Hindawi
Security and Communication Networks
Volume 2021, Article ID 8661083, 10 pages
https://doi.org/10.1155/2021/8661083

Research Article

Dual-Tree Complex Wavelet Transform-Based Direction Correlation for Face Forgery Detection

Shichao Gao, Ming Xia, and Gaobo Yang

College of Computer Science and Electronic Engineering, Hunan University, Changsha 410082, China

Correspondence should be addressed to Gaobo Yang; yanggaobo@hnu.edu.cn

Received 17 June 2021; Accepted 28 August 2021; Published 29 September 2021

Academic Editor: Beijing Chen

With the rapid development of face synthesis techniques, things are going from bad to worse as high-quality fake face images are unnoticeable by human eyes, which has brought serious public confidence and security problems. Thus, effective detection of face image forgeries is in urgent need. We observe that some subtle artificial artifacts in spatial domain can be easily recognized in transformation domain, and most facial features have an inherent directional correlation, and generative models would ruffle this kind of distribution pattern. Inspired by this, we propose a two-stream dual-tree complex wavelet-based face forgery network (DCWNet) to expose face image forgeries. Specifically, dual-tree complex wavelet transform is exploited to obtain six directional features ($\pm 75^{\circ}$, $\pm 45^{\circ}$, $\pm 15^{\circ}$) of different frequency components from original images, and a direction correlation extraction (DCE) block is presented to capture the direction correlation. Then, the direction pattern-aware clues and the original image are taken as two complementary network inputs. We also explore how specific frequency components work in face forgery detection and propose a new multiscale channel attention mechanism for features fusion. The experimental results prove that the proposed DCWNet outperforms the state-of-the-art methods in open datasets such as FaceForensics++ and achieves high robustness against lossy image compression.

1. Introduction

In recent years, various deep learning technologies such as FaceSwap [1], Deepfake [2], and Face2Face [3] have presented for facial image manipulations which change the attributes of face images. Besides, some generative adversarial network- (GAN-) [4] based works can even create fake faces without target images. As shown in Figure 1, these artificial products seem scarily real that it is difficult to find fake face images from real ones by naked eyes. This brings great threats to public information security. For example, these techniques might be used to produce pornographic videos or scams. Thus, how to distinguish real and fake face images has attracted more attentions in the community of image content security.

Many works have been proposed to use artificial intelligence (AI) to fight with AI, namely, using deep learning methods to differentiate real images from fake ones. Among them, some sophisticated convolutional neural network (CNN) structures [7–10] were proposed or they were combined with hand-crafted features [11–13] to achieve better performance. However, what makes CNNs be much more perceptive than humans? Some researchers tried to provide some explanations to this from frequency domain [14–17]. Nevertheless, the conventional frequency-domain transformation methods, such as FFT [18] and DCT [19], do not keep well the spatial information of the original image. That is, the images with distinct visual contents might have the same spectral amplitudes. Thus, vanilla CNN structures might be inapplicable. In [16], the frequency features extracted by frequency-aware decomposition (FAD) and local frequency statistics (LFS) were combined with sliding window DCT (SWDCT) to preserve the spatial structure of the image to some extent.

(a)

(b)

Figure 1: Real and fake face images. (a) Real face images. (b) From left to right, fake face images generated by Deepfake, FaceSwap, Face2Face, Neural Textures [5], and StyleGAN [6].

Wavelet transform has been widely used in various image applications such as denoising, compression, and texture classification. Compared with fast Fourier transform (FFT) and other transforms, wavelet transform preserves well multiscale image spatial structure, which makes it to be known as textual microscope. This motivates us that wavelet transform might be compatible with CNN for face forgery detection tasks.

The direction-related details such as facial contour, wrinkles, and light-shadow cross lines are intuitive yet effective for face image forensics. Dual-tree complex wavelet transformation (DTCWT) was proposed to overcome the translation sensitivity, which has higher directional selectivity than traditional wavelets [20]. We exploit the DTCWT to reveal the correlation between facial features in different directions. Moreover, wavelet transformation decomposes the original image into multiple scales. Among them, the low-level features provide richer details, whereas the high-level features provide more semantics information. It is well-known that both low-frequency and high-frequency information is useful for image classification tasks [21]. Is it the same for face image forensics? If so, what is the role each component plays in face forgery detection and how can we fuse multiscale features?

In this work, we propose a novel two-stream deep network for face image forgery detection. One stream exploits DTCWT to learn multiscale directional features. In Figure 2, we show the results of the two-stage DTCWT on the original face image. Each stage contains six different directional features. The other stream takes the original image as input which provides low-frequency and pixel-level information for the network. Moreover, to fully exploit different frequency components, we propose a multiscale channel attention (MSCA) mechanism to fuse multiscale frequency-domain features from direction correlation extraction (DCE) block. The main works and contributions are three-fold: (1) DTCWT is combined with CNN for face image forensics. It addresses face forgery detection from a new perspective, in which a novel DCE block is proposed to extract the correlation features. (2) A MSCA mechanism is proposed to improve feature fusion efficiency. (3) We demonstrate that face image forensics is different from image classification, and the influence of various frequency components on face forgery detection is well studied.

The remainder of this paper is organized as follows: Section 2 summarizes the related works. Section 3 presents the proposed DCWNet. Section 4 reports the experimental results, and conclusion is given in Section 5.

2. Related Work

The recent AI-enabled face forgeries can generate fake face images without any noticeable artificial artifacts. CNNs have achieved great success compared with the earlier works which exploit hand-crafted features [22, 23]. Many face forgery detection works have been presented for better accuracy or interpretability.

2.1. Pixel-Level Forgery Detection. The most widely used method is to input the original images into CNN, either in RGB or HSV color space. In [24], Dang et al. proposed a CNN-based approach integrated with an attention mechanism to improve the feature maps. Inspired by image steganalysis, Nataraj et al. proposed to combine pixel cooccurrence matrices with CNN for face forgery detection [13]. The model was trained on the dataset generated by CycleGAN [25] and had an extra test on face images generated by different GAN structures (StarGAN [26]). The experimental results showed that their work has good generalization capability. Afchar et al. proposed to use two existing networks, namely, Meso-4 and Meso-Inception-4, to exploit the mesoscopic properties of the images [27]. They achieved an accuracy of the ACC up to 98.4%. Guo et al. proposed an adaptive manipulation trace extraction network (AMTEN) [14]. It predicts manipulation traces by an adaptive convolution layer, which are also reused to

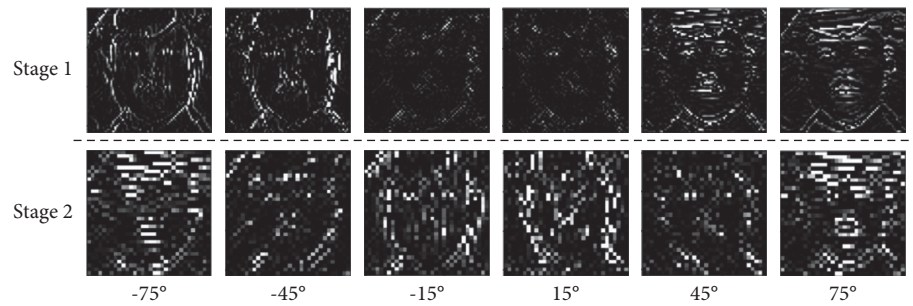

Figure 2: Results of two-stage dual-tree complex wavelet transform. Stage 1 is the result of the first wavelet transform on the original image and stage 2 is the second. Each stage contains six different direction features (±75°, ±45°, ±15°).

maximize manipulation artifacts. For various face forgeries, AMTEN achieved an average accuracy of 98.52%. Nirkin et al. thought that Deepfake methods produce discrepancies between faces and their context. Their approach involved two networks and used the recognition signals from these two networks to detect such discrepancies [28]. In addition, recurrent neural network (RNN) was also exploited by considering face images with temporal properties [29–31]. Some other works exploited visual artifacts such as 3D head poses incoherence for better explanations [32–34]. Chen et al. proposed an improved Xception model for GAN-generated faces [35]. They removed the four residual blocks of Xception to avoid the overfitting problem, and the dilated convolution is used to replace the common convolution layer. The proposed model performed well on their locally GAN-based generated face (LGGF) dataset.

2.2. Frequency-Based Forgery Detection. Image transformation refers to transforming an original image from the spatial domain to other domains such as frequency. The common image transformations include discrete cosine transform [19], fast Fourier transform [18], and wavelet transform [36], which are widely used in various image applications such as edge enhancement, image smoothing, and texture analysis.

In recent years, transform domain processing has been introduced into face forensics. Qian et al. proposed a novel F^3-Net [16], which exploits frequency-aware decomposed image components and local frequency statistics. F^3-Net performs well on the FaceForensics++ dataset, especially for low-quality images. Liu et al. found that the phase spectrum is more sensitive to the up-sample operation than the amplitude spectrum and proposed to expose the up-sample traces by exploiting the phase spectrum [37]. Gong et al. exploited 2D DCT for each RGB channel of the original image and then used AutoGAN [38] to synthesize GAN artifacts in any image without pretrained model [15].

2.3. Attention Mechanism. The attention mechanism generates a set of weighting coefficients, which are often adaptively weighted to strengthen interested regions and suppress irrelevant background regions. There are three common attention mechanisms. The first one is the channel attention. In SENet [39], global average pooling is used to obtain the mean value of the channels as the input of the following fully connected layer. In ECANet [40], 1×1 convolutions replace the fully connected layer to pay more attention to the relationship between adjacent channels. The second one is the spatial attention mechanism which reinforces local areas in each channel. One of the most outstanding works is CBAM [41]. The third one is the self-attention [42], which models the global context through the self-attention mechanism and effectively captures long-distance feature dependencies.

3. Our Approach

3.1. Direction Correlation Extraction Block. Face images have rich directional information such as wrinkles, facial contours, and light and shadow boundaries. They have distribution patterns under specific facial movements. That is, there are spatial correlations among them. The AI-generated fake faces might have weak relevances. This can be used as the clue for face forensics, which motivates us to design a DCE block to expose this, as shown in Figure 3. Conv means convolution operation, BN represents batch normalization, and ReLu is the activation function.

Directional correlation contains two parts: (1) local correlation inside each direction map. (2) Correlation among different direction maps. For local features, we applied 3×3 convolutions on each type of directional feature maps, respectively.

$$f_{n,i} = I_n * \left[C_{1,i}, C_{2,i}, \ldots, C_{m,i}\right], \quad n = \{1, 2, \ldots, m\}, \; i = \{1, 2 \ldots k\}, \tag{1}$$

where I_n are the face feature maps of the nth direction obtained by DTCWT; C_i denotes the convolution kernels; and $f_{n,i}$ represents the features extracted with C_i in direction n. In this work, both m and k are set to 6. For each input, we obtain the feature maps of six channels, which are concatenated to obtain F_{local}.

$$F_{\text{local}} = \text{concat}\left(\left[f_{1,1} \ldots f_{1,k}\right], \ldots \left[f_{m,1} \ldots f_{m,k}\right]\right). \tag{2}$$

The SE block [39] is an existing channel attention method. The input multichannel feature maps are taken into the global average pooling to obtain the weight array. Considering the characteristics of the wavelet coefficients,

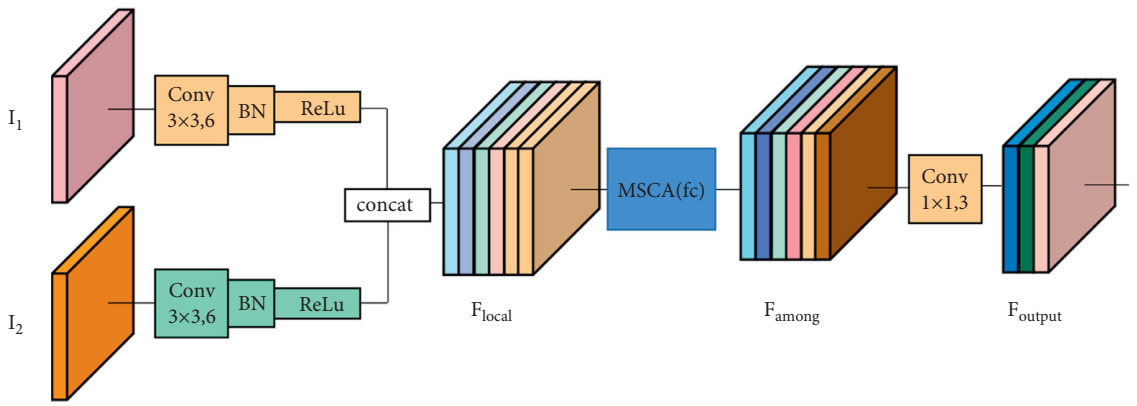

Figure 3: Directional correlation extraction block.

MSCA is adopted to extract features among directional channels (we will demonstrate MSCA in Subsection 3.2.2).

$$F_{\text{ouput}} = c_{1\times1} * \text{MSCA}_{fc}(F_{\text{local}}). \tag{3}$$

Note that the original 1×1 convolution in MSCA is replaced with a fully connected layer (MSCA_{fc}). The reason behind this is that the 1×1 convolution pays more attention to the correlation among adjacent channels. In contrast, the fully connected layer is a point-to-multipoint relationship, which comprehensively describes the relationship between interval channels. Besides extracting the correlation between channels, the MSCA_{fc} block also reduces redundant information in local features. Thus, DCE focuses on directional components. Then, we apply a 1×1 convolution operation $C_1 \times_1$ to further exploit interchannel correlation. In this manner, the same directional features share the convolution kernel in wavelet transform.

3.2. Attention-Based Multiscale Feature Fusion. In essence, multiscale wavelet transform is the stepped dichotomization of the original image frequency. How each frequency component works for face forensics task and how to effectively fuse the directional features obtained from the multiscale wavelet transform? Thus, we proposed a new attention-based feature fusion method.

3.2.1. The Impacts of Frequency Components on Face Forensics. Face forgery detection is different from the traditional image classification tasks. As claimed in [21], the deep network models for image classification exploit both low-frequency and high-frequency information, both contribute to final classification. We conduct a preliminary experiment by selecting 10k face images in which real and fake ratios are half. The fake face images are generated by four face image forgeries. ResNet18 is exploited for experiments. These images are reconstructed by FFT with r as the radius to keep the centre frequency component (Figure 4(a)). The training and testing processes are recorded in Figure 4(b). The horizontal axis is the number of epochs trained, and the vertical axis is the ACC. r is the radius of masking. The larger the r is, the more the high-frequency components are retained. From it, we can observe the following: (1) for low-frequency images, the network converges much quickly, and three epochs are enough. (2) The initial accuracy is continuously improved with the increasing of the high-frequency components. (3) With the introduction of higher frequency components, the network benefits less, and even the accuracy drops.

From the above observations (1) and (2), the network should learn some features from low-frequency components. Note that the frequency components are exploited in parallel, which is different from the conventional image classification [21]. Actually, this is also consistent with our common sense. As we know, image classification is usually of semantic level, whereas face tampering detection is a fine-grained classification task. From the observation (3), since the image often contains some noises that usually exist in the high-frequency components, the accumulation of high-frequency components also brings some difficulties to network learning.

3.2.2. Multiscale Channel Attention. Wavelet transform can provide multiscale image description due to diverse frequency components. Both high-frequency and low-frequency components benefit for face forgery detection. Thus, fusing features is a key issue. The weights of the conventional channel attention mechanisms are based on the mean values of channels, e.g., SENet [39]. Although they work, yet ignore some important local information in the subimportant feature channels. This drawback inhibits wavelet transform from exerting its capability of detail representation. Inspired by the receptive field of human visual cortex neurons, we propose a multiscale channel attention (MSCA) mechanism, which considers the importance of local features and minimizes the side effect of noises. Figure 5 shows the proposed MSCA. C_n denotes different DCE feature maps. They are concentrated as C_a.

$$C_a = \text{concat}(C_1, C_2, \ldots, C_n). \tag{4}$$

We perform maximum pooling with the kernels of 3×3, 5×5, and 7×7 on C_a. For each pooling, we get a 1×1 channel array by global average pooling.

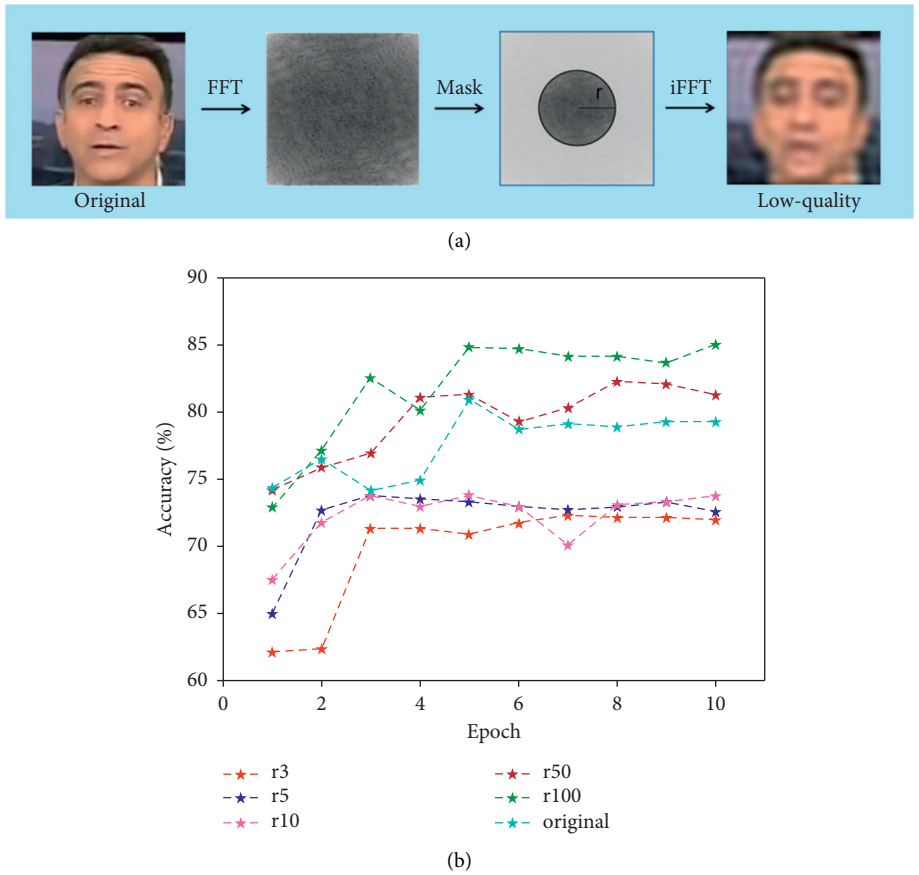

FIGURE 4: Exploring the effectiveness of different frequencies. (a) The original image is transformed by FFT, and we retain and reconstruct frequencies within the circle of radius (r). (b) The accuracy variation when using different frequency components during training.

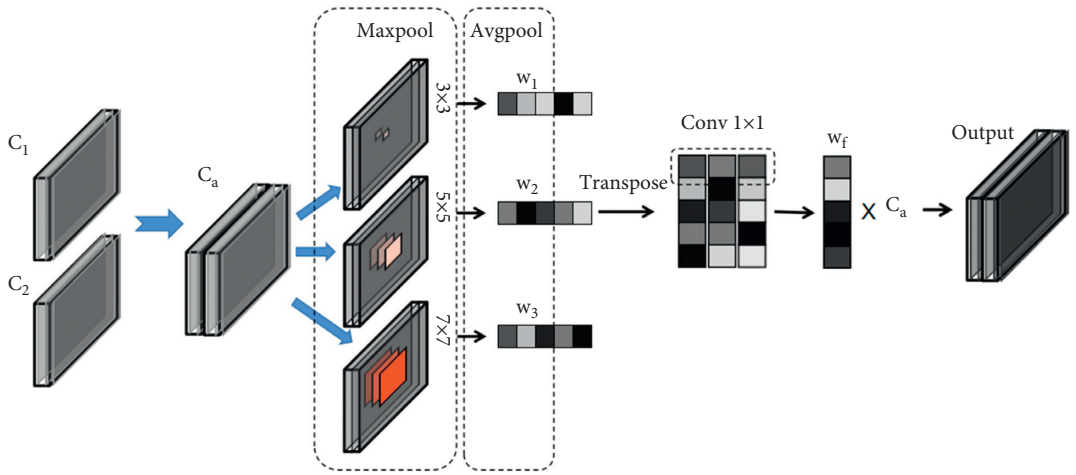

FIGURE 5: Multiscale channel attention (MSCA).

$$w_s = \mathrm{Avg}(C_a * \mathrm{Maxpool}_{s\times s}), \quad s = \{3, 5, 7\}. \tag{5}$$

Next, we transpose and concentrate them to 3×1 channels, then we use a 1×1 convolutional operation ($C_{1\times1}$) to obtain w_f. The final output is obtained by multiplying C_a with w_f.

$$w_f = \mathrm{concat}(w_3, w_5, w_7) * C_{1\times1}, \tag{6}$$

$$\mathrm{output} = w_f \odot C_a. \tag{7}$$

The maximum pooling strategy strengthens local features, while average pooling highlights global information.

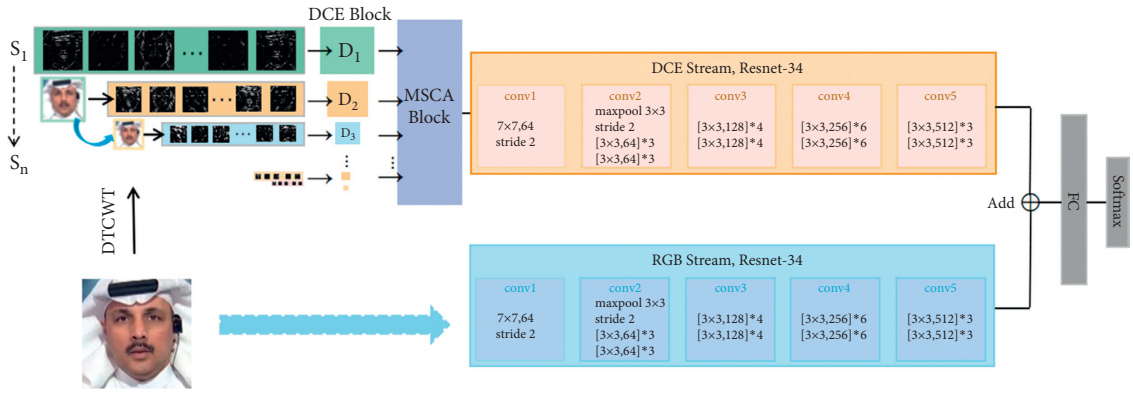

Figure 6: The framework of the proposed DCWNet.

Thus, the assignment of the weights for each channel is comprehensively considered by using MSCA. Please note that the directional features use high-frequency components. The experiment in Subsection 3.2.1 proves that the low-frequency components also play a role in the model training. Thus, we use a two-stream network to exploit the low-frequency information and pixel-level features simultaneously.

Based on the above methods, we proposed our DCWNet, and Figure 6 shows the framework of the complete work.

4. Experimental Results and Analysis

4.1. Experimental Setting. Image Dataset. FaceForensics++ is the most recent face manipulation dataset, which has been widely used in existing works [33, 43]. It is expanded from the FaceForensics dataset with three quality levels, namely, RAW (raw), HQ (high quality), and LQ (low quality). For the FaceForensics dataset, each level includes 1,000 videos, which are directly collected from YouTube without tampering. The same amounts of fake videos are generated by four face forgeries including Deepfake, Face2Face, FaceSwap, and Neural Textures. In addition, the FaceForensics++ dataset also contains 363 real videos from 28 actors under 16 scenes. Thus, the FaceForensics++ dataset has 1,363 real videos and 4,000 fake videos for each quality. We extract 60 frames for each real video at equal interval and 16 frames for each fake video. The MTCNN [44] is used to crop the face images. Thus, we have 63k fake face images and 63k real face images, totally 126k face images. We divide them into 85k, 35k, and 6k face images as the training set, the testing set, and the validation set, respectively. In addition, the DFDC preview [45] dataset, which is a preview dataset of the Deepfake Detection Challenge, is also used for experiments. It contains 1131 real videos and 4119 fake videos. We obtain 120k face images from the DFDC preview dataset.

Evaluation Metrics. To evaluate the effectiveness of our model, we exploit two widely used metrics, namely, classification accuracy (ACC) and area under receiver operating characteristic curve (AUC). The closer the ACC is to 100% and the AUC is to 1, the better the performance the network achieves.

Experiment Details. The ResNet34, which was pretrained on ImageNet [46], is exploited as the backbone for two streams. The Kaiming Batch Normalization is used for initialization. The networks are optimized via SGD with 0.9 as the momentum and 0.0005 as the weight decay. We set the base learning rate as 0.02 and use StepLR as the learning rate scheduler with half the learning rate per step. The batch size is 64 and we train the model for about 14k iterations. The whole work is completed upon PyTorch 1.1.0 with two Nvidia GeForce GTX 1080 Ti GPUs. To speed up the training process, we save the results of wavelet transform into local disk in NumPy format.

4.2. Comparisons with the Existing Works. The proposed DCWNet is tested on different quality image datasets that consist of fake images produced by different image tampering methods. Experimental comparisons are made among the proposed approach and the existing works. For the FaceForensics++ dataset, the experimental results are shown in Table 1. Apparently, the proposed DCWNet achieves a pretty high ACC (98.73%) and AUC (0.999) on the FaceForensics++ (HQ) dataset.

For the LQ dataset, DCWNet also achieves desirable results with the ACC of 97.91% and the AUC of 0.994. Compared to the baseline networks (ResNet34), DCWNet achieves the improvement of ACC about 2.05%. This proves that the DCE block is effective. Figure 7 reports the ROC curves for different face forgery detection methods. We also conduct the experiments on the DFDC preview dataset with the same experimental setting. Table 2 reports the experimental results.

For different face manipulations, we also test our model. Specifically, there are four face manipulations for the fake images in the FaceForensics++ dataset. Each face manipulation has 31k images. Among them, 22k, 8k, and 1k are used for training, testing, and validation, respectively. Similar experimental results are obtained, which are reported in Table 3.

Table 1: Results on the FaceForensics++ dataset with LQ and HQ.

Methods	ACC (LQ) (%)	AUC (LQ)	ACC (HQ) (%)	AUC (HQ)
Meso-4 [27]	54.38	0.542	60.63	0.660
Meso-Incep [27]	58.30	0.694	64.49	0.734
HP-CNN [11]	62.59	0.683	64.09	0.712
Constrained Conv [47]	80.05	0.883	83.40	0.920
AMTEN [14]	83.76	0.868	85.69	0.917
XceptionNet [9]	88.04	0.974	92.29	0.985
ResNet34 [8]	93.93	0.753	96.68	0.803
DCWNet(ResNet34)	97.91	0.994	98.73	0.999

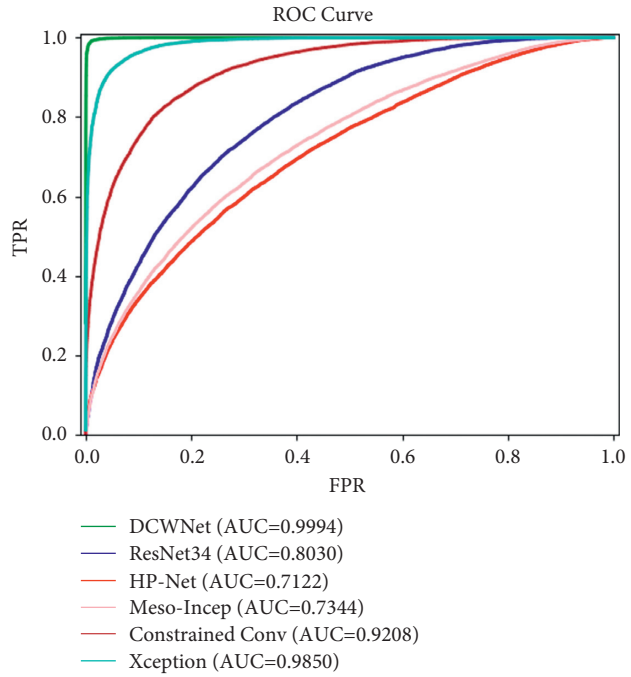

Figure 7: ROC curve for different face forgery detection methods.

Table 2: The experimental results on the DFDC preview dataset.

Methods	ACC (%)	AUC
Meso-4 [27]	53.71	0.553
Meso-Incep [27]	58.16	0.654
HP-CNN [11]	61.49	0.675
Constrained Conv [47]	81.01	0.877
AMTEN [14]	88.83	0.892
XceptionNet [9]	89.37	0.969
ResNet34 [8]	94.52	0.736
DCWNet(ResNet34)	97.31	0.920

4.3. Ablation Study

4.3.1. The DCE Block. To prove the contribution of the proposed DCWNet, ablation study is conducted. We first explore the influence of the number of directions, and the experimental results are recorded in Table 4. Even with features from one direction, the DCE stream achieves high ACC and AUC. This proves that the DCE block is powerful for local feature representation. With more features from multiple directions, the detection accuracies improve greatly. This implies that the features extracted from different directions are complimentary to each other. We also compare the effect of the FC layer and 1×1 convolution used in MSCA. We observe that with the using of more directions, FC is better than 1×1 convolution.

Figure 8 shows some feature maps extracted from the DCE block. We can notice that the attention responses of the fake images are distracted, whereas those of the real images are compact. The reason behind this is that the directional features are not strongly correlated in fake face images, while they are more uniform for real face images.

Table 3: Detection results for different face manipulations.

Methods	Deepfake (%)	Face2Face (%)	FaceSwap (%)	Neural Textures (%)
Meso-4 [27]	53.31	61.80	62.08	50.33
Meso-Incep [27]	76.01	71.12	71.69	50.30
HP-CNN [11]	86.03	81.48	89.30	77.07
Constrained Conv [47]	82.39	81.63	88.57	79.15
AMTEN [14]	86.56	84.76	80.12	76.07
XceptionNet [9]	97.51	97.24	97.11	79.41
ResNet34 [8]	98.32	98.35	97.90	95.90
DCWNet(ResNet34)	99.54	99.55	98.84	96.24

Table 4: Ablation study of the DCE block for different number of directions.

Direction	Conv 1 × 1		FC	
	ACC (%)	AUC	ACC (%)	AUC
(+15°)	89.24	0.908	90.03	0.898
(+15°, +45°)	91.19	0.932	90.42	0.906
(+15°, +45°, +75°)	92.88	0.929	92.74	0.938
(+15°, +45°, +75°, −15°)	92.87	0.923	93.28	0.942
(+15°, +45°, +75°, −15°, +45°)	93.40	0.946	94.38	0.948
(+15°, +45°, +75°, −15°, −45°, +75°)	93.77	0.956	95.34	0.962

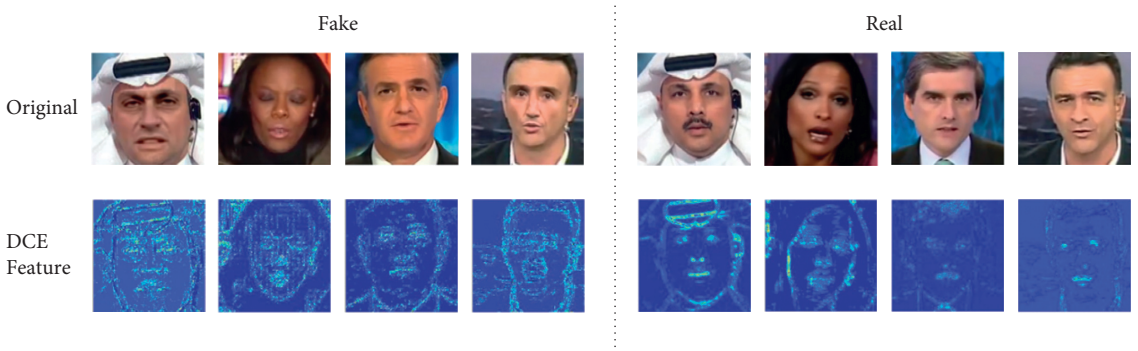

Figure 8: Feature maps from DCE block.

Table 5: Comparisons among three feature fusion methods.

Components	S1 (%)	S2 (%)	Addition (S1, S2) (%)	SE (S1, S2) (%)	MSCA (S1, S2) (%)
ACC	95.34	94.98	95.28	95.46	96.81

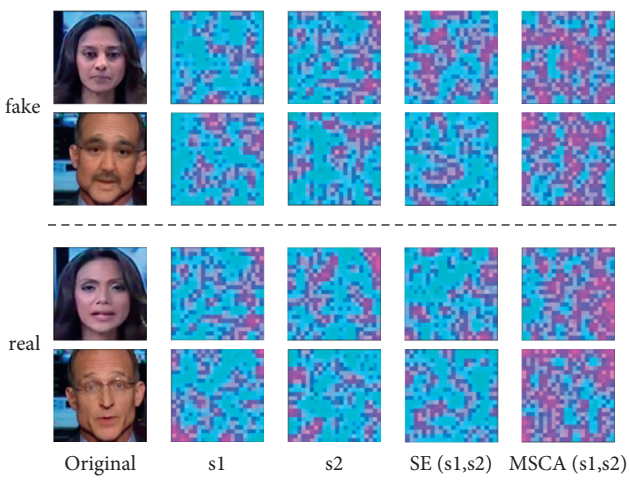

Figure 9: The feature maps from different fusion methods.

4.3.2. MSCA. To prove the effectiveness of MSCA, we use different feature fusion methods for the DCE feature maps. The experimental results are reported in Table 5. Specifically, we conduct experiments for the first (S1) and second (S2) stages of the wavelet transform, respectively. The element-wise addition, self-attention (SE), and MSCA are used for feature fusion. From Table 5, the MSCA achieves the best feature fusion. Figure 9 also compares the feature maps from the DCE stream between SE and MSCA.

5. Conclusion

In this work, we propose a two-stream DCWNet for face forgery detection. One stream uses the DCE block to exploit the multiscale directional correlation. To fuse the DCE feature maps of different scales, MSCA is proposed. The other stream uses the original image as input. The experimental results showed that DCWNet achieves desirable

results on the FaceForensics++ and DFDC preview datasets. From the ablation study, we observe that real and fake faces have different feature maps that learned from the DCE block. This proves that the correlation of direction distribution is valuable for face forgery detection. Moreover, the effectiveness of the proposed MSCA is verified by comparisons with existing feature fusion methods. We also explore how different frequency components contribute to face forgery detection, which provides some interpretability for face forensics.

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.

References

[1] FaceSwap: https://github.com/MarekKowalski/FaceSwap.

[2] Deepfakes github: https://github.com/deepfakes/faceswap.

[3] J. Thies, M. Zollhofer, M. Stamminger, C. Theobalt, and M. Nießner, "Face2face: real-time face capture and reenactment of rgb videos," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2387–2395, San Francisco, CA, USA, August 2016.

[4] I. Goodfellow, J. Pouget-Abadie, M. Mirza et al., "Generative adversarial nets," in *Proceedings of the 28th International Conference on Neural Information Processing Systems (NIPS)*, Montreal, Canada, June 2014.

[5] 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.

[6] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4401–4410, Los Angeles, USA, June 2019.

[7] A. G. Howard, M. Zhu, B. Chen et al., "Mobilenets: efficient convolutional neural networks for mobile vision applications," Available: https://arxiv.org/abs/1704.04861, 2017.

[8] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, San Francisco, CA, USA, August 2016.

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

[10] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4700–4708, Honolulu, HI, USA, July, 2017.

[11] H. Mo, B. Chen, and W. Luo, "Fake faces identification via convolutional neural network," in *Proceedings of the 6th ACM Workshop on Information Hiding and Multimedia Security (IH&MMSec)*, New York, USA, June 2018.

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

[13] L. Nataraj, T. M. Mohammed, and B. S. Manjunath, "Detecting GAN generated fake images using co-occurrence matrices," *Journal of Electronic Imaging*, vol. 5, pp. 1–7, 2019.

[14] Z. Guo, G. Yang, J. Chen, and X. Sun, "Fake face detection via adaptive manipulation traces extraction network," *Computer Vision and Image Understanding*, vol. 204, Article ID 103170, 2021.

[15] X. Zhang, S. Karaman, and S.-F. Chang, "Detecting and simulating artifacts in gan fake images," in *Proceedings of the 11th IEEE International Workshop on Information Forensics and Security (WIFS)*, Delft, The Netherlands, December 2019.

[16] Y. Qian, G. Yin, L. Sheng, Z. Chen, and J. Shao, "Thinking in frequency: face forgery detection by mining frequency-aware clues," in *Proceedings of the European Conference on Computer Vision*, pp. 86–103, Glasgow, UK, August 2020.

[17] Z. Qin, P. Zhang, F. Wu, and X. Li, "Fcanet: frequency channel attention networks," Available: https://arxiv.org/abs/2012.11879, 2020.

[18] D. F. Elliott and K. R. Rao, *Fast Fourier Transform and Convolution Algorithms*, Springer-Verlag, New York, 1982.

[19] N. Ahmed, T. Natarajan, and K. R. Rao, "Discrete cosine transform," *IEEE Signal Processing Magazine*, vol. 100, no. 1, pp. 90–93, 1974.

[20] I. W. Selesnick, R. G. Baraniuk, and N. C. Kingsbury, "The dual-tree complex wavelet transform," *IEEE Transactions on Computers*, vol. 22, no. 6, pp. 123–151, 2005.

[21] H. Wang, X. Wu, and Z. Huang, "High-frequency component helps explain the generalization of convolutional neural networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8684–8694, Seattle, WA, USA, June 2020.

[22] P. Ferrara, T. Bianchi, A. De Rosa, and A. Piva, "Image forgery localization via fine-grained analysis of CFA artifacts," *IEEE Transactions on Information Forensics and Security*, vol. 7, no. 5, pp. 1566–1577, 2012.

[23] X. Pan, X. Zhang, and S. Lyu, "Exposing image splicing with inconsistent local noise variances," in *Proceedings of the IEEE International Conference on Computational Photography (ICCP)*, pp. 1–10, Seattle, WA, USA, April 2012.

[24] H. Dang, F. Liu, and J. Stehouwer, "On the detection of digit manipulation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Seattle, WA, USA, June 2020.

[25] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Venice, Italy, October 2017.

[26] Y. Choi, M. Choi, M. Kim, J.-W. Ha, S. Kim, and J. Choo, "StarGAN: unified generative adversarial networks for multi-domain image-to-image translation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Salt Lake City, USA, June 2018.

[27] D. Afchar, V. Nozick, J. Yamagishi, and I. Echizen, "Mesonet: a compact facial video forgery detection network," in *Proceedings of the 2018 IEEE International Workshop on Information Forensics and Security (WIFS)*, pp. 1–7, IEEE, Hong Kong, China, December 2018.

[28] Y. Nirkin, L. Wolf, Y. Keller, and T. Hassner, "DeepFake detection based on discrepancies between faces and their context," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 1, no. 1, p. 99, 2020.

[29] P. Korshunov and S. Marcel, "Deepfakes: a new threat to face recognition? assessment and detection," 2018, https://arxiv.org/abs/1812.08685.

[30] 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 (AVSS)*, pp. 1–6, IEEE, Auckland, New Zealand, November 2018.

[31] E. Sabir, J. Cheng, A. Jaiswal, W. AbdAlmageed, I. Masi, and P. Natarajan, "Recurrent convolutional strategies for face manipulation detection in videos," *Interfaces*, vol. 3, no. 1, pp. 80–87, 2019.

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

[33] F. Matern, C. Riess, and M. Stamminger, "Exploiting visual artifacts to expose deepfakes and face manipulations," in *Proceedings of the 2009 IEEE Winter Applications of Computer Vision Workshops (WACVW)*, pp. 83–92, IEEE, Snowbird, UT, USA, December 2009.

[34] N. Yu, L. Davis, and M. Fritz, "Attributing fake images to gans: analyzing fingerprints in generated images," 2018, https://arxiv.org/abs/1811.08180.

[35] B. Chen, X. Ju, B. Xiao, W. Ding, Y. Zheng, and V. H. C. De Albuquerque, "Locally GAN-generated face detection based on an improved Xception," *Information Sciences*, vol. 572, pp. 16–28, 2021.

[36] M. Antonini, M. Barlaud, P. Mathieu, and I. Daubechies, "Image coding using wavelet transform," *IEEE Transactions on Image Processing*, vol. 1, no. 2, pp. 205–220, 1992.

[37] H. Liu, X. Li, W. Zhou et al., "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, Montreal, QC, Canada, October 2021.

[38] X. Gong, S. Chang, Y. Jiang, and Z. Wang, "Autogan: neural architecture search for generative adversarial networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3224–3234, Los Angeles, USA, June 2019.

[39] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7132–7141, Salt Lake City, UT, USA, June 2018.

[40] Q. Wang, B. Wu, P. Zhu, L. Peihua, Z. Wangmeng, and H. Qinghua, "ECA-Net: efficient channel attention for deep convolutional neural networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Seattle, WA, USA, June 2020.

[41] S. Woo, J. Park, J. Y. Lee, and S. I. Kweon, "Cbam: convolutional block attention module," in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 3–19, Munich, Germany, September 2018.

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

[43] H. H. Nguyen, F. Fang, J. Yamagishi, and I. Echizen, "Multi-task learning for detecting and segmenting manipulated facial images and videos,," in *Proceedings of the IEEE 10th International Conference on Biometrics Theory, Applications and Systems*, pp. 1–8, Tampa, USA, 2019.

[44] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Processing Letters*, vol. 23, no. 10, pp. 1499–1503, 2016.

[45] B. Dolhansky, R. Howes, B. Pflaum, N. Baram, and C. C. Ferrer, "The deepfake detection challenge (DFDC) preview dataset," 2019, https://arxiv.org/abs/1910.08854.

[46] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and F.-F. Li, "Imagenet: a large-scale hierarchical image database," in *Proceedings of the I2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, Miami, FL, USA, June 2009.

[47] B. Bayar and M. C. Stamm, "A deep learning approach to universal image manipulation detection using a new convolutional layer," in *Proceedings of the 4th ACM Workshop on Information Hiding and Multimedia Security*, pp. 5–10, Web Tokyo, Japan, 2016.

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