Hindawi
Security and Communication Networks
Volume 2021, Article ID 4280328, 10 pages
https://doi.org/10.1155/2021/4280328

Research Article

A Novel Defensive Strategy for Facial Manipulation Detection Combining Bilateral Filtering and Joint Adversarial Training

Yifan Luo,[1,2] Feng Ye,[1,2] Bin Weng,[1,2] Shan Du,[3] and Tianqiang Huang[1,2]

[1]*College of Mathematics and Informatics, Fujian Normal University, Fuzhou 350117, China*
[2]*Digital Fujian Institute of Big Data Security Technology, Fuzhou 350117, China*
[3]*Department of Computer Science, Mathematics, Physics and Statistics, The University of British Columbia, Vancouver, Okanagan, Canada*

Correspondence should be addressed to Bin Weng; binweng@fjnu.edu.cn

Received 29 April 2021; Revised 24 May 2021; Accepted 27 July 2021; Published 3 August 2021

Academic Editor: Zhili Zhou

Facial manipulation enables facial expressions to be tampered with or facial identities to be replaced in videos. The fake videos are so realistic that they are even difficult for human eyes to distinguish. This poses a great threat to social and public information security. A number of facial manipulation detectors have been proposed to address this threat. However, previous studies have shown that the accuracy of these detectors is sensitive to adversarial examples. The existing defense methods are very limited in terms of applicable scenes and defense effects. This paper proposes a new defense strategy for facial manipulation detectors, which combines a passive defense method, bilateral filtering, and a proactive defense method, joint adversarial training, to mitigate the vulnerability of facial manipulation detectors against adversarial examples. The bilateral filtering method is applied in the preprocessing stage of the model without any modification to denoise the input adversarial examples. The joint adversarial training starts from the training stage of the model, which mixes various adversarial examples and original examples to train the model. The introduction of joint adversarial training can train a model that defends against multiple adversarial attacks. The experimental results show that the proposed defense strategy positively helps facial manipulation detectors counter adversarial examples.

1. Introduction

Facial manipulation refers to swapping the target face with the source face, containing both forms of identity exchange and expression exchange. Identity exchange means swapping the entire face of the target and the source characters, which can change the identity. The expression exchange, on the other hand, only changes the facial expression but does not change the identity. Figure 1 shows two examples of real-world applications of facial manipulation [1, 2]. These two examples in Figure 1 reveal the threat posed by facial manipulation videos to the field of security detection. With the continuous evolvement of facial manipulation, the techniques required for facial manipulation methods are becoming cheaper; the training data sets required are becoming smaller; and the resulting fake videos are becoming more realistic. Timely and effective countermeasures are needed; otherwise, the consequences will be unbearable.

To cope with the threat posed by facial manipulation videos, a number of facial manipulation detectors have been proposed. Facial manipulation detectors can be grouped into two broad categories. One is based on the manual feature extraction [3–5], and the other is on various deep neural networks [6–10]. Compared with traditional feature extraction methods, deep neural network-based methods generally have better detection performance. However, existing deep neural network-based facial manipulation detection models [11–13] are highly vulnerable to adversarial attacks. Obviously, the deep neural network models of facial manipulation detection have security vulnerabilities. Because of the security problem of detection, there are five

(a) (b)

Figure 1: Examples of facial manipulation technology in reality. The first row contains screenshots applying the expression exchange technique. The expression action in (b) is migrated to (a) by Adobe After Effects and the facial manipulation tool, FakeApp. Screenshots of the second row refer to the identity exchange technology. The face in (a) is swapped to the face in (b) using FaceSwap.

adversarial attack methods are used in the paper [13] to attack two kinds of networks [3, 14]. Gandhi and Jain [11] used the classic fast gradient sign method (FGSM) [15] and the Carlini and Wagner Attack (C&W) [16] methods against facial manipulation detectors, ResNet [17] and VGG [18]. Neekhara et al. [12] performed white- and black-box attacks on XceptionNet [19] and MesoNet [7] using a gradient sign-based approach for perturbation optimization. If these models are applied, they will inevitably entail huge risks and irreparable losses. Therefore, it is necessary to improve the security of facial manipulation detectors. However, Gandhi and Jain [11] only applied two defense methods, Lipschitz regularization and deep image prior (DIP), to enhance the security of face swap detectors. And the accuracy that these two methods can improve is very limited. There is no other defense against adversarial attacks. Therefore, the research on defense strategies for facial manipulation detectors needs to be further explored.

To address the above problems, this paper proposes a new defense strategy for facial manipulation detectors. This strategy designs effective methods to defend against adversarial example attacks from passive and proactive defenses. Passive defense means defending against the adversarial attacks without modifying the structure and parameters of the model, while proactive defense needs to train a new model so that the new model can exhibit strong robustness against the attacks. The defense strategy proposed in this paper is a passive defense method based on bilateral filtering and a proactive defense method based on joint adversarial training. The contributions of this work can be summarized as follows:

(1) Three adversarial attack methods, basic iterative method (BIM) [20], projected gradient descent (PGD) [21], and fast FGSM (FFGSM) [22], have been used to perform white- and black-box attacks on the trained Xception model [19].

(2) The bilateral filtering method is introduced as the passive defense method in our proposed defense strategy. It does not require additional model training and only simple processing of input data in the data preprocessing stage. To the best of our knowledge [11], we are the first ones to propose a passive defense method for facial manipulation detectors.

(3) The joint adversarial training method is the proactive defense method in our proposed defense strategy. This method enables facial manipulation detectors to have the ability to defend against multiple adversarial attacks.

The results show the state-of-art performance for the security for the facial manipulation detectors.

2. Related Works

2.1. Facial Manipulation Methods. In terms of identity exchange, classic facial manipulation algorithms include DeepFakes [23], FaceSwap [24], FSGAN [25], FaceShifter [26], and so forth. Among them, DeepFakes [23] was designed based on VAE (variational autoencoder) and GAN (generative adversarial networks). The idea of this algorithm was based on the unsupervised image-to-image transformation proposed by Liu et al. [27]. However, this method needs to train a model with a large number of facial images specifically for both target and source characters, which is very time-consuming. FaceSwap [24] was based on the traditional method of facial region extraction and exchange, which was more lightweight compared to DeepFakes.

FSGAN [25] was a recursive neural network-based face reconstruction algorithm. It can transform the target face based on the source face on the pretrained face model with high efficiency and convenience. FaceShifter [26] can fully extract and adaptively integrate target attributes to generate a realistic face by extracting multilevel target face attributes and adding adaptive attentional denormalization layers. It also introduces a heuristic error acknowledging refinement network that can effectively solve the face occlusion problem.

In terms of expression exchange, Thies et al. [28] used an RGB-D camera to track and reconstruct 3D models of two people's faces and realized facial reconstruction and expression exchange. Face2Face [29] optimized the expression exchange algorithm by combining 3D reconstruction and video reproduction technology. Thies et al. [30] further proposed another neural texture method; this method made the expression exchange more realistic and natural by rerendering the 3D content.

With the improvement of the reconstruction quality of the facial manipulation algorithms, the generated fake videos have reached the level of genuine ones. It is hard to identify accurately the authenticity of these videos with the human eye alone. This poses a threat to the information security of society and the public. If effective measures are not taken, it will certainly bring great harm to social security and stability. Therefore, facial manipulation video detection technology is a hot research content with certain social and practical value.

2.2. Facial Manipulation Detection. Facial manipulation detection algorithms have been developed to address the threat from facial manipulation videos. Some of the researchers detect fake videos by manual feature extraction. Li et al. [31] detected facial manipulation videos based on the biological signal of blink in the videos. Yang et al. [32] suggested that by estimating the 3D head pose in facial images, combined with SVM (support vector machine), classifier can effectively detect fake videos. Amerini et al. [33] used the optical flow method to detect facial manipulation videos. Durall et al. [34] found that facial manipulation videos can be detected by simple frequency domain analysis only.

On the other hand, deep neural networks have also been applied in detecting face-manipulated videos. Li et al. [35] proposed to use the convolutional neural network (CNN) to detect artifacts generated during facial manipulation. Mesonet [7], inspired by InceptionNet, effectively detected fake videos generated by DeepFakes [23] and Face2Face [29]. Nguyen et al. [10] proposed to use the capsule network [36] for facial manipulation video detection, which can effectively detect multiple types of fake videos. In [37], the authors used Xception to detect face-manipulated videos and showed excellent results, so Xception is also one of the basic models used by various new methods for comparison [38–41]. Compared with the methods based on artificial features, the methods based on the deep neural network generally have higher detection accuracy [42–44].

2.3. Adversarial Attacks and Defenses. The adversarial attack [15] is when an attacker generates a corresponding adversarial example [45] by maliciously adding a small perturbation to the original example. Such perturbations are not only undetectable by human eyes but also can lead to misclassification of trained models. Deep neural networks and many other pattern recognition models were found to be vulnerable to adversarial attacks in previous studies [15]. It is also verified that in the field of facial manipulation detection, various methods are also vulnerable to attacks from adversarial examples. Neekhara et al. [12] used the L_∞ distortion metric as the constraint for adding perturbations and optimized it using a gradient sign-based approach. They studied robust attack methods for facial manipulation detection networks from two aspects of white- and black-box attacks. Gandhi and Jain [11] used the classical FGSM and CW attack methods to attack VGG [18] and ResNet [17], which found that the classification accuracy of the two networks was reduced to less than 27%. By modifying the potential space of the generator, Carlini and Farid [13], respectively, carried out white- and black-box attacks against two kinds of networks [3, 14]. Similarly, the accuracy of classification is obviously reduced by this method. Most of the above methods apply classical adversarial attack methods, such as FGSM [15], which are prone to generate visible noise when generating adversarial examples, while methods such as CW [16] are less efficient in attacking. In this paper, we use three improved methods, BIM [20], PGD [21], and FFGSM [22], to attack the facial manipulation detectors. These three methods not only generate less noisy adversarial examples but also are more efficient and thus more difficult to defend.

In order to defend against adversarial attacks effectively, a variety of defense methods have been investigated. Wang et al. [46] defended against adversarial examples by randomly ablating features in the original examples. Papernot et al. [47] used a distillation network for soft label training, which can effectively defend against examples. Goodfellow et al. [15] first proposed that adversarial training could be performed by adding adversarial examples to the training set to enhance the robustness of the models. Bhagoji et al. [48] proposed to use dimension-reduction techniques such as principal component analysis (PCA) for defense. However, in the field of facial manipulation detection, only Gandhi and Jain [11] applied Lipschitz regularization [49] and depth image prior (DIP) [50] to resist the attacks of adversarial examples. The Lipschitz regularization method [49] enhances the robustness of the models against adversarial attacks by constraining the gradient of the detector with respect to the inputs. However, this method is limited by the gradient calculation method and can only be applied to part of the networks (only to ResNet in the paper). Moreover, the accuracy improvement of this method is very limited (only a 10% improvement in detection accuracy), and there is a bottleneck in the application of real scenarios. The DIP method [50] was an unsupervised technique to eliminate interference by preprocessing the data before feeding it to the classifier. However, the optimal accuracy can only be achieved after 6,000 iterations of the model, so it is very

time-consuming. The defense methods mentioned above are only considered from the perspective of proactive defense of the models, which suffered from time-consuming, limited application scenarios and accuracy improvement. In this paper, we propose a new defense strategy. This strategy designs two effective defense methods, the bilateral filtering and the joint adversarial training, from the perspectives of both passive and proactive defenses, respectively.

3. Materials and Methods

3.1. Adversarial Attack. Adversarial attacks against machine learning models can be classified into two types, namely white- and black-box attacks. The classification of the two depends on whether the attacker has access to the prior knowledge of the models. Specifically, adversarial examples are generated by model A to attack model B. If both are the same model, it is a white-box attack; otherwise, it is a black-box attack.

In this paper, we use the following three typical adversarial attack methods on facial manipulation detectors.

The basic iterative method (BIM) [20] is also known as the iterative FGSM (I-FGSM) algorithm. Compared with the classical FGSM, this method uses an iterative approach to find the perturbations of each pixel, rather than making all the pixels change greatly at once. BIM can effectively reduce the disturbance noise.

The projected gradient descent (PGD) [21] is also an iterative implementation of the FGSM algorithm. However, compared with BIM, PGD further increases the number of iterations and adds a layer of randomization, which was initialized with uniform random noise. PGD is very effective against both linear and nonlinear models. And it is one of the most powerful first-order attack methods available.

Wong et al. [22] proposed fast FGSM (FFGSM) attack method, which was used in the fast adversarial training method using the FGSM attack. Compared with the traditional FGSM algorithm, this method combines random initialization. Through simple random initialization operation, FFGSM not only can accelerate the generation of adversarial examples but also can have a strong attack effect.

Figure 2 shows the framework of the adversarial attack process in this paper. We use two types of attacks on the facial manipulation detectors, white- and black-box, respectively. In this case, the target model is Xception [19], and the substitute model in the black-box attack is Meso-Inception [7]. It is important to note that both the white- and the black-box attacks occur in the testing stage of the model. This means that both the substitute and the target models are trained models, and both are trained under the same training set to ensure the transferability of the generated adversarial examples.

In this paper, we propose a new defense strategy that can effectively defend against adversarial attacks for facial manipulation detectors. Specifically, we use bilateral filtering as the passive defense method and the joint adversarial training as the proactive defense method. We will describe the proposed approach in detail in the next section. The overall framework of the defense strategy proposed in this paper is shown in Figure 3.

3.2. Passive Defense. In this paper, we use the bilateral filtering method as the passive defense method. Passive defense occurs in the preprocessing stage of the model on the inputs, and it is very simple and effective to enhance the robustness of the models without retraining.

Bilateral filtering is a method of spatial smoothing. Its main purpose is to deal with image noise reduction. A bilateral filter is a kind of nonlinear filter. This filter is a combination of spatial proximity of images and similarity of pixel values. The bilateral filtering method considers both spatial proximity information and color similarity information. It removes noise and smoothes the image while maintaining edge detail. The formula is as follows:

$$g(i,j) = \frac{\sum_{(k,l)\in S(i,j)} f(k,l) w(i,j,k,l)}{\sum_{(k,l)\in S(i,j)} w(i,j,k,l)}, \tag{1}$$

where (i, j) is the corresponding pixel position, $g(i, j)$ denotes the output image, $f(k, l)$denotes the input image, and $w(i, j, k, l)$ is the value calculated by the two Gaussian functions. The basic idea of the bilateral filter is that the weights calculated by spatial proximity and those calculated by pixel similarity are multiplied, and then the weights are convolved with the image to achieve the effect of keeping the edges to remove noise.

Through comparative experiments, we found that the bilateral filtering method not only can retain the edge information of the image well but also has the best performance in enhancing the robustness of the detector. As shown in Figure 4, compared with the original images, the images processed by median filtering, mean filtering, and Gaussian filtering show the phenomenon of edge blurring. However, the images processed by bilateral filtering can effectively retain the edge information while denoising.

The specific defense process of the bilateral filtering method is shown in the green dotted box in Figure 3. The attacker generates adversarial examples in the test stage of the target model. We add a bilateral filter in the data preprocessing stage so that the adversarial examples can pass the bilateral filter, so as to achieve noise reduction. The noise-reduced examples are then fed into the target model so that the target model can resist the attack of the adversarial examples.

3.3. Proactive Defense. In this paper, we use joint adversarial training as the active defense method. Its principle can be summarized as follows:

$$\min_{\theta} \mathrm{E}_{(Z,y)\sim D}\left[\max_{\delta\le\varepsilon} L\left(f_\theta(X+\delta), y\right)\right], \tag{2}$$

where X is the input of the data, δ denotes the perturbation superimposed on the input, f_θ is the neural network function, and y is the label of example. $L(f_\theta(X+\delta), y)$ is the loss obtained by superimposing a perturbation δ on example X and then comparing it with the label Y through the neural network function.$\max(L)$ denotes the optimization goal, namely to find perturbation to maximize the loss function. The outer layer of formula (2) is the minimization formula

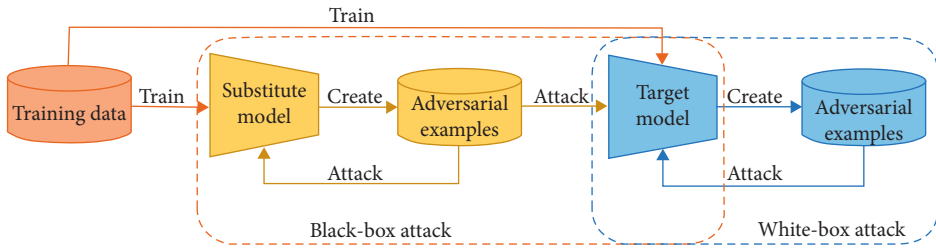

Figure 2: The overall framework of the adversarial attacks. The white-box attack process is shown in the blue dotted box, and the black-box attack process is shown in the yellow dotted box.

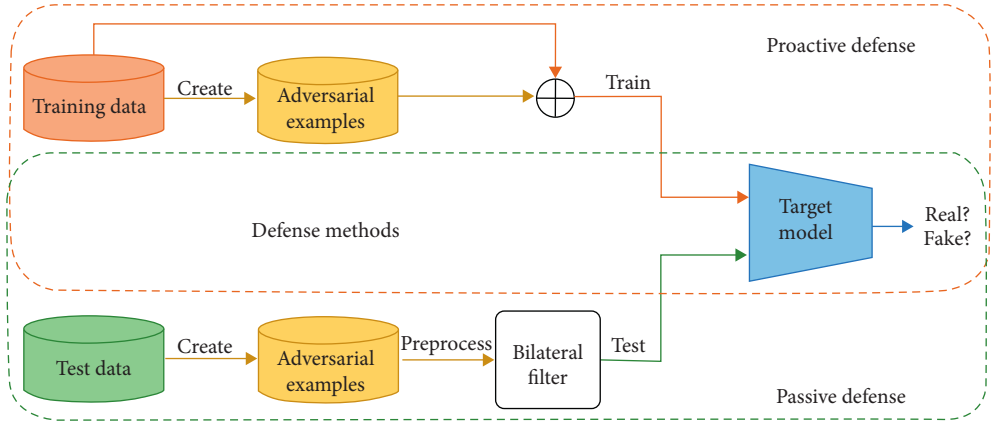

Figure 3: The proposed framework of the defense strategy process. The orange dotted box denotes the proactive defense process based on the joint adversarial training method, and the green dotted box denotes the passive defense process based on the bilateral filtering method.

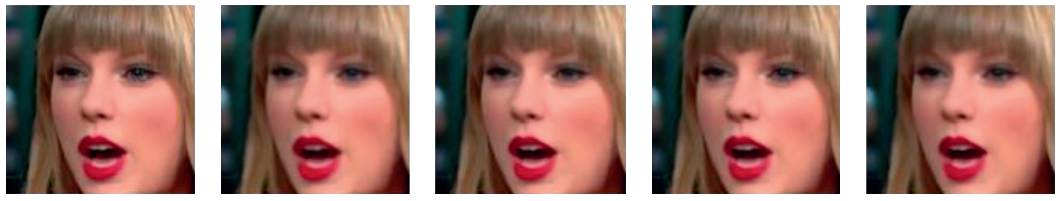

Figure 4: The original images and the images after four kinds of filtering processes: (a) original, (b) mean filter, (c) median filter, (d) Gaussian filter, and (e) bilateral filter.

optimized for the neural network, that is, when the perturbation is fixed, the training neural network model minimizes the loss of training data. In other words, the model has been made robust enough to accommodate such perturbations.

The adversarial training method can only be trained for a particular adversarial attack method. After training, the model can only defend against this kind of adversarial attack method and cannot defend against other kinds of adversarial attack methods. Aiming at the defect of traditional adversarial training, this paper designs a proactive defense method that can defend against multiple adversarial attack methods, that is, joint adversarial training method.

The framework of the joint adversarial training method is shown in the orange dotted box in Figure 3. Specifically, in the training stage, we first carry out the first kind of adversarial attack on the model and mix the generated adversarial examples into the training set. After that, the second adversarial attack is performed on the model, and the generated adversarial examples are mixed into the training set, and until the N-th adversarial attack's adversarial examples are mixed. After mixing the original examples and the N kinds of adversarial examples, the model is then trained. The model trained by this method can effectively defend against multiple adversarial attacks at the same time, and the detection accuracy of the original examples is not affected much.

3.4. Experiment Setting. In the experiment, we use the FaceForensics++ benchmark [3] as the data set for model training and testing. FaceForensics++ contains 1,000 real videos and 5,000 fake videos generated by five facial

manipulation algorithms, with the videos divided into three quality levels. We used three subsets of FaceForensics++. These three subsets are facial manipulation videos generated by three facial manipulation algorithms, namely DeepFakes (DF), Face2Face (F2F), and FaceSwap (FS). Each subset contains 1,000 fake videos. The data set has 1,000 real videos crawled from YouTube. We divided each subset of videos into the training set, testing set, and validation set in the ratio of 7:2:1, respectively, that is, they consist of 700, 200, and 100 real videos and fake videos generated by their corresponding facial manipulation algorithms. Then, we randomly extract some frames from the video. The face detector in the Dlib library is used to extract the images of the face region, which are used as the inputs for model training, validation, and testing. For the models, we use the Xception [4] as the target model for attack and defense and the MesoInception [5] as the substitute model for the Xception.

During the white-box attack experiments, we first pretrained the target model using three subsets of FaceForensics++ separately. Then, in the test stage, BIM [6], PGD [7], and FFGSM [8] are, respectively, used to carry out white-box attacks on the trained target model and generate adversarial examples. In addition, we set the attack intensity of BIM, PGD, and FFGSM as $\varepsilon = 1/255$, $\varepsilon = 1/255$, and $\varepsilon = 1/255$, respectively. During the black-box attack experiments, we first pretrained the target and substitute models using the same three subsets of FaceForensics++. Then three adversarial attack methods were used to carry out white-box attacks on the trained substitute model to generate adversarial examples. Since the black-box transfer attack is less effective than the white-box attack, we appropriately increase the attack intensity when carrying out the white-box attack on the substitute model to ensure the success of the black-box attack. Specifically, we increased the attack intensity of BIM, PGD, and FFGSM to $\varepsilon = 8/255$, $\varepsilon = 8/255$, and $\varepsilon = 8/255$, respectively. Then, the black-box attack is carried out on the target model using the generated adversarial examples. The white-box attack effects of the three attack methods are shown in Figure 5. In this figure, rows represent the type of adversarial attacks to which the images are subjected, and columns represent the type of facial manipulation methods to which the images are subjected, where the first row is the original images that have not been attacked, and the first column is the real images. It can be seen that adversarial examples generated by the three attack methods all have minimal noise. The original examples and the adversarial examples cannot be distinguished by naked eyes only.

For the passive defense experiment, we carried out a comparative experiment of filtering in the test stage of the target model. In other words, the adversarial examples generated by attacking the target model are input into the target model with and without bilateral filtering to test its defensive performance. In the experiment, we set the neighborhood diameter of bilateral filtering as 9, and the standard deviations of spatial Gaussian function and gray similarity Gaussian function are both 75. For the proactive defense experiment, three attack methods, BIM, PGD, and FFGSM, are used in the training process of the target model to carry out joint adversarial training, that is, three adversarial examples are generated simultaneously in the training process. And the generated three adversarial examples are mixed with the original examples as the training set of the target model for joint adversarial training. After the training is completed, in the test stage, three kinds of adversarial attack methods are used to attack the model. The test evaluated the robustness of the target model that defends against adversarial attacks after joint adversarial training.

4. Results and Discussion

We first tested the model accuracy of the trained Xception model under the original examples, and the results are shown in Tables 1 and 2. We can see that the Xception model has great detection accuracy under all three unperturbed subsets of the FaceForensics++ data set. When we apply the three adversarial attack methods to white- and black-box attacks, we can find that the performance of the target model will decline sharply.

Next, we will show the defensive performance of the bilateral filtering method and joint adversarial training method in resisting white- and black-box attacks, respectively, from the perspective of passive and proactive defenses.

4.1. Passive Defense. In the experiment, we first test the performance of four spatial smoothing methods, namely mean filtering, median filtering, Gaussian filtering, and bilateral filtering, in resisting the white-box attack. In Table 1, boldfaced numbers indicate the best precision indexes. From Table 1, the four kinds of spatial filtering can improve the robustness of the target model that defends against white-box attacks. However, the bilateral filtering method shows the optimal effect under all kinds of adversarial attacks. Next, we further experiment with the effect of bilateral filtering defend against black-box attack. The results are shown in Table 2. Similarly, from Table 2, it can be found that the introduced bilateral filtering can well improve the robustness of the target model against black-box attack.

When performing adversarial attacks, attackers usually try to introduce noise perturbation that is difficult to detect with naked eyes, causing the target model to misclassify. From the results of Tables 1 and 2, the proposed passive defense method based on bilateral filtering can effectively perform noise reduction on the input images. The perturbations generated by the attacker are counteracted, thus rendering the adversarial attack ineffective.

As we all know, the detection of existing facial manipulation detectors mostly relies on artifacts in the image, and the bilateral filtering method proposed in this paper is mainly used to denoise the image. Therefore, after the experiment, we conducted statistics on the detection of the samples. We found that the bilateral filtering method proposed in this paper does not increase the number of false-positive and false-negative samples.

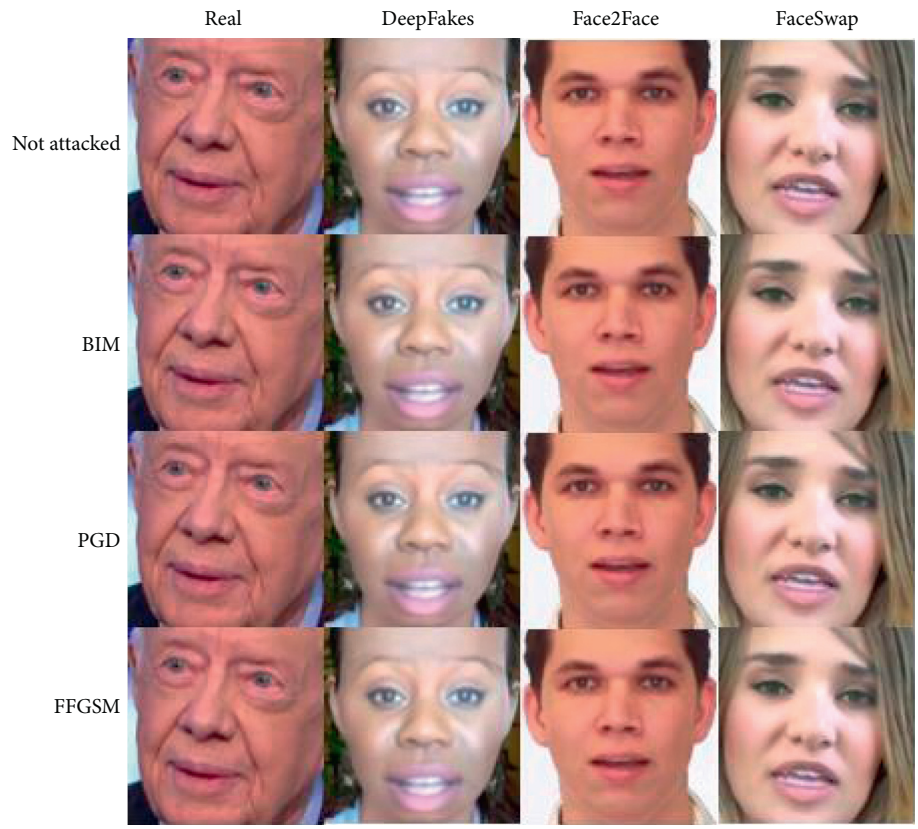

Figure 5: The original examples of the real images and the facial manipulation images and the adversarial examples generated by the three attack methods.

Table 1: Performance comparison of using four filters to defend against white-box attacks of three adversarial attack methods.

Attack	Filter	DF	F2F	FS
None	None	96.10	97.59	98.22
BIM	None	35.77	35.93	32.25
	Mean filter	77.46	86.99	75.79
	Median filter	69.67	82.52	69.78
	Gaussian filter	69.26	81.99	66.41
	Bilateral filter	**83.61**	**88.62**	**77.21**
PGD	None	52.21	51.21	51.67
	Mean filter	83.61	88.62	77.21
	Median filter	74.79	85.58	72.38
	Gaussian filter	74.74	88.18	71.00
	Bilateral filter	**86.62**	**90.69**	**87.98**
FFGSM	None	5.59	7.24	8.49
	Mean filter	60.44	64.55	66.84
	Median filter	36.87	53.37	31.00
	Gaussian filter	35.32	47.31	35.51
	Bilateral filter	**74.80**	**82.47**	**74.26**

4.2. Proactive Defense. One criterion for evaluating the performance of adversarial training is that the model after adversarial training is not only an effective defense against adversarial attacks but also has great classification accuracy for the original examples. At the same time, for our proactive defense experiment, we expect the model to have a good defense against all three adversarial attack methods after joint adversarial training. As shown in Tables 3 and 4, the Xception model after joint adversarial training shows good accuracy against all three white- and black-box attacks while maintaining the classification accuracy of the original examples as much as possible. It can be seen that the use of joint adversarial training as a proactive defense method effectively improves the robustness of the Xception model.

Table 2: Performance of bilateral filtering defending against black-box attacks of three adversarial attack methods.

Attack	Filter	DF	F2F	FS
None	None	96.10	97.59	98.22
BIM	None	52.70	58.34	37.55
	Bilateral filter	86.49	90.44	89.62
PGD	None	62.74	59.58	50.42
	Bilateral filter	87.10	90.85	89.32
FFGSM	None	69.80	59.21	50.28
	Bilateral filter	84.75	89.10	87.69

Finally, we test the performance of the two existing defense methods (the Lipschitz regularization method and the deep image prior method) and the proposed two defense methods against unseen adversarial attacks. To be specific, we used the FGSM attack method to carry out white- and black-box attacks on the Xception model and then used four defense methods against the attack. The experimental results are shown in Table 5. It can be seen from the data in the table that the two defense methods proposed in this paper still

Table 3: Performance of Xception model after joint adversarial training on original examples and defending against white-box attacks.

Attack	DF		F2F		FS	
	Original examples	Adversarial examples	Original examples	Adversarial examples	Original examples	Adversarial examples
BIM	93.31	91.04	94.25	93.17	94.17	92.83
PGD		91.19		92.89		92.65
FFGSM		85.19		86.92		88.15

Table 4: Performance of Xception model after joint adversarial training on defending against black-box attacks.

Attack	DF	F2F	FS
BIM	89.21	91.79	91.72
PGD	90.67	90.42	93.25
FFGSM	90.13	90.04	92.79

Table 5: Performance of Xception model with defense methods on defending against FGSM attack.

Attack	Defense method	DF	F2F	FS
None	None	96.10	97.59	98.22
FGSM_white-box	None	54.35	55.03	51.25
	Lipschitz regularization	56.23	56.91	55.84
	Deep image prior	67.28	68.10	64.91
	Bilateral filter	88.70	92.11	90.56
	Joint adversarial training	74.38	76.09	73.80
FGSM_black-box	None	65.23	61.44	55.71
	Lipschitz regularization	66.80	63.78	61.02
	Deep image prior	75.71	72.19	72.01
	Bilateral filter	86.39	89.82	90.79
	Joint adversarial training	78.66	76.50	76.12

have better defense effects against the unseen adversarial attack compared with the existing defense methods.

5. Conclusions

Various facial manipulation detectors have been introduced to address the security issues associated with facial manipulation. However, most existing models are vulnerable to adversarial example attacks and obviously have security vulnerabilities. In this paper, a new defense strategy for facial manipulation detectors has been proposed to address the vulnerability of detectors defend against adversarial example attacks. Specifically, two defense methods, bilateral filtering and joint adversarial training, are introduced from both passive and proactive defenses. The bilateral filtering method can be used instantly without any modification to the model, which is very convenient and effective. While the joint adversarial training method can effectively defend against multiple adversarial attacks to make the facial manipulation detection model have better robustness. The effectiveness of the two methods is demonstrated through various comparative experiments as well as analyses. The reasons for the defense failure of a small number of adversarial examples are also analyzed qualitatively from the example perspective.

For future work, we will continue to address the reasons for defense failures. And powerful defense methods will be introduced to make facial manipulation detection models more robust.

Data Availability

The data set used to support the findings of this study can be obtained by contacting the authors of [37].

Conflicts of Interest

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

Acknowledgments

This work was supported by the National Key R&D Program Special Fund (grant no. 2018YFC1505805), the National Natural Science Foundation of China (nos. 62072106 and 61070062), and the general project of the Natural Science Foundation in Fujian Province (no. 2020J01168).

References

[1] Buzz Feed Video, "You won't believe what Obama says in this video!," 2018, https://www.youtube.com/watch?v=cQ54GDm1eL0.

[2] Birb Fakes, "Jennifer Lawrence-Buscemi on her favorite housewives [Deepfake]," 2019, https://www.youtube.com/watch?v=r1jng79a5xc.

[3] J. Frank, T. Eisenhofer, L. Schönherr et al., "Leveraging frequency analysis for deep fake image recognition," 2020, https://arxiv.org/abs/2003.08685.

[4] H. Li, B. Li, S. Tan et al., "Detection of deep network generated images using disparities in color components," 2018, https://arxiv.org/abs/1808.07276.

[5] M. Đorđević, M. Milivojević, and A. J. a. Gavrovska, "DeepFake video production and SIFT-based analysis," in *Proceedings of the 2019 27th Telecommunications Forum TELFOR*, vol. 3, p. 6, Belgrade, Serbia, November 2019.

[6] J. H. Bappy, C. Simons, L. Nataraj et al., "Hybrid LSTM and encoder–decoder architecture for detection of image forgeries," *IEEE Transactions on Image Processing*, vol. 28, no. 7, pp. 3286–3300, 2019.

[7] D. Afchar, V. Nozick, J. Yamagishi et al., "MesoNet: a compact facial video forgery detection network," in *Proceedings of the 2018 IEEE International Workshop on Information Forensics and Security*, Hong Kong, China, December 2018.

[8] D. Güera and E. J. Delp, "Deepfake video detection using recurrent neural networks," in *Proceedings of the 2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance*, Auckland, New Zealand, November 2018.

[9] X. Chang, J. Wu, T. Yang, and G. Feng, "DeepFake face image detection based on improved VGG convolutional neural network," in *Proceedings of the 2020 39th Chinese Control Conference*, Shenyang, China, July 2020.

[10] H. H. Nguyen, J. Yamagishi, and I. Echizen, "Use of a capsule network to detect fake images and videos," 2019, https://arxiv.org/abs/1910.12467.

[11] A. Gandhi and S. Jain, "Adversarial perturbations fool Deepfake detectors," 2020, https://arxiv.org/abs/2003.10596.

[12] P. Neekhara, S. Hussain, M. Jere et al., "Adversarial Deepfakes: evaluating vulnerability of Deepfake detectors to adversarial examples," 2020, https://arxiv.org/abs/2002.12749.

[13] N. Carlini and H. Farid, "Evading Deepfake-image detectors with white-and black-box Attacks," 2020, https://arxiv.org/abs/2004.00622.

[14] S.-Y. Wang, O. Wang, R. Zhang et al., "CNN-generated images are surprisingly easy to spot. . . for now," 2019, https://arxiv.org/abs/1912.11035.

[15] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," 2014, https://arxiv.org/abs/1412.6572.

[16] N. Carlini and D. Wagner, "Towards evaluating the robustness of neural networks," in *Proceedings of the 2017 IEEE Symposium on Security and Privacy*, pp. 39–57, San Jose, CA, USA, May 2017.

[17] K. He, X. Zhang, S. Ren, C. Szegedy et al., "Deep residual learning for image recognition," 2015, https://arxiv.org/abs/1512.03385.

[18] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, https://arxiv.org/abs/1409.1556.

[19] F. Chollet, "Xception: deep learning with depthwise separable convolutions," in *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1251–1258, Honolulu, HI, USA, July 2017.

[20] A. Kurakin, I. Goodfellow, and S. Bengio, "Adversarial examples in the physical world," 2016, https://arxiv.org/abs/1607.02533.

[21] A. Madry, A. Makelov, L. Schmidt et al., "Towards deep learning models resistant to adversarial attacks," 2017, https://arxiv.org/abs/1706.06083.

[22] E. Wong, L. Rice, and J. Zico Kolter, "Fast is better than free: revisiting adversarial training," 2020, https://arxiv.org/abs/2001.03994.

[23] Deepfakes, "DeepFakes," 2020, https://github.com/deepfakes/faceswap.

[24] M. Kowalski, "FaceSwap," 2021, https://github.com/MarekKowalski/FaceSwap/.

[25] Y. Nirkin, Y. Keller, and T. Hassner, "FSGAN: subject agnostic face swapping and reenactment," 2019, https://arxiv.org/abs/1908.05932.

[26] L. Li, J. Bao, H. Yang, D. Chen, and F. Wen, "FaceShifter: towards high fidelity and occlusion aware face swapping," 2019, https://arxiv.org/abs/1912.13457.

[27] M.-Y. Liu, T. Breuel, and J. Kautz, "Unsupervised image-to-image translation networks," 2017, https://arxiv.org/abs/1703.00848.

[28] J. Thies, M. Zollhöfer, M. Nießner et al., "Real-time expression transfer for facial reenactment," *ACM Trans. Graph*, vol. 34, 2015.

[29] J. Thies, M. Zollhofer, M. Stamminger et al., "Face2face: real-time face capture and reenactment of RGB videos," in *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2387–2395, Las Vegas, NV, USA, June 2016.

[30] J. Thies, M. Zollhöfer, and M. Nießner, "Deferred neural rendering: image synthesis using neural textures," *ACM Transactions on Graphics*, vol. 38, no. 4, pp. 1–12, 2019.

[31] Y. Li, M.-C. Chang, and S. Lyu, "In Ictu Oculi: exposing AI generated fake face videos by detecting eye blinking," 2018, https://arxiv.org/abs/1806.02877.

[32] X. Yang, Y. Li, and S. Lyu, "Exposing deep fakes using inconsistent head poses," in *Proceedings of the 2019 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 8261–8265, Brighton, UK, May 2019.

[33] I. Amerini, L. Galteri, R. Caldelli et al., "Deepfake video detection through optical flow based CNN," in *Proceedings of the 2019 IEEE International Conference on Computer Vision Workshops*, Seoul, South Korea, October 2019.

[34] R. Durall, M. Keuper, F.-J. Pfreundt et al., "Unmasking DeepFakes with simple features," 2019, https://arxiv.org/abs/1911.00686.

[35] Y. Li and S. Lyu, "Exposing DeepFake videos by detecting face warping artifacts," 2018, https://arxiv.org/abs/1811.00656.

[36] S. Sabour, N. Frosst, and G. Hinton, "Dynamic routing between capsules," 2017, https://arxiv.org/abs/1710.09829.

[37] A. Rössler, D. Cozzolino, L. Verdoliva et al., "FaceForensics++: learning to detect manipulated facial images," 2019, https://arxiv.org/abs/1901.08971.

[38] X. Zhu, H. Wang, H. Fei et al., "Face forgery detection by 3D decomposition," 2020, https://arxiv.org/abs/2011.09737.

[39] X. Li, Y. Lang, Y. Chen et al., "Sharp multiple instance learning for DeepFake video detection," 2020, https://arxiv.org/abs/2008.04585.

[40] Y. Yu, R. Ni, and Y. Zhao, "Mining generalized features for detecting AI-manipulated fake faces," 2020, https://arxiv.org/abs/2010.14129.

[41] I. Ganiyusufoglu, L. M. Ngô, and N. Savov, "Spatio-temporal features for generalized detection of Deepfake videos," 2020, https://arxiv.org/abs/2010.11844.

[42] J. Li, H. Xie, J. Li et al., "Frequency-aware discriminative feature learning supervised by single-center loss for face forgery detection," 2021, https://arxiv.org/abs/2103.09096.

[43] H. Zhao, W. Zhou, D. Chen et al., "Multi-attentional Deepfake detection," 2021, https://arxiv.org/abs/2103.02406.

[44] Y. Nirkin, L. Wolf, Y. Keller et al., "DeepFake detection based on the discrepancy between the face and its context," 2021, https://arxiv.org/abs/2008.12262.

[45] C. Szegedy, W. Zaremba, I. Sutskever et al., "Intriguing properties of neural networks," 2013, https://arxiv.org/abs/1312.6199.

[46] Q. Wang, W. Guo, K. Zhang et al., "Adversary resistant deep neural networks with an application to malware detection," in *Proceedings of the 2017 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1145–1153, Halifax NS Canada, August 2017.

[47] N. Papernot, P. McDaniel, X. Wu et al., "Distillation as a defense to adversarial perturbations against deep neural networks," in *Proceedings of the 2017 IEEE Symposium on Security and Privacy*, pp. 582–597, San Jose, CA, USA, May 2017.

[48] A. N. Bhagoji, D. Cullina, C. Sitawarin et al., "Enhancing robustness of machine learning systems via data transformations," in *Proceedings of the 2018 52nd Annual Conference on Information Sciences and Systems*, pp. 1–5, Princeton, NJ, USA, March 2018.

[49] W. Woods, J. Chen, and C. Teuscher, "Adversarial explanations for understanding image classification decisions and improved neural network robustness," 2019, https://arxiv.org/abs/1906.02896.

[50] D. Ulyanov, A. Vedaldi, and V. Lempitsky, "Deep image prior," in *Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 9446–9454, Salt Lake City, UT, USA, June 2018.