mathematics

MDPI

Article

Frequency Domain Filtered Residual Network for Deepfake Detection

Bo Wang [1], Xiaohan Wu [1], Yeling Tang [1], Yanyan Ma [1], Zihao Shan [2] and Fei Wei [3,*]

[1] School of Information and Communication Engineering, Dalian University of Technology, Dalian 116024, China
[2] Independent Researcher, 1 Hacker Way, Menlo Park, CA 94560, USA
[3] School of Computing, National University of Singapore, 13 Computing Drive, Singapore 117417, Singapore
* Correspondence: feiwei@nus.edu.sg

Abstract: As deepfake becomes more sophisticated, the demand for fake facial image detection is increasing. Although great progress has been made in deepfake detection, the performance of most existing deepfake detection methods degrade significantly when these methods are applied to detect low-quality images for the disappearance of key clues during the compression process. In this work, we mine frequency domain and RGB domain information to specifically improve the detection of low-quality compressed deepfake images. Our method consists of two modules: (1) a preprocessing module and (2) a classification module. In the preprocessing module, we utilize the Haar wavelet transform and residual calculation to obtain the mid-high frequency joint information and fuse the frequency map with the RGB input. In the classification module, the image obtained by concatenation is fed to the convolutional neural network for classification. Because of the combination of RGB and frequency domain, the robustness of the model has been greatly improved. Our extensive experimental results demonstrate that our approach can not only achieve excellent performance when detecting low-quality compressed deepfake images, but also maintain great performance with high-quality images.

Keywords: deepfake detection; neural networks; wavelet transform; frequency domain features; feature fusion

MSC: 68T09

check for updates

Citation: Wang, B.; Wu, X.; Tang, Y.; Ma, Y.; Shan, Z.; Wei, F. Frequency Domain Filtered Residual Network for Deepfake Detection. *Mathematics* **2023**, *11*, 816. https://doi.org/10.3390/math11040816

Academic Editors: Sheng Li, Zhenjun Tang and Guorui Feng

Received: 1 January 2023
Revised: 27 January 2023
Accepted: 2 February 2023
Published: 6 February 2023

1. Introduction

With the development of deep learning, great strides have been made in image processing techniques. One of these techniques is called deepfake, which is used to create incredibly realistic fake images by replacing the face of a source image with a target face. In particular, the birth of Generative Adversarial Network (GAN) [1] makes fake images more difficult to distinguish. At present, deepfake has positive applications in many fields, such as education, medical care and entertainment. However, there are also a lot of malicious applications of deepfake. Deepfake porn is one of such malicious applications. Rana Ayyub, a female journalist in India, was haunted by a sex video faked with her face. During the Russian–Ukrainian conflict, a video of Ukrainian President Volodymyr Zelensky was widely circulated, in which Zelensky called on Ukrainian soldiers to lay down their arms, and the video was later confirmed as a rumor. This was the first time deepfake was used in war. Facial information plays a huge role in our lives, so once deepfake is abused and maliciously spread, it will pose a great threat to the personal information security and social stability. Therefore, it is urgent to design efficient and accurate methods to detect these deepfake contents.

Current deepfake detection methods, whether based on handcrafted features or deep features extracted by deep neural networks, are essentially mining the differences between real face images and fake face images, and then using the differences for classification. It is worth mentioning that most of the methods [2–9] utilize the texture information in the RGB domain of the image. These detection methods can achieve average detection accuracy of more than 90% on uncompressed datasets, but the detection accuracy will decrease significantly on compressed datasets, especially on highly compressed datasets. The detection accuracy of some methods testing on compressed datasets even drops to around 50%. That is to say, the detection method is completely invalid and cannot distinguish between real and fake face images. Since the purpose of a deepfake generation model is to produce RGB images that are difficult for the human eye to distinguish, more attention is paid to the adjustment of the RGB domain during the fine-tuning stage to erase the forgery traces. This postprocessing operation is one of the most important reasons for the drastic drop of performance degradation. In addition, another reason is that the detection model mainly focuses on the RGB domain, and when the image is highly compressed, the image quality is greatly degraded, leading to the weakening or even the disappearance of critical RGB textures used for detection.

In our work, we propose a deepfake detection method based on frequency-domain filtered residuals. The low frequency information of the human face is obtained by using Haar wavelet transform. Then, the residual calculation is made between the low frequency information map and the gray scale of the original image. After that, we obtain the mid-high frequency residual map of the original image. Finally, the original image concatenated with the residual image is fed into the convolutional neural network for classification.

The rest of this paper is arranged as follows: in Section 2, related works are briefly summarized and the proposed method is described in detail. Experimental results and conclusion are presented in Sections 3 and 4, respectively.

The contributions of our work can be outlined as follows:

- We develop an easily adaptable module that extracts a mid-high frequency map and fuse them with RGB images, which fully mines the features of forged face images in both the frequency and RGB domain.
- Instead of utilizing medium frequency or high frequency information directly, we obtain medium and high frequency joint information by residual operation, which is more comprehensive.
- The combination of the frequency and RGB domain allows the model to maintain its great performance on high-quality images and improve its generalization ability on low-quality images.
- We empirically demonstrate that our method outperforms baseline approaches on deepfake benchmark datasets with 88.09% average accuracy on low-quality deepfakes.

2. Materials and Methods

2.1. Related Work

2.1.1. Deepfake Generation

In general, deepfake generation algorithms can be divided into four main categories, which are deepfake [10], Face2Face [11], FaceSwap [12] and NeuralTexture [13].

(1) DeepFake. The generation process of DeepFake has two stages, the training stage and the generation stage. In the training phase, the weight sharing encoders are used to extract the latent features of the face, and the decoders are used to reconstruct the image. In the generation phase, the decoders are swapped to obtain the face swap image. There are many open-source methods of DeepFake, such as DFaker [14], DeepFaceLab [15], DeepFake-tf [16] and so on.

(2) Face2Face. Face2Face is a forgery technique of expression tampering, which enables the person in the source video to control the facial expressions and postures of the person in the target video or image.

(3) FaceSwap. FaceSwap is a tampering technique based on graphics. Firstly, we extract the face region using facial landmarks. Then a fitted 3D template model is back

projected to the target image. Finally, mixing and color correction are performed to the target image.

(4) Neural Texture. Neural texture is a new graphics primitive proposed by Justus Thies et al. [13]. It obtains a tampered image by learning the texture features of the target face and combining it with the background and identity information of the source face.

2.1.2. Deepfake Detection

Deepfake detection is essentially a classification task, and the purpose is to distinguish whether a given face image or video is real or fake. The generation process of deepfake can be roughly divided into two processing stages, namely the forgery stage and the postprocessing stage. In the forgery stage, although different forgery methods use different generation methods, these methods all leave unique operation traces in the face images. In the postprocessing stage, the generated face needs to be further refined and corrected and then fused with the background image. This stage usually leads to some boundary effect between the central facial area and the background area of the fake image. In order to eliminate the boundary effect, the color correction or smoothing of the facial area will further destroy the original underlying data distribution of the image. Therefore, these inherent flaws in deepfake will bring special forgery traces and underlying noise distribution to deepfake face images and videos.

Whether for the image or for the video, the general processing pipelines of deepfake detection are similar: (1) preprocessing the input (e.g., extract the frame of the video, crop out the face region); (2) extracting features from the preprocessed, the core part of the approach, in which commonly used features include handcrafted features extracted by classifiers such as Support Vector Machines(SVM) and deep features extracted by deep neural networks; and (3) the extracted features are used for classification and finally output as "real" or "fake".

Handcrafted features are widely used in the field of image classification [17–20]. With the development of deep learning techniques, deep features are used more and more widely in deepfake detection. The features in [21–25] are based on physiological signals such as eye blinking, phoneme-lip correspondence and so on. These methods are more realistic for identity video detection due to the use of high-level semantic features. Most fake videos (images) are different from real data in inter-frame coherence (texture consistency within an image). These inconsistencies are often used as evidence of face forgery [5–9]. Recently, transformer [26] has become a hit in computer vision, which leads to the widespread application of the attention mechanism in the deepfake detection field. Dang et al. [27] proposed a multi-task learning model with an attention mechanism used to locate the forged regions by training the learned attention map. Luo et al. [28] treated deepfake detection as a fine-grained classification problem. They utilize multiple spatial attention heads to focus on local regions and attention maps to fuse low-level features and high-level features. In [29], the authors propose a deepfake detection network fusing RGB features and textural information extracted by neural networks and signal processing methods, including an attention module. In [30], the authors proposed an attention-based deepfake detection method and achieved excellent performance on public datasets.

Frequency-based deepfake detection [7,28,31–33] is also a popular research direction recently because spatial-based features are not obvious enough to mine or not enough to prove the authenticity of the image. Luo et al. [28] proposed a multi-scale high-frequency feature extraction module, a residual-guided spatial attention module and a cross-modality attention module to fuse high-frequency information and RGB information. Qian et al. [32] proposed an effective method named F^3-Net, using frequency-aware image decomposition and local frequency statistics, but the number of parameters from its backbone doubled. Binh M. et al. [34] proposed frequency attention and multi-view-based knowledge distillation to detect low-quality compressed deepfake images. The frequency domain information and multi-view information of the teacher model are transferred by knowledge distillation. However, they do not mention the generalization ability of cross-dataset testing.

2.1.3. Wavelet Transform

A wavelet is a wave-like oscillation with an amplitude that begins at zero, increases or decreases and then returns to zero one or more times. Wavelets are termed a "brief oscillation". The basic idea of wavelet transform is to represent a function or signal by scaling and translating a set of functions. The Haar wavelet transform [35] is one of the oldest transform functions, proposed in 1910 by the Hungarian mathematician Alfred Haar. It is found effective in applications such as signal and image compression in electrical and computer engineering as it provides a simple and computationally efficient approach for analyzing the local aspects of a signal.

At present, wavelet transform has been widely used in computer vision tasks. Jaejun Yoo et al. [36] embedded the wavelet transform into the neural network based on the Whitening and Coloring Transforms (WCT2), and proposed a style transfer method with high temporal stability. J. M. Fortuna-Cervantes et al. [37] proposed a first approach for object detection with a repetitive pattern and binary classification in the image plane based on wavelet analysis. Huang et al. [38], utilizing wavelet transform, presented a generative adversarial approach called WaveletSRGAN for multi-scale face hallucination.

In our work, we chose Haar wavelet transform instead of Discrete Cosine Transform (DCT). The reason is that after the Haar wavelet transform, the wavelet domain of the image is divided into four sub-bands, each sub-band includes not only the frequency domain component of the image but also its spatial domain component, so containing richer information, while the DCT transformed image only contains frequency domain information.

2.2. Method

In this section, we will first introduce the motivation for the proposed method and then introduce the method in detail.

2.2.1. Motivation

In the process of deepfake, upsampling is a non-negligible step, and continuous upsampling operations usually leave traces in the image frequency domain. Research shows that in the highly compressed images and videos, although the forged face in the RGB domain tampering are significantly weakened, the tampering traces of the frequency domain are affected relatively less, which means that the frequency domain features can show the tampering traces more obviously. Compared with real images, there are usually some obviously abnormal frequency distribution forms in fake images [31]. Figure 1 shows the comparison of real images and two types of fake images (neural textures and Face2Face) under high compression in the RGB domain and the frequency domain, respectively. We obtain the frequency domain information through three different filters. As can be seen from Figure 1, it is difficult to detect the forged traces in the RGB domain, while in the frequency domain, in the central area of the face, especially the area containing the nose and mouth marked by the red frame in the figure, the forged images lack some texture and boundary information of the real images. Therefore, these forgery traces weakened or eliminated in the RGB domain can be preserved and captured in the frequency domain, especially in the mid-high frequency domain, making it easier to detect forgery traces [28]. Due to the above findings, we propose a method based on frequency domain filter residuals.

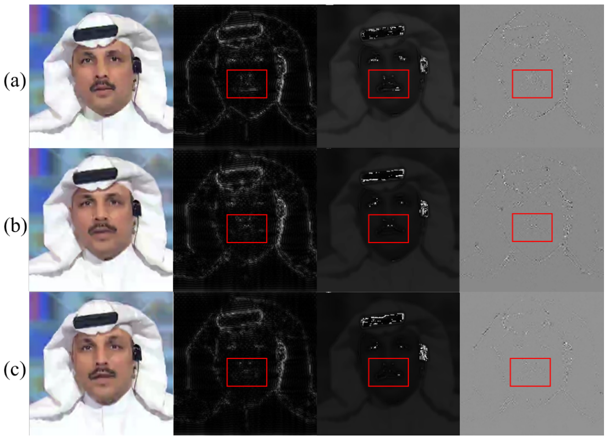

Figure 1. Frequency domain analysis of highly compressed face images: (**a**) real face images; (**b**) fake face images by neural textures; (**c**) fake face images by Face2Face. The forged images (**b**,**c**) lack some textures and boundary information of the real images (**a**), especially in the area containing the nose and mouth.

2.2.2. Frequency Domain Filtered Residual Network

An overview of our approach is illustrated in Figure 2. The framework of our method is simple, consisting of two modules: (1) a preprocessing module and (2) a classification module. The function of the preprocessing module is to obtain mid-high frequency maps and concatenate them with the RGB domain input. The Haar wavelet transform is first performed on the input image. We take the low-frequency subgraph after the Haar wavelet transform for bilinear interpolation so that the size of the low-frequency map is consistent with the input image. Then, the grayscale image of the input and the low-frequency information map after bilinear interpolation are used for residual calculation. By calculating the residuals from the input and low-frequency maps, we obtain a mid-high frequency map, which is then concatenated in the channel direction with the RGB domain input. In the classification module, concatenated maps are fed into the convolutional neural network for classification. We use Xception as the backbone network for the classification module. It is worth mentioning that the first module of our method acts as a preprocessing and does not affect the structure of the backbone network. So the computational complexity of our method does not change much compared to Xception.

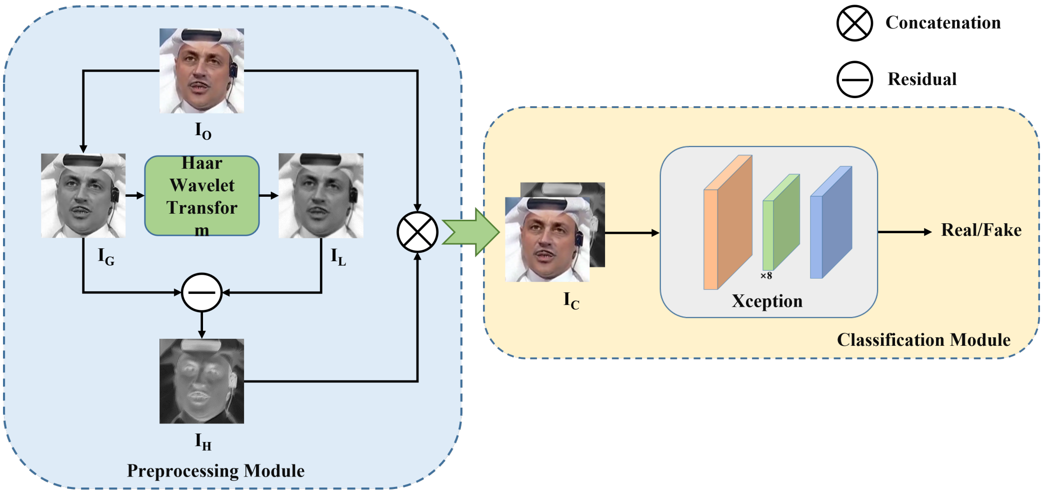

Figure 2. The framework of the algorithm.

Haar wavelet transform. First, the original image I_O is grayscaled to obtain I_G. This step can be expressed as a formula:

$$I_G = g(I_O) \tag{1}$$

where $g(\cdot)$ is the grayscale function. Then, we crop I_G into 2×2 subimages represented as I_{Gi} and perform the calculation shown in Equation (2) for all subimages.

$$\hat{I}_{Gi} = H^T I_{Gi} H \tag{2}$$

where H is the Haar transition matrix, as shown in Equation (3). The output of the Haar wavelet transform is indicated by $\hat{I}_{Gi}$ and shown in Figure 3.

$$H = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} \tag{3}$$

Bilinear interpolation. The low-frequency information map I_{L1} obtained after the Haar wavelet transformation is enlarged to the original image size by the bilinear interpolation, and the enlarged image is represented by I_{L2}. If not enlarged, the resolution of the four subimages is half of the original, which means that we cannot obtain the residuals with the original image. Let the size of I_{L1} be $W_{L1} \times H_{L1}$ and the size of I_{L2} be $W_{L2} \times H_{L2}$. Let the coordinate of the target pixel of I_{L2} be (X_{L2}, Y_{L2}) and the coordinate of the pixels mapped back to I_{L1} be (X_{L1}, Y_{L1}). The two coordinates satisfy the following formula:

$$X_{L1} = (X_{L2} + 0.5) \times \frac{W_{L1}}{W_{L2}} - 0.5 \tag{4}$$

$$Y_{L1} = (Y_{L2} + 0.5) \times \frac{H_{L1}}{H_{L2}} - 0.5 \tag{5}$$

Then, the grayscale images I_G and I_{L2} are used to obtain the residual image I_H of the mid-high frequency:

$$I_H = I_G - I_{L2} \tag{6}$$

The reason it is necessary to perform residual calculation on the grayscale image and the low-frequency information map after bilinear interpolation is that the proposed method aims to use the mid-high frequency information of the face image to mine forgery clues. The mid-high frequency information map contains the detailed information of the texture and edge of the image, while the low-frequency information map contains most of the original semantic information of the image, which can also be obtained in the RGB domain. In the detailed information of the mid-high frequency map, there is a greater probability to find the subtle differences of the real and fake face images so as to ensure that the features used for classification are more robust. However, the three high-frequency subimages directly generated after the Haar wavelet transform only provide the grayscale change information and edge texture information in the vertical, horizontal and diagonal directions of the image, respectively. Our method requires relatively complete mid-high frequency information, so the mid-high frequency information obtained after the residual is used instead of using the frequency information map after the Haar wavelet transform directly.

Concatenation After obtaining the mid-high frequency information residual map I_H, the original RGB image I_O and I_H are spliced together along the image channel direction. The process of obtaining the concatenated image I_C can be represented by Equation (7).

$$I_C = I_O \otimes I_H \tag{7}$$

where $\otimes$ represents the concatenation. The concatenation not only retains the rich semantic information of face images, which can ensure the accuracy when detecting uncompressed images, but also combines a large amount of mid-high frequency textures, which can be provided to the classification module when detecting compressed images. In this way, more forgery traces and feature selection are provided to the classification module, thereby increasing the robustness of the model.

In summary, the process of the preprocessing module can be summarized as follows:

$$I_C = I_O \otimes (g(I_O) - f_{BI}(h_L(g(I_O)))) \tag{8}$$

where $f_{BI}(\cdot)$ stands for the bilinear interpolation function, and $h_L(\cdot)$ represents the low frequency information subimage after the Haar wavelet transform.

Classification. The classification module takes the I_C output from the preprocessing module as input and obtains the classification result by training a convolutional neural network end-to-end. We choose Xception [39] as the backbone and cross entropy (Equation (9)) loss as the loss function.

$$Loss = -\left[Y \cdot \log(\hat{Y}) + (1-Y) \cdot \log(1-\hat{Y})\right] \tag{9}$$

where Y represents the real label, and $\hat{Y}$ represents the predicted label.

3. Results

Dataset Setting. To demonstrate the strong generality of our method, we perform the evaluation on multiple different datasets: FaceForensics++ (FF++) [40], Celeb-DeepFake (Celeb-DF) [41] and UADFV [19]. The videos from FF++ are compressed into two versions: medium compression (c23) and high compression (c40), using the H.264 codec with a constant rate quantization parameter of 23, and 40, respectively. For each subset of the FF++ dataset, we randomly selected 720 videos as the training set, 140 videos as the validation set, and 140 videos as the test set. Then we randomly extracted 50 frames from each video in the training set and 100 frames from the verification set and the test set. In the Celeb-DF dataset, we divided the training set, verification set and test set according to the ratio of 6:1:1. Since the number of real videos was much smaller than the number of forged videos, in order to ensure the balance between real data and forged data in the training process, the number of frames extracted from each video in real videos was larger than that of forged videos. For the UADFV dataset, 37 videos were randomly selected as the training set, 6 videos as the verification set and 6 videos as the test set. Then, we randomly extracted 50 frames from each video in the training set and 150 frames from the videos in the verification set and the test set. In this paper, we detected and extracted faces in videos of three datasets, and all image sizes were cut to 256 × 256.

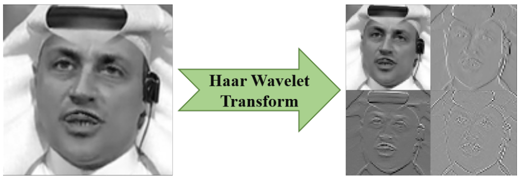

Figure 3. The result of the Haar wavelet transform. The low frequency information map in the upper left corner saves most of the information of the image, and the other three subimages reflect the edge texture information in different directions.

Implementation Details. We implemented our model with Pytorch. We set the initial value of the learning rate to 10^{-4}. Batchsize was set to 32, and the Adam optimizer [42] was used to train the network. We trained 20 epochs and multiplied the learning rate by 0.1 for every 5 epochs. To evaluate our model more comprehensively, we chose Accuracy (ACC) and Area Under Curve (AUC) as evaluation metrics. AUC is the measure of the ability of a classifier to distinguish between classes and is used as a summary of the ROC curve. The higher the AUC and ACC, the better the performance of the model at distinguishing between the real and fake classes. The difference between the two is that AUC represents the ability to classify or sort and has nothing to do with the classification threshold, while the ACC rate is related to the threshold.

3.1. Intra-Dataset Evaluation

Experimental Results under Uncompressed Scene (C0). Table 1 shows the ACC experimental results of various methods in the uncompressed scenario. The average accuracy of our method is more than 98% in the uncompressed scene. Although it is not optimal on some datasets, considering that the main contribution of the method proposed in this paper is in the compressed scene, the method proposed in this paper maintains its competitiveness in the uncompressed scene.

Table 1. The ACC results under uncompressed scene (C0) (%).

Methods	FF++ (C0)				Celeb-DF	UADFV
	DF	F2	FS	NT		
D-CNN [43]	98.03	98.96	98.94	96.06	-	-
Meso-4 [44]	96.37	97.95	98.17	93.30	87.10	82.67
MesoIn-4 [44]	88.34	97.65	97.81	92.52	88.10	96.33
Xception [39]	98.31	97.75	97.10	96.45	90.78	99.33
ours	98.85	99.08	98.19	98.87	92.25	99.94

Experimental Results under Lightly Compressed Scene (C23). Table 2 shows the ACC experimental results of various methods in the light compression scene. Our method achieves 90.35% detection accuracy on neural textures, which is the best result among all comparison methods. However, it performed poorly on the Face2Face and FaceSwap dataset, even worse than the Xception method in the benchmark method. The possible reason is that both of these two methods are expression tampering methods, and the tampered area is smaller than that of the forgery method of face replacement type, so the forgery trace left is lower. As a result, the proposed method cannot fully mine the forgery trace of the corresponding image, resulting in low detection accuracy. In this case, the global information including the boundaries of the nose and eyes presented by the mid-high frequency is not conducive to the experimental results.

Table 2. The ACC results under light compression scene (C23) (%).

Methods	FF++ (C23)			
	DF	F2	FS	NT
Local descriptors [45]	81.78	85.32	85.69	80.60
D-CNN [43]	82.16	93.48	92.51	75.18
Steg.Features [46]	77.12	74.68	79.51	76.94
NCL [47]	90.18	94.93	93.14	86.04
Meso-4 [44]	89.77	94.25	95.50	78.70
MesoIn-4 [44]	83.74	91.48	94.34	75.06
Xception [39]	95.15	97.07	95.96	87.99
ours	95.52	93.49	95.16	90.35

Experimental Results under Highly Compressed Scene (C40). Table 3 shows the ACC experimental results of various methods in the high compression scene. The results show that the accuracy of our method is lower than that of Xception and higher than that of the benchmark method testing on Face2Face. Our method combines features of the RGB domain and the frequency domain. In forgery detection, we can extract richer features. However, Face2Face forgery focuses on the lip area, so there are very few forgery traces, and even in the frequency domain, the number of forgery clues that can be recorded is limited. So, the detection accuracy decreases.

Table 3. The ACC results under high compression scene (C40) (%).

Methods	FF++ (C40)			
	DF	**F2**	**FS**	**NT**
Local descriptors [45]	68.26	59.38	62.08	62.42
D-CNN [43]	73.25	62.33	67.08	62.59
Steg.Features [46]	65.58	57.55	60.58	60.69
NCL [47]	80.95	77.30	76.83	72.38
Simple features [48]	71.69	65.66	65.43	59.34
OC-FakeDect [49]	88.40	71.20	86.10	97.50
Meso-4 [44]	77.68	83.65	79.92	77.74
MesoIn-4 [44]	74.20	78.75	79.72	67.94
Xception [39]	83.70	87.21	83.17	87.90
ours	89.28	84.85	86.90	91.32

3.2. Generalization Ability Evaluation

Although the method proposed in this paper is mainly to solve the problem of poor detection performance of existing methods for deepfake, in order to have a more comprehensive evaluation, we also evaluate the generalization ability of the proposed method. The training set of the proposed method and the baseline method is the Deepfakes sub-dataset of the lightly compressed version of the FF++ dataset, and then the cross-database test is performed on the Celeb-DF dataset. The experimental results are listed in their percentage value of the AUC. As described in Table 4, it can be seen that the generalization ability of the method based on the frequency domain residual map proposed in this paper exceeds the baseline method and most comparison methods in the fields, and the cross-database AUC test reaches 71.27%, illustrating the competitiveness of the proposed method in generalization ability.

Table 4. The AUC results of cross-database testing (%).

Methods	AUC
Simple features [48]	54.34
SMIL [50]	56.30
Capsule [51]	57.50
F^3-Net [32]	65.17
Multi [52]	67.44
MTD [4]	70.12
Ensemble of cnns [53]	71.14
Dual Network [7]	72.30
TSDA [54]	73.40
SPSL [31]	76.88
Meso-4 [44]	54.80
MesoIn-4 [44]	53.60
ours	72.27

3.3. Ablation Study

We conducted experiments on different variants of our model to prove the effectiveness of the proposed method.

Haar wavelet transform. The Wavelet transform is different from the Fourier transform, and the result of the wavelet transform is different according to the choice of the wavelet basis. The commonly used wavelet bases are Haar, Symlets, Coiflets, Daubechies, etc. We conducted experiments with different wavelet bases under the high compression scene (C40), and the results obtained are shown in Table 5. It can be seen that the model performance is similar when tested with different wavelet bases. Considering the average performance and the simplicity of the wavelet function, the Haar wavelet transform is the most suitable.

Table 5. The ACC results of different wavelet bases under the high compression scene (C40).

Wavelet Basis	FF++ (C0)				Average
	DF	F2	FS	NT	
Sym2	89.08	81.14	88.40	89.63	87.06
Coif2	89.02	81.79	88.14	80.06	84.75
Db2	91.71	82.60	88.09	88.79	87.80
Haar	89.28	84.85	86.90	91.96	88.25

Concatenation. In the proposed method, we concatenate the residual image with the original image. To demonstrate the effectiveness of the concatenation, we conduct experiments on FF++, Celeb-DF v2 and UADFV datasets with different compression scenes. The experimental results are shown in Table 6. Experimental results show that the model combining frequency domain with RGB domain shows better performance in most cases. The reason for the performance degradation on the F2F dataset is that F2F only tampers locally (especially the mouth), resulting in the information brought by the frequency domain misleading the classifier.

Table 6. Performance of different domains.

Domain	FF++ (C0)				FF++ (C23)				FF++ (C40)				Celeb-DF v2	UADFV
	DF	F2	FS	NT	DF	F2	FS	NT	DF	F2	FS	NT		
RGB	98.31	97.75	98.10	96.45	95.15	97.07	95.96	87.99	83.70	87.21	83.17	87.90	90.78	99.33
Frequency	98.65	98.04	95.86	98.04	91.49	92.11	92.74	83.72	89.51	80.67	79.29	77.29	88.87	91.39
RGB+Frequency	98.85	99.08	98.19	98.87	95.52	93.49	95.16	90.35	89.28	84.85	86.90	91.32	92.25	99.94

4. Conclusions

In this paper, we propose a novel spatial-frequency domain deepfake detection methodology for improving the detection of low-quality compressed deepfake images. By performing residual operations on a grayscale image of the original image and the low-frequency information image after Haar wavelet transformation, unlike some methods that utilize the output of wavelet transform directly, the residual image of the medium and high frequency information of the image is obtained. Then, the original image and the residual image of the medium and high frequency information are spliced together and input into the convolution neural network for detection of fake images. The experimental results show that the detection performance of the proposed method for most forgery techniques is better than that of the benchmark methods in the three compression scenarios, and the average detection performance drop of all forgery techniques in the compression scenario is better than those of the benchmark methods.

Since most of the images and videos on social media are compressed, our research is very relevant. In addition, our simple framework is easy to implement. However, our approach still has some limitations: (1) when performing generalization tests, our method performed poorly on the Face2Face and FaceSwap dataset, even worse than the baseline method, and (2) performance degradation with low-quality compressed deepfake images improved but was still there. In addition, due to the limited public dataset and the large number of video and image processing operations on the Internet, our method needs to be adjusted in practical application. Therefore, in the future, the efficient and accurate detection of low-quality images is still the main direction of research. In addition, we will explore other effective applications of wavelet transforms in deepfakes detection.

Author Contributions: Conceptualization, B.W.; methodology, X.W.; software, Y.T.; validation, B.W. and Y.T.; formal analysis, Y.M.; investigation, Y.M.; data curation, F.W.; writing—original draft preparation, X.W. and Y.M.; writing—review and editing, X.W.; supervision, F.W.; project administration, Z.S.; funding acquisition, Z.S. All authors have read and agreed to the published version of the manuscript.

Funding: This work is supported by the National Natural Science Foundation of China (No. U1936117, No. 62106037, No. 62076052), the Science and Technology Innovation Foundation of Dalian (No. 2021JJ12GX018), the Application Fundamental Research Project of Liaoning Province (2022JH2/101300262), the Open Project Program of the National Laboratory of Pattern Recognition (NLPR) (No. 202100032), and the Fundamental Research Funds for the Central Universities (DUT21GF303, DUT20TD110, DUT20RC(3)088).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data will be available upon reasonable request to the corresponding author.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

GAN	Generative Adversarial Network
SVM	Support Vector Machines
WCT2	Whitening and Coloring Transforms
DCT	Discrete Cosine Transform
FF++	FaceForensics++
ACC	Accuracy
AUC	Area Under Curve

References

1. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative adversarial nets. *Adv. Neural Inf. Process. Syst.* **2014**, *27*, 53–65.
2. Wang, B.; Li, Y.; Wu, X.; Ma, Y.; Song, Z.; Wu, M. Face forgery detection based on the improved siamese network. *Secur. Commun. Netw.* **2022**, *2022*, 5169873. [CrossRef]
3. Yang, J.; Xiao, S.; Li, A.; Lan, G.; Wang, H. Detecting fake images by identifying potential texture difference. *Future Gener. Comput. Syst.* **2021**, *125*, 127–135. [CrossRef]
4. Yang, J.; Li, A.; Xiao, S.; Lu, W.; Gao, X. MTD-Net: Learning to detect deepfakes images by multi-scale texture difference. *IEEE Trans. Inf. Forensics Secur.* **2021**, *16*, 4234–4245. [CrossRef]
5. Li, Y.; Lyu, S. Exposing DeepFake Videos By Detecting Face Warping Artifacts. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, Long Beach, CA, USA, 16 June 2019.
6. Matern, F.; Riess, C.; Stamminger, M. Exploiting Visual Artifacts to Expose Deepfakes and Face Manipulations. In Proceedings of the 2019 IEEE Winter Applications of Computer Vision Workshops (WACVW), Waikoloa Village, HI, USA, 7–11 January 2019; pp. 83–92. [CrossRef]
7. Jia, G.; Zheng, M.; Hu, C.; Ma, X.; Xu, Y.; Liu, L.; Deng, Y.; He, R. Inconsistency-Aware Wavelet Dual-Branch Network for Face Forgery Detection. *IEEE Trans. Biom. Behav. Identity Sci.* **2021**, *3*, 308–319. [CrossRef]
8. Sun, Z.; Han, Y.; Hua, Z.; Ruan, N.; Jia, W. Improving the Efficiency and Robustness of Deepfakes Detection through Precise Geometric Features. In Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021; pp. 3608–3617. [CrossRef]
9. Haliassos, A.; Vougioukas, K.; Petridis, S.; Pantic, M. Lips Don't Lie: A Generalisable and Robust Approach to Face Forgery Detection. In Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021; pp. 5037–5047. [CrossRef]
10. Deepfakes. 2019. Available online: https://github.com/deepfakes/faceswap (accessed on 10 October 2022).
11. Thies, J.; Zollhöfer, M.; Stamminger, M.; Theobalt, C.; Nießner, M. Face2Face: Real-Time Face Capture and Reenactment of RGB Videos. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 2387–2395. [CrossRef]
12. Faceswap. 2019. Available online: https://www.github.com/MarekKowalski/FaceSwap (accessed on 10 October 2022).
13. Thies, J.; Zollhöfer, M.; Nießner, M. Deferred neural rendering: Image synthesis using neural textures. *ACM Trans. Graph.* **2019**, *38*, 1–12. [CrossRef]
14. DFaker. 2019. Available online: https://github.com/dfaker/df (accessed on 10 October 2022).
15. DeepFaceLab. 2019. Available online: https://github.com/iperov/DeepFaceLab (accessed on 10 October 2022).
16. DeepFake-tf. 2019. Available online: https://github.com/StromWine/DeepFake-tf (accessed on 15 October 2022).

17. Chen, H.; Miao, F.; Chen, Y.; Xiong, Y.; Chen, T. A Hyperspectral Image Classification Method Using Multifeature Vectors and Optimized KELM. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2021**, *14*, 2781–2795. [CrossRef]
18. Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A. Going deeper with convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 1–9.
19. Yang, X.; Li, Y.; Lyu, S. Exposing deep fakes using inconsistent head poses. In Proceedings of the ICASSP 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 12–17 May 2019; pp. 8261–8265.
20. Huang, C.; Zhou, X.; Ran, X.; Liu, Y.; Deng, W.; Deng, W. Co-evolutionary competitive swarm optimizer with three-phase for large-scale complex optimization problem. *Inf. Sci.* **2023**, *619*, 2–18. [CrossRef]
21. Li, Y.; Chang, M.C.; Lyu, S. In 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), Hong Kong, China, 11–13 December 2018; pp. 1–7. [CrossRef]
22. Jung, T.; Kim, S.; Kim, K. DeepVision: Deepfakes Detection Using Human Eye Blinking Pattern. *IEEE Access* **2020**, *8*, 83144–83154. [CrossRef]
23. Ciftci, U.A.; Demir, I.; Yin, L. FakeCatcher: Detection of Synthetic Portrait Videos using Biological Signals. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, 1. [CrossRef] [PubMed]
24. Agarwal, S.; Farid, H.; El-Gaaly, T.; Lim, S.N. Detecting Deep-Fake Videos from Appearance and Behavior. In Proceedings of the 2020 IEEE International Workshop on Information Forensics and Security (WIFS), New York, NY, USA, 6–11 December 2020; pp. 1–6. [CrossRef]
25. Lin, J.; Zhou, W.; Liu, H.; Zhou, H.; Zhang, W.; Yu, N. Lip Forgery Video Detection via Multi-Phoneme Selection. In Proceedings of the 2021 International Workshop on Safety and Security of Deep Learning, New York, NY, USA, 19–20 August 2021.
26. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention Is All You Need. *arXiv* **2017**, arXiv:1706.03762.
27. Dang, H.; Liu, F.; Stehouwer, J.; Liu, X.; Jain, A.K. On the Detection of Digital Face Manipulation. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 5780–5789. [CrossRef]
28. Luo, Y.; Zhang, Y.; Yan, J.; Liu, W. Generalizing Face Forgery Detection with High-frequency Features. In Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021; pp. 16312–16321.
29. Zhao, L.; Zhang, M.; Ding, H.; Cui, X. MFF-Net: Deepfake detection network based on multi-feature fusion. *Entropy* **2021**, *23*, 1692. [CrossRef] [PubMed]
30. Khormali, A.; Yuan, J.S. Add: Attention-based deepfake detection approach. *Big Data Cogn. Comput.* **2021**, *5*, 49. [CrossRef]
31. Liu, H.; Li, X.; Zhou, W.; Chen, Y.; He, Y.; Xue, H.; Zhang, W.; Yu, N. Spatial-Phase Shallow Learning: Rethinking Face Forgery Detection in Frequency Domain. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021; pp. 772–781.
32. Qian, Y.; Yin, G.; Sheng, L.; Chen, Z.; Shao, J. Thinking in Frequency: Face Forgery Detection by Mining Frequency-Aware Clues. In Proceedings of the Computer Vision—ECCV 2020, Glasgow, UK, 23–28 August 2020; Vedaldi, A., Bischof, H., Brox, T., Frahm, J.M., Eds.; Springer International Publishing: Cham, Switzerland, 2020; pp. 86–103.
33. Li, J.; Xie, H.; Li, J.; Wang, Z.; Zhang, Y. Frequency-aware Discriminative Feature Learning Supervised by Single-Center Loss for Face Forgery Detection. In Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021; pp. 6454–6463. [CrossRef]
34. Woo, S.; Le, B.M. ADD: Frequency Attention and Multi-View Based Knowledge Distillation to Detect Low-Quality Compressed Deepfake Images. In Proceedings of the AAAI Conference on Artificial Intelligence, Online, 22 February–1 March 2022; Volume 36, pp. 122–130.
35. Haar, A. *Zur Theorie der Orthogonalen Funktionensysteme*; Georg-August-Universitat: Gottingen, Germany, 1909.
36. Yoo, J.; Uh, Y.; Chun, S.; Kang, B.; Ha, J.W. Photorealistic style transfer via wavelet transforms. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Korea, 27 October–2 November 2019; pp. 9036–9045.
37. Fortuna-Cervantes, J.M.; Ramírez-Torres, M.T.; Martínez-Carranza, J.; Murguía-Ibarra, J.S.; Mejía-Carlos, M. Object Detection in Aerial Navigation using Wavelet Transform and Convolutional Neural Networks: A First Approach. In Proceedings of the Program Computer Soft, Madrid, Spain, 13–17 July 2020; Volume 46, pp. 536–547.
38. Huang, H.; He, R.; Sun, Z.; Tan, T. Wavelet domain generative adversarial network for multi-scale face hallucination. *Int. J. Comput. Vis.* **2019**, *127*, 763–784. [CrossRef]
39. Chollet, F. Xception: Deep learning with depthwise separable convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 1251–1258.
40. Rossler, A.; Cozzolino, D.; Verdoliva, L.; Riess, C.; Thies, J.; Niessner, M. FaceForensics++: Learning to Detect Manipulated Facial Images. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea, 27 October–2 November 2019; pp. 1–11. [CrossRef]

41. Li, Y.; Yang, X.; Sun, P.; Qi, H.; Lyu, S. Celeb-DF: A Large-Scale Challenging Dataset for DeepFake Forensics. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 3204–3213. [CrossRef]
42. Kingma, D.; Ba, J. Adam: A Method for Stochastic Optimization. *arXiv* **2014**, arXiv:1412.6980.
43. Rahmouni, N.; Nozick, V.; Yamagishi, J.; Echizen, I. Distinguishing computer graphics from natural images using convolution neural networks. In Proceedings of the 2017 IEEE Workshop on Information Forensics and Security (WIFS), Rennes, France, 4–7 December 2017; pp. 1–6. [CrossRef]
44. Afchar, D.; Nozick, V.; Yamagishi, J.; Echizen, I. Mesonet: A compact facial video forgery detection network. In Proceedings of the 2018 IEEE International Workshop on Information Forensics and Security (WIFS), Hong Kong, China, 11–13 December 2018; pp. 1–7.
45. Cozzolino, D.; Poggi, G.; Verdoliva, L. Recasting Residual-based Local Descriptors as Convolutional Neural Networks: An Application to Image Forgery Detection. *arXiv* **2017**, arXiv:1703.04615.
46. Fridrich, J.; Kodovsky, J. Rich Models for Steganalysis of Digital Images. *IEEE Trans. Inf. Forensics Secur.* **2012**, *7*, 868–882. [CrossRef]
47. Bayar, B. A Deep Learning Approach to Universal Image Manipulation Detection Using a New Convolutional Layer. In Proceedings of the ACM, Tacoma, WA, USA, 18–20 August 2016.
48. Durall, R.; Keuper, M.; Pfreundt, F.J.; Keuper, J. Unmasking DeepFakes with simple Features. *arXiv* **2019**, arXiv:1911.00686.
49. Khalid, H.; Woo, S.S. OC-FakeDect: Classifying Deepfakes Using One-class Variational Autoencoder. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Seattle, WA, USA, 14–19 June 2020; pp. 2794–2803. [CrossRef]
50. Li, X.; Lang, Y.; Chen, Y.; Mao, X.; He, Y.; Wang, S.; Xue, H.; Lu, Q. Sharp multiple instance learning for deepfake video detection. In Proceedings of the 28th ACM International Conference on Multimedia, Seattle, WA, USA, 12–16 October 2020; pp. 1864–1872.
51. Nguyen, H.H.; Yamagishi, J.; Echizen, I. Capsule-forensics: Using capsule networks to detect forged images and videos. In Proceedings of the ICASSP 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 12–17 May 2019; pp. 2307–2311.
52. Zhao, H.; Zhou, W.; Chen, D.; Wei, T.; Zhang, W.; Yu, N. Multi-attentional deepfake detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 2185–2194.
53. Bonettini, N.; Cannas, E.D.; Mandelli, S.; Bondi, L.; Bestagini, P.; Tubaro, S. Video face manipulation detection through ensemble of cnns. In Proceedings of the 2020 25th International Conference on Pattern Recognition (ICPR), Milan, Italy, 10–15 January 2021; pp. 5012–5019.
54. Bondi, L.; Cannas, E.D.; Bestagini, P.; Tubaro, S. Training strategies and data augmentations in cnn-based deepfake video detection. In Proceedings of the 2020 IEEE International Workshop on Information Forensics and Security (WIFS), New York, NY, USA, 6–11 December 2020; pp. 1–6.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.