Wiley
International Journal of Intelligent Systems
Volume 2024, Article ID 7578036, 12 pages
https://doi.org/10.1155/2024/7578036

WILEY

Research Article

Deepfake Detection Based on the Adaptive Fusion of Spatial-Frequency Features

Fei Wang[1], **Qile Chen**[1], **Botao Jing**[1], **Yeling Tang**[1], **Zengren Song**[2] and **Bo Wang**[1]

[1]*School of Information and Communication Engineering, Dalian University of Technology, Dalian, China*
[2]*National Computer Network Emergency Response Technical Team, Coordination Center of China, Beijing, China*

Correspondence should be addressed to Zengren Song; songzr0518@163.com

Received 12 June 2024; Accepted 14 October 2024

Academic Editor: Beijing Chen

Detecting deepfake media remains an ongoing challenge, particularly as forgery techniques rapidly evolve and become increasingly diverse. Existing face forgery detection models typically attempt to discriminate fake images by identifying either spatial artifacts (e.g., generative distortions and blending inconsistencies) or predominantly frequency-based artifacts (e.g., GAN fingerprints). However, a singular focus on a single type of forgery cue can lead to limited model performance. In this work, we propose a novel cross-domain approach that leverages a combination of both spatial and frequency-aware cues to enhance deepfake detection. First, we extract wavelet features using wavelet transformation and residual features using a specialized frequency domain filter. These complementary feature representations are then concatenated to obtain a composite frequency domain feature set. Furthermore, we introduce an adaptive feature fusion module that integrates the RGB color features of the image with the composite frequency domain features, resulting in a rich, multifaceted set of classification features. Extensive experiments conducted on benchmark deepfake detection datasets demonstrate the effectiveness of our method. Notably, the accuracy of our method on the challenging FF++ dataset is mostly above 98%, showcasing its strong performance in reliably identifying deepfake images across diverse forgery techniques.

Keywords: adaptive fusion; deepfake detection; spatial and frequency domain; wavelet transform

1. Introduction

The rapid proliferation of deepfake technologies, which use deep learning techniques to generate highly realistic yet fabricated media, poses a significant threat to the integrity and credibility of digital content [1, 2]. The mainstream methods of deepfakes are shown in Figure 1. The forged images are highly realistic, to the extent that they can be mistaken for genuine. The ability to create convincing fake images, videos, and audio has far-reaching implications, from the spread of misinformation and undermining of trust in media to the potential for malicious exploitation in areas such as fraud, blackmail, and election interference. In the recent Russia–Ukraine conflict, many videos or images of Zelensky's remarks circulated on the Internet have been proven to be fake. As these forgery techniques continue to evolve and become more sophisticated, the challenge of accurately detecting deepfake content has emerged as a critical research problem.

In recent years, the abuse of deep forgery technology has brought the prosperity of deepfake detection methods. Some methods are based on the inherent patterns of images, such as speeded-up robust features (SURFs) [3], photo response nonuniformity (PRNU) [4], and local binary patterns (LBPs) [5]. Some methods reveal inconsistent features of images, such as face X-ray [6], subtle manipulation patterns [7], implicit identity [8, 9], and global texture extraction [10]. Some methods focus on designing network structures, such as MesoNet [11], multitask model [12], and attention mechanism [13]. The others use the continuity anomaly

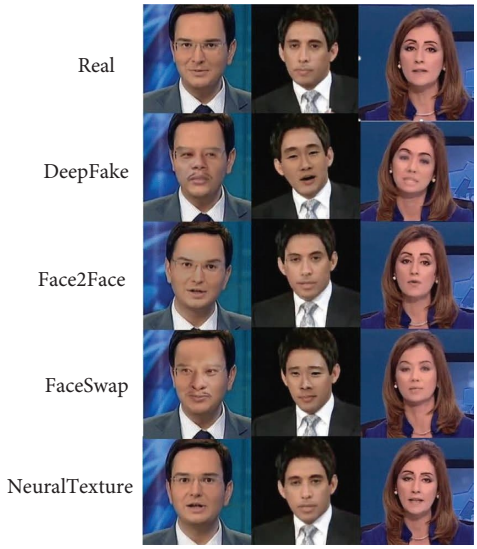

Figure 1: Comparison of four types of deepfake images from FaceForensics++ datasets. They are DeepFake, Face2Face, FaceSwap, and NeuralTexture. Among them, DeepFake and FaceSwap are face identity replacements, while Face2Face and NeuralTexture are face attribute edits.

facial posture between continuous frames of video to detect image tampering, such as eye-blinking detection [14], head-pose-changing detection [15], and phoneme-vision-mismatches detection [16]. Existing detection approaches often focus on identifying either spatial artifacts or frequency-based cues, but a singular reliance on a single type of forgery indicator can limit the effectiveness of these methods.

Considering the above issues, in this work, we propose a novel network architecture called the cross-domain fusion network with one spatial and two frequency domain (CFNSFD) features. Our approach combines color domain features and both low-level and high-level frequency domain features of images, leveraging a cross-domain fusion module to integrate these complementary representations. First, we extract the color feature in the spatial domain, which can comprehensively capture the original image characteristics. We then obtain two distinct frequency domain features. The shallow frequency feature is extracted directly from the input image using wavelet transformation. The deep frequency domain feature, on the other hand, is obtained by first filtering the input to produce a residual image and then extracting features from this residual using a specialized convolutional extractor. The residual images highlight the edge information of the images, which can be more discriminative for deepfake detection. We concatenate the wavelet-based frequency features and the residual-based frequency features to form a composite frequency domain feature set. Finally, we fuse this composite frequency domain feature with the color features in the spatial domain using an adaptive feature fusion module. The resulting fused features are then input to a fully connected layer for end-to-end deepfake classification. Extensive experiments on the FF++ dataset demonstrate the effectiveness of our approach, with most of the accuracy above 98%.

Our contributions can be concluded as follows:

- We propose a cross-domain method that combines spatial domain and frequency domain features. We extract a comprehensive set of features including color domain features, shallow frequency domain features, and deep frequency domain features.
- We propose a feature adaptive fusion module that enables the features of different branches to be fully expressed in the classification stage.
- Extensive experiments demonstrate the effectiveness of our method.

The remainder of this paper is organized as follows. Section 2 reviews the related work on deepfake detection. Section 3 details the proposed cross-domain detection framework, including the feature extraction and fusion strategies. Section 4 presents the experimental setup and results, along with ablation studies and discussions. Finally, Section 5 concludes the paper and outlines future research directions.

2. Related Work

2.1. Face Tampering. Among all the face tampering methods, we usually divide them into four categories according to the implementation principle [17], including DeepFake, Face2Face, FaceSwap, and NeuralTexture. Examples of these four types of face-tampering images are depicted in Figure 1. DeepFake and FaceSwap are face-replacement methods, which use a source image to replace the face of target images. Face2Face and NeuralTexture are facial reenactment methods, using a source image to reenact the facial expressions of target images. Furthermore, DeepFake and NeuralTexture are learning-based methods, and the other two are graphics-based methods.

DeepFake is generated by two autoencoders with shared weights. First, we train two autoencoders to reconstruct two input images separately. These two encoders have the same weights but are different in the decoder. After training, the input image of the class is encoded by the share-weight encoder but decoded to reconstruct as another class.

FaceSwap uses information on facial landmarks to get sparse knowledge of faces. Then, it uses these facial landmarks to do 3D effect transformation. After that, it is projected back to the target face by minimizing their distance.

Face2Face uses a dense reenactment system with the whole image pixel by pixel. The target images' pose, change, and expression move according to the source images totally.

In NeuralTexture, first, the neural texture pattern is trained from source images and then the target images are adjusted to match the neural texture pattern from the source images. Finally, photometric reconstruction loss and adversarial loss are used during the training stage.

2.2. Face Tampering Detection. Face forgery based on simple copy move can be detected by simple machine learning methods. But with the increasingly sophisticated developments in computer graphics and neural networks,

these methods have experienced a significant decline in effectiveness. Correspondingly, face forgery detection has been gradually updated. In this part, we introduce the current progress in the field of face-tampering detection.

2.2.1. Spatial-Based Tampering Detection. In recent years, many researchers have tried to detect forgery contents based on neural networks. Most of them focused on mining the color features of the image, such as RGB or CMKY pattern, whether it is based on manual features or models. Matern et al. [18] ignored the detailed facial characteristics but focused on the global artifacts, such as the color of irises and weird shadows on the face. Yang et al. [15] exploited head-pose reconstruction. After 2D and 3D head-pose reconstruction, if the angle between the two directional vectors exceeds a certain threshold, then it is identified as a forged image. Li et al. [5] observed the fusion boundary of the image. If there are color abnormal points on the fusion boundary, through face X-ray, the conversion binary image will have an obvious fusion boundary. Khalid et al. [19] trained a one-class classifier (autoencoder), which can reconstruct images, and use the reconstructed score to judge whether it is an abnormal image (forged). Zhu et al. [20] proposed a deepfake detection approach that extracts and fuses features from the YCbCr and RGB color spaces. Coccomini et al. [21] applied the spatial features from the original video frames to EfficientNet, Vision Transformer (ViT), and Cross-ViT. The AltFreezing [22] adopts both spatial and temporal artifacts to achieve face forgery detection. Chen et al. [23] utilized spatial domain features and adopted an adversarial network to train the generator and discriminator simultaneously. Zhang et al. [24] divided the original image into several blocks of the same size before extracting spatial features, forcing the model to explore more discriminative forgery traces. Other studies such as [11, 25] used neural networks to extract high-level information from the spatial domain of images.

2.2.2. Frequency-Based Tampering Detection. Knowledge of the frequency domain is very important for forming an image. They have been used in image classification [26, 27], superpixel [28], and so on. Durall et al. [29] found that in the frequency domain, the artifacts are distinguished since the spectrum of real and fake images has different distributions. Because the visual artifacts of tampering images are difficult to detect, many studies turn to using frequency domain features to detect forged images and have achieved a lot of success. Qian et al. [7] found two frequency-aware cues. One is to decompose image components in the frequency domain, and the other is to calculate local frequency statistics and propose an F^3 network. Liu et al. [30] used discrete Fourier transform (DFT) to get the spectrum. They considered not only amplitude but also phase and presented a spatial-phase shallow learning method to capture the upsampling artifacts. Luo et al. [31] utilized high-frequency noises for face forgery detection. They proposed a complex network to extract high-frequency noise characteristics. Masi et al. [32] used the Laplacian of Gaussian (LoG) algorithm for frequency-based feature enhancement, suppressing the image content which is represented by low-level features. Luo et al. [33] found that in the image frequency domain, high-frequency signals that eliminate color textures are more effective in distinguishing real and forged videos than low-frequency signals. Tan et al. [34] rethought the upsampling artifacts in the frequency domain. ADD [35] developed two distillation modules for detecting highly compressed deepfake, including frequency attention distillation and multiview attention distillation. BiHPF [36] amplified the size of artifacts through two high-pass filters. FreGAN [37] observed that unique frequency-level artifacts in the generated images lead to overfitting on the training source. Therefore, FreGAN mitigated the impact of frequency-level artifacts through frequency-level perturbation. Tan et al. [38] improved the generalizability through frequency-domain learning. Doloriel and Cheung [39] proposed a novel deepfake detector through frequency masking.

In our work, we think that it is not sufficient to use only RGB color texture or frequency domain characteristics. Our detection network must not lose information from both sides. In addition, compared with existing frequency-based methods, we believe that it is not sufficient to obtain the spectrum of the image by using a fixed filter because the image forgery is tricky and difficult to capture by using a fixed paradigm. So, the feature extraction module with dynamic characteristics is also under consideration.

3. Proposed Method

In this section, we first comprehensively review the motivation behind our approach. Then, we provide a detailed description of our method. As shown in Figure 2, the input image is divided into three branches for separate processing. This includes directly extracting RGB color features from the original input image, extracting wavelet features (shallow-frequency features) using wavelet transform, and extracting residual features (deep-frequency domain features) using residual filters. The RGB features are directly extracted by inputting the original color image into the ResNet-34 network. The extraction methods for wavelet features and residual features will be detailed in the following sections. The composite frequency domain features are formed by concatenating the wavelet frequency domain features and the residual frequency domain features. The classification features are obtained by fusing the composite features and the color features through the adaptive feature fusion module. Finally, the fused features are fed into a fully connected layer for end-to-end classification.

In this paper, we utilize the ResNet-34 network [40] as the backbone network to implement face forgery detection. ResNet-34 introduces residual connections, which make it

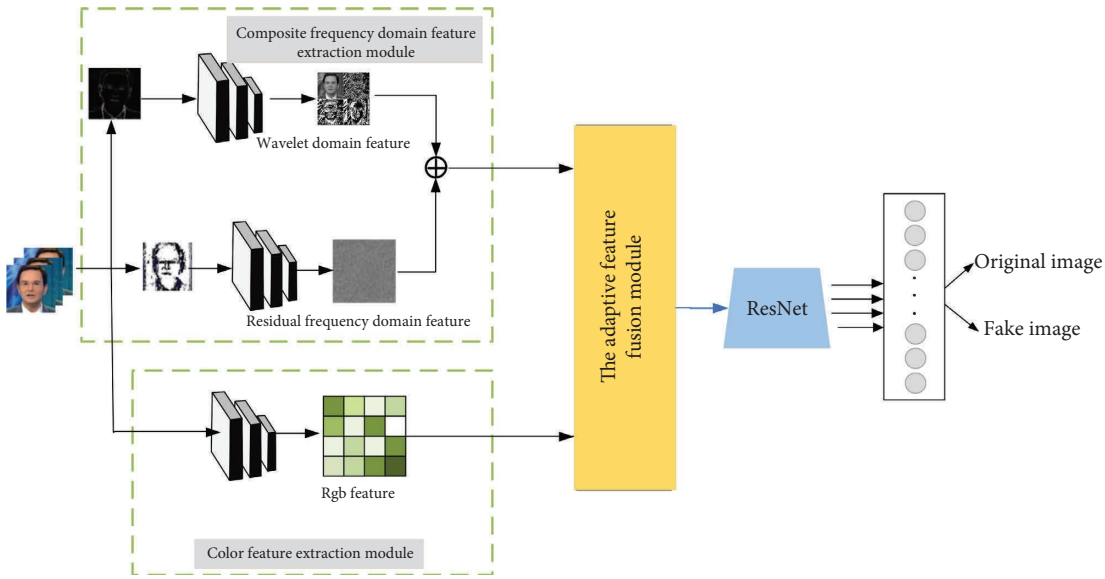

FIGURE 2: The diagram of our cross-domain fusion network with one spacial and two frequency domain (CFNSFD) feature method. We extract wavelet domain features, residual frequency domain features, and RGB features, respectively, while concatenating the first two and then fusing with the last one in adaptive feature fusion modules.

easier for the network to learn identity mappings during training, thus alleviating the problem of degradation in deep networks. Compared to ResNet-18, ResNet-34 has more layers to learn features from images, allowing it to better capture the complex details and textures present in forged images, thereby enhancing its feature learning capabilities. In contrast to even deeper network architectures, ResNet-34 offers lower computational and memory requirements.

3.1. Motivation. Frequency domain analysis can capture the global information of the entire image. However, image compression can potentially alter its frequency domain characteristics, which may result in a performance decline when processing compressed images using frequency domain analysis. In contrast to frequency domain analysis, spatial domain analysis is less susceptible to the effects of image compression because it directly operates on the image's pixels. However, spatial domain analysis methods may struggle to capture very covert and finely detailed forgery information. Therefore, we combine spatial domain and frequency domain analyses to provide a more comprehensive description of image features. This comprehensive analysis allows the model to better understand the image content and can enhance its robustness in detecting various forgery techniques.

Face forgery methods proposed in recent years have done rigorous postprocessing, including boundary fuzzy and interpolation, which make the mixed boundaries look invisible, especially in the RGB domain where faces can be directly observed by human eyes, while most forgery detection methods also focus on this. However, as shown in Figure 3, we observed that although the real images and the tampering images have almost invisible differences in the RGB domain, their statistical characteristics are significantly distinguished when we convert them to the frequency domain. Therefore, we propose a face tampering detection method based on the cross-domain feature fusion in the color domain and frequency domain. Because the high-frequency band of the face implies the image fusion information during face tampering, we use filters with dynamic characteristics to filter the original image to get the shallow frequency-domain cues. Moreover, we employ three filter combinations to filter the residual maps and extract deep frequency domain cues. In addition, we propose an adaptive feature fusion module with gated convolutions [41] to dynamically fuse the color domain and frequency domain features. This module adaptively integrates the two domains based on their relevance and importance in the fusion process.

3.2. Wavelet Feature. Wavelet transform, similar to the short-time Fourier transform, allows us to obtain the time–frequency domain characteristics of an image and perform localized analysis. The short-time Fourier transform achieves this by using "Windows" to decompose the image into segments, performing the Fourier transform on each segment and then concatenating them in the time domain. However, this approach faces the challenge of selecting an appropriate window size. The window size needs to be fixed within the signal period, making it difficult to represent the characteristics of signals that exhibit rapid changes. In contrast, wavelet transform utilizes a scale parameter to decompose the image into different frequency resolutions, enabling multiscale refinement of the image and achieving dynamic frequency decomposition.

Within the wavelet transform family, there are over a dozen functions to choose from. Although the Haar wavelet transform is the most commonly used one, its scaling function and wavelet function are discontinuous, resulting in lower smoothness and higher computational

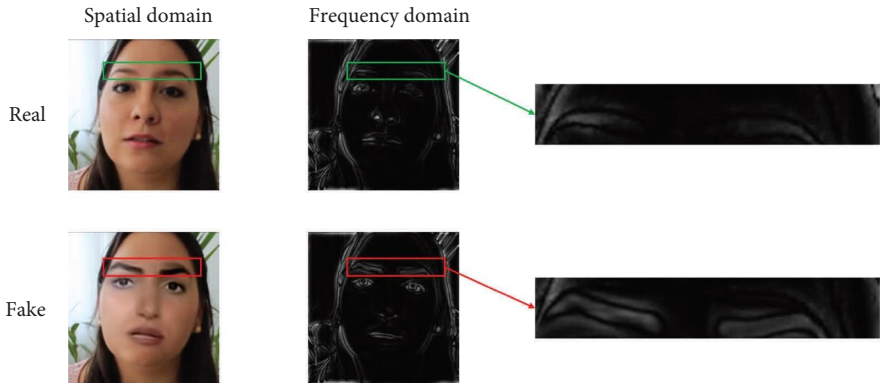

Figure 3: Comparison of spatial and frequency domain of real and fake images.

cost. In contrast, the Daubechies wavelet transform [42] exhibits better spectral response characteristics and lower tail effects, allowing for improved differentiation of different frequencies. Daubechies wavelet has been proven to generate accurate results and has excellent computational efficiency [43]. Compared with other wavelet transforms, Daubechies wavelet transform has the following advantages: (1) orthogonality: Daubechies wavelet functions are orthogonal, meaning that their analysis and synthesis functions are mutually orthogonal. Orthogonality ensures the integrity of information and allows for a complete reconstruction of the original signal without the need for redundant information. (2) compact support: Daubechies wavelet functions have a compact support region, which allows them to better localize signal features and thus capture signal discontinuities and abrupt changes more effectively. (3) higher-order continuity: Daubechies wavelet functions of different orders have varying degrees of continuity. Higher-order Daubechies wavelet functions have greater continuity, which makes them more suitable for smooth signal processing and image processing. Therefore, we utilize the Daubechies wavelet function as the basis function for the wavelet transform.

From a mathematical perspective, wavelet transform is performed by convolving a signal with a wavelet function. The Daubechies wavelet transform is defined by an integer N and a set of wavelet filter coefficients $\{a_i, i = 0, 1, 2, \ldots, N-1\}$ through a two-scale relation as follows:

$$\begin{aligned}\phi(x) &= \sum_{l=0}^{N-1} a_l \phi(2x-l),\\ \psi(x) &= \sum_{l=0}^{M-1} b_l \psi(2x-l).\end{aligned} \tag{1}$$

In the equations, $\phi(x)$ represents the scaling function, and $(a_0, a_1, \ldots, a_{N-1})$ are the scaling coefficients. $\psi(x)$ is the wavelet function, and $(b_0, b_1, \ldots, b_{M-1})$ are the wavelet coefficients. The conditions that the coefficients a_l need to satisfy can be found in [44]. We will have N orthonormality conditions if we do a Daubechies wavelet transform with order N. So, it is easy for us to get a_l from solving the $A(w)$. $A(w)$ can be written as the following equation for the guarantee of orthonormality [45]:

$$A(w) = C \cdot \left(\frac{1+e^{-iw}}{2}\right)^2 \cdot \left(e^{-iw} - (2-\sqrt{3})\right), \tag{2}$$

where C is a constant value. From solving equation (2), we can get our coefficient such as four orders as follows:

$$\begin{aligned}h(0) &= \frac{1+\sqrt{3}}{4\sqrt{2}} = 0.4830,\\ h(1) &= \frac{3+\sqrt{3}}{4\sqrt{2}} = 0.8356,\\ h(2) &= \frac{3-\sqrt{3}}{4\sqrt{2}} = 0.2242,\\ h(4) &= \frac{1-\sqrt{3}}{4\sqrt{2}} = -0.1294.\end{aligned} \tag{3}$$

3.3. Residual Feature. In the shallow-frequency domain feature extraction branch introduced in the previous section, we utilized the Daubechies wavelet transform extractor directly applied to the original input image. In this section, we will take the residual image as the basic input. After obtaining the residual map through the filter, we will use the convolution-styled deep frequency domain feature extraction method to obtain the deep-frequency domain information of the image. The deep- and shallow-frequency domain features we mentioned here are different from the high and low frequencies distinguished by the range of frequency bands. We distinguish them based on the layer of hidden space feature position and the feature space from which we extract them.

It is not difficult to understand that the frequency features extracted from the original image combined with color reflect the details of the face image, and the deep frequency-based cues extracted from residuals reflect the edge information of the face image. In face forgery, the mainstream methods can be concluded as extracting a source facial image, transferring the central area of the face to the target image, making the target image look like the source one, and

eliminating the mixed boundary by using various image mixing postprocessing algorithms or the generation-adversarial manner of GAN network. However, from a higher dimensional perspective, the mixed boundary cannot be completely eliminated. Therefore, in this branch, we extract the residual map reflecting the information of the mixed boundary and then map the original narrow channel information to the high-dimensional space according to the residual map. We use a neural network with a pure convolution layer to obtain deep-frequency domain features for the next stage of adaptive fusion.

In detailed implementation steps, we use the filter from [46] to capture different types of dependencies among neighboring pixels. The advantage of the residual image is that the image content is largely suppressed, allowing a more compact and robust description. We form it as noise residuals as follows:

$$R_{ij} = \widehat{X}_{ij}\left(N_{ij}\right) - cX_{ij}, \tag{4}$$

where X_{ij} is the value of the current pixel we calculate, and N_{ij} is the local neighborhood of pixel X_{ij}. c is the dynamic value according to the filter type. $\widehat{X}_{ij}$ is the prediction value of X which is the neighbor domain of X_{ij}. We chose three different types of filters to get the residual image, as shown in Figure 4. First, the image is convolved with Filter 1 to capture the key edge features. Second, it is then convolved with Filter 2 to further enhance the texture. Next, it is convolved with a second-order horizontal filter, Filter 3, to obtain the residual values that capture fine-grained details. These three filters have been carefully selected to be the most effective combination for our face tampering detection task.

Lastly, we truncate the calculated residual value in order to curb the residual range and quantization to make the residual more sensitive to the edge or discontinuity in the image. The quantification coefficient of the three filters is 4, 12, and 2 separately. We consider them according to the central filter kernel's values. We truncate the values using HardTanh to curb the residuals between 0 and 1. The expression of HardTanh is as follows:

$$\text{HardTanh}(x) = \begin{cases} +3, & \text{if } x > +3, \\ -3, & \text{if } x < -3, \\ x, & \text{otherwise.} \end{cases} \tag{5}$$

3.4. Adaptive Fusion. After the previous steps, our method extracts two types of features: the composite frequency domain features obtained by combining the Daubechies wavelet domain features and the residual frequency domain features, as well as the color features of the image. In the fusion stage, to prevent one feature from overpowering and interfering with another feature, we do not simply add, multiply, or concatenate the output features of each branch. Instead, we use a fusion network to achieve adaptive feature fusion. In the next section, the experimental analysis will demonstrate that the fusion network proposed in this section outperforms other simple feature fusion methods.

As shown in Figure 5, the feature fusion module has two inputs: the composite frequency domain feature map $F_{\text{comp}} \in R^{H\times W\times 6}$ and the color feature map $F_{\text{rgb}} \in R^{H\times W\times C}$. $H \times W$ represents the spatial dimensions, and C represents the number of channels. First, the composite feature F_{comp} undergoes a 3×3 convolution to transform it into a new composite feature map with dimensions $H\times W\times C$. Then, one branch undergoes another 3×3 convolution to obtain the feature map R_c, which is then added to the color feature F_{rgb}, resulting in the first fusion feature map F_{rc} of the dual branch. F_{rc} undergoes gated convolution and activation function to obtain the adaptive map F_{ad} with dimensions $H\times W\times 2C$. Next, the adaptive map F_{ad} is pixelwise multiplied with the composite feature F_{comp}, which has undergone gated convolution transformation, resulting in the mixed feature map F_{mixture} with class attention-guided mechanism, with dimensions $H\times W\times 2C$. Finally, the mixed feature map F_{mixture} is added to the composite feature F_{comp} to obtain the output feature F_{out}, with dimensions $H\times W\times 2C$. This output feature is then fed into the backbone network for end-to-end training.

This method utilizes gated convolution to ensure the adaptive nature of the fusion module. In conventional convolution operations, each pixel is treated equally. However, in gated convolution, a learnable dynamic feature selection mechanism is applied to all spatial positions in the feature map. In deepfake detection for faces, the generated face images undergo various preprocessing steps, such as face image region extraction [47] and clipping. As a result, the face region occupies the majority of the pixel area in the image (over 90%). In addition, visual artifacts are often fixed to key neighboring pixels, both in the spatial and frequency domain. Therefore, this approach employs gated convolution to capture and localize these key pixels instead of treating the entire feature map equally. Furthermore, F_{comp} and F_{rgb} are used to localize the most discriminative regions within the face.

In gated convolution, the specific implementation is as follows. First, the C input channels are mapped to a feature latent space of dimension $2C$. Half of the dimensions are used as gates, while the other half are used as features. Sigmoid and ReLU activation functions are applied to the gates and features, respectively, to constrain their outputs. Finally, a pointwise matrix multiplication is performed. The implementation is shown in the following equation:

$$\text{output}_{y,x} = \text{Sigmoid}\left(\sum\sum W_{\text{gates}}\right) \cdot \text{ReLU}\left(\sum\sum W_{\text{features}}\right). \tag{6}$$

Through gated convolution, the method implements a dynamic feature selection mechanism similar to attention in the convolutional layers.

4. Experiments

In this section, we first introduce the overall experimental settings and then present adequate experimental results to demonstrate the superiority of our approach.

0	0	0	0	0
0	−1	2	−1	0
0	2	4	2	0
0	−1	2	−1	0
0	0	0	0	0

Filter 1

−1	2	−2	2	−1
2	−6	8	−6	2
−2	8	−12	8	−2
2	−6	8	−6	2
−1	2	−2	2	−1

Filter 2

0	0	0	0	0
0	0	0	0	0
0	1	−2	1	0
0	0	0	0	0
0	0	0	0	0

Filter 3

Figure 4: The filters we chose for getting residual images.

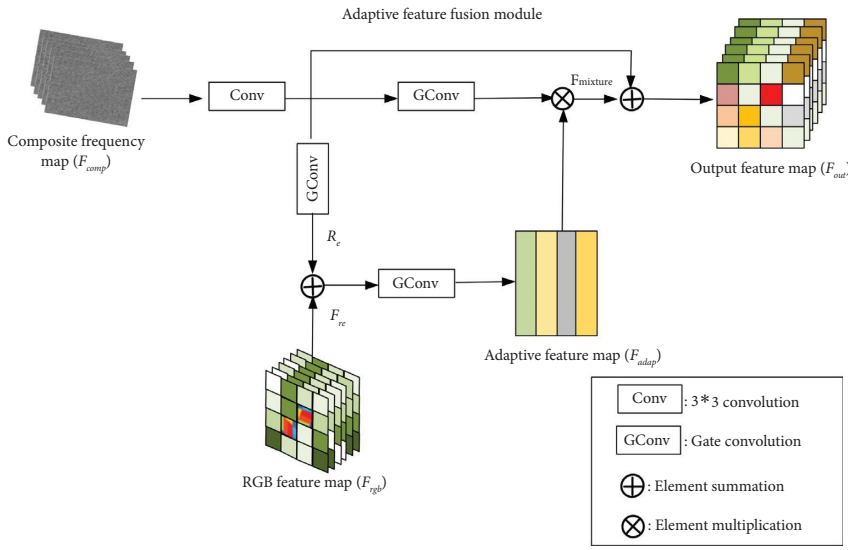

Figure 5: The adaptive feature fusion module. We use gate convolution to treat the convolution layer as an "adaptive processor" [41].

4.1. Datasets. Similar to most other methods of deepfake detection, we conduct our experiments on the three benchmark public deepfake datasets: FaceForensics++ (FF++) [17], Celeb-DF [48], UADFA [49], and DeepFake Detection (DFD) [50]. All of them are large-scale datasets and contain pristine and manipulated videos of human faces.

4.1.1. FaceForensics++. Faceforensics++ [17] is a large scale and widely used public facial forgery database that consists of 1000 real portrait videos and 1000 manipulated videos for each manipulation type. Most real portrait videos are collected from YouTube with the consent of the subjects. Each real video is manipulated by four manipulation methods, including DeepFake, FaceSwap, Face2Face, and NeuralTexture. Output videos are developed with three quality levels: raw, C23, and C40, corresponding to raw quality, low-compression (high resolution) quality, and high-compression (low resolution) quality, respectively. We followed the previous work to partition the database to compare it with other methods. For 1000 videos in each subdatabase, we used 720 videos for training, 140 videos for validation, and 140 videos for testing. We sampled 200–400 frames from each training video and 100 frames from each validation and testing video. Our performance report was implemented on C23 and C40 video quality.

4.1.2. Celeb-DF. Celeb-DF [48] is a new large scale and challenging deepfake detection video database. The Celeb-DF database aims to generate fake videos of better visual quality compared with the previous database. This database contains 590 real videos extracted from YouTube, with a variety of diversity. These videos exhibit an extensive range of aspects, such as face sizes, lighting conditions, and backgrounds. As for fake videos, a total of 5639 videos are created swapping faces using DeepFake technology. The final videos are in MPEG4.0 format. We followed the set in previous work to partition the database and when constructing the database, we ensured that the ratio of real and forged categories is close to 1:1. We use the testing set provided by the database itself, and we randomly selected 15% of the videos as a validation set, with the remaining 85% for training.

4.1.3. UADFA. The UADFA dataset [49] was released in 2018. It consists of 49 real videos and 49 manipulated videos, each with a duration of approximately 11 s and a resolution

of 294×500 pixels. It is the first publicly available dataset in the field of deepfake detection, although it has a relatively small scale. In this study, 35 videos from each category were selected as the training set, 7 videos as the validation set, and 7 videos as the test set. Each video was randomly sampled to extract 200 frames for analysis.

4.1.4. DFD. The DFD dataset [50] is a widely used benchmark for evaluating deepfake detection algorithms. It was released in 2019 by researchers from the University of Albany and the University of Southern California. The dataset contains over 100,000 video clips, including both real and deepfake samples. The deepfake videos were generated using state-of-the-art techniques such as FaceSwap and Face2Face, covering a diverse range of subjects, poses, and scenarios. The dataset is designed to be challenging, with subtle manipulations that can be difficult to detect, making it a valuable resource for developing and assessing the robustness of deepfake detection models.

4.2. Implementation Details. Our model is trained in a supervised manner, and it is an end-to-end model. We train our model by minimizing the dual classification cross-entropy loss function as follows:

$$L = \frac{1}{N}\sum_{i}^{N} -\left[y_i \cdot \log\left(p_i\right) + \left(1 - y_i\right) \cdot \log\left(1 - p_i\right)\right], \quad (7)$$

where L represents the loss value, y_i represents the true label (0 or 1), p_i represents the predicted probability by the model, and N represents the number of classes for classification (which is 2 in this case).

In our experiment preparation, we use the facial identification tool MTCNN [51] to extract the facial area from videos frame by frame and choose the image that has the facial information with appropriate and consistent size with hardly any occlusion. We slightly enlarged the face identification boxing about 1.2 times to ensure that the whole face is included in the cut input because our forged contour also includes hair or ears. The images are aligned and resized to 256×256 pixels. The batch size is set to 64. We use the Adam [52] optimization algorithm for optimizing with the learning rate fixed to 0.0001. We trained our network for 32 epochs. We applied various data augmentation techniques for our datasets. Examples of data augmentation are shown in Figure 6, including the original image, horizontal flip, vertical flip, brightness change, hue change, and saturation change. In the data normalization stage, we normalized the distribution with a mean and standard deviation of 0.5. We apply ACC and AUC (area under the ROC curve) as our evaluation metrics to do a complete analysis of the experimental results. ACC can help us to analyze the classification ability of our model and AUC to comprehensively measure the method effect of the classification threshold.

4.3. Experimental Results

4.3.1. In-Dataset Evaluation. We conducted extensive experiments on the FaceForensics++ dataset to evaluate the performance of our method. In this experiment, we trained the network using the training set of the FaceForensics++ dataset and selected the model that achieved the highest accuracy score on the validation set for testing. The experiment was conducted on three groups: the original quality (C0), low compression quality (C23), and heavy compression quality (C40). The generalization ability of the model was evaluated on multiple datasets, including the ability to detect different types of manipulations and the ability to generalize across different datasets.

Tables 1 and 2 present the ACC and AUC results for the detection on the FaceForensics++ dataset [11, 55]. These are classic methods in the field of image forgery detection. In both original quality and low-quality datasets, our method shows a significant improvement in accuracy compared to these two classic methods. In terms of overall performance, our method outperforms the comparison methods significantly on both ACC and AUC evaluation metrics. However, it is worth noting that we have also identified some challenges when detecting DeepFake, a type of facial identity replacement forgery. Nevertheless, our detection performance still ranks within the top 4. This may be attributed to the fact that DeepFake forgery techniques typically involve advanced image generation and synthesis methods to achieve identity replacement, so detection models need to recognize finer details and inconsistencies.

For images of original quality, machine learning approaches that extract intrinsic attribute features from the images generally exhibit better performance, achieving accuracy rates of over 98% for detection in all four manipulation types. Our proposed method achieves the highest ACC accuracy on the Face2Face and FaceSwap datasets, as well as the highest AUC accuracy across all four manipulation datasets. In the low compression scenario, our method achieves accuracies of 97.09% on the FaceSwap datasets, ranking first among all compared methods. However, it is 3.19% lower than Xception on the DeepFake dataset. In the high compression dataset, our proposed method using composite frequency domain features demonstrates significantly better experimental performance on the FaceSwap and NeuralTexture datasets, achieving ACC of 88.43% and 93.38%, respectively, surpassing other methods. The abovementioned experiments demonstrate that the proposed method using composite frequency domain features still maintains good performance on compressed datasets, achieving ACC of over 90% on most compressed datasets.

Furthermore, we evaluated our method on two additional well-known publicly available deepfake detection datasets, Celeb-DF and UADFA, as shown in Table 3. We trained the network using the training sets of the Celeb-DF dataset and UADFA dataset, respectively, and tested the model performance on the corresponding test sets of the two datasets. The data from the table indicate that our method achieved a detection performance of over 90% on both of these datasets. Particularly, on the UADFA dataset, which has a smaller scale and relatively coarse manipulation techniques, the method's ability to detect deepfakes is enhanced due to the larger residual artifacts left behind. This makes it easier for deep learning-based deepfake detection

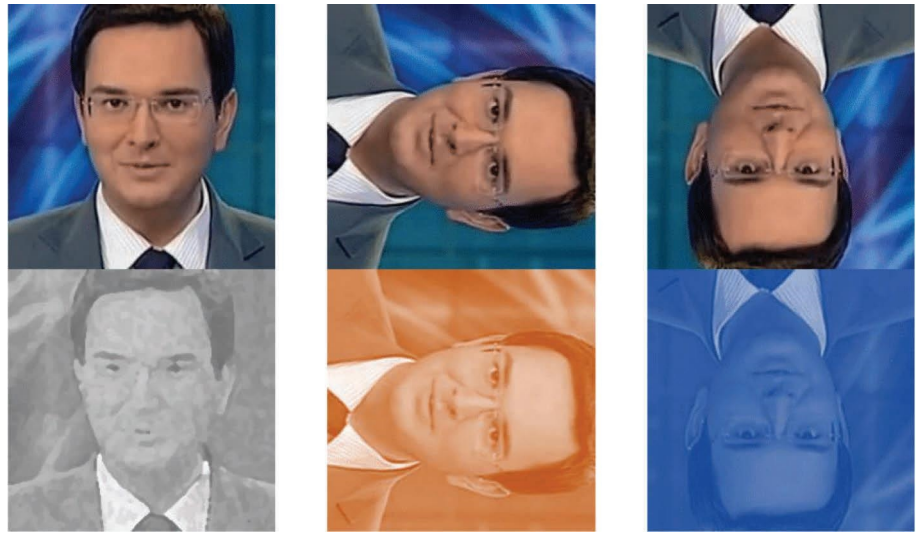

Figure 6: Some of the data augmentation approaches implemented in our experiments.

Table 1: The ACC performance (%) of our method and other classical deepfake detection methods on the FaceForensics++ dataset; the compression intensity is C0, C23, and C40.

Method	C0				C23				C40			
	DP	F2F	FS	NT	DP	F2F	FS	NT	DP	F2F	FS	NT
R^2LD [53]	98.83	98.56	98.89	**99.88**	81.78	85.32	85.69	80.60	68.26	59.38	62.08	62.42
RMSD [46]	99.03	99.13	98.27	**99.88**	77.12	74.68	79.51	76.94	65.58	57.55	60.58	60.69
D-CNN [54]	98.03	98.96	98.94	96.06	82.16	93.48	92.51	75.18	73.25	62.33	67.08	62.59
MesoNet [11]	96.37	97.95	98.17	93.30	89.77	94.25	95.50	78.70	77.68	83.65	79.92	77.74
Xception [55]	98.31	97.75	97.10	96.45	95.15	97.07	95.96	87.99	83.70	**87.21**	83.17	87.90
SupCon [56]	**99.11**	96.57	81.57	98.22	**95.46**	96.17	73.56	90.38	86.02	65.81	52.63	90.38
ADD [35]	94.09	93.22	92.39	92.39	93.62	93.63	94.02	92.08	**89.51**	80.75	84.83	92.35
ETD [57]	98.89	97.56	98.86	94.97	94.99	**97.24**	96.70	**93.21**	88.46	80.97	84.50	80.56
Ours	98.80	**99.16**	**98.94**	96.96	91.96	97.21	**97.09**	92.17	81.51	76.18	**88.43**	**93.38**

Note: The bold values mean the best results.

Table 2: The AUC performance (%) of our method and other classical deepfake detection methods on the FaceForensics++ dataset; the compression intensity is C0, C23, and C40.

Method	C0				C23				C40			
	DP	F2F	FS	NT	DP	F2F	FS	NT	DP	F2F	FS	NT
MesoNet [11]	98.72	98.60	99.37	96.87	95.79	97.57	98.97	86.56	85.32	92.15	87.46	85.72
MesoInception [11]	98.82	98.90	99.48	97.14	94.07	96.87	98.30	83.25	83.75	90.47	94.03	83.03
Xception [55]	98.98	99.15	98.04	97.49	97.77	98.92	99.27	92.52	90.93	**95.01**	94.13	95.21
SupCon [56]	99.27	99.34	86.50	**99.59**	**98.85**	99.12	78.42	97.59	92.69	71.81	53.52	95.93
ADD [35]	98.28	98.13	97.58	94.96	97.80	97.35	98.03	94.70	**96.15**	88.44	84.83	97.91
Ours	**99.47**	**99.79**	**99.67**	99.41	97.62	**99.27**	**99.40**	**97.73**	89.72	84.61	**95.16**	**98.30**

Note: The bold values mean the best results.

Table 3: The ACC and AUC performance (%) of our method on the Celeb-DF and UADFA datasets.

Method	Celeb-DF		UADFA	
	ACC	AUC	ACC	AUC
Ours	96.33	99.25	100.00	99.99

algorithms to identify them. As for Celeb-DF, which is a high-resolution dataset generated using the DeepFake technique, our method also demonstrates good detection performance despite being a dataset generated using a single manipulation technique.

4.3.2. Cross-Dataset Evaluation. Given the rapid advancements in forgery generation techniques, it is crucial to develop a detector that can accurately classify samples from novel and previously unseen manipulation methods. To simulate this scenario, we aim to train a detector using a set

Table 4: Cross-dataset generalization. The ACC performance (%) on the Celeb-DF and DFD datasets when trained on the FF++ dataset.

Training dataset	Method	Testing dataset	
		Celeb-DF	DFD
FF++	Xception [55]	79.54	66.81
	F3Net [7]	77.69	88.36
	ETD [57]	80.66	85.18
	SBI [58]	63.77	72.04
	MADD [59]	82.68	86.56
	QAD [60]	81.93	85.10
	RECCE [61]	80.12	87.98
	Ours	**82.70**	**89.35**

Note: The bold values mean the best results.

Table 5: Ablation study results.

Method	ACC	AUC
ResNet-34	90.59	94.98
ResNet-34 + shallow frequency cues	77.25	83.61
ResNet-34 + RGB cues + shallow frequency cues	92.58	96.62
ResNet-34 + deep frequency cues	62.50	66.74
ResNet-34 + RGB cues + deep frequency cues	89.35	93.96
Ours	**93.27**	**98.91**

Note: The bold values mean the best results.

Table 6: Experimental results for the effect of gated convolution within the adaptive fusion module.

Method	ACC	AUC
w/o gated convolution	97.11	99.04
w/gated convolution (ours)	**97.21**	**99.27**

Note: The bold values mean the best results.

of known manipulated samples and evaluate its performance on unfamiliar and unknown manipulation techniques. To evaluate cross-dataset generalization, we trained on FF++ (all four methods) and tested on Celeb-DF and DFD. The experimental results are shown in Table 4. Our approach achieved the best generalization performance on both the Celeb-DF and DFD datasets, attaining accuracy rates of 82.70% and 89.35%, respectively. These results demonstrate that our method consistently outperforms other approaches, showcasing its superiority in terms of accuracy, which provides promise for its performance on future improved forgeries.

4.3.3. Ablation Study. In order to assess the contribution of each module to the final experimental results, we incrementally added and removed our branch model blocks. The experimental results are shown in Table 5. We use ResNet-34 as our backbone. Although the single shallow frequency cues and residual cues suffer from a sharp decline, when they are combined with the RGB cues branch, they can get a better performance than the convention ResNet-34. Finally, after fetching them together, we can get our best model, which has an ACC of 93.27% and an AUC of 98.91%. The ablation experiments have substantiated the effectiveness of our approach. While individual shallow-level frequency domain features (wavelet frequency domain features) and deep-level frequency domain features (residual frequency domain features) may experience a significant decline in performance, they still retain valuable information. When combined with RGB features, they complement each other, enhancing the model's robustness. Simultaneously, RGB features to capture color and texture information in the images, while shallow-level frequency domain features and deep-level frequency domain features provide crucial insights into image structure and variations. The integration of RGB features, shallow-level frequency domain features, and deep-level frequency domain features effectively introduces multiple sources of features, with each source offering unique insights into different aspects of forged images. This diversity of features enables a more comprehensive representation of image information, contributing to improved detection performance.

In addition, to assess the efficacy of the gated convolution within the adaptive fusion module, we conducted experiments on the F2F (C23) dataset, evaluating the detection results both with and without the gated convolution. As shown in Table 6, the inclusion of the gated convolution led to measurable improvements in the experimental outcomes, with the ACC and AUC metrics increasing by 0.10% and 0.23%, respectively. These results effectively demonstrate the effectiveness of the gated convolution component.

5. Conclusion

In conclusion, the proliferation of deepfake technologies poses a significant threat to the integrity and trustworthiness of digital media. In this work, we have proposed a novel deep learning-based approach, the CFNSFD features, to address the critical challenge of deepfake detection. By comprehensively extracting color features in the spatial domain as well as shallow- and deep-frequency domain features and then adaptively fusing these complementary representations, our method is able to capture a rich set of forgery cues that are highly discriminative for distinguishing fake content. The superior performance of our method, demonstrated through extensive experiments on benchmark datasets, underscores the importance of leveraging both spatial and frequency-based information for robust deepfake detection. As deepfake generation techniques continue to evolve, maintaining the trustworthiness of digital media will remain a crucial societal and technological imperative. Our work represents an important step forward in this direction, providing a strong foundation for future research on generalized and reliable deepfake detection.

Data Availability Statement

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

Conflicts of Interest

The authors declare no conflicts of interest.

Funding

This work was supported by the National Natural Science Foundation of China (nos. 62076052 and 62106037) and the Application Fundamental Research Project of Liaoning Province (2022JH2/101300262).

References

[1] H. Zhang, B. Chen, J. Wang, and G. Zhao, "A Local Perturbation Generation Method for gan-generated Face Anti-forensics," *IEEE Transactions on Circuits and Systems for Video Technology* 33, no. 2 (2023): 661–676, https://doi.org/10.1109/tcsvt.2022.3207310.

[2] S. Dong, B. Chen, K. Ma, and G. Zhao, "Active Defense Against Voice Conversion through Generative Adversarial Network," *IEEE Signal Processing Letters* 31 (2024): 706–710, https://doi.org/10.1109/lsp.2024.3365034.

[3] Y. Zhang, L. Zheng, and V. L. L. Thing, "Automated Face Swapping and its Detection," in *IEEE 2nd international conference on signal and image processing (ICSIP)* (IEEE, August 2017), 15–19.

[4] M. Koopman, A. M. Rodriguez, and Z. Geradts, "Detection of Deepfake Video Manipulation," in *The 20th Irish Machine Vision and Image Processing Conference (IMVIP)* (August 2018), 133–136.

[5] A. Khodabakhsh, R. Ramachandra, K. Raja, et al., "Fake Face Detection Methods: Can They Be Generalized?" in *International Conference of the Biometrics Special Interest Group (BIOSIG)* (IEEE, September 2018), 1–6.

[6] L. Li, J. Bao, T. Zhang, et al., "Face X-Ray for More General Face Forgery Detection," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (June 2020), 5000–5009.

[7] Y. Qian, G. Yin, L. Sheng, et al., "Thinking in Frequency: Face Forgery Detection by Mining Frequency-Aware Clues," (2020), https://arxiv.org/abs/2007.09355.

[8] S. Dong, J. Wang, R. Ji, J. Liang, H. Fan, and G. Zheng, "Implicit Identity Leakage: The Stumbling Block to Improving Deepfake Detection Generalization," (2023), 3994–4004, https://arxiv.org/abs/2210.14457.

[9] B. Huang, Z. Wang, J. Yang, et al., "Implicit Identity Driven Deepfake Face Swapping Detection," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (June 2023), 4490–4499.

[10] Z. Liu, X. Qi, J. Jia, and P. H. S. Torr, "Global Texture Enhancement for Fake Face Detection in the Wild," (2020), 8057–8066, https://arxiv.org/abs/2002.00133.

[11] D. Afchar, V. Nozick, J. Yamagishi, et al., "Mesonet: A Compact Facial Video Forgery Detection Network," (2018), 1–7, https://arxiv.org/abs/1809.00888.

[12] H. H. Nguyen, F. Fang, J. Yamagishi, et al., "Multi-Task Learning for Detecting and Segmenting Manipulated Facial Images and Videos," (2019), 1–8, https://arxiv.org/abs/1906.06876.

[13] H. Dang, F. Liu, J. Stehouwer, et al., "On the Detection of Digital Face Manipulation," (2020), 5781–5790, https://arxiv.org/abs/1910.01717.

[14] Y. Li, M. C. Chang, and S. Lyu, "In Ictu Oculi: Exposing Ai Created Fake Videos by Detecting Eye Blinking," in *2018 IEEE International workshop on information forensics and security (WIFS)* (December 2018), 1–7.

[15] X. Yang, Y. Li, and S. Lyu, "Exposing Deep Fakes Using Inconsistent Head Poses," (2019), 8261–8265, https://arxiv.org/abs/1811.00661.

[16] S. Agarwal, H. Farid, O. Fried, et al., "Detecting Deep-Fake Videos from Phoneme-Viseme Mismatches," in *IEEE/CVF conference on computer vision and pattern recognition workshops* (June 2020), 660–661.

[17] A. Rossler, D. Cozzolino, L. Verdoliva, et al., "Faceforensics++: Learning to Detect Manipulated Facial Images," (2019), 1–11, https://arxiv.org/abs/1901.08971.

[18] F. Matern, C. Riess, and M. Stamminger, "Exploiting Visual Artifacts to Expose Deepfakes and Face Manipulations," in *IEEE Winter Applications of Computer Vision Workshops (WACVW)* (IEEE, January 2019), 83–92.

[19] H. Khalid and S. S. Woo, "OC-FakeDect: Classifying Deepfakes Using One-Class Variational Autoencoder," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)* (June 2020), 2794–2803.

[20] X. T. Zhu, Y. Q. Tang, and P. Z. Geng, "Detection Algorithm of Tamper and Deepfake Image Based on Feature Fusion," *Netinfo Security* 21, no. 8 (2021): 70–81.

[21] D. A. Coccomini, N. Messina, C. Gennaro, et al., "Combining Efficientnet and Vision Transformers for Video Deepfake Detection," (2022), 219–229, https://arxiv.org/abs/2107.02612.

[22] Z. Wang, J. Bao, W. Zhou, et al., "Altfreezing for More General Video Face Forgery Detection," (2023), 4129–4138, https://arxiv.org/abs/2307.08317.

[23] L. Chen, Y. Zhang, Y. Song, et al., "Self-supervised Learning of Adversarial Example: Towards Good Generalizations for Deepfake Detection," (2022), 18710–18719, https://arxiv.org/abs/2203.12208.

[24] J. Zhang, J. Ni, F. Nie, and J. Huang, "Domain-invariant and Patch-Discriminative Feature Learning for General Deepfake Detection," *ACM Transactions on Multimedia Computing, Communications, and Applications* (2024): https://doi.org/10.1145/3657297.

[25] N. Yu, L. Davis, and M. Fritz, "Attributing Fake Images to GANs: Learning and Analyzing GAN Fingerprints," (2019), 7555–7565, https://arxiv.org/abs/1811.08180.

[26] A. Sarlashkar, M. Bodruzzaman, and M. Malkani, "Feature Extraction Using Wavelet Transform for Neural Network Based Image Classification," *Proceedings of Thirtieth Southeastern Symposium on System Theory* (1998): 412–416, https://doi.org/10.1109/ssst.1998.660107.

[27] J. A. Stuchi, M. A. Angeloni, R. F. Pereira, et al., "Improving Image Classification with Frequency Domain Layers for Feature Extraction," in *IEEE 27th International Workshop on Machine Learning for Signal Processing (MLSP)* (September 2017), 1–6, https://doi.org/10.1109/mlsp.2017.8168168.

[28] H. Huang, R. He, Z. Sun, and T. Tan, "Wavelet-SRNet: A Wavelet-Based CNN for Multi-Scale Face Super Resolution," in *IEEE International Conference on Computer Vision (ICCV)* (October 2017), 1698–1706.

[29] R. Durall, M. Keuper, F. J. Pfreundt, et al., "Unmasking Deepfakes with Simple Features," (2019), https://arxiv.org/abs/1911.00686.

[30] H. Liu, X. Li, W. Zhou, et al., "Spatial-Phase Shallow Learning: Rethinking Face Forgery Detection in Frequency Domain," (2021), 772–781, https://arxiv.org/abs/2103.01856, https://doi.org/10.1109/cvpr46437.2021.00083.

[31] Y. Luo, Y. Zhang, J. Yan, and W. Liu, "Generalizing Face Forgery Detection with High-Frequency Features," (2021), 16312–16321, https://arxiv.org/abs/2103.01856.

[32] I. Masi, A. Killekar, R. M. Mascarenhas, et al., "Two-branch Recurrent Network for Isolating Deepfakes in Videos," (2020), 667–684, https://arxiv.org/abs/2008.03412.

[33] Y. Luo, Y. Zhang, J. Yan, et al., "Generalizing Face Forgery Detection with High-Frequency Features," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (June 2021), 16317–16326.

[34] C. Tan, Y. Zhao, S. Wei, et al., "Rethinking the Up-Sampling Operations in Cnn-Based Generative Network for Generalizable Deepfake Detection," (2024), 28130–28139, https://arxiv.org/abs/2312.10461.

[35] L. M. Binh and S. Woo, "ADD: Frequency Attention and Multi-View Based Knowledge Distillation to Detect Low-Quality Compressed Deepfake Images," *Proceedings of the AAAI Conference on Artificial Intelligence* 36, no. 1 (2022): 122–130, https://doi.org/10.1609/aaai.v36i1.19886.

[36] Y. Jeong, D. Kim, S. Min, et al., "Bihpf: Bilateral High-Pass Filters for Robust Deepfake Detection," (2022), 48–57, https://arxiv.org/abs/2109.00911.

[37] Y. Jeong, D. Kim, Y. Ro, and J. Choi, "FrePGAN: Robust Deepfake Detection Using Frequency-Level Perturbations," *Proceedings of the AAAI Conference on Artificial Intelligence* 36, no. 1 (2022): 1060–1068, https://doi.org/10.1609/aaai.v36i1.19990.

[38] C. Tan, Y. Zhao, S. Wei, G. Gu, P. Liu, and Y. Wei, "Frequency-Aware Deepfake Detection: Improving Generalizability through Frequency Space Domain Learning," *Proceedings of the AAAI Conference on Artificial Intelligence* 38, no. 5 (2024): 5052–5060, https://doi.org/10.1609/aaai.v38i5.28310.

[39] C. T. Doloriel and N. M. Cheung, "Frequency Masking for Universal Deepfake Detection," (2024), 13466–13470, https://arxiv.org/abs/2401.06506.

[40] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (June 2016).

[41] J. Yu, Z. L. Lin, J. Yang, X. Shen, X. Lu, and T. S. Huang, "Free-Form Image Inpainting with Gated Convolution," (2019), 4470–4479, https://arxiv.org/abs/1806.03589.

[42] C. Vonesch, T. Blu, and M. Unser, "Generalized Daubechies Wavelet Families," *IEEE Transactions on Signal Processing* 55, no. 9 (2007): 4415–4429, https://doi.org/10.1109/tsp.2007.896255.

[43] L. E. Ratcliff, W. Dawson, G. Fisicaro, et al., "Flexibilities of Wavelets as a Computational Basis Set for Large-Scale Electronic Structure Calculations," *The Journal of Chemical Physics* 152, no. 19 (2020): 194110, https://doi.org/10.1063/5.0004792.

[44] M. Q. Chen, C. Hwang, and Y. P. Shih, "The Computation of wavelet-Galerkin Approximation on a Bounded Interval," *International Journal for Numerical Methods in Engineering* 39, no. 17 (1996): 2921–2944.

[45] A. Boggess and F. J. Narcowich, *A First Course in Wavelets with Fourier Analysis* (John Wiley & Sons, 2015).

[46] J. J. Fridrich and J. Kodovský, "Rich Models for Steganalysis of Digital Images," *IEEE Transactions on Information Forensics and Security* 7, no. 3 (2012): 868–882, https://doi.org/10.1109/tifs.2012.2190402.

[47] S. Chaichulee, M. Villarroel, J. Jorge, et al., "Multi-Task Convolutional Neural Network for Patient Detection and Skin Segmentation in Continuous Non-Contact Vital Sign Monitoring," in *12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)* (May 2017), 266–272, https://doi.org/10.1109/fg.2017.41.

[48] Y. Li, X. Yang, P. Sun, et al., "Celeb-df: A Large-Scale Challenging Dataset for Deepfake Forensics," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (June 2020), 3207–3216.

[49] X. Yang, Y. Li, and S. Lyu, "Exposing Deep Fakes Using Inconsistent Head Poses," (2019), 8261–8265, https://arxiv.org/abs/1811.00661.

[50] Google Ai Blog, "Contributing Data to Deepfake Detection Research," (2024), https://ai.googleblog.com/2019/09/contributing-data-to-deepfake-detection.html.

[51] J. Xiang and G. Zhu, "Joint Face Detection and Facial Expression Recognition with MTCNN," in *4th International Conference on Information Science and Control Engineering (ICISCE)* (July 2017), 424–427.

[52] D. Kingma and J. Ba. Adam, "A Method for Stochastic Optimization," *Computer Science* (2014).

[53] D. Cozzolino, G. Poggi, and L. Verdoliva, "Recasting Residual-Based Local Descriptors as Convolutional Neural Networks: an Application to Image Forgery Detection," (2017), 159–164, https://arxiv.org/abs/1703.04615.

[54] N. Rahmouni, V. Nozick, J. Yamagishi, et al., "Distinguishing Computer Graphics from Natural Images Using Convolution Neural Networks," in *IEEE Workshop on Information Forensics and Security (WIFS)* (IEEE, December 2017), 1–6.

[55] F. Chollet, "Xception: Deep Learning with Depthwise Separable Convolutions," (2017), 1251–1258, https://arxiv.org/abs/1610.02357.

[56] Y. Xu, K. Raja, and M. Pedersen, "Supervised Contrastive Learning for Generalizable and Explainable Deepfakes Detection," in *IEEE/CVF Winter Conference on Applications of Computer Vision* (January 2022), 379–389.

[57] Z. Ba, Q. Liu, Z. Liu, et al., "Exposing the Deception: Uncovering More Forgery Clues for Deepfake Detection," *Proceedings of the AAAI Conference on Artificial Intelligence* 38, no. 2 (2024): 719–728, https://doi.org/10.1609/aaai.v38i2.27829.

[58] K. Shiohara and T. Yamasaki, "Detecting Deepfakes with Self-Blended Images," (2022), 18720–18729, https://arxiv.org/abs/2204.08376.

[59] H. Zhao, W. Zhou, D. Chen, et al., "Multi-attentional Deepfake Detection," (2021), 2185–2194, https://arxiv.org/abs/2103.02406.

[60] B. M. Le and S. S. Woo, "Quality-agnostic Deepfake Detection with Intra-model Collaborative Learning," (2023), 22378–22389, https://arxiv.org/abs/2309.05911.

[61] J. Cao, C. Ma, T. Yao, et al., "End-to-end Reconstruction-Classification Learning for Face Forgery Detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (June 2022), 4113–4122.