}, \quad (1)$$

FIGURE 2: The framework of the adversarial attacks and defence process of the face antispoofing detection model.

where z represents the original input image with label t , f represents the predicted classification probability under the model parameters ω , L is the cross-entropy loss, δ_i indicates the different sizes of disturbance added in the adversarial example, and its perturbation threshold should not be greater than ϵ . The first term of equation (1) prompts the model to learn features of clean examples, while the second term allows the model to learn multiple perturbation features, where the trade-off between different adversarial examples is regulated by α_i and the trade-off between the original and adversarial examples is regulated by ρ .

Since the hybrid adversarial training method learns the characteristics of the original image while also learning the perturbation features of the adversarial example, the detection accuracy of the model on original examples after the hybrid adversarial training is lower than that of the original network. For this reason, we proposed a combination method of active defence and passive defence to improve the robustness of the model. For the passive defence, we adopt the differentiable high-frequency suppression module proposed by Zhang et al. [31]; that is, given an input image, first convert the image to the frequency domain through the Discrete Fourier Transform (DFT), then suppress high-frequency components in the frequency domain, and finally convert the corrected image back to the spatial domain. Our work is the first to apply this method to face antispoofing detection to resist adversarial attacks.

Specifically, the input image is expressed as $z \in \mathbb{R}^{M \times N}$; then the corresponding frequency domain map $z' \in \mathbb{C}^{M \times N}$ can be calculated as

$$z'_{u,v} = \sum_{a=0}^{M-1} \sum_{b=0}^{N-1} z_{a,b} e^{-j2\pi((u/M)a + (v/N)b)}, \quad (2)$$

where (a, b) is the coordinate of space domain and (u, v) represents the coordinate of frequency domain.

The suppression module can be expressed as

$$z' \leftarrow P \odot z', \quad (3)$$

where $\odot$ represents the dot product of the pixel and $P \in \mathbb{R}^{M \times N}$ is set as a high-frequency suppression box with the radius of r :

$$P_{u,v} = \begin{cases} 1, & 0 \leq |u|, |v| \leq r, \\ 0, & \text{else.} \end{cases} \quad (4)$$

We set $P_{-u,-v} = P_{u,v}$ and $P_{.,-v} = P_{.,N-v}$. Formula (4) indicates that, for high-frequency pixels, $P_{u,v}$ is set as 0, and, for low-frequency pixels, the original characteristics should be retained, so $P_{u,v}$ is set as 1. Therefore, the final suppression module can be expressed as

$$z \leftarrow F^{-1}(P \odot F(z)). \quad (5)$$

In the above formula, $F(\cdot)$ and $F^{-1}(\cdot)$ are the Fourier transform and inverse Fourier function, respectively. The proposed method fuses the diffusible high-frequency suppression module as a preprocessing layer into the face antispoofing detection model to suppress high-frequency disturbance components.

4. Results and Discussion

4.1. Dataset and Face Antispoofing Model. We use CASIA-SURF dataset published by Zhang et al. [32] for experiments. Each sample contains three types of images, namely, RGB, Depth, and IR. These three modal images are segmented from the complex background of the face area with cropping and alignment; an example of the dataset is shown in Figure 3. The dataset contains 21,000 video clips of 1,000 Chinese people. Each person has recorded multiple samples. Each sample contains 1 real recorded video clip, and 6 fake

FIGURE 3: Examples of CASIA-SURF dataset.

face videos are produced by different deception methods. In these different deception methods, the eyes, the nose, the mouth, and other parts will be cut off in the printed pictures or curved pictures, or they will be recombined. In this paper, we focus on the criminals in the real situation through deception to make the detection model recognize the attacking pictures as real faces to obtain illegitimate benefits. Therefore, the testing set only selects the images containing six spoofing methods to conduct adversarial attacks and defensive experiments.

In this paper, we evaluated the results of adversarial attacks and defence based on the model proposed by Shen et al. [1]. As shown in Figure 4, the model has two versions, single-modality input and multimodality input. The single-modality version inputs RGB, Depth, and IR images into the model, respectively, and the multimodality version inputs three image types into the model in parallel. On the input side, the model does not input the entire face image into the network but randomly extracts N small patches of learning features in the image. At the output end, the model judges whether the complete face is a live face through the average of the class probabilities obtained by the prediction of N image patches. In our experiments, we set N as 36. In the design of Shen et al. [1], the multimodality model differs from the single-modality model in that the spatial features extracted from multiple modalities are feature-fused after the third residual block and combined with a multimodality feature erasure operation that erases randomly selected modal features to enhance feature learning and prevent overfitting. For a more realistic demonstration of the experimental effect, we use the pretrained model of Shen et al. [1] to conduct attack experiments. The parameter settings remain consistent with the original. The size of the image patch is 32×32 , which is the highest detection accuracy in the model.

4.2. Experimental Environment and Evaluation. In our experiments, the running environment is Ubuntu 16.04 system, and four NVIDIA Tesla P100 GPUs are used to train and test the model. Our code is implemented on PyTorch 1.3.0 deep learning framework.

To evaluate the performance of the model before and after being attacked by the adversarial examples, we introduce the accuracy metric to evaluate the model. This metric

indicates that the accurate numbers of live faces and spoof faces are predicted to account for all tested live faces and spoof faces, and the metric can be computed as follows:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (6)$$

where TP indicates that the live face is predicted to be a live face, TN indicates that the spoof face is predicted to be a spoof face, FP indicates that the spoof face is predicted to be a live face, and FN indicates that the live face is predicted to be a spoof face.

4.3. Adversarial Attacks on Single-Modality and Multimodality Models. In previous works [18, 19] on face anti-spoofing models against attacks, people used relatively basic single-modality convolutional networks to conduct experiments and achieved good attack results. However, the above research did not test the attack effect on Depth and IR images, such as [21]. Meanwhile, as the robustness models in this field continue to be researched in depth, whether existing models, especially multimodality models, can resist adversarial attacks remains to be experimented. Therefore, the experiment in this section first conducts adversarial attacks on a single-modality model based on RGB, Depth, and IR images and a multimodality model fused with three image types to explore the robustness of face anti-spoofing detection models against adversarial examples.

In this paper, we choose FGSM [28] and DeepFool [33] methods to carry out white-box attacks and use NATTACK method [34] to carry out black-box attack. The attack methods of these three methods cover single-step attacks, iterative attacks, and attacks based on probability density distribution. Through experiments, the perturbation value added by FGSM is finally set to 0.01, the maximum number of iterations of DeepFool is set to 50, the standard deviation σ of NATTACK is set to 0.1, and the learning rate α is 0.2. We used the same training parameters provided by the original author [1], and the experimental results are shown in Table 1. It can be seen from Table 1 that, in the attacks on the single-modality face anti-spoofing model, adding small perturbations to the spoofed RGB, Depth, and IR images causes the detection accuracy of the model to drop sharply, with an average accuracy of only about 14%, which indicates that the single-modality model is very sensitive to the adversarial attacks. At the same time, it can be seen from Table 1 that when the model faces the FGSM adversarial examples formed on the Depth images, its detection accuracy is 63.9%, which is relatively small compared to the other two modalities. Therefore, we further increase the perturbation value for the FGSM attack on the Depth images. When the perturbation value is set to 0.02, the model's detection accuracy drops to 34.5%; and, in the attacks on the multimodality detection model, adding small perturbations to all three types of images with parallel inputs can reduce the detection accuracy of the model from 99.5% to about 10% on average. At the same time, the results found that the attack success rate of NATTACK attacks is not as good as that of white-box attacks. The experiment in Table 1 shows

FIGURE 4: The structure of the multimodality face antispoofing model proposed by Shen et al. [1].

TABLE 1: The accuracy of single-modality and multimodality models under attacks.

Data stream type	Original sample (%)	FGSM, 0.01 perturbation (%)	DeepFool (%)	NATTACK (%)
RGB	94.1	2.5	8	24.3
Depth	98.5	63.9	4.9	7.6
IR	96.3	1	10.7	12.4
Fusion	99.5	2.7	0.7	31.5

that adversarial attack is still a major challenge in this field, and the robustness of the multimodality face antispoofing model needs to be further improved.

Figures 5–7 further show examples of single-modality model attacks against RGB, Depth, and IR images, respectively. Since the black-box attack is not accessible to the model, only the input and output class probability vectors of the network are known, and the output of this paper is the class probability prediction mean of 36 image patches extracted from the complete image, so we only show the example of the white-box attacks. Through observation, it can be found that, after adding a small perturbation, the attacked patches and the original patches are almost visually consistent.

4.4. Multimodality Model Resists Single-Modality Adversarial Attacks. The above experiments demonstrate that the multimodality model is difficult to resist the attack of adversarial examples, but these experiments are performed on all three data streams, and the attack cost is relatively high. For this reason, we also performed white-box and black-box attacks on a single-modality model (only one image type is attacked at a time, and the rest of the input streams remain unchanged) to explore whether it can achieve similar attack effects. The experiments of the FGSM attack used multiple perturbation values in this section; the results are shown in Table 2.

From Table 2, we find that, in the multimodality model, attacking only one modality cannot successfully deceive the network. When the perturbation is added to 0.2, the accuracy of the model drops significantly, but the images have

lost their original semantic information. As shown in Figure 8, compared with the original image patches, the structural information at the nose of the face in the FGSM adversarial examples with perturbation of 0.2 is difficult to distinguish.

4.5. Attacks on Different Number Patches. Since the experiments in this paper used a face antispoofing model based on image patches extraction features proposed by Shen et al. [1], while the above experiments are conducted against attacks on all 36 image patches randomly extracted from the complete image, we conjecture whether only adding perturbations to part of the 36 image patches can achieve the same attack effect. We performed white-box and black-box attacks on 36 32×32 random image patches extracted from the input image of the model and selected 20, 26, 28, and 30 image patches, respectively. The results are shown in Figure 9.

The results of Figure 9 show that when perturbations are randomly added to 20 image patches, an obvious attack effect cannot be achieved. When the attacked image patches reach 26 blocks, the detection accuracy of the model drops significantly. But, for Depth images, reducing the number of attacked image patches allows the model to maintain high detection accuracy. Consequently, we believe that it is necessary to add perturbation to 36 randomly extracted image patches in this paper to achieve a higher attack success rate. Since the model has shown a certain degree of robustness in the attack experiments on some image patches, this paper does not explore the defence methods of different number patches.

FIGURE 5: An example of single-modality model attacking the RGB image, where (a) represents FGSM attack with the perturbation value of 0.01 and (b) represents DeepFool attack.

FIGURE 6: An example of single-modality model attacking the Depth image, where (a) represents FGSM attack with the perturbation value of 0.01 and (b) represents DeepFool attack.

FIGURE 7: An example of single-modality model attacking the IR image, where (a) represents FGSM attack with the perturbation value of 0.01 and (b) represents DeepFool attack.

TABLE 2: The accuracy of multimodality model after resisting the single-modality adversarial attacks.

Attack method	Attack RGB (%)	Attack Depth (%)	Attack IR (%)
Original sample	99.5	99.5	99.5
FGSM 0.01	99.2	99.1	99.2
FGSM 0.1	99	88.4	98.1
FGSM 0.2	46.4	58.3	83.2
FGSM 0.25	33.3	37.1	65.7
DeepFool	99.3	99.1	99.2
NATTACK	97.1	93.7	97.2

FIGURE 8: RGB, Depth, and IR images formed under the FGSM attack with perturbations of 0.1 and 0.2, respectively. The first line represents the original image patches, and the second line represents the FGSM attacks with perturbation of 0.1. The third line represents the FGSM attacks with perturbation of 0.2.

FIGURE 9: Choosing different numbers of image patches for white-box and black-box attacks. The abscissa represents the number of attacked image patches, and the ordinate represents the detection accuracy of the model in the face of the attacks. The first line shows the results on RGB images and Depth images, respectively; the second row shows the results on IR images and multimodality images, respectively. (a) RGB. (b) Depth. (c) IR. (d) Fusion.

4.6. Research on the Defence of Single-Modality and Multimodality Models. In previous defensive studies of face anti-spoofing detection, there are few defensive methods for multimodality models. The defence methods proposed in [25, 27] are applied in the field of image classification and object detection. In this paper, we propose a new defence method that combines hybrid adversarial training and diffusible high-frequency suppression modules to improve model security.

We first use the differentiable high-frequency suppression module proposed by Zhang et al. [31] to filter the disturbance characteristics. In the experiment, the hyperparameter r is set as 4. The results in Tables 3–6 show that the Fourier transform is used to suppress high-frequency information. It can be seen that this method is not very effective in improving the detection accuracy of the model after being attacked. The average detection accuracy is only about 50%, but this method can make the attacked model maintain an average detection accuracy of about 94% on original examples.

Aiming at the problem that the differentiable high-frequency suppression module is not effective in improving the security of the model in the face of adversarial attacks, we propose a new hybrid adversarial training method to learn perturbation features more directly, so we choose to generate the adversarial examples in a manner that is consistent with the attack phase. The optimal parameters are selected through experiments, so finally the perturbation value of the FGSM adversarial examples is set to 0.001, the maximum iteration number of the DeepFool adversarial examples is set to 10, the standard deviation σ of the NATTACK adversarial examples is set to 0.1, the learning rate α is 0.2, the hyperparameter in the mixed adversarial training optimization objective is set to 1, and the weight ratio parameters among the three adversarial examples of FGSM (α_1), DeepFool (α_2), and Nattack (α_3) are 0.3, 0.3, and 0.4, respectively. From the results in Tables 3–6, it is shown that the mixed adversarial training method can make the detection accuracy of the model when facing adversarial examples

TABLE 3: The accuracy of the model before and after adding the defence module to the RGB images.

	The original authors' pretrained model (%)	After adding the diffusible high-frequency suppression (%)	After using the mixed adversarial training (%)	Combination of the high-frequency suppression module and the mixed adversarial training (%)
Original	94.1	94.4	82.3	92.7
FGSM 0.01	2.5	8.9	70.8	81.5
DeepFool	8	60.3	72.6	81
NATTACK	24.3	76.6	63.25	85.96

TABLE 4: The accuracy of the model before and after adding the defence module to the Depth images.

	The original authors' pretrained model (%)	After adding the diffusible high-frequency suppression (%)	After using the mixed adversarial training (%)	Combination of the high-frequency suppression module and the mixed adversarial training (%)
Original	98.5	88.3	93.65	90.3
FGSM 0.01	63.9	70.2	82.3	89.7
DeepFool	4.9	50.8	64.5	76.6
NATTACK	7.6	55.98	91.2	84.5

TABLE 5: The accuracy of the model before and after adding the defence module to the IR images.

	The original authors' pretrained model (%)	After adding the diffusible high-frequency suppression (%)	After using the mixed adversarial training (%)	Combination of the high-frequency suppression module and the mixed adversarial training (%)
Original	96.3	95.2	77	91.9
FGSM 0.01	1	57.3	66	79.1
DeepFool	10.7	14.5	67.6	82.3
NATTACK	12.4	46.3	78.1	89.6

TABLE 6: The accuracy of the model before and after adding the defence module to the multimodality model.

	The original authors' pretrained model (%)	After adding the diffusible high-frequency suppression (%)	After using the mixed adversarial training (%)	Combination of the high-frequency suppression module and the mixed adversarial training (%)
Original	99.5	97.58	96.8	99.7
FGSM 0.01	2.7	25.8	95.2	92
DeepFool	0.7	64.1	93.5	81.4
NATTACK	31.5	68	91.9	96

have a certain improvement compared with the method of the differentiable high-frequency suppression module; especially for multimodality models, the improvement effect is most obvious, with an average improvement of 40% in accuracy. However, this method would reduce the detection accuracy of the model on original examples to a certain extent, and the detection accuracy on the single-modality model will drop by up to 20%.

For this reason, we attempt to effectively combine the method of the diffusible high-frequency suppression module with the proposed hybrid adversarial training method to improve the model's ability to resist attacks from adversarial examples while maintaining the model's high detection accuracy on original examples. The combined result is also shown in Tables 3–6.

The experimental results in Tables 3–6 show that the defence method using the combination of the diffusible high-frequency suppression module and hybrid adversarial training can keep the detection accuracy of the model above 90% on all original examples, where the accuracy in the

multimodality model is 0.2% higher than that on the pre-trained model of the original authors. At the same time, this defence method has a good effect on improving the security of the model in the face of white-box and black-box attacks, which can make the detection accuracy of the model remain mostly above 80%. The experiment in Tables 3–6 further validates the effectiveness and the progressiveness of the proposed method.

5. Conclusions

Face antispoofing detection aims to identify whether the user's face identity information is legal. However, the existing face antispoofing detection methods have security vulnerabilities. This paper proposes a new method that combines mixed adversarial training and differentiable high-frequency suppression modules to effectively improve model safety. First, we demonstrate the sensitivity of the method in the face of adversarial examples by performing white-box and black-box attacks on the single-modality and

multimodality face antispoofing models of Shen et al. with excellent performance. Second, by performing single-modality white-box and black-box attacks on the multimodality detection model, the results show that the multimodality model has stronger robustness compared to the single-modality model. Third, we try to perform white-box and black-box attacks on different number of patches of the input model and find that although there is an attack effect, the model still has a certain detection ability. Finally, the effectiveness and superiority of the proposed defence method are verified by experiments.

The adversarial attacks and defence experiments in this paper show that adversarial attacks are still a major challenge in the field of face antispoofing detection. In the future, we can try to apply more attack types and defence methods in a variety of face antispoofing models to evaluate existing models more comprehensively.

Data Availability

The CASIA-SURF dataset used to support the findings of this study can be obtained by request to the authors of the paper “CASIA-SURF: A Large-Scale Multi-Modal Benchmark for Face Anti-Spoofing.”

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

This paper was supported by National Key R&D Program Special Fund (Grant no. 2018YFC1505805), the National Natural Science Foundation of China (nos. 62072106 and 61070062), General Project of Natural Science Foundation in Fujian Province (no. 2020J01168), Innovation Strategy Research Project of Fujian Province (no. 2020R0178), and Open Project of Fujian Key Laboratory of Severe Weather (no. 2020KFKT04).

References

- [1] T. Shen, Y. Huang, and Z. Tong, “Facebagnet: bag-of-local-features model for multi-modal face anti-spoofing,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, Long Beach, CA, USA, June 2019.
- [2] Z. Yu, Y. Qin, X. Li et al., “Multi-modal face anti-spoofing based on central difference networks,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 650–651, Seattle, WA, USA, June 2020.
- [3] Q. Yang, X. Zhu, J. K. Fwu et al., “PipeNet: selective modal pipeline of fusion network for multi-modal face anti-spoofing,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 644–645, Seattle, WA, USA, June 2020.
- [4] J. Määttä, A. Hadid, and M. Pietikäinen, “face spoofing detection from single images using micro-texture analysis,” in *Proceedings of the 2011 International Joint Conference on Biometrics*, pp. 1–7, Washington, DC, USA, October 2011.
- [5] J. Komulainen, A. Hadid, and M. Pietikäinen, “Context based face anti-spoofing,” in *Proceedings of the 2013 IEEE Sixth International Conference on Biometrics*, pp. 1–8, Arlington, VA, USA, September 2013.
- [6] K. Patel, H. Han, and A. K. Jain, “Secure face unlock: spoof detection on smartphones,” *IEEE Transactions on Information Forensics and Security*, vol. 11, no. 10, pp. 2268–2283, 2016.
- [7] Z. Boulkenafet, J. Komulainen, and A. Hadid, “Face anti-spoofing based on color texture analysis,” in *Proceedings of the 2015 IEEE International Conference on Image Processing*, pp. 2636–2640, Quebec City, Canada, September 2015.
- [8] J. Li, Y. Wang, T. Tan, and A. K. Jain, “Live face detection based on the analysis of fourier spectra,” in *Biometric Technology For Human Identification*, pp. 296–303, International Society for Optics and Photonics, Bellingham, WA, USA, 2004.
- [9] G. Pan, L. Sun, Z. Wu, and S. Lao, “Eyeblink-based anti-spoofing in face recognition from a generic webcam,” in *Proceedings of the 2007 IEEE 11th International Conference on Computer Vision*, pp. 1–8, Rio de Janeiro, Brazil, October 2007.
- [10] K. Kollreider, H. Fronthaler, M. I. Faraj, and J. Bigun, “Real-time face detection and motion analysis with application in “I” a,” *IEEE Transactions on Information Forensics and Security*, vol. 2, no. 3, pp. 548–558, 2007.
- [11] J. Komulainen, A. Hadid, and M. Pietikäinen, “Face spoofing detection using dynamic texture,” in *Proceedings of the Asian Conference on Computer Vision*, pp. 146–157, Daejeon, Korea, November 2012.
- [12] W. Bao, H. Li, N. Li, and J. Wei, “A liveness detection method for face recognition based on optical flow field,” in *Proceedings of the 2009 International Conference On Image Analysis And Signal Processing*, pp. 233–236, Linhai, China, 2009.
- [13] L. Li, X. Feng, Z. Boulkenafet et al., “An original face anti-spoofing approach using partial convolutional neural network,” in *Proceedings of the 2016 Sixth International Conference On Image Processing Theory*, pp. 1–6, Oulu, Finland, 2016.
- [14] J. Yang, Z. Lei, and S. Z. Li, “Learn convolutional neural network for face anti-spoofing,” 2014, <https://arxiv.org/abs/1408.5601>.
- [15] L. Feng, L.-M. Po, Y. Li et al., “Integration of image quality and motion cues for face anti-spoofing: a neural network approach,” *Journal of Visual Communication and Image Representation*, vol. 38, pp. 451–460, 2016.
- [16] Z. Xu, S. Li, and W. Deng, “Learning temporal features using LSTM-CNN architecture for face anti-spoofing,” in *Proceedings of the 2015 3rd IAPR Asian Conference On Pattern Recognition*, pp. 141–145, Kuala Lumpur, Malaysia, November 2015.
- [17] Y. Liu, A. Jourabloo, and X. Liu, “Learning deep models for face anti-spoofing: binary or auxiliary supervision,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 389–398, Salt Lake City, UT, USA, June 2018.
- [18] Y. Ma and L. Wu, “Algorithm to generate adversarial examples for face-spoofing detection,” *Ruan Jian Xue Bao/ Journal of Software*, vol. 30, no. 2, pp. 469–480, 2019.
- [19] G. Goswami, A. Agarwal, N. Ratha, R. Singh, and M. Vatsa, “Detecting and mitigating adversarial perturbations for robust face recognition,” *International Journal of Computer Vision*, vol. 127, no. 6, pp. 719–742, 2019.

- [20] N. Bousnina, L. Zheng, M. Mikram, S. Ghouzali, and K. Minaoui, "Unraveling robustness of deep face anti-spoofing models against pixel attacks," *Multimedia Tools and Applications*, vol. 80, no. 5, pp. 7229–7246, 2021.
- [21] L. Yang, Q. Song, and Y. Wu, "Attacks on state-of-the-art face recognition using attentional adversarial attack generative network," *Multimedia Tools and Applications*, vol. 80, no. 1, pp. 855–875, 2021.
- [22] L. Yang, J. Yang, M. Peng et al., "Sparse adversarial attack based on ℓ_q -norm for fooling the face anti-spoofing neural networks," *Journal of Electronic Imaging*, vol. 30, no. 2, 2021.
- [23] B. Zhang, B. Tondi, and M. Barni, "Attacking CNN-based anti-spoofing face authentication in the physical domain," 2019, <https://arxiv.org/abs/1910.00327>.
- [24] C. Xie, J. Wang, Z. Zhang, Z. Ren, and A. Yuille, "Mitigating adversarial effects through randomization," 2017, <https://arxiv.org/abs/1711.01991>.
- [25] G. S. Dhillon, K. Azizzadenesheli, Z. C. Lipton et al., "Stochastic activation pruning for robust adversarial defense," 2018, <https://arxiv.org/abs/1803.01442>.
- [26] F. Liao, M. Liang, Y. Dong, T. Pang, X. Hu, and J. Zhu, "Defense against adversarial attacks using high-level representation guided denoiser," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1778–1787, Salt Lake City, UT, USA, June 2018.
- [27] J. Chen, H. Zheng, R. Chen, and H. Xiong, "RCA-SOC: a novel adversarial defense by refocusing on critical areas and strengthening object contours," *Computers & Security*, vol. 96, pp. 101–106, 2020.
- [28] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," 2014, <https://arxiv.org/abs/1412.6572>.
- [29] A. Shafahi, M. Najibi, A. Ghiasi et al., "Adversarial training for free!," 2019, <https://arxiv.org/abs/1904.12843>.
- [30] Y. Ma, L. Wu, Z. Li, and F. Liu, "A novel face presentation attack detection scheme based on multi-regional convolutional neural networks," *Pattern Recognition Letters*, vol. 131, pp. 261–267, 2020.
- [31] Z. Zhang, C. Jung, and X. Liang, "Adversarial defense by suppressing high-frequency components," 2019, <https://arxiv.org/abs/1908.06566>.
- [32] S. Zhang, A. Liu, J. Wan et al., "Casia-surf: a large-scale multi-modal benchmark for face anti-spoofing," *IEEE Transactions on Biometrics, Behavior, and Identity Science*, vol. 2, no. 2, pp. 182–193, 2020.
- [33] S. M. Moosavi-Dezfooli, A. Fawzi, and P. Frossard, "Deepfool: a simple and accurate method to fool deep neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2574–2582, Las Vegas, NV, USA, July 2016.
- [34] Y. Li, L. Li, L. Wang, T. Zhang, and B. Gong, "Nattack: learning the distributions of adversarial examples for an improved black-box attack on deep neural networks," in *Proceedings of the International Conference on Machine Learning*, pp. 3866–3876, Long Beach, CA, USA, June 2019.