Frontiers in Computing and Intelligent Systems
ISSN: 2832-6024 | Vol. 11, No. 3, 2025

Deepfake Detection Technology Integrating Spatial Domain and Frequency Domain

Haoqi Geng *, Tianliang Lu, Wanxin Huang and Bowen Ding

People's Public Security University of China, Beijing, China

*** Corresponding author:** Haoqi Geng

Abstract: In recent years, with the rapid development of deep forgery technology, its verisimilitude is increasing day by day, and its social impact is becoming more and more serious. However, although a variety of face deep forgery video detection algorithms have been proposed and have shown certain detection capabilities on open source data sets, in the face of increasingly sophisticated deep forgery technology, the differences between genuine and fake videos are gradually difficult to be captured by the naked eye, and existing detection methods generally have problems such as low cross-compressibility detection and poor robustness. Therefore, in order to improve the detection accuracy and model robustness, a deep forged video detection method named SFDT is proposed. In this scheme, the framework structure of fusion of air frequency domain is adopted. Firstly, feature extraction is enhanced by improving MVIT in the airspace and using ASFF adaptive module; secondly, frequency domain features are extracted by dynamic filter in the frequency domain; then FECAM is used to reduce the loss caused by frequency domain information transmission; finally, multi-mode fusion module is used for feature fusion. This air-frequency fusion detection scheme can not only improve the cross-compression detection performance of the model, but also effectively deal with various interference in the transmission of video, and improve the robustness of the model.

Keywords: Deepfake Detection; Spatial Domain; Frequency Domain; Fecam.

1. Introduction

Deepfake technology [1] is one of the key technologies for making fake videos, also known as AI face changing technology. Once this technology is maliciously used by attackers, it will cause unpredictable consequences to the politics and economy of the society. A typical case is that the deep fake videos of the presidents of Russia and Ukraine are maliciously uploaded to social media, thus promoting false narratives against the people and national interests of the target countries and impacting the psychology of the people of the target countries. What is more, the faces of some well-known singers, movie stars and other public figures are "transferred" to porn stars, falsifying porn for illegal profit, which constitutes a serious violation of personal reputation and portrait rights. At present, the mainstream method of forgery is to use machine learning algorithms such as Generative adversarial network (GAN) [2] or Convolutional Neural Network (CNN) through a large number of coding and decoding training. The robustness of the model is enhanced, and finally the facial features of one target individual are successfully "transplanted" to the face of another individual. Since videos are made up of successive frames, simply replacing the faces in each picture can result in a new, very realistic fake video. To be specific, first of all, the video of the imitated object is transformed frame by frame to obtain a large number of photos, then the face of the object is replaced by the face of the imitated object, and finally the fake video resynthesis of the replaced photo is carried out to deceive the human eye and produce wrong cognition.

Although the current mainstream detection methods are largely based on deep learning technology, the inherent limitations of deep learning have become a key factor restricting the progress of deep fake video detection technology. The method can be classified as "feature extraction network + classification network", where in feature extraction network is generally a deep structure and has the inherent defect of low learning priority for high-frequency components, so the utilization rate of high-frequency information is low in the final classification decision. However, due to the special synthesis method of Deepfake technology, discontinuous features will be generated between the synthetic face and the surrounding pixels, and such edge discontinuous features often exist in the high-frequency information in the frequency domain. See Fig. 1 with the development and iteration of Deepfake technology, it is difficult to identify such edge discontinuity features based on the spatial domain. Therefore, it is an effective and feasible method to detect deep forged face video by using spatial information and frequency domain information fusion

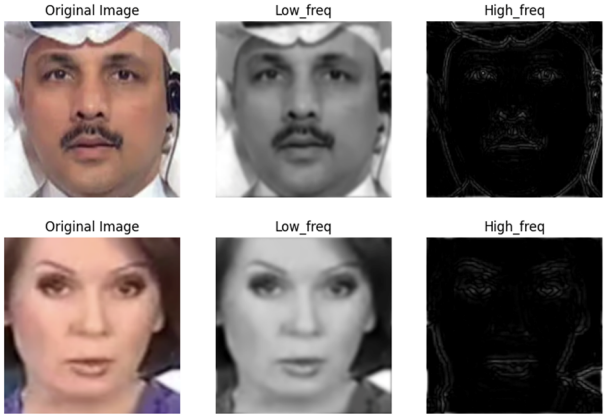

Fig 1. Example of separating low frequency information from high frequency information

2. Related Work

The detection [3] and prevention of deepfake videos has become a key research in the field of network security, and its core task is to quickly and accurately identify the traces of

forgery in videos. For mainstream forged audio and video, traditional detection methods usually use face recognition algorithms, convert facial data into feature vectors, and distinguish the authenticity of the video by machine learning techniques such as Support Vector Machine (SVM)[4].

Several techniques that have been introduced rely on CNNs for a frame-by-frame examination. For example, a method known as MesoNet[5] has been developed, which is a less complex CNN designed to identify synthetic faces. However, a research [6] demonstrated that this network is surpassed in performance by a retrained XceptionNet designed for the same purpose.

Recently, The F3-Net [7]employs a dual-branch framework, with one branch dedicated to identifying forgery patterns through frequency cues, while the other branch focuses on discerning the discrepancies in frequency statistical properties between authentic and counterfeit images.

Other methods leverage the temporal progression of video frames by utilizing Long Short-Term Memory (LSTM) [8]networks for analysis, where frame-specific features are initially extracted and subsequently integrated using a recurrent neural network architecture.

This paper implements a model framework that fuses the spatial-frequency domain to detect local inconsistencies at various scales for forgery identification. Furthermore, we incorporate frequency modality to enhance the resilience of our approach against diverse image compression techniques.

3. Approh

In this paper, we propose a novel deepfake video detection scheme fusing spatial-frequency domain features, named SFDT. The model consists of two key layers: Spatial domain feature extraction layer and frequency domain feature extraction layer. Firstly, several convolution layers are used to extract the features of the input image as M1. The purpose of convolution is mainly to speed up the convergence speed, and then they are sent to the space-frequency domain for feature processing. The spatial feature extraction layer is mainly composed of the backbone network MViT. By embedding the ASFF spatial adaptive module in each Patch output position in the MViT structure, the network can better capture the multi-scale information of the target by dynamically selecting and fusing the multi-layer feature maps. In the frequency domain feature extraction layer, the spatial domain is transformed by two-dimensional FFT, which is combined with a filter to simulate the correlation of different frequency band components. By introducing the FECAM frequency enhancement module, the problem of high frequency noise caused by Gibbs phenomenon in Fast Fourier transform FFT was improved, so as to optimize the feature output. Finally, the CMF feature fusion module was used to fuse M1 and spatial-frequency domain features for discrimination.

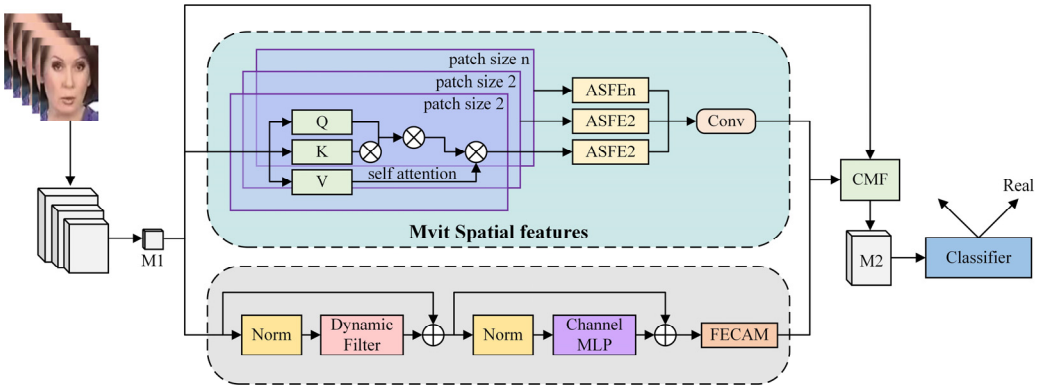

Fig 2. SFDT overall model framework

3.1. Mvit_R

While MVit have shown potential in modeling interactions between tokens, they focus on content rather than structure. Spatial-temporal structure modeling relies only on absolute location embeddings to provide location information. This ignores the fundamental principle of translation invariance of vision. That is, the way MViT models the interaction between two patches changes with their absolute position in the image, even if the relative position does not change. To solve this problem, the Self-Attention module is embedded by inputing the relative position between elements i and j:

$$\mathrm{Atten}(Q,K,V) = \mathrm{Softmax}\left(\frac{QK+E}{\sqrt{d}}\right)V \quad (1)$$

$$E = \mathrm{Q}R_{p(i)p(j)} \quad (2)$$

where p(i) and p(j) represent the spatial-temporal positions of elements i and j.MViT has a larger step size on the K and V tensors than on the Q tensor, which is downsampled only if the resolution of the output sequence changes across stages. This motivates the addition of residual pooling connections to the Q tensor to increase the information flow, driving the training and convergence of attention.

3.2. ASFF

ASFF is an improvement based on the Faster R-CNN framework, which aims to improve the effect of feature fusion, thereby improving the accuracy and performance of object detection[9]. The ASFF network introduces an adaptive structural feature fusion module to fuse multiple feature maps at different levels.Traditional object detection networks usually use only a single level of feature maps, while ASFF network dynamically selects and fuses multi-level feature maps, so that the network can better capture multi-scale information of the target. The ASFF network first extracts feature maps at different levels through the backbone network. Then, it uses an adaptive attention mechanism to perform weighted fusion of these feature maps, so as to better retain important feature information. Finally, the fused features are passed to the object detection head for object classification and localization. The key idea is to adaptively learn the fusion spatial weight of each scale feature map in two steps: identity scaling and adaptive fusion.

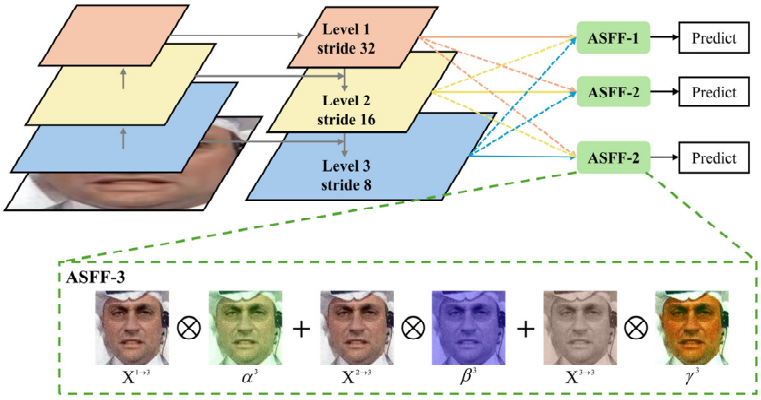

Fig 3. Overall structure of ASFF

Taking ASFF-3 as an example, the green part of the figure describes the process of feature fusion as:

$$\mathbf{y}_{ij}^{3} = \alpha_{ij}^{3}\mathbf{x}_{ij}^{1\to 3} + \beta_{ij}^{3}\mathbf{x}_{ij}^{2\to 3} + \gamma_{ij}^{3}\mathbf{x}_{ij}^{3\to 3} \tag{3}$$

Where i and j represent the spatial position, x is the feature from different levels, and is the feature from different levels multiplied by the weight parameter and added, the new fusion feature ASFF-3 can be obtained. The output adaptive spatial feature is defined as f_s.

3.3. Dynamic Filter

To leverage frequency information, Auto-former [10] uses FFT to efficiently compute autocorrelation functions, FNO [11] is used as an internal block of the network to perform representation learning in the low frequency domain, and DCTnet [12] uses discrete cosine transform to compress information to retain more original image information in CV tasks. Most of these works are based on Fourier transform, which is helpful to extract frequency features. However, most Fourier transform-based methods use the Fourier transform to obtain the frequency information and the inverse Fourier transform to reconstruct the time information to avoid complex training, which introduces a new amount of computation that can be avoided if the time-frequency transform is performed using DCT. More importantly, the implicit periodicity of the DFT causes boundary discontinuities. The high-frequency component that leads to significant is known as the Gibbson phenomenon. After quantization, Gibbs phenomenon causes boundary points to take wrong values. Image classification and downstream tasks Dynamic filters are generated based on global filters. For the 2D discrete Fourier transform, given a 2D signal x(h,w), the 2D-DFT is defined as follows.

$$\tilde{x}(h',w') = \sum_{h=0}^{H-1}\sum_{w=0}^{W-1}\frac{x(h,w)e^{-2\pi j\left(\frac{h}{H}+\frac{w}{W}\right)}}{\sqrt{HW}} \tag{4}$$

Where $\tilde{x}(h',w')$ is the result of the transformation and is usually expressed as a value in the frequency domain of the complex number, h and w are the row and column indices in the spatial domain, H and W represent the number of rows and columns of the image in the spatial domain, j is the imaginary unit. This exponential function is used to associate points in the spatial domain with points in the frequency domain. $\sqrt{HW}$is the normalization factor used to ensure that the energy remains the same before and after the transformation.

The inverse 2D-DFT exists and is known as the two-dimensional inverse discrete Fourier transform F^{-1} (2D-IDFT). Moreover, the frequency domain has an important property that multiplication in the frequency domain is equivalent to cyclic convolution in the original domain, known as the convolution theorem. Define a global filter G for feature X, and the global filter is defined as:

Where G(X) denotes the output filter, $\odot$ signifies matrix point-wise multiplication, K is a learnable filter, and F^{-1} represents the inverse transformation of 2D-FFT. The neural network within the dynamic filter dynamically determines a sufficient number of global filters to be included, and the overall structure can be seen as a dual-stream architecture. First, the input is processed through multiple layers of perception and a softmax function to extract features. Then, a Filter of size N is used to dynamically adjust the weights, which is set to 4 in the experiments of this chapter. Next, high-frequency features are extracted through 2D inverse discrete Fourier transform. Finally, the output is fused by weighted summation and 2D inverse discrete Fourier transform. The definition of the dynamic filter is as follows:

$$D(X) = F^{-1}\left(K_M(X) \odot \mathcal{F} \circ A(X)\right) \tag{5}$$

Where K_M denotes the function that determines the dynamic filter. A is the continuous real map, including point convolution and identity map, $\odot$ denotes matrix dot multiplication, $\circ$ denotes composite map.

3.4. Fecam

In the process of frequency domain feature extraction, the loss of key information is inevitable with the increase of calculation. This is the problem with the model when it comes to capturing frequency information. At present, the mainstream frequency information extraction methods are based on Fourier transform. However, the use of the Fourier transform is problematic due to the Gibbs phenomenon, and the discrete Fourier transform presented in the previous section creates the same problem. If the values on the two sides of the sequence are very different, an oscillatory approximation is observed around the sides and high-frequency noise is introduced. Therefore, this chapter uses the FECAM frequency enhancement module to ameliorate this problem by adaptively modeling the interdependence between frequencies based on the discrete cosine transform, which will essentially avoid the problems caused by high-frequency noise during the Fourier transform.

FECAM [13]works by first applying DCT to the time series data of each channel to convert it into a frequency domain representation. Then, through the channel attention mechanism, the frequency information is weighted, focusing

on those frequency components that contribute most to the prediction task. This enhanced frequency information is then incorporated into the model to generate more accurate predictions. As shown in Figure 4, FECAM divides the input feature map into n groups along the channel dimension. Subsequently, the groups are processed by the corresponding DCT frequency components from low frequency to high frequency, and each individual channel is processed by the same frequency components to obtain the Freq sequence, which is defined as:

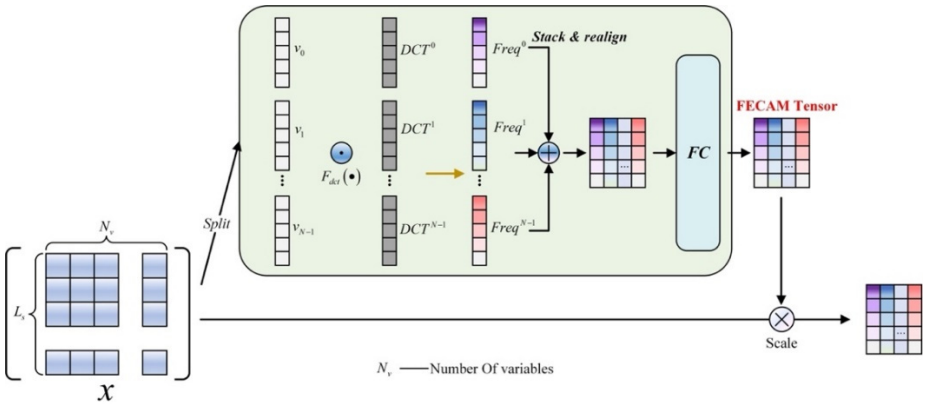

Fig 4. Overall structure of ASFF[13]

$$Freq^i = DCT^i \odot V^i \quad (6)$$

$$Freq = stack([Freq^0, Freq^1, \cdots, Freq^{n-1}]) \quad (7)$$

Where i is the group number from 0 to N-1, DCT is the DCT transform applied, V is the feature map group, and then $Freq$ can be fused by stack stacking.In this way, each channel feature interacts with each frequency component to comprehensively obtain important time domain information from the frequency domain, which will prompt the network to enhance the diversity of extracted features and define the output frequency domain feature as f_{fq}.

4. Result and Conclusion

4.1. Dataset

In order to effectively prove the robustness of the model proposed in this chapter, multiple data sets are selected for testing. The datasets selected for this chapter are FF++, Celeb-DFV2, BioDeepAV, DeepForensic, and FaceShifter datasets[14].

FF++ [15] is a large-scale and diverse fake video dataset, which contains a variety of attack methods, such as deepfake, facial replacement and expression imitation. The dataset covers real videos of different races, genders and ages, as well as synthetic fake videos. In addition, the FF++ dataset provides a variety of scene, lighting, and camera conditions to enhance its generalization ability in practical applications.

Celeb-DFV2 [16] is a high-quality dataset specifically for deepfake detection, containing 590 real videos and 5639 fake videos. These videos are produced using advanced Generative Adversarial Network (GAN) technology and have high visual quality and deception. The Celeb-DFV2 dataset covers a variety of expressions, poses and lighting conditions, which provides rich training and testing samples for deepfake detection tasks.

DeeperForensics [17] represents the largest face forgery detection dataset to date, containing 60,000 videos consisting of 17.6 million frames. More challenging benchmarks that apply a wide range of real-world perturbations to obtain larger scale and higher diversity. All source videos in the dataset are carefully collected and fake videos are generated by the newly proposed end-to-end face exchange framework. As verified by the user study, the quality of the generated videos is better than that in existing datasets.

FaceShifter [18] is a fake video dataset based on face swapping technology, which consists of two networks, AEI-Net and HEAR-Net. AEI-Net produces preliminary face swapping results, while HEAR-Net refines this output. It contains 1000 real videos and 1000 fake videos. These fake videos are generated by face swapping technique and have high visual quality and deception. The FaceShifter dataset covers people of multiple races, genders, and ages, as well as multiple expressions, poses, and lighting conditions, which provides important data support for studying the application of faceshifter technology in forgery video detection.

BioDeepAV [19] is a multimodal benchmark dataset to evaluate the performance of deepfake detectors in the face of unknown generators. Contains over 1600 deepfake videos, and the videos are generated using four state-of-the-art methods specifically for video synthesis. The dataset covers multiple identities and expressions, sampling real videos from HDTF and TalkingHead-1KH datasets, and using face images from sources such as RealVisXL, LAION-Face and HDTF, as well as English dialects, HDTF datasets.

To ensure the effectiveness of the amount of experimental data, 1000 real and fake videos are randomly selected for the above five data sets, and 32 frames are selected at equal intervals. Among them, Celeb-DFV2 selects 590 real videos and divides the training set and the test set with the ratio of 7:3. The training set of each dataset contains 22,400 real video frame images and 22,400 fake video frame images. The test set contains 9600 real video frame images and 9600 fake video frame images, and the experimental data has 320,000 face images.

4.2. Evaluation Metrics and Experimental Setup

The selected evaluation metrics in the experiment are accuracy (ACC) and area under the Curve (AUC) as the key metrics to evaluate the performance of the model, and the specific calculation formulas of these two-evaluation metrics have been elaborated in the third chapter of the report. ACC measures how well the model classifies the positive and negative samples, while AUC measures how well the model classifies the positive and negative samples.

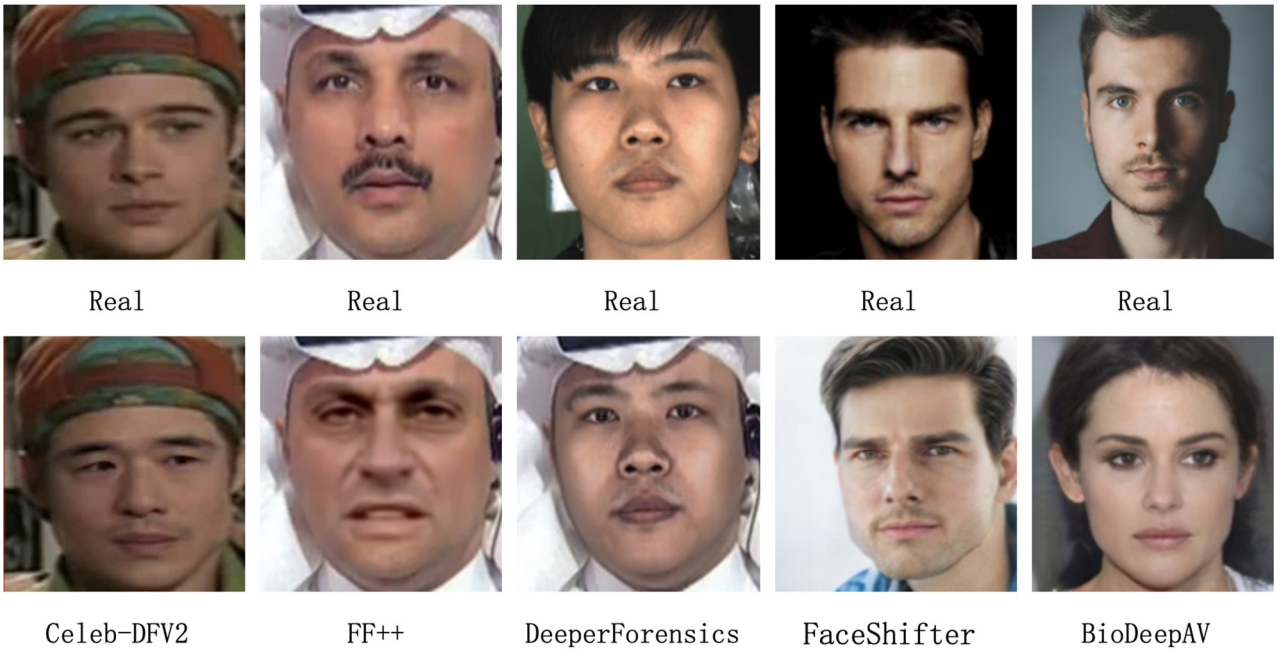

Fig 5. Examples of datasets used for experiments

The Learning Rate determines the step size at which the model updates the weights during the optimization process. An appropriate learning rate can speed up the convergence of the model while avoiding oscillations or overstepping the optimal solution during the training process. In this experiment, we set the learning rate to 0.001, which is a starting value that is often used in practice. The Batch Size defines the number of samples used in each training iteration. A large Batch Size can reduce the variance in the training process and improve the efficiency of memory utilization, but it may reduce the generalization ability of the model. A smaller Batch Size may increase the training noise, but it helps the model capture more feature details. In this experiment, we chose 128 as the value of Batch Size. The Number of Epochs represents the number of times the model iterates over the entire training set. A sufficient number of iterations ensures that the model sufficiently learns the patterns in the training data, but too many iterations may cause the model to overfit. In the experiments, we set the number of iterations as 30, which is an empirically based choice designed to balance training time and model performance. The Optimizer determines the strategy for updating the model weights. In this experiment, we adopted the Adam optimization algorithm because it combines the momentum method and the first moment estimation of the gradient, which is generally able to provide good performance in a variety of different optimization problems.

In addition, to mitigate the possible overfitting of the model, we employ Dropout technique. By randomly dropping a part of the neurons in the network during training, Dropout effectively reduces the dependence of the model on specific training samples, thereby improving the generalization ability of the model. In this experiment, we apply Dropout appropriately to ensure that the model can maintain good prediction performance even on unseen data. With these meticulous hyperparameter tuning and policy application, we expect to be able to train a model that is both accurate and robust.

Table 1. Setting of experimental hyperparameters

Parameter	Value
Learning Rate	0.001
Batch Size	128
Epochs	30
Optimizer	Adam
Dropout	0.5

4.3. Ablation Experiment

In order to analyze the gain of different network modules for model checking, ablation experiments are carried out for verification. Firstly, the Baseline models MVIT and MVIT_R are used as the Baseline, and then the ASFF adaptive module is introduced in the spatial domain and the FECAM frequency enhancement module is introduced in the frequency domain to discuss the gain of each part on the accuracy, and the influence of different frameworks is explored. Keep the parameters and other conditions consistent for each experiment. The model was trained and tested on the FF++, Celeb-DFV2 and BioDeepAV datasets respectively, and ACC (Video level) was used as the evaluation index. The final experimental results are shown in the table.

As can be seen from the table, the accuracy of detection generally keeps increasing with the combination of modules. Whether MVIT or MVIT_R is used as the skeleton, their detection accuracy is generally improved after adding different modules. This indicates that the module of this chapter is effective for enhancing the accuracy of forgery detection.

On the MVIT architecture, after adding the ASFF module, the accuracy on DeepFakes dataset is improved from 90.94% to 91.55%, and FaceSwap is improved from 90.29% to 91.82%. The accuracy on Celeb-DFV2 and BioDeepAV datasets is improved from 92.19% to 95.98% and 75.92% to 77.73%, respectively. When the FECAM module is added alone, the accuracy on DeepFakes dataset is significantly improved to 95.46%, FaceSwap is improved from 92.32% to 94.19%, and the accuracy on Celeb-DFV2 and BioDeepAV datasets is improved to 95.89% and 80.78%, respectively. For

the MVIT_R architecture acting on modules alone, we can observe a similar trend. It is particularly worth noting that when the MVIT_R skeleton combines the two modules ASFF and FECAM, the accuracy on DeepFakes, FaceSwap and Celeb-DFV2 datasets reaches the highest 97.89%, 98.11% and 96.23%. The accuracy on BioDeepAV dataset reaches 83.89%. This indicates that the fusion of multiple modules can produce a synergistic effect to achieve better detection performance.

Table 2. Ablation experiment result

Method				DeepFakes	FaceSwap	Celeb-DFV2	BioDeepAV
MVIT	MVIT_R	ASFF	FECAM	ACC	ACC	ACC	ACC
√				90.94%	90.29%	92.19%	75.92%
√		√		91.55%	92.32%	95.98%	77.73%
√			√	95.46%	94.19%	95.89%	80.78%
√		√	√	97.32%	96.88%	96.02%	**84.63%**
	√			93.54%	93.86%	91.26%	77.05%
	√	√		91.25%	94.28%	94.23%	79.91%
	√		√	96.46%	96.48%	95.51%	78.59%
	√	√	√	**97.89%**	**98.11%**	**96.23%**	83.89%

The results show that MVIT and MVIT_R, as two different frameworks, can significantly improve the detection accuracy of deepfake videos after combining ASFF and FECAM modules.

4.4. Cross Compression Experiment

Most of the existing deepfake detection models are trained on finely constructed high-quality datasets. However, in practical applications, fake face videos are often compressed in the process of generation and dissemination on social media. This process will lead to the loss of a large number of details during the coding and compression of the image, making the forged traces become unclear. Therefore, when these models trained on high-quality datasets are applied to detect low-quality data, their accuracy often degrades significantly. Cross-compression ratio test is a challenge often encountered in practical detection environments, and it is also a key indicator to test the generalization ability of the algorithm. To evaluate the performance of our model in dealing with compression, we conduct a series of cross-compression experiments and compare it with the state of the art deepfake detection techniques. In terms of experimental data selection, we chose to train the model on the high-quality C23 compression rate FF++ dataset. The trained model is used to test DeepFakes (DF), FaceSwap (FS), Face2Face (F2F) and NeuralTextures (NT) on the FF++ dataset with low quality C40 compression rate. In this section, the video-level accuracy Acc is used as the standard to measure the performance, and the specific experimental results are shown in Table 3.

Table 3. Results of the cross-compression experiment[20]

	Pretrain on FF++ C23				
Model	DF (C40)	FS(C40)	F2F(C40)	NT(C40)	AVG
Xception	63.60%	55.83%	50.05%	55.22%	56.18%
Nguyen	67.46%	54.23%	55.90%	53.85%	57.86%
MesoNet	78.39%	59.89%	69.47%	53.94%	65.42%
CN-RN	71.97%	52.29%	50.26%	51.50%	56.51%
3D-CNN	79.99%	54.65%	69.69%	59.33%	65.92%
CrossV	63.50%	67.53%	50.85%	66.24%	62.03%
MADD	88.96%	83.88%	62.04%	52.87%	71.94%
SFDT	93.47%	86.01%	75.11%	63.68%	79.57%

According to the data in the table, the SFDT method has shown significant advantages in cross-compression (C23-C40) detection experiments, and its accuracy has been greatly improved in four different FF++ datasets. Compared with other methods in the table, SFDT achieves 93.47%, 86.01%, and 75.11% cross-compression accuracy on DeepFakes, FaceSwap, and Face2Face, respectively, which are the highest values. Compared with the second-best method MADD, SFDT improves the first three tasks by 4.51 percentage points, 2.13 percentage points, and 13.07 percentage points, respectively. On average, SFDT achieves an average improvement of 8.52 percentage points on the four tasks. This result shows that SFDT has higher cross-compression detection performance when dealing with different types of deepfake technologies, benefiting from the blessing of fusing spatial-frequency domain features.

4.5. Robustness Experiments

Since deepfake videos can be easily modified on various media platforms, and deepfake detectors are highly vulnerable to interference from adversarial attacks (deepfake detection adversarial robustness enhancement algorithm), Therefore, the robustness of the model is detected by adding Saturation, Contrast, Noise, Blur, Pixel and so on to the Deepfakes fake video data set in F++. These jamming methods can effectively simulate the natural interference or human damage that face forgery images may encounter in real detection scenarios. The corresponding interference visualization used in the experiment is shown in the figure:

In order to deeply verify the advantages of the proposed algorithm in this chapter in terms of robustness, this section designs a comparison experiment with the robustness test of mainstream algorithms. In view of the fact that previous deepfake research has paid relatively little attention to robustness testing and lacks comparable data, this chapter selected mainstream face forgery detection models, including Xception, FaceXray, LipForensics, CADDM, LAA-Net, as the comparison object. These models are pre-trained on a standard dataset of DeepFakes and subsequently evaluated for

performance on perturbed data. In this chapter, Acc (Video level) is used as the evaluation index in the performance test on the standard data

Fig 6. Disturbance visualization

Table 4. Robustness experiments result[21]

Method	Fake	Saturation	Contrast	Noise	Blur	Pixel
Xception	✓	99.31%	98.62%	53.85%	60.22%	74.27%
FaceXray	✓	97.62%	88.55%	49.83%	63.86%	88.62%
LipForensics	✓	99.92%	99.61%	73.89%	96.14%	95.63%
CADDM	✓	99.63%	99.86%	87.41%	99.03%	98.86%
LAA-Net	✓	99.95%	99.92%	53.96%	98.22%	99.81%
SFDT	✓	99.97%	99.87%	89.27%	98.48%	99.91%

Compared with other algorithms, the color-related changes such as saturation and contrast have little impact on the overall detection results, and the proposed SFDT algorithm achieves 99.9% and 99.8% accuracy. Compared with CADDM method, it improves the noise processing by 1.8%. It achieves 98.4% accuracy in fuzzy processing, which still has room for improvement compared with the mainstream algorithms. The pixel perturbation has a weak impact on the detection as a whole, and the proposed method achieves an accuracy of 99.9%, which proves that the method in this chapter has certain advantages in the robustness test of forged videos.

4.6. Visual Analysis

In order to visually show how the proposed detection model makes decisions, this chapter visualizes the classification results of the input face image by the detection model and the final feature vector input to the classifier, revealing the basis for the detection model to determine the truth or falsehood. This is exactly the face region that the detection model focuses on when processing the input samples, and it is also the region that the model thinks may have deepfake traces. In this chapter, the model visualization analysis is carried out by Grad cam[22].

CelebDF-v2 has superior forgery methods compared with other datasets. It uses advanced Generative Adversarial Network (GAN) technology to generate high-quality fake videos covering multiple well-known public figures, which ensures the diversity and representation of the dataset and is more difficult to be detected by the naked eyes. Therefore, the forged data in the CelebDF-v2 dataset are selected for testing. By showing the activation of different channels downstream of the model, we can observe that the model smooths the skin texture, and pays more attention to the key edge parts of the face downstream, revealing possible regions of forgery traces. As shown in Figure 7:

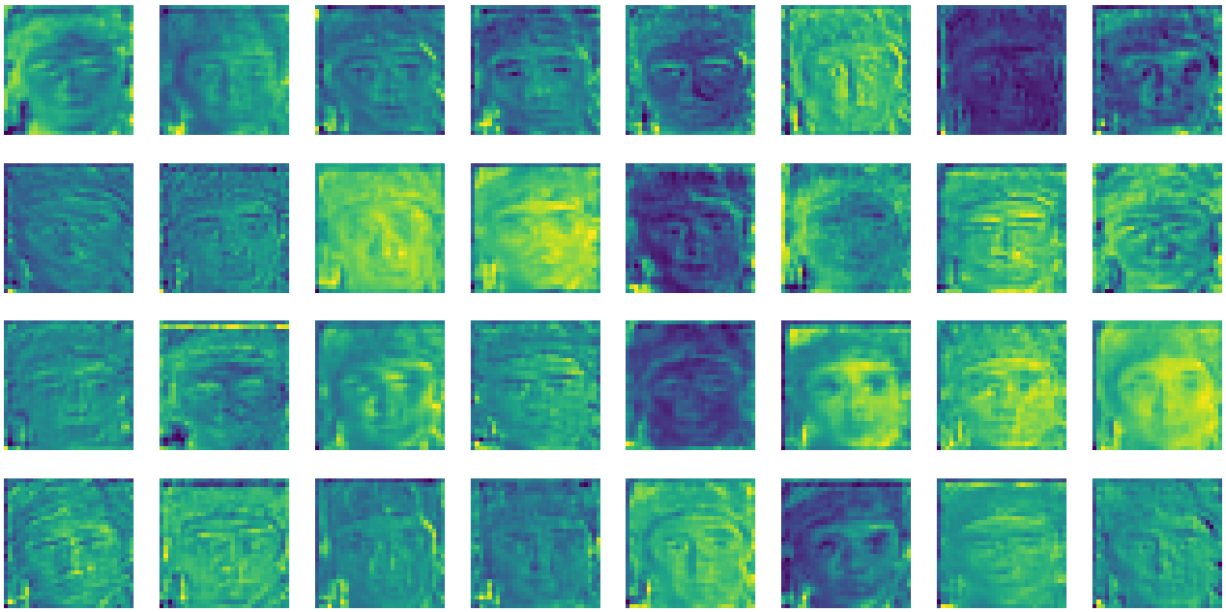

Fig 7. Activation map visualization downstream of the model

Grad cam is usually used to explain the partial feature contribution of a single instance, and it can show the contribution of the features in the space to the prediction of the current sample through the heat map. With this partial visualization, it's possible to see at a glance which features play a key role in individual predictions.

Fig 8. Falsified data with SFDT area of concern visualization

As can be observed from the figure, for different samples, the SFDT detection model proposed in this chapter accurately focuses on the fake facial regions, such as the lower jaw, the forehead, the nose, the border after the face change, etc. It is easy to see that the forgery regions focused on by the proposed model in deepfake detection are more detailed and accurate, and can almost perfectly correspond to the information abnormal regions in the input samples. The visual experimental results fully prove the reliability and persuadability of the detection model.

5. Conclusion

In this paper, to enhance the robustness of deepfake detection model, this paper proposes a detection algorithm that fuses spatial-frequency domain features. The algorithm combines spatial domain and frequency domain information to improve the detection effect of cross-compression and anti-interference. Firstly, this paper expounds the research motivation of the algorithm, and describes the overall process of the model, the spatial domain feature extraction part, the frequency domain information extraction part, and the CMF feature fusion module in detail in the overall algorithm architecture. Through a series of experiments to verify the performance of the algorithm in this chapter, the influence of each module is analyzed through ablation experiments, the transferability evaluation on cross-data sets, the cross-compression ability test under different compression rates, the model is presented in the form of visualization, which shows the superiority of the algorithm in the generalization ability. robustness test of anti-interference, and the attention to the

References

[1] Mohammad Rami Koujan, Michail Christos Doukas, Anastasios Roussos, and Stefanos Zafeiriou. 2020. Head2head: Video-based neural head synthesis. arXiv preprint arXiv: 2005. 10954 (2020). Mohammad Rami Koujan, Michail Christos Doukas, Anastasios Roussos, and Stefanos Zafeiriou. 2020. Head2head: Video-based neural head synthesis. arXiv preprint arXiv:2005.10954.

[2] Goodfellow I, Pouget-Abadie J, Mirza M, et al. Generative adversarial networks[J]. Communications of the ACM, 2020, 63(11): 139-144.

[3] Rana M S, Nobi M N, Murali B, et al. Deepfake detection: A systematic literature review[J]. IEEE access, 2022, 10: 25494-25513.

[4] Suthaharan S, Suthaharan S. Support vector machine[J]. Machine learning models and algorithms for big data classification: thinking with examples for effective learning, 2016: 207-235.

[5] Afchar D, Nozick V, Yamagishi J, et al. Mesonet: a compact facial video forgery detection network[C]//2018 IEEE international workshop on information forensics and security (WIFS). IEEE, 2018: 1-7.

[6] Lu X, Firoozeh Abolhasani Zadeh Y A. Deep Learning-Based Classification for Melanoma Detection Using XceptionNet [J]. Journal of Healthcare Engineering, 2022, 2022(1): 2196096.

[7] Sun B, Liu G, Yuan Y. F3-Net: Multiview scene matching for drone-based geo-localization[J]. IEEE Transactions on Geoscience and Remote Sensing, 2023, 61: 1-11.

[8] Yu Y, Si X, Hu C, et al. A review of recurrent neural networks: LSTM cells and network architectures[J]. Neural computation, 2019, 31(7): 1235-1270.

[9] Fan H, Xiong B, Mangalam K, et al. Multiscale vision transformers[C]//Proceedings of the IEEE/CVF international conference on computer vision. 2021: 6824-6835.

[10] Chen M, Peng H, Fu J, et al. Autoformer: Searching transformers for visual recognition[C]//Proceedings of the IEEE/CVF international conference on computer vision. 2021: 12270-12280.

[11] Wen G, Li Z, Azizzadenesheli K, et al. U-FNO—An enhanced Fourier neural operator-based deep-learning model for multiphase flow[J]. Advances in Water Resources, 2022, 163: 104180.

[12] Men Y, Yao Y, Cui M, et al. Dct-net: domain-calibrated translation for portrait stylization[J]. ACM Transactions on Graphics (TOG), 2022, 41(4): 1-9.

[13] Jiang M, Zeng P, Wang K, et al. FECAM: Frequency enhanced channel attention mechanism for time series forecasting[J]. Advanced Engineering Informatics, 2023, 58: 102158.

[14] Testa R L, Machado-Lima A, Nunes F L S. Deepfake detection on videos based on ratio images[C]//2022 12th International Congress on Advanced Applied Informatics (IIAI-AAI). IEEE, 2022: 403-408.

[15] Rossler A, Cozzolino D, Verdoliva L, et al. Faceforensics++: Learning to detect manipulated facial images[C]// Proceedings

of the IEEE/CVF international conference on computer vision. 2019: 1-11.

[16] Li, Yuezun, et al. "Celeb-df: A large-scale challenging dataset for deepfake forensics." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020.

[17] Jiang, Liming, et al. "Deeperforensics-1.0: A large-scale dataset for real-world face forgery detection." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020.

[18] Li, Lingzhi, et al. "Faceshifter: Towards high fidelity and occlusion aware face swapping." arXiv preprint arXiv: 1912. 13457 (2019).

[19] Croitoru F A, Hiji A I, Hondru V, et al. Deepfake Media Generation and Detection in the Generative AI Era: A Survey and Outlook[J]. arXiv preprint arXiv:2411.19537, 2024.

[20] Wan Da. Research on face forgery image detection method based on deep learning [D]. People's Public Security University of China,2024.DOI:10. 27634/d. cnki. gzrgu. 2024. 000268.

[21] Nguyen D, Mejri N, Singh I P, et al. Laa-net: Localized artifact attention network for quality-agnostic and generalizable deepfake detection[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024: 17395-17405.

[22] Selvaraju R R, Cogswell M, Das A, et al. Grad-cam: Visual explanations from deep networks via gradient-based localization[C]//Proceedings of the IEEE international conference on computer vision. 2017: 618-626.