Hindawi
Security and Communication Networks
Volume 2022, Article ID 8609661, 8 pages
https://doi.org/10.1155/2022/8609661

Research Article

Manipulated Faces Detection with Adaptive Filter

Chaoyang Peng, Tanfeng Sun, Zhongjie Mi, and Lihong Yao

School of Electronic Information and Electrical Engineering, Shanghai Jiao Tong University, Shanghai, China

Correspondence should be addressed to Tanfeng Sun; tfsun@sjtu.edu.cn

Received 24 February 2022; Accepted 19 August 2022; Published 15 October 2022

Academic Editor: Jegatha Deborah

With the progress of face manipulation techniques, synthesized faces are spreading on the Internet, which raises concerns about potential threats. To prevent these techniques' abuse, various detection algorithms have been proposed. In this paper, we consider the image's frequency information, then propose an adaptive filtering algorithm named spatial and adaptive filtering (SAF) Network. SAF is a dual-stream network that considers spatial and frequency domains. In the frequency domain, wavelet transform is used to divide the image into different frequency bands, then an adaptive filter is introduced, which aims to capture more decisive information by giving different weights to different frequencies. To fuse spatial and frequency features, spatial pyramid pooling fusion (SPPF) is proposed, which solves the mismatch of feature maps, and considers the relationship between different patches by attention mechanism. Experiment results show that the performance of SAF is better than the comparison algorithm.

1. Introduction

With the rapid development of Deepfakes technology, a large number of manipulated faces have emerged on the Internet. Similiar to text semantics [1], images also have semantic information, so the content of images may be modified. From the forgery results, the tampering methods can be divided into two categories: tampering with some specific character attributes [2, 3] or generating an entire face [4].

In order to detect tampered faces, many forensics algorithms have been proposed. These algorithm can be roughly divided into three categories. First, detection based on biometrics [5, 6]. Second, detection based on spatial domain [7, 8]. Third, detection based on frequency domain [9–11]. Although the existing algorithm has achieved good detection results on public datasets, there are still some problems to be solved. On the one hand, new face manipulation methods are proposed constantly, and the quality of generated faces is higher and higher, which increases the difficulty of detection. Therefore, the detection ability of the previous algorithm may be reduced. On the other hand, the problem of detecting forged faces training and testing in the same dataset is already reasonably solved, so the real challenge is to train on one dataset but test on another with totally different methods.

The current detection algorithms basically focus on deep learning. Most of them use Convolutional neural network (CNN) [12, 13] to detect directly in the spatial domain. They regard deepfake detection as an image classification problem and use CNN to extract features. However, some image post-processing methods will reduce the performance [9, 14, 15] in the RGB domain, such as Gaussian noise, JPEG compression, and median filtering. In the frequency domain, previous work has used filters to preprocess the images. For example, Stuchi et al. [16] designed multiple frequency band filters to operate on the image and manually set the parameters based on experience. However, this method of manually designing filters is inappropriate in some situations because it is difficult for filters with fixed parameters to adaptively capture information of different frequencies.

This paper proposes an adaptive filter to solve the disadvantage of manual filters. Every color component of images is split into different frequencies, and then they are concentrated together to get a multi-channel input, so each channel represents a specific frequency band of a color component. Squeeze-and-Excitation Networks (SENet) [17] can assign weights to different channels to achieve an adaptive filter.

This paper designs a dual-stream network, with one branch used to extract spatial features and the other branch used to extract frequency features adaptively. In extracting the frequency features, we use wavelet transform. According to the properties of the wavelet transform, the image size will be reduced by half after the wavelet transform. So even if the same network is used for both branches, the size of the extracted features in the spatial and frequency domains will be different. If different networks are used, it is more challenging to ensure the consistency of the shape of the feature map.

Spatial pyramid pooling can solve the inconsistency problem of input images and get a fixed size output no matter how large the input image size is. In order to fuse the features extracted from the two branches, we propose spatial pyramid pooling fusion (SPPF). After SPPF, the spatial features and frequency features are fused, and finally, the fused features are passed through the fully connected layer to discriminate the results, real or fake.

2. Related Work

2.1. Manipulated Faces Generation. As for manipulated faces generation, there has been extensive research. Face2Face [18] is known as face reenactment, which modifies the facial expression of the target face. In Face2Face, an actor animates the facial expressions of the target video, and then a manipulated output video is generated. Neural Textures [19] also modifies the facial expression of the source actor. Feature maps are trained as part of the scene capture process, and the training process is end-to-end. StyleGAN [4] can learn high-level attributes automatically, which allows it finely control face properties. ICface [3] proposes a face animator, a data-driven system. It is implemented as a two-stage neural networks, which can mix information from multiple sources. Li et al. [20] introduced a deepfake-based method that solved some problems.

Manipulated faces datasets are the significant benchmark for detecting algorithms. Some popular datasets are listed here. FaceForensics++ (FF++) [21], Celeb-DF [20], Google DFD [22], DFFD [23], Deeperforensics-1.0 (DF-1.0) [24] and DFDC [25]. Examples are shown in Figure 1.

2.2. Manipulated Faces Detection. In order to detect tampered faces, many forensics algorithms have been proposed. The simplest way is to start with biological features and look for defects in visual effects [5, 6, 26]. Li et al. [5] detected manipulated faces by blink. This method combines Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN), which can capture the feature of the human eye blinking in videos. Because eyes in training pictures are generally open, the blinking frequency of the actors in the obtained video will be lower than that of normal people. In addition, people's blinking mainly includes the eye-closing stage and eye-opening stage. These two stages are a gradual process, which is easily ignored in video generation. Li et al. [6] Detected the false face boundary. For the face change performed by deepfake, the edge area of the replacement face will leave traces. Therefore, the authenticity of the face can be judged by detecting the area around the face. Haliassos et al. [26] used high-level semantic irregularities in mouth movement as a feature, which are common in manipulated videos.

The current detection algorithms basically focus on deep learning. Most of them use CNN to detect directly in the spatial domain. Li et al. [7] found a more noticeable difference between the real image and the manipulated image in the YCrCb domain compared with the RGB domain. In this method, the residuals of the YCrCb domain are used as input, and a classifier is trained by CNN. Liu et al. [8] proposed Gram-Net, which used the global texture features. It has strong robustness and generalization ability. CNN is good at classification tasks. For example, Wang et al. [15] directly used the Resnet50 pre-trained on the Imagenet as the backbone, achieving good detection performance. Gowda et al. [27] compared three neural net models and showed that the ensemble method works better.

Some algorithms also use frequency information for detection [9, 11, 16, 28]. Frank et al. [28] comprehensively investigated the characteristics of different GAN structures in the frequency domain. They found that noticeable grid artifacts will be introduced due to the upsampling. Qian et al. [9] extracted two kinds of frequency features, frequency aware decomposition (FAD) and local frequency statistics (LFS), then proposed the F^3-Net. Stuchi et al. [16] designed multiple frequency band filters to operate on the image and manually set the parameters based on experience. However, this method of manually designing filters is inappropriate in some situations because it is difficult for filters with fixed parameters to adaptively capture information of different frequencies.

3. Proposed Method

3.1. Framework. The framework of the proposed algorithm is shown in Figure 2, which fuses spatial and frequency features. The whole process consists of four parts. (1) Preprocessing. The spatial domain image is wavelet transformed, and each color component is decomposed into approximation (LL), horizontal (LH), vertical (HL), and diagonal (HH), a total of 12 feature inputs. After the wavelet transform, the size of the image is halved. That is, if the input image size is $w \times h$, then the size of the wavelet image is $w/2 \times h/2$. (2) Feature extraction. The original spatial domain image is fed into the pre-trained Resnet50 network to extract the spatial features, and the wavelet transformed frequency domain image is fed into the pre-trained SE_Resnet to extract the frequency domain features. (3) Fusion. Since the size of the input frequency domain image is half of the spatial domain image, the size of the feature map obtained after feature extraction should also be halved. In order to make the feature map size consistent, spatial pyramid pooling (SPP) [29] is adapted. During the process of feature fusion, attention mechanism [30] is used here. (4) Classification. The fusion feature is flattened, then binary classification is performed by a fully connected layer. Finally, the detection result will be given.

Figure 1: Examples of manipulated faces (a) Celeb-DF[20] (b) FF++[21] (c) DFD[22] (d) DFFD[23] (e) DF-1.0[24] (f) DFDC[25].

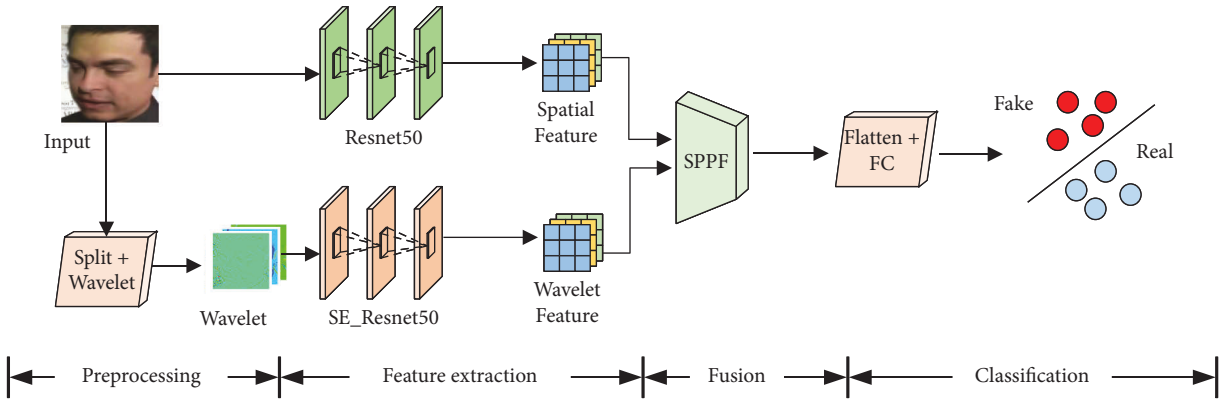

Figure 2: Framework of the proposed SAF. The input is a face to be detected. Spatial and frequency information is extracted by CNN, then they are fused by SPPF. The output is the result (real or fake).

3.2. Frequency Analysis. The natural image consists of three color components: *R*, *G* and *B*. But for the original image format inside the camera, each position has only one component. These colors are arranged in Bayer format [31]. Figure 3 shows the color matrix. Because the human eye is most sensitive to green light, the green component is the sum of the blue and red components. In order to convert the Bayer format to a natural image, CFA interpolation algorithm [31] is adopted. According to the principle of interpolation, the high-frequency information of the three components is similar. Given an image, wavelet transform is carried out. Figure 4 shows the scatter diagram of wavelet detail coefficients in HH. The three coordinate axes represent *R*, *G*, and *B*, respectively. The scatter diagram is distributed in a straight line, and the vector direction is (1, 1, 1), which means that the high-frequency components are approximately equal.

For an image, each color component can be decomposed into high-frequency and low-frequency information (1)

$$C = C^l + C^h. \tag{1}$$

Because of the similarity of high-frequency components, the difference channel C1-C2 can be represented by Equation (2). Therefore, for real images, the high-frequency component is filtered out. However, for manipulated faces, an interpolation algorithm is not adopted so that some high-frequency information will be left.

$$C_1 - C_2 = C_1^l + C_1^h - C_2^l - C_2^h \approx C_1^l - C_2^l. \tag{2}$$

3.3. Adaptive Filter. The image has low and high-frequency contents. Although manipulated faces already have sound visual effects, the details are still lacking, so the difference between real and manipulated faces is more evident in high-frequency. Based on this premise, we studied the imaging process of the camera and found that the high-frequency

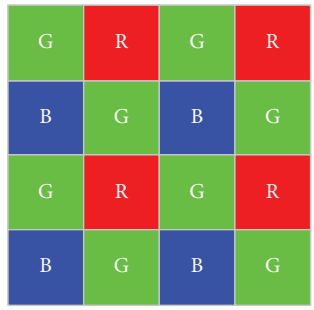

Figure 3: Bayer color filter array pattern.

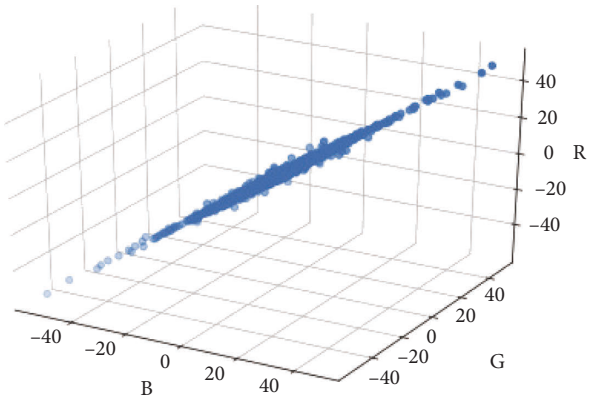

Figure 4: Scatter plots of detail wavelet coefficients.

components of different channels in natural images have a strong correlation, while this property is relatively weak in forged images. Therefore, the difference in high-frequency information is the key to our algorithm. Although the low-frequency information of the natural and forged images are relatively similar, it contains the central semantics of the images. For some manipulated faces with poor visual effects, they can be distinguished clearly with human eyes. So low-frequency information is also taken into account.

Section 1 shows that most previous works [16] design filters manually. This paper gives an adaptive filter. First, color components are split into *R*, G, and B. Then, high and low-frequency information from three channels is divided. Each color component is transformed into LL, LH, HL, and HH. LL is low-frequency information, while LH, HL, and HH are high-frequency information. To simplify the problem, the Haar wavelet is operated only once, so there will be a 12 channels image, and each channel represents a different frequency and color. SENet [17] considers the relationship between channels, which gives different weights to different channels. Therefore, an adaptive filter is achieved.

3.4. Spatial Pyramid Pooling Fusion. The process of spatial pyramid pooling fusion (SPPF) is shown in Figure 5. After it, the features of the two branches are fused. Even if the output sizes are inconsistent, the method can also realize the fusion. For Figure 5, several explanations are given here. (1) After feature extraction, there will be two feature maps. Here, it is assumed that the dimensions are $M\times M\times$ Ch1 and $N\times N\times$ Ch2. Figure 5 shows that Ch1 is equal to 3 and Ch2 is equal to 2, which is only an example. (2) Spatial pyramid pooling [29] ignores the input size and compresses each channel into a vector whose length is L. For example, the length shown in Figure 5 is 4. (3) Using attention [30] to capture the global information and calculate the relationship between various regions. (4) two branches are mixed by multiplying each other, then stacked.

SPPF has two obvious advantages: (1) It solves the inconsistency of feature maps to realize fusion. (2) Since each element of SPP corresponds to a patch in the original map, the relationship between different patches can be reflected when using the attention mechanism, and the size of patches does not need to be the same.

Features of spatial (IS) and frequency (IF) are extracted by networks. For IS, Resnet50 is adopted here, shown in Equation (3). For IF, channels are divided firstly, then perform wavelet transform on them. Next, SE_Resnet50 is used to extract frequency information, shown in Equation (4) and (5).

$$\mathrm{IS}_{M\times M\times \mathrm{Ch1}} = \mathrm{Resnet50}\left(\mathrm{Input}_{R,G,B}\right), \tag{3}$$

$$\mathrm{IW} = \left[\mathrm{Wavelet}\left(\mathrm{Input}_{R}\right), \mathrm{Wavelet}\left(\mathrm{Input}_{G}\right), \mathrm{Wavelet}\left(\mathrm{Input}_{B}\right)\right], \tag{4}$$

$$\mathrm{IS}_{N\times N\times \mathrm{Ch2}} = \mathrm{SE\_Resnet50}\left(\mathrm{IW}\right). \tag{5}$$

Due to their different shapes, the extracted features need to be fused. When it comes to frequency features, the wavelet changes the size of the input image in half, so the size of the feature map of frequency is also half compared with spatial's. In addition, SPP levels (set to 4 in this paper) determine the times of pooling, and pooling type represents pooling mode (max-pooling is adopted). Here, the attention mechanism is used to capture the correlation between patches. After crossing the information of IS and IF, Fusion1 and Fusion2 have the same columns, so they can be concentrated to get the fusion feature (FF). The specific process is shown in Algorithm 1.

4. Experiment Analysis

4.1. Setup

4.1.1. Dataset. Manipulated image datasets are important benchmarks to evaluate the effect of the detection algorithm. In this paper, Celeb-DF [20], FaceForensics++ (FF++) [21] are selected.

(i) Celeb-DF [20]: The second-generation Deepfakes dataset, containing 590 real and 5639 Deepfakes videos.

(ii) FF++ [21]: FF++ is the most widely studied, which includes 1k real and 4k fake videos generated by four methods (Deepfakes (https://github.com/deepfakes/faceswap), Face2Face [18], FaceSwap

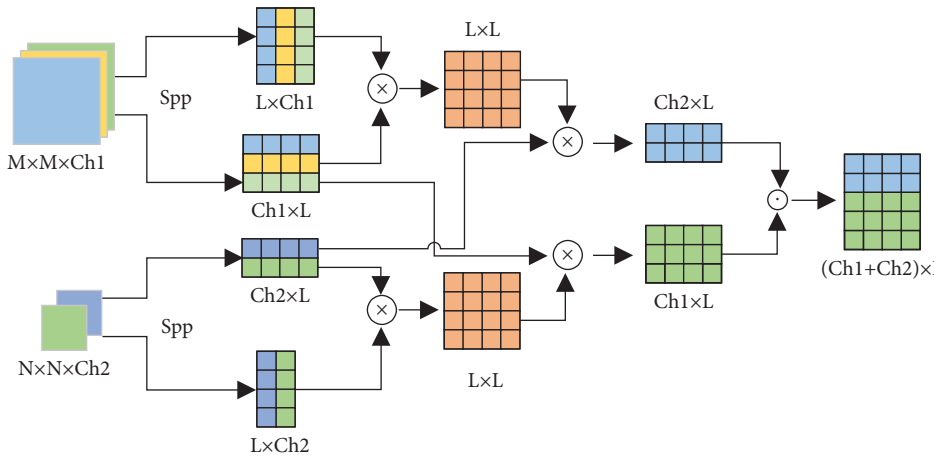

Figure 5: Process of spatial pyramid pool fusion.

(i) **Input:** spatial feature *IS* (M ×*M*× Ch1);
(ii) frequency feature *IF* (N × N × Ch2);
(iii) SPP levels *L*;
(iv) pooling type *T*
(v) **Output:** fusion feature *FF*
(1) *cnt* = *0*;
(2) *S* = *[]*;
(3) *F* = *[]*;
(4) **while** *cnt* < *L* **do**
(5) *Cnt* + = *1*;
(6) *Kernel_S* = *(M/cnt, M/cnt)*;
(7) *Kernel_F*= *(N/cnt, N/cnt)*;
(8) *S* = *[S, Pooling(IS, T, Kernel_S)]*;
(9) *F* = *[F, Pooling(IF, T, Kernel_F)]*;
(10) end
(11) $Attention\_S = S \cdot S^T$;
(12) $Attention\_F = F \cdot F^T$;
(13) $Fusion1 = Attention\_S \cdot F^T$;
(14) $Fusion2 = Attention\_F \cdot S^T$;
(15) FF = [Fusion1, Fusion2]

Algorithm 1: Spatial Pyramid Pooling Fusion.

(https://github.com/MarekKowalski/FaceSwap/) and Neural Textures [19]). All videos in FF++ have three resolutions: raw quality (c0), high-quality (c23), and low quality (c40).

(iii) DFD: Google DFD [22] is the supplement of FF++, with 363 real and 3068 fake videos. They are generated by publicly available methods (https://ai.googleblog.com/2019/09/contributing-data-to-deepfake-detection.html)

4.1.2. Evaluation Metrics. The receiver operating characteristic curve can easily find out the recognition ability of a classifier to samples at a certain threshold. In this paper, The area under the curve (AUC) values are taken as an evaluation metric, which is widely used in manipulated face detection [9, 10, 32].

4.1.3. Implementation Details. Resnet50 is used as the spatial backbone in the experiments to extract features and is loaded with the imagenet pre-training model. For the adaptive frequency filter, to reuse Resnet50, the SE layer is added on top of Resnet50 to get SE_Resnet50, and the imagenet pre-training model is loaded. The learning rate is set as 5e-5 of Adam optimizer. In this paper, the image input size is 224 × 224, so for the spatial domain, the feature map shape is 7 × 7, while the image size after wavelet transform is halved to 112 × 112, so for the frequency branch, the feature map shape is 4 × 4. Haar is selected as the wavelet basis, and the order is 1. The detailed output shape of every layer is listed in Table 1.

4.2. Intra-dataset Emperiments. To prove the effectiveness of SAF, intra-dataset experiments are conducted. FF++ and Celeb-DF are selected as the test database.

4.2.1. Evaluation on FF++. FF++ is the most widely used dataset. Therefore, the proposed method is compared with the previous algorithm. Our experiments are conducted on high-quality (c23) videos, and all four types are used. Each methods provides 10k images, and the ratio of traing set to testing set is 4 : 1. In the experiment, the positive and negative samples are balanced, that is, the ratio of real image to forged image is 1 : 1. Several recent works are compared with our method, including: i.e., (i) Face X-ray [33], which detects manipulated faces across blending boundaries, (ii) F^3-Net [9], which uses frequency features as clues, (iii) Two Branch [34], which proposes a two-branch structure: original and frequency information, (iv) SPSL [10], which combines spatial image and phase spectrum to capture the upsampling artifacts, (v) EfficientNet-B4 (Eff-B4) [35], which is popular in the DeepFake Detection Challenge due to its performance, (vi) Capsule [36], which uses capsule network to detect spoofs, such as replay attacks and deepfakes, (vii)Xception [21], which has good performance in manipulated faces detection and can significantly reduce the number of parameters, (viii) MaDD [32], which captures artifacts by multiple attentional map.

Table 1: Network structure.

Layer	Output size	
Input	224 ×224× 3 (spatial)	112 ×112× 12 (frequence)
Backbone	7 ×7× 2048 (Resnet50)	4 ×4× 2048 (SE_Resnet50)
SPPF	4096 × 30	
Flatten + FC	2	

Table 2: Evaluation on FF++.

Method	AUC score (%)
Face X-ray [37]	87.4
F^3-net [9]	98.1
Two branch [34]	98.7
SPSL [10]	95.3
Eff-B4 [35]	99.2
Capsule [36]	96.6
Xception [21]	**99.7**
MaDD [32]	99.3
Ours	**99.7**

The results are shown in Table 2 and data are cited directly from [10, 32, 38]. The AUC of the proposed method achieves 99.4%, whose performance is better than the comparison algorithm. The AUC of Face X-ray is only 87.4%, and the proposed method is 12% higher than it. Xception [21] performs best in comparison methods, whose AUC is 99.7%. The proposed method can also reach it.

4.2.2. Evaluation on Celeb-DF. Compared with FF++, the forged videos in Celeb-DF have a better visual effect. So we conduct experiments on it. Due to the data imbalance, 60 and 8 are set as the sampling rates for real and manipulated faces respectively, which are set according to SE_EDNet [14]. Table 3 gives the comparison with previous methods. Capsule [36] has been introduced in the last section. I3D [33] is a spatiotemporal network whose convolution and pooling kernels are 3D. Triplet [39] uses a triplet network architecture to detect Deepfakes. SE_EDNet [14] use Euclidean distance to reflect the similarity between vectors, and a new calculation method of attention mechanism is proposed. EfficientNet-B4 (Eff-B4) [35] is popular in the DeepFake Detection Challenge due to its performance. Compared with these methods, the AUC of the proposed algorithm performs better, which achieves 99.9%.

4.3. Cross-Dataset Emperiments. Although the proposed method outperforms the comparison algorithm, we have only made some slight improvements. As seen from section 3.2, the problem of detecting Deepfakes training and testing in the same dataset is already reasonably solved. The real challenge is to train on one dataset and test on another. The detection algorithm does not know the manipulated methods in the actual scene, so it is necessary to evaluate generalization. 16k images are sampled from Celeb-DF [20] (8k for real and 8k for forged), and the DFD [22] is same as it.

Table 3: Evaluation on Celeb-DF.

Method	AUC score (%)
Capsule [36]	93.2
I3D [33]	97.6
Triplet [39]	99.2
SE_EDNet [14]	99.7
Eff-B4 [35]	99.8
Ours	**99.9**

4.3.1. Cross-Dataset Evaluation on Celeb-DF. This section analyses the generalization ability of SAF on unseen data and gives the comparison results. The model is trained on FF++ (all four methods) but evaluated on Celeb-DF. The experimental results are shown in Table 4. Results of previous methods are directly cited from MaDD [32] or original papers. As demonstrated in Table 2, the proposed algorithm performs best in intra-dataset experiments compared to several published methods, whose AUC reaches 99.7%. Although the AUC score of Xception is equal to ours (shown in Table 2), it performs slightly worse than the proposed algorithm when testing on Celeb-DF. That is, the proposed algorithm has stronger transferability.

4.3.2. Cross-Dataset Evaluation on DFD. Besides Celeb-DF, we also conduct experiments on DFD [22]. The results are shown in Table 5, which are cited from [41]. FD^2 Net [41] use facial detail as the clue, which is the combination of light and identity texture. Table 5 indicates that the AUC of the proposed method reaches 84.8%, which outperforms previous algorithms. The previous algorithm with the strongest detection performance is FD^2 Net [41], but its AUC is still 5.7% lower than ours.

4.4. Ablation Study. Four sets of ablation experiments are conducted to analyze the effectiveness of wavelet adaptive filter and SPPF. Experiments results are shown in Table 6, and △ refers to the difference in AUC score between Spatial.

(i) Spatial. Spatial information (original image) is used as input and is sent to Resnet50 directly, which is the baseline.

(ii) Wavelet. Wavelet image is used as input, which is sent to SE_Resnet50.

(iii) Mixing + Cat. mixing wavelet and spatial information by cat, which simply combines the channels of dual-stream outputs.

(iv) Mixing + SPPF. the input is same as Mixing + Cat, but SPPF is introduced to replace cat.

Three conclusions can be drawn from the results in Table 6: (1) proposed wavelet adaptive filter can detect manipulated faces well. When only using wavelet, although AUC (98.4%) is lower than Spatial (99.5%), it outperforms some previous methods in Table 2, such as Capsule [36] and I3D [33]. (2) Mixing spatial and wavelet information is helpful. It performs better than pure wavelet and pure spatial. (3) Proposed SPPF does better in fusing features than combining features directly by Cat.

Table 4: Cross-dataset evaluation on Celeb-DF.

Method	Celeb-DF
Face X-ray [37]	74.2
F^3-net [9]	65.2
Two branch [34]	73.4
SPSL [10]	76.9
Eff-B4 [35]	64.3
Capsule [36]	57.5
Xception [21]	65.3
MaDD [32]	67.4
Ours	**77.1**

Table 5: Cross-dataset evaluation on DFD.

Method	Celeb-DF
Xception [21]	65.6
Eff-B4 ensemble [40]	72.8
FD^2 net [41]	79.1
Ours	**84.8**

Table 6: Intra-dataset Experiments on Celeb-DF (*denotes baseline).

Method	AUC score (%)	△
Spatial*	99.5	—
Wavelet	98.4	−1.1
Mixing + Cat	99.7	0.2
Mixing + SPPF	**99.9**	**0.4**

5. Conclusion

This paper proposes a manipulated faces detection algorithm (SAF), which considers both spatial and frequency information. In the frequency domain, different frequencies are arranged into different channels, and then the channel weighting function of SENet is used for the adaptive filter. In addition, SPPF is proposed to fuse spatial and frequency features, which solves the problem of feature fusion of different shapes. Extensive experiments show the good detection and generalization ability of SAF.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

Ant Group of China supports the collaborative education project of industry university cooperation of the Ministry of Education (the second batch in 2021) (No. 21h010303219).

References

[1] L. Jegatha Deborah, R. Baskaran, and A. Kannan, "Visualizing domain ontology using enhanced anaphora resolution algorithm," *International Journal of Database Management Systems*, vol. 3, no. 3, pp. 110–120, 2011.

[2] H. Zhu, C. Fu, Q. Wu, W. Wu, C. Qian, and R. He, "Aot: appearance optimal transport based identity swapping for forgery detection," *Advances in Neural Information Processing Systems*, vol. 33, pp. 21699–21712, 2020.

[3] S. Tripathy, J. Kannala, and E. Rahtu, "Icface: interpretable and controllable face reenactment using gans," in *Proceedings of the IEEE/CVF winter Conference on Applications of Computer Vision*, pp. 3385–3394, Snowmass, CO, USA, March 2020.

[4] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, "Analyzing and improving the image quality of stylegan," in *Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8107–8116, Seattle, WA, USA, June 2020.

[5] Y. Li, M.-C. Chang, and S. Lyu, "Ictu oculi: exposing ai created fake videos by detecting eye blinking," in *Proceedings of the 2018 IEEE International Workshop on Information Forensics and Security (WIFS)*, pp. 1–7, IEEE, Hong Kong, China, December 2018.

[6] Y. Li and S. Lyu, "Exposing deepfake videos by detecting face warping artifacts," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 46–52, Long Beach, CA, USA, June 2019.

[7] H. Li, B. Li, S. Tan, and J. Huang, "Identification of deep network generated images using disparities in color components," *Signal Processing*, vol. 174, Article ID 107616, pp.1–12, 2020.

[8] Z. Liu, X. Qi, and P. Torr, "Global texture enhancement for fake face detection in the wild," in *Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8057–8066, June 2020.

[9] Y. Qian, G. Yin, S. Lu, Z. Chen, and J. Shao, "Thinking in frequency: face forgery detection by mining frequency-aware clues," in *European Conference on Computer Vision*, pp. 86–103, Springer, Germany, 2020.

[10] H. Liu, X. Li, W. Zhou, and Y. Chen, "Spatial-phase shallow learning: rethinking face forgery detection in frequency domain," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 772–781, Nashville, TN, USA, June 2021.

[11] J. Li, H. Xie, J. Li, Z. Wang, and Y. Zhang, "Frequency-aware discriminative feature learning supervised by single-center loss for face forgery detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6458–6467, Nashville, TN, USA, June 2021.

[12] C. Wang and W. Deng, "Representative forgery mining for fake face detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Article ID 14923, Nashville, TN, USA, June 2021.

[13] F. Maher Salman and S. S. Abu-Naser, "Classification of real and fake human faces using deep learning," *International Journal of Applied Engineering Research*, vol. 6, no. 3, 2022.

[14] C. Peng, L. Yao, T. Sun, X. Jiang, and Z. Mi, "Se_ednet: a robust manipulated faces detection algorithm," in *Computer Graphics International Conference*, pp. 80–88, Springer, Germany, 2021.

[15] S.-Yu Wang, O. Wang, R. Zhang, A. Owens, and A. A. Efros, "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, Nashville, TN, USA, June 2020.

[16] J. A. Stuchi, M. A. Angeloni, R. F. Pereira et al., "Improving image classification with frequency domain layers for feature extraction," in *Proceedings of the 2017 IEEE 27th International Workshop on Machine Learning for Signal Processing (MLSP)*, pp. 1–6, IEEE, Tokyo, Japan, September 2017.

[17] J. Hu, Li Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7132–7141, Nashville, TN, USA, June 2018.

[18] J. Thies, M. Zollhöfer, M. Stamminger, C. Theobalt, and M. Nießner, "Face2face: real-time face capture and reenactment of rgb videos," *Communications of the ACM*, vol. 62, no. 1, pp. 96–104, 2018.

[19] 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.

[20] Y. Li, X. Yang, Pu Sun, H. Qi, and S. Lyu, "Celeb-df: a large-scale challenging dataset for deepfake forensics," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3207–3216, Seattle, WA, USA, June 2020.

[21] A. Rossler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Niessner, "Faceforensics++: learning to detect manipulated facial images," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1–11, Seoul, Korea (South), November 2019.

[22] N. Dufour and A. Gully, "Contributing data to deepfake detection research," *Google AI Blog*, vol. 1, no. 3, 2019.

[23] H. Dang, F. Liu, Joel Stehouwer, X. Liu, and A. K. Jain, "On the detection of digital face manipulation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5781–5790, Seattle, WA, USA, June 2020.

[24] L. Jiang, Li Ren, W. Wu, C. Qian, and C. C. Loy, "Deeperforensics-1.0: a large-scale dataset for real-world face forgery detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2889–2898, Seattle, WA, USA, June 2020.

[25] B. Dolhansky, J. Bitton, P. Ben et al., "The Deepfake Detection challenge (Dfdc) Dataset," 2020, https://arxiv.org/abs/2006.07397.

[26] A. Haliassos, K. Vougioukas, S. Petridis, and M. Pantic, "Lips don't lie: a generalisable and robust approach to face forgery detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5039–5049, Nashville, TN, USA, June 2021.

[27] A. Gowda and N. Thillaiarasu, "Investigation of comparison on modified cnn techniques to classify fake face in deepfake videos,"vol. 1, pp. 702–707, in *Proceedings of the 2022 8th International Conference on Advanced Computing and Communication Systems (ICACCS)*, vol. 1, IEEE, Coimbatore, India, March 2022.

[28] J. Frank, T. Eisenhofer, L. Schönherr, and A. Fischer, "Leveraging frequency analysis for deep fake image recognition," in *Proceedings of the 37th International Confer ence on Machine Learning*, pp. 3247–3258, PMLR, Germany, July 2020.

[29] L. Shi, Z. Zhou, and Z. Guo, "Face anti-spoofing using spatial pyramid pooling," in *Proceedings of the 2020 25th International Conference on Pattern Recognition (ICPR)*, pp. 2126–2133, IEEE, Milan, Italy, January2021.

[30] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7794–7803, Salt Lake City, UT, USA, June 2018.

[31] B. Sun, N. Yuan, and Z. Zhao, "A hybrid demosaicking algorithm for area scan industrial camera based on fuzzy edge strength and residual interpolation," *IEEE Transactions on Industrial Informatics*, vol. 16, no. 6, pp. 4038–4048, 2020.

[32] H. Zhao, W. Zhou, D. Chen, T. Wei, W. Zhang, and N. Yu, "Multi-attentional deepfake detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2185–2194, Nashville, TN, USA, March 2021.

[33] J. Carreira and A. Zisserman, "Quo vadis, action recognition? a new model and the kinetics dataset," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6299–6308, Honolulu, HI, USA, July 2017.

[34] I. Masi, A. Killekar, R. M. Mascarenhas, S. P. Gurudatt, and W. AbdAlmageed, "Two-branch recurrent network for isolating deepfakes in videos," in *European Conference on Computer Vision*, vol. 12352, pp. 667–684, Springer, Germany, 2020.

[35] M. Tan and Q. Le, "Efficientnet: rethinking model scaling for convolutional neural networks," in *Proceedings of the International Conference on Machine Learning*, pp. 6105–6114, PMLR, Long Beach, CA, USA, May 2019.

[36] H. H. Nguyen, J. Yamagishi, and I. Echizen, "Capsule-forensics: using capsule networks to detect forged images and videos," in *Proceedings of the ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2307–2311, IEEE, Brighton, UK, May 2019.

[37] L. Li, J. Bao, T. Zhang et al., "Face x-ray for more general face forgery detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5001–5010, Seattle, WA, USA, June 2020.

[38] J. Wang, Z. Wu, J. Chen, and Yu-G. Jiang, "M2tr: multi-modal multi-scale transformers for deepfake detection," 2021, https://arxiv.org/abs/2104.09770.

[39] A. Kumar, A. Bhavsar, and R. Verma, "Detecting deepfakes with metric learning," in *Proceedings of the 2020 8th international workshop on biometrics and forensics (IWBF)*, pp. 1–6, IEEE, April 2020.

[40] N. Bonettini, E. D. Cannas, S. Mandelli, L. Bondi, P. Bestagini, and S. Tubaro, "Video face manipulation detection through ensemble of cnns," in *Proceedings of the 2020 25th International Conference on Pattern Recognition (ICPR)*, pp. 5012–5019, IEEE, Milan, Italy, January 2021.

[41] X. Zhu, H. Wang, H. Fei, Z. Lei, and S. Z. Li, Face forgery detection by 3d decomposition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2929–2939, Nashville, TN, USA, June 2021.